BIG DATA IN RELIABILITY

Vianney Bordeau, François Escudié, Gilles Debache, Martin Le Loc

BIG DATA IN RELIABILITY

Vianney Bordeau [1], François Escudié [2], Gilles Debache [3], Martin Le Loc [4]

[1] RATP, Fontenay-sous-Bois, France
[2] LGM, Vélizy-Villacoublay, France
[3] Dassault Aviation, Saint-Cloud, France
[4] QUANTMETRY, Paris, France
vianney.bordeau@ratp.fr
01 58 78 30 41

Mots-clés : *Big Data, Fiabilité, Retour d'expérience ; Maintenance prévisionnelle*

Résumé

Cette communication présente les travaux du projet IMdR P17-4 « Big Data in Reliability » qui s'est déroulé entre 2018 et 2019, et auquel ont participé les 11 souscripteurs suivants : ALSTOM, DASSAULT AVIATION, DGA, EDF, GRTgaz, INRS, IRSN, NEXTER, RATP, SNCF et THALES. Ce projet est né à la suite d'échanges au sein du groupe de travail « Retour d'expérience technique » de l'IMdR piloté par N. Dechy (IRSN) et E. Rémy (EDF), et à l'initiative de A. Lannoy (IMdR) à la suite du constat d'une démocratisation des nouvelles technologies liées au « digital » ayant bouleversé les méthodes traditionnelles de création de valeur des entreprises.

Summary

This communication presents the work and results of the IMdR project P17-4: "Big Data in Reliability" which took place between 2018 and 2019 and for which the 11 subscribers participated: ALSTOM, DASSAULT AVIATION, DGA, EDF, GRTgaz, INRS, IRSN, NEXTER, RATP, SNCF and THALES. This project was born as a result of discussions within the IMdR's "Technical feedback" working group led by N. Dechy (IRSN) and E. Rémy (EDF), and at the initiative of A Lannoy (IMdR) following the observation of a democratization of new technologies linked to "digital" having upset traditional methods of value creation in companies.

1. Introduction

Les outils digitaux se sont insérés dans de très nombreux processus métier du quotidien et ont donc créé de la donnée d'utilisation. Cette information produite est depuis devenue une ressource importante dans la recherche de l'excellence opérationnelle. Les entreprises désirent maintenant optimiser chacun des maillons de leurs chaînes de valeur. Cet objectif de performance a engendré une dynamique d'émulation responsable à la fois d'une augmentation des performances purement techniques du matériel informatique (en termes de capacité de stockage et de vitesse de calcul) ainsi qu'une réduction des coûts nécessaires à son implémentation.

Conjugué à la baisse en général du coût des capteurs, les acteurs industriels se sont donc retrouvés avec la capacité de stocker massivement des données d'intérêt puis de les analyser dans le cadre de leurs activités de fiabilité et de maintenabilité (ex. : conditions d'exploitation du matériel, détection de franchissement d'un seuil d'alerte par un signal, etc.) notamment grâce à l'implémentation des fonctions HUMS (Health and Usage Monitoring System).

Le contexte d'analyse de risque et de modélisation de la sûreté de fonctionnement va donc se trouver fortement bouleversé. Le retour d'expérience (sa collecte et son traitement) est le premier affecté et va profondément influer sur toutes les autres thématiques de la maîtrise des risques et de la sûreté de fonctionnement. Ces facilités technologiques, une collecte plus facile, de nouveaux outils de traitement vont nous faire passer d'une période au trop peu de données où la modélisation est plutôt statique, à une période au trop plein de données où une approche dynamique du comportement fiabiliste peut être envisagée. On passe d'un retour d'expérience statique à un retour d'expérience dynamique.

Il est probable que ces nouveaux outils vont nous conduire à une meilleure connaissance du profil d'usage (ou profil de fonctionnement), des dégradations de toute nature, des défaillances, des incidents et des dysfonctionnements. Ce bouleversement peut cependant remettre en cause, sous certains aspects, tous les modèles et toutes les approches actuellement utilisés, notamment les pratiques de validation des données brutes. Les pratiques industrielles issues de l'expérience et acceptées par les autorités de tutelle peuvent se trouver impactées.

Ce projet a visé à mettre en évidence les conséquences de ce bouleversement et les pistes de R&D qu'il conviendrait d'engager dès à présent. Il a reposé sur quatre principaux centres d'intérêt : méthodes Big data, estimation de la fiabilité opérationnelle et maintenance prévisionnelle, anticipation et confiance dans les résultats.

1.1. Démarche et analyse du besoin des souscripteurs

L'analyse des besoins des souscripteurs, via questionnaire et entretiens, a permis de mieux connaître la situation de chacun vis-à-vis de l'utilisation du Big Data et ainsi de mieux comprendre les attentes pour la réalisation de l'étude. Le tableau ci-dessous résume les points abordés et dans quelles tâches du projet ces points seront traités. Cette étape a été très importante pour accorder les souscripteurs et les rédacteurs sur un sujet nouveau et vaste avec des niveaux de maturité dans cette technologie très différents entre souscripteurs. Les besoins qui ressortent correspondent bien au Cahier des Charges, avec en plus un point abordé sur la sécurité des données et leur partage avec des prestataires externes.

Tâches de l'étude	Thèmes à traiter
	Règles de bonnes pratiques
	Exemple d'application dans l'industrie
	Technique de traitement des données (sélection d'échantillon, exploration des données,…)
	Limite de l'exploitation des données
	Comprendre les algorithmes, Intelligibilité des modèles, interpréter les résultats
	Fiabilité mécanique et électronique
	Lien entre Big Data et estimation de la fiabilité
	Automatisation du Rex, nouvelles approche et méthodes pour le Rex
	Maintenance prédictive/prévisionnelle
	Organisation des données
	Sensibiliser à l'intérêt de la qualité des données
	Standardisation des process
	Pouvoir gérer la confidentialité des données
	Qualité des données
T1 : Etat de l'art	
T2 : Impact du Big Data	
T3 : Impact du Big Data sur l'estimation de la fiabilité	
T4 : Impact du Big Data sur les méthodes de maintenance	
T5 : impact du Big Data sur l'aide à la décision	
T6 : Organisation et impact	

	Partie où le sujet sera principalement abordé
	Partie où le sujet sera traité mais moins en profondeur

2. Etat de l'art

2.1. Pourquoi parle-t-on de big data ?

La démocratisation des nouvelles technologies liées au « digital » a bouleversé les méthodes traditionnelles de création de valeur des entreprises. En effet, les outils digitaux se sont insérés dans de très nombreux processus métier du quotidien et ont donc créé de la donnée d'utilisation. Cette information produite est depuis devenue une ressource importante dans la recherche de l'excellence opérationnelle. Les entreprises désirent maintenant optimiser chacun des maillons de leurs chaînes de valeur. Cet objectif de performance a engendré une dynamique d'émulation responsable à la fois d'une augmentation des performances purement techniques du matériel informatique (en termes de capacité de stockage et de vitesse de calcul) ainsi que d'une réduction des coûts nécessaires à son implémentation.

Conjugué à la baisse en général du coût des capteurs, les acteurs industriels se sont donc retrouvés avec la capacité de stocker massivement des données d'intérêt puis de les analyser dans le cadre de leurs activités de fiabilité et maintenabilité (ex. : conditions d'exploitation du matériel, détection de franchissement d'un seuil d'alerte par un signal, etc.) notamment grâce à l'implémentation des fonctions HUMS (Health and Usage Monitoring System).

Cette explosion exponentielle du nombre de données à disposition des entreprises est à l'origine même de la création du terme Big Data. La littérature résumait ce phénomène à travers ce qu'on appelait communément « les 3V du Big Data » pour Volume, Vélocité, Variété. On entend aujourd'hui souvent parler des « 7V du Big Data » avec les ajouts des concepts de Variabilité, Véracité, Visualisation et Valeur. Certains auteurs contemporains continuent encore de proposer de nouveaux V pour lire la transformation Big Data avec « les 10V du Big Data » (Cf. Figure 1) comprenant en plus la Vulnérabilité (notion de cyber-sécurité), la Volatilité (péremption de l'intérêt d'une donnée) et la Validité (qualité de la donnée).

Figure 1 – Principales caractéristiques du BIG DATA

Indépendamment du nombre exact de V à considérer (croissant au fil du temps), l'essentiel à retenir est que le terme Big Data décrit un objet de nature protéiforme comprenant à la fois le stockage, l'infrastructure technique, les algorithmes de traitement et les informations produites. Au vu des tendances scientifiques actuelles, le prochain axe de lecture à inclure concernera très sûrement l'intelligibilité des modèles et donc implicitement leur éthique.

2.2. Et l'Intelligence Artificielle dans tout ça ?

La principale valeur ajoutée du Big Data est avant tout de réussir à créer un lien entre des données hétérogènes en apparence afin d'élargir sa vision d'ensemble d'un processus. Il est important de garder à l'esprit que le Big Data n'a pas vocation à « prédire » l'avenir. Sa plus-value principale réside dans l'analyse d'un grand volume de données brutes par la recherche de règles implicites en son sein grâce à une approche probabiliste.

La machine est capable de déceler et quantifier des corrélations entre des données produisant ainsi une nouvelle information mais elle ne maîtrise pas le concept de causalité. Les systèmes d'IA utilisent aujourd'hui une approche dite connexionniste avec une logique inductive peu explicable par opposition à une approche dite symbolique où chaque étape de la démarche logique serait démontrable. Afin d'illustrer ces divergences entre les démarches, prenons à chaque fois un exemple commun où l'on chercherait à identifier un opérateur au sein d'une image, suivi d'un exemple plus proche du monde industriel.

- l'approche connexionniste donnerait pour résultat la probabilité que l'image contienne un humain basée sur son expérience de ce qu'était un homme ou non (sa donnée d'entraînement composé de milliers de photo d'hommes). La réponse de l'algorithme ne sera pas argumentée logiquement et fortement dépendante de son entraînement. Avec la même méthodologie, la détection de défauts d'un roulement à billes peut-être automatisée grâce à un apprentissage basé sur des analyses vibratoires. L'algorithme extrait des caractéristiques vibratoires depuis le signal brut et identifie ensuite les défauts possibles (fissures des bagues extérieures ou intérieures et corrosion des billes par piqûre). Les comportements chaotiques ont ainsi pu être fortement corrélés avec un défaut de la bague interne ou une corrosion des billes [1].
- l'approche symbolique répondrait oui ou non et justifierait sa réponse par la présence de tous les éléments logiques permettant d'identifier avec certitude un humain (visage, morphologie, cheveux, vêtements, etc.). La réponse fournie est directement interprétable car on connaît explicitement la chaîne logique pour y parvenir. Cela présuppose donc une formalisation rigoureuse de l'environnement et des perturbations en son sein. Une approche de système expert afin de classifier des perturbations transitoires du réseau électrique est décrite dans P.K. Dash *et al.* [2].

La distinction entre ces deux approches permet d'expliquer l'une des caractéristiques essentielles aujourd'hui pour implémenter l'intelligence artificielle connexionniste et comprendre son succès : il n'y a pas besoin d'implémenter de règles complexes décrivant formellement le monde réel. On parle d'apprentissage automatique, ou machine learning, pour décrire la capacité du système à extraire des modèles à partir de données brutes. On s'appuie alors sur une modélisation dite data-driven par opposition avec une approche de simulation numérique qui serait basée sur un modèle formel physique avec des équations décrivant le système.

L'intelligence artificielle est un terme aujourd'hui largement plébiscité par les journaux et dans le discours marketing des entreprises. Nous sommes encore très loin du fantasme hollywoodien du « Terminator » et de son comportement autonome. Elle prend aujourd'hui la forme d'un programme informatique, capable de réaliser automatiquement des tâches de « faible complexité » pour un opérateur humain mais souvent extrêmement chronophages car répétitives. L'automatisation de ces tâches « bas niveau » procure alors un gain de productivité évident pour l'employé en lui permettant de se consacrer à des tâches avec une plus haute valeur ajoutée.

L'implémentation de l'IA est étroitement liée au Big Data car il est en mesure de lui fournir les informations dont elle a besoin pour « prendre » une décision (à savoir obéir à des règles métiers implémentées par l'informaticien). Une fois créé, le programme est capable d'ingérer régulièrement des quantités de données massives avant d'exécuter la tâche pour laquelle il a été spécifiquement défini.

2.3. L'émergence d'un besoin d'intelligibilité des modèles

Les propositions de solution d'un algorithme de Deep Learning sont basées sur des corrélations comprises au sein des données d'entrée et suivies d'une optimisation mathématique de la vraisemblance de la prédiction ou prévision. Les volumes de données concernés sont souvent bien trop importants pour qu'un homme puisse en réaliser une synthèse similaire en un temps raisonnable. Il faut donc prêter attention aux questions d'intelligibilité de l'algorithme employé, soit notre « capacité à expliquer ou à présenter son fonctionnement en des termes compréhensibles par l'humain ». En effet, des questions essentielles se cachent derrière: quel degré d'autonomie sommes-nous prêts à déléguer à la machine ? Comment se fier à la sortie obtenue pour prendre une décision ?

Il existe en réalité deux niveaux d'interprétabilité, un local et un global. L'interprétabilité globale vise à donner des indications sur comment les différentes variables d'entrée vont impacter les prédictions du modèle, alors que l'interprétabilité locale se joue à l'échelle individuelle (un individu, une machine, etc.) et vise à donner des indications sur comment les valeurs prises par les variables d'entrée impactent une prédiction isolée. Ces notions sont faciles à appréhender sur un modèle purement linéaire mais sont loin d'être évidentes dans un réseau de neurones composés de plusieurs couches interagissant entre elles… Ainsi plus un modèle est complexe, plus il va falloir de travail pour le rendre intelligible. Afin de réussir à surmonter cet obstacle d'intelligibilité, de nombreuses initiatives scientifiques sont actuellement en cours de développement, notamment les méthodes LIME (Local Interpretable Model-agnostic Explanation) et les méthodes SHAP (SHapley Additive Explanation).

2.4. La méthodologie d'un projet data

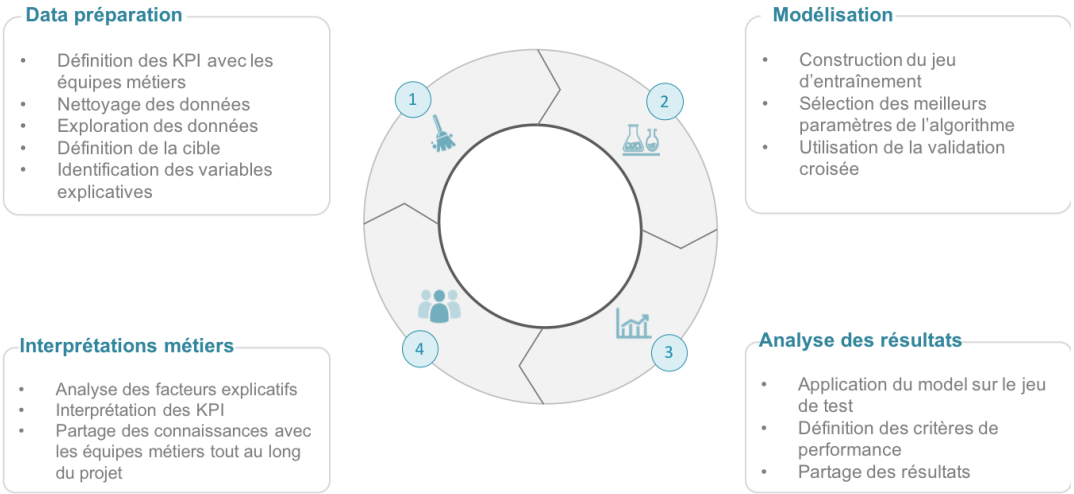


Figure 2 – Le processus itératif du projet data

Le développement d'une réponse basée sur l'exploitation de la donnée à une problématique métier est souvent mené au sein d'un processus itératif qui touche autant les métiers que les équipes techniques/data, de l'ingestion de la donnée à la restitution des résultats :

1. **Préparation des données** : les métiers définissent un axe d'étude, formulent une problématique. Les équipes data récupèrent la donnée susceptible de répondre à cette problématique, la nettoient et l'explorent. Le but est de construire un pont entre des problématiques métier concrètes et la théorie mathématique / informatique. Ceci passe par le cadrage du problème, la définition des méthodes à appliquer et d'une métrique objective pour évaluer les performances de la future réponse théorique, la création de valeur en créant de nouvelles variables à partir de l'existant.

2. **Modélisation** : il s'agit de la phase de modélisation statistique. Sur un jeu de données dit d'« entraînement » (i.e. jeu de données d'où l'on tire les exemples sur lesquels notre algorithme va construire son savoir), une méthode sélectionnée en

amont est appliquée, et ses différents paramètres sont réglés afin d'optimiser le critère d'évaluation défini dans la tâche 1. Des recommandations existent pour la sélection des jeux de données d'entraînement, de validation ou de test [3].

3. **Analyse des résultats** : l'algorithme entraîné à l'étape précédente est appliqué à un jeu de données dit de « test » (i.e. jeu de données sur lequel l'algorithme n'a pas appris). De manière générale, on évalue la capacité de l'algorithme à généraliser son apprentissage sur le jeu d'entraînement au jeu de test. Les résultats sont ensuite mis en forme pour un partage avec les métiers (visualisation, vulgarisation des méthodes).

4. **Interprétation métiers** : évaluation de la pertinence des développements réalisés à partir de la donnée, et réflexion sur des pistes d'améliorations.

5. **Itérations** : ce processus est généralement itératif, dans le but de créer une amélioration continue de la réponse aux métiers. A chaque nouvelle boucle, une nouvelle source de données peut être intégrée, la métrique d'évaluation peut être affinée, un nouvel algorithme peut être testé, etc.

Ainsi, l'état de l'art de la fiabilité portera surtout sur la partie 2 (Modélisation) de ce processus itératif.

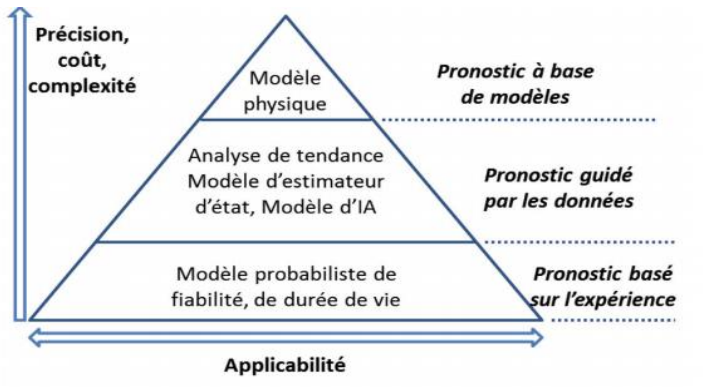


Figure 4 – Typologies des approches de pronostic
extraite de M. Weinechter, IMdR 2018 [4]

Dans le cadre de la fiabilité, ce type de réponse guidé par les données correspond à la partie centrale de la pyramide ci-dessus, entre le modèle physique et le modèle de durée de vie. Des approches hybrides sont bien sûr possibles et peuvent figurer dans l'état de l'art de la fiabilité [5].

2.5. Travaux de normalisation du big data

Les enjeux de la normalisation volontaire dans le domaine du Big Data sont importants pour définir un vocabulaire commun, une architecture fonctionnelle et technique, ou des standards favorisant l'interopérabilité des processus, un langage pour décrire les métadonnées adapté à divers métiers ou secteurs, ou encore décrire les rôles associés à la gouvernance de la donnée.

Les travaux de normalisation sur le Big Data et le Cloud Computing ont été initialisés par le National Institute of Standards and Technology pour les USA et par l'Union internationale des télécommunications pour les Nations unies. Ils sont relayés au sein de l'ISO/CEI par le JTC 1 (Joint Technical Committee 1), dont les travaux sont présentés dans son rapport préliminaire de 2014 [6], et sont suivis en France par l'AFNOR et son Comité stratégique information et communication numérique (COS ICN). Ce dernier a publié en 2015 un Livre blanc [7] qui dresse un état des lieux et une cartographie des travaux de normalisation pour le domaine du Big Data et surtout les attentes et enjeux de ces travaux pour les acteurs économiques et publics.

Les travaux de normalisation du Big Data sont principalement organisés autour de six axes :

- la gouvernance de la donnée,
- la qualité et l'identification,
- les données ouvertes (open data),
- les opérateurs d'infrastructures,
- les opérateurs de service,
- la normalisation technique.

Pour ce qui est de l'IA, l'AFNOR, toujours via son COS ICN, a publié l'an dernier (avril 2018) un autre Livre blanc dédié à la normalisation : [8] L'impact et les attentes pour la normalisation dans l'Intelligence Artificielle.

Les travaux associés sont plus récents que ceux relatifs au Big Data et l'élaboration de normes volontaires dans ce domaine semble plus délicate du fait que l'IA est un domaine aux frontières larges et imprécises, avec des technologies variées, des enjeux complexes aux conséquences parfois critiques ([8] *ibid.* p.47) et doit prendre en compte les réglementations existantes.

Au sein de l'ISO les travaux sont dirigés par l'ISO/IEC JTC 1 qui a créé un sous-comité dédié à l'IA, le SC 42. L'AFNOR a constitué un groupe miroir à ce SC42 pour l'élaboration de textes à portée nationale pour répondre à des besoins à court terme des entreprises mettant en place des dispositifs avec de l'intelligence artificielle, en particulier sur les volets robustesse et éthique (en anticipation d'un volet européen et/ou réglementaire notamment), et participer aux travaux du SC42 pour asseoir une contribution française forte sur les segments où il est nécessaire de faire valoir les propositions françaises permettant de promouvoir les intérêts des acteurs français ([8] *ibid.* p.49).

Les travaux de l'ISO/IEC JTC 1 SC42 ont pour objectifs de clarifier :

- les fondamentaux de l'IA (les concepts, les définitions, le cadre d'architecture, les cas d'usage…),
- les composants caractéristiques de l'IA (les algorithmes, les données, les apprentissages, la validation, etc.),
- les méthodes et les techniques permettant de développer des solutions d'IA optimisées (le filtrage des données, le choix des algorithmes, la robustesse et la précision des modèles…),
- les différences avec la programmation classique et les difficultés induites,
- les méthodes de limitation des risques.

2.6. Conclusions de l'état de l'art

La généralisation de l'usage du numérique au sein des activités industrielles entraîne une croissance et une variété importantes des sources de données à disposition des industriels. La conjonction d'une baisse du coût des technologies nécessaires à leur traitement ainsi que de nouvelles techniques augmentant la capacité à les exploiter est une véritable opportunité à saisir pour améliorer les stratégies et les méthodes de fiabilité.

Afin de réussir cette transformation, il est important d'appréhender et de mettre en œuvre, en complément des pratiques actuelles, de nouvelles techniques disponibles (Big Data et techniques d'apprentissage machine notamment) dont l'état de l'art est développé dans cette étude.

Pour tirer pleinement parti de ces méthodes, l'amélioration de la fiabilité par le Big Data nécessite un partage de l'information entre tous les acteurs de la fiabilité, depuis la conception du produit jusqu'à son exploitation, en passant par sa fabrication. Afin d'y parvenir, nous recommandons d'engager en complément des actions individuelles une stratégie de partage systématique de données fournisseurs (tests de qualité), fabricants (protocoles de fabrication, mesure de la qualité en ligne, essais) et exploitants (usages, profils de fonctionnement, retours d'expérience).

3. Impact sur le REX et son analyse

Un des impacts majeurs de l'adoption du Big Data sur le retour d'expérience est de transformer la connaissance du fonctionnement et du dysfonctionnement d'un système d'un mode statique à un mode dynamique et continu.

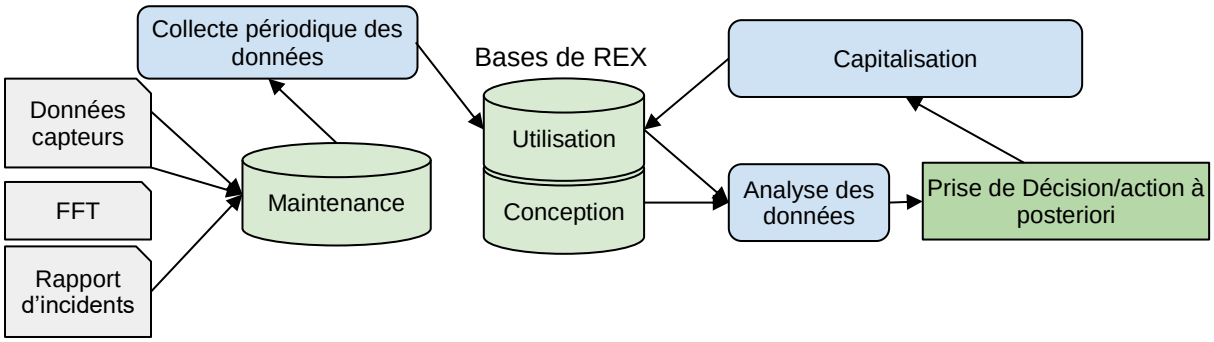


Figure 1 - Processus simplifié d'exploitation des données dans le cadre du REX traditionnel

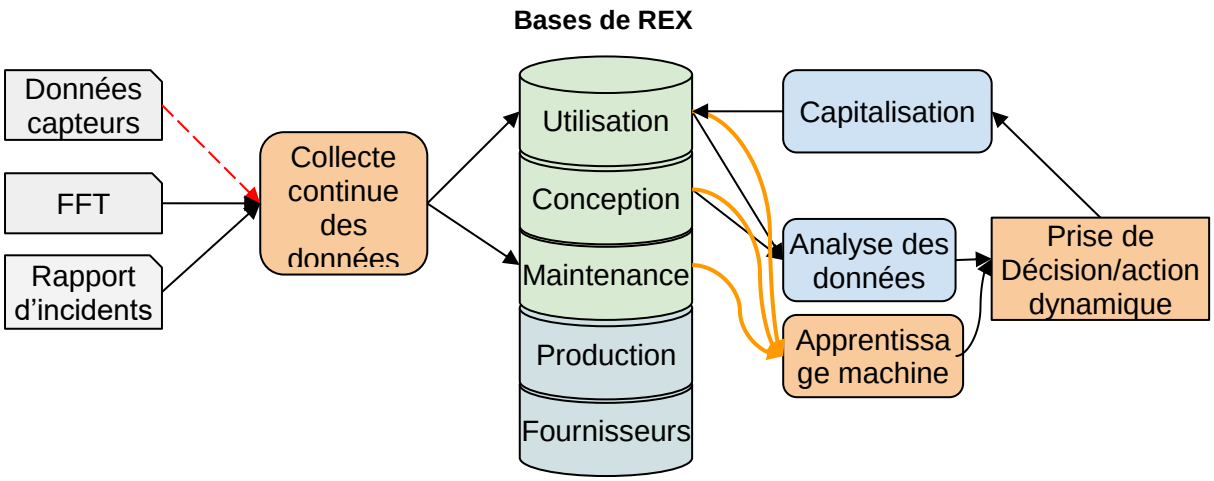


Figure 2 - Processus simplifié d'exploitation des données dans le cadre du REX en environnement *Big Data*

Les techniques de Big Data vont ainsi généraliser l'utilisation systématique d'un panel de données plus larges et plus fréquentes, voire l'utilisation de données jugées jusqu'alors inexploitables de par leur complexité à être rapprochées des données habituellement utilisées dans le cadre du retour d'expérience.

La connaissance dynamique et continue des systèmes va permettre à l'industriel ou à l'exploitant de modéliser un jumeau numérique de son système, pouvant aller jusqu'à un jumeau numérique individualisé de chacun de ses systèmes. L'objectif est de passer progressivement d'une approche globale de REX à une approche personnalisée et individualisée.

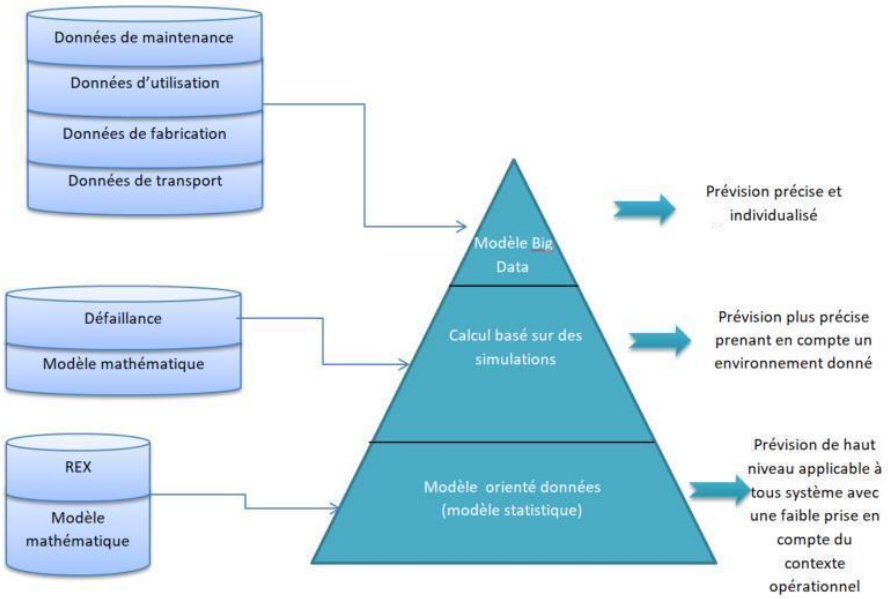


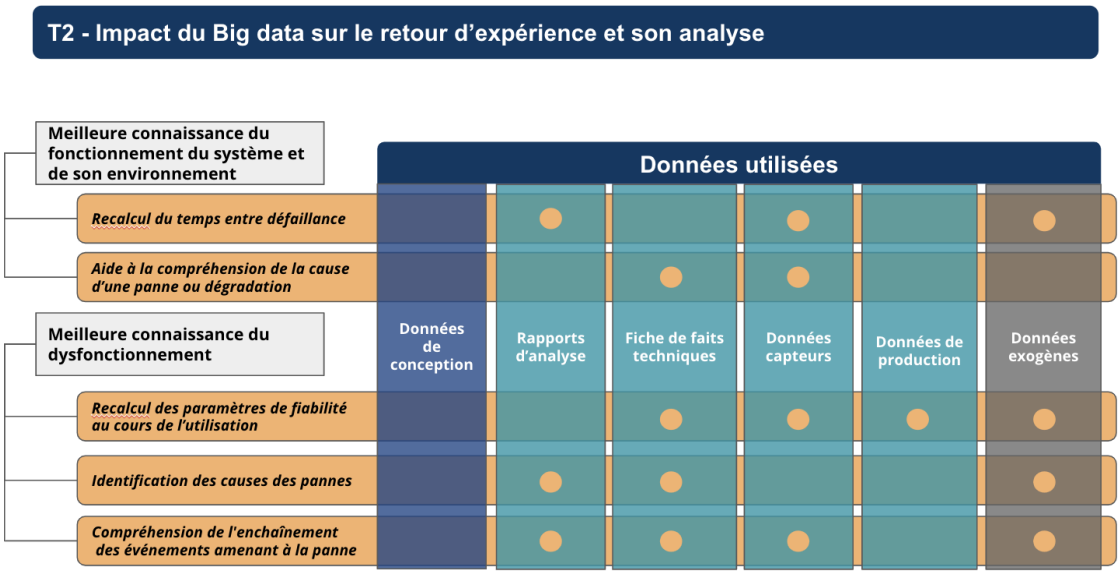
Figure 5 : un modèle orienté DATA du modèle Big Data

Le rapport du Projet de l'IMdR n°P15-2 "Health and Usage Monitoring System (HUMS) – Health Monitoring" (2017) concluait que les nouvelles technologies HUMS PHM généraient une importante quantité de données et qu'un futur challenge à relever serait le nettoyage, le traitement et le stockage de ces données.

La gouvernance de ces données de capteurs et de l'ensemble des données utilisées dans le retour d'expérience est ainsi identifiée aujourd'hui comme un prérequis indispensable à l'utilisation du Big Data et de l'Intelligence Artificielle.

Il reste nécessaire de souligner les freins (ou contraintes) ou les bénéfices (ou apports) que l'on peut rencontrer lors de la mise en place d'un projet de Big Data. Ainsi le rôle de l'expert dans une telle démarche s'avère primordial, que ce soit lors de la mise en place du projet (quelles données surveiller et collecter ? À quelles fréquences ?), ou lors de la phase d'analyse des données (labellisation du jeu de données d'entraînement par exemple). Ces nouveaux modes d'analyse nécessitent un investissement en ressources humaines et financières supérieur, tout du moins dans un premier temps, au niveau actuel requis par le REX.

Les cas d'usage :



4. Traitement d'un cas test

Un cas test a été réalisé dans le cadre du projet afin d'en illustrer les différentes tâches et opportunités.

L'objet de ce cas test, réalisé en partenariat avec Dassault Aviation, est la création de certains modules d'un outil d'aide à la décision destiné à évaluer la probabilité d'une pièce soupçonnée défaillante de l'être réellement, de façon à éviter la non-détection de défaut lors du renvoi au fournisseur (NFF).

Cette aide à la décision nécessite la conception et la mise en œuvre d'un algorithme supervisé de classification qui évalue la probabilité de NFF en fonction d'un certain nombre de critères issus de données connues (trames électroniques de l'avion, …). Certaines données disponibles sont structurées (données de description, données issues de capteurs…), d'autres non structurées (texte libre sur fiches scannées, comptes rendus de réparation…).

L'étude s'est concentrée sur la proposition de solutions pour répondre aux **deux contraintes majeures** dans la préparation des données :

- Exploitation de données non structurées : extraction d'informations depuis les comptes rendus de réparation "authorised release certificate",
- Rapprochement d'événements sans clés naturelles d'association : création de continuité numérique pour identifier et rapprocher les événements décrivant une défaillance et l'envoi en réparation de la pièce associée.

4.1. Exploitation de données non structurées issues de rapports numérisés

Les comptes rendus de réparation contenant de l'information utile pour caractériser et dater la défaillance et son évaluation par le fournisseur prenait la forme suivante (flouté pour raison de confidentialité) :

Grâce à l'utilisation de librairies open-source de prétraitement et d'extraction de texte (OpenCV et Tesseract notamment), six champs ont été extraits avec des précisions allant de 70 à 90%.

Une fois les champs extraits et grâce à des techniques d'analyse de texte, nous avons pu extraire de l'information complémentaire pour plus de 50% des défaillances.

Ces opérations d'extraction de données depuis des sources non structurées (image et texte) permettent d'enrichir le jeu de donnée d'apprentissage du modèle supervisé pour ne pas se limiter uniquement aux données issues des trames de l'avion.

4.2. Rapprochement d'événements sans clés naturelles d'association

Les variables explicatives du problème étaient contenues dans 8 fichiers différents. Chaque fichier contient des « informations unitaires » qu'il faut relier entre elles pour constituer des « macro-événements » décrivant une défaillance soupçonnée. Certaines liaisons entre fichiers sont définies par des clés de jointures claires et fiables (ex : rapprochement par l'identifiant de pièce) mais d'autres liaisons ne sont pas intuitives et nécessite un rapprochement temporel (ex : lien entre une défaillance observée dans une trame avion et sa déclaration par le client au service garantie).

Pour réaliser ces rapprochements et créer du lien entre des jeux de données distincts, nous avons utilisé un algorithme permettant la constitution de regroupements temporels (DBSCAN).

Ces opérations de rapprochement de données non triviales permettent d'enrichir les informations disponibles pour construire de nouvelles variables explicatives pertinentes dans l'apprentissage du modèle.

L'emphase donnée par l'étude sur ces deux techniques a pour objectif d'illustrer les capacités disponibles aujourd'hui pour exploiter des données jusqu'alors souvent ignorées car trop difficile d'exploitation (ici les rapports fournisseurs numérisés) ou trop difficile à rapprocher (ici les informations de garantie client) afin de répondre aux enjeux de la fiabilité.

5. Conclusion

Ce projet a d'abord permis de démystifier les champs du big data pour mieux en comprendre l'impact potentiel sur la maîtrise des risques. Il a permis de mettre en exergue qu'il n'y avait pas de réponse générique et unique, mais la nécessité de co construire, à l'image du « fait maison », ou « fait main ». Il a aussi mis en évidence l'énorme travail qu'il fallait fournir pour

conduire un projet big data et en remplir les conditions de son succès. En effet, il faut penser la conduite du changement autour des compétences, des besoins de partage des données, des structures des systèmes d'information, des ontologies, etc. Il y a par conséquent de nombreuses conditions de succès.

Le projet a convaincu l'ensemble des souscripteurs des bénéfices et intérêts en soulignant que la conduite d'un projet big data revêt une profondeur transformative sur les organisations afin d'obtenir les résultats souhaités. Ce n'est pas à traiter comme un « plug ».

La donnée n'est pas objective, tout à une sémantique ; tout ce qui est dénoté est connoté ; il y a des effets boîtes noires. Les experts ne sont pas en voie de disparition, bien au contraire, ils ont toute leur place dans ce changement et la traduction d'IA par Ingénieur Augmenté serait appropriée pour caractériser l'expert SdF immergé dans un projet big data.

Il y a des opportunités mais également des risques. Et c'est à travers ce genre de projet que pour l'IMdR les réflexions naissent et murissent avec ce fil conducteur qu'est la maîtrise des risques. Du reste plusieurs perspectives ont été identifiées pour prolonger cette étude :

- Etude du développement d'une méthode pour systématiser une étude de maintenance prévisionnelle en conception.
- Comment équiper en capteurs un système existant, quelles phases respecter, quelles exigences de fiabilité pour les capteurs ?
- Estimer et comprendre les différences et points communs entre approche de fiabilité traditionnelle et approche Big Data / IA pour les études de fiabilité prévisionnelle ?
- Estimer l'impact de l'utilisation de la méthodologie Big data sur le binôme expert - Data scientist : analyse de la situation de travail, impact du changement, rôle de chacun, efficacité.
- Assurer une veille sur la cartographie des normes et la standardisation en lien avec le Big Data ou l'IA et les applications pour la sûreté de fonctionnement.

Références

[1] P.K. Kankar *et al.*, *Fault diagnosis of ball bearings using machine learning models*, in Expert Systems with Applications, Volume 38, Issue 3, March 2011, 1876-1886.

[2] P.K. Dash *et al.*, *Classification of power system disturbances using a fuzzy expert system and a Fourier linear combiner*, in IEEE Transactions on Power Delivery, Volume: 15, Issue: 2, Apr 2000.

[3] T. Borovicka *et al.*, Selecting Representative Data Sets in "Advances in Data Mining Knowledge Discovery and Applications", edited by Adem Karahoca, 2012 - ISBN: 978-953-51-5710-6.

[4] *M. Weinechter*, R. Parouty, *HUMS, Health & Usage Monitoring System – état de l'art et opportunités,* IMdR Congrès Lambda Mu 21 (2018).

[5] Y. Fan *et al.*, *Transfer learning for remaining useful life prediction based on consensus self-organizing models*, 10/2019.

[6] ISO/IEC JTC 1 Information technology, Big data, Preliminary Report 2014.

[7] AFNOR, Comité stratégique information et communication numérique, *Livre blanc - Données massives - Big Data Impact et attentes pour la normalisation.*

[8] AFNOR, Comité stratégique information et communication numérique, *Livre blanc, l'impact et les attentes pour la normalisation dans l'Intelligence Artificielle*, 2018.