



ÉCOLE
CENTRALELYON



Big Data et Sport Synthèse bibliographique

BASSO-BERT Yanis
DETRE Louis
GOBERT Grégoire
HASSANALY Ravi
PIGE Alexis

Décembre 2019

Table des matières

1	Introduction	1
2	Les données sportives intéressantes	2
3	Big Data et sport professionnel	4
3.1	Influence sur la stratégie de jeu	4
3.2	Suivi physique des athlètes	5
4	Big Data et sport amateur	6
4.1	La data, l'avenir du sport amateur ?	6
4.2	La démocratisation des objets connectés et de la data	6
4.3	Les flux de data	7
4.4	La data, une valeur marchande	7
4.5	Utilisation de la data sportive	8
5	La Data au service de l'extra-sportif	9
5.1	Sport et réseaux sociaux	9
5.2	Diffusion des évènements sportifs	9
5.3	"Marketing" sportif	9
5.4	Côte et paris sportifs	10
6	Questionnements éthiques	11
6.1	Le RGPD - Règlement général sur la protection des données	11
6.2	La transparence de l'utilisation des données	12
6.3	L'enjeu sécuritaire	12
6.4	Gamification et fair-play sportif	12
7	Conclusion	13

1 Introduction

Lors des derniers Jeux olympiques de Rio en 2016, l'équipe britannique a remporté de nombreuses médailles, en partie grâce à un investissement massif dans l'analyse des données récoltées lors de l'entraînement [1]. De son côté, le ministère des Sports français a lancé le projet "Sciences 2024" [2], un plan d'investissement pour la recherche scientifique orientée vers la performance sportive, dans le but d'obtenir de meilleurs résultats sportifs lors des jeux olympiques de Paris en 2024. La récolte et l'analyse de grands volumes de données physiologiques des athlètes sont des axes de recherche de ce projet.

Depuis plusieurs années, la récolte et l'analyse de grands volumes de données (aussi appelé "Big Data") ont touché de nombreux domaines, et le domaine du Sport n'y a pas échappé, d'autant plus qu'il se prête particulièrement bien à cela [Section 2]. En effet, une place sur un podium se joue généralement à peu de choses et une amélioration, ne serait-ce qu'infime, des performances réalisées grâce à ces analyses de données peuvent permettre de décrocher une médaille. Par conséquent, les retombées économiques d'une telle pratique peuvent s'avérer très rentables pour les structures sportives professionnelles. [Section 3]

Mais le Sport professionnel n'a pas le monopole du "Big Data" dans le domaine du Sport. De plus en plus de services d'analyse de données sont proposés aux sportifs amateurs, notamment via certaines applications pour smartphones ou des montres connectées. Dans ce cas, les acteurs et les modèles économiques de ce secteur sont différents : les services sont proposés, la plupart du temps, gratuitement aux sportifs amateurs, en contrepartie d'une exploitation lucrative des données [Section 4]. L'accès à de grands volumes de données sportives intéresse aussi d'autres domaines. Citons par exemple les diffuseurs d'évènements sportifs, les entreprises de pari sportif et leurs utilisateurs, etc. [Section 5]

Ainsi, le monde du Sport, qu'il soit professionnel, amateur ou extra-sportif, a connu un bouleversement radical ces dernières années avec le développement du "Big Data". Ce dernier ne va pas sans quelques questionnements éthiques sur de nombreux points tels que l'exploitation de données physiologiques sensibles ou la dénaturation de l'esprit sportif. [Section 6]

2 Les données sportives intéressantes

Un réel changement au niveau de la collection et de l'exploitation des données dans l'industrie du sport s'est effectué au début des années 2000. Il a été initié par l'ancien directeur sportif de l'équipe de baseball d'Oakland, Michael Lewis et plus particulièrement développé dans son ouvrage *MoneyBall the art of winning an unfair game* [3]. Il est le premier à proposer d'évaluer les performances sportives à partir d'évidences terrains et d'outils d'analyse statistique plutôt que d'intuition et de tradition. Cette tendance se manifeste aujourd'hui avec une collecte massive de données pour produire des modèles d'analyse de plus en plus performants (modèle prédictif ou modèle prescriptif). Cela va de l'analyse pour le recrutement de sportifs, à l'analyse des performances jusqu'au renforcement de l'engagement des fans.

Collections de données selon les 3V du Big Data

Pour parvenir à extraire des données à haute valeur ajoutée et créer des modèles prédictifs ou prescriptifs, de grandes quantités de données sont collectées selon les 3V du Big Data : Volume, Variété et Velocités.

Données spatio-temporelles pour l'étude d'évènement sportif

Les premières concernent les données spatio-temporelles [4], c'est à dire les données liées à la location dans l'espace et le temps des différents phénomènes. Cela concerne dans un premier temps le suivi de trajectoires de sportifs et de leur mouvements. Ces données sont captées sous différentes formes et de manière continue lors d'entraînements ou de compétitions : images avec suivi optique par caméra ou point de position grâce à des puces GPS ou RFID sur des systèmes embarqués. Par exemple, la ligue de basket NBA utilise le système **sportVU** composé de 6 caméras à prise de vue 25 images/secondes. On considère qu'un match de basket permet de collecter environ 1TB de données brutes. Quant aux puces GPS, le rythme d'acquisition peut monter jusqu'à 1000 points par secondes par joueur pour la technologie **ClearSky**. On notera que les données collectées par GPS sont de l'ordre de quelques bytes et que malgré la haute fréquence d'acquisition elle fournit un réel avantage en terme de stockage contrairement à l'analyse vidéo [5].

Dans un second temps, les données concernant les évènements sont massivement collectées (pour le football par exemple : passes, tirs, tackles ...)

Moniteur d'activité physique grand public

En ce concerne les moniteurs d'activités physiques disponibles au grand public (*Cf.* partie 4), les données brutes collectées sont moins volumineuses pour répondre aux impératifs des objets connectés. Elles concernent les données de mouvements via l'utilisation d'accéléromètres ou des données physiologiques comme la mesure du rythme cardiaque par cardiofréquence-mètre. Pour ces deux cas, les données sont acquises en temps réel et constamment renouvelées. On considère pour les applications disposant de 30 millions d'utilisateurs que le nombre de tuples collectés peut monter jusqu'à 25 000 tuples/seconde [5]. Cependant l'utilisateur est aussi invité à entrer lui même des données sur ses habitudes et pratiques (alimentation, poids, ...). Nous ne sommes plus dans le cas d'acquisition d'information en grande quantité mais directement sur la collecte d'information à haute valeur ajoutée.

Surveillance des réseaux sociaux

Au delà des données liées aux performances sportives ou à la santé des utilisateurs, une collecte de données de nature très différente est au coeur des enjeux du big data dans l'industrie du sport : la surveillance des réseaux sociaux [6]. Les réseaux sociaux sont la plus grande source d'informations pour les supporters et les fans. Grâce aux réactions des supporters sur les différentes plate-formes, chaque organisation peut avoir une idée très claire de ses attentes, de ses besoins et ainsi toujours avoir un moyen de surveiller et d'améliorer la communication et les relations avec les fans. Cela passe par du **Text-mining** de l'ensemble des informations produites sur les réseaux lors de grands événements. Si on prend comme exemple la coupe du monde de football de 2018, pas moins de 115 milliard de tweets ont été publiés sur l'ensemble des 64 matchs. L'information extraite joue un rôle clé dans la stratégie des entreprises sportives pour fidéliser les supporters et développer leurs outils de marketing.

En somme, l'industrie du sport est très friande de Data. Des informations de natures très différentes sont récoltées de manière continue et en grande quantité. Que se soit des images, des séries-temporelles ou du texte, les informations sont méticuleusement choisies afin d'adresser au mieux les besoins des sportifs ou à des fins de marketing.

3 Big Data et sport professionnel

Le sport professionnel, collectif ou individuel, est un univers aux enjeux économiques importants et aux résultats incertains. Ces deux éléments mènent naturellement à l'utilisation du Big Data. En effet, de plus en plus de clubs, représentant une grande variété de disciplines sportives, ont recours à l'analyse de grandes quantités de données, à différents niveaux, afin d'améliorer les résultats sportifs. Il s'agit principalement d'aides à la décision en ce qui concerne la stratégie adaptée et du suivi physique des athlètes.

3.1 Influence sur la stratégie de jeu

Certaines disciplines sportives se prêtent mieux à l'utilisation de grandes quantités de données. En effet, plusieurs éléments vont rendre cette tâche plus ou moins compliquée :

- La facilité à acquérir des données. Certaines données sont plus facilement accessibles que d'autres. Ainsi, dans le sport automobile, les données mécaniques peuvent être aisément acquises grâce à l'utilisation de différents capteurs dédiés ajoutés au véhicule. Dans d'autres sports comme le football, l'acquisition de données élémentaires comme la position d'un joueur nécessite un équipement supplémentaire non naturel pour le joueur.
- La quantité de données nécessaire. Cette notion est assez subjective et dépend de la complexité et du niveau de l'analyse des données qui est réalisée. Mais il peut être nécessaire d'analyser les données « à la volée » à des fins stratégiques, ce qui requiert des technologies de transfert de données performantes.
- La rapidité de l'analyse. Certains sports peuvent nécessiter de prendre des décisions très rapidement, comme lors d'une course automobile, alors que d'autres sports laissent un intervalle de temps plus important entre l'acquisition et le résultat de l'analyse. Cependant, au sein d'un même sport, cette notion peut varier en fonction des objectifs de l'analyse de données. A titre d'exemple, un suivi en temps réel de la zone du terrain accessible par un des joueurs d'une équipe et la prédiction de la réaction d'une équipe adverse à une tactique précise n'ont pas les mêmes exigences temporelles.

Ces quelques éléments expliquent en partie pourquoi l'usage du Big Data dans le domaine du sport professionnel est très varié. Nous allons étudier quelques exemples caractéristiques qui illustrent le potentiel et les enjeux du Big Data dans le monde du sport professionnel.

En Formule 1, les différentes écuries ont recours au Big Data depuis plusieurs années. Les monoplaces sont équipées d'environ 300 capteurs (réduits au nombre de 200 le jour de la course en raison de leur poids) qui fournissent une grande quantité de données variées : pression des pneus, température, etc. Sur un week-end de grand prix, pas moins de 500 Go de données brutes sont récoltés pour chaque véhicule. Lors de l'arrêt au stand, les données sont transmises, via l'utilisation de la 5G et de matériel dédié, à un centre d'analyse de données propre à chaque écurie, où une dizaine de Data Analysts (accompagnés de grosses structures de calcul) vont analyser les données et essayer d'optimiser la course : à quel moment changer les pneus ? Quelle quantité de carburant embarquer ? Quand freiner pour mieux négocier un virage ? L'écurie Mercedes a même conclu un partenariat avec une société spécialisée dans

l'analyse de données. Dans un sport où le facteur matériel est prépondérant et où la victoire se joue souvent à quelques dixièmes de secondes, une bonne analyse des données peut s'avérer être décisive.[7]

En ce qui concerne les sports collectifs, les enjeux sont différents. Si certaines données accessibles en continu, comme la distance totale parcourue par un joueur, l'évaluation de son niveau de fatigue ou même son positionnement peuvent être utiles, les données récoltées pendant un match sont souvent analysées à froid et utilisées dans un processus d'amélioration continue, afin d'appliquer une stratégie plus adaptée lors du match suivant [8]. De plus, les perspectives en terme d'analyses prédictives sont bien plus intéressantes. Ainsi, certains grands clubs de football comme le Paris Saint Germain tentent de relever ce défi. Le club, en collaboration avec l'Ecole Polytechnique, a créé un concours dont le but était de mettre au point un algorithme d'intelligence artificielle, entraîné sur les données d'une saison entière de Ligue 1 (données récoltées par la société spécialisée Opta), afin de prédire la position du ballon « 10 faits de jeu plus tard » [9]. D'autres sujets d'étude comme la simulation de la réaction d'une équipe donnée à une tactique adverse (technique appelée « ghosting ») voient le jour [10]. Le potentiel tactique, aussi bien offensif que défensif, d'un tel dispositif est immense.

Bien sûr, d'autres sports sont aussi très actifs sur le sujet, comme le football américain [11, 12], le tennis [13], le rugby [8, 14], etc.

3.2 Suivi physique des athlètes

L'autre grande problématique du sport professionnel pour laquelle le Big data offre des solutions intéressantes est le suivi de la forme physique des athlètes. En effet, les blessures sont fréquentes dans ce domaine et les répercussions sportives et économiques peuvent être importantes. De plus, la charge d'entraînement et de compétition étant toujours plus importante, il est nécessaire de protéger les athlètes contre les risques de blessure.

Un des sports pionniers dans ce domaine est le rugby. En effet, depuis une dizaine d'années, les joueurs sont équipés, à l'entraînement et en match, de boîtiers situés au sommet de leur dos contenant une puce GPS et un accéléromètre. Ainsi, les préparateurs physiques ont accès aux distances parcourues par les joueurs et surtout à l'intensité des plaquages subis. De plus, ils ont la possibilité de détecter une dissymétrie dans la manière de se déplacer d'un joueur, synonyme de risque de blessure [15, 8]. D'autres données plus subjectives sont aussi analysées : les joueurs répondent à des questions sur leur alimentation, sur la qualité de leur sommeil, sur leur moral, etc.[16]

Le Big Data est donc utilisé dans de nombreux sports dans un but de prévention des blessures, mais il peut aussi être utilisé à la suite d'une blessure, lors de la période de retour à la compétition, afin de personnaliser la préparation physique du joueur, comme c'est le cas dans certains clubs de football américain.[11]

4 Big Data et sport amateur

4.1 La data, l'avenir du sport amateur ?

La data est très présente dans le sport aussi bien professionnel qu'amateur. La tendance paraît profonde, plus qu'un effet de mode, car au-delà de ce qui pourrait pour certains apparaître comme du « gadget », de nombreux acteurs travaillent à l'exploitation et à la valorisation de ces données. Pour le haut niveau, comme cela a été développé précédemment, cela se transforme en un élément de management autant qu'en un outil tactique, permettant une meilleure connaissance de soi.

Pour le sport de tous les jours, cela devient un moyen de motiver, de « gamifier » la pratique sportive, à l'instar de ce que proposent Strava ou Zwift par exemple. Ainsi chaque sportif peut voir évoluer ses capacités, ses performances, ce qui lui permet de se comparer à d'autres sportifs professionnels ou amateurs, connus ou inconnus. Par ailleurs ces données personnelles permettent de créer des entraînements personnalisés sans avoir besoin d'un coach personnel. La data, bien analysée, permet à chacun de se préparer comme de véritables professionnels tout en restant maître de son emploi du temps [17].

4.2 La démocratisation des objets connectés et de la data

Les équipements technologiques à destination des sportifs professionnels et amateurs se ressemblent de plus en plus. Les sportifs amateurs développent les mêmes besoins en termes de composants de l'outil. La performance devient un élément accessible au plus grand nombre, sous une forme ludique.

Ainsi, on voit apparaître de nombreuses startups qui tentent de s'attaquer à ce nouveau marché. On retiendra en particulier Strava, Mojjo, Swimbot, et Zwift pour comparer le fonctionnement de ces nouvelles applications et la gestion de la donnée.

- Installée sur le marché du tennis, Mojjo permet d'obtenir les vidéos et statistiques des matchs. Via une caméra, l'algorithme de détection analyse les mouvements, les échanges de la rencontre. Mojjo propose ensuite un service de statistiques complet, une feuille de match détaillée et une vidéo qui résume les moments forts. Mojjo s'adresse essentiellement aux amateurs et aux joueurs de 4^{ème} série (66% de leurs clients). Le service que développe cette startup a pour but d'offrir à tous les joueurs de tennis la possibilité de progresser en s'appropriant les outils des professionnels [18].

- Swimbot démontre que les objets connectés parviennent à toucher tous les sports. Swimbot a développé un nouvel objet connecté à destination des nageurs. Doté d'un simple petit boîtier et d'une paire d'écouteurs, l'utilisateur peut optimiser ses performances. En plaçant l'appareil derrière le bonnet de bain, les écouteurs sur les oreilles, des informations lui sont transmises instantanément. Le swimbot analyse tous les mouvements et les corriger en direct.

- Zwift est une application développée pour Home trainer, c'est-à-dire pour toutes les personnes souhaitant s'entraîner au vélo chez eux, et ceci en combinant l'intensité de l'entraînement et le plaisir du jeu. En effet, Zwift offre à ses clients des mondes virtuels et une communauté qui incitent ses membres à rouler en leur fournissant les statistiques de leur performance, et en leur permettant de rouler aux

côtés de sportifs d'autres nationalités. Chacun peut alors challenger la communauté et ainsi amener la communauté à se dépasser [19].

- Enfin les objets connectés ont surtout envahi un domaine en particulier : le running. On peut désormais facilement repérer des coureurs équipés de montres ou de bracelets connectés. Des objets discrets et performants qui se sont progressivement intégrés au mode de vie de la population. Cette tendance se traduit par la notion de « *quantified-self* », qui est la collecte et l'utilisation d'informations personnelles. L'emploi de ces outils ne se limite plus seulement à la pratique sportive mais s'intègre parfaitement au mode de vie des personnes. On les utilise pour courir afin de connaître la distance parcourue, la vitesse moyenne, le nombre de calories dépensées, mais aussi au jour le jour, afin d'obtenir un détail de son activité. Strava fournit une application mobile permettant d'enregistrer diverses données (position, vitesse, rythme cardiaque, etc.) durant une séance de sport (course à pied, cyclisme, natation, etc.) [20].

Ainsi, des outils réservés initialement aux sportifs professionnels sont adaptés pour les sportifs amateurs et leurs permettent de se jauger, de se motiver, et de s'amuser. La performance sportive ne devient donc plus seulement la propriété des professionnels mais aussi celle de tous les amateurs. On assiste alors à des flux de données entre les différents acteurs (utilisateurs, serveurs, communautés, applications, acteurs publiques). Cette data est maintenant considérée comme normale, mais elle possède une vraie valeur marchande. Ainsi, les sportifs sont prêts à vendre leurs informations personnelles pour disposer de services gratuits comme le montre l'application Strava et sa politique de confidentialité [20]. L'utilisation de la data est alors démocratisée et cela ne choque plus personne que les données soient utilisées par d'autres pour peu que celles-ci soient anonymisées.

4.3 Les flux de data

Afin de fournir le meilleur service possible, les applications sportives tissent un réseau de flux de data entre elles. En effet les applications sont de plus en plus connectées entre elles, et s'échangent des informations. Ainsi, Strava récupère les données des montres GPS (Garmin, Polar, etc.) pour les réutiliser et offrir une plateforme plus attrayante car plus complète.

Par exemple, un million de sportifs rejoignent Strava tous les 30 jours, et plus de 32 millions d'athlètes Strava téléchargent des activités dans 195 pays. 85 millions de commentaires et de "kudos" sont échangés chaque semaine par les athlètes Strava.

4.4 La data, une valeur marchande

On distingue différents business modèles pour ces startups. Celles-ci ont différentes stratégies pour gagner de l'argent. Elles peuvent :

- Faire payer la publicité qu'elles diffusent sur l'application ou le site web.
- Faire payer le service ou une partie du service (système de freemium, premium, vente de matériel etc. . .).
- Vendre les données récoltées.

Les applications gratuites, par exemple, utilisent souvent le troisième levier bien qu'elles puissent aussi utiliser les publicités. Par exemple, l'application Strava est gratuite pour l'utilisateur. Cependant, le fait de devoir stocker les données et de les analyser assez précisément engendre des coûts en terme d'infrastructures et de

calcul pour l'entreprise. Celle-ci a donc besoin d'une rentrée d'argent. Or comme Strava a décidé de ne pas inclure de publicités dans son application ou son site internet, elle fait donc rentrer de l'argent dans ses caisses de 2 manières :

- **Via la vente d'un ou plusieurs abonnements "Premium".**
- **Via *Strava Metro*[21], i.e en revendant les informations personnelles (parcours de course à pied et de vélo, fréquence d'activités) aux villes.** Il s'agit d'un service mis à la disposition des villes du monde entier et qui est destiné à analyser les données issues de l'application pour la ville en question. Par exemple, la ville de Lyon pourrait avoir accès au nombre de cyclistes (utilisant l'application Strava) ayant emprunté le tunnel de la Croix-Rousse au mois de Septembre 2019 et ainsi envisager un aménagement des infrastructures adapté aux besoins. Il est aussi possible pour les villes de comparer l'effet de la mise en place d'une infrastructure (avant/après). Enfin, les données sont fournies dans un format standard (format SIG) qui permet aux villes de les analyser plus facilement et de les intégrer à d'autres jeux de données.

4.5 Utilisation de la data sportive

Les données rassemblées concernant les sportifs amateurs ou "sportifs du dimanche" intéressent plus d'un domaine. En effet ces données donnent accès :

- à des comportements généraux comme les flux de coureurs dans les villes (Strava Métro), ce qui intéresse les agglomérations qui peuvent donc savoir quels sont les lieux fréquentés par les coureurs et les horaires de sorties.
- aux profils individuels, ce qui intéresse par exemple les assurances maladie, puisque ceci leur permettrait d'adapter leur prix en fonction des pratiques sportives de l'individu.

5 La Data au service de l'extra-sportif

Au delà de l'amélioration des performances sportives et du suivi des athlètes professionnels ou amateurs, toutes ces données produites ont de nombreuses applications dans tout le cadre extra-sportif : diffusions d'évènements, paris sportifs, publicité... Nous avons vu précédemment que les données récoltées sont très variées : données sur les joueurs, statistiques sur les sports en question, données des réseaux sociaux... Elles peuvent donc être utilisées pour de nombreuses et diverses applications. Cela provoque évidemment un changement de l'économie des activité entourant le sport. En effet, toutes ces données sont utilisées pour mieux cibler la clientèle potentielle, les horaires de diffusion, les matchs à évènements à diffuser en priorité...

5.1 Sport et réseaux sociaux

Y-a-t'il un meilleur moyen que d'utiliser les réseaux sociaux pour suivre et supporter son équipe favorite au quotidien ? Sportifs et clubs sont suivis par des millions de fans sur les différents réseaux sociaux : la page facebook du FC Barcelone est suivie par plus de 100 millions de personnes et Cristiano Ronaldo atteint même les 120 millions de fans sur cette même plate-forme. S'il s'agit d'un très bon moyen pour le public de suivre et montrer son soutien à son équipe, les diffuseurs, annonceurs et entreprises en profitent largement pour récolter un maximum de données afin d'attirer plus de monde et augmenter leurs ventes [22]. Les données des très nombreux sportifs amateurs sont aussi utilisables, en effet nombre d'utilisateur aime partager leurs entraînements et performances quotidiens ce qui génère un grand volume de données.

5.2 Diffusion des évènements sportifs

Pour les diffuseurs (chaînes de télévision principalement) les nouvelles technologies peuvent être utilisées pour aider à la réalisation de certaines séquences diffusées : extraction automatique de données, découpage automatique de séquences, personnalisation des contenus, fabrication de résumés prêts à être diffusés au cours du direct, création d'images intermédiaires pour les ralentis, outils de visualisation 3D pour les compte-rendus, calculs statistiques (par exemple compter le nombres de passes d'un joueur)...[23] Elle permettent grandement l'amélioration de l'expérience fournie aux téléspectateurs et donc de fidéliser ceux-ci ou d'en attirer de nouveaux.

5.3 "Marketing" sportif

De même, les marques utilisent largement ce phénomène et récoltent beaucoup de données afin de créer les nouvelles tendances, des équipements plus performants et de mieux cibler leurs clients. En effet le Big Data est très utilisé depuis plusieurs années maintenant dans le domaine du marketing, et les marques de sport n'échappe pas à ce phénomène. Dans ce domaine les données des clients sont utilisées à plusieurs fins : mieux cibler les publicités sur internet afin de données envie à l'utilisateur d'acheter le produit, utiliser des algorithmes de recommandation sur les site de vente en ligne afin de vendre plus de produits au consommateur, ou encore mieux choisir les effigies de la marque afin de donner un maximum de visibilité. Sur ce dernier point, on sait que le marketing sur les réseaux sociaux se développe à grande vitesse, et pour mieux choisir quel influenceur peut représenter sa marque, celle-ci étudient grâce au Big Data les différents profils pour obtenir leur audience (tranche d'âge,

pays, genre...), les engagements (si la communauté de l'influenceur est active et suit réellement la personne), et grâce à toute ces données, les marques peuvent choisir les influenceurs qui optimiseront le retour sur investissement. Par ailleurs, comme vu précédemment, certaines marques utilisent le Big Data pour développer de nouvelles technologies (montres sportives, applications...) et l'utilisation de machine learning apparaît comme un argument de vente supplémentaire. Le Big Data est aussi utile pour développer des produit plus performant et donc se distinguer de la concurrence pour mieux vendre son produit.

5.4 Côte et paris sportifs

En 2018, EA sports avait simulé le déroulement de la Coupe du Monde de football sur son célèbre jeu de simulation FIFA 2018, et avait correctement prédit que la France allait gagner ainsi qu'une bonne partie du tableau final [24]. La prédiction de résultats sportifs est très importante pour les bookmakers et pour les parieurs. En effet les parieurs ont misés près de 3.9 milliards d'euros en France sur l'année 2018 et ce marché ne cesse de croître et de gagner en popularité notamment via les plate-formes de paris en ligne [25]. Il s'agit donc d'une activité très rentable pour les bookmakers qui cherchent le moyen d'augmenter le nombre de parieurs en proposant des côtes avantageuses sans prendre de risques. En parallèle, le nombre de parieurs augmente et ceux-ci cherchent le meilleur moyen de gagner leurs paris. C'est là qu'intervient le Big Data et les algorithmes de prédiction : après avoir récolté la masse de données disponible à propos d'un sport (classement, historique des rencontres, tendances, actualités des équipes/joueurs...) il est possible de créer un modèle simple qui prends ces données en entrée et détermine en sortie les probabilités (par exemple) pour les trois évènements suivants : *l'équipe à domicile gagne, l'équipe qui se déplace gagne, ou égalité entre les deux équipes* [26]. Il est évidemment possible de complexifier ce modèle en utilisant plus de données, des données plus variées, des données à analyser en temps réel. Par exemple certaines plate-formes de paris proposent des paris en direct avec des côtes qui varient en temps réel (la côte s'adapte en fonction du score, de la possession...). On peut également tenter de prédire des évènements plus spécifiques comme le score d'une rencontre, l'écart de points d'un match de basketball ou encore les buteurs d'un match de football. Les plate-formes de paris en ligne utilisent également le big data pour garder les clients : elles scrutent chacun des clics des utilisateur afin de mieux cibler leurs habitudes de paris et de les inonder de publicités très personnalisées [27].

6 Questionnements éthiques

Bill Gates a dit un jour : "Nous surestimons toujours le changement qui se produira au cours des deux prochaines années et sous-estimons le changement qui se produira au cours des dix prochaines. Ne vous laissez pas bercer par l'inaction."

Ce sentiment n'a jamais été aussi vrai que lorsqu'il a été appliqué à la Big Data qui transforme l'utilisation et la valeur des données dans le sport. Les progrès de la technologie ne servent pas seulement à améliorer le sport en tant que produit de divertissement grand public, mais ils peuvent aussi être utilisés pour améliorer la performance des athlètes et des utilisateurs d'application grand public. Le Big Data et leur analyse permettent le profilage des individus en traitant de grandes quantités de leurs données pour en dégager des tendances et parfois prévoir les comportements futurs. Les données sont diverses et de nature sensible puisqu'elles portent en partie sur des données biologiques et peuvent ainsi entrer dans le domaine du secret médicale. Elles illustrent de nombreux aspects de la performance d'un individu tels que la distance, la vitesse, mais aussi la température du corps, la fréquence cardiaque, le rythme cardiaque, le rythme du sommeil ou encore l'apport calorique. Elles font partie intégrante de l'entraînement et de l'analyse de la performance et sont en pleine croissance, aussi bien chez les athlètes que les amateurs [28], comme cela a pu être développé précédemment.

Plusieurs questions peuvent ainsi être posées à cause de la sensibilité et du volume de ces données. Que se passe-t-il si des concurrents s'emparent de ces informations et les utilisent pour, par exemple, exploiter les faiblesses de leurs concurrents ? Que se passe-t-il si les joueurs sont soumis à une prise de décision automatisée, basée sur des algorithmes avec des informations pouvant être inexactes, tout en ayant un impact à long terme sur leur vie et leur carrière ? Qu'arrive-t-il si les informations les plus personnelles sont exploitées à des fins commerciales ?

6.1 Le RGPD - Règlement général sur la protection des données

Le RGPD, en vigueur dans toute l'Union Européenne à partir du 25 mai 2018, est conçu pour répondre en partie à ces questions. Le RGPD introduit des exigences plus strictes de protection des données. Les entreprises devront désormais faire face à des sanctions financières qui pourraient atteindre 4% de leur chiffre d'affaire annuel mondial en cas de non-respect [29].

Prenons l'exemple d'une équipe de football qui recueille des données sur la performance et la santé de son équipe, pendant l'entraînement et les matchs, à partir de dispositifs de surveillance portés par ses joueurs. Ces données sont ensuite utilisées pour prendre certaines décisions concernant un joueur, par exemple s'il doit être remplacé dans le prochain jeu, ce que ce joueur vaut sur le marché des transferts, et si ce joueur devrait être vendu [28].

Le RGPD régleme aussi la manière dont l'équipe entreprend cette activité de récolte et de traitement des données, avec le risque d'amendes importantes et de relations publiques défavorables, si elles ne respectent pas les règles. Dans un premier temps, dans le cadre du RGPD, l'équipe devra procéder à une évaluation de

l'atteinte à la vie privée avant d'entreprendre la récolte de données et en informer le joueur, lors de la demande de consentement. La transparence de l'algorithme doit aussi être montrée. Le RGPD accorde, en parallèle, de l'importance aussi à la sécurité des données [29, 28].

6.2 La transparence de l'utilisation des données

Les utilisateurs des applications grand public comme Strava ou Runkeeper ont très peu conscience de la façon dont leurs données peuvent être utilisées par ces applications. Ainsi Strava revend les données de géolocalisation, certes anonymisées, à des acteurs publics dans le cadre d'évaluation de politique de transport par exemple, comme développé en section 4. Mais d'autres applications peuvent utiliser les données amassées pour de la publicité ciblée, en fonction du profil de chaque utilisateur et de ses habitudes [30]. En effet, récemment, les principales applications ont été rachetées par des équipementiers sportifs tels que Runkeeper (comptant 33 millions d'utilisateurs) racheté par Asics ou encore MyFitnessPal par Under Armour [31].

6.3 L'enjeu sécuritaire

Chez Strava, mais c'est également le cas des autres applications, l'ensemble des trajets des athlètes est visualisable sur une carte ouverte à tous les utilisateurs : la *heatmap*. Ainsi Strava a révélé involontairement en 2017 l'emplacement de bases américaines et françaises en Irak et au Niger qui sont isolées géographiquement dans des zones désertiques. A partir de la densité du trafic et des points de convergence, il est même possible d'avoir une idée relativement précise de la localisation des quartiers dans les campements où résident les militaires et permet de connaître les routines de ces utilisateurs [32]. Cependant, les données fournies par Strava sont anonymisées et il est impossible d'extrapoler des horaires précis, ou de dater les courses [33]. Cela est néanmoins discutable car les profils des utilisateurs sont paramétrés à la base pour être accessibles à tous, sans restriction, et ainsi voir les informations enregistrées. Sur ces profils, les utilisateurs peuvent par ailleurs mentionner des données personnelles comme leur régiment, dans le cas des militaires, ou encore leur profession de manière plus générale.

6.4 Gamification et fair-play sportif

Un procédé de gamification est intégré dans les applications sportives grand public pour fidéliser leurs utilisateurs : quand les utilisateurs battent leurs propres records, ils reçoivent des médailles virtuelles les récompensant. En parallèle un classement des performances réalisées sur un même parcours entre les différents utilisateurs est instauré.

Avec l'émergence de la gamification à-travers les classements, un changement de comportement a été établi. Ainsi, des utilisateurs peuvent être poussés à courir pour le classement, c'est-à-dire, à courir seulement sur les segments où sont réalisés les classements, ce qui les éloigne finalement de l'esprit sportif, du dépassement de soi et du fair-play [34].

7 Conclusion

Ainsi l'usage de la Big Data s'est durablement implanté dans l'ensemble du paysage sportif en une dizaine d'années. Les acteurs du Big Data ont établi des modèles économiques particuliers pour se développer et, en parallèle, ont pu prouver l'efficacité et l'intérêt de la collecte des données et de leurs analyses. Le Big Data n'a pas encore fini son développement, l'amélioration du temps d'analyse des données ouvre des voies prometteuses à l'utilisation du Big Data lors des compétitions par exemple. Cependant on ne peut pas réduire la performance d'un athlète à l'analyse de ses données. Le Big Data n'est qu'une partie de l'ensemble du travail fourni par l'athlète et son entraîneur, dont le résultat final dépend d'une multitude de facteurs. L'usage du Big Data dans le sport amateur a modifié le comportement et l'intérêt porté au sport en apportant un aspect ludique quant à la visualisation des données (avec les classements ou les *heatmaps*).

L'exploitation des données a pour but final la recherche de profit économique, et se réalise dans un environnement juridique encore balbutiant. Les questions éthiques et de protection des données commencent à devenir des enjeux cruciaux du fait de la taille des données collectées de plus en plus importantes et détaillées, grâce à l'évolution des capteurs. Aujourd'hui les enjeux éthiques et légaux ont pu rattraper une partie de leur retard sur le développement du Big Data dans le sport grâce à la sensibilisation des usagers et l'établissement du RGPD en Europe.

Références

- [1] Robin DELASSUS Consultant mc2i Groupe. Jo paris 2024 : la business intelligence, le big data et les objets connectés sont les nouveaux dopants. 05-10-2017. <https://www.usine-digitale.fr/article/jo-paris-2024-la-business-intelligence-le-big-data-et-les-objets-connectes-sont-les-nouveaux-dopants.N596438>.
- [2] Ecole polytechnique. Sports et sciences, une alliance en or. <https://sciences2024.polytechnique.fr/>.
- [3] Big data analysis is changing the nature of sports science. <https://www.technologyreview.com/s/600957/big-data-analysis-is-changing-the-nature-of-sports-science/>.
- [4] Michael Horton Joachim Gudmundsson. Spatio-temporal analysis of team sports – a survey. <https://arxiv.org/pdf/1602.06994.pdf>.
- [5] Olivier Marin Pierre Sens Rudyar Cortés, Xavier Bonnaire. Sport trackers and big data : Studying user traces to identify opportunities and challenges. <https://hal.inria.fr/hal-01092242/document>.
- [6] Dr David Marlin. The growing necessity of big data in the sports market. <https://www.arioneo.com/en/big-data-competitive-sports-market/>.
- [7] Stéphanie Mundubeltz-Gendron. Fast data, analyse prédictive... la formule 1, une course de vitesse à tous les niveaux. *L'Usine Digitale*, 2018. <https://www.usine-digitale.fr/article/fast-data-analyse-predictive-la-formule-1-une-course-de-vitesse-a-tous-les-niveaux.N693729>.
- [8] Yannick Cochenec. Comment le gps et le big data changent le rugby. *Slate*, 2014. <http://www.slate.fr/story/82935/comment-gps-big-data-changent-rugby>.
- [9] Bastien L. Le psg mise sur l'intelligence artificielle pour gagner la ligue des champions. *Le Big Data*, 2019. <https://www.lebigdata.fr/psg-intelligence-artificielle>.
- [10] DisneyResearchHub. Data-driven ghosting using deep imitation learning. 2017. <https://www.youtube.com/watch?v=WI-WL2cj0CA>.
- [11] John Pollard. La nfl entre dans une nouvelle ère d'analyse en temps réel. *L'Usine Digitale*, 2019. <https://www.usine-digitale.fr/article/la-nfl-entre-dans-une-nouvelle-ere-d-analyse-en-temps-reelle.N816250>.
- [12] Bastien L. Sport et big data – quand la science des données donne l'avantage sur le terrain. *Le Big Data*, 2017. <https://www.lebigdata.fr/sport-et-big-data>.
- [13] Léna Corot. Data, intelligence artificielle... infosys mène la transformation digitale de roland-garros. *L'Usine Digitale*, 2019. <https://www.usine-digitale.fr/article/data-et-intelligence-artificielle-infosys-mene-la-transformation-digitale-de-roland-garros.N821340>.
- [14] V.P-L. La mêlée révolutionnée. *L'équipe*, 2010. <https://www.lequipe.fr/Rugby/Actualites/La-melee-revolutionnee/111964>.
- [15] David Decloux. Sport data : la compo gagnante ? *Culture Data*, 2017. <https://culture-data.cartegie.com/big-data/sport-data-la-compo-gagnante>.
- [16] ASM Rugby. Le "big data" ou comment prévenir plutôt que guérir ! *ASM Rugby*, 2018. <https://www.asm-rugby.com/actus/le-big-data-ou-comment-prevenir-plutot-que-guerir>.

- [17] Data et performance sportive. 2016. <https://usineasite.parisandco.com/Le-Tremplin/Actus/Tendances-Sport-Innovation/Data-et-performance-sportive>.
- [18] Mojjo. Site officiel mojjo. 2019. <https://www.mojjo.fr/>.
- [19] Zwift. Site officiel zwift. 2019. <https://zwift.com/eu>.
- [20] Strava. Politique de confidentialité. 2019. <https://www.strava.com/legal/privacy>.
- [21] Strava. Strava métró. 2016. <https://metro.strava.com/>.
- [22] Ryan Ayers. How big data is revolutionizing sports. *Dataconomy*, 2018. <http://dataconomy.com/2018/01/big-data-revolutionizing-favorite-sports-teams/>.
- [23] Bernard Poiseuil. Le sport s'éveille à l'intelligence artificielle. *Mediakwest*, 2018. <http://www.mediakwest.com/broadcast/item/mk28-le-sport-s-veille-a-l-intelligence-artificielle.html>.
- [24] NC. Ea sports prévoit une victoire de la france pour la 2018 fifa world cup. *EA Sport*, 2018. <https://www.ea.com/fr-fr/games/fifa/news/ea-sports-predicts-world-cup-fifa-18>.
- [25] NC. Année record pour les paris sportifs, portés par de la coupe du monde de football. *Le Figaro*, 2019. <https://www.lefigaro.fr/conso/2019/02/15/20010-20190215ARTFIG00097-annee-record-pour-les-paris-sportifs-portes-par-de-la-coupe-du-monde-de-football>.
- [26] Charles Malafosse. Machine learning for sports betting : It's not a basic classification problem. *Towards Data Science*, 2018. <https://towardsdatascience.com/machine-learning-for-sports-betting-not-a-basic-classification-problem-b42ae490>.
- [27] NC. Revealed : how bookies use ai to keep gamblers hooked. *The Guardian*, 2018. <https://www.theguardian.com/technology/2018/apr/30/bookies-using-ai-to-keep-gamblers-hooked-insiders-say>.
- [28] Warren Phelops and Andrew Gilchrist. The legal implications for big data, sports analytics and player metrics under the GDPR - LawInSport. <https://www.lawinsport.com/content/articles/item/the-legal-implications-for-big-data-sports-analytics-and-player-metrics-under-the-gdpr>.
- [29] Commission européenne. EUGDPR – information portal. <https://eugdpr.org/>.
- [30] L. Bastien. Publicité : comment le big data résout les problèmes des annonceurs. <https://www.lebigdata.fr/publicite-big-data-annonceurs-2305>.
- [31] Samuel Gibbs. Runkeeper bought by asics in latest sports brand app acquisition. <https://www.theguardian.com/technology/2016/feb/12/runkeeper-asics-sports-brand-app-acquisition>.
- [32] Nicolas Six. Une application de jogging menace la sécurité des bases militaires. https://www.lemonde.fr/pixels/article/2018/01/29/la-securite-des-bases-militaires-menacee-par-une-application-de-jogging_5248885_4408996.html.
- [33] Strava. Données et vie privée. <http://support.strava.com/hc/fr/articles/360001487844-Données-et-vie-privée>.
- [34] Strava. Strava engineering. <https://medium.com/strava-engineering>.

- [35] Ban500. Football big data. *Étonnante Époque*, 2019. <https://www.etonnante-epoque.fr/football-big-data/>.
- [36] C.Y.Chen N.Kitiseni, T.Y.Lee. Big data sport. <http://tassm.org.tw/wp-content/uploads/2018/10/Big-Data-Sport.pdf>.