***NOTE*** *: Cet article a été écrit en* [*anglais*](https://towardsdatascience.com/data-science-trends-for-2019-11b2397bd16b) *et a été traduit en français à l'aide de* [*DeepL*](https://www.deepl.com/home)*, un outil qui forme à l'intelligence artificielle pour comprendre et traduire des textes.*

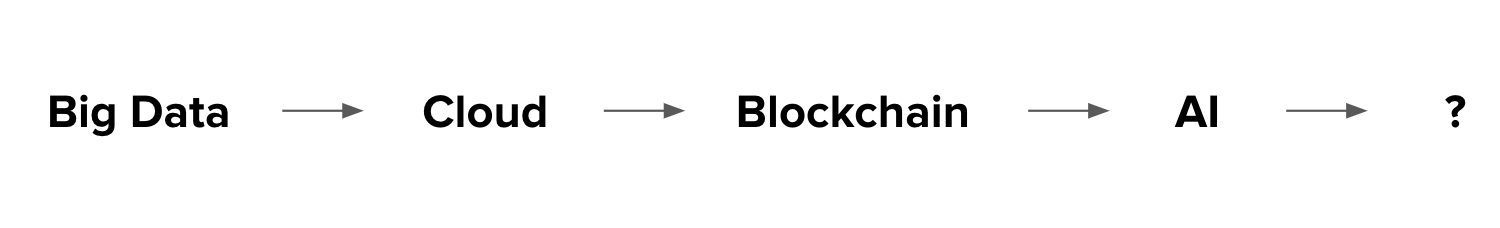
# **Tendances de la science des données pour 2019**

### **Avec tout ce qui se passe, que s'est-il passé en 2018 et qu'est-ce qui semble prometteur pour 2019?**

### 26 déc. 2018. Sept minute de lecture.

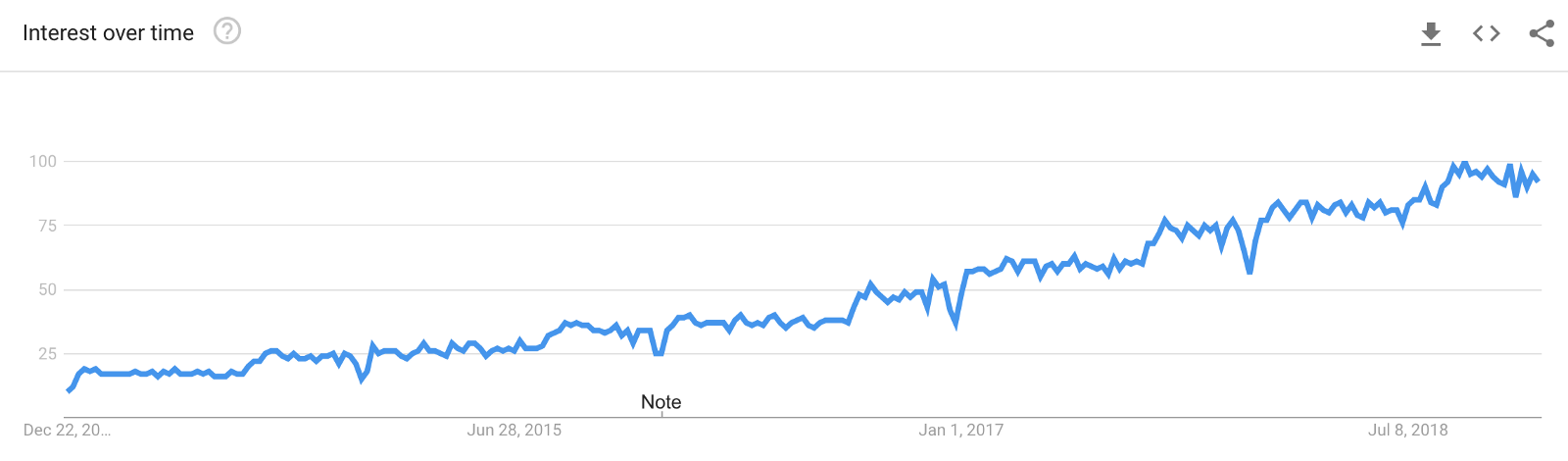
**Cette année peut être considérée comme le boom de l'intelligence artificielle (IA)**. Il suffit de regarder le nombre de startups avec le terme AI dans leurs slogans, où les acquisitions des grandes entreprises se sont concentrées et les sujets abordés lors des plus grandes conférences technologiques. L'IA est partout, même si c'est un mot à la mode. Mais qu'est-ce que l'IA ? C'est une question assez difficile et controversée à laquelle il est difficile de répondre.

Un sac de mots à la mode. Ne nous concentrons pas sur les mots à la mode, mais sur ce que les technologies sous-jacentes peuvent réellement résoudre.



Ce qui est clair, c'est que la science des données résout les problèmes. Les données sont partout, et les usages que nous en faisons (la science) augmentent et ont un impact de plus en plus grand sur la société. Concentrons-nous sur la science des données, tandis que d'autres philosopheront sur la meilleure définition de l'IA.

Alors que d'autres mots à la mode continuent de prospérer, *comment va la science des données ?*

**

L'intérêt n'est pas mauvais du tout ! Je ne cesse de faire valoir que la science des données n'est pas un mot à la mode. Même pour les personnes qui se joignent maintenant à la science des données - et il y en a beaucoup - vous n'avez qu'à faire une recherche d'emploi rapide sur LinkedIn et vous serez étonné par le nombre d'offres. Commençons par regarder ce qui s'est passé en 2018, puis concentrons-nous sur les sujets d'actualité pour 2019.

# **Aujourd'hui**

L'an dernier, j'ai publié un article sur mes attentes concernant les tendances en sciences des données pour 2018. Les principaux développements que j'ai mentionnés sont : l'automatisation des flux de travail, l'explicabilité, l'équité, la banalisation de la science des données et l'amélioration des outils d'ingénierie et de nettoyage des fonctions.

En ce qui concerne l'automatisation, le travail des spécialistes des données est, très souvent, l'automatisation de leur propre travail. De nos jours, il est courant pour les entreprises d'ouvrir leurs propres pipelines d'automatisation. D'autres, continuez à le vendre, mais chaque jour avec plus de concurrence (par exemple, Microsoft Azure, H2O, Data Robot, Google).

Heureusement, la science des données est une discipline transversale et les mêmes algorithmes que ceux utilisés dans les soins de santé peuvent être utilisés, avec quelques ajustements, en agriculture. Ainsi, si une entreprise fait faillite dans un secteur vertical, ses développements peuvent être rapidement adaptés à un autre domaine.

Ces outils sont en train de devenir des produits de base que vous n'avez même pas besoin de savoir coder pour les utiliser. Certains d'entre eux sont nés il y a quelques années de la rareté des talents en sciences de l'information et ont été transformés en produits rentables par la suite. Cela rappelle l'un des principes du *Rework book - vendre vos sous-produits*.



Les moyens d'amener les humains à faire confiance aux machines sont en train d'être pavés :

L'explicabilité et l'équité ont connu de grands développements en 2018. Il y a maintenant beaucoup plus de ressources disponibles. Les outils qui n'étaient que des versions alpha de Python ont mûri (par exemple, SHAP). En outre, vous pouvez facilement trouver des livres structurés et soutenus sur le sujet, tels que le livre Interpretable Machine Learning, par Christoph Molnar. Comprendre des modèles très complexes va dans la bonne direction en réduisant les obstacles - l'outil d'analyse d'hypothèses de Google en est un excellent exemple.

L'ingénierie des fonctionnalités reste l'une des principales sauces secrètes des solutions Data Science - jetez un coup d'œil à la description de la solution gagnante pour le risque de défaut de crédit immobilier dans Kaggle. Bien que la plupart des meilleures fonctionnalités soient encore créées manuellement, Feature Tools est devenu l'une des principales bibliothèques d'ingénierie de fonctionnalités cette année, pour les scientifiques des données paresseux (intelligents ?). Le problème de ces outils est que vous avez besoin de normes de données pour l'ensemble de votre entreprise, c'est-à-dire que si l'un de vos clients fournit des données dans un format, vous devez vous assurer que le second client suit la même procédure - sinon, vous allez avoir beaucoup de travail manuel indésirable.

Enfin, si nous délivrions des Oscars aux langages de programmation, Python en recevrait probablement quelques-uns. C'est aujourd'hui le langage de programmation majeur à la croissance la plus rapide et le langage le plus recherché pour la deuxième année consécutive, selon Stack Overflow. A ce rythme, il devient rapidement le langage de programmation le plus utilisé.

# **Demain**

Alors, qu'est-ce qu'on fait ensuite ? Que peut-on encore faire ?

Il y a beaucoup à faire dans les domaines ci-dessus. Et ils continueront d'être l'un des

principaux centres d'intérêt des spécialistes des données en 2019 et dans les années à venir.

L'accent sera mis sur la maturation des technologies tout en répondant aux questions :

* Comment pouvons-nous réduire au minimum le temps passé par les spécialistes des données à nettoyer les données et à concevoir des fonctionnalités ?
* Comment définir la confiance dans le contexte de l'apprentissage machine ?
* Si nous disons qu'un modèle de machine est juste, quelles sont ses caractéristiques ?



*Quels sont les principes selon lesquels on peut dire qu'on fait confiance à un robot ?*

Mais, en dehors de ces méta-questions, auxquelles il est difficile de répondre, quels sont les sujets prometteurs ?

L'apprentissage de renforcement a pu passer beaucoup d'hivers au cours de sa vie. Cependant, il semble que nous approchons d'un autre printemps. Un bon exemple est la fantastique performance de Dota 2. Il y a beaucoup à faire, et beaucoup de puissance de calcul sera nécessaire... Mais, quoi qu'il en soit, le renforcement de l'apprentissage est le comportement d'apprentissage le plus humain que nous ayons actuellement et c'est excitant de voir ses applications.

Nous commencerons très probablement à voir ces preuves de concepts se transformer en produits réels. Si vous en avez le temps, jetez un coup d'œil et utilisez gym pour les développer.

*Considérant 71 du GDPR : La personne concernée devrait avoir "le droit... d'obtenir une explication de la décision prise... et de la contester".*

Le règlement général sur la protection des données (GDPR) est en vigueur dans l'UE depuis le 25 mai 2018 et affecte directement la science des données.

Le problème est que les entreprises comprennent encore les limites de cette nouvelle réglementation. Deux des principaux sujets ouverts sont :

* Confidentialité des données. Les entreprises qui manipulent mal les données personnelles sont maintenant menacées d'amendes énormes. Cela signifie-t-il que l'accès aux données deviendra plus difficile pour la recherche ? Verrons-nous de nouveaux développements dans la synthèse des données ? Pouvons-nous vraiment anonymiser les données ?
* Droit à l'explication. Les décisions entièrement automatisées doivent pouvoir être expliquées. Eh bien, c'est génial... Mais qu'est-ce que cela signifie en fait "explicable" ? Va-t-on assister à la standardisation d'un algorithme d'interprétabilité d'apprentissage machine ? Il n'y a pas de réponse des entités de l'UE à ce sujet - nous attendons probablement la plus grosse amende jamais exécutée.

*L'IA digne de confiance comporte deux volets : (1) elle devrait respecter les droits fondamentaux, la réglementation applicable et les principes et valeurs fondamentaux, en assurant un "but éthique" et (2) elle devrait être techniquement solide et fiable puisque, même avec de bonnes intentions, un manque de maîtrise technologique peut causer un préjudice involontaire[éthique de l'IA dans l'UE].*

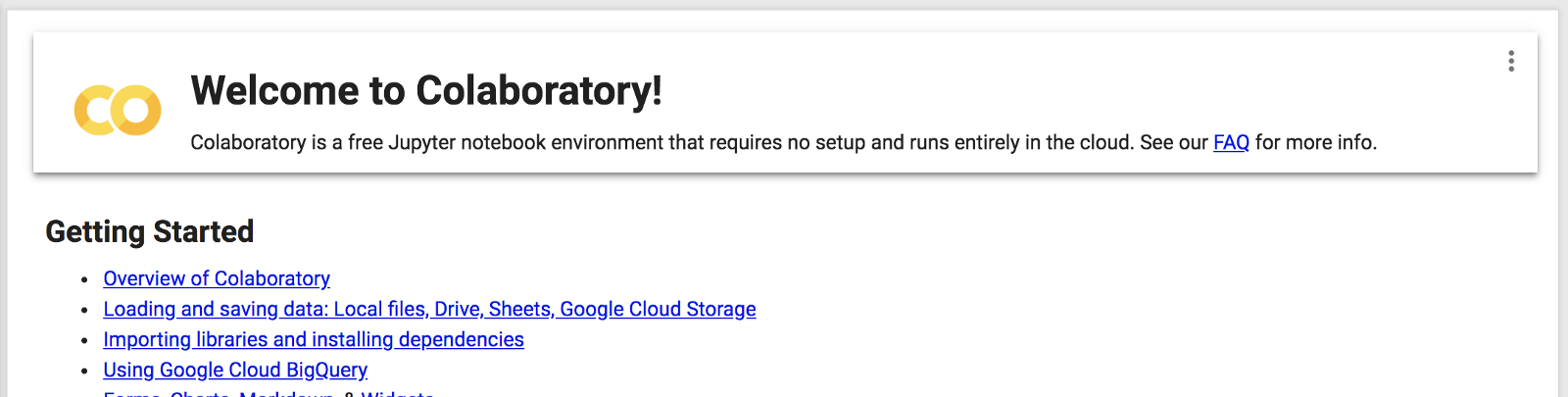
Comme les algorithmes affectent davantage la société, nous avons le droit de nous assurer que les biais sont atténués et que leur utilisation profite à l'ensemble et non seulement à quelques-uns. Heureusement, des entreprises et des institutions y travaillent. Le projet d'éthique de l'IA de l'UE et les principes de l'IA de Google en sont de parfaits exemples. Il y a encore un long chemin à parcourir en matière d'éthique, mais il s'agit maintenant d'un sujet de discussion récurrent - et c'est une bonne chose.

[](https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai)

Le projet de l'UE sur l'éthique de l'IA est un exemple de la façon dont les institutions gouvernementales abordent le sujet.

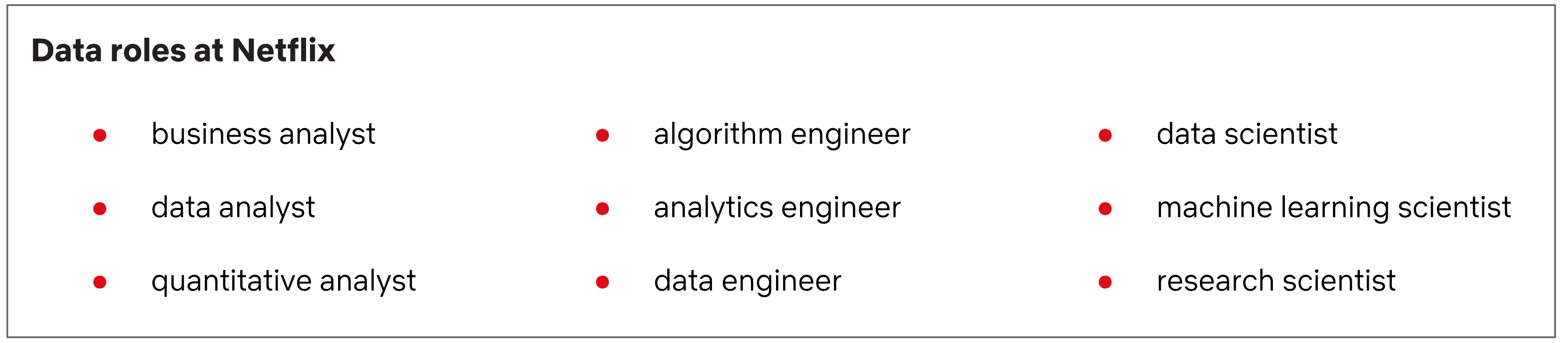
Comme les algorithmes deviennent de plus en plus complexes et qu'il y a de plus en plus de données facilement accessibles (chaque gadget génère maintenant des données, n'est-ce pas ?), moins de gens utiliseront leur ordinateur portable pour faire de la science des données.

Nous utiliserons des solutions basées sur le cloud, même pour les projets les plus simples (par exemple, Google Colab). Le temps presse, les GPU ne le sont pas... Les ordinateurs portables n'évoluent pas assez vite pour suivre le rythme de la puissance de calcul requise.

[](https://colab.research.google.com/notebooks/welcome.ipynb)

Google Colab : faciliter le partage des ordinateurs portables et utiliser plus de puissance de calcul. Maintenant, imaginez que vous voyez une entreprise avec un poste vacant pour le poste d'"Ingénieur" - exactement cela. C'est génial... Mais il y a une centaine de types d'ingénieurs de nos jours. Est-ce que c'est un ingénieur en mécanique ? Aérospatiale ? Logiciel ? "Ingénieur" est trop généraliste.

Il y a un ou deux ans, les entreprises publiaient simplement une offre d'emploi en tant que "Data Scientist". Eh bien, ça commence à être incomplet. Et si vous débutez dans ce domaine, devenir un spécialiste des données générales peut s'avérer trop accablant. Après avoir saisi ce domaine, vous feriez mieux de vous concentrer sur un sujet particulier. Prenons par exemple Netflix, qui a neuf rôles de données :

[](https://medium.com/netflix-techblog/notebook-innovation-591ee3221233)

Il y a beaucoup de domaines de spécialisation qui n'existaient pas auparavant et il devient de plus en plus important pour les scientifiques des données de se concentrer sur un seul pour prendre position. Il est temps de trouver le vôtre, si ce n'est déjà fait. De mon point de vue, les compétences en ingénierie des données sont les plus intéressantes pour les prochaines années. Si vous ne les avez pas dans votre équipe, vous ne faites probablement que jouer à la science des données dans les ordinateurs portables Jupyter. Et les entreprises s'en rendent compte.

2019 sera une année incroyable, encore une fois. Il y a beaucoup à faire, et ce n'est pas seulement de la technique et du nerd ! De vrais problèmes à résoudre sont à venir.

En guise de conclusion, n'oubliez pas que le temps est notre plus grand atout. Chaque seconde que vous passez à faire sans valeur est une seconde que vous venez de perdre en ne faisant pas quelque chose de bien. Choisissez votre sujet et ne considérez pas votre travail comme si de rien n'était.

***NOTE*** *: Cet article a été écrit en* [*anglais*](https://towardsdatascience.com/data-science-trends-for-2019-11b2397bd16b) *et a été traduit en français à l'aide de* [*DeepL*](https://www.deepl.com/home)*, un outil qui forme à l'intelligence artificielle pour comprendre et traduire des textes.*