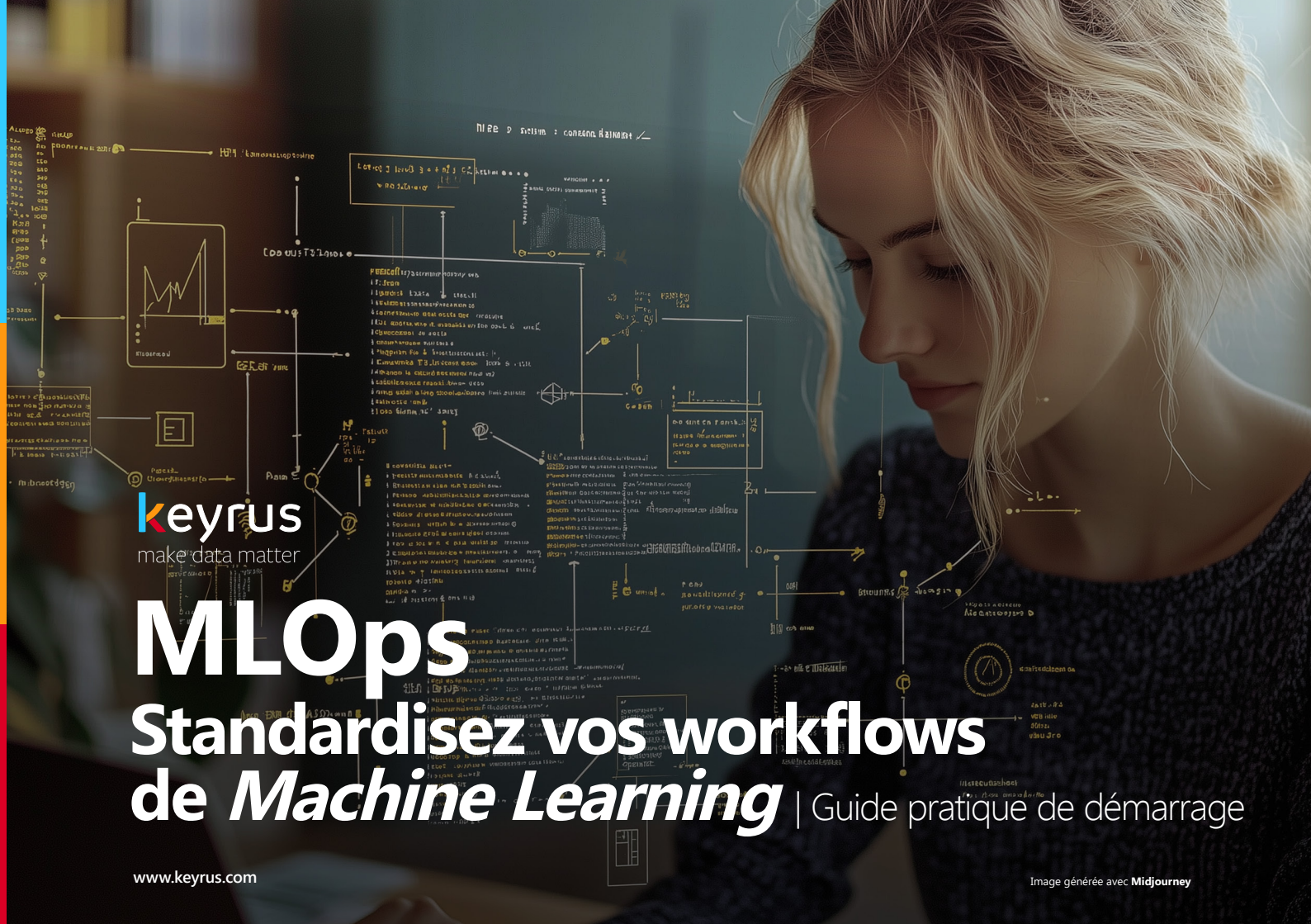


keyrus
make data matter

MLOps

Standardisez vos workflows de *Machine Learning* | Guide pratique de démarrage

www.keyrus.com

Image générée avec **Midjourney**

Standardisez vos workflows de *Machine Learning*

Guide pratique de démarrage MLOps

Définir le MLOps peut s'avérer plus complexe qu'il n'y paraît. Ce concept dépasse largement le simple déploiement et la surveillance des modèles de *machine learning* en production. Le MLOps vise en réalité à **standardiser** et à **optimiser** la gestion complète du cycle de vie des modèles d'apprentissage automatique, de leur conception à leur surveillance continue, en passant par leur mise en production et leur gouvernance.

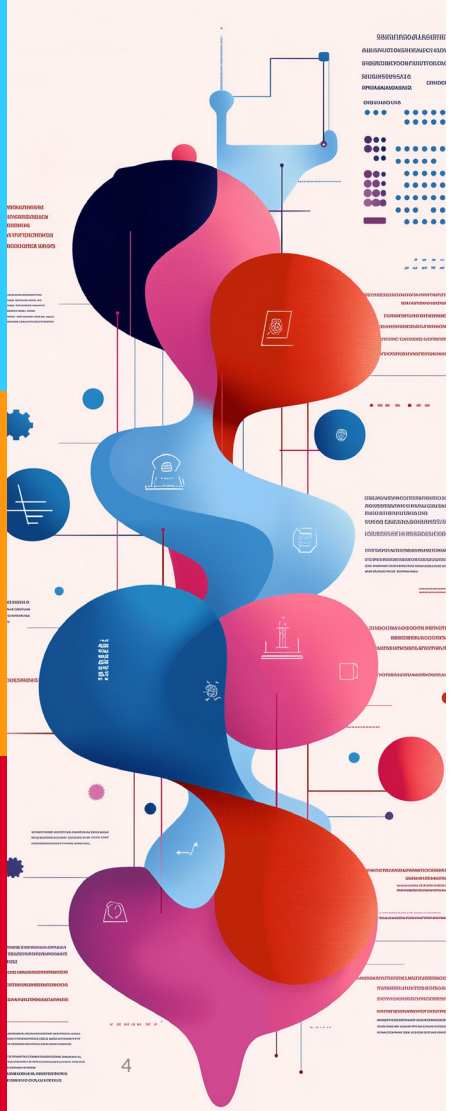
Les modèles de *machine learning* ont une durée de vie limitée en production en raison de phénomènes comme le *data drift*.

Au fil du temps, les données évoluent, ce qui peut entraîner une diminution des **performances** des modèles, rendant nécessaire leur mise à jour régulière. Gérer ces ajustements de manière manuelle peut rapidement devenir fastidieux et peu viable à grande échelle.

Le MLOps répond à ce besoin en offrant **un cadre d'automatisation qui facilite la gestion des différentes étapes du cycle de vie des modèles**. Cela inclut la **conception**, la **construction**, le **déploiement**, ainsi que la **surveillance** continue et la **gouvernance**.

En adoptant une approche MLOps, les organisations peuvent non seulement maintenir leurs modèles en production, mais aussi les améliorer de manière continue, en réduisant les interventions manuelles tout en assurant **robustesse** et **évolutivité** dans des environnements complexes.

Pourquoi le **MLOps** est-il important ?



Le MLOps pour le passage à l'échelle

Le MLOps est essentiel car il permet de structurer et d'automatiser les différentes étapes du cycle de vie des modèles de *machine learning*, garantissant ainsi leur efficacité et leur évolutivité. Voici quelques raisons clés pour lesquelles le MLOps joue un rôle central dans la gestion des projets ML.

1 Suivi et gestion

Dans la phase de conception, de nombreuses expériences sont menées pour optimiser les modèles. Le MLOps permet de garder une trace des différentes versions des modèles, ainsi que des ensembles de données utilisés, pour garantir une gestion rigoureuse des itérations. Cela aide à savoir précisément quelle version du modèle a été déployée et à comprendre l'impact des modifications.

2 Évaluation continue

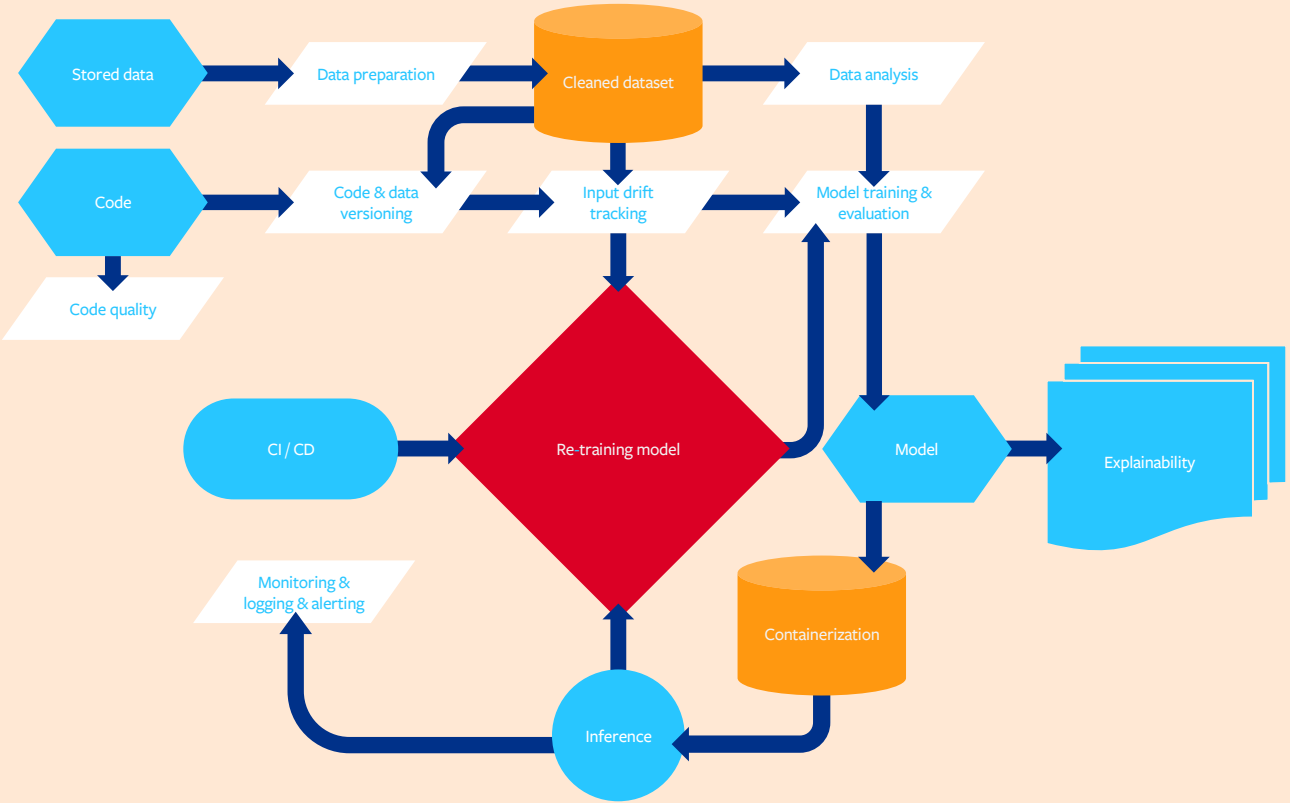
Avec le MLOps, les équipes peuvent évaluer si les modèles recyclés (re-entraînés avec de nouvelles données) surpassent réellement les versions précédentes. Cela permet de promouvoir uniquement les modèles les plus performants en production, en se basant sur des métriques claires et objectives. Il est possible de mettre en place des tests A/B pour s'assurer que le modèle chalenger soit bien le meilleur.

3 Surveillance proactive

Les performances des modèles peuvent se dégrader avec le temps à cause de l'évolution des données *data drift*. Le MLOps met en place des processus automatiques pour surveiller les performances des modèles à intervalles réguliers (quotidiennement, mensuellement, etc.), afin de détecter toute baisse de qualité et réagir rapidement, par exemple en re-entraînant le modèle avec de nouvelles données.

MLOps Workflow

Pour les définitions des termes employés, se référer au glossaire, page 20.



Vous êtes sur la piste du **MLOps** !

Principales fonctionnalités du MLOps

Source de données : il est important d'identifier des ensembles de données pertinents et fiables, de faciliter l'accès aux données et de vérifier leurs propriétés, telles que leur disponibilité en temps réel et leur mise à jour après le déploiement du modèle. Il est également essentiel de garantir la conformité avec les conditions d'utilisation et la protection des données personnelles (PII), tout en s'assurant que les populations minoritaires sont correctement représentées.

Type de déploiement : les modèles peuvent être déployés via des API REST ou dans des formats portables tels que ONNX. Dockeriser les applications permettent de simplifier le déploiement. Il est crucial de respecter les normes de codage, de valider la précision des modèles et leur explicabilité, et de s'assurer du respect des exigences de gouvernance et de la qualité des artefacts.

Reproductibilité : les paramètres et configurations des modèles doivent être sauvegardés pour permettre la reproduction des expériences.

Entraînement et évaluation : il est nécessaire de suivre et de comparer les résultats de chaque expérience à l'aide de métriques claires, en utilisant des outils de gestion pour assurer un suivi rigoureux et faciliter les comparaisons.

IA responsable : la compréhension et l'explicabilité des modèles doivent être assurées à travers une documentation automatisée et l'analyse des sous-populations. Des méthodes d'explicabilité, comme la valeur Shapley et l'ICE, permettent d'analyser l'impact des caractéristiques du modèle. Il est également important de limiter la complexité des modèles et de s'assurer que les jeux de données soient équilibrés.

Surveillance : les métriques liées à la vitesse d'exécution, à la consommation des ressources (CPU et mémoire), ainsi qu'aux performances des modèles et des données doivent être surveillées en continu.

Vers la **production** et au-delà...

Préparation à la production

Environnements d'exécution

Dans le cadre de MLOps, un système idéal permet un déploiement rapide et automatisé des modèles, particulièrement pour les processus nécessitant beaucoup d'efforts manuels. Les environnements de production peuvent varier : des services sur mesure, des plateformes de data science, des services dédiés comme TensorFlow Serving, ou des infrastructures basées sur des clusters Kubernetes et des machines virtuelles Java (JVM) sur des systèmes embarqués.

Conversion des modèles

Il est souvent nécessaire de convertir les modèles pour les adapter à l'environnement de production. Par exemple, un modèle développé en Python avec scikit-learn peut être converti au format ONNX (*Open Neural Network Exchange*) pour une production en C++. Cette étape garantit la compatibilité entre l'environnement de développement et celui de production.

Quantification des modèles

Par ailleurs, la quantification dans ONNX réduit la taille du modèle. Pendant ce processus, les valeurs flottante sont mappées sur un espace de quantification 8 bits. Cela diminue l'usage de mémoire et accélère les calculs, tout en préservant une précision suffisante, rendant les modèles plus adaptés aux environnements à ressources limitées.

Accès aux données avant la validation

L'environnement de production doit disposer d'un accès fiable aux bases de données et des pilotes nécessaires pour la communication. Il est également essentiel d'assurer un stockage suffisant. Si les données proviennent de sources externes, une copie doit être effectuée pour garantir l'accès en cas d'indisponibilité future des données.

Évaluation des risques du modèle

Il est indispensable d'évaluer les risques liés à l'utilisation du modèle en production. Cela inclut les scénarios où le modèle pourrait échouer ou être manipulé pour révéler des données sensibles. Les risques incluent des erreurs dans la conception, le faible alignement entre les données de production et celles d'entraînement, ou encore les attaques adversariales. Des questions juridiques, éthiques, ou de réputation doivent également être prises en compte.

Reproductibilité et auditabilité

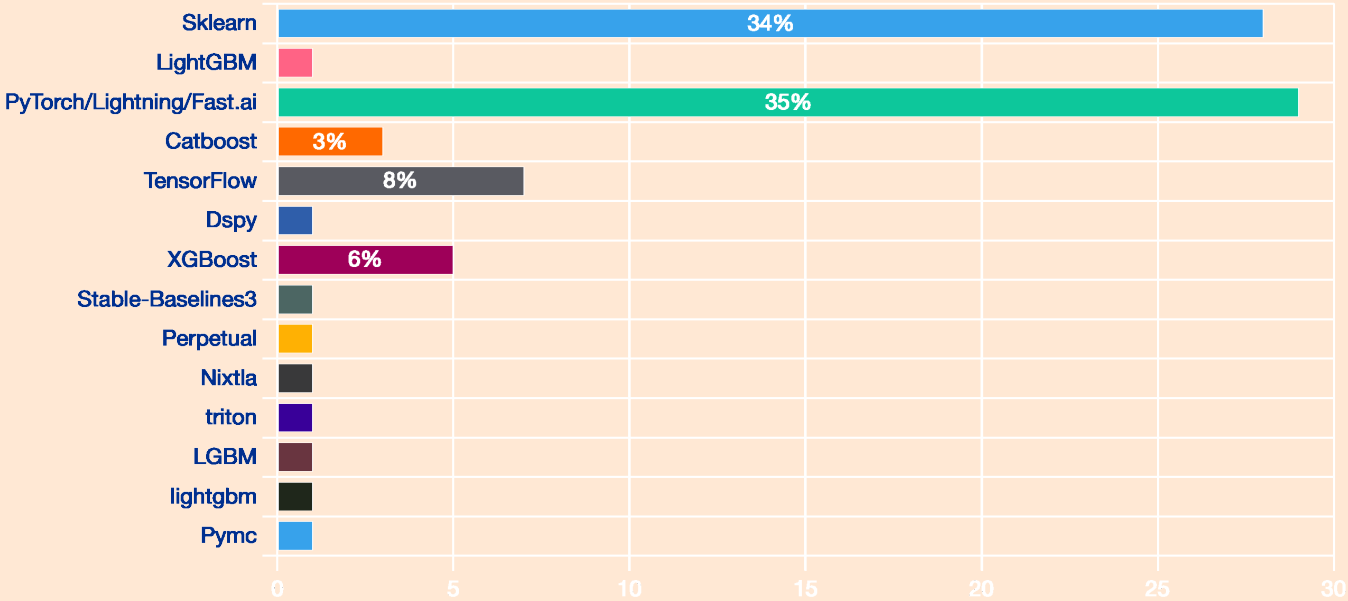
La reproductibilité en MLOps consiste à pouvoir reproduire exactement la même expérience avec le même modèle, la même documentation et les mêmes données. Cela permet non seulement de prouver les résultats du modèle, mais aussi de déboguer ou d'améliorer une expérience précédente. L'auditabilité, quant à elle, exige que l'historique complet du *pipeline* ML soit accessible, avec toute la documentation, les *artefacts*, et les résultats de tests associés, garantissant ainsi un suivi et une gestion fiables du modèle à chaque étape.

Conformité légale et préparation à l'IA Act

Dans un contexte de plus en plus régulé par des lois comme l'IA Act, ces principes de reproductibilité et d'auditabilité deviennent essentiels pour assurer la conformité légale des systèmes d'IA. En effet, l'IA Act, qui vise à encadrer les usages de l'IA dans l'Union Européenne, mettra en avant la transparence, la traçabilité, et la responsabilité des acteurs qui développent et déploient des systèmes d'IA. La reproductibilité permet de prouver le comportement du modèle et de le valider face aux régulations, tandis que l'auditabilité garantit que chaque décision ou action prise par le modèle peut être examinée et justifiée a posteriori. Ainsi, se conformer à ces exigences facilite non seulement la gestion des risques, mais aussi le respect des obligations légales.

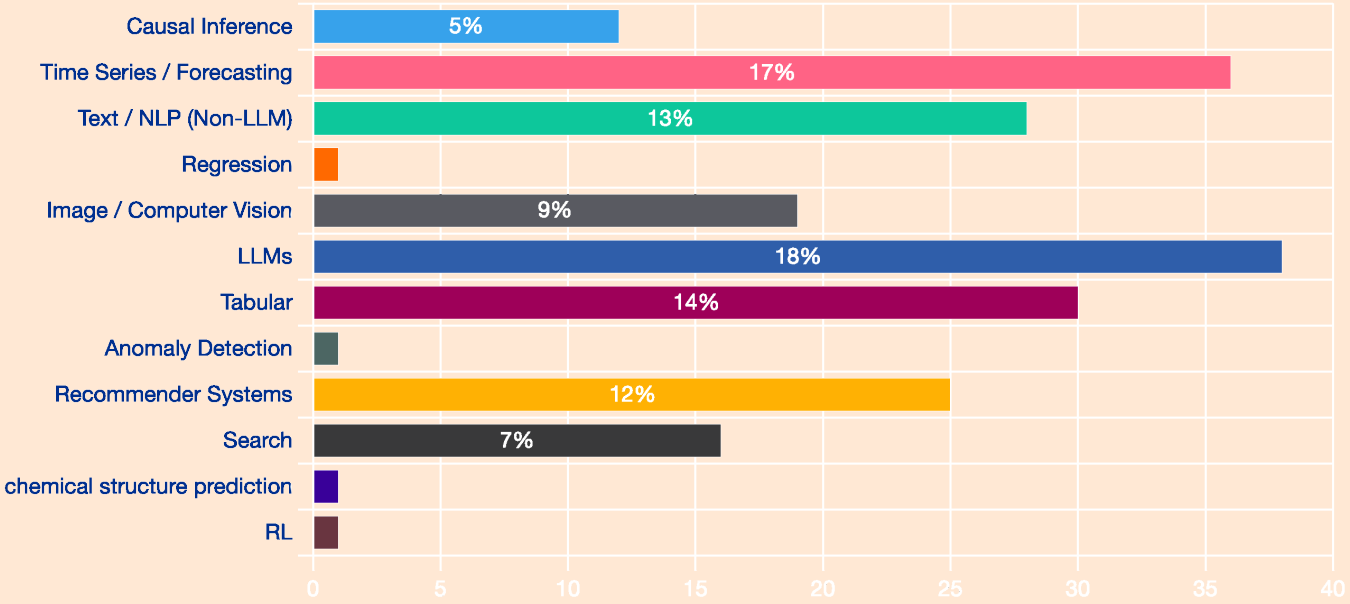
Quelle bibliothèque de *machines* utilisez-vous le plus ?

Source : The Institute for Ethical AI & Machine Learning, State of Machine learning 2024



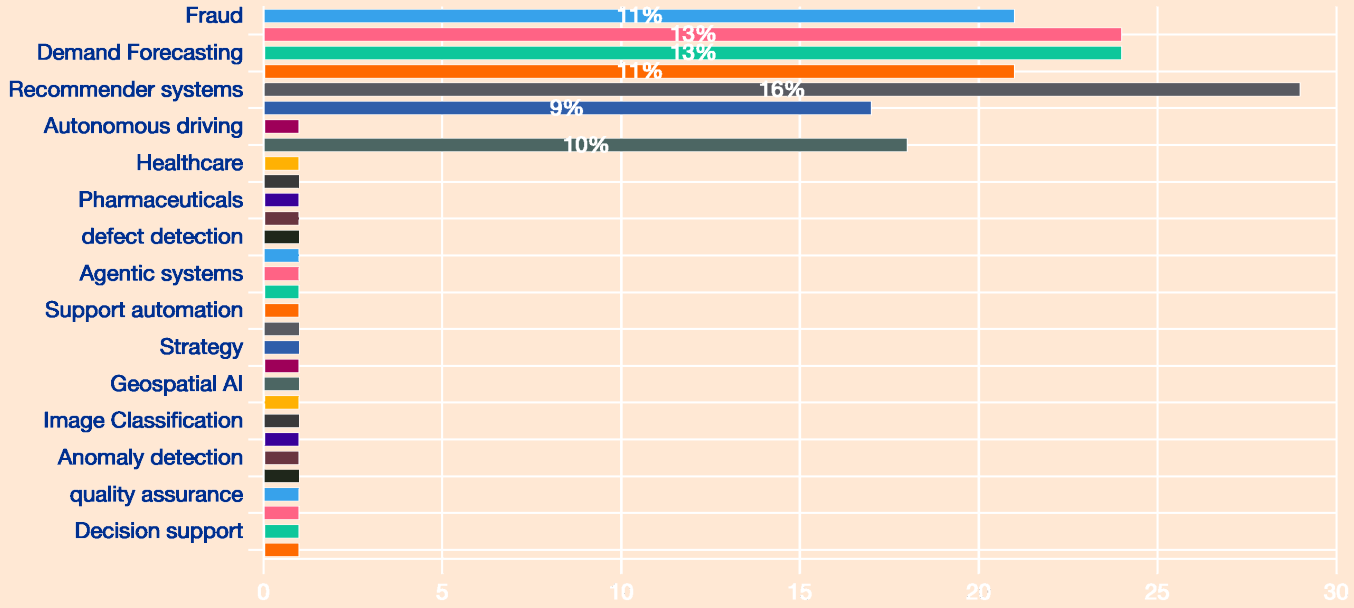
Sur quels domaines ou modalités du *machine learning* / science votre équipe travaille-t-elle ?

Source : The Institute for Ethical AI & Machine Learning, State of machine learning 2024



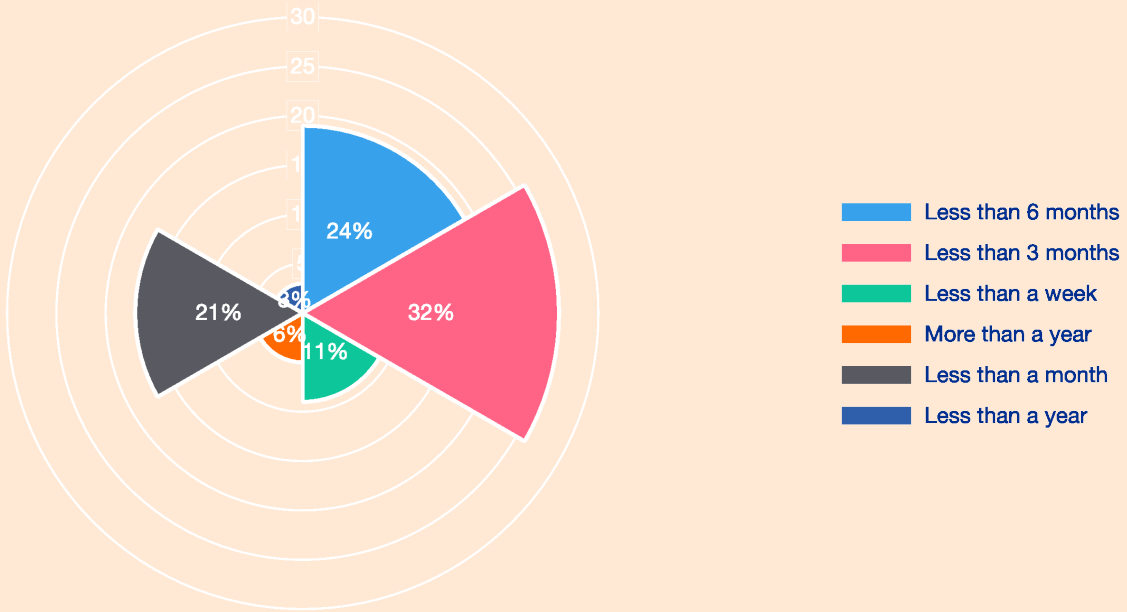
Pour quels cas d'utilisation votre équipe utilise-t-elle le *machine learning* ?

Source : The Institute for Ethical AI & Machine Learning, State of machine learning 2024



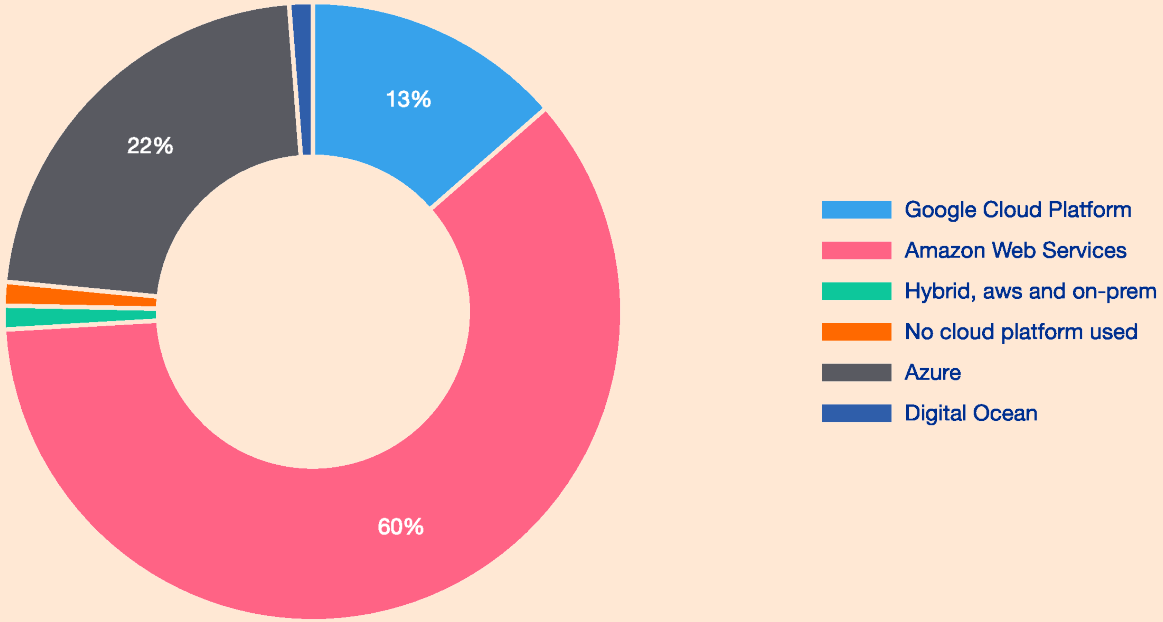
Combien de temps faut-il à votre équipe pour produire un modèle (y compris l'ingestion de données, etc.) ?

Source : The Institute for Ethical AI & Machine Learning, State of machine learning 2024



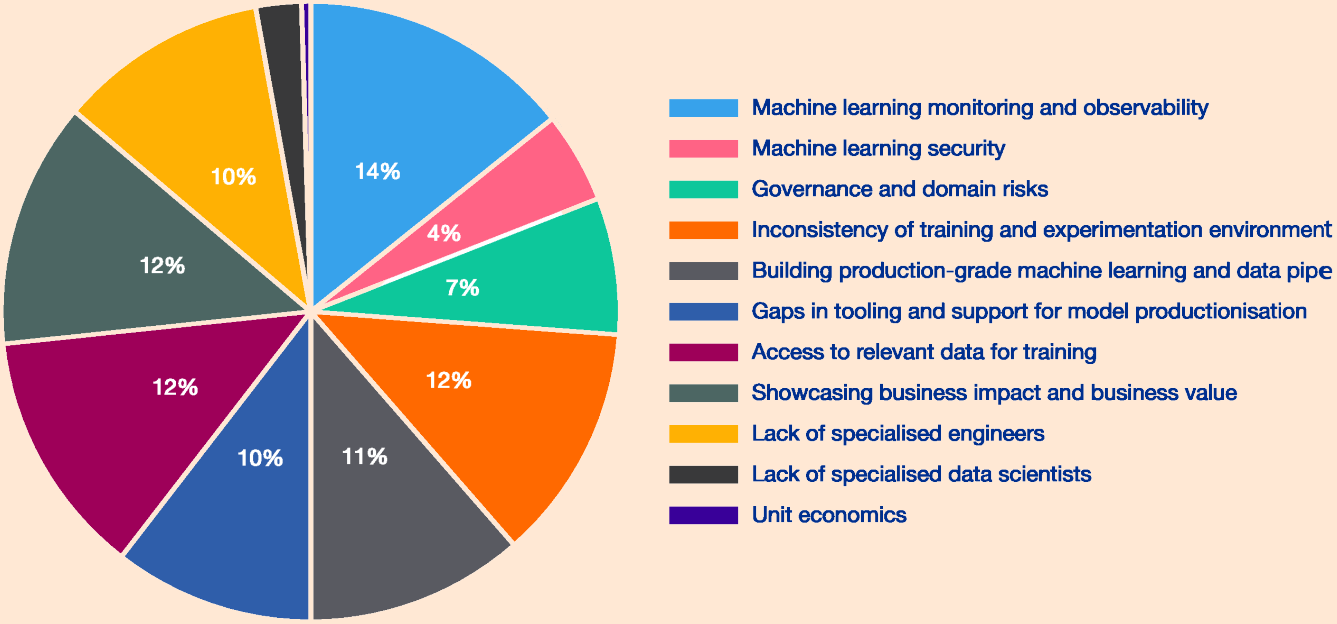
Quelle plateforme cloud utilisez-vous le plus ?

Source : The Institute for Ethical AI & Machine Learning, State of machine learning 2024



Sélectionnez les 3 plus grands défis auxquels vous êtes confrontés lors de la production de vos modèlesde *machine learning*

Source : The Institute for Ethical AI & Machine Learning, State of machine learning 2024



Les acteurs du MLOps

Au cœur du processus se trouvent les data scientists, responsables de la conception, de l'expérimentation, de la formation et de l'évaluation des modèles. Ils collaborent étroitement avec les data engineers, qui s'occupent de l'acquisition, de la préparation des données et de l'ingénierie des features, garantissant que les données sont prêtes pour entraîner les modèles de manière optimale. Les experts métiers jouent également un rôle clé en fournissant des questions commerciales pertinentes et en validant les modèles afin de s'assurer qu'ils répondent aux besoins spécifiques de l'entreprise.

Les équipes DevOps et MLOps (ou ModelOps) sont impliquées dans les phases de déploiement, veillant à l'automatisation, à la mise en conteneurs des modèles (par exemple via Docker), et à leur scalabilité. Après le déploiement en production, les pratiques DataOps assurent la surveillance continue, notamment en suivant les dérives des données (*input drift*) et des performances des modèles.

Ces indicateurs permettent aux équipes de réagir rapidement grâce à des alertes et des journaux d'activité, garantissant ainsi la performance continue des modèles.

Les ML engineers ou architects jouent un rôle crucial dans la préparation des modèles à la production, en s'assurant qu'ils respectent les normes de robustesse et de qualité. Ce flux de travail, qui forme une boucle fermée allant de la conception à la surveillance, permet une collaboration efficace entre les équipes, garantissant des modèles performants, fiables et évolutifs à chaque étape du cycle de vie.

Gouvernance & éthique, une IA responsable et transparente

IA responsable

La notion d'IA responsable influence considérablement la réflexion collective au sein de la communauté de l'IA, notamment en façonnant les approches et cadres réglementaires. Elle plaide pour **une approche collaborative, éthique et transparente du développement et du déploiement de l'IA**, garantissant que les technologies sont utilisées de manière à bénéficier à la société tout en minimisant les **biais** et les **risques**.

Les principales composantes de l'IA responsable incluent :

- **Intentionnalité** : les modèles d'IA doivent être conçus pour atteindre des objectifs précis et être compréhensibles au-delà de leurs développeurs.
- **Responsabilité** : un contrôle centralisé des efforts d'IA est nécessaire pour éviter les pratiques non conformes, avec une attention particulière à la traçabilité des données et à la conformité réglementaire.
- **Approche centrée sur l'humain** : il est crucial d'équiper les individus des outils et de la formation nécessaires pour appliquer les principes d'intentionnalité et de responsabilité.

Importance et éléments clés de l'IA responsable

L'IA responsable influence la manière dont les réglementations et les cadres sont élaborés, en plaidant pour une approche éthique et collaborative dans le développement de l'IA. Les éléments clés de cette gouvernance comprennent :

- **Équité** : les systèmes d'IA doivent être conçus de manière à ne pas introduire de biais discriminatoires basés sur le genre, l'ethnicité ou d'autres facteurs qui pourraient causer des inégalités. Des outils d'interprétation des modèles aident à identifier et à corriger ces biais pour garantir une prise de décision juste.
- **Fiabilité et sécurité** : les systèmes d'IA doivent être fiables et sécurisés, notamment dans des domaines critiques comme les véhicules autonomes ou les diagnostics médicaux. Des processus de tests rigoureux sont essentiels pour garantir que ces systèmes fonctionnent correctement avant leur déploiement.
- **Confidentialité et sécurité** : la protection de la vie privée est primordiale pour les systèmes d'IA, qui manipulent souvent des données sensibles. La confidentialité doit être assurée à toutes les étapes du cycle de vie des données, même après la mise en production.
- **Inclusivité** : les systèmes d'IA doivent être conçus pour bénéficier à l'ensemble de la société, sans distinction liée à l'origine ethnique, au genre ou aux capacités physiques. L'inclusivité est une composante clé du développement de l'IA.
- **Transparence** : les utilisateurs doivent comprendre comment fonctionnent les systèmes d'IA et connaître leurs objectifs, ainsi que leurs limites. Cela permet une meilleure adoption et une utilisation responsable de ces technologies.
- **Responsabilité** : les développeurs doivent être responsables de leurs systèmes d'IA et s'assurer qu'ils respectent les normes éthiques et légales. La responsabilité implique de suivre des cadres de gouvernance stricts pour éviter les abus et garantir un usage éthique de l'IA.

Les clés de la gouvernance

Comprendre et classifier les cas d'usage analytiques : identifier et catégoriser les cas d'usage analytiques en fonction des exigences réglementaires, du type de clientèle, des besoins en disponibilité, des impacts potentiels des erreurs, et d'autres facteurs critiques.

Définir une position éthique : établir la position de l'entreprise sur des questions éthiques comme l'équité, la confidentialité, les droits de l'homme et la responsabilité, influençant ainsi le comportement des modèles et les processus de gouvernance.

Déterminer les responsabilités : identifier les personnes clés dans les différents départements pour superviser la gouvernance, en s'appuyant sur des structures existantes et en favorisant une large implication.

Développer des politiques de gouvernance : Élaborer des politiques de base pour le processus MLOps, en ajustant les mesures de gouvernance en fonction du profil de risque et des exigences réglementaires de chaque initiative analytique.

Intégrer les politiques dans le processus MLOps : intégrer les politiques de gouvernance dans le flux de travail MLOps, en définissant les étapes du processus et en assignant les responsabilités pour chaque mesure de gouvernance.

Choisir des outils pour la gestion centralisée de la gouvernance : sélectionner des outils qui facilitent la gestion centralisée des processus de gouvernance, garantissant la traçabilité, la conformité et la collaboration entre les équipes.

Impliquer et former : communiquer l'importance de la gouvernance MLOps, en offrant des formations et de la documentation pour que chacun comprenne ses rôles et responsabilités.

Surveiller et ajuster : surveiller en continu l'efficacité des processus de gouvernance, ajustant les politiques et pratiques en fonction des indicateurs de performance, des évolutions réglementaires et des enseignements tirés.

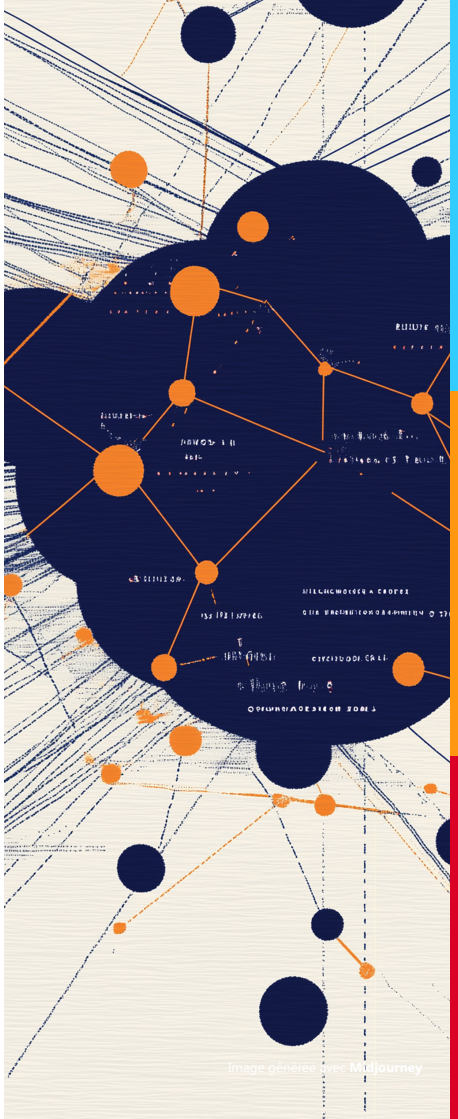
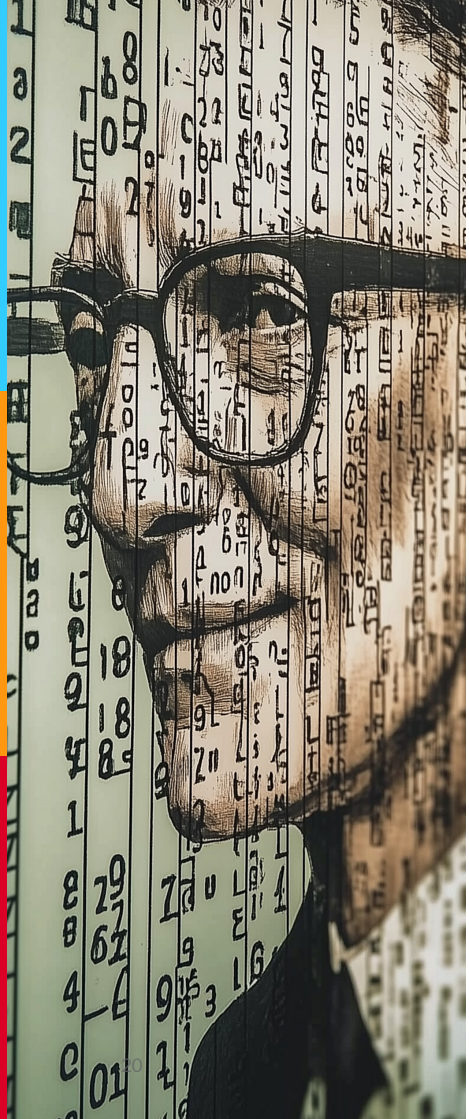


Image générée avec Midjourney



Glossaire

- ***Stored data*** : données brutes stockées avant tout traitement ou transformation.
- ***Data preparation*** : préparation des données pour les rendre exploitables par un modèle, incluant le nettoyage, la normalisation et la transformation des données.
- ***Data analysis*** : exploration des données pour en extraire des informations utiles, comme des tendances, des corrélations ou des insights.
- ***Feature engineering*** : création de nouvelles variables ou caractéristiques à partir des données brutes pour améliorer la performance du modèle.
- ***Code & data versioning*** : gestion des différentes versions du code et des données pour assurer la traçabilité et la reproductibilité des expériences.
- ***Code quality*** : mesure de la clarté, de la maintenabilité et de la performance du code, incluant des pratiques comme le linting, les tests unitaires et la revue de code.
- ***Input drift tracking*** : détection de la dérive des données d'entrée au fil du temps, pour identifier si les données utilisées pour le modèle changent significativement.
- ***CI / CD*** : intégration continue / Déploiement continu pour automatiser la livraison de nouvelles versions de code et de modèles en production.
- ***Re-training model*** : décision de recyclage du modèle lorsque sa performance diminue, souvent en réponse à un changement dans les données.
- ***Explainability*** : documentation expliquant comment un modèle fonctionne et pourquoi il prend certaines décisions, pour une meilleure transparence et compréhension.
- ***Model evaluation*** : évaluation de la performance d'un modèle à l'aide de métriques comme la précision, le rappel, la F1-score, etc.
- ***Monitoring, logging & alerting*** : suivi du système, enregistrement des événements, et alertes en cas d'anomalies pour garantir la fiabilité des modèles en production.
- ***Containerization*** : technique d'encapsulation d'applications dans des conteneurs isolés pour assurer portabilité, cohérence et déploiement rapide sur différents environnements.
- ***Scalability*** : capacité à adapter les ressources et les infrastructures pour traiter un volume croissant de données ou de demandes d'inférence.
- ***Inference*** : exécution d'un modèle pour produire des estimations, prévisions ou classifications sur de nouvelles données.
- ***Model deployment*** : Processus de mise en production d'un modèle, permettant son utilisation en temps réel ou par batch pour des prédictions.

keyrus

make data matter

Acteur international du conseil et des technologies, Keyrus a pour mission de donner du sens aux données, en révélant toute leur portée, notamment sous un angle humain.

Parce que ce ne sont pas tant les données elles-mêmes qui importent, mais les opportunités que nous pouvons développer en les apprivoisant vraiment, nous nous efforçons constamment de comprendre les objectifs que nos clients souhaitent atteindre. Nous explorons et mesurons les comportements, nous les comprenons et les traduisons en un résultat concret. Nous donnons un sens aux réalités que les données portent afin d'aider nos clients à prendre des décisions plus efficaces.

Les données, qu'elles soient grandes, petites, humaines, complexes, historiques ou prospectives, n'ont de sens que lorsqu'elles sont utilisées pour développer les expériences, affiner la compréhension du quotidien et prendre les meilleures décisions.

Notre proposition de valeur est fondée sur cinq grands groupes de services, chacun comprenant des offres multiples :

- **Automatisation et intelligence artificielle** : nous fournissons à nos clients les moyens d'améliorer leur productivité et leur précision sur l'ensemble de leurs processus, afin de se concentrer sur le travail à plus forte valeur ajoutée.
- **Expérience numérique centrée sur l'humain** : la relation avec les clients et l'engagement des collaborateurs constituent deux des plus grands contributeurs au succès global des entreprises. Nous aidons les entreprises à imaginer et à créer des expériences numériques multimodales et fluides pour atteindre leurs objectifs.
- **Mise en œuvre des données et des analyses** : les données sont une clé incontestable du succès pour les entreprises. Lorsqu'elles sont utilisées intelligemment, elles ouvrent des opportunités uniques pour faire face aux défis actuels et futurs. Nous permettons aux organisations de déployer tout le potentiel de leurs données : nous mettons la science des données au profit du développement de l'entreprise.
- **Cloud et sécurité** : le Cloud et les plateformes numériques ont le potentiel de révolutionner la façon dont les données sont transformées en valeur, tout en portant l'extensibilité et la flexibilité à un niveau supérieur. Nous sécurisons l'ensemble de vos données et veillons à ce qu'elles soient protégées et confidentielles.
- **Transformation et innovation** : pour prospérer dans l'écosystème actuel, chaque entreprise doit non seulement accélérer sa transformation numérique, mais aussi acquérir des compétences pour stimuler son adaptabilité, sa résilience et sa compétitivité. Nous aidons nos clients à se transformer avec succès pour développer un meilleur futur.

S'appuyant sur l'expérience cumulée de plus de 3 500 collaborateurs et présent dans 27 pays sur 4 continents, Keyrus est l'un des principaux experts internationaux en matière de données, de conseil et de technologie.

Pour en savoir plus : **www.keyrus.fr**

Jean-Philippe CLAIR
Directeur Marketing, Communication & Expérience client
jean-philippe.clair@keyrus.com