Qu'apportent les métiers de la *data* à la statistique publique ?

Lino Galiana[*] *et Olivier Lefebvre*[**]

La numérisation progressive des données a profondément transformé la production d'information, dans un contexte où les progrès technologiques et méthodologiques ont amélioré l'efficience des outils informatiques. Parallèlement, la société s'est acculturée à disposer en permanence de statistiques dans le débat public et dans le quotidien, et les entreprises utilisent de plus en plus de données dans leurs processus de décision. Face à ces besoins, il existe une forte demande sur le marché du travail pour des profils spécialisés dans la valorisation de données : *data analysts*, *data scientists*, *data engineers*, *data journalists*…

Ces profils favorisent un rapprochement des compétences entre directions métiers et informatiques, ce qui a de nombreuses vertus : possibilité de réduire les délais de livraison, capacité à associer des données de formes et de natures diverses pour répondre à de nouveaux besoins, production plus transparente… Mais cela représente aussi un défi d'adaptation continue, dans un environnement où les besoins, les outils et les méthodes évoluent rapidement.

La statistique publique tire pleinement parti de ces transformations au service de missions inchangées (éclairer le débat public et aider à la décision publique) et de valeurs essentielles (rigueur, transparence, respect de la confidentialité…). L'Insee et les services statistiques ministériels se sont dotés d'infrastructures et d'organisations pour explorer, expérimenter, industrialiser et maintenir ces nouveaux processus.

The progressive digitisation of data has profoundly transformed the production of information, at a time when technical and methodological advances have improved the efficiency of computing tools. At the same time, society has grown accustomed to constantly having statistics at its disposal in public consultations and in its everyday life. Companies also use more and more data in their decision process. To meet those needs, there is a high demand on the market for profiles specialised in maximising the value of data: data analysts, data scientists, data engineers, data journalists…

These profiles foster skill alignment between business-decision actors and IT functions, bringing many benefits: opportunities to reduce delivery time, matching different types and forms of data to meet new needs, producing more transparently… But this also represents a challenge of constant adaptation in an environment where needs, tools and methods change rapidly.

Official statistics make the most of these transformations while serving the same missions (enlightening public debate and supporting public decision-making) and core values such as discipline, transparency, respect of privacy. INSEE and the Ministerial Statistical Offices have developed infrastructures and organisational frameworks to explore, experiment, industrialise and keep those new processes sustainable.

* Chef de la section « Expertise de la fonction publique », Insee.
lino.galiana@insee.fr

** Direction des statistiques démographiques et sociales, Insee.
olivier.lefebvre@insee.fr

> ***La généralisation des traces numériques a profondément transformé les métiers de la donnée.***

Affirmer que le métier de *data scientist* est « le boulot le plus sexy du XXIe siècle » (Davenport et Patil, 2012), ou que la donnée est « le pétrole du XXIe siècle », est quelque peu emphatique et provocateur. Mais cela reflète une réalité. La donnée est d'abord devenue un objet du quotidien. Puis, progressivement et moyennant des investissements conséquents, elle est devenue un actif central dans les économies numérisées. Son analyse ouvre des champs d'innovation stimulants et l'opportunité d'un enrichissement des connaissances.

La généralisation des traces numériques[1], alliée aux progrès techniques et à un appétit croissant pour les indicateurs chiffrés, a profondément transformé les métiers de la donnée (ou *data*). La donnée est désormais omniprésente et guide des politiques publiques, des stratégies d'entreprises, des choix individuels et des décisions collectives. Elle est davantage qu'un simple intrant informatique : elle constitue un matériau à exploiter, à transformer et à façonner pour créer de la valeur.

Cette évolution a fait émerger de nouveaux profils, comme les *data scientists* et les *data engineers*, qui viennent compléter les compétences des statisticiens (Volle, 1980), des analystes et des développeurs. La coordination de ces multiples profils est désormais indispensable pour traiter des données massives, hétérogènes et souvent non structurées, et pour en faire un levier d'innovation et de connaissance. Ce matériau est protéiforme. Il impose au statisticien d'adapter ses outils en permanence, et d'intervenir plus en amont de la chaîne de fabrication de statistiques.

Dans le champ de la statistique publique, cette hybridation des compétences, jointe à la construction d'environnements de traitement adaptés, a permis d'élargir les méthodes, de mieux exploiter de nouvelles sources de données, sans sacrifier à la transparence. Cette dernière est en effet la garantie, in fine, de la confiance dans les chiffres produits.

▶ La *data science* pour tirer parti d'un univers de données en expansion et en évolution

Un stock de données, de toutes formes, en explosion

Les baisses continues des coûts de stockage et de traitement ont joué un rôle de premier plan dans la massification de la production de données. Par ailleurs, Internet, les smartphones et les objets connectés créent des traces numériques directes qui peuvent être accumulées dans des entrepôts de données. Ces derniers sont devenus peu coûteux[2] : un térabyte de stockage[3] coûtait seulement 11 dollars en 2023, contre 6,5 millions de dollars en 1990[4].

1 Informations enregistrées par un dispositif numérique sur l'activité ou l'identité de ses utilisateurs (source : Wikipédia).
2 La logique d'accumulation à faible coût de données, sans définir en amont les besoins liés aux cas d'usage, a créé un grand engouement pour le big data pendant les années 2010. Ce paradigme est aujourd'hui critiqué par des acteurs comme Jordan Tigani (créateur de BigQuery, l'offre d'outil big data de Google). Pour en savoir plus, voir Tigani (2023).
3 C'est l'unité de mesure la plus grande pour un usage courant. Les centres informatiques nécessitent quant à eux des unités de mesure encore plus grandes.
4 Voir le lien suivant : *https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage*.

Les outils numériques, devenus omniprésents, sont aussi créateurs de données diversifiées : valeurs quantitatives classiques mais à haute fréquence, textes, images, vidéos, positions GPS, traces informatiques (*log files*[5])… Cette diversification des types de données et des contenus a permis d'ouvrir l'éventail des thèmes concernés par la *data* et des acteurs économiques impliqués dans sa valorisation. Par ailleurs (opportunité liée à la disponibilité de ce type d'information ou souci d'objectivation ?), nous faisons une consommation effrénée, voire boulimique, de données dans notre vie quotidienne : nous comptons nos pas, choisissons nos restaurants en fonction de la note moyenne donnée par les consommateurs, etc.

Un besoin croissant en ressources de calcul

Cette explosion du volume de données a entraîné un besoin croissant en ressources de calcul. Elle a favorisé des innovations comme, à partir de la fin des années 2000 :

- le calcul distribué, qui permet de résoudre des problèmes complexes en répartissant le travail entre plusieurs ordinateurs ;
- le calcul scientifique sur GPU (pour *Graphics Processing Unit* en anglais, ou processeur graphique en français), qui permet d'accélérer les traitements en parallélisant les tâches nécessitant peu de mémoire.

Les années 2010 ont par ailleurs été marquées par un recours de plus en plus fréquent à des infrastructures aux ressources mutualisées, notamment dans le *cloud*[6]. Ces dernières permettent de bénéficier d'une plus grande puissance de calcul et d'une sécurité accrue par rapport aux micro-ordinateurs.

> ***La convergence entre accumulation de données, augmentation de la puissance de calcul et progrès des logiciels de programmation a redynamisé des champs de recherche restés à l'état de promesses.***

La convergence entre accumulation de données, augmentation de la puissance de calcul et progrès des logiciels de programmation a redynamisé des champs de recherche existant dès les années 1950, mais longtemps restés à l'état de promesses faute de moyens technologiques adéquats. C'est notamment le cas des réseaux de neurones, une des grandes méthodes d'intelligence artificielle[7]. Entre les premières formalisations académiques des années 1950 et leur adoption industrielle à la fin des années 2000, ils ont eu peu de succès. En effet, peu d'infrastructures étaient suffisamment puissantes pour les mettre en œuvre, mais aussi, rares étaient les cas d'usage justifiant une telle complexité de modélisation.

5 Un *log file* ou *log* est un fichier informatique contenant l'historique des événements affectant un processus.
6 Ensemble des serveurs distants proposant des services accessibles par le réseau Internet (source : Le Robert).
7 Méthode permettant la résolution de problèmes complexes, comme la vision par ordinateur. Voir le glossaire mis en ligne par la Commission nationale de l'informatique et des libertés (Cnil) à l'adresse suivante : *https://www.cnil.fr/fr/glossaire*. Voir Russell et Norvig (2021) pour une histoire détaillée des cas d'usage de l'intelligence artificielle.

Les grands modèles de langage (LLM)[8] qui prennent aujourd'hui place dans notre quotidien, par exemple pour répondre à des questions administratives ou commerciales à la place de conseillers humains, s'inscrivent dans ce mouvement. Ils sont dans la continuité directe des travaux des années 2010, une décennie riche d'innovations dans le domaine de l'analyse textuelle, elles-mêmes rendues possibles par l'entraînement de modèles sur des corpus textuels de grande ampleur. Certaines tâches, comme l'entraînement de modèles de fondation[9], nécessitent néanmoins une puissance financière importante. Ainsi, l'entraînement du modèle de langage GPT-4, développé par la société OpenAI en 2023, aurait coûté 80 millions de dollars (Maslej et al., 2025).

Une donnée qui circule beaucoup plus

Les quinze dernières années sont marquées par une ouverture progressive des données (open data), et une multiplication des acteurs associés, en France comme à l'international (Kitchin, 2021). Ce mouvement concerne notamment les données du secteur public, considérées comme un levier pour éclairer le débat citoyen et les décisions de politiques publiques (Bothorel, 2020).

Les quinze dernières années sont marquées par une ouverture progressive des données (open data), et une multiplication des acteurs associés, en France comme à l'international (Kitchin, 2021).

En France, cette dynamique s'appuie à la fois sur :

- des dispositifs institutionnels : création d'Etalab[10] et de la plateforme de partage de données publiques data.gouv.fr en 2011, créations de postes d'administrateurs ministériels des données, des algorithmes et des codes sources à la fin des années 2010 ;
- des outils et cadres techniques, comme les interfaces de programmation applicative (API pour *application programming interface* en anglais)[11] ou le Référentiel général d'interopérabilité (RGI)[12] ;
- un cadre législatif (loi du 7 octobre 2016 pour une République numérique[13]) qui facilite l'accès des citoyens, chercheurs et entreprises aux données produites par les administrations, tout en garantissant la protection de la vie privée, en conformité avec la loi Informatique et Libertés[14] et le règlement général pour la protection des données (RGPD)[15].

8 LLM pour *Large Language Models* en anglais. Un modèle de langage peut par exemple prédire le mot suivant dans une séquence de mots (voir le glossaire de la Cnil).
9 Modèles d'intelligence artificielle « généralistes », qui peuvent être adaptés à une grande variété de tâches (voir le glossaire de la Cnil).
10 À la suite du décret nº 2011-194 du 21 février 2011 portant création d'une mission « Etalab » chargée de la création d'un portail unique interministériel de données publiques. Voir les références juridiques en fin d'article.
11 Une API est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service, afin d'échanger des données et des fonctionnalités.
12 Le RGI est un cadre de recommandations référençant des normes et standards qui favorisent l'interopérabilité au sein des systèmes d'information de l'administration.
13 Voir les références juridiques en fin d'article.
14 Loi nº 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Voir les références juridiques en fin d'article.
15 Règlement (UE) nº 2016/679 du Parlement européen et du Conseil du 27 avril 2016. Voir les références juridiques en fin d'article.

La loi pour une République numérique favorise par ailleurs la transmission, pour la production de statistiques publiques, de données détenues par des entreprises privées. Une disposition du même ordre figure depuis fin 2024 dans le règlement européen sur la production de statistiques[16].

Pour les données individuelles, plus sensibles, la diffusion à des fins de recherche a été encouragée, dans un cadre simplifié, tout en conservant un haut niveau de sécurité, grâce notamment au Centre d'accès sécurisé aux données (CASD) (Gadouche, 2019). La diversification des producteurs de données s'est aussi traduite par une ouverture parallèle de données par des entreprises privées, que ce soit gratuitement ou dans une logique de monétisation (Lesur, 2025).

L'effort de normalisation et d'harmonisation constitue une condition préalable pour tirer parti de données de plus en plus nombreuses et hétérogènes. Au-delà de l'interopérabilité technique[17], qui a progressé avec la diffusion des API et de formats de données standardisés (Dondon et Lamarche, 2023), la réutilisation effective suppose aussi une contextualisation suffisante. Il faut en effet pouvoir connaître un minimum l'univers métier dans lequel sont conçues les données pour éviter des incohérences, des interprétations erronées voire des mésusages, ce que la littérature désigne par le terme anglais *data friction* (Edwards et al., 2011). Extraire de l'information pertinente d'une photo, d'un scan de documents, d'un texte ou d'un graphe de relations est en effet plus difficile que si l'on part d'une information déjà organisée et construite pour une finalité bien définie.

Mais au-delà de ces enjeux techniques, l'augmentation et la diversification des producteurs de données transforment également les pratiques : elles ouvrent la voie à des usages plus empiriques et variés.

Des approches plus empiriques, partant davantage de la donnée que de la théorie

Ces données, plus nombreuses, plus accessibles, créent des occasions de valorisation à saisir, parfois imprévues dans leur usage initial.

Le premier usage opportuniste consiste en une mutualisation de données entre plusieurs acteurs lorsque les conditions de partage le permettent. C'est notamment le cas dans le domaine des démarches administratives avec le principe du « Dites-le-nous une fois ». En permettant des transferts de pièces justificatives entre administrations, il simplifie les formalités des citoyens[18] et réduit les risques d'erreurs ainsi que les délais de traitement.

Les données peuvent aussi bénéficier d'une seconde vie en étant réutilisées, souvent par de nouveaux acteurs, à des fins différentes de celles prévues lors de leur conception. Parmi les exemples classiques, on trouve l'utilisation de données de téléphonie mobile pour estimer les flux de population. On détourne alors les traces numériques de leur

16 Règlement (UE) n° 2024/3018 modifiant le règlement (CE) n° 223/2009 relatif aux statistiques européennes. Voir les références juridiques en fin d'article.
17 Capacité de matériels, de logiciels ou de protocoles différents à fonctionner ensemble et à partager des informations (source : Larousse).
18 Par exemple quand une attestation de quotient familial de la caisse d'allocations familiales est nécessaire pour bénéficier de la gratuité d'un service public local.

usage premier – le bornage aux antennes, qui permet d'acheminer les données vers l'utilisateur – vers un autre usage – par exemple l'estimation du nombre de visiteurs d'un événement sportif ou festif (voir aussi Joubert, 2025, pour différents usages intéressant la statistique publique).

> ***Les données peuvent bénéficier d'une seconde vie en étant réutilisées à des fins différentes de celles prévues lors de leur conception.***

Cette logique empirique se retrouve aussi dans les méthodes de traitement, notamment à travers le succès des algorithmes prédictifs. En l'occurrence, l'empirisme renvoie au fait que ces méthodes, comme l'apprentissage automatique (machine learning)[19], ne reposent pas sur une théorie préalable mais sur la détection de corrélations utiles à la création de règles de décision. Leur mise en œuvre repose en partie sur de nouveaux métiers, avec une frontière plus ou moins marquée entre les rôles selon les organisations. Les ***data scientists*** développent les expérimentations de modèles, les ***data engineers*** assurent leur intégration dans les systèmes et leur mise à l'échelle, tandis que les ***data analysts*** font le lien entre les indicateurs découlant des traitements et les recommandations à des fins de décision (***figure 1***).

▶ Une évolution des manières de travailler

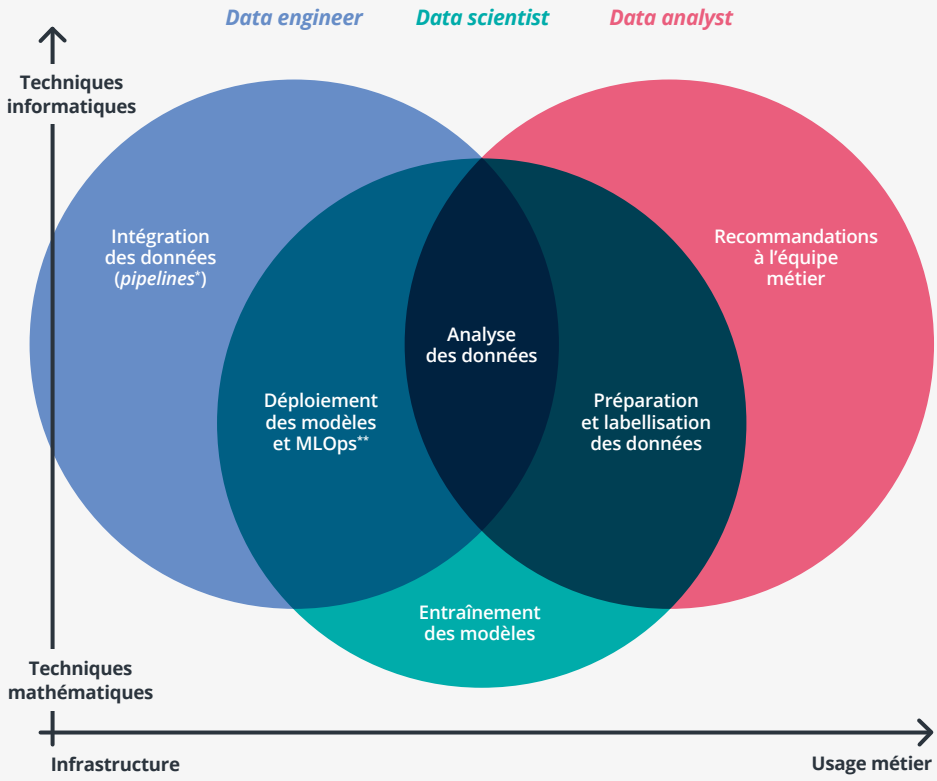
Pour que la statistique publique bénéficie pleinement de ces opportunités, les manières de travailler doivent évoluer. En particulier, deux exigences s'imposent, qui sont étroitement liées : celle de la réplicabilité des traitements et celle de l'ouverture des programmes informatiques (open source). En effet, les données produites doivent être fiables, comparables et suivies dans le temps et les méthodes qu'emploient les statisticiens publics gagnent à être transparentes.

L'exigence de réplicabilité au cœur des chaînes de traitement

Dans un contexte où données, environnements de travail et outils évoluent constamment, pouvoir répliquer un traitement ou reproduire des résultats dans des conditions plus ou moins différentes de leur élaboration d'origine s'avère indispensable pour mettre en production de manière fluide des projets expérimentaux. Les acteurs numériques recourent de plus en plus souvent à des environnements *cloud*, où le principe du microservice prévaut, c'est-à-dire à des environnements logiciels temporaires recréés à la demande. Dans ce contexte, l'adoption de pratiques favorisant la réplicabilité devient indispensable sous peine de ne pas être en mesure de reproduire dans le temps une chaîne de production, ou de simuler l'impact de modifications, en vue d'une optimisation du processus ou d'une analyse de la qualité des données produites.

19 Champ d'étude de l'intelligence artificielle qui vise à donner aux machines la capacité d'« apprendre » à partir de données, via des modèles mathématiques (voir le glossaire de la Cnil).

▶ Figure 1 - Le halo des métiers de la *data*



* Traitements de données standardisés.
** MLOps : Voir ***encadré 1***.
Note : Selon les organisations, le positionnement et la taille des cercles peut varier.

Parmi les solutions techniques permettant de répliquer une série de programmes dans son intégralité, la conteneurisation (Comte et al., 2022) occupe une place centrale. Cette approche consiste à créer, pour chaque traitement, un environnement technique maîtrisé de A à Z garantissant sa reproductibilité. Cette approche renforce l'autonomie technique des *data scientists* et *data engineers* ; les premiers sont autonomes pour définir précisément l'environnement utilisé par leurs prototypes que les seconds peuvent ainsi industrialiser plus facilement.

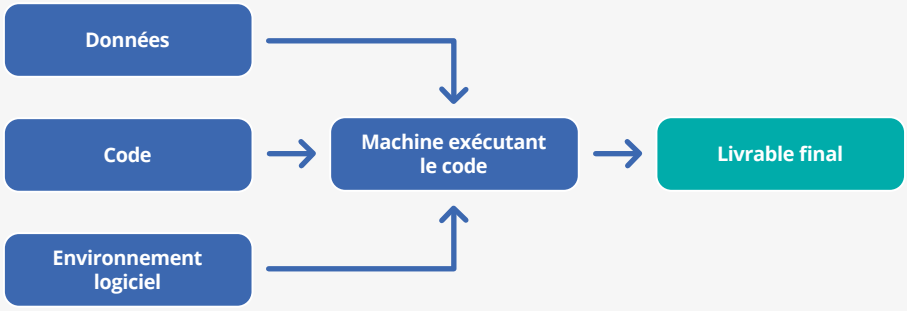
Plutôt que de reconstruire un environnement sur la base de spécifications, les *data engineers* peuvent donc l'adapter directement pour le passage à l'échelle. Le rôle d'un *data engineer* se distingue ainsi de celui d'un développeur applicatif classique ; il ne s'agit pas de développer une chaîne répondant aux besoins du métier (*data scientists* ou *data analysts*), mais d'intégrer et d'organiser les données au sein d'un écosystème plus vaste de services (entrepôts de données consommables par des logiciels statistiques, tableaux de bord, etc.), offrant aux utilisateurs finaux une plus grande autonomie dans l'exploitation de l'information brute disponible.

> ***Parmi les solutions techniques permettant de répliquer une série de programmes dans son intégralité, la conteneurisation occupe une place centrale.***

Afin d'offrir la flexibilité attendue dans un écosystème mouvant, les pratiques actuelles reposent sur la séparation conceptuelle des trois composantes de l'exécution du traitement : stockage des données, stockage du code et définition de l'environnement logiciel nécessaire au fonctionnement de la chaîne ***(figure 2***). Grâce à cette organisation, il est possible de faire évoluer séparément chaque composante sans fragiliser l'ensemble. Notamment, des solutions techniques permettent d'expérimenter un traitement nouveau en parallèle de la stabilisation d'une version en production.

L'articulation de ces éléments demande néanmoins un savoir-faire spécifique, ce qui explique le rapprochement entre les compétences d'analyse de la donnée et celles de développement informatique, tout comme la fourniture d'un environnement qui en permette la mise en œuvre.

Figure 2 - Architecture d'une infrastructure *cloud* moderne au service des *data scientists* et des *data engineers*



L'open source, un levier pour des chaînes de traitement flexibles et lisibles

Dans un contexte où les intrants d'une chaîne de production (notamment les données) évoluent fréquemment, il est nécessaire d'être en capacité d'adapter en permanence les outils de traitement. À cet égard, l'open source[20] se révèle être un levier stratégique de première importance. Utiliser des solutions construites et éprouvées par une communauté aux besoins similaires permet d'adopter des outils plus agiles, évolutifs et collaboratifs que les solutions propriétaires[21]. Ces dernières sont en effet souvent figées pour garantir leur stabilité (ou la « captivité » de leurs utilisateurs).

20 Un logiciel open source se caractérise par le fait que son code source est libre d'accès, réutilisable et modifiable.
21 Un logiciel propriétaire se caractérise par le fait que son code source est fermé, i.e. n'est pas accessible au public.

L'avantage des logiciels open source va bien au-delà de leur gratuité. En premier lieu, ils peuvent s'adapter plus vite aux besoins de leurs utilisateurs, lesquels peuvent eux-mêmes développer des modules, accélérant ainsi le cycle d'intégration de fonctionnalités. Surtout, ces logiciels apportent de la transparence par rapport aux logiciels propriétaires. Contrairement à ces derniers, qui ne communiquent pas forcément sur les choix méthodologiques, ni sur la manière dont le code les met en œuvre, les outils open source permettent de voir et de contrôler l'ensemble du processus.

> ***Il est nécessaire d'être encapacité d'adapter en permanence les outils de traitement. À cet égard, l'open source se révèle être un levier stratégique de première importance.***

Cette logique fait du code un objet de communication entre des équipes amenées à collaborer sur une chaîne de valeur de données. Ceci explique que les praticiens de la donnée se soient appropriés des outils et des pratiques développés à l'origine dans la communauté du développement logiciel. C'est notamment le cas du logiciel Git[22], qui est progressivement devenu un outil du quotidien. Il permet d'abord de tracer de manière fine les évolutions dans le temps d'un projet et de fiabiliser les échanges de programmes informatiques, du développement au déploiement. Mais aussi, il constitue une porte d'entrée pour les statisticiens et *data scientists* vers l'adoption de pratiques agiles[23] et collaboratives, propices à des projets plus transparents, de meilleure qualité, et plus réactifs. Ce recours à des outils facilitant la réplicabilité fait par ailleurs écho à un débat scientifique plus large sur la « crise de la reproductibilité » (Ioannidis, 2005 ; Kapoor et Narayanan, 2023). Ce débat a amené à l'adoption, par la communauté académique, de règles et pratiques plus exigeantes pour favoriser la transparence et la réplicabilité des publications (Vilhuber, 2020).

La transparence permise par l'open source nécessite cependant une vigilance accrue sur la qualité des composants intermédiaires. Là où un logiciel propriétaire propose, pour un besoin donné, un nombre restreint de manières d'y répondre et des solutions guidées par l'éditeur du logiciel, l'écosystème open source peut mettre à disposition plusieurs outils. Ces derniers n'intègrent pas forcément les mêmes fonctionnalités ou ne sont pas de qualité équivalente. Cette logique responsabilise ainsi les équipes développant des codes. Car la maintenance d'une chaîne de production ne passe pas que par les mises à jour de l'infrastructure, mais aussi par le suivi des différentes briques d'un projet : il convient de s'assurer que les solutions envisagées à un moment restent pertinentes et opérationnelles.

22 Git est un logiciel libre qui permet de gérer les différentes versions d'une application informatique et qui facilite la collaboration entre les développeurs.
23 La méthode agile est une méthode de gestion de projet favorisant la collaboration, la flexibilité et l'adaptabilité. Elle se concentre sur des cycles de développement courts et itératifs.

De formidables opportunités à mettre au service de la statistique publique

L'univers de la donnée et de l'intelligence artificielle connaît une expansion accélérée, stimulée par des avancées technologiques qui en facilitent la collecte, la circulation et l'utilisation. Cette profusion de données ou d'outils impose de nouvelles manières de travailler et de s'organiser, et accélère les transformations qui sont en œuvre.

Il y a quelques décennies, le modèle dominant était celui des enquêtes. Dans ce dernier, le statisticien était maître de son « destin statistique », à savoir la conception de l'enquête, l'échantillonnage, le plan d'exploitation, l'analyse et la diffusion. En revanche, il était dépendant de collègues informaticiens pour la réalisation des outils de traitement et des produits de diffusion.

Progressivement, les enquêtes se sont vues complétées par des données administratives, par exemple sur les salaires, les revenus ou les résultats des entreprises. Les prendre en compte demandait un temps important de préparation et de maturation, à la fois politique et juridique (pour accéder aux données), sémantique (pour passer de concepts administratifs à des concepts statistiques) et technique (pour transmettre et traiter ces données, bien plus volumineuses que des données d'enquêtes)[24]. Parallèlement, les statisticiens ont commencé à s'approprier un outil de traitement statistique des données, qui leur a permis de gagner en autonomie et en flexibilité sur l'analyse des données produites, voire sur la construction de certaines étapes de traitement (par exemple les redressements ou la création de variables supplémentaires par combinaison de variables collectées). L'outil propriétaire en question, bien que riche, était assez peu évolutif, ce qui présentait l'avantage d'un univers bien balisé, relativement stable, mais l'inconvénient d'une dépendance marquée et d'une capacité d'innovation limitée.

> ***Loin d'être en rupture avec les valeurs du statisticien public, les apports de la* data science *peuvent les renforcer, en réconciliant innovation technique et exigence scientifique et déontologique.***

Dans le contexte actuel, les sources de données foisonnent, et la technique, telle que décrite précédemment, permet de les tester assez facilement, et de les intégrer dans les chaînes de production. On gardera en tête que, si la technique permet beaucoup de choses, cela n'exonère pas le statisticien de réflexions et d'échanges sur la légitimité du traitement, la capacité juridique à accéder aux données et la protection de la vie privée.

Le statisticien public se trouve face à une opportunité majeure : enrichir ses outils et méthodes, diversifier ses sources et thématiques d'analyses, tout en restant fidèle à ses principes fondamentaux que sont l'indépendance, la rigueur méthodologique,

24 Plus largement, Lamarche et Rivière (2025) comparent le temps de préparation des processus de production de statistiques publiques selon le type de données sur lesquelles ils sont fondés : données d'enquêtes, données administratives ou données détenues par des opérateurs privés.

le respect de la confidentialité[25] et la transparence. Loin d'être en rupture avec ces valeurs, les apports de la *data science* peuvent les renforcer, en réconciliant innovation technique et exigence scientifique et déontologique.

Intégrer la *data science* : une question d'infrastructures et de compétences

Le secteur privé a été précurseur dans l'intégration de profils aux compétences multiples au sein des chaînes de valeur de la donnée. Mais le secteur public – et la statistique publique en particulier – a également pris ce virage depuis plus d'une dizaine d'années[26]. L'Insee est confronté depuis longtemps aux enjeux du traitement de données produites pour une finalité de gestion plutôt que statistique. Néanmoins, l'exploitation accrue de celles-ci à différents stades du processus de production statistique et les premiers travaux visant à valoriser des données privées (à partir du début des années 2010) ont renforcé les besoins d'infrastructures et de compétences adaptées.

Dans son plan stratégique « Insee 2025 » adopté en 2015, l'Insee retenait parmi ses quatre grandes orientations pour la période 2016-2025 : « Innover et être en première ligne sur les sources de données ». Cela impliquait d'investir dans trois directions indispensables et complémentaires :

- des infrastructures informatiques permettant de traiter ces nouvelles données de manière souple et sécurisée ;
- le développement de compétences pour le traitement de ces données ;
- la mise en place d'un environnement d'innovation de type « *lab* »[27], permettant d'explorer de nouvelles sources de données et de conduire des expérimentations en matière de *data science*.

D'un point de vue organisationnel, cela s'est traduit par la mise en place de **deux équipes, l'une positionnée à la direction du système d'information, l'autre à la direction de la méthodologie statistique**. Ces deux équipes interagissent étroitement entre elles et, du fait de leur positionnement transversal, avec les utilisateurs et les porteurs d'innovations. Ce mode de fonctionnement facilite à la fois le travail collectif et la pratique du mentorat propre à développer les compétences et diffuser les bonnes pratiques.

Porté par la **première équipe**, l'Insee a lancé le SSPCloud[28] (Comte et al. 2022), une infrastructure de *data science* pensée pour lever les freins à l'innovation dans le secteur public (manque d'autonomie et d'écosystème adapté à l'expérimentation, outils inadaptés, cloisonnements entre directions métiers et direction du système d'information). Conçu comme un bac à sable contrôlé, le SSPCloud permet aux *data scientists* d'expérimenter librement, pour pérenniser ensuite les expérimentations grâce à la conteneurisation. Cette

25 Voir à ce sujet l'article du blog de l'Insee consacré à la manière dont l'institut protège les données qu'il collecte (Lefebvre, 2025).
26 La direction interministérielle du numérique (DINUM) et l'Insee ont consacré un rapport à l'évaluation des besoins de l'État en compétences et expertises en matière de donnée (Bourlange et al., 2021). Y sont évoqués les besoins de recrutement de *data scientists* dans divers domaines d'application pour s'adapter à l'offre de données croissante. Dans une perspective plus centrée sur la statistique publique, voir également Beaurepaire et al. (2022).
27 Abréviation anglaise de *laboratory* (laboratoire).
28 Le SSPCloud est disponible au lien suivant : *https://datalab.sspcloud.fr/*. Le logiciel sous-jacent, Onyxia, évoqué ultérieurement, est rendu public au lien suivant : *https://github.com/InseeFrLab/onyxia*.

infrastructure est devenue un outil central de formation aux écoles ENSAI et ENSAE[29] et, dans le cadre de la formation continue, pour les statisticiens publics. Ses outils couvrent l'ensemble des compétences en science et ingénierie des données.

En complément des logiciels d'analyse de données (R et Python), le SSPCloud permet également de disposer d'une large palette d'outils open source couvrant des besoins variés. Ceux-ci sont mis à disposition sous la forme de services préconfigurés et interfaçables (i.e. pouvant être connectés à d'autres systèmes). Ils favorisent la modularité des chaînes de valeur de données en permettant de décomposer celles-ci en maillons, lesquels peuvent évoluer selon les besoins ou l'apparition de nouvelles briques techniques.

Pour favoriser la création de plateformes identiques dans d'autres environnements, l'Insee a développé pour l'occasion le logiciel open source Onyxia (Avouac et al., 2025). Celui-ci permet à d'autres organisations de déployer et de gérer une plateforme équivalente au SSPCloud sur leur propre infrastructure. Onyxia a bénéficié du soutien d'acteurs tels que la direction interministérielle du numérique (Dinum), et d'une communauté active de développement issue, notamment, d'autres instituts nationaux de statistique. Parmi les organisations utilisant le logiciel Onyxia, on trouve des institutions telles que l'institut d'océanographie Mercator Ocean, l'agence publique de coopération technique Expertise France, des opérateurs de l'État, des établissements d'enseignement supérieur et de recherche et des instituts de statistique européens. Leurs missions sont assez diverses, mais ces organisations partagent le besoin commun de mettre à disposition, auprès des utilisateurs de leurs données, des outils flexibles et construits à partir des technologies open source les plus récentes.

Pour renforcer l'autonomie de ses statisticiens dans la production statistique sur données confidentielles, l'Insee a déployé début 2024 une nouvelle infrastructure interne, appelée LS3. Elle offre, dans un cadre conteneurisé, des environnements R/Python et des outils spécialisés (stockage S3[30], MLFlow[31], etc.), tout en assurant la sécurité des données traitées. L'infrastructure, encore récente, répond déjà aux besoins d'environ 10 % des statisticiens de l'institut.

> ***Conjointement au développement d'infrastructures informatiques, l'Insee a mis en place un* lab *de* data science *dédié à l'expérimentation.***

Conjointement au développement d'infrastructures informatiques, l'Insee a mis en place un *lab* de *data science* (nommé SSP Lab), dédié à l'expérimentation et à l'exploration de nouvelles sources ou de nouveaux usages de la donnée. C'est la **deuxième équipe** évoquée précédemment, dédiée, elle, à l'innovation sur le plan de la méthodologie statistique. Son rôle est d'accompagner les équipes de production dans leurs démarches d'innovation, de diffuser les pratiques de *data science* dans la communauté des statisticiens publics et de mutualiser les retours d'expérience.

29 ENSAI : école nationale de la statistique et de l'analyse de l'information ; ENSAE : école nationale de la statistique et de l'administration économique.
30 S3 pour *Simple Storage Service* en anglais (soit « moyen simple de stockage » en français). Il s'agit d'un service de stockage de données sur Internet.
31 Plateforme open source au service de la traçabilité des projets utilisant du machine learning.

Cette équipe a notamment piloté les premières expérimentations de modèles de machine learning à l'Insee et a accompagné ensuite leur mise en production. Ce fut le cas par exemple pour le modèle destiné à déterminer le code de l'activité principale exercée (APE) des nouvelles entreprises à partir des libellés d'activité déclarés lors de leur création. Le déploiement rapide du premier prototype, ainsi que l'amélioration continue du modèle, ont été rendus possibles grâce à l'adoption des principes du MLOps (***encadré 1***). Depuis 2022, cette équipe anime également le réseau SSPHub[32], ouvert à tous les *data scientists* du secteur public, à travers des ateliers, formations, conférences et échanges de pratiques innovantes.

Contrairement aux apparences, la répartition des missions entre deux équipes ne reproduit pas un cloisonnement entre statisticiens et informaticiens, mais favorise au contraire leur collaboration étroite, indispensable à la diffusion de l'innovation dans un environnement où métiers informatiques et statistiques étaient historiquement assez disjoints. Leur montée en compétences s'est faite de manière coordonnée, en s'enrichissant mutuellement. Par ailleurs, l'appartenance de ces deux équipes à deux directions distinctes permet également de se rapprocher d'utilisateurs de profils différents.

D'autres organisations de la statistique publique ont opté pour une approche différente, en concentrant leurs compétences au sein d'une seule équipe dédiée à la *data science*. C'est le cas de la direction de la recherche, des études, de l'évaluation et des statistiques (Drees)[33], avec son Lab Santé fortement mobilisé pendant la crise sanitaire, ou de la sous-direction des systèmes d'information et études statistiques (Sies)[34], dont l'équipe de *data science* développe le « baromètre de la science ouverte » en mobilisant des méthodes d'apprentissage automatique sur des publications de chercheurs. Cette approche de concentration des compétences dans une seule équipe a également été adoptée par les services *data* de l'inspection générale des finances (Bolliet et al. 2025), de l'inspection générale des affaires sociales (Berthe, 2025) ou de la Cour des comptes[35].

Une hybridation des compétences nécessaire à l'aboutissement de projets statistiques de grande ampleur

La crise sanitaire a représenté pour plusieurs institutions un accélérateur d'innovation. La production et la diffusion quotidiennes par la Drees de statistiques sur l'épidémie de Covid (hospitalisation, tests positifs…) a représenté un véritable défi. Pour le relever, il est apparu indispensable d'adopter une forme d'agilité et d'arriver à automatiser le processus malgré un environnement instable. La Drees n'a pas désinvesti ce sujet après la crise sanitaire. Pour valoriser le système national des données de santé (SNDS)[36], qui permet de produire des indicateurs ou des bases de données telles que celle des restes à charge, plusieurs profils sont mobilisés. Il y a d'abord des ingénieurs capables de traiter les données fournies par la caisse nationale de l'assurance maladie (Cnam). Il y a ensuite des experts issus de différents champs disciplinaires (*data scientists*, statisticiens ou médecins), pour donner un sens socioéconomique aux données traitées.

32 Les ressources communautaires développées dans le cadre de ce réseau sont recensées au lien suivant : *https://ssphub.netlify.app/*.
33 Service statistique ministériel dans le domaine de la santé et des solidarités.
34 Service statistique ministériel dans le domaine de l'enseignement supérieur et de la recherche.
35 Voir l'article de Belrhomari et al. dans ce même numéro.
36 Voir l'encadré consacré au SNDS dans Berthe (2025).

Intégrer certaines sources de données transactionnelles dans les processus de production représente parfois un véritable défi informatique et méthodologique. C'est le cas de l'intégration de données de caisse[37] dans l'indice des prix à la consommation (Leclair, 2019). Le temps important d'expérimentation avant le passage en production s'explique d'abord par le besoin de créer une infrastructure informatique capable de traiter la charge (plus de 1,7 milliard d'enregistrements quotidiens par mois). Il s'explique aussi par la nécessité de s'assurer que l'utilisation de cette source n'introduise pas de

▶ Encadré 1. La production à partir de machine learning, un travail d'amélioration continue

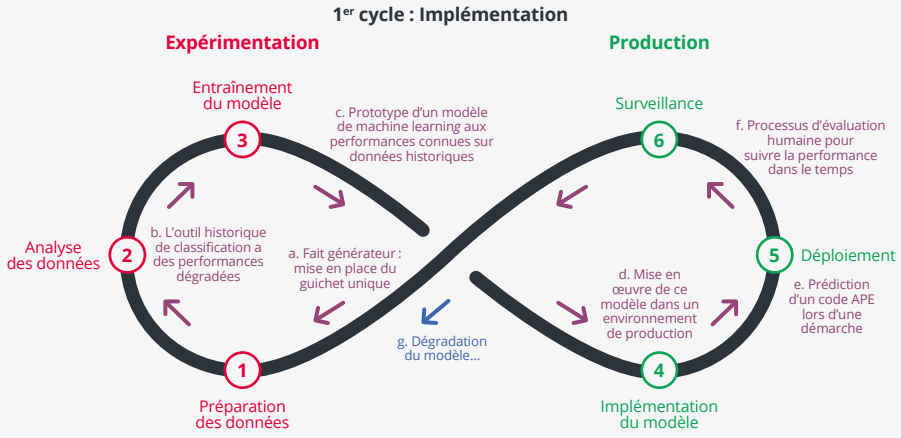
Dans le domaine du développement logiciel, l'approche DevOps vise à favoriser la collaboration entre les équipes de développement (Dev) et celles en charge des opérations (Ops) c'est-à-dire de l'exploitation opérationnelle. Cela passe principalement par l'intégration du cycle de vie d'un projet dans un continuum de phases de travail permettant la simultanéité, et non l'enchaînement séquentiel, de périodes d'expérimentation et de production. L'objectif est de réduire le « mur de la production », c'est-à-dire le risque qu'un projet échoue lors du passage en production par manque d'anticipation, en intégrant dès la conception les enjeux liés au déploiement, à la maintenance et à la remontée de bugs d'un logiciel.

Comme les projets impliquant un modèle de machine learning s'inscrivent dans un travail collaboratif entre différents profils *data* et requièrent des cycles itératifs, l'approche correspondante, dite MLOps (ML pour machine learning), a adopté les principes DevOps dans ce domaine. La différence de champ d'application (développement logiciel versus machine learning) nécessite cependant des adaptations pour certaines phases. C'est notamment le cas pour l'évaluation, qui mobilise des critères plus larges que la seule satisfaction d'un utilisateur de service.

L'intérêt de l'approche MLOps peut être illustré à partir du projet de codification automatique de l'activité principale exercée (APE) à partir de déclarations textuelles d'activités (en langage naturel) faites dans le cadre des formalités d'entreprise (Avouac et al., 2025) (***figure encadré***).

Dans une première phase d'**expérimentation**, un corpus d'entraînement issu des déclarations passées a permis de tester plusieurs modèles. Cette étape a mobilisé des analystes métiers et gestionnaires d'applications pour définir les besoins, ainsi que des statisticiens et *data scientists* pour concevoir et évaluer le modèle le plus adapté.

Cycles de vie d'un modèle de machine learning : l'exemple de l'activité principale exercée (APE)



37 Informations recueillies par les enseignes du commerce de détail, au moment où les clients passe à la caisse, sur les produits achetés et les prix payés.

rupture de série, dans le contexte d'un indice produit à partir de plusieurs sources (données de caisse, *web scraping*[38], relevés d'enquête, données administratives).

L'expérience acquise sur ces projets a permis d'adopter pour les chantiers plus récents une démarche d'ingénierie moderne, tenant compte au mieux de l'état de l'art en la matière. C'est le cas par exemple pour le répertoire statistique des individus et des logements (Résil) de l'Insee (Lefebvre, 2024). Celui-ci mobilise pour sa mise à jour des

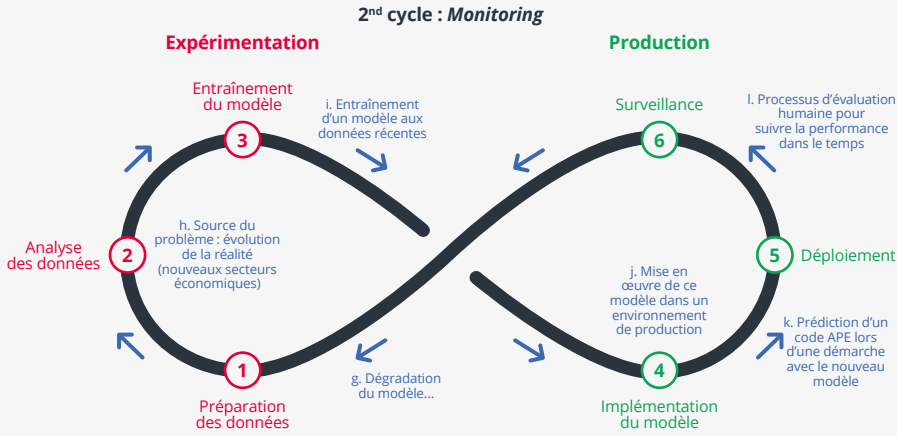
La **mise en production** a consisté ensuite à entraîner le modèle retenu, puis à l'utiliser pour prédire le code APE des nouvelles déclarations. Elle s'est déroulée de manière fluide grâce à l'adoption d'une démarche MLOps et à l'utilisation d'une infrastructure conteneurisée. Ce dispositif a permis d'éviter une réécriture du code par les *data engineers* : le métier conserve la maîtrise du traitement des données, tandis que les ingénieurs garantissent l'intégration du modèle dans un environnement de production stable et sécurisé.

L'un des enjeux majeurs du machine learning en production réside dans la surveillance continue (*monitoring*) du modèle. Un modèle entraîné sur des données historiques peut perdre en pertinence face à des phénomènes nouveaux. Par exemple, l'émergence de nouvelles activités liées à l'économie des plateformes ou à la livraison à domicile change les termes employés dans les déclarations. La surveillance du modèle via la collecte de *logs* et de données annotées* constitue donc un élément essentiel de la démarche MLOps. Elle assure à la fois la qualité des prédictions en production et l'alimentation des futures phases d'expérimentation, permettant ainsi un cycle d'amélioration continue.

L'adoption de plateformes conteneurisées comme Onyxia contribue par ailleurs à fluidifier ce processus : elles offrent un environnement cohérent entre l'expérimentation (logiciel Python associé à des ressources de calcul adaptées, entrepôt de modèles, stockage de données) et la production, renforçant la reproductibilité et la portabilité (capacité à fonctionner sur un environnement différent) des traitements.

Reste enfin à définir les choix stratégiques : fréquence de réentraînement des modèles, modalités de collecte de nouvelles données annotées, ou seuils déclenchant une nouvelle phase d'expérimentation. Ces décisions, à la fois organisationnelles et techniques, conditionnent la capacité à accélérer les itérations et à améliorer durablement la qualité des productions statistiques fondées sur le machine learning.

* *L'annotation est le procédé par lequel les données sont décrites manuellement afin d'être caractérisées (voir le glossaire de la Cnil).*



38 Le *web scraping*, ou moissonnage en français, est une technique d'extraction automatisée de données de sites web.

données issues de diverses sources administratives, souvent volumineuses et parfois incomplètes, à harmoniser et à traiter dans un temps contraint. Ainsi, la fabrication de l'univers de référence de Résil implique le traitement de bases de données dont la taille cumulée atteint 12 milliards de lignes, mais la génération d'un univers de référence ne prend que 30 minutes environ. Ceci permet de simuler facilement l'impact de la modification de règles de calcul ou l'ajout de telle ou telle donnée. Les *data scientists* peuvent ainsi concevoir des traitements reproductibles permettant de mettre en relation plusieurs sources dont la mise en production est fluidifiée grâce aux infrastructures conteneurisées de l'Insee.

Un apport dans l'ensemble du cycle de valorisation de la donnée

Au-delà du défi technique, qui peut être relevé grâce à des infrastructures adaptées et à l'acquisition de compétences en ingénierie logicielle, l'adoption pérenne de la science et de l'ingénierie des données pose surtout la question de la répartition des rôles tout au long du cycle de vie d'une production statistique. Les premières étapes d'un projet de valorisation (la phase expérimentale) mobilisent déjà des compétences proches de celles requises pour un développement applicatif, mais la phase de production n'en exige pas moins : elle suppose un suivi continu, garant de la robustesse et de la pertinence des résultats dans le temps.

Ce besoin de suivi, qui mobilise à la fois des compétences techniques et une expertise métier, concerne également les chaînes de production de données. Alors que l'informatique classique repose sur les tests unitaires[39], le monde de la donnée exige des validations spécifiques : contrôles de cohérence technique (présence de variables selon la situation), vérifications statistiques (absence de valeurs aberrantes) ou encore comparaisons d'ordres de grandeur avec d'autres sources. Une partie de ces tests, ceux qui s'appuient sur des vérifications statistiques ou des validations métiers, sont généralement conçus par les *data scientists* et *data analysts* et intégrés dans des systèmes à plus grande échelle par les *data engineers*. Disposer des différents profils capables de les mettre en œuvre est ainsi déterminant pour les organisations désirant sécuriser la production et la diffusion de résultats fondés sur les données. Faire en sorte que ces profils partagent une culture et un langage communs est une condition de succès de la spécialisation des rôles. À cet égard, le nouveau nom du corps des administrateurs de l'Insee souligne l'élargissement de la culture scientifique des statisticiens publics aux enjeux propres au traitement de la donnée (***encadré 2***).

▶ Des progrès indéniables pour la statistique publique dans le respect de ses valeurs essentielles

Les transformations structurelles de l'univers de la donnée et de l'intelligence artificielle permettent de construire des analyses à partir de sources plus nombreuses et plus diverses, d'être en mesure de réagir plus vite aux situations exceptionnelles, d'améliorer la précision de certaines productions. Elles peuvent aussi conduire à quelques gains

39 Méthode consistant, pour vérifier le bon fonctionnement d'un programme complexe, à effectuer des tests séparés sur des parties plus élémentaires (unités).

▶ Encadré 2. Le nouveau corps des « ingénieurs de la statistique, de l'économie et de la donnée »

Le métier de statisticien s'enrichit de nouvelles facettes. Il s'agit à présent de créer des outils de traitement de la donnée et non plus uniquement de les spécifier. Et il s'agit de veiller à ce que ces outils, et les chaînes de traitement qui découlent de leur assemblage, soient robustes, performantes, suffisamment souples pour s'adapter à des transformations de données ou des nouvelles contraintes, tout en assurant la traçabilité et la reproductibilité des résultats. Ces facettes, ou ces compétences nouvelles ne sont pas sans lien avec des compétences d'ingénieur, dont on attend à la fois des capacités de conception, mais aussi de mise en place d'outils de fabrication.

À cet égard, le changement du nom du corps des administrateurs de l'Insee en ingénieurs de la statistique, de l'économie et de la donnée (Insed)* est révélateur. Cette nouvelle dénomination, sans renier les fondamentaux d'origine (statistique et économie) reconnaît que les compétences des membres de ce corps, au service de la prise de décision ou de la conception de chaînes de valeur de données, exigent une maîtrise de concepts scientifiques et opérationnels. Elle traduit également le caractère interministériel de ce corps, dont les agents essaiment très largement au-delà de l'Insee, et même de la statistique publique. Dans de nombreux champs de l'action publique voire dans le secteur privé, ils apportent leur rigueur scientifique ainsi que leur expertise en analyse économique et en méthodes quantitatives de valorisation de la donnée.

Cette évolution de dénomination est également cohérente avec le fait que l'ENSAI et l'ENSAE, les deux écoles de formation initiale respectivement des attachés de l'Insee et des Insed, soient reconnues par la Commission des titres d'ingénieur comme écoles d'ingénieurs à part entière. Des enseignements consacrés spécifiquement aux enjeux propres à la mise en production et au MLOps sont présents dans le parcours de formation de ces écoles**.

* *Ce changement est créé par le chapitre V du décret n° 2025-822 du 12 août 2025. Voir les références juridiques en fin d'article.*

** *Le contenu de ces enseignements est à l'image des compétences attendues des statisticiens publics : à l'interface de l'informatique et de la statistique. Voir par exemple, au lien suivant, le cours de mise en production de projets de* data science *construit par Romain Avouac et Lino Galiana pour le cursus d'ingénieurs de la donnée de l'ENSAE : https://ensae-reproductibilite.github.io/website/.*

sur les délais de production[40]. La statistique publique, même si elle a des exigences spécifiques, est confrontée à des défis proches de ceux du secteur privé. Elle a ainsi connu un mouvement comparable à ce dernier, qu'il s'agisse de l'évolution de ses outils de travail (logiciels et infrastructures) ou des modes d'organisation du travail.

Comme le reste du marché du travail, la statistique publique a commencé à recruter des *data scientists* au début des années 2010 pour en faire un profil très polyvalent, intervenant de l'exploration à la mise en production. Néanmoins, ce profil touche-à-tout montre ses limites dès lors qu'interviennent des enjeux techniques complexes liés à la production. Industrialiser des traitements, fiabiliser l'intégration de données dans un entrepôt ou s'assurer du bon fonctionnement d'une application de production nécessitent des compétences différentes de celles des *data scientists* en charge, par exemple, de la conception d'une méthode d'imputation de valeurs manquantes.

Dans le secteur privé, une réponse à cette complexification a consisté à identifier des rôles distincts (*data engineer*, *data scientist*, *data analyst*), une évolution facilitée par l'adoption progressive de technologies *cloud*. La statistique publique, qui a besoin de concilier innovation méthodologique et exigences de robustesse, traçabilité et pérennité des productions, a tout à gagner à s'inscrire dans ce mouvement. Elle s'est longtemps appuyée, comme de nombreuses organisations, sur une séparation nette entre

40 Cela ne peut pas être des gains très significatifs, l'essentiel des facteurs explicatifs du temps de production étant ailleurs (Lamarche et al., 2025).

fonctions informatiques et fonctions statistiques. L'essor du *data scientist* a constitué un premier virage en faisant émerger un profil hybride, à l'interface entre ces deux mondes. Une seconde étape de rapprochement est l'ingénierie des données, qui offre un ensemble d'outils et de pratiques pour fiabiliser l'orchestration et la supervision de chaînes de production, depuis l'acquisition des données brutes à la mise à disposition de données transformées pour les utilisateurs. Pour tirer pleinement profit de cette nouvelle science, une partie des profils d'informaticiens doivent ainsi évoluer vers des profils de *data engineers*.

Dans un contexte où les sources de la statistique publique évoluent fréquemment et où le public demande des statistiques plus fines, plus nombreuses et diffusées plus rapidement, allier flexibilité et traçabilité passe certes par l'adoption de nouveaux outils, mais aussi par l'acculturation des acteurs de la statistique publique aux pratiques de l'ingénierie de données. De la même manière que la statistique publique est nourrie par des savoirs scientifiques en évolution, elle bénéficie de compétences techniques elles aussi diversifiées et en constante adaptation, au service de nos missions et dans le respect de nos valeurs.

▶ Fondements juridiques

- Règlement (UE) n° 2024/3018 du Parlement européen et du Conseil du 27 novembre 2024 modifiant le règlement (CE) n° 223/2009 relatif aux statistiques européennes. In : *site de l'Union européenne*. [en ligne]. [Consulté le 05 février 2026]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ%3AL_202403018*.

- Règlement (UE) n° 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données). In : *site de l'Union européenne*. [en ligne]. [Consulté le 20 avril 2026]. Disponible à l'adresse : *https://eur-lex.europa.eu/FR/legal-content/summary/general-data-protection-regulation-gdpr.html*.

- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (loi Informatique et Libertés). In : *site de Légifrance*. [en ligne]. [Consulté le 20 avril 2026]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006068624*.

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

- Décret n° 2011-194 du 21 février 2011 portant création d'une mission « Etalab » chargée de la création d'un portail unique interministériel des données publiques. In : *site de Légifrance*. [en ligne]. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000023619063/*.

- Décret n° 2025-822 du 12 août 2025 portant dispositions statutaires communes et particulières aux corps interministériels d'ingénieurs de l'État ayant vocation à exercer des fonctions d'encadrement supérieur. In : *site de Légifrance*. [en ligne]. [Consulté le 15 avril 2026]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000052102609*.

▶ Bibliographie

- AVOUAC, Romain, FARIA, Thomas et COMTE, Frédéric, 2025. L'apport des technologies cloud pour industrialiser le processus d'innovation statistique. In : *Documents de travail*. [en ligne]. Juillet 2025. Insee. N° M2025-05. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/8614378*.

- BEAUREPAIRE, Camille, COQUET, François, LE SAOUT, Ronan, 2022. Qu'est-ce qu'un•e cadre de la statistique publique ? Analyse historique des formations et analyse statistique des fiches de poste. In : *Journées de méthodologie de statistique de l'Insee 2022*. [en ligne]. Mars 2022. Insee. [Consulté le 20 avril 2026]. Disponible à l'adresse : *https://hal.science/hal-05208159v1/document*.

- BERTHE, Juliette, 2025. Le défi des données pour l'inspection générale des affaires sociales. In : *Courrier des statistiques*. [en ligne]. 23 juin 2025. Insee. N° N13, pp. 29-47. [Consulté le 20 avril 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8546937?sommaire=8546949*.

- BOLLIET, Quentin, FLOYRAC, Aymeric, MAILLARD, Sophie et ROSENZWEIG, Agathe, 2025. L'inspection générale des finances et la science des données – Quelles méthodes pour quels usages ? In : *Courrier des statistiques*. [en ligne]. 23 juin 2025. Insee. N° N13, pp. 7-28. [Consulté le 20 avril 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8546934?sommaire=8546949*.

- BOTHOREL, Éric, 2020. Pour une politique publique de la donnée. In : *Rapport de la mission sur la politique publique de la donnée, des algorithmes et des codes sources*. [en ligne]. 23 décembre 2020. [Consulté le 05 février 2026]. Disponible à l'adresse : *https://www.info.gouv.fr/rapport/11979-rapport-sur-la-politique-publique-de-la-donnee-des-algorithmes-et-des-codes-sources*.

- BOURLANGE, Danielle, BRUNET, François, CHIGNARD, Simon et EIDELMAN, Alexis, 2021. Évaluation des besoins de l'État en compétences et expertises en matière de donnée. In : *Rapport de la mission auprès de la DINUM et de l'Insee*. [en ligne]. 15 juin 2021. [Consulté le 21 avril 2026]. Disponible à l'adresse : *https://www.vie-publique.fr/rapport/281669-besoins-de-etat-en-competences-expertises-en-matiere-de-donnees*.

- COMTE, Frédéric, DEGORRE, Arnaud et LESUR, Romain, 2022. Le SSPCloud : une fabrique créative pour accompagner les expérimentations des statisticiens publics. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 68-85. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035940?sommaire=6035950*.

- DAVENPORT, Thomas H. et Patil, DJ, 2012. Data Scientist: The Sexiest Job of the 21st Century. In : *Harvard Business Review*. [en ligne]. Octobre 2012. Volume 90, pp. 70-76. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century*.

- DONDON, Alexis et LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 86-103. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.

- EDWARDS, Paul N., MAYERNIK, Matthew S., BATCHELLER, Archer L., BOWKER, Geoffrey C. et BORGMAN, Christine L., 2011. Science friction: Data, metadata, and collaboration. In : *Social Studies of Science*. [en ligne]. Volume 41, n° 5, pp. 667-690. [Consulté le 21 avril 2026]. Disponible à l'adresse : *https://www.researchgate.net/publication/51874125_Science_Friction_Data_Metadata_and_Collaboration*.

- GADOUCHE, Kamel, 2019. Le centre d'accès sécurisé aux données (CASD), un service pour la *data science* et la recherche scientifique. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N° N3, pp. 76-92. [Consulté le 21 avril 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254227?sommaire=4254170*.

- IOANNIDIS, John P. A., 2005. Why Most Published Research Findings Are False. In : *PLOS Medicine*. [en ligne]. 30 août 2005. Volume 2, n° 8. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://doi.org/10.1371/journal.pmed.0020124*.

- JOUBERT, Marie-Pierre, 2025. Les données de téléphonie mobile – Une source de connaissance sur la population et ses déplacements. In : *Courrier des statistiques*. [en ligne]. 23 juin 2025. Insee. N° N13, pp. 95-118. [Consulté le 21 avril 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8546945?sommaire=8546949*.

- KAPOOR, Sayash et NARAYANAN, Arvind, 2023. Leakage and the reproducibility crisis in machine-learning-based science. In : *Patterns*. [en ligne]. 8 septembre 2023. Volume 4, n° 9. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://doi.org/10.1016/j.patter.2023.100804*.

- KITCHIN, Rob, 2021. *Data Lives : How Data Are Made and Shape Our World*. 3 février 2021. Policy Press. ISBN : 9781529215144.

- LAMARCHE, Pierre et RIVIÈRE, Pascal, 2025. À propos du temps de production des statistiques publiques. In : *Courrier des statistiques*. [en ligne]. 15 décembre 2025. N° N14, pp. 95-118. [Consulté le 21 avril 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8675585?sommaire=8667496*.

- LECLAIR, Marie, 2019. Utiliser les données de caisses pour le calcul de l'indice des prix à la consommation. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 61-75. [Consulté le 21 avril 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254225?sommaire=4254170*.

- LEFEBVRE, Olivier, 2024. Le Répertoire Statistique des Individus et des Logements (Résil) – Un nouvel univers de référence pour les statistiques démographiques et sociales. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2024. Insee. N° N11, pp. 73-94. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8203040?sommaire=8203072*.

- LEFEBVRE, Olivier, 2025. Comment l'Insee protège-t-il les données qu'il collecte ? In : *Le blog de l'Insee*. [en ligne]. 23 avril 2025. [Consulté le 27 avril 2026]. Disponible à l'adresse : *https://blog.insee.fr/comment-l-insee-protege-les-donnees-collectees/.*

- LESUR, Romain, 2025. Sources de données privées : panorama et perspectives. In : *Courrier des statistiques*. [en ligne]. 23 juin 2025. N° N13, pp. 73-94. [Consulté le 27 avril 2026]. Disponible à l'adresse : *https://www.insee.fr/fr/information/8546943?sommaire=8546949.*

- MASLEJ, Nestor, FATTORINI, Loredana, PERRAULT, Raymond et al., 2025. In : *Artificial Intelligence Index Report 2025*. [en ligne]. 8 avril 2025. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://doi.org/10.48550/arXiv.2504.07139.*

- RUSSELL, Stuart et NORVIG, Peter, 2021. *Intelligence artificielle – Une approche moderne.* 26 novembre 2021. Pearson. ISBN : 978-2-3260-0221-0.

- TIGANI, Jordan, 2023. Big Data Is Dead. In : *MotherDuck*. [en ligne]. 07 février 2023. [Consulté le 21 janvier 2026]. Disponible à l'adresse : *https://motherduck.com/blog/big-data-is-dead/.*

- VILHUBER, Lars, 2020. Reproducibility and Replicability in Economics. In : *Harvard Data Science Review*. [en ligne]. 21 décembre 2020. Volume 2-4. [Consulté le 21 avril 2026]. Disponible à l'adresse : *https://hdsr.mitpress.mit.edu/pub/fgpmpj1l/release/5.*

- VOLLE, Michel, 1980. *Le métier de statisticien*. Hachette Littérature. L'esprit critique. ISBN 2-01-004529-7.