# Autonomy is all you need

Romain Avouac, Frédéric Comte and Thomas Faria

**Abstract** Abstract here

## 1 Introduction

In recent years, the European Statistical System (ESS) has committed to leverage non-traditional data sources in order to improve the process of statistical production, an evolution that is encapsulated by the concept of Trusted Smart Statistics [**?**]. This evolution is accompanied by innovations in the statistical processes, so as to be able to take advantage of the great potential of these new sources (greater timeliness, increased spatio-temporal resolution, etc.), but also to cope with their complexity or imperfections. At the forefront of these innovations are machine-learning methods and their promising uses in the coding and classification fields, data editing and imputation [**?**]. The multiple challenges faced by statistical institutes because of this evolution are addressed in the Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics), which predicts that "the variety of new data sources, computational paradigms and tools will require amendments to the statistical business architecture, processes, production models, IT infrastructures, methodological and quality frameworks, and the corresponding governance structures", and consequently invites the ESS to assess the required adaptations and prioritize them [**?**].

In line with these recommendations, much work has been done in the context of successive projects at the European level in order to operationalize the use of non-traditional data sources in the production of official statistics. Within the scope of the ESSnet Big Data II project (2018-2020), National Statistical Offices (NSOs) have

Romain Avouac
Insee, 88 avenue François Verdier, Montrouge, e-mail: romain.avouac@insee.fr

Thomas Faria
Insee, 88 avenue François Verdier, Montrouge, e-mail: thomas.faria@insee.fr

been working across a wide range of themes (online job vacancies, smart energy, tracking ships, etc.) in order to put together the building blocks for using these sources in actual production processes and identify their limitations [**?**]. However, while a substantial amount of work has been devoted to developing methodological frameworks [**?**, **?**], quality guidelines [**?**] as well as devising business architectures that make third-party data acquisition more secure [**?**], not much has been said about the IT infrastructures and skills needed to properly deal with these new objects.

Big data sources, which are at the heart of Trusted Smart Statistics, have characteristics that, due to their volume, their velocity (speed of creation or renewal) or their variety (structured but also unstructured data, such as text and images), make them particularly complex to process. Besides, the "skills and competencies to automate, analyse, and optimize such complex systems are often not part of the traditional skill set of most National Statistical Offices" [**?**]. Not incidentally, an increasing number of public statisticians trained as data scientists have joined NSOs in recent years. Within its multiple meanings, the term “data scientist” reflects the increased involvement of statisticians in the IT development and orchestration of their data processing operations, beyond merely the design or validation phases [**?**]. However, the ability of these new data professionals to derive value from big data sources and/or machine learning methods is limited by several challenges.

A first challenge is related to the lack of proper IT infrastructures to tackle the new data sources that NSOs now have access to as well as the accompanying need for new statistical methods. For instance, big data sources require huge storage capacities and often rely on distributed computing frameworks to be processed, which generally cannot be provided by traditional IT infrastructures [**?**]. Similarly, the adoption of new statistical methods based on machine learning algorithms often require IT capacities (notably, graphical processing units - GPUs) to massively parallelize computations [**?**].

Another major challenge is related to the difficulty of transitioning from innovative experiments to production-ready solutions. Even when statisticians have access to development environments in which they can readily experiment, the step towards putting the application or model in production is generally very large. Such examples highlight the need to make statisticians more autonomous regarding the orchestration of their processings as well as fostering a more direct collaboration between teams, as advocated by DevOps and DataOps approaches.

A third challenge is to foster reproducibility in official statistics production. This quality criterion involves devising processing solutions that can produce reproducible statistics on the one hand, and that can be shared with peers on the other hand.

- Final challenge : encourage and faciliate collaboration - Against that background, we argue that common theme : fostering autonomy - ref innovation plateformes blabla - choix technologiques qui favorisent l’autonomie et la scalabilité - make cloud resources easily available - retex : insee + ssp - MLOps case study to illustrate - open-source project - one-stop-shop - blueprint for building other similar data science platforms

# 2 Context

## 2.1 Freins à l'innovation

- Thème général : donner de l'autonomie
- Limites du poste de travail : littérature sur scaling horizontal / vertical
- Observation commune aux différents INS :
  - Insee / SSM : homogénéité des parcours, pourtant grande diversité d'infra, de moyens DSI → difficulté à partager des environnements, des formations → idée de fournir une "sandbox", un commun technologique (2020) [NB : dans la continuité, sandbox à l'échelle européenne via le one-stop-shop (2024)]
  - Visions/incitations différentes DSI/statisticien → sécurité avant le fonctionnel
- Inspirations : DevOps, DataOps

## 2.2 Innovation technologique

Observation : convergence d'éco-sytèmes.
Axe : big data is dead → architecture découplage.

- Transition éco big data → éco découplage : co-localisation plus très justifiée
- Stockage objet
- Infra BD tradi très spécialisées (calcul distribué). Aujourd'hui avec ML etc cas d'usages bcp plus diversifiés → outils d'automatisation, MLOPS, GPUs
- Insee : déjà culture fichier SAS + volumétries limitées → sauté l'étape BDD (cf. big data is dead)

Axe : conteneurisation comme moyen d'autonomisation.

- Conteneurisation = light virtualization vs. VM
- Tendance DevOps → DataOps, MLOps
- Reproductibilité des traitements

# 3 Implementation

## 3.1 Onyxia

Axe : mise à dispo des technos cloud → favoriser l'autonomie.

- Convergence des choix d'archi. Mais suffisant pour garantir l'autonomie : non → les outils de l'éco-système s'adressent plutôt à des informaticiens (ex : difficulté de configurer Spark sur du stockage objet en mode kube)

- Eco système découplé, mais exigeant → compétences diverses.
- Enjeu : faciliter l'acces aux ressources cloud pour les statisticiens (qui doit déjà s'acculturer à la reproductibilité → convergence avec les outils des développeurs) → double décalage qui demande une assistance
- IHM Onyxia comme liant technique

Axe : principes

- production-ready : outils d'automatisation (-> autonomie)
- no vendor-lockin (enfermement de la structure → coût (licences) et des pratiques → fige les compétences)
- cloud-native : onyxia n'est pas le choix fondamental, le parti pris est sur le choix sous-jacent : conteneurisation + stockage objet

### 3.2 SSP Cloud

- Orientation plateforme : instance vivante d'Onyxia, ouverte, collaborative, sandbox (cf. ref papier SSP Cloud sur l'aspect plateforme)
- Innovation ouverte → littérature
- Open-data
- Instance de partage : formations reproductibles + utilisation dans les écoles de stats + hackathons (organisation annuelle du funathon cf. one-stop-shop)
- A catalog of services which covers the entire lifecycle of a data science project
- Acculturation aux bonnes pratiques par l'usage

## 4 A case-study : MLOps APE

- Illustration de la diversité des taches nécessaires dans un projet de ML et modularité indispensable de l'infra utilisée (reprendre infra BD trop spécifique et onyxia cool)
- Possible car équipe au pied du mur → Innovation possible mais pas voulu
- Notebook avec méthode de data science classique
- Avantages données ouvertes → utilisation ssp cloud possible
- Parler des problèmes liés à ce type de développement → MLOps pour les résoudre (model versionning, logging parameters)
- Logiciel qui permet de suivre cette approche = MLflow et c'est dispo sur ssp cloud
- Distribution des entrainements : scaling horizontal (argo workflow)
- Déploiement API avec fastAPI → conteneurisation (liberté vis à vis de l'informatique)
- Automatisation des processus avec argoCD pour déploiement API et dashboard de monitoring
- Environnement dev et production très proche → passage en prod facilité

  - Transmission d'une image
  - Transmission d'une API
- Monitoring indispensable

# 5 Discussion

## 5.1 Future

- Onyxia, un bien commun opensource largement réutilisé (Insee, SSB) → faciliter les contributions pour la postérité du projet open-source, qui dépasse l'Insee
- One-stop-shop : SSP Cloud comme plateforme de référence pour les projets de ML → croissance de l'offre de formation (+ traduction)
- Accompagner les réinstanciations (datafid, POCs dans le secteur privé)
- Multiplication des projets qui passent en prod (applications de dataviz, modèles de ML avec MLOps, webscraping : Jocas/WINs)

## 5.2 Discussion

- Cout d'entrée important pour l'organisation : stockage objet, cluster kube/conteneurisation
  - Choix fondamental d'archi → limite à la diffusion d'onyxia
  - Assumer le choix : compétences, organisation ...
  - Mais globalement : tendance favorable car beaucoup d'orga et INS font ce choix
- Cout d'entrée important pour le statisticien :
  - Non-persistence de l'environnement → git + stockage objet
  - Travail dans un conteneur → perte de repères sur l'environnement
  - Mais formation : bonnes pratiques + écoles de formation Insee + accompagnements
- SSP Cloud :
  - Instance ouverte → absence de données sensibles → grosse limitation des cas d'usage réalisables + frustrations → en résumé, difficile de maximiser à la fois innovation et sécurité (pb sur-contraint)
  - → résolution via le choix de l'innovation max car sujet des échanges inter-administration de données complexe + le SSP Cloud a pavé la voie à des instances internes, plus fermées → stratégie assumée "platform-as-a-package" : projet open-source packagé → facilité ++ de réinstanciation
  - Pas une plateforme de diffusion de données → pas de stratégie globale de gouvernance → le sujet de la méta-donnée n'est pas abordé.

- Gouvernance :
  - Quelle organisation ? Equipe DS centralisée qui vient en appui ou data scientists dans les orgas métiers ? Collaboration avec les équipes infos ? (cf. graphique orga/compétences de Romain)

## Appendix