FONDS NATIONAL SUISSE
DE LA RECHERCHE SCIENTIFIQUE

Division Carrières
Bourses de mobilité
fellowships@snf.ch

Rapport scientifique et général

Ce rapport concerne le subside n°:	**P2LAP1-184113**
Nom et prénom :	**Jaton Florian**
Il est établi pour la période du	**01.01.2019 au 31.08.2020**
Titre du projet de recherche :	Designing "ground truths" in an asset management firm: An inquiry into algorithmic finance
Institution hôte :	CSI – Mines ParisTech
Nationalité :	Suisse
Téléphone direct :	0788525278
E-mail :	florian.jaton@gmail.com
Date de remise du rapport :	16.11.2020

1. Rapport scientifique

☐ **intermédiaire** ☒ **final**

Le rapport doit être soumis via ***my*SNF** en format PDF (un envoi par courrier postal n'est pas nécessaire). Un rapport est considéré comme final lorsque la durée du subside est écoulée. Prière d'établir un résumé et de répondre à chaque question de manière à éviter tout malentendu. Le rapport général peut être rédigé en français ou en anglais. Le rapport scientifique doit être rédigé dans la langue du plan de recherche.

Brève description (10-15 lignes) des résultats obtenus et de leur signification; ce texte peut être mis à la disposition de milieux intéressés.

The first result of this qualitative investigation of quantitative finance concerns machine learning algorithms and their relations to "ground truths" – referential databases that define the terms of computationally solvable problems. In view of my observations, it seems likely that, at least for the restricted field of asset management, ground truths are necessary for the shaping of both supervised and unsupervised learning algorithms. Indeed, without referring to a database whose content can operate as a set of correct values, computer scientists working in corporate finance cannot quantitatively measure the performances of their algorithms and therefore cannot make them exist and circulate in extended networks. The second result is that there is some uncertainty about the actual use of machine learning algorithms – and thus also the ground truths that attest to their performance and veracity – by fund and asset managers. It seems indeed that their theoretical and organizational environment tends to go against the untimely use of predictive models based on non-certified benchmarks. The third result concerns the effects of the work of designing new machine learning algorithms within asset management firms. Far from leading to audacious computational models that are used by managers, the practical

work of shaping new machine learning algorithms contributes to equipping the information systems of the firms. Asset management, as a professional activity, may then – as a result of the infrastructural work carried out by computer science specialists under the banner of machine learning – be about to be constituted as an asset (whose rent modalities have yet to be further identified).

Résumé (2-4 pages) des résultats obtenus et de leur signification:

1.1 Décrivez brièvement la problématique abordée et citez les principaux résultats acquis.

This research was a qualitative investigation of computer science professionals working for an asset management firm in Geneva and Paris. Its main goal was to further document the significance of referential databases – often called "ground truths" – for the development of statistical learning algorithms – often called "machine learning" – in quantitative finance. On the basis of the collected materials, the research makes three main observations which suggest, but by no means demonstrate, the existence of collective phenomena that would deserve further consideration.

I shall start by quickly introducing the notion of ground truth as it constitutes my main entry point in the field of algorithmic finance and machine learning methods. Roughly put, ground truths take the shape of databases gathering input-data and desired output-targets and, as such, define the terms and solutions of algorithmic problems. These databases do not appear ex nihilo: they must be manually designed and shaped during what I call problematization processes which engage habits, desires, skills, and values. Once assembled, these ground truths are divided into two subsets: a training set and an evaluation set. The training set is used to extract numerical features capable of automating the transformation of the set's input-data into its correlated output-targets. Once extracted from the training set and translated into machine-readable lists of instructions, the transformation of input-data into output-targets – which often borrows from certified mathematical claims and, in the case of machine learning, can be partially automated – is confronted with the evaluation set. This confrontation produces performance indicators generally expressed in terms of precision and recall statistical measures. The results of these performance evaluations are then used to attest to the efficiency of algorithms in papers, reports, presentations, or other promotional materials. The centrality of ground truths for the design, evaluation, publication, and, as such, the instauration of algorithms makes me suggest that, to a certain extent, *we get the algorithms of our ground truths*. This powerful proposition was the touchstone of this postdoctoral project that attempted to answer the following questions: Do computer scientists engaged in algorithmic finance need ground truths (and training and evaluation sets, that is)? And if they do, what are the consequences of these practical processes? How do they impact the design and shaping of new algorithms in asset management firms?

The first observation I made during this postdoc project concerns financial machine learning algorithms and their subordination to ground-truth databases. My investigation suggests that, at least for the restricted domain of asset management, ground truths are necessary for both supervised *and* unsupervised learning algorithms. That supervised learning algorithms depend on labeled data in order to formulate approximation functions is nothing new: as many scholars have shown, in order to detect the organization rules of data, supervised learning algorithms are based on – and therefore supervised by – dedicated datasets. But as far as unsupervised learning algorithms are concerned, saying that they too depend on curated ground-truth databases may sound odd at first: are these algorithms not explicitly presented as relying only on input-data to detect patterns and regularities? Yet rather than a mathematical necessity – unsupervised learning algorithms do not, in effect, need any external supervision to learn their function – this subordination to ground truths comes from a practical imperative: without referring to a database whose content can operate as correct values, computer scientists working in finance cannot quantitatively measure the performances of their unsupervised learning algorithms and therefore cannot make them exist and circulate within their firm. This practical necessity is linked to the fact that unsupervised learning algorithms are not intended to remain theoretical: They are designed to be ultimately

used and worked upon, which implies comparing them to benchmarked ground-truth datasets in order to show their relevance and efficiency. And if ground truths – together with their labels (and their biases) – are not necessary for the definition of the algorithms' learning functions, they remain necessary to make them exist as devices producing valuable results.

The second main observation is that there is uncertainty about the actual use of machine learning algorithms – and thus also the ground truths that attest to their performance and veracity – by fund and asset managers. If common supervised (e.g. Gradient Boosting Machine) and unsupervised (e.g. Density-Based Spatial Clustering) learning algorithms are, sometimes, used for statistical arbitrage and thus engage with and support market-related actions, it is not clear whether the highly specialized supervised and unsupervised learning algorithms shaped by in-house computer scientists contribute effectively to investment decisions. This element, sometimes a source of frustration for my computer scientist informants, may derive, at least in part, from a mistrust towards newcomers who, in the case of the data analytics teams that I followed, were mostly hired without any prior financial training (most of them were computer scientists specialized in signal processing or data science and holding a Master's degree or, for some of them, a PhD). But this presumed mistrust towards the results of computer scientists' complex modeling work is also, perhaps, to be put in relation with divergent work and thinking habits. It may indeed be, but I can in no way affirm it, that the epistemic culture shared among financial managers is currently not suitable to algorithmic prediction schema that may look close to technical analyses (and even charting) due to their dependence on past compiled entries. Portfolio managers tend indeed to assume that markets are efficient and, therefore, that they react randomly according to unpredictable (yet bounded) fluctuations: it is by acting *as if* the markets contain all the available information that they can be modelled as randomly unfolding phenomena capable of being grasped mathematically by specific branches of (axiomatized) probability theory. In short, it is by describing markets as efficient and modellable through random processes whose only foundations are measures of volatility that modern portfolio theory – and its numerous curricula and manuals – has been established as a professional activity and even gradually associated with the inner mechanisms of liberal democracy (especially with regard to annuities and funded pensions). Instead of making predictions about the evolution of stocks – as for example algorithmic trading on extremely short time scales or specialized hedge funds on risky classes can do – asset management often consists of reacting to random market fluctuations by equipping itself with theoretical, sometimes algorithmic, devices that feed on this randomness. Hence a certain sense of expertise which involves coping with, instead of delegating to, multiple views, devices, and technologies. In short, the predictive views generated by machine learning algorithms – and relying on ground truths for their training and evaluation – may appear as subsidiary resources, among many others, for navigating the contingency of financial markets and making effective investment decisions.

The project's third proposition concerns the actual effects of the work of designing new machine learning algorithms within asset management firms. Rather than leading to the shaping of audacious computational models used without question by managers to ground their investment decisions, the practical work of shaping new machine learning algorithms contributes to at least two dynamics. The first one is to support a discourse, currently positively valued, centered around the new potentials of data collection, compilation, and organization, notably through high-level programming languages such as Python, Matlab, or R. This discursive environment – which refers indeed to part of the work carried out by expert staff recruited for this purpose and which is put forward in many firms' promotional rhetoric – participates in turn in what Stefan Leins calls *stories of capitalism*: a regime of speech and writing which encourages investment and thus promotes profitable services offered by the financial industry. The second effect of the practical work of shaping new machine learning algorithms to equip (loosely) some asset management decisions is more silent, and therefore all the more interesting. It appears indeed that an important part of the work subtending the shaping of new machine learning algorithms in asset management firms consists in equipping and reinforcing the information systems of these firms. Again, this is a practical necessity: In order to train machine learning algorithms with the hope that they will be used for investment decisions, it is necessary to set up dedicated infrastructures enabling the collection, storage, organization, and processing of large volumes of data (notably via parallel computing

procedures). The implementation of these data and processing infrastructures allows, in turn, the aggregation of ground truths that refer to market data (e.g. Bloomberg data) or some of its assumed proxies (e.g. Twitter data). However, and this is the aspect that seemed the most promising in the eyes of my informants, this ground-truth infrastructure also allows the aggregation of signals referring to asset management activity itself. In short, and quite surprisingly, firms investing in "Big Data", "machine learning" or "artificial intelligence" – and thus setting up dedicated teams of computer scientists to design in-house models – generate a data collection capacity that also concerns their own activities. Asset management, as a professional activity, is then perhaps – as a result of the infrastructural work carried out by computer science specialists under the banner of machine learning – about to be constituted as an asset (whose rent modalities have yet to be identified).

Taken together, these three observations – the supervision requirements of unsupervised learning algorithms; the uncertainty of the actual use of machine learning models in investment decisions; the progressive and indirect assetization of asset management activity – indicate that quantitative finance is indeed being traversed by new phenomena related to the advent of statistical learning methods and the referential databases they require. And these dynamics, which concern us since they also concern corporate finance, would benefit from further exploration.

1.2 Commentez le cas échéant les modifications significatives apportées au plan de recherche approuvé.

There were no significant changes to the original research plan.

1.3 Publications résultant du subside de la bourse.

Jaton F (2020a, submitted) Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society.*

Jaton F (2020b, submitted) Éléments pour une microsociologie des pratiques de codage : Alignements, chaînes de référence et indexations. *RESET*, *Recherches en sciences sociales sur internet*.

Jaton F (2021, in press) *The Constitution of Algorithms: Ground-Truthing, Programming, Formulating*. Cambridge, MA: MIT Press.

Jaton F and Vinck D (2020, submitted) Politicizing algorithms by other means: Toward inquiries for affective dissensions. *Perspectives on Science*

Veuillez utiliser des feuilles séparées et indiquer ci-après le nombre total de pages : 3

2. Rapport général

Toutes les informations que vous fournirez ci-après seront traitées de façon confidentielle. Elles sont réservées à un usage interne, notamment pour pouvoir offrir les meilleures prestations possibles aux bénéficiaires de bourses FNS.

2.1 Questions administratives

Avez-vous connu d'éventuels problèmes administratifs (assurances, visa, ambassade,…) ? Si oui, lesquels ?

Non

2.2 Impôts

La bourse a-t-elle été imposée ? ☒ oui ☐ non
Si oui :

- en Suisse : veuillez s.v.pl. indiquer dans quel canton Vaud
- à l'étranger : veuillez s.v.pl. indiquer dans quel pays

Commentaire :

2.3 Assurance vieillesse (p. ex. AVS)

Avez-vous retiré vos papiers lors de votre départ de la Suisse? ☐ oui ☒ non
Avez-vous pu cotiser à une caisse d'assurance vieillesse durant la période de la bourse?
☐ non ☒ oui, en Suisse ☐ oui, à l'étranger

Commentaire :

2.4 Montant de la bourse

Veuillez indiquer : 81'230 chf

- la ville et le pays de séjour : Paris
- ☒ célibataire ☐ marié/ménage commun
- enfant(s) : 0 (nombre)

2.4.1 Etablissez une liste sommaire des coûts que vous avez dû couvrir avec le montant de la bourse.

Loyer, assurances, trajets, communications, subsistance, impôts. À noter que ces coûts ont été supérieurs au montant de la bourse.

2.4.2 Si vous avez des enfants : avez-vous fait garder vos enfants par des tiers ? ☐ oui ☐ non

Si oui, à combien se sont élevées vos dépenses (coûts mensuels, combien de jours par semaine, quel type de prise en charge avez-vous choisi) ?

Si non, comment vous êtes-vous organisé-e ?

2.4.3 Remarques sur le montant de la bourse

Comment jugez-vous en résumé ce soutien financier, notamment par rapport à des chercheuses et chercheurs de même profil au bénéfice de bourses octroyées par d'autres organisations ? Dans la mesure du possible, merci de nous indiquer le salaire net (après déduction des impôts) obtenus par des chercheuses et chercheurs occupant une position similaire dans votre institut hôte.

Indiquez également si vous avez étiez plutôt insatisfait ☐, satisfait ☒ ou très satisfait ☐ du montant de la bourse.

2.4.4 Fonds de tiers

Avez-vous obtenu des fonds supplémentaires provenant d'autres organismes de financement, salaires pour charge de cours ou autres subsides pendant la période de la bourse ?

☐ oui ☒ non

Si oui, pourriez-vous nous indiquer le montant exact par activité (net) ?

CHF_____activité_____

2.5 Commentaire sur l'institut hôte

Comment jugez-vous le soutien scientifique de l'institut hôte ? Est-ce que le matériel de consommation et l'infrastructure étaient suffisants ? Est-ce que l'institut hôte vous a soutenu financièrement pour la participation à des conférences, congrès, workshops, etc. ? Quels autres coûts ont été pris en charge par l'institution d'accueil ? Recommanderiez-vous l'institut hôte ? Dans le cas contraire, exposez brièvement vos raisons.

Indiquez également si dans l'ensemble, vous étiez plutôt insatisfait ☐, satisfait ☐ ou très satisfait ☒ de votre institut d'accueil.

2.6 Activité après la bourse et perspectives professionnelles (à répondre seulement s'il s'agit du rapport final)

Allez-vous poursuivre votre séjour à l'étranger ? Quelles sont les prochaines étapes prévues selon votre plan de carrière ? Comment voyez-vous votre retour en Suisse ? Veuillez actualiser si nécessaire votre adresse dans mySNF.

Je suis actuellement engagé comme postdoc pour un projet FNS Sinergia conduit à l'Université de Lausanne.

Division Carrières
Bourses de mobilité
fellowships@snf.ch

2.7 Autres remarques

Quels avantages vous a apporté cette mobilité lors de votre séjour de recherche à l'étranger ? Comment estimez-vous l'utilité de votre séjour de recherche pour l'obtention d'un nouvel emploi et, plus généralement, pour la suite de votre carrière scientifique ou académique ?

Indiquez également si dans l'ensemble, vous étiez plutôt insatisfait ☐, satisfait ☒ ou très satisfait ☐ de votre séjour de recherche.

Mai 2017