Les *big data*[1] changent-ils la donne en finance ?[2]

Mathieu Rosenbaum[3]

Prisme N°35

Mars 2017

[1] L'usage du féminin pluriel ou du masculin singulier en français est fréquent pour cet anglicisme, ce texte choisit le pluriel masculin, conformément au pluriel latin du neutre *datum*.

[2] Retranscription par Jean-Gabriel Brin de l'intervention à l'atelier du Centre Cournot sur les mégadonnées. La vidéo est disponible sur le site du Centre : www.centre-cournot.org. Le Centre remercie Serge Chanchole pour sa relecture attentive.

[3] **Mathieu Rosenbaum** est professeur à l'École polytechnique, où il est notamment responsable du Master 2 de probabilités et finance, coorganisé avec l'Université Pierre-et-Marie-Curie, université dans laquelle il était précédemment professeur. Cofondateur et éditeur en chef du journal *Market Microstructure and Liquidity*, il est également responsable éditorial de *Quantitative Finance* et co-éditeur des revues *Electronic Journal of Statistics, Journal of Applied Probability, Mathematical Finance, Mathematics and Financial Economics, Statistical Inference for Stochastic Processes, SIAM Journal on Financial Mathematics* et *Statistics & Risk Modeling*. Il a reçu en 2014 le prix Europlace du meilleur jeune chercheur en finance et obtenu une allocation de recherche du Conseil européen de la recherche (ERC) en 2015. La recherche de Mathieu Rosenbaum se concentre sur les problèmes de la finance statistique, comme la modélisation de la microstructure des marchés ou la construction des procédures statistiques pour les données haute fréquence. Sa recherche concerne également les questions de réglementation, notamment dans le contexte du *trading* à haute fréquence.

Résumé

La disponibilité croissante et massive de données numériques pose des questions nouvelles aux disciplines mathématiques. Si l'afflux de *big data* modifie l'accès et la gestion des données, les méthodes quantitatives restent, elles, largement inchangées, même si elles doivent s'appliquer en très grandes dimensions et sur un temps très court. L'enjeu de ce *Prisme* est de donner un sens à la reformulation que les mégadonnées imposent dans les mathématiques et aux approfondissements qu'elles rendent possibles en finance.

Décrire en quelques pages l'ensemble des activités financières est inenvisageable. Il est cependant possible de mettre en perspective certains de leurs développements. L'objectif de ce texte est de dépeindre l'un d'entre eux, considéré comme le plus marquant de ces dernières années : le développement des *big data*, celui des données de masse ou « mégadonnées ». Les données massives se sont introduites dans de nombreux domaines : leurs techniques de collecte et de traitement ont fait l'objet de dépenses massives et ces dépenses concernent aussi bien le matériel que la publicité. De nombreux acteurs se sont engagés sur des marchés où l'idée d'une révolution en cours doit être entretenue. A première vue, l'expression *big data* qualifie d'abord une promesse commerciale. Le cadre général des mathématiques qui les étudient n'a en effet pas changé. En revanche, les approches et les méthodes de la finance ont été touchées par l'arrivée des données massives : comment décrire cette transformation des procédés et comment la mettre en perspective ? Ce texte propose de suivre l'histoire récente des méthodes appliquées sur les marchés financiers pour définir et qualifier les changements qui sont intervenus dans les pratiques.

Revenir brièvement sur le développement des produits dérivés permet de suivre l'évolution du traitement de l'incertitude sur les marchés financiers, aujourd'hui en partie confié aux *big data*. L'apparition des produits dérivés sous leur forme actuelle date des années 1970. L'année 1971 marque en effet la fin de la convertibilité en or du dollar des États-Unis. Le premier marché de contrats d'options est créé deux ans plus tard. Le *Chicago Board Options Exchange* doit permettre de répondre aux incertitudes que fait naître la fin d'un régime de changes fixes, celui qui avait été instauré à Bretton Woods en 1944. En 1973 également, un modèle de valorisation des options négociables est mis au point par Fischer Black, Robert Merton et Myron Scholes. La démarche de ces chercheurs, fondée sur les processus stochastiques en temps continu, permet d'évaluer des produits dérivés financiers. Ces produits assurent la gestion des nouvelles incertitudes auxquelles font face les entreprises. En France, l'idée d'un marché dérivé des actions est promue en 1978, mais les premiers marchés organisés de contrats à terme n'ouvrent qu'en 1986. Rapidement, les développements probabilistes s'imposent et l'idée de Black, Scholes et de Merton se diffuse : ils permettent de créer des produits de couverture et d'investissement, dont ils savent parfaitement maîtriser les risques grâce à une

gestion dynamique[4].

Un exemple concret permet de comprendre la logique de cette idée. Soit un acheteur qui souhaite se fournir en kérosène dans un an, à un prix fixé aujourd'hui, disons une tonne, à 1 euro le kilo. Le client et l'institution financière signent un contrat, appelé option d'achat, dans lequel le client peut s'il le souhaite, au bout d'un an, exiger la tonne de kérosène au prix fixé dans le contrat (1 euro le kilo). Si la tonne de kérosène coûte moins d'1 euro le kilo une fois l'année passée, le client n'utilise pas l'option ; si elle est plus chère, il exerce l'option. Imaginons que le prix du kilo du kérosène soit dans un an égal à s ; si $s > 1$, l'option est exercée, mais pas dans le cas contraire. Le client récupère au bout d'un an un profit de $1000*(s-1)$ si $s - 1 > 0$, 0 sinon. Évidemment, l'institution financière est rémunérée le jour de la signature du contrat : c'est le prix de l'option. Comment fixer ce prix de l'option, c ? La théorie de Black-Scholes[5] et Merton[6] permet de calculer c, de sorte qu'en investissant c sur les marchés de manière dynamique, l'institution financière est certaine de pouvoir rémunérer le client au bout d'un an, c'est-à-dire de transformer c aujourd'hui en $1000*(s-1)$, si $s-1$ est positif demain, 0 sinon. Il n'y a donc plus d'aléa dans le cadre des hypothèses de ce modèle probabiliste. Cette manière de faire disparaître le risque a été une révolution scientifique sur les marchés.

Dans tous les cas, l'unité de temps reste fondamentale. Dans le monde des produits dérivés en effet, et pour des raisons structurelles, le temps naturel du marché est la journée. Jusqu'à la fin des années 1990, il s'agit d'enregistrer une donnée quotidienne, par exemple tous les jours à 17 heures. Dans l'activité de l'opérateur, il faut modéliser un prix observé une fois par jour, à une heure donnée, puis le lendemain, etc. Le bilan est fait tous les soirs et chaque jour est un autre jour. Quand le prix est enregistré tous les jours à 17 heures pendant deux ans, cela s'apparente à

[4] El Karoui N. (2009), *Un moment de l'expérience probabiliste Théorie des processus stochastiques et pratique dans les marchés financiers,* Prisme du Centre Cournot. Ce texte met en perspective l'histoire du développement des produits dérivés.

[5] Black, Fischer & Myron Scholes (1973). « The Pricing of Options and Corporate Liabilities ». *Journal of Political Economy.* 81 (3) : 637-54.

[6] Merton, Robert (1973). « Theory of Rational Option Pricing ». *Bell Journal of Economics and Management Science.* The RAND Corporation. 4 (1)) : 141-83.

ceci :

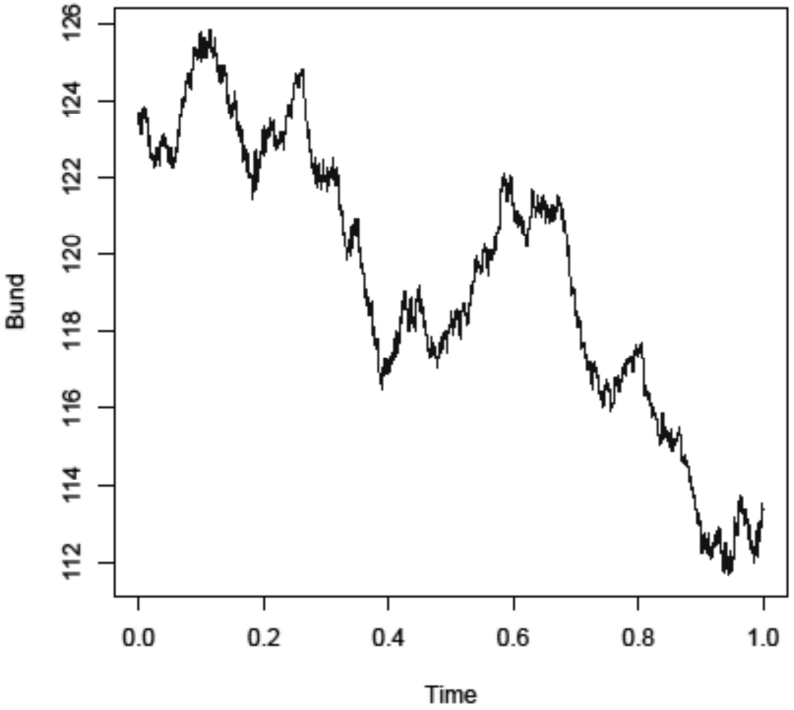


Figure 1 : Graphique type du prix d'action sur 2 ans avec relevé quotidien

Ce tracé ressemble à celui d'un mouvement aléatoire particulier : le mouvement brownien. Avant l'arrivée des données massives, l'ensemble de la finance mathématique, pilotée par le questionnement sur les produits dérivés, a conservé comme modélisation de base celle de ce mouvement. Au niveau statistique, comme pour la perception visuelle, le mouvement brownien est en effet un processus approprié. De plus, certaines de ses propriétés théoriques l'imposent comme le processus pertinent à cette échelle de temps.

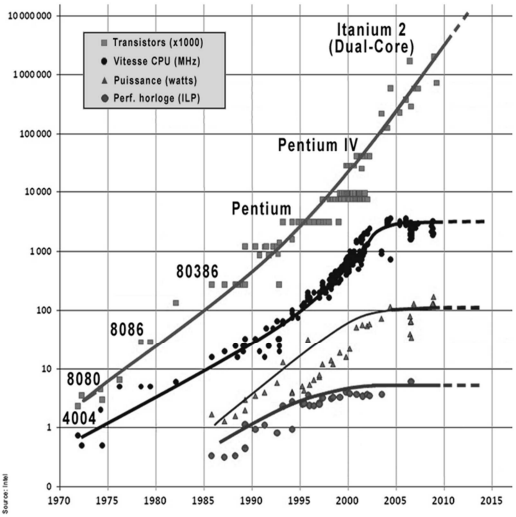
La modernisation de la finance issue des travaux de Black-Scholes et Merton va de pair avec l'arrivée de supports informatiques de plus en plus puissants. La mise sur le marché progressive de machines plus rapides et l'arrivée d'ordinateurs personnels entraînent la disparition, à partir du milieu des années 1980, du gros

ordinateur IBM qui centralise les calculs. Ce remplacement s'accompagne de la prolifération des données à disposition des opérateurs. De nouveaux logiciels sont élaborés et de nouvelles professions naissent : informaticiens et ingénieurs affluent dans les salles de marché dans les années 1990. L'arrivée de données massives est perceptible au début des années 2000. Progressivement, le lexique de la profession est lui aussi touché. « Statistiques », par exemple, disparaît et laisse sa place à « sciences des données » ou plutôt à son anglicisme « data science ». Dans les Grandes Écoles, il n'y a plus de voie statistique, il n'y a plus de cours de statistiques : il y a les filières de ***data science***. La « classification » n'existe plus non plus, le ***clustering*** s'impose. L'exploration de données, le ***data mining***, qui était vraiment poussiéreux, a laissé sa place à l'apprentissage automatique, plus connu en tant que « machine learning ». Il y a une composante de mode dans les ***big data***, et la question qu'il faut poser porte sur ce que cache cette vogue. Où se trouvent les transformations tangibles ?

Il convient pour commencer de définir les ***big data*** en finance et de poser les bonnes hypothèses pour comprendre les conséquences de leur utilisation. Avant tout, les données de masse proviennent de l'enregistrement systématique des activités prises en compte sur les marchés. Ces données sont conservées et facilement consultables. Concrètement, tous les prix, toutes les transactions, tous les carnets d'ordres sont mémorisés de manière systématique. Les processus d'enregistrements permettent de conserver toutes les versions, de les horodater ou d'en marquer tous les accès, ce qui les rend plus facilement utilisables par les opérationnels et offre une traçabilité complète. Ces enregistrements sont effectués car, potentiellement, et il faut souligner potentiellement, de l'information est contenue dans les données passées. Les données n'ont d'intérêt que si des systèmes permettent d'y accéder facilement et de les analyser de manière pertinente. La prise en compte des ***big data*** consiste en grande partie à bien ranger les données. Ainsi, l'accumulation de données se révèle par exemple d'une grande utilité dans le ***trading*** à haute fréquence.

Les transactions à haute fréquence relèvent d'une intervention quasi-permanente sur le marché, où un très grand nombre d'échanges a lieu en un temps très court. Elles s'inscrivent logiquement dans le développement ininterrompu des capacités informatiques. Ce développement est illustré par les deux graphiques

suivants :



Source : Intel

Figure 2 : Croissance des capacités de traitement de données 1970-2015

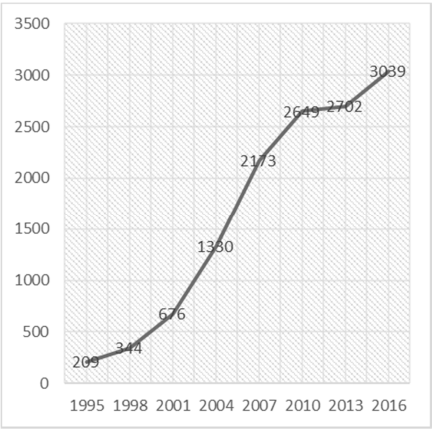


Figure 3 : Chiffre d'affaire du marché de gré à gré des produits dérivés d'après le *BIS Statistics Explorer* (avril 2016, moyennes journalières, en milliards de dollars des E.U.)

Les interventions quasi-permanentes sur le marché ne résultent pas la plupart du temps, comme on peut parfois l'entendre, de l'action mal intentionnée d'informaticiens qui auraient pour projet de dérober l'argent des investisseurs. Elles concernent aujourd'hui toutes les pratiques de la finance de marchés. Avec des ordinateurs puissants, tout le monde peut d'ailleurs avoir accès à la haute, ou plus ou moins haute fréquence, et les marchés financiers s'ajustent. L'ordre de grandeur est la milliseconde, voire la nanoseconde. A cette vitesse, toutes les transactions d'un carnet d'ordres peuvent être enregistrées, c'est sur elles que les opérateurs se concentrent. Le marché est alors, à l'instant *t*, ce lieu virtuel où des acheteurs et des vendeurs souhaitent échanger un actif. Comment cela se passe-t-il concrètement ?

Prenons, par exemple, l'action de l'entreprise X. A l'instant *t*, il y a des acquéreurs qui sont prêts à acheter des actions de X à 110,53€, d'autres souhaitent acheter à 110,52€, d'autres à 110,51€, etc. Il y a d'un côté des prix, et en face, il y a des quantités.

Tableau : Carnet d'ordres simplifié

Prix acheteurs	Quantités d'actions à l'achat	Prix vendeurs	Quantités d'actions à la vente
110,53€	95	110,54€	17
110,52€	90	110,55€	50
110,51€	5	110,56€	117
110,50€	2	110,57€	18
110,49€	10	110,58€	100

Selon ce carnet, il y a sans doute un opérateur qui se dit prêt à acheter « 95 actions X au prix de 110,53€ ». Peut-être y a-t-il deux opérateurs, l'un est prêt à en acheter 40 et l'autre à en acheter 55. Le plus souvent, le détail n'est pas explicite, mais en tout cas, il y a bien 95 actions X qui sont susceptibles d'un achat à 110,50€. C'est la même chose du côté ventes : il y a des vendeurs prêts à intervenir à 110,54€ ; à 110,55€ ; à 110,60€…

Une transaction a lieu lorsque quelqu'un dit, par exemple: « Je suis prêt à acheter à 110,53€ ». Si vous êtes prêt à vendre à 110,53€, vous avez trouvé un acheteur. L'autre dit « je suis prêt à acheter à 110,54€ » et un vendeur est trouvé. C'est cette opération que l'on observe dans les salles de marché, où des opérateurs devant leurs écrans scrutent leur carnet d'ordres. En gros, dès qu'une donnée change, une petite lumière s'allume et elle clignote. La banque enregistre ainsi l'ensemble des ordres sur les milliers d'actifs disponibles. Des téraoctets de données sont ainsi conservés. Dans ce cas sont enregistrées les données de quatre colonnes pour chaque actif : ce qui doit être enregistré est de dimension $5+5+5+5=20$ et se modifie à peu près toutes les millisecondes. Les données sont donc vraiment massives et arrivent à très haute fréquence.

Il est alors raisonnable de demander à une banque de quoi se composait le carnet d'ordres d'une certaine action trois jours auparavant, à 16h 43min 55s 95ms, car la banque est techniquement en mesure de répondre à la question. Cette grande nouveauté en matière de stockage s'accompagne d'une compréhension du marché plus fine puisque sa substance devient plus accessible. Comment améliorer et tester les modèles pour les adapter à ces nouvelles hypothèses ? S'il y a plus de données pour stimuler les chercheurs, il y a aussi beaucoup plus de données pour calibrer leurs nouveaux modèles et les mettre à l'épreuve, dans des conditions beaucoup plus difficiles.

L'exploitation de données enregistrées toutes les millisecondes, voire toutes les nanosecondes, n'est pas possible dans le cadre du modèle brownien évoqué plus haut. En effet, celui-ci n'est pas pertinent sur des horizons de temps court, or de nombreux problèmes de *trading* se posent sur des intervalles brefs. Un client peut appeler un courtier et passer un ordre : « vous avez deux heures pour me vendre 10 millions d'actions ». Vendre 10 millions d'actions ne peut pas se faire n'importe

comment et sûrement pas d'un coup. Comment maximiser le gain pour le client ? Si l'opérateur suit strictement le carnet d'ordres du tableau, alors, il vend 95 lots à 110,53€, 90 lots à 110,52€ et ainsi de suite, si bien que le prix moyen de vente par action vaut, disons par exemple 100€. Le client, qui regarde le marché, observe que le prix de l'action est à peu près de 110,53€, il ne peut accepter que le prix moyen de vente ressorte à 100€. Il faut que l'opérateur organise sa stratégie de vente sur l'intervalle de deux heures.

Les nouveaux problèmes de modélisation qui apparaissent sont donc directement issus de la compréhension des données et de leur utilisation, en grande dimension et sur un temps très court. Même si l'on considère simplement la première ligne d'un carnet d'ordres, soit 110,53€ dans le tableau, les deux heures imparties à la transaction vont voir le prix évoluer selon une trajectoire qui n'est en rien brownienne :

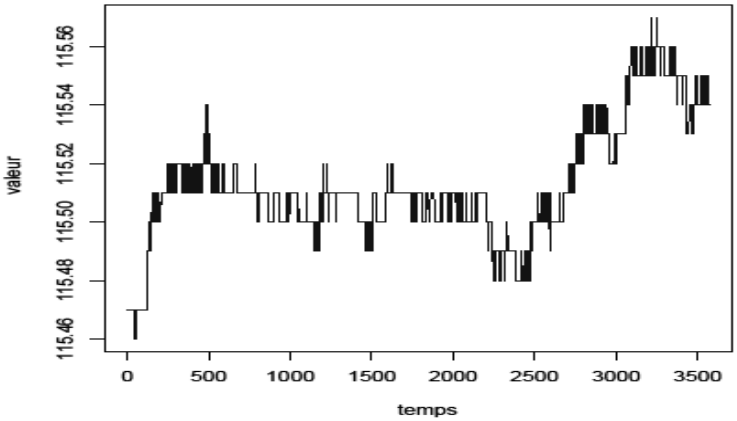


Figure 4 : Évolution typique d'un actif sur une heure

Les enjeux de l'explosion des ***big data*** en finance sont donc d'abord ceux des mathématiques, qui doivent répondre aux questions posées par les praticiens. Parmi celles-ci, la compréhension de la volatilité est primordiale : on sait ce que signifie la volatilité du marché sur la courbe de la figure 1, elle est mesurable dans un cadre brownien. Sur la seconde courbe, on est démuni, or un concept de volatilité est nécessaire dans le calcul du risque à haute fréquence. Une autre question difficile concerne la corrélation entre deux actifs. Il y a dix ans, il était possible de faire la corrélation empirique des accroissements de prix observés quotidiennement, afin d'obtenir une mesure grossière de la dépense entre actifs. Sur deux heures, cette opération est beaucoup plus délicate. Même si les évolutions de prix sont comparables, le fait que les moments de transaction concernant deux entreprises soient différents induit des problèmes de discontinuité temporelle, d'une sérieuse difficulté pour le statisticien. Sont ainsi observés deux processus qui sautent, mais pas aux mêmes dates, c'est d'ailleurs une observation triviale. La non-synchronicité constitue à elle seule un obstacle qui rend la mesure de la corrélation extrêmement difficile, or le besoin d'une telle mesure reste très fort en haute fréquence.

Les données en grande dimension en finance ont posé de nouvelles questions à la recherche et entrainé de nouveaux développements mathématiques, notamment des progrès en contrôle stochastique, avec les questions que pose la haute fréquence. L'observation des pratiques dans les banques invite en fait à redécouvrir, à travers le prisme des données massives, des techniques et des problèmes connus depuis la mise au point des produits dérivés. Ce phénomène trouve une illustration dans la construction de nouveaux programmes dans l'enseignement universitaire. Quand j'étais professeur à l'Université Paris VI, il nous a été demandé de monter une filière ***big data***. Une fois passés en revue les cours dispensés en mathématiques à l'Université, il est apparu qu'une trentaine d'enseignements disponibles à l'Université Paris VI pouvaient déjà être libellés ***big data***. Ce qui manquait, c'était une espèce de chapeau commun et surtout une compréhension de l'interdépendance entre les statistiques et d'autres disciplines numériques, et notamment informatiques[7]. Dans cet exemple, j'insiste sur la prise de conscience de ce que la composante

[7] Voir, par exemple, El Karoui, Noureddine (2015), *Les statistiques peuvent-elles se passer d'une théorie des probabilités ?*, Prisme N°30, octobre, Paris : Centre Cournot.

informatique doit être conçue et enseignée en synergie avec les autres composantes.

Il faut d'ailleurs dissiper une confusion qui naît de la mauvaise perception des utilisateurs des mégadonnées. L'un des dangers que court la finance est la croyance dans le tout *big data* et l'idée que ceux-ci constituent une solution à tous les problèmes. L'un des effets les plus marqués de cette croyance est la valorisation de nouvelles qualifications, et en premier lieu la capacité à traiter ces données numériques. Les fiches de postes montrent un véritable intérêt pour ceux qui peuvent mobiliser ces sources. Il est évidemment nécessaire que des compétences en statistiques et des compétences en informatique soient requises pour travailler en finance. Néanmoins, la tendance aujourd'hui laisse croire que les modèles ne servent à rien, que celui qui est super fort en informatique, qui sait faire de l'apprentissage automatique est compétent en finance. C'est faux. Il ne faut pas oublier que la finance est un système avec de l'économie en arrière-plan et des agents qui interagissent, il y a aussi des institutions, même si tout cela est parfois peu perceptible ! Si l'on oublie cela, à un moment, on le paie. Dans l'industrie financière, les données en grandes dimensions, seules, ne servent d'ailleurs à rien, elles sont un moyen et pas une fin. Il est intéressant de regarder par exemple le sort des structures qui n'en ont pas pris conscience : la durée de vie des nouvelles entreprises de transactions haute fréquence aujourd'hui (se fondant uniquement sur des techniques informatiques) est de six mois en moyenne.

Le défi que lance l'abondance de donnés consiste, finalement, à mettre au point de bons modèles haute fréquence. Comment y parvenir ? Un bon modèle haute fréquence doit bien reproduire les caractéristiques empiriques haute fréquence des marchés, mais ce doit être aussi un modèle utile pour résoudre un problème de *trading* important (comme celui de notre client souhaitant vendre, par exemple) ou pour améliorer notre compréhension des marchés. La combinaison de ces deux vertus n'est pas triviale. Se contenter, comme on le voit parfois, de modèles désincarnés reproduisant uniquement les phénomènes abstraits, ne sert pas à grand-chose.

Je n'identifie donc pas de discontinuité scientifique derrière les *big data*. Avec l'arrivée massive de données, je vois une discipline qui avance : les problèmes sont mieux formulés, les questions sont mieux posées, les modélisations des interactions sont plus précises, l'optimisation numérique est plus performante. Dans

l'ensemble, beaucoup de choses progressent vite. Il n'y a pas de nouveauté scientifique, mais des recherches qui s'approfondissent et qui avancent. Il y a par ailleurs une prise de conscience des acteurs en aval, industrie et services. Les salariés doivent avoir les qualifications pour aborder les grandes quantités de données qu'il faut savoir mieux enregistrer, trier et traiter. Assez rapidement, l'industrie a souhaité que soient créées des formations générales abordant l'ensemble des aspects des ***big data***, et principalement l'aspect statistique, l'aspect analytique et l'aspect informatique. L'adaptation des enseignements à des techniques déjà éprouvées constitue un changement structurel, en raison de la mise en place de synergies entre les disciplines. Il n'y a cependant pas de changement de paradigme.

Il est important de rappeler que les ***big data*** ne sont pas nés d'un nouveau théorème dont la force impose que son apprentissage est tout d'un coup nécessaire pour travailler dans la finance. C'est pour ces raisons que, pour moi, il y a vraiment une différence entre l'éclosion des produits dérivés et la mise à disposition des ***big data***. Au fond, les ***big data*** ont permis de comprendre que la notion de temps en soi n'est pas déterminante, c'est la dimension nécessaire des données qui permet de résoudre des problèmes pratiques et cette dimension dépend de l'échelle de temps. Le modèle de Black-Scholes et Merton est un modèle en temps continu pour une pratique basse fréquence en petite dimension ; or aujourd'hui, les opérateurs travaillent réellement en haute fréquence. La modélisation des pères fondateurs n'est plus adaptée et il devient nécessaire de considérer des objets en grandes dimensions.

Pour conclure, il faut garder à l'esprit que la disponibilité croissante et massive de données numériques a d'abord posé des questions nouvelles à la finance: accès et gestion des données, objets mathématiques et rapport à ses théories réorganisations disciplinaires et transformations des méthodes. Dans les banques, les ***big data*** ont déclenché une prise de conscience du besoin de mieux accéder aux données, avant de mieux les organiser et de les traiter correctement. La disponibilité des données massives invite aussi à regarder l'histoire des développements mathématiques avec plus d'acuité. Il faut suivre le développement des méthodes probabilistes pour comprendre les transformations que les ***big data*** ont imposées. Leur entrée en jeu a conduit à de nouvelles questions aux chercheurs, aux opérateurs et aux banques elles-mêmes, les invitant à revoir un certain nombre de faits qui

semblaient établis depuis les années 1990. Elle n'est en rien comparable à la révolution induite par les applications des idées de Black-Scholes et Merton pour les produits dérivés.