Sébastien PREYS

L'expérience d'Ondalys dans la formation continue aux outils opérationnels de la chimiométrie et du *machine learning*

Sébastien Preys, Chef de projet Data Science et Machine Learning[1] chez Ondalys.
Monsieur Sébastien Preys est, depuis 2006, docteur en chimiométrie[2] à l'INRA (Institut national de la recherche agronomique) de Montpellier. Sa thèse portait sur l'analyse de données multi-blocs[3] combinant différents jeux de données multivariées, provenant de différentes techniques analytiques. Depuis 15 ans, il travaille chez Ondalys pour fournir des services et des formations.

1. *Machine learning* : apprentissage automatique.
2. La chimiométrie est un outil utilisé afin d'extraire de l'information pertinente et utile à partir de données physicochimiques mesurées ou connues brutes. Il est basé sur la construction, puis l'exploitation d'un modèle de comportement à l'aide d'outils statistiques.
3. La chimiométrie est un outil utilisé afin d'extraire de l'information pertinente et utile à partir de données physicochimiques mesurées ou connues brutes. Il est basé sur la construction, puis l'exploitation d'un modèle de comportement à l'aide d'outils statistiques.

Introduction

Qu'est-ce qu'Ondalys ? C'est une entreprise créée il y a 20 ans (***Figure 1***), au départ, une jeune entreprise innovante et maintenant une petite équipe d'une dizaine de *data scientists*[4] avec de l'expérience, un leader dans la chimiométrie et le *machine learning* en France. Ondalys travaille principalement pour l'industrie de process[5] : pharmaceutique, biotechnologie, chimie, agroalimentaire, cosmétique, etc. Son activité principale est d'accompagner les industriels dans la mise en place des outils de *machine learning* et d'intelligence artificielle.

Ondalys conduit aussi une activité de formation continue, qui fait l'objet du présent chapitre. Enfin, Ondalys distribue également les logiciels d'analyse de données de certains de ses partenaires.

Voici quelques mots-clés que nous aborderons dans ce chapitre : *data mining*[6], calibration spectroscopique, modèle prédictif, combinaison de capteurs, ou monitoring[7], qui implique une supervision de procédés en continu[8] ou

4. *Data scientist* : scientifique des données, exploite les données de l'entreprise.
5. Industrie de process : industrie dans laquelle les matières premières subissent une transformation chimique et/ou physique.
6. *Data mining* : exploration de données, analyse de données depuis différentes perspectives et le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns.
7. Monitoring : supervision.
8. Procédé continu : mode de production industriel destiné à fabriquer, construire ou traiter des matériaux sans interruption.

Figure 1

Résumé des informations générales sur l'entreprise Ondalys.

en batch[9], et plans d'expériences[10]. Nous sommes une équipe d'ingénieurs qui a développé un réseau de partenaires de logiciels sur lesquels nous formons des professionnels.

Un projet de *machine learning* ou de data science[11] peut s'aborder au niveau de l'audit chez le client pour voir ce qui a été fait et donner notre avis (***Figure 2***). L'activité principale d'Ondalys réside dans le conseil et l'étude de faisabilité ; le *proof of concept*[12], pour tester les outils de *machine learning* sur des applications précises. Si le projet est satisfaisant, l'idée est d'accompagner le client à développer et valider des modèles et à réaliser l'implémentation logicielle. Comme un modèle possède un cycle de vie, il faut vérifier qu'il ne dérive pas, qu'il continue à bien prédire et cela sous-tend toute une activité de maintenance de modèle. La formation continue, qui est le thème de ce chapitre, représente à peu près 20 % de notre activité, en incluant du coaching après une formation accompagnée.

Ce chapitre revient d'abord sur la sémantique et sur certains concepts de chimiométrie et de *machine learning* qui font partie du vaste champ de l'intelligence

9. Procédé batch : mode de production industriel par lots avec interruption.
10. Plans d'expériences : consiste à sélectionner et ordonner les essais afin d'identifier, à moindres coûts, les effets des paramètres sur la réponse du produit.
11. Data science : science des données.
12. *Proof of concept* : preuve de concept.

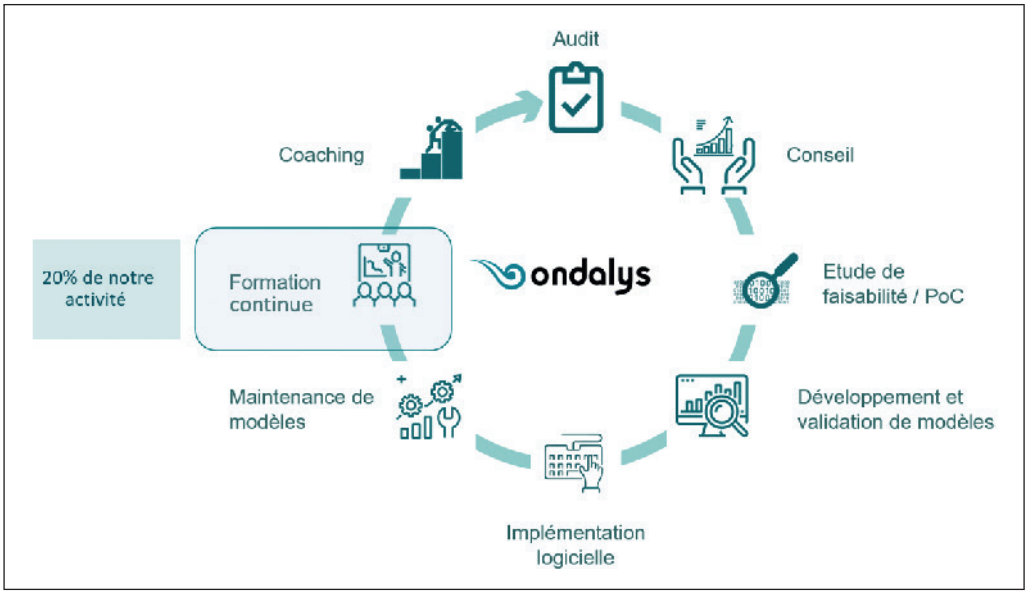


Figure 2

Différentes entrées de l'entreprise dans un projet.

artificielle. Il développe ensuite les activités de formation continue aux outils opérationnels et termine par la présentation de trois applications rendues possibles par ces outils.

1 Chimiométrie, *machine learning* et intelligence artificielle : sémantique et concepts

1.1. Sémantique

Le *deep learning* est une application du *machine learning* utilisant des algorithmes complexes de type réseaux de neurones. La chimiométrie quant à elle utilise des algorithmes linéaires multivariés. Enfin, d'autres outils non linéaires, comme les SVM (*Super Vector Machines*)[13], les arbres de régression, de classification[14], et les forêts aléatoires[15] font également partie du *machine learning*.

On peut signaler un certain nombre de mots-clés, signalés sur la ***Figure 3*** : *Big data*[16], *data science*, *data analytics*[17], MVA (*multivariate analysis*) – en français, analyse de données multivariées – *metabolomics* (analyse du métabolome[18]), etc. Mais retenons

13. *Super Vector Machines* : machines à vecteurs de support, ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression.

14. Arbres de régression, de classification : techniques de groupement des données.

15. Forêt aléatoire : regroupement d'arbres.

16. *Big data* : données massives, ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment travailler.

17. *Data analytics* : analyse de données.

18. Métabolome : ensemble des métabolites, des petites molécules telles que les intermédiaires métaboliques, les hormones et autres molécules signal ainsi que les métabolites secondaires, qui peuvent être trouvées dans un échantillon biologique.

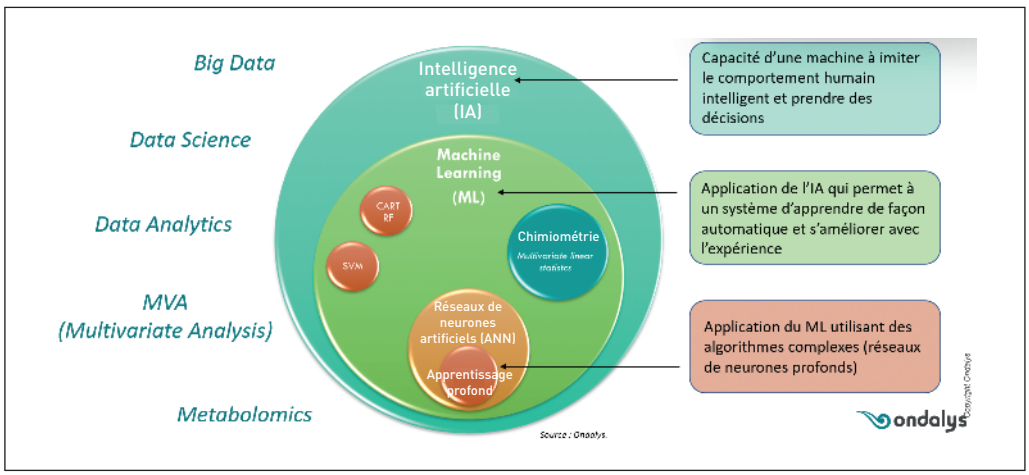


Figure 3

Distinction entre les notions importantes liées à l'intelligence artificielle.

que l'intelligence artificielle est une activité beaucoup plus vaste qui utilise des outils de *machine learning* et prend des décisions autonomes.

1.2. Exemples de concepts mis en œuvre

1.2.1. Signal multivarié

Pour (entre autres) la chimiométrie ou le *machine learning*, on peut utiliser différents types de présentation pour les données : soit la **description univariée**, des paramètres isolés, soit une présentation multidimensionnelle dite en **données multivariées**.

À gauche (*Figure 4*), est présenté un tableau de données consistant en une seule colonne : une variable ou un paramètre est décrit avec des valeurs continues ou qualitatives sur un certain nombre d'échantillons ou sur une population d'individus ; cela correspond à ce qu'on appelle un vecteur. C'est un objet mathématique bien précis, qui va être traité par des statistiques classiques, dites univariées.

Pour traiter simultanément plusieurs variables, mesurées de façon concomitante sur les mêmes échantillons, on établit un tableau (une matrice) de données avec plusieurs colonnes et toujours les mêmes lignes (voir à droite *Figure 4*)**.** Cette présentation permet de travailler sur les statistiques multivariées, utilisant les méthodes de l'algèbre linéaire et le calcul matriciel.

À quoi ressemblent ces données en chimie, ces signaux multivariés ? Il peut s'agir des paramètres procédés (*Figure 5* en haut à gauche) qu'ils soient mesurés *at-line*[19] ou *in-line*[20] sur un réacteur chimique. C'est classiquement le pH, la température, la pression, etc.

Mises bout à bout, ces mesures constituent un signal

19. *At-line* : près de la ligne.
20. *In-line* : dans le milieu.

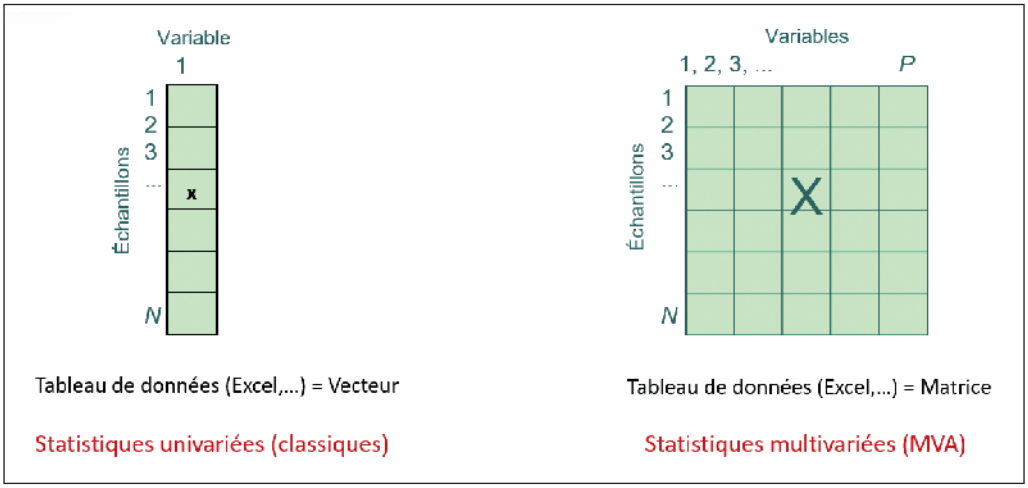


Figure 4

Comparaison entre les données univariées et multivariées.

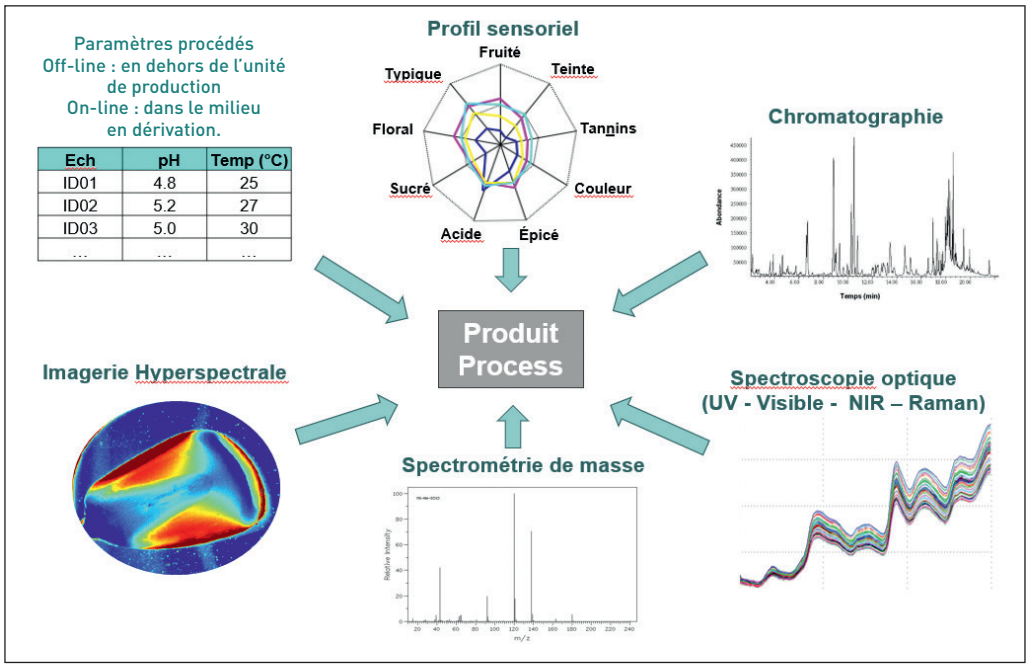


Figure 5

Différents types de données multivariées.

multivarié. En haut au centre, c'est l'exemple d'un profil sensoriel, rare en chimie (quoique peut-être au niveau des odeurs dans le milieu de l'automobile) mais très fréquent en agroalimentaire. À droite, c'est un chromatogramme issu de séparation sur colonne chromatographique. On présente aussi des spectres de masse, des spectres optiques type UV visible ou proche infrarouge, Raman[21] et de l'imagerie hyperspectrale[22]. Avec la spectroscopie optique, on enregistre une mesure d'intensité d'absorbance pour chacune des longueurs d'onde et on peut en extraire un « signal multivarié » (qu'on peut aussi appeler une empreinte ou un *fingerprint*[23]).

Les données de type chromatographie et spectrométrie de masse permettent l'utilisation de la chimiométrie et du *machine learning* en permettant de constituer des **données**

21. Spectroscopie Raman : méthodes non destructives d'observation et de caractérisation de la composition moléculaire et de la structure externe d'un matériau, qui exploite le phénomène physique selon lequel un milieu modifie légèrement la fréquence de la lumière y circulant.

22. Imagerie hyperspectrale : technologie permettant d'obtenir l'image d'une scène dans un grand nombre (généralement plus d'une centaine) de bandes spectrales à la fois étroites et contiguës.

23. *Fingerprint* : empreinte digitale.

métabolomiques[24] ; elles permettent de disposer de signaux synthétiques très utiles pour la comparaison d'échantillons.

1.2.2. Corrélation

Le concept de corrélation est majeur et d'utilisation constante.

Prenons l'exemple du suivi d'un paramètre, par exemple la température sur un procédé physicochimique. On suit la température dans le temps et on la reporte sur une carte de contrôle. L'ensemble de la population de points mesurés est représenté sur l'axe des abscisses (***Figure 6***) pour identifier les points compris entre des limites de contrôle (en l'occurrence fixées ici à plus ou moins 3 écarts-types), pour détecter des dérives du procédé (des non-conformités). Sur ce cas assez simple (axe des abscisses sur la ***Figure 6***), on a deux points, un à température un peu basse et un autre à température un peu haute, qui vont générer des alertes.

Si on suit un deuxième paramètre ou une deuxième variable, par exemple la pression, mesurée au même moment sur le même échantillon, on établit une carte analogue – ici sur l'axe des ordonnées (***Figure 6***) – permettant éventuellement d'observer des points anormaux, en l'occurrence un à pression inférieure et un autre à pression supérieure qui constituent des alarmes.

En considérant simultanément les deux paramètres, on a des informations beaucoup plus puissantes. Par exemple, le premier point situé à gauche (***Figure 6***) paraît conforme en température et en pression si on le considère isolément, mais comme atypique par rapport à la population étudiée car il s'écarte du nuage général. C'est typiquement le cas d'une anormalité **non détectée en univarié, mais détectée en multivarié**, c'est-à-dire, ici, en considérant simultanément la température et la pression. Une situation en sens inverse, un cas détecté non conforme à tort en univarié apparaît aussi sur le diagramme : c'est le cas d'un point détecté conforme sur la carte de contrôle multivariée mais qui déclencherait en univarié une alerte dans l'autre sens.

Ces exemples illustrent l'intérêt de travailler en multidimensionnel, tenant compte de la structure de corrélation entre les variables, et donc de l'utilisation des outils de chimiométrie et de *machine learning*.

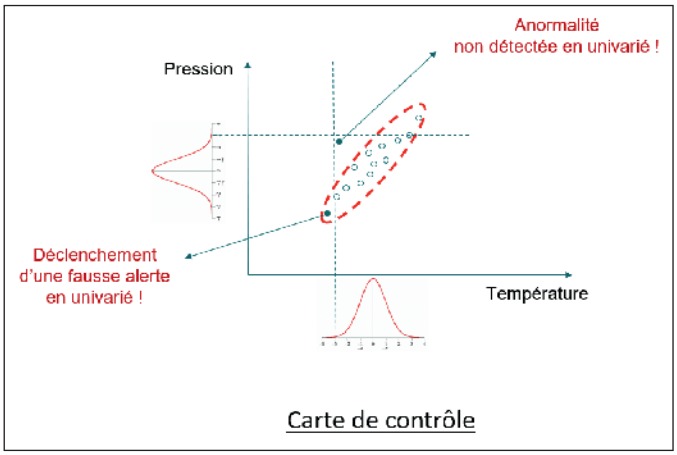


Figure 6

Carte de contrôle de la pression et de la température.

24. Étude des métabolites issus de l'organisme ou provenant de l'environnement.

RECONNAÎTRE AUTOMATIQUEMENT L'IMAGE D'UN CHAT

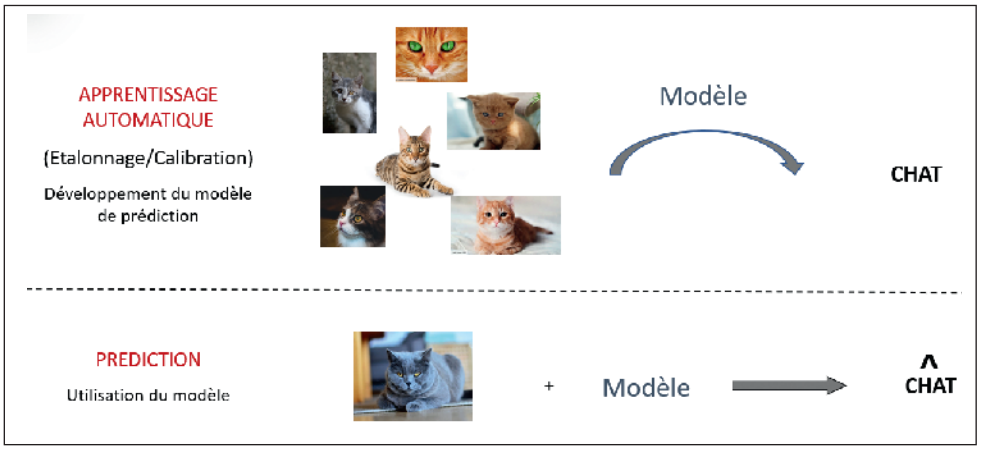


Figure 7

L'automatique pour reconnaître des images de chat.

Nous l'illustrons ici par l'apprentissage d'un algorithme sur des images de chats. Il s'agit de reconnaître ce qu'est un chat, au moyen d'un jeu d'étalonnage, d'entraînement ou de calibration, pour développer un modèle de prédiction (***Figure 7***). Une fois entraîné et optimisé, **le modèle s'appliquera sur l'image d'un nouveau chat : s'il marche bien, il prédira correctement qu'elle correspond bien à un chat.**

1.2.3. Apprentissage automatique

Le concept d'apprentissage automatique du *machine learning* est souvent évoqué (voir Encart : Reconnaître automatiquement l'image d'un chat).

On peut l'aborder en soulignant une correspondance entre données traitées traditionnellement et approches de *machine learning*. Dans les premières, dans le domaine de la chimie, on s'appuie sur des spectres qui se ressemblent tous et construisent une base à partir de différents échantillons. De la même façon, les outils de *machine learning* permettent d'établir une base de calibration spectrale (***Figure 8***) à partir d'un grand nombre de spectres et d'optimiser un modèle de prédiction sur une valeur quantitative ou qualitative. Quantitativement, cela peut être la viscosité, une teneur en molécule, alors que qualitativement, cela peut être une conformité, une qualité.

Le « modèle prédictif » sera appliqué à un nouveau spectre pour prédire le paramètre d'intérêt. C'est l'application typique de développement de calibration spectroscopique utilisant le *machine learning*.

Les deux grandes catégories d'apprentissage automatique sont : les **apprentissages non**

supervisés, utilisés pour le *data mining* ou le *clustering*[25] pour explorer les données de façon multivariée, et les **apprentissages supervisés** pour développer des modèles prédictifs, quantitatifs ou qualitatifs (***Figure 9***).

25. *Data clustering* : partitionnement de données, méthode d'analyse de données.

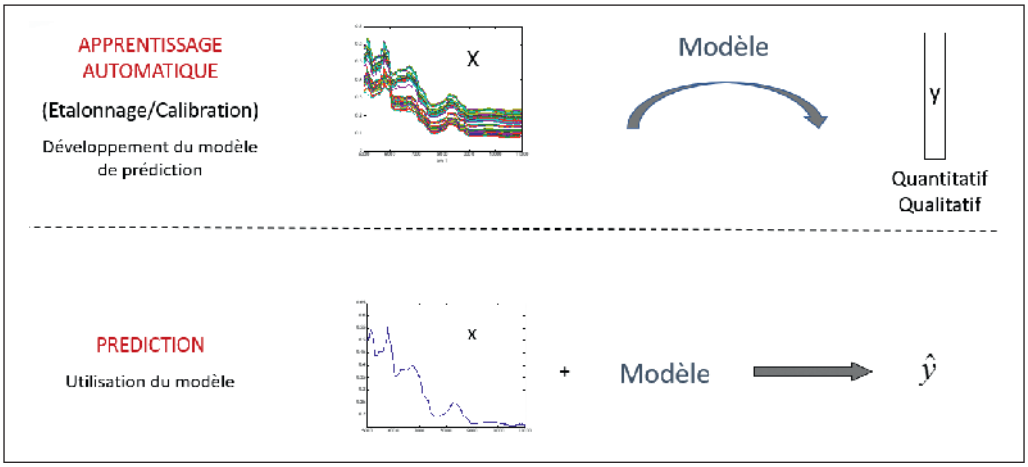


Figure 8

Concept de l'apprentissage automatique pour traiter des spectres.

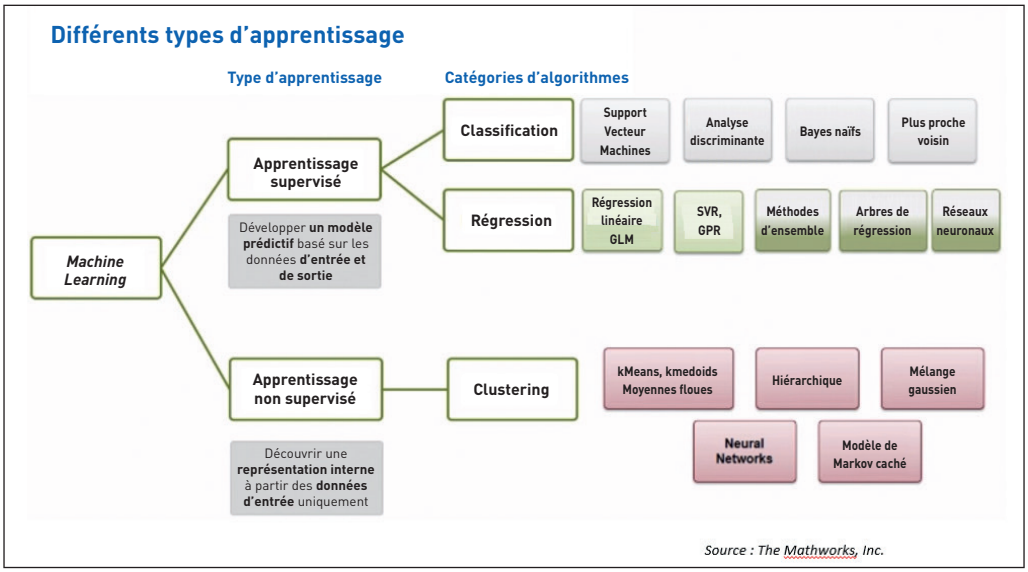


Figure 9

Différents types d'apprentissage automatique.

Souvent on passe beaucoup de temps pour trouver le meilleur algorithme de *machine learning*, mais il faut souligner qu'une étape préalable reste tout à fait indispensable : c'est **le nettoyage et l'exploration des données qui seront utilisées,** processus appelé « conciliation ».

On appelle « conciliation des données » tout ce qui est alignement et synchronisation des données provenant de différents instruments ou capteurs. Le nettoyage des données utilise des statistiques classiques (distribution, quartiles, moyenne, etc.) ; il utilise aussi des statistiques multivariées de la même manière avec, notamment, le choix de la composante principale pour faire ressortir ce qui est outlier[26], tendance, cluster[27], etc.

De nouveau, il y a lieu d'insister sur la tâche de fiabilisation des données (***Figure 10***) à effectuer au préalable et qui est nécessaire pour augmenter la performance et faciliter l'interprétation ultérieure des résultats. On voit là qu'il y a quand même une « valeur ajoutée de l'humain » – on parlera de *human learning*[28] pour illustrer cette notion avec un peu d'humour –, c'est en fait une étape très chronophage dont on ne doit pas se passer avant d'entrer toutes les données dans les algorithmes de *machine learning*.

26. Outlier : individu atypique.
27. Cluster : groupe.

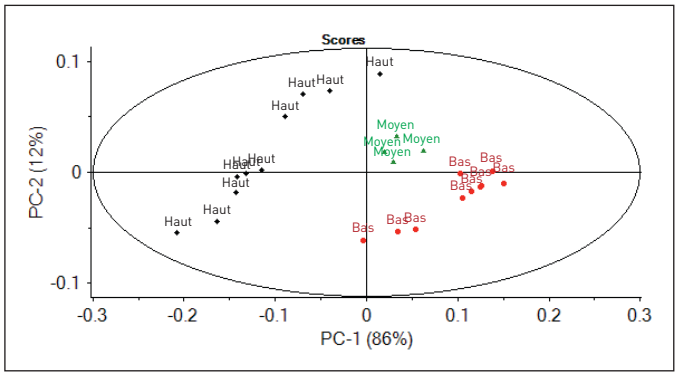


Figure 10

Graphique permettant de juger la fiabilisation de nos données.
Source : The Mathworks, Inc.

2 La formation continue aux outils opérationnels de chimiométrie et *machine learning* chez Ondalys

La partie principale du chapitre a pour but de partager l'expérience développée chez Ondalys dans la formation continue sur ces outils de chimiométrie et de *machine learning*.

2.1. Présentation de la formation

Nous sommes cinq formateurs au sein d'Ondalys. La petite valise au centre (***Figure 11***) représente les compétences multidisciplinaires et c'est un message adressé aux étudiants : quand vous avez les choix à faire dans vos études à un moment donné, pensez aussi multidisciplinarité. Il ne s'agit pas d'avoir seulement les compétences en mathématiques, en statistiques, en modélisation ou en algorithmique ; il faut aussi connaître « les métiers », soit tout ce qui

28. *Human learning* : apprentissage humain.

Figure 11

Équipe de formateurs d'Ondalys.

est procédés, produits, production, laboratoire, R&D, et puis au-delà, les techniques de codage et les langages de programmation. En chimiométrie et *machine learning*, on est à l'interface de plusieurs disciplines entre le métier, la chimie, les compétences en mathématiques, statistiques et informatique.

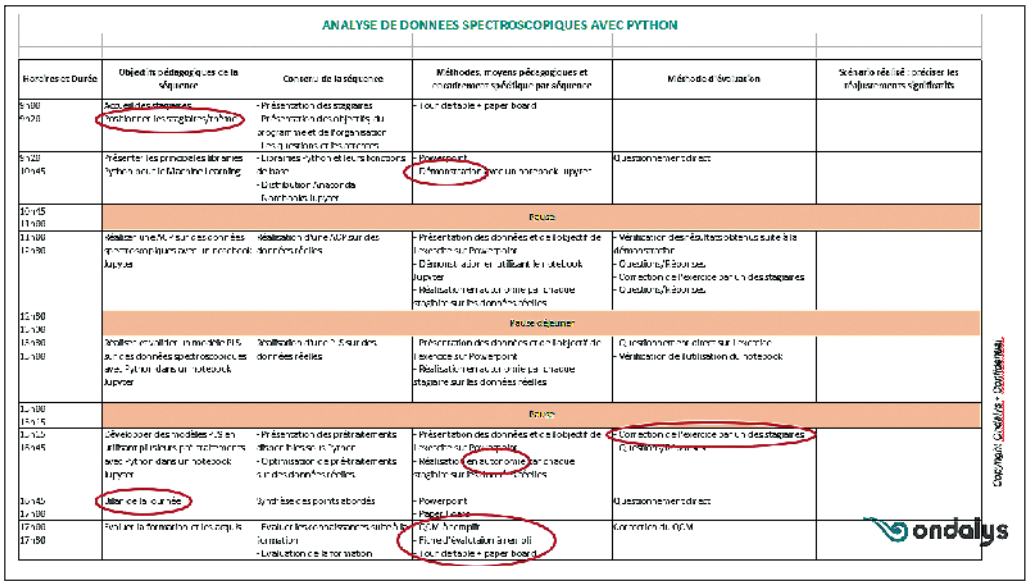
La ***Figure 12*** présente un exemple de déroulé pédagogique sur la formation en analyse de données spectroscopiques de type proche infrarouge ou Raman sur « Python™ ». Le processus est certifié Qualiopi, et comporte une quantité de milestones[29] et de points à surveiller, notamment positionner les besoins des différents stagiaires en début de formation. La formation mélange la théorie, la méthodologie, les petites astuces dans le traitement de données et la pratique, car il y a beaucoup de pratiques à acquérir sur des logiciels partenaires ou sur des langages de programmation type Python™.

La formation est ponctuée de démonstrations du formateur, ainsi que d'exercices réalisés en autonomie et corrigés par un des stagiaires. L'idée est que le stagiaire sorte de cette formation, qui dure en général deux ou trois jours, en étant capable de traiter les données lui-même sur un logiciel ou avec un langage de programmation. Elle est adressée à des débutants et se conclut par un bilan effectué avec un QCM à remplir sur l'évaluation du contenu, une fiche d'évaluation sur la formation et sur le formateur, puis un tour de table final.

2.2. Formation sur mesure

On propose des programmes de formation sur mesure, notamment en intra-entreprise

29. Milestones : jalons.



ANALYSE DE DONNEES SPECTROSCOPIQUES AVEC PYTHON

Horaires et Durée	Objectifs pédagogiques de la séquence	Contenu de la séquence	Méthodes, moyens pédagogiques et ordonnancement synthétique par séquence	Méthodes d'évaluation	Scénario réalisé : préciser les aménagements significatifs
9h00 9h20	Accueil des stagiaires Présentation des stagiaires/Python	- Présentation des stagiaires Présentation des objectifs du programme et de l'organisation Les questions préliminaires	- Tour de table + paper board		
9h20 10h45	Présenter les principales librairies Python pour le Machine Learning	- Librairies Python et leurs fonctions de base - Distribution Anaconda - Notebooks Jupyter	- Powerpoint - Démonstration avec un notebook Jupyter	Questionnement direct	
10h45 11h00	Pause				
11h00 12h30	Réaliser une ACP sur des données spectroscopiques avec un notebook Jupyter	Réalisation d'une ACP sur des données réelles	- Présentation des données et de l'objectif de l'exercice sur Powerpoint - Démonstration et utilisation du notebook Jupyter - Réalisation individuelle par chaque stagiaire sur les données réelles	- Vérification des résultats obtenus suite à la démonstration - Questions/Réponses - Correction de l'exercice par les stagiaires - Questions/réponses	
12h30 13h30	Pause déjeuner				
13h30 15h00	Réaliser une régression PLS sur des données spectroscopiques avec Python dans un notebook Jupyter	Réalisation d'une PLS sur des données réelles	Présentation des données et de l'objectif de l'exercice sur Powerpoint - Réalisation individuelle par chaque stagiaire sur les données réelles	Questionnement direct sur l'exercice - Vérification de l'utilisation du notebook	
15h00 15h15	Pause				
15h15 16h45	Développer des modèles PLS en utilisant plusieurs prétraitements avec Python dans un notebook Jupyter	- Présentation des prétraitements disponibles sous Python - Optimisation de prétraitements sur des données réelles	- Présentation des données et de l'objectif de l'exercice sur Powerpoint - Réalisation en autonomie par chaque stagiaire sur les données réelles	- Correction de l'exercice par les stagiaires - Questions/réponses	
16h45 17h00	Aborder les limites	Synthèse des points abordés	- Powerpoint - Paper board	Questionnement direct	
17h00 17h30	Evaluer la formation et les acquis	Evaluation des connaissances suite à la formation - Evaluation de la formation	- QCM Scrumpill - Fiche d'évaluation individuelle - Tour de table + paper board	Correction du QCM	

ondalys

Figure 12

Déroulé pédagogique d'une formation en analyse de données spectroscopiques avec Python™.

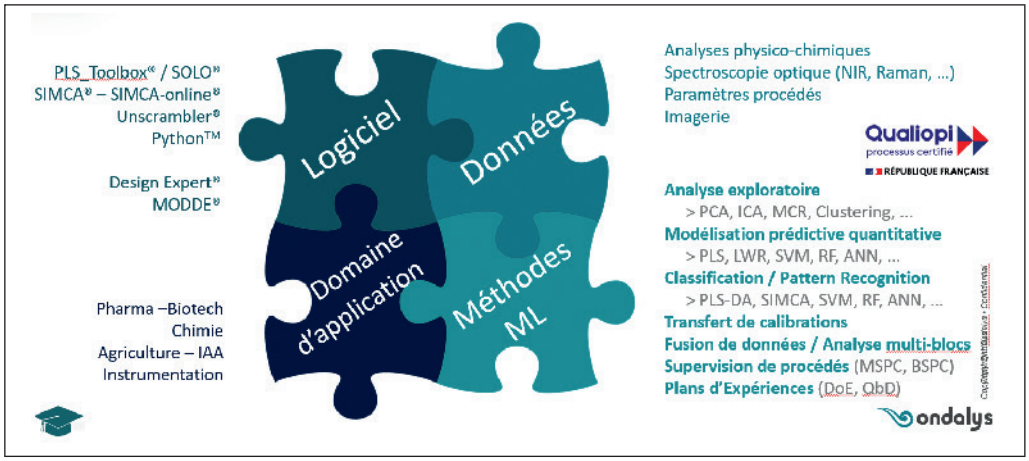


Figure 13

Les quatre composantes de la formation continue.

sur site, qui sont le résultat de quatre composantes (***Figure 13***) :

- le type des données qui seront traitées ultérieurement, que ce soit des analyses physico-chimiques, de la spectroscopie optique, des paramètres procédés, de l'imagerie ;
- le domaine dans lequel le client intervient : en pharma,

biotech, chimie ou autre industrie ;

• les différentes thématiques de *machine learning*, que ce soit l'analyse exploratoire, la modélisation prédictive quantitative, classification, fusion de données, analyses multiblocs, supervision de procédés continus ou en batchs, ou plans d'expériences ;

• les pratiques variées sur des logiciels abordés dans les formations ; le client doit choisir le logiciel sur lequel il veut être formé ; s'il n'a pas d'idées, il est conseillé.

2.3. Logiciels de formations

La ***Figure 14*** montre les différents logiciels sur lesquels sont effectuées les formations. Il y en a deux grandes catégories : les logiciels à interface graphique, qui intéressent plutôt les débutants ; et les langages de programmation, type Matlab ou Python™, qui intéressent des ingénieurs confirmés ou qui veulent en faire leur métier. Dans les logiciels à interface graphique, nous avons quatre partenaires : Eigenvector Research, Sartorius, Aspentech et StatEase, qui proposent chacun différents logiciels ou des suites de logiciels pour adresser le développement de modèles de *machine learning*, leur mise en œuvre et éventuellement le développement de plans d'expériences.

L'offre de formation continue est proposée en intra-entreprise sur site. Elle consiste en formations sur mesure ; alternativement des thématiques fixées avec des dates, des lieux et des logiciels prédéfinis sont proposées. Depuis la période Covid, nous proposons aussi des formations « online ». Le public concerné est constitué d'industriels, mais également de centres de recherche

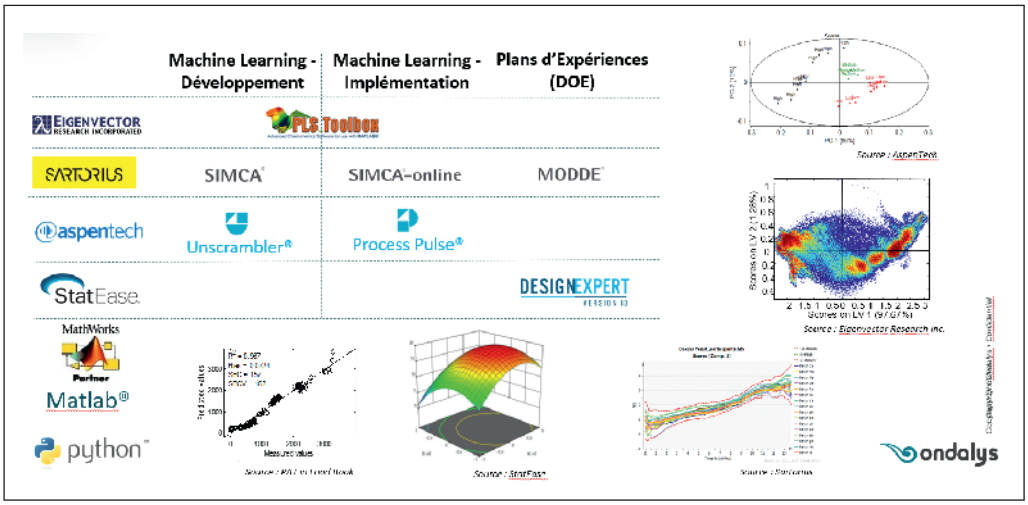


Figure 14

Logiciels partenaires.

publics et de centres techniques ; il est constitué de chercheurs, de techniciens, d'ingénieurs et occasionnellement de managers. Les formations sont données en français ou en anglais.

On propose également après la formation des services associés de type coaching, pour accompagner les stagiaires dans la mise en œuvre de ces outils avec support client, visio ou téléphone, ainsi que des prestations de traitement de données. Le gros de notre activité dans ce domaine est de traiter les données des clients sur des problématiques un peu complexes. On distribue également certains logiciels de nos partenaires.

La société Ondalys est référencée dans différents catalogues de partenaires pour la formation continue, notamment CPE Lyon, EASE Training et MabDesign pour la biotech. Quelques chiffres issus des questionnaires d'évaluation de la formation montrent que les gens sont « satisfaits » pour l'instant de ce que nous proposons depuis plus d'une quinzaine d'années maintenant. Quelques références qui nous ont fait confiance pour la formation continue dans les différents domaines sont listées sur la ***Figure 15*** : pharma, biotech, chimie, agriculture, agroalimentaire, instrumentation et centre de recherche.

3 Pour quoi faire ? Quelques exemples d'applications

Dans sa dernière partie, le chapitre présente des applications très différentes qui ont été développées avec les outils de *machine learning* et de chimiométrie.

3.1. Calibrations spectroscopiques

La première application traite du **développement de**

Figure 15

Clients d'Ondalys dans les différents domaines.

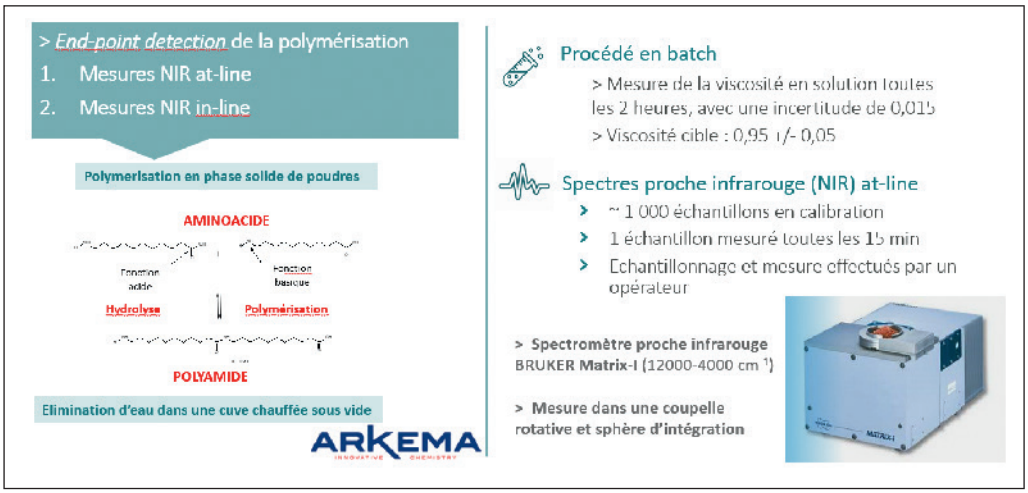


Figure 16

Résumé des besoins d'Arkema dans le modèle at-line.

calibrations spectroscopiques mises au point avec l'industriel Arkema en Normandie (***Figure 16***). Le but était d'avoir une aide au pilotage d'un procédé de polymérisation, notamment en gérant ce qu'on appelle le End-point[30] (fin du procédé) où il faut arrêter le process. Le procédé est en batch[31] classique, ce qui demande une mesure de la viscosité toutes les deux heures. Le besoin était d'avoir une mesure plus rapide, même éventuellement indirecte sur un spectromètre NIR (*Near InfraRed*)[32] en faisant usage dans un premier temps d'un échantillonnage opérateur par prélèvement (modèle at-line).

La base de calibration (d'entraînement) était constituée d'environ mille échantillons pour lesquels on disposait à la fois des spectres NIR et de la viscosité mesurée en laboratoire. Le modèle a été entraîné avec des outils de *machine learning* classiques et a fourni le modèle présenté ***Figure 17***, qui a permis de suivre la cinétique. On voit en rouge les mesures de référence au labo et les petites croix violettes qui correspondent aux prédictions proche infrarouge.

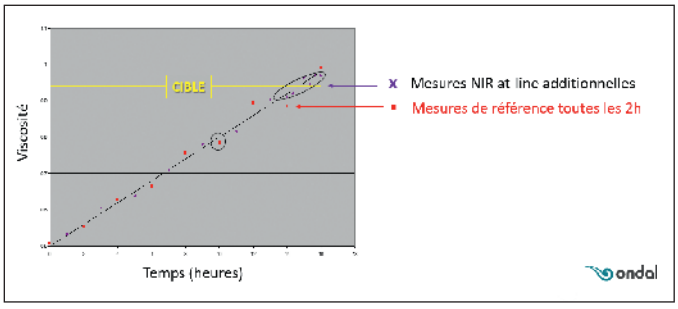


Figure 17

Détection du End-point de la polymérisation avec des mesures proche infrarouge at-line.

30. End-point : fin du procédé.
31. Traitement industriel par lots dans lequel le produit fini est obtenu à la suite d'une série de tâches plutôt qu'en continu.
32. NIR (*Near InfraRed*) : proche infrarouge.

Le modèle fonctionnait bien, avec une faisabilité qui donnait satisfaction ; à ceci près que cela nécessitait toujours un échantillonnage manuel et le fait d'analyser les échantillons en laboratoire engendrait une influence de la température et de l'humidité du milieu environnant.

Pour améliorer la performance, on a regardé si les contrôles pouvaient marcher également sur un analyseur NIR « en ligne » (***Figure 18***). Une deuxième base de données d'étalonnage a été construite en mesurant environ 10 000 échantillons avec, cette fois-ci, un instrument de process, équipé d'une sonde et d'une fibre optique rattachée, permettant une mesure en ligne toutes les deux minutes.

La ***Figure 19*** montre la cinétique de prédiction. On voit la montée en viscosité du produit sur un batch jusqu'à la cible, qui permet de décider du Endpoint. En rose, est présentée la température, puis, en bleu foncé, ce qu'on appelle la distance de Mahalanobis qui est une mesure de la proximité spectrale d'un nouvel échantillon par rapport à la base d'étalonnage et permet de déterminer le domaine de validité du modèle de prédiction ; ici, il est valide au-dessus d'une certaine température. Cette application a très bien marché et ces analyseurs proche-infrarouge ou Raman sont de plus en plus montés sur les installations en chimie, en pharma, en biotech. Nous faisons face à beaucoup de demandes.

3.2. Chromatographie couplée à la spectrométrie de masse

La deuxième application abordée concerne un type de données complètement différent : il s'agit de traiter des données de laboratoire issues d'une chromatographie couplée à

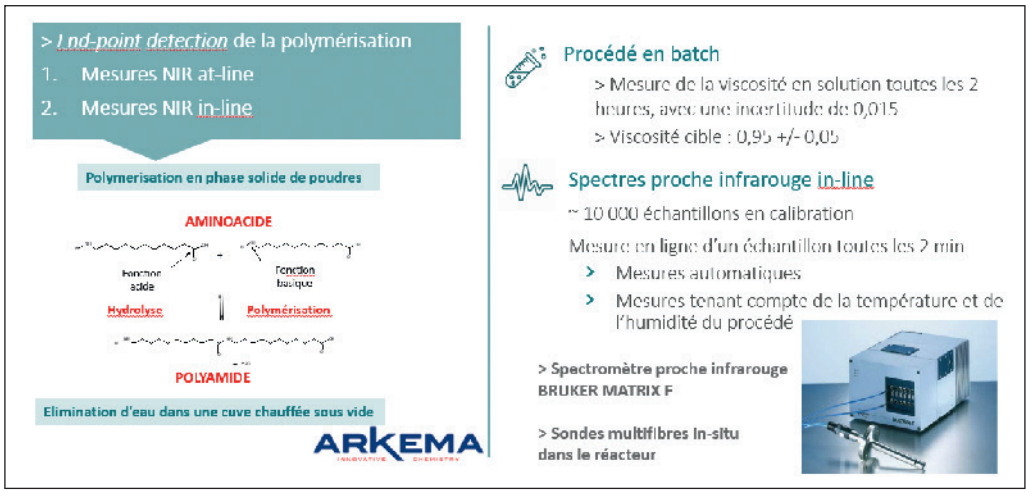


Figure 18

Résumé des besoins d'Arkema dans le modèle in-line.

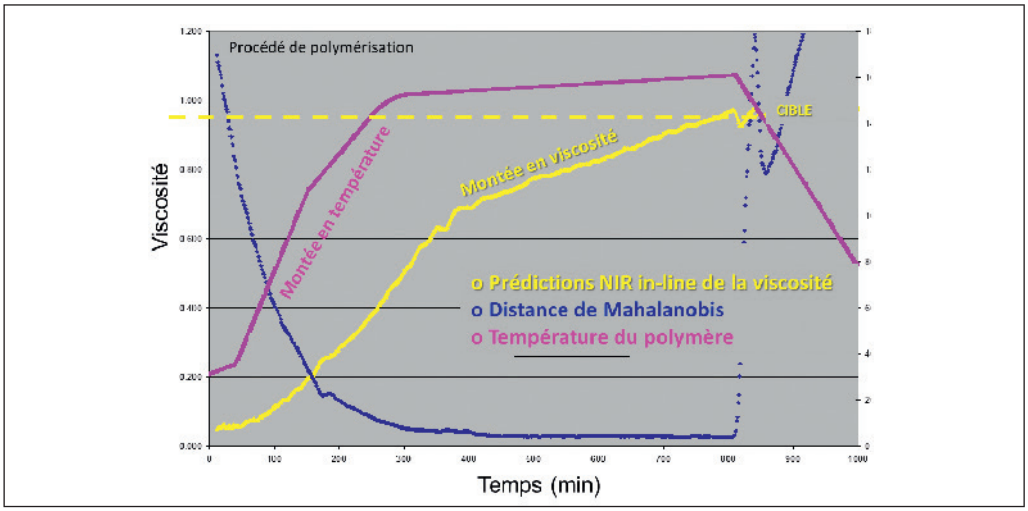


Figure 19

Détection du End-point de la polymérisation avec des mesures proche infrarouge at-line.

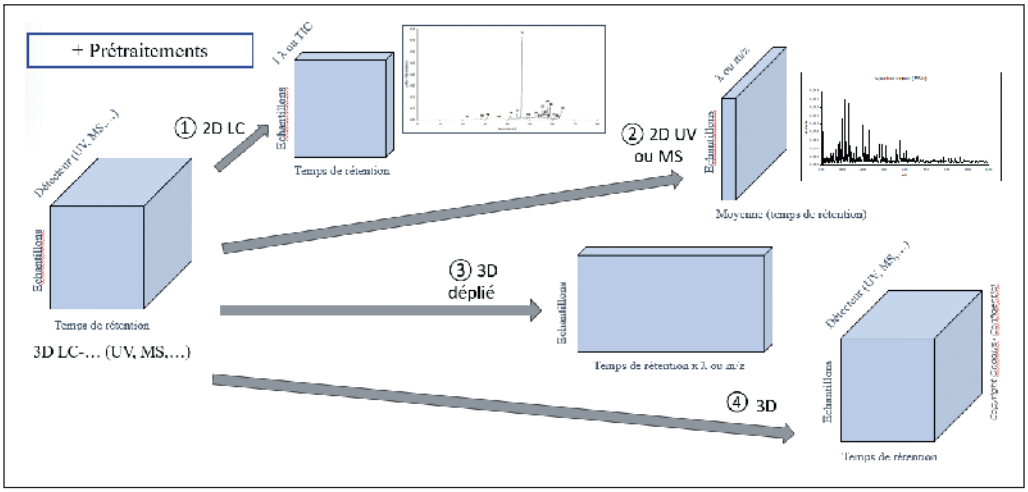


Figure 20

Différentes stratégies pour traiter les données de la chromatographie par machine learning.

une spectrométrie de masse. Cette technique génère une masse de données importantes très complexes présentées ici sur trois dimensions (***Figure 20***) : la dimension échantillon, une dimension chromatographique (évolution du temps de rétention) et une dimension détectrice, dans le cas présent un détecteur masse, qui donne un spectre de masse à chaque temps de rétention du chromatographe. Les mesures fournissent des cubes de données dont il faut

déterminer s'ils peuvent être traités facilement par de la chimiométrie et du *machine learning*. La réponse étant non, cela a conduit à adopter plusieurs stratégies alternatives.

Parmi les stratégies présentées (***Figure 20***), nous avons choisi de travailler sur la dimension chromatographique (voie 1), sur la dimension masse (voie 2), ou sur du déplié (voie 3), mais d'éviter de traiter directement les données 3D (voie 4) très complexes dans le cas présent.

Dans la stratégie 1 (***Figure 20***) – dimension chromatographique – le signal retenu ici correspond à un résumé de

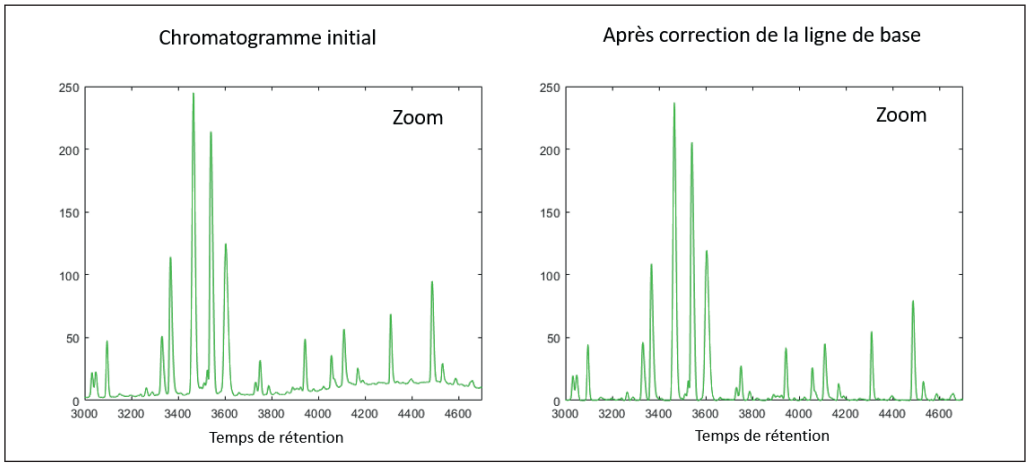


Figure 21

Correction de la ligne de base.

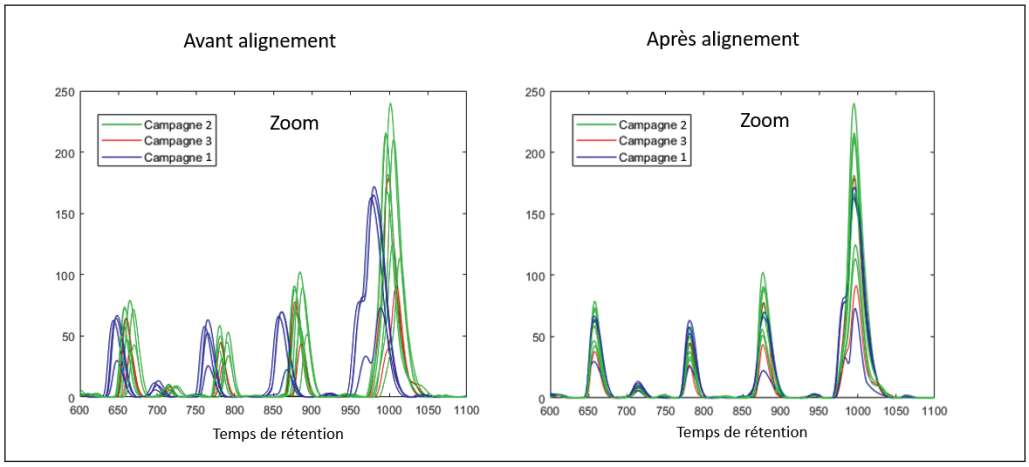


Figure 22

Correction de l'alignement.

l'information de masse TIC (*Total Ion Current*)[33], synthèse des m/z[34] recueillis. On obtient ce genre de signaux représentés sur la ***Figure 21*** qui comportent une forêt de pics.

Un certain nombre de prétraitements sont appliqués pour nettoyer le signal, avant de le soumettre au *machine learning*. Ce peut être par exemple une correction ligne de base (***Figure 21***) ou une correction d'alignement (***Figure 22***).

Il faut aussi aligner les chromatogrammes entre eux (***Figure 23***), parce qu'il y a des « décalages » qui se créent avec la colonne et qu'avant de rentrer ces données dans les algorithmes de chimiométrie et de *machine learning*, il est nécessaire d'aligner tous ces pics.

L'application d'un algorithme, qui s'appelle SIMCA (*Soft Independent Modeling for Class Analogies*)[35], a permis de réaliser la discrimination. Dans ce cas-là, les petits triangles bleus (***Figure 23***) correspondent aux échantillons de calibration, des échantillons conformes qui ont permis d'entraîner le modèle et qui sont bien qualifiés comme conformes. Ce modèle a été testé sur des échantillons conformes en vert, et des échantillons non conformes, en rouge. La validation a été satisfaisante : le cas fonctionne bien.

Pour cette application, nous avons utilisé la dimension chromatographique comme signal global. Cette approche diffère de l'approche classique

33. *Total Ion Current* : courant ionique total.
34. m : masse, z : charge.

35. *Soft Independent Modeling for Class Analogies* : Modélisation douce et indépendante pour les analogies de classes.

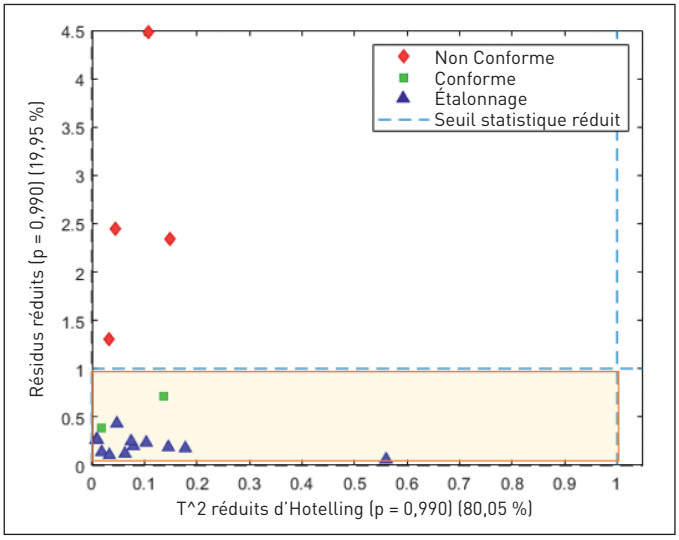


Figure 23

Modélisation de la conformité.

où l'on utilise la masse pour identifier les pics, et où l'on intègre ensuite les aires des pics identifiés pour effectuer de la quantification.

3.3. Modélisation de procédés batch

Le dernier exemple est encore très différent : il concerne la modélisation de procédés batchs. Un certain nombre de variables de procédés (éthanol, température, niveau, temps, pH, etc.) ont été mesurées tout au long de la cinétique, ce qui donne lieu à un cube de données (***Figure 24***). Sur ces mêmes batchs, des paramètres « qualité », caractérisant le produit fini, ont été mesurés ; des conditions initiales sur des variables caractérisant les matières premières entrant dans les batchs ont également été intégrées. Nous disposons alors d'une collection de données hétérogènes, très différentes entre elles et la question est : comment combiner ces données et permettre soit de gagner en compréhension, soit de faire des modèles prédictifs, soit de faire la supervision du procédé en multivarié, ce qui donne le résultat de la ***Figure 25***. Finalement, toutes ces variables sont résumées de façon multivariée par ce qu'on appelle des scores dans une enveloppe statistique, qu'on appelle aussi le « *golden batch* ». L'objectif consiste soit de piloter les procédés en temps réel, soit en différé de faire du troubleshooting pour voir pourquoi un batch de production s'est mal passé. On utilise ainsi une méthode qu'on appelle le « *contribution plot* » : on va cliquer à un endroit où le procédé peut dériver et on va vérifier les variables qui peuvent être responsables de cette dérive (***Figure 26***).

Ces trois exemples montrent bien la diversité des applications qui peuvent être mises en œuvre dans l'industrie de process avec des outils de chimiométrie et de *machine learning*.

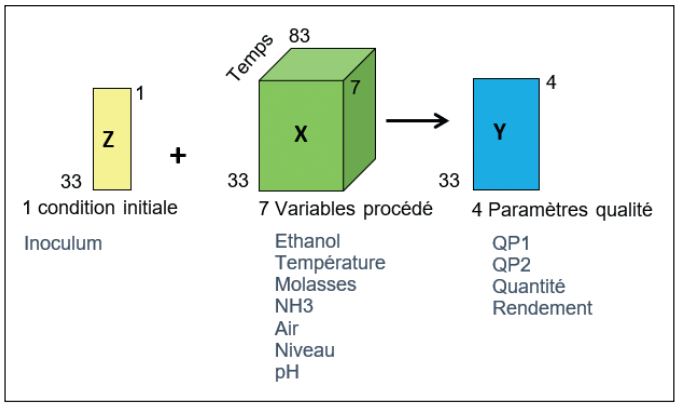


Figure 24

Modélisation d'un procédé batch.

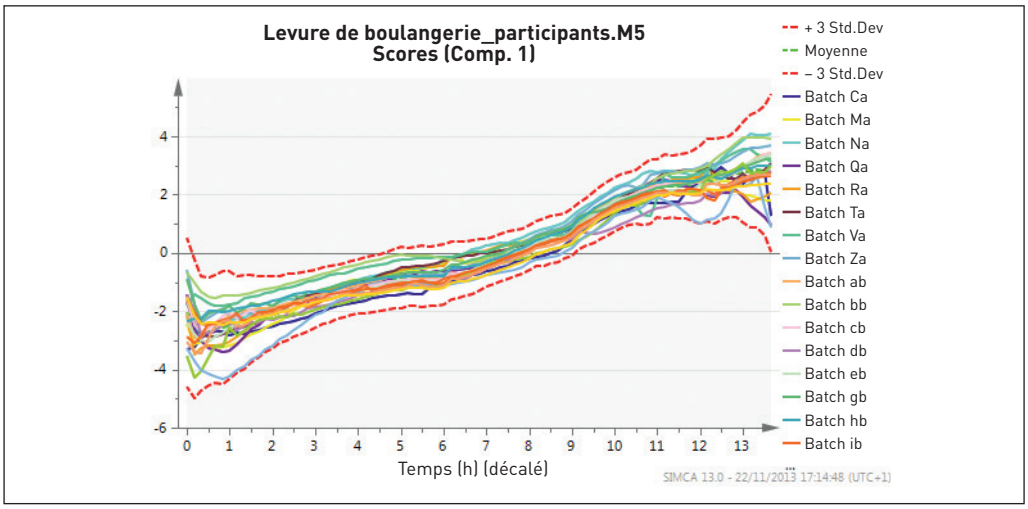


Figure 25

Exemple de process monitoring (supervision de procédé).

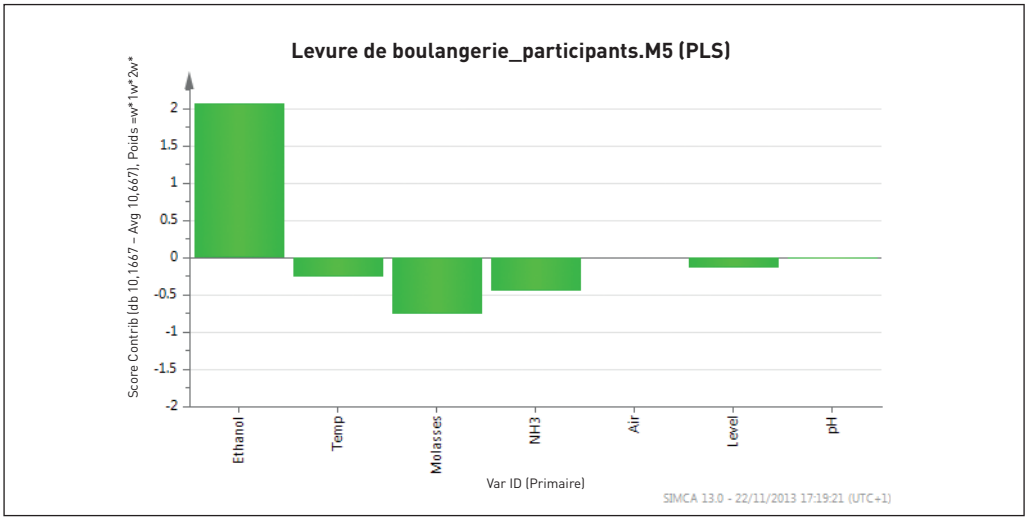


Figure 26

« Contribution plot ».

Conclusion
En chimie comme ailleurs, l'intelligence artificielle n'a pas dit son dernier mot !

Pour terminer, je ne résiste pas à l'envie de vous montrer une petite expérience de la semaine dernière sur des outils d'intelligence artificielle qui sont devenus très à la mode tout d'un coup. C'est l'interface qui s'appelle Dall-E d'OpenAI, c'est un peu un analogue de Chat GPT, mais qui traite les images au lieu de traiter les textes.

Par curiosité, j'ai mis la définition de *machine learning* sur Wikipédia et le logiciel m'a transformé la définition en image. Cela donne la ***Figure 27*** : chacun aura un avis sur cette image ! J'ai fait une deuxième expérience en faisant un résumé en anglais de mon intervention présente sur la formation continue (parce que ça marche mieux pour l'instant en anglais

Figure 27

Image représentant la définition de « Machine Learning » par Dall-E.

qu'en français), et il m'a sorti l'image de la ***Figure 28***, qui est, finalement, assez réaliste malgré les quelques défauts du visage.

L'engouement actuel du grand public pour ces outils d'intelligence artificielle a beau être tout jeune, on le voit déjà, après quelques mois, s'amplifier à toute vitesse. Il s'agit là d'exemples ludiques, mais je crois avoir montré dans ce chapitre que la dynamique est tout aussi vigoureuse pour l'industrie en général et **pour l'industrie chimique en particulier**. Des résultats très nouveaux sont déjà obtenus. Ils sont annonciateurs de performances extraordinaires qu'on n'ose même pas évoquer en craignant de quitter le réalisme… Mais ces techniques d'intelligence artificielle sont parties et nous aurons à travailler pour les suivre et en tirer des bénéfices inimaginables ! Allons-y !

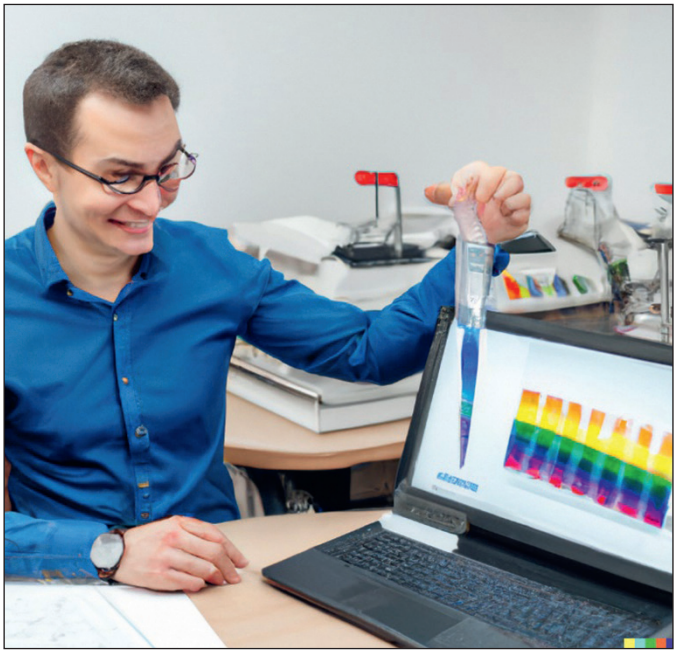


Figure 28

Image représentant la formation continue à la chimiométrie et au « Machine Learning ».