L’EDA comme acte fondateur de la data science

Dans la représentation la plus courante de la data science, l’analyse exploratoire des données, ou EDA, est souvent présentée comme une étape préparatoire. On commence par regarder les données, on trace quelques graphiques, on calcule quelques statistiques descriptives, on vérifie les valeurs manquantes, on examine les corrélations, puis l’on passe à ce qui serait supposé constituer le cœur du travail : le modèle, l’algorithme, la prédiction, la décision. Dans cette vision, l’EDA a une fonction presque subalterne. Elle sert à préparer le terrain avant l’entrée en scène des “vrais” outils.

Une telle représentation est, à mes yeux, profondément insuffisante. Elle ne décrit pas ce qui se joue réellement lorsque l’on cherche à comprendre un système complexe à partir de données. Elle passe à côté du moment le plus important, celui où le chercheur, l’analyste ou le data scientist ne plaque pas encore une forme théorique sur le réel, mais commence au contraire à découvrir la logique propre du phénomène observé. L’EDA n’est pas seulement une étape de préparation. Elle est le moment où les données commencent à devenir intelligibles. Elle est le point de passage entre l’accumulation brute des observations et l’émergence d’une structure.

Cette idée mérite d’être développée, car elle change profondément la manière de concevoir la data science. Si l’on réduit cette dernière à un ensemble d’algorithmes, de méthodes d’apprentissage ou de techniques d’optimisation, on risque d’oublier que les données ne parlent pas d’elles-mêmes et que les modèles ne valent que par leur capacité à respecter les formes réelles du phénomène. La question essentielle n’est donc pas seulement : quel algorithme vais-je appliquer ? Elle est d’abord : qu’est-ce que les données révèlent du système ? Quels sont les invariants empiriques ? Quelles sont les asymétries, les persistances, les intermittences, les régimes, les échelles temporelles, les événements extrêmes, les formes de mémoire ou les directions dominantes qui se dégagent de l’observation ?

C’est là que l’EDA prend toute son importance. Elle n’est pas une opération périphérique. Elle est l’acte fondateur de la connaissance en data science.

Cette position s’enracine dans une conviction simple. On ne comprend pas un système complexe en lui imposant immédiatement un modèle. On commence par le regarder. Mais regarder, ici, ne signifie pas simplement visualiser. Cela signifie apprendre à reconnaître des formes, à discerner ce qui est stable dans l’irrégulier, à repérer ce qui revient à travers la variabilité apparente. L’EDA devient alors une discipline du regard. Elle est le lieu où l’on suspend provisoirement le désir de conclure pour laisser apparaître les contraintes du réel.

Ce point est capital, car un système complexe présente presque toujours une double apparence. D’un côté, il semble désordonné. Les fluctuations sont nombreuses, les trajectoires paraissent irrégulières, les événements se succèdent sans schéma évident. D’un autre côté, à mesure que l’on approfondit l’examen, certaines formes reviennent. Des distributions particulières se dessinent. Des corrélations apparaissent ou, au contraire,

disparaissent. Des changements d'échelle révèlent des régularités inattendues. Des alternances de phases calmes et agitées se manifestent. Des structures de mémoire ou de dépendance deviennent visibles. Le travail de l'EDA consiste précisément à faire apparaître cette seconde couche du réel, celle qui n'est pas immédiatement visible à l'œil nu mais qui organise pourtant le comportement du système.

Dans cette perspective, l'EDA a une fonction de vérité. Elle oblige le chercheur à se confronter à ce qui est là, avant de décider ce qu'il veut y voir. Elle joue un rôle presque critique : elle empêche de plaquer trop vite une théorie séduisante mais inadéquate. Elle rappelle que le réel possède ses propres exigences, et que tout modèle sérieux doit s'y soumettre. Ce rôle est d'autant plus important que les outils contemporains de calcul permettent d'ajuster presque n'importe quoi à presque n'importe quoi. Il est toujours possible de construire un dispositif mathématique ou algorithmique qui produise une performance locale acceptable. Mais la question n'est pas là. La vraie question est de savoir si ce dispositif respecte la structure profonde du phénomène.

L'EDA intervient justement à cet endroit. Elle permet de distinguer entre ajustement et compréhension.

Cette distinction prend une importance particulière lorsque l'on travaille sur des systèmes où l'aléa, la mémoire et les changements de régime jouent un rôle central. Dans de tels cas, les statistiques classiques peuvent être trompeuses si elles sont interprétées trop rapidement. Une série temporelle peut sembler chaotique à première vue alors qu'elle recèle une organisation lente dans l'intensité de ses fluctuations. Un signal peut paraître imprévisible à courte échelle tout en révélant des structures stables lorsqu'on l'examine à d'autres résolutions. Un ensemble de données peut suggérer un bruit diffus alors qu'il correspond en réalité à une alternance entre plusieurs régimes dynamiques. Sans EDA approfondie, on risque de traiter comme du simple hasard ce qui relève d'une architecture cachée.

C'est précisément ce type de situation qui apparaît lorsqu'on examine certaines grandes séries financières, comme celles du CAC 40. L'observation attentive de ces données montre que les rendements journaliers possèdent peu d'autocorrélation, ce qui signifie qu'ils se comportent en première approximation comme une agitation difficilement prévisible à court terme. Mais cette apparente absence de structure est trompeuse. Dès que l'on s'intéresse non plus au signe ou à la valeur immédiate des variations, mais à leur intensité, une autre image apparaît. Le carré des rendements, ou toute mesure voisine de la volatilité, présente une persistance notable. Les grandes fluctuations tendent à se regrouper. Les périodes de calme et les périodes d'agitation se succèdent par plages. Cette propriété, essentielle, n'apparaît que si l'on accepte de regarder au bon niveau.

Ainsi, l'EDA ne se contente pas de décrire la série. Elle révèle qu'il existe en réalité deux temps du phénomène. Un temps rapide, où les variations directionnelles semblent proches d'un bruit peu corrélé, et un temps lent, où l'intensité même de l'agitation porte une mémoire. Cette découverte est décisive, car elle transforme immédiatement la nature du problème. Si la volatilité a une mémoire longue, alors le bruit lui-même doit être pensé

comme organisé. Le désordre apparent n'est plus un désordre pur. Il devient une agitation structurée, modulée, porteuse de temporalités emboîtées.

Ce simple exemple montre pourquoi l'EDA est bien davantage qu'une étape préliminaire. Elle est ce qui permet de formuler correctement la question. Avant même de chercher une équation, elle oblige à demander : quel type de bruit est à l'œuvre ? Y a-t-il mémoire ? Y a-t-il intermittence ? Y a-t-il alternance de régimes ? Y a-t-il auto-similarité ? Y a-t-il une variable dominante qui concentre l'essentiel du comportement global ? Tant que ces questions ne sont pas posées, la modélisation reste aveugle.

L'un des apports majeurs de l'EDA est précisément de faire émerger ce que l'on pourrait appeler les faits stylisés ou les invariants empiriques du système. Un invariant empirique n'est pas une loi fondamentale au sens strict. C'est une propriété suffisamment robuste pour s'imposer comme contrainte à toute modélisation sérieuse. Dans les séries financières, par exemple, la présence de queues lourdes dans les distributions de variations constitue un tel invariant. Les événements extrêmes y sont plus fréquents que dans un cadre gaussien simple. Cette propriété change tout. Elle signifie qu'un modèle de diffusion élémentaire, à variance constante, est insuffisant. De même, la mémoire longue observée dans l'intensité des fluctuations interdit de traiter le bruit comme un simple bruit blanc homogène. Enfin, la possibilité de dégager une direction dominante par réduction de dimension, par exemple via une composante principale expliquant une part importante de la variance, suggère l'existence d'une variable d'ordre ou au moins d'un sous-espace privilégié.

Ces constats ne sont pas accessoires. Ils constituent la matière première de la modélisation. L'EDA n'est donc pas ce qui vient avant le modèle au sens chronologique seulement. Elle est ce qui produit les exigences auxquelles le modèle devra répondre. Elle fonctionne comme une phase de diagnostic structurel.

Cette fonction diagnostique est au cœur de votre manière de penser la data science. Vous semblez considérer que le travail de l'analyste ne consiste pas d'abord à optimiser une performance prédictive, mais à comprendre de quel type de système il s'agit. Est-on en présence d'une dynamique dans un paysage comportant plusieurs puits de stabilité ? Y a-t-il une dérive effective que l'on peut relier à un potentiel ? Le système présente-t-il des barrières, des transitions, des franchissements de seuil ? Le bruit est-il additif, multiplicatif, ou modulé par une variable lente ? Le phénomène change-t-il d'allure selon l'échelle d'observation ? Répondre à ces questions exige une EDA qui ne soit pas seulement statistique, mais déjà physique au sens large.

C'est ici qu'apparaît l'un des traits les plus originaux de votre approche. Vous ne lisez pas les résultats exploratoires comme de simples résumés numériques. Vous les lisez comme les traces d'un mécanisme. Une densité empirique n'est pas pour vous une courbe parmi d'autres ; elle peut suggérer un paysage de potentiel à travers la relation entre probabilité et structure des états. Une corrélation lente du carré des variations n'est pas simplement un fait descriptif ; elle suggère l'existence d'une dynamique propre de la volatilité. Une structure spectrale en loi de puissance peut signaler une organisation multi-échelle ou un effet de mémoire profonde. Une composante principale dominante peut devenir le

candidat naturel pour une variable d'ordre. En ce sens, votre EDA est déjà un travail de dérivation conceptuelle. Elle prépare l'équation avant que celle-ci ne soit écrite.

Cette manière d'aborder les données conduit à renverser le statut habituel de l'EDA. Dans beaucoup de pratiques standard, on considère que la partie noble de la data science réside dans le choix d'un algorithme, l'entraînement d'un modèle, l'évaluation d'une métrique ou l'optimisation d'une architecture. Dans votre perspective, ces étapes ne prennent sens qu'à condition qu'un travail beaucoup plus fondamental ait déjà été accompli : avoir appris à voir. L'EDA devient alors non seulement la première étape, mais la plus importante. Elle conditionne la qualité de tout ce qui suit.

On peut aller plus loin encore. L'EDA joue aussi un rôle décisif dans la recherche de modèles simples, pédagogiques et interprétables. Cette idée traverse fortement vos réflexions. Vous ne cherchez pas des modèles complexes pour le plaisir de la sophistication. Vous cherchez des formes réduites, à peu de paramètres, dont chacun possède une signification physique ou systémique. Or une telle simplicité ne peut être obtenue de manière légitime que si l'EDA a permis d'identifier les mécanismes vraiment indispensables. Sans EDA, un modèle simple risque d'être arbitraire. Avec une EDA approfondie, la simplicité devient une réduction fondée. On peut alors justifier pourquoi il faut, par exemple, un terme de dérive non linéaire, pourquoi il faut une volatilité à mémoire longue, pourquoi un bruit additif résiduel doit être conservé, et pourquoi certains détails peuvent être laissés de côté à l'échelle choisie.

Autrement dit, l'EDA est ce qui permet la parcimonie intelligente. Elle aide à distinguer les termes essentiels des raffinements superflus. Elle transforme la simplicité en condensation du réel plutôt qu'en simplification abusive.

Cette fonction est particulièrement visible dans la manière dont peut émerger une équation différentielle stochastique réduite pour décrire un système complexe. Une telle équation ne surgit pas comme une hypothèse gratuite. Elle se construit progressivement à partir des indices fournis par l'exploration des données. Si l'on observe des régimes ou des formes stationnaires non triviales, on peut être conduit à introduire une dynamique dans un potentiel. Si l'on observe une mémoire longue dans l'intensité des fluctuations, il devient nécessaire de modéliser une volatilité lente ou corrélée. Si l'on observe des événements extrêmes plus fréquents que prévu, il faut penser un mécanisme de modulation ou de mélange d'échelles. Si une variable dominante émerge, cette variable devient le support naturel de la dynamique réduite. Le passage de l'EDA au modèle ne relève donc pas du saut arbitraire. Il ressemble à une montée en abstraction guidée par les propriétés observées.

Dans cette perspective, l'EDA constitue aussi une protection contre certaines illusions de la data science contemporaine. L'une de ces illusions consiste à croire que les données, traitées par des outils suffisamment puissants, livreront automatiquement la vérité du système. Une autre consiste à confondre prédiction locale et compréhension globale. Une troisième consiste à penser qu'un bon score suffit à valider un modèle. Votre approche suggère au contraire qu'aucune performance algorithmique ne dispense du travail préalable de reconnaissance des formes empiriques fondamentales. L'EDA rappelle qu'un

système n’est pas réductible à la sortie d’un algorithme. Elle oblige à se demander ce qui, dans les données, mérite d’être expliqué et non seulement reproduit.

En cela, elle joue un rôle de vigilance épistémologique. Elle maintient ouverte la question du sens.

Cette vigilance est d’autant plus importante que votre travail ne se limite pas à un seul domaine d’application. Vous avez cherché à établir des ponts entre la finance, la turbulence, le clutter radar, la fatigue, les projets, la cliodynamique et les systèmes complexes adaptatifs. Dans un tel cadre transdisciplinaire, l’EDA devient un langage commun. Elle permet de retrouver, au-delà des objets particuliers, des signatures partagées : mémoire, intermittence, régimes, queues lourdes, structures de dépendance, échelles de temps emboîtées, directions dominantes. C’est parce que l’EDA fait apparaître ces formes communes qu’elle peut servir de passerelle entre des domaines qui, vus de loin, paraissent hétérogènes.

Elle devient ainsi le premier niveau d’une science générale des systèmes complexes. Avant même la théorie unifiée, il y a la reconnaissance exploratoire des invariants de comportement.

On comprend alors pourquoi l’EDA occupe, dans votre réflexion, une place si centrale. Elle n’est pas seulement une technique ; elle est une méthode de découverte. Elle permet de passer de la donnée brute à la structure, de la structure au mécanisme plausible, et du mécanisme plausible à l’équation réduite. Elle est le point de départ de la modélisation, mais aussi la condition de sa légitimité. Elle prépare la réduction de dimension, elle oriente le choix des variables, elle met au jour les contraintes statistiques et dynamiques, elle relie l’empirie à la physique, elle fonde la simplicité interprétable du modèle.

On pourrait donc dire que, dans votre manière de concevoir la data science, l’EDA joue le rôle qu’occupe l’observation dans les grandes sciences de la nature. Elle n’est pas ce que l’on fait avant de penser ; elle est une manière de penser à partir du réel. Elle est le moment où les données cessent d’être un amas de mesures pour devenir l’expression d’une dynamique possible.

Cette formulation permet sans doute de résumer votre position de la manière la plus juste. L’EDA n’est pas un préambule. Elle est le moment fondateur où l’on découvre la forme cachée du système avant d’écrire l’équation qui pourra en rendre compte.

En ce sens, elle est bien la partie primordiale de la data science.