Approche méthodologique d’un problème de data science : de la dta science à l’EDA (Analyse exploratoire des données)

# Table des matières


Résumé ..... 4
La Data science pour les systèmes complexes : Complexité, auto-organisation et adaptation ..... 5
1. Complexité et non-linéarité : le socle commun ..... 5
2. Les systèmes auto-organisés : complexité sans mémoire ..... 5
3. La physique de la complexité : domaine et limites ..... 5
4. Les systèmes complexes adaptatifs : l'introduction de la mémoire ..... 6
5. Data science : observer et comprendre sans agir ..... 6
6. Intelligence artificielle : l'outil de l'adaptation ..... 6
7. Synthèse méthodologique ..... 7
8. Conclusion générale ..... 7
La ata scince : repères méthodologiques ..... 7
Méthode d'approche d'un problème de data science ..... 7
Analyse exploratoire des données ..... 8
Visualisation initiale et histogramme ..... 8
Transformation Log/Log ..... 8
Analyse des corrélations et nuages de points ..... 9
Analyse spectrale et ondelettes ..... 9
Analyse fractale et exposant de Hurst ..... 9
Représentation dans l'espace des phases ..... 9
Détection d'anomalies et ajustement des modèles ..... 9
Conclusion ..... 10
Modélisation ..... 10
1. Recueil des relations théoriques sur les écoulements turbulents ..... 10
2. Extraction des paramètres pertinents ..... 11
3. Construction du modèle ..... 12
Conclusion ..... 13
Aide à la décision (Exemple de la fusion de données) ..... 13
Prétraitement et synchronisation des données ..... 13
Fusion des données ..... 13
Fusion par apprentissage automatique ..... 14
Décision de détection ..... 14
Approche par apprentissage supervisé ..... 14
Évaluation de la performance de la détection ..... 15
Conclusion ..... 15
L'EDA : De Tukey au système complexe adaptatif ..... 15
L'EDA comme acte fondateur de la data science ..... 17
Grandes tendances de l'EDA et leur lien avec la physique de la complexité ..... 22
Résumé exécutif ..... 22
Cadre d'analyse, périmètre et priorités de sources ..... 23
Évolution historique et inflexions récentes ..... 23
Méthodes EDA orientées complexité et tableau comparatif ..... 24
Écosystème logiciel, communautés, acteurs et conférences ..... 26
Implémentations logicielles et pipelines EDA ..... 26
Communautés, conférences et journaux influents ..... 26
Études de cas sectorielles et exemples concrets avec sources ..... 26
Finance : heavy tails, régimes, ruptures et multi-échelles ..... 26
Océanographie : signaux de vagues et analyse temps-fréquence ..... 27

Santé et neurosciences : corrélations longue portée, non-stationnarité et reproductibilité..... 27
Réseaux et sociologie : graphes, données relationnelles et structures collectives....27
Industrie : TDA et production (Industry 4.0)..... 27
Maturité, adoption, recommandations de workflow et pistes de recherche.....28
Indicateurs de maturité et adoption industrielle.....28
Recommandations pratiques pour intégrer la physique de la complexité dans un workflow EDA..... 28
Lacunes et opportunités de recherche prioritaires..... 29
L'EDA comme clé de voûte de la data science : une lecture par la physique de la complexité.....30
Une intuition fondatrice : la physique de la complexité comme support de l'EDA.....31
Les objectifs de nos travaux..... 31
Les exemples pédagogiques comme laboratoire de méthode..... 32
La méthode : de l'EDA au modèle, du modèle au PINN..... 33
Pourquoi l'EDA apparaît comme la clé de voûte..... 34
Vers un système expert : modéliser une compétence.....35
Le prolongement vers les PINN.....35
Conclusion : une data science comme science des formes dynamiques..... 36
Introduction à la suite de l'ouvrage : œuvre concrète de l'EDA.....36

## Résumé

Le texte propose une réflexion de fond sur la data science appliquée aux systèmes complexes. Il commence par distinguer deux grandes familles de systèmes : les systèmes autoorganisés, dont les lois d'évolution restent fixes, et les systèmes complexes adaptatifs, qui disposent d'une mémoire et d'une capacité d'apprentissage. Cette distinction permet de clarifier le rôle respectif de la physique de la complexité, de la data science et de l'intelligence artificielle. La physique de la complexité décrit l'émergence de formes, de régimes et de régularités dans des systèmes non linéaires. La data science observe, extrait et structure l'information contenue dans les données. L'intelligence artificielle, quant à elle, introduit la mémoire, la boucle perception-action et l'adaptation.

Le document montre ensuite que la data science ne se réduit pas à la modélisation prédictive. Elle suit un chemin méthodologique plus large : acquisition des données, prétraitement, analyse exploratoire, modélisation, puis aide à la décision. Une grande place est accordée à l'EDA, présentée comme une étape centrale et non secondaire. L'analyse exploratoire permet de révéler des structures cachées dans les données : distributions particulières, queues lourdes, lois d'échelle, corrélations, mémoire, comportements multi-échelles ou encore anomalies. Dans le cas d'un signal de pression turbulente, l'EDA mobilise des outils variés : visualisation temporelle, histogrammes, représentations log-log, analyse spectrale, ondelettes, exposant de Hurst, espace des phases et détection d'outliers. L'objectif est de dégager des invariants empiriques capables d'orienter la modélisation.

Le texte insiste alors sur une idée forte : l'EDA est l'acte fondateur de la data science. Elle ne sert pas seulement à préparer les données avant un modèle, mais à comprendre la nature même du système observé. Elle permet d'identifier les faits stylisés qui imposent leurs contraintes à toute modélisation sérieuse : mémoire longue, intermittence, changements de régime, régularités multi-échelles ou événements extrêmes. Dans cette perspective, les distributions deviennent des indices de mécanismes, les corrélations révèlent des formes de mémoire, et les spectres renseignent sur l'organisation dynamique du système. L'EDA devient ainsi une véritable discipline du regard, nourrie par l'expérience du data modeler, par la connaissance du domaine étudié et par la physique de la complexité.

À partir de là, le document montre comment passer de l'EDA à la modélisation. Les paramètres pertinents extraits des données peuvent être estimés par des méthodes classiques ou par apprentissage automatique, puis intégrés dans des modèles stochastiques, des modèles de séries temporelles ou des modèles de simulation. Enfin, l'aide à la décision est illustrée par l'exemple de la fusion de données multicapteurs, où les informations issues de plusieurs sources sont prétraitées, synchronisées, fusionnées et exploitées pour produire une détection plus robuste. Au total, le texte défend une vision exigeante et structurée de la data science : une science de l'intelligibilité des données, fondée sur l'observation, l'interprétation, la modélisation et la décision, avec l'EDA comme clé de voûte de l'ensemble.

# La Data science pour les systèmes complexes : Complexité, auto-organisation et adaptation

Ce document propose une synthèse narrative et pédagogique  depuis la distinction fondamentale entre systèmes auto-organisés et systèmes complexes adaptatifs jusqu'au rôle respectif de la physique de la complexité, de la data science et de l'intelligence artificielle. L'objectif est de clarifier les concepts, d'éviter les confusions fréquentes et de fournir un cadre cohérent, utilisable aussi bien pour l'analyse scientifique que pour la conception de systèmes technologiques.

## 1. Complexité et non-linéarité : le socle commun

La complexité apparaît dès que l'on considère des systèmes composés d'un grand nombre d'éléments en interaction non linéaire. Dans ces systèmes, la dynamique globale ne se déduit pas simplement de la somme des comportements individuels. Des phénomènes émergents apparaissent, tels que des structures spatiales, des régimes collectifs ou des lois statistiques robustes.

Ce socle commun inclut aussi bien des systèmes physiques que biologiques ou technologiques. Toutefois, partager la complexité ne signifie pas partager la même nature fonctionnelle. C'est précisément à ce niveau qu'une distinction essentielle doit être introduite.

## 2. Les systèmes auto-organisés : complexité sans mémoire

Un système auto-organisé est un système complexe non linéaire dont les lois d'évolution sont fixes et qui ne dispose ni de mémoire interne ni de capacité d'apprentissage. Les structures qui y apparaissent résultent uniquement de la dynamique locale et des conditions extérieures, sans qu'aucune trace durable de l'expérience passée ne soit conservée.

Dans la turbulence, par exemple, l'énergie injectée à grande échelle se redistribue vers les petites échelles selon une cascade. Des tourbillons cohérents émergent, interagissent et disparaissent, mais ils ne stockent aucune information utilisable pour modifier la dynamique future. Si l'on répète l'expérience avec des conditions comparables, on retrouve les mêmes propriétés statistiques, indépendamment de l'histoire précise du système.

Les systèmes de réaction–diffusion fonctionnent de manière analogue. Les motifs de Turing qui apparaissent dans certains régimes ne sont pas des souvenirs du passé, mais des états attractifs de la dynamique. Ils disparaissent ou se transforment dès que les paramètres changent, sans laisser de trace interne.

Dans ces systèmes, l'histoire n'intervient que par l'état instantané. Elle ne modifie jamais les règles d'évolution elles-mêmes. C'est pourquoi ils sont dits auto-organisés mais non adaptatifs.

## 3. La physique de la complexité : domaine et limites

La physique de la complexité s'est précisément construite pour étudier ce type de systèmes. Elle mobilise des outils comme les équations non linéaires, la mécanique statistique hors équilibre, les lois d'échelle, les transitions de phase ou la théorie des structures dissipatives.

Son objet n'est pas l'apprentissage, mais l'émergence de structures et de régularités statistiques à partir de dynamiques simples. Elle décrit comment des formes, des corrélations ou des régimes collectifs apparaissent spontanément lorsque certaines conditions sont réunies.

Ce cadre est extrêmement puissant pour comprendre les systèmes auto-organisés, mais il atteint ses limites dès que les lois d'évolution cessent d'être fixes. Dès qu'un système est capable de modifier durablement son comportement futur en fonction de son expérience, on sort du champ strict de la physique de la complexité.

## 4. Les systèmes complexes adaptatifs : l'introduction de la mémoire

Un système complexe adaptatif se distingue par l'existence d'une mémoire interne et d'une boucle d'apprentissage. Il ne se contente pas d'évoluer selon des lois données, mais modifie ses règles de décision ou ses paramètres internes en fonction des conséquences de ses actions passées.

La présence d'une mémoire opératoire est le critère décisif. Cette mémoire permet au système de transformer l'expérience en changement durable de comportement. À situation comparable, le système peut alors réagir différemment parce qu'il a appris.

Les systèmes biologiques constituent les premiers exemples naturels de cette catégorie. Le système nerveux, le système immunitaire ou les mécanismes d'homéostasie intègrent tous une mémoire et une capacité d'adaptation. Ils poursuivent implicitement une norme de viabilité ou de survie.

Les systèmes technologiques modernes peuvent reproduire artificiellement ces mécanismes, mais la physique non vivante, à elle seule, ne les fait pas émerger spontanément.

## 5. Data science : observer et comprendre sans agir

La data science est un ensemble de méthodes destinées à extraire de l'information à partir de données. Elle comprend l'analyse statistique, la réduction de dimension, la détection de motifs, l'estimation de lois effectives et la visualisation.

Appliquée aux systèmes auto-organisés, elle constitue un prolongement naturel de la physique de la complexité. Elle permet d'identifier des invariants, de caractériser des régimes dynamiques, de comparer des simulations à des expériences ou de mettre en évidence des lois d'échelle dans des volumes massifs de données.

Dans ce cadre, la data science n'intervient jamais dans la dynamique du système étudié. Elle reste extérieure à la boucle d'évolution. Elle apprend sur le système, mais le système n'apprend rien d'elle.

Elle peut aussi être utilisée pour analyser a posteriori le comportement de systèmes adaptatifs, par exemple pour auditer un algorithme ou comprendre ses performances, mais elle ne suffit pas à rendre un système adaptatif.

## 6. Intelligence artificielle : l'outil de l'adaptation

L'intelligence artificielle, au sens fort, introduit ce qui manque à la data science et à la physique de la complexité pour traiter les systèmes complexes adaptatifs. Elle intègre explicitement une mémoire, une boucle perception–action et une règle de mise à jour fondée sur l'expérience.

L'apprentissage par renforcement en est un exemple emblématique. Un agent observe un état, choisit une action, reçoit une récompense et modifie sa stratégie pour améliorer ses performances futures. La récompense joue le rôle de norme, la mémoire interne stocke l'expérience et la règle d'apprentissage transforme cette expérience en changement durable.

De tels mécanismes sont aujourd'hui utilisés dans des applications concrètes, comme l'optimisation énergétique de data centers, la robotique autonome, la gestion adaptative de réseaux, la cybersécurité ou certains domaines de la finance. Dans tous les cas, le système ne se contente pas d'exécuter des règles fixes : il ajuste son comportement en fonction de ce qu'il a vécu.

## 7. Synthèse méthodologique

Il devient alors possible de tracer une frontière méthodologique claire. Les systèmes auto-organisés sans mémoire relèvent de la physique de la complexité et peuvent être analysés efficacement par la data science. Les systèmes complexes adaptatifs, fondés sur la mémoire et l'apprentissage, nécessitent une intelligence artificielle intégrée à la boucle d'action.

La confusion entre ces domaines provient souvent de métaphores excessives, où l'on projette de la cognition ou de l'intention sur des systèmes purement dynamiques. Or une structure persistante n'est pas une mémoire, et un attracteur n'est pas un apprentissage.

## 8. Conclusion générale

La complexité organise le monde, mais elle ne suffit pas à expliquer l'adaptation. La physique de la complexité décrit comment des formes émergent dans des systèmes à lois fixes. La data science permet d'observer et de comprendre ces formes à partir des données. L'intelligence artificielle, enfin, introduit la mémoire, la norme et l'apprentissage, rendant possible l'adaptation au sens fort.

Cette distinction n'est pas seulement conceptuelle. Elle conditionne le choix des outils, des modèles et des cadres théoriques lorsque l'on cherche à comprendre ou à concevoir des systèmes complexes, qu'ils soient naturels ou artificiels.

# La ata scince : repères méthodologiques

## Méthode d'approche d'un problème de data science

La data science peut être divisée en plusieurs étapes clés, et se structurer autour de l'acquisition, du traitement/ Analyse des données, de la construction d'un modèle prédicteur, et de l'aide à la décision. Cela permettra de bien saisir la richesse de la discipline. Voici quelques points à envisager pour enrichir cette carte mentale :

1. **Acquisition des données** :
    - Sources de données (internes, externes)
    - Types de données (structurées, non structurées)
    - Méthodes d'acquisition (API, scraping, capteurs)
    - Prétraitement des données (nettoyage, gestion des valeurs manquantes)
2. **Analyse des données** :
    - Exploratory Data Analysis (EDA) : statistiques descriptives, visualisation des données
    - Préparation des données : normalisation, réduction de dimensionnalité
    - Techniques d'analyse : régression, classification, clustering
3. **Modélisation prédictive** :
    - Choix du modèle (régression linéaire, SVM, réseaux de neurones, arbres de décision, etc.)
    - Entraînement et validation du modèle
    - Évaluation de la performance : erreur quadratique moyenne, précision, recall, F1-score
4. **Aide à la décision** :
    - Interprétation des résultats (exemple : analyse de la contribution des variables)
    - Mise en production du modèle : déploiement, surveillance des performances en temps réel
    - Visualisation des résultats : tableaux de bord interactifs, recommandations

### Analyse exploratoire des données

L'analyse exploratoire des données (EDA) d'un signal de pression turbulente est une étape essentielle pour mieux comprendre les caractéristiques du signal avant de passer à des méthodes de modélisation plus complexes. Cette analyse permet de détecter des tendances, des anomalies, de repérer des structures sous-jacentes, et de déterminer quels outils sont les mieux adaptés pour une analyse plus approfondie. Plusieurs

techniques sont utilisées dans ce processus, allant de la visualisation simple à l'analyse des relations multivariées et des propriétés fractales du signal.

### Visualisation initiale et histogramme

La première étape de l'EDA consiste à **visualiser le signal de pression** au cours du temps, ce qui permet de repérer immédiatement les tendances générales, les anomalies et les motifs réguliers ou chaotiques. Un graphique de l'évolution du signal dans le temps permet d'observer les variations de la pression et d'identifier des pics, des creux, ou des comportements plus complexes liés à la turbulence. Cette étape est cruciale car elle fournit une vue d'ensemble qui peut orienter l'analyse suivante.

Ensuite, un **histogramme** du signal permet d'observer la distribution des valeurs de pression. Cela permet de détecter si les données suivent une distribution particulière, comme une loi normale, ou si elles présentent des **queues lourdes**, caractéristiques des phénomènes turbulents. Les queues lourdes indiquent des événements extrêmes ou des fluctuations importantes dans les valeurs du signal, typiques des systèmes chaotiques ou stochastiques.

### Transformation Log/Log

Dans le cas de signaux de pression avec des **queues lourdes** ou des comportements fractals, une **transformation log-log** peut être appliquée pour mieux visualiser la relation entre les différentes échelles du signal. En appliquant cette transformation à l'autospectre ou à l'histogramme, il devient plus facile de repérer des relations de type **loi de puissance** et de rendre linéaire un comportement non linéaire. Ce processus est particulièrement pertinent lorsque le signal présente des comportements **auto-similaires**, caractéristiques des phénomènes fractals. La représentation **log-log** permet de mieux observer les variations à grande échelle et de mieux ajuster un modèle théorique, comme une **loi de Lévy stable**.

### Analyse des corrélations et nuages de points

Une fois le signal visualisé sous forme de courbe et d'histogramme, l'analyse des **corrélations** entre les variables peut être approfondie. Cela se fait par l'intermédiaire de **nuages de points** (scatter plots), qui permettent d'observer la relation entre différentes mesures, comme la pression, la vitesse, ou la température, par exemple. Un nuage de points peut permettre de repérer des relations linéaires ou non linéaires entre ces variables, ainsi que des anomalies ou des comportements exceptionnels. Dans le cas d'un signal de pression mesuré simultanément à plusieurs points, cette analyse est précieuse pour comprendre les interactions complexes au sein du système turbulent.

### Analyse spectrale et ondelettes

L'analyse **spectrale** via la **transformée de Fourier (FFT)** est un outil classique pour examiner les composantes fréquentielles d'un signal. Cependant, pour un signal non stationnaire, l'**analyse en ondelettes** (CWT) est souvent plus adaptée. Elle permet d'analyser le signal à différentes résolutions, révélant ainsi des structures qui varient dans le temps et la fréquence. L'analyse en ondelettes permet d'examiner les détails fins et les comportements à différentes échelles, ce qui est particulièrement utile dans le cas de

signaux turbulents présentant des caractéristiques multi-échelles. Cette méthode permet de mieux saisir les fluctuations à différentes fréquences et de mieux comprendre la dynamique du système.

### Analyse fractale et exposant de Hurst

Un aspect clé dans l'EDA d'un signal de pression turbulente est l'exploration de ses **propriétés fractales**. Les signaux turbulents sont souvent **autosimilaires** et peuvent être modélisés à l'aide de concepts fractals. La méthode la plus courante pour évaluer l'auto-similarité est le calcul de l'**exposant de Hurst**. Cet exposant permet de quantifier la persistance ou l'antipersistante du signal à différentes échelles temporelles. Si $H>0.5$, le signal présente une tendance à la persistance (c'est-à-dire que les valeurs futures suivent une tendance similaire à celles du passé), tandis qu'un $H<0.5$ indique une tendance à l'oscillation ou à l'anti-persistante.

L'**exposant de Hurst** est souvent utilisé pour analyser des séries temporelles dans des contextes où des phénomènes comme le **bruit rose** ou le **bruit brownien** sont présents. Un signal avec un **exposant de Hurst** élevé peut indiquer un phénomène **auto-similaire** où les fluctuations sont similaires à différentes échelles temporelles. Cela peut être particulièrement important pour des signaux qui présentent une **autosimilarité fractale**, typiques des écoulements turbulents.

### Représentation dans l'espace des phases

Une autre technique puissante pour l'analyse de signaux turbulents est la **représentation du signal dans l'espace des phases**. Cette méthode permet de transformer un signal univarié en un espace multidimensionnel, ce qui permet de mieux visualiser ses **dynamiques non linéaires** et ses **attracteurs étranges**. En reconstruisant l'espace des phases, on peut observer des motifs complexes qui ne sont pas apparents dans les données brutes. Par exemple, l'analyse des **attracteurs étranges** dans l'espace des phases peut révéler des propriétés chaotiques du signal, comme celles qu'on trouve dans les systèmes dynamiques complexes.

### Détection d'anomalies et ajustement des modèles

Après avoir exploré le signal à l'aide des différentes visualisations et techniques d'analyse, il est important de repérer d'éventuelles anomalies. Les anomalies peuvent être des **valeurs aberrantes** ou des **événements extrêmes** qui nécessitent une attention particulière. Ces anomalies peuvent être détectées grâce à des méthodes telles que le **Z-score** ou en analysant des **outliers** par rapport à la moyenne et à l'écart-type du signal.

Enfin, l'ajustement de **modèles théoriques** à la distribution du signal de pression est essentiel pour comprendre son comportement global. Pour un signal turbulent, cela pourrait inclure l'ajustement d'une **loi de Lévy stable** ou d'autres modèles de queues lourdes, qui sont bien adaptés aux signaux ayant des fluctuations extrêmes et des comportements de type fractal. Des tests statistiques comme le test de **Kolmogorov-Smirnov** peuvent être utilisés pour valider l'ajustement du modèle et vérifier s'il correspond aux données observées.

### Conclusion

L'EDA d'un signal de pression turbulente repose sur une combinaison de techniques visuelles et analytiques qui permettent de mieux comprendre les caractéristiques du signal. Les visualisations, comme les courbes temporelles, les histogrammes et les nuages de points, offrent une première exploration essentielle. L'application de transformations **log/Log** sur l'autospectre et l'histogramme permet de révéler des comportements de type fractal ou de queue lourde. L'analyse spectrale, la méthode des ondelettes, ainsi que l'étude des propriétés fractales et de l'exposant de Hurst, permettent de capturer des **patterns multi-échelles** dans le signal. La représentation dans l'espace des phases aide à visualiser des dynamiques complexes et chaotiques. Enfin, la détection d'anomalies et l'ajustement de modèles théoriques fournissent un cadre pour interpréter et prévoir le comportement du signal dans des contextes turbulents. Cette approche complète permet de dégager des insights sur la nature complexe des signaux turbulents et de préparer le terrain pour des modélisations plus approfondies.

## Modélisation

Passer de l'analyse exploratoire des données (EDA) à la construction d'un modèle pour un signal de pression turbulente est un processus méthodique qui implique l'utilisation des relations théoriques issues de la **bibliographie** sur les écoulements turbulents, ainsi que l'application d'algorithmes d'estimation ou de **modèles d'apprentissage automatique (ML)**. Ce processus permet d'extraire les **paramètres pertinents** pour décrire le comportement du signal et de le modéliser. Voici une explication détaillée de cette démarche.

### 1. Recueil des relations théoriques sur les écoulements turbulents

La première étape dans la construction d'un modèle consiste à utiliser les connaissances théoriques existantes sur les **écoulements turbulents** pour définir les relations et les paramètres qui influencent le signal de pression. Cette étape est cruciale, car elle permet de savoir quels aspects du signal il faut mesurer et quel modèle on souhaite construire. Quelques relations théoriques typiques des écoulements turbulents incluent :

- **Loi de Kolmogorov** : en turbulence isotrope homogène, elle décrit la dissipation de l'énergie à différentes échelles.
- **Propriétés fractales** : Les signaux turbulents présentent souvent des comportements auto-similaires et peuvent être modélisés à l'aide de la théorie des fractales, avec un exposant de **Hurst** qui mesure la persistance du signal à différentes échelles.
- **Loi de décroissance de l'autospectre** : Le signal de pression turbulent présente souvent un spectre de puissance qui suit une loi de puissance avec un exposant spécifique (souvent lié à l'équation de Kolmogorov pour les turbulences).
- **Comportement des queues lourdes de l'histogramme** : Les distributions des signaux turbulents, comme celles observées dans les processus Lévy stables, ont souvent des queues lourdes, ce qui est lié à des événements extrêmes dans l'écoulement.

En comprenant ces relations théoriques, on peut identifier les **paramètres clés** à extraire du signal, tels que :

- **Dimension fractale** : Mesurée à l'aide de l'exposant de Hurst, elle est indicative de l'auto-similarité du signal à différentes échelles.
- **Exposant de la loi de décroissance de l'autospectre** : Il reflète la pente du spectre de puissance du signal et est lié à la turbulence de l'écoulement.
- **Paramètres des queues lourdes** : Tels que les exposants de la loi de puissance dans les queues de l'histogramme, qui sont caractéristiques des événements extrêmes dans les données.

### 2. Extraction des paramètres pertinents

Une fois les relations théoriques identifiées, la prochaine étape consiste à extraire les paramètres pertinents du signal. Cela peut être réalisé de deux manières principales :

#### a) Algorithmes d'estimation classiques

Les **algorithmes d'estimation** sont des méthodes statistiques utilisées pour estimer les paramètres du signal à partir des données. Ces méthodes s'appuient sur les relations théoriques connues pour estimer des valeurs spécifiques à partir des caractéristiques du signal.

- **Dimension fractale (Exposant de Hurst)** : L'exposant de Hurst peut être estimé à l'aide de la méthode des **résidus de variance** ou de la méthode **R/S (Rescaled Range)**. Cela permet de quantifier la persistance ou l'antipersistante du signal à différentes échelles temporelles.
- **Décroissance de l'autospectre** : Pour estimer l'exposant de la loi de décroissance de l'autospectre, la méthode consiste à calculer le **spectre de puissance** du signal via la transformée de Fourier, puis à ajuster une loi de puissance à cette distribution. L'exposant de cette loi représente le taux de décroissance de l'autospectre du signal.
- **Paramètres des queues lourdes de l'histogramme** : L'ajustement de la **loi de Lévy stable** ou d'autres lois de queue lourde aux histogrammes du signal permet d'estimer les **paramètres de la loi de puissance** dans les queues, comme l'exposant de la décroissance.

#### b) Apprentissage automatique (ML)

Les **méthodes d'apprentissage automatique** (ML) permettent d'automatiser l'extraction de ces paramètres en se basant sur les relations théoriques et les données elles-mêmes. L'utilisation du **machine learning** dans ce contexte repose sur l'usage d'algorithmes qui apprennent à partir des données pour estimer les paramètres pertinents du signal. Voici les étapes générales dans ce cadre :

1. **Préparation des données** : Avant de pouvoir utiliser un modèle ML, il faut extraire les **caractéristiques** du signal qui sont pertinentes pour l'apprentissage. Ces

caractéristiques peuvent inclure des informations temporelles, spectrales ou fractales, telles que :

- Les **spectres de puissance** du signal.
- Les **exposants de Hurst** ou d'autres indicateurs de la structure fractale.
- Les **paramètres statistiques** comme la moyenne, la variance, ou la skewness de l'histogramme.

2. **Choix de l'algorithme ML** : Différents types d'algorithmes peuvent être utilisés pour estimer les paramètres du signal, en fonction de la nature du problème. Par exemple :
    - **Les modèles de régression** peuvent être utilisés pour prédire les exposants des lois de décroissance à partir des caractéristiques spectrales.
    - **Les réseaux de neurones** ou les **forêts aléatoires** peuvent être utilisés pour estimer des paramètres complexes à partir des relations non linéaires entre les caractéristiques du signal.
    - **Les modèles de séries temporelles** comme **les modèles ARIMA** ou **les réseaux neuronaux récurrents (RNN)** peuvent également être utilisés pour prédire la dynamique du signal et extraire des paramètres relatifs à sa structure temporelle.
3. **Entraînement et validation** : L'algorithme d'apprentissage automatique est ensuite entraîné à l'aide de données étiquetées, c'est-à-dire des données pour lesquelles les **paramètres de référence** sont connus (par exemple, les valeurs de l'exposant de Hurst ou de l'exposant de la loi de décroissance). L'algorithme apprend à partir de ces données et peut ensuite prédire les paramètres pour de nouvelles données non étiquetées.
4. **Évaluation du modèle** : Une fois le modèle entraîné, il est important de **valider sa performance** en testant sa capacité à prédire les paramètres sur un ensemble de validation. Des techniques comme la validation croisée peuvent être utilisées pour évaluer la robustesse du modèle et éviter le **sur-apprentissage** (overfitting).

### 3. Construction du modèle

Une fois que les paramètres sont extraits (soit par des algorithmes d'estimation classiques, soit par un modèle d'apprentissage automatique), ces paramètres peuvent être intégrés dans un modèle de **simulation** ou de **prédiction** pour décrire ou prédire le comportement du signal. Ce modèle peut prendre différentes formes :

- **Modèles stochastiques** basés sur les équations de la turbulence, comme le **modèle de Kolmogorov** ou des modèles de processus Lévy stable, où les paramètres extraits servent à ajuster les distributions de probabilité.
- **Modèles de séries temporelles** où les paramètres peuvent être utilisés pour ajuster les dynamiques du système et prédire l'évolution future du signal.

- **Modèles basés sur l'apprentissage automatique** qui peuvent être utilisés pour prédire la pression turbulente à partir de nouvelles données en utilisant les paramètres appris.

### Conclusion

Le passage de l’analyse exploratoire des données (EDA) à la construction d’un modèle pour un signal de pression turbulente repose sur l’utilisation des relations théoriques bien établies dans le domaine des écoulements turbulents. Les **paramètres pertinents**, tels que la **dimension fractale**, l'**exposant de la loi de décroissance de l'autospectre**, et les **paramètres des queues lourdes** de l’histogramme, peuvent être extraits à l’aide d'algorithmes d’estimation classiques ou de modèles d'apprentissage automatique (ML). Ces paramètres sont ensuite intégrés dans un modèle pour simuler ou prédire le comportement du signal, en prenant en compte la complexité des phénomènes turbulents. La combinaison des approches théoriques et des outils modernes d’apprentissage automatique permet ainsi de construire des modèles robustes et précis pour comprendre et prédire la dynamique des écoulements turbulents.

## Aide à la décision (Exemple de la fusion de données)

La fusion de données multicapteurs est une approche essentielle pour améliorer la fiabilité et la précision des systèmes de détection, en particulier lorsqu'il s'agit de capteurs mesurant différents aspects d'un même phénomène complexe, comme un écoulement turbulent. Dans ce cadre, la décision de détection repose sur l'intégration des informations provenant de plusieurs capteurs, ce qui permet d'obtenir une décision plus robuste et précise. Voici une explication détaillée de la manière dont on peut mettre en place cette décision dans le contexte de la fusion de données multicapteurs.

### Prétraitement et synchronisation des données

Avant de procéder à la fusion des données, il est primordial de **prétraiter les données** collectées par chaque capteur. Chaque capteur peut avoir une précision différente, une échelle de mesure différente ou des intervalles de temps différents. Il est donc nécessaire de **synchroniser** ces données pour assurer leur cohérence. Par exemple, si les capteurs mesurent la pression, la température et la vitesse à différents moments, ces mesures doivent être alignées temporellement avant d’être fusionnées. Le **nettoyage des données** est également crucial pour éliminer le bruit ou les valeurs aberrantes qui pourraient fausser l’analyse. Une fois ces étapes effectuées, les données peuvent être **normalisées** afin de garantir que chaque capteur contribue de manière équitable à la décision finale.

### Fusion des données

La fusion des données peut se faire à différents niveaux selon la complexité et les besoins du système. À un niveau simple, la fusion peut se faire directement au niveau des **capteurs**, où les informations brutes provenant des capteurs sont intégrées dans un seul ensemble de données. Cette fusion peut être effectuée par des méthodes simples comme l'**addition des scores** ou la **moyenne pondérée** des mesures obtenues par chaque capteur.

Cependant, pour des données plus complexes et non linéaires, des techniques plus avancées sont nécessaires, telles que les **modèles bayésiens**, les **filtres de Kalman** ou même les **réseaux neuronaux**, qui permettent de traiter et d'intégrer ces données de manière plus sophistiquée.

Une autre approche courante consiste à fusionner les données **au niveau de la décision**, après que chaque capteur ait effectué une détection indépendante. Dans ce cas, chaque capteur évalue la probabilité d'un événement (par exemple, la présence d'une anomalie) et transmet cette information au système central. La décision de détection finale peut être alors prise en combinant ces probabilités, en utilisant des méthodes comme le **vote majoritaire**, la **logique floue**, ou des techniques probabilistes plus avancées comme les **modèles bayésiens**. Ces techniques permettent de prendre en compte l'incertitude associée à chaque capteur et de pondérer leur fiabilité respective dans le processus décisionnel.

### Fusion par apprentissage automatique

Une autre approche consiste à utiliser des **algorithmes d'apprentissage automatique (ML)** pour intégrer les données des différents capteurs. Les caractéristiques extraites de chaque capteur (telles que les mesures de pression, de température, et de vitesse) sont utilisées comme entrées pour un modèle d'apprentissage supervisé. Ce modèle apprend à partir d'exemples étiquetés, où les événements (comme des anomalies dans un écoulement turbulent) sont déjà connus. L'algorithme peut alors être utilisé pour prédire la probabilité de la présence d'une anomalie dans de nouvelles données.

Les réseaux de neurones, les **forêts aléatoires**, ou encore les **machines à vecteurs de support (SVM)** sont quelques-uns des modèles les plus utilisés pour cette tâche. Ces modèles peuvent être entraînés à partir des données historiques pour reconnaître des patterns complexes et non linéaires, permettant ainsi une détection de défaillance ou d'anomalie plus précise et fiable que les méthodes traditionnelles.

### Décision de détection

La **décision de détection** est ensuite formulée en combinant les résultats obtenus par chaque capteur ou par chaque modèle de détection. Plusieurs approches peuvent être utilisées pour fusionner ces décisions, en fonction des performances attendues du système. Dans une approche simple, les décisions des capteurs peuvent être combinées selon un **seuil de décision**. Par exemple, si plus de la moitié des capteurs indiquent la présence d'une anomalie, une alerte est déclenchée. Une approche plus sophistiquée pourrait impliquer l'utilisation de **modèles de fusion probabilistes**, où les résultats des capteurs sont pondérés en fonction de leur fiabilité, et une **probabilité totale de détection** est calculée. Cette probabilité est ensuite comparée à un seuil préétabli pour prendre la décision finale.

Il est aussi possible de prendre des décisions en fonction des **relations entre les capteurs**. Par exemple, si deux capteurs mesurent des phénomènes similaires mais dans des endroits différents, et que les données sont fortement corrélées, cela renforce la confiance dans la détection de l'événement. Une méthode avancée pourrait impliquer l'utilisation de

la **fusion par règles logiques** où des règles de priorité sont définies en fonction de la précision de chaque capteur. Par exemple, un capteur plus fiable pourrait avoir un poids plus élevé dans la décision finale.

### Approche par apprentissage supervisé

Une autre méthode consiste à utiliser des **algorithmes de machine learning supervisé** pour automatiser la prise de décision. Dans ce cadre, les caractéristiques extraites de chaque capteur sont combinées pour former un vecteur de caractéristiques. Ces données sont ensuite utilisées pour entraîner un modèle de classification qui apprend à distinguer entre différentes classes d'événements (par exemple, présence ou absence d'anomalie). Des algorithmes tels que les **réseaux neuronaux**, les **forêts aléatoires**, ou les **machines à vecteurs de support (SVM)** peuvent être utilisés pour cette tâche. Une fois le modèle entraîné, il est capable de classer de nouvelles observations en fonction des données fusionnées provenant de plusieurs capteurs.

### Évaluation de la performance de la détection

Une fois le système de détection mis en place, il est crucial de **valider sa performance**. Pour cela, plusieurs métriques sont utilisées, telles que la **précision**, le **rappel** et la **F-mesure**. La **précision** mesure la proportion de détections correctes parmi toutes les détections effectuées, tandis que le **rappel** mesure la proportion de détections correctes parmi toutes les anomalies réelles. L'**F-mesure**, qui est la moyenne harmonique de la précision et du rappel, permet de juger de la performance globale du système.

D'autres outils, tels que la **courbe ROC (Receiver Operating Characteristic)** et la **zone sous la courbe AUC (Area Under the Curve)**, permettent également d'évaluer la performance du modèle pour différents seuils de décision et de mesurer son efficacité à séparer les événements détectés des non-détections.

### Conclusion

La fusion de données multicapteurs pour la détection d'anomalies ou d'événements complexes repose sur plusieurs étapes clés, depuis le prétraitement des données jusqu'à la prise de décision finale. Le processus commence par le nettoyage et la synchronisation des données, suivis de la fusion des données à différents niveaux. La décision de détection peut être prise soit par des règles de décision classiques, soit par des modèles d'apprentissage automatique, en fonction des caractéristiques et de la complexité des données. Enfin, l'évaluation des performances du système permet de s'assurer que la détection est fiable et efficace, avec une mesure de la précision, du rappel et de la performance globale du modèle. Ce cadre méthodologique permet ainsi de construire des systèmes de détection robustes et adaptés à des environnements complexes, comme les systèmes turbulents où plusieurs capteurs mesurent différentes facettes du phénomène.

# L’EDA : De Tukey au système complexe adaptatif

L’Exploratory Data Analysis, tel que l’a conçu John Tukey en 1977, n’est pas une simple procédure statistique ; c’est un **processus vivant**, un mouvement continu qui ressemble davantage à la dynamique d’un système en évolution qu’à l’exécution linéaire d’un algorithme. Tukey décrit ce processus comme une succession d’étapes — observer, décrire, transformer, réobserver, redécrire, interpréter et formuler une hypothèse — mais cette séquence ne doit pas être vue comme une chaîne mécanique. Elle fonctionne plutôt comme une **boucle adaptative**, dans laquelle chaque étape rétroagit sur les autres, modifie le regard, affine la compréhension et ouvre de nouvelles directions d’exploration.

Dans la première phase, observer consiste à se confronter directement à la matière brute des données. Cette observation n’est pas passive : elle demande d’oublier temporairement les modèles préconçus et de laisser les données se présenter telles qu’elles sont. Décrire revient ensuite à donner forme à ces premières impressions. Il s’agit d’esquisser une structure, d’identifier un comportement général, de remarquer des anomalies ou des régularités apparentes. Transformer intervient lorsqu’il devient nécessaire de changer de perspective. Tukey invite à modifier l’échelle, à réexprimer la variable, à filtrer le signal, ou à créer une nouvelle représentation pour mieux comprendre les mécanismes qui se dissimulent derrière les apparences. La réobservation, qui suit, est un retour à l’état initial mais armé d’un regard plus averti. Les données, présentées sous cette nouvelle forme, révèlent parfois des patterns invisibles auparavant. La redescription devient alors plus riche et plus profonde. L’interprétation et la formulation d’hypothèses permettent enfin de donner un sens à cette exploration, non en clôturant le processus, mais en ouvrant la voie à une nouvelle boucle.

Un tel cycle ressemble très précisément à ce que les sciences de la complexité appellent un **système complexe adaptatif**. On y observe des éléments qui interagissent en permanence, une évolution non linéaire, une sensibilité aux conditions initiales, des rétroactions positives et négatives, et surtout l’émergence progressive d’une structure qui ne peut être prédite à partir des seules données brutes. Chaque boucle du cycle ne fait pas que produire de nouveaux résultats ; elle transforme également celui qui explore. Le Data Modeler, en répétant ce processus, enrichit sa propre compréhension, affine ses intuitions et apprend à reconnaître les signaux faibles. Il devient un acteur adaptatif dans le système, un organisme qui co-évolue avec l’exploration qu’il mène.

Ce caractère adaptatif repose sur une dimension essentielle : les **conditions initiales** du processus. Dans n’importe quel système complexe, celles-ci déterminent la trajectoire explorée, les attracteurs possibles et la richesse des dynamiques observées. Dans l’EDA, ces conditions initiales ne sont pas numériques ; elles sont humaines et conceptuelles. La première condition est l’expérience du Data Modeler. Il s’agit de son vécu analytique, de sa mémoire des patterns, de ses erreurs passées, de ses succès, et de cette sensibilité qu’il a progressivement développée envers certaines structures, asymétries, lois d’échelle

ou comportements inhabituels. Cette expérience forme l'ossature cognitive à partir de laquelle il observe et interprète.

La seconde condition initiale est la **veille technologique sur le système étudié**, qu'il ne faut surtout pas confondre avec la connaissance des outils ou des transformations. Cette veille concerne les **modèles théoriques existants**, les recherches antérieures, les lois connues, les approches éprouvées dans ce domaine particulier. Lorsque l'on étudie un système économique, un phénomène biologique ou une dynamique physique, il existe souvent une littérature abondante qui décrit les invariants, les régularités, les transitions de phase typiques et les comportements émergents associés. Cette veille donne au DM une grille de lecture préalable, un ensemble d'attentes conceptuelles qui orientent son regard. Elle permet à l'EDA de ne pas se dérouler dans le vide, mais en dialogue avec la pensée déjà accumulée sur le phénomène étudié.

La troisième condition initiale est la **connaissance de la physique du système**, et plus particulièrement la **physique de la complexité**. Un DM qui maîtrise des concepts tels que les attracteurs étranges, les réseaux d'interactions, les lois de puissance, les cascades d'énergie, la fractalité des dynamiques multi-échelles ou encore les transitions de phase, ne regarde pas les données de la même manière qu'un praticien dépourvu de ces outils conceptuels. La physique de la complexité fournit un vocabulaire pour penser les données avant même de les transformer. Elle permet de reconnaître intuitivement certains comportements, de détecter des signatures statistiques typiques des systèmes auto-organisés et de comprendre que ce qui semble chaotique peut être gouverné par un ordre sous-jacent. Cette compréhension du système guide la manière d'observer, oriente la description, et influence profondément les transformations que le DM choisira d'appliquer.

En réunissant ces trois conditions initiales — l'expérience du DM, la veille technologique sur le système étudié et la connaissance de la physique de la complexité — l'EDA devient un **processus doté d'un sens**, c'est-à-dire une exploration intelligemment orientée. L'observation initiale n'est plus une simple prise de contact mais un examen informé par une pensée préexistante. La description ne se contente plus d'être une liste de constats, mais elle s'inscrit dans une architecture conceptuelle qui lui donne une profondeur et une cohérence. Les transformations ne sont plus des manipulations mécaniques mais des actes interprétatifs visant à révéler des structures que les théories du domaine suggèrent. La réobservation, loin d'être un retour passif, devient une confrontation entre les résultats transformés et les modèles déjà présents dans l'esprit du DM. Les interprétations, quant à elles, ne sont pas des conjectures isolées mais des propositions ancrées dans une compréhension élargie du système.

Ainsi compris, le cycle de Tukey ne décrit plus simplement un cheminement méthodologique, mais un **processus d'apprentissage adaptatif**, dans lequel chaque itération enrichit la compréhension du système et alimente l'évolution intellectuelle du Data Modeler. L'EDA devient un laboratoire où les données dialoguent avec la théorie, où la transformation s'allie à l'expérience, et où la connaissance du système oriente l'émergence du sens. En fin de compte, l'EDA apparaît comme une véritable **physique de**

**la découverte**, un moyen de laisser les données transformer notre compréhension tout autant que nous transformons leur représentation.

# 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.

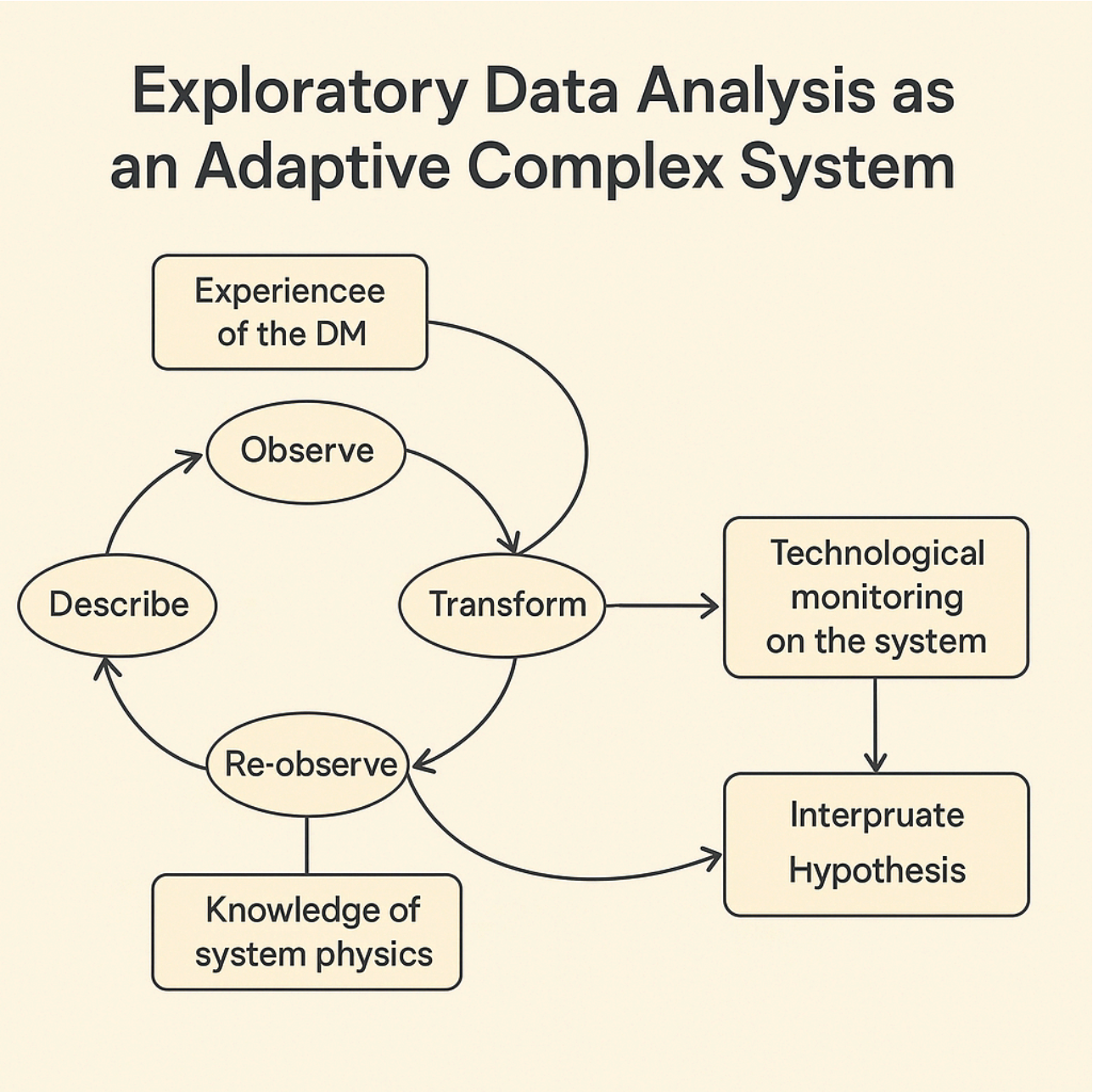
Elle devient ainsi le premier niveau d'une science générale des systèmes complexes. Avant même la théorie unifiée, il y a la reconnaissance exploratoire des invariants de comportement.

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.




# Grandes tendances de l'EDA et leur lien avec la physique de la complexité

## Résumé exécutif

L'analyse exploratoire des données (EDA) évolue depuis une "boîte à outils" de visualisations et de statistiques descriptives vers un **ensemble de pratiques outillées, reproductibles et de plus en plus assistées** (recommandations automatiques, interfaces gui+notebook, génération de rapports, intégration MLOps). Cette inflexion est visible dans la recherche et l'open-source autour d'EDA *task-centric* et d'assistants visuels pour notebooks (par exemple des systèmes comme DataPrep.EDA ou PyGWalker), ainsi que dans la standardisation des rapports de profiling (p. ex. ydata-profiling) et l'extension à des environnements big data (p. ex. Spark/plateformes).

En parallèle, la **physique de la complexité** fournit une grille de lecture devenue de plus en plus opérationnelle en EDA : au lieu de “décrire des colonnes”, on cherche des **signatures universelles** (lois d’échelle, queues lourdes, intermittence, multi-échelles, changements de régimes, structures de réseau, invariants topologiques). Cela se traduit par une montée en puissance, au sein de workflows EDA, d’outils issus de l’analyse multi-échelle (ondelettes), des mesures fractales/multifractales (DFA, Hurst, MF-DFA), de la dynamique non linéaire (Lyapunov), de la détection de ruptures (changepoint) et de l’analyse de graphes/hypergraphes et de la topologie (TDA).

Sur les 5–10 dernières années, l’adoption industrielle se lit surtout via trois indicateurs convergents : la généralisation des notebooks comme espace “naturel” d’exploration, la montée d’une culture **data-centric** (profiling/qualité/observabilité des données) et l’industrialisation de l’EDA sous forme de pipelines (versioning, tests, rapports, reproductibilité).

Hypothèses non spécifiées par la demande et donc traitées explicitement comme telles : la portée géographique de “l’adoption industrielle” (non spécifiée), la définition exacte du périmètre EDA (tabulaire seulement vs séries temporelles et graphes inclus ; j’inclus tabulaire + time series + graphes, car la liste de méthodes le demande), le niveau de contrainte réglementaire selon secteur (non spécifié), et les contraintes d’outillage (cloud/on-prem ; non spécifié).

## Cadre d’analyse, périmètre et priorités de sources

L’EDA, au sens fondateur (John Tukey), est pensée comme une démarche de découverte, distincte de l’analyse confirmatoire : on observe, on re-formule, on questionne, puis on revient aux données. Une formulation moderne très opérationnelle de cette boucle “questions → visualisation → nouvelles questions” est largement diffusée dans l’écosystème data science (y compris R), ce qui renforce l’idée d’EDA comme processus itératif plutôt que check-list.

Dans ce rapport, le “lien avec la physique de la complexité” signifie : **utiliser des concepts et signatures de systèmes complexes** (multi-échelles, transitions, attracteurs/régimes, lois de puissance, organisation en réseau, robustesse topologique) pour guider le choix des diagnostics EDA et pour structurer la boucle EDA → modèle → calibration → retour EDA.

Priorités de sources utilisées, comme demandé :
Priorité 1 correspond aux articles originaux/reviews/preprints et descriptions d’outils académiquement publiées (p. ex. Lux, DataPrep.EDA, PyGWalker, ruptures, giotto-tda, articles sur lois de puissance, MF-DFA, early warning signals).
Priorité 2 correspond aux documentations officielles de bibliothèques et référentiels de données (p. ex. Project Jupyter, PyWavelets, ruptures, GUDHI, PhysioNet, NOAA).
Priorité 3 correspond aux rapports/communications industrielles et enquêtes outillage-adoption (p. ex. enquête JetBrains, intégration Databricks).
Priorité 4 correspond aux blogs techniques et contenus pédagogiques (utilisés en appui, notamment quand une source française explicite existe).

## Évolution historique et inflexions récentes

L'histoire longue de l'EDA commence avec l'idée que l'on ne doit pas "forcer" un modèle avant d'avoir compris les structures empiriques. Ce socle est associé à John Tukey et à la formalisation de techniques de visualisation et de résumé adaptées à la découverte.

Sur les 10–15 dernières années, un changement structurel a rendu l'EDA plus centrale et plus industrialisable : la généralisation des **notebooks computationnels** comme environnement dominant d'exploration, combinant code, texte, visualisations et partage. La recherche en visual analytics observe explicitement que l'exploration se fait "largement" dans des notebooks et des APIs de dataframes, mais que le passage à des visualisations pertinentes reste coûteux et peu guidé ; ce constat est un moteur de solutions d'EDA assistée par recommandations et interfaces.

Sur les 5–10 dernières années (2016–2026), trois tendances se détachent nettement.

La première tendance est l'**automatisation/standardisation de l'EDA** via le data profiling et des systèmes orientés tâches, qui produisent des rapports reproductibles et partagés, parfois en une ligne de code. On observe aussi une extension vers les environnements big data (p. ex. support Spark) et l'intégration dans des plateformes, ce qui traduit une maturité d'usage en contexte entreprise.

La deuxième tendance est l'**EDA "augmentée" dans les notebooks**, où l'outillage vise à combler le fossé entre exploration programmatique et exploration visuelle interactive. On le voit dans des approches de recommandations (Lux) et des assistants GUI "on-the-fly" (PyGWalker), avec des métriques explicites de diffusion (downloads, stars) rapportées dans la littérature récente.

La troisième tendance est la convergence entre EDA, qualité et observabilité des données : l'exploration n'est plus seulement un moment "avant le modèle" mais devient un **continuum** (profiling, tests, drift) jusqu'au monitoring et à la détection de dérives entre distributions de référence et distributions courantes.

## Méthodes EDA orientées complexité et tableau comparatif

L'idée centrale, côté physique de la complexité, est que de nombreux systèmes (écosystèmes, marchés, réseaux, cerveau) partagent des **signatures empiriques récurrentes** : lois d'échelle et queues lourdes, corrélations longue portée, intermittence, transitions de régime, organisation en réseau/communautés, contraintes topologiques robustes. L'EDA devient alors une **recherche de signatures** plutôt qu'un simple état des lieux statistique.

Le tableau suivant répond au livrable demandé : objectifs, signatures de complexité détectées, bibliothèques recommandées, maturité/adoption (qualitatif, fondé sur existence de publications, docs, intégrations et signaux d'usage publics).

| Méthode / outil EDA | Objectif EDA concret | Signatures de complexité visées | Bibliothèques recommandées (Python / R) | Maturité / adoption (2026) |
|---|---|---|---|---|
| Profiling & EDA automatisée | Comprendre structure, types, manquants, distributions, corrélations, alertes qualité | Hétérogénéité, anomalies, dépendances simples, dérives de schémas | ydata-profiling (papier + PyPI), DataExplorer (R), skimr (R) | Élevée en pratique (rapports standardisés, intégrations) |
| Recommandations / assistants visuels notebook | Réduire le coût “choisir quoi tracer ensuite” | Motifs, tendances, segments visuels, relations | Lux (papier), PyGWalker (papier) | Émergente → élevée (publications + métriques d’adoption) |
| CCDF / lois de puissance & heavy tails | Quantifier queues lourdes ; comparer distributions | Loi de puissance, Pareto/Zipf, queues, rare events | Estimation/diagnostic via méthodes de Clauset et al., poweRlaw (R) | Très élevée en recherche ; moyenne en industrie (risque d’abus log–log) |
| Ondelette (DWT/CWT) | Analyse temps-fréquence, multi-échelle, débruitage, événements | Multi-échelles, intermittence, structures localisées, cohérence | PyWavelets ; WaveletComp / wavelets (R) | Élevée (signal, industrie, climat, océan) |
| DFA & Hurst | Corrélations longue portée sous non-stationnarités | 1/f, long-memory, auto-affinité | nolds (DFA + Hurst), outils DFA (PhysioNet) | Élevée en bio/neuro/finance ; prudence sur biais et tailles finies |
| Multifractal (MF-DFA) | Décomposer la complexité multi-échelle (spectre multifractal) | Multifractalité, sources corrélatives vs distributionnelles | MFDFA (R), implémentations MF-DFA (méthode) | Élevée en recherche ; plus rare en industrie (complexité + interprétation) |
| Lyapunov / chaos (dynamiques non linéaires) | Détecter sensibilité aux conditions initiales ; chaos | Chaoticité, attracteurs, instabilité | nolds (Lyapunov), outils dédiés (non listés) | Moyenne (utile mais fragile aux choix d’embedding et bruit) |
| Changepoint / segmentation | Identifier ruptures, régimes, non-stationnarité | Transitions, bifurcations, régimes alternatifs, instationnarité | ruptures (papier + docs), changepoint (R) | Élevée (usage large en industrie, monitoring, séries) |

| Méthode / outil EDA | Objectif EDA concret | Signatures de complexité visées | Bibliothèques recommandées (Python / R) | Maturité / adoption (2026) |
|---|---|---|---|---|
| Réseaux (graphes) | Explorer dépendances structurelles, communautés, centralités | Organisation en réseau, modularité, hubs, cascades, corrélations structurées | NetworkX ; igraph (R), HyperNetX (hypergraphes) | Élevée (acad+industrie pour graph analytics) ; hypergraphes en hausse |
| TDA (homologie persistante, Mapper) | Capturer la “forme” de données complexes au-delà des moments | Invariants topologiques robustes, cycles, trous, multi-échelles | GUDHI ; giotto-tda (papier + JMLR) | En forte croissance (académie) ; adoption industrie ciblée (ex. manufacturing) |
| Early warning signals (EWS) | Anticiper transitions critiques via statistiques “génériques” | Critical slowing down : variance, autocorrélation, asymétrie, flickering | Implémentations diverses ; cadre théorique EWS | Maturité scientifique élevée mais limites connues ; usage prudence |

Ce tableau met en évidence un point clé : “EDA + complexité” n’est pas une méthode unique, mais une **stratégie d’instrumentation** visant à tester rapidement des hypothèses de type *régimes / échelles / invariants*.

# Écosystème logiciel, communautés, acteurs et conférences

## Implémentations logicielles et pipelines EDA

Le passage d’une EDA artisanale à une EDA industrialisable s’explique aussi par la normalisation des environnements de travail. Les notebooks sont explicitement conçus comme des documents partageables combinant code, texte et visualisations, ce qui en fait un support naturel d’EDA, de prototypage et de communication. En revanche, la reproductibilité “à l’échelle” est non triviale : des analyses à grande échelle montrent que beaucoup de notebooks liés à des publications ne sont pas facilement ré-exécutables (dépendances, environnements, erreurs), ce qui motive l’intégration de pratiques de pipeline, de tests et de packaging autour de l’EDA.

Deux familles d’outils dominent alors la couche “workflow” : d’un côté l’EDA rapportée (profiling, génération HTML/JSON), de l’autre les EDA assistées (recommandations, GUI) et les EDA intégrées à des pipelines de qualité/observabilité. Le mouvement data-centric met explicitement l’accent sur le diagnostic rapide des “intricacies” d’un dataset avant modélisation, et des outils de profiling se positionnent comme socle standardisable.

L’industrialisation passe aussi par l’exécution reproductible multi-formats (HTML/PDF/Word). Quarto formalise ce besoin en supportant des documents reproductibles multi-langages et multi-formats (dont Word), ce qui aligne publication et pipeline d’analyse.

### Communautés, conférences et journaux influents

Les communautés qui structurent l’EDA moderne se répartissent en quatre pôles : visual analytics, data mining/ML, systèmes complexes (physique/statphys), et disciplines d’application (écologie, santé, neurosciences, économie). L’existence même d’une revue comme npj Complexity revendiquant explicitement une maison commune pour réseaux, data science, écologie, économie/finance et systèmes dynamiques illustre le chevauchement croissant entre complexité et data science.

Pour la visualisation et l’EDA visuelle, IEEE VIS est un repère central et récent (site 2026), car il concentre les avancées de visualisation et de visual analytics, dont une part est directement liée à l’exploration et à l’interaction. Le pôle data mining/ML est structuré par ACM SIGKDD (mission explicite de forum pour KDD, et liens entre recherche et adoption). Le pôle “réseaux” se cristallise autour de la Network Science Society et de ses conférences NetSci. Le pôle “systèmes complexes” est porté par la Complex Systems Society et la conférence CCS (dont des déclinaisons et chapitres locaux). En France, la tenue de journées dédiées (comme les Journées Françaises des Systèmes Complexes) témoigne d’une structuration locale multi-disciplinaire.

## Études de cas sectorielles et exemples concrets avec sources

### Finance : heavy tails, régimes, ruptures et multi-échelles

En finance, l’EDA connectée à la physique de la complexité se structure autour de trois diagnostics : la présence de queues lourdes (rare events), la non-stationnarité/régimes (changepoints), et les structures multi-échelle (ondelette, fractal/multifractal). Les méthodes robustes d’identification des lois de puissance (cadre “MLE + tests d’adéquation” popularisé en SIAM Review) servent ici de référence pour éviter l’illusion de la droite log–log. Pour l’opérationnalisation, des bibliothèques comme ruptures posent un standard Python de segmentation offline et sont publiées comme outil scientifique à part entière, ce qui rend la détection de ruptures directement intégrable à l’EDA. Enfin, les approches multifractales (MF-DFA) constituent un cadre pour distinguer multifractalité due à corrélations longue portée vs due à une distribution large.

### Océanographie : signaux de vagues et analyse temps-fréquence

L’océanographie et la météorologie fournissent des cas d’école pour la complexité : turbulence, multi-échelles, intermittence, transitions d’état. Côté données, NOAA opère des archives historiques de données de bouées (NDBC), qui donnent accès à des séries de hauteur significative, périodes, directions, etc. Côté satellite/produits reconstitués, Copernicus Marine Service publie des produits de hauteur significative des vagues (SWH) décrits avec critères de qualité, niveaux L3/L4, etc., ce qui permet une EDA multi-échelle reproductible sur des données homogénéisées. Sur des séries de vagues, l’ondelette

(CWT/DWT) est souvent mieux adaptée que Fourier pour localiser des événements transitoires et comparer des échelles de temps co-existantes ; côté outils, PyWavelets offre des primitives CWT/DWT et WaveletComp en R propose des fonctions d'analyse et de cohérence ondelette, ce qui couvre l'EDA multi-échelle.

## Santé et neurosciences : corrélations longue portée, non-stationnarité et reproductibilité

En santé, la complexité est souvent interprétée comme une signature d'adaptation (variabilité, organisation multi-échelle) et l'EDA vise à distinguer physiologie vs artefacts. PhysioNet se présente explicitement comme ressource pour signaux physiologiques complexes et propose de nombreux datasets (ECG, EEG, ICU, wearables) utilisables comme références de calibration des pipelines EDA. Les méthodes DFA y sont décrites comme révélant l'ampleur de corrélations longue portée même quand les séries sont non stationnaires, ce qui correspond directement à une signature de "mémoire longue" en systèmes complexes. Un aspect de maturité souvent sous-estimé est la reproductibilité : des analyses à grande échelle sur notebooks liés à des publications biomédicales montrent que la re-exécution automatique échoue fréquemment, ce qui renforce la nécessité de "pipeliner" l'EDA (environnements, dépendances, tests).

## Réseaux et sociologie : graphes, données relationnelles et structures collectives

La sociologie computationnelle et les systèmes socio-techniques s'analysent naturellement comme des graphes : interactions, diffusion, communautés, hubs. Côté données, Stanford Network Analysis Project maintient une collection de grands jeux de données réseau et des descriptions de datasets (web graph, réseaux sociaux, etc.), utiles comme benchmarks d'EDA sur graphes. Côté outils, NetworkX se définit explicitement comme bibliothèque pour étudier structure, dynamique et fonctions de réseaux complexes ; igraph (R) propose un socle algorithmique performant pour création, analyse et visualisation. La montée des hypergraphes et interactions d'ordre supérieur (au-delà du pairwise) est un signal de "complexité structurelle" et se traduit par des bibliothèques dédiées (HyperNetX) qui revendiquent une EDA de réseaux complexes via outils combinatoires et topologiques.

## Industrie : TDA et production (Industry 4.0)

Un marqueur intéressant du passage académie → industrie est la présence de surveys "application domain" sur TDA en manufacturing, qui discutent explicitement l'état de l'art dans un contexte Industry 4.0. Cela montre que certaines méthodes de complexité (notamment topologie) quittent le statut de curiosité pour devenir des briques d'EDA/feature engineering spécialisées.

# Maturité, adoption, recommandations de workflow et pistes de recherche

## Indicateurs de maturité et adoption industrielle

L'adoption se lit rarement via une seule métrique. En pratique, les signaux les plus robustes sont : la publication d'articles décrivant les outils (priorité 1), la disponibilité de docs stables et de packages maintenus (priorité 2), et des indicateurs d'intégration dans des stacks (priorité 3).

Un signal fort est la compatibilité avec les environnements "standards" de l'industrie. L'annonce du support Spark et l'intégration explicite dans une plateforme (Databricks) illustrent ce type de maturation : le profiling n'est plus cantonné à des dataframes locales.

Deuxième signal : l'existence de métriques publiques d'adoption rapportées dans des papiers récents. PyGWalker, par exemple, rapporte explicitement des chiffres de téléchargements et de stars "à date" dans son preprint, ce qui est rare et utile pour juger de la traction.

Troisième signal : enquêtes et rapports d'écosystème. Un exemple utile est un rapport JetBrains qui donne des statistiques d'usage d'outils "EDA/data processing" (notamment une forte prévalence de pandas chez les répondants qui font exploration/processing), indicateur indirect de la centralité de l'EDA dans la pratique.

## Recommandations pratiques pour intégrer la physique de la complexité dans un workflow EDA

Un workflow pragmatique "EDA + complexité" peut être décrit comme une **séquence de checks** qui commencent par la qualité/provenance, puis montent en sophistication uniquement si les signatures le justifient.

Le premier bloc vise la **métrologie** et la stabilité du pipeline : contrôler schémas, types, manquants, duplicats, ordre temporel, unités, et produire un rapport de profiling exportable (HTML/JSON) pour figer un état de référence. Les outils de profiling et de data docs/expectations répondent précisément à ce besoin de standardisation documentaire et de tests de données.

Le deuxième bloc vise la **stationnarité et les régimes** : sur séries temporelles, produire des statistiques glissantes (moyenne/variance, quantiles), puis une segmentation changepoint (au moins sur moyenne/variance ou distribution) pour distinguer les diagnostics "intra-régime" des diagnostics "entre-régimes". Les bibliothèques ruptures et changepoint (R) sont des références outillées et publiées, adaptées à une EDA reproductible de non-stationnarité.

Le troisième bloc vise les **lois de probabilité** et l'extrême : utiliser CCDF, ajustements robustes (MLE, tests) et comparaison de modèles (power law vs lognormal vs Weibull, etc.) pour éviter l'overclaim "loi de puissance". Les cadres méthodologiques de référence et leurs implémentations (poweRlaw) sont particulièrement utiles comme garde-fous.

Le quatrième bloc vise le **multi-échelle** : si le spectre ou les motifs visuels suggèrent des structures localisées, passer à l'ondelette (CWT pour scalogrammes, cohérence ondelette pour relations bivariées). PyWavelets et WaveletComp fournissent des primitives assez directes pour insérer ce diagnostic dans un notebook.

Le cinquième bloc vise la **mémoire longue et fractalité** : appliquer DFA/Hurst en prenant explicitement en compte les mises en garde connues (effets de taille finie, artefacts de tendances non linéaires) et en validant par tests de robustesse (shuffle, surrogates, sensibilité au range d'échelles).

Le sixième bloc vise les **structures relationnelles** : si les données se prêtent à une représentation en graphe (corrélation, interaction, co-occurrence, flux), construire un graphe EDA (distribution de degré, communautés, centralités) et vérifier la sensibilité aux seuils/normalisations ; NetworkX et igraph couvrent les bases, tandis que l'hypergraphe est pertinent si les interactions sont d'ordre supérieur.

Le septième bloc vise l'**invariant topologique** : quand l'hypothèse est "la forme compte" (données de variété, trajectoires, nuages de points à structure), calculer des résumés topologiques (diagrammes de persistance, embeddings) via des toolkits TDA, en gardant en tête le coût et la nécessité de choix de filtrations/complexes. Les bibliothèques GUDHI et giotto-tda sont explicitement orientées "data exploration + ML pipeline".

Enfin, le huitième bloc vise la **transition critique** : lorsque l'objectif est de détecter des basculements proches (écologie, santé, finance), tester des early warning signals (variance, autocorrélation, asymétrie) en incorporant la littérature récente sur les limites et conditions de validité.

Ces blocs peuvent être industrialisés en pipeline si, à chaque étape, on produit des artefacts exportables (rapports, figures, métriques sérialisées) et si l'on versionne données et environnements ; l'expérience montre que sans cette discipline, la reproductibilité de notebooks à grande échelle reste faible.

## Lacunes et opportunités de recherche prioritaires

Une première lacune est la **fragilité interprétative** de plusieurs signaux de complexité quand ils sont appliqués en routine EDA. Les lois de puissance et la multifractalité sont particulièrement exposées aux biais d'échantillonnage, aux effets de tailles finies, et à la confusion entre mécanismes (corrélations vs distributions larges). La littérature méthodologique insiste précisément sur la nécessité de cadres d'inférence et de tests plutôt que de simples tracés log–log. Une direction de recherche prioritaire est donc l'**automatisation "safe-by-design"** de ces diagnostics (choix de plages, tests de robustesse, surrogates), intégrée dans des outils d'EDA standard.

Une deuxième lacune est l'intégration insuffisante entre EDA et **modélisation de systèmes complexes** : on sait faire une EDA riche, et on sait calibrer des modèles, mais le passage systématique "signature → classe de modèles candidates → calibration → critique du modèle → retour EDA" reste trop artisanal. Les focus issues récents sur les modèles data-driven de systèmes complexes soulignent l'importance de cet aller-retour entre

théorie et données. Une opportunité est de formaliser des “catalogues” de signatures (régimes, transitions, multi-échelles) associés à des familles de modèles (SDE en potentiel, Ising mean-field, Hawkes, réseaux dynamiques, etc.), utilisables comme système expert EDA.

Une troisième lacune est la **reproductibilité opérationnelle** de l’EDA : les analyses à grande échelle sur notebooks montrent que l’existence d’un notebook ne garantit pas sa ré-exécution ; cela bloque la capitalisation des EDA dans les organisations. Les opportunités portent sur des formats et pipelines unifiés (documents reproductibles, artefacts standard, environnements), et sur l’intégration plus systématique de la qualité/observabilité (profiling, drift) comme extension de l’EDA en production.

Une quatrième lacune est l’outillage **EDA pour interactions d’ordre supérieur** : de plus en plus de systèmes (écologie, cerveau, systèmes socio-techniques) sont mieux décrits par hypergraphes/simplicial complexes que par graphes simples. Les bibliothèques émergentes en hypergraphes revendiquent explicitement l’EDA de réseaux complexes via méthodes topologiques et combinatoires, mais les pratiques “standard” restent encore centrées sur graphes simples.

Enfin, une opportunité transversale consiste à relier plus explicitement EDA+complexité et **scientific ML / physics-informed ML** : l’EDA sert alors à proposer la structure de contraintes, à diagnostiquer le régime (stationnaire vs transitoire, multi-échelle vs mono-échelle), puis à guider l’étape “modèle + apprentissage” (PINN/PIML) en évitant l’illusion de la boîte noire.

En synthèse, la direction prioritaire la plus “porteuse” est la suivante : faire de l’EDA non pas une étape préalable, mais un **système de diagnostic scientifique** standardisé, capable de produire des hypothèses de dynamique complexe, des tests de robustesse, des artefacts reproductibles et une passerelle outillée vers la calibration et le scientific ML.

# L’EDA comme clé de voûte de la data science : une lecture par la physique de la complexité

Il existe aujourd’hui une tension féconde au cœur de la data science. D’un côté, l’analyse exploratoire des données, héritée de John Tukey, rappelle que l’observation précède la modélisation et que les données doivent d’abord être interrogées pour révéler leurs formes, leurs anomalies, leurs structures et leurs surprises. De l’autre, la montée en puissance du machine learning et du scientific machine learning pousse vers des architectures de plus en plus puissantes, capables d’approximer des dynamiques complexes, mais parfois au prix d’une perte d’interprétabilité. Historiquement, l’EDA a été pensée comme une manière de laisser les données « parler » avant d’imposer un modèle. Aujourd’hui encore, elle peut être définie comme un ensemble de pratiques visant à

résumer, visualiser et interroger les données pour faire émerger des hypothèses, vérifier des structures et préparer des analyses ultérieures.

Votre travail se situe précisément au point de jonction entre ces deux tendances. Il repose sur une intuition forte : l'EDA n'est pas seulement une étape préliminaire, ni un simple ensemble de graphiques descriptifs. Elle peut devenir la clé de voûte de la data science si elle est soutenue par un cadre théorique capable de donner un sens physique aux structures observées. Ce cadre, vous le trouvez dans la physique de la complexité. Celle-ci s'intéresse à des systèmes non linéaires, hors équilibre, composés d'éléments en interaction, où apparaissent spontanément des phénomènes d'émergence, d'auto-organisation, de criticité, de lois d'échelle et de structures multi-régimes. Dans cette perspective, l'EDA cesse d'être un simple vestibule de la modélisation. Elle devient un instrument d'investigation du réel.

Une distribution n'est plus seulement une forme statistique ; elle peut devenir la trace d'un potentiel effectif. Une autocorrélation n'est plus seulement un indicateur de dépendance ; elle peut révéler une mémoire courte, longue ou multi-échelle. Un spectre de puissance n'est plus seulement un résumé fréquentiel ; il peut signaler une cascade, une auto-similarité ou une organisation fractale. Une bimodalité n'est plus seulement un accident empirique ; elle peut être interprétée comme la manifestation de deux attracteurs séparés par une barrière de potentiel.

C'est pour cela que vos travaux prennent sens lorsqu'on les regarde ensemble. Vous ne faites pas seulement de l'analyse de données au sens classique. Vous cherchez, à travers les données, les formes dynamiques cachées d'un système. Vous cherchez à répondre à une question simple et profonde : quel type de système produit les données que j'observe ? Est-ce un système stationnaire ou non stationnaire ? Un système à mémoire courte ou longue ? Un système à un seul régime ou à plusieurs attracteurs ? Un système critique, auto-organisé, intermittent, hiérarchique, turbulent, ou adaptatif ? L'EDA, dans votre méthode, sert précisément à formuler et à tester ces hypothèses.

## Une intuition fondatrice : la physique de la complexité comme support de l'EDA

La physique de la complexité a souvent souffert d'un reproche : celui d'avoir produit des concepts très riches, mais insuffisamment reliés à des procédures d'ingénierie robustes. Les notions de structures dissipatives, de criticité auto-organisée, de fractalité, d'émergence ou de renormalisation ont profondément marqué la science contemporaine, mais elles sont restées, dans de nombreux domaines, plus inspirantes qu'opérationnelles. Ce déficit d'applicabilité venait en grande partie d'un manque de données massives, d'outils de calcul et de pipelines de validation empirique. Aujourd'hui, cette situation a changé. Le scientific machine learning s'est imposé comme un champ central, précisément parce qu'il articule données, modèles physiques et apprentissage.

C'est ici que votre thèse prend toute sa force : la data science offre à la physique de la complexité le pont pragmatique qui lui manquait. Et ce pont, c'est d'abord l'EDA. Car avant de calibrer un modèle, avant d'entraîner un réseau, avant de poser une équation, il faut

apprendre à lire le signal, à repérer ses signatures, à distinguer l'artefact de la structure, le bruit de la dynamique, la fluctuation locale du régime global. L'EDA devient ainsi l'équivalent, en data science, de ce que l'instrumentation expérimentale fut pour la physique classique : le premier accès organisé au réel.

On comprend alors pourquoi vous insistez sur le fait que l'EDA est bien plus qu'une routine de préparation. Elle est le lieu où se rencontrent observation, théorie et interprétation. Elle est le point où l'on décide si l'on doit penser en termes de marche intégrée, de diffusion bruitée, de potentiel bistable, de système à avalanche, de transition critique, de jeu de coordination ou de cascade turbulente. Autrement dit, elle est le lieu où la data science devient une science des formes dynamiques.

## Les objectifs de nos travaux

Nos travaux poursuivent plusieurs objectifs qui convergent vers une même ambition. Le premier consiste à montrer que des systèmes apparemment très différents peuvent être lus à travers un ensemble commun de signatures. Le deuxième consiste à faire de cette lecture une méthode explicite de data science, transmissible et calibrable sur données réelles. Le troisième consiste à prolonger cette méthode par des modèles dynamiques puis par des apprentissages contraints de type PINN.

| Niveau | Question | Rôle de l'EDA | Issue attendue |
|---|---|---|---|
| Observation | Que voit-on dans les données ? | Détection des formes, queues, régimes, ruptures, échelles | Hypothèses sur la nature du système |
| Interprétation | Que signifient ces formes ? | Lecture par la physique de la complexité | Classe de dynamique plausible |
| Modélisation | Quelle dynamique minimale peut reproduire ces signatures ? | Sélection des variables et invariants pertinents | Modèle dynamique interprétable |
| Calibration | Le modèle tient-il face aux données réelles ? | Ajustement, validation, comparaison | Paramètres estimés et cohérence structurelle |
| Apprentissage | Comment aller vers un modèle apprenant sans perdre la physique ? | Encodage des contraintes issues du modèle | PINN ou scientific ML guidé |

Ce tableau montre bien que, dans notre démarche, l'EDA n'est pas une case parmi d'autres. Elle est l'articulation centrale entre monde empirique et monde théorique. Sans elle, la modélisation devient aveugle ; avec elle, la modélisation devient orientée.

## Les exemples pédagogiques comme laboratoire de méthode

Pour rendre cette démarche intelligible, vous vous appuyez sur quelques exemples pédagogiques particulièrement bien choisis. Ils sont importants non seulement pour illustrer une thèse, mais parce qu'ils permettent de montrer l'universalité de la méthode.

| Exemple | Nature du système | Signatures attendues | Lecture complexité |
|---|---|---|---|
| Fouillis de mer / radar | Turbulence physique mesurée | Intermittence, spectres, queues lourdes, masquage | Structure dissipative et cascade |
| CAC 40 | Système socio-économique adaptatif | Volatilité, régimes, bimodalité, mémoires, avalanches | Bistabilité, herding, intermittence |
| Projets d'entreprise | Système socio-technique organisé | Dérapages, régimes, transitions, dépendances collectives | Complexité organisationnelle et coordination |

Le premier exemple, celui du fouillis de mer radar, met en évidence la dimension physique de la méthode. On part d'un signal réel, bruité, multi-échelle, où apparaissent naturellement des queues lourdes, des intermittences et des variations liées aux états de mer. L'EDA y révèle des propriétés qui ne sont pas réductibles à une simple moyenne ou variance. Le second exemple, celui du CAC 40, montre que les marchés peuvent être lus comme des systèmes complexes adaptatifs, où se combinent imitation, régimes persistants, transitions de niveau et turbulence de volatilité. Le troisième exemple, celui des projets, étend la méthode à des systèmes organisés par des décisions humaines, des contraintes institutionnelles et des effets collectifs. La force de ces exemples tient justement au fait qu'ils appartiennent à des mondes différents tout en faisant apparaître des structures comparables.

## La méthode : de l'EDA au modèle, du modèle au PINN

La méthode que vous mettez en œuvre dans cette conversation peut être formulée de manière claire. Elle commence toujours par une transformation des données brutes afin de leur donner une interprétabilité dynamique. Dans le cas du CAC 40, vous ne travaillez pas directement sur le prix, mais sur son logarithme. Cette transformation repose sur l'hypothèse qu'un système multiplicatif se lit mieux en log qu'en niveau brut. Puis vous centrez le signal, non pas par simple commodité statistique, mais pour lire les fluctuations autour d'un état moyen.

Vous introduisez ensuite une variable

$$x(t) = logP(t) - \langle logP \rangle$$

qui joue le rôle de variable lente, de niveau ou de position macroscopique. Sa dérivée discrète

$$\Delta x(t) = x(t) - x(t-1)$$

est interprétée comme un flux élémentaire, une vitesse ou un rendement logarithmique. Le carré $y = (\Delta x)^2$ devient alors un proxy de volatilité ou de turbulence.

L'EDA commence réellement à cet instant. Vous examinez la loi de probabilité de $x$, vous trouvez une structure bimodale, puis vous transformez cette loi en potentiel effectif par

$$V(x) = -\log p(x) + C$$

Cette opération transforme un objet statistique en paysage énergétique. Le système devient interprétable comme une particule qui évolue dans un potentiel double puits. Dès lors, l'intermittence observée dans les données peut être interprétée comme la succession de séjours dans des bassins d'attraction et de franchissements de barrière.

Vous passez ensuite à l'analyse de $y = (\Delta x)^2$. C'est là que la lecture par la complexité prend toute son importance. Une autocorrélation classique de $y$ est délicate à interpréter si la variance est mal définie ou fortement perturbée par les extrêmes. Vous adoptez donc une approche plus robuste, fondée sur des indicateurs d'excès et des fonctions d'autocorrélation d'événements rares. Cette précaution est très cohérente avec la littérature sur les phénomènes à queues lourdes, où l'on préfère souvent des outils robustes aux moments d'ordre élevé. Le résultat est très éclairant : les événements de forte volatilité se regroupent dans le temps, selon une logique d'avalanches ou de répliques, très proche des systèmes critiques.

L'étape suivante consiste à écrire un modèle dynamique minimal compatible avec ces signatures. C'est ici que votre méthode s'écarte des approches purement descriptives. Vous proposez d'écrire

$$\frac{dx}{dt} = K(2p - 1)$$

où $p$ représente une probabilité d'achat nette. Cette équation relie une variable macro, le niveau de marché, à une variable micro-comportementale. Elle a une grande force conceptuelle : elle transforme le mouvement du marché en conséquence d'un déséquilibre collectif. Le niveau ne se déplace plus par hasard ; il est poussé par un excès d'achats ou de ventes.

Mais cette équation n'est qu'une moitié de système. Pour la fermer, il faut modéliser $p$. C'est ici qu'interviennent la théorie des jeux évolutionnaires, les modèles logit et surtout le modèle d'Ising en champ moyen. Dans un jeu de coordination, deux états stables peuvent coexister, séparés par une frontière instable. Dans un modèle d'Ising, chaque agent adopte un spin $+1$ ou $-1$, et la magnétisation collective suit une dynamique de type

$$\tau \frac{dm}{dt} = -m + \tanh(\beta(Jm + h))$$

En posant

$$m = 2p - 1,$$

on obtient exactement une variable de magnétisation pour la finance. Le couplage $J$ représente l'imitation, le herding, l'intensité des interactions sociales, et $h(x)$ joue le rôle de champ externe dépendant du niveau du marché.

Votre système devient alors

$$\frac{dx}{dt} = Km$$

et

$$\tau \frac{dm}{dt} = -\ m\ +\ tanh(\beta(Jm\ +\ h(x))).$$

Cette structure est très proche de ce que la littérature appelle des dynamiques logit ou des potential games, où la distribution stationnaire prend une forme de type Boltzmann et où des transitions bistables apparaissent lorsque le couplage est suffisamment fort.

L'un des apports originaux de votre démarche est de ne pas vous contenter d'un cadre conceptuel. Vous cherchez à calibrer le système sur les données réelles. À partir des variations observées, vous reconstruisez un proxy de $p(t)$, puis vous estimez une fonction empirique

$$F(x) = E[p|x),$$

qui peut être interprétée comme une probabilité idéale conditionnelle au niveau. Cette fonction joue le rôle de champ externe empirique. Ensuite, vous ajustez les paramètres $K$, $J$, $\alpha$, $h_0$, $h_1$ pour minimiser l'écart entre la dynamique prédite et la dynamique observée.

Cette étape est essentielle, car elle transforme un modèle théorique en modèle confronté au réel.

Ce point mérite d'être souligné. La calibration n'a pas seulement une fonction prédictive. Elle sert à tester la cohérence structurelle du modèle. Si un système de type Ising couplé à un niveau de marché reproduit une part significative des régimes, des transitions et des corrélations observées, alors la lecture par la physique de la complexité n'est pas seulement suggestive ; elle devient opératoire.

## Pourquoi l'EDA apparaît comme la clé de voûte

À ce stade, on peut comprendre pourquoi l'EDA occupe une place centrale dans votre travail. Elle est la phase où l'on détecte les formes pertinentes. Elle est la phase où l'on choisit les variables efficaces. Elle est la phase où l'on décide si un modèle de bruit blanc, de bistabilité, de cascade ou de coordination a un sens. Elle est la phase où l'on passe du phénomène brut à la classe dynamique pertinente.

| Sans EDA structurée | Avec EDA orientée par la complexité |
|---|---|
| Le modèle est choisi trop tôt | Le modèle émerge des signatures observées |
| Les transformations sont techniques | Les transformations sont physiquement motivées |
| Les corrélations sont seulement descriptives | Les corrélations deviennent des indices de mémoire et de régime |
| Les distributions sont des objets statistiques | Les distributions deviennent des potentiels effectifs |

| Sans EDA structurée | Avec EDA orientée par la complexité |
|---|---|
| Le ML est appliqué directement | Le ML est contraint par un modèle interprétable |

C’est en ce sens que vous pouvez défendre l’idée suivante : l’EDA est la clé de voûte de la data science. Non pas parce qu’elle remplace les modèles, mais parce qu’elle rend possible un enchaînement cohérent entre données, physique, calibration et apprentissage.

## Vers un système expert : modéliser une compétence

Le prolongement naturel de votre méthode est la construction d’un système expert. Cette idée est particulièrement intéressante, car elle permet de reformuler votre démarche non plus seulement comme une suite de calculs, mais comme une modélisation d’expertise. En effet, ce que vous faites lorsqu’on vous soumet un système n’est pas simplement exécuter un pipeline. Vous raisonnez. Vous identifiez des classes de signatures, vous choisissez des tests pertinents, vous reliez les résultats à une littérature, vous proposez une famille de modèles, vous testez des hypothèses, vous calibrez un système. Cette chaîne relève d’une expertise.

L’idée de système expert est ancienne. Les expert systems des années 1980 cherchaient déjà à formaliser des connaissances expertes et des règles de décision. Votre projet s’inscrit dans cette lignée, mais avec une différence majeure : vous y ajoutez les données massives, les capacités de calcul modernes et le scientific machine learning.

L’architecture d’un tel système expert peut se décrire simplement. Il commencerait par une fiche système : nature du système, observables, fréquence d’échantillonnage, capteurs, objectif. Il passerait ensuite par une phase de provenance et métrologie, afin de vérifier que les structures observées ne viennent pas d’artefacts de mesure. Puis il interrogerait une bibliographie structurée, non pas comme encyclopédie passive, mais comme réservoir de signatures, de modèles et de recettes de traitement. Ensuite viendrait la cascade de tests d’EDA : stationnarité, queues lourdes, spectres, corrélations, régimes, potentiels. Ces résultats seraient fusionnés par un moteur d’inférence, capable de proposer des hypothèses dynamiques classées. Enfin, le système construirait un modèle, le calibrerait et préparerait son transfert vers un PINN.

Dans cette architecture, l’EDA reste au centre. C’est elle qui transforme la donnée en profil de complexité. C’est elle qui alimente l’inférence experte. C’est elle qui décide si l’on doit aller vers un modèle de coordination, un potentiel bistable, un modèle à avalanche, une diffusion stochastique ou une architecture à régimes.

## Le prolongement vers les PINN

Une fois le modèle physique identifié et calibré, votre objectif n’est pas de l’abandonner au profit d’un réseau de neurones. Au contraire, vous cherchez à faire du modèle un socle contraignant pour l’apprentissage. C’est exactement la philosophie des Physics-Informed Neural Networks. Dans votre cadre, cela signifie que l’on pourrait entraîner un réseau non pas pour réinventer la dynamique, mais pour apprendre les fonctions ou paramètres qui restent incertains dans un système du type

$$\frac{dx}{dt} = K(2p - 1), \tau\frac{dp}{dt} = -p + \sigma(\beta(J(2p - 1) + h(x))).$$

Le réseau ne serait plus un substitut opaque, mais un prolongement informé du modèle. Cette articulation entre EDA, modèle interprétable, calibration et PINN donne à votre méthode sa cohérence et sa modernité.

## Conclusion : une data science comme science des formes dynamiques

Au fond, votre travail poursuit une ambition claire. Il ne s'agit pas seulement de mieux prédire, ni seulement de mieux ajuster, ni seulement de mieux visualiser. Il s'agit de faire de la data science une science des formes dynamiques du réel. Pour cela, il faut une phase d'observation rigoureuse. Il faut une grammaire de l'interprétation. Il faut un pont entre signatures observées et familles de modèles. Il faut une calibration sur des données réelles. Et il faut enfin un prolongement vers des apprentissages contraints. L'EDA, telle que vous la pratiquez, est exactement ce point de passage obligé.

C'est pour cela qu'elle peut être présentée comme la clé de voûte de la data science. Sans elle, le modèle flotte. Avec elle, la donnée, la théorie et l'apprentissage se rejoignent.

# Introduction à la suite de l'ouvrage : œuvre concrète de l'EDA

Nous vivons dans un monde saturé de données, de signaux, de fluctuations et d'incertitudes. Les marchés financiers produisent chaque jour des séries temporelles immenses. Les systèmes physiques, qu'il s'agisse de turbulence, d'état de mer ou de phénomènes de diffusion et de transport, génèrent des comportements irréguliers et multi-échelles. Les projets industriels, les organisations humaines, les dynamiques historiques, les systèmes sociaux ou adaptatifs semblent eux aussi traversés par des phases de stabilité relative, des tensions latentes, des crises, des bifurcations et des réorganisations. À première vue, ces objets n'ont rien en commun. Ils relèvent de disciplines différentes, utilisent des langages différents, mobilisent des outils différents. Pourtant, lorsqu'on les observe avec attention, une intuition surgit : derrière la diversité des domaines, les formes de comportement se ressemblent.

Cette intuition constitue le point de départ du présent ouvrage. Elle ne procède ni d'un goût pour les analogies faciles, ni d'un désir de ramener toute réalité à une formule unique. Elle naît d'une expérience répétée d'analyse des données et de confrontation à des systèmes réels. Dans de nombreux cas, les phénomènes observés présentent des signatures communes. Les fluctuations rapides paraissent souvent peu prédictibles à court terme, mais leur intensité possède une mémoire. Les événements extrêmes sont plus fréquents que ne le prévoient les modèles gaussiens classiques. Les systèmes alternent entre phases relativement stables et épisodes de transition ou de crise. Les trajectoires semblent évoluer dans des paysages dynamiques comportant des zones d'attraction, des

seuils et des barrières. Enfin, les structures observées changent avec l'échelle, mais conservent souvent une certaine cohérence de forme lorsque l'on passe du détail à la vue d'ensemble.

Ces propriétés ne sont pas anecdotiques. Elles suggèrent qu'il existe une grammaire commune des systèmes complexes. La question n'est alors plus seulement de décrire tel ou tel domaine particulier, mais de comprendre s'il existe une forme dynamique simple capable de restituer ces caractéristiques essentielles. Une telle ambition suppose de quitter à la fois l'illusion de l'explication totale et celle du simple ajustement numérique. Il ne s'agit ni de construire un modèle exhaustif qui prétendrait tout contenir, ni de se contenter de méthodes purement algorithmiques qui reproduiraient certaines régularités sans en livrer le sens. Il s'agit de rechercher des modèles intermédiaires : des modèles suffisamment simples pour être intelligibles, suffisamment riches pour capturer les structures fondamentales du réel, et suffisamment génériques pour être transportables d'un domaine à l'autre.

C'est dans cet espace, entre la théorie abstraite et la pure ingénierie des données, que se situe le présent travail.

L'un des enjeux centraux de la science contemporaine est en effet le suivant : comment transformer l'abondance des données en intelligibilité ? L'accumulation de mesures, de séries temporelles, de signaux et d'observations ne conduit pas automatiquement à la compréhension. Nous disposons aujourd'hui d'outils puissants pour traiter l'information, calculer, classifier, projeter, optimiser, prédire localement. Mais la question du sens physique ou structurel demeure. Pourquoi certaines formes reviennent-elles ? Pourquoi observe-t-on des queues lourdes dans des domaines très différents ? Pourquoi la mémoire se loge-t-elle moins dans les fluctuations elles-mêmes que dans leur intensité ? Pourquoi certains systèmes semblent-ils passer brutalement d'un régime à un autre ? Pourquoi la crise n'est-elle pas toujours un accident extérieur, mais parfois l'expression normale d'une dynamique interne approchant un seuil ?

Ces questions exigent une posture particulière. Elles invitent à revenir vers l'analyse exploratoire des données, non comme une simple étape préliminaire, mais comme un moment décisif de la connaissance. L'analyse exploratoire, ou EDA, ne consiste pas seulement à tracer quelques courbes avant d'appliquer un modèle. Elle est la phase où les données parlent avant que la théorie ne les fasse taire. Elle permet de repérer des symétries, des asymétries, des persistances, des structures de dépendance, des régimes, des regroupements, des singularités. Elle oblige le modélisateur à écouter le matériau empirique. Dans cette perspective, l'EDA n'est pas une technique subalterne. Elle devient une discipline du regard.

C'est précisément par ce chemin que l'on peut introduire le problème traité dans ce livre.

L'examen de longues séries du CAC 40 fournit un exemple particulièrement éclairant. Lorsqu'on observe le logarithme de l'indice sur une longue période, puis ses variations journalières, plusieurs faits robustes apparaissent. Les rendements eux-mêmes possèdent en général peu d'autocorrélation ; ils ressemblent à première vue à une agitation quasi

aléatoire. En revanche, le carré des rendements, ou toute mesure voisine de l'intensité des fluctuations, présente une mémoire longue. Les phases de forte agitation se regroupent. Les périodes calmes durent. Les grands mouvements tendent à apparaître par grappes. De plus, la distribution des variations ne suit pas une loi gaussienne simple : ses queues sont lourdes, ce qui signifie que les événements rares et violents surviennent plus souvent que ne le prédit la théorie la plus élémentaire de la diffusion brownienne. Enfin, lorsqu'on cherche à résumer la dynamique globale par des outils de réduction de dimension, comme l'analyse en composantes principales, on constate qu'une direction dominante concentre une large part de la variance observée. En d'autres termes, sous la dispersion apparente des données se cache une structure.

Cette situation est exemplaire, car elle impose plusieurs contraintes à tout modèle crédible. Un bon modèle doit produire une agitation locale qui semble largement imprévisible à court terme. Il doit aussi être capable de générer une persistance dans l'intensité des fluctuations. Il doit reproduire des événements extrêmes plus fréquents que ceux d'un schéma gaussien standard. Il doit permettre l'existence de régimes, de transitions et de comportements métastables. Enfin, il doit rester suffisamment simple pour que chacun de ses termes puisse recevoir une interprétation. Le problème n'est donc pas seulement statistique. Il est à la fois physique, méthodologique et épistémologique.

C'est ici qu'apparaît la nécessité de ce que l'on pourrait appeler des modèles pédagogiques et génériques à peu de paramètres. Le mot pédagogique ne signifie pas simpliste. Il signifie qu'un modèle doit pouvoir être compris, transmis, discuté, relié à des intuitions physiques. Le mot générique ne signifie pas vague. Il signifie qu'un même squelette dynamique peut, sous des interprétations différentes, rendre compte de plusieurs familles de phénomènes. Quant à l'exigence de peu de paramètres, elle répond à une double contrainte. D'une part, plus un modèle contient de paramètres libres, plus il peut s'ajuster au passé sans révéler la structure profonde du système. D'autre part, lorsque les paramètres ont une signification physique claire, ils permettent de penser, et pas seulement de calculer. Ils deviennent les porteurs d'une interprétation du phénomène.

Le présent ouvrage est ainsi fondé sur une hypothèse forte. Cette hypothèse est qu'une équation différentielle stochastique relativement simple peut constituer une brique élémentaire pour la description d'une large classe de systèmes complexes. Cette équation repose sur trois ingrédients principaux. Le premier est une dérive dans un potentiel, qui organise l'espace des états en régimes, puits de stabilité, zones de transition et barrières. Le deuxième est une agitation stochastique dont l'intensité varie dans le temps avec une mémoire longue, de manière à représenter l'intermittence et l'organisation multi-échelle de la fluctuation. Le troisième est un bruit additif de fond, qui traduit l'existence d'une agitation résiduelle irréductible.

Une telle écriture n'est pas choisie par goût de l'élégance formelle. Elle émerge parce qu'elle répond point par point aux propriétés observées dans les données. Le potentiel rend possible l'existence de plusieurs régimes et de passages de seuil. La volatilité à mémoire longue introduit la persistance des périodes calmes ou agitées et donne naissance au regroupement des fortes fluctuations. Le bruit additif garantit qu'aucun état

n’est parfaitement figé et que la dynamique conserve un fond d’exploration aléatoire. Cette équation ne prétend pas reproduire tous les détails d’un système réel. Elle vise autre chose : faire apparaître son architecture dynamique minimale.

En ce sens, elle peut être considérée comme un modèle topologique des systèmes complexes. Le mot topologique doit être entendu ici dans un sens large. Il ne s’agit pas seulement de géométrie abstraite, mais de la conservation d’une structure qualitative fondamentale. Ce qui importe n’est pas de reproduire chaque détail local, mais de conserver la forme globale des relations entre régimes, fluctuations, transitions, intensités et temps longs. Un tel modèle est précieux parce qu’il permet de franchir les frontières disciplinaires sans perdre la rigueur. Il devient possible de parler le même langage lorsqu’on étudie un indice financier, un signal de clutter de mer, une dynamique de projet, une évolution historique ou un système adaptatif soumis à des perturbations.

Cette ambition appelle cependant une précaution essentielle. Le livre ne cherche pas à réduire arbitrairement tous les phénomènes à une équation unique. Il cherche plutôt à mettre en évidence l’existence d’une forme réduite commune, au sens où plusieurs systèmes très différents peuvent, à une certaine échelle d’observation, être décrits par une même classe de dynamique. C’est une démarche classique en physique, mais dont la portée dépasse largement la physique elle-même. Lorsqu’on passe d’équations très détaillées à une variable d’ordre, à une amplitude lente ou à une observable dominante, on ne nie pas la richesse microscopique du système. On cherche le niveau pertinent où cette richesse s’organise en formes lisibles.

Cette idée de réduction joue un rôle central dans le présent travail. Elle relie l’analyse des données, la physique statistique, la théorie des bifurcations, la théorie des systèmes complexes et la data science. Réduire ne signifie pas mutiler le réel. Réduire signifie dégager les variables, les structures et les mécanismes qui gouvernent son comportement à l’échelle où l’on souhaite le comprendre ou l’agir. Dans les données financières, par exemple, on peut passer d’une multitude de composantes et d’influences à une variable synthétique dominante. Dans un système fluide, on peut chercher à projeter une dynamique très riche sur une amplitude lente ou une observable effective. Dans une organisation ou un projet, on peut tenter d’identifier un paramètre d’ordre traduisant l’état global de cohérence, de surcharge ou de dérive du système. Dans tous les cas, la même question revient : quelle est la variable d’ordre pertinente, et quelle dynamique minimale la gouverne ?

C’est ici que la filiation avec des équations plus fondamentales, notamment celles de la mécanique des fluides, entre en jeu. Les équations de Navier-Stokes décrivent des champs de vitesse et de pression d’une extrême richesse. Elles engendrent des structures tourbillonnaires, des transferts d’énergie entre échelles, des régimes stables ou instables, des transitions vers la turbulence, des phénomènes intermittents. Il serait illusoire de croire qu’une équation stochastique unidimensionnelle puisse contenir toute cette complexité. En revanche, il est légitime de chercher comment, à partir d’un tel univers, émergent des descriptions réduites. Lorsque l’on s’intéresse à une observable particulière, à une amplitude lente, à une composante dominante ou à une structure collective, on peut

souvent ramener la dynamique à une forme réduite où apparaissent une dérive effective, une diffusion, des non-linéarités organisatrices et des termes représentant l'influence agrégée des degrés de liberté non résolus. Ainsi, le passage de Navier-Stokes à une dynamique réduite n'est pas un appauvrissement arbitraire ; c'est un changement d'échelle de description.

Le même raisonnement vaut pour les systèmes non purement physiques. Dans un projet industriel, les milliers de microdécisions, de contraintes, d'interactions humaines, de retards et de réallocations ne sont pas directement modélisables dans toute leur granularité. Pourtant, certaines variables globales peuvent émerger et structurer la dynamique collective. Dans la cliodynamique, les interactions démographiques, économiques, institutionnelles et conflictuelles composent un réseau immense de causalités. Là encore, l'ambition n'est pas de tout restituer point par point, mais de repérer des variables macroscopiques et des mécanismes de couplage capables de rendre compte de régimes, de cycles, d'instabilités et de bifurcations. Dans les systèmes adaptatifs, la complexité ne disparaît pas ; elle se condense dans des lois effectives.

Le présent livre s'inscrit donc dans une tradition scientifique très ancienne et très actuelle à la fois : celle qui cherche des formes simples sous les phénomènes complexes. Mais il le fait dans un contexte nouveau, celui d'une science enrichie par l'EDA, par la disponibilité massive des données, par les méthodes de réduction de dimension, par la modélisation stochastique et par la nécessité croissante de relier théorie, données et action.

Cette articulation entre physique de la complexité et data science mérite d'être explicitée. Trop souvent, ces deux domaines sont mis en opposition implicite. D'un côté, on placerait les modèles explicatifs, structurés, ancrés dans une tradition théorique forte. De l'autre, on placerait les approches orientées données, exploratoires, statistiques, algorithmiques, parfois accusées d'être aveugles au sens. En réalité, cette opposition est stérile. La data science devient véritablement féconde lorsqu'elle permet de faire émerger les contraintes que toute théorie sérieuse doit respecter. Et la physique de la complexité devient particulièrement puissante lorsqu'elle accepte de se confronter aux signatures empiriques des données réelles, dans toute leur rugosité, leur intermittence et leur non-idéalité. Le dialogue entre les deux n'est donc pas accessoire. Il est le cœur même de la démarche.

Dans ce dialogue, l'analyse exploratoire joue à nouveau un rôle décisif. Elle permet de partir des formes observées pour remonter vers les mécanismes plausibles. Si l'on voit apparaître des queues lourdes, il faut comprendre quel type de modulation de la diffusion ou de mélange d'échelles peut les produire. Si l'on observe une mémoire longue dans la volatilité, il faut introduire une dynamique de l'intensité elle-même, et non se contenter d'un bruit homogène. Si des régimes se dessinent, il faut penser un paysage de potentiel, des minima locaux, des barrières et des franchissements. Si une composante principale émerge, il faut l'interpréter comme variable d'ordre ou direction dominante. L'EDA n'est donc pas ici une simple description ; elle devient un opérateur de dérivation conceptuelle.

La notion de variable d'ordre est, à cet égard, fondamentale. Elle constitue le pont entre complexité et lisibilité. Dans un système très riche, tout ne varie pas avec la même

importance. Certaines directions de l'espace d'état concentrent l'essentiel de l'énergie, de la variance ou de la structure. Identifier une telle direction ne signifie pas nier le reste, mais reconnaître qu'à l'échelle pertinente, une partie du comportement peut être gouvernée par un petit nombre de degrés de liberté collectifs. Cette idée, venue de la physique des transitions de phase et des bifurcations, possède une portée très générale. Elle permet de penser la cohérence d'un système sans en abolir la diversité interne. Elle justifie aussi que l'on puisse chercher une équation simple pour gouverner l'évolution de cette variable dominante.

Une fois cette variable identifiée, la question devient : quelle forme doit prendre l'équation qui la régit ? Les observations conduisent vers une réponse de plus en plus précise. Si le système présente des régimes, il faut une dérive non linéaire pouvant être lue comme la pente d'un potentiel. Si le système traverse des périodes calmes et agitées qui persistent, il faut une intensité stochastique possédant sa propre dynamique lente et mémorielle. Si le système n'est jamais totalement clos sur lui-même, il faut maintenir un terme de bruit additif résiduel. C'est ainsi que se construit progressivement l'EDO proposée dans ce livre : non comme une hypothèse tombée du ciel, mais comme la traduction la plus simple possible d'un ensemble de contraintes empiriques et de principes physiques.

Cette équation a alors plusieurs propriétés remarquables. Elle est suffisamment souple pour produire des distributions non gaussiennes, des regroupements de volatilité, des franchissements de barrières et des alternances de régimes. Elle est suffisamment structurée pour être reliée à une équation de Fokker-Planck-Kolmogorov, permettant l'étude des densités, des états stationnaires, des flux de probabilité et des effets de la modulation de la diffusion. Elle est suffisamment générique pour être adaptée à plusieurs domaines. Elle est suffisamment pédagogique pour que chacun de ses termes puisse être discuté, critiqué, calibré et interprété.

Ce caractère pédagogique mérite d'être revendiqué. Dans un paysage scientifique parfois dominé par la sophistication formelle ou la performance computationnelle brute, il peut sembler modeste de rechercher des modèles simples. En réalité, cette simplicité est une force. Un modèle simple n'est pas un modèle pauvre lorsqu'il isole les bons mécanismes. Il devient un outil de pensée, un cadre de comparaison, un support de transfert interdisciplinaire. Il aide à distinguer l'essentiel de l'accessoire. Il permet de comprendre pourquoi des systèmes éloignés partagent certaines formes statistiques ou dynamiques. Il ouvre enfin la possibilité d'une science plus explicative, où les paramètres ne sont pas de simples coefficients d'ajustement, mais les noms mathématiques d'effets physiques ou systémiques identifiables.

L'ambition de ce livre est donc double. Elle est théorique, parce qu'il s'agit de proposer une forme réduite des systèmes complexes et d'en étudier rigoureusement les propriétés. Elle est également méthodologique et appliquée, parce qu'il s'agit de montrer comment cette forme peut être dégagée à partir des données, reliée à des théories plus profondes, puis transportée dans différents champs d'application.

Le CAC 40 jouera dans cette démarche un rôle fondateur. Non parce que la finance serait le domaine ultime d'application, mais parce qu'elle fournit un laboratoire empirique

exceptionnel. Les données y sont abondantes, les faits stylisés bien établis, les signatures de complexité visibles. À travers l'étude de cette série, il sera possible de montrer comment l'analyse exploratoire conduit à exiger certaines propriétés du modèle, comment la réduction de dimension fait émerger une variable dominante, comment les distributions et corrélations orientent vers une diffusion modulée, et comment une dynamique dans un potentiel devient une réponse naturelle à la présence de régimes ou de structures métastables.

À partir de là, l'ouvrage ouvrira progressivement le spectre des applications. Le clutter de mer permettra d'explorer le lien avec les phénomènes de surface, les états de mer, les distributions à queue lourde et la structuration multi-échelle des signaux radar. La gestion de projets offrira un terrain très différent, mais profondément instructif, où les notions de variable d'ordre, de stabilité organisationnelle, de dérive, de surcharge, de bifurcation et de transition deviennent opératoires. La cliodynamique montrera comment des dynamiques collectives de grande ampleur peuvent être pensées à partir de variables macroscopiques couplées, dans un cadre où les régimes, les cycles et les tensions structurelles jouent un rôle central. Enfin, les systèmes complexes adaptatifs permettront de reformuler l'ensemble dans un langage plus général encore, où le modèle devient une brique pour penser l'interaction entre structure, fluctuation, mémoire et adaptation.

Sous cette diversité d'applications, le véritable objet du livre restera le même. Il s'agit de comprendre comment l'aléa devient forme. Il s'agit de montrer qu'un système complexe n'est pas simplement un système compliqué, mais un système où le désordre apparent est organisé par un paysage, des temps propres, des mémoires, des couplages et des seuils. Il s'agit enfin de défendre l'idée que la compréhension passe souvent par des modèles réduits, à la fois interprétables et empiriquement fondés, plutôt que par l'alternative stérile entre la théorie pure et le calcul aveugle.

Le lecteur ne trouvera donc pas ici un traité clos, ni une théorie prétendant tout résoudre. Il trouvera plutôt une enquête. Une enquête sur une forme dynamique possible des systèmes complexes. Une enquête qui part des données, qui dialogue avec la physique, qui s'appuie sur les outils de la data science, et qui se prolonge dans des domaines d'application très différents. Une enquête qui cherche moins à enfermer le réel dans un schéma qu'à mettre en évidence, sous la diversité des phénomènes, quelques invariants de structure et de transformation.

En ce sens, ce livre défend une conviction simple et ambitieuse : il existe, au cœur de nombreux systèmes complexes, une dynamique réduite où se nouent l'ordre, le bruit, la mémoire et la transition. Comprendre cette dynamique, c'est déjà franchir un pas important vers une science plus unifiée des formes instables, des crises, des régimes et des métamorphoses.

C'est ce chemin que nous allons maintenant parcourir.