

Statistique et Apprentissage Automatique en Science des Données

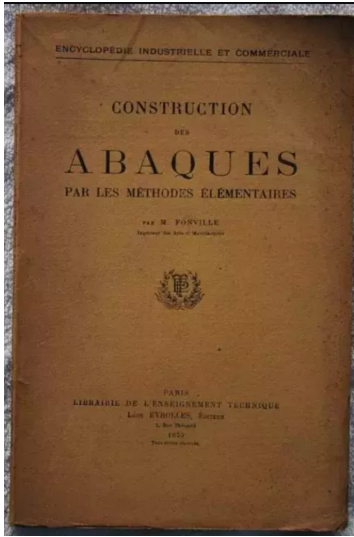
Partie 1 : Historique

SAA-SD - Historique

- 1940-70** : Statistiques classiques — Questions associées à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.
- 1970s** : Généralisation des premiers outils informatiques — Exploration des données plus volumineuses à l’aide de systèmes experts.
- 1980s** : Emergence de l’apprentissage automatique en alternative aux systèmes experts.
- 1990s** : Les données ne sont plus planifiées mais sont préalablement acquises — From Data Mining to Knowledge Discovery.
- 2000s** : Le nombre de variables p explose, notamment avec les données omiques où $p \gg n$ — La qualité de prévision devient plus importante que la réalité du modèle devenu boîte noire. Fléau de la dimension.
- 2010s** : Le nombre d'observations n explose dans le e-commerce, la géo-localisation, ... — Structuration des bases de données structurées en cloud et moyens de calculs regroupés en clusters. La notion de rapidité des algorithmes devient critique.
- 2020s** : Explosion des usages dû à la facilité d’accès à des données massives et à des ressources calcul puissantes — explicabilité des décisions à fort impact social ou économique, robustesse des décisions dans un cadre critique, embarquabilité et frugalité des modèles, cybersécurité, choix de modèles adaptés pour des contextes small à big data.
- 2025 - .** : Utilisation massive des I.A. génératives — Questions de loyauté, souveraineté, cybersécurité, frugalité, impact social, ...



et



vers



SAA-SD - Historique

1940-70 : Statistiques classiques — Questions associées à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques — Exploration des données plus volumineuses à l’aide de systèmes experts.

1980s : Emergence de l’apprentissage automatique en alternative aux systèmes experts.

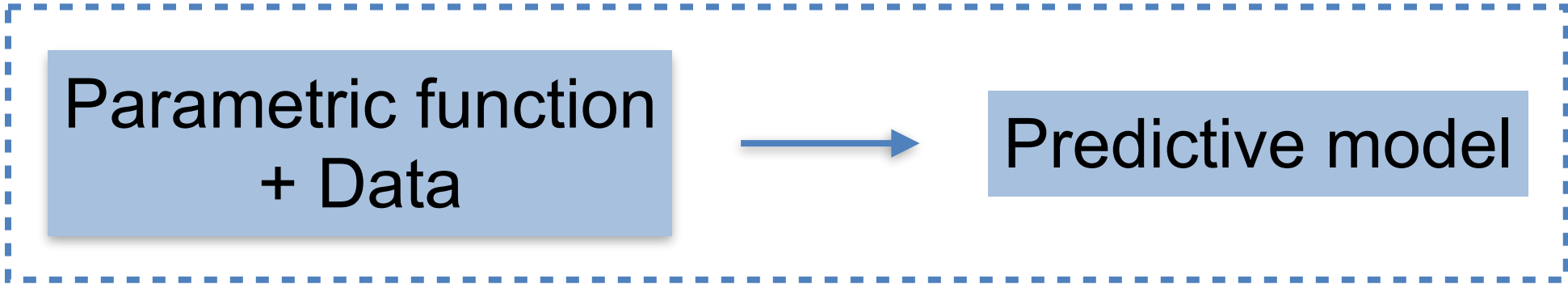
1990s : Les données ne sont plus planifiées mais sont préalablement acquises — From Data Mining to Knowledge Discovery.

2000s : Le nombre de variables p explose, notamment avec les données omiques où $p \gg n$ — La qualité de prévision devient plus importante que la réalité du modèle devenu boîte noire. Fléau de la dimension.

2010s : Le nombre d'observations n explose dans le e-commerce, la géo-localisation, ... — Structuration des bases de données structurées en cloud et moyens de calculs regroupés en clusters. La notion de rapidité des algorithmes devient critique.

2020s : Explosion des usages dû à la facilité d’accès à des données massives et à des ressources calcul puissantes — explicabilité des décisions à fort impact social ou économique, robustesse des décisions dans un cadre critique, embarquabilité et frugalité des modèles, cybersécurité, choix de modèles adaptés pour des contextes small à big data.

2025 - . : Utilisation massive des I.A. génératives — Questions de loyauté, souveraineté, cybersécurité, frugalité, impact social, ...



SAA-SD - Historique

1940-70 : Statistiques classiques — Questions associées à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques — Exploration des données plus volumineuses à l'aide de systèmes experts.

1980s : Emergence de l'apprentissage automatique en alternative aux systèmes experts.

1990s : Les données ne sont plus planifiées mais sont préalablement acquises — From Data Mining to Knowledge Discovery.

2000s : Le nombre de variables p explose, notamment avec les données omiques où $p \gg n$ — La qualité de prévision devient plus importante que la réalité du modèle devenu boîte noire. Fléau de la dimension.

2010s : Le nombre d'observations n explose dans le e-commerce, la géo-localisation, ... — Structuration des bases de données structurées en cloud et moyens de calculs regroupés en clusters. La notion de rapidité des algorithmes devient critique.

2020s : Explosion des usages dû à la facilité d'accès à des données massives et à des ressources calcul puissantes — explicabilité des décisions à fort impact social ou économique, robustesse des décisions dans un cadre critique, embarquabilité et frugalité des modèles, cybersécurité, choix de modèles adaptés pour des contextes small à big data.

2025 - . : Utilisation massive des I.A. génératives — Questions de loyauté, souveraineté, cybersécurité, frugalité, impact social, ...



amazon.com

SAA-SD - Historique

1940-70 : Statistiques classiques — Questions associées à une hypothèse expérimentalement réfutable avec $n \approx 30$ observations et $p < 10$ variables.

1970s : Généralisation des premiers outils informatiques — Exploration des données plus volumineuses à l'aide de systèmes experts.

1980s : Emergence de l'apprentissage automatique en alternative aux systèmes experts.

1990s : Les données ne sont plus planifiées mais sont préalablement acquises — From Data Mining to Knowledge Discovery.

2000s : Le nombre de variables p explose, notamment avec les données omiques où $p \gg n$ — La qualité de prévision devient plus importante que la réalité du modèle devenu boîte noire. Fléau de la dimension.

2010s : Le nombre d'observations n explose dans le e-commerce, la géo-localisation, ... — Structuration des bases de données structurées en cloud et moyens de calculs regroupés en clusters. La notion de rapidité des algorithmes devient critique.

2020s : Explosion des usages dû à la facilité d'accès à des données massives et à des ressources calcul puissantes — explicabilité des décisions à fort impact social ou économique, robustesse des décisions dans un cadre critique, embarquabilité et frugalité des modèles, cybersécurité, choix de modèles adaptés pour des contextes small à big data.

2025 - . : Utilisation massive des I.A. génératives — Questions de loyauté, souveraineté, cybersécurité, frugalité, impact social, ...

