Des maths pour l'exloitation des données

V.THOUVENOT - vincent.thouvenot@thalesgroup.com

Thales Communication & Security

12/02/2018

Sommaire

- CPGE - Spécialité Maths Physique
- ENSAI (Ecole Nationale de Statistique et de l'Analyse de l'Information) - Spécialité Génie Statistique - Rennes
- Master 2 de Statistique et d'Econométrie - Université de Rennes 1
- Thèse de Mathématiques Appliquées sous la direction d'Antoniadis, de Poggi et de Goude (tuteur industriel) - Université d'Orsay et EDF R&D
- Data Scientist - Thales Communication & Security

- CPGE - Spécialité Maths Physique
- ENSAI (Ecole Nationale de Statistique et de l'Analyse de l'Information) - Spécialité Génie Statistique - Rennes
- Master 2 de Statistique et d'Econométrie - Université de Rennes 1
- Thèse de Mathématiques Appliquées sous la direction d'Antoniadis, de Poggi et de Goude (tuteur industriel) - Université d'Orsay et EDF R&D
- Data Scientist - Thales Communication & Security

- CPGE - Spécialité Maths Physique
- ENSAI (Ecole Nationale de Statistique et de l'Analyse de l'Information) - Spécialité Génie Statistique - Rennes
- Master 2 de Statistique et d'Econométrie - Université de Rennes 1
- Thèse de Mathématiques Appliquées sous la direction d'Antoniadis, de Poggi et de Goude (tuteur industriel) - Université d'Orsay et EDF R&D
- Data Scientist - Thales Communication & Security

- CPGE - Spécialité Maths Physique
- ENSAI (Ecole Nationale de Statistique et de l'Analyse de l'Information) - Spécialité Génie Statistique - Rennes
- Master 2 de Statistique et d'Econométrie - Université de Rennes 1
- Thèse de Mathématiques Appliquées sous la direction d'Antoniadis, de Poggi et de Goude (tuteur industriel) - Université d'Orsay et EDF R&D
- Data Scientist - Thales Communication & Security

- CPGE - Spécialité Maths Physique
- ENSAI (Ecole Nationale de Statistique et de l'Analyse de l'Information) - Spécialité Génie Statistique - Rennes
- Master 2 de Statistique et d'Econométrie - Université de Rennes 1
- Thèse de Mathématiques Appliquées sous la direction d'Antoniadis, de Poggi et de Goude (tuteur industriel) - Université d'Orsay et EDF R&D
- Data Scientist - Thales Communication & Security

ENSAI

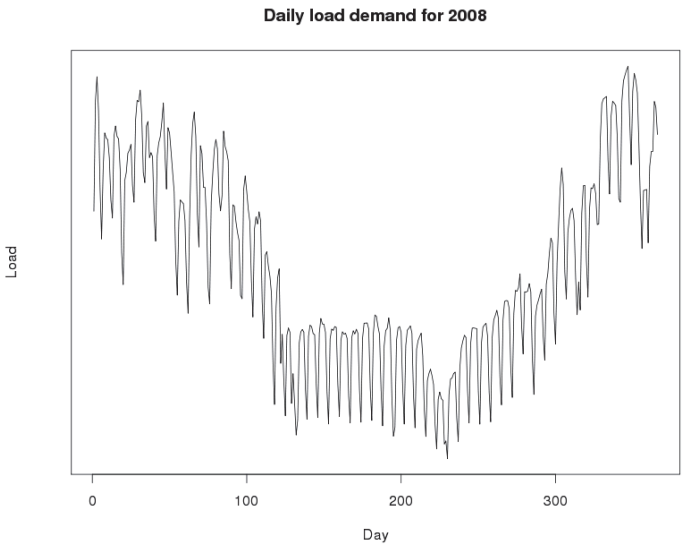
- Campus de Ker Lann proche de Rennes
- Habilitée par la CTI à délivrer le diplome d'ingénieur conférant le grade de master
- Triple compétence statistique-économétrie-informatique
- Deux cycles
  - Fonctionnaire (INSEE et Ministères)
  - Ingénieur
- 6 filières de spécialisation pour la 3ème année des ingénieurs
  - Marketing Quantitatif et Revenue Management
  - Statistique et Ingénieurie des données
  - Génie Statistique
  - Statistique pour les sciences de la vie
  - Gestion des risques et ingénierie financière
  - Ingénierie statistique des territoires et de la santé
- Recrutement à partir de bac + 2 : 50 % de MP, 15 % de B/L (économie et Sciences sociales) et ENS Cachan D2 (économie et Gestion), 35 % sur titres (DUT Stid, Licence de mathématique, étudiants internationaux, etc.)

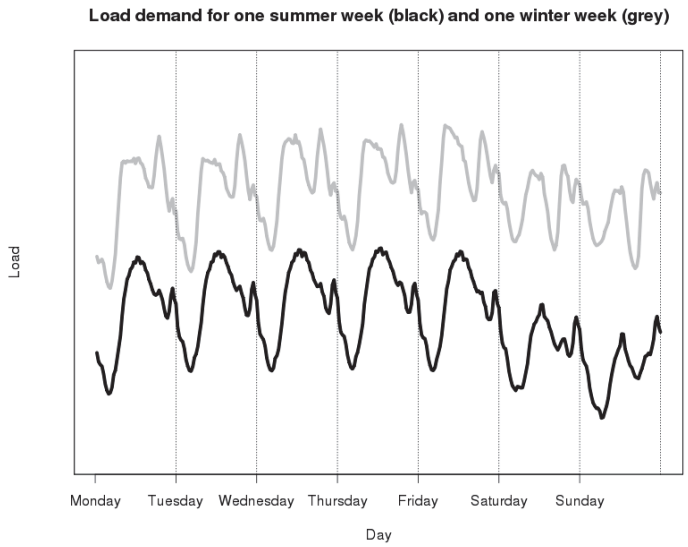
Contribution des Statistiques en entreprises : illustration avec les filières de l'ENSAI

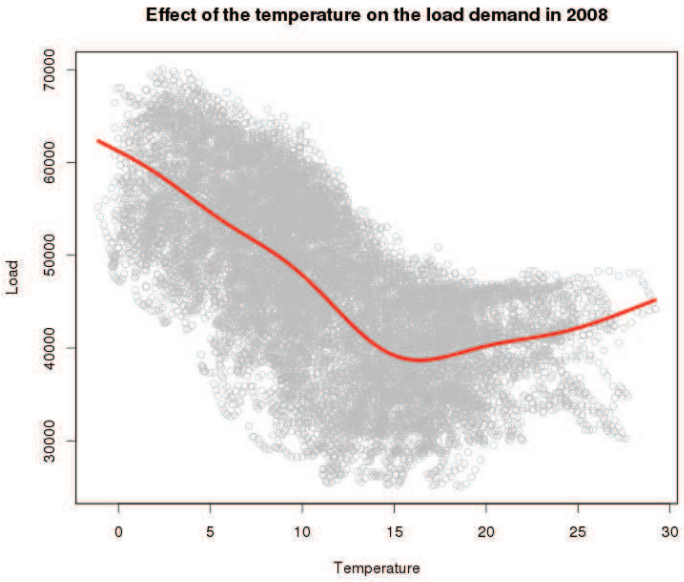
- Marketing Quantitatif et Revenue Management
  - Etude comportement de clients d'un grand fournisseur de jeu vidéo
  - Ciblage de clients dans le luxe
- Statistique et Ingénieurie des données
  - Implémentation d'algorithmes passant à l'échelle pour de la cybersécurité
- Statistique pour les sciences de la vie
  - Etude de validation de médicaments
- Gestion des risques et ingénierie financière
  - Modélisation des défauts de paiement pour une banque
  - Scoring de clients pour une banque
- Ingénierie statistique des territoires et de la santé
  - Evaluation économique des produits de sante
- Génie Statistique
  - Maintenance prédictive pour l'industrie
  - Modélisation de la consommation électrique

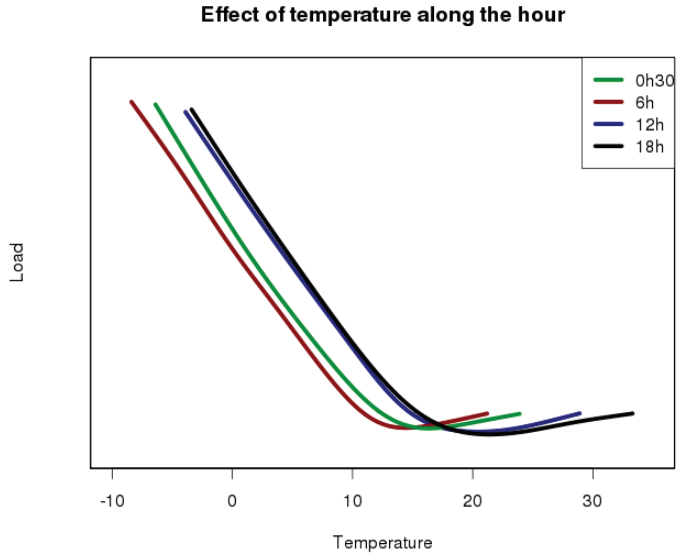
Sommaire

Courbe de charge française







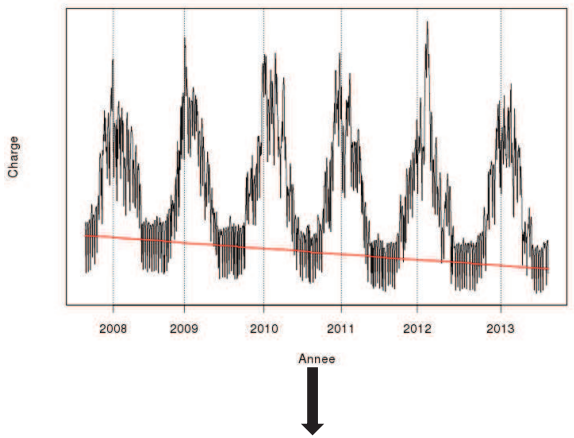


L'électricité ne se stocke pas

Production ⚖ Consommation

Ouverture du marché de l'électricité

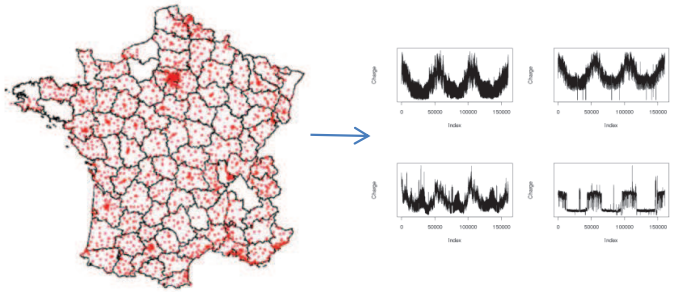
Perte/Gain de clients



Besoin de méthodes adaptatives temporellement

Développement de nouvelles technologies de mesure

Exemple: poste source



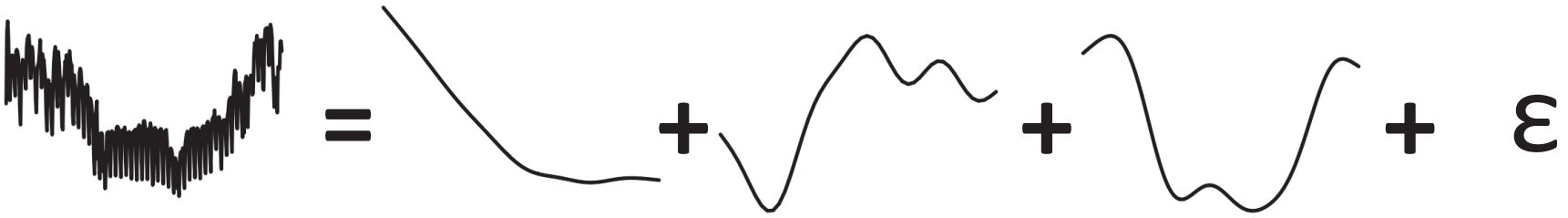
Besoin de développer des méthodes automatiques d'estimation de modèles statistiques

Quelle famille de modèles utiliser ?

- Modélisation non paramétrique $\Rightarrow$ *Beaucoup de variables explicatives*
- Modélisation paramétrique : modèle Météhore [3] $\Rightarrow$ *Beaucoup de paramètres à estimer et peu évolutif*
- Modéle additif (Hastie et Tibshirani [8]) :

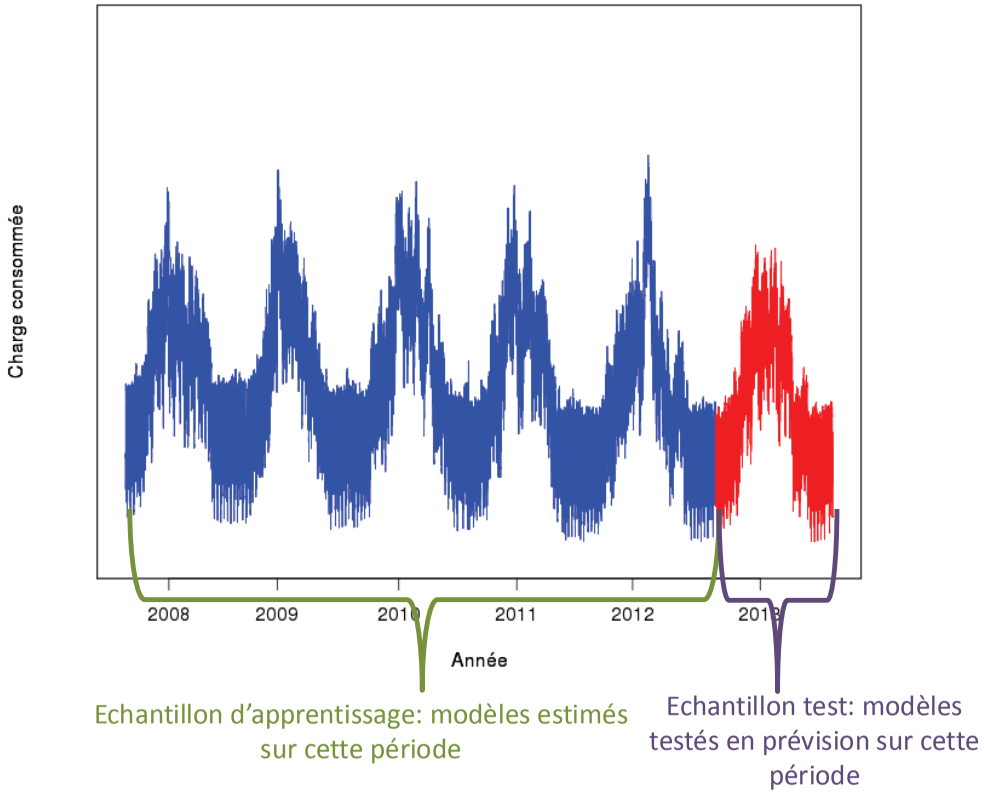
$$Y_i = \beta_0 + f_1(X_{1,i}) + \cdots + f_d(X_{d,i}) + \cdots + \varepsilon_i$$

  - Application à la prévision de consommation électrique au niveau agrégé : Pierrot et Goude [13], Fan et Hyndman [6]
  - Application à la prévision de consommation électrique au niveau local : Goude, Nedellec et Kong [7], Nedellec, Cugliari et Goude [12]



Load = f(T) + g(H) + h(D) + ε

Données du portefeuille d'un fournisseur



Quel dictionnaire de covariables ?

Météorologie instantanée

- Température brute nationale
- Vent
- Nébulosité
- ...

Température journalière

- Température brute moyenne journalière
- Température brute maximale journalière
- ...

Température lissée

- Lissage exponentiel de la température brute nationale
- Lissage exponentiel de la température moyenne journalière
- ...

Variables calendaires

- Moment dans l'année
- Nombre de jours
- Type de jours
- Offset
- ...

Modélisation de la consommation nationale

$$
\begin{aligned}
Y_t =& s_1(Toy_t) + \sum_{i=1}^{9}\sum_{j=1}^{3} \alpha_{i,j} \mathbb{1}_{DayType_t=i} \mathbb{1}_{Offset_t=j} \\
&+ \sum_{i=1}^{12} \beta_i T_{t-6} \mathbb{1}_{Month_t=i} + s_2(CC_t) + s_3(W_t) + s_4(T_t) \\
&+ s_5(\theta_t^{0.996}) + s_6(\theta_t^{Max,0.983}) \\
&+ s_7(\theta_t^{max,0.983}) + \gamma \Gamma_t + s_8(t) + \varepsilon_t
\end{aligned}
$$

- Toy : Moment dans l'année
- $DayType$: Type de jour
- $Offset$: Période homogène en termes de consommation
- T_{t-6} : Température retardée de 6 heures
- T : Température
- $\theta_t^{0.996}$: Lissage exponentiel de la température de coefficient 0.996
- Γ : Tendance associée à la part chauffage

Comment automatiser la sélection de variables et l'estimation des modèles ?

Modélisation de la consommation nationale

$$
\begin{aligned}
Y_t =& s_1(Toy_t) + \sum_{i=1}^{9}\sum_{j=1}^{3} \alpha_{i,j} \mathbb{1}_{DayType_t=i} \mathbb{1}_{Offset_t=j} \\
&+ \sum_{i=1}^{12} \beta_i T_{t-6} \mathbb{1}_{Month_t=i} + s_2(CC_t) + s_3(W_t) + s_4(T_t) \\
&+ s_5(\theta_t^{0.996}) + s_6(\theta_t^{Max,0.983}) \\
&+ s_7(\theta_t^{max,0.983}) + \gamma \Gamma_t + s_8(t) + \varepsilon_t
\end{aligned}
$$

- Modèle obtenu à l'aide de méthode stepwise à dire d'un expert
- Long, fastidieux, probablement non optimal et **difficilement généralisable à d'autres jeux de données**

Comment automatiser la sélection de variables et l'estimation des modèles ?

Modélisation de la consommation nationale

$$
\begin{aligned}
Y_t =& s_1(Toy_t) + \sum_{i=1}^{9}\sum_{j=1}^{3} \alpha_{i,j} \mathbb{1}_{DayType_t=i} \mathbb{1}_{Offset_t=j} \\
&+ \sum_{i=1}^{12} \beta_i T_{t-6} \mathbb{1}_{Month_t=i} + s_2(CC_t) + s_3(W_t) + s_4(T_t) \\
&+ s_5(\theta_t^{0.996}) + s_6(\theta_t^{Max,0.983}) \\
&+ s_7(\theta_t^{max,0.983}) + \gamma \Gamma_t + s_8(t) + \varepsilon_t
\end{aligned}
$$

- Modèle obtenu à l'aide de méthode stepwise à dire d'un expert
- Long, fastidieux, probablement non optimal et **difficilement généralisable à d'autres jeux de données**

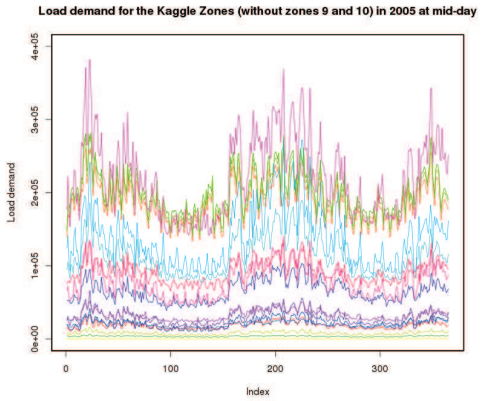
Comment automatiser la sélection de variables et l'estimation des modèles ?

Etude d'une maille locale

Poste source

- Liaison entre les réseaux de transport et de distribution
- Environ 2200
- Forts enjeux industriels

Compétition Kaggle GEFCom 2012

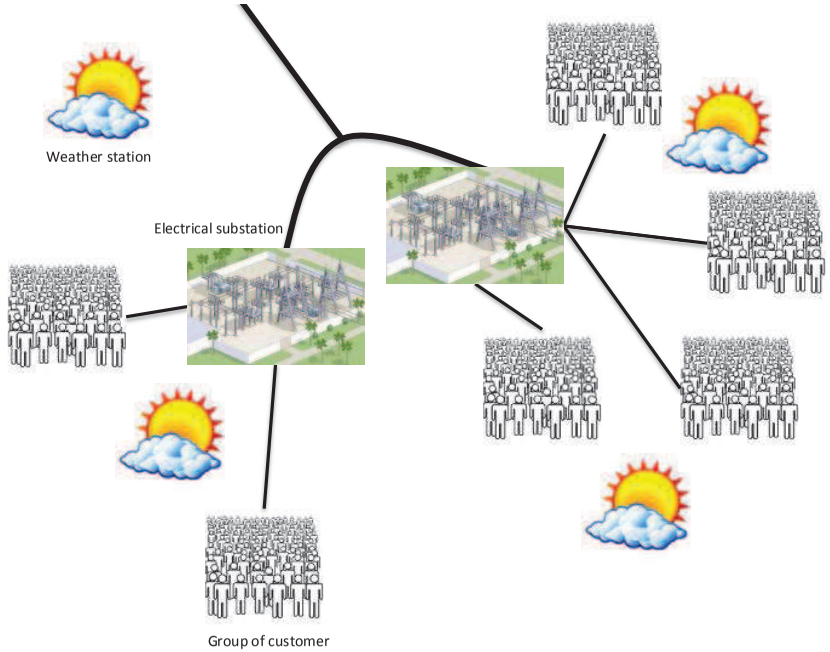


- Compétition présentée par Hong, Pinson et Fan [9]
- 17 zones électriques à prévoir, 11 stations météorologiques, aucune information géographique
- Version idéalisée mais riche en enseignements

Modélisation de la consommation locale

Modèle inspiré par Nedellec *et al.* [12] et par Goude *et al.* [7] :

$$Y_{t,j} = \sum_{q=1}^{11} m_q \mathbb{1}_{DayType_t=q} + h(Toy_t) + k(t) + \sum_{i \in \mathbf{S^j}} l_i(T_{t,S_i}) + \sum_{i \in \mathbf{S^j}} g_i(\theta_{t,S_i}) + \varepsilon_t$$

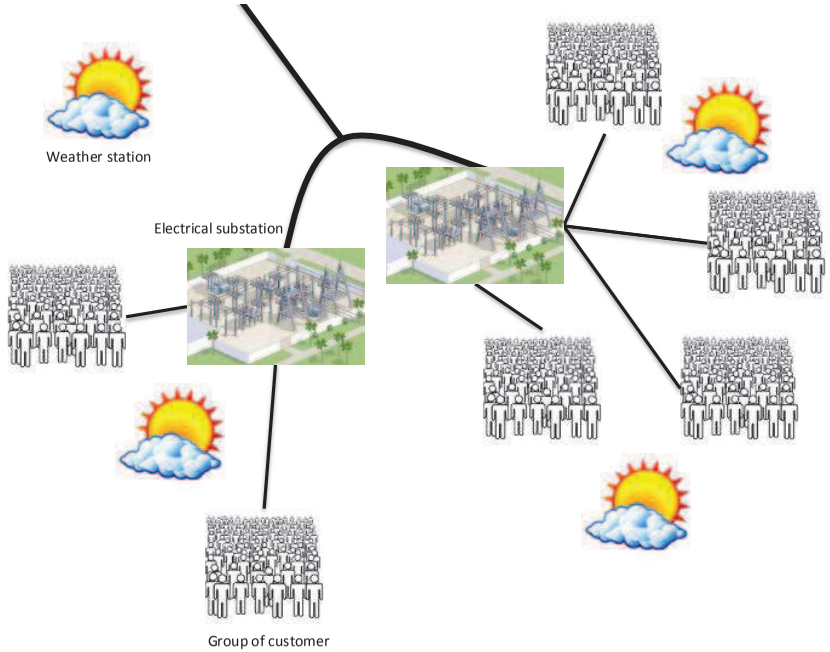


Comment procéder **automatiquement sans s'imposer un nombre fixe** de stations météorologiques ?

Modélisation de la consommation locale

Modèle inspiré par Nedellec *et al.* [12] et par Goude *et al.* [7] :

$$Y_{t,j} = \sum_{q=1}^{11} m_q \mathbb{1}_{DayType_t=q} + h(Toy_t) + k(t) + \sum_{i \in \mathbf{S^j}} l_i(T_{t,S_i}) + \sum_{i \in \mathbf{S^j}} g_i(\theta_{t,S_i}) + \varepsilon_t$$



Comment procéder **automatiquement sans s'imposer un nombre fixe** de stations météorologiques ?

Régression pénalisée

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + f_3(X_3) + \cdots + f_d(X_d) + \varepsilon$$

- Approximation : B-Splines (De Boor, 1978 [4])
- Estimation : P-Splines (Eilers et Marx, 1996 [5]) :
- Sélection : Group LASSO (Yuan et Lin, 2006 [14]) :

Régression pénalisée

$$Y = \beta_0 + \sum_{k=1}^{m} \beta_{1,k} B^q_{1,k}(X_1) + \sum_{k=1}^{m} \beta_{2,k} B^q_{2,k}(X_3) + \sum_{k=1}^{m} \beta_{3,k} B^q_{3,k}(X_3) + \cdots + \sum_{k=1}^{m} \beta_{d,k} B^q_{d,k}(X_d) + \varepsilon$$

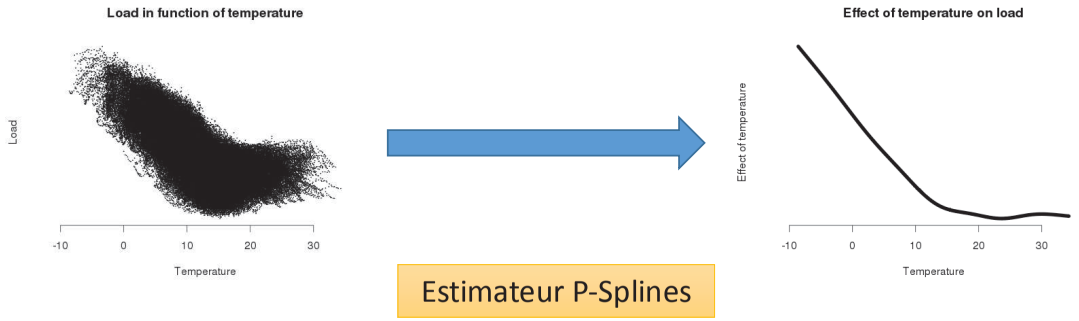
- Approximation : B-Splines (De Boor, 1978 [4])
- Estimation : P-Splines (Eilers et Marx, 1996 [5]) :
- Sélection : Group LASSO (Yuan et Lin, 2006 [14]) :

Régression pénalisée

$$Y = \beta_0 + \sum_{k=1}^{m} \beta_{1,k} B_{1,k}^q(X_1) + \sum_{k=1}^{m} \beta_{2,k} B_{2,k}^q(X_3) + \sum_{k=1}^{m} \beta_{3,k} B_{3,k}^q(X_3) + \cdots + \sum_{k=1}^{m} \beta_{d,k} B_{d,k}^q(X_d) + \varepsilon$$

- Approximation : B-Splines (De Boor, 1978 [4])
- Estimation : P-Splines (Eilers et Marx, 1996 [5]) :

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \sum_{k=1}^{m} \beta_{j,k} B_{j,k}^q(X_{i,j}) \right)^2 + \sum_{j=1}^{d} \lambda_j \|D_k \beta_{\mathbf{j}}\|_2^2$$



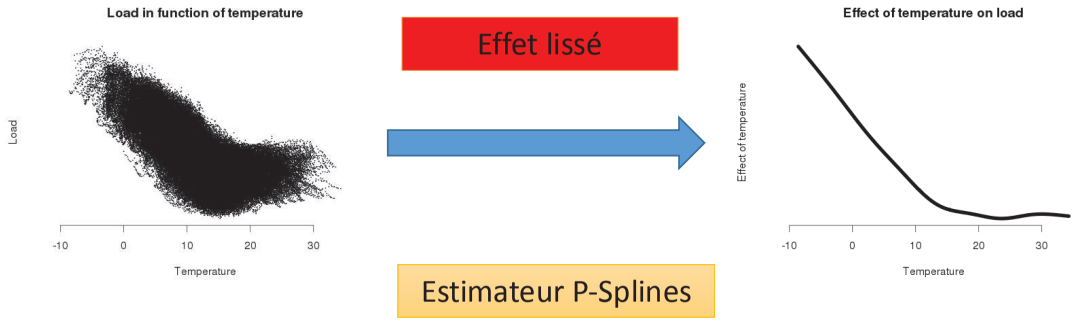
- Sélection : Group LASSO (Yuan et Lin, 2006 [14]) :

Régression pénalisée

$$Y = \beta_0 + \sum_{k=1}^{m} \beta_{1,k} B_{1,k}^q(X_1) + \sum_{k=1}^{m} \beta_{2,k} B_{2,k}^q(X_3) + \sum_{k=1}^{m} \beta_{3,k} B_{3,k}^q(X_3) + \cdots + \sum_{k=1}^{m} \beta_{d,k} B_{d,k}^q(X_d) + \varepsilon$$

- Approximation : B-Splines (De Boor, 1978 [4])
- Estimation : P-Splines (Eilers et Marx, 1996 [5]) :

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \sum_{k=1}^{m} \beta_{j,k} B_{j,k}^q(X_{i,j}) \right)^2 + \sum_{j=1}^{d} \lambda_j \|D_k \beta_{\mathbf{j}}\|_2^2$$



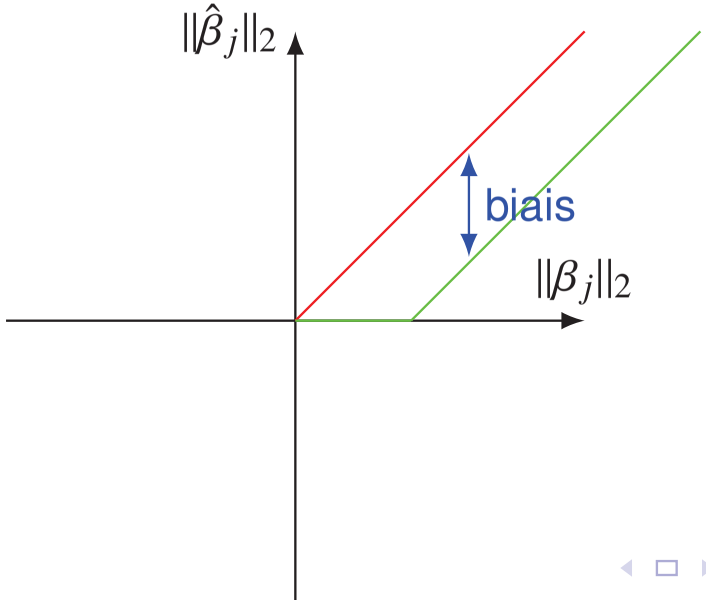
- Sélection : Group LASSO (Yuan et Lin, 2006 [14]) :

Régression pénalisée

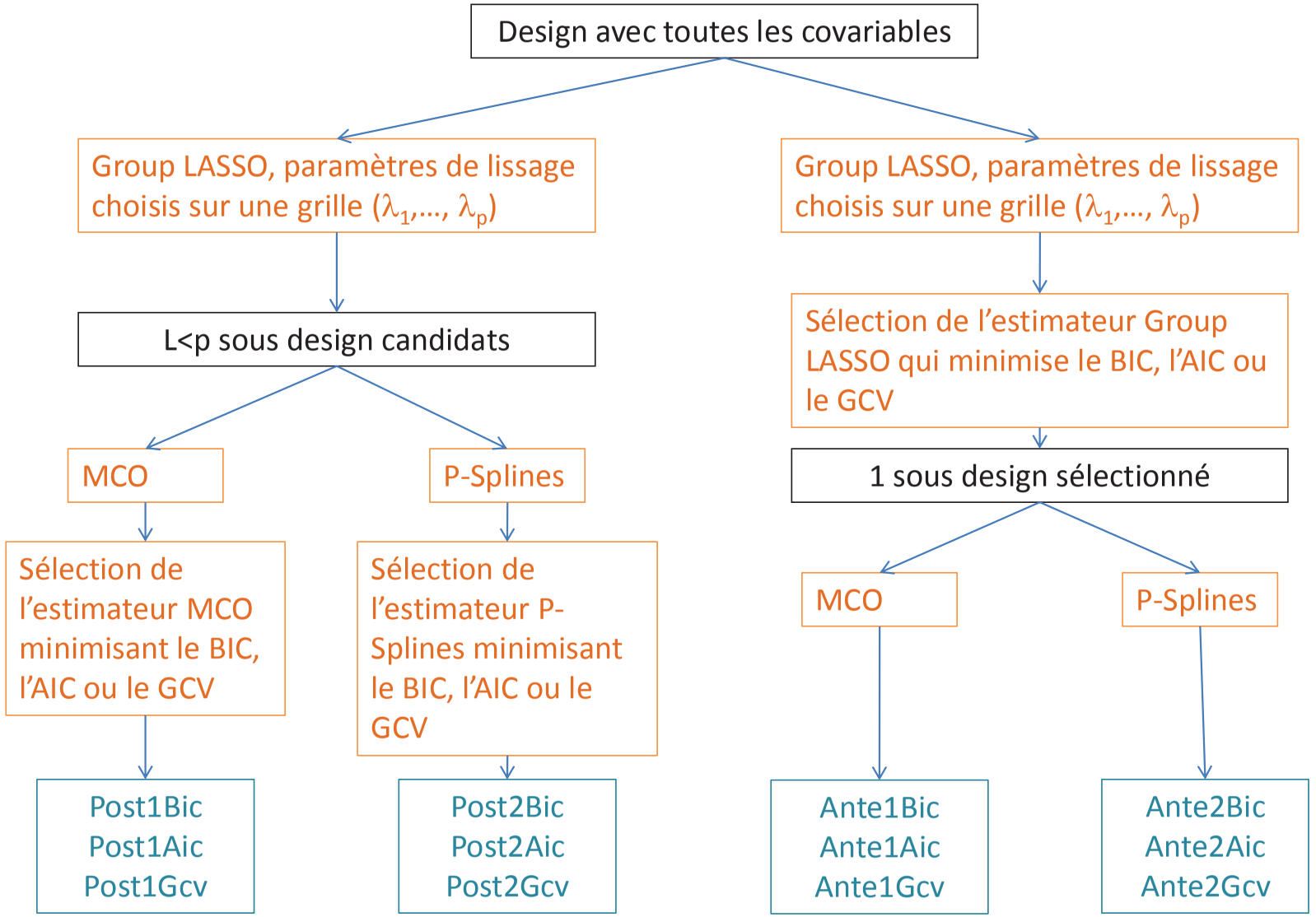
$$Y = \beta_0 + \cancel{\sum_{k=1}^{m} \beta_{1,k} B^q_{1,k}(X_1)} + \sum_{k=1}^{m} \beta_{2,k} B^q_{2,k}(X_3) + \cancel{\sum_{k=1}^{m} \beta_{3,k} B^q_{3,k}(X_3)} + \cdots + \sum_{k=1}^{m} \beta_{d,k} B^q_{d,k}(X_d) + \varepsilon$$

- Approximation : B-Splines (De Boor, 1978 [4])
- Estimation : P-Splines (Eilers et Marx, 1996 [5]) :
- Sélection : Group LASSO (Yuan et Lin, 2006 [14]) :

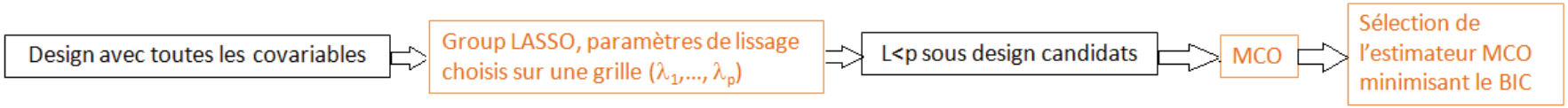
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \sum_{k=1}^{m} \beta_{j,k} B^q_{j,k}(X_{i,j}) \right)^2 + \lambda \sqrt{m} \sum_{j=1}^{d} \|\boldsymbol{\beta}_\mathbf{j}\|_2$$



Estimateurs en plusieurs étapes



Résultat asymptotique pour Post1Bic



Théorème

Sous nos hypothèses, soit $\lambda_n \sim \sqrt{n \log(m_n d)/m_n}$, *alors*

$$P(S_{\lambda_n} = S^*) \to 1,$$

et pour tout λ *tel que* $S_\lambda \neq S^*$ *et* $card(S_\lambda) \leq M$

$$P(BIC(\lambda) - BIC(\lambda_n) > 0) \to 1.$$

Voir Antoniadis, Goude, Poggi et Thouvenot [1] pour la démonstration. De plus, sous les hypothèses posées [1], l'estimateur **converge en erreur quadratique moyenne**.

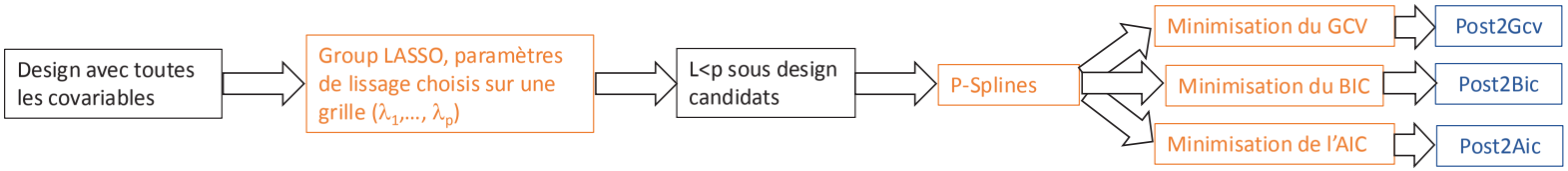
Portefeuille global d'un fournisseur

- Tendance retirée
- Modèle de base (Variables toujours présentes) :

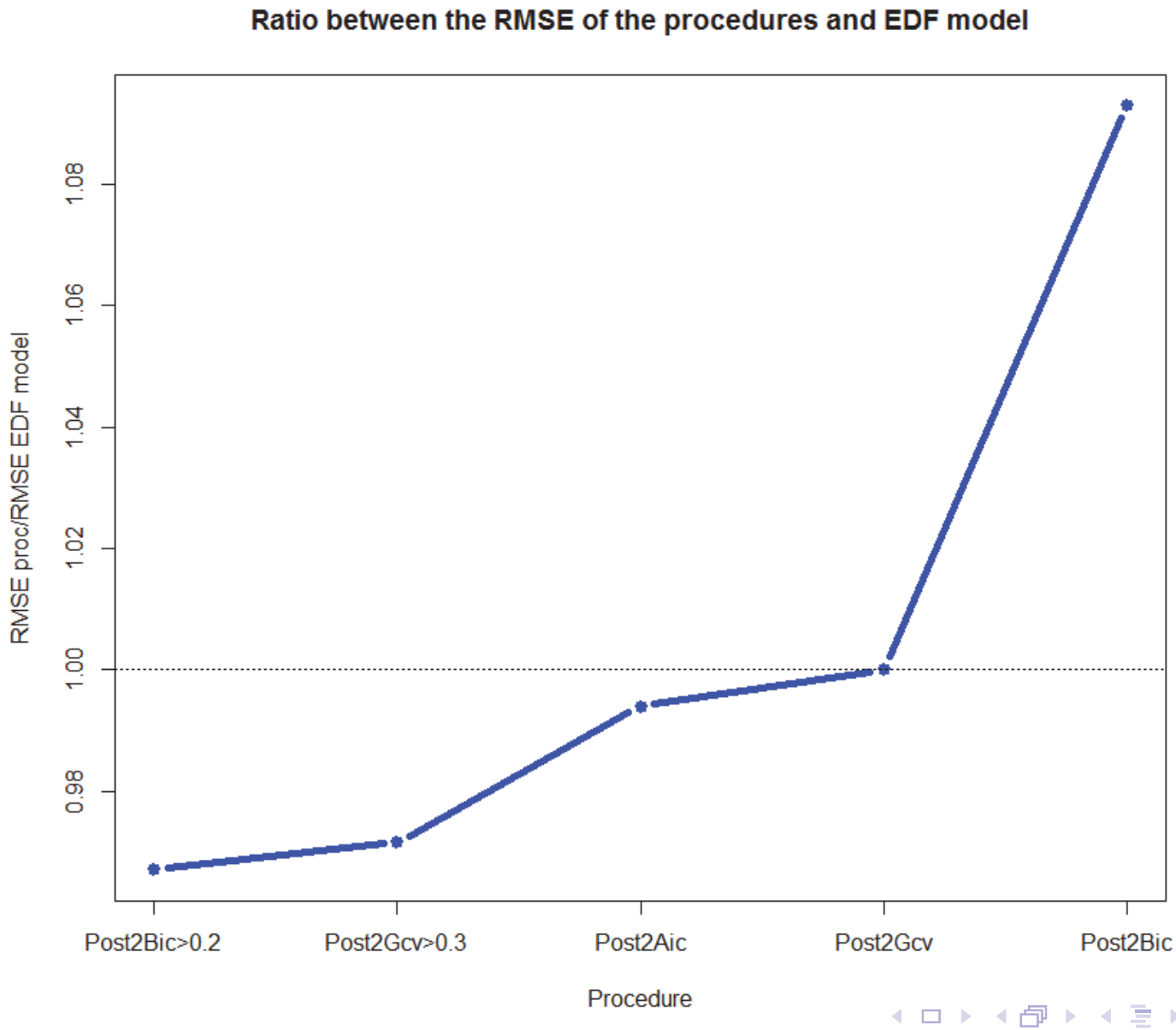
$$Y_t = s_1(Toy_t) + \sum_{i=1}^{9}\sum_{j=1}^{3} \alpha_{i,j} \mathbb{1}_{DayType_t=i} \mathbb{1}_{Offset_t=j}$$

$$+ \sum_{i=1}^{12} \beta_i T_{t-6} \mathbb{1}_{Month_t=1} + \varepsilon_t$$

- Objectif : sélectionner parmi 76 variables un modèle permettant de prédire la période test
- Prévision à un horizon d'un an (prévoir la période test entière)
- Trois procédures appliquées : Post2Gcv, Post2Bic, et Post2Aic



Performance à moyen terme



Application locale - Détails de l'application GEFCom 2012

- Modèles **moyen terme** appris en 2004-2007, testés sur les 6 premiers mois de 2008
- Modèles horaires inspirés de Nedellec *et al.* [12] :

$$Y_{t,j} = \sum_{q=1}^{11} m_q \mathbb{1}_{DayType_t=q} + h(Toy_t) + k(t) + \sum_{i \in \mathbf{S^j}} l_i(T_{t,S_i}) + \sum_{i \in \mathbf{S^j}} g_i(\theta_{t,S_i}) + \varepsilon_t$$

- A chaque station météorologique, des températures brutes et lissées sont associées
- Quelle stratégie adopter face au nombre de stations météorologiques ?
- Application de Post2Bic et Post2Aic

Passage aux 17 zones

Trois modèles de référence :

Modèle saturé

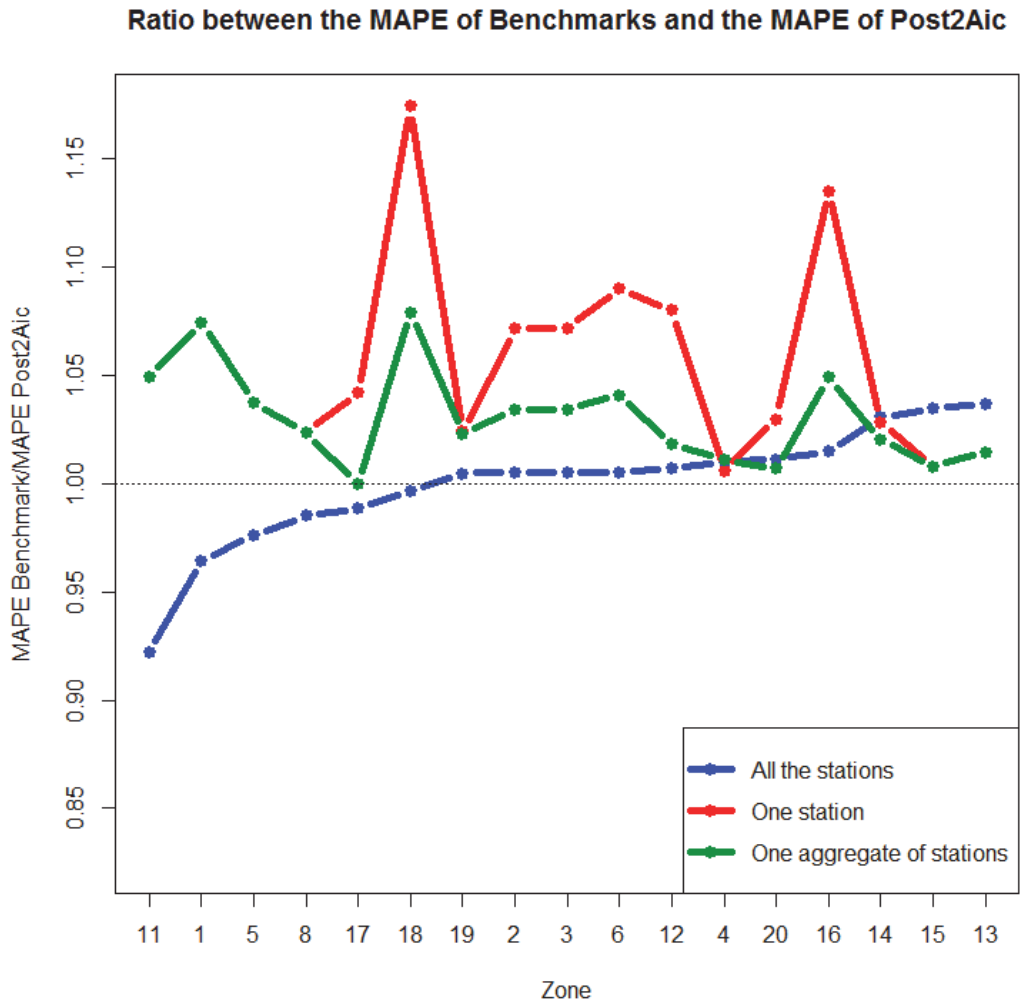
$$Y_{t,j} = \sum_{q=1}^{11} m_q \mathbb{1}_{DayType_t=q} + h(Toy_t) + k(t) + \sum_{b=1}^{11} l_b(T_{t,b}) + \sum_{b=1}^{11} g_b(\theta_{t,b}) + \varepsilon_t$$

Nombre fixe de stations météorologiques sélectionnées, s'inspire de Lloyd [11]

Modèle avec un signal de température

$$Y_{t,j} = \sum_{q=1}^{11} m_q \mathbb{1}_{DayType_t=q} + h(Toy_t) + k(t) + l(T_{t,j}) + g(\theta_{t,j}) + \varepsilon_t$$

- Sélection d'une station météorologique minimisant le MAPE sur un échantillon de validation : **nombre fixe** de stations météorologiques, s'inspire de Ben Taieb *et al.* [2]
- Sélection de type pas à pas de moyennes agrégées de températures : **nombre variable** de stations météorologiques, s'inspire de Hong *et al.* [10]



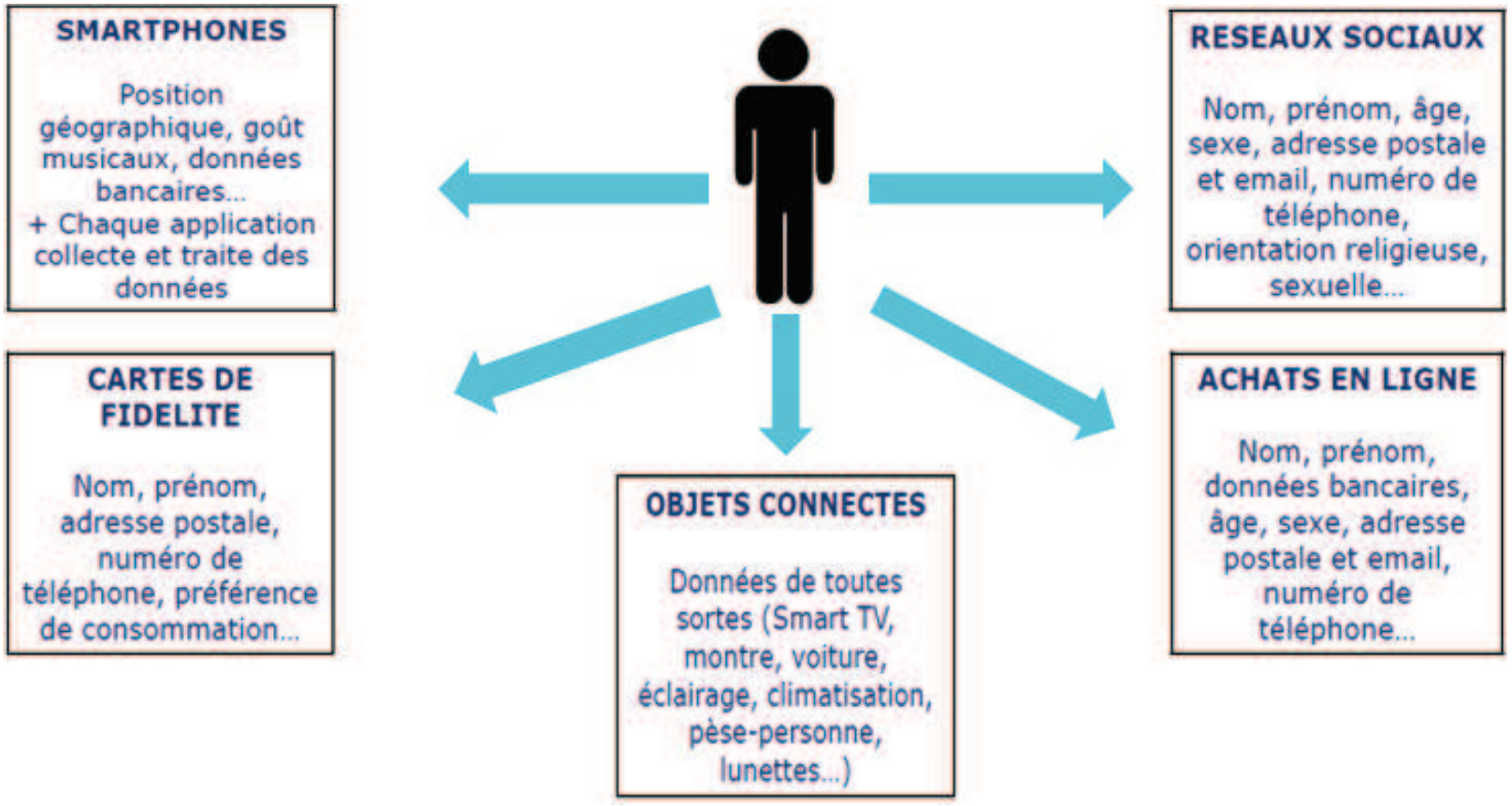
- Représentation des MAPE des modèles de référence relativement à celui de Post2Aic
- Valeur 1 $\Rightarrow$ le MAPE du modèle considéré = MAPE de Post2Aic
- Valeur $< 1 \Rightarrow$ le modèle de référence a des meilleures performances prédictives d'après le critère MAPE

Apport des Mathématiques

- Modélisation et son approximation
- Algorithmes d'optimisation pour estimer les coefficients associés à la modélisation
- Obtention de méthodes ayant automatiquement des performances équivalentes à des méthodes manuelles
- Assurances théoriques sur la justesse en sélection et en estimation des estimateurs obtenus
- Mesure de la qualité des prévisions

Sommaire

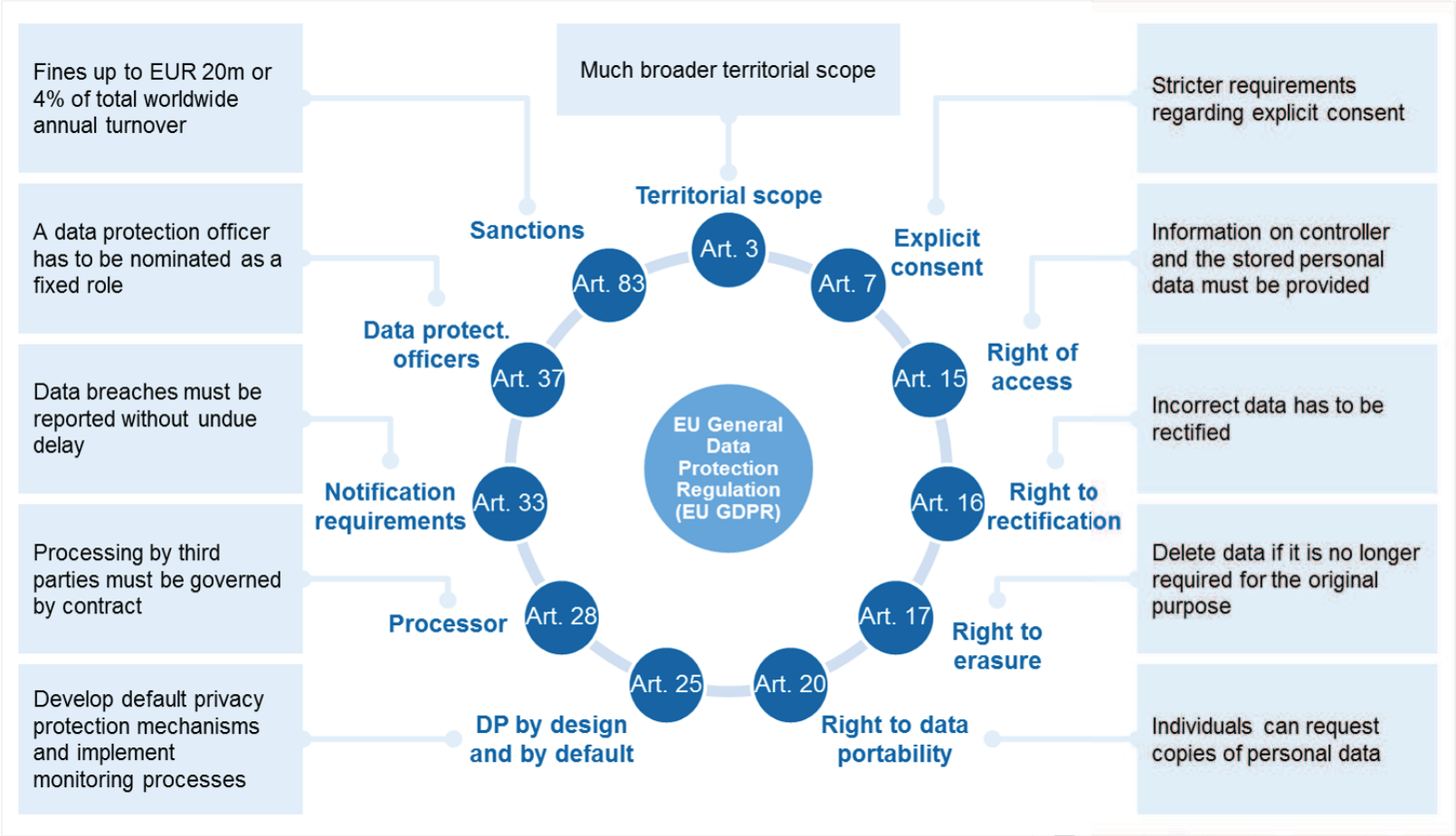
Données personnelles



Les données, une mine d'or

- Transformation numérique/digitale
  - Désilotage des données
  - Nouveaux marchés
- Open Data
  - Forte exigence réglementaire
  - Favorise l'ouverture et la circulation des données et du savoir
  - Permet de garantir un environnement numérique ouvert
  - Doit être respectueux de la vie privée
- Big Data
  - Permet de nouvelles méthodes de croisement de données et donc de ré-identifier des individus

Cadre réglementaire - RGPD



Protection des données personnelles ?

Protéger les données personnelles, c'est . . .

- Empêcher un individu d'être isolé
- Empêcher la possibilité de corréler les informations sur un individu de plusieurs jeux de données
- Empêcher d'obtenir des informations sur les individus grâce à des variables exogènes

. . .et concerne

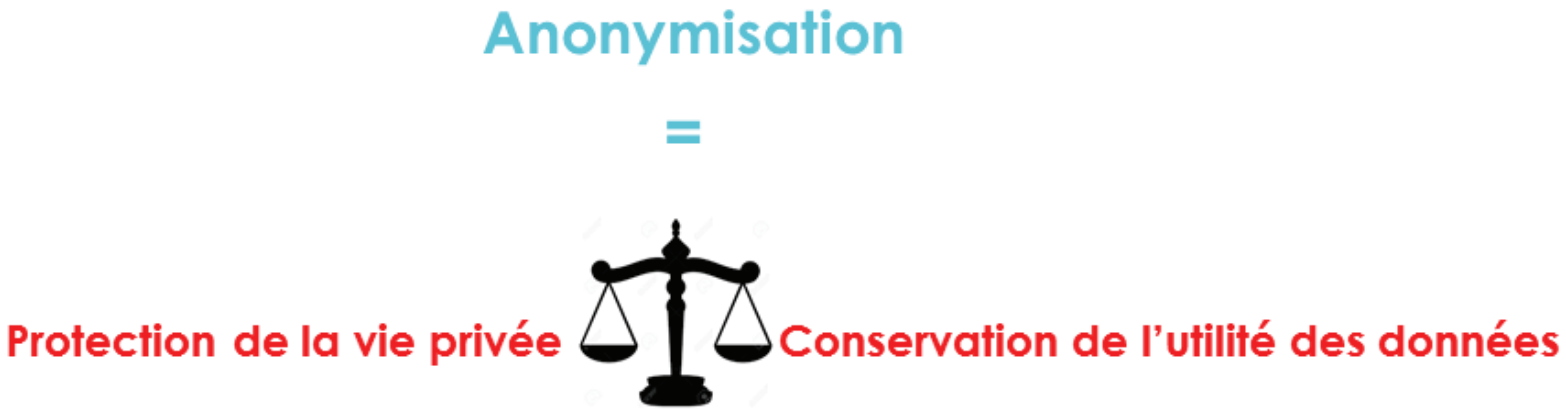
- des informations identifiant directement ou quasi directement des individus (nom, numéro de téléphone, adresse, etc.)
- des informations sensibles portant sur les opinions politiques ou religieuses, la santé et les habitudes, etc. des individus

Protéger les données par la pseudonymisation

La pseudonymisation

- consiste à chiffrer ou supprimer les éléments directement identifiants. . .
- . . .et est souvent **insuffisante** car des élements peuvent être croisées ou des éléments extérieures pour permettre la ré-identification
- Par exemple AOL en 2006
  - Mise en Open Data de 20 millions de requêtes de recherche dont l'identifiant des utilisateurs avaient été chiffrés
  - Par recoupement géographique et utilisation d'informations se retrouvant avec les recherches effectuées (centre d'intérêt, âge, etc.), certains utilisateurs ont été ré-identifiés
  - Le CTO et plusieurs salariés d'AOL renvoyés suite à cette mise en Open Data

Anonymisation : Réalisation d'un compromis utilité/protection



L'anonymisation est fortement dépendante

- De la typologie des données ⇒ *données d'une table médicale ≠ données de réseaux sociaux par exemple*
- Des traitements futures ⇒ *besoin de données fines individuellement ou seulement de groupes d'individus ?*
- Du temps ⇒ *de nouveaux jeux de données et des nouvelles méthodes permettent de ré-identifier*

L'anonymisation, un sujet complexe

- Pas de solution universelle
  - Réalisation d'un compromis utilité/protection
  - Fortement dépendante du **type de données**, du **type d'application** et du **temps**
- Pas de risque zéro avec l'anonymisation
  - Implique l'étude du risque ré-identification
  - Risque résiduelle doit être faible par rapport aux bénéfices
- Procédé global
  - Comprenant des composants de **dé-identification**, de **ré-identification** et d'**évaluation de la qualité des données**
  - Auditable

Généralité sur la dé-identification - Trois familles de techniques

- L'anonymisation peut se partager en trois familles de techniques
  - Transformer les données pour avoir des individus fictifs
  - Agréger et généraliser les individus pour fournir des agrégats symbolisant des groupes d'individus
  - Générer entièrement des données fictives
- Pour plus de protection, les techniques peuvent être combinées

Généralité sur la dé-identification - Techniques de dé-identification

- Techniques de permutation et de puzzling
  - Approches déstructurant et modifiant l'architecture des données
  - Individus ainsi crées peuvent être irréalistes
- Application de bruits additifs ou multiplicatifs
  - Ajout de bruit : ré-identification par filtrage, MAP, etc.
  - Bruits multiplicatifs : ré-identification possible lorsque l'attaquant connait un sous-échantillon
  - Confidentialité différentielle
- Généralisation des similaires
- Microagrégation

Généralité sur la dé-identification - Techniques de dé-identification

- Techniques de permutation et de puzzling
  - Approches déstructurant et modifiant l'architecture des données
  - Individus ainsi crées peuvent être irréalistes
- Application de bruits additifs ou multiplicatifs
  - Ajout de bruit : ré-identification par filtrage, MAP, etc.
  - Bruits multiplicatifs : ré-identification possible losrque l'attaquant connait un sous-échantillon
  - **Confidentialité différentielle**
- **Généralisation des similaires**
- **Microagrégation**

Confidentialité différentielle - Principe [?]

Notons ε un réel positif, M un mécanisme d'anonymisation, Im_M l'image de M. Alors M est dit ε-différentiellement confidentielle si pour tout D_1 et D_2 jeux de données adjacents et pour tout $S \subset Im_M$:

$$P\{M(D_1) \in S\} \leq \exp(\varepsilon) P\{M(D_2) \in S\}$$

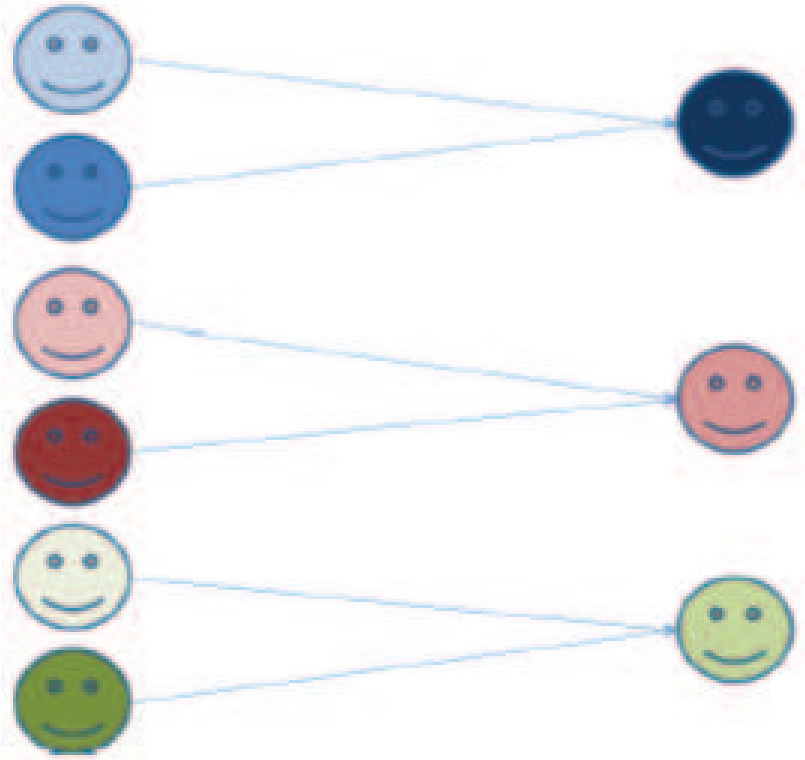
- La définition assure que la présence ou l'absence d'un individu n'influencera pas significativement la sortie du mécanisme
- Plus ε est petit, plus il est compliqué de distinguer la présence ou l'absence d'un individu
- Lorsqu'il est nul, appliqué aux deux jeux de sortie adjacents, le mécanisme a la même probabilité de produire la sortie S. On ne peut donc plus distinguer la présence ou l'absence de l'individu différent

Généralisation des similaires - Principe

- K-anonymisation [**?**] : un jeu de données est K-anonyme si au moins K enregistrements existent pour chaque combinaison de quasi-identifiant
- L-Diversité [**?**] : un jeu de données repecte la L-Diversité si au moins K enregistrements existent pour chaque combinaison de quasi-identifiant et si dans chaque classe, la variable sensible prend au moins L valeurs différentes
- T-Proximité [**?**] : un jeu de données respecte la L-Diversité si au moins K enregistrements existent pour chaque combinaison de quasi-identifiant et si la répartition des valeurs sensibles respectent la répartition de celles-ci dans la population totale

Micro-agrégation - Principe

- Objectif : traiter différents types de variables (continues, catégorielle, etc.) et agréger les individus de manière la plus optimale possible.
- Deux phases
  - Partitionnement des données en des sous espaces de taille k au minimum
  - Application d'un opérateur d'agrégation
- Paramètres
  - Le choix de la taille k minimum
  - L'algorithme et les distances utilisés pour partitionner les données
  - Le choix de la méthode d'agrégation
- Risques de ré-identification : déduction d'informations probables sur un individu
- Utilité des données : métriques d'homogéneité (indices de Bouldin-Davies, Silhouette, etc.) et indicateurs de qualité globaux (RMSE, MAPE, etc.)



Ré-identification

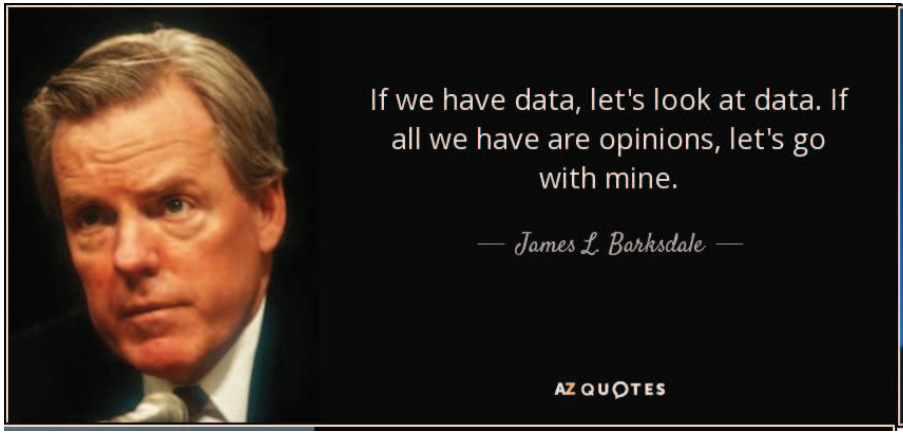
- Plusieurs scénarios de ré-identification sont possibles
  - Isoler et identifier un ou quelques individus
  - Identifier l'ensemble des individus
  - Déduire des informations sensibles sur un ou plusieurs individus
- Plusieurs catégories de méthodes de ré-identification
  - Utilisation d'un jeu de données non anonymisé pour ré-identifier un jeu de données anonymisé
    - Par exemple, utilisation d'un réseau social non anonymisé pour attaquer un réseau social anonymisé
  - Utilisation d'informations présentes dans le jeu de données anonymisée
    - Par exemple, le contenu pour attaquer un réseau social anonymisé
- Utilisation de méthodes classiques de Machine Learning et d'optimisation combinatoire

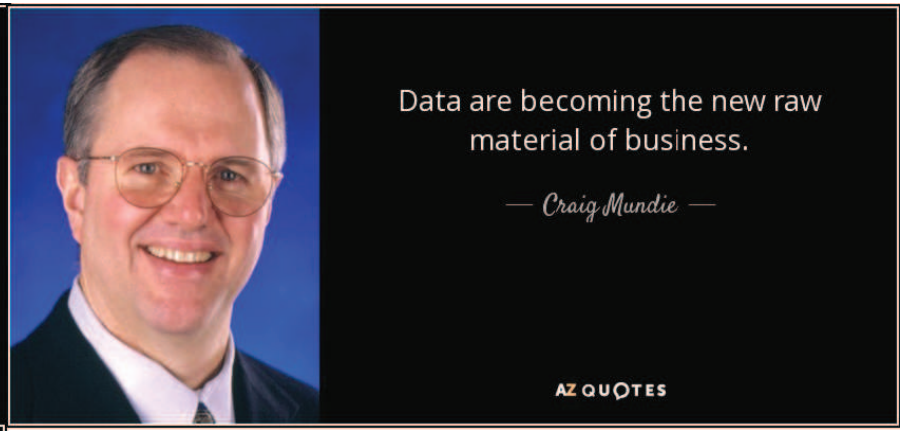
Apport des Mathématiques

- Cryptologie et chiffrement
- Pré-traitement des données
- Etude de Probabilité
- Algorithmes de partitionnement et de généralisation
- Métriques de perte d'information
- Mesure de la "Privacy"

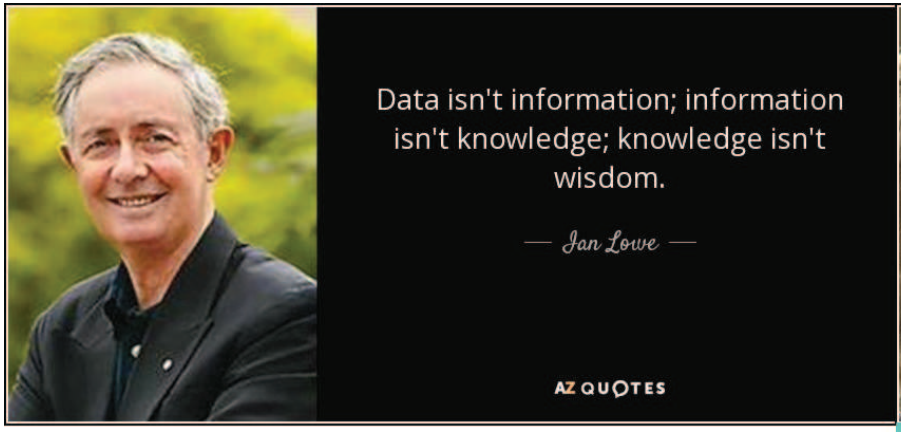
Sommaire

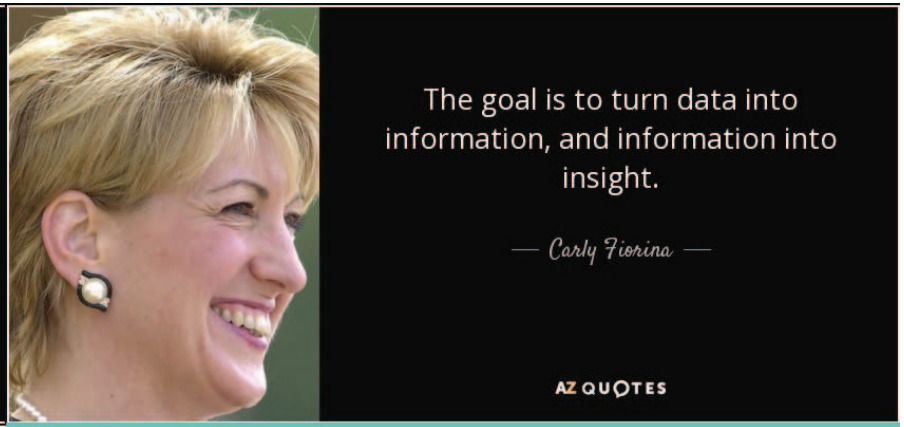
Des données au Machine Learning

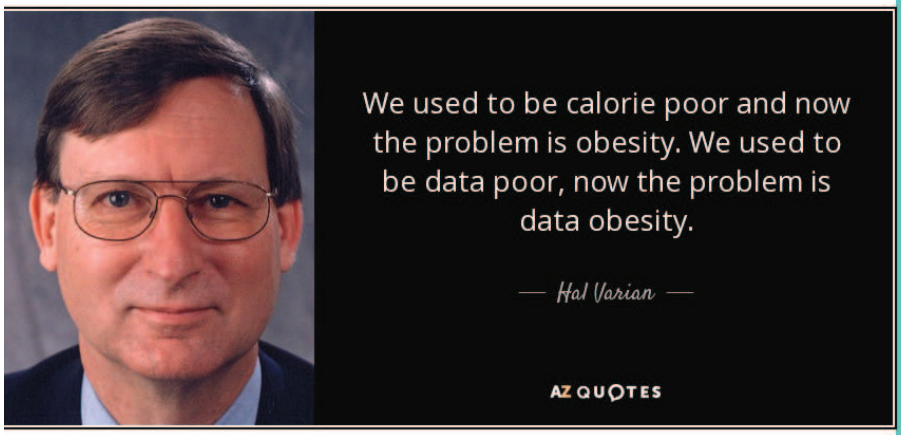


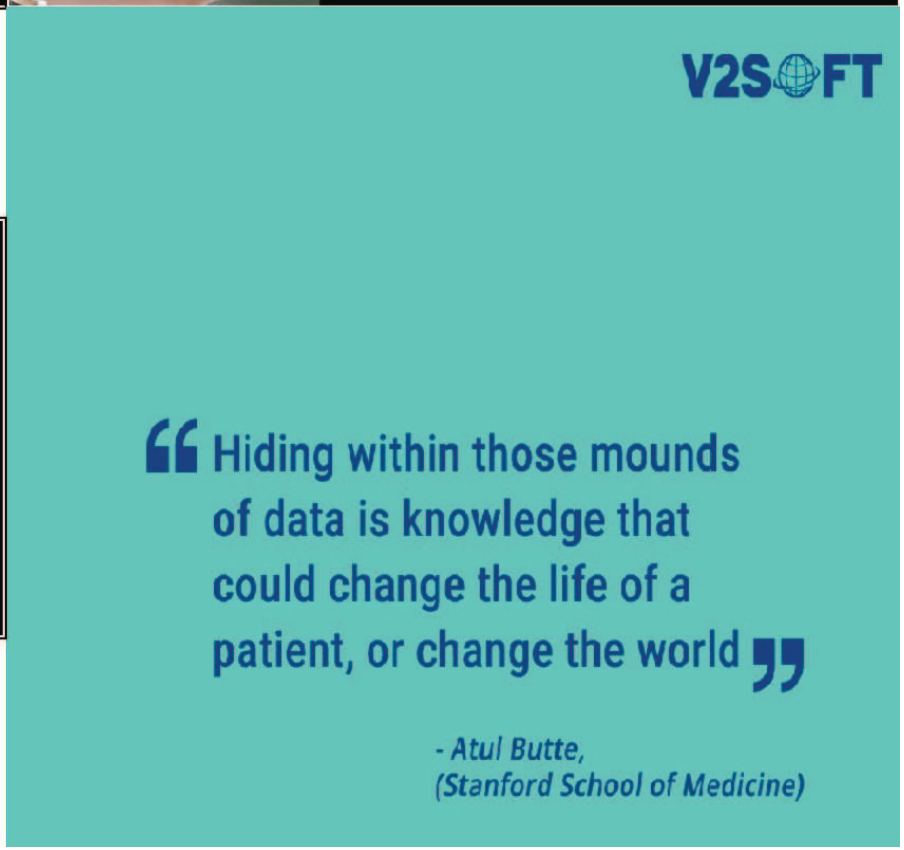


Des données au Machine Learning

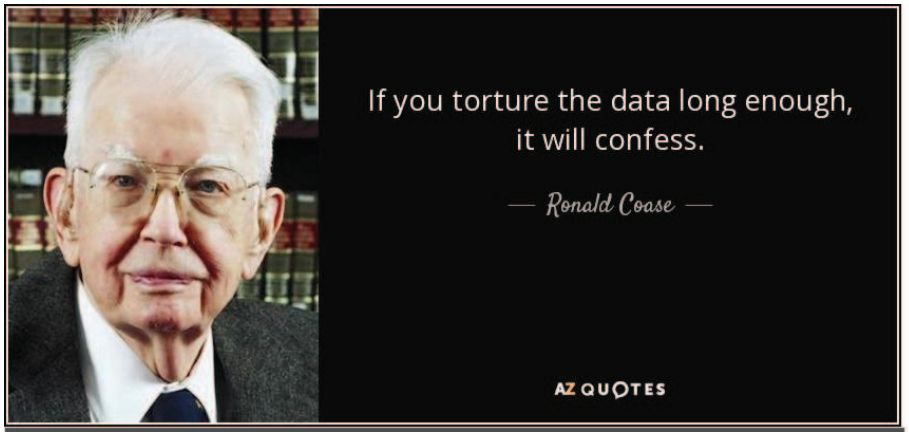


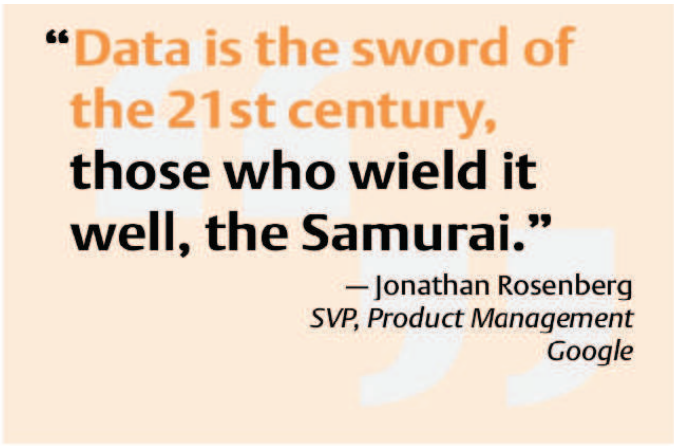


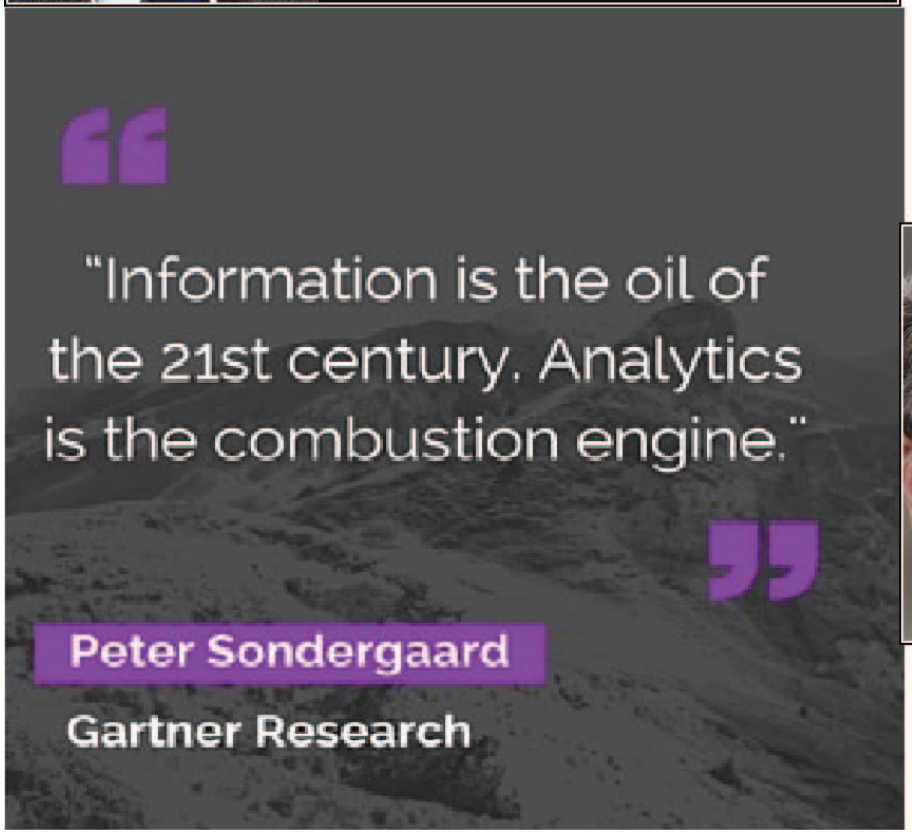


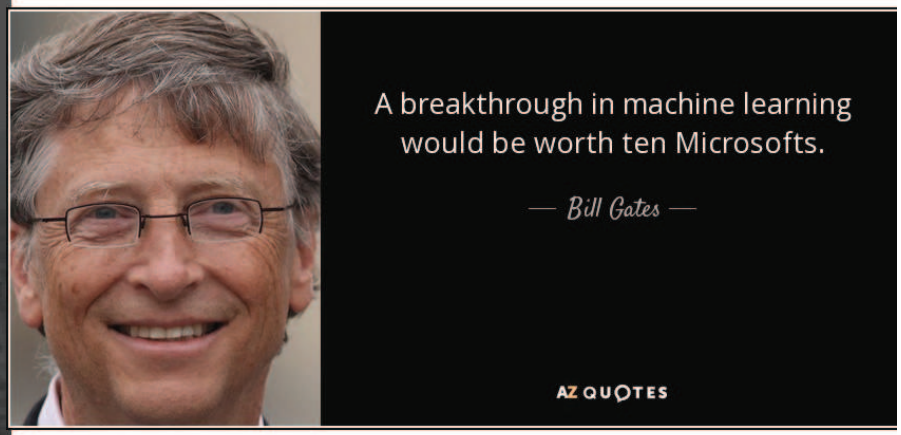


Des données au Machine Learning

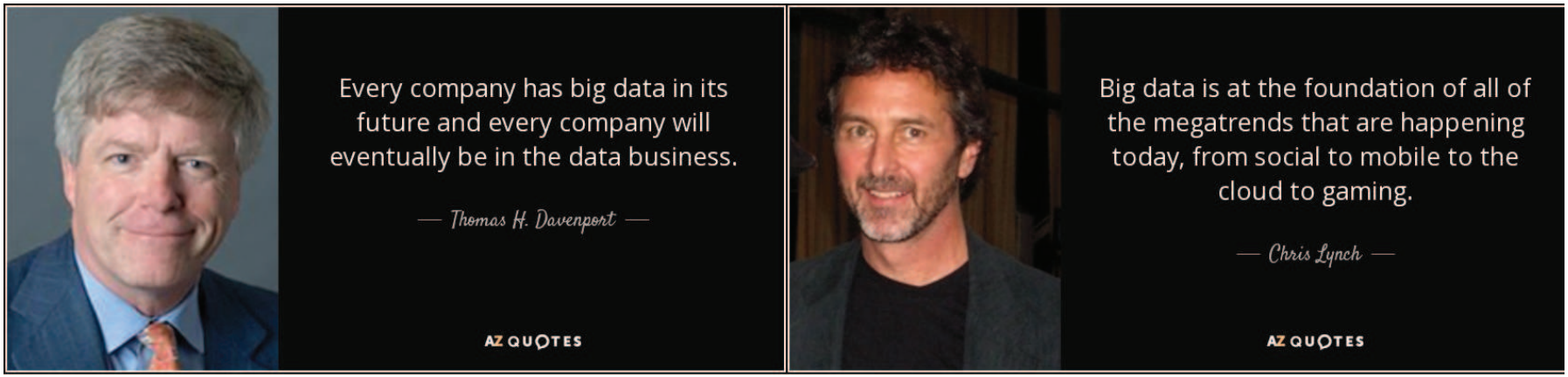




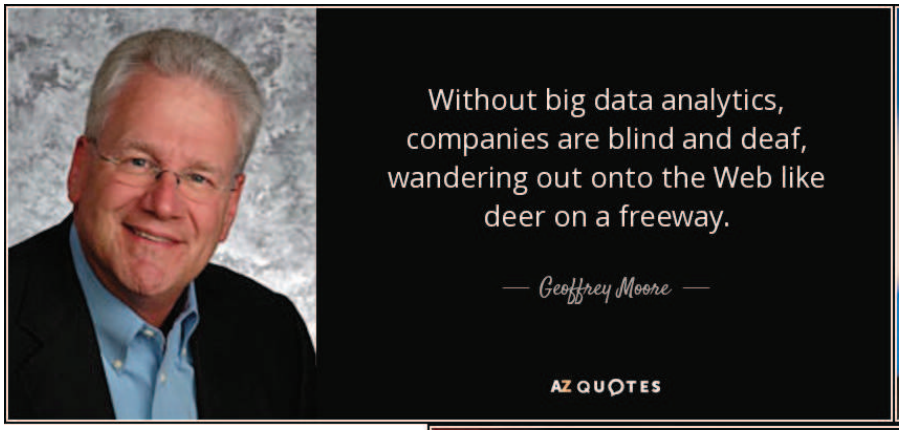


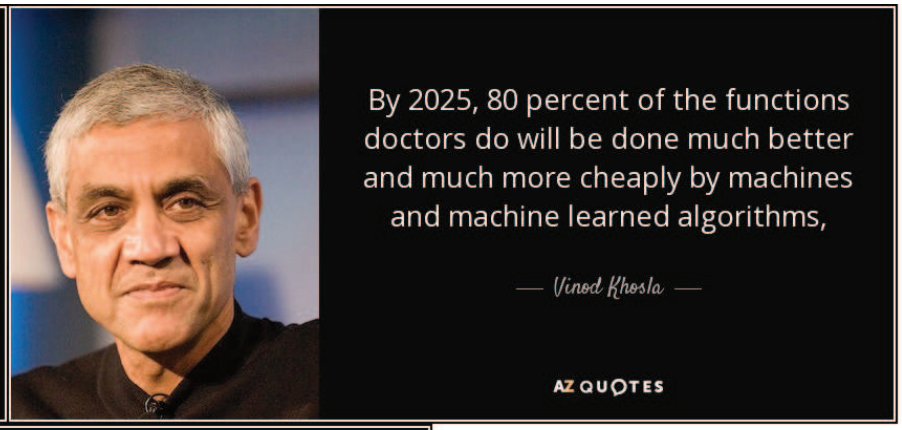


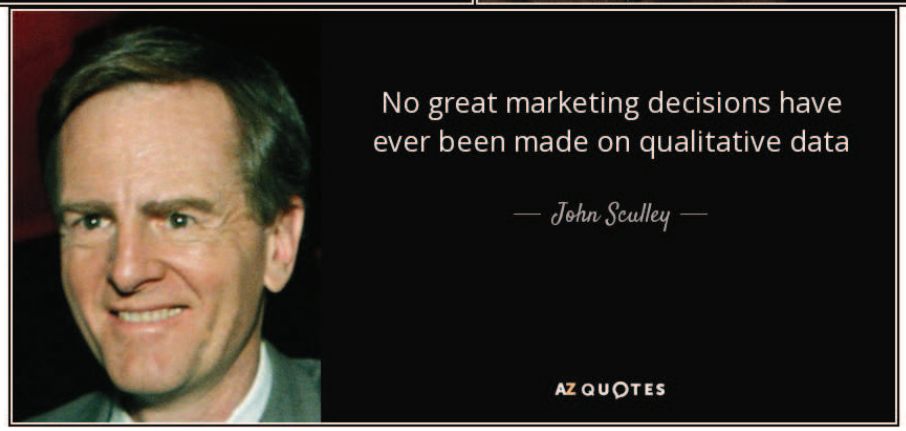
Des données au Machine Learning



Des données au Machine Learning

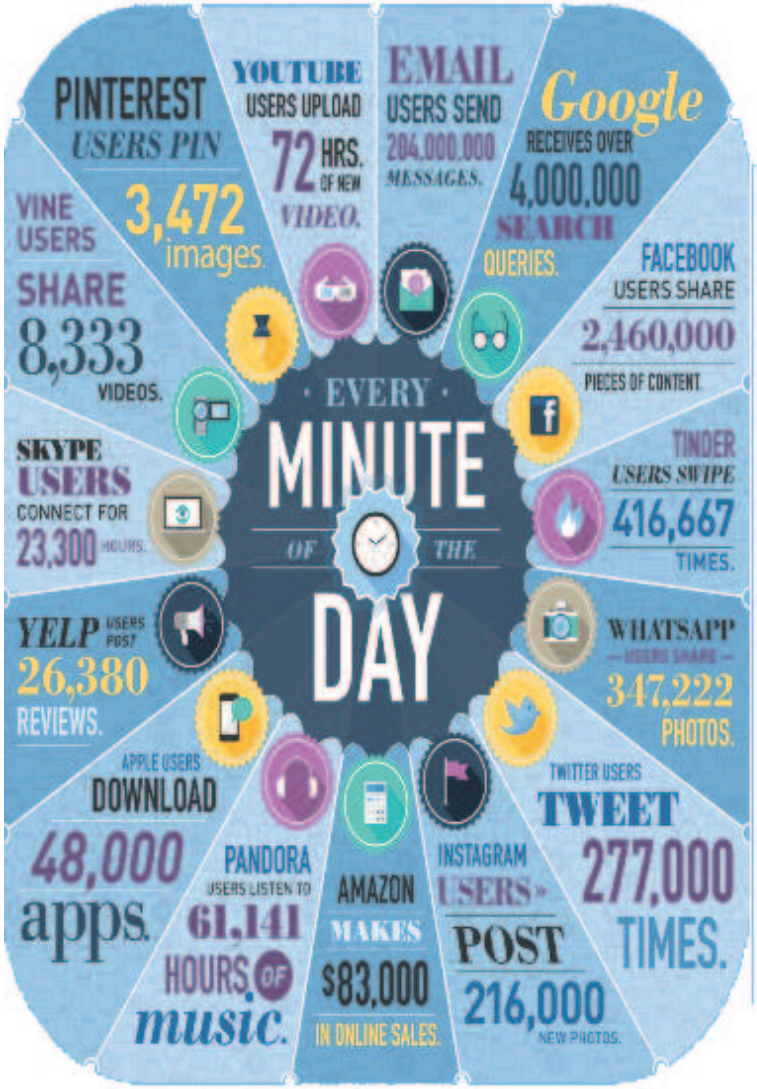




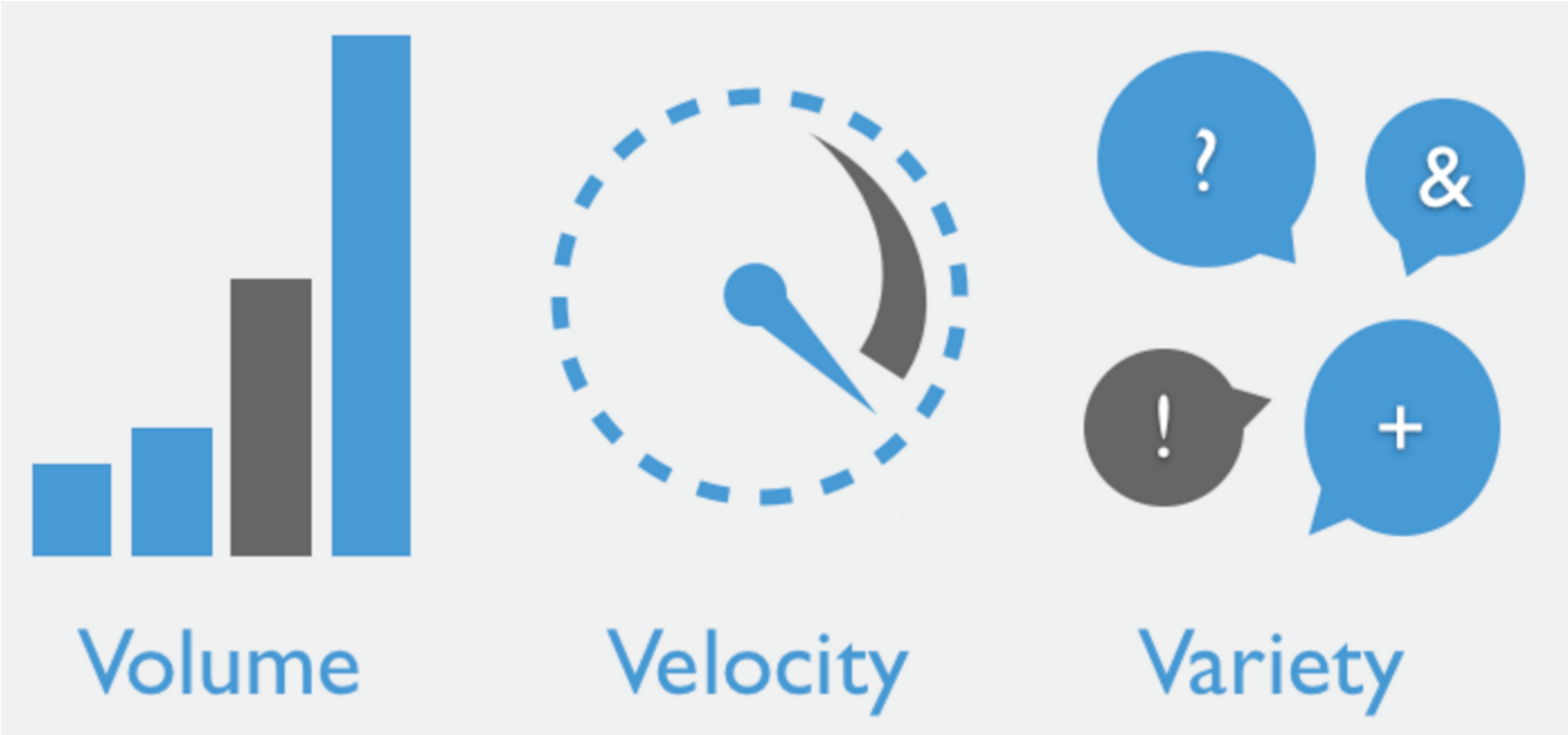


Des données au Machine Learning

Le Big Data : Pourquoi ?



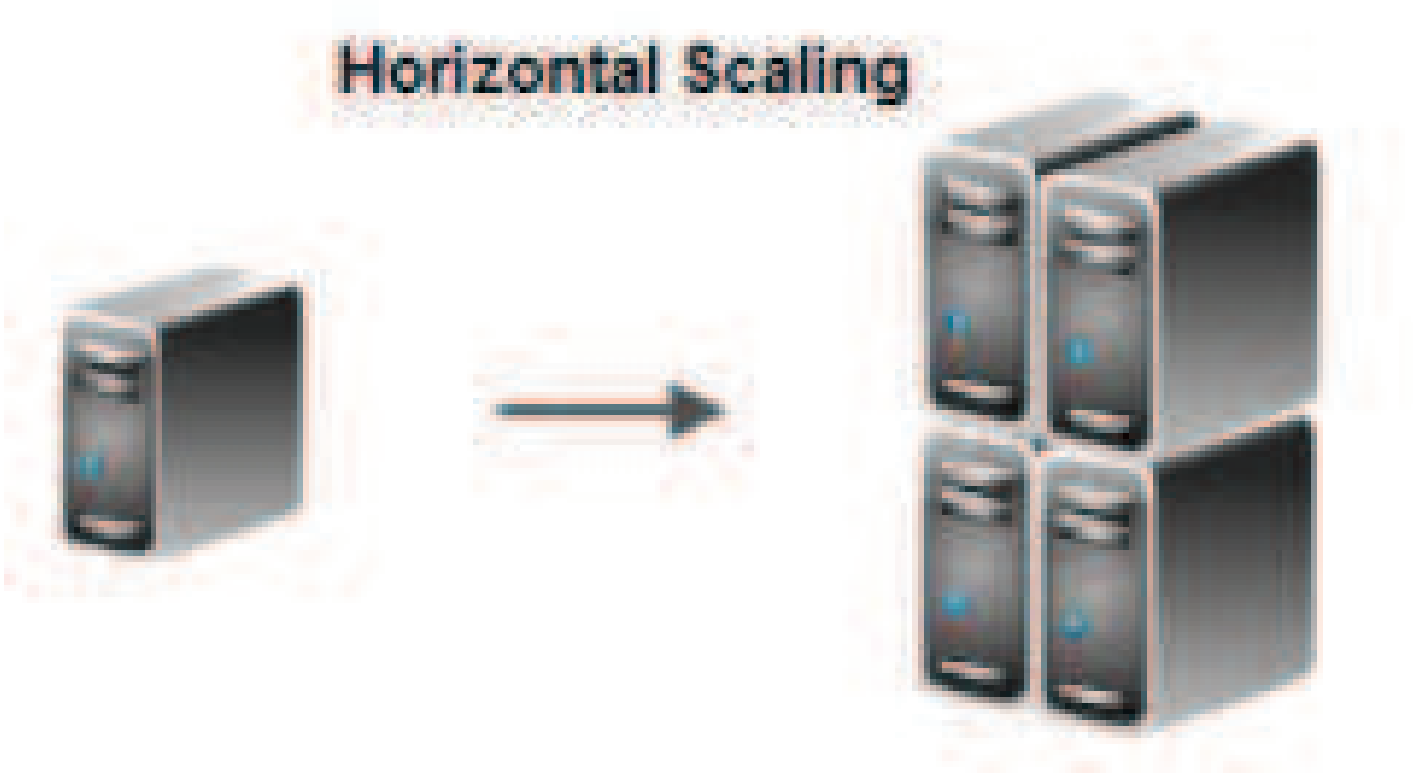
Le Big Data : les 3 V



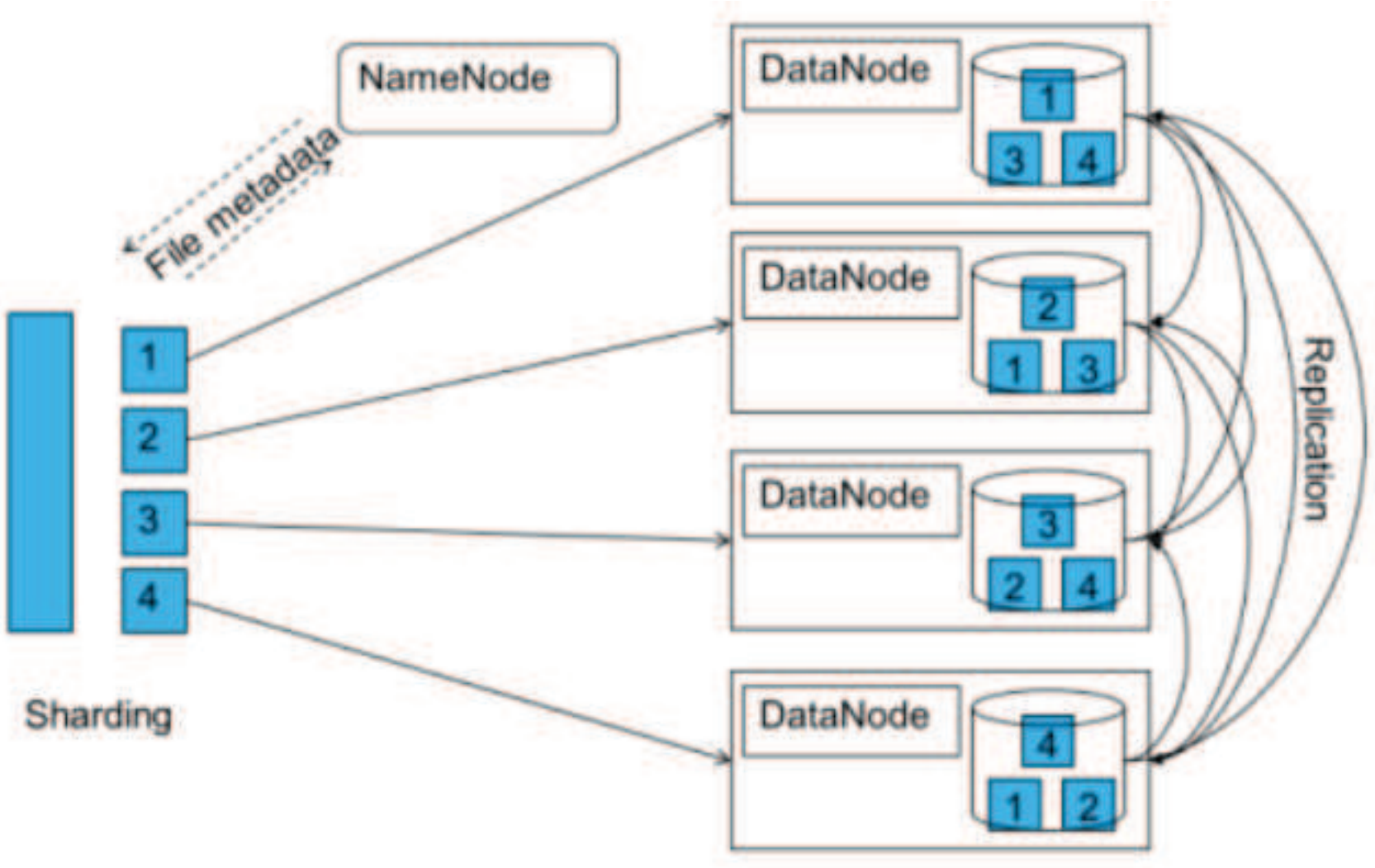
Le Big Data n'est pas . . .

Utilisation de technologies Big Data : scaling horizontal

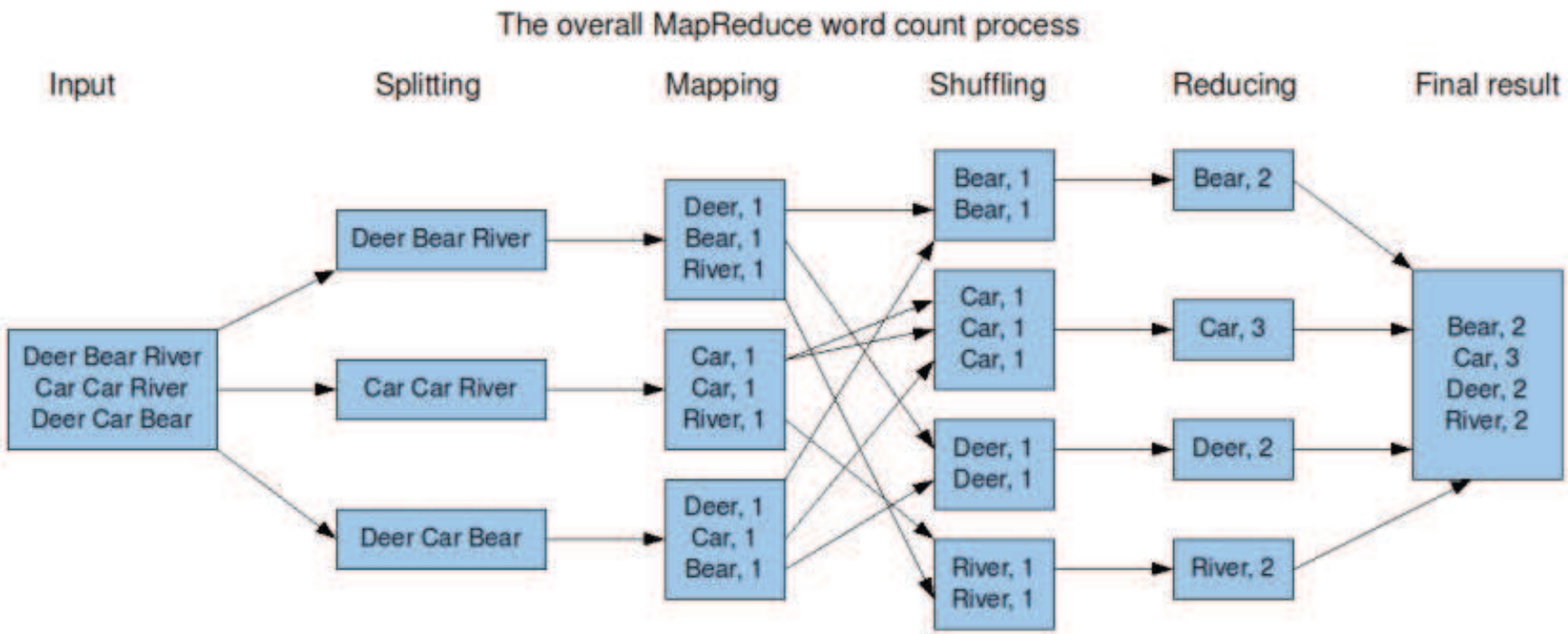
Lorsque le volume des données augmente, augmentation du nombre de serveurs ⇒ plus souple et moins couteux qu'un hyper calculateur



Système de fichiers distribués



Map Reduce


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Apport des Mathématiques

- Le Big Data met en jeu un grand nombre de domaines des mathématiques appliquées : statistiques au sens large, traitement du signal et des images, optimisation convexe et non convexe, probabilités appliquées (graphes aléatoires, algorithmes stochastiques...).
- Toutes les techniques statistiques doivent être passées à l'échelle. Par exemple, si on considère la recherche de la solution de l'estimation des MCO d'une régression linéaire sous hypothèses classiques
  - Classiquement un produit matricielle impliquant une inversion de matrice $\Rightarrow$ Comment l'obtenir dans un contexte de données distribuées ?
  - Utilisation d'un algorithme de descente de gradient stochastique par exemple en contexte Big Data

Sommaire

- Les Mathématiques appliquées aux données, utilisant notamment des Statistiques et des Probabilités appliquées, des algorithmes d'optimisation, de recommendation, de traitement du signal, etc., sont de plus en plus demandées
  - Dans tous les secteurs : banque et assurance, industrie lourde et pharmaceutique, Transport, Energie, etc.
  - Pour de nombreuse applications : cyber sécurité, connaissance client (comportement et ciblage), maintenance, dimensionnement de réseau, prévision de série temporelle, gestion de risque, etc.
- Nombreuse sources de recrutement : ENSAE, ENSAI, ISUP, Université, Masters spécialisés d'école d'ingénieur, etc.

[1] A. Antoniadis, Y. Goude, J.-M. Poggi, and V. Thouvenot, *Sélection de variables dans les modèles additifs avec des estimateurs en plusieurs étapes*, tech. report.
https://hal.archives-ouvertes.fr/hal-01116100,.

[2] S. Ben Taieb and R. Hyndman, *A gradient boosting approach to the kaggle load forecasting competition*, International Journal of Forecasting, 30 (2014), pp. 382 – 394.

[3] E. Bruhns, G. Deurveilher, and J.-S. Roy, *A non-linear regression model for mid-term load forecasting and improvements in seasonality*, The 15th Power Systems Computation Conference, Liege, Belgium, (2005).

[4] C. De Boor, *A practical guide to splines*, Springer-Verlag New York, 1978.

[5] P. Eilers and B. Marx, *Flexible smoothing with b-splines and penalties*, Statistical Science, 11 (1996), pp. 89–121.

[6] S. Fan and R. Hyndman, *Short-term load forecasting based on a semi-parametric additive model*, Power Systems, IEEE Transactions on, 27 (2012), pp. 134–141.

[7] Y. Goude, R. Nedellec, and N. Kong, *Local short and middle term electricity load forecasting with semi-parametric additive models*, Smart Grid, IEEE Transactions on, 5 (2014), pp. 440–446.

[8] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*, London : Chapman & Hall, 1990.

[9] T. Hong, P. Pinson, and S. Fan, *Global energy forecasting competition 2012*, International Journal of Forecasting, 30 (2014), pp. 357 – 363.

[10] T. Hong, P. Wang, and L. White, *Weather station selection for electric load forecasting*, International Journal of Forecasting, (To appear).

[11] J. Lloyd, *Gefcom2012 hierarchical load forecasting : Gradient boosting machines and gaussian processes*, International Journal of Forecasting, 30 (2014), pp. 369–374.

[12] R. Nedellec, J. Cugliari, and Y. Goude, *Gefcom2012 : Electric load forecasting and backcasting with semi-parametric models*, International Journal of Forecasting, 30 (2014), pp. 375 – 381.

[13] A. Pierrot and Y. Goude, *Short-term electricity load forecasting with generalized additive models*, Proceedings of ISAP power, (2011), pp. 593–600.

[14] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society, Series B, 68 (2006), pp. 49–67.

Merci de votre attention