esiea
ECOLE D'INGENIEURS DU MONDE NUMERIQUE

TECHNIQUES D'EXTRACTION DE CONNAISSANCES (DATA MINING, BIG DATA) : DÉFINITIONS ET ENJEUX.

Éric FILIOL

Laboratoire de Cryptologie et de Virologie Opérationnelles (C + V)^{O}

filiol@esiea.fr

AGENDA

- Introduction
  - Que sont le Big Data et techniques assimilées ?
- Méthodes et techniques
  - Enjeux, renseignement, conceptions d’attaques
- Conclusion

AGENDA

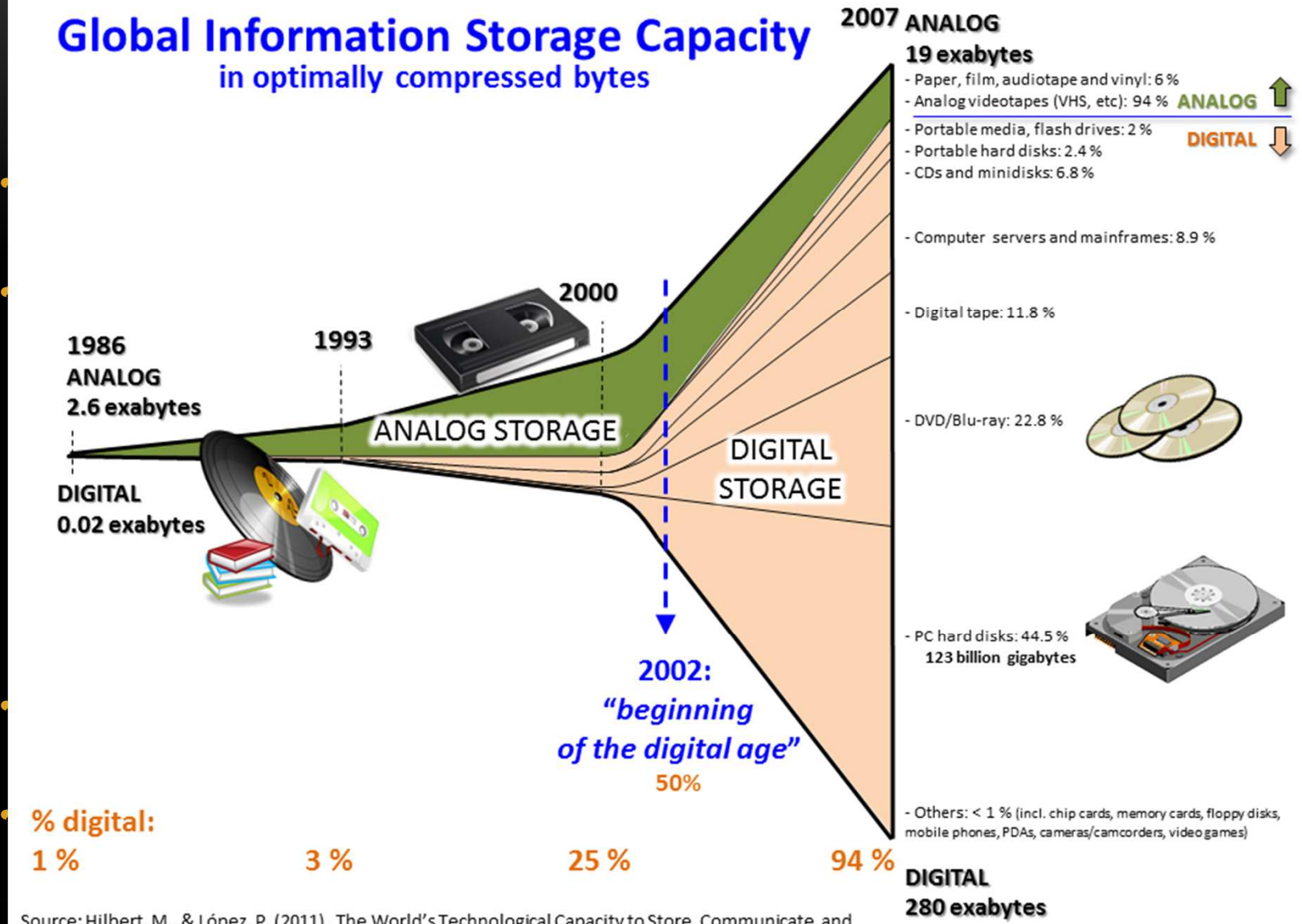
- Introduction
  - Que sont le Big Data et techniques assimilées ?
- Méthodes et techniques
  - Enjeux, renseignement, conceptions d’attaques
- Conclusion

CONTEXTE ET PROBLÉMATIQUE

- Techniques existant depuis les années 60 (extraction de données) boostées par les attentats du 9/11
- L'explosion du volume et de la nature des données a donné naissance (fin 90s, début 2000s) au big data
  - Volume traité par la NSA par Projet Muskular (2013 40 Go/j [estimation WP])
  - 2013 : Twitter produit 7 To/j – Facebook 10 To/j
  - *Square Kilometre Array* production annoncée de 50 To/j (données analysées) soit un 7 000 To/sec (données brutes)
  - Estimation Internet (2020) 40.10^{21} octets
- Explosion de la variété des données : données, métadonnées, textes, images, sons, vidéos, signaux EM….
- Problématiques nouvelles de stockage, de traitement/analyse, de visualisation…

Global Information Storage Capacity

in optimally compressed bytes



2007 **ANALOG**
19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % **ANALOG**

DIGITAL

- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %

- DVD/Blu-ray: 22.8 %

- PC hard disks: 44.5 %
  123 billion gigabytes
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

DIGITAL
280 exabytes

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

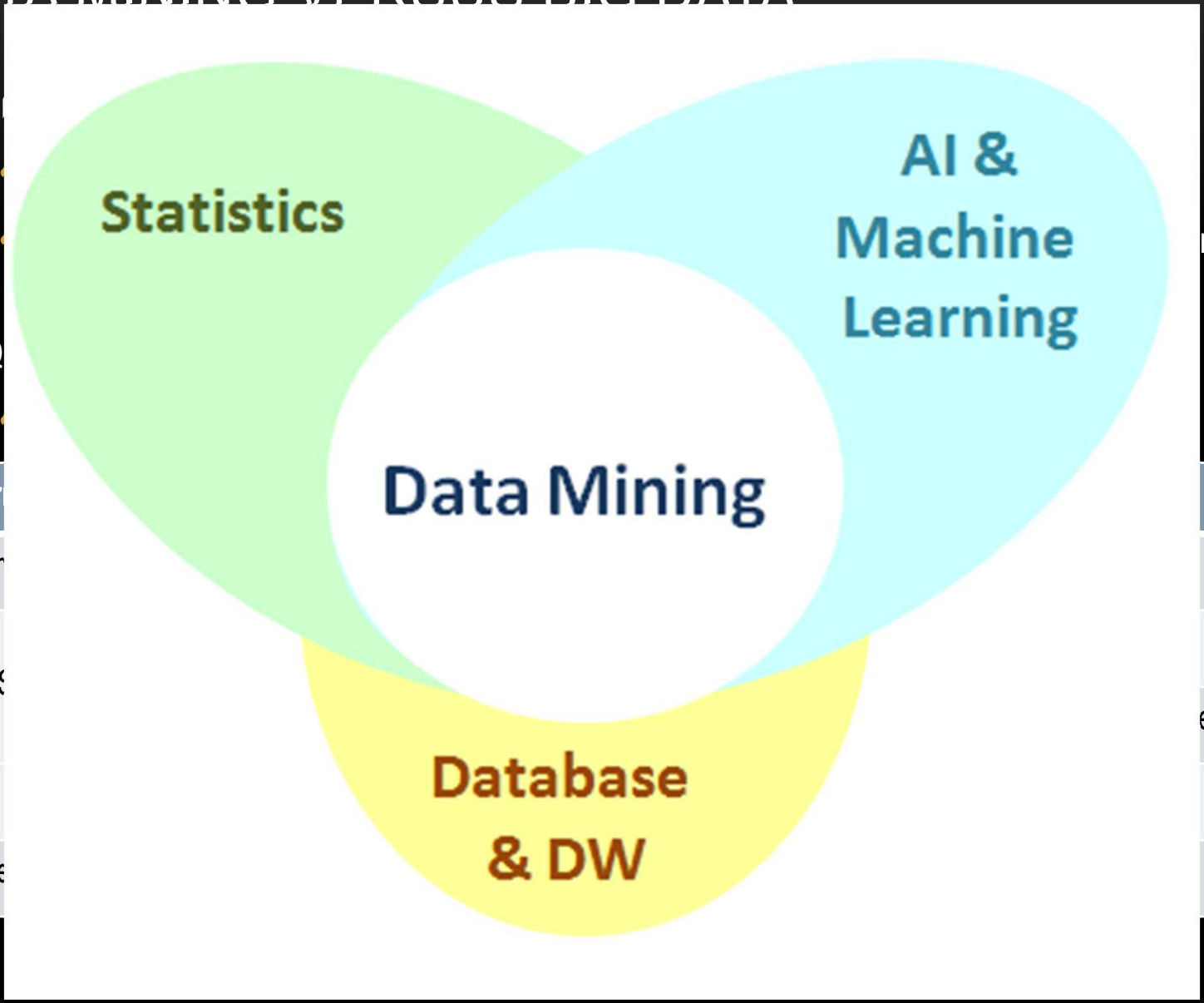
DATA MINING VERSUS BIG DATA

- Analyse de données brutes, en grand nombre et souvent informes dans le but de
  - comprendre des données par rapport à un ou plusieurs critères,
  - pouvoir prendre une décision par rapport au contexte et aux éléments constituant ce contexte.
- Quand parle t-on de big data ?
  - Pour des données dont la taille dépasse les quelques To.

<table>
<tr><th>Informatique décisionnelle</th><th>Data Mining</th><th>Big Data</th></tr>
<tr><td>Données à forte entropie</td><td>Entropie faible à moyenne</td><td>Entropie faible</td></tr>
<tr><td rowspan="2">Statistique descriptive</td><td colspan="2">Statistique inférentielle</td></tr>
<tr><td>Algorithmique</td><td>Algorithmique distribuée</td></tr>
<tr><td>Logiciel R</td><td>Weka</td><td>Hadoop/MapReduce</td></tr>
<tr><td>Existence d’un modèle absolu</td><td colspan="2">Extraction d’un modèle local et subjectif</td></tr>
</table>

DATA MINING VERSUS BIG DATA

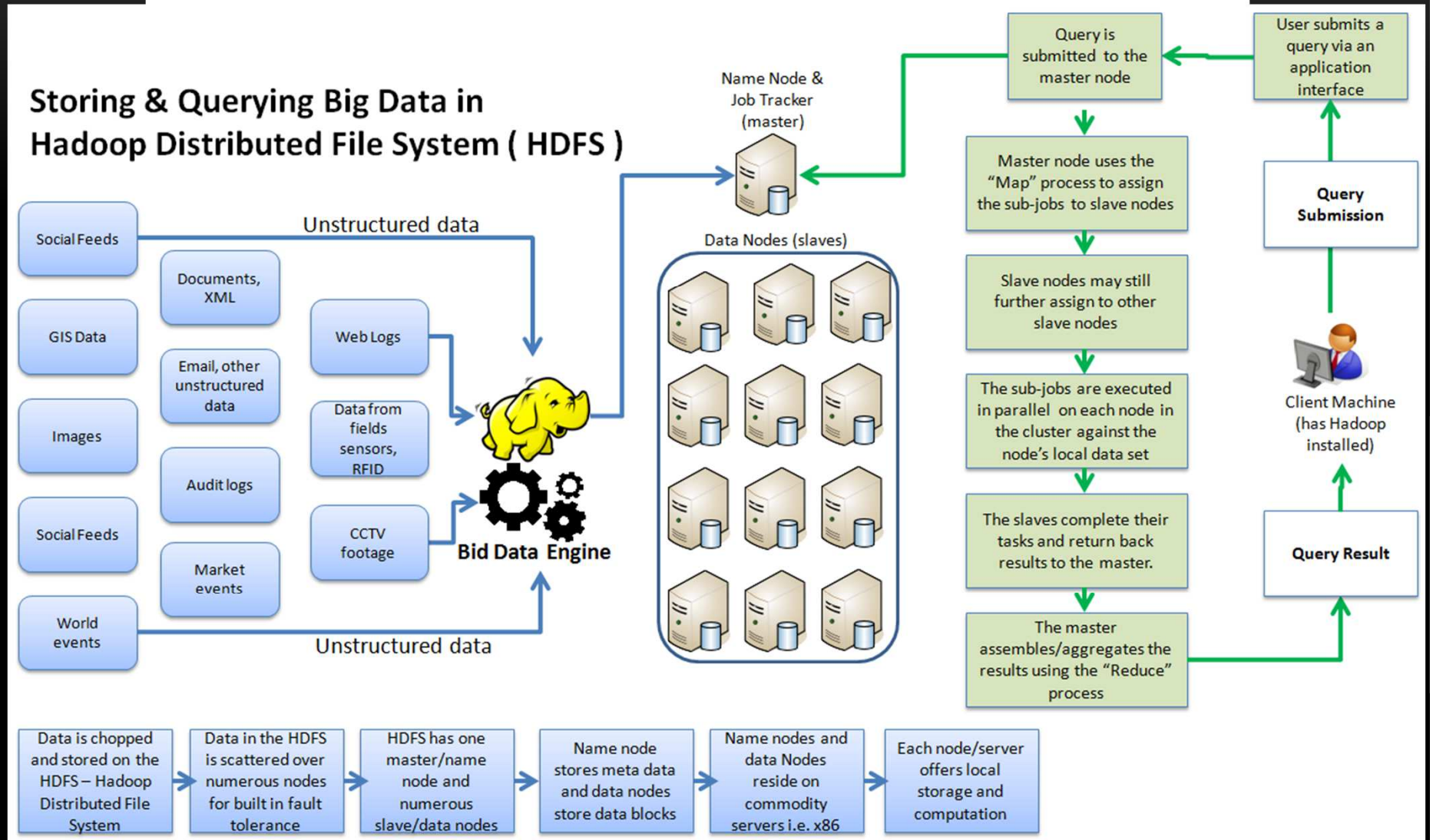
- A
  -
  -
- Q
  -



nt ce

Infor	
Dor	
S	e
Existe	

DATA MINING VERSUS BIG DATA



MYTHES ET REALITE

- Le big data/data mining tente
  - d'expliquer le passé par l'exploration des données (partie statistiques),
  - de « prédire le futur » par la modélisation des données (partie inférentielle).
- MAIS
  - Ce n'est pas une boule de cristal!
  - Importance critique de l'expertise en amont (bien définir le problème) et en aval (savoir correctement interpréter le résultat et surtout le confronter à la réalité [problème de la distribution d'échantillonnage])
  - Toujours subordonner la technique à la dimension opérationnelle.
  - BD et DM ne sont ni « bons » ni « mauvais » : c'est leur utilisation par l'homme qui peut poser problème
  - Il n'existe par de « super modèle » ou de « super équation » unificateurs

MYTHES ET REALITE (2)

- Avec le big data nous sommes juste passés d'une connaissance avec un référentiel absolu (ex : loi normale) à une connaissance à référentiel statistique (exemple : traduction Google)

« *Nous voilà trois contre toi, il faut que tu te rendes à l'opinion de la majorité. La majorité, vois-tu, mon ami, c'est plus fort que tout le monde cela. Mets dix philosophes d'un côté et onze imbéciles de l'autre, les imbéciles l'emporteront* »

Claude Tillier (1842)

- Bien utilisés et bien compris BD et DM peuvent être des outils très puissants… à ne pas mettre dans toutes les mains (voir plus loin)
- Les modèles fournis pour le BD/DM doivent être confrontés au temps et réactualisés en permanence… surtout dans le champ humain.

RÉALITÉ

- Par nature big data et data mining visent à extraire des informations potentiellement sensibles, dissimulées dans des données en apparence anodines
  - Exemple : la liste de vos achats pendant une période de quelques mois peut permettre de « deviner » vos habitudes de vie les plus intimes (religieuses, sexuelles, culturelles, idéologiques…)

AGENDA

- Introduction
  - Que sont le Big Data et techniques assimilées ?
- Méthodes et techniques
  - Enjeux, renseignement, conceptions d’attaques
- Conclusion

APPRENTISSAGES

- Pour tout problème (décidable!) il faut un algorithme.
- Dans beaucoup de situations, cet algorithme n'est pas disponible (trop complexe, non connu...)
  - Exemples : détection d'un terroriste.
- Les « données » vont servir pour « apprendre » ce qu'est un terroriste et l'ordinateur va automatiquement extraire l'algorithme pour résoudre ce problème.
- L'apprentissage consiste donc à programmer un ordinateur pour optimiser un critère de performance en utilisant des données produites par l'expérience.
- Définition d'un « modèle » conditionné par un certain nombre de paramètres. L'apprentissage consiste donc à optimiser les paramètres de ce modèle.
- Les termes « *données* », « *apprendre* », « *optimiser* », « *expérience* » et « *paramètres* » soulignent le caractère subjectif et local du modèle obtenu.

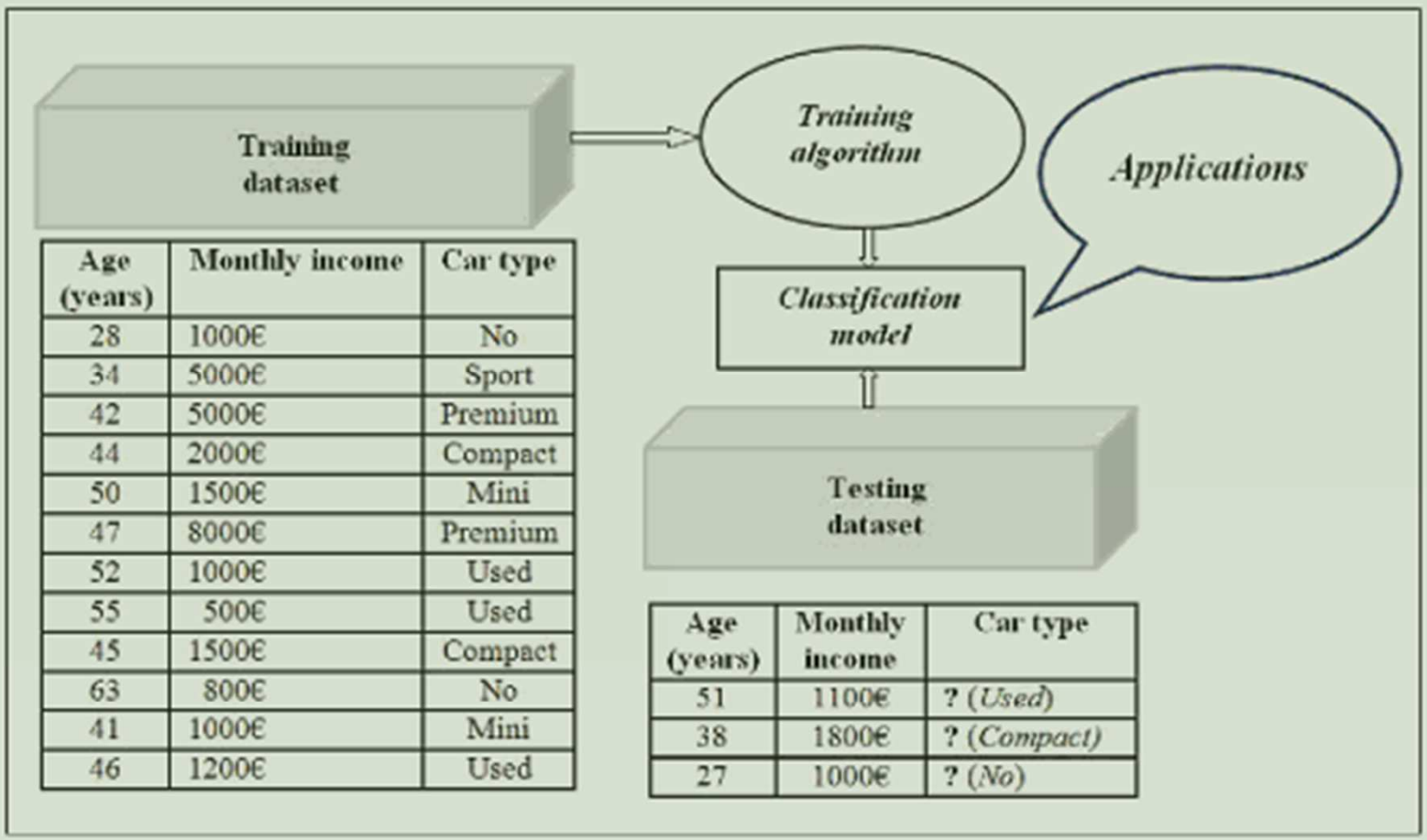
APPRENTISSAGES (2)

- Deux types d'apprentissage :
- **Apprentissage supervisé** : les données sont déjà caractérisées (traitées et validées) par un superviseur (données avec label, base de données d'apprentissage).
  - Méthode des k plus proches voisins, SVM, classification naïve bayésienne, arbres de décision, réseau de neurones...
- **Apprentissage non supervisé** : technique d'apprentissage automatique où les données sont brutes (sans label) et non interprétées. Il faut les classifier selon un ou plusieurs critères de similarité.
  - Techniques d'estimation de densité (*clustering*), réseaux de neurones (par exemple : cartes auto-adaptatives)...

PRINCIPALES ÉTAPES EN DM/BD

1. Traiter, extraire et explorer les données. On part de données brutes, souvent malformées, partiellement non pertinentes... pour obtenir des données utilisables pour le traitement.
    - bruit (e.g. traitement de signal nécessaire), outliers/anomalies, données dupliquées ou redondantes, données obsolètes, fausses, manquantes...
2. Construction et validation d'un modèle (description d'un ensemble d'objets, de leurs propriétés et relations/interactions par un programme). Entre plusieurs modèles, il faut pouvoir choisir le plus adapté pour notre problème, dans le contexte opérationnel donné (nécessité de mesures de performances).
3. Application du modèle à de nouvelles données pour une approche prédictive/décisionnelle relative au problème.
4. Evaluation de la performances : cout (temps, mémoire, impact opérationnel d'une mauvaise décision…), robustesse, rapidité, efficacité, sensibilité, scalabilité…

Univariate Analysis - Numerical			
Statistics	**Visualization**	**Equation**	**Description**
Count	Histogram	N	The number of values (observations) of the variable.
Minimum	Box Plot	Min	The smallest value of the variable.

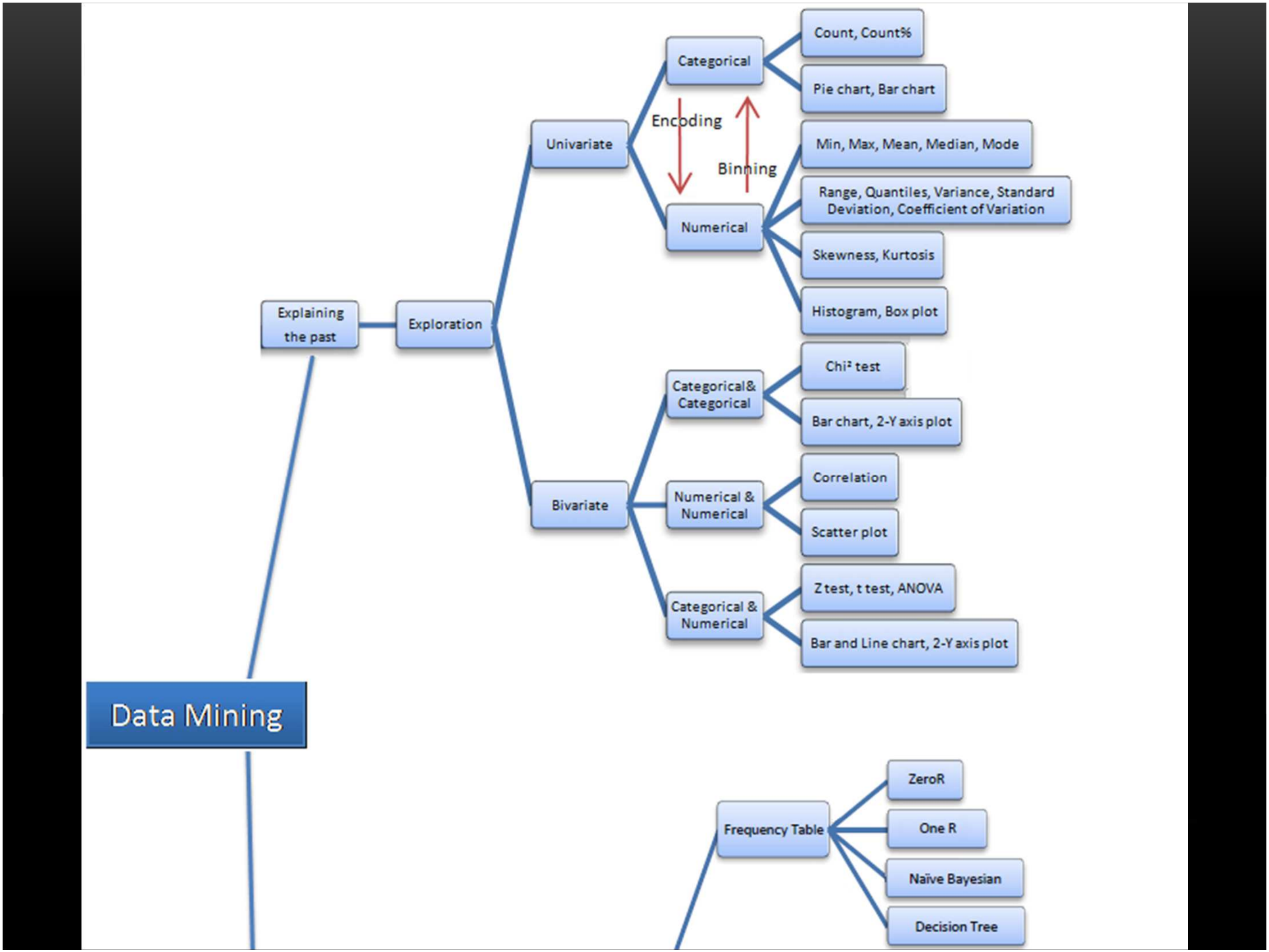


Age (years)	Monthly income	Car type
28	1000€	No
34	5000€	Sport
42	5000€	Premium
44	2000€	Compact
50	1500€	Mini
47	8000€	Premium
52	1000€	Used
55	500€	Used
45	1500€	Compact
63	800€	No
41	1000€	Mini
46	1200€	Used

Age (years)	Monthly income	Car type
51	1100€	? (*Used*)
38	1800€	? (*Compact*)
27	1000€	? (*No*)

Kurtosis	Histogram	$\left[\frac{}{(N-1)(N-2)(N-3)}\sum\left(\frac{}{S}\right)\right]$ $-\frac{3(N-1)^2}{(N-2)(N-3)}$	A measure of whether the data are peaked or flat relative to a normal distribution.

Univariate Analysis - Numerical			
Statistics	**Visualization**	**Equation**	**Description**
Count	Histogram	N	The number of values (observations) of the variable.
Minimum	Box Plot	Min	The smallest value of the variable.
Maximum	Box Plot	Max	The largest value of the variable.
Mean	Box Plot	$\bar{X} = \frac{\sum X}{N}$	The sum of the values divided by the count.
Median	Box Plot	$\tilde{X}$	The middle value. Below and above median lies an equal number of values.
Mode	Histogram		The most frequent value. There can be more than one mode.
Quantile	Box Plot	Q_k	A set of 'cut points' that divide a set of data into groups containing equal numbers of values (Quartile, Quintile, Percentile, ...).
Range	Box Plot	$Max-Min$	The difference between maximum and minimum.
Variance	Histogram	$S^2 = \frac{\sum(X-\bar{X})^2}{N-1}$	A measure of data dispersion.
Standard Deviation	Histogram	$S = \sqrt{S^2}$	The square root of variance.
Coefficient of Deviation	Histogram	$CV = \frac{S}{\bar{X}} \times 100\%$	A measure of data dispersion divided by mean.
Skewness	Histogram	$\frac{N}{(N-1)(N-2)} \sum \left(\frac{X-\bar{X}}{S}\right)^3$	A measure of symmetry or asymmetry in the distribution of data.
Kurtosis	Histogram	$\left[\frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum \left(\frac{X-\bar{X}}{S}\right)^4\right] - \frac{3(N-1)^2}{(N-2)(N-3)}$	A measure of whether the data are peaked or flat relative to a normal distribution.

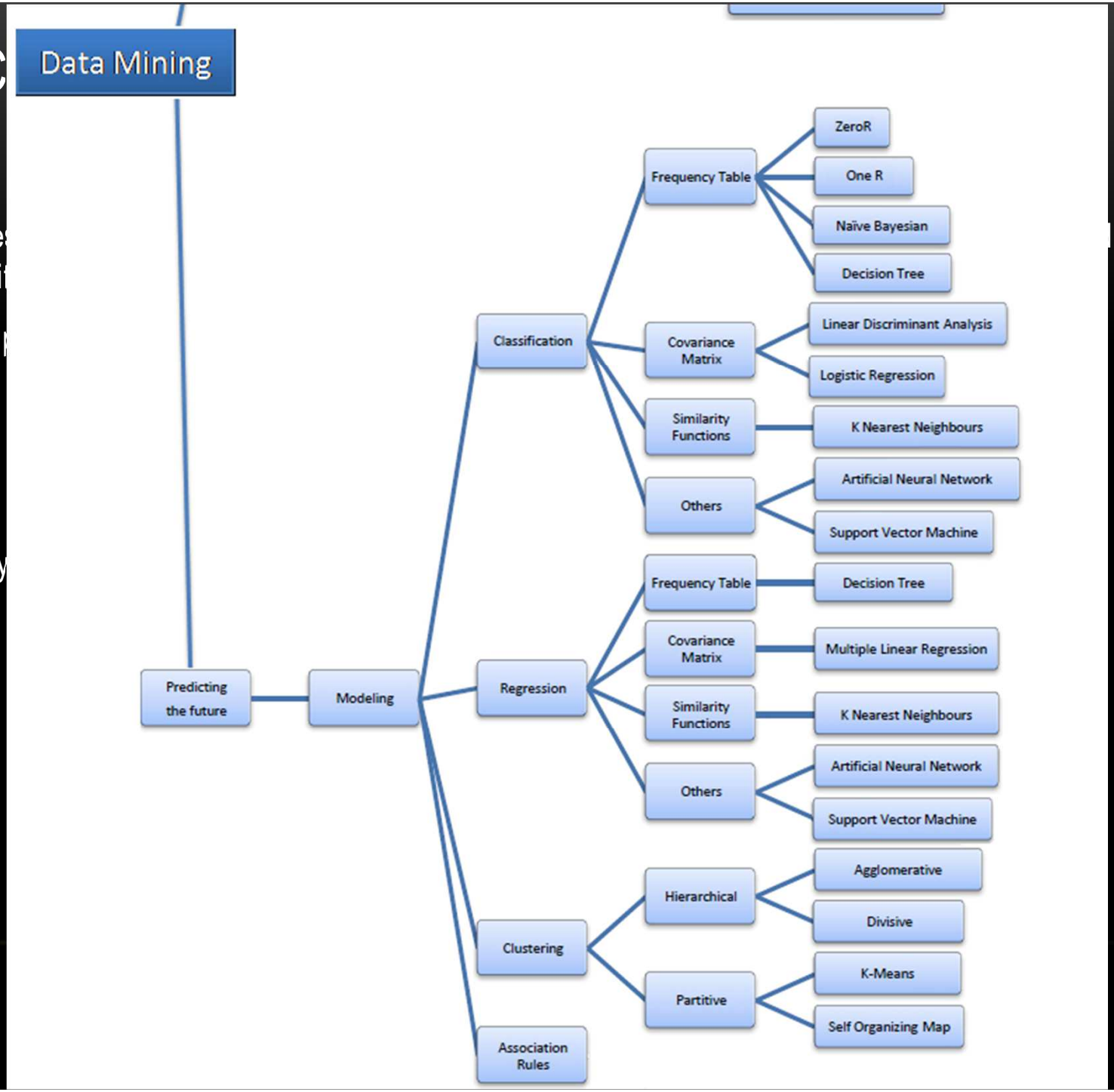

Data Mining
Explaining the past
Exploration
Univariate
Categorical
Count, Count%
Pie chart, Bar chart
Encoding
Binning
Numerical
Min, Max, Mean, Median, Mode
Range, Quantiles, Variance, Standard Deviation, Coefficient of Variation
Skewness, Kurtosis
Histogram, Box plot
Bivariate
Categorical& Categorical
Chi² test
Bar chart, 2-Y axis plot
Numerical & Numerical
Correlation
Scatter plot
Categorical & Numerical
Z test, t test, ANOVA
Bar and Line chart, 2-Y axis plot
Frequency Table
ZeroR
One R
Naïve Bayesian
Decision Tree

TECHNIQUES DE BD/DM

- Très nombreuses techniques selon la nature (supervisée ou non) de l'apprentissage et du traitement
- Le passage au big data impose une vision algorithmique totalement différente
  - Exemple : en data mining comparaison de N objets → $O(N^2)$. En big data on passe sur des méthodes proches de $O(N)$ (*minhashing*, techniques LSH…)
  - Techniques de multihashing dans le domaine des règles d'association
- Voyons quelques exemples
  - Partitionnement (*clustering*)
  - Règles d'association (*association rules*)

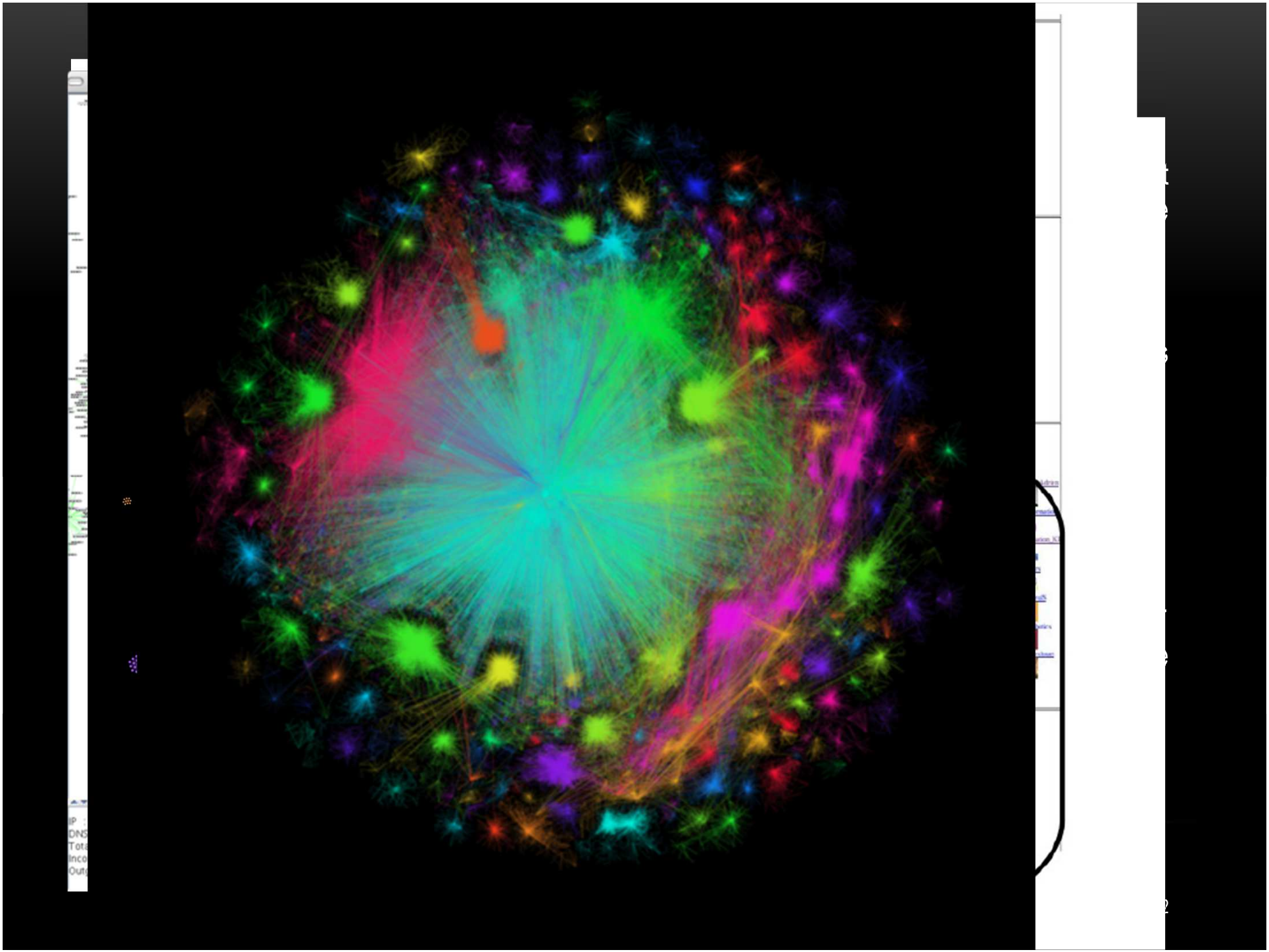
TEC

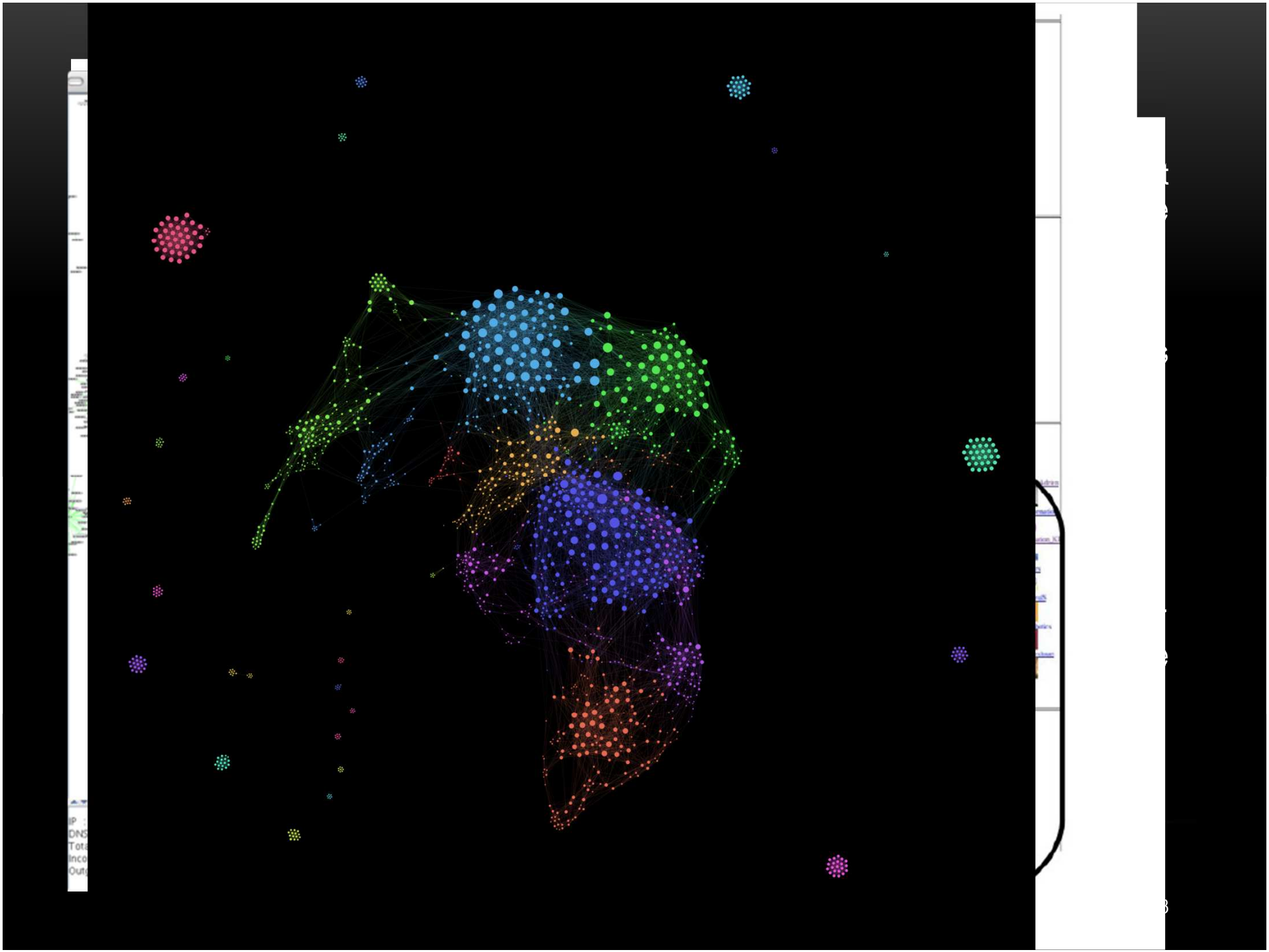
- Très
  trait
- Le
  -
  -
- Voy
  -
  -

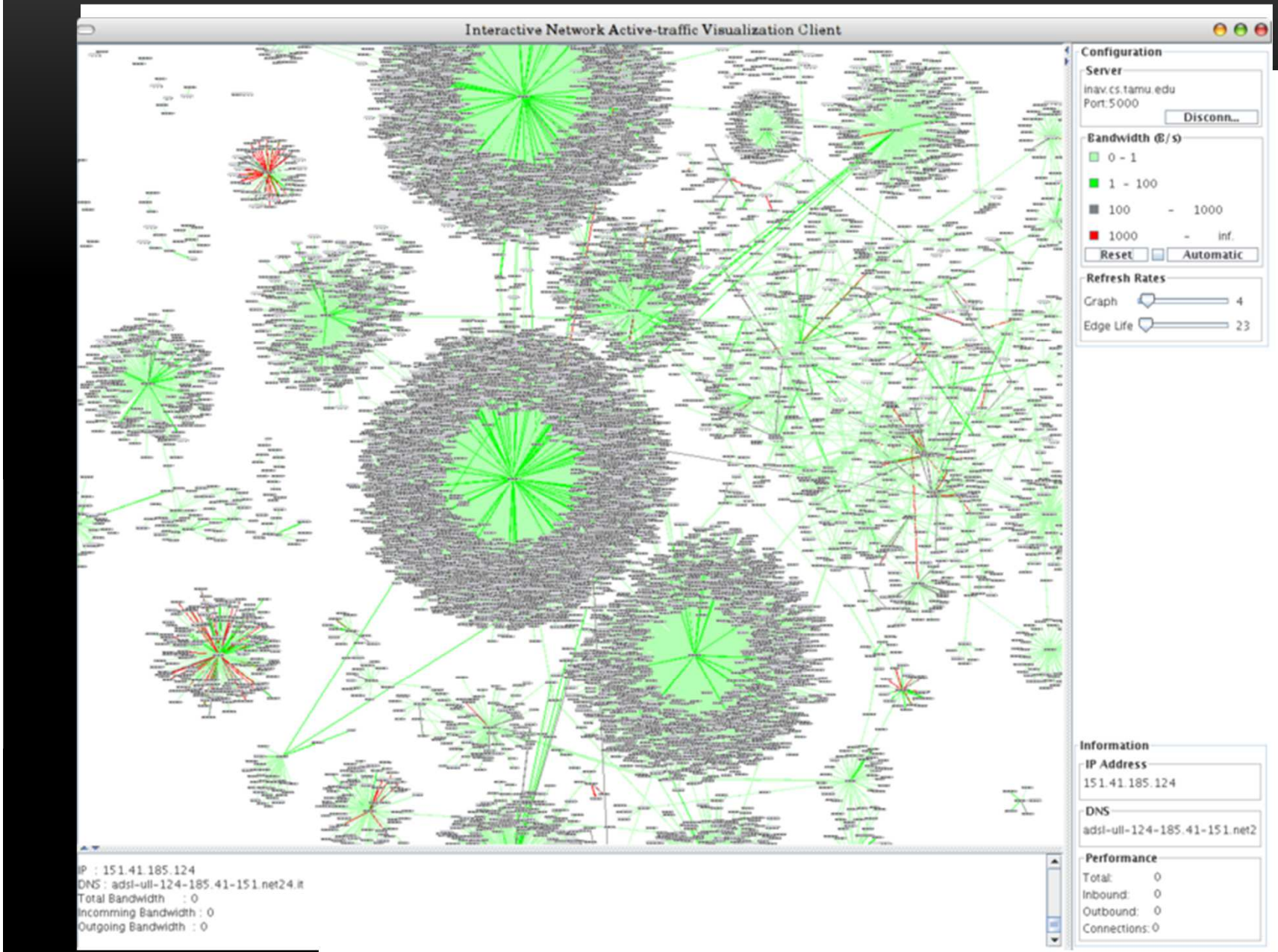


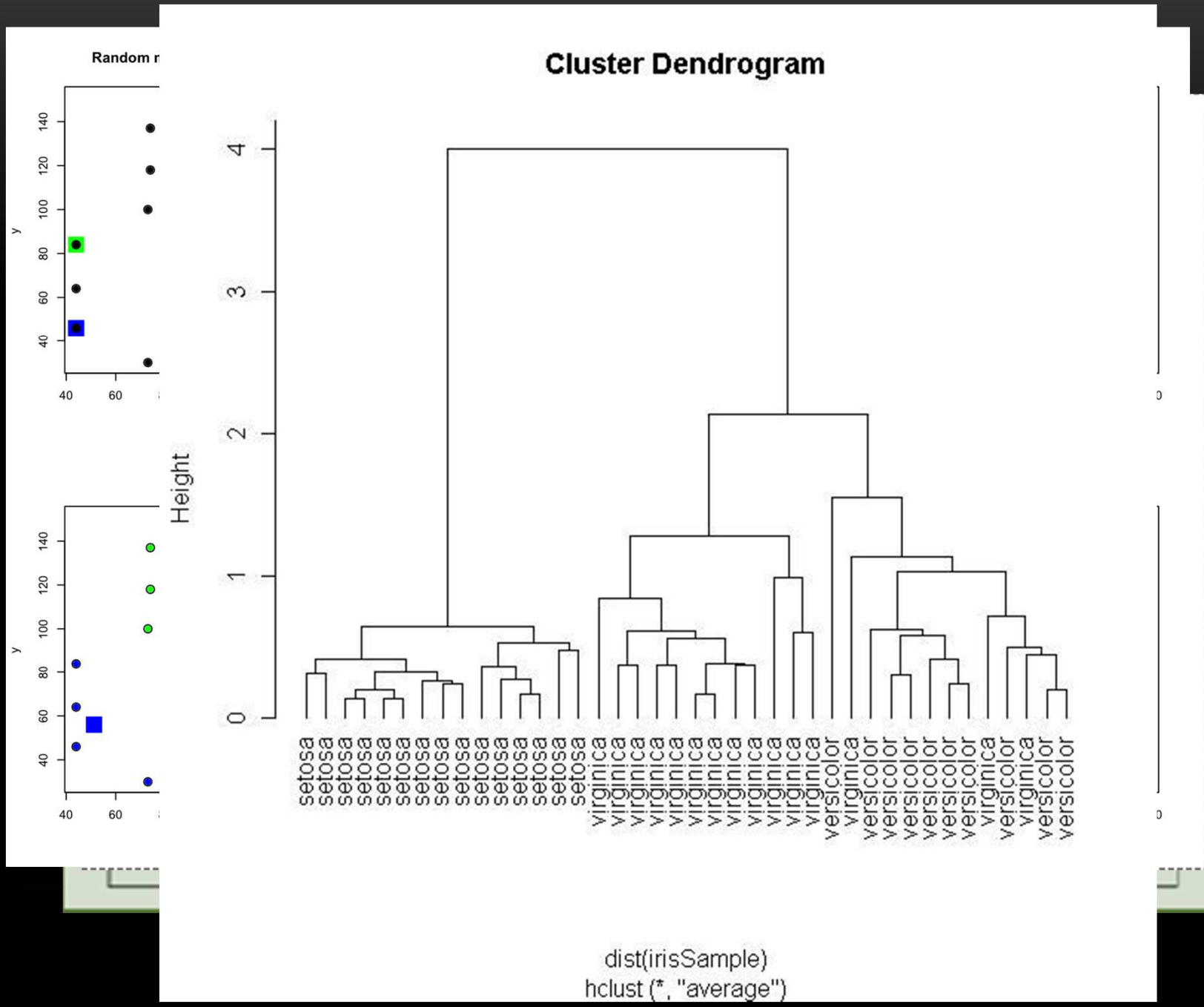
TECHNIQUES DE PARTITIONNEMENT

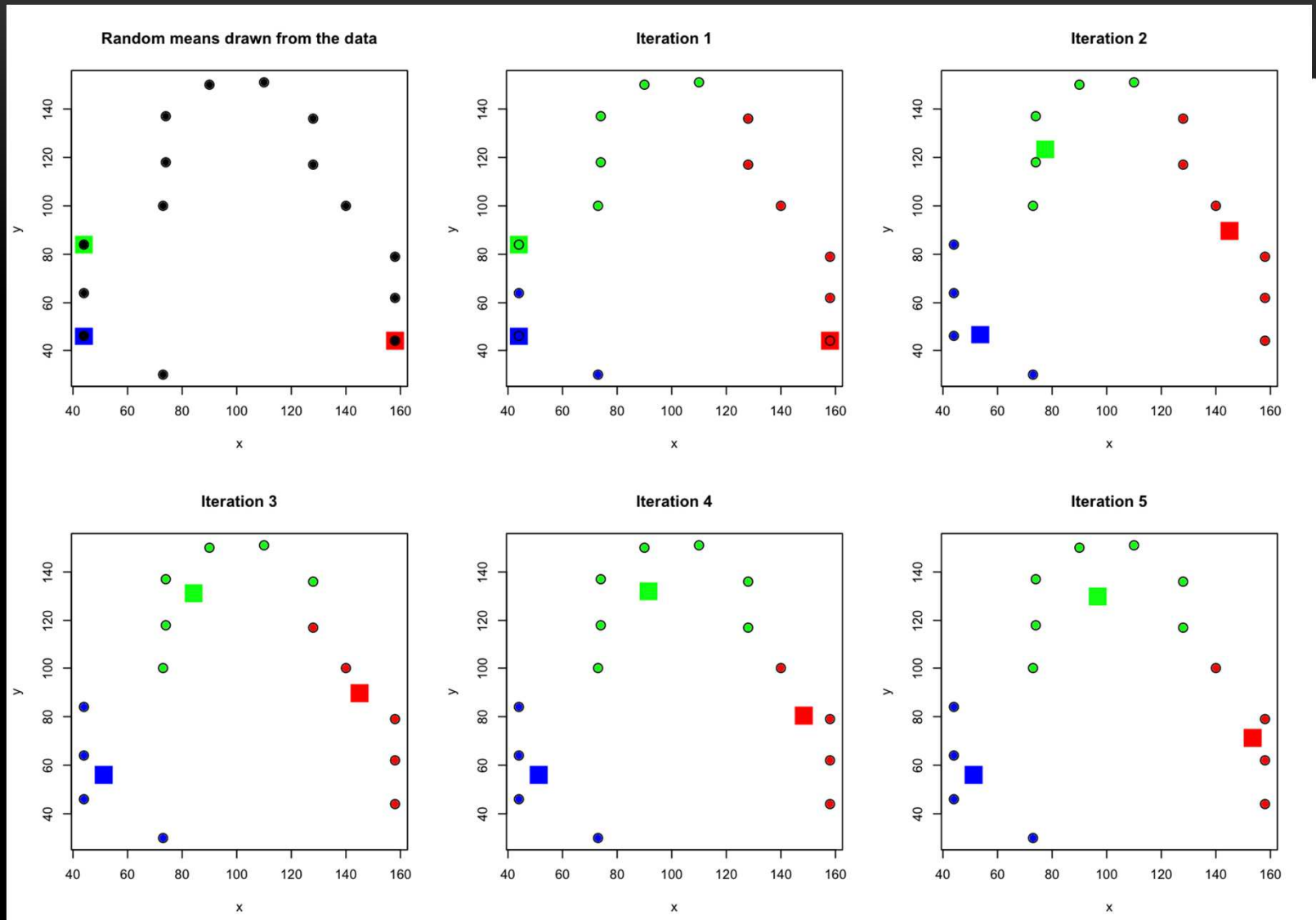
- Méthodes non supervisées, les objets (caractérisés par un ensemble d'attributs), doivent être répartis dans des groupes homogènes par rapport à des propriétés de similarité prédéterminées.
  - Pas de labellisation des objets (les groupes ne sont pas *a priori* connus).
  - Le nombre de classes (clusters) n'est *a priori* pas connu (utilisation d'arbres hiérarchiques ascendants)
  - Utilisation d'une ou plusieurs mesures de similarité ou de distances.
- A la fin du partitionnement :
  - Les objets d'un même groupe sont plus similaires les uns par rapport aux autres.
  - Les objets de groupes différents sont moins similaires les uns par rapport aux autres.
- Une étape additionnelle (manuelle, supervisée, statistique...) peut permettre de caractériser (labelliser) les classes.
- Analyse des outliers



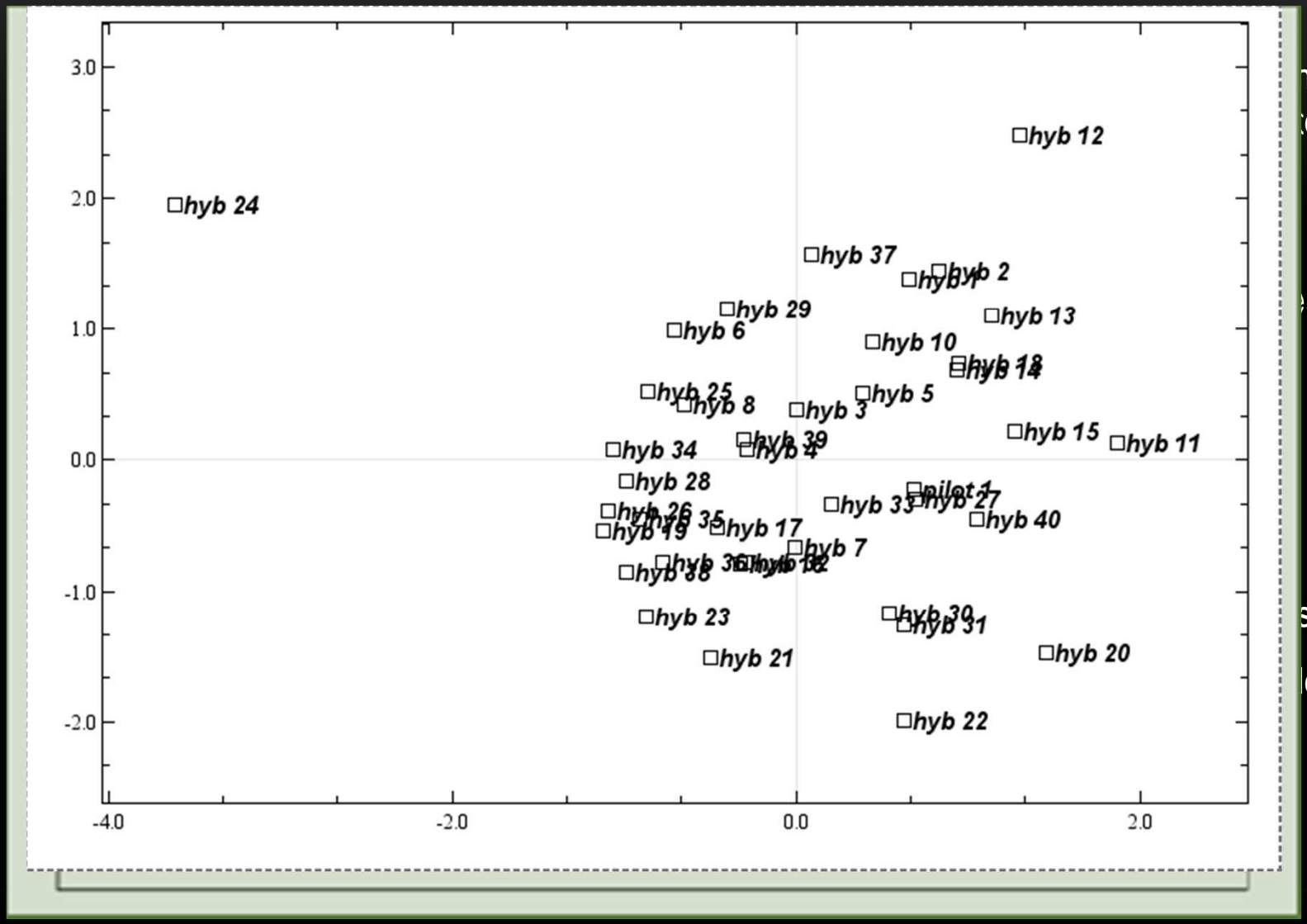



Interactive Network Active-traffic Visualization Client
Configuration
Server
inav.cs.tamu.edu
Port:5000
Disconn...
Bandwidth (B/s)
0 - 1
1 - 100
100 - 1000
1000 - inf.
Reset
Automatic
Refresh Rates
Graph
4
Edge Life
23
Information
IP Address
151.41.185.124
DNS
adsl-ull-124-185.41-151.net2
Performance
Total: 0
Inbound: 0
Outbound: 0
Connections: 0
IP : 151.41.185.124
DNS : adsl-ull-124-185.41-151.net24.it
Total Bandwidth : 0
Incomming Bandwidth : 0
Outgoing Bandwidth : 0

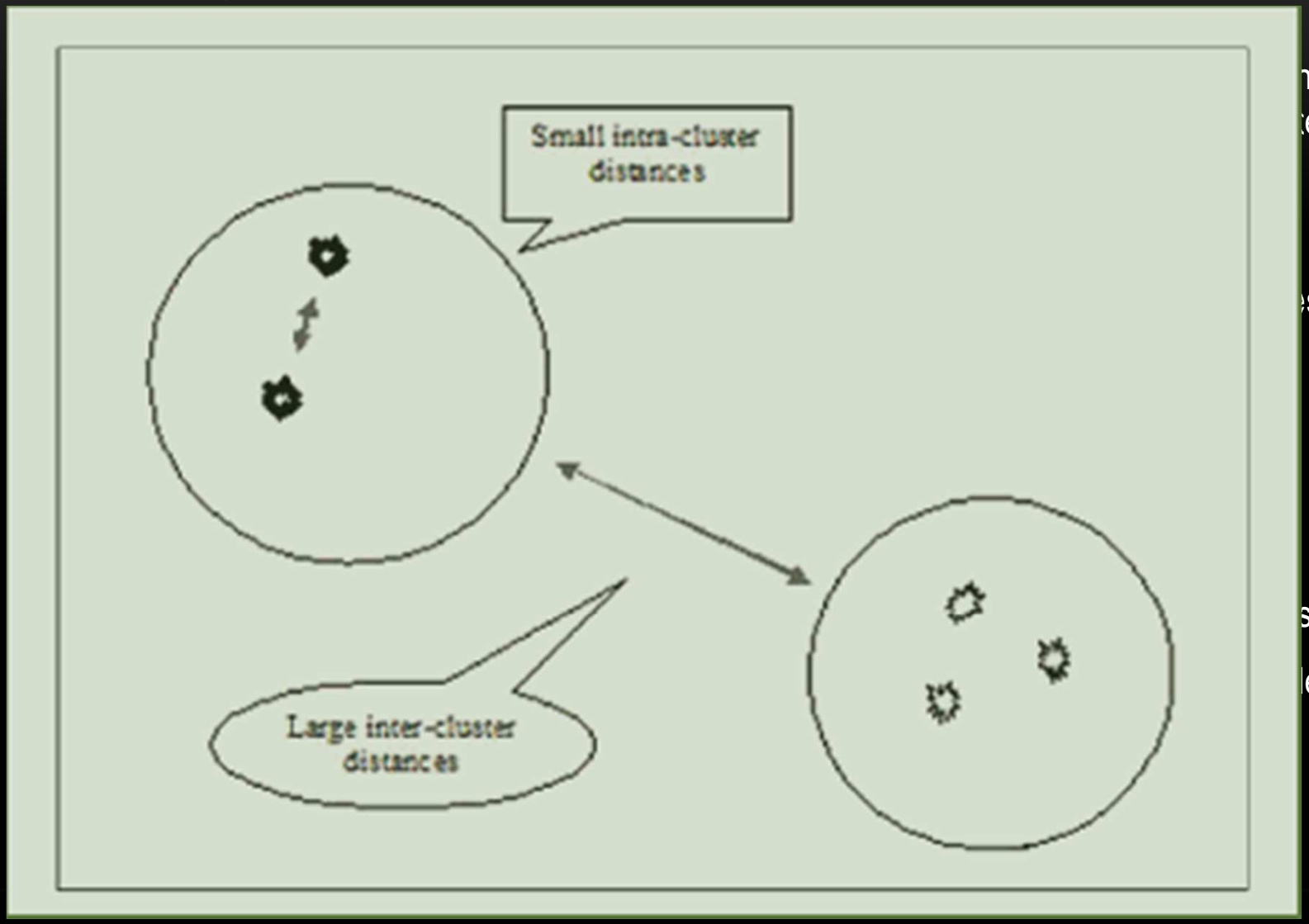

Random n
y
40
60
80
100
120
140
Cluster Dendrogram
Height
0
1
2
3
4
setosa
virginica
versicolor
dist(irisSample)
hclust (*, "average")


Random means drawn from the data
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
x
y
40
60
80
100
120
140
160

TECHNIQUES DE PARTITIONNEMENT



TECHNIQUES DE PARTITIONNEMENT



ASPECTS COMBINATOIRES DU PARTITIONNEMENT

- Pour trouver le meilleur partitionnement, on peut imaginer explorer toutes les partitions possibles (à nombre fixe de classes ou non) et choisir celles qui optimisent le critère (classifieur) considéré.
- L'explosion combinatoire rend cette tâche impossible : en traitant 10^9 partitions par seconde, il faudrait 126 ans pour traiter toutes les partitions d'un ensemble de 25 objets.
- Nombre de partitions de n objets en k classes (nombre de Stirling de 2nde espèce)

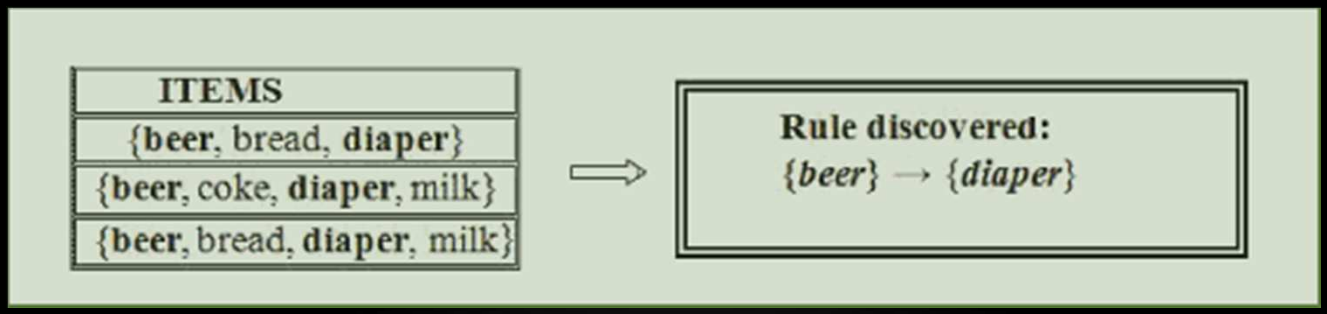
$$P_{n,1} = P_{n,n} = 1 \qquad P_{n,n-1} = \frac{n(n-1)}{2} \qquad P_{n,2} = 2^{n-1} - 1$$

$$P_{n,k} = P_{n-1,k-1} + k.P_{n-1,k} \qquad P_{n,k} = \frac{1}{k!} \sum_{i=1} k \binom{k}{i} (-1)^{k-1} i^n$$

- Dans les problèmes classiques de data mining/big data, n est de l'ordre de quelques dizaines de milliers à quelques dizaines de milliards d'objets.
- BD/DM ne fournissent qu'une approximation (problème du choix des vecteurs initiaux de classes)

RÈGLES D'ASSOCIATION

- Processus d'identification des règles de dépendance entre différents groupes de phénomènes.
- Ces règles permettent ensuite de prédire les occurrences d'un phénomène en fonction d'un autre (exemple : « suggestions d'Amazon »)
- En big data : problème algorithmique d'identification des « *frequent item sets* » (techniques de hashing et multihashing)
- Exemple (historique ? Légende urbaine ?) : Wal-mart



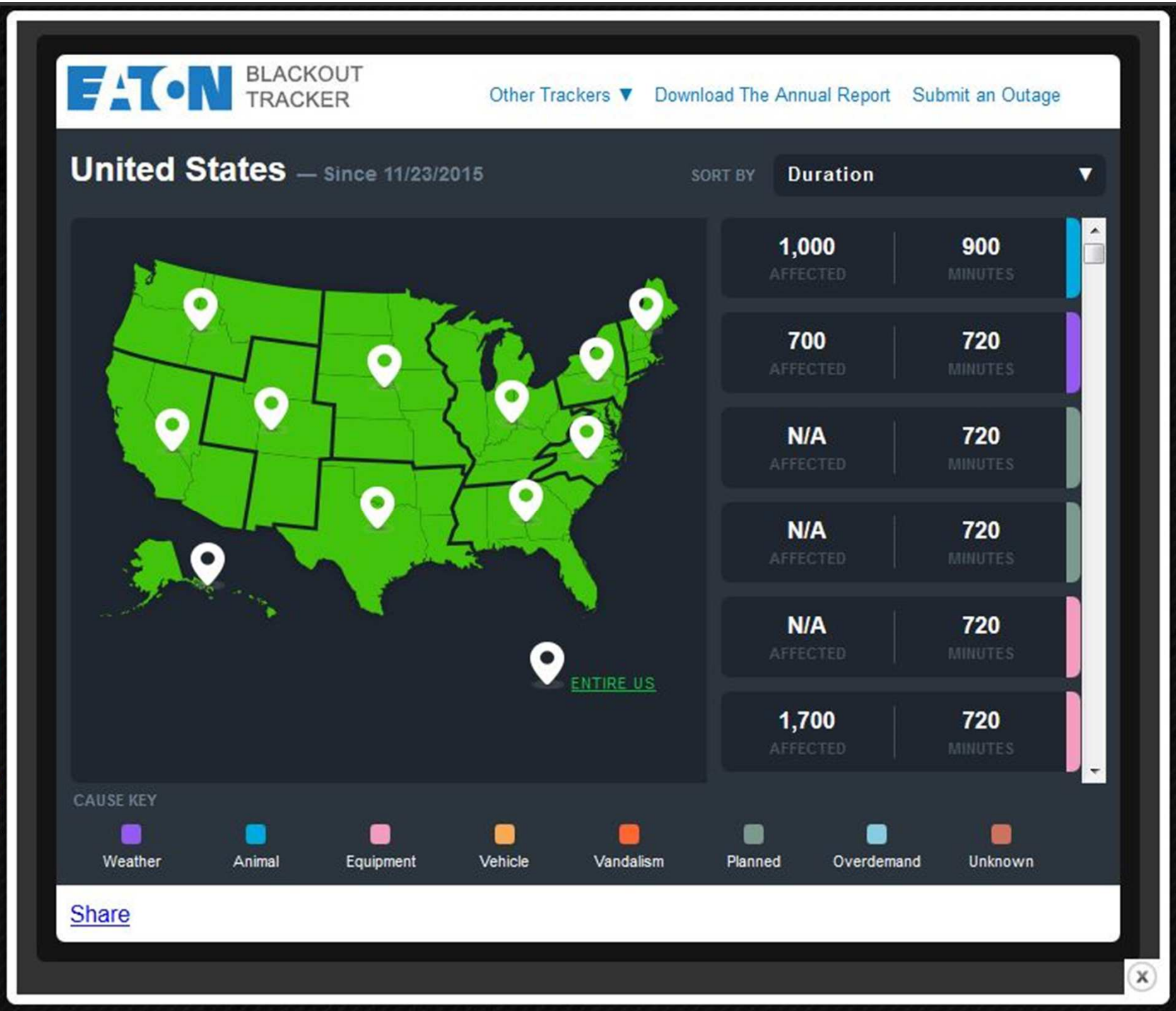
RÈGLES D'ASSOCIATION - EXEMPLE

- T = 1 000 000 consultations de sites web « sensibles »
  - 40 000 consultations concernent le site A (4 %)
  - 60 000 consultations concernent le site B (6 %)
  - 20 000 consultations concernent a la fois A et B (2 %)
- Sur un ensemble de d objets, on peut extraire $R = (3^d – 2^{d+1} + 1)$ règles du type $X \rightarrow Y$
- Ici d = 2 soit deux règles possibles $A \rightarrow B$ ou $B \rightarrow A$
- Règle $A \rightarrow B$: 50 % de confiance (20 000/40 000). On en déduit que ceux qui consultent le site A ont 8.33 (0,5/0.06) plus de chances de consulter aussi le site B.
- Comme 8.33 >> 1, il y a une très forte corrélation entre la consultation du site A et celle du site B
- NB : dans $X \rightarrow Y$, X et Y décrivent des sous-ensembles d'« objets » (item sets)

APPLICATION : ATTAQUE D'UN RÉSEAU ÉLECTRIQUE

- Etude publiée dans les Cahiers de la Revue de Défense Nationale et présentée à ICCWS 2016 (Boston)
- Optimisation d'une attaque conventionnelle contre le réseau électrique US par la phase de renseignement et de planification
  - Utilisation de données ouvertes (media et open data)
- Aspect critique : déterminer les points sensibles du réseau (pylônes, sous-stations…) nécessitant un temps de réparation plus long que pour les autres points sensibles (but : prolongement des effets de l'attaque)
- Règles d'association
  - {critères géographiques, critères climatiques, largeur des routes, critères démographiques, équipements, ressources de sécurité} $\rightarrow$ {durée de réparation}

APPLICATION : ATTAQUE D'UN RÉSEAU ÉLECTRIQUE

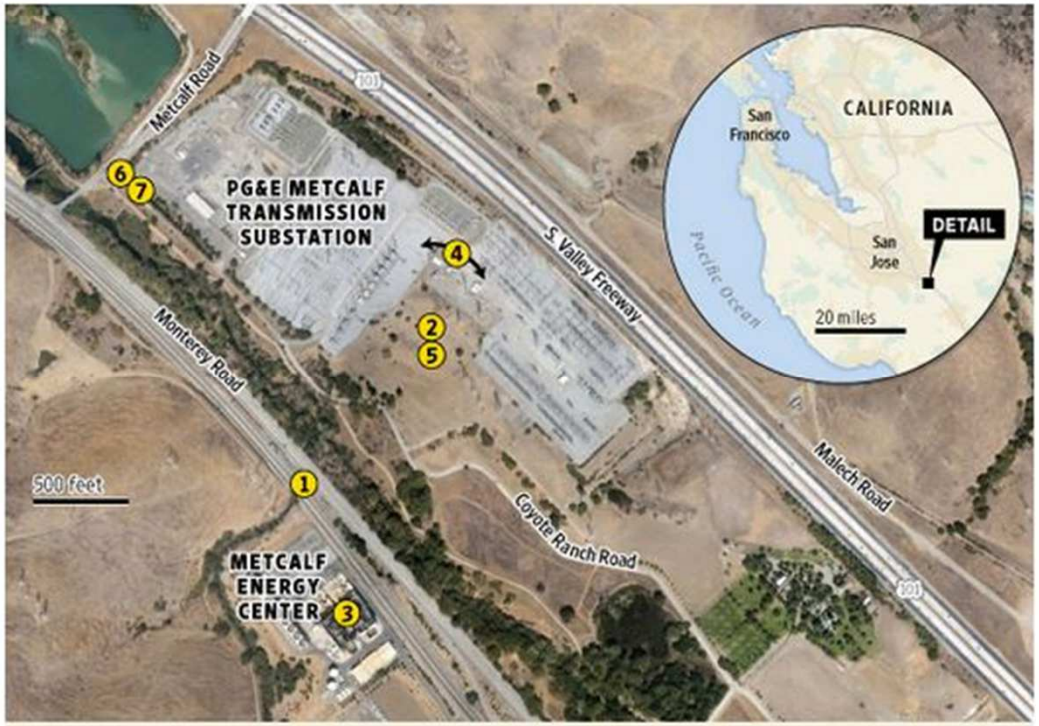

EATON BLACKOUT TRACKER
Other Trackers ▼
Download The Annual Report
Submit an Outage
United States — Since 11/23/2015
SORT BY
Duration
1,000 AFFECTED 900 MINUTES
700 AFFECTED 720 MINUTES
N/A AFFECTED 720 MINUTES
N/A AFFECTED 720 MINUTES
N/A AFFECTED 720 MINUTES
1,700 AFFECTED 720 MINUTES
ENTIRE US
CAUSE KEY
Weather
Animal
Equipment
Vehicle
Vandalism
Planned
Overdemand
Unknown
Share

Assault on California Power Station Raises Alarm on Potential for Terrorism

April Sniper Attack Knocked Out Substation, Raises Concern for Country's Power Grid

By REBECCA SMITH CONNECT

Feb. 4, 2014 10:30 p.m. ET



Shots in the Dark

A look at the April 16 attack on PG&E's Metcalf Transmission Substation

1	2	3	4	5	6	7
12:58 a.m., 1:07 a.m. Attackers cut telephone cables	**1:31 a.m.** Attackers open fire on substation	**1:41 a.m.** First 911 call from power plant operator	**1:45 a.m.** Transformers all over the substation start crashing	**1:50 a.m.** Attack ends and gunmen leave	**1:51 a.m.** Police arrive but can't enter the locked substation	**3:15 a.m.** Utility electrician arrives

Sources: PG&E; Santa Clara County Sheriff's Dept.; California Independent System Operator; California Public Utilities Commission; Google (image)
The Wall Street Journal

Popular Now

ARTICLES

1 **Best Places to Find Love**

2 **Comcast Acquiring Time Warner Cable In All-Stock Deal**

3 **$2.2 Billion Bird-Scorching Solar Project**

4 **Assault on Power Grid Raises Alarms**

5 **BNP Hit by $1.1 Billion Provision**

AGENDA

- Introduction
  - Que sont le Big Data et techniques assimilées ?
- Méthodes et techniques
  - Enjeux, renseignement, conceptions d’attaques
- Conclusion

CONCLUSION

- Ces techniques offrent un potentiel énorme d'un point de vue opérationnel
  - Mais au-delà des outils et des techniques, cela requiert avant tout une expertise métier et/ou opérationnelle.
  - Il faut correctement définir le problème et surtout correctement interpréter les « réponses » fournies
  - Ces « réponses » sont subjectives par nature !
- Les risques pour la vie privée et les libertés (surveillance sans contrôles des citoyens) sont à la hauteur de ce potentiel et sont bien réels
  - Capteurs divers (Windows 10, imsi-catcher, « boitiers noirs », smart compteur…)

  « Si vous voyez un seul article qui remet en cause la liberté publique, dites-le moi. En revanche, il y a des articles qui remettent en cause la vie privée »

  Bernard Cazeneuve 14 Avril 2015, Assemblée Nationale
- Nécessité de préserver le droit à la solitude, à l'intimité, à l'anonymat dans la foule et à réserve (Askland, 2006)

CONCL

- Ces tech
  - Mai… tise
    méti
  - Il fa… s
    « ré
  - Ces
- Les risqu… ens)
  sont à la
  - Cap… r…)

  « Si vou… oi. En
  revanch…

- Nécessit… ale
  réserve (… e et à

CONCL


Pour Cazeneuve, la vie privée n'est pas une liberté !
Guillaume Champeau - 14 avril 2015 - Politique
Accueil > Politique > Pour Cazeneuve, la vie privée n'est pas une liberté !

17:45


Le compteur Linky
dévoile votre vie privée

(PRÉ)TRAITEMENT DES DONNÉES CLIENTS

- Pour l'UE (G29, directive 95/46/CE), les données personnelles doivent être anonymisées lorsque susceptibles de faire objet d'analyse de type *big-data* (ce que font certains opérateurs)
- Trois critères pour évaluer une solution d'anonymisation
  - L'individualisation : est-il toujours possible d'isoler un individu ?
  - La corrélation : est-il possible de relier entre eux des ensembles de données distincts concernant un même individu ?
  - L'inférence : peut-on déduire de l'information sur un individu ?
- Voir site de la CNIL (http://www.cnil.fr/documentation/fiches-pratiques/fiche/article/cloud-computing-les-7-etapes-cles-pour-garantir-la-confidentialite-des-donnees)

BIBLIOGRAPHIE (1)

- Leskovec, J., Rajaraman, A. et Ullman J. D. (2014) *Mining of Massive Datasets* (cours de Standford CS 246). Cambridge University Press. Voir http://http://mmds.org/
- Aggarwal, C. C. (2013). Outlier Analysis. Springer Verlag
- Mohanty, H., Bhuyan, P., Chenthati, . (2015). *Big data* – A Primer. Springer Verlag
- Gorunescu, F. (2011). *Data Mining – Concepts, Models and Techniques*. Springer Verlag
- Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press
- Nakache, J.-P. et Confais, J. (2005). *Approches pragmatiques de la classification*. Technip
- Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning – Data Mining, Inference and Prediction*. Springer Verlag.

BIBLIOGRAPHIE (2)

- Andrew Askland (2006) *What, Me Worry: The Multi-Front Assault on Privacy*. St. Louis University Public Law Review, Vol. 25, No. 1, pp. 33--61
- Andrew Askland (2009). *Science and Socially Responsible Freedom*. Science and Engineering Ethics, Vol. 15, No. 3, pp. 343--349.
- Gaston Bachelard (1927). *Essai sur la connaissance approchée*. Jean Vrin éd.
- Auguste Comte (1830). *Cours de Philosophie Positive*. http://gallica.bnf.fr/ark:/12148/bpt6k76267p/
- #DATAGUEULE (2014). *Big data : données, données, donnez-moi* ! https://www.youtube.com/watch?v=5otaBKsz7k4
- #DATAGUEULE (2015). Privés de vie privée ? https://www.youtube.com/watch?v=wShQYeH9qJk

esiea
ECOLE D'INGENIEURS
DU MONDE NUMERIQUE

Merci de votre attention

QUESTIONS & RÉPONSES