



INSTITUT
de MATHEMATIQUES
de TOULOUSE



Mathématiques et

*BIG
DATA*

"LES DONNÉES SONT LE NOUVEL OR NOIR."

Depuis les premières mesures jusqu'en 2012, nous avons généré **5 exaoctets de données.**

En 2011, cette quantité était générée en 2 jours.
En 2012, cette quantité sera générée en 10 minutes.

Le trafic internet généré en 1 heure pourrait remplir environ **7 milliards de DVD.**

En 2008, un entrepreneur britannique spécialiste de la commercialisation de données, cette célèbre phrase a été reprise par le Forum Economique Mondial dans son rapport 2011. Ce dernier a présenté les données comme étant un atout économique, au même titre que le pétrole.

Il y a presque autant de données et d'informations dans l'univers digital qu'il y a d'étoiles dans l'univers réel.

En août 2012, on comptait plus de **4 millions d'articles en anglais sur Wikipedia.**

Il y a **133 millions de BLOGS** sur internet.

80% de la population mondiale possède un téléphone mobile. Dans le monde, on compte 1 milliard de smartphones. Les 3 milliards pairs de téléphones mobiles à Singapour, 84% des citoyens possèdent un smartphone.

L'anglais est la langue dominante sur Internet. Cependant, étant donné son actuelle croissance, **le Chinois** aura pris le prix d'ici 2014.

Langues les plus utilisées sur Internet (Mai 2011)



247 milliards D'E-MAILS sont envoyés chaque jour (80% d'entre eux sont des spam).

Le trading haute fréquence.

De la même façon que l'étude de l'activité sur Twitter a donné aux publicités, agences et journalistes des informations précieuses sur le traitement de texte et le tonnerre au Japon, grâce à des algorithmes, utilise le Big Data pour suivre les tendances et saisir le plus vite possible toutes les opportunités.

En une fraction de seconde, ces algorithmes spécifiques prennent la décision d'acheter ou de vendre une marchandise.

L'installation de nouveaux câbles sous l'Atlantique va permettre de gagner **5 millisecondes** sur les 62 millisecondes actuellement nécessaires à un ordre de transaction pour voyager de New-York à Londres.

Avec de nouveaux câbles optiques.

L'aller-retour Londres-New-York va nécessiter plus que 99,9 millisecondes.

Cette économie de 3 millisecondes représente plusieurs millions de dollars pour les entreprises de trading qui utilisent cette nouvelle fibre (et qui sont prêtes à investir des millions pour en profiter).

Comment économiser 3 millisecondes

Le profondeur de l'océan Atlantique varie.

Les nouveaux câbles seront positionnés sur des fonds océaniques moins profonds qu'actuellement, avec une différence pouvant aller jusqu'à 300 mètres. Cette nouvelle route permettra de réduire la longueur de câble nécessaire et donc de gagner du temps dans le transport de l'information.

Les nouvelles fibres sous-marines sont installées sous un profond et s'enfoncent dans une route plus directe.

10% de toutes les e-mails jamais envoyés (jusqu'en 2011)

60% de l'humanité (5,4 milliards de personnes) utilise activement les sms.

En 2012, 150 000 messages ont été envoyés chaque semaine.

50% des e-mails commerciaux envoyés de 3 ans ont accès à un smartphone.



1970s KO Analyse des données (multivariate analysis)

•1980s MO IA, réseaux de neurones, Stat. fonctionnelle

•1990s GO 1^{er} changement de paradigme : **Data Mining** Modèles prédictifs et **données pré-acquises**

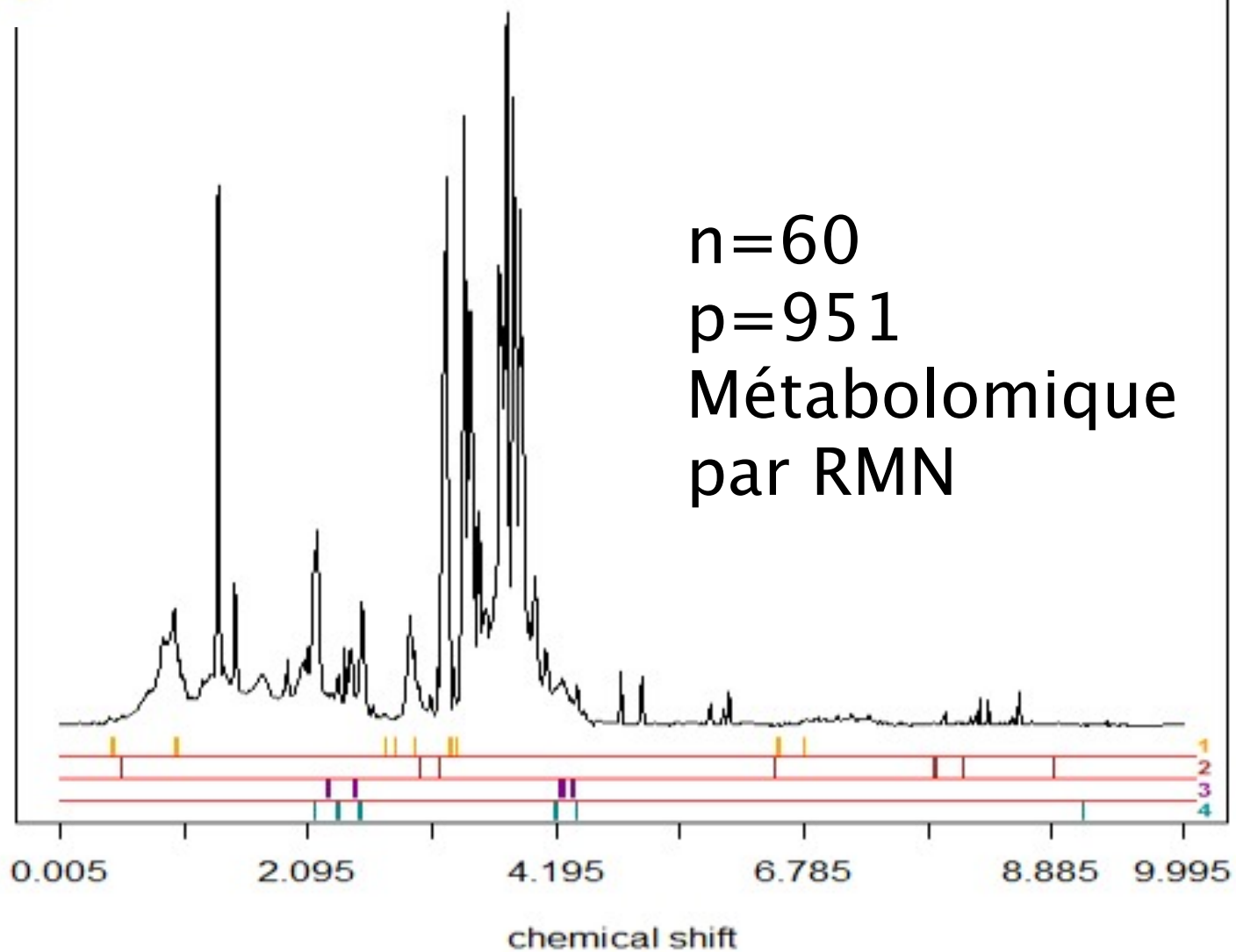
•2000s TO 2^{ème} chang. de paradigme : **Apprent. Statistique** $p \gg n$, parcimonie, Biais² + Variance

•2010s PO 3^{ème} chang. de paradigme : **Big Data**
E(approximation)+E(estimation)+**E(optimisation)**



Apprentissage Statistique et omiques

- $p \gg n$ ($p \# 10^4$, $n \# 10$)
- Tests multiples
- Sélection de modèles par pénalisation
- Sélection de variables (biomarqueurs)
- SVM, boosting, random forest...
- Inférence de graphes





Applications du Big Data

- E-commerce (1000 merci, tinyclues...)
- Géolocalisation (Médiamobile, Datasio...)
- Industries (Total, Geosys...)

- Structures complexes des données industrielles
 - graphes, signaux, images, fonctions...
- Vitesse d'acquisition (VVV)

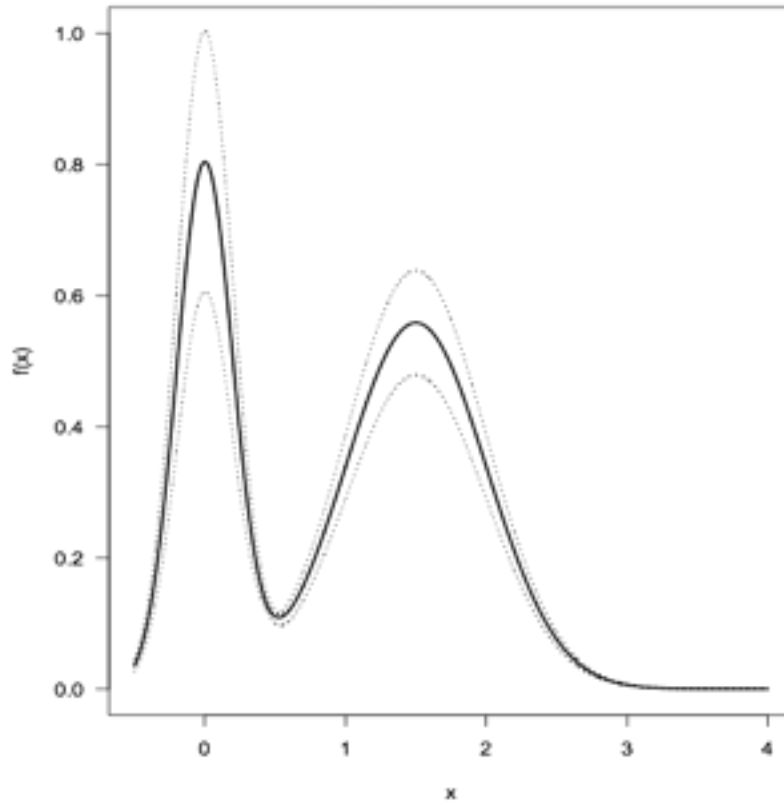


Problèmes mathématiques et/ou Informatiques

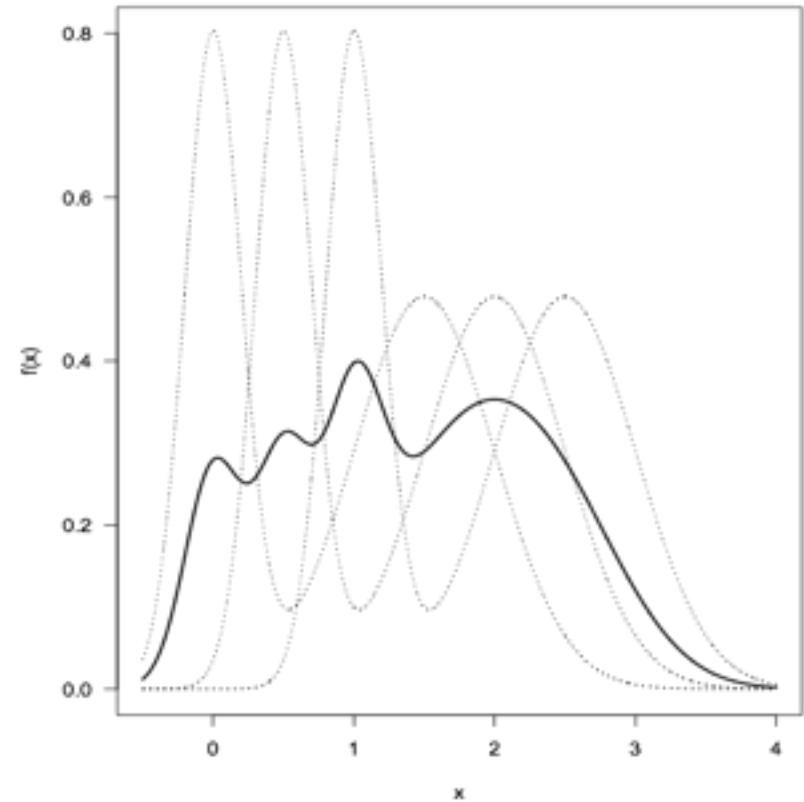
- E(Approximation) + E(Estimation) + E(Optimisation)
- Déterminer La structure mathématique « géométrie » ou distance adaptée
- Trouver des invariants ou comportements reproductibles
- Vitesse d'acquisition vs. Décision séquentielle ou adaptative

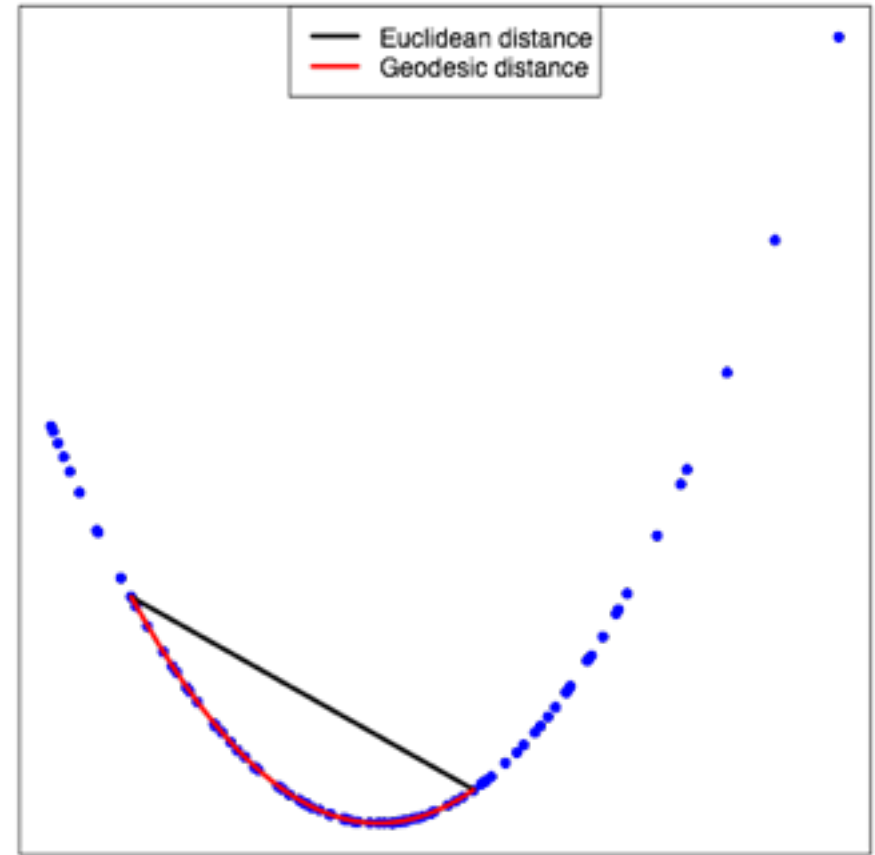
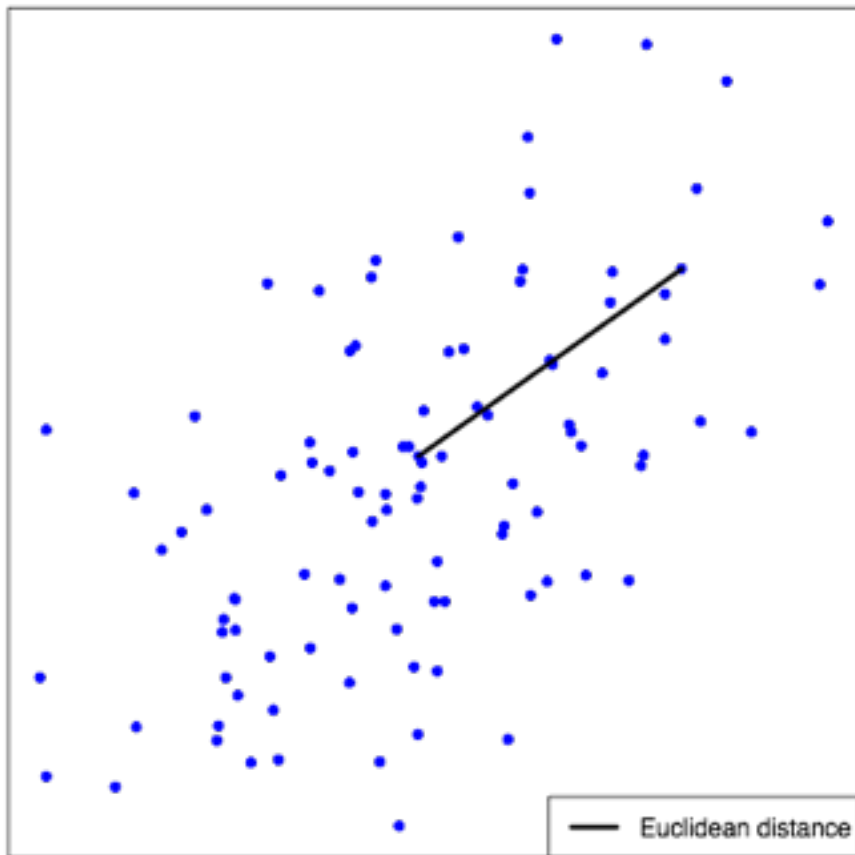


Amplitude variation



Phase variation







DATASIO

Géolocalisation des taxis à SF



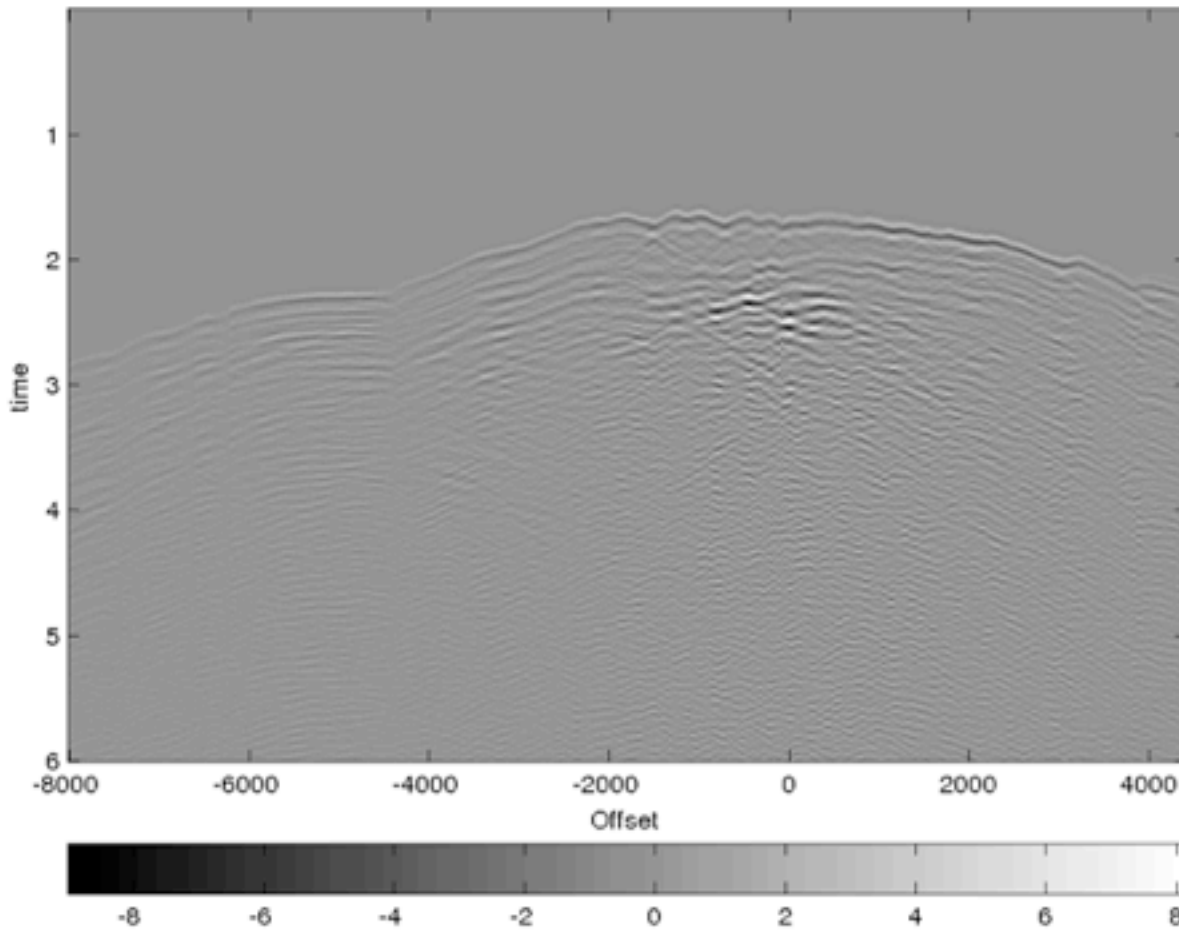
Nb cluster= 10 et Nb trajet tot= 122





Caractérisation de simulations de tirs

Energie mesurée pour la source 1 pour tous les capteurs alignés en ligne lointaine





Systemes de recommandation multi agents





Mathématiques utilisées 1/2

- **Géométrie** : trouver le bon espace ou un modèle de déformations
- **Statistique** : estimer la distance géodésique
- **Théorie des graphes** : estimation de la distance en trouvant les « chemins minimaux » entre les points
- **Optimisation** : calculer la distance minimale



Mathématiques utilisées 2/2

- **Clustering** : trouver des groupes similaires au sens de la distance
- **Statistique** : étudier la variabilité, plans d'expérience dynamiques, numériques
- **Apprentissage** (on line): prévoir l'appartenance au groupe, apprentissage renforcé
- **Probabilités** : analyse de graphes, matrices aléatoires, chaînes de Markov...



Conclusion très provisoire

- Beaucoup de sous-disciplines des mathématiques et d'interactions
- Connexions Mathématiques / Informatique
- Problème industriel = problème de recherche