

Le point de vue d'un algorithmicien sur Science des Données

Michel Habib

habib@liafa.univ-paris-diderot.fr

<http://www.liafa.univ-paris-diderot.fr/~habib>

Réunion des 4 Pôles USPC, 30 novembre 2015

Plan de l'exposé

Remarques liminaires

Plan de l'exposé

Remarques liminaires

Une évolution naturelle de l'informatique

Plan de l'exposé

Remarques liminaires

Une évolution naturelle de l'informatique

La rupture

Plan de l'exposé

Remarques liminaires

Une évolution naturelle de l'informatique

La rupture

Statistiques sur les graphes

Plan de l'exposé

Remarques liminaires

Une évolution naturelle de l'informatique

La rupture

Statistiques sur les graphes

Conclusions

Remarques liminaires

Une évolution naturelle de l'informatique

La rupture

Statistiques sur les graphes

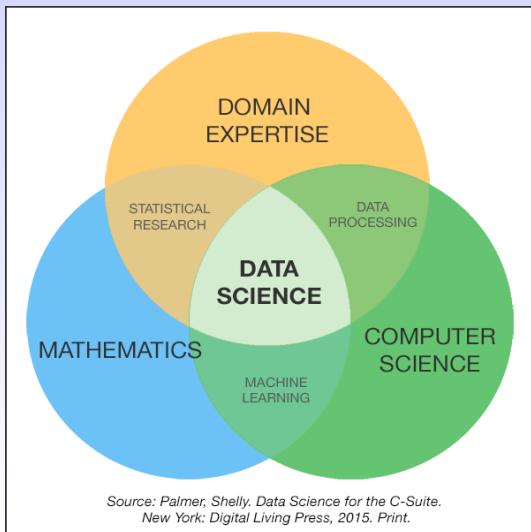
Conclusions

- ▶ Science des données : éviter les recettes de cuisine et les effets de mode, trouver les concepts durables.

- ▶ Science des données : éviter les recettes de cuisine et les effets de mode, trouver les concepts durables.
- ▶ i.e., l'apprentissage de technologies trop rapidement périssables
(cf. en micro-électronique selon pôle emploi, un ingénieur nouvellement formé n'est compétent que pendant 5 ans)

- ▶ Science des données : éviter les recettes de cuisine et les effets de mode, trouver les concepts durables.
- ▶ i.e., l'apprentissage de technologies trop rapidement périssables
(cf. en micro-électronique selon pôle emploi, un ingénieur nouvellement formé n'est compétent que pendant 5 ans)
- ▶ L'outil est éphémère le concept est durable ...

Une définition pas très pluridisciplinaire



Remarques liminaires

Une évolution naturelle de l'informatique

La rupture

Statistiques sur les graphes

Conclusions

- ▶ S'adapter aux données dont les tailles grandissent vertigineusement

- ▶ S'adapter aux données dont les tailles grandissent vertigineusement
- ▶ Cela fait 70 ans que les informaticiens le font.

Un disque de 5 méga-octets en 1956



Remarques liminaires

Une évolution naturelle de l'informatique

La rupture

Statistiques sur les graphes

Conclusions

- ▶ Toutefois la croissance exponentielle de la collecte des données a changé la donne.

- ▶ Toutefois la croissance exponentielle de la collecte des données a changé la donne.
- ▶ Il faut une nouvelle informatique pour traiter ces données gigantesques.

Le blues d'un agent de la NSA

Avalanche de données sur l'internet en 1s :

- ▶ 220 000 messages sur Whatsapp

Le blues d'un agent de la NSA

Avalanche de données sur l'internet en 1s :

- ▶ 220 000 messages sur Whatsapp
- ▶ 23 000 minutes de vidéos échangées sur Skype

Le blues d'un agent de la NSA

Avalanche de données sur l'internet en 1s :

- ▶ 220 000 messages sur Whatsapp
- ▶ 23 000 minutes de vidéos échangées sur Skype
- ▶ 52 000 likes sur Facebook

Le blues d'un agent de la NSA

Avalanche de données sur l'internet en 1s :

- ▶ 220 000 messages sur Whatsapp
- ▶ 23 000 minutes de vidéos échangées sur Skype
- ▶ 52 000 likes sur Facebook
- ▶ On peut parler de masse de données.

Le blues d'un agent de la NSA

Avalanche de données sur l'internet en 1s :

- ▶ 220 000 messages sur Whatsapp
- ▶ 23 000 minutes de vidéos échangées sur Skype
- ▶ 52 000 likes sur Facebook
- ▶ On peut parler de masse de données.
- ▶ Sans compter les messages électroniques, toutes les images des caméras et autres satellites

Ce qu'il faudrait développer tant en recherche qu'en enseignement

- ▶ De très grandes machines parallèles (peu de recherche en informatique sur ce point au sein d'USPC (C. Cerin P13), mais sujet présent à INRIA)

Ce qu'il faudrait développer tant en recherche qu'en enseignement

- ▶ De très grandes machines parallèles (peu de recherche en informatique sur ce point au sein d'USPC (C. Cerin P13), mais sujet présent à INRIA)
- ▶ Des moyens de communication pour échanger ces données (les débits de communication ne suivent pas la croissance des données)

Ce qu'il faudrait développer tant en recherche qu'en enseignement

- ▶ De très grandes machines parallèles (peu de recherche en informatique sur ce point au sein d'USPC (C. Cerin P13), mais sujet présent à INRIA)
- ▶ Des moyens de communication pour échanger ces données (les débits de communication ne suivent pas la croissance des données)
- ▶ De nouvelles méthodes de stockage, de compression et d'accès à ces données (par exemple données distribuées accessibles à l'aide de Hadoop, Map Reduce) Peu d'expérience sur ce sujet dans le monde académique.

Ce qu'il faudrait développer tant en recherche qu'en enseignement

- ▶ De très grandes machines parallèles (peu de recherche en informatique sur ce point au sein d'USPC (C. Cerin P13), mais sujet présent à INRIA)
- ▶ Des moyens de communication pour échanger ces données (les débits de communication ne suivent pas la croissance des données)
- ▶ De nouvelles méthodes de stockage, de compression et d'accès à ces données (par exemple données distribuées accessibles à l'aide de Hadoop, Map Reduce) Peu d'expérience sur ce sujet dans le monde académique.
- ▶ Les méthodes d'analyse des données : fouille de données (Themis Palpanas P5) et l'apprentissage (Yann Chevaleyre P13, Mohamed Nadif P5)

Mais aussi et la liste n'est pas ordonnée

- ▶ Les méthodes de visualisation de données

Mais aussi et la liste n'est pas ordonnée

- ▶ Les méthodes de visualisation de données
- ▶ L'algorithmique, telle par exemple l'algorithmique à la volée (streaming, Frédéric Magniez au LIAFA)

Mais aussi et la liste n'est pas ordonnée

- ▶ Les méthodes de visualisation de données
- ▶ L'algorithmique, telle par exemple l'algorithmique à la volée (streaming, Frédéric Magniez au LIAFA)
- ▶ La protection des données : les méthodes de cryptographie et d'anonymisation (est-ce vraiment possible ?)

Mais aussi et la liste n'est pas ordonnée

- ▶ Les méthodes de visualisation de données
- ▶ L'algorithmique, telle par exemple l'algorithmique à la volée (streaming, Frédéric Magniez au LIAFA)
- ▶ La protection des données : les méthodes de cryptographie et d'anonymisation (est-ce vraiment possible ?)
- ▶ Les libertés individuelles (Pourquoi Google archive tout sur nous ?) Moi j'ai plein de choses à cacher.

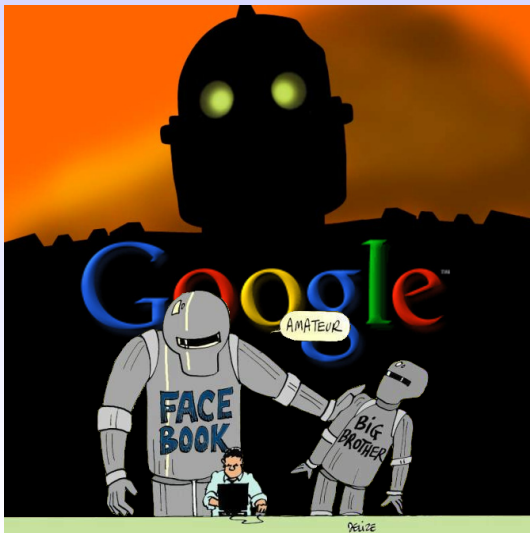
Mais aussi et la liste n'est pas ordonnée

- ▶ Les méthodes de visualisation de données
- ▶ L'algorithmique, telle par exemple l'algorithmique à la volée (streaming, Frédéric Magniez au LIAFA)
- ▶ La protection des données : les méthodes de cryptographie et d'anonymisation (est-ce vraiment possible ?)
- ▶ Les libertés individuelles (Pourquoi Google archive tout sur nous ?) Moi j'ai plein de choses à cacher.
- ▶ Quels sont les modèles économiques compatibles avec la démocratie ?

Mais aussi et la liste n'est pas ordonnée

- ▶ Les méthodes de visualisation de données
- ▶ L'algorithmique, telle par exemple l'algorithmique à la volée (streaming, Frédéric Magniez au LIAFA)
- ▶ La protection des données : les méthodes de cryptographie et d'anonymisation (est-ce vraiment possible ?)
- ▶ Les libertés individuelles (Pourquoi Google archive tout sur nous ?) Moi j'ai plein de choses à cacher.
- ▶ Quels sont les modèles économiques compatibles avec la démocratie ?
- ▶ ...

Big Brother est déjà installé



Pourquoi je me sens concerné par le sujet

- ▶ Spécialiste d'algorithmique de graphes je m'intéresse aux algorithmes sur les très grands graphes (par ex. graphe des amis de Facebook)

Pourquoi je me sens concerné par le sujet

- ▶ Spécialiste d'algorithmique de graphes je m'intéresse aux algorithmes sur les très grands graphes (par ex. graphe des amis de Facebook)
- ▶ Calcul de diamètre et centres de grands graphes
<https://who.rocq.inria.fr/Laurent.Viennot/road/>

Pourquoi je me sens concerné par le sujet

- ▶ Spécialiste d'algorithmique de graphes je m'intéresse aux algorithmes sur les très grands graphes (par ex. graphe des amis de Facebook)
- ▶ Calcul de diamètre et centres de grands graphes
<https://who.rocq.inria.fr/Laurent.Viennot/road/>
- ▶ Participation au projet ANR Algopol avec le sociologue Dominique Cardon (Orange)

Pourquoi je me sens concerné par le sujet

- ▶ Spécialiste d'algorithmique de graphes je m'intéresse aux algorithmes sur les très grands graphes (par ex. graphe des amis de Facebook)
- ▶ Calcul de diamètre et centres de grands graphes
<https://who.rocq.inria.fr/Laurent.Viennot/road/>
- ▶ Participation au projet ANR Algopol avec le sociologue Dominique Cardon (Orange)
- ▶ 2 cours en Master Informatique P7 sur des sujets Data Science Grands Réseaux d'interaction, Méthodes et Algorithmes pour l'accès à l'Information Numérique (Moteur de recherche, algorithmes de recommandation ...)

Remarques liminaires

Une évolution naturelle de l'informatique

La rupture

Statistiques sur les graphes

Conclusions

Statistiques sur les graphes, un joli sujet de recherche

- Echantillonnage de graphes, "Property testing" = statistiques sur des graphes plutôt que sur des nombres (points forts actuels : US et Israel).

Statistiques sur les graphes, un joli sujet de recherche

- ▶ Echantillonnage de graphes, "Property testing" = statistiques sur des graphes plutôt que sur des nombres (points forts actuels : US et Israel).
- ▶ Limite de graphes : un sujet difficile quelques résultats fondamentaux : Livre de L. Lovász

Statistiques sur les graphes, un joli sujet de recherche

- ▶ Echantillonnage de graphes, "Property testing" = statistiques sur des graphes plutôt que sur des nombres (points forts actuels : US et Israel).
- ▶ Limite de graphes : un sujet difficile quelques résultats fondamentaux : Livre de L. Lovász
- ▶ Lemme de régularité de E. Szemerédi, Prix Abel 2012.

Statistiques sur les graphes, un joli sujet de recherche

- ▶ Echantillonnage de graphes, "Property testing" = statistiques sur des graphes plutôt que sur des nombres (points forts actuels : US et Israel).
- ▶ Limite de graphes : un sujet difficile quelques résultats fondamentaux : Livre de L. Lovász
- ▶ **Lemme de régularité de E. Szemerédi**, Prix Abel 2012.
- ▶ Une théorie des graphes clairsemés ou peu denses, P. Ossona de Mendès et J. Nešetřil (Collaboration avec Prague, LEA CNRS)

Statistiques sur les graphes, un joli sujet de recherche

- ▶ Echantillonnage de graphes, "Property testing" = statistiques sur des graphes plutôt que sur des nombres (points forts actuels : US et Israel).
- ▶ Limite de graphes : un sujet difficile quelques résultats fondamentaux : Livre de L. Lovász
- ▶ **Lemme de régularité de E. Szemerédi**, Prix Abel 2012.
- ▶ Une théorie des graphes clairsemés ou peu denses, P. Ossona de Mendes et J. Nešetřil (Collaboration avec Prague, LEA CNRS)
- ▶ Modèles de graphes aléatoires (apports récents très important des physiciens)

Remarques liminaires

Une évolution naturelle de l'informatique

La rupture

Statistiques sur les graphes

Conclusions

Science des données : un sujet pluridisciplinaire, cf. Sergueï Brin et Larry Page (Google)

1. **Sociologie** : Ils se sont inspiré d'une idée de sociologues sur les indices de citations (G. Pinsky et F. Narin 1976), qu'ils ont appliqué aux pages web.

Science des données : un sujet pluridisciplinaire, cf. Sergueï Brin et Larry Page (Google)

1. **Sociologie** : Ils se sont inspiré d'une idée de sociologues sur les indices de citations (G. Pinsky et F. Narin 1976), qu'ils ont appliqué aux pages web.
2. **Mathématiques** : Ils ont formalisé cette idée et l'ont ramené à un calcul de valeurs propres, ont croisé le théorème de Perron-Frobenius (1900), les chaînes de Markov.

Science des données : un sujet pluridisciplinaire, cf. Sergueï Brin et Larry Page (Google)

1. **Sociologie** : Ils se sont inspiré d'une idée de sociologues sur les indices de citations (G. Pinsky et F. Narin 1976), qu'ils ont appliqué aux pages web.
2. **Mathématiques** : Ils ont formalisé cette idée et l'ont ramené à un calcul de valeurs propres, ont croisé le théorème de Perron-Frobenius (1900), les chaînes de Markov.
3. **Informatique** : Puis ils se sont construit un système informatique distribué, efficace et adapté.

Deux notes optimistes

1. Les mathématiciens et informaticiens de P7 sont prêts à avoir un rôle moteur dans un projet de master Science des données USPC.

Deux notes optimistes

1. Les mathématiciens et informaticiens de P7 sont prêts à avoir un rôle moteur dans un projet de master Science des données USPC.
2. Dans le cadre des projets USPC, nous avons obtenu une allocation pluridisciplinaire (Math-Info) de doctorat pour étudier avec E. Birmele (P5) : Les algorithmes de classification de grandes masses de graphes avec comme premier jeu de données des graphes venant de la biologie.