Trois exemples tirés d'un article paru dans Libération pour se mettre en bouche

(http://www.liberation.fr/economie/2012/12/03/donnees-le-vertige_864585)

«A Singapour, une étude menée en 2012 a croisé les données GPS de 16 000 taxis avec les relevés météo et montré que les chauffeurs s'arrêtent de rouler dès les premières gouttes de peur d'être impliqués dans un accident et de devoir payer un malus d'assurance élevé»

«un Américain a découvert la grossesse de sa fille en voyant la teneur des publicités hyperciblées, envoyées par les commerçants sur la base de l'examen de ses tickets de caisse»

«Les téléspectateurs font souvent autre chose pendant qu'ils regardent une émission : ils vérifient les informations diffusées, commentent sur les réseaux sociaux… Et nous arrivons à dire combien tweetent en regardant le Grand Journal, puis zappent sur Secret Story»

Big data, potentiel actuel et futur d'un phénomène émergent

- Big data : de quoi parle-t-on exactement ?
- La question du potentiel aujourd'hui et dans le futur : pas qu'une question de capacité de traitement
- Les limites actuellement perceptibles : technologiques, socio-institutionnelles, juridiques, auto-construites
- Les perspectives et enjeux, aujourd'hui, mais aussi demain
- Les marchés porteurs, tendances et options envisageables dans le futur

1. Big data : de quoi parle-t-on exactement ?

- Définitions: déjà les premiers problèmes:

  Prop. => Capacité de traitement et domaines d'application mettant en jeu des volumes de données et des méthodes de traitement sortant de l'ordinaire de l'offre commerciale en software/hardware et des savoir-faire, méthodes, de traitement courants pour un temps donné

- Gartner: la tri-dimensionnalité des 3 «V»s: (volume, vélocité et variété)

Problème # 1: l'obsolescence

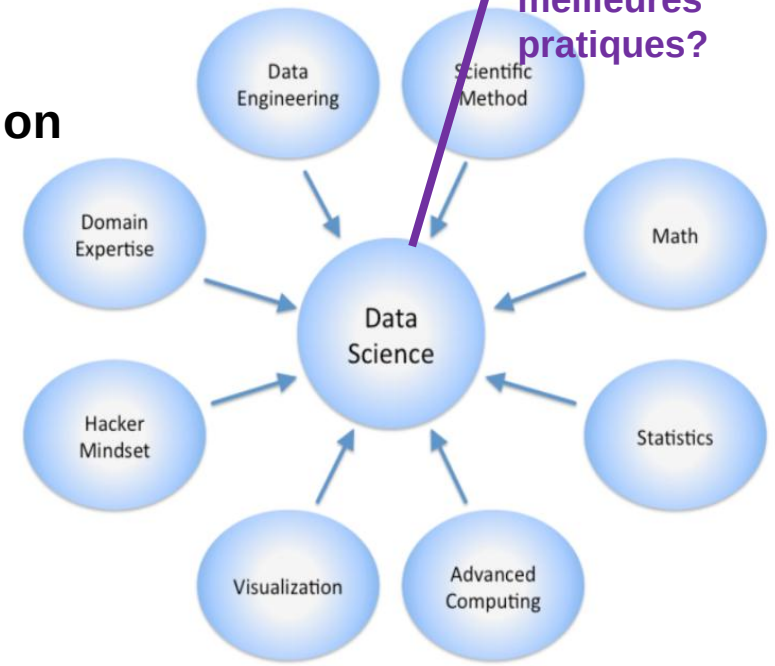
- D'autres ont essayé d'ajouter: la variabilité, la complexité ou … la véracité

Problèmes # 2 et 3: capacités discutables…… vers des super-pouvoirs?

Plus concrètement: un phénomène conceptuel et applicatif mettant en valeurs les «data sciences»

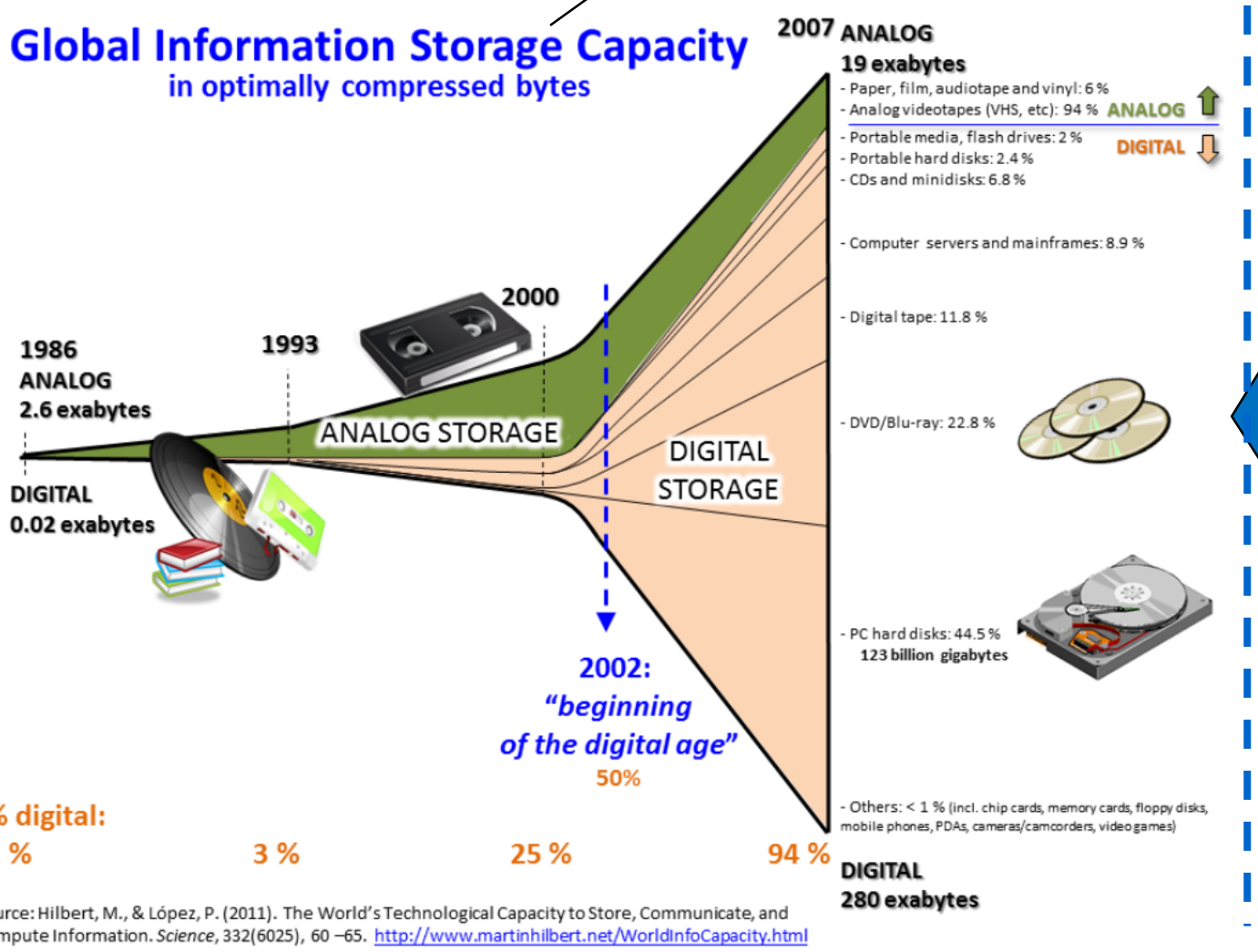
- Des volumes très variables (un facteur > 1000 de différence!),
- Une variété de sources … plutôt finie,
- **Une vélocité de traitement en augmentation rapide, basée sur <u>une combinaison de capacités</u>**:
  - Hardware en constante augmentation de puissance (traitement, transfert, entreposage, sécurisation)
  - Software, architectures, paradigmes de travail et multiples méthodes combinées pour gérer de gros volumes de données

«Tricks» ou procédures standard et meilleures pratiques?



- **De très nombreux domaines d'application bénéficiaires de ces capacités**: science (physique des particules, protéomique, séquençages génétiques, etc.), criminologie, médecine prédictive et autres questions de santé, météorologie, finance, marketing, réseaux sociaux, gestion de trafic dans plusieurs domaines comme l'énergie, les télécoms, les transports, la logistique, etc.

L'évolution de l'acquisition et des sources



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

Il faudra ajouter la production de données de **l'Internet des objets, > 20 millards en 2020**

2. La question du potentiel aujourd'hui et dans le futur : pas qu'une question de capacité de traitement

Big =

- Bcp de données, mais pas unique critère
- Méthodes multiples, notion de donnée innovante*
- Multiples supports et formats de «sourcing»
- Concept liant capacités de traitement et domaines d'applications
- «Narrative» de prédictibilité, véracité et pouvoirs analytiques sans limites

Mais aussi, co-tendances:

- Data sciences
- «Quantified self»
- Clouds et data centers
- Open data et transparence
- Internet of Things
- Lois de Moore et Koomey
- Fight against terrorism, etc.

> «anonymisables»

*
- Déduction > induction (ou statistique inférentielle)
- Données à basse densité
- « Curation" (identification, contrôle qualité…) de données dans leur format le plus natif possible pour accroître leur partage par des communautés différentes
- Visualisation

3. Les limites actuellement perceptibles : technologiques, socio-institutionnelles, auto-construites

Technologiques:

- Gourmandise du data science pour le grand nombre vs. rasoir d'Ockham,
- Données = le passé
- Prédictibilité dans les problèmes complexes = discutable (corrélations ≠ causalité)

Socio-institutionnelles (juridiques?)

- Sphère privée bousculée, «ownership» = problème multi-facettes !
- Opacité (absence de traçabilité complète pour «sources data» et méthodes)!
- Contrôle/gouvernance des processus Big Data, risque anti-démocratiqueü

Auto-construites:

- Le succès s'imite, produit de la convergence, limitant l'avantage compétitif
- Le paradoxe du forecast qui construit les conditions de son contournement
- Arrogance, leurres? on peut produire des chiffres impressionnants pour des problèmes dont on ne peut définir raisonnablement les conditions initiales

Difficultés juridiques actuelles

- Propriété/protection des données individuelles OK, mais: quid des corrélations et des métadonnées?
- Anonymisation irréversible: techniques existent, mais statut juridique (ex: recommandations UE) ?
- Discrimination en raison de l'appartenance à un groupe traité comme comportement de masse = difficile en Suisse car pas de «class action» possible
- Difficulté d'établir la responsabilité vis-à-vis d'algorithmes et du travail des robots informatiques

De la prédiction

Certains problèmes conviennent à l'exigence de prédictibilité (simulations, prédictions statistiques basées sur des données fiables), **d'autres moins** (problème vraiment complexes), d'autre encore … dans une certaine mesure seulement (météorologie)

Pour simplifier le débat technique, cette question tend à opposer aujourd'hui **ceux qui croient aux dragons-rois et ceux qui croient aux signes noirs** ou encore les approches plutôt déterministes d'une part et les approches plutôt constructivistes d'autre part

Mais le Big Data ne s'embarrasse pas souvent de ces nuances et s'affirme plutôt comme pluri-potent:

- Quand c'est vrai: potentiel formidable et marchés durables
- Quand ça l'est moins: dangers et marchés fragiles (marketing trop optimiste, risque d'erreur grave, points de basculement)

En résumé: du potentiel… mais aussi des carrefours, des virages dangereux et peut-être des impasses /1

Du potentiel actuel (plusieurs modèles d'affaire):

- **Traitement des grands nombres de données** (sciences)
- **Monitoring trafics** en tous genres (transports, énergie, consommation d'eau, déchets, télécoms, assurances, santé),
- **Prédictions probabilistes** efficace dans des domaines très variés (logistique, médecine, criminalité, météo, processus industriels, agriculture, etc.
- **Identifications de formes:** de type «signatures», **mais aussi de corrélations cachées, de motifs peu visibles** (santé, comportements de masse, consommation, sécurité, analyse de risques, environnement, etc.) => passé et temps réel «épais et riche» (avec des acteurs comme Google ou Twitters comme «listening posts» ou même peer-to-peer)

En résumé: du potentiel… mais aussi des carrefours, des virages dangereux et peut-être des impasses /2

Du potentiel futur:

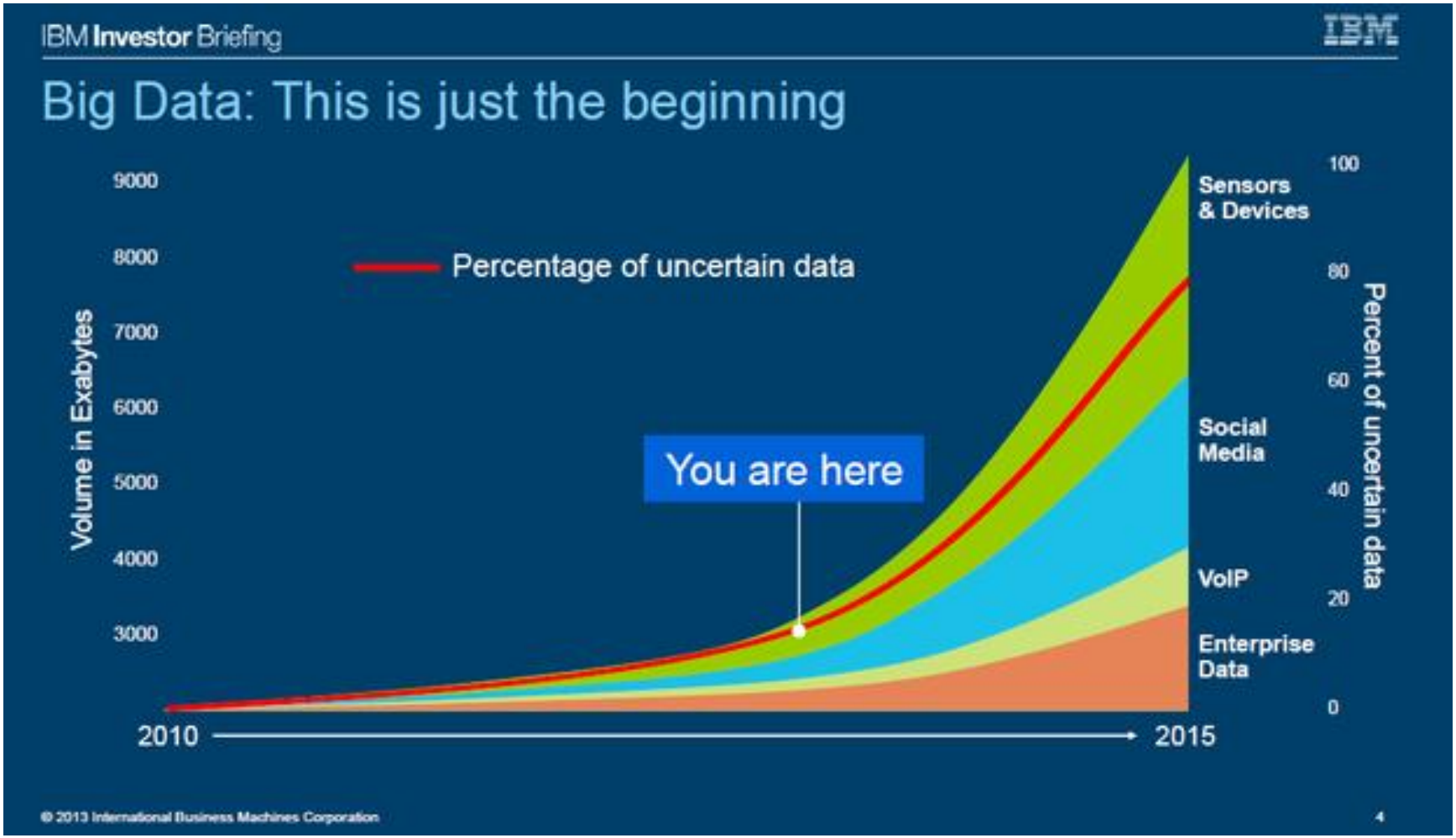
- **Toujours plus de la même chose: la puissance et l'efficience technologique continuera de s'accroître***, dans un monde hyper-connecté et quantifié

  - Gartner: les entreprises qui auront intégré toutes les dimensions du Big Data d'ici à 2015 seront plus performantes de 20% par rapport à leurs concurrentes
  - Gartner: 40 zettaoctets (1 million de millards de millards) stockés en 2020
  - AFDEL (2013): le Big Data devrait représenter 8% du PIB européen en 2020

Mais, nécessité d'apprentissage vis-à-vis des erreurs, leurres, paradoxes et dérives possible: p. ex: faire du Big Data pour spéculer «à la baisse», détecter des failles et faiblesses et en profiter, détourner des données déjà collectées

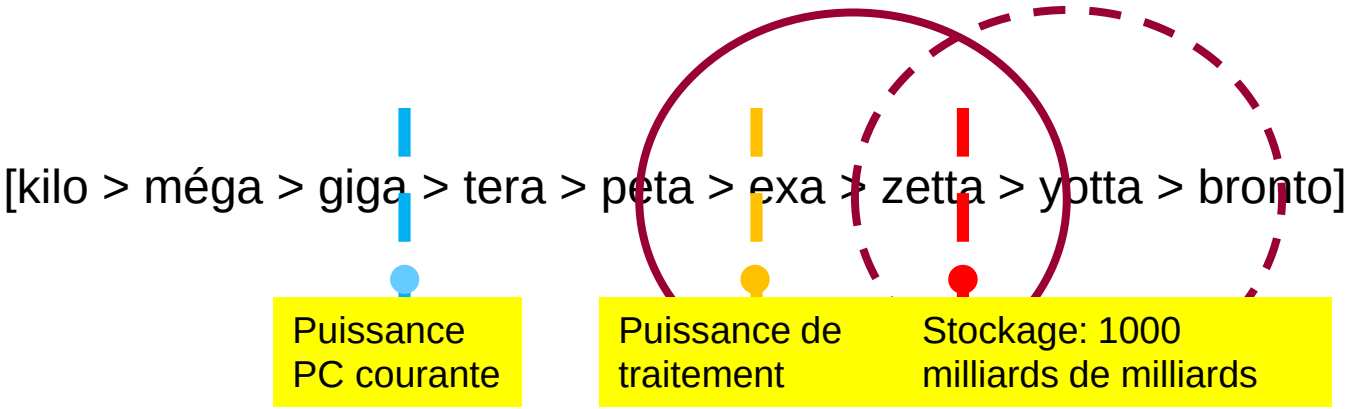
*
- Informatique (nano-, opto-, quantique, ADN, …)
- Matériaux (nano-tubes)
- Architectures innovantes, combinantes, flexibles
- Nouveaux paradigmes (super-grid, brain)

De quoi méditer un peu, avec IBM!



http://www.zdnet.com/ibm-eyes-china-south-america-africa-and-big-data-for-2015-growth-7000011976/

4. Les perspectives et enjeux, aujourd'hui, mais aussi demain



Principaux enjeux:

- La gestion des robots (hard et soft)
- La traçabilité, les déclarations de sources et procédés, et les bonnes pratiques (notamment vis-à-vis des brokers): «surveiller les surveilleurs»
- La mémoire et l'éternité des données
- Pays dominants / pays dépendants
- L'empreinte carbone et hydrique du Big Data !!!
- Pour une gouvernance plus pro-active que réactive, besoin de recherche

5. Les marchés porteurs, tendances et options envisageables dans le futur

- Domaines phares: **santé, agriculture, questions d'environnement, sécurité, énergie, déchets, logistique**
- **Capacités du Big Data appliquées aux micro-marchés** («Small Big Data»)
- Ré-appropriation des données par «l'usager» (en réalité le fournisseur), vers des **options Big Data2.0** ?
- Applications du **Big Data dans des domaines de plus en plus qualitatif**, avec du potentiel innovant, mais aussi dérives possibles

Merci de votre attention!