Big Data
Opportunity & Risk

Nozha Boujemaa
Director of Research
Advisor to The CEO of Inria in Big Data
Member of BDVA Board Of Directors

Introduction

Data-driven Digital Transformation

- Data-driven research and innovation have led to the rise of novel economical paradigms.
- The historical leaders of existing economic sectors (transportation, insurance, tourism etc.) no longer have the guarantee to remain in their leading positions given the agility of new competitors with data-based innovative economic models.
- Data technologies are not only transforming traditional industries, they are changing our way of life, our social structures and even our entire social landscape.
- Using data for the benefit of mankind is not only a technological, but also an ethical, legal, economic and social challenge.

Emergence of Big Data Technologies

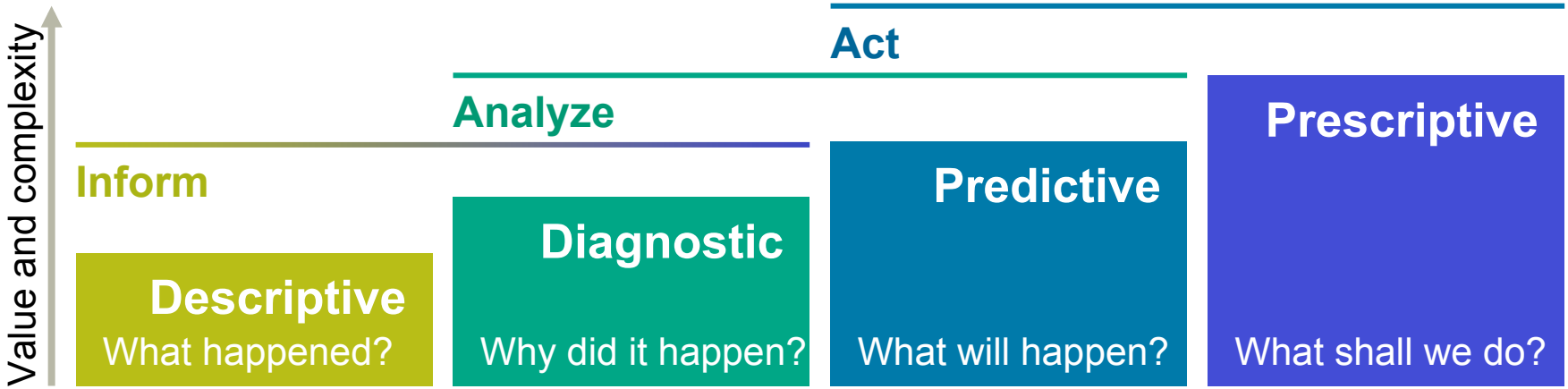
Convergence of three factors:

- **Data Tsunami**
- **Affordable/Powerful Computing Facilities,** including open-source software framework
- Advanced Machine Learning algorithms and paradigms, mainly « **Deep Learning** » registering significant performance gain (about 15% wrt SoA techniques since 2 years)

These are **enablers** for Artificial Intelligence (AI) capabilities

From « Data Analytics » to « Cognitive Systems »

Focus of data analytics is changing – From description of past to decision support



Examples

Descriptive	Diagnostic	Predictive	Prescriptive
– Plant operation report – Fault report	– Alarm management – Root cause identification	– Power consumption prediction – Fault prediction	– Operation point optimization – Load balancing

Gartner 2013 - N. Gauss/Siemens - 2015

Opportunités

Quelques domaines d'application phares:

- Marketing digital/CRM, analyse de traces pour ciblage publicitaire (preuve de concept apportée par les GAFA), recommandations dans les plateformes de services en ligne
- Maintenance prédictive pour l'industrie 4.0, Véhicule connecté
- « Smart Cities », « Smart Factories », « Smart-Home »,
- Energie: production, distribution,
- Santé: aide au diagnostic médical, « quantified-self », etc
- Observation de la Terre, environnement, ressources naturelles
- Sécurité: détection de signaux faibles
- Etc

Drivers et freins

Drivers:

- Les performances croissantes des sciences des données:
- Les modèles d'affaires pour la création de valeur

=> Un pré-requis: la dualité donnée-algorithme

Freins:

- **Confiance** et appropriation: **gouvernance** des données, la **transparence** des algorithmes
  - La véracité des données (bruit intentionnel ou non), le temps réel,
  - La qualité et les conditions d'utilisation des algorithmes
- Les compétences interdisciplinaires en science des données

Vulnérabilités

- Certaines plateformes dominantes sur le marché joue un rôle de « prescripteur » en orientant une grande part du trafic utilisateur, les mécanismes de tri, de sélection de contenus proposés parfois opaques
- L'utilisation faites des données personnelles des utilisateurs n'est pas toujours explicite, peut conduire à des pratiques illicites de différentiation des prix

- Quel garanties d'impartialité à des facteurs commerciaux? (appli. GPS, etc), Quel impact en temps de crise?
- Quelle garantie de non discrimination (droits citoyens)?,
  - *technique: l'apprentissage supervisé est un système à réaction positive*
  - *Quels critères, quelles données*
- Concurrence déloyale et position dominante des grands acteurs

DATA SCIENCE
PILLARS & CHALLENGES

Data Science
5 Pillars

5 Pilars for Data Science*

1- **Data Management**: unstructured and semi-structured

- Semantic interoperability of heterogeneous sources and representations, Data quality, Data lifecycle, Data provenance, Integration of data and business process

2- **Data Processing Architecture** :

- Scalability, Real-Time analytics and on-the-fly processing, Event processing (Hadoop, Spark, etc)
- Decentralization (Cloud/Fog etc)
- Performance, Low-energy consumption through better integration between hardware and software

* Inspired by BDVA SRIA technical priorities

5 Pilars for Data Science

3- **Data Analytics**:

- Semantic and Knowledge-based Analysis (i.e. sentiment, semantics, ontology engineering etc.), Real-Time interlinking and automatic annotation, scalable and incremental reasoning, linked data mining, cognitive computing
- Content Validation: Implementation of veracity (source reliability / information credibility) models for validating content (exp: exploiting content recommendations from unknown users), Outliers detection
- Predictive and Presciptive Analytics: Machine (Deep) Learning, clustering, pattern mining, network analysis and hypothesis testing techniques applied on extremely large graphs containing sparse, uncertain and incomplete data.

5 Pilars for Data Science

4- **Data Protection**:

- Data protection and anonymization is a major issue in the areas of Big Data and data analytics wrt person-specific and sensitive information
- Development of privacy-enhancing models and techniques, such as differential privacy, private information retrieval, syntactic anonymity, homomorphic encryption, secure search encryption, and secure multiparty computation, PIMS etc

⇒ Robusteness (against reversibility) and scalability,

⇒ Privacy-preserving data mining algorithms

5 Pilars for Data Science

5- **Data Visualization**

- Interactive visual analytics of multiple scale data: Appropriate scales of analysis are not always clear in advance, and single optimal solutions are unlikely to exist
- Collaborative, intuitive, and interactive visual interfaces: is required in order to foster effective exploitation of the information and knowledge that analytics can deliver.

  => effective communication and visualisation to enable collaborative decision making processes

- Cross-platform data visualization frameworks, including augmented reality visualization of data on mobile devices

Challenges

Challenges for Data Science
1- Progressive user-centric analytics

- **What –** Having analytics technology targeting the user needs and expectations, allowing the user to drive the analytics process effortlessly
  - real-time analytics and decision making
  - interactive mining, learning, visualization
  - On-line learning with few examples
  - user modeling and user intention models

- **Why** – Seamless cooperation between the machine and the analysts will facilitate the adoption of big data technology and the semantic effectiveness

Challenges for Data Science
2- Processing Architecture & Big Data,

- Optimized Architecture for energy consumption reduction
- Utilization within Embedded-Systems
- Less dependent to remote computing facilities (Cloud/Data Centers)
- Specialized Processors, GAFAM still pioneers: Google first announced such optimized architecture for TensorFlow (Its Open-Source Machine Learning Library) => Not for sale!

Challenges for Data Science
3- Responsible/Ethical Data Management and Analytics

It is often assumed that big data techniques are unbiased:

- because of the scale of the data
- because the techniques are implemented through algorithmic systems.

⇒ it is a mistake to assume they are objective simply because they are data-driven *

Consensus is emerging to develop methods and Tools to **build Trust & Transparency for Data and Algorithms**

* White House - Office of Science and Technology Policy Report « Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights », May 2016

- Federal Trade Commission Report: "Big Data: A Tool for Inclusion or Exclusion? January 2016

Challenges for Data Science
3- Responsible/Ethical Data Management and Analytics

Challenges of promoting fairness and overcoming the discriminatory effects of data technologies

1. Challenges relating to ***data used as inputs*** to an algorithm

2. Challenges related to ***the inner workings of the algorithm itself***

Challenges for Data Science
3- Responsible/Ethical Data Management and Analytics

Challenges 1: Data Inputs to an Algorithm

- *Poorly selected data*
- *Incomplete, incorrect, or outdated data*
- *Data sets that lack information or disproportionately represent certain populations*

Challenges for Data Science
3- Responsible/Ethical Data Management and Analytics

Challenges 2: The Design of Algorithmic Systems and Machine Learning

- *Poorly designed matching systems*
- *Personalization and recommendation services that narrow instead of expand user options*
- *Decision-making systems that assume correlation necessarily implies causation*
- *Unintentional perpetuation and promotion of historical biases*

Challenges for Data Science
3- Responsible/Ethical Data Management and Analytics

"**Data fundamentalism**"—the belief that numbers cannot lie and always represent objective truth—can present serious and obfuscated bias problems that negatively impact people's lives

=> Implementing the "**equal opportunity by design**" principle

Challenges for Data Science
3- Responsible/Ethical Data Management and Analytics

Contributions to AFNOR/DIN WG: **3 dimensions**

1. Trust and Transparency of data (**Provenance**): What information/data was used and where does it come from? Governance of data chain, who owns what, who can make value of what?
2. Trust and Transparency of data used and produced by algorithms (**Control**) : What data comes in and out of algorithms which are used in the big data pipeline?

Challenges for Data Science
3- Responsible/Ethical Data Management and Analytics

3- Trust and Transparency of computer-aided decision-making process **(decision responsibility)**: What are the different criteria/steps/settings that have led to the specific decision in order to understand the global path for the reasoning?

- "How can I trust Machine Learning prediction?" it happens to build the model of the pattern context rather the pattern itself
- Decision explanation and tractability
- Robustness to bias/diversion/corruption

Challenges for Data Science
3- Responsible/Ethical Data Management and Analytics

Several Open Source available

- Software reuse
- Inventor is not the user who deploy in a given context and application domain
- How to be confident in the relevance of parameters and mastering the piece of software?

3- Responsible/Ethical Data Management and Analytics :

- Très peu de travaux en France et en Europe sur le sujet. Un des aspects a été abordé dans le projet *CNIL-Inria* **Mobilitics**
- Consultation récente menée par le CGE missionné par le cabinet d'Axelle Lemaire (dans le cadre de la *loi pour la république numérique*) qui remis un rapport « Modalités de régulation des algorithmes de traitement des contenus », parmi les propositions:
  - La mise en place d'une Plateforme scientifique de test des algorithmes en vue de leur régulation/gouvernance.
  - La création d'une cellule de contrôle spécialisée « bureau des technologies de contrôle de l'économie numérique » pour l'ensemble des pouvoirs publics

3- Responsible/Ethical Data Management and Analytics :

- Groupe franco-allemand (AFNOR-DGE/DIN-BMWi) mis en place à la suite du sommet sur le numérique (Nov 2015) Merckel-Hollande => Recommandations « Best practices » sur les outils technologiques de la transparence, franco-anglais
- USA: OSTP (Office for Sciene and Technology Policy – White House), FTC (Federal Trade Commission),
- « Data Transparency Lab » depuis 1,5 an (MIT, Telefonica et Mozilla au board + Inria et Columbia en cours).
- Workshops « Fairness, Accountability, Transparency in Machine Learning » et « Data Transparency » co-localisé avec DTL'16
- NIPS'16: workshop Machine Learning & Law

En France:

- *En cours: Après ISN*… Proposition d'**Institut Convergence I2-DRIVE** : « **I**nterdisciplinary **I**nstitute for **D**ata **R**esearch: **I**ntelligence, **V**alue and **E**thics » , un de ses axes est « Trust & Transparency »
- Programme de travail sur **10 ans**, portée par l'Univ. Paris-Saclay, partenaires: Inria, l'X, CentraleSupelec, CEA, Institut Mines-Telecom, HEC, ENSAE, Univ. Paris-Sud, Univ. Versailles St-Quentin et Inra
- **Sciences des données et ses interfaces:**
  - **Interdisciplinaires** : ***couplage fort*** entre mathématiques et informatique, économie de la donnée, juridiction de l'immatériel, sociologie quantitative, etc
  - **Applicatives** : énergie, mobilité, finance, santé et bien-être, etc
- **Couplage fort** entre Recherche – ***Formation*** (créer de nouveaux parcours interdisciplinaires qui manquent cruellement) – Innovation
- **P**rogramme d'**A**ffiliation **I**ndustrielle (PAI) – 10 soutiens au dépôt

Responsible/Ethical Data Management and Analytics

- La transparence des algorithmes est un facteur essentiel de la confiance numérique
- Outils pour « l'empowerment » du citoyen
- Outils pour le régulateur pour l'application de la loi
- La Transparence un avantage compétitif?

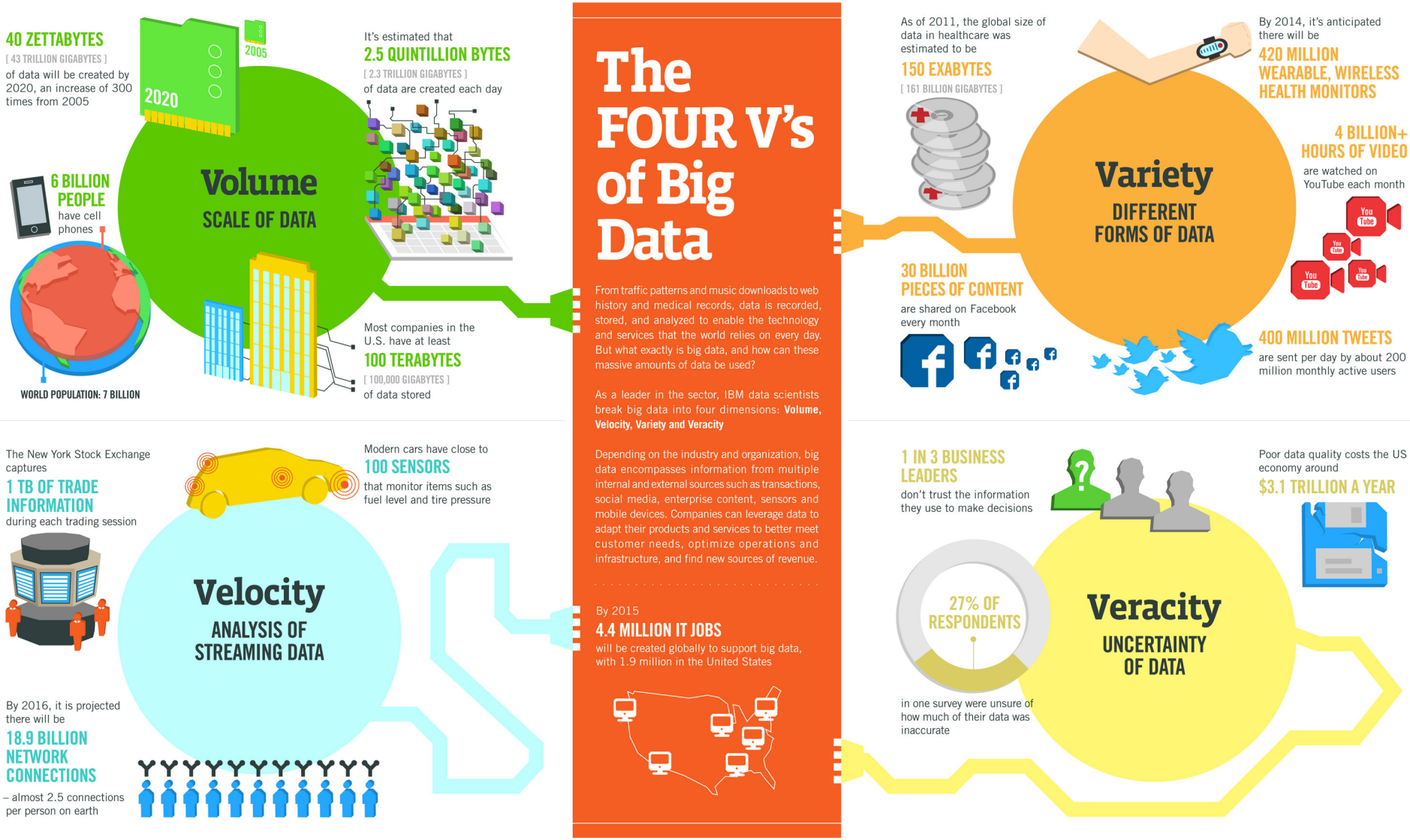
Conclusion

- Les Big Data représentent une opportunité économique et technologique permettant à l'humain de mieux maitriser son environnement (personnalisation des services, optimisation des ressources etc)
- A condition d'en maitriser l'écosystème: fournir les outils et les méthodes pour la transparence des données et des algorithmes à fin de mesurer la conformité et le respect des droits civiques
- Le volet scientifique et technologique reste à développer

Thank you for your attention

Nozha.Boujemaa@inria.fr

1- Introduction



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS