Towards Trusted AI
Impact on Language Technologies

Nozha Boujemaa
Director at DATAIA Institute
Research Director at Inria
Member of The BoD of BDVA
nozha.boujemaa@inria.fr

November 2018

Data & Algorithms

« 2 sides of the same coin »

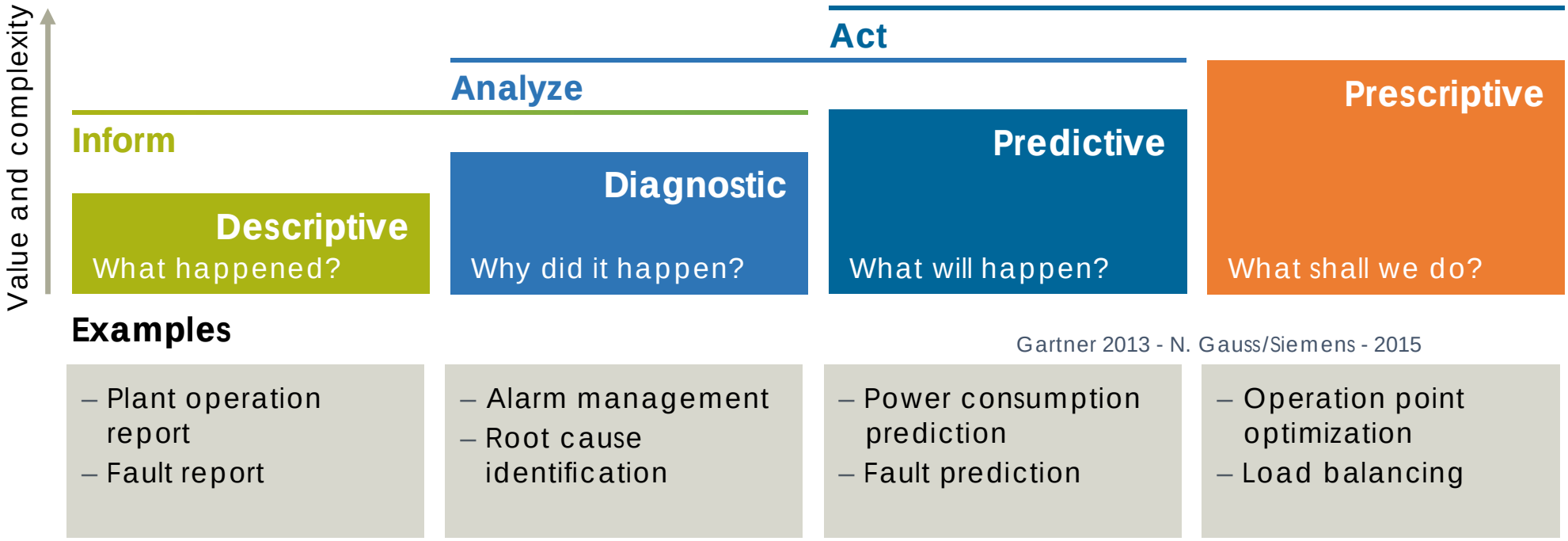
- Data are everywhere in personal and professional environment
- Algorithms making sense from these data are pervasive in more and more digital services.
- Algorithmic-based decisions are embedded from the processing of personal data to sensitive data in critical industrial systems : autonomous cars, conversational agents, health-care and well-being, public services etc.
- **Big Data** Technologies, agnostic to applications, are enablers for **AI capabilities** in real-life services

Data & Algorithms

« 2 sides of the same coin »

- Rising benefits from Big Data and AI technologies have wide impact on our economy and social organization ;
- Transparency and trust of such Algorithmic Systems (data & algorithms) becoming competitiveness factors for Data-driven economy ;
- Data analytics is changing from description of past to predictive and prescriptive analytics for decision support ;
- Importance of remedying the **information asymmetry** between the producer of the digital service and its consumer, be it citizen or professional – ***B2C or B2B*** => **civil rights**, **competition**, **sovereignty**.

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



Big Data Technologies are **enablers** for **AI capabilities**

5 Pillars for Data Science*

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

- Semantic interoperability of heterogeneous sources and representations, Data quality, Content Validation, Data provenance,

2- **Data Processing Architecture** :

- Scalability, Decentralization (Cloud/Fog etc), Low-energy consumption

3- **Data Analytics, Machine learning** :

- Machine Learning, Semantic Analysis (including NLP&U), Predictive/Presciptive Analytics

4- **Data Protection**:

- Privacy-enhancing models and techniques, Robusteness against reversibility

5- **Data Visualization:**

- Interactive visual analytics, Collaborative, Cross-platform data frameworks

* Inspired by BDVA SRIA technical priorities

Algorithmic systems in every day life

- Some dominant platforms on the market play a role of "prescriber" by directing a large share of user traffic:
  - Ranking mechanisms (search engine),
  - Recommendation mechanisms and content selection

  Product or service recommendation: is it most appropriate for the consumer (personalization) or the most appropriate to the seller (given the stock)?

- Opacity of the use made of the personal data and how they are processed,
  - What about the **consent**? Is it always respected?
  - Credit scoring, how fair is it?
  - Predictive justice?

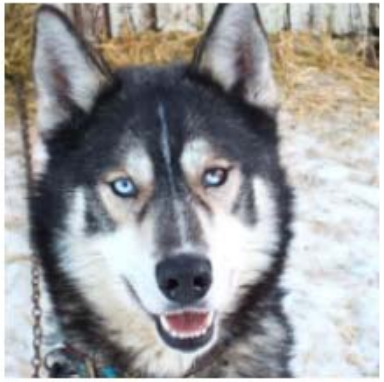
⇒ **New discrimination between those who know how algorithms work ad who do not**

In addition to economical and geostrategic effects on persons and societies

Transparent and Accountable Data Management and Analytics

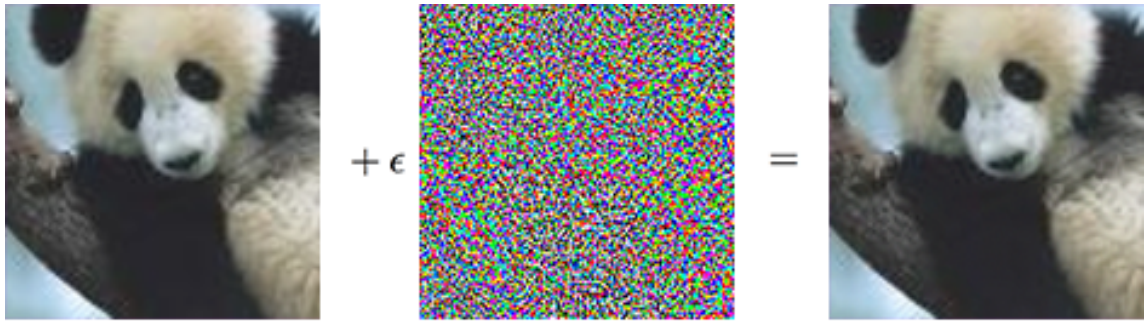
- Decision explanation and tractability: Trust and Transparency of computer-aided decision-making process **(decision responsibility)**: what are the different criteria/data/settings that have led to the specific decision in order to understand the global path for the reasoning?
- "How Can I trust Machine Learning prediction?" it happens to build the model of the object context rather the object itself
- Robustness to bias/diversion/corruption
- Careful software reuse

Safe AI: Robustness and Explanation



(a) Husky classified as wolf

(b) Explanation



"panda"
57.7% confidence

"gibbon"
99.3% confidence

Explanation:

Ribeiro et al. 2016, LIME: Why should I trust you? Explaining the predictions of any classifier

Robustness:

Goodfellow, Shlens and Szegedy 2015, "Explaining and Harnessing Adversarial Examples"

Challenges

- It is a mistake to assume they are objective simply because they are data-driven
- Implementing the *"Transparent-by-design":* fairness/equity, loyalty, neutrality, etc.
- **Mastering** the accuracy and robustness of Big Data & AI techniques: bias, diversion/corruption, reproducibility, source of unintentional discrimination

Algorithmic Systems Bias

Mastering Big Data Technologies: Bias problems could impact data technologies **accuracy** and people's lives

Challenges 1: **Data** Inputs to an Algorithm

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

Challenges 2: The Design of **Algorithmic** Systems and Machine Learning

- *Poorly designed matching systems*
- *Unintentional perpetuation and promotion of historical biases*
- *Decision-making systems that assume correlation implies causation*

Challenges / Efforts

- Algorithms are **encapsulated opinions** through decision parameters and learning data
- Mastering the accuracy and robustness of Big Data & AI techniques: bias, reproducibility, source of **unintentional discrimination**
- Implementing the *"Transparent-by-design":* fairness/equity, loyalty, neutrality, etc.
- **Interdisciplinary co-conception** of solutions, How responsible is a ML algorithm?
- **Interdisciplinary training** for Data Scientists: law, sociology and economy, **Careful software reuse** => mastering information leaks (SRE)

Challenges / Efforts

- Complex concepts, Dependent on cultural context, law context, etc.
  Transparency, Asymmetry, Accountability, Loyalty, Fairness, Equity, Intelligibility, Explainability, Traceability, Auditability, Proof and Certification, Performance, Ethics, Responsibility

  ➡ Ethical ≠ Responsible, *Transparent* ≠ Make available the source code

  ➡ International collaboration is key (AI HLG- EC, OECD, UNESCO etc)

- Pedagogy and explanation, awareness rising, uses-cases, (all public! Including scientists)

Challenges / Efforts

- **Trusted AI:**
  - **Responsible: Compliance with Policy and with Social Values/Ethics (democracy, human dignity etc),**
  - **Robust and safe: against bias, corruption, noise, reproducibility etc**
- **Auditability and Transparent-by-Design (Values-by-Design)** tools and algorithms for **socio-economic empowerment**
- **AI is part of the solution and not only the law! Algorithmic tools to monitor the behavior of AI technologies (traceability, explanability, intelligibility etc)**
- **Governance of Data is key, ML algorithms are shared in open-source but NOT Data**

➡ **Transparency Tools vs GDPR vs Having the Choice**

➡ **Cloud Act (Clarifying Lawful Overseas Use of Data Act)**

Challenges in Language Industry

Technical issues

- Diversity and representativeness of learning data (context, minorities and multi-layered cultural nuances)
- Reproducibility and robustness of learning algorithms
- Traceability

Application issues

- Chatbots and Nudging (L. Devillers &all - DATAIA)
- Chatbots/emotions and related business models => Need for Ethical Guidelines
- etc

International Efforts – **AI HLEG EC**

Artificial Intelligence - High Level Expert Group of the European Commission

- 52 independent experts multiple expertise (computer science, law, ethicist, philosopher, entrepreneur,) and background: industry, academia, consumer associations

General objective : Support the implementation of the **European strategy on AI**.

- Elaboration of recommendations on future AI-related **policy** development and on **ethical**, **legal** and **societal issues** related to AI, including socio-economic challenges
- Elaboration of recommendations on **AI Policy & Investment Recommendations**
- Serve as the **steering group for the European AI Alliance's work**
- AI HLEG **Chair**: Pekka Ala-Pietilä, **2 Vice-Chairs**: Barry O'Sullivan & Nozha Boujemaa

Trusted AI

Intent

Regulated Ex Ante:
Roman Law Style

Ensure ethical intent when developing/using AI, in line with:

Core Values & Principles

Implementation

Regulated Ex post:
Common Law Style

Ensure proper implementation of values & principles when developing/using AI

Responsible/Compliant & Robust AI

Mechanisms

Before: Design

During: Auditability

After: Traceability

Comprehensive Check List/Guidelines based on Use Cases

Red Lines

International Efforts – **AIGO**

Artificial Intelligence Expert Group at the OECD

- 36 members: OECD governments representatives + Experts (MIT, Harvard, Inria, IEEE, Civil Society)
- Report AI for Society November 2019:
  - build a shared understanding of AI
  - map economic / social impacts of AI applications.
  - discuss policies that influence adoption of AI and policies to address its consequences.
  - help coordination and consistency with discussions in other international fora and among OECD policy

Need for Interdisciplinary & International efforts

THANK YOU

nozha.boujemaa@inria.fr