

# Human Values-Centric AI, Technological Challenges

**Nozha Boujemaa**

Director at DATAIA Institute  
Research Director at Inria

[nozha.boujemaa@inria.fr](mailto:nozha.boujemaa@inria.fr)

*inria*  
informatics mathematics

July 2018

université  
PARIS-SACLAY



# 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 or public services etc.
- **Big Data** Technologies, **agnostic** to applications, are **enablers** for **AI capabilities** in real-life services



# Data & Algorithms

---

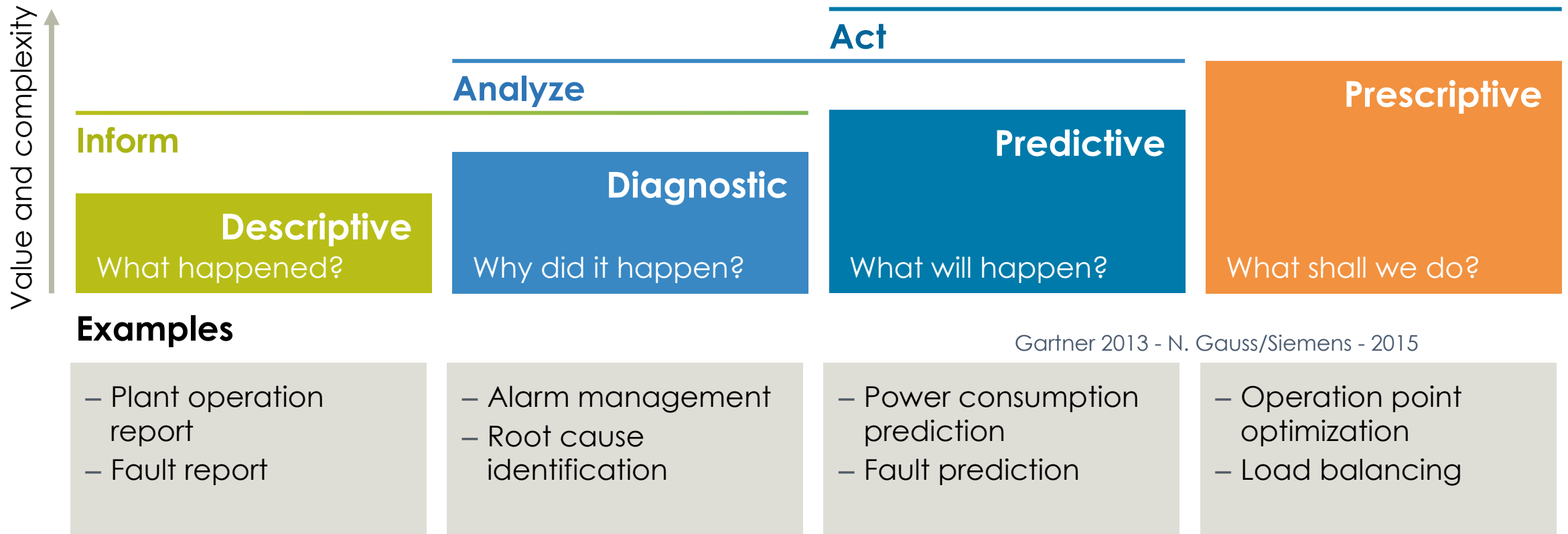
« 2 sides of the same coin »



- 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**, **Predictive/Prescriptive Analytics**
- 4- **Data Protection:**
  - **Privacy-enhancing models and techniques**, Robustness against reversibility
- 5- **Data Visualization:**
  - **Interactive** visual analytics, **Collaborative**, **Cross-platform data frameworks**

\* Inspired by BDVA SRIA technical priorities

# *Towards responsible and Safe AI*



# 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:

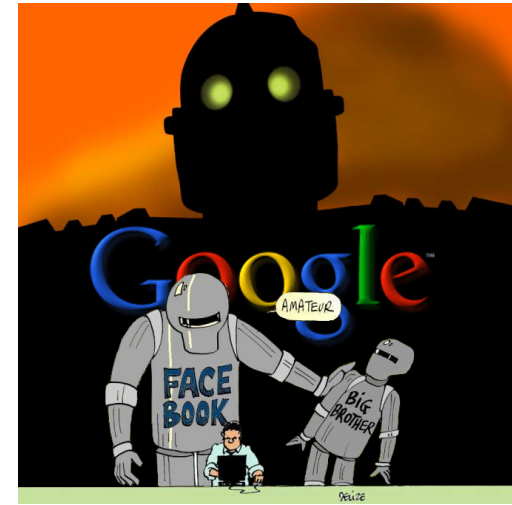
- **Ranking** mechanisms (search engine),
- **Recommendation** mechanisms and content selection

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 and who do not

**In addition to economical and geostrategic effects on persons and societies**







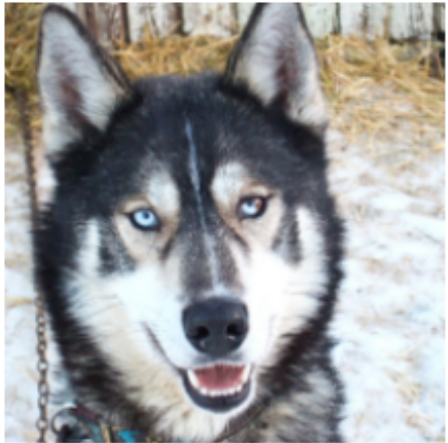
# Transparent and Accountable Data Management and Analytics

- **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

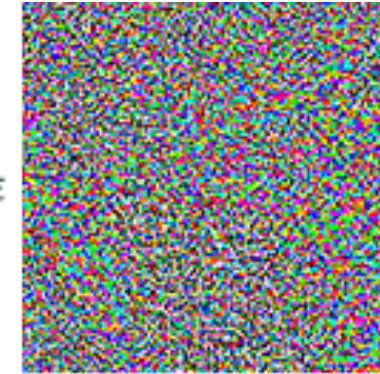


(b) Explanation



"panda"  
57.7% confidence

+  $\epsilon$



=



"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  $\neq$  Responsible, *Transparent*  $\neq$  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

---

- **Auditability and Transparent-by-Design** tools and algorithms for **socio-economic empowerment**
- **AI is part of the solution and not only the law!**
- **Governance of Data is key, ML algorithms are shared in open-source but NOT Data (governance of AS!)**
  - ➔ **Transparency Tools vs GDPR vs Having the Choice**
  - ➔ **Cloud Act (Clarifying Lawful Overseas Use of Data Act)**



# DATAIA Institute

## 4 Overarching Challenges:

1. From Data to Knowledge, from Data to Decision,
  2. From Machine Learning to Artificial Intelligence,
  3. Transparency, responsible AI & Ethics,
  4. Data protection, regulation and economy.
- **Scientific and disciplinary foundations:** Data Science, Management and Economy, Social Sciences, Legal Sciences
  - **Industry Affiliation Program, International Collaboration Program**
  - **Application domains:** Internet of people and things, Urbanization 4.0, Optimal Energy Management, Business Analytics
    - *Roadmap for 8 years, 150 M€ Budget, with 14 academic founding institutions*
    - **Kick-off** => 15 February 2018







# DATAIA Ecosystem



1. From algorithms to proofs of concepts
2. Joining the human sciences and the digital revolution
3. Designing a humanly sustainable data science

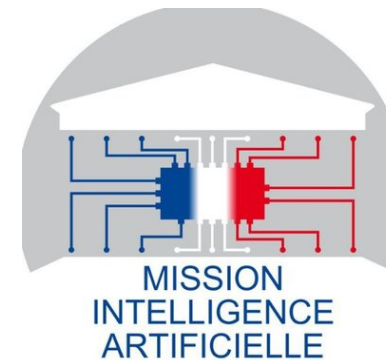
Disciplinary foundations + Interdisciplinary Challenges + Application Domains



# REMISE DU Rapport VILLANI: Notre écosystème EN AVANCE DE PHASE SUR LES **I3A**



« Nous devons prendre nos responsabilités pour former et garder nos compétences en IA. Pour cela, il faut faire émerger 4 à 5 pôles en France, dont certains existent déjà comme l'institut **Data IA** et **DigiHall**. »



Réseaux d'Instituts Interdisciplinaires de l'IA (**3IA**)



# Overarching challenges

---

## 1- Machine Learning Toward AI

- Disruptive machine learning and AI: common sense, adaptability, generalization, unsupervised, weakly supervised
- Deep learning and adversarial learning
- Reinforcement learning and online Learning
- Reproducibility and robust learning
- Machine learning and hyper-optimization, combinatorial optimization
- Statistical inference and validation, causality
- Compositionality of deep architectures



# Overarching challenges

---

## 2- From Data to Knowledge, from Data to Decision

- Fast big data: structuring the data in order to exploit it, heterogeneous, complex, incomplete, semi-structured and / or uncertain data
  - Improved storage, low energy consumption calculation
- Content Analytics and Understanding: NLP&U, speech, image & video
- Human-machine coevolution in autonomous systems: conversational agents, autonomous vehicles, social robots
  - Modeling interactions between agents (human or artificial) by game theory
- Multi-scale and multimodal representation and algorithms
- Theoretical analysis of heuristic methods (complexity theory, information geometry, Markov chains theory)



# Overarching challenges

---

## 3- Transparency, Responsible AI and Ethics

- Audit of algorithmic systems: non-discrimination, loyalty, technical bias, neutrality, equity
- Responsibility by-design, Transparency-by-design, Explicability-by-design, Equity-by-design
- "Progressive user-centric-analytics" (interactive monitoring of decision systems)
- Responsibility for information processing and decision making: data use control and fact-checking
- Causal discovery, traceability of inferences from source data, interpretability of deep architectures



# Overarching challenges

---

## 4- Protection, regulation and data economy

- « Privacy-by-design », GDPR,
- Privacy-friendly learning ("differential privacy")
- Development of ethically responsible methodologies and technologies to regulate the collection, use and process of personal data, and the exploitation of the knowledge derived from this data
- Computer security of data processing chains
- Security/crypto: block-chain and trusted third parties



# Projets en cours de lancement

- **PEPER** : Prédiction de la Prosommation d'énergie renouvelable par apprentissage, *Hossam AFIFI, Télécom SudParis; Jordi BADOSA, Ecole polytechnique; Florence OSSART, CentraleSupélec GEEPS*
- **VADORE** : Valorisation des Données pour la Recherche d'Emploi, *Bruno CREPON, ENSAE; Michele SEBAG, CNRS; Marco CUTURI, ENSAE; Christophe GALLAC, ENSAE; Philippe CAILLOU, LRI*
- **Bad Nudge - Bad Robot ?** : Nudge et Éthique dans l'interaction verbale homme-machine, *Laurence DEVILLERS, LIMSI-CNRS ; Serge PAJAK, Université Paris-Sud*
- **RGPD et Cloud Personnel** : de l'Empowerment à la REsponsabilité (GDP-ERE), *Célia ZOLYNSKI, UVSQ ; Mélanie CLEMENT-FONTAINE, UVSQ ; Nicolas ANCIAUX, INRIA ; Philippe PUCHERAL, USVQ/INRIA ; Guillaume SCERRI, UVSQ/Inria*
- **MissingBigData** : missing data in the big data era  
*Julie JOSSE, CMAP/CNRS ; Gaël VAROQUAUX, INRIA*







## Projets en cours de lancement

---

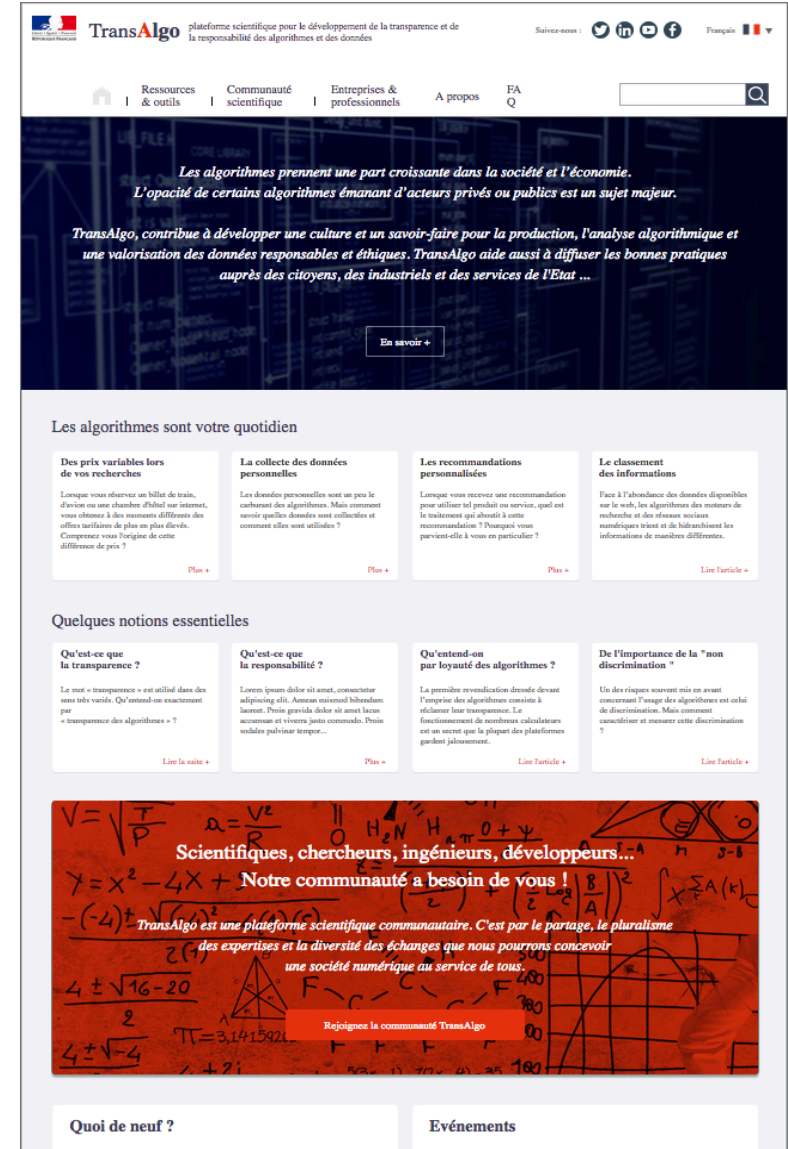
- **Smart Lawyer**: Rating Legal Services in the Courtroom  
*David RESTREPO AMARILES, HEC Paris ; Michalis VAZIRGIANNIS, École polytechnique*
- **HistorIA** : Grandes bases de données historiques. Fouille de données, exploration et explicabilité  
*Christophe PRIEUR, Telecom ParisTech ; Jean-Daniel FEKETE, Inria*
- **StreamOps** : Plateforme Open Source pour la Recherche et l'Intégration d'Algorithmes pour l'Analyse de Flux de Séries Temporelles Massives  
*Cédric GOUY-PAILLER, CEA; Karine ZEITOUNI, Université de Versailles-St-Quentin-en-Yvelines*



Une école de l'IMT



- National Scientific Platform for Transparency & Accountability Tools and Methods for Data and Algorithms (Fairness, Neutrality, Loyalty); B2B & B2C.
- Support of The new “*Law for Digital Republic*”: the right to the explainability of algorithmic decision of public services (APB service stopped!)
- Contributors: Regulation authorities (DGCCRF, CSA, ARCEP, CNIL), besides academia, industries and associations,



TransAlgo plateforme scientifique pour le développement de la transparence et de la responsabilité des algorithmes et des données

Les algorithmes prennent une part croissante dans la société et l'économie. L'opacité de certains algorithmes émanant d'acteurs privés ou publics est un sujet majeur.

TransAlgo, contribue à développer une culture et un savoir-faire pour la production, l'analyse algorithmique et une valorisation des données responsables et éthiques. TransAlgo aide aussi à diffuser les bonnes pratiques auprès des citoyens, des industriels et des services de l'Etat ...

En savoir +

Les algorithmes sont votre quotidien

- Des prix variables lors de vos recherches**  
Lorsque vous observez un billet de train, évaluez ou une chambre d'hôtel sur internet, vous obtenez à des moments différents des offres tarifaires de plus en plus élevées. Comprenez vous l'origine de cette différence de prix ? [Plus +](#)
- La collecte des données personnelles**  
Les données personnelles sont un peu le carburant des algorithmes. Mais comment savoir quelles données sont collectées et comment elles sont utilisées ? [Plus +](#)
- Les recommandations personnalisées**  
Lorsque vous recevez une recommandation pour utiliser tel produit ou service, quel est le traitement qui abouti à cette recommandation ? Pourquoi vous parvenez-elle à vous en particulier ? [Plus +](#)
- Le classement des informations**  
Face à l'abondance des données disponibles sur le web, les algorithmes des moteurs de recherche et des réseaux sociaux trient et hiérarchisent les informations de manières différentes. [Lire l'article +](#)

Quelques notions essentielles

- Qu'est-ce que la transparence ?**  
Le mot « transparence » est utilisé dans des sens très variés. Qu'entend-on exactement par « transparence des algorithmes » ? [Lire la suite +](#)
- Qu'est-ce que la responsabilité ?**  
Lorsqu'un algorithme agit, comment savoir si on peut lui reprocher un comportement inapproprié ? [Plus +](#)
- Qu'entend-on par loyauté des algorithmes ?**  
La première recommandation demandée devant l'opacité des algorithmes consiste à réclamer leur transparence. Le fonctionnement de nombreux calculateurs est un secret que le pluspart des plateformes gardent jalousement. [Lire l'article +](#)
- De l'importance de la "non discrimination"**  
Un des risques souvent mis en avant concernant l'usage des algorithmes est celui de discrimination. Mais comment caractériser et mesurer cette discrimination ? [Lire l'article +](#)

Scientifiques, chercheurs, ingénieurs, développeurs...  
Notre communauté a besoin de vous !  
TransAlgo est une plateforme scientifique communautaire. C'est par le partage, le pluralisme des expertises et la diversité des échanges que nous pourrions concevoir une société numérique au service de tous.  
Rejoignez la communauté TransAlgo

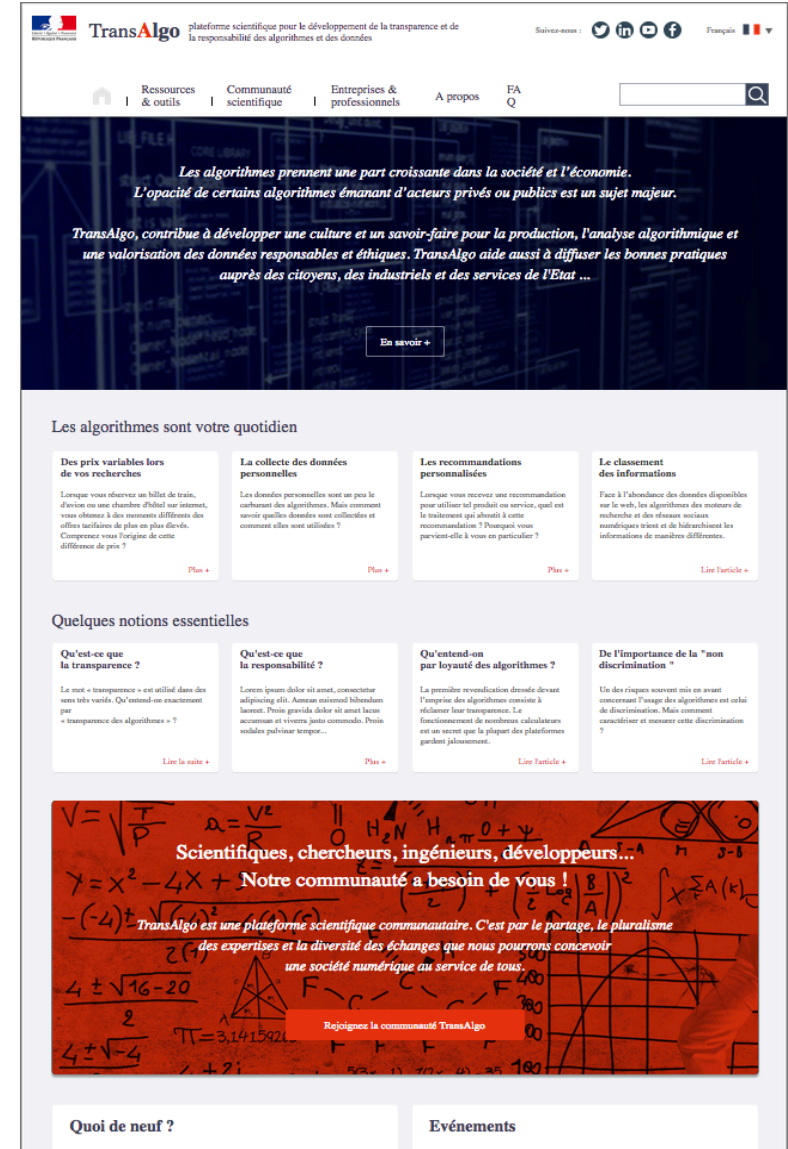
Quoi de neuf ? Événements

## Objectives:



- Resource center, Empowerment tools: reports, publications, software, controlled data sets & testing protocols ;
- Awareness rising: workshops & Moocs ;
- Best practices recommendation & sharing ;
- Research & Dev. Programs.

## Working Groups :

- Auditability of Recommendation and Ranking systems ;
- Explainability, Reproducibility and Bias of ML ;
- Privacy, Data Usage Control & Information-flow-monitoring ;



TransAlgo plateforme scientifique pour le développement de la transparence et de la responsabilité des algorithmes et des données

Suivez-nous :  Français 

Ressources & outils | Communauté scientifique | Entreprises & professionnels | A propos | FAQ

*Les algorithmes prennent une part croissante dans la société et l'économie. L'opacité de certains algorithmes émanant d'acteurs privés ou publics est un sujet majeur.*

*TransAlgo, contribue à développer une culture et un savoir-faire pour la production, l'analyse algorithmique et une valorisation des données responsables et éthiques. TransAlgo aide aussi à diffuser les bonnes pratiques auprès des citoyens, des industriels et des services de l'Etat ...*

[En savoir +](#)

Les algorithmes sont votre quotidien

- Des prix variables lors de vos recherches**  
Lorsque vous réservez un billet de train, évaluez ou une chambre d'hôtel sur internet, vous obtenez à des moments différents des offres tarifaires de plus en plus élevées. Comprenez vous l'origine de cette différence de prix ? [Plus +](#)
- La collecte des données personnelles**  
Les données personnelles sont un peu le carburant des algorithmes. Mais comment savoir quelles données sont collectées et comment elles sont utilisées ? [Plus +](#)
- Les recommandations personnalisées**  
Lorsque vous recevez une recommandation pour utiliser tel produit ou service, quel est le traitement qui aboutit à cette recommandation ? Pourquoi vous parvenez-elle à vous en rendre compte ? [Plus +](#)
- Le classement des informations**  
Face à l'abondance des données disponibles sur le web, les algorithmes des moteurs de recherche et des réseaux sociaux manipulent et trient et hiérarchisent les informations de manières différentes. [Lire l'article +](#)

Quelques notions essentielles

- Qu'est-ce que la transparence ?**  
Le mot « transparence » est utilisé dans des sens très variés. Qu'entend-on exactement par « transparence des algorithmes » ? [Lire la suite +](#)
- Qu'est-ce que la responsabilité ?**  
Lorsqu'un produit ou un service, conçu par un algorithme, a des conséquences négatives sur la vie des personnes, comment pouvons-nous nous en rendre compte ? [Plus +](#)
- Qu'entend-on par loyauté des algorithmes ?**  
La première recommandation proposée devant l'écran des algorithmes consiste à refuser leur transparence. Le fonctionnement de nombreux calculateurs est un secret que le plus grand des plateformes gardent jalousement. [Lire l'article +](#)
- De l'importance de la "non discrimination"**  
Un des risques souvent mis en avant concernant l'usage des algorithmes est celui de discrimination. Mais comment caractériser et mesurer cette discrimination ? [Lire l'article +](#)

**Scientifiques, chercheurs, ingénieurs, développeurs...**  
Notre communauté a besoin de vous !  
TransAlgo est une plateforme scientifique communautaire. C'est par le partage, le pluralisme des expertises et la diversité des échanges que nous pourrions concevoir une société numérique au service de tous.  
[Rejoignez la communauté TransAlgo](#)

Quoi de neuf ? | Événements



# Summer School

- **DATAIA Institute co-organizes the DS3 Summer School with École polytechnique :**

- Speakers confirmed: Cédric Villani, Yann Le Cun, Adrian Weller, Krishna Gummadi, Mireille Hildebrandt, Jean-Philippe Vert ...

## Fairness in ML, Interpretable ML, Law and ML Design

- Attendees: 760 applications for 400 (students, academics and professionals), 35 countries, 39 speakers

<http://www.ds3-datascience-polytechnique.fr/>



DATA SCIENCE SUMMER SCHOOL

**JUNE  
25-29  
2018**

**OPENING** by **Cédric VILLANI**

**TUTORIALS ON**

Deep Learning

**Yann LECUN** [Facebook - New York University]

Interpretable Machine Learning

**Adrian WELLER** [University of Cambridge - Alan Turing Institute]

Fairness in Machine Learning

**Krishna GUMMADI** [Max Planck Institute]

Probabilistic Numerical Methods

**Mark GIROLAMI** [Imperial College London]

Online Learning Algorithms

**Nicolò CESA-BIANCHI** [University of Milano]

Non-convex Optimization

**Suvrit SRA** [MIT]

... other speakers will be confirmed soon

**AT CAMPUS  
POLYTECHNIQUE**

**PARALLEL SESSIONS**

on Health and Social Sciences

**PRACTICAL SESSIONS**

on Deep Learning, Reinforcement

Learning, Recommender Systems, Precision Medicine...

**POSTER SESSION**

**ROUND TABLE DISCUSSION**

Targeted for students, academics and professionals

More information to come on:

[www.ds3-datascience-polytechnique.fr](http://www.ds3-datascience-polytechnique.fr)





# France-Japan Symposium

- The DATAIA Institute co-organize with JST, in partnership with the Ministry of Higher Education, Research and Innovation (MESRI) and the French Embassy in Japan: **"Data Science and AI, Core Technologies and Applications For a New Society"**
- Dates: from **11 -12 July**, 2018
  - Location: Paris - Amphitheatre of MESRI
  - With PIs **Pr. Kitsuregawa**, **Pr. Tanaka** and **Pr. Hagita** research program directors from 3 CREST/JST Programs (equivalent to ERC Advanced G.) supervisor <http://dataia.eu/en/news/dataia-jst-international-symposium-data-science-and-ai-0>



MINISTÈRE  
DE L'ENSEIGNEMENT  
SUPÉRIEUR,  
DE LA RECHERCHE  
ET DE L'INNOVATION



AMBASSADE DE FRANCE  
AU JAPON



# Need for Interdisciplinary efforts

THANK YOU

[nozha.boujemaa@inria.fr](mailto:nozha.boujemaa@inria.fr)