Loyauté des Décisions Algorithmiques

Philippe Besse

Université de Toulouse – INSA
Institut de Mathématiques – UMR CNRS 5219
CIMI – Projet AOC

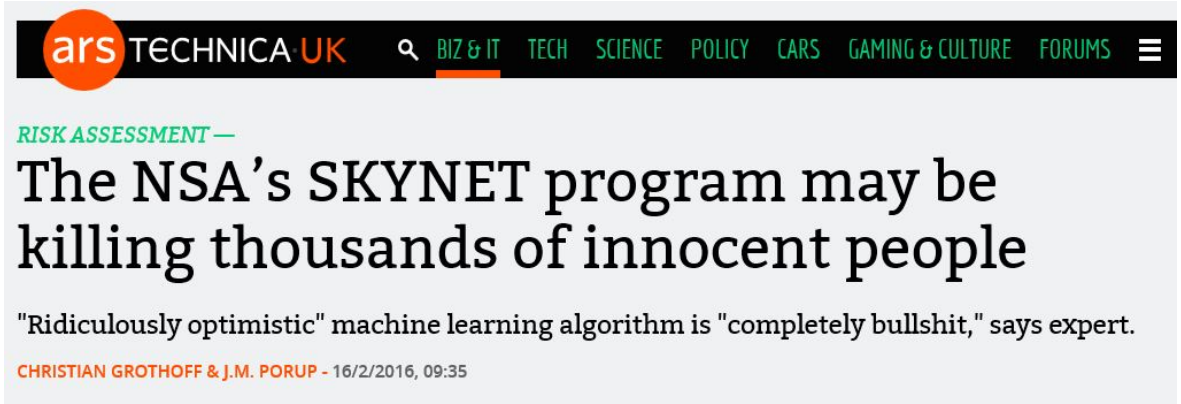
Éthique et Technologies

- *La technologie elle elle neutre ?* : qui y a l'accès ?
- Pour quel usage ?
  Pharmakon digital : Serres et al. (2014), Stiegler et Kyrou, (2015)


M PIXELS CHRONIQUES DES (R)ÉVOLUTIONS NUMÉRIQUES

Après IBM et Google, Microsoft s'intéresse de près au cancer

LE MONDE | 21.09.2016 à 15h47


ars TECHNICA UK BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE FORUMS

RISK ASSESSMENT —

The NSA's SKYNET program may be killing thousands of innocent people

"Ridiculously optimistic" machine learning algorithm is "completely bullshit," says expert.

CHRISTIAN GROTHOFF & J.M. PORUP - 16/2/2016, 09:35

Éthique, Algorithmes et Grosses Data

- Contributions : *Gouvernementalité algorithmique* (Rouvroy et Berns 2013),
  Mort de la politique (Morozov, 2014),
  - Intermédiation technologique (Uber, Blablacar, Airbnb...)
  - Application ”automatique” des lois
- *Open data*, anonymisation, fin du consentement libre et éclairé
  - Projet `care.data` NHS et Royaume Uni : (*Social License*)
  - Projet *Data Science Initiative* : $\mathcal{X}$ et CNAM, base Sniiram
- Entraves à la concurrence, *algorithmic pricing*, comparateurs
  Virtual Competition (Ezrachi et Stucke ; 2016)

Notions importantes

- *Trustworthiness* : loyauté, fiabilité, crédibilité, pour mériter la confiance
- *Accountability* : responsabilité, capacité à rendre compte

Éthique, Déontologie scientifique et Statistique

- Statistique : usages, abus, fraudes et déluge de données
  - KO Planification expérimentale et Test d'hypothèse
  - MO Données préalables, fouille et *data snooping*
  - GO Données omiques avec $p >> n$: indétermination
  - TO Grosses data : n très grand, Prévision au lieu de Tests (tous significatifs)
- Problème de Reproductibilité des résultats scientifiques
- Bannissement de la p-valeur (Trafimow et Marks ; 2015)
- Apprentissage Machine et Science des Données (2008)
  - D. Patil (LinkedIn) et J. Hammerbacher (Facebook) : *Analyste, ça fait trop Wall Street ; statisticien, ça agace les économistes ; chercheur scientifique, ça fait trop académique. Pourquoi pas "data scientist" ?*
  - J. Wills (Cloudera) :*Data scientist (n) : Person who is better at statistics than any software engineer and better at software than any statistician*

Éthique, Épistémologie et Science des Données

- Fin de la théorie et obsolescence de la démarche scientifique (Anderson ; 2008)
- Hypothèse déduite des données pas d'une théorie
- Validation de la recherche : cas de la base Sniiram
  1. Test d'une hypothèse a priori : effets indésirables et risque d'un médicament
  2. Fouille systématique : corrélation, co-occurrence, motifs
     Vérification : retour à (1). Sinon : risque d'artefact (data snooping)
- Évolution méthodologique mais pas épistémologique

Open data & open Science

- Risque de *fake* (pseudo) Sciences : climatosceptique (Allègre, 2010)
- Astronomes amateurs et traque des astéroïdes
- Accès aux données, aux codes et réfutabilité des hypothèses.

Cadre juridique

Loi n°78-17 du 6/01/1978 relative à l'informatique, aux fichiers et aux libertés

Article 10 Aucune décision produisant des effets juridiques à l'égard d'une personne ne peut être prise sur le seul fondement d'un traitement automatisé de données destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité.

Loi n° 2015-912 du 24/07/2015 relative au renseignement

Art. L. 851-3 C Il peut être imposé aux opérateurs ... la mise en œuvre sur leurs réseaux de traitements automatisés destinés ... à détecter des connexions susceptibles de révéler une menace terroriste.

Loi n°2016-1321 du 7/10/2016 pour une République Numérique

Article 6 Sous réserve des secrets protégés, les administrations ... publient en ligne les règles définissant les principaux traitements algorithmiques utilisés dans l'accomplissement de leurs missions lorsqu'ils fondent des décisions individuelles.

Article 50 Les opérateurs de plateformes en ligne dont l'activité dépasse un seuil de nombre de connexions défini par décret élaborent et diffusent aux consommateurs des bonnes pratiques visant à renforcer les obligations de clarté, de transparence et de loyauté.

Décret du 16/03/2017 Art. R. 311-3-1-2.

L'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :

1. Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;
2. Les données traitées et leurs sources ;
3. Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;
4. Les opérations effectuées par le traitement.

Règlement 2016/679/EU sur la protection des données personnelles

Article 22 Décision individuelle automatisée, y compris le profilage

1 La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative...

3 Le responsable du traitement met en œuvre des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée, au moins du droit de la personne concernée d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision.

4 Les décisions visées ne peuvent être fondées sur les catégories particulières de données à caractère personnel (cf. article 9 : biométriques, génétiques, de santé, ethniques ; orientation politique, syndicale, sexuelle, religieuse, philosophique).

Effectif en mai 2018

Décision algorithmique

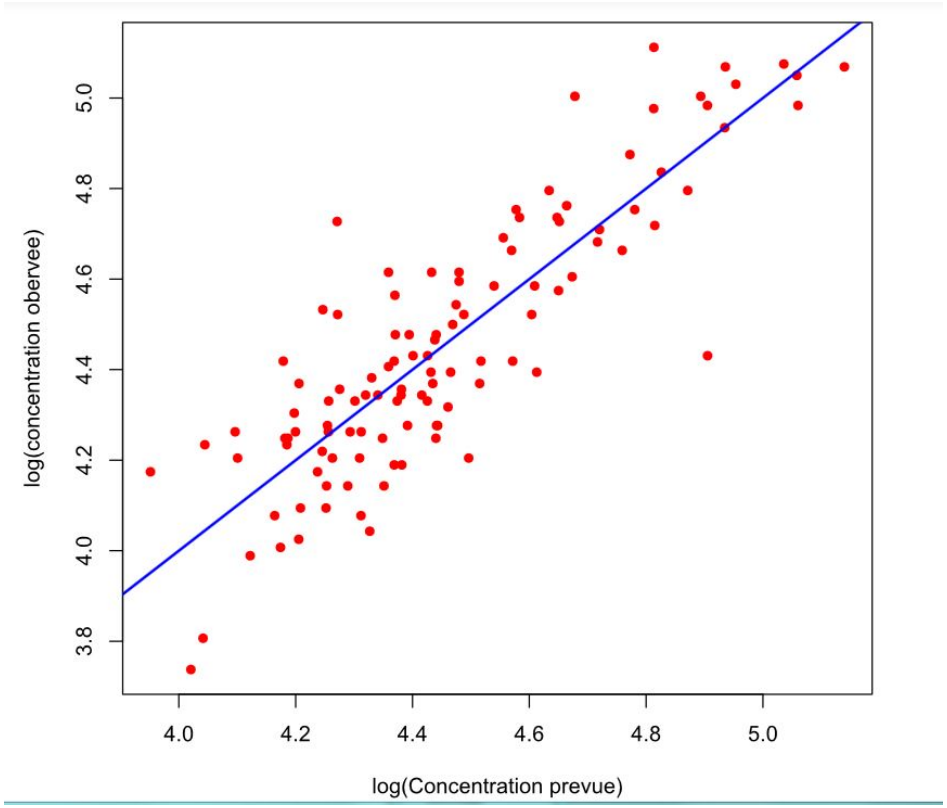
- Décision issue d'un traitement automatisée
- Algorithme procédural type APB (admission post-bac)
- Algorithme par Apprentissage Statistique ou Machine
- Choix de :
  Traitement, action commerciale, maintenance préventive, accord de crédit, mise sous surveillance, d'un produit...
- Prévision d'une probabilité ou risque de :
  Déclencher une maladie, départ d'un client, défaillance d'un système, défaut de paiement, radicalisation, d'appétence...
- Décision découle d'un Modèle ou Algorithme :
  - Estimé sur un *échantillon d'apprentissage*
  - Optimisé par *validation croisée*
  - Validé sur un *échantillon test* indépendant

Loyauté des Algorithmes

- *Accountability* et *Trustworthiness*
- Se traduisent et s'évaluent par leur :
  - Explicabilité et transparence
  - Qualité de prévision et justesse de décision
  - Biais et discrimination

Explicabilité : modèle linéaire du ”siècle dernier”

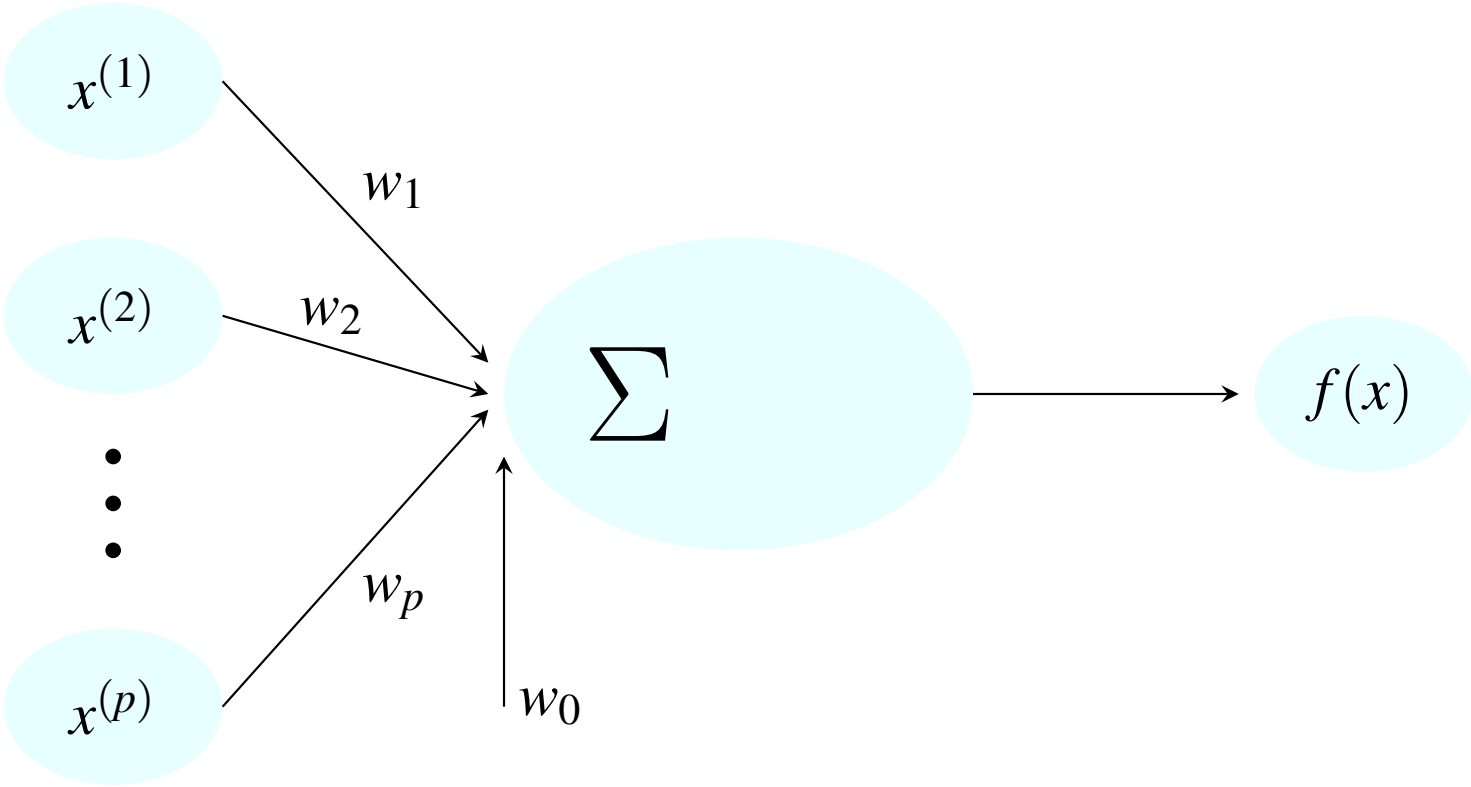
Prévoir la Concentration en Ozone



$$\begin{aligned} \log(\text{ConcODemain}) &= 2,4 + 0,35 \times \log(\text{ConcOJour}) + 0,05 \times \text{Sec} + \\ &+ 0,03 \times \text{T12} - 0,03 \times \text{Ne9} + 0.1 \times \text{Vx9} \end{aligned}$$

Modèle / Neurone Linéaire

Modéliser / prévoir une variable quantitative

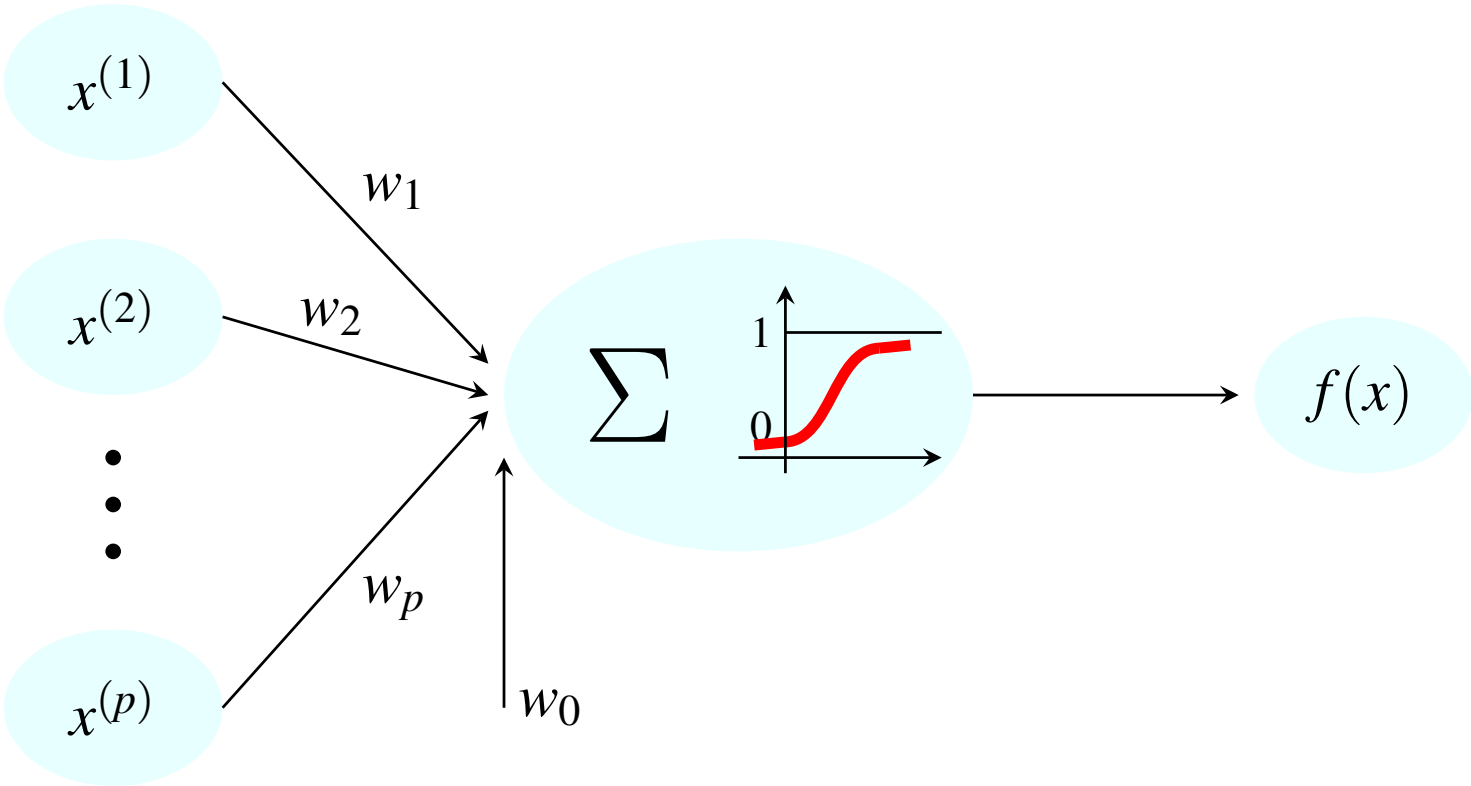


$$f(x) = w_0 + w_1 \times x^{(1)} + w_2 \times x^{(2)} + \cdots + w_p \times x^{(p)}$$

Compléments sur `wikistat.fr`

Modèle / **Neurone** *logistique*

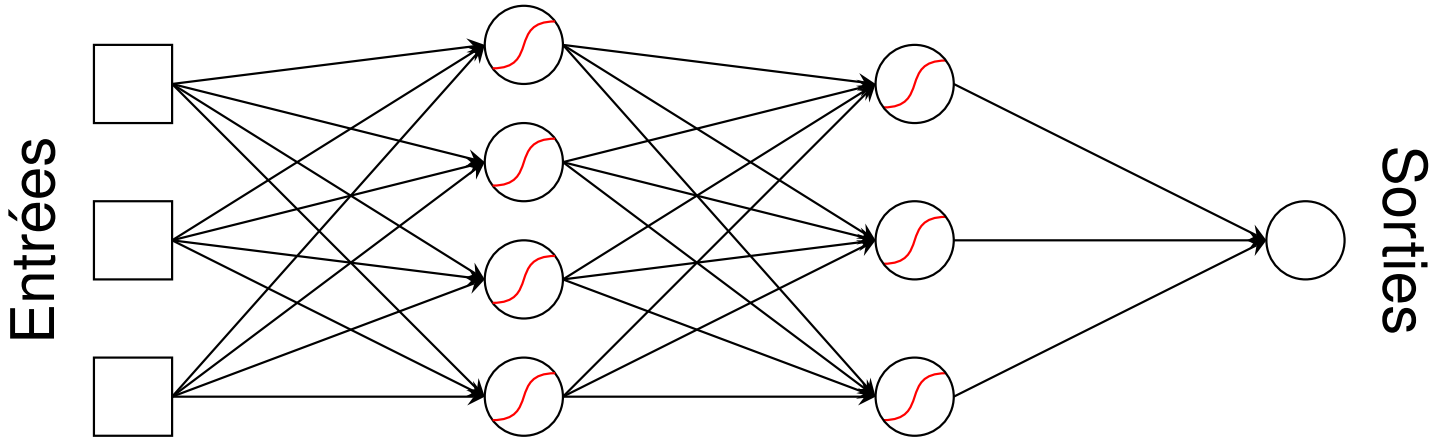
Variable binaire : Maladie, Panne, Départ, Faillite...



Exemple en épidémiologie : évaluer les facteurs de risque

Compléments sur `wikistat.fr`

Explicabilité : réseau de neurones : (Perceptron)



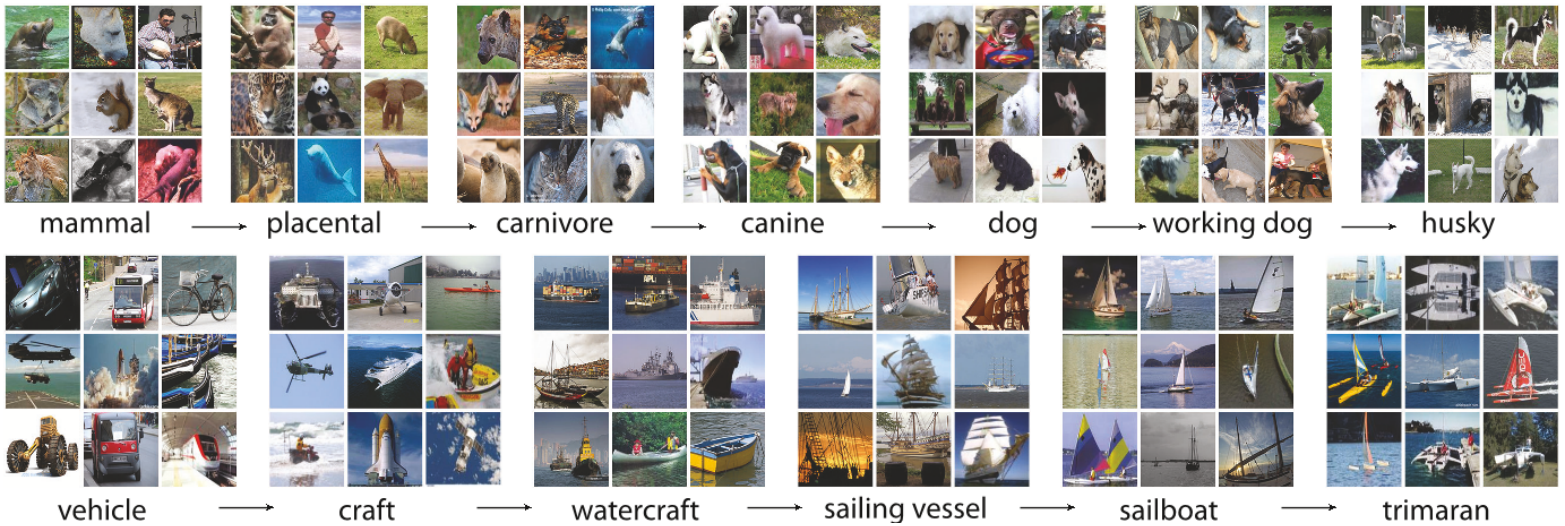
$x = (x^{(1)}, \ldots, x^{(p)})$ *Couche 1* *Couche 2* $y = F(x)$

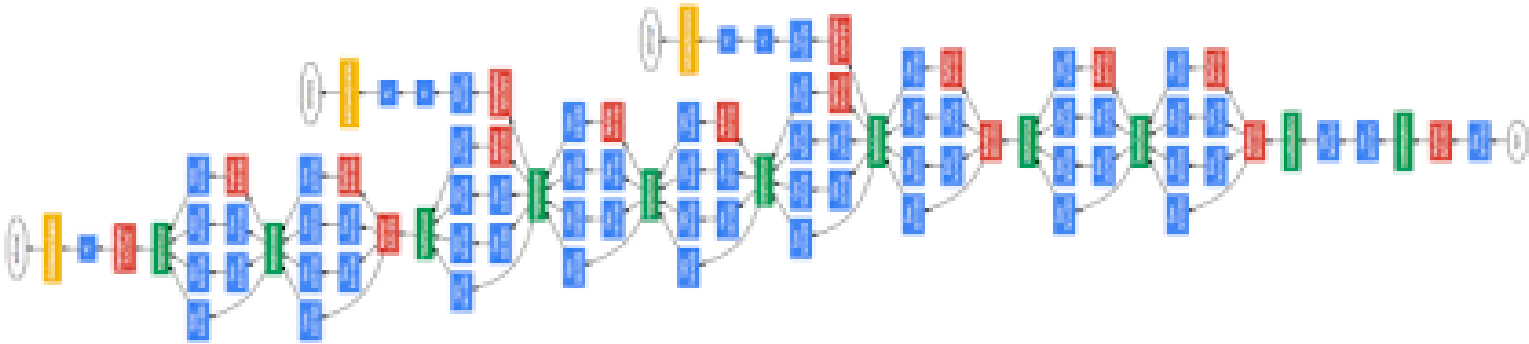
Problème : Boîte Noire — Explication impossible

Compléments sur `wikistat.fr`

Explicabilité : Deep Learning 1

Exemple : base de données ImageNet :
15 millions d'images, 22000 catégories
2016 : 152 couches et mieux que l'expert humain





Éthique ”industrielle”

Amazon, Google, Facebook, IBM, Microsoft, Apple...

1. We believe that artificial intelligence technologies hold great promise for raising the quality of people's lives and can be leveraged to help humanity address important global challenges such as climate change, food, inequality, health, and education.

...

7. We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.

Grosses données et qualité de prévision

- Plus de données entraine-t-il une meilleure prévision ?
- *L'efficacité prédictive sera d'autant plus grande qu'elle sera le fruit de l'agrégation de données massives*
  in *La Gouvernementalité Algorithmique* (Rouvroy et Berns, 2013)
- **Vrai** et **Faux**
- Rappel : deux sources d'erreur de prévision
  - Biais
  - "Variance"
- Ne pas confondre estimation / prévision d'une moyenne
  (*loi des grands nombres*)
  et celle d'un comportement individuel
- Taux d'erreur élevés en marketing (15 à 30%)

Fiabilité des algorithmes

Exemples : *Google Flu Trend* (2008-2015), Médiamétrie

FINAL FINAL

POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

Entre échantillonnage et big data, les nouveaux enjeux de la mesure d'audience

28 août 2013

Variance résiduelle

- Exemple jouet : Choix d'une couleur de cravate

  Monsieur Météo suit la prévision météo du lendemain
  Monsieur Aléa suit un tirage à pile ou face
  Facebook peut-il prévoir la couleur ?
- Ne pas confondre estimation / prévision d'une moyenne (*loi des grands nombres*)
  et celle d'un comportement individuel
- Taux d'erreur élevés en marketing (15 à 30%)
- Conclusion : Ne pas idéaliser les capacités prédictives d'un comportement

Qualité : Deep Learning 2

M PIXELS | CHRONIQUES DES (R)ÉVOLUTIONS NUMÉRIQUES

Derrière les dérapages racistes de l'intelligence artificielle de Microsoft, une opération organisée

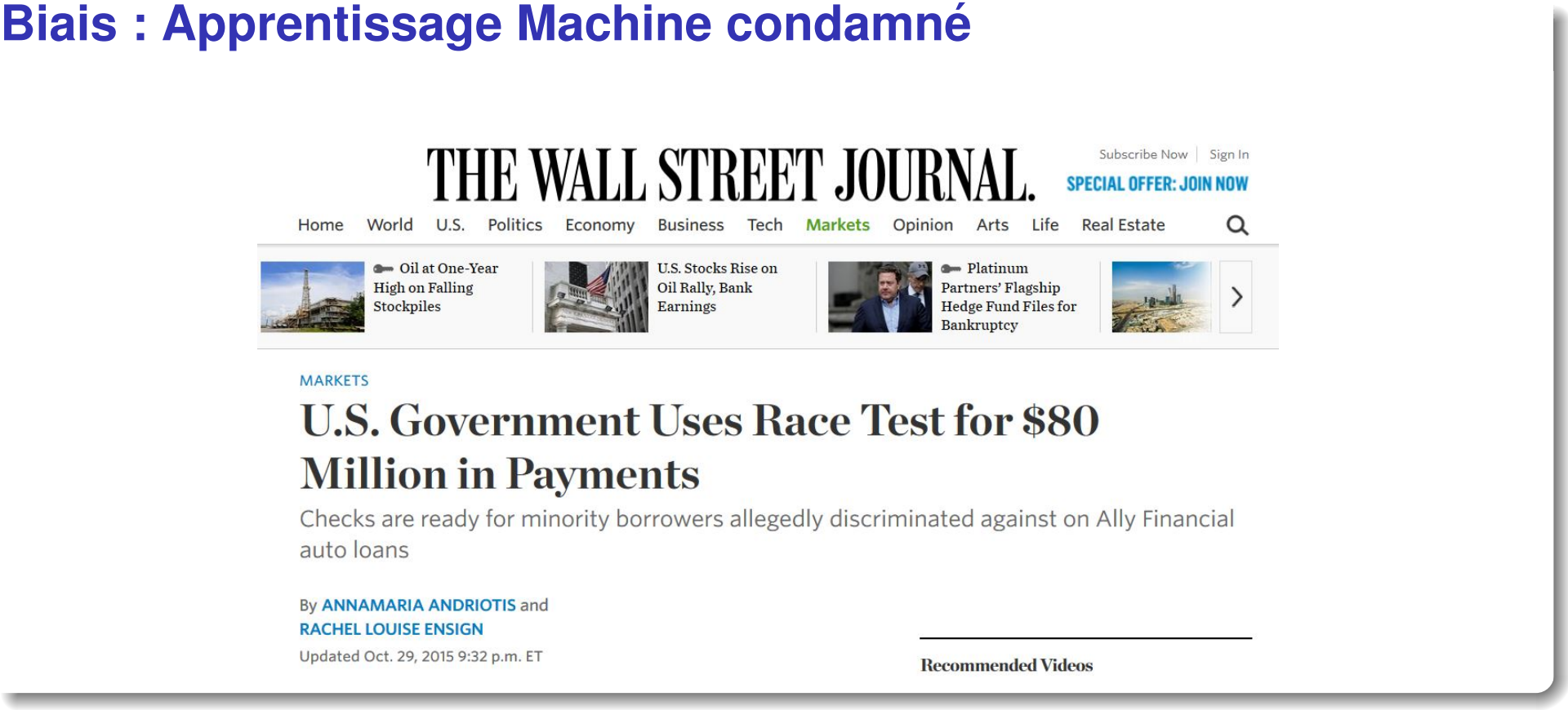
Partisans de Donald Trump, soutiens du GamerGate, ou simples internautes adeptes du chaos se sont associés pour transformer Tay, un programme censé imiter la conversation d'une adolescente, en nazie.

Le Monde.fr | 25.03.2016 à 15h50 • Mis à jour le 25.03.2016 à 17h04 |

Par William Audureau

Abonnez vous à partir de 1 € Réagir Ajouter

Biais : Apprentissage Machine condamné



THE WALL STREET JOURNAL.

Subscribe Now | Sign In

SPECIAL OFFER: JOIN NOW

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Oil at One-Year High on Falling Stockpiles

U.S. Stocks Rise on Oil Rally, Bank Earnings

Platinum Partners' Flagship Hedge Fund Files for Bankruptcy

MARKETS

U.S. Government Uses Race Test for $80 Million in Payments

Checks are ready for minority borrowers allegedly discriminated against on Ally Financial auto loans

By ANNAMARIA ANDRIOTIS and RACHEL LOUISE ENSIGN

Updated Oct. 29, 2015 9:32 p.m. ET

Recommended Videos

Mesures de discrimination

Table : Appartenance au groupe × Nature de la décision

Groupe Protégé	Décision Positive	Décision Négative	
Oui	a	b	n_1
Non	c	d	n_2
	m_1	m_2	n

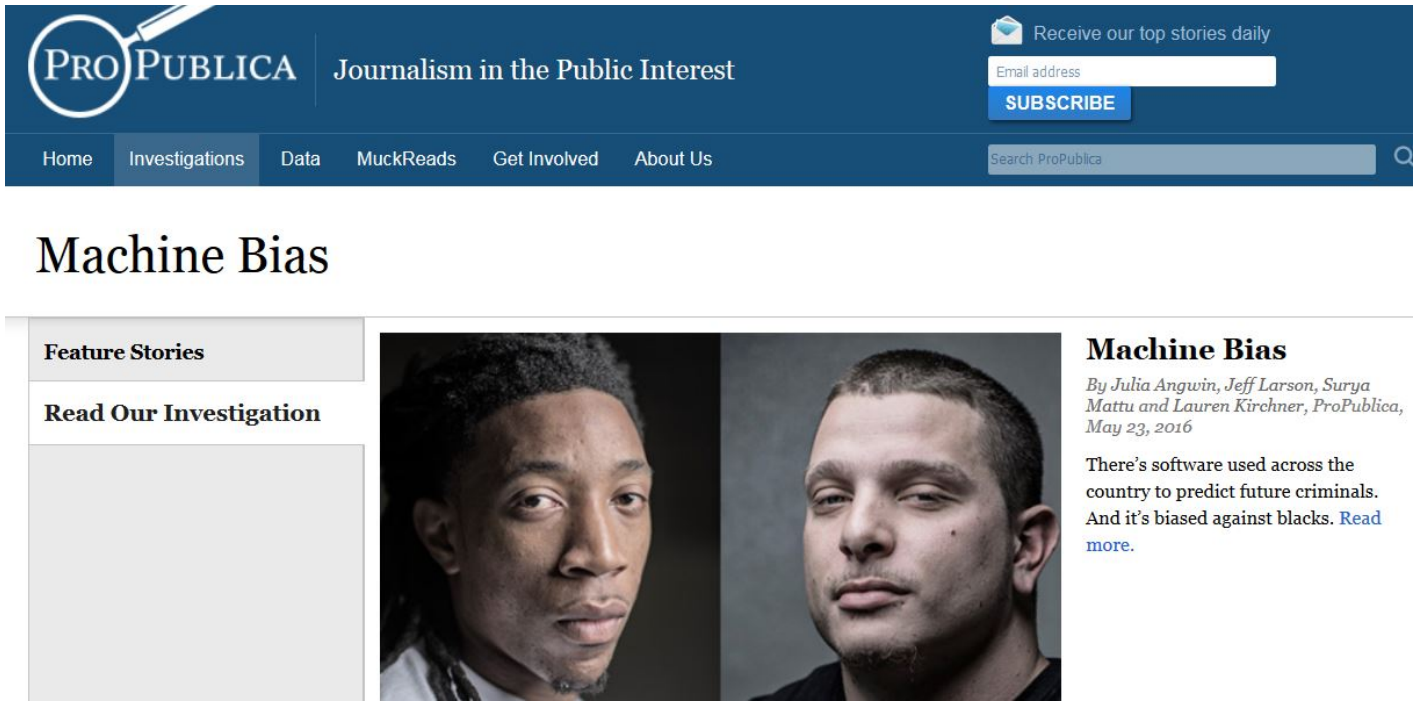
Proportions : $p_1 = a/n_1$, $p2 = c/n_2$, $p = m_1/n$

Mesures simples (Pedreschi et al. 2012)

- Différence de risque : $DR = p_1 - p_2$
- Risque relatif : $RR = p_1/p_2$
- Chance relative : $CR = (1 - p_1)/(1 - p_2)$
- Rapport de cote (*odds ratio*) : RR/CR

Autres mesures : Zliobaité (2015)

Justice prédictives : ProPublica vs. Equivant (NorthPointe Inc.)



Angwin et al. 2016

ProPublica vs. Equivant (NorthPointe Inc.)

Matrice de confusion

Observation / Récidive	Score: Faible	Score: Élevé	
Oui	FN	VP	q_1
Non	VN	FP	q_2
	m_1	m_2	n

- Absence de discrimination selon NorthPointe Inc.
  - Distributions des scores ($m1$ et $m2$) similaires
  - Taux d'erreur ($FN + FP/n$) similaires
- Discrimination selon ProPublica
  - Taux de faux positifs$= FP/q_2$
    afro-américain (45%) vs. caucasiens (25%)
- Taux de récidive afro-américain plus élevé (Chouldechova ; 2016)
- Taux d'erreur très élevé (40%)

Objectif : respect de la loi et acceptabilité des technologies

Actions a priori

- Qualité de prévision : optimiser et informer (cf.sondages)
- Compromis interprétation / qualité (Zeng et al. 2016)
- Algorithme le moins biaisé

Débiaiser une décision algorithmique

- Débiaiser l'échantillon d'apprentissage (variable sensible connue)
  - Supprimer des variables
  - Pondérer les observations (Kamiran et Calders, 2011)
  - Rapprocher les distributions conditionnelles (Feldman et al. 2015)
  - Biais et confidentialité différentielle (Ruggiery, 2014),Hajian et al. (2014)
- Débiaiser l'algorithme
  - Adapter une méthode : arbre (Kamiran et al. 2010)
  - Ajouter une contrainte de "loyauté" (Zafar et al. 2017)

Actions a posteriori

- Face aux Disruptions technologiques
- Faire évoluer le cadre juridique
- Auditer, contrôler un algorithme
- Comment ?
  - Vérifier la précision, l'explicabilité ou l'interprétabilité
  - Détecter un biais (Ruggieri et al. 2010).
- Par qui ?
  - CNIL, DGCCRF (répression des fraudes)
  - Plateformes collaboratives : *Data transparency lab*, TransAlgo (INRIA)
  - Médias : *ProPublica*
  - Associations : *Bayes Impact*
  - ... ?

Conclusion

- Enjeu : Acceptabilité ou rejet (cf. nanostech., OGM, care.data...)
- *Trustworthiness* et *Accountability*
  - Précision, explicabilité, biais des algorithmes
  - Cadre juridique imprécis : proposer des modifications ?
  - Difficile à contrôler
  - Par qui ?

Références

- Allègre C. (2010). *L ?imposture climatique ou la fausse écologie*, Plon.
- Angwin J., Larson J., Mattu S., Kirchner L. (2016). How we analyzed the compas recidivism algorithm. ProPublica, en ligne consulté le 28/04/2017.
- Anserson, C. (2008). *The End of Theory : The Data Deluge Makes the Scientific Method Obsolete*, Wired.
- Chouldechova A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, arXiv pre-print.
- Datta A., Sen S., Zick Y. (2016). Algorithmic Transparency via Quantitative Input Influence : Theory and Experiments with Learning Systems, in IEEE Symposium on Security and Privacy.
- Ezrachi A., Stucke M. (2016). *Virtual Competition The promise and perils of algorithmic-driven economy*, Harward University Press.
- Feldman M., Friedler S., Moeller J., Scheidegger C., Venkatasubramanian S. (2015). Certifying and removing disparate impact, arXiv-preprint.
- Goodman B. (2016). A Step Towards Accountable Algorithms?:Algorithmic Discrimination and the European Union General Data Protection, in 29th Conference on Neural Information Processing Systems (NIPS 2016).
- Goodman B., Flaxman S. (2016). EU regulations on algorithmic decision-making and a ”right to explanation”, ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York.
- Hajian S., Domingo-Ferrer J., Farràs O. (2014). Generalization-based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining, Data Mining and Knowledge Discovery 28 (5-6), 1158-1188.
- Kamiran F., Calders T. (2011). Data Pre-Processing Techniques for Classification without Discrimination, Knowledge and Information Systems 33(1).
- Kamiran F., Calders T, Pechenizkiy M. (2010). Discrimination Aware Decision Tree Learning in ICDM, 869-874.

Références – suite

- Morozov E. (2014). The rise of data and the death of politics, The Observer.
- Rouvroy A., Berns T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation, *Réseaux*, 177, 163-196.
- Ruggieri S. (2014). Using t-closeness anonymity to control for non-discrimination, Transaction on Data Privacy, 7, 99-129.
- Serres M., Legros M., Ortoli, S. (2014). *Pantopie : de Hermès à Petite Poucette*, Le Pommier.
- Stiegler B., Kyrou A. (2015). *L'Emploi est Mort, Vive le Travail !*, Fayard.
- Trafimow, D., Marks, M. (2015), Editorial, *Basic and Social Psychology*, 37, 1 ?2
- Zafar M., Valera I., Rodriguez M., Gummadi K. (2017). Fairness Constraints : Mechanisms for Fair Classication in International Conference on Articial Intelligence and Statistics (AISTATS), vol. 5.
- Wachter S., Mittelstadt B., Floridi L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, à paraître.
- Zeng J., Ustuny B., Rudin C. (2016). Interpretable Classification Models for Recidivism Prediction, arXiv pre-print.
- Zliobaité I. (2015). A survey on measuring indirect discrimination in machine learning. arXiv pre-print.

Quelques sites consultés

- Le Monde: IBM, Google, Microsoft et le cancer
- ArsTechnika: Programme Skynet de la NSA
- Mesure d'audience
- Le Monde: révélation du code source d'APB
- Predpol: Police prédictive
- NorthPointe : prévision de la récidive
- ProPublica : article sur le biais des "machines"
- Le Monde: *Tay, ChatBot* raciste de Microsoft
- Village de la Justice: Justice prédictive
- Wall Street Journal: condamnation du gouvernement
- Partenariat sur l'IA