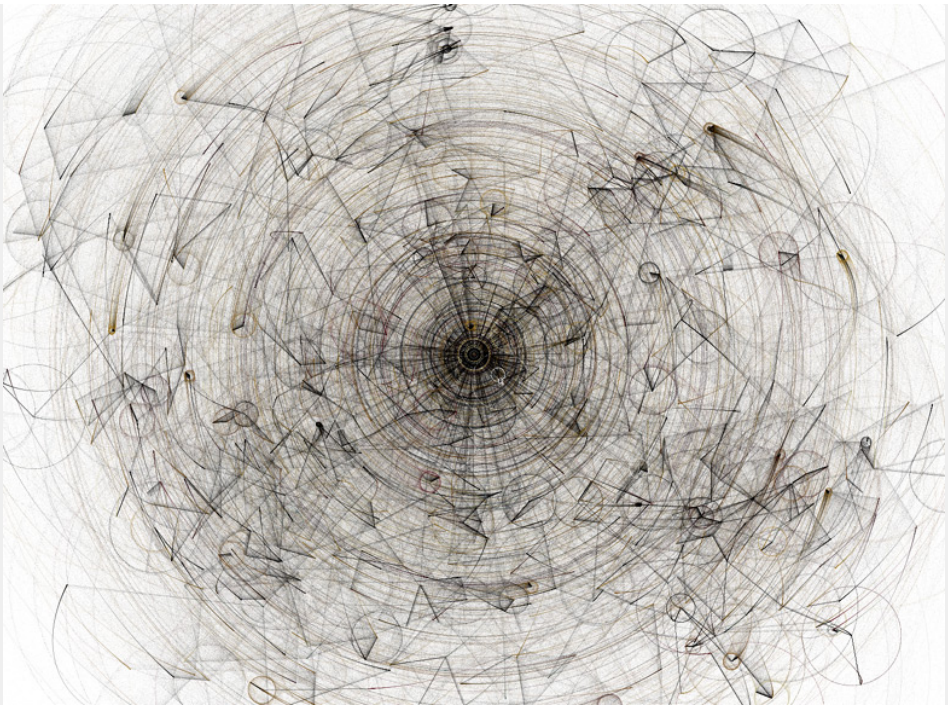
De la centralité des bases de données "ground truth" dans la construction des systèmes algorithmiques



Les méthodes informatiques de calcul – souvent appelées « algorithmes » – alimentent le fonctionnement de dispositifs ordinaires. Moteurs de recherche, réseaux sociaux, systèmes de surveillance, plateforme d'achats en ligne : qu'on le veuille ou non, nous ne cessons d'interagir avec ces systèmes numériques ainsi qu'avec les algorithmes qui fondent leurs mécanismes internes. Comment reprendre la main sur ces entités souvent présentées (à raison) comme opaques, arbitraires et vectrices d'inégalités ? En se basant sur les résultats d'enquêtes ethnographiques récentes, Florian Jaton propose de se focaliser sur les bases de données référentielles – souvent appelées *ground truth* – qui sont au fondement du travail de construction algorithmique.

La cause semble étendue : du fait notamment de l'avènement de l'informatique mobile et distribuée, nos vies deviennent de plus en plus algorithmiques, parfois pour le meilleur, souvent pour le pire[1]. Que faire à partir de là ? Dans cette note, je souhaiterais souligner l'importance des bases de données référentielles – souvent appelées *ground truth* dans la littérature technique – qui fondent la véridiction, relative, des algorithmes. L'argument sociotechnique est le suivant : la réalité des algorithmes est structurée par la réalité des bases de données *ground truth* desquelles ils émanent. L'argument politique est le suivant : si nous – consommateur·rice·s, militant·e·s, ingénieur·e·s, politicien·ne·s – parvenions à saisir les bases de données *ground truths*, peut-être parviendrions-nous à saisir les algorithmes, pour mieux les changer.

Les bases de données ground truth, ou le fondement des algorithmes

Les méthodes informatiques de calcul – souvent appelées « algorithmes » – ne tombent pas du ciel (même si l'on peut parfois en avoir l'impression) : comme tous les autres dispositifs techniques, ce sont des constructions culturelles. Et parmi les nombreuses pratiques et matériaux concrets dont ces constructions ont besoin pour advenir et circuler, il y a des bases de données référentielles. Ces bases de données – souvent appelées *ground truth* – organisent, et donc mettent en relation, au moins deux sous-ensembles : les données d'entrée (« input-data », ce que le futur algorithme devra traiter) et les cibles de sortie (« output-targets », ce que le futur algorithme devra produire). En tant que telles, ces bases de données *ground truth* définissent l'objectif fondamental des algorithmes : ils doivent se rapprocher, au mieux, de la fonction qui organise les relations entre les données d'entrée et les cibles de sortie (voir figure 1). Mais comme ces bases de données servent également à évaluer les performances des algorithmes et à les comparer entre eux (voir figure 2), elles constituent également leur limite : les algorithmes – quoi qu'on en dise – ne peuvent aller au-delà des bases de données dont ils émanent. En ce sens, les *ground truths* sont à la fois les matrices fondamentales des algorithmes et leurs horizons indépassables.

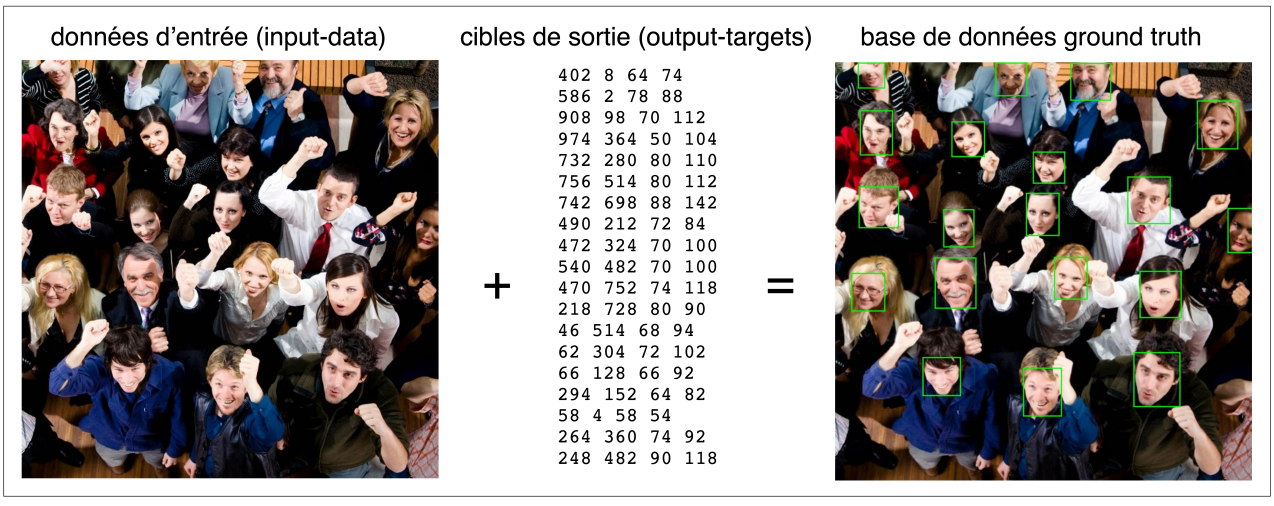


Figure 1. Tiré de Yang et al. (2016), échantillon de la base de données ground truth WIDER FACE pour la recherche en détection faciale. À gauche, une parmi les 32 203 images de ce jeu de données accessible publiquement. Au milieu, les annotations de visage pour cette image spécifique. Comme chaque annotation appartient à l'espace de coordonnées de l'image, elle peut être exprimée par un ensemble de quatre

valeurs, les deux premières exprimant la position de départ de l'annotation le long des axes x et y, la troisième exprimant le nombre de pixels de largeur, la quatrième exprimant le nombre de pixels de hauteur. Ces informations, qui correspondent aux rectangles verts de l'image de droite, ont été produites manuellement par un annotateur humain et vérifiées par deux autres (Yang et al., 2016 : 5527). En tant que telles, ces annotations constituent les cibles de sorties (output-targets) de l'image pour ce qui est de la détection faciale ; elles viennent s'additionner aux données d'entrée afin de fournir quelque chose à apprendre et à formuler. Les données d'entrée (input-data) et leurs cibles de sorties (output-targets) peuvent ensuite être utilisées pour construire des algorithmes de détection faciale, la base de données ground truth opérant comme la liste des meilleures réponses pour cette tâche précise. Source : Florian Jaton, 2021bi

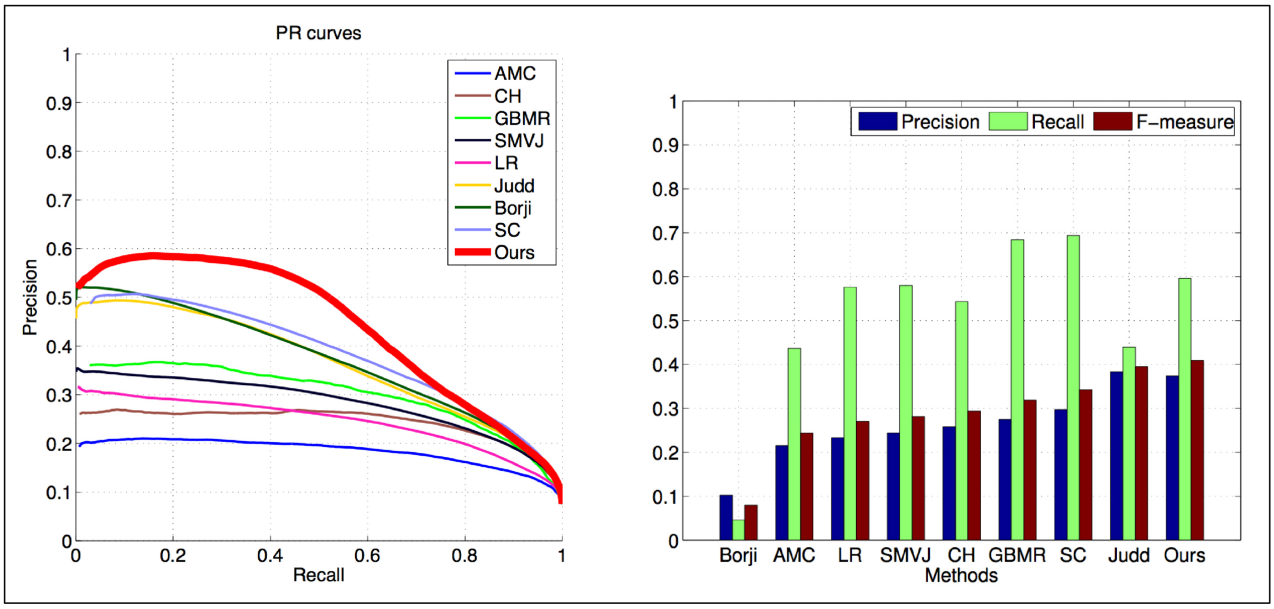


Figure 2. Tiré de Jaton (2021 : 75), deux graphiques qui comparent les performances de plusieurs algorithmes de détection (« AMC », « CH », « Ours », etc.) sur une base de données ground truth en mobilisant une série de mesures statistiques (« precision », « recall », etc.). L'important ici est de constater que ces deux graphiques sont rendus possibles grâce à une base de données ground truth qui définit les valeurs maximales indépassables (coordonnée [1;1] pour le graphique de gauche et valeur [1] sur l'histogramme de droite).

Tout comme les algorithmes qu'elles engendrent et confinent, les bases de données *ground truth* ne préexistent pas : elles découlent de processus collectifs concrets. Ces processus dits de *ground-truthing* engagent plus ou moins de personnes, d'efforts et de ressources[2]. Pour autant, sans même prendre en compte le problème important des conditions de travail des précaires « travailleur·euse·s du clic » souvent mobilisé·e·s pour la production des cibles de sortie[3], les produits de ces processus (c'est-à-dire les bases de données *ground truth*) restent limités (même si parfois massives, elles restent moins riches que les phénomènes qu'elles tentent de décrire), arbitraires (elles pourraient être différentes) et orientés socioculturellement (les positions et habitudes des personnes impliquées dans leur construction influencent leur contenu). Par conséquent, les algorithmes – en tant que dispositifs visant à approximer les relations entre les éléments des bases de données *ground truth* – sont également limités, arbitraires et socioculturellement

orientés. L'argument, évident dès lors que l'on s'engage dans des enquêtes ethnographiques sur la construction concrète des algorithmes, a le mérite de la simplicité : nous récoltons les algorithmes de nos (ou plutôt *leurs*) bases de données *ground truth*[4].

Bases de données ground truth et algorithmes d'apprentissage machine

Mais qu'en est-il des algorithmes dits d'apprentissage machine (*machine learning*), ces entités qui sont aujourd'hui de plus en plus affiliées à de l'intelligence artificielle (ce qui n'aide pas à l'éclaircissement du débat) ? Dans la littérature spécialisée, il est convenu de distinguer deux grands types d'algorithmes d'apprentissage machine : les algorithmes dits « supervisés », et les algorithmes dits « non-supervisés »[5].

Le cas des algorithmes supervisés est relativement simple. Ces méthodes informatiques de calcul qui mobilisent statistiques et probabilités pour approximer des fonctions d'apprentissage ont besoin de cibles de sortie, c'est-à-dire de réponses étiquetées comme « correctes ». En ce sens, ces algorithmes dépendent également de bases de données *ground truth* limitées, arbitraires et socioculturellement orientées. Par conséquent, ces algorithmes *sont* également limités, arbitraires et socioculturellement orientés. C'est là une caractéristique établie, démontrée par de nombreuses études sociales sur la construction des algorithmes[6].

Le cas des algorithmes non-supervisés (ou auto-supervisés) est plus délicat. Même si ces algorithmes spécifiques (et de plus en plus sophistiqués) s'appuient uniquement sur les données d'entrée des bases de données *ground truth* pour définir leur fonction d'apprentissage, ils doivent néanmoins être confrontés aux cibles de sortie pour être évalués et comparés. Plus qu'un impératif technique, il s'agit là d'une nécessité pratique : sans la possibilité de s'en référer à des cibles de sortie opérant comme valeurs de référence, les ingénieur·euse·s informatiques sont dans l'incapacité de mesurer (et de faire valoir) les performances de leurs algorithmes d'apprentissage non/auto-supervisés, ce qui est une condition à leur mise en existence et en circulation[7]. Cette nécessité pratique est liée au fait que les algorithmes d'apprentissage non/auto-supervisés ne sont généralement pas destinés à rester théoriques : ils sont conçus pour être utilisés et travaillés, ce qui implique de les comparer à des bases de données *ground truth* afin de démontrer leur pertinence et leur efficacité. En somme, et similairement à tous les autres algorithmes informatiques, les algorithmes d'apprentissage non/auto-supervisés dépendent de bases de données *ground truth* limitées, arbitraires et orientées socioculturellement. En ce sens, et malgré certains discours promotionnels peu avisés, les algorithmes d'apprentissage non/auto-supervisées *sont* également limités, arbitraires et socioculturellement orientés.

Se saisir des algorithmes revient (en partie) à se saisir des bases de données ground truth

On l'aura compris, le poids des bases de données *ground truth* dans l'existence des algorithmes – qu'ils soient dits d'apprentissage machine, d'intelligence artificielle, ou encore de Big Data – est conséquent. Quelles conclusions (provisoires) tirer de cette observation, tout à fait banale pour les ingénieur·euse·s informatiques qui travaillent chaque jour à l'élaboration de nouveaux algorithmes, mais souvent surprenante pour tous.tes les autres ?

La première serait sans doute de considérer les algorithmes non plus, classiquement, comme des dispositifs informatiques qui résolvent des problèmes, mais bien comme des dispositifs qui retrouvent des réponses à des problèmes définis en amont, au sein des bases de données *ground truth* qui fondent leur but et horizon.

Dès qu'il serait question d'un algorithme problématique, l'attention se focaliserait dès lors sur la base de données *ground truth* concrète de laquelle cet algorithme a émané. Quand est-ce que cette base de données a été construite ? Et par qui ? Et selon quelle modalité ? Et somme, ce réflexe – qu'il s'agirait d'instaurer – permettrait de ramener la question des algorithmes à des considérations matérielles et situées.

D'où la deuxième conclusion, qui suggère un potentiel levier d'action. Si les mathématiques mobilisées pour la construction des algorithmes (c'est-à-dire pour approximer la fonction qui organise les relations entre les données d'entrée et les cibles de sortie des bases de données *ground truth*) ainsi que le code nécessaire à leur expression au sein de machines informatiques apparaissent – à raison – comme hautement spécialisés et relativement opaques, les bases de données *ground truth* sont beaucoup plus intuitives. Il s'agit en effet de simples listes de données produites via des processus collectifs assignables qui peuvent être décrits. D'où une capacité accrue d'audit, les ressorts sociaux participant à l'élaboration de ces bases de données limitées, arbitraires et socioculturellement orientées pouvant être évaluées par des personnes ne provenant pas seulement du champ restreint de l'informatique appliquée. Revendiquer ce droit de regard critique sur ce qui fonde la véridiction des algorithmes qui contribuent à façonner nos vies constitue, selon moi, un élément central à la politisation des processus algorithmiques.

Notes de bas de page :

[1] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, Cambridge, Massachusetts, 2016; Mazzotti, Massimo, « Algorithmic Life », *Los Angeles Review of Books*, 2017.

[2] Florian Jaton, *The Constitution of Algorithms: Ground-Truthing, Programming, Formulating*, MIT Press, Cambridge, Massachusetts, 2021.

[3] Antonio A. Casilli, *En attendant les robots. Enquête sur le travail du clic*, Paris, Seuil, 2019.

[4] Florian Jaton, « We get the algorithms of our ground truths: Designing referential databases in digital image processing », *Social Studies of Science*, 2017, vol. 47, n° 6, pp. 811–840.

[5] Ethem Alpaydin, *Machine Learning: The New AI*, MIT Press, Cambridge, Massachusetts, 2016. Dans cette note, je ne parlerai pas de l'apprentissage par renforcement (*reinforcement learning*), aujourd'hui encore marginal.

[6] Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press, New Haven, 2021; Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, New York University Press, 2018.

[7] Anja Bechmann, Geoffrey C. Bowker, « Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media », *Big Data & Society*, 2019, vol. 6, n° 1.

Pour citer cet article :

Florian Jaton, «De la centralité des bases de données *ground truth* dans la construction des systèmes algorithmiques», *Silomag*, n°15, juillet 2022. URL: https://silogora.org/de-la-centralite-des-bases-de-donnees-ground-truth-dans-la-construction-des-systemes-algorithmiques/