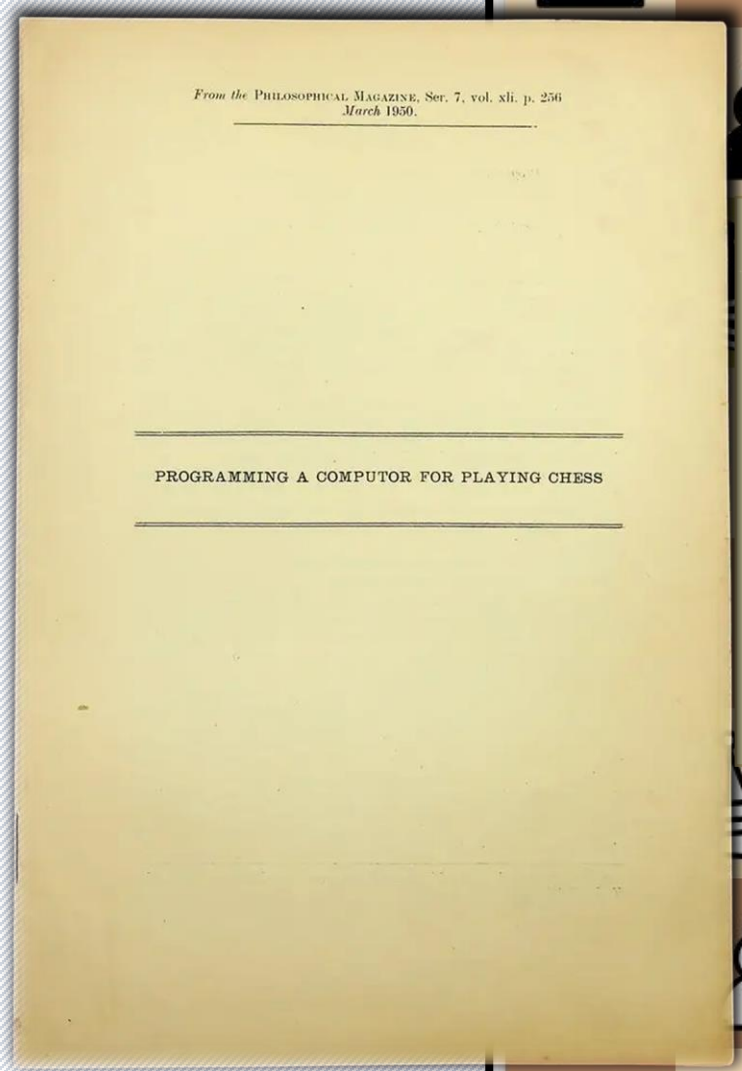
Comprendre et évaluer les intelligences artificielles

Adrien Foucart

Skeptics in the Pub Liège

8 décembre 2023



From the PHILOSOPHICAL MAGAZINE, Ser. 7, vol. xli. p. 256
March 1950.

PROGRAMMING A COMPUTOR FOR PLAYING CHESS

On considère généralement que les échecs nécessitent de la « pensée » pour y jouer avec talent; une solution à ce problème nous forcera soit à admettre la possibilité d'une pensée mécanisée, soit à restreindre plus notre concept de la « pensée ».

-- *Claude Shannon, 1950*

[My First Decisive Game! | Qatar Masters Round 3], @eric-rosen on YouTube

[My First Decisive Game! | Qatar Masters Round 3], @eric-rosen on YouTube

Deep Blue

Murray Campbell [a,*], A. Joseph Hoane Jr. [b], Feng-hsiung Hsu [c]

[a] *IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA*
[b] *Sandbridge Technologies, 1 N. Lexington Avenue, White Plains, NY 10601, USA*
[c] *Compaq Computer Corporation, Western Research Laboratory, 250 University Avenue, Palo Alto, CA 94301, USA*

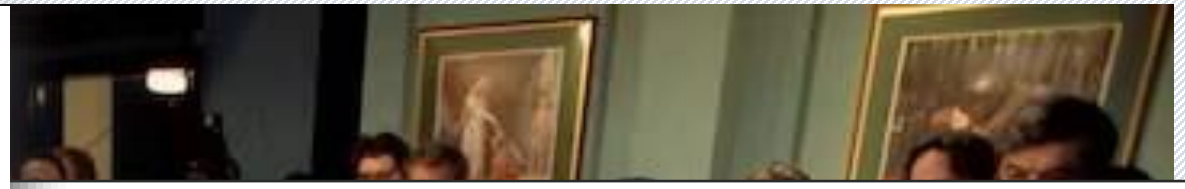
Abstract

Deep Blue is the chess machine that defeated then-reigning World Chess Champion Garry Kasparov in a six-game match in 1997. There were a number of factors that contributed to this success, including:

- a single-chip chess search engine,
- a massively parallel system with multiple levels of parallelism,
- a strong emphasis on search extensions,
- a complex evaluation function, and
- effective use of a Grandmaster game database.

This paper describes the Deep Blue system, and gives some of the rationale that went into the design decisions behind Deep Blue. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Computer chess; Game tree search; Parallel search; Selective search; Search extensions; Evaluation function



« Le livre des ouvertures pour Deep Blue était créé à la main, principalement par le Grand Maître Joel Benjamin, avec l'assistance des Grands Maîtres Nick De Firmian, John Fedorowicz et Miguel Illescas (…) Avant la partie, un répertoire particulier était choisi pour Deep Blue. Il y avait un certain nombre de répertoires parmi lesquels choisir, et le choix était fait selon la situation du match et les expériences précédentes en jouant avec la même couleur. Des changements de dernières minutes pouvaient être mises dans un petit livret "override". »

-- M. Campbell et al., 2002

$$f(P) = 200(K-K') + 9(Q-Q') + 5(R-R') + 3(B-B'+N-N') + (P-P') - 0.5(D-D'+S-S'+I-I') + 0.1(M-M') + \ldots$$

Claude Shannon, 1950

Jeff Christensen/Reuters
https://theconversation.com/twenty-years-on-from-deep-blue-vs-kasparov-how-a-chess-match-started-the-big-data-revolution-76882

Deep Blue : phase d'**ouverture** prédéterminée par des humains, phase de **recherche** par énumération des positions possibles (100 millions par seconde) sur base d'une **fonction d'évaluation** prédéterminée par des humains.

Qui a battu Kasparov ? Une machine… ou des humains aidés d'une puissante calculatrice ?

Jeff Christensen/Reuters
https://theconversation.com/twenty-years-on-from-deep-blue-vs-kasparov-how-a-chess-match-started-the-big-data-revolution-76882

Comment ça marche, une « intelligence artificielle » moderne ?

Deep learning

Algorithmes d'apprentissage sur des modèles constitués de grands réseaux de neurones artificiels avec des millions de paramètres.

Entrée(s) → Modèle → Sortie(s)

Images, sons, documents, mesures…

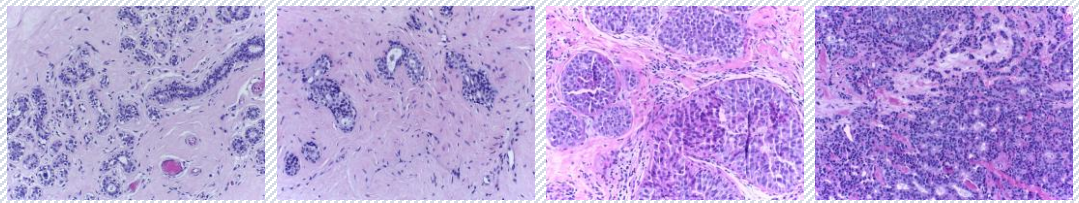
Décisions, prédictions, complétions…

x y z

$ax + by + cz + d \geq 0 \longrightarrow c_1$

sinon $\longrightarrow c_2$

Paramètres à « apprendre »



Normal | Benign | *In situ* | Invasive

Images: ICIAR BACH Challenge 2018.

Comment ça marche, une « intelligence artificielle » moderne ?

Deep learning

Algorithmes d'apprentissage sur des modèles constitués de grands réseaux de neurones artificiels avec des millions de paramètres.

$$f(\boldsymbol{x}; \boldsymbol{w}) = f_2(f_1(f_0(\vec{x}; \vec{w}_0); \vec{w}_1); \vec{w}_2)$$

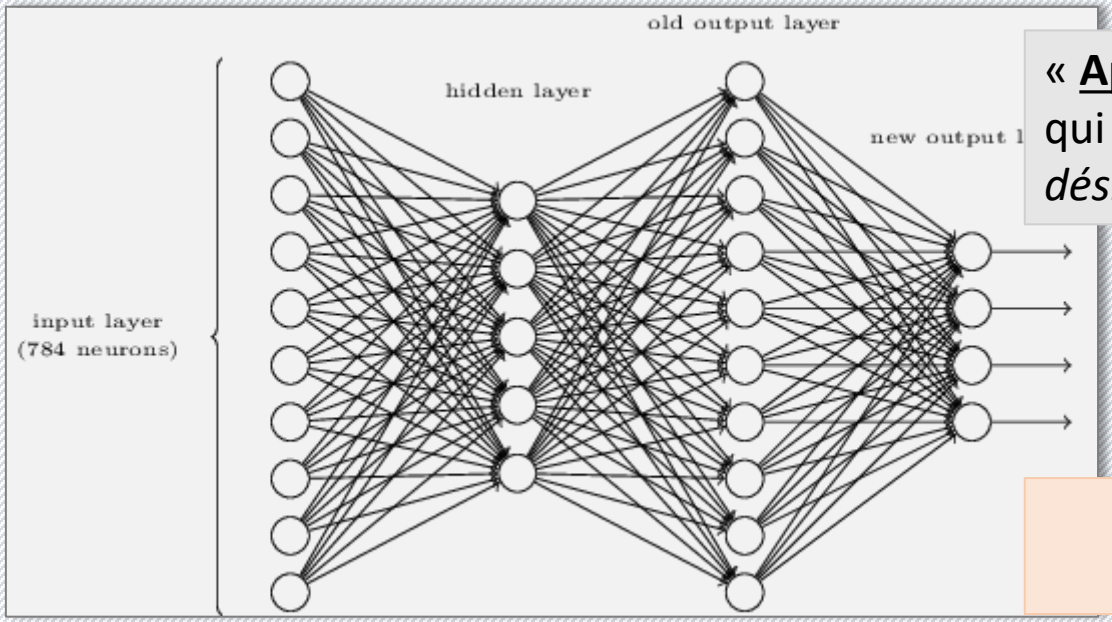


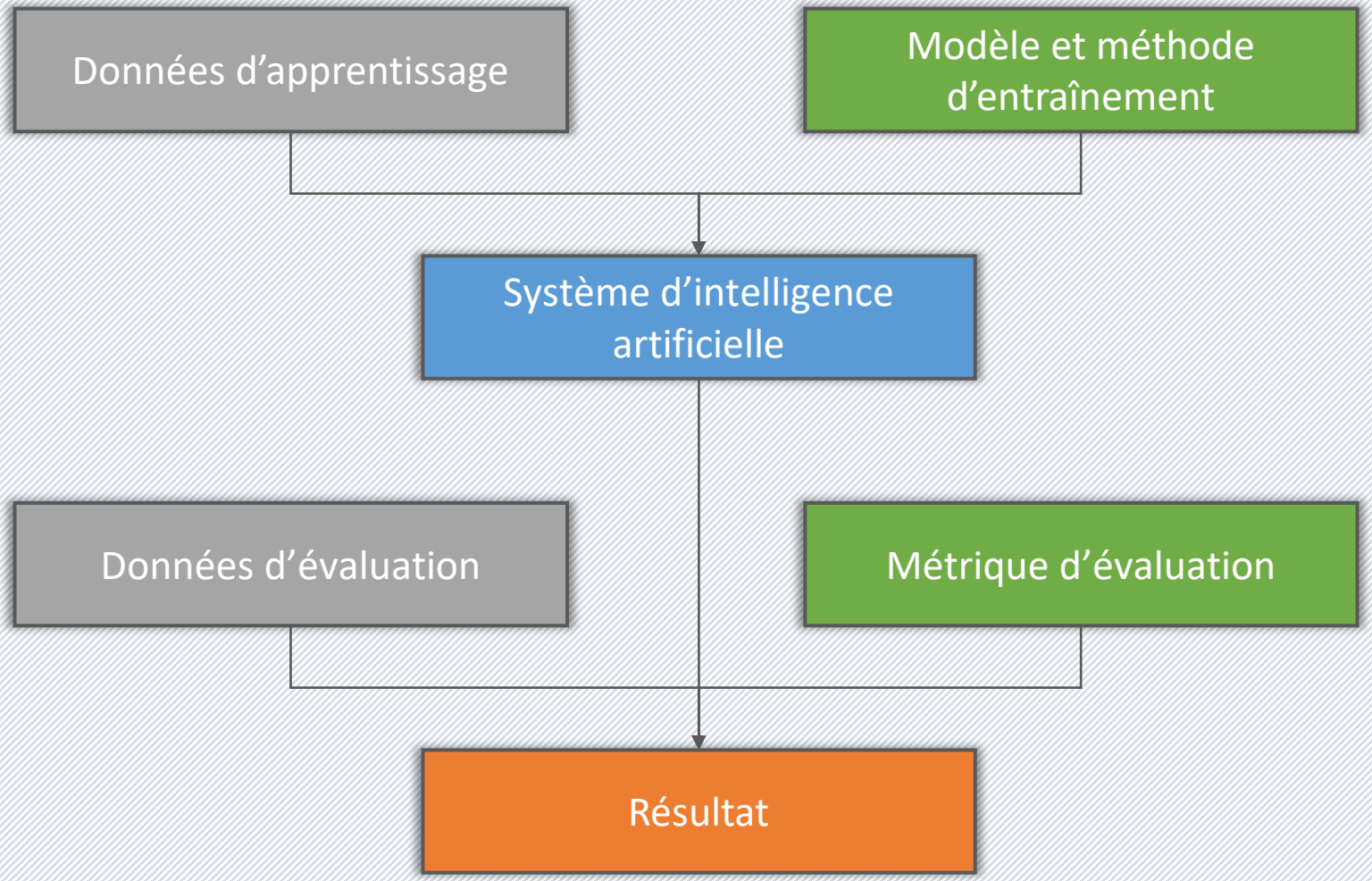
Image: Michael A. Nielsen, *Neural Networks and Deep Learning*, 2015

« **Apprentissage** » = **trouver les paramètres** du modèle qui permettent *le mieux possible* d'obtenir la *sortie désirée du système* pour un *ensemble de données*.

« **Évaluation** » = **estimation** de la *performance du système* sur un *autre ensemble de données*.

Modèle de **réseau de neurone**, non-linéaire
chaque connexion = 1 paramètre

Apprentissage et évaluation: le scénario idéal



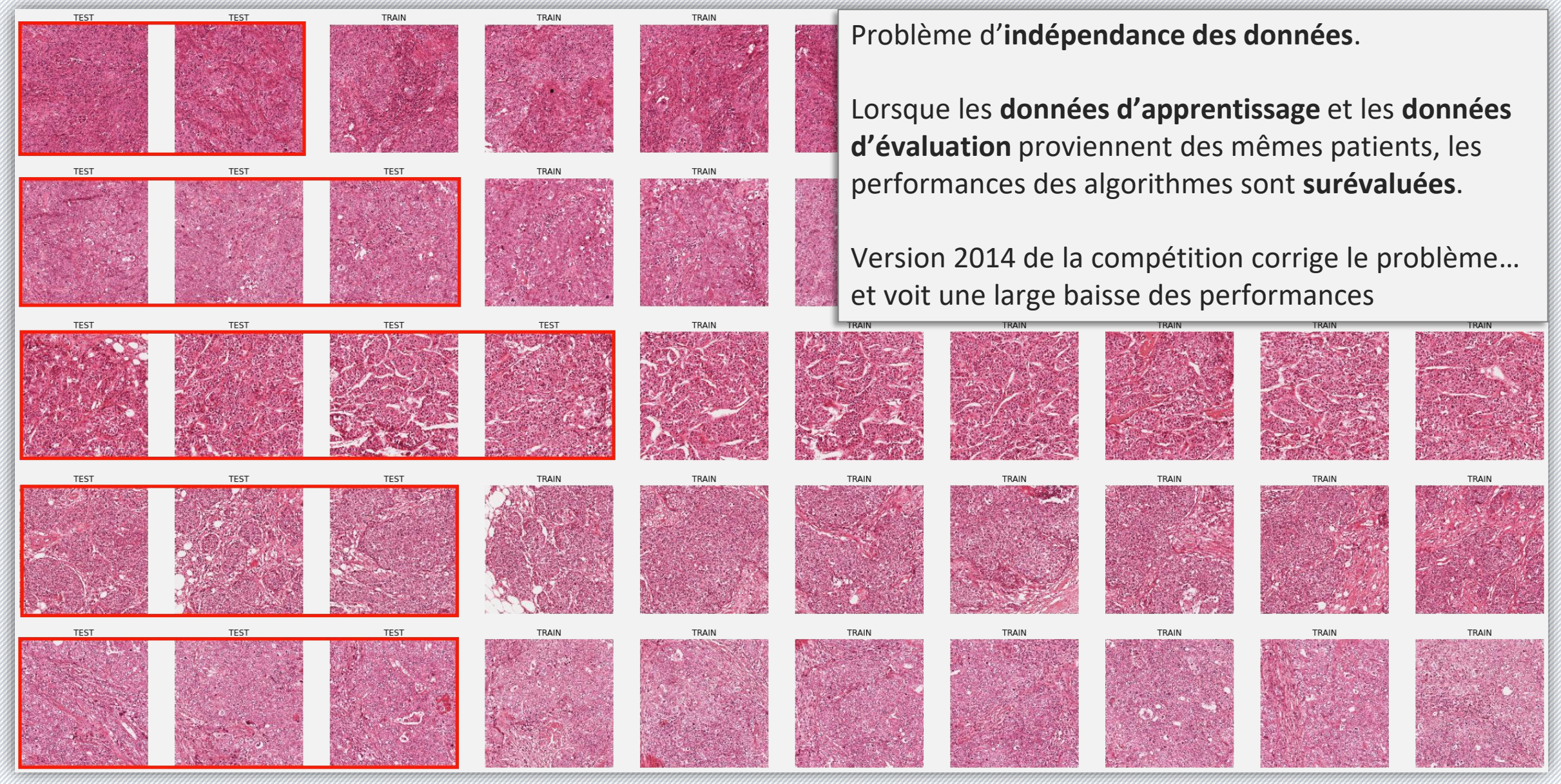
⛔ Retour en arrière interdit ! (modifier et réévaluer, c'est tricher).

📎 Plus les données d'évaluation sont **indépendantes** des données d'apprentissage, plus les résultats sont exploitables.

📏 La **métrique** doit être **corrélée** au succès de l'algorithme « dans la vraie vie ». La **tâche** à résoudre doit être bien identifiée.

🙈 Les données d'évaluation ne sont **pas accessibles** à celleux qui créent/entraînent l'algorithme avant l'évaluation, et l'évaluation est faite par quelqu'un d'autre que celleux-ci.

Quand ça ne va pas: l'exemple de MITOS12



Problème d'**indépendance des données**.

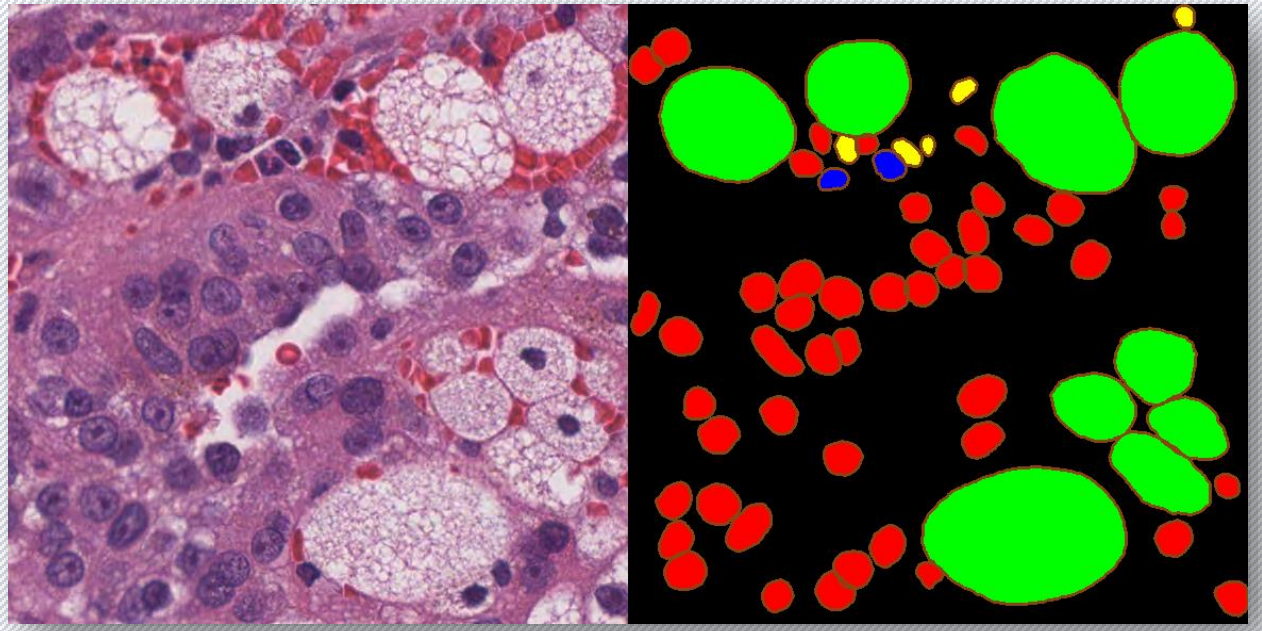
Lorsque les **données d'apprentissage** et les **données d'évaluation** proviennent des mêmes patients, les performances des algorithmes sont **surévaluées**.

Version 2014 de la compétition corrige le problème… et voit une large baisse des performances

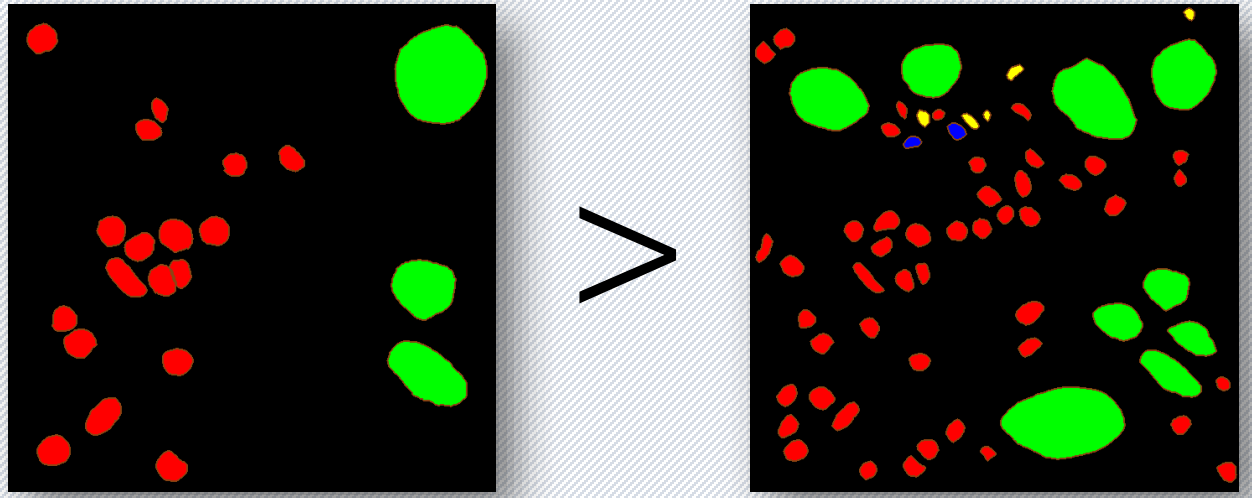
Quand ça ne va pas: l'exemple de MoNuSAC20

Problème n°1
Erreur dans le *code* de la métrique d'évaluation.

Problème n°2
Choix de la métrique peut amener à des mauvaises conclusions.



$$PQ = \frac{\sum_{(x,y)\in TP} IoU(x,y)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$



Révolution dans le monde de la tech: une
nou
per
tâc
GP

THE COURT: How did you go about finding the cases that you cited in your memoranda?

MR. SCHWARTZ: First, I went to Fastcase, which is the research tool that our office subscribes to. It did not have access to federal cases that I needed to find, so I began to attempt to try to find another source to find the cases. I tried Google. Again, I didn't have access to Westlaw or Lexis. And it had occurred to me that I heard about this new site which I assumed -- I falsely assumed was like a super search engine called ChatGPT, and that's what I used.

Mata v Avianca, #52

23.
re

IN THE UNITED ST
FOR THE SOUTHE
ROBERTA MATA,
-against
AVIANCA, INC.,
Defendant(s).

Mata v Avianca, #21

itations. Although
ndersigned has been
osition, and the few
positions for which

Avianca, #24

GPT-4 : évaluation et marketing

Exam	GPT-4
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)
LSAT	163 (~88th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)
SAT Math	700 / 800 (~89th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)
USNCO Local Section Exam 2022	36 / 60
Medical Knowledge Self-Assessment Program	75 %

GPT-4 Technical Paper, OpenAI.
https://arxiv.org/abs/2303.08774

Évaluations fournies par OpenAI…

- Ne sont pas revues par des pairs.
- Ne donnent pas de détails sur les données d'entraînement et de test.
- Ne donnent pas de détails sur le fonctionnement du modèle.
- Ne font pas le lien entre la méthode d'évaluation et des cas d'utilisation réels du système.

De telles évaluations relèvent du domaine du **marketing**, pas de la **recherche scientifique**.

then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

Évaluer une « IA » : quelles questions se poser ?

1 Quelle est la **tâche réelle / précise** que l'on veut réaliser avec le système ?

- Fournir du conseil légal ?
- Aider à de la recherche documentaire ?
- Passer un examen ?

Vendre des abonnements ChatGPT, convaincre Microsoft d'investir des milliards, « battre » Google.

2 Quelles sont les **types d'erreurs** possible ? Quelle est la **gravité relative** de ces erreurs ?

- Se tromper sur un texte de loi.
- Inventer des citations.
- Donner de mauvais conseils.

Mettre OpenAI dans une situation où ils sont légalement responsables d'erreurs du système.

3 Quelles sont les **données** à notre disposition ? Sont-elles **représentatives** de tous les cas d'utilisation de notre tâche ? Est-ce que des **biais** ou **zones d'ombre** existent dans ces données ?

- Wikipedia ?
- Bases de données légales ?
- Forums anonymes ?

Tout ce que OpenAI a pu récupérer sur le web, avec un processus de filtrage minime.

Du déploiement abusif des intelligences artificielles

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By **Jeffrey Dastin**

October 11, 2018 2:50 AM GMT+2 · Updated 5 years ago

Reuteurs, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/

Amazon declined to comment on the technology's challenges, but said the tool "was never used by Amazon recruiters to evaluate candidates." The company did not elaborate further. It did not dispute that recruiters looked at the recommendations generated by the recruiting engine.

Du déploiement abusif des intelligences artificielles

Dissecting racial bias in an algorithm used to manage the health of populations

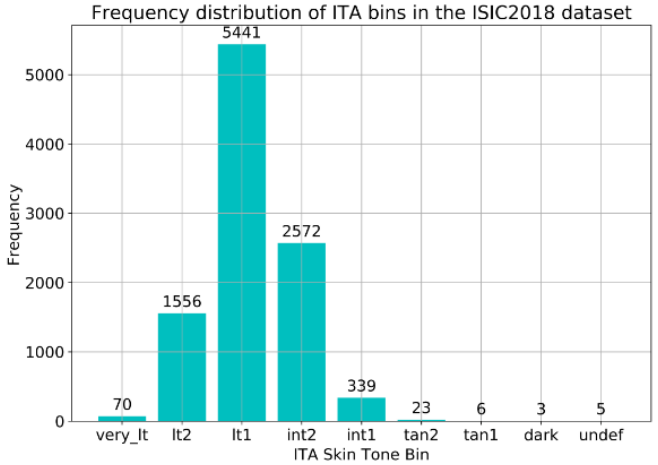
Obermeyer et al., 2019. *Science* 366 (6464).

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

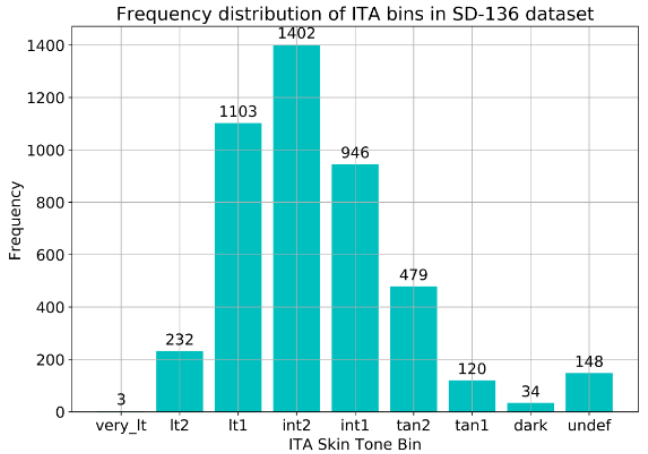
Du déploiement abusif des intelligences artificielles

Fairness of Classifiers Across Skin Tones in Dermatology

Kinyanjui et al. *Proc. MICCAI 2020*.



(a)



(b)

Fig. 3. Skin tone distribution for (a) ISIC2018, and (b) SD-136 entire datasets.

Points clés

🧠 L'**intelligence** est un concept fondamentalement humain. On évalue les **machines** sur base de leur capacité à **résoudre des tâches** bien définies.

✅ **Évaluer** un algorithme d'intelligence artificielle, c'est évaluer **ses résultats** mais aussi **la qualité des données** d'apprentissage *et de test*. Si les données sont biaisées, les résultats seront biaisés aussi.

💰 Derrière les sociétés qui font de l'IA se trouvent des **enjeux financiers énormes**. Une certaine dose de **scepticisme** est nécessaire face aux **résultats annoncés**. Les startups sont financées sur le « hype », pas sur les résultats.

Contact et liens

✉ Adrien.Foucart@ulb.be

📎 **Research blog**: https://research.adfoucart.be

📎 **Blog d'opinions**: https://adfoucart.be/blog

https://social.sciences.re/@AFoucart