Ecole d'hiver é-EGC
11h30-13h00

Machine Learning and interpretability : examples in precision medicine

JEAN-DANIEL ZUCKER
DR IRD

UMMISCO
Unité Mixte Internationale

Travail en collaboration E. Prifti (IRD), Y. Chevaleyre (Dauphine), B. Hanczar (Evry), K. Clément (INSERM) & N. Sokolovska (SU)

75 ans — Sorbonne Université — IntegrOmics — ANR — Investissements d'avenir — iCAN

IRD IS A PUBLIC INSTITUTE THAT UNDERTAKES RESEARCH & TRAINING ACTIVITIES IN PARTNERSHIP TO ADDRESS THE CHALLENGES OF SUSTAINABLE DEVELOPMENT

- 65 research units (jointly with other institutions)
- 2250 agents and a community of 7000 French researchers
- Over 1500 publications/year, 50% in co-publication with partners

ECOBIO Ecology, biodiversity and functioning of continental ecosystems
OCEANS Oceans, Climate and Resources
DISCO Internal dynamics and continents surface
SOC Societies and globalisation
SAS Health and Societies

IRD Institut de Recherche pour le Développement

Why you shouldn't trust today's skin cancer (free) apps diagnosing melanoma

1. **You cannot trust the findings:** Studies indicate that skin cancer apps have poor diagnostic accuracy for melanoma. **When patients do a self-examination of the skin, reported sensitivity (correctly identified) may increase from 25% to 93% and specificity (correctly identified as not) ranges from 83% to 97%**
Conclusion: it's better to trust your own findings than those of an app.

2. **Apps don't pick up every symptom**
Without specialist input, apps **may not recognize rare or unusual cancers**. ↗**false negatives +false sense of security.**

3. **Photographs don't show and tell**
When screening for skin cancer, dermatologists take special dermoscopic images of the skin, using a dermatoscope. Dermoscopic images can unveil e.g. blue-white pigmentation or asymmetries that suggest melanoma. **These clues can hardly be seen in photos (clinical images) alone.** These apps use standard photos taken with a smartphone camera.

4. **No compliance with medical regulations**
Researchers say that skin cancer apps vary in quality and that some have not been tested properly to show that they work and are safe5. **In the US, the app needs to be cleared by the FDA.**

5. **Apps can cause anxiety**
As skin cancer apps have a moderate-to-high sensitivity but only moderate specificity, they might increase the risk of **unnecessary removal of pigmented skin lesions** and create more dermatologist visit→ harmful and expensive to society.

https://www.barco.com/en/news/2019-05-23-skin-cancer-apps-for-diagnosing-melanoma
https://www.medicalnewstoday.com/articles/285751.php

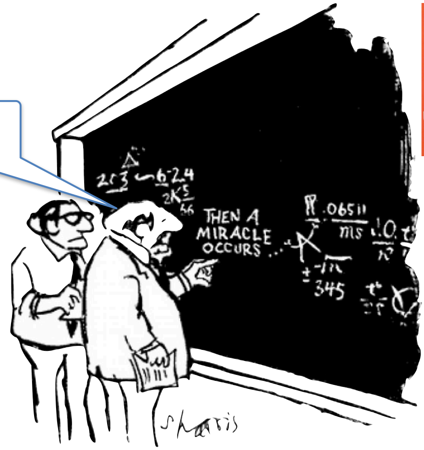
January 23rd 2020

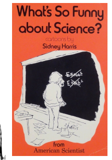
By Jonny Evans, Computerworld | JAN 23, 2020 5:22 AM PST
NEWS
Apple says it's healthy to be skeptical about digital health
Apple's vice president for health, Dr. Sumbul Desai, says it's important to question what digital health solutions can do.

Machine Learning and Interpretability: why bother ?

- Model misuse,
- Model ethics,
- Model bias,
- Model regulatory requirements
- Model trust
- Model understanding
- …

« I think you should be more explicit here in step two. »



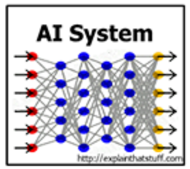


Why making AIs fair, accountable and transparent is crucial

- In October 2017, lawsuit of American teachers with their school district → **computer program that assessed their performance.**
- The system **rated teachers in Houston** by comparing their students' test scores against state averages.
  high ratings → **won praise and even bonuses**. poor ratings→ **faced the sack**.
- **No way of checking if the program was fair or faulty**: *the company that built the software, the SAS Institute, regards its algorithm a trade secret and would not disclose its workings.*
- A federal judge **ruled** that use of the EVAAS (Educational Value Added Assessment System) program may **violate their civil rights.**
- → the school district **paid** the teachers' fees & **stop using the software.**

https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial

The need for Explainable Artificial Intelligence (XAI)



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

DoD and non-DoD Applications

Transportation
Security
Medicine
Finance
Legal
Military

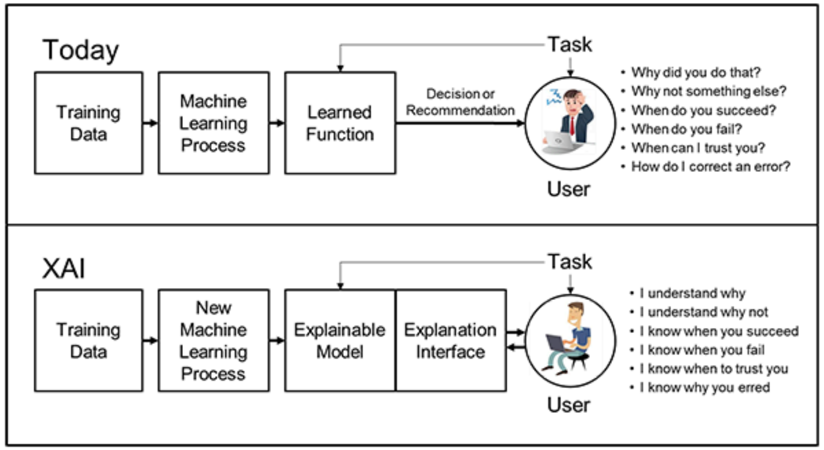
Why did you do that?
Why not something else?
When do you succeed?
When do you fail?
When can I trust you?
How do I correct an error?

The Explainable AI (XAI) program aims to create a suite of machine learning techniques that:

- **Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and**
- **Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.**

https://www.darpa.mil/program/explainable-artificial-intelligence

XAI Concept



https://www.xpowerpoint.com/explainable-artificial-intelligence-xai-darpa--PPT.html

https://www.darpa.mil/program/explainable-artificial-intelligence

Première vue: utilisons les boites noires pour ce qu'elles sont…

ARTIFICIAL INTELLIGENCE

In defense of the black box

Black box algorithms can be useful in science and engineering

By **Elizabeth A. Holm**

The science fiction writer Douglas Adams imagined the greatest computer ever built, Deep Thought, programmed to answer the deepest question ever asked: the Great Question of Life, the Universe, and Everything. After 7.5 million years of processing, Deep Thought revealed its answer: Forty-two (*1*). As artificial intelligence (AI) systems enter every sector of human endeavor—including science, engineering, and health—humanity is confronted by the same conundrum that Adams encapsulated so succinctly: What good is knowing the answer when it is unclear why it is the answer? What good is a black box?

In an informal survey of my colleagues in the physical sciences and engineering, the top reason for not using AI methods such as deep learning, voiced by a substantial majority, was that they did not know how to interpret the results. This is an important objection, with implications that range from practical to ethical to legal (*2*). The goal of scientists and the responsibility of engineers is not just to predict what happens but to understand why it happens. Both an engineer and an AI system may learn to predict whether a bridge will collapse. But only the engineer can explain that decision in terms of physical models that can be communicated to and evaluated by others. Whose bridge would you rather cross?

5 APRIL 2019 • VOL 364 ISSUE 6435
sciencemag.org SCIENCE

- **Nous ne pouvons pas utiliser les boîtes noires en IA pour trouver des liens de causalité, ou de compréhension.**
- **Cette tache est pour l'intelligence humaine et l'IA interprétable.**
- **Mais acceptons les boîte noires en ce qu'elles fournissent une valeur prédictive, qu'elles fournissent d'excellents résultats et …**

Rétinopathie diabétique

Deuxième vue: n'utilisons pas les boites noires pour la santé et la justice !

PERSPECTIVE | nature machine intelligence

https://doi.org/10.1038/s42256-019-0048-x

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

NATURE MACHINE INTELLIGENCE | VOL 1 | MAY 2019 | 206–215 | www.nature.com/natmachintell

Prise de décisions à enjeux élevés **santé**, justice pénale, etc.
La voie à suivre est de concevoir des modèles qui sont intrinsèquement interprétables

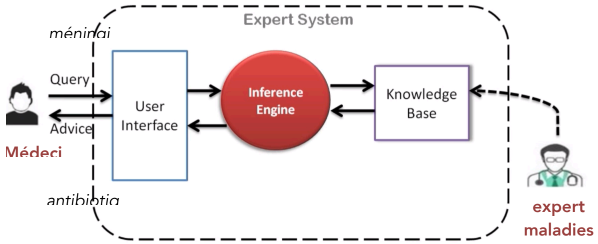
Today: understanding the SOA and issues

Plan

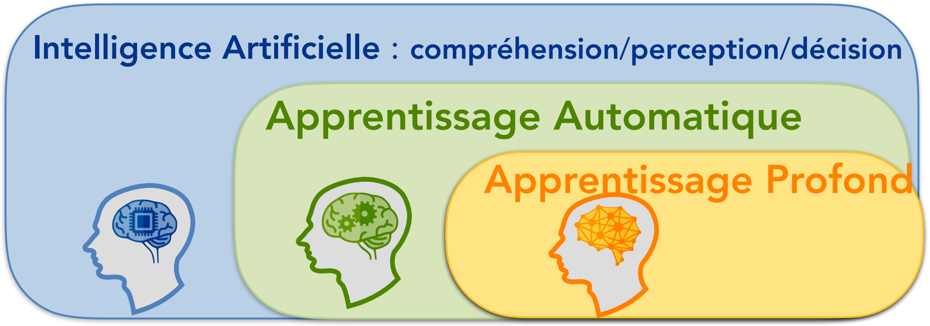
IA et médecine… une longue histoire

IA : compréhension/perception/décision

Le système expert MYCIN (1970) Watson for Oncology (2013)



…mais l'IA et les données massives



… **transforme la médecine… c'est déjà presque une vielle nouvelle !**

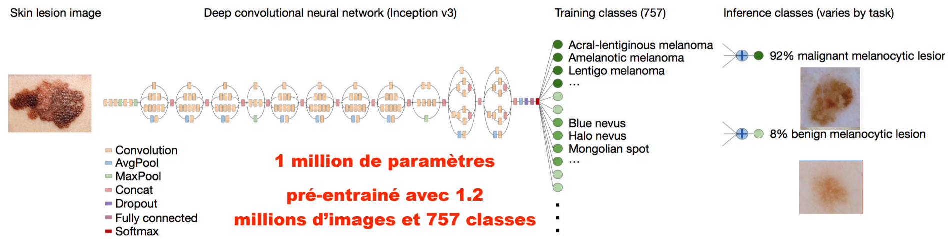
3 predictions in **2016** on Machine Learning as a disruptive technology for Medicine in the next few years.

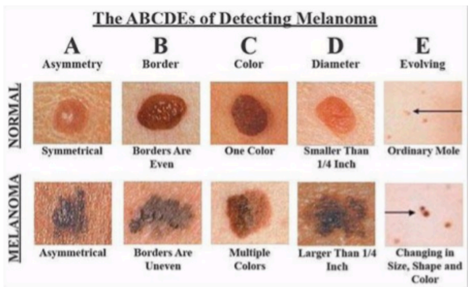
Obermeyer, Z. & Emanuel, E.J., **2016**. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine, 375(13), pp.1216–1219.

- First, ML will **dramatically improve prognosis**. We can precisely identify large patient subgroups with mortality rates approaching 100% and others with rates as low as 10%.
  prediction ➞ come into use in the next 5 years.
- Second, ML will **displace much of the work of radiologists and anatomical pathologists**. Algorithms will also monitor and interpret streaming physiological data, replacing aspects of anesthesiology and critical care.
  prediction ➞ disruptions is within next years, not decades.
- Third, ML will **improve diagnostic accuracy**.
  Obstacles: a) gold standard for diagnosis unclear ➞ harder to train algorithms. b) high-value EHR data are often stored in unstructured formats c) models need to be built and validated individually for each diagnosis.

  prediction ➞ to develop, over the next decade.

Classification des cancer de la peau du niveau d'un expert dermatologue (Nature,2017)

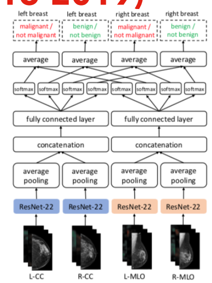




A. Esteva, et al.,"Dermatologist-level classification of skin cancer with deep neural networks," Nature, 2017.

Les performances des « IA » dépassent régulièrement celles des radiologues et anatho-pathologistes (2016-2019)

- diagnostiquer les cancers du sein mieux que les radiologues (Nan Wu, et al. 2019). Trained and evaluated on over 200,000 exams (over **1,000,000 images**). AUC of 0.895/



- diagnostiquer la rétinopathie diabétique comme les **ophtalmologistes** (Gulshan,JAMA,2016.) **128 000** images

Caveat

- To avoid « **disillusionment** » a stronger appreciation of the **technology's capabilities** and limitations is needed.
- **Combining** machine-learning software with the **best human clinician "hardware"** will permit delivery of care that outperforms what either can do alone.

Chen, J. H. & Asch, NEJM 376 (2017).

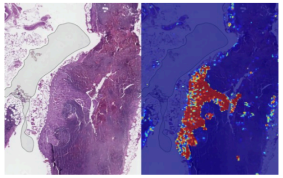
Des systèmes basés sur une collaboration homme-machine peuvent faire mieux que l'IA seule…

INTELLIGENCE ARTIFICIELLE SCIENCE

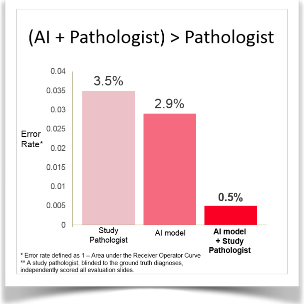
Google détecte le cancer du sein métastatique avec une précision de 99%

Google a largement investi dans le développement des applications de santé.

PAR VALENTIN CIMINO - @CIMINIX - 15 OCTOBRE 2018



https://siecledigital.fr/2018/10/15/une-ia-by-google-detecte-le-cancer-avec-une-precision-de-99/



INTELLIGENCE ARTIFICIELLE

Google a créé une intelligence artificielle capable de détecter le cancer du poumon

Les chercheurs de Google et l'hôpital universitaire de Northwestern se sont unis pour élaborer une IA capable de détecter le cancer du poumon

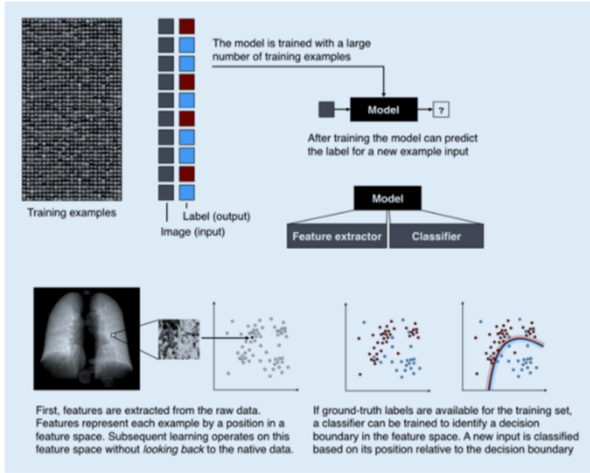
PAR CAMILLE ZAGHET - 21 MAI 2019

Deep Learning Drops Error Rate for Breast Cancer Diagnoses by 85%

JAMA, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.

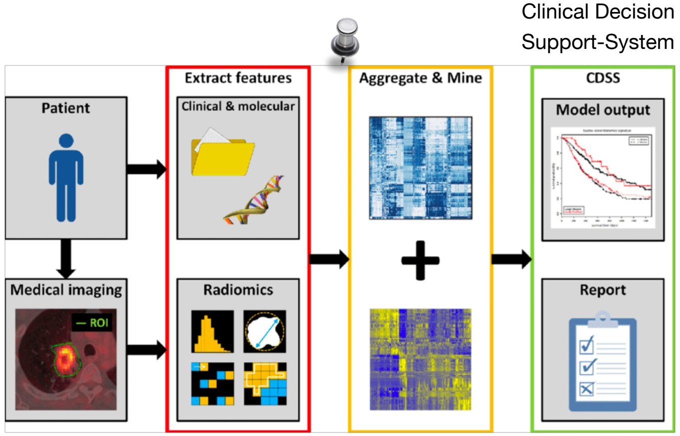
The explosion of medical imaging data creates an environment ideal for machine-learning and data-based science (1/2)

Radiomics, the high-throughput mining of quantitative image features from standard-of-care medical imaging that enables data to be extracted and applied within clinical-decision support systems (CDSS) to improve **diagnostic, prognostic, and predictive accuracy**, is gaining importance in **cancer research.**



S. Röhrich,"Machine learning: from radiomics to discovery and routine," 2018.

The explosion of medical imaging data creates an environment ideal for machine-learning and data-based science (2/2)



P. Lambin, et al., "Radiomics: the bridge between medical imaging and personalized medicine," Nature, 14(12), 2017.

L'approbation des usages médicaux de l'IA est en marche… forte des performances en prédiction…

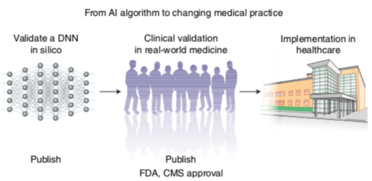


Table 1 | Peer-reviewed publications of AI algorithms compared with doctors

Specialty	Images	Publication
Radiology/ neurology	CT head, acute neurological events	Titano et al.[27]
Pathology	Breast cancer	Ehteshami Bejnordi et al.[41]
	Lung cancer (+driver mutation)	Coudray et al.[33]
	Brain tumors (+methylation)	Capper et al.[45]
Dermatology	Skin cancers	Esteva et al.[47]
	Melanoma	Haenssle et al.[48]
	Skin lesions	Han et al.[49]
Ophthalmology	Diabetic retinopathy	Gulshan et al.[51]
Gastroenterology	Polyps at colonoscopy*	Mori et al.[36]
	Polyps at colonoscopy	Wang et al.[37]
Cardiology	Echocardiography	Madani et al.[23]
	Echocardiography	Zhang et al.[24]

Table 2 | FDA AI approvals are accelerating

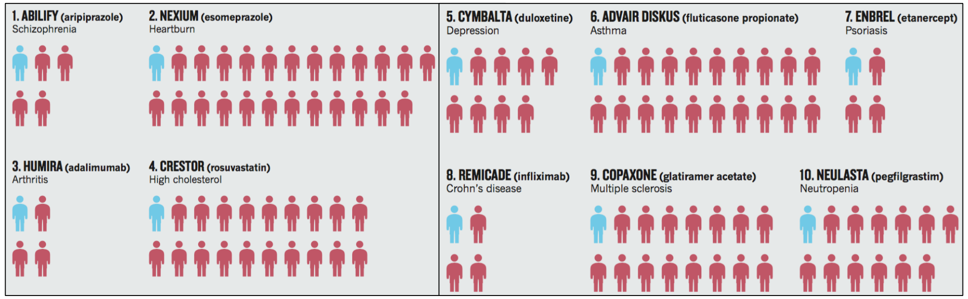
Company	FDA Approval	Indication
Apple	September 2018	Atrial fibrillation detection
Aidoc	August 2018	CT brain bleed diagnosis
iCAD	August 2018	Breast density via mammography
Zebra Medical	July 2018	Coronary calcium scoring
Bay Labs	June 2018	Echocardiogram EF determination
Neural Analytics	May 2018	Device for paramedic stroke diagnosis
IDx	April 2018	Diabetic retinopathy diagnosis
Icometrix	April 2018	MRI brain interpretation
Imagen	March 2018	X-ray wrist fracture diagnosis
Viz.ai	February 2018	CT stroke diagnosis
Arterys	February 2018	Liver and lung cancer (MRI, CT) diagnosis
MaxQ-AI	January 2018	CT brain bleed diagnosis
Alivecor	November 2017	Atrial fibrillation detection via Apple Watch
Arterys	January 2017	MRI heart interpretation

E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nat Med, pp. 1–13, Jan. 2019.

Plan

Malgré ses succès la médecine reste « imprécise »…

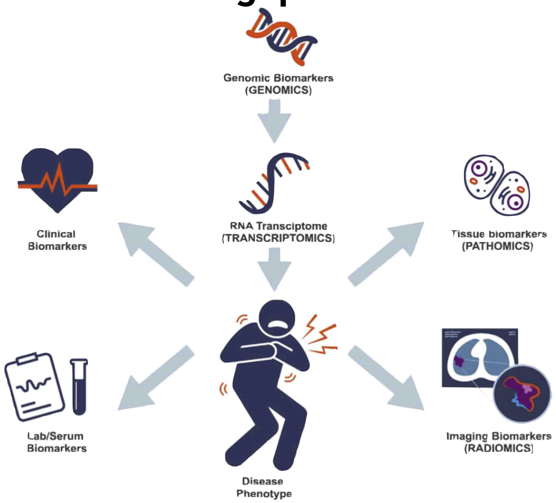
Pour **chaque** personne qu'ils aident (bleu), les dix médicaments les plus lucratifs aux États-Unis ne parviennent pas à améliorer les conditions d'entre **3** et **24** personnes (rouge).



Schork, N.J., 2015. Personalized medicine: time for one-person trials. Nature.

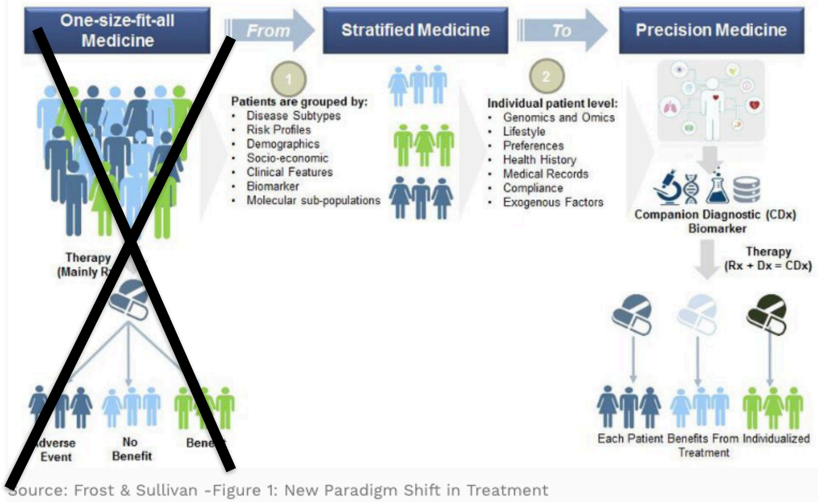
Emergence de la médecine de précision…

→ fournir les meilleurs soins disponibles à chaque patient, sur la base d'une stratification en sous-classes de maladies présentant une base biologique commune.



Shaikh et al., "Translational Radiomics: Defining the Strategy Pipeline and Considerations for Application-Part 1: From Methodology to Clinical Implementation," JACR, 2018.

… et la fin de la médecine « one-size-fit all »



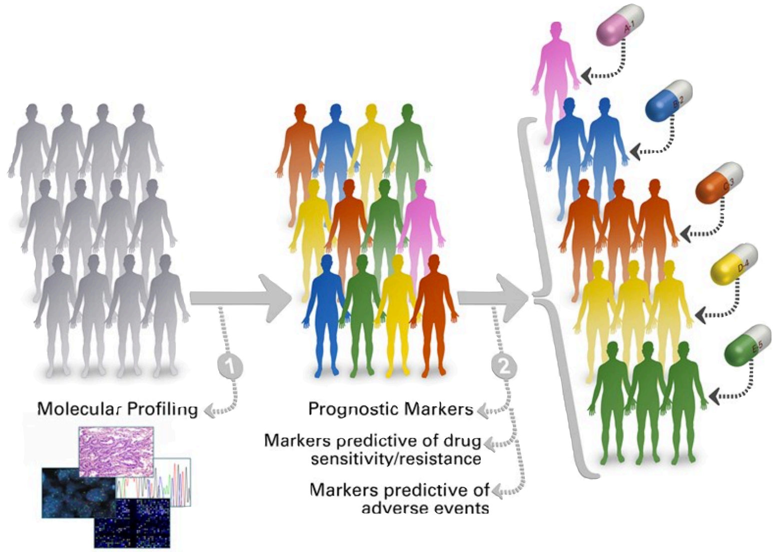
→ la bonne intervention au bon patient au bon moment.

Médecine personnalisée vs. de précision

Médecine personnalisée	→ dédié à 1 patient
Médecine de précision	→ stratification fine des patients
Médecine ciblée	→ spécifique à une cible thérap.
Médecine translationelle	→ boucle R&D : Bed2Bench2Bed

Feldman, A. M. (2015). Bench-to-Bedside; Clinical and Translational Research; Personalized Medicine; **Precision Medicine-What's in a Name?** Clinical and Translational Science, 8(3), 171–173. http://doi.org/10.1111/cts.12302

Médecine de précision pour le cancer

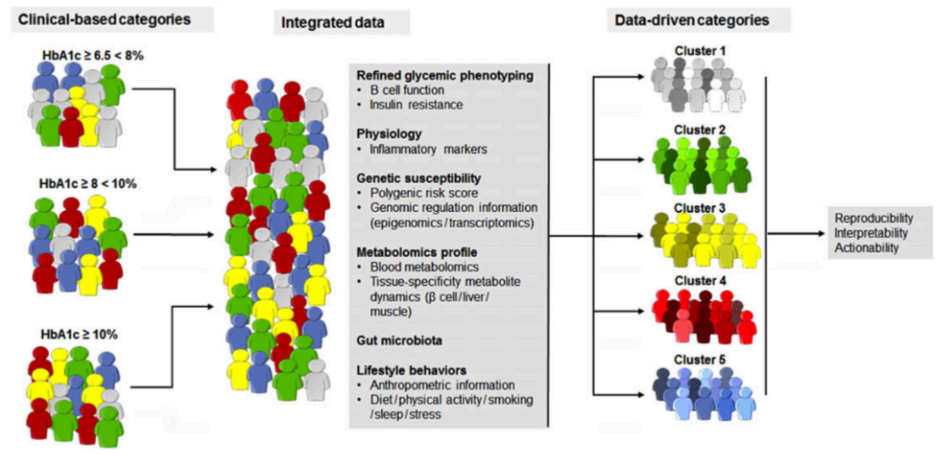


https://pct.mdanderson.org/

Other chronic disease are strongly multi-factorial : Cardio-metabolic diseases (CMD)

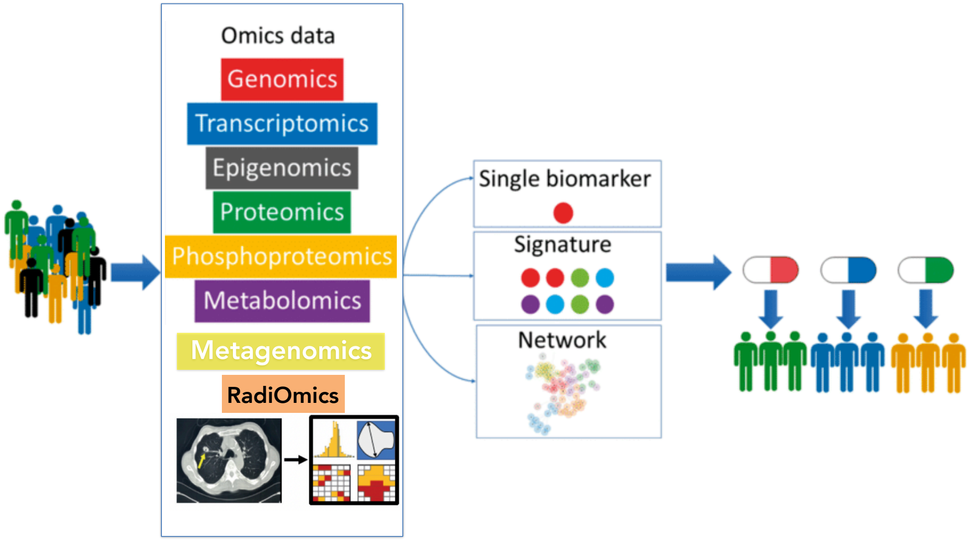
- Overweight = BMI > 25 and Obesity = BMI > 30 (BMI=Weight/Height^2)
- Obesity is a **chronic disease of pandemic evolution** → increased risk of many pathologies (cardiometabolic) pathologies (dyslipidemias, T2 diabetes, arterial hypertension) and articular depression and many cancers.
- World Prevalence of overweight **or** obese is **37%** for men and **38%** for women.
- In France, 2012 overweight **or** obese ~ half of population (Obese 15%~6.9 **millions**).
- **In Africa, diabetes (5.7% of the adult population in Africa is now affected) and** [cardiovascular diseases kill more than AIDS.]()
- **How to improve treatments ?**

Médecine de précision pour le diabète



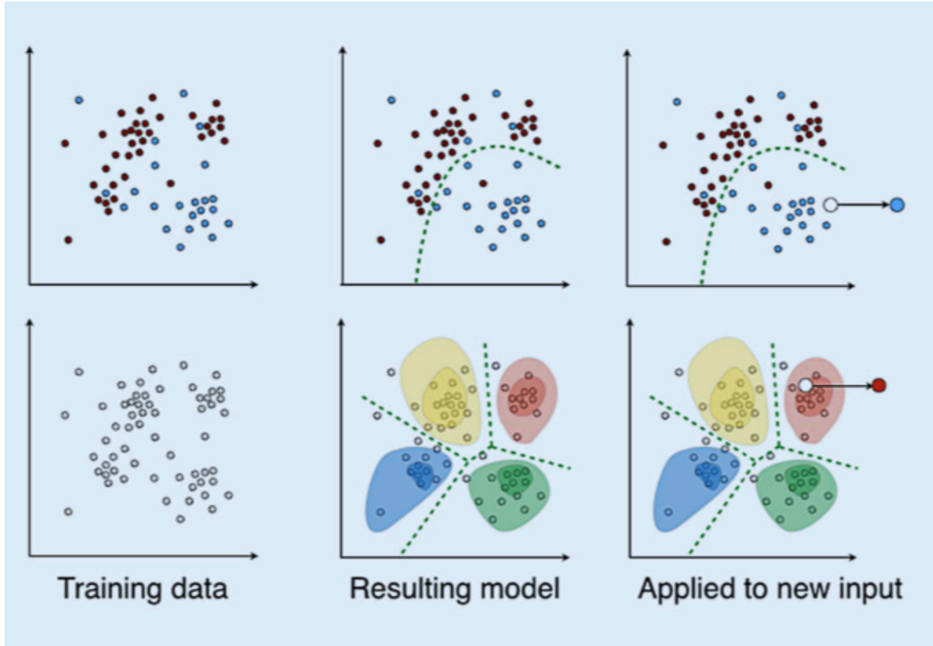
J. Merino and J. C. Florez, "Precision medicine in diabetes: an opportunity for clinical translation," Ann. N.Y. Acad. Sci., vol. 1411, no. 1, pp. 140–152, Jan. 2018.

On détermine les meilleures options thérapeutiques en fonction des caractéristiques biologiques et génétiques d'une personne.



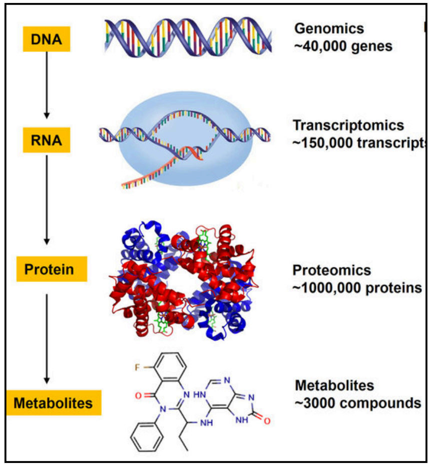
Adapté de G. Giudice and E. Petsalaki, "Proteomics and phosphoproteomics in precision medicine: applications and challenges," Brief Bioinformatics, vol. 1, no. 2, pp. 129–12, Oct. 2017.

Médecine de précision et apprentissage automatique

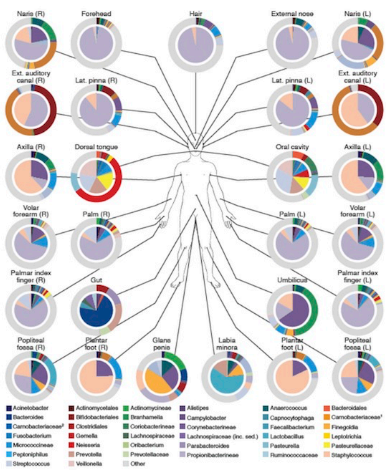


Les données « Omics » permettent de nous caractériser très finement, nous et… nos hôtes.

Analyser nos propre cellules



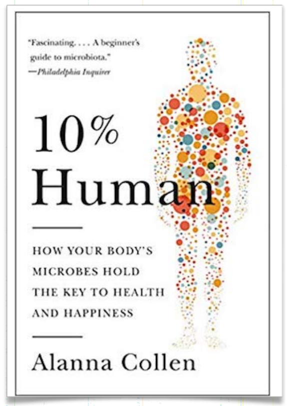
Analyser nos bactéries

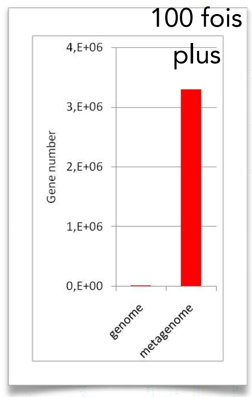


F. S. MD, et al., "Translational Radiomics: Defining the Strategy Pipeline and Considerations for Application-Part 1: From Methodology to Clinical Implementation," Journal of the American College of Radiology, 2018

Microbiote intestinal humain : un organe oublié

Du bébé "stérile" à la naissance → 2 kg de micro-organismes, sur les 100 billions de cellules du corps humain, seule 1 sur 10 est humaine.



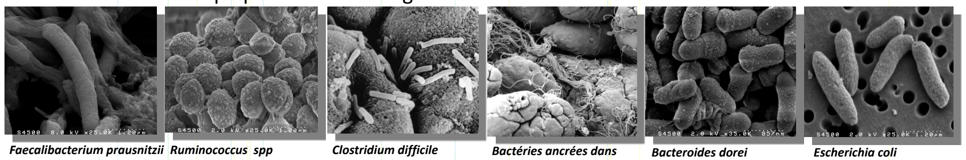


Métabolisme
production de vitamines
dégradation des aliments
extraction énergétique

Système immunitaire
l'"éducation" des défenses immunitaires innées

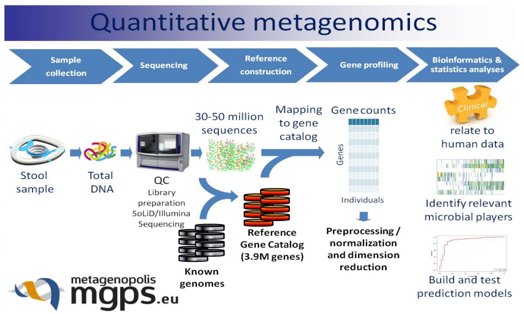
Quantification de notre microbiome

La plupart des micro-organismes sont inconnus et non cultivables…



Faecalibacterium prausnitzii — *Ruminococcus spp* — *Clostridium difficile en caecum souris* — *Bactéries ancrées dans une Plaque de Peyer, Intestin de souris* — *Bacteroides dorei* — *Escherichia coli*

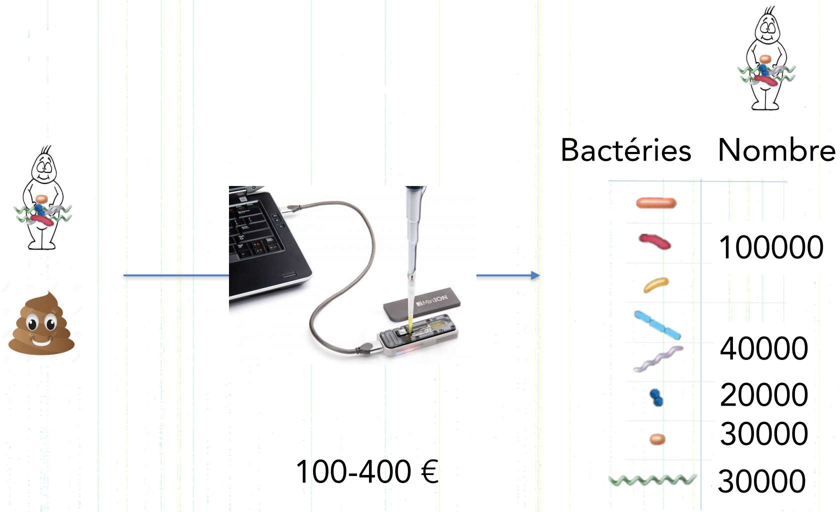
Photos UEPSD



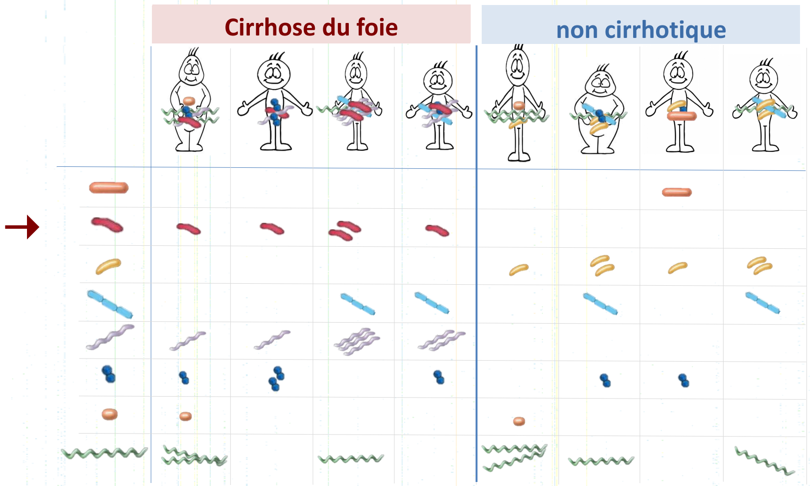
A Powerful Microscope to Scan the neglected organ

Ehrlich SD ©

Quantification de notre microbiome (vision simple)



Vers une médecine de précision des maladies intégrant la métagénomique.



Signature géniques impliquant plusieurs gènes bactériens parmi des **millions**…

For the classification tasks there are metagenomic datasets from the ExperimentHub

Dataset Name	Disease	# features (*species, genus, family, order, class, phylum, whole_tax, marker, pathway*)	# cases	# controls	Average Reads (std) (M)	Type of Task
cirrhosis1	Liver cirrhosis stage 1	462, 151, 52, 22, 15, 9, 1252, 128224, 310	98	83	51.6 (30.9)	classification
cirrhosis2	Liver cirrhosis stage 2	408, 118, 53, 22, 15, 9, 990, 86308, 306	25	31	51.6 (30.9)	classification
ibd	Inflammatory bowel disease	719, 299, 141, 64, 33, 21, 1934, 222837, 427	148	248	53.9 (20.2)	classification
t2dw	Type 2 diabetes	381, 142, 39, 29, 24, 14, 943, 91102, 430	53	43	31.0 (17.6)	classification
t2d	Type 2 diabetes	505, 222, 98, 45, 14, 8, 1463, 131309, 431	170	174	40.2 (11.8)	classification
obesity	Obesity	429, 243, 121, 60, 31, 20, 1365, 128510, 418	167	96	69.0 (23.2)	classification
microbaria	Bariatric surgery	558	24	-	41.83 (19)	regression

INTEGROMICS

State of the art: RF/SVM or linear models

(Pasolli et al., 2016)

		Cirrhosis	Colorectal	IBD	Obesity	T2D	WT2D
	#samples	232	121	110	253	344	96
Species abundance	#species	542	503	443	465	572	381
Species abundance	RF	0.945 (0.036)	0.873 (0.071)	0.890 (0.078)	0.655 (0.079)	0.744 (0.056)	0.762 (0.111)
Species abundance	RF-FS:Emb	**0.946 (0.035)**	**0.881 (0.067)**	**0.893 (0.080)**	**0.656 (0.072)**	**0.745 (0.056)**	**0.772 (0.116)**
Species abundance	SVM	0.922 (0.041)	0.809 (0.086)	0.862 (0.083)	0.648 (0.071)	0.663 (0.066)	0.664 (0.126)
Marker presence	#markers	120553	108034	91756	99568	119792	83456
Marker presence	RF	0.945 (0.032)	0.844 (0.087)	0.905 (0.071)	0.631 (0.076)	0.747 (0.056)	0.739 (0.113)
Marker presence	RF-FS:Emb	0.935 (0.035)	**0.845 (0.085)**	0.884 (0.081)	0.621 (0.080)	0.718 (0.062)	0.681 (0.126)
Marker presence	SVM	**0.963 (0.027)**	0.823 (0.087)	**0.914 (0.084)**	**0.659 (0.073)**	**0.757 (0.056)**	**0.785 (0.104)**

AUC: 0.5 – 1.0

	Bins	Model	Color	CIR	COL	IBD	OBE	T2D	WT2	AVG
MetAML - RF				0.877	0.805	0.809	0.644	0.664	0.703	0.750
Fill-up	SPB	CNN	gray	**0.905**	0.793	**0.868**	**0.680**	0.651	0.705	0.767
Fill-up *	SPB	CNN	grays	**0.903**	0.768	**0.840**	**0.673**	0.666	**0.739**	0.765
TSNE	QTF	CNN	jet	0.870	0.748	0.820	0.654	**0.690**	0.660	0.741

T. H. Nguyen, Y. Chevaleyre, E. Prifti, N. Sokolovska, and J.-D. Zucker, "Deep Learning for Metagenomic Data: using 2D Embeddings and Convolutional Neural Networks," 2017.

Met2Img (Thanh-Hai et al. 2018) outperforms MetAML [Pasolli, 2016] for 5 out of 6 datasets

E. Pasolli, D. T. Truong, F. Malik, L. Waldron & N. Segata; "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights"; PLoS Comput". Biol. 12, p. e1004 977 (2016)

Results with 1D data

Framework	Model	CIR	COL	IBD	OBE	T2D	WT2	AVG
MetAML	RF	0.877	0.805	0.809	0.644	0.664	0.703	0.750
	SVM	0.834	0.743	0.809	0.636	0.613	0.596	0.705
Met2Img	RF	0.877	0.812	0.808	0.645	0.672	0.703	0.753
	SVM- Sigmoid	0.509	0.603	0.775	0.648	0.515	0.553	0.600
	SVM- Radial	0.529	0.603	0.775	0.648	0.593	0.553	0.617
	SVM- Linear	0.766	0.666	0.792	0.612	0.634	0.676	0.691
	FC	0.776	0.685	0.775	**0.656**	0.665	0.607	0.694
	CNN1D	0.775	0.722	**0.842**	**0.663**	0.668	0.618	0.715

Results with synthetic images

Phylogenetic ordering

	Bins	Model	Color	CIR	COL	IBD	OBE	T2D	WT2	AVG
MetAML - RF				0.877	0.805	0.809	0.644	0.664	0.703	0.750
Fill-up	SPB	CNN	gray	**0.905**	0.793	**0.868**	**0.680**	0.651	0.705	0.767
Fill-up *	SPB	CNN	grays	**0.903**	0.768	**0.840**	**0.673**	0.666	**0.739**	0.765
TSNE	QTF	CNN	jet	0.870	0.748	0.820	0.654	**0.690**	0.660	0.741

Random ordering

Precision medicine directed at the microbiota could inform physicians about prognosis and therapy.

CANCER THERAPY

Precision medicine using microbiota

Intestinal microbiota influence cancer patient responses to immunotherapy

Jobin, C. (2018). Precision medicine using microbiota. Science, 359(6371), 32–34.

One could view the microbiota as a treasure trove for next-generation medicine, and tapping into this network may produce new therapeutic insights.

Plan

I. Apprentissage Artificiel et médecine

II. Médecine de précision

III. Pourquoi des modèles interprétables en médecine ?

IV. Machine Learning interpretable trois approches

V. Deux exemples de modèles interprétables

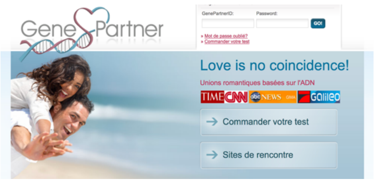
VI. Conclusion

RGPD et modèles interprétables : droit et confiance

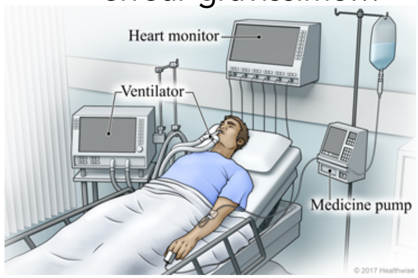
- Règlements de l'UE (Règlement général sur la protection des données (GDPR) en vigueur le 25 mai 2018) sur la prise de décision algorithmique et un "droit d'explication". Goodman, B. & Flaxman, S. R. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI magazine, 2017
- Une explication de la prédiction est désirée par médecins et patients lorsque un **modèle** doit être validé avant d'être déployé en routine→ **confiance**

Vanthienen, et al. Performance of classification models from a user perspective. *Decision Support Systems* **51**, 782- 793,(2011).

erreur pas (trop) grave…



erreur gravissime…



A. Vellido, et al., "Machine learning in critical care: state-of-the-art and a sepsis case study," BMEO,2018.

Equité/Fairness : l'IA est biaisée par les données

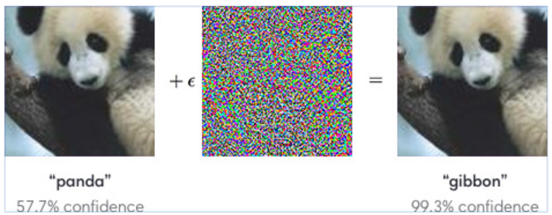
Une étude récente a révélé que certains programmes de reconnaissance faciale **classent incorrectement** **moins de 1 % des hommes à la peau claire**, mais plus d'**un tiers des femmes à la peau foncée**.

Que se passe-t-il lorsque l'on se fie à de tels algorithmes pour diagnostiquer le mélanome sur une peau claire ou foncée… ?

Un programme apprends à partir des données qu'on lui donne et qui peuvent être … biaisées

Les « aversarials attacks » sont maintenant connues, mais..

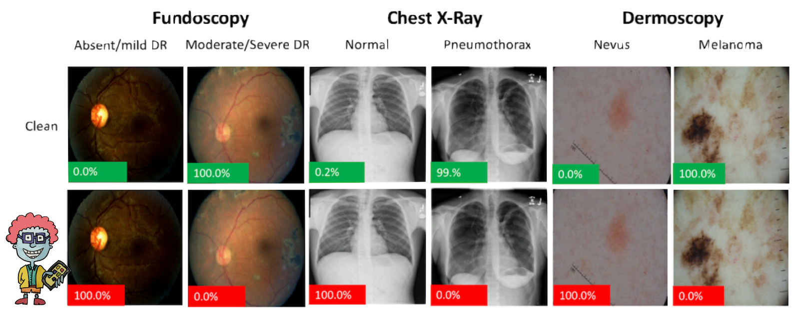
J. Su, et al. "One pixel attack for fooling deep neural networks.," CoRR, 2017.



Akhtar & Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," arXiv.org, 02-Jan-2018.

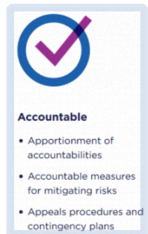
…se pose la problème de la « responsabilité » des algorithmes

… notamment en cas d'attaques d'images médicales.

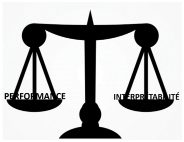


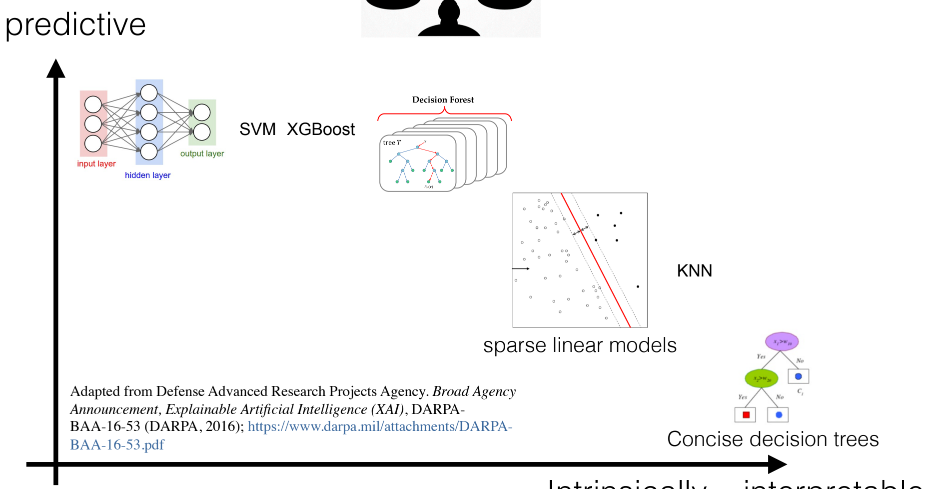
Finlayson, et al., "Adversarial Attacks Against Medical Deep Learning Systems.," arXiv, 2018.

Qui est responsable en cas d'erreur ?



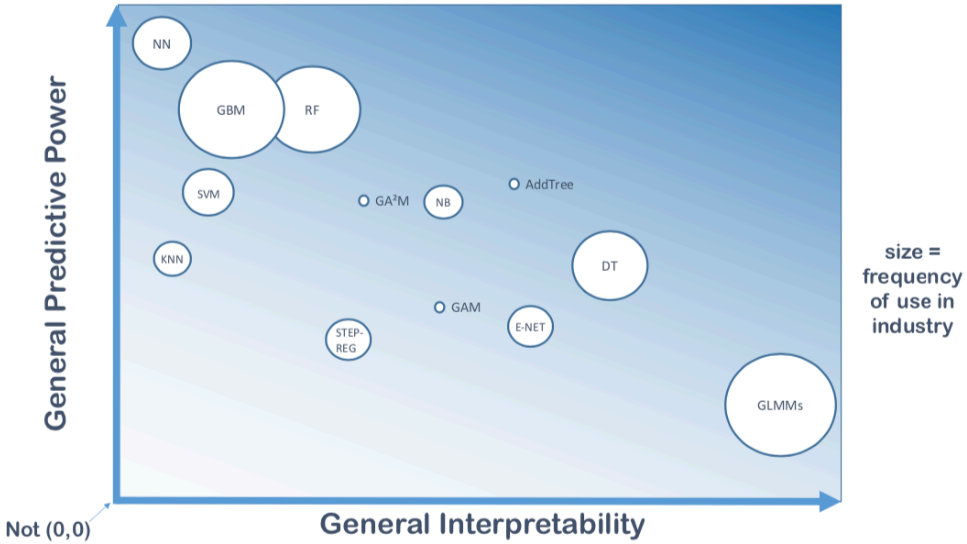
Interpretability vs Predictive power





Adapted from Defense Advanced Research Projects Agency. *Broad Agency Announcement, Explainable Artificial Intelligence (XAI)*, DARPA-BAA-16-53 (DARPA, 2016); https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

Intrepretability/Accuracy and Usage



A Data Science Playbook for xAI - Navigating Interpretable and Predictive Models Josh Poduska, Chief Data Scientist, Domino Data Lab

Plan

The PDR Framework : 3 desiderata should be used to select interpretation methods

☑ **Predictive accuracy** : the *quality of a model's fit* measured with test-set accuracy (the data used to check for predictive accuracy must resemble the population of interest, distribution of predictions matters,…)

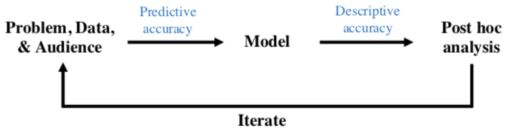
☑ **Descriptive accuracy**: the *degree* to which an interpretation method *objectively captures the relationships learned* by machine-learning models.

☑ **Relevancy** : an interpretation that *provides insight for a particular audience* into a chosen domain problem

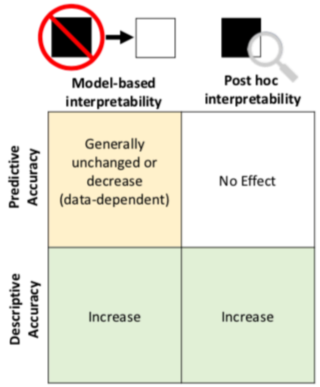
W. J. Murdoch, C. Singh, K. K. P. O. the, 2019, "Definitions, methods, and applications in interpretable machine learning," *PNAS*

Interpretability in Machine Learning concepts

Predictive and Descriptive accuracy



Impact of interpretability methods on descriptive and predictive accuracies.

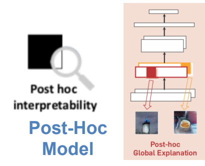


W. J. Murdoch, C. Singh, K. K. P. O. the, 2019, "Definitions, methods, and applications in interpretable machine learning," *National Acad Sciences*

Interpretability in Machine Learning

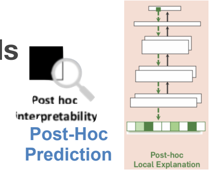
Type A - Interpreting black-box models

What was learned and is hidden in the model ?



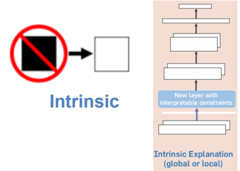
Type B - Interpreting predictions from black-box models

Why this individual has been classified this way ?



Type C - Learning interpretable models

How do we intrinsically explain the model ?

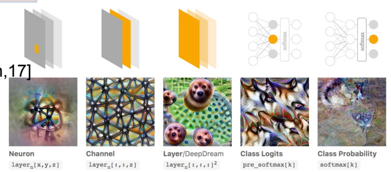


Interpretability in Machine Learning

Type A - Interpreting black-box models

Looking into the black box

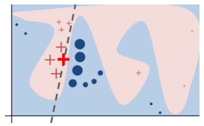
Model distillation (soft DT)[Frosst&Hinton,17]



[Olah, distill.pub]

Type B - Interpreting predictions from black-box models

Attribution methods: e.g. LIME



"Why Should I Trust You?" Explaining the Predictions of Any Classifier [Ribeiro et al. '16]

Type C - Learning interpretable models

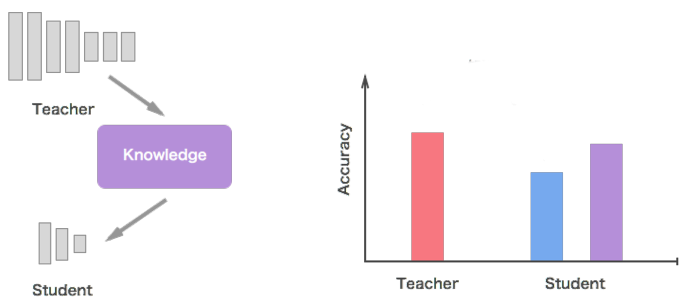
Decision tree, Rules, linear model, scoring model, …

Looking into the black box: A detail view of an activation atlas from one of the layers of the InceptionV1 vision classification network.

It reveals many of the visual detectors that the network uses to classify images, such as different types of fruit-like textures, honeycomb patterns and fabric-like textures.

https://ai.googleblog.com/2019/03/exploring-neural-networks.html

Model Distillation: Principles

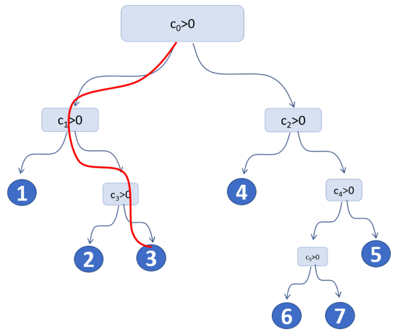


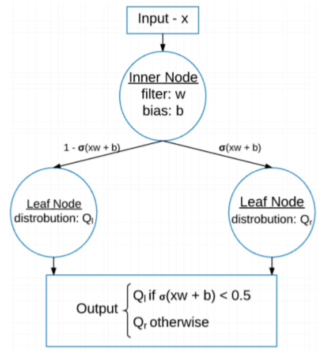
Student = Soft Decision Tree→ for explaining a particular classification decision on a particular test case

Student = Smaller Network→ for improving the performance of deep learning models on mobile devices

[PDF] Distilling the Knowledge in a Neural Network - University of ...
https://www.cs.toronto.edu › ~hinton › absps › distillation ▾
by G Hinton - 2015 - Cited by 2800 - Related articles
arXiv:1503.02531v1 [stat.ML] 9 Mar 2015. **Distilling** the Knowledge in a Neural Network.

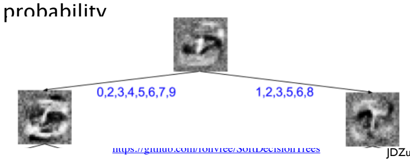
A Soft Decision Tree





Basic Tree. Each data–point travels through the tree until one of the leafs. **The path is determined by the split conditions**, which are functions of the features. The leafs determine the prediction target.

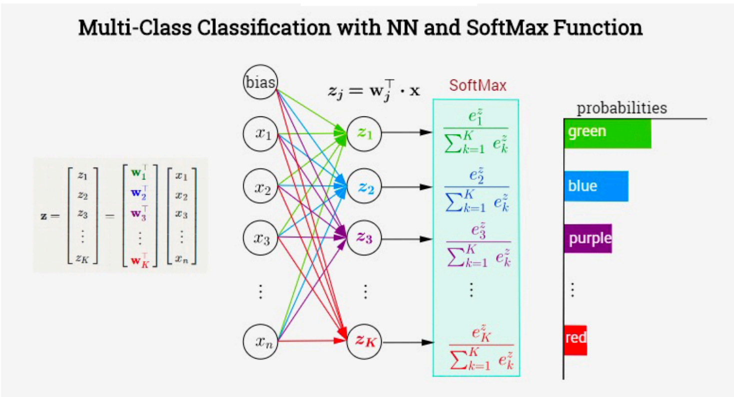
Soft Decision Tree. Each data point does not have a unique path through the tree. They now belong to every leaf of the tree, with a certain probability, i.e. the path probability



The distillation approach: back to the softmax

$$a_i = \frac{e^{z_i}}{\sum_{k=1}^{c} e^{z_k}}$$

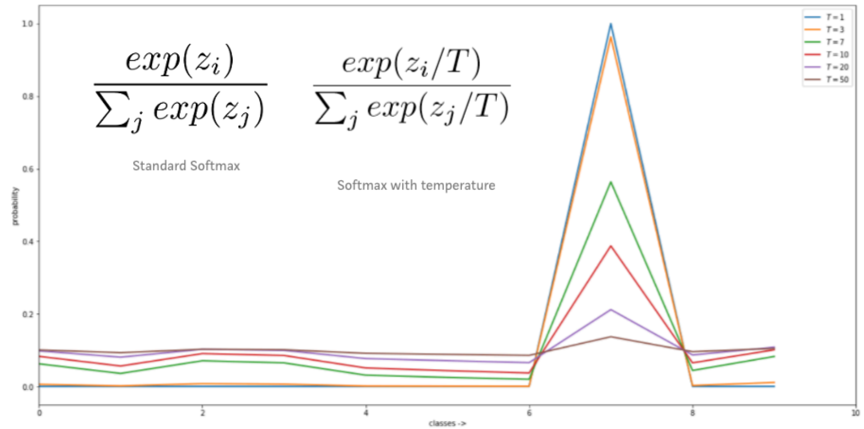
$$\text{where } \sum_{i=1}^{c} a_i = 1$$



In practice, the model will output « green » but cannot say like 'red' is much closer to 'green'.This is because the target output class will have high probability and all other classes will have probability closer to zero

http://www.adeveloperdiary.com/data-science/deep-learning/neural-network-with-softmax-in-python/

Detecting the « dark knowledge »



https://medium.com/analytics-vidhya/knowledge-distillation-dark-knowledge-of-neural-network-9c1dfb418e6a

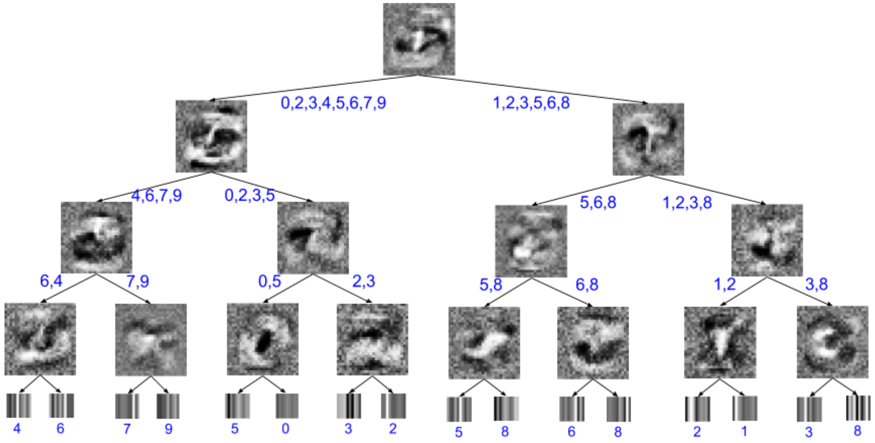
Predictive probabilities with different values of temperature T

To extract this dark knowledge we used ensemble of models in practice. So we turned into knowledge distillation where a complex model (Teacher model) will be used to distill its knowledge to the small model (Student model) .The student model can be as complex as teacher model or lesser. In practice we use less complex model as student model.

https://medium.com/@ahmdtaha/distilling-the-knowledge-in-a-neural-network-77b7232bb631

Distilling a Neural Network Into a Soft Decision Tree

The images at the inner nodes are the learned filters
The images at the leaves are visualizations of the learned probability distribution over classes



A type of soft decision tree that generalizes better than one learned directly from the training data of NMIST

[Frosst and Hinton]

The soft decision tree trained improves accuracy

It reaches a test accuracy of 96.76% which is about halfway between the neural net and the soft decision tree trained directly on the data.

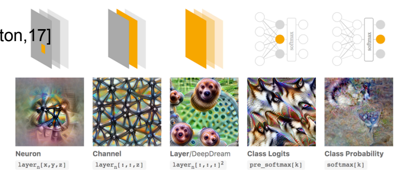
Dataset	Accuracy		
	SDT with true targets	Neural Network	SDT with soft targets
MNIST	94.45%	99.21%	96.76%
Connect4	78.63%	NA	80.60%
Letter	78%	95.9%	81%

https://medium.com/razorthink-ai/distilling-a-neural-network-into-a-soft-decision-tree-1d1818dc1c4f

Interpretability in Machine Learning

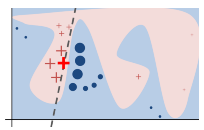
Type A - Interpreting black-box models

Model distillation (soft DT) [Frosst&Hinton,17]
Looking into the black box



[Olah, distill.pub]

Type B - Interpreting predictions from black-box models
- **Activation Maps**
- **Attribution methods: e.g. LIME**
- **Feature relationships**
- **Feature importance scores**



"Why Should I Trust You?" Explaining the Predictions of Any Classifier [Ribeiro et al. '16]

Type C - Learning interpretable models
Decision tree, Rules, linear model, scoring model, …

Interpretability in Machine Learning

Type B - Interpreting predictions from black-box models

Classification of the methods:

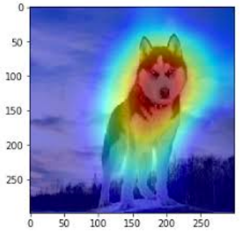
Global (whole dataset) vs. Local (one instance) methods

Model-Agnostic (any learner) or Model-specific methods

Class Activation maps :
localy interpretable & Model-specific Explanations

Allows to spot the region where neurons are particularly activated when fed with a specific input image.

InceptionV3 model.



The red region represents the area of the image on which the network focuses to class

https://jacobgil.github.io/deeplearning/class-activation-maps

https://edebrouwer.github.io/deeplearning/carvision/visualization/neural/networks/learning/2017/08/09/Deep_Visualization.html

« Attention Maps » for medicine: Single retinal fundus image and different classes predicted (age,gender, smoking, HbA1C, BMI)

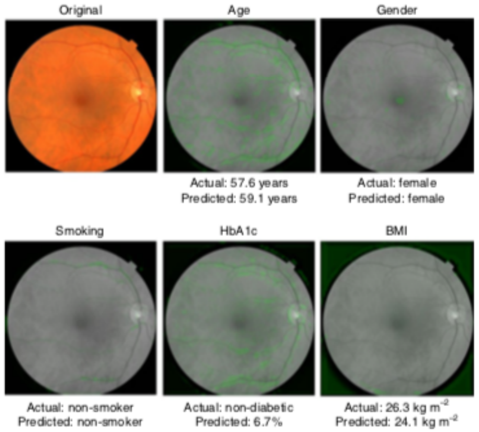


Table 6 | Percentage of the 100 attention heat maps for which doctors agreed that the heat map highlighted the given feature

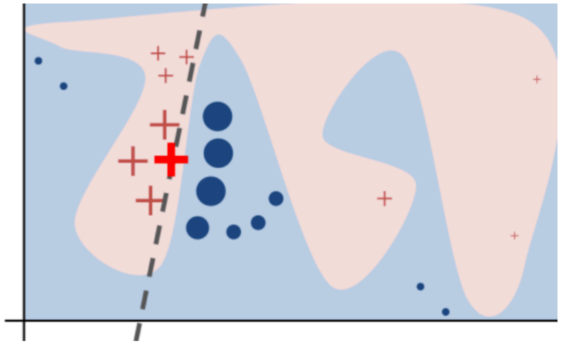
Risk factor	Vessels (%)	Optic disc (%)	Non-specific features
Age	95	33	38
Gender	71	78	50
Current smoker	91	25	38
HbA1c	78	32	46
SBP	98	14	54
DBP	29	5	97
BMI	1	6	99

Heat maps (n= 100) were generated for each risk factor and then presented to three ophthalmologists who were asked to check the features highlighted in each image (n=300 responses for each risk factor). The images were shuffled and presented as a set of 700, and the ophthalmologists were blinded to the output prediction of the heat maps and the ground-truth label. For the variables that were present in both datasets (age and gender), the most commonly highlighted features were identical in both datasets.

The top left image is a sample retinal image in colour from the UK Biobank dataset. The remaining images show the same retinal image, but in black and white.

LIME :
Local Interpretable Model-agnostic Explanation

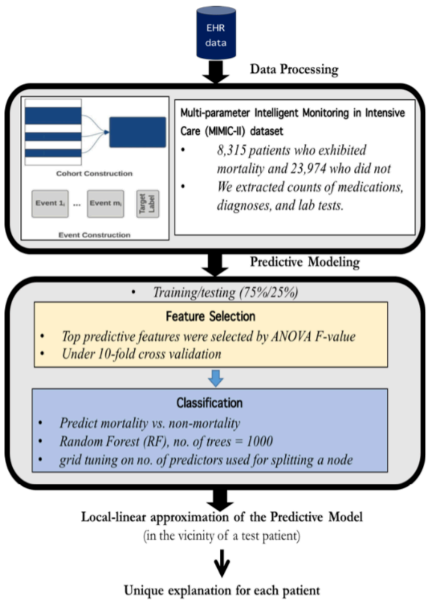
Objectif: convertir les prédictions en un modèle **interprétable : séparateur linéaire.**



Le graphique représente les zones possibles de prédiction en rouge et bleue, la croix rouge en gras et la prédiction initiale, les axes représentent des variables les autres points (rond bleu ou croix rouge) sont les **prédictions obtenues après modification des valeurs des variable**s.
Par exemple, un point situé à droite de la prédiction originale aura été modifiée uniquement sur la variable qui correspond à l'axe des abscisses.
Enfin plus un point possède une grande taille, plus il est "proche" (en distance) du point initial.

M. T. Ribeiro, S. Singh, and C. Guestrin, **"'Why Should I Trust You?',"** presented at the the 22nd ACM SIGKDD International Conference, New York, New York, USA, 2016, pp. 1135–1144.

LIME for Precision Medicine (ICU)



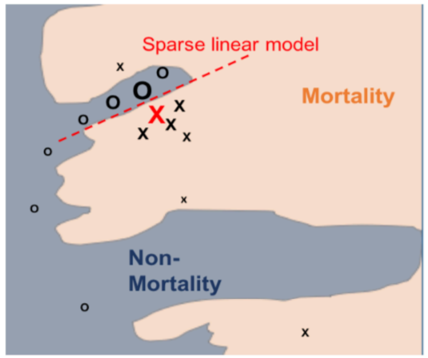


Figure 2: Non-linear decision function of the complex predictive model is represented by the orange/blue background. The red cross is the test patient being explained (let's call it X). Perturbed instances around X weighted by their proximity to X are fed into the model. A sparse linear model (red dashed line) is fitted for the model's prediction on these perturbed instances. This linear model approximates the non-linear decision function of the predictive model, locally in the neighborhood of X.

G. J. Katuwal and R. Chen, "Machine Learning Model Interpretability for Precision Medicine," arXiv.org, vol. q-bio.QM. 28-Oct-2016.

… to answer to the Why question ?

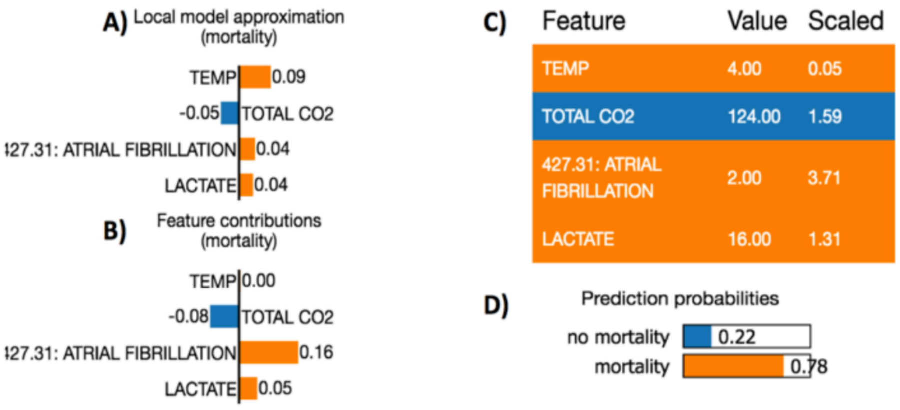


Figure 3: Patient specific model interpretation. A) Local model approximation in the vicinity of the patient: correlation of the features to mortality. *Temperature, atrial fibrillation, and lactate level are positively correlated with mortality*. B) Feature contributions for prediction. *Higher counts of atrial fibrillation and higher lactate level contribute towards mortality of this particular patient.* C) Value: original value for each feature and Scaled: scaled value, D) Class prediction probabilities. *The Random Forest model predicts 78% mortality for this particular test patient.*

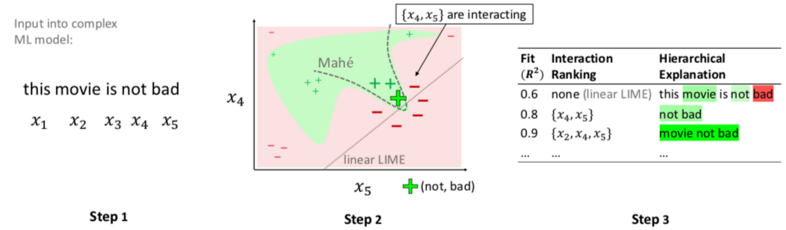
Beyond « Why Should I Trust You? »… « Can I trust you more » ?

MAHE Model-Agnostic Hierarchical Explanation

☑ Interactions such as double negation in sentences and scene interactions in images are common forms of complex dependencies captured by state-of-the-art machine learning models.

☑ MAHE explains how powerful machine learning models capture these interactions

☑ MAHE fits a neural network to learn the highly nonlinear decision boundary used to classify the instance.



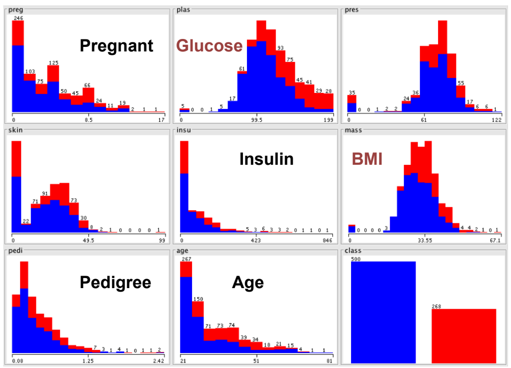
☑ Attribution scores of those interactions can then be shown for the data instance, as displayed in Step 3
→ The film is positively rated (green) in spite of the word bad being there which is explained by the interaction « not bad »

M. Tsang, Y. Sun, D. Ren, and Y. L. 0002, "Can I trust you more? Model-Agnostic Hierarchical Explanations.," arXiv, vol. stat.ML, 2018.

Partial Dependancy Plots : they show the marginal effect of values of one or two variables

Feature label	Variable type	Range
Number of times pregnant	Integer	0–17
Plasma glucose concentration in a 2 h oral glucose tolerance test	Real	0–199
Diastolic blood pressure	Real	0–122
Triceps skin fold thickness	Real	0–99
2 h serum insulin	Real	0–846
Body mass index	Real	0–67.1
Diabetes pedigree function	Real	0.078–2.42
Age	Integer	21–81
Class	Binary	Tested positive for diabetes = 1

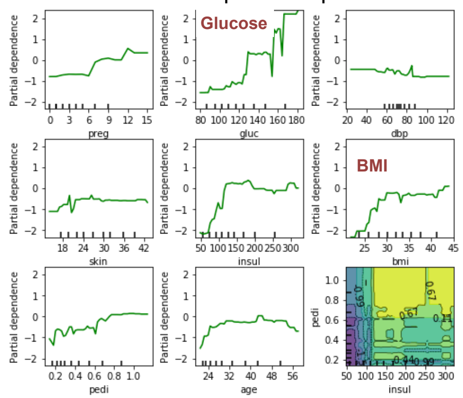
https://data.world/data-society/pima-indians-diabetes-database



Partial Dependancy Plots : they show the marginal effect of values of one or more variables

☑ If you are familiar with linear or logistic regression models, partial dependence plots can be interpreted similarly to the coefficients in those models.
☑ But partial dependence plots can capture more complex patterns from your data, and they can be used with any model.

Y axis: « diabetes partial dependance»

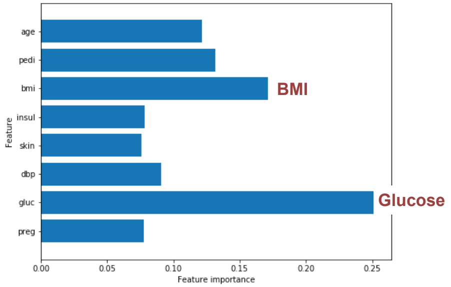


https://briangriner.github.io/Partial_Dependence_Plots_presentation-BrianGriner-PrincetonPublicLibrary-4.14.18-updated-4.22.18.html

Variable Importance: Global, Model-Agnostic or not

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way.

- ☑ To measure the importance of the ith feature after training, the values of the i-th feature are permuted among the training data **and the out-of-bag error is again computed on this perturbed data set**.
- ☑ The importance score for the i-th feature is computed by averaging **the difference in out-of-bag error before and after the permutation over all trees**.
- ☑ The score is normalized by the standard deviation of these differences.



h

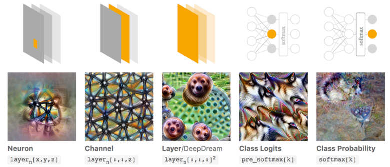
Code Python https://briangriner.github.io/Partial_Dependence_Plots_presentation-BrianGriner-PrincetonPublicLibrary-4.14.18-updated-4.22.18.html

Interpretability in Machine Learning

Type A - Interpreting black-box models

Model distillation (soft DT) [Frosst&Hinton,17]

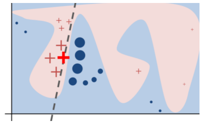
Looking into the black box



[Olah, distill.pub]

Type B - Interpreting predictions from black-box models

Attribution methods: e.g. LIME



"Why Should I Trust You?" Explaining the Predictions of Any Classifier [Ribeiro et al. '16]

Type C - Learning interpretable models

Decision tree, rules, linear model, scoring model, … prototypes

Encouraging Interpretability as part of the obj. funct.

Des arbres plus performant mais tjs interpretables

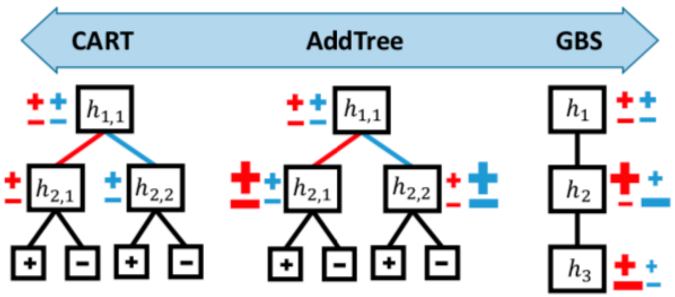


Fig. 1. A depiction of the continuum relating CART, GBS, and our AddTree. Each algorithm has been given the same 4 training instances (blue and red symbols); the symbol's size depicts its weight when used to train the adjacent node.

J. M. Luna, E. D. Gennatas, L. H. Ungar, E. Eaton, E. S. Diffenderfer, S. T. Jensen, C. B. Simone, J. H. Friedman, T. D. Solberg, and G. Valdes, "Building more accurate decision trees with the additive tree.," PNAS, vol. 116, no. 40, pp. 19887–19893, Oct. 2019.

Constructing optimal logical models.

Table 3 | Scoring system for risk of recidivism

1.	Prior arrests ≥ 2	1 point	···
2.	Prior arrests ≥ 5	1 point	+···
3.	Prior arrests for local ordinance	1 point	+···
4.	Age at release between 18 to 24	1 point	+···
5.	Age at release ≥ 40	−1 point	+···
		Score	= ···

Score	−1	0	1	2	3	4
Risk (%)	11.9	26.9	50.0	73.1	88.1	95.3

This system is from ref. [21], which was developed from refs. [29,46]. The model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

RiskSLIM (Risk-Supersparse-Linear-Integer- Models) algorithm

C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 1–10, May 2019.

Define interpretability for specific domains and create methods accordingly, including computer vision

Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

Interpretable deep learning : « 'This look like that' because its reasoning process considers whether 'this' part of the image looks like 'that' prototype.

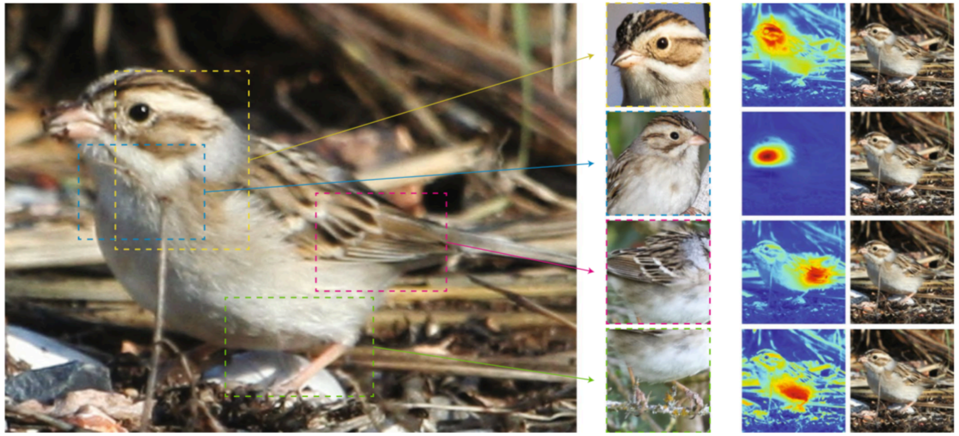


Fig. 3 | Image from the authors of ref. [48], indicating that parts of the test image on the left are similar to prototypical parts of training examples.

C. Chen, O. Li, A. Barnett, J. Su, and C. Rudin, "This looks like that - deep learning for interpretable image recognition.," *arXiv*, vol. cs.LG, 2018.

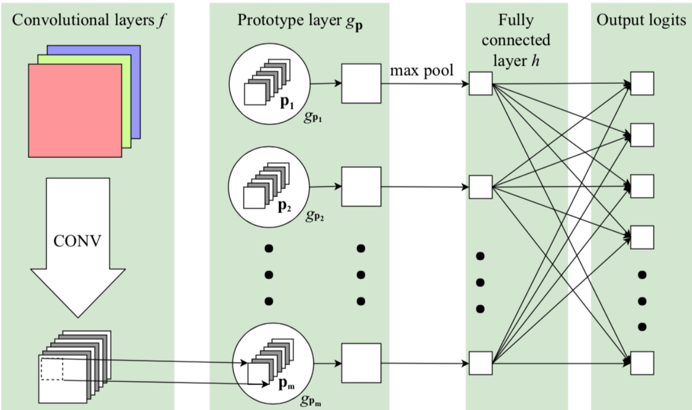
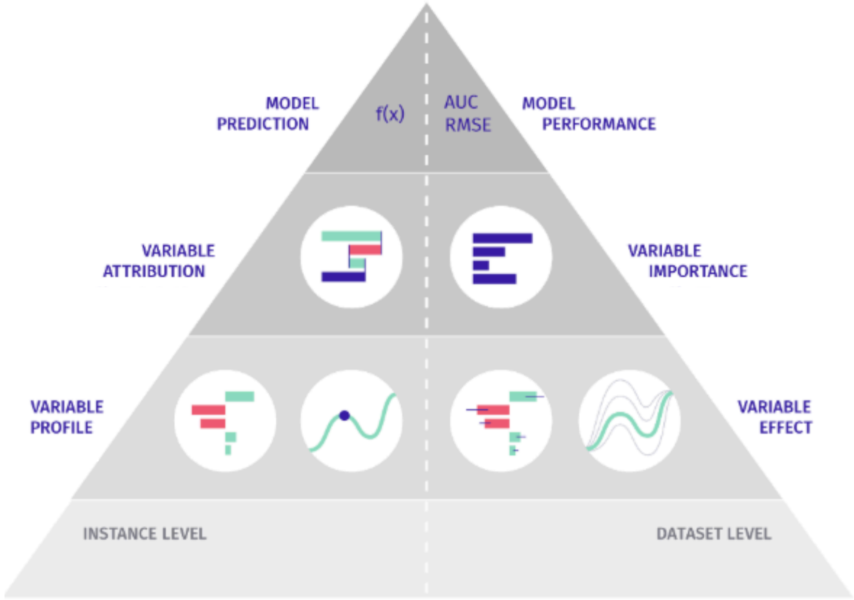


Figure 2: The network architecture.

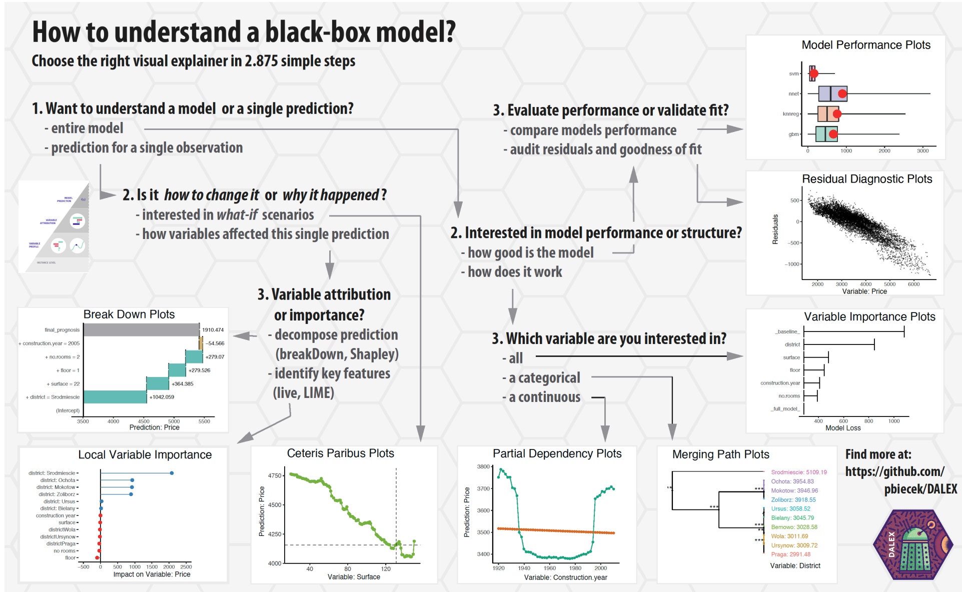
Many packages and libraries

- LIME(Local Interpretable Model-Agnostic Explanations) package
- breakDown : Outil agnostique de décomposition des prédictions des boîtes noires. Break Down Table montre les contributions de chaque variable à une prédiction finale. Break Down Plot présente les contributions des variables de manière graphique et concise. Ce package fonctionne pour les classificateurs binaires et les modèles de régression générale.
- DALEX (Descriptive mAchine Learning EXplanations) : L'ensemble Dalex contient divers explicatifs qui aident à comprendre le lien entre les variables d'entrée et la sortie du modèle. https://github.com/ModelOriented/DALEX
- IML(Interpretable Machine Learning) : Agnostic-model explanation tool.
- ceterisParibus R package
- « What-if » tool in Google TensorBoard

Model Exploration Stack

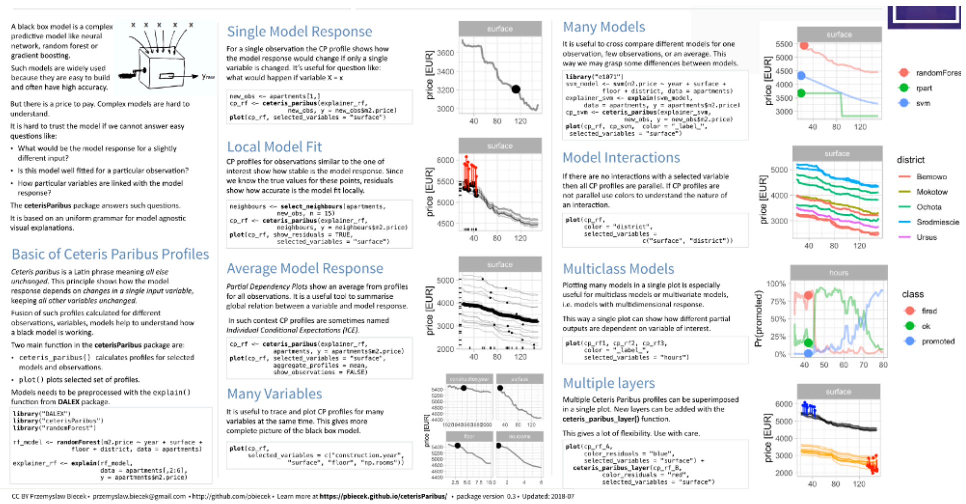


Descriptive mAchine Learning(DALEX)



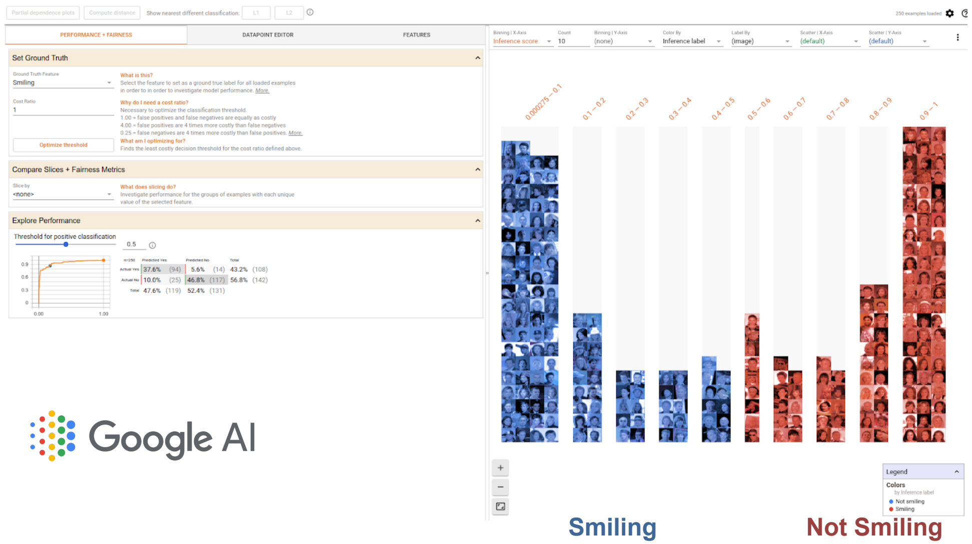
ceterisParibus: an R package for model agnostic visual exploration

Les diagrammes Ceteris Paribus (Toutes choses étant égales par ailleurs) sont conçus pour présenter des réponses modèles autour d'un **point unique dans l'espace des caractéristiques.**



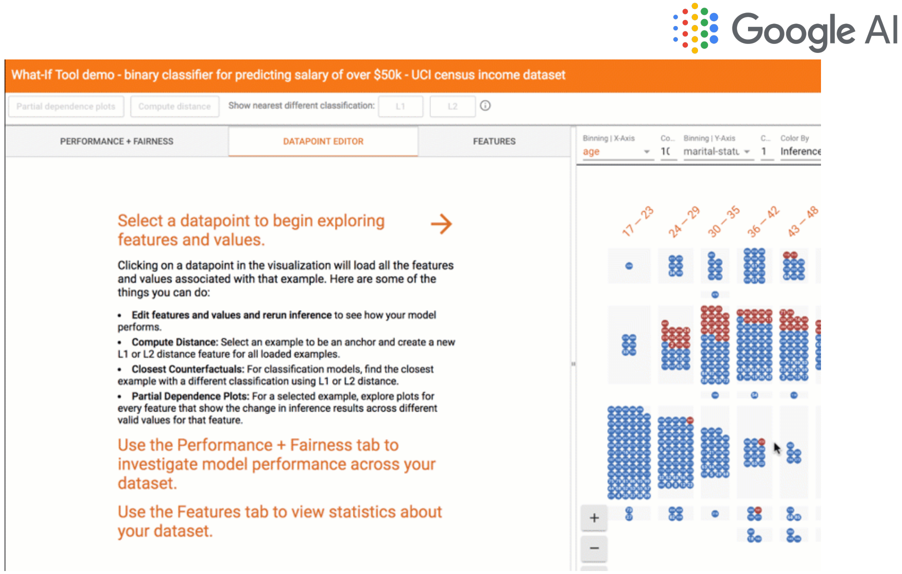
What if tool in TensorBoard: e.g. Smiling

a new feature of the open-source TensorBoard web application, which let users analyze an ML model without writing code. Given pointers to a TensorFlow model and a dataset, the What-If Tool offers an interactive visual interface for exploring model results.



https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

The What-If Tool: Code-Free Probing of Machine Learning



https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

A bit of R code to compute variable importance

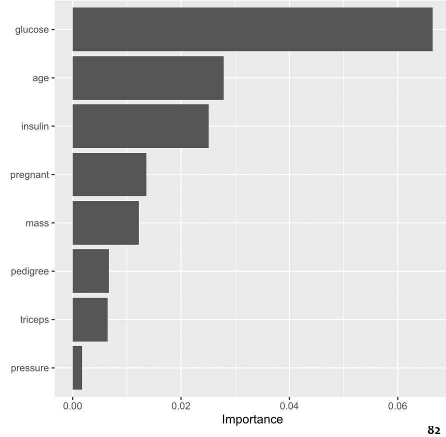
```
> head(pima)
   pregnant glucose pressure triceps insulin mass pedigree age diabetes
4         1      89       66      23      94 28.1    0.167  21      neg
5         0     137       40      35     168 43.1    2.288  33      pos
7         3      78       50      32      88 31.0    0.248  26      pos
9         2     197       70      45     543 30.5    0.158  53      pos
14        1     189       60      23     846 30.1    0.398  59      pos
15        5     166       72      19     175 25.8    0.587  51      pos
```

```
# Load required packages
# library for random forest
library(ranger)
# library for variable importance
library(vip)
# Load the Pima indians diabetes data
data(pima, package = "pdp")
pima <- na.omit(pima)  # remove records with
 missing values

# Fit a random forest
set.seed(1322)  # for reproducibility
rfo1 <- ranger(diabetes ~ ., data = pima,
 importance = "permutation")

# Plot VI scores
p1 <- vip(rfo1)  # model-specific

plot(p1)
```



Conclusion on interprétations

- ☑ **Predictive accuracy** : well addressed by both literature and tools
- ☑ **Descriptive accuracy**: more and more approaches (GAFA, R package, Python Library, …)
- ☑ **Relevancy** : « A major limitation of existing work on interpretable machine learning is that the explanations are designed based on the intuition of researchers rather than focusing on the demands of endusers »

M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," Communications of the ACM, vol. 63, no. 1, pp. 68–77, Dec. 2019.

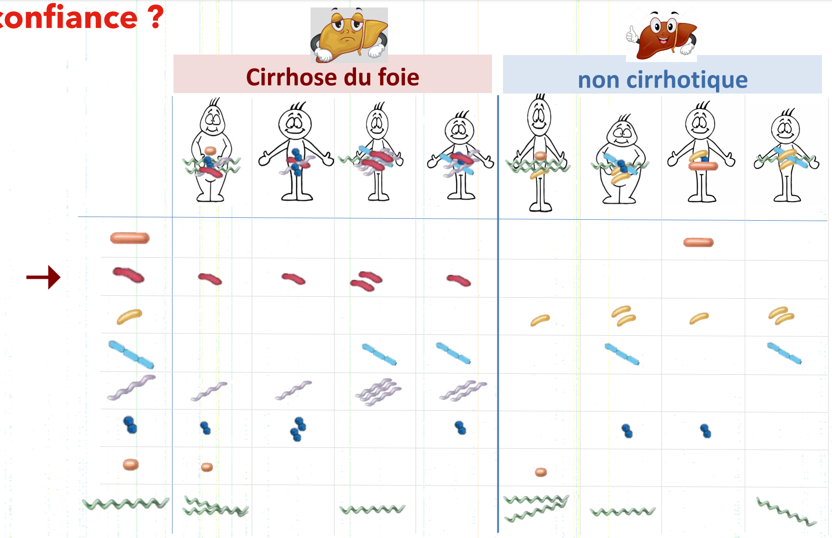
Explanation formats that might be more understandable and friendly to users

- ☑ **Contrastive explanations.** "Why Q rather than R?" The user may compare with another real case and raise question: "Why didn't I get a MRI when my neighbor did?" On the other hand, the user may ask: "Why was my request for X treatment rejected ?" Since it is compared to an event that has not happened, thus the desirable explanation here can also be called **counterfactual explanation**."Your MRI would be accepted if your invalidity score was Y"
- ☑ **Selective explanations**. Usually, users do not expect an explanation can cover the complete cause of a decision. A **sparse explanation**, which includes a minimal set of features that help justify the prediction is preferred, although incompletely.
- ☑ **Credible explanations.** Good explanation might be **consistent with prior knowledge** of general users. Low credibility could be caused by the poor fidelity of explanation to the original model.
- ☑ **Conversational explanations**. Explanations might be delivered as a conversation between the explainer and explanation receivers. It means we must consider the social context, that is, to whom an explanation is provided, in order to determine the content and formats of explanations.

M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," Communications of the ACM, vol. 63, no. 1, pp. 68–77, Dec. 2019.

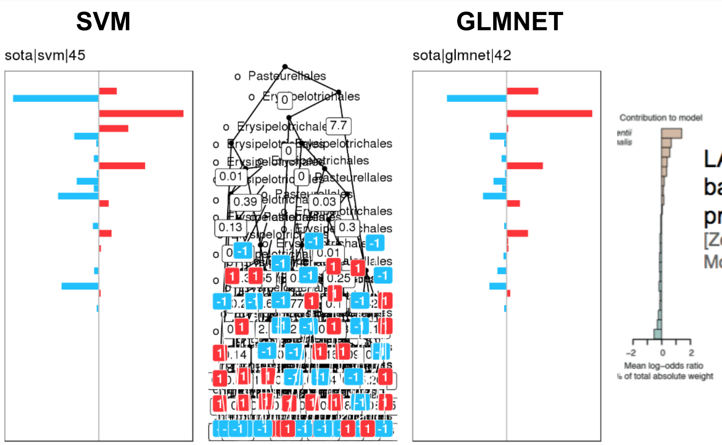
Plan

Médecine de précision basée sur la métagénomique : quelle confiance ?



Signature géniques impliquant plusieurs gènes bactériens parmi des **millions**…

State of the art models are not easy to interpret



LASSO model based on species profiles (Specl)
[Zeller*, Tap*, Voigt* et al., Mol. Syst. Biol. 2014]

Accurate but black boxes …
High-Dimension compatibility
Rely on a large number of genes (or species, or functions, or taxonomic level)

Objective→ develop algorithms to learn interpretable models as accurate as the state of the art on average

Inspired by microbial ecosystem interactions

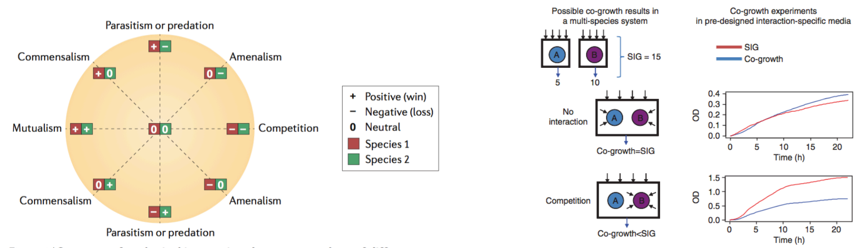


Figure 1 | **Summary of ecological interactions between members of different species.** The wheel display introduced by Lidicker[1] has been adapted to summarize all possible pairwise interactions. For each interaction partner, there are three possible outcomes: positive (+), negative (–) and neutral (0). For instance, in parasitism, the parasite benefits from the relationship (+), whereas the host is harmed (–); this relationship is thus represented by the symbol pair + –.

- **Microbial ecosystem interactions: the addition, subtraction, and ratio of microbial taxon abundances may become signature.**
- **Binary models tests whether the cumulated abundance of a set of species is below or above a certain threshold.**
- **Ternary tests whether the difference of cumulated abundance of a two sets of species is below or above a certain threshold.**
- **Ratio model tests whether the ratio of two sets of cumulated abundance is above a given threshold.**

« Intrinsically » Interpretable Models

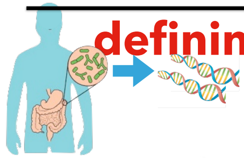
- **Interpretability criteria**
  - Conciseness
  - Models that can be applied « manually » to get a decision
  - simple operations (+,-,*, opérateurs logiques), integer values

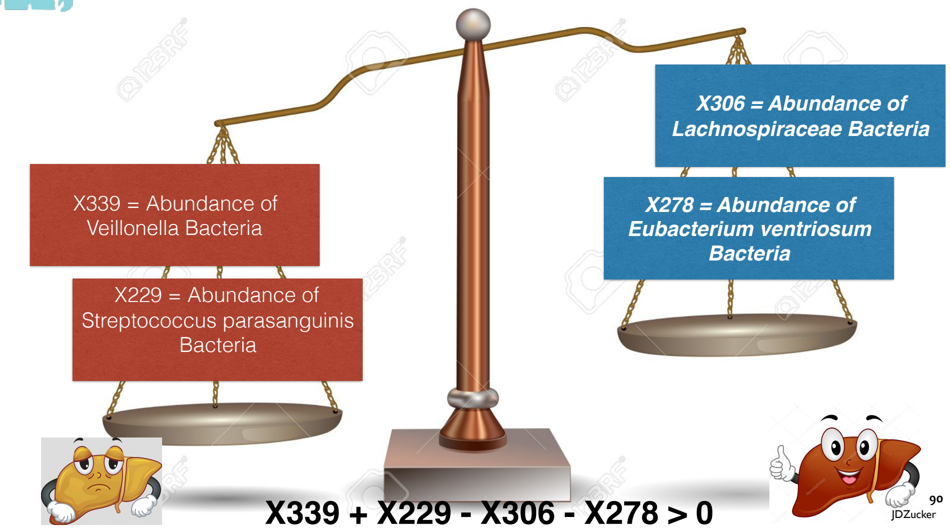
- **Example 1:** Discrete linear models

$$y \sim x_1 - x_4 + x_5 + x_8 - x_{14}$$

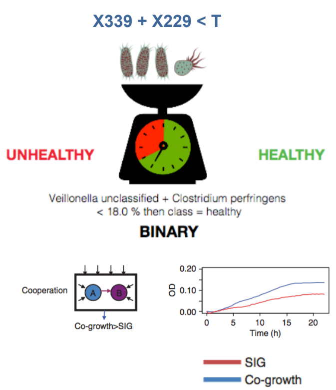
- **Example 2:** Scoring Models

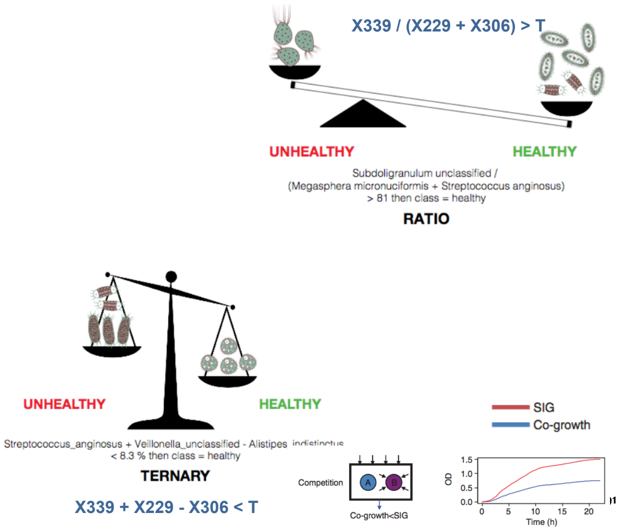
Commensurability of data supports defining easy to interpret models : BTR





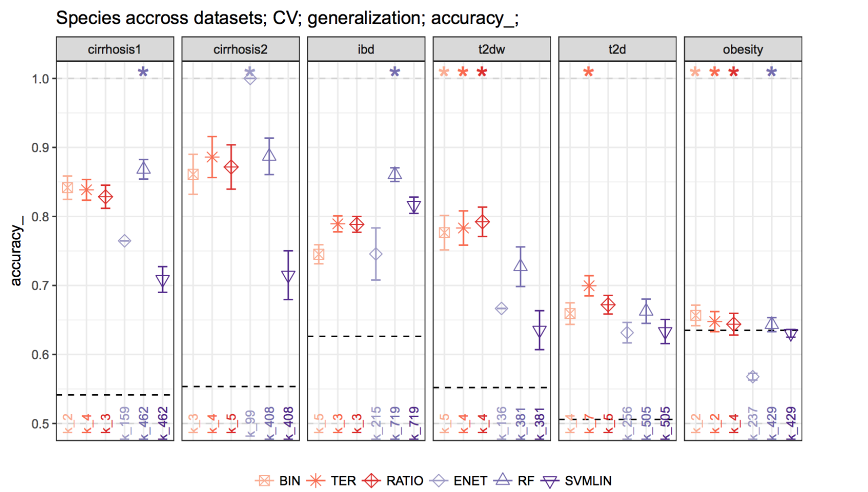
The three balance concepts depicting the BTR models inspired from microbial ecosystem





Prifti, E., Y. Chevaleyre, ... and J.-D. Zucker
(In press Gigascience 2020)

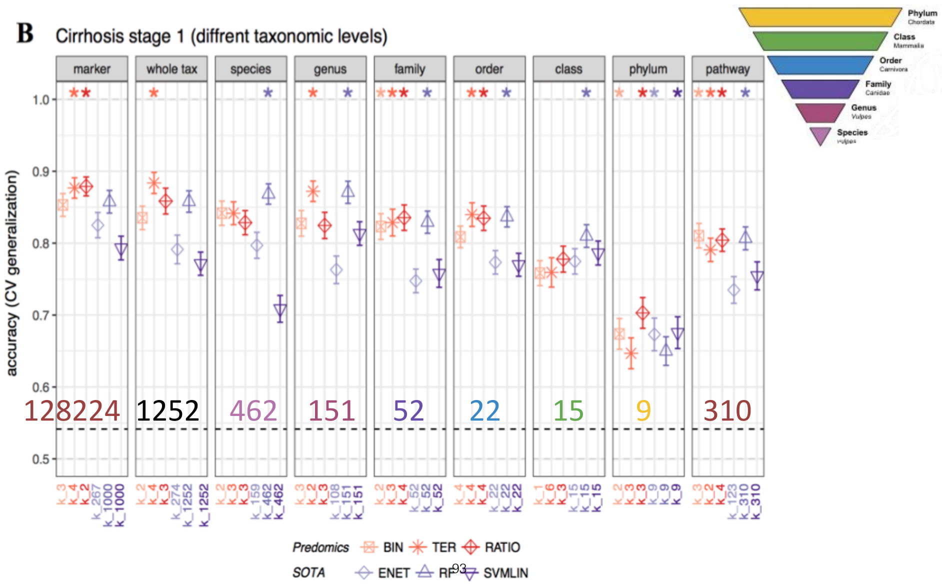
Machine Learning to learn super-sparse, interpretable signatures as precise as state of the art on average



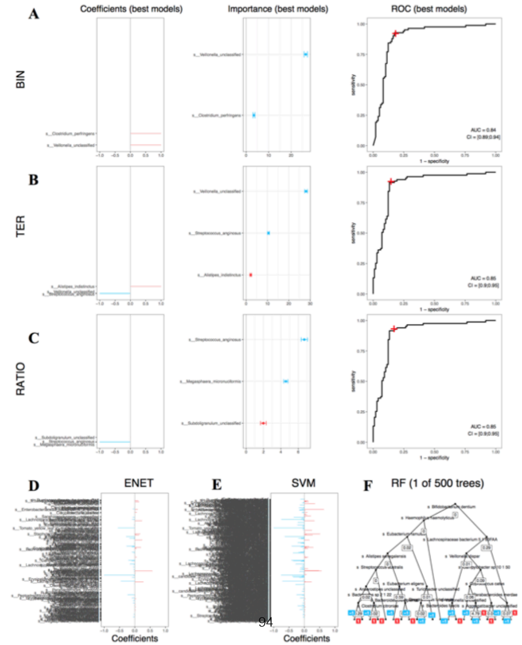
Prifti, E., Y. Chevaleyre, B. Hanczar, E. Belda Cuesta, K. Clement, A. Danchin and J.-D. Zucker (In press Gigascience)

BTR and SOTA performance across different disease and taxonomic levels in presence/absence data

BTR performed at least as well as SOTA in 43/54 (80%) of the experiments and outperformed SOTA in 14/54 (26%), while the SOTA outperformed BTR in 11/54 (20%) of the cases



BTR models are interpretable compared to state-of-the-art...



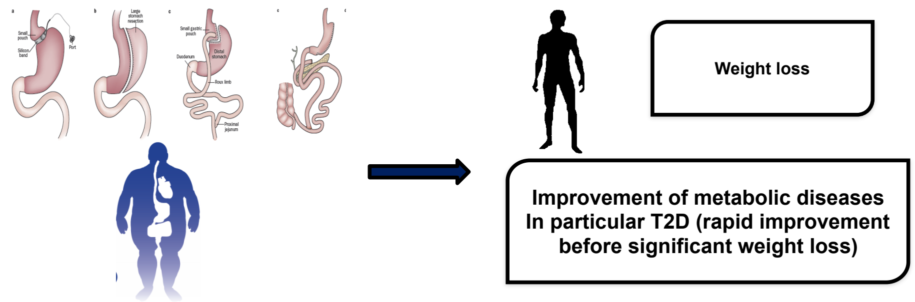
Models are also biologically « justifiable »

(S8) g__Coprococcus - g__Veillonella > -0.1 ***then*** *class = healthy*

(S9) (g__Eubacterium + g__Bacteroides) / g__Veillonella > 140 ***then*** *class = healthy*

- The potential competition between oral and gut microbes in the progression to cirrhosis reported in previous studies is reflected in best by Ter and Ratio models with **genus abundance data**, that combine *Veillonella* (oral bacteria; opportunistic pathogen) **enriched in liver cirrhosis patients** at one side

Bariatric surgery improves Type 2 Diabetes (T2D)



The **DiaRem** score is used to predict remission

Diarem score: was best score (2013); validated on independent cohorts

	Score
Age (years)	
<40	0
40–49	1
50–59	2
≥60	3
HbA_{1c} (%)	
<6·5%	0
6·5–6·9%	2
7·0–8·9%	4
≥9·0%	6
Other diabetes drugs	
No sulfonylureas or insulin-sensitising agent other than metformin	0
Sulfonylureas and insulin-sensitising agent other than metformin	3
Treatment with insulin	
No	0
Yes	10

Total score calculated by adding scores for each of the four variables.

Table 5: Calculation of DiaRem score for prediction of the probability of diabetes remission after Roux-en-Y gastric bypass surgery

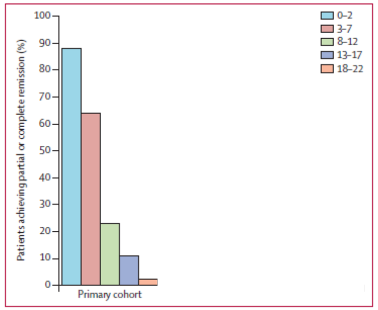


Figure 4: Proportion of patients in each cohort achieving partial or complete remission at 14 months after surgery, by DiaRem score

Table 2 Prediction errors of the different models in percentage

	Train error	Test error	Cross-validation error
Hayes - SL	13.4	31	
Hayes - J48DT	12.6	34.5	
Dixon - LR6	16	40.5	
Dixon - LR7	12	22.6	
Lee - Score	27.3[a]	16.7	16.7
Still - Score	19.4[b]	15.5	15.9
LR	7.1		19.9
DT	13.1		17.6
Lasso	14.3		18.9
EN[c]	13.1		17.2

Still et al Lancet endocrinology 2013;

Automated Score Re-Construction of Diarem

1. Identification of related clinical variables

age	glycated hemoglobin	insuline	other drugs

2. Meaningful thresholds for clinical variables

age				glycated hemoglobin				insuline		other drugs	
<40	40–49	50 – 59	>60	<6.5	6.5 – 6.9	7 – 8.9	> 9	yes	no	yes	no

3. Optimization of weights for sub-groups of the variables

age				glycated hemoglobin				insuline		other drugs	
<40	40–49	50 – 59	>60	<6.5	6.5 – 6.9	7 – 8.9	> 9	yes	no	yes	no
0	1	2	3	0	2	4	6	10	0	3	0

4. Find an optimal separator between two classes

Classify as `Remission` if sum of scores < 7
Classify as `Non-remission` if sum of scores ≥ 7

Score Construction as an Optimization Problem

We minimise the hinge loss penalized by the Fused Lasso:

$$\sum_{i=1}^{N} \ell(y_i, \theta \cdot \bar{x}_i + b) + \lambda \sum_{j=1}^{\bar{d}-1} |\theta_j - \theta_{j+1}|.$$

The linear programming formulation of the problem:

$$\begin{aligned}
&\min \Big(\sum_{i=1}^{N} \xi_i + \sum_{j=1}^{\bar{d}} \eta_j\Big), \text{ such that} \\
&\text{for all } i,\ y_i(\theta \cdot \bar{x}_i + b) \geq 1 - \xi_i, \\
&\text{for all } j,\ \ -\lambda\eta_j \leq \theta_j - \theta_{j+1} \leq \lambda\eta_j, \\
&\qquad \xi_i \geq 0, \theta_i \in \mathbb{N} \text{ for all } i.
\end{aligned}$$

Fully Corrective Binning (FCB) algorithm

Nataliya Sokolovska, Yann Chevaleyre and Z. Jean Daniel (AISTATS 2018).
Sokolovska, N., Y. Chevaleyre and J.-D. Zucker (DA2PL'2016)

The AdDiaRem

Using our Fully Corrective Binning (FCB) algorithm

Age	
[15 – 41]	0
(41 – 52]	3
(52 – 69]	5
HbA1c	
[4.5 – 6.9]	0
(6.9 – 7.4]	2
(7.4 – 18.4]	4
Insuline	
No	0
Yes	3

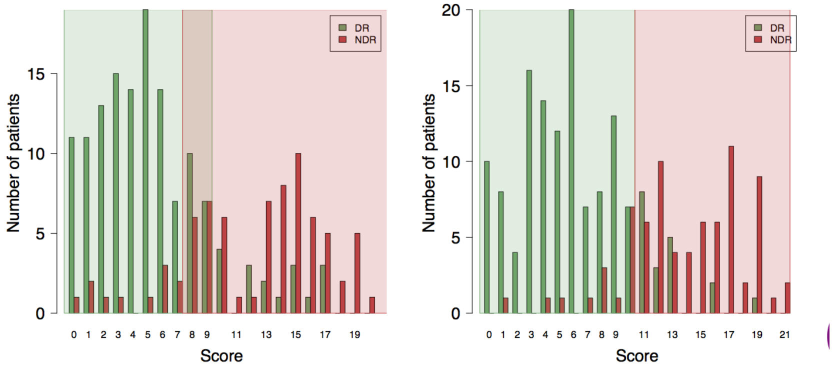
Other glucose-lowering drugs	
No	0
Yes	1
Number of glucose-lowering drugs	
0	0
1	1
2	2
≥ 3	3
Diabetes duration	
[0 – 6.9]	0
(6.9 – 14]	3
≥ 14	5

The training procedure relies on the IBM ILOG CPLEX Optimization Studio[2] which efficiently performs the constrained optimization. In particular, integrity constraints are added to the optimisation problem to obtain integer solutions.

ANR DiagnoLearn N. Sokolovska (PI) ANR 2018-2020

The AdDiaRem

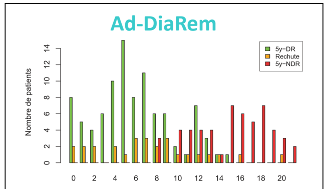
- New biomarkers (diabetes duration, number of drugs taken)



The distributions of the DiaRem and AdDiaRem scores

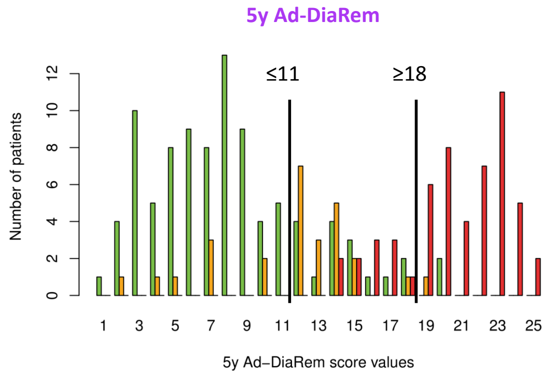
J. Aron-Wisnewsky et al., Diabetologia, 2017

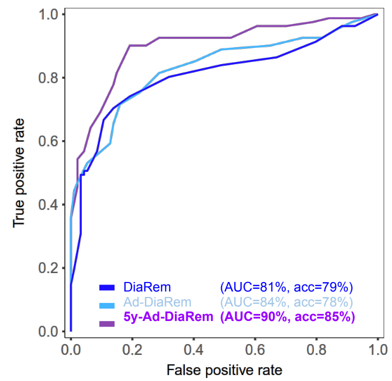
Another score dedicated to 5y T2D Remission proposed 5yAd-DiaRem (n=175)



Précision = 0.79 / AUROC = 0.84

Debedat, J., N. Sokolovska, ...J. D. Zucker, K. Clement and J. Aron-Wisnewsky (2018). "Long-term Relapse of Type 2 Diabetes After Roux-en-Y Gastric Bypass: Prediction and Clinical Relevance." **Diabetes Care.**





Open question : use of AdDiarem & 5YAdDiarem (at 1Y) to improve the follow-up patients prognosed to relapse.

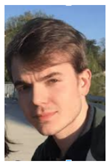
Glayn et al. Precision medicine in the management of type 2 diabetes. *THE LANCET Diabetes & Endocrinology* 2018

Collaboration with:

Pr Karine Clément

Dr Judith Aron-Wisnewsky

Dr Nataliya Sokolovska



Jean Debédat

Dr Michèle Guerre-Millo

Garance de Turenne

Service de nutrition	**BARICAN – ICAN**	**Hôpital Louis-Mourier**	**Chirurgiens**
Pr Jean-Michel Oppert Pr Christine Poitou	Valentine Lemoine, ARC Dr Florence Marchelli	Dr Séverine Ledoux Dr Muriel Coupaye	Pr Jean-Luc Bouillot Dr Laurent Genser

ICAN Institute of Cardiometabolism And Nutrition — Assistance Publique Hôpitaux de Paris — Inserm — Sorbonne Université

NUTRIOMICS — L'institut Benjamin Delessert

Plan

I. Apprentissage Artificiel et médecine

II. Médecine de précision

III. Pourquoi des modèles interprétables en médecine ?

IV. Machine Learning interpretable trois approches

V. Deux exemples de modèles interprétables

VI. Conclusion

Conclusions on precision medicine and AI

- **La médecine de précision** annonce un bouleversement dans la prise en charge des patients, leur parcours de soin et leur suivi grâce à l'IA.
- **Nouveaux diagnostics moléculaires** (omiques) et **d'imagerie**
  - **stratification** des maladies → meilleurs diagnostic,
  - aide au **prognostic** → meilleurs choix des traitements,
  - **désert médicaux** → tri des patients les plus à risques.
- Progrès de l'IA et du **Deep Learning** posent des **questions éthiques** sur son adoption en médecine : équité/confiance/transparence/**interprétabilité**
- **L'IA doit aider les cliniciens** (pas se substituer) à être plus efficace mais **l'interpretabilité** est indispensable pour éviter les erreurs et contribuer à la recherche de l'étiologie …
- Explications souvent pour des **experts**… et **non des utilisateurs finaux**…

Future of interpretability in ML

T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019.

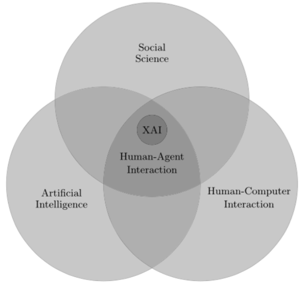
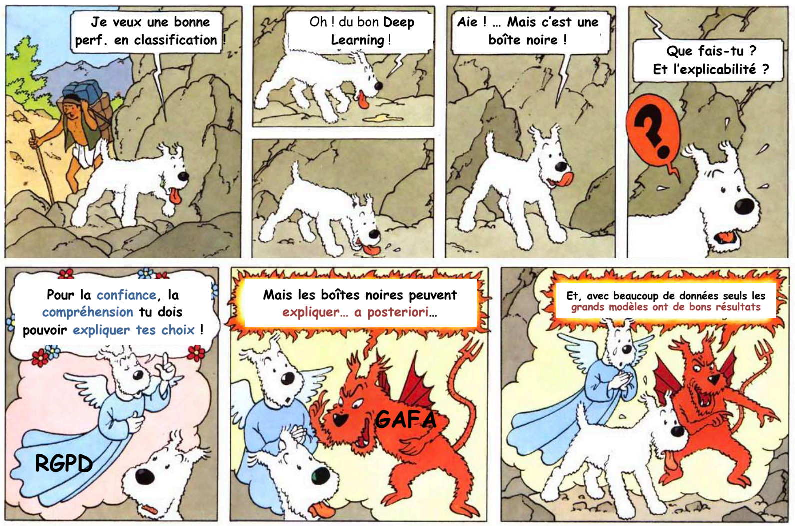


Fig. 1. Scope of explainable artificial intelligence.

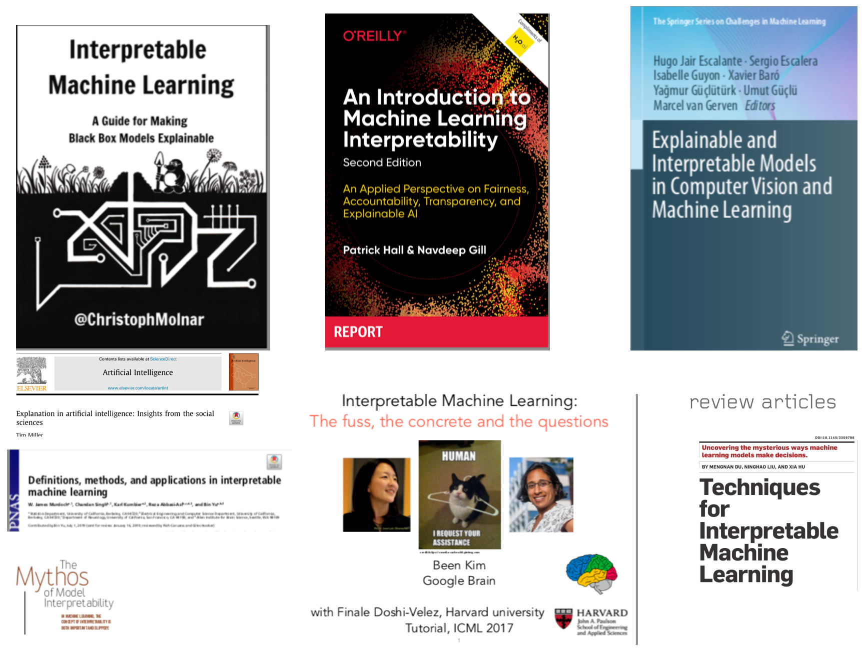
- ☑ Explanations are **contrastive** — People rather ask why event P happened instead of some event Q.→ social and computational consequences for XAI
- ☑ Explanation are **selected (in a biased manner)** — Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be **THE explanation**.
- ☑ Explanation using **probabilities probably don't matter so much** — statistical relationships in explanation is not as effective **as referring to causes.**

Milou, l'interpretabilité et sa consommation de Deep Learning

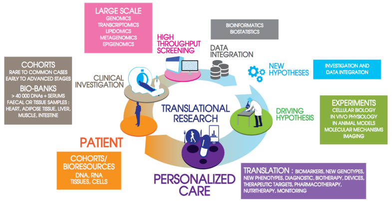


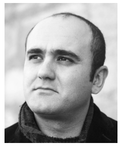
Jean-Daniel Zucker

Bibliography



Merci





Dr. Edi Prifti