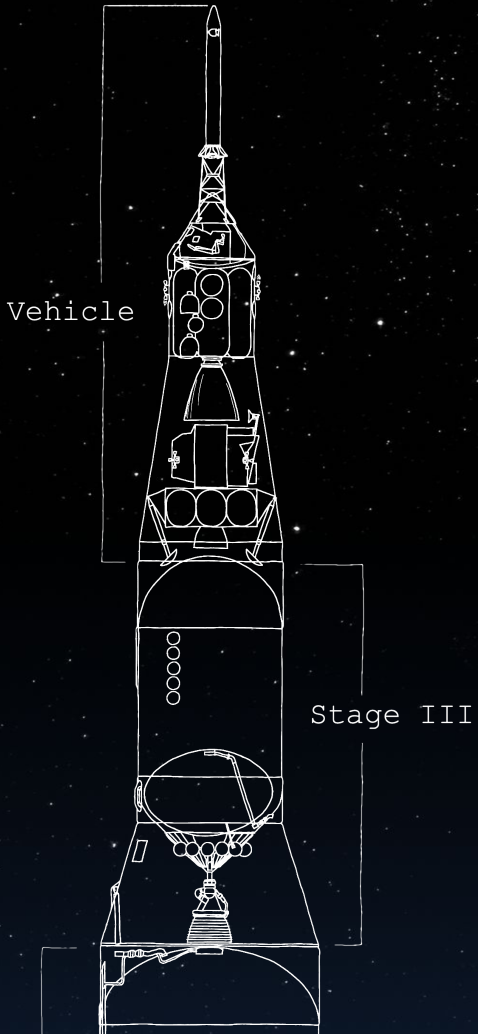
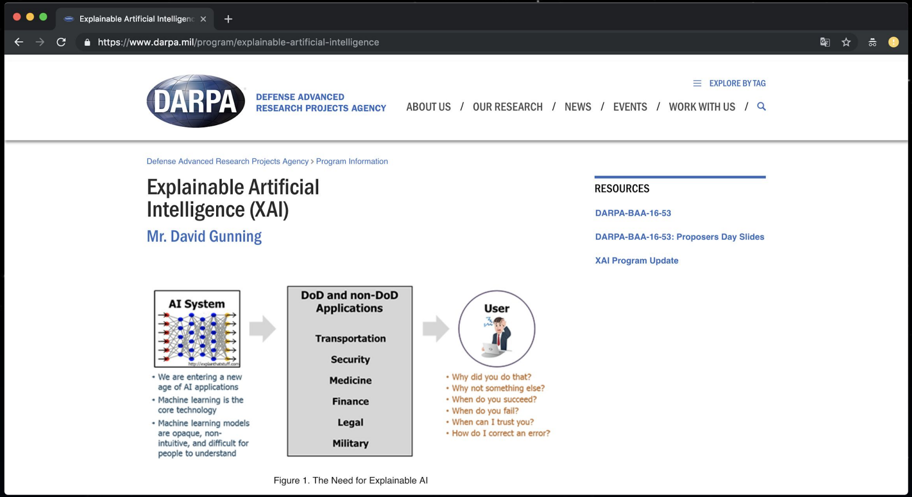


XAI

THE Key to AI operationalization

Clodéric Mars - CTO @ craft ai

17/10/2019 - Séminaire Aristote l'IA est-elle explicable

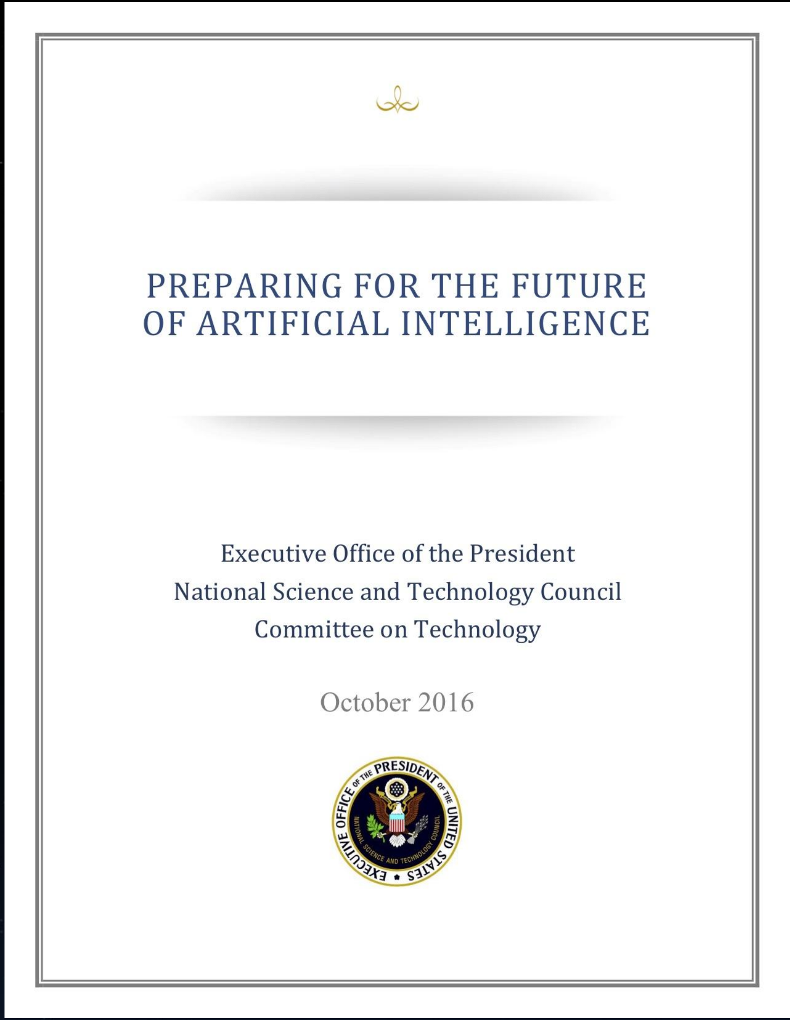


> The current generation of AI systems offer tremendous benefits, but **their effectiveness will be limited by the machine's inability to explain its decisions and actions to users**

\- *David Gunning - DARPA/I2O XAI Program Update November 2017*

Les systèmes d'IA ont vocation à interagir avec des utilisateurs humains : **ils doivent donc être capables d'expliquer leur comportement**, de justifier d'une certaine manière les décisions qu'ils prennent afin que les utilisateurs humains puissent comprendre leurs actions et leurs motivations.

- *Intelligence Artificielle. Les défis actuels et l'action d'Inria*


PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE

Executive Office of the President
National Science and Technology Council
Committee on Technology

October 2016

Federal agencies [...] should [...] ensure that AI-based products or services purchased with Federal grant funds **produce results in a sufficiently transparent fashion** and are supported by evidence of efficacy and fairness.

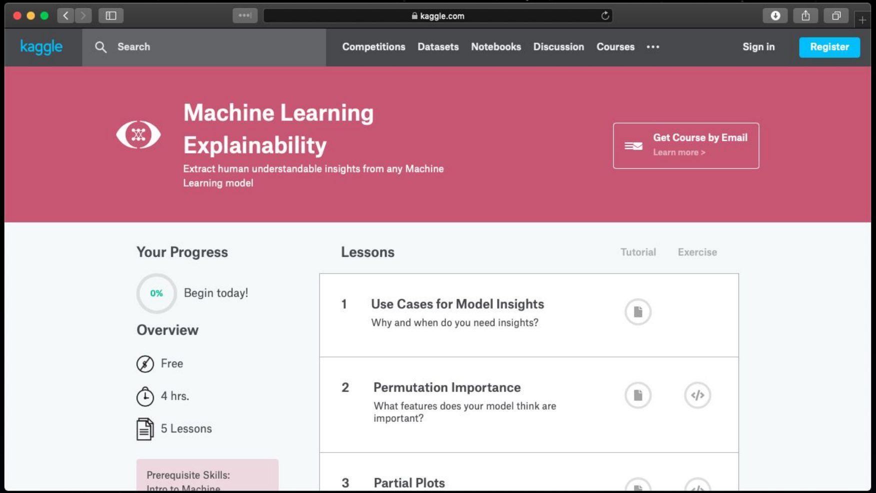
- *Preparing for the future of Artificial Intelligence*

> En premier lieu, il faut accroître la transparence et l'auditabilité des systèmes autonomes d'une part, en développant les capacités nécessaires pour observer, comprendre et auditer leur fonctionnement et, d'autre part, en **investissant massivement dans la recherche sur l'explicabilité.**

- *Cédric Villani - Donner Un Sens À L'Intelligence Artificielle*

> L'explicabilité est importante parce qu'elle apporte la confiance. **Si on est capable d'expliquer pourquoi et comment on fait les choses, ça enlève le côté magique.**

- *Luc Julia - L'intelligence artificielle n'existe pas*



Many important decisions are made by humans. For these decisions, **insights can be more valuable than predictions.**

In practice, **showing insights [...] will help build trust, even among people with little deep knowledge of data science.**

- *Dan Becker - Data Scientist, Instructor @ Kaggle*

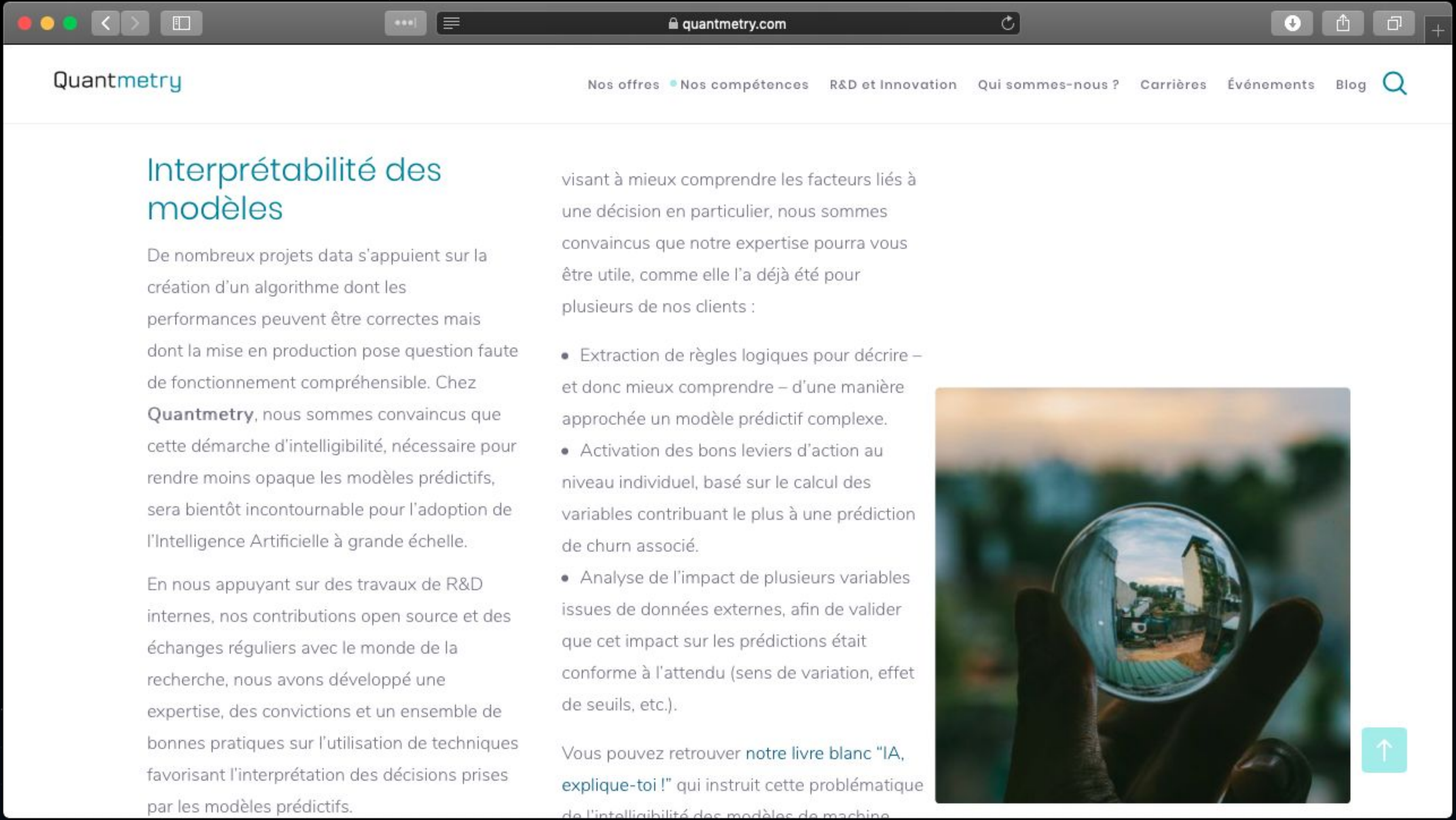
> **Explanations are mandatory** when AI empowers humans to perform complex tasks

- *Antoine Buhl - XAI, a game changer for AI in production @ AI Night 2019*

> When it comes to create AI for critical systems, **trustability and certifiability** are mandatory.

- *David Sadek - XAI, a game changer for AI in production @ AI Night 2019*


sites.google.com
IJCAI 2019 Workshop on Explainable Artifi...
IJCAI 2019 WORKSHOP ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)
11 August, 2019. Macau, China
https://www.ijcai19.org/


quantmetry.com

Quantmetry

Nos offres · Nos compétences · R&D et Innovation · Qui sommes-nous ? · Carrières · Événements · Blog

Interprétabilité des modèles

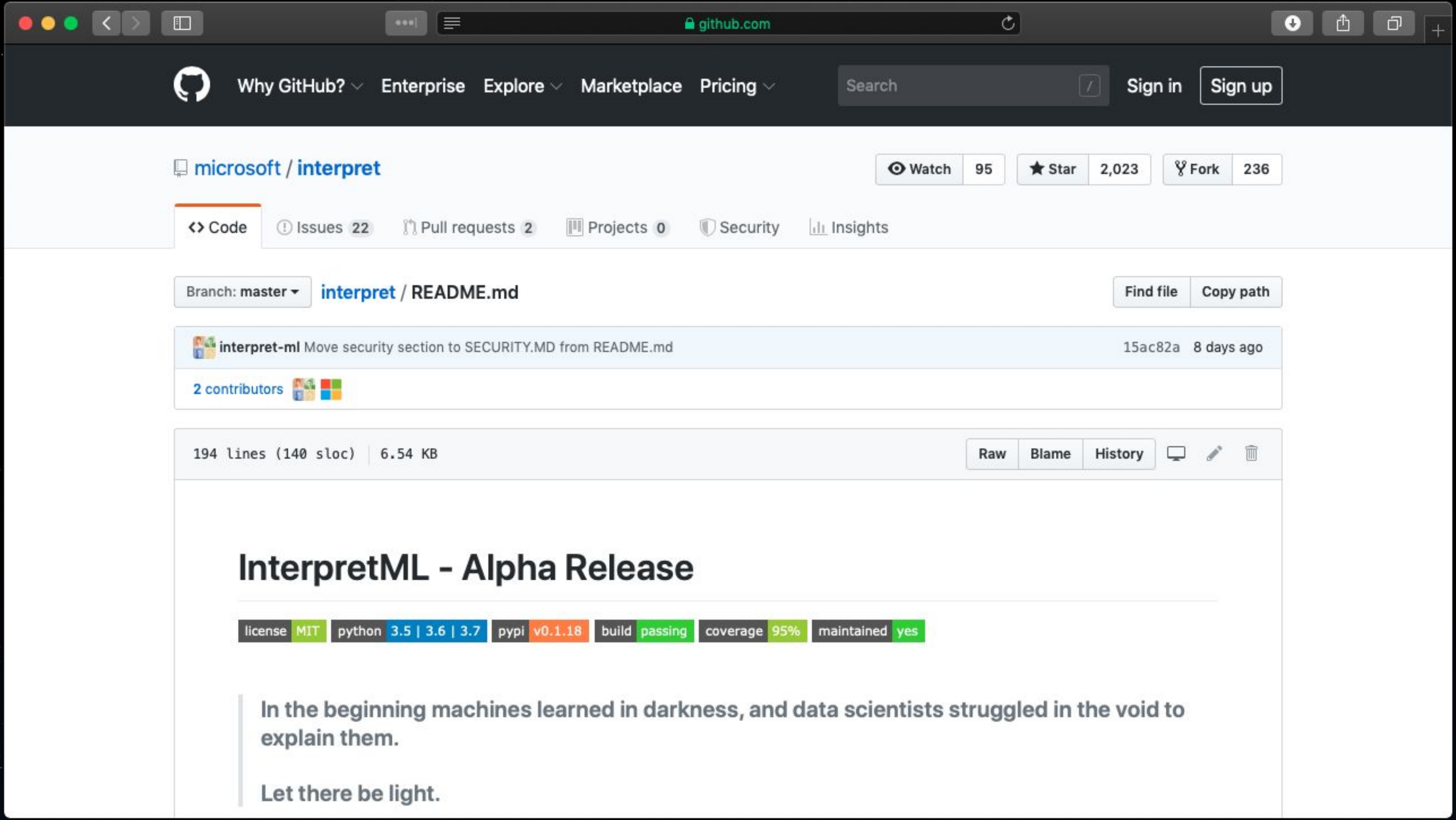
De nombreux projets data s'appuient sur la création d'un algorithme dont les performances peuvent être correctes mais dont la mise en production pose question faute de fonctionnement compréhensible. Chez **Quantmetry**, nous sommes convaincus que cette démarche d'intelligibilité, nécessaire pour rendre moins opaque les modèles prédictifs, sera bientôt incontournable pour l'adoption de l'Intelligence Artificielle à grande échelle.

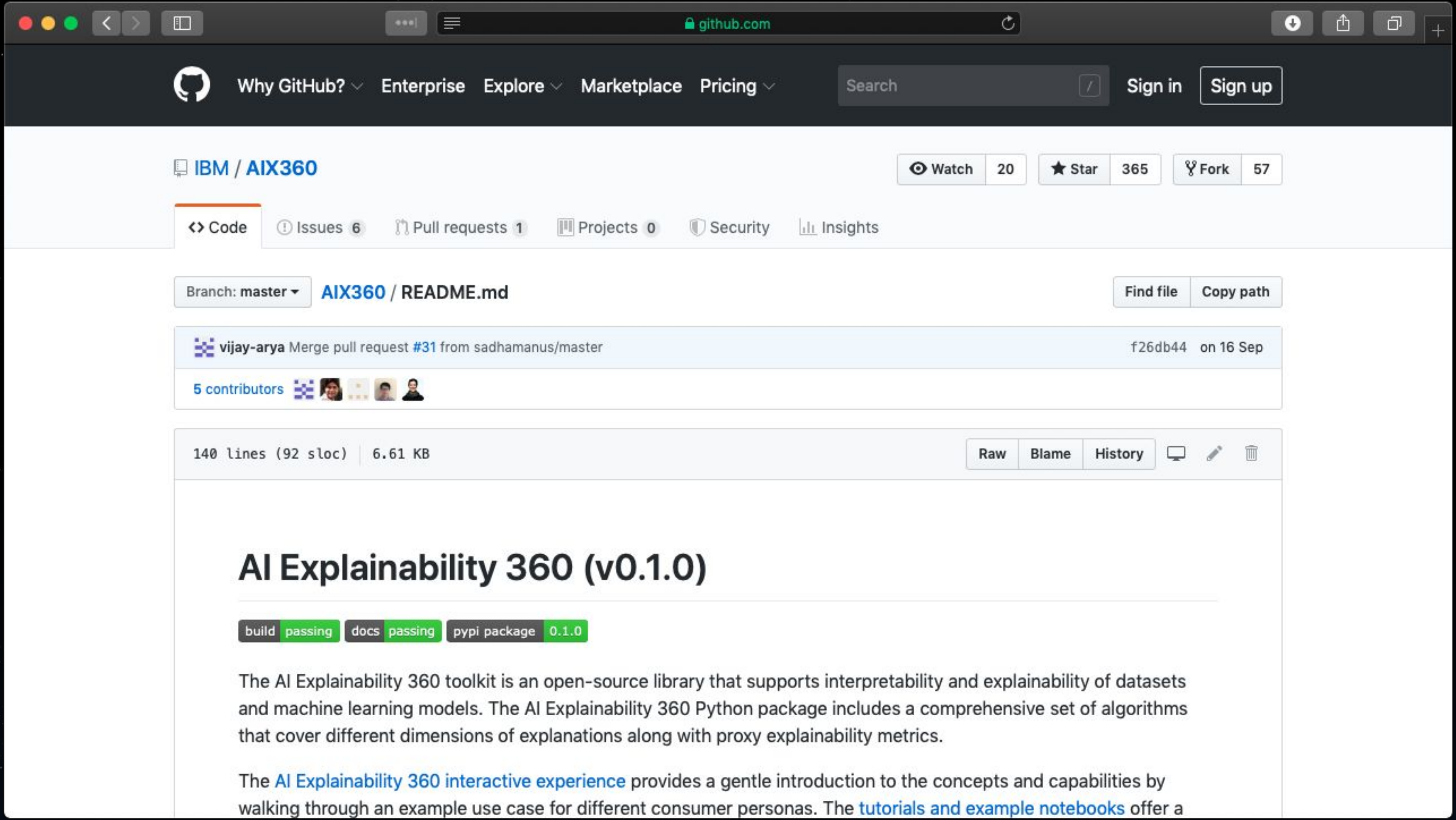
En nous appuyant sur des travaux de R&D internes, nos contributions open source et des échanges réguliers avec le monde de la recherche, nous avons développé une expertise, des convictions et un ensemble de bonnes pratiques sur l'utilisation de techniques favorisant l'interprétation des décisions prises par les modèles prédictifs visant à mieux comprendre les facteurs liés à une décision en particulier, nous sommes convaincus que notre expertise pourra vous être utile, comme elle l'a déjà été pour plusieurs de nos clients :

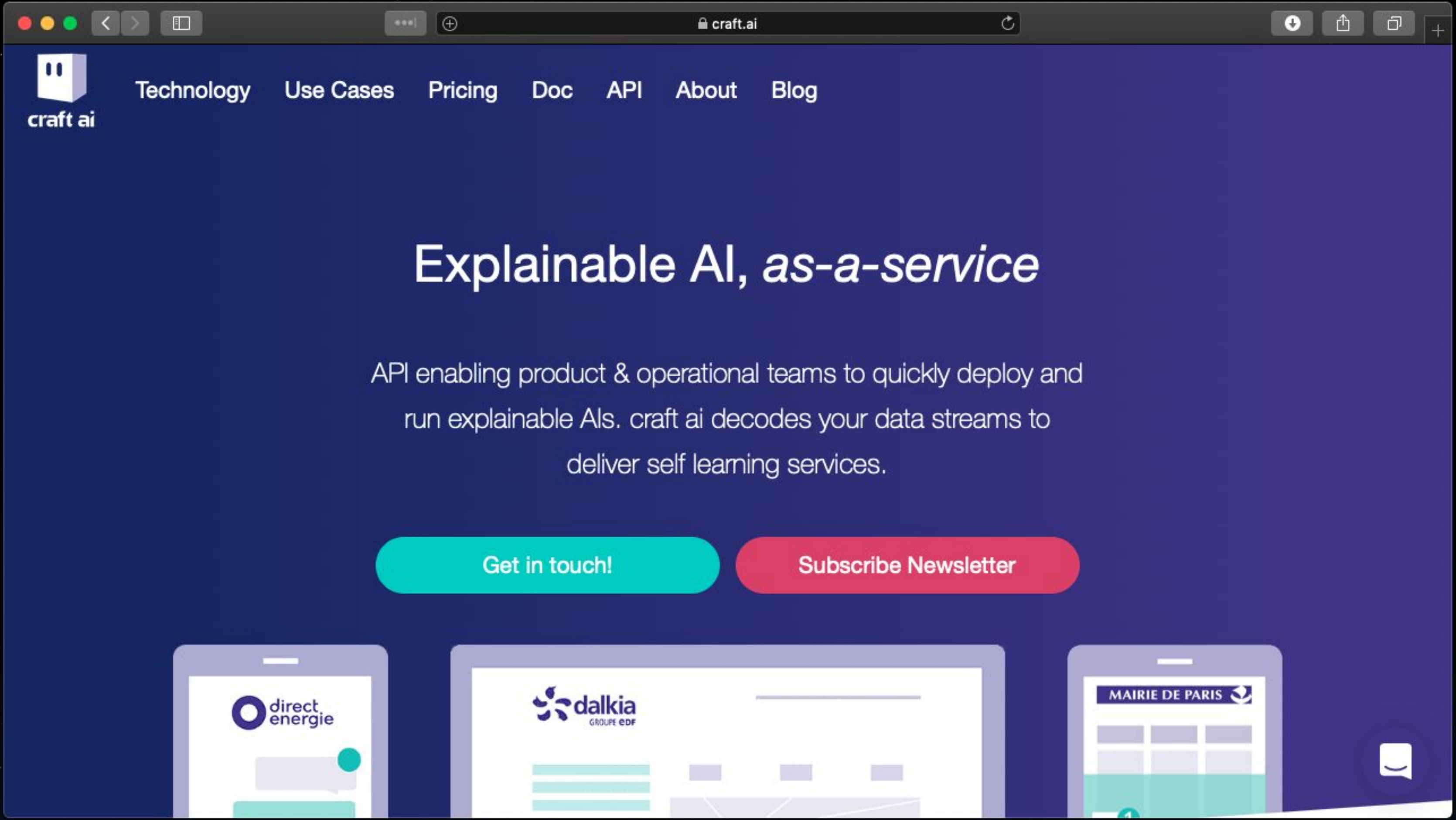
- Extraction de règles logiques pour décrire – et donc mieux comprendre – d'une manière approchée un modèle prédictif complexe.
- Activation des bons leviers d'action au niveau individuel, basé sur le calcul des variables contribuant le plus à une prédiction de churn associé.
- Analyse de l'impact de plusieurs variables issues de données externes, afin de valider que cet impact sur les prédictions était conforme à l'attendu (sens de variation, effet de seuils, etc.).

Vous pouvez retrouver notre livre blanc "IA, explique-toi !" qui instruit cette problématique

IA Explique toi ! Quand la performance ne suffit pas


github.com
Why GitHub?
Enterprise
Explore
Marketplace
Pricing
Search
Sign in
Sign up
microsoft / interpret
Watch 95
Star 2,023
Fork 236
Code
Issues 22
Pull requests 2
Projects 0
Security
Insights
Branch: master
interpret / README.md
Find file
Copy path
interpret-ml Move security section to SECURITY.MD from README.md
15ac82a 8 days ago
2 contributors
194 lines (140 sloc) 6.54 KB
Raw
Blame
History
InterpretML - Alpha Release
license MIT
python 3.5 | 3.6 | 3.7
pypi v0.1.18
build passing
coverage 95%
maintained yes
In the beginning machines learned in darkness, and data scientists struggled in the void to explain them.
Let there be light.


github.com
Why GitHub?
Enterprise
Explore
Marketplace
Pricing
Search
Sign in
Sign up
IBM / AIX360
Watch 20
Star 365
Fork 57
Code
Issues 6
Pull requests 1
Projects 0
Security
Insights
Branch: master
AIX360 / README.md
Find file
Copy path
vijay-arya Merge pull request #31 from sadhamanus/master
f26db44 on 16 Sep
5 contributors
140 lines (92 sloc) 6.61 KB
Raw
Blame
History
AI Explainability 360 (v0.1.0)
build passing
docs passing
pypi package 0.1.0
The AI Explainability 360 toolkit is an open-source library that supports interpretability and explainability of datasets and machine learning models. The AI Explainability 360 Python package includes a comprehensive set of algorithms that cover different dimensions of explanations along with proxy explainability metrics.
The AI Explainability 360 interactive experience provides a gentle introduction to the concepts and capabilities by walking through an example use case for different consumer personas. The tutorials and example notebooks offer a


craft.ai
craft ai
Technology
Use Cases
Pricing
Doc
API
About
Blog
Explainable AI, *as-a-service*
API enabling product & operational teams to quickly deploy and run explainable AIs. craft ai decodes your data streams to deliver self learning services.
Get in touch!
Subscribe Newsletter
direct energie
dalkia
GROUPE EDF
MAIRIE DE PARIS

How XAI makes a difference?

What’s an explanation?

Tim Miller - Explanation in Artificial Intelligence: Insights from the Social Sciences

> People look for explanations to improve their understanding of someone or something so that they can derive stable model that can be used for prediction and control

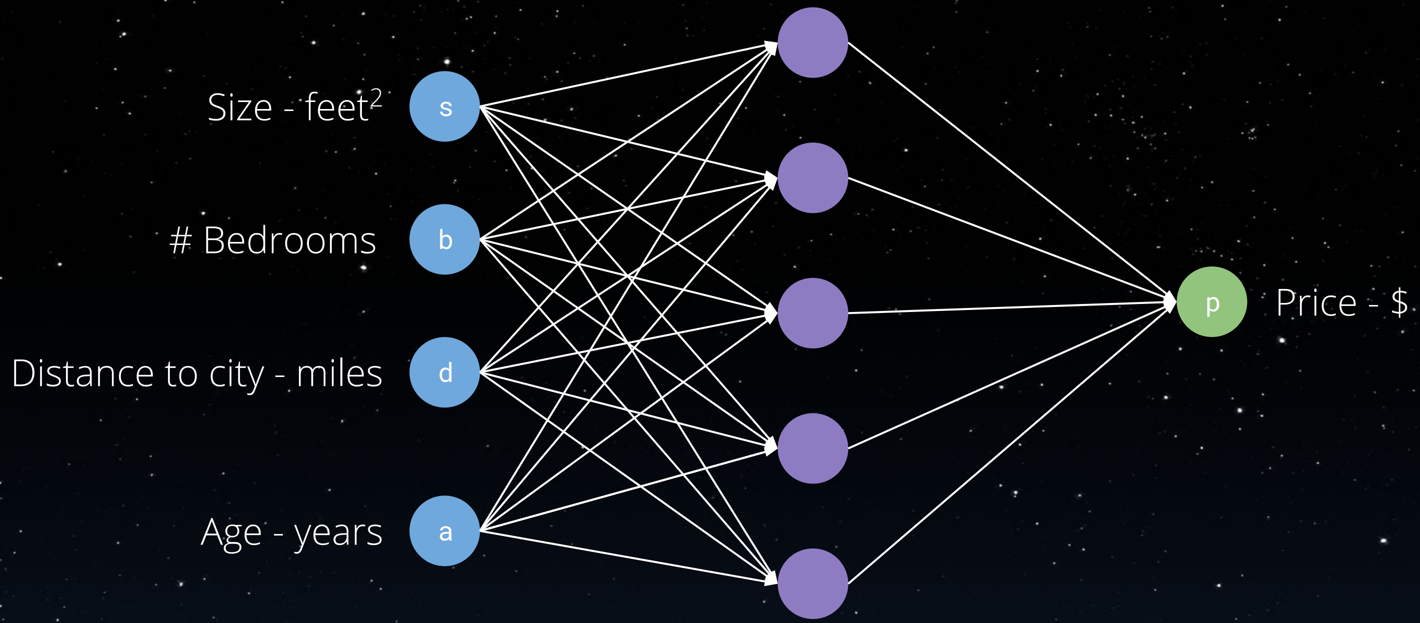
- Fritz Heider, Australian psychologist

1. Answer to “Why?” questions
2. Answer with contrastive explanations
3. Biased towards the explainee

AI, not explainable?

Anatomy of a Neural Network

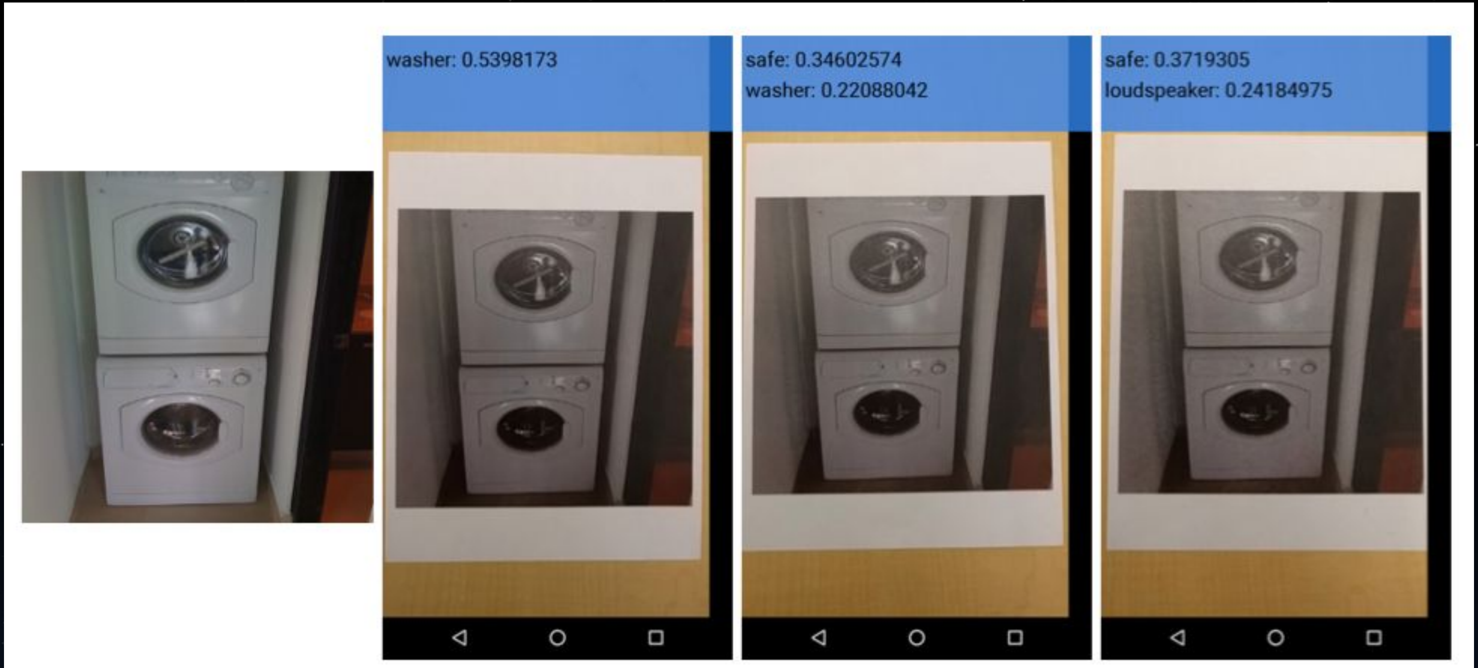
SuperDataScience - Artificial Neural Networks - How do Neural Networks Work?



$$p = f(w_1 f(w_{11}s + w_{21}b + w_{31}d + w_{41}a) + w_2 f(w_{12}s + w_{22}b + w_{32}d + w_{42}a) + \ldots)$$

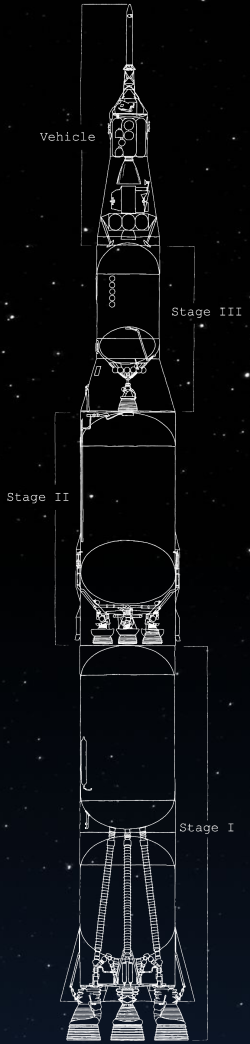
Adversarial attacks

Alexey Kurakin, Ian Goodfellow, Samy Bengio - Adversarial examples in the physical world

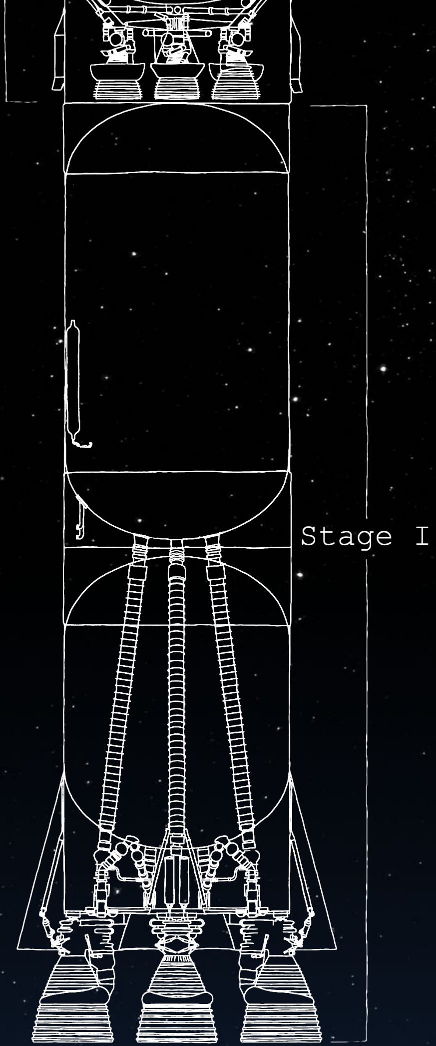


(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

How XAI makes a difference?



The 3 stages of XAI

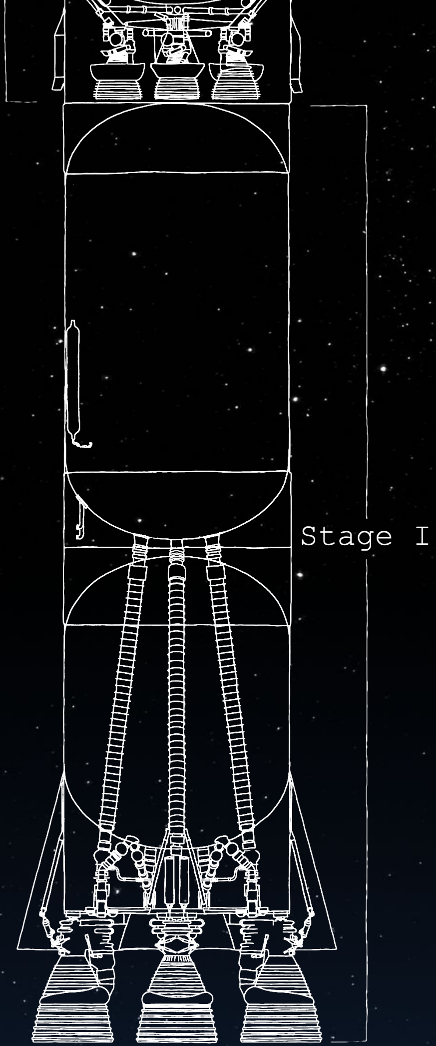


Stage I

Explainable building process

Involve Business Experts

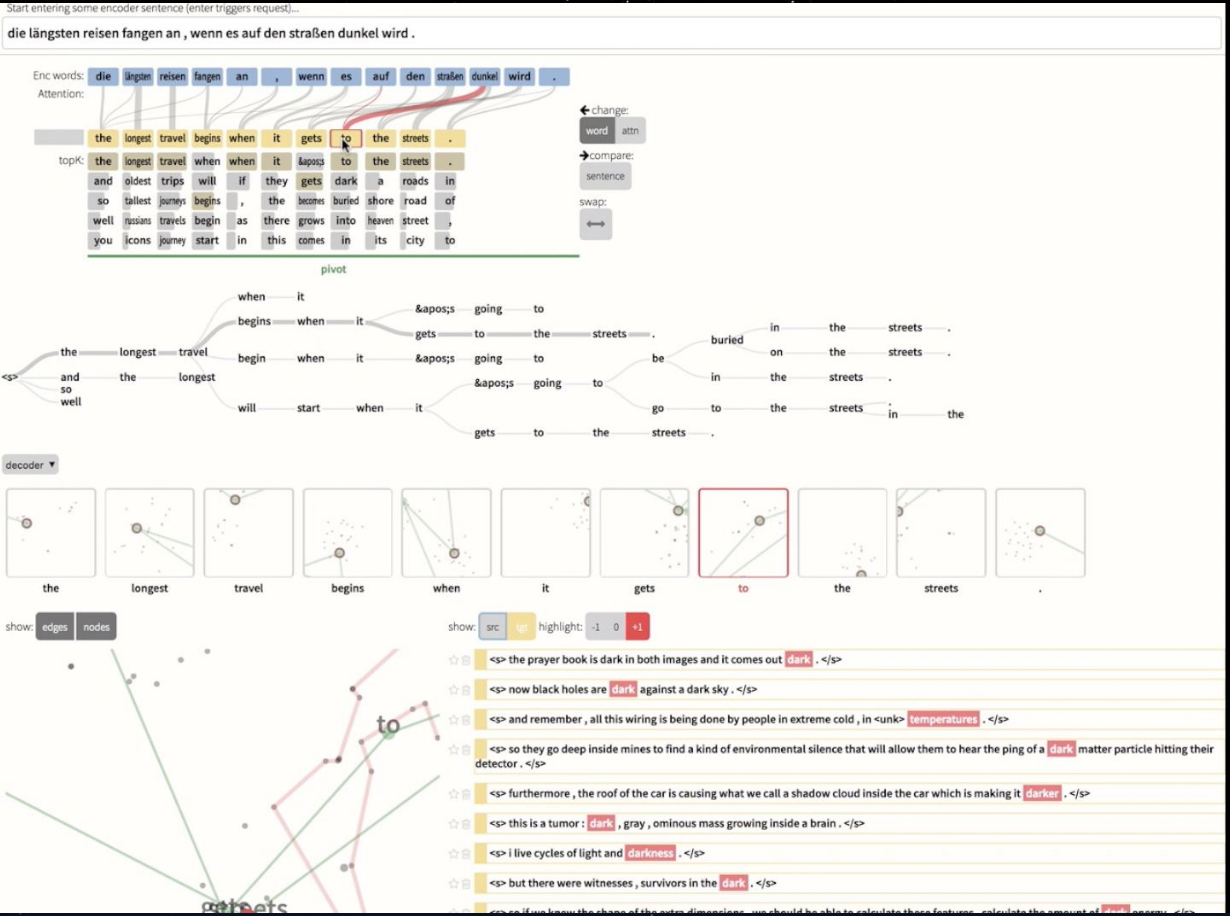
Acceptability + Performances



Stage I

The tools

Results Visualization
Model Exploration
Simulation

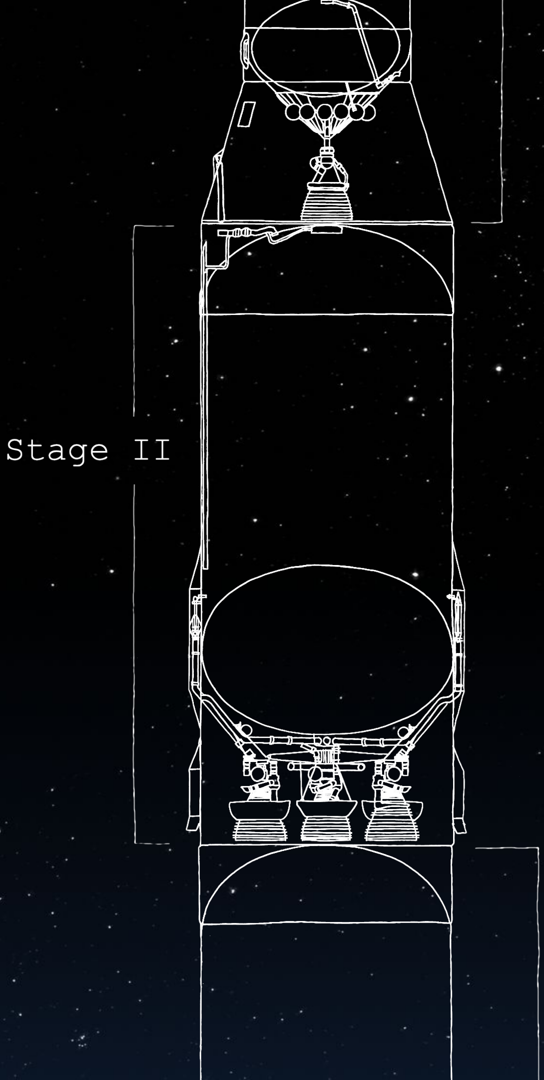


Stage I

Seq2Seq-Vis

[Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, Alexander M. Rush - arXiv 2018]

Translation system debugger & visualizer

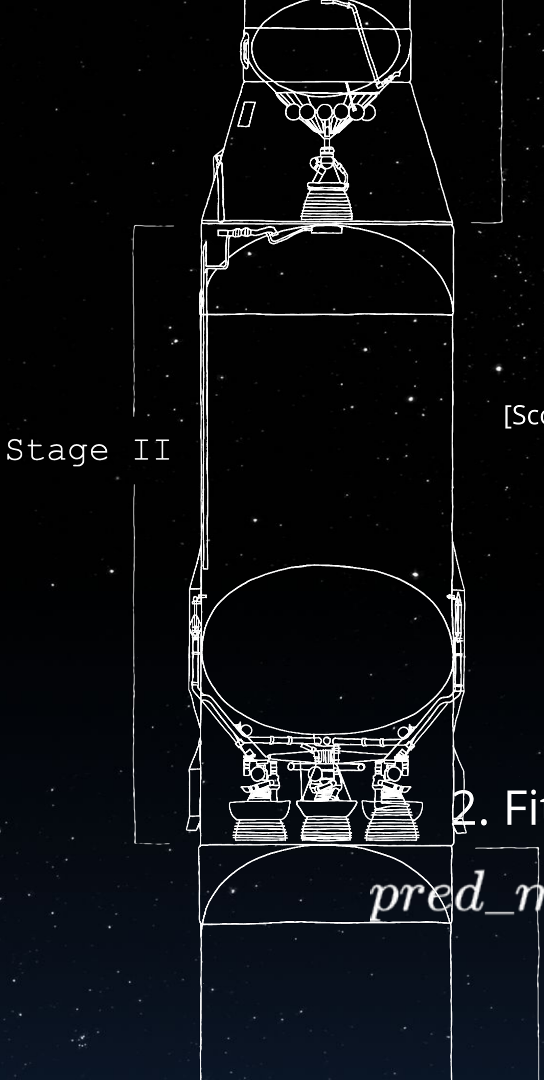


Stage II

Explainable decisions

Follow Least Surprise Principle

Trust + Traceability



Stage II

SHAP

[Scott M. Lundberg, Su-In Lee - A Unified Approach to Interpreting Model Predictions - NeurIPS 2017]

1. Select business understandable features

$$x' = simpl\_feat(x)$$

2. Fit a explainable model approximating the actual model

$$pred\_model(x) \approx expl\_model(x') = \phi_0 + \sum_{i=1}^{dim(x')} \phi_i x'_i$$

Stage II

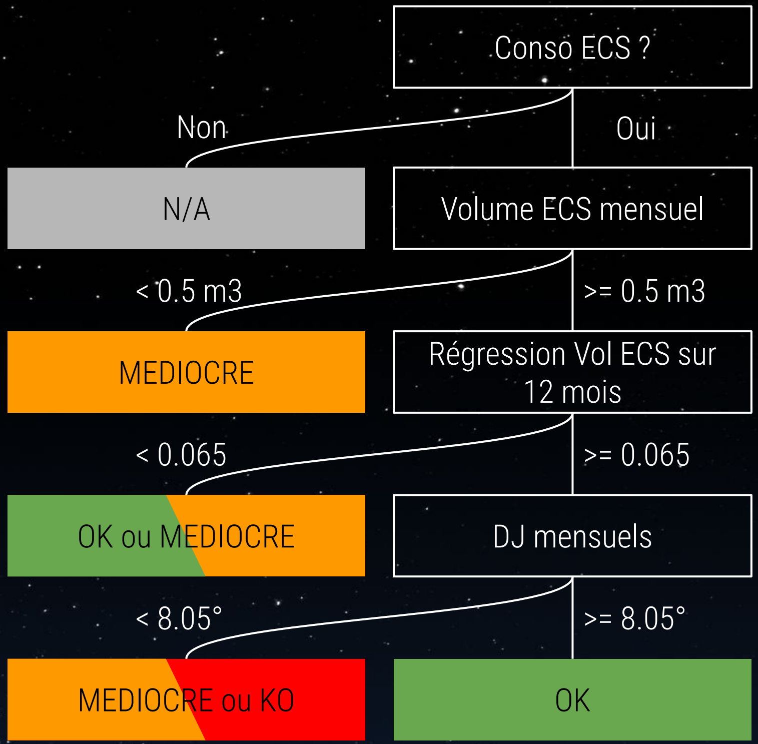
SHAP Usage example

Fraud pattern recognition

Explainability = Insights + Traceability

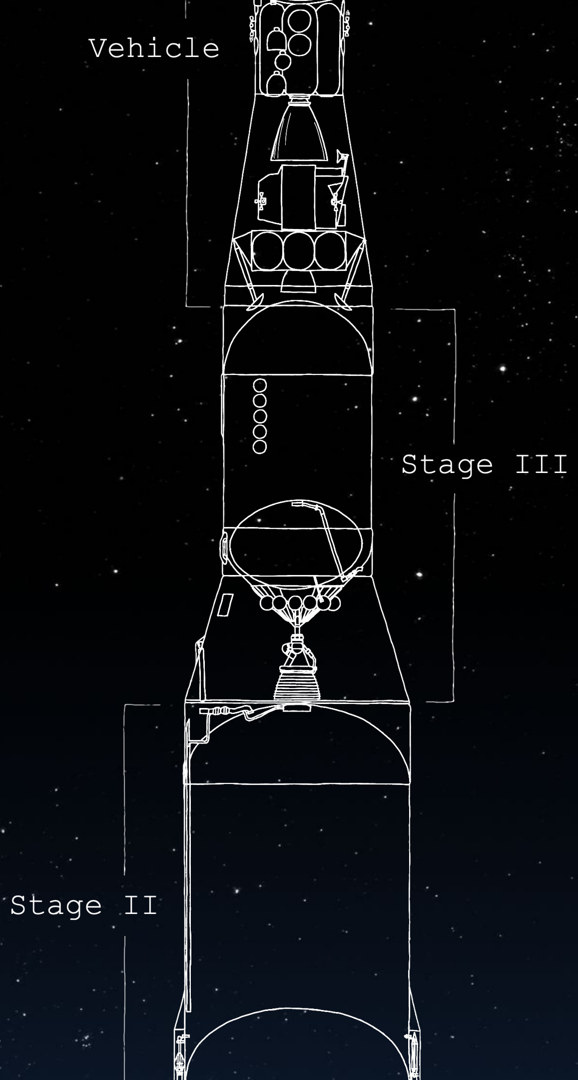
Stage II

Decision Tree to generate contrastive explanations



Energy manager assistant

Explainability = Productivity

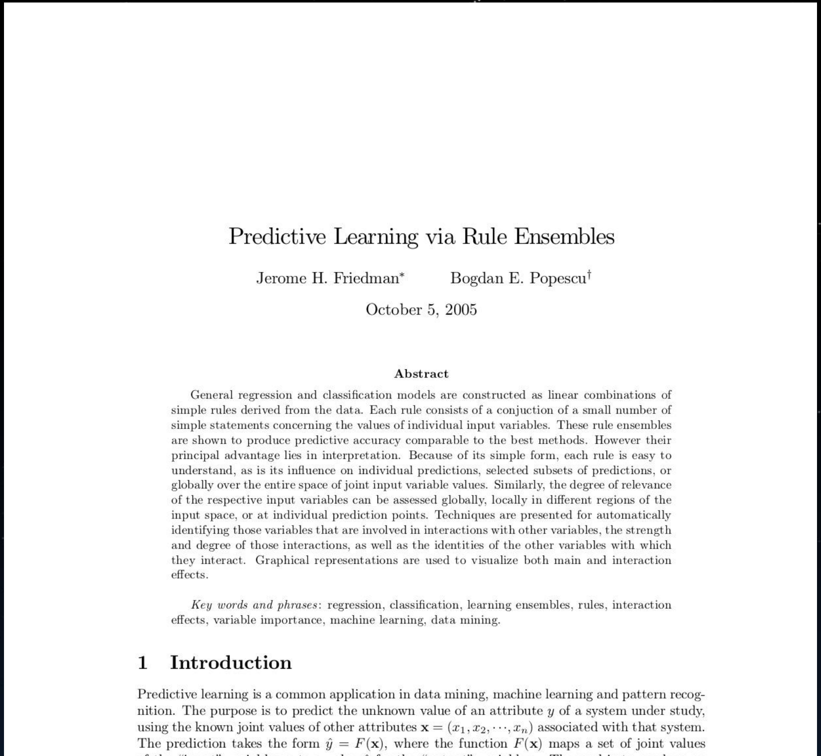


Stage III

Explainable decision process

Enable programmatic inspection of the decision process and underlying concepts

Automation + Certifiability



Predictive Learning via Rule Ensembles

Jerome H. Friedman* Bogdan E. Popescu†

October 5, 2005

Abstract

General regression and classification models are constructed as linear combinations of simple rules derived from the data. Each rule consists of a conjuction of a small number of simple statements concerning the values of individual input variables. These rule ensembles are shown to produce predictive accuracy comparable to the best methods. However their principal advantage lies in interpretation. Because of its simple form, each rule is easy to understand, as is its influence on individual predictions, selected subsets of predictions, or globally over the entire space of joint input variable values. Similarly, the degree of relevance of the respective input variables can be assessed globally, locally in different regions of the input space, or at individual prediction points. Techniques are presented for automatically identifying those variables that are involved in interactions with other variables, the strength and degree of those interactions, as well as the identities of the other variables with which they interact. Graphical representations are used to visualize both main and interaction effects.

Key words and phrases: regression, classification, learning ensembles, rules, interaction effects, variable importance, machine learning, data mining.

1 Introduction

Predictive learning is a common application in data mining, machine learning and pattern recognition. The purpose is to predict the unknown value of an attribute y of a system under study, using the known joint values of other attributes $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ associated with that system. The prediction takes the form $\hat{y} = F(\mathbf{x})$, where the function $F(\mathbf{x})$ maps a set of joint values

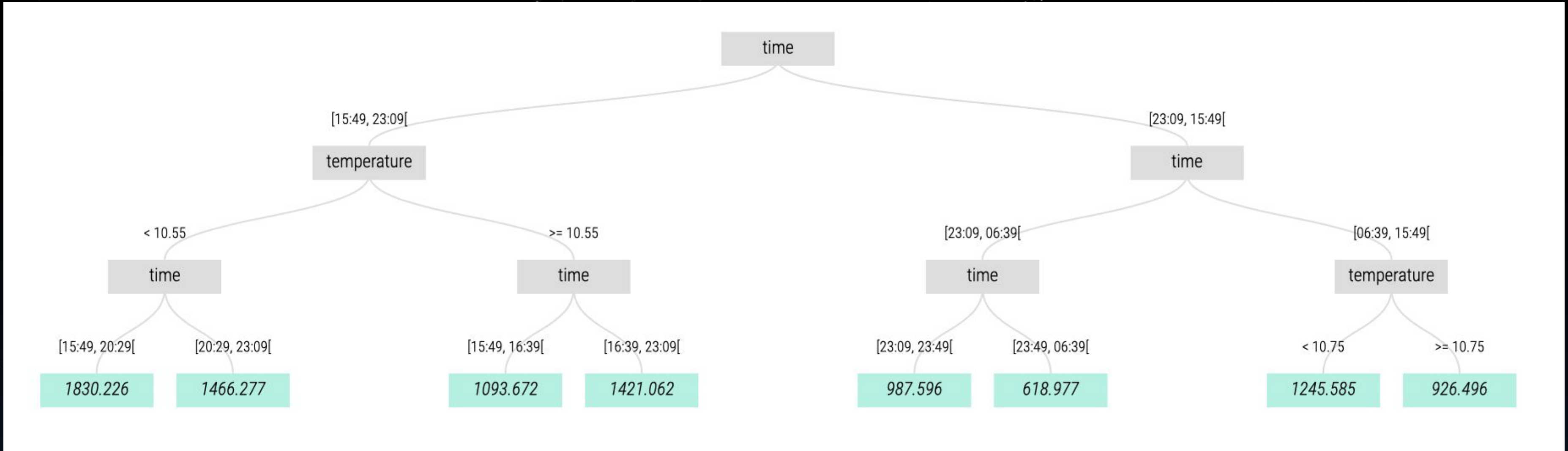
Stage III

RuleFit

Generate a minimal set of rules from a trained Random Forest

Stage III

Decision Trees



Stage III

Energy coach

Household consumption predictive model used as a knowledge base

Explainability = Interface with symbolic systems

Challenges

> Not because they are easy, but because they are hard

- John F. Kennedy

How can we evaluate explainability?

Challenges

> Not because they are easy, but because they are hard

- John F. Kennedy

Improve the performances of XAI

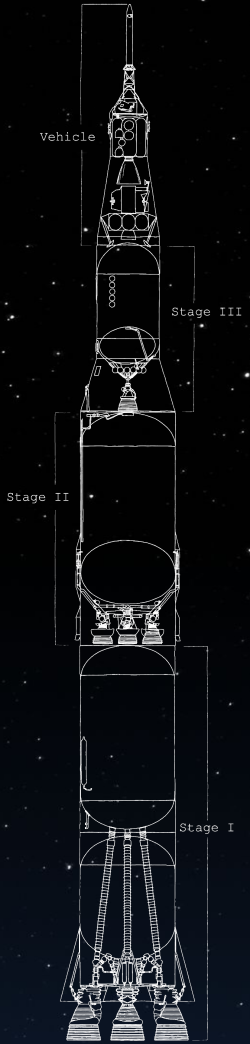
Improve Data Engineering
New Machine Learning approaches
Hybrid models

Takeaways

> We set sail on this new sea because there is new knowledge to be gained, and new rights to be won, and they must be won and used for the progress of all people.

- John F. Kennedy

Explainability is an opportunity
Regulation is coming
No system in production below stage II


Vehicle
Stage III
Stage II
Stage I


S-IV