

The 3 stages of

# XAI

How explainability facilitates real world deployment of AI

Explainable Artificial Intelligence

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

**DARPA** DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

EXPLORE BY TAG

ABOUT US / OUR RESEARCH / NEWS / EVENTS / WORK WITH US /

Defense Advanced Research Projects Agency > Program Information

## Explainable Artificial Intelligence (XAI)

Mr. David Gunning

**AI System**

- 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

→

**User**

- 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?

**RESOURCES**

[DARPA-BAA-16-53](#)

[DARPA-BAA-16-53: Proposers Day Slides](#)

[XAI Program Update](#)

Figure 1. The Need for Explainable AI

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/20 XAI Program Update November 2017*

# Intelligence Artificielle

Les défis actuels et l'action d'Inria



*Inria* | LIVRE BLANC | N°01

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*



CÉDRIC VILLANI

Mathématicien et député de l'Essonne

## DONNER UN SENS À L'INTELLIGENCE ARTIFICIELLE

POUR UNE STRATÉGIE  
NATIONALE ET EUROPÉENNE

Composition de la mission

**Marc Schoenauer** Directeur de recherche INRIA • **Yann Bonnet** Secrétaire général du Conseil national du numérique • **Charly Berthet** Responsable juridique et institutionnel du Conseil national du numérique • **Anne-Charlotte Cornut** Rapporteur au Conseil national du numérique • **François Levin** Responsable des affaires économiques et sociales du Conseil national du numérique • **Bertrand Rondepierre** Ingénieur de l'armement, Direction générale de l'armement.

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*



LUC JULIA

# L'Intelligence artificielle n'existe pas



Le cocréateur de *Siri*  
déconstruit le mythe de l'IA !

FIRST  
EDITIONS



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*

The image shows a browser window displaying the Kaggle website. The page is for the 'Machine Learning Explainability' course. The header is pink and contains the course title, a sub-header 'Extract human understandable insights from any Machine Learning model', and a 'Get Course by Email' button. Below the header, there is a 'Your Progress' section showing 0% completion and 'Begin today!'. An 'Overview' section lists course details: 'Free', '4 hrs.', and '5 Lessons'. A 'Prerequisite Skills' section is partially visible. The main content area is titled 'Lessons' and lists three lessons: '1 Use Cases for Model Insights', '2 Permutation Importance', and '3 Partial Plots'. Each lesson has a brief description and icons for 'Tutorial' and 'Exercise'.

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*





IJCAI 2019 Workshop on Explainable Artifi...

# IJCAI 2019 WORKSHOP ON EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

11 August, 2019. Macau, China

<https://www.ijcai19.org/>

# Ai?



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 de l'intelligibilité des modèles de machine



↑

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



[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



How XAI makes a  
difference?

# What's an explanation?

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

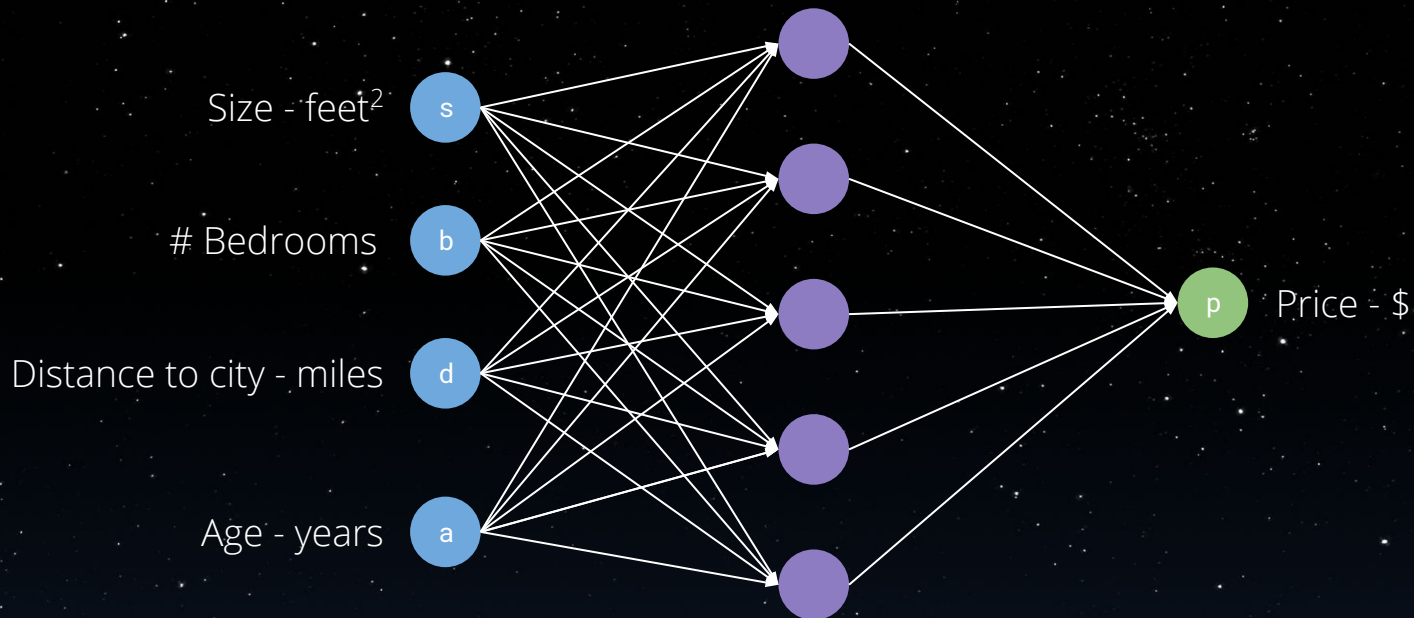
People look for explanations to improve their understanding of someone or something so that they can derive stable model that can be used for prediction and control

- Fritz Heider, Australian psychologist

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

# Non-explainable AI != Non-deterministic AI

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

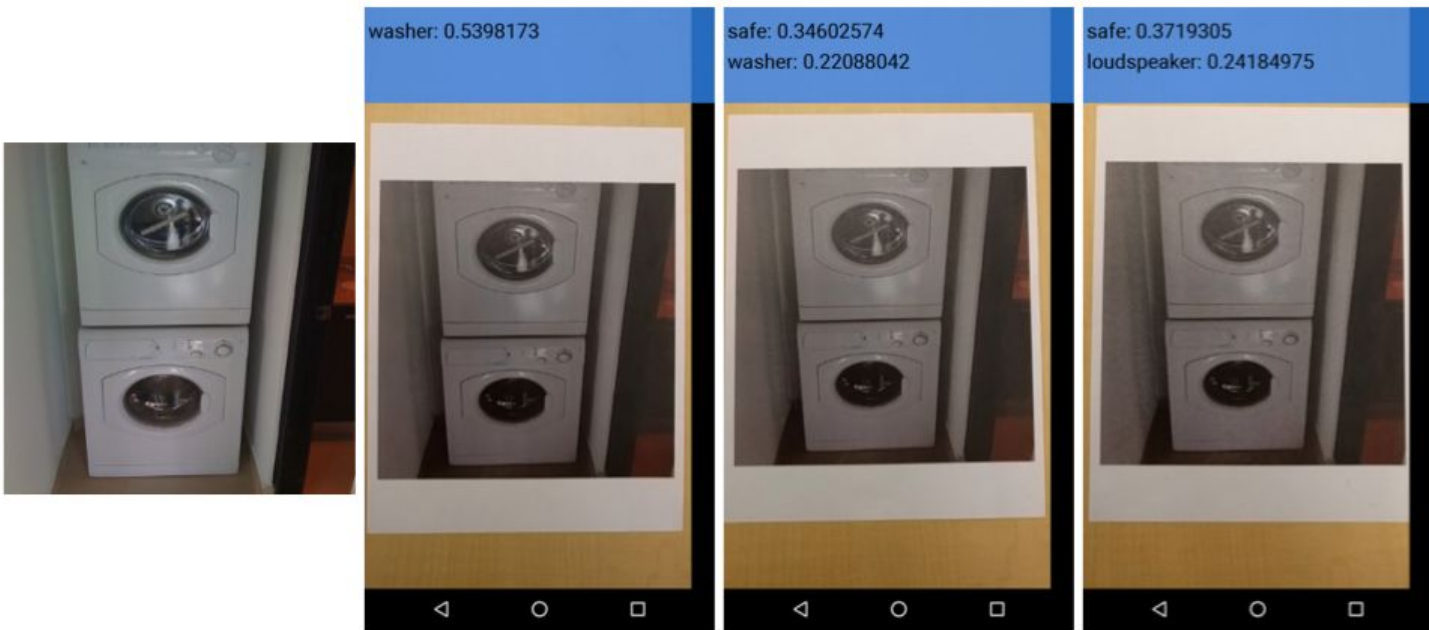


$$p = f(w_1 f(w_{11}s + w_{21}b + w_{31}d + w_{41}a) + w_2 f(w_{12}s + w_{22}b + w_{32}d + w_{42}a) + \dots)$$



# Non-explainable AI != Non-deterministic AI

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

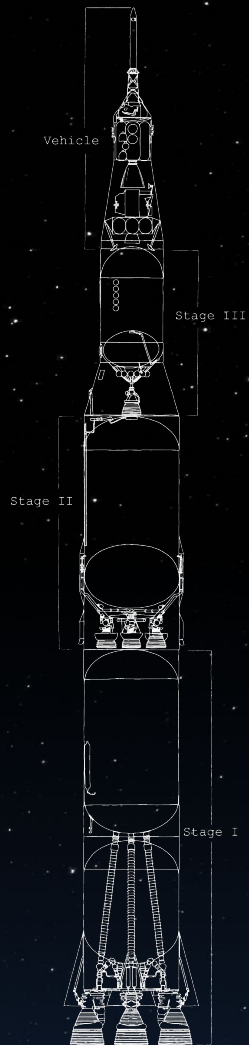


(a) Image from dataset

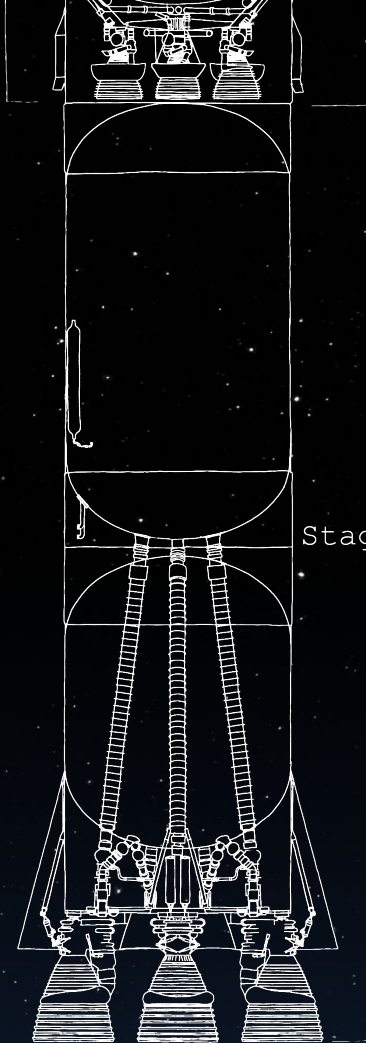
(b) Clean image

(c) Adv. image,  $\epsilon = 4$

(d) Adv. image,  $\epsilon = 8$



# The 3 stages of XAI



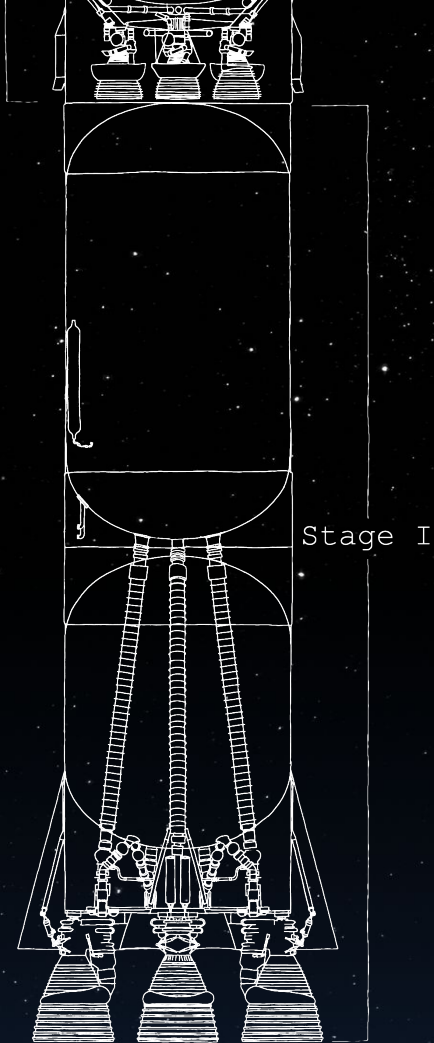
Stage I

Stage I

Explainable building process

Involve Business Experts

**Acceptability + Performances**



Stage I

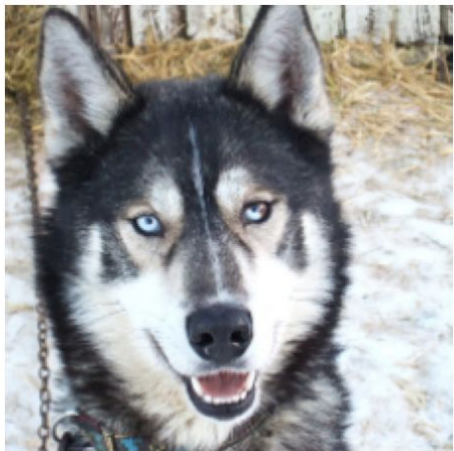
Some tools

Visualization

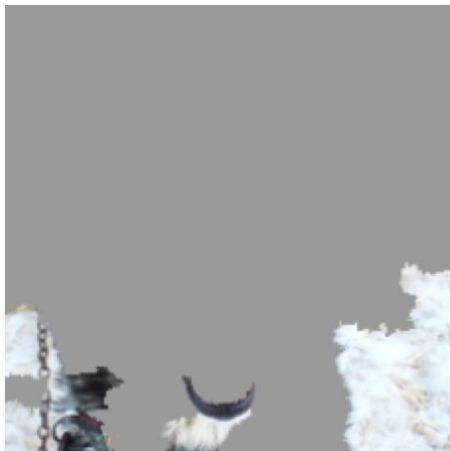
Simulation

Offline debugging tools

+ *upper stages tools*



(a) Husky classified as wolf



(b) Explanation

Stage I

## Saliency Maps

[Ribeiro, M. T., S. Singh, et C. Guestrin (2016)]

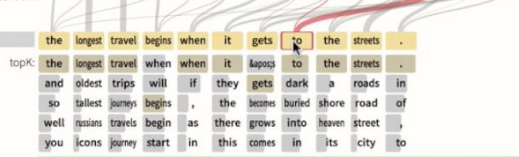
Visualize important  
regions of an image

Start entering some encoder sentence (enter triggers request)...

die längsten reisen fangen an , wenn es auf den straßen dunkel wird .

Enc words: die längsten reisen fangen an , wenn es auf den straßen dunkel wird .

Attention:



← change:

word attn

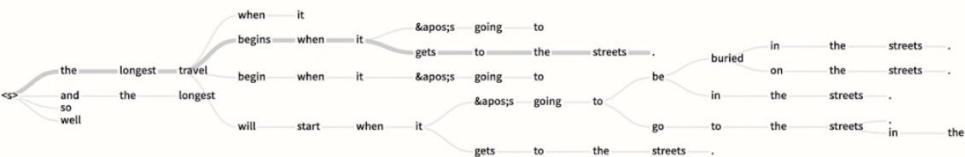
→ compare:

sentence

swap:

↔

pivot



decoder ▾



show:

edges nodes



show:

src tgt

highlight: -1 0 +1

- <S> the prayer book is dark in both images and it comes out **dark** . </S>
- <S> now black holes are **dark** against a dark sky . </S>
- <S> and remember , all this wiring is being done by people in extreme cold , in -cun- <b>temperatures</b> . </S>
- <S> so they go deep inside mines to find a kind of environmental silence that will allow them to hear the ping of a **dark** matter particle hitting their detector . </S>
- <S> furthermore , the roof of the car is causing what we call a shadow cloud inside the car which is making it **darker** . </S>
- <S> this is a tumor : **dark** , gray , ominous mass growing inside a brain . </S>
- <S> i live cycles of light and **darkness** . </S>
- <S> but there were witnesses , survivors in the **dark** . </S>

Stage I

# Seq2Seq-Vis

[Strobelt, H., S. Gehrmann, M. Behrisch, A. Pèrer, H. Pfister, et A. M. Rush (2018)]

Translation system  
debugger &  
visualizer



Stage II

Stage II

Explainable decisions

Follow Least Surprise Principle

**Trust + Traceability**



Stage II

Stage II

Some tools

Local explanation generation  
(TreeInterpreter, LIME, SHAP...)

[Saabas, A. (2014); Ribeiro, M. T., S. Singh, et C. Guestrin (2016a); Lundberg, S. M. et S.-I. Lee (2017)]

+ *upper stages tools*





Stage II

## Stage II SHAP

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

1. Select business understandable features  
 $x' = \text{simpl\_feat}(x)$

2. Fit an explainable model approximating the actual model

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



BLECKWEN

18/04/2019 - 07:28 | Alert creation: fraud suspicion

⚙️ An alert has been created for suspicion of fraud.

#### RULES FEEDBACK

❌ ml

❌ RULE\_MODEL

MODEL\_SCORE > 0.5 => BLOCK

[See all rules](#)

#### MACHINE LEARNING FEEDBACK

78 %



[See all the features](#)

Stage II

# Banking fraud detection

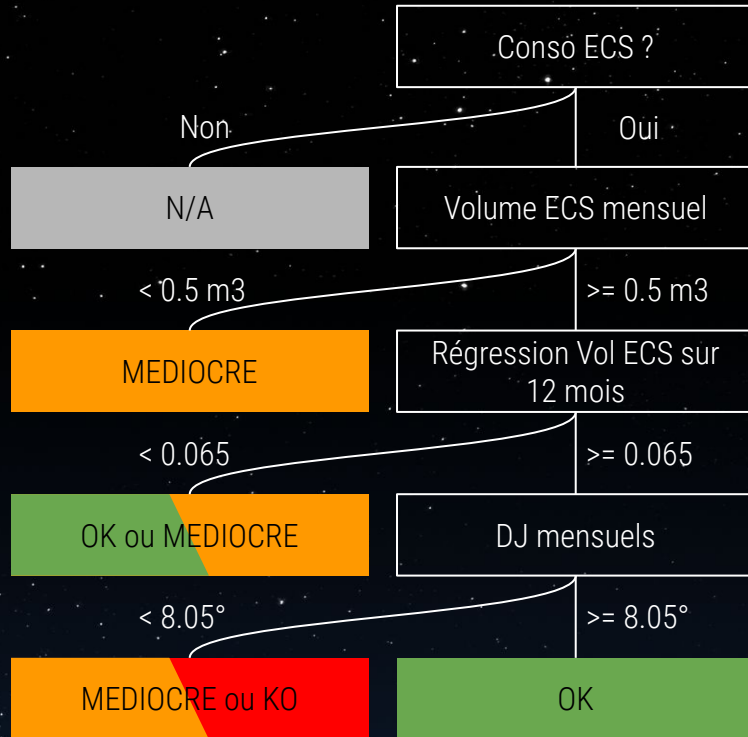
## Leveraging SHAP

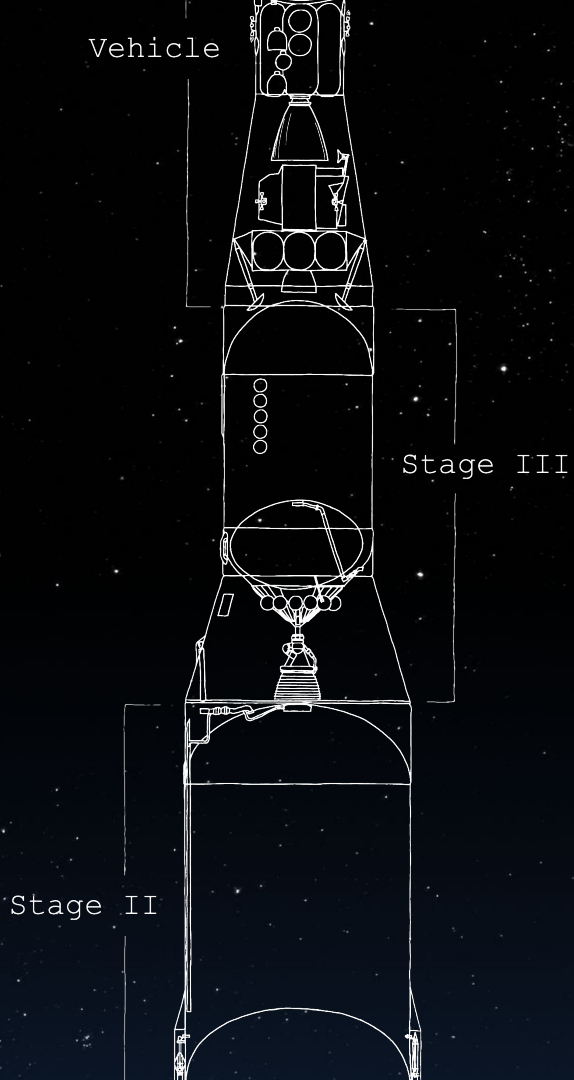
# Explainability = Insights + Traceability

# Energy manager assistant

Decision Tree to generate  
contrastive explanations

**Explainability = Productivity**



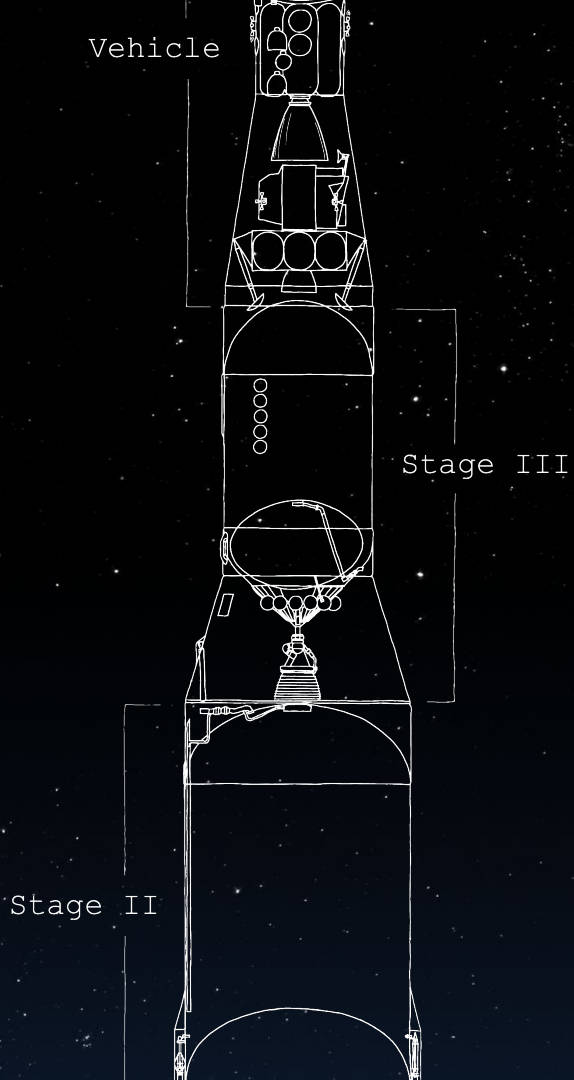


Stage III

**Explainable decision process**

Enable interoperability with business logic

**Automation + Certifiability**



Stage III

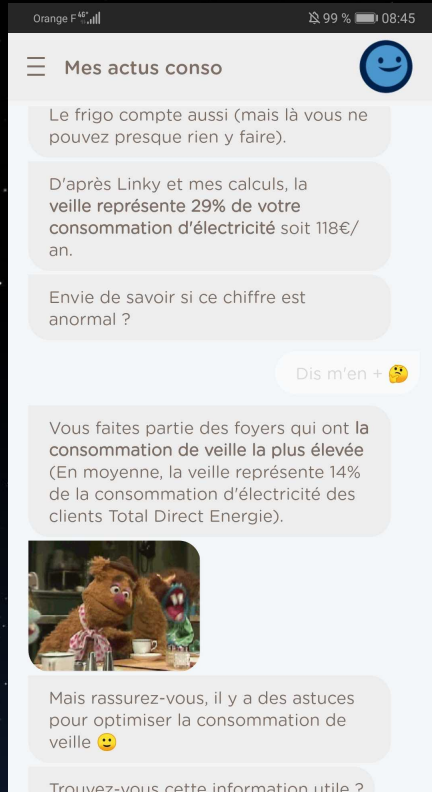
Some tools

Globally explainable models  
(Decision Trees, Business Rules,  
Regressions, ...)

# 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?

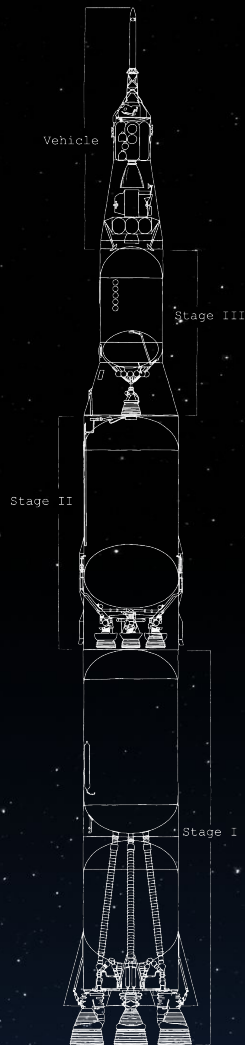
# Challenges

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





## Stage III

# Explainable decision process

Enable interoperability with business logic

---

## Stage II

# Explainable decisions

Foster trusts with users & supervisors

---

## Stage I

# Explainable building process

Facilitate acceptance

# 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

