January 2022

# ACPR Tech Sprint on the explainability of artificial intelligence

## Summary report

Author: Laurent Dupont, Fintech-Innovation Hub, ACPR


*A participant summarises the challenges proposed by the ACPR's Tech Sprint on AI explainability.*

The ACPR held its first Tech Sprint in June-July 2021. The challenge: generating explanations to understand the behaviour of credit risk predictive models based on artificial intelligence (AI) and only accessible as "black boxes"[1].

The ACPR's Fintech-Innovation Hub acted as creator, organiser and facilitator of the event – sometimes known as a regulatory hackathon. In doing so, it collaborated with 4 voluntary credit institutions (the "Partners") which designed and trained machine learning (ML) models on an agreed-upon use case, namely predicting which loans to individual customers are likely to default[2].

Tech Sprint participants included professionals from fintechs, banks or other financial actors, as well as researchers and students in data or computer science. Those participants teamed up to play the role of "Analysts". Their primary task was explaining the behaviour of the predictive models and elucidating their nature and characteristics.

This report describes the origin of the Tech Sprint, the event itself, then summarises the key lessons learned by the ACPR from the challenge, and finally outlines potential further work on AI explainability and adjacent themes.

# Table of Contents



[1] The term "black box" designates a model whose internal workings are hidden from the observer, and for which only the input (in this case, the credit application or current loan) and output data (e.g. the predicted probability of default) are visible. By extension, "black box" also applies to observable models whose design is too complex to fully grasp (typically, a deep neural network in contrast with a linear regression model).

[2] More precisely, the models' prediction consists of either directly estimating their probability of default, or computing a proxy indicator (in this case, the individual's presence in a credit risk registry).

# 1. Design of the Tech Sprint

## 1.1 The 2020 report

The Tech Sprint followed in the footsteps of the Fintech-Innovation Hub's prior work on AI, primarily the discussion document on AI governance in finance, which was published by the ACPR in June 2020 and was accompanied by a public consultation. The document focused around two areas, namely the evaluation and the governance of AI algorithms.

Our work led to identifying four interdependent criteria in the design and evaluation of AI algorithms and tools in finance, which aimed to help introduce AI into business processes while limiting associated risks: appropriate data management, system performance, stability, and explainability.

Furthermore, as the introduction of AI in finance inevitably affects the governance of associated procedures, we recommended to particularly focus, as early as the algorithm's design phase, on its integration into business processes, on interactions between human and algorithm, on security and outsourcing aspects, on the initial and continuous validation processes, and on internal or external auditability of such systems.

The public consultation associated to the document, whose results were summarised and published in December 2020, confirmed the relevance of those design and evaluation principles, but also of the proposed governance principles.

## 1.2 The explainability topic

Among the AI design and evaluation principles, explainability deserved particular attention, as it is the most distinctive feature of AI and – when properly used for both internal control and external audit – it underlies controlled adoption of AI in finance.

**Definition**

For the purpose of this work on AI, explainability encompasses two questions. On the one hand the "why" i.e. the question of transparency: the main associated issue is auditability. On the other hand the "how" i.e. the question of interpretability, which affects the intelligibility of the system's behaviour by human operators interacting with it and by customers, as well as social or ethical acceptance.

The question of explainability arises very concretely as early as in the design phase of an AI component, with a continuum between two extremes, namely black box AI and fully-interpretable AI[3].

[3] This idea is promoted for example by Cynthia Rudin who advocates for avoiding black boxes, which is often an ideal situation but unreachable in practice. See "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." Cynthia Rudin (2018).

**Explanation levels**

The 2020 discussion document proposed the following 4-level scale to characterize an algorithmic explanation:

**Level-1 explanation: observation**

Such an explanation answers technically-speaking the question *"How does the algorithm work?"* and functionally-speaking the question *"What is the algorithm's purpose?"* This level can be achieved:

- Empirically, by observing the algorithm's output (individually or as a whole) as a function of input data and of the environment
- Analytically, via an information sheet for the algorithm, the model, and the data used, without requiring the analysis of the code and data themselves.

**Level-2 explanation: justification**

Such an explanation answers the question: *"Why does the algorithm produce such a result?"* (in general or in a specific situation). This level can be achieved:

- Either by presenting in a simplified form some explanatory elements from higher levels (3 and 4), possibly accompanied with counterfactual explanations.
- Or by having the ML model itself it has been trained to produce.

**Level-3 explanation: approximation**

Such an explanation provides an – often inductive – answer to the question *"How does the algorithm work?"* This level of explanation can be achieved, in addition to level-1 and 2 explanations:

- By using explanatory methods which operate on the model being analysed.
- Via a structural analysis of the algorithm, the resulting model and the data used. This analysis will be all the more fruitful if the algorithm is designed by composition of multiple ML building blocks (hyper-parameter tuning or auto-tuning, ensemble methods, boosting, etc.).

**Level-4 explanation: replication**

Such an explanation provides a demonstrable answer to the question *"How to prove that the algorithm works correctly?"*

This level of explanation can be achieved, in addition to level-1 to 3 methods, by detailed analysis of the algorithm, model and data. In practice, this is feasible only by doing a line-by-line review of the source code, a comprehensive analysis of all datasets used, and an examination of the model and its parameters.

**Recipients of an explanation**

One of the main factors dictating the expected explanation level is the type of recipient targeted. This is because the relevant form under which an explanation should be proposed in order to be effective depends both on their technical and business proficiency and on their intrinsic motives for demanding an explanation. Hence different explanation levels could be applied to the same algorithm, depending on whether the explanations serve an end user (who tries to check that they have not been treated unfairly by the system, and for whom an explanation has to be intuitively intelligible) or an auditor (who needs to understand the system's technical architecture in detail and who is subjected to rigorous regulatory requirements).

Our discussion document thus recommended adapting the objectives of an explanation (and thus, its form and its content) to the audience considered:

- For human operators who interact with the AI system: to understand its behaviour
- For individuals affected by the system's predictions or decisions (such as customers in a sales context): to understand the underlying motives
- For those who designed the system or are tasked with checking its compliance: to assess its social and ethical acceptability, in order (among other things) to prove the absence of discriminatory bias in its decision-making process.
- For internal control procedures led by the regulated firms themselves on such systems, but also for external audits conducted by supervisory authorities: to inspect systems containing at least one AI-based model, some of which may only be exposed as a "black box".

## 1.3 The credit risk use case

The use case of credit risk predictive models is relatively common among ML hackathons. In the case of the Tech Sprint, the choice nevertheless quickly converged on retail credit risk due to the many high-stake topics involved: financial stability, public consumption, commercial objectives, and socio-economic issues such as financial inclusion.

The European Commission published in the interim (April 2021) its proposal for an AI regulation. The proposal offered a 4-level risk scale and mentioned credit risk modelling as the only use case in the financial sector categorised as "high risk", which made that use case all the more relevant for the ACPR.

## 1.4 Main stakes

The main goal pursued by the ACPR in organising the Tech Sprint in summer 2021 was thus to shed light on regulatory challenges linked to AI and ML, from risk management to process governance and consumer protection.

From a technical perspective, the Tech Sprint aimed at exploring which explainability methods apply to a concrete use case, and which kinds of explanations are most suitable to a variety of stakeholders (auditors, domain experts, technical staff, or customers). It also contributed to the ACPR's policy of promoting knowledge sharing and fostering collaboration initiatives among actors of the financial sector.

In parallel, the Fintech-Innovation Hub had started considering how to audit an AI-based system during both on-site and off-site supervisory missions – a very nascent field of research for which a practical application scenario appeared necessary.

# 2. The event

## 2.1 Objectives

**Main objective: explaining the behaviour of the models**

The primary objective of Tech Sprint participants was to make each model's behaviour intelligible, i.e. both why a particular prediction is made and the overall behaviour on all possible input data.

**Bonus task: assessing algorithmic equity of the models**

A secondary objective, presented to the participants as a bonus question, was to assess each predictive model's fairness, i.e. to describe and quantify the different types of statistical or other biases they contain, as well as how they reflect or reinforce any undesired biases already present in the data.

**Non-objectives: evaluating performance and soundness of the models**

In order to avoid any misunderstanding, the Analyst Guide also indicated that unlike most traditional hackathons, the measure of model performance (i.e. how accurate they are) and the evaluation of their soundness (i.e. whether their predictions are justified and the resulting decisions can be contested) were not part of the Tech Sprint objectives: they were qualified as "non-objectives".

## 2.2 Programme

The preparatory phase of the Tech Sprint lasted about 6 months, during which the Fintech-Innovation Hub and the Tech Sprint partners collaborated on designing and training the predictive models. Those models were selected for their diverse structures and varying levels of complexity, while remaining representative of models currently or soon-to-be deployed in production by banking institutions.

Eventually, the ACPR packaged them as black boxes and deployed them for the Tech Sprint: after each model had been trained by the partner bank on a real dataset, they were installed on Banque de France's internal cloud. The participants would then have no direct access to the models nor to their training data, but only to an API enabling them to query each model using a set of characteristics for one or more customers, which would return the probability(ies) of default estimated by the model. It should also be noted that each partner bank had provided a test dataset, containing anonymised but realistic data, so that participants were able to tackle the challenge without spending too much time generating synthetic data for querying the models.

The Tech Sprint event itself consisted of two sessions, enabling both professionals and students to tackle the explanation challenge. Each session spanned two days.

The first day was set up as a typical hackathon, featuring activities related to its theme such as a talk on the socio-cognitive factors involved in AI explanations. On the second day, each team of analysts prepared and performed the presentation of their work in front of a jury composed of executives from Banque de France and the ACPR. Presentations were limited to 5 minutes[4] but their format was left to the creativity of each team: slideware, interactive demo, or in one team's case a theatrical performance of the explanations produced by their method.

Within each session, the 3 teams with the most convincing works were rewarded by the jury on the basis of the following criteria:

[4] Thus longer than the '240 seconds of glory' format used for NASA's Space App Challenge, but only slightly since a dozen teams were presenting their results in each session.

- Their technical achievements
- Their methodological and scientific innovation
- The quality of the presentation of their works
- Their contribution to regulatory and business issues (for example how to aid credit risk experts in interpreting the algorithm's decisions, or facilitating their maintenance by bringing to light unexpected behaviour).

## 2.3 The challenge

**A road paved with obstacles**

The difficulty of the competition proposed by the Tech sprint cannot be overstated. The challenge had indeed been designed under a triple perspective: technical (a generic explainability challenge), functional (the practical use case of retail credit risk models), and regulatory (the black box audit situation).

The Tech Sprint therefore presented various kinds of obstacles, including the following:

- The latency time due to model hosting and construction (i.e. round-trip time plus prediction time) may affect the explanatory method used by limiting the maximum number of predictions per time unit.
- Test data, which are imperfect in terms both of quality of completeness (precisely because test datasets provided by the partner banks were representative of real-world data), raise the question of synthetic data generation.
- Available explanatory toolkits present some limitations. Common solutions such as DiCE and SHAP are very costly in computation time, a phenomenon compounded in the case of a complex model, which is less interpretable than simple models while having a potentially higher inference time.
- A certain level of domain expertise is necessary to conceptualise explanations (i.e. to produce explanations using higher-level concepts[5] than the sole predictive variables).
- The black box audit situation may have proven the most challenging as it precludes any knowledge of the training algorithms for the models considered.

**Multidisciplinary approach**

The Tech Sprint demonstrated the practical benefits of adopting an interdisciplinary approach. Indeed the teams combining data science and data engineering expertise along with data visualisation skills delivered the most complete and accurate explanations, as well as the most persuasive visual renderings of those explanations.

Further, familiarity with socio-cognitive concepts involved in information processing can prove particularly useful in tailoring explanations to different audience types.

**Unveiling the black boxes**

Within each Tech Sprint session, the content of each black box was unveiled to the entire audience, right after all participants had presented their works so as to keep the mystery intact up until that time.

As previously mentioned, the models studied represented a variety of algorithm classes (from linear models to neural networks, including model stacking), of input data schema, and of hosting setups (on

[5] Understood here as concepts meaningful to an expert of the business domain (in this case retail credit, more particularly in the context of a specific banking institution), as opposed to lower-level concepts suitable to technical experts like data scientists (who manipulate the "raw" predictive variables).

the Banque de France internal cloud or in an external environment). The models were nevertheless realistic and representative of production or experimentation systems within partner banking institutions.

The content of each black box can be summarised as follows:

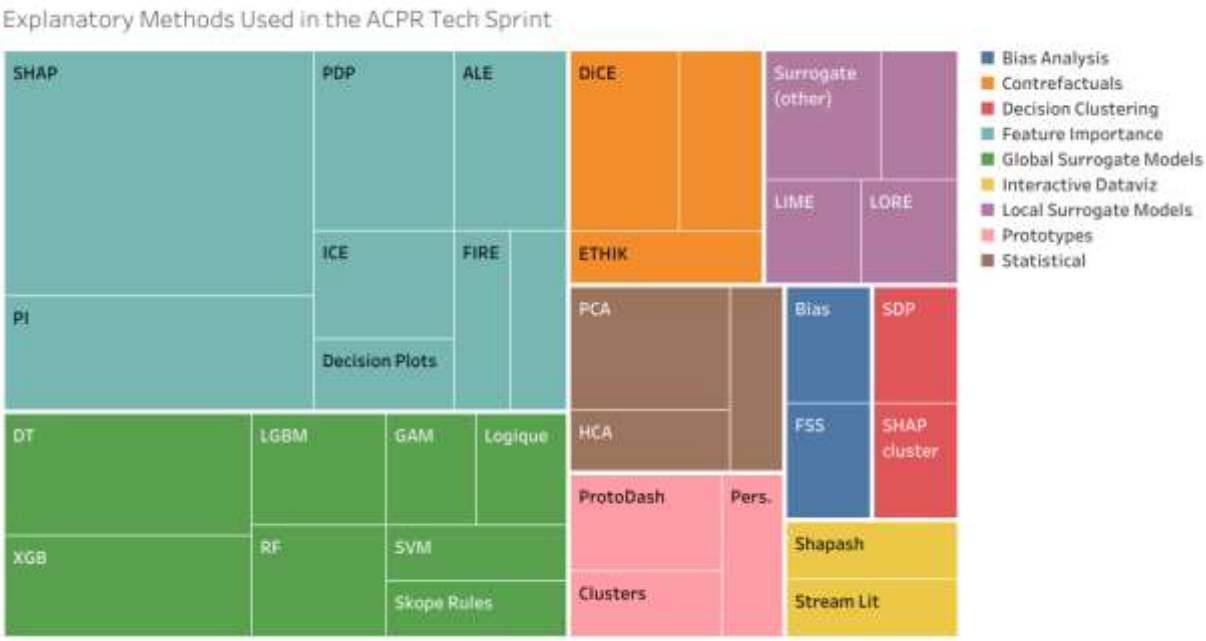
- Black box A contained three models with 7 predictive variables: one XGBoost model, one MLP (multi-layer perceptron), and a model combining one XGBoost and a random forest using a logistic regression.
- Black box B contained three models with 63 predictive variables: a regulatory model based on stacking multiple logistic regressions, and a pair of random forests for which a comparative analysis was suggested to participants so as to determine what distinguished them (namely a bias introduced in their training data).
- Black box C contained one model with 322 predictive variables: a random forest.
- Black box D contained two models with 8 predictive variables: a simple GBT (Gradient Boosted Tree) and an aggregated model of 3 GBT. A functional difference in that black box was the target variable, which aimed to predict an individual's being registered in a credit risk file maintained by Banque de France.

Besides, models also differed from one another by their treatment of missing input variables: a more or less rudimentary imputation was operated on all variables as appropriate, furthermore the list of actually used predictive factors was provided to analysts but not for all models, in which case an extra challenge consisted in trying to infer those factors.

# 3. Explanatory methods

## 3.1 Methods used to explain a model

The following diagram shows the explanatory methods used during the Tech Sprint.




*1. Explanatory methods used in the Tech Sprint*

Participating teams thus deployed a broad range of methods, wherein post-hoc model-agnostic techniques are well represented. Most prominent among those was SHAP, which seduces by its plug-and-play nature and its relatively intuitive visual rendering – even to non-experts – at the cost of the aforementioned computational cost. The ubiquity of SHAP visualisations may also yield an impression of *déjà-vu*, and its methodological interest is also downplayed by several experts, as evidenced by the following quote:

> *"Mathematical problems arise when Shapley values are used for feature importance and that the solutions to mitigate these necessarily induce further complexity, such as the need for causal reasoning […] Shapley values are not a natural solution to the human-centric goals of explainability."* Elizabeth Kumar (2020)

The most commonly used category (which includes SHAP) is feature importance: such methods enable to rank variables from the most to the least predictive, or to visualise the impact of a variable or pair of variables on the predicted value (in this case, the probability of default).

Tech Sprint participants often resorted to global surrogate models. Such models are trained to approximate the predictions of the original black box model, while being more easily interpretable by their very nature. The preponderance of this choice is somewhat surprising insofar as training such models requires high availability of the API, and does not always result in a net gain if the underlying black box model turns out to be a simple one (for example a decision tree). It may however be explained by the overall difficulty met by analysts for inferring the class of models contained in the black boxes, which prevented the use of more targeted methods.

Analysts also used local surrogate models. Those methods, which include the well-know LIME technique, have the drawback of being relatively unfaithful to non-linear or non-continuous models.

The use of several methods based on prototypes should be noted, which represent in a synthetic manner a subset of the entire domain of possible input data. In the specific application scenario considered, namely credit risk models, this usage is in line with the principle of segmentation of retail credit consumers. Contrarily, methods based on "criticisms" – i.e. those outlining areas which are poorly represented in input data – were neglected, although they may have revealed interesting anomalies in low-density areas of the input domain.

Several purely statistical methods also deserved a mention, such as principal component analysis techniques as well as a measure of permutation feature importance with an unusual implementation. This shows that it is possible – and sometimes relevant – to recreate existing methods, either from scratch or by adding variations.

Decision clustering methods, such as SDP (Same Decision Probability) or SHAP clustering, were also used. Their benefit will discussed in the "Intelligibility principle" section.

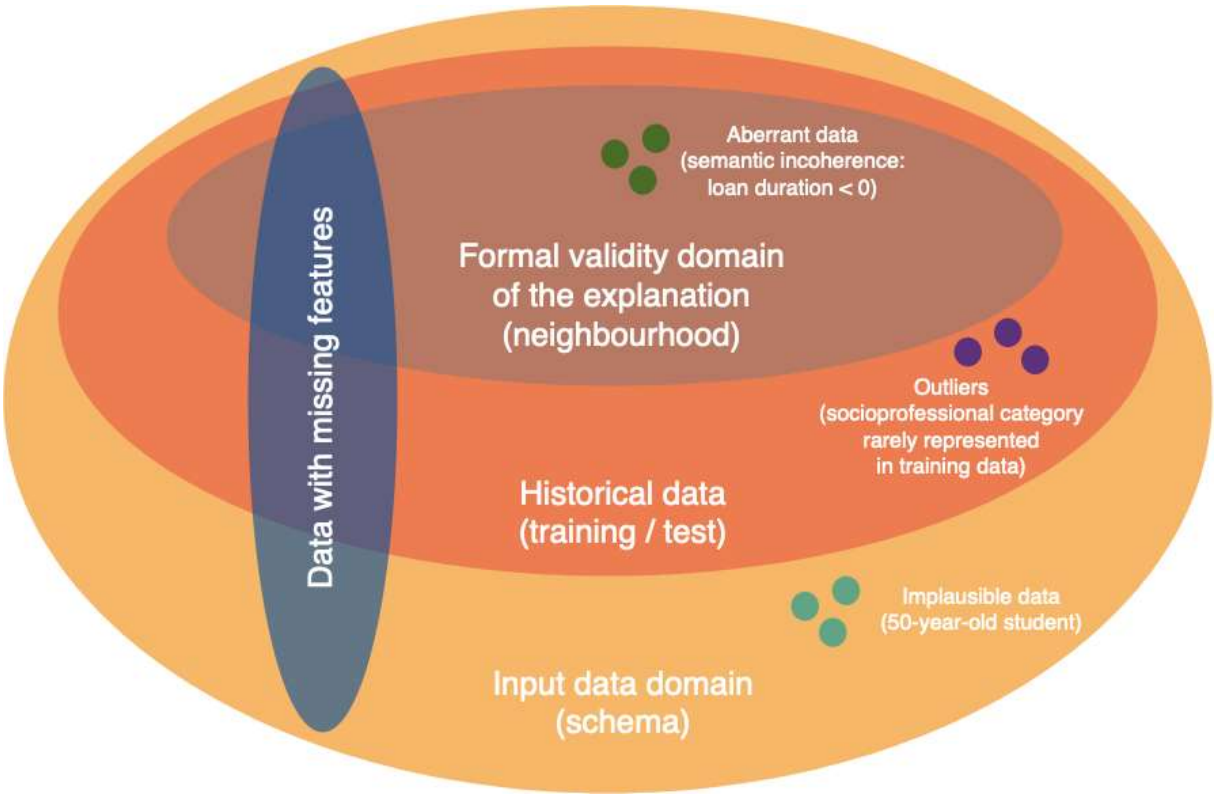
Lastly, with respect to efficiently delivering explanations to their audience, participants put the emphasis on producing explanations which are as intelligible as possible, rather than on the visualisation of those explanations – including interactive ones. This last point will be discussed in the "User interaction principle" section.

### 3.2 Explaining a model on "any" input data: generalisation power

The generalisation power of explanations produced is an uncontroversial goal of an explanatory method. Generalisation here means trying to explain a model's behaviour on any kind of input data imaginable. The question however remains of where to put the cursor on this requirement:

a. Should an explanation be required to remain valid when staying within the training data statistical distribution, or also in never-observed regions?
b. Should the model behaviour also be explained on incoherent or semantically aberrant data points?
c. Should an explanation cover unrealistic data?
d. Should it cover data points with missing attribute values?

The following figure summarises the various possible perimeters for the generalisation power of an algorithmic explanation:




*2. Possible validity perimeters for an algorithmic explanation*

This issue was tackled in some participants' works. They posed subtle questions such as whether an ideal explanatory method should be applicable to atypical data points (of the kinds b and c above): they addressed for instance the question of data points implausible from a business perspective (such as a 50-year-old student), which in some cases yielded significant effects.

A problem related to generalisation power, which participants also pointed out, is the lack of realism in the data used by some methods, such as PDPs (Partial Dependence Plots). Indeed, by relying on marginal distributions, PDPs hide the actual input data distribution (which may lead to overemphasising the model impact on low-density regions), furthermore they assume conditional independence between the predictive variable under investigation and the remaining variables.

### 3.3 Explaining beyond the model

Tech Sprint participants primarily focused on explaining each model's behaviour, by answering questions such as:

- Why was a particular individual prediction made?
- How can this individual prediction be changed?
- How does the model behave in certain regions (anomalous, low-density, etc.)?
- How do perturbations (minor changes) in input data affect the output prediction?

They also avoided answering the question of justification ("Where is the model right / wrong?") and rarely tackled the bonus question of algorithmic fairness ("Is the model fair or discriminatory?").

However, besides explaining the model itself, the Analyst Guide given to participants hinted at the relevance of explaining the training data, as well as the model building process. The remainder of this section shall summarise the works produced to this aim and the difficulties encountered.

**Explaining the training data**

The question here is what can be inferred about data used to train the model (their value domain, their statistical distribution, etc.) including for variables which have little or no predictive value.

In this regard, analysts mostly computed feature importance values, without deducing any further knowledge about training data. At best, they gleaned interesting clues, e.g. assuming (without being able to prove it) that surprising behaviours – such as a risk plateau beyond 10 years of loan duration – were linked to a low training data density.

Besides, in order to propose a challenge in training data explanation, a pair of models was proposed wherein the only difference was in their respective training datasets: defaulted loans without any prior banking incident were oversampled in model B3 compared to model B2. This difference could be inferred from the lower importance of incident features in model B3, although this last behaviour could also be attributed to a structural difference between both models[6]. This extra challenge proved too tricky to solve within the time imparted to the Tech Sprint.

**Explaining the machine learning procedure**

The question here is what can be inferred about the ML algorithm: its nature (is it a linear regression, a neural network, etc.), its general composition (number of parameters and hyperparameters), and other detailed composition (value of parameters and hyperparameters).

Various methods, often quite creative, have been tested by participants to this aim: visualisations such as PDP and ICE (Independent Conditional Expectations) to detect decision trees, "hacking" of the black box documentation to deduce its content, etc.

The main takeaway is that inferring the ML algorithm is extremely difficult for a black box model. One can only guess that a reliable, efficient solution (which to the best of our knowledge is not yet available) will likely combine complex heuristics in order to be applicable to the most common algorithm classes.

## 3.4 Methodological takeaways

This section lists the main lessons learned from the ACPR's first Tech Sprint.

These takeaways may be summarised as follows: an audit mission focused on an AI algorithm requires assembling a team equipped with a variety of methods in its toolbox, and maintaining an agile approach over the entire duration of the mission – especially where black box ML models may be involved.

**Build vs. buy explainability**

The Tech Sprint showed that the "build vs. buy" dilemma, well known in software engineering, is also a valid question for AI explainability.

[6] This alternative hypothesis could in turn be disproved via a comparative feature importance analysis between the two models.

Indeed several teams, who represented small or medium-size technology providers, tackled the challenge using their own data science platform. Such platforms usually enable performing more or less a “one-click” model analysis. The benefits typically associated are:

- Robustness of an enterprise solution, with a battle-tested process for generating explanations which can be relaunched in case of failure.
- Scalability: since models are treated indifferently, their analysis can usually be parallelized.
- Pre-built features such as automated report generation which speed up the entire process.

The drawbacks of those tools is that they cannot handle models with fewer than 10 predictive variables differently from those with hundreds of variables; they rarely provide a semantic abstraction layer on top of the models under investigation; and their visualisation capabilities are somewhat limited or too rigid.

Other teams developed their solutions *ab initio*, often relying on popular open source software components. Those teams were better able to customise the explanations produced:

- To the specific use case, e.g. by integrating domain knowledge into the explanations generated in the form of high-level concepts or an abstraction layer on top of the raw predictive variables.
- To the black box situation, e.g. by tailoring the implementation between highly responsive models which enabled training a global surrogate model, and those less responsive ones which were more amenable to multivariate analysis of predictive variables.

The drawbacks of custom solutions were obviously the time to design and implement them (which was incompatible with the very limited timeframe of the Tech Sprint), the relative brittleness of the resulting tools, and the practical impossibility to handle all 9 models with equal attention.

**Cutting-edge R&D**

The Tech Sprint was an opportunity to showcase French expertise and tradecraft in explainable AI, and more generally in data science.

Besides popular methods, participants indeed leveraged the state of the art in AI explainability, including inventions by the teams themselves – thus confirming the excellence of French scientific research in this field. Among those:

- The « Active Coalitions of Variables » method consists of computing Shapley values only for the most influential variables, while ensuring robustness of the resulting explanations. This leads to more easily interpretable and visualized explanations (at the cost of an additive constraint imposed on those explanations, which makes the method not necessarily suitable for all use cases).
- The Shapash library, created by the MAIF Group, delivers explanatory elements such as feature importance values but in a visually simple form, making them accessible to all types of audience.
- The Skope-rules method., whose designers include the BPCE Group, aims to learn simple, logical rules so as to optimise the detection of input data areas mapped to a given output prediction.

It should be noted that the innovation within these various works pertains not only to the production of explanations, but also to the quality of their delivery (particularly via a conciseness criterion, detailed in the “Intelligibility principle” section).

**Simplicity does not equal transparency**

The Tech Sprint illustrated that, when they are available as black boxes only, models based on random forests are sometimes as difficult to explain as the more complex ones (such as stacked models or GBMs).

Rather, the complexity of the input data – particularly the number of predictive variables – was the main factor in how difficult each black box proved to be. This observation applies both:

- To the generation of explanations. Indeed, it is much more challenging to select relevant features when potential predictive factors are in large number. In particular, the impact of a given variable may be non-monotonic between two local explanations, or between a local and a global one, which is generally characteristic of less robust explanations.
- To the appropriate reception of the explanations produced. Indeed, a prediction is often more difficult to motivate when many factors are at play, especially insofar as those factors are statistically and semantically correlated. The recipient of an explanation implicitly expects causal relationships, which is infeasible in the case of pure ML models without any causal constraint embedded in them.

# 4. Delivery of the explanations

The Tech Sprint thus required leveraging one or several explanatory methods suitable – both scientifically and in terms of technical feasibility – to the specific challenge of black box auditing. The quality of the delivery was nonetheless as important an evaluation criterion as the method used, since the final objective was ensuring proper and complete understanding of an explanation by its recipient(s).

This sections attempts summarising the means used by participants to deliver the explanations produced as adequately as possible. Two key principles seem to emerge:

- Striving to make explanations as intelligible as possible by limiting the cognitive load associated to their interpretation.
- Designing user interfaces enabling some form of interaction with the explanations.

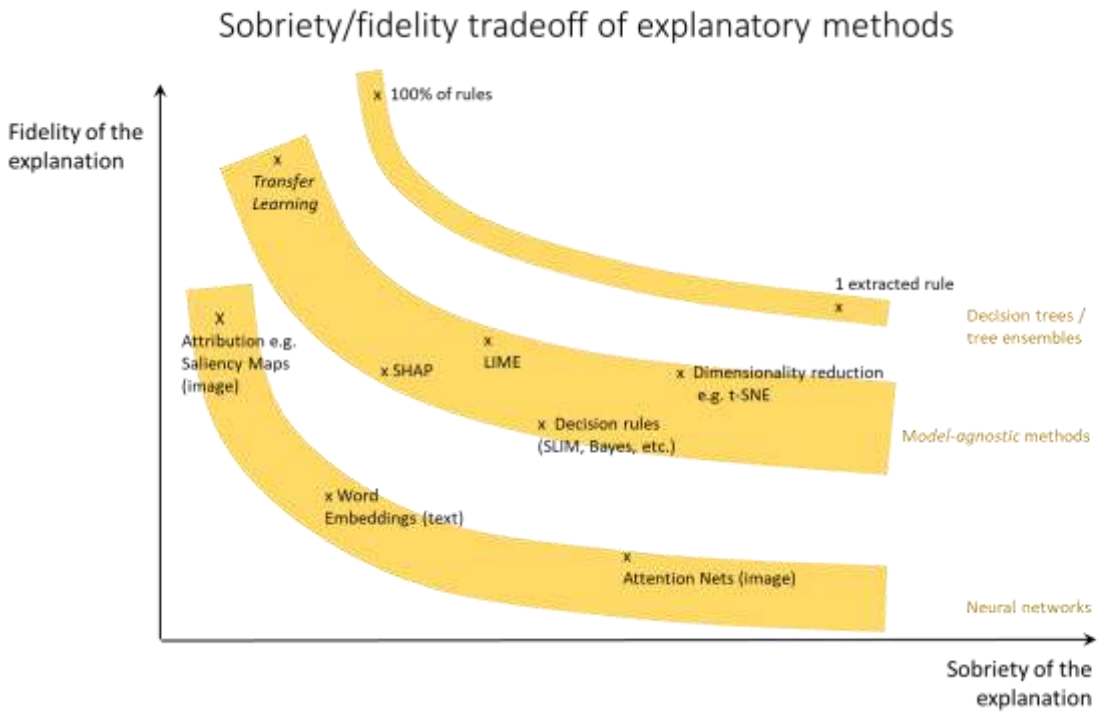
## 4.1 Intelligibility principle

An algorithmic explanation should be intelligible by its intended recipients, suitable for the use case considered, and proportionate to the risk associated to the business process.

**Sobriety / fidelity trade-off**

The ACPR's discussion paper on governance emphasised a trade-off between an explanation's sobriety and its fidelity to the underlying model. On the one hand, the explanation's fidelity (with respect to the algorithm which produced a given prediction) is imperfect since the algorithm's behaviour is necessarily simplified when its output is explained in terms of certain characteristics of the individual or transaction considered. On the other hand, the sobriety of the explanation, that is its intuitiveness and intelligibility by a layperson, is both subjective and constrained in practice.

The following diagram, excerpted from the 2020 governance document, attempts to represent the sobriety/fidelity trade-off for an explanation according to the type of ML algorithm and the type of explanatory method. A few "corridors" are drawn to show that for a given algorithm type, some explanatory methods will deviate slightly from the general trends.




*3. Trade-off between sobriety and fidelity of an explanatory method*

**Conciseness**

Research literature in cognitive sciences related to the production and reception of algorithmic explanations puts the cognitive limitations of the recipient at the forefront of the issues. Among commonly accepted principles for taking those limitations into account are the constraint on the number of cognitive chunks[7] to display and their temporal sequencing, on-demand displaying of the explanations, and the progressive disappearance of their components over time.

Conciseness of the explanatory elements presented to the user at a given point in time is thus recognized as a key element of a good explanation. Conciseness is however, as explained in the previous section, in tension with the objective of fidelity to the underlying model, and is also notoriously difficult to measure insofar as it relies on the ambiguous notion of a cognitive chunk. A Tech Sprint participant put it as follows: “The conciseness metric should be adapted to each use case, depending on who the main stakeholders are.”

Tech Sprint participants strived to adapt their explanations’ conciseness to the use case presented to them. The method generally used to this aim was to group variables either by behaviour or by semantics:

- Semantic categorisation consists for instance of grouping financial variables (household income, monthly loan payments, etc.), sociodemographic variables (marital situation, professional situation, etc.), and account management variables (date and frequency of reimbursement incidents, etc.) The limitations of this method are its aggregation of variables

[7] A maximum of 3 to 5 cognitive chunks, in any case no more than 7, is typically recommended. For a general definition of a cognitive chunk, see the seminal paper by G. A. Miller: “The magical number seven, plus or minus two: Some limits on our capacity for processing information”. The precise definition is context-specific but usually corresponds to a conceptual cluster, i.e. a group of concepts that are mutually similar while being semantically distant from other concepts involved in the explanation.

with potentially diverging behaviour, and its sometimes arbitrary groupings (in this example, where to put the housing situation?).
- Categorisation of variables according to their behaviour is thus often preferable, in which variables with similar trends are grouped together – and often turn out to be semantically related anyway (for example mortgage information in the case of the Tech Sprint on retail credit applications).

Another particularly innovative method attempted to reconcile conciseness with robustness: the approach in question enables measuring the stability of a group of variables, and thus to derive the smallest "stable" subset of predictive variables in the sense of leaving the outcome unchanged with a very high degree of probability[8]. The constraint relaxation resulting from this probabilistic thresholding enables reducing the number of features presented in a given explanation while only sacrificing those with a negligible impact or affecting a small minority of individuals.

**Bridging the gap between local and global explanations**

Some participants tried to bridge the gap between local and global explanations, and thereby to remedy their respective flaws. Indeed, local explanations can be faithful to the underlying model which they attempt to approximate or whose predictive features they attempt ranking by importance, but only in the neighbourhood of a single input data point: an explanation thus only explains a single prediction at a time. As for global explanations, they aim to approximate the overall behaviour of the model, again either by using a surrogate model or by computing feature importance, which leads for all models but the simplest ones to a somewhat unfaithful representation of reality.

Three distinct approaches enabled several teams to define explanatory scopes going beyond this local/global dichotomy: *personae*, simple clustering, and clustering of Shapley values.

Personae

This approach consists of building from scratch a series of archetypes, or avatars, representative of a population segment. Each *persona* is typically not a real, unique individual, and its characteristics often result from the combination of statistical criteria (to obtain as representative avatars as possible) and business criteria (to avoid grouping together customer profiles that are totally unrelated).

The benefit of this method is to suggest a limited number of realistic individual consumer profiles, then to explain the model's behaviour on each profile. The mental representation by explanation recipients becomes simpler as the user considers the treatment of a specific individual. On the other hand, mapping of all users to a persona can yield approximate results, due to the small number of *personae* typically defined. Thus, the actual model behaviour on individuals far away from the *persona* may differ from what the analysis suggests, hence a relative lack of fidelity of this method. Similarly, the set of predefined *personae* may not correctly account for data points in low-density areas.

Simple clustering

In this approach (the most classical of the three), individuals are grouped according to the similarity of their intrinsic characteristics, for example homeowners aged 30 to 45. The definition of the corresponding classes (or clusters) is an emergent property from training and test datasets which aims to summarise high-density areas in those data, rather than based on criteria defined subject matter experts (as in the case of *personae*).

[8] The method is called *Active Coalitions of Variables* and uses the SDP (Same Decision Probability) technique.

The advantage of this method compared to *personae* is that clusters reflect the statistical distribution of input data: a given cluster will usually not contain two points too remote from each other; furthermore, a representative individual (real and not an artefact as with *personae*) can usually be inferred from each cluster. Conversely, clusters can be in larger number than the expert-defined *personae*, and those representative data points rarely match the most relevant profiles from a business standpoint.

Clustering of Shapley values

In this method, individuals are grouped according to the similarity of their most important features. This does not imply a similar distribution of their intrinsic characteristics as in the case of simple clustering: for example, individuals younger than 25 or older than 60, with no current credit or more than 3 current credits, might be grouped together by this method due to the importance of factors "age" and "number of current credits". A tool in this category, developed by one of the Tech Sprint teams, produces so-called "regional explanations".

This method distinguishes itself from the others in that it groups individuals according to their treatment by the model, and not according to intrinsic features that may well have no significant impact in the predictive model. Conversely, the mental representation and the description of these clusters is less intuitive than for the other methods since they are defined in terms of importance of each variable and not in terms of their value.

### 4.2 User interaction principle

The previous section "Intelligibility principle" started from the necessity of taking into account cognitive limits of the recipient of an algorithmic explanation. Another established result from socio-cognitive research is that an algorithmic explanation works best when it aims to reproduce or accompany the human process for formulating an explanation.

The best practices entailed are adapting an explanation to its recipient's objectives, selecting predictive features according to criteria close to those used by humans (notably abnormality and intentionality), iteratively adjusting displayed elements based on the user's reaction, and favouring actionable variables. Some of these points have been particularly examined by Tech Sprint participants.

**Actionability**

The analysts' works also raised the question of how actionable the variables presented in an explanation should be. Actionable variables are – in the present context – the factors that may be affected by a change in the financial services consumer's behaviour: for example, the loan amount may be somewhat actionable in some cases, but not for a consumer with a well-defined project; the household revenue is not actionable.

Some teams deemed important to include actionable variables both within explanations intended for the consumers themselves and within those displayed to client advisors. Conversely, other teams pointed out situations requiring non-actionable variables: for example if a customer's loan application has been rejected due (at least partly) to their age, this should be made transparent since loan approval decisions may not be systematically and exclusively based on age.

**Contextualisation**

Finally, Tech Sprint works have emphasised that an explanation is never a standalone piece of information that can be interpreted without any additional context. For example, diagrams should be

accompanied by domain knowledge. This might include a description of each predictive variable's semantics – what it means in the real world, not just in the model. Information about the model itself may also be relevant in some cases, e.g. to make explicit the difference between a median interest rate and the interest rate offered to the consumer.

**Explanatory dialogue**

Participants have explored by various means the fundamentally interactive nature of the process of receiving an explanation.

Some teams' work included features enabling navigation within an explanation: on-demand visualisation of the most important features (clicking on a variable shows a detailed display of its distribution and its effects), zoom on a subpopulation or even a specific individual to show local explanations, interactive filtering of the number and type of variables (e.g. to only retain the actionable or non-actionables ones), etc.

One team included a feedback loop into its explanatory mechanism so as to continuously improve the quality of the explanations produced over time: higher-level concepts integrated into explanations are then updated based on user feedback.

## 5. Future work topics

For its first Tech Sprint, the ACPR thus decided to focus on one of the foundational principles identified for proper governance of AI algorithms, namely model explainability.

Among potential future ACPR works on AI explainability, the following can be mentioned:

- As a follow-up work on AI explainability, the ACPR may focus on other AI implementation and governance principles that have only been tackled at a theoretical level, such as algorithmic fairness.
- The Fintech-Innovation Hub is also pursuing work on interactions between AI-based systems and human operators, both under an academic angle (a literature review of the human factors at play in the reception of an algorithmic explanation) and under a practical one (an experimental study on robo-advisors used for life insurance).

Besides, the Fintech-Innovation Hub is working on the audit of AI algorithms in general. The specific question of an appropriate evaluation process arises naturally: based on best practices or high-level regulatory requirements, the analysis of a concrete use case may shed light on the most suitable method and tooling for this evaluation task. Interactions between humans and the algorithm are a key issue: as both the ACPR's publication on AI governance and the European Commission's proposal for AI regulation have emphasised, human should play an active role for any AI-based system, hence the necessity to include their actions and behaviour in the evaluation protocol.

Lastly, as a continuation of the Tech Sprint, one of the winning teams is currently expanding on its work from the event and building a demo application showcasing the methods used, as well as their results on example datasets from the Tech Sprint and beyond. This prolonged impact of the ACPR Tech Sprint on the works of various actors within the financial sector is a very positive signal, confirming the value of such collaborative events held under the ACPR's umbrella.