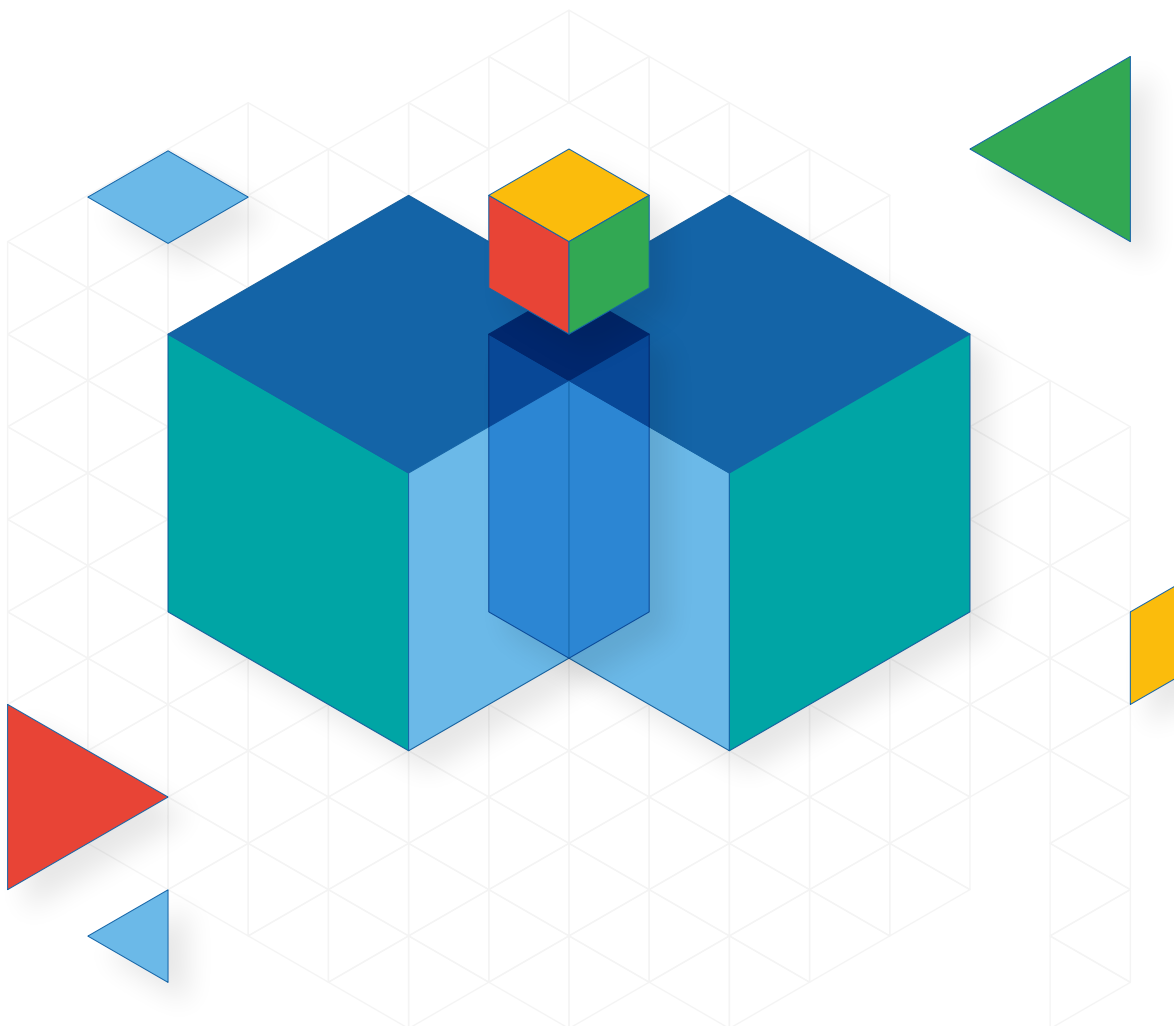


Delivering  
**Advanced Artificial Intelligence**  
in the banking industry





Delivering  
**Advanced Artificial Intelligence**  
in the banking industry



## Data & Analytics

### Delivering Advanced Artificial Intelligence in the banking industry

*whitepaper*

#### **BBVA Team:**

Roberto Maestre, *PhD.*

José Antonio Rodríguez, *PhD.*

Jordi Nin, *PhD.*

Axel Brando, *Industrial PhD Candidate.*

Irene Unceta, *Industrial PhD Candidate.*

Alberto Hernández, *AI Manager, Innovation Labs.*

#### **Google Team:**

José Carmona. *Cloud Consultant.*

Lorenzo Caggioni. *Strategic Cloud Engineering.*

Stefan Hosein. *Strategic Cloud Engineering.*

---

#### **Copy editing**

Jairo Mejía

#### **Design**

Israel Viadest

Joan Llop

#### **BBVA Data & Analytics**

Av. de Burgos, 16D, 28036 Madrid

Contact:

[roberto.maestre@bbvadata.com](mailto:roberto.maestre@bbvadata.com)

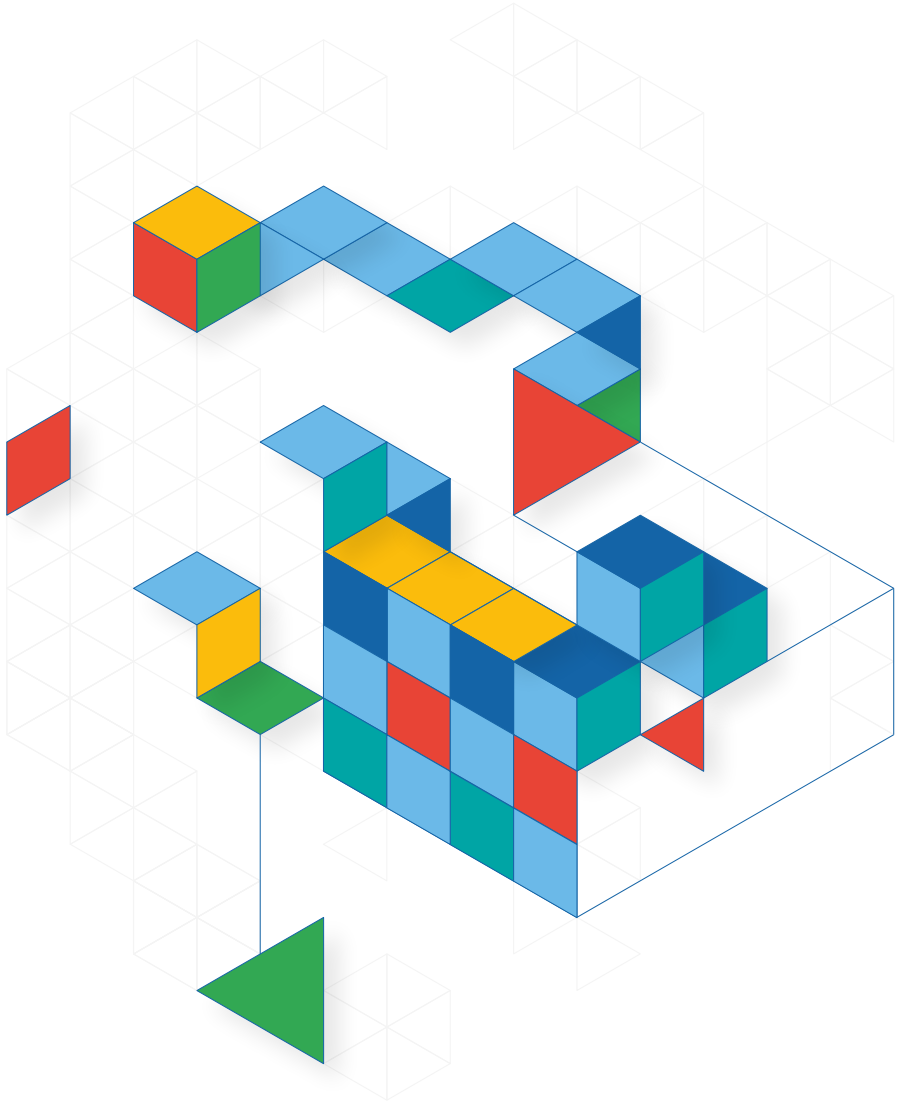
October 2018 | Madrid, Spain



Scan this QR code to download the digital version of this whitepaper

# INDEX

<b>EXECUTIVE SUMMARY</b>	5
<b>1. ARTIFICIAL INTELLIGENCE IN THE HANDS OF CUSTOMERS</b>	11
<b>1.1 Current product</b>	11
<b>1.2 Product boosted with AI capabilities</b>	12
<b>2. SCIENTIFIC AND TECHNICAL SOLUTION</b>	15
<b>2.1 Recurrent Neural Networks</b>	15
<i>2.1.1 Uncertainty</i>	16
<i>2.1.2 Deep Networks for modeling uncertainty</i>	17
<b>2.2 Privacy</b>	19
<i>2.2.1 Solution and recommendations</i>	19
<b>2.3 Technology</b>	21
<i>2.3.1 A hybrid approach</i>	21
<b>3. RESULTS AND CONCLUSIONS</b>	27
<b>3.1 Evaluation metrics and baselines comparison</b>	27
<b>3.2 Conclusions</b>	28
<b>4. REFERENCES</b>	31




# EXECUTIVE SUMMARY

Delivering real AI applications in digital financial retail is complex from both a scientific and technical perspectives. This is due to the difficulty in providing credible and advanced mathematics to solve a real business problem while, at the same time, being able to integrate these capabilities into a production-ready system. BBVA and its center of excellence in analytics, BBVA Data & Analytics, are delivering a **new engine to forecast expenses and incomings of customers**, based on deep learning techniques that take into account the uncertainty of the forecast and addresses privacy concerns. With the support of the Google Cloud Platform (GCP) team, we have deployed to production the mathematical solution integrally developed at BBVA D&A [1] as Minimum Viable Model (MVM). This provides us with a roadmap to update the current “future expenses tracker” functionality boosted by real AI.

The current budget forecasting engine is part of the new suite of experiences in the BBVA mobile app. A customer can visualize an estimate of outflows and inflows for the coming month in several day-to-day categories, such as groceries, gas, or cash withdrawals.

As an example, the forecasting engine must not only provide the following month’s expenses related to a given category (e.g. cash withdrawals), but also perform this task in a way the customer understands the “confidence” of the prediction, e.g.: 200 +/- 50 (meaning that the range of variation in the real forecasting can vary from 150 to 250). Being transparent about uncertainty has enormous implications in the customer experience because it provides a **more trusting interaction**.



...being transparent about uncertainty has enormous implications because it reinforces trust

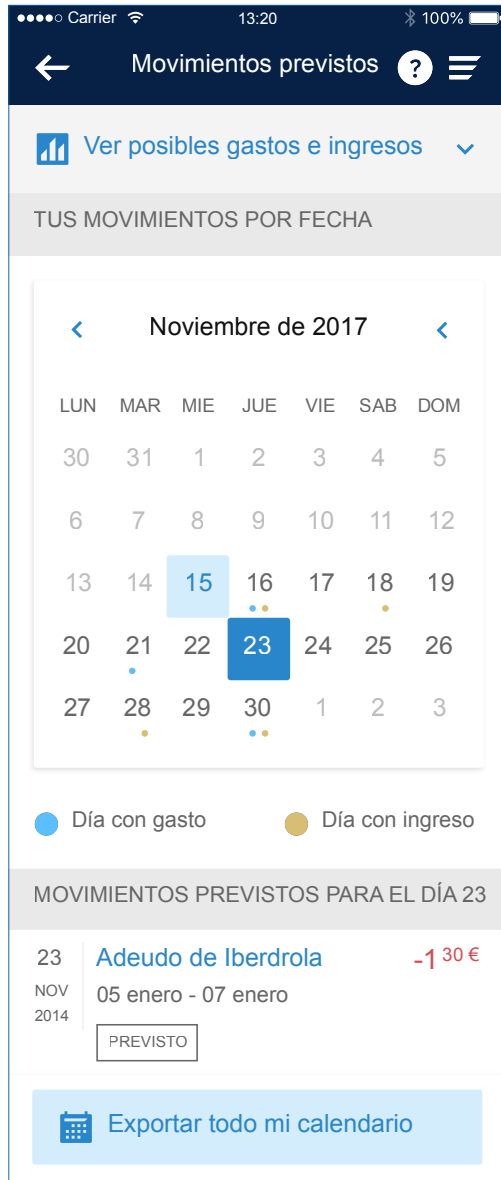
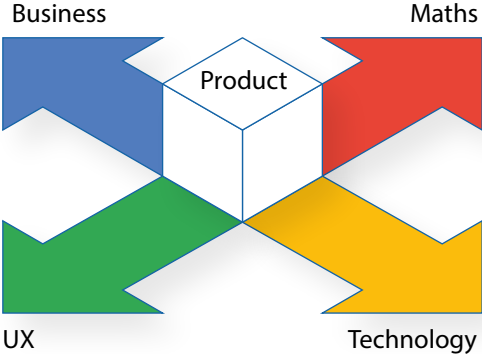


Figure 1: Screenshot of the calendar for expected transactions



In addition to the uncertainty above, another issue of capital importance when delivering real AI to customers is privacy preservation. Even when the dataset model has been trained – and possibly sanitized - the interaction with the model itself may lead to private data leakages [2]. In the worst case scenario, such a leakage may reveal private information about individuals in the dataset. To overcome this issue, in the present project we envisage specific countermeasures to mitigate the problem.

We believe that a well balanced AI-driven product is subject to four essential forces: **Business, UX, Mathematics and Technology**. Each force pulls in its own cardinal direction, which defines how the product is shaped. In order to avoid deformation and an unbalanced product, we have to find the correct trade-offs among the different facets outlined here.



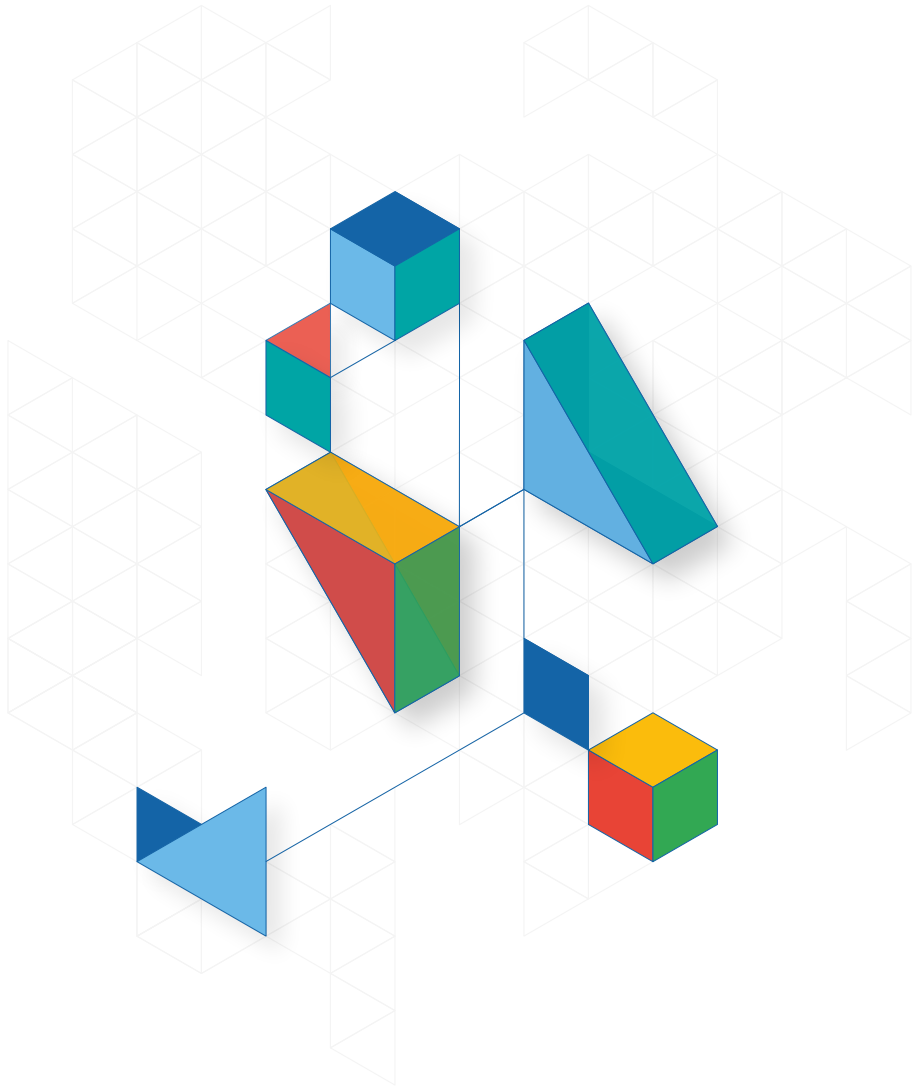
*Figure 2: The four pillars in which viable products are supported*

During the development of this project we have answered questions for each pillar, as follows:

- **How can our customers get more valuable information from their own generated data?** By using AI models able to learn complex patterns from all raw datasets, we have provided a highly reliable prediction for each customer.
- **What is the correct way to interact with a forecasting engine from a UX perspective?** By providing uncertainty in each prediction, more transparent and clearer knowledge is provided. We can also gain trust from a customer perspective as each customer can accept or reject predictions based on the level of uncertainty.
- **How can we solve a large-scale time-dependent optimization problem?** By training large Deep Neural Networks with uncertainty capability at GCP, we have been able to deploy state-of-the-art AI in the hands of our customers.
- **What is the method to ensure privacy on a trained model?** By designing privacy statements and providing clear metrics to ensure privacy when an AI model is deployed, we are certain that we can deploy models safely.

Moreover, the team has created an excellent way to translate and communicate ideas across disciplines, which leads to one single team working together. As a result, the team has been able to deliver an end-to-end AI product that helps our customers manage future expenses, gauging the uncertainty of the prediction. The product ensures privacy when deployed in the cloud.



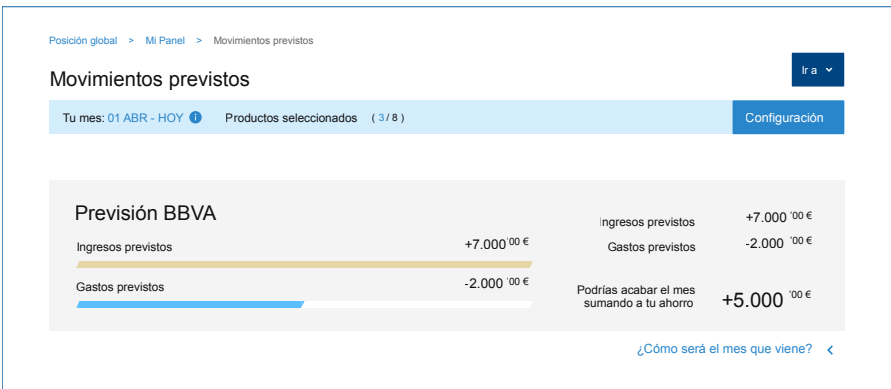


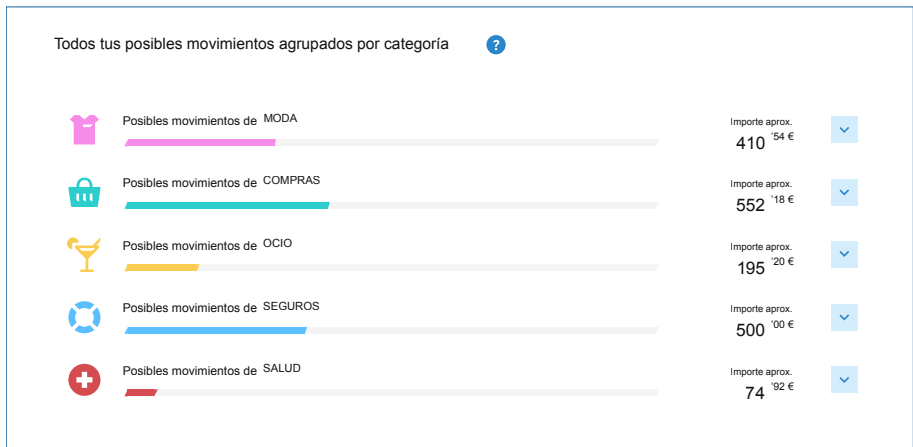
# 1. ARTIFICIAL INTELLIGENCE IN THE HANDS OF CUSTOMERS

## 1.1 Current Product

The Forecasting Expenses Engine is one of the most advanced features included in the BBVA mobile app. It provides a calendar forecasting upcoming expenses of a given customer. Customers are able to view two types of forecasts. Firstly, the total amount of expenses and incomings expected in a set of financial categories (eg. groceries, education, or leisure) for the current month. And secondly, operations which present a strong temporal recurrent pattern are displayed in a calendar together with a projected amount. This feature gives customers a picture of when an electricity bill or a mortgage payment is due, allowing them to prepare for events that may have an impact on their account balance.

In the following figures, the BBVA mobile app offers a snapshot of forecasted inflows and outflows in an account. The second figure goes even further, and predicts the amount expected in terms of transaction categories, such as groceries or restaurants.





Figures 3 and 4: Detailed examples of the current expenses tracker application

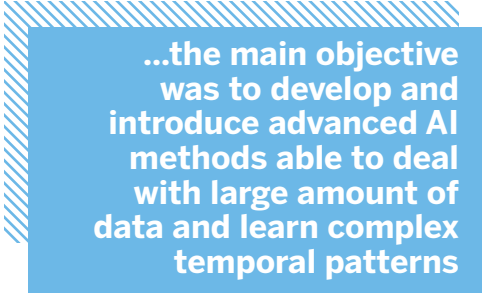
The two above views help customers not only to make future financial decisions, but also to compare the forecasted expenses with current patterns. Comparing real and forecasted transactions has proven to be a powerful method to detect anomalies in recurrent bills, or cash withdrawals. This way, customers are able to make decisions based on their own data. The new modeling implementation presented here expands the sophistication and potential of this popular forecasting engine tool.

### 1.2 Product boosted with AI capabilities

Previous studies have shown a path in which classic approaches can be improved specifically in error precision. By using the large amount of information generated by customers, the budget forecasting error metrics can be enhanced [3], providing a more reliable product. However, AI approaches such as Deep Learning allow us to study the patterns of all users simultaneously and therefore exploit much more information during training (which is usually translated into an improvement in accuracy for all users).

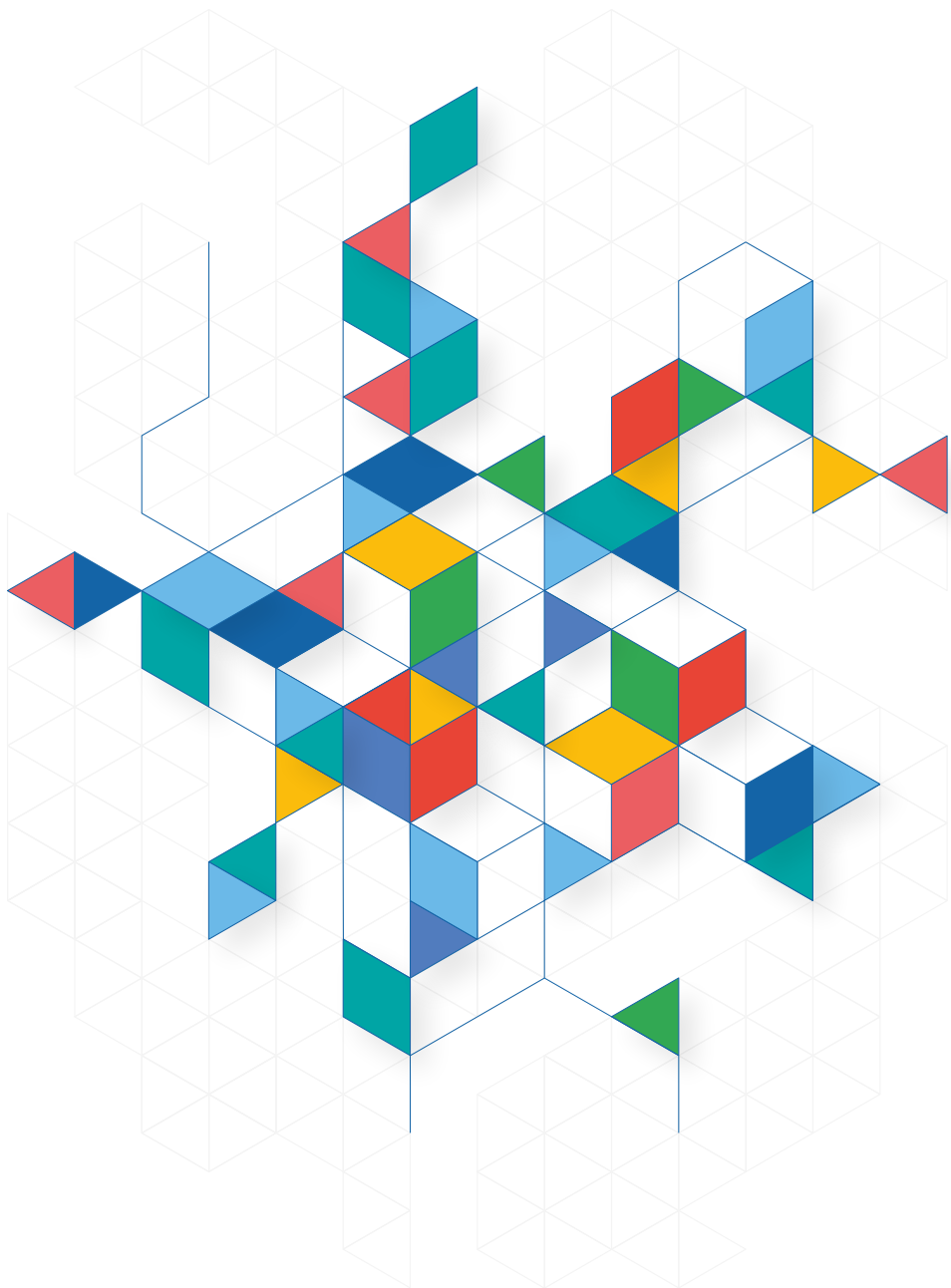
By learning multiple patterns, applied AI is able to improve forecasting for all customers as demonstrated in the evaluation section.

The model selected is called "Recurrent Neural Networks". This model requires advanced technology to be trained. Therefore, together with the Google Professional Services team, several approaches were explored in order to find the correct trade-off between computation performance in both training and inference, and data privacy.



**...the main objective was to develop and introduce advanced AI methods able to deal with large amount of data and learn complex temporal patterns**

Achieving a reduced error metrics directly impacts the user experience because it provides accurate information well ahead of time. Such information can improve trust and reliability in terms of financial actions in order to correct a given financial course.





# 2. SCIENTIFIC AND TECHNICAL SOLUTION

## 2.1 Recurrent Neural Networks

Recurrent Neural Networks provide an excellent approach to modeling long-short temporal relations from time-dependent datasets by using internal states (memory) to capture patterns from a sequence of inputs. The following figure represents how a RNN works from a high level perspective:

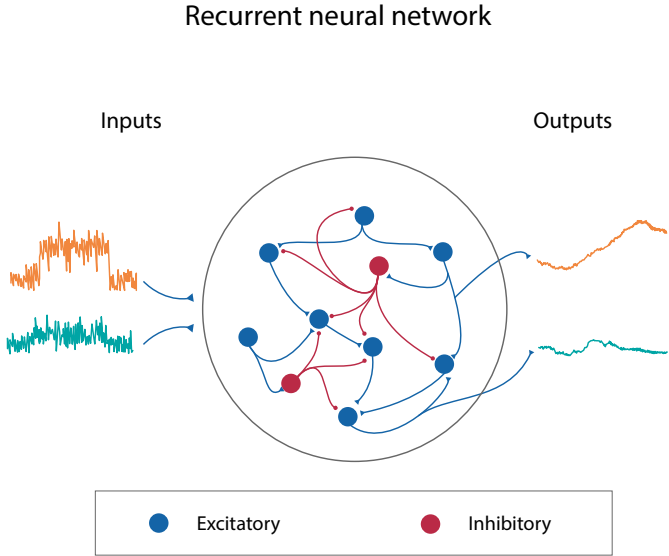


Figure 5: A conceptual scheme for a RNN

Inputs are time series related to “income and expenses”, aggregated and anonymized. This massive amount of data will be passed through a huge set of neurons. These neurons, by means of “Excitatory and Inhibitory” mechanisms (non-linear transformations), will react by producing an output: the prediction of the following

value. We believe that this specific training process will benefit from the capabilities of Google Cloud due to the amount of computation required for the task at hand.

### 2.1.1 Uncertainty

From a standard deviation perspective, uncertainty can be seen as a measurement or estimation of the amount of variation given an input pattern. It is extremely important to take uncertainty into account; while some human expenses series exhibit a degree of recurrence and are predictable, other series contain a number of variations, unexpected expenses or reflect erratic behavior. Modeling the uncertainty can help the prediction system to output a range instead of a single value and, if the range is too large, discard the forecast, since it most likely does not correspond to a recurrent or predictable expense.

The following figure represents:

- in **blue**, lines the value of expenses (historical data)
- in **red**, current expenses to be predicted

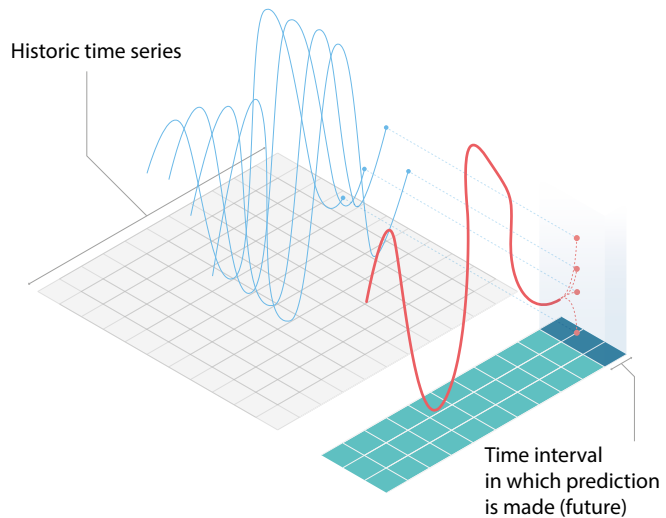
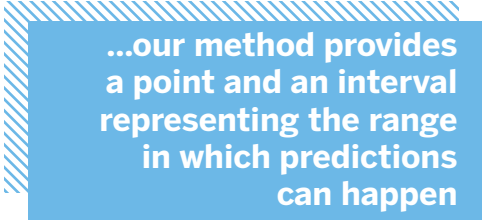


Figure 6: A conceptual scheme of time series forecasting

The main question is: What is the future value of the red line? (the red dot on the right).

If we observe the historical data, we can see that the shape of the blue lines are similar to the red line. Thus, we can use this past information for forecasting the red line. The principal Machine Learning methods will provide a single point in the middle of the blue dots (an average as a predictor with a minimum error).



**...our method provides a point and an interval representing the range in which predictions can happen**

Customers can also manage every expense and income category sorted by uncertainty. For instance, gas bills could be stable over time (perhaps with slight stationary variations). Cash withdrawal, on the other hand, exhibit greater variations.

### ***2.1.2 Deep Networks for modeling uncertainty***

Deep Learning offers a wide range of solutions for forecasting, in addition to other machine learning problems. This is because it is a really efficient way to execute a mathematical function when provided with a large data set. Therefore, Deep Learning is used as a “black box” first approach to handle a variety of cases that require a prediction.

At first glance, this is a powerful tool. However, we also need to consider the challenges it creates. For instance, what happens when it is more important to avoid making mistakes than for something to be predicted all the time? In most of the real world problems it is preferable to measure the risk surrounding the consequence of an action than to indiscriminately accept it. This indicates that the benefit of a correct decision versus the loss of a bad decision is not the same.

The straightforward application of Deep Learning models does not provide the confidence in its

predictions. To solve this limitation, we must look into already developed techniques for statistics that attempt to capture the concept of uncertainty [4-6].

The following proposal represents a groundbreaking approach in the banking industry that combines Deep Learning techniques (as a good Mathematical function approximator) with statistical modelling to demonstrate that not only is it important to take uncertainty into account in the problem explained, but that it is also possible to find a reliable solution with state-of-the-art techniques [1,7].

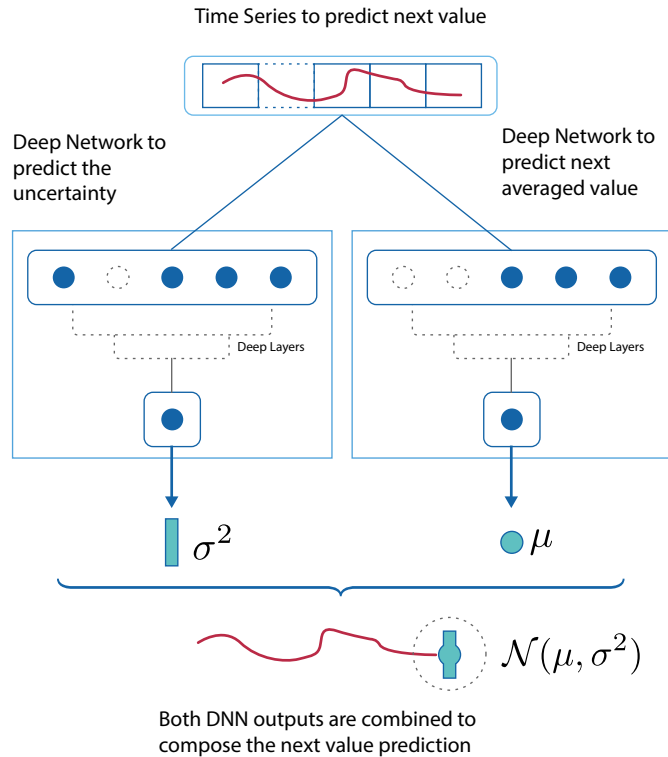


Figure 7: An schematic figure for a Deep RNN modeling heteroscedastic uncertainty

## **2.2 Privacy**

Preservation of privacy is an essential aspect of safeguarding when public data dissemination is required, even when data is only exposed as an output of a machine learning model.

When publicly released, AI solutions can be potentially compromised by malicious users that exploit query accesses to a model's interface. In the worst scenario, this practice results in the leaking of the underlying dataset on which the model was trained, and consequently allowing the attacker to recover sensitive features. When such leaking occurs, it is possible to recover information not only about the general training data distribution but, more worryingly, also about individual data records.

An important consideration when tackling such attacks is that they may occur independently of the security measures put in place during the data storage, pre-processing and fitting. Indeed, while specific privacy preserving protocols may be implemented at the different stages of data management, this type of privacy limitation is related to the interaction with the ML models themselves. As recently shown, many machine learning techniques learn the data to the extent to which they often "remember" specific attributes of the training distribution. To overcome this issue, not only do we need to ensure that the data be private, but also that the models themselves be immune to private data extraction.

### **2.2.1 Solution and recommendations**

In recent years, different solutions have been put forward by the scientific community for building privacy-preserving approaches such as: 1) ensuring the k-anonymity of datasets and/or 2) differential privacy [8].

A simpler, and perhaps more straightforward approach to tackle this problem, is to limit either the number or the types of queries allowed, or both. When interacting with a model interface, the amount of information output by the model in response to a given query is critical to evaluate the disclosure risk. In many adversarial learning examples it has been shown that models with information-rich predictions are susceptible to being attacked by malicious users. Thus, by establishing allowed queries to be of a membership type, or by limiting the total amount of queries, can be positive approaches for mitigating the potential damage of data leakage. On the downside, these countermeasures notably impact user experience.

**... one approach that we have been focusing on is the use of synthetic datasets to conceal training data information**

One approach that we have been focusing on at BBVA Data & Analytics is the use of synthetic datasets to conceal training data information. After a model has been trained, a surrogate model can be built using artificially generated data and later disclosed to the

public. In this setting, the original model is queried to obtain a representation of the learned decision function. Query points are generated following a pre-defined heuristic which ensures a good coverage of the whole attribute space while allowing no knowledge of the original data distribution.

The question remains however of how to efficiently build such synthetic datasets, a task that has revealed highly non-trivial for high dimensional environments.

These solutions aim to mitigate the risk of revealing confidential information about person connected to the data, and for which a certain model has been trained. Nevertheless, while convenient in specific situations, none of the above provide a holistic solution

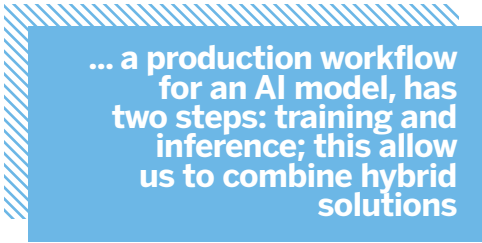
to the problem of data leakage. Admittedly, the optimal approach for each particular case may involve the usage of more than one of these countermeasures simultaneously.

## **2.3 Technology**

### **2.3.1 A hybrid approach**

Several solutions have been explored together with the Google Professional Services team to apply an strong technology approach [9]. Because a common AI model in production has two main stages, a hybrid productivization can be achieved providing considerable flexibility.

On the one hand, the learning procedure needs to read from the historical dataset and requires a huge amount of computation. For the second stage, the inference step only needs to see single data instances and can be computed faster with fewer computer resources.



**... a production workflow for an AI model, has two steps: training and inference; this allow us to combine hybrid solutions**

The following figure shows the main stages on a common AI model production: at the top (Training), an iterative procedure to reduce the training error provides a model which effectively captures relations presented in the data (represented by  $x$  and  $y$ ). Once a model is verified from a scientist's point of view -i.e.: cross validated error, overfitting control, common-sense concerning the correlation, fairness metrics and regularizations- the model is ready to be "pushed" to production. In production, the model infers new data providing a new kind of error: the inference error.

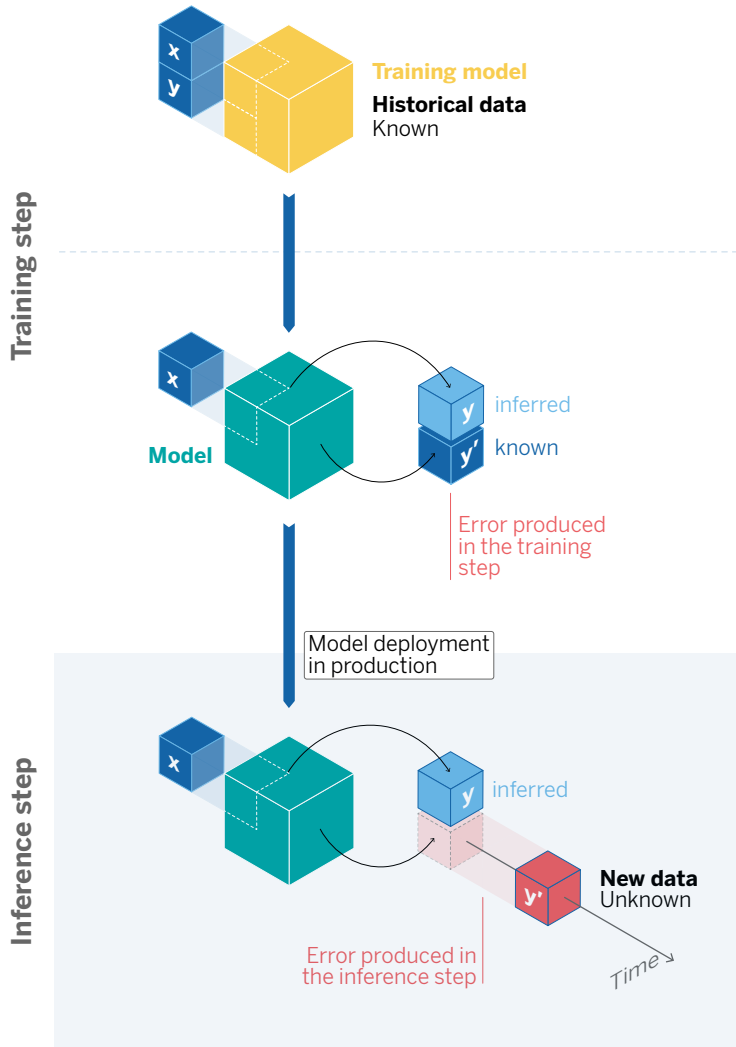


Figure 8: Main stages on the AI model production

The main technological scenarios in which an hybrid solution can be accomplished -taking into account the prior classification step- is summarized in the following figure:



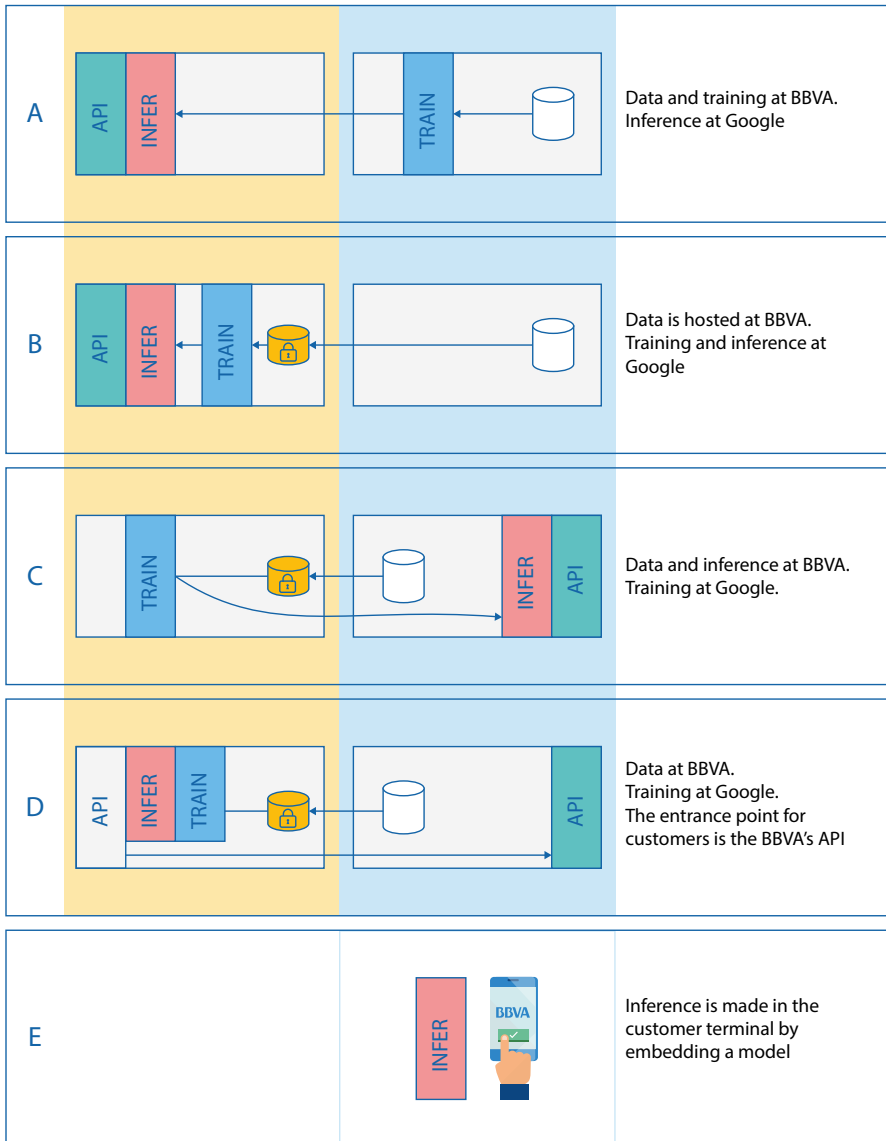


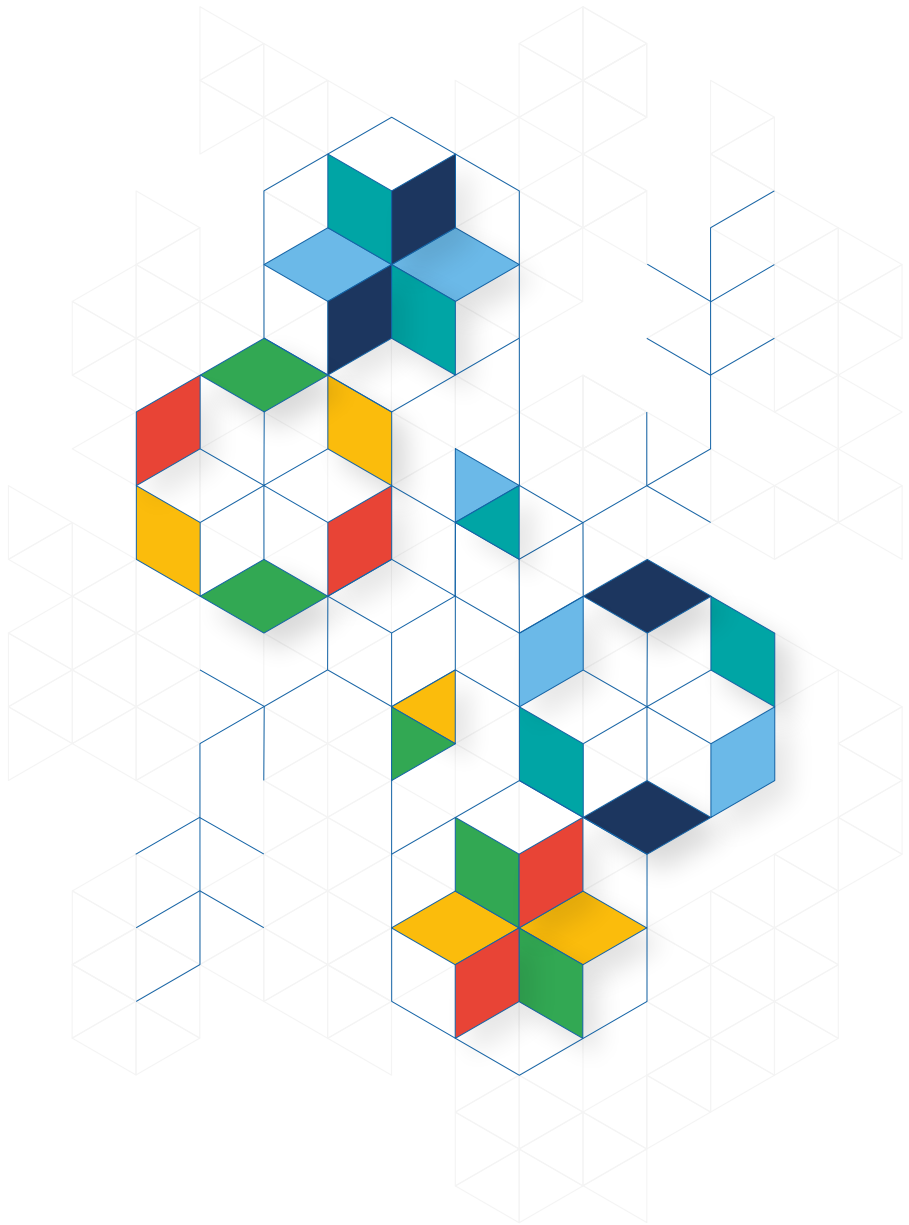
Figure 9: Main technological scenarios

The team has tested the five main possible scenarios, with the conclusions shown in the pro/cons table below:

Scenario	Pros	Cons
A	Raw Data is hosted at BBVA and the training model can learn from the raw data.  Scalability at the inference step on demand.	The GCP is not fully used in the training step, losing features such as the hyperparameter search.
B	Full GCP power both in training and inference.  Scalability at the inference step on demand.  Google tools for model managing i.e. versioning, rollbacks, are provided both in training and inference.	Because training is hosted at GCP, Raw Data is aggregated and anonymized losing power expressivity.
C	Full GCP power in training.	Because training is hosted at GCP, Raw Data is aggregated and anonymized losing power expressivity
D	Same as scenario B. However, all customer calls to the Inference API are made through the BBVA API to ensure in-house security mechanisms.	
E	Related to the training stage, scenarios A,B,C or D are suitable to apply here.	Issue considerations related to privacy must be taken into account because the model is integrated within the customer's application.

As a conclusion, Scenario D provides the best trade-off between data privacy and model performance in this use case, though having five available scenarios for other use cases and their peculiarities provides a huge richness by itself. From this study arises another important feature regarding the GCP ad-hoc hardware (TPUs). The training step is optimized directly on the vendor's hardware, thus providing a competitive advantage.





# 3. RESULTS AND CONCLUSIONS

## 3.1 Evaluation metrics and baselines comparison

To evaluate the error of the “expense tracker” AI model, we use the well-known mean absolute error. Therefore, the customer can set a threshold to select the level of uncertainty which incorporates the deviation between the real value and the forecasting. This threshold can inform customers of non-reliable forecasts.

Common Error Metric

Error with Uncertainty integrated

$$\text{MAE}(\mathbf{y}, \mathbf{x}) = \frac{1}{N} \sum_{i=1}^N |\mathbf{y}_i - \phi(\mathbf{x}_i)| \quad \blacktriangleright \quad \text{MAE}_{\kappa}(\kappa; \mathbf{y}, \mathbf{x}) = \frac{\sum_{i=1}^N |\mathbf{y}_i - \phi(\mathbf{x}_i)| \cdot [v_i < \kappa]}{\sum_{i=1}^N [v_i < \kappa]}$$

Three models have been compared: a trivial baseline (by using the average value or the last value available), a Random Forest, GAM modeling non-linear patterns and uncertainty, and a regular Deep Learning Neural Network. The following figure represents the error comparison between models, confirming that the best performing one is the proposed LSTM network (more details can be accessed directly from the research paper).

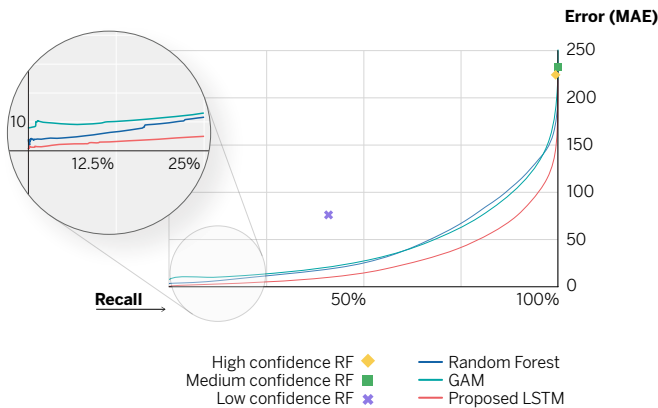


Figure 10: Models comparison

### **3.2 Conclusions**

We have applied an advanced AI algorithm, designed at BBVA Data & Analytics for a problem of critical business relevance (expense forecasting) and deployed in an end-to-end MVM pipeline in Google Cloud.

This exercise illustrates the roadmap to deliver real AI in a company in which, in our opinion, builds on four strong capability pillars: Business, UX, Mathematics, and Technology.

A clear channel of communication between teams responsible for each pillar must be effectively accomplished to ensure the correct integration of each solution component.

The algorithmic outcomes are the result of years of investment in specific disciplines of data science: time series forecasting and deep learning. This investment results in algorithms such as those already deployed in the BBVA app, as well as more advanced AI algorithms including the one presented in this paper. All of these can yield even more impact in the future or build assets that solve a diverse set of business problems.







## 4. REFERENCES

**[1] Brando, A., Rodríguez, J., Ciprian, M., Maestre, R., Vitrià, J.**, "Uncertainty Modelling in Deep Networks: Forecasting Short and Noisy Series". *Applied Data Science Track of ECML-PKDD*, 2018.

**[2] Tramèr, F., Zhang, F., Juels, A., Reiter, M., Ristenpart, T.**, "Stealing machine learning models via prediction APIs" in *25th USENIX Security Symposium*, 2016.

**[3] Alonso, A., et al**, "Forecasting financial short time series". *Electronic Journal of Applied Statistical Analysis*, vol. 11, no. 1, 2018.

**[4] Kendall, A., et al**, "What uncertainties do we need in bayesian deep learning for computer vision?" in: *NIPS*. pp. 5580–5590, 2017.

**[5] Bishop, C.M.**, "Mixture density networks", 1994.

**[6] Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.**, "Weight uncertainty in neural networks", *ICML*, 2015.

**[7] Ciprian, M., et al**, "Evaluating uncertainty scores for deep regression networks in financial short time series forecasting". *Workshop on Machine Learning for Spatiotemporal Forecasting*, in *NIPS*, 2016.

**[8] Dwork, C.**, "Differential privacy", a *ICALP*, LNCS 4052, Springer, p. 1-12., 2016.

**[9] Abadi, M., et al**, "TensorFlow: Large-scale machine learning on heterogeneous systems", 2015.

**BBVA**  
Data & Analytics



**BBVA**

Data & Analytics

*whitepaper*

