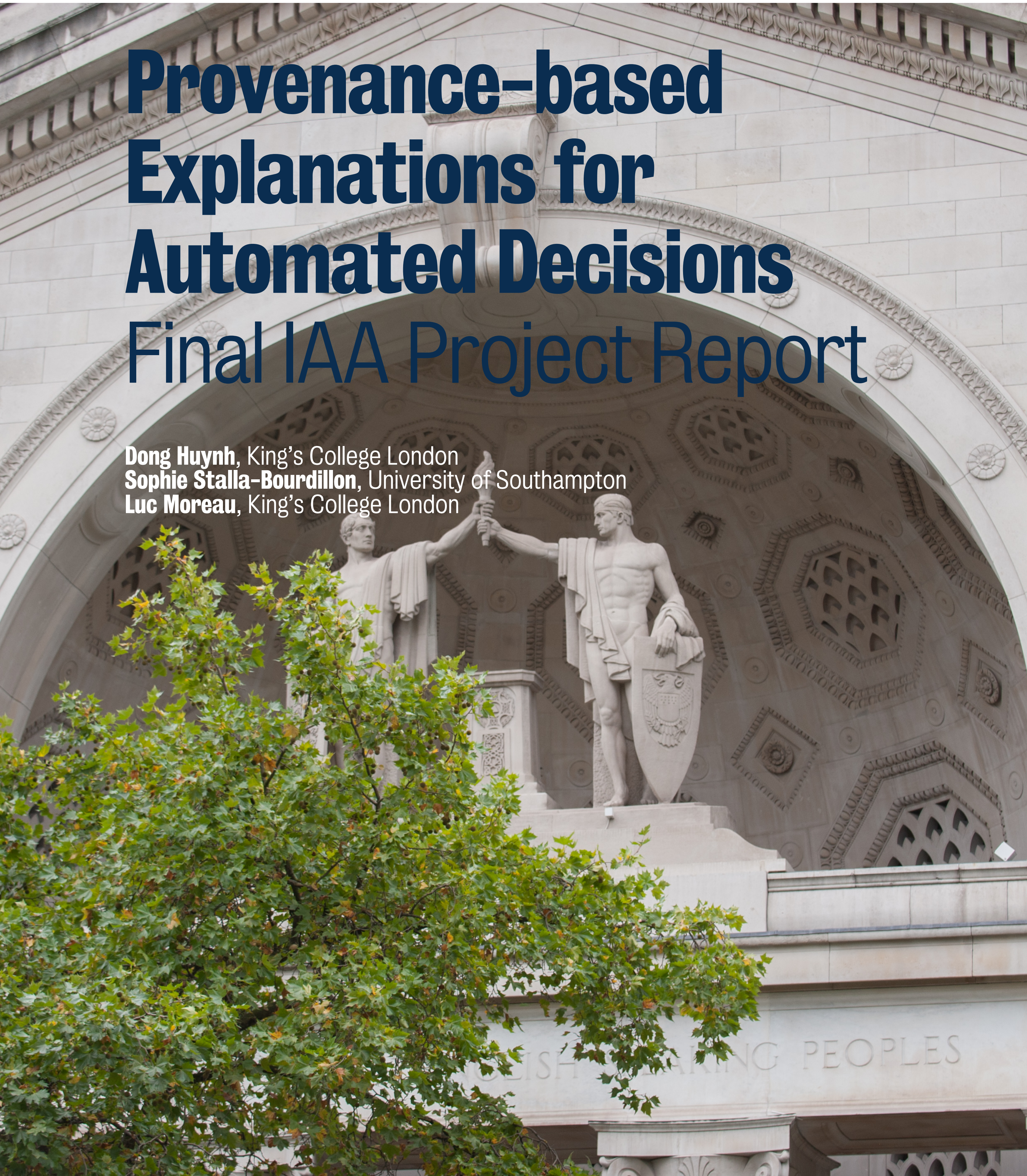
UNIVERSITY OF Southampton

KING'S College LONDON



Provenance-based Explanations for Automated Decisions

Final IAA Project Report

Dong Huynh, King's College London
Sophie Stalla-Bourdillon, University of Southampton
Luc Moreau, King's College London

Table of Content

AI-based automated decisions are increasingly used as part of new services being deployed to the general public. This approach to building services presents significant potential benefits, such as the reduced speed of execution, increased accuracy, lower cost, and ability to learn from a wide variety of situations. Of course, equally significant concerns have been raised and are now well documented such as concerns about privacy, fairness, bias and ethics.

Opportunities and Challenges

Several regulatory and legal frameworks have emerged across the world to address some of the concerns arising from automated services. Of interest to us in this project is the European Union's General Data Protection Regulation (GDPR) (European Union, 2016), a framework that codifies some rights for data subjects (the users who have provided data in return for those services) and obligations on data controllers (the organisations that are providing these services).

A key challenge is that regulatory frameworks remain high-level and do not specify practical ways of becoming compliant. For instance, how to determine if a decision is solely based on automated processing (Article 22 of the GDPR), how should the 'logic' of the processing be derived and expressed (Article 22 of the GDPR), or what is actually required in terms of transparency/accountability obligations, or whether transparency necessary leads to fit-for-purpose explanations (Article 12 of the GDPR).

Technology: Part of the Solution

While technology underpinning automated decision-making is the source of concerns, technology also has a place to help address these concerns. We are not suggesting the solution should only be technological but, instead, that technology must be part of the solution, in particular, because compliance should also be performed speedily, with accuracy, and low cost. Otherwise, the benefits of technology in the first place will be greatly diminished.

As there is increased interest in tightened governance frameworks for automated decisions, including steps for generating explanations pertaining to decisions, our focus is on what we refer to as *explainable computing*, including not only explainable AI, but also explainable security, explainable workflows, and any form of computing activity requiring explanations.

Provenance-based Explanations

Thus, "*a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering*" an automated decision is an incredibly valuable source of data from which to generate explanations. This is precisely the definition of W3C PROV provenance (Moreau & Missier, 2013), a standardised form of knowledge graph providing an account of what a system performed. It includes references to people, data sets, and organisations involved in decisions; attribution of data; and data derivation. It is suggested that provenance can assist with Information Accountability (Weitzner et al., 2008) and is listed as a fundamental principle in the US ACM statement on Algorithmic Transparency and Accountability.

To promote the suitability of provenance as a technological solution to aid in the construction of explanations, we designed and built a demonstrator that produces explanations for decisions related to a fictitious loan scenario. It allows a user to impersonate an individual applying for a loan, obtaining a decision, and being able to request explanations pertaining to the whole process. The demonstrator generates explanations in real-time.

This work is undertaken by a multi-disciplinary team, involving Sophie Stalla-Bourdillon providing a legal perspective, Dong Huynh and Luc Moreau for the computational aspects, and the Information Commissioner's Office, the UK regulator for data protection.

1. Explanations as Detective Controls

Automated decision-making has been and continues to be the subject of attention since the adoption of the GDPR (European Union, 2016). Academics from different disciplines have discussed at length the domain and effect of a right to explanation, exploring in particular two routes to generate explanations. Edwards and Veale explain that at a high level, explanations can be grouped into two classes: *"model-centric explanations (MCEs) and subject-centric explanations (SCEs)"* (Edwards & Veale, 2017). They observe that explanations might not be particularly useful to vindicate data subject rights but could prove useful to increase the level of trust in Machine Learning (ML) models or as pedagogical tools. S. Wachter *et al.* highlight the benefits of counterfactual explanations, which make it possible to generate explanations without opening algorithmic black boxes (Wachter, Mittelstadt, & Russell, 2017).

In this project on constructing explanations from provenance data, we build upon Selbst and Powles' approach, who argue that *"a flexible approach, guided by… functional requirements…, can best effectuate [the right to explanation] while preserving the ability of technologists to innovate in ML and AI"* (Selbst & Powles, 2017).

Explanations as detective controls

Selbst and Powles make the point that information is meaningful where it serves a *"functional"* purpose, i.e. when it has *"instrumental value"* such as the facilitation of a data subject's right to contest a decision. **We broaden the claim and argue that explanations have the potential to serve all "*data protection goals*,"** to use the expression at the core of the approach of the German Supervisory Authorities and expanded upon in the Standard Data Protection Model.[1] We, therefore, start from the assumption that information and, in particular, provenance-related information, is useful or meaningful each time it helps data controllers to demonstrate compliance with Article 5's data protection principles[2] and other related-data protection requirements or it helps data subjects to exercise their rights (the list of rights should go beyond Article 22,[3] i.e. the right not to be subject to a decision based solely on automated processing).

Explanations are, therefore, conceived as **detective controls** – measures facilitating the detection of potential compliance issues or of opportunities to exercise a right – which can benefit both data controllers and data subjects and which should be at the core of any data protection-by design strategy. In fact, to be able to bake data protection principles within systems as early as possible and meet the requirement of data protection by design introduced by the GDPR in its Article 25,[4] data controllers should embed a mix of preventive, directive, detective, and corrective controls (i.e. organisational or technical measures aimed at implementing data protection principles as early as possible). Detective controls are important to meet the principle of accountability, which requires that the data controller be in a position to demonstrate compliance with the principles of purpose limitation, data minimisation, accuracy, fairness, transparency… Detective controls are also important to *"integrate the necessary safeguards into the processing [of data] in order to… protect the rights of data subjects"* (see Article 25).

[1] Available at https://www.datenschutzzentrum.de/sdm/.

[2] See http://www.privacy-regulation.eu/en/article-5-principles-relating-to-processing-of-personal-data-GDPR.htm.

[3] See http://www.privacy-regulation.eu/en/article-22-automated-individual-decision-making-including-profiling-GDPR.htm.

[4] See http://www.privacy-regulation.eu/en/article-25-data-protection-by-design-and-by-default-GDPR.htm.

Let us take an example. In the context of a loan decision, an explanation generating information relating to the freshness of the data could take the following form:

> The data sources were a credit reference (credit_history/128350251) provided by the credit agency (credit_agency) at <2019-01-10T14:10:16> and a fico score (fico_score/128350251) provided by the credit agency (fico) at <2019-03-10T11:00:00>."

Such an explanation would thus make it possible for the data subject to assess whether up-to-date data has been used and, if necessary, react and ask for correction.

From a data subject's standpoint (the same is also true from a data controller's standpoint) both the data query itself and the resulted explanation are therefore important to detect whether accurate data has been processed. They are both key components of an effective detective control or safeguard.

Conceiving explanations as detective controls makes it possible to go beyond the debate whether the GDPR introduces a right to explanation to the benefits of data subjects, and to give data controllers an incentive to consider introducing explanatory methods each time they are using automated models to support decision-making.

Broadly defined, explanations can thus be described as the details or reasons provided to make an automated decision clearer or easier to understand. For our purposes, a decision can be either partially or fully automated; and in many cases, it is likely that a human will be included in the loop.

Linking explanations to goals and audience in context

Linking explanations to functional purposes or data protection goals is critical because, contrary to what is often assumed, explanations can be multiform. In other words, a wide variety of explanations can be generated, and their usefulness or meaningfulness is intrinsically dependent upon the specific goal they pursue. Counterfactual explanations, which usually take the form: "If X was different, Y would have been different/the same," are thus only one form of subject-centric explanations, which are not useful or meaningful in all circumstances.

Our initial list of goals, which will be refined over time, comprises: *transparency*, *accuracy*, *data minimisation*, *fairness* (with three sub-goals: automation, profile-related fairness, discrimination-related fairness), *intervenability* (with four sub-goals: access, rectification, portability, contestability of purely automated decision-making), *accountability* (with four sub-goals: performance, responsibility, process, on-going monitoring).

What is more, it is also important to understand who the recipient of the information is and what his or her expectations are. In other words, it is important to precisely identify the needs of what we call the **audience** of the explanations in addition to its goal.

By precisely determining the content of the data controller's obligations or the effect of the data subject's rights and thereby deriving what we call the **rationale** for the explanation, it should be possible to assess whether the explanation generated has reached an acceptable level of usefulness or meaningfulness. In order to do so, analysing data controller obligations and data subject rights in **context** is particularly helpful. Hence our attempt to build a prototype for a particular scenario in which an application for a loan is submitted to the data controller.

2. A Scenario of Automated Decision-Making

Credit applications nowadays are typically assessed by automated systems and often approved or rejected within seconds, without human intervention. In this project, we create a hypothetical loan assessment scenario that allows us to simulate such an automated decision pipeline with the aim to explore potential questions one may ask about its decision output.

The Loan Assessment Scenario

Loan Company is a credit institution that offers short-term unsecured loans to borrowers. In order to minimise loss from charge-off, i.e. when a loan is unlikely to be repaid by the borrower, the institution developed a machine-learning pipeline that predicts the probability of a charge-off from a loan application. Based on this probability, an automated recommendation is made on whether the application should be approved or rejected.

The pipeline was trained and tested on the company's past loan performance data and was shown to perform reasonably well. It was approved for deployment to access all incoming loan applications and is enabled to make automatic decisions in clear-cut cases without the attention of a loan officer:

- If the probability of charge-off is higher than 50%, the loan application is automatically rejected
- If the probability is less than 20%, it is automatically approved.

A loan officer has to examine the remaining cases (i.e. where the probability is between 20% and 50%) and make the final decision. Note that such a human decision is simulated in our implemented loan pipeline to streamline the whole process.

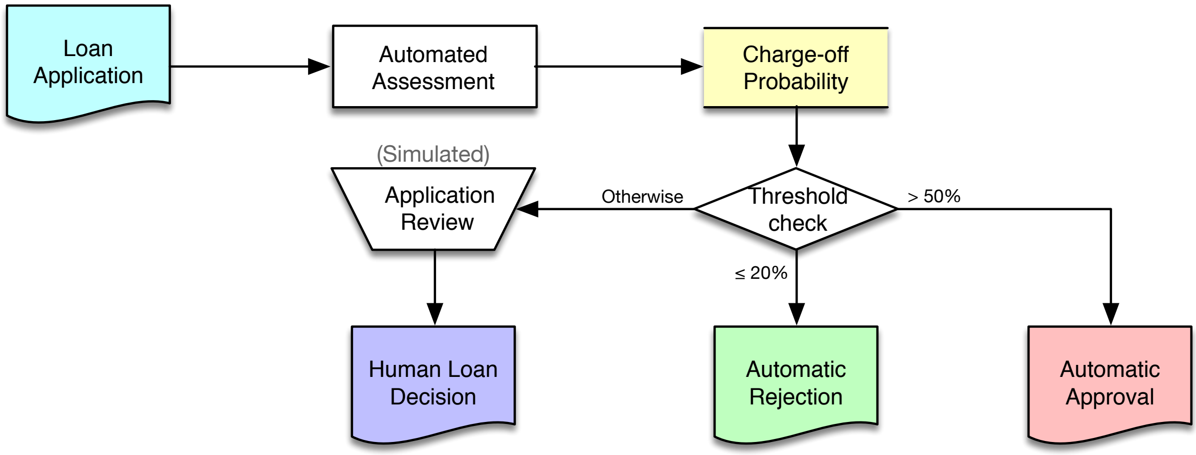


Figure 1. The flowchart of the simulated loan decision pipeline.

Building the Automated Decision Pipeline

To add some realism to the loan scenario, we use a real-world loan performance dataset[5] originally published by LendingClub, a US credit institution, to build the decision pipeline. The dataset went through typical machine learning analysis, filtering, and transformation steps:

1) Data filtering and selection:
    a) Only loans that have finished, either as *fully paid* or *charged off*, are retained.
    b) Loan features with significant missing data (i.e. over 30% of the dataset) or that are not available before a loan is approved are removed.
2) Data preparation and transformation:
    a) Remove loan features that are clearly not useful as predictors for charge-off:
        i) All values are unique or too many different values
        ii) Features that are already included in another (duplication)
    b) Convert feature values to those suitable for machine learning
        i) Loan status (fully paid/charged off) to 0/1 labels
        ii) Replace categorical features with dummy labels (0/1) for each of their categories
3) Split data into train and test sets according to the loan date: 90% and 10%
4) Create a machine learning pipeline with the Python Scikit-learn library to combine imputation and decision tree classification
5) Train the pipeline with the training dataset (90%)
6) Validate its accuracy with the test set (10%)

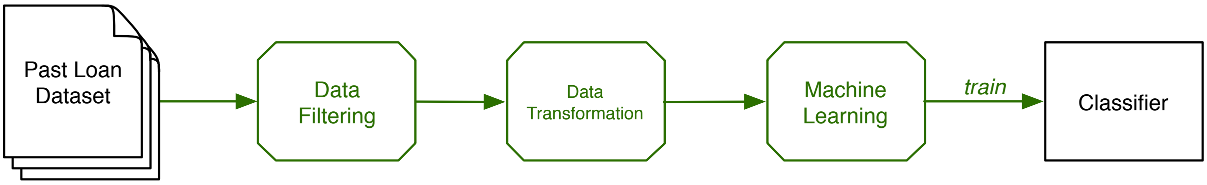


Figure 2. The multiple steps involved in creating the loan decision pipeline from data of past loans.

[5] Available at https://www.kaggle.com/wordsforthewise/lending-club.

3. Categories of Explanations

The above scenario of automated decisions provides us with the necessary details to think concretely about the types of questions one would ask with respect to a decision coming out of such a pipeline. In a workshop with the Information Commissioner's Office, we brainstormed on a variety of such questions. The questions can be loosely categorised into those that address the concerns of an individual data subject and those that address the concerns of the data controller (summarised in Table 1 below).

Table 1. Categories of explanations.

Individual Concerns	Institutional Concerns
Automation	Performance
Data Inclusion	Responsibility
Data Exclusion	Process
Data Source	Systemic Discrimination or Bias
Data Accuracy	Ongoing Monitoring
Data Currency	
Profile-related Fairness	
Discrimination-related Fairness	

In the following sections, we describe the above categories of explanations in more details. In particular, we identify the target *audience* of the explanation, the corresponding *questions* they may have, the *rationale* for an organisation to provide such an explanation and provide *example explanations* generated from the provenance recorded from the aforementioned loan decision pipeline.

Automation

Audience	Data subjects
Questions	Has the loan decision been reached solely via automated means?
Description	Whether a decision made solely by automated means without any *meaningful* human involvement.
Rationale	This explanation helps determine whether GDPR Article 22 is applicable and thereby the prohibition applies: "*The data subject shall have the right not to be subject to a decision based solely on automated processing…*" It is therefore relevant for demonstrating compliance with Article 5(1)(a) (fairness principle) and Article 5(2) (principle of accountability). This explanation should also help understand when best practice as unfolded in Recital 71 is met, e.g. to determine whether both child data and solely automated means have been used. This explanation could also help determine whether the information provided to the data subject as per Article 13, 14 and 15 is adequate.
Examples	`The automated recommendation was reviewed by a credit officer (staff/112) whose decision was based on your application (applications/34), the automated recommendation (recommendation/34)`

	itself, a credit reference (credit_history/34) and a fico score (fico_score/34). The loan application was automatically approved based on a combination of the borrower loan application and third-party data: the borrower credit reference and the borrower FICO score.

Data Inclusion

Audience	Data subjects
Questions	What types of data were used to assess my loan application?
Description	A loan application assessment may consider several types of data about the applicant, such as credit scores, previous interaction with the credit provider, employment data or other publicly available information.
Rationale	This explanation would help determine whether the data is ultimately relevant to the processing purposes as per Article 5(1)(c) (although this would only be the first step) and inform requests for access (Article 15), rectification (Article 16) and portability (Article 20). Ultimately its implementation would be useful for accountability purposes.
Examples	The data Loan Company considered for the borrower loan application is the number of mortgage accounts, the number of derogatory public records, the number of public record bankruptcies, the number of open credit lines in the borrower credit file, total credit revolving balance, the month the borrower earliest reported credit line was opened, revolving line utilization rate, the listed amount of the loan applied for by the borrower, the self-reported annual income provided by the borrower during registration, the purpose of the loan, the state, the address, the self-reported job title, the loan title provided by the borrower, employment length in years, type of application, the number of payments on the loan, the home ownership status provided by the borrower during registration, the borrower lower FICO score and the borrower higher FICO score.

Data Exclusion

Audience	Data subjects
Questions	Which data was excluded from the decision process?
Description	An automated decision pipeline is significantly influenced by the data used to build it. Therefore, it is important to understand why certain types or slices of data were included *and* excluded while creating the pipeline as those decisions have a direct impact on whether the pipeline is approved for service and how it behaves in deployment.
Rationale	This explanation would help determine whether the data is ultimately relevant to the processing purposes as per Article 5(1)(c) (although this would only be the first step). Ultimately its implementation would be useful for accountability purposes.
Examples	The data that the Loan Company excluded for the processing of the borrower loan application are the loan title provided by the borrower, the self-reported job title, the home ownership status provided by the borrower during registration, type of application, the address, the number of derogatory public records and the number of public record bankruptcies.

Data Source

Audience	Data subjects
Questions	Where did you get those data about me?
Description	Data considered by a credit institution may come from a variety of sources. Some examples are: the applicant, internal data from previous interactions with the applicant or their partner, credit agency. Some data may be purchased from a data broker.
Rationale	This explanation would help assess the lawfulness, fairness and transparency of the processing (Art. 5(1)(a)), the accuracy of the data (Article 5(1)(d)) and inform requests for access (Article 15) and rectification (16). Ultimately its implementation would be useful for accountability purposes.
Examples	`The data sources were the borrower FICO score (fico_score/35) provided by the credit referencing agency (fico) at <2019-06-21T06:25:13.426223>, the borrower loan application (applications/35) provided by the loan applicant (applicants/35) and the borrower credit reference (credit_history/35) provided by the credit referencing agency (credit_agency) at <2019-06-23T01:50:09.266708>.`

Data Accuracy

Audience	Data subjects
Questions	Are the data used for assessing my loan application correct?
Description	Data correctness may not be guaranteed: the applicants may have made a typo in their application; credit ratings may be of a different person of a similar name or adversely affected by wrong information.
Rationale	This explanation would help determine whether the data is ultimately relevant to the processing purposes over time as per Article 5(1)(e) and inform requests for access (Article 15), rectification (Article 16) and erasure (Article 17). Ultimately its implementation would be useful for accountability purposes.
Examples	You can check the data supplied in your original application for any inaccuracy. In addition, we obtained the following data from third-party providers: • Accounts: 12.0 The number of open credit lines in the borrower's credit file • Mortgages: 5.0 Number of mortgage accounts • Revolving Balance: 11771.0 Total credit revolving balance • Credit Utilization Rate: 25.4 Revolving line utilization rate, or the amount of credit the borrower is using • Earliest Credit Line: Oct-2002 The month the borrower's earliest reported credit line was opened • Public Records: 0.0 Number of derogatory public records • Bankruptcies: 0.0 Number of public record bankruptcies We also calculated the following data from the above and your supplied application data:

	• Interest Rate: 7.21 Interest rate on the loan • Instalment: 1238.93 The monthly payment owed by the borrower when the loan starts • Debt to Income: 17.94 The ratio of debt over income • Grade: A Loan grade • Subgrade: A3 Loan subgrade

Data Currency

Audience	Data subjects
Questions	How timely relevant is the data used for assessing my loan? Is the data used for assessing my loan up to date?
Description	Data used in loan decision making may be collected a long time ago and no longer relevant.
Rationale	This explanation would help determine whether the data is ultimately relevant to the processing purposes over time as per Article 5(1)(e) and inform requests for access (Article 15), rectification (Article 16) and erasure (Article 17). Ultimately its implementation would be useful for accountability purposes.
Examples	`The external data sources were the borrower FICO score (fico_score/35) provided by the credit referencing agency (fico) at <2019-06-21T06:25:13.426223> and the borrower credit reference (credit_history/35) provided by the credit referencing agency (credit_agency) at <2019-06-23T01:50:09.266708>.`

Profile-related Fairness

Audience	Data subjects
Questions	Have I been treated similarly to others having the same profile?
Description	People of a similar profile going through the same process should be treated similarly.
Rationale	This explanation would help determine whether the data is processed fairly (Article 5(1)(a)) (although this would not lead to a complete fairness assessment). Ultimately its implementation would be useful for accountability purposes. This explanation would help the data subject to determine whether their application was treated differently than the rest and to decide whether to seek further explanations.
Examples	`In the last month, applicants having the same income range as yours who applied for similar amounts were successful in getting the loans 73% of the time.`

Discrimination-related Fairness

Audience	Data subjects
Questions	I believe I was discriminated due to my gender/disability; can you prove otherwise? Is there bias introduced in the decision by my home ownership status?

Description	Identify any bias against a protected characteristic or the use of data fields correlated with protected characteristics (e.g. post code, consumer behaviours).
Rationale	This explanation would help determine whether the data is processed lawfully and fairly (Article 5(1)(a)) (although this would not lead to a complete fairness assessment). This explanation would also help determine whether special categories of data as defined in Article 9 have been processed. Ultimately its implementation would be useful for accountability purposes. The UK Equality Act prohibits differential treatments based on protected characteristics (i.e. age, disability, gender reassignment, marriage and civil partnership, race, religion or belief, sex, sexual orientation), unless exceptions apply.
Examples	We simulated alternative loan applications for all possible values of home_ownership, i.e., OTHER, OWN and RENT. In these simulations, the alternate applications would result in the following decisions: 'approval', 'approval' and 'approval' for values OTHER, OWN and RENT, respectively.

Performance

Audience	Data controller
Questions	Is the decision pipeline sufficiently accurate? How its performance was evaluated?
Description	Organisations must assure that the performance of the system is satisfactory.
Rationale	A performant decision pipeline will contribute to the efficiency of a business and, ultimately, its profitability. This explanation would help determine whether the data is processed fairly (Article 5(1)(a)) (although this would not lead to a complete fairness assessment). Ultimately its implementation would be useful for accountability purposes.
Examples	The company pipeline was assessed by a data engineer (staff/259) and has the level of accuracy of 79.58%.

Responsibility

Audience	Data controller, Regulator
Questions	Who were responsible for the final decision pipeline? Who set the threshold value for automated decisions? Who decided how the data was selected? Who approved the pipeline for deployment?
Description	As part of their own governance and to support accountability, organisations must keep track of who did what and when in their internal processes.
Rationale	This explanation would help determine whether the data is processed fairly and transparently (Article 5(1)(a)) (although this would not lead to a complete fairness assessment). Ultimately its implementation would be useful for accountability purposes.
Examples	Responsibilities for the AI pipeline were that data engineer (staff/259) selected file (loans_filtered.xz), that data engineer (staff/259) split file (loans_train.xz), that manager (staff/37) approved the company pipeline (pipeline/1) and that data engineer

	(staff/259) fit data for the company pipeline (1558649326/5011959424).

Process

Audience	Data controller, Regulator
Questions	What is the process for choosing the threshold value?
Description	Organisations need to understand how their business is run in practice.
Rationale	This explanation would help determine whether the data is processed fairly (Article 5(1)(a)) (although this would not lead to a complete fairness assessment). Ultimately its implementation would be useful for accountability purposes.
Examples	A committee consisting of a director (staff/34), a loan officer (staff/65), and a data engineer (staff/83) met on 31/10/2018 and approved the loan review policy (loan:policy/loan_review/2019).

Systemic Discrimination/Bias

Audience	Data controller
Questions	Has an equalities review carried out on the past loan applications?
Description	An automated decision pipeline may exhibit systematic and repeatable unfair treatment to a particular group of data subjects, which is often unintended and unanticipated.
Rationale	[Similar to discrimination-related fairness]
Examples	[No example currently available]

On-going Monitoring

Audience	Data controller, Regulator
Questions	When was the last time the decision pipeline revalidated? What is the current accuracy level of the loan decision pipeline? How often the accuracy is checked?
Description	Organisations must ensure that the automated decision pipeline maintain a reasonable performance and no new discrimination or bias is introduced to their systems over time.
Rationale	[Similar to the rationale for Performance and Systemic Discrimination/Bias explanations]
Examples	The performance of the decision pipeline was last assessed on 23/05/2019 by data engineer (staff/259) and approved by a manager (staff/37)

4. Tracking Provenance in the Loan Decision Pipeline

As we argue that the provenance of automated decisions is a valuable source of data from which explanations about those decisions can be generated. In order to explore how to realise that capability, we first need to record the provenance of such decisions in our hypothetical loan scenario.

Since the loan decision pipeline was implemented in Python, we use the PROV Python package[6] to record provenance assertions that are compliant to the PROV recommendations (Moreau & Groth, 2013) by the World Wide Web Consortium. In brief, the PROV Data Model defines three basic concepts, *Entity*, *Activity* and *Agent* (see Table 2 below) and various relations between them (as shown in Figure 3). For instance, an entity can be used by an activity to generate some new entity; the activity itself may be influenced in some ways by agents.

Table 2. PROV core concepts.

PROV concept	Description	Examples
Entity	A thing, either physical, digital or conceptual, whose provenance we want to describe	piece of information, decision, document, plan, dataset, trained machine learning model
Activity	Occurs over a period of time and acts upon or with entities	actions such as planning, monitoring, approving, training, classifying
Agent	Bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity	person, machine, service, system, organisation, collective

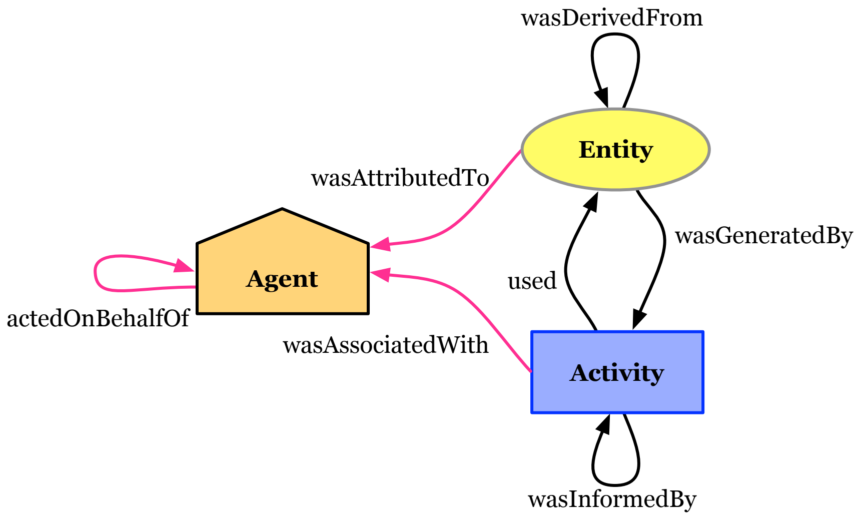


Figure 3. Main types of relations between PROV concepts.

[6] PROV Python package https://pypi.org/project/prov/

In the following sections, we present the recorded provenance of a loan decision step-by-step, from the loan application made by a borrower to its classification by the machine learning pipeline and the final decision, either automated or made by a loan officer.

Provenance of Input Data

The first piece of inputs to the pipeline is the loan application (`loan:applications/35`) made by an applicant (see entity `loan:applicants/35` in Figure below). The application entity contains the data provided by the applicant in the application form (shown in the white box linked to the entity by a dotted line). In addition, the Loan Company obtains the applicant's credit history (`loan:credit_history/35`) and credit score (`loan:fico_score/35`) from third-party organisations (`loan:credit_agency` and `loan:fico`, respectively). Each of the input entities is attributed to the responsible agent via an attribution relation.

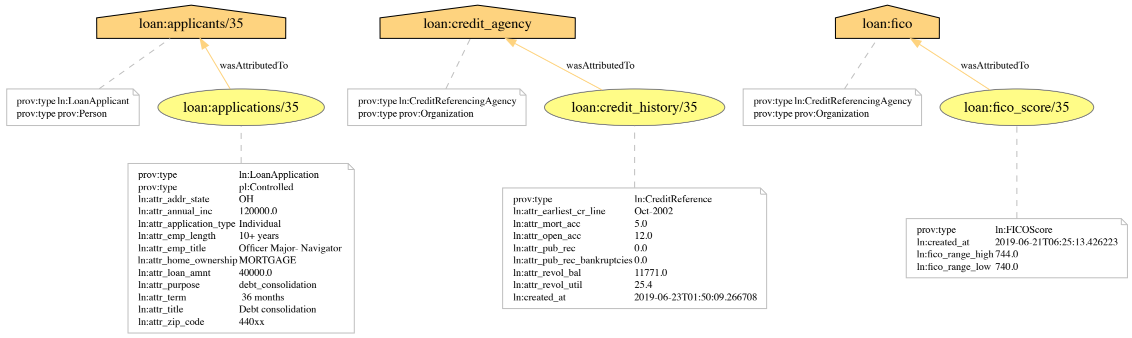

loan:applicants/35
wasAttributedTo
prov:type ln:LoanApplicant
prov:type prov:Person
loan:applications/35
prov:type ln:LoanApplication
prov:type pl:Controlled
ln:attr_addr_state OH
ln:attr_annual_inc 120000.0
ln:attr_application_type Individual
ln:attr_emp_length 10+ years
ln:attr_emp_title Officer Major- Navigator
ln:attr_home_ownership MORTGAGE
ln:attr_loan_amnt 40000.0
ln:attr_purpose debt_consolidation
ln:attr_term 36 months
ln:attr_title Debt consolidation
ln:attr_zip_code 440xx
loan:credit_agency
wasAttributedTo
prov:type ln:CreditReferencingAgency
prov:type prov:Organization
loan:credit_history/35
prov:type ln:CreditReference
ln:attr_earliest_cr_line Oct-2002
ln:attr_mort_acc 5.0
ln:attr_open_acc 12.0
ln:attr_pub_rec 0.0
ln:attr_pub_rec_bankruptcies 0.0
ln:attr_revol_bal 11771.0
ln:attr_revol_util 25.4
ln:created_at 2019-06-23T01:50:09.266708
loan:fico
wasAttributedTo
prov:type ln:CreditReferencingAgency
prov:type prov:Organization
loan:fico_score/35
prov:type ln:FICOScore
ln:created_at 2019-06-21T06:25:13.426223
ln:fico_range_high 744.0
ln:fico_range_low 740.0

Figure 4. Provenance of inputs data into the pipeline.

Classifying a loan application

The input data are then transformed into a set of loan features, modelled as an entity, that is suitable for processing by the machine learning pipeline (see Figure 5, `py:loan_features/35`). The provenance of the loan features entity is explicitly asserted by the derivation relations linking it to the input entities from which it was produced.

Using the pipeline (`loan:pipeline/1`), a computer (`ex:machine/8e7425f366a0`) assesses the loan features and produces an automated recommendation (`ex:recommendation/35`) for the loan application. It does so, however, on behalf of the Loan Company (`loan:institution`). The process of classification is modelled as an activity (`ex:classify_loans/35`), which has a start time and an end time; it uses the loan features as inputs and generates the recommendation as the output (see Figure 6).

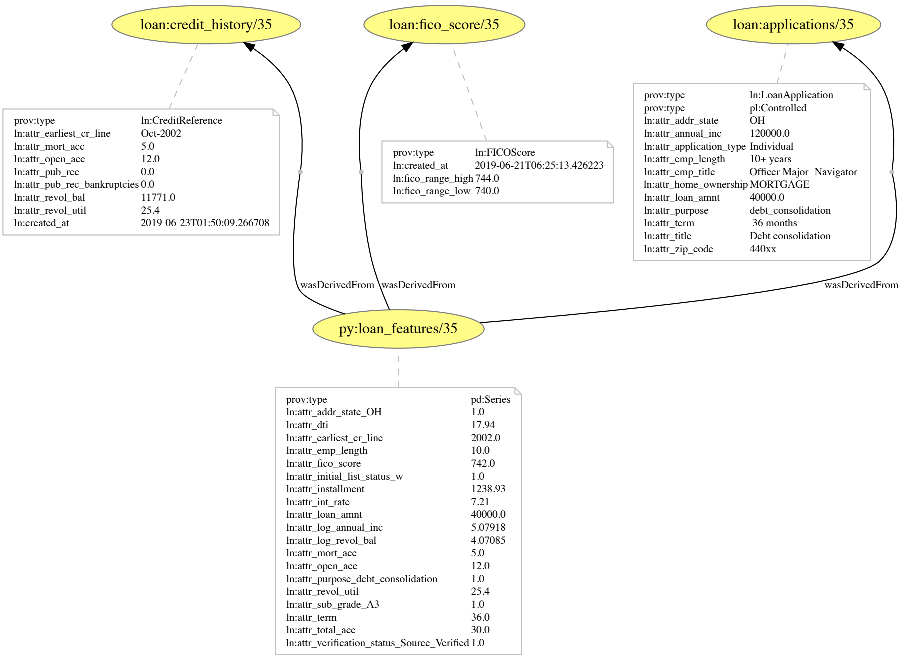

loan:credit_history/35
loan:fico_score/35
loan:applications/35
prov:type ln:CreditReference
ln:attr_earliest_cr_line Oct-2002
ln:attr_mort_acc 5.0
ln:attr_open_acc 12.0
ln:attr_pub_rec 0.0
ln:attr_pub_rec_bankruptcies 0.0
ln:attr_revol_bal 11771.0
ln:attr_revol_util 25.4
ln:created_at 2019-06-23T01:50:09.266708
prov:type ln:FICOScore
ln:created_at 2019-06-21T06:25:13.426223
ln:fico_range_high 744.0
ln:fico_range_low 740.0
prov:type ln:LoanApplication
prov:type pl:Controlled
ln:attr_addr_state OH
ln:attr_annual_inc 120000.0
ln:attr_application_type Individual
ln:attr_emp_length 10+ years
ln:attr_emp_title Officer Major- Navigator
ln:attr_home_ownership MORTGAGE
ln:attr_loan_amnt 40000.0
ln:attr_purpose debt_consolidation
ln:attr_term 36 months
ln:attr_title Debt consolidation
ln:attr_zip_code 440xx
wasDerivedFrom
wasDerivedFrom
wasDerivedFrom
py:loan_features/35
prov:type pd:Series
ln:attr_addr_state_OH 1.0
ln:attr_dti 17.94
ln:attr_earliest_cr_line 2002.0
ln:attr_emp_length 10.0
ln:attr_fico_score 742.0
ln:attr_initial_list_status_w 1.0
ln:attr_installment 1238.93
ln:attr_int_rate 7.21
ln:attr_loan_amnt 40000.0
ln:attr_log_annual_inc 5.07918
ln:attr_log_revol_bal 4.07085
ln:attr_mort_acc 5.0
ln:attr_open_acc 12.0
ln:attr_purpose_debt_consolidation 1.0
ln:attr_revol_util 25.4
ln:attr_sub_grade_A3 1.0
ln:attr_term 36.0
ln:attr_total_acc 30.0
ln:attr_verification_status_Source_Verified 1.0

Figure 5. Provenance of machine learning features for a loan application.

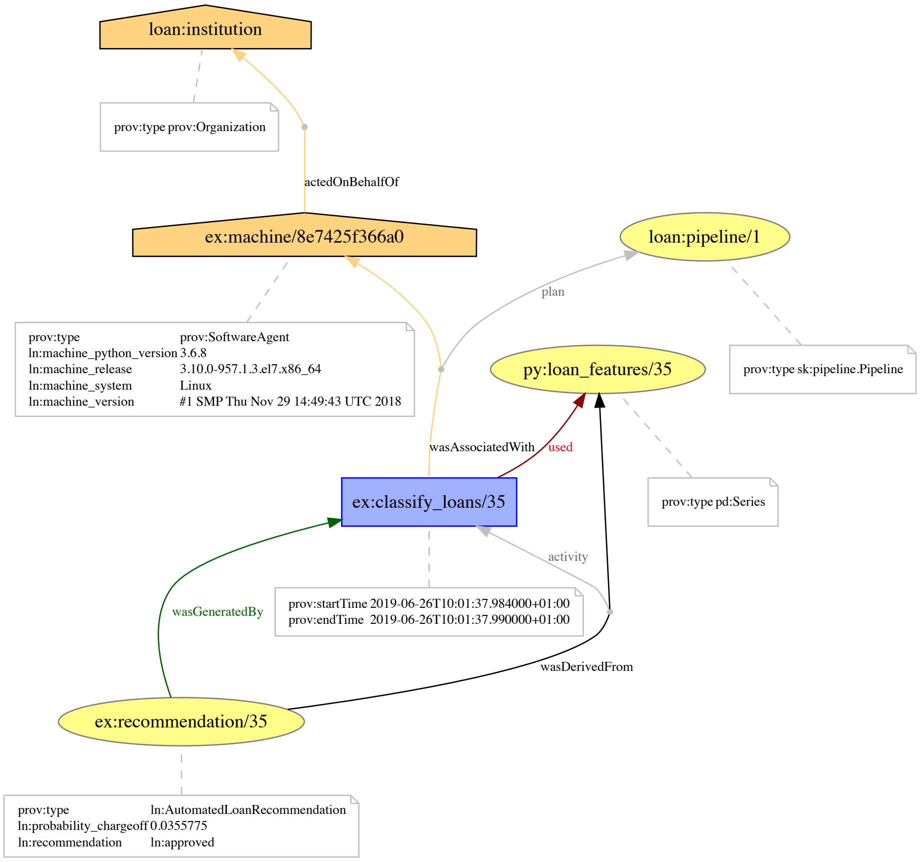

loan:institution
prov:type prov:Organization
actedOnBehalfOf
ex:machine/8e7425f366a0
loan:pipeline/1
prov:type prov:SoftwareAgent
ln:machine_python_version 3.6.8
ln:machine_release 3.10.0-957.1.3.el7.x86_64
ln:machine_system Linux
ln:machine_version #1 SMP Thu Nov 29 14:49:43 UTC 2018
plan
py:loan_features/35
prov:type sk:pipeline.Pipeline
wasAssociatedWith
used
ex:classify_loans/35
prov:type pd:Series
activity
prov:startTime 2019-06-26T10:01:37.984000+01:00
prov:endTime 2019-06-26T10:01:37.990000+01:00
wasGeneratedBy
wasDerivedFrom
ex:recommendation/35
prov:type ln:AutomatedLoanRecommendation
ln:probability_chargeoff 0.0355775
ln:recommendation ln:approved

Figure 6. Provenance of an automated recommendation of the pipeline.

Making a loan decision

Let us consider the case in which the pipeline predicts that the probability of charge-off is very low (3.5%) and, hence, an automated approval decision is generated directly from the recommendation from the pipeline (Figure 7). The decision is attributed to the computer running the pipeline; however, the chain of responsibility is made clear: the computer produces the decision on behalf of the Loan Company.

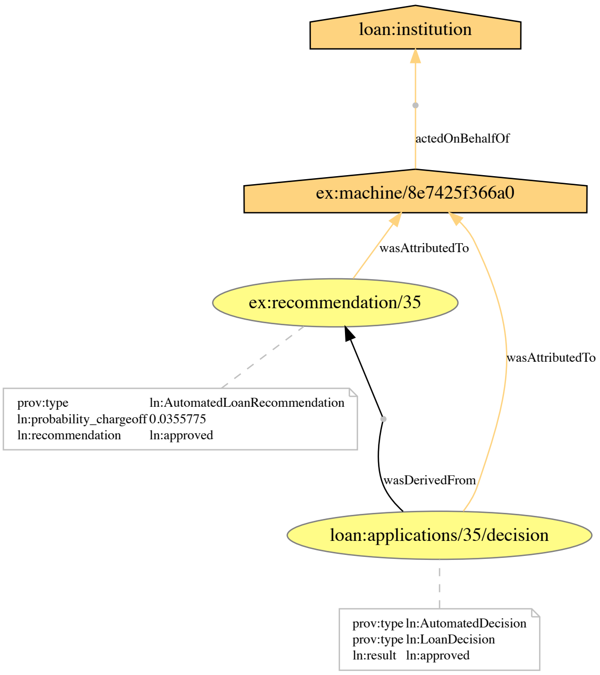

loan:institution
actedOnBehalfOf
ex:machine/8e7425f366a0
wasAttributedTo
ex:recommendation/35
wasAttributedTo
prov:type ln:AutomatedLoanRecommendation
ln:probability_chargeoff 0.0355775
ln:recommendation ln:approved
wasDerivedFrom
loan:applications/35/decision
prov:type ln:AutomatedDecision
prov:type ln:LoanDecision
ln:result ln:approved

Figure 7. Provenance of an automated loan decision.

In another case, the probability of charge-off is on the borderline (25.4%), the automated recommendation is escalated to be reviewed by a loan officer (`loan:staff/112`). The resulted loan decision is now attributed not to the computer but to the officer, who also acts on behalf of the Loan Company. Compared to the previous automated case, the provenance in this case (Figure 8) shows that the review activity takes into account the loan application, the credit history, and the credit score of the applicant in addition to the automated recommendation produced by the pipeline (Figure 8).

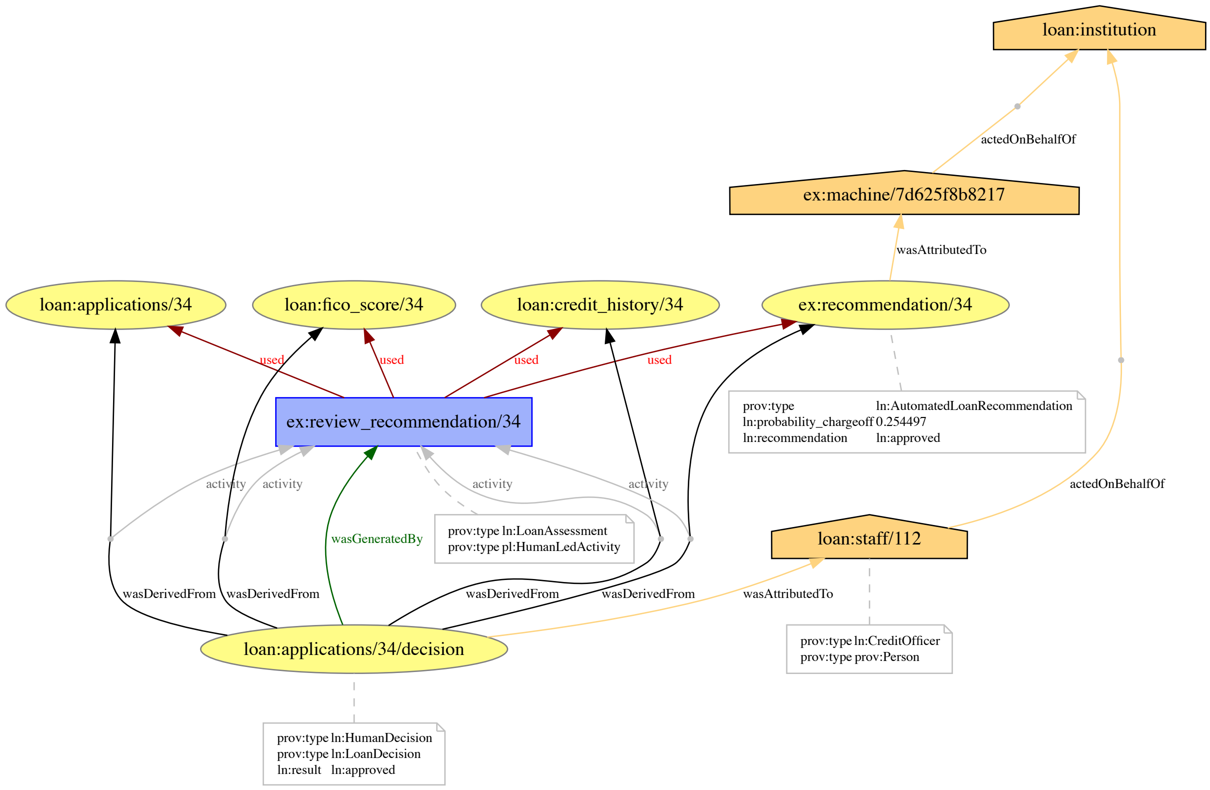


Figure 8. Provenance of a loan decision by a human.

Discussion

For the sake of brevity, we present the provenance of a decision here in small, digestible snippets. The full provenance of a decision is recorded as a single directed graph allowing one to trace via the provenance back to the input data and to identify the responsibility for each of the activities found along the way.

Each of the entities, activities, and agents in the provenance is annotated by types using one or more `prov:type` attributes. Most types are application-specific such as `ln:LoanApplication`, `ln:FICOScore`, and `ln:CreditOfficer`. In addition, we tag certain entities with types that will be useful for identifying relevant data in support explanation generation: `pl:Controlled`, `pl:HumanLedActivity`, `prov:SoftwareAgent`, `prov:Person` and so on. In the next section, we discuss the technical approach to generate explanations from the recorded provenance by querying the data for those annotated types.

5. Constructing Explanations from Provenance

In this section, we present the mechanism we use to construct explanations from the above provenance. Given the limited space, instead of reporting the technical details in full, we provide a summary of the overall approach (depicted in Figure 9) and illustrate it with a specific example.

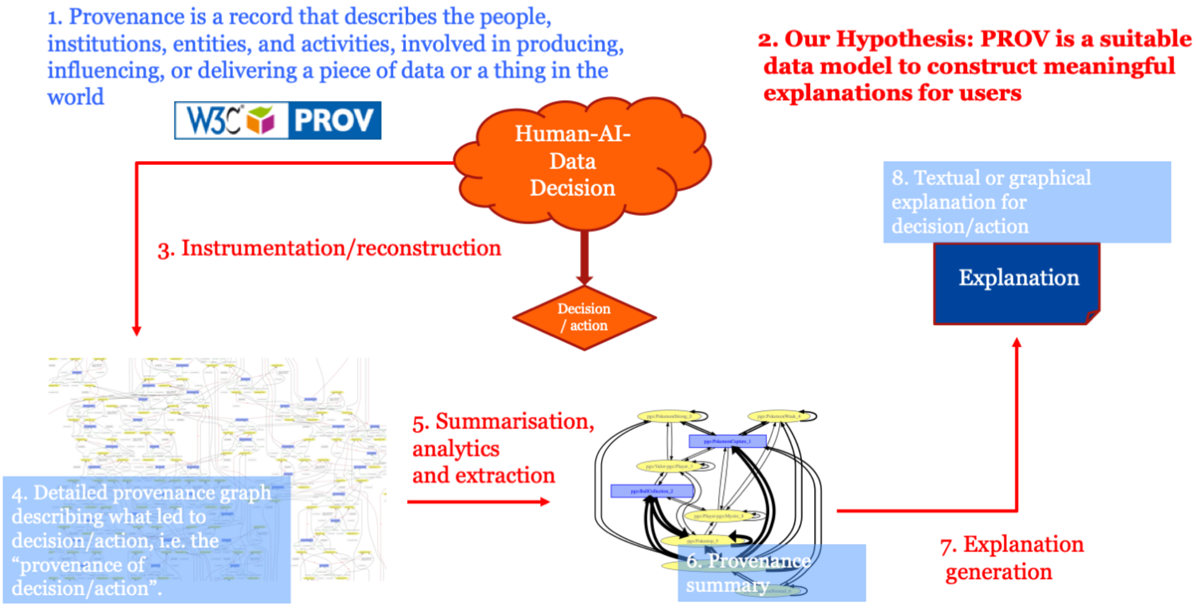

1. Provenance is a record that describes the people, institutions, entities, and activities, involved in producing, influencing, or delivering a piece of data or a thing in the world
W3C PROV
2. Our Hypothesis: PROV is a suitable data model to construct meaningful explanations for users
Human-AI-Data Decision
3. Instrumentation/reconstruction
Decision / action
8. Textual or graphical explanation for decision/action
Explanation
4. Detailed provenance graph describing what led to decision/action, i.e. the "provenance of decision/action".
5. Summarisation, analytics and extraction
6. Provenance summary
7. Explanation generation

Figure 9. Our approach to generate explanations from provenance.

As introduced earlier, provenance is defined as *"a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a pied of data or a thing in the world"*. Our focus here is the provenance of an AI-based decision (Note 1, Figure 9). It is our view that such a provenance record is an excellent starting point to construct an explanation specific to a decision, helping a data subject understand it and take action in response to it, appropriate for the context (see Note 2). Thus, we are assuming that the AI-based application has been instrumented in order to log provenance (see Note 3). To this end, we have been developing various techniques by which code can be instrumented (Moreau, Batlajery, Huynh, Michaelides, & Packer, 2018), some libraries have been instrumented, and even provenance can be reconstructed from logged data (as shown by Ramchurn et al., 2016).

In this project, we instrumented the loan decision pipeline described earlier to record the full provenance of a loan decision in the scenario. The result is a provenance graph (Note 4) providing full details of the inputs and processing leading to a decision, including data items, processes, agents that have influenced decisions. In long-running applications, and in applications processing a vast amount of data, potentially this provenance can become very large.

This full record in itself is not conducive to construct an explanation directly, since it may contain too many details that a data subject may find irrelevant, tedious, or overwhelming. Instead, the provenance needs to be processed (Note 5) to produce relevant information nuggets in support a specific explanation's purpose: this may involve summarisation (Moreau, 2015) and analytics (Huynh, Ebden, Fischer, Roberts, & Moreau, 2018), to extract the essence of provenance. The result is what we refer to as a provenance summary (Note 6).

The provenance summary is the input to a generation component (Note 7), converting relevant summary information into an explanation, which in our case would be textual. Alternatively, or additionally, the same extracted information could be represented in a graphical way to further aid its consumption. The outcome is an explanation which could be targeted to the data subject (and would typically refer to the data subject and their data, using “you” or “your application”) or to the data controller (and then would typically use “the borrower” or “the borrower’s application” instead).

An illustration

We illustrate the above approach with the explanation *Data Currency* from the loan scenario.

In this instance, a question a data subject may want to ask is “How timely relevant is the data used for assessing my loan?” so they can check if the latest data are being used (since they may believe that their circumstances have recently changed in their favour).

This question can be addressed by extracting all data items that have affected a decision. In the instance of a loan decision, they consist of 3 items: the loan application made by the borrower, and two credit references obtained from two distinct external credit referencing agencies. It is these that are of interest to the data subject, as they want to ensure that they are the most recent.

A suitable query over the provenance of the decision can be designed and executed to extract the following subgraph (Figure 10), out of which the relevant information can be found to construct the explanation. In the graph shown below, at the bottom, we find the decision (yellow ellipsis), and three influencing entities from which it was derived: the loan application, a fico score and a credit history. The latter two are provided by external agencies (fico) and (credit_agency), represented as orange pentagons.

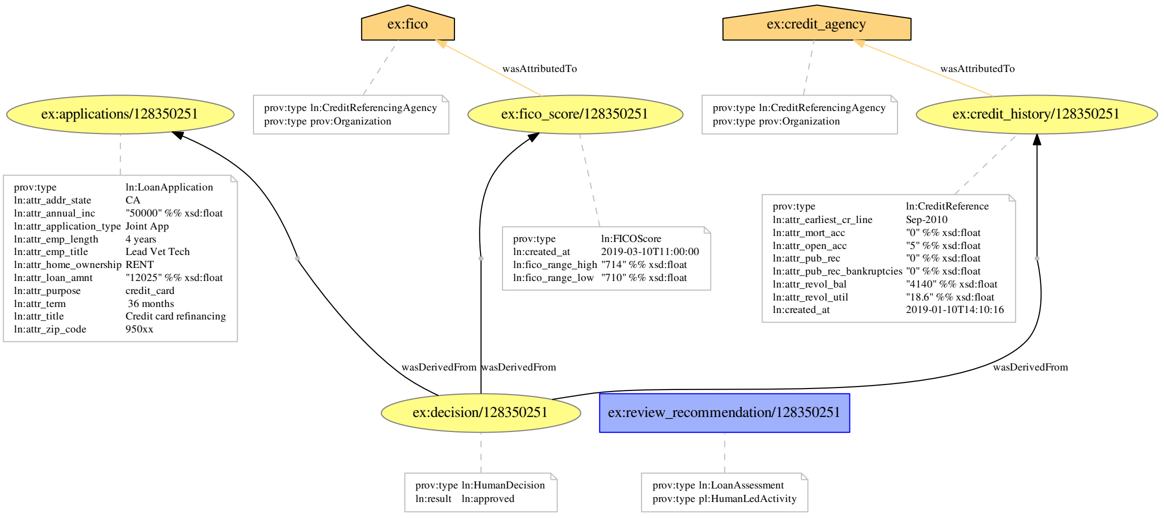

ex:fico
wasAttributedTo
ex:credit_agency
wasAttributedTo
ex:applications/128350251
prov:type ln:CreditReferencingAgency
prov:type prov:Organization
ex:fico_score/128350251
prov:type ln:CreditReferencingAgency
prov:type prov:Organization
ex:credit_history/128350251
prov:type ln:LoanApplication
ln:attr_addr_state CA
ln:attr_annual_inc "50000" %% xsd:float
ln:attr_application_type Joint App
ln:attr_emp_length 4 years
ln:attr_emp_title Lead Vet Tech
ln:attr_home_ownership RENT
ln:attr_loan_amnt "12025" %% xsd:float
ln:attr_purpose credit_card
ln:attr_term 36 months
ln:attr_title Credit card refinancing
ln:attr_zip_code 950xx
prov:type ln:FICOScore
ln:created_at 2019-03-10T11:00:00
ln:fico_range_high "714" %% xsd:float
ln:fico_range_low "710" %% xsd:float
prov:type ln:CreditReference
ln:attr_earliest_cr_line Sep-2010
ln:attr_mort_acc "0" %% xsd:float
ln:attr_open_acc "5" %% xsd:float
ln:attr_pub_rec "0" %% xsd:float
ln:attr_pub_rec_bankruptcies "0" %% xsd:float
ln:attr_revol_bal "4140" %% xsd:float
ln:attr_revol_util "18.6" %% xsd:float
ln:created_at 2019-01-10T14:10:16
wasDerivedFrom
wasDerivedFrom
wasDerivedFrom
ex:decision/128350251
ex:review_recommendation/128350251
prov:type ln:HumanDecision
ln:result ln:approved
prov:type ln:LoanAssessment
prov:type pl:HumanLedActivity

Figure 10. Provenance graph extracted by the query for Data Currency in the loan scenario.

Out of the extracted provenance subgraph, one can then construct the following explanation. It explicitly lists the external credit referencing agencies, the credit references they provided, and the time at which such credit references were obtained.

> The external data sources were the borrower FICO score (fico_score/29) provided by the credit referencing agency (fico) at <2019-06-15T20:46:39.921182> and the borrower credit reference (credit_history/29) provided by the credit referencing agency (credit_agency) at <2019-06-20T12:43:31.114156>.

The explanation can then further be enriched by providing contact details for these external agencies. Taking our approach of providing *explanations as detective controls* (see Section 1), the explanation allows a data subject to decide if external information is timely; the explanation is actionable, as a data subject can use the contact details to approach these agencies, query the credit reference, and potentially have it fixed.

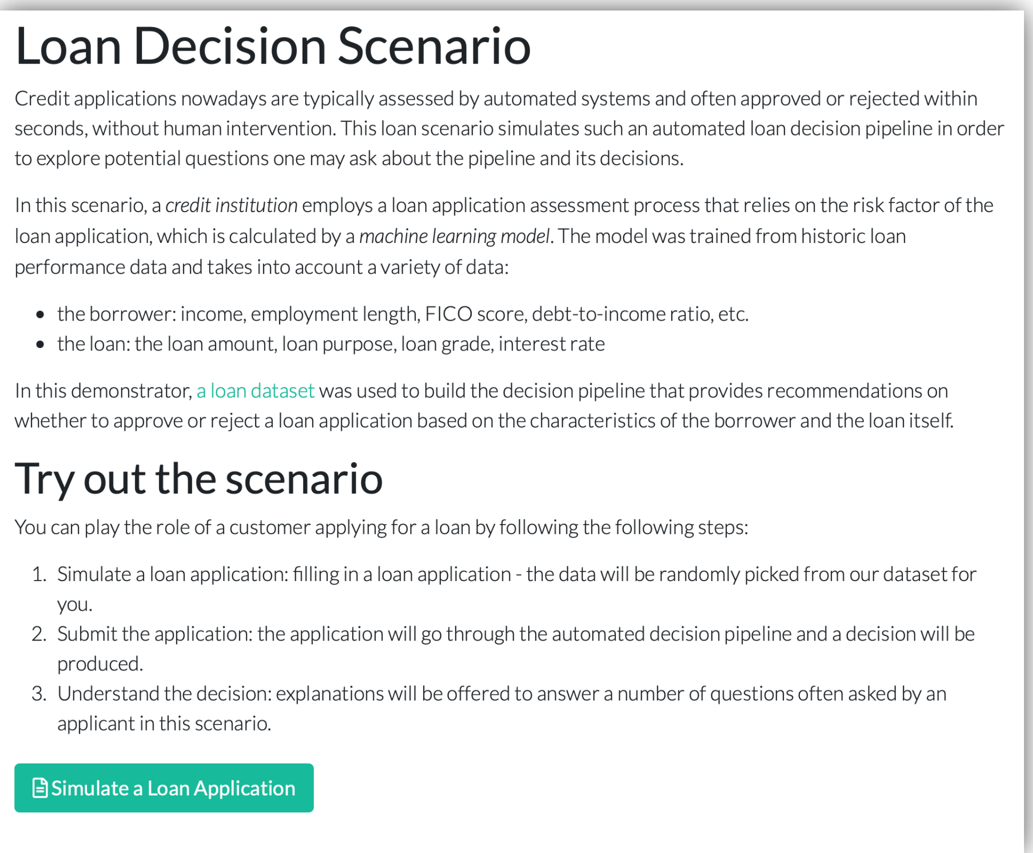
A standard such as PROV allows multiple organisations to share provenance information in an interoperable manner. In this simulated example, the provenance is created by the loan company, based on the processes it follows. If the loan company and credit referencing agencies inter-operate properly, we can envisage that not only they exchange credit reference but also (some of) the provenance of these credit references. Credit reference agencies themselves rely on some other external organisations to compile these references. All that provenance information can be transformed into actionable information for the data subject.

6. The Demonstrator

Following the mentioned approach, we have implemented a demonstrator to generate explanations for most of the explanation categories we identify in Section 3 above for the loan decision scenario. The demonstrator is deployed online at the below web address:

`explain.openprovenance.org`

The website currently presents a single scenario, the loan scenario (Section 2), in which you can play the role of a borrower submitting a loan application to a lender. At the end of the process, you will receive a (simulated) decision, at which point you will be also offered a number of explanations with respect to the decision.



Loan Decision Scenario

Credit applications nowadays are typically assessed by automated systems and often approved or rejected within seconds, without human intervention. This loan scenario simulates such an automated loan decision pipeline in order to explore potential questions one may ask about the pipeline and its decisions.

In this scenario, a *credit institution* employs a loan application assessment process that relies on the risk factor of the loan application, which is calculated by a *machine learning model*. The model was trained from historic loan performance data and takes into account a variety of data:

- the borrower: income, employment length, FICO score, debt-to-income ratio, etc.
- the loan: the loan amount, loan purpose, loan grade, interest rate

In this demonstrator, a loan dataset was used to build the decision pipeline that provides recommendations on whether to approve or reject a loan application based on the characteristics of the borrower and the loan itself.

Try out the scenario

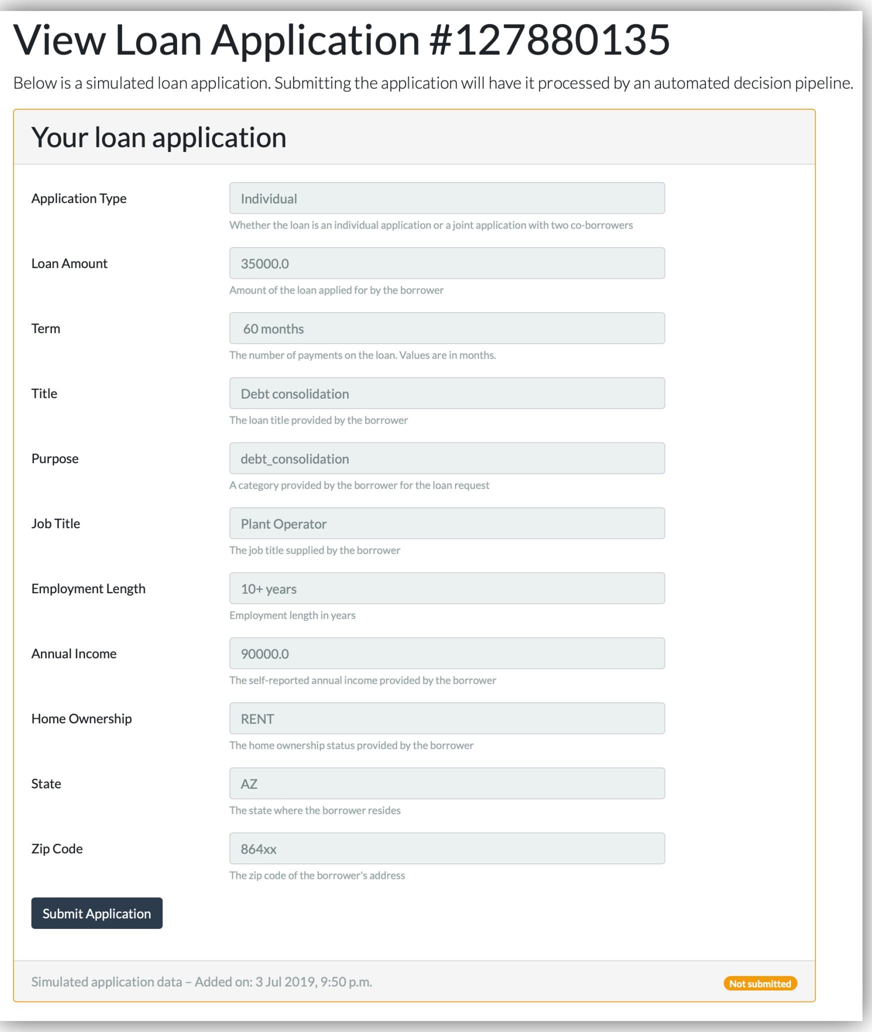
You can play the role of a customer applying for a loan by following the following steps:

1. Simulate a loan application: filling in a loan application - the data will be randomly picked from our dataset for you.
2. Submit the application: the application will go through the automated decision pipeline and a decision will be produced.
3. Understand the decision: explanations will be offered to answer a number of questions often asked by an applicant in this scenario.

Figure 11. The loan decision scenario provided by the demonstrator.

Simulate a loan application

Clicking on the *Simulate a Loan Application* button (Figure 11) will present you with a new loan application form whose data fields are readily filled in for you (as shown in Figure 12). You can then submit the application when ready.



View Loan Application #127880135

Below is a simulated loan application. Submitting the application will have it processed by an automated decision pipeline.

Your loan application

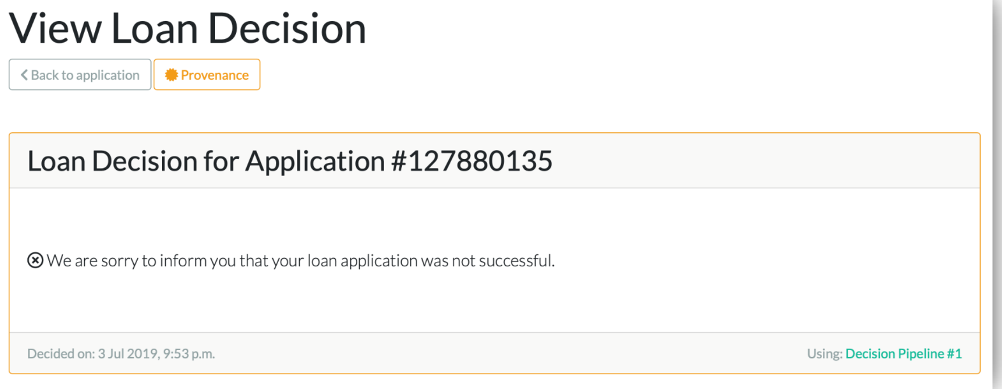
Application Type	Individual	Whether the loan is an individual application or a joint application with two co-borrowers
Loan Amount	35000.0	Amount of the loan applied for by the borrower
Term	60 months	The number of payments on the loan. Values are in months.
Title	Debt consolidation	The loan title provided by the borrower
Purpose	debt_consolidation	A category provided by the borrower for the loan request
Job Title	Plant Operator	The job title supplied by the borrower
Employment Length	10+ years	Employment length in years
Annual Income	90000.0	The self-reported annual income provided by the borrower
Home Ownership	RENT	The home ownership status provided by the borrower
State	AZ	The state where the borrower resides
Zip Code	864xx	The zip code of the borrower's address

Simulated application data – Added on: 3 Jul 2019, 9:50 p.m.

Figure 12. An example loan application form.

View the loan decision and its explanations

The loan decision will be presented shortly after an application is submitted (Figure 13).



View Loan Decision

Loan Decision for Application #127880135

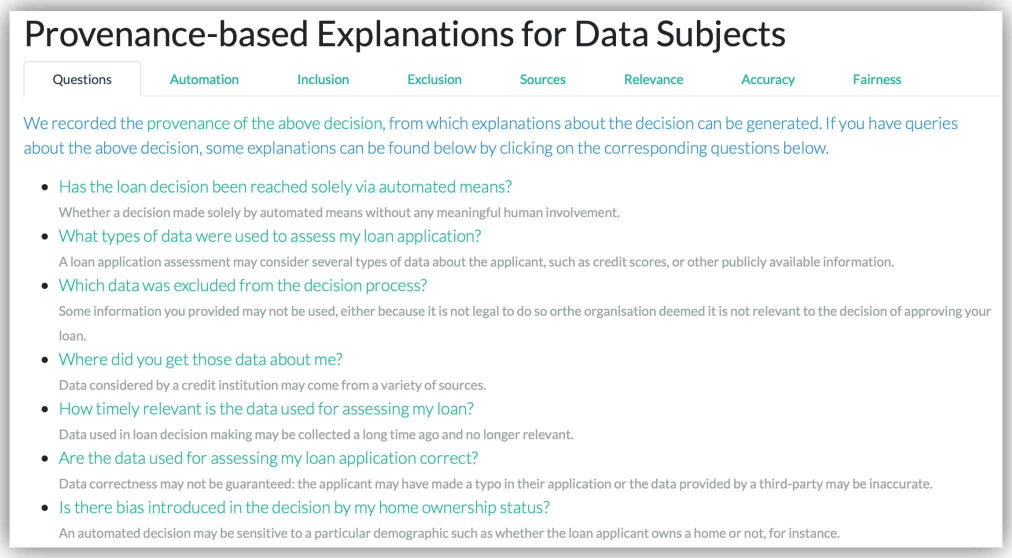
⊗ We are sorry to inform you that your loan application was not successful.

Decided on: 3 Jul 2019, 9:53 p.m.

Using: Decision Pipeline #1

Figure 13. The resulted decision of a loan application. The full provenance of the decision is available via the ***Provenance*** *button.*

Below the loan decision, you will find a list of questions (Figure 14) that an applicant may inquire about its various aspects.



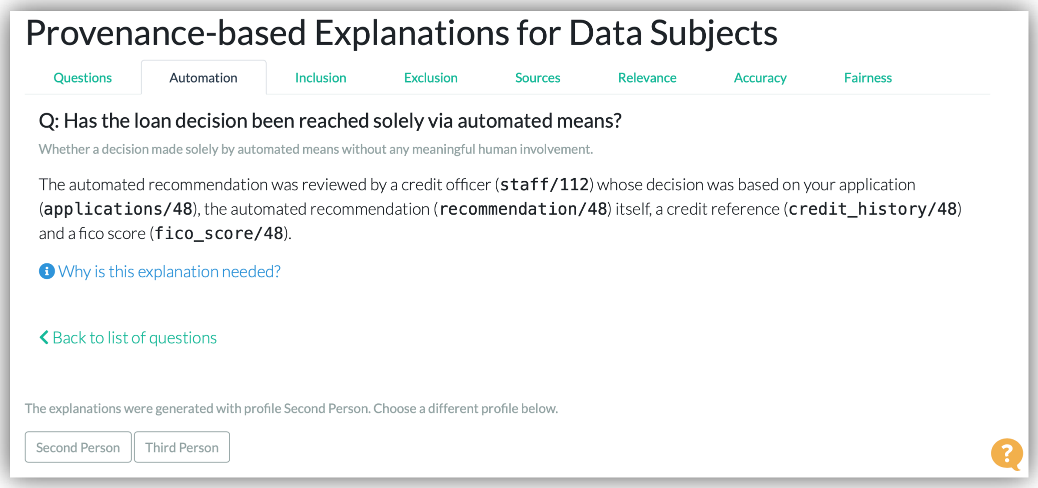
Provenance-based Explanations for Data Subjects

We recorded the provenance of the above decision, from which explanations about the decision can be generated. If you have queries about the above decision, some explanations can be found below by clicking on the corresponding questions below.

- Has the loan decision been reached solely via automated means?
  Whether a decision made solely by automated means without any meaningful human involvement.
- What types of data were used to assess my loan application?
  A loan application assessment may consider several types of data about the applicant, such as credit scores, or other publicly available information.
- Which data was excluded from the decision process?
  Some information you provided may not be used, either because it is not legal to do so orthe organisation deemed it is not relevant to the decision of approving your loan.
- Where did you get those data about me?
  Data considered by a credit institution may come from a variety of sources.
- How timely relevant is the data used for assessing my loan?
  Data used in loan decision making may be collected a long time ago and no longer relevant.
- Are the data used for assessing my loan application correct?
  Data correctness may not be guaranteed: the applicant may have made a typo in their application or the data provided by a third-party may be inaccurate.
- Is there bias introduced in the decision by my home ownership status?
  An automated decision may be sensitive to a particular demographic such as whether the loan applicant owns a home or not, for instance.

Figure 14. The list of questions an applicant may ask about a loan decision.

For instance, they may ask whether the decision was solely automated. The *Automation* tab (also accessible by clicking on the corresponding question) will provide the answer, which is generated from the provenance data of the decision (see Figure 15). Below each explanation, we provide its legal and business contexts that call for the explanation.



Provenance-based Explanations for Data Subjects

Q: Has the loan decision been reached solely via automated means?

Whether a decision made solely by automated means without any meaningful human involvement.

The automated recommendation was reviewed by a credit officer (`staff/112`) whose decision was based on your application (`applications/48`), the automated recommendation (`recommendation/48`) itself, a credit reference (`credit_history/48`) and a fico score (`fico_score/48`).

ℹ Why is this explanation needed?

‹ Back to list of questions

The explanations were generated with profile Second Person. Choose a different profile below.

*Figure 15. The **Automation** explanation.*

7. Conclusions

Over this EPSRC impact acceleration project, over a period of three months, we have implemented the Loan Decision scenario, instrumented the pipeline so that it produces provenance, categorised explanations according to their audience and their purpose, built an explanation-generation prototype, and wrapped the whole system in an online demonstrator. This work aimed to demonstrate that provenance, defined as *a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering* a decision, is a solid foundation for generating its explanations.

Given the limited time available, however, there are inevitably some limitations.

- First, we designed this prototype for one application scenario, for one machine learning pipeline, for one specific regulatory framework (GDPR), and for a subset of requirements from this framework. It is our intent to generalise the approach to other scenarios, regulations and requirements.

- Second, the approach is predicated on finding certain mark-ups in the provenance to be able to construct the relevant explanations. Besides the above generalisation, there is also a clear need to document such mark-ups, so that data controllers can adapt their system to produce suitably annotated provenance. It has to be understood by data controllers that a failure to generate provenance with the right mark-ups will result in the system's inability of constructing some explanations.

- Third, adequate tools need to be provided to assist data controllers in producing the right provenance information and in checking that it addresses data protection (or others) requirements they are under the obligation to meet.

- Fourth, explanations can and should be refined to fully meet their purposes. Suitable requirement capturing and user studies will help validate these.

- Fifth, it is our belief that explanations could be viewed as more than just one paragraph communicated to the data subject in a single request-response interaction. We envisage explanations potentially as part of a dialogue between the system and its targeted recipients. A mechanism to design such an explanation service would, therefore, be required.

- Finally, some aspects of the decision-making pipeline are currently not explained. It is particularly the case of the machine learning algorithm itself, which remains a black-box: the algorithm was used to create a model, and the model was used to classify some input data. Both the model creation and classification are modelled by activities in the provenance. If some libraries are able to generate further provenance, this, in turn, can be turned into explanations.

In addition, the work has also opened up a number of interesting research questions that require further investigation.

- **Automation**. We generate different explanations for automated and human decisions. Something to investigate is how *meaningful* the human involvement is. How much is added by the human on top of the automated recommendation they proceed? Can the meaningfulness be determined automatically? Which semantic mark-up in the provenance would help with this task?

- **Exclusion**. We were able to demonstrate that some loan application characteristics (or elements of third-party data such as credit reference) were not used by the decision-making pipeline. This information, while certainly useful, is looking at "syntactic usage": certain data may have been passed to the pipeline but may or may not have been effectively used to reach the decision. In other words, the data may or may not have had an influence on the final decision. However, such information can only be surfaced if we gain a better understanding of the black box.

- **Counter-factual explanations**. We have demonstrated that it is possible to construct simple counter-factual explanations out of provenance. By simply considering alternate loan applications in a counter-factual world (e.g. loans for a different purpose, for a different amount, for a data subject with different profile), and applying the pipeline, we obtain counter-factual decisions. By marking the original loan application and associated decision, as well as alternate applications and theirs, we were able to construct an example of counter-factual explanation. This approach needs to be generalised and the nature of explanations that can be supported needs to be studied.

Next Step

This limited proof-of-concept exercise is only the start of a journey. With the new EPSRC-funded project to start in September 2019 on *Provenance-driven and Legally-grounded Explanations for Automated Decisions* (PLEAD), we are about to embark in novel research to address some of the above concerns. Information about the PLEAD project is available at the following website, where new research findings will also be published.

plead-project.org

8. Acknowledgement

This project is funded in part by the EPSRC Impact Acceleration Account from King’s College London.

The project was conducted in collaboration with members of the Information Commissioner’s Office. The authors want to thank the following for their constructive contributions: Carl Wiper, Robert McCombe, Elanor McCombe, Alex Hubbard, Valeria Gallo, Joel Talman, and Reuben Binns.

9. References

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review*, *16*(1), 1–65. https://doi.org/10.2139/ssrn.2972855

European Union. (2016). Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, *59*(L 199), 1–88. Retrieved from http://data.europa.eu/eli/reg/2016/679/oj

Huynh, T. D., Ebden, M., Fischer, J., Roberts, S., & Moreau, L. (2018). Provenance Network Analytics. *Data Mining and Knowledge Discovery*, *32*(3), 708–735. https://doi.org/10.1007/s10618-017-0549-3

Moreau, L. (2015). Aggregation by Provenance Types: A Technique for Summarising Provenance Graphs. *Graphs as Models 2015 (An ETAPS'15 Workshop), Electronic Proceedings in Theoretical Computer Science*, 129–144. https://doi.org/10.4204/EPTCS.181.9

Moreau, L., Batlajery, B. V., Huynh, T. D., Michaelides, D., & Packer, H. (2018). A Templating System to Generate Provenance. *IEEE Transactions on Software Engineering*, *44*(2), 103–121. https://doi.org/10.1109/TSE.2017.2659745

Moreau, L., & Groth, P. (Eds.). (2013). *PROV-Overview: An Overview of the PROV Family of Documents*. Retrieved from http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Moreau, L., & Missier, P. (2013). PROV-DM: The PROV Data Model. In *W3C Recommendation*. Retrieved from http://www.w3.org/TR/prov-dm/

Ramchurn, S. D., Huynh, T. D., Wu, F., Ikuno, Y., Flann, J., Moreau, L., … Jennings, N. R. (2016). A Disaster Response System based on Human-Agent Collectives. *Journal of Artificial Intelligence Research*, *57*, 661–708. https://doi.org/10.1613/jair.5098

Selbst, A. D., & Powles, J. (2017). Meaningful Information and the Right to Explanation. *International Data Privacy Law*, *7*(4), 233–242. Retrieved from https://ssrn.com/abstract=3039125

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2). https://doi.org/10.2139/ssrn.3063289

Weitzner, D. J., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., & Sussman, G. J. (2008). Information accountability. *Communications of the ACM*, *51*(6), 82–87. https://doi.org/10.1145/1349026.1349043