Explanation by Example

The OptaPlanner way

Daniele Zonca
Principal Software Engineer

Tommaso Teofili
Principal Software Engineer

Rui Vieira
Software Engineer

Red Hat

Business Automation

Decision

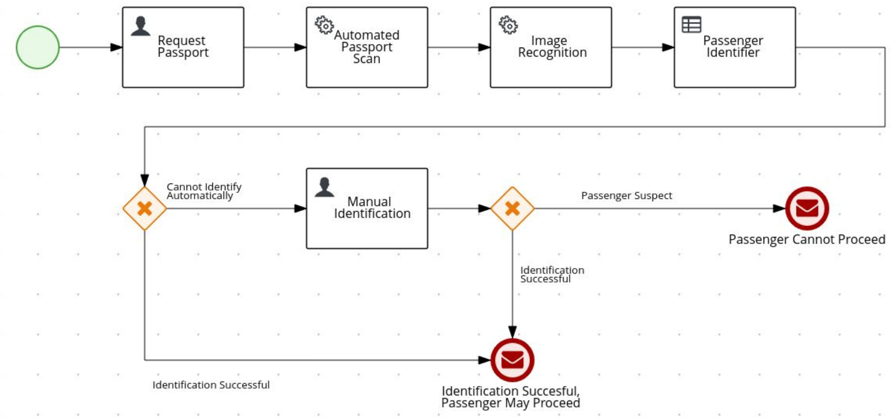
Drools

Automatically processed? *(Decision Table)*

F	Calculate Trip risk *(number)*	Image score *(number)*	Automatically processed? *(boolean)*	Description
1	>0.8	-	false	Trip risk is too high
2	-	<0.7	false	Passport image is too different
3	-	-	true	Fine to proceed

Process

jBPM



Vehicle Routing

Mathematical Optimization

OptaPlanner

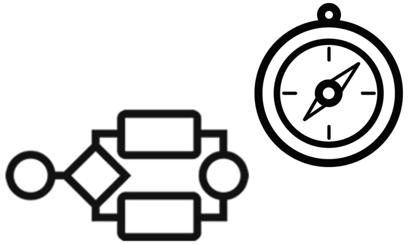
Red Hat

Next-gen Cloud-Native Business Automation

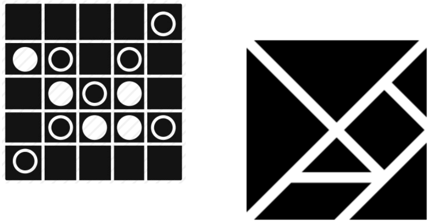
Cloud-Native Business Automation for building intelligent applications, backed by battle-tested capabilities

Cloud-Native Business Automation

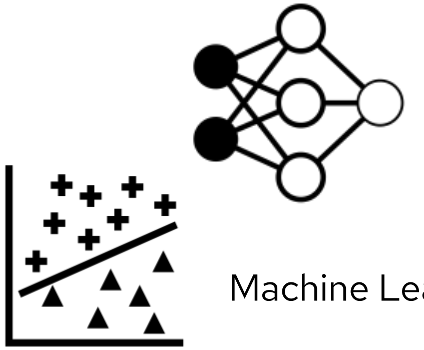
Knowledge as a Service



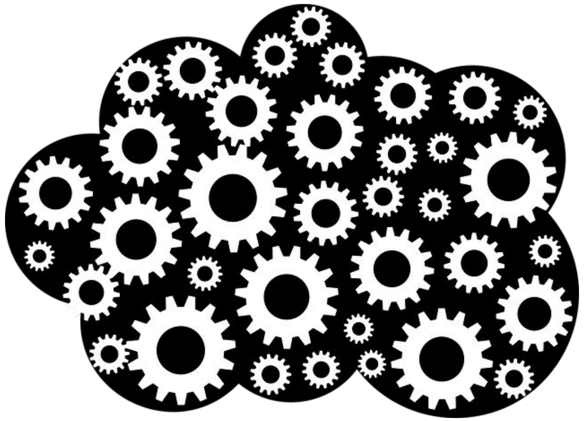
Workflow and Digital Decisioning



Mathematical Optimization



Machine Learning



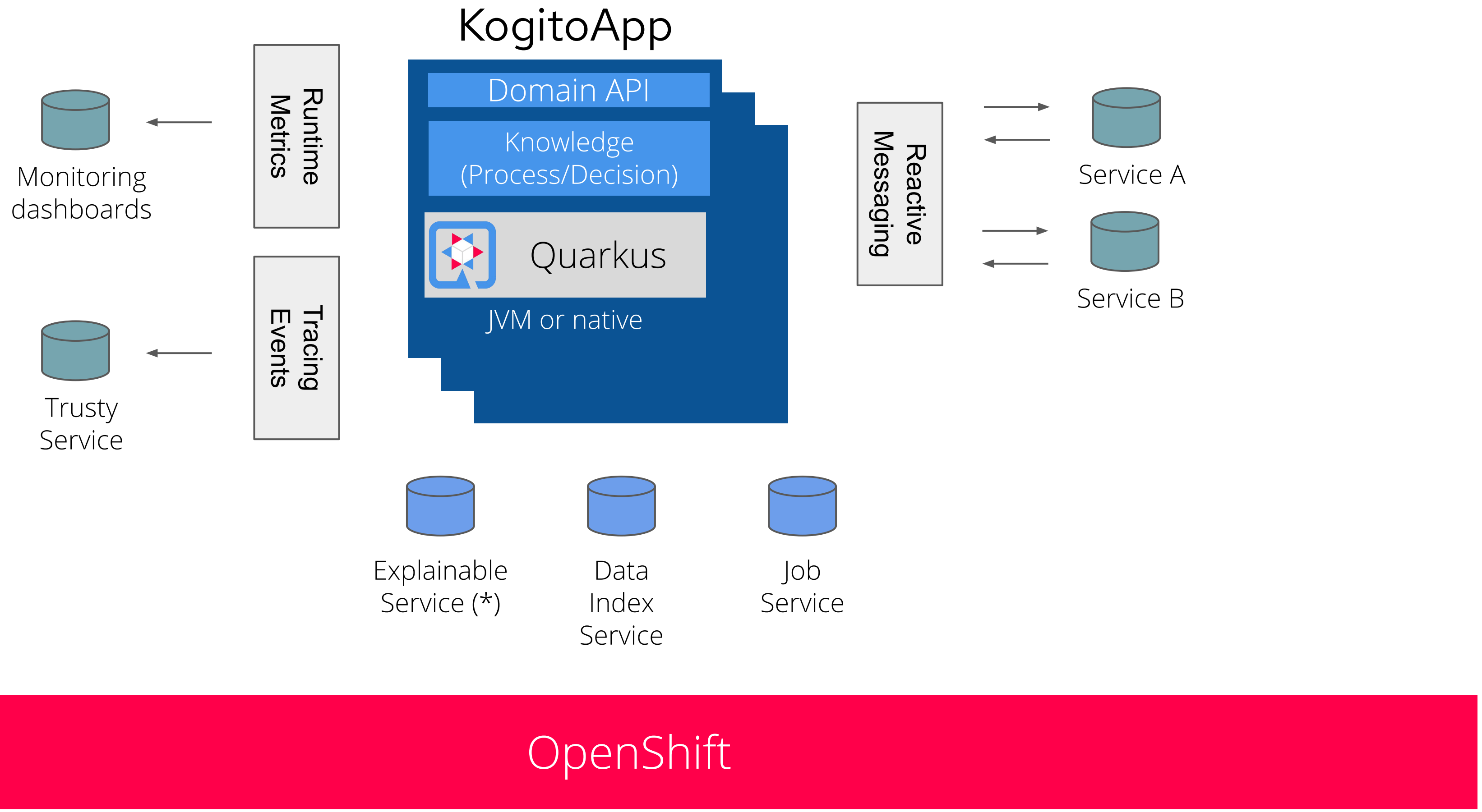
Knowledge as a Service

TrustyAI

Offer value-added services for Business Automation.

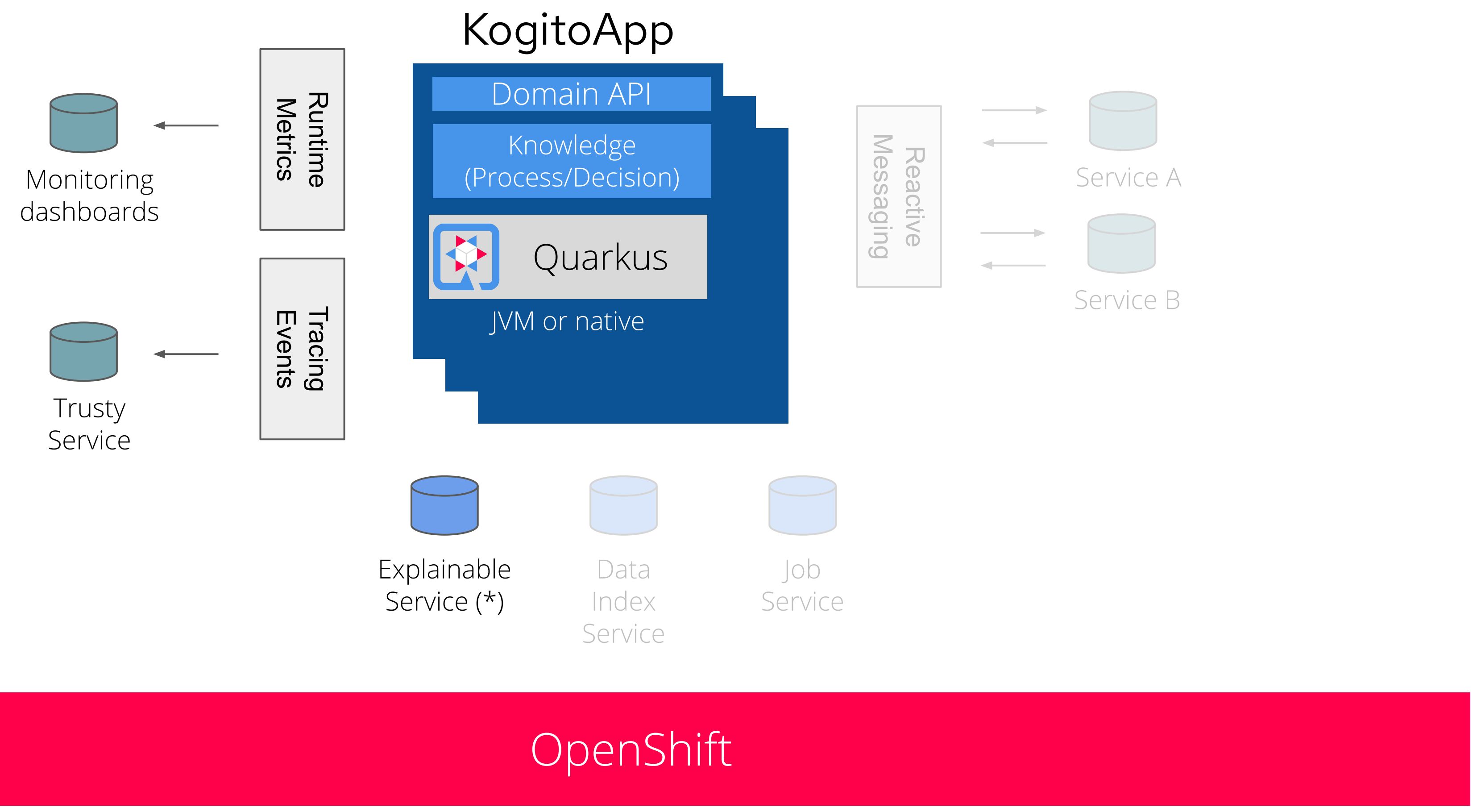
- **Runtime Monitoring Service**
  - dashboard for business runtime monitoring
- **Tracing and Accountability Service**
  - extract, collect and publish metadata for auditing and compliance
- **Explanation Service**
  - XAI algorithms to enrich model execution information

Runtime Ecosystem



(*) Kogito 0.15 (~middle September)

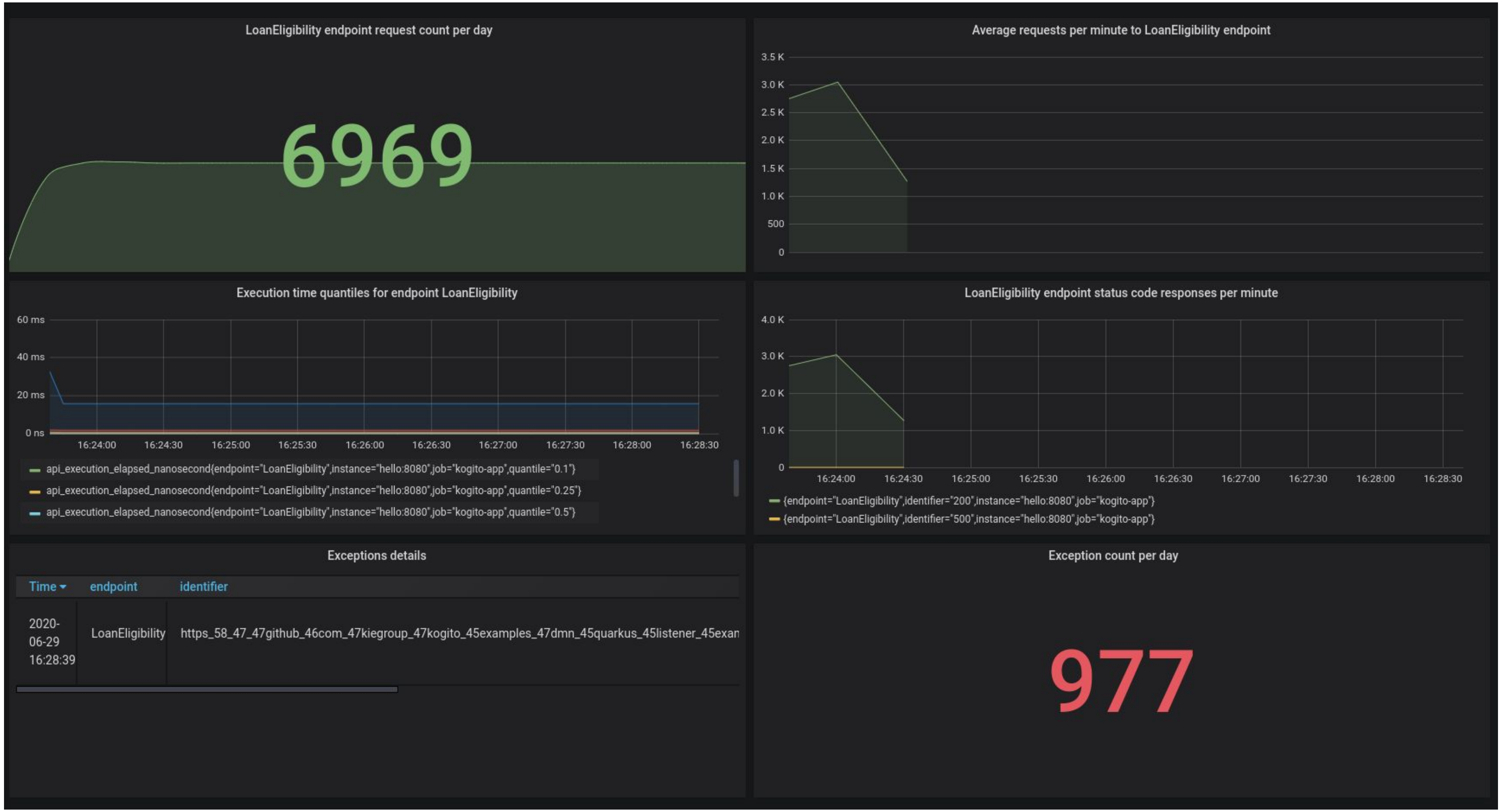
TrustyAI Services



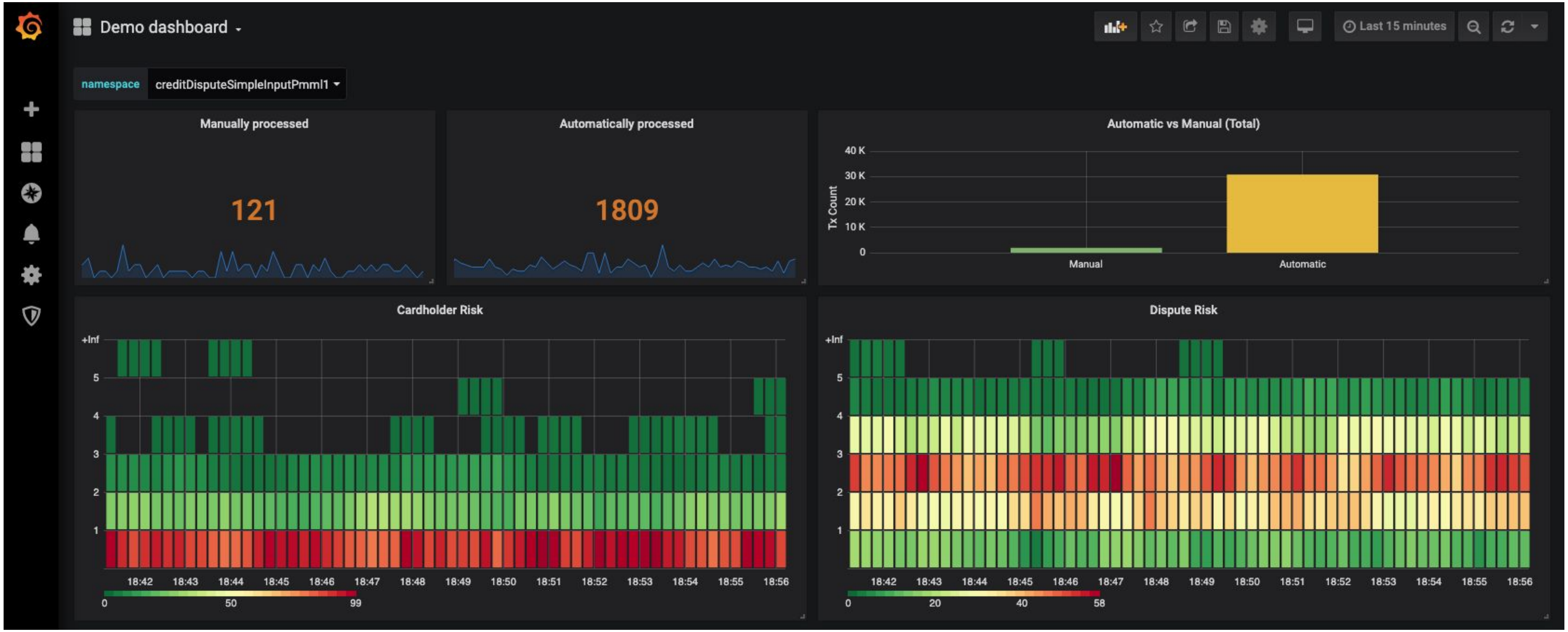
(*) Kogito 0.15 (~middle September)

Red Hat

DevOps Monitoring



Business Monitoring

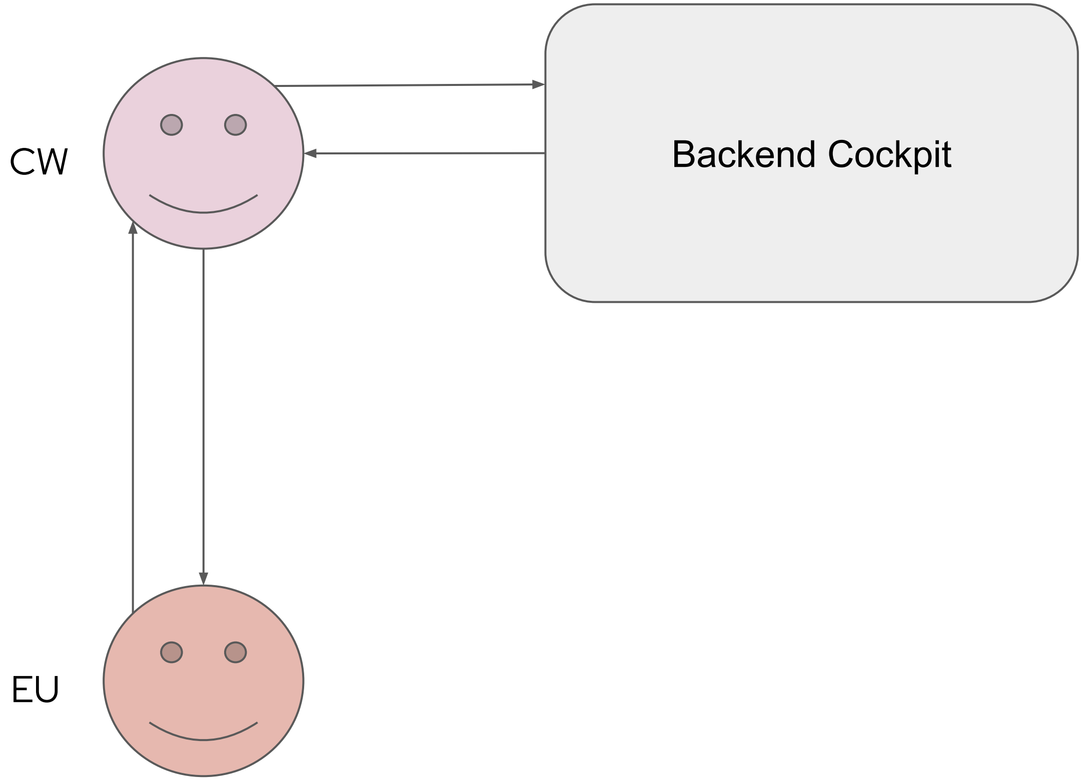


Red Hat

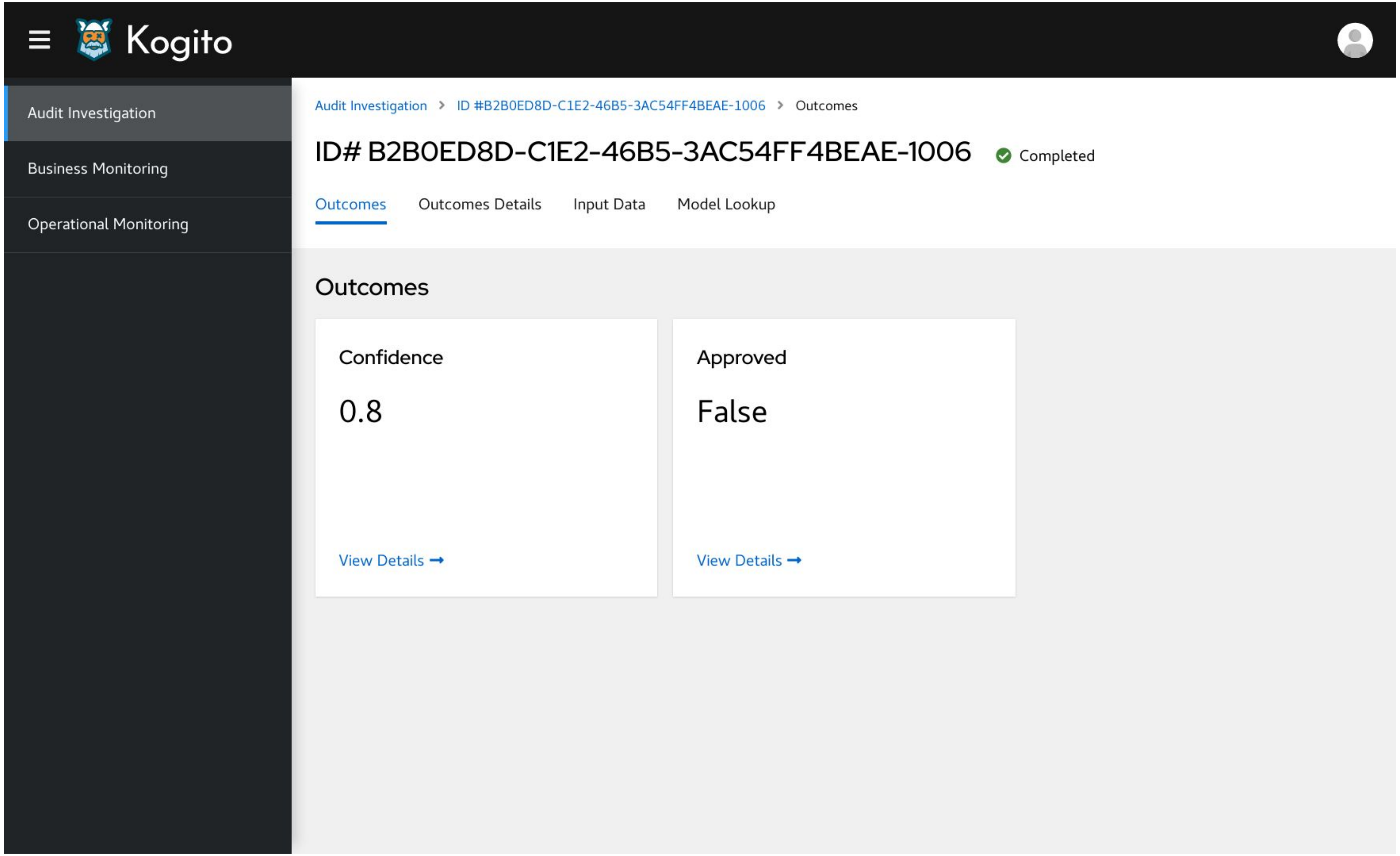
Use case: Credit card approval

"As a case worker (CW) I want to be able to **explain** to end user (EU) **why** that credit card request was rejected or accepted."

"As a case worker (CW) I want to provide information to my end user (EU) about **what is needed** to get it accepted."



Trusty UI (*)

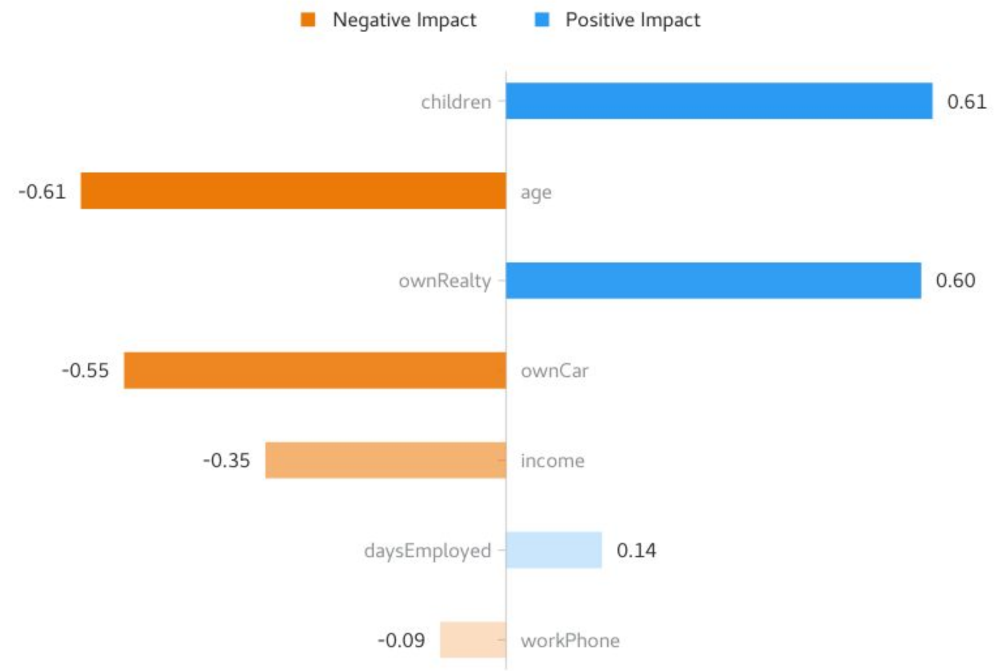


(*) Kogito 0.15 (~middle September)

Trusty UI (*)

Explanation

Features Score Chart



Features Weight

Positive Weight	Score
children	0.61
ownRealty	0.60
daysEmployed	0.14

Negative Weight	Score
age	-0.61
ownCar	-0.55
income	-0.35
workPhone	-0.09

(*) Kogito 0.15 (~middle September)

Red Hat

Explainability - Goals

- Establish **trust** in automated business processes
- **Transparent** decision making when black box models are involved
    - More fine grained **understanding** of specific **predictions**
    - Coarse grained **model behaviour** understanding
- **Accountability**
    - Track changes in model behaviour across versions

My black box model is...

Transparent

A model is considered to be transparent if by itself the model makes a human understand how it works without any need for explaining its internal structure or algorithms

Explainable

A model is explainable if it provides an interface with humans that is both accurate with respect the decision taken and comprehensible to humans

Trustworthy

A model is considered trustworthy when humans are confident that the model will act as intended when facing a given problem

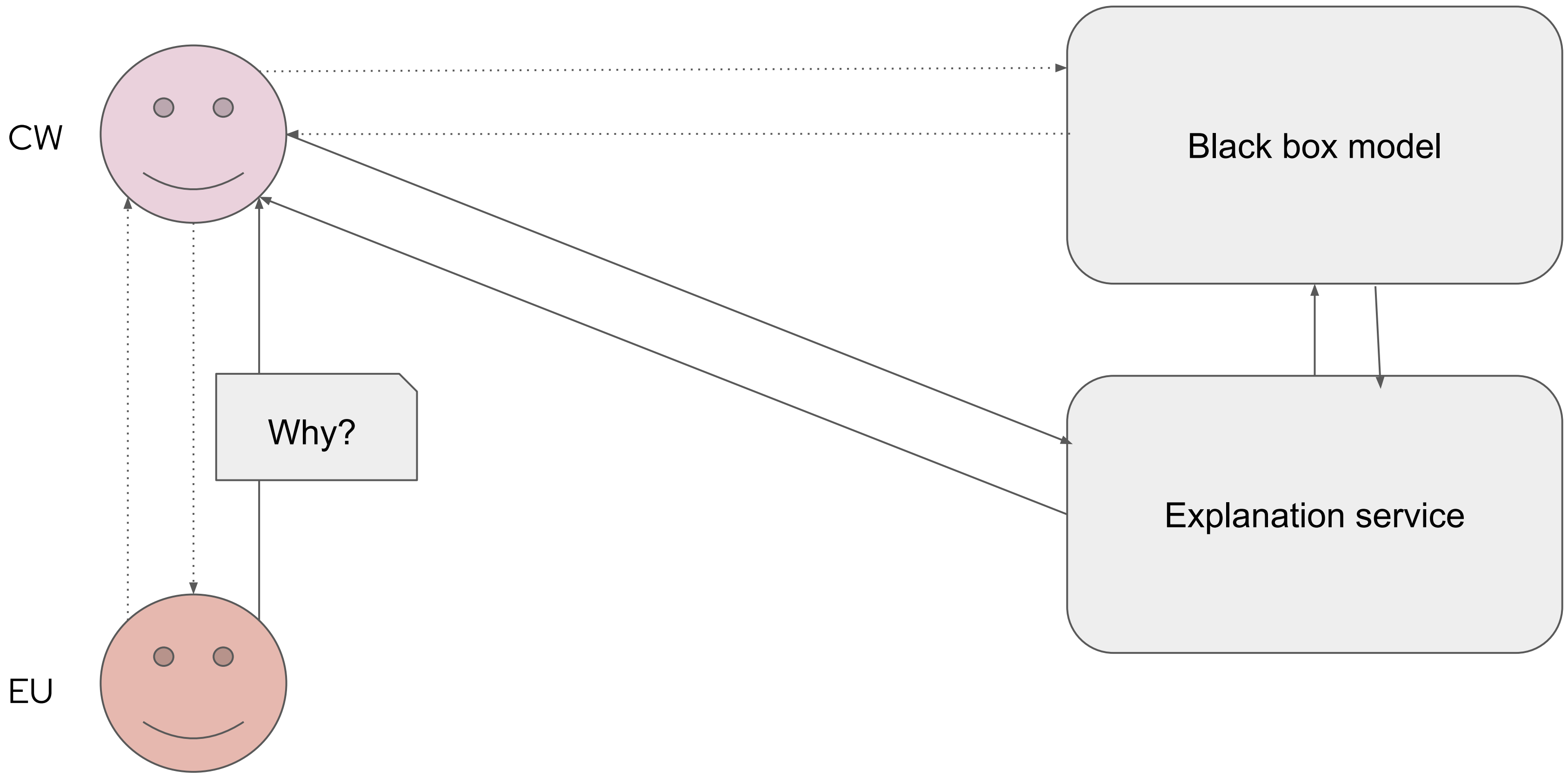
The right explanation to the right stakeholder

- **Case worker**
  - Good domain knowledge, case by case
  - No technical knowledge
- **Compliance worker**
  - Good high level domain knowledge
  - No technical knowledge
- **Data scientist**
  - No/limited domain knowledge
  - Good technical knowledge

The right explanation to the right stakeholder

- **Case worker**
  - Needs explanations on a case by case basis to support end users
    - Local explanations
- **Compliance worker**
  - Needs explanation from a high level perspective (regulations, business objectives, etc.)
    - Global explanations
- **Data scientist**
  - Needs explanations to understand model behavior and debug
    - Global and local explanations

Case worker

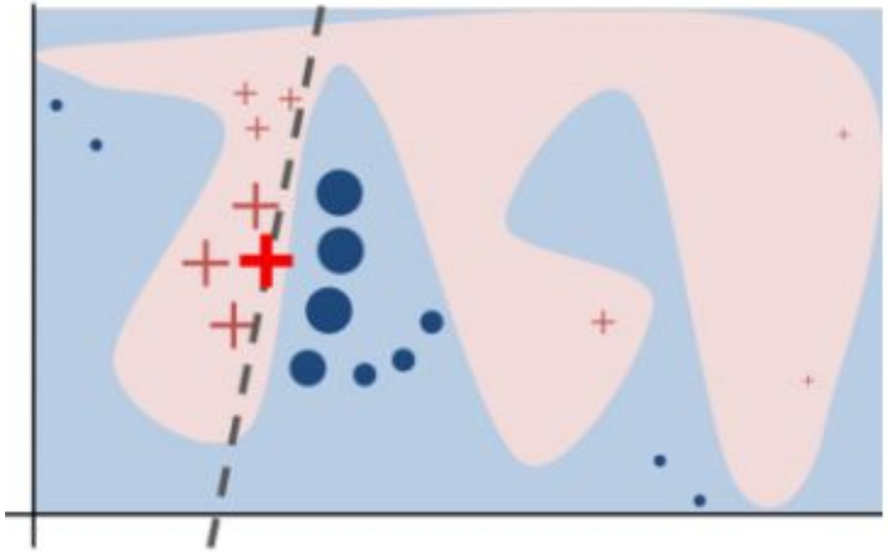


Case worker – Why

- **Need**
  - Which *inputs* does the model give more importance to decide whether to grant the credit card or not?
- **Explanation**
  - **Saliency** explanations give *feature importance* scores for a *single* prediction
    - The value of *children* plays a **positive** role for granting the credit card
    - The value of age *plays* a **negative** role for granting the credit card

Red Hat

LIME



(a) Husky classified as wolf (b) Explanation

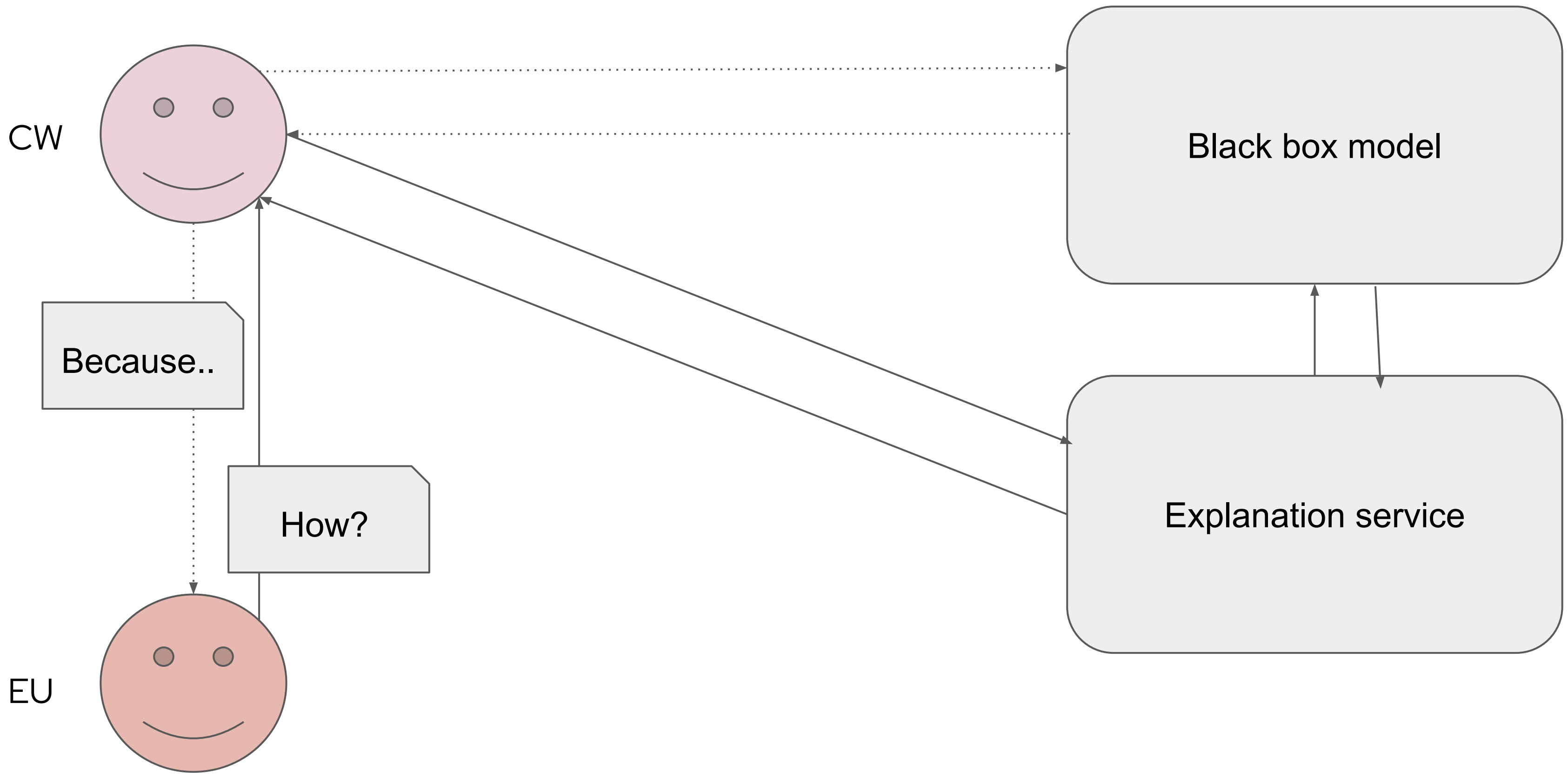
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD, 2016

LIME (*)

- LIME tests what happens to the prediction when you provide *perturbed* versions of the input to the black box model
- Trains an **interpretable** model (e.g. a linear classifier) to separate perturbed data points by label
- The *weights* of the linear model (one for each feature) are used as **feature importance** scores

(*) Kogito 0.15 (~middle September)

Case worker



Case worker - How

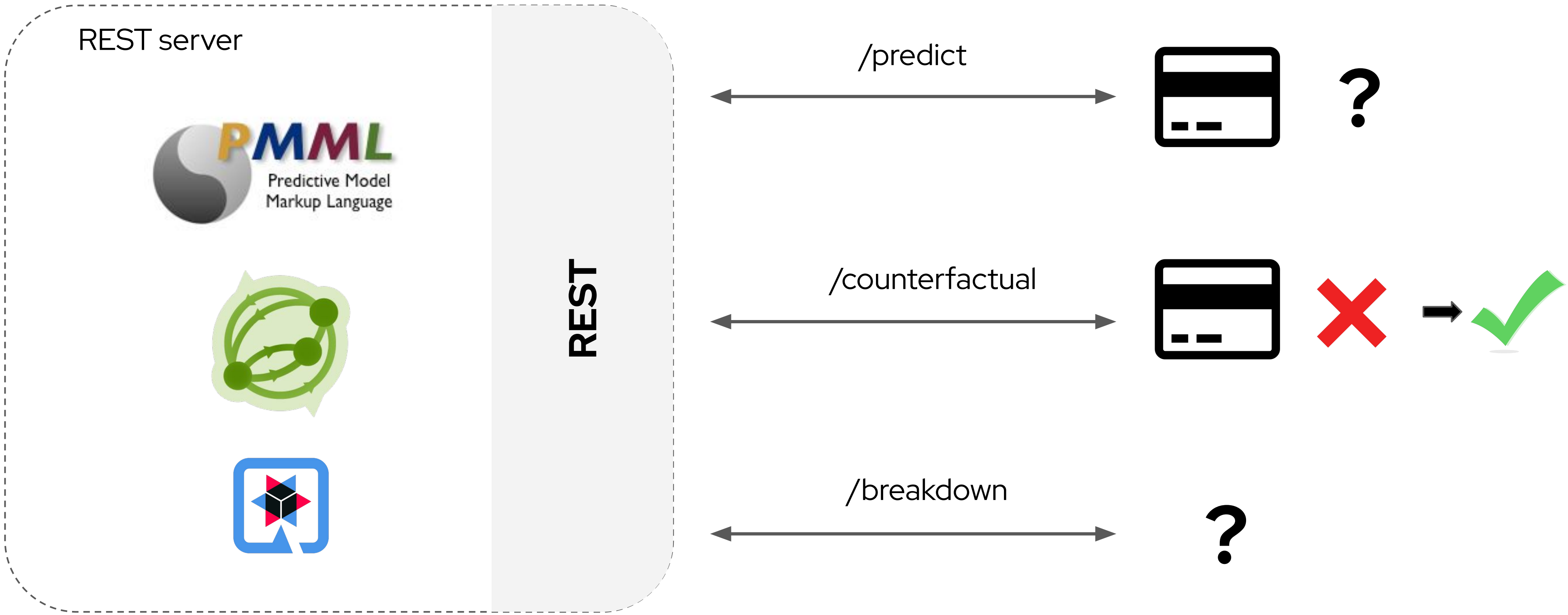
- **Need**
  - What should the end user *change* to get the credit card (similar input, flipped prediction) ?
- **Explanation**
  - **Exemplar** explanations provide explanations for single predictions by means of **examples** (in the input space)
    - **Counterfactual explanations** provide examples that
      - Have a *desired* prediction, according to the black box model
      - Are as *close* as possible to the original input
    - How should the user change its inputs in order to get a formerly rejected credit card request granted ?

Counterfactual explanations

- Usually work by **minimizing** two cost functions
  - **Input cost**
    - representing the distance between the original input and a new input
  - **Target cost**
    - representing the distance between the desired output and the output generated by querying the model with the new input
- Huge **search space**
  - High dimensional inputs
  - Numerical features
  - Out of distribution problems
- **Hard constraints** make the problem worse
  - Some things cannot be (easily) changed by the end user

Demo

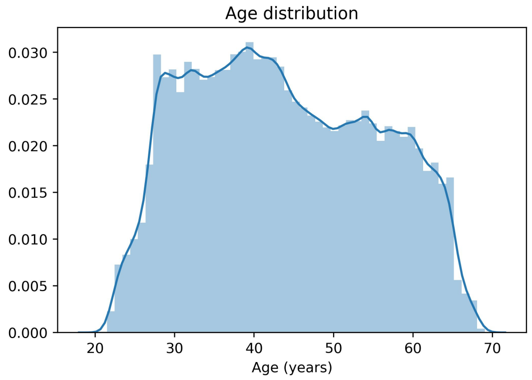
Demo architecture

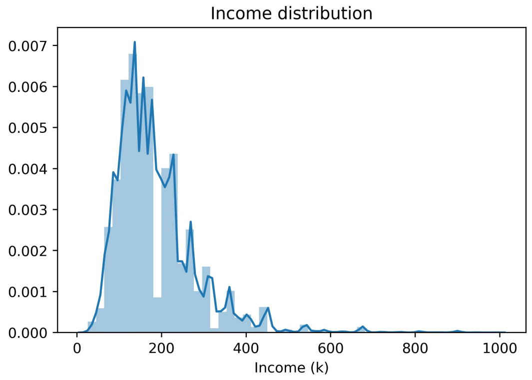


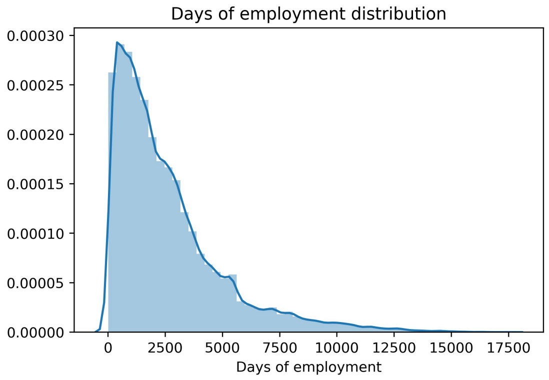
Red Hat

Training dataset

age	income	# children	employment days	owns realty	has work phone	owns car



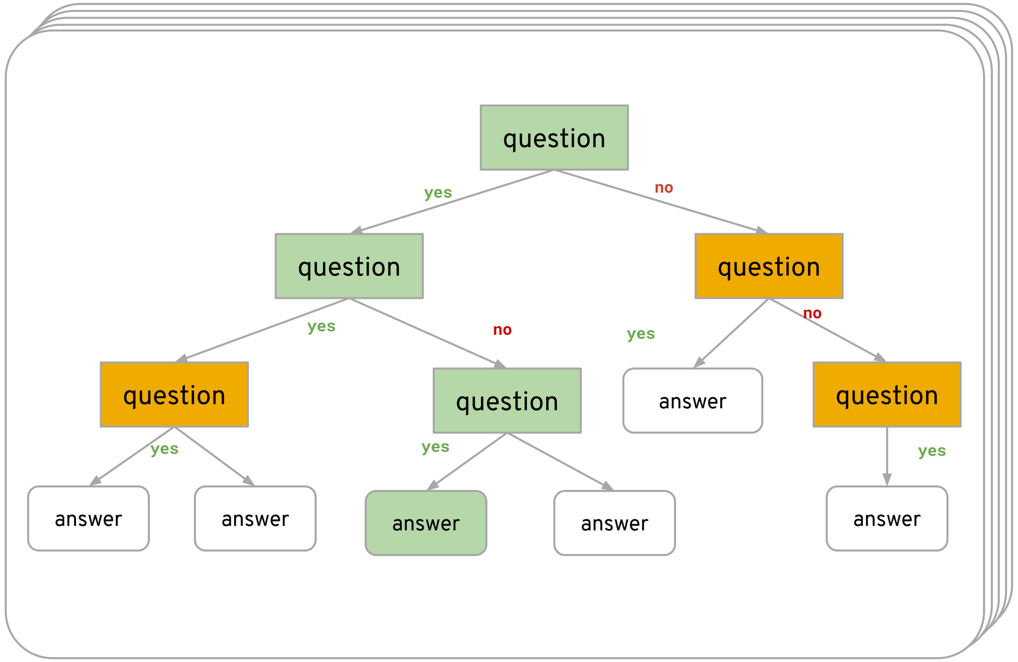




Building the predictive model

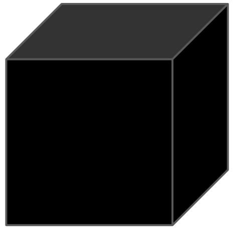
age	income	# children	employment days	owns realty	has work phone	owns car

Random forest classifier

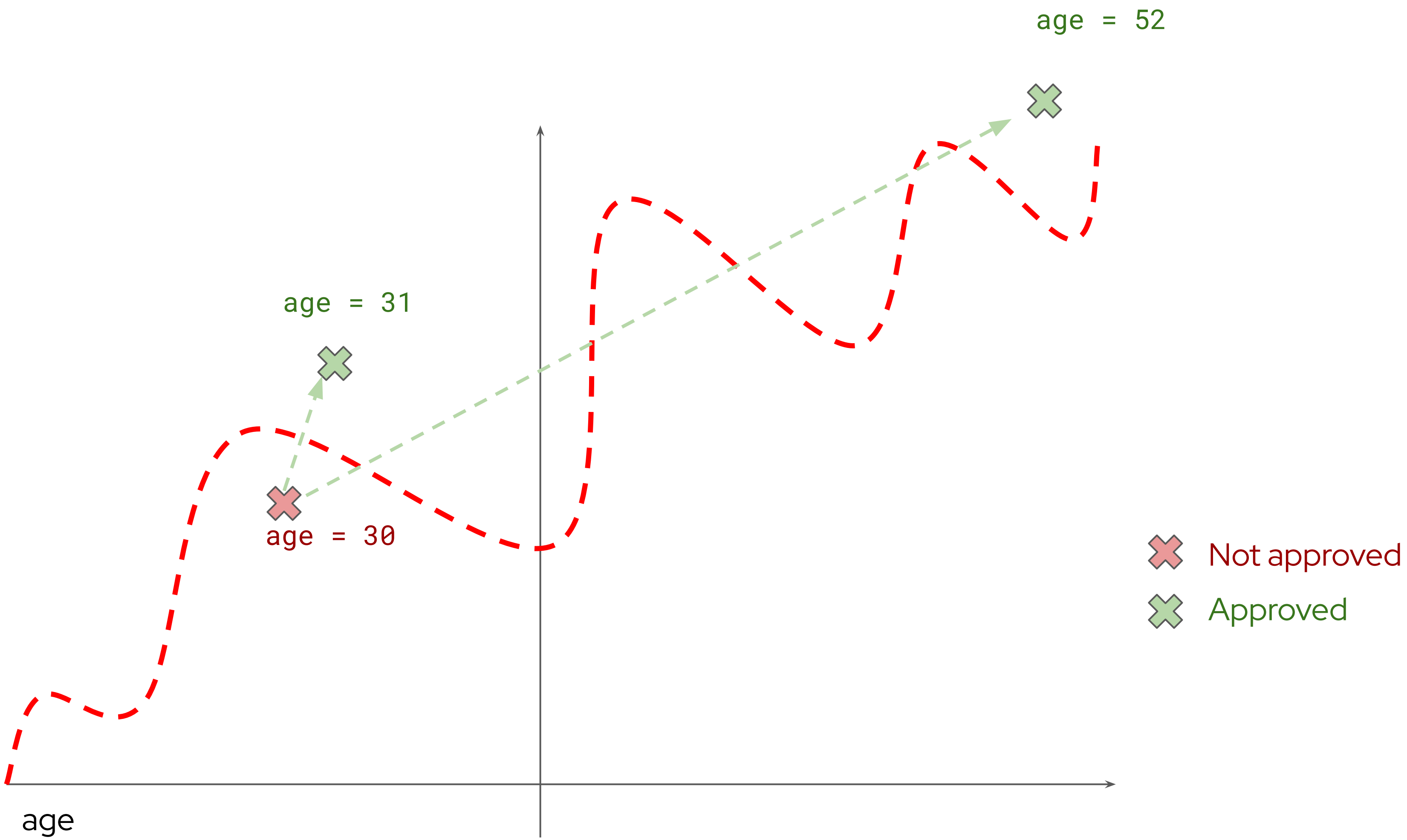


sklearn2pmml

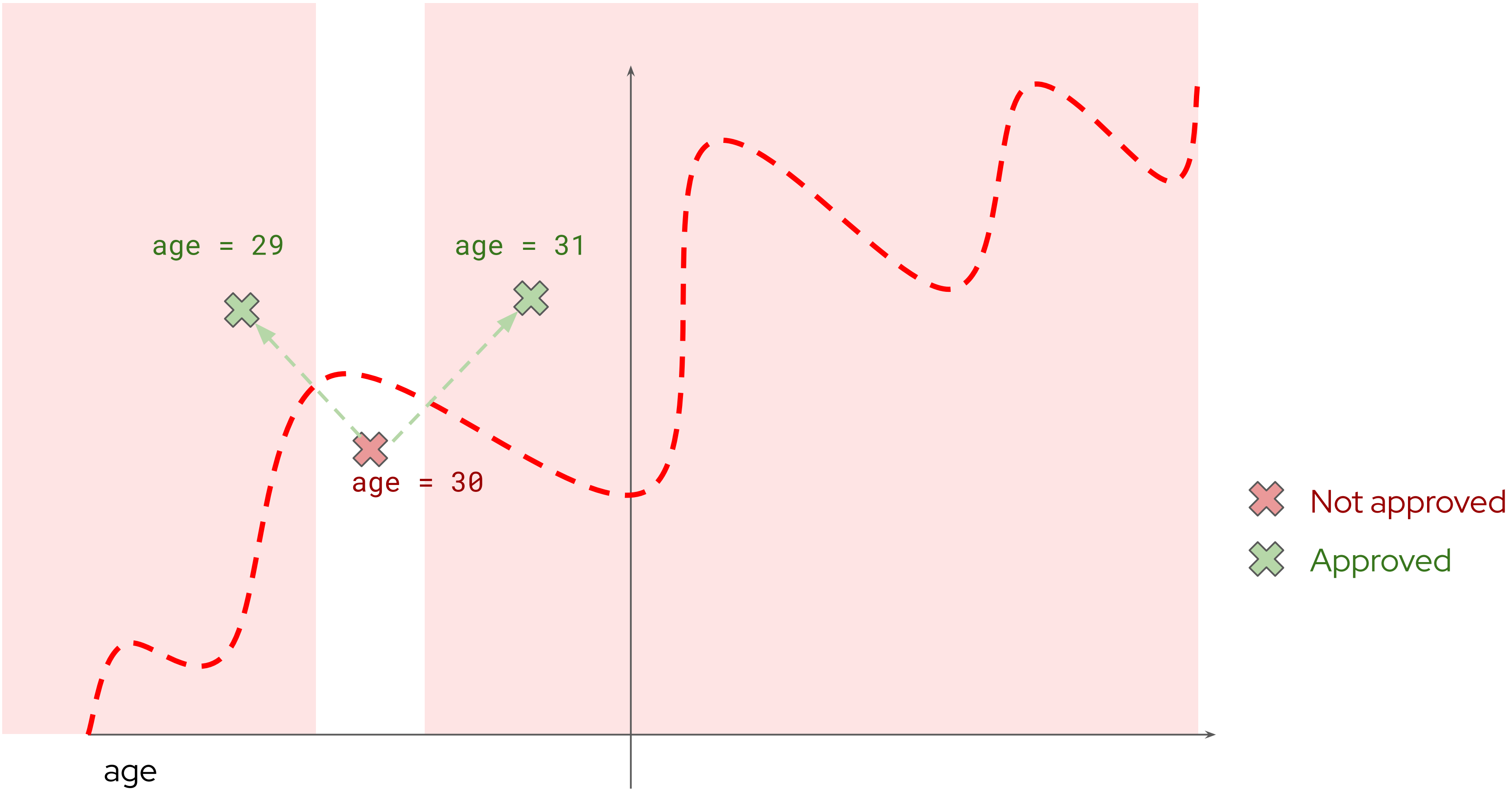
PMML



Domain search space



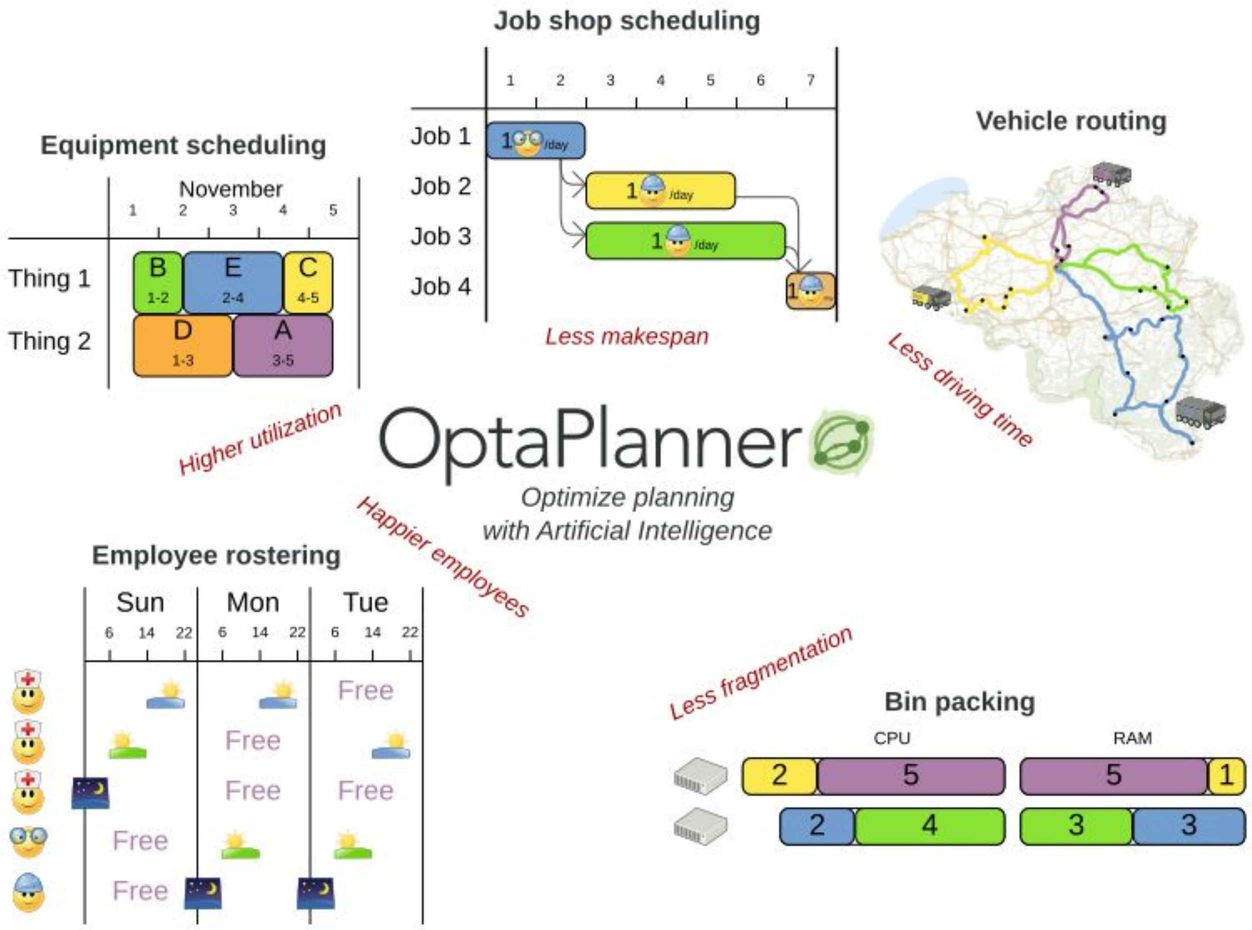
Fixed inputs constraint



Red Hat

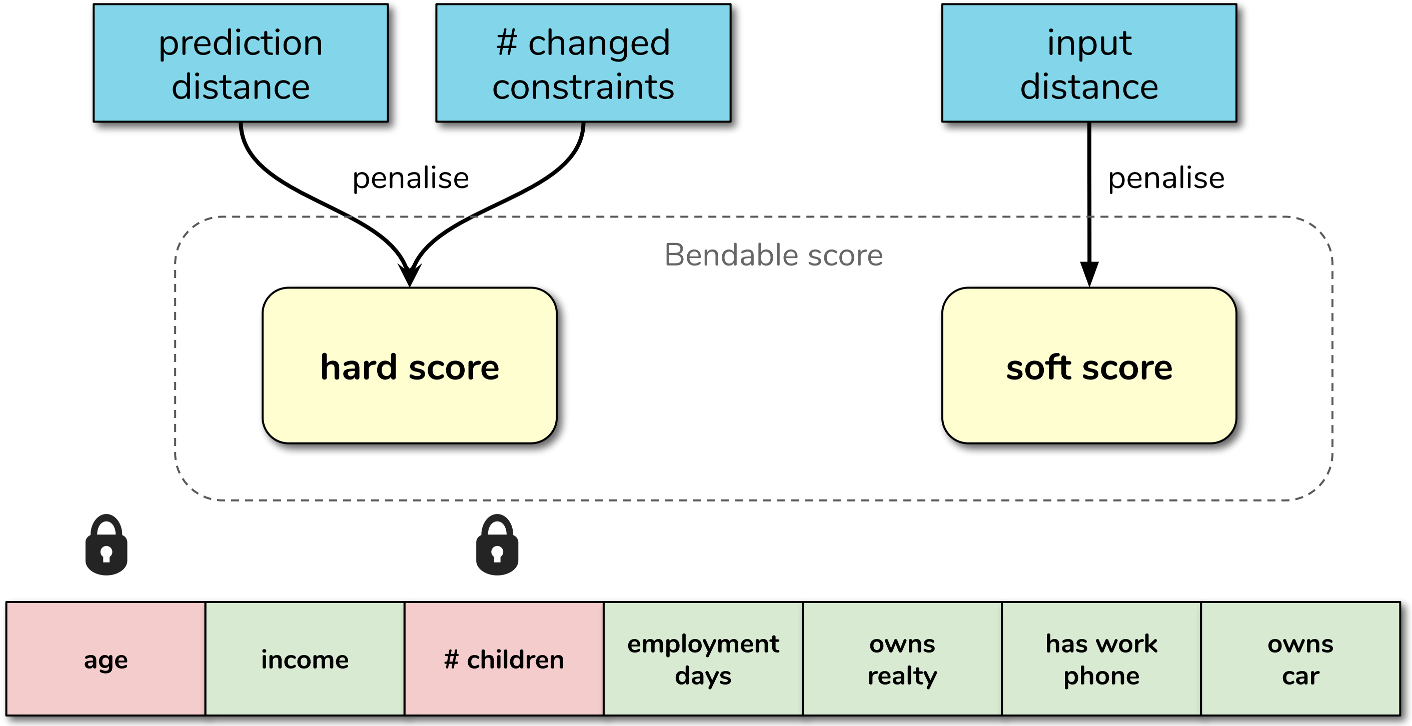
OptaPlanner

- Open source
- Battle-tested constraint solver
- Express complex constraints



Red Hat

Counterfactual solution scoring



Defining constraints and domain

Constraint streaming API

```
public class ApprovalContraintsProvider implements ConstraintProvider{

      private Constraint changedAge(ConstraintFactory constraintFactory) {
       return constraintFactory
           .from(CreditCardApprovalEntity.class)
           .filter(entity -> !entity.getAge().equals(Facts.input.getAge()))
           .penalize(
               "Changed age",
               BendableBigDecimalScore.ofHard(
                   HARD_LEVELS_SIZE, SOFT_LEVELS_SIZE, 1, BigDecimal.valueOf(1)));
      }

}
```

Planning variables

```
@PlanningVariable(valueRangeProviderRefs = {"ageRange"})
public Integer getAge() {
 return age;
}

@PlanningVariable(valueRangeProviderRefs = {"incomeRange"})
public Double getIncome() {
 return income;
}

@PlanningVariable(valueRangeProviderRefs = {"childrenRange"})
public Integer getChildren() {
 return children;
}

...
```

age	18 - 100
income	0 - 1000k
# children	0 - 20
employment days	0 - 18250 (50 years)
owns car	True / False

Red Hat

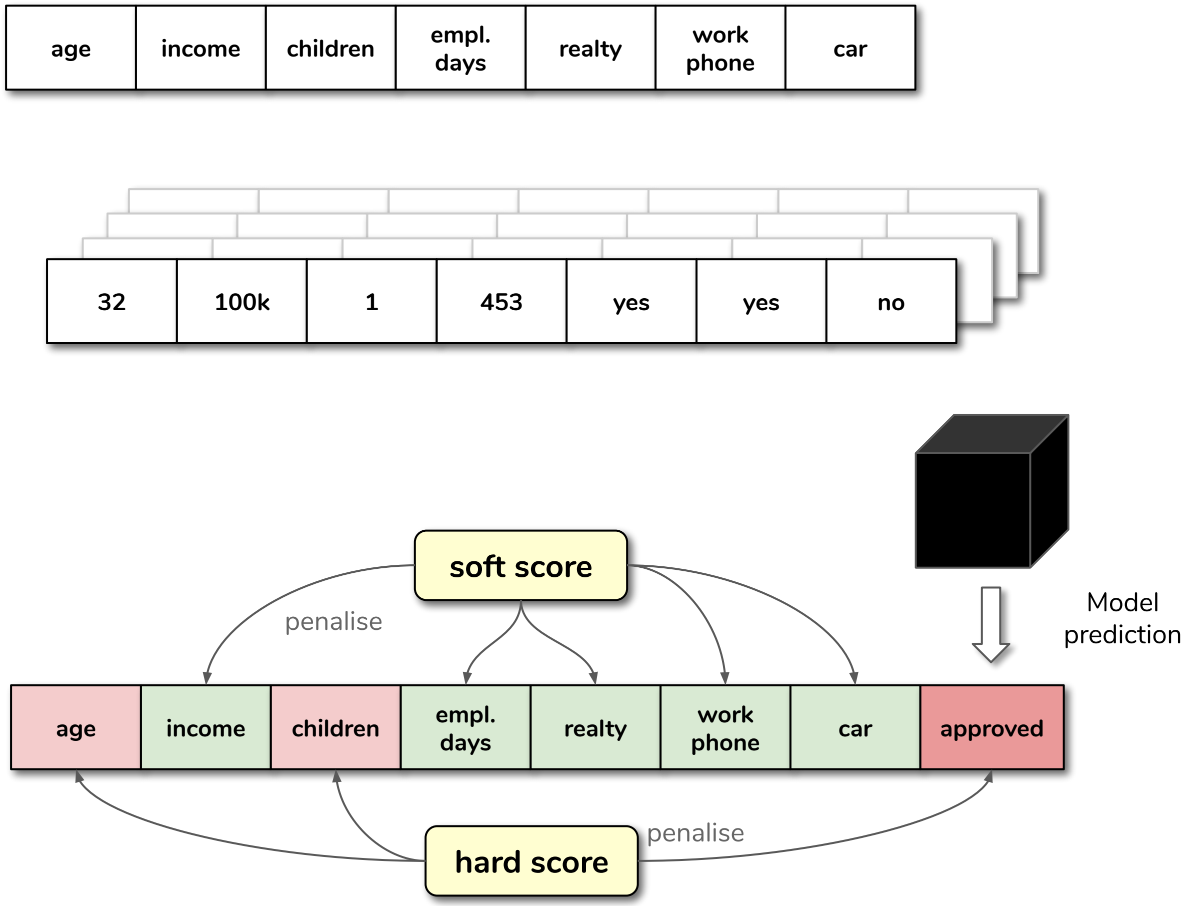
Searching for counterfactuals

Construction heuristics

search space

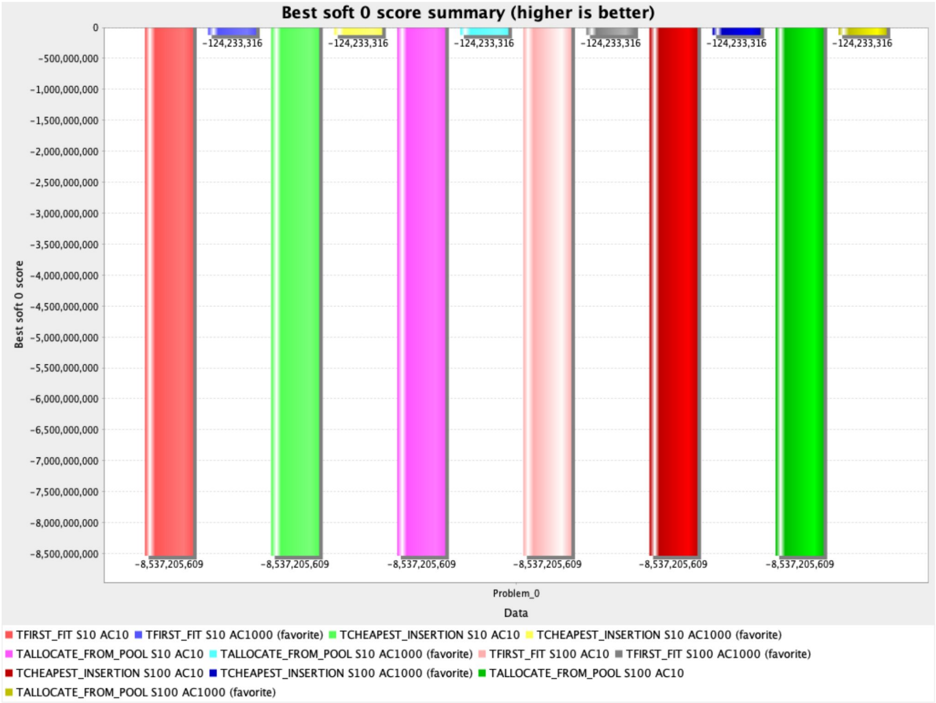
Local search

scoring



Red Hat

Benchmarking



Solver	Total	Average	Standard Deviation	Problem	
				Problem_0	Problem_0
TFIRST_FIT S10 AC10 6 !	[0/-1]hard /[-8537205609.0]soft	[0/-1]hard /[-8537205609.0]soft	0.0E0/0.0E0 /0.0E0	[0/-1]hard /[-8537205609.0]soft 0 !	
TFIRST_FIT S10 AC1000 0	[0/0]hard /[-124233316.0]soft	[0/0]hard /[-124233316.0]soft	0.0E0/0.0E0 /0.0E0		[0/0]hard /[-124233316.0]soft 0
TCHEAPEST_INSERTION S10 AC10 6 !	[0/-1]hard /[-8537205609.0]soft	[0/-1]hard /[-8537205609.0]soft	0.0E0/0.0E0 /0.0E0		
TCHEAPEST_INSERTION S10 AC1000 0	[0/0]hard /[-124233316.0]soft	[0/0]hard /[-124233316.0]soft	0.0E0/0.0E0 /0.0E0		
TALLOCATE_FROM_POOL S10 AC10 6 !	[0/-1]hard /[-8537205609.0]soft	[0/-1]hard /[-8537205609.0]soft	0.0E0/0.0E0 /0.0E0		
TALLOCATE_FROM_POOL S10 AC1000 0	[0/0]hard /[-124233316.0]soft	[0/0]hard /[-124233316.0]soft	0.0E0/0.0E0 /0.0E0		
TFIRST_FIT S100 AC10 6 !	[0/-1]hard /[-8537205609.0]soft	[0/-1]hard /[-8537205609.0]soft	0.0E0/0.0E0 /0.0E0		
TFIRST_FIT S100 AC1000 0	[0/0]hard /[-124233316.0]soft	[0/0]hard /[-124233316.0]soft	0.0E0/0.0E0 /0.0E0		
TCHEAPEST_INSERTION S100 AC10 6 !	[0/-1]hard /[-8537205609.0]soft	[0/-1]hard /[-8537205609.0]soft	0.0E0/0.0E0 /0.0E0		

Takeaways

- **Kogito** makes **Business Automation** working well in cloud environment
- **TrustyAI** adds value-added services to Kogito to enable **tracing**, **explainability** and **monitoring**
- **Explainability** is needed to **establish trust** in automated business processes
- **Counterfactual explanations** provide examples to explain how to obtain the **desired result**
- **OptaPlanner** is a really **powerful** and **flexible** constraint solver
- **OptaPlanner** can be used to **score a prediction**
- It is possible to do **optimization** on **top of predictions** to explain/enrich them

Resources

- **Kogito** - http://kogito.kie.org/
- **TrustyAI** introduction: https://blog.kie.org/2020/06/trusty-ai-introduction.html
- **TrustyAI** aspects: https://blog.kie.org/2020/06/trusty-ai-aspects.html
- **Demo code**: https://github.com/kiegroup/trusty-ai-sandbox/tree/master/counterfactual-op
- **Example-Based Explanations**: https://christophm.github.io/interpretable-ml-book/example-based.html

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat