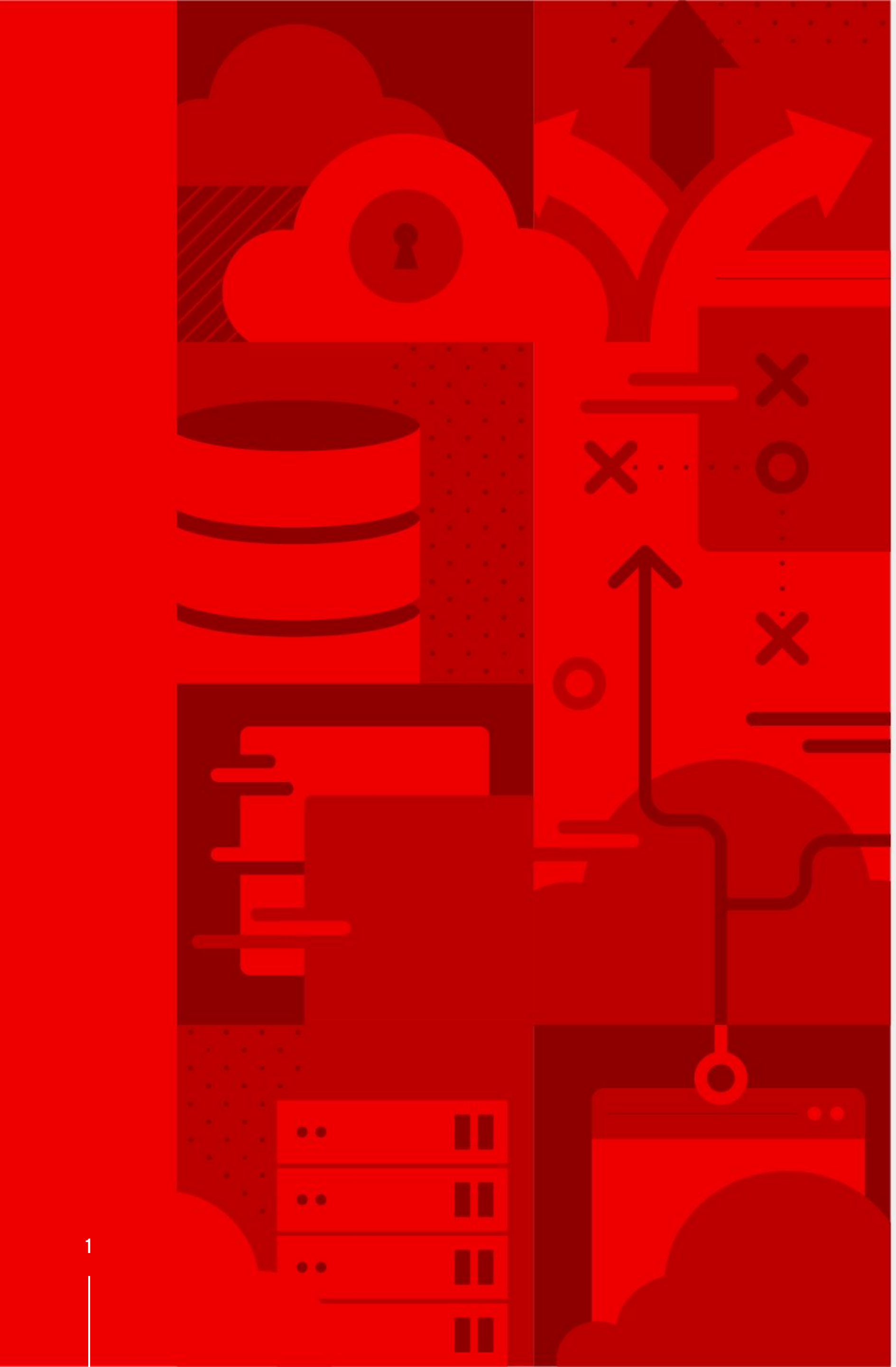


Can you trust your AI?

How TrustyAI toolbox can help you to understand your AI based automation

Daniele Zonca
Architect
TrustyAI

What is Artificial Intelligence?

In computer science, artificial intelligence (AI) is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans (Wikipedia)

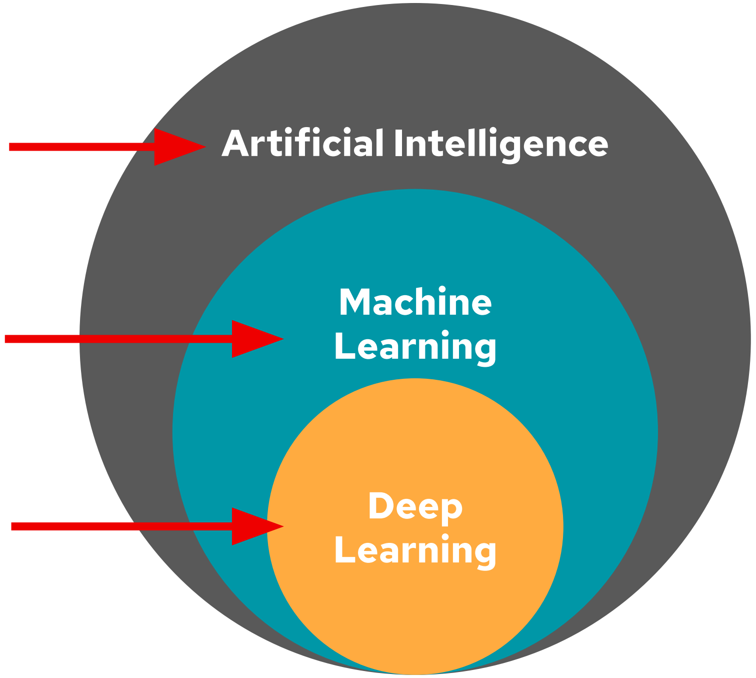
Two main approaches:

- Symbolic: logic/rule based
- Sub-symbolic: statistical learning

Artificial Intelligence
Any technique which enables computers to mimic human behavior.

Machine Learning
Subset of AI which use statistical methods to enable machines to improve with experiences.

Deep Learning
Subset of ML which use multi-layer neural networks.



Prolog (1972) - Symbolic AI

Predicates/Rules:

sibling(X, Y)   :- **parent_child**(Z, X), **parent_child**(Z, Y).

parent_child(X, Y) :- **father_child**(X, Y).

parent_child(X, Y) :- **mother_child**(X, Y).

Facts:

mother_child(*trude*, *sally*).

father_child(*tom*, *sally*).

father_child(*tom*, *erica*).

father_child(*mike*, *tom*).

Query

?- **sibling**(*sally*, *erica*).

<u>Yes</u>

Drools

Rules:

```
rule "validate holiday"
when
   $h1 : Month( name == "july" )
then
   drools.insert(new HolidayNotification($h1));
end
```

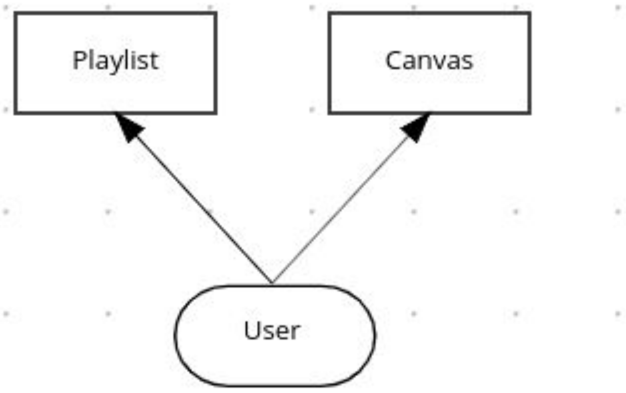
Facts:

```
drools.insert(new Month("july"))
drools.insert(new Month("may"))
```

Query

```
query "checkHolidayNotification" (String monthName)
   holiday := HolidayNotification(month.name == monthName )
end
```

DMN



Canvas *(Decision Table)*

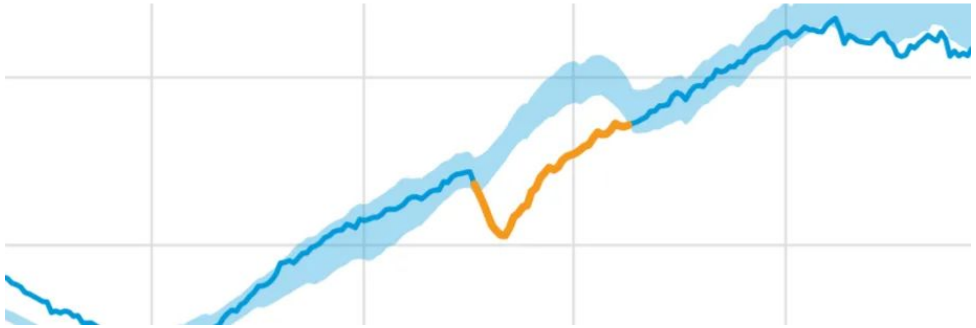
F	User *(string)*	Canvas *(string)*	
1	["bradPitt", "angelina	"Picasso"	
2	"nicolasCage"	"Monet"	
3	-	"Alarm"	

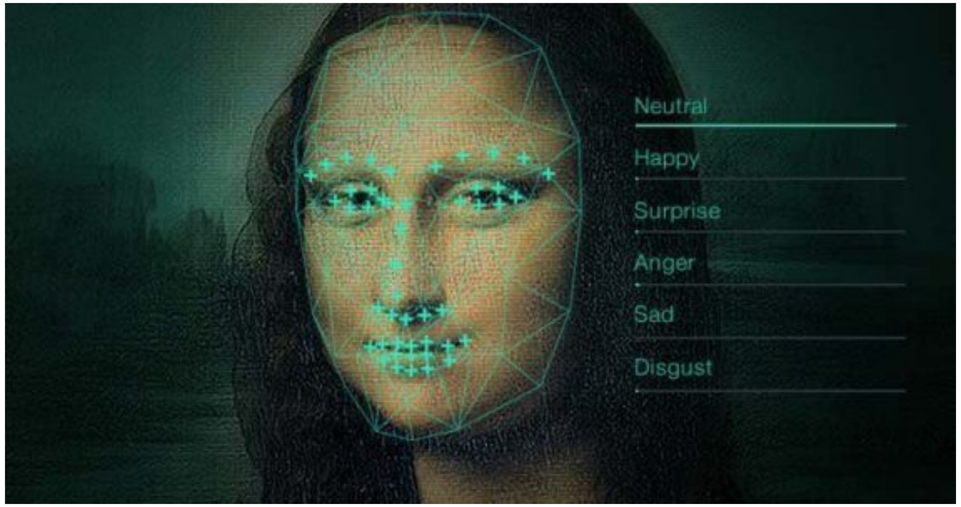
Playlist *(Decision Table)*

F	User *(string)*	Playlist *(string)*	Description
1	"bradPitt"	"Tina Turner"	
2	"angelinaJolie"	"Frank Sinatra"	
3	"nicolasCage"	"Enrico Caruso"	
4	-	"Alarm"	

Is this enough to cover all use cases?

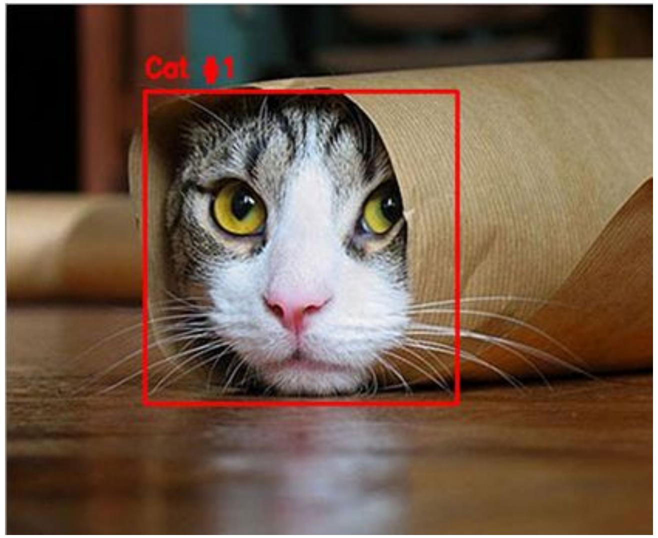
- Image recognition
- Speech recognition
- Anomaly detection



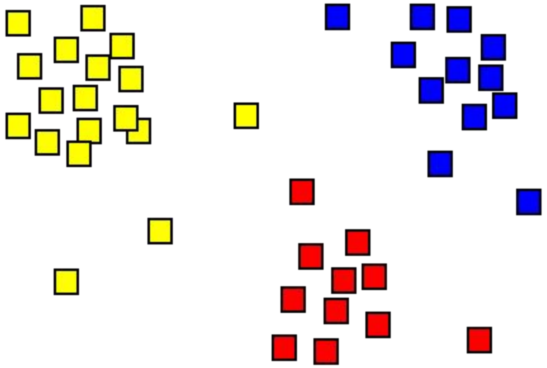


Many different ML algorithms

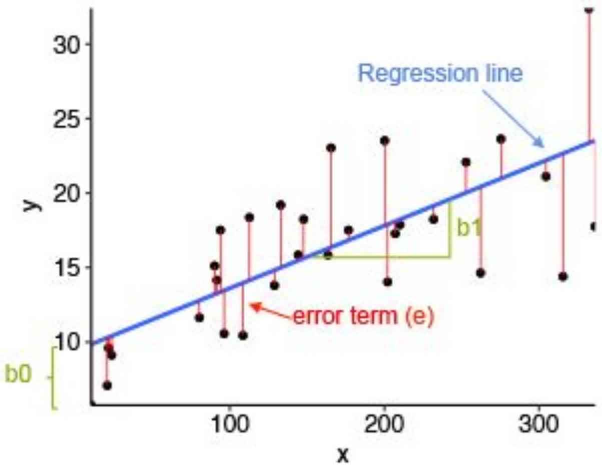
Neural Network



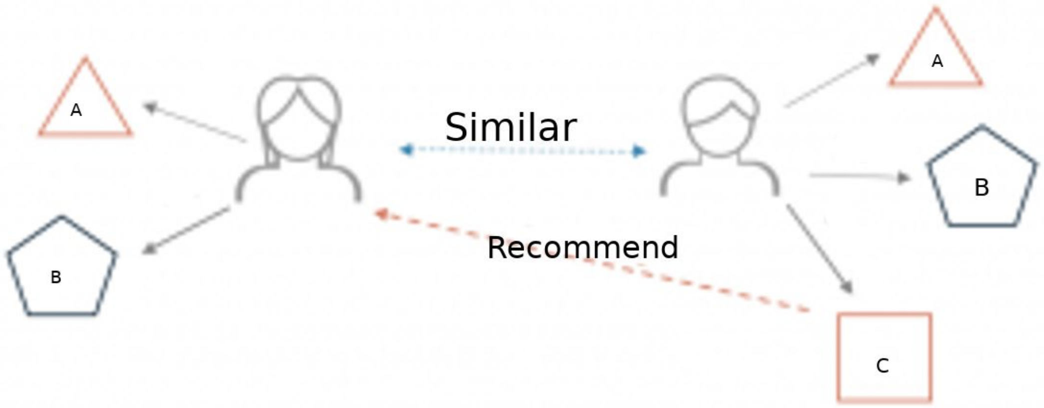
Clustering



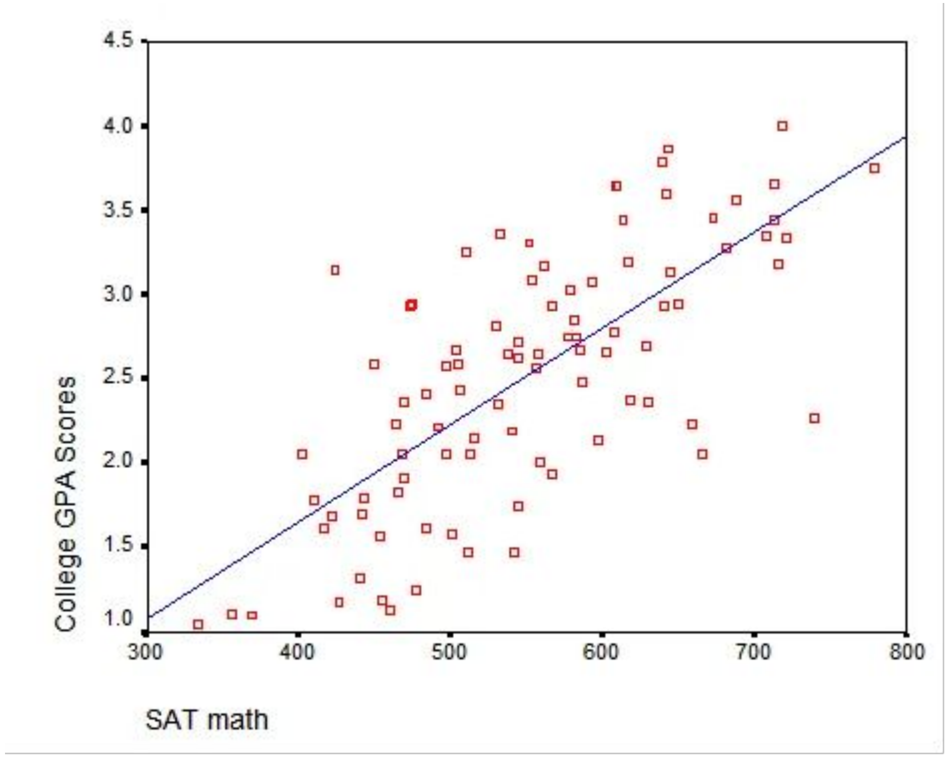
Linear Regression

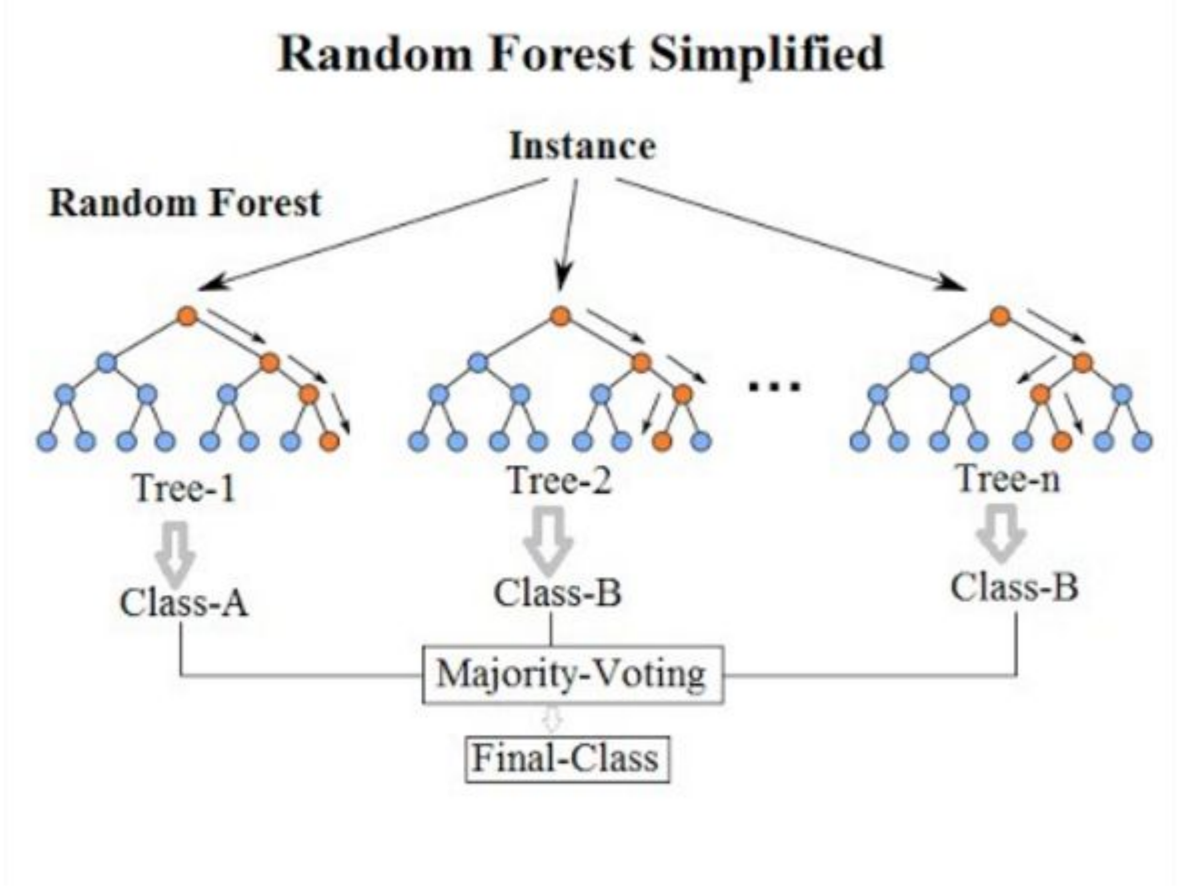


Learn from data

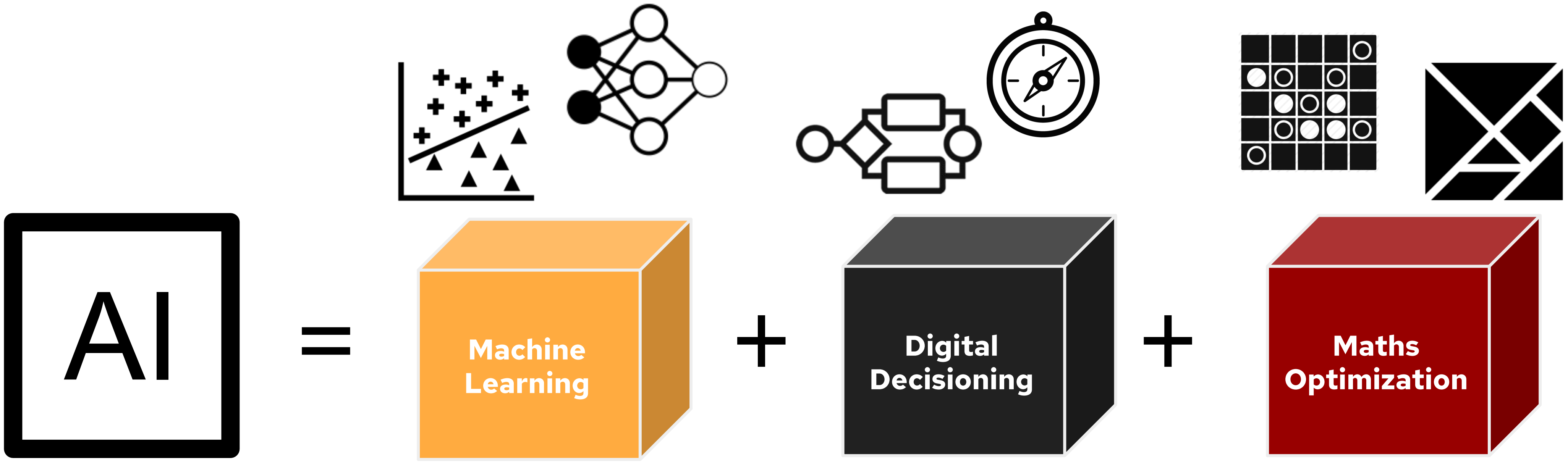


Handle noisy data





Pragmatic Approach to Predictive Decision Automation



Extract information from data analysis

+

Model the human knowledge and expertise

+

Solve complex problems to better resources allocations

Red Hat

Ref: Forrester Research, Inc., “The Future of Enterprise AI and Digital Decisions”, BRAIN 2019, Bolzano Rules and Artificial Intelligence Summit Sep 2019

From Business Automation To Machine Learning

Business Automation

CMMN (2014)

BPMN2 (2011)

DMN (2015)

PMML (1999)

Machine Learning

Done!
Thank you

Well... not really

47,525 views | Jul 1, 2015, 01:42pm

Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software

Maggie Zhang Forbes Staff
Tech
I write about technology, innovation, and startups.

This article is more than 2 years old.

TOM SIMONITE BUSINESS 01.11.2018 07:00 AM

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By James Vincent | Jan 12, 2018, 10:35am EST

SHARE

REUTERS World Business Markets Politics TV

Midterm Elections | Imprisoned In Myanmar | Sectors Up Close | Breakingviews | Investing | Future of Money | Charged: The Future of Aut

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

Amazon scraps secret AI showed bias against wom

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's specialists uncovered a big problem: their

The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.

Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist

THE VERGE TECH SCIENCE C

TECH AMAZON ARTIFICIAL INTELLIGENCE

21

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

AI-powered camera used to replace humans during soccer games confuses referee's bald head with the ball during a game, denying viewers a look at field

- **Scotland's Inverness Caledonian Thistle football club uses AI to record games**
- **The system got confused by a ref's bald head, repeatedly mistaking it for the ball**
- **A commentator apologized to fans for the error**
- **Many smaller teams use AI cameras, as professional crews are too pricey**

By DAN AVERY FOR DAILYMAIL.COM
PUBLISHED: 19:46 GMT, 28 October 2020 | **UPDATED:** 14:10 GMT, 30 October 2020

Share 33k shares 12 View comments

An AI camera at a soccer game in Scotland kept tracking a bald referee instead of the ball during a game.

Inverness Caledonian Thistle played Ayr United on Saturday in a home game at the Caledonian Stadium.

The team doesn't use a cameraman to film games; instead the group relies on an automated camera system to follow the action.

The AI cameras at Caledonian Stadium in Inverness, Scotland, kept mistaking this referee's bald head for the soccer ball. A color commentator for the Inverness Caledonian Thistle apologized for the error


GDPR
GDPR
Data Protection
Officer (DPO)
Compliance
25 May 2018
Data Breaches
Personal Data

Articles 13-15 of the regulation

*"**meaningful information** about the logic involved"*

"the significance and the envisaged consequences"

Article 22 of the regulation

that data subjects have the right not to be subject to such decisions when they'd have the type of impact described above

Recital 71 (part of a non-binding commentary included in the regulation)

States that data subjects are entitled to ***an explanation*** of automated decisions after they are made, in addition to ***being able to challenge*** those decisions.

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

Next-gen Cloud-Native Business Automation

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

Kogito

kubernetes OPENSHIFT

OPERATOR FRAMEWORK

QUARKUS

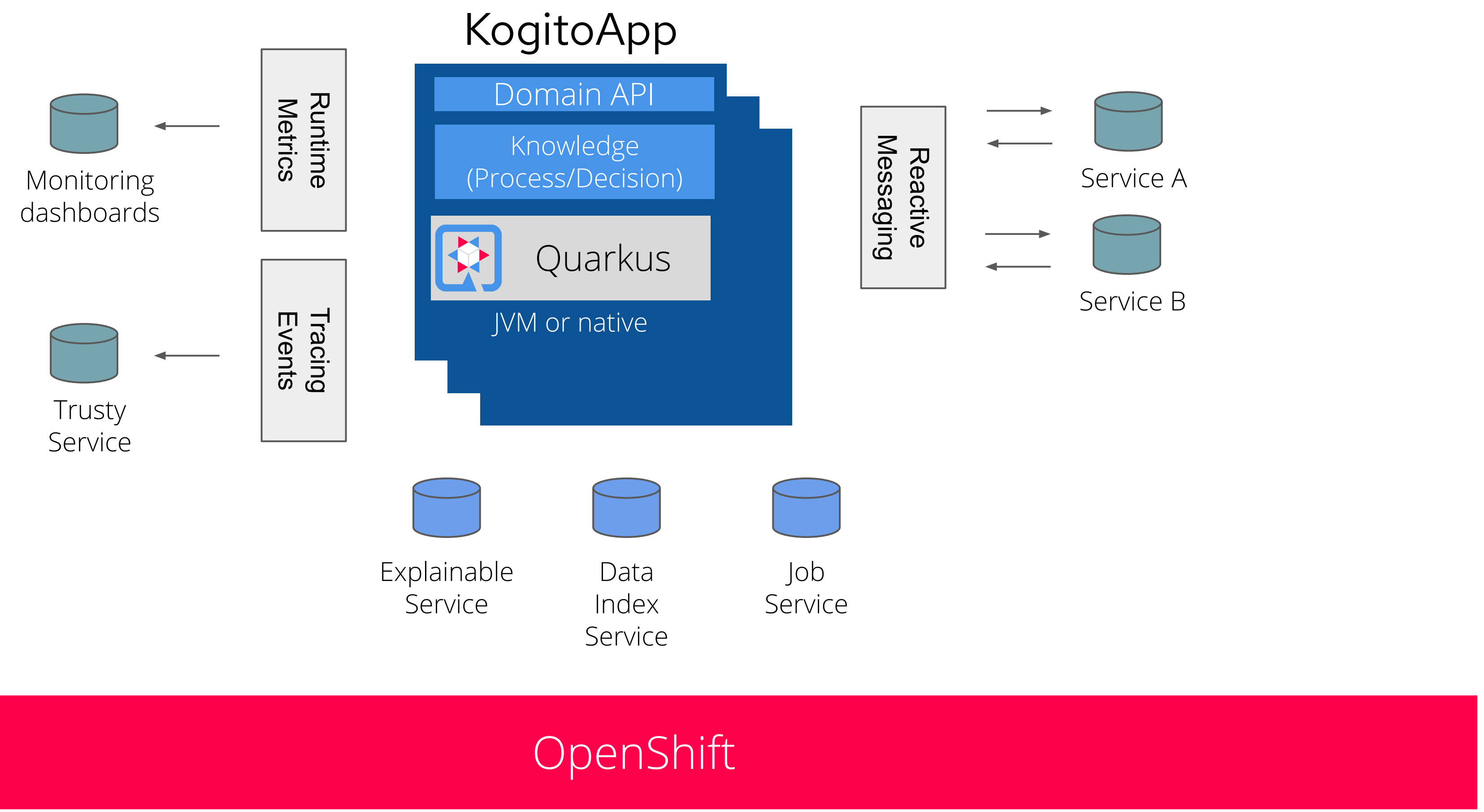
spring boot

Infinispan

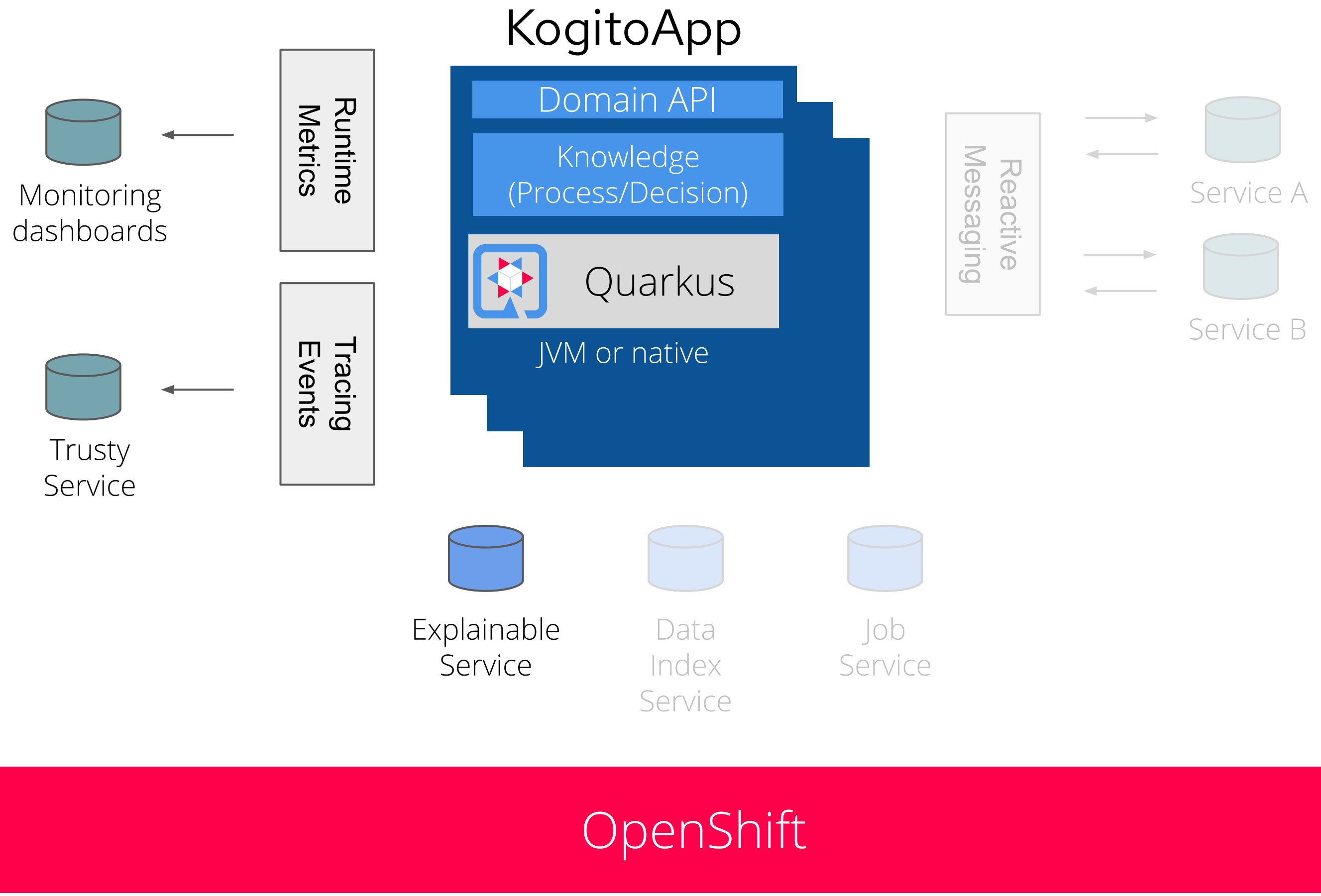
APACHE kafka

Grafana

Runtime Ecosystem



TrustyAI Services

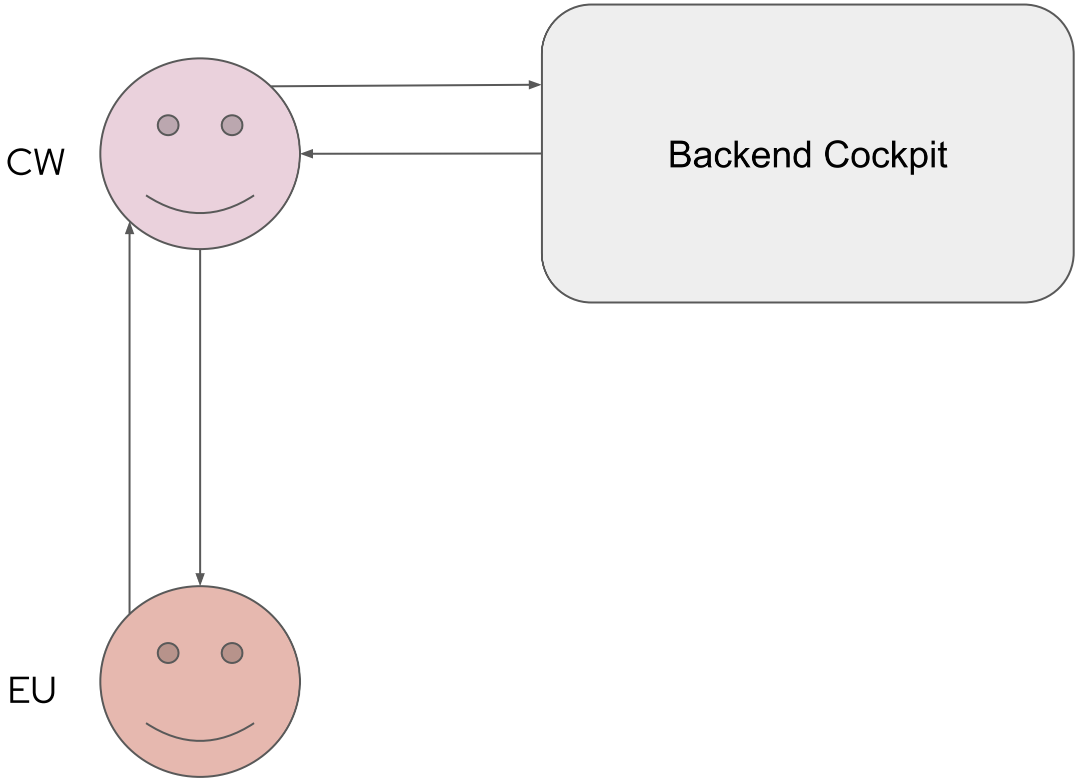


How to empower a Use case with Trusty AI

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."



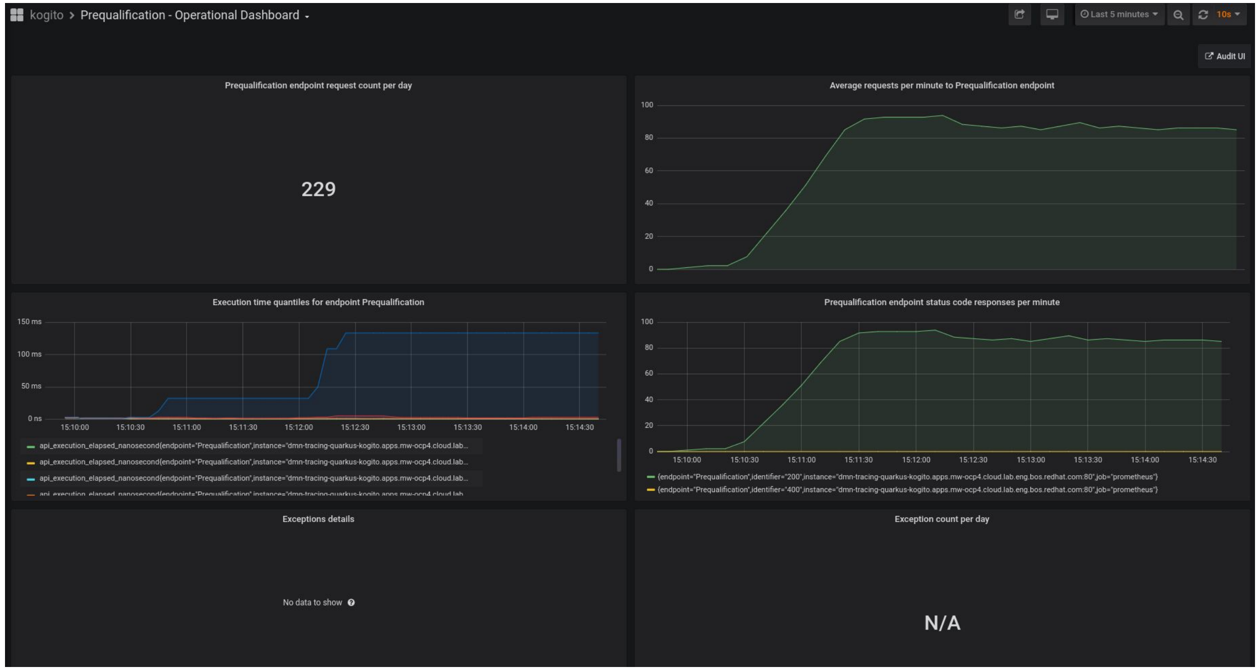
The right tool to the right stakeholder

- **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

Business Monitoring

- Real time business metrics.
- Monitors decision making to ensure it is correct.
- Displays metrics based on model decisions.
- Stakeholders can then monitor the system for business risk and optimization opportunities.

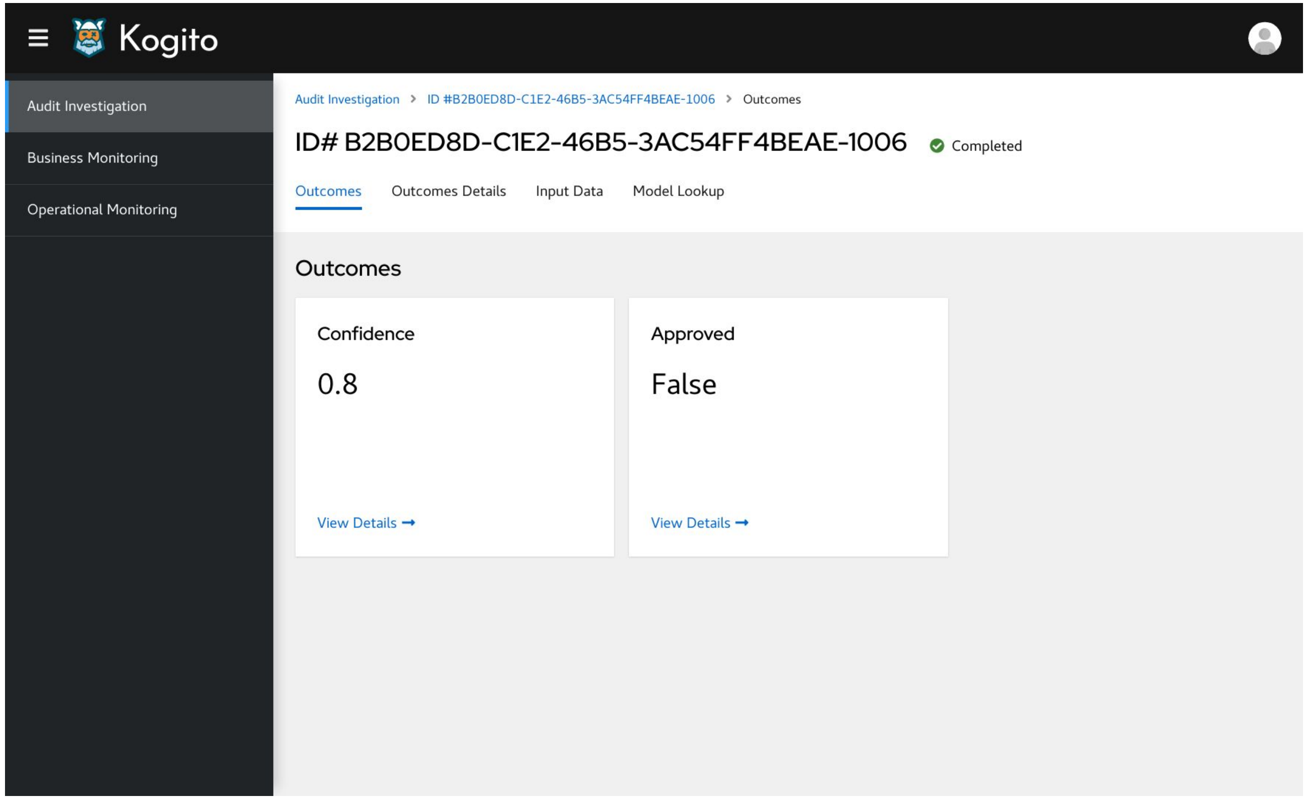
Operational Monitoring



- Real time monitoring service for operational metrics.
- Provides execution monitoring for the decisions.
- Devops engineers can check for correct deployment and system health.

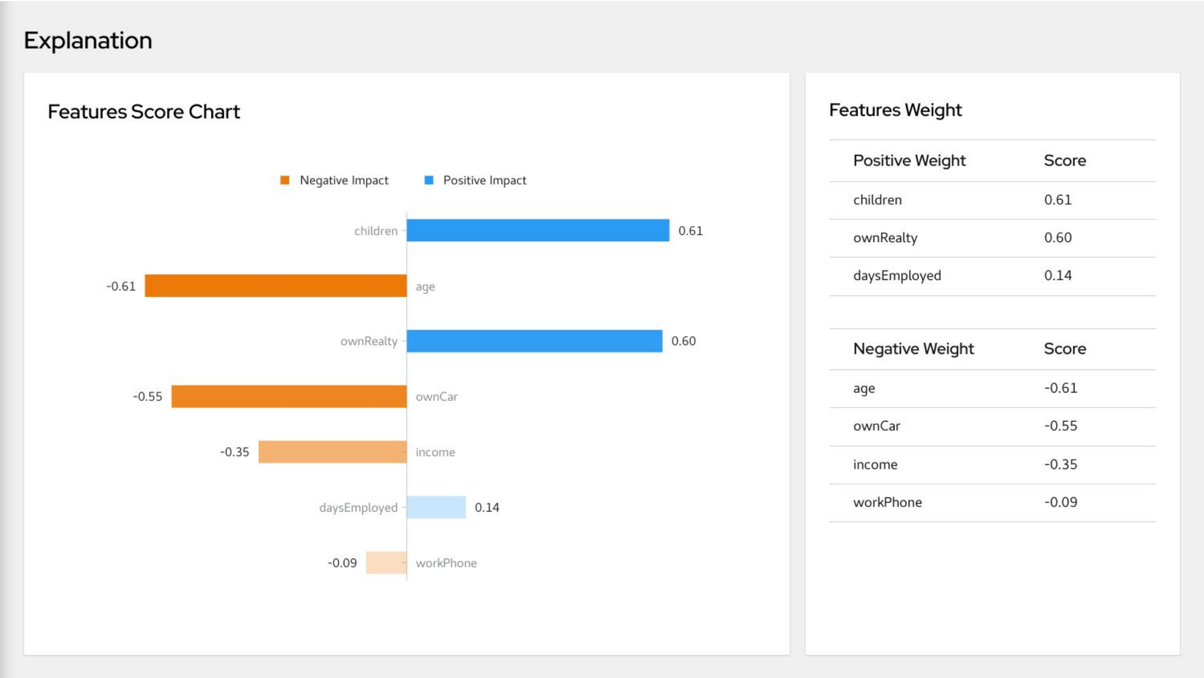
Red Hat

Audit UI



- Trace decision execution
- Provides ability to query historic decisions
- Introspection of each individual decision made within the system
- Details of decision outcomes
- Provides model metadata for auditing purposes

Audit UI



- Explainability is shown for each of the decisions
- Being able to say *why* a decision was made helps with the accountability of the system

Red Hat

My 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

Types of explanations

- **Local** vs **global**
  - local explanation is used for describing the behaviour of a single prediction while a global explanation is used for describing the behaviour of the entire model
- **Directly interpretable** vs **post-hoc**
  - when an explanation is understandable by most consumers whereas a post-hoc explanation is one that involves an auxiliary method to explain a model after it has been trained
- **Surrogate**
  - involves a second, usually directly interpretable, model that approximates a more complex (and less interpretable) one
- **Static** vs **interactive**
  - a static explanation doesn’t change while interactive explanations allow consumers to drill down or ask for different types of explanations

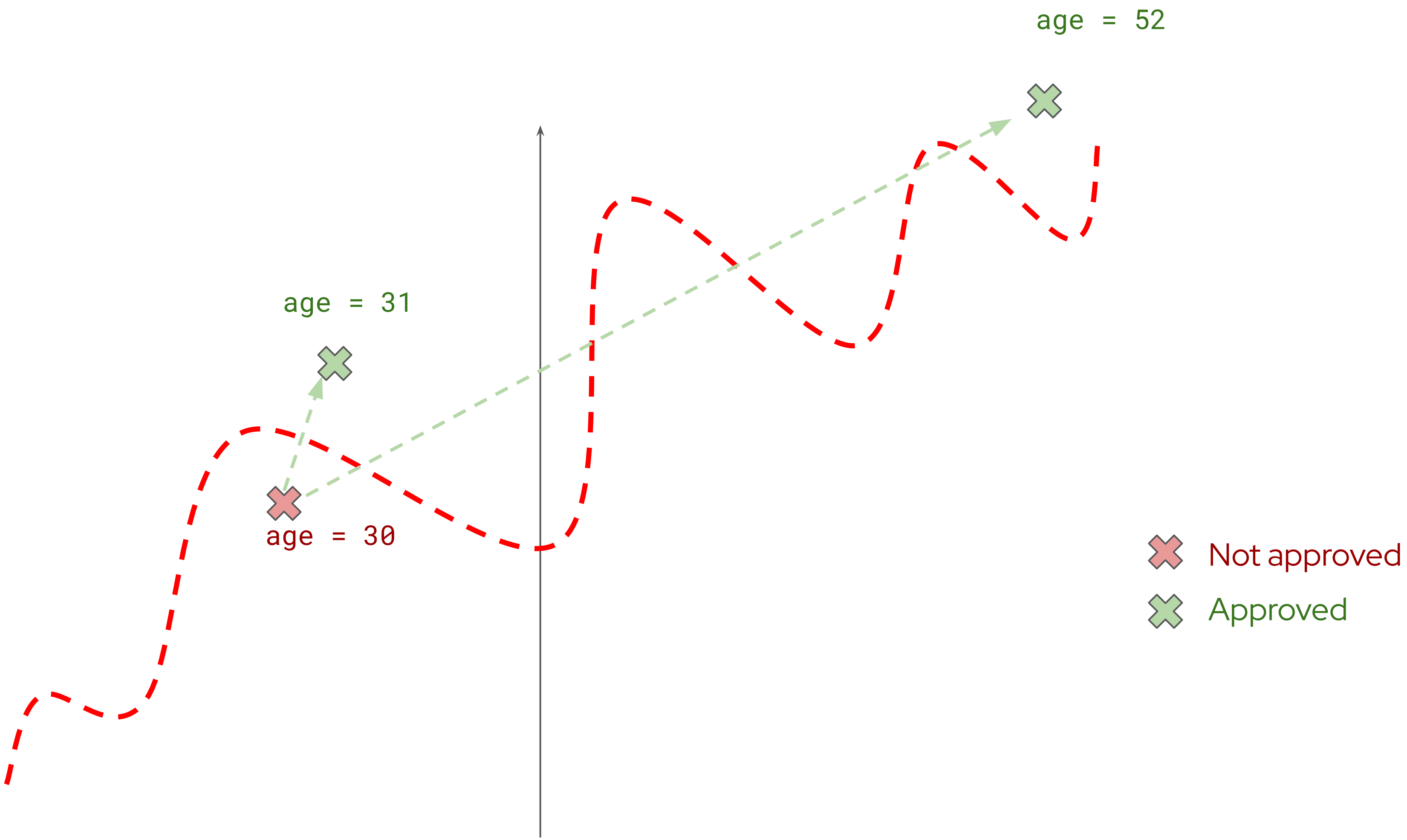
LIME

- 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

Counterfactual explanations

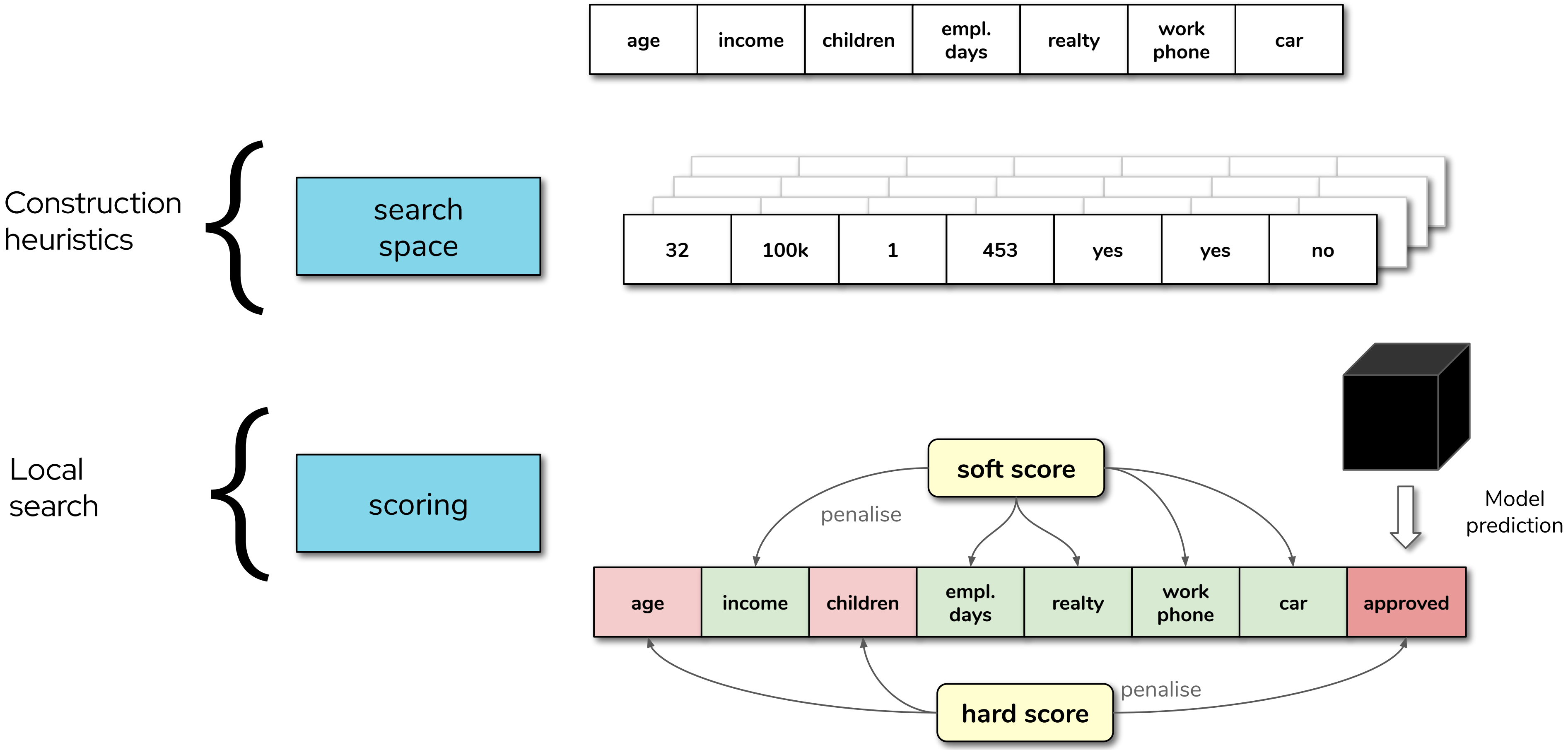
- **Exemplar** explanations provide explanations for single predictions by means of **examples** (in the input space)
  - **Counterfactual explanations** provide examples that
    - Have a *desired* prediction, according to the black box model
    - Are as *close* as possible to the original input
 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

Domain search space



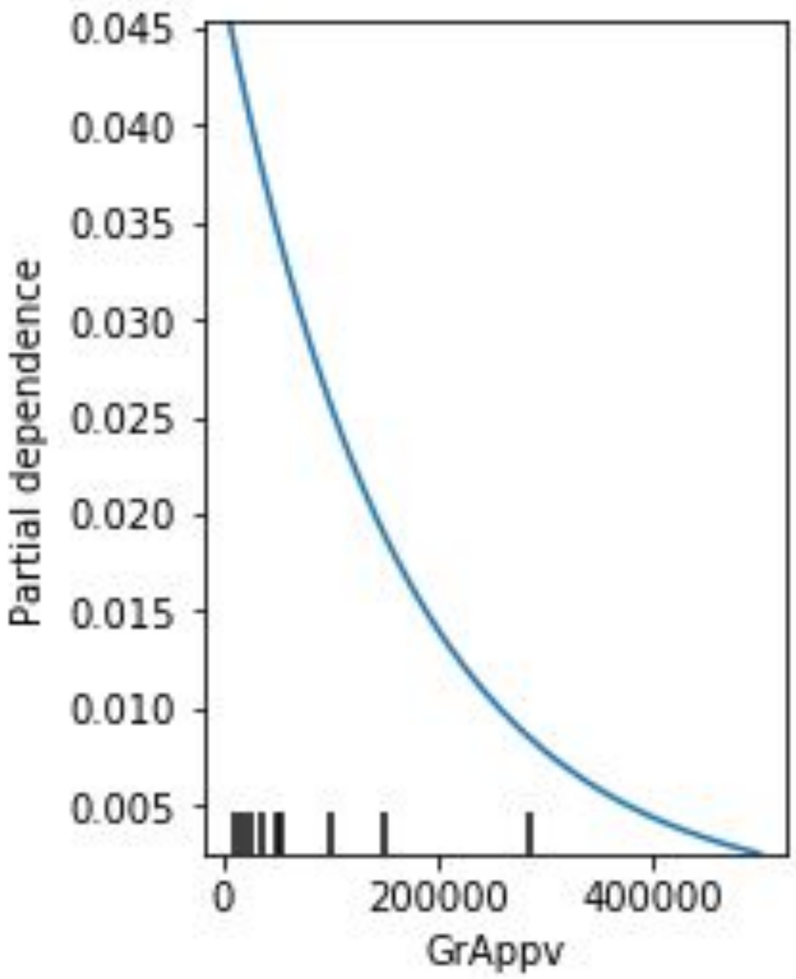
Red Hat

Searching for counterfactuals



Feature relevance methods - PDPs

- Observe how the changes in a certain feature influences the prediction, on average, when all other features are left fixed
- Visualization based explanation



TrustyAI - Explainability

- Explanation Library
  - Algorithms and tools to explain black box models
- Explainability ITs
  - Integration tests to check functionalities, performance and stability of explainability algorithms on different types of models
    - DMN
    - PMML
    - OpenNLP Language Detector
- Explainability Service
  - Exposes explainability algorithms as a service
    - Currently connects to the model to explain via a remote endpoint

TrustyAI - Explainability

- Explanation Library provides implementation of
    - LIME
        - Local post-hoc explanation (saliency method)
    - PDP
        - Global post-hoc explanation (feature relevance method)
    - Explainability evaluation metrics
    - Counterfactual explanation (WIP)
    - Aggregated LIME global explanation (WIP)
    - Integration with
        - DMN models
        - PMML models

What's next?

- Fairness analysis, for accountability (e.g. change in code improved model removed geographical bias in predictions)
- Global explanation (SHAP)
- Interpretability analysis, for model selection (e.g. given a task, I want to use the most interpretable one)
- Simplicity analysis, for model selection (e.g. given data and model, does a similar but simpler, and more interpretable, model with comparable performance exist ?)
- End to end accountability (e.g. keep track from requirement definition to the solution in production)

References

TrustyAI introduction: https://bit.ly/35Yfs7M + https://bit.ly/2THWSLA

End to end demo instructions: https://git.io/JT5bI

Sandbox repo: https://github.com/kiegroup/trusty-ai-sandbox

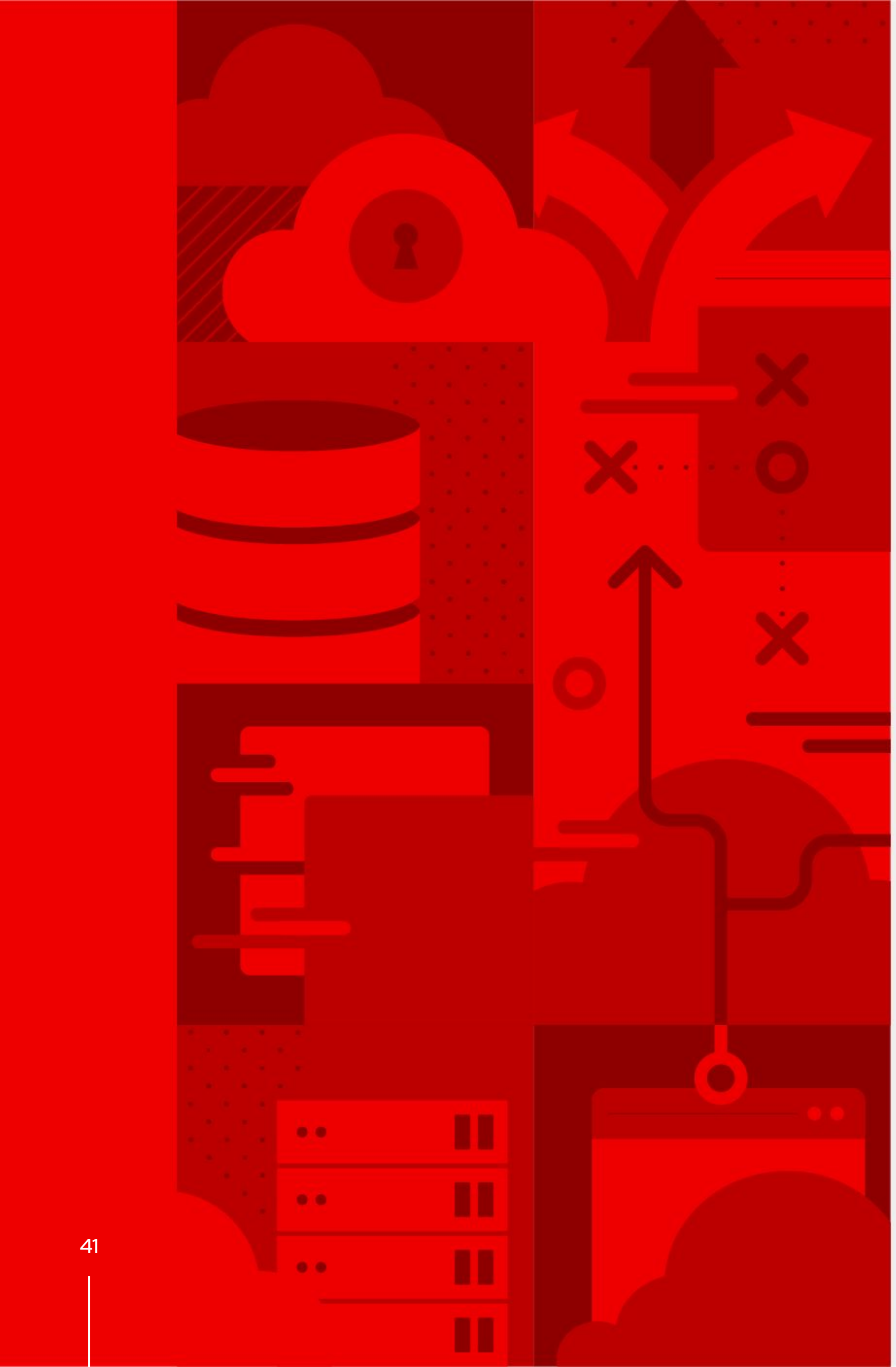
Counterfactual POC: https://youtu.be/4H3U6xyCgMI + https://bit.ly/3mL5Kg0

Blogpost Explainability: https://bit.ly/38aLm3w

Blogpost Monitoring: https://bit.ly/322Mm5W

TrustyAI Zulip chat https://kie.zulipchat.com/#narrow/stream/232681-trusty-ai

Red Hat



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

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat

The right A.I. for the job

One Artificial Intelligence algorithm does not fit all use cases.

Vector Space Model

Full text search

"cat"

⇩

The secret life of felines
felines.pdf
Felines, or **cats** as they are more commonly known, are carnivorous ...

Other use cases include:
recommendations, similarties, ...

Implemented by:

Lucene

Neural Net

Image recognition

⇩

"Dog"

Other use cases include:
voice recognition, machine translation, ...

Implemented by:

TensorFlow, Deeplearning4j

Constraint Solver

Vehicle routing problem

⇩

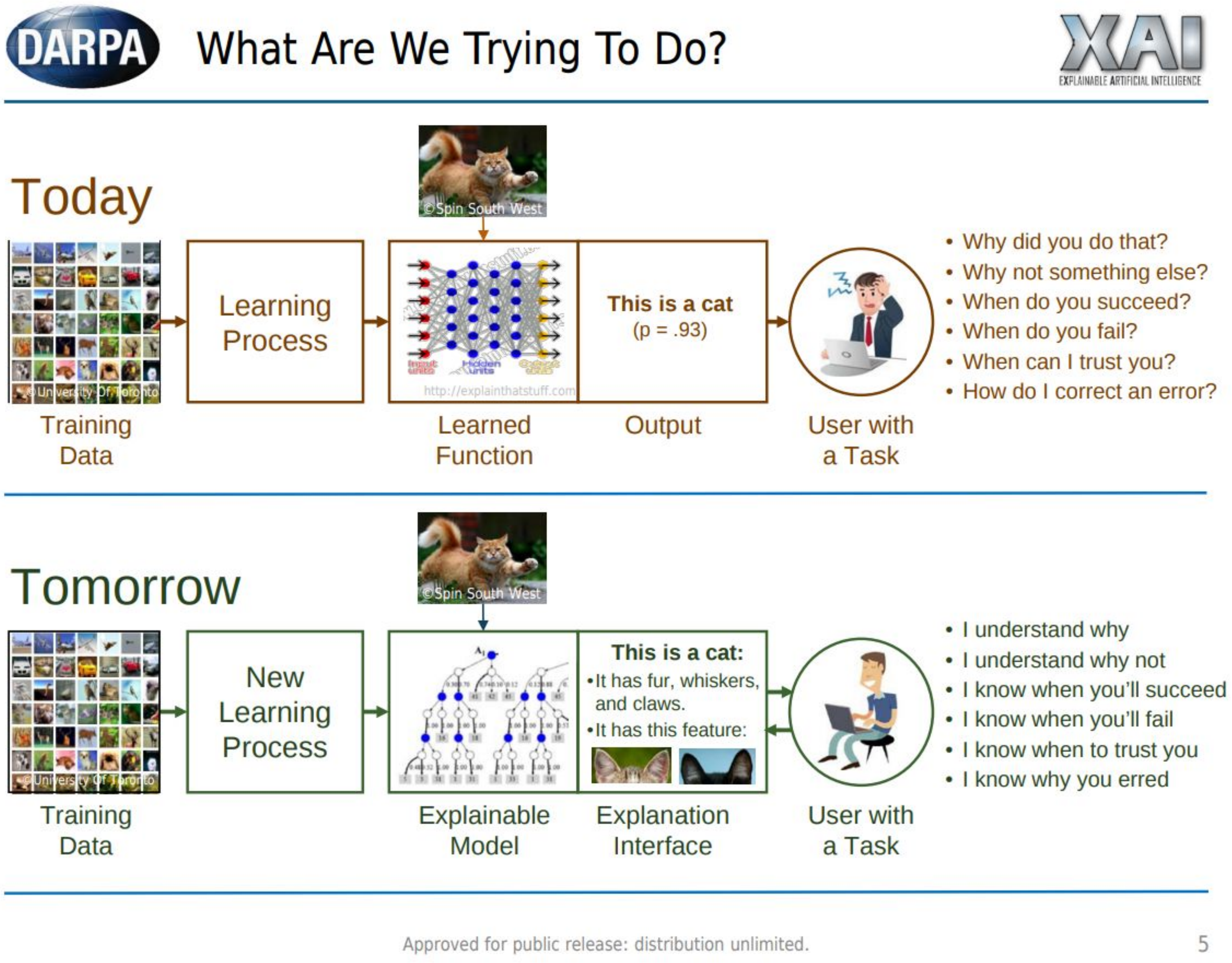
15% less driving time

Other use cases include:
employee rostering, job scheduling, ...

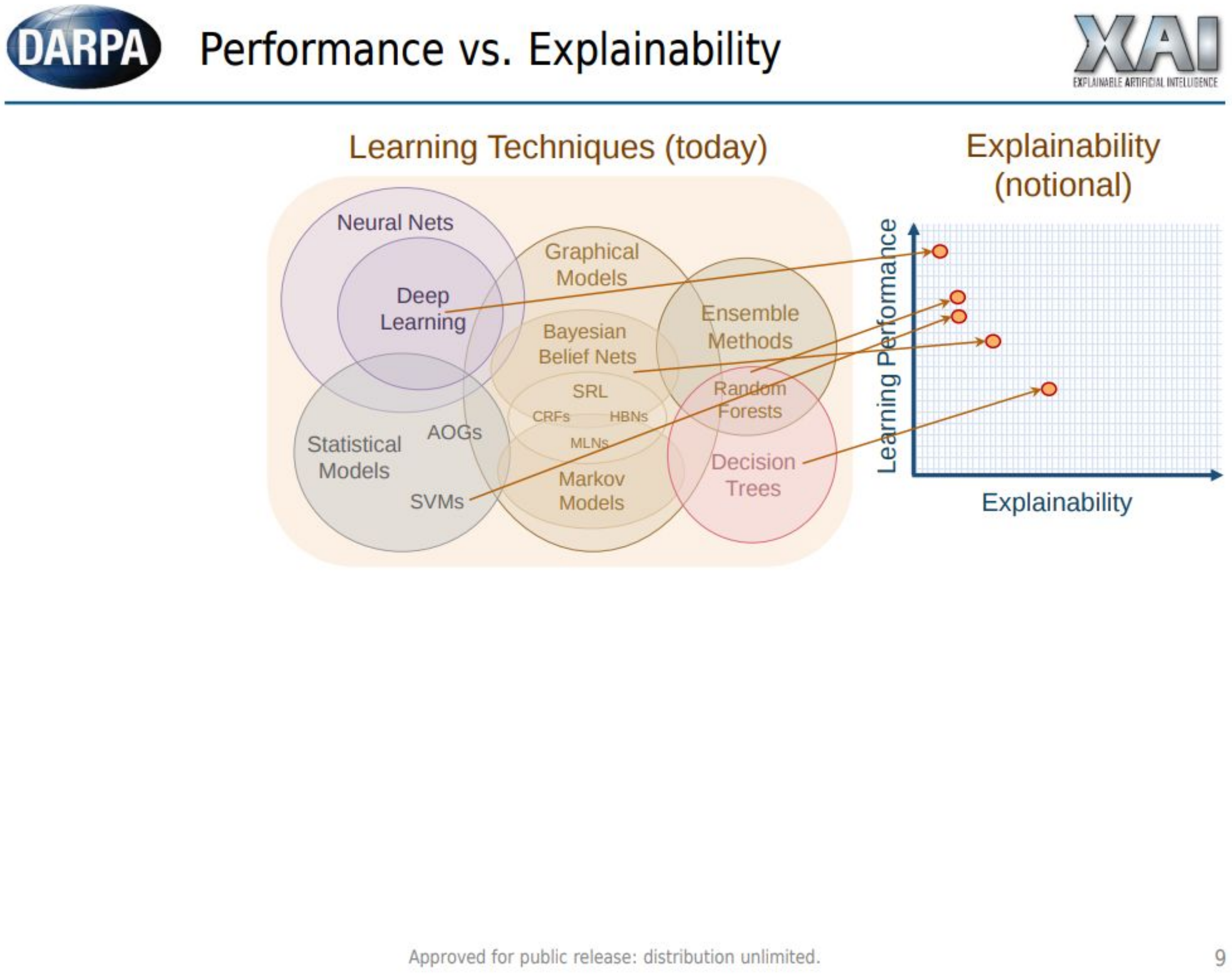
Implemented by:

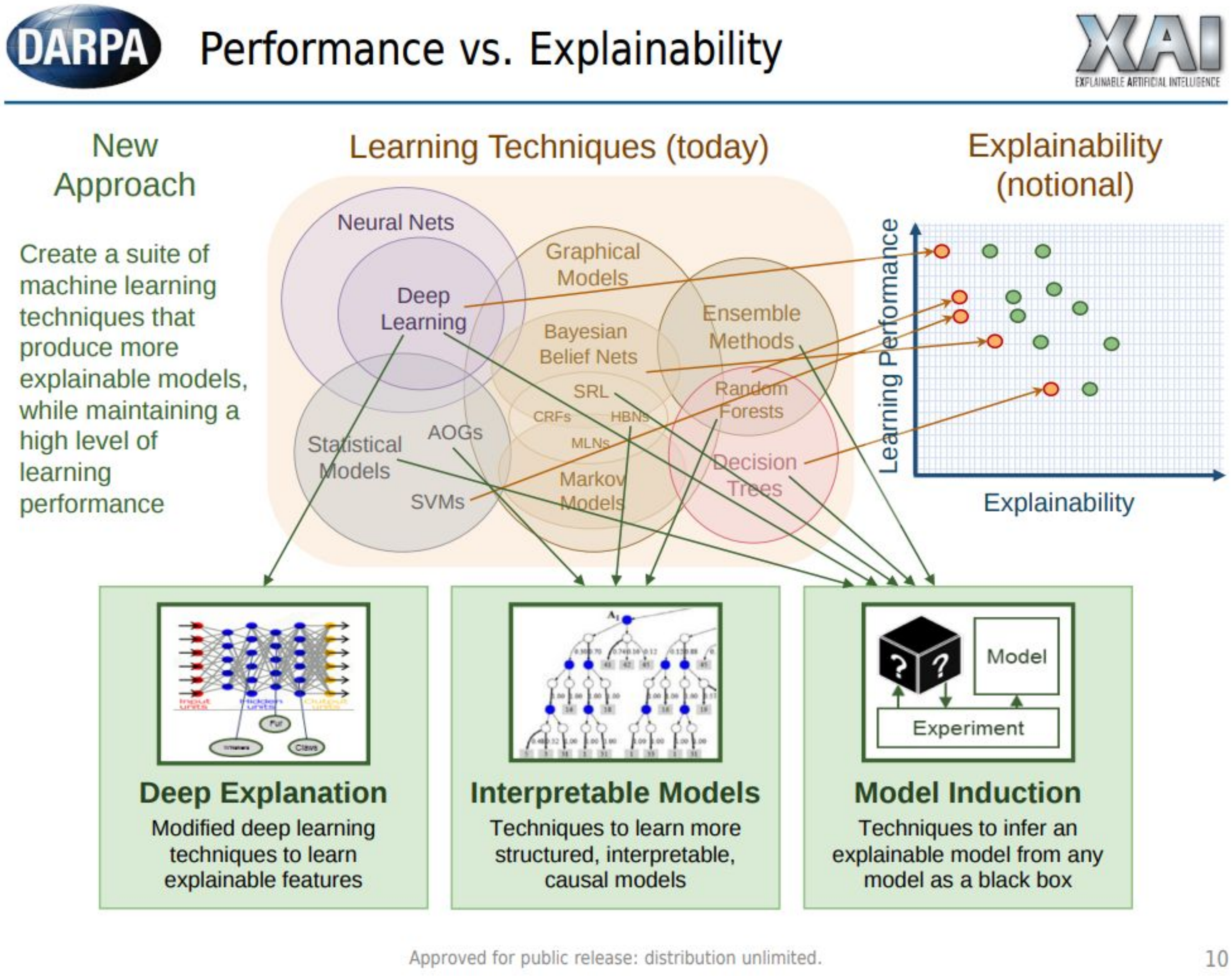
OptaPlanner

Other algorithms for other use cases:
A* Search for pathfinding, Rete/Phreak for production rule systems, k-means for cluster analysis, ...


DARPA
What Are We Trying To Do?
XAI
EXPLAINABLE ARTIFICIAL INTELLIGENCE
Today
©Spin South West
© University Of Toronto
Training Data
Learning Process
Learned Function
http://explainthatstuff.com
This is a cat (p = .93)
Output
User with a Task
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
Tomorrow
©Spin South West
© University Of Toronto
Training Data
New Learning Process
Explainable Model
This is a cat:
•It has fur, whiskers, and claws.
•It has this feature:
Explanation Interface
User with a Task
• I understand why
• I understand why not
• I know when you'll succeed
• I know when you'll fail
• I know when to trust you
• I know why you erred
Approved for public release: distribution unlimited.
5

Red Hat

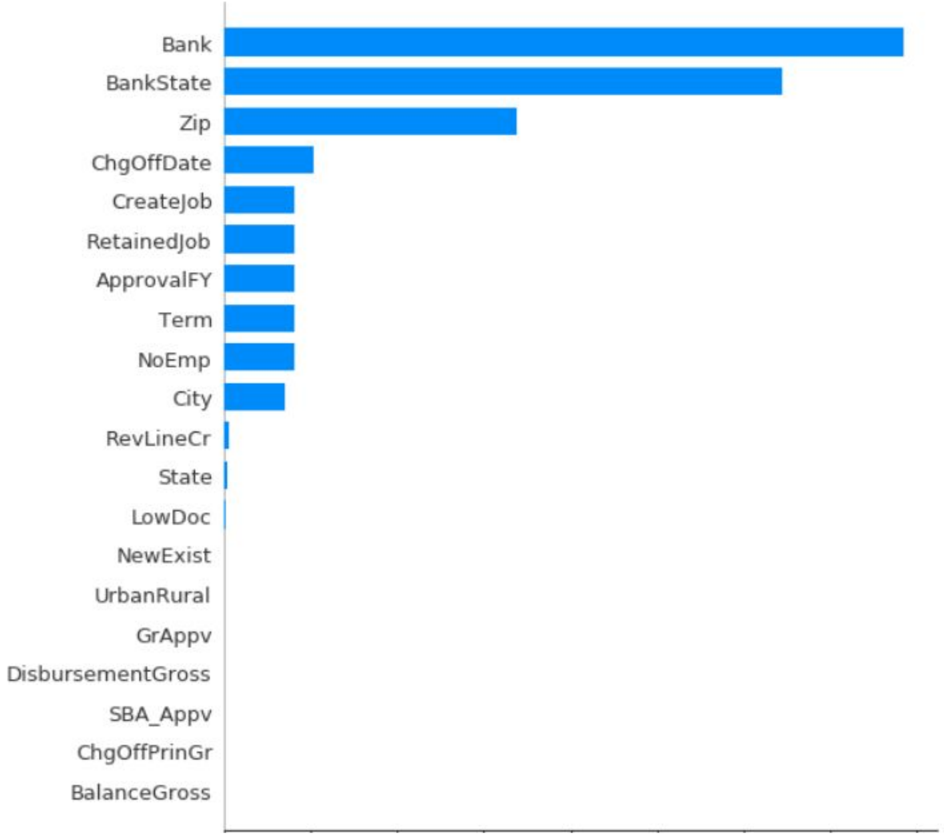

DARPA
Performance vs. Explainability
XAI
EXPLAINABLE ARTIFICIAL INTELLIGENCE
Learning Techniques (today)
Neural Nets
Deep Learning
Graphical Models
Bayesian Belief Nets
SRL
CRFs
HBNs
MLNs
Markov Models
Ensemble Methods
Random Forests
Decision Trees
Statistical Models
AOGs
SVMs
Explainability (notional)
Learning Performance
Explainability
Approved for public release: distribution unlimited.
9


DARPA
Performance vs. Explainability
XAI
EXPLAINABLE ARTIFICIAL INTELLIGENCE
New Approach
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance
Learning Techniques (today)
Neural Nets
Deep Learning
Graphical Models
Bayesian Belief Nets
Ensemble Methods
Random Forests
Decision Trees
SRL
CRFs
HBNs
MLNs
Markov Models
Statistical Models
AOGs
SVMs
Explainability (notional)
Learning Performance
Explainability
Deep Explanation
Modified deep learning techniques to learn explainable features
Interpretable Models
Techniques to learn more structured, interpretable, causal models
Model Induction
Techniques to infer an explainable model from any model as a black box
Model
Experiment
Approved for public release: distribution unlimited.
10

Saliency methods - SHAP

- Explains the prediction of an instance by computing the contribution of each feature to the prediction
- Computes Shapley values for each feature (the average marginal contribution of a feature value across all possible coalitions)
- Additive feature attribution method

```
shap.summary_plot(shap_values, test_exp, plot_type="bar")
```



Red Hat