QuantUniversity, LLC

www.quantuniversity.com

A Master Class in Building Production-Grade NLP Pipelines

Presented By:

Sri Krishnamurthy, CFA, CAP

sri@quantuniversity.com

www.quantuniversity.com

10/15/2019

MathWorks Conference

New York, NY

Speaker bio

QuantUniversity, LLC
www.quantuniversity.com

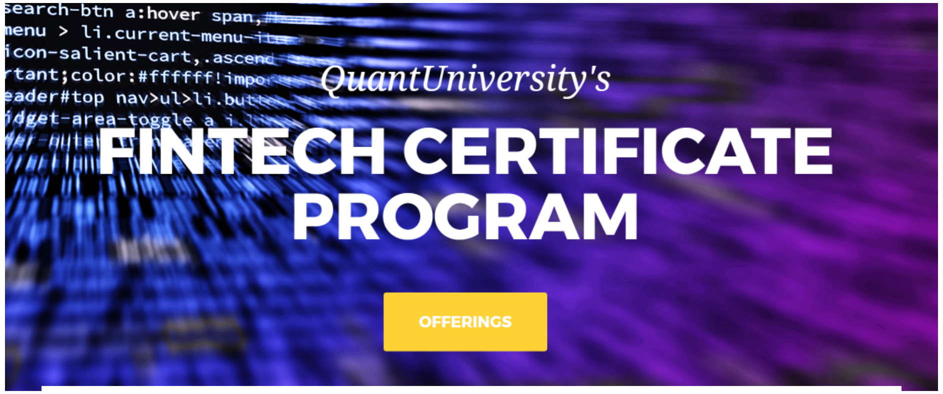
Sri Krishnamurthy
Founder and CEO
QuantUniversity

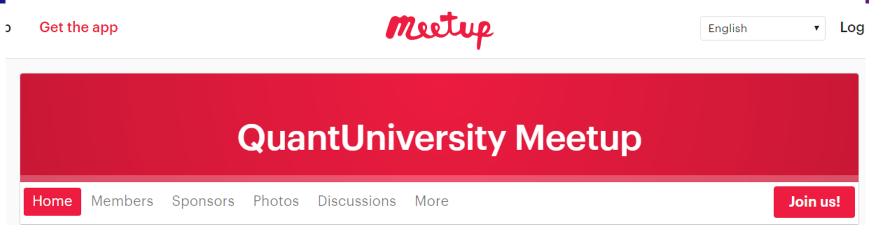
BABSON MBA

- Quant, Data Science & ML practitioner
- Prior Experience at MathWorks, Citigroup and Endeca and 25+ financial services and energy customers.
- Columnist for the Wilmott Magazine
- Teaches Data Science/AI at Northeastern University, Boston
- Reviewer: Journal of Asset Management

About QuantUniversity

- Boston-based Data Science, Quant Finance and Machine Learning training and consulting advisory
- Trained more than 1000 students in Quantitative methods, Data Science, ML and Big Data Technologies
- Building QuSandbox a platform for operationalizing AI and Machine Learning in the Enterprise



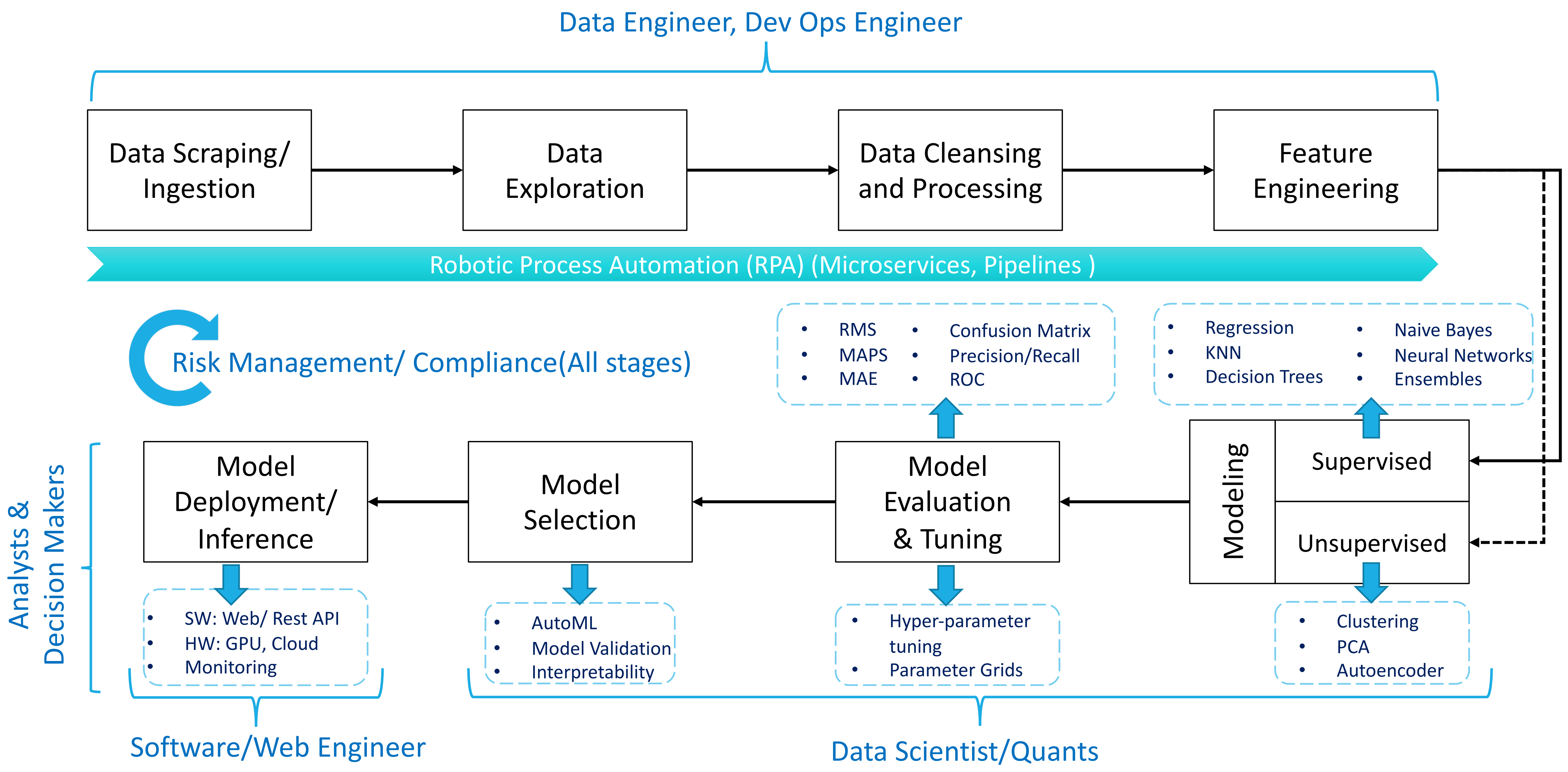


QuantUniversity, LLC
www.quantuniversity.com

Agenda

1. Model Life Cycle Management & Pipelines
2. Productionizing Pipelines: An NLP Case study

Machine Learning Workflow



Challenges

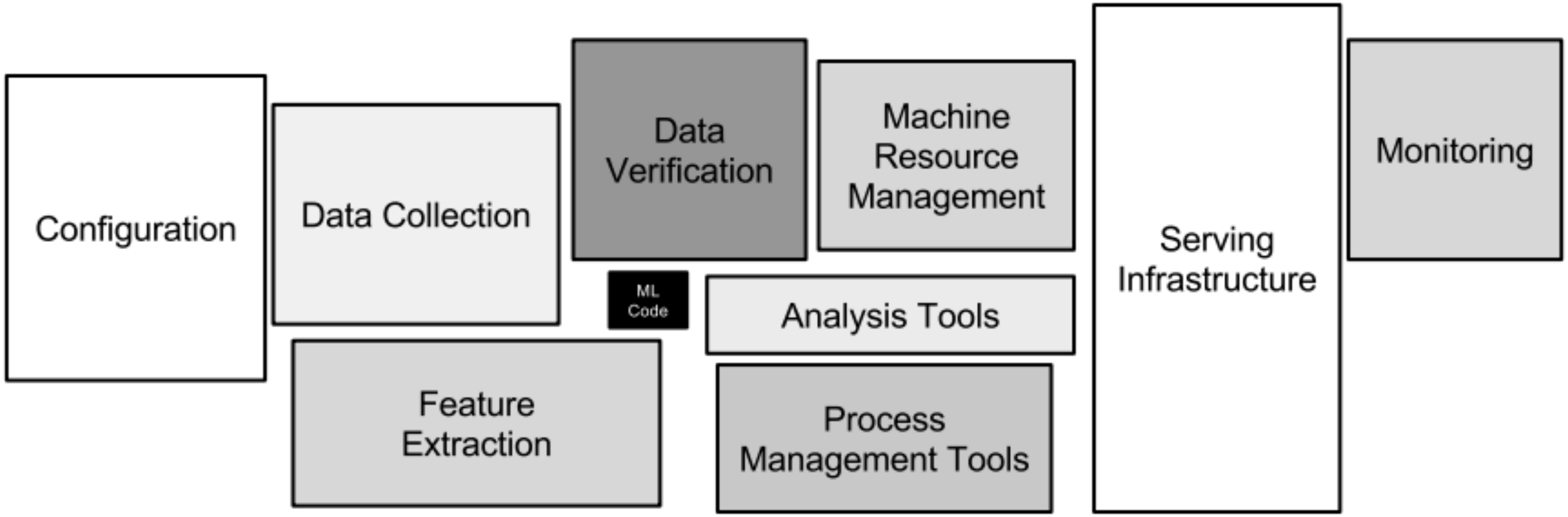
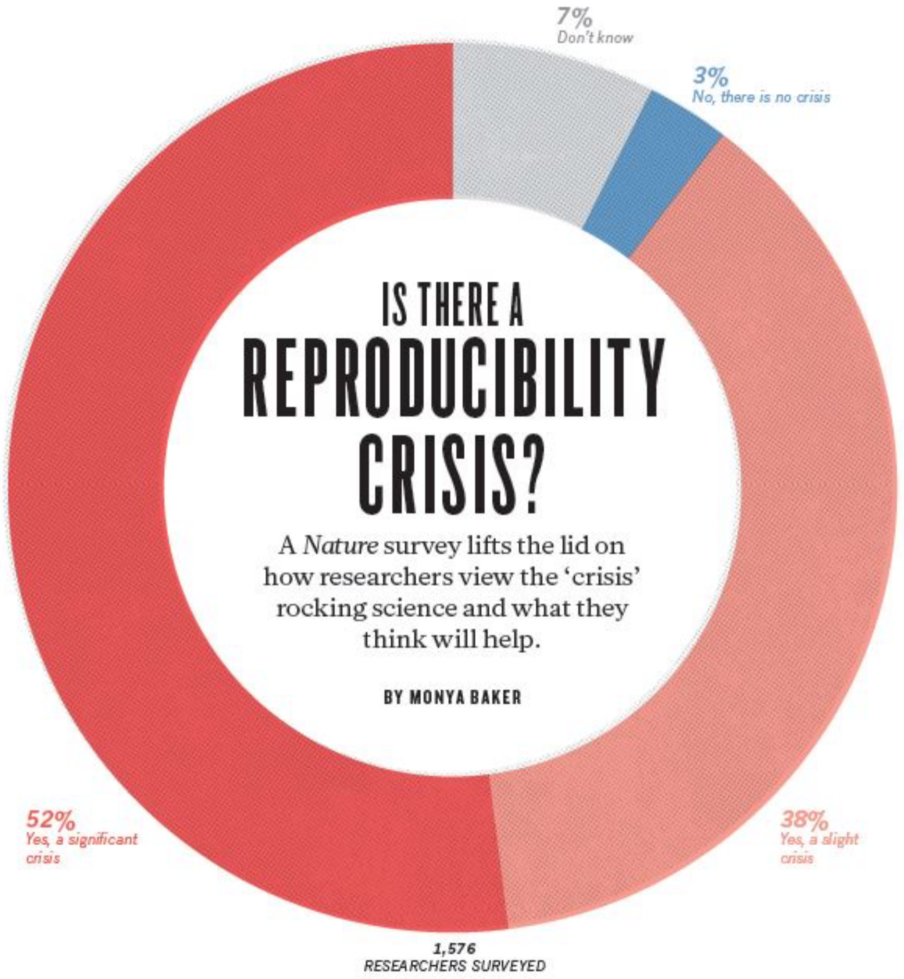
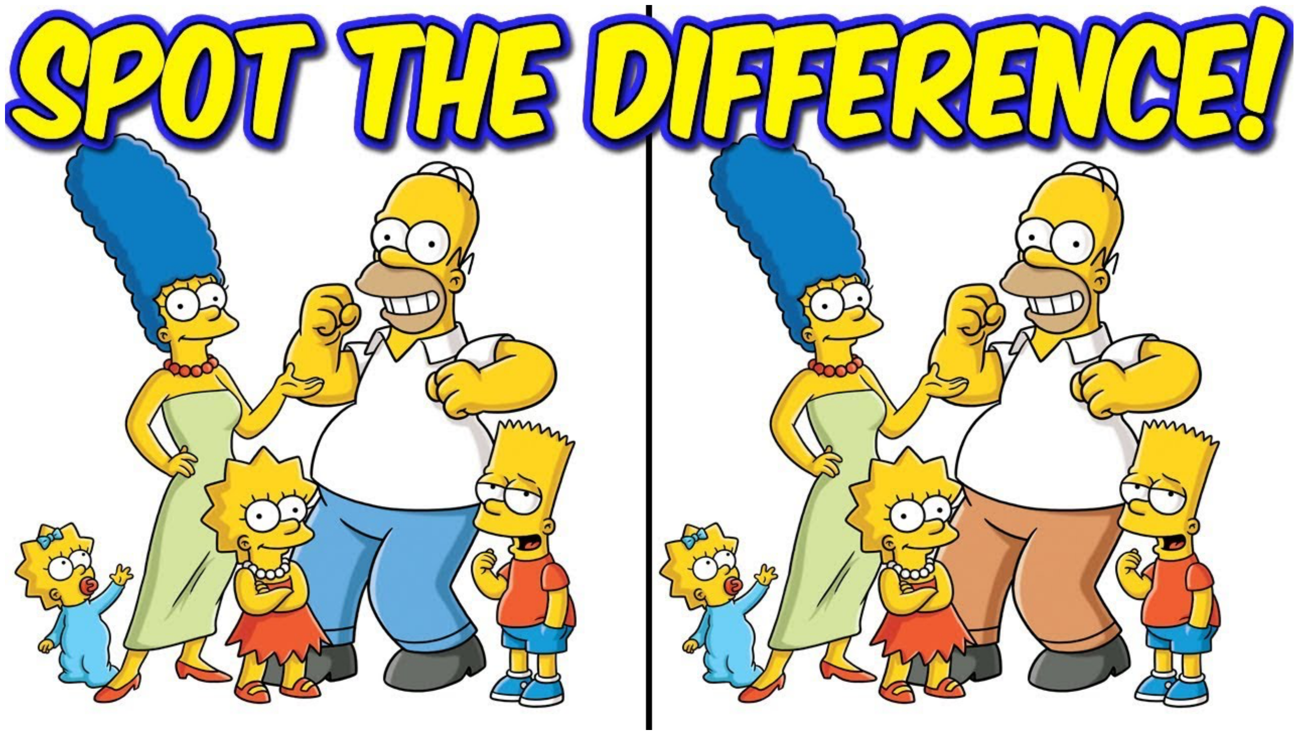


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Source: Sculley et al., 2015 *"Hidden Technical Debt in Machine Learning Systems"*

The reproducibility challenge





https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Repeatable or Reproducible or Replicable

- Repeatability (Same team, same experimental setup)
  - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- Replicability (Different team, same experimental setup)
  - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- Reproducibility (Different team, different experimental setup)
  - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

https://www.acm.org/publications/policies/artifact-review-badging

Many choices

Languages

R

Platforms

hadoop

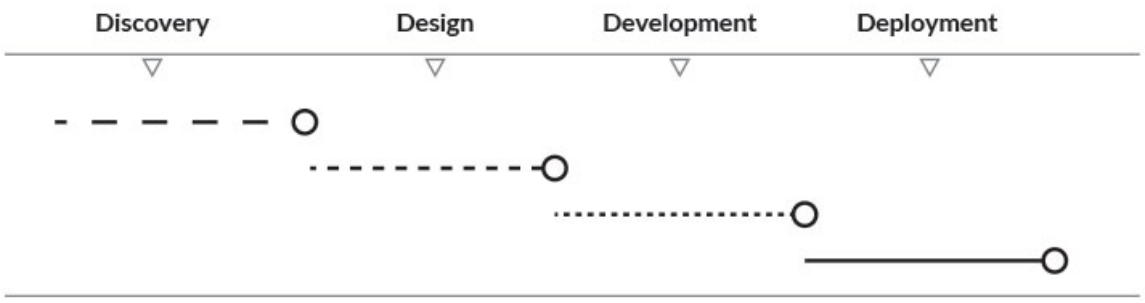
APACHE Spark

APACHE kafka
A distributed streaming platform

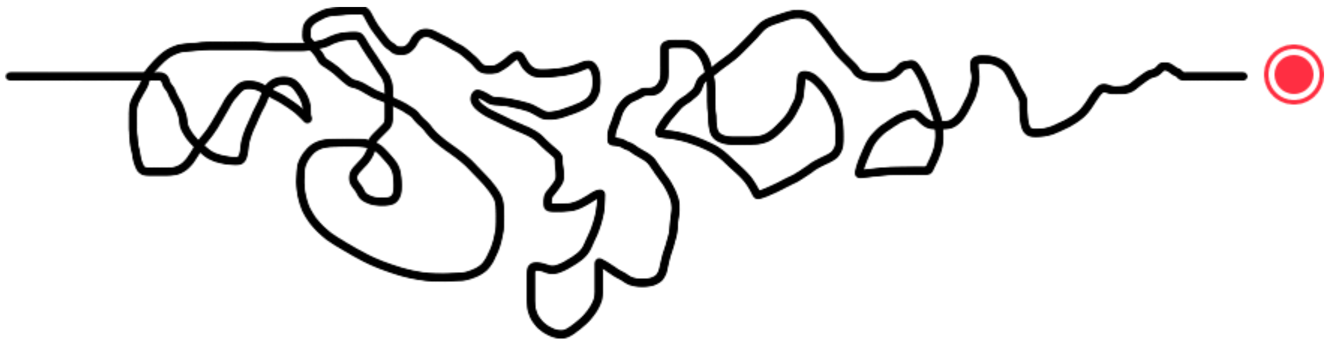
Frameworks

Caffe | Chainer | DL4J Deeplearning4j
MINERVA | mxnet | Purine
KERAS | Microsoft CNTK | MatConvNet
TensorFlow | theano | torch

Processes are chaotic



Planning



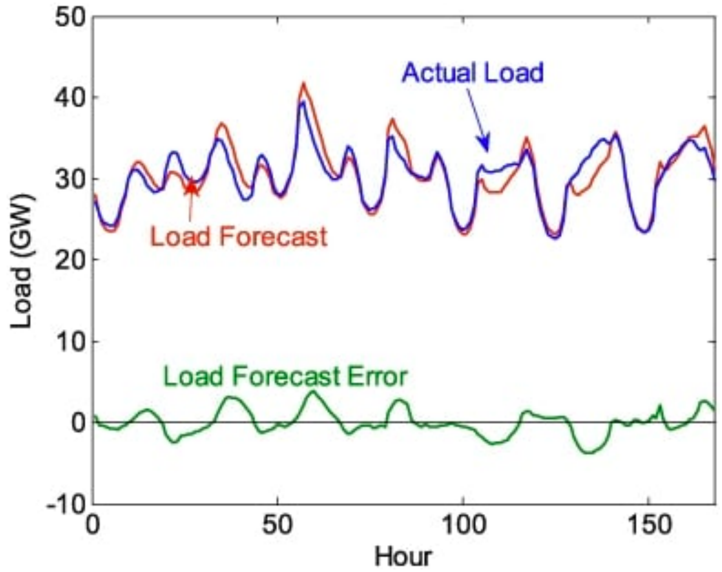
Reality

Multiple stakeholders

Engineering/IT	Quants/Data Scientists
• Scaling • Structuring • Design of Experiments • Data Parallel/Task Parallel	• New Algorithms • Try new methods • Effect of Parameters and Hyper Parameters

Which Model to choose ?

Client Objective:
- Build the best forecasting model that has a MAPE of 5% or less



Source: http://engineering.electrical-equipment.org/electrical-distribution/electric-load-forecasting-advantages-challenges.html

Result:
- Regression – 7% MAPE
- Neural Networks – 4% MAPE
- Random Forest – 5% MAPE

Client choice:
- Regression despite being the worst of the top-3 models
- *"I won't deploy anything that I don't understand"*

Elements of Model Risk Management

1. **Model Governance structure**: Addresses regulatory requirements, roles, responsibilities, oversight, control and escalation procedures
2. **Model Lifecycle management**: Addresses the processes involved in the design, development, testing, deployment and use of models. Also addresses testing and documentation plans and change management.
3. **Model Review and Validation Process**: Addresses internal and external model review, verification, validation and ongoing monitoring of models (both qualitative and quantitative)

AI Governance is gaining focus

AI system: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

AI system lifecycle: AI system lifecycle phases involve: *i)* 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; *ii)* 'verification and validation'; *iii)* 'deployment'; and *iv)* 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.

AI knowledge: AI knowledge refers to the skills and resources, such as data, code, algorithms, models, research, know-how, training programmes, governance, processes and best practices, required to understand and participate in the AI system lifecycle.

AI actors: AI actors are those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI.

Stakeholders: Stakeholders encompass all organisations and individuals involved in, or affected by, AI systems, directly or indirectly. AI actors are a subset of stakeholders.

https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

NLP pipeline

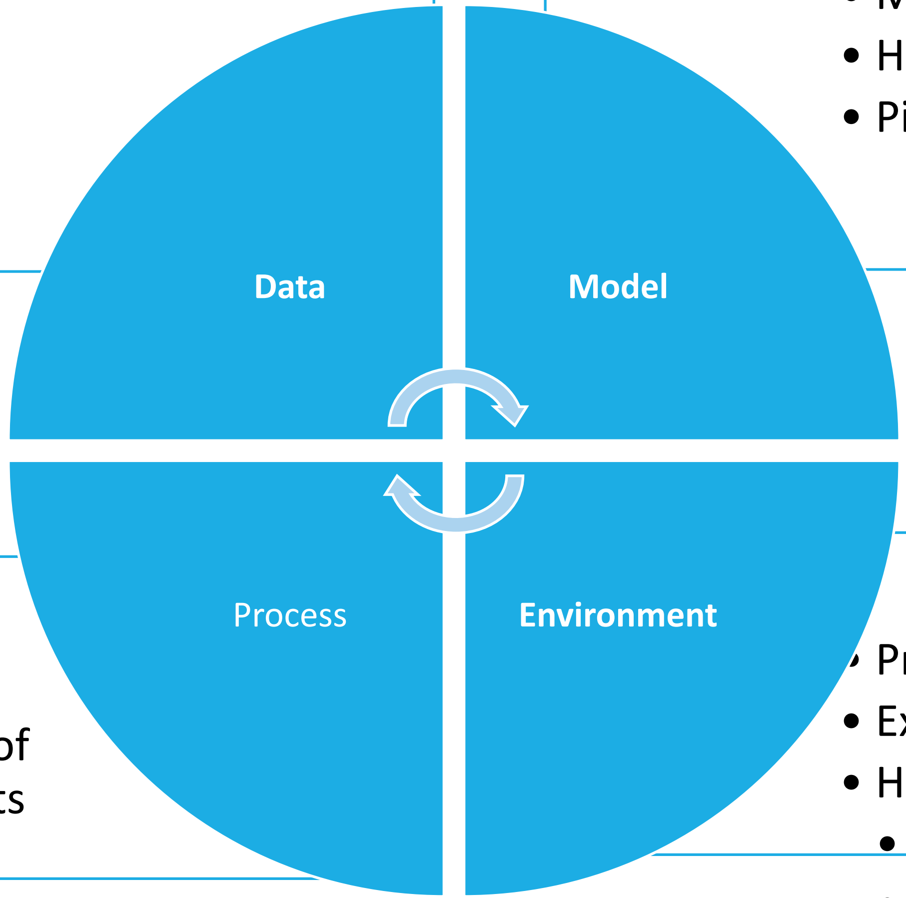
Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Data Ingestion from Edgar	Pre-Processing	Invoking APIs to label data	Compare APIs	Build a new model for sentiment Analysis

- Amazon Comprehend API
- Google API
- Watson API
- Azure API

Components that needs to be tracked

- Model specific
- Tests
- Data versions

- Model params
- Hyper parameters
- Pipeline specifications



- Dependencies
- Lineage/Provenance of individual components

- Programming environment
- Execution environment
- Hardware specs
  - Cloud
  - GPU

Provenance and Lineage of pipelines

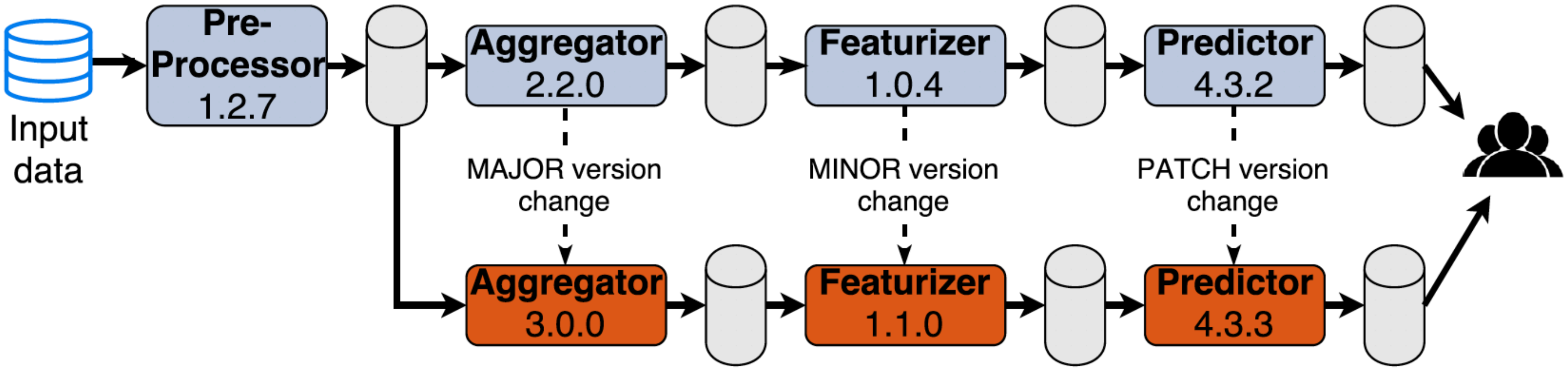


Figure 3: Running multiple pipeline versions

Source: T. van derWeide, O. Smirnov, M. Zielinski, D. Papadopoulos, and T. van Kasteren. Versioned machine learning pipelines for batch experimentation. In ML Systems, Workshop NIPS 2016, 2016.

Schemas proposed

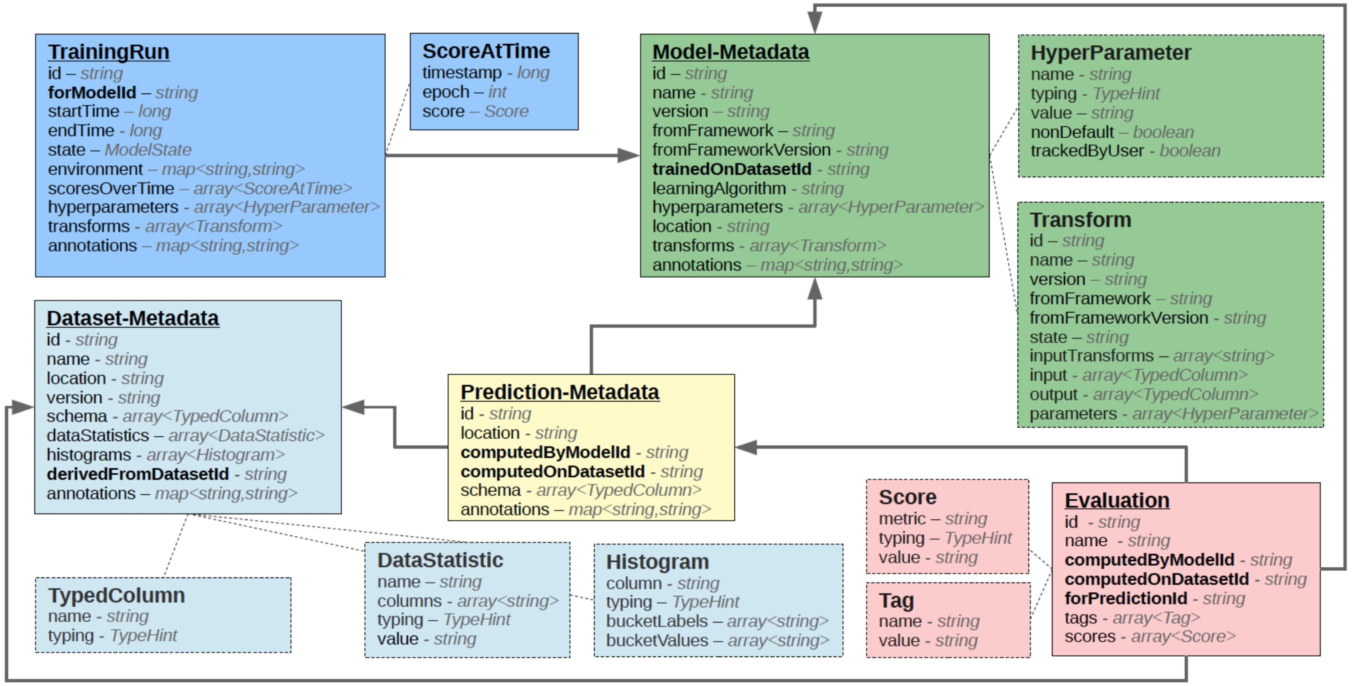


Figure 2: Simplified version of our schema used to store artifact metadata and lineage information. A detailed version available under Apache license can be found at `https://github.com/awslabs/ml-experiments-schema`. Bold attributes indicate lineage relationships and every entity can be extended via arbitrary key-value pairs stored in the annotation attribute.

Sebastian Schelter, Joos-Hendrik Boese, Johannes Kirschnick, Thoralf Klein, and Stephan Seufert. Automatically Tracking Metadata and Provenance of Machine Learning Experiments. NIPS Workshop on Machine Learning Systems, 2017.

Schemas proposed

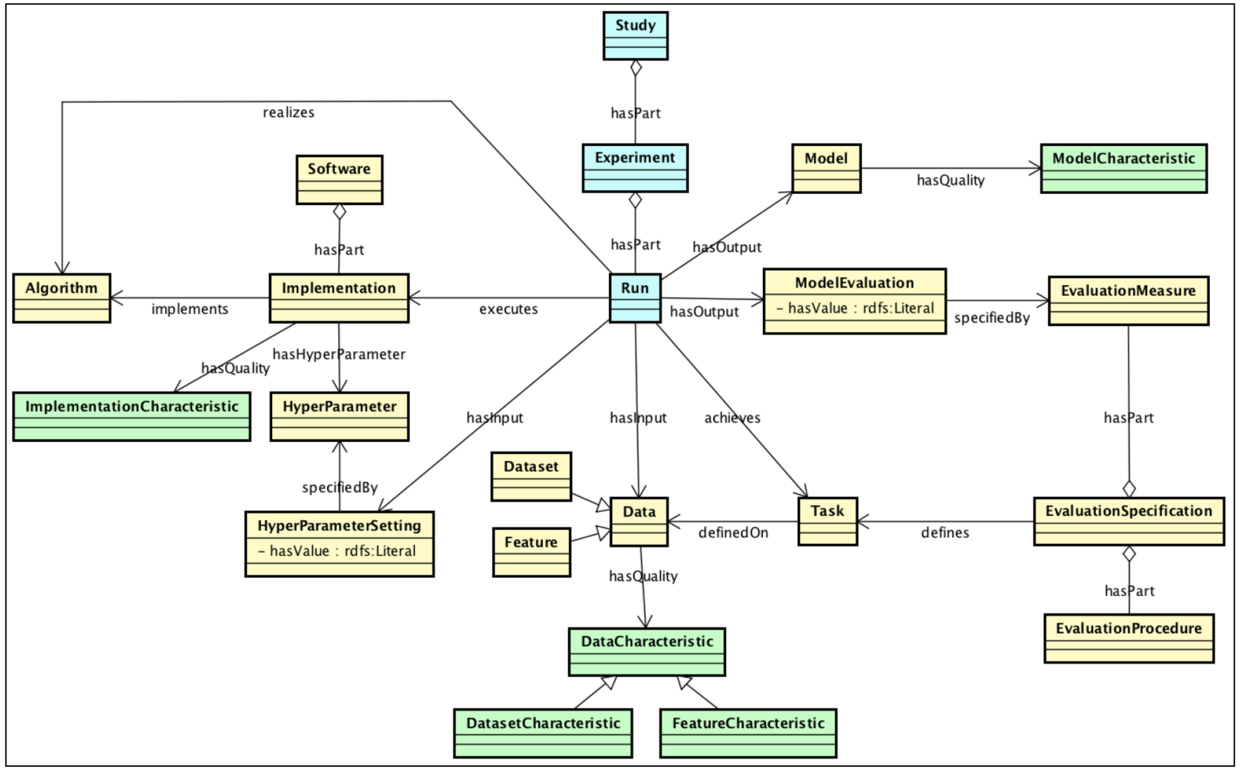


Figure 2. ML Schema Core classes. Boxes represent classes. Arrows without filled heads represent properties, arrows with empty heads represent subclass relations, and arrows with diamonds represent part-of relations.

G. C. Publio, D. Esteves, and H. Zafar, "ML-Schema : Exposing the Semantics of Machine Learning with Schemas and Ontologies," in Reproducibility in ML Workshop, ICML'18, 2018.

QuantUniversity, LLC
www.quantuniversity.com

MLFlow

mlflow GitHub Docs

Default > Run bf34330b7ebc4a07abe335b9a2a6ce2a

Date: 2019-10-05 20:56:09

Source: sklearn_elasticnet_wine

Entry Point: main

Duration: 1.6min

Run ID: bf34330b7ebc4a07abe335b9a2a6ce2a

Git Commit: 60c71fae70aeff9841d60f63e023f129c02dec19

User: srimacpro

Run Command

```
mlflow run file:///Users/srimacpro/mlflow#examples/sklearn_elasticnet_wine -v 60c71fae70aeff9841d60f63e023f129c02dec19 -P alpha=0.5 -P l1_ratio=0.1
```

Notes

None

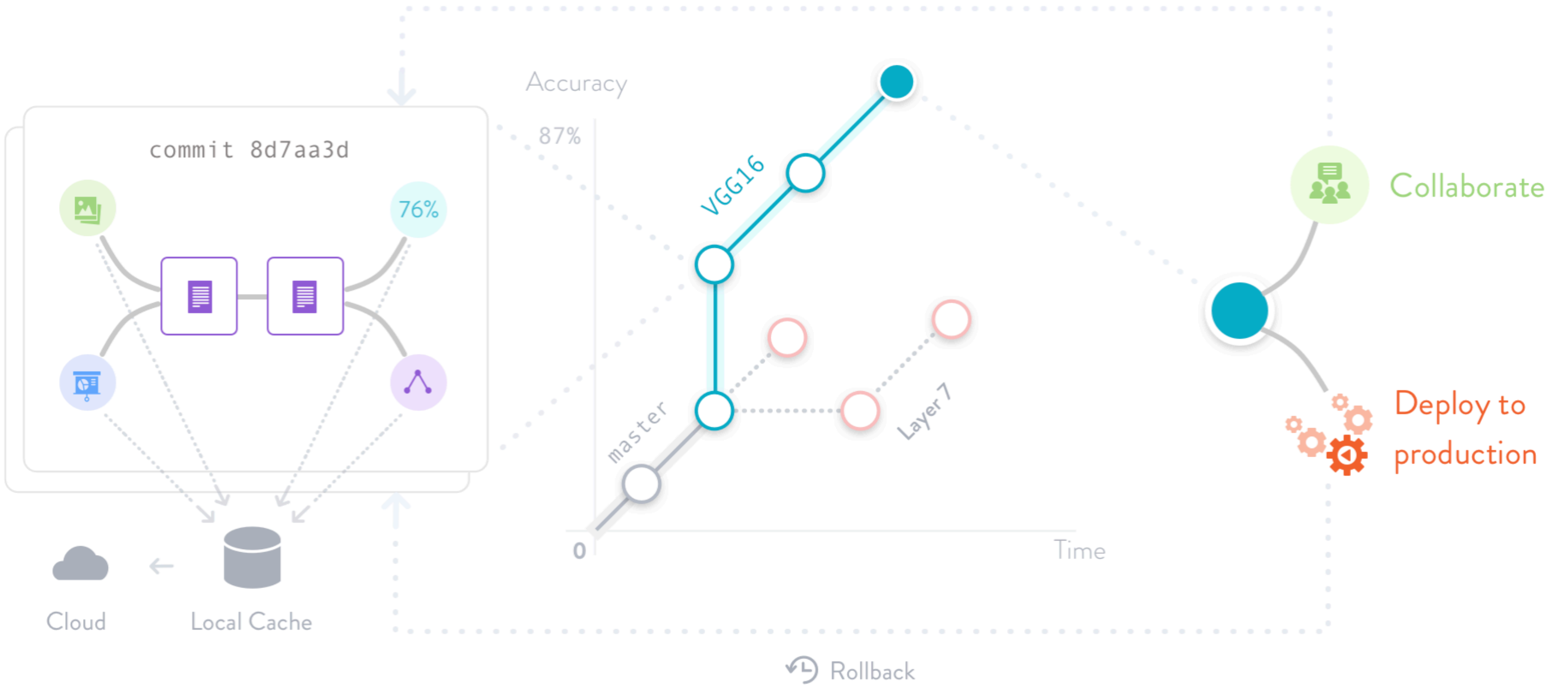
Parameters

Name	Value
alpha	0.5
l1_ratio	0.1

DVC

DVC tracks ML models and data sets

DVC is built to make ML models shareable and reproducible. It is designed to handle large files, data sets, machine learning models, and metrics as well as code.



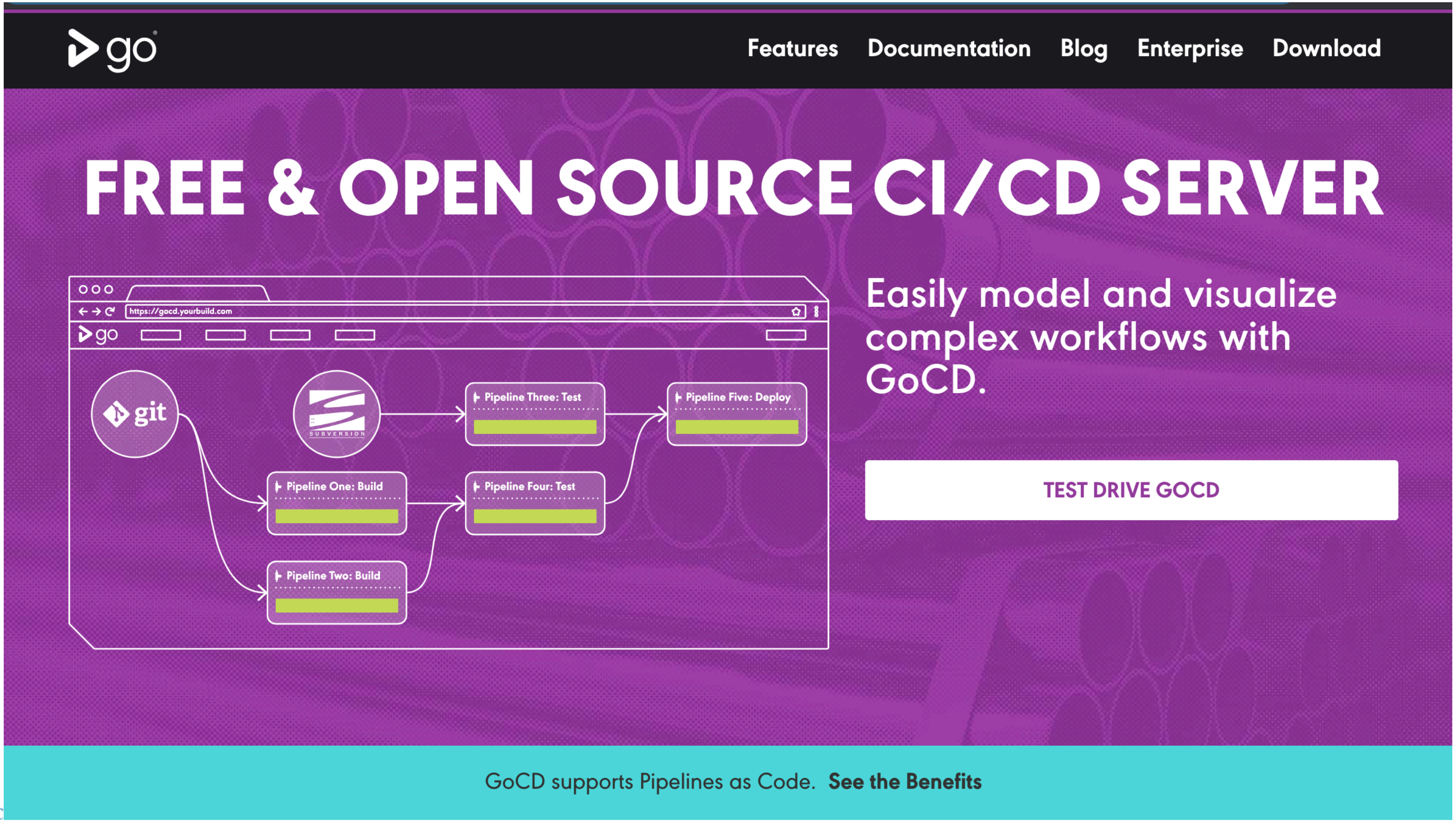
ML project version control

ML experiment management

Deployment & Collaboration

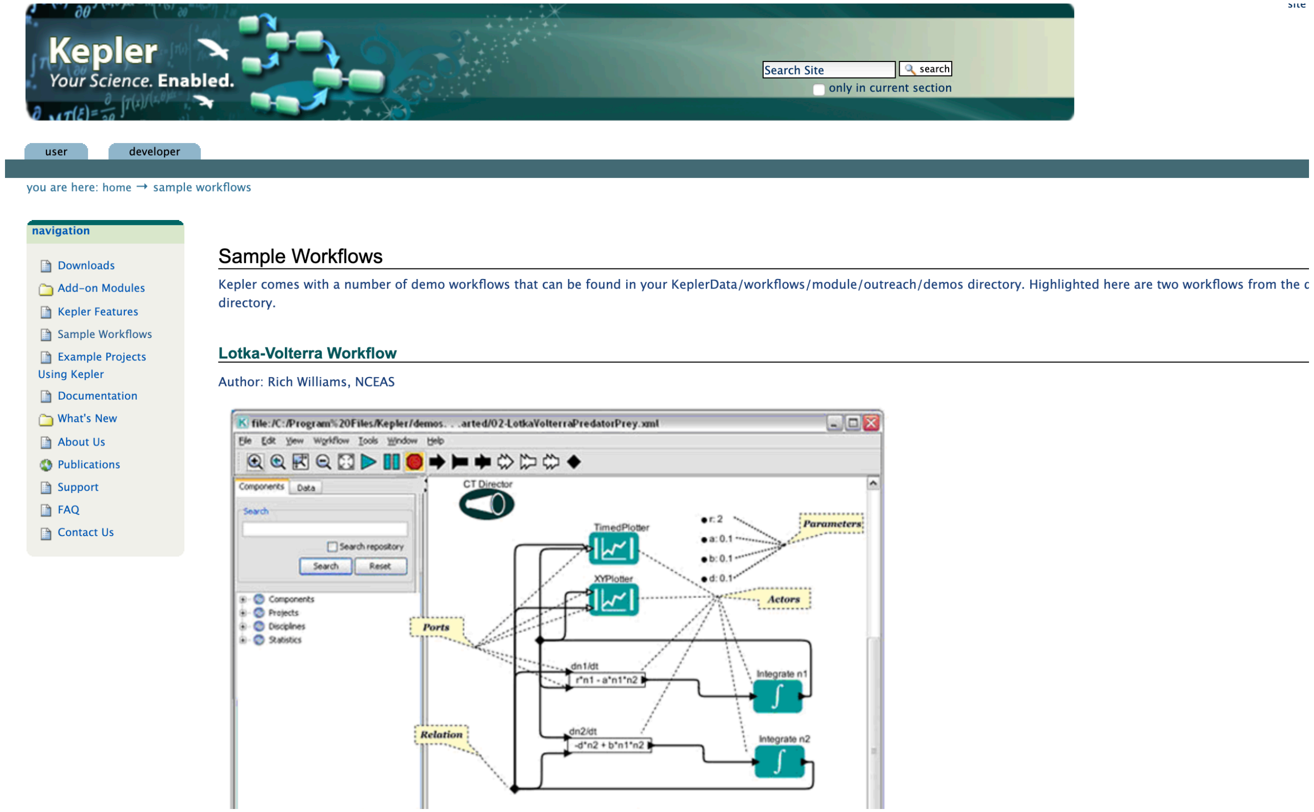
QuantUniversity, LLC
www.quantuniversity.com

GoCD


go
Features
Documentation
Blog
Enterprise
Download
FREE & OPEN SOURCE CI/CD SERVER
https://gocd.yourbuild.com
go
git
SUBVERSION
Pipeline Three: Test
Pipeline Five: Deploy
Pipeline One: Build
Pipeline Four: Test
Pipeline Two: Build
Easily model and visualize complex workflows with GoCD.
TEST DRIVE GOCD
GoCD supports Pipelines as Code. See the Benefits

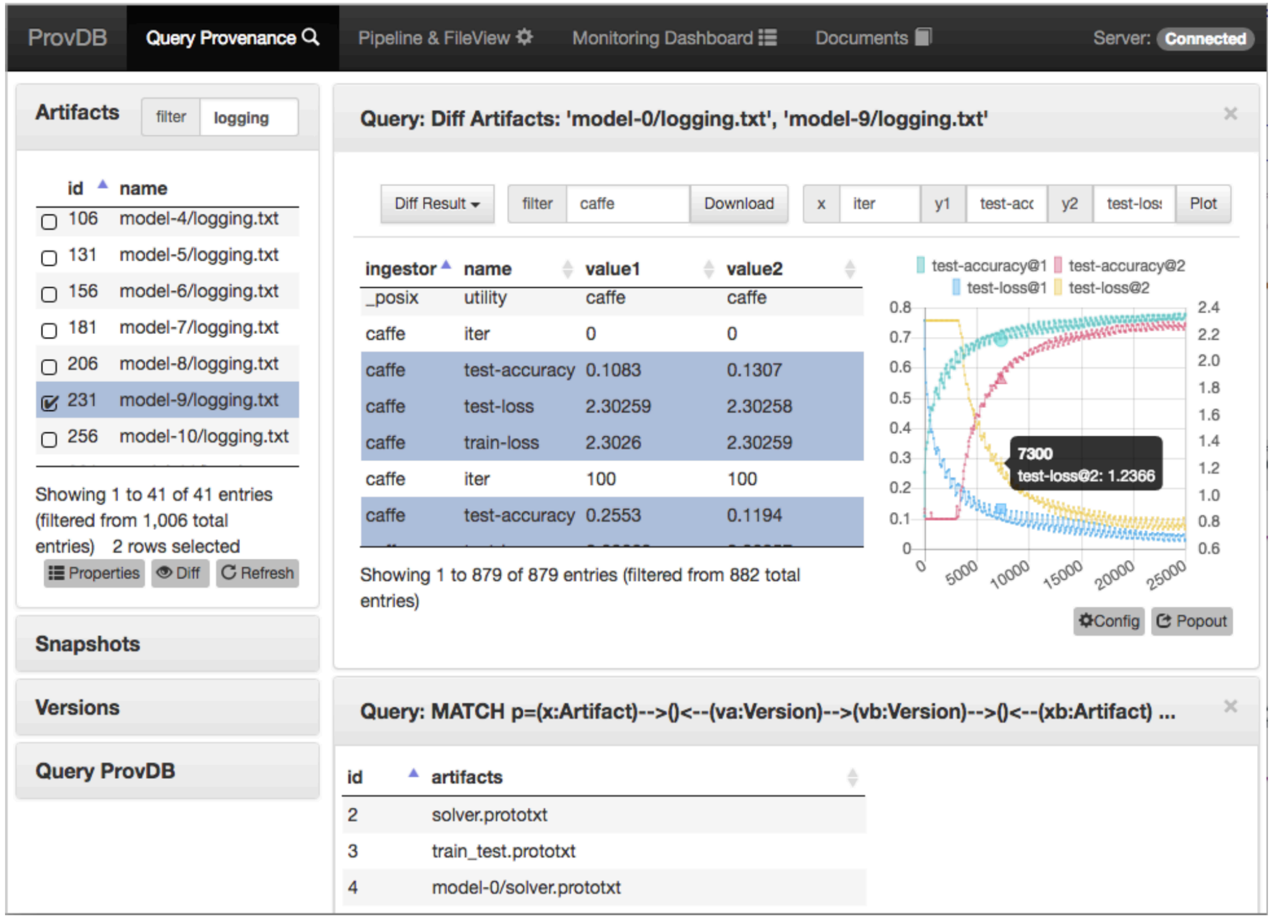
Implementation Approaches

Current approaches



I. Altintas, O. Barney, and E. Jaeger-Frank. Provenance collection support in the Kepler scientific workflow system. In Provenance and annotation of data, pages 118–132.

Current approaches



Miao, Hui & Chavan, Amit & Deshpande, Amol. (2016). ProvDB: A System for Lifecycle Management of Collaborative Analysis Workflows.

Related work

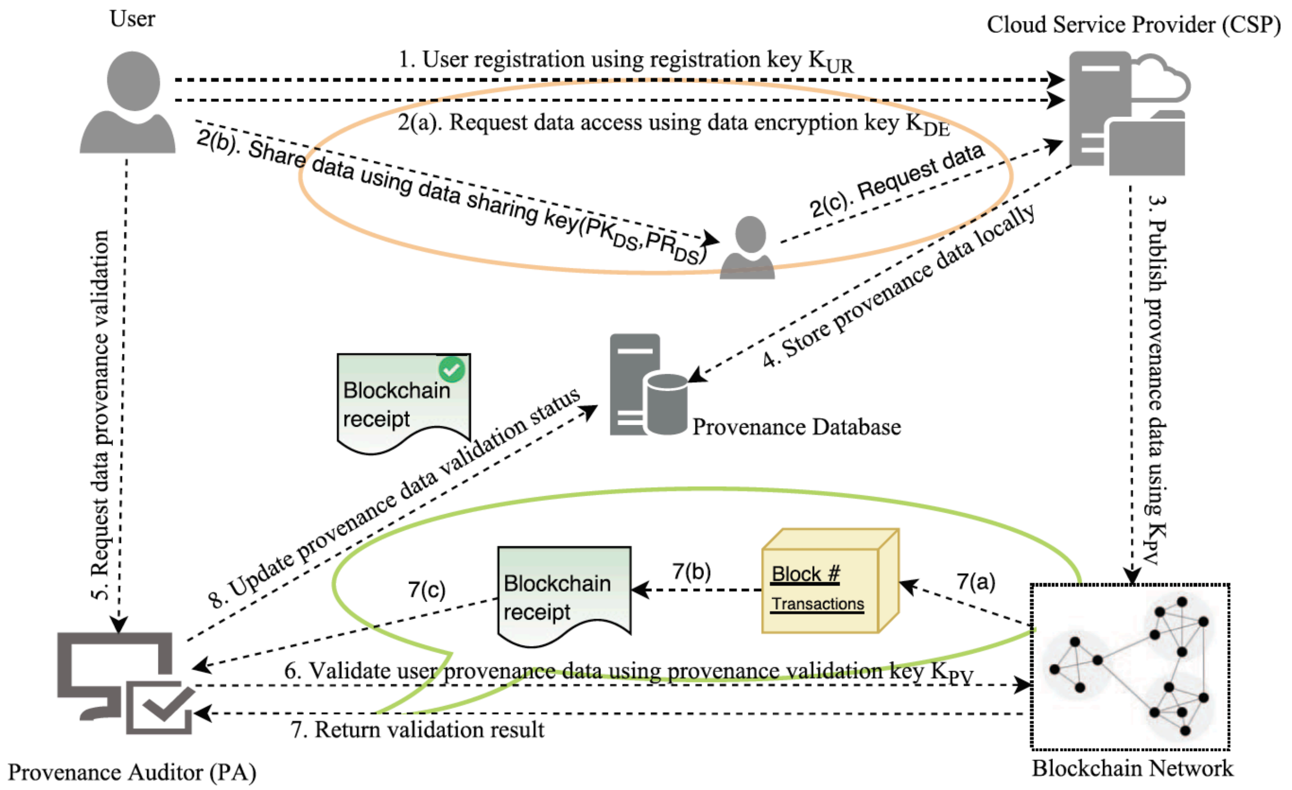


Figure 1: ProvChain System Interaction.

Focus on Cloud data provenance using Blockchain

Xueping Liang, Sachin Shetty, Deepak Tosh, Charles Kamhoua, Kevin Kwiat, and Laurent Njilla. 2017. ProvChain: A Blockchain-based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. In Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '17). IEEE Press, Piscataway, NJ, USA, 468-477. DOI: https://doi.org/10.1109/CCGRID.2017.8

Related work

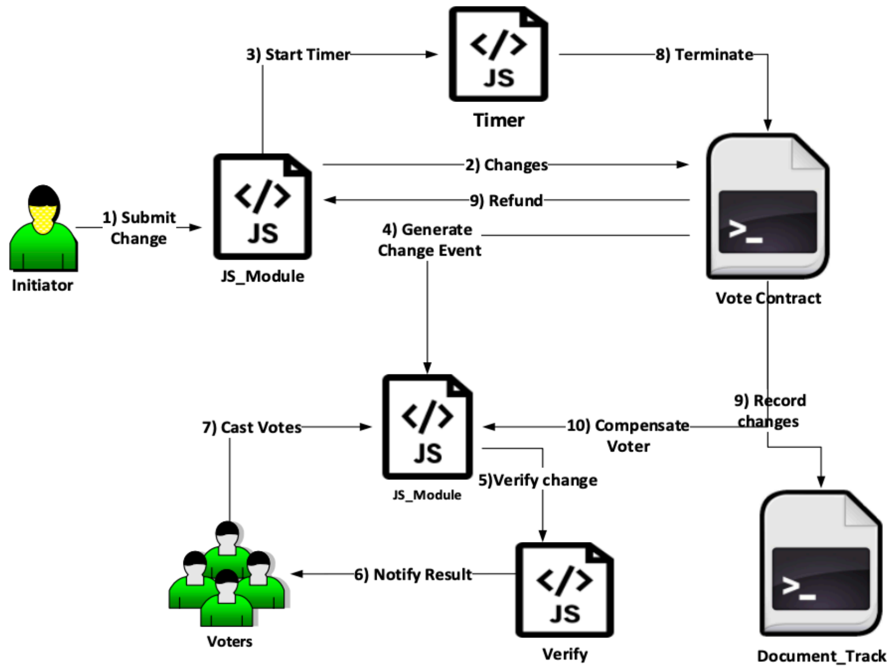


Figure 2: Voting procedure for a Document change.

DataProv: Built on top of Ethereum, the platform utilizes smart contracts and open provenance model (OPM) to record immutable data trails.

Ramachandran, Aravind & Kantarcioglu, Dr. (2017). Using Blockchain and smart contracts for secure data provenance management.

Related work

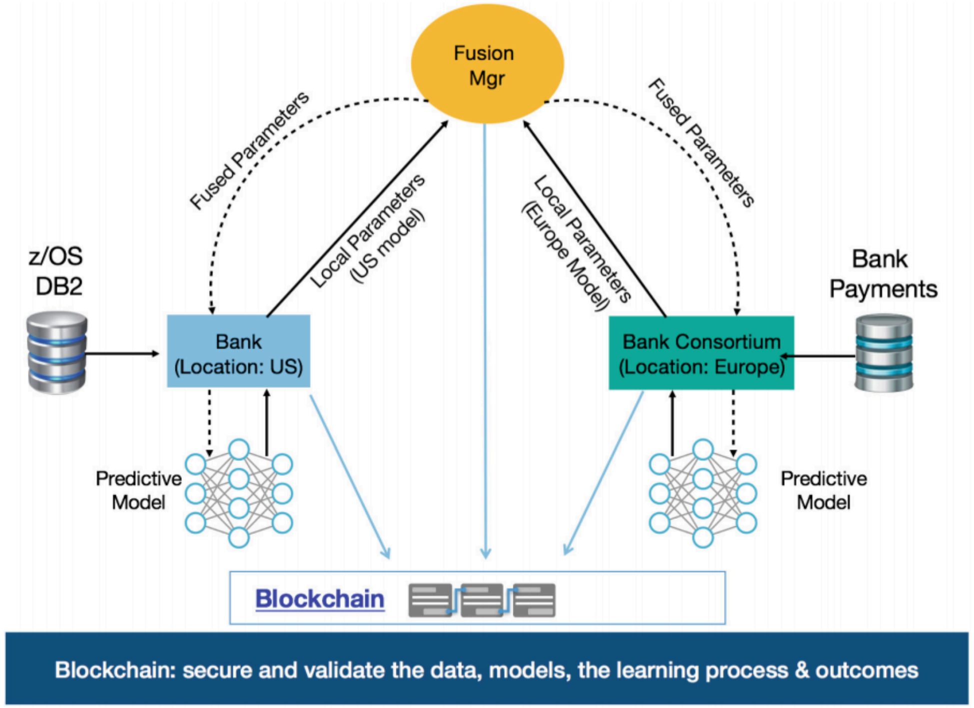


Fig. 2. Trusted federated learning: the basic setting

Trusted AI and provenance of AI models

Sarpatwar, Kanthi & Vaculín, Roman & Min, Hong & Su, Gong & Heath, Terry & Ganapavarapu, Giridhar & Dillenberger, Donna. (2019). Towards Enabling Trusted Artificial Intelligence via Blockchain. 10.1007/978-3-030-17277-0_8.

NLP Case study

Goal

- Understanding sentiments in Earnings call transcripts

sec.gov

CORPORATEPARTICIPANTS

Dana Quattrochi athenahealth, Inc. - IR
Jonathan Bush athenahealth, Inc - Chairman and CEO
Tim Adams athenahealth, Inc - CFO
Andy Hurd Epocrates - President and CEO
Rob Cosinuke athenahealth, Inc. - Chief Marketing Officer

CONFERENCECALLPARTICIPANTS

Sean Wieland Piper Jaffray & Co. - Analyst
Jamie Stockton Wells Fargo Securities, LLC - Analyst
George Hill Citigroup - Analyst
Greg Bolan Sterne, Agee & Leach - Analyst
Ryan Daniels William Blair & Company - Analyst
Rich Close Avondale Partners - Analyst
Sandy Draper Raymond James - Analyst
David Bayer Northland Securities - Analyst
Dave Windley Jefferies & Co. - Analyst
Charles Rhyee Cowen and Company - Analyst
Bret Jones Oppenheimer & Co. - Analyst
Michael Cherny ISI Group - Analyst
Tony Bartsch Park West Asset Management - Analyst

PRESENTATION

Operator

Welcome to the athenahealth conference call. I would now like to turn the call over to Ms. Dana Quattrochi. You may now begin.

Dana Quattrochi - *athenahealth, Inc. - IR*

Good morning and thank you for joining us. With me on the call today is Jonathan Bush, our Chairman and CEO; Tim Adams, our Chief Financial Officer; Rob Cosinuke, our Chief Marketing Officer; and Andy Hurd, President and CEO of Epocrates.

Challenges

- Interpreting emotions
- Labeling data

<u>Options</u>

- APIs
- Human Insight
- Expert Knowledge
- Build your own

NLP pipeline

Stage 1 – Data Ingestion from Edgar

Stage 2 – Pre-Processing

Stage 3 – Invoking APIs to label data

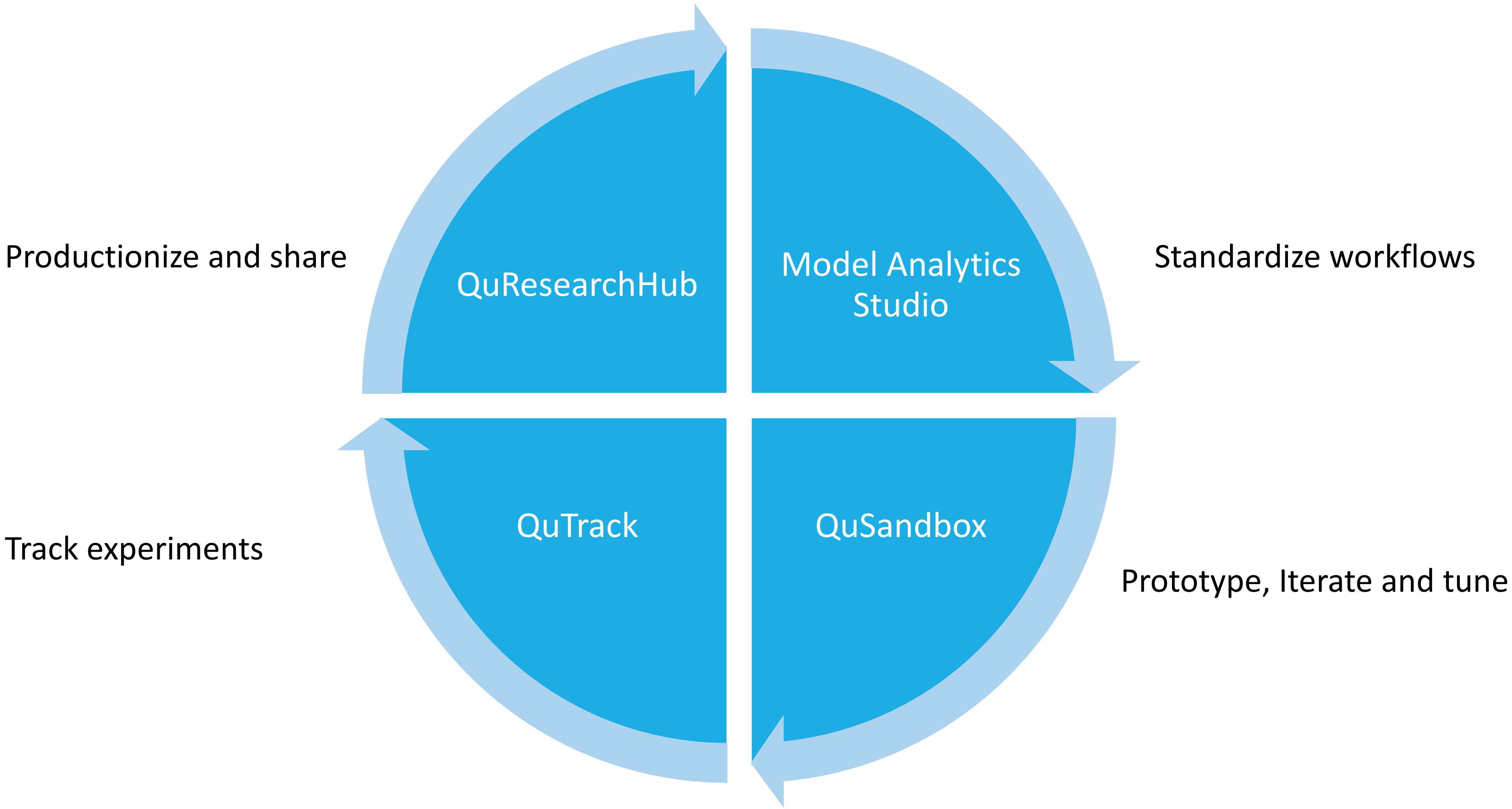
Stage 4 – Compare APIs

Stage 5 – Build a new model for sentiment Analysis

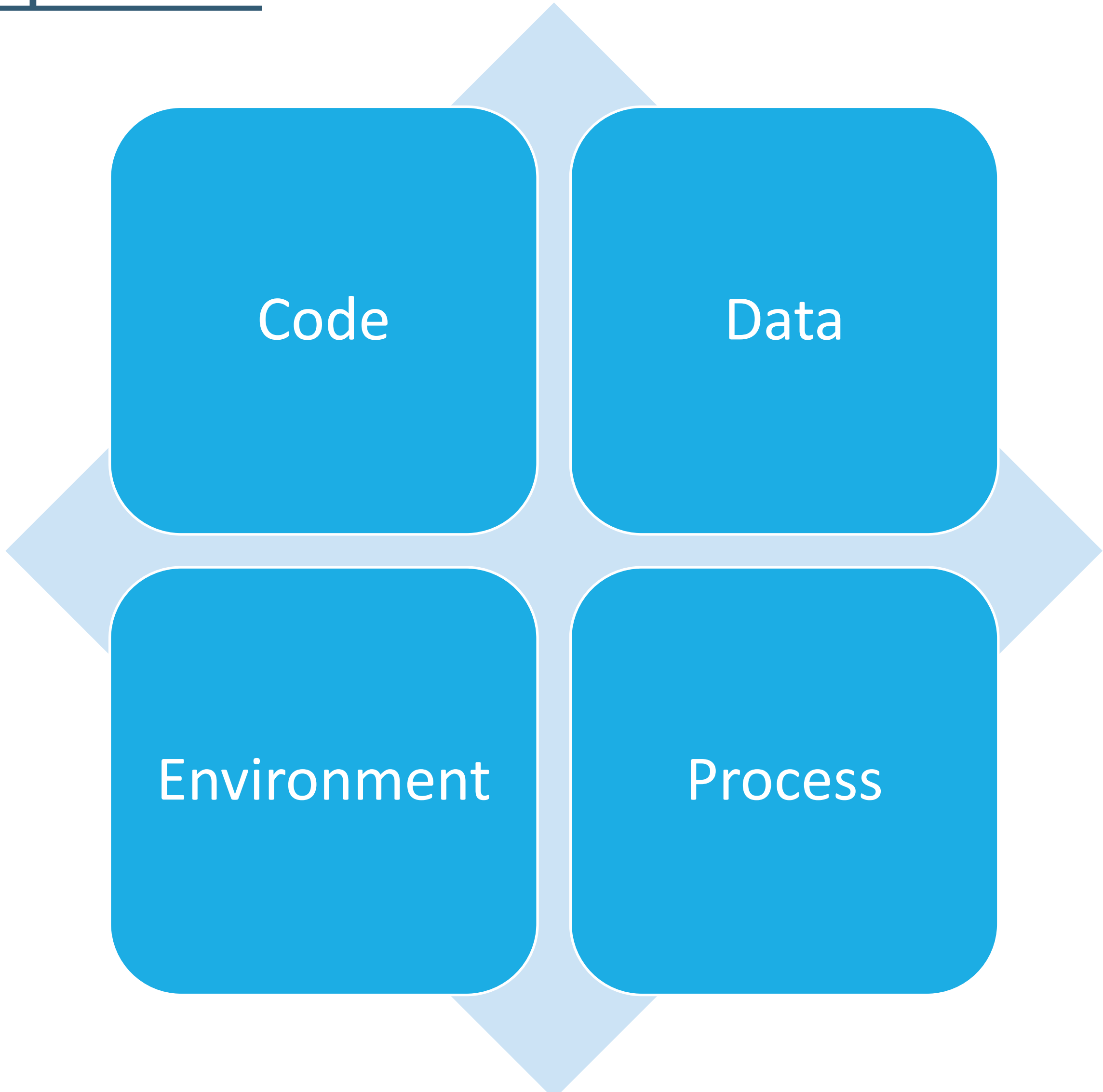
- Amazon Comprehend API
- Google API
- Watson API
- Azure API

QuSandbox- The platform for governing Data Science and AI workflows in the Enterprise

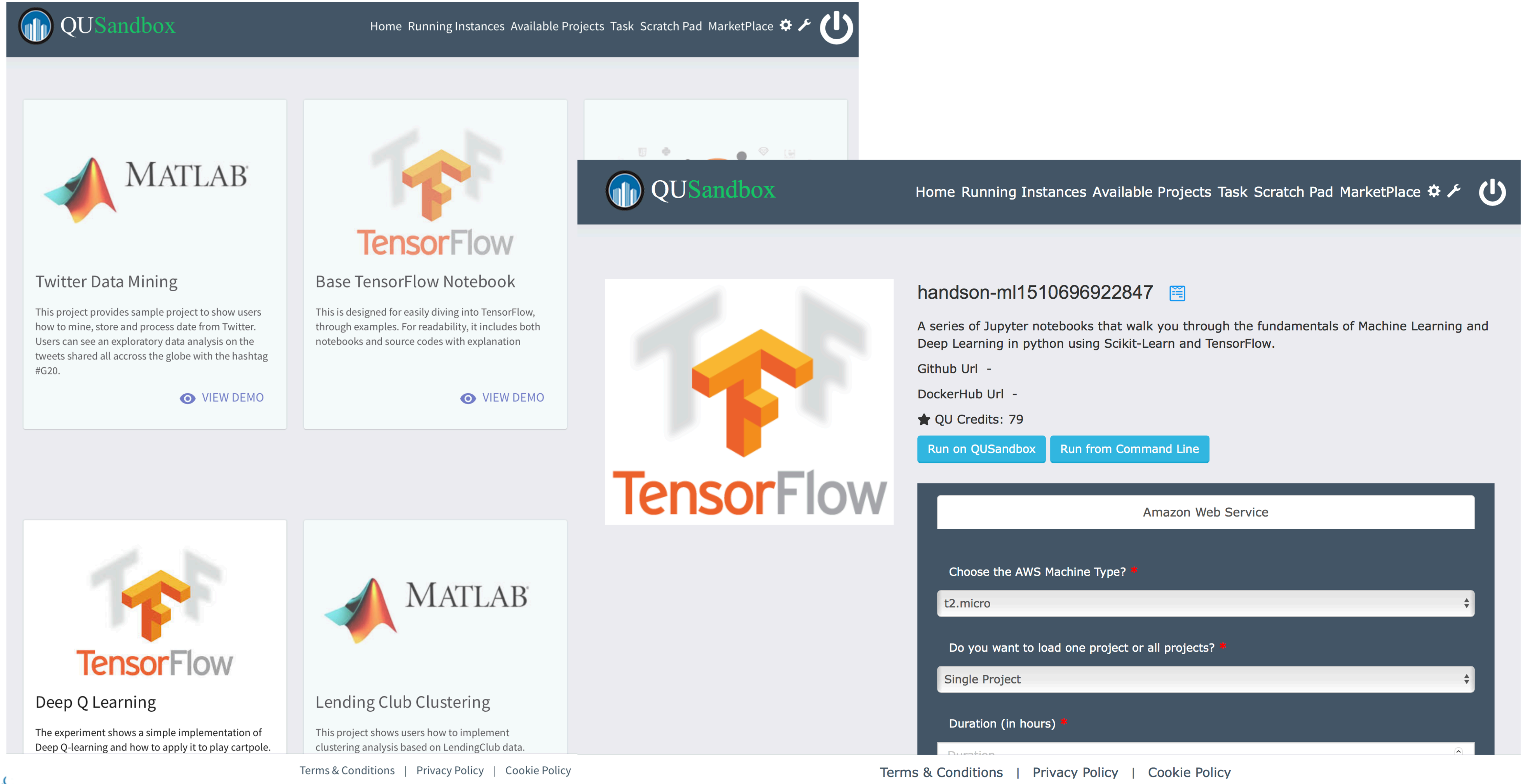
QuSandbox research suite

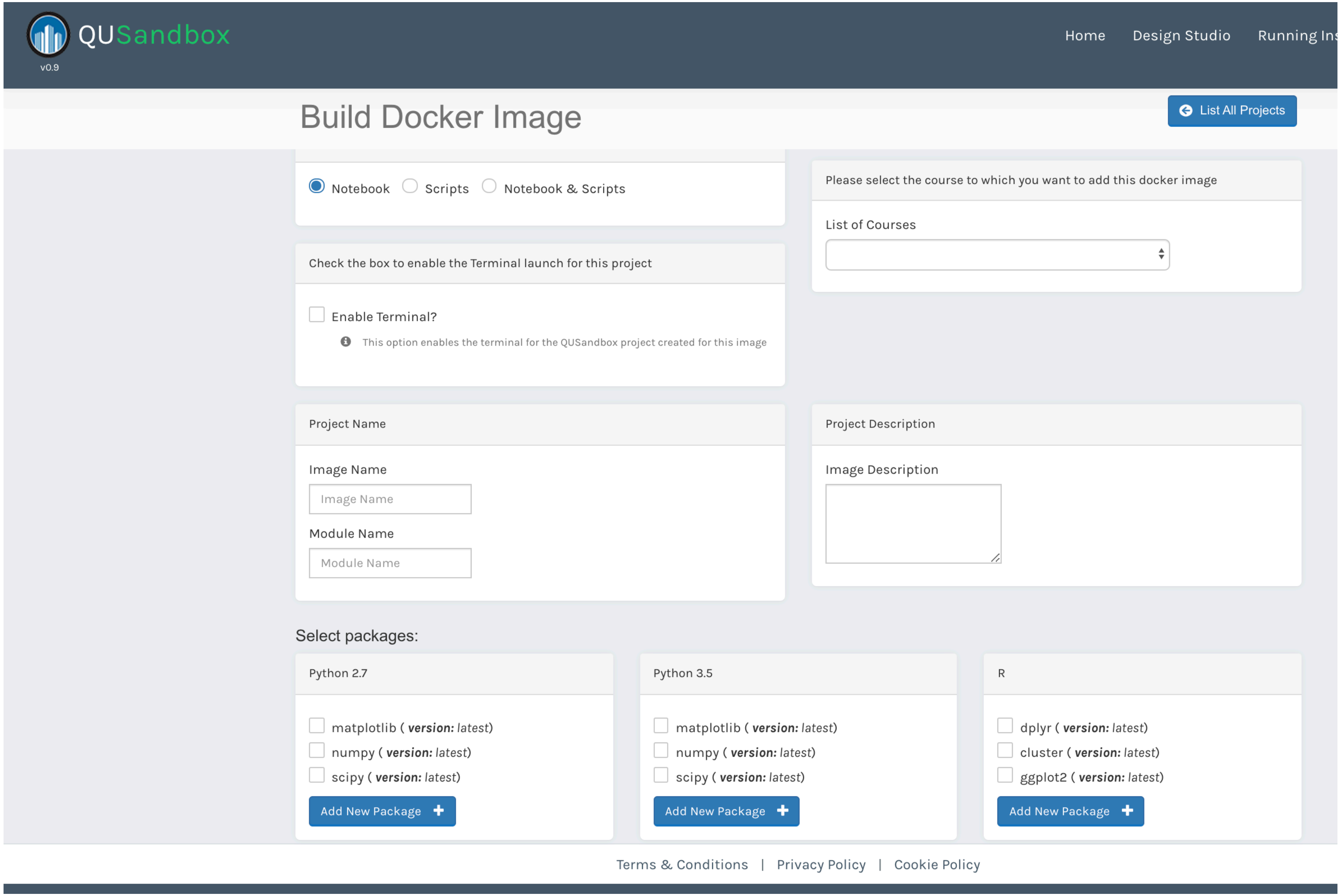


The four components that need to be encapsulated for reproducible pipelines

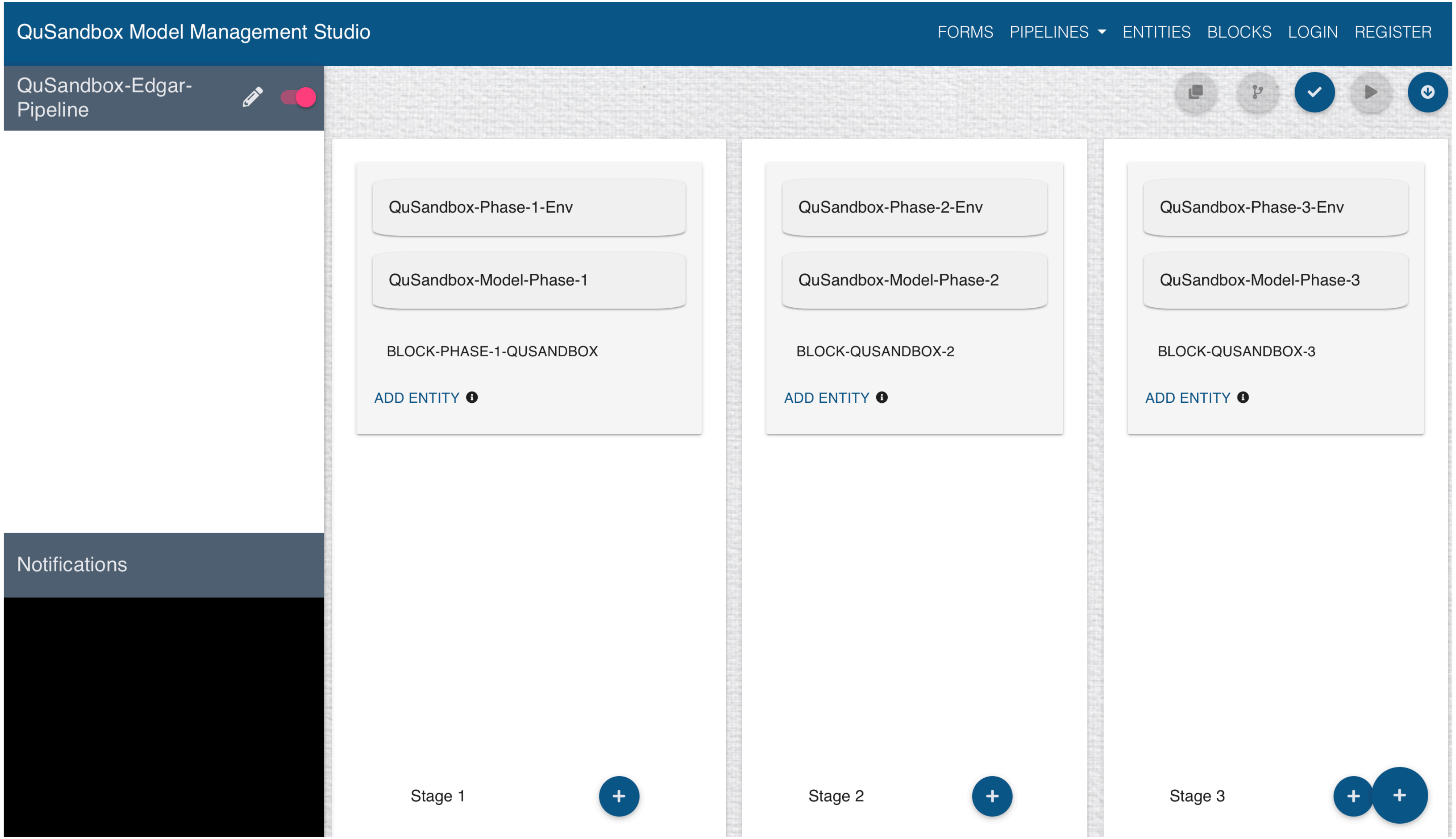


QuSandbox


QUSandbox
Home Running Instances Available Projects Task Scratch Pad MarketPlace
MATLAB
Twitter Data Mining
This project provides sample project to show users how to mine, store and process date from Twitter. Users can see an exploratory data analysis on the tweets shared all accross the globe with the hashtag #G20.
VIEW DEMO
TensorFlow
Base TensorFlow Notebook
This is designed for easily diving into TensorFlow, through examples. For readability, it includes both notebooks and source codes with explanation
VIEW DEMO
TensorFlow
Deep Q Learning
The experiment shows a simple implementation of Deep Q-learning and how to apply it to play cartpole.
MATLAB
Lending Club Clustering
This project shows users how to implement clustering analysis based on LendingClub data.
Terms & Conditions | Privacy Policy | Cookie Policy
QUSandbox
Home Running Instances Available Projects Task Scratch Pad MarketPlace
TensorFlow
handson-ml1510696922847
A series of Jupyter notebooks that walk you through the fundamentals of Machine Learning and Deep Learning in python using Scikit-Learn and TensorFlow.
Github Url -
DockerHub Url -
QU Credits: 79
Run on QUSandbox
Run from Command Line
Amazon Web Service
Choose the AWS Machine Type? *
t2.micro
Do you want to load one project or all projects? *
Single Project
Duration (in hours) *
Terms & Conditions | Privacy Policy | Cookie Policy


QUSandbox
v0.9
Home
Design Studio
Running Ins
Build Docker Image
List All Projects
Notebook
Scripts
Notebook & Scripts
Please select the course to which you want to add this docker image
List of Courses
Check the box to enable the Terminal launch for this project
Enable Terminal?
This option enables the terminal for the QUSandbox project created for this image
Project Name
Image Name
Image Name
Module Name
Module Name
Project Description
Image Description
Select packages:
Python 2.7
matplotlib (version: latest)
numpy (version: latest)
scipy (version: latest)
Add New Package
Python 3.5
matplotlib (version: latest)
numpy (version: latest)
scipy (version: latest)
Add New Package
R
dplyr (version: latest)
cluster (version: latest)
ggplot2 (version: latest)
Add New Package
Terms & Conditions | Privacy Policy | Cookie Policy

Model Management Studio



Terms

- **JDF**: Job Definition File; A DSL for representing Model Pipelines
- **Stage**
- **Entity**
  - Model
  - Data
  - Environment
- **Version format**
  - **M:m:p -> Major Version: Minor Version: Patch**

JDF- DSL

jdf-QuSandbox-Edgar-Pipeline.json


```
{"status":"active","_id":{"version":1,"id":20},"name":"QuSandbox-Edgar-Pipeline","access_role":"public","connections":[],"stages":[[{"status":"active","name":"Block-
Phase-1-QuSandbox","access_role":"public","output_number":0,"entities":[{"status":"active","description":"","access_role":"admin","_id":{"version":1,"id":63},"type":
{"username":"ubuntu","src":"AWS","name":"Environment","single_project":true,"os":"ubuntu","start_from":"DOCKERHUB","docker_hub_url":"jdadaptivealgo/
sentiment_analysis_1","duration":3,"use_ac":false,"config":{"callback":"","post_execution":"","pre_execution":"","parameters":
[],"run_command":""},"size":"t2.micro"},"name":"QuSandbox-Phase-1-Env"},{"status":"active","description":"","access_role":"admin","_id":{"version":1,"id":65},"type":
{"src":"sentiment_analysis_scraping1530543528937","config":{"callback":"","post_execution":"","pre_execution":"","parameters":
[],"run_command":""},"name":"Model","src_type":"QuSandbox"},"name":"QuSandbox-Model-Phase-1"}],"input_number":0,"_id":{"version":1,"id":27},"description":""}],
[{"status":"active","name":"Block-QuSandbox-2","access_role":"admin","output_number":0,"entities":[{"status":"active","description":"","access_role":"admin","_id":
{"version":1,"id":68},"type":
{"username":"ubuntu","src":"AWS","name":"Environment","single_project":true,"os":"Ubuntu","start_from":"DOCKERHUB","docker_hub_url":"jdadaptivealgo/
sentiment_analysis_2","duration":3,"use_ac":false,"config":{"callback":"","post_execution":"","pre_execution":"","parameters":
[],"run_command":""},"size":"t2.micro"},"name":"QuSandbox-Phase-2-Env"},{"status":"active","description":"","access_role":"admin","_id":{"version":1,"id":64},"type":
{"src":"sentiment_analysis_pre_processing1530543608233","config":{"callback":"","post_execution":"","pre_execution":"","parameters":
[],"run_command":""},"name":"Model","src_type":"QuSandbox"},"name":"QuSandbox-Model-Phase-2"}],"input_number":0,"_id":{"version":1,"id":28},"description":""}],
[{"status":"active","name":"Block-QuSandbox-3","access_role":"admin","output_number":0,"entities":[{"status":"active","description":"","access_role":"admin","_id":
{"version":1,"id":66},"type":
{"username":"ubuntu","src":"AWS","name":"Environment","single_project":true,"os":"Ubuntu","start_from":"DOCKERHUB","docker_hub_url":"jdadaptivealgo/
sentiment_analysis_3","duration":3,"use_ac":false,"config":{"callback":"","post_execution":"","pre_execution":"","parameters":
[],"run_command":""},"size":"t2.micro"},"name":"QuSandbox-Phase-3-Env"},{"status":"active","description":"","access_role":"admin","_id":{"version":1,"id":67},"type":
{"src":"sentiment_analysis_stage31531847777475","config":{"callback":"","post_execution":"","pre_execution":"","parameters":
[],"run_command":""},"name":"Model","src_type":"QuSandbox"},"name":"QuSandbox-Model-Phase-3"}],"input_number":0,"_id":{"version":1,"id":
29},"description":""}]],"description":""}
```

QuResearchHub

QUResearchHub
Powered By QUSandbox

Home Profile

Table of Contents

Back to the Projects

Sentiment Analysis with QuSandbox

QuantUniversity Team

Abstract: EDGAR, the Electronic Data Gathering, Analysis, and Retrieval system, performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the U.S. Securities and Exchange Commission (the "SEC"). The database contains a wealth of information about the Commission and the securities industry which is freely available to the public via the Internet (HTTPS).[1] In this project, we intend to analyze the sentiment of each paragraph in a form-425 file of a specific company downloaded from Edgar. This form-425 file is phone call transaction which contains a lot of dialogues between operator and clients. Thus, Sentiment analysis will be applied on each paragraph in the file and return either a sentiment label or a sentiment score or both of them to the user. Also, we want to test and compare the performances of different NLP APIs in this project. Amazon Comprehend, Microsoft Azure, Google Cloud and IBM Watson will be used in the project to analysis the same file. Moreover, before the actual analysis, a crawler and a preprocessing phase need to be designed in order to have a analysable data set.

This sprint involved integration of the Model management Studio, QuSandbox and the ResearchHub

QULink

Additional Links

1. **Model Management workflow**
   We illustrate how QuSandbox can be used to set up a pipeline to enable crawling, pre-processing and prediction of sentiments. The workflow is al illustration of the key concepts and terminologies used to structure the pipeline

2. **Using CLI tools to automate the workflow**
   This is to illustrate how the CLI can be used to invoke actions on the QuSandbox. The CLI tools enable access to the QuSandbox without the need of the Model Management Studio. This enables integration with third party scheduling tools.

3. **Sentiment analysis workflow**
   This experiment has 3 stages. In the first stage, we crawl the form 425 data from the Edgar website and store it to a Amazon S3 bucket. In the second stage, we pre-process the data and store it back to the Amazon S3 bucket. In the third stage, we let the quant perform sentiment analysis using the API of his/her choice. We provide Jupiter notebooks for 4 APIs. 1. Amazon's Comprehend API, 2. Google API, 3. IBM Watson API, 4. Microsoft API

Architecture : What's tracked ?

Metadata

- Data about the information to be tracked
- Includes version number, timestamps, user information, MD5 of the artifacts and high-level notes

Data

- Pipelines, custom DSL, standard formats for representing models
- Events (Updates, rollbacks
- JSON, Amazon ION, YAML,

Artifacts

- Model Pickle files, ONYX, COREML, Model params
- Data, blobs etc.

Architectures supported

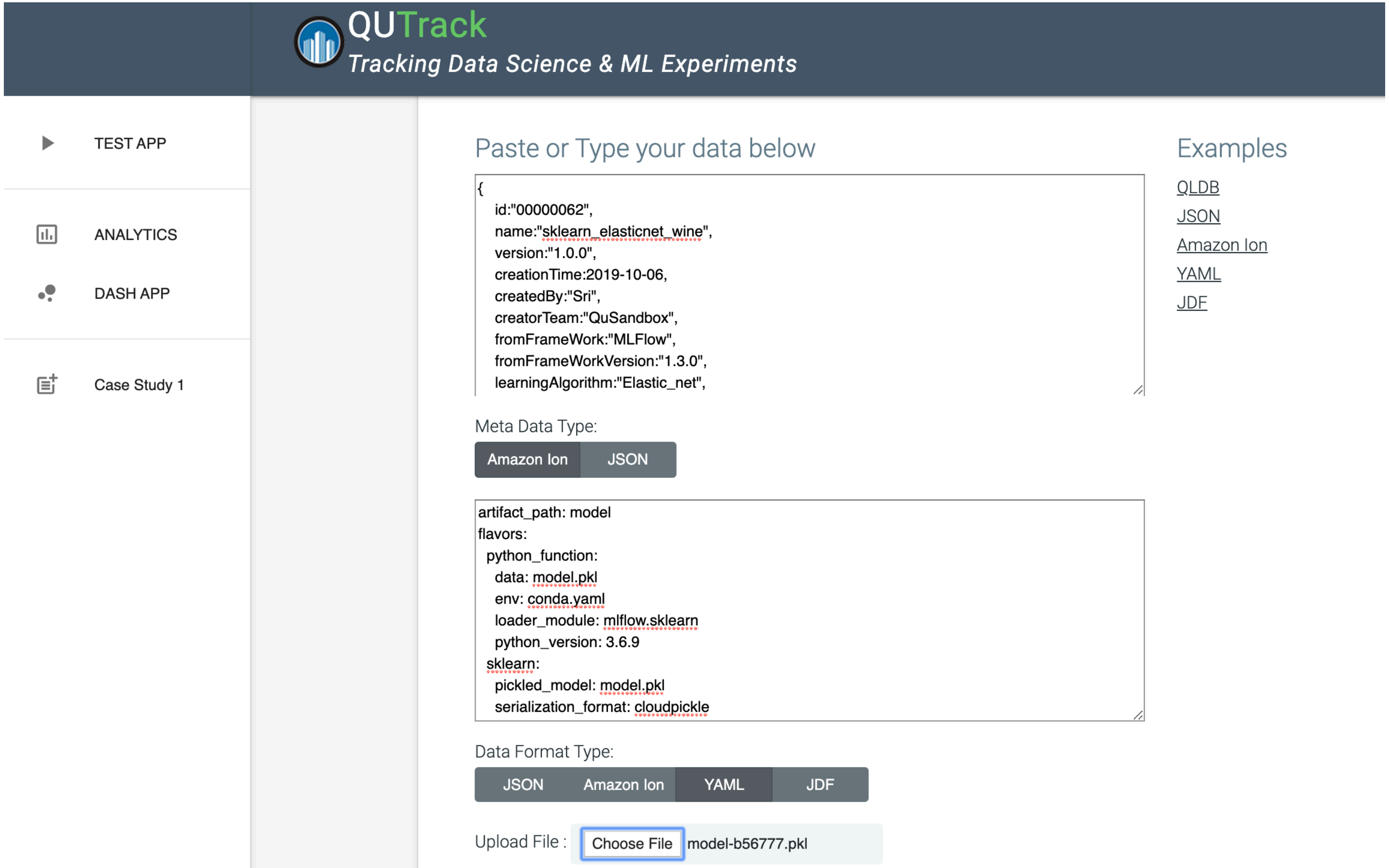
Blockchain-based:

- QLDB
- Ethereum

Non-Blockchain-based:

- MongoDB

QuTrack


QUTrack
Tracking Data Science & ML Experiments
TEST APP
ANALYTICS
DASH APP
Case Study 1
Paste or Type your data below
{
id:"00000062",
name:"sklearn_elasticnet_wine",
version:"1.0.0",
creationTime:2019-10-06,
createdBy:"Sri",
creatorTeam:"QuSandbox",
fromFrameWork:"MLFlow",
fromFrameWorkVersion:"1.3.0",
learningAlgorithm:"Elastic_net",
Meta Data Type:
Amazon Ion
JSON
artifact_path: model
flavors:
python_function:
data: model.pkl
env: conda.yaml
loader_module: mlflow.sklearn
python_version: 3.6.9
sklearn:
pickled_model: model.pkl
serialization_format: cloudpickle
Data Format Type:
JSON
Amazon Ion
YAML
JDF
Upload File :
Choose File
model-b56777.pkl
Examples
QLDB
JSON
Amazon Ion
YAML
JDF

Demo

Future work

- Support for ONYX, CoreML
- Integration with:
  - MLFlow, DVC, GoCD
- Integration with SCM systems
  - Github, SVM
- Tracking Back tests
- Push Architecture -> Event-Driven Architecture
- Enriched Analytics
- Roles and Authorization

QuantUniversity, LLC

www.quantuniversity.com

Sign up for Updates at

www.qusandbox.com

LinkedIn srikrishnamurthy

www.QuantUniversity.com
www.analyticscertificate.com