



QuantUniversity, LLC

[www.quantuniversity.com](http://www.quantuniversity.com)

# A Master Class in Building Production-Grade NLP Pipelines

## Presented By:

Sri Krishnamurthy, CFA, CAP

[sri@quantuniversity.com](mailto:sri@quantuniversity.com)

[www.quantuniversity.com](http://www.quantuniversity.com)

10/15/2019

MathWorks Conference

New York, NY

# Speaker bio



QuantUniversity, LLC

www.quantuniversity.com




Sri Krishnamurthy  
Founder and CEO  
QuantUniversity

- 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  a platform for operationalizing AI and Machine Learning in the Enterprise



Get the app  English

QuantUniversity Meetup

[Home](#) [Members](#) [Sponsors](#) [Photos](#) [Discussions](#) [More](#)

[Join us!](#)



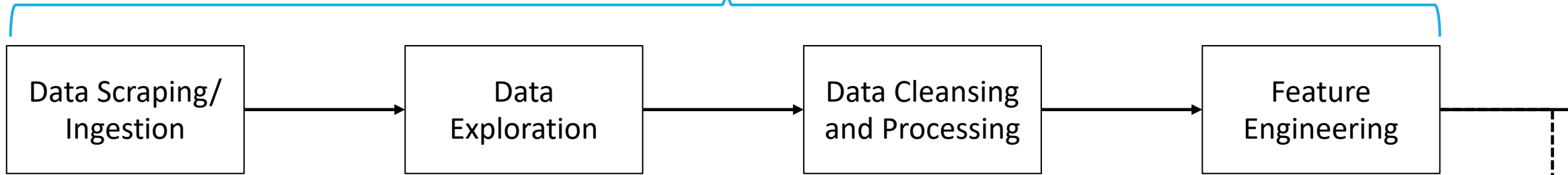
# Agenda

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



# Machine Learning Workflow

Data Engineer, Dev Ops Engineer



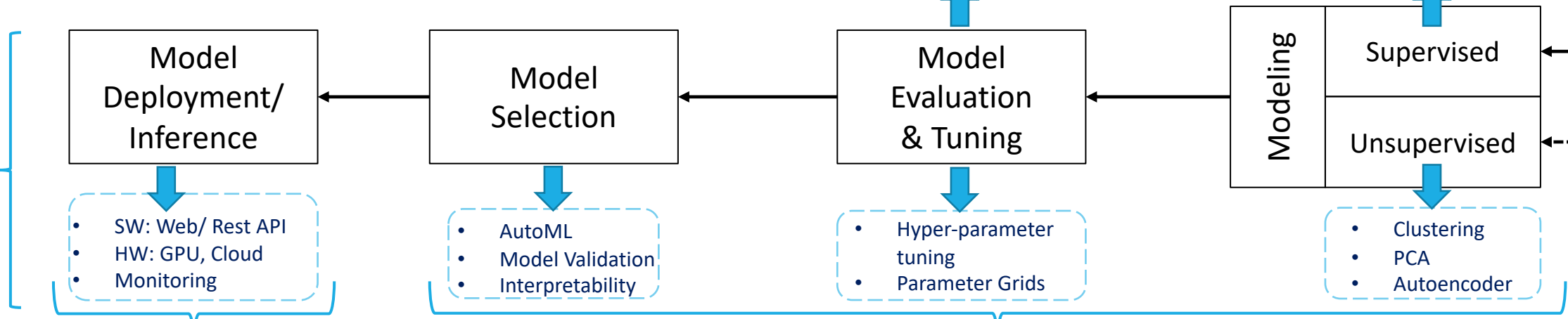
Robotic Process Automation (RPA) (Microservices, Pipelines)

 Risk Management/ Compliance(All stages)

- RMS
- MAPS
- MAE
- Confusion Matrix
- Precision/Recall
- ROC

- Regression
- KNN
- Decision Trees
- Naive Bayes
- Neural Networks
- Ensembles

Analysts & Decision Makers



Software/Web Engineer

Data Scientist/Quants

# Challenges

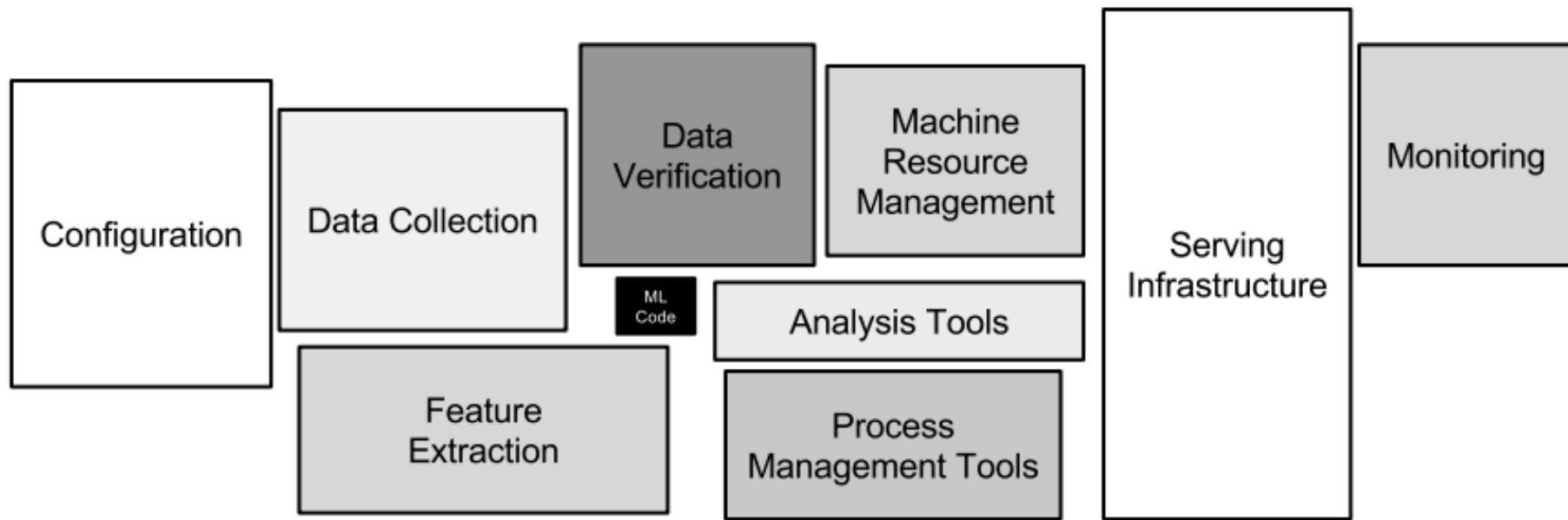
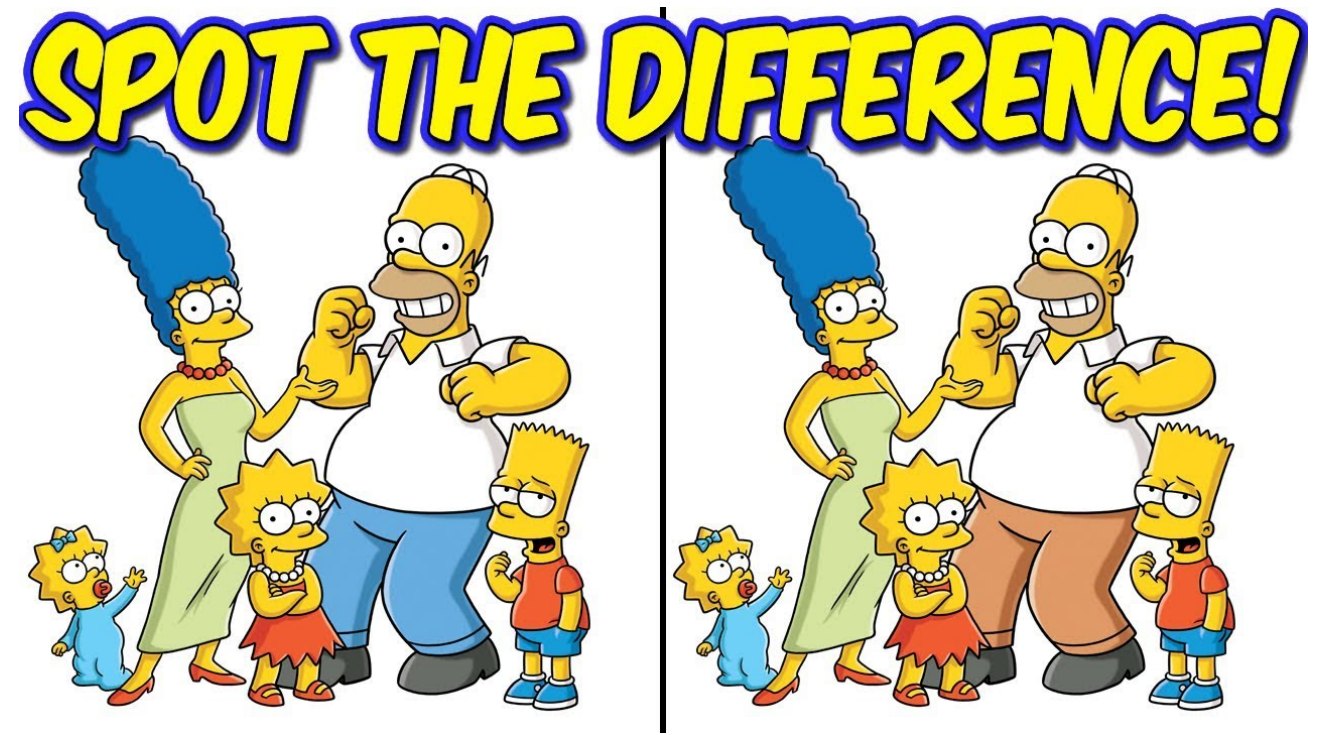
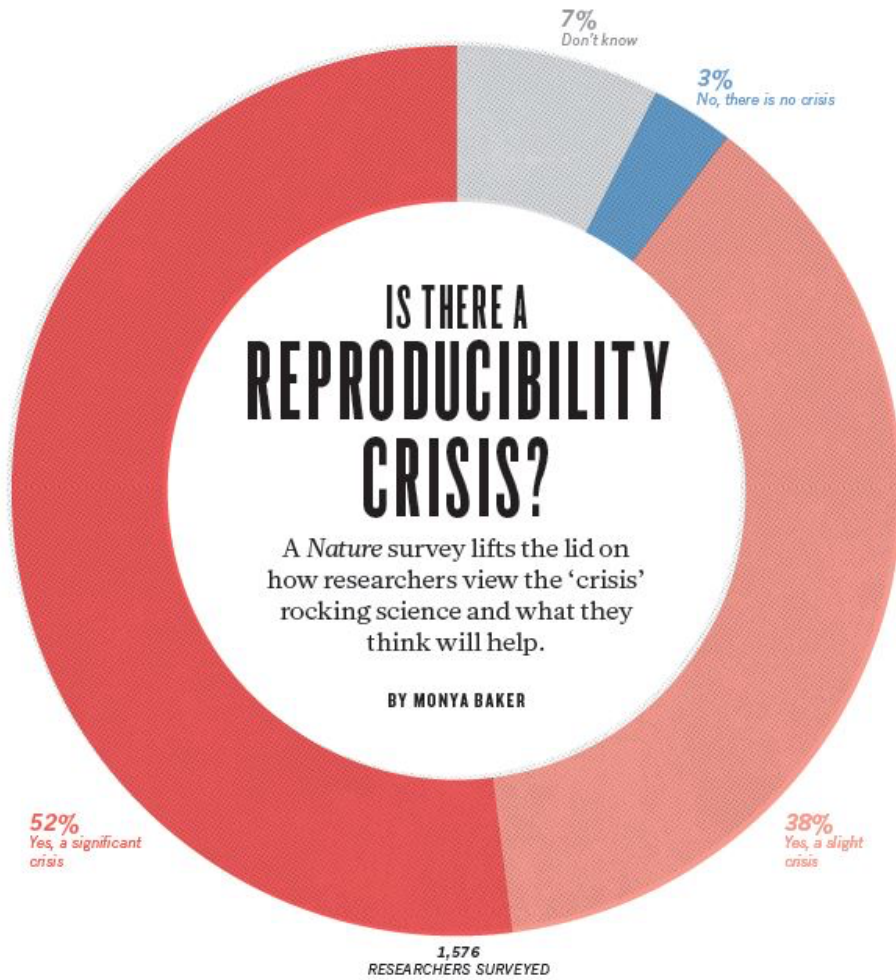


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

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.

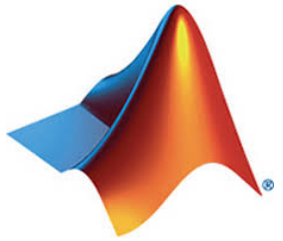
<https://www.acm.org/publications/policies/artifact-review-badging>





# Many choices

## Languages



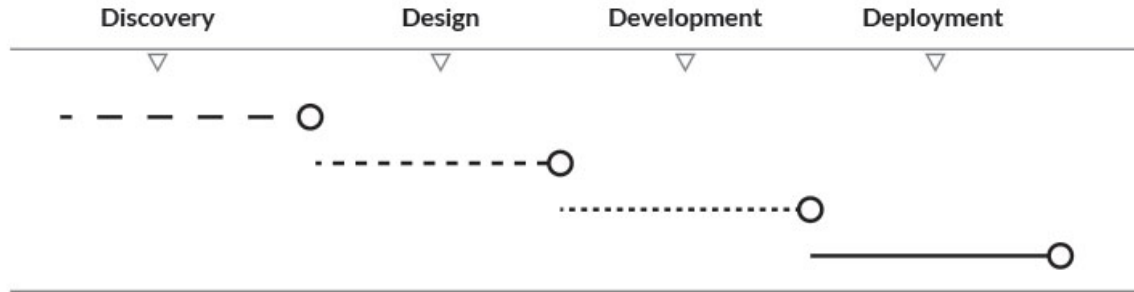
## Platforms



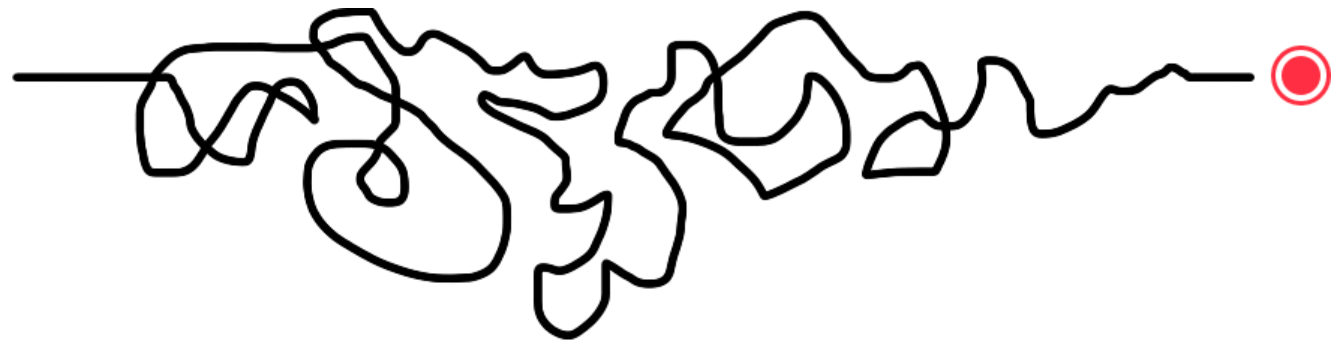
## Frameworks



# Processes are chaotic



Planning



Reality



## Multiple stakeholders

### Engineering/IT

- Scaling
- Structuring
- Design of Experiments
- Data Parallel/Task Parallel

### Quants/Data Scientists

- New Algorithms
- Try new methods
- Effect of Parameters and Hyper Parameters



## Which Model to choose ?

### Client Objective:

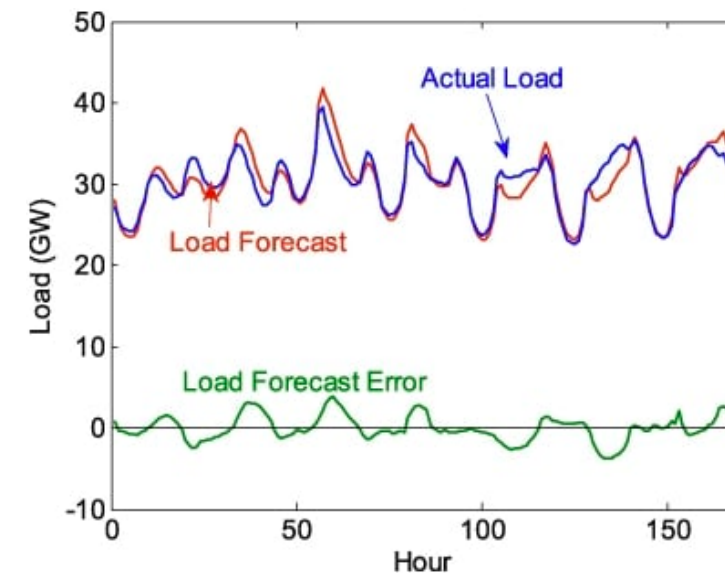
- Build the best forecasting model that has a MAPE of 5% or less

### 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"*



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

# 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

Data Ingestion  
from Edgar

Stage 2

Pre-Processing

Stage 3

Invoking APIs to  
label data

Stage 4

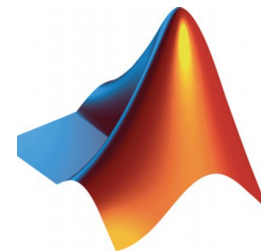
Compare APIs

Stage 5

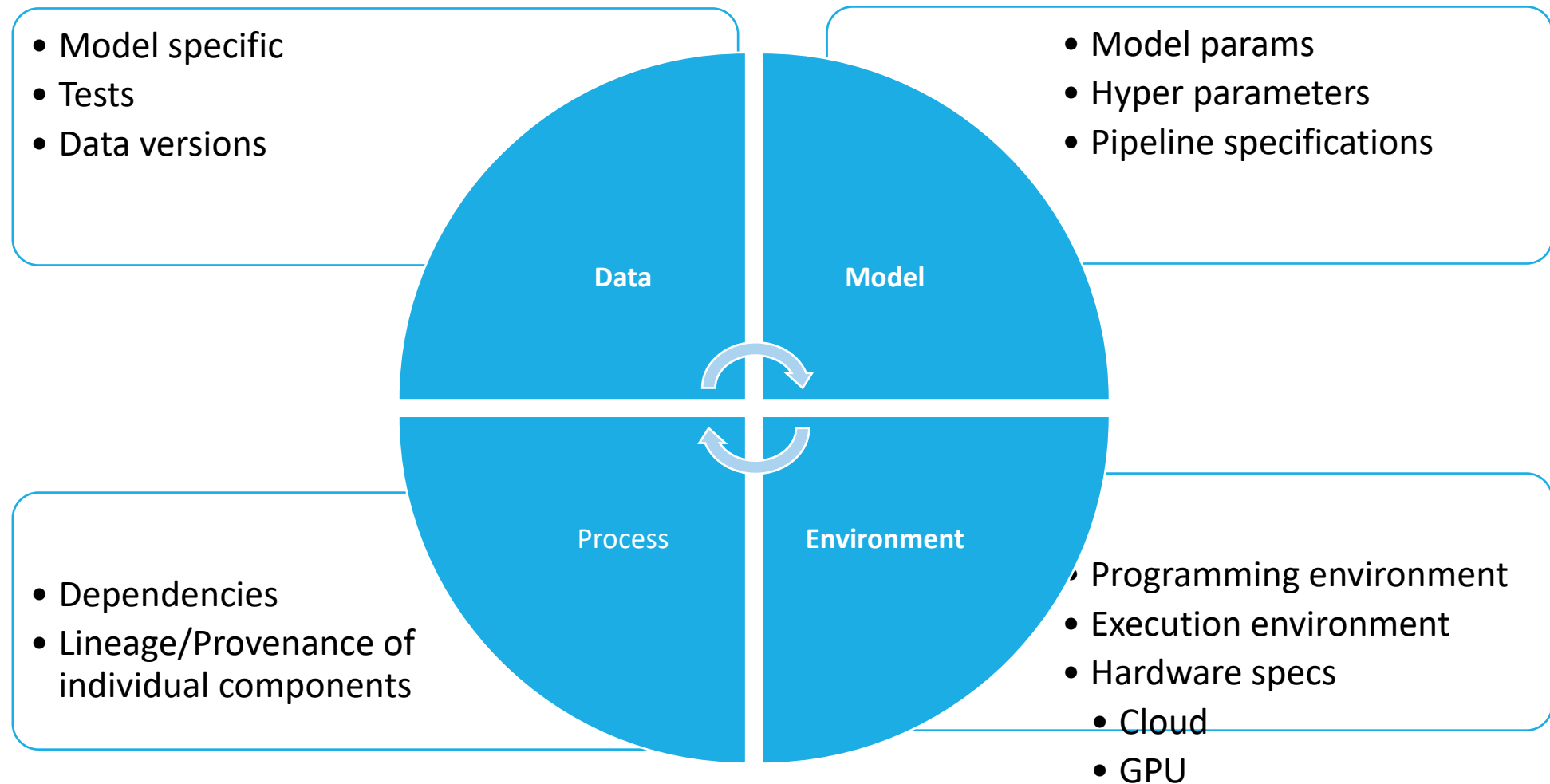
Build a new  
model for  
sentiment  
Analysis



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



# Components that needs to be tracked





# Provenance and Lineage of pipelines

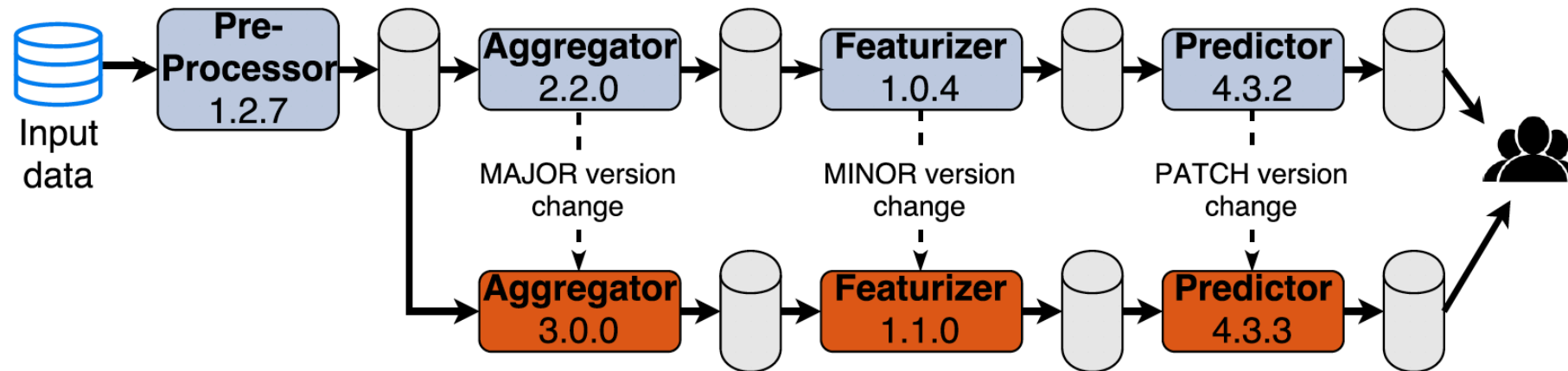


Figure 3: Running multiple pipeline versions

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

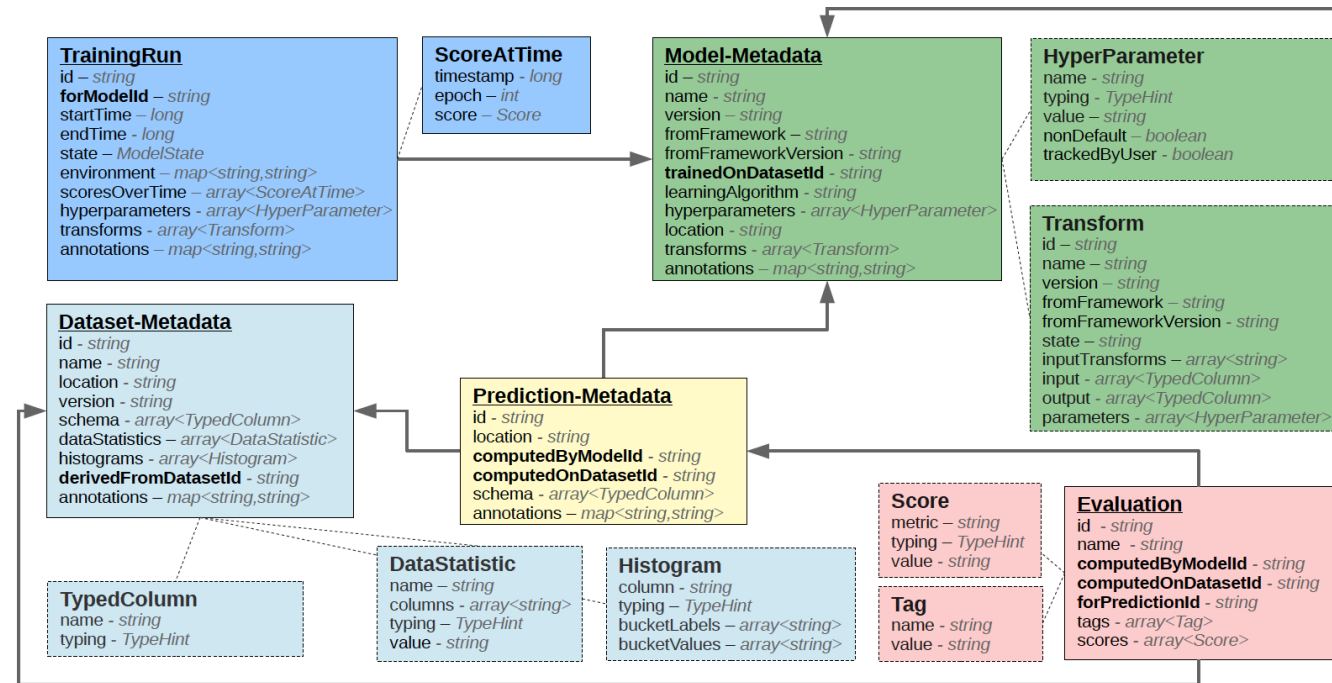
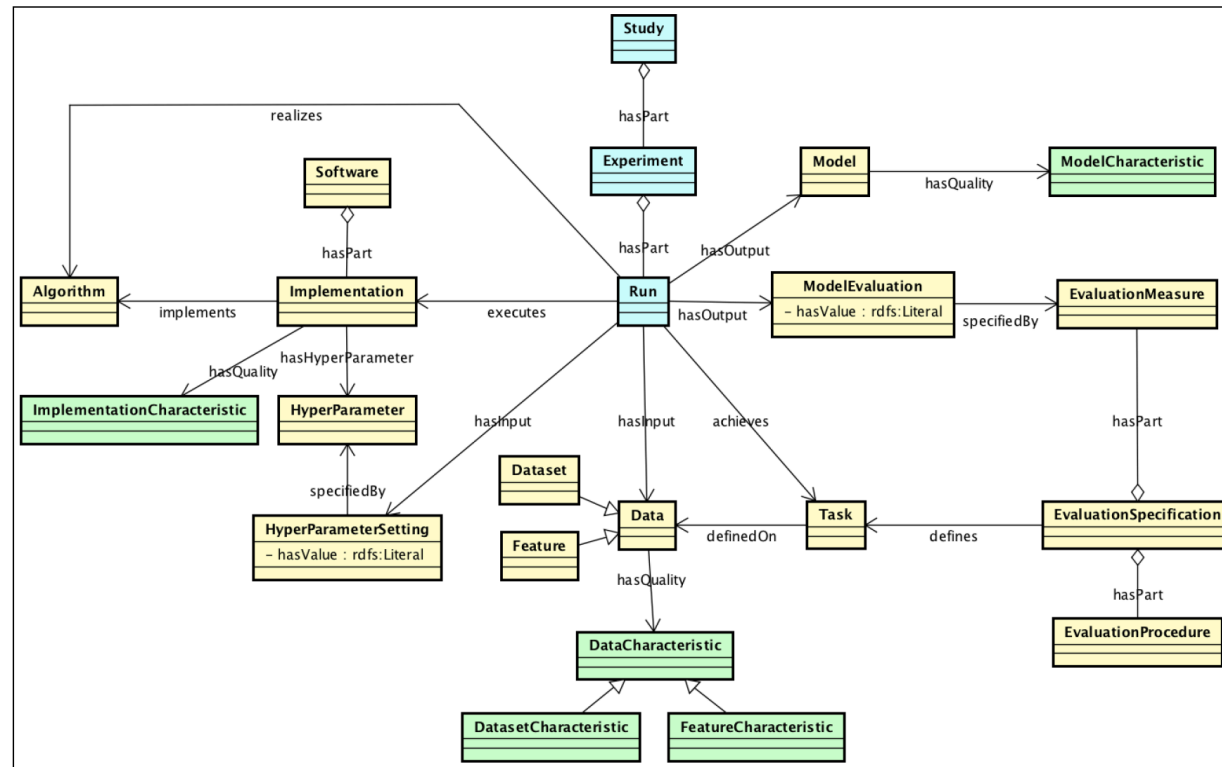


Figure 2: Simplified version of our schema used to store artifact metadata and lineage information. A detailed version available under Apache license can be found at <https://github.com/awsml/ml-experiments-schema>. Bold attributes indicate lineage relationships and every entity can be extended via arbitrary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



# 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.

# MLFlow



## Default > Run bf34330b7ebc4a07abe335b9a2a6ce2a ▾

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

Run ID: bf34330b7ebc4a07abe335b9a2a6ce2a

Source:  sklearn\_elasticnet\_wine

Git Commit: 60c71fae70aeff9841d60f63e023f129c02dec19

Entry Point: main

User: srimacpro

Duration: 1.6min

### 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

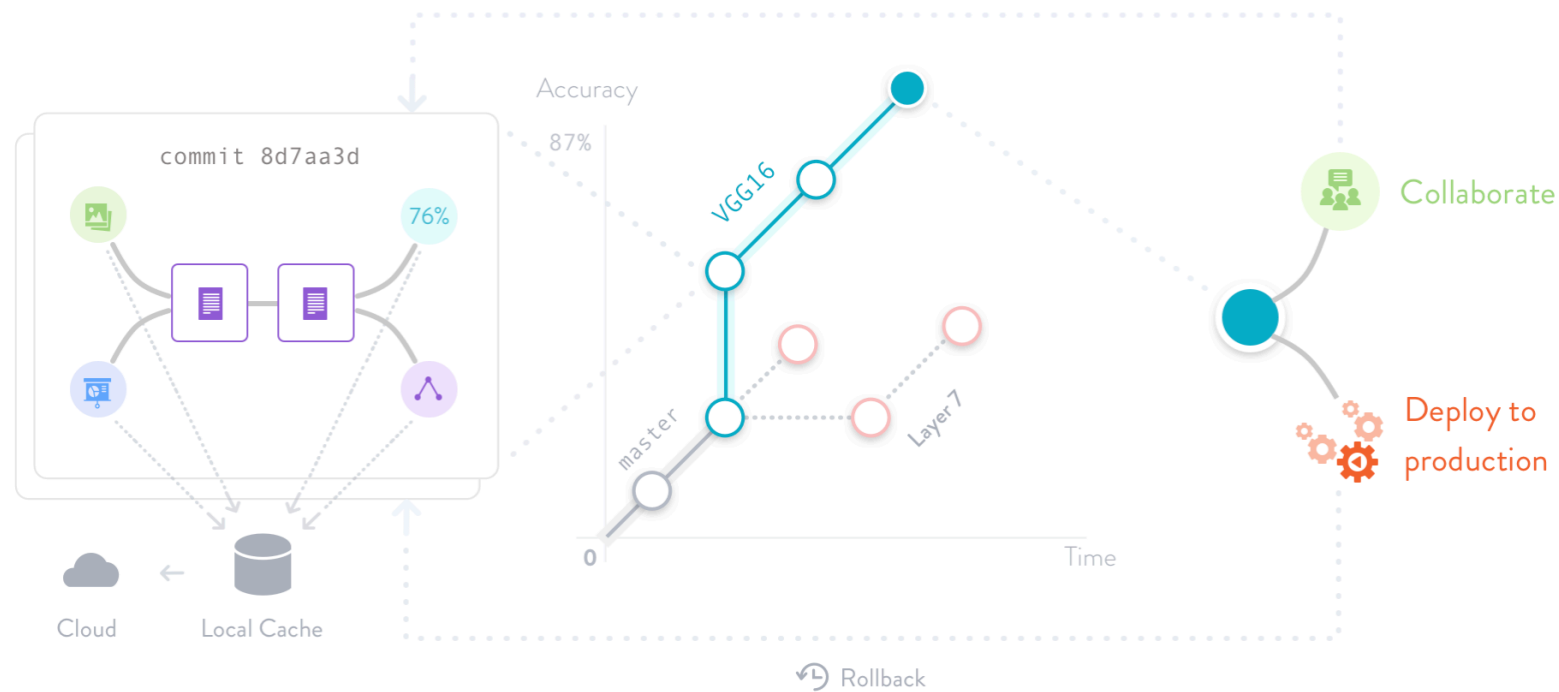
Name	Value
alpha	0.5
l1_ratio	0.1



# 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.

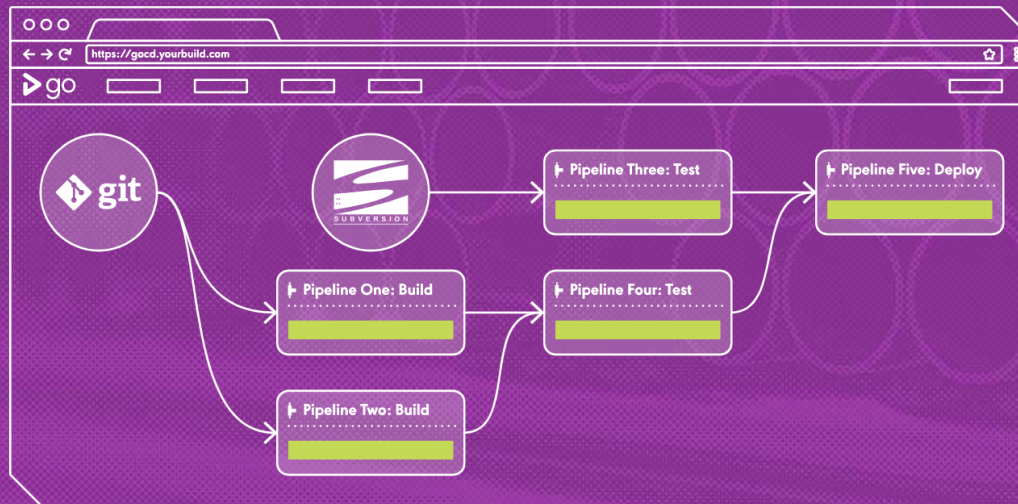


# GoCD



Features Documentation Blog Enterprise Download

## FREE & OPEN SOURCE CI/CD SERVER



Easily model and visualize complex workflows with GoCD.

TEST DRIVE GOCD

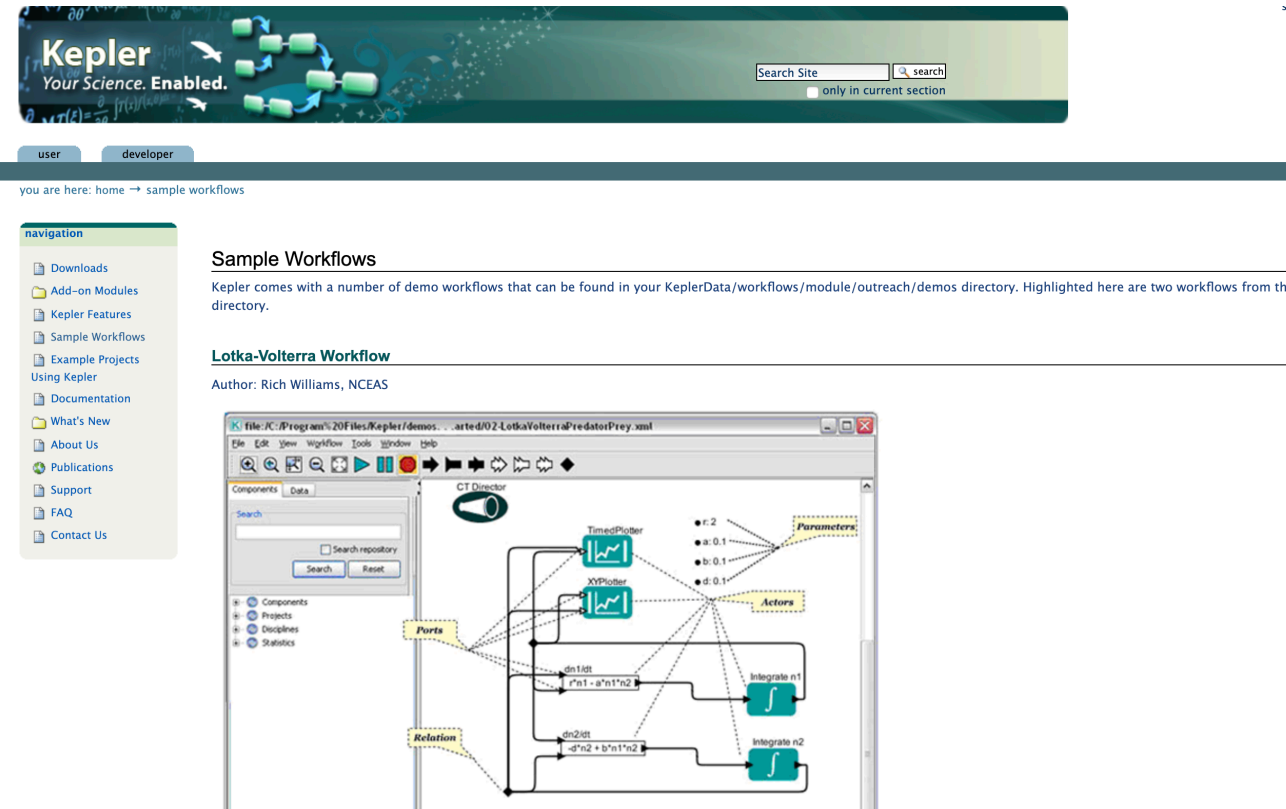
GoCD supports Pipelines as Code. [See the Benefits](#)



# Implementation Approaches



# Current approaches



The image shows a screenshot of the Kepler website and a sample workflow diagram. The website header features the Kepler logo with the tagline "Your Science. Enabled." and a search bar. Below the header, there are navigation tabs for "user" and "developer", and a breadcrumb trail "you are here: home → sample workflows". A navigation sidebar on the left lists various links such as "Downloads", "Add-on Modules", "Kepler Features", "Sample Workflows", "Example Projects Using Kepler", "Documentation", "What's New", "About Us", "Publications", "Support", "FAQ", and "Contact Us". The main content area is titled "Sample Workflows" and contains a paragraph explaining that Kepler includes demo workflows in the directory `KeplerData/workflows/module/outreach/demos`. Below this, the "Lotka-Volterra Workflow" is highlighted, with the author listed as "Rich Williams, NCEAS". The workflow diagram itself is a complex graph showing a "CT Director" component connected to several other components: "TimedPlotter", "XYPlotter", "Integrate n1", and "Integrate n2". It also includes mathematical expressions for differential equations:  $\frac{dn1}{dt} = r*n1 - a*n1*n2$  and  $\frac{dn2}{dt} = -d*n2 + b*n1*n2$ . The diagram is annotated with yellow callouts for "Ports", "Relation", "Parameters", and "Actors".

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

The screenshot displays the ProvDB Query Provenance interface. The top navigation bar includes 'ProvDB', 'Query Provenance', 'Pipeline & FileView', 'Monitoring Dashboard', 'Documents', and 'Server: Connected'.

**Artifacts Panel:** A list of artifacts is shown with columns 'id' and 'name'. The selected artifact is '231 model-9/logging.txt'. Below the list, it indicates 'Showing 1 to 41 of 41 entries (filtered from 1,006 total entries) 2 rows selected'.

**Query: Diff Artifacts: 'model-0/logging.txt', 'model-9/logging.txt'**

The diff result table shows a comparison between two versions of the 'caffe' artifact. The table has columns: 'ingestor', 'name', 'value1', and 'value2'.

ingestor	name	value1	value2
_posix	utility	caffe	caffe
caffe	iter	0	0
caffe	test-accuracy	0.1083	0.1307
caffe	test-loss	2.30259	2.30258
caffe	train-loss	2.3026	2.30259
caffe	iter	100	100
caffe	test-accuracy	0.2553	0.1194

Below the table, it indicates 'Showing 1 to 879 of 879 entries (filtered from 882 total entries)'.

**Plot:** A line graph showing performance metrics over iterations. The x-axis represents iterations (0 to 25000), and the y-axis represents accuracy and loss (0 to 2.4). The plot includes four series: 'test-accuracy@1' (cyan), 'test-accuracy@2' (pink), 'test-loss@1' (blue), and 'test-loss@2' (yellow). A callout box highlights a point at iteration 7300 with 'test-loss@2: 1.2366'.

**Query: MATCH p=(x:Artifact)-->()<--(va:Version)-->(vb:Version)-->()<--(xb:Artifact) ...**

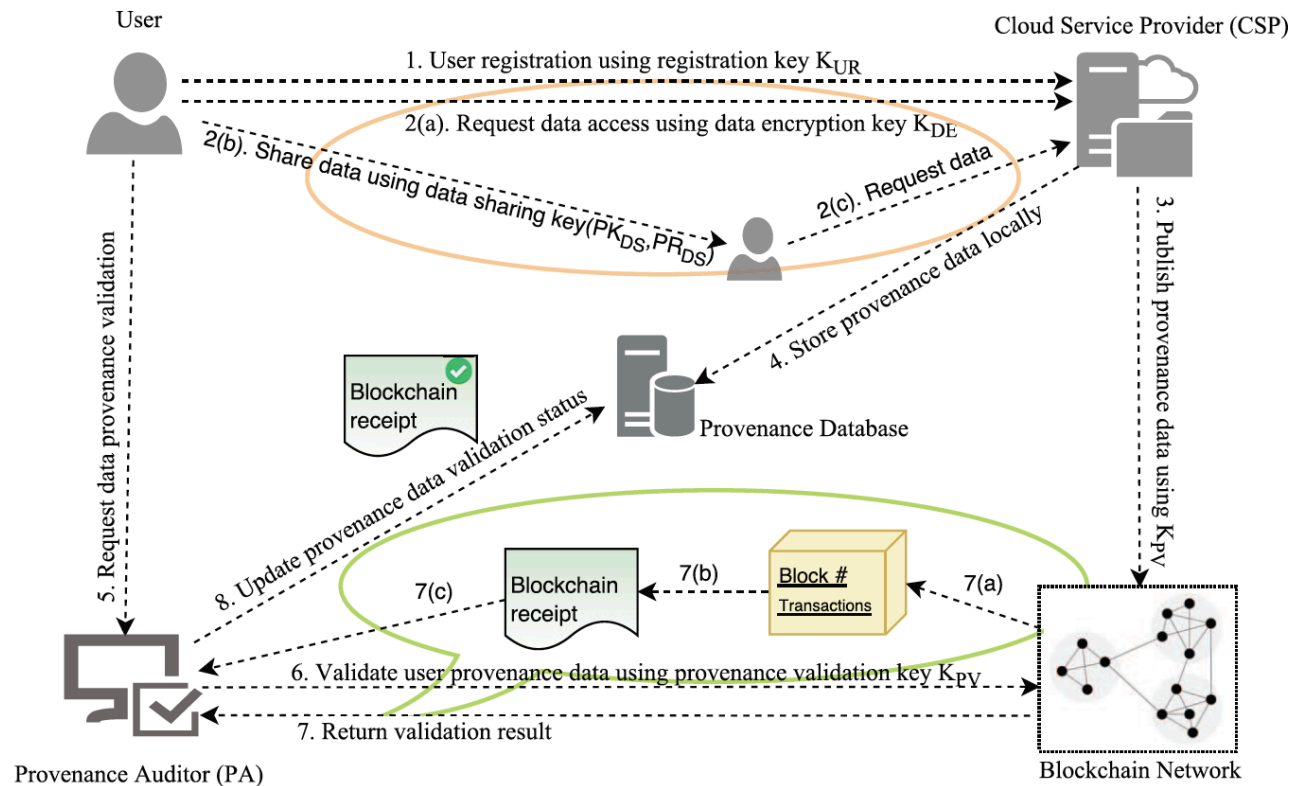
The query results table shows the following artifacts:

id	artifacts
2	solver.prototxt
3	train_test.prototxt
4	model-0/solver.prototxt

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



# Related work



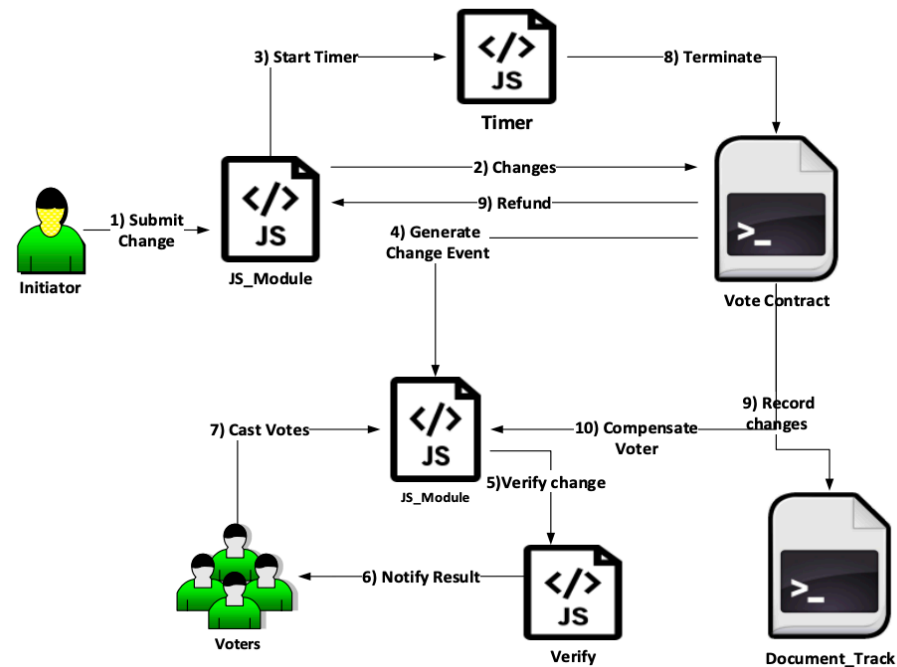
Focus on Cloud data  
provenance using  
Blockchain

Figure 1: ProvChain System Interaction.

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



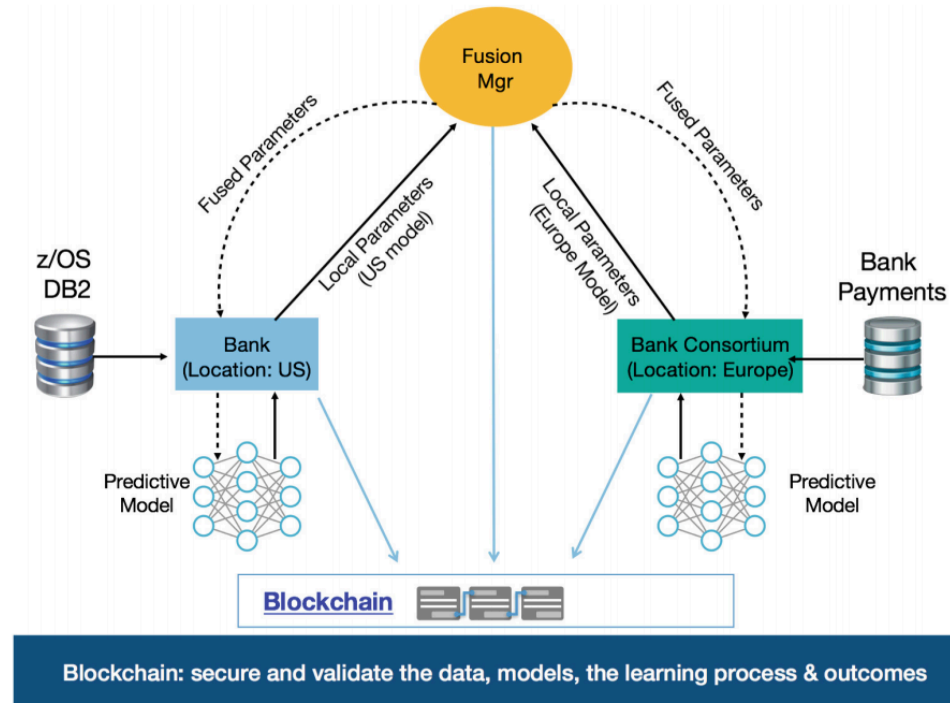
**Figure 2: Voting procedure for a Document change.**

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



Trusted AI and  
provenance of AI models

Fig. 2. Trusted federated learning: the basic setting

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



## CORPORATE PARTICIPANTS

**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

## CONFERENCE CALL PARTICIPANTS

**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

## Options

- 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

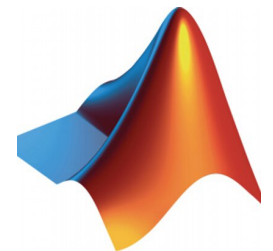
Compare APIs

Stage 5

Build a new  
model for  
sentiment  
Analysis



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





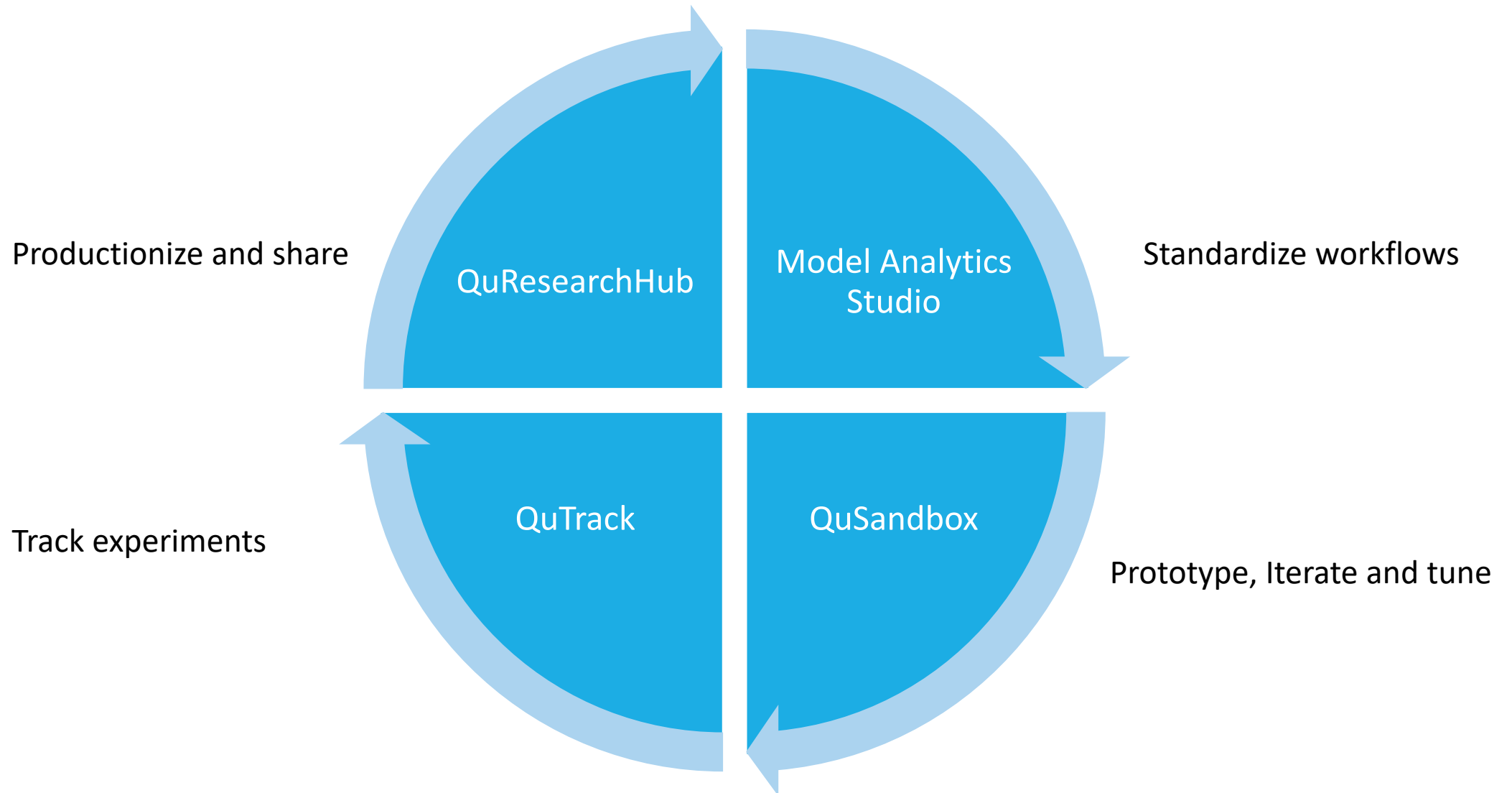


QuantUniversity, LLC

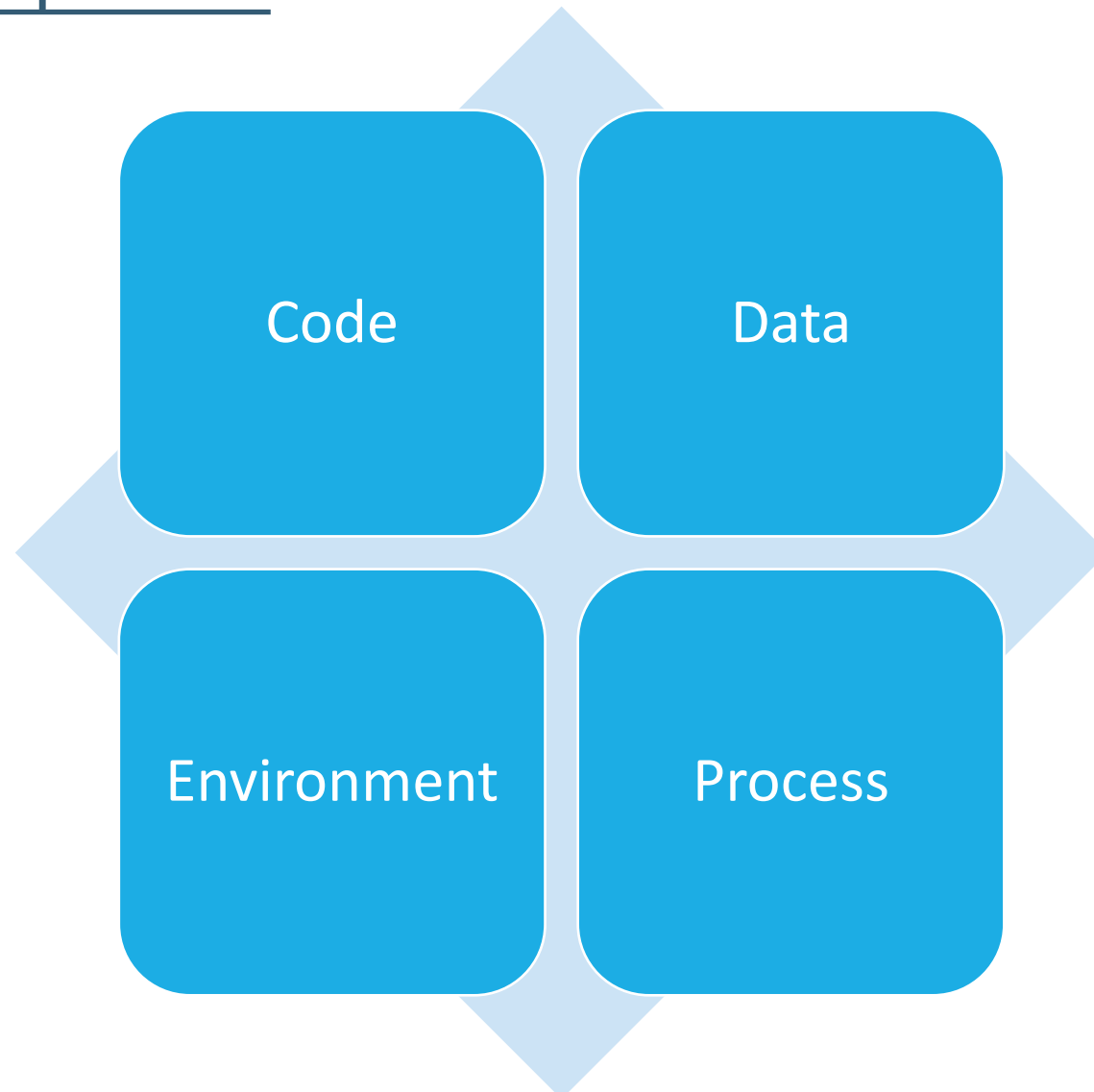
[www.quantuniversity.com](http://www.quantuniversity.com)

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


# QuSandbox research suite

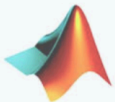


## The four components that need to be encapsulated for reproducible pipelines



# QuSandbox


QU Sandbox
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 data from Twitter. Users can see an exploratory data analysis on the tweets shared all across 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.


QU Sandbox
Home Running Instances Available Projects Task Scratch Pad MarketPlace



## TensorFlow

### hands-on-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) \*



## Build Docker Image

[← List All Projects](#)
 Notebook
  Scripts
  Notebook & Scripts

Check the box to enable the Terminal launch for this project

 Enable Terminal?

**i** This option enables the terminal for the QUSandbox project created for this image

Project Name

Image Name

Module Name

Please select the course to which you want to add this docker image

List of Courses

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 +



# Model Management Studio

QuSandbox Model Management Studio

FORMS PIPELINES ▾ ENTITIES BLOCKS LOGIN REGISTER

QuSandbox-Edgar-Pipeline

QuSandbox-Phase-1-Env  
QuSandbox-Model-Phase-1  
BLOCK-PHASE-1-QUSANDBOX  
ADD ENTITY ⓘ

QuSandbox-Phase-2-Env  
QuSandbox-Model-Phase-2  
BLOCK-QUSANDBOX-2  
ADD ENTITY ⓘ

QuSandbox-Phase-3-Env  
QuSandbox-Model-Phase-3  
BLOCK-QUSANDBOX-3  
ADD ENTITY ⓘ

Notifications

Stage 1 +

Stage 2 +

Stage 3 + +

# 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":"idadaptivealgo/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":"idadaptivealgo/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":"idadaptivealgo/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

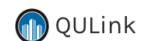
Back to the Projects

## 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



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 :  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](http://www.quantuniversity.com)

Sign up for Updates at

[www.qusandbox.com](http://www.qusandbox.com)

LinkedIn  [srikrishnamurthy](#)

[www.QuantUniversity.com](http://www.QuantUniversity.com)

[www.analyticscertificate.com](http://www.analyticscertificate.com)