VANDERBILT UNIVERSITY

STRATUM: A Serverless Framework for the Lifecycle Management of Machine Learning-based Data Analytics Tasks

- ***Anirban Bhattacharjee***
- Yogesh Barve
- Shweta Khare
- Shunxing Bao
- Aniruddha Gokhale
- Thomas Damiano

- 2019 USENIX Conference on Operational Machine Learning
- MAY 20, 2019

Acknowledgments

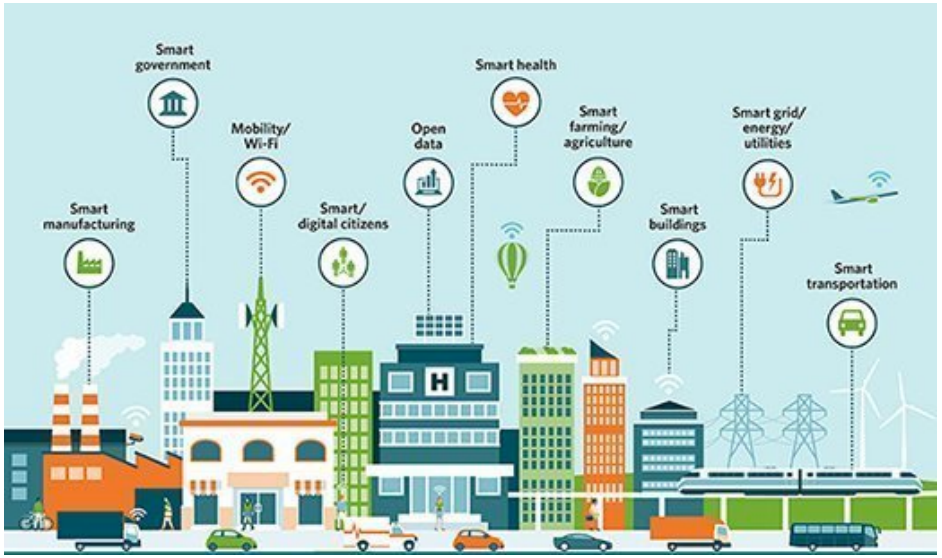
This work was supported in part by NSF US Ignite CNS 1531079, AFOSR DDDAS FA9550-18-1-0126 and AFRL/Lockheed Martin StreamlinedML program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF, AFOSR or AFRL.

LOCKHEED MARTIN

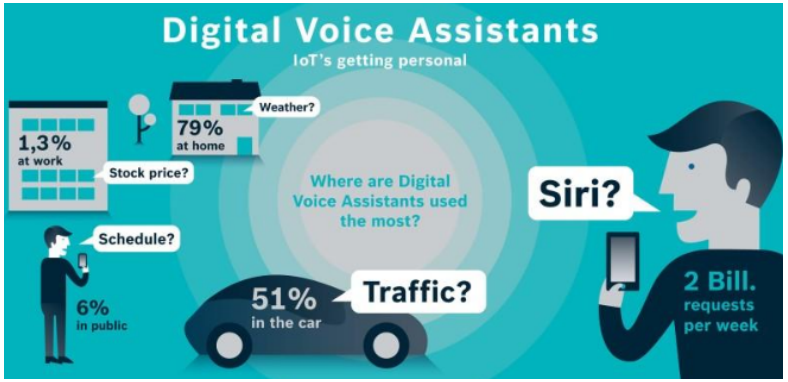
Data Analytics Trends

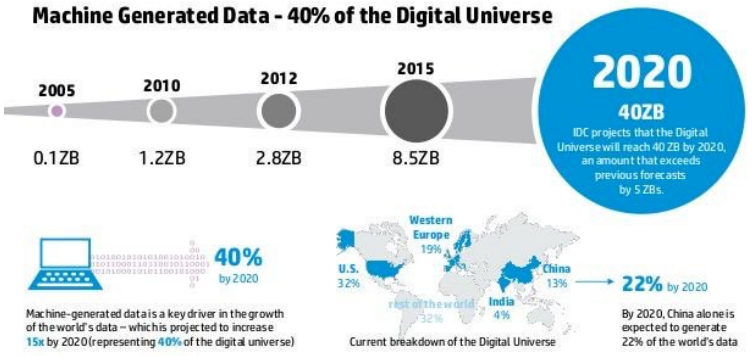
The world is changing and accelerating

Internet of Things (IoT) applications, such as cognitive assistance, voice assistance, patient health monitoring, and connected vehicles are increasingly using the cloud and edge analytics.

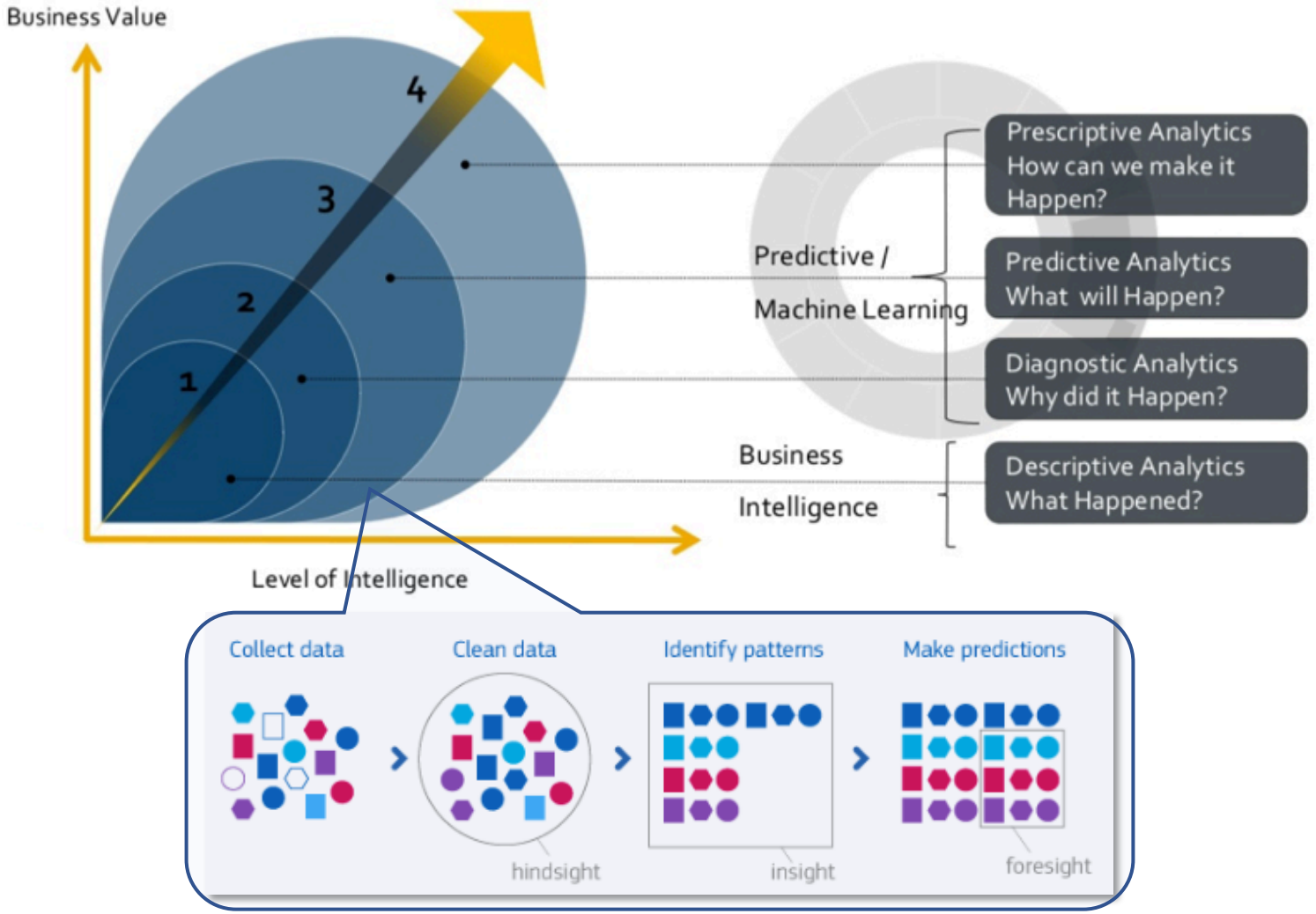


Smart IoT devices generates data in volume and velocity, which needs to be analyzed to get valuable insight.

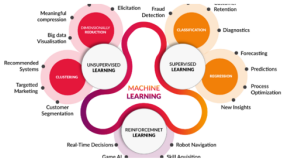




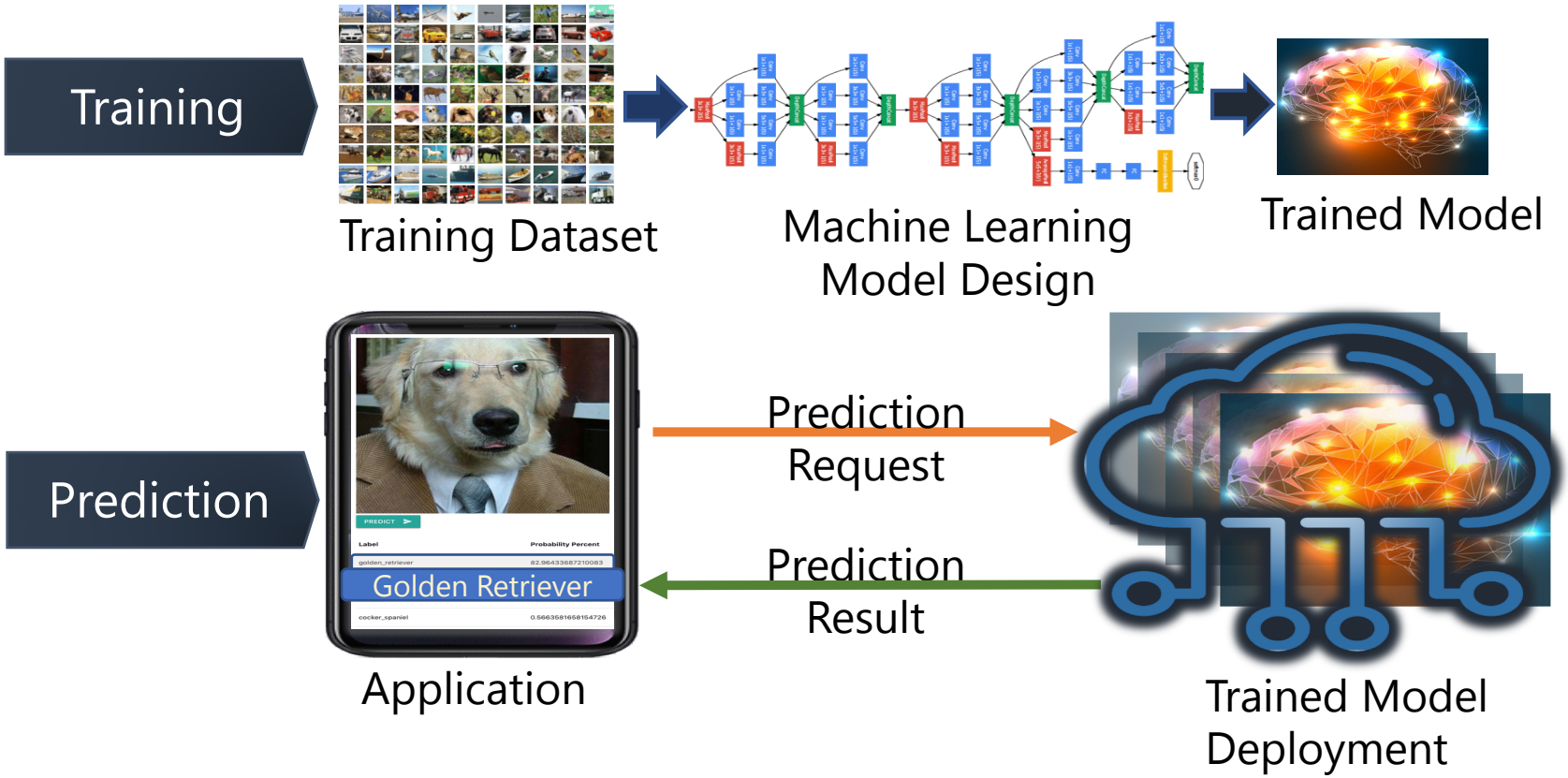
Big Data Value Model



ML-based Predictive Analytics

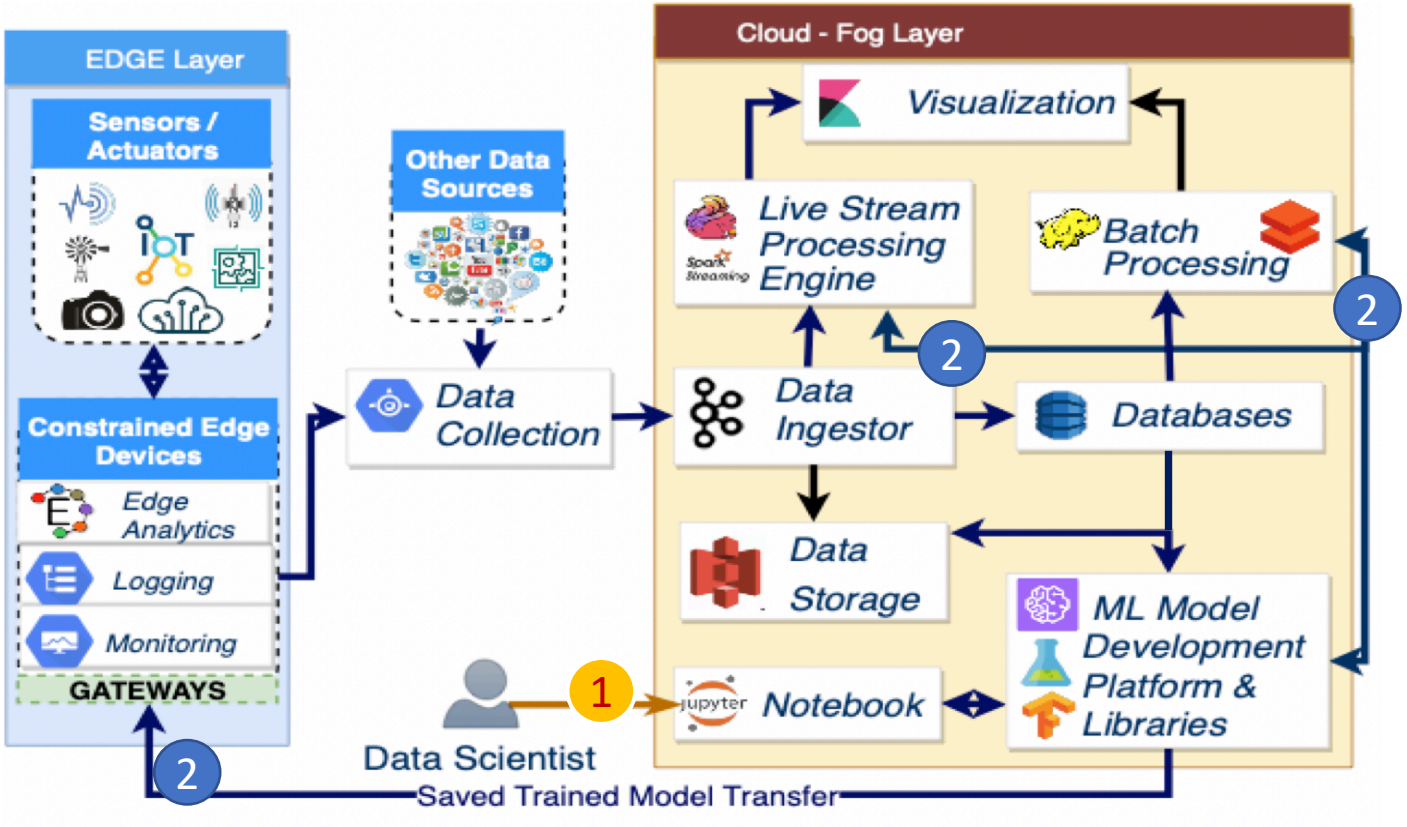


- **Predictive analytics**
  - It uses various statistical modelling and machine learning techniques to analyze past data and predict the future outcomes.



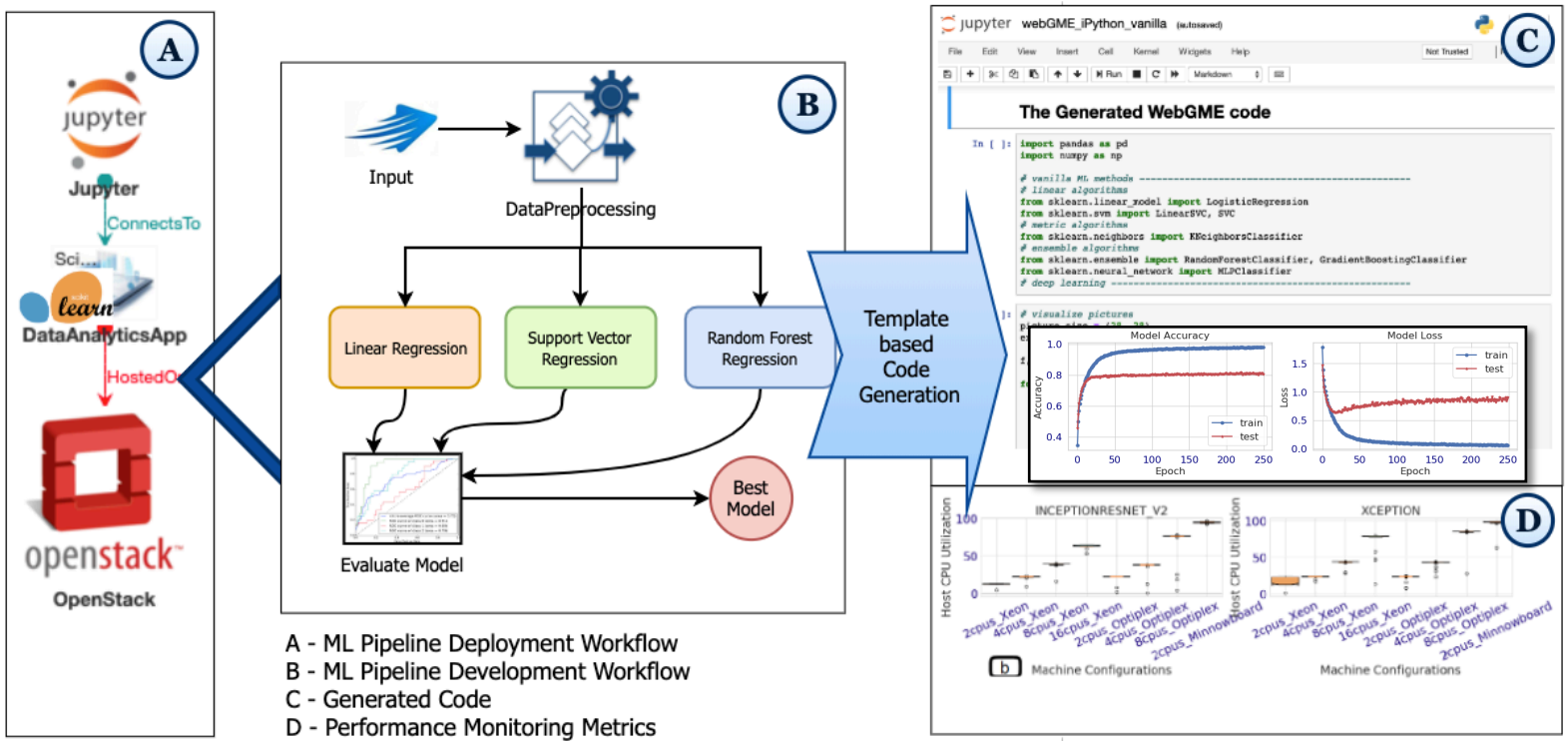
ML-based Predictive Analytics

- **Predictive analytics**
  - It uses various statistical modelling and machine learning techniques to analyze past data and predict the future outcomes.



Requirement

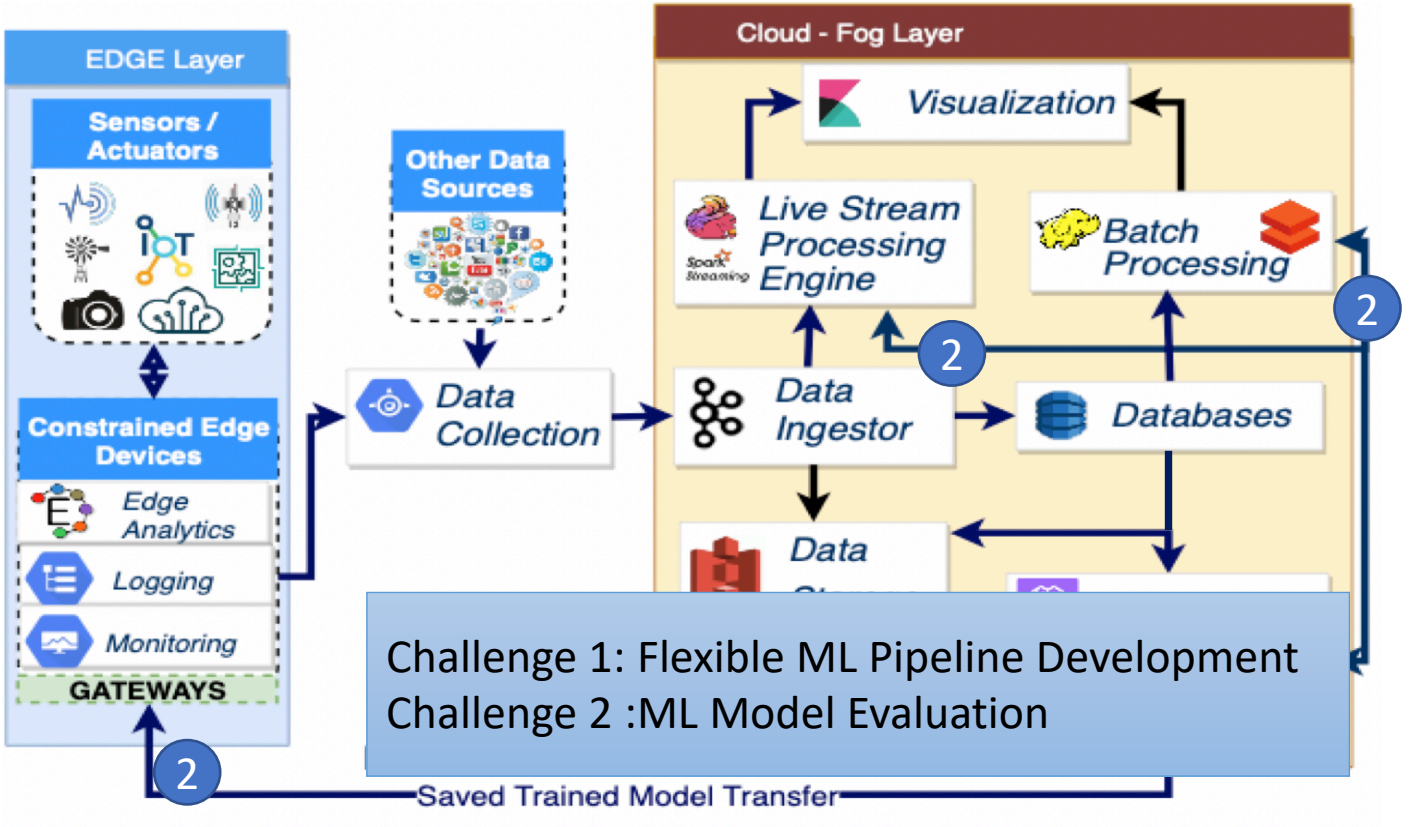
- Automation of Machine Learning (ML) Model Development and Deployment
  - Alleviate ML developers from writing the code from scratch
  - ML Library and Framework Agnostic



A - ML Pipeline Deployment Workflow
B - ML Pipeline Development Workflow
C - Generated Code
D - Performance Monitoring Metrics

ML-based Predictive Analytics

- **Predictive analytics**
  - It uses various statistical modelling and machine learning techniques to analyze past data and predict the future outcomes.



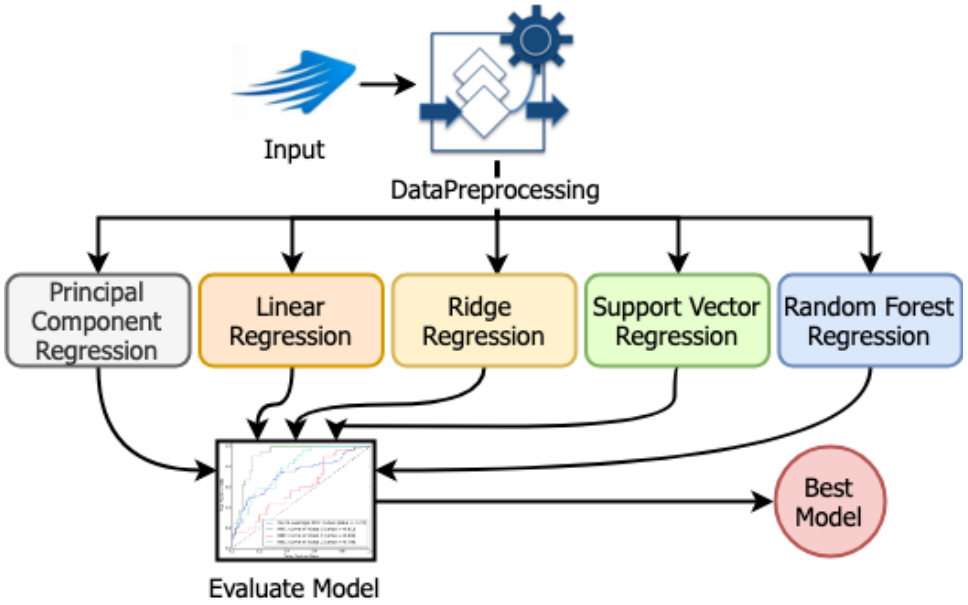
Challenges[1/3]

Flexible ML Pipeline Development

- A diverse set of ML algorithms -
  - classification (logistic regression, naive Bayes), regression, decision trees, random forests, and gradient-boosted trees, recommendation (ALS), clustering (K-means, GMMs), and many others.
- A diverse set of different ML libraries and frameworks –
  - Scikit-learn, Spark MLlib, TensorFlow etc.
- ML pipeline capabilities needs to be captured, and abstracted in the metamodel.
  - Attributes of ML algorithms, data preprocessing strategies, evaluation methods etc.

Challenges[2/3]

ML Model Evaluation

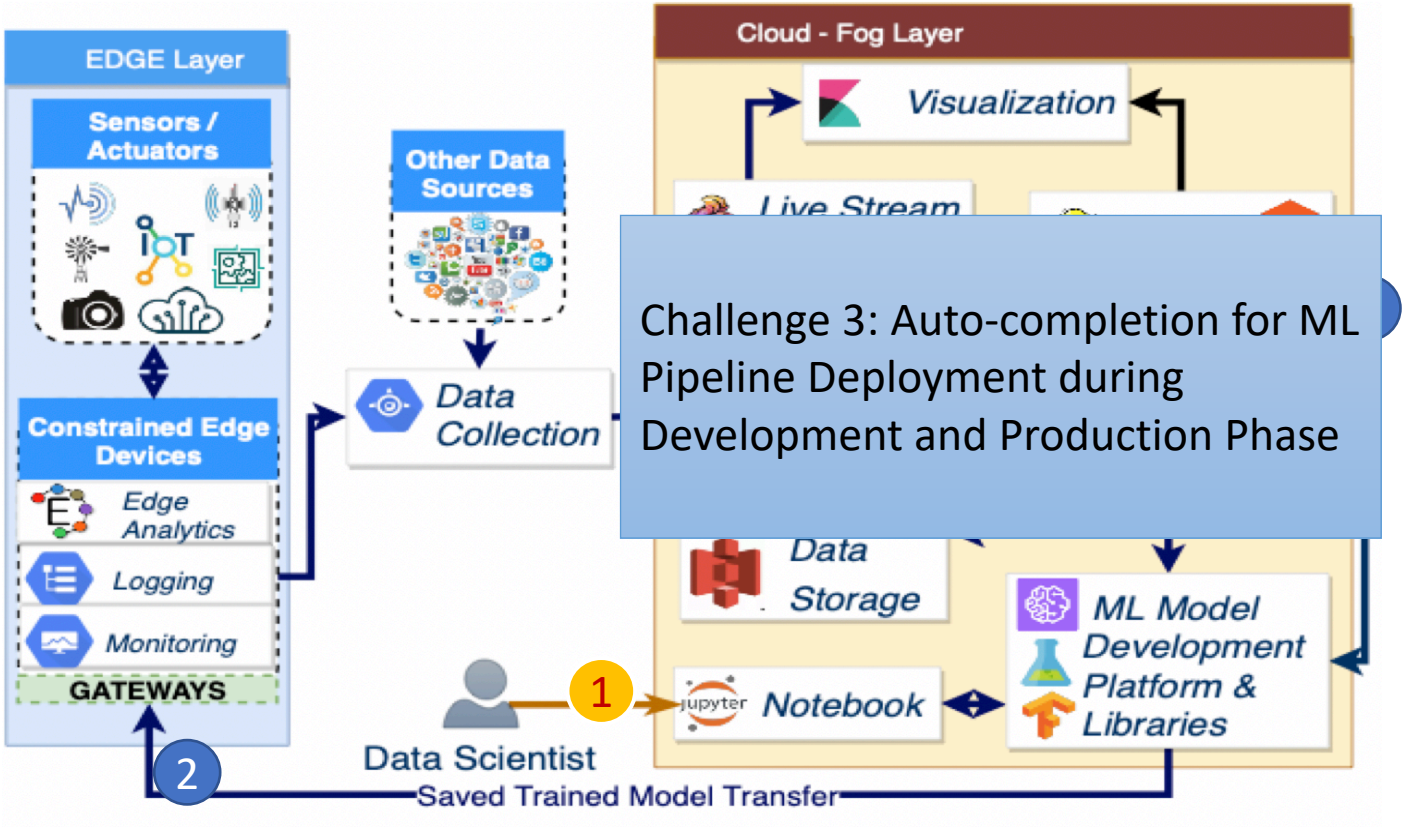


- After training the ML models with the diverse set of ML algorithms, the best model for the dataset needs to selected.
  - save it for prediction jobs.
- The model can be evaluated based on different scoring methods such accuracy, f1 score, precision, r2 score, mean square error which is captured in metamodel.
- To speedup the training process, the ML models with different algorithms can be distributed.

ML Trained Model with all the software dependencies needs to be encapsulated in a container on the specific hardware.

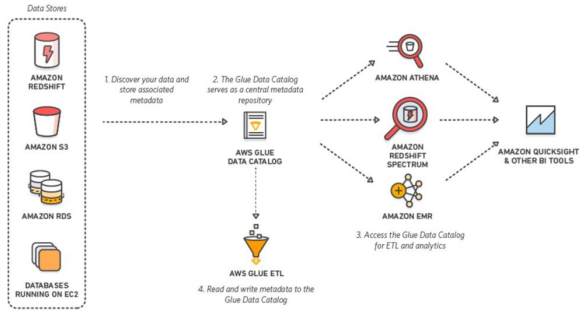
ML-based Predictive Analytics

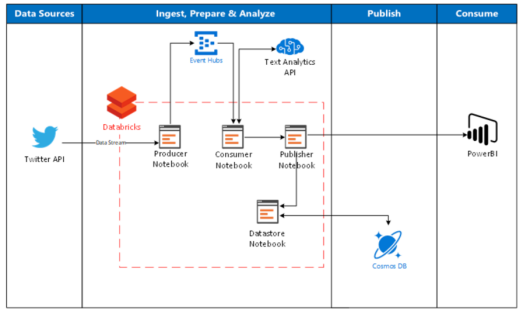
- **Predictive analytics**
  - It uses various statistical modelling and machine learning techniques to analyze past data and predict the future outcomes.



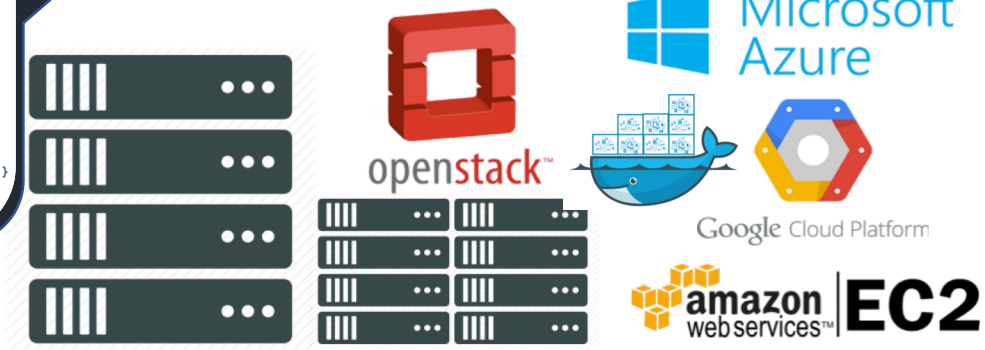
Challenges[3/3]

Auto-completion for ML Pipeline Deployment



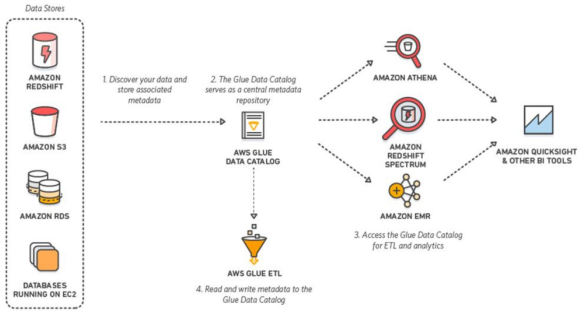


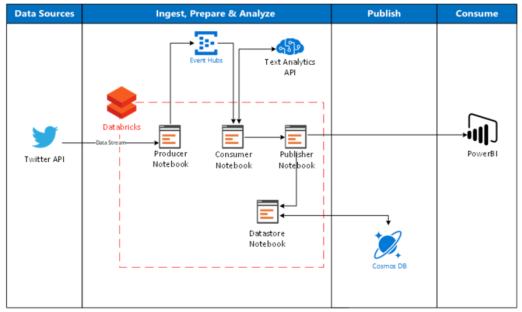

```yaml
---
- name: Install PHP-FPM
  apt:
    name: "{{ item }}"
    state: latest
    update_cache: yes
  with_items: "{{ ubuntu_php_pkgs }}"

- name: Copy the templates to their respestive destination
  template:
    src: "{{ item.src }}"
    dest: "{{ item.dest }}"
    owner: root
    group: root
  with_items:
    - { src: 'www.conf.j2', dest: '/etc/php/7.0/fpm/pool.d/www.conf' }
    - { src: 'php.ini.j2', dest: '/etc/php/7.0/fpm/php.ini' }
  notify:
     - restart apache2
```



Challenges[3/3]

Auto-completion for ML Pipeline Deployment






```
- name: Install PHP-FPM
  apt:
    name: "{{ item }}"
    state: latest
    update_cache: yes
  with_items: "{{ ubuntu_php_pkgs }}"

- name: Copy the templates to their respestive destination
  templa
    src:
    dest
    owne
    grou
  with_i
    - {
    - {
  notify
      -
```

Service Providers Concerns

How to deploy and maintain the application components with ease to increase productivity and usability while reducing the time-to-market?

STRATUM: ML Pipeline Automation

FRAMEWORK DESIGN

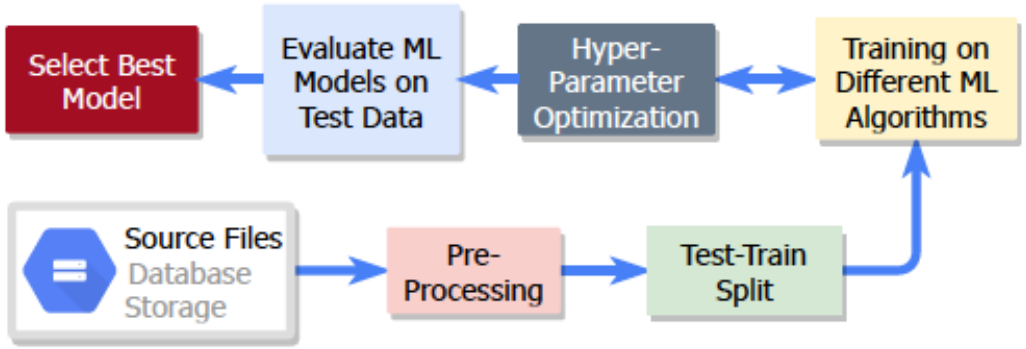
EVALUATION RESULTS

STRATUM: ML Pipeline Automation

FRAMEWORK DESIGN

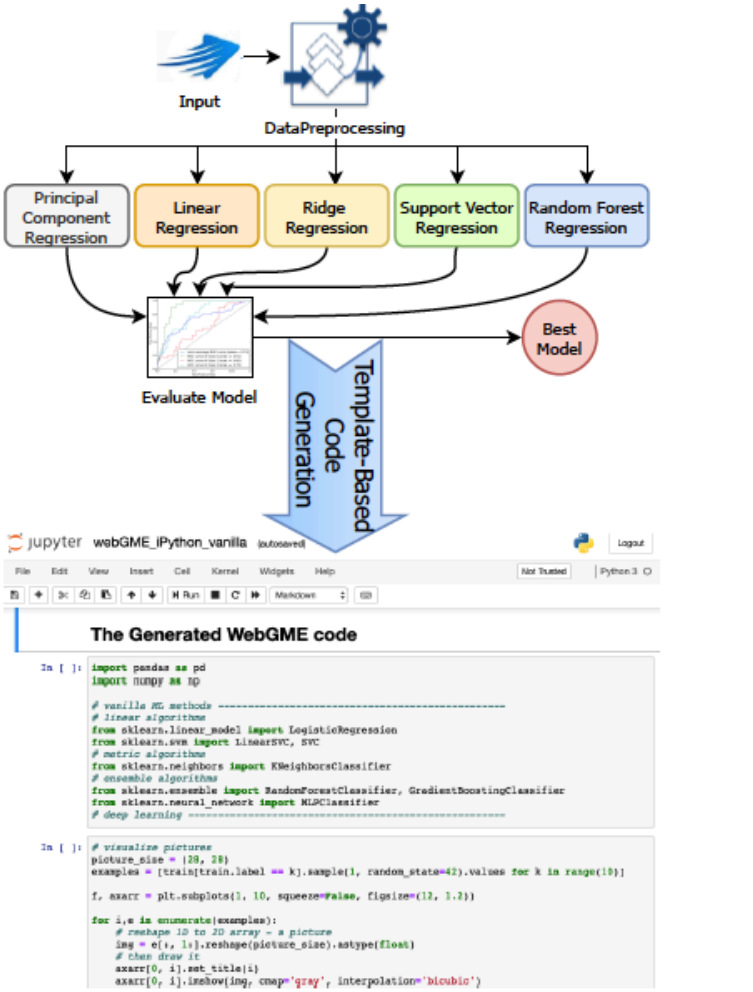
EVALUATION RESULTS

ML Model Evaluation



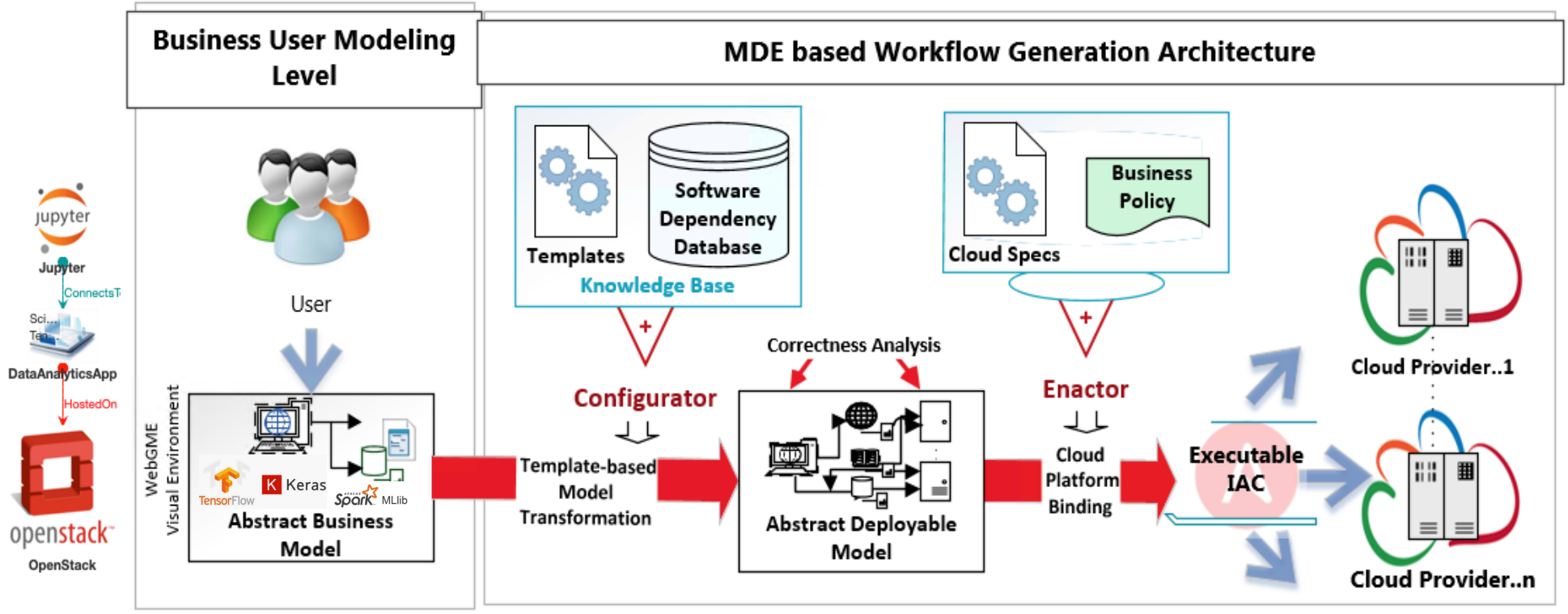
Sample Machine Learning Pipeline

- The framework provide the GUI for ML pipeline construction, model evaluation, and hyper-parameter tuning capabilities, which forms the basis for continuous evaluation.
- We also integrated Jupyter Notebook (notebook-based environment) to provide data-scientists the ability to train their models interactively.
- The ML execution pipeline needs to be bind with a specific library or framework such as Scikit-learn or TensorFlow.



Sample generative capabilities of Stratum

DSML for STRATUM Deployment Framework



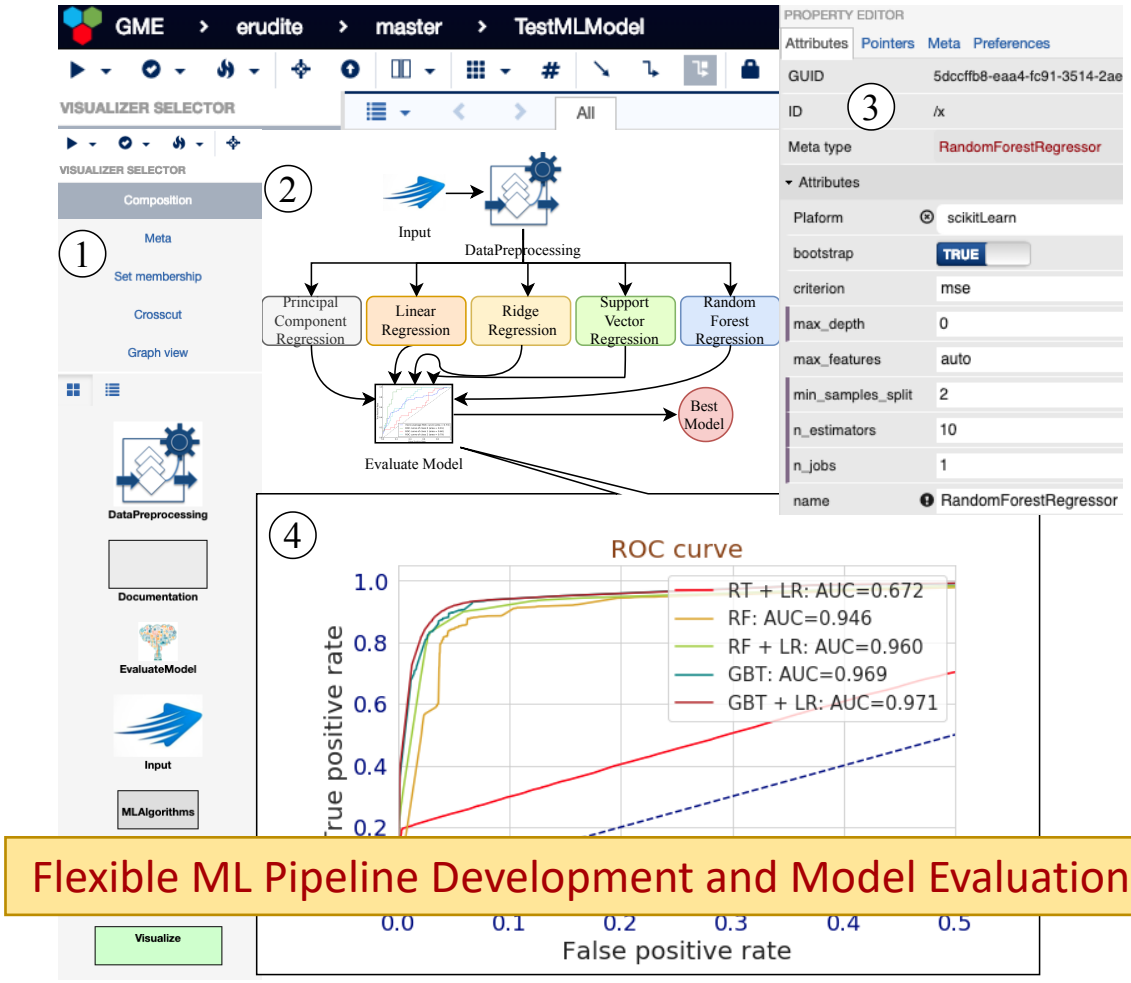
- ***Abstraction of Model:*** Deployers provision by selecting only business-relevant components.
- ***Configurator:*** Transforms abstract service components to Ansible-specific automation tasks using DSML.
- ***Enactor:*** Generates IAC by integrating automation code, cloud-specs & inter-component connection types.
- ***Knowledge Base:*** Software dependencies for service component types are stored in RDBMS table.

STRATUM: ML Pipeline Automation

FRAMEWORK DESIGN

EVALUATION RESULTS

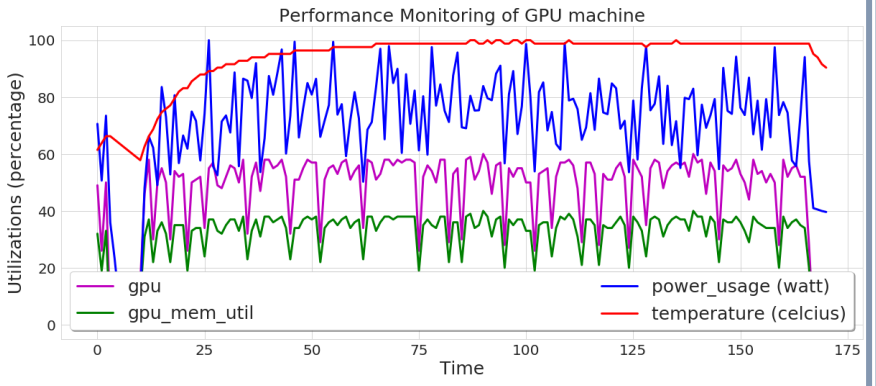
Usability of STRATUM Framework



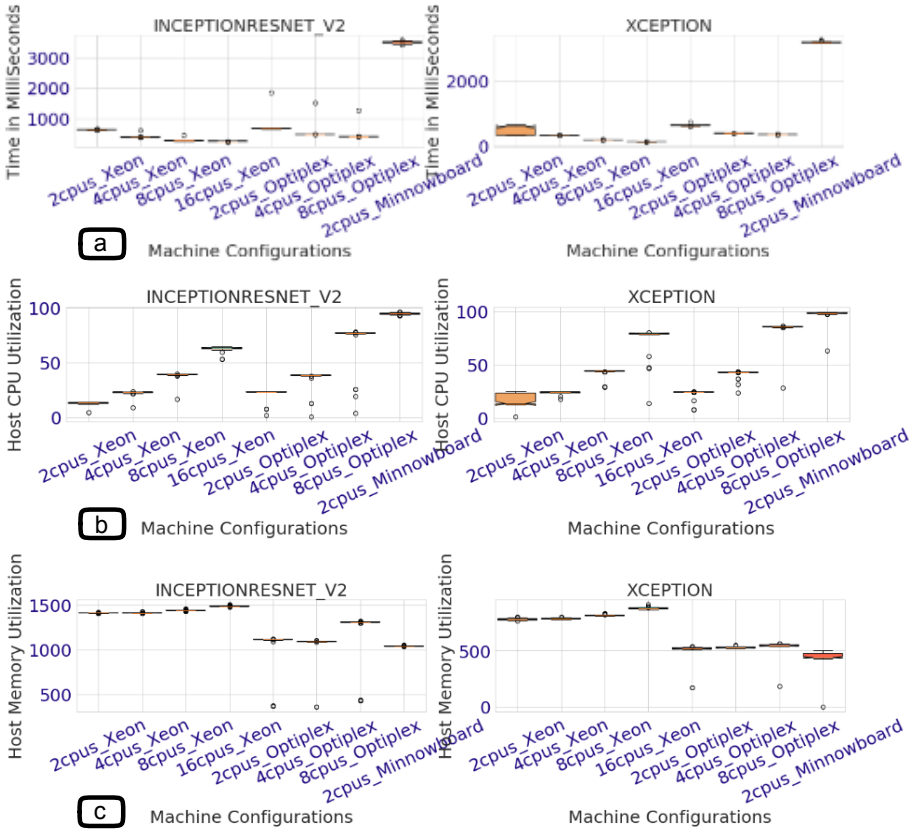
Flexible ML Pipeline Development and Model Evaluation

- The ML model can be developed by dragging and dropping the build blocks (from box 1→2).
- All the attributes of the selected ML algorithms such as max_depth, criteria need to be specified by the user (box 3).
- The Erudite model transformer can distribute different jobs with different ML techniques over a cluster of connected machines.
- It aids the developer to select the best model or ensemble of models based on the user's choice of evaluation methods.

Performance Monitoring



Sample Deep Learning Training on GPU Machine



Performance Monitoring of the prediction services (a)The execution latency of InceptionResnetV2 and Xception model on different ML containers with variable configurations, (b) Host CPU utilization of the ML containers (c) Host Memory utilization of ML containers (in MB).

Summary

- We presented a model-driven engineering and generative programming approach for automated development of ML pipeline.
- We integrated a monitoring framework to analyze the performance of ML pipeline during training and prediction phase.
- We proposed a ML pipeline deployment methodology across cloud-fog-edge spectrum.

https://doc-vu.github.io/Stratum/