Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Autodeploy

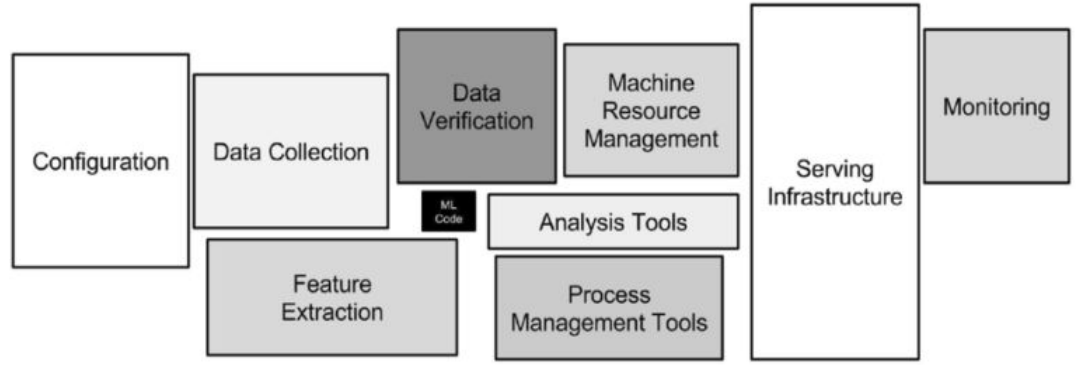
Scalable library for model management

Lenovo-BSC collaboration

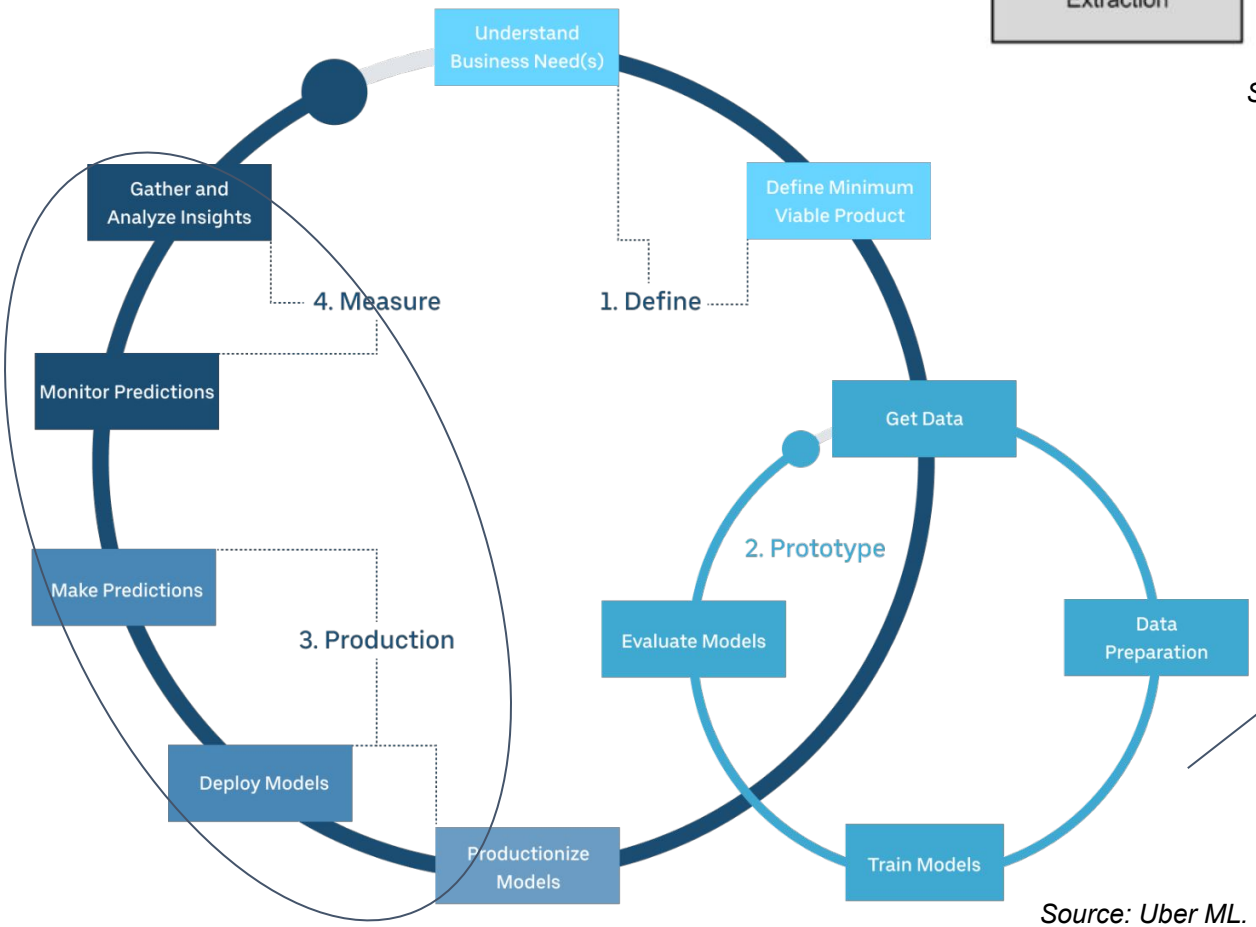
January 2020

Applied Machine Learning: Workflow

Deployment: The ML code is a small part of the complete pipeline.



Source: Sculley D., Holt. G.



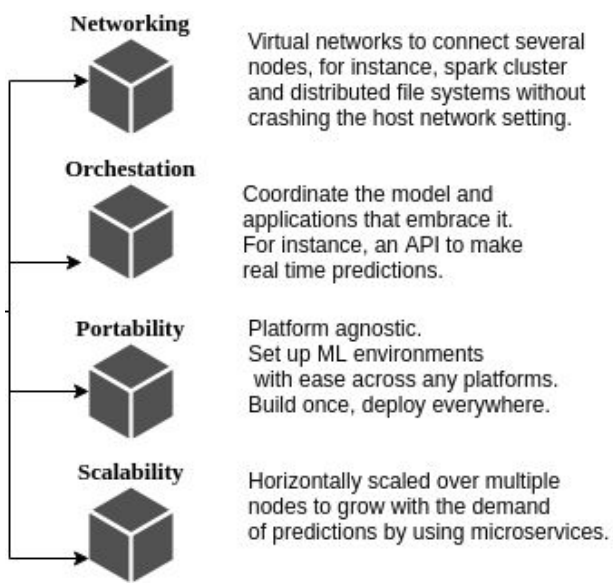
Applied ML workflow: There are more post-modeling steps.

Source: Uber ML.

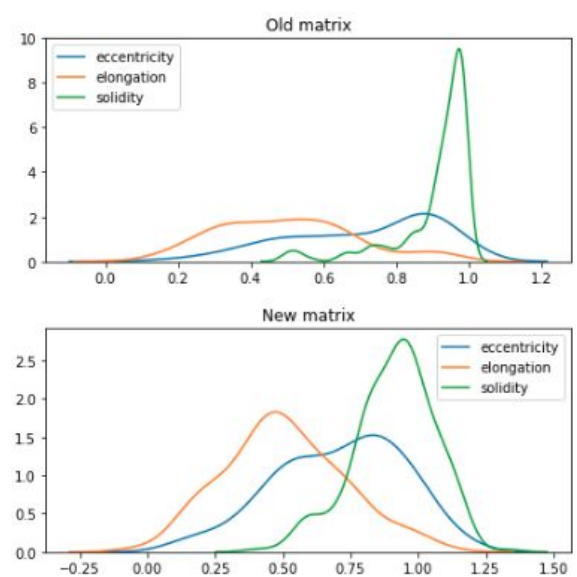
Applied Machine Learning: Issues

- Moving a complete workflow from development platform to another new platform can break things, e.g, operating system, libraries, dependencies, etc.
- Some steps in a workflow need different amount and type of computational resources, e.g, RAM, Storage, CPU, GPU.
- The complete workflow might scale from a single node to a cluster.
- The dataset distribution might change (normal, poison, etc, or different patterns). It is called distribution drift. An anomaly detector might help this.
- Some feature levels and balance of classes might change (categories, e.g, before {red, blue}, after {red, blue, black}. Classes, e.g, before {30% men, 70%women}. after {60% men, 40%women}).

Flexibility, portability, scalability

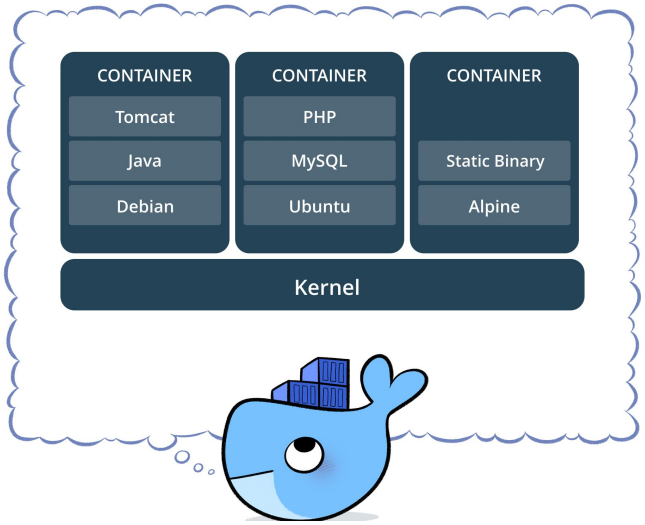


Distribution drift



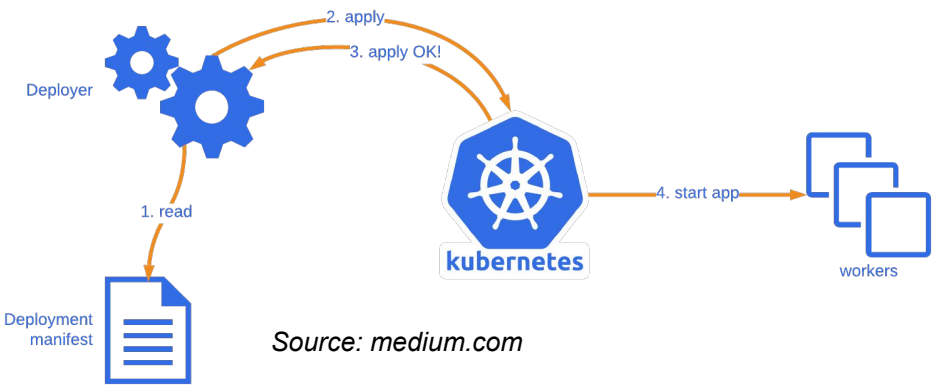
Applied Machine Learning: Recipes

Docker: OS-level virtualization



Source: arquitectoit.com

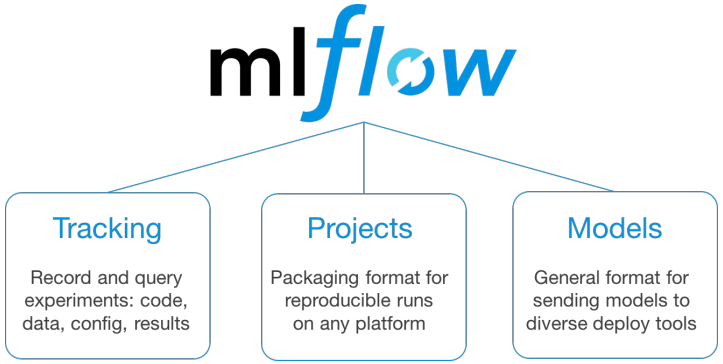
Kubernetes: container-orchestration



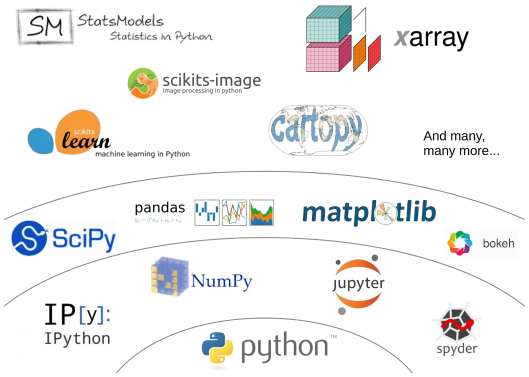
Source: medium.com

Python: Works quickly and integrate systems more effectively. Robust AI ecosystem

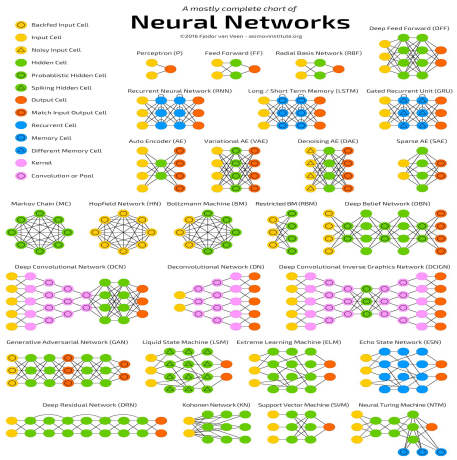
MLflow: ML Lifecycle Platform



Source: mlflow.org

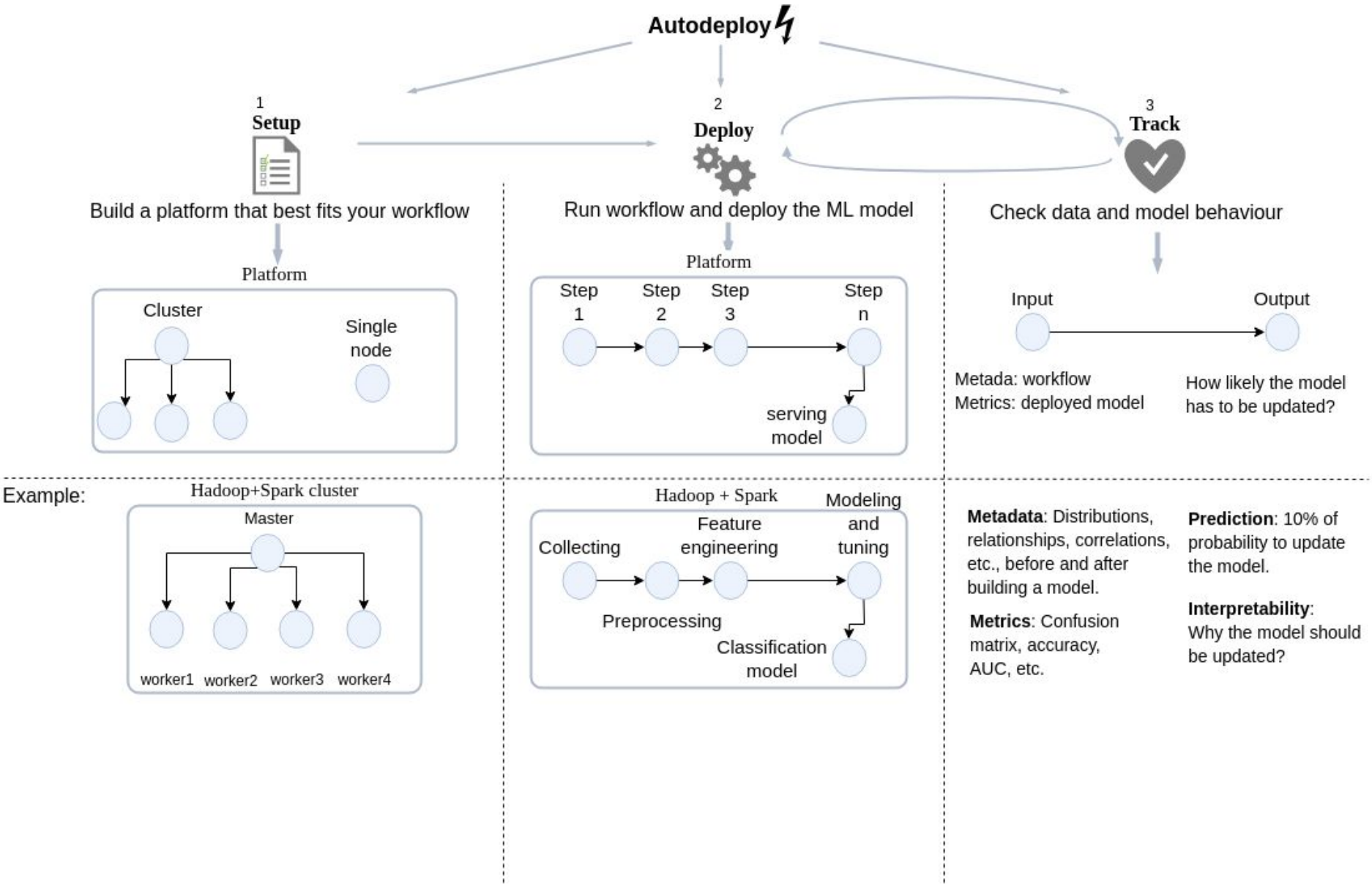


Source: leblancfg.com

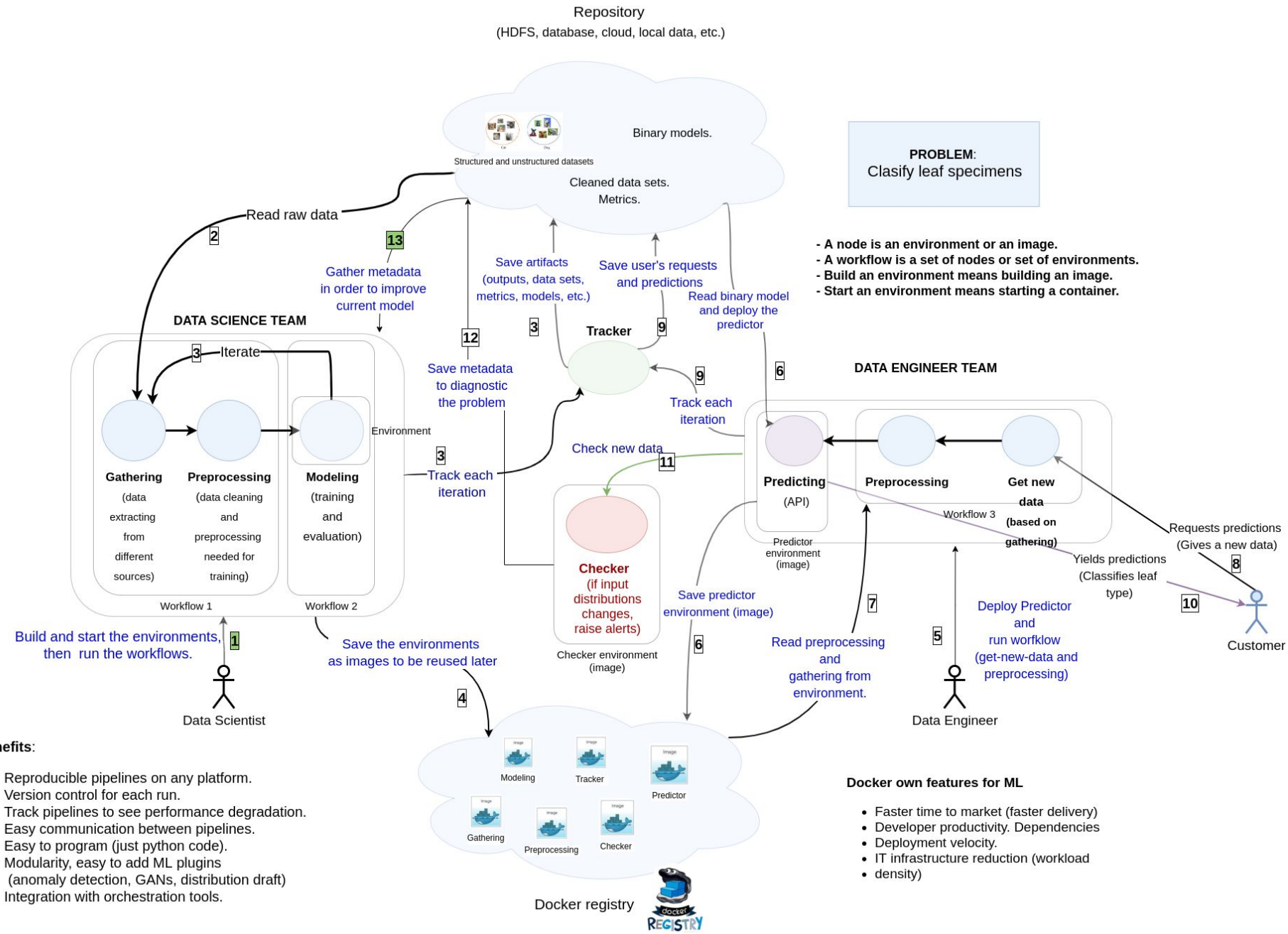


Source: fjodor van veen - asimovinstitute.org

Autodeploy: High-level overview



MACHINE LEARNING PIPELINE WITH AUTODEPLOY



Benefits:

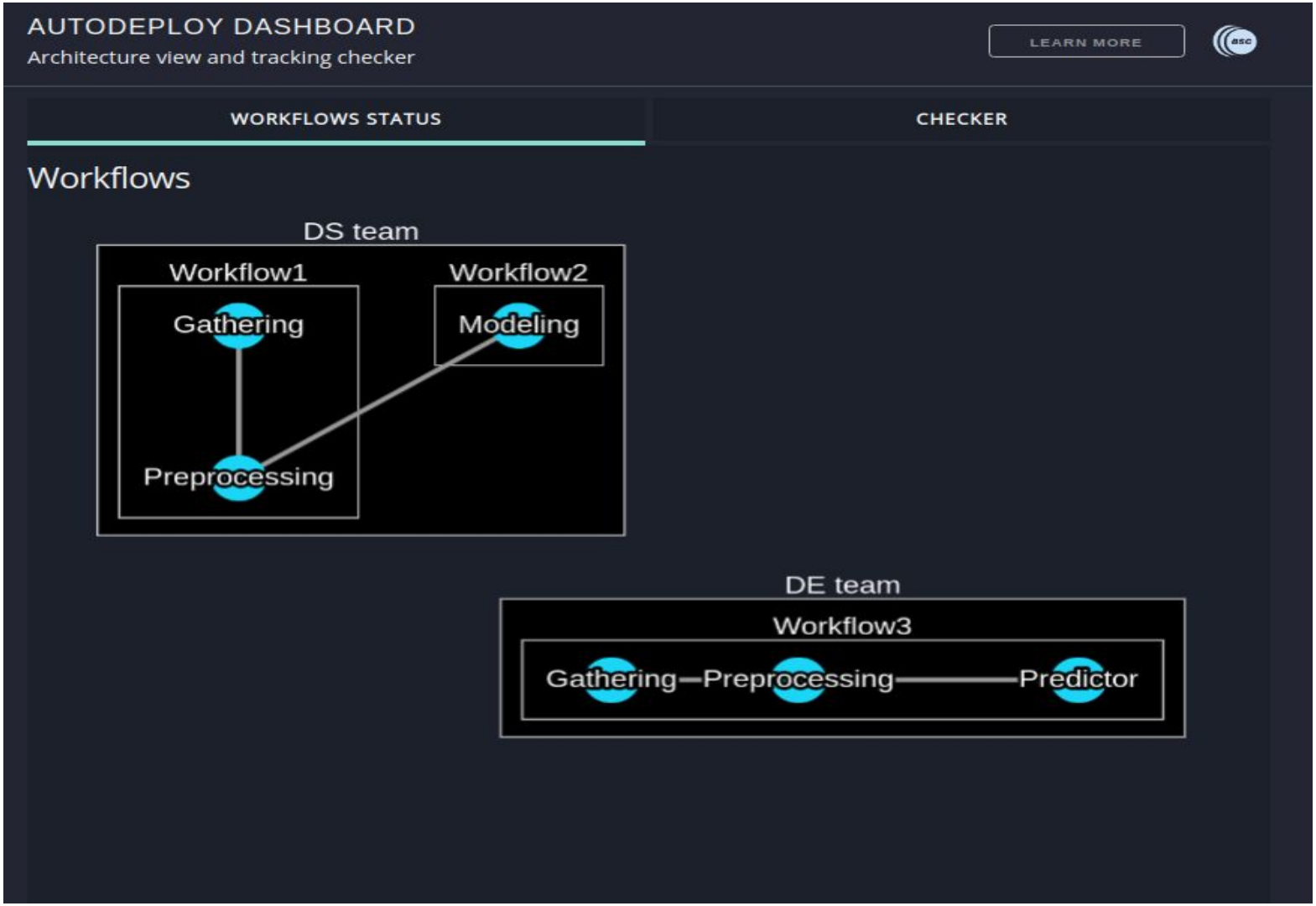
- Reproducible pipelines on any platform.
- Version control for each run.
- Track pipelines to see performance degradation.
- Easy communication between pipelines.
- Easy to program (just python code).
- Modularity, easy to add ML plugins
- (anomaly detection, GANs, distribution draft)
- Integration with orchestration tools.

Docker own features for ML

- Faster time to market (faster delivery)
- Developer productivity. Dependencies
- Deployment velocity.
- IT infrastructure reduction (workload
- density)

Setup and deployment

Design and start your workflows



Tracking

Track each workflow

mlflow GitHub Docs

Default

Experiment ID: 0 Artifact Location: /mlflow/mlruns/0

Description:

Search Runs: metrics.rmse < 1 and params.model = "tree" State: Active Search

Filter Params: alpha, lr Filter Metrics: rmse, r2 Clear

Showing 6 matching runs Compare Delete Download CSV

	Date	User	Run Name	Source	Versi…	Tags	Parameters
	2020-01-07 15:34:01	root	preprocessi…	prepro…			dtypes: {'species': 'int64',… n_classes: 30 n_features: 14 n_samples: 340 problem_type: classification
	2020-01-07 15:33:59	root	gathering	gather…			

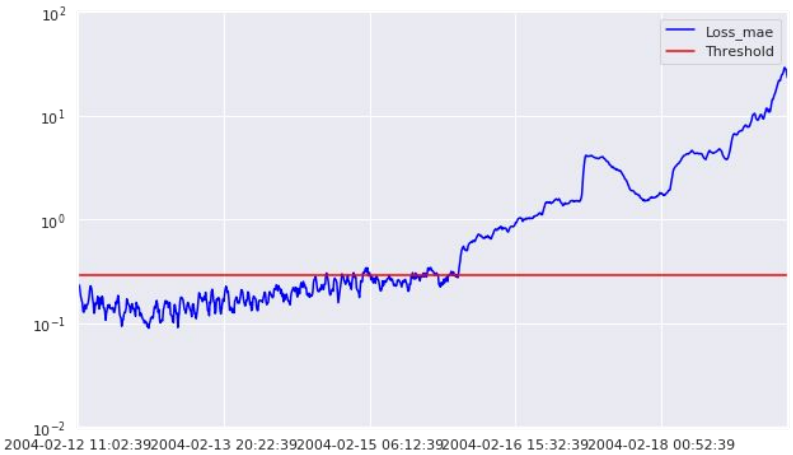
Checking

Check for anomaly patterns

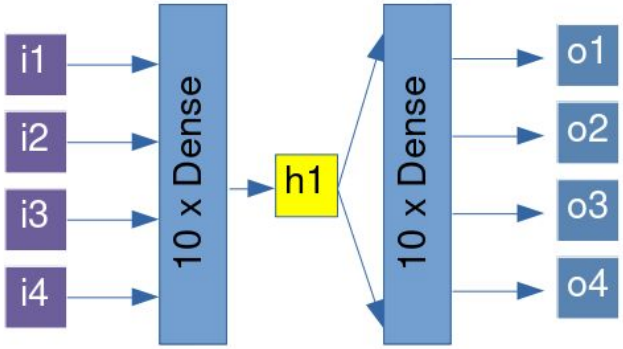
Checker architecture

Input

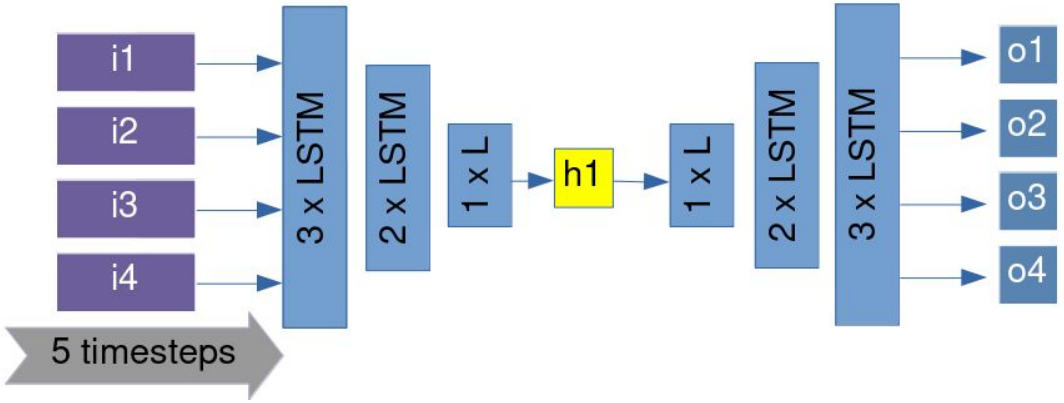
Output



Naive Autoencoder

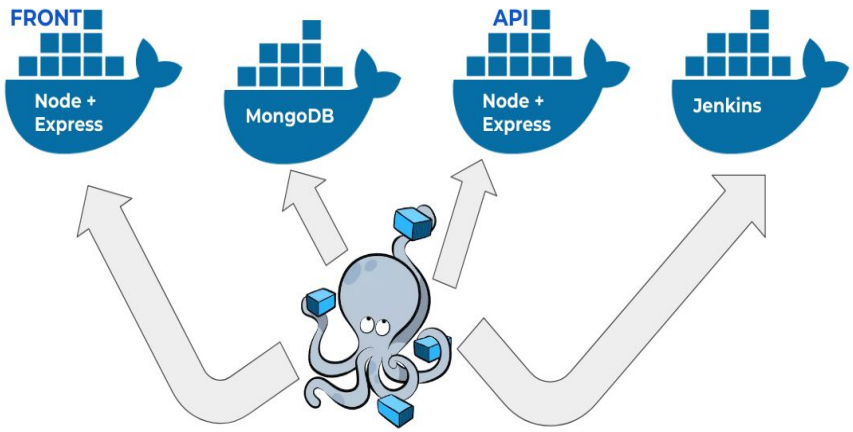


LSTM Autoencoder



Integration: Docker compose

Docker compose behaviour



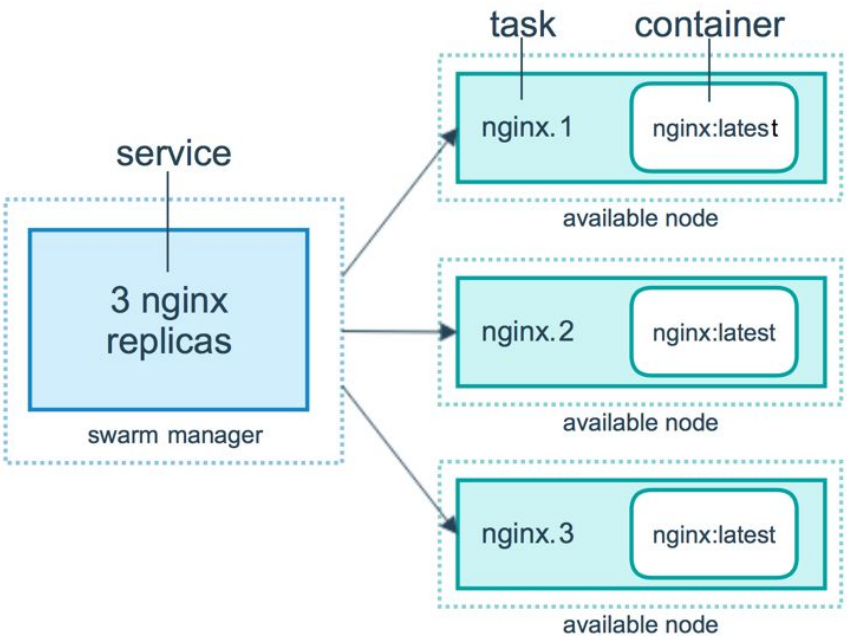
Source: medium.com

docker-compose.yml

```yaml
version: '3'
services:
  get_new_data:
    image: get_new_data
    container_name: get_new_data-20200126033621
    networks:
    - network-workflow3
    depends_on:
    - tracker-workflow3
    environment:
      MLFLOW_TRACKING_URI: http://tracker-workflow3:8003
    volumes:
    - /home/guess/Desktop/autodeploy/examples/demo2/data-
eng/:/app
    - /home/guess/Desktop/autodeploy/examples/demo2/data-
eng/ad-stuff/ad-tracker/tracker-workflow3:/mlflow
    tty: 'true'
  preprocessing_new_data:
    image: preprocessing_new_data
    container_name: preprocessing_new_data-20200126033621
    networks:
    - network-workflow3
    depends_on:
    - tracker-workflow3
    environment:
      MLFLOW_TRACKING_URI: http://tracker-workflow3:8003
    volumes:
    - /home/guess/Desktop/autodeploy/examples/demo2/data-
eng/:/app
    - /home/guess/Desktop/autodeploy/examples/demo2/data-
eng/ad-stuff/ad-tracker/tracker-workflow3:/mlflow
    tty: 'true'
  tracker-workflow3:
    image: tracker-workflow3
    container_name: tracker-workflow3-20200126033621
    networks:
    - network-workflow3
    volumes:
    - /home/guess/Desktop/autodeploy/examples/demo2/data-
eng/ad-stuff/ad-tracker/tracker-workflow3:/mlflow
    ports:
    - 8008:8003
networks:
  network_workflow3: null
```

Integration: Docker Swarm

Docker Swarm behaviour



Source: filepicker.io

Docker Swarm console

```
[xgbravo@nxt2027 compose_repo]$ docker service ls
ID              NAME                            MODE         REPLICAS    IMAGE                              PORTS
5r1l6yx27pfk    my_swarm_gathering              replicated   1/1         gathering:latest
xke5uf9aqdh4    my_swarm_modeling               replicated   1/1         modeling:latest
ro9haibzt9ma    my_swarm_preprocessing          replicated   1/1         preprocessing:latest
pvdud6whi4pg    my_swarm_tracker_workflow1      replicated   1/1         tracker_workflow1:latest           *:8006->8001/tcp
alnxiq8y3tfs    my_swarm_tracker_workflow1_scale replicated  0/5         my_swarm:latest
2dh4zpwu48od    my_swarm_tracker_workflow1_scale2 replicated 0/5         my_swarm_tracker_workflow1:latest
ilzhcp546xft    my_swarm_tracker_workflow1_scale3 replicated 5/5         tracker_workflow1:latest
ygwulqaah8a8    my_swarm_tracker_workflow2      replicated   1/1         tracker_workflow2:latest           *:8007->8002/tcp
```

Integration: Kubernetes

Kubernetes dashboard

kubernetes

Search

Discovery and Load Balancing > Services

Cluster

- Cluster Roles
- Namespaces
- Nodes
- Persistent Volumes
- Storage Classes

Namespace

default

Overview

Workloads

- Cron Jobs
- Daemon Sets
- Deployments

Services

	Name	Namespace	Labels	Cluster IP	Internal Endpoints	External Endpoints	Age
✓	tracker-workflow1	default	io.kompose.service: tracker-workflow1	10.152.183.	tracker-workflow1:8 TCP tracker-workflow1:0 TCP	-	8 minutes
✓	tracker-workflow2	default	io.kompose.service: tracker-workflow2	10.152.183.	tracker-workflow2:8 TCP tracker-workflow2:0 TCP	-	8 minutes
✓	kubernetes	default	component: apiserver provider: kubernetes	10.152.183.	kubernetes: TCP kubernetes: TCP	-	2 hours

1 – 3 of 3

Is it relevant?. AI predictions for 2020.

Creator of pytorch: … " place more value on AI model performance beyond accuracy. "

Celeste Kidd, psychologist at the University of California, Berkeley: … " increased awareness of the real-life implications of tech tools … "

Jeff Dean, Google AI chief: … " he wants to see less of an emphasis on slight state-of-the-art advances in favor of creating more robust models. "

Anima, Anandkumar, NVIDIA: … " self-supervision, and self-training methods of training models, which are the kinds of models that can improve through self-training with unlabeled data. "

Dario gil, IBM: …" focus on metrics beyond accuracy to consider the value of models deployed in production. Shifting the field toward building trusted systems instead of prioritizing accuracy above all else will be a central pillar to the continued adoption of AI. "

Keywords: **robust models, interpretable models, trusted models, self-supervision (automatic)**.

Source: https://venturebeat.com/2020/01/02/top-minds-in-machine-learning-predict-where-ai-is-going-in-2020/