MLOps Intro

August | 2022

Intro & Agenda

Jonathan Cosme

AI/ML Solutions Architect
Jonathan.Cosme@run.ai

Agenda

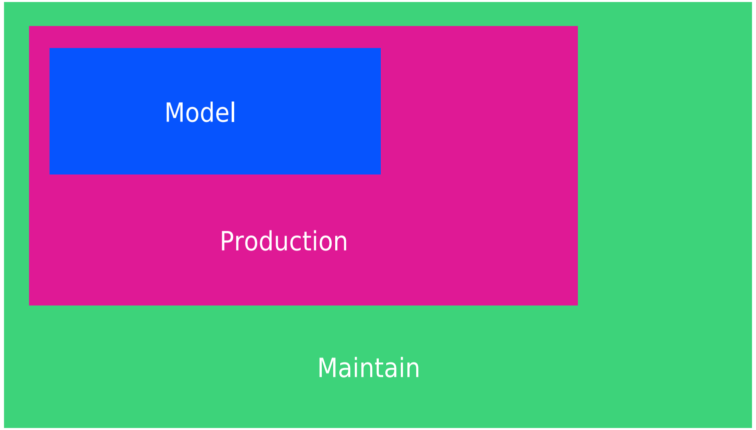
1. MLOps overview
2. Roles & responsibilities
3. Maintenance step
4. Serving step
5. Build step

MLOps Overview

Goal of MLOps

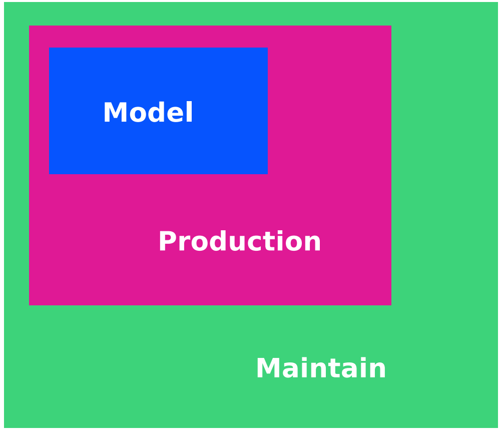
End Step:

Maintain a model in production

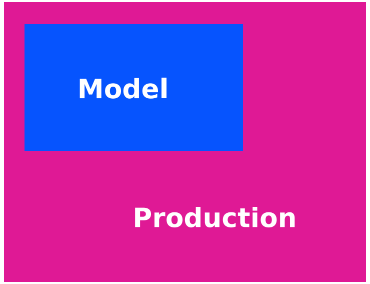


Steps of MLOps

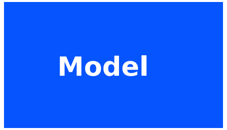
End Step:
Maintain a model in production



End Step - 1:
Serve a model in production



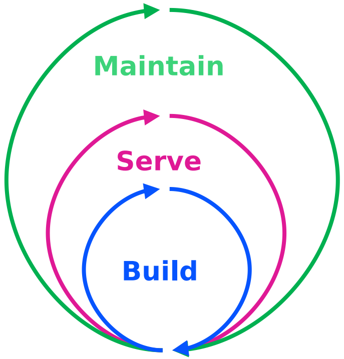
End Step - 2:
Build a model



Process of MLOps

The process of MLOps is **not linear**

The process of MLOps is a **feedback loop.**



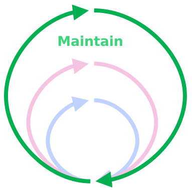
Role Responsibilities

Who’s responsibility is it?

Build	• Data Scientist	It is the role of the **data scientist** to **deliver** trained and optimized **models**.
Serve	• ML Engineer • Software Developer	The **machine learning engineer** arguably has the most difficult job: **putting** a **model into production**. Often, they will work alongside software developers/engineers to achieve this. A **common mistake** organizations make is hiring **too many data scientist**, and **too few ML engineers.**
Maintain	• ML Engineer • Data Scientist • Stakeholder	Maintenance is a coordinated effort between the ML engineer, data scientist, and stakeholder.

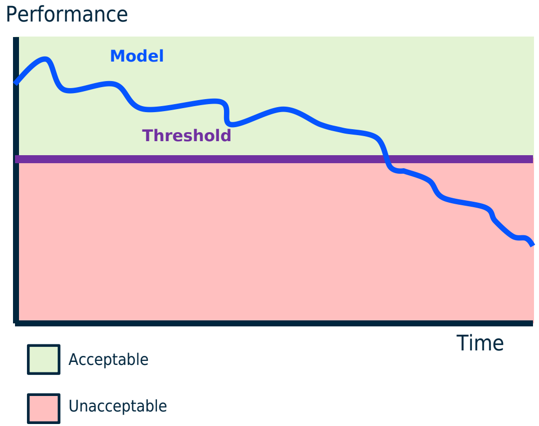
Maintenance Step

Maintenance



What do we want to do?

Make sure our model is **meeting** our **minimum** threshold of **performance**.

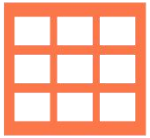


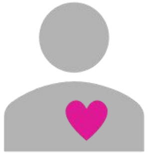
How do we do this?

1) **Monitoring:**

Model Predictions

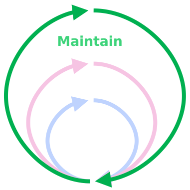
Input Data





User Feedback

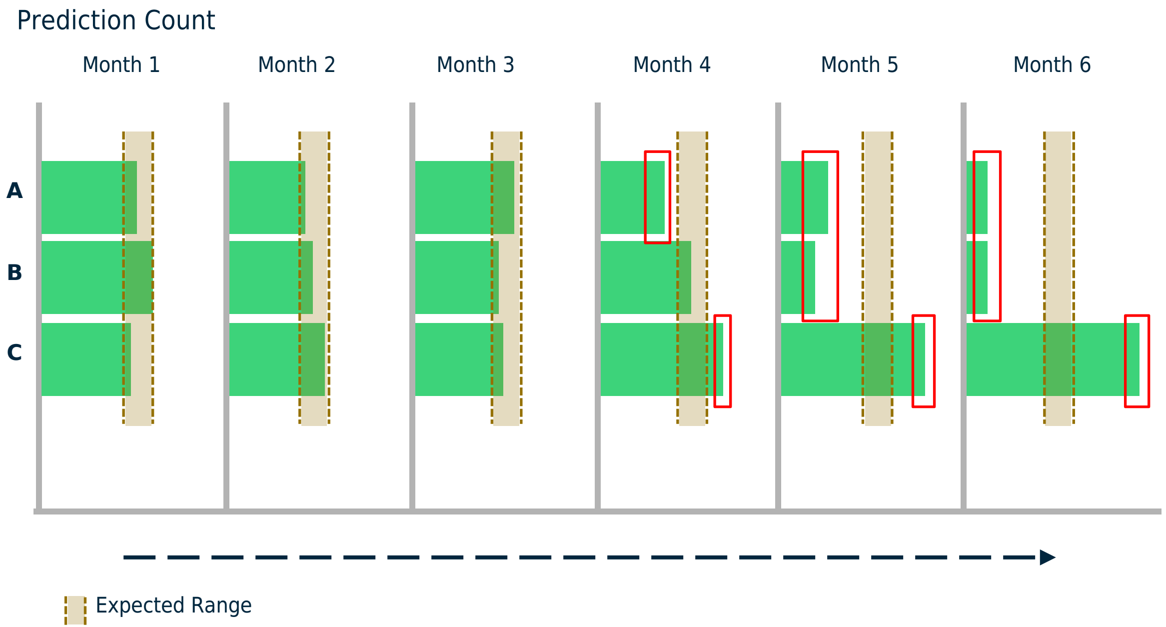
2) **Retraining**

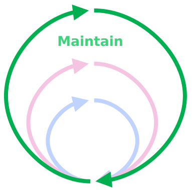


Model Predictions

Watching the distribution and frequencies of model predictions over time can **help identify warning signs** of model performance degradation.

In this example, we can see that over time, the **frequencies** of the model predictions **move further** and further away **from our expected range.**

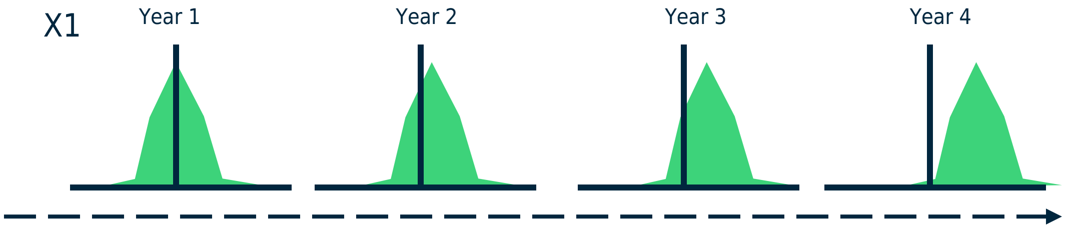




Input Data

The **underlying distribution of data** can (and often) **shifts with time** It is crucial to **keep an eye on data drift**.

The figure for input data variable **X1** illustrates this. Since it is common practice to normalize input data before feeding it to a model, it's important to track distributions.

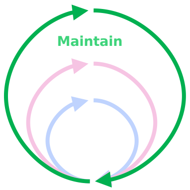


As time goes by, it is common for a particular variable to be **able to take on new values**, that previously did not exist. The figure for input data variable **X2** illustrates this. Since encodings for models are finite, the frequency and count of new unknown values must be monitored

X2

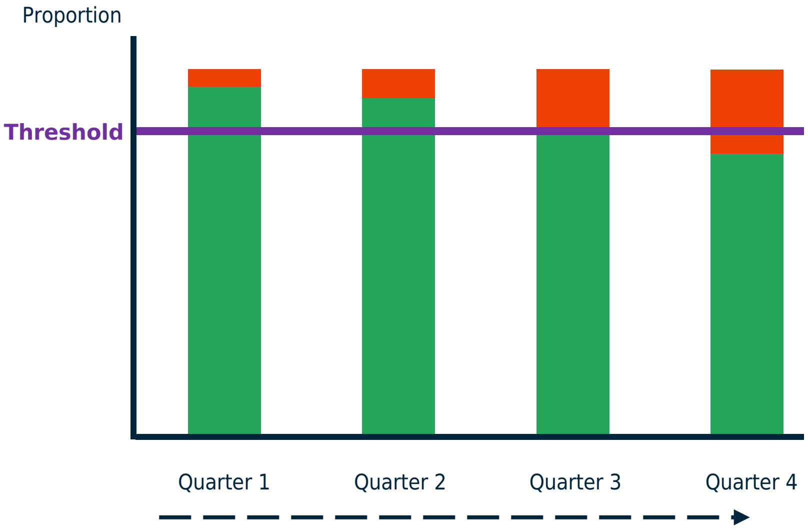
Decade 1	Decade 2	Decade 3
Genres	Genres	Genres
Drama	Drama	Drama
Action	Action	Action
Comedy	Adventure	Adventure
Romance	Comedy	Comedy
	Dark Comedy	Dark Comedy
	Slapstick Comedy	Slapstick Comedy
	Romance	Romance
	Romantic Comedy	Romantic Comedy
	Documentary	Documentary
		Foreign

User Feedback

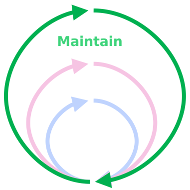


In order to gauge the effectiveness of a model, many AI **systems have a method for obtaining user feedback** built in.

A common feature is the **thumbs-up or thumbs-down** buttons available **on recommender systems.** Music and movie streaming services are heavy users of this. The proportion of feedback should be tracked, with the goal of trying to **maintain as high a positive feedback proportion as possible**.

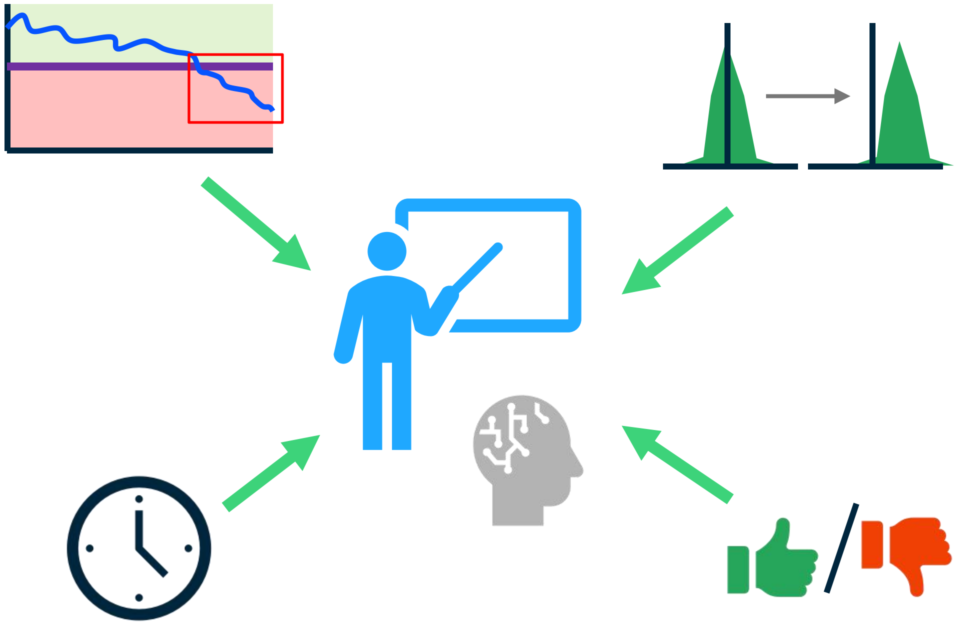


Retraining



Retraining your model **is critical** to the success of a machine learning or AI effort.

Retraining should occur when model **performance is below par**, when there are **changes in data**, after **user feedback is received**, and at **regularly scheduled intervals**.



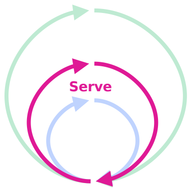
Who's responsibility is it?

Build	• Data Scientist	It is the role of the **data scientist** to **deliver** trained and optimized **models**.
Serve	• ML Engineer • Software Developer	The **machine learning engineer** arguably has the most difficult job: **putting** a **model into production**. Often, they will work alongside software developers/engineers to achieve this. A **common mistake** organizations make is hiring **too many data scientist**, and **too few ML engineers.**
Maintain	• ML Engineer • Data Scientist • Stakeholder	Maintenance is a coordinated effort between the ML engineer, data scientist, and stakeholder.

run:ai

Serve Step

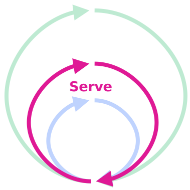
Serve



What is needed to put a model into production?

- A **model**, optimized for inference
- A **location** to put the model
- Ways to **get data** into the model
- Ways to get model **outputs to users**
- Ability to **collect user feedback**
- Model **retraining** capabilities
- Data, metadata, and model version **tracking**
- As much **automation** as possible

Model

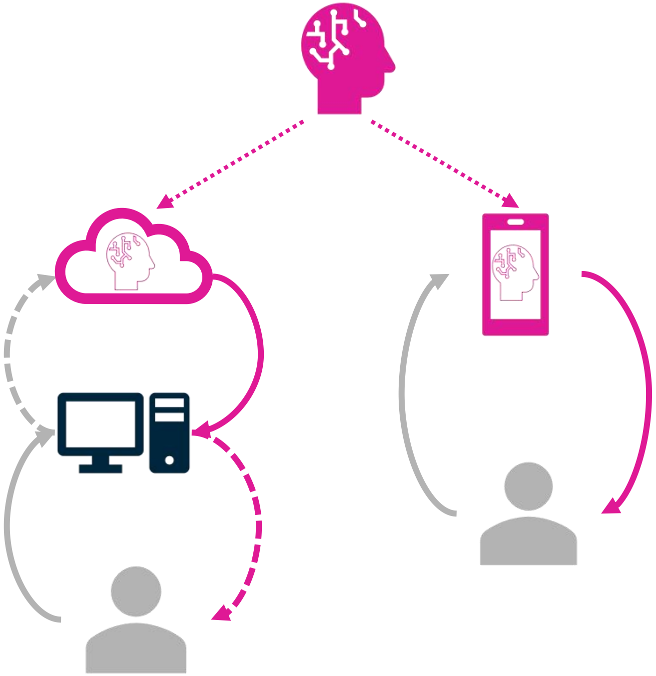


A trained **model**, optimized **for inference**

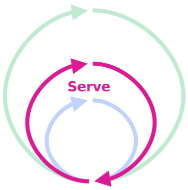
A place to **put the model**. You can either keep your model on a **server**, **or** on **edge** devices.

Maintaining models on **a server is easier** to set up an maintain, but has the disadvantage of **not** being **accessible without a connection** to the server.

Models on **edge** devices can perform **inference at any time**, **without needing** to contact a **server**. **Maintaining** models on edge devices is significantly **more difficult** however, as it involves maintaining any number of models, and potentially creating custom applications or software for the edge device.



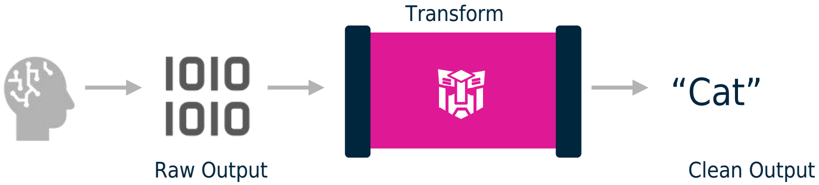
Data



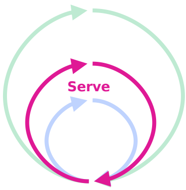
Models require very specific data, in a very specific way.

In order to get to this clean data that a model can use, **data preprocessing pipelines** have to be created. They should be able to take raw data, **clean** it, **transform** it, and **normalize** it.

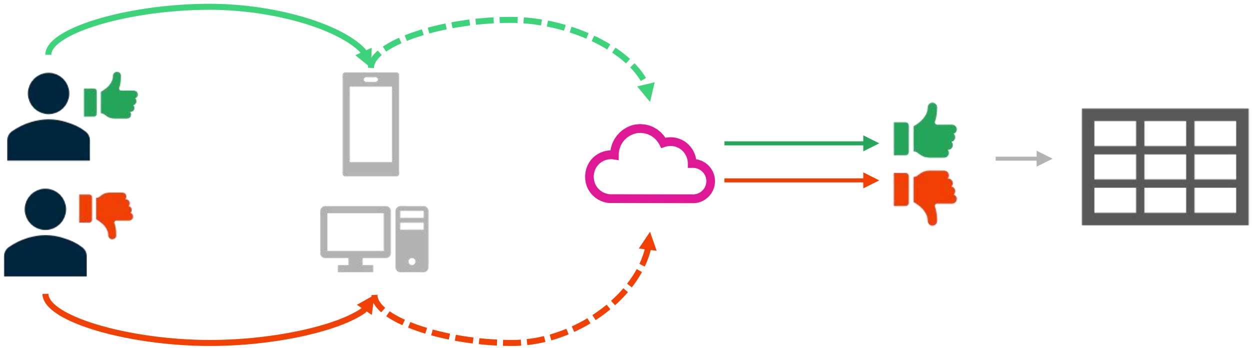
Sometimes a model will also give results that are not too user friendly. In these cases, a **post processing pipeline** must be created.



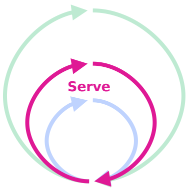
Feedback



For recommender systems, user feedback is extremely important. A **pipeline** should be created for **capturing user feedback**.



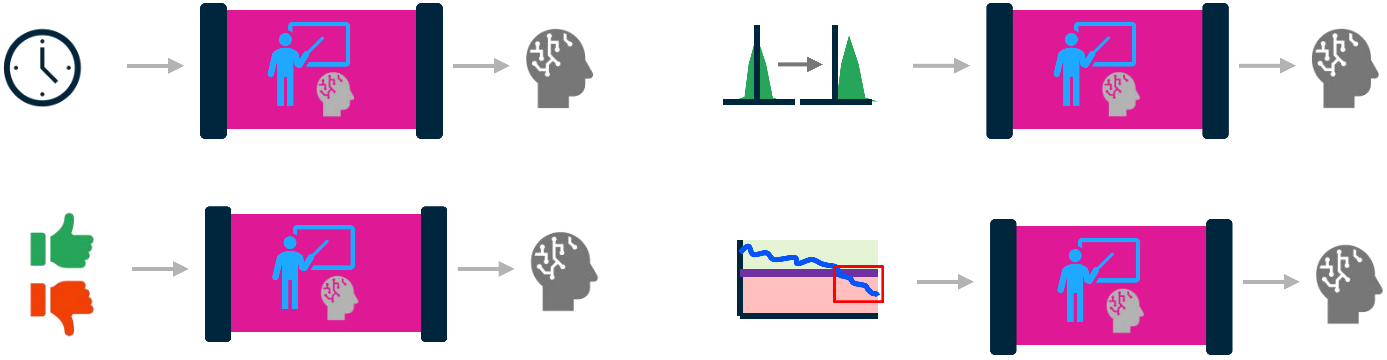
Retraining



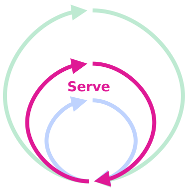
Pipelines for retraining

- on regular intervals,
- on changed data,
- on user feedback,
- or for model performance,

should be created.

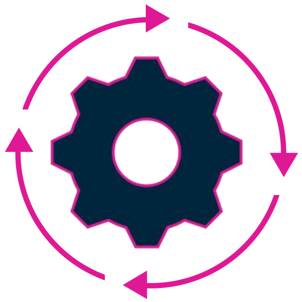


Tracking & Automation



It’s important to set **up tracking of model changes and data schemas.** Sometime case model reversion is necessary and we use tracking to do it.

Finally, **automate as much as possible**. It will still always be necessary to have some level of human input, but automation decreases time to production, and increases productivity

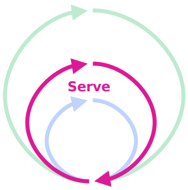


Who’s responsibility is it?

Build	• Data Scientist	It is the role of the **data scientist** to **deliver** trained and optimized **models**.
Serve	• ML Engineer • Software Developer	The **machine learning engineer** arguably has the most difficult job: **putting** a **model into production**. Often, they will work alongside software developers/engineers to achieve this. A **common mistake** organizations make is hiring **too many data scientist**, and **too few ML engineers.**
Maintain	• ML Engineer • Data Scientist • Stakeholder	Maintenance is a coordinated effort between the ML engineer, data scientist, and stakeholder.

Build Step

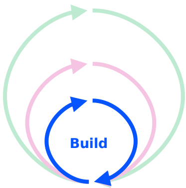
Build



What is needed to build a model?

- Data
    - labels
    - ETL
- Model
    - architecture
    - tuning
- Reproducibility
    - experiment tracking

Data



In order to get our model to accept our data, it is necessary to apply data transformations to the raw data

Common data transformations include cleaning, scaling, encoding, and feature engineering

Cleaning

Example ID	Color
1	Red
2	NULL
3	Green

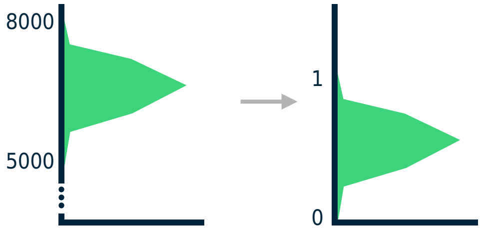
Example ID	Color
1	Red
3	Green

Label encoding

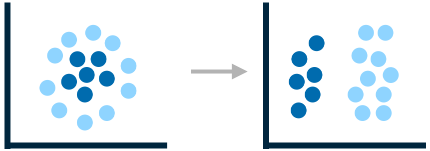
Example ID	Color
1	Red
2	Blue
3	Green

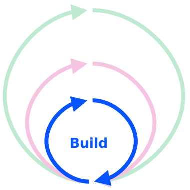
Example ID	Color
1	0
2	1
3	2

Scaling



Feature engineering





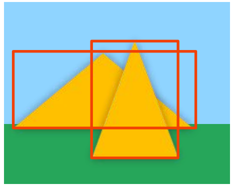
Data

An often overlooked aspect of model building is data **label consistency**.
Inconsistent labels can severely impact model performance

In this example, we can see that the same set of instructions can be interpreted many ways by annotators.

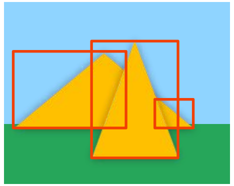
Draw a box around triangles

Annotator 1



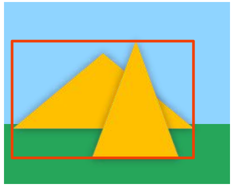
(**2** boxes)

Annotator 2



(**3** boxes)

Annotator 3



(**1** box)

In this example, the ambiguity of the word “bad” can lead different annotators to label the same sentence as opposite sentiments

Mark label as positive or negative sentiment

Annotator 1

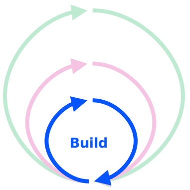
User ID	Text	Label
1	He is a bad dude.	+
2	He is a bad dude!	+
3	He is a bad dude…	+

Annotator 2

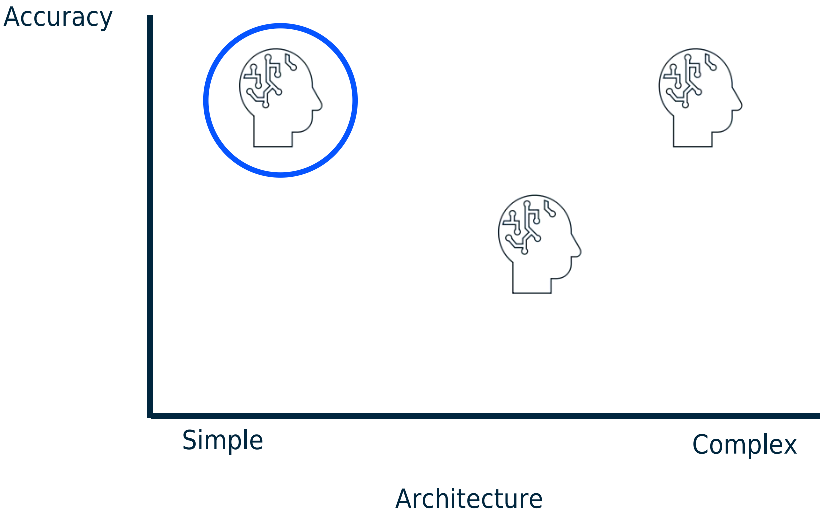
User ID	Text	Label
1	He is a bad dude.	-
2	He is a bad dude!	-
3	He is a bad dude…	-

Model

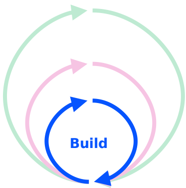
The model architecture selection(s) will depend on the problem



All else equal, we prefer a simpler model

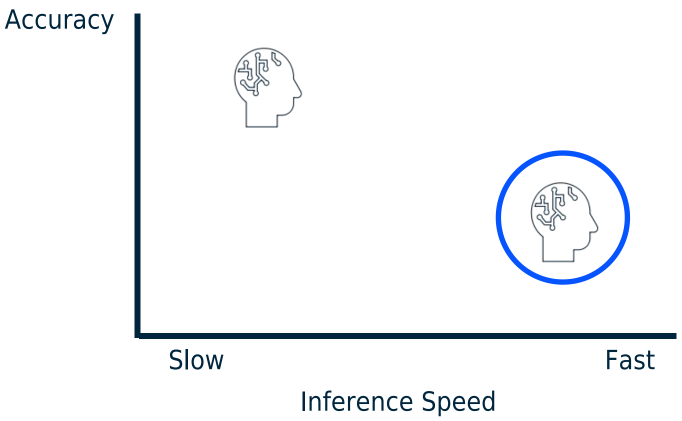


Model

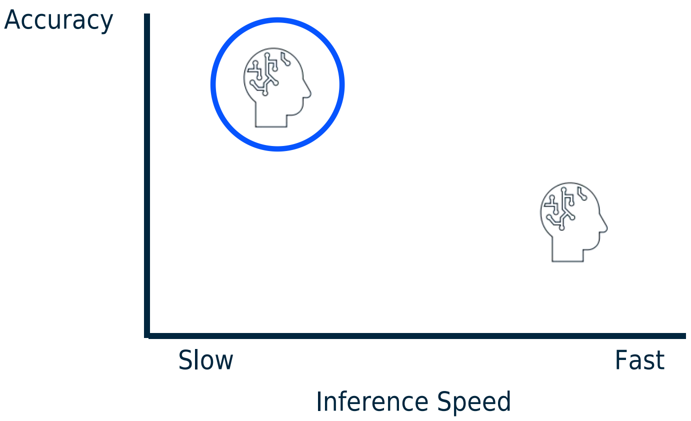


The model architecture selection(s) will depend on the problem

If our problem needs to be solved in **real time** (like **autonomous driving**), we will need to select a model with fast inference speed



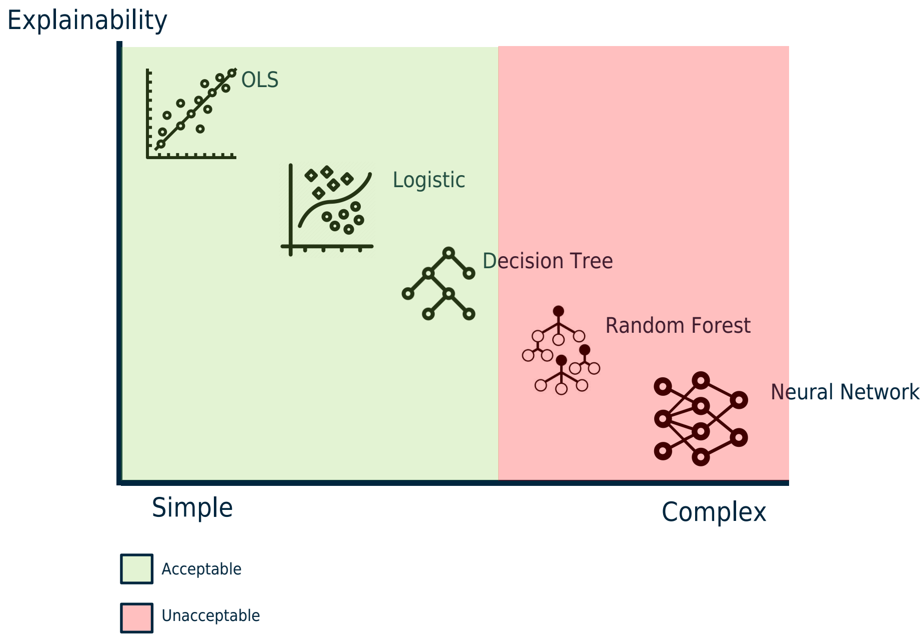
If **accuracy is the top priority** (like in **medical imaging**), then we will focus on architectures with high accuracy



Model

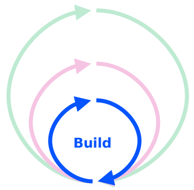
The model architecture selection(s) will depend on the problem

If **model fairness** is an important metric, there will be a limitation on the architectures that can be used based on **explainability**,



Model

Hyperparameter tuning



Trying out different combinations of hyperparameters to arrive at the best model is a normal part of model building

Relu Activation

Batch size	Learning Rate: 0.01	0.001	0.0001
16			
32		Accuracy	
64			

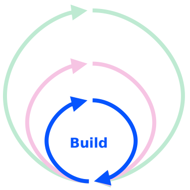
Selu Activation

Batch size	Learning Rate: 0.01	0.001	0.0001
16			
32		Accuracy	
64			

Gelu Activation

Batch size	Learning Rate: 0.01	0.001	0.0001
16			
32		Accuracy	
64			

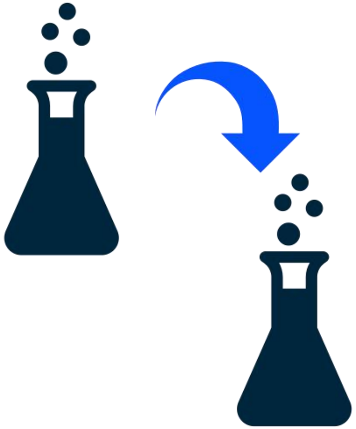
Reproducibility



We want to be able to reproduce the results of any single experiment

Meticulous records of all aspects of model building must be kept

- Data splits
- Transformation pipelines
- Models
- Hyperparamters
- Metrics
- Evaluations



Who's responsibility is it?

Build	• Data Scientist	It is the role of the **data scientist** to **deliver** trained and optimized **models**.
Serve	• ML Engineer • Software Developer	The **machine learning engineer** arguably has the most difficult job: **putting** a **model into production**. Often, they will work alongside software developers/engineers to achieve this. A **common mistake** organizations make is hiring **too many data scientist**, and **too few ML engineers.**
Maintain	• ML Engineer • Data Scientist • Stakeholder	Maintenance is a coordinated effort between the ML engineer, data scientist, and stakeholder.

run: ai

Thank you!

info@run.ai | www.run.ai