

**Kubernetes Clusters running Seldon Core:-**

1)**Kubernetes API**:- It communicates to the operator to deploy the Machine learning models.

2)**Operator**: - it is operator inside the cluster that manages the components like Kubernetes pods, Deployment, and which pulls inference graphs.

3)**Docker Registry**:- All the models would be in docker images.(as per the architecture pulling the image from docker registry)

**Deployment Controller**:-(Kubectl,CI/CD, Seldon Deploy):- Kubectl-(kuberentes commands) – It communicates with Kuberenetes API.

Rest API:- A)**REST API** is a way of accessing web services in a simple and flexible way without having any processing.

B)An API is a set of definitions and protocols for building and integrating application software.

**Inference Graph:-**

A graph specification that describes how your components are joined together.

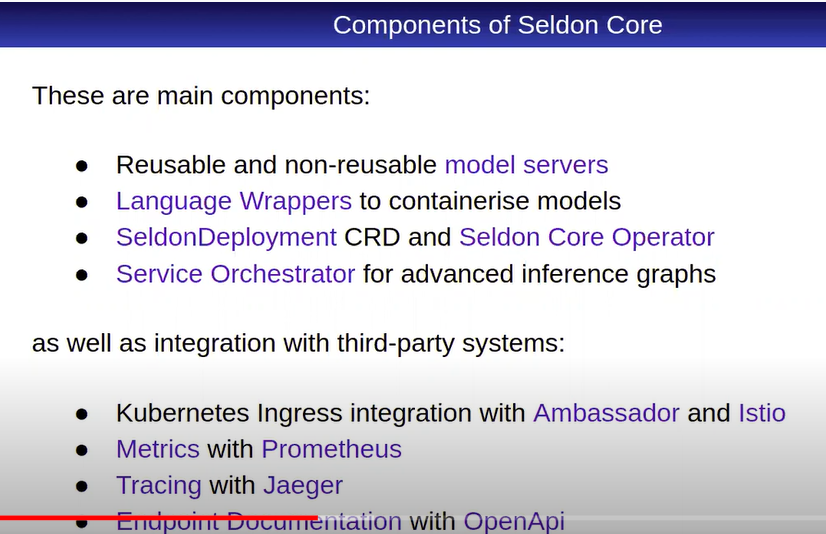
**Kubernetes ingress:** is a collection of routing rules that govern how external users access services running in a Kubernetes cluster.

**Ambassador:-**

1)Ambassador is a Kubernetes-native API Gateway built on Envoy Proxy.

2)Managed entirely via Kubernetes Custom Resource Definitions,

3)Ambassador provides powerful capabilities for traffic management, authentication, and observability.



**Language Wrappers** :- Language wrappers allows Seldon Core users to build models.

**Seldon Deployment CRD(Custom Resource Definition):-**it allows to deploy inference model to the Kubernetes cluster.

CRD:-It is extension of Kubernetes API they allow to create custom combination of basic Kubernetes objects and it defines the inference graph through manifest YAML File.

**Seldon Operator**:- it controls the seldon deployements.it reads the CRD and it applies to the cluster and it take care of required components like pods and services.

**Service orchestration**:- It is responsible for Managing the intra graph traffic.it reads inference graph structure from CRD and when inference request is received it makes sure that it is passed to each node of the graph in the right order.

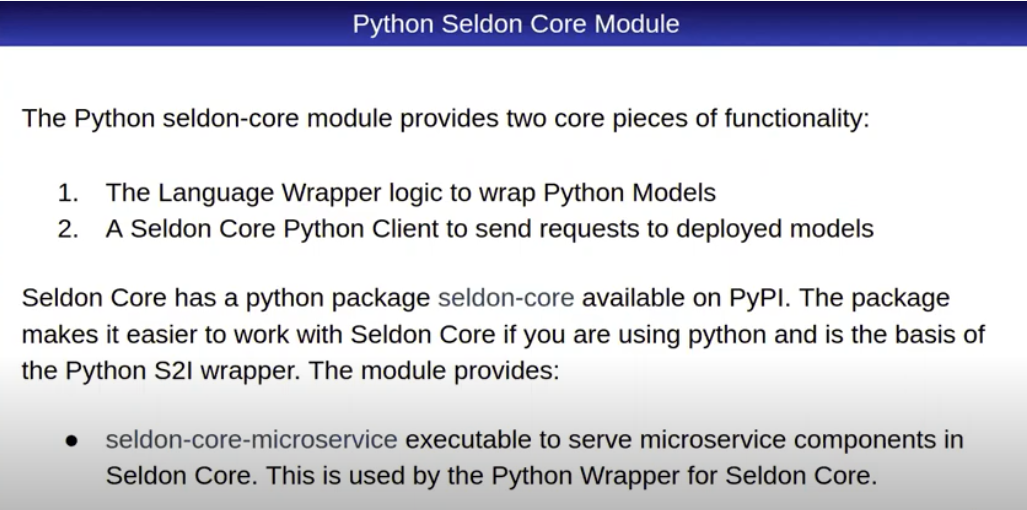
Third party System.

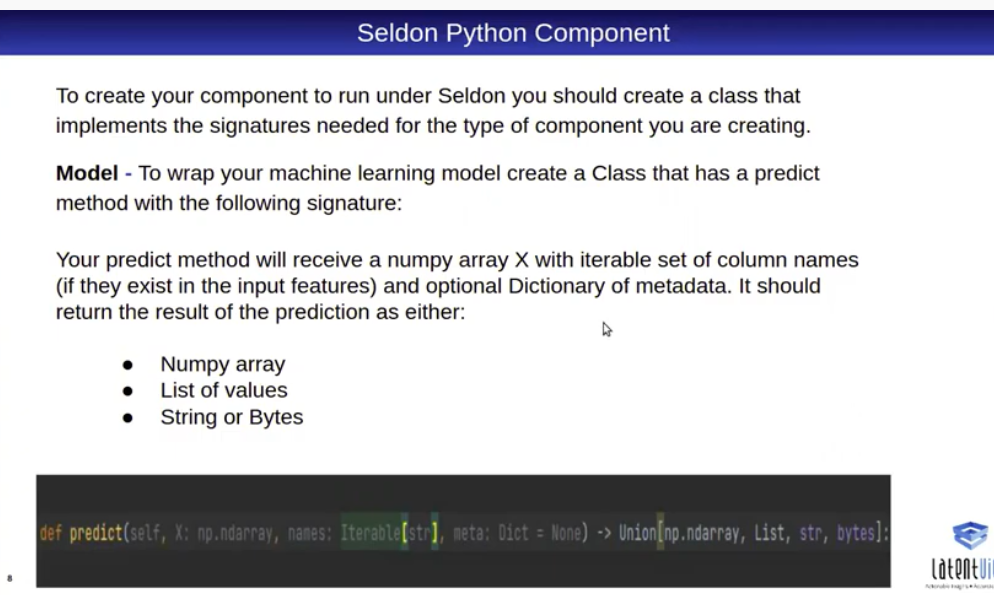
1)Kubernetes ingress integration with **ambassador** and **istio(Ingress tools)**

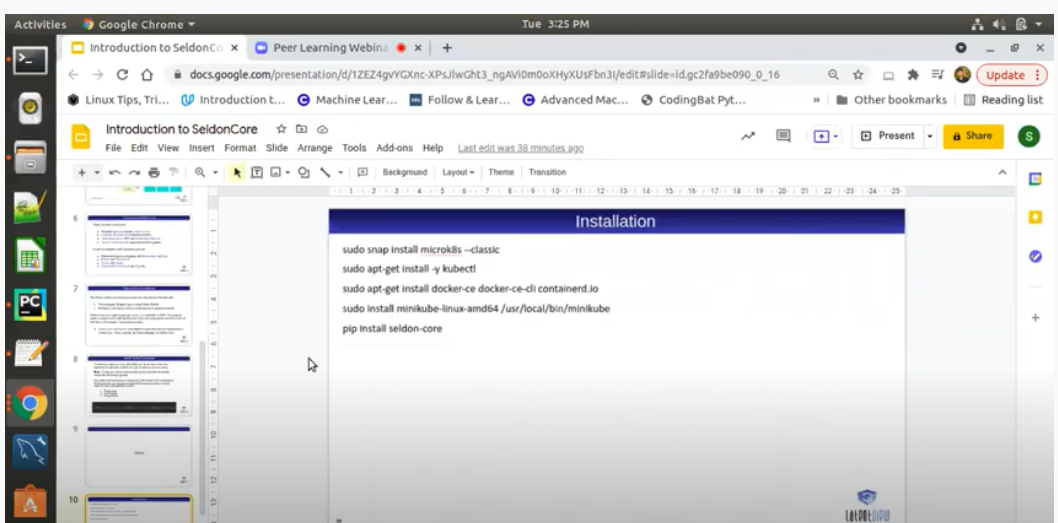
2)tracing with **jaeger**(troubleshoot with jaeger)

3)we can also metrics using Prometheus.(Monitor Tools)

4)Endpoint Documntation with **OpenAPI**:- We can also integrate with open API.







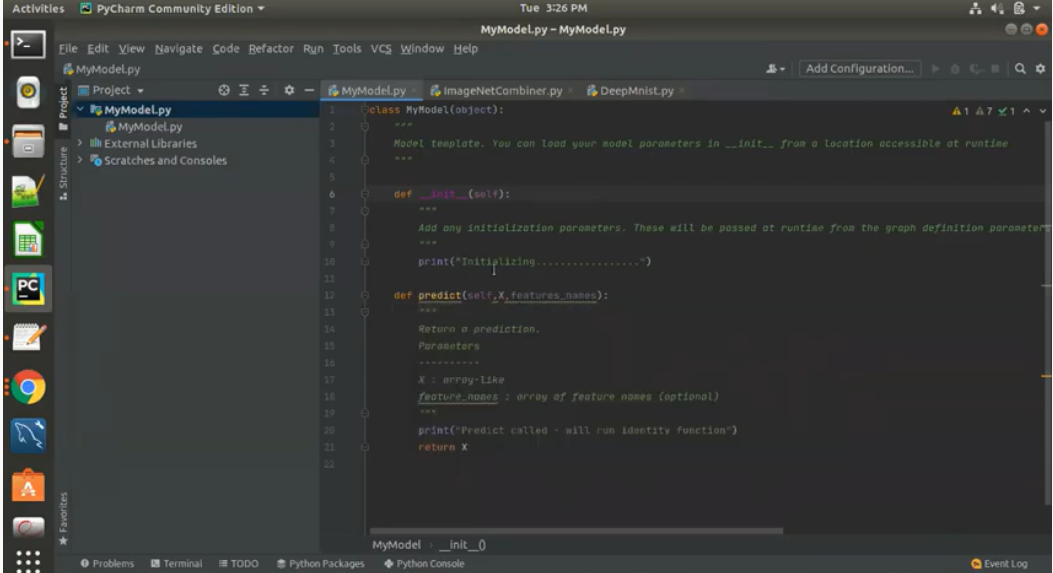
**Pre-requisites:**

1)Install Kubernetes

2)install docker

3)install seldon-core

**Basic Model:**



**Init** :- anything pre-loaded before calling actual function(predict).that can be done init function.

**Predict Function:-** Which define the models or any scoring logic

**Commands to Install:-**

