



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

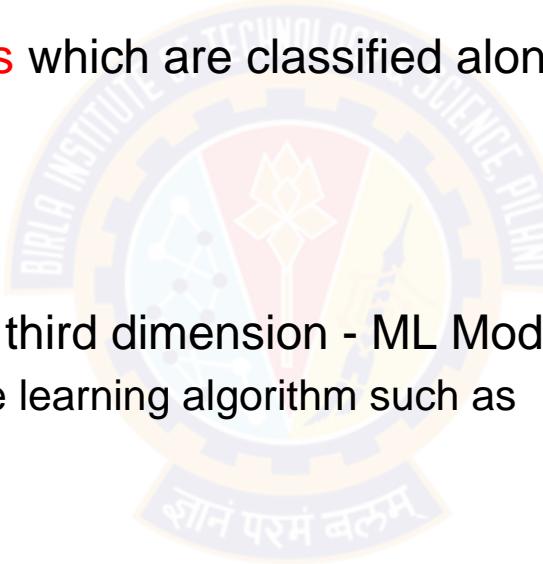
Different forms of ML workflows

Pravin Y Pawar

Extracted from
[Three Levels of ML Software](#)

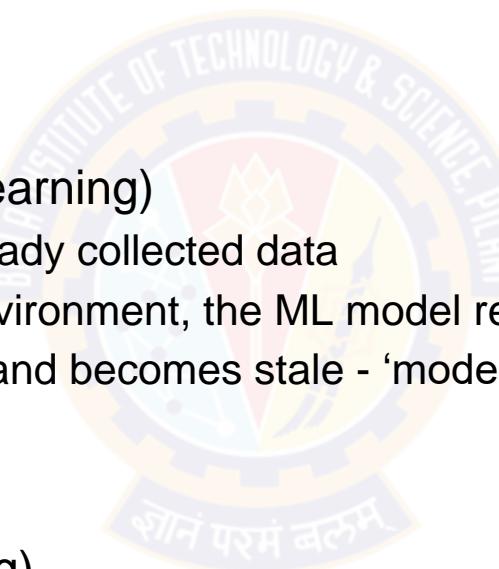
Different forms of ML workflows

- Operating an ML model might assume **several architectural styles**
- Primarily **Four architectural patterns** which are classified along two dimensions:
 - ML Model Training
 - ML Model Prediction
- For sake of simplicity disregard the third dimension - **ML Model Type**
 - which denotes the type of machine learning algorithm such as
 - Supervised
 - Unsupervised
 - Semi-supervised
 - and Reinforcement Learning



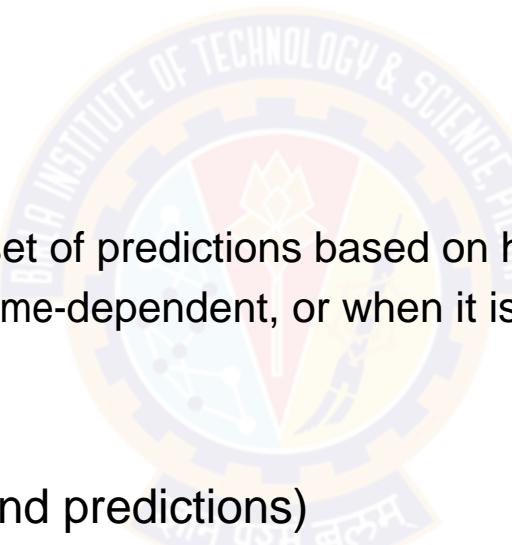
Model Training Patterns

- Two ways to perform ML Model Training:
 - Offline learning
 - Online learning
- Offline learning (aka batch or static learning)
 - The model is trained on a set of already collected data
 - After deploying to the production environment, the ML model remains constant until it re-trained
 - Model will see a lot of real-live data and becomes stale - 'model decay' and should be carefully monitored
- Online learning (aka dynamic learning)
 - The model is regularly being re-trained as new data arrives, e.g. as data streams
 - Usually the case for ML systems that use time-series data, such as sensor, or stock trading data to accommodate the temporal effects in the ML model



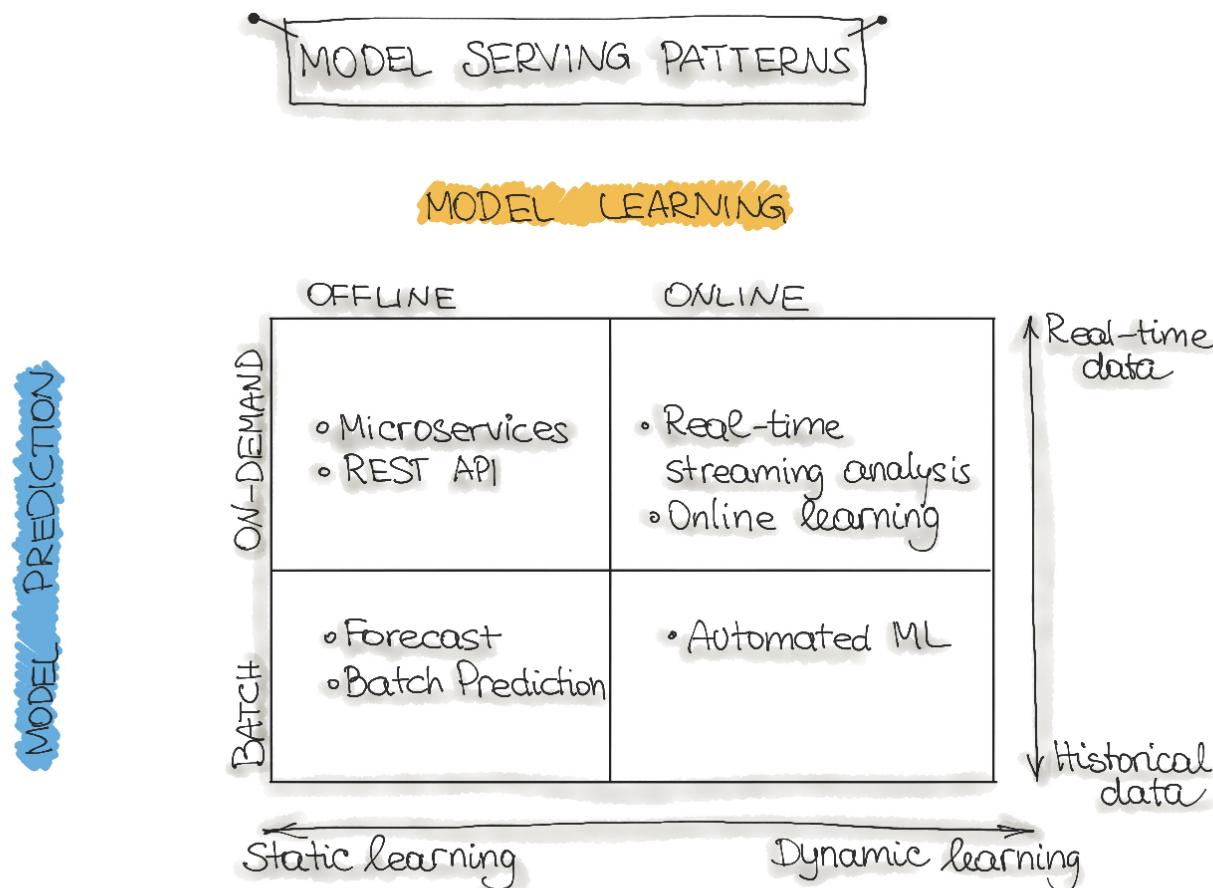
Model Prediction Patterns

- Two modes to denote the mechanics of the ML model to makes predictions
 - Batch predictions
 - Real-time predictions
- Batch predictions
 - The deployed ML model makes a set of predictions based on historical input data
 - often sufficient for data that is not time-dependent, or when it is not critical to obtain real-time predictions as output
- Real-time predictions (aka on-demand predictions)
 - Predictions are generated in real-time using the input data that is available at the time of the request



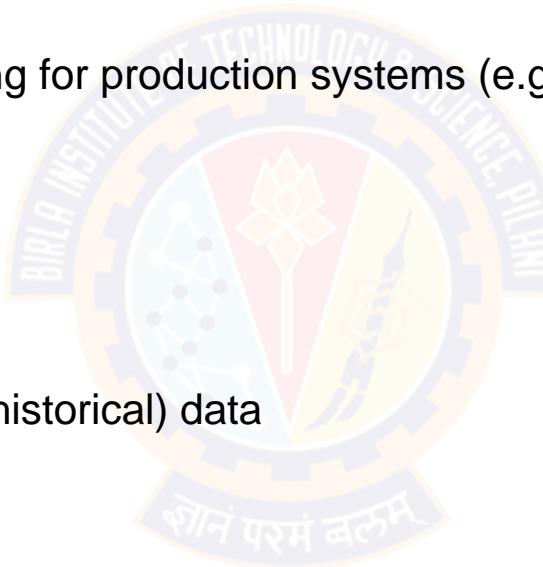
ML architecture patterns

Forecast, Web-Service, Online Learning, and AutoML



Forecast

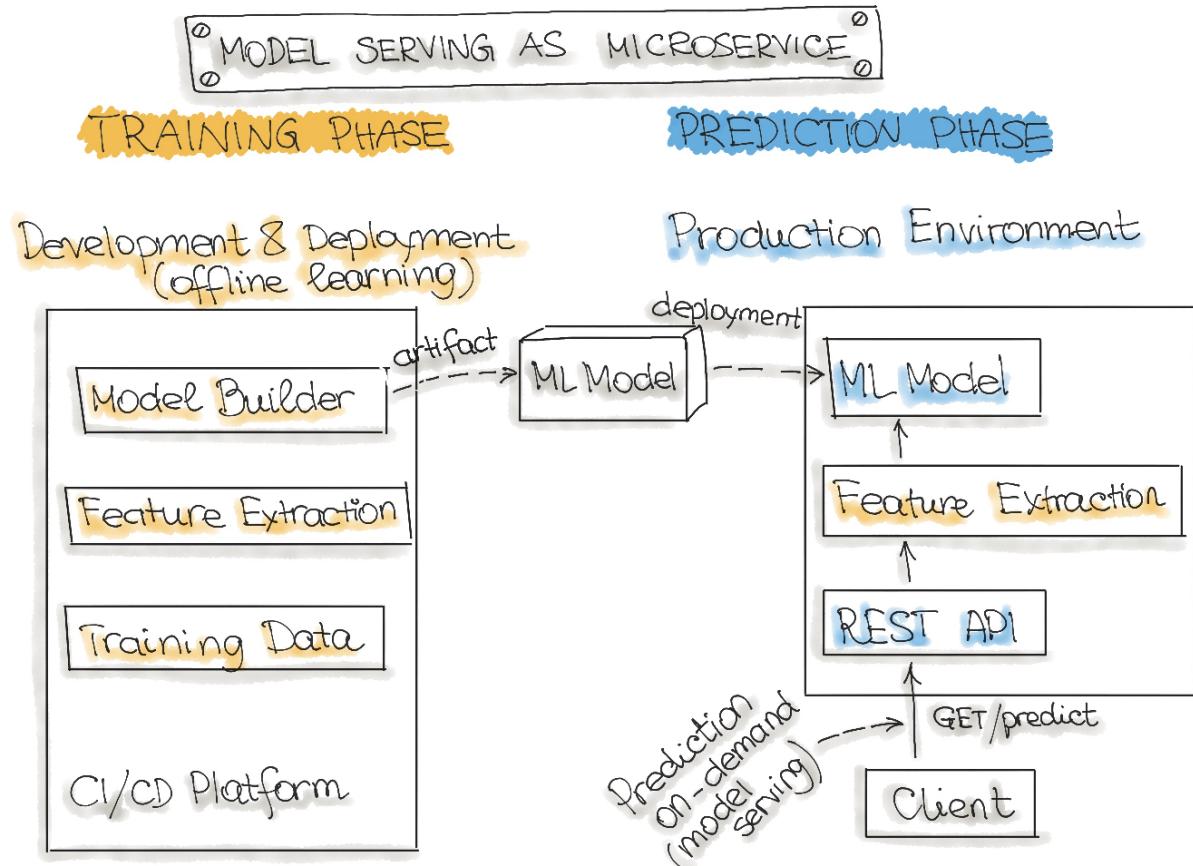
- Widely spread in academic research or data science education (e.g., Kaggle or DataCamp)
 - used to experiment with ML algorithms and data as it is the easiest way to create a machine learning system
 - not very useful in an industry setting for production systems (e.g. mobile applications)
- Usually involves steps
 - take an available dataset
 - train the ML model
 - run this model on another (mostly historical) data
 - makes predictions



Web-Service (or Microservices)

Architecture for wrapping trained models as deployable services

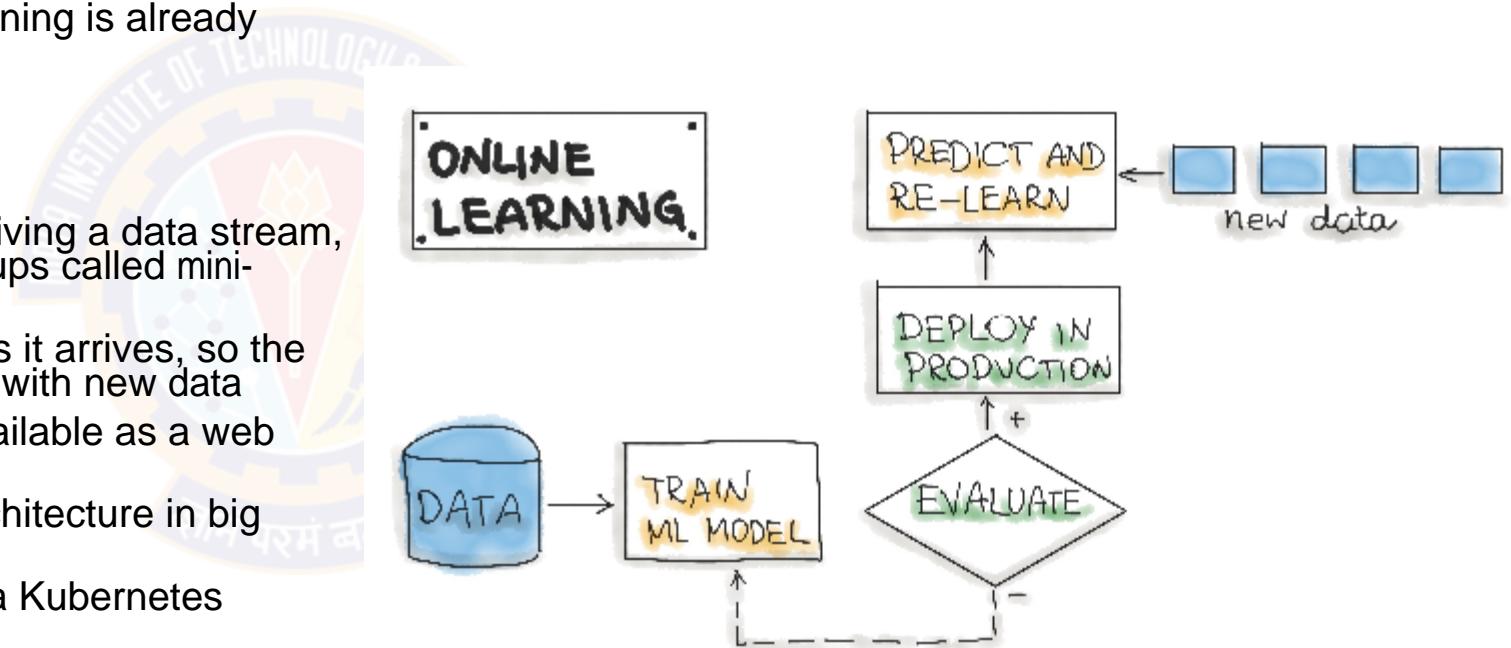
- The most commonly described deployment architecture for ML models
- The web service takes input data and outputs a prediction for the input data points
 - Model is trained offline on historical data, but it uses real-live data to make predictions
 - Model remains constant until it is re-trained and re-deployed into the production system.
- The difference from a forecast (batch predictions) is that
 - the ML model runs near real-time
 - handles a single record at a time instead of processing all the data at once



Online Learning (Real-time streaming analytics)

Most dynamic way to embed machine learning into a production system

- Confusing name because the core learning or ML model training is usually not performed on the live system
 - call it incremental learning- term online learning is already established within the ML community
- In this type of ML workflow
 - ML learning algorithm is continuously receiving a data stream, either as single data points or in small groups called mini-batches
 - System learns about new data on the fly as it arrives, so the ML model is incrementally being re-trained with new data
 - Continually re-trained model is instantly available as a web service
 - Technically works well with the lambda architecture in big data systems
 - Model would typically run as a service on a Kubernetes cluster or similar
- A big difficulty with the online learning system in production is that
 - if bad data is entering the system, the ML model, as well as the whole system performance, will increasingly decline



AutoML

Sophisticated version of online learning

- AutoML is getting a lot of attention
 - considered the next advance for enterprise ML
 - promises training ML models with minimal effort and without machine learning expertise
 - User needs to provide data, and the AutoML system automatically selects an ML algorithm and configures the selected algorithm
 - very experimental way to implement ML workflows
 - usually provided by big cloud providers, such as Google or MS Azure
 - Instead of updating the model, need to execute an entire ML model training pipeline in production that results in new models on the fly



Thank You!

In our next session: