ARTIFICIAL INTELLIGENCE SOFTWARE DEVELOPMENT

Week 5 Lecture 1

Dr. Hari M Koduvely

MLOPS

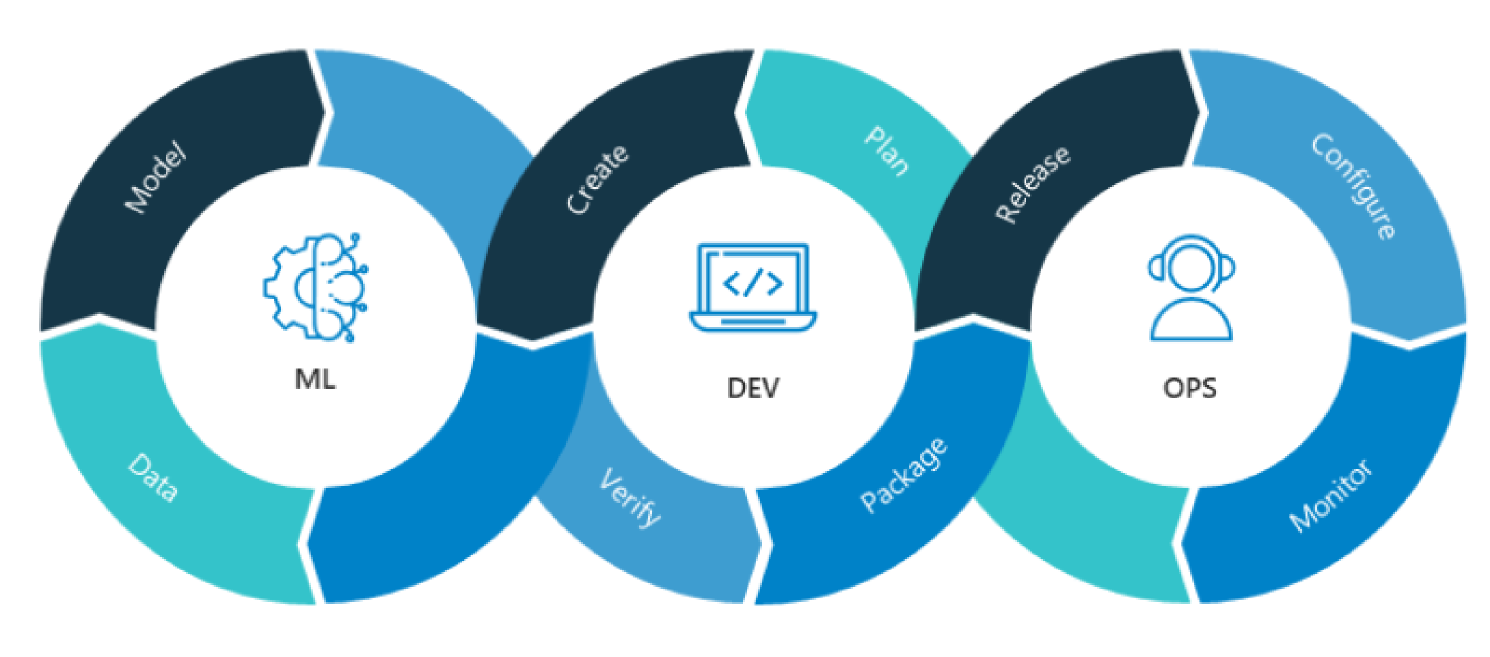


Image Source
https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/

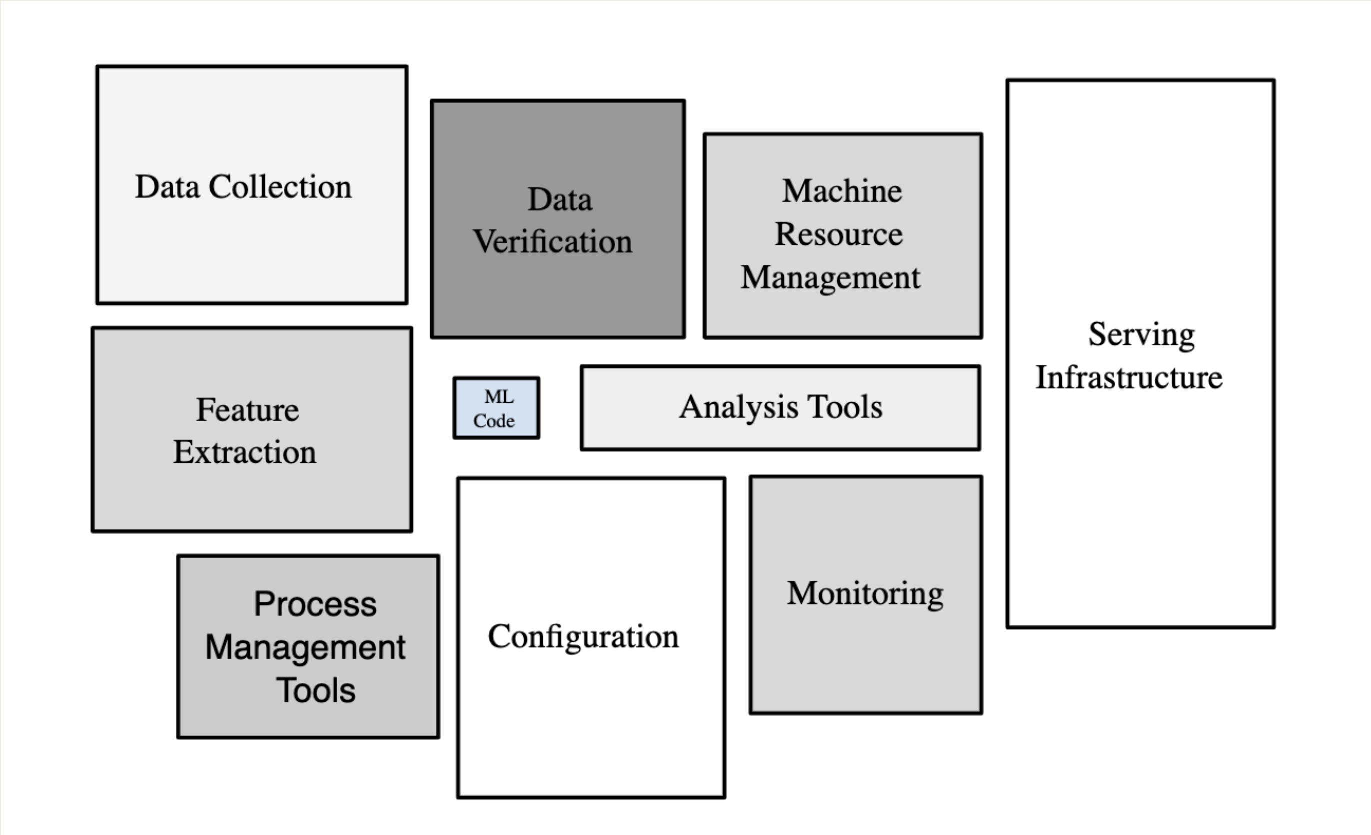
References

- Designing Machine Learning Systems – Chip Huyen

What is MLOps ?

- MLOps is a set of Tools and Best Practices for bringing ML into production
- Similar to DevOps – Developments and Operations
- Treats ML System Holistically

ML System Overview



- Real world production ML Systems are large software ecosystems
- ML Agorithm or Model code is only < 10% of all the code

Image source https://developers.google.com/machine-learning/crash-course/production-ml-systems

ML in Research vs Production

	Research	Production
Requirements	Best model performance on benchmark datasets	Depends on the Stake Holder
Computational Priority	Fast Training, High Throughput	Fast Inference, Low Latency
Data	Static	Constantly Changing
Ethical Aspects	Often not a focus	Must be considered
Interpretability	Often not a focus	Must be considered

ML in Research vs Production - Requirements

- Production requirements vary from stakeholder to stakeholder
- *e.g.* Mobile app recommending restaurants to users.
  - ML Engineers want a models that provides good quality recommendations
  - Sales team wants a model that recommends more expensive restaurants
  - Product team wants a model that returns recommendations in < 100 ms
- Two different objectives:
  - Recommending restaurants that are most likely to be clicked by users
  - Recommending restaurants that brings more revenue to app

ML in Research vs Production - Requirements

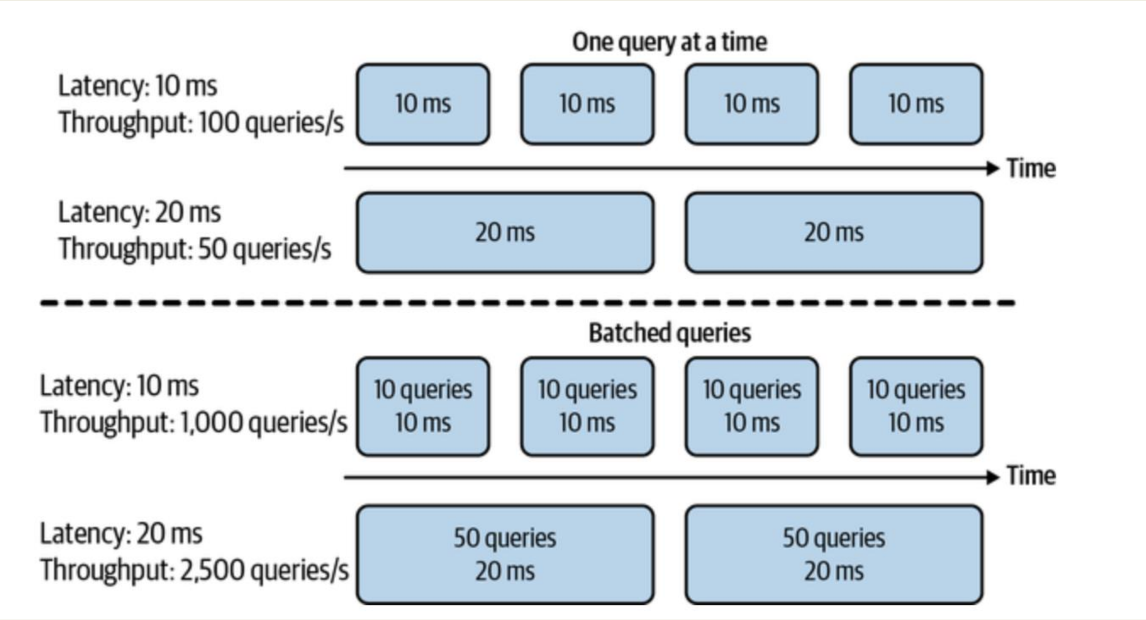
- Understand the ***strict requirements vs good to have***
  - Latency could be a strict requirement
  - Quality of recommendations could be a good to have
- Understand the ***impact of performance improvements***
  - 0.1 % increase in CTR for online ads can increase the revenue significantly
  - 0.1 % increase in image classification accuracy is not very significant
- Understand the ***impact of model complexity***
  - Ensemble models are commonly used to improve model performance
  - Ensemble models have higher computational costs and less interpretability

ML in Research vs Production - Computational Priority

- Latency Vs Throughput
  - Latency is the time between receiving an inference request to returning the results
  - Throughput is the number of inference requests processed in a specific amount of time
- For systems processing one request each time:
  higher latency => lesser throughput
- For systems that process requests in a batch:
  higher latency => higher throughput

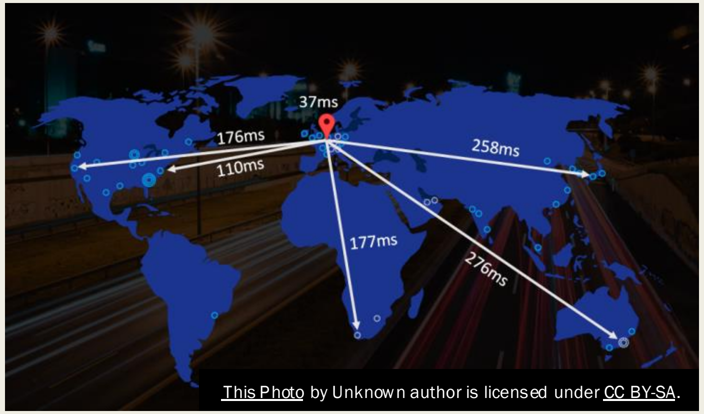
ML in Research vs Production - Computational Priority

- Latency Vs Throughput



ML in Research vs Production – Computational Priority

- Latency is very important factor for a good customer experience
  - Increase of 30% in latency can reduce conversion rates by 0.5% (Booking.com 2019)
  - 50% of the mobile users will leave a page if it takes more than 3 secs to load (Google 2016)



This Photo by Unknown author is licensed under CC BY-SA.

ML in Research vs Production - Data

- Research datasets are often clean and well formatted
- Many of them are standard benchmark datasets used by several researchers
- Issues about the datasets are known and often fixed
- Scripts to process them are easily available
- Production datasets are:
  - Messy, noisy
  - Not structured
  - Biased, constantly shifting
  - Issues are not fully known or documented
  - Privacy and confidential information exposed
  - Partially labeled, imbalanced classes
  - Constantly generated by Users, Systems and Logs

ML in Research vs Production - Data

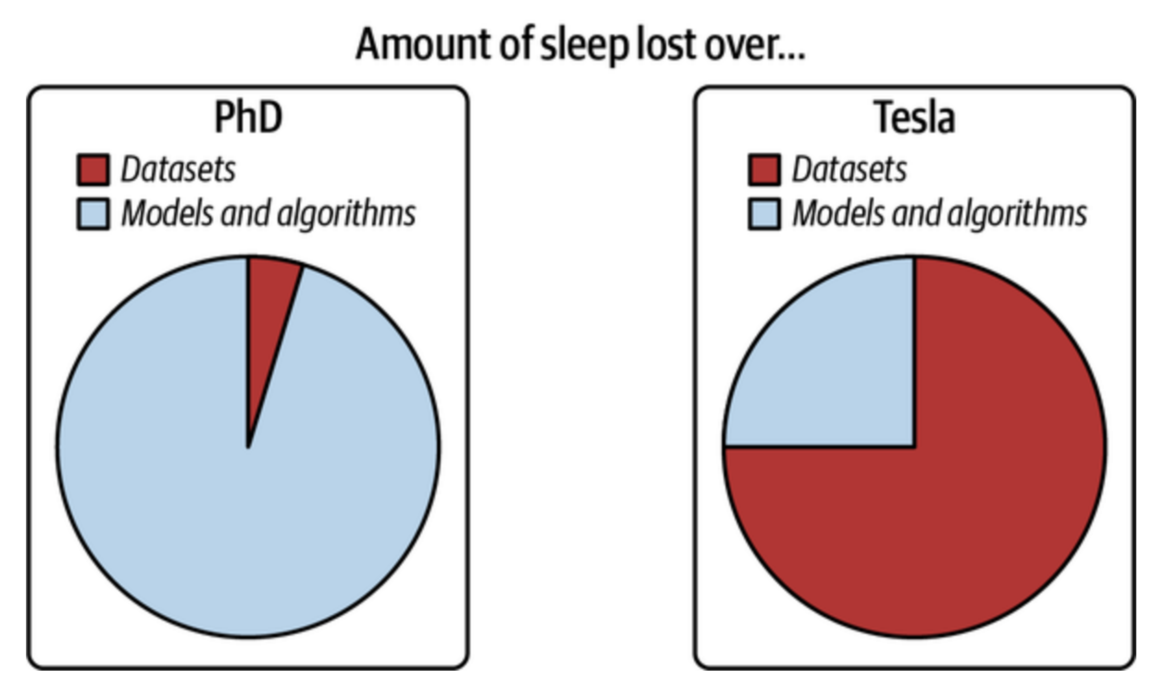


Figure 1-5. Data in research versus data in production. Source: Adapted from an image by Andrej Karpathy[24]

ML in Research vs Production - Ethical Aspects

- During research phase models are rarely used on people:
  - Ethical aspects are overlooked or
  - Their implementations are postponed to production stage
- Monitoring for ethical aspects in production alone is not sufficient
- Some examples :
  - Rejection of loan application
  - 1.3 million creditworthy Black and Latino people have been rejected loans between 2008 and 2015 (Berkely study, 2019)

ML in Research vs Production - Ethical Aspects

- During research phase models are rarely used on people:
  - Ethical aspects are overlooked or
  - Their implementations are postponed to production stage
- Monitoring for ethical aspects in production alone is not sufficient
- Some examples :
  - Rejection of loan application
  - 1.3 million creditworthy Black and Latino people have been rejected loans between 2008 and 2015 (Berkely study, 2019)
  - When racial identifying features were removed from model, their mortgage applications were accepted

ML in Research vs Production - Interpretability

- Question: Who would you choose between?
  - A Human surgeon who cures 80% of cancer patients
  - A Black Box AI surgeon who cures 90% of cancer patients
- Interpretability is important to understand why a certain decision was made
- It will help to build trust among users
- It can expose potential biases
- Important for developers to debug and improve models
- It can be hard to interpret models such as deep neural networks and ensemble models

ML Systems Vs Traditional Software

- Why not just use the proven best practices from software development to ML systems development?
- ML System pipelines are different from software development pipelines

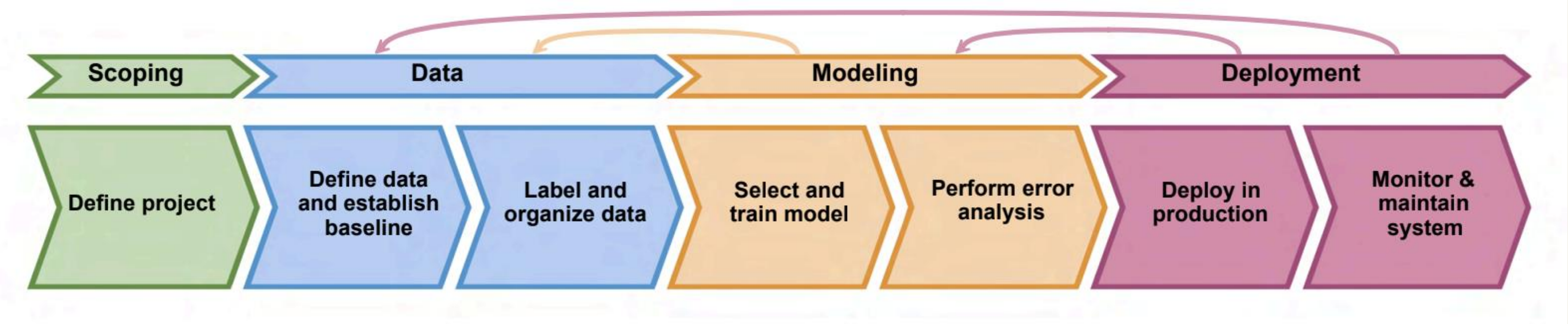
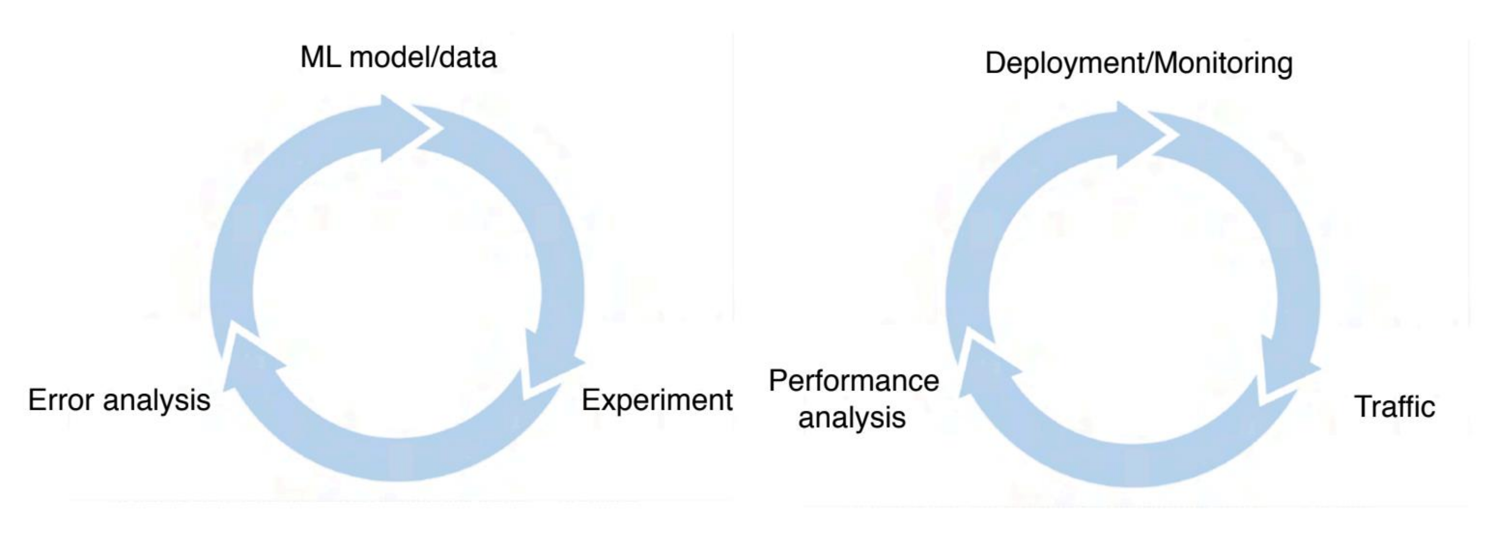


Image Source: DeepLearning.AI

ML Systems Vs Traditional Software

- Why not just use the proven best practices from software development to ML systems development?
- Just as ML model development is iterative so is model deployment

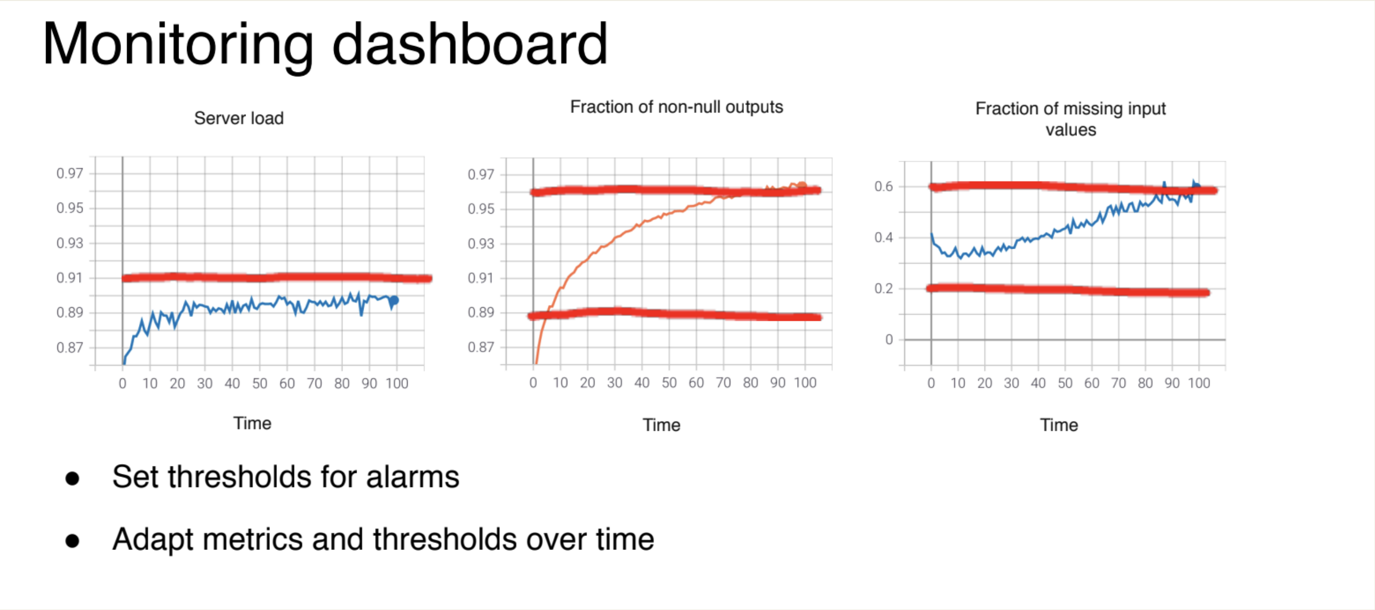


Iterative process to choose the right set of metrics to monitor.

Image Source: DeepLearning.AI

ML Systems Vs Traditional Software

- Why not just use the proven best practices from software development to ML systems development?
- Just as ML model development is iterative so is model deployment



Monitoring dashboard

- Set thresholds for alarms
- Adapt metrics and thresholds over time

Image Source: DeepLearning.AI

ML Systems Vs Traditional Software

- Why not just use the proven best practices from software development to ML systems development?
- Many challenges are unique to ML Systems and it requires unique tools
  - Traditional software development assumes Data and Code are separated
  - ML Systems are part code, part data and part generated from both
  - Systems can be improved by improving the data and not code
  - Need to be adaptive to changing environments
  - Need to do testing and versioning of Data also
  - Not all data samples are equal, some are more valuable than others
  - Model sizes are large to load on to RAMs
  - Concept Shift and Data Drift