Session -9

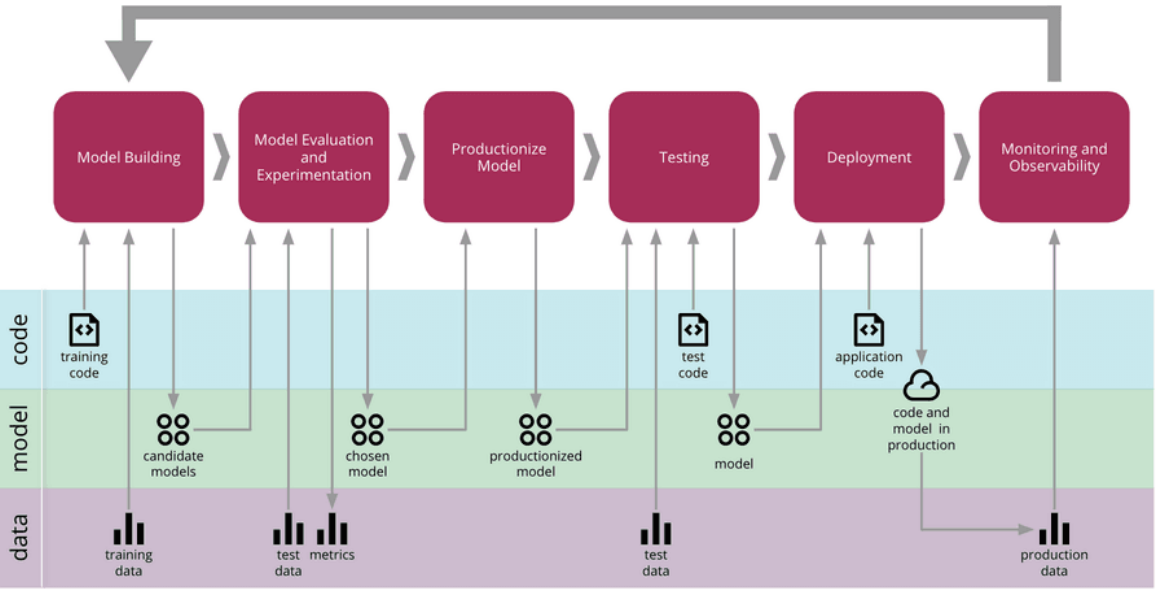
# ML - System Design -1

June 23, 2025



## Agenda

* What does ML system design includes.




60 → EDA / Data cleaning

10% → model dev

30 → maintenance

## Things to observe post deployment:

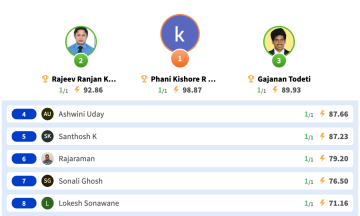
1. Online Accuracy Metrics
2. Latency shouldn't increase beyond the threshold.
3. Server Load (typically this should be less than 50%)
4. Memory Consumption (should be below threshold)
5. Throughput - (Requests you can serve per second, this shouldn't decrease, beyond a certain threshold)

**Which of these is NOT a system health/performance metric typically monitored after deploying an ML model?**

1 user has participated

| | | |
|---|---|---|
| A | Requests per second | 0% |
| ✓ B | Accuracy of the model's predictions | 100% |
| C | CPU/memory/data utilization | 0% |
| D | Latency per request | 0% |

End Quiz Now



**What is the purpose of the 'Model Evaluation and Experimentation' phase in the ML model lifecycle?**

0 users have participated

| | | |
|---|---|---|
| A | Deploying the model to a production environment | 0% |
| B | Preparing data and initial code formulation | 0% |
| ✓ C | Feature selection, hyperparameter tuning, and algorithm comparison | 0% |
| D | Monitoring and observing the model in a live environment | 0% |

End Quiz Now

Based on all quizzes from the session

| | | |
|---|---|---|
| 1 | Phani Kishore R ... | 2/2 ⚡ 190.00 |
| 2 | Rajeev Ranjan K... | 2/2 ⚡ 187.10 |
| 3 | Santhosh K | 2/2 ⚡ 182.43 |
| 4 | Ashwini Uday | 2/2 ⚡ 177.76 |
| 5 | Gajanan Todeti | 2/2 ⚡ 177.73 |
| 6 | Sonali Ghosh | 2/2 ⚡ 163.17 |
| 7 | Arindam Maji | 1/2 ⚡ 85.00 |
| 8 | Chirag Bhatt | 1/2 ⚡ 82.73 |
| 9 | Rajaraman | 1/2 ⚡ 79.20 |
| 10 | Lokesh Sonawane | 1/2 ⚡ 71.16 |

# Drift in Data/Model

→ Something has change → (1) Target (2) Features.

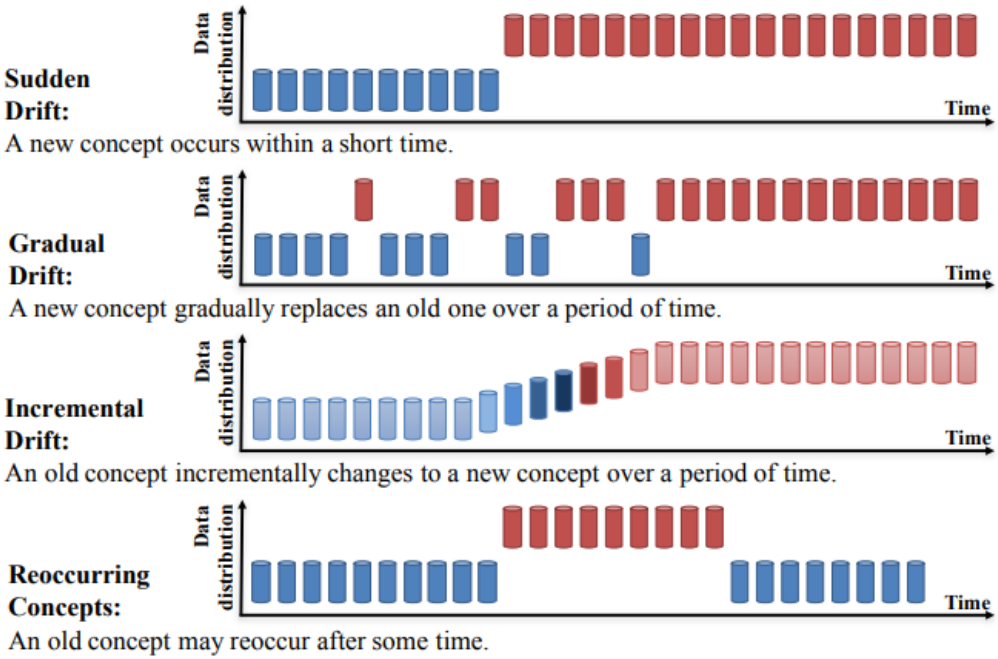
(1) Concept drift: properties of target has changed.

(a). Fraud detection

Soln: Train the model with more recent data

(2) Data drift: Properties of input feature has changed.

Soln: Engineer new features.




## How to Notice Drift

1. Visual Inspection: Input data over time
   1. distribution changes
      → mean
      → median (over months)
   2. Plot data & See if there's big chang

2. Statistical test For Same column over diff periods

② Model performance monitoring – Key offline metrics

① Over past 6 months

**What is drift in the context of machine learning?**

1 user has participated

| | | |
|---|---|---|
| A | The movement of an autonomous vehicle | 0% |
| B | The physical displacement of a computing server | 0% |
| ✓ C | The change in the statistical properties of model data over time | 100% |
| D | The time it takes for a model to make a prediction | 0% |

End Quiz Now

Based on all quizzes from the session

| | | |
|---|---|---|
| 2 | Rajeev Ranjan K... | 3/3 ⚡ 279.30 |
| 1 | Phani Kishore R ... | 3/3 ⚡ 288.43 |
| 3 | Gajanan Todeti | 3/3 ⚡ 269.63 |
| 4 | Ashwini Uday | 3/3 ⚡ 269.13 |
| 5 | Santhosh K | 3/3 ⚡ 247.47 |
| 6 | Arindam Maji | 2/3 ⚡ 178.37 |
| 7 | Rajaraman | 2/3 ⚡ 173.87 |
| 8 | Sonali Ghosh | 2/3 ⚡ 163.17 |
| 9 | Lokesh Sonawane | 2/3 ⚡ 162.70 |
| 10 | Sayyid Thajudheen Thangal K C | 2/3 ⚡ 130.99 |

January

train

test

model

train

Prediction

test

MSE

$R^2$

Feb

Prediction

MSE

$R^2$

After observing change in model metrics:

1. First check if my features have changed, their distribution.
2. If they've changed, you know the perp, otherwise it's concept drift.

**What type of drift occurs when there's a change in the relationship between input data 'x' and output data 'y'?**

13 users have participated

| | | |
|---|---|---|
| A | Data drift | 46% |
| B | Business drift | 0% |
| C | Algorithm drift | 8% |
| ✓ D | Concept drift | 46% |

End Quiz Now

Leaderboard

Based on all quizzes from the session

| | | |
|---|---|---|
| 2 | Ashwini Uday | 4/4 359.02 |
| 1 | Phani Kishore R ... | 4/4 386.52 |
| 3 | Santhosh K | 4/4 342.66 |
| 4 | Rajeev Ranjan Kumar Singh | 3/4 279.30 |
| 5 | Gajanan Todeti | 3/4 269.63 |
| 6 | Rajaraman | 3/4 267.87 |
| 7 | Sonali Ghosh | 3/4 251.83 |
| 8 | Lokesh Sonawane | 3/4 248.03 |
| 9 | Arindam Maji | 3/4 243.26 |
| 10 | Sagar Deka | 2/4 169.00 |

# MULTIPROCESSING

Python applications → Process → multiple workers (thread)

→ Reserves RAM

" compute

$P_1$ | $P_2$

global Interpreter Lock,

GIL

Process 1

thread 1

thread 2

Process 2

thread 1

thread 2

Inside a process only one thread is active at any instan

In C / C++ / Java → a process

multiple threads

can execute

concurrently

multi-processing

multi-threads

multi-processes

Website – 1K images

$T_1 \xrightarrow{req} Server_1 \quad Im_1$

$T_2 \xrightarrow{req} Ser$

$\vdots$

$T_{1000} \xrightarrow{req_1} Ser$

compiled obj