Data Drift (ML concept)

- Data drift happens when the data your model sees in real world changes compared to the data it was trained on (the world changed, but your model didn't)
  - your models assumptions about what is normal is no longer true leading to worse predictions or accuracy dropping over time

- Ex (housing price model)
  - you trained a model in 2020 on housing data where:
    - intrest rates where 1-2% • avg house price was 900$ per sqrt Foot
  - Now its 2025:
    - intrest rate is 6.5% • Construction cost doubled • new neighbours

=) The relationship between "sqr Feet" and "price" changed
that's data drift - your models world changed

* types of drift	Discription	Example
covariate drift	input features (x) changes but relationship x→y might stay same	Different age distribution of patients, but same desease pattern
Lable drift (Prior drift)	target lable (y) changes Frequency	more Fraudulent transactions then before
Concept Drift	relationship between X and y changes	Bigger houses used to mean higher prices, mostly not true now

- how its detected
  - minor Feature distribution over time using statistical test (kolmogorov-test)
  - compare model preformance metrics week-week • monitor input Features mean/variance
  - Use population stability index (PSI) or KL-divergence to see how distribution changes

- How to Fix
  1) Detect early (monitor input data/predictions) 2) retrain model periodicall w new data
  3) Revalidate Features - some may no longer be predictive 4) use adaptive or online learning (learning models that update incrimentally) ↗ Not Req