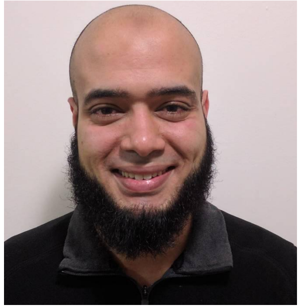
Machine Learning

Data Distribution Shift

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



ML Failures in Production

- Maybe 50% or more of ML pipelines come from normal SWE errors
- However there are specific ML failures
    - Real data is way bigger than your dataset
        - Hence, we break the **‘same distribution’** assumption
    - We assume the data is **stationary**
        - In many problems the data changes over time!
        - This cause “**Data Distribution Shift**” Challenge
    - Your system fails in **edge cases** in critical safety apps (e.g. autonomous driving)
        - Edge cases (rare but possible scenarios) != outliers (significantly different from rest)
    - You failed to properly monitor your system / update when necessary
    - A degenerate feedback loop!

Degenerate Feedback Loop

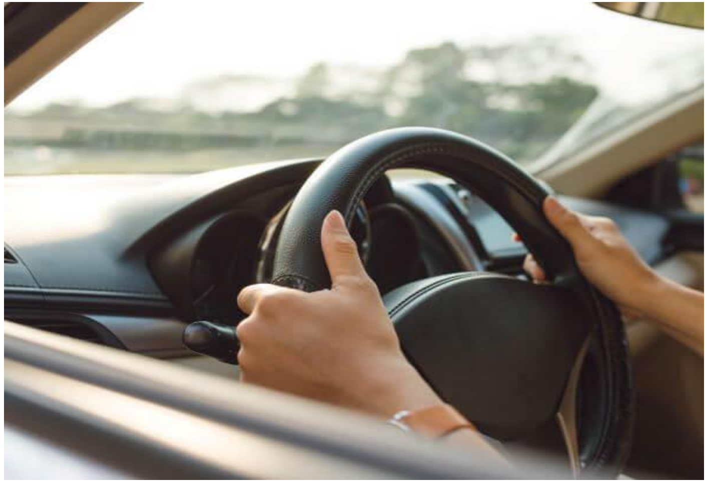
- A degenerate feedback loop in machine learning refers to a situation where the model's **outputs** influence **future inputs** in a way that degrades the model's performance over time
    - Natural labels from **Recommender Systems**: user interacts with your very top suggestions
        - Ending up the system keep favouring them with big gap than others
    - **Policing Algorithms: “**predict crime in various neighborhoods and police are then dispatched based on these predictions Officers arrest more people and the model with new inputs keep focusing in these crimes in these areas”
    - **Credit Scoring**: “A model that predicts lower creditworthiness for individuals from a particular demographic could lead to **fewer loans** being extended to those individuals, which in turn could affect **credit-building opportunities**”
- Large language models probably will suffer from the loop in future!
- One needs to think in ways to fight this loop with natural labels
    - For example, some randomization effect?

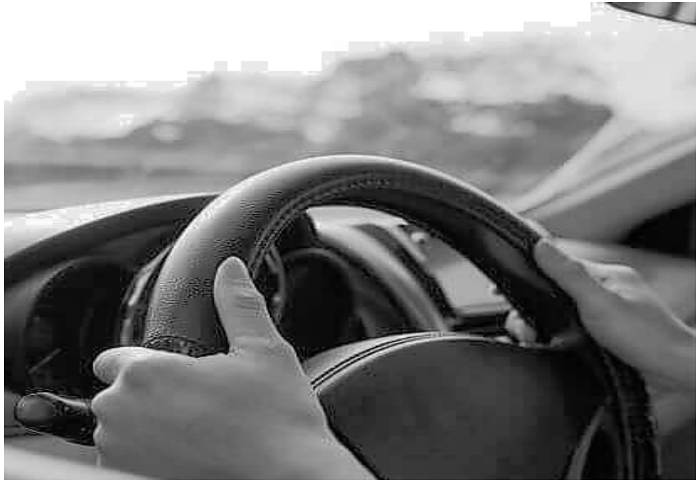
Data Distribution Shift

- A production challenge where the statistical properties of the data change over time as the result the model performance degrades!
- There are several types of shifts
  - Covariate Shift
  - Label (Target) Shift
  - Concept Drift
- Let’s first introduce what is a distribution change: input and output

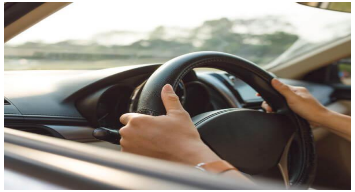
Compare these 2 images

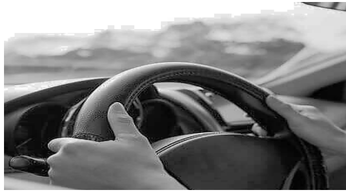
- In the 'Hands-on-wheel' context
  - Compare the input
  - Compare the model's expected output





Distribution Shift





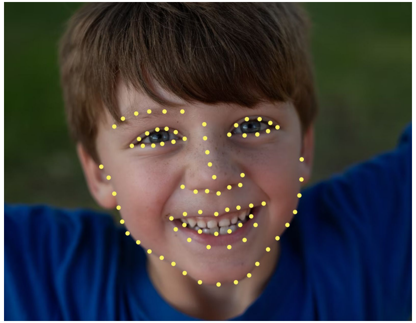
- Input Distribution Shift
  - The input itself (feature space) is changed
  - Imagine You trained your vision model with camera XX1234
  - After 2 years, the camera supplier stopped this line of production.
    - We used a new camera that generates gray-like images and smaller field of view
  - Although the image is still classified e.g. hands on wheel, there is input features change
- Output Distribution Shift
  - Imagine you oversampled your data that was 10% cats and 90% dogs into 50% cats and 50% dogs
  - This means the distribution of the labels got changed
  - This has implications on multi-label classifier (see calibration later)

Covariate Shift

- When the distribution of input features changes, **but** the conditional distribution of the target given the inputs **remains the same**.
  - In self-driving, this might happen when the model encounters different types of roads or lighting conditions than those it was trained on.
- In below example, both are hands on (no output change) but the input changed severely

Label Shift

- When the output distribution changes but the input doesn't, we call it **label shift**, such as oversampling case
- Example: Facial Emotion Recognition
  - We use facial landmarks to represent faces
  - We can train a model to extract facial landmarks and classify their emotion (e.g. happy/angry)
  - The model is deployed in a bad supermarket during a recession where many people have faces showing anger or frustration
  - The distribution of emotions (**labels**) changes in the deployment scenario, even though the features (facial expressions) are the same types as in the training data.



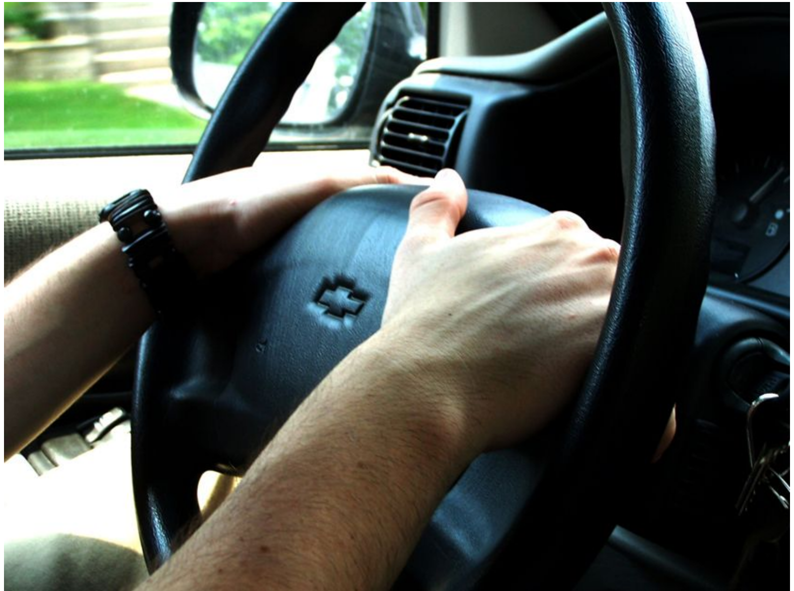
Img src

Concept Drift

- The **relationship** between the features and the target variable changes over time
    - Due to some crisis (e.g. floods, COVID), the prices of homes changed severely and house predictor models fail
    - Assume the dataset has home X for 100,000, but now it become for 1000,000
        - We need to update our dataset to reflect the
- Seasonal Data
    - In some problems, prices could change in cycles
        - Weekends e.g. in rideshare prices
        - Special events sales like Black Friday
    - In problems where summer and winter sales are different

Concept Drift

- Another example: Assume the customer said when the user put his hands on the spoke area (see below), this is hands on
- Later, after deploying the system, the customer said this is actually wrong, we need them to be hands off
- The hands on/off concept changed
- We need to relabel our dataset



Misc

- **Detection**: we might monitor the features statistics (e.g. min, max, percentiles). Nowadays frameworks support that
  - Sometimes, it is good to monitor changes **over a time window**
- **Mitigation**:
  - we might try to use weighting strategy
  - Train on large dataset or fine-tune
  - Human-in-the-Loop: Use human experts to validate model decisions, especially when the model is uncertain
- **Domain Adaptation**: To shorten a known gap
  - You can’t get data from a private military area. Train your object detector in a way it can adapt to the new data distribution
- **Model Re-calibration** for probabilities if needed

Selection Bias Shift

- This is a different story
- Data Collection: some **selection sample bias** makes train data shifted toward something
    - Collected data from medical research facility that deals primarily with **severe** cases
    - Later, you deployed the disease model in public hospitals (more normal cases)
- Active Learning Process: the **selection process** focuses on the most informative samples so we end up with a mismatch

“Acquire knowledge and impart it to the people.”

“Seek knowledge from the Cradle to the Grave.”