

Machine Learning Model Validation Framework

Overview of model validation framework and techniques for machine learning models.



The motivation and the topics

- The objective is not to create a new model validation structure but to enhance the existing risk model validation framework with insights on machine learning models. This enhancement will address specific challenges associated with ML and might also consider issues that are shared with conventional modeling approaches.
- While the European Union (with GDPR) and the United States have made significant strides in addressing bias and fairness, these concerns haven't been as pronounced in the UAE. Nevertheless, considering the financial and reputational setbacks experienced by prominent tech firms, it's timely to investigate these matters
- Data-related challenges are prevalent in banking, including issues of privacy, quality, bias, drift.

What is model validation and Why do we need it?



Ensure model operates as expected

Model validation verifies the model performs according to specifications and requirements.



Evaluate risks

Model validation identifies potential risks like bias and monitors model performance over time.



Meet business needs

Validation guarantees the model aligns with and achieves business objectives.



Manage risks

It puts controls in place to mitigate risks that could impact model performance or lead to issues.

In summary, model validation is crucial to ensure models operate correctly, evaluate risks, meet business goals, and manage those risks effectively, mandated by cbuae

What is model validation and Why do we need it? - II

MPW · AMAZON

Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

BY DAVID MEYER

October 10, 2018 at 2:00 PM GMT+4



The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

 Share full article



 168

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

WASHINGTON — The Department of Housing and Urban Development [sued Facebook on Thursday for engaging in housing discrimination](#) by allowing advertisers to restrict who is able to see ads on the platform based on characteristics like race, religion and national origin.

acmqueue

Discrimination in Online Ad Delivery

Google ads, black names and white names, racial discrimination, and click advertising

Latanya Sweeney

Do online ads suggestive of arrest records appear more often with searches of black-sounding names than white-sounding names? What is a black-sounding name or white-sounding name, anyway? How many more times would an ad have to appear adversely affecting one racial group for it to be considered discrimination? Is online activity so ubiquitous that computer scientists have to think about societal consequences such as structural racism in technology design? If so, how is this technology to be built? Let's take a scientific dive into online ad delivery to find answers.

"Have you ever been arrested?" Imagine this question appearing whenever someone enters your name in a search engine. Perhaps you are in competition for an award, a scholarship, an appointment a promotion, or a new job, or maybe you are in a position of trust, such as a professor, a physician, a banker, a judge, a manager, or a volunteer. Perhaps you are completing a rental application, selling goods, applying for a loan, joining a social club, making new friends, dating, or engaged in any one of hundreds of circumstances for which someone wants to learn more about you online. Appearing alongside your list of accomplishments is an advertisement implying you may have a criminal record, whether you actually have one or not. Worse, the ads may not appear for your competitors.

Why AI/ML



Improves through experience

Machine learning models improve performance on specific metrics through experience from data.



Data has great value

Data is like oil for ML models, generating value through training.



Widely used including banking

ML is used in many sectors like search, phones, agriculture and banking for efficiency.

Machine learning leverages data to improve through experience and is widely used including in banking. Tools like auto ml (H2o), causal inference(DoWhy) make life easier.

Why ML models need validation



ML models have risks

Depending on the use case, ML models can have different levels of risk



High risk models

Models like credit scoring systems face stricter obligations from regulators

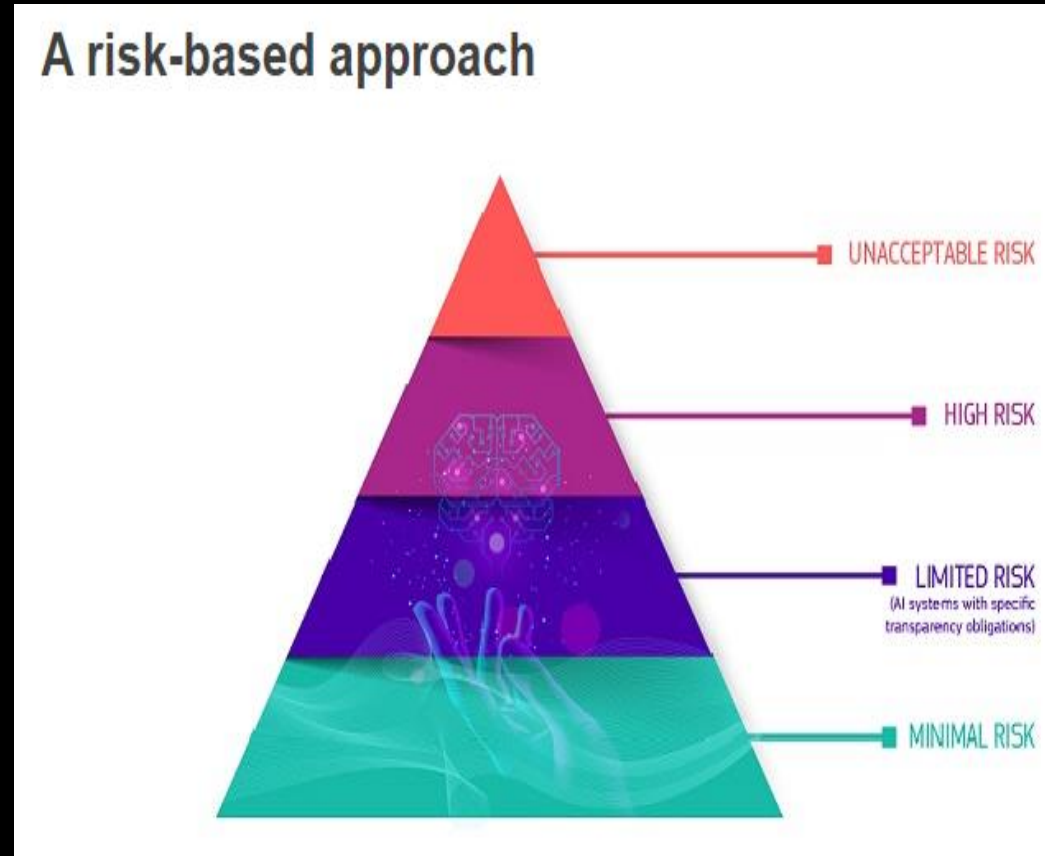


Regulations

Regulators mandate fairness, accountability and transparency(FAT) for high-risk models

ML models, especially high-risk ones, need validation to ensure they meet regulatory and performance requirements including risk mitigating before put to the market

Why ML models need validation



European commission share the risk tier for machine learning product as above figure.

What issues AI/ML present



Data issues

Data collection, quality, sampling, transform, lineage, monitor, privacy

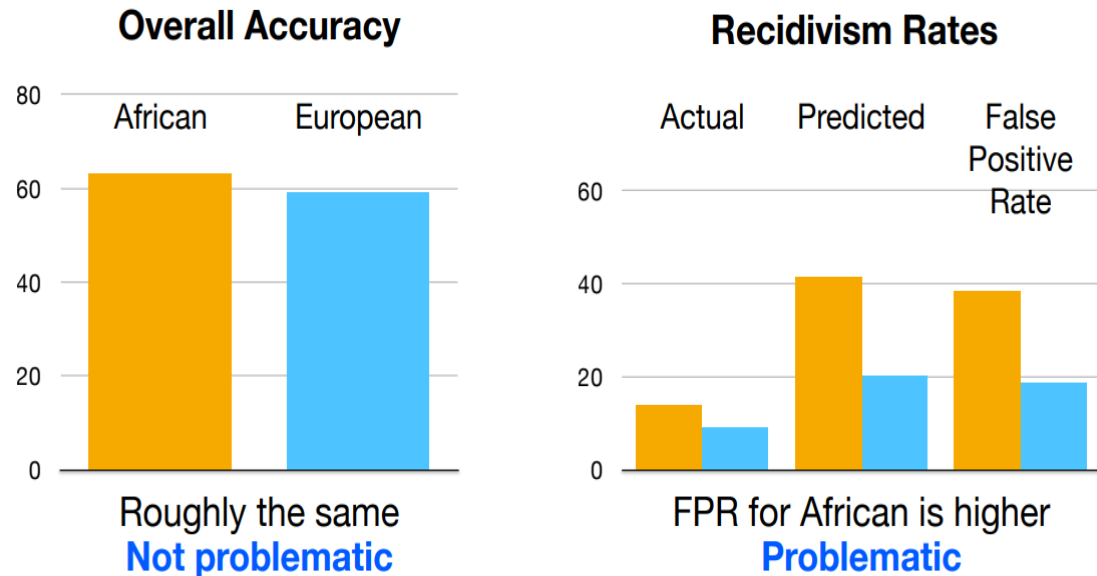


Model issues

Pre-modeling, in-modeling like bias, black box, post-modeling monitor

Proper data and model validation needed throughout the ML pipeline to ensure quality.

Defendants of African descents were often predicted to be more risky than they actually were, and vice versa



* **FPR (false positive ratio)** = ratio of # of actually non-recidivated to # of people predicted to recidivate

Compas Case Study

The Compas recidivism model, developed by Northpointe, aimed to predict defendants' risk of reoffending. However, investigations found that the model was biased against African-Americans, who were often assigned higher risk scores than white defendants with similar profiles.

Data – I

- Data collection bias

Measurement bias, coverage bias, social bias exist in historical data

- Data completeness

Evaluate time period, data sources, distribution, labeling definition

- Data sampling bias

Data may not represent the population if only approved loans available for the outcome (ri)

- Data lineage

Track data transforms to understand how bias propagates

- Data drift

Monitor statistical distribution changes over time

- Data privacy

Techniques like PII, K-Anonymity, Differential Privacy to protect sensitive data

Data Issues II



Insufficient data

If the data set is inadequate and can't represent the whole data set, then training can lead to poor model performance.



Imbalanced data

Class imbalances can affect the model's ability to learn patterns for the minor classes.



Data quality issues

Missing values, outliers, duplicates and noise in the data can impact model performance.

Careful data collection, sampling, cleaning and balancing is critical for building effective machine learning models.

Model - I



Black box nature

AI/ML models are often complex black boxes, making it hard to understand how they arrive at decisions



Internal validation

Validate model on design, hyperparameters, loss function, metrics aligned with business objectives



Explainability

Explain model outputs to ensure they make sense and identify potential biases



Unseen data testing

Evaluate model performance on new, unseen data to ensure robustness

Evaluating AI/ML from multiple angles is crucial to ensure fairness, accountability and transparency.

Model - II

- Methodology alignment

The methodology should align with business needs.

- Hyperparameter tuning

Tune hyperparameters to improve model performance.

- Optimization function

Choose an appropriate optimization function for training.

- Loss function

Select a loss function that matches the business objective.

- Performance metrics

Metrics should align with or proxy the business objective.

- Model stability

Assess model stability via weight confidence and cross-validation.

- Sensitivity analysis

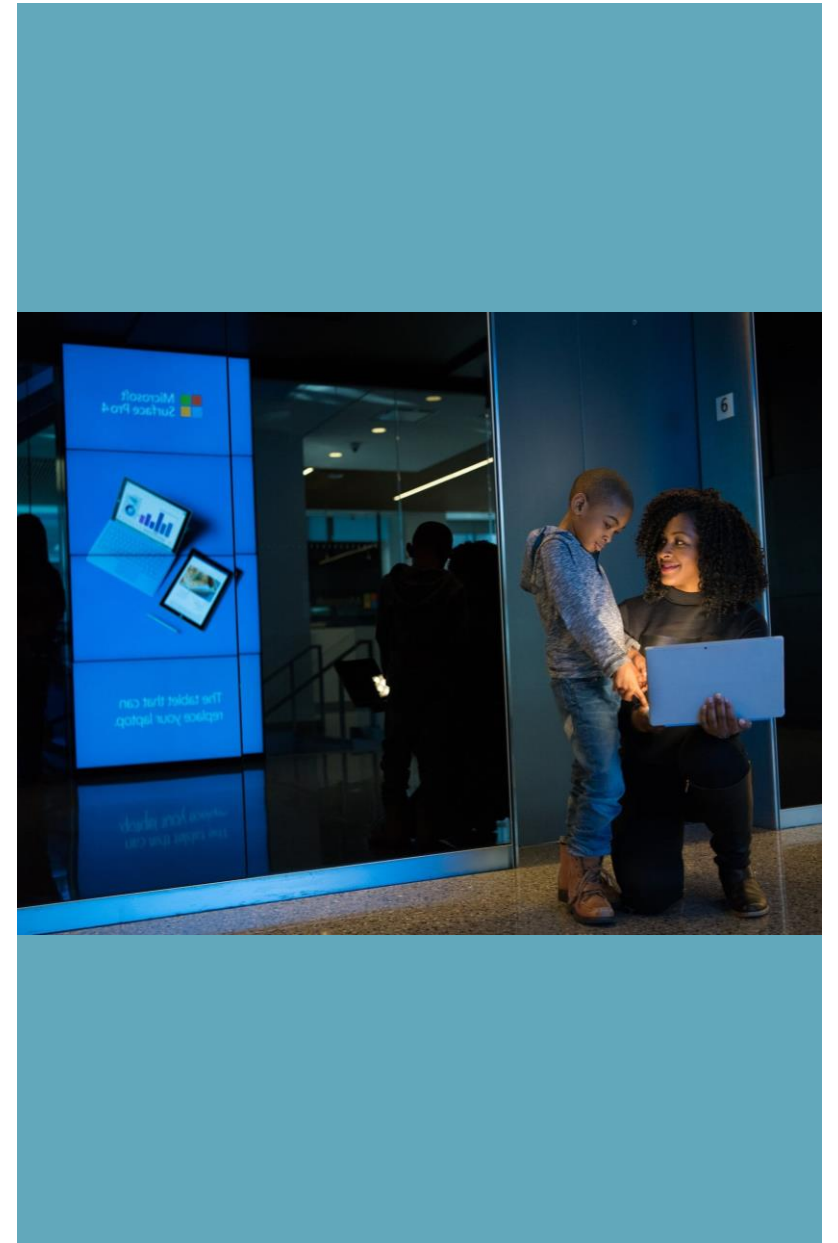
Analyze sensitivity with methods like PDP plots.

- Scenario testing

Simulate scenarios and stress test the model.

Model - III

Model explainability is important in areas like finance and healthcare where decisions can have major impacts on people's lives. It allows humans to understand why a model makes certain predictions and helps identify potential biases.

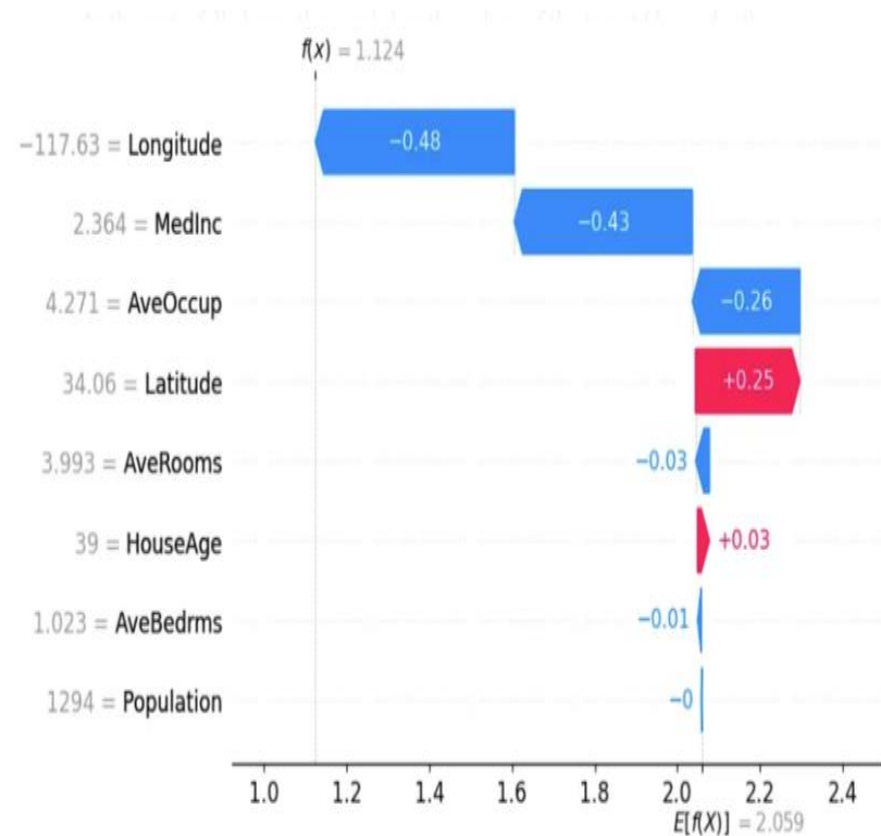
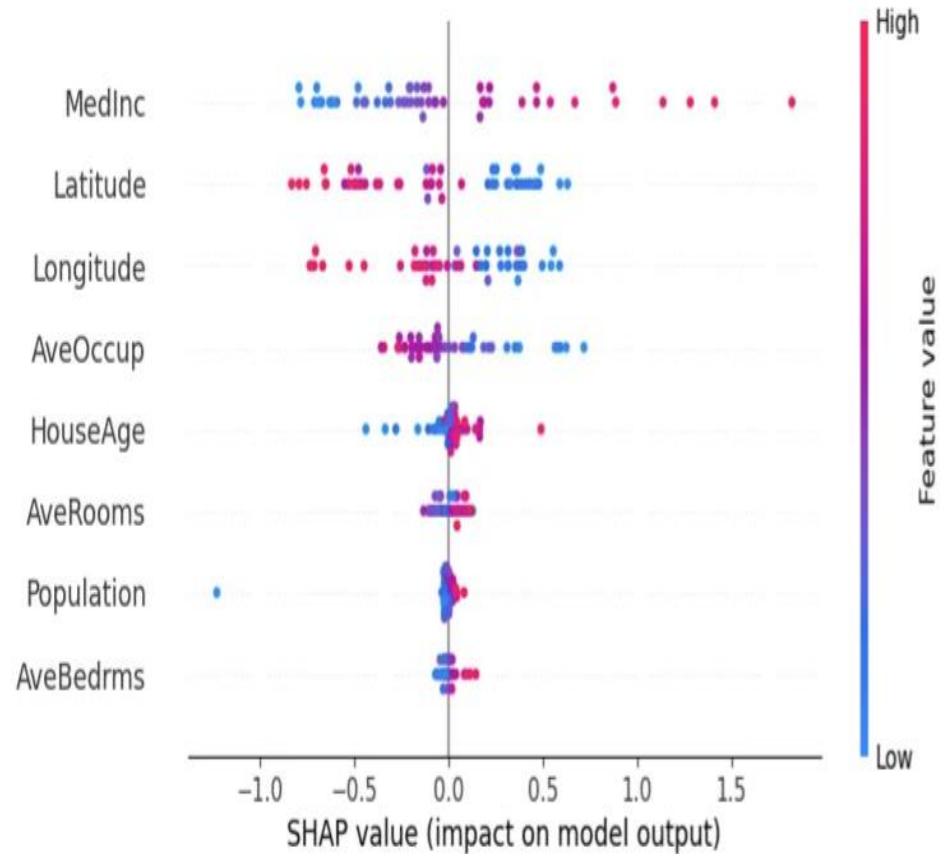


Model – IV

TABLE 1 Examples of approaches to interpretability of prediction regression models

	Global	Local
Model-specific	<ul style="list-style-type: none"> - Decision trees (depends on depth and number of terminal nodes; Hastie, Tibshirani, & Friedman, 2009; Stiglic, Kocbek, Pernek, & Kokol, 2012), - Linear and logistic regression models (Harrell Jr, 2015), - Generalized linear models (GLM) and generalized additive models (GAM; Hastie et al., 2009), - Naive Bayes classifier (Kononenko, 1993), - GNNExplainer (Ying et al., 2019) 	<ul style="list-style-type: none"> - Set of rules (for specific individual; Visweswaran, Ferreira, Ribeiro, Oliveira, & Cooper, 2015), - Decision trees (by tree -decomposition; Visweswaran et al., 2015), - Visual analytics-based approaches (interactive visualization techniques for interpretation focusing on individual prediction), - k-Nearest neighbors (k-NN; depends on the number of important features, retrieving k-nearest neighbors for interpretation; Yuwono et al., 2015), - GNNExplainer (Ying et al., 2019)
Model-agnostic	<ul style="list-style-type: none"> - Different variants of model compression/knowledge distillation/global surrogate models (Elshaw, Al-Mallah, & Sakr, 2019), - Partial Dependence Plots (PDP; Elshaw, Al-Mallah, & Sakr, 2019), - Individual Conditional Expectation (ICE) plots (Elshaw, Al-Mallah, & Sakr, 2019), - Black Box Explanations through Transparent Approximations (BETA; Lakkaraju, Kamar, Caruana, & Leskovec, 2017) - Model understanding through subspace explanations (MUSE; Lakkaraju et al., 2019). 	<ul style="list-style-type: none"> - Local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016), - Shapley additive explanations (SHAP; Lundberg & Lee, 2017), - Anchors (Ribeiro et al., 2018), - Attention map visualization, - Model understanding through subspace explanations (MUSE; Lakkaraju et al., 2019).

Model – V



NOTE: THESE SHAP VALUES ARE ONLY FOR THIS OBSERVATION ONLY, WITH OTHER

Model - VI

Language/Tool Used

Python code converted to Kedro pipeline and deployed using GitHub Actions by data science platform team.

Version Control

GitHub used for distributed version control.

Micro-service/Container

Flask API implemented as service, deployed through Docker containers.

Documentation

Model Documentation using confluence.

Infrastructure

Kedro pipelines and GitHub Actions used for workflow automation.

Model - VII



Decision Monitor (Approval rate, Accept rate etc)

Performance monitor (Default rate)

Data Monitor (Drift detection)

Three lines of defense which is widely used in risk team

Model – VIII






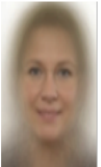
Three Lines of Defence

Model - IX

- Data Poison
- Membership inference attacks
- Model Extraction/Stealing



Error rates (1 - TPR) in a gender prediction from facial images

	darker male 	darker female 	lighter male 	lighter female 
Microsoft	6.0%	20.8%	0.0%	1.7%
IBM	12.0%	34.7%	0.3%	7.1%
Face++	0.7%	34.5%	0.8%	7.1%

Error rates for darker females are generally worse than lighter males



Bias and Fairness

- Bias is systematic errors in the data and/or model that leads to potentially unfair outcomes.
- Bias existed in the industry; even top technology companies also suffered from the bias problem.
- Many bias problems are discussed before, we will brief mention inductive bias.

Bias and Fairness – Inductive Bias



Minority loan repayment records ignored

ML algorithms may assume minority individuals are less likely to repay loans, ignoring examples of those who have repaid



Biased assumptions

ML algorithms can make biased assumptions about minority groups that lead to unfair outcomes

Inductive bias occurs when assumptions in ML algorithms cause biased and unfair treatment.

Inductive Bias

- Occam's Razor

Prefer simpler models, but may miss exceptions

- Smoothness

Prefer smoother decision boundaries and curves

- Sparseness

Prefer hypotheses with fewer features

- Model Bias

Learned model may not match target model

Fairness Concepts



Equal treatment

All individuals should be treated equally and without bias, regardless of protected attributes



Protected attributes

Attributes like race,, religion, national original, gender, marital status, age etc. should not be used to discriminate against people.



Fairness metrics

Metrics like demographic parity and equality of opportunity help evaluate model fairness across groups.

Ensuring fairness and mitigating bias is crucial for building ethical, trustworthy ML systems.

Fairness Metrics



Demographic parity

The model predictions should be consistent across various demographic subgroups.



Equality of odds

The true positive rate and true negative rate should be equal for different subgroups.



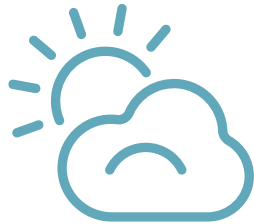
Equality of opportunity

The probability of true positives should be equalized across groupings of a protected category.

Using metrics like these can help measure and improve fairness in machine learning models.

More fairness metric, please refer to the working paper from Amazon [Fairness Measures for Machine Learning in Finance](#)

Data Debiasing



Fairness through Unawareness

Remove features correlated with protected attributes



Equalize data points

Use sampling to have equal data points (or ratio) for each protected group



Data augmentation

Generate more minority data synthetically to balance the dataset

By removing correlated features, equalizing and augmenting data, we can reduce bias caused by imbalanced training data.

Model Debiasing



Change model type

Trying different model types like neural networks, random forests etc to see which has lower bias



Ensemble modeling

Using multiple models together and combining predictions to reduce bias



Multi-task learning

Training model on multiple objectives to learn more generalized representations

By changing the model architecture and training process, we can reduce inherent bias and improve fairness.



Drift

The statistical properties of a dataset change over time. This concept drift can impact the performance and accuracy of machine learning models, which are often trained on historical data. When new data deviates from the training distribution, predictive performance may decline.

Causes of Drift

Seasonal Variations

Data distributions can change due to seasonal factors like holidays, weather, etc.

Changing User Behavior

User preferences and demographics evolve over time, altering data patterns.

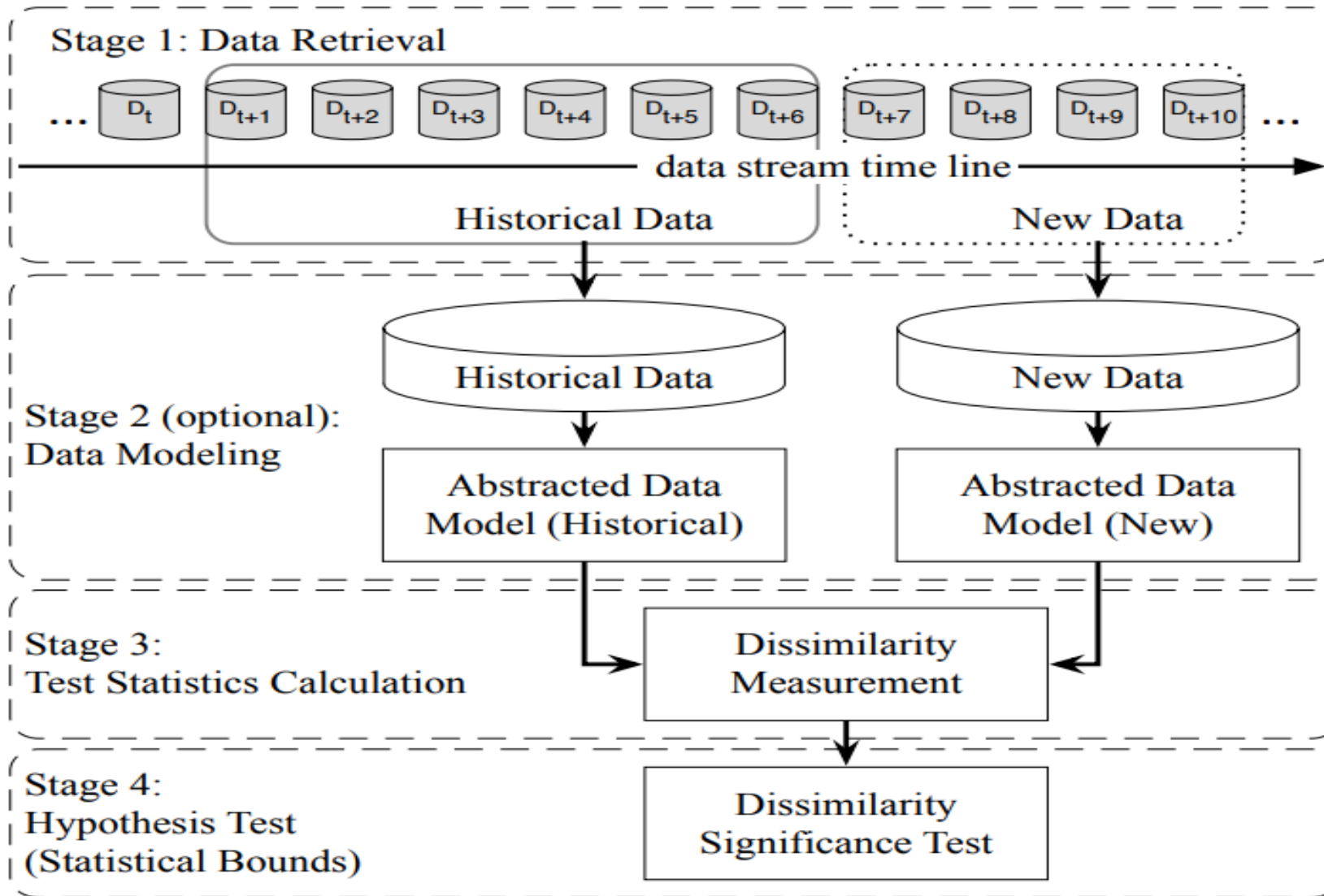
Instrumentation Changes

Modifications to data collection tools and processes introduce variability.

External Events

Major events like regulations, economic shifts and crises impact data.

Drift Detection



Drift Detection Methods



Statistical tests

KS Test and PSI can detect distributional shifts in datasets, other KL divergence family (JS divergence) can also be applied



Model-based approach

Train a classification model on reference dataset and monitor performance on test data

Statistical tests and classification models help detect and quantify data drift

Mitigation Strategies



Retrain models frequently

Update model incrementally

Weight recent data higher

Ensemble models on new/old data