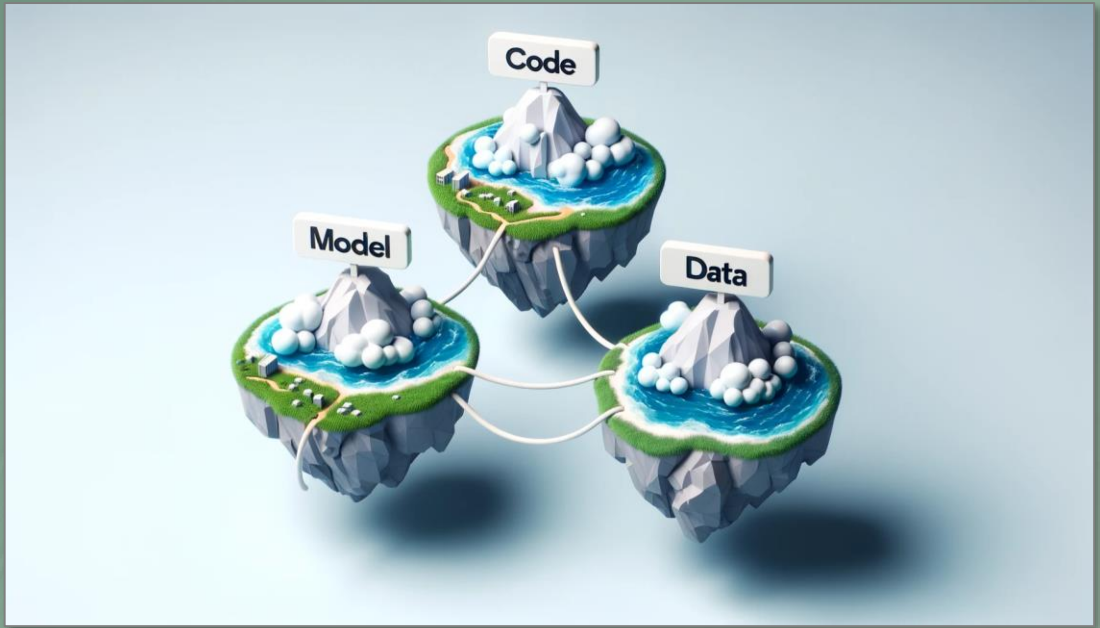
Model Drift

Advanced Software-Engineering
Dr. Harald Stein, Prof. Dr.-Ing. Stefan Edlich
Jan 2024



Agenda

- What is Model Drift?
- Types and Triggers
- Drift Detection techniques
- Infrastructure for monitoring, alerting, retraining
- Examples with classification and segmentation

What is Model Drift?

Definition

Model Drift in the Context of ML Ops

What is Model Drift?

… occurs when the statistical properties of input features or target variable, which predictive model is trying to estimate, change over time.

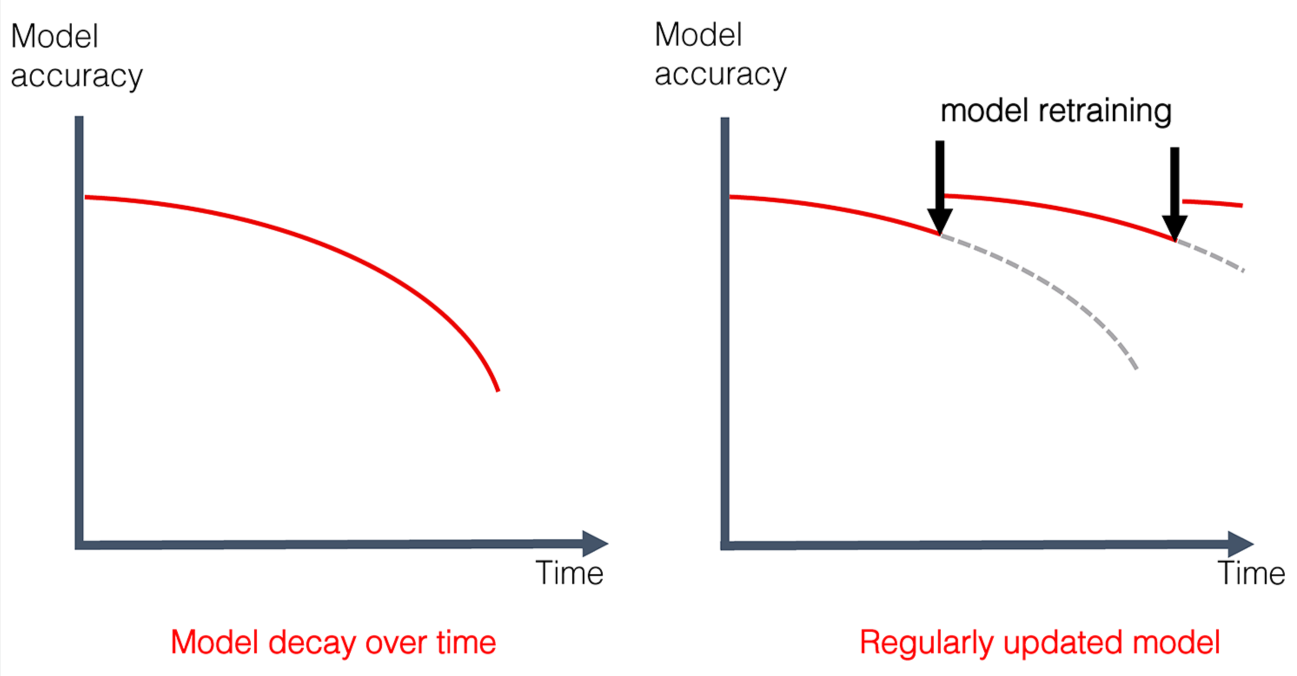
This leads to decrease in model accuracy and effectiveness, as learned patterns no longer represent current situation.

Causes

- Changes in underlying data patterns
- due to various factors like
    - evolving trends
    - user behavior
    - or environmental shifts

Recognizing and addressing model drift

- by regularly updating the model
- is crucial for maintaining reliability, accuracy of machine learning models in real-world applications



Model Drift in the Context of MLOps

Model drift is a critical factor in MLOps, emphasizing the need for continuous monitoring and updating of ML models in production.

Lifecycle Management

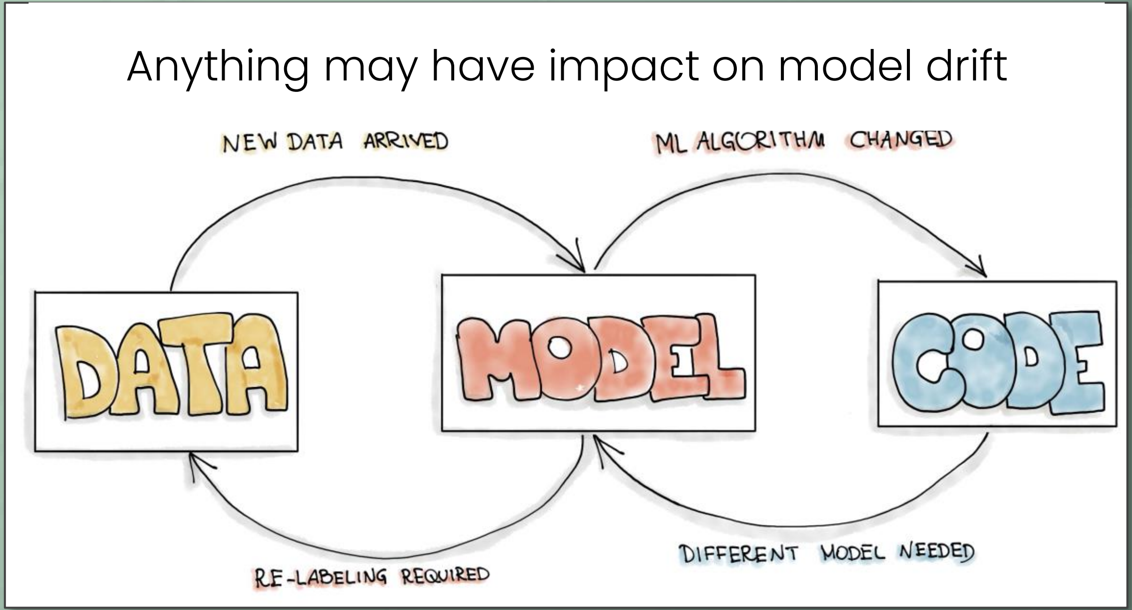
MLOps involves

- … entire lifecycle of a model
- from development to deployment and maintenance
- where identifying, addressing model drift is a key maintenance activity.

Automated Responses

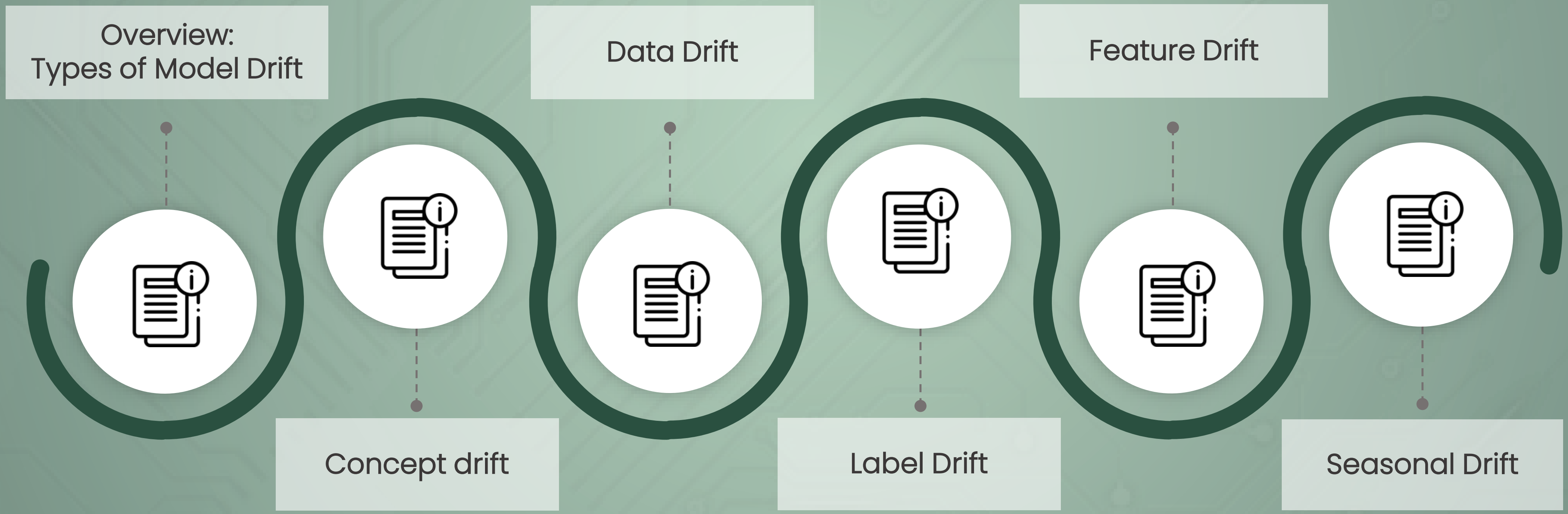
MLOps frameworks often include

- … automated systems to detect and respond to model drift
- ensuring models remain accurate and relevant.



Types and Triggers

… include concept, data, label and feature drift, and are often triggered by changes in external factors, evolving user behaviors, or alterations in the underlying data distribution



Overview: Types of Model Drift

… all of them may lead to decrease in model performance

Concept Drift

changes in external factors influence

- Influence statistical properties
- Make model's predictions less accurate

Data Drift

changes in model's data due to shifts

- In population or sample
- affecting model's accuracy

Seasonal Drift

special type of drift

- seasonal, or cyclical changes in data
- which are predictable, recurring

(Output data) Label Drift

changes in

- Distribution/frequency of outcomes (output variables / labels)
- Special case: Prior probability drift
  Baseline of model accuracy changes

(Input data) Feature Drift

changes in relationship

- Distribution/frequency of outcomes
- Special case: covariate drift
  relationship among input features changes, not relationship to output

Concept drift

… occurs due to external factors that influence statistical properties

Causes

- Evolving consumer behaviors, social trends
- Changes in economic conditions, technology

Impact

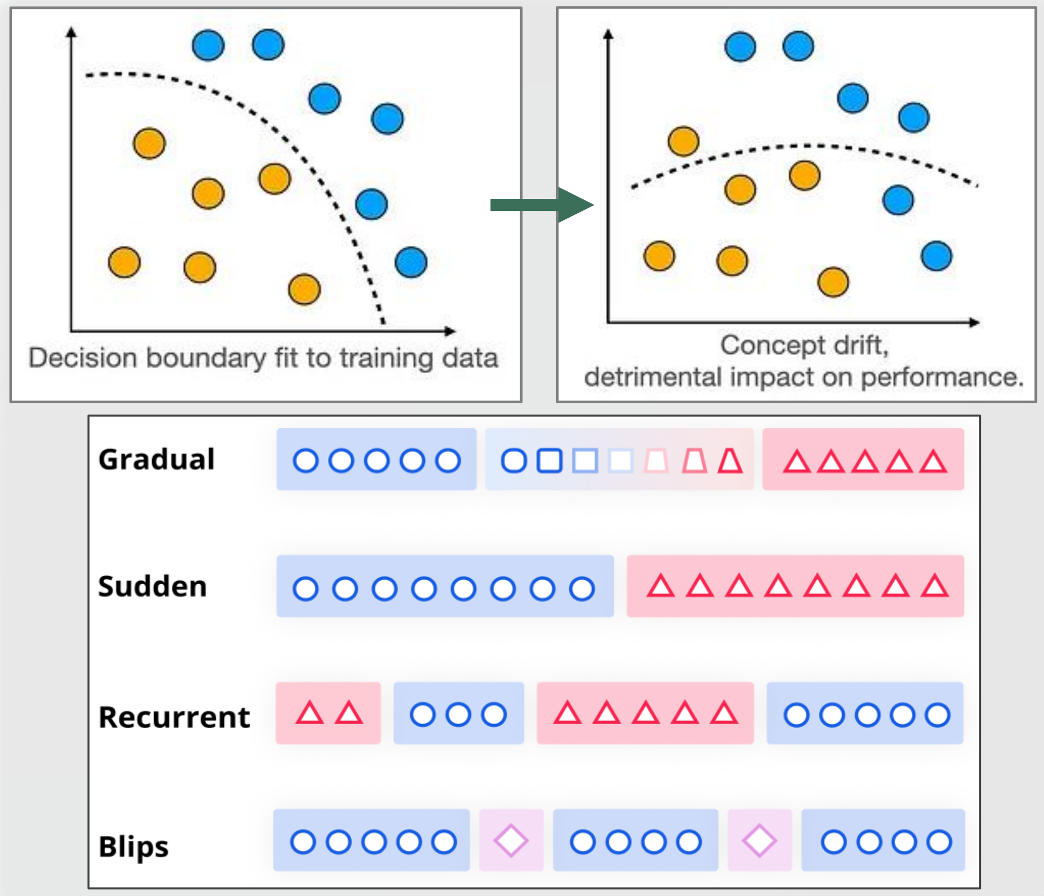
- Decreased model accuracy and predictive reliability.
- Misalignment between learned patterns and current data.

Detection

- Monitoring performance metrics.
- Analyzing prediction errors over time.

Response

- Regular model retraining with updated data.
- Implementing adaptive learning algorithms.



Gradual Concept Drift

Gradual or incremental drift is expected phenomenon that explains how world's changes lead to models' aging and decline in quality

Examples of Gradual Drift

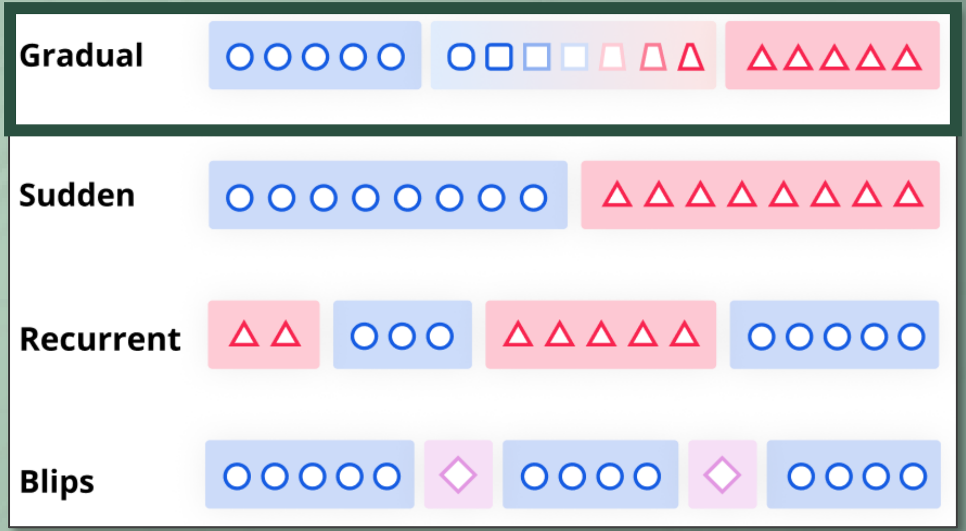
- New competitors alter consumer choices, impacting sales forecasts.
- Macroeconomic shifts redefine credit risk, affecting scoring models.

Observing Shifts

- Shifts detected at individual feature level in tasks like churn prediction.
- Stable feature distribution, but increasing target class proportion in certain ranges signals new patterns.

Model Lifespan and Updates

- Model aging varies
- testing with historical data can guide update frequency.
- Retraining intervals ensure models accuracy.



Sudden concept drift

Sudden or drastic external changes are easily noticeable disruptions

Pandemics

COVID-19 pandemic exemplified such changes, significantly altering mobility and shopping patterns.

Financial Models

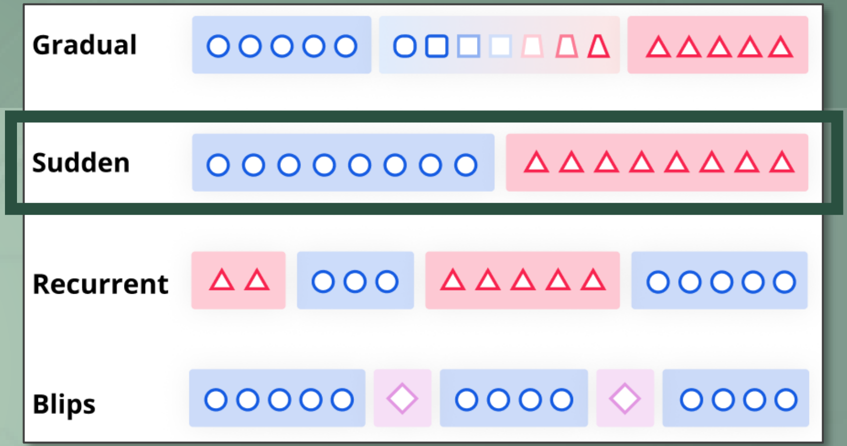
Interest rate adjustments by central banks can immediately render existing financial models outdated.

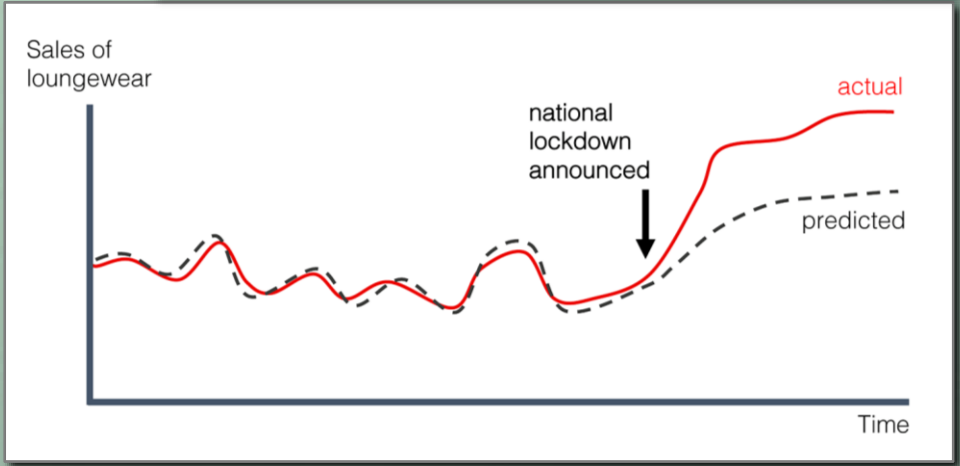
Manufacturing Processes

Updates in production lines necessitate new predictive maintenance strategies due to changes in equipment failure modes.

Digital Platforms

Major app interface overhauls make historical user interaction data obsolete, requiring a reevaluation of user journey models.





Data Drift

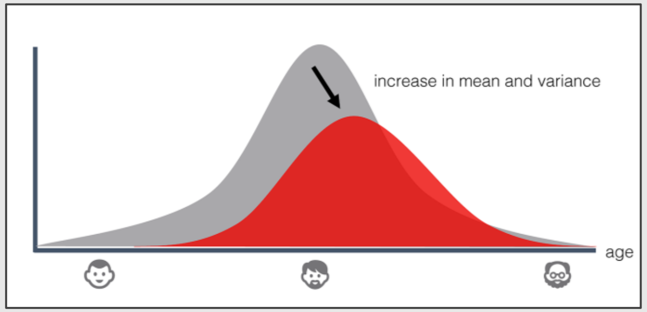
… change in input data's distribution or characteristics over time, affecting model performance.

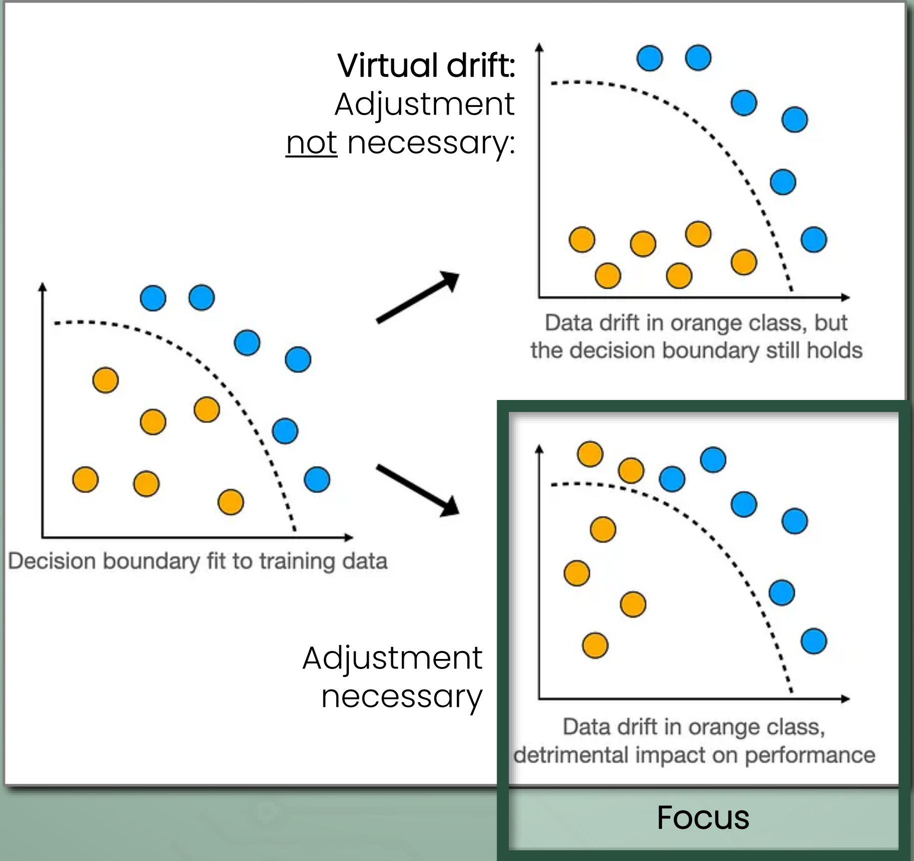
Causes: Shift or changes in

- … user behavior or preferences.
- … data collection methods.
- … population demographics.

Impact

- Decreased model prediction accuracy





Detection by regular monitoring of

- … changes in data statistics, distributions.
- … model inputs compared with training data.

Response

- Retraining model with current data.
- Using different input features

Label Drift as kind of data drift

… changes in the distribution or definition of labels in the dataset over time

Causes: Changing data due to

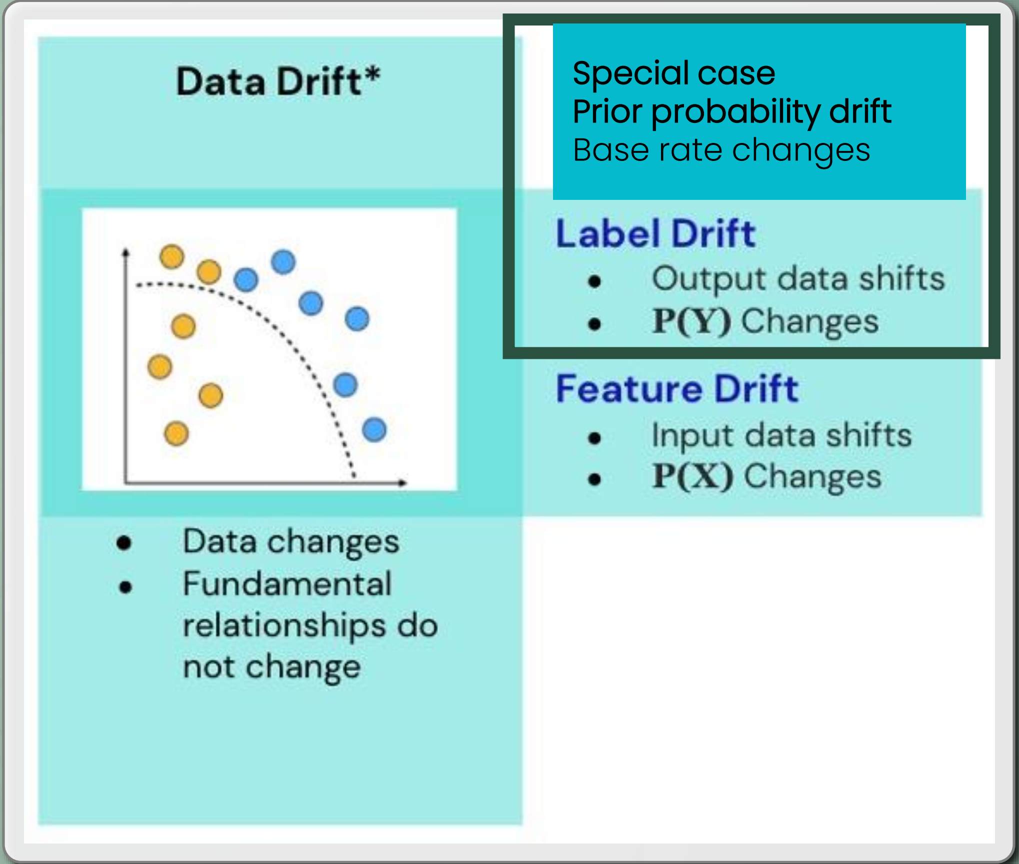
- … changing societal norms or definitions.
- … evolving user behavior or preferences.
- … modifications in labeling criteria or errors.

Label Drift, example: Fashion Trend Forecasting

- Trend prediction model may suffer from label drift
- as certain styles gain or lose popularity.

Prior Probability Drift

- Email Spam Filters:
  Outbreak of spam emails due to new spamming technique, prior probability of any email being spam increases
- Loan Default Prediction:
  During economic downturn, base rate of loan defaults may rise, representing shift in prior probability of default.



Feature Drift as kind of data drift

… changes in relationship between input features and target variable over time.

Causes

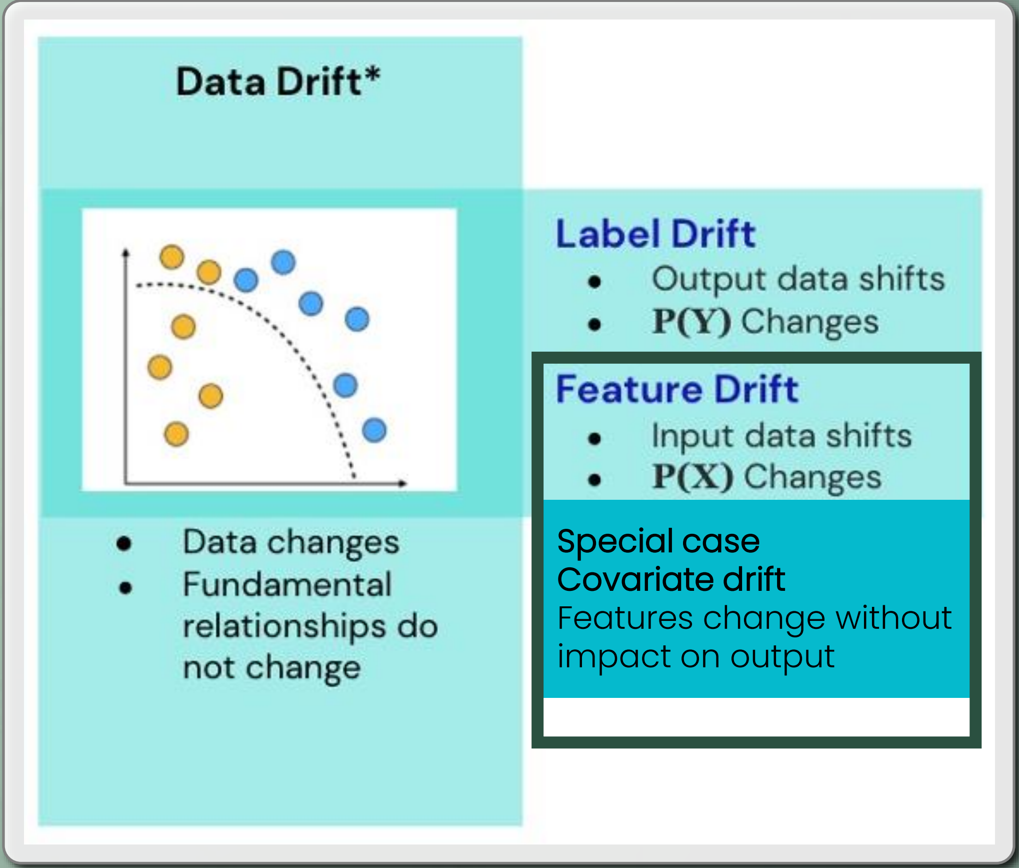
- Evolution in underlying data patterns.
- Technological advancements impacting data collection.
- Shifts in environmental or socio-economic factors.

Feature Drift, example: Consumer Behavior Tracking

- Online retailer's customer segmentation model may face feature drift
- as consumer spending patterns shift due to changing economic conditions.

Covariate Drift, example: Ad Click Prediction

- Advertising model sees shift in type of users
- e.g., younger demographics
- clicking on ads without change in click-through rate



Seasonal Drift

… predictable and recurring changes in data patterns due to seasonal factors

Causes

- Regular events: holidays, weather changes, economic cycles, etc.
- Season-specific consumer behaviors or activities

Impact

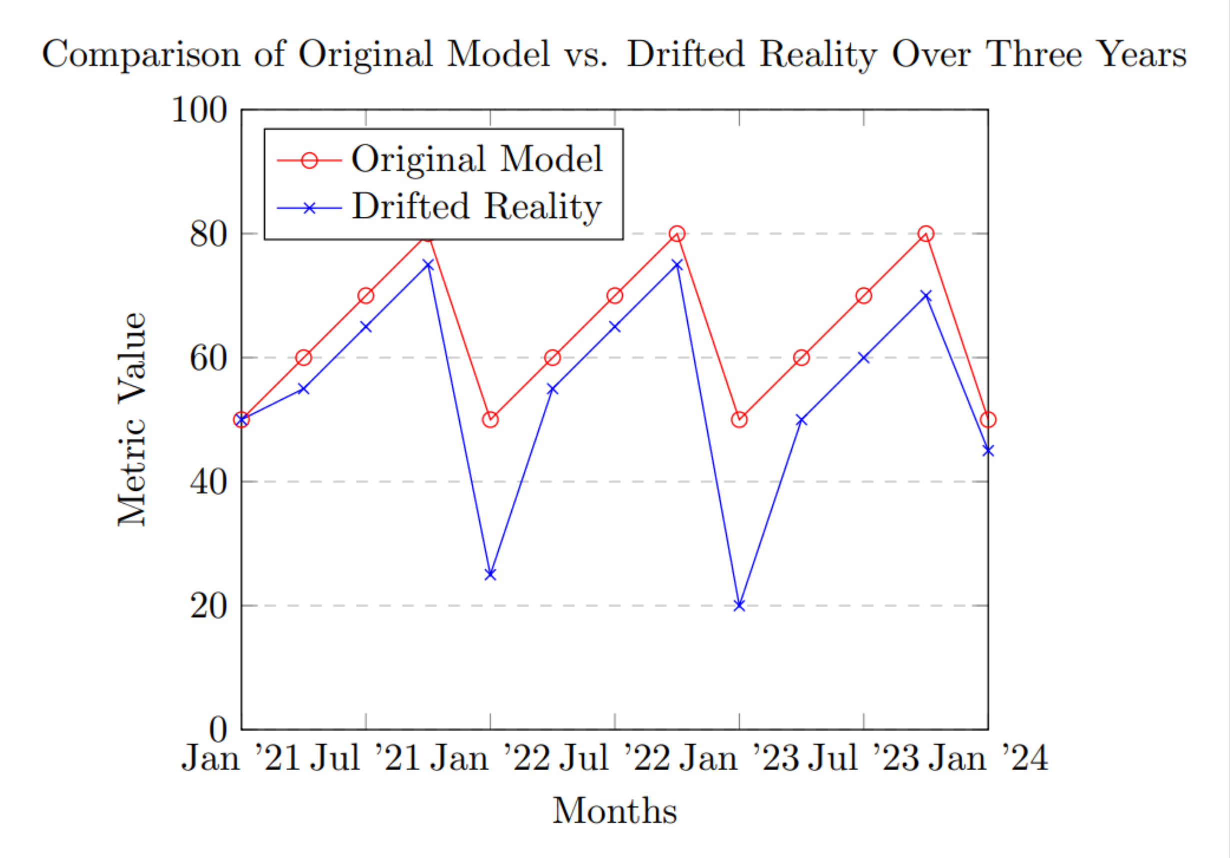
- Temporary shifts in model performance aligned with seasons.
- Potential misinterpretation as long-term trends

Detection

- Analyzing data trends over multiple seasonal cycles.
- Comparing model performance across different times of year.

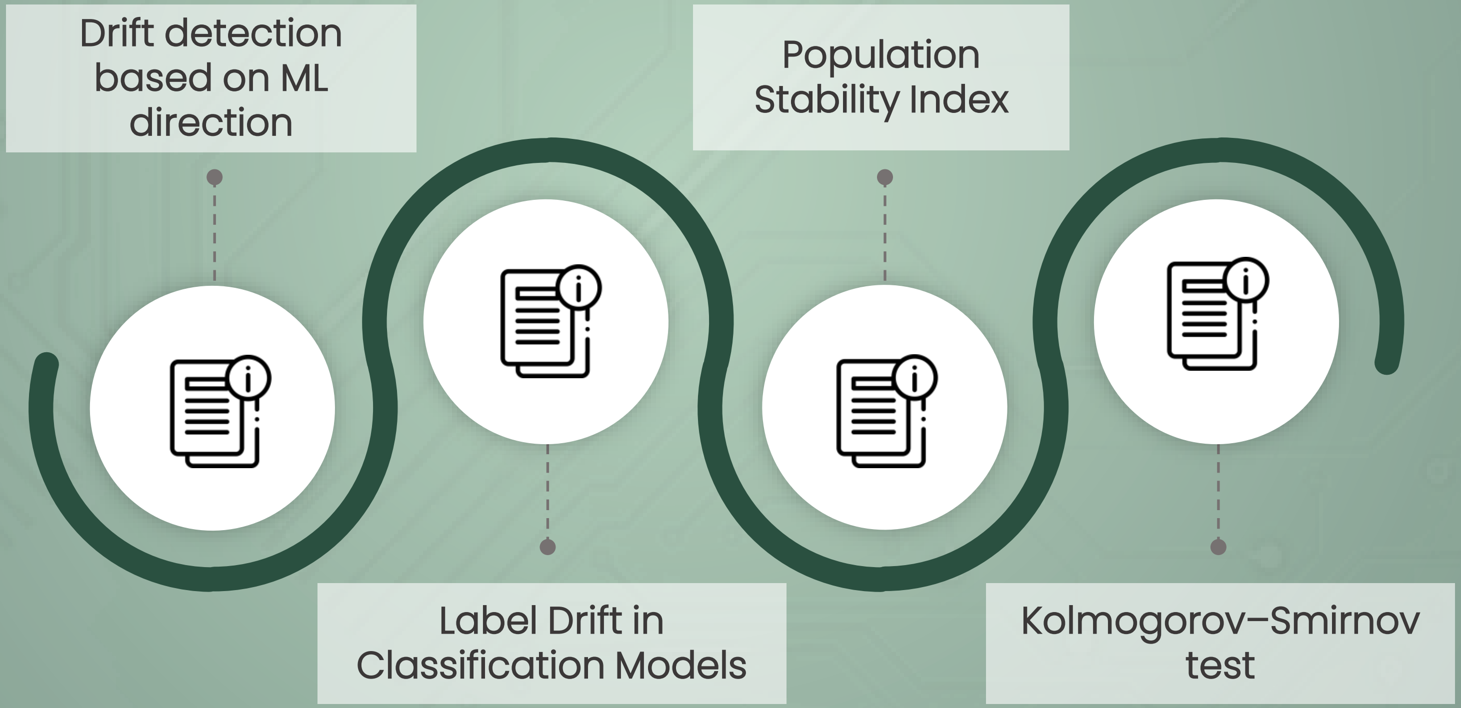
Response

- Incorporating seasonal factors into model design.
- Using time series analysis techniques for better prediction.

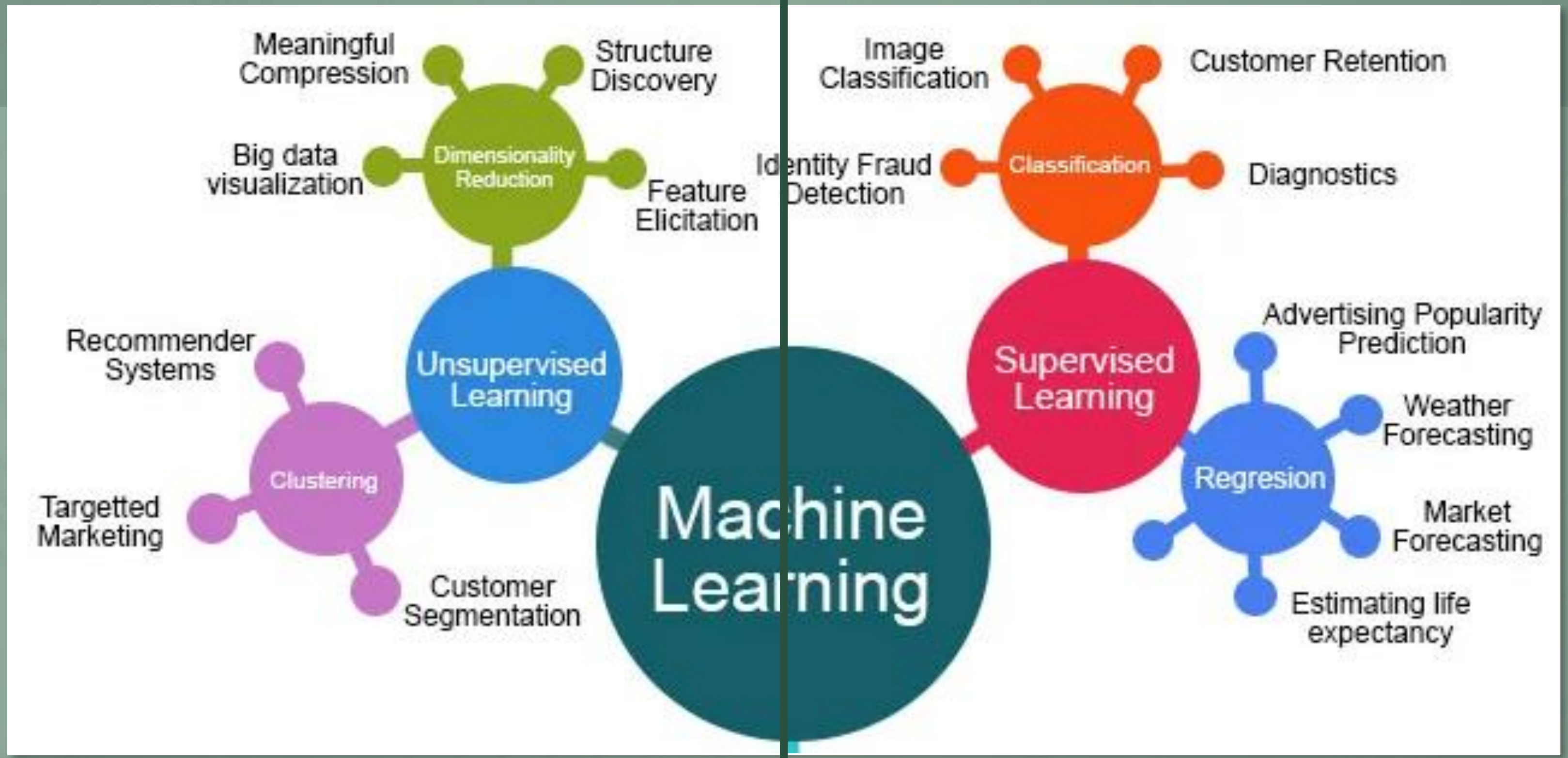


Drift Detection techniques

… for supervised and unsupervised learning



Drift detection based on ML direction



Unsupervised learning:
Data Drift Monitoring

- Population Stability Index
- Kolmogorov–Smirnov test

Supervised learning:
Performance Monitoring

- accuracy decrease & false positives increase

Detecting Label Drift in Classification Models

… occurs when distribution of classes/labels in dataset changes over time

Example spam detection model

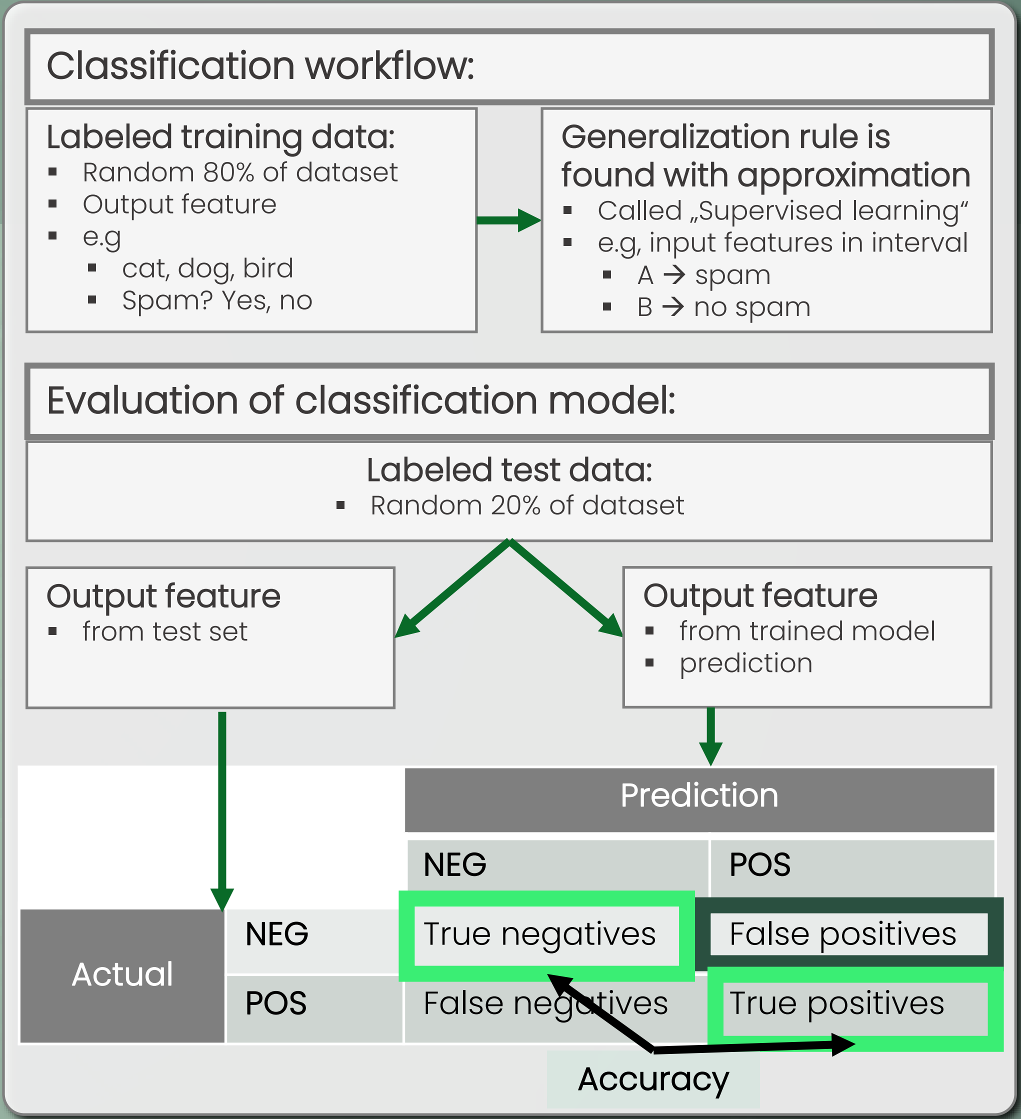
- sudden increase in legitimate emails classified as spam (false positives)
- coupled with decrease in overall accuracy could signal label drift

Tracking Accuracy and False Positives

- Accuracy: Drop in overall accuracy can indicate label drift, especially if model was previously well-calibrated.
- False Positives: Increase in false positive rates can be strong indicator of label drift in specific classes.

Actionable Responses

- Data Analysis:
  Examine distribution of classes in recent data vs. training set.
- Model Update:
  Retraining model with updated data reflecting new class distribution.



Detecting Label Drift in Classification Models

Example spam detection model. sudden increase in legitimate emails being classified as spam (false positives) coupled with decrease in overall accuracy could signal label drift

Example numbers: Ex ante		Prediction	
		NEG	POS
Actual	NEG	True negatives: 30%	False positives: 10%
	POS	False negatives: 20%	True positives: 40%

Example numbers: Ex post		Prediction	
		NEG	POS
Actual	NEG	True negatives: 25%	False positives: 15%
	POS	False negatives: 30%	True positives: 30%

False positives:	10% → 15%
Accuracy:	70% → 55%

Population Stability Index (PSI)

… is statistical measure for determining stability of model's population over time. It quantifies how much input variables' distribution has shifted: training dataset ⇔ new dataset.

- PSI is calculated by dividing population into buckets
- based on variable distribution and
- comparing distribution in each bucket between training and new datasets

$$PSI = \sum \left(\%_{actual} - \%_{expected}\right) \times \ln \frac{\%_{actual}}{\%_{expected}}$$

Interpreting PSI

- Low PSI value (typically < 0.1) indicates little to no shift in population, suggesting stable model performance.
- Higher PSI values suggest significant shifts, indicating potential model drift, need for model recalibration

Input Feature, e.g.: income, size, speed

Score Range	Decile	Scoring% (A)	Training% (B)	A - B	ln(A/B)	PSI
>720	1	12%	11%	1%	0.09	0.001
671–720	2	11%	11%	0%	0.00	0.000
641–670	3	14%	12%	2%	0.15	0.003
611–640	4	12%	13%	-1%	-0.08	0.001
581–610	5	12%	11%	1%	0.09	0.001
551–580	6	10%	11%	-1%	-0.10	0.001
521–550	7	12%	13%	-1%	-0.08	0.001
491–520	8	6%	5%	1%	0.18	0.002
451–490	9	6%	7%	-1%	-0.15	0.002
< 451	10	5%	6%	-1%	-0.18	0.002
						0.013

Population Stability Index

Kolmogorov-Smirnov (KS) test with Bonferroni correction

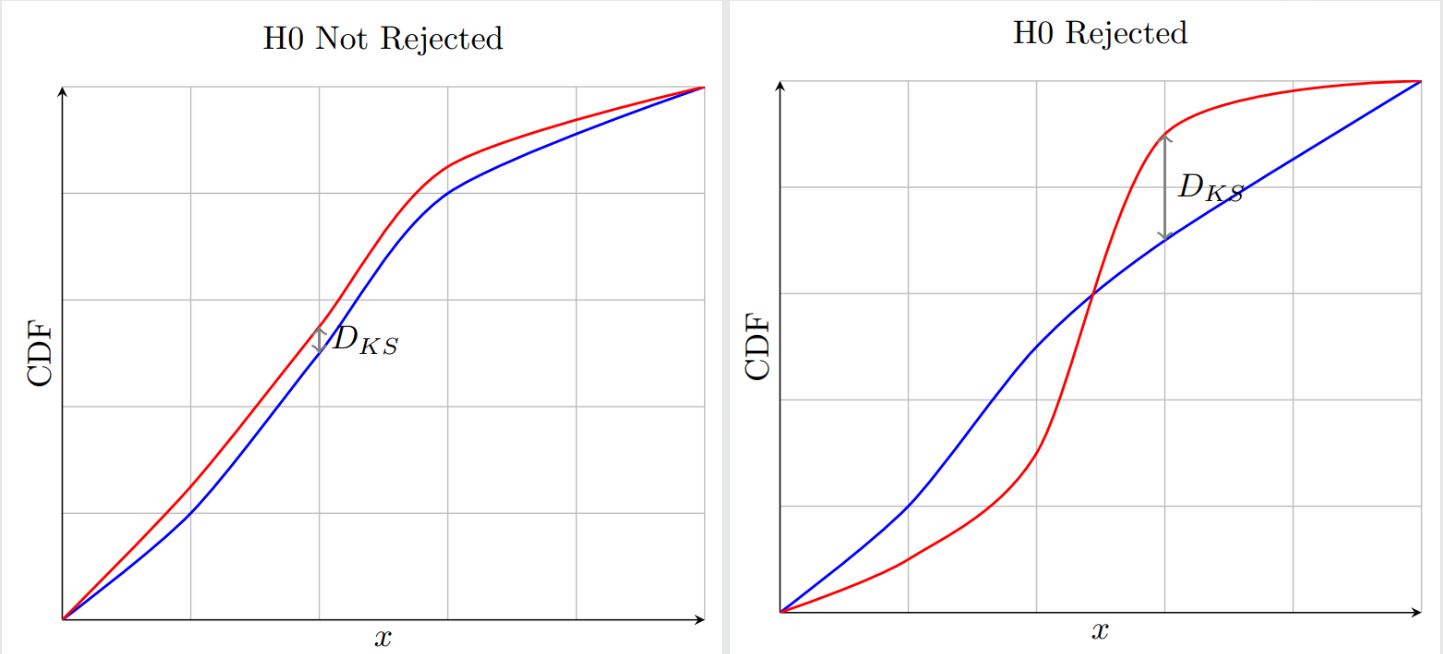
is non-parametric test used to compare distributions of two datasets which measures largest distance between cumulative distribution functions of two datasets.

Null hypothesis (H_0)

- Distributions x, y come from same population
- If KS statistic has p-value < α, reject H_0

Bonferroni correction

- Adjusts α level to reduce false positives
- α_new = α_original / n, where n = total number of feature comparisons



red: ex ante, blue: ex post

red: ex ante, blue: ex post

Infrastructure for monitoring, alerting, retraining

… involves automated systems for performance tracking and processes for periodic model updates

- CRISP-ML(Q): The ML Lifecycle Process
- Retraining-Validation-Monitoring Loop
- Setting Alerts and Thresholds
- Roles and Responsibilities
- Software Tools

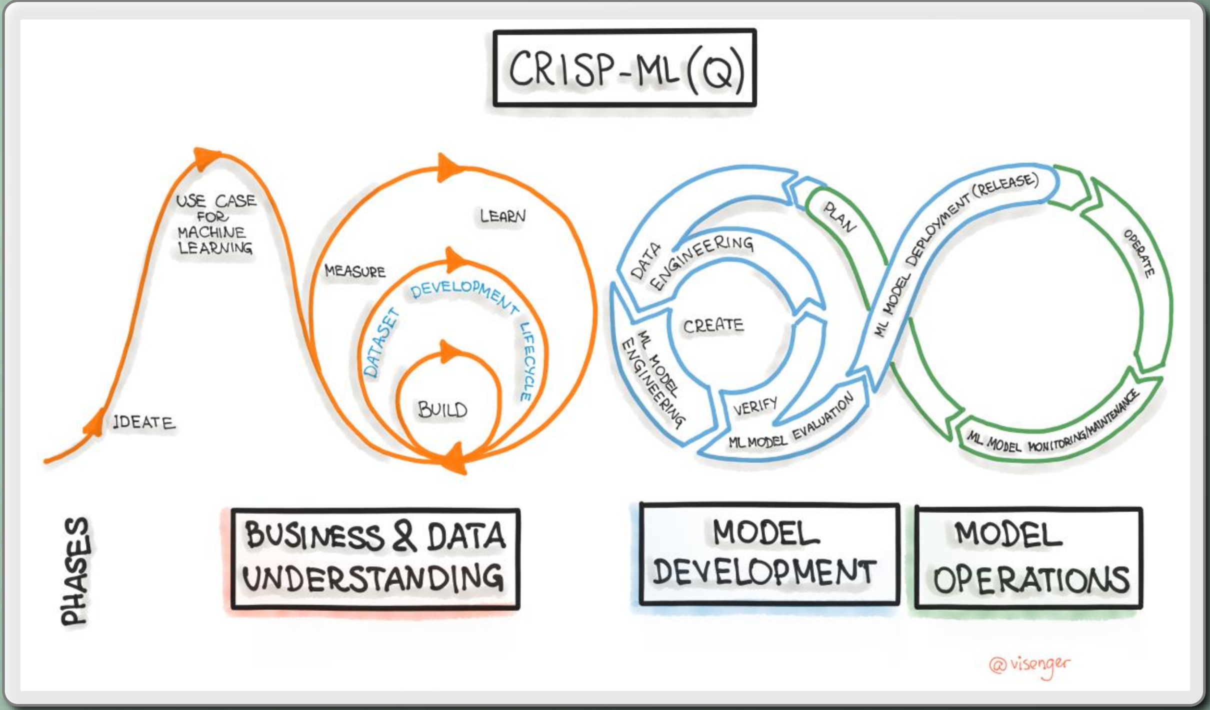
CRISP-ML(Q): The ML Lifecycle Process

… proposed standard to structure, enhance the organization of ML and data science projects.

Phase No.	Phase Name	Description
1	Business and Data Understanding	Setting context, understanding data sources.
2	Data Engineering (Data Preparation)	Processing and transforming raw data.
3	Machine Learning Model Engineering	Algorithm selection, training, and tuning.
4	Quality Assurance for ML Applications	Ensuring model robustness, performance, etc.
5	Deployment	Launching the model into a production or operational environment.
6	Monitoring and Maintenance	Continuous monitoring / drift detection and updating the model as needed.

Retraining, validation

Monitoring

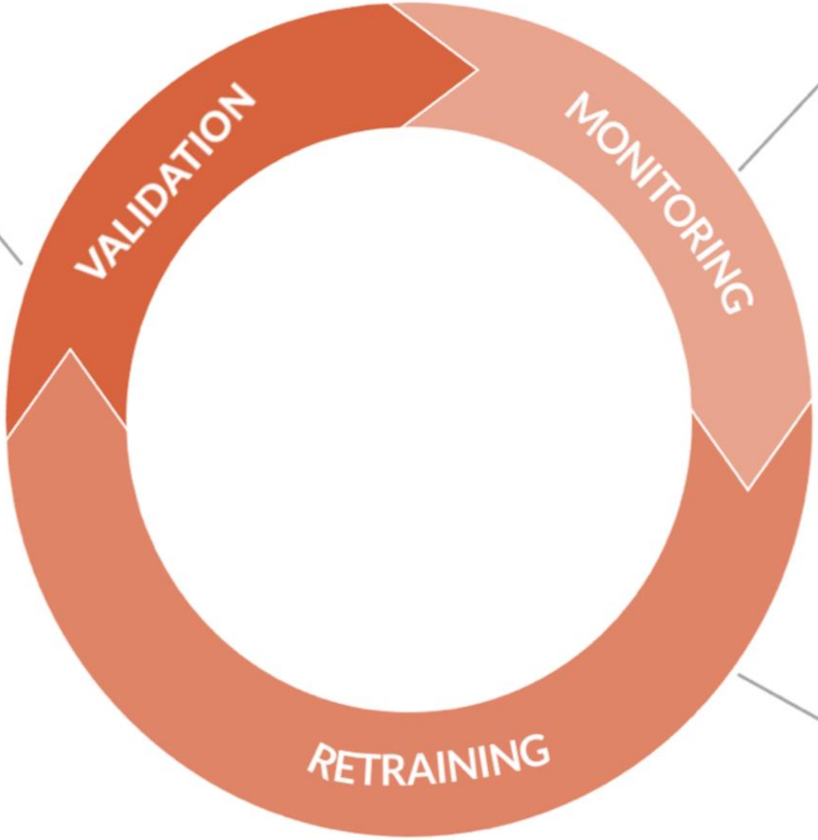


Retraining-Validation-Monitoring Loop

… is loop of monitoring, retraining, validation or drift detection

VALIDATION

- Model Evaluation:
  Assessing model performance against set of criteria
- Model Validation:
  Confirming accuracy and reliability of model.
  … or opposite. Drift detection
- Inference Serving:
  Deploying model to make predictions or inferences.



MONITORING

- Data Schema:
  Ensuring structure of incoming data remains consistent.
- Data Distribution Monitoring:
  Watching for changes in data that could signal drift.
- Prediction Distribution:
  Checking distribution of predictions to detect shifts.
- Anomaly Detection:
  Identifying unusual patterns that may indicate problems.

RETRAINING

- Data Extraction:
  Pulling new data to refresh the training dataset.
- Data Validation:
  Verifying the quality and relevance of the new data.
- Data Preparation:
  Preprocessing and cleaning data for training.
- Model Selection:
  Choosing an appropriate model framework or algorithm.
- Model Training:
  Updating the model with new data to maintain accuracy.

Setting Alerts and Thresholds

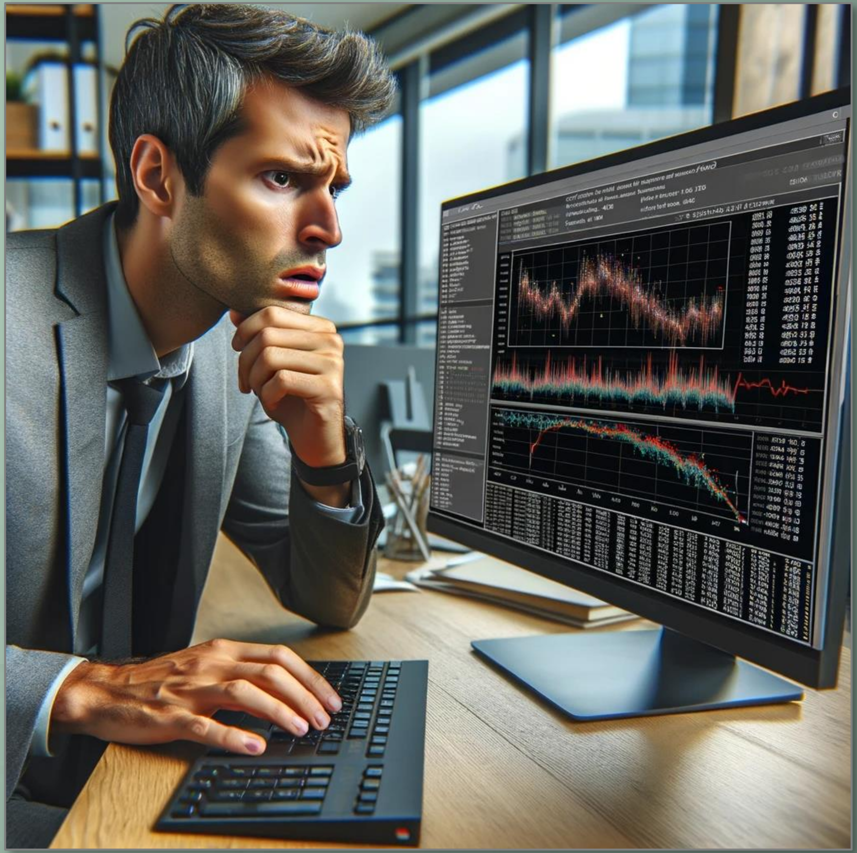
Strategy for alerts should be according to potential business impact of model performance issues. Efficient alerting, threshold settings are balancing act

Alert Mechanisms

- Utilizing email alerts as primary communication channel for stakeholders.
- Selecting relevant measures like model accuracy, KS-test, etc. for alerting.

Threshold Definition

- Align thresholds with business impact to avoid overburdening the system.
- Ensures thresholds are neither too sensitive (causing false alarms) nor too insensitive (missing important alerts).
    - Sensitive thresholds for high-stakes decisions, e.g., medical diagnoses.
    - More relaxed thresholds for less critical misclassifications, e.g., marketing campaigns.



Roles and Responsibilities

Strategy for alerts should be according to potential business impact of model performance issues. Efficient alerting, threshold settings are balancing act

Alert Recipients

- Determine who within organization needs to be notified
- Typically includes data engineers and data scientists involved in model development and data management.

Technical Involvement

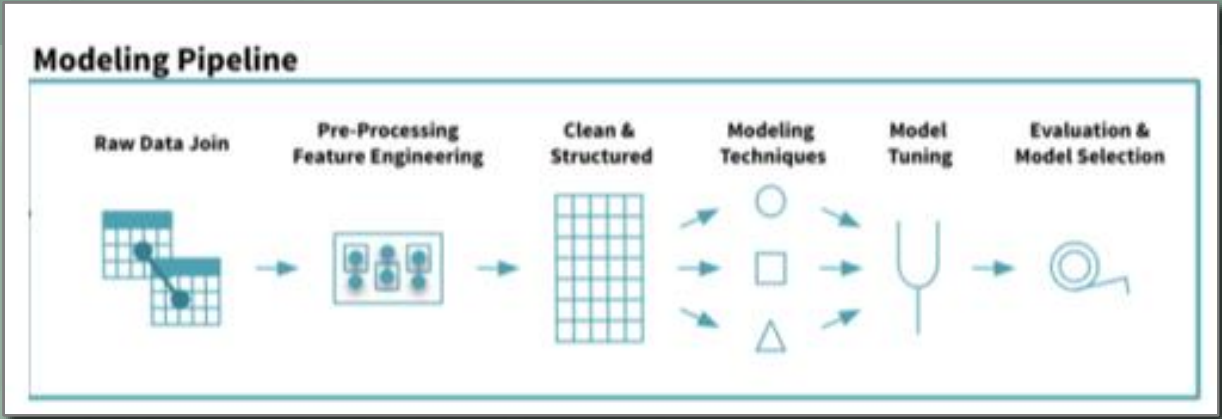
- Ensure technical teams are poised to analyze and address root drift causes
- Automation in ML pipeline may handle some issues, but human oversight remains crucial for unanticipated problems.

Documentation of changes

- Record all modifications made in response to drift to maintain a history of actions and their outcomes.
- Inform affected shakeholders of changes in ML models, data, pipelines, etc.

Software Tools: MLflow

Open-source platform designed to manage end-to-end machine learning lifecycle, encompassing creation, deployment and maintenance stages.



Benefits

- Unified Interface:
  Streamlines workflow for different ML tasks.
- Cross-platform:
  Compatible with various ML frameworks and languages.
- Community-Driven:
  Contributions from a wide range of users and companies.

Use Cases

- Ideal for teams seeking to streamline experimentation, reproducibility, and deployment of ML models.

mlflow

Tracking
Record and query experiments: code, data, config, results

Projects
Packaging format for reproducible runs on any platform

Models
General format for sending models to diverse deploy tools

Software Tools: Python

… is used for handling model drift because it offers extensive libraries and tools for data analysis, machine learning, and statistical testing

Editors

- IDE: Spyder
- Notebook: Jupyter

Basic Python packages

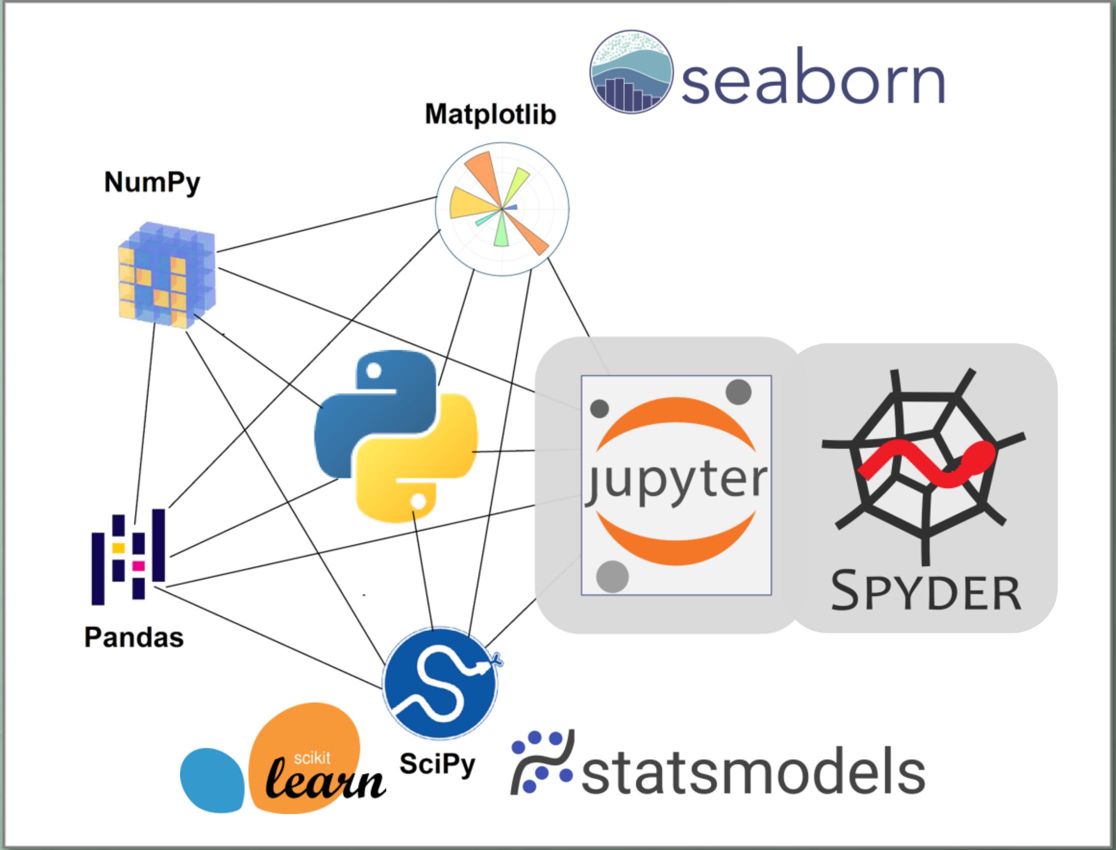
- Linear Algebra, general math: Numpy
- Dataframes: Pandas

Statistics

- Hypothesis tests: Scipy
- Machine learning: SKLearn, Scipy

Visualization

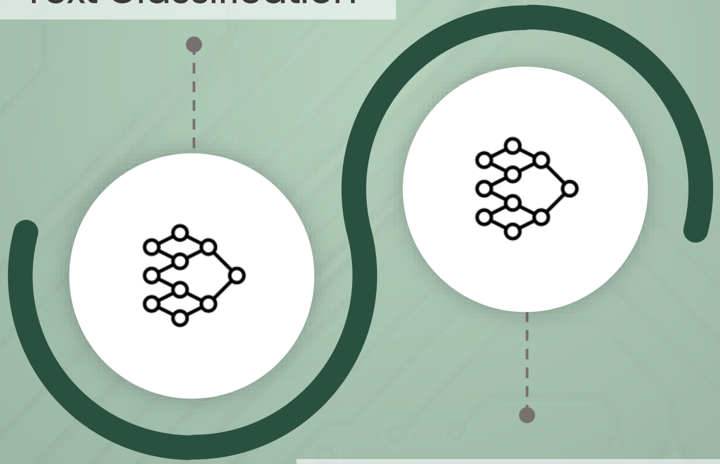
- Matplotlib
- Seaborn



Examples with classification and segmentation

… i.e. with supervised learning (classification) and unsupervised learning (segmentation)

Example: Model Drift in Text Classification



Example: Model Drift in Segmentation

Example: Model Drift in Text Classification

Sentiments of cinema reviews are classified. Later in production model performance will deteriorate

Model Overview

- Model is trained on cinema review comments
- In production (e.g. as tool running on website):
  - trained model gets new, unplanned comments as input
  - makes classification based on training

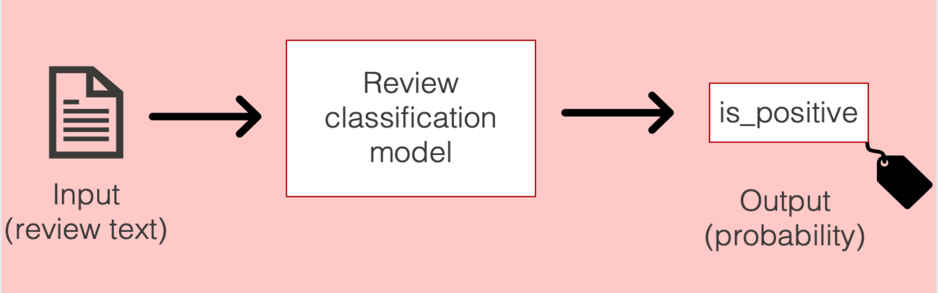
Performance Evaluation

- Focus on accuracy
  (true positives + false negatives, assumption: balanced classed)
- It is expected that accuracy will deteriorate

Model Drift Consideration

- Potential for drift as new, unseen data is introduced.
- Need for continuous monitoring of these performance metrics to detect and address any degradation over time.

Use case:



Labeled training data:

Evaluation of model performance:

Classification Model Performance. Target: 'is_positive'

Current: Model Quality Metrics

Accuracy	Precision	Recall	F1	ROC AUC	LogLoss
0.836	0.858	0.94	0.897	0.728	5.243

Focus

Example: Model Drift in Text Classification

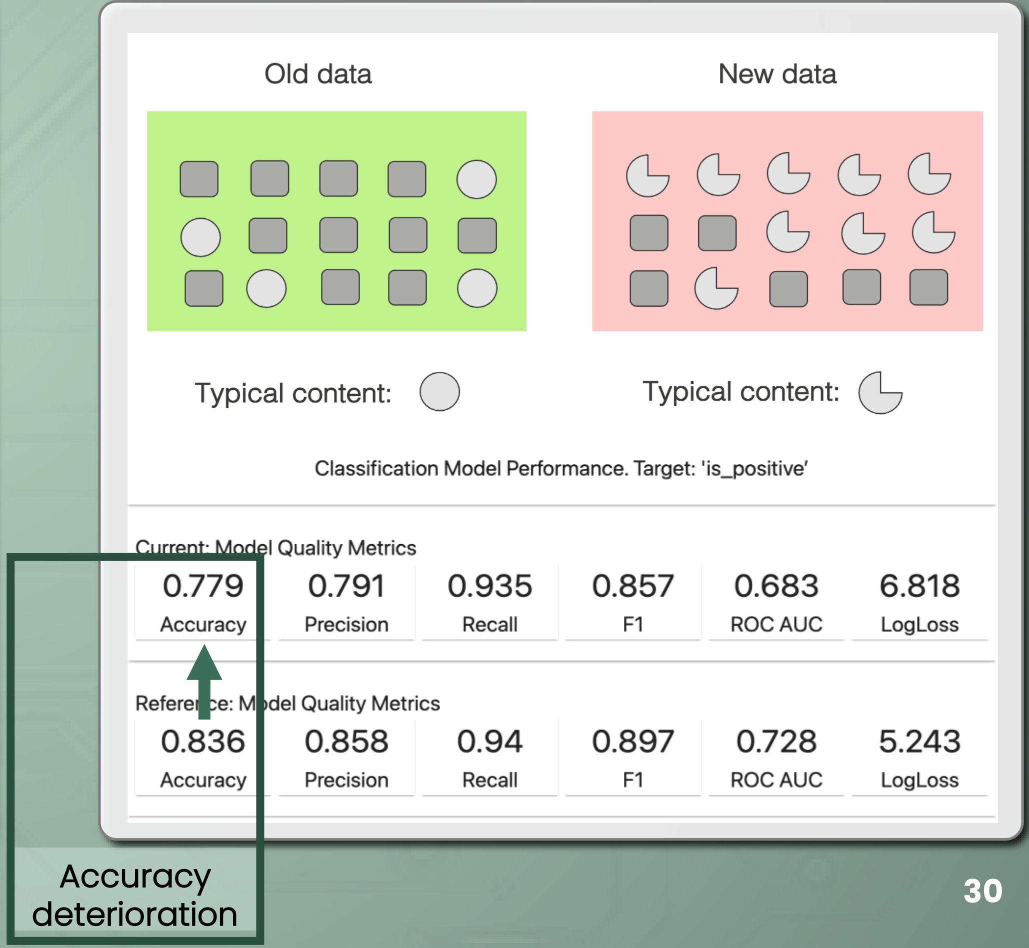
Changes in kind of text content can lead to model drift, affecting model performance

Types of Drift

- Concept drift:
    - Changing word meanings
    - evolution of movie genres.
    - Shifted vocabulary: E.g. Multiple reviews contain over 30% of out-of-vocabulary words
    - Changing text lengths
- Data Drift: New data differing from training data, disrupting learned patterns, maybe due to used data sources.

Consequences

- Checking if drift is permanent, systematic
- If yes, retraining based on thresholds or schedules



Drift Management in E-commerce Recommendation Systems

Ensuring recommendation systems adapt to changing customer preferences, product trends, and seasonal influences.

KS Test for Drift Detection:

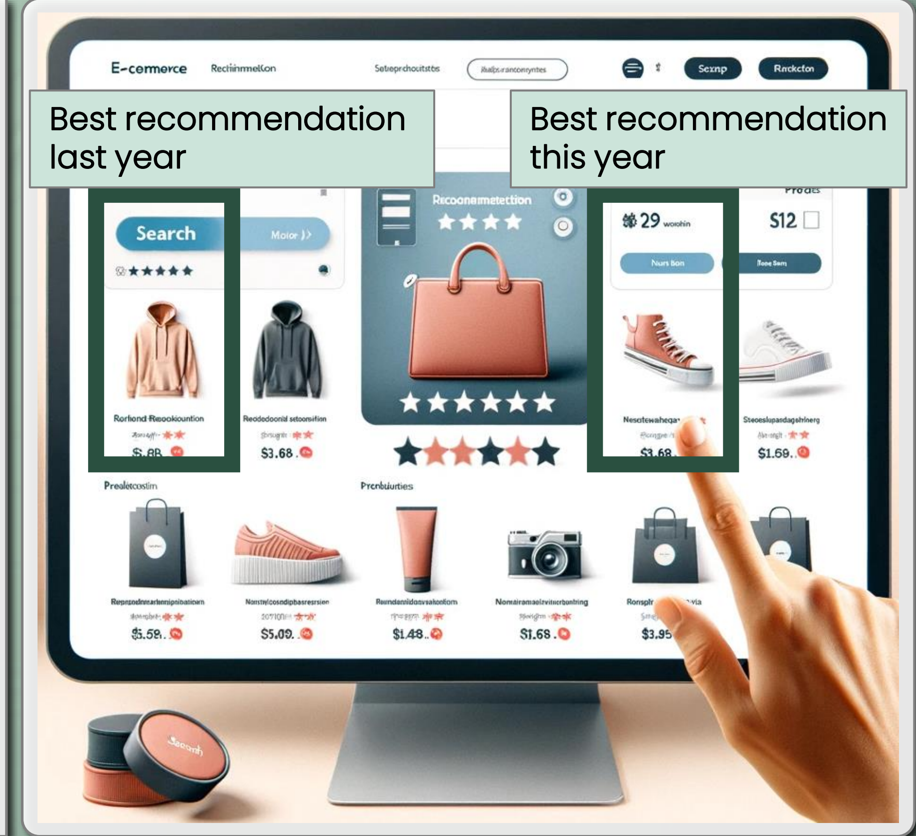
- Monthly KS tests compare current user interaction data with reference dataset to monitor recommendation relevancy.
- Multiple comparisons across product categories necessitate Bonferroni correction to maintain statistical rigor.

Drift Identification and Response

- Sales numbers of highest recommended sales items decrease
- Analysis follows to understand market trends influencing detected drift
- KS test detects significant shift in preferences

Adaptive Model Update Strategy

- Update and retrain models with recent data reflecting current trends.
- Validate improvements through A/B testing against original model.
- Continual adaptation cycle initiated by regular application of KS test.



Links

i.e. sources for self-learning

	Title	Link
Model Drift: Overview	ML Drift: Identifying Issues Before You Have a Problem	https://www.youtube.com/watch?v=uOG685WFO00
	A unifying view on dataset shift in classification	https://www.sciencedirect.com/science/article/abs/pii/S0031320311002901
	Concept Drift Detection in Data Stream Mining : A literature review	https://www.sciencedirect.com/science/article/pii/S1319157821003062
	Understanding Dataset Shift	https://towardsdatascience.com/understanding-dataset-shift-f2a5a262a766
	Model Drift in Machine Learning: How to Detect and Avoid It	https://blog.nimblebox.ai/machine-learning-model-drift
	Machine Learning Monitoring, Part 5: Why You Should Care About Data and Concept Drift	https://www.evidentlyai.com/blog/machine-learning-monitoring-data-and-concept-drift
	End To End ML Life cycle	https://www.linkedin.com/pulse/end-ml-life-cycle-satya-pati/
	The Value Proposition for ML Ops	https://www.credera.com/insights/the-value-proposition-for-mlops
	A Survey on Concept Drift Adaptation	https://www.win.tue.nl/~mpechen/publications/pubs/Gama_ACMCS_AdaptationCD_accepted.pdf

Links

i.e. sources for self-learning

	Title	Link
Types and triggers	Don't let your model's quality drift away	https://towardsdatascience.com/dont-let-your-model-s-quality-drift-away-53d2f7899c09
	What Are Drifts and How to Detect Them?	https://www.youtube.com/watch?v=5KjpZCj853k
	Model Drift in Machine Learning — Data Science	https://priyanka-dalmia.medium.com/model-drift-in-machine-learning-395313b655c2
	An introduction to Model drift in machine learning	https://ubiops.com/an-introduction-to-model-drift-in-machine-learning/
	Drift in Machine Learning: How to Identify Issues Before You Have a Problem	https://www.fiddler.ai/blog/drift-in-machine-learning-how-to-identify-issues-before-you-have-a-problem
	Machine Learning Monitoring, Part 5: Why You Should Care About Data and Concept Drift	https://www.evidentlyai.com/blog/machine-learning-monitoring-data-and-concept-drift
	Machine Learning Concept Drift – What is it and Five Steps to Deal With it	https://www.seldon.io/machine-learning-concept-drift
	A Survey on Concept Drift Adaptation	https://www.researchgate.net/publication/261961254_A_Survey_on_Concept_Drift_Adaptation
	Everything you need to know about drift in machine learning	https://superwise.ai/blog/everything-you-need-to-know-about-drift-in-machine-learning/
	Data Drift	https://medium.com/@evertongomede/data-drift-dee73dcb8b6b

Links

i.e. sources for self-learning

	Title	Link
Drift Detection techniques	From concept drift to model degradation: An overview on performance-aware drift detectors	https://www.sciencedirect.com/science/article/pii/S0950705122002854
	Shift happens: we compared 5 methods to detect drift in ML embeddings	https://www.evidentlyai.com/blog/embedding-drift-detection
	Q&A: ML drift that matters. "How to interpret data and prediction drift together?"	https://www.evidentlyai.com/blog/data-and-prediction-drift
	Take My Drift Away	https://arize.com/blog/take-my-drift-away/
	Inferring Concept Drift Without Labeled Data	https://concept-drift.fastforwardlabs.com/
	Production Machine Learning Monitoring: Outliers, Drift, Explainers & Statistical Performance	https://towardsdatascience.com/production-machine-learning-monitoring-outliers-drift-explainers-statistical-performance-d9b1d02ac158
	Detecting Data Drift with KS Test Using Attention Map	https://link.springer.com/chapter/10.1007/978-3-031-47634-1_6
	Understanding Kolmogorov-Smirnov (KS) Tests for Data Drift on Profiled Data	https://towardsdatascience.com/understanding-kolmogorov-smirnov-ks-tests-for-data-drift-on-profiled-data-5c8317796f78
	Monitoring NLP models in production: a tutorial on detecting drift in text data	https://www.evidentlyai.com/blog/tutorial-detecting-drift-in-text-data
	Survey of distancemeasures for quantifying concept drift and shift in numeric data	https://www.researchgate.net/publication/327539525_Survey_of_distance_measures_for_quantifying_concept_drift_and_shift_in_numeric_data
	Population Stability Index (PSI)	https://medium.com/model-monitoring-psi/population-stability-index-psi-ab133b0a5d42

Links

i.e. sources for self-learning

	Title	Link
Infrastructure: Monitoring, Alerts, Mitigation	Staying on Track: The Impact of Data Drift on Machine Learning and How to Overcome It	https://www.linkedin.com/pulse/staying-track-impact-data-drift-machine-learning-how-iain-brown-ph-d-/
	Chapter 7. Monitoring and Feedback Loop	https://www.oreilly.com/library/view/introducing-mlops/9781492083283/ch07.html#:~:text=Monitoring%20and%20feedback%20loop%20highlighted,correctly%20in%20the%20production%20environment.
	The Value Proposition for ML Ops	https://www.credera.com/insights/the-value-proposition-for-mlops
	MLOps Principles	https://ml-ops.org/content/mlops-principles

	Title	Link
Tools and Technologies	Getting a Grip on Data and Model Drift with Azure Machine Learning	https://towardsdatascience.com/getting-a-grip-on-data-and-model-drift-with-azure-machine-learning-ebd240176b8b
	Deploy Your Own MLflow Workspace On-Premise with Docker	https://towardsdatascience.com/deploy-your-own-mlflow-workspace-on-premise-with-docker-b54294676f0b
	Mary Grace Moesta - MLOps Deployment Patterns with Delta Lake and MLflow | PyData Seattle 2023	https://www.youtube.com/watch?v=BkUuH51zSYo
	I built an awesome ML model with Spark, Delta and MLFlow. How do I get the right people to use it?	https://fukumaruuu.medium.com/i-built-an-awesome-ml-model-with-spark-delta-and-mlflow-how-do-i-get-the-right-people-to-use-it-db0c34e46605
	MLflow: A Primer Why/how to transform on-premise ML frameworks into a unified one	https://towardsdatascience.com/mlflow-a-primer-6dfe6be48353
	Drifting Away: Testing ML Models in Production	https://www.youtube.com/watch?v=tGckE83S-4s

ChatGPT/Dall-E3 Prompts

Photo of three floating islands, each labeled 'Code', 'Model', and 'Data'. The 'Model' island is slightly shifted from its original position, indicating drift. Connecting bridges show the interdependence. The title 'Model Drift: When Code, Model, or Data Shifts' is displayed above.

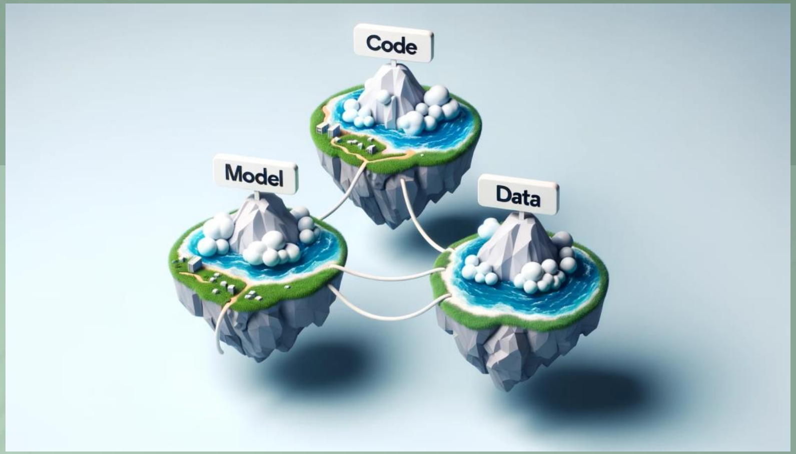


Photo of a frustrated data scientist sitting at a desk, staring at a computer screen displaying a prediction model's performance metrics. The graph on the screen shows a noticeable downward trend in model accuracy and other key performance indicators. Papers with printed charts and graphs are scattered across the desk, and a crumpled paper can be seen in the corner, symbolizing disappointment and concern over the deteriorating model performance. The room is dimly lit to emphasize the challenging situation.

About me

Dr. Harald Stein

- Data Scientist ~ 7 years experience
- Algotrader ~ 4 years experience
- Ph.D. in Economics, Game Theory

- LinkedIn: https://www.linkedin.com/in/harald-stein-phd-1648b51a
- ResearchGate: https://www.researchgate.net/profile/Harald-Stein