Data Quality Management

Discussion Paper

EY

Building a better working world

Table of Contents

Overview

What is Data Quality?

The general definition of data quality is "fitness for use"

Data Quality Definitions in Literatures

From a business perspective, data quality is the **capability of data to satisfy the stated business**, system, and technical requirements of an enterprise.[1]

From a consumer perspective, data quality is an insight into or an evaluation of data's **fitness of use** by data consumers.[2]

From a standards-based perspective, data quality is the **usefulness, accuracy, and correctness of data** for its application.[3]

The general definition of data quality is "fitness for use", or more specifically, to what extent some data successfully serve the purpose of the user.

Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features."

Illustration of the desired characteristics for data that make them fit for purpose

Free of defects	Desired features
Correct	Contextual
Complete	Pertinent
Valid	Comprehensive
Reliable	Easy to read
Consistent	Unambiguous
Unique	Easy to understand
Current	Right level of details

1. Mahanti, R. (2019). "Chapter 1: Data, Data Quality, and Cost of Poor Data Quality". *Data Quality: Dimensions, Measurement, Strategy, Management, and Governance*. Quality Press. pp. 5-6.
2. *Fürber, C. (2015). "3. Data Quality". Data Quality Management with Semantic Technologies. Springer. pp. 20–55.*
3. NIST Big Data Public Working Group, Definitions and Taxonomies Subgroup (October 2019). "NIST Big Data Interoperability Framework: Volume 4, Security and Privacy" (PDF). *NIST Special Publication 1500-4r2* (3rd ed.).

Why is Data Quality Important?

Poor data Quality can result in a variety of negative consequences

Data Quality is crucial because it can have a significant impact on the accuracy, effectiveness, and reliability of business decisions and operations. Poor data quality can result in a variety of negative consequences:

- **Poor business decision and Missing Opportunities:** Low-quality data can produce inaccurate or incomplete insights and analysis, which can lead to incorrect business decisions and to missing out on potential opportunities for growth or improvement
- **Wasted time and resources:** Cleaning and fixing low-quality data can be time-consuming and costly, resulting in wasted resources.
- **Damage to reputation:** Poor data quality can cause customer dissatisfaction and damage a company's reputation.
- **Non-compliance risks:** Organizations can face legal and regulatory sanctions for using inaccurate or incomplete data.

Poor data quality can lead to failures.

In 1999, NASA's Mars Climate Orbiter mission failed due to a miscalculation of the spacecraft's trajectory as the mission team in US and Europe used different units to express force.

In 2012, JPMorgan Chase suffered a significant loss due to risky trading practices by a group of traders known as the "London Whale." The root cause of the issue was traced back to poor data quality management practices that led to inaccurate risk assessments.

During the COVID-19 pandemic, several countries faced challenges in accurately reporting infection rates, death counts, and vaccination data such as inconsistencies in data collection methods, delays in reporting, and data entry errors. Poor data quality hampered the ability to track the virus's spread, allocate resources, and make informed policy decisions.

In 2016, the polling data used to predict the outcome of the U.S. presidential election was flawed due to low response rates and sampling errors, resulting in inaccurate predictions and analysis.

Facilitating high-quality data is essential for effective decision-making, efficient operations, and maintaining a competitive edge in today's data-driven business landscape

How to Measure Data Quality?

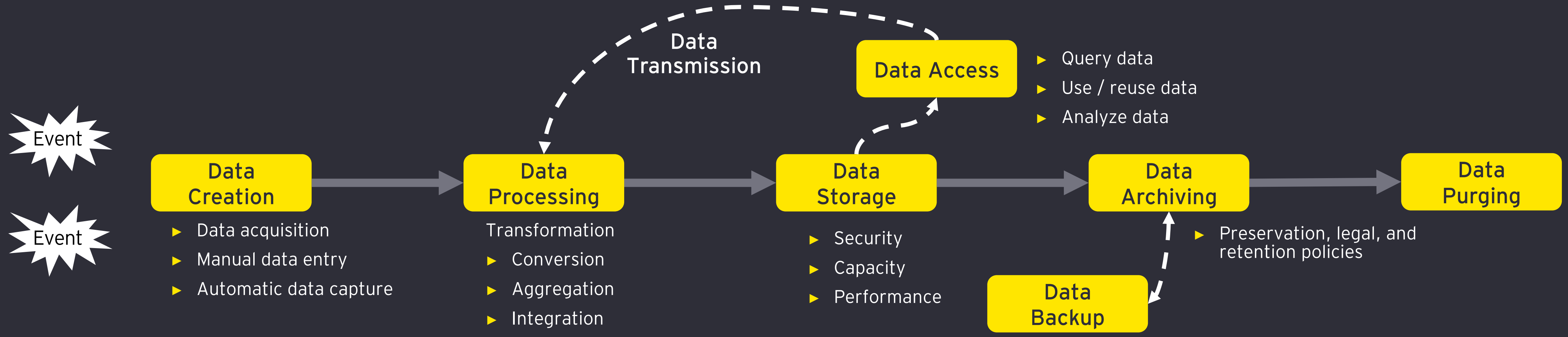
Data Quality Dimensions

- The famous adage "**What gets measured gets managed**" applies to data quality management.
- Despite the fact that fitness for use or purpose does capture the principle of quality, it is abstract, and hence it is a challenge to measure data quality using this holistic construct or definition.
- To be able to assess such a construct, we need to operationalize it into measurable variables -- Data Quality Dimensions. The diagram below captures some of data quality dimensions that are commonly used to describe the characteristic of data.

Possible Causes of Bad Data Quality

Bad data quality can originate from any phase of the data lifecycle

- Data issues can sneak into every phase of the data lifecycle, starting from initial data creation and collection to data processing, transfer, storage, archiving and purging.
- Thus, data quality is impacted by activities in all of the phases in the data lifecycle as illustrated below



COMMON CAUSES OF BAD DATA QUALITY

- Errors in manual data entry
- Inadequate validation in the data capture process
- Aging of data/data decay
- Errors in data migration or conversion
- Mistakes in data integration processes
- Bugs in data cleansing programs
- Errors in data purging processes
- Organization changes, such as M. & A.
- System upgrades
- Loss of expertise
- Inefficient process management and design
- Lack of common data standards, data dictionary, and metadata
- Data ownership and governance issues
- Data corruption by hackers

Data Quality Management Framework

Data quality issues should be addressed in three pillars: people, processes, and technology

- Data quality issues can be mitigated in three pillars of people, processes, and technology as outlined below.
- By implementing these three pillars, organizations can effectively address data quality issues and ensure their data is accurate, reliable, and useful for decision making.

Data Quality Management Framework

People
- Clearly-defined roles and responsibilities
- Talent management and training

Processes
- Data quality standards and guidelines
- Well-established data governance structure
- Policies and Procedures
- Regular data quality checks and audits

Technology
- Data profiling and data cleansing tools
- Data quality monitoring and reporting tools
- Issue remediation with advanced techniques (ML)

Proposed Process Flow for Data Quality Issue Identification and Remediation

The Six Sigma methodology (DMAIC) can be applied to data quality management

1 DEFINE
Define the problem and set goals

- Help define the intended use of the data
- Help identify the required charac-teristics of the data for the intended use
- Help identify the relevant data quality dimensions based on the required characteristics

2 MEASURE
Identify metrics and measure data quality

- Help define the metrics for the selected data quality dimensions to measure the quality of the data
- Help define the requirements /criteria based on the desired characteristics and purpose
- Measure the data quality with defined metrics

3 ANALYZE
Analyze results and identify root cause

- Help analyze the results to identify any issues associated with the data quality
- Help identify the root causes of data quality issues, such as data entry errors, process issues, or system problems

4 IMPROVE
Develop and implement solutions

- Help develop solutions to address the root causes of data quality issues
- Help implement the developed solutions and monitor the impact on data quality

5 CONTROL
Monitor and sustain improvements

- Help create visual dashboards for continuous and comprehensive monitoring
- Document the processes that led to the improvements
- Help train staff on the new processes and procedures to ensure that they are followed consistently

EY

Deep Dive: Data Quality Issue Identification and Remediation

Completeness and Coherence

Deep Dive: Completeness

A general procedure for data completeness checks and remediation

Objective

- To identify missing or incomplete data in a dataset
- To quantify the extent of missing or incomplete data in each variable or record
- To assess the impact of missing or incomplete data on analysis result
- To inform decisions about how to handle missing or incomplete data, such as imputing missing values or excluding incomplete records
- To monitor the completeness of data over time and identify changes that may require action

Procedures for Data Completeness Checks and Remediation

1. Define the completeness criteria

Identify what constitutes complete data for each variable, for example, a range of values, a specific format, or the presence of certain fields.

2. Data completeness measurements

Determine the completeness metrics for each variable, such as the percentage of missing values or the number of incomplete records.

3. Analyze measurement results

Examine the completeness metrics to identify any patterns or trends that may affect your analysis, such as variables with high levels of missing data or records with incomplete data across multiple variables.

4. Remediate completeness issues

Decide what action to take based on the completeness metrics. This could involve imputing missing data, excluding records with incomplete data, or adjusting your analysis to account for the missing data.

5. Monitor completeness

Regularly monitor completeness metrics to ensure that data remains complete over time and to identify any changes that may require action.

Deep Dive: Completeness (contd.)

Considerations on addressing missing values

Type of Missing Values

Missing values in tabular data can be classified into the following three categories

1. Missing Completely at Random (MCAR): Data is MCAR if the probability of being missing is the same in all cases; the cause of the missing data is unrelated to the data and thus complexities arising from the missing data, other than the loss of information, can be ignored
   - e.g. Weighting scale that runs out of batteries
   - e.g. Each member of a population has the same chance of being included in a random sample
2. Missing at Random (MAR): The probability of being missing is the same only within groups defined by the observed data
   - e.g. The probability of being included in a sample taken of a population depends on some known property
   - MAR is more general and realistic than MCAR; modern missing data methods generally start from this assumption
3. Missing Not at Random (MNAR): The probability of being missing varies for unknown reasons
   - e.g. In public opinion research, those with weaker opinions respond less
   - Strategies to handle MNAR data include finding more data about the causes for the missingness or performing what-if analyses to determine the sensitivity of results under various scenarios

Testing for Types of Missing Values

- Testing for the distinction between Missing at Random and Missing not at Random without knowledge of the end to end capture process is *extremely difficult*
- One possibility, specific to modelling use cases, is to use the conditional distribution of the target variable missing versus non missing. More rigorously, we conjecture if:

  $$(Y|X\_i=Null) \sim (Y|\ X\_i\ \neq Null)$$

  then the feature X_i is missing at random and if:

  $$(Y|X\_i=Null) \neq (Y|\ X\_i\ \neq Null)$$

  then the feature is missing not at random.

- This allows us to statistically test for the equality of the conditional distribution within a modelling context
- Multiple tests are available (Chi-Square, Kolmogorov-Smirnov, Epps-Singleton etc.)
- We can also use distance measures like the Kullback-Liebler divergence to establish heuristic thresholds

Missing Value Imputation Techniques

Standard imputation techniques:

- When missing is at random, using Mean or Mode for imputation
- When missing is not at random:
  - Frequency Encoding: Data categories are encoded with values between 0 and 1 based on their relative frequency.
  - Target Encoding: Data categories are encoded with the mean of the data's target variable in the range of 0 to 1. Mathematically, data category x is encoded by the value
    $Encode(x)=(Sum\ of\ Target)/(Count\ of\ Target)$
  - Weight of Evidence (WOE): Tells the predictive power of an independent random variable in relation to the dependent variable. Let p be the event occurrence percentage and q be the non-event occurrence percentage. Then the WOE is given by the following formula $WOE=\ln\ (p/q)$

Model-based imputation techniques:

- Classification and Regression
  - For each categorical column fit a classifier using the rest of the dataset
  - For each numerical column fit a regressor using the rest of the dataset
    - Highly flexible with respect to algorithm choice. The user could specify a novel classification or regression per column
    - May cause problems due to correlation structure of missing values (what happens if 80% of a single row is missing?) which will dramatically affect model performance and quality
- Multiple Imputation by Chained Equation (MICE)
  - Solves the problem above as it crudely imputes all values with mean or mode except for the column it is trying to make predictions on
  - Allows for the construction of confidence intervals depending on the number of iterations used
  - Dramatically more computationally complex

EY

Deep Dive: Coherence

A general procedure for data anomaly checks and remediation

Objective

- To identify any anomaly in the dataset
- To quantify the impact of data anomalies in the decision process
- To inform the decision makers about how to detect and remediate data anomalies
- To monitor the data coherence over time and identify changes that may require action

Procedures for Data anomalies Checks and Remediation

1. Define data anomaly

At a univariate and multivariate level, provide a definition of data anomaly. E.g. in a time series variable, (upper and lower) thresholds can be specified to identify outliers in the data.

2. Data anomaly detection

Provide data anomaly detection methods depending on the type of data (structured and unstructured) in evaluation. Consider different methods for univariate and multivariate anomalies detection.

3. Results Analysis

Confirm the anomalies by investigating the underlying causes. This may involve performing root cause analysis or engaging domain experts to investigate the anomalies.

4. Remediate anomalies issues

Decide what action to take based on the data anomaly detection analysis. This could consist of removing in the dataset outliers or to incorporate them in the whole data set with different weight to reduce their effect on the decision process.

5. Monitor data anomalies

Regularly monitor data anomalies to ensure that they are detected and treated over time and to identify changes that may require action.

EY

Deep Dive: Coherence – Cont.

Anomalies detection and remediation

Univariate anomaly detection and remediation methods

Z-score technique

1. Description

Scale the univariate feature distribution to ensure it has unit variance $z = \frac{x-\mu}{\sigma}$. An outlier may be defined as any observation where $z > 3$.

2. Assumption

The transformed distribution must be approximately normal with unit variance $z \sim N(0,1)$

3. Advantages

- Simple to implement
- Probabilistic in nature – the outlier threshold can correspond to a probability of observing a more extreme value based on the cumulative distribution

4. Disadvantages

- Most feature spaces do not satisfy the normality assumption even after transformation
- The outlier threshold is not determined analytically

Adjusted Tukey Method

1. Description:

Data points lying outside the bounds given by a function of the data's first and third quartiles, IQR, and medcouple (MC) are determined to be outliers.

$$MC = med\frac{(x_j - med(X_n)) - (med(X_n) - x_i)}{x_j - x_i}, x_i \le med(X_n) \le x_j, x_i \ne x_j$$

$$[Q_1 - 1.5e^{-4MC}IQR, Q_3 + 1.5e^{3MC}IQR] \; if \; MC > 0$$
$$[Q_1 - 1.5e^{-3MC}IQR, Q_3 + 1.5e^{4MC}IQR] \; if \; MC < 0$$

2. Assumption: No major assumption

3. Advantages

- Applicable to skewed or non-mound-shaped data since the method makes no distributional assumptions and does not depend on mean or standard deviation

4. Disadvantages

- May not be appropriate for small sample size
- Fails if empirical distribution has too many peaks

Multivariate anomaly detection and remediation methods

Multivariate Kernel density estimators

1. Description: Estimates a multivariate probability density function of a finite sample of n-dimensional data with the following form

$$\hat{f}_H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i)$$

2. Assumption: Choice of the kernel (Gaussian is standard)

3. Advantages

- Lack of histogram binding grid eliminates the density's dependency on the anchor point
- No dependency on bandwidth
- Easier to interpret

4. Disadvantages

- More difficult to compute than histograms

Unsupervised Learnings methods

1. Description: Use unsupervised learning algorithms to identify anomalies at the observation level. Some unsupervised learners support univariate clustering however, generally we assume that the feature space has a dimension of at least two

2. Assumption: Each potential algorithm has a differing set of assumptions and key hyperparameters that need to be tuned

3. Advantages

- Capable of detecting outliers among rich feature sets
- Probabilistic in nature – the outlier threshold can correspond to a probability of observing a more extreme value based on the cumulative distribution

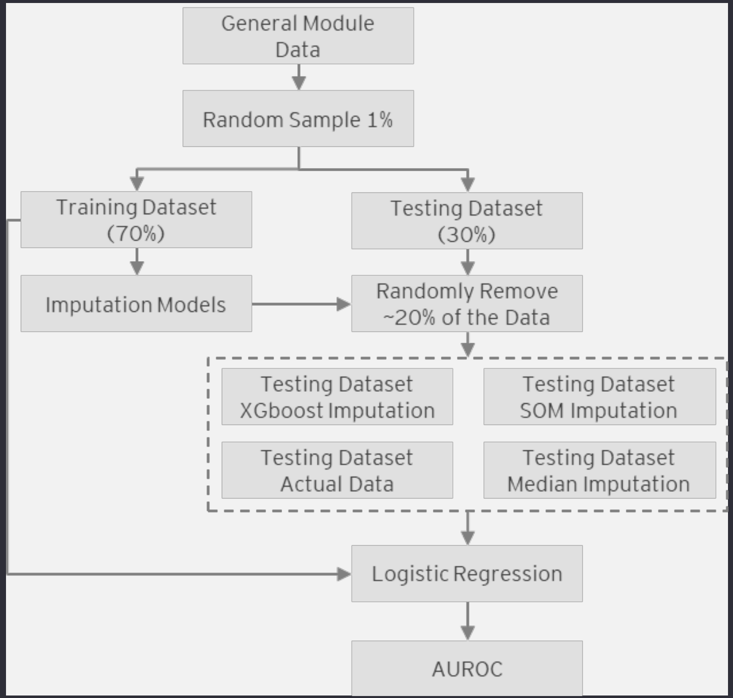
4. Disadvantages

- Computationally much more complex than simple outlier detection
- Interpreting the root cause of the anomaly is opaque for large feature spaces

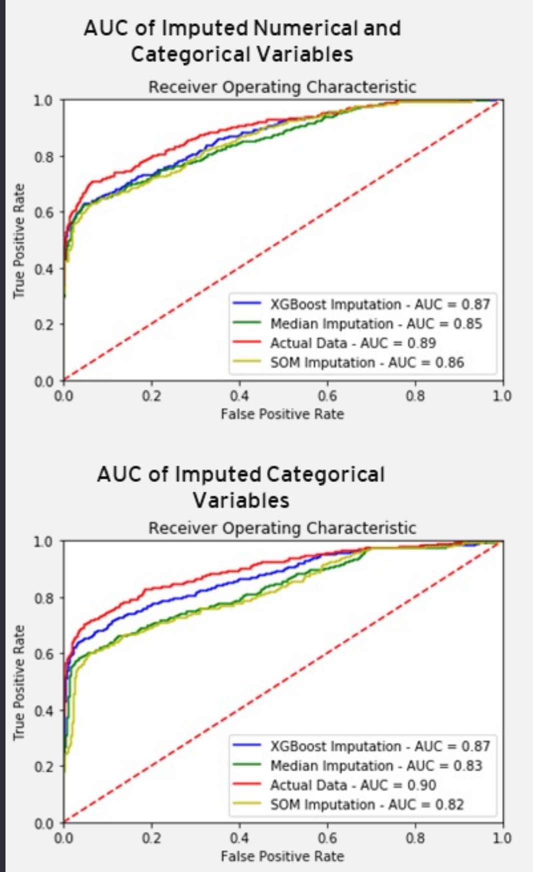
Data Quality Use Cases

Missing Value Imputation for a Large Financial Institution

- The key objective of the Proof of Concept is to enhance data quality (DQ) through missing value imputation and anomaly treatment using Machine Learning (ML) methodologies
- Enhancing data quality for a credit risk model development data set leads to better model performance and better business decision making
- It sets a foundation for establishing a data quality and governance framework that can be scaled and applied to other model development data sets



Missing value imputation based on selected features



Key Findings

- Gradient Boosting Trees imputed test set has **2% higher AUROC** than baseline model and an AUROC closest to that of the actual test set, which is the optimal desired outcome.
- Gradient Boosting Trees are **better at imputing categorical missing values** than the baseline model, which replaces categorical missing values with their mode.

Anomaly Detection for a Large Financial Institution

Anomaly Detection

The anomaly detection Proof of Concept consists of the following steps:

1. Four anomaly detection models are developed
2. For each use case dataset, apply all four anomaly detection models; obtain four sets of anomaly indicators e.g., Gaussian Mixture Model (GMM) anomaly, Isolation Forest (IF) anomaly etc.
3. Samples are reviewed by business subject matter experts (SME) for feedback
4. Use feedback to enhance anomaly detection with Active Learning

ID	GMM	IF	DBSCAN	SOM
1	●	●	●	●
2	●	●	○	●
⋮				
100	○	○	○	○

Anomaly ●

No Anomaly ○

1 Gaussian Mixture Model

2 Isolation Forest

3 DBSCAN

4 Self-Organizing Map

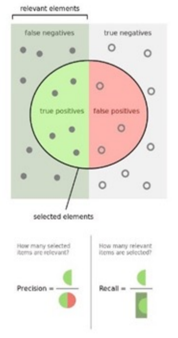
Insights gained during model development

- All datasets involved sampling to limit dataset size per Sandbox constraints
- Robust computational resources are required for anomaly detection models
- Business SME feedback is critical to guide model tuning and DQ success
- The ML based solution is a record level outlier detection model that considers relationships among all available fields and these relations should be curated by SME expertise.
- In certain cases, it is about detecting erroneous records in one time snapshot, in others, it is about detecting inconsistencies through time

Anomaly Identification Evaluation Metric

Account	Static Rule (Current)	ML Identification	Expert Verification
1	Anomaly	Anomaly	True Anomaly
2	Non - anomaly	Anomaly	True Anomaly
3	Non - anomaly	Non - anomaly	Normal
4	Non - anomaly	Anomaly	True Anomaly
⋮	⋮	⋮	⋮

Machine Learning (ML) Model	Baseline (BL) Model	
	True Positive/ True Positive	True Positive/ False Positive
	False Positive/ True Positive	False Positive/ False Positive



Precision $= \dfrac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$

Recall $= \dfrac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$

* ML identification is obtained based on a bagging model across the four models

EY

Anomaly Detection for a Large Financial Institute (contd.)

Active Learning – Approach and Lessons Learned

Active Learning: N-step model that is ideal for unlabelled datasets and maximally converges to the performance of a supervised learning algorithm

1 Step 1

- 3 subsets of data: labelled, unlabelled & hold-out
- Train Support Vector Machines (SVM) on existing data labels
- E.g. *anomaly, non-anomaly flag provided in SME feedback* or *anomalies flagged by DQ rules*

2 Step 2

- SVM is used to score N samples from unlabelled dataset
- Sample is based on 'entropy-score' (confidence-level)
- High entropy -> uncertain
- Low entropy -> certain
- Sample based on entropy threshold

3 Step 3

- Additional SME feedback on Step 2 sample
- Add labels to Step 2 sample: true positive/false positive
- Retrain model

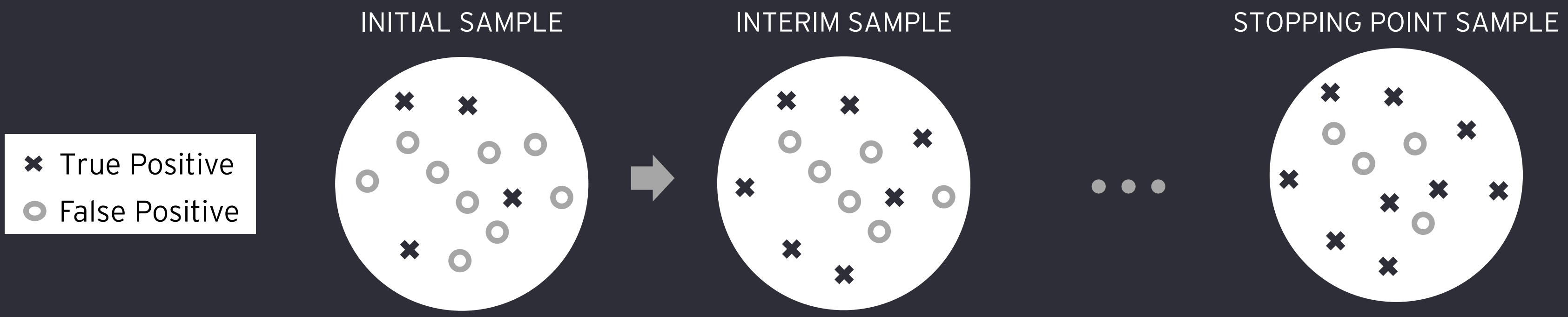
. . .

N Step N

- Repeat Step 2 & Step 3 until Active Learning model converges
- Rate of true positives much higher at stopping point Step N

Key Takeaways:

- Limit the dimensions of the feature space
- Prioritize numeric fields over categorical
- Ensure sufficient balance of the classes prior to the first round of sampling
- Parametrize the uncertainty of SME or "Oracle" feedback



EY

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

ey.com/ca

Authors

Mario Schlener

Partner, Lead Financial Services Risk Management Practice and Enterprise Risk Strategy, EY Canada

EY Global FS Risk Technology, Alliance, Innovation Lead

mario.schlener@ca.ey.com

Tarek Elguebaly, Ph.D.

Executive Director, Data And AI and Risk Tech Lead, Data and analytics, EY Canada

tarek.Elguebaly@ca.ey.com

Yara Elias, Ph.D.

Senior Manager, AI Risk Lead, Financial Services Risk Management, EY Canada

yara.elias@ca.ey.com

Liang Hu, Ph.D.

Manager, Responsible AI and AI Risk , Financial Service Risk Management, EY Canada

liang.Hu@ca.ey.com