

Data Quality Management Guidebook

– Maximize the value of data and Artificial Intelligence –

AISI Japan
AI Safety
Institute

IPA Information-technology
Promotion
Agency, Japan
Digital Infrastructure Center

Garbage in, Garbage out.

AI is amplifying existing data problems.
Quality of data directly impacts on AI outputs.

About This Guide - Background

- ♦ AI systems are becoming deeply integrated into a wide range of social and industrial domains, influencing our daily lives and shaping important decisions. At the foundation of their reliability and safety lies, above all, the quality of the data they rely on. No matter how advanced an AI system may be, inaccurate or unreliable data can lead to incorrect or harmful outcomes. Therefore, **ensuring data quality is the first and most essential step in securing trustworthy AI.**
- ♦ However, approaches to evaluating and managing data quality differ across organizations and sectors. In addition, the nature of data used in AI systems changes throughout its life cycle, meaning that a single, uniform standard is rarely sufficient.
- ♦ **This guidebook aims to provide a shared framework and way of thinking for the practical implementation of data quality management.** It is informed by international standards while also incorporating practical columns and AI-specific considerations, with the goal of making it applicable and useful in real-world settings.

About This Guide – How to use

- ♦ **As a guideline for policy development**
 - Use this guide as a foundational reference or checklist when developing organizational policies for ensuring data quality.
 - Examples include data governance policies, AI ethics principles, and quality management plans.
- ♦ **As a practical tool for project operations**
 - Use it to clarify what needs to be checked and who is responsible at each stage of the AI development process—such as data collection, annotation, and model training.
- ♦ **As educational and shared learning material**
 - Use it as training content to help all stakeholders involved in handling data—providers, holders, developers, and users—build a common understanding.

This document is not a legally binding standard, but a guidebook intended to share practical knowledge on data quality. We hope readers will refer to it flexibly—according to their roles, needs, and purposes—and adapt it to best suit their own organizations and projects.

About This Guide – Target Audience

- ◆ Data quality management cannot be completed by a single role or department; it requires collaboration across diverse functions both inside and outside the organization.

Role	Expected Use (example)
Executives / Policy decision-makers	Use as guidance for establishing policies and organizational structures to ensure data quality across the entire organization and society.
Data managers / Data governance officers	Use as standards and process design references for maintaining and evaluating data quality.
Data and AI engineers	Use as practical guidance for data quality management and validation throughout the AI development life cycle.
Business planners / Service designers	Refer to it when defining and aligning necessary data quality requirements in service design and risk management.
Auditors / Researchers / Educators	Use as foundational material for evaluating, educating, and researching data quality management.

Contents

I. Background and Summary	p. 8	III. Perspective	p. 38
① Background	p. 9	III. Perspective - Process View(Operation)	p. 39
② Data Quality and Trust in Society	p. 10	Process	p. 40
③ AI System and Data Quality	p. 11	Role of stakeholders	p. 41
④ AI Safety and Data Quality	p. 12	1.Data Planning	p. 42
⑤ Benefits from High-quality Data	p. 13	Column: Reference mode	p. 48
⑥ Risks from Low-quality Data	p. 14	Column: Data dictionary	p. 49
⑦ Causes of Low-quality data	p. 15	Column: Data modeling	p. 50
⑧ Objectives and Goals	p. 16	Column: DCAT	p. 51
⑨ Principles	p. 17	2.Data Acquisition	p. 52
⑩ Scope	p. 19	Column: Improvement of input method	p. 56
⑪ Stakeholders in Data and AI	p. 20	Column: Provenance data	p. 57
⑫ Methods to Ensure Quality	p. 21	3. Data Preparation	p. 58
⑬ Relationship between ISO Standards and Framework	p. 22	Column: IDs	p. 64
⑭ Quality-related Guidelines in Japan	p. 23	Column: Data cleansing	p. 65
Column: Service Quality in Japan	p. 25	Column: Data matching	p. 66
II. Concept of Data Quality Management Framework	p. 26	Column: Data completion	p. 67
① Management, Governance and Maturity	p. 27	Column: Synthetic data	p. 68
② Framework Overview	p. 28	Column: Content Transparency Tech.	p. 70
③ Three Views of the Framework	p. 29	Column: Indication of data quality	p. 71
④ Characteristics	p. 30	4. Data Processing	p. 72
⑤ Process	p. 31	5. AI System	p. 75
⑥ Contents and Data Management	p. 32	Column: Bias	p. 78
⑦ Data Quality Management for AI	p. 36	Column: Model collapse	p. 79
		6.Evaluation of Output	p. 80
		Column: ASOT	p. 83
		7. Deliver the Result	p. 84
		Column: Social and human impact	p. 87
		8. Decommissioning	p. 88
		9. Action throughout the Life Cycle	p. 91
		Column: Access control	p. 94
		Column: Improving data quality using AI	p. 95

Continued on the next page

Contents

III. Perspective - Governance Cycle View

Governance Cycle	p. 96
Data Quality Planning	p. 97
Data Quality Control	p. 98
Data Quality Assurance	p. 100
Data Quality Improvement	p. 102
Data-related Support	p. 104
Resource Provision	p. 106

Column: Human resource and team

III. Perspective - Gateway View(Characteristics)

Verification of Data Quality Characteristics.	p. 110
List of Characteristics	p. 111
Accuracy	p. 112
Completeness	p. 113
Consistency	p. 114
Credibility	p. 115
Currentness	p. 116
Accessibility	p. 117
Compliance	p. 118
Confidentiality	p. 119
Efficiency	p. 120
Precision	p. 121
Traceability	p. 122
Understandability	p. 123
Availability	p. 124
Portability	p. 125
Recoverability	p. 126
Auditability	p. 127
Balance	p. 128
Diversity	p. 129
Effectiveness	p. 130
Identifiability	p. 131
Relevance	p. 132

Representativeness	p. 135
Similarity	p. 136
Timeliness	p. 137
Sensor data	p. 138
Characteristics of Sensor Data	p. 139
Time Characteristics of Sensor Data	p. 140
Challenges in using Multiple Sensors	p. 141
Processing with Cloud-Edge-IoT	p. 144

Postscript

Message	p. 145
About us and this document	p. 146
References	p. 147
Version History	p. 148

I. Background and Summary

① Background

- ◆ AI is rapidly transforming society by enabling useful and efficient services.
- ◆ However, many people remain concerned about the use of AI, and ensuring the accuracy of AI outputs is essential for safe adoption. To achieve this, it is necessary to improve the quality of the data used for training and processing.
- ◆ AI society
 - AI systems rely on large volumes of data for training and for producing outputs such as predictions, recommendations, and explanations.
 - The quality of data directly impacts AI system performance and effectiveness.
- ◆ Data-driven society
 - High data quality is essential in a data-driven society because it ensures accurate and reliable information. This supports informed decision-making, reduces errors, and minimizes risks.
 - Quality data also improves customer satisfaction and helps organizations comply with regulations, protecting them from legal and financial issues.
- ◆ Ensuring data quality is essential to ensuring the reliability of AI services

② Data Quality and Trust in Society

- ◆ Ensuring trust is crucial for the use of AI and data.
- ◆ Data is the foundation for AI. If that is not correct, the reliability of the entire process will be compromised.

Element of Trust in the data-driven society
<ul style="list-style-type: none"> • Service quality • Ethics • Transparency • Accountability • Privacy protection • Security • Data quality • Partnership and collaboration

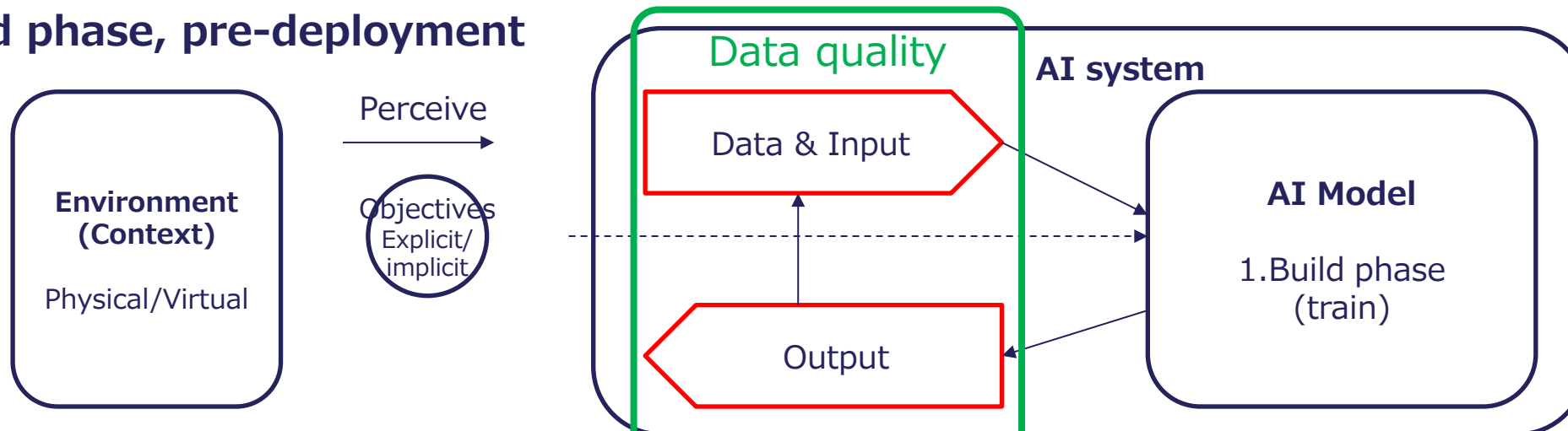
Elements of Trustworthy AI
<ul style="list-style-type: none"> • Accuracy and Reliability • Ethics • Transparency • Accountability • User-Centric Design • Safety • Continuous Improvement

- ◆ Data quality is the foundation of AI excellence, enabling trustworthy AI, which drives user adoption and engagement.

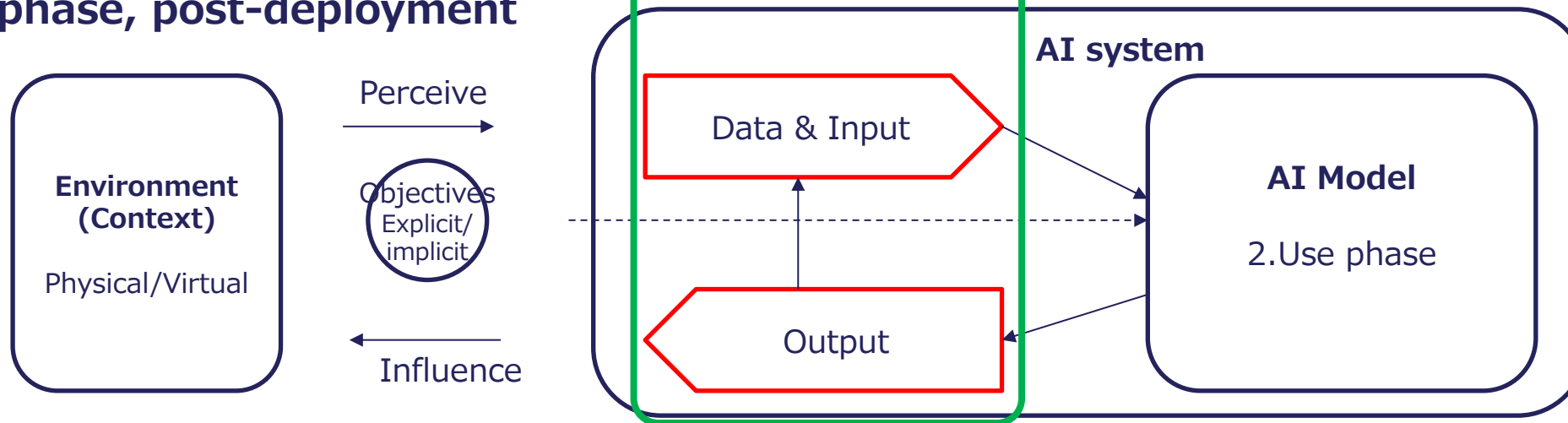
③ AI System and Data Quality

- ◆ In AI systems, data plays an important role not only at the time of use, but also in the preparatory stage.

Build phase, pre-deployment



Use phase, post-deployment



[Updates to the OECD's definition of an AI system explained - OECD.AI](#)

④ AI Safety and Data Quality

- Low-quality input data affects many critical aspects, including the performance of AI models and the credibility, consistency, and accuracy of outputs. As a result, public trust may be undermined.
- Data quality management is essential for ensuring the appropriate and safe use of AI.
- These perspectives illustrate that data quality underpins multiple dimensions of AI safety—including robustness, fairness, verifiability, and security.

Guide to Evaluation Perspectives on AI Safety
(2024-09-25, Japan AISI (2025-03-31 Revised))

Evaluation Perspectives on AI Safety	
Control of Toxic Output	Ensuring Security
Prevention of Misinformation, Disinformation and Manipulation	Explainability
Fairness and Inclusion	Robustness
Addressing to High-risk Use and Unintended Use	Data Quality
Privacy Protection	Verifiability

https://aisi.go.jp/output/output_framework/guide_to_evaluation_perspective_on_ai_safety/

⑤ Benefits from High-quality Data

1. Enhanced Accuracy

- Higher data quality results in more accurate and consistent AI predictions and decisions, reducing the risk of flawed outcomes.

2. Better Insights

- Quality data enables the identification of actionable insights and reliable trends, empowering better decision-making processes.

3. Increased Efficiency

- Clean, high-quality data reduces time spent on preprocessing tasks such as cleaning, correction, and normalization, allowing teams to focus on higher-value activities.

4. Improved User Experience

- Accurate and relevant data enhances the overall experience for end-users by delivering precise, personalized, and timely results.

5. Reduced Errors

- Minimizing inconsistencies and errors in data reduces inaccuracies in AI outputs, improving reliability and performance.

6. Cost Savings

- Investing in high-quality data decreases costs related to error correction, reprocessing, and mitigating downstream issues caused by low-quality data.

7. Compliance and Security

- Maintaining high-quality data ensures adherence to data governance, privacy regulations, and security standards, safeguarding organizational integrity.

8. Enhanced Trust

- Consistently high-quality data builds trust in AI systems among stakeholders, fostering confidence in automated processes and decisions.

9. Scalability

- High-quality data serves as a robust foundation for scaling AI systems, ensuring consistent performance as the system expands.

10. Facilitates Model Training and Updates

- Clean and structured data simplifies the process of training and fine-tuning AI models, accelerating development cycles and improving model adaptability.

11. Competitive Advantage

- Organizations leveraging high-quality data gain a significant competitive edge by delivering superior AI-driven products and services.

12. Ecosystem Integration

- High-quality data enables seamless integration with other systems and platforms, ensuring interoperability and streamlined workflows.

⑥ Risks from Low-quality Data

1. Decision-Making Errors

- Inaccurate or incomplete data can lead to flawed analyses, resulting in incorrect conclusions, strategic missteps, and poor decision-making at all levels of the organization.

2. Operational Inefficiency

- Time and resources must be allocated to cleaning and correcting low-quality data, delaying AI model deployment and reducing operational effectiveness.

3. Decreased Customer Satisfaction

- Inaccurate or incomplete customer data can lead to poor personalization, unmet expectations, and diminished trust in services or products.

4. Increased Costs

- Low-quality data increases costs through error correction, reprocessing, re-training of AI models, and potential financial losses due to incorrect predictions or recommendations.

5. Legal and Regulatory Risks

- Non-compliance with data protection laws and regulations (e.g., GDPR, CCPA) can lead to significant fines and lawsuits.

6. Reputation Damage

- Errors caused by low-quality data can damage trust and credibility with customers, partners, and stakeholders, potentially leading to long-term brand erosion.

7. Competitive Disadvantage

- Competitors with better-quality data can outperform in key areas such as customer insights, operational efficiency, and market responsiveness, leaving organizations lagging behind.

8. Lost Opportunities

- Missed insights and trends due to low-quality data can result in failure to seize new market opportunities or innovate effectively.

9. Model Performance Degradation

- AI models trained on low-quality data may exhibit bias, unreliability, or harmful behavior, leading to ethical concerns and reduced effectiveness in real-world applications.

10. Security Risks

- low-quality data may inadvertently include vulnerabilities or errors that malicious actors can exploit, leading to potential security breaches or misuse of sensitive information.

11. Stakeholder Distrust

- Internal teams and external stakeholders may lose confidence in AI systems if the data quality consistently undermines reliability and results.

⑦ Causes of Low-quality Data

- ♦ **Human error:** Mistakes made during data entry or data processing.
- ♦ **Lack of standardization:** Inconsistent formats and standards across data sources.
- ♦ **Inadequate data governance:** Lack of policies and procedures for data management.
- ♦ **Outdated systems:** Using legacy systems that are not equipped to handle modern data requirements.
- ♦ **Incomplete data collection:** Missing or incomplete data due to inadequate data collection processes.
- ♦ **Insufficient training:** Lack of proper training for personnel handling data.
- ♦ **Inadequate data integration:** Issues arising from merging data from different sources.
- ♦ **Limited quality control:** Inadequate checks and validation processes for data quality.
- ♦ **Unverified sources:** Using data from unreliable or unverified sources.
- ♦ **Bias in data collection:** Collecting data that reflects inherent biases.
- ♦ **Technical glitches:** Errors caused by software bugs or hardware failures.
- ♦ **Lack of documentation (e.g., data dictionaries, data lineage diagrams):** Inadequate documentation leading to misunderstandings or misinterpretations of data.
- ♦ **Misunderstandings:** Data entry errors due to misunderstanding the requirements or instructions.
- ♦ **Malicious actions:** Deliberate tampering or poisoning of data by someone with harmful intent.

⑧ Objectives and Goals

The objectives and goals of data quality management are as follows:

◆ Objective

Enhancing data quality for AI ensures that decisions, predictions, and recommendations are based on accurate, consistent, and reliable information. **High-quality data empowers organizations to leverage AI effectively, reduce risks, and unlock its full potential.** On a societal level, **improving data quality fosters trust in AI systems, promotes ethical data use, and ensures fairness.** It establishes a foundation for innovative applications in healthcare, education, and public services, ultimately **contributing to economic growth and a better quality of life for all.**

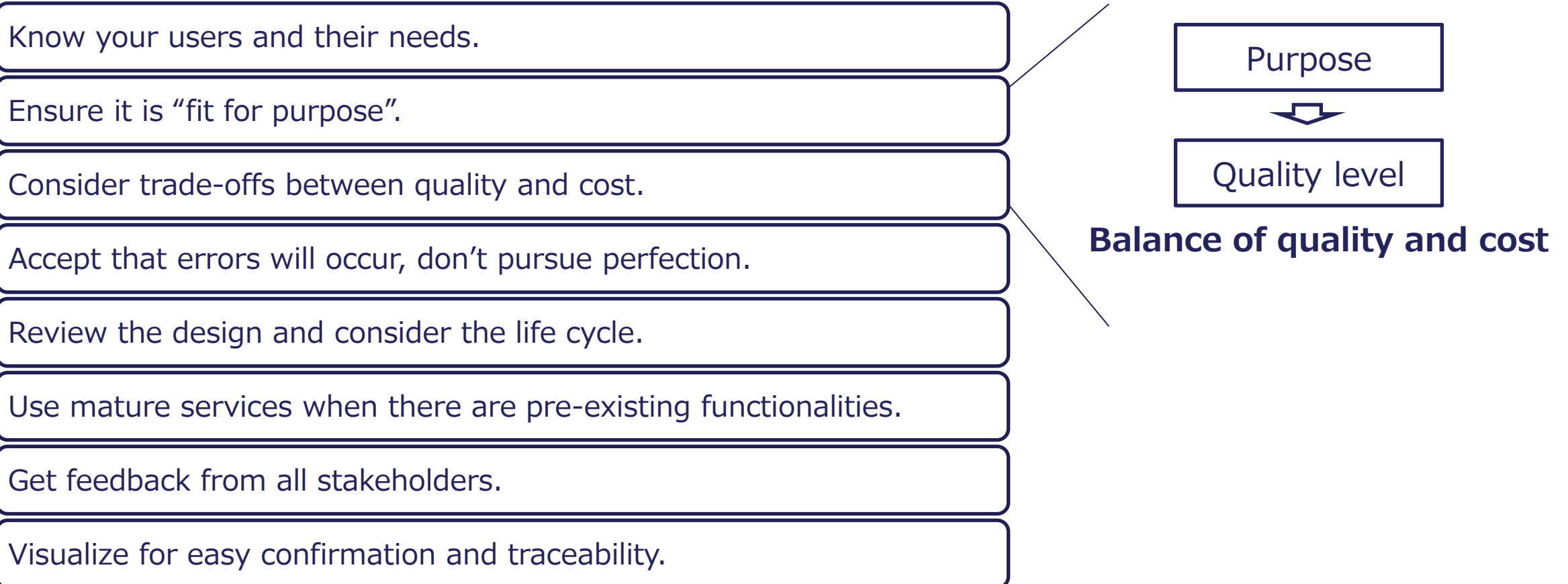
◆ Goal

The goal of improving data quality is to create a robust ecosystem where accurate, secure, and accessible data enables AI to achieve its best performance.

This involves establishing standardized practices for data collection, cleaning, and validation. At the societal level, the aim is to democratize data-driven innovation, ensuring that all sectors benefit equally. By achieving this, we can promote transparency, enhance decision-making processes, and accelerate progress toward a sustainable and inclusive digital society where technology serves humanity responsibly and effectively.

⑨ Principles

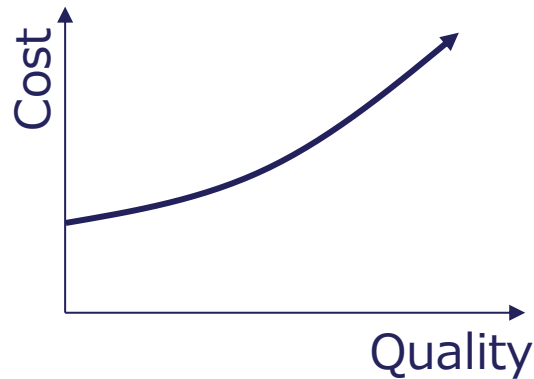
- ◆ This guidebook considers the following as the principles of Data Quality.



⑨ Principles - Quality and Cost

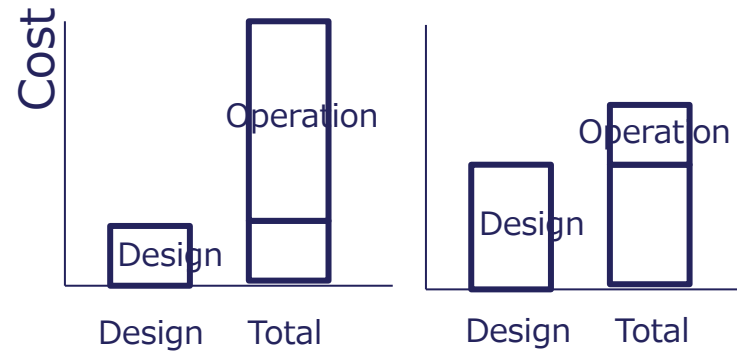
- ◆ When planning for quality, costs must be considered.

Higher quality usually costs more.



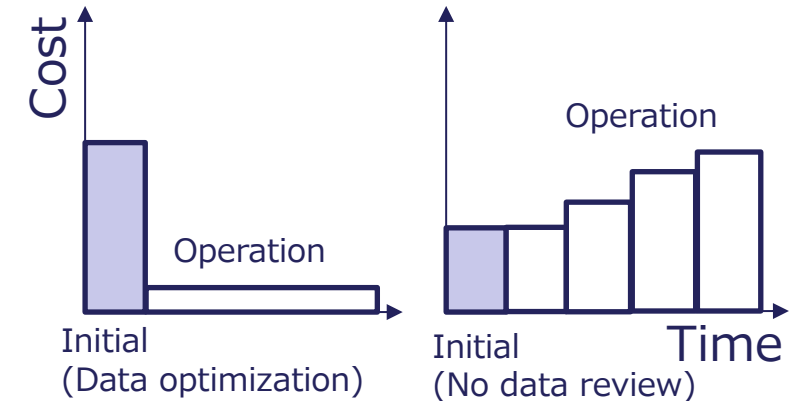
If you raise the quality above the target, it will cost more.

It is important to incorporate quality in early processes such as design



If the design is appropriate, operating costs and total costs will decrease, and you will be able to respond quickly to changes.

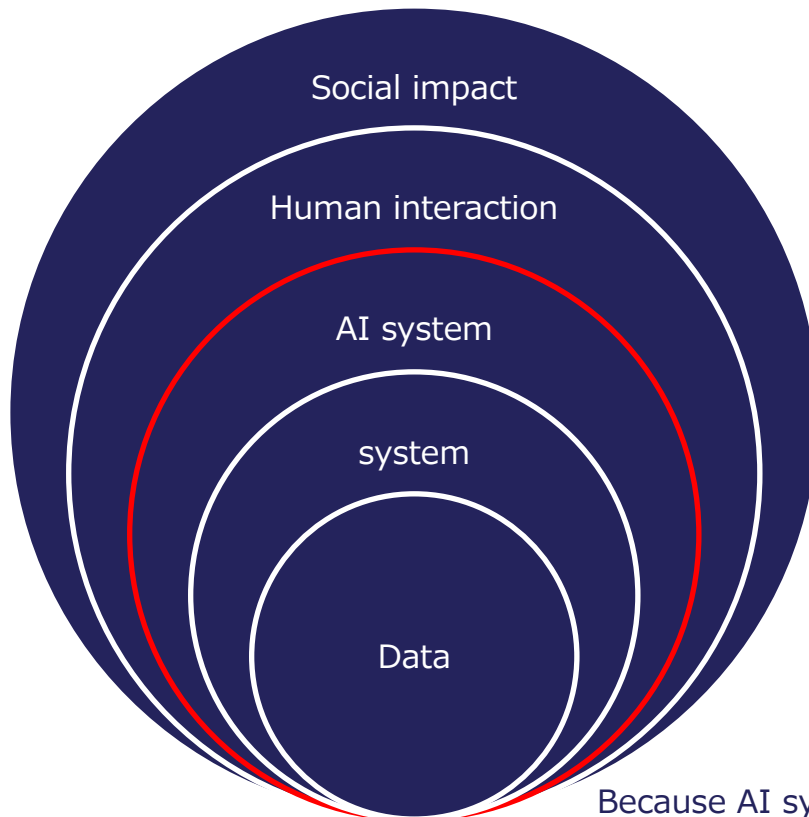
Consider not only initial costs but also operational costs



If you optimize your data when you implement the system, your operating costs can be reduced.

⑩ Scope

- ◆ AI and data are important parts of society, but the scope is broad. They need to be considered in the context of the whole picture of society.



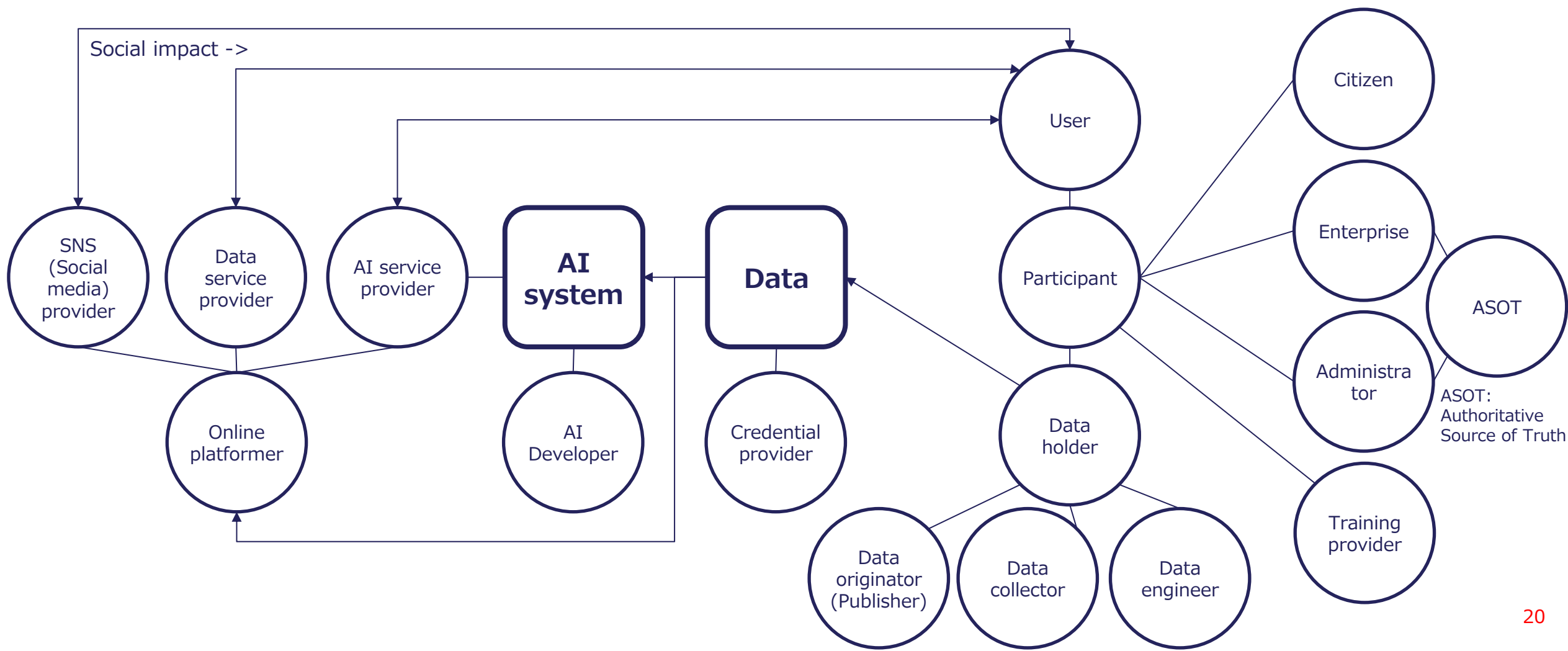
- ◆ Data types
 - Text
 - Numeric
 - Image
 - Video
 - Sound
- ◆ AI systems
 - Narrow AI(task-specific)
 - Generative AI / Foundation Models (large-scale, multi-task)

Note: Generative AI can also be used for task-specific applications.

Because AI systems include training data, it can handle a wider range of data than traditional systems.

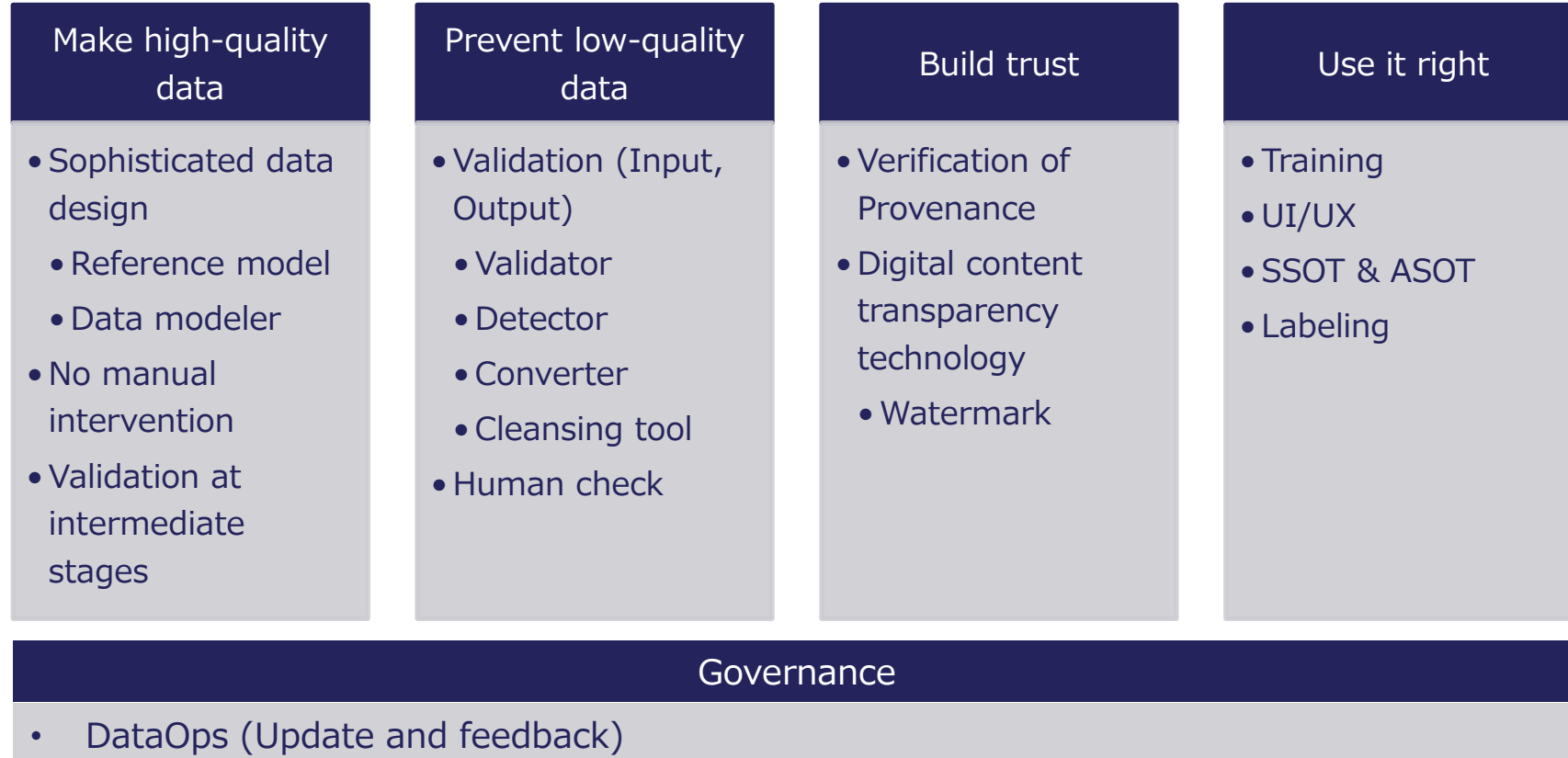
⑪ Stakeholders in Data and AI

- ◆ Stakeholders are diverse. They are characterized by being both users of AI and suppliers of data.



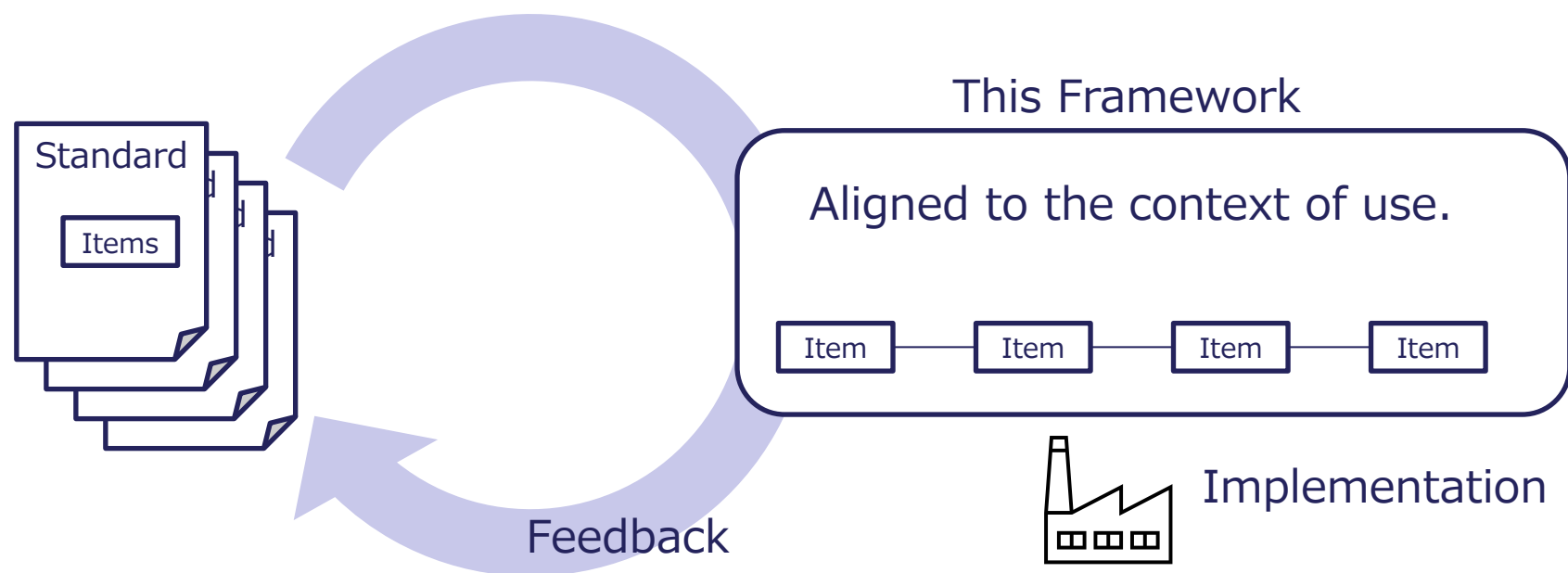
⑫ Methods to Ensure Quality

- ◆ The following methods are the main approaches to ensure data quality.



⑬ Relationship between ISO Standards and Framework

- ◆ This framework will organize and guide many standards on data quality in a way that makes them easy to implement.
- ◆ It does not create new standards, but is only a practical guide, and synergies can be achieved by feeding the results of implementation back into the standardization process.



⑭ Quality-related Guidelines in Japan

- ◆ In Japan, several guidelines concerning AI quality have been established.

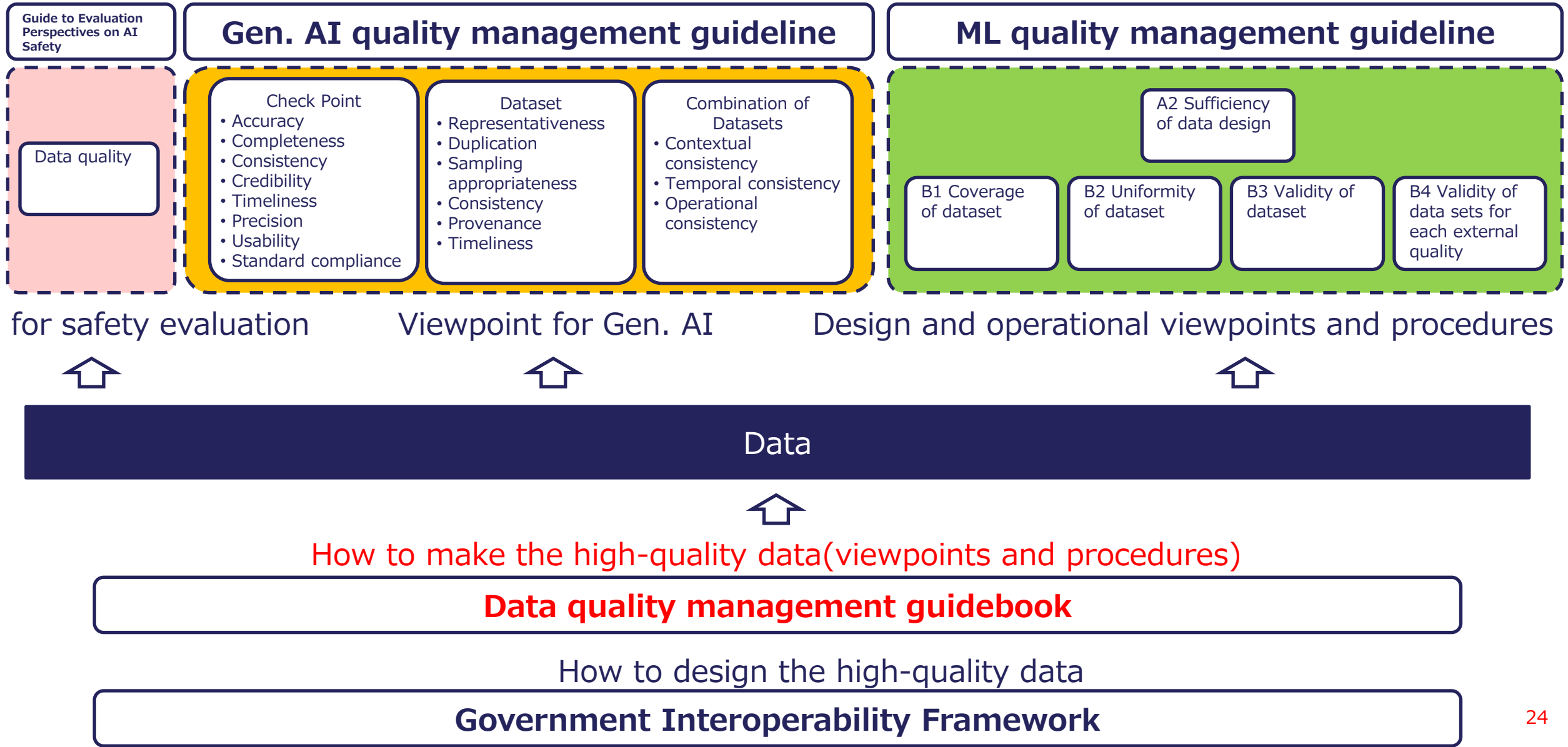
AI Guidelines for Business (MIAC, METI)

		Purpose of AI system	
		Generation	Recognition and prediction
Type of AI model	Generation	<ul style="list-style-type: none"> • Gen.AI Quality Management Guideline(AIST) • Guide to Evaluation Perspectives on AI Safety(AISI) 	Both AI system guideline
	Recognition & prediction	Both AI system guideline	Machine Learning Quality Management Guideline(AIST)

Data quality Management Guideline (AISII, IPA)

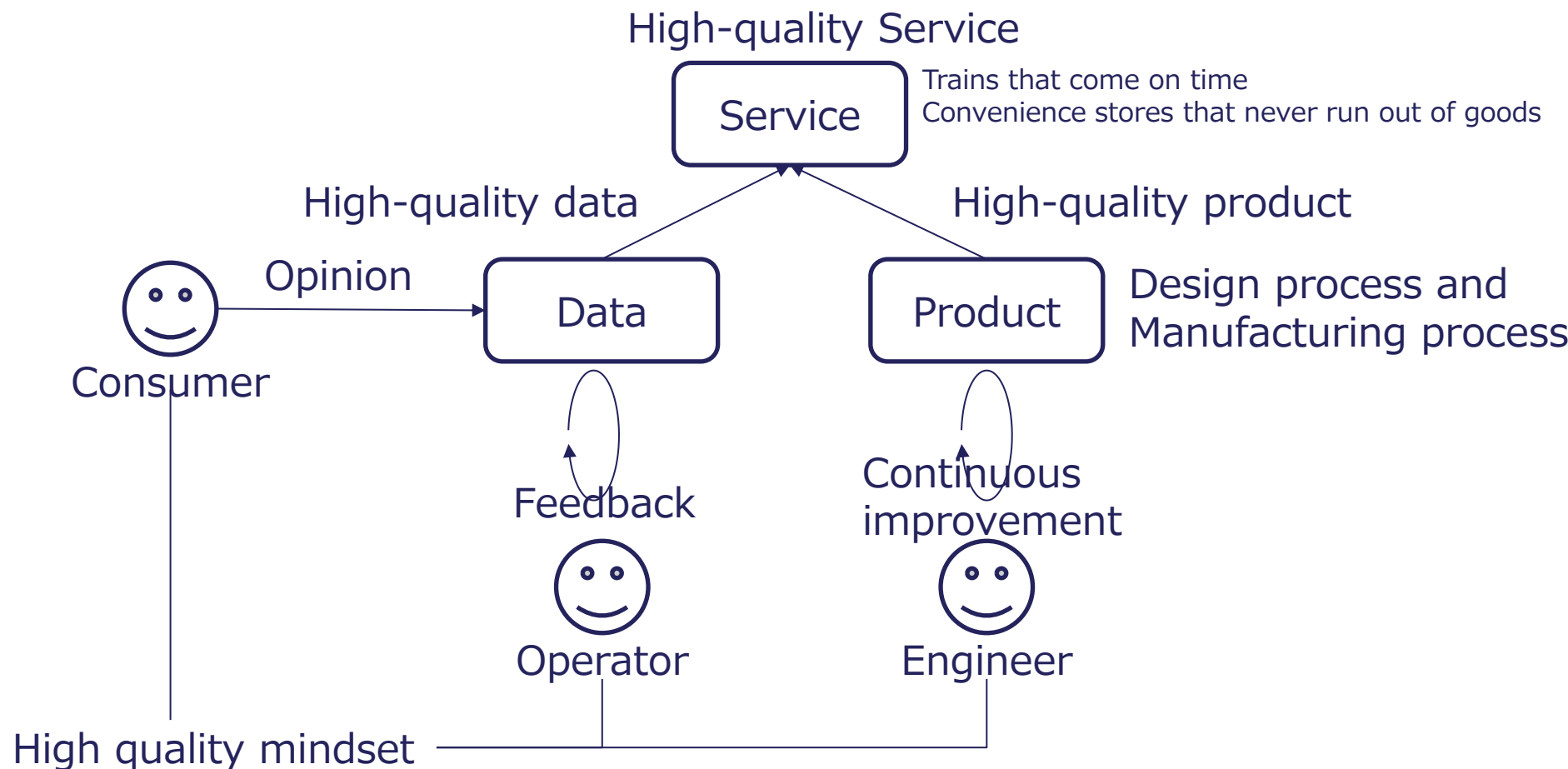
Government Interoperability Framework (Digital Agency)

⑭ Quality-related Guidelines in Japan - Data View



Column: Service Quality in Japan

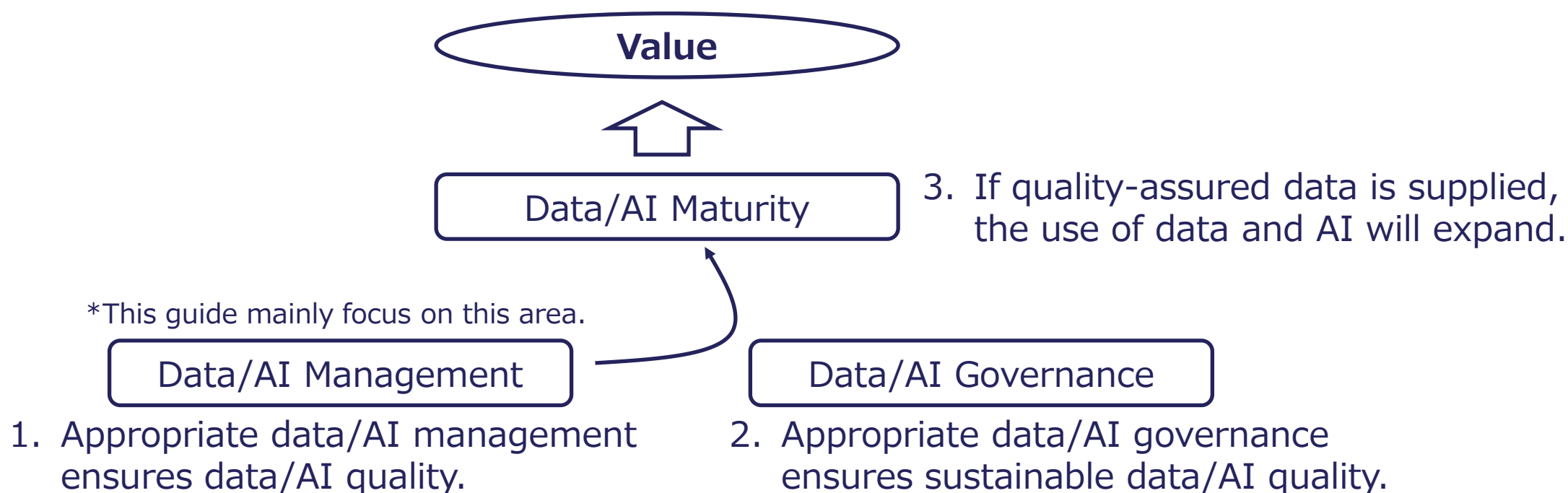
Japan's high service quality is realized based on the quality of its products, people and data.



II. Concept of Data Quality Management Framework

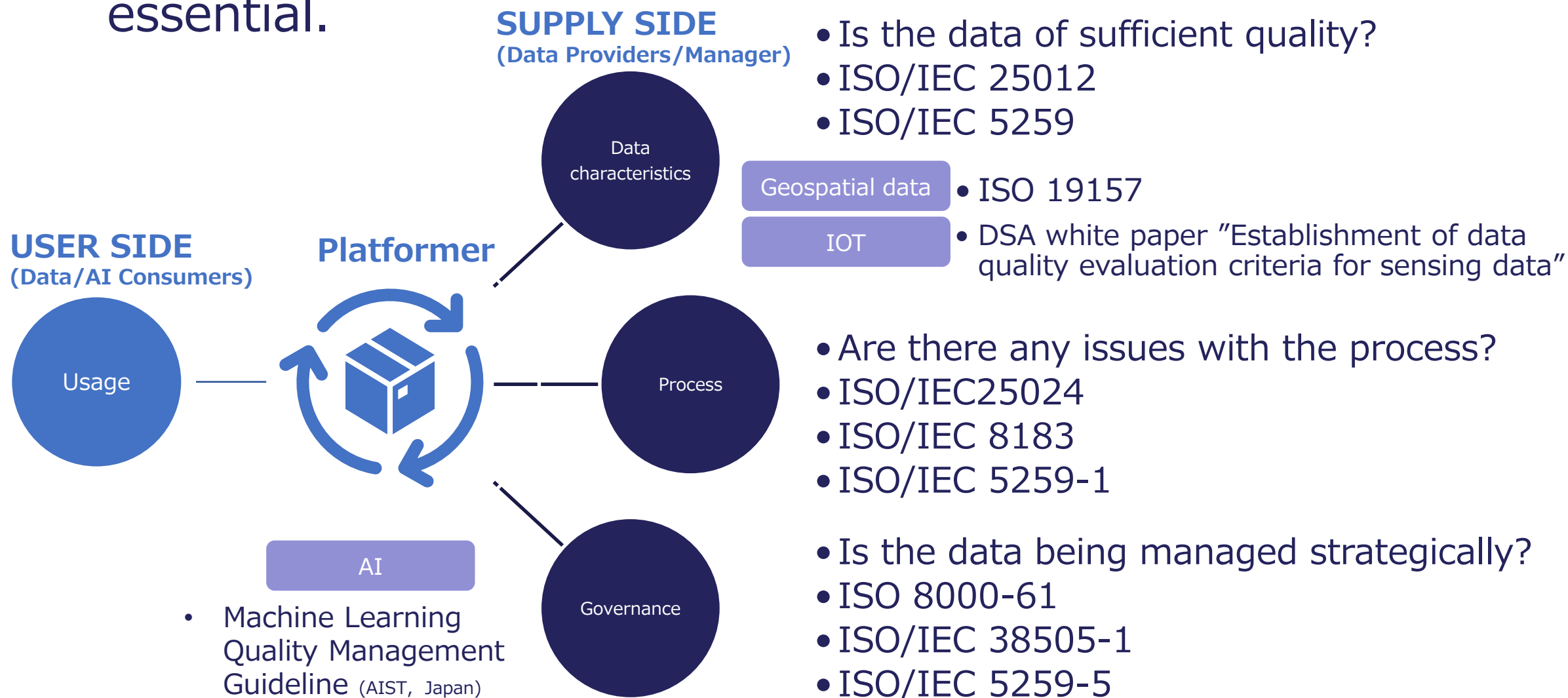
① Management, Governance and Maturity

- ◆ Continuously providing high-quality data is essential to promote the widespread adoption of AI. Effective management and governance are key to maintaining data quality and achieving maturity in data and AI utilization.



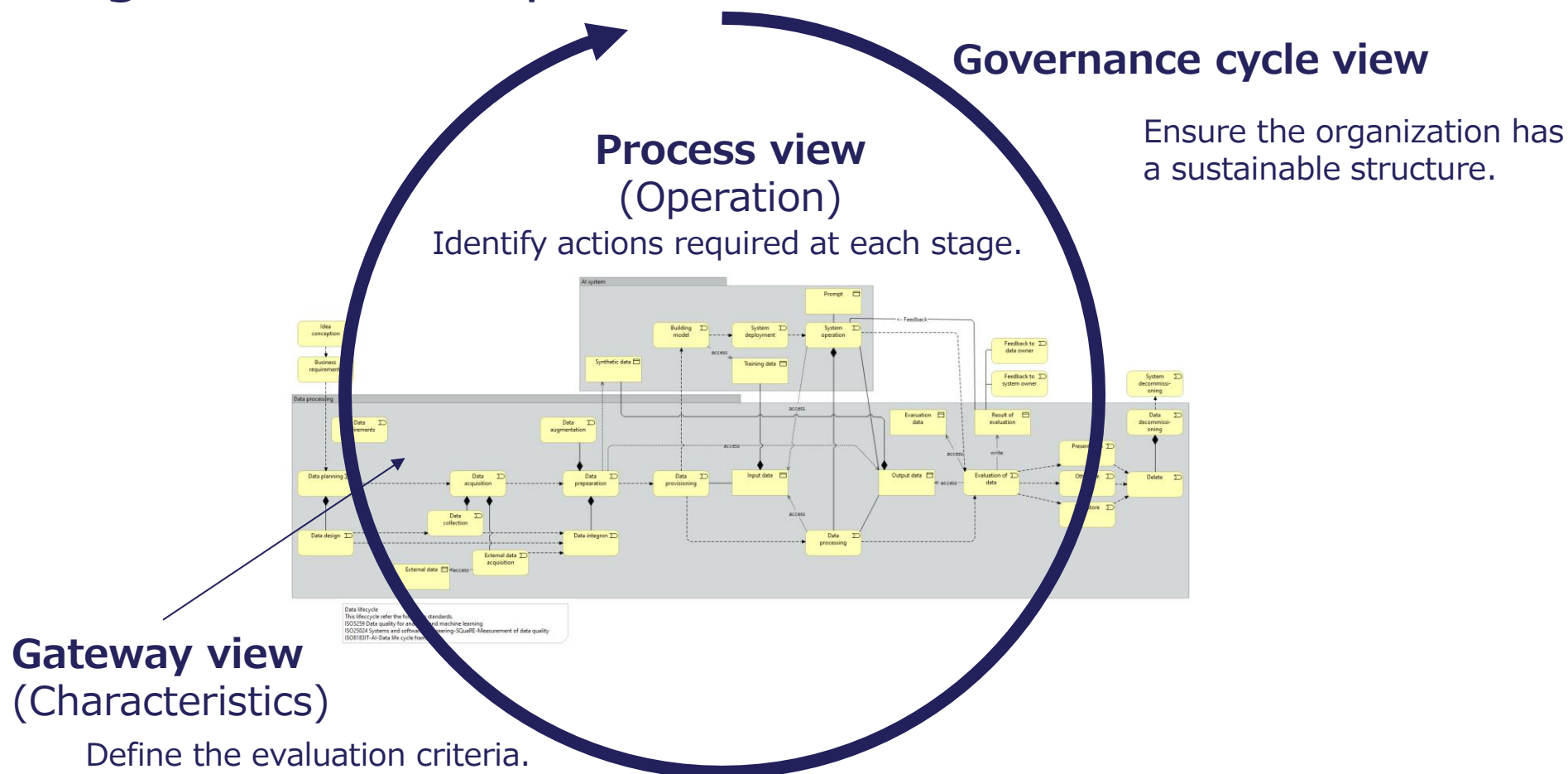
② Framework Overview

- ◆ Data connects to the world, so global interoperability is essential.



③ Three Views of the Framework

- ◆ This framework consists of 3 views, covering both management and operations.



④ Characteristics

Problems

- Inaccurate data
- Unformatted data
- Out-of-date data
- Data from unknown sources
- Poisoned data
- Spike data
- Non-calibrated data
- Mis/Dis/Mal-information
- Bias
- Incomplete data
- Duplicate data
- Inconsistent data
- Irrelevant data
- Corrupted data
- Unclear data



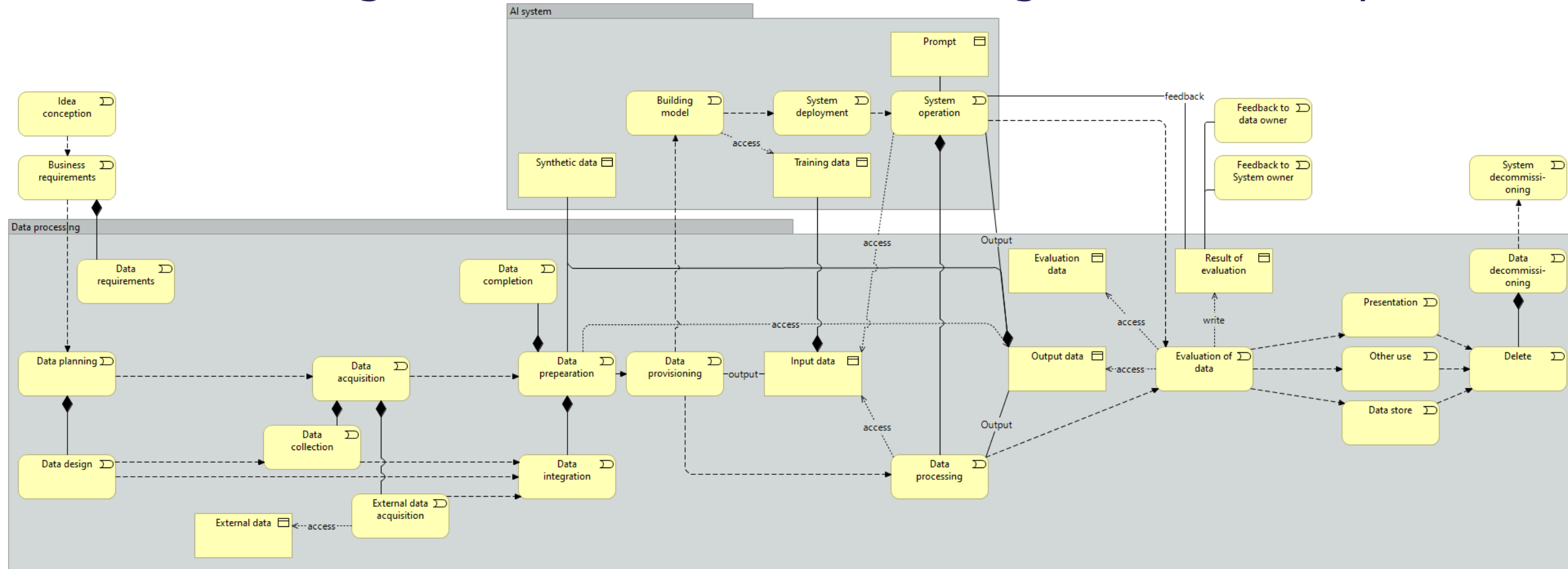
Data Quality Characteristics

(ISO/IEC 25012 and ISO/IEC 5259-2)

- | | |
|-----------------------|------------------------|
| 1. Accuracy | 13. Availability |
| 2. Completeness | 14. Portability |
| 3. Consistency | 15. Recoverability |
| 4. Credibility | 16. Auditability |
| 5. Currentness | 17. Balance |
| 6. Accessibility | 18. Diversity |
| 7. Compliance | 19. Effectiveness |
| 8. Confidentiality | 20. Identifiability |
| 9. Efficiency | 21. Relevance |
| 10. Precision | 22. Representativeness |
| 11. Traceability | 23. Similarity |
| 12. Understandability | 24. Timeliness |

⑤ Process

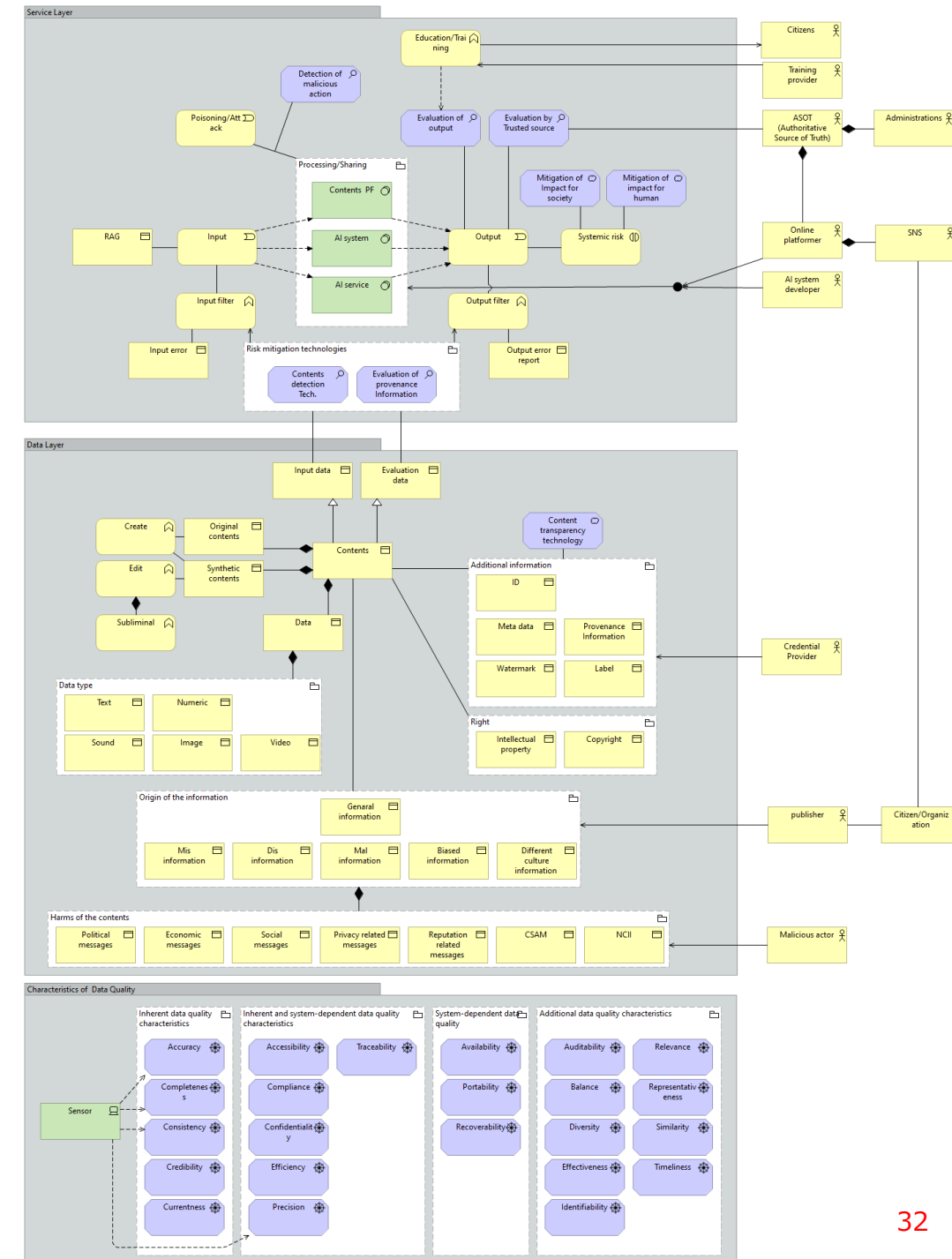
- There are various standards for defining the life cycle. Here, we will organize them into the following detailed life cycle.



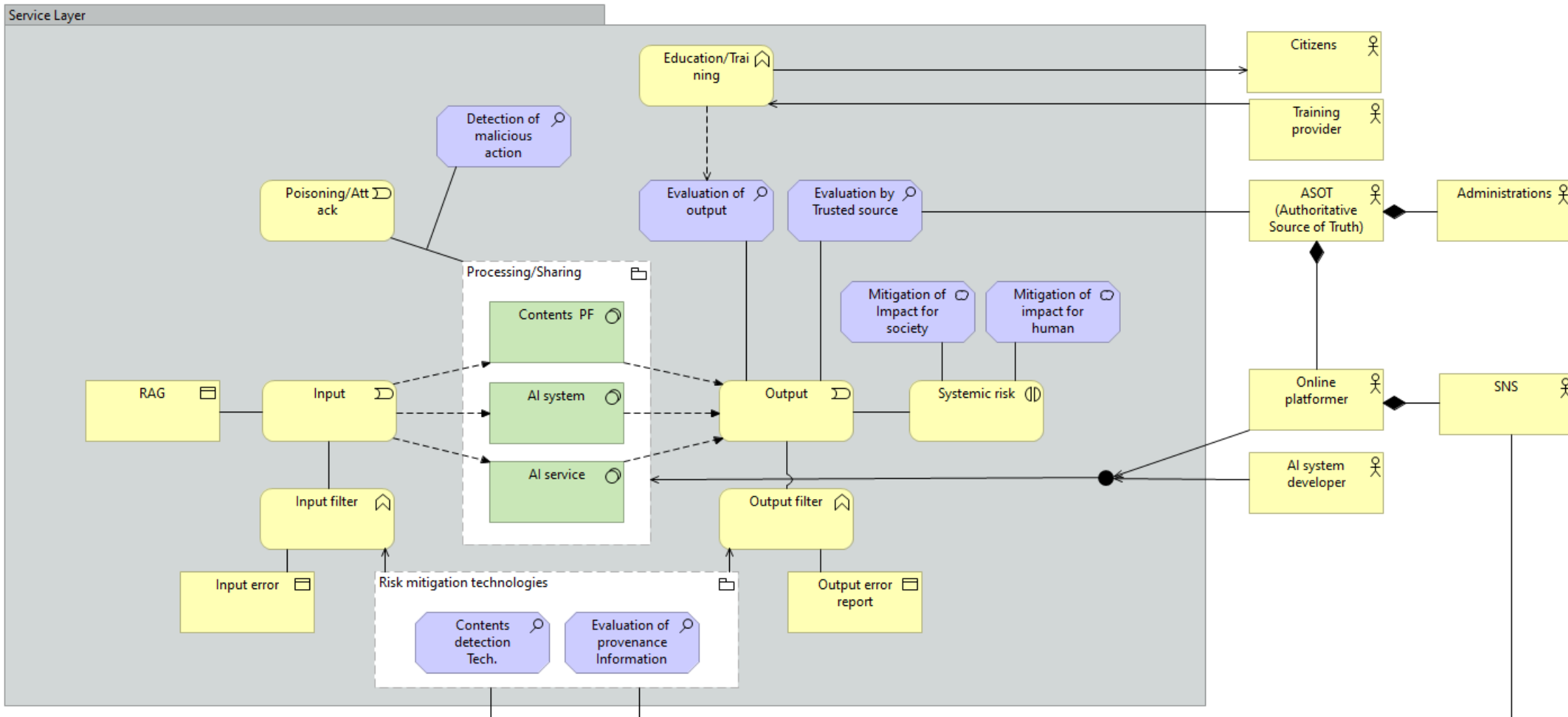
⑥ Contents and data management

- ♦ AI handles various data, and its stakeholders are diverse.
- ♦ This diagram illustrates the multilayer structure of the AI and data ecosystem in relation to data quality, transparency, and risk management.
- ♦ This diagram is meant for overview understanding, showing how data quality, transparency, and governance relate across the ecosystem.

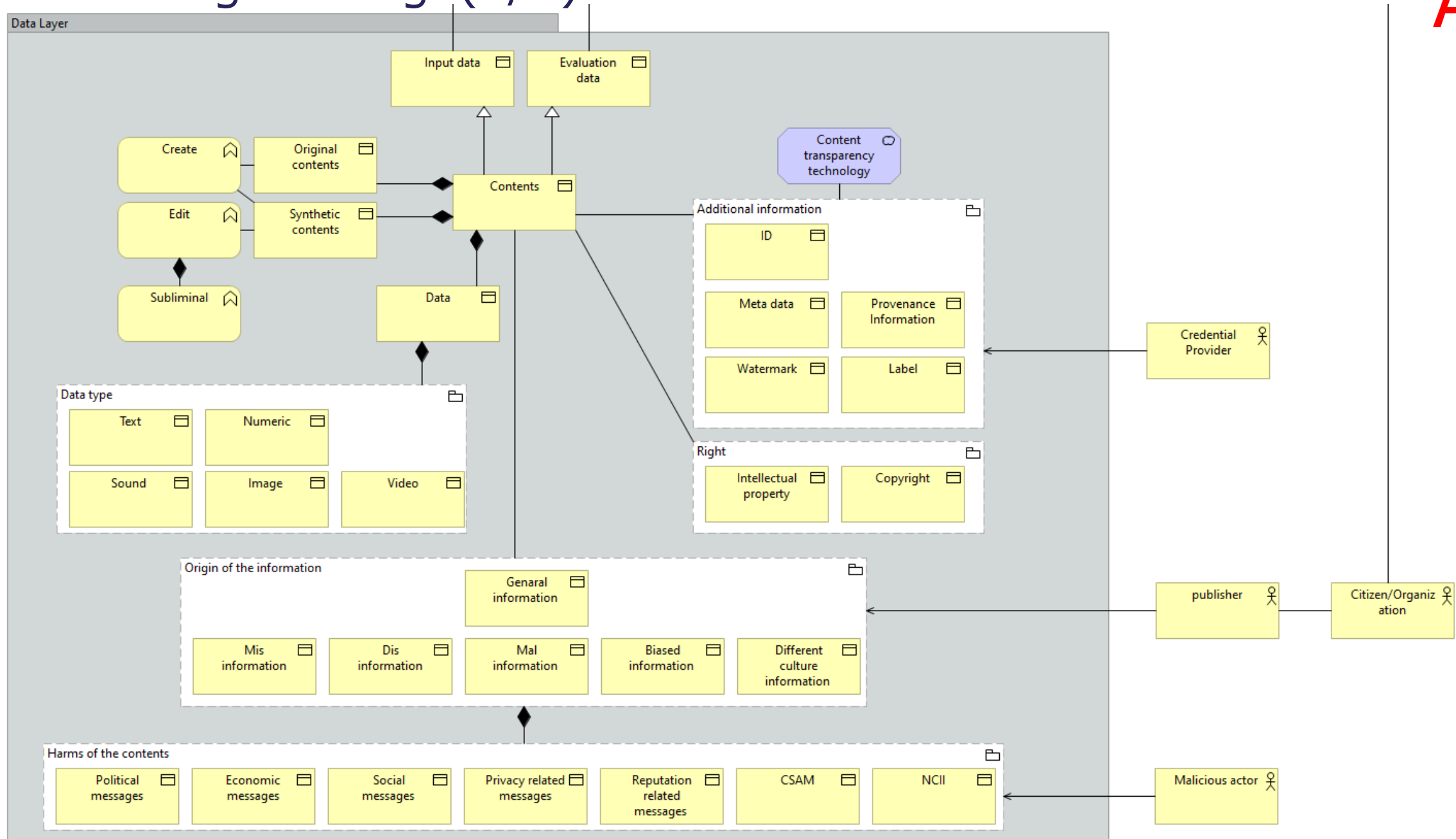
Enlarged image on the next pages



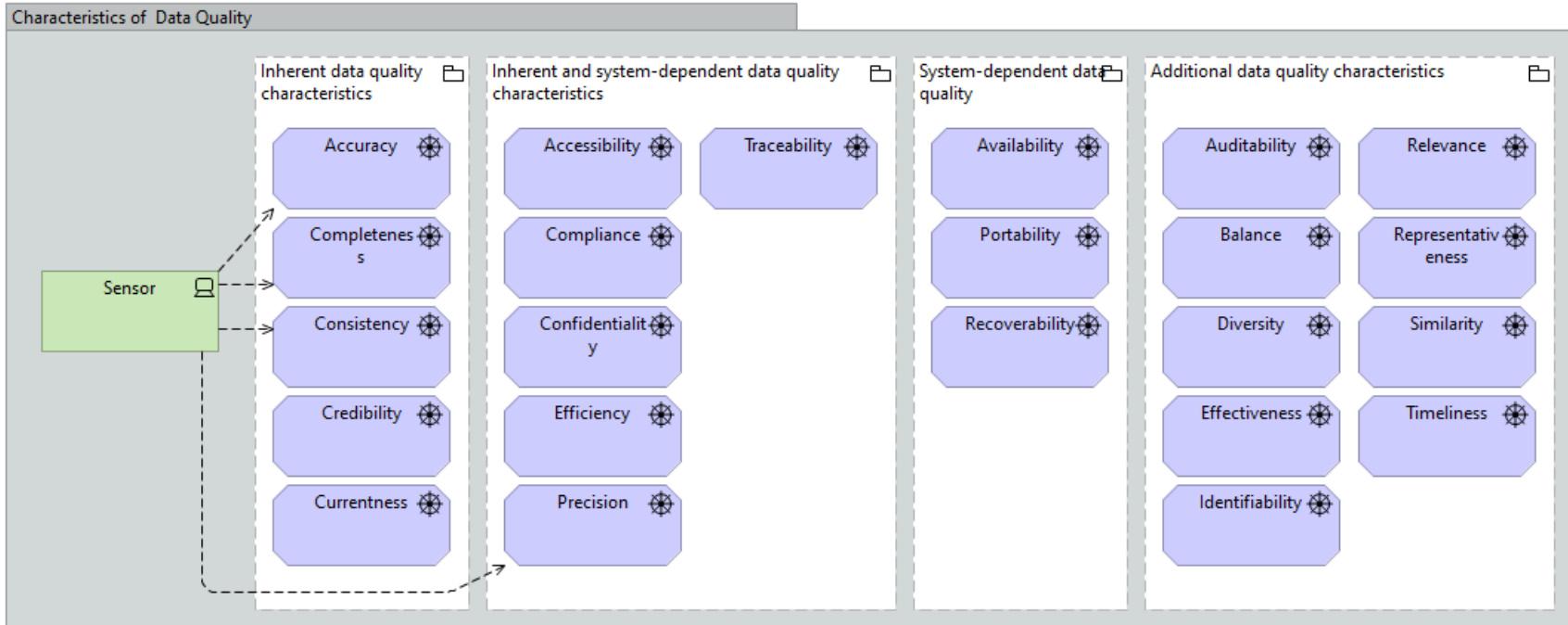
Enlarged image(1/3)



Enlarged image(2/3)



Enlarged image(3/3)



⑦ Data Quality Management for AI

- ◆ When managing data quality, accuracy, completeness, and timeliness are important. However, when data is used for AI, additional considerations are required, including the following:
 1. Annotation (Labeling) Quality Management
 - For machine learning, manual or automated labeling of data is essential. If there are many incorrect or inconsistent labels, it can lead to reduced model accuracy and skewed training results.
 2. Checking for Bias and Ensuring Fairness
 - It is essential to verify that the data does not disproportionately represent certain attributes (e.g., gender, race, region) and to prevent such biases from being amplified during the training process.
 3. Privacy and Security Measures
 - When handling data that includes personal or confidential information, proper anonymization and masking must be performed, and secure storage and processing must be ensured. Additionally, privacy protection technologies (e.g., differential privacy) should be considered to prevent personal information from being inferred using trained models.

⑦ Data Quality Management for AI(Cont.)

4. Version Control and Drift Management

- AI models are trained based on the data distribution at the time of training. Over time, the data distribution —or the underlying relationships between input and output— may change (“data drift” or “concept drift”), which can cause a sudden drop in model performance. It is important to implement data version control, monitor distributions, retrain or refine models as needed.

5. Handling Synthetic Data and Generated Content

- In cases where real-world data is insufficient, synthetic data may be used. However, clarity regarding its generation and quality is critical. Improper use of synthetic data can lead to flawed training and incorrect model outcomes.

6. Ensuring Explainability and Transparency

- AI models should not operate as “black boxes.” Explainable AI (XAI) methods helps clarify the decision-making process behind critical outcomes, thereby enhancing trust and accountability.

7. Establishing Operational and Monitoring Frameworks

- Continuous monitoring and maintenance are essential not only during data and model development but also throughout deployment. It is important to have systems in place that can respond quickly if unexpected bias or performance decline arises.

III. Perspective

III. Perspective

- Process View(Operation)

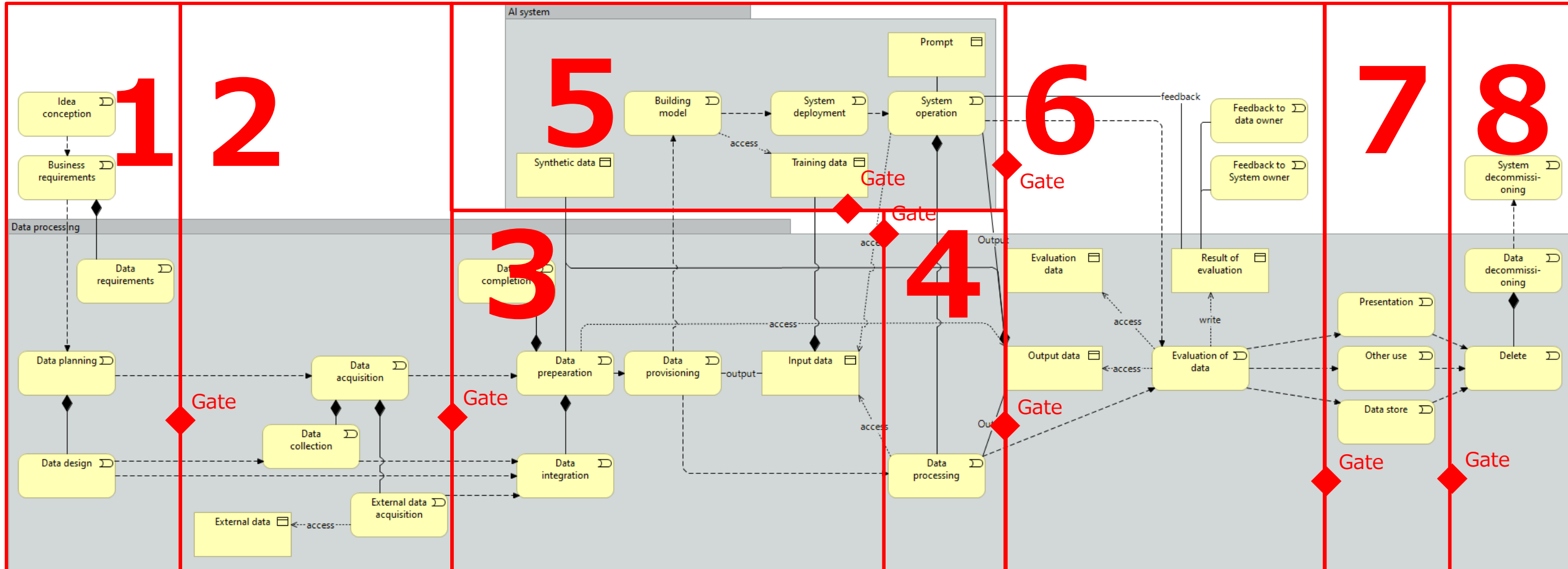
Process

1.Data planning
2.Data acquisition

3.Data preparation
4.Data processing
5.AI system

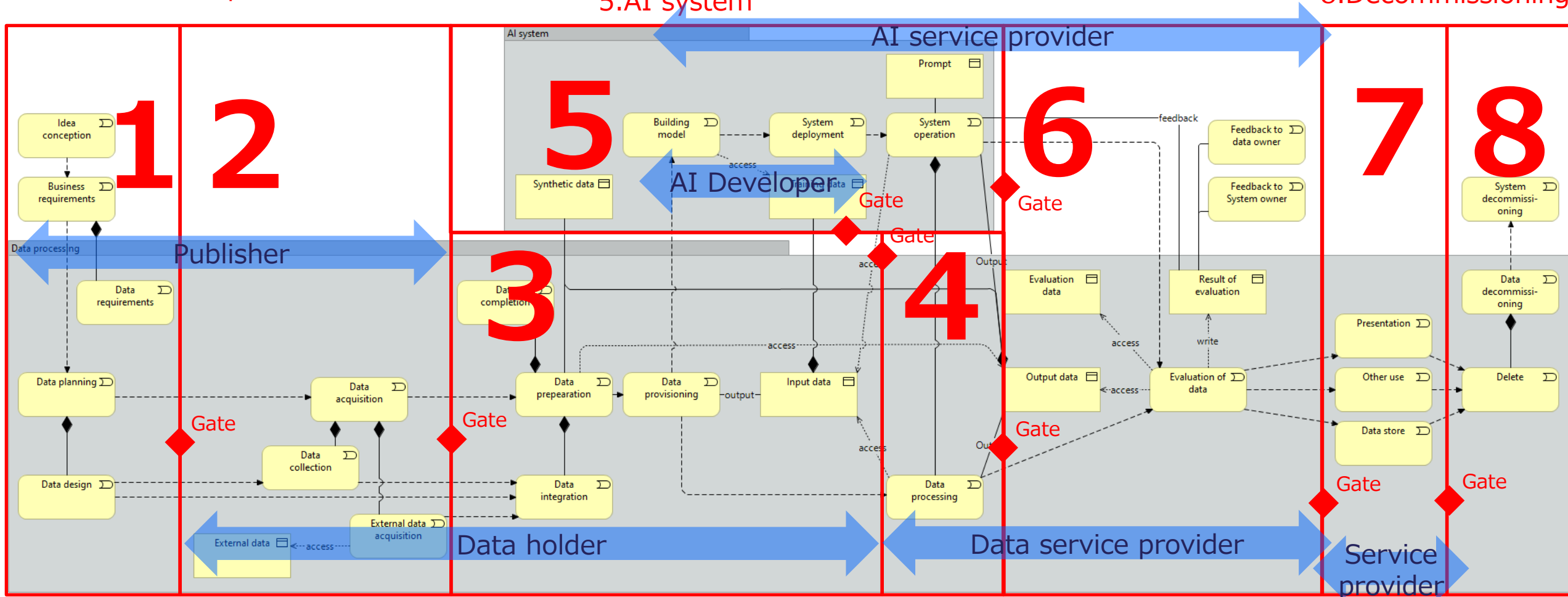
6.Evaluation of output

7.Deliver the result
8.Decommissioning



Role of stakeholders

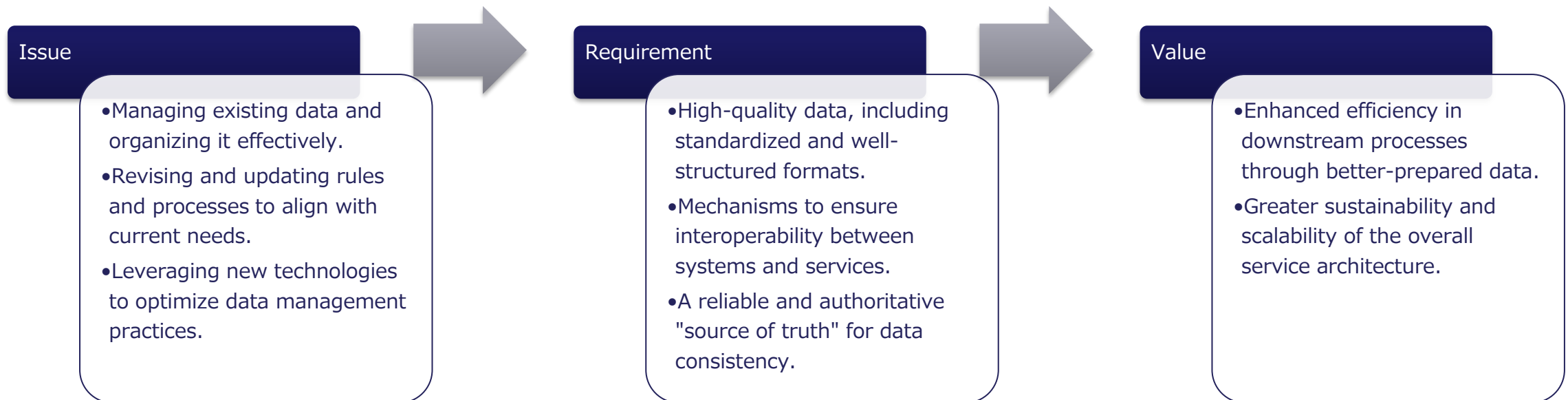
1. Data planning
2. Data acquisition
3. Data preparation
4. Data processing
5. AI system
6. Evaluation of output
7. Deliver the result
8. Decommissioning



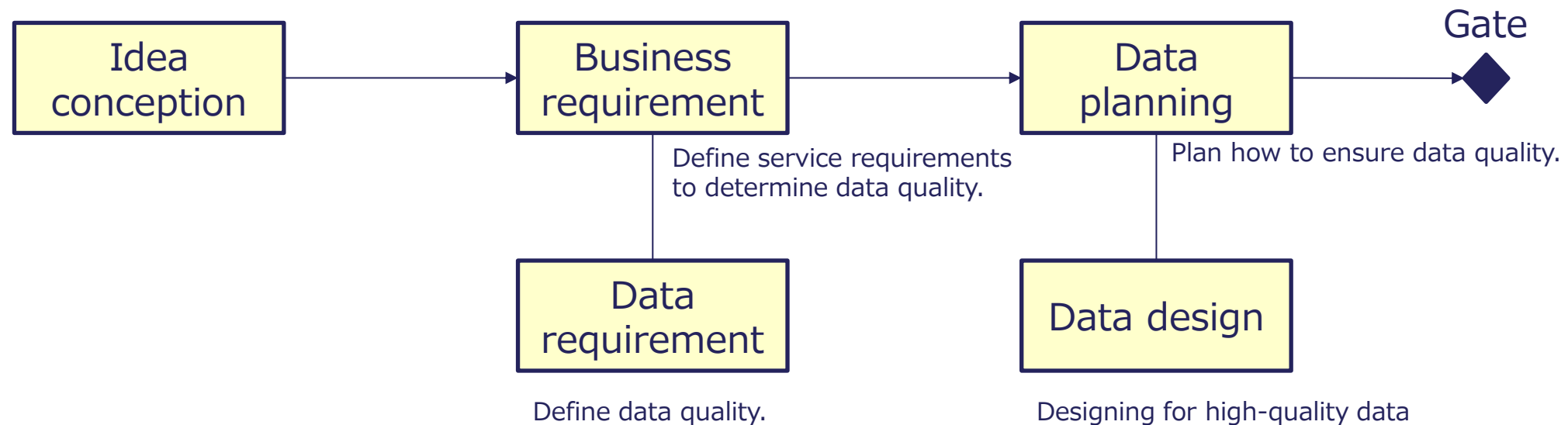
1. Data Planning

Description

- ♦ This is a critical phase in the data life cycle.
- ♦ Clearly define the intentions and motivations behind the need for the data.
- ♦ Establish interoperability across the entire service and ensure scalability for future growth.
- ♦ Specify the data life cycle, including acquisition methods, evaluation methods, and disposal processes.



- ◆ Quality is built in three steps, from concept creation to requirements definition and design.



Determine methods for data validation, change management, configuration management, and risk management throughout the data life cycle.

Procedure and checkpoint

Procedure

Idea conception

1. Gather the user needs
2. Create a concept
3. Define the business and data policy
 - Align with the organizational policies
 - Check the future trend
4. List the stakeholders
5. Check the feasibility

Business requirement

1. Define the business objective and scope
2. Define the value, requirement, constraint, and risk.

Checkpoint

- ❑ Do decision-makers understand the overall system concept?
- ❑ Do decision-makers understand the benefits of high-quality data?
- ❑ Do decision-makers understand the risks posed by low data quality?
- ❑ Do decision-makers understand the costs of requiring higher data quality than necessary and the importance of balancing effectiveness and cost?
- ❑ Do you have organizational capacity and skilled personnel?
- ❑ Is service quality clearly defined? (speed, cost, etc.)

Procedure and checkpoint

Procedure

Data requirement

1. Define the relevant data

- Input, Output, Reference
- Constraint, Interface, Statistical data

2. Specify the data specifications

- Description, Goal, Requirement, Matters to consider
- Refer to ISO/IEC5259-3 for data specification

3. Define the required quality level

- Accuracy, Completeness, Consistency, Credibility, Currentness
- Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability, Understandability
- Availability, Portability, Recoverability
- Auditability, Balance, Diversity, Effectiveness, Identifiability, Relevance, Representativeness, Similarity, Timeliness
- Refer ISO/IEC5259-2 characteristics

Checkpoint

- ▣ Are the data required by the business listed?
- ▣ Have quality control items been defined for each data?
- ▣ Are the data quality requirement levels for each data defined?

Procedure and checkpoint

Procedure

Data planning

1. Check the existing data.
2. Gather the data needs
3. Define the master data
4. Define the data architecture
5. Define the design policy and methodology
 - Structure, Location, Modeling, Documentation
6. Define and find the data source
7. Check whether any the data-related legislation applies.

Checkpoint

- ❑ Do decision-makers agree to convert existing data into the new model?
- ❑ Are the data needs of stakeholders for their operations and decision-making understood?
- ❑ Are the data architecture and design policies documented?
- ❑ Are the required data defined and available?
- ❑ Have legal constraints on the use of the data been checked?
- ❑ Does it contain any privacy and/or sensitive data?
- ❑ Does it contain any intellectual property data?

Procedure and checkpoint

Procedure

Data design

1. Check the reference models and taxonomy
2. Design the data models and the taxonomy
3. Design the metadata and labels
4. Design the rules
5. Check the compliance for legislation

Note:

In Japan, GIF (Government Interoperability Framework) provides reference models.

Checkpoint

- ❑ Do you refer to a data reference model or standardized taxonomy?
- ❑ Do you use modeling tools?
- ❑ Is the metadata designed based on a DCAT basis?
- ❑ Do you refer to general rules for utilization and access?
- ❑ Are they designed to comply with legislation?

Column: Reference model

- A reference model is a conceptual framework that provides standardized structures, processes, and data frameworks for a specific industry or domain. It serves as a "template" for designing and implementing systems or processes, offering unified terminology, definitions, and methods. This enhances interoperability between organizations and systems, improving efficiency in development and operations. Due to its high level of abstraction, a reference model can be applied to a wide range of scenarios and customized to meet specific requirements.
- How a Reference Model Improves Data Quality
 - A reference model establishes a foundation for data consistency and standardization.
 - By adopting common data definitions and naming conventions, it prevents inconsistencies across different systems.
 - Additionally, using a reference model eliminates data duplication and redundancy, enabling the creation of an efficient and streamlined data structure.
 - By embedding governance rules and constraints aligned with business processes, it enhances the accuracy and integrity of the data.
- Reference models play a critical role in ensuring data completeness during the design phase and reducing errors and costs of correction during the operational phase.

Column: Data dictionary

- A Data Dictionary is a repository that systematically organizes and manages the names, definitions, formats, and relationships of data elements used in information systems and databases. It defines attribute names, data types, lengths, constraints, and business rules for each data element, providing consistent standards across the organization.
- This shared understanding of data meaning and structure among stakeholders—such as developers and operators—helps streamline system development and operations, improving efficiency and quality.
- Moreover, during data mapping and system integration, it enhances the accuracy of data migration and reporting between different systems, making it easier to ensure consistency.
- In terms of data governance, a Data Dictionary serves as the foundation for security requirements, access rights, and regulatory compliance.
- By keeping it up to date, organizations can flexibly accommodate new data requirements and swiftly adapt to changing business needs, ensuring a robust and responsive system environment.

Column: Data modeling

- Data modeling is the process of visually and logically organizing the data used in information systems and databases, representing the structure and relationships of the data in the form of a model. It involves defining entities (data objects), attributes (data elements), and relationships (connections between data). Data modeling typically includes three levels: conceptual data models (representing data from a business perspective), logical data models (defining detailed structures from a technical perspective), and physical data models (defining the database structure concretely).
- Data Modeling Contributes to Improved Data Quality.
 1. Ensures Consistency and Standardization
 2. Eliminates Duplication and Redundancy
 3. Ensures Data Integrity and Consistency
 4. Enhances Data Reusability
 5. Improves Data Driven Decision Making
- Data modeling is not just a technical process; it is a critical activity that underpins data quality management and data governance.

* Apart from data modeling, there is AI modeling; AI modeling refers to AI training data.

Column: DCAT

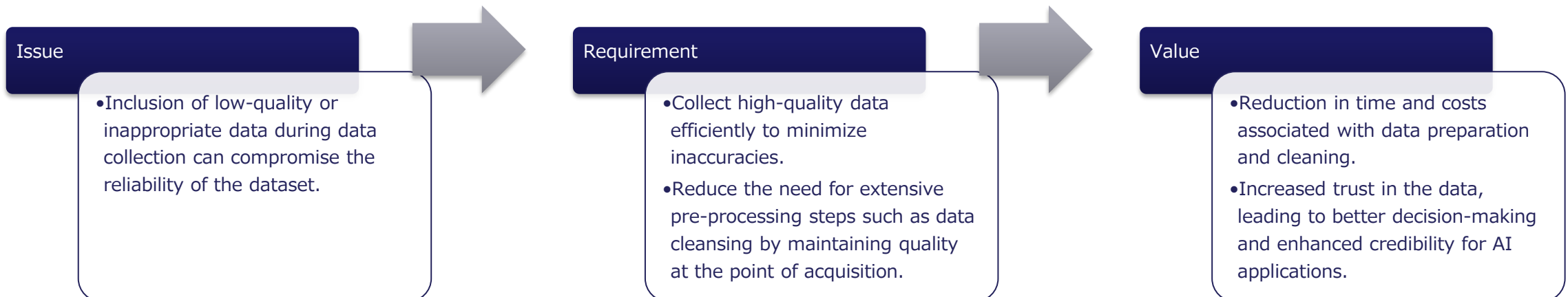
- DCAT (Data Catalog Vocabulary) is a W3C standard vocabulary designed to facilitate interoperability between data catalogs published on the internet.
- It enables publishers to describe datasets and data catalogs in a standardized way, making them easier to find, share, and reuse across platforms.
- DCAT supports metadata such as titles, descriptions, URLs, and licensing.
- By promoting consistent metadata, DCAT improves the discoverability of datasets and supports data integration and open data initiatives.
- The current version, DCAT v3, extends support for data services, versions, and provenance, aligning with FAIR data principles.
 - FAIR: Findable, Accessible, Interoperable, Re-usable

<https://www.w3.org/TR/vocab-dcat-3/>

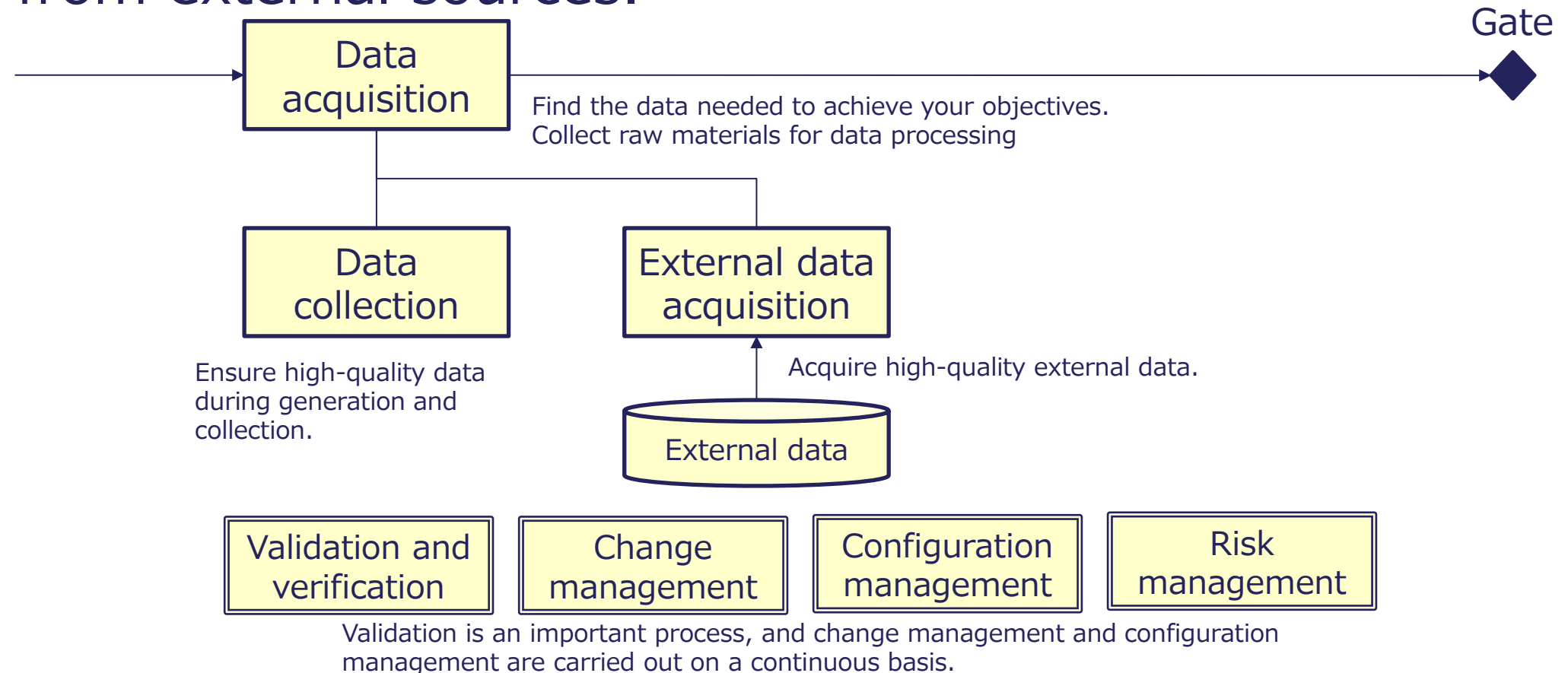
2. Data Acquisition

Description

- ♦ This step involves generating, collecting, and sourcing data from external systems or entities.
- ♦ Prevent errors during data acquisition, whether from sensors or manual data input.
- ♦ Ensure that low-quality data is excluded when obtaining data from external sources.
- ♦ When supplementing data with synthetic data, clearly document the process and prevent the inclusion of inappropriate synthetic content.
- ♦ Add metadata, such as data profiles and provenance information, to ensure traceability and quality of data.



- ♦ Generate and collect the data required to achieve the objective, and if the data is not available within the organization, acquire it from external sources.



Procedure and checkpoint

Procedure

Data acquisition

1. Identify and locate necessary data
2. Check the provenance information
3. Check the condition for use

Data collection

1. Check the device (Sensor)
2. Collect or input the data
 - Prevent errors by using the web forms and APIs
3. Verify the data
 - Removal out-of-range and inappropriate data
 - Correct inconsistent data
4. Anonymize and mask
5. Create metadata
 - General information
 - Quality information
 - Method of measuring or gathering data

Checkpoint

- ❑ Is data obtained from reliable sources?
- ❑ Are there any problems with the data's provenance information?
- ❑ Are there any restrictions on how the data can be used?

- ❑ Are you taking steps to prevent inappropriate data?
- ❑ Do you ensure that out-of-range data is not entered?
- ❑ In the case of sensor data, is it calibrated?
- ❑ Do you check that the data is consistent?
- ❑ Do you refer to DCAT for metadata?
- ❑ Do you describe the method of measurement and collection of data?

Procedure and checkpoint

Procedure

External data acquisition

1. Verify the metadata and provenance information
2. Verify the data
 - Checking quality characteristics
 - Finding inappropriate content (including synthetic data)
3. Add metadata and provenance information

Checkpoint

- ❑ Is the source of the data reliable?
- ❑ Is the data clear about provenance information?
- ❑ Does the quality of the data meet the specifications?
- ❑ Does the data meet the requirements for "Portable Data," allowing for effective exchange and transfer with the target system without loss of content or meaning?
- ❑ Does it contain data inappropriate to the system's objectives?
- ❑ Does the data contain important information other than the data items to be imported (such as PII)?
- ❑ Does it contain synthetic data that is not clearly identified as synthetic data?

Column: Improvement of input method

- By linking data via an interface between systems, we can prevent input errors by users.



Accurate and quick
No delivery errors

- When a person is entering data, using an input form allows the user to correct or adjust the data as they enter it.



- Checking the data formats
- Checking the accuracy of the data
- Preventing input variations
- Using controlled vocabulary

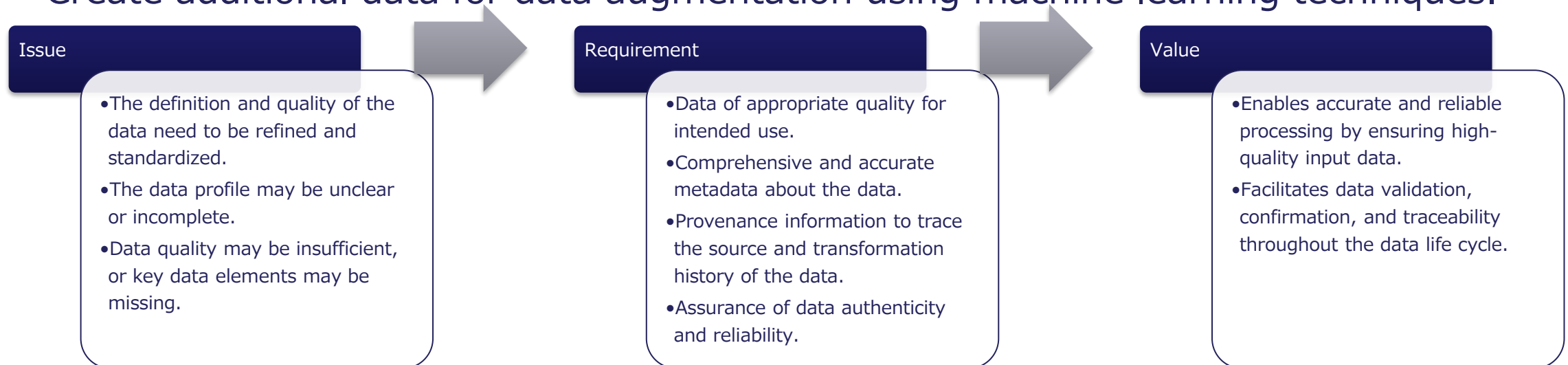
Column: Provenance data

- Provenance data refers to information that records the history and origin of data, including its creation, transformation, movement, and usage. It tracks who created, modified, or used the data, as well as when and how these actions occurred. Provenance data serves as metadata to enhance transparency in data processing, making it particularly valuable in fields such as healthcare, finance, and scientific research where data reliability and integrity are critical. This information typically includes the data's source, processing history, applied business rules or algorithms, and associated systems or users.
- How Provenance Data Improves Data Quality
 - Provenance data ensures data reliability by making its history transparent.
 - By clarifying the origins and life cycle of the data, it becomes easier to validate its accuracy and appropriateness, and to identify errors or tampering.
 - It also provides insights into data creation and updates, enabling the correction of inconsistencies or incomplete records.
 - Additionally, in audits and compliance processes, provenance data serves as evidence that the data has been handled appropriately, boosting confidence in its reliability.
- This enhances the accuracy and effectiveness of data analysis and decision-making, while also mitigating potential risks and strengthening data governance.

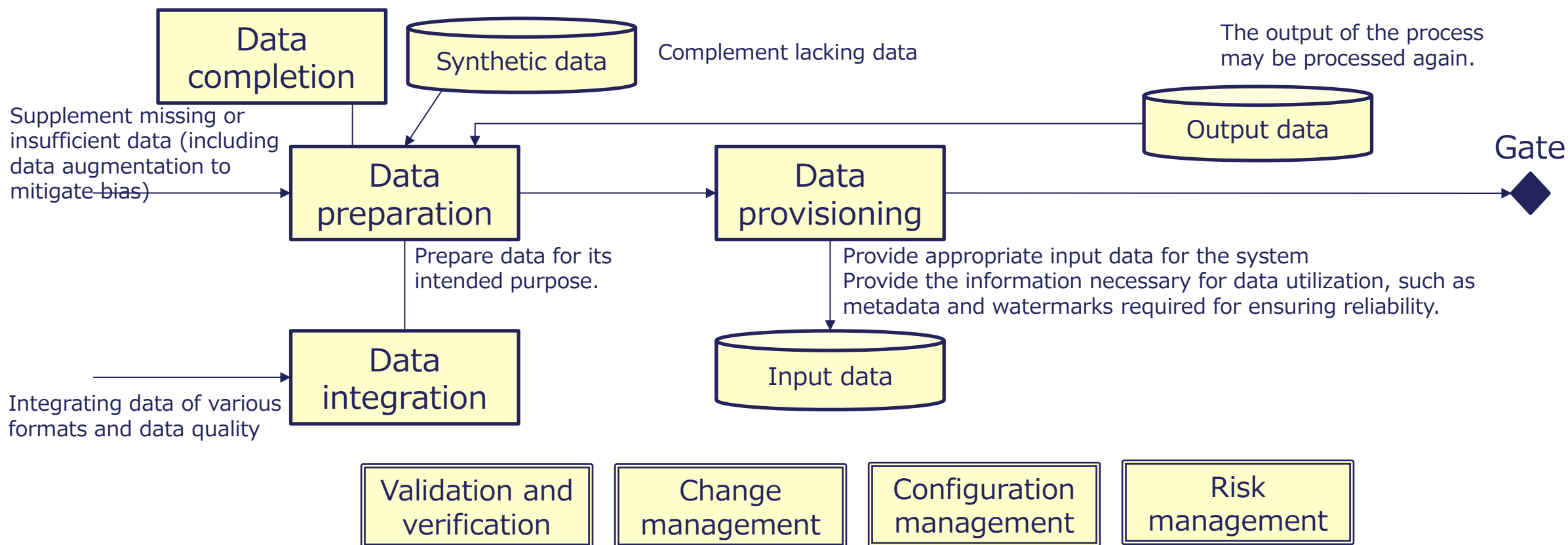
3. Data Preparation

Description

- ♦ It is the final stage of providing quality data.
- ♦ Clean the prepared data: Remove any erroneous or out-of-range data and adjust taxonomy conversions, semantics, and granularity.
- ♦ Supplement lacking data as necessary.
- ♦ Integrate the cleaned data into a unified dataset.
- ♦ Add watermarks to the data, when necessary, to ensure data authenticity and integrity.
- ♦ Create additional data for data augmentation using machine learning techniques.



- ◆ Various datasets are integrated to create input data, which is used for training AI models or further processing.



Change management and configuration management become important. In particular, sensor data is real-time data, so consideration is needed for its management.

Procedure and checkpoint

Procedure

Data preparation

1. Document the collected data
2. Define on the data integration policy
3. Define the data supplementation policy
4. Clean the data
5. Add labels

Checkpoint

- ❑ Is all necessary data documented?
- ❑ Is a data integration policy in place?
- ❑ Is there a defined policy for data supplementation?
- ❑ Is the division of data done in a logical way for AI training and validation?

Procedure and checkpoint

Procedure

Data integration

1. Define the post-integration data model
2. Create conversion tables for taxonomies and controlled vocabularies
3. Verify the semantics of data items and determine the matching method
4. Harmonize accuracy and units
5. Convert file formats (e.g. XML, JSON, CSV)
6. Document semantic differences
7. Create integration-related metadata

Checkpoint

- ❑ Has the conversion table been created taking into account the meaning of the data items?
- ❑ Are there priority rules for duplicate data?
- ❑ Is blank data treated as no data?
- ❑ Is data conversion carried out automatically by tools?
- ❑ Is the consistency of integrated data checked?
- ❑ Is metadata created by DCAT?

Procedure and checkpoint

Procedure

Data completion

1. Analyze the state of the data and determine whether it needs to be supplemented.
2. Add complementary data

Data augmentation (for AI system)

1. Prepare the base data
2. Generate training data based on the base data

Synthetic data acquisition

1. Check the metadata and provenance information
2. Validate the data

Checkpoint

- ❑ Are you using appropriate data to supplement the data?
- ❑ Is there any bias in the base data?
- ❑ Is there a bias caused by reliance on a small number of base data?
- ❑ Is it clearly indicated that the content is synthetic?
- ❑ Does it contain inappropriate or non-consensual content?

Procedure and checkpoint

Procedure

Data provisioning

1. Register catalog information
2. Provide data interface
3. Control the version
4. Provide data samples
5. Add content protection information, such as watermarks
6. Operate the access control function
7. Track the usage, if necessary

Checkpoint

- ❑ Is the data provided in a machine-readable interface?
- ❑ Do you include mechanisms such as watermarking to make it difficult for data to be misused?
- ❑ Are access controls in place to prevent unauthorized use?

Column: IDs

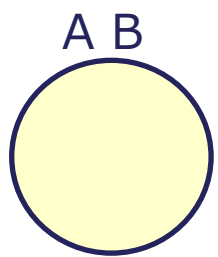
- Assigning unique identifiers (IDs) to data is critical for improving data quality. By tagging each record with an ID, you make it possible to integrate information from multiple sources and ensure consistent data management across systems.
- This prevents confusion or duplication when similar or related pieces of data need to be merged or compared. For example, if two departments track customers in separate databases, using the same customer ID allows the two datasets to be combined accurately.
- In addition to aiding data integration and consistency, assigning IDs provides several other benefits:
 - 1. Traceability and Auditability:** A unique ID makes it easier to trace the origin and history of a data record, which helps in audits and compliance checks.
 - 2. Efficient Updates:** When you need to update or delete specific data, having an ID allows quick and precise identification of the target records.
 - 3. Scalability:** As data grows, unique IDs ensure that new records can be added without conflicts or confusion, helping maintain clarity in large-scale databases.
 - 4. Enhanced Data Governance:** IDs support clear ownership and governance policies, as you can assign responsibility for specific records to the appropriate teams or systems.
- Overall, using IDs is a fundamental practice that not only enables reliable data integration and consistency but also improves the overall manageability and value of your data.

Column: Data cleansing

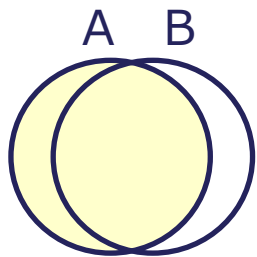
- Data cleansing is the process of identifying, correcting, or removing errors, inconsistencies, and inaccuracies in datasets to improve data quality and reliability. It ensures that data is complete, accurate, and consistent, making it more suitable for analysis, reporting, and decision-making. Data cleansing involves addressing issues such as missing values, duplicate records, incorrect formatting, and outliers. This process is crucial for maintaining the integrity of data, especially when integrating multiple datasets or preparing data for advanced analytics.
- Data Cleansing Techniques
 - Removing Duplicates: Identifying and eliminating duplicate records to avoid redundant or misleading data.
 - Handling Missing Data: Filling in missing values using techniques like interpolation, imputation, or deletion where necessary.
 - Standardizing Data: Ensuring data follows consistent formats, such as standardizing date formats or capitalization.
 - Validating Data: Checking data against predefined rules or reference datasets to ensure accuracy and consistency.
 - Correcting Errors: Identifying and rectifying typos, incorrect entries, or mismatched values.
- Effective data cleansing ensures high-quality data, which leads to better insights and more accurate decision-making.

Column: Data matching

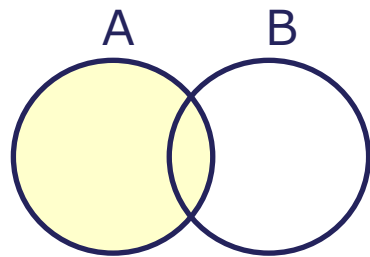
- Data matching is the process of comparing and identifying records from different datasets -or within the same dataset- to determine whether they represent the same entity. It is widely used in scenarios such as deduplication, linking customer information across systems, fraud detection, and data integration. Data matching involves evaluating attributes such as names, addresses, phone numbers, or other identifiers to detect matches, even when there are slight variations or inconsistencies in the data. The level of matching is sometimes expressed as a score.
- By consolidating duplicate or related records, data matching improves data quality and ensures accurate analysis and reporting.



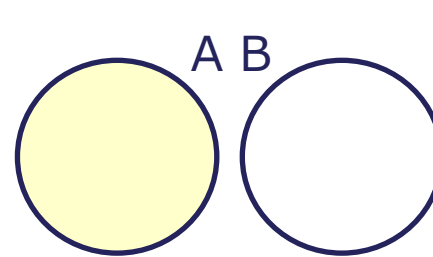
Exact match



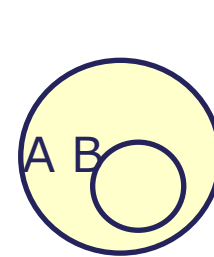
Close match



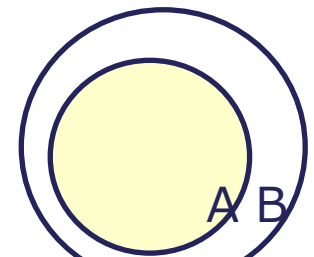
Related match



No match



Narrow match



Broad match

Column: Data completion

- **Data augmentation** is a technique used to enhance the size and diversity of a dataset by applying various transformations to existing data. These transformations can include, for example, rotating, flipping, cropping, adjusting brightness and contrast, or adding noise for image data, paraphrasing, or back-translation for text data, or other transformations for different types of data. The goal is to increase the variety of training data, which helps improve the performance and robustness of machine learning models.
- **Data enrichment** is the process of enhancing existing data by supplementing it with additional information from external or internal sources. This process adds value to raw data, making it more comprehensive, accurate, and useful for analysis or decision-making. Enrichment often includes adding missing data points, improving data accuracy, or integrating contextual information such as demographic, geographic, or behavioral data.

Column: Synthetic data – Positive aspects

The use of synthetic data generated by AI offers significant benefits and opportunities.

1. Enhancing Data Privacy

Synthetic data enables analysis and model training without exposing real personal information, helping to protect individual privacy.

2. Expanding Data Availability and Safe Testing

Synthetic data can fill gaps where real data are scarce, sensitive, or difficult to collect, and enables safe testing and validation of AI systems under controlled conditions.

3. Accelerating AI Development

By generating diverse and high-quality training data, synthetic data helps improve AI performance, robustness, and fairness.

4. Reducing Costs and Risks

Using synthetic data lowers the cost of data collection and annotation, while minimizing regulatory and security risks associated with handling real data.

5. Promoting Innovation and Collaboration

Openly shared synthetic datasets can encourage cross-industry collaboration and faster research progress while maintaining confidentiality.

Column: Synthetic data – Negative aspects

The synthetic content generated by AI has a significant impact

1. Spread of Fake News and Misinformation

Synthetic content can generate highly realistic images, audio, and text, which may be intentionally misused to spread false information.

2. Invasion of Privacy

Technologies like deepfakes can create content that mimics an individual's face or voice, potentially violating his or her privacy.

3. Damage to Brand or Corporate Reputation

Maliciously created synthetic content can tarnish a company or brand reputation.

4. Legal Risks and Ethical Challenges

The creation and distribution of synthetic content can lead to legal and ethical issues such as copyright infringement, violation of publicity rights, or discriminatory content.

5. Decline in Trust

An abundance of synthetic content makes it harder for consumers and businesses to trust digital content.

6. National Security and Political Risks

Synthetic content might be politically exploited.

Column: Content Transparency Tech

- Content Transparency Technologies refer to tools, frameworks, or systems designed to provide clarity, authenticity, and accountability in digital content. These technologies aim to ensure that users can understand the origin, context, and reliability of the information they encounter. They are often used in media, advertising, social platforms, and data-driven industries to address issues such as misinformation, biased content, or non-transparent algorithms.
- Key features of Content Transparency Technologies include:
 - Source Verification: Ensuring content authenticity by verifying the origin or creator of the material.
 - Traceability: Providing a clear history of edits, ownership, and dissemination of the content.
 - Content Labeling: Adding metadata or markers to indicate whether content is sponsored, user-generated, or machine-generated.

Column: Indication of data quality

The level of data quality indicators should naturally differ between internal management and public use.

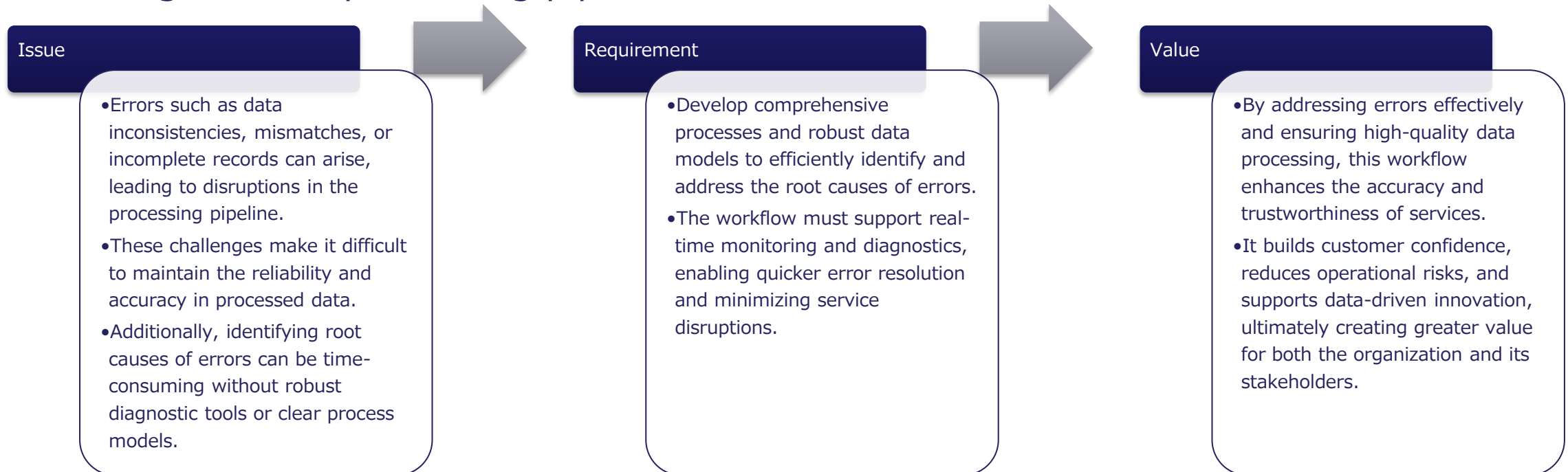
- Data quality is often managed in detail within organizations by data producers and users. However, when data circulates more broadly in society, **simpler and easier-to-understand indicators** become necessary.
- This is similar to **how home appliances are handled**: detailed specifications and quality checks are used in manufacturing and logistics, but consumers only see simple indicators such as efficiency labels or safety marks at stores.
- Likewise, in data quality management, it is important to design different levels of indicators —detailed ones for internal management and simple ones for external sharing or trading —according to the purpose and stakeholders involved.



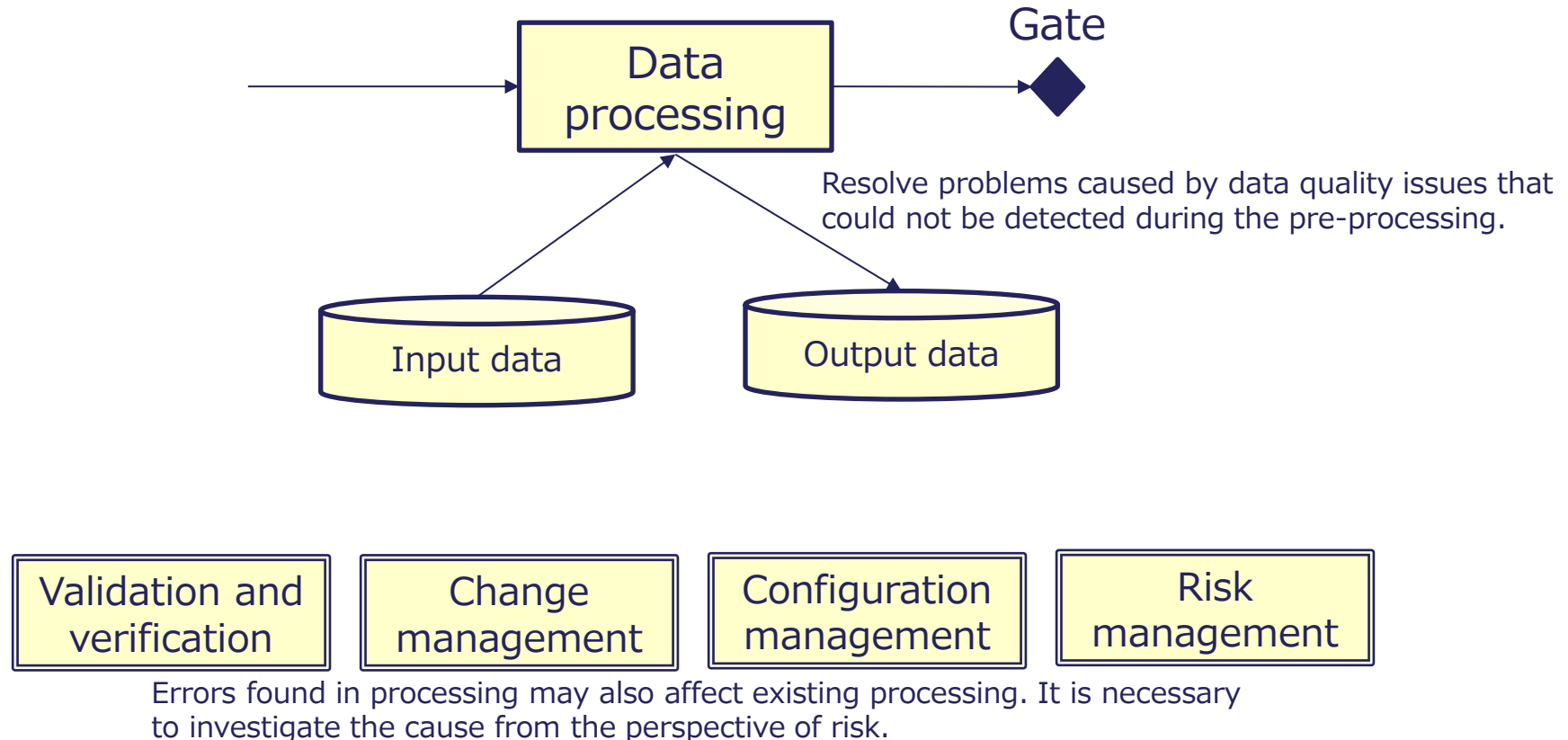
4. Data Processing

Description

- ♦ Create service value by processing data effectively and efficiently.
- ♦ In the event of an error, identify its root cause within the data path and provide actionable feedback to ensure prompt resolution.
- ♦ The workflow emphasizes maintaining data integrity and seamless operation throughout the processing pipeline.



- ◆ Process the data that has been entered. As there is a possibility of data inconsistencies occurring during the processing, a feedback mechanism is required.



Procedure and checkpoint

Procedure

Data Processing

1. Check consistency
2. Process data
3. Report errors in processing

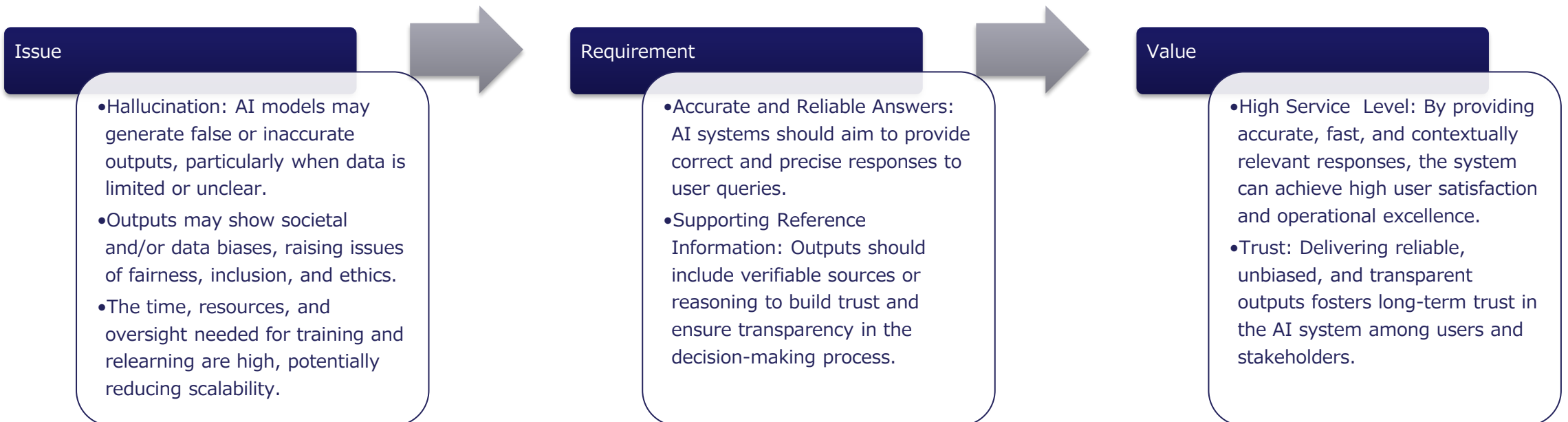
Checkpoint

- ❑ Is an error notification provided when there is an anomaly in the data?
- ❑ Is the data processing visualized for verification?

5. AI System

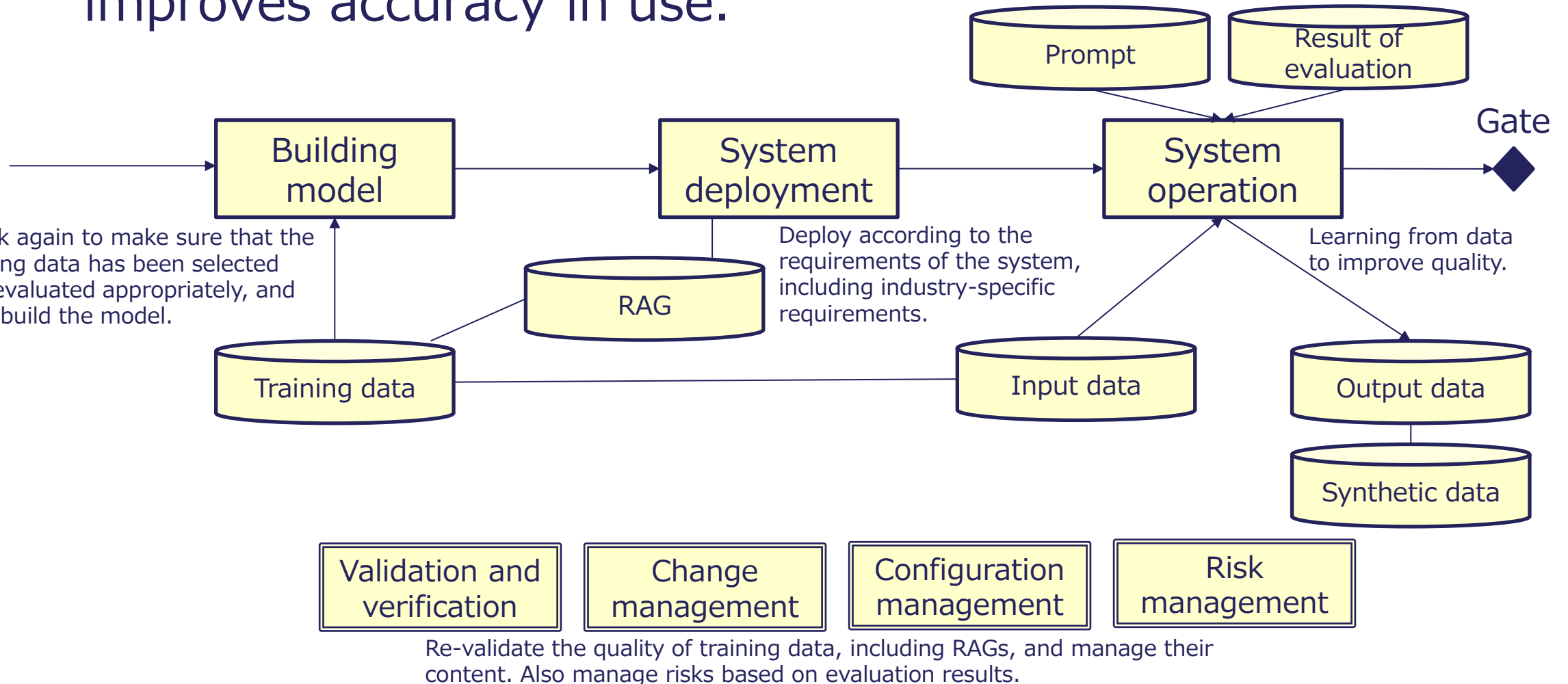
Description

- ♦ Train AI models using curated datasets.
- ♦ Implement, utilize, and operate these models.
- ♦ Continuously retrain and refine models based on output evaluations.
- ♦ For example, Retrieval-Augmented Generation (RAG) techniques can be used to enhance the accuracy and reliability by referencing external knowledge sources.



5. AI System Actions

- ◆ Input of high-quality data that is fit for purpose and further improves accuracy in use.



Procedure and checkpoint

Procedure

Building model

1. Check the required level for AI
2. Decide on the training data
 - Prevent biased data and inappropriate data
3. Decide whether to use RAGs*
 - *RAG(Retrieval-Augmented Generation)
4. Check whether there is any personal or intellectual property information
5. Generate the model
6. Test the model

System deployment

1. Deploy the system
2. User training

System operation

1. Operate the system
2. Continuous monitoring

Checkpoint

- ❑ Do you use reliable data for your training data?
- ❑ Have you implemented RAGs to improve accuracy?
- ❑ If there is information that could lead to personal data or intellectual property, do you take measures to prevent this information from being displayed, for example by excluding the necessary information from the training data?
- ❑ Do you have a deployment policy and follow it?
- ❑ Does reusing the output introduce any bias into the AI system?

Column: Bias

Reducing bias is essential for implementing AI. By addressing bias, AI systems can foster broader societal acceptance, ensure ethical use, and maintain long-term reliability and sustainability.

1. Preventing Unfairness and Discrimination

AI trained on biased data may make decisions that disadvantage specific groups based on gender, age, race, or location, leading to ethical concerns.

2. Ensuring Trust and Fairness

Bias undermines trust in AI systems. To create trustworthy AI solutions, fairness and impartiality must be maintained.

3. Avoiding Business Risks

Bias-related controversies can harm a company's reputation and expose it to legal risks.

4. Minimizing Societal Impact

The widespread adoption of AI amplifies its societal influence. If biased, AI can exacerbate inequalities or deepen societal divides.

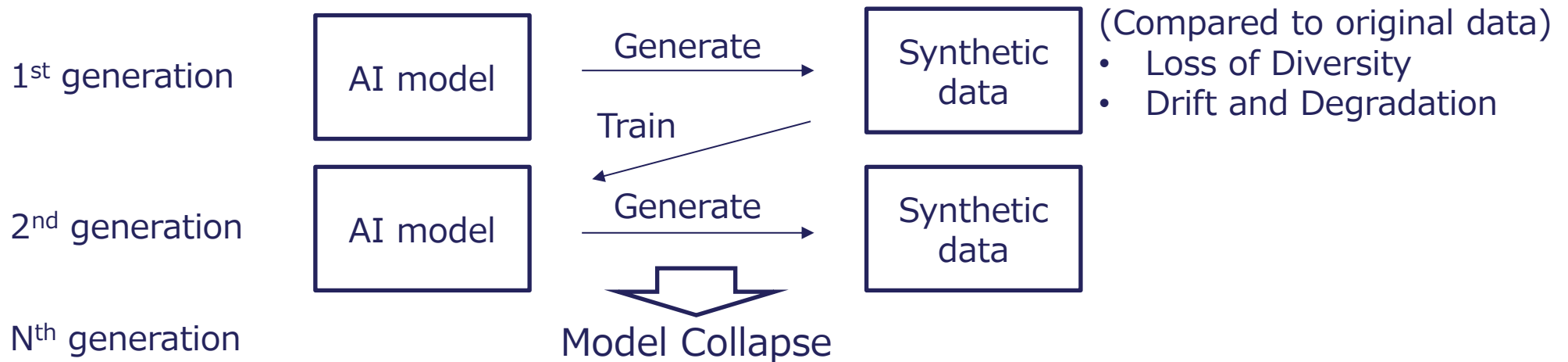
5. Compliance with Legal and Ethical Standards

Increasing regulations around AI emphasize fairness and bias mitigation. Non-compliance can lead to penalties or exclusion from specific markets.

Column: Model collapse

What is model collapse?

- A phenomenon where generative AI models trained repeatedly on synthetic data (rather than real human data) gradually lose diversity and accuracy over time.



Prevention and mitigation

- Ensure clear distinction and proper governance between real and synthetic data.
- Periodically retrain using verified real-world datasets.
- Implement continuous quality monitoring and feedback loops for dataset integrity.

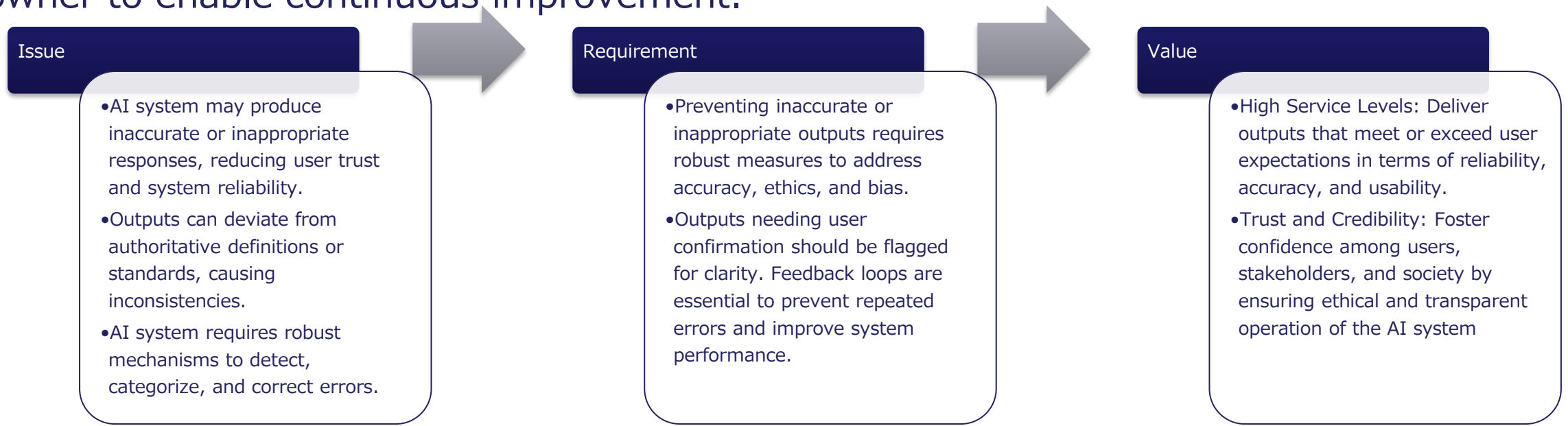
(ref) Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).

<https://doi.org/10.1038/s41586-024-07566-y>

6. Evaluation of output

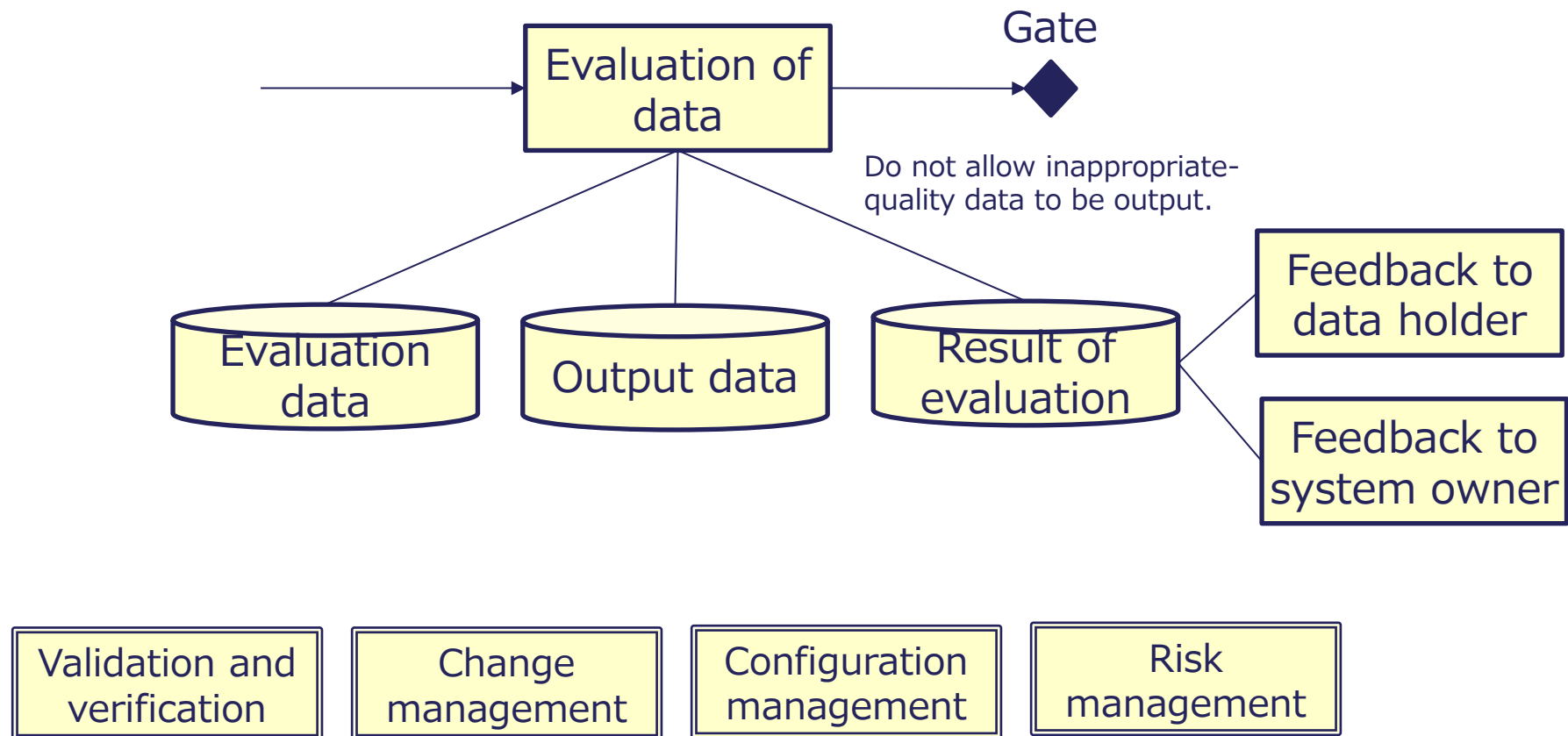
Description

- ◆ This phase serves as the final verification before the AI system's outputs are provided to end-users. It includes assessing the outputs for ethical considerations, information accuracy, potential biases, and alignment with intended purposes.
- ◆ Both automated systems and human reviewers evaluate outputs using reliable information sources and pre-established evaluation datasets.
- ◆ Errors identified during this process should be reported back to the data holder or system owner to enable continuous improvement.



6. Evaluation of Output Actions

- ♦ Evaluation of outputs is important as a safeguard for the system. Prevent inappropriate responses.



Risk management is important, and mechanisms are needed to assess data that cannot be clearly judged as inappropriate

Procedure and checkpoint

Procedure

Evaluation of output

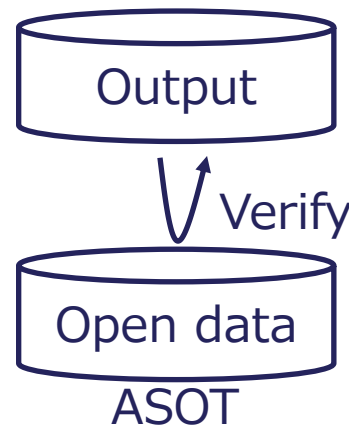
1. Prepare the evaluation data
2. Confirm using reliable data
3. Remove inappropriate responses based on the evaluation data
4. Create error reports
5. Provide feedback to data holders
6. Provide feedback to system owners

Checkpoint

- ▣ Are safeguards in place to prevent unethical responses?
- ▣ If a decision needs to be made on a response, does the process involve human intervention?
- ▣ Do you have mechanisms in place to investigate the causes of inappropriate responses and provide feedback?
- ▣ Are there any outputs that do not match the reliable data?

Column: ASOT

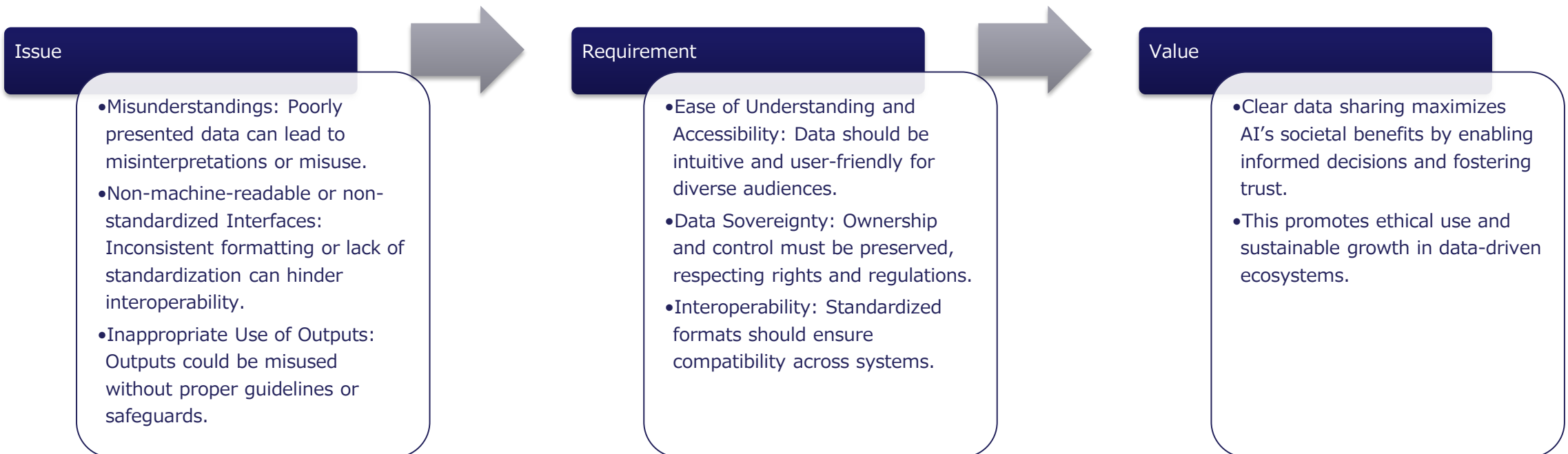
- The **Authoritative Source Of Truth** (ASOT) refers to the definitive, trusted source of information for a specific system, process, or context. It is the location where accurate, up-to-date, and complete data is maintained, ensuring consistency and reliability across systems or teams that rely on that information.
- By verifying data using ASOT, you can eliminate incorrect data and improve data quality. Open data provided by the government serves as essential social infrastructure for enhancing data quality.



7. Deliver the Result

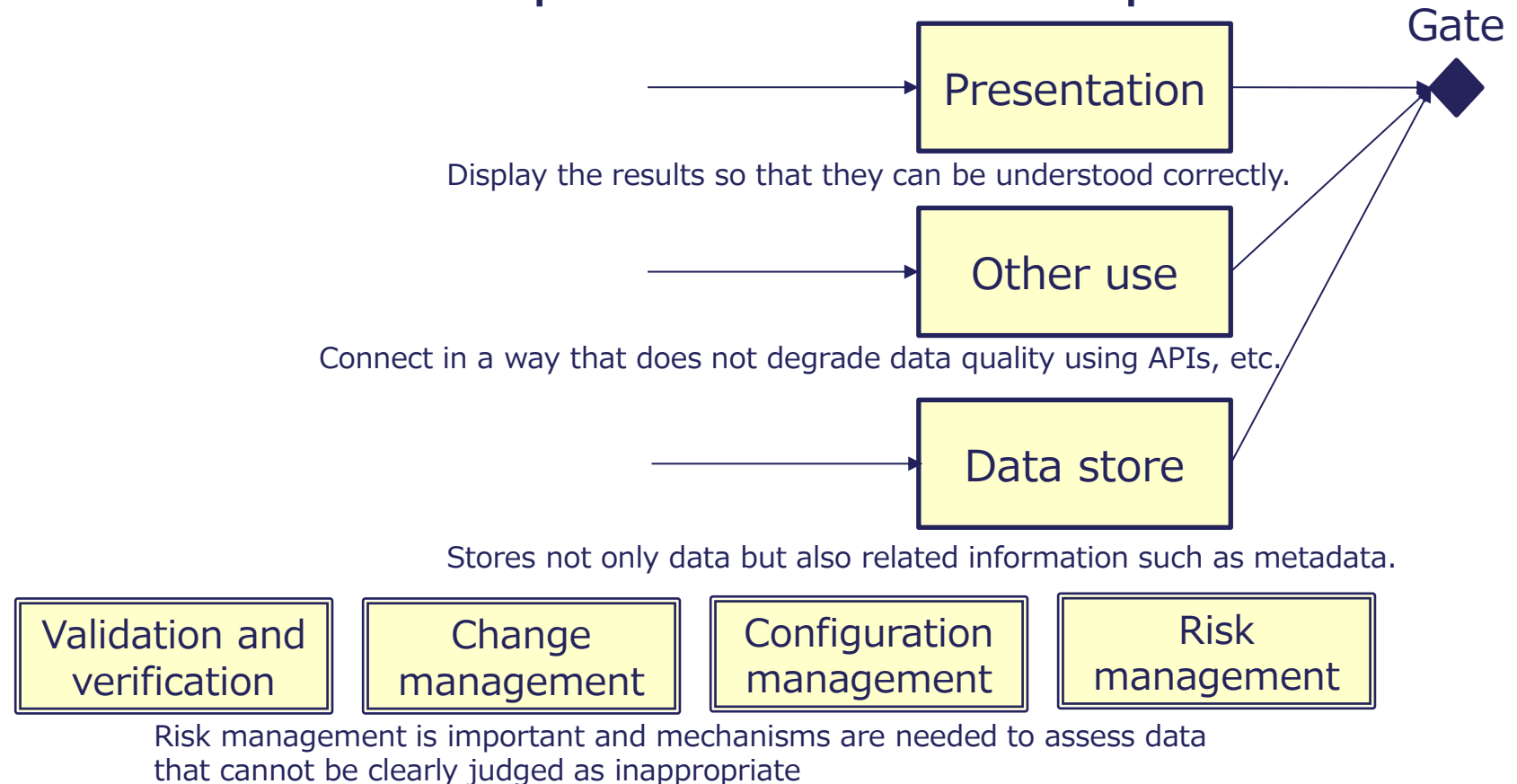
Description

- ♦ This phase involves delivering processed data and AI outputs to end-users in a clear and accessible manner.
- ♦ Proper presentation prevents misunderstandings, ensuring trust and usability.
- ♦ Metadata adds clarity by providing context and reliability.
- ♦ Watermarking can also be used to secure data and protect intellectual property.



Actions

- ♦ Outputs should be made publicly available or provided for use in other systems. It is important to prevent misinterpretation and its potential social impact



Procedure and checkpoint

Procedure

Presentation

1. Operate the access control function
2. Show the processing results
 - Easy to understand
 - Visualization
3. Gather the user's opinion

Other use

1. Operate the access control function
2. Provide API and relevant information

Data store

1. Store the processing results
2. Back up the data
3. Optimize data storage

Checkpoint

- ❑ Are the answers and expressions misleading?
- ❑ Is there a mechanism to check the basis of the answers, for example, by making the provenance information available for checking?
- ❑ Does it provide a machine-readable interface?
- ❑ Are protective measures taken to ensure that data is not lost?

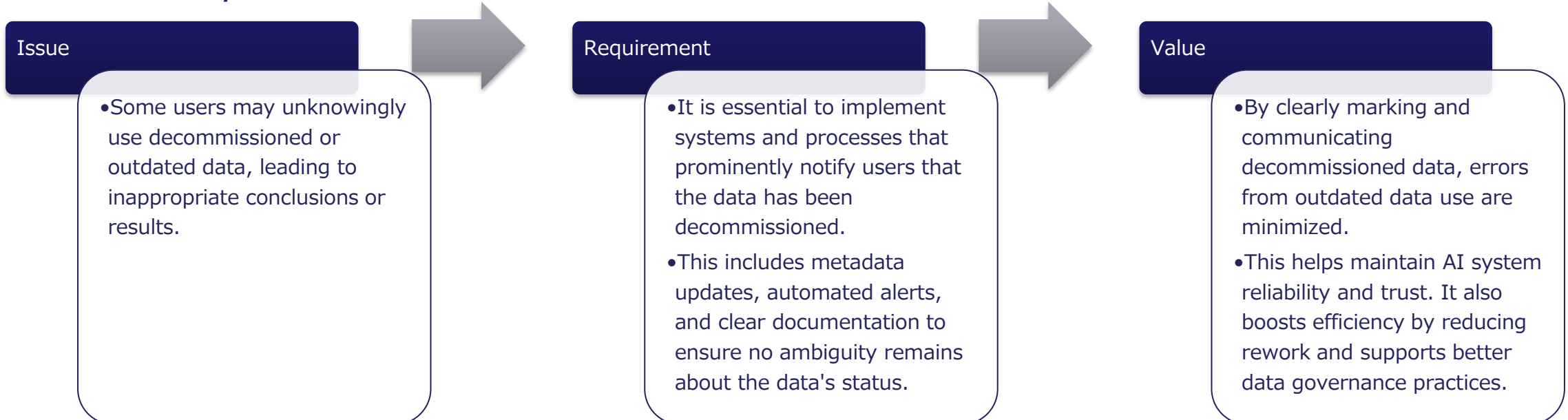
Column: Social and human impact

- Due to low-quality data, the system could provide incorrect data, which may shock or disadvantage users.
- Furthermore, due to low-quality data, the system might provide incorrect data, which could lead to accidents or impact on the economy.
- In addition to evaluating data before it is provided, it is also important to prepare countermeasures in case any negative impact occurs.
- If incorrect information is released, it is important to have reliable sources that can provide and verify accurate data.

8. Decommissioning

Description

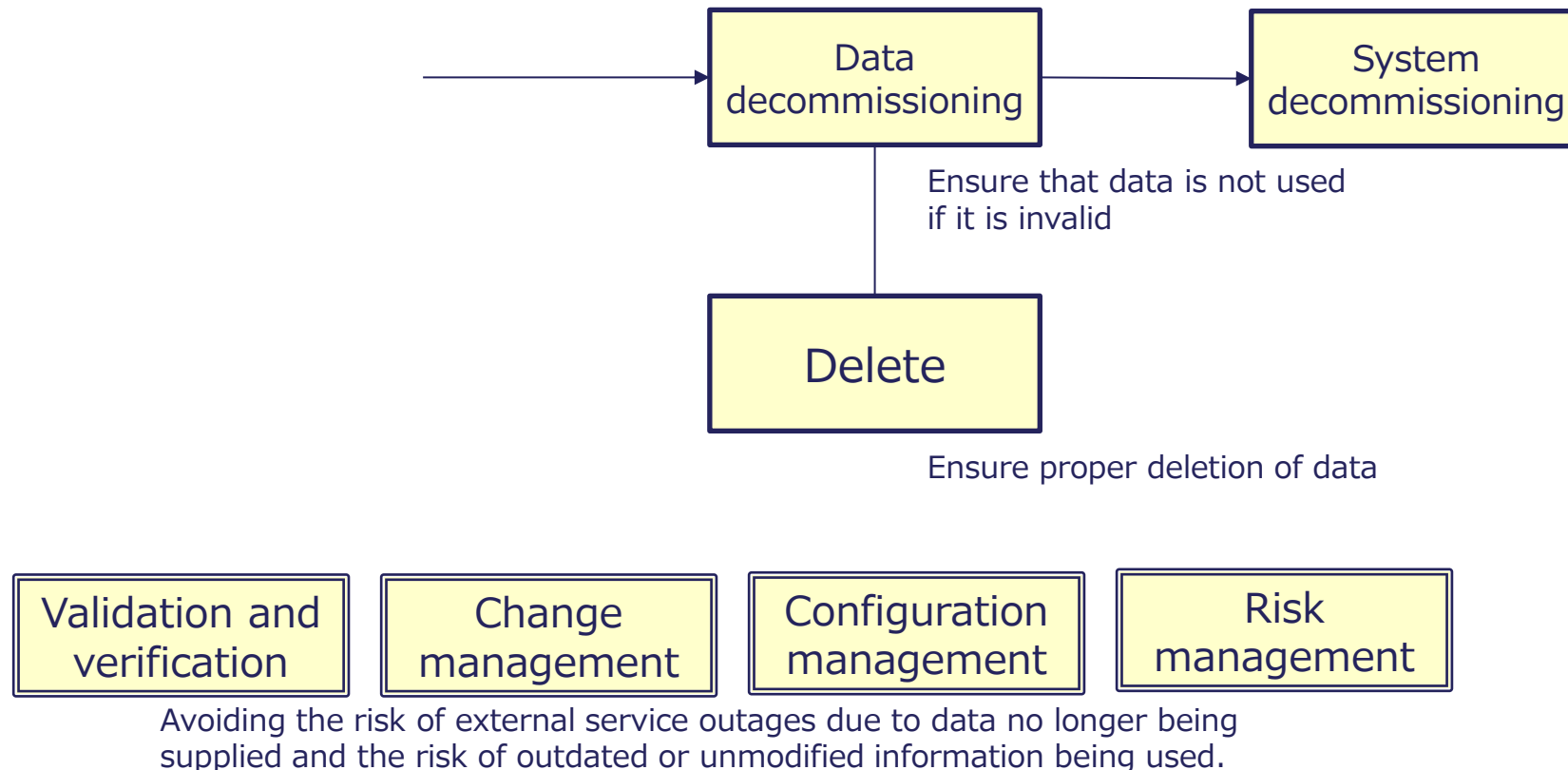
- ♦ The data quality control process is completed, and the data is officially decommissioned.
- ♦ Clear measures are taken to mark it as decommissioned to prevent accidental or unintended use.
- ♦ This step is essential for maintaining the integrity of AI systems and avoiding errors caused by outdated data.



8. Decommissioning

Actions

- ♦ Stop providing data. Provide advance notice of decommissioning and indicate that the data is no longer guaranteed thereafter.



Procedure and checkpoint

Procedure

Data decommissioning

1. Notify users of the decommissioning of the data
2. Provide information on data that has been decommissioned.
3. When transferring data, the data, metadata and related documents are transferred.

Delete

1. If necessary, archive the data.
2. Erase the data so that it cannot be restored
3. When deleting parts, give prior notice.

System decommissioning

1. Dispose of system components.

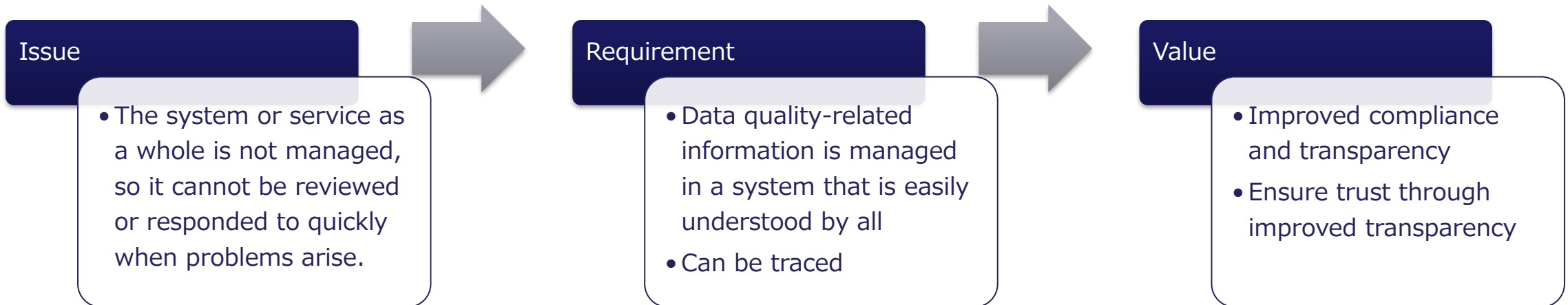
Checkpoint

- ❑ Did you give sufficient notice before suspending the data?
- ❑ If you are transferring to another party, did you provide sufficient information for them to take over?
- ❑ Did you check the results of the deletion?
- ❑ Are system components properly decommissioned?

9. Action throughout the Life Cycle

Description

- ♦ It is necessary to implement 'validation and verification', 'change management', 'configuration management' and 'risk management' throughout the data life cycle.
- ♦ Through these initiatives, we will improve data quality across processes.



Procedure and checkpoint

Procedure

Validation and verification

1. Define the characteristics, targets and tolerance levels required for the system's purpose
2. Manage the system throughout the life cycle

Change management

1. Define a basic policy for changes, such as data integration, processing, and modification
2. Define policies for data degradation over time, e.g., data refreshing
3. Record changes

Checkpoint

- ▣ Is data quality control becoming a burden?
- ▣ Is the change history readily available?

Procedure and checkpoint

Procedure

Configuration management

1. Manage software configurations
2. Manage data configurations
3. Manage the configurations of relevant documents

Risk management

1. Define data quality risks and response policies
2. Manage data quality risks using access control and continuous monitoring
3. Ensure that essential risk factors are addressed
4. Develop Business Continuity Plan (BCP)

Checkpoint

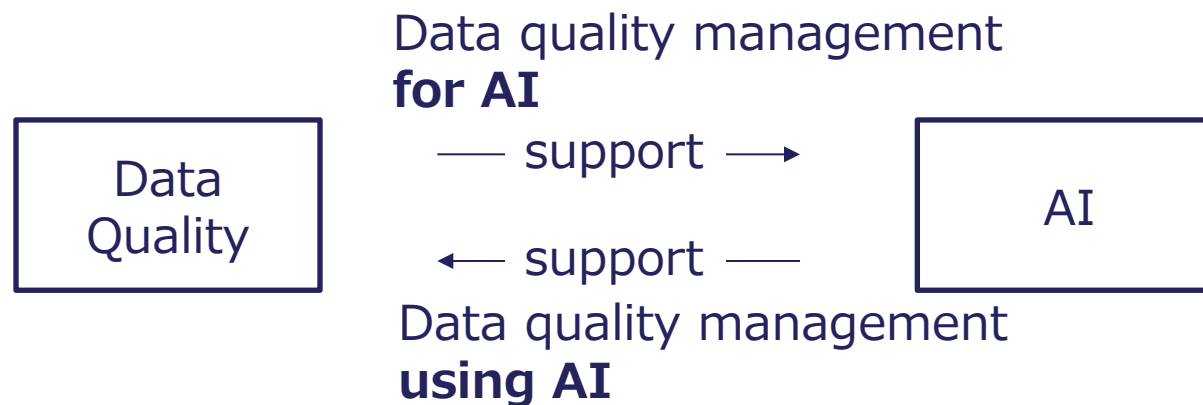
- ❑ Do you manage the list of software, data and documents?
- ❑ Do you understand the risk of your system or service?
- ❑ When a risk is discovered, does your organization have a culture of reviewing the situation from the root cause?

Column: Access control

- Access control is essential for maintaining data quality by preventing data poisoning and other forms of unauthorized manipulation.
- By defining and regulating which users, devices, or processes are permitted to access, modify, or delete data, organizations can significantly reduce the risk of malicious attacks and accidental errors.
- Key elements of effective access control include implementation of clearly defined user roles and privileges, enforcing multi-factor authentication, regularly auditing access logs, and continuously monitoring for unusual or unauthorized activities.
- This structured approach ensures data remains accurate, consistent, and secure throughout its entire life cycle.

Column: Improving data quality using AI

- AI-assisted data quality management involves using artificial intelligence to detect, correct, and prevent data errors.
- AI algorithms can automatically identify inconsistencies, missing values, and duplicates by learning from data patterns.
- They also suggest improvements, monitor data quality over time, and adapt to changing data environments. This helps ensure accurate, reliable, and up-to-date information for decision-making.



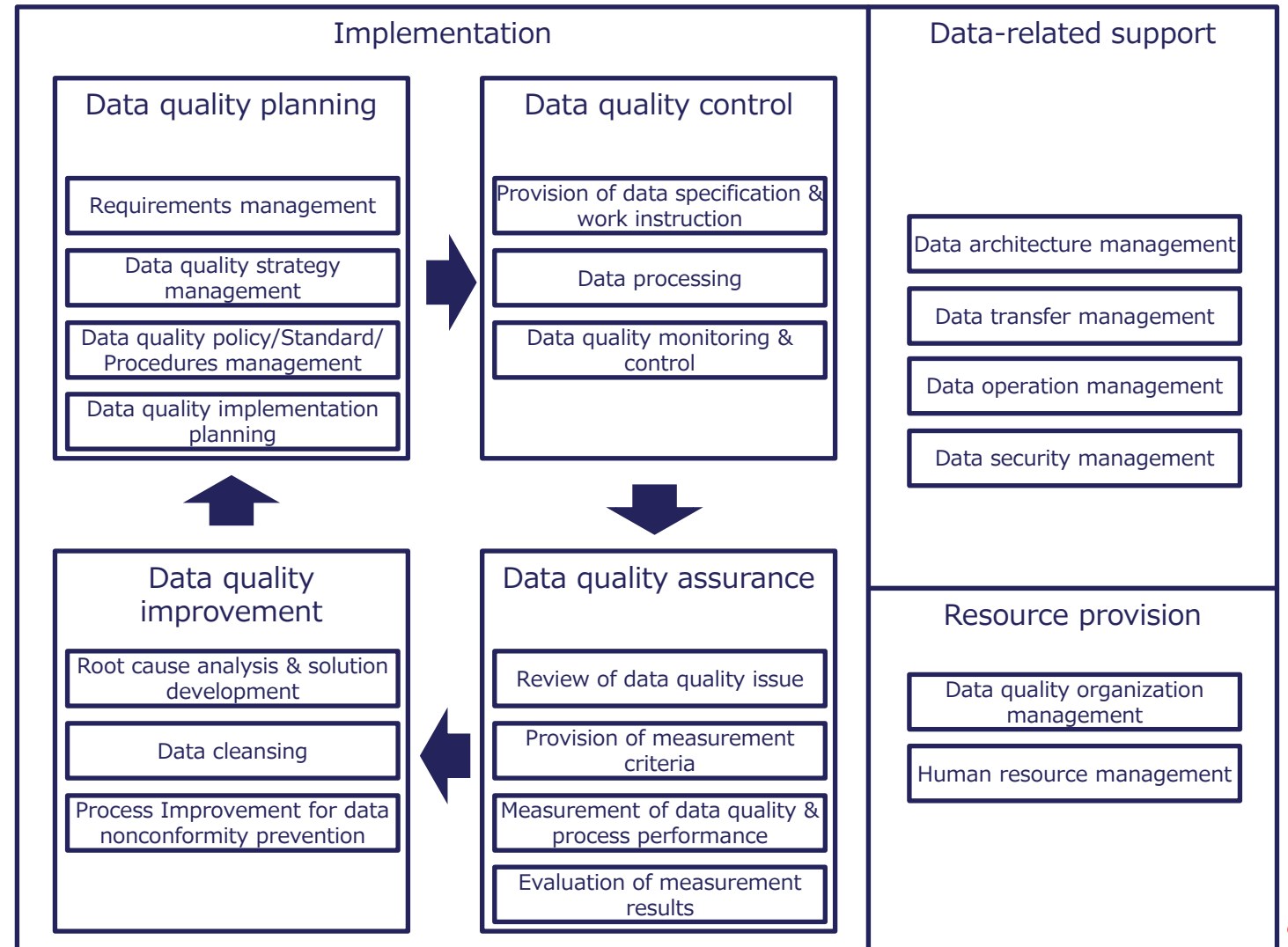
III. Perspective

- Governance Cycle View

Governance Cycle

Managing data quality involves performing individual activities correctly, as well as organizational controls to ensure they are sustainable and smooth operation.

Data governance aligned with ISO 8000-61.



Data Quality Planning - Overview

- ♦ Data quality planning is essential for ensuring that AI systems are built on accurate, reliable, and consistent data.
- ♦ High-quality data enables AI models to perform effectively and ethically, reducing the risks of bias, errors, and poor decision-making.
- ♦ The process involves a structured approach to identify data requirements, define quality standards, establish policies, and implement actionable plans to manage data quality throughout the AI life cycle.
- ♦ At this stage, it is important to ensure that executives understand the importance of data quality.

Data Quality Planning - Elements

Requirements

- Identify Business Needs: determine the specific objectives of the AI system and the data needed to achieve them.
- Define Data Quality Dimensions: Establish the critical dimensions (e.g., accuracy, completeness, consistency, timeliness) relevant to the project.
- Assess Current Data: Conduct a gap analysis of existing data to identify areas needing improvement.
- Document Requirements: Clearly document the data quality requirements to guide subsequent processes.

Data Quality Strategy

- Develop a Vision: Define the long-term goals for data quality aligned with the AI system's objectives.
- Set Measurable Targets: Establish key performance indicators (KPIs) to evaluate data quality over time.
- Stakeholder Engagement: Ensure alignment with all stakeholders, including data engineers, analysts, and decision-makers.
- Risk Management: Identify and mitigate potential risks associated with low-quality data in AI applications.

Data Quality Policy/Standards/Procedures

- Define Policies: Establish organizational policies for data governance and quality management.
- Create Standards: Develop specific standards for data collection, storage, processing, and validation to ensure consistency.
- Document Procedures: Outline detailed procedures for maintaining and improving data quality, such as periodic audits and validation protocols.
- Compliance Check: Ensure that all policies and standards align with relevant legal and regulatory requirements.

Data Quality Implementation Planning

- Develop an Action Plan: Create a step-by-step plan for implementing data quality measures, including timelines and responsibilities.
- Assign Roles: Define clear roles and responsibilities for team members involved in data quality management.
- Implement Tools and Processes: Deploy tools for data cleansing, validation, monitoring, and reporting.
- Monitor and Refine :Continuously monitor data quality and refine the implementation plan as needed based on feedback and performance metrics.

Data Quality Control - Overview

- ♦ Data quality control ensures that the data used in AI systems is accurate, consistent, and reliable.
- ♦ This process is critical for optimizing AI performance, reducing biases, and achieving trustworthy results.
- ♦ Effective data quality control involves clear specifications, robust processing techniques, and continuous monitoring to identify and address quality issues.
- ♦ The important thing at this stage is to avoid placing a burden on the workplace. It is necessary to make technical innovations, such as automatic checks.

Data Quality Control - Elements

Provision of Data Specifications and Work Instructions

- Define clear data requirements, including formats, structures, and acceptable value ranges.
- Provide detailed instructions for data collection, labeling, and processing to ensure uniformity.
- Establish guidelines for metadata creation to track data sources, timestamps, and context information.
- Develop a validation checklist for contributors to ensure adherence to data quality standards.

Data Processing

- Clean and preprocess data to remove inconsistencies, duplicates, and irrelevant information.
- Normalize data to ensure compatibility across systems (e.g., format conversions, scaling).
- Annotate and label data accurately for supervised learning models.
- Implement automated data validation tools to identify anomalies and errors during processing.
- Document each processing step to maintain traceability and reproducibility.

Data Quality Monitoring and Control

- Establish key quality metrics (e.g., accuracy, completeness, consistency, timeliness).
- Implement real-time monitoring systems to detect data quality issues promptly.
- Schedule regular audits to review and validate data integrity.
- Use AI tools to identify patterns of errors or potential biases in datasets.
- Create feedback loops to continuously improve data quality standards based on findings.

Data Quality Assurance - Overview

- ♦ Data quality assurance (DQA) ensures that the data used in AI development and operations meets required standards of accuracy, consistency, completeness, and reliability.
- ♦ This process is essential for preventing errors, biases, and inefficiencies in AI systems.
- ♦ DQA involves identifying potential data quality issues, defining evaluation criteria, measuring data quality, and analyzing results to guide continuous improvements.

Data Quality Assurance - Elements

Review of Data Quality Issues

- Identify potential issues such as missing values, inconsistencies, inaccuracies, or redundancies in the dataset.
- Analyze root causes of data quality problems, including errors in data collection, processing, or storage.
- Document known issues and assess their potential impact on AI model performance.

Provision of Measurement Criteria

- Define clear and measurable criteria for data quality attributes (e.g., accuracy, completeness, timeliness, consistency, and validity).
- Establish benchmarks and thresholds that data must meet to be considered suitable for use.
- Align criteria with the specific requirements of AI models and business objectives.

Measurement of Data Quality and Process Performance

- Conduct systematic evaluations of datasets against the established measurement criteria.
- Use data profiling tools to detect anomalies and assess quality attributes.
- Monitor data processing workflows to ensure consistent adherence to quality standards.

Evaluation of Measurement Results

- Analyze measurement outcomes to identify gaps and areas for improvement.
- Quantify the impact of data quality issues on AI system performance and decision-making.
- Generate reports and dashboards to communicate findings to stakeholders.

Data Quality Improvement - Overview

- ♦ Data quality improvement is a critical process for ensuring the reliability and effectiveness of AI systems.
- ♦ It involves identifying and resolving data issues, enhancing data consistency, and implementing preventive measures to maintain high-quality data over time.
- ♦ This process ensures that AI models can deliver accurate, unbiased, and actionable insights.
- ♦ Rather than merely modifying the data, it is essential to address the root causes, such as difficulties of data entry or inefficiencies in the process itself, to achieve sustainable improvement.

Data Quality Improvement - Elements

Root Cause Analysis and Solution Development

- Identify the Source of Issues: Investigate the underlying causes of data quality problems, such as incorrect data entry, system errors, or outdated data.
- Develop Targeted Solutions: Design and implement specific corrective actions to address identified root causes, such as updating workflows, improving validation rules, or automating data entry processes.
- Monitor and Validate Results: Continuously evaluate the effectiveness of implemented solutions to ensure the issues are resolved and do not reoccur.

Data Cleansing

- Identify Inaccuracies: Detect duplicate, incomplete, or inconsistent records within the dataset.
- Standardize Data Formats: Ensure uniformity in data presentation, such as consistent date formats, unit measurements, and naming conventions.
- Correct or Remove Errors: Modify inaccurate entries, fill missing values, or remove irrelevant or outdated data to maintain dataset integrity.
- Automate Cleansing Processes: Leverage tools and algorithms to automate repetitive data cleansing tasks and reduce manual errors.

Process Improvement for Data Nonconformity Prevention

- Establish Data Governance Policies: Define and enforce clear policies for data collection, management, and usage to prevent quality issues.
- Enhance Data Validation Mechanisms: Implement real-time validation checks during data entry or ingestion to identify nonconformities early.
- Train Stakeholders: Educate employees and data handlers about best practices for maintaining data quality and the importance of adhering to standards.
- Monitor and Audit Data Processes: Regularly assess data workflows to detect potential sources of nonconformity and ensure compliance with established policies.

Data-related Support - Overview

- ♦ Effective data quality governance is critical for AI development and utilization.
- ♦ Data-related support encompasses a comprehensive set of activities designed to ensure the integrity, consistency, and reliability of data across its life cycle.
- ♦ These efforts focus on establishing frameworks, processes, and tools to optimize data management while addressing compliance, security, and operational needs.

Data-related Support - Elements

Data Architecture Management

- Define data models, schemas, and standards to ensure consistency. Develop and manage metadata repositories to enhance data discoverability and traceability.
- Implement data lineage tracking to monitor the flow and transformation of data.
- Ensure scalability and flexibility to accommodate evolving AI demands.
- Align architecture with organizational goals and compliance requirements.

Data Transfer Management

- Establish protocols for secure data transfer (e.g., encryption, VPNs).
- Monitor and optimize data flow to prevent bottlenecks and ensure real-time availability. Maintain logs and audits of data transfers for accountability and traceability.
- Define policies for cross-border data transfers to comply with legal regulations (e.g., GDPR, CCPA).
- Automate data transfer processes where applicable to reduce manual errors.

Data Operations Management

- Conduct regular data validation and cleansing to maintain accuracy.
- Manage data storage solutions to ensure accessibility and scalability.
- Establish workflows for data ingestion, processing, and integration. Monitor system performance and resolve issues related to data operations.
- Implement version control for datasets to track changes and ensure consistency.

Data Security Management

- Implement access controls, including role-based permissions and multifactor authentication.
- Conduct regular security audits and vulnerability assessments.
- Deploy encryption protocols for data at rest and in transit. Monitor for suspicious activities and respond to security incidents promptly.
- Ensure compliance with global and local data protection regulations.

Resource Provision - Overview

- ♦ Resource provision supports effective data quality governance by ensuring that the necessary organizational structures, human resources, and operational capabilities are in place.
- ♦ This ensures high-quality data for AI systems, minimizing risks and maximizing the accuracy, reliability, and fairness of AI models.
- ♦ The process involves structuring teams, allocating skilled personnel, and defining clear roles and responsibilities.

Resource Provision - Elements

Data Quality Organization Management

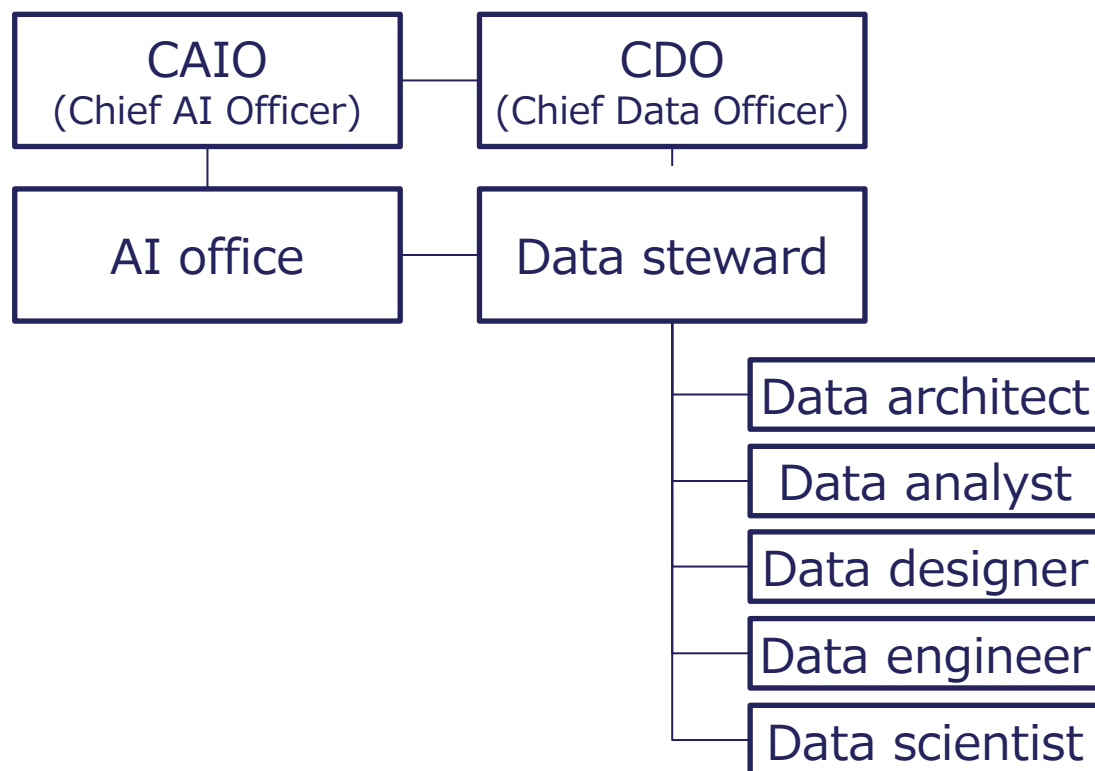
- Establishing a dedicated Data Quality Governance team or council.
- Defining roles and responsibilities for data quality management (e.g., Data Stewards, Data Quality Analysts).
- Setting up clear reporting lines and accountability for data quality issues.
- Creating cross-functional collaboration frameworks to involve stakeholders from various departments.
- Regularly reviewing and updating organizational structures to adapt to evolving data and AI needs.

Human Resource Management

- Recruiting data quality professionals with expertise in data governance, data architecture, and AI integration.
- Providing training programs to enhance employees' skills in data quality assessment, cleansing, and validation.
- Defining career paths and development plans for data governance professionals to improve retention.
- Ensuring adequate staffing levels for ongoing data quality monitoring and improvement tasks.
- Promoting a culture of data quality awareness and accountability across the organization.

Column: Human resource and team

- The data governance team will work with the AI team to improve data quality.



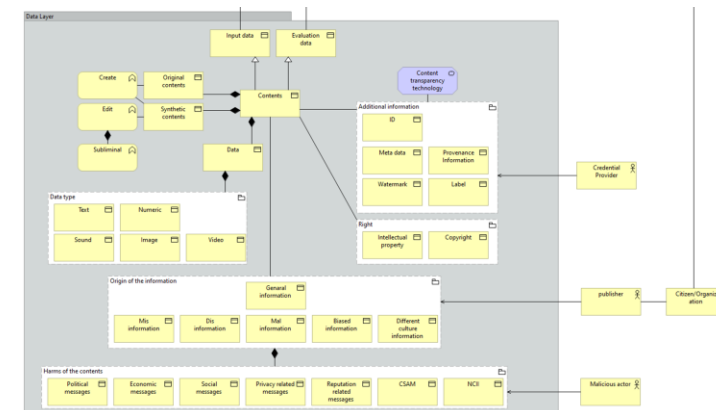
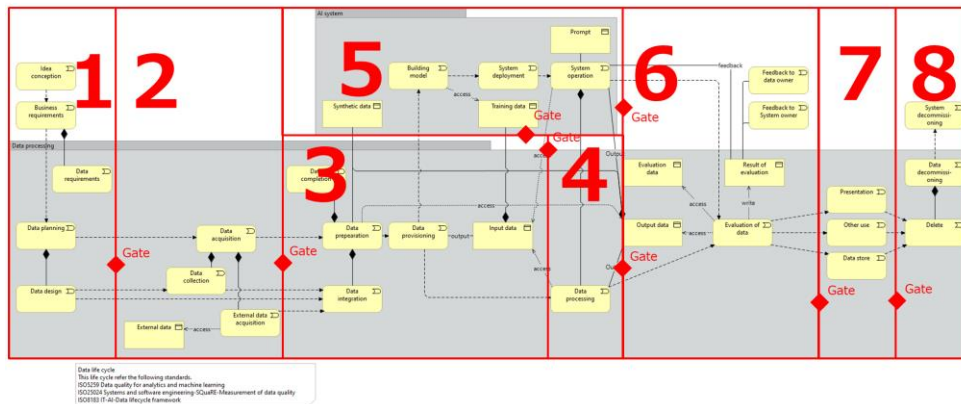
- Business fields using the data also need literacy training, including data quality.

III. Perspective

- Gateway View(Characteristics)

Verification of Data Quality Characteristics

- ♦ To maintain data quality, data quality needs to be checked at process boundaries and transactions.
- ♦ In the case of sensor data, this may be a periodic inspection.
- ♦ During the inspection process, each characteristic should be checked, assuming the type of data and events that may degrade the quality of the data.
- ♦ The checking of characteristics should be automated where possible, so as not to increase the workload on the site.



List of Characteristics

- ◆ Utilizes the required quality characteristics from among many characteristics.

Inherent data quality characteristics

- Accuracy
- Completeness
- Consistency
- Credibility
- Currentness

Inherent and system-dependent data quality characteristics

- Accessibility
- Compliance
- Confidentiality
- Efficiency
- Precision
- Traceability
- Understandability

System-dependent data quality

- Availability
- Portability
- Recoverability

Additional data quality characteristics for AI/ML

- Auditability
- Balance
- Diversity
- Effectiveness
- Identifiability
- Relevance
- Representativeness
- Similarity
- Timeliness

Sensor data quality characteristics

- Accuracy
- Completeness
- Consistency
- Precision

Accuracy

- Accuracy ensures that the data reflect real-world values correctly. High accuracy is vital for AI to produce reliable outcomes, avoiding false predictions or flawed decisions.
- Evaluation Points
 - Data match verified sources.
 - Error rate is within acceptable thresholds.
 - Outliers are justified or corrected.
- Inappropriate Examples
 - Incorrect labels in training datasets.
 - Mismatched units in numerical data.
 - Spelling errors in text data.

Completeness

- Completeness measures whether all required data points are present. Missing values can distort AI predictions and compromise system effectiveness.
- Evaluation Points
 - All critical fields are filled.
 - Null or missing rates are minimal.
 - Completeness checks align with AI model requirements.
- Inappropriate Examples
 - Missing demographic attributes in user profiles.
 - Partial transaction records.
 - Null values in key input variables.

Consistency

- Consistency ensures that data values are uniform across datasets and time. Inconsistencies can cause AI to misinterpret patterns.
- Evaluation Points
 - Uniform formatting and units.
 - Synchronized datasets over time.
 - No conflicting entries within linked data.
- Inappropriate Examples
 - Different date formats in a dataset.
 - Duplicate records with conflicting details.
 - Mismatched entries in integrated datasets.

Credibility

- Credibility measures the trustworthiness of data sources. Reliable sources improve AI model validity and trust.
- Evaluation Points
 - Verified, reputable data origins.
 - Data provenance is documented.
 - Peer-reviewed or authenticated sources.
- Inappropriate Examples
 - Data from unknown/unverified sources.
 - Fake or manipulated datasets.
 - User-generated content without validation.

Currentness

- Currentness assesses how up-to-date data is. Stale data can result in outdated AI insights or actions.
- Evaluation Points
 - Data aligns with the latest context.
 - Timestamping practices ensure traceability.
 - Regular updates are maintained.
- Inappropriate Examples
 - Outdated stock prices in financial models.
 - Old weather data in climate predictions.
 - Historical user preferences in real-time applications.

Accessibility

- Accessibility ensures that data is available to authorized users and systems without barriers. It involves proper storage, robust APIs, and universal design principles for all users. Accessibility supports efficient AI model training and application by enabling smooth data flow.
- Evaluation Points
 - Data can be accessed across platforms and devices.
 - APIs are well-documented and error-free.
 - Meets accessibility standards for all user demographics.
- Inappropriate Examples
 - Data locked in proprietary formats.
 - Missing API documentation.
 - Non-compliance with ADA accessibility standards.

Compliance

- Compliance ensures that data management adheres to laws, regulations, and industry standards like GDPR or CCPA. Proper compliance avoids legal risks, protects privacy, and builds trust, which is critical for AI system reliability and public acceptance.
- Evaluation Points
 - Meets all relevant legal standards.
 - Implements proper data consent mechanisms.
 - Regular compliance audits are conducted.
- Inappropriate Examples
 - Collecting data without consent.
 - Ignoring jurisdictional privacy laws.
 - Lacking audit trails for data handling.

Confidentiality

- Confidentiality ensures that sensitive data is protected from unauthorized access or breaches. Essential measures include encryption, access controls, and anonymization techniques. Maintaining confidentiality safeguards trust and reduces the risks associated with data misuse.
- Evaluation Points
 - Strong encryption for data at rest and transit.
 - Role-based access control systems.
 - Regularly updated security protocols.
- Inappropriate Examples
 - Storing sensitive data in plaintext.
 - Sharing private data without user consent.
 - Weak or outdated access controls.

Efficiency

- Efficiency in data management minimizes processing time and resource usage without compromising quality. Efficient data ensures faster AI training and deployment, cost savings, and scalability.
- Evaluation Points
 - Optimized data storage structures.
 - Minimal latency in data access.
 - Efficient ETL pipelines.
- Inappropriate Examples
 - Redundant data processing steps.
 - High latency in API calls.
 - Excessive resource consumption for simple tasks.

Precision

- Precision ensures the degree of detail or granularity in data. The higher the precision, the more exact and detailed the information that can be obtained.
- Evaluation Points
 - Data values are recorded with sufficient detail.
 - Units and scales are clearly specified
 - Temporal and spatial information is recorded with appropriate granularity.
- Inappropriate Examples
 - Latitude and longitude values are too coarse.
 - Values with inconsistent levels of precision, making calculations difficult.
 - Excessively high resolution increases the processing or storage burden.

Traceability

- Traceability tracks the origin, transformations, and use of data, ensuring accountability and reproducibility. Detailed logs and metadata enhance transparency, critical for debugging and compliance.
- Evaluation Points
 - Comprehensive data lineage documentation.
 - Transformation logs maintained.
 - Unique IDs for traceability.
- Inappropriate Examples
 - Missing source details for datasets.
 - Unlogged data modifications.
 - Inconsistent versioning of datasets.

Understandability

- Understandability ensures that data can be interpreted correctly by both humans and machines. Clear labeling, metadata, and intuitive structures improve usability, essential for effective AI training.
- Evaluation Points
 - Data labeled clearly and comprehensively.
 - Metadata aligns with schema standards.
 - Consistent and logical data structure.
- Inappropriate Examples
 - Vague or missing labels.
 - Complex, undocumented structures.
 - Misleading metadata annotations.

Availability

- Availability ensures that AI data is accessible whenever needed. Reliable systems minimize downtime and ensure continuous operation, crucial for real-time AI applications.
- Evaluation Points
 - Uptime percentage meets service-level agreements (SLAs).
 - Redundancy mechanisms to prevent single points of failure.
 - Regular monitoring and alerts for accessibility issues.
- Inappropriate Examples
 - Frequent server outages disrupting data access.
 - No backups, leading to inaccessible data during hardware failure.
 - Delays in resolving access issues during critical tasks.

Portability

- Portability refers to the ability to transfer AI data seamlessly across platforms, systems, or environments without compatibility issues. It ensures flexibility and adaptability.
- Evaluation Points
 - Data is stored in widely accepted formats (e.g., CSV, JSON).
 - Use of standardized APIs for data exchange.
 - Adequate documentation for data migration.
- Inappropriate Examples
 - Proprietary formats requiring specialized software.
 - Inconsistent data structures across systems.
 - Lack of metadata, causing misinterpretation during migration.

Recoverability

- Recoverability focuses on the system's ability to restore AI data quickly and accurately after unexpected disruptions or failures, ensuring minimal data loss.
- Evaluation Points
 - Regular backups with multiple restore points.
 - Disaster recovery plans tested periodically.
 - Redundant storage to avoid single points of failure.
- Inappropriate Examples
 - Outdated backups causing irreversible data loss.
 - Recovery processes requiring significant manual intervention.
 - Failure to test recovery plans, leading to delays.

Auditability

- Auditability ensures data traceability and the ability to review the processes of data collection and usage. It enables accountability and compliance with ethical and legal standards.
- Evaluation Points
 - Clear documentation of data sources.
 - Well-defined data collection processes.
 - Availability of data lineage records.
- Inappropriate Examples
 - Missing metadata for data sources.
 - Ambiguous data provenance.
 - Inaccessible logs for key data processes.

Balance

- Balance ensures that data represents all relevant categories or outcomes proportionally, minimizing bias in AI models.
- Evaluation Points
 - Equal representation of categories.
 - Avoidance of over/under-sampling.
 - Consistency across datasets.
- Inappropriate Examples
 - Gender imbalance in a dataset.
 - Skewed representation of geographical regions.
 - Over-representation of one age group.

Diversity

- Diversity ensures that datasets include a wide range of perspectives, scenarios, and variations for better generalization in AI systems.
- Evaluation Points
 - Coverage of different cultural contexts.
 - Inclusion of varied scenarios and demographics.
 - Range of linguistic expressions.
- Inappropriate Examples
 - Excluding minority dialects.
 - Homogeneous data in multilingual settings.
 - Ignoring varied environmental factors.

Effectiveness

- Effectiveness measures whether the data fulfill its intended purpose and support the desired model outcomes.
- Evaluation Points
 - Alignment with model objectives.
 - Testing for model improvements.
 - Appropriate feature representation.
- Inappropriate Examples
 - Redundant data with no value.
 - Data leading to incorrect predictions.
 - Irrelevant features lowering model performance.

Identifiability

- Identifiability ensures that sensitive or personal data are anonymized to protect privacy while maintaining utility.
- Evaluation Points
 - Adherence to anonymization standards.
 - Removal of personal identifiers.
 - Assessment of re-identification risks.
- Inappropriate Examples
 - Retention of identifiable personal details.
 - Incomplete data masking.
 - Reversible anonymization techniques.

Relevance

- Relevance ensures that data align with the use case, covering only necessary and meaningful information.
- Evaluation Points
 - Targeted data for the AI model's domain.
 - Exclusion of irrelevant variables.
 - Avoidance of redundant information.
- Inappropriate Examples
 - Outdated information in dynamic applications.
 - Including unrelated features.
 - Excessive focus on non-critical variables.

Representativeness

- Representativeness ensures that data reflect the real-world population or scenarios the AI model will encounter.
- Evaluation Points
 - Alignment with target population characteristics.
 - Coverage of expected conditions.
 - Avoidance of sampling bias.
- Inappropriate Examples
 - Over-sampling urban populations in national studies.
 - Ignoring rare but critical conditions.
 - Incomplete geographic coverage.

Similarity

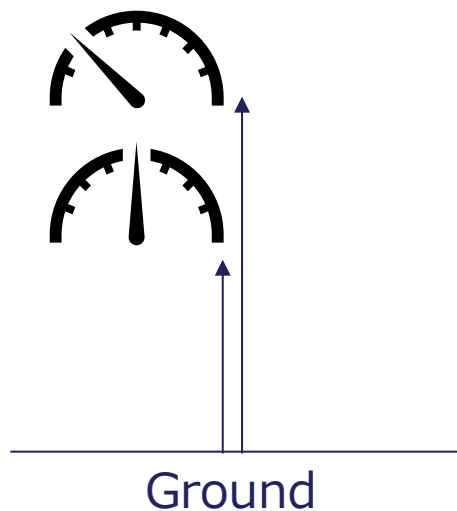
- Similarity measures the consistency between training data and real-world data, reducing model performance gaps.
- Evaluation Points
 - Alignment with deployment environment.
 - Testing against real-world scenarios.
 - Identification of data mismatches.
- Inappropriate Examples
 - Synthetic data misrepresents reality.
 - Inconsistent formatting across datasets.
 - Differences in distributions between datasets.

Timeliness

- Timeliness ensures that data is up-to-date and reflects current conditions, crucial for accuracy in dynamic models.
- Evaluation Points
 - Regular data updates.
 - Tracking time-sensitive variables.
 - Expiration checks on older datasets.
- Inappropriate Examples
 - Using outdated population statistics.
 - Training on obsolete customer trends.
 - Neglecting updates in real-time applications.

Sensor data

- Sensors are used in various environments, including outdoors and inside equipment. In addition, the measurement method will also make a difference, so it is necessary to manage the data collection according to the type of sensor.



The same place, but the values differ depending on the height measured and the device used.

Category	Device-dependent quality metric	Description
Design information	Device information	Level of understanding of the measurement principles, processing methods, etc. for the physical quantities (light, sound, etc.) input to the device
	Fault-tolerance	Level of device operation
	Durability	Level of decline in serviceable parts
	Security measures	Level of implementation of security measures
	Communication stability	Level of operation without communication interruption or delay
Installation and adjustment	Appropriateness of installation method	Level of implementation of appropriate installation according to conditions
Operation and maintenance	System stability	Level of operation stability
	System environment monitoring	Level of installation status monitoring
	Appropriateness of updates	Level of appropriate software version operation

Characteristics of Sensor Data

- ♦ Although each piece of sensor data is small, when aggregated in real time, it can become a large amount of data.
- ♦ Many of them are incorporated into devices and services and are directly linked to safety. In addition, they are often set up in various environments, such as outdoors, and are easily affected by the external environment.
- ♦ On the other hand, when there are many sensors, if one fails, it can be supplemented using data from surrounding sensors, and in some cases, data can be supplemented using data from before and after in chronological order.
- ♦ For these reasons, the following four points need to be managed as particularly important data quality items.
 - Accuracy, Completeness, Consistency, Precision

Time Characteristics of Sensor Data

- ♦ Sensor data may change over time and data characteristics may change, requiring compensation measures.

Offset

Constant deviation from true value.

Drift

Gradual change over time.

Trim

Adjustment to correct error.

Spike

Sudden, short-lived jump.

Noise

Random variations in data.

Data loss

Missing data points or gaps.

Lack of data

Insufficient data collected.

Shift

Sudden baseline change.

Drop or Rise

Abrupt decrease or increase.

Stuck

Repeated constant readings.

Bounded oscillation

Regular, limited fluctuations.

Inconsistent frequency

Irregular data intervals.

Different resolution

Varying data granularity.

Incorrect timestamp

Misaligned time in records.

Latency

Delay between event and record.

Challenges in using Multiple Sensors

When using multiple sensors, the following data quality issues can arise:

1. Inconsistent Data

- **Mismatch in Sensor Outputs:** Different sensors measuring the same phenomenon might provide inconsistent values due to calibration errors or varying sensitivity.
- **Different Units or Scales:** Sensors may report data in incompatible units or scales.

2. Temporal Misalignment

- **Clock Drift:** Sensors might not be synchronized, causing timestamps to be out of alignment.
- **Latency Variability:** Different sensors may have varying delays in recording or transmitting data.

3. Spatial Misalignment

- **Incorrect Sensor Positioning:** Sensors may not be placed in the correct locations, leading to inaccurate measurements for a specific area or system.

Challenges in using Multiple Sensors(Cont.)

4. Noise and Interference

- Environmental Noise: Sensors may pick up irrelevant signals from the environment, degrading the data quality.
- Cross-Talk: Signals from one sensor interfere with another, causing erroneous readings.

5. Resolution and Precision Variability

- Different Granularities: Sensors may record data at different levels of precision or resolution, making integration challenging.
- Sampling Rate Disparities: Inconsistent sampling rates across sensors lead to incomplete or redundant data points.

6. Faults and Anomalies

- Sensor Drift: Some sensors might degrade over time, resulting in inaccurate data.
- Stuck or Frozen Sensors: Sensors might repeatedly report the same value due to hardware issues.
- Data Gaps: Sensor failures or transmission errors lead to missing data.

Challenges in using Multiple Sensors(Cont.)

7. Redundancy and Overlap

- Duplicate Data: Overlapping sensor coverage might lead to redundant data, increasing processing complexity.
- Conflicting Data: Redundant sensors may report contradictory values.

8. Incorrect Calibration

- Bias Errors: Calibration issues lead to consistent offsets or scaling errors in the measurements.

9. Integration Challenges

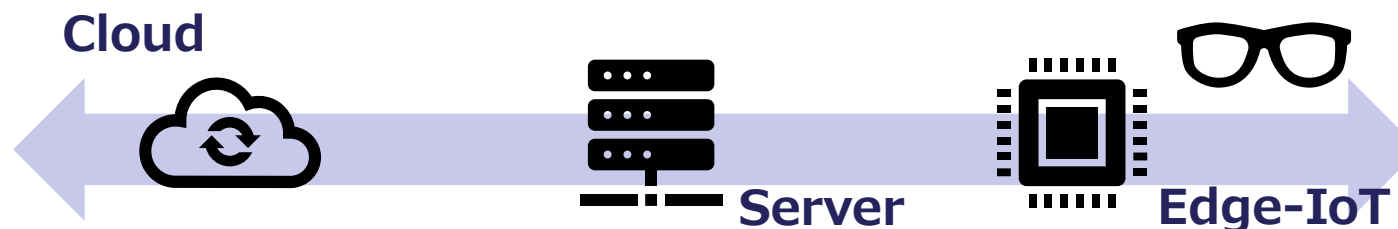
- Different Protocols: Sensors using different communication protocols might make data integration difficult.
- Heterogeneous Data Formats: Sensors may produce data in varied formats that need standardization.

10. Environmental Impact

- Temperature, Humidity, or Pressure: Environmental conditions might affect sensor performance differently.

Processing with Cloud-Edge-IoT

- ◆ Data collected by IoT devices and Edge may be processed on devices such as Edge AI, at data aggregation points, or collected in the cloud for mass data processing, requiring data quality measures at each location.
 - Cloud: Detects unusual or biased data in the course of processing large volumes of data.
 - Aggregation point: Data is aggregated in areas. Data conversion and integration, if necessary; some instructions may be sent to edge.
 - Edge: At the edge, data processing such as data cleansing, recognition and anonymization processes are carried out. Corrections such as sensor-specific offsets may also be performed.



Postscript

Message

- ◆ People are focusing on the cutting-edge technology of AI, but to ensure that we can use AI in society in a sustainable and secure manner, it is important to properly manage the quality of data.
- ◆ Keep in mind the principle of “garbage in, garbage out” at all times, and let's maximize the value of AI.

About us and this document

- ♦ AISI is the government initiative responsible for ensuring safety as a foundation for accelerating AI and AI-based innovation.
- ♦ IPA is a government-funded organization focused on digital technologies that participates in AISI's activities.
- ♦ This document has been jointly produced by AISI's standards team and a team of data experts from IPA's Digital Infrastructure Center, with valuable contributions from AISI's Data Quality Sub-Working Group.
- ♦ This is a living document. We look forward to hearing from you.

References

- ♦ ISO/IEC 25012: SQuaRE, Data quality model
- ♦ ISO/IEC 25024: SQuaRE, Measurement of data quality
- ♦ ISO 8000: Data quality
- ♦ ISO/IEC 5259: Data quality for analytics and machine learning (ML)
- ♦ ISO/IEC 8183: Data life cycle framework
- ♦ ISO/IEC 38505-1: Governance of IT, Governance of data
- ♦ ISO 19157: Geographic information, Data quality

- ♦ DAMA-DMBOK (2nd edition, 2017), DAMA International

- ♦ Government Interoperability Framework: Data quality management guide, Digital Agency, Japan
 - <https://github.com/JDA-DM/GIF>
- ♦ Data quality guidebook for data sharing, Cabinet Office, Japan
 - https://www.chisou.go.jp/tiiki/kokusentoc/supercity/supercity_230926_guidebook.html
- ♦ White Paper “Study for Formulating Guidelines for Evaluating the Quality Level of Sensing Data”, Data Sharing Association, Japan
 - <https://data-society-alliance.org/survey-research/data-quality-evaluation-standards/>
- ♦ Machine Learning Quality Management Guideline 4th Edition, AIST, Japan
 - <https://www.digiarc.aist.go.jp/publication/aiqm/>

Version History

Version	Release Date	Summary of Revisions
1.01	2025-12-02	<ul style="list-style-type: none">- Revised wording and adjusted page order to improve accuracy and readability without affecting the overall content.- Added About This Guide before the main sections.- Content previously listed as IV. Implementation in Version 1.00 contained only limited material, so the relevant parts were integrated into Section III. In future updates, the Implementation section will be reorganized and expanded.
1.00	2025-3-31	Initial release.

Appendix - 1

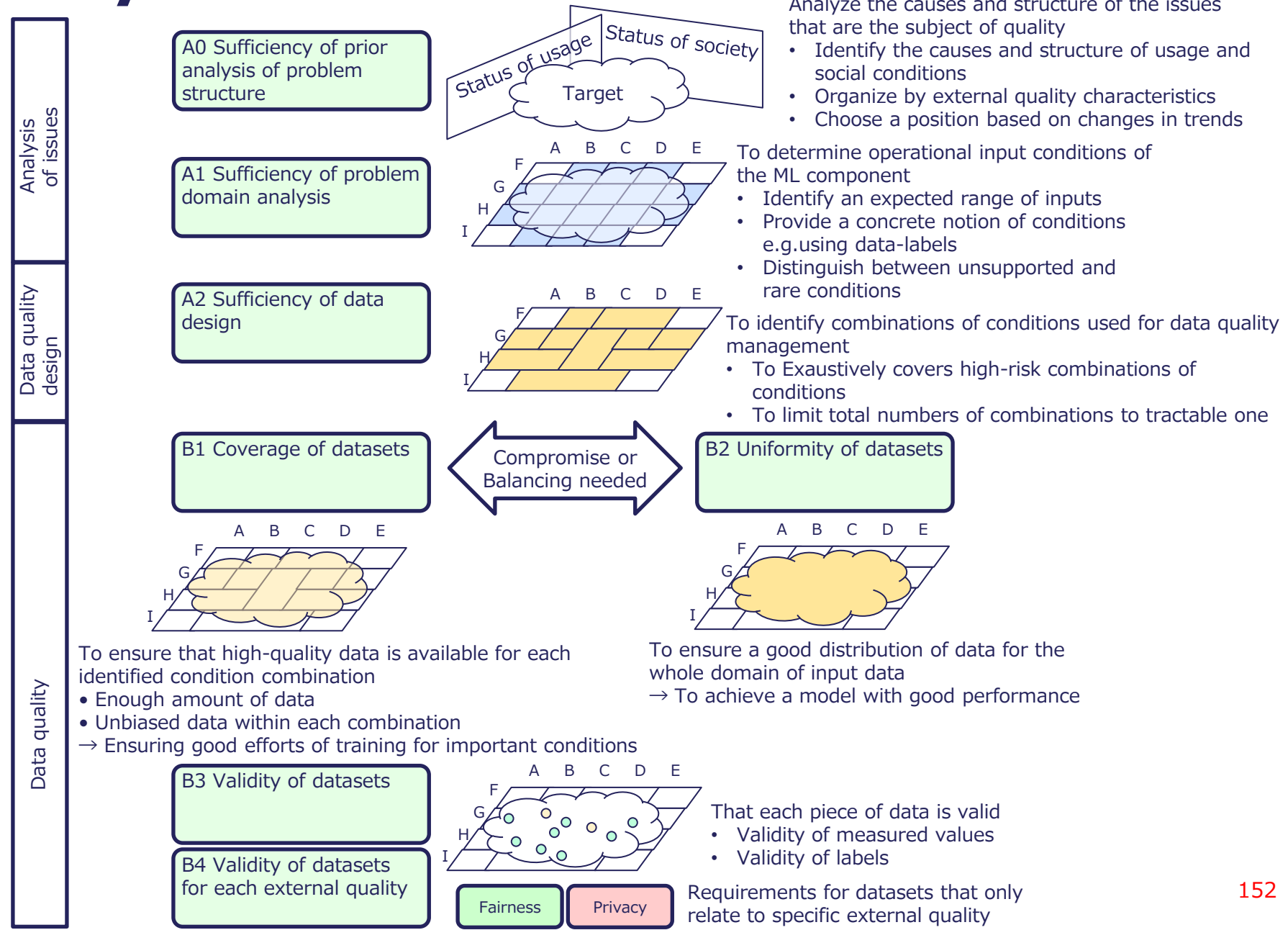
Machine Learning Quality Management Guideline 4th Edition

Technical Report Artificial Intelligence Research Center
National Institute of Advanced Industrial Science and Technology (AIST)
2023-12-12

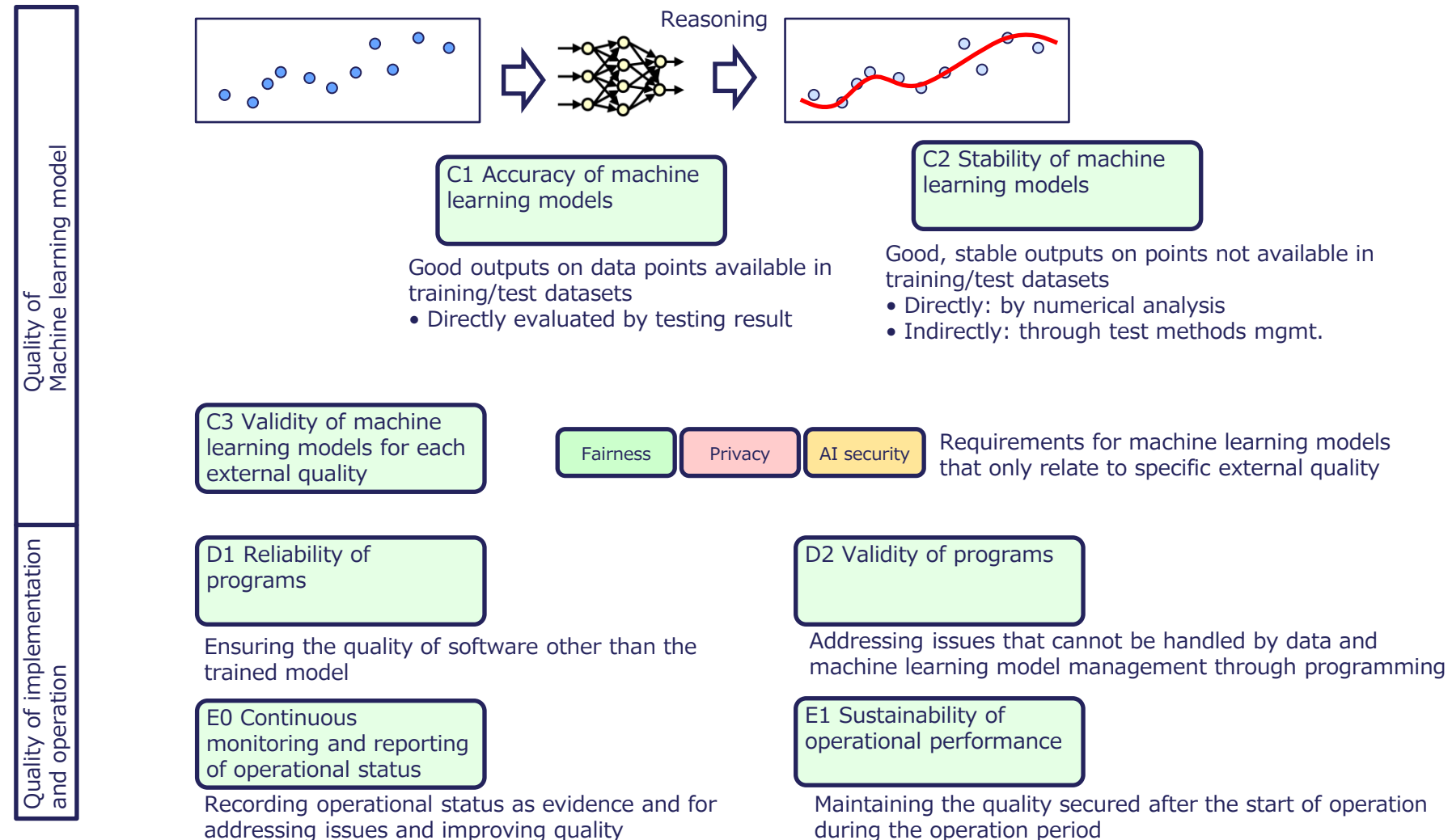


Internal quality characteristics

Data is a core element in AI. This guideline offers guidance on data quality management issues and responses to consider when promoting machine learning.



Internal quality characteristics(Cont.)





Disclaimer

While every effort has been made to ensure the accuracy and reliability of the information provided in this guidebook, We make no warranties, either express or implied, regarding its contents. We shall not be held liable for any losses, damages, or claims arising from the use of this guidebook. All information is provided “as is,” and the reader assumes full responsibility for any actions taken based on its content.