Munich Personal RePEc Archive

Unlocking Hidden Value: A Framework for Transforming Dark Data in Organizational Decision-Making

Leogrande, Angelo

LUM University Giuseppe Degennaro

2024

Online at https://mpra.ub.uni-muenchen.de/122776/
MPRA Paper No. 122776, posted 25 Nov 2024 15:05 UTC

Unlocking Hidden Value: A Framework for Transforming Dark Data in Organizational Decision-Making

Angelo Leogrande, LUM University Giuseppe Degennaro, Casamassima, Bari, Puglia, Italy, EU, leogrande.cultore@lum.it

Abstract

In today's data-driven world, organizations generate and collect vast amounts of information, yet not all data is managed or utilized with the same degree of efficiency and purpose. This paper investigates the taxonomy and distinctions among *white data*, *grey data*, and *dark data*, offering a comprehensive analytical framework to better understand their characteristics, value, and implications. *White data* refers to structured, accessible, and actively managed information that supports strategic decision-making and operational processes. In contrast, *grey data* occupies an intermediate space, representing semi-structured or unstructured data that, while not fully optimized, holds potential value when properly integrated into organizational practices. Lastly, *dark data* comprises the large quantities of information that remain unexploited, often due to a lack of resources, awareness, or technology. By mapping these categories, this paper aims to highlight the importance of a systematic approach in managing diverse data types, underscoring both the risks and opportunities associated with each. The study ultimately provides practical insights and recommendations for organizations seeking to maximize the value of their data assets through effective taxonomy and governance strategies.

JEL CODE: C8, C80, C81, C82, C83, C87.

KEYWORDS: Dark Data, White Data, Grey Data, Warehouse Management.

1. Introduction

In today's data-driven world, organizations generate and collect vast amounts of information, often spanning various formats, sources, and levels of structure. However, not all data is managed, utilized, or leveraged with the same degree of efficiency and purpose. The emergence of diverse data types has led to an increasingly complex information ecosystem, where the value and potential impact of data can vary significantly based on how it is categorized and managed. Understanding these variations is crucial for organizations aiming to capitalize on their data assets while mitigating associated risks. This paper investigates the taxonomy and distinctions among *white data*, *grey data*, and *dark data*, offering a comprehensive analytical framework to better understand their characteristics, value, and implications. *White data* refers to structured, accessible, and actively managed information. It is characterized by its organization and availability, making it a critical resource for supporting strategic decision-making and optimizing operational processes. Examples of white data include structured customer databases, sales reports, and regulated compliance documents, all of which are integrated into a company's information systems to deliver actionable insights and maintain competitive advantage. In contrast, *grey data* occupies an intermediate space in the data taxonomy. It often includes semi-structured or unstructured information such as emails, meeting notes, social media interactions, and sensor data. While this type of data may not be fully optimized or readily accessible, it holds latent value that can be unlocked when appropriately integrated into organizational practices. The management of grey data poses a challenge as it requires sophisticated

tools and strategies to transform these less structured sources into actionable intelligence. As companies navigate the digital age, the capacity to effectively harness grey data can provide critical insights into emerging patterns, customer preferences, and operational efficiencies. Lastly, the concept of *dark data* highlights the large quantities of information that remain unexploited within organizations. Dark data encompasses a wide array of sources, including old system logs, archived documents, and sensor data from IoT devices. Often overlooked due to a lack of resources, awareness, or appropriate technology, dark data represents a potentially valuable yet underutilized asset. However, it also poses risks, such as storage costs, compliance challenges, and security vulnerabilities, if not managed properly. By mapping these categories, this paper aims to highlight the importance of a systematic approach in managing diverse data types, underscoring both the risks and opportunities associated with each. Establishing a clear taxonomy not only aids in optimizing data utilization but also ensures organizations can navigate the complexities of data governance effectively. The study ultimately provides practical insights and recommendations for organizations seeking to maximize the value of their data assets through effective taxonomy and governance strategies. This work contributes to the broader discourse on data management by demonstrating how a structured approach can bridge the gap between data generation and data utilization, turning information into a strategic resource.

The originality of this idea lies in its development of a structured analytical framework that categorizes organizational data into three distinct types: white data, grey data, and dark data. While the concept of data management is not new, this paper introduces an innovative taxonomy that enables organizations to better understand and manage the full spectrum of their data assets. By clearly defining and distinguishing these categories, the framework provides organizations with a practical and strategic tool for optimizing their data utilization. White data, as characterized in the study, refers to actively managed, structured data that supports decision-making and operations—information that organizations typically prioritize. The novel aspect of the paper is its introduction and emphasis on grey data, which occupies a middle ground between structured and unstructured data. Unlike previous studies that primarily focus on either structured or unstructured data, this paper highlights the often-overlooked potential of grey data, demonstrating that it can be a significant resource if integrated effectively into business practices. Additionally, the paper's exploration of dark data, which remains unexploited due to various barriers such as technology and awareness, offers a fresh perspective on the untapped value that lies within most organizations. By connecting these categories and mapping them into a cohesive taxonomy, the paper encourages a holistic and systematic approach to data management. The originality also extends to the practical recommendations provided, which are tailored to help organizations implement governance strategies that maximize the value of all data types. This innovative categorization and the actionable insights proposed make the paper a significant contribution to data management literature, addressing both the risks and opportunities that come with managing diverse data forms in a modern, data-centric environment.

2. Literature Review

2.1 Dark Data

Dark Data Management & Governance. Ajis, et al., (2022) develop a dark data management framework using grounded theory, offering a structured approach to managing dark data by investigating real-world scenarios. Their approach is empirical and context-specific, providing insights into the implementation of systematic frameworks for organizations dealing with large volumes of dark data. By employing grounded theory, they establish a structured methodology that considers organizational needs and technical capacities, underscoring the need for customized frameworks rather than a one-size-fits-all solution. This approach is particularly significant as it highlights the adaptability required in dark data management, depending on organizational structure

and industry-specific challenges. Expanding on this context-specific perspective, Ajis, et al., (2022) focus on small and medium enterprises (SMEs) in Malaysia, exploring the characteristics of dark data in this specific setting. They argue that SMEs, due to their limited resources and capacities, face unique challenges in managing dark data. Their findings indicate that understanding the characteristics of dark data specific to SMEs is crucial for developing effective management strategies. This complements Ajis et al. (2022) by demonstrating that the management framework should be adapted to fit the scale and resources of the business. It also reflects a broader understanding that dark data management needs to be versatile and flexible, acknowledging the diversity of business environments. Dimitrov and Chikalanov (2016) emphasize the governance aspect of dark data, particularly its role in reducing security risks. They argue that an effective governance structure is essential for managing dark data securely, thus preventing data breaches and minimizing risks associated with unmonitored or unutilized data. This perspective aligns with Ajis et al. (2022) but takes a more focused approach by highlighting security as a critical driver for implementing dark data management frameworks. The emphasis on governance suggests that beyond technical solutions, organizations must establish policies and procedures that ensure dark data is managed with a focus on security and compliance. Similarly, Corallo et al. (2021) delve into the manufacturing industry, focusing on how dark data management can be integrated into industrial processes. They argue that in the context of manufacturing, understanding and defining dark data is crucial for optimizing production and operational efficiency. Their study emphasizes that the manufacturing sector, with its reliance on large datasets, must develop industry-specific frameworks to handle dark data. This aligns with Dimitrov and Chikalanov (2016) in underscoring governance but extends the discussion by demonstrating how dark data, if managed correctly, can contribute to improving industry performance. DiMatteo (2021) offers another dimension by exploring the role of cloud services in managing dark data. The study suggests that cloud services can be an effective tool for organizations to manage and process dark data, enabling real-time access and reducing the burden of maintaining in-house data management systems. This technological approach complements the governance and framework-oriented discussions by providing a practical solution that integrates technology into the process of dark data management. Keller-Fröhlich (2022) evaluates the impact of dark data on value-driven data management strategies in the manufacturing sector. Their study builds on Corallo et al. (2021) by arguing that the management of dark data should be strategically integrated into value-driven frameworks to enhance the overall efficiency and output of the manufacturing process. The study suggests that organizations should view dark data not merely as a challenge but as an opportunity to extract value and drive business performance. The concept of dark data, defined as unused or underutilized information accumulated by organizations, is increasingly gaining attention in the realm of data management and analytics.

The articles by Maju and Prakasi (2022), Shetty (2021), Ross (2021), Raca (2021), and others provide a comprehensive understanding of how organizations can integrate, manage, and mitigate the risks associated with dark data. Maju and Prakasi (2022) focus on developing a decision-making model that integrates dark data from hybrid sectors, emphasizing the importance of utilizing this data effectively for strategic and operational decision-making. This model underscores the potential of dark data as an asset, contrasting with traditional views of it as a liability. This aligns with Shetty's (2021) insights from Gartner, which outline strategies for addressing the challenges posed by dark data. Shetty's work serves as a practical guide for organizations, highlighting steps such as data classification, governance, and the application of AI to make dark data more accessible and valuable. The article emphasizes the need for organizations to develop a robust framework to tackle the ever-growing volumes of unstructured and unclassified data. Ross (2021) takes a critical approach in exploring the fear organizations have towards dark data, particularly regarding cybersecurity risks. His article underscores the importance of governance and compliance measures, demonstrating how poorly managed dark data can expose organizations to significant security threats and regulatory penalties. This perspective is echoed in Dimitrov and Chikalanov's (2016) earlier work, which

emphasizes the role of dark data governance in reducing security risks. They present a detailed analysis of the measures organizations can implement to mitigate these risks, such as encryption and strict access controls, reinforcing the need for proactive management strategies. Raca (2021) and Lager (2021) further explore the characteristics and potential uses of dark data. Raca's chapter presents methods and applications for classifying and analyzing dark data, providing an academic and methodological perspective on how to extract value from this data. Lager's (2021) expands on these ideas by detailing the factors influencing dark data utilization and proposing steps for organizations to leverage this untapped resource. The thesis discusses the technological and organizational changes required to harness dark data's potential, such as investing in data infrastructure and developing expertise in data analytics. Ajis's (2023) doctoral dissertation offers a grounded theory analysis of dark data management among Malaysian SMEs, highlighting the unique challenges and opportunities within this context. His findings, echoed in subsequent works by Ajis, et al., (2023), indicate that while dark data can be leveraged for business growth and sustainability, it also poses a risk to small and medium enterprises if not managed properly. Their study provides practical insights into how SMEs can develop strategies to manage dark data, improve data literacy, and ensure long-term sustainability. Another work by Ajis et al. (2024) further elaborates the Malaysian data crisis and its link to dark data management. This study demystifies the factors contributing to dark data accumulation and suggests innovative solutions to transform dark data into an organizational catalyst for growth. Chant (2023) also contributes to the discourse by discussing how organizations can "shine a light" on dark data through knowledge management practices. This aligns with the evolving perspective on dark data as a source of value rather than just a liability, a shift reflected in the broader academic and industry discourse. Chant advocates for integrating knowledge management with data governance practices to unlock the strategic potential of dark data. Banafa (2022) and Ajis, et al., (2024) provide additional insights into understanding dark data's role within organizations. Banafa's publication delves into the fundamentals of dark data, offering a comprehensive explanation of its characteristics and implications. Ajis et al. (2024) build on these foundations, elucidating the theories behind the Malaysian data crisis and proposing frameworks for demystifying and managing dark data effectively.

Dark Data and Business Value. Akbar, et al., (2018) focus on the strategic alignment of IS/IT with business operations to optimize the utilization of dark data. Their study highlights that, while organizations often accumulate vast amounts of dark data, these resources remain underutilized due to the lack of integration between technology and business strategies. The authors argue that aligning IS/IT with business objectives enables organizations to systematically access, process, and analyze dark data, transforming it into a valuable asset. By doing so, businesses can gain insights that inform decision-making, ultimately enhancing operational efficiency and competitive advantage. Gimpel (2021) furthers this discourse by framing dark data as an invisible resource with the potential to enhance organizational performance when properly managed. Gimpel (2021) emphasizes that dark data, despite being abundant within organizations, remains overlooked due to its perceived complexity and unstructured nature. However, the study argues that with appropriate management practices, dark data can provide crucial information that complements existing data, offering a more comprehensive understanding of business processes. Gimpel suggests that companies need to invest in innovative technologies and data management strategies that can identify, categorize, and analyze dark data, allowing them to unlock its value. This perspective aligns with Akbar et al. (2018) by stressing the need for a strategic approach but further underscores the necessity for technological investment and innovation as key enablers of dark data utilization. George, et al., (2023) present a comprehensive framework aimed at unlocking the business value of dark data. Their study builds on the concept that dark data can be a significant asset if organizations establish structured processes for its identification and analysis. The proposed framework integrates business and technological approaches, emphasizing the importance of cross-departmental collaboration and technological adoption. The framework includes steps for identifying, processing, and integrating dark data into

business operations, allowing organizations to uncover patterns, trends, and insights that would otherwise remain hidden. The authors advocate for a holistic approach that combines technology with a supportive organizational culture, where employees are trained and incentivized to engage in data-driven practices. This perspective complements the views of Akbar et al. (2018) and Gimpel (2021) by emphasizing the importance of aligning technology with business strategy while also recognizing the need for a supportive culture and structured framework. Matanović, et al., (2022) take a different approach by examining the application of dark data in a scientific and environmental context, specifically focusing on the Late Miocene freshwater cockles housed in the Croatian Natural History Museum (CNHM). Their study assesses the validity of dark data in contributing to scientific understanding and business value. The authors argue that dark data, even in highly specialized scientific domains, can provide valuable insights when correctly interpreted and contextualized. Although the study is not directly focused on business organizations, it highlights the broader applicability of dark data, showing that the principles of data utilization and integration can transcend sectors. The research demonstrates that with the proper tools and methodologies, dark data can significantly enhance knowledge and decision-making, contributing to organizational or sectoral advancements. This perspective aligns with the business-focused approaches of the other articles by illustrating how systematic data utilization can drive value beyond conventional business settings. Together, these articles provide a comprehensive understanding of how dark data can be leveraged to create business value. They emphasize the importance of aligning IS/IT strategies with business goals, investing in innovative technologies, and developing structured frameworks for data integration. Moreover, they collectively highlight the transformative potential of dark data across diverse contexts, reinforcing the idea that organizations, regardless of their sector, stand to benefit significantly from effectively managing and utilizing dark data. The articles collectively examine the critical role dark data plays in organizations, exploring both the opportunities and challenges associated with its management and utilization. Taulli (2019) provides an overview of dark data in the business context, highlighting the immense volume of unstructured and unused data that organizations accumulate. He discusses the risks, such as increased storage costs and security vulnerabilities, while also emphasizing its potential as a valuable resource for enhancing business intelligence. His practical approach sets the stage for understanding why organizations must act proactively to manage and harness dark data.

Priya, et al., (2022) build on this concept by presenting a balanced perspective on dark data's potential. Their work, suggests that dark data, when processed and analyzed effectively, can uncover significant business insights that might otherwise remain hidden. They emphasize that organizations need to view dark data not merely as a liability but as an asset that, when illuminated, can enhance operational efficiencies, customer relations, and overall decision-making capabilities. This perspective aligns with Munot, et al., (2019) findings. The authors highlight the importance of using advanced data analytics tools and AI to convert dark data into a valuable resource, suggesting that such technologies can uncover hidden patterns and opportunities in otherwise unutilized data. Moumeni et al. (2021) offer a more futuristic and comprehensive view, proposing that dark data is not only a challenge but also a key driver for improving business performance. Their research reviews the current state of dark data management and offers perspectives on its future integration into business operations. They argue that organizations that invest in proper data governance frameworks and technology will be better positioned to transform dark data into actionable business intelligence, driving growth and innovation. Gimpel and Alter (2021) discuss the intersection of dark data and the Internet of Things (IoT), emphasizing how IoT systems generate massive amounts of data that often remain unexamined. They propose that by accessing and analyzing this data, businesses can optimize their operations in real-time and gain a competitive edge. This aligns with the broader narrative that dark data, if managed correctly, holds significant untapped value. Forker (2023) investigates the potential of dark data to predict future firm performance, highlighting its informativeness beyond traditional metrics. His study underscores the predictive value dark data can provide when it is

systematically integrated into business forecasting models. By correlating dark data trends with firm performance metrics, Forker demonstrates that organizations can gain a deeper understanding of market behaviors and anticipate future business outcomes, thus strengthening their strategic planning processes. In the same vein, Faghih, et al., (2021) explore the entrepreneurial implications of dark data, arguing that it is crucial for enhancing entrepreneurial motivation and decision-making. Their work demonstrates that dark data can reveal market trends, consumer behavior patterns, and other insights crucial for startups and entrepreneurs seeking to innovate and grow in competitive markets. Bhatia and Alojail (2022) take an innovative approach to the subject by proposing a new method for extracting value from big data using dark data. Their study develops a framework for deciphering dark data to reveal hidden business value. They suggest that through AI and machine learning algorithms, organizations can not only automate the extraction of insights but also minimize the costs associated with storing and managing large volumes of data. This aligns with the earlier perspectives of Munot et al. (2019) and Moumeni et al. (2021), who also advocate for technological solutions as key enablers for unlocking dark data's potential. Banafa's (2022) exploration of different types of data—big, dark, thick, and small—contextualizes dark data within the broader spectrum of data types that organizations manage. By comparing dark data to other forms of data, he emphasizes that while each type presents unique challenges, dark data, in particular, holds a strategic advantage due to its volume and potential hidden insights. His work underscores the need for a more nuanced approach to data management, one that not only distinguishes between these data types but also integrates their management for maximum organizational benefit.

Dark Data in Cybersecurity and Privacy. Dimitrov, et al., (2018) offer an in-depth examination of various types of dark data and the hidden cybersecurity risks associated with them. Their study highlights how dark data, due to its unstructured and often unmonitored nature, can create vulnerabilities within organizational systems. By categorizing these risks, they make it clear that an organization's lack of visibility over its dark data poses significant security challenges, such as unauthorized access, data breaches, and compliance violations. They argue that to mitigate these risks, organizations need to implement comprehensive data governance frameworks that encompass dark data management, emphasizing the importance of visibility, access control, and regular audits. Building on this perspective, Hobart (2020) frames the dark data issue as a "conundrum" due to the dual-edged nature of this data. While dark data can pose significant security risks, Hobart notes that it also offers substantial untapped value if appropriately managed. Hobart's analysis aligns with Dimitrov et al. (2018) by emphasizing the importance of visibility and management but goes a step further by focusing on the ethical and strategic dimensions of dark data. Organizations are encouraged to not only secure their dark data but also to assess its potential value carefully. Hobart suggests that with the right data management practices, companies could transform this liability into an asset, but the challenges are primarily ethical and technical. This nuanced view introduces a moral imperative: organizations must balance data utilization with privacy considerations and ensure they are not compromising sensitive information while exploring dark data's potential. Jackson and Hodgkinson (2022) expand the discussion of dark data beyond cybersecurity to environmental concerns, arguing that dark data contributes to digital pollution and negatively impacts sustainability efforts. They claim that the energy consumed in storing and maintaining unnecessary or unused data has a significant carbon footprint, underscoring a need for what they call "digital decarbonization." This perspective introduces a new dimension to the discourse, connecting cybersecurity practices with environmental sustainability. While traditional dark data discussions focus on security risks, Jackson and Hodgkinson argue that responsible data management must also include minimizing environmental impact. This approach challenges organizations to adopt sustainable data practices, such as deleting redundant or unnecessary data and utilizing energy-efficient storage solutions. This perspective not only adds depth to the understanding of dark data but also broadens the scope of responsibility for organizations, pushing them to consider both digital security and environmental impact as part of their data management strategies. Teymourlouei and Jackson (2021) take a pragmatic approach by

exploring how managing dark data can not only address cybersecurity challenges but also generate organizational benefits. Their study underscores the potential of dark data as a valuable asset if organizations implement effective data management strategies. This aligns with Hobart's (2020) notion that dark data has untapped potential, but Teymourlouei and Jackson focus specifically on the technological and strategic approaches that can be employed to harness this value while addressing security risks. They advocate for organizations to adopt advanced data management technologies such as artificial intelligence and machine learning to analyze and secure dark data efficiently. These technologies can transform dark data from a risk factor into a source of actionable insights that can drive decision-making and innovation. By integrating these advanced solutions, organizations not only enhance their cybersecurity posture but also unlock new opportunities for growth and efficiency. Collectively, these articles provide a comprehensive analysis of dark data's impact on cybersecurity and privacy, while also expanding the conversation to include environmental and strategic considerations. The authors converge on the idea that effective dark data management is essential not only to mitigate cybersecurity risks but also to unlock potential value and support sustainable practices. They highlight the need for organizations to adopt a proactive, integrated approach that combines technology, governance, and sustainability measures to manage the complexities associated with dark data. The three articles offer insights into the challenges and opportunities surrounding dark data, particularly within the context of emerging technologies like IoT and artificial intelligence. Gianna (2021) addresses the risks associated with dark data in big IoT environments, emphasizing the growing challenge of managing massive, unstructured data generated by IoT devices. Gianna (2021) highlights the importance of risk mitigation strategies, such as implementing robust data governance frameworks and enhancing data security protocols. Gianna (2021) argues that without these measures, organizations risk not only data breaches but also compliance violations, which can have severe financial and reputational consequences. Ge (2022) explores the role of artificial intelligence (AI) and machine learning in managing dark data. Ge (2022) suggests that AI and machine learning algorithms are pivotal in processing and making sense of vast amounts of unstructured information. By leveraging these technologies, organizations can transform dark data into valuable insights, driving efficiency and innovation in various industries. Ge (2022) work underscores the necessity of integrating these advanced technologies to keep pace with the increasing complexity and volume of data. Gautam (2023) offers a unique perspective by investigating the personal safety risks associated with dark data. His study focuses on the implications of dark data for individuals, particularly in the context of data privacy and cybersecurity. Gautam (2023) emphasizes the need for individuals and organizations to be proactive in managing their digital footprints and protecting sensitive information.

Biomedical and Healthcare Dark Data. Almeida, Torres-Espin, et al. (2022) discuss the principles of FAIR data (Findable, Accessible, Interoperable, and Reusable) through the example of the Multicenter Animal Spinal Cord Injury Study (MASCIS). They emphasize the importance of structuring and excavating dark data to align with FAIR principles, which helps make previously isolated or inaccessible data useful for broader scientific communities. Their work shows how dark data, once integrated into a cohesive framework, can enhance the reproducibility and reliability of scientific research. By focusing on animal spinal cord injury studies, they provide a case study that demonstrates the challenges and rewards of converting dark data into FAIR data. This transition not only improves collaboration across research institutions but also ensures that valuable data does not remain underutilized, enabling researchers to build upon previous findings and accelerate the pace of discovery. Ahmed and Verma (2024) explore the role of artificial intelligence in the classification of dark web data using neural networks, a method that is increasingly relevant in the field of biomedical research due to the rise of complex and unstructured datasets. Although their study is specifically focused on dark web data, the application of neural networks as a classification tool is broadly applicable to biomedical data analysis. By using neural networks to classify dark data efficiently, the authors highlight how AI can automate the process of structuring and analyzing large volumes of

unorganized information. Their findings underscore the potential for AI to transform dark data into actionable insights, which is crucial in medical fields where rapid data interpretation can have life-saving implications. This approach resonates with the call for greater technological integration in the healthcare sector, especially in the management of electronic health records and other large datasets that contain hidden, valuable information. Maju and Gnana Prakasi (2022) delve into a specific medical application by utilizing dark data from electronic health records (EHRs) for the early detection of Alzheimer's disease. Their work exemplifies the direct clinical benefits that can emerge from effectively managing dark data. By applying advanced analytical techniques to EHR data, they identify early indicators of Alzheimer's, demonstrating the potential for dark data to improve diagnostic capabilities. This study underscores the value of integrating dark data into medical practice, particularly in preventative care. The use of dark data in this context not only enhances early diagnosis but also highlights the importance of leveraging existing data to develop predictive models, which can ultimately lead to more personalized and effective treatment plans for patients. The study's focus on EHRs is especially pertinent, given the vast amounts of patient information that remain underutilized in healthcare systems globally. Murphy and Thomas (2024) emphasize the importance of addressing and utilizing "negative" data—data that does not show expected outcomes or significant results—in spinal cord medicine. Often overlooked or discarded, such data represent a significant portion of dark data in biomedical research. By illuminating these datasets, the authors argue that "negative" results are crucial for providing a comprehensive understanding of medical phenomena and improving clinical practices. Their study not only validates the importance of data transparency but also advocates for a cultural shift within the medical community to value all forms of data, regardless of the perceived outcome. The findings of this article suggest that integrating negative data into medical records and research archives can prevent the repetition of unsuccessful trials and direct resources toward more promising avenues of study. Together, these articles present a cohesive argument for the integration of dark data into healthcare and medical research, demonstrating that when managed and utilized properly, dark data can significantly advance medical knowledge, improve patient outcomes, and increase the efficiency of healthcare systems. The emphasis on technologies such as AI and the adherence to frameworks like FAIR highlights the evolving landscape of biomedical data management, calling for continued innovation and collaboration across disciplines to fully harness the potential of dark data. The articles collectively explore the intricate domain of "dark data" within medical and biomedical contexts, shedding light on the challenges and opportunities of harnessing such data. Suzen et al. (2023) delve into the issue of "medicine's dark matter," focusing on learning from missing data within medical practices. The authors argue that the medical field holds a significant volume of unutilized or underutilized data, including missing data that could still be informative if appropriately managed. They propose advanced machine learning methods to effectively work with incomplete datasets, highlighting the potential to improve patient outcomes and medical decision-making. This approach illustrates the transformative power of big data and machine learning in medicine, emphasizing that even incomplete or 'dark' data can contribute valuable insights when managed and analyzed systematically. Jha and Singh (2022) explore a broader perspective of human-machine convergence, focusing on its impact on socio-cognitive capabilities. Their article is not strictly confined to the medical domain, their insights into how machines process and analyze data relate closely to the challenges of dark data in medicine. They emphasize the disruption of cognitive processes due to increased reliance on automated data processing, raising concerns about how medical professionals interpret and engage with data, including data that is incomplete or underexplored. Their analysis suggests that while technology offers tools to manage dark data, it also necessitates a rethinking of human-machine interactions to maintain critical cognitive capabilities. Hawkins et al. (2020) specifically address the issue of traumatic brain injury (TBI) and the dark data associated with this medical field. The study discusses the "long tail" of data resulting from inconsistent data collection, storage issues, and variability in data quality. They advocate for improved data dissemination techniques to make this wealth of underutilized data more accessible to researchers and clinicians. This approach underscores the

importance of collaboration and transparency in medical data practices, suggesting that improved access and dissemination could bridge the gap between data collection and actionable insights. Aggarwal and Singh (2020) contribute two significant works exploring visual analytics and knowledge discovery from biomedical dark data. The first article addresses the application of visual analytics tools to manage and interpret biomedical data that remains unexamined due to its complex nature. Their emphasis on visual tools highlights the necessity of simplifying the interaction with complex data sets, making them more accessible for medical professionals who may not have advanced technical skills. The second work focuses on the potential of visual exploration methods for knowledge discovery from biomedical dark data. By leveraging visual and exploratory techniques, they demonstrate how even fragmented or incomplete biomedical data can be transformed into comprehensible, actionable knowledge, offering new avenues for research and patient care.

Machine Learning and AI in Dark Data Analysis. Liu et al. (2021) focus on using deep learning techniques, specifically deep hash-based relevance-aware models, to assess data quality in image dark data. Their approach is significant in that it not only addresses the issue of managing vast amounts of unstructured image data but also incorporates relevance-awareness, ensuring that the data deemed useful is prioritized for analysis. This work highlights the role of machine learning algorithms in effectively filtering and structuring dark data, allowing organizations to extract meaningful patterns from enormous image datasets that would otherwise remain underutilized. By improving the efficiency and accuracy of data quality assessments, their approach sets a foundation for scalable solutions in industries such as healthcare, where image analysis plays a critical role. The integration of relevance-aware mechanisms further enhances the practicality of their model, ensuring that only the most pertinent information is processed, which is crucial for resource-intensive operations. Seki et al. (2023) extend the discussion to the realm of audio data, particularly focusing on text-to-speech synthesis using dark data with evaluation-in-the-loop data selection. This approach emphasizes the dynamic and adaptive nature of ML models in optimizing data usage. Their method incorporates real-time evaluation, allowing for continuous refinement and improvement of the text-to-speech system. This work is notable for its emphasis on the iterative nature of ML models, demonstrating how feedback loops can be utilized to enhance performance continually. By targeting the adaptation of dark data for speech synthesis, Seki et al. (2023) contribute to the growing body of work aimed at improving the quality and efficiency of human-computer interaction technologies. The ability to dynamically select and utilize relevant data ensures that the system remains efficient, particularly when handling large, unstructured datasets—a recurring challenge in AI and ML applications. Ravindranathan et al., (2024) take a broader perspective by exploring how dark analytics, powered by ML and AI, can be leveraged to gain insights and manage risks. Their study examines the application of dark data analysis in various domains, emphasizing the potential of ML models to uncover hidden patterns and correlations that traditional methods may overlook. They argue that by illuminating dark data, organizations can gain a competitive advantage, make informed decisions, and proactively manage risks. This perspective aligns with the practical applications seen in Liu et al. (2021) and Seki et al. (2023), but it expands the discussion to include broader implications for organizational strategy and risk management. By demonstrating the versatility of dark analytics, the authors provide a framework for businesses looking to integrate AI into their data management practices, highlighting how these technologies can transform not just data analysis but also strategic decision-making. Zhou and Song (2021) provide an overarching view by introducing a special issue dedicated to learning-based support for data science applications, situating ML and AI as essential tools in the contemporary landscape of data analysis. Their introduction underscores the growing importance of these technologies in not only managing dark data but also enhancing the entire field of data science. They emphasize that the integration of ML in data science applications is becoming increasingly indispensable for handling the vast amounts of data generated across industries. By framing their discussion in the context of a broader academic and industrial movement, Zhou and Song reinforce the necessity for continual advancements in ML and AI to keep pace with the growing

complexity of dark data. Their work aligns with the other articles in highlighting that as data volumes and complexities increase, the role of AI becomes more critical in extracting valuable insights and maintaining data quality. Slimani et al. (2022) examine the role of automated machine learning (AutoML) in contemporary data science challenges. Their study emphasizes how AutoML represents a pivotal shift in data science, automating complex processes that traditionally required human expertise. They highlight the relevance of AutoML in handling large volumes of unstructured data, including dark data, by automating data cleaning, model selection, and hyperparameter tuning. This automation accelerates the process of turning dark data into valuable insights, demonstrating its crucial role in modern data-driven industries. Singh (2021) explores the novel application of Intuitionistic Plithogenic graphs for dark data analysis. The article introduces an innovative approach that combines neutrosophic logic with graphical models to process and analyze dark data. Singh (2021) argues that such advanced mathematical frameworks are essential for effectively managing the uncertainty and imprecision characteristic of dark data. This approach allows for a more nuanced understanding of datasets that are often incomplete or ambiguous, demonstrating that incorporating intuitionistic and plithogenic methods can significantly enhance the accuracy and depth of dark data analysis. The study bridges theoretical mathematics and practical application, illustrating the evolving methodologies used to address the challenges posed by dark data. Similarly, Singh, et al., (2021) contribute to the field by presenting a comparative analysis of regression-based machine learning techniques. Their work assesses various regression models and their applicability to handling different forms of dark data. They emphasize the importance of selecting appropriate machine learning algorithms based on data characteristics and the desired outcomes. Their comparative analysis provides valuable insights into optimizing regression models for unstructured and partially labeled datasets, underscoring the significance of algorithm selection in maximizing the value extracted from dark data. Shah et al. (2024) focus on the application of machine learning in the context of accident prediction using dark data. The study implements AdaBoost and Random Forest algorithms to enhance predictive accuracy. By leveraging dark data, such as unreported incidents or non-standardized accident reports, the authors demonstrate how these ML algorithms can refine models and yield more precise predictions. This work highlights the real-world implications of dark data analytics, showing how effectively managed and analyzed dark data can play a vital role in public safety and decision-making. Meil (2021) offers a different perspective by focusing on programmatic labeling techniques for dark data in spatial informatics. Meil (2021) discusses how programmatic labeling of dark data can be crucial for training AI models. By using automation to label spatial data that might otherwise remain unclassified, Meil's approach facilitates the transformation of dark data into structured information, making it usable for AI applications in geospatial analysis. This method not only optimizes the use of data but also illustrates the importance of integrating labeling techniques into dark data management.

Environmental and Industrial Dark Data. Sundarraj and Natrajan (2019) propose a sustainable method for managing dark data within smart factories, demonstrating the potential for efficient data usage in industrial settings. The study emphasizes the importance of integrating dark data management practices into the digital ecosystems of smart factories, which are increasingly reliant on large volumes of sensor and machine-generated data. By applying sustainable practices, they argue that smart factories can minimize data wastage, streamline processes, and improve productivity. Their approach aligns with the broader trend of promoting sustainability in industrial operations, showing that efficient dark data management is not only about optimizing performance but also about reducing the environmental impact of unnecessary data storage and processing. This aligns with current industry goals of achieving both economic and environmental efficiencies, highlighting the importance of sustainable practices as integral to modern smart factory operations. Pawlewitz, et al. (2020) expand on the industrial application of dark data by exploring the concept of digital twins in brownfield environments. Their study examines how digital twin technology—a digital replica of physical systems—can be utilized to manage dark data effectively. In brownfield environments,

which involve the redevelopment or improvement of existing industrial sites, dark data often remains underutilized due to the complexity and age of systems in place. By employing digital twin technology, the study shows how organizations can create a comprehensive, real-time view of their operations, allowing them to identify and integrate dark data into decision-making processes. This integration enhances efficiency and operational oversight, particularly in legacy systems that otherwise lack data transparency. The application of digital twin technology is thus presented as a viable method for unlocking the potential of dark data, making it a valuable tool for organizations looking to modernize and optimize their operations within existing infrastructures. Chan, et al., (2020) take a different approach by focusing on the adaptation of a PMU (Phasor Measurement Unit) Time Series Module to reduce dark data in energy production processes, specifically ethanol fuel production. The study highlights how dark data, when left unmanaged, can compromise the accuracy and quality of analytics in energy production systems. By modifying the PMU module, the authors demonstrate that energy production systems can reduce the accumulation of dark data, thereby improving the accuracy of real-time monitoring and control. This approach emphasizes the importance of customizing technology solutions to fit the specific needs of energy industries, ensuring that data is effectively managed and utilized. The adaptation of PMU technology not only improves the quality of analytics but also enhances the overall efficiency of the energy production process, showcasing a practical application of dark data management that can have immediate economic and environmental benefits. Purss et al. (2015) explore the management of dark data from an environmental perspective, focusing on the Australian Landsat archive. Their study examines how dark data from satellite imagery, which had been largely underutilized, can be unlocked and converted into high-performance data infrastructures. By developing a system that integrates this dark data into accessible and usable formats, the authors show how satellite imagery data can be transformed into valuable resources for environmental monitoring and decision-making. This approach illustrates the potential of dark data in contributing to large-scale environmental projects, where previously inaccessible data can provide critical insights for ecological assessments and planning. The transformation of the Landsat archive into a high-performance infrastructure highlights the importance of building systems that not only store but also make data usable and interoperable, enabling long-term benefits for environmental research and monitoring efforts. Da Costa and Barrett (2021) discuss the advancements in cathodic protection monitoring using Industrial Internet of Things (IIoT) and big data technologies. Presented at the NACE CORROSION conference, their work highlights the transformation in monitoring practices within the corrosion industry. They explain how integrating IIoT sensors and big data analytics enhances the accuracy, efficiency, and real-time capabilities of monitoring systems, enabling proactive management of infrastructure. The authors emphasize that these technological advances not only improve data collection and analysis but also optimize maintenance strategies, reducing costs and extending the lifespan of critical assets. This paper underscores the importance of leveraging digital technologies to enhance traditional engineering practices, demonstrating how IIoT and big data can revolutionize infrastructure monitoring and maintenance in various industries, particularly those involving metal structures vulnerable to corrosion.

Dark Data in Scientific Research and FAIR Data. Hand (2020) provides a comprehensive overview of the concept of dark data, discussing its implications in scientific research and emphasizing why it matters. Hand argues that dark data—information that exists but remains unutilized—can often lead to biases, incomplete conclusions, and missed opportunities if not adequately addressed. In his exploration of the scientific community, he emphasizes that the presence of dark data is not just a technical challenge but a fundamental issue that affects the integrity and reliability of research outcomes. Hand's work highlights the need for systematic efforts to identify and integrate dark data into existing research frameworks, ensuring that the insights drawn from scientific investigations are comprehensive and not skewed by unseen information. This perspective lays the groundwork for why FAIR principles are essential in managing dark data, as they offer a systematic approach to

transforming these overlooked resources into valuable assets for scientific progress. Lovato and Zimmerman (2021) expand on Hand's foundational arguments by exploring how dark data can be made FAIR. They discuss the application of FAIR principles to dark data management, emphasizing that simply acknowledging dark data's existence is insufficient; there must be a concerted effort to convert it into usable formats. They advocate for initiatives that aim to make dark data findable and accessible, not only to individual researchers but to the entire scientific community. Their work underscores the importance of collaboration and open data practices, which are essential for ensuring that dark data does not remain siloed or inaccessible. By aligning dark data management with FAIR principles, Lovato and Zimmerman illustrate how making data interoperable and reusable can facilitate cross-disciplinary research, opening new avenues for discovery and innovation. Their approach is particularly relevant in an era where data is generated at unprecedented rates, necessitating frameworks that prioritize transparency, accessibility, and efficiency. Stahlman (2020) provides a case study that applies these ideas within the field of astronomy. Stahlman's work highlights the importance of mining dark data in astronomy through a mixed-methods approach, examining how previously inaccessible data can be transformed into valuable resources for scientific inquiry. By focusing on the "long tail" of astronomy data—information that is often overlooked because it does not fit within conventional data models—Stahlman shows how integrating dark data into broader research frameworks can fill knowledge gaps and provide a more nuanced understanding of astronomical phenomena. His approach demonstrates that dark data, when managed systematically and aligned with FAIR principles, can significantly enhance the scope and depth of scientific exploration. It reinforces the idea that dark data is not just a repository of untapped potential but an essential component of holistic scientific analysis. Upham, Poelen, and their colleagues (2021) bring these concepts into the domain of biological research, specifically focusing on host-virus relationships. Their study demonstrates how biological dark data, when liberated from isolated datasets and integrated into interoperable frameworks, can enhance our understanding of viruses and their hosts, which is critical for fields like epidemiology and conservation biology. By making host-virus data FAIR, the researchers not only facilitate knowledge sharing among biologists but also enable the broader application of this data in predicting and managing disease outbreaks. This study underscores the importance of converting dark data into actionable insights that have real-world implications, reinforcing the need for collaborative frameworks that promote transparency and accessibility. Upham et al. (2021) focus on host-virus relationships, addressing the wealth of unexamined biological data that remains inaccessible and unutilized. The article emphasizes the need for more open access and integration of such data to advance global health initiatives, especially as pandemics and zoonotic diseases continue to emerge. The authors argue that liberating this "biological dark data" is critical for constructing a comprehensive knowledge base, enabling researchers to identify and predict virus-host dynamics effectively. This perspective underscores the urgency of developing infrastructures and collaborative platforms that facilitate the sharing and integration of biological data, which is often siloed and underused. Stahlman, et al., (2018) present another perspective on dark data in astronomy through the Astrolabe Project. Their research focuses on astronomical data that remains uncataloged and inaccessible despite its potential to enhance our understanding of the universe. By developing cyberinfrastructure resources to identify and curate these datasets, they aim to convert this neglected data into valuable assets for the scientific community. The authors highlight the need for innovative technological solutions to curate and manage dark data efficiently, suggesting that without such infrastructure, valuable insights might remain hidden. Their work demonstrates that dark data is not only an issue of availability but also one of organization, curation, and accessibility. Schembera and Durán (2020) delve into the philosophical and technical implications of dark data in big data science, proposing the concept of the "Scientific Data Officer" to manage this growing challenge. Their article emphasizes the increasing complexity and volume of data that goes unused or unnoticed, suggesting that it requires specialized roles to oversee and manage it. They argue that this new professional role would be responsible for data governance, ethical considerations, and the optimization of dark data for

scientific purposes. This perspective highlights the evolving nature of scientific data management and the necessity for multidisciplinary approaches to extract value from vast, underutilized data reserves. Mohr et al. (2008) discuss the dark data issue focusing on data management systems necessary to handle astronomical data efficiently. The paper outlines the development of a robust data management system to process, analyze, and store the extensive observational data collected during the survey. The authors emphasize that effective data management systems are vital in turning dark data into useful information for scientific exploration. Their work provides a practical example of how integrating data management with large-scale observational projects can reduce the accumulation of dark data, ensuring that valuable information is not overlooked. Lastly, Matanović, et al., (2022) investigate the validity of dark data in geology, specifically regarding Late Miocene freshwater cockle specimens at the CNHM. The authors explore the challenges of verifying and utilizing historical geological data that has not been systematically catalogued. They stress the importance of validating such data to enhance its scientific credibility and utility. Their work highlights that, even within geology, dark data can hold significant insights if properly evaluated and integrated into contemporary research frameworks.

Policy and Ethical Implications of Dark Data. Chakrabarty and Joshi (2020) delve into the social dimensions of dark data, examining its role in enhancing "people to people" recovery, particularly in disaster management and crisis response scenarios. They argue that dark data, which includes unstructured information from social media, local communications, and other informal networks, can be harnessed to provide real-time insights and facilitate recovery efforts. The study highlights the potential of dark data to bridge the gap between formal institutional responses and local, on-the-ground realities. By utilizing dark data, organizations and governments can make more informed decisions that are aligned with the needs and circumstances of affected communities. Chakrabarty and Joshi emphasize that the integration of such data is critical for building responsive and resilient systems that are capable of adapting to the dynamic nature of disasters and emergencies. This work underscores the importance of recognizing and leveraging the social value embedded in dark data, which is often overlooked due to its informal or unstructured origins. Giest and Samuels (2020) address the issue of data gaps in the context of big data and dark data, particularly focusing on how these gaps can distort decision-making processes and policy development. They argue that while big data offers unprecedented access to information, the absence of certain datasets or the presence of unstructured dark data can create blind spots that mislead analysis. Their work highlights the need for policymakers and organizations to identify and fill these data gaps by actively incorporating dark data sources, which are often neglected because they do not fit neatly into structured, conventional databases. Giest and Samuels advocate for a more inclusive approach to data collection and analysis, one that accounts for the diversity and complexity of dark data sources. By doing so, organizations can create a more comprehensive and accurate picture of the issues they aim to address, leading to better-informed policies and strategies. Gimpel (2020) explores the organizational potential of dark data, particularly in the context of the Internet of Things (IoT). The study emphasizes that many organizations fail to capitalize on the wealth of data generated by IoT devices, much of which remains unstructured and unutilized as dark data. Gimpel argues that by effectively managing and integrating this data into organizational processes, companies can unlock valuable insights that were previously inaccessible. The work showcases how dark data, when properly harnessed, can enhance operational efficiency, improve decision-making, and create competitive advantages. Gimpel's perspective aligns with a growing recognition in the industry that digital transformation efforts must include robust dark data management practices. This proactive approach enables organizations to maximize the value of their existing data resources, thus turning what was once seen as a burden into an asset. Ingólfsdóttir (2023) addresses the societal implications of dark data, particularly in the context of disinformation and digital transparency. The study highlights how dark data, when disclosed and integrated into public information systems, can act as a powerful tool in the fight against misinformation. Ingólfsdóttir argues that transparency and access to previously hidden or obscured data are crucial for

creating a well-informed public, which is necessary for a healthy democratic process. By shedding light on dark data, institutions can provide a more accurate and comprehensive narrative, countering the spread of false information. This work adds a critical dimension to the discourse on dark data, emphasizing that beyond its technical and organizational value, dark data also plays an essential role in upholding the integrity of information ecosystems and supporting societal resilience against digital threats. Shin and Kwon (2023) provide a review of David J. Hand's book entitled Dark Data: Why What You Don't Know Matters. They highlight Hand's argument that dark data—uncollected, ignored, or misinterpreted data—has a profound impact on decision-making processes. The review emphasizes the need for awareness and strategies to address these hidden data gaps, underscoring how unrecognized dark data can lead to biased outcomes in various fields, from business to healthcare. Shave (2023) emphasizes the importance of lifelong learning, particularly in the context of data management. She discusses how professionals must adapt to emerging challenges, including the growing complexities associated with managing and interpreting dark data. This aligns with the broader call for continuous skill development to stay competent in the digital age, where data governance and understanding are crucial. Choi (2021) investigates the use of dark data to identify policy demands among innovators at Daedeok Innopolis. By analyzing civil complaint data, Choi demonstrates how dark data can reveal hidden patterns and demands, informing better policymaking. This study illustrates the practical application of dark data analytics in understanding and responding to public needs, showing its value beyond mere theory.

Technological Frameworks and Architectures for Dark Data. Ahlawat, Borgman, et al. (2023) focus on the development of a new architecture aimed at managing data costs and complexity. Their approach addresses the financial and operational challenges associated with dark data, particularly the burden of storing and processing large volumes of unstructured information. They propose a system architecture that optimizes data flows, minimizes redundancies, and strategically integrates dark data into existing databases, thus reducing costs. By highlighting the importance of efficiency, the authors make a case for organizations to rethink their data management strategies, ensuring they not only accumulate data but also manage it in ways that enhance value. This study is particularly relevant in the era of big data, where the sheer volume and complexity of information often lead to inflated costs if not managed appropriately. Their architecture offers a solution for companies aiming to harness dark data while maintaining financial and operational control, highlighting the balance between technological advancement and cost efficiency. Roman, Prodan, et al. (2022) expand on the theme of efficient dark data management by focusing on big data pipelines within the computing continuum. They explore how these pipelines can be optimized to access and integrate dark data, bridging the gap between data storage and analysis across distributed computing environments. Their work emphasizes the need for robust and scalable data pipelines that can handle the diverse nature of dark data, ensuring that it is accessible and usable at different points in the computing continuum, from edge to cloud computing. By developing a comprehensive approach to managing data pipelines, the authors highlight how organizations can transform dark data into actionable insights in real-time, ultimately improving decision-making and enhancing the agility of business operations. This perspective aligns with the growing trend of integrating AI and machine learning capabilities into data management systems, demonstrating that efficient pipelines are essential for organizations seeking to maximize the potential of dark data while maintaining flexibility and scalability. Benvenuti (2023) proposes a framework for data pipeline discovery, emphasizing the importance of identifying and understanding the components that make up data pipelines to manage dark data effectively. The framework seeks to streamline the process of integrating and analyzing dark data by making the architecture of data pipelines more transparent and adaptable. By identifying key stages in the data flow and pinpointing where dark data can be integrated, the framework provides a structured method for transforming raw, unstructured data into valuable information that can be used in decision-making processes. This study contributes to the discourse on dark data management by focusing on the foundational aspect of data architecture, offering a method that organizations can use to design and

optimize their data systems. By emphasizing the discovery and refinement of data pipelines, Benvenuti's framework aligns with the need for adaptive, modular systems that can respond to evolving data landscapes. Zhong, et al. (2024) extend the conversation into the industrial domain by addressing the construction of a sustainable knowledge management system that incorporates dark data for industrial maintenance. Their study highlights the value of dark data in predictive maintenance and operational efficiency within industrial environments. By integrating dark data into knowledge management systems, they argue that companies can optimize maintenance schedules, reduce downtime, and increase the longevity of industrial assets. Their approach underscores the need for sustainable and integrated solutions that not only store dark data but actively utilize it to enhance operational processes. This study reinforces the importance of sustainability in data management, particularly in industries where efficiency and cost-effectiveness are crucial. It also demonstrates how dark data, when properly harnessed, can serve as a strategic asset that supports long-term industrial goals. The selected articles explore diverse approaches and methodologies for managing and utilizing dark data across various domains, including high-performance computing (HPC), cyber-physical systems, blockchain, and image processing. Schembera (2021) emphasizes the importance of metadata annotation in HPC applications. The article discusses the vast amounts of data generated by HPC systems, much of which remains unstructured or underutilized. Schembera argues that effective metadata management is crucial for transforming this dark data into valuable information. By enhancing data cataloging and accessibility through metadata, HPC applications can maximize their computational resources and improve the efficiency of their data analysis processes. This focus on metadata underscores the need for a systematic approach to managing and leveraging dark data. Nguyen et al. (2022) address the challenges of dark data in cyber-physical systems (CPS) through a combination of software engineering and AI. Their report highlights the SEA4DQ'21 workshop discussions on data quality in CPS. The authors stress the importance of developing AI-driven frameworks to enhance data quality, reduce inconsistencies, and process the vast amounts of unstructured data produced in CPS environments. This approach shows that AI has the potential to revolutionize how dark data is managed and optimized for real-time applications in CPS, enhancing system performance and decision-making capabilities. Neha and Pahwa (2020) propose using blockchain technology for dark data analytics, as presented in their conference proceeding. They argue that blockchain's immutable and transparent nature makes it ideal for managing and tracking dark data. By integrating blockchain, organizations can maintain data integrity, track its provenance, and improve data transparency, which is crucial for industries that rely heavily on data authenticity and traceability. Their approach demonstrates how blockchain can be an innovative solution for enhancing the reliability and accessibility of dark data. Maju and Prakasi (2022) present a decision-making model for integrating dark data from hybrid sectors. Their work offers a model that utilizes analytics to incorporate dark data into organizational decision-making processes. They emphasize that organizations across various sectors accumulate dark data without fully utilizing its potential. By employing their decision-making model, businesses can integrate this data, gaining insights that improve operational efficiencies and strategic planning. This model highlights the growing recognition of dark data's value as a strategic asset in various sectors. Liu et al. (2019) focus on a framework for assessing image-based dark data. Presented at the APWeb-WAIM 2019 conference, their research provides a systematic method for evaluating the quality of image dark data using big data and AI tools. Their approach is crucial for industries reliant on visual data, as it allows organizations to extract useful information from underutilized image repositories, demonstrating the importance of specialized frameworks for managing dark data in different contexts. Chaudhari and Pund (2020) explore techniques for visualizing uncertainties in dark data. They argue that visualization methods are essential for interpreting ambiguous or noisy data, which is common in dark data scenarios. Their work highlights that visual tools can bridge the gap between complex data sets and actionable insights, making it easier for organizations to extract value from dark data. Lastly, Benvenuti (2023) emphasizes the need for efficient data pipelines to optimize dark data management. Presented at the 2023 International Conference on Management of Data, the study suggests

frameworks for discovering and managing data pipelines, ensuring that dark data is accessible and effectively integrated into data workflows. This approach underscores the need for proactive strategies in data management, promoting connectivity and accessibility throughout the data lifecycle.

Macro-Theme	Articles
Dark Data Management & Governance	Ajis, (2023); Ajis et al., (2023); Ajis et al., (2024) ; Ajis et al., (2022); Banafa (2022); Chant (2023); Corallo, et al. (2021). DiMatteo (2021); Dimitrov and Chikalanov (2016); Keller-Fröhlich (2022); Lager (2021); Maju and Prakasi (2022); Raca, (2021); Ross (2021); Shetty (2021).
Dark Data and Business Value	Akbar et al., (2018); George et al. (2023); Gimpel, (2021); Matanović et al., (2022); Taulli, (2019).
Dark Data in Cybersecurity and Privacy	Teymourlouei and Jackson, (2021); Jackson and Hodgkinson, (2022); Hobart, (2020); Gianna, (2021); Ge, (2022); Gautam, (2023); Dimitrov, (2018).
Biomedical and Healthcare Dark Data	Taulli, (2019); Suzen, et al. (2023); Murphy and Thomas, (2024); Matanović e,t al., (2022); Maju and Gnana Prakasi (2022); Jha and Singh, (2022); Hawkins et al., (2020). George and Sujatha, (2023); Gimpel, (2021); Choi, (2021); Almeida, et al. (2022); Ahmed and Verma, (2024); Aggarwal and Singh (2020); Aggarwal and Singh, (2020).
Machine Learning and AI in Dark Data Analysis	Liu, et al. (2021); Seki, et al. (2023); Ravindranathan, et al. (2024); Zhou and Song, (2021); Slimani, et al. (2022); Singh, (2021); Singh et al., (2021); Shah, et al. (2024); Meil, (2021).
Environmental and Industrial Dark Data	Sundarraj and Natrajan, (2019); Pawlewitz, et al. (2020); Chan et al., (2020); Purss, et al. (2015); Da Costa and Barrett (2021).
Dark Data in Scientific Research and FAIR Data	Hand (2020); Lovato and Zimmerman, (2021); Stahlman (2020); Upham et al., (2021); Stahlman et al., (2018); Schembera and Durán (2020); Mohr, et al. (2008); Matanović et al., (2022).
Policy and Ethical Implications of Dark Data	Chakrabarty and Joshi, (2020); Giest and Samuels, (2020); Gimpel (2020); Ingólfsdóttir (2023); Shin and Kwon (2023); Shave (2023); Choi (2021).
Technological Frameworks and Architectures for Dark Data	Ahlawat, et al. (2023); Roman, et al. (2022); Benvenuti, (2023); Zhong, et al. (2024); Schembera (2021); Nguyen, et al. (2022); Neha and Pahwa, (2020); Maju, and Prakasi, (2022); Liu, et al. (2019); Chaudhari and Pund, (2020); Benvenuti, (2023).

2.2 Grey Data

Fundamental and Theoretical Advancements in Grey System Theory. Liu, et al., (2017) explore the foundational aspects of grey data analysis, focusing on its ability to model and interpret incomplete or ambiguous data in a structured manner. They provide insights into the development of GST methodologies and their practical implementation, emphasizing the theory's relevance in fields where data is often imprecise, such as engineering, economics, and environmental science. This work serves as a crucial entry point for understanding how grey data analysis operates, offering a comprehensive explanation of its core principles and methods. Building on this foundation, Liu et al. (2020) delve into spectral analysis within grey system theory and introduce new directions for research. They expand the scope of GST by integrating spectral analysis, demonstrating its capability to process and analyze time-series data more effectively. This integration enhances the predictive power of GST, providing a robust approach to dealing with temporal uncertainties in complex systems. Furthermore, the authors propose novel methodologies and applications, showcasing GST's potential in addressing emerging challenges. By pushing the boundaries of traditional grey analysis, Liu and colleagues position GST as a dynamic and evolving field with promising prospects for future research. In contrast, Yang, et al., (2019) focus on uncertainty within grey data analytics, a core issue GST seeks to address. They argue that the power of GST lies in its ability to quantify and manage uncertainty, distinguishing it from other data analysis methods that may fall short in scenarios where information is incomplete or unclear. Their work offers a detailed examination of uncertainty modeling techniques, providing both theoretical insights and practical applications. By emphasizing the need for precise models to handle indeterminate information, this article advances the conversation around uncertainty, solidifying GST's role as a critical tool in modern data analytics. Ng (1994), one of the earlier works in this collection, discusses the fundamentals of grey systems and the grey relational

model (GRM), which are crucial components of GST. GRM serves as a means to measure relationships between uncertain factors within a system, enabling decision-making processes even when information is incomplete or lacking. This foundational work laid the groundwork for subsequent advancements in grey systems, including the development of models and techniques for dealing with various types of uncertainty in data. Ng's contribution is particularly important for understanding the origins and early theoretical underpinnings of GST. Finally, Shimizu, et al., (1998) present an introduction to time series data analysis using grey system theory. This article highlights the practical applications of GST in time-series data, a common scenario where data is prone to uncertainty due to temporal variations. The authors provide examples of how GST can be applied to time series data, demonstrating its effectiveness in forecasting and trend analysis. By focusing on practical applications, this article illustrates GST's utility beyond theoretical constructs, showing its adaptability in real-world scenarios.

Applications in Clustering and Data Mining. Chang and Yeh (2005) focus on the application of GRA for data clustering, presenting it as an effective technique for grouping similar data points even in situations where data is incomplete or ambiguous. The authors explain how GRA is particularly useful in clustering analysis due to its ability to quantify the degree of similarity or relevance among multiple attributes, even when the data exhibits uncertainty. This foundational work establishes the utility of GRA as a reliable tool in clustering tasks, setting the stage for further development in this field. By leveraging the core principles of grey system theory, Chang and Yeh demonstrate how GRA can provide more accurate and efficient clustering results compared to traditional methods, especially when dealing with incomplete or noisy data. Building on this, Li, et al., (2015) introduce an improved version of GRA for panel data clustering. Their approach addresses limitations in traditional GRA, such as handling high-dimensional data and temporal dependencies inherent in panel datasets. By enhancing the GRA framework, the authors create a method that can better capture complex relationships in panel data, thus improving clustering accuracy and performance. This paper marks a significant advancement in the field, demonstrating the adaptability of GRA when combined with modifications tailored to specific data types. This improved methodology enhances the robustness of clustering analysis, particularly in socioeconomic and environmental studies where panel data is common. Zhang and Zhou (2015) take a different approach by integrating the Grey Wolf Optimizer (GWO) with clustering analysis, demonstrating the innovative ways in which grey systems theory can be hybridized with optimization algorithms. Their study introduces a bio-inspired technique for clustering that mimics the hunting behavior of grey wolves, effectively optimizing cluster formation. The integration of GWO with GRA enhances the clustering process by refining the search for optimal clusters, ensuring that solutions are not only efficient but also globally optimal. This fusion highlights the flexibility of GRA when combined with other algorithms, making it a powerful tool for complex clustering problems that demand precision and adaptability. Huang and Sun (2016) shift focus to data imputation, applying a GRA-based k-nearest neighbor (k-NN) approach for missing data imputation. Their method extends the use of GRA beyond clustering, demonstrating its versatility in other critical data preprocessing tasks. By using GRA to identify relationships between data points with missing values and their closest neighbors, Huang and Sun's approach improves the imputation accuracy, which is essential for ensuring the reliability of data-driven analyses. This work emphasizes the broad applicability of GRA, showcasing how it can be adapted and combined with other machine learning techniques for robust data handling solutions. Lastly, Zhang and Li (2006) apply GRA for gene selection in microarray data, illustrating the method's potential in the field of bioinformatics. In microarray studies, selecting relevant genes is crucial for disease classification and understanding biological processes. The authors demonstrate that GRA can effectively identify the most significant genes from high-dimensional microarray data by measuring the relevance between gene expression profiles. This application highlights GRA's ability to perform feature selection tasks in scenarios where data is not only high-dimensional but also highly complex. By focusing on gene selection,

Zhang and Li extend the application of GRA to the domain of health sciences, showcasing its utility in advanced biomedical research.

Optimization and Decision-Making Models. Al-Refaie (2010) investigates the integration of grey data envelopment analysis (DEA) within the Taguchi method framework. This study applies GRA to enhance the Taguchi method, which is traditionally used for optimizing quality and performance in manufacturing and engineering processes. By incorporating grey DEA, Al-Refaie addresses the limitations of conventional Taguchi methods in situations where data is incomplete or ambiguous. The grey DEA technique is particularly valuable for identifying and quantifying relationships between input factors and outputs under uncertain conditions, enabling more precise optimization. This work highlights the potential of integrating GRA with established optimization frameworks to enhance their capabilities, especially in quality engineering where data uncertainty is a critical challenge. Tzeng et al. (2009) similarly combine GRA with the Taguchi method but focus on optimization applications across different engineering fields. Their study showcases how the integration of GRA helps quantify the relationships between control factors and performance measures in a systematic manner, leading to more accurate optimization outcomes. The authors demonstrate that the hybrid approach enhances the Taguchi method's robustness, making it more effective in managing the influence of noise factors and variability in experiments. This integration of GRA ensures that even when complete information is not available, the optimization process remains reliable, supporting its use in real-world engineering applications where data conditions may not always be ideal. Wang, et al., (2007) apply GRA and DEA to evaluate hospital efficiency in China. This study underscores the versatility of GRA in performance evaluation, particularly in the healthcare sector. By combining GRA with DEA, the authors develop a model that can handle the complexity and uncertainty of hospital data, which often involves multiple and interrelated performance indicators such as service quality, patient satisfaction, and resource utilization. The integration of GRA allows for a more nuanced understanding of efficiency and productivity, especially when dealing with incomplete or vague information. This research exemplifies the use of GRA as a valuable tool for decision-making and performance assessment in complex organizational systems like hospitals, where standard evaluation methods might be inadequate. Goyal and Grover (2012) extend the application of GRA into advanced manufacturing systems, incorporating fuzzy logic to rank these systems. By utilizing fuzzy GRA, the authors enhance the method's ability to manage ambiguity and subjectivity, which are common in evaluating advanced manufacturing environments characterized by multiple conflicting criteria. The fuzzy GRA approach provides a systematic and precise ranking of manufacturing systems, allowing decision-makers to identify the best options under uncertain conditions. This study demonstrates the adaptability of GRA when fused with fuzzy logic, making it suitable for environments that require both quantitative and qualitative evaluations. Finally, Zeng et al. (2007) employ hierarchical GRA for optimizing wastewater treatment processes. Their work focuses on the environmental management sector, showing how GRA can optimize treatment techniques by evaluating multiple, often conflicting, environmental and operational parameters. The hierarchical GRA method provides a structured approach to assessing and prioritizing various treatment options based on performance, cost, and environmental impact, even when the data is incomplete or uncertain. This study illustrates how GRA can be adapted for environmental applications, helping policymakers and engineers make more informed and effective decisions in the context of wastewater management.

Forecasting and Prediction Models. Hsu and Wang (2007) explore the use of improved grey models (IGM) for forecasting in the integrated circuit (IC) industry, a sector where rapid technological advancements and market fluctuations create highly volatile conditions. They propose enhancements to traditional grey models to improve forecasting accuracy for IC production and demand. By refining the model to better capture the dynamic nature of the industry, they illustrate how grey models can be adapted to provide more accurate predictions in high-tech environments. This research

demonstrates the potential of grey models in business forecasting, especially when precise historical data is scarce or unreliable, thus making it a valuable tool for industries characterized by uncertainty and rapid change. Rajesh (2024) expands the application of grey models to decision-making during pandemics, illustrating the relevance of GST in public health and crisis management. By utilizing grey models, the study aims to predict pandemic trajectories, evaluate intervention effectiveness, and aid policymakers in making informed decisions under conditions of limited and evolving data. This research is particularly timely, as it aligns with the need for adaptable and reliable models that can handle the unpredictability and data scarcity typical in pandemic scenarios. The study highlights how grey models can serve as a critical tool for real-time analysis, providing valuable insights to optimize responses during health emergencies and other crisis situations. Hu (2020) applies grey prediction models to bankruptcy forecasting, addressing the financial sector's need for effective risk management tools. By employing grey models, Hu illustrates how firms' financial health and potential bankruptcy risks can be assessed even when historical financial data is incomplete or limited. The model's ability to generate accurate predictions under these conditions makes it a useful asset for stakeholders and policymakers in finance. This application shows the utility of grey models in economic forecasting and risk analysis, proving that GST can be a powerful tool in predicting financial outcomes when data availability is constrained. Huang, et al., (2019) focus on carbon emissions forecasting using a hybrid model that combines GRA with Long Short-Term Memory (LSTM) neural networks. By integrating GRA's capacity to measure relationships and LSTM's proficiency in handling sequential data, the authors create a robust forecasting model for carbon emissions. This hybrid approach effectively leverages the strengths of both methods, demonstrating how grey models can be combined with modern machine learning techniques to enhance predictive accuracy. This study underscores the versatility of GRA when adapted to address environmental challenges, emphasizing its relevance in sustainability forecasting and carbon management strategies. Lastly, Tsaur (2008) proposes a fuzzy grey regression model for forecasting limited time series data. This approach integrates fuzzy logic with grey models to manage the uncertainty and imprecision inherent in short or incomplete datasets. By applying this hybrid method, Tsaur shows how grey regression can enhance time series analysis when traditional statistical models fall short due to limited data. This innovation is particularly useful in economic forecasting and other fields where time series data is often incomplete or scarce. The study highlights the flexibility of grey models, demonstrating their capacity to adapt and improve through integration with other mathematical frameworks like fuzzy logic.

Risk Analysis and Evaluation Studies. Baghery, et al., (2018) apply GRA to prioritize risks in the auto parts manufacturing industry. Their study focuses on identifying, evaluating, and ranking the potential risks associated with production processes, supply chain management, and operational activities. By using GRA, they are able to systematically quantify and compare the severity and probability of different risks, even when data is incomplete or uncertain. This allows for a more objective assessment, helping managers allocate resources effectively to mitigate the most critical threats. The study underscores GRA's utility in complex manufacturing environments where decision-making is often challenged by dynamic and uncertain conditions. It highlights the method's effectiveness in providing clear, actionable insights for risk management, ensuring efficiency and safety in manufacturing operations. Feng et al. (2019) extend the application of GRA to assess security risks in small reservoirs. This study focuses on evaluating the factors that pose threats to the safety and integrity of reservoirs, such as structural issues, natural hazards, and operational vulnerabilities. By utilizing GRA, the researchers are able to rank the identified risk factors according to their impact and likelihood, offering a systematic approach for prioritizing interventions. The application of GRA in this context demonstrates its capability to manage uncertainty in environmental and infrastructural systems, where data may be limited or variable. Feng and colleagues' work showcases how GRA can be effectively used in the environmental management sector to develop more robust risk mitigation strategies, ensuring the safety and sustainability of critical infrastructure like reservoirs. Fu et al.

(2001) investigate the use of GRA in corrosion failure analysis. This study is particularly significant in the field of material science, where understanding and preventing corrosion is critical to maintaining the longevity and safety of equipment and structures. By applying GRA, the authors assess various factors contributing to corrosion failure, such as material properties, environmental conditions, and operational parameters. GRA allows them to prioritize these factors based on their influence on corrosion, providing a comprehensive framework for developing effective prevention strategies. This application highlights the adaptability of GRA in technical fields, where precise and reliable analysis of multiple interacting variables is essential. The study emphasizes the importance of using GRA to optimize failure analysis and material selection processes, ensuring reliability in industries that depend on material durability, such as construction and manufacturing. Chang and Lin (1999) apply GRA to analyze CO2 emissions in Taiwan, focusing on the environmental impact of various industries and sectors. The study explores the relationship between economic activities and emissions, using GRA to determine which sectors contribute most significantly to CO2 emissions. This approach allows for an objective comparison, helping policymakers and researchers understand the key drivers of emissions in Taiwan's economy. By identifying the most influential factors, the study provides a basis for developing targeted policies and strategies aimed at reducing emissions and mitigating climate change impacts. This research illustrates how GRA can be effectively applied in environmental policy analysis, where understanding complex interactions between economic activities and environmental outcomes is crucial. It highlights GRA's role in facilitating data-driven policy-making, ensuring that interventions are both efficient and effective.

Applications in Medical and Healthcare Fields. Xuerui and Yuguang (2004) explore the application of GRA in medical data analysis, focusing on its use for analyzing complex medical datasets where information may be incomplete or inconsistent. The study demonstrates how GRA can effectively identify relationships among various medical variables, helping to uncover patterns that might be difficult to detect using traditional statistical methods. The authors emphasize that GRA's ability to work with uncertain or ambiguous data makes it a suitable method for analyzing clinical information and supporting medical decision-making. This approach is particularly valuable in scenarios where precise and complete medical records may not be available, such as in early diagnostic stages or in settings with limited resources. The study underlines GRA's potential in medical research, paving the way for more effective diagnostic models and patient monitoring systems. Javed and Liu (2018) apply GRA to evaluate outpatient satisfaction in healthcare projects, emphasizing the importance of patient feedback in improving healthcare services. Their study uses GRA to systematically analyze patient satisfaction data, identifying the most influential factors that affect outpatient experiences. By quantifying the relationships between various service attributes—such as waiting time, service quality, and facility environment—and overall patient satisfaction, the authors are able to rank these factors in order of importance. This prioritization provides healthcare managers with clear insights into which areas need improvement to enhance patient satisfaction. The use of GRA in this context highlights its effectiveness in handling subjective and often qualitative data, transforming it into actionable insights that can guide service improvement initiatives in healthcare settings. Building upon their earlier work, Javed et al. (2019) extend the application of GRA to evaluate healthcare service quality in Pakistan. The study incorporates various dimensions of healthcare service quality—such as reliability, responsiveness, and empathy—and uses GRA to rank these dimensions based on patient perceptions. The authors demonstrate how GRA can be applied in the healthcare sector to provide a comprehensive evaluation of service quality, even in developing countries where data limitations and resource constraints are common challenges. The study's findings offer valuable insights for healthcare administrators in Pakistan, guiding them to prioritize improvements in specific service areas. The research underscores GRA's role as a powerful tool for healthcare management, helping to enhance service quality and patient satisfaction, particularly in regions where healthcare systems are still evolving and improving. Lin (2008) presents another innovative application of GRA, using it for electrocardiogram (ECG) beat discrimination with GRA-based classifiers. The study

demonstrates how GRA can be employed in diagnostic processes, specifically in distinguishing between different types of ECG beats, which is crucial for detecting arrhythmias and other heart conditions. By applying GRA to ECG data, Lin develops a classifier that can accurately differentiate between normal and abnormal heartbeats, thus aiding early diagnosis and treatment planning. The study emphasizes the suitability of GRA for medical diagnostics, particularly in scenarios involving time-sensitive and incomplete data. The ability of GRA to process such data in a precise manner makes it an effective tool for developing automated diagnostic systems that can support clinicians in real-time patient monitoring and decision-making.

Energy and Environmental Applications. Guo et al. (2020) focus on energy conservation efforts in underdeveloped regions of China. The study applies GRA to evaluate and prioritize different factors influencing energy conservation strategies. Given the socio-economic disparities in these regions, developing efficient and sustainable energy policies is crucial. By using GRA, the authors are able to identify key variables—such as economic support, infrastructure development, and technological implementation—that significantly impact energy efficiency outcomes. This prioritization allows policymakers to allocate resources effectively and tailor their strategies to the specific needs of these regions. The study underscores the utility of GRA in sustainable development and policy-making, particularly in underdeveloped areas where data limitations are common. Xia et al. (2022) explore the use of GRA for detecting electricity theft, a significant problem in the energy sector that leads to financial losses and operational inefficiencies. The study employs GRA to identify irregular patterns in electricity consumption, helping to distinguish between legitimate usage and theft. By quantifying the relationships between various indicators, such as abnormal load patterns and meter tampering signals, GRA provides a systematic approach for evaluating and detecting anomalies. This application demonstrates GRA's effectiveness in enhancing the accuracy of monitoring systems, reducing revenue losses, and ensuring the stability of power supply networks. The study highlights GRA's potential in optimizing energy management, showcasing how it can be adapted for real-time monitoring and fraud detection. Huang and Wang (2016) investigate CO2 emissions in China using GRA, focusing on the relationship between economic activities, energy consumption, and carbon emissions. The study applies GRA to analyze how different sectors contribute to emissions levels and to prioritize strategies for reducing carbon output. By comparing the impact of various economic activities on CO2 emissions, the researchers provide insights that help policymakers design targeted emission reduction policies. This study emphasizes the value of GRA in environmental policy and management, as it allows for a nuanced understanding of the complex interactions between economic growth and environmental sustainability. The research shows how GRA can facilitate data-driven decision-making in addressing climate change, especially in rapidly developing economies like China. Yin, et al., (2017) apply GRA in the context of storm-tide disaster loss analysis, focusing on quantifying and ranking the factors contributing to disaster severity. GRA is used to analyze various variables such as geographical characteristics, economic exposure, and meteorological data. By prioritizing these factors, the study provides crucial information for disaster preparedness and response planning, enabling authorities to allocate resources and implement measures where they are most needed. This application of GRA demonstrates its capacity to manage and analyze complex environmental data, making it a valuable tool in disaster risk management and mitigation planning. Wang et al. (2016) use GRA to evaluate productivity in Vietnamese agroforestry systems, a sector crucial to the country's rural economy and food security. The study applies GRA to assess different factors influencing agroforestry productivity, such as soil quality, climate conditions, and farming techniques. By ranking these factors, the authors provide insights into optimizing agroforestry practices to enhance productivity. The use of GRA in this study highlights its relevance in agricultural management, especially in developing countries where maximizing productivity and ensuring food security are paramount. The method offers a structured approach to understanding the complexities of agroforestry systems, helping policymakers and farmers make informed decisions that enhance efficiency and sustainability

Macro-Theme	References
Fundamental and Theoretical Advancements in Grey System Theory	Liu et al., (2017); Liu, et al., (2020); Yang, et al., (2019); Ng, (1994); Shimizu, et al., (1998).
Applications in Clustering and Data Mining	Chang, et al., (2005); Li, et al., (2015); Zhang, and Zhou, (2015); Huang and Sun, (2016), Zhang and Li, (2006)
Optimization and Decision-Making Models	Al-Refaie, (2010); Tzeng, et al., (2009); Wang, et al., (2007); Goyal and Grover, (2012); Zeng et al. (2007).
Forecasting and Prediction Models	Hsu, and Wang, (2007); Rajesh, (2024); Hu, (2020); Huang, et al., (2019); Tsaur, (2008).
Risk Analysis and Evaluation Studies	Baghery, et al., (2018); Feng, et al., (2019); Fu, et al., (2001); Chang, and Lin, (1999).
Applications in Medical and Healthcare Fields	Xuerui, and Yuguang (2004); Javed and Liu (2018); Javed, et al., (2019); Lin (2008).
Energy and Environmental Applications	Guo, et al., (2020); Xia, et al., (2022); Huang, and Wang, (2016); Yin, et al., (2017); Wang, et al.,. (2016).

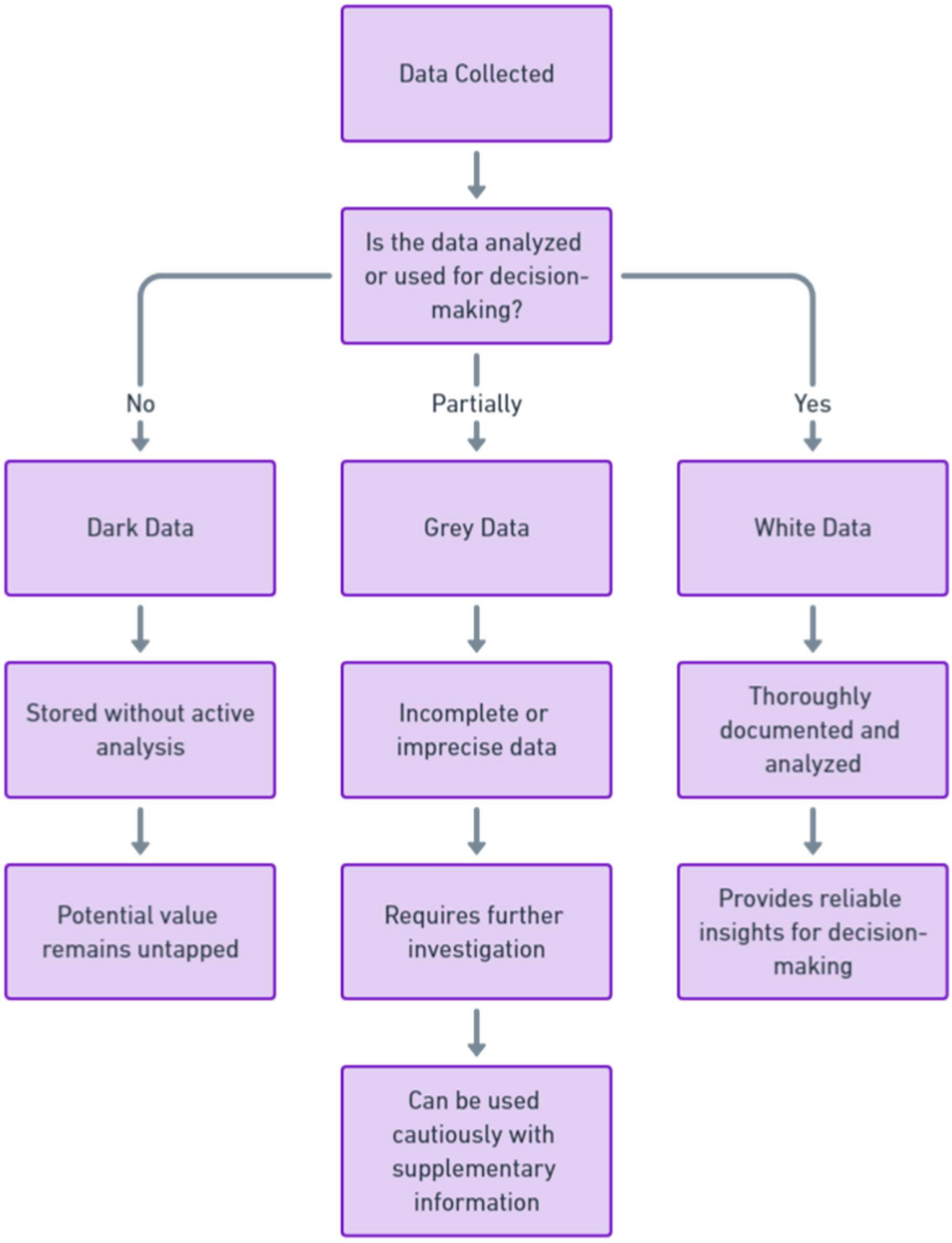
3. Taxonomy and characteristics of the White, Grey and Dark Data

Definition. The definition of data can be categorized into three distinct types: dark data, grey data, and white data, each with its own characteristics and implications for decision-making.

Dark data refers to information that is collected but remains unused for decision-making or analysis. This data is often stored in archives or databases without being actively analyzed, which means its potential value is untapped. It could include information like server logs, email archives, or unused survey responses that organizations might overlook. Though it may hold valuable insights, its underutilization makes it a missed opportunity for gaining strategic advantage or improving processes.

Grey data, on the other hand, represents information that is incomplete, imprecise, or uncertain, necessitating further investigation or refinement. This type of data is partially analyzed but lacks the clarity or accuracy needed to be fully reliable on its own. It may include preliminary results from surveys, estimates, or incomplete customer feedback forms. Organizations can use grey data for decision-making, but they must exercise caution due to its inherent uncertainties and may need to supplement it with more information or advanced analytical methods.

White data is the most actionable and valuable type. It is fully accessible, thoroughly documented, and actively used for decision-making and analysis. White data provides a clear, reliable foundation for generating insights and guiding strategic actions. Examples include validated customer records, financial reports, and thoroughly completed survey results. This type of data is structured, consistent, and maintained with a high degree of governance, ensuring its accuracy and relevance.



Data visibility. Data visibility refers to the ability of an organization to access, monitor, and understand its data across various systems, ensuring transparency, accuracy, and accessibility for effective decision-making, compliance, and optimization of business operations.

Dark data is information that remains hidden or overlooked within databases. Often, this data is collected passively or archived without a clear intention for its use, leading to its neglect. It may include system logs, email archives, and other unstructured data that, while potentially valuable,

remains unused and unnoticed. The lack of visibility means that organizations may miss out on insights that could be derived from this data if it were properly processed and analyzed. Without efforts to bring this data to light, it remains a lost opportunity, leading to storage costs without corresponding benefits.

Grey data sits in between; it is partially visible, but its quality or completeness is often uncertain. This type of data may be recognized and accessible within databases, but it might not be in a readily usable form or may lack full documentation. Examples include incomplete survey responses, estimated metrics, or preliminary research data. While grey data can be utilized in decision-making, it usually requires additional validation, refinement, or supplementation with other sources. The partial visibility of grey data means that organizations need to invest resources to fill in gaps and assess its reliability before it can be fully integrated into analytical processes.

White data represents the most visible and accessible form of information. It is fully transparent, well-documented, and available to stakeholders. White data is often structured and maintained in an organized manner, allowing for efficient access and integration into decision-making and analysis. Examples include financial records, validated customer databases, and other well-governed datasets. Its complete visibility ensures that it plays a crucial role in supporting operational and strategic decisions, maximizing its value for the organization.

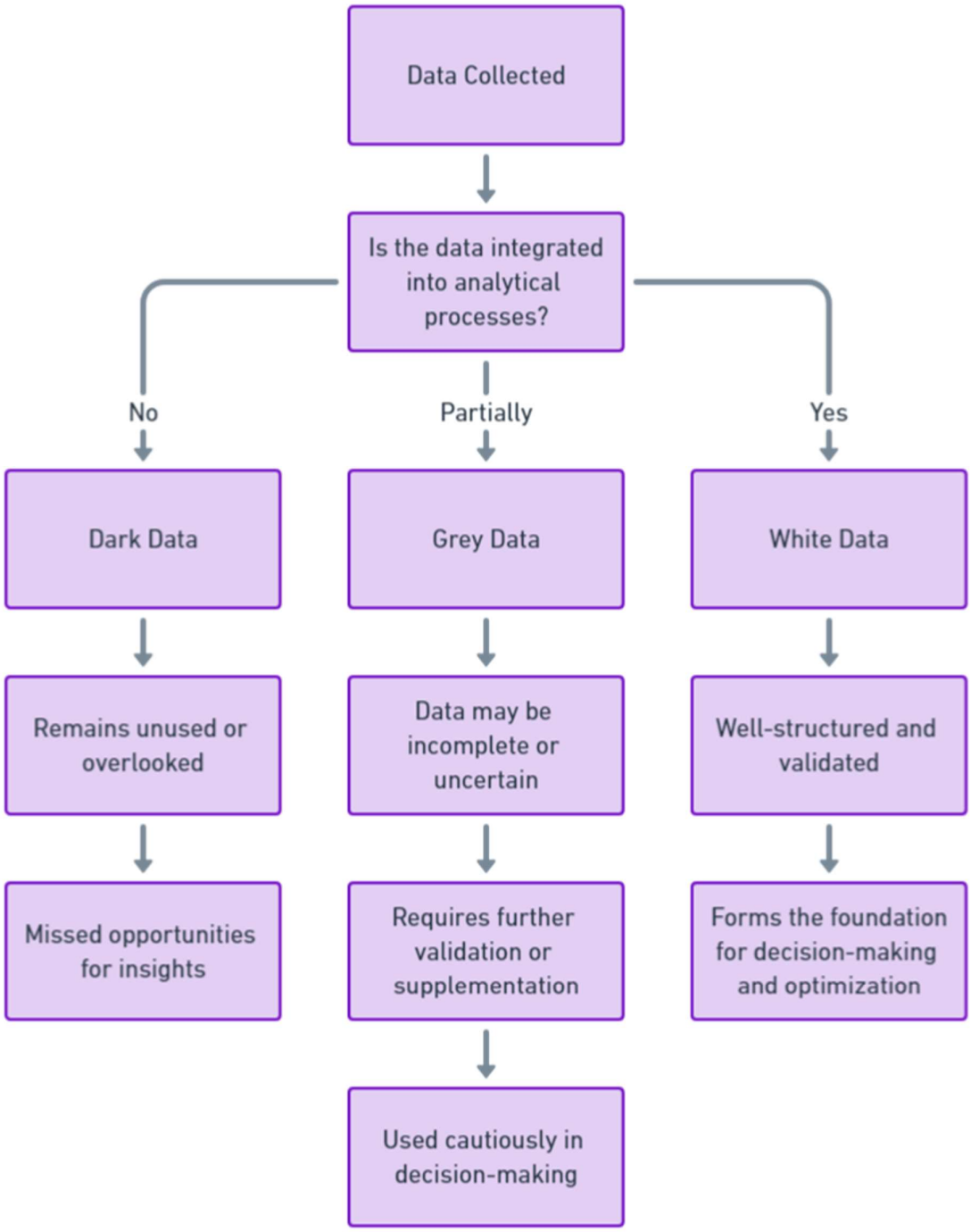
Usage in Analysis. Refers to the application of various techniques, tools, and methods to examine, interpret, and derive meaningful insights from data, supporting decision-making, forecasting, and optimizing processes within organizations.

Dark data is often unused in analysis or decision-making. This type of data, despite being collected and stored, remains untapped, either due to a lack of awareness or because it is not deemed relevant. Examples include log files, unused survey results, or archived records that are not integrated into the

organization's analytical processes. The lack of usage can result from data being unstructured, siloed, or lacking visibility within the organization. As a consequence, this data's potential value remains unrealized, leading to missed opportunities for insights or improvements that could be gleaned if it were properly analyzed.

Grey data, on the other hand, is used with caution and may require further validation before it can be relied upon. This data may be partially analyzed or recognized within the organization, but its reliability may be in question due to its incomplete or uncertain nature. Examples include estimated figures, preliminary survey data, or partially documented customer feedback. While grey data can inform decision-making when there is no better alternative, organizations must supplement it with additional sources or validate it through further analysis to reduce the risk of misinformed decisions.

White data is the most actively used type of data in analysis and decision-making. It is well-structured, validated, and integrated into the organization's workflows, forming the foundation for most analytical processes. Examples of white data include verified financial reports, complete customer databases, and thoroughly documented operational metrics. This data's reliability and accessibility make it a critical asset, enabling organizations to make data-driven decisions confidently and efficiently, optimizing both strategic planning and operational performance.



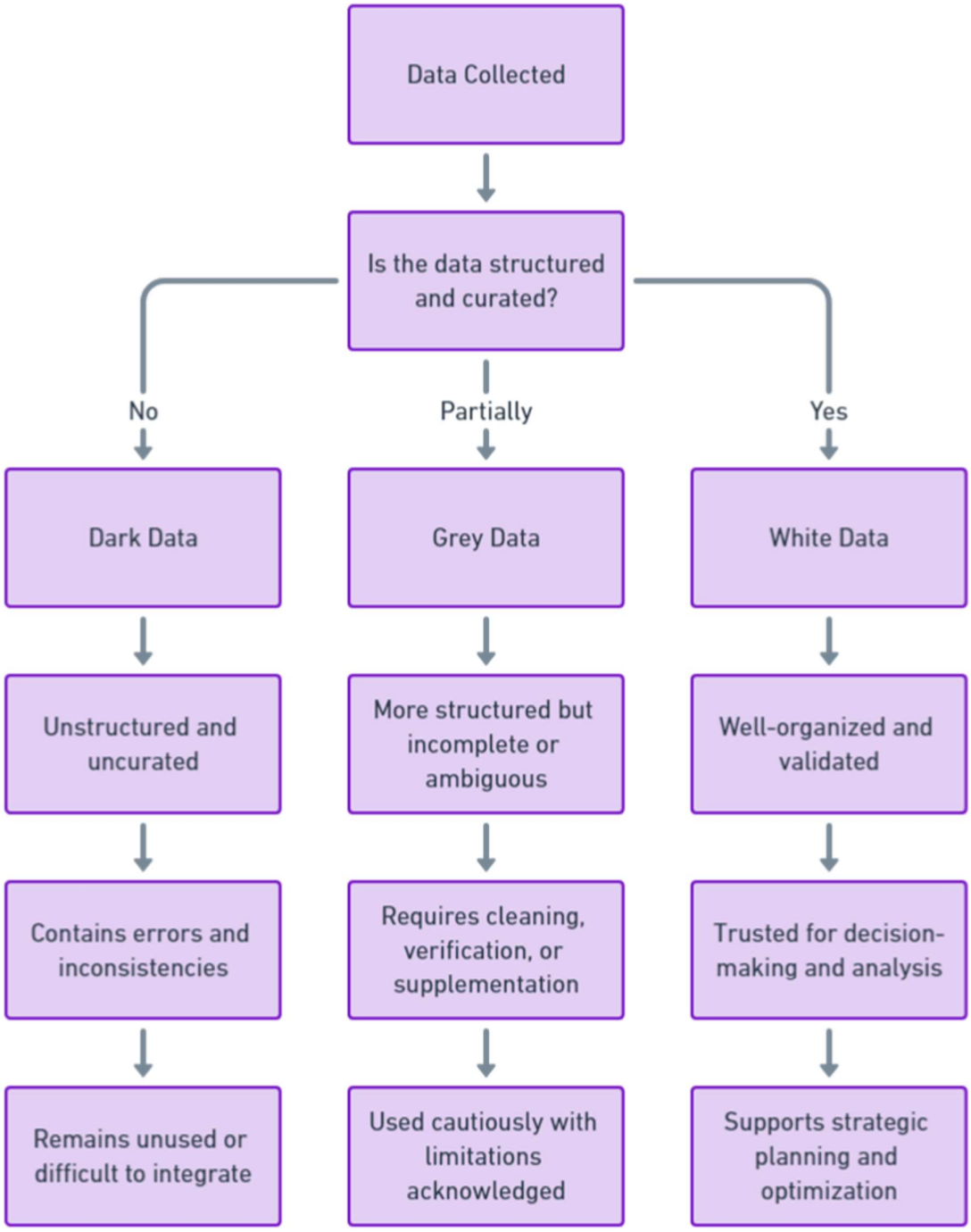
Quality and Completeness. Refers to the accuracy, consistency, and comprehensiveness of data, ensuring it is error-free, well-documented, and fully populated to support reliable analysis, decision-making, and organizational processes.

Dark data is typically unstructured and uncurated, leading to significant variations in quality. Often collected passively, such as through logs, emails, or raw sensor data, dark data remains largely unmanaged and unorganized. Its lack of structure makes it difficult to interpret, process, or integrate into analytical workflows. Without curation, this data can contain errors, inconsistencies, or

redundancies, reducing its reliability. As a result, dark data often remains unused, despite its potential value if properly processed and structured.

Grey data, in contrast, is more structured than dark data but remains incomplete or ambiguous. This type of data may be collected intentionally, such as through surveys or initial research efforts, but often lacks the necessary detail, consistency, or context needed for accurate interpretation. While grey data is accessible, it typically requires further verification, cleaning, or supplementation before it can be fully integrated into analysis. Organizations may use grey data cautiously, acknowledging its limitations while seeking additional sources or conducting further investigations to fill gaps and enhance its reliability

White data represents the highest standard in terms of quality and completeness. It is well-organized, validated, and thoroughly documented, ensuring its consistency and reliability. This type of data is actively curated and maintained, often following strict data governance protocols. Examples include audited financial reports, validated customer information, and complete operational performance metrics. Because of its high quality and completeness, white data is trusted and widely used in decision-making processes. It serves as a critical resource for organizations, providing a solid foundation for accurate analysis, strategic planning, and operational optimization.



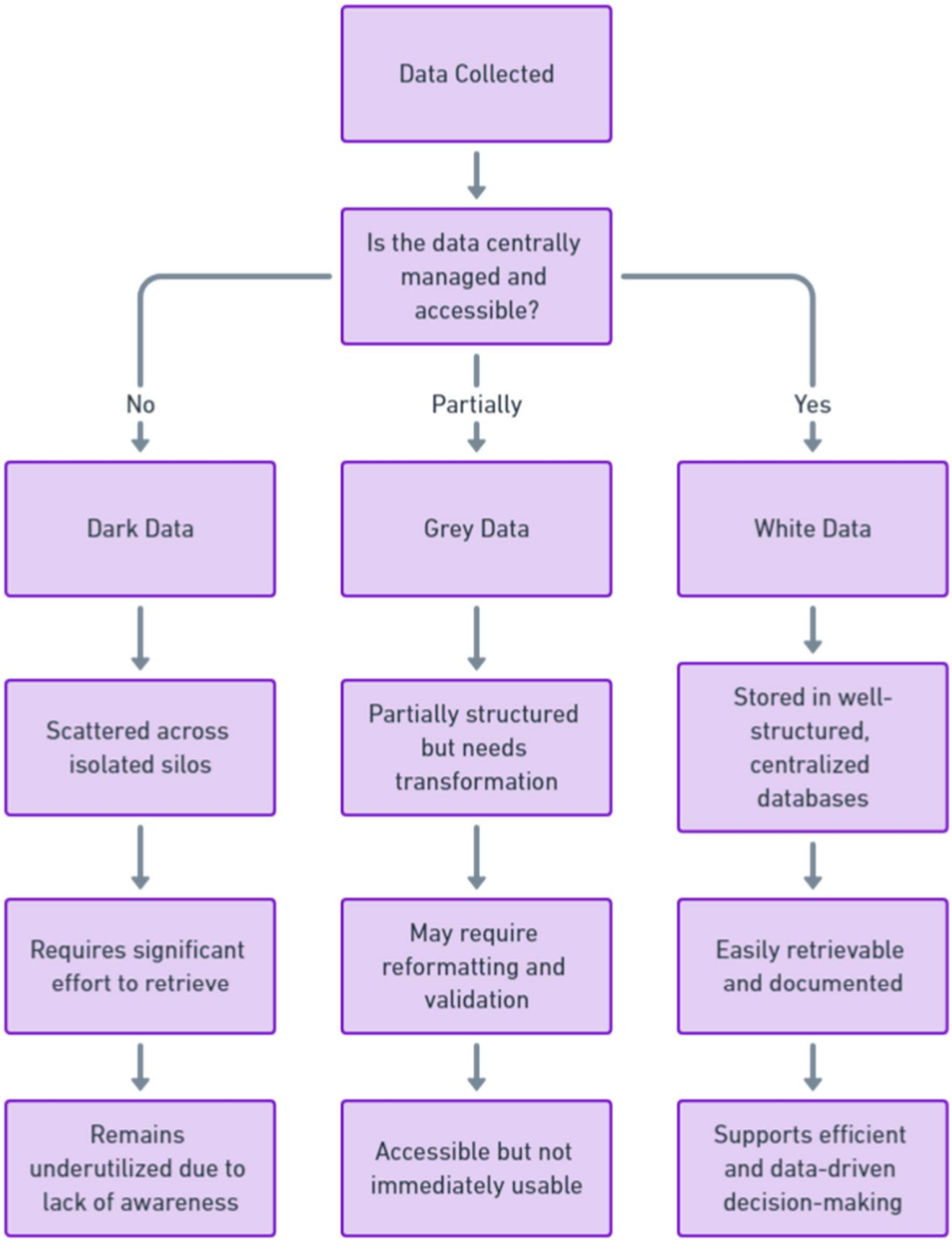
Access Level. Refers to the degree of availability and control over data, determining who can view, modify, or manage information within an organization, based on their roles, permissions, and security protocols.

Dark data is often difficult to access due to a lack of awareness or because it resides in isolated silos. This data may be stored across various unconnected systems, departments, or databases, making it challenging for analysts and decision-makers to retrieve or identify. Examples of dark data include old email archives, server logs, or unused customer records that accumulate over time. Organizations

may not even be aware of its existence, or it may be scattered across different platforms without any centralized management. Consequently, accessing this data often requires significant effort, and its potential value remains unrealized.

Grey data is more accessible than dark data but may not be in the ideal format for immediate use. This data could be stored in systems where it is partially structured or available in formats that are not readily compatible with analytical tools. Examples include preliminary research results or incomplete data files that may require cleaning or transformation before they are usable. Although grey data is recognized within the organization, accessing and preparing it for analysis may involve additional steps like reformatting, filling in gaps, or validating its accuracy. This process can slow down decision-making and increase the resources needed for effective data utilization.

White data is the most accessible type, being readily available and fully integrated into organizational systems. It is stored in well-structured databases, where it is systematically organized, documented, and easily retrievable. Organizations invest in maintaining and integrating white data within centralized data management systems, making it accessible for various stakeholders and ensuring compatibility with analytical tools. This high level of access allows for efficient use of the data in decision-making processes, supporting a data-driven approach and promoting organizational efficiency.



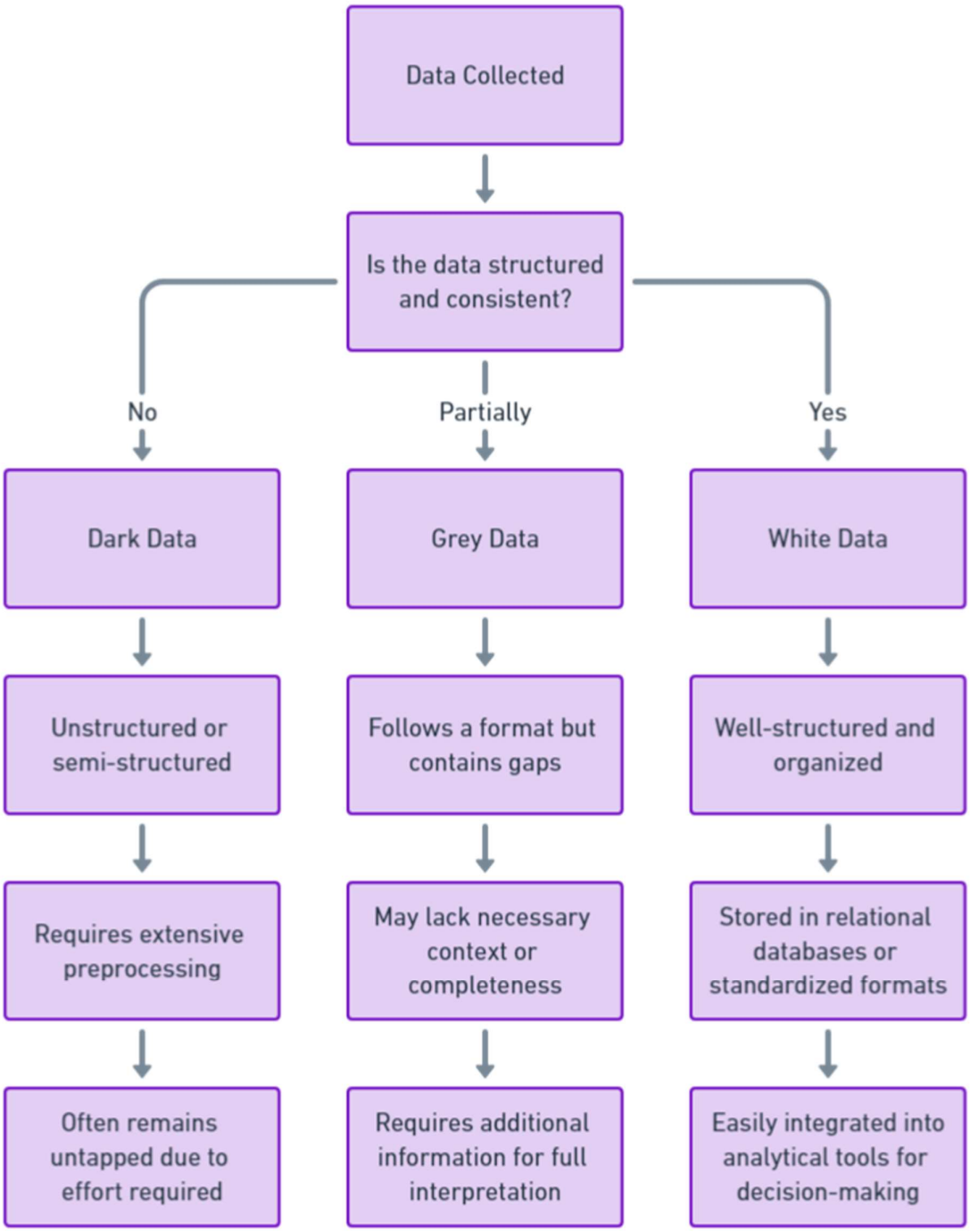
Structure. Refers to the organization and format of data, defining how it is stored, processed, and accessed, typically involving elements like records, fields, and the relationships between data points.

Dark data is typically unstructured or semi-structured, making it challenging to organize and interpret. This type of data often comes in the form of logs, emails, social media feeds, sensor data, or images, which do not follow a consistent format. Without a predefined structure, dark data lacks the organization needed for easy analysis and integration into systems. Its unstructured nature requires

significant preprocessing, such as parsing, cleaning, and transforming, before it becomes usable. As a result, dark data frequently remains untapped, despite its potential value, because it demands extensive resources and effort to convert into a structured format suitable for analysis.

Grey data, on the other hand, is more structured but still contains gaps or lacks the necessary context. Examples include preliminary survey results, partial customer feedback records, or early research findings that follow a defined format but may be incomplete or missing relevant information. While grey data is more organized than dark data, its structural limitations mean that it may not be immediately suitable for analysis. Analysts often need to supplement grey data with additional information or contextual details to make it fully interpretable. This level of structure provides a starting point for analysis, but it requires further refinement to become actionable.

White data is well-structured and easy to interpret, offering a high level of organization and consistency. It is stored in formats such as relational databases, well-documented spreadsheets, or standardized data management systems. Examples include validated financial records, complete customer profiles, and operational performance metrics. White data's clear structure ensures that it can be seamlessly integrated into analytical tools, making it readily usable for decision-making processes. This level of structure supports efficient analysis and facilitates a data-driven approach, maximizing its value for the organization.



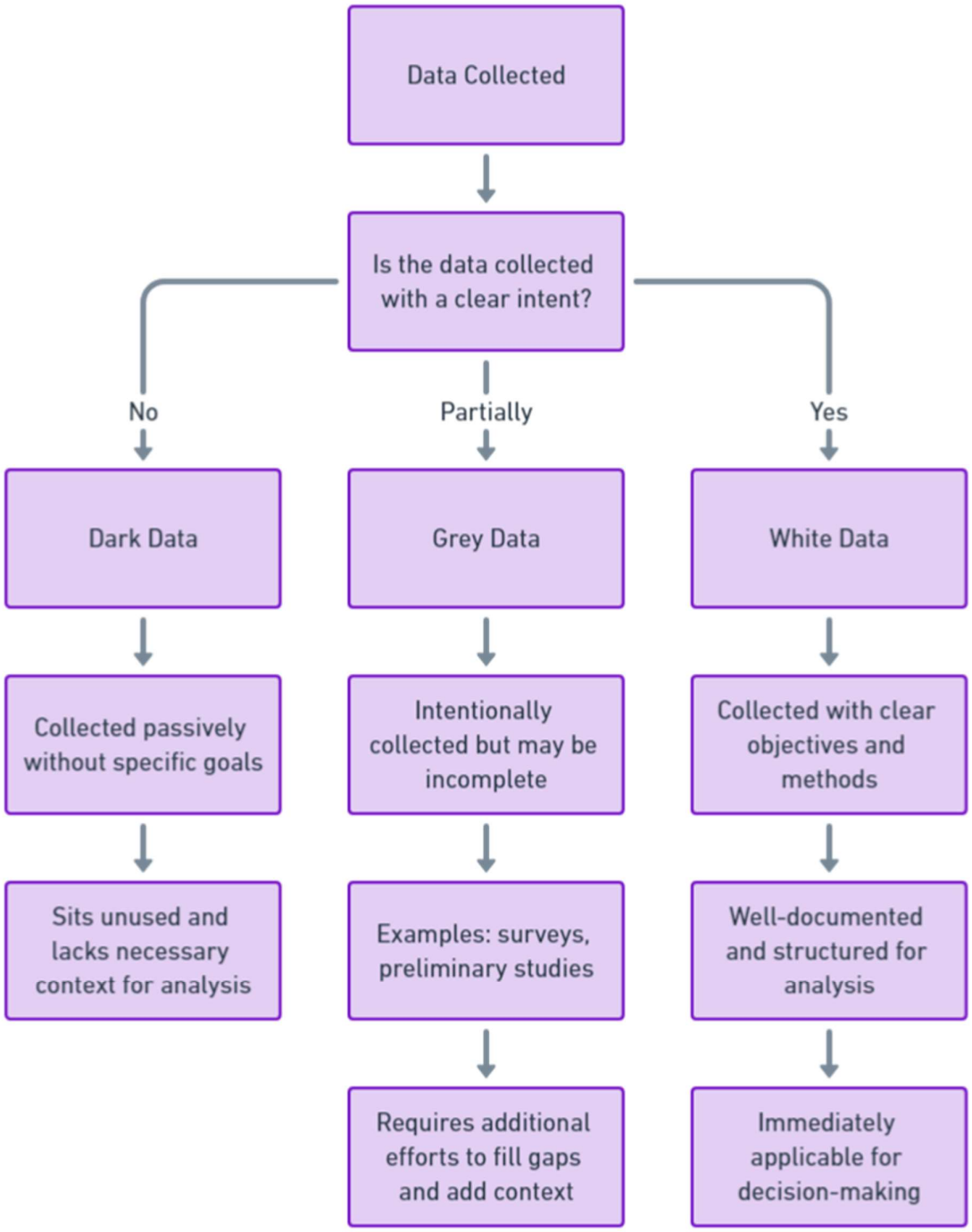
Source of Data. Refers to the origin or provider from which data is collected, such as databases, surveys, sensors, documents, or online platforms, serving as the initial point for data gathering and analysis.

Dark data is typically collected passively, often without a clear intent for its use. Examples include server logs, system event records, or security camera footage. Organizations gather this data as a byproduct of regular operations, usually for compliance, security, or archival purposes, but it is rarely collected with specific analytical or decision-making goals in mind. As a result, dark data often sits

unused in databases, with its potential value overlooked. Because it is collected without a focus on actionable insights, it may lack the necessary context or structure needed for meaningful analysis.

Grey data, in contrast, is intentionally collected but may be incomplete or lack certain details. This type of data is often gathered during surveys, preliminary studies, or pilot programs, where the intent is clear, but the information may be collected under conditions that limit its completeness or accuracy. For instance, early-stage research data, initial feedback forms, or estimated sales figures fall into this category. While grey data is more focused than dark data, the methods used to collect it might not be thorough enough, leading to gaps or ambiguities that need to be filled before it can be used effectively.

White data is actively collected with well-defined objectives and methods. Organizations intentionally gather this data to meet specific goals, such as understanding customer behavior, monitoring financial performance, or tracking operational efficiency. Examples include validated survey results, comprehensive sales records, and quality-controlled production data. Because white data is collected with a clear purpose and under controlled conditions, it is reliable and immediately applicable for decision-making and analysis. This proactive approach ensures that white data is comprehensive, accurate, and ready for integration into analytical models, maximizing its utility and impact.



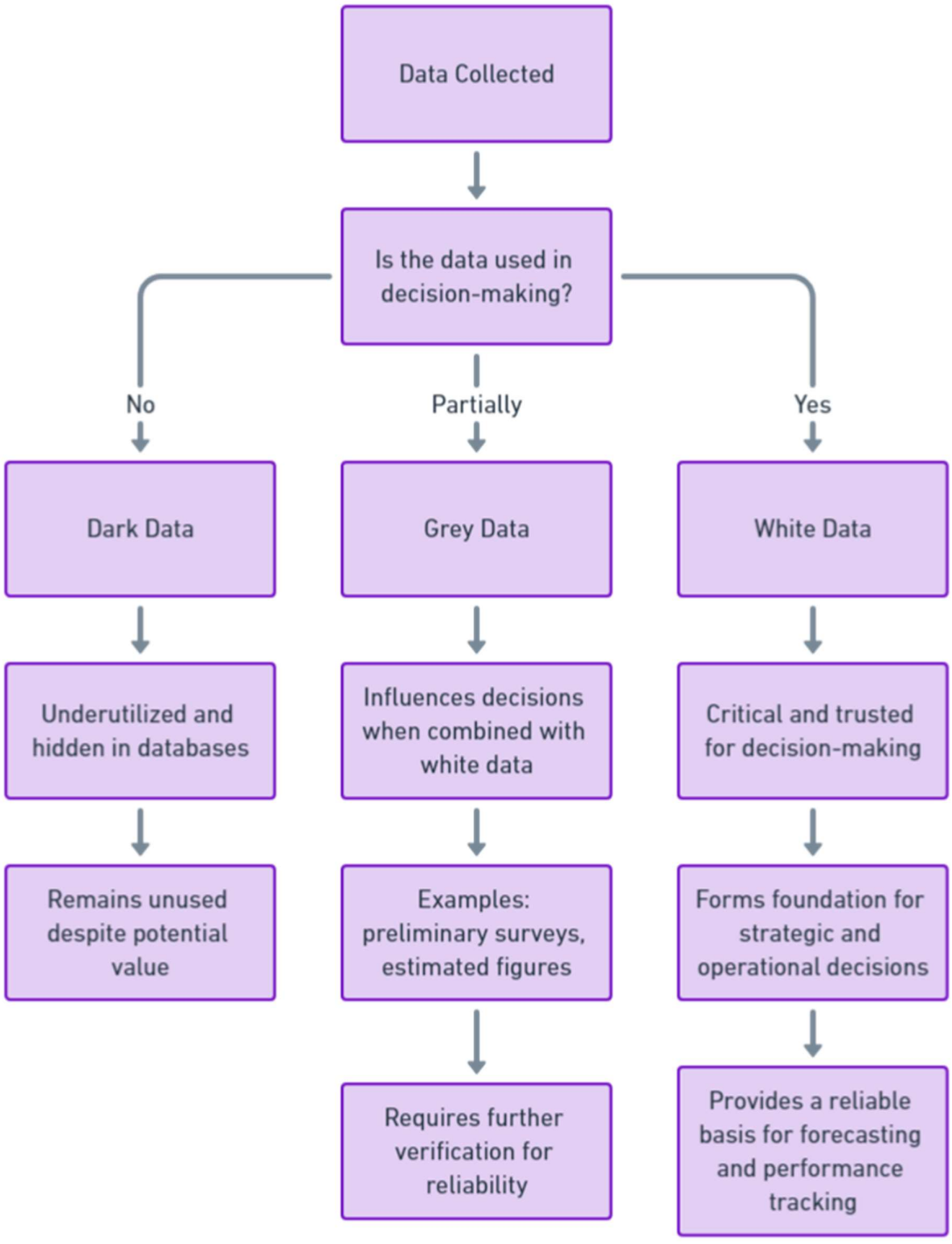
Role in Decision-Making. Refers to the influence and contribution of a person, group, or data type in guiding and shaping choices, strategies, and actions taken within an organization or decision-making process.

Dark data plays no role in decision-making due to its underutilization. Despite being collected and stored, dark data remains largely unused because it is often hidden in databases or stored in unstructured formats that make it difficult to access and analyze. Examples include system logs, archived emails, and raw sensor data, which may hold valuable insights but are not actively integrated

into decision-making processes. The lack of awareness or resources dedicated to processing this data means that its potential impact on strategic planning, process improvement, or performance evaluation remains unrealized.

Grey data has a more moderate role in decision-making. It may influence decisions when combined with white data, providing supplementary information or a broader context. However, grey data is often incomplete or uncertain, which limits its reliability when used alone. For instance, preliminary survey results, estimated financial figures, or partial feedback data can inform decisions, but they require further verification or cross-referencing with more accurate, validated sources. Organizations may use grey data to form hypotheses or support initial planning stages, but they usually exercise caution, acknowledging its limitations and the need for validation.

White data is critical for decision-making and analysis. It is high-quality, fully validated, and systematically collected, making it the most trusted source of information for organizations. Examples include comprehensive financial reports, customer databases, and performance metrics that are regularly analyzed and integrated into business intelligence systems. White data forms the foundation for strategic and operational decisions, providing a reliable basis for accurate forecasting, performance tracking, and risk assessment. Its structured and validated nature ensures that organizations can make data-driven decisions with confidence, optimizing efficiency and achieving their objectives.



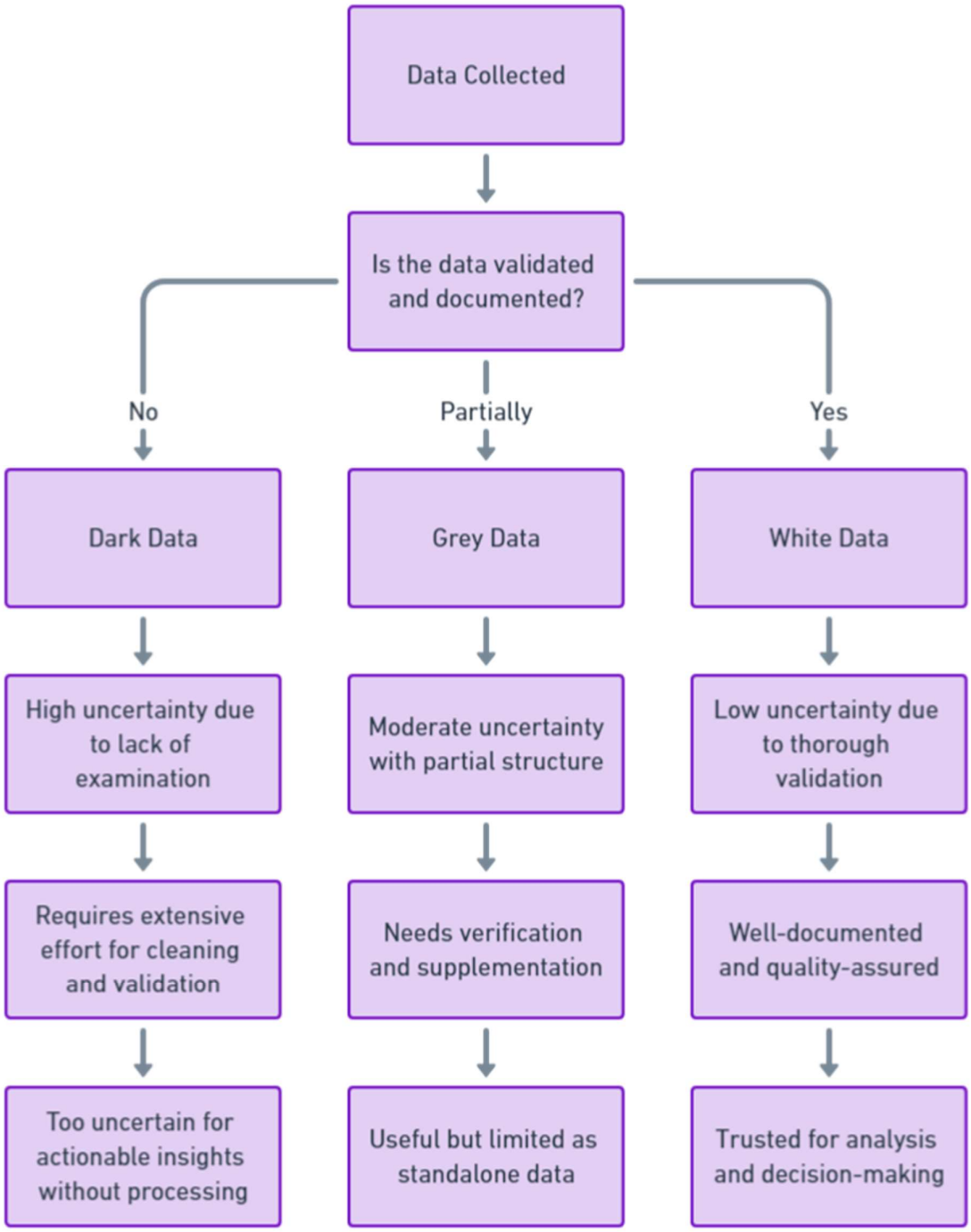
Data Uncertainty. Refers to the lack of precision, accuracy, or completeness in data, often resulting from measurement errors, missing values, or inconsistent documentation, making it challenging to derive reliable insights and conclusions.

Dark data is characterized by high uncertainty because it is rarely, if ever, examined or validated. This data, which may include server logs, old email archives, or sensor recordings, is often collected passively without specific objectives for its use. As a result, it lacks the documentation and context

needed to assess its accuracy or relevance. The absence of a structured review process further compounds the uncertainty, making it unreliable for analysis or decision-making. This high level of ambiguity means that, without significant effort to clean, structure, and validate it, dark data remains too uncertain to be actionable.

Grey data has moderate uncertainty. While it is more structured and has been collected with some intent, it still requires further verification to be fully reliable. Examples of grey data include preliminary research results, incomplete surveys, or early feedback forms. These datasets may provide valuable insights, but their partial nature means there are gaps or inconsistencies that could affect their accuracy. Organizations must verify grey data by cross-referencing it with more reliable sources or supplementing it with additional data. Though it can be useful, the moderate uncertainty of grey data limits its standalone value in decision-making.

White data exhibits low uncertainty. This data is thoroughly documented, well-validated, and systematically reviewed to ensure its accuracy. White data, such as audited financial statements, comprehensive customer records, or quality-assured operational metrics, is collected with rigorous controls and standards. Its low uncertainty makes it the most trusted and reliable data for analysis, enabling organizations to base critical decisions and strategic planning on it with a high degree of confidence.



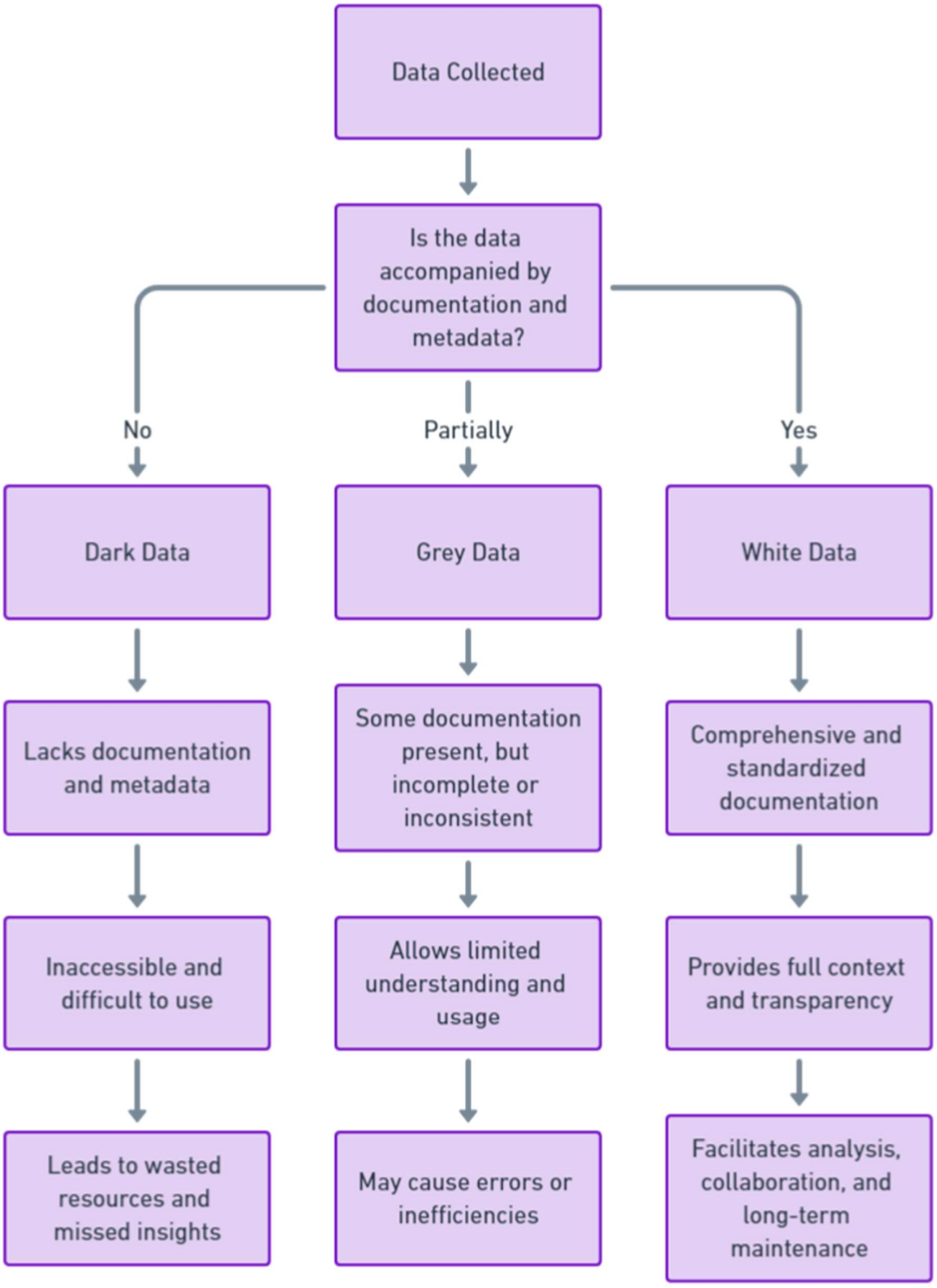
Documentation Level. Refers to the extent and quality of information recorded about data or systems, detailing its structure, usage, and context. It influences data accessibility, usability, and compliance, ranging from minimal to comprehensive documentation.

Dark Data is characterized by its lack of documentation and metadata, making it difficult or impossible to use effectively. This data remains largely inaccessible because essential information about its structure, origin, or context is missing. It is often stored but not analyzed or applied, resulting in wasted storage resources and lost opportunities for valuable insights. In research and organizational

contexts, dark data presents significant challenges, as the absence of documentation makes it hard to evaluate, replicate, or extend findings derived from it.

Grey Data occupies a middle ground, where some documentation is present, but it may be incomplete or inconsistent. At this level, partial metadata is available, allowing for some degree of understanding and usage; however, it lacks the comprehensive and uniform approach needed for full transparency. This inconsistency can cause confusion or errors when different teams or individuals attempt to interpret or work with the data, leading to inefficiencies and potential misinterpretations. While grey data is more accessible than dark data, its fragmented documentation still limits its overall utility and reliability.

White Data represents the most usable and accessible type of data, accompanied by full and comprehensive documentation. At this level, the data is meticulously described with standardized and complete metadata, providing users with all the necessary context to understand its origin, structure, and variables. Such thorough documentation ensures that the data can be effectively interpreted, analyzed, and shared. It facilitates collaboration, replication, and long-term maintenance, as users have access to a well-organized and transparent data resource. White data exemplifies best practices in data management, maximizing its potential for generating insights and supporting decision-making processes.



Accessibility for tools. Refers to how easily data can be accessed, processed, and utilized by analytical or software tools, often depending on its structure, format, and the availability of compatible technologies.

Dark Data is the most challenging to access with analysis tools due to its unorganized nature or incompatible formats. This type of data is often hidden within systems, databases, or archives without the necessary structure for integration into analytical workflows. It may be stored in legacy systems, undocumented formats, or even as unstructured data, such as email archives or raw sensor logs. Because of its inaccessible state, dark data requires extensive effort to clean, reformat, or even identify before it becomes usable. As a result, this data remains largely untapped and represents a missed opportunity for deriving insights.

Grey Data is somewhat more accessible but still poses challenges for direct analysis. This data may be in partially structured formats, like inconsistent spreadsheets or databases lacking uniformity. While tools can often read and process grey data, it frequently requires preprocessing steps such as cleaning, transforming, or standardizing formats to make it fully compatible with analysis software. Analysts and data scientists may need to apply additional transformations or harmonize the data to resolve inconsistencies. Though it is more usable than dark data, the preprocessing required can consume time and resources, reducing efficiency.

White Data on the other hand, is fully structured, standardized, and designed for easy access by analysis tools. This data is stored in compatible formats, such as structured databases or clean, annotated spreadsheets, which integrate seamlessly with analytics platforms. Its compatibility ensures that analysts can quickly and efficiently load, visualize, and analyze the data without extensive preprocessing. White data exemplifies best practices in data management, supporting efficient workflows and maximizing the potential for generating insights with minimal delays or obstacles.

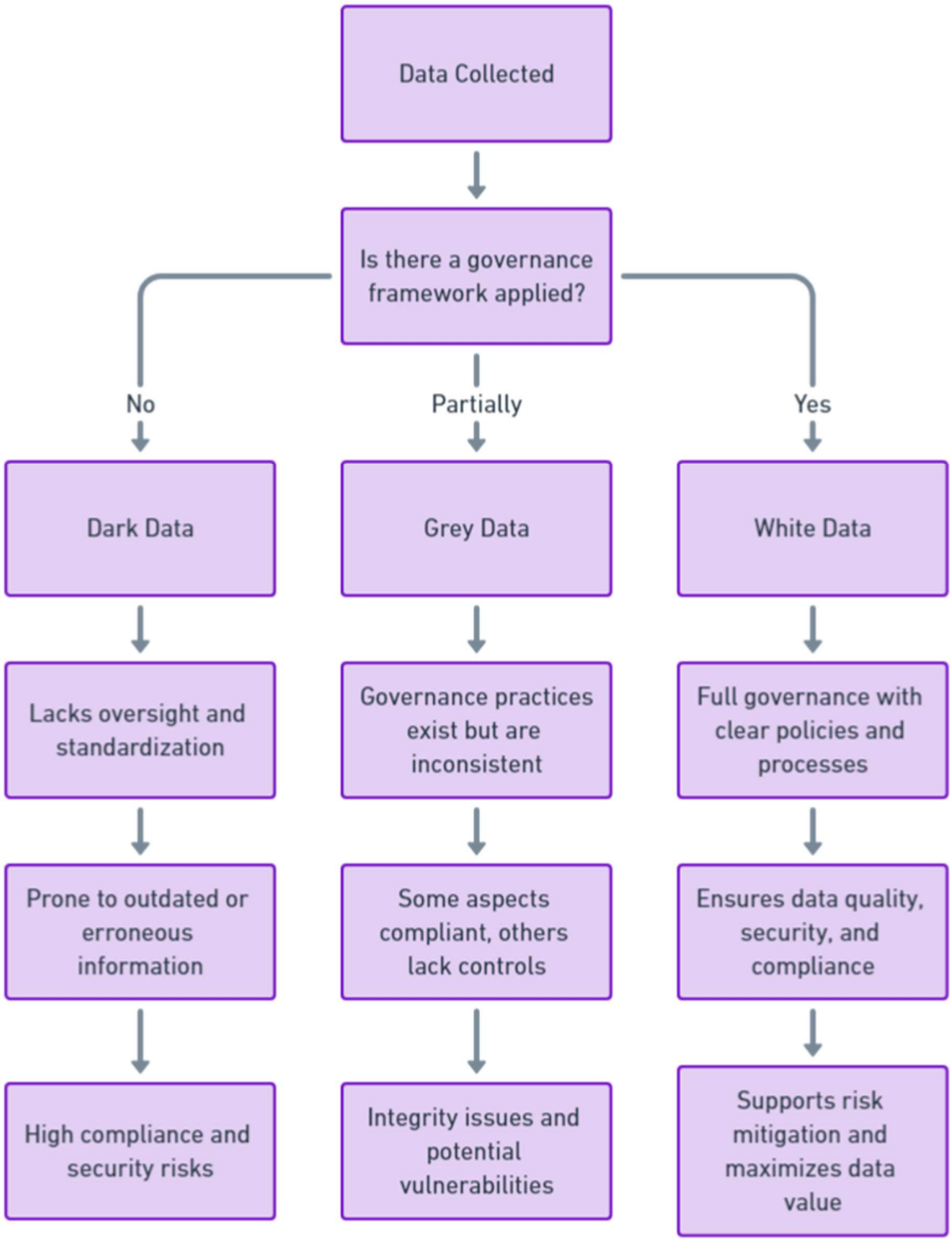
Data governance. Is the process of managing data's availability, usability, integrity, and security within an organization, ensuring compliance with regulations and policies while optimizing data usage to support business operations and strategic decision-making.

Dark Data is poorly governed, often lacking the necessary oversight to ensure its security, quality, and compliance. Because this data is frequently unmanaged, it is prone to being stored in disparate locations without standardized processes for access or updates. The absence of formal governance

frameworks means that data protection measures, such as encryption or access controls, may not be consistently applied, increasing the risk of breaches or unauthorized access. Furthermore, the lack of documentation and oversight makes it difficult to maintain data quality, resulting in outdated, redundant, or erroneous information. This chaotic environment exposes organizations to significant compliance risks, particularly when regulations such as GDPR or HIPAA are in play, as they may not even be aware of the sensitive data they hold.

Grey Data falls into a middle ground where governance practices exist but are inconsistent or only partially implemented. Organizations may have basic policies for managing and securing grey data, but these policies are not uniformly applied or enforced across all data sources. While some aspects of grey data may be compliant with regulations, other parts may lack the necessary controls or documentation, creating vulnerabilities. Inconsistencies in governance can lead to data integrity issues, with discrepancies between data sets or variations in how data is collected, stored, or accessed. This partial governance may be sufficient for less sensitive data, but it still leaves gaps that can cause problems, especially if the organization attempts to scale its data usage.

White Data represents the pinnacle of effective data governance. It is fully governed with clear policies and processes that ensure data quality, security, and compliance. White data is well-documented, regularly audited, and maintained according to industry standards and regulatory requirements. Robust access controls and encryption measures are applied, and the data is consistently monitored and updated to maintain its accuracy and relevance. This level of governance supports compliance efforts, mitigates risk, and ensures that the organization can confidently use and share data, maximizing its value while maintaining legal and ethical standards.



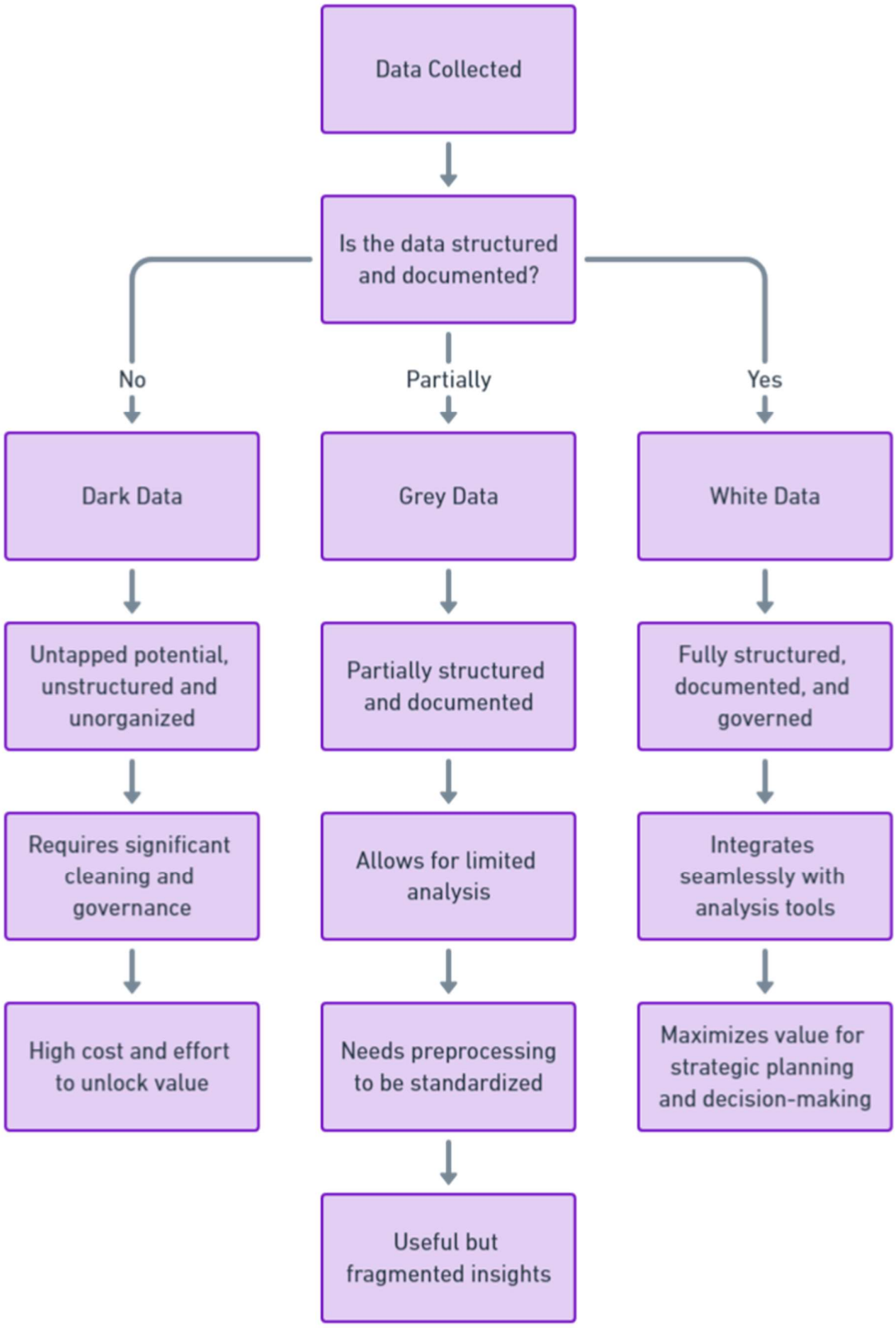
Value of organization. Refers to the benefits gained when data is systematically structured, managed, and documented, enabling efficient access, analysis, and utilization for informed decision-making, operational efficiency, and strategic advantage.

Dark Data possesses untapped potential but remains hidden and underutilized. This type of data is often collected passively and stored without structure, organization, or proper documentation. Examples include unstructured logs, old transaction records, or unused sensor data. Since it is not

integrated into the organization's analytic processes, dark data is often neglected, leading to missed opportunities for valuable insights. Companies may not even be fully aware of the extent or content of their dark data, making it difficult to transform this raw information into actionable knowledge. To unlock its value, significant efforts in data cleaning, structuring, and governance would be required, which is often seen as a cost-prohibitive and labor-intensive endeavor.

Grey Data sits in the middle of the spectrum, providing some value but often lacking the consistency needed for it to be fully actionable. It may be partially structured and documented, which allows for some level of analysis. However, grey data often requires considerable preprocessing, such as cleaning and transforming, to standardize it for effective use. While it may be suitable for limited or exploratory projects, the inconsistency and partial nature of grey data restrict its capacity to support larger-scale, strategic decision-making. Consequently, the insights derived from grey data may be useful but are often fragmented or not comprehensive enough for long-term planning and growth.

White Data offers the highest value to an organization, serving as a reliable and actionable foundation for generating insights. It is fully structured, well-documented, and governed, ensuring that it integrates seamlessly with analysis tools. This makes white data ideal for strategic planning, forecasting, and operational improvements. Because of its high quality and accessibility, white data allows organizations to make informed, data-driven decisions, maximizing its potential to drive innovation and long-term success.

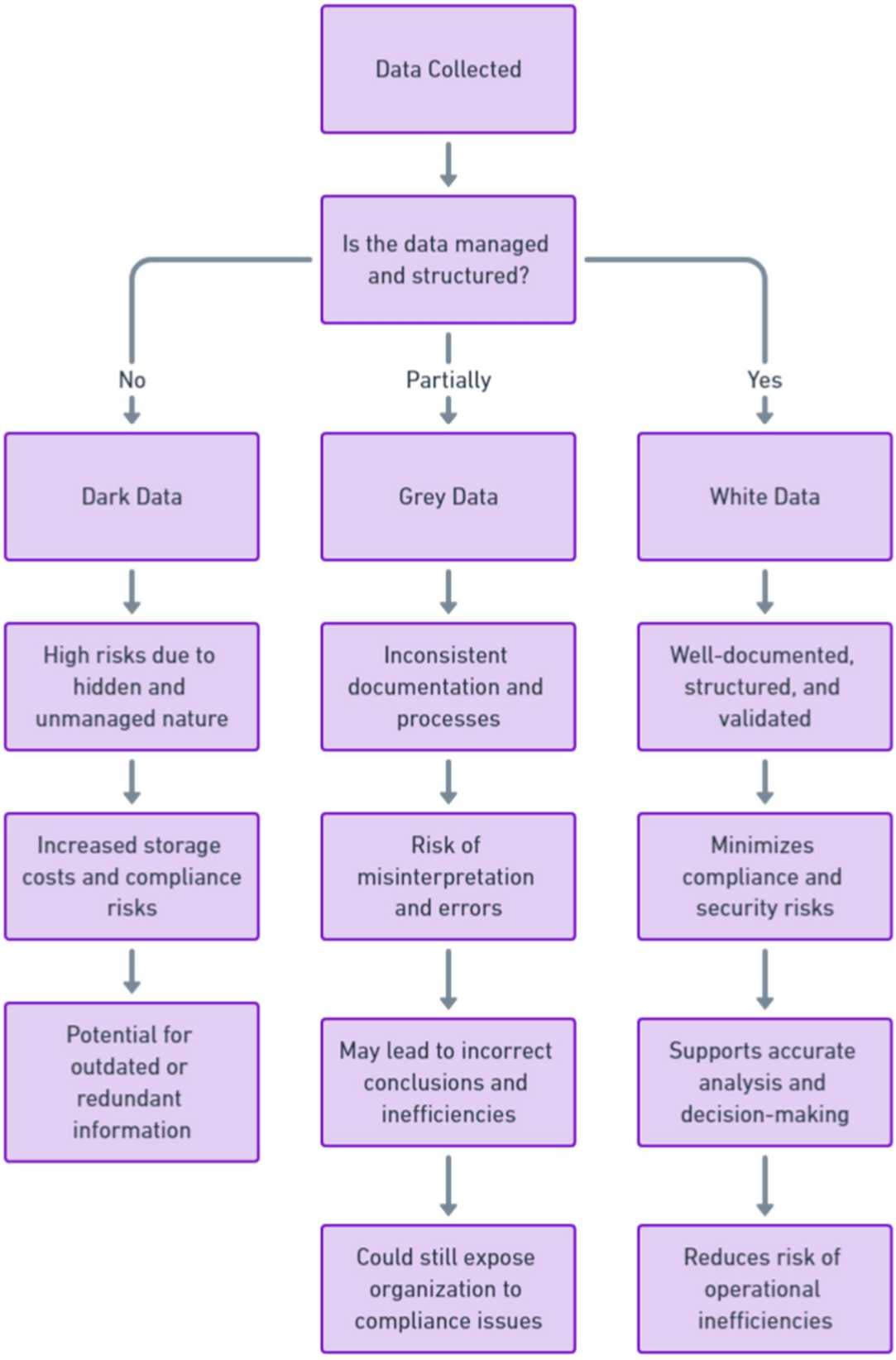

Data Collected
Is the data structured and documented?
No
Partially
Yes
Dark Data
Grey Data
White Data
Untapped potential, unstructured and unorganized
Partially structured and documented
Fully structured, documented, and governed
Requires significant cleaning and governance
Allows for limited analysis
Integrates seamlessly with analysis tools
High cost and effort to unlock value
Needs preprocessing to be standardized
Maximizes value for strategic planning and decision-making
Useful but fragmented insights

Associated Risks. Refer to potential negative outcomes or hazards linked to a particular action, decision, or situation, which could lead to adverse effects such as financial loss, reputational damage, safety issues, or compliance breaches.

Dark Data carries significant risks, primarily due to its hidden, unmanaged nature. Organizations often accumulate vast amounts of dark data without knowing its content, which can lead to increased storage costs. This data might include sensitive information that, without proper governance, poses substantial compliance risks. For instance, regulations like GDPR or HIPAA require organizations to know and control personal data. Without visibility or proper management, dark data can become a liability if it contains sensitive information that is stored insecurely or used without proper authorization. Additionally, dark data is prone to becoming outdated or redundant, making it both costly and risky to maintain without providing any real value in return.

Grey Data, while somewhat more accessible and partially structured, also presents risks, particularly related to misinterpretation or incorrect use. Due to inconsistent documentation and a lack of standardized processes, grey data can lead to errors when being analyzed. For example, inconsistent definitions or formats may cause discrepancies in interpretation, leading to incorrect conclusions or flawed decision-making. If grey data is used to inform strategic decisions without proper validation or preprocessing, the organization could face operational inefficiencies or other unintended consequences. Moreover, grey data may still contain sensitive information that, if not fully governed, could expose the organization to regulatory and compliance issues.

White Data is associated with the lowest risk because it is well-documented, structured, and validated through systematic processes. This data is stored securely and managed in compliance with relevant regulations, minimizing the chances of security breaches or misuse. The clarity and reliability of white data reduce the risk of misinterpretation, as it is consistently maintained and validated. By ensuring that the data is accurate, standardized, and accessible, organizations can use white data confidently for strategic planning, decision-making, and operational improvements, minimizing compliance and operational risks.

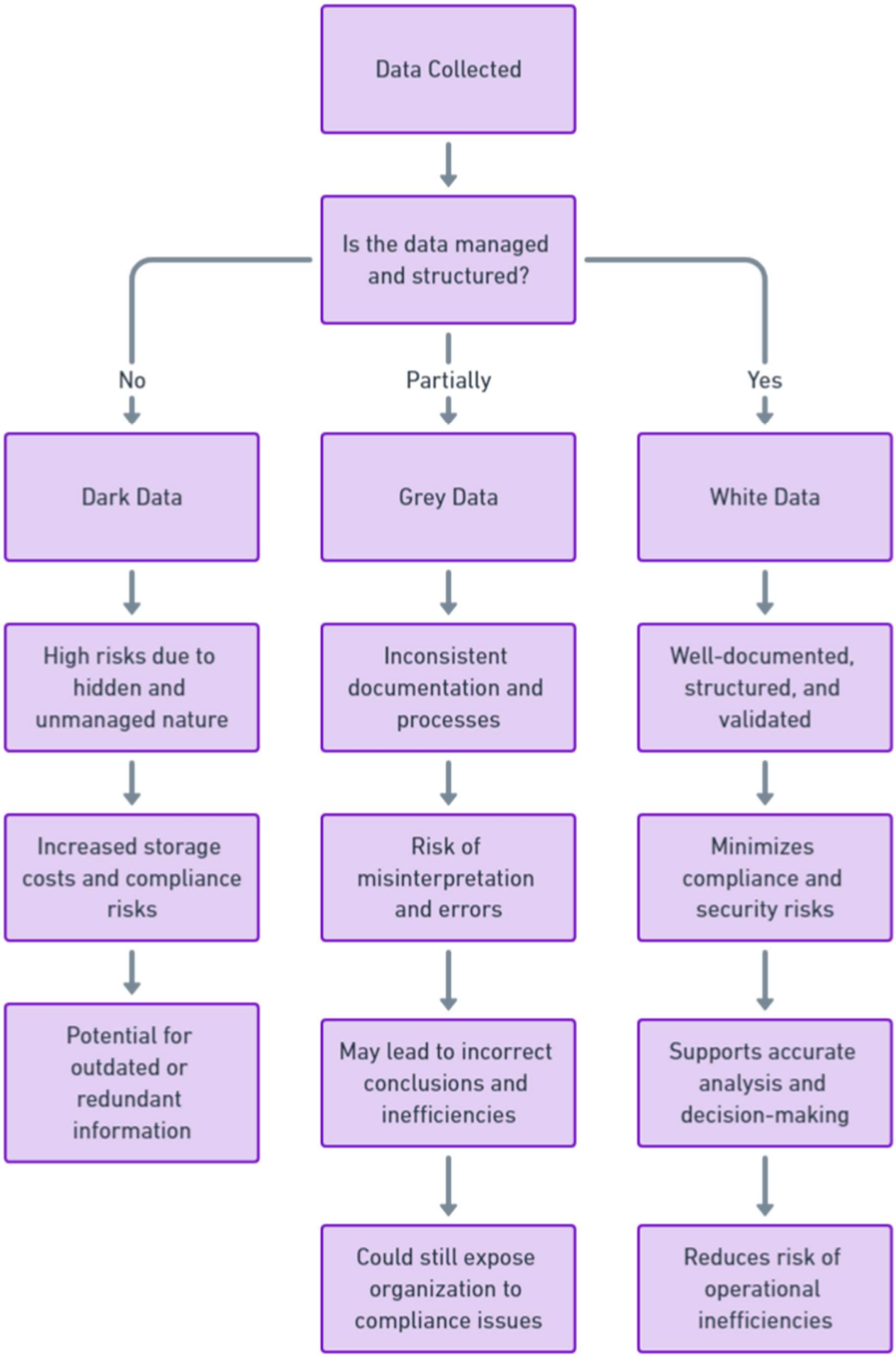

Data Collected
Is the data managed and structured?
No
Partially
Yes
Dark Data
Grey Data
White Data
High risks due to hidden and unmanaged nature
Inconsistent documentation and processes
Well-documented, structured, and validated
Increased storage costs and compliance risks
Risk of misinterpretation and errors
Minimizes compliance and security risks
Potential for outdated or redundant information
May lead to incorrect conclusions and inefficiencies
Supports accurate analysis and decision-making
Could still expose organization to compliance issues
Reduces risk of operational inefficiencies

Data Maintenance Effort. Refers to the resources, time, and processes required to manage, update, clean, and secure data, ensuring its accuracy, consistency, and accessibility throughout its lifecycle for effective use and analysis.

Dark Data is rarely maintained, leading to storage bloat and inefficiencies. This type of data is often accumulated passively and stored without much oversight or organization. As a result, it frequently includes outdated, redundant, or irrelevant information that occupies significant storage space without providing value. Organizations may not even know the extent or content of their dark data, which makes maintenance challenging and costly. The lack of structure or documentation further complicates any efforts to clean, organize, or delete unnecessary files, meaning that storage costs continue to rise while the data remains essentially unusable. Without proactive maintenance, dark data remains a growing liability rather than an asset.

Grey Data is maintained sporadically, requiring extra effort to keep it usable. This data may have some structure and organization, but it lacks the consistency needed for streamlined maintenance. Organizations often invest time in ad-hoc cleaning or updating processes when they need to use grey data for specific projects, rather than having a regular maintenance schedule. This reactive approach means that each time grey data is needed, additional effort is required to preprocess or standardize it. While this allows for some use of the data, it is not an efficient or sustainable method for long-term data management, as inconsistencies and gaps in the maintenance process can still lead to errors and inefficiencies.

White Data is actively maintained and regularly updated, making it the most efficient and valuable type of data for an organization. This data is structured and managed according to well-established protocols, ensuring that it remains accurate, relevant, and compliant with regulatory standards. Regular audits and updates ensure that any changes in the data environment or new information are integrated seamlessly. This systematic maintenance approach maximizes the data's usability and reliability, supporting its use in strategic decision-making and ensuring that it remains an asset rather than a burden.

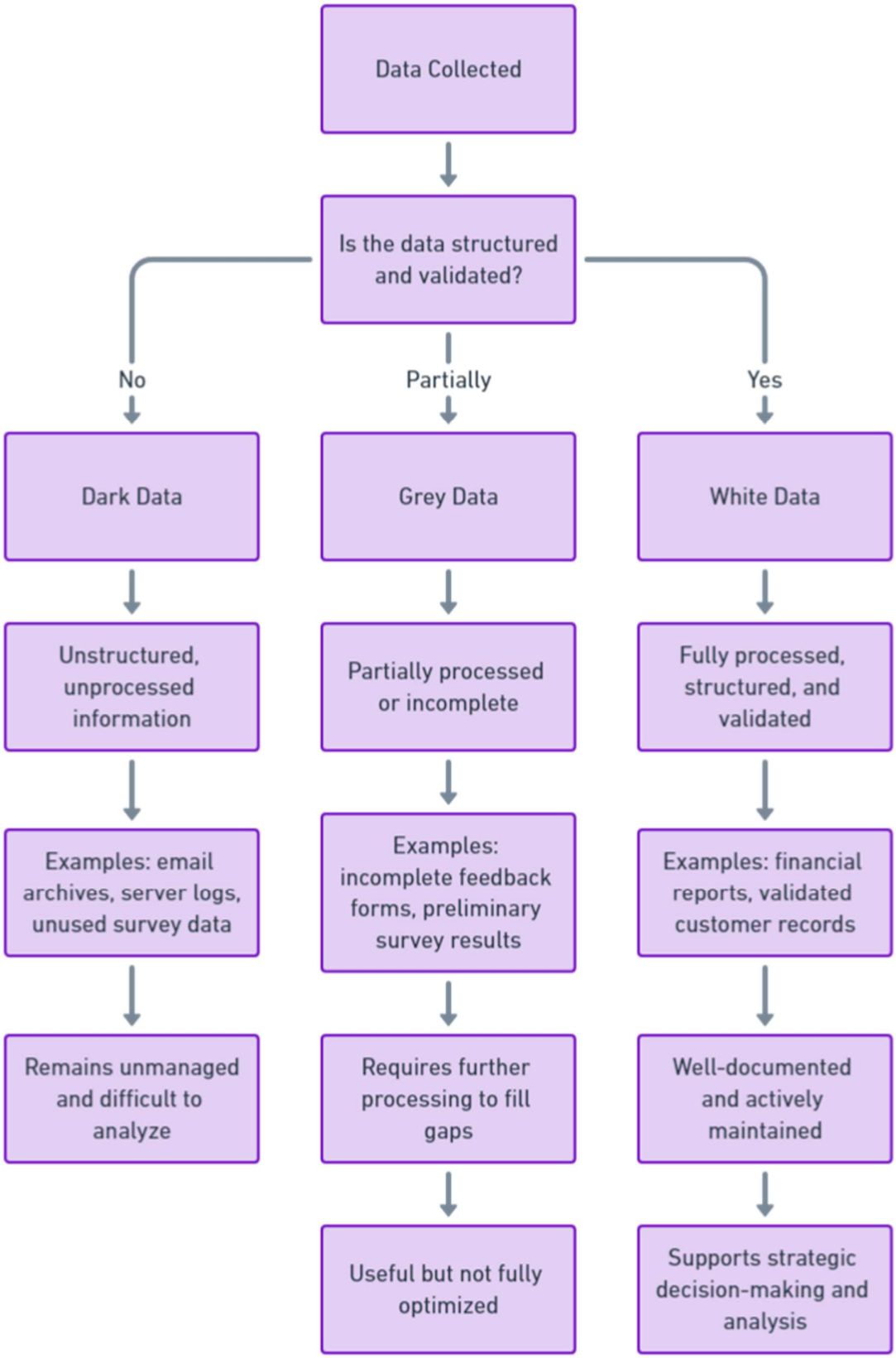

Data Collected
Is the data managed and structured?
No
Partially
Yes
Dark Data
Grey Data
White Data
High risks due to hidden and unmanaged nature
Inconsistent documentation and processes
Well-documented, structured, and validated
Increased storage costs and compliance risks
Risk of misinterpretation and errors
Minimizes compliance and security risks
Potential for outdated or redundant information
May lead to incorrect conclusions and inefficiencies
Supports accurate analysis and decision-making
Could still expose organization to compliance issues
Reduces risk of operational inefficiencies

Example Types. Refers to categories or instances of data used to illustrate various data management practices, such as structured, unstructured, and semi-structured data, highlighting their unique characteristics, usage, and management strategies.

Dark Data consists of unstructured or unprocessed information that organizations collect and store but do not actively use. Examples include email archives, server logs, and unused survey data. These types of data are often accumulated passively and remain unmanaged, leading to a lack of organization or structure that makes them difficult to analyze. Email archives, for instance, contain vast amounts of communication data, but without proper sorting, indexing, or documentation, they are often inaccessible and unutilized. Similarly, server logs and raw survey data may hold valuable insights about customer behavior or system performance, yet they remain dormant due to the complexity and effort required to process them. Dark data, therefore, represents a missed opportunity for organizations, as it is stored without a clear plan for its usage or analysis.

Grey Data includes partially processed or incomplete information that has some potential for use but is not fully optimized. Examples are incomplete feedback forms or preliminary survey results. Such data is more accessible than dark data, as it may have some structure or documentation, but it often lacks the consistency or comprehensiveness needed for effective analysis. For instance, feedback forms that are not entirely filled out or surveys that are only partially completed may provide insights, but they require further processing to fill gaps and standardize formats before they become fully actionable.

White Data is fully processed, structured, and validated, representing the most valuable and usable type of information. Examples include financial reports and validated customer records. These data types are well-documented and actively maintained, ensuring their accuracy and relevance. Financial reports provide structured, verified information that supports strategic decision-making, while validated customer records offer reliable data for analysis, enabling organizations to make informed, data-driven decisions efficiently. White data is therefore an essential asset, actively contributing to organizational goals and insights.

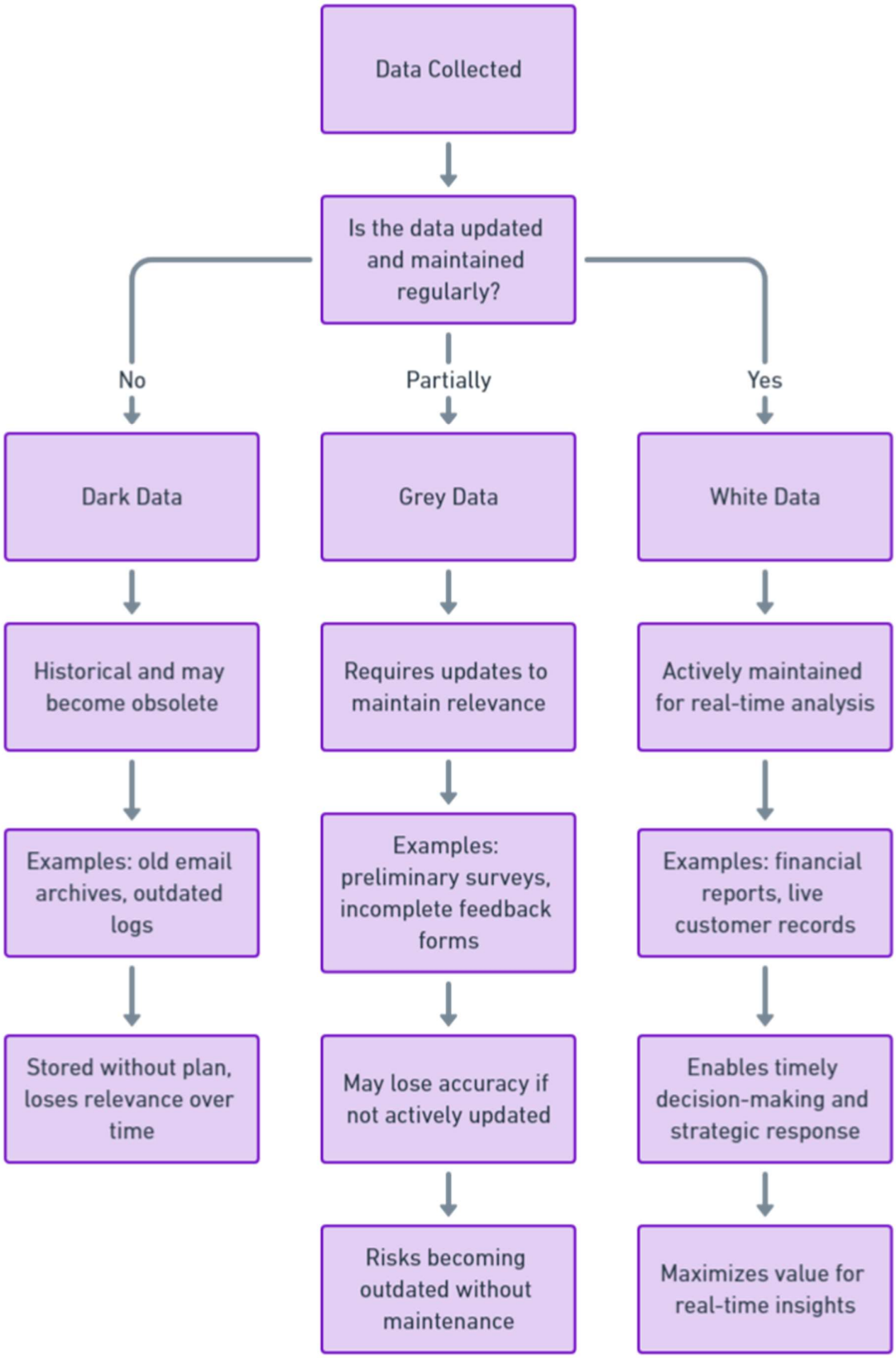

Data Collected
Is the data structured and validated?
No
Partially
Yes
Dark Data
Grey Data
White Data
Unstructured, unprocessed information
Partially processed or incomplete
Fully processed, structured, and validated
Examples: email archives, server logs, unused survey data
Examples: incomplete feedback forms, preliminary survey results
Examples: financial reports, validated customer records
Remains unmanaged and difficult to analyze
Requires further processing to fill gaps
Well-documented and actively maintained
Useful but not fully optimized
Supports strategic decision-making and analysis

Time sensitivity. Refers to the degree to which the value or relevance of data diminishes over time, emphasizing the need for timely processing and analysis to extract accurate and actionable insights.

Dark Data is often historical in nature and may become obsolete over time. Examples include old email archives, outdated server logs, and past survey data that were collected but never utilized. This type of data is typically stored without an immediate plan for analysis or use, leading to its accumulation over months or even years. Because it is rarely updated or maintained, dark data can lose its relevance, making any potential insights it contains less valuable as time progresses. Organizations might store such data for compliance purposes or for potential future use, but without regular updates, the data can quickly become outdated, offering little to no actionable information in the present.

Grey Data tends to be more time-sensitive, meaning that it may require updates to maintain its relevance. This data is often collected for specific projects or tasks, such as preliminary survey results or incomplete feedback forms. While it has some structure and may offer useful information, it is not always maintained systematically. As time passes, grey data might lose its accuracy unless it is actively updated or validated. Organizations that rely on grey data may need to invest time and resources in keeping it current, ensuring that it remains a useful asset. If neglected, grey data risks becoming outdated, losing its relevance and actionable value.

White Data is designed to be up-to-date, often supporting real-time analysis and decision-making. This category includes actively maintained and validated information such as financial reports, live customer records, or real-time sales data. Because white data is updated regularly, it maintains its relevance and accuracy, enabling organizations to respond swiftly to changes or emerging trends. This ensures that white data is not only current but also aligned with strategic objectives, maximizing its value as a resource for real-time insights and actions.

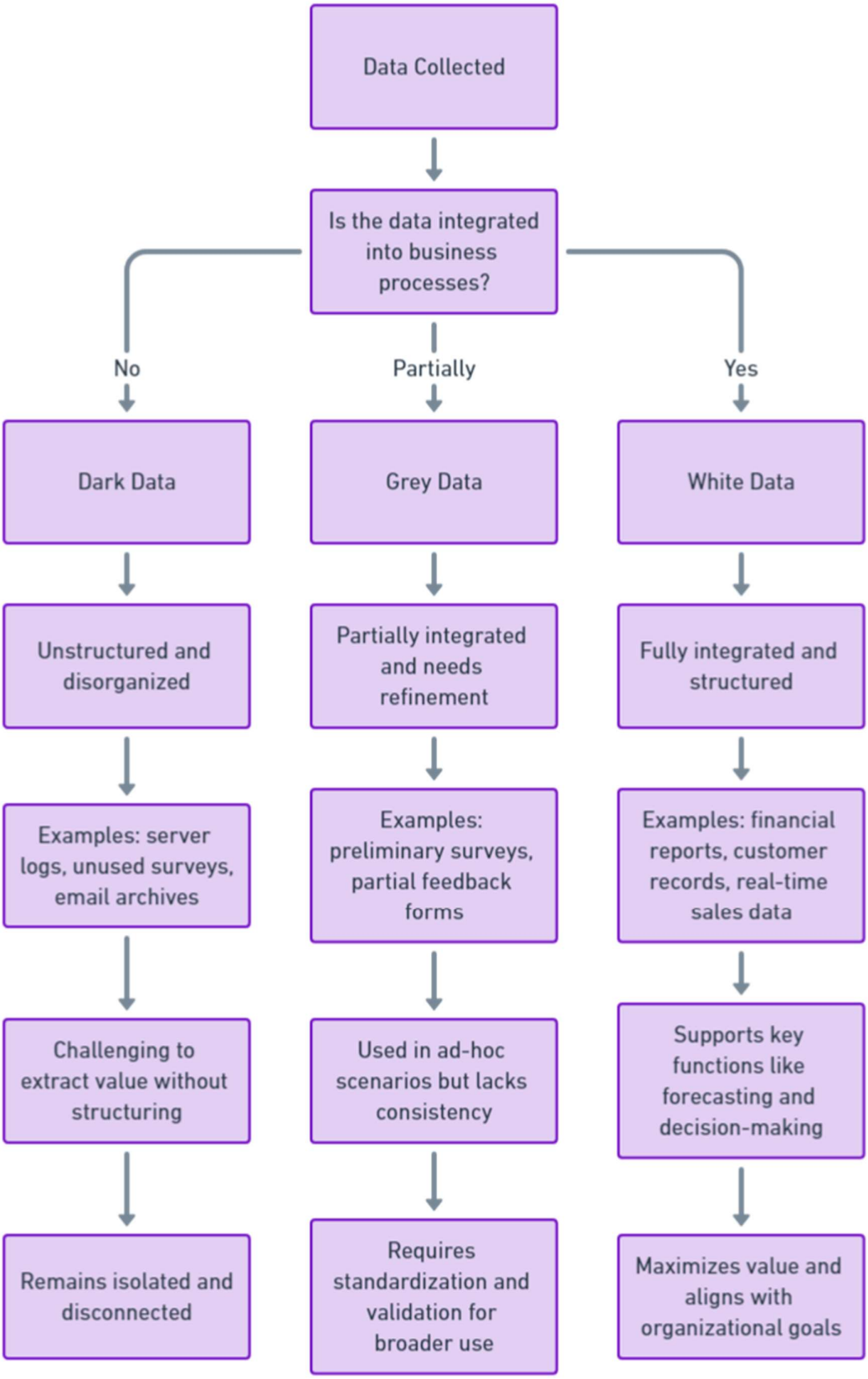

Data Collected
Is the data updated and maintained regularly?
No
Partially
Yes
Dark Data
Grey Data
White Data
Historical and may become obsolete
Requires updates to maintain relevance
Actively maintained for real-time analysis
Examples: old email archives, outdated logs
Examples: preliminary surveys, incomplete feedback forms
Examples: financial reports, live customer records
Stored without plan, loses relevance over time
May lose accuracy if not actively updated
Enables timely decision-making and strategic response
Risks becoming outdated without maintenance
Maximizes value for real-time insights

Integration into Processes. Refers to the incorporation of data into organizational workflows, ensuring that information is systematically used to enhance decision-making, optimize operations, and improve efficiency throughout various business activities.

Dark Data is rarely integrated into business processes due to its unstructured nature and lack of organization. This data is often stored in disparate systems without a clear plan for usage, such as raw server logs, unused survey data, or email archives. The absence of structure and documentation makes it challenging for organizations to extract value or insights from dark data. As a result, this data type remains isolated and disconnected from daily operations or strategic decision-making processes. The unstructured and disorganized state of dark data means that, without significant investment in cleaning, structuring, and integrating it, this resource remains underutilized and dormant.

Grey Data is partially integrated into processes, but often needs further refinement and alignment to be effectively utilized. Examples include preliminary survey results or partially completed feedback forms, which may offer some insights but lack consistency. Grey data is often used in specific, ad-hoc scenarios where quick analyses or exploratory efforts are needed. However, for grey data to be fully embedded into broader organizational processes, it requires additional work—such as standardization, validation, and alignment with existing frameworks. This refinement process may be time-consuming, making it difficult to integrate grey data seamlessly into systems that rely on consistency and accuracy, such as predictive modeling or long-term strategy planning.

White Data is fully integrated and essential for an organization's operations and strategic initiatives. Examples of white data include financial reports, validated customer records, and real-time sales data. This data is actively maintained, well-documented, and structured, allowing it to seamlessly connect with business processes. White data supports key functions like performance monitoring, forecasting, and decision-making, as it is designed to align with the organization's needs and goals. By being consistently updated and verified, white data serves as a dependable foundation for automated systems, real-time analytics, and evidence-based strategies, maximizing its value across the organization.

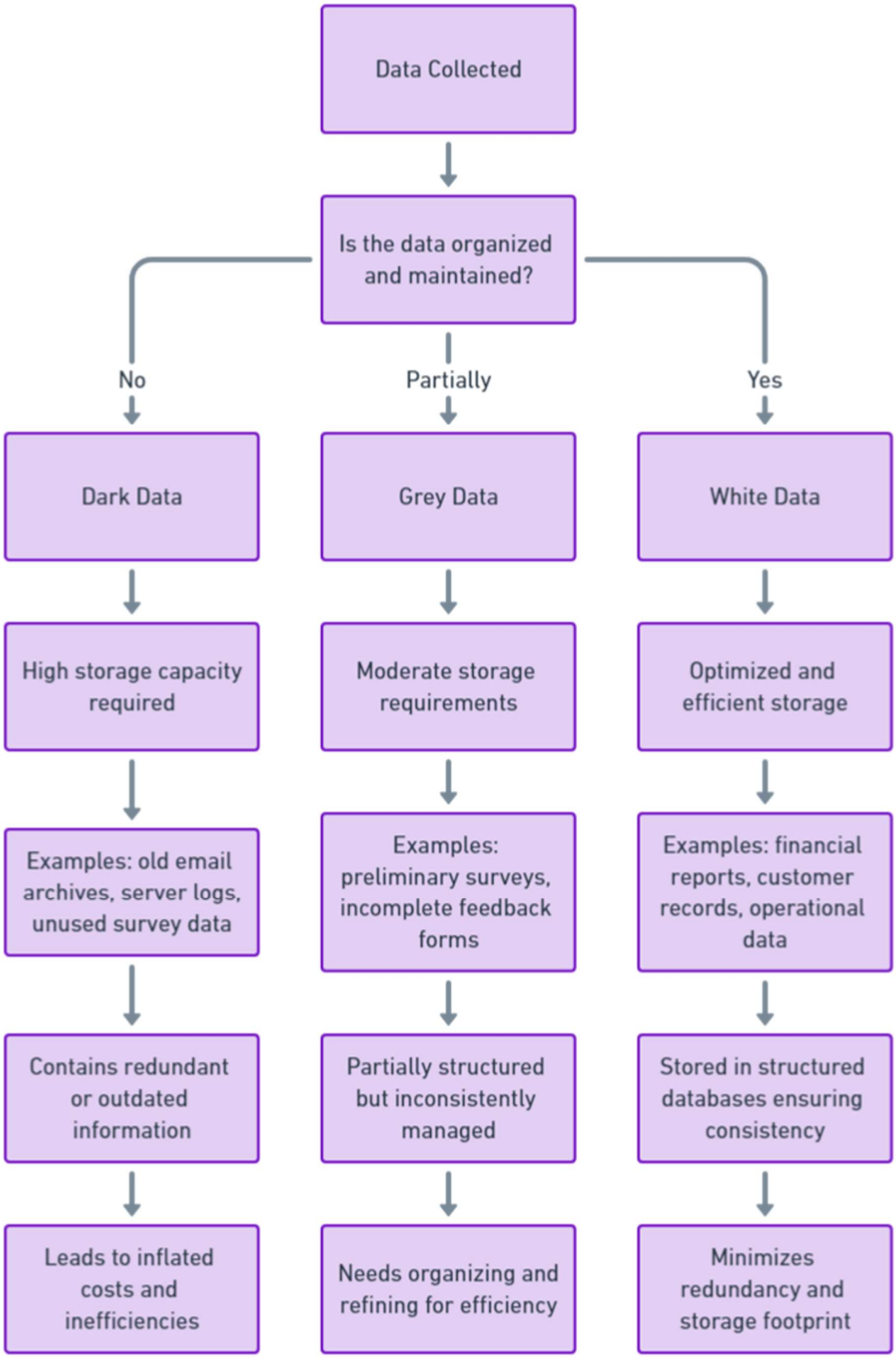

Data Collected
Is the data integrated into business processes?
No
Partially
Yes
Dark Data
Grey Data
White Data
Unstructured and disorganized
Partially integrated and needs refinement
Fully integrated and structured
Examples: server logs, unused surveys, email archives
Examples: preliminary surveys, partial feedback forms
Examples: financial reports, customer records, real-time sales data
Challenging to extract value without structuring
Used in ad-hoc scenarios but lacks consistency
Supports key functions like forecasting and decision-making
Remains isolated and disconnected
Requires standardization and validation for broader use
Maximizes value and aligns with organizational goals

Storage Requirements. Refers to the information detailing the necessary specifications, space, and conditions needed to store digital or physical assets, ensuring accessibility, security, and preservation of data integrity over time.

Dark Data typically requires high storage capacity due to its unstructured and unmanaged nature. This category includes information that is collected and stored but rarely accessed or used, such as old email archives, raw server logs, or unused survey data. Since dark data is not actively maintained or organized, it often contains redundant, outdated, or irrelevant information that continues to accumulate over time. The result is a significant increase in storage needs, leading to inflated costs and inefficiencies. Organizations may not even be aware of the full extent of their dark data, as it is often spread across various systems and lacks central oversight. Without proactive management, dark data continues to grow, consuming valuable storage resources while offering little in return.

Grey Data has moderate storage requirements, as it is partially integrated into active systems but not fully optimized for use. Grey data may include preliminary survey results or incomplete feedback forms—data that has some structure and relevance but lacks consistency and full documentation. It occupies a middle ground where it is not entirely dormant like dark data but is also not systematically used or refined like white data. As a result, grey data still requires considerable storage space, particularly as it may be duplicated or inconsistently managed. Organizations may find it difficult to efficiently store grey data unless they invest in organizing and refining it, aligning it with their existing databases and processes.

White Data is stored in an optimized and efficient manner, reflecting its structured and fully integrated nature. Examples include financial reports, validated customer records, and real-time operational data. White data is actively managed within structured databases that ensure consistency, accessibility, and relevance. These systems are designed to minimize redundancy, making storage highly efficient and cost-effective. By being stored in an organized and well-documented manner, white data maximizes its accessibility while minimizing the storage footprint, enabling organizations to maintain large volumes of valuable data without excessive resource consumption.

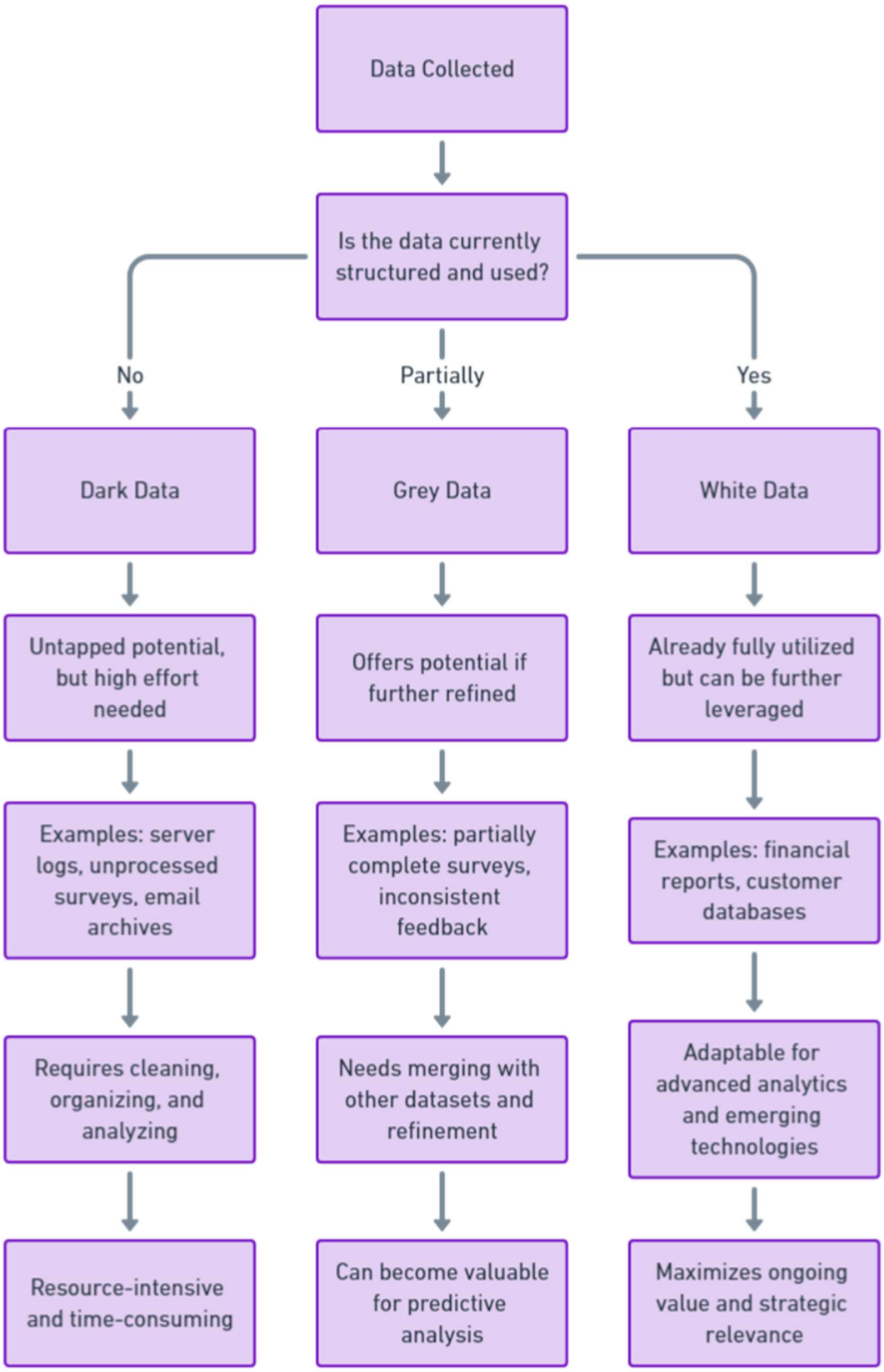

Data Collected
Is the data organized and maintained?
No
Partially
Yes
Dark Data
Grey Data
White Data
High storage capacity required
Moderate storage requirements
Optimized and efficient storage
Examples: old email archives, server logs, unused survey data
Examples: preliminary surveys, incomplete feedback forms
Examples: financial reports, customer records, operational data
Contains redundant or outdated information
Partially structured but inconsistently managed
Stored in structured databases ensuring consistency
Leads to inflated costs and inefficiencies
Needs organizing and refining for efficiency
Minimizes redundancy and storage footprint

Potential for Future Use. Refers to information currently unutilized or under-analyzed but holds potential value if processed, structured, or integrated, enabling it to provide insights and support future decision-making and operations.

Dark Data has significant untapped potential, but realizing this value requires substantial effort and investment. This category comprises vast amounts of information that organizations collect and store without using, such as server logs, unprocessed survey data, or email archives. While this data may contain valuable insights—like patterns in customer behavior or operational inefficiencies—it remains hidden due to its unstructured nature and lack of documentation. To unlock its potential, organizations must undertake extensive processes of cleaning, organizing, and analyzing dark data. This is both time-consuming and resource-intensive, making it a challenge for companies to justify such efforts without clear evidence of immediate benefits. However, for organizations that invest strategically in this area, dark data could provide a wealth of previously undiscovered insights that can inform decision-making and innovation.

Grey Data offers some potential for future use, provided that it is further refined or supplemented. This type of data, such as partially complete surveys or inconsistent feedback forms, has some structure and can be partially analyzed. However, it often lacks the completeness and consistency required for comprehensive insights. Organizations may need to enhance grey data by merging it with other datasets or by refining its structure and documentation. If this refinement process is undertaken, grey data can become a valuable resource for predictive analysis and strategic planning.

White Data is already fully maximized, yet it still offers opportunities for further leverage. This category includes validated, structured, and well-documented data that organizations actively use, such as financial reports and customer databases. While its current value is high, white data can continue to be a foundation for new insights, especially when integrated with advanced analytics, machine learning, or other emerging technologies. White data's structured nature makes it adaptable for future analytical models, ensuring its ongoing relevance and maximizing its utility in strategic contexts.


Data Collected
Is the data currently structured and used?
No
Partially
Yes
Dark Data
Grey Data
White Data
Untapped potential, but high effort needed
Offers potential if further refined
Already fully utilized but can be further leveraged
Examples: server logs, unprocessed surveys, email archives
Examples: partially complete surveys, inconsistent feedback
Examples: financial reports, customer databases
Requires cleaning, organizing, and analyzing
Needs merging with other datasets and refinement
Adaptable for advanced analytics and emerging technologies
Resource-intensive and time-consuming
Can become valuable for predictive analysis
Maximizes ongoing value and strategic relevance

Aspect	Dark Data	Grey Data	White Data
Definition	Data collected but not used for decision-making or analysis.	Incomplete, imprecise, or uncertain information needing further analysis.	Fully accessible, documented, and used for decision-making.
Visibility	Hidden or overlooked within databases.	Partially visible but quality or completeness may be uncertain.	Fully visible and accessible to stakeholders.
Usage in Analysis	Often unused in analysis or decision-making.	Used with caution, may require further validation.	Actively used as the basis for analysis and decision-making.
Quality and Completeness	Unstructured and uncurated, quality varies.	Incomplete or ambiguous, needs further verification.	High-quality, complete, and validated.
Access Level	Difficult to access due to lack of awareness or silos.	Accessible but may not be in the ideal format.	Readily available and integrated into systems.
Structure	Often unstructured or semi-structured (e.g., logs).	Structured but with gaps or lacking context.	Well-structured and easy to interpret.
Source of Data	Collected passively, often without intent for use.	Intentionally collected but may be incomplete.	Actively collected with clear objectives and methods.
Role in Decision-Making	Plays no role due to underutilization.	May influence decisions if combined with white data.	Critical for decision-making and analysis.
Data Uncertainty	High uncertainty due to lack of examination.	Moderate uncertainty, requires further verification.	Low uncertainty, well-documented and accurate.
Documentation Level	Lacks documentation and metadata.	Partial or inconsistent documentation.	Fully documented with comprehensive metadata.
Accessibility for Tools	Not easily accessible by analysis tools.	Accessible but may need preprocessing.	Readily accessible and compatible with analysis tools.
Data Governance	Poorly governed, leading to compliance risks.	Moderately governed but may have inconsistencies.	Well-governed, ensuring data standards and compliance.
Value to the Organization	Untapped potential, remains hidden.	Provides some value but may not be fully actionable.	High value, forms a reliable foundation for insights.
Associated Risks	Storage costs, potential compliance issues.	Risk of misinterpretation or incorrect use.	Low risk, as it is validated and used systematically.
Data Maintenance Effort	Rarely maintained, leading to storage bloat.	Maintained sporadically, often needing extra effort.	Actively maintained and regularly updated.
Example Types	Email archives, server logs, unused survey data.	Incomplete feedback forms, preliminary survey results.	Financial reports, validated customer records.
Time Sensitivity	Often historical, may become obsolete.	Time-sensitive, may need updates for relevance.	Up-to-date, often used in real-time analysis.
Integration into Processes	Rarely integrated, unstructured nature hinders use.	Partially integrated, needs refinement for alignment.	Fully integrated, essential for operations and strategies.
Storage Requirements	High due to unused and redundant information.	Moderate, part of active systems but not fully used.	Optimized, efficiently stored within databases.
Potential for Future Use	Significant untapped potential if analyzed.	Potential value if further refined or supplemented.	Fully maximized, can be further leveraged for insights.

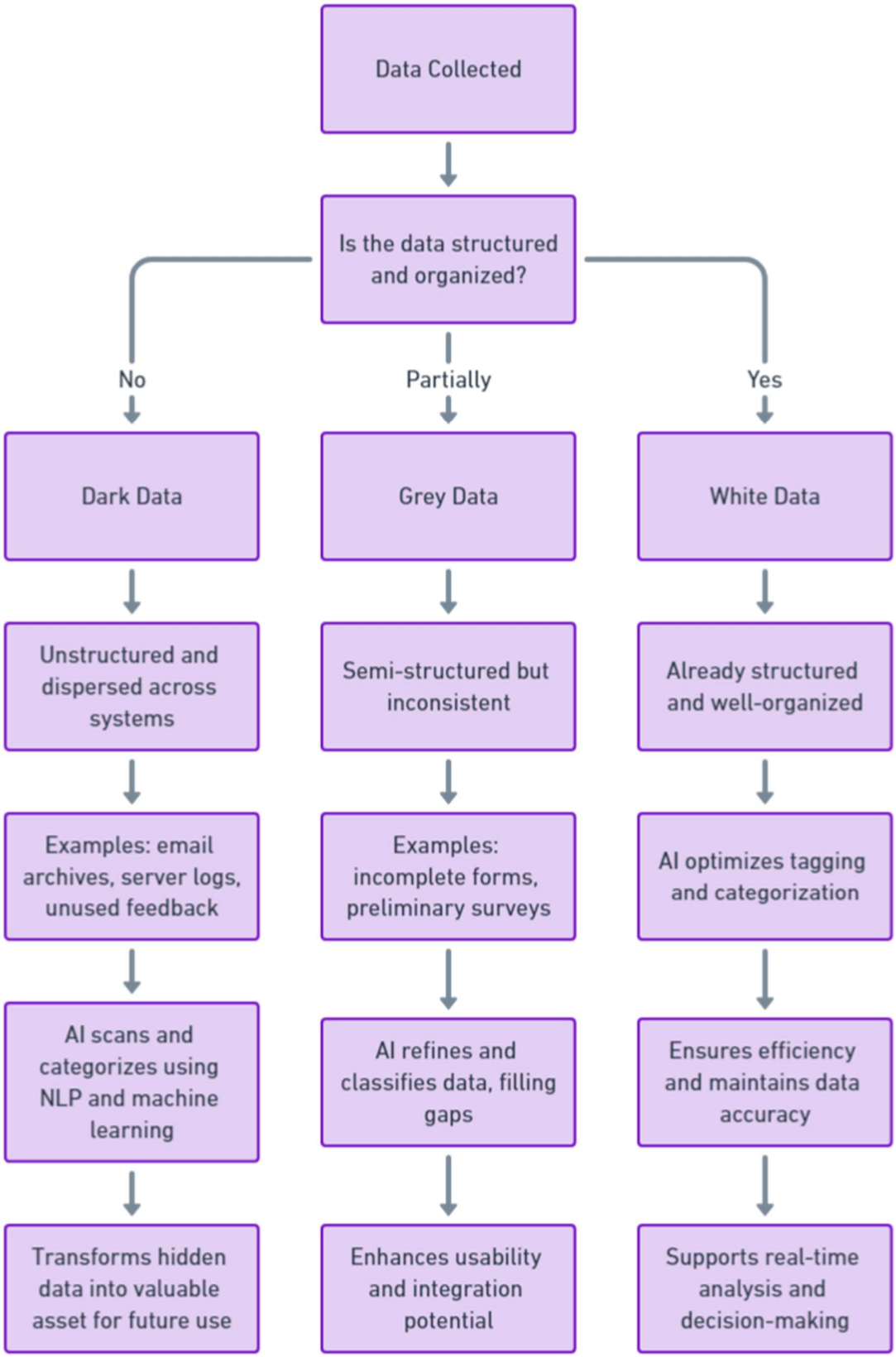
4. Data management and artificial intelligence

Data Discovery and Classification. Is the process of identifying, categorizing, and labeling data based on its sensitivity and type, enabling organizations to manage, protect, and utilize information effectively while ensuring compliance and security.

Dark Data is often unstructured and stored across various systems without organization or documentation. AI can automatically scan these storage systems to identify and categorize unstructured data, which may include email archives, server logs, and unused customer feedback. Using natural language processing (NLP) and machine learning algorithms, AI identifies patterns, topics, and key terms within these vast datasets. This automated classification helps organizations understand what data they possess, enabling them to sort and prioritize it for further analysis. By shedding light on previously hidden or ignored information, AI transforms dark data into a potentially valuable asset, setting the stage for future data utilization and integration.

Grey Data typically has some structure but is not fully organized or consistent, such as incomplete forms or preliminary survey results. AI plays a crucial role here by refining and classifying this semi-structured data, making it easier to manage and access. For example, AI can group and tag similar datasets, identify gaps, and suggest ways to standardize and complete the data. By organizing grey data, AI reduces inconsistencies and enhances the usability of this information, making it more accessible for analysis and integration into business processes.

White Data is already structured and well-organized, but AI optimizes its classification further to enhance efficiency. AI systems can automate data tagging and categorization, ensuring that white data is efficiently managed and easily retrievable. By continuously classifying and reclassifying data based on usage patterns and evolving organizational needs, AI maintains data accuracy, ensuring that it remains an up-to-date and valuable resource for real-time analysis, strategic planning, and decision-making.


Data Collected
Is the data structured and organized?
No
Partially
Yes
Dark Data
Grey Data
White Data
Unstructured and dispersed across systems
Semi-structured but inconsistent
Already structured and well-organized
Examples: email archives, server logs, unused feedback
Examples: incomplete forms, preliminary surveys
AI optimizes tagging and categorization
AI scans and categorizes using NLP and machine learning
AI refines and classifies data, filling gaps
Ensures efficiency and maintains data accuracy
Transforms hidden data into valuable asset for future use
Enhances usability and integration potential
Supports real-time analysis and decision-making

Data Cleaning and Structuring. Is the process of identifying, correcting, or removing inaccuracies and inconsistencies in data, and organizing it into a consistent format for accurate analysis and efficient utilization.

Dark Data, often stored in unstructured forms like server logs, email archives, or raw text, presents significant challenges in its original state. AI technologies such as natural language processing (NLP) and machine learning algorithms can be employed to clean and structure this unorganized data. AI scans through large volumes of information, detecting patterns, removing redundancies, and categorizing content. For example, AI can extract key information from raw text or logs, converting them into structured formats like databases or spreadsheets. By automating these tasks, AI significantly reduces the time and effort required to transform dark data into usable forms, making it accessible for further analysis and decision-making.

Grey Data has some level of structure but is often incomplete or inconsistent. AI refines and fills gaps within these datasets, transforming them into fully structured and coherent information. For example, AI can identify missing fields in survey results or feedback forms and use predictive modeling to fill in these gaps based on existing data patterns. Additionally, AI can standardize data formats and merge information from different sources, ensuring consistency across datasets. This refinement process is crucial in making grey data more usable and reliable for analysis, helping businesses make more accurate and data-driven decisions.

White Data, already structured and validated, benefits from AI's ability to maintain data quality over time. AI monitors white data for anomalies or discrepancies, ensuring that updates and modifications are consistent with predefined standards. By continuously cleaning and verifying data as it is updated, AI guarantees that white data remains accurate and relevant, enhancing its reliability for real-time analysis and long-term strategic planning.

Automated Data Integration. Is the process of using software tools to automatically collect, transform, and merge data from various sources into a unified system, ensuring consistency, accuracy, and efficiency without manual intervention.

Dark Data, which often resides in unstructured forms such as raw logs, emails, or archived files, typically remains isolated from the rest of an organization's structured data systems. AI-driven tools, using natural language processing (NLP) and advanced machine learning algorithms, can extract relevant information from these unstructured sources and convert it into structured formats. Once processed, AI integrates this newly organized data into existing databases or data lakes, aligning it with other structured datasets. This process enables organizations to unlock valuable insights hidden within dark data by merging it with existing analytics and reporting systems, transforming previously dormant data into a useful asset.

Grey Data, which may have partial structure but lacks consistency or completeness, benefits greatly from AI's integration capabilities. AI aligns grey data with established data frameworks by detecting patterns and standardizing formats, ensuring that semi-structured data aligns with the organization's data architecture. For example, AI can clean and standardize incomplete survey results or feedback forms, then integrate them into the company's existing customer relationship management (CRM) or business intelligence systems. This alignment process ensures that grey data, once fragmented or inconsistent, becomes part of a cohesive data ecosystem that enhances its value and usability.

White Data, already well-organized and structured, is further optimized through AI's integration capabilities. AI enhances the connectivity of white data across various business systems by ensuring consistency and real-time updates. By integrating white data seamlessly across platforms such as

CRM, enterprise resource planning (ERP), and business intelligence tools, AI maintains synchronization and accessibility. This automated, streamlined integration allows organizations to leverage their structured data for more efficient analytics, reporting, and decision-making processes, ensuring that data is consistently available where it is most needed.

Predictive Data Analysis. Involves using historical data, statistical algorithms, and machine learning techniques to forecast future outcomes, trends, or behaviors, enabling organizations to make data-driven decisions and optimize strategies proactively.

Dark Data, being unstructured and often unprocessed, presents significant challenges for predictive analysis. Without structure or context, AI models cannot accurately identify trends or predict outcomes. For dark data to be useful in predictive analysis, it first needs to be cleaned, organized, and transformed into a structured format. AI technologies such as natural language processing (NLP) and machine learning (ML) can help with this process, converting dark data into a format suitable for analysis. Once structured, AI can then begin to uncover hidden patterns or correlations that were previously inaccessible. However, the initial transformation process is crucial; otherwise, dark data remains unsuitable for predictive modeling.

Grey Data has a partial structure, making it somewhat ready for predictive analysis, but it often requires refinement. AI tools can fill gaps, standardize formats, and clean inconsistencies, enabling predictive models to operate more effectively. For example, AI can analyze semi-structured survey results or customer feedback to identify emerging trends and patterns. By refining grey data, AI enables the organization to derive meaningful insights and make informed predictions, though the process still requires attention to data quality and consistency. This predictive capability allows businesses to use grey data for strategic planning and performance evaluation, provided that the data is sufficiently processed and refined.

White Data, being fully structured and validated, is ideal for predictive analysis. AI enhances the capabilities of white data by integrating it into advanced predictive models, allowing for real-time insights and accurate forecasting. Whether it's financial data for economic forecasting or customer data for behavior prediction, AI utilizes structured white data to make precise and actionable predictions. This capability supports real-time decision-making and long-term strategic planning, helping organizations optimize operations, reduce risks, and capitalize on opportunities based on reliable forecasts and data-driven insights.

Data Privacy and Compliance Management. Involves implementing policies and practices to protect sensitive information, ensuring that data handling, storage, and processing adhere to legal regulations and industry standards to maintain security and user trust.

Dark Data, which is often unstructured and stored without much oversight, poses significant privacy and compliance risks. Hidden within email archives, server logs, or unprocessed customer feedback, this data may contain sensitive information such as personally identifiable information (PII) or financial details. AI tools equipped with natural language processing (NLP) and pattern recognition capabilities can scan these unstructured sources to identify and flag sensitive information. By tagging and classifying this data, AI provides organizations with visibility into their dark data, highlighting compliance risks and enabling further action, such as encryption, restricted access, or secure deletion, to protect privacy and align with regulations like GDPR or HIPAA.

Grey Data, with its partial structure, often includes data sets that may contain sensitive information but are not consistently managed. Examples include incomplete surveys, customer feedback, or semi-structured reports. AI helps manage compliance by identifying and securing sensitive elements within

these datasets, filling gaps, and applying data protection measures like encryption or anonymization. AI can also standardize grey data, ensuring that it meets compliance standards across all records. This automated approach not only protects sensitive information but also allows organizations to refine grey data into a more consistent and compliant format, reducing risks associated with regulatory violations.

White Data, which is fully structured and maintained, benefits from AI's ability to continuously monitor and ensure compliance automatically. AI systems can validate data in real-time, ensuring that it aligns with the latest regulations and industry standards. For instance, AI can check and verify that customer records are anonymized where necessary or that financial data adheres to specific compliance protocols. By automating compliance processes, AI minimizes human error and keeps white data secure, up-to-date, and legally compliant, enabling businesses to operate confidently within regulatory frameworks and build trust with customers and stakeholders.

Data Lifecycle Management. Is the process of managing data from creation to deletion, ensuring it is stored, accessed, and disposed of efficiently while maintaining compliance, security, and integrity throughout its lifecycle.

Dark Data typically consists of unstructured and outdated information stored across various systems without organization or management. AI can automate the archiving process for such data, scanning and identifying information that is no longer relevant or has become obsolete. By tagging this data and moving it to long-term storage or secure deletion systems, AI helps organizations minimize storage costs and reduce data bloat. Automated archiving also ensures that dark data does not consume unnecessary resources, while securely storing it for potential future use or regulatory compliance purposes. This approach streamlines the data management process, transforming dark data from an unmanaged burden into an efficiently stored resource.

Grey Data is more active and partially structured, requiring regular updates and maintenance to remain useful. AI tools can monitor grey data, identifying areas where information needs to be filled in, updated, or standardized. For instance, AI can automatically update incomplete records or align data formats to ensure consistency with existing frameworks. This proactive management keeps grey data relevant and ready for analysis, reducing the time and resources needed to refine and update it manually. By maintaining data consistency and quality, AI ensures grey data is consistently accessible and usable for business operations.

White Data, being fully structured and integrated into business processes, benefits from AI's ability to keep it up-to-date and optimized. AI continuously monitors and validates white data, applying any necessary updates or corrections in real-time. This ensures that the data remains accurate, relevant, and compliant with organizational standards. By automating these processes, AI maintains the quality of white data, allowing organizations to reliably use it for strategic planning, decision-making, and analytics.

Text and NLP Processing. Involves analyzing and manipulating written language using computational techniques. It encompasses tasks like text classification, sentiment analysis, language translation, and information extraction to enable machines to understand and generate human language.

Dark Data often includes vast amounts of unstructured text, such as server logs, email archives, and documents that remain unused due to their disorganized state. NLP tools powered by AI can scan these sources to identify relevant information, extract key phrases, and convert raw text into structured data formats like databases or spreadsheets. By automating this process, AI unlocks the

value hidden within dark data, turning it into a resource that can be analyzed for trends, patterns, and insights. For instance, AI can process historical email communications to identify common issues or customer preferences, making this previously inaccessible data actionable and valuable for the organization.

Grey Data typically consists of semi-structured documents, such as incomplete feedback forms, partially filled surveys, or preliminary reports. AI and NLP systems can efficiently process these documents, extracting and categorizing relevant information to fill in gaps or organize data more effectively. For example, NLP can analyze customer feedback to identify recurring themes or sentiments, even if the data is inconsistent or incomplete. By processing semi-structured text data, AI provides organizations with a clearer understanding of patterns and trends, enhancing the value and usability of grey data for targeted insights and decision-making.

White Data is already structured and may include validated reports, customer records, or other standardized documents. AI, through NLP, can further enhance the value of this structured data by extracting deeper insights. For instance, NLP algorithms can analyze large volumes of structured customer feedback, detecting subtle shifts in sentiment over time or identifying emerging topics relevant to business strategy. AI can also summarize and categorize information, making it easier for organizations to leverage white data for real-time analysis and strategic forecasting. This advanced processing ensures that structured text data continues to provide meaningful and up-to-date insights.

Real-Time Data Monitoring. Refers to the continuous observation and analysis of data as it is generated, providing immediate insights and updates, enabling timely responses and decision-making in dynamic, time-sensitive environments.

Dark Data is often unstructured and unmanaged, making real-time monitoring largely inapplicable unless the data is first processed and structured. Typically, dark data consists of archives like server logs, raw documents, or emails that have accumulated over time without proper organization. Before real-time monitoring can be effective, AI needs to clean and structure this data, converting it into a format suitable for analysis. Once this transformation is achieved, AI can begin to analyze and monitor patterns, detecting anomalies or important events that were previously hidden. However, without this foundational step, real-time monitoring is impractical for dark data, as its original state does not support continuous analysis.

Grey Data, which is partially structured, can be monitored for real-time relevance with AI intervention. This type of data may include preliminary survey responses, incomplete transaction records, or semi-processed reports. AI systems can monitor these sources, identifying updates or changes and processing them in near real-time. By filling in gaps and standardizing the data, AI makes it possible to extract and analyze information as it becomes available. For example, AI can track customer feedback updates or evolving market trends, providing organizations with timely insights that inform strategic decisions. While the grey data may require further refinement, AI monitoring ensures that the data remains relevant and actionable.

White Data is ideal for AI-driven real-time monitoring because it is already structured and validated, making it well-suited for instant analysis. AI systems continuously track and evaluate live data streams, such as financial transactions, sales figures, or sensor readings, and provide immediate feedback or alerts. This allows organizations to react quickly to opportunities, risks, or operational changes. By integrating white data into real-time monitoring systems, businesses can maximize their responsiveness and agility, ensuring that they are constantly aligned with the most current information available.

Automated Metadata Generation. Is the process of using software tools or algorithms to automatically create metadata, providing descriptions, tags, or categorization for data, documents, or digital assets, enhancing organization, searchability, and management.

Dark Data often remains unstructured and undocumented, making it difficult for organizations to utilize or even understand what they have stored. Examples include raw server logs, email archives, or unprocessed textual data. AI systems can scan and analyze these unstructured sources to automatically generate relevant metadata. Using natural language processing (NLP) and pattern recognition, AI identifies key characteristics such as date ranges, topics, and entities within these datasets. By creating metadata tags, descriptions, and classifications, AI transforms dark data into a more searchable and manageable resource. This automated process helps businesses organize vast amounts of dormant data, making it accessible for future use or analysis without the need for manual tagging.

Grey Data is often partially structured but may have incomplete or inconsistent documentation. Examples include draft reports, partially filled surveys, or customer feedback forms with missing fields. AI can identify these gaps and automatically fill in the missing metadata, ensuring that the data aligns more closely with organizational standards. For instance, AI can infer missing variables, categorize data points, and generate contextual tags that enhance data organization. By completing the metadata, AI helps convert grey data into a more consistent and usable format, facilitating easier integration into business systems and more effective analysis.

White Data is already well-organized and structured, but AI plays a crucial role in maintaining its metadata comprehensively and accurately. AI systems continuously monitor and update metadata for white data, ensuring it reflects the most current state of the data. For example, when customer records or financial reports are updated, AI automatically logs changes and updates metadata tags, preserving data consistency. This automation supports compliance with regulatory requirements and enhances the data's accessibility, making white data an efficient and reliable resource for decision-making and strategic planning.

Data Visualization and Reporting. Refers to the process of presenting data in graphical or visual formats, such as charts and dashboards, to make complex information easily understandable, enabling effective analysis, communication, and decision-making.

Dark Data often contains hidden patterns and insights that are difficult to uncover due to its unstructured nature. AI-powered visualization tools can process this raw data—such as email archives, server logs, or historical records—and transform it into structured visual formats like graphs, charts, and heatmaps. By identifying key trends or anomalies within unstructured data, AI helps organizations see correlations or issues that were previously invisible. For instance, AI can analyze email communication patterns to reveal organizational bottlenecks or common customer complaints. These visual representations make it possible for businesses to unlock the value in dark data, guiding decision-makers with newfound insights.

Grey Data, which is partially structured but often inconsistent or incomplete, also benefits significantly from AI-driven visualization. AI tools can analyze this data, filling in gaps and aligning it with existing frameworks, then presenting it in visual formats that highlight trends and patterns. For example, AI can aggregate survey data with missing responses to show overall trends while clearly indicating the gaps in information. Such visualizations provide a clearer understanding of the dataset's integrity, helping businesses identify areas that need refinement or further investigation. By converting grey data into actionable insights, AI makes it easier for organizations to develop targeted strategies and improvements.

White Data is well-structured and ready for advanced visualization. AI enhances this by creating real-time dashboards and interactive reports that provide strategic insights. Whether tracking financial performance, monitoring customer behavior, or analyzing sales figures, AI visualizes this data dynamically, allowing organizations to respond instantly to changes and trends. By offering a comprehensive, real-time view of critical metrics, AI ensures that white data remains a vital tool for decision-making, enabling leaders to stay informed and agile in a competitive business environment.

AI Capability	Dark Data	Grey Data	White Data
Data Discovery and Classification	AI scans and identifies unstructured data across storage systems.	AI classifies data with partial structure for better organization.	AI optimizes classification for efficient data management.
Data Cleaning and Structuring	AI cleans and structures unstructured data for usability.	AI refines and fills gaps to convert it into structured formats.	Maintains data quality and ensures consistency.
Automated Data Integration	AI integrates dark data with structured systems.	AI aligns grey data with existing data frameworks.	Enhances integration of white data across business systems.
Predictive Data Analysis	Limited application unless data is structured first.	Analyzes refined data for trends and predictive insights.	Enhances analysis for real-time insights and forecasting.
Data Privacy and Compliance Management	AI identifies and flags sensitive information in dark data.	AI manages compliance by securing partially sensitive data.	Ensures data is compliant with regulations automatically.
Data Lifecycle Management	Automates archiving of outdated dark data.	AI updates and maintains relevance of active grey data.	Keeps white data up-to-date and optimized for use.
Text and NLP Processing	Converts unstructured text (e.g., logs, emails) into structured data.	Processes semi-structured documents for data insights.	Enhances insights from structured text data sources.
Real-Time Data Monitoring	Not typically applicable until dark data is structured.	AI monitors semi-processed grey data for real-time relevance.	Monitors live data streams for instant analysis.
Automated Metadata Generation	AI generates metadata for unstructured and undocumented data.	Fills in gaps in documentation for partial datasets.	Maintains comprehensive metadata for structured data.
Data Visualization and Reporting	Transforms hidden patterns into visual insights.	Visualizes grey data gaps and trends for better clarity.	Provides real-time dashboards for strategic insights.

5. Dark Data, Grey Data and White Data in the Context of Warehouse Management

Inventory Accuracy. Refers to the degree to which the recorded inventory levels match the actual physical stock, ensuring precise tracking, minimizing discrepancies, and optimizing operational efficiency and decision-making.

Dark Data represents untapped potential, as it remains hidden and unused. This data is often collected and stored in large quantities but lacks structure, accessibility, or proper documentation, preventing it from being analyzed or integrated into business intelligence systems. Examples of dark data include outdated records, unstructured logs, emails, and archived documents, which sit idle in storage without contributing to organizational objectives. The lack of processing and governance means that any insights embedded in dark data are effectively lost, representing a missed opportunity for innovation, efficiency, and growth. To unlock its value, organizations would need to invest significantly in organizing, documenting, and processing this data—a costly and time-consuming endeavor that many companies hesitate to undertake.

Grey Data, on the other hand, offers some value but is not fully optimized or actionable. This type of data is partially structured and accessible, allowing organizations to extract insights, though these insights may be limited due to the data's inconsistencies or incomplete nature. Grey data might be suitable for smaller-scale projects or exploratory analyses, where partial information can still offer some guidance. However, due to its lack of comprehensive structure or documentation, the reliability

and depth of insights derived from grey data are often insufficient for making significant, strategic decisions. Organizations may need to preprocess or clean grey data extensively to maximize its utility, which can be time-consuming and costly.

White Data is highly valuable and forms a reliable foundation for generating insights. Fully structured, documented, and governed, white data is readily available for analysis, making it ideal for driving strategic planning, operational improvements, and market analysis. Its consistency and high quality allow organizations to build robust models, perform accurate forecasting, and make data-driven decisions with confidence. This maximizes the data's potential, transforming it into a critical asset that supports long-term growth and innovation.

Operational Efficiency. Refers to optimizing processes, resources, and workflows to minimize costs, reduce errors, and maximize productivity, ensuring accurate, timely storage and distribution of goods while maintaining high service levels and safety standards.

Dark Data is characterized by its unstructured and underutilized nature, making it challenging to enhance operational efficiency. In this context, historical machinery usage data might be collected but remains unprocessed and unstructured, rendering it unusable for any meaningful analysis or process optimization. Without converting this data into an accessible and structured format, organizations miss opportunities to identify patterns, predict maintenance needs, or optimize machinery usage. The inability to leverage this data results in inefficiencies, as decision-makers cannot access valuable insights that could inform improvements or reduce downtime.

Grey Data offers a middle ground, where some operational data is processed but not fully aligned or standardized, limiting its usefulness for streamlining processes. Semi-processed workflow logs, for instance, may capture daily operations and performance metrics, but without proper alignment or consistent structuring, this data remains difficult to interpret and apply effectively. While it can be used for preliminary analyses or smaller-scale optimizations, inconsistencies prevent it from providing a comprehensive view of operations. As a result, organizations must invest time and resources in further aligning and processing grey data to extract actionable insights that would lead to meaningful efficiency gains.

White Data, on the other hand, represents the ideal state for operational efficiency, where structured and consistent data is available and actively used. In this scenario, daily operations data is systematically collected, organized, and documented, making it easy to analyze and integrate into process optimization strategies. This structured approach enables organizations to monitor performance in real-time, identify bottlenecks, and implement data-driven improvements. With reliable and accurate data at their disposal, businesses can streamline workflows, predict maintenance needs, and optimize resource allocation, ultimately achieving higher productivity and cost savings. White data, thus, is essential for maximizing operational efficiency and supporting continuous improvement.

Demand forecasting. Involves predicting future inventory needs based on historical data, sales trends, and market analysis to optimize stock levels, minimize storage costs, and ensure efficient and timely product availability.

Dark Data in the context of demand forecasting represents untapped potential. Organizations often have past sales and shipment data that could be valuable for predicting future demand; however, this data remains dormant because it is unstructured, unorganized, or not integrated into analytical systems. Without being processed or cleaned, it is impossible to extract meaningful insights from this information. Consequently, the data sits unused, preventing companies from leveraging historical

trends to make accurate forecasts. This lack of action results in missed opportunities to align inventory levels with demand, optimize supply chains, or plan effectively for seasonal fluctuations.

Grey Data provides more opportunity but still presents challenges for accurate forecasting. In this case, sales data might be partially processed but inconsistent, with variations in how it is recorded or stored. Data cleansing becomes a crucial step to eliminate discrepancies, such as missing values or different formats, that could otherwise lead to inaccurate forecasts. Organizations relying on grey data can only produce moderately reliable demand projections unless they invest in the necessary preprocessing and alignment efforts. While the potential for demand forecasting is greater than with dark data, the inefficiencies in working with grey data can still hinder swift and precise decision-making, leading to suboptimal planning.

White Data is the ideal scenario for demand forecasting. When sales and shipment data are clean, consistent, and well-documented, they become highly valuable for generating precise, data-driven forecasts. Structured data allows for seamless integration into forecasting models, enabling businesses to predict demand trends accurately and respond proactively to changes in the market. This data-driven approach supports optimized inventory management, efficient resource allocation, and improved profitability. White data maximizes the potential of demand forecasting, transforming historical data into a strategic asset that drives business growth.

Space Optimization. Space optimization in a warehouse refers to efficiently utilizing available storage space through strategic layout planning, organization, and inventory management, maximizing capacity, and ensuring easy access while minimizing wasted space and operational inefficiencies.

Dark Data in the context of space optimization refers to historical storage data that, while potentially useful, remains unprocessed and inaccessible in its raw state. Organizations may have collected years' worth of information on storage usage and patterns, but without organizing and structuring this data, it cannot be effectively analyzed or applied. This lack of processing means that the data sits idle, preventing decision-makers from understanding past space utilization trends that could guide reorganization efforts. To unlock its value, companies would need to invest time and resources into structuring this dark data, making it suitable for analysis and space optimization planning. Until this processing occurs, opportunities for improving space efficiency remain untapped.

Grey Data represents a step up, as it is partially processed and documented, providing some insights into storage trends but still requiring further analysis. This type of data may highlight general patterns in how space is utilized, such as seasonal fluctuations or areas of underutilization; however, inconsistencies or gaps in the data limit its full potential. Organizations may need to clean and align grey data further to extract actionable insights. Though the data may support preliminary reorganization efforts, the lack of consistency and depth means that space optimization strategies based on grey data are likely to be less precise or comprehensive.

White Data is the most valuable for space optimization, as it consists of real-time, fully structured, and documented information on space usage. This type of data allows organizations to monitor space efficiency dynamically, adjusting storage and space allocations based on up-to-date information. With accurate and consistent data feeding into analytical models, businesses can maximize storage efficiency, minimize wasted space, and adapt quickly to changing needs. White data thus enables a proactive approach to space management, ensuring that organizations make the most of their available resources and continuously optimize their operational environments.

Labor Management. In warehouse management involves overseeing, optimizing, and coordinating workforce activities to ensure efficient operations, productivity, and resource allocation. It includes tracking performance, managing shifts, and aligning staffing levels with operational demands.

Dark Data in labor management consists of performance logs, attendance records, or other workforce-related information that remains unstructured and, consequently, underutilized. Organizations may have accumulated significant amounts of such data over time, but if it is not processed or organized systematically, it cannot be leveraged to optimize labor management. Unstructured data might be scattered across different systems or stored in formats that prevent efficient analysis. As a result, crucial insights—such as patterns in absenteeism, performance metrics, or peak demand periods—remain hidden, making it impossible to align staffing levels with business needs effectively. This unutilized data represents a missed opportunity to improve labor productivity and manage workforce costs more effectively.

Grey Data represents a step forward, where data, such as shift logs or employee records, is partially documented but remains incomplete or inconsistently formatted. While this data provides some insights, such as basic information on shift patterns or employee availability, it often requires further refinement and alignment for it to be genuinely useful. Inconsistencies in recording practices or gaps in documentation can lead to partial or inaccurate analyses, limiting the ability of managers to make informed decisions. For example, grey data might indicate general workforce availability but lack the detail necessary to adjust staffing dynamically in response to fluctuations in demand. This leads to inefficiencies in scheduling and resource allocation.

White Data, on the other hand, offers the ideal scenario for labor management. This fully structured and documented data provides comprehensive insights into employee performance, attendance, and staffing needs. By aligning workforce data with operational demand in real-time, organizations can optimize staffing levels, ensuring that they have the right number of employees with the necessary skills at the right time. This data-driven approach not only improves productivity but also enhances employee satisfaction by creating efficient and fair schedules. White data enables proactive labor management, minimizing labor costs and maximizing efficiency through precise, informed decision-making.

Order Fulfillment Accuracy. Refers to the precision with which a warehouse processes and delivers orders, ensuring that the correct products are picked, packed, and shipped accurately to meet customer requirements.

Dark Data in the context of order fulfillment comprises archived records of fulfillment errors that remain unstructured and, therefore, unusable for improving accuracy. Over time, companies may collect extensive data on issues such as incorrect shipments, delays, or inventory discrepancies, but if these records are not organized or structured, they cannot be analyzed for actionable insights. Without processing this data into a consistent format, businesses miss opportunities to identify patterns or root causes of errors. Consequently, these unresolved issues continue to disrupt order accuracy, preventing improvements in the fulfillment process. Investing in structuring this dark data could reveal valuable insights to reduce errors and improve efficiency, but until then, the potential remains untapped.

Grey Data represents an intermediate state where tracking information is collected but remains inconsistent and needs further processing. In this scenario, businesses may have data on shipments, inventory levels, and delivery timelines, but discrepancies in recording methods or incomplete logs create challenges in using this data effectively. This partial information can provide some understanding of fulfillment issues but requires additional work to clean, align, and standardize the

data for it to become actionable. Organizations relying on grey data may struggle with inconsistency, leading to delays or inaccuracies in order fulfillment and necessitating further refinement to improve operational effectiveness.

White Data is the optimal state for order fulfillment, as it involves real-time, structured information that significantly enhances accuracy and customer satisfaction. With comprehensive and up-to-date tracking of orders, inventory, and delivery processes, businesses can monitor and adjust their operations dynamically. This data-driven approach ensures that orders are fulfilled correctly and on time, reducing errors and improving customer experiences. By leveraging white data, companies can proactively manage their order fulfillment processes, continuously refining operations to meet customer expectations and maintain high levels of satisfaction.

Risk Management. Risk management in warehouse management involves identifying, assessing, and mitigating potential risks such as theft, damage, safety hazards, and operational disruptions to ensure the safety, efficiency, and security of inventory and operations.

Dark Data in risk management includes historical incident reports and other safety records that remain unprocessed and unstructured. Although this data holds significant potential for understanding past risks and preventing future incidents, it remains underutilized because it is stored in formats that are not easily accessible or analyzable. For example, old safety logs or incident records may be stored in paper archives or legacy systems without standardization, making it challenging to extract insights. Without processing and organizing this data, organizations miss opportunities to identify trends or recurring safety issues that could inform risk prevention strategies. Properly structuring and analyzing this dark data could provide valuable lessons and inform the development of comprehensive risk management plans, but until that happens, the data's value remains unrealized.

Grey Data is a step closer to being actionable but still presents challenges. In the case of incomplete or inconsistently filled safety check forms, organizations have some data on risk factors but lack the full picture needed for accurate risk assessments. This type of data may highlight some patterns or issues, but gaps and inconsistencies limit its effectiveness in drawing meaningful conclusions. Grey data often requires extensive cleaning and standardization to align it with current safety protocols and to be used in risk analysis models. While it provides more insight than dark data, it still falls short of enabling proactive risk management, requiring further refinement.

White Data represents the most valuable state for risk management, as it includes real-time, structured risk assessment data. This data enables organizations to monitor risks continuously, providing up-to-date information on safety incidents, hazard levels, and compliance. By integrating this real-time data into monitoring systems, organizations can proactively identify emerging risks and implement mitigation measures swiftly, reducing the likelihood of incidents. White data's accuracy and timeliness allow for informed decision-making and continuous improvement in risk management strategies, supporting a proactive, rather than reactive, approach to organizational safety.

Cost Management. In warehouse management involves monitoring, controlling, and optimizing expenses associated with storage, handling, and distribution of goods, aiming to improve efficiency, reduce costs, and maximize profitability within warehouse operations.

Dark Data in cost management refers to unstructured and unutilized expense records or logistics data. Organizations often accumulate vast amounts of financial information, such as invoices, supplier contracts, transportation costs, or maintenance expenses, but if this data is not organized or processed, it remains inaccessible for analysis. This lack of structure prevents companies from gaining insights into their cost patterns, inefficiencies, or opportunities for savings. Unstructured data, stored across

various systems and formats, often becomes siloed, making it challenging to consolidate or interpret. As a result, organizations miss out on valuable opportunities to optimize spending or negotiate better supplier terms, ultimately leading to wasted resources and inefficient cost control.

Grey Data represents a middle ground, where cost logs or financial records are collected and partially processed but lack consistency and standardization. Semi-processed data might provide some insight into organizational expenses, but discrepancies in logging practices or incomplete entries prevent comprehensive analysis. For instance, cost data may be recorded in different formats or systems, creating gaps and inconsistencies that complicate efforts to gain a clear understanding of total expenses. To make grey data actionable, organizations must invest time in aligning, cleaning, and standardizing it. This process can reveal important trends, but the incomplete nature of grey data still limits the precision and effectiveness of cost management strategies.

White Data is the most valuable for effective cost management, as it involves structured and consistent financial data. When expense records and logistics data are well-organized and fully documented, they provide a comprehensive view of all costs incurred. This data can be readily integrated into financial management systems, allowing organizations to monitor spending in real-time, identify cost-saving opportunities, and optimize resource allocation. White data enables precise, data-driven decision-making, ensuring that companies can manage their expenses efficiently and maintain financial health. By leveraging such structured data, organizations can implement targeted strategies that enhance cost control and profitability.

Compliance and Auditing. Involves ensuring warehouse operations adhere to regulations and standards, systematically documenting activities, and conducting regular audits to verify accuracy, safety, and compliance with legal and organizational policies.

Dark Data in the realm of compliance and auditing includes historical logs, records, and compliance documents that remain unstructured and inaccessible. Organizations often collect large volumes of compliance-related information, such as policy adherence records, incident reports, and regulatory filings, but these remain scattered and unorganized across various databases or archives. Without structuring and centralizing this data, it becomes difficult to access or review, making compliance audits time-consuming and inefficient. Moreover, the inability to extract or analyze information from this unstructured data increases the risk of non-compliance, as potential issues remain undetected. By not utilizing this data effectively, organizations expose themselves to potential legal and regulatory risks and reduce their ability to respond proactively.

Grey Data represents a partially organized state where compliance documentation exists but is incomplete or inconsistently maintained. This data may include records of inspections, compliance checklists, or training logs that are partially processed but not standardized. While some information is accessible, inconsistencies and gaps make it challenging to form a comprehensive view of compliance. Incomplete data creates uncertainty, making it difficult to identify potential areas of non-compliance or prepare for audits confidently. To make grey data useful, further processing is necessary to fill in missing information, align formats, and verify records, all of which require significant effort and resources.

White Data is the optimal state for compliance and auditing, consisting of fully structured and well-maintained compliance records. Properly documented and centralized, this data provides a clear and complete view of the organization's adherence to policies and regulations. It is regularly updated and audited to ensure accuracy, making it easy to access and review for both internal and external audits. With white data, organizations can quickly demonstrate compliance, minimizing legal risks and maintaining audit readiness. This proactive approach not only supports regulatory adherence but also

enhances the organization's reputation and operational integrity, ensuring smooth, efficient compliance processes.

Customer satisfaction. Customer satisfaction in warehouse management refers to the fulfillment of customer expectations through efficient storage, accurate order fulfillment, timely deliveries, and effective inventory management, ensuring a seamless experience and boosting overall satisfaction with services.

Dark Data in customer satisfaction consists of unused feedback logs, such as customer surveys, complaint records, or service call logs that remain unstructured and inaccessible. These logs often sit in various formats—emails, handwritten notes, or unprocessed digital entries—making it difficult to aggregate and analyze them meaningfully. Without proper structuring, this feedback cannot be used to identify recurring issues or understand customer needs, preventing the organization from acting on valuable insights. Consequently, dark data represents a missed opportunity to address customer concerns, improve service offerings, and enhance overall satisfaction. Investing in structuring and processing this data could unlock its potential, but until then, its value remains untapped, and customer dissatisfaction may persist due to unresolved issues.

Grey Data offers partial visibility into customer satisfaction trends, as it includes partially collected and somewhat organized feedback data. This might consist of online surveys, social media comments, or product reviews that are collected but inconsistently documented or maintained. While grey data can reveal general patterns and insights, such as common areas of satisfaction or dissatisfaction, it often lacks completeness and uniformity. The gaps and inconsistencies within grey data make it challenging to fully understand customer sentiment or take decisive actions. Organizations need to invest time and resources into filling these gaps, standardizing formats, and ensuring that feedback collection is consistent to create a complete picture of customer experiences.

White Data is the most valuable state, where customer feedback is fully structured, consistent, and readily accessible. This includes comprehensive records from multiple sources, such as surveys, customer support interactions, and social media, all aggregated in a standardized manner. White data provides a clear and accurate view of customer sentiment, allowing organizations to quickly identify trends, track customer needs, and respond proactively to improve service quality. By leveraging white data, businesses can implement targeted strategies to enhance customer satisfaction, build loyalty, and drive growth, ensuring that customer insights translate into meaningful action and improvement.

Aspect	Dark Data	Grey Data	White Data
Inventory Accuracy	Unused sensor logs and shipment records remain untapped unless processed.	Incomplete order logs may lead to stock inaccuracies, needing refinement.	Real-time inventory data ensures precise stock tracking and minimal errors.
Operational Efficiency	Historical machinery usage data is unutilized without structuring.	Semi-processed workflow logs need alignment for process optimization.	Structured daily operations data helps streamline and optimize processes.
Demand Forecasting	Past sales and shipment data have forecasting potential but remain dormant.	Inconsistent sales data needs cleansing for accurate forecasting.	Clean sales data enables precise, data-driven demand forecasts.
Space Optimization	Historical storage data guides reorganization but requires processing.	Partially documented data shows trends but needs further analysis.	Real-time space usage data maximizes storage efficiency dynamically.
Labor Management	Performance logs or attendance records are underutilized if unstructured.	Incomplete shift logs provide limited insights, needing refinement.	Structured labor data ensures optimal staffing based on demand.
Order Fulfillment Accuracy	Archived fulfillment error records need structuring for actionable insights.	Inconsistent tracking information needs processing to resolve challenges.	Real-time order data improves accuracy and customer satisfaction.

Risk Management	Historical incident reports hold value but remain unused without processing.	Incomplete safety check forms provide limited risk insights.	Real-time risk assessment data enables proactive monitoring and mitigation.
Cost Management	Expense records/logistics data remain unstructured and unusable.	Semi-processed cost logs need consistency for effective management.	Structured financial data provides insights for efficient cost control.
Compliance and Auditing	Historical compliance logs need structuring for access and review.	Incomplete documentation creates gaps and requires further processing.	Properly maintained compliance data minimizes legal risks and ensures audit readiness.
Customer Satisfaction	Unused feedback logs remain inaccessible without structuring.	Partially collected feedback shows trends but needs completion.	Structured customer data supports improved service quality and satisfaction.

6. Conclusions

In conclusion, the taxonomy of dark, grey, and white data provides a structured framework for understanding how organizations can manage and leverage different types of data. Each type of data carries its unique characteristics, implications, and potential value, reflecting the complexity of modern data environments. The categorization offers an analytical perspective that organizations can use to optimize their data governance strategies and transform unused information into actionable insights. Dark data represents the large quantities of information that organizations collect but fail to utilize. This data remains hidden due to various factors such as unstructured formats, legacy systems, and the absence of adequate processing tools or organizational awareness. Despite its vast potential, dark data poses risks including compliance issues, security vulnerabilities, and increased storage costs. The challenge with dark data lies in its unprocessed and unstructured nature, which prevents it from being easily integrated into business intelligence systems or strategic decision-making processes. However, when organizations invest in structuring and processing dark data, its latent value can be unlocked. The potential insights derived from analyzing historical records, sensor data, and other previously dormant sources can lead to improvements in areas like risk management, operational efficiency, and market understanding. This transition requires a systematic approach involving advanced technologies like artificial intelligence (AI), cloud computing, and data analytics frameworks designed to classify, process, and secure data. By effectively managing dark data, organizations can convert what was once seen as a liability into a strategic asset that enhances decision-making and organizational resilience. Grey data occupies an intermediate space between structured and unstructured information. It often includes semi-structured data like emails, social media interactions, meeting notes, or workflow logs that are collected but not fully optimized for use. Grey data holds significant potential, but its inconsistent documentation and formatting present challenges for analysis. This data type is partially accessible and provides some insights but often requires further processing to be integrated into organizational practices effectively. The management of grey data involves addressing inconsistencies and filling in information gaps. By investing in data alignment, cleansing, and standardization, organizations can make grey data more actionable. Tools such as machine learning algorithms and data visualization techniques can be particularly effective in this process, helping businesses to identify patterns and emerging trends. If integrated properly, grey data can support better forecasting models, enhance customer satisfaction analysis, and optimize processes such as labor management. Organizations that develop strategies to transform grey data into structured, actionable information gain a critical advantage, as they tap into data resources that are often overlooked. White data represents the pinnacle of effective data management. This type of data is fully structured, documented, and readily accessible, enabling organizations to leverage it for strategic decision-making, operational efficiencies, and customer satisfaction. White data includes organized customer databases, compliance records, and performance logs that are consistently updated and maintained according to industry standards and regulatory requirements. The availability and accuracy of this data make it a foundational asset for business intelligence systems. The use of

white data ensures that organizations can make informed decisions quickly and accurately. For instance, structured sales data supports precise demand forecasting, while comprehensive customer feedback records enable targeted service improvements. Furthermore, white data plays a critical role in compliance and auditing processes, as its well-documented nature guarantees audit readiness and minimizes legal risks. By maintaining a high standard of data governance, organizations can fully capitalize on the value of white data, using it as a tool for achieving long-term growth and operational excellence. The taxonomy highlights the importance of integrating and managing diverse data types. Organizations cannot solely rely on white data, as grey and dark data hold significant untapped value. A comprehensive data strategy requires businesses to bridge the gap between these data forms, transforming dark and grey data into structured, usable information that complements existing white data assets. This transformation process is central to creating a data-driven organization that maximizes the value of its information resources. Technological advancements, such as AI, machine learning, and cloud services, play a crucial role in this integration. These technologies offer scalable solutions that enable organizations to analyze large volumes of unstructured data efficiently, uncovering insights from grey and dark data that can be integrated into decision-making processes. Moreover, investing in data governance frameworks ensures that all types of data are managed securely and efficiently, minimizing risks while maximizing the potential for value extraction. The conclusions drawn from this taxonomy emphasize the need for a proactive, systematic approach to data management. Organizations must develop governance strategies that encompass all three data types, ensuring that dark, grey, and white data are managed within a cohesive framework. Such strategies should prioritize data quality, consistency, and security, enabling businesses to make data-driven decisions that are reliable and compliant with regulations. Data governance frameworks should also focus on promoting data transparency and accessibility. By structuring and documenting grey and dark data effectively, organizations can not only enhance their compliance capabilities but also unlock new avenues for business intelligence and innovation. This proactive stance transforms data into a central pillar of organizational strategy, shifting the view of data management from a technical challenge to a strategic enabler. In conclusion, the structured approach to categorizing and managing data as white, grey, and dark offers a valuable framework for organizations aiming to optimize their data resources. While white data provides the ideal state for structured information, dark and grey data offer significant but often unrealized potential. By investing in the tools, technologies, and strategies needed to transform these data types into structured, usable forms, organizations can gain a competitive edge, improve operational efficiencies, and drive long-term growth. The evolution from dark and grey to white data represents not just a technical journey but a strategic transformation. Organizations that adopt a holistic view of data management, integrating technological solutions with strong governance practices, will be better positioned to navigate the complexities of the modern data environment. Ultimately, the taxonomy of dark, grey, and white data serves as both a guide and a call to action for organizations seeking to maximize the value of their information assets in an increasingly data-centric world.

7. References

Aggarwal, S., & Singh, R. (2020). Visual Analytics on Biomedical Dark Data.

Aggarwal, S., & Singh, R. (2020). Visual Exploration and Knowledge Discovery from Biomedical Dark Data. arXiv preprint arXiv:2009.13059.

Ahlawat, P., Borgman, J., Eden, S., Huels, S., Iandiorio, J., Kumar, A., & Zakahi, P. (2023). A new architecture to manage data costs and complexity. Boston Consulting Group (BCG), 1-12.

Ahmadi, R., Ekbatanifard, G., & Bayat, P. (2021). A modified grey wolf optimizer based data clustering algorithm. Applied Artificial Intelligence, 35(1), 63-79.

Ahmed, M. B., & Verma, M. R. (2024). DARK WEB DATA CLASSIFICATION USING NEURAL NETWORK. PARADIGM SHIFT: MULTIDISCIPLINARY RESEARCH FOR A CHANGING WORLD, VOLUME-2, 124.

Ajis, A. F. M., Ibrahim, A. A. A., & Verma, M. K. (2024, August). Elucidating Theory of Malaysian Data Crisis in Demystifying Dark Data Catalyst. In 2024 IEEE 6th Symposium on Computers & Informatics (ISCI) (pp. 30-36). IEEE.

Ajis, A. F. M., Jali, J. M., Ishak, I., & Harun, Q. N. (2023). Enlightening the Repercussion of Dark Data Management towards Malaysian SMEs Sustainability. Environment-Behaviour Proceedings Journal, 8(SI15), 223-229.

Akbar, L. S., Al-Mutahr, K., & Nazeh, M. (2018). Aligning IS/IT with Business Allows Organizations to Utilize Dark Data. International Journal of Innovative Technology and Exploring Engineering, 8(2), 80-85.

Almeida, C. A., Torres-Espin, A., Huie, J. R., Sun, D., Noble-Haeusslein, L. J., Young, W., ... & Ferguson, A. R. (2022). Excavating FAIR data: the case of the Multicenter Animal Spinal Cord Injury Study (MASCIS), blood pressure, and neuro-recovery. Neuroinformatics, 1-14.

Al-Refaie, A. (2010). Grey-data envelopment analysis approach for solving the multi-response problem in the Taguchi method. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 224(1), 147-158.

Baghery, M., Yousefi, S., & Rezaee, M. J. (2018). Risk measurement and prioritization of auto parts manufacturing processes based on process failure analysis, interval data envelopment analysis and grey relational analysis. Journal of Intelligent Manufacturing, 29(8), 1803-1825.

Banafa, A. (2022). 13 Understanding Dark Data.

Banafa, A. (2022). Part 3 Big Data, Dark Data, Thick Data, and Small Data.

Benvenuti, D. (2023, June). Towards a Framework for Data Pipeline Discovery. In Companion of the 2023 International Conference on Management of Data (pp. 293-294).

Bhatia, S., & Alojail, M. (2022). A Novel Approach for Deciphering Big Data Value Using Dark Data. Intelligent Automation & Soft Computing, 33(2).

Bin, S., Ping, Y., Yunbai, L., & Xishan, W. (2002, October). Study on the fault diagnosis of transformer based on the grey relational analysis. In Proceedings. International Conference on Power System Technology (Vol. 4, pp. 2231-2234). IEEE.

Chakrabarty, S., & Joshi, R. S. (2020). Dark Data: People to People Recovery. In ICT Analysis and Applications: Proceedings of ICT4SD 2019, Volume 2 (pp. 247-254). Springer Singapore.

Chan, S., Oktavianti, I., & Nopphawan, P. (2020, October). PMU Time Series Module Adapted for Reduction of Dark Data and the Ensuing Enhanced Analytics for Higher Quality Yields of Ethanol Fuel Production. In 2020 8th International Conference on Condition Monitoring and Diagnosis (CMD) (pp. 412-415). IEEE.

Chang, K. C., & Yeh, M. F. (2005). Grey relational analysis based approach for data clustering. IEE Proceedings-Vision, Image and Signal Processing, 152(2), 165-172.

Chang, T. C., & Lin, S. J. (1999). Grey relation analysis of carbon dioxide emissions from industrial production and energy uses in Taiwan. Journal of Environmental Management, 56(4), 247-257.

Chant, G. G. (2023, May). Dealing with Dark Data–Shining a Light. In International Conference on Knowledge Management in Organizations (pp. 149-160). Cham: Springer Nature Switzerland.

Chaudhari, A. A., & Pund, M. A. (2020). Visualization of Uncertainties and Noise in Dark Data: Methods & Techniques. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 11(3), 2076-2083.

Choi, H. O. (2021). A study on identifying the policy demand by innovators using dark data-Daedeok Innopolis using civil complaint. Journal of Digital Contents Society, 22(10), 1645-1652.

da Costa, T., & Barrett, M. (2021, April). Improving Cathodic Protection Monitoring Data in the Time of IIoT and Big Data. In NACE CORROSION (p. D051S026R001). NACE.

Dimitrov, W., & Chikalanov, A. (2016). Dark Data Governance Reduces Security Risks. Big Data, Knowledge and Control Systems Engineering, 81.

Dimitrov, W., Сярова, С., & Petkova, L. (2018). Types of dark data and hidden cybersecurity risks. Epub ahead of print. doi, 10.

Dormann, C. F., Purschke, O., Marquez, J. R. G., Lautenbach, S., & Schroeder, B. (2008). Components of uncertainty in species distribution analysis: a case study of the great grey shrike. Ecology, 89(12), 3371-3386.

Faghih, N., Bonyadi, E., & Sarreshtehdari, L. (2021). Entrepreneurial Motivation Index: importance of dark data. Journal of Global Entrepreneurship Research, 1-13.

Feng, J. C., Huang, H. A., Yin, Y., & Zhang, K. (2019). Comprehensive security risk factor identification for small reservoirs with heterogeneous data based on grey relational analysis model. Water Science and Engineering, 12(4), 330-338.

Forker, E. (2023). The Informativeness of Dark Data for Future Firm Performance.

Fu, C., Zheng, J., Zhao, J., & Xu, W. (2001). Application of grey relational analysis for corrosion failure of oil tubes. Corrosion Science, 43(5), 881-889.

Gautam, A. (2023). Navigating the Risks of Dark Data: An Investigation into Personal Safety.

Ge, Z. (2022). Artificial Intelligence and Machine Learning in Data Management. Future And Fintech, The: Abcdi And Beyond, 281.

George, A. S., Sujatha, V., George, A. H., & Baskar, T. (2023). Bringing Light to Dark Data: A Framework for Unlocking Hidden Business Value. Partners Universal International Innovation Journal, 1(4), 35-60.

George, D. A. S., Sujatha, D. V., George, A. H., & Baskar, D. T. (2023). Bringing Light to Dark Data: A Framework for Unlocking Hidden Business Value. Partners Universal International Innovation Journal, 1 (4), 35–60.

Gianna, D. A. (2021). Dark Data Risk Mitigation in Big IoT Data (Doctoral dissertation, Capitol Technology University).

Giest, S., & Samuels, A. (2020). 'For good measure': data gaps in a big data world. Policy Sciences, 53(3), 559-569.

Gimpel, G. (2020). Bringing dark data into the light: Illuminating existing IoT data lost within your organization. Business Horizons, 63(4), 519-530.

Gimpel, G. (2021). Dark data: the invisible resource that can drive performance now. Journal of Business Strategy, 42(4), 223-232.

Gimpel, G., & Alter, A. (2021). Benefit from the internet of things right now by accessing dark data. IT Professional, 23(2), 45-49.

Goyal, S., & Grover, S. (2012). Applying fuzzy grey relational analysis for ranking the advanced manufacturing systems. Grey Systems: Theory and Application, 2(2), 284-298.

Guo, H., Deng, S., Yang, J., Liu, J., & Nie, C. (2020). Analysis and prediction of industrial energy conservation in underdeveloped regions of China using a data pre-processing grey model. Energy policy, 139, 111244.

Hajiagha, S. H. R., Zavadskas, E. K., & Hashemi, S. S. (2013). Application of stepwise data envelopment analysis and grey incidence analysis to evaluate the effectiveness of export promotion programs. Journal of business economics and management, 14(3), 638-650.

Hampton, P. (2020). Keeping a secure hold on data through modern electronic content management. Network Security, 2020(6), 8-11.

Han, M., Zhang, R., Qiu, T., Xu, M., & Ren, W. (2017). Multivariate chaotic time series prediction based on improved grey relational analysis. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(10), 2144-2154.

Hand, D. J. (2020). Dark data: Why what you don't know matters. Princeton University Press.

Hawkins, B. E., Huie, J. R., Almeida, C., Chen, J., & Ferguson, A. R. (2020). Data dissemination: shortening the long tail of traumatic brain injury dark data. Journal of neurotrauma, 37(22), 2414-2423.

Hobart, M. (2020). The 'dark data'conundrum. Computer Fraud & Security, 2020(7), 13-16.

Hsia, K. H., Chen, M. Y., & Chang, M. C. (2004). Comments on data pre-processing for grey relational analysis. Journal of Grey System, 7(1), 15-20.

Hsu, L. C., & Wang, C. H. (2007). Forecasting the output of integrated circuit industry using a grey model improved by the Bayesian analysis. Technological Forecasting and Social Change, 74(6), 843-853.

Hu, Y. C. (2020). A multivariate grey prediction model with grey relational analysis for bankruptcy prediction problems. Soft Computing, 24(6), 4259-4268.

Hu, Y. C. (2020). Constructing grey prediction models using grey relational analysis and neural networks for magnesium material demand forecasting. Applied Soft Computing, 93, 106398.

Huang, J., & Sun, H. (2016, August). Grey relational analysis based k nearest neighbor missing data imputation for software quality datasets. In 2016 IEEE International Conference on Software Quality, Reliability and Security (QRS) (pp. 86-91). IEEE.

Huang, M., & Wang, B. (2016). Factors influencing CO2 emissions in China based on grey relational analysis. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 38(4), 555-561.

Huang, Y., Shen, L., & Liu, H. (2019). Grey relational analysis, principal component analysis and forecasting of carbon emissions based on long short-term memory in China. Journal of Cleaner Production, 209, 415-423.

Ikram, M., Sroufe, R., Rehman, E., Shah, S. Z. A., & Mahmoudi, A. (2020). Do quality, environmental, and social (QES) certifications improve international trade? A comparative grey relation analysis of developing vs. developed countries. Physica A: Statistical Mechanics and its Applications, 545, 123486.

Ingólfsdóttir, I. K. (2023). Dark data disclosure: defence against disinformation (Doctoral dissertation).

Jackson, T., & Hodgkinson, I. R. (2022). 'Dark data'is killing the planet–we need digital decarbonisation. The Conservation online.

Javed, S. A., & Liu, S. (2018). Evaluation of outpatient satisfaction and service quality of Pakistani healthcare projects: application of a novel synthetic grey incidence analysis model. Grey Systems: Theory and Application, 8(4), 462-480.

Javed, S. A., Liu, S., Mahmoudi, A., & Nawaz, M. (2019). Patients' satisfaction and public and private sectors' health care service quality in Pakistan: Application of grey decision analysis approaches. The International journal of health planning and management, 34(1), e168-e182.

Jha, A., & Singh, S. R. (2022). Human-Machine Convergence and Disruption of Socio-Cognitive Capabilities. International Journal of Next-Generation Computing, 13(3).

Jin, X., Xu, X., Song, X., Li, Z., Wang, J., & Guo, W. (2013). Estimation of leaf water content in winter wheat using grey relational analysis–partial least squares modeling with hyperspectral data. Agronomy Journal, 105(5), 1385-1392.

Jolliffe, I. (2021). Dark data: Why what you don't know matters.

Keller-Fröhlich, M. (2022, November). IMPACT OF DARK DATA ON THE VALUE DRIVEN DATA MANAGEMENT STRATEGY OF MANUFACTURERS–A LITERATURE REVIEW. In 26th European Scientific Conference of Doctoral Students (p. 59).

Kim, H. S. (2024). Dark Data in Real-World Evidence: Challenges, Implications, and the Imperative of Data Literacy in Medical Research. Journal of Korean Medical Science, 39(9).

Kucukonder, H., Demirarslan, P. C., Burgut, A., & Boga, M. (2019). A hybrid approach of data envelopment analysis based grey relational analysis: a study on egg yield.

Lager, O. (2021). Dark Data: Characteristics, steps to utilize it & factors influencing its utilization (Master's thesis).

Lee, Y. T. (2016). Principle study of head meridian acupoint massage to stress release via grey data model analysis. Evidence-Based Complementary and Alternative Medicine, 2016(1), 4943204.

Li, X., Hipel, K. W., & Dang, Y. (2015). An improved grey relational analysis approach for panel data clustering. Expert Systems with Applications, 42(23), 9105-9116.

Lin, C. H. (2008). Frequency-domain features for ECG beat discrimination using grey relational analysis-based classifier. Computers & Mathematics with Applications, 55(4), 680-690.

Liu, S., Lin, C., Tao, L., Javed, S. A., Fang, Z., & Yang, Y. (2020). On Spectral Analysis and New Research Directions in Grey System Theory. Journal of Grey System, 32(1).

Liu, S., Yang, Y., & Forrest, J. (2017). Grey data analysis. Springer Singapore, Singapore, Doi, 10(1007), 978-981.

Liu, X., Liu, H., Zhao, X., Han, Z., Cui, Y., & Yu, M. (2022). A novel neural network and grey correlation analysis method for computation of the heat transfer limit of a loop heat pipe (LHP). Energy, 259, 124830.

Liu, Y., Du, J. L., Zhang, R. S., & Forrest, J. Y. L. (2019). Three way decisions based grey incidence analysis clustering approach for panel data and its application. Kybernetes, 48(9), 2117-2137.

Liu, Y., Wang, Y., Gao, L., Guo, C., Xie, Y., & Xiao, Z. (2021). Deep hash-based relevance-aware data quality assessment for image dark data. ACM/IMS Transactions on Data Science, 2(2), 1-26.

Liu, Y., Wang, Y., Zhou, K., Yang, Y., Liu, Y., Song, J., & Xiao, Z. (2019). A framework for image dark data assessment. In Web and Big Data: Third International Joint Conference, APWeb-WAIM 2019, Chengdu, China, August 1–3, 2019, Proceedings, Part I 3 (pp. 3-18). Springer International Publishing.

Lovato, J., & Zimmerman, J. (2021). Dark Data: making dark data FAIR.

Maju, S. V., & Gnana Prakasi, O. S. (2022). Utilization of Dark Data from Electronic Health Records for the Early Detection of Alzheimer's Disease. In Recent Advances in Artificial Intelligence and Data Engineering: Select Proceedings of AIDE 2020 (pp. 195-203). Springer Singapore.

Maju, S. V., & Prakasi, O. S. (2022). Design of a Decision Making Model for Integrating Dark Data from Hybrid Sectors. Grenze International Journal of Engineering & Technology (GIJET), 8(1).

Matanović, A. J., Bošnjak, M., & Sremac, J. (2022, July). Testing the validity of" dark data" on the Late Miocene freshwater cockles housed in the CNHM. In Mathematical methods and terminology in geology 2022.

Md Ajis, A. F. (2023). Dark data management among Malaysian small and medium enterprises: a grounded theory analysis (Doctoral dissertation, Universiti Teknologi MARA (UiTM)).

Meil, J. (2021, April). Programmatic Labeling of Dark Data for Artificial Intelligence in Spatial Informatics. In EGU General Assembly Conference Abstracts (pp. EGU21-16326).

Mohr, J. J., Adams, D., Barkhouse, W., Beldica, C., Bertin, E., Cai, Y. D., ... & Stoughton, C. (2008, July). The dark energy survey data management system. In Observatory Operations: Strategies, Processes, and Systems II (Vol. 7016, pp. 176-191). SPIE.

Moumeni, L., Slimani, I., El Farissi, I., Saber, M., & Belkasmi, M. G. (2021, June). Dark data as a new challenge to improve business performances: review and perspectives. In 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA) (pp. 216-220). IEEE.

Munot, K., Mehta, N., Mishra, S., & Khanna, B. (2019, March). Importance of dark data and its applications. In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-6). IEEE.

Murphy, C., & Thomas, F. P. (2024). Illuminating dark data: Advancing spinal cord medicine through reporting on "negative" data. The Journal of Spinal Cord Medicine, 47(1), 1-2.

Neha, & Pahwa, P. (2020). Dark Data Analytics Using Blockchain Technology. In Advances in Data Sciences, Security and Applications: Proceedings of ICDSSA 2019 (pp. 467-474). Springer Singapore.

Ng, D. K. (1994). Grey system and grey relational model. ACM SIGICE Bulletin, 20(2), 2-9.

Nguyen, P. H., Sen, S., Jourdan, N., Cassoli, B., Myrseth, P., Armendia, M., & Myklebust, O. (2022). Software engineering and AI for data quality in cyber-physical systems-sea4dq'21 workshop report. ACM SIGSOFT Software Engineering Notes, 47(1), 26-29.

Niu, B., Shi, M., Zhang, Z., Li, Y., Cao, Y., & Pan, S. (2022). Multi-objective optimization of supply air jet enhancing airflow uniformity in data center with Taguchi-based grey relational analysis. Building and Environment, 208, 108606.

Pakkar, M. S. (2016). An integrated approach to grey relational analysis, analytic hierarchy process and data envelopment analysis. Journal of Centrum Cathedra, 9(1), 71-86.

Pawlewitz, J., Mankel, A., Jacquin, S., & Basile, N. (2020, May). The digital twin in a brownfield environment: How to manage dark data. In Offshore Technology Conference (p. D021S018R002). OTC.

Perini, D. J., Batarseh, F. A., Tolman, A., Anuga, A., & Nguyen, M. (2023). Bringing dark data to light with AI for evidence-based policymaking. In AI Assurance (pp. 531-557). Academic Press.

Priya, S., Vidyapeeth, L. S., & Mahajan, S. (2022). Unveiling the Silver Lining of Dark Data for Organizations. Amity Journal of Strategic Management. Vol.-05, Issue-02. July-Dec., 2022

Purss, M. B., Lewis, A., Oliver, S., Ip, A., Sixsmith, J., Evans, B., ... & Chan, T. (2015). Unlocking the Australian landsat archive–from dark data to high performance data infrastructures. GeoResJ, 6, 135-140.

Raca, K. (2021). Enterprise Dark Data. In Data Analysis and Classification: Methods and Applications 29 (pp. 119-131). Springer International Publishing.

Rajesh, R. (2024). Grey models for data analysis and decision-making in uncertainty during pandemics. International Journal of Disaster Risk Reduction, 104881.

Ravindranathan, P., Ashok, P., & Prabhu, S. (2024, January). Illuminating the Dark: Gaining Insights and Managing Risks with Dark Analytics. In 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) (pp. 1-6). IEEE.

Roman, D., Prodan, R., Nikolov, N., Soylu, A., Matskin, M., Marrella, A., ... & Kharlamov, E. (2022). Big data pipelines on the computing continuum: tapping the dark data. Computer, 55(11), 74-84.

Ross, S. J. (2021). Afraid of the Dark (Data). ISACA Journal, (6).

Sallehuddin, R., Shamsuddin, S. M. H., & Hashim, S. Z. M. (2008, November). Application of grey relational analysis for multivariate time series. In 2008 Eighth International Conference on Intelligent Systems Design and Applications (Vol. 2, pp. 432-437). IEEE.

Schembera, B. (2021). Like a rainbow in the dark: metadata annotation for HPC applications in the age of dark data. The Journal of Supercomputing, 77(8), 8946-8966.

Schembera, B., & Durán, J. M. (2020). Dark data as the new challenge for big data science and the introduction of the scientific data officer. Philosophy & Technology, 33, 93-115.

Seki, K., Takamichi, S., Saeki, T., & Saruwatari, H. (2023, June). Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Shah, M., Malik, F., Suliman, M., Rahman, N., Ullah, I., Ullah, S., ... & Alam, S. (2024). Dark Data in Accident Prediction: Using AdaBoost and Random Forest for Improved Accuracy. Journal of Computing & Biomedical Informatics, 7(02).

Shave, L. (2023). The need for lifelong learning. IQ: The RIMPA Quarterly Magazine, 39(4), 26-30.

Shetty, S. (2021). How to tackle dark data. Available via Gartner: https://www. gartner. com/sma rterwithgartner/how-to-tackle-dark-data. Accessed, 20.

Shimizu, N., Ueno, O., & Komata, C. (1998, April). Introduction of time series data analysis using grey system theory. In 1998 Second International Conference. Knowledge-Based Intelligent Electronic Systems. Proceedings KES'98 (Cat. No. 98EX111) (Vol. 2, pp. 67-72). IEEE.

Shin, S. I., & Kwon, M. M. (2023). Dark data: Why What You Don't Know Matters: Dark Data: Why What You Don't Know Matters, by David J. Hand, New Jersey, US, Princeton University Press, 2020, 330 pp., $29.95 (hardback), ISBN: 9780691182377.

Singh, J., Upadhyay, D., Singh, M., & Sagar, P. (2021). Towards a Comparative Analysis of Regression Based Machine Learning Techniques. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC).

Singh, P. K. (2021). Dark data analysis using Intuitionistic Plithogenic graphs. International Journal of Neutrosophic Sciences, 16(2), 80-100.

Škrinjarić, T., & Šego, B. (2021). Evaluating business performance using data envelopment analysis and grey relational analysis. In Handbook of Research on Engineering, Business, and Healthcare Applications of Data Science and Analytics (pp. 115-148). IGI Global.

Slimani, I., Slimani, N., Achchab, S., Saber, M., El Farissi, I., Sbiti, N., & Amghar, M. (2022). Automated machine learning: the new data science challenge. Int. J. Electr. Comput. Eng., 12(4), 4243-4252.

Song, Q., & Shepperd, M. (2011). Predicting software project effort: A grey relational analysis based method. Expert Systems with Applications, 38(6), 7302-7316.

Song, Q., Shepperd, M., & Mair, C. (2005, September). Using grey relational analysis to predict software effort with small data sets. In 11th IEEE International Software Metrics Symposium (METRICS'05) (pp. 10-pp). IEEE.

Stahlman, G. R. (2020). Exploring the long tail of astronomy: A mixed-methods approach to searching for dark data (Doctoral dissertation, The University of Arizona).

Stahlman, G., Heidorn, P. B., & Steffen, J. (2018). The astrolabe project: identifying and curating astronomical 'dark data'through development of cyberinfrastructure resources. In EPJ Web of Conferences (Vol. 186, p. 03003). EDP Sciences.

Sundarraj, M., & Natrajan, R. (2019). A sustainable method to handle dark data in a smart factory. Software Quality Professional, 21(4), 21-33.

Suzen, N., Mirkes, E. M., Roland, D., Levesley, J., Gorban, A. N., & Coats, T. J. (2023, December). What is Hiding in Medicine's Dark Matter? Learning with Missing Data in Medical Practices. In 2023 IEEE International Conference on Big Data (BigData) (pp. 4979-4986). IEEE.

Taulli, T. (2019). What You Need To Know About Dark Data. Forbes, October, 27.

Teymourlouei, H., & Jackson, L. (2021). Dark data: managing cybersecurity challenges and generating benefits. In Advances in Parallel & Distributed Processing, and Applications: Proceedings from PDPTA'20, CSC'20, MSV'20, and GCC'20 (pp. 91-104). Springer International Publishing.

Tsaur, R. C. (2008). Forecasting analysis by using fuzzy grey regression model for solving limited time series data. Soft Computing, 12(11), 1105-1113.

Tsolas, I. E. (2019). Utility exchange traded fund performance evaluation. A comparative approach using grey relational analysis and data envelopment analysis Modelling. International Journal of Financial Studies, 7(4), 67.

Tzeng, C. J., Lin, Y. H., Yang, Y. K., & Jeng, M. C. (2009). Optimization of turning operations with multiple performance characteristics using the Taguchi method and Grey relational analysis. Journal of materials processing technology, 209(6), 2753-2759.

Upham, N. S., Poelen, J. H., Paul, D., Groom, Q. J., Simmons, N. B., Vanhove, M. P., ... & Agosti, D. (2021). Liberating host–virus knowledge from biological dark data. The Lancet Planetary Health, 5(10), e746-e750.

Wang, C. N., Dang, T. T., Nguyen, N. A. T., & Le, T. T. H. (2020). Supporting better decision-making: A combined grey model and data envelopment analysis for efficiency evaluation in e-commerce marketplaces. Sustainability, 12(24), 10385.

Wang, C. N., Lin, H. S., Hsu, H. P., Le, V. T., & Lin, T. F. (2016). Applying data envelopment analysis and grey model for the productivity evaluation of Vietnamese agroforestry industry. Sustainability, 8(11), 1139.

Wang, S., Ma, Q., & Guan, Z. (2007, November). Measuring hospital efficiency in China using grey relational analysis and data envelopment analysis. In 2007 IEEE International Conference on Grey Systems and Intelligent Services (pp. 135-139). IEEE.

Warangal, T. A Review on machine learning models used for anomaly detection.

Wu, C. H. (2007). On the application of grey relational analysis and RIDIT analysis to Likert scale surv Hsia, K. H., Chen, M. Y., & Chang, M. C. (2004). Comments on data pre-processing for grey relational analysis. Journal of Grey System, 7(1), 15-20.

Xia, R., Gao, Y., Zhu, Y., Gu, D., & Wang, J. (2022). An efficient method combined data-driven for detecting electricity theft with stacking structure based on grey relation analysis. Energies, 15(19), 7423.

Xuerui, T., & Yuguang, L. (2004). Using grey relational analysis to analyze the medical data. Kybernetes, 33(2), 355-362.

Xuerui, T., Julong, D., Hongxing, P., & Sifeng, L. (2007, November). Grey system and grey data management in medicine. In 2007 IEEE International Conference on Grey Systems and Intelligent Services (pp. 163-166). IEEE.

Yang, W., & Wu, Y. (2019). A Novel TOPSIS Method Based on Improved Grey Relational Analysis for Multiattribute Decision-Making Problem. Mathematical Problems in Engineering, 2019(1), 8761681.

Yang, Y., Liu, S., & Xie, N. (2019). Uncertainty and grey data analytics. Marine Economics and Management, 2(2), 73-86.

Yin, K., Zhang, Y., & Li, X. (2017). Research on storm-tide disaster losses in China using a new grey relational analysis model with the dispersion of panel data. International journal of environmental research and public health, 14(11), 1330.

Yu, X., Skeie, K. S., Knudsen, M. D., Ren, Z., Imsland, L., & Georges, L. (2022). Influence of data pre-processing and sensor dynamics on grey-box models for space-heating: Analysis using field measurements. Building and Environment, 212, 108832.

Zeng, G., Jiang, R., Huang, G., Xu, M., & Li, J. (2007). Optimization of wastewater treatment alternative selection by hierarchy grey relational analysis. Journal of environmental management, 82(2), 250-259.

Zhai, L. Y., Khoo, L. P., & Zhong, Z. W. (2009). Design concept evaluation in product development using rough sets and grey relation analysis. Expert systems with applications, 36(3), 7072-7079.

Zhang, L. J., & Li, Z. J. (2006). Gene selection for classifying microarray data using grey relation analysis. In Discovery Science: 9th International Conference, DS 2006, Barcelona, Spain, October 7-10, 2006. Proceedings 9 (pp. 378-382). Springer Berlin Heidelberg.

Zhang, S., & Zhou, Y. (2015). Grey wolf optimizer based on Powell local optimization method for clustering analysis. Discrete Dynamics in Nature and Society, 2015(1), 481360.

Zhang, Z., Wang, Y., & Xie, L. (2018). A novel data integrity attack detection algorithm based on improved grey relational analysis. IEEE Access, 6, 73423-73433.

Zhong, K., Jackson, T., West, A., & Cosma, G. (2024, June). Building a Sustainable Knowledge Management System from Dark Data in Industrial Maintenance. In International Conference on Knowledge Management in Organizations (pp. 263-274). Cham: Springer Nature Switzerland.

Zhou, K., & Song, J. (2021). Introduction to the Special Issue on Learning-based Support for Data Science Applications. ACM/IMS Transactions on Data Science, 2(2), 1-1.