Fujitsu AI White Paper

Fujitsu Causal Knowledge Graph

Transform to Data-Driven Decision-Making based on Logical Reasoning

Artificial Intelligence Laboratory,
Fujitsu Research Unit,
Fujitsu Ltd.
October 2024

Table of Contents

1. Introduction

At Fujitsu, we aim to create a world where efficient decision-making can be achieved based on diverse data, even under uncertain conditions. To make efficient decisions in an era of rapid change, also known as VUCA (Volatility, Uncertainty, Complexity, Ambiguity), it is essential to not only use the data within one's own organization but also to integrate and analyze multiple data sources, including trends from surrounding organizations, regions, and other fields. Logical reasoning based on the causal relationships between data items is indispensable.

In recent years, large language models (LLMs) have emerged, which are trained on vast amounts of various types of documents. These models can utilize a wide range of knowledge through natural language interfaces and are becoming useful tools for supporting decision-making. However, since the judgments made by LLMs are based on the general knowledge they encompass, they are not suitable for decision-making based on the latest data in specific specialized fields. For example, in a use case aiming to improve a company's profit margin, even if data such as employee satisfaction, number of stores, climate, and customer reviews—which are important factors for the company's sales—are provided to the LLM, effective support cannot be expected if the causal relationships between these factors are not clear.

To address this issue, we propose the 'Causal Knowledge Graph.' This concept combines two technologies that Fujitsu has developed over many years: Knowledge Graph (KG) that visualizes the relationships between data items to support information exploration and reasoning, and causal analysis that estimates the causal relationships inherent in the data. The Causal Knowledge Graph integrates and stores the causal relationships contained in the data along with related knowledge. By using this graph, efficient decision-making based on data becomes possible.

Furthermore, by accumulating Causal Knowledge Graphs from various fields, we can support deep analysis and decision-making across a wide range of domains. For example, in fields such as management, healthcare, sports, and manufacturing, it becomes possible to clarify the interrelationships and causal relationships among various related data, enabling more accurate analysis and predictions. This allows for customized data analysis and decision-making support tailored to the needs of customers in different fields.

In this white paper, we will introduce the concept of the 'Causal Knowledge Graph,' which is indispensable for data-driven decision-making. Additionally, we will discuss the underlying technological trends, the technologies for constructing and utilizing the Causal Knowledge Graph, and its use cases.

Contents of each Chapter

In Chapter 2, we will explain the technological trends of 'Causal Analysis' and 'Knowledge Graph (KG),' which are important technical elements of the Causal Knowledge Graph. Causal Analysis has primarily developed as a cutting-edge technology in data mining and statistical analysis, while KG has evolved as a form of knowledge representation in the internet age. We will discuss the fundamentals, research trends, and examples of the mutual utilization of these two technologies with the latest advancements, such as LLMs.

In Chapter 3, we will explain Fujitsu's concept of the Causal Knowledge Graph, which aims to transform data-driven decision-making. This includes the basic concept, the definition of the Causal Knowledge Graph, and its use cases. Through this, we will elucidate why the Causal Knowledge Graph is necessary, its internal structure, and the value it provides.

In Chapter 4, we will explain Fujitsu's suite of technologies for constructing and utilizing the Causal Knowledge Graph. Broadly, these technologies include: extracting known causal relationships from existing documents (Causal Extraction from documents), estimating unknown causal relationships from numerical data (Statistical Causal Discovery), estimating causal relationships from multiple numerical datasets (Integrated Causal Discovery), using known causal relationships as prior knowledge to estimate causal relationships (Causal Knowledge Transfer), integrating known and unknown causal relationships into a Causal Knowledge Graph (Causal Knowledge Creation), and utilizing the Causal Knowledge Graph (Root Cause Analysis, Causal Decision-making, and Causal Knowledge Graph Reasoning).

In Chapter 5, we will introduce case studies where the suite of technologies presented in Chapter 4 has been applied to three fields: 'ICT support,' 'human resources,' and 'healthcare and wellness.'

2. Technological Trends in Causal Analysis and Knowledge Graph

Before delving into the explanation of the Causal Knowledge Graph that Fujitsu aims to achieve, this chapter will describe the technological trends of 'Causal Analysis' and 'Knowledge Graph (KG)' which are important technical elements of the Causal Knowledge Graph. Causal Analysis has primarily developed as a cutting-edge technology in data mining and statistical analysis, while KG has evolved as a form of knowledge representation in the internet age. In the following sections, we will explain the basic knowledge and technological trends in Causal Analysis and KG, as well as examples of their mutual utilization with the latest advancements, such as large language models (LLMs).

2.1. Trends in Causal Analysis

Basics of Causal Analysis

Causal Analysis is a method used to clarify the relationship between cause and effect, aiming to understand and predict how specific causes influence outcomes. By elucidating the causal relationships of 'things' and 'events,' it is possible to unravel complex structures and mechanisms. As a result, causal analysis is applied in a wide range of fields, including science, medicine, economics, and social sciences, and is also used in policy-making and business strategy formulation.

When analyzing known causal relationships, the analyst predefines the causal relationships, setting the 'direction of causality' (which is the cause, and which is the effect) and defining the causal relationships to ensure that 'hidden common factors' do not appear. They then verify the presence and extent of the actual effects. However, to investigate unknown causal relationships, such prior settings by the analyst are not possible. Therefore, it is necessary to use experimental approaches or statistical approaches.

A representative example of a statistical approach is 'statistical causal discovery.' This technique uses statistical processing to estimate causal relationships from data for events where the causal relationships are not obvious. This allows analysts to perform causal analysis even on things and events where the causal relationships are not self-evident.

The most important challenge in causal analysis is to scrutinize the 'correct causal relationships.' Even when using statistical causal discovery as mentioned above, there is no guarantee that the correct causal relationships can always be estimated, and traditionally, manual verification by experts was necessary. Recently, with the advent of LLMs, there have been attempts to replace the expert verification process with LLMs. Below, we will explain the technological trends of the representative method, statistical causal discovery, and examples of using LLMs in causal analysis.

Technological Trends in Statistical Causal Discovery

Statistical causal discovery is a method that uses statistical approaches to estimate the cause-and-effect relationships from observed numerical data. This method goes beyond mere correlations to clarify how specific variables influence other variables. By doing so, it enables the scientific verification of the impact of strategies and the effects of interventions in areas such as business strategy formulation and treatment planning.

Statistical causal discovery outputs a causal graph where each item in the data (e.g., diet, exercise, health) is represented as a node, and the estimated cause-and-effect relationships are represented as edges connecting the nodes (see Figure 1). The causal graph is a directed graph, where the node at the base of the arrow represents the cause, and the node at the tip of the arrow represents the effect. Additionally, the edges are annotated with 'causal effects,' which indicate the degree of influence the cause has on the effect. Typically, causal effects are represented as numerical values with positive or negative signs. A positive value means that as the cause increases, the effect also increases, while a negative value means that as the cause increases, the effect decreases.

Representative methods of statistical causal discovery include the Structural Equation Model (SEM) introduced by Pearl [1] and the Potential Outcomes Framework introduced by Rubin [2]. The Structural Equation Model allows for the visual and precise formulation of causal relationships by describing the causal effects from causes to effects as equations, thereby capturing the flow of causality.

Under various assumptions, the causal discovery algorithms for observational data that follow the Structural Equation Model can be broadly divided into the following three types.

- **Constraint-based methods** : These methods replace the problem of causal relationships with the problem of conditional independence and estimate the Markov equivalence class (a class that shares the same conditional independence structure). The PC algorithm [3] is well-known for this approach.
- **Score-based methods** : These methods optimize a score defined for the causal graph to estimate the causal graph underlying the observational data. The GES algorithm [4] is an example of this approach.
- **Function-based methods** : These methods directly estimate the equations that describe the Structural Equation Model. Models that are linear, additive, and have non-Gaussian error variables are known as LiNGAM, with DirectLiNGAM [5] being well-known. In recent years, extensions to various aspects, such as non-linear models [6], have also been recognized.

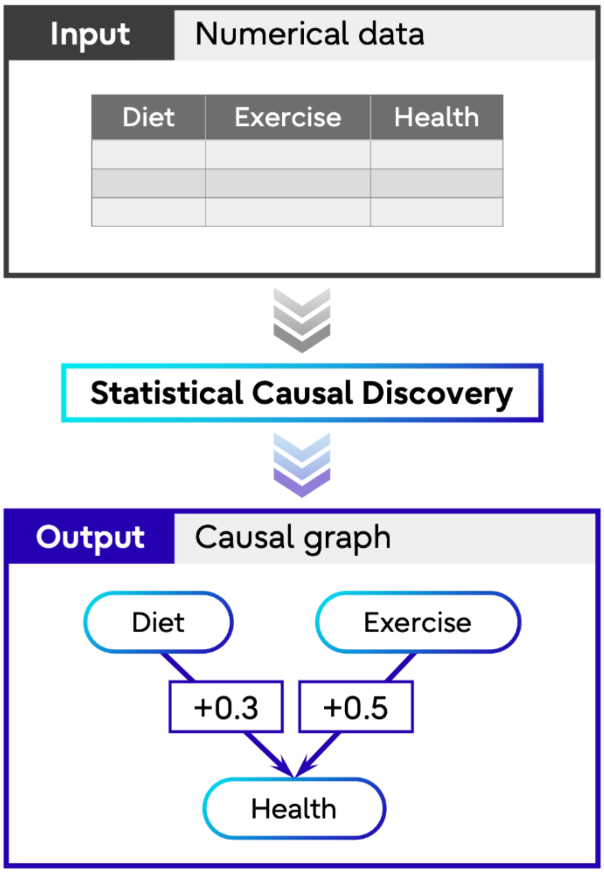


Figure 1: Input and Output of Statistical Causal Discovery

Practical Examples and Service Cases of Causal Analysis

Currently, many companies offer causal analysis services using statistical causal discovery, but the causal discovery techniques supported by each company vary. For example, some companies apply techniques that are also applicable to non-linear models, while others adopt a constraint-based approach for causal discovery algorithms. Some companies perform causal discovery based on Rubin's Potential Outcomes Framework rather than Pearl's Structural Equation Model. There are also companies that place more emphasis on estimating intervention effects and optimal intervention values rather than on causal discovery itself.

In addition, causal discovery support products are also being offered. For example, user-friendly interfaces are provided that allow beginners to select and execute various discovery algorithms on their own. Furthermore, features that enable users to specify the presence or absence of specific edges, thereby leveraging their specialized knowledge, are also available. In this way, there are multiple players offering statistical causal analysis as a service. However, the basic functionality generally involves estimating causal structures from individual observational data held by users or providing consulting services.

Examples of Using Large Language Models (LLMs) in Causal Analysis

In general, causal analysis requires interpreting the estimated causal structures between data items based on expert knowledge. Additionally, when estimating causal structures, any causal relationships that have been confirmed as expert knowledge need to be incorporated as prior knowledge. Traditionally, such tasks were labor-intensive and performed by experts. However, with the rapid development of LLMs in recent years, several studies have emerged that aim to support these tasks using LLMs.

For example, by repeatedly querying the LLM about the validity of causal relationships during statistical causal discovery, it has been reported that it is possible to estimate causal structures that do not contradict expert knowledge [7]. However, there are also reports that current LLMs perform significantly poorly in logically distinguishing between correlation and causation, which is one of the most important challenges in statistical causal discovery [8].

As for services provided by companies, at present, the use of LLMs is limited to supporting the inputs and outputs of statistical causal discovery. There are no confirmed cases where LLMs have significantly advanced causal discovery algorithms. For example, using LLMs as an interface to answer questions like 'What is the cause of the churn rate?' based on causal graphs, having LLMs suggest prior knowledge for statistical causal discovery or candidates for unobserved common causes, or having LLMs explain the results of causal analysis—these are mostly aimed at improving the usability of causal analysis rather than advancing the causal discovery itself. There are still many challenges to be addressed in significantly advancing statistical causal discovery through the combination with LLMs.

2.2. Trends in Knowledge Graph

Basics of Knowledge Graph

Knowledge Graph (KG) have been used as a term since around 1972 and have been continuously researched and developed as a form of knowledge representation since the 1980s. There are various definitions, but comprehensively, they can be described as graph-structured data aimed at accumulating and transmitting real-world knowledge. In KG, the nodes represent entities of interest (such as people, places, objects, concepts, etc.), and the edges represent the relationships between these entities (for example, 'X is part of Y,' 'X owns Z,' etc., see Figure 2).

The main advantages of representing information or data as KG are as follows.

- **Integration, Sharing, and Updating of Knowledge:** By representing information from multiple data sources in a unified graph structure, it becomes easier to integrate and share data that differ in format and structure while maintaining compatibility. Additionally, the addition and updating of new data can be handled flexibly.
- **Representation of Complex Relationships:** It can easily represent complex relationships between entities, including many-to-many relationships, cyclic relationships, and hierarchical relationships, which are difficult to express in relational databases (RDB).
- **Efficient and Flexible Search:** It can efficiently execute complex queries that span multiple tables and include relationships concerning entities, which is challenging in relational databases (RDB).
- **Enhanced Reasoning and Knowledge Discovery:** It can apply both deductive reasoning, which derives new knowledge from existing knowledge using relationships, and inductive reasoning, which finds regularities in the graph structure to extract new knowledge. This enables knowledge discovery and the reasoning of unknown information.

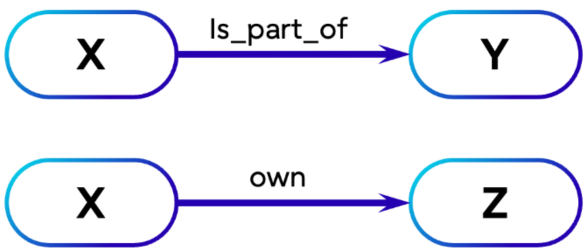


Figure 2: simple example of Knowledge Graph

In the following sections, we will explain the research trends in KG and examples of using LLMs in causal analysis.

Technological Trends in Knowledge Graph

The field of KG is expansive, and the research areas can be largely divided into three categories: knowledge representation, knowledge acquisition, and reasoning.

Knowledge representation is a technology that represents information and concepts on a computer so that the computer can understand and reason with human knowledge. Key knowledge representation formats used in KG include RDF (Resource Description Framework) [9] and property graphs. RDF is standardized by the W3C and represents data in triples of 'subject-predicate-object.' For example, 'X is part of Y' is represented as '<X>-<is_part_of>-<Y>.' It also identifies resources with URIs and uses standard vocabularies such as ontologies to define the meaning of data, enhancing interoperability and reusability. Property graphs are graph data with nodes and edges assigned multiple key-value pairs called properties. Properties can be added flexibly, and they have high compatibility with graph databases. Recently, as an extension of RDF, RDF-star has been proposed, which, although not yet standardized by the W3C, allows for the definition of meta-information for triples and has the expressive power of hypergraphs.

Knowledge acquisition is a technology that extracts entities and their relationships from texts and databases to construct and expand KG. It is researched from three perspectives: relation extraction, entity discovery and entity linking, and KG completion and link prediction. Relation extraction is a technology that extracts relationships concerning entities from texts and is used for the automatic construction of KG. Entity discovery and entity linking are technologies that extract entities from texts and link them to the KG, thereby resolving entity ambiguities.

Reasoning in KG is a reasoning technology that derives new knowledge or infers unknown information from the KG. Deductive reasoning is a technique that derives new knowledge deductively from existing knowledge, such as 'Socrates is a human' and 'All humans will eventually die,' leading to the new knowledge that 'Socrates will eventually die.' Inductive reasoning is a technique that applies machine learning to discover patterns and regularities on the KG to generate new knowledge. Recently, graph embedding techniques that use deep learning to embed entities and relationships into vector spaces and infer based on their similarities have been researched and applied. Additionally, technologies that explain these reasoning results in a human-understandable form using relationships between entities are also being researched.

In recent years, with the rapid development of LLMs, numerous studies and applications combining KGs and LLMs have been published.

Practical Examples and Service Cases of Knowledge Graph

Practical examples of using KG include search services, question-answering systems, and recommendation systems. In search services, they are used to present information about entities included in search queries. In question-answering systems, tracing the relationships between entities on the KG helps provide more appropriate answers. In recommendation systems, KG of product information are used to suggest related products based on information such as the genres of products the user has previously purchased.

In addition, several KGs containing general information or information specific to certain fields have already been made public. For general information, there are DBpedia [10], which is based on data extracted from Wikipedia; YAGO [11], which integrates Wikipedia and WordNet; and the Google Knowledge Graph [12], which was built including DBpedia and FreeBase, the latter being based on human contributions. Examples of publicly available KGs in specific fields include DrugBank [13], which compiles information about drugs in the biomedical field; ROBOKOP [14] and Bio2RDF [15], which integrate DrugBank with PubMed article information and other data. Other examples include the Financial Industry Business Ontology [16], which provides a standard common language for the financial industry, and GOV.UK [17], which provides government-related information from the UK. Many of these KGs available on the internet are described in RDF, which ensures resource identifiability and high interoperability.

Databases for storing KGs can be broadly categorized into two types corresponding to knowledge representation: RDF stores and graph databases. Representative RDF stores include Virtuoso [18] and GraphDB [19]. Representative graph databases include Neo4j [20] and Amazon Neptune [21]. The query languages they support differ: RDF stores are compatible with SPARQL, which allows for searching triples, while Neo4j and similar databases support Cypher, which is designed for graph databases.

Examples of Using Large Language Models (LLMs) in Knowledge Graph (KG)

As a problem with LLMs such as GPT-4, it has been pointed out that while LLMs excel at answering simple questions, their accuracy significantly decreases when answering complex questions that require multi-step reasoning with related information. Additionally, it is difficult to investigate the process and basis by which the answers were derived, making it challenging to determine whether the answers obtained from LLMs are correct.

On the other hand, KGs are inherently difficult to construct, and existing methods particularly struggle with generating new facts and representing unknown knowledge. Therefore, research and development are progressing to integrate KGs and LLMs, leveraging their respective strengths and compensating for their weaknesses.

According to a recent survey paper on the integration of LLM and KG [22], the forms of integration can be classified into the following three categories (see Figure 3).

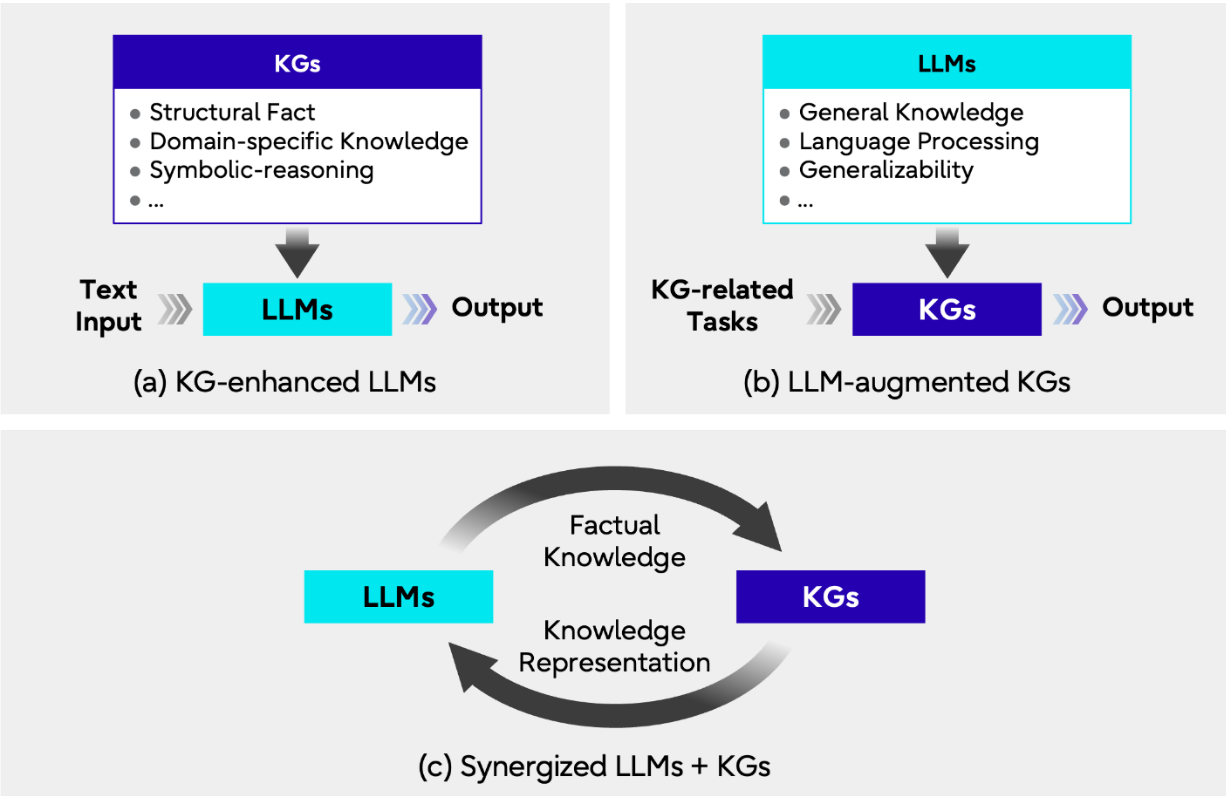


Figure 3: Integration patterns of LLM and KG
Source: Created based on Fig. 6 from [22]

(a) KG-enhanced LLMs: Improving the performance and interpretability of LLM using KGs
- Applying KGs during pre-training, using KGs for LLM inference, and utilizing KGs to understand and interpret the knowledge learned by LLMs and their reasoning processes.

(b) LLM-augmented KGs: Enhancing KG tasks using by LLMs
- KG embedding using LLMs, KG completion using LLMs, KG generation using LLMs, text generation from KGs using LLMs, and KG-based question answering using LLMs

(c) Synergized LLMs + KGs: Synergistic effects of LLMs and KGs

- Integrating LLMs and KGs into a unified framework to mutually enhance each other

As mentioned above, in the field of basic research, the development of technologies that integrate KGs and LLMs to complement each other is attracting attention. On the other hand, although still few, there are some examples in the business field where both are integrated and utilized. For example, the application of LLMs to interactive AI assistants for enterprise search is a representative case. By leveraging KGs, it is possible to derive appropriate answers even for questions that require generating responses from multiple pieces of evidence [23] [24].

Another application example is the development of technologies that use KGs to verify the answers output by LLMs. For instance, KGs created from open clinical data are being used to improve the accuracy of responses from medical LLMs [25]. These application examples correspond to 'a. KG-enhanced LLMs' in Figure 3, where KGs are utilized to address the challenges of LLMs.

Similarly, there are examples corresponding to 'b. LLM-augmented KGs' in Figure 3, where LLMs are utilized to address the challenges of KGs. Recently, services have been offered that use generative AI, including LLMs, to quickly create KGs from unstructured data such as PDFs, web pages, and documents [26]. In addition to these examples, the integration and utilization of LLMs and KGs are expected to be applied in various areas, including risk assessment of financial data, summarization of legal documents, analysis of traffic patterns and social media, recommendation systems for products and content in e-commerce and retail, and the provision of personalized education and healthcare in the education and healthcare sectors [27].

3. Causal Knowledge Graph proposed by Fujitsu

In this chapter, we will explain the concept of Fujitsu's Causal Knowledge Graph, which aims to transform data-driven decision-making. This includes the basic concept, the contents of the Causal Knowledge Graph, and its use cases (provided value).

3.1. Basic Concept of Causal Knowledge Graph: Why Causal Knowledge Graph is Necessary?

In corporate management, achieving data-driven decision-making requires more than just the 'correlations' extracted through conventional data mining and statistical analysis. Correlations often involve spurious correlations or latent common factors, making it impossible to make logical decisions based on them. For example, if a correlation is found between the number of Nobel Laureates and chocolate consumption (see Figure 4), it would be meaningless to implement a policy of distributing free chocolate to employees with the aim of producing Nobel Laureates from within the company based on that result.

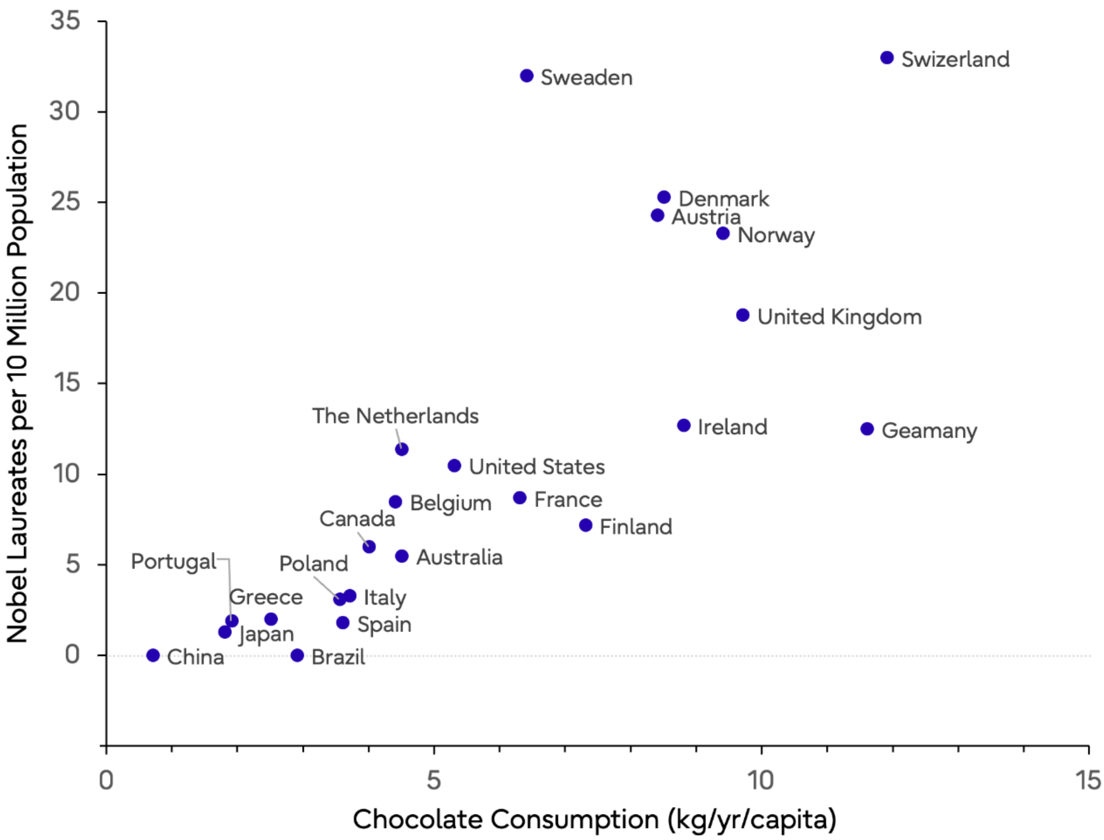


Figure 4: Scatter Plot of Chocolate Consumption and Nobel Laureates
Source: Created based on Fig.1 in [28]

Such measures are based on superficial correlations that ignore causal relationships, making it unlikely to achieve actual results. Therefore, for companies to truly make data-driven decisions, it is important to accurately capture causal relationships beyond mere correlations. By understanding the correct causal relationships, it becomes possible to make logical and effective decisions based on data, leading to business success.

Here, there are significant challenges in capturing causal relationships. To establish accurate causal relationships, extensive verification work by experts is usually required, which involves considerable cost and time. There are two main reasons for this. First, much of the known causal relationships are described in natural language (referred to as unstructured data), making them not easily processable by computers. Experts need to structure the information and concepts so that they can be understood by computers. Second, in many cases, clear causal relationships are not known, necessitating additional analysis and verification by experts to discover these relationships. These challenges make it difficult to accurately capture the causal relationships that support data-driven decision-making.

In response to the two reasons mentioned above, as we have seen in Chapter 2, the following two technologies have been developed. For known causal relationships, there is a technology that automatically extracts causal relationships from documents using LLMs. For unknown causal relationships, there is a technology that automatically estimates causal relationships between data items using statistical causal discovery.

However, the current two technologies each have the following limitations. First (Limitation 1), in the extraction of causal relationships from documents, all extracted causal relationships are qualitative. Therefore, it is not possible to predict the probability or impact of another event (effect) occurring when a certain event (cause) happens. Next (Limitation 2), in statistical causal discovery from data, since all causal relationships are estimated from the data, biases within the data or a lack of data can lead to the derivation of causal relationships where the true cause and effect are reversed. Furthermore (Limitation 3), current statistical causal discovery can only derive causal relationships within the scope of the given data. As a result, it cannot combine multiple datasets with different observation conditions (e.g., different observed subjects, different observation items, different measurement conditions at the time of observation) to derive causal relationships. Overcoming these limitations requires more advanced technologies and approaches.

To overcome these limitations and achieve a transformation in data-driven decision-making, Fujitsu is working on the following technological developments.

- **Technology to overcome Limitation 1:** By integrating causal relationships extracted from documents with statistical causal relationships estimated from data, it becomes possible to measure causal effects and impact even for causal relationships that were previously understood only qualitatively. For example, in employee engagement, the causal relationships between engagement perspectives and initiatives can be understood from organizational behavior textbooks. By combining this with the causal effects estimated from the actual engagement survey results of the company, it becomes possible to formulate initiatives that include quantitative expected effects.

- **Technology to overcome Limitation 2:** Accumulate known causal relationships extracted from documents and set these known causal relationships in advance when performing statistical causal discovery on numerical data. This approach allows for the estimation of unknown causal relationships while resolving contradictions between the causal relationships derived from statistical causal discovery and the known causal relationships.

- **Technology to overcome Limitation 3:** Develop a technology that automatically determines the data distribution common to different domains, enabling the combination of numerical data with different observation conditions and granularity for statistical causal discovery. This will allow for causal discovery that learns and continues to grow from large datasets.

To achieve the above, we will adopt the 'Knowledge Graph (KG)' as the data structure. KG is a form of knowledge representation, and it is particularly suitable for naturally expressing causal relationships, making it ideal for our purpose. Additionally, one of the application areas of KG is data integration, which offers the advantage of easily integrating causal relationships extracted from multiple data sources such as documents and numerical data.

In summary, the 'Causal Knowledge Graph' that Fujitsu aims to achieve, equipped with the technologies and data structures, is designed to transform data-driven decision-making in corporate management.

3.2. Definition of Causal Knowledge Graph: What are the Contents of Causal Knowledge Graphs?

Figure 5 shows an image of the 'Causal Knowledge Graph' that Fujitsu aims to achieve.

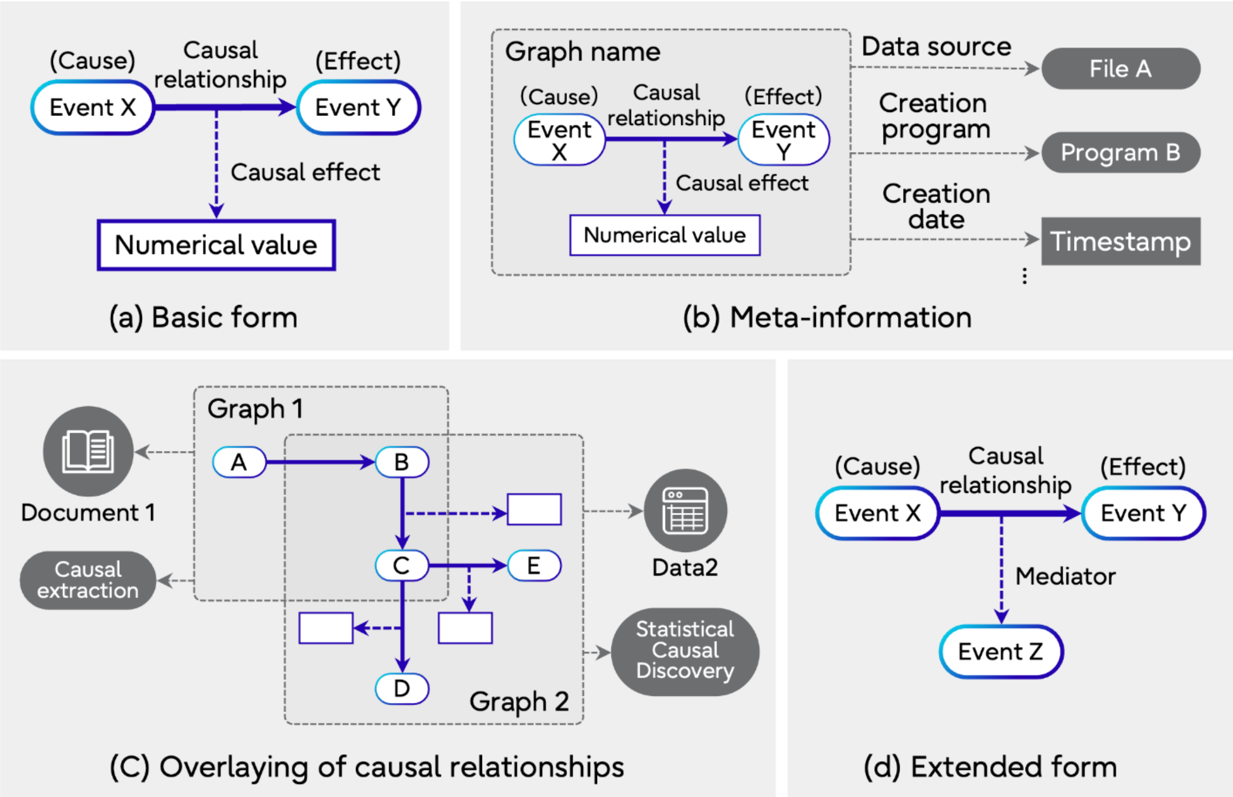


Figure 5: Fujitsu's Causal Knowledge Graph (Conceptual Diagram)

As shown in Figure 5(a), since the KG is the basic form, the basic unit is a triple consisting of two nodes (ellipses) connected by a single edge (arrow). The nodes represent events, with the event at the base of the arrow being the cause and the event at the tip of the arrow being the effect. Therefore, the solid line in Figure 5(a) represents 'Event X (cause) causes Event Y (effect) (there is a causal relationship),' which is the basic form.

The dotted lines in Figure 5 represent optional additional information, and Figure 5(a) shows an example where 'causal effect' is added. Since causal effects are usually represented numerically, they are shown as string types (squares) rather than entity types (ellipses).

Figure 5(b) is an example where meta-information is added. You can add a name (graph name) to the causal relationship or include information about the data source or the program used to create the causal relationship. Additionally, if it is numerical data, you can include conditions at the time of observation or statistical values such as mean and variance.

Figure 5(c) is an example where several causal relationships are overlaid. Here, the meta-information indicates that 'Graph1 was created from Document1 using a causal extraction program,' and it shows that two causal relationships were extracted: 'Event A (cause) → Event B (effect)' and 'Event B (cause) → Event C (effect).' Similarly, for Graph2, it indicates that 'Graph2 was created from Data2 using a statistical causal discovery program,' and it shows that the causal relationships 'Event B (cause) → Event C (effect),' 'Event C (cause) → Event D (effect),' and 'Event C (cause) → Event E (effect)' were estimated with causal effects.

Here, 'Event B (cause) → Event C (effect)' is extracted or estimated from both Document1 and Data2. By overlaying this, it is possible to integrate causal relationships extracted from multiple data sources. Notably, the 'Event B (cause) → Event C (effect)' extracted from Document1 serves to resolve the 'contradictions in statistical causal discovery,' which is Limitation 2 mentioned in the previous section. Meanwhile, the 'Event B (cause) → Event C (effect)' with causal effects estimated from Data2 corresponds to the 'integration of causal relationships extracted from documents and statistical causal relationships estimated from data,' which is Limitation 1 mentioned in the previous section.

Figure 5(d) is an example where additional information other than causal effects is attached to the causal relationships themselves as an option. This allows for the description of mediators or side effects that occur when a certain causal relationship takes place.

3.3. Use Cases: What Value provided by the Causal Knowledge Graph?

In this section, we will explain the value provided by the Causal Knowledge Graph through two future scenarios.

Future Scenario 1: The Case of Ms. Tanaka in the Human Resources Department

Ms. Tanaka in the human resources department is supporting action-taking aimed at improving workplace engagement. She provides guidance on how to interpret the results of the Engagement Survey (ES) and offers advice on initiatives to various department in her company, but she is also concerned about whether she is proposing initiatives that are suitable for each department. Today, she received a consultation from Mr. Sato, the division manager, who said, 'I'm told to increase 'motivation' and 'fulfillment' to boost engagement, but I don't really know what to do in practice.'

So, Ms. Tanaka decided to try Fujitsu's Causal Knowledge Graph. The Causal Knowledge Graph has pre-analyzed organizational management textbooks and stored causal relationships such as 'supporting skill development' and 'assigning to growth areas' increase 'career opportunities,' and 'clarification of the vision' and 'assigning to growth areas' enhance 'the company's future' (upper part of Figure 6). In addition to this, Ms. Tanaka had the engagement survey results of Mr. Sato's department analyzed.

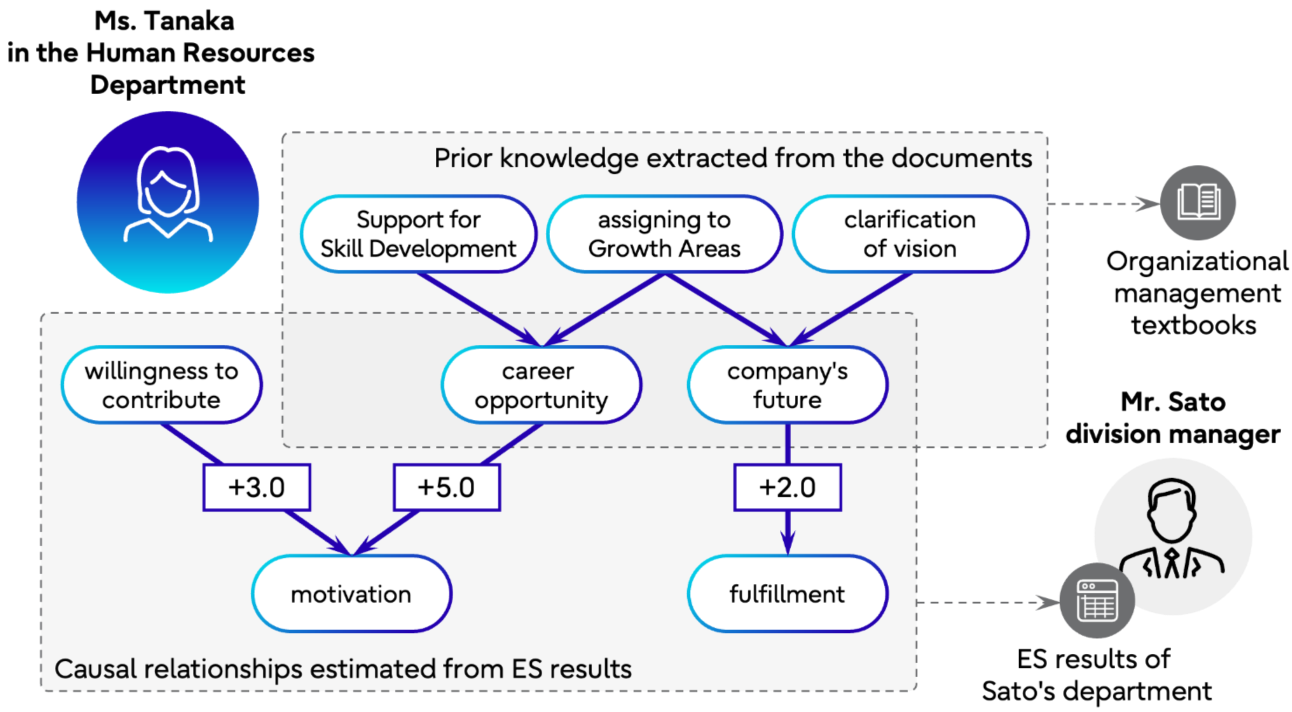


Figure 6: Causal Knowledge Graph generated from Management Textbooks and Engagement Surveys

With Fujitsu's Causal Knowledge Graph, it is possible to estimate the causal relationships between each item in the engagement survey. It was found that 'to increase "motivation," "willingness to contribute" and "career opportunities" are the causes,' and 'to increase "fulfillment," "the company's future" is the cause,' among other things (lower part of Figure 6). Here, 'career opportunities' and 'the company's future' become key, and by combining the known causal relationships with the estimated causal relationships, it was determined that three initiatives—support of skill development, assigning to growth areas, and clarification of the vision—are effective in increasing "motivation" and "fulfillment."

So, Ms. Tanaka advised Mr. Sato on two initiatives: 'assigning to growth areas' and 'clarification of the vision.' Additionally, when she analyzed the engagement survey results of other departments, she found that in every department, 'career opportunities' were the cause of increasing 'motivation.' Therefore, she decided to propose to the HR director the enhancement of e-learning education as an initiative to support skill development.

The HR director asked, 'Why is support of skill development important?' and 'What is its effect?' Ms. Tanaka, while showing the Causal Knowledge Graph, was able to explain that 'it is an important initiative to increase motivation' and that 'by increasing career opportunities, motivation can be improved by 5 points.'

Future Scenario 2: The Case of Mr. Suzuki in Sales Department

Mr. Suzuki in sales department is busy working as a manager every day. Recently, inquiries about the products he handles have increased, and the number of business negotiations has doubled compared to before. However, since the products are in a relatively new field, the win rate of the negotiations is not very high. The department head has instructed him to 'analyze the factors behind the wins and losses in the negotiations.' Meanwhile, the results of the Engagement Survey in December have come back, and the HR department is urging him to 'analyze the Engagement Survey results and come up with initiatives for each team.' Honestly, Mr. Suzuki feels that 'I am too busy with daily tasks to have time for analysis.

One day, Mr. Suzuki heard the phrase 'Enhancing employee engagement is the key to business success' at a lecture. Although he was skeptical, thinking 'Really?', he decided to use Fujitsu's Causal

Knowledge Graph to simultaneously analyze the engagement survey results and business negotiation results from his department (right side of Figure 7).

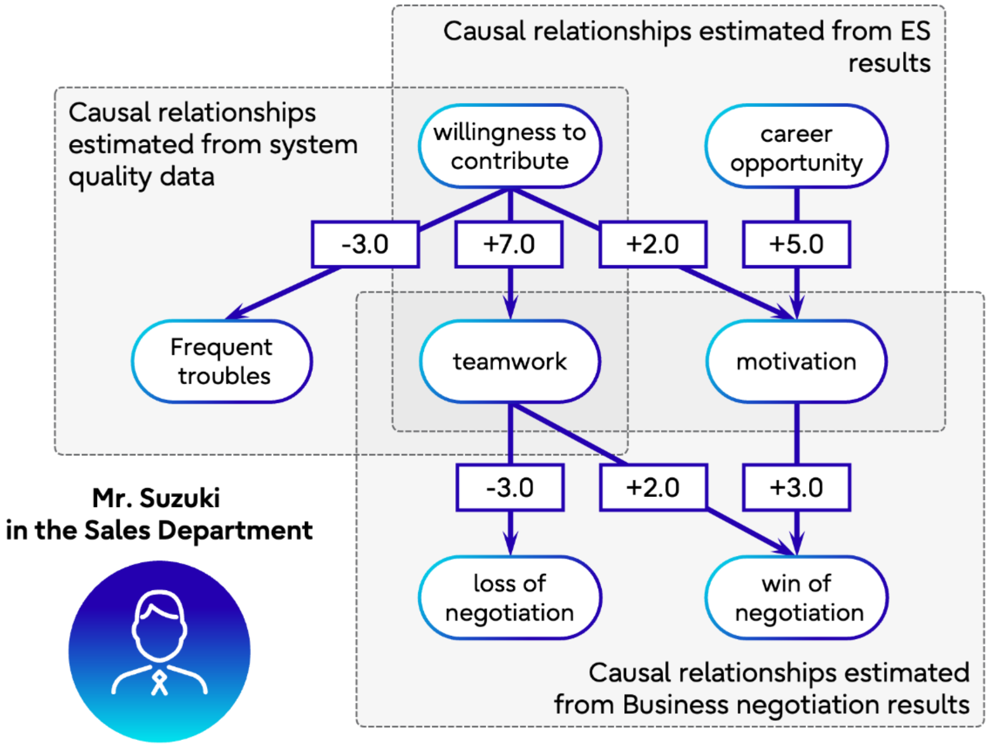


Figure 7: Causal Knowledge Graph generated from three data sources.

As a result, it became clear that 'teamwork' and 'motivation' influence the 'wins' and 'losses' in business negotiations, and that 'to improve "teamwork" and "motivation," it is necessary to increase "willingness to contribute" and "career opportunity."' Indeed, it seems there is some truth to the idea that 'enhancing employee engagement is the key to business success.'

Furthermore, when Mr. Suzuki added and analyzed the system quality data that manages the issues and risks of the project, he found that 'a decrease in "willingness to contribute" is causing "frequent troubles"' (left side of Figure 7).

Here, using the causal decision-making function of Fujitsu's Causal Knowledge Graph (which will be discussed later in Chapter 4), Mr. Suzuki analyzed the factors that could simultaneously increase the number of 'wins of negotiation' and reduce 'frequent troubles.' He found that the most effective initiative was 'improving willingness to contribute.'

Mr. Suzuki shared these analysis results with the department head and decided to make 'improving willingness to contribute' the top priority for his team. He promptly went to consult with Ms. Tanaka in the HR department about initiatives to increase 'willingness to contribute.

4. Fujitsu's Technologies related Causal Knowledge Graph

In this chapter, we will explain Fujitsu's suite of technologies for constructing and utilizing the Causal Knowledge Graph. Broadly, these technologies include: extracting known causal relationships from existing documents (Causal Extraction from document), estimating unknown causal relationships from numerical data (Statistical Causal Discovery), integrating causal relationships estimated from multiple numerical datasets (Integrated Causal Discovery), using known causal relationships as prior knowledge to estimate causal relationships (Causal Knowledge Transfer), integrating known and unknown causal relationships into a Causal Knowledge Graph (Causal Knowledge Creation), and utilizing the Causal Knowledge Graph (Root Cause Analysis, Causal Decision-making, and Causal Knowledge Graph Reasoning). It should be noted that the Causal Knowledge Graph-related technologies introduced here are currently in the conceptual stage, with only some basic functions implemented at this time.

4.1. Positioning of each Technology

Figure 8 shows the related technologies and overall structure of the Causal Knowledge Graph. First, known causal relationships are extracted from various documents and stored in a graph database. This forms part of the Causal Knowledge Graph and is used as prior knowledge for statistical causal discovery (overcoming Limitation 2). Next, for individual numerical data, unknown causal relationships are estimated using statistical causal discovery and similarly stored in the graph database. For multiple numerical datasets, integrated causal discovery is used to estimate causal relationships by combining their respective preconditions and statistical values such as mean and variance (overcoming Limitation 3), and these are stored in the graph database. Finally, by entity-izing the nodes as 'events' for both the causal relationships extracted from documents and those estimated from numerical data, the Causal Knowledge Graph is constructed. Currently, events that appear in both documents and numerical data are represented as the same event, integrating them into the Causal Knowledge Graph (overcoming Limitation 1). The constructed Causal Knowledge Graph is used not only for the utilization technologies described later but also as known causal relationships, serving as prior knowledge for statistical causal discovery and integrated causal discovery (overcoming Limitation 2).

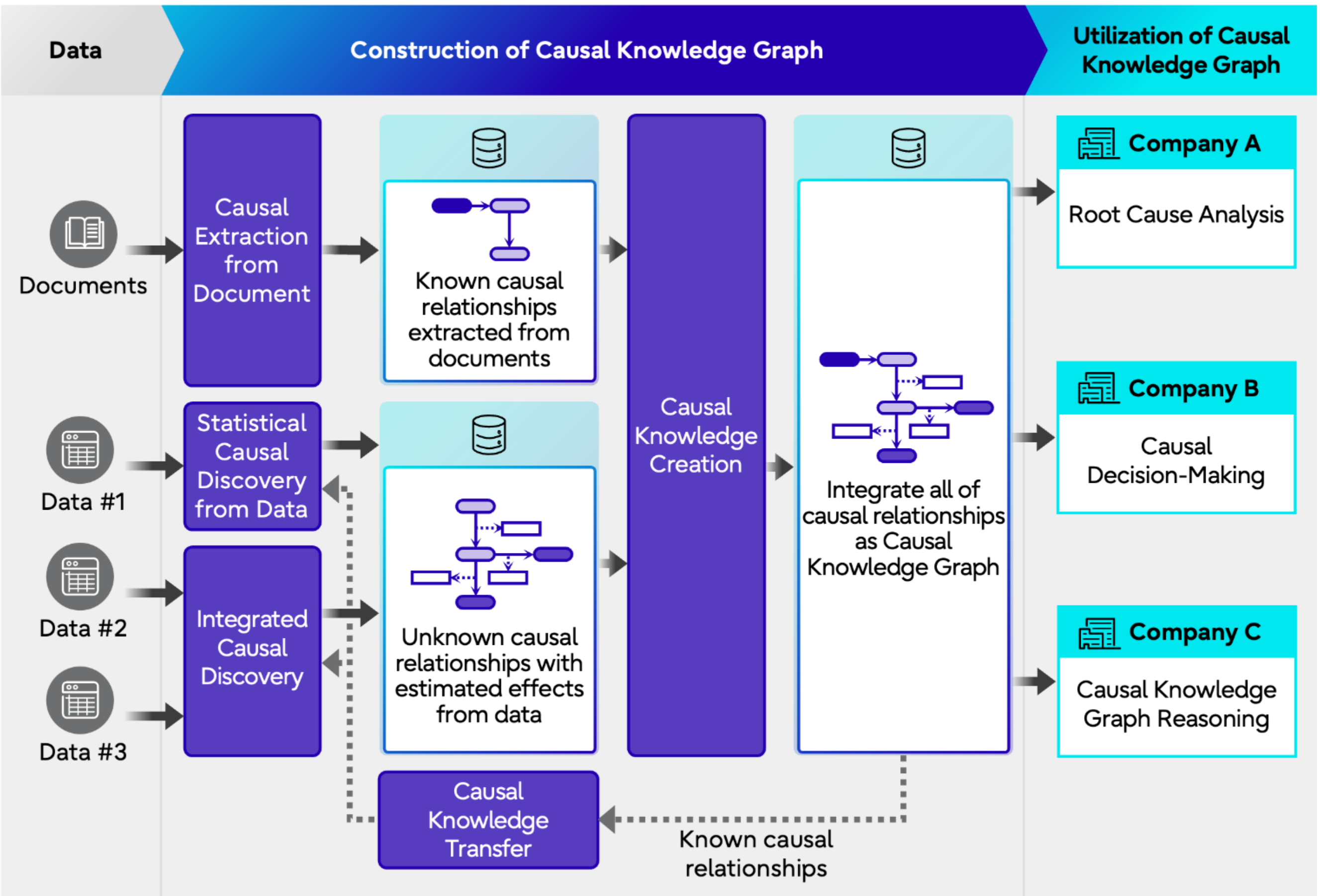


Figure 8: System Structure of Causal Knowledge Graph-Related Technologies

Next, we discuss the utilization technologies of the Causal Knowledge Graph. There are a wide range of applications, including Root Cause Analysis (RCA), which investigates the fundamental causes by tracing the chain of causality backward; Causal Decision-making, which considers side effects and adverse impacts of the occurrence of certain events; and Causal Knowledge Graph Reasoning, which derives new causal relationships from multiple existing causal relationships. Additionally, by similarly converting customer data into a Causal Knowledge Graph, it is possible to develop applications that reflect the customer's data.

Table 1 provides an overview of each technology.

Table 1: Overview of Causal Knowledge Graph-Related Technologies

	Technology	Overview	Input/Output
Construction of Causal Knowledge Graph	**Causal Extraction from document**	This technology enables the automatic extraction of causal relationships from documents and the creation of causal knowledge graphs by automatically converting processing flows and prompts to generate KGs based on document structure types.	**Input:** Document Data **Output:** Causal Knowledge Graph
	Statistical Causal Discovery from Data	There is a high demand for understanding causal relationships that apply only to data under specific conditions, rather than general causal relationships that apply to the entire dataset. Therefore, this technology efficiently explores conditions under which unknown causal relationships emerge by rapidly enumerating all conditions to cover all variations of causal relationships.	**Input:** Multivariate Numerical Data **Output:** Causal Graph
	Integrated Causal Discovery (overcoming Limitation 3)	By linking and analyzing data with different granularities and items of measurement at the time of observation, an integrated causal graph is constructed. This technology enables causal inference on large datasets.	**Input:** Multivariate Numerical Data **Output:** Causal Graph
	Causal Knowledge Transfer (overcoming Limitation 2)	By automatically identifying causal relationships stored in the existing causal knowledge graph that can be utilized as prior knowledge for the given numerical data, a more valid causal graph is constructed.	**Input:** Multivariate Numerical Data **Output:** Causal Graph
	Causal Knowledge Creation (overcoming Limitation 1)	This technology constructs an integrated causal knowledge graph by relating entities that appear in causal relationships extracted from documents and statistical causal relationships estimated from numerical data. The constructed causal knowledge graph complements newly input document data and numerical data to enhance root cause analysis and causal decision-making and is also used to infer new causal relationships in the KG.	**Input:** Causal Relationships Extracted from documents and Causal Relationships Estimated from Numerical Data **Output:** Causal Knowledge Graph
Utilization of Causal Knowledge Graphs	**Root Cause Analysis**	Using the causal knowledge graph, logically explain the basis of measures. It is possible to generate high-precision answers even for complex events that would fail in RAG searches.	**Input:** Causal Knowledge Graph **Output:** Root Cause (Event Entity)
	Causal Decision-Making	This technology recommends the most effective measures (events) to achieve goals based on all possible causal relationships discovered from data. The measures recommended by this technology are optimized to achieve the goals at the lowest cost while minimizing the impact on other items.	**Input:** Causal Knowledge Graph, Attributes to be Changed, and Their Target Values **Output:** Most Effective Measures (Event Entity)
	Causal Knowledge Graph Reasoning (Question answering using LLM)	This technology uses LLMs to provide answers reflecting the causal relationships in the causal knowledge graph to user questions. Dividing the causal knowledge graph into subgraphs and making them the search targets of RAG, improves answer accuracy.	**Input:** Causal Knowledge Graph, Question Text **Output:** Answer Text

4.2. Construction Technologies of Causal Knowledge Graph

Causal Extraction from Documents

By extracting triples (two words/events/information and their relationship) from documents and creating a KG, it is possible to generate high-precision answers to queries related to the documents in a short time. For example, in the case of complex queries that require referencing multiple documents, the usual Retrieval Augmented Generation (RAG) in LLMs struggles to correctly reference multiple documents, leading to a decrease in answer accuracy. However, by using a KG that reflects the content of multiple documents, it is possible to prevent the decrease in answer accuracy. Fujitsu refers to this as 'KG-enhanced RAG.'

While KGs are useful, constructing them manually requires experts to invest a significant amount of time and effort, which has been a challenge. Therefore, automatic generation technologies have been widely explored. One approach to creating a general-purpose KG involves applying prompt engineering techniques (such as in-context learning) to LLMs (like GPT-4) to comprehensively extract nouns from documents and their relationships. Although, general-purpose KGs can be applied to tasks like question-answering, they often generate a large number of triples unrelated to causality or separate causal relationships due to paraphrasing of events. As a result, extensive extraction and transformation work by experts is still required. Additionally, if one wants to generate a KG tailored to a specific purpose, an approach could involve training a language model (like BERT) using a dataset with similar characteristics to the desired KG. However, preparing such datasets generally requires a significant amount of effort. As described above, traditional methods still require expert work, which has prevented the rapid decision-making needed in real business scenarios.

Therefore, we have developed the Causal Knowledge Graph extraction technology that automatically extracts causal relationships from documents without the need for training data. This technology leverages LLMs. While LLMs excel at understanding grammatical structures such as word extraction, it was necessary to appropriately guide the focus of the LLM to extract causal relationships between events from the entire document. As shown in Figure 9, we pre-defined the structure type of the Causal Knowledge Graph as KG schema and devised a framework to analyze input documents according to this schema, enabling the output of KGs that indicate causal relationships. In this framework, document analysis using LLMs is repeatedly performed according to the processing flow, gradually constructing the KG. This is achieved through a flow conversion technology that automatically converts the causal knowledge graph schema into a processing flow and a prompt conversion technology that automatically converts it into specific LLM instruction prompts.

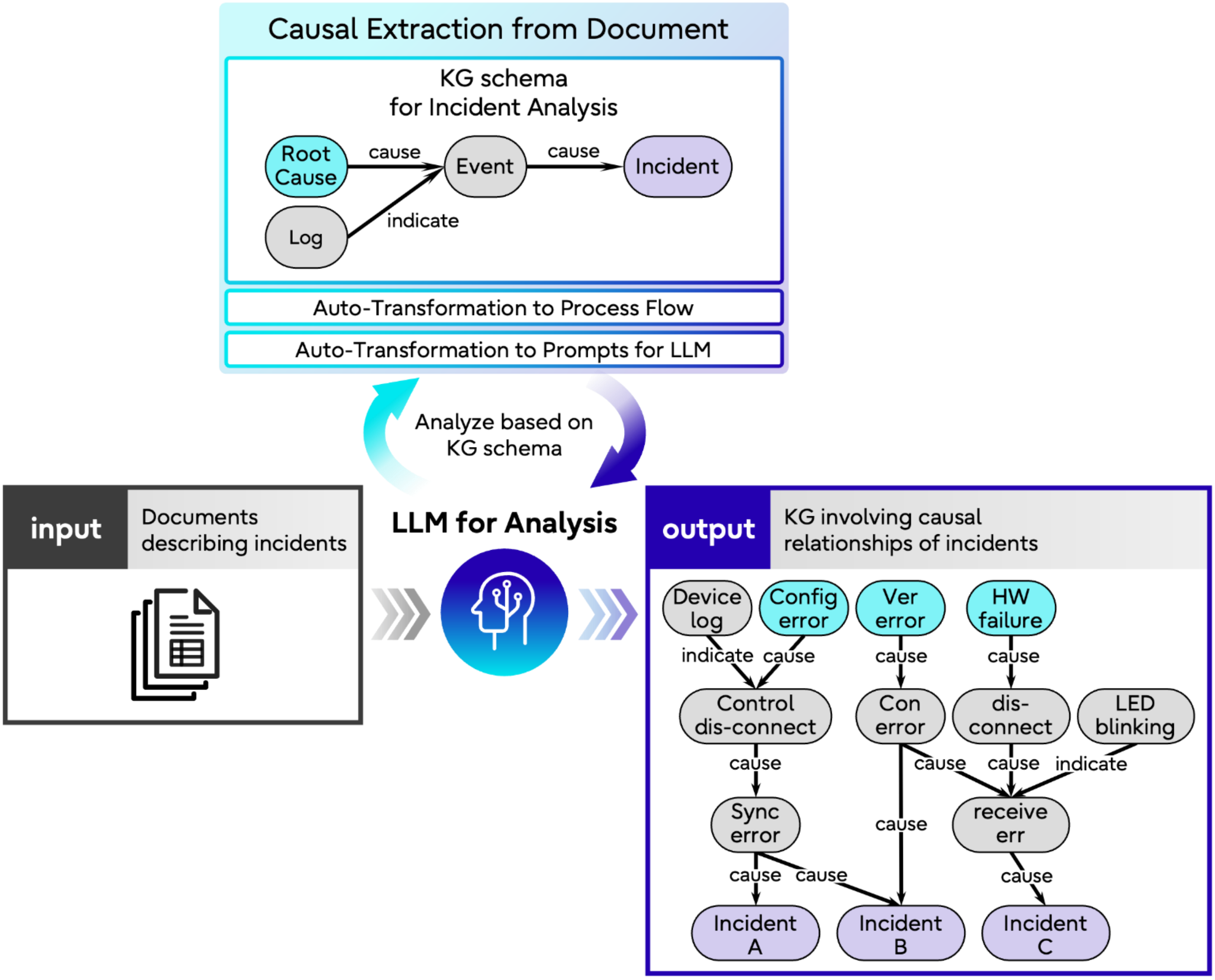


Figure 9: Overview of Causality Extraction Techniques from Documents

Statistical Causal Discovery from Numerical Data

Here, we introduce several technologies developed by Fujitsu to enhance the practicality of statistical causal analysis.

There is a high demand for understanding causal relationships that apply not to the entire dataset but only to data under specific conditions. In the medical field, knowing the drug reactions specific to patients with certain characteristics can accelerate insights and facilitate the repurposing of existing drugs. In the retail sector, understanding the differences in promotional effects among individuals can help in devising detailed sales strategies.

However, such causal relationships can only be known by collecting data under specific conditions, which creates a dilemma as these conditions cannot be known in advance. Therefore, we have developed a technology that rapidly enumerates all conditions to cover all variations of causal relationships and efficiently explores the conditions under which unknown causal relationships appear. We applied this technology to gene expression data in colorectal cancer and healthy colon tissues and successfully identified genes that are considered important for the classification of colorectal cancer automatically. Another challenge is that when the number of variables reaches thousands or tens of thousands, the enumeration and causal discovery require enormous computational power. By implementing this technology on Fugaku, we were able to complete calculations that previously would have taken 4,000 years in less than a day. Applying this technology to approximately 20,000 human genes, we successfully derived new insights related to drug resistance in lung cancer.

In addition, to improve practicality from different perspectives, we are also developing causal discovery technologies that capture more complex causal relationships. Traditional technologies primarily assumed that the cause and effect have simple linear relationships. However, in complex real-world systems and phenomena, there can be intricate causal relationships that cannot be expressed by linear relationships. Therefore, by utilizing nonlinear regression with neural networks, we have developed a technology to estimate more complex causal relationships among multiple variables [6].

In addition, to address the challenge of not having sufficient data for statistical causal discovery, Fujitsu has proposed a fusion technology that combines statistical causal discovery with agent-based modeling (ABM). By using ABM to model individual behaviors and interactions, it becomes possible to understand social phenomena and experiment with 'To-Be' scenarios. This technology can identify causal relationships between individual objectives, behaviors, environmental factors, and the analysis target using individual-level data and macro-level counterfactual data generated by ABM simulations. This supports the formulation and evaluation of realistic measures. To date, we have successfully proposed indirect measures to improve passenger experience at airports and presented their impacts [29], as well as generated explainable store product placement patterns that optimize customer experience and sales in retail stores [30].

Integrated Causal Discovery

In the current state of statistical causal discovery, it is difficult to integrate multiple datasets to estimate causal relationships. However, it is necessary to comprehensively determine causal relationships from numerous datasets observed under various conditions. For example, each hospital collects data on various test items, but it is unrealistic for a single hospital to conduct all the tests. Typically, in addition to items measured at any hospital, such as blood pressure, height, and weight, there are test items that are only conducted at specific hospitals. Although the measurement environments differ among hospitals, if these data could be integrated for causal discovery, it would enable cross-sectional analysis of the test items across different hospitals. To achieve this, we are developing a technology called 'integrated causal discovery,' which allows for causal discovery across multiple datasets.

Many approaches to discovering causal relationships are designed to learn a fixed causal relationship model from a single dataset, assuming that there are no hidden confounding factors. However, due to the cost of data acquisition, only a subset of variables may be measured for analysis. Therefore, there is a need for methods to discover causal relationships from multiple datasets with different variable sets, such as health checkup data measured by different hospitals or HR data collected by different companies. Fujitsu's integrated causal discovery can estimate the underlying data distribution using common variables as keys across datasets and construct a Causal Knowledge Graph that integrates datasets from different domains.

Causal Knowledge Transfer

Current statistical causal discovery technologies derive causal relationships based on a given dataset, but it is challenging to utilize those causal relationships for causal discovery in other datasets. One reason is that different datasets may not necessarily be generated according to the same causal structure, making it difficult to determine the extent to which causal relationships can be reused. While it is possible to use specific causal relationships as prior knowledge in causal discovery, the validity of such prior knowledge is currently judged manually by experts. However, when the number of variables exceeds several dozen, there are limitations to expert judgment. Fujitsu's causal knowledge transfer technology overcomes the limitations of current causal discovery, which is confined to the given dataset, by automatically obtaining the causal relationships with the least contradictions from the Causal Knowledge Graph as prior knowledge for the dataset whose causal structure is to be estimated. This technology enables the realization of a causal AI that continues to grow through the learning of large amounts of datasets.

4.3. Utilization Technologies of Causal Knowledge Graph

Root Cause Analysis

As one of the utilization technologies of the Causal Knowledge Graph, we are developing a technology to automatically analyze the root cause of network failures.

Figure 10 shows an example of utilizing the Causal Knowledge Graph extracted in Figure 9 for root cause analysis. By exploring and referencing the KG, appropriate causal relationships can be extracted, and potential failure causes can be enumerated comprehensively, considering various scenarios. This enables the generation of high-precision answers even for complex events that would typically fail in standard RAG. Additionally, it is possible to present the impact of each failure cause, verification procedures, and the knowledge necessary for narrowing down and recovery.

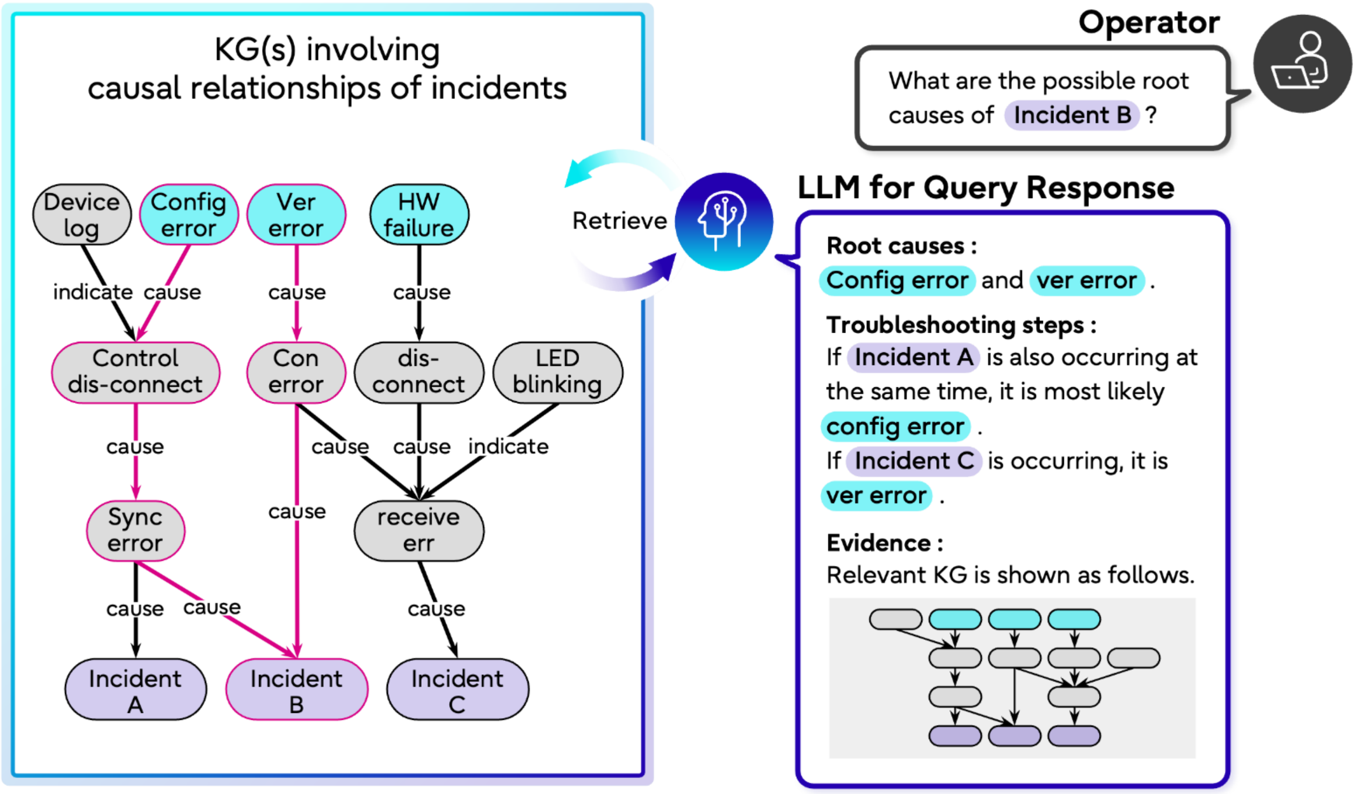


Figure 10: Overview of Root Cause Analysis Using Causal Knowledge Graph

Causal Decision-Making

Once the causal relationships between items in a given dataset are identified, it becomes possible to determine which items should be intervened on to achieve the desired goals, i.e., to formulate strategies. For example, in the HR domain, identifying the causal relationships between items from the results of an employee engagement survey can help formulate strategies to improve trust in management while maintaining employee productivity. However, even if the causal relationships between items are identified, it is difficult to manually derive effective strategies to achieve the goals. Fujitsu's causal decision-making recommends the most effective strategies to achieve the goals based on all possible causal relationships estimated from the data. The strategies recommended by this technology are optimized to achieve the goals at the lowest cost while minimizing adverse effects on items other than the goals. Applying this technology to the results of an engagement survey, we successfully derived new insights into strategies to improve employee productivity and trust in management.

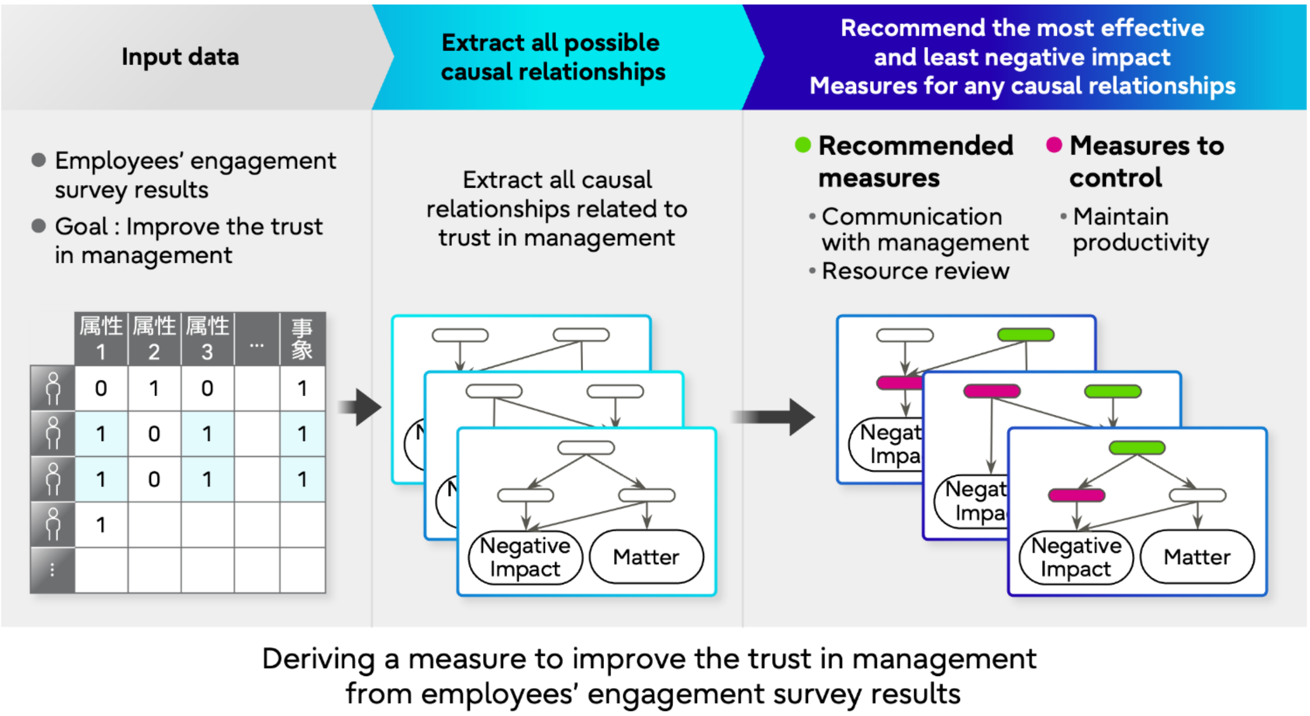


Figure 11: Example of Causal Decision-Making

Causal Knowledge Graph Reasoning (Question answering using LLM)

A Causal Knowledge Graph constructed from a vast number of documents and numerical data is expected to serve as a foundation for scientific discoveries and important decision-making. However, as the scale of the Causal Knowledge Graph increases and the relationships between items become more complex, interpreting its content, and utilizing it for reasoning becomes increasingly difficult.

Therefore, we are developing a reasoning technology for large-scale Causal Knowledge Graphs using LLMs. This technology divides the Causal Knowledge Graph into subgraphs and stores them in the LLM's RAG. By referencing these subgraphs during question answering, it can handle large-scale graphs and improve the accuracy of the answers generated by the LLM through a division suitable for reasoning. This corresponds to 'a. KG-enhanced LLMs' in the integration technologies of LLMs and KGs shown in Figure 3. Fujitsu refers to this format as 'KG-enhanced RAG.

Figure 12 is an example of reasoning based on a Causal Knowledge Graph generated from health checkup results and questionnaire data. In response to the question 'How can obesity be prevented?', the answer 'Maintain regular bowel movements' is provided based on the Causal Knowledge Graph, which is not included in general knowledge.

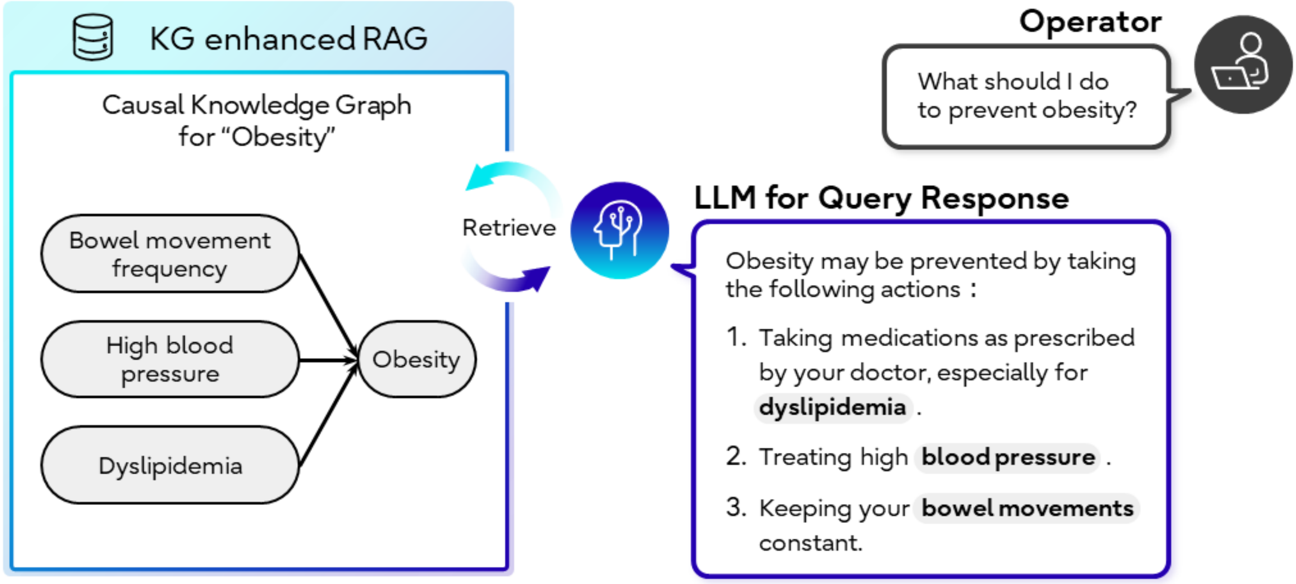


Figure 12: Example of Question Answering using Causal Knowledge Graph and LLM

5. Case Studies of Fujitsu's Causal Knowledge Graph

In this chapter, we introduce case studies where the Causal Knowledge Graph has been applied to three fields: 'ICT support,' 'human resources,' and 'healthcare and wellness.' In the ICT support field, a Causal Knowledge Graph of network failures is constructed from multiple communication device specifications and troubleshooting guides to analyze root causes. In the human resources field, integrated causal discovery is conducted across business negotiation data and employee engagement survey data to analyze the impact of engagement items on business wins. In the healthcare and wellness field, causal relationships derived from the health checkup big data promoted by Hirosaki University are used as prior knowledge to estimate valid causal relationships for other health checkup data.

5.1. Case Study 1: Analysis of Network Failure Causes

Overview:

In network deployment and operation, it is necessary to handle various devices connected in complex topologies, making it difficult to identify the causes of issues such as transmission errors or delays when they occur. On the other hand, the impact of a single failure can be significant, with potential losses amounting to billions of yen. Therefore, preventing failures and quickly recovering from them are critical challenges for network operators. Although some automation efforts are currently underway, the analysis and investigation of the diverse causes of failures—such as configuration errors, version incompatibilities, and hardware degradation—require a significant amount of effort from experts. As a result, there is a strong demand for technologies that can automate and support the analysis and investigation processes.

Therefore, to realize the cause analysis of network failures, we applied the Causal Knowledge Graph to automate the analysis of complex causal relationships of failures. In the future, we plan to expand the application to causal analysis in various fields, including cloud systems and factory equipment operations.

Datasets:

Troubleshooting guides, specifications, manuals, and other documents

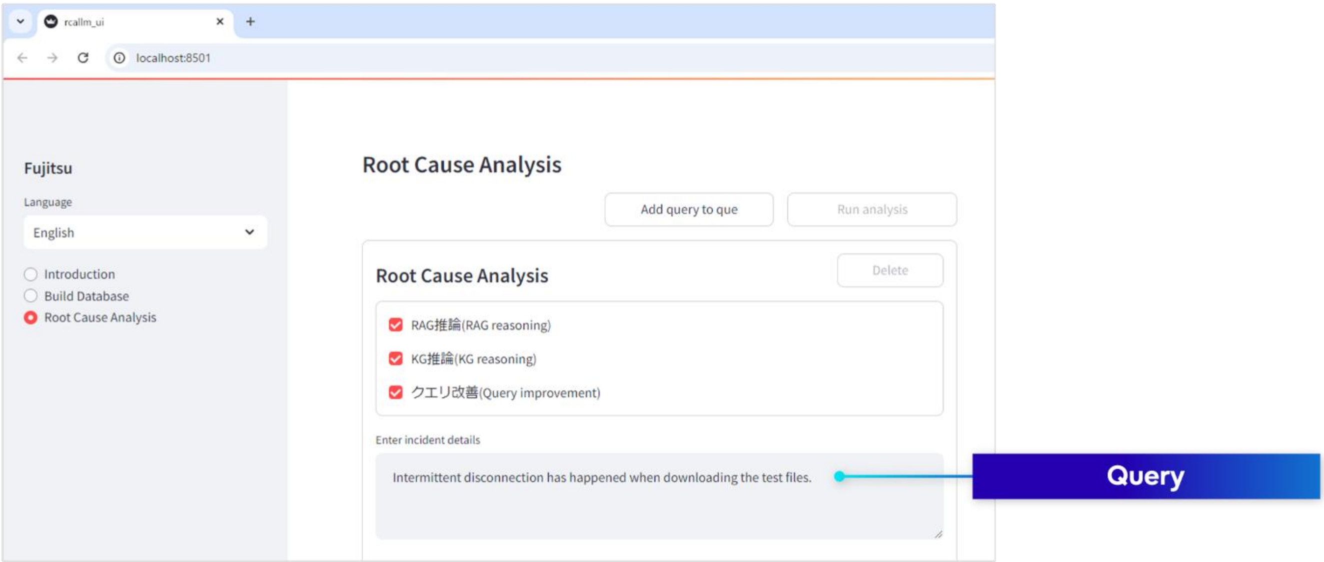
Technologies:

Causal Extraction from Documents, Root Cause Analysis

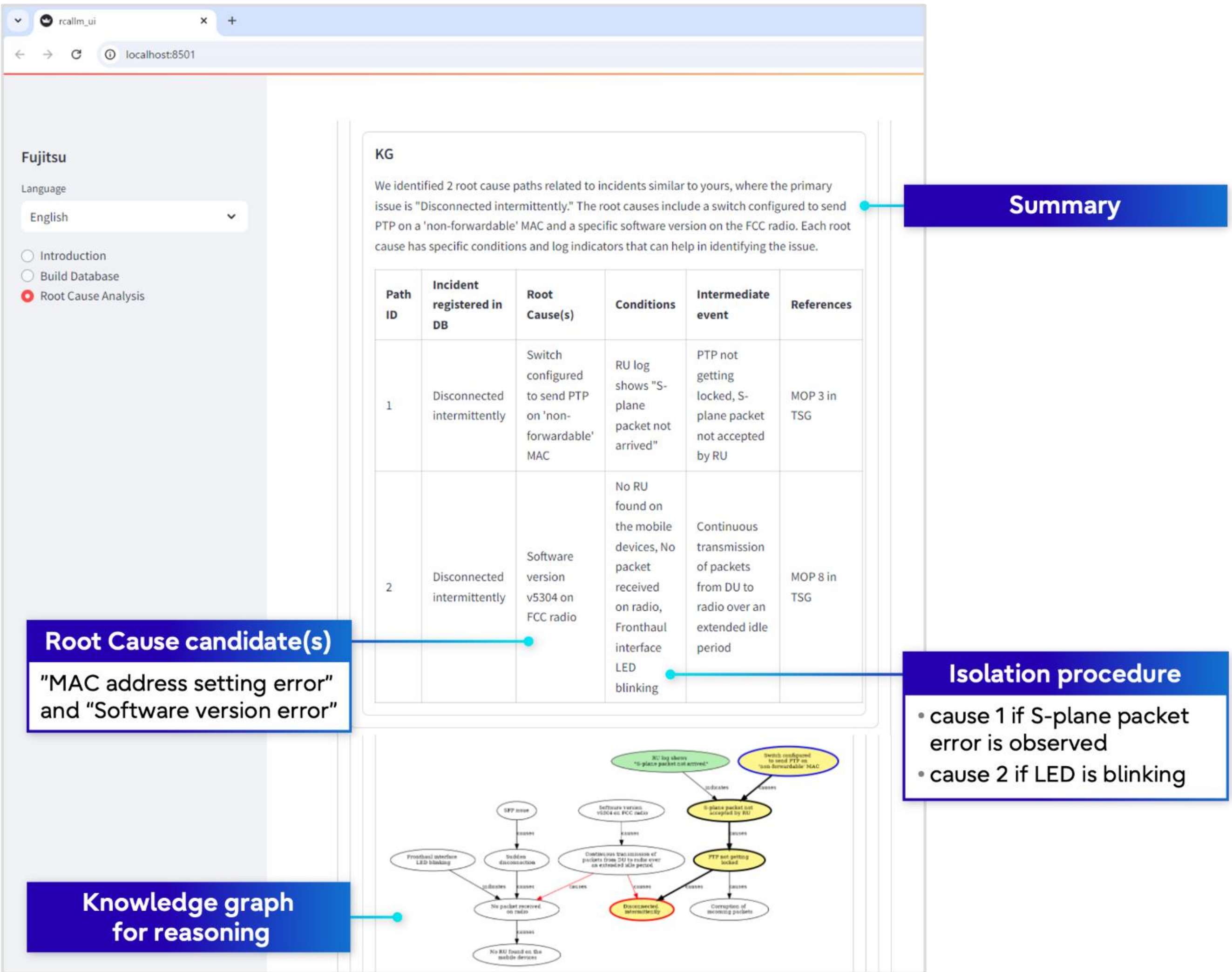
Detail Explanation of Actual Output:

In Figure 10 explained in the previous section, the operational overview of the case study applied to network failure cause analysis is shown. The left side of Figure 10 shows the Causal Knowledge Graph automatically extracted from troubleshooting guides and specifications. By exploring and referencing this Causal Knowledge Graph to extract appropriate knowledge, it is possible to enumerate potential failure causes comprehensively, considering various scenarios. This enables the generation of high-precision answers even for complex events that would typically fail in standard RAG. Additionally, it is possible to present the impact of each failure cause, verification procedures, and the knowledge necessary for narrowing down and recovery.

This technology is available as a web application called 'Fujitsu KG-enhanced RAG for Root Cause Analysis' in Fujitsu's PoC environment, 'Kozuchi.' Figure 13 shows an example screen of this web application. When the failure details are entered as a query, as shown in the upper part of Figure 13, the root cause analysis results are obtained along with the KG that serves as the basis for the judgment, as shown in the lower part of Figure 13.



Query input



Response

Figure 13: Example Screens of a Web Application for Root Cause Analysis Using a Causal Knowledge Graph (Top: Query Input Screen, Bottom: Answer Viewing Screen)

This allows experts to obtain investigation results immediately instead of reading and analyzing a large number of documents. Additionally, it can present more comprehensive and error-free recovery procedures, significantly reducing the time required for failure recovery.

5.2. Case Study 2: Causal Analysis Integrating Business Negotiation Data and Engagement Survey

Overview:

How 'improving employee engagement affects business wins' and ' how the results of business negotiations affect engagement survey scores' are of interest to the HR department. Additionally, in the sales field, there is a demand for initiatives that increase the win rate of business negotiations.

However, when formulating initiatives to improve the win rate of business negotiations using simple correlation analysis, there is a risk of identifying factors that are merely spurious correlations rather than actual causes of the win rate. Therefore, to minimize this risk, it is important to apply causal analysis techniques that clarify the cause-and-effect relationships between factors and allow for the verification of the extent to which each factor influences the outcome.

It is difficult to construct an integrated causal graph that spans different domains, such as business negotiation data and engagement survey data, using traditional methods. Therefore, we decided to use Fujitsu's integrated causal discovery technology to output a causal graph that spans both datasets. This technology estimates the causal graph by using common items as hints and assuming that both datasets are generated from a common causal structure. As a result, we were able to identify the relationships between business negotiations and engagement surveys. Furthermore, by applying causal decision-making to this causal graph, we were able to formulate initiatives to improve the win rate of business negotiations.

Datasets:

Data recording the details and outcomes of business negotiations, and employee engagement survey data

Technologies:

Statistical Causal Discovery, Integrated Causal Discovery, Causal Decision-Making

Detail Explanation of Actual Output:

Figure 14 is an example of the cross-analysis results of business negotiation data and engagement survey data (this is a model case and does not represent actual data). The red nodes represent items from the business negotiation data, and the blue nodes represent items from the engagement survey data. By observing the causal relationships between these two sets of items, we can understand the relationships between them.

Additionally, Figure 15 is an example of the initiatives output by causal decision-making to increase the business win rate ("Negotiation_Success") by 10%. By constraining the three causes—" Dedicated_Staff", "Customer_Meetings", and "Psychologica_Safety"—the actions to influence the items in the engagement survey are output.

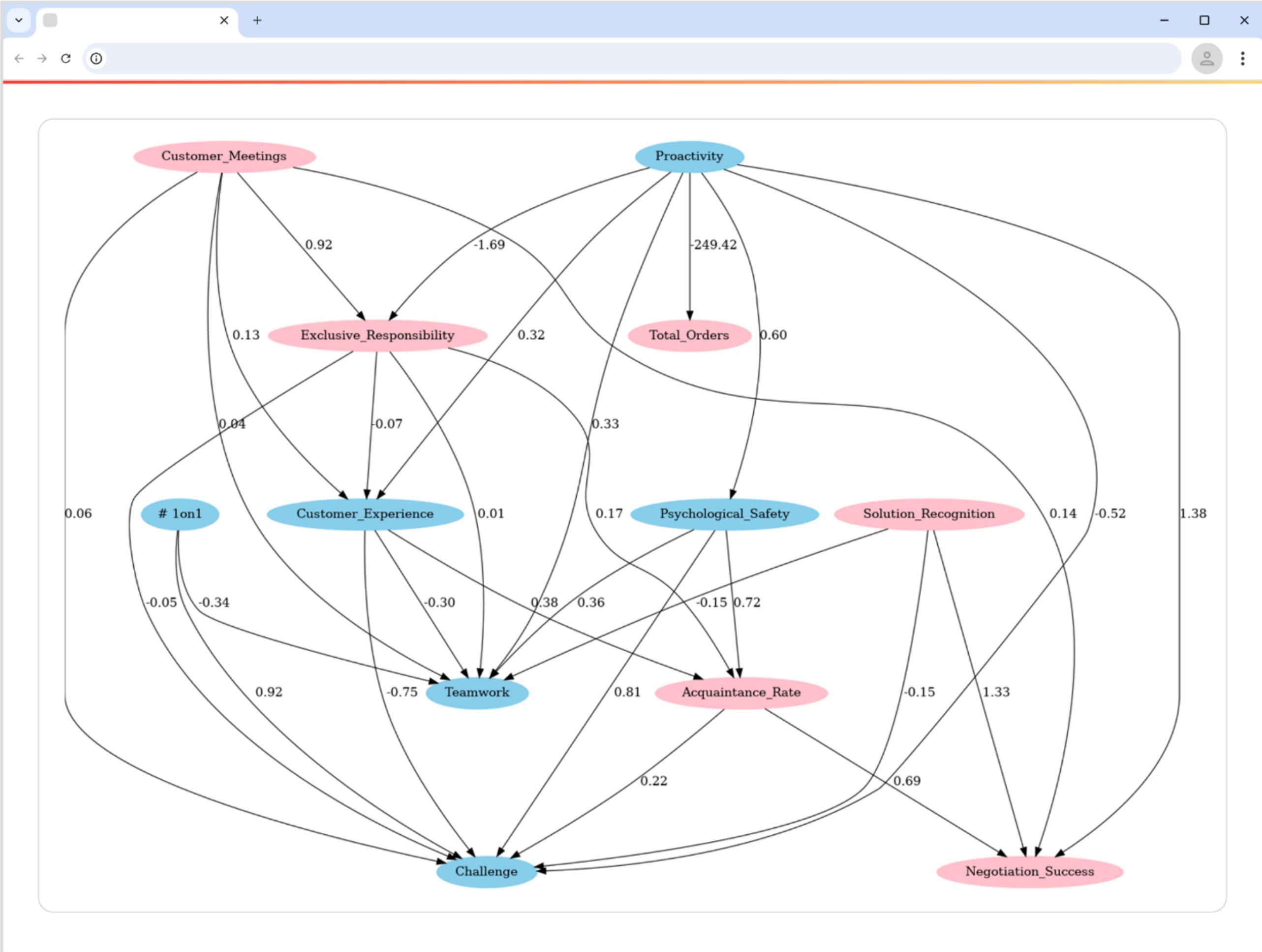


Figure 14: Integrated Causal Graph of Engagement and Business Negotiation Data

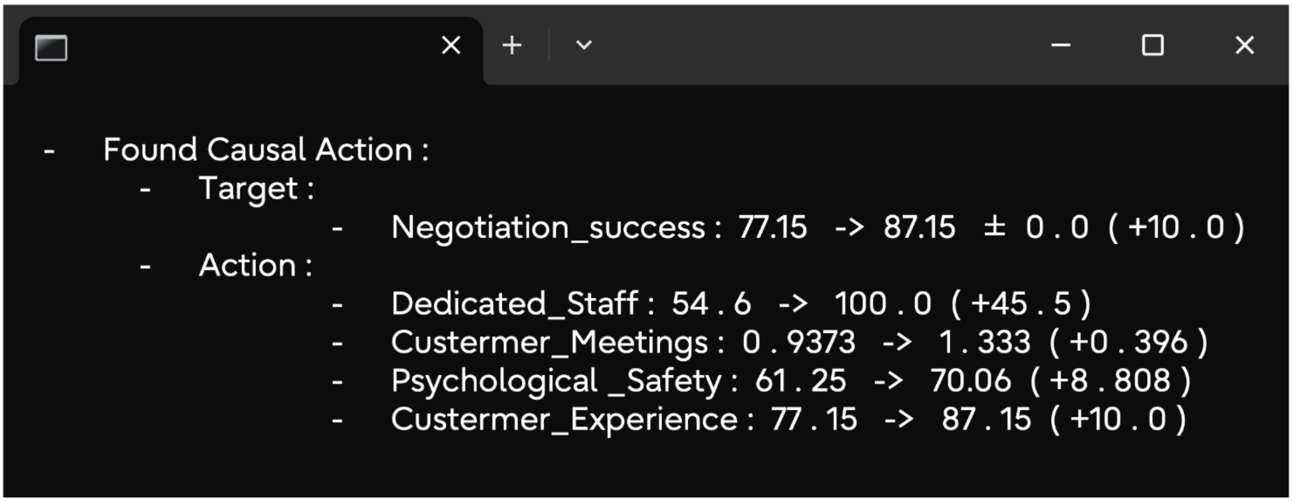


Figure 15: Example Output from Causal Decision-Making

5.3. Case Study 3: Estimation of Causal Relationships in Data on Sleep and Lifestyle

Overview:

Good quality sleep is widely known to be an important factor in supporting a healthy lifestyle, including reducing the risk of high blood pressure and heart disease, alleviating mental stress, regulating hormone balance, and suppressing overeating. So, what specifically should be done to improve sleep quality? To answer this question, it is essential to go beyond mere correlations and focus on the causal relationships with factors that affect sleep quality through data analysis. By clarifying which aspects of lifestyle are the 'causes' that influence sleep quality, more effective improvement measures can be identified.

However, identifying causal relationships based on a single dataset often involves various issues, such as insufficient sample size, the influence of confounding factors (unknown factors not included in the dataset), and biases in the data collection process. These issues can reduce the reliability of the analysis results. Fujitsu's causal knowledge transfer technology uses highly reliable known causal networks and converts them into information about causal relationships in unknown datasets, thereby improving the accuracy of causal discovery in unknown datasets.

This time, Kyoto University's research group applied their unique Bayesian network technology to the large-scale health checkup data led by Hirosaki University, 'Hirosaki University COI-NEXT Iwaki Health Promotion Project Health Checkup Big Data' [31], and constructed a highly reliable 'Hirosaki (Iwaki) Health Checkup Causal Network.' By using Fujitsu's causal knowledge transfer to apply this causal network to the causal discovery of a dataset on sleep and lifestyle, we were able to derive more valid causal relationships compared to when the Hirosaki Health Checkup Causal Network was not utilized.

Datasets:

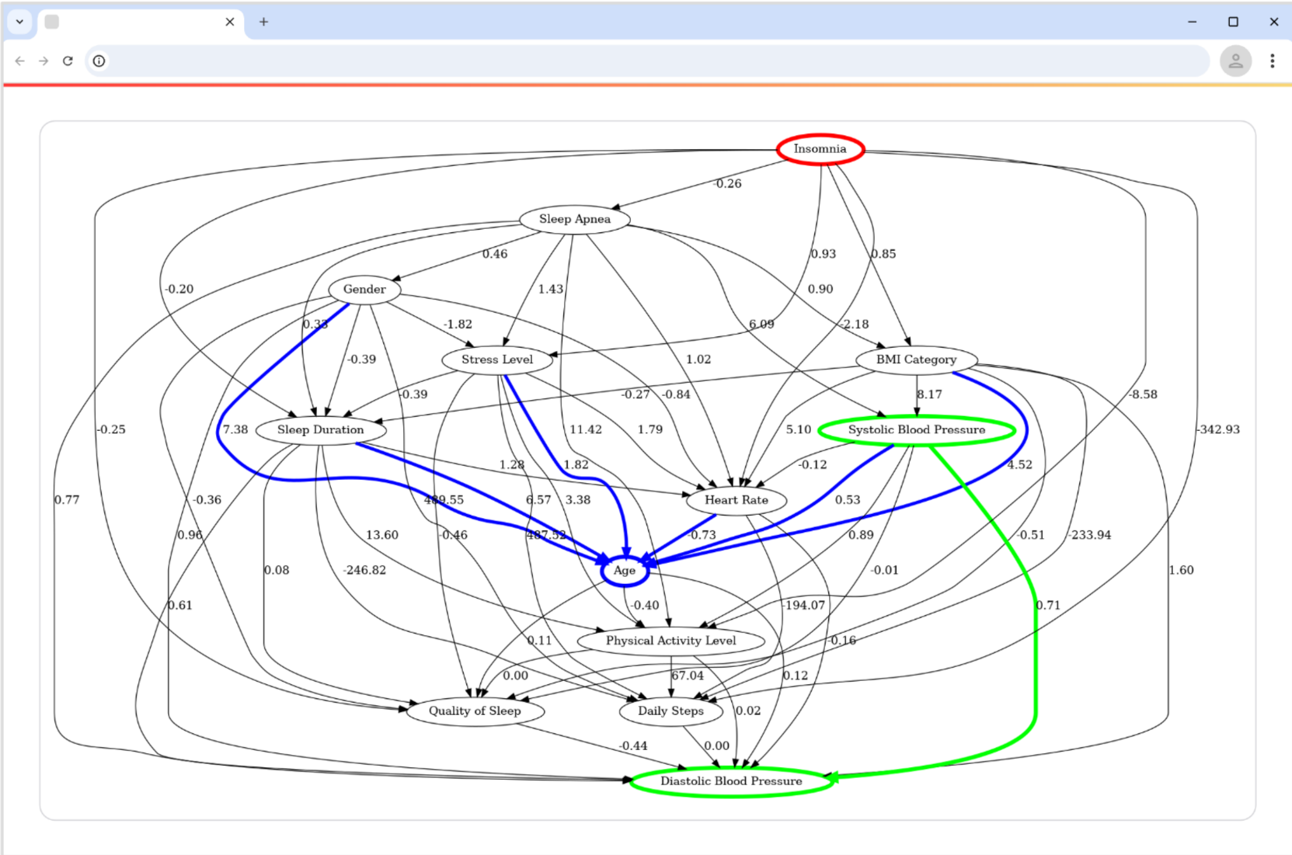
Hirosaki Health Checkup Causal Network, Sleep Health and Lifestyle Dataset [32]

Technologies:

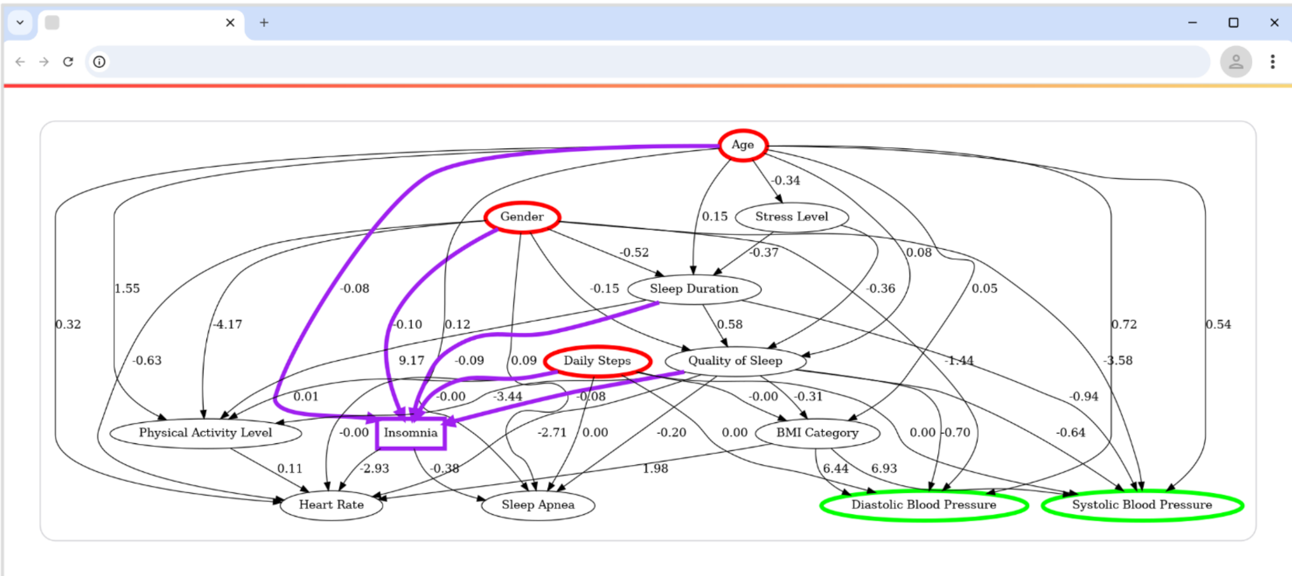
Statistical Causal Discovery, Causal Knowledge Transfer

Detail Explanation of Actual Output:

This time, we will execute two patterns when estimating causal relationships in the open data on sleep and lifestyle, 'Sleep Health and Lifestyle Dataset' (hereinafter referred to as SH data): one using causal knowledge transfer and one without using it. Figure 16 shows the execution results. Figure 16 (1) is the result of estimating causal relationships using only the SH data, and Figure 16 (2) is the result of estimating causal relationships in the SH data using the 'Hirosaki Health Checkup Causal Network' through causal knowledge transfer. By comparing the respective causal graphs, the following observations can be made.



(1) When estimating causal relationships using only the open data on sleep and lifestyle (SH data)



(2) When estimating causal relationships in the SH data
using the Hirosaki Health Checkup Causal Network as prior knowledge

Figure 16: Differences in the Estimation of Causal Relationships in Sleep and Lifestyle with and without the Utilization of the Hirosaki Health Checkup Causal Network

- The nodes enclosed by thick red lines are nodes that do not have parent nodes (nodes that are direct causes of the corresponding nodes) in each causal graph. In the causal graph utilizing the Hirosaki Health Checkup Causal Network (Figure 16 (2)), 'age,' 'gender,' and 'number of steps' are estimated as exogenous factors without parent nodes, which aligns with the intuition that these factors are not influenced by other nodes. On the other hand, in the causal graph not utilizing the Hirosaki Health Checkup Causal Network (Figure 16 (1)), only 'insomnia' is estimated as an exogenous factor, and 'age' and 'gender,' which were exogenous factors in the utilized causal graph, have parent nodes.
- In the causal graph without utilization (Figure 16 (1)), the causal relationships where 'age' is the direct cause are indicated in blue. For example, an obviously invalid causal relationship is estimated, such as 'sleep duration' being caused by 'age.'
- In the causal graph with utilization (Figure 16 (2)), the causal relationships where 'insomnia' is the direct cause are indicated in purple. It can estimate valid results, such as 'sleep duration' and 'sleep quality' being directly influenced by 'insomnia.'
- The nodes 'systolic blood pressure' and 'diastolic blood pressure' enclosed in green are values detected as a result of various health conditions, and it is counterintuitive for these to have a causal relationship with each other. In the causal graph without utilization (Figure 16 (1)), a causal relationship is detected from 'systolic blood pressure' to 'diastolic blood pressure,' but in the causal graph with utilization (Figure 16 (2)), these do not have a direct causal relationship.

In this way, by using causal knowledge transfer to leverage the information on causal relationships from the highly reliable known 'Hirosaki Health Checkup Causal Network,' it becomes possible to estimate more valid causal relationships even for datasets like the SH data, which would otherwise yield less reliable results in causal discovery.

6. Conclusion

In this white paper, we discussed the challenges faced by LLMs and the potential of Causal Knowledge Graphs as a promising solution. While LLMs are revolutionary in natural language processing, truly reliable decision support requires logical reasoning based on causal relationships, not just data correlations. Causal Knowledge Graphs have the potential to overcome this challenge and evolve AI from a mere information-providing tool to a more advanced decision support tool.

In Chapters 2 and 3, we detailed the technical background of 'causal analysis' and 'Knowledge Graph' that constitute the Causal Knowledge Graph, as well as the definition, use cases, and necessity of the Causal Knowledge Graph as envisioned by Fujitsu. In Chapter 4, we introduced the suite of technologies Fujitsu is working on for constructing and utilizing the Causal Knowledge Graph, with a particular focus on foundational technologies such as causal extraction from documents statistical causal discovery, and integrated causal discovery. In Chapter 5, we presented case studies of constructing and utilizing the Causal Knowledge Graph in three fields: 'ICT support,' 'human resources,' and 'healthcare and wellness.

Through these technologies, Fujitsu will advance the construction and application of Causal Knowledge Graphs in various fields. This will support more accurate and reliable decision-making in a wide range of areas, including strategic business decisions, treatment planning in healthcare, and development planning in sports. By combining cross-industry causal relationships into a Causal Knowledge Graph, Fujitsu aims to solve complex social issues through data-driven logical decision-making, thereby contributing to the realization of a better future society.

Lead Author :
Cognitive Graph Core Project
Kaoru Yokoo (Senior Project Director)

Authors :
Masatoshi Ogawa (Research Director);
Koji Maruhashi (Research Director);
Yuiko Ohta (Senior Research Manager);
Takuya Takagi (Senior Research Manager);
Kento Uemura; Takanori Ukai; Yoshihiro Okawa; Seiji Okajima; Ryosuke Ozeki;
Kentaro Kanamori; Tatsuya Kikuzuki; Hirofumi Suzuki; Shuang Chang;
Teruhisa Ninomiya; Shun Yanashima; Akihiro Wada

Supervisor :
Nobuyuki Igata (Senior Research Director)

A. References

[1] J. Pearl, Causality, Cambridge university press, 2009.

[2] D.B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions.," *Journal of the American Statistical Association,* vol. 100, no. 469, pp. 322-331, 2005.

[3] P. Spirtes, C. Glymour and R. Scheines, Causation, prediction, and search, MIT press, 2001.

[4] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research,* vol. 3, no. Nov, pp. 507-554, 2002.

[5] S. Shimizu, "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model," *Journal of Machine Learning Research-JMLR,* vol. 2, no. Apr, pp. 1225-1248, 2011.

[6] K. Uemura, "A multivariate causal discovery based on post-nonlinear model," Conference on Causal Learning and Reasoning, 2022.

[7] M. Takayama, T. Okuda, T. Pham, T. Ikenoue, S. Fukuma, S. Shimizu and A. Sannai, "Integrating Large Language Models in Causal Discovery: A Statistical Causal Approach.," CoRR abs/2402.01454, 2024.

[8] Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. T. Diab and B. Schölkopf, "Can Large Language Models Infer Causation from Correlation?," CoRR abs/2306.05836, 2023.

[9] W3C, "RDF (Resource Description Framework)," [Online]. Available: https://www.w3.org/RDF/.

[10] dbpedia, 2024. [Online]. Available: https://wiki.dbpedia.org/.

[11] yago, 2024. [Online]. Available: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/.

[12] googlekg, 2024. [Online]. Available: https://developers.google.com/knowledge-graph/.

[13] drugbank, 2024. [Online]. Available: https://www.drugbank.ca/.

[14] robokop, 2024. [Online]. Available: https://robokop.renci.org/.

[15] bio2rdf, 2024. [Online]. Available: http://bio2rdf.org/.

[16] fibo, 2024. [Online]. Available: https://www.omg.org/spec/FIBO/.

[17] govuk, 2024. [Online]. Available: https://www.gov.uk/.

[18] Openlink, "Virtuoso," [Online]. Available: https://virtuoso.openlinksw.com/.

[19] OntoText, "GraphDB," [Online]. Available: https://graphdb.ontotext.com/.

[20] Neo4j, "Neo4j Database," [Online]. Available: https://neo4j.com/.

[21] Amazon, "Amazon Neptune," [Online]. Available: https://aws.amazon.com/neptune/.

[22] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, "Unifying Large Language Models and Knowledge Graphs: A Roadmap," 2024. [Online]. Available: https://arxiv.org/abs/2306.08302.

[23] Glean Technologies, Inc., "How Glean's knowledge graph powers great enterprise search and generative AI," [Online]. Available: https://cdn.prod.website-files.com/6127a84dfe068e153ef20572/65f9da1c15f95fbf6866c112_Glean%20Knowledge%20Graph.pdf.

[24] T. Bratanič, "Knowledge Graphs & LLMs: Multi-Hop Question Answering," 2024. [Online]. Available: https://neo4j.com/developer-blog/knowledge-graphs-llms-multi-hop-question-answering/.

[25] C. Bharadwaj and S. Devi, "Building a Clinical Intelligence Engine using MedLM augmented Clinical Knowledge Graphs," 2024. [Online]. Available: https://cloud.google.com/blog/topics/healthcare-life-sciences/building-a-clinical-intelligence-engine-using-medlm?hl=en.

[26] Neo4j, Inc., "Neo4j Partners with Google Cloud to Launch New GraphRAG Capabilities for GenAI Applications," 2024. [Online]. Available: https://neo4j.com/press-releases/neo4j-google-cloud-graphrag-for-genai/.

[27] N. Maryam, "Synergizing Knowledge Graphs with Large Language Models (LLMs): A Path to Semantically Enhanced Intelligence," 2024. [Online]. Available: https://enterprise-knowledge.com/synergizing-knowledge-graphs-with-large-language-models-llms/.

[28] F. H. Messerl, "Chocolate Consumption, Cognitive Function, and Nobel Laureates," the New England Journal of Medicine, 367, 1562-1564., 2012.

[29] S. Chang, T. Kato, Y. Koyanagi, K. Uemura and K. Maruhashi, "An Iterative Analysis Method Using Causal Discovery Algorithms to Enhance ABM as a Policy Tool," 2023 Winter Simulation Conference (WSC), pp. 138-149, 2023.

[30] S. Chang, S. Yamane and K. Maruhashi, "Causal-based Rack Layout Optimization in Retail: Incorporating ABM and Causal Discovery," Winter Simulation Conference 2024, Florida, U.S., Dec. 2024, 2024.

[31] Hirosaki Univercity CO-NEXT, "Iwaki Health Promotion Project Health Checkup Big Data," 2024. [Online]. Available: https://coi.hirosaki-u.ac.jp/.

[32] Kaggle, "Sleep Health and Lifestyle Dataset," 2024. [Online]. Available: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset.

B. Fujitsu's Research Achievements

Research Achievements in Causal Analysis:

- Mar., 2017 : Fujitsu in Plans for RIKEN AIP-FUJITSU Collaboration Center
- Dec., 2020 : Fujitsu Develops Technology to Discover Characteristic Causal Relationships of Individual Data in Medicine, Marketing, and More
- Feb., 2021 : Fujitsu and Hokkaido University Develop "Explainable AI" Technology Providing Users with Concrete Steps to Achieve Desired Outcomes
- Feb., 2021 : Developing a new AI technology to recommend the optimal order of actions based on a counterfactual explanation
- Mar., 2022 : Fujitsu and Tokyo Medical and Dental University leverage world's fastest supercomputer and AI technology for scientific discovery to shed light on drug resistance in cancer treatment
- Apr., 2022 : Fujitsu and Atmonia leverage HPC and AI technology in joint project to contribute to carbon neutrality
- Feb., 2023 : Fujitsu and Atmonia succeed in development of new technology that accelerates search for disruptive catalyst for enabling sustainable ammonia production
- May, 2023 : Fujitsu, Kyoto University, and Chordia Therapeutics launch AI trials to discover biomarkers for new cancer drugs
- Oct., 2023 : Fujitsu and Atmonia discover a novel catalyst candidate for clean ammonia synthesis leveraging high-speed quantum chemical calculations

Research Achievements in Knowledge Graph

- Apr., 2013 : Fujitsu and DERI Revolutionize Access to Open Data by Jointly Developing Technology for Linked Open Data
- Jan., 2014 : Fujitsu Laboratories Develops Technology for Automatically Linking with Open Data throughout the World
- Feb., 2014 : Fujitsu Develops First-of-Its-Kind Assessment Tool that Visualizes a Community's Characteristics
- Sep., 2017 : Fujitsu Fuses Deep Tensor with Knowledge Graph to Explain Reason and Basis Behind AI-Generated Findings
- Nov., 2019 : Fujitsu Improves Efficiency in Cancer Genomic Medicine in Joint AI Research with the Institute of Medical Science at the University of Tokyo
- Oct., 2021 : Fujitsu and Aichi Cancer Center Develop AI System to Offer Patients Personalized Cancer Treatment
- May, 2024 : Fujitsu introduces 'explainable AI' for use in genomic medicine and cancer treatment planning
- May, 2024 : Fujitsu chosen for GENIAC project, starts development of large language models for logical reasoning