CRISP-DM 2.0 for the Semiconductor Industry and Other Complex Domains

Juan Manuel Gonzalez Huesca[1], **Mykola Pechenizkiy**[2]

[1]ASML Netherlands B.V., the Netherlands
[2]Department of Computer Science, TU Eindhoven, the Netherlands
juanmanuel.gonzalezhuesca@asml.com, m.pechenizkiy@tue.nl

Abstract

Data science plays a transformative role in the semiconductor industry by optimizing manufacturing processes, enhancing chip design, and strengthening global supply chains. It enables predictive maintenance, increases production yields, and accelerates R&D, helping the industry innovate to meet the demands of AI, IoT, and high-performance computing. While data science is accelerating scientific discovery and engineering design, the current data-centric approach is suboptimal, leading to many failed projects in this complex environment.

This paper challenges that prevailing data-centric approach, introducing a new paradigm that focuses on the understanding of the business challenges and research problems over the data itself. This shift, while seemingly subtle, demands a fundamental change in mindset, requiring data scientists to develop strong subject matter expertise and a deep understanding of the business needs. By effectively identifying and defining the business and research problems, data scientists can collaborate with their teams to develop solutions that create value, thereby unlocking the true transformative power of data science.

Introduction

Data science projects are unique due to their data-driven, interdisciplinary, collaborative, and exploratory nature, with a strong focus on innovation and iterative experimentation, especially in the semiconductor industry and other complex domains. Unlike software engineering projects, data science emphasizes exploration over engineering, involving a complex data lifecycle, specialized skills for model training and deployment, and management challenges due to uncertainty and unpredictable dependencies. The initial phase, involving business understanding and requirements gathering, is particularly challenging and requires multiple iterations, specialized expertise, and even feasibility studies to refine the project goals and success criteria, which may evolve over time.

Despite this complexity, data science projects are highly valuable for the semiconductor industry, in different areas, ranging from predictive maintenance, yield prediction and optimization, equipment health monitoring, fault detection and classification, and process control and optimization, among others. Currently, 42% of IT professionals at large organizations report that they have already deployed data-driven solutions, while an additional 40% are in the process to do it in the near future(Benchaita, Sarah 2024).

Background and Related Work

The strategic value of data science projects is recognized but many organizations struggle to scale these projects beyond the pilot phase, with McKinsey reporting that only 15% have successfully automated processes across multiple areas and only 36% have moved projects beyond pilot deployment(Panikkar, Rohit and Saleh, Tamim and Szybowski, Maxime and Whiteman, Rob 2021). These failures are often not due to technical limitations but to organizational challenges like lack of management understanding, resistance to change, and poor alignment with business strategy. Gartner estimates an 85% failure rate in data science projects, largely due to cultural misalignment and lack of clear strategic vision (Asay, Matt 2017), while Boston Consulting Group (BCG) indicate that 74% of organizations fail to achieve full data science potential due to a fragmented strategy, lack of organizational alignment, weak integration with the business needs and lack of a data-driven culture(Gregoire, Eric 2024). Even for the latest technologies, Gartner predicts that by 2025, 30% of Generative AI projects will not go beyond the proof of concept phase due to poor data quality, inadequate risk controls, escalating costs or unclear business value(Keen, Emma 2024). Despite high failure rates, organizations continue investing heavily in data science due to its transformative potential, competitive pressures, and the expectation of future advancements.

There are several structured methodologies to address some of these challenges and increase the success rate of data science projects, with the Cross-Industry Standard Process for Data Mining (CRISP-DM)1 being one of the most widely used for guiding data science projects. However, CRISP-DM has several limitations, particularly in high-stakes projects with significant uncertainty and risk, such as those in the semiconductor industry. These challenges include complex data (high-dimensional and evolving) and context (changing environment, dynamic production processes, per-batch needs), and scarcity of subject matter expertise and a effective feedback loop(Lijffijt et al. 2022),

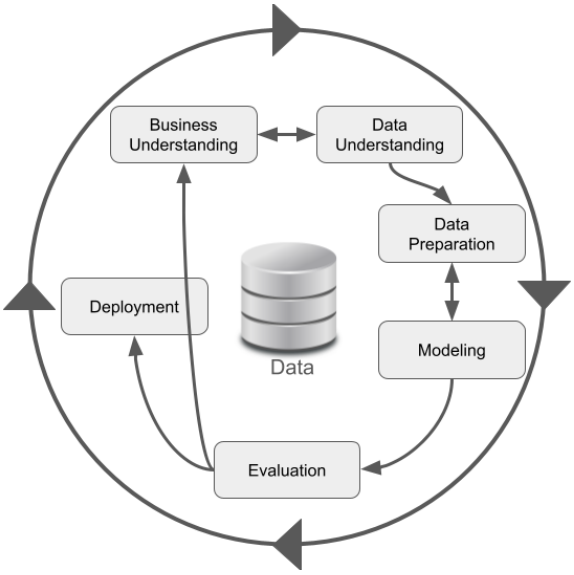

Figure 1: Original CRISP-DM diagram

challenges that require a closer, constant and frequent collaboration between data scientists and Subject Matter Experts (SMEs), while having all the time the business challenges and research problems as the primary project driver.

Argumentation for the Position

While data is an important element for success in data science projects, it's only an incentive, the most crucial factor is the understanding of the business challenges and the research problems. This understanding shapes the projects by enabling accurate problem framing, defining relevant research questions, deciding the expected outcome, selecting the most suitable success metrics, and describing necessary data characteristics, especially in the semiconductor industry and other complex domains where there is a high degree of uncertainty and business needs may evolve based on new findings and insights.

Recognizing the importance of business understanding and the limitations of CRISP-DM, we propose a new, enhanced framework that emphasizes placing the formulation of business challenges and research problems as the core of the process (2). This shift acknowledges that understanding and continuously refining these challenges and problems throughout the project lifecycle is critical for success. By integrating the problem formulation with each phase, this upgraded framework aims to increase the project's success rate and improve project outcomes, creating greater value. It facilitates a dynamic connection between evolving business needs and the corresponding data science efforts, ensuring that research questions adapt alongside changing project goals and objectives.

The problem formulation is central to the data science process and influences every phase through a bidirectional flow. Additionally, an essential aspect of this framework is the integration of ethics, morality, and legal considerations, which guide the responsible development of data-driven solutions by ensuring fairness, protecting personal data privacy, and promoting transparency and accountability. While the benefits of data science are often highlighted, the potential harm that models can inflict is frequently neglected. As data-driven solutions gain wider adoption across industries and domains, the importance of ethical considerations will increase, together with the compliance needs with new regulations.

This concept is not new. For example, "The Golden Circle", introduced by Simon Sinek in *Start With Why*, highlights a common issue in organizations: the tendency to prioritize "What" and "How" over "Why." Sinek argues that starting with "Why" fosters stronger connections by emphasizing purpose and meaning. Although not originally intended for data science, this concept is equally relevant, as it helps clarify the core purpose of a project. By focusing on the underlying reason for a data science initiative, such as improving yield prediction to enhance production efficiency and reduce waste, teams can ensure alignment with business objectives and avoid treating the project as merely a technical exercise. Our experience in the semiconductor industry shows that considering the project purpose while navigating through complexity helps improve the decision making process, and that establishing stretch goals can drive innovation and productivity, especially when broken down into concrete, clear and specific SMART goals(Doran 1981).

By applying this principle in data science, we acknowledge the importance of asking the right questions, especially "why?", to discover the real user needs. Data scientists should invest time and have longer planning phases as they prove being essential to build business understanding, learn domain-specific knowledge, and collaboratively prototype with stakeholders. To carry this process effectively, the importance of requirements elicitation and gathering should be emphasized, highlighting a range of methods to use, in-

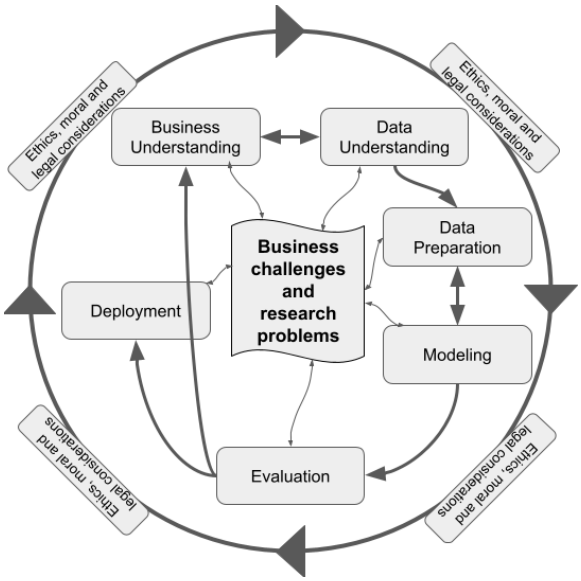

Figure 2: CRISP-DM 2.0: A new paradigm in data science projects

cluding but not limited to interviews, surveys, and observational methods with process mining and design thinking, ensuring clear and accurate captured needs.

Data scientists add value by clarifying the core business challenges and user needs, a task that often reveals that users and stakeholders, though familiar with data-driven tools, might lack a deep understanding of data science possibilities and limitations, which is why having a bidirectional knowledge sharing process (from stakeholders to data scientists and vice versa) within data science projects is very important. The latter, in addition to detailed and structured documentation, is crucial to guide project development, support knowledge sharing, and serve as a foundation for future work by recording technical details and facilitating compliance. Proper documentation encourages collaboration, aligns the team on project goals, and promotes accountability.

On the other side, although the key driver of success is the business understanding, having quality data can play an important role. Overall speaking, the more quality data the better, but challenges like small datasets can now be addressed with techniques such as few-shot learning and data augmentation.

It is essential for data scientists to establish a clear connection between the data and the problem being addressed. This involves critical assessments about data quality, availability, and the specific characteristics necessary for achieving the project goals and objectives. For example, sometimes data scientists start cleaning sensor data by removing outliers, assuming that they are errors or they are not relevant, this can lead to loss of critical data points, such as temperature spikes, which actually signaled early equipment malfunctions. This information can only be obtained through close collaboration with maintenance experts to understand the domain and map it to the project's needs.

Within the realm of data science, we believe that the project success should extend beyond simply achieving performance metrics. While metrics are useful, they are merely proxies and can risk misalignment with business goals if not continually aligned with the problem's real-world context. Data scientists must balance meeting current project goals with discovering new insights and findings that may lead to additional opportunities. This involves redefining success iteratively, as project goals and research questions evolve. This is why the relationship between model performance and business metrics needs to be validated and monitored over time to ensure the model's impact remains relevant and valuable.

Implications and Impact

This new framework requires a different mindset and skill set. To unlock data science's full value, data scientists must play a key role by developing new skills that connect technical expertise with business understanding, framing projects around actual business problems, and keeping adapting them throughout the entire project. Developing domain knowledge and refining problem-definition skills are essential, as highlighted by Foster Provost and Tom Fawcett in *Data Science for Business*, to create effective, strategically aligned solutions.

As stated before, the success of a data science project depends on understanding business challenges and research problems rather than merely focusing on data or machine learning models. Key elements include effectively framing problems with appropriate research questions, clear outcomes, and suitable success metrics while defining data requirements. To enable and professionalize this process, a new role has been introduced in some organizations: the analytics translator, with McKinsey predicting a demand for two to four million of these professionals in the U.S. by

2026(Henke, Nicolaus and Levine, Jordan and McInerney, Paul 2018). Although not yet widespread in data science teams, data scientists should develop essential skills of this role, which acts as a bridge between business goals and technical requirements.

To illustrate this concept with an example, imagine working in a predictive maintenance project that is focused on estimating the remaining useful life (RUL) of a chemical vapor deposition (CVD) chamber to proactively schedule maintenance. However, by understanding the business needs, the real business challenge may not be the timing of component failure but rather contamination in the chamber, which led to reduced product yield and wafer rejections. Shifting the focus from RUL prediction to real-time fault diagnosis can allow to identify contamination early, pinpoint its causes, and take corrective actions to maintain process stability.

The successful implementation of this framework requires effective and efficient knowledge sharing, strong collaboration between data scientists and all stakeholders, and a flexible, adaptive way of working.

The interdisciplinary nature of data science demands robust bidirectional knowledge-sharing, communication and collaboration processes to create impactful solutions. Formal and informal educational efforts help bridge knowledge gaps, allowing data scientists to communicate complex ideas in simple terms. Effective collaboration is central to data science projects, enhancing work quality by leveraging diverse expertise and maintaining alignment on evolving project goals, as outlined in CRISP-DM 2.0.

Successful data science projects rely not only on structured phases but also on effective working methods. Best practices suggest that combining a flexible project lifecycle with an agile approach yields optimal results. Agile methodologies enable teams to iterate quickly and adapt to shifting requirements, evolving business challenges, and new insights discovered during the project. Additionally, employing the scientific method is crucial, as it provides a systematic framework guiding data scientists from hypothesis formulation to conclusion based on rigorous data analysis. While there are general guidelines, they should be adapted, as each team, organization, and industry has its unique working dynamics shaped by culture, values, and leadership.

SMEs are essential to the success of data science projects, especially in the initial phases of business understanding and requirements gathering. SMEs define key research questions, which data scientists then translate into data science terms. Because data science projects are often exploratory, SMEs' adaptability is valuable as project goals evolve with new insights. Their involvement fosters trust, adoption, and stronger solutions by aligning models with practical needs and embedding expert feedback into model development. Furthermore, they bring critical domain insights for data labeling, bias detection, and model fairness.

Conclusions

The iterative, exploratory, and experimental nature of data science projects introduces significant complexity and uncertainty, particularly in domains like the semiconductor industry. Major challenges, such as inadequate translation between business needs and technical solutions and a lack of adaptability to evolving requirements or insights, often lead to project failures. To address these issues, we propose an updated CRISP-DM framework that places business challenges and research problems at its core, rather than data, to enhance project success rates and unlock the true transformative power of data science. Achieving this requires a mindset change, the development of new skills, and a stronger collaboration with all stakeholders, especially with SMEs. This new paradigm not only redefines the approach to data science projects but also fosters the creation of more impactful and sustainable solutions in an evolving and increasingly complex environment driven by the growing demand for smaller microchips.

References

Asay, Matt. 2017. 85% of big data projects fail, but your developers can help yours succeed. https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/. Accessed: 2024-11-17.

Benchaita, Sarah. 2024. IBM Global AI Adoption Index 2023. https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters. Accessed: 2024-11-17.

Doran, G. T. 1981. There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review*, 70(11): 35–36.

Gregoire, Eric. 2024. AI Adoption in 2024: 74% of Companies Struggle to Achieve and Scale Value. https://www.bcg.com/press/24october2024-ai-adoption-in-2024-74-of-companies-struggle-to-achieve-and-scale-value. Accessed: 2024-11-17.

Henke, Nicolaus and Levine, Jordan and McInerney, Paul. 2018. Analytics translator: The new must-have role. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Analytics%20translator/Analytics-translator-The-new-must-have-role.pdf. Accessed: 2024-11-17.

Keen, Emma. 2024. Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025. https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025. Accessed: 2024-11-17.

Lijffijt, J.; Gkorou, D.; Van Hertum, P.; Ypma, A.; Pechenizkiy, M.; and Vanschoren, J. 2022. Introduction to the Special Section on AI in Manufacturing: Current Trends and Challenges. *SIGKDD Explor. Newsl.*, 24(2): 81–85.

Panikkar, Rohit and Saleh, Tamim and Szybowski, Maxime and Whiteman, Rob. 2021. Operationalizing machine learning in processes. https://www.mckinsey.com/capabilities/operations/our-insights/operationalizing-machine-learning-in-processes. Accessed: 2024-11-17.