INSTITUTE
OF ECONOMICS

Scuola Superiore
Sant'Anna

LEM | Laboratory of Economics and Management

Institute of Economics
Scuola Superiore Sant'Anna

Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy
ph. +39 050 88.33.43
institute.economics@sssup.it

LEM
WORKING PAPER SERIES

Measuring Data-Driven Innovation in Firms: A Multi-Source Approach Using Large Language Models

Diletta Abbonato [1,2], Gianluca Biggi [3], Carlo Bottai [3], Lisa Crosato [4], Roberto Fontana [5], Marco Guerzoni [2], Caterina Liberati [2], Arianna Martinelli [3], Noman Mustafa [3], Massimiliano Nuccio [4], Prince Oguguo [5,6], Stefania Scrofani [3,7], Federico Tamagni [3]

[1] University of Turin, Italy
[2] University of Milan-Bicocca, Italy
[3] Scuola Superiore Sant'Anna, Pisa, Italy
[4] University of Venice Ca' Foscari, Italy
[5] University of Pavia, Italy
[6] Neoma Business School, France
[7] University of Tubingen, Germany

2026/27 September 2026
ISSN(ONLINE) 2284-0400
DOI: 10.57838/sssa/z66h-8g34

Measuring Data-Driven Innovation in Firms: A Multi-Source Approach Using Large Language Models*

Diletta Abbonato[a,b], Gianluca Biggi[c], Carlo Bottai[c], Lisa Crosato[d], Roberto Fontana[e], Marco Guerzoni[b], Caterina Liberati[b], Arianna Martinelli[c], Noman Mustafa[c], Massimiliano Nuccio[d], Prince Oguguo[f,e], Stefania Scrofani[g,c], and Federico Tamagni[c]

[a]University of Turin,Italy
[b]University of Milan-Bicocca, Italy
[c]Scuola Superiore Sant'Anna, Italy
[d]Ca' Foscari University of Venice, Italy
[e]University of Pavia, Italy
[f]Neoma Business School, France
[g]University of Tubingen, Germany

Abstract

This study presents a comprehensive framework for identifying and measuring Data-Driven Innovation (DDI) at the firm level. We develop a conceptual definition of DDI grounded in the chain-linked model of innovation, where data infrastructure and advanced analytics augment each stage of the innovation process. Because DDI is intangible, embedded, complex, and heterogeneous, traditional innovation metrics fail to capture it. We therefore propose a multi-source measurement strategy exploiting Large Language Models (LLMs) applied to four distinct textual data sources: corporate communications, patents, company websites, and startup descriptions. The identification strategy relies on detecting the co-presence of *data-driven means* (i.e., technologies, infrastructure, algorithms) and *innovative purpose* (creating, improving, or optimizing products, processes, or services) within the same textual unit. Applying this framework to datasets covering listed firms in Italy (102 firms) and the US (3,714 firms), 3.4 million USPTO patent filings, 134,208 Italian SME websites, and 9,705 Crunchbase startup profiles, we document the diffusion, intensity, and heterogeneity of DDI across firm size, age, industries, and countries.

Keywords: Data-Driven Innovation, text analysis via LLMs, firm characteristics
JEL codes: 030, D22, C55

*This study present part of the results gathered within the framework of the "DDI-EMPIRICS" project, financed by the Italian Ministry of University and Research under the PRIN22 initiative.

1 Introduction

Data have become a central input into contemporary innovation processes. Firms increasingly collect information generated by customers, production systems, connected products, transactions, and research activities, and use analytical techniques to transform these data into new products, services, processes, and business models. This development has given rise to what is commonly described as Data-Driven Innovation (DDI hereafter): the systematic use of data and analytics to support the creation or significant improvement of economic value (Kusiak, 2009; OECD, 2015; Luo, 2022). DDI can affect virtually every stage of innovation, from identifying unmet market needs and exploring alternative designs to testing prototypes, optimising production, and learning from product use after commercialisation.

The economic significance of DDI extends beyond the digital and ICT sectors. Manufacturing firms use sensors and real-time analytics to improve production and develop predictive-maintenance services. Healthcare organisations analyse clinical data to design treatments and diagnostic tools. Agricultural firms exploit satellite and sensor information to optimise production. Retailers use transaction and behavioural data to customise products and services. Data-driven approaches may therefore operate as a general enabling capability whose applications vary across sectors, organisational functions, and types of firms. A growing literature accordingly associates data and analytical capabilities with innovation, productivity, and competitive performance (Brynjolfsson et al., 2011; OECD, 2015; Fosso Wamba et al., 2015; Mikalef et al., 2019; Ghasemaghaei and Calic, 2020).

Despite its growing importance, DDI remains difficult to observe and measure at the firm level. Existing indicators fall short of capturing the specific role played by data in the innovation process. R&D expenditure does not indicate whether investments are data-driven. Patent counts exclude many organisational, service, process, and business-model innovations and are especially incomplete for smaller firms and firms operating in sectors with a low propensity to patent. Measures of digital-technology adoption, meanwhile, may identify whether a firm uses cloud computing, artificial intelligence, data storage, or Internet of Things technologies, but not whether these technologies are employed for innovation rather than for routine operational or administrative purposes.

The intangible, embedded, and heterogeneous character of DDI complicates measurement. DDI may take the form of an algorithm incorporated into an existing product, an analytical capability used to redesign a production process, a system for learning from customer behaviour, or a new value proposition revolving around data. Such activities frequently leave no trace. At the same time, simple references to "data," "analytics," "cloud," or "artificial intelligence" cannot be treated as sufficient evidence of DDI. A firm may acquire cloud infrastructure without changing its products or processes, while another may launch a new product without relying on data-intensive methods. Identifying DDI therefore requires determining both whether data-related technologies and innovation activities are mentioned, and whether the former are functionally connected to the latter.

Existing text-based indicators only partially resolve this problem. Keyword dictionaries, patent classifications, and term-frequency measures have proved valuable for mapping technological change and firm-level digitalisation (Cockburn et al., 2018; WIPO, 2019; Webb, 2020; Kinne and Lenz, 2021; Arts et al., 2021; Hain et al., 2022; Ashouri et al., 2024). However, dictionary-based approaches are tied to predetermined vocabularies. They may fail to identify new terminology and semantically equivalent expressions. They may also generate false positives when terms appear in an unrelated context. More importantly, conventional keyword approaches are not well suited to evaluating whether a data-driven means is semantically linked to an innovative purpose.

This report addresses these limitations by developing a Large Language Model–based framework for identifying and measuring DDI across heterogeneous firm populations and textual sources. We define a textual unit as containing evidence of DDI when three conditions are jointly satisfied. First, the text must refer to a data-driven means, including data assets, analytical methods, or technological infrastructure. Second, it must describe an innovative purpose involving the creation, improvement, or optimisation of a product, process, service, or business model. Third, the two elements must be semantically dependent, meaning that the text must indicate that the data-driven means enables, supports, or is applied to the innovative purpose. DDI is therefore distinguished both from the routine adoption of digital technologies and from innovation that does not rely on data.

The conceptual framework is grounded in the chain-linked model of innovation developed by Kline and Rosenberg (1986). Rather than treating innovation as a linear progression from research to commercialisation, the chain-linked model emphasises interactions among market identification, design, testing, production, distribution, and multiple feedback mechanisms. We conceptualise DDI as an augmentation of these stages and linkages. Data and analytics may help firms infer emerging demand, expand the space of potential designs, conduct simulation and virtual testing, optimise production in real time, personalise market offerings, and accelerate feedback from users to design and market identification. This perspective shifts attention from individual digital technologies to the organisational function they perform within the innovation process.

The theoretical framework has a direct measurement implication. Because data can support innovation at different stages and in different organisational contexts, DDI cannot be captured adequately through a single technological proxy or documentary source. Different firms leave observable traces of DDI in different places. Listed firms discuss strategic initiatives in public communications. Patenting firms codify data-driven technical solutions in patent documents. Small and medium-sized enterprises describe their products and capabilities on corporate websites. Startups present their technologies and value propositions on entrepreneurial information platforms. A multi-source strategy is therefore needed to observe complementary manifestations of DDI.

We implement this strategy using four large textual datasets. The first consists of corporate communications produced by 102 Italian and 3,714 US listed firms over the period 2008–2022. The second contains approximately 3.4 million USPTO patent documents covering the period 2010–2021. The third comprises the websites of 134,208 Italian small and medium-sized enterprises. The fourth includes company descriptions for 9,705 startups listed on Crunchbase. Together, these sources cover established and newly founded firms, large listed corporations and SMEs, patenting and non-patenting organisations, and both technical and market-facing representations of innovation.

The empirical implementation combines two complementary Natural Language Processing strategies. For corporate communications, where references to DDI are dispersed across long narratives, we fine-tune a RoBERTa encoder model to classify sentences and construct a continuous firm-year measure of DDI intensity. For patent documents, corporate websites, and Crunchbase descriptions, which provide more focused accounts of technologies, products, and capabilities, we use a prompt-engineered Gemma model. The implementation varies with the structure of each source, but both approaches apply the same underlying identification principle based on the semantic linkage between data-driven means and innovative purposes.

The analysis documents several common patterns. First, DDI is becoming increasingly prominent. The mean share of DDI-related text in US corporate communications rises over the sample period, while the proportion of USPTO patents classified as DDI approximately doubles, from 9% in 2010 to around 18% at the end of the period. The Crunchbase evidence similarly indicates that DDI is more prevalent among recent founding cohorts. The convergence of these trends across technical documents, corporate disclosures, and entrepreneurial descriptions suggests that the increasing visibility of DDI is not specific to a single population or communication setting.

Second, DDI is pervasive but unevenly distributed. Information and Communication consistently appears among the most DDI-intensive sectors, but substantial DDI activity is also detected in manufacturing, professional services, construction, utilities, transportation, agriculture, accommodation, healthcare, and wholesale and retail. This broad sectoral reach supports the interpretation of DDI as an enabling capability rather than a phenomenon confined to firms producing digital technologies.

Third, DDI is broader than artificial intelligence. Among the 447,827 patents classified as DDI, 59% do not satisfy the WIPO criteria for AI patents. AI and machine learning are important instruments for DDI, but DDI also encompasses statistical analytics, cloud-based data infrastructures, sensor systems, real-time optimisation, automation, and other methods of extracting innovative value from data. Equating DDI with AI would consequently exclude a large part of the phenomenon that this report seeks to measure.

Fourth, the relationship between DDI and firm characteristics depends on the margin being observed. Larger firms tend to exhibit greater absolute DDI activity and stronger DDI visibility in corporate communications. Among patenting firms, however, smaller and younger firms display a higher share of DDI patents in their portfolios, even though large and mature firms produce more DDI patents in absolute terms. This distinction between the scale of DDI and specialisation in DDI helps reconcile apparently different size and age patterns across sources. Established firms may incorporate DDI into broad and diversified innovation portfolios, while smaller and younger firms may build more narrowly specialised technological and market positions around data.

Given the heterogeneity of the data sources, these findings may not be directly comparable. However, as no individual source provides a complete measure of DDI, their combination reveals patterns that would remain hidden or potentially distorted when relying on a single indicator. Corporate communications capture strategic attention and external signaling. Patents capture codified technological output. Websites capture market-facing capabilities. Crunchbase descriptions capture entrepreneurial orientation and venture identity.

This study makes three main contributions. First, it provides a conceptually grounded definition of DDI that distinguishes the innovative use of data from generic digitalisation, routine analytics, and artificial intelligence as a technology category. By embedding DDI in the chain-linked model, the framework identifies the functions that data and analytics can perform throughout the innovation process. Second, from the methodological viewpoint, the report introduces a semantic identification strategy based on the functional relationship between data-driven means and innovative purposes. The use of LLMs makes it possible to evaluate contextual meaning and semantic dependency rather than relying exclusively on fixed keyword lists. The multi-source design further addresses the selective coverage of individual innovation indicators and provides a framework that can be adapted to different firm populations and documentary settings. Third, the report provides a large-scale empirical mapping of DDI across countries, industries, firm sizes, firm ages, and organisational forms.

All in all, the results provide a descriptive foundation for future research on the determinants and consequences of DDI, including its relationships with productivity, employment, knowledge spillovers, sustainability, and firm performance. They also indicate that policies supporting DDI should consider both the broad sectoral applicability of data-driven approaches and the different constraints faced by large diversified firms and smaller specialised organisations.

The remainder of the report is organised as follows. Section 2 reviews the background literature on DDI, its economic implications, and its relationship with artificial intelligence and related concepts. Section 3 develops the theoretical framework, explaining how data and analytics augment the stages and feedback mechanisms of the chain-linked innovation model. Section 4 presents the LLM-based measurement strategy and the two complementary methodological implementations. Section 5 describes the four data sources. Section 6 reports the descriptive evidence from corporate communications, patent documents, Italian SME websites, and Crunchbase profiles, and compares the resulting patterns across sources. Section 7 concludes by discussing the principal findings, policy implications, limitations, and directions for future research.

2 Data-Driven Innovation: Concept and Literature

The concept of Data-Driven Innovation (DDI) has emerged over the past two decades as a fundamental paradigm in innovation studies, reflecting the growing role of data as a strategic asset in value creation. The term was first formally introduced by Kusiak (2009) who proposed a "data-driven approach" to innovation in which large volumes of data generated by modern production systems serve as the primary input for discovering patterns, optimising processes, and generating novel solutions. In this view, data should not be seen as a mere by-product of business operations but an essential resource that directly feeds the innovation process.

The most widely cited institutional definition comes from the OECD report *Data-Driven Innovation: Big Data for Growth and Well-Being*, which defines DDI as "the use of data and analytics to develop or foster new products, processes, organizational methods and markets." OECD (2015, p. 21). This broad definition encompasses the full spectrum of Schumpeterian innovation activities – from product and process innovation to organisational and market innovation – provided that data and analytics play a central role. The OECD report emphasises that DDI is beginning to "transform all sectors in the economy, including low-tech industries and manufacturing" OECD (2015, p. 17), thereby challenging the view that data-centric innovation is confined to high-technology sectors.

A critical conceptual contribution was made by Trabucchi and Buganza (2019) who articulated DDI as a perspective switch in which "data is no longer only the output of a value proposition but rather becomes an input for the generation of a new value proposition" (p. 25). Whereas traditional firms generate data as a side effect of operations (e.g., transaction records, sensor logs), data-driven innovators deliberately collect, process, and analyse data as the starting point for creating new value. This reframing is fundamental to distinguish DDI from the mere digitalisation of existing processes.

More recently, Luo (2022) has provided a systematic conceptualisation, arguing that DDI should be understood as a distinct innovation paradigm characterised by the use of data as the *core input*, advanced analytics as the *processing mechanism*, and novel value creation as the *main output*. Visvizi et al. (2022) further explored DDI in the context of innovative start-ups, emphasising the role of a "data-driven orientation" as a strategic capability that shapes organisational transformation.

A common source of confusion in discussions of DDI is its relationship to Artificial Intelligence (AI). While the two concepts are closely intertwined in practice, they are conceptually distinct. DDI refers to a broader innovation paradigm where data is systematically exploited to *generate* innovative ideas, solutions, and decisions. AI refers to a set of technologies and algorithms – such as machine learning, deep learning, and neural networks – that *enable* machines to perform tasks requiring human-like intelligence.

AI often serves as a *tool* within the DDI toolkit. Advanced AI algorithms can analyse big data at scale, detecting patterns or making predictions that humans alone could not easily discern. However, one can have data-driven innovation without deploying cutting-edge AI (e.g., a firm using simple statistical analysis on customer datasets to inform a new marketing strategy), and conversely, AI can be applied in contexts that do not fundamentally aim at innovation (e.g., automating routine processes). The essence of DDI lies in its approach to innovation – the strategic use of data as a core input – not necessarily in the novelty of the algorithms employed. The overlap is that modern DDI initiatives increasingly take advantage of AI techniques to handle big data. Many contemporary examples of DDI, such as recommendation engines, predictive maintenance systems, personalised medicine, are powered by AI analytics. Nevertheless, DDI also involves managerial and strategic dimensions beyond the algorithms themselves, including data governance, organisational capabilities, and business model design.

The empirical distinction between DDI and AI is not merely conceptual. As we demonstrate in Section 6, among 447,827 data-driven patents identified at the USPTO, 59% do not qualify as AI patents under the WIPO classification, while only 14% of non-DDI patents are classified as AI. This asymmetry confirms that DDI encompasses a substantially broader set of data-driven approaches —from sensor-based process optimisation to analytics-informed business model design— that do not necessarily involve artificial intelligence techniques. Conflating the two concepts risks both overestimating DDI (by equating it with AI adoption) and underestimating it (by ignoring non-AI data-driven innovation).

The rise of DDI has far-reaching economic implications across virtually every industry. By exploiting large datasets —whether generated from user activities, sensors in manufacturing, transactions in finance, or R&D processes— organisations can unlock new sources of value and efficiency (OECD, 2015). DDI enables firms to derive insights that inform better decision-making, customise offerings, optimise operations, and predict future trends.

What is particularly striking is that even traditionally "low-tech" sectors are being transformed by data. Precision agriculture uses sensor data to improve yields; hospitals analyse patient data to improve treatments; online retailers leverage customer data to develop personalised products and services.

Evidence consistently links the adoption of data-driven innovation (DDI) to higher productivity and competitive performance. Firms that successfully harness big data analytics tend to outperform their peers: companies with a strong data-driven decision-making culture exhibit output and productivity levels approximately 5% higher than expected (Brynjolfsson et al., 2011) while OECD (2015) evidence suggests that firms adopting DDI experience productivity growth approximately 5–10% faster than non-adopters.

These gains arise because data-rich firms can identify inefficiencies, learn faster from feedback, and tailor their offerings more precisely to market demand. A growing body of empirical evidence links DDI adoption to firm performance. Ghasemaghaei and Calic (2020) assessed the impact of big data on firm innovation performance, finding that data quality and analytical capability matter more than data volume alone. Mikalef et al. (2019) demonstrated that big data analytics capabilities enhance innovation through the mediating role of dynamic capabilities, while Fosso Wamba et al. (2015) documented through longitudinal analysis that firms systematically leveraging big data achieve measurable performance gains.

3 A Theory of DDI: Data Along the Innovation Chain

To develop a systematic understanding of how DDI operates, we anchor our theoretical framework in the chain-linked model of innovation proposed by Kline and Rosenberg (1986). This model represents the innovation process not as a linear sequence from basic research to market, but as a complex chain of interconnected stages with multiple feedback loops and links to the existing stock of knowledge and new research.

The chain-linked model identifies five central stages in the innovation process: (1) **Market Finding** – the identification of a potential market or need; (2) **Synthetic Design** – the inventive or analytic design of a new process or product; (3) **Detailed Design and Testing** – the detailed design, prototyping, and testing; (4) **Production** – the redesign and production phase; and (5) **Distribution and Marketing** – the distribution to and interaction with the market. Crucially, the model also identifies several types of *links*: short feedback loops (from production back to design), long feedback loops (from users back to market identification), links to existing knowledge (science and technology), and links to new research.

In parallel with the notion that AI might represent the invention of a new method of invention (Cockburn et al., 2018), following the original invention of the method of invention represented by the institutionalization of R&D (Whitehead, 1925), we propose that data-driven innovation

Table 1: Data-Driven Augmentation of Innovation Chain Elements

Chain Element	DDI Function	Typical Data Sources
1. Market Finding	Supervised & Unsupervised Learning: Inferring future needs via pattern recognition in user behaviour and market signals.	Social Media, Search Logs, Transaction Data
2. Synthetic Design	Generative Exploration: Broadening the search space of possible solutions via algorithmic generation and evaluation.	Patent Databases, Open-Source Repositories
3. Detailed Design & Testing	Simulation & Digital Twins: Virtual validation replacing or augmenting physical prototyping.	Physics Engines, Historical Test Data
4. Production	Real-Time Optimisation: Continuous error correction and process optimisation through sensor data feedback.	IoT Sensors, ERP Logs
5. Distribution & Marketing	Personalisation: Micro-targeting user segments and customising offerings based on behavioural data.	CRM Systems, Clickstream Data

represents "the invention of new method of innovating", unfolding through a systematic augmentation of each element of the chain-linked model. In fact, the availability of large-scale data and advanced analytical capabilities transforms how the innovation process is carried out at each stage. Table 1 summarises how DDI functions map onto the elements of the innovation chain.

At the market-finding stage, supervised and unsupervised learning techniques allow firms to identify emerging demand and infer unmet customer needs from behavioural, transactional, and market data. In this case, data complement conventional market research and enable firms to detect patterns that may not be visible through direct observation or surveys.

During synthetic design, data-driven and generative methods broaden the scope of possible technological or product solutions. Firms may use information drawn from patent databases, scientific repositories, and open-source materials to identify relevant combinations of knowledge and to generate and evaluate alternative designs.

At the detailed design and testing stage, simulations, digital twins creation, and analysis of historical test data allow firms to assess the performance of alternative designs before committing resources to physical prototypes. These techniques can reduce experimentation costs, accelerate testing, and allow a larger number of design alternatives to be evaluated.

In production, real-time information generated by sensors and enterprise systems supports continuous monitoring, anomaly detection, and process optimization. In this case, data are used both to control existing operations and to generate improvements in production methods, quality, reliability, and resource efficiency.

Finally, at the distribution and marketing stage, customer and usage data allow firms to segment demand more precisely, personalise offerings, and adjust products and services to observed user behaviour. Data-driven analysis thus strengthens the connection between market interaction and the commercial development of innovations.

Beyond the central chain elements, DDI also transforms the *feedback mechanisms and knowl-*

Table 2: Data-Driven Augmentation of Innovation Links

Link Type	DDI Function	Typical Data Sources
Short Feedback (Production → Design)	Diagnosis: Instant anomaly detection enabling rapid correction cycles.	Quality Cameras, Process Sensors
Long Feedback (User → Market)	Usage Analytics: Observing product behaviour "in the wild" through connected devices.	Connected Products (IoT)
Link to Knowledge (Science/Technology)	Semantic Search: Mining unstructured scientific and technical text for relevant knowledge.	Digital Libraries, Wikis
Link to Research (New Science)	Automated Discovery: High-throughput experimentation and data-driven hypothesis generation.	Lab Automation, AI Models

edge links that characterise the innovation process, as we summarize in Table 2.

The first mechanism concerns short feedback loops from production to design. Data generated by production equipment, quality-control systems, and process sensors permit rapid diagnosis of errors and performance deviations. Instead of waiting for periodic inspections or ex post assessments, firms can identify problems as they emerge and use this information to modify product or process design in real time. DDI therefore increases the speed and frequency of corrective learning cycles.

The second mechanism involves long feedback loops from users and markets. Connected products and digital services generate information about how products are actually used after commercialisation. Usage analytics allow firms to observe performance "in the wild," identify unexpected applications or shortcomings, and incorporate this information into subsequent market analysis and product development. This makes post-market interaction a continuous source of innovation rather than merely an endpoint of the innovation process.

The third mechanism concerns the link to the existing stock of scientific and technological knowledge. Semantic-search tools and text-mining techniques help firms identify relevant knowledge contained in patents, scientific publications, technical documents, and digital repositories. By improving the retrieval and recombination of dispersed knowledge, data-driven tools can reduce search costs and expand the range of information available to innovators.

The fourth mechanism is the link to new research. Automated experimentation, high-throughput screening, laboratory automation, and data-driven hypothesis generation can contribute directly to the production of new scientific and technical knowledge. In this case, data and algorithms help firms to retrieve existing knowledge and become instruments for generating new discoveries and testing new conjectures.

Overall, therefore, DDI affects not only what firms do at each stage of innovation, but also how information circulates among stages and between innovation activities, existing knowledge, and new research. Data-driven technologies can make feedback more immediate, systematic, and scalable, thereby increasing the iterative character of innovation described by the chain-linked model.

The DDI augmentation of the innovation chain rests on two interconnected enabling elements:

- **Data Infrastructure (the "Big Data Stack")**. The technological infrastructure required to collect, store, and process large volumes of data at speed. This includes: cloud computing platforms, distributed processing frameworks (e.g., MapReduce, Spark), and scalable storage architectures (e.g., data lakes, NoSQL databases). Without this infrastructure, the volume, velocity, and variety of data that characterise modern DDI would be unmanageable.

- **Algorithms as Instruments**. The analytical methods that transform raw data into action-able insights. These include: (i) *Prediction* – forecasting outcomes from patterns using machine learning and AI; (ii) *Pattern Recognition* – extracting structure from unstructured data (text, images, signals); and (iii) *Optimisation and Exploration* – searching vast solution spaces using genetic algorithms, reinforcement learning, adversarial networks, and related techniques.

The critical insight is that both the technological infrastructure and the analytical capabilities underpinning DDI leave observable traces. The adoption of cloud computing, machine learning, IoT sensor networks, or other data-intensive technologies is reflected in corporate communications, patent filings, websites, and related digital documents. These traces provide observable evidence of firms' underlying data capabilities and therefore constitute the empirical basis for our measurement strategy.

As data can augment every stage of the innovation chain, DDI should be understood as a general organisational capability rather than a phenomenon confined to specific industries or firm types. Its underlying mechanisms are potentially applicable across heterogeneous organisational settings (Bresnahan and Trajtenberg, 1995). Its implementation varies according to firms' technological opportunities, organisational capabilities, and strategic objectives (Vial, 2019). In particular:

- **Across sectors**: DDI is adopted in both high- and low-technology industries, although through different technological configurations and innovation processes.
- **Across firm types**: Both digital native firms and incumbent organisations undergoing digital transformation leverage data to support innovation, albeit with different strategic objectives and organisational capabilities.
- **Across firm sizes**: Large firms typically invest in integrated data infrastructures, whereas SMEs often rely on more targeted or cloud-based data analytics solutions.

This heterogeneity implies that DDI cannot be adequately captured through a single technology or industry-specific proxy. Instead, its measurement requires an approach capable of identifying multiple observable manifestations of data-driven innovation across diverse organisational and industrial contexts.

4 Why Large Language Models? A Strategy to Measuring DDI

4.1 The Measurement Challenge

Despite widespread recognition of DDI's economic importance, significant challenges persist in measuring it at the firm level. Traditional innovation metrics, such as R&D expenses, patent counts, product launch counts, are not designed to capture the nuanced ways in which data-driven approaches permeate organisations. Several properties of DDI make it particularly elusive to standard measurement:

1. **Intangibility**: DDI often manifests as an improved algorithm, a new analytical capability, or a data-driven decision process. These may not be patented, marketed, or formally announced, yet they significantly enhance a firm's innovation capacity.

2. **Embeddedness**: Data-driven capabilities are typically embedded within existing products, services, and processes rather than constituting stand-alone innovations. A machine learning model that improves a supply chain is invisible in standard measures and statistics.

3. **Complexity**: DDI spans across business functions – from marketing analytics to IoT-based operations to AI-driven R&D. No single metric can capture this cross-functional impact.

4. **Heterogeneity**: DDI encompasses algorithmic implementations, business model adjustments, organisational changes, and more. The same "DDI" label covers vastly different realities across firms and sectors.

Existing approaches to measuring data-related innovation through text have relied predominantly on keyword dictionaries and patent classification codes. Cockburn et al. (2018) identify AI-related inventions using combinations of CPC codes and keyword lists, while the WIPO (2019) methodology refines this approach with curated keyword sets applied to patent abstracts. Webb (2020) maps technology exposure by matching patent text to occupational descriptions using tf-idf similarity. In the broader text-as-data tradition (Gentzkow et al., 2019) dictionary-based methods, that count occurrences of pre-specified terms, have been the workhorse for measuring latent constructs in economic texts, including recent applications to firm-level technology adoption through web-scraped data (Ashouri et al., 2024) and patent similarity (Hain et al., 2022).

These approaches are valuable but face inherent limitations when applied to DDI. First, keyword dictionaries are rigid: they miss paraphrases, novel terminology, and context-dependent usage. A term like "platform" may signal DDI in one context ("our data platform enables predictive analytics") but not another ("we launched a new retail platform"). Second, dictionary methods cannot assess *semantic dependency* between means and purpose: they detect the presence of individual terms but not whether those terms are functionally linked within the text. Third, dictionaries require continuous manual updating as the technological vocabulary evolves, creating a lag between practice and measurement.

These limitations motivate our adoption of Large Language Models, which can assess context, resolve ambiguity, and detect the co-presence of means and purpose without relying on fixed term lists.

The further specific feature of our measurement strategy is to apply LLMs to the analysis of different types of text documents. Indeed, intangibility and heterogeneity of DDI imply that no single datasource can provide a comprehensive picture. Different firm types leave DDI traces in different places:

- **Large listed firms** produce extensive public communications (e.g., earnings calls, annual reports) where they discuss strategic priorities including data and analytics initiatives. These documents are a rich, yet largely unexploited, source for inferring DDI involvement. How-ever, they cover only a fraction of the economy.

- **Patenting firms** protect their technical innovations through intellectual property filings. Patent text provides detailed descriptions of data-driven technologies and their applications. Yet patents are filed predominantly by larger firms and are concentrated in certain sectors; many SMEs and service-sector firms do not patent at all.

- **SMEs** may not patent and may not be subject to public disclosure requirements, but they increasingly maintain corporate websites that may describe their DDI capabilities, products, and technological orientation. Recent research has demonstrated the value of web-scraped data as a source of firm-level innovation indicators (Kinne and Lenz, 2021).

- **Startups** describe themselves on platforms, such as Crunchbase, where they present their business model, technological base, and value proposition to potential investors. These descriptions reveal how new ventures position data and analytics within their innovation strategies.

By analysing multiple sources in parallel, we can capture DDI manifestations across different types of firms, contexts, and communication purposes. The multi-source approach reflects what firms *do* (as revealed by patents), what firms *say* to markets and investors (as revealed by communications and websites), and how firms *position themselves* in the entrepreneurial ecosystem (as revealed by Crunchbase profiles).

4.2 Semantic Identification of DDI Using Large Language Models

The core empirical challenge in measuring DDI through textual analysis is that neither the presence of data-related technologies nor the mention of innovation objectives is, by itself, sufficient to identify DDI.
Two types of ambiguity must be resolved:

1. **Ambiguity of Means.** A firm may mention data infrastructure, cloud services, or advance analytics without using them for innovation. For example, a sentence such as "We replaced our server with a Google Cloud solution" describes technology acquisition but not necessarily innovation. Thus, the mere detection of data-driven *means* (i.e., infrastructure, algorithms, big data stack) does not imply DDI.

2. **Ambiguity of Purpose.** A firm may describe innovation activities (e.g., improving, inventing, optimising, creating) without these being data-driven. For example, a sentence such as "We developed a new app for our customers" describes an innovation but may involve no data-driven component. The detection of innovative *purpose* alone does not imply DDI.

We formalise the identification rule as summarised in the following box:

Definition: Data-Driven Innovation in Text. A textual unit t is classified as containing evidence of DDI if and only if:

1. t contains reference to at least one data-driven *means* $m \in \mathcal{M}$, where $\mathcal{M}$ encompasses data infrastructure (cloud computing, data lakes, IoT sensors), analytical methods (machine learning, statistical modeling, optimisation algorithms), and data assets (real-time data streams, large-scale datasets, sensor data);

2. t contains reference to at least one innovative *purpose* $p \in \mathcal{P}$, where $\mathcal{P}$ encompasses the creation, improvement, or optimisation of products, processes, services, or business models; and

3. m and p are *semantically dependent* within t — that is, the text describes the means as enabling, supporting, or being applied toward the purpose.

Condition (3) is what distinguishes our approach from keyword counting: DDI requires not just co-occurrence but functional linkage. Consider the following contrasting examples:

- **Example 1 — DDI detected** (all three conditions met):

  "We **developed a new app**, based on the analysis of the **real-time data** summarised with a **machine learning** model deployed on **Microsoft Azure**."

  The innovative purpose (developing a new app) is explicitly enabled by data-driven means (real-time data analysis, machine learning, cloud deployment).

- **Example 2 — Means only, no DDI** (condition 2 fails):

  "We migrated our legacy systems to **Google Cloud** and now store all transaction records in a **data lake**."

  Data infrastructure is present, but no innovative purpose is described: this is technology adoption, not innovation.

- **Example 3 — Purpose only, no DDI** (condition 1 fails): "We **launched a new line** of organic skincare products targeting the European market."

  Innovation is present (new product launch), but no data-driven means are involved.

- **Example 4 — Co-occurrence without dependency, no DDI** (condition 3 fails):

  "Our company uses **cloud computing** for email hosting. Separately, we **developed a new** packaging design based on customer focus groups."

  Both means and purpose appear in the text, but they are not semantically linked: the cloud service does not enable the innovation.

Of course, edge cases are possible, as several boundary situations deserve attention, such as:

- *Process optimisation without novelty*: a firm using predictive analytics to reduce energy costs is engaging in DDI only if the optimisation represents a meaningful improvement, not merely routine monitoring.

- *Data as product*: firms that sell datasets or analytics services (e.g., credit scoring bureaus) may or may not qualify, depending on whether the data processing itself constitutes an innovative offering.

- *Aspirational language*: statements like "we plan to leverage AI to transform our operations" describe intent rather than realised DDI; despite classifiers are can be trained to distinguish aspiration from implementation, this remains a source of measurement noise.

The above definition and framing of DDI have a direct methodological implication: the analytical tool must be capable of understanding context and semantic relationships within text, not merely counting keywords. Simple keyword matching would generate excessive false positives (e.g., any mention of "cloud" or "data") and false negatives (e.g., paraphrased descriptions of DDI without standard key-words). This is where Large Language Models become essential.

The methodological foundation of our measurement approach rests on the significant advancements in Natural Language Processing (NLP) driven by transformer-based neural network architectures. The transformer architecture, introduced by Vaswani et al. (2017) employs a self-attention mechanism that enables models to weigh the importance of different parts of an input sequence when producing each element of the output. This allows the model to capture complex contextual relationships and long-range dependencies within text.

The evolution of transformer-based models has produced two broad families relevant to our work:

- **Encoder Models (e.g., BERT, RoBERTa).** These models are trained to build rich contextual representations of text. BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2018) pioneered the pre-training/fine-tuning paradigm. The model is first pre-trained on massive text corpora using masked language modeling objectives, then fine-tuned on domain-specific labeled data for downstream classification tasks. RoBERTa (Liu et al., 2019) refined this approach with longer pre-training, dynamic masking, and larger batch sizes, achieving state-of-the-art performance in text classification.

- **Decoder Models (e.g., GPT, Gemma).** These generative models are trained to produce text, and can be directed to perform specific analytical tasks through prompt engineering – carefully designed instructions that elicit structured responses. Large Language Models (LLMs) such as GPT-4 and Gemma exhibit remarkable capabilities in following instructions, performing complex reasoning, and extracting structured information from unstructured text (Brown et al., 2020).

In the specific context of DDI identification, transformer-based models offer critical advantages over traditional text analysis:

1. **Contextual understanding**: The attention mechanism considers full context when interpreting words and phrases, enabling accurate identification of DDI concepts expressed in varied ways.

2. **Transfer learning**: Pre-training on massive general-purpose corpora followed by fine-tuning on domain-specific data allows the model to leverage general language understanding while adapting to specialised text types.

3. **Scalability**: Parallel processing enables efficient analysis of large document collections across multiple data sources.

4. **Adaptability**: The same architecture can be adapted to different tasks through fine-tuning, prompt engineering, or few-shot learning, enabling methodological consistency across diverse data sources.

4.3 Methodological Implementation: Two Complementary Approaches

The choice between encoder and decoder approaches is driven by the nature of the text. Starting from the characteristics of our datasources, we employ two complementary strategies. Econder-based fine tuning is applied to corporate communications, since they are narrative and require learning of subtle patterns from labeled data. A decoder-based prompt-engineering approach is instead used to analyse patents, websites and Crunchbase firm descriptions, as these texts are more formulaic and can be effectively analysed through well-designed prompts. We provide some further detail here below. In both cases, the underlying goal is the same: detecting the semantic co-presence of data-driven means and innovative purpose.

Encoder-based Fine-Tuning

For corporate communications – representing narrative, abstract and dispersed texts, where DDI signals are diffuse – we implement a supervised classification approach using a fine-tuned RoBERTa model. The analysis operates at the *sentence level*, then aggregates to the document/firm/year level, producing a continuous DDI score.
The training data ($N = 2{,}900$) was constructed as follows:

- **Positive labels (DDI)**: 1,450 innovative case studies from major cloud providers (Google Cloud, AWS, Alibaba, Microsoft Azure, IBM), rewritten via ChatGPT-4 to mimic the grammatical structure and tone of earnings calls (*style transfer*).

- **Negative labels (Non-DDI)**: 1,450 random sentences from historic earnings calls over the period 2004–2007, selected from a period prior to the emergence of the DDI concept.

The use of style-transferred case studies as positive training examples requires justification. The fundamental challenge is that no large, pre-labeled corpus of DDI sentences from earnings calls exists. Cloud providers' case studies were selected because they describe verified instances of firms using data infrastructure and analytics for innovative purposes, precisely the means–purpose combination that defines DDI. The style transfer step is necessary because case studies and earnings calls differ markedly in register, sentence structure, and vocabulary (e.g., case studies use promotional language while earnings calls adopt formal financial disclosure style). By prompting ChatGPT-4 to rewrite each case study "as if it were a sentence from a quarterly earnings call" we preserve the semantic content (the DDI signal) while adapting the surface form to match the target domain. We verified the quality of the style transfer through manual inspection of 200 randomly selected rewritten sentences, confirming that: (a) the DDI content was preserved; (b) the financial register was convincingly adopted; and (c) no hallucinated content was introduced. The choice of the 2004–2007 period for negative examples ensures temporal separation from the DDI phenomenon, minimising label noise in the negative class.

The model's performance was validated through three complementary approaches: (i) manual inspection of 500 high-confidence classifications (top and bottom deciles of predicted probability; (ii) token association analysis confirming that the model attends to semantically relevant terms rather than spurious correlates; (iii) comparison with independent ChatGPT-4-generated classifications on a held-out sample. The fine-tuned model achieves 79% accuracy with 85% precision and 78% recall on the test set, indicating that the model is somewhat conservative (fewer false positives than false negatives), which is appropriate for an indicator intended to measure the *intensity* of DDI rather than merely its presence.

Decoder-based Prompt Engineering

For patents, websites, and Crunchbase profiles – texts that are more descriptive, homogeneous, and technical – we implement a prompt engineering approach using the Gemma model (mannix/gemma2-9b-simpo). The model is instructed to classify each text as DDI-related or not, providing structured output with concise reasoning.

The prompt design embeds the means–purpose framework directly into the classification criteria, as presented in the following box:

```
Analyze the following text. Respond EXCLUSIVELY in the format:
YES|concise reasoning
or
NO|concise reasoning

Criteria for Data-Driven Innovation:
- Use of data as main resource for innovation
- Analytics/AI/ML applied in practical contexts
- Data-driven automation/optimization
- IoT for data collection
- Value creation through data processing

DO NOT add explanations or additional text.
Text to analyze:
{text}
```

Model parameters were set to maximise deterministic responses: temperature at 0, top k at 2, and top p at 0.10. The analysis was executed locally on high-performance hardware (NVIDIA Tesla V100-PCIE-16GB) using Ollama, ensuring data security and eliminating API costs.

For patent analysis specifically, an extended prompt was employed to extract in a single pass the following multiple dimensions: advanced analytics technologies mentioned, product vs. process orientation of the underlying invention, sustainability characteristics.

Validation and Measurement Uncertainty

Any text-based measurement strategy faces the risk of false positives (non-DDI content classified as DDI) and false negatives (genuine DDI missed by the classifier). Several sources of measurement noise deserve explicit discussion:

- *Marketing inflation.* Corporate communications and websites may use data-related buzzwords (e.g., "AI-powered," "data-driven") as marketing signals without corresponding substantive innovation. Our means–purpose framework mitigates this risk by requiring both a technology component *and* an innovation component linked within the same textual unit. However, aspirational or inflated language can still produce false positives.
- *Routine analytics vs. innovation.* Firms routinely use data for operational purposes (e.g., monitoring dashboards, standard reporting) that do not constitute innovation. The boundary between routine data use and DDI might be inherently fuzzy. Our classifiers are trained on examples that emphasise novelty and improvement, but some borderline cases will inevitably be misclassified.
- *Omission bias.* Firms may engage in DDI without documenting it in the textual sources we analyse. This is particularly likely for small firms, which may lack the communication infrastructure to publicize their activities, and for process innovations, as they are less visible than product innovations. Our multi-source strategy partially addresses this by casting a wide net, but systematic under-reporting remains difficult to fully tackle.
- *Source-specific biases.* Each datasource has its own biases. For instance, corporate communications are shaped by disclosure regulation and investor expectations; patents are filtered by patentability criteria and strategic considerations; websites reflect marketing choices; and start-up Crunchbase profiles are curated for investor audiences. These biases do not invalidate the measures but must be kept in mind when making interpretations and cross-source comparisons.

We narrow down these concerns through three safeguards: (i) the means–purpose framework itself, which imposes a structural filter beyond simple keyword detection; (ii) validation against human judgment and independent LLM classifications; (iii) the use of multiple sources, which allows triangulation and reduces reliance on any single, potentially biased signal.

5 Data Sources

Table 3 provides an overview of the four data sources employed in this study.

Table 3: Data Sources Overview

Source	Geographic Coverage	Time Coverage	Observations
Communications	USA	2008–2022	28,756 docs / 3,714 firms
	Italy	2008–2022	2,985 docs / 102 firms
Patents	Global (USPTO)	2010–2021	3,398,767 patent docs
Websites	Italy (SMEs)	2018–2023	134,208 SMEs
Crunchbase	Global (Startups)	2011–2021	9,705 firm descriptions

As a first source, we draw from publicly available **corporate communications**. By their nature, these documents capture how companies communicate and signal their data-related innovation activities to the public. We compiled two datasets, including companies listed in the Italian and US stock markets:

- **Italy**: Compulsory communications filed by companies listed on the Milan stock exchange, retrieved from the *FactSet* platform, for a total of 2,985 documents from 102 companies, over the period 2008–2022.

- **US**: 10-K reports filed by companies listed on the New York stock exchange, retrieved from the *SEC-EDGAR* system, for a total of 28,756 documents from 3,714 companies, over the period 2008–2022.[1] The DDI identification produces a firm-specific, time-varying index ("DDI-Share"), reflecting the proportion of sentences within a firm's communications in a given year classified as DDI-related by the RoBERTa model. We further define "DDI-Intense" firms as those with DDI-Share above the annual cross-sectional average, and "DDI-Persistent" (DDI-PERS) firms as those classified as DDI-Intense in every year of observation. By examining verbs associated with DDI mentions, we also distinguish between acquisition-oriented DDI (integrate, deploy, introduce, implement, acquire, adopt) and development-focused DDI (create, improve, design, promote, generate).

The encoder-based identification of DDI described above, produces the firm-and-year specific index "DDI-Share", reflecting the proportion of sentences within the corpus of a firm's communications in a given year, which are classified as DDI-related by the RoBERTa model. We exploit this in several ways to identify various groups of firms involved in DDI to a different extent (see below).

Second, we resort to patent documents, which provide detailed insights into the technical implementation of DDI. We selected the "Brief Summary Text" (encompassing Field of Invention, Background, and Summary) of all pre-grant publications filed at the USPTO over the period January 2011 to December 2021, amounting to 3,398,767 distinct patent documents retrieved from the *USPTO-PatentsView* platform. The prompt-based analysis described above, classifies each patent as DDI-related or not, and additionally identifies whether the technology relates to a physical product, a service, or a production process. Patents are matched to firms using the *ORBIS-IP* concordance, enabling analysis by firm characteristics.

Third, we analyse corporate websites. They represent important external communication channel where firms articulate technological capabilities and strategic orientations. Our sample comprises 134,208 Italian SMEs (limited liability companies) active in 2018, identified from *BvD-AIDA*, for which a certified website could be found. Website text was retrieved from the Wayback Machine (Bottai et al., 2024) and processed through the LLM-based prompt as described above. The output is a binary classification of whether each website, and hence each firm, is DDI or Non-DDI.

Lastly, the *Crunchbase* dataset allows to access structured information about startups and emerging companies. Our sample comprises the description (or company profile) of 9,705 firms founded over the period 2011–2021. Company descriptions were analysed using the prompt engineering approach described above, to classify each firm as DDI or Non-DDI.

6 Descriptive Evidence: DDI Patterns Across Sources

6.1 Corporate Communications

DDI Diffusion and Intensity

Table 4 reports descriptive statistics of DDI-Share for both the Italian and US samples. In Italy, the mean DDI-Share shows a gradual increase from 0.001 in 2008–2015 to 0.002–0.003 in 2016–2022, with considerable heterogeneity. While the mean indicates that only 0.1–0.3% of sentences deal with DDI on average, the maximum reaches about 3.5%. These values are considerably lower

[1] 10-K reports include audited financial statements, management's discussion and analysis, and disclosures about the company's operations, risk factors, and market conditions.

Table 4: Corporate communications: share of DDI-related text

	Variable: DDI-Share							
	Italy sample				**US sample**			
Year	**Mean**	**St.Dev.**	**Min**	**Max**	**Mean**	**St.Dev.**	**Min**	**Max**
2008	0.001	0.002	0	0.014	0.046	0.039	0	0.368
2009	0.001	0.002	0	0.010	0.046	0.040	0	0.430
2010	0.001	0.002	0	0.014	0.046	0.041	0	0.497
2011	0.001	0.002	0	0.020	0.047	0.041	0	0.479
2012	0.001	0.002	0	0.015	0.048	0.042	0	0.466
2013	0.001	0.002	0	0.010	0.049	0.043	0	0.483
2014	0.001	0.002	0	0.014	0.050	0.043	0	0.522
2015	0.001	0.002	0	0.010	0.052	0.043	0	0.486
2016	0.001	0.004	0	0.033	0.051	0.043	0	0.488
2017	0.002	0.004	0	0.025	0.051	0.041	0	0.521
2018	0.002	0.004	0	0.033	0.051	0.041	0	0.514
2019	0.002	0.004	0	0.026	0.051	0.041	0	0.493
2020	0.003	0.005	0	0.028	0.051	0.040	0	0.359
2021	0.002	0.005	0	0.035	0.055	0.039	0	0.275
2022	0.002	0.005	0	0.027	0.059	0.040	0	0.305

Notes: Descriptive statistics of DDI-Share (share of DDI-related sentences) aggregated at the firm level, by year. Italy: quarterly earnings calls. US: 10-K reports.

Table 5: Corporate communications: firm-level DDI classification

	Italy sample			**US sample**		
Year	**Total Firms**	**No-DDI Firms**	**DDI-Intense Firms**	**Total Firms**	**No-DDI Firms**	**DDI-Intense Firms**
2008	91	78	13	2,813	40	1,025
2009	117	97	20	3,306	77	1,224
2010	139	113	26	3,279	97	1,207
2011	153	129	24	3,266	99	1,226
2012	156	130	26	3,252	80	1,218
2013	153	121	32	3,233	100	1,229
2014	171	136	35	3,253	95	1,224
2015	194	145	49	3,261	98	1,223
2016	202	158	44	3,206	98	1,215
2017	216	156	58	3,181	81	1,235
2018	244	164	73	3,167	79	1,215
2019	245	182	63	3,131	89	1,197
2020	248	158	78	3,053	65	1,174
2021	246	154	76	2,953	58	1,161
2022	245	160	70	2,836	42	1,152

Notes: For each year: (a) total number of firms; (b) No-DDI firms (DDI-Share = 0); and (c) DDI-Intense firms (DDI-Share above the annual cross-sectional average).

than the US sample, where mean DDI-Share ranges from 4.6% to 5.9%, reflecting differences in both the nature of documents (quarterly earnings calls vs. comprehensive 10-K reports) and the diffusion of DDI practices.

Firm-level DDI Classification

Starting from the yearly, firm-specific values of DDI-Share, we identify two groups of firms involved in DDI to a different degree. We define (i) a time-varying indicator for "DDI-Intense" firms, indicating firms whose DDI-Share value in a given year is larger than the annual cross-sectional average of DDI-Share; and (ii) a time-invariant indicator for "DDI-Persistent" (DDI-PERS) firms, identifying firms which are classified as DDI-Intense over the entire period of observation.

Table 5 reports the total number of firms in the Italian and US samples, then broken down by No-DDI firms and DDI-Intense firms, per year. In Italy, the proportion of No-DDI firms ranges from approximately 65% to 85%, meaning that the majority of Italian listed companies do not mention DDI-related content in their communications. This contrasts sharply with the US, where only 1–3% of firms show zero DDI content. The number of DDI-Intense firms grows over time in both samples: from 13 to about 70 in Italy, and consistently around one-third of the US sample each year. In the US sample, very few firm-year observations have DDI-Share equal to zero, indicating that DDI is widespread among listed companies. This finding suggests that for listed companies, the relevant analytical question is not whether firms engage in DDI, but how intensely they do so. For the Italian sample and the other data sources covering more heterogeneous firm populations, the dichotomous DDI/No-DDI distinction remains informative.

For DDI-Persistent firms, the US sample identifies 667 firms that remain DDI-Intense across all years, while the Italian sample yields only 31 firm-year observations qualifying as DDI-PERS, reflecting the smaller and more heterogeneous sample.

DDI by Firm Size and Age

Tables 6 and 7 compare size and age characteristics of DDI-Intense and DDI-PERS firms with the rest of the sample, for Italy and the US respectively.

In the **Italian sample**, DDI-Intense firms are slightly smaller on average (10,156 employees vs. 12,913 for others), although the vast majority (88%) are still large firms (250 employees). DDI-Intense firms are younger on average (58 vs. 68 years) but concentrated among mature firms (10 years: 90%). The DDI-PERS sub-group is even smaller and younger, with 43% being young firms (≤10 years), though sample sizes are small.

In the **US sample**, DDI-Intense firms are clearly larger on average (16,955 vs. 9,647 employees), with about 72% being large firms. They are younger on average (30 vs. 35 years) but concentrated among mature firms (91%). DDI-PERS firms amplify these patterns: even larger (19,944 mean employees) and younger (26 years on average).

Table 6: Italian communications: DDI by firm size and age

	Firm Size			**Firm Age**		
	Mean Empl.	SMEs	Large	Mean Age	Young	Mature
DDI-Intense:	10,156	22	160	58.21	11	96
		(12.09%)	(87.91%)		(10.28%)	(89.72%)
Others:	12,913	80	300	67.76	32	174
		(21.05%)	(78.95%)		(15.53%)	(84.47%)
DDI-PERS:	9,080	1	10	43.29	3	4
		(9.09%)	(90.91%)		(42.86%)	(57.14%)
Others:	12,079	101	450	64.98	40	266
		(18.33%)	(81.67%)		(13.07%)	(86.93%)

Notes: Statistics on firm-year observations. Firm size: SMEs (≤250 employees) vs. Large (>250). Age: Young (≤10 years) vs. Mature (>10 years).

DDI by Sector

Tables 8 and 9 report the sectoral distribution of DDI for Italy and the US respectively. In both countries, Information and Communication (sector J) leads in DDI intensity: 55.4% DDI-Intense in Italy and 80.4% in the US. Professional, scientific and technical activities (sector M) and Manufacturing (sector C) also show high DDI intensity. The US sample reveals a broader sectoral diffusion of DDI, with Education (P: 73.8%), Administrative services (N: 49.2%), and Wholesale/Retail (G: 47.2%) also featuring prominently.

Table 7: US communications: DDI by firm size and age

	Firm Size			Firm Age		
	Mean Empl.	SMEs	Large	Mean Age	Young	Mature
DDI-Intense:	16,955	4,018	10,444	29.74	1,625	15,597
		(27.78%)	(72.22%)		(9.44%)	(90.56%)
Others:	9,647	5,382	13,694	35.33	2,798	24,199
		(28.21%)	(71.79%)		(10.36%)	(89.64%)
DDI-PERS:	19,944	1,625	5,361	25.98	722	7,103
		(23.26%)	(76.64%)		(9.23%)	(90.77%)
Others:	10,918	7,775	18,777	34.70	3,701	32,623
		(29.28%)	(70.72%)		(10.17%)	(89.83%)

Notes: Statistics on firm-year observations. Firm size: SMEs (≤250 employees) vs. Large (>250). Age: Young (≤10 years) vs. Mature (>10 years).

Table 8: Italian communications: DDI by sector

NACE	**Total Obs**	**% DDI Intense**	**% DDI PERS**
C - Manufacturing	131	41.2	3.8
D - Electricity, gas, steam and air conditioning supply	44	25.0	0.0
F - Construction	30	40.0	0.0
G - Wholesale and retail; repair motor-vehicles/cycles	129	27.9	0.0
H - Transportation and storage	7	0.0	0.0
J - Information and communication	56	55.4	26.8
K - Financial and insurance activities	219	24.7	1.8
L - Real estate activities	32	31.2	0.0
M - Professional, scientific and technical activities	63	44.4	0.0
N - Administrative and support service activities	3	0.0	0.0
O - Public administration and defence	2	100.0	100.0

Notes: Firm-year observations by NACE 1-digit sectors. Sectors with no observations omitted.

Table 9: US communications: DDI by sector

NACE	**Total Obs**	**% DDI Intense**	**% DDI PERS**
A - Agriculture, forestry and fishing	207	19.8	3.4
B - Mining and quarrying	1917	7.1	1.5
C - Manufacturing	16758	48.9	19.7
D - Electricity, gas, steam and air conditioning supply	2289	6.1	1.2
E - Water; sewerage, waste manag. and remediation act.	391	17.1	4.6
F - Construction	541	17.0	4.3
G - Wholesale and retail; repair motor-vehicles/cycles	2882	47.2	16.8
H - Transportation and storage	973	26.4	7.0
I - Accommodation and food service activities	764	33.8	5.8
J - Information and communication	3749	80.4	59.7
K - Financial and insurance activities	7594	17.6	4.6
L - Real estate activities	579	7.8	3.1
M - Professional, scientific and technical activities	2293	61.8	33.3
N - Administrative and support service activities	1240	49.2	26.4
P - Education	172	73.8	50.6
Q - Human health and social work activities	623	38.2	14.6
R - Arts, entertainment and recreation	249	29.7	6.0
S - Other service activities	1790	2.0	0.0

Notes: Firm-year observations by NACE 1-digit sector.

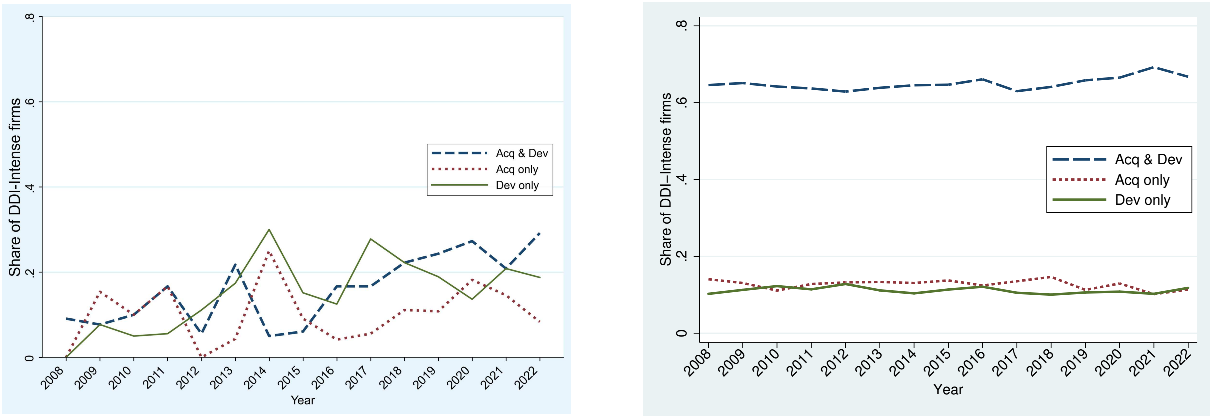

Figure 1: Share of DDI-Intense firms per year by DDI strategy: Development Only, Acquisition Only, and Development & Acquisition. Italy (left) vs. US (right).

DDI Acquisition vs. Development

By examining verbs associated with the sentences which are classified as DDI mentions by the encoder, we distinguish between a terminology oriented to DDI acquisition (i.e., integrate, deploy, introduce, implement, acquire, adopt) and a terminology hinting at DDI development (i.e., create, improve, design, promote, generate). This is used to create three further categories of firms: DDI-Acquisition Only, DDI-Development Only, and DDI-Acquisition & Development, respectively indicating whether DDI-related sentences in firms' communications are indeed only referring to either acquisition or development terminology, or both.

As shown in Figure 1, in the US about 60% of DDI-Intense firms combine internal development and external acquisition. Acquisition-Only and Development-Only firms each account for about 18% instead. These patterns are stable over time. In contrast, the Italian sample exhibits considerably more volatility, reflecting smaller sample sizes and less consolidated DDI strategies. Development-Only firms show peaks around 30% in certain years, while recent years (2020–2022) suggest a possible convergence toward the mixed strategy that characterises the US market.

6.2 Patent Documents

DDI Incidence Over Time

Out of 3,398,767 USPTO patent documents analysed, 447,827 (13.2%) are classified as DDI-patents. Table 10 and Figure 2 show that the incidence of DDI almost doubles over the sample period, from 9.0% in 2010 to 18.1% in 2020.

DDI by Technology Field

Figure 3 reports the distribution of DDI-patents across WIPO technology fields. The large majority (about 40%) relates to computer technologies, followed by digital communication, medical, and measurement technologies (about 10% each). In Table 11 we further disaggregate by CPC subclasses. DDI-patents are heavily concentrated in G06 (about 40%), particularly G06F (Electronic digital data processing, 18.3%). Six CPC codes appear in the top-10 of both DDI and Non-DDI patents, but DDI patents are disproportionately concentrated in computing-related classes.

Table 10: USPTO patents: DDI incidence over time

Year	Total Patents	DDI Patents	% Share
2010	201,464	18,097	9.0
2011	226,246	21,063	9.3
2012	261,194	26,481	10.1
2013	316,538	34,429	10.9
2014	328,616	37,138	11.3
2015	332,838	40,354	12.1
2016	336,743	44,214	13.1
2017	347,999	48,954	14.1
2018	349,338	53,542	15.3
2019	343,981	59,636	17.3
2020	244,091	44,175	18.1
2021	109,719	19,744	18.0
Overall	3,398,767	447,827	13.2

Notes: Total patents from USPTO-PatentsView and associated % share of DDI-patents, by year. Last two years affected by truncation.

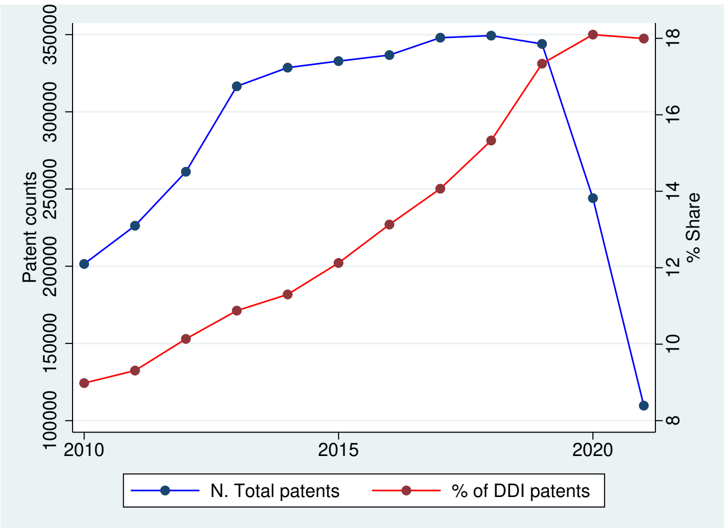

Figure 2: USPTO patents: total patents per year (left axis) and % share of DDI-patents (right axis).

Table 11: USPTO patents: DDI vs. Non-DDI by top CPC subclasses

DDI patents		Non-DDI patents	
Subclass code and label	**%**	**Subclass code and label**	**%**
G06F - Electronic digital data processing	18.3	G06F - Electronic digital data processing	8.5
G06Q - ICT technologies	7.0	H01L - Semiconduc. devices not under H10	7.1
H04L - Transmission of digital info	5.6	H04L - Transmission of digital info	4.2
G06K - Graphical data (read and present)	5.6	H04W - Wireless communication networks	3.8
A61B - Diagnosis, surgery	5.4	A61K - Prepar. for medical purposes	3.8
G06T - Image data (process or generation)	5.3	H04N - Pictures communication	2.8
H04N - Pictures communication	4.6	A61B - Diagnosis, surgery	2.7
G06N - Computing, specific comp. models	3.7	G06Q - ICT technologies	1.7
G01N - Investig. materials' chem/phys prop.	3.4	G02B - Optical elements, systems and apparatus	1.5
G10L - Speech analysis (recogn., process)	2.1	G01N - Investig. materials' chem/phys prop.	1.5

Notes: Top-10 CPC subclasses, frequencies within DDI and Non-DDI patents.

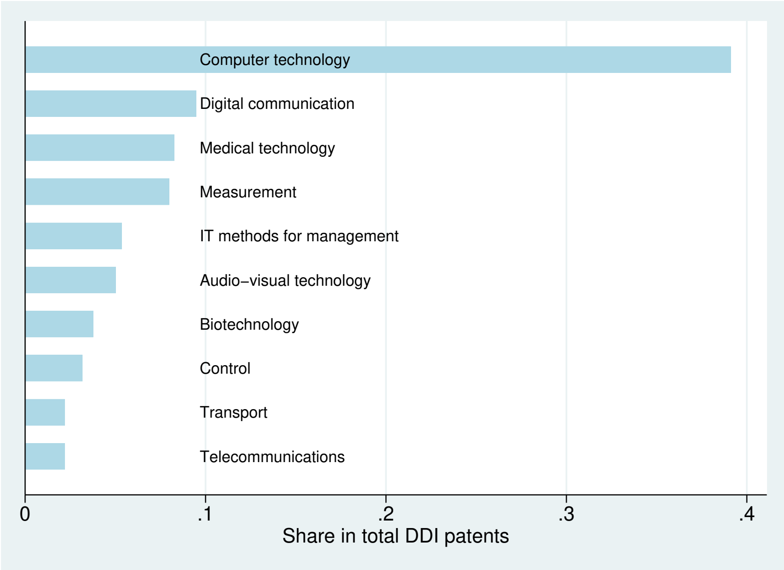

Figure 3: DDI-patents across WIPO technology fields: shares over total DDI-patents.

DDI vs. AI Patents

We apply the WIPO procedure (CPC codes + keywords) to identify AI patents in our sample and compare them with DDI-patents (Table 12). Of the 447,827 DDI-patents, 41% are also classified as AI, while 59% are DDI but not AI. Conversely, among the 2,950,940 Non-DDI patents, only 14% are classified as AI. This confirms that DDI is a broader concept than AI. Most DDI-patents involve data-driven innovation without necessarily qualifying as AI, while AI technologies are disproportionately associated with DDI.

Table 12: USPTO patents: DDI vs. AI cross-tabulation

	DDI	**No-DDI**	Total
AI:	183,950 (41%)	400,451 (14%)	584,401
No-AI:	263,877 (59%)	2,550,489 (86%)	2,814,366
Total:	447,827	2,950,940	3,398,767

DDI by Product vs. Process Innovation

As mentioned, extending the basic prompt, each patent can be classified as product-oriented, process-oriented, or both. Table 13 and Figure 4 show the distribution of DDI-patents by type. DDI-process patents represent 32% of total DDI-patents, DDI-product patents account for an 8.9%, while 49.1% are DDI-mix patents (i.e., are classified as both product and process). In the initial years, the share of DDI-product and DDI-process patents increased while DDI-mix patents declined. Relative shares stabilised from around 2013 onward.

Geography of DDI Patents

To explore the geographical distribution of DDI-patents across countries, we applied a standard fractional counting assignment, based on the address of the inventors involved in each patent, thus capturing the location where underlying knowledge is developed. Figure 5 shows the top-

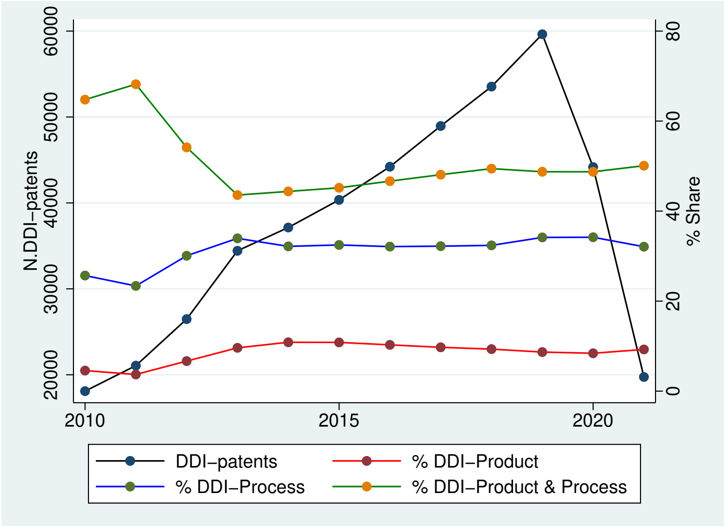

Figure 4: DDI-Product, DDI-Process and DDI-Mix patents: count (left axis) and share of total DDI-patents (right axis), by year.

Table 13: USPTO patents: DDI-Product vs. DDI-Process patents

		DDI-Product		DDI-Process		DDI-Mix	
Year	**DDI pat.**	**N.**	**%**	**N.**	**%**	**N.**	**%**
2010	18,097	826	4.6	4648	25.7	11721	64.8
2011	21,063	779	3.7	4924	23.4	14364	68.2
2012	26,481	1768	6.7	7966	30.1	14343	54.2
2013	34,429	3312	9.6	11689	34.0	14999	43.6
2014	37,138	4031	10.9	11947	32.2	16480	44.4
2015	40,354	4371	10.8	13108	32.5	18233	45.2
2016	44,214	4545	10.3	14199	32.1	20629	46.7
2017	48,954	4767	9.7	15771	32.2	23544	48.1
2018	53,542	4998	9.3	17339	32.4	26469	49.4
2019	59,636	5176	8.7	20372	34.2	29071	48.7
2020	44,175	3717	8.4	15105	34.2	21529	48.7
2021	19,744	1829	9.3	6337	32.1	9889	50.1
Overall	447,827	40,119	8.9	143,405	32.0	221,271	49.1

Notes: DDI-patents classified as Product-Only, Process-Only, or Product&Process (DDI-Mix).

10 countries by share in total DDI-patents per year. The US dominate over the entire period, followed by Japan, Germany, and Korea. China has sensibly increased its role over time.

DDI patents in Firms

Patent documents are matched to 230,441 firms in the ORBIS dataset via ORBIS-IP. Table 14 provides descriptive statistics. Pooling over time, 50,033 firms have at least one DDI-patent, while 180,408 do not. Labeling firms having at least one DDI-patent as DDI-firms, we find that they are substantially more active in patenting overall: an average of 59.43 patents per firm vs. 3.59 for Non-DDI firms. DDI-firms file on average about 9 DDI-patents, with a median of 1, and the share of DDI-patents in their portfolio increases from about 28% in 2010 to 43% in 2019.

DDI by firm size. Table 15 shows that Non-DDI firms are concentrated among small firms (54%), while DDI-firms are equally split between small and large firms (about 40% each). Small DDI-firms have fewer DDI-patents on average (3.43) but a higher DDI share in their portfolio

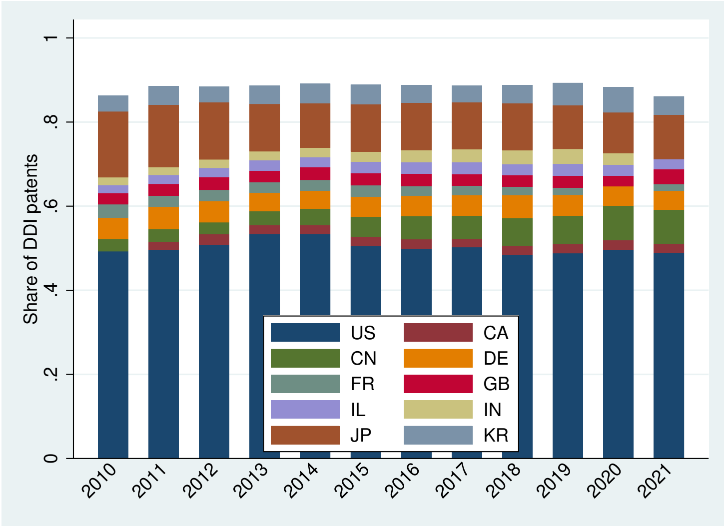

Figure 5: Top-10 countries by share in total DDI-patents per year (fractional counting based on inventors' address).

Table 14: USPTO patents: firm-level DDI

No-DDI firms			**DDI firms**					
	Patents per firm			**Patents per firm**		**DDI-Patents per firm**		
N.Firms	**Mean**	**Median**	**N.Firms**	**Mean**	**Median**	**Mean**	**Median**	**Share**
180,408	3.59	1	50,033	59.43	4	9.45	1	0.59

Table 15: USPTO patents: DDI and firm size

	DDI firms by size		**DDI-patents by size**		
	Non-DDI	DDI	Mean	Median	Share
Small:	159,753	36,015	3.43	1	0.66
	(54.34%)	(40.24%)			
Medium:	68,767	16,523	6.68	2	0.44
	(23.39%)	(18.46%)			
Large:	65,453	36,955	54.98	3	0.25
	(22.26%)	(41.29%)			

Notes: DDI-firms defined as having ≥1 DDI-patent over the period. Size: Small (≤49 employees), Medium (50–249), Large (>249).

(0.66) compared to large firms (54.98 DDI-patents, share of 0.25). This suggests that while large firms patent more in DDI, small firms are relatively more specialised in DDI.

DDI by firm age. Table 16 shows that the age distribution is similar across DDI and Non-DDI firms, but mature firms (10 years) have more DDI-patents on average (21.12 vs. 5.17 for young firms). However, young firms have a substantially higher DDI share (0.65 vs. 0.43), mirroring the size result: smaller and younger firms are relatively more DDI-specialised.

DDI by sector. Table 17 shows the distribution of DDI-firms across NACE sections. DDI-activity is particularly frequent (shares above 40%) in Information and Communication (J: 46.7%), Education (P: 48.8%), and Public Administration (O: 42.7%), reflecting intensity of DDI-related research in ICT, universities and defence. Health (Q: 30.7%) and Professional/Scientific activities (M: 23.9%) also show high DDI presence.

Table 16: USPTO patents: DDI and firm age

	DDI firms by age		**DDI-patents by age**		
	Non-DDI	DDI	Mean	Median	Share
Young:	369,477 (42.44%)	90,010 (43.25%)	5.17	1	0.65
Mature:	501,126 (57.56%)	118,084 (56.75%)	21.12	2	0.43

Notes: Young (≤10 years) vs. Mature (>10 years). DDI-firms: ≥1 DDI-patent.

Table 17: USPTO patents: DDI firms by sector

NACE Sections	**N.Firms**	**DDI-Firms**	**% DDI-Firms**
A - Agriculture, forestry and fishing	2,031	306	15.07
B - Mining and quarrying	1,262	242	19.18
C - Manufacturing	71,644	11,650	16.26
D - Electricity, gas, steam & air conditioning	959	206	21.48
E - Water; sewerage, waste manag.	749	78	10.41
F - Construction	4,948	645	13.04
G - Wholesale and retail	21,965	2,854	12.99
H - Transportation and storage	2,091	552	26.40
I - Accommodation and food service	1,112	169	15.20
J - Information and communication	20,511	9,570	46.66
K - Financial and insurance	7,880	1,811	22.98
L - Real estate activities	2,113	311	14.72
M - Professional, scientific & technical	37,094	8,851	23.86
N - Administrative and support service	5,666	1,306	23.05
O - Public admin. and defence	487	208	42.71
P - Education	2,734	1,333	48.76
Q - Human health and social work	5,008	1,536	30.67
R - Arts, entertainment and recreation	1,200	242	20.17
S - Other service activities	1,860	375	20.16

Notes: Total firms by NACE 1-digit sectors and share of DDI-firms (≥1 DDI-patent, 2010–2021).

6.3 Italian SME Websites

The analysis of 134,208 Italian SME websites yields a binary classification telling whether each firm is a DDI-firm or not. Overall, about 32% of analysed company websites contained explicit references to DDI-related approaches and capabilities.

DDI by Firm Size. Table 18 shows DDI prevalence by firm size class. We observe that DDI shares increase with firm size. Indeed, micro firms (≤10 employees) show the lowest DDI share, while medium firms (50–249 employees) have the highest proportion classified as DDI.

Table 18: Italian SME websites: DDI by firm size

	Micro (≤10)	Small (11–49)	Medium (50–249)
DDI-No:	66,744	35,988	7,484
DDI-Yes:	22,209	12,338	4,357

Notes: Firm size defined by number of employees: Micro (≤10), Small (11–49), Medium (50–249).

DDI by Sector. Table 19 reports DDI prevalence by NACE 1-digit sector, while Table 20 shows the top-10 NACE 2-digit sectors. Information and Communication (J) leads with 49.1% DDI, followed by Water/Sewerage (E: 42.2%) and Professional activities (M: 39%). At the

2-digit level, Computer Programming (J62) stands out at 60.5% DDI, followed by Machinery Manufacturing (C28: 37.7%) and Specialised Construction (F43: 31.7%).

Table 19: Italian SME websites: DDI by NACE 1-digit sector

NACE	N.Firms	%DDI
J - Information and communication	11,111	49.1
D - Electricity, gas, steam & air cond. supply	571	45.0
E - Water; sewerage, waste manag. & remed. act.	1,451	42.2
M - Professional, scientific & technical activities	13,415	39.0
F - Construction	12,612	28.2
C - Manufacturing	42,091	26.0
N - Administrative and support service activities	7,913	22.9
G - Wholesale and retail	38,149	20.0
H - Transportation and storage	4,669	18.1
L - Real estate activities	5,744	18.0
I - Accommodation and food service activities	9,993	12.3

Table 20: Italian SME websites: top-10 NACE 2-digit sectors by DDI prevalence

NACE	N.Firms	% DDI
G46 - Wholesale (exc. motor-vehicles/cycles)	21,631	23.7
G47 - Retail (exc. motor-vehicles/cycles)	11,197	15.5
C25 - Metal products (exc. Machin. & Equip.)	8,624	20.2
F43 - Specialised construction activities	7,569	31.7
J62 - Computer programming and related act.	6,403	60.5
C28 - Manuf. Machinery & Equip.	6,020	37.7
L68 - Real estate activities	5,744	18.3
I56 - Food & Beverage service activities	5,336	8.6
G45 - Wholesale and retail motor-vehicles/cycles	5,321	13.3
I55 - Accommodation	4,657	15.4

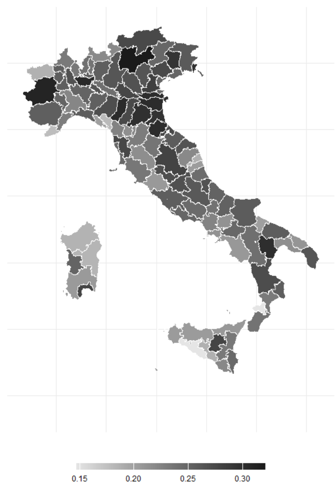

Figure 6: Distribution (% share) of DDI firms by Italian province (NUTS3).

DDI by Geography. Figure 6 shows the geographical distribution of DDI firms across Italian provinces (NUTS3). DDI frequencies follow the overall distribution of firms closely (correlation above 90%), but with meaningful variation in intensity. Table 21 reports the top-10 and bottom-10 provinces by DDI prevalence. Top provinces include Trento (32.1%), Torino

(31.0%), and Milano (30.5%), while the lowest-ranked provinces (Vibo Valentia, Agrigento, Imperia) show DDI prevalence around 15–18%.

Table 21: Italian SME websites: top and bottom provinces by DDI prevalence

TOP provinces				BOTTOM provinces			
Rank	Province	N.Firms	% DDI	Rank	Province	N.Firms	% DDI
1	Trento	1,637	32.1	1	Vibo Valentia	136	14.7
2	Torino	5,338	31.0	2	Agrigento	333	14.7
3	Milano	16,921	30.5	3	Imperia	287	17.8
4	Reggio Emilia	1,812	30.3	4	Massa-Carrara	450	18.0
5	Forli'-Cesena	1,150	30.3	5	Nuoro	185	18.4
6	Ferrara	749	30.2	6	Sassari	751	18.5
7	Rovigo	492	30.1	7	Fermo	525	18.7
8	Matera	283	30.0	8	Valle d'Aosta	251	18.7
9	Gorizia	258	29.8	9	Caltanissetta	250	18.8
10	Bologna	3,909	29.8	10	La Spezia	403	20.1

6.4 Crunchbase data on Startups

DDI Prevalence. Of the 9,705 firms in the Crunchbase sample, 5,461 (56.3%) are classified as DDI, while 4,244 (43.7%) are classified as No-DDI. The higher DDI prevalence compared to other sources reflects the nature of the platform, which attracts technology-oriented startups.

Table 22: Crunchbase: DDI vs. Non-DDI companies (firms with status and funding data)

DDI Status	N	Pct	Succ. Exit (%)	Fail Exit (%)	Operating (%)	Mean Funding ($M)
DDI	3,656	63.7	1.6	2.9	90.3	267.25
Non-DDI	2,084	36.3	0.8	7.2	89.7	103.02

Notes: Restricted to 5,740 firms with available status and funding information. DDI prevalence in the full sample (9,705 firms) is 56.3%.

Table 22 provides a comparison between DDI and No-DDI companies for the subset of 5,740 firms with available status and funding information. Within this subset, DDI companies represent 63.7%, have a higher successful exit rate (1.6% vs. 0.8%), a lower failure rate (2.9% vs. 7.2%), and substantially higher mean funding ($267M vs. $103M).[2] The funding differential is particularly striking. DDI startups attract on average 2.6 times more capital than their no-DDI counterparts, consistent with investors' recognition of data-driven approaches as value-creating strategies.

DDI by Foundation Year. Figure 7 shows the distribution of companies by founding year and the percentage of DDI firms over time. The share of DDI firms increases among more recently founded companies, consistent with the growing centrality of DDI in new venture creation.

DDI by Firm Size. Table 23 and Figure 8 show that DDI prevalence increases sharply with firm size: from 48% among micro firms to 91.5% among large firms. This pattern likely reflects that larger startups tend to be more data-intensive and technology-focused.

DDI by Sectors and Technologies. Unlike the other data sources, Crunchbase does not provide NACE sector codes, precluding a direct sectoral comparison. However, we could extract the technology categories identified by the LLM as justification for DDI assignment, which can serve to proxy for the innovation domains in which DDI startups operate.

[2] A successful exit is defined when a start-up is acquired for an amount higher than the total capital raised or undergoes an IPO.

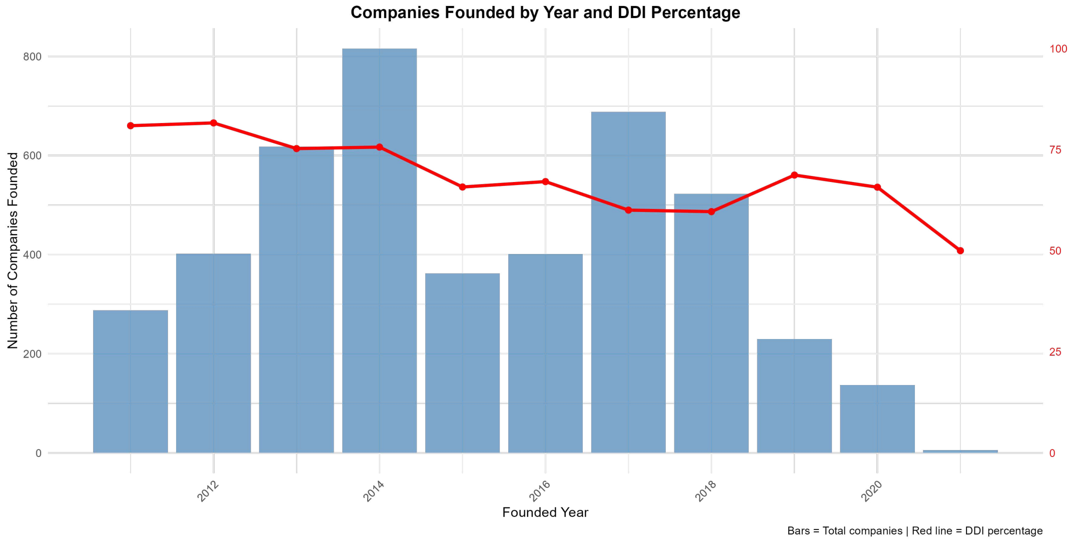

Figure 7: Crunchbase: N. of firms founded (left axis) and % of DDI-firms (right axis), by year.

Table 23: Crunchbase: DDI by employee cohort

Employee Cohort	Total	DDI Count	DDI %	Success Rate DDI (%)
Micro	2,527	1,212	48.0	0.7
Small	820	541	66.0	1.3
Medium	414	338	81.6	3.6
Large	1,248	1,142	91.5	1.9

Notes: Micro (≤10), Small (11–50), Medium (51–200), Large (>200 employees).

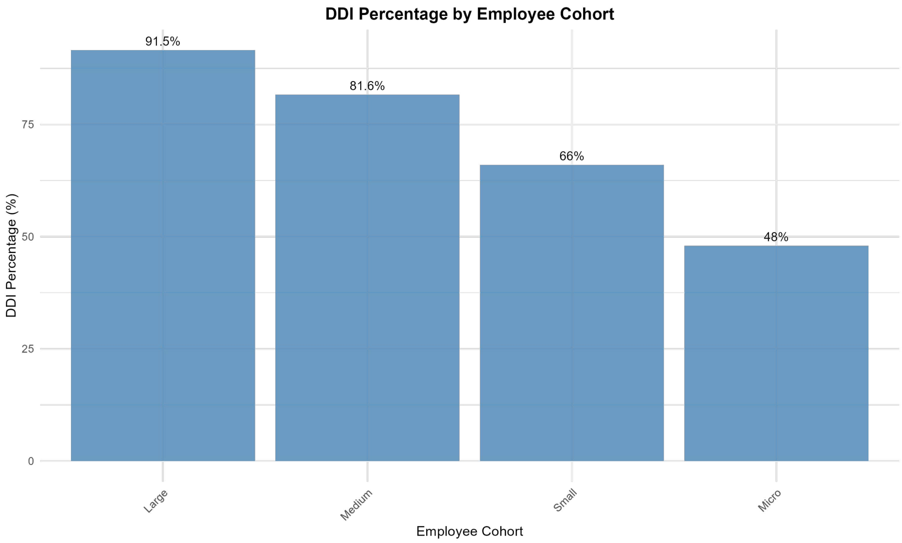

Figure 8: Crunchbase: DDI percentage by employee cohort.

Table 24 reports the technology categories identified in DDI justifications. Data/Analytics is the dominant category (87.8%), followed by AI/ML (42.5%), Automation (12.9%), and Blockchain/Crypto (8.9%).[3] This hints at a concentration of DDI activity in Data/Analytics and AI/ML categories, consistent with the sectoral patterns observed in patents and communications, where Information and Communication leads. However, the presence of Healthtech

[3]The percentages sum to more than 100% because a single firm's description may reference multiple technology categories.

Table 24: Crunchbase: technology categories in DDI justifications

Technology	Count	Percentage
Data/Analytics	3,211	87.8
AI/ML	1,554	42.5
Automation	471	12.9
Blockchain/Crypto	325	8.9
IoT/Hardware	279	7.6
Mobile/App	273	7.5
Cloud/SaaS	161	4.4
Healthtech	63	1.7
Fintech	28	0.8
Sustainability	23	0.6
E-commerce	13	0.4

Notes: Based on 3,656 DDI-classified firms with available data. Percentages sum to more than 100% because firms may span multiple categories.

(1.7%), Fintech (0.8%), and E-commerce (0.4%) as distinct categories, corroborates that DDI diffuses across multiple industry verticals, even when measured through a platform that attracts predominantly tech-oriented firms. Overall, this evidence suggests that, in the startup ecosystem, AI plays a major but not exclusive role, echoing the patent-level finding that the majority of DDI activity does not involve AI. The presence of IoT/Hardware (7.6%) and Cloud/SaaS (4.4%) further highlights the infrastructure dimension of DDI identified in our theoretical framework (recall Section 3).

6.5 Cross-Source Comparison

The four datasources examined in this study capture different facets of DDI across different dimensions: the distinct firm populations covered, the geographic and temporal scope, the unit of observation and the function performed by the document types. Corporate communications reveal the prominence that listed firms assigned to DDI when reporting their strategies and activities to investors. Patents capture the technical embodiment of data-driven inventive activity. Corporate websites indicate how SMEs present their capabilities, products and services to customers and other market participants. Crunchbase descriptions reveal the extent to which data and analytics are central to the identity and value propositions of entrepreneurial ventures.

Given the heterogeneity of the empirical settings, cross-source comparison should focus primarily on the direction of the patterns and the dimensions of heterogeneity, not on the ranking of populations according to their raw DDI prevalence. Table 25 summarises the key findings along broadly comparable dimensions. Several patterns are remarkably consistent across sources but, at the same time, the sources reveal instructive divergences.

Common Patterns

The first common result is the *increasing importance of DDI over time* in the datasources allowing to track intertemporal dynamics. In US corporate communications, mean DDI-Share –measuring average relevance of DDI-related sentences– increases from approximately 4.6% in 2008 to 5.9% in 2022. In the patent sample, the proportion of DDI patents doubles from 9% in 2010 to approximately 18% in 2020–2021. Crunchbase provides cohort rather than panel evidence, but the proportion of firms classified as DDI is generally higher among more recently founded ventures. As a result, the sources consistently indicate that the use of data and analytics for innovation purposes has become increasingly central to both established firms and new ventures.

These commonalities are relevant because each source captures a different stage or man-

Table 25: Cross-source comparison of DDI findings

Dimension	Communications	Patents	Websites	Crunchbase
Population	Listed firms (IT, US)	Global patenting firms	Italian SMEs	Global startups
DDI incidence	97–99% (US); 15–35% (IT)	13.2% of patents; 21.7% of firms	32% of firms	56.3% of firms
Trend	Gradual increase	Doubles (9%→18%)	Cross-section	Rising with cohort
Leading sectors	Info & Comm (J); Education(P)	Info & Comm (J)	Info & Comm (J)	Data/Analytics
Firm Size	Larger firms more intense (US)	Larger firms: more patents; smaller firms: higher DDI share	Increases with size	Sharply increases
Firm Age	Younger more DDI-intense	Younger: higher DDI share	n/a	Rising in recent cohorts

Notes: Comparable dimensions across data sources. "DDI incidence" indicates the share of observations (documents or firms) classified as DDI.

ifestation of the innovation process. The patent trend reflects growth in technically codified data-driven inventions, while the corporate communications trend reflects the increasing strategic attention assigned to DDI by listed firms. The Crunchbase cohort pattern suggests that DDI is becoming more deeply embedded in the business models of newly established firms. The simultaneous emergence of these patterns reduces the likelihood that the observed increase might solely be driven by changing terminology or by the characteristics of one particular textual source.

The second common pattern is the *association between DDI and Information and Communication activities*. Information and Communication (NACE section J) is the most DDI-intensive sector in the corporate communications and website samples, reaching 80.4% of DDI-intense observations in US communications and 49.1% DDI prevalence among Italian SMEs' websites. In the patent data, Information and Communication is also among the leading sectors, with 46.7% of firms holding at least one DDI patent (although Education records a slightly higher share of 48.8%). The concentration of Crunchbase descriptions in Data/Analytics and AI/ML provides further, though not directly sector-comparable, evidence of the central role of information-related technologies.

At the same time, our findings across datasources agree that DDI is not confined to ICT-producing industries but it can be detected in manufacturing, construction, transportation, utilities, wholesale and retail, accommodation, health, and other service sectors. This broad diffusion is consistent with the conceptualisation developed in Section 3 according to which data and analytics constitute enabling factors that can augment multiple stages of the innovation process and can therefore be applied across widely different technological and market environments.

The third common result is that *DDI cannot be reduced to the adoption of artificial intelligence*. In the patent sample, 59% of patents classified as DDI do not satisfy the WIPO criteria for AI patents. Similarly, although AI/ML appears in 42.5% of the DDI-classified Crunchbase descriptions, the broader Data/Analytics category appears in 87.8%. This evidence suggests that AI is a major component of contemporary DDI, but DDI includes statistical analytics, sensor-based optimisation, data infrastructures, automation, cloud services, and other forms of data-enabled value creation that do not necessarily rely on artificial intelligence.

Heterogeneities

A first diverging pattern regards the relationship between firm size and DDI, which seems less consistent across data sources. On the one hand, in US corporate communications, larger firms exhibit higher DDI intensity, consistent with their greater availability of financial, organizational, and technical resources. DDI prevalence also increases with firm size in the Italian website sample and among Crunchbase ventures, where it rises from 48% for micro firms to 91.5% for large firms.

On the other hand, the patent evidence, reveals a distinction between the scale and the specialisation of DDI. With an average of approximately 55 patents, compared with 3.4 among small firms, large patenting firms file more DDI patents in absolute terms. Nevertheless, DDI accounts for only 25% of the patent portfolios of large DDI firms, compared with 66% among small DDI firms.

This extensive vs. intensive margin distinction helps reconcile the apparently divergent findings. Larger firms possess the complementary assets required to invest in data infrastructure, recruit specialised personnel, undertake multiple DDI projects, and communicate these initiatives to external stakeholders. Smaller firms have fewer projects in absolute terms but may be founded around a narrower data-driven technological or market opportunity. They may therefore display *greater DDI specialisation* without matching the aggregate DDI output of larger organisations. The Crunchbase pattern is also likely to signal a selection process. Ventures that successfully grow may be those that have developed sufficiently strong data capabilities, attracted external funding, or scaled data-intensive business models.

A similar distinction emerges for firm age. Younger firms are relatively more DDI-intensive in corporate communications. Also, young patenting firms have a substantially higher DDI share in their patent portfolios than mature firms, despite producing fewer DDI patents in absolute terms. Together with the Crunchbase cohort evidence, this pattern suggests that DDI may be more central to the technological identity of younger organisations, whereas established firms integrate DDI into broader and more diversified portfolios. Differences in sectoral composition, firm survival, disclosure practices, and sample selection may however contribute to the observed relationship.

All in all, these inconsistencies are just ostensible and dissolve when one considers that each source captures a different margin of DDI activity. Corporate communications reflect *strategic emphasis and external signaling* (how prominently a firm signals DDI to investors); patents measure *inventive output* (the technical embodiment of DDI); websites signal *market facing capabilities and commercial* (how a firm presents data-driven capabilities to clients; Crunchbase profiles reveal *entrepreneurial orientation and venture identity* (whether data is central to the venture's identity). A large firm may discuss DDI prominently in earnings calls (high communication intensity) while maintaining a broad patent portfolio in which DDI is a modest share. A small startup may be entirely DDI-focused yet produce no patents.

These complementarities underscore the nuances of the DDI concept as well as the value of a multi-source approach. As DDI activity is heterogeneous, no single source can capture its full spectrum. Reliance on any one source would produce a partial and potentially misleading picture of DDI diffusion across the economy.

7 Conclusion

Measuring DDI at the firm level is a challenging task, especially when the aim is to obtain a broad, large-scale view beyond small ad hoc surveys or individual case studies. This study has addressed this challenge by exploiting the power and flexibility of Large Language Models to extract information about a complex construct – "using big data and data analytics for the specific purpose of innovation" – from multiple textual data sources.

The multi-source framework developed here encompasses corporate communications, patent documents, SME websites, and startup descriptions. It provides complementary angles on DDI in firms. Each source captures a different facet. Corporate communications reveal how listed companies signal DDI strategies to financial markets. Patents document the technical implementation of data-driven inventions. Websites reflects how SMEs present their data-driven capabilities to the market and potential customers. Crunchbase profiles reveal how startups position data and analytics within their business models.

The descriptive evidence presented here allows to document empirical facts about DDI, in turn having implications for policy and future studies.

Key findings:

The main empirical patterns, consistent across datasources, reveal that:

- **DDI is growing**: Whether measured through corporate communication intensity, patent shares, or startup prevalence, DDI shows a clear upward trend over the period examined.
- **DDI is pervasive**: It appears across all sectors, though with considerable variation. Information and Communication consistently leads, but even traditionally low-tech sectors show meaningful DDI activity.
- **DDI is heterogeneous**: The relationship between DDI and firm characteristics varies by data source. In patents, smaller and younger firms show higher DDI specialisation (DDI share in portfolio). In communications, larger US firms are more DDI-intense. Among startups, DDI prevalence increases strongly with firm size.
- **DDI is broader than AI**: The cross-tabulation of DDI and AI patents confirms that the majority (59%) of DDI-patents do not qualify as AI, underscoring that DDI encompasses data-driven approaches beyond artificial intelligence.
- **DDI strategies differ**: The acquisition vs. development analysis reveals that most US firms combine both strategies, while Italian firms show more volatile patterns, suggesting less consolidated DDI approaches.

These findings provide a rich empirical mapping of DDI that can serve as a foundation for subsequent analytical work.

Policy Implications:

The pervasiveness of DDI across sectors and across firm types has direct implications for innovation policy:

- Policies aimed at fostering data-driven innovation should not be confined to high-tech sectors. Our evidence shows meaningful DDI activity in construction, agriculture, wholesale, and accommodation. Sector-neutral instruments, such as digital skills training, data infrastructure subsidies, and regulatory sandboxes for data sharing, may be more effective than targeted sectoral programmes.
- Policies facilitating access by of data-specialised small/medium and young firms to cloud computing, open data, and analytics-as-a-service could help unlock the innovation potential of these firms. In fact, the finding that smaller and younger firms show higher DDI specialisation (particularly in patents) suggests that these firms face binding constraints in scaling their DDI activities.
- The marked gap between Italy and the US in DDI intensity among listed firms points to structural differences in digital readiness, managerial capabilities, and ecosystem maturity that merit targeted intervention.

Research Agenda:

The firm-level DDI measures developed in this report open several avenues for future research:

- **DDI and firm performance**: Do firms with higher DDI intensity achieve superior productivity, profitability, or market valuation? The panel structure of corporate communications and patent data enables causal identification strategies exploiting within-firm variation over time. The available information on the sources and structure of funding rounds for start-ups enable an analysis of the relationship between DDI, selection into funding and performance outcome.

- **DDI and labour markets**: How does DDI affect employment composition, skill demand, and wage structures within firms? Linking DDI measures to matched employer-employee data could illuminate the labour-market implications of data-driven innovation.
- **DDI and knowledge spillovers**: Do DDI-active firms generate knowledge spillovers to geographically or technologically proximate firms? Patent citation networks and co-location analysis provide natural empirical strategies.
- **DDI determinants**: What firm-level, industry-level, and institutional factors predict DDI adoption? The cross-country and cross-sector variation documented here provides a starting point for multivariate analysis.
- **DDI and sustainability**: The overlap between data-driven optimisation and resource efficiency suggests a link between DDI and environmental performance that remains largely unexplored.

Methodological Extensions:

The LLM-based measurement framework is not static. Several extensions are natural as the underlying technology and data landscape evolve. First, more recent and capable models (e.g., GPT-4, Llama 3) may improve classification accuracy, particularly for edge cases involving aspirational language or ambiguous means–purpose linkages. Second, the framework could be extended to additional textual sources such as job postings (revealing DDI-related hiring), news articles (capturing market-facing DDI narratives), or social media (reflecting real-time technology adoption signals). Third, the binary DDI/Non-DDI classification could be enriched with a multi-dimensional taxonomy, distinguishing for example between product DDI, process DDI, and business-model DDI within a single text. Finally, the temporal dimension could be strengthened by applying the framework to historical corpora, enabling longer-run analysis of DDI diffusion since the early 2000s.

References

Arts, S., Hou, J., and Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.

Ashouri, S., Hajikhani, A., Suominen, A., Pukelis, L., and Cunningham, S. W. (2024). Measuring digitalization at scale using web scraped data. *Technological Forecasting and Social Change*, 207:123618.

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., and Liberati, C. (2024). Scraping innovativeness from corporate websites: Empirical evidence on italian manufacturing smes. *Technological Forecasting and Social Change*, 207:123597.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies "engines of growth"? *Journal of Econometrics*, 65(1):83–108.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.

Cockburn, I. M., Henderson, R., and Stern, S. (2018). The impact of artificial intelligence on innovation: An exploratory analysis. Working Paper 24449, National Bureau of Economic Research.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., and Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165:234–246.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Ghasemaghaei, M. and Calic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research*, 108:147–162.

Hain, D. S., Jurowetzki, R., Buchmann, T., and Wolf, P. (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177:121559.

Kinne, J. and Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLOS ONE*, 16(4):e0249071.

Kline, S. J. and Rosenberg, N. (1986). An overview of innovation. In Landau, R. and Rosenberg, N., editors, *The Positive Sum Strategy: Harnessing Technology for Economic Growth*, pages 275–305. National Academy Press, Washington, DC.

Kusiak, A. (2009). Innovation: A data-driven approach. *International Journal of Production Economics*, 122(1):440–448.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, J. (2022). Data-driven innovation: What is it? *IEEE Transactions on Engineering Management*, 70(2):784–790.

Mikalef, P., Boura, M., Lekakos, G., and Krogstie, J. (2019). Big data analytics capabilities and innovation: The mediating role of dynamic capabilities and moderating effect of the environment. *British Journal of Management*, 30(2):272–298.

OECD (2015). *Data-Driven Innovation: Big Data for Growth and Well-Being*. OECD Publishing, Paris.

Trabucchi, D. and Buganza, T. (2019). Data-driven innovation: switching the perspective on big data. *European Journal of Innovation Management*, 22(1):23–40.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2):118–144.

Visvizi, A., Troisi, O., Grimaldi, M., and Loia, F. (2022). Think human, act digital: activating data-driven orientation in innovative start-ups. *European Journal of Innovation Management*, 25(6):452–478.

Webb, M. (2020). The impact of artificial intelligence on the labor market. Technical report, Stanford University.

Whitehead, A. N. (1925). *Science and the Modern World*. Macmillan, New York.

WIPO (2019). *WIPO Technology Trends 2019: Artificial Intelligence*. World Intellectual Property Organization, Geneva.