Making official statistics AI-ready with SDMX

Stats Café: Data for AI: making official statistics AI-ready

Jens Dossé (OECD, SDMX-TWG)

OECD

sdmx
Statistical Data and Metadata eXchange

Unclassified - Non classifié

Outline

- What is AI?
- Common SDMX – AI use cases
- Criteria for AI-readiness of statistical data
- AI-readiness of SDMX
- TWG closing gaps in AI-readiness of SDMX
- SDMX-readiness of AI
- Annexes



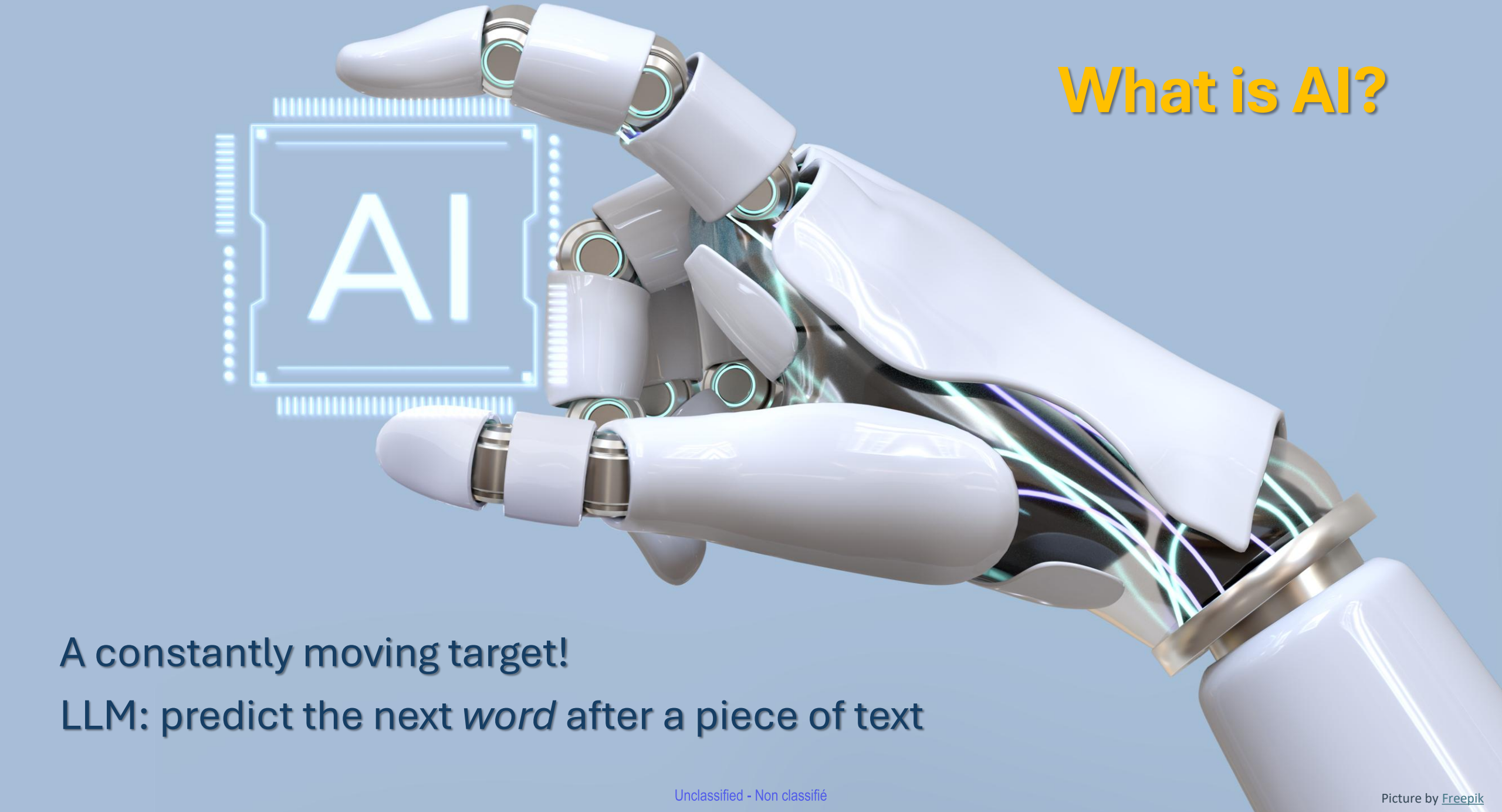
OECD

sdmx
Statistical Data and Metadata eXchange

Unclassified - Non classifié

Picture by Freepik

What is AI?



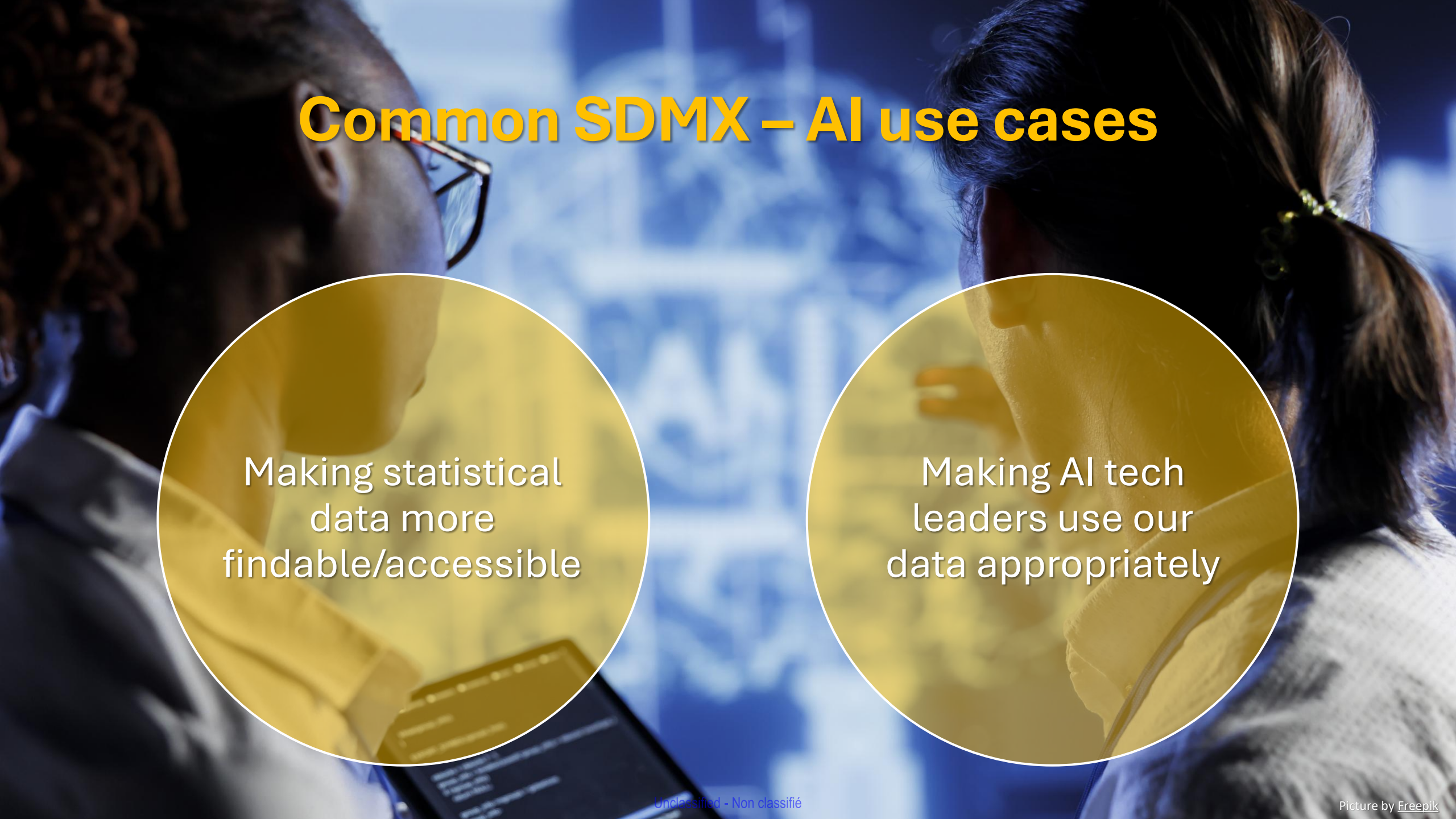
A constantly moving target!

LLM: predict the next *word* after a piece of text

Unclassified - Non classifié

Picture by Freepik

Common SDMX – AI use cases



Unclassified - Non classifié

Picture by Freepik

Today

What would be the criteria for statistical data to be AI-ready?

Copilot

To make statistical data **AI-ready**, it must meet several criteria that ensure it can be effectively used for training, validating, or deploying AI models. These criteria span **data quality**, **structure**, **context**, and **compliance**. Here's a breakdown:

---

1. Data Quality

- **Completeness**: Minimal missing values; if missing, they should be handled or imputed appropriately.
- **Accuracy**: Data should reflect real-world values or measurements reliably.
- **Consistency**: Uniform formats, units, and naming conventions across datasets.
- **Timeliness**: Up-to-date data, especially for time-sensitive applications.

---

2. Data Structure & Format

- **Structured Format**: Preferably in tabular form (e.g., CSV, Excel, SQL) with clear column headers and data types.
- **Standardized Encoding**: Categorical variables should be encoded (e.g., one-hot, label encoding) for model compatibility.
- **Normalized/Scaled Values**: Numerical data often needs normalization or standardization for model performance.

Criteria for AI-readiness of statistical data

(according to AI 😉)

1. Data Quality

- Completeness: Minimal missing values; if missing, they should be handled or imputed appropriately.
- Accuracy: Data should reflect real-world values or measurements reliably.
- Consistency: Uniform formats, units, and naming conventions across datasets.
- Timeliness: Up-to-date data, especially for time-sensitive applications.

2. Data Structure & Format

- Structured Format: Preferably in clear text-readable tabular form (e.g., CSV, Excel, JSON) or embedded in well-structured tables within documents with clear column headers and data types, consistent rows, and no nested tables, merged cells or ambiguous formatting.
- Standardized Encoding: Categorical variables should be encoded (e.g., one-hot, label encoding) for model compatibility.
- Multilingual clarity: If multilingual, clearly separate or label languages. Use UTF-8 encoding to ensure the file uses standard character encoding to avoid misinterpretation of accents or symbols.
- Normalized/Scaled Values: Numerical data often needs normalization or standardization for model performance.
- Consistent formatting: Dates, numbers, and categories should follow a consistent format.

3. Semantic Clarity through Metadata & Documentation

- Clear Definitions: Each column name or variable should be meaningful (or have a documented meaning using understandable categories) and include unit and scales.
- Natural language compatibility: Avoid jargon-heavy or overly technical formatting unless explained.
- Data Provenance: Information about how and where the data was collected, processed and published.
- Versioning: Track changes to datasets over time for reproducibility.
- Narrative context: If possible, embed the data within a short explanatory paragraph.
- Example queries: Optionally include sample questions or use cases to guide interpretation.

4. Contextual Relevance

- Domain-Specific Suitability: Data should be relevant to the problem the AI is solving.
- Balanced Representation: Avoid bias by ensuring diverse and representative samples.
- Temporal Context: For time-series or evolving data, timestamps and sequence integrity are crucial.

5. Legal & Ethical Compliance

- Privacy & Anonymization: Personally identifiable information (PII) must be removed or anonymized.
- Consent & Licensing: Ensure data usage complies with legal agreements and ethical standards.
- Bias Auditing: Check for and mitigate harmful biases in the dataset or note any known limitations or biases in the dataset.

6. Technical Compatibility

- Machine-Readable Format: Easily ingestible by AI pipelines (e.g., JSON, Parquet, CSV).
- Interoperability: Compatible with common AI frameworks (TensorFlow, PyTorch, Scikit-learn).
- Scalability: Suitable for processing at scale, especially for big data applications.

AI readiness...



= AI optimisation (extending SEO)

~ human readiness

Unclassified - Non classifié

Picture by Freepik

So, is SDMX data AI-ready?

Statistical Data and Metadata eXchange



1) Standardised container

- Information model, technical formats & protocols (API) for streamlined data processing, transmission and storage
- Model Context Protocol (MCP) as middleware to AI agents

→ discoverable, machine-accessible and up-to-date data under public domain/open license or with protected access

→ interoperability, timeliness, accessibility, security, efficiency

Image by freepng Image by freepik Icon by Vectors Market - Flaticon Image in public domain

So, is SDMX data AI-ready?



2) Standardised content

- Common language and rules for multidimensional data modelling according to (well-governed) quality standards
- Intra-organisational and international data harmonisation

→ well-structured, well-identified, semantically versioned, harmonised, high-quality data

→ relevance, accuracy, credibility/reliability, coherence, consistency & comparability

Image by pngkey Image by freepik Icon by Vectors Market - Flaticon Image in public domain

So, is SDMX data AI-ready?



3) Augmented content

- Descriptive metadata (clear definitions and contexts at all levels, explanatory documentation, incl. for known issues)
- Standardised semantic relationships (SDMX Ontology, knowledge graphs as linked open data)

→ well understandable, inter-related, AI-ready data

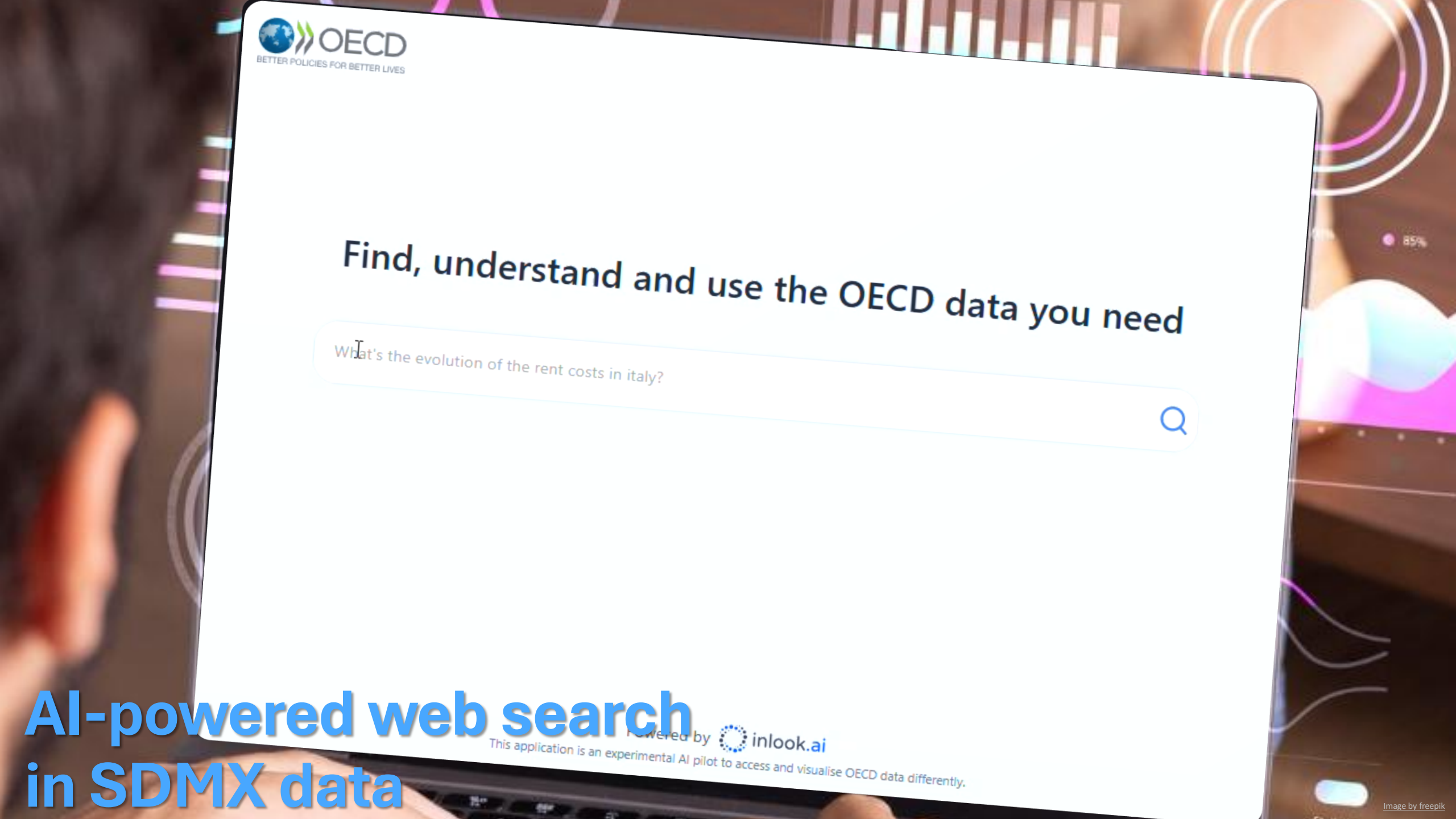
→ interpretability, transparency, reproducibility

Image by pngkey Image by freepik Icon by Vectors Market - Flaticon Image in public domain

AI-powered web search in SDMX data



Image by freepik



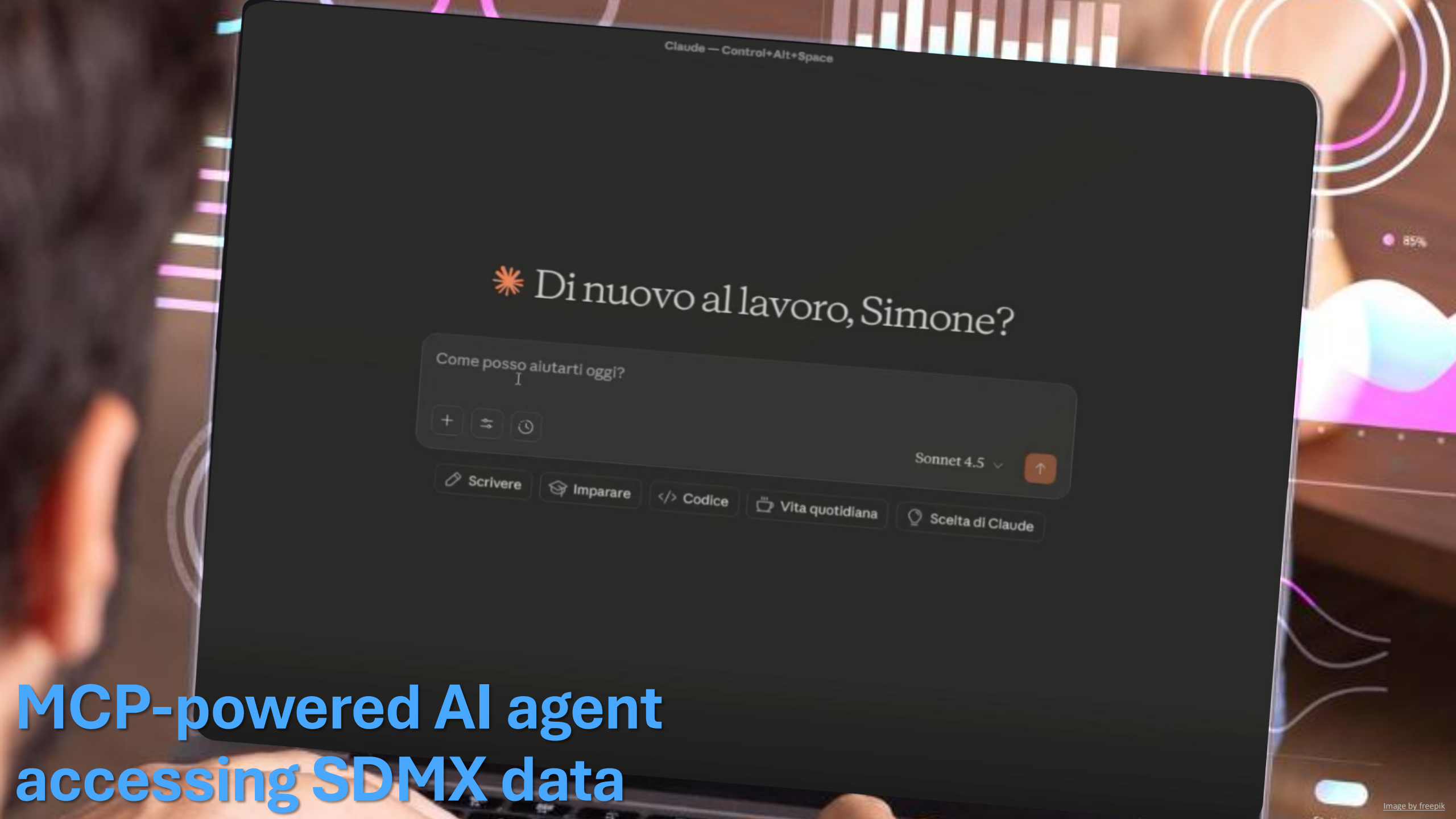
AI-powered web search in SDMX data

Image by freepik

MCP-powered AI agent accessing SDMX data

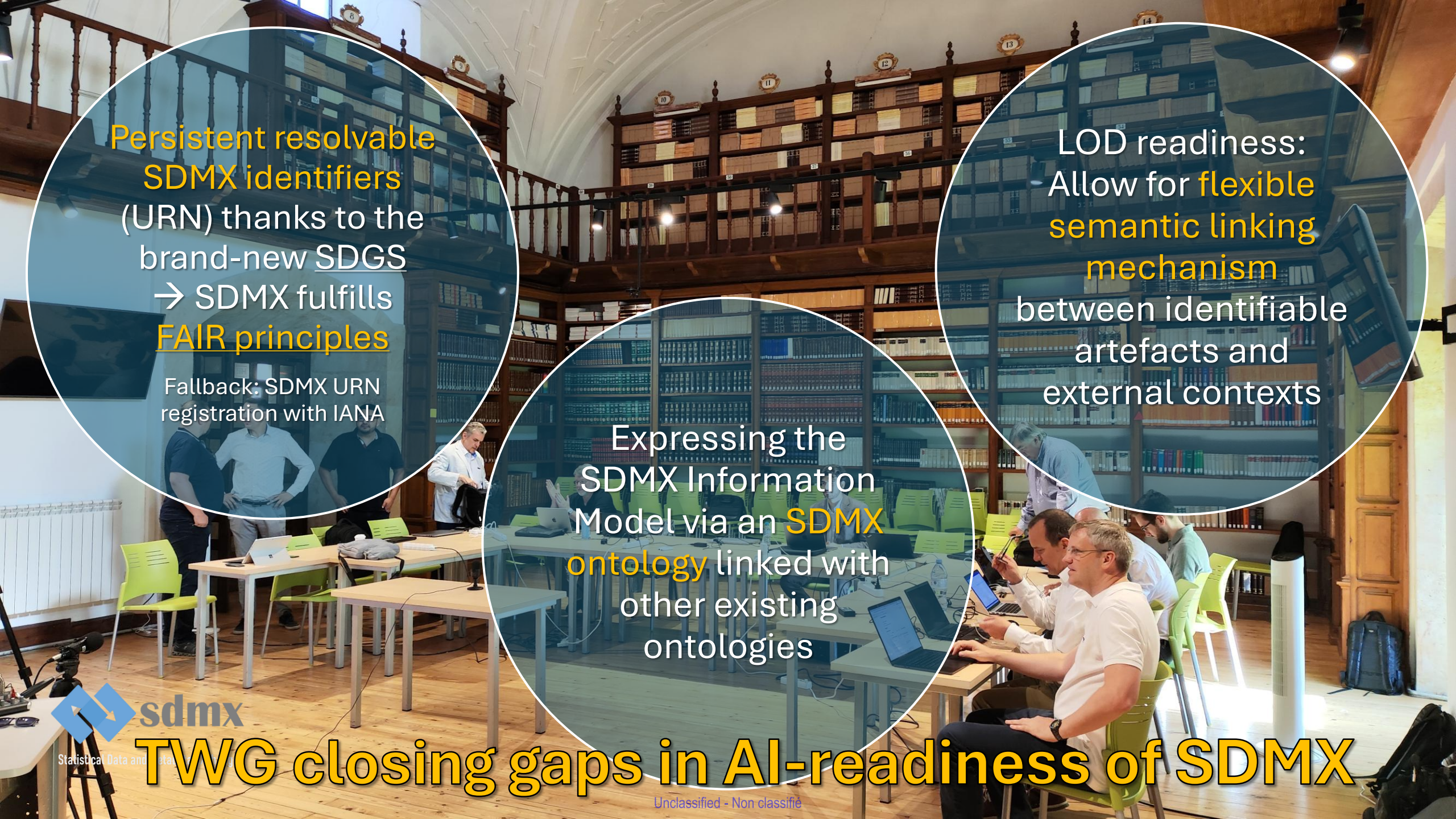


Image by freepik



MCP-powered AI agent accessing SDMX data

Image by freepik



Unclassified - Non classifié

And is AI also SDMX-ready?

AI agents can connect to:

🎉 Google announced in Sep 2025 that Data Commons will allow importing data in SDMX!

SDMX-API

SDMX-powered **dataset catalogues** on the web

SDMX-powered **Model-Context-Protocol** (MCP) services

What else is needed?

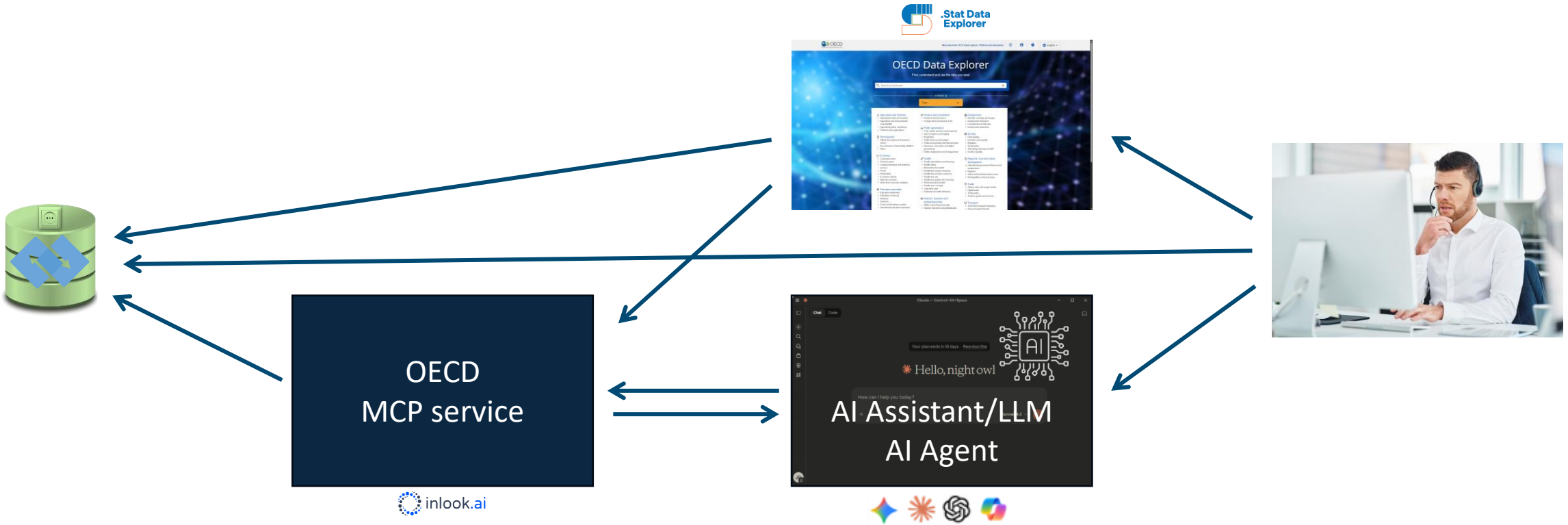
Collaboration & governance Compliancy with agreed rules

Returned services: visibility, feedback & support

Unclassified - Non classifié

Picture by Freepik

What's the future of data dissemination?



Artificial intelligence icons created by juicy_fish - Flaticon



OECD-KOSTAT Global Conference 2025

Annex - Overcoming common challenges in implementing AI-based systems

- Find expertise for the implementation → Partner with advanced public or private organisations
- High costs (depending on model and usage), high environmental impact → Cost-benefit analysis
- Keeping up with quickly advancing technology → Keep solutions simply and cost-effective, maintain and develop expertise, be ready for quick changes
- Probabilistic nature of genAI (LLMs) lead to hallucinations (making things up!)
  - → Avoid misinformation to maintain credibility (zero hallucinations!)
  - → Put responsibility on the user: Let the user validate/decide; choose the document, dataset, indicator, …
- Findability use case:
  - Data availability issues (sparseness/gaps) → Combine AI with checks using SDMX data availability
  - Knowing what data users really need! → Test prototypes with real users, collect feedback

Unclassified - Non classifié

Annex - Capacity building for AI-readiness: institutions, systems, individuals

Institutions (governance & legal frameworks)

- Strategic vision, data governance, innovation, data quality frameworks
- Community of practice approach
- International collaboration, governance at global level
- Agility

Systems (infrastructure & investment)

- SDMX-based, user-centric data lifecycle management platforms, sharing, co-operation
- Workflows, pipelines, automation

Individuals (workforce skills)

- In-depth knowledge of high-quality data modelling (SDMX) and awareness of required metadata content for AI readiness
- Online SDMX training courses

Source: François Fonteneau, OECD, Paris21, at KOSTAT-OECD Global Conference 2025

Video by Vecteezy

Annex - Frequent needs by NSOs in the global south

- State & ecosystem: connectivity, reliable power, cloud access
- Policy/governance: enabling laws, clear mandate, terms of use, open licenses
- Data & platform layer: importance of metadata & modern data pipelines, APIs
- People skills: upskill statisticians, working with data engineers & data scientists, leadership skills for top management
- Value for money: Return on Investment

Source: François Fonteneau, OECD, Paris21, at KOSTAT-OECD Global Conference 2025

Unclassified - Non classifié



Thank you!

Jens Dossé, OECD & SDMX-TWG

OECD

sdmx

Statistical Data and Metadata eXchange

Unclassified - Non classifié

Picture by Freepik