National Artificial Intelligence Centre

Preparing small and medium businesses for AI success

In today's data-driven world, Small and Medium Enterprises (SMEs) must harness the power of Artificial Intelligence (AI) to remain competitive. Before they can fully embrace AI technologies, it's essential for SMEs to ensure that their data is AI-ready. This involves understanding the various types of data they possess and preparing it to train their AI or machine learning models.

Understanding data types

Typically three primary types of data is collected:

1. **Structured data:** This type of data is highly organised and easily interpretable by both business users and machine learning algorithms. It includes quantitative data such as dates, names, addresses, and other customer information.
2. **Unstructured data:** Unstructured data is qualitative and includes raw audio, emails, and Internet of Things (IoT) sensor data. While it provides deeper insights, it can be challenging to process and annotate.
3. **Semi-structured data:** Falling between structured and unstructured data, this type includes formats like HTML and JSON, often used for data transmission between servers and web pages.

The AI data lifecycle

To embark on a successful AI project, businesses should follow a four-stage AI data lifecycle:

1. **Data sourcing:** Good quality data may come from various sources, including publicly available open-source data, purchased commercial data, or data collected in-house.
2. **Data preparation:** This phase involves structuring and labelling raw data, annotating it, and adding metadata to inform and train machine learning models.
3. **Model training and deployment:** AI models and machine learning algorithms are trained using rich, quality data that are sourced and prepared.
4. **Model evaluation by humans:** Data patterns can change over time, so ongoing human monitoring, retraining, and refinement with new data is crucial for improving AI models.

Without a clear understanding of data and strong data governance, you won't be able to effectively feed the AI systems. Data quality and accuracy are paramount, whether you are building or procuring AI solutions.

Aurelie Jaquet, CSIRO Australia's National Science Agency, emphasises the importance of data in AI.

Overcoming barriers to AI

SMEs often face significant barriers when implementing AI, including the high cost and time investment required to start from scratch. To address these challenges, SMEs have three paths to consider: building AI capabilities in-house, purchasing AI solutions, or partnering with third-party experts.

Common obstacles for SMEs include sourcing and preparing high-quality data, which can be slow and expensive. Access to data annotators can also be a challenge. SMEs can tap into the expertise of larger companies or explore emerging technologies like self-supervised or weakly supervised learning to overcome these barriers.

Key lessons for SMEs

SMEs must focus on the following key aspects:

Benchmarking and human evaluation: Good governance involves ensuring the accuracy of AI modeling through benchmarking and human evaluation. Blindly adapting models is unlikely to produce accurate results.

Data handling: SMEs should handle data containing personally identifiable information with care, adhering to data retention and privacy regulations. Safe and secure data storage is essential, and transparency with customers or clients about data usage is a must.

Achieving data culture and maturity

Building a strong data culture within your organisation begins with assessing existing skill capabilities and investing in digital literacy training. A data-mature organization leverages data effectively to achieve its goals, fostering a data-first culture that empowers data use in discussions and decision-making.

Case study: Whisper

Unstructured Data and Weakly Supervised Learning
Whisper, a remarkable open-source neural net AI model, is being trained to transcribe and translate speech audio from 97 different languages. This automated speech recognition (ASR) software was trained on 680,000 hours of unstructured audio data collected from the web, utilising weakly supervised learning. The approach enabled Whisper to achieve high accuracy, even for audio it was not initially trained on.

It's crucial to ensure that AI models work effectively for all users, as highlighted by MingKuan Liu from Appen, given the challenges faced by speech recognition engines in real-world scenarios.

Next steps

When looking to harness the power of AI, understanding the importance of data readiness, data governance, and AI data lifecycle stages is the first step towards a successful AI journey. Collaboration, smart data handling, and ongoing evaluation are key to overcoming barriers and building a data-centric culture.

The National AI Centre is at the forefront of building Australia's responsible and inclusive AI future: **csiro.au/naic**

The National AI Centre is building Australia's responsible and inclusive AI future.

For further information

National AI Centre
1300 363 400
+61 3 9545 2176
naic@csiro.au
csiro.au/naic

Acknowledgements: The National AI Centre is funded by the Australian Government and coordinated by CSIRO.

CSIRO acknowledges CEDA and Google as Foundation Partners of the National AI Centre.

ceda
committee for economic development of australia

Google