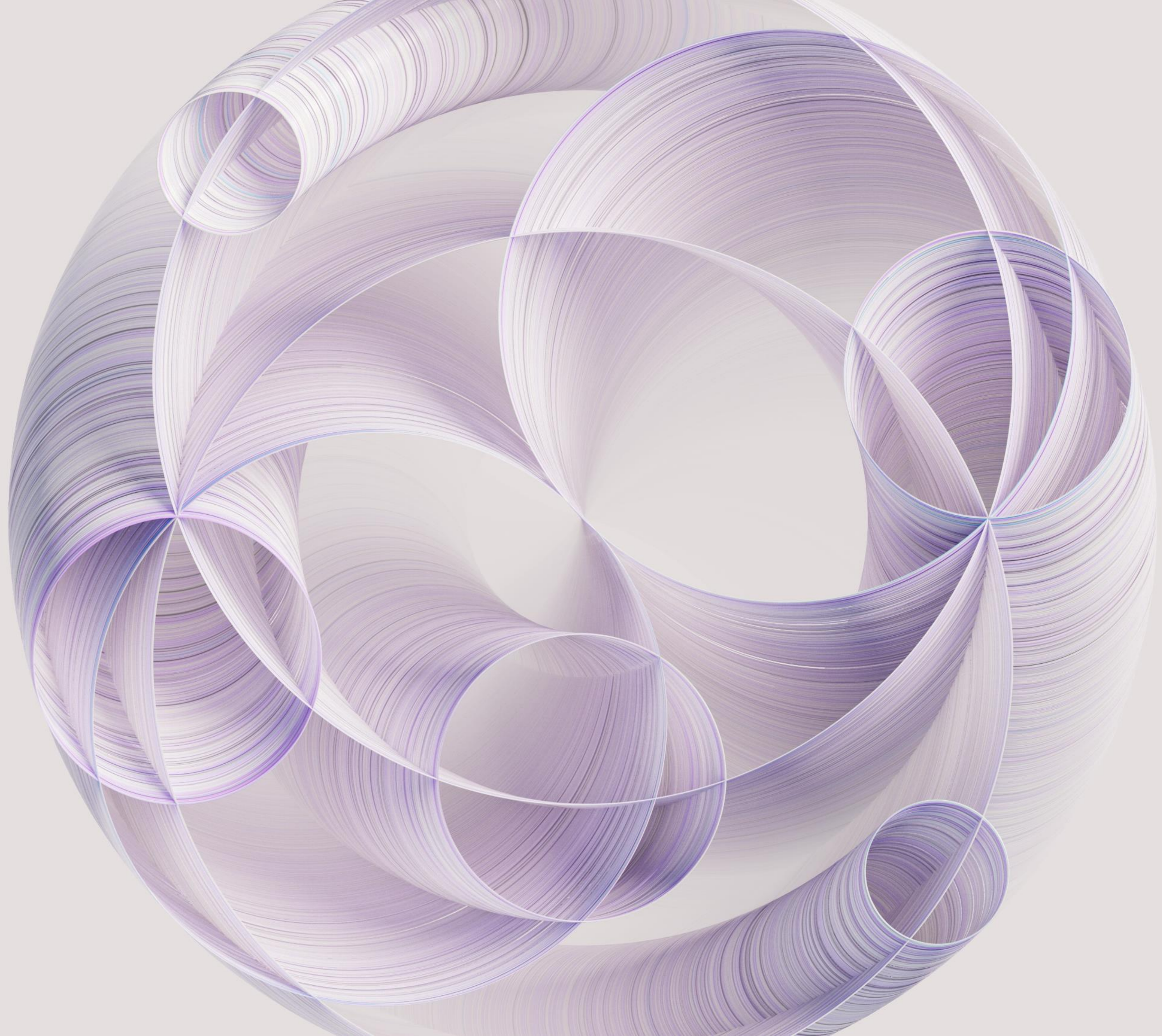


# A quick history of AI at IBM and introducing watsonx.ai

Brett Coffman

[brett.coffman@ibm.com](mailto:brett.coffman@ibm.com)

Principal, Data & AI Technical Specialist



# History of AI at IBM – its not new, but continually changing



**watsonx**



1950s

1980s

1990s

2011

2010's

2020s

**Arthur Samuel's checkers player**  
Self learning player

**Gerald Tesauro's (TD-Gammon)**  
Reinforcement learning (RL) solving complex Real-world problems

**Deep Blue – Computer Chess**  
Combined position, intelligent assessments, beat Kasparov

**Watson – Jeopardy The IBM Challenge**  
Natural Language Understanding, deep learning, speed and accuracy

**Watson Services**  
Speech-to-Text, Text to Speech, Language Translator, Language Classifier, Language Understanding, and open-source tools for use in data science

**Watson Machine Learning**  
Enterprise assistant powered with AI, cloud and the IoT accessed via voice or text interaction

**Watson Assistant**  
Develop real-time, data-backed business models with cognitive analytics and IoT connectivity

**Watson IoT Platform**  
Train, validate, tune and deploy foundation and machine learning models

**watsonx.ai**  
Scale AI workloads for your data, anywhere with a data lakehouse that is open, hybrid, and governed

**watsonx.data**  
Accelerate responsibility, transparency and explainability in your data, and governed

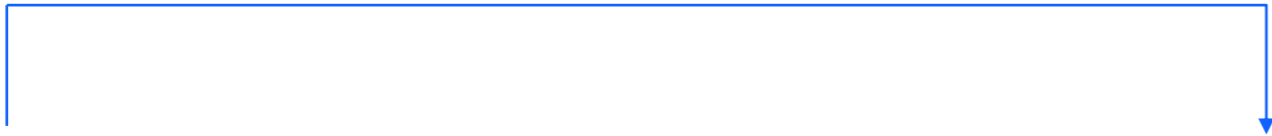
**watsonx.governance**  
Foundation Models, Code-Assistants, Models, Assistants and Chatbots designed enterprise, optimized for business use cases



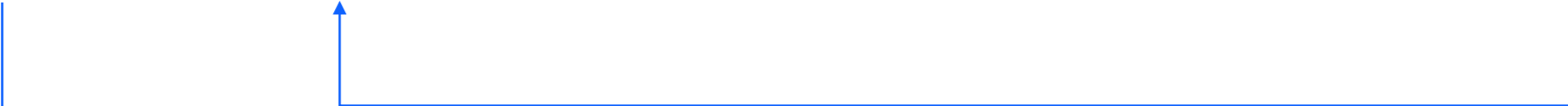
# Put AI to work with watsonx

Scale and accelerate the impact of AI with trusted data

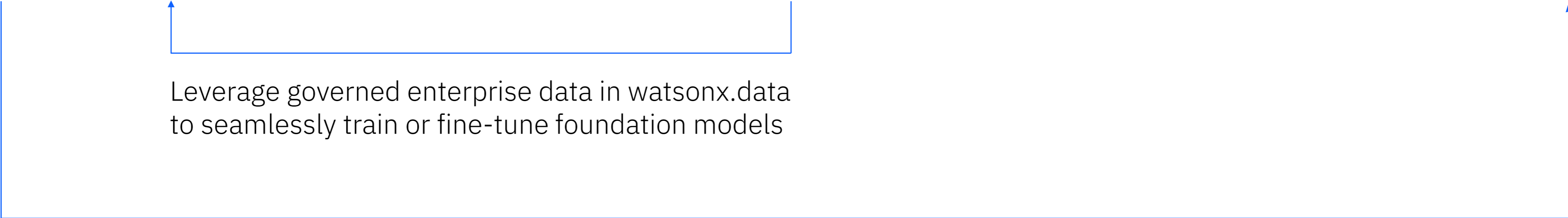
Leverage foundation models to automate data search, discovery, and linking in watsonx.data



Leverage governed enterprise data in watsonx.data to seamlessly train or fine-tune foundation models



Enable fine-tuned models to be managed through market-leading governance and lifecycle management capabilities



# watsonx.ai

## generative AI tools

### Prompt Engineering

- Learn and develop Generative AI skills
- Experience different models and parameters
- Learn from examples or experiment with shot prompting

### Chat UI

- Provides an interactive chat experience with LLMs
- Sample Chat instructions to achieve good completion
- Chat in a RAG use case

### Code Samples

- Sample code generation and translation
- Sample codes in Curl, Node.js, and Python
- Prompt save as a template, prompt session, or as a Jupyter Notebook

### Prompt/Fine-tuning

- Use labeled data to train LLMs to support specific downstream tasks
- Enhance models with business-specific knowledge and information

### InstructLab

- Allows SMEs to collaborate and provide expert input into an LLM
- Improve the knowledge and skill levels of the LLM
- Deploy the enhanced LLM to solve business issues

### Synthetic Data

- Generate data for testing, developing AI applications
- Generate using a specified schema and distributions
- Generate by mimicking an existing data set
- Privacy and security

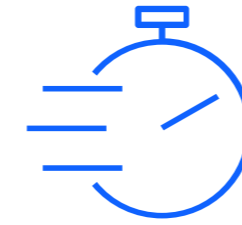
# watsonx.data

Scale AI workloads,  
for all your data,  
anywhere

A fit-for-purpose data  
store, based on an open  
lakehouse architecture,  
supported by querying,  
governance and open  
data formats to access  
and share data



Access all your data  
through a single point  
of entry across all  
clouds and on-premises  
environments.



Get started in  
minutes with built-in  
governance, security  
and automation.



Reduce the cost of  
your data warehouse  
by up to 50%\*  
through workload  
optimization across  
multiple query engines  
and storage tiers.

\*When comparing published 2023 list prices normalized for VPC hours of IBM watsonx.data to several major cloud data warehouse vendors. Savings may vary depending on configurations, workloads and vendors.

# watsonx.governance AI ethics and governance

## **Explainability and interpretability**

- Explainable AI is a set of processes and methods that enables human users to interpret, comprehend and trust the results and output created by algorithms

## **Fairness and inclusion**

- Machine learning, inherently a type of statistical discrimination, becomes problematic when it consistently benefits privileged groups while disadvantaging unprivileged ones.

## **Robustness and security**

- Robust AI manages unusual conditions, like abnormal inputs or malicious attacks, without causing unintended harm.

## **Accountability and transparency**

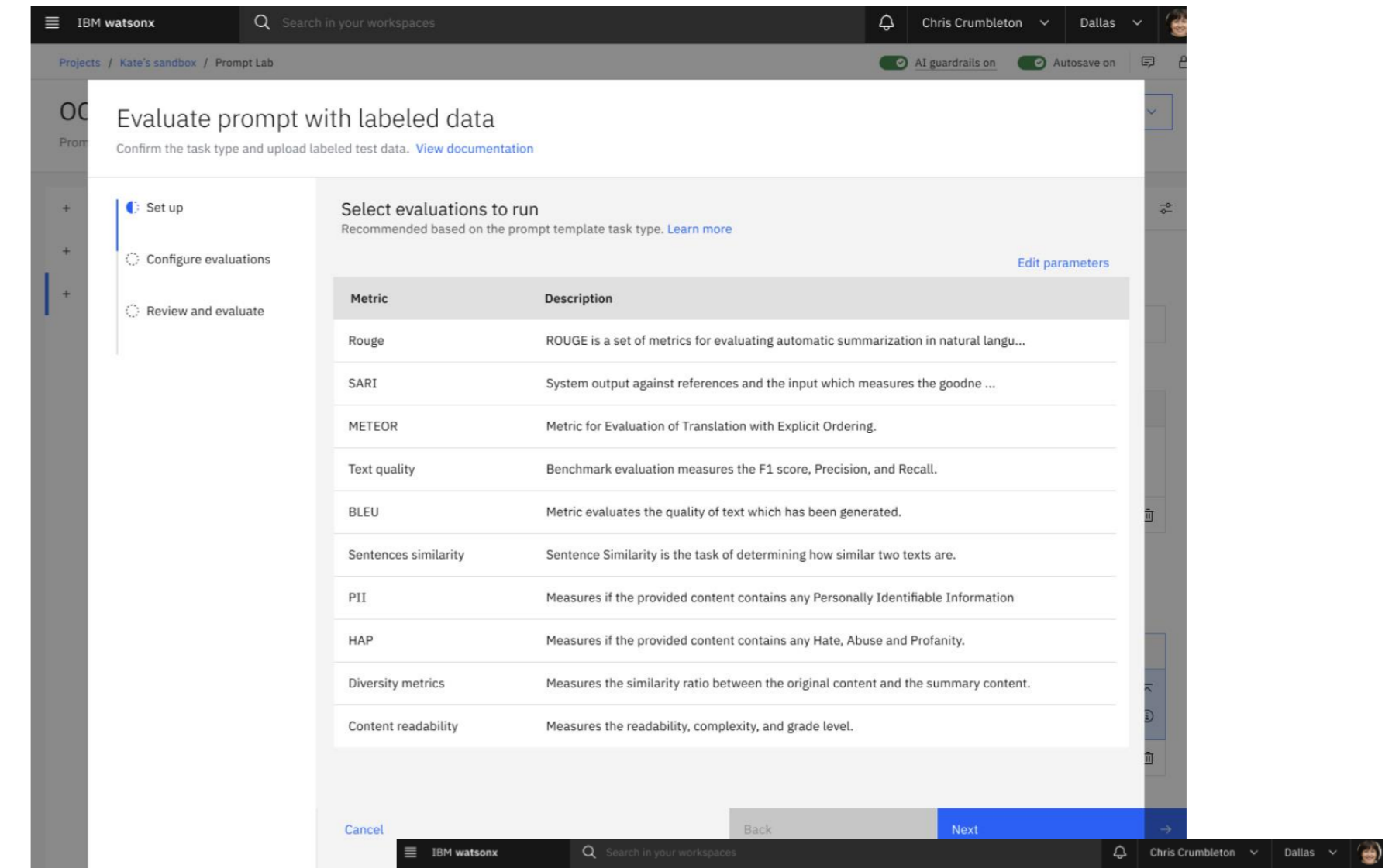
- Organizations need clear responsibilities and governance for AI development, deployment, and outcomes.

## **Privacy and compliance**

- It's essential to protect AI models containing personal information, control the data used, and create flexible systems that can adapt to evolving regulations and ethical standards in AI.

# Monitoring of new LLM Metrics

- Users will be able to 'Evaluate' their Prompt template for LLMs in watsonx.ai through the Prompt Lab within their Project itself during pre-production, as well as for continuous monitoring in production
- Users can evaluate their prompts for various metrics for use cases like Text Summarization, Text Classification, Language Translation, Content generation and Q&A.
- Some of the metrics planned to be supported out-of-the-box for LLMs in watsonx.ai are:
  - HAP, PII Detection, Stigma/Social bias, Faithfulness
  - Text Summarization metrics – ROUGE, SARI, sacrebleu, BLEURT, METEOR, SUPER GLUE
  - Text Classification metrics – Accuracy, Precision, Recall, ROC AUC, F1, Brier Score, GLUE metrics, Matthew's Correlation coefficient, Label Skew
  - Entity Extraction – Seq Eval
  - Content Generation, Q&A Evaluation metrics – BLEU, exact\_match
  - Other metrics – Sentence similarity: Jaccard & cosine, div\_metrics, flesch



The screenshot shows the 'OCCS Prompt' evaluation results in the IBM watsonx interface. The screen displays a summary of the evaluation, including the test data set, task type, and a table of metrics and scores. The 'Next' button is highlighted in blue.

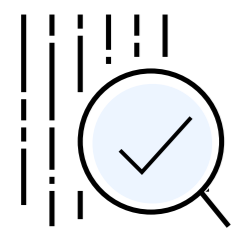
Metric	Score	Violation
ROUGE-1	0.83	None
ROUGE-2	0.5	None
ROUGE-L	0.83	None
ROUGE-SUM	0.92	None
SARI	80	None
METEOR	0.93	None
Normalized F1	0.97	None
Normalized precision	0.42	None
Normalized recall	0.93	None
BLEU	0.83	None
Jaccard similarity	0.93	None
Cosine similarity	0.96	None
PII	0.96	None
HAP	0.96	None
Diversity score	0.96	None
Content readability	90 - very easy to read	None

# IBM watsonx.ai Foundation Model Library

Model variety to cover different enterprise use cases and compliance requirements

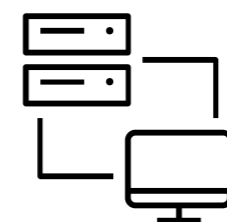
## IBM models

IBM's suite of foundation models is designed to ensure model trust and efficiency in business applications. The suite of models features:



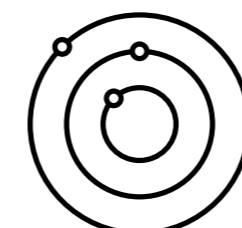
### Transparent pre-training on IBM's trusted Data Lake

- One of the largest repositories of enterprise-relevant training data
- Verified legal and safety reviews by IBM
- Full, auditable data lineage available for any IBM Model



### Compute-Optimal Model Training and Architectures

- Granite Decoder only transformers
- Granite code model
- Granite multi-lingual model



### Efficient domain and task specialization

Models Coming Soon:

- Finance
- Cybersecurity
- Legal, etc.

## Open source models

Experiment with open source models



IBM and Hugging Face partnership demonstrates the shared *commitment to delivering to clients an open ecosystem approach* that allows them to define the best models for their business needs.

## OEM partnership



- OEM partnership with Mistral brings best in class models (such as the mixtral-8x7b-instruct-v01 and the merlinite-7b models).

## Bring-your-own-model

- Clients can bring models that they have been working with to leverage the capabilities of watsonx.ai to develop and deploy applications

# IBM watsonx.ai: Synthetic Data Generator

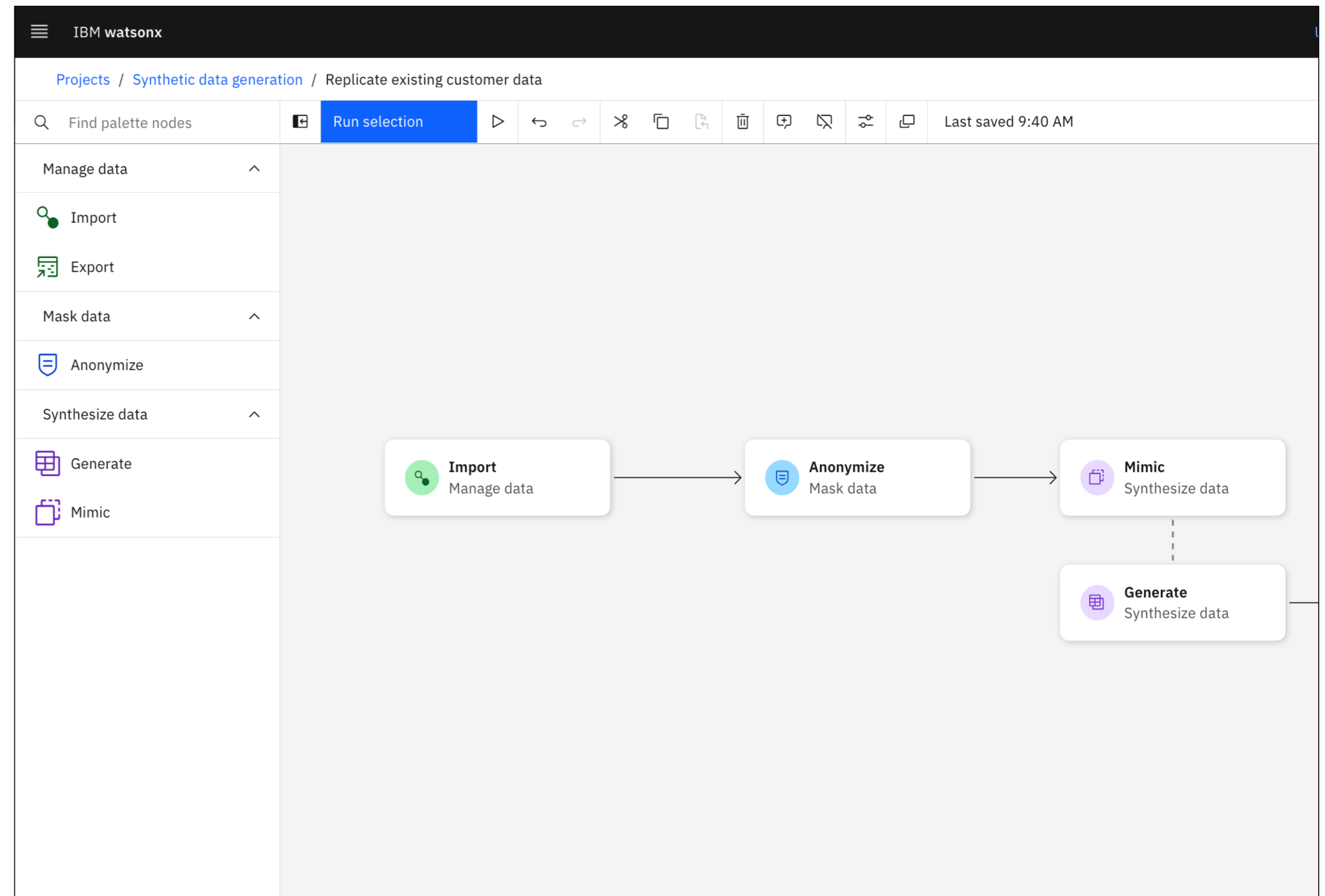
Generate synthetic tabular data to address your data gaps

## Create synthetic data at scale

- Unlock your valuable insights by using synthetic data. Create synthetic data by:
  - Using your existing data in a database or by uploading a file.
  - Designing your own data schema.

## Select your model and privacy needs

- Clients can select from multiple IBM models\* to create their synthetic tabular data.
- When using existing data, IBM models apply differential privacy to minimize your privacy risk and give you control over the level of privacy protection that is required for your organization.



# The generative AI stack



AI assistants

watsonx Assistant  
watsonx Code Assistant  
watsonx Discovery  
watsonx Orchestrate  
+++

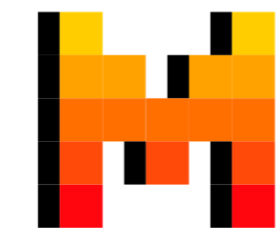
Gen AI platform

watsonx.ai  
wastonx.governance  
watsonx.data

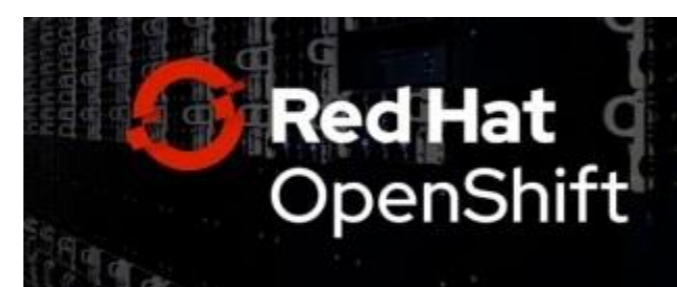


Foundation models

granite-7b  
granite-13b  
granite-20b  
more ...



Infrastructure



HYBRID CLOUD

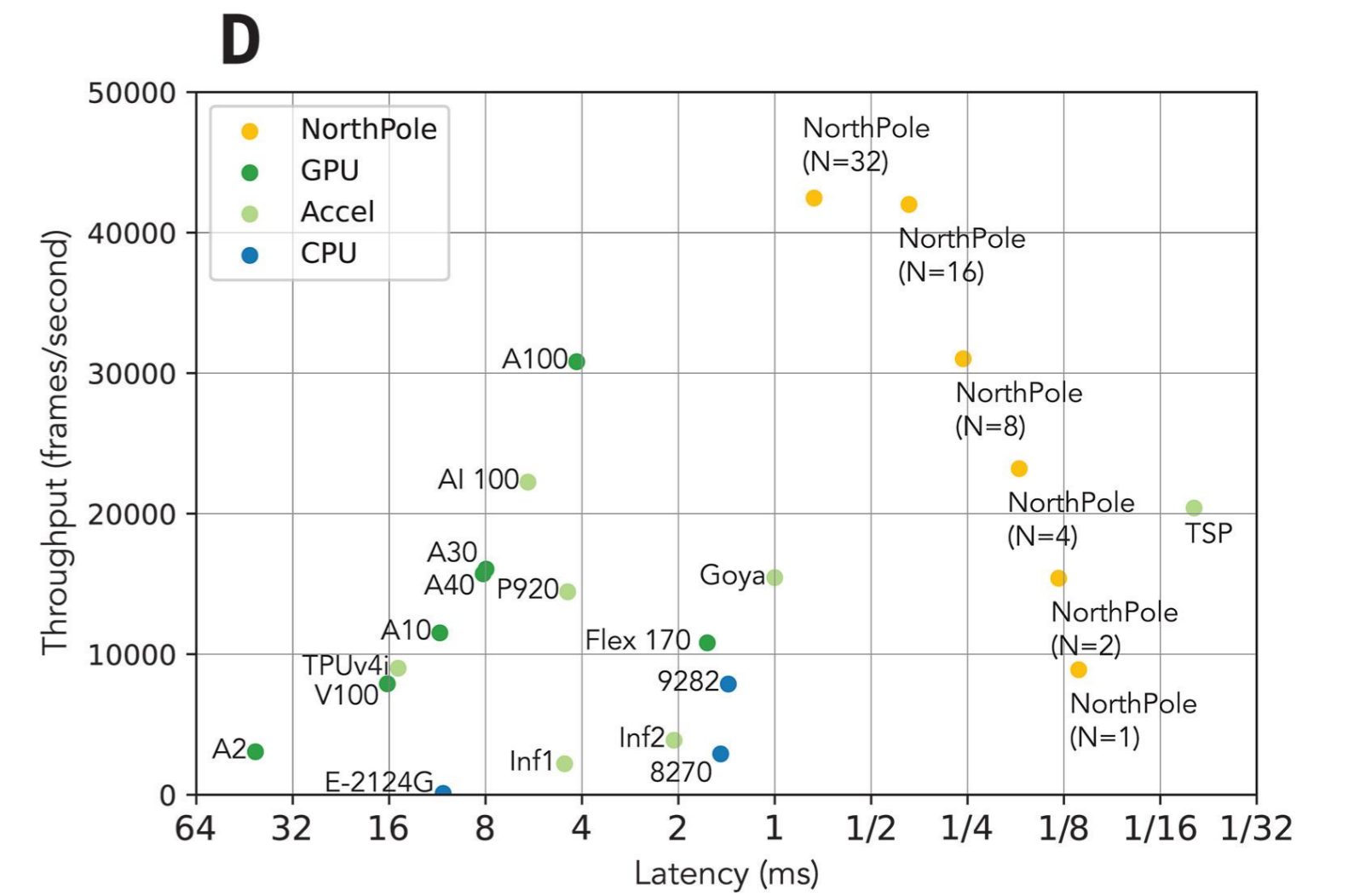
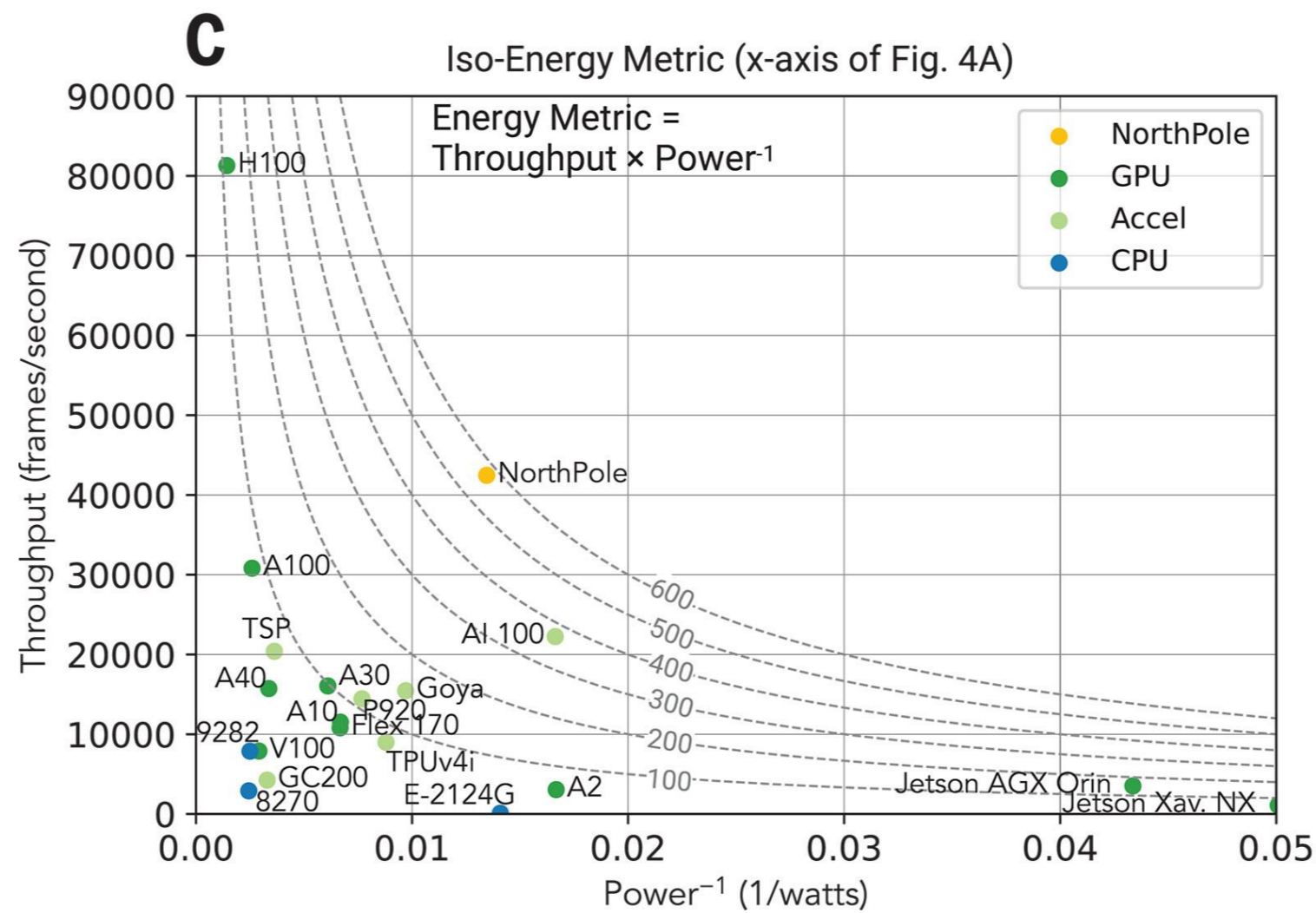
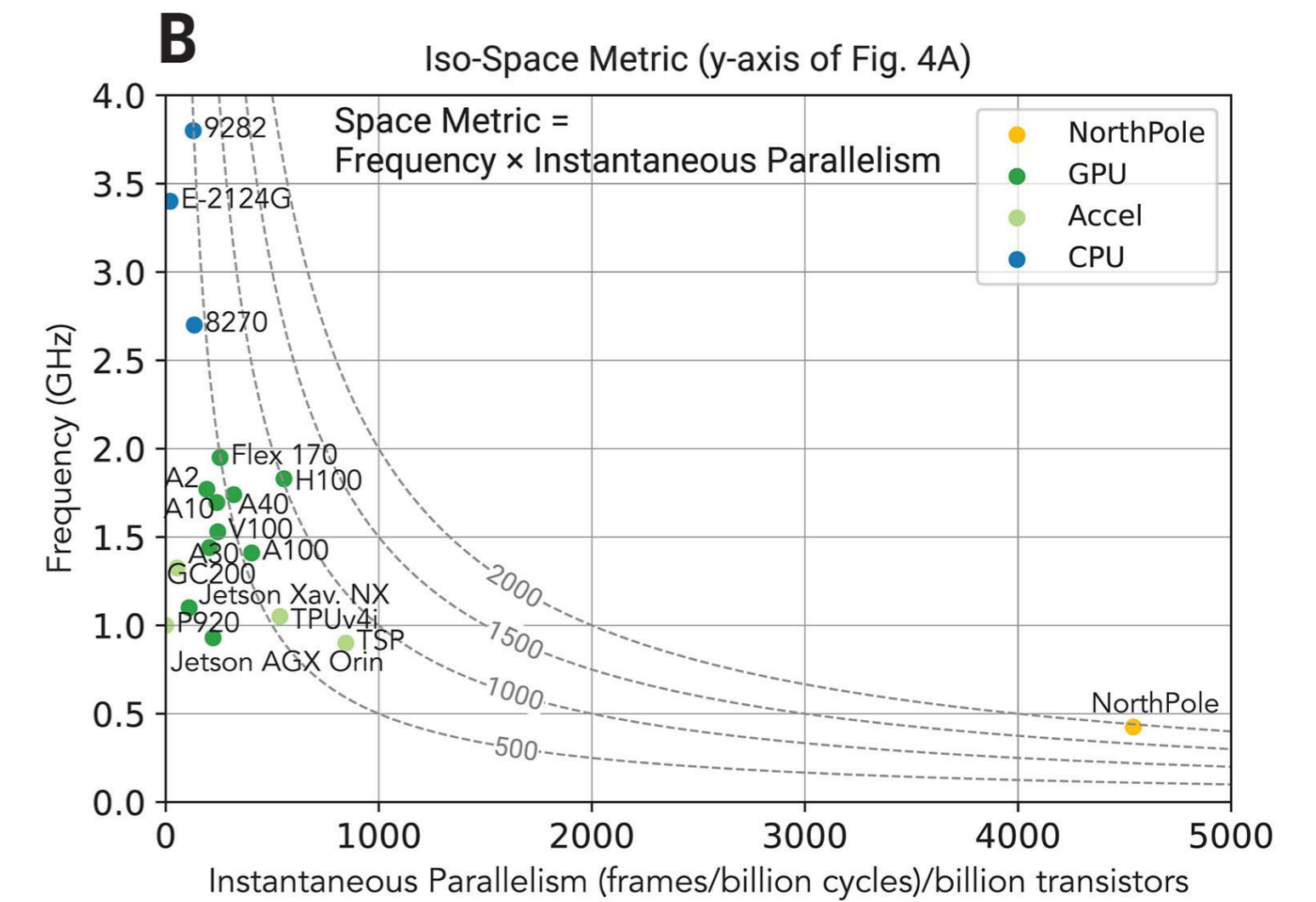
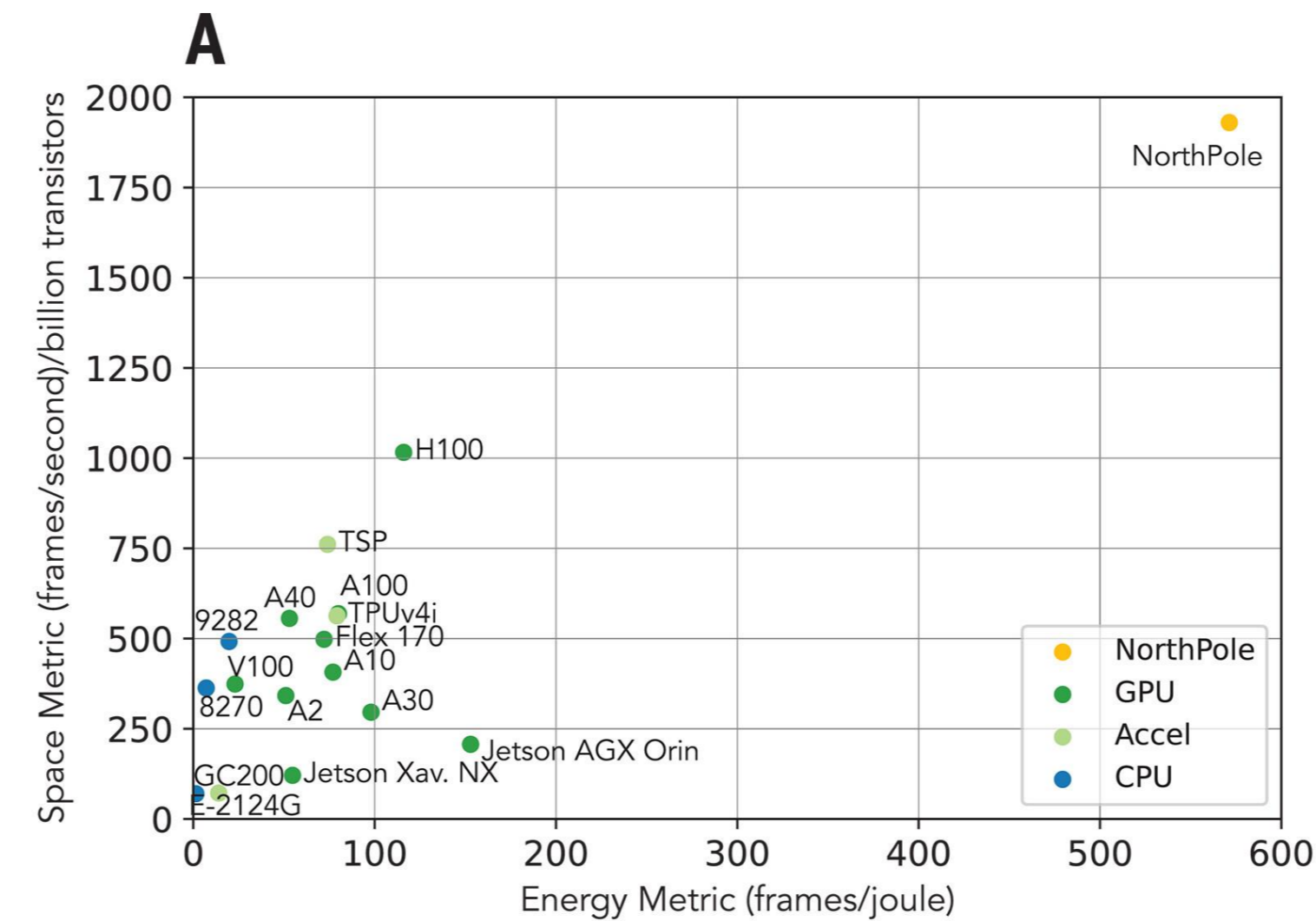


# IBM Research's lab NorthPole AI Chip prototype

A new chip architecture points to faster, more energy-efficient AI



<https://research.ibm.com/blog/northpole-ibm-ai-chip>



Source: Neural inference at the frontier of energy, space, and time - <https://www.science.org/doi/full/10.1126/science.adh1174>

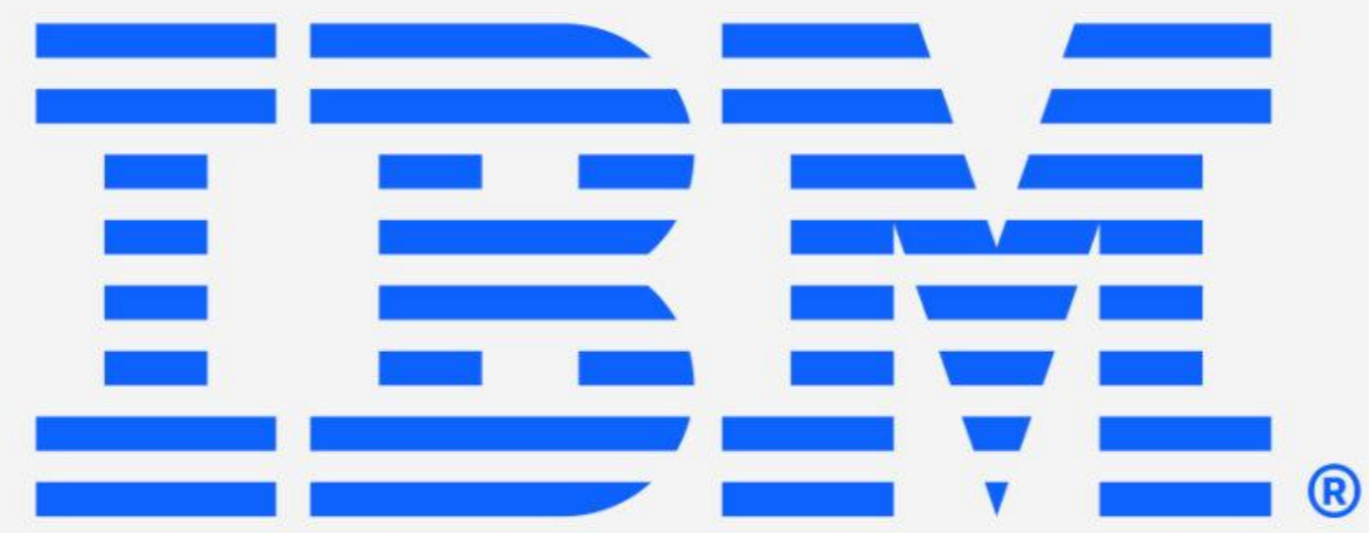
# Additional resources

## IBM articles

- [The games that helped AI evolve](#)
- [IBM Watson: The inside story of how the Jeopardy-winning supercomputer was born, and what it wants to do next](#)
- [What is artificial intelligence \(AI\)?](#)
- [Meet watsonx](#)
- [Foundation Models in watsonx.ai](#)
- [IBM Granite LLMs](#)

## Industry articles

- [Technology Modernization Fund \(Call for AI Proposals\)](#)
- [NASA and IBM Openly Release Geospatial AI Foundation Model for NASA Earth Observation Data](#)
- [IBM and KONE - Watson IoT Gives Lift To Innovation In Smart Buildings](#)



Let's  
create |