

watsonx

workshop

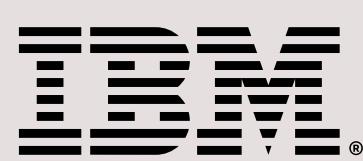
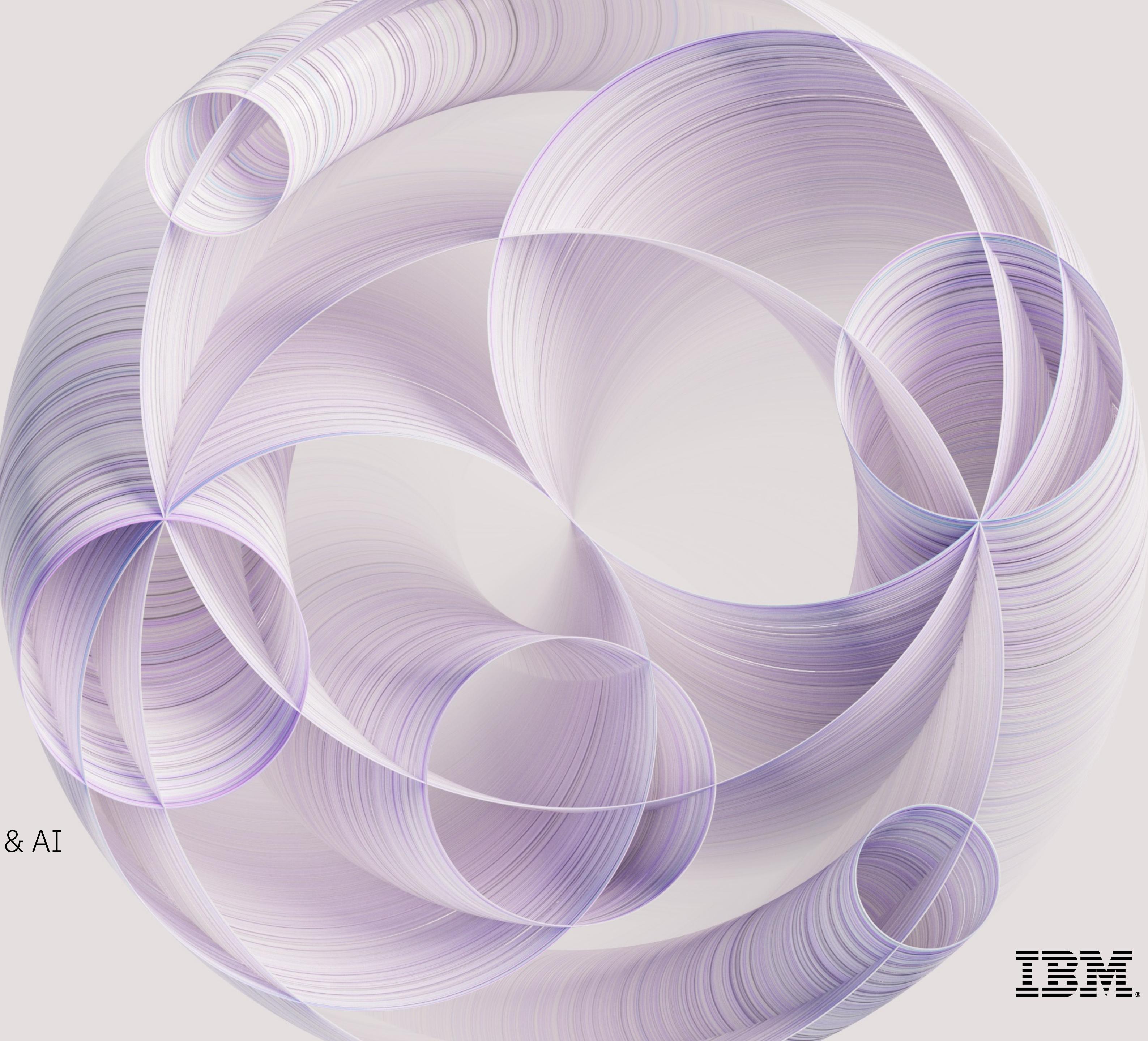
—
October 2nd, 2023

Prepared for:



Institute for
Data Science & Informatics
University of Missouri

— Ryan Kather
Sr. AI Engineer - Client Engineering, IBM Data & AI
e: ryan.Kather@ibm.com



Agenda

- The impact of generative AI:
The opportunity
- AI Development: From +AI to AI+
- Enterprise considerations
- What IBM offers:
Open, trusted, targeted,
empowering AI solutions
- Getting started

The speed, scope,
and scale of
generative AI
impact is
unprecedented

Massive early adoption

80%

of enterprises are working
with or planning to
leverage foundation models
and adopt generative AI

Broad-reaching and deep impact

Generative AI could raise
global GDP by

7% within 10 years

Critical focus of AI activity and investment

Generative AI expected
to represent

30%

of overall market by 2025

The impact of generative AI

The most common generative AI tasks implemented today

Retrieval-Augmented Generation

Based on a documents or dynamic content, create a chatbot or question-answering feature.

Building a Q&A resource from a broad knowledge base, providing customer service assistance

Summarization

Transform text with domain-specific content into personalized overviews that capture key points.

Conversation summaries, insurance coverage, meeting transcripts, contract information

Content Generation

Generate text content for a specific purpose.

Marketing campaigns, job descriptions, blog posts and articles, email drafting support

Named Entity Recognition

Identify and extract essential information from unstructured text.

Audit acceleration, SEC 10K fact extraction

Insight Extraction

Analyze existing unstructured text content to surface insights in specialized domain areas.

Medical diagnosis support, user research findings

Classification

Read and classify written input with as few as zero examples.

Sorting of customer complaints, threat and vulnerability classification, sentiment analysis, customer segmentation

The impact of generative AI
| +AI to AI+ |

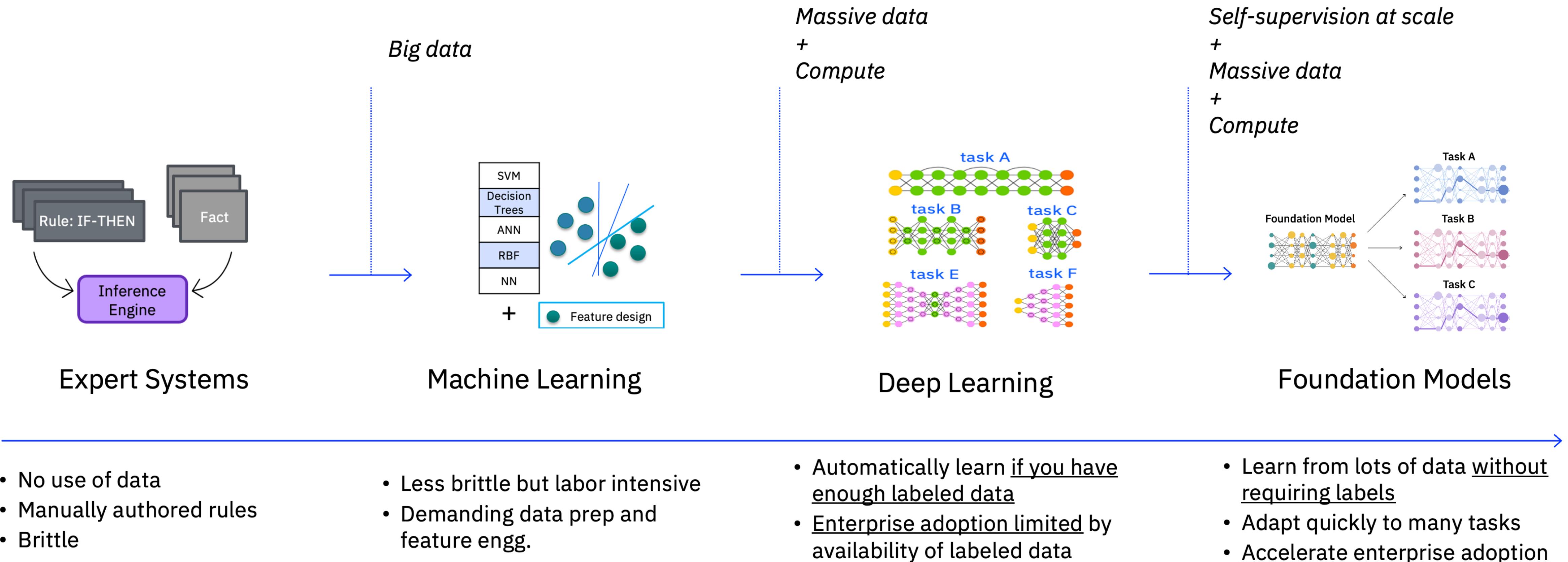
IBM is actively engaging with enterprise clients across a broad set of business domains

NON-EXHAUSTIVE

Customer-facing functions and experiences	HR, Finance, and Supply Chain functions	IT development and operations	Core business operations
Customer service Empower customers to find solutions with easy, compelling experiences. Automate answers with 95% accuracy	HR automation Reduce manual work and automate recruiting, sourcing and nurturing job candidates. Reduce employee mobility processing time by 50%	App modernization, migration Generate code, tune code generation response in real time. Deliver faster development output	Threat management Reduce incident response times from hours to minutes or seconds. Contain potential threats 8x faster
Marketing Increase personalization, improve efficiency across the content supply chain. Reduce content creation costs by up to 40%	Supply chain Automate source to pay processes, reduce resource needs and improve cycle times. Reduce cost per invoice by up to 50%	IT automation Identify deployment issues, avoiding incidents, optimize application demand to supply. Reduce mean time to repair (MTTR) by 50%+	Asset management Optimize critical asset performance and operations while delivering sustainable outcomes. Reduce unplanned downtime by 43%
Content creation Ex. Enhance digital sports viewing with auto-generated spoken AI commentary. Scale live viewing experiences cost effectively	Planning and analysis Make smarter decisions, focus on higher value tasks with automated workflows and AI. Process planning data up to 80% faster	AIOps Assure continuous, cost-effective performance and connectivity across applications. Reduce application support tickets by 70%	Product development Ex. Expedite drug discovery by inferring structure with AI from simple molecular representations. Faster and less expensive drug discovery
Knowledge worker Enable higher value work, improve decision making, and increase productivity. Reduce 90% of text reading and analysis work	Regulatory compliance Support compliance based on requirements / risks, proactively respond to regulatory changes. Reduce time spent responding to issues	Data platform engineering Redesign the approach for data integration using generative AI. Reduce data integration time by 30%+	Environmental intelligence Provide intelligence to proactively plan and manage impact of severe weather and climate. Increase manufacturing output by 25%

The impact of generative AI | The opportunity

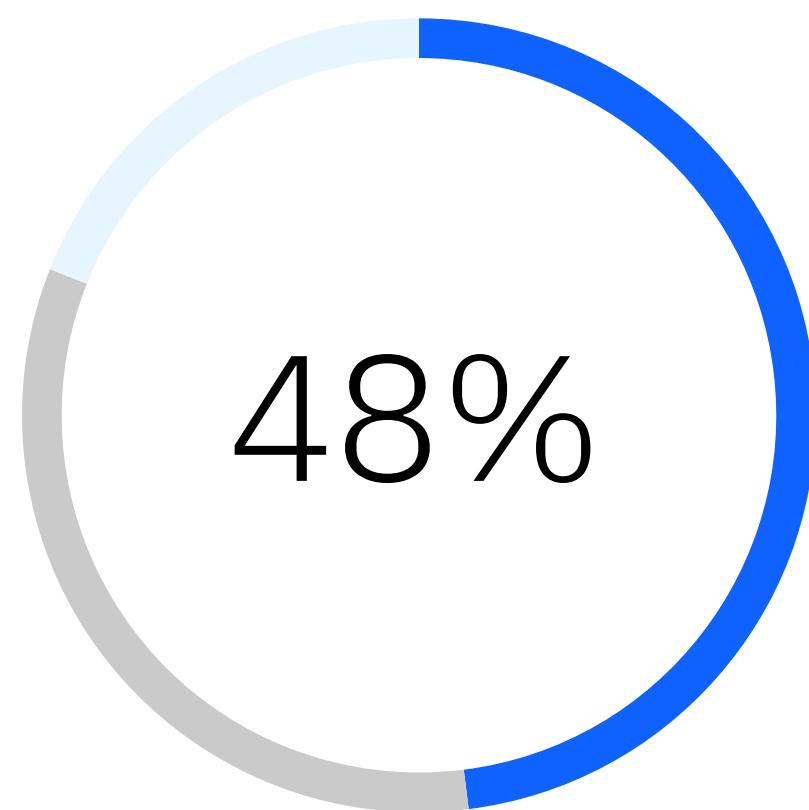
Foundation models can help accelerate enterprise AI adoption



Generative AI adoption considerations, inhibitors and fears

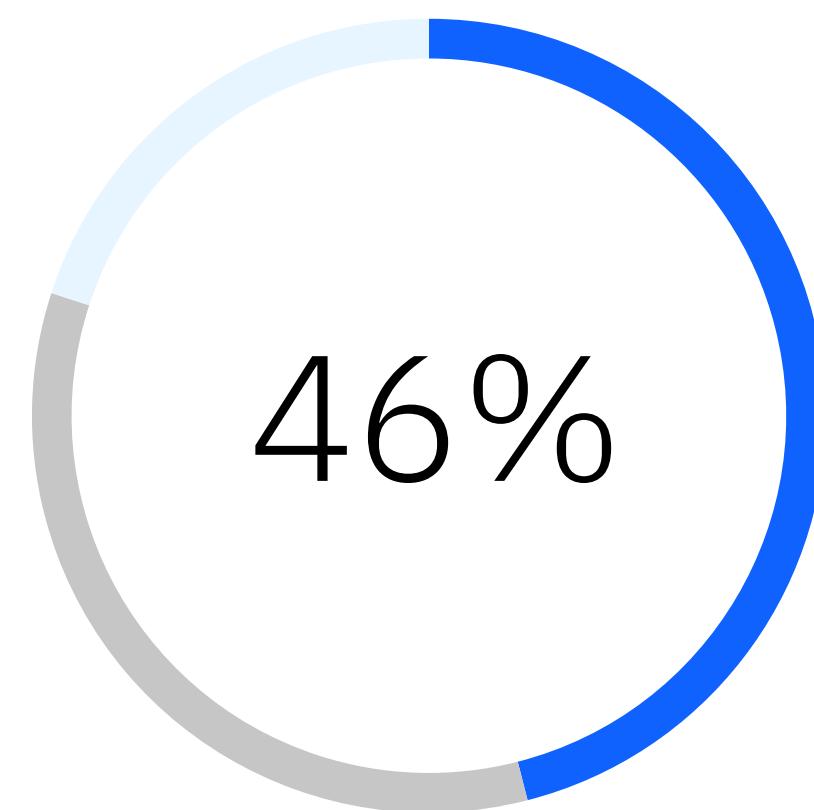
80% of business leaders see at least one of these ethical issues as a major concern

Explainability



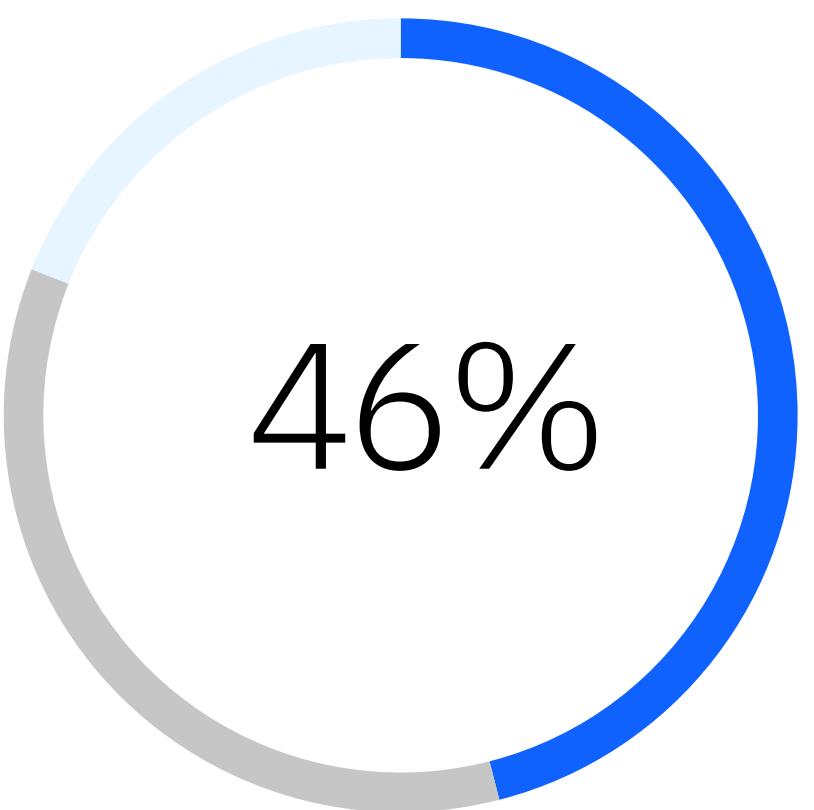
Believe decisions made by Generative AI are not sufficiently **explainable**.

Ethics



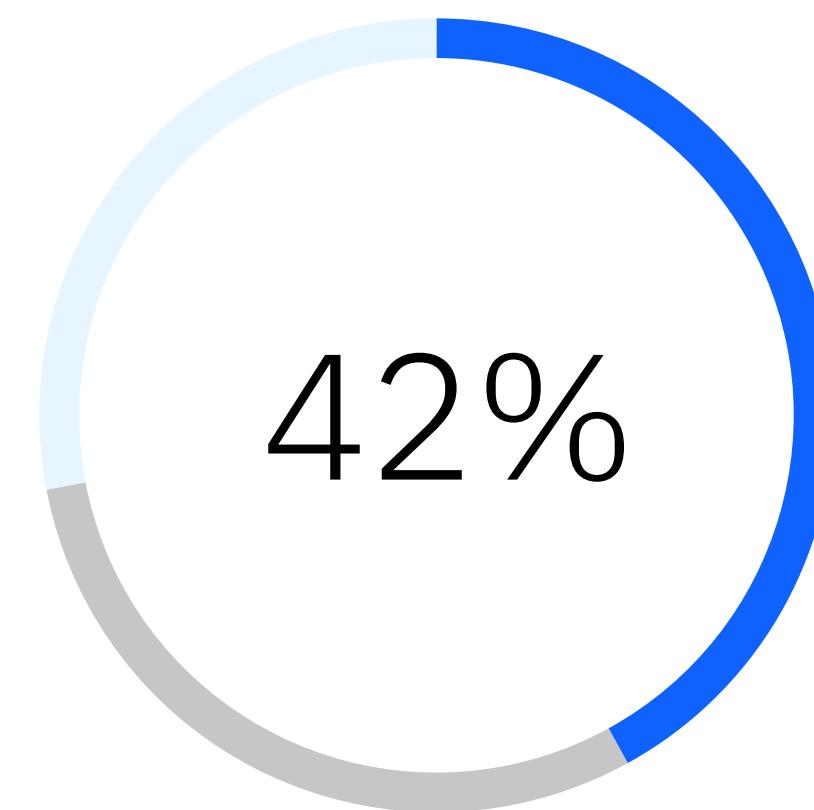
Concerned about the safety and **ethical** aspects of Generative AI.

Bias



Believe that Generative AI will propagate established **biases**.

Trust



Believe Generative AI cannot be **trusted**.

Simpler Times

Data Science is now a team sport.

Data and ML landscapes have evolved.

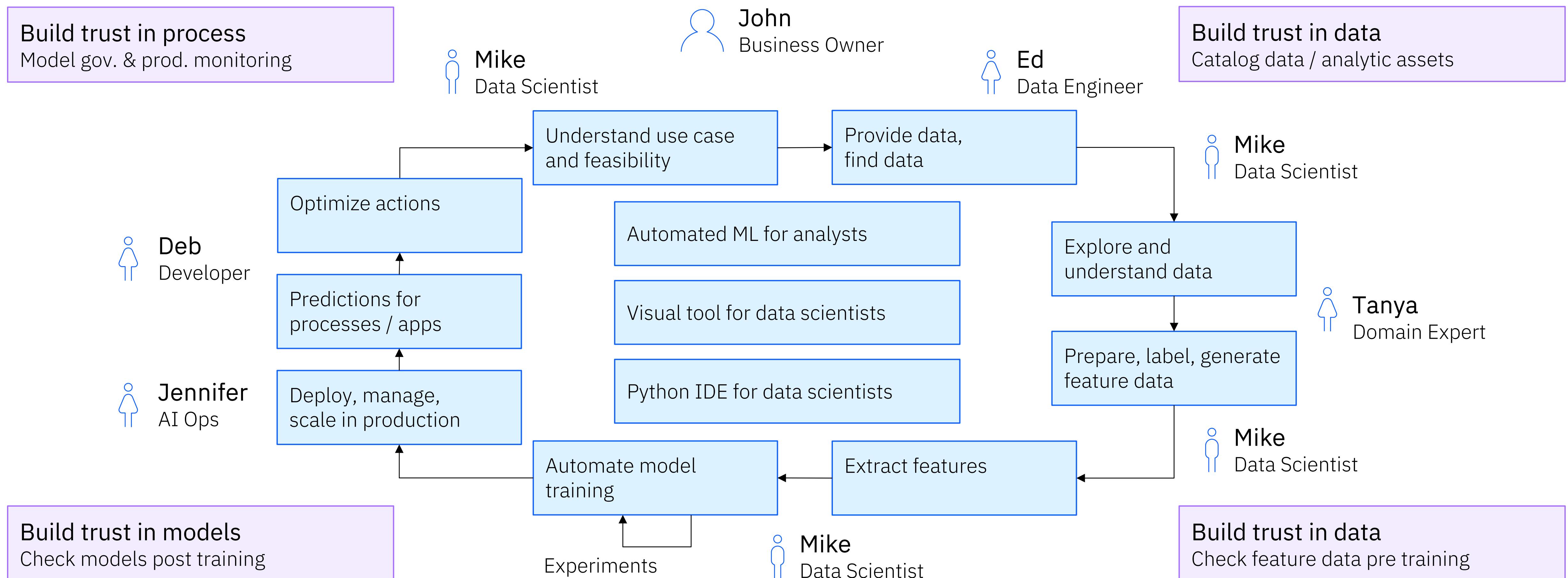
Factors to Account

- Repeatability
- Accountability
- Transparency
- Explainability
- Risk Exposure
- Robustness
- Performance



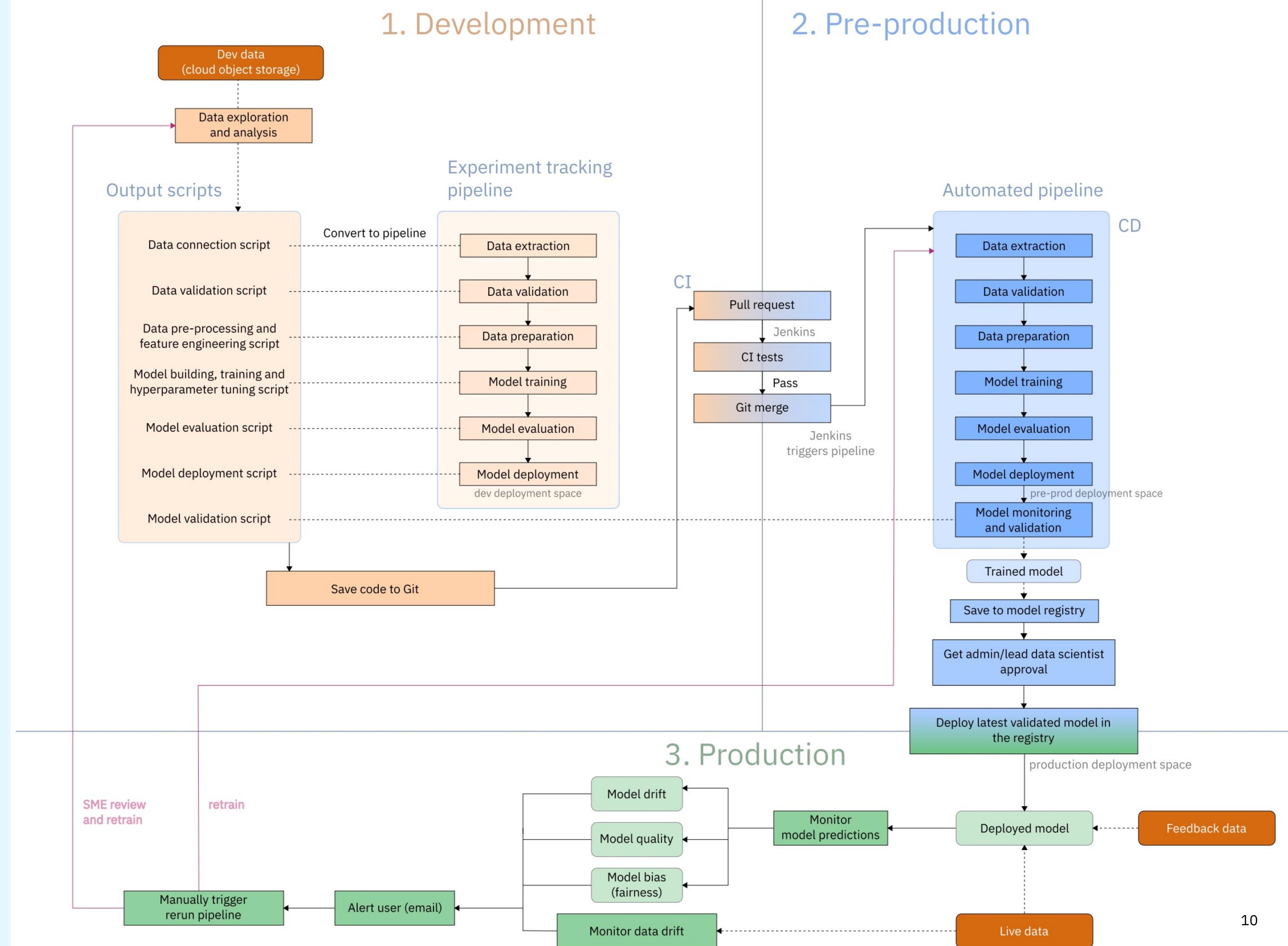
DS Process

CRISP-DM NIST Methodology

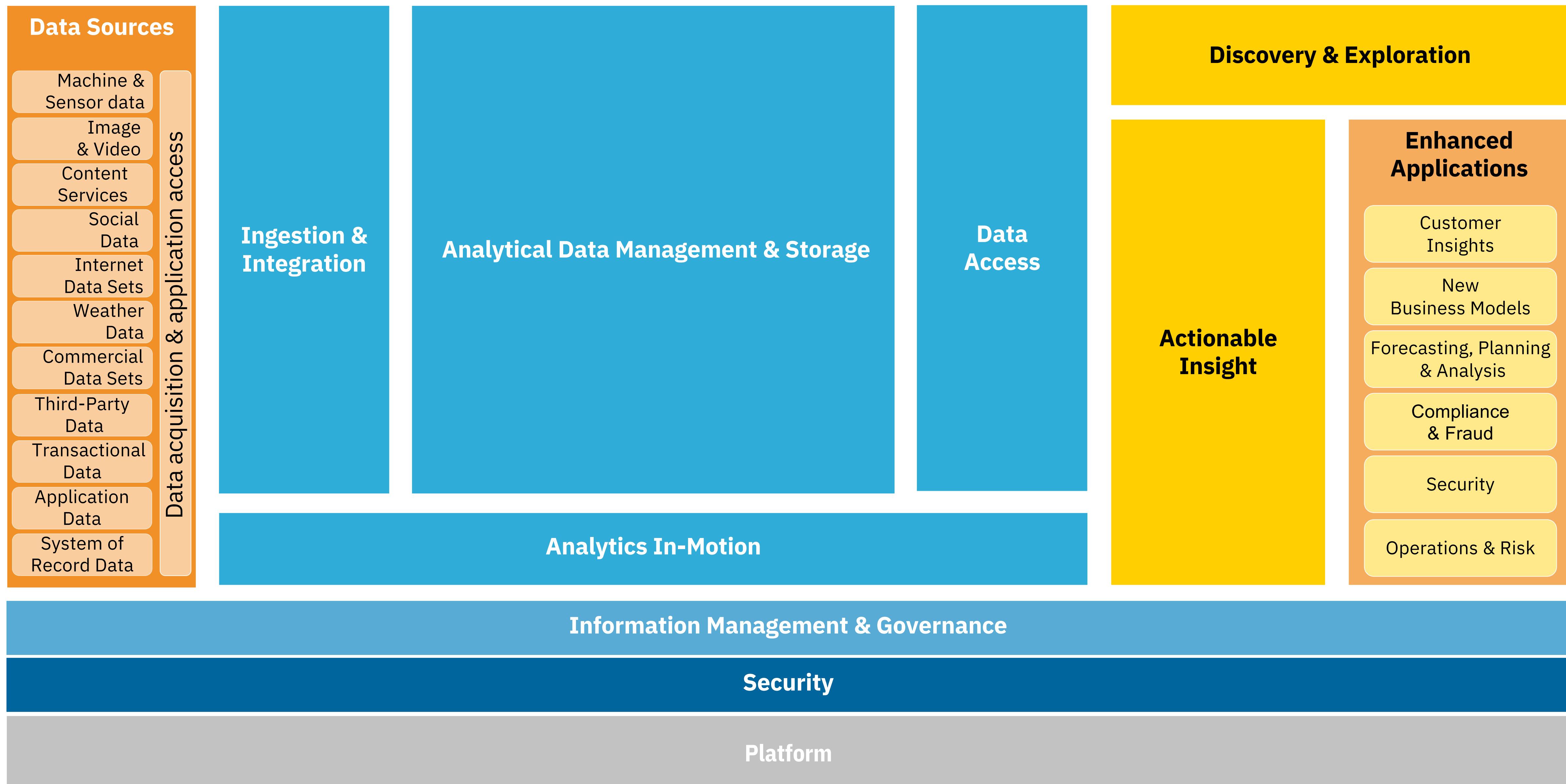


MLOps

Modern Watson Studio
MLOps Workflow from
Dev to Prod with CI-CD



Data & Analytics Reference Architecture



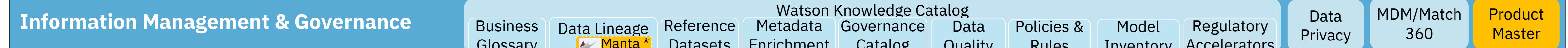
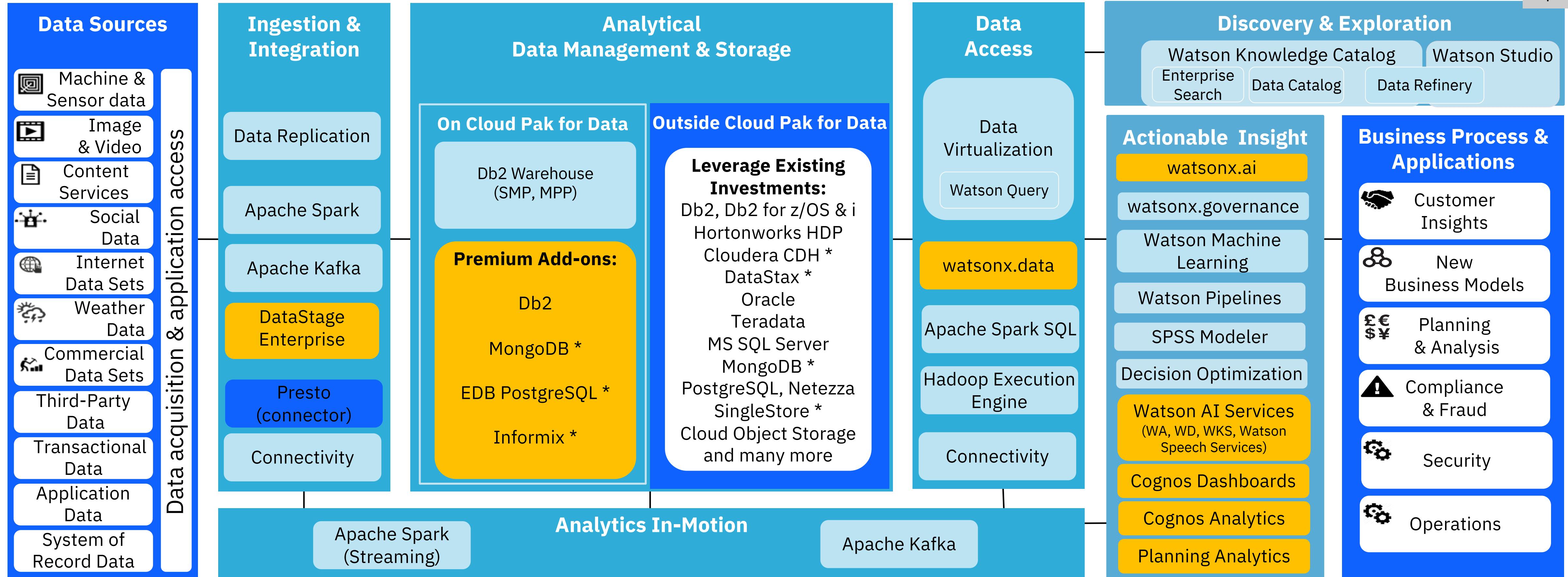
Reference Architecture – Product View

IBM Cloud Pak for Data (Base)

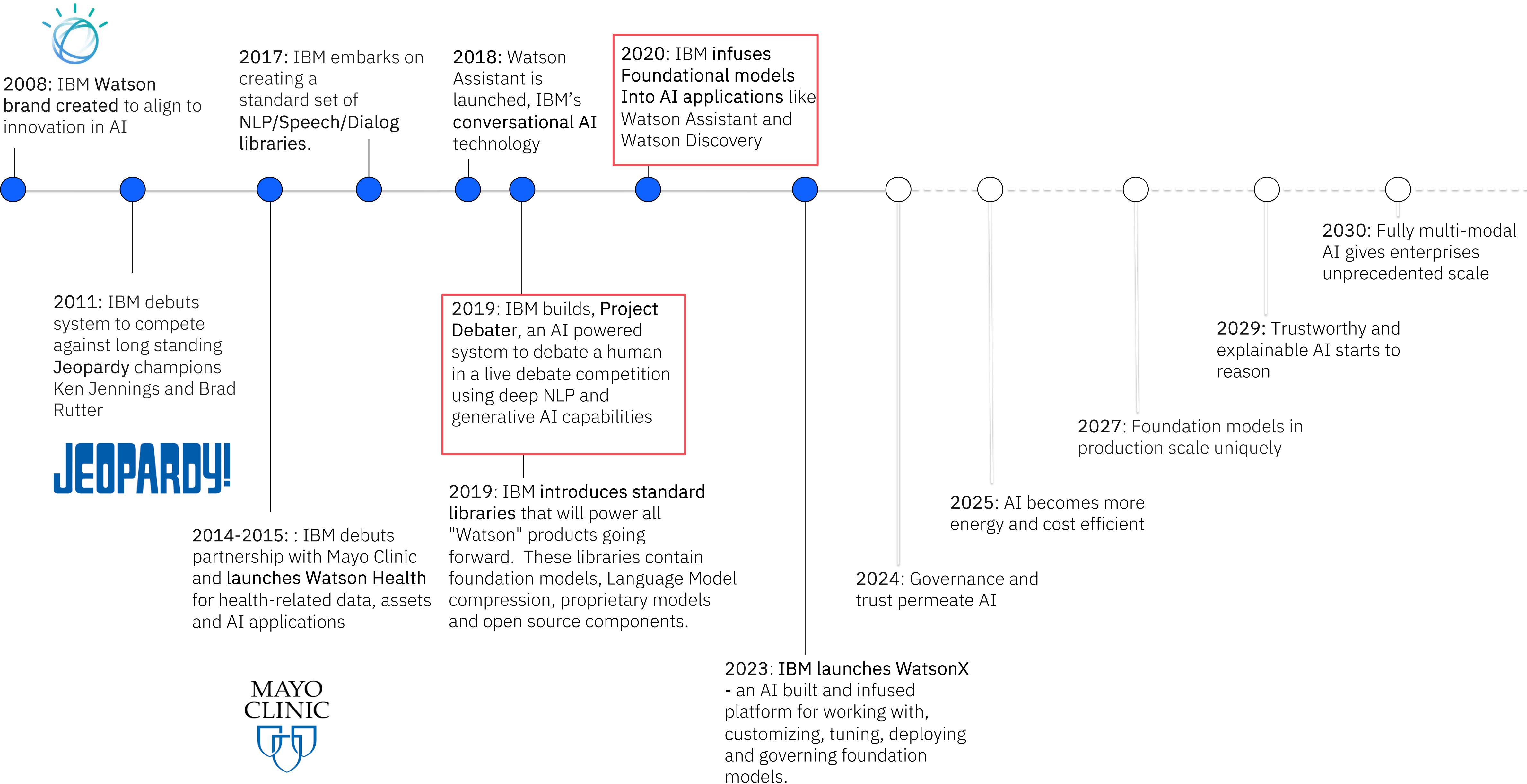
IBM Cloud Pak for Data – Premium Add-ons

Customer Investments Outside Cloud Pak for Data

* partner



IBM's Evolution and Leadership in AI:



Generative AI must be tailored to the enterprise (TOTE)

Trusted

- Offering security and data protection.
- Built with governance, transparency, and ethics that support increasing regulatory compliance demands.

Open

- Based on the best open technologies available.
- Giving access to the innovation of the open community and multiple models.

Targeted

- Designed for targeted for business use cases, that unlock new value.
- Including models that can be tuned to your proprietary data.

Empowering

- On a platform to bring your own data and AI models that you tune, train, deploy, and govern.
- Running anywhere, designed for scale and widespread adoption.

The platform
for AI and data

watsonX

Scale and
accelerate the
impact of AI with
trusted data.

watsonX.ai

Train, validate, tune and
deploy AI models

A next generation enterprise studio
for AI builders to train, validate, tune,
and deploy both traditional machine
learning and new generative AI
capabilities powered by foundation
models. It enables you to build AI
applications in a fraction of the time
with a fraction of the data.

watsonX.data

Scale AI workloads, for all
your data, anywhere

Fit-for-purpose data store optimized
for governed data and AI workloads,
supported by querying, governance
and open data formats to access and
share data.

watsonX.governance

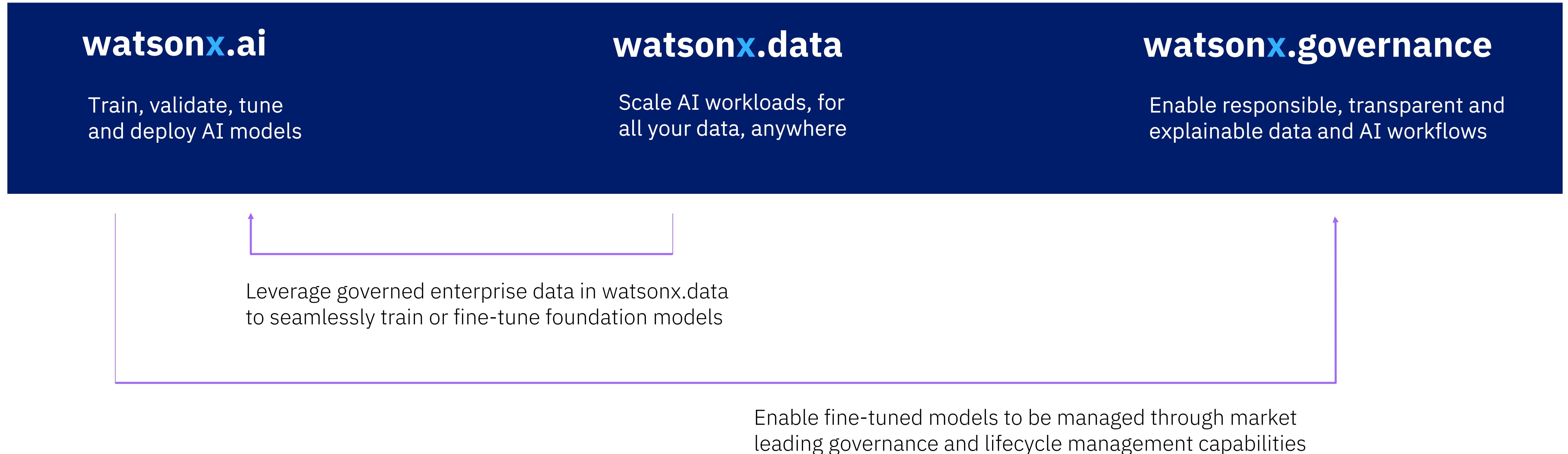
Enable responsible,
transparent and explainable
data and AI workflows

End-to-end toolkit encompassing
both data and AI governance to
enable responsible, transparent, and
explainable AI workflows.

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

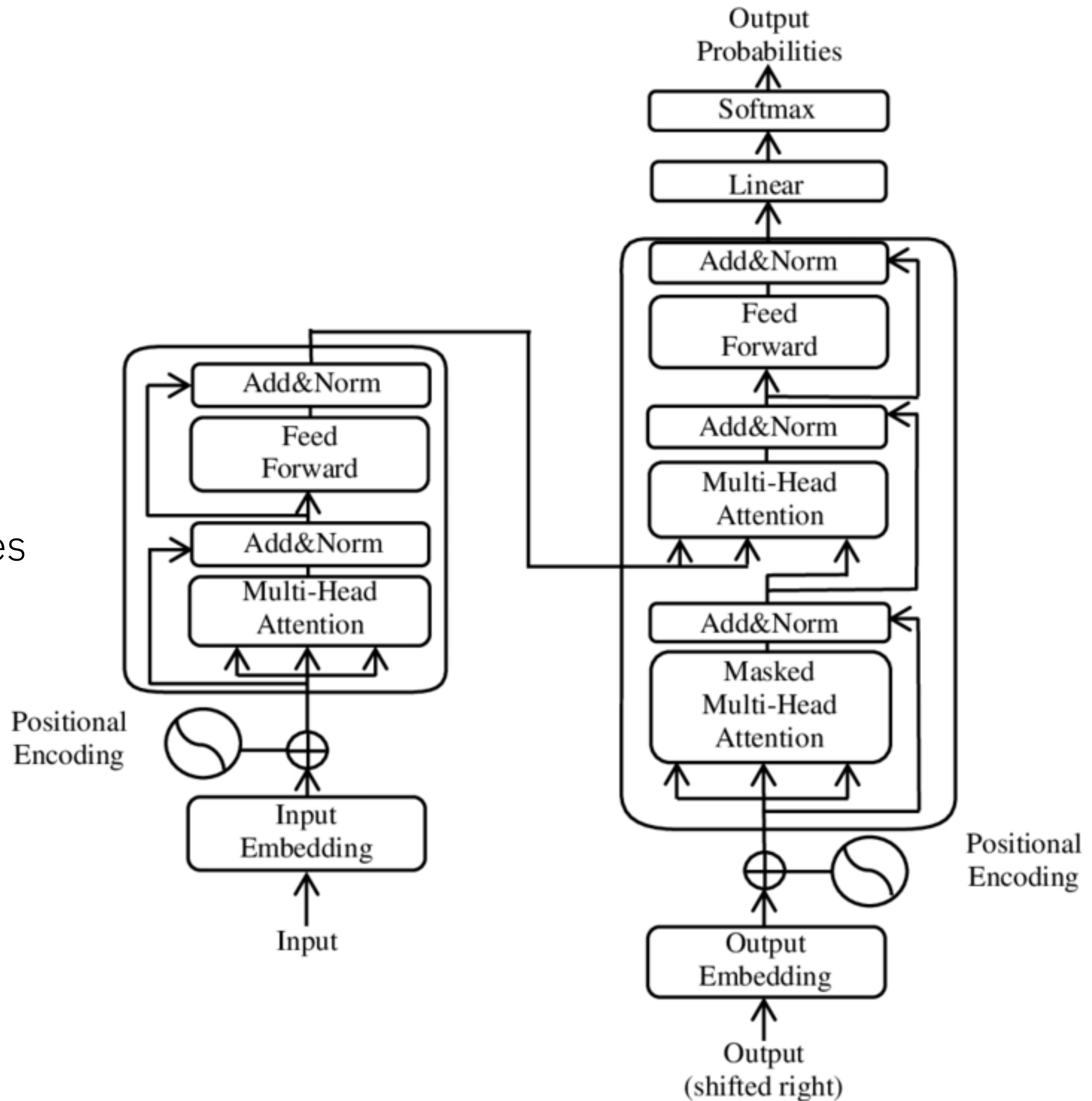


Transformer Model Architecture

- NLP: Transformer Based Architecture

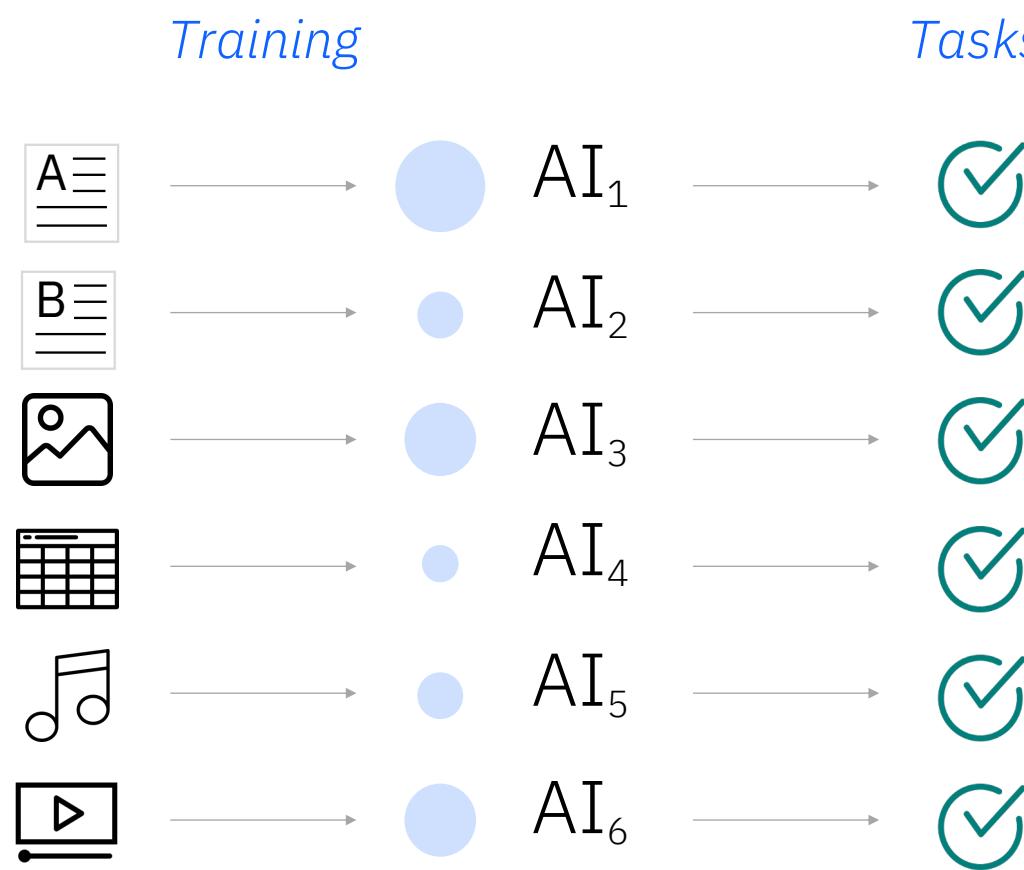
Encoding and Decoding Neural Stages
Operates with Embeddings and Attention

Two Papers:
[“Attention Is All You Need”](#)
[“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”](#)



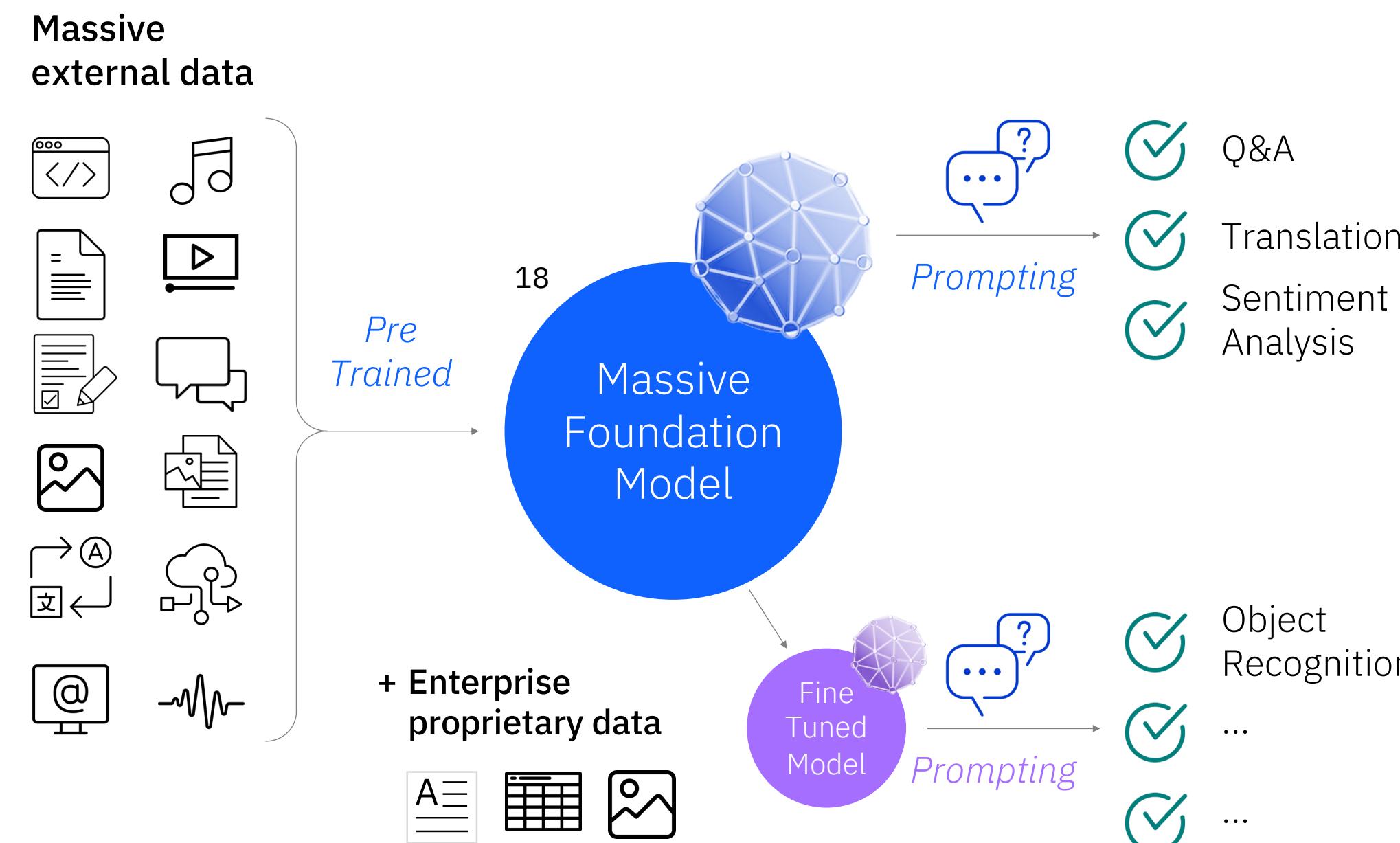
Foundation models establish a new paradigm for AI capabilities

Traditional AI models



- Individual siloed models
- Require task specific training
- Lots of human supervised training

Foundation Models



- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning

Enhanced capabilities

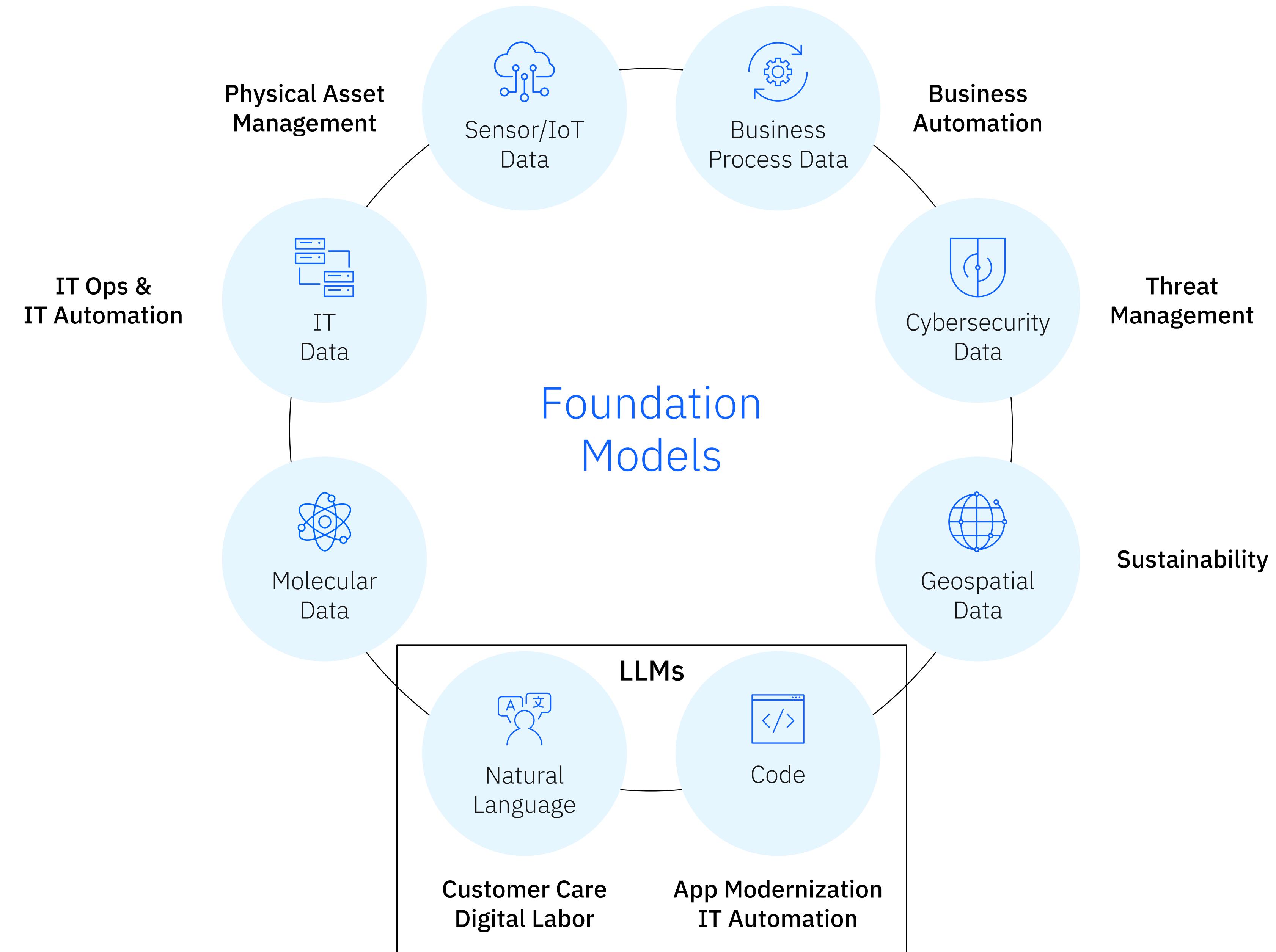
- Summarization
- Conversational Knowledge
- Content Creation
- Code Co-Creation

Key advantages

- Lower upfront costs through less labeling
- Faster deployment through fine tuning and inferencing
- Equal or better accuracy for multiple use cases
- Incremental revenue. through better performance

up to **70% reduction** in certain NLP tasks

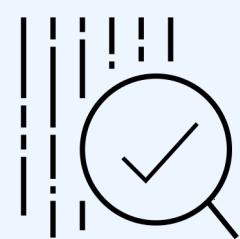
Opportunity to unlock business advantage with foundation models trained across the breadth of enterprise data



Model variety to cover enterprise use cases and compliance requirements

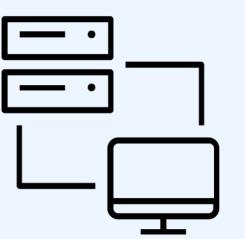
IBM models

IBM's suite of foundation models are designed to ensure model trust and efficiency in business applications. Our suite of models feature:



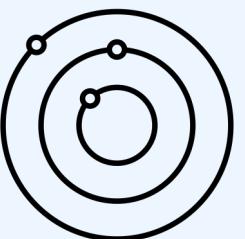
Transparent Pre-Training on IBM's trusted Data Lake

- Largest known repository 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
- Sandstone
Encoder-decoder transformers
- Obsidian (in progress)
Sparse universal transformers



Efficient Domain and Task Specialization

- Models Coming Soon:
- fm.geospatial
 - granite.3b.finance
 - granite.3b.cybersecurity
 - And more! (legal, climate, etc.)

Opensource models

Experiment with opensource models

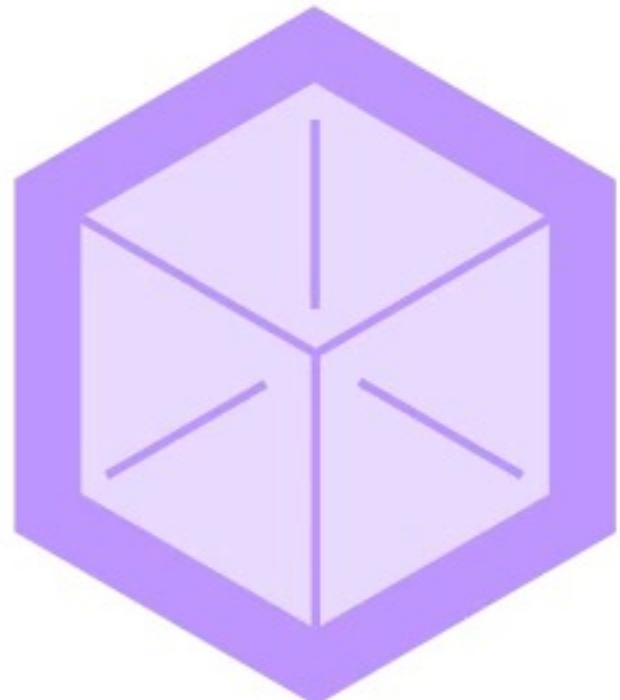
IBM and Hugging Face partnership demonstrates our shared *commitment to delivering to clients an open ecosystem approach* that allows them to define the best models for their business needs.



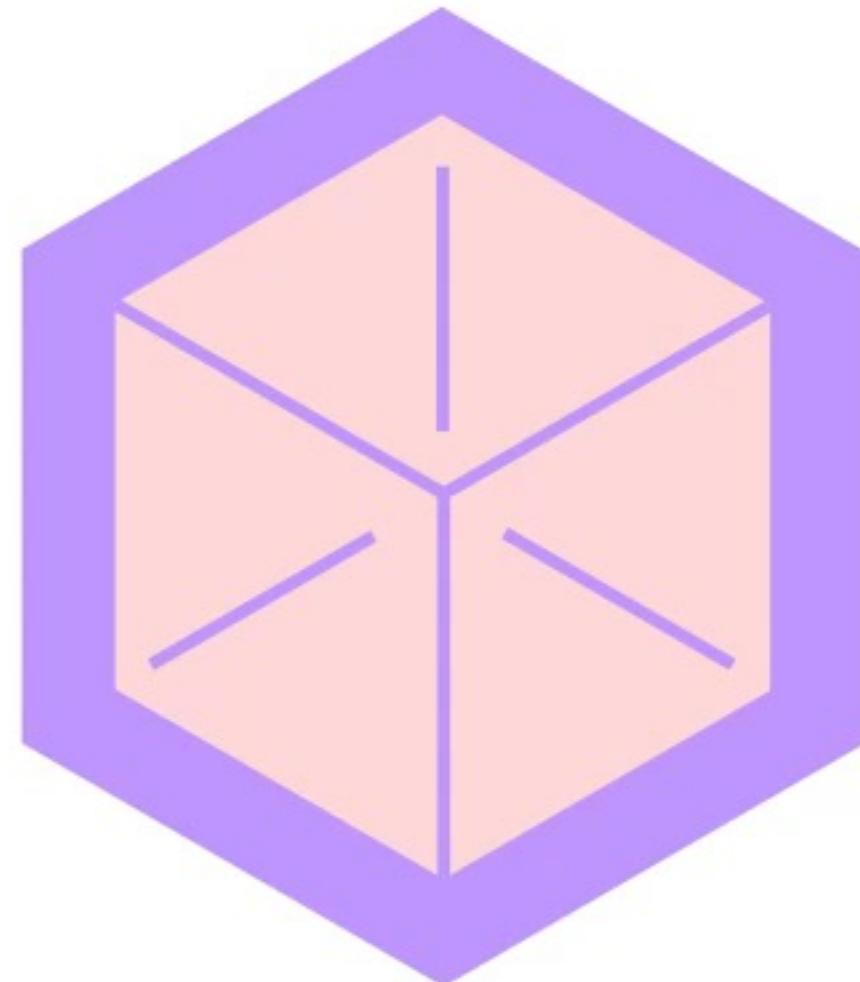
BYO models (optional add on)

Partner with IBM Research to pre-train your own foundation models.

Model architectures



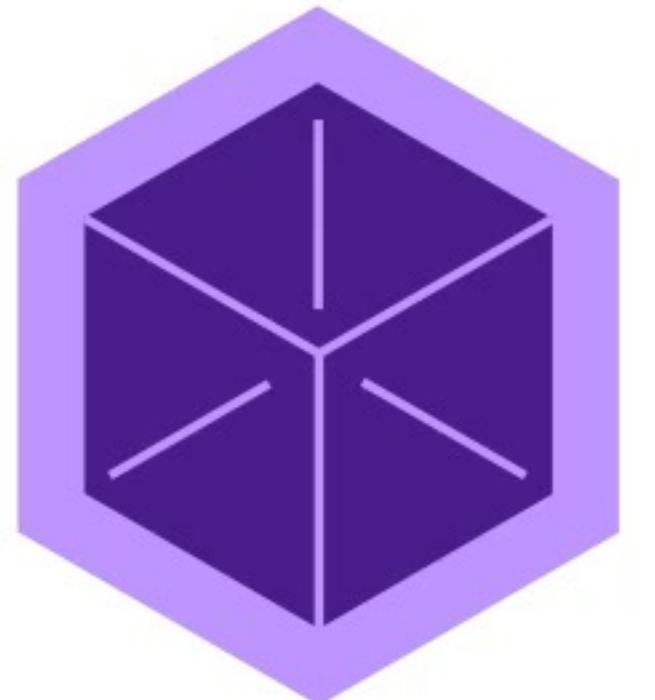
Slate
Non-generative
encoder-only
architecture



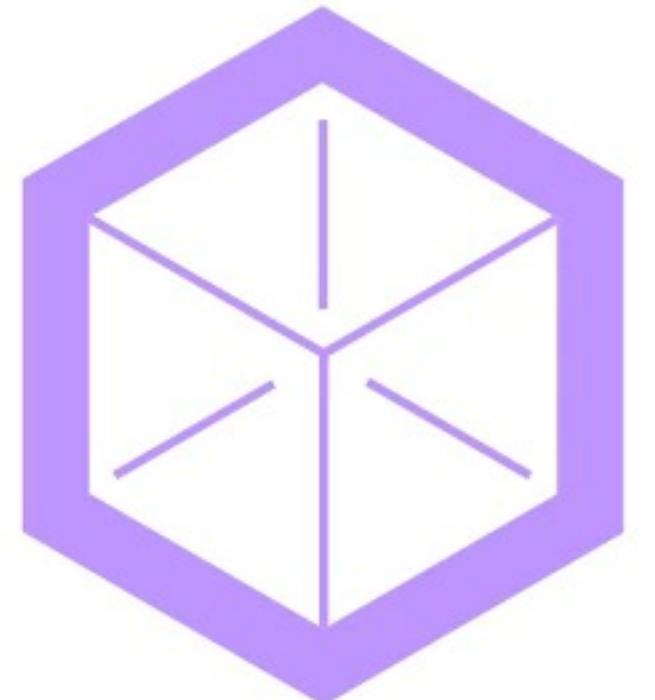
Sandstone
Lightweight
encoder-decoder
architecture



Granite
Decoder-only
architecture



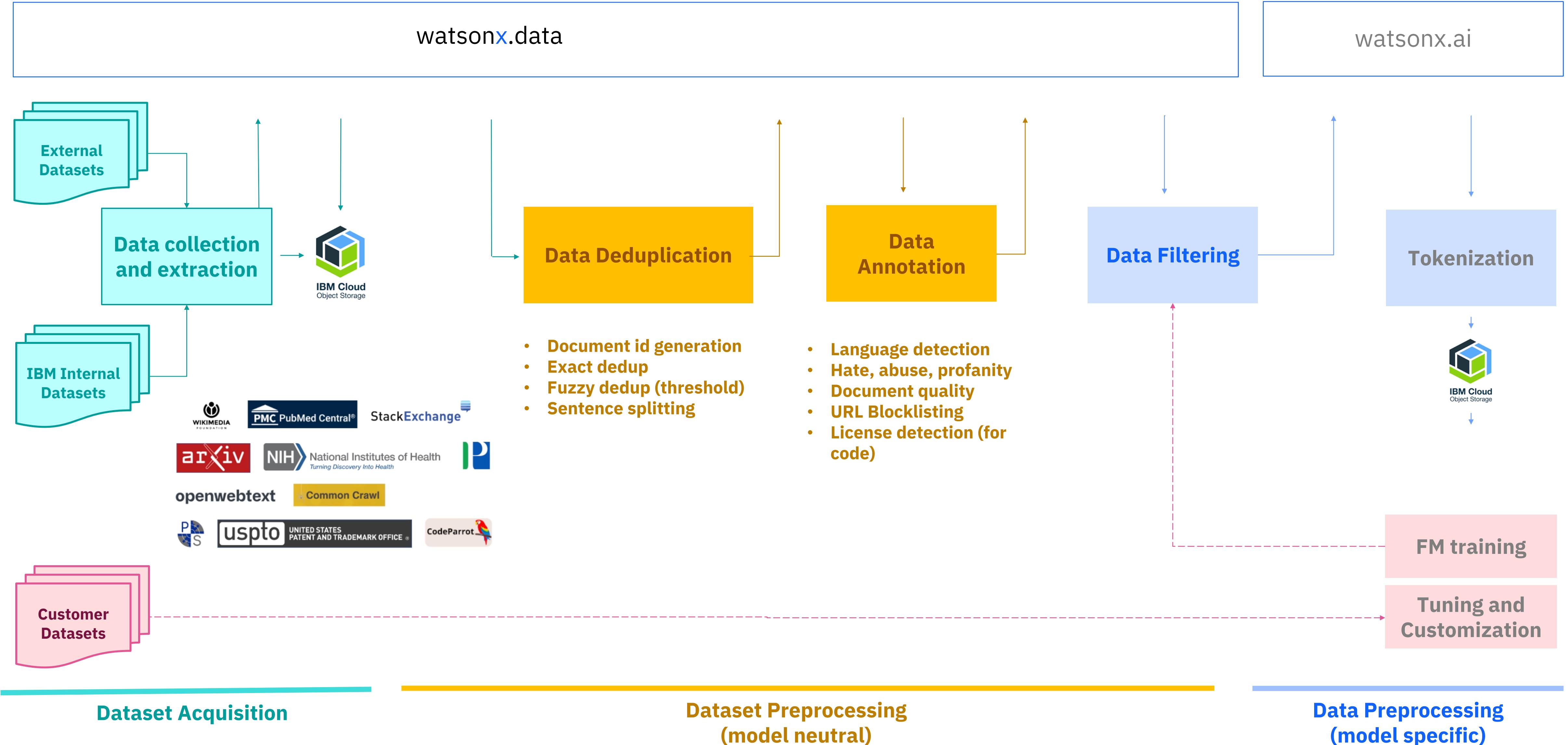
Obsidian
Novel-sparse
universal transformer
architecture



Moonstone
Novel architecture
based on dense
associative memory

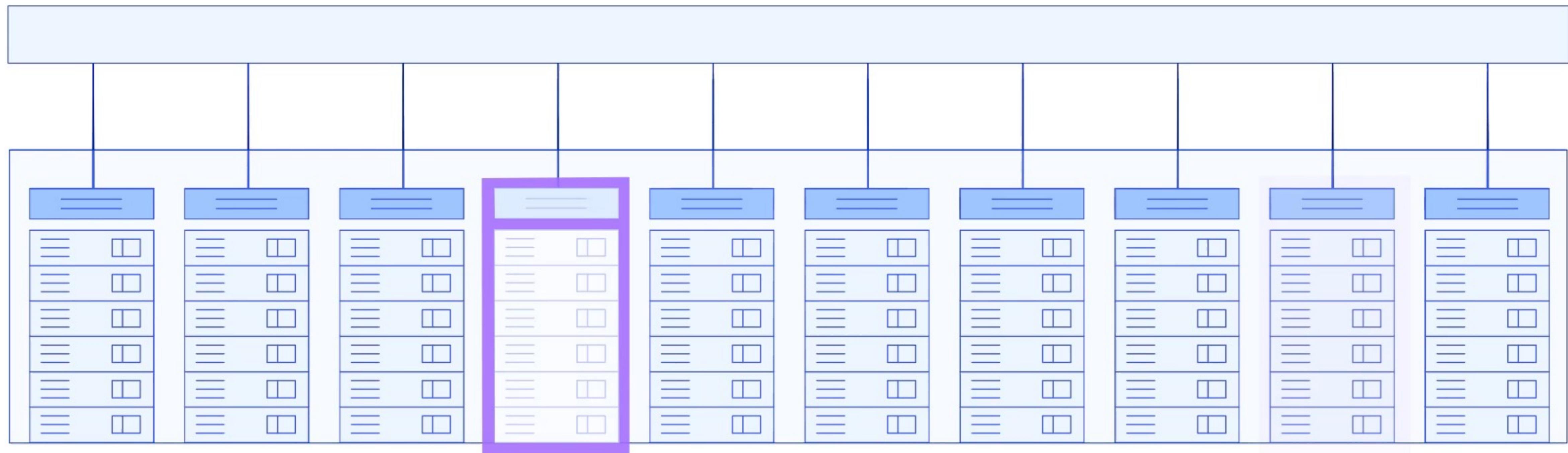
watsonx IBM Data Pile

Enterprise-ready data acquisition, curation, provenance, and governance



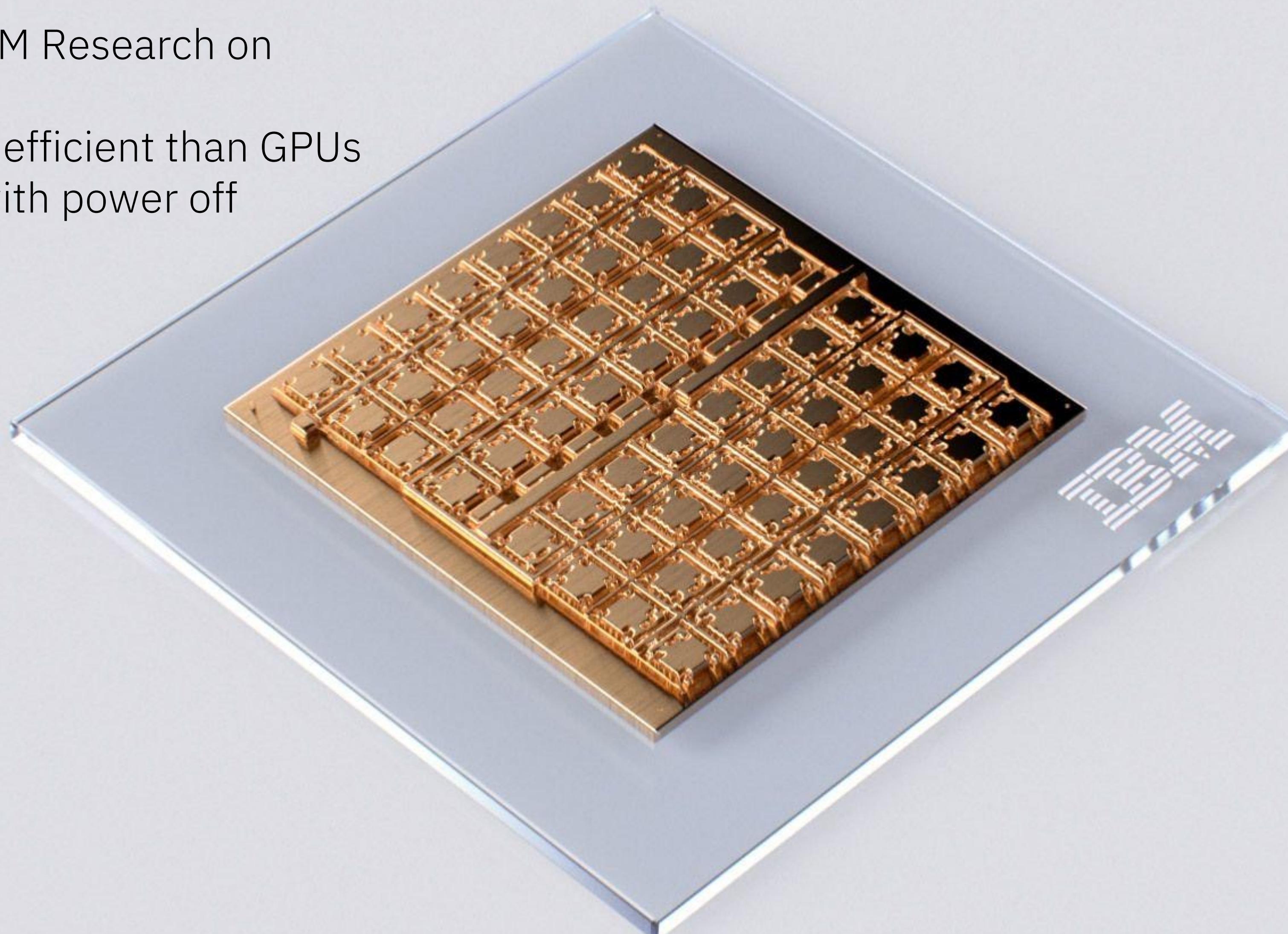
IBM Vela

- Bare-metal performance in the cloud
- < 5% virtualization overhead
- 90%+ GPU efficiency



IBM Analog AI Chip

- Announced by IBM Research on August 23, 2023
- 14x more energy efficient than GPUs
- Maintains state with power off
- Edge enabler

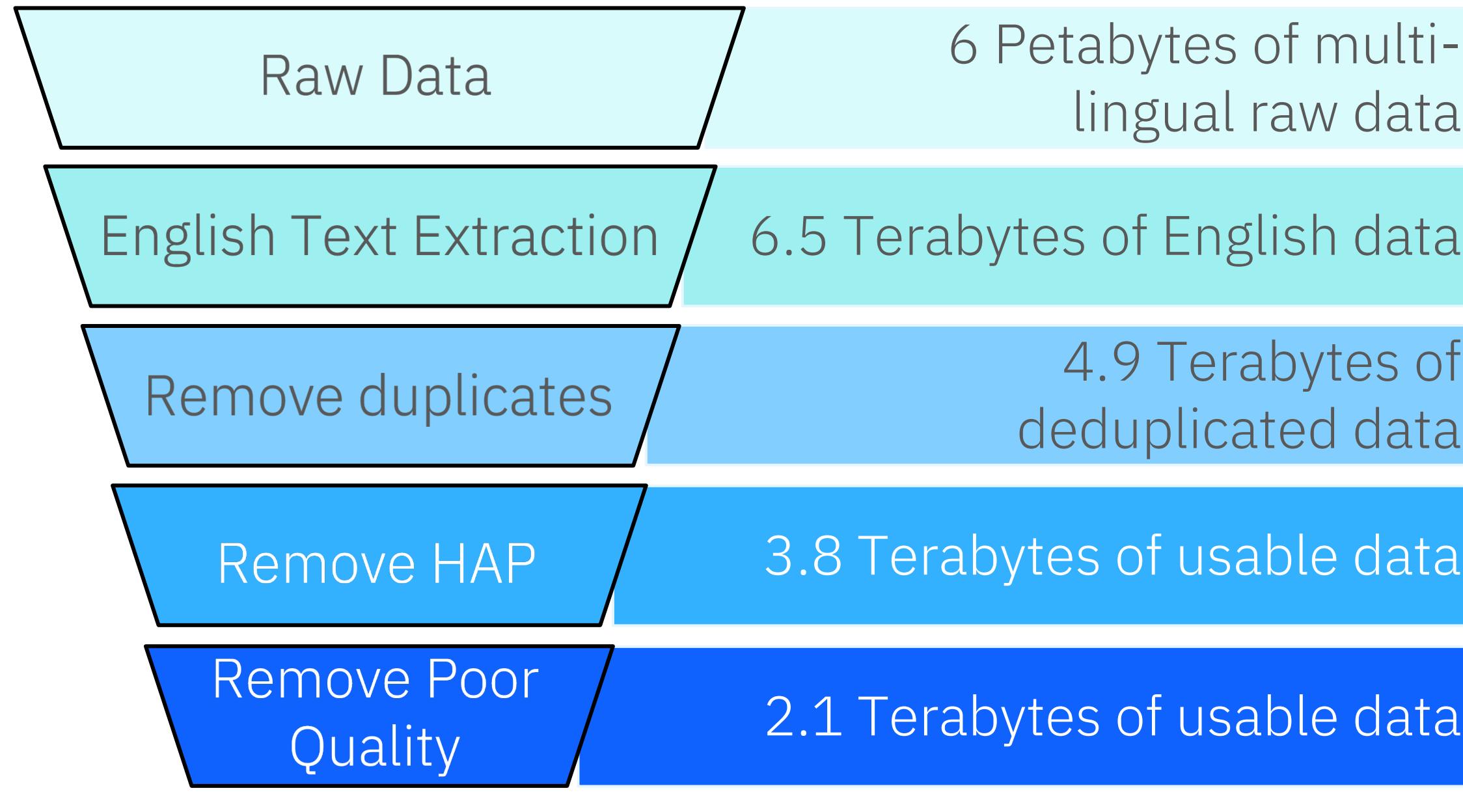


Granite on watsonx.ai

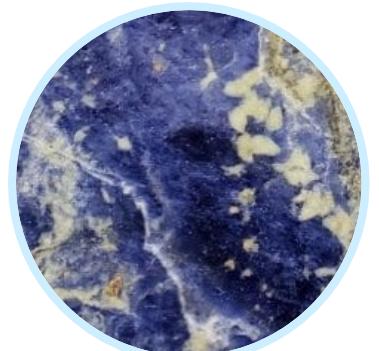
Trusted

IBM's approach to AI model development is grounded in core principles of trust and transparency.

What were the datasets and sources used?



1T Tokens of data for training granite.13b

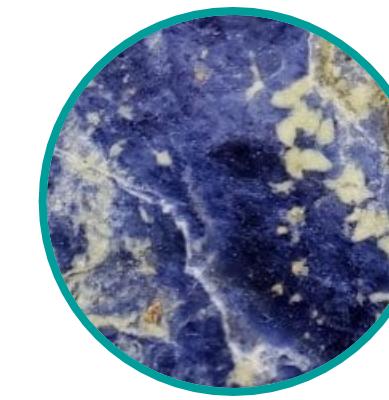
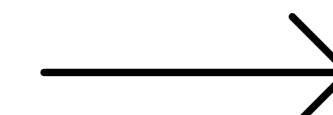


Granite
Decoder-only Transformer
Model
e.g. Llama, Falcon, GPT-3

What makes IBM models safe for enterprise use?

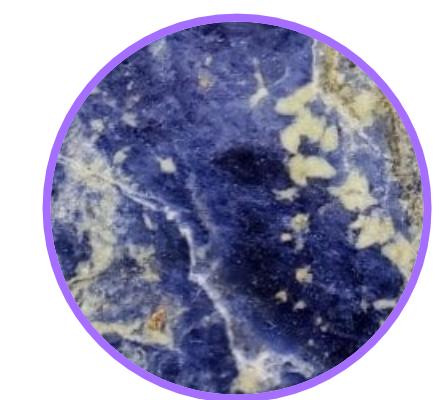
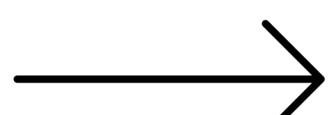
- Models were reviewed against IBM's extensive data governance practices, corresponding to data clearance and acquisition; document quality checks; pre-processing data pipelines, including tokenization, data de-duplication, etc.
- Granite models were trained on data scrutinized by IBM's own HAP detector – to detect and root out objectionable content, benchmarked against internal and public models
- IBM deploys regular, ongoing data protection safeguards, including monitoring for websites known for piracy or other offensive materials, and avoid those websites

granite.13b.instruct



Supervised Fine Tuned

granite.13b.chat



Contrastive Fine Tuned

[Granite Model Paper](#)

IBM Granite models – Enterprise performance delivered faster

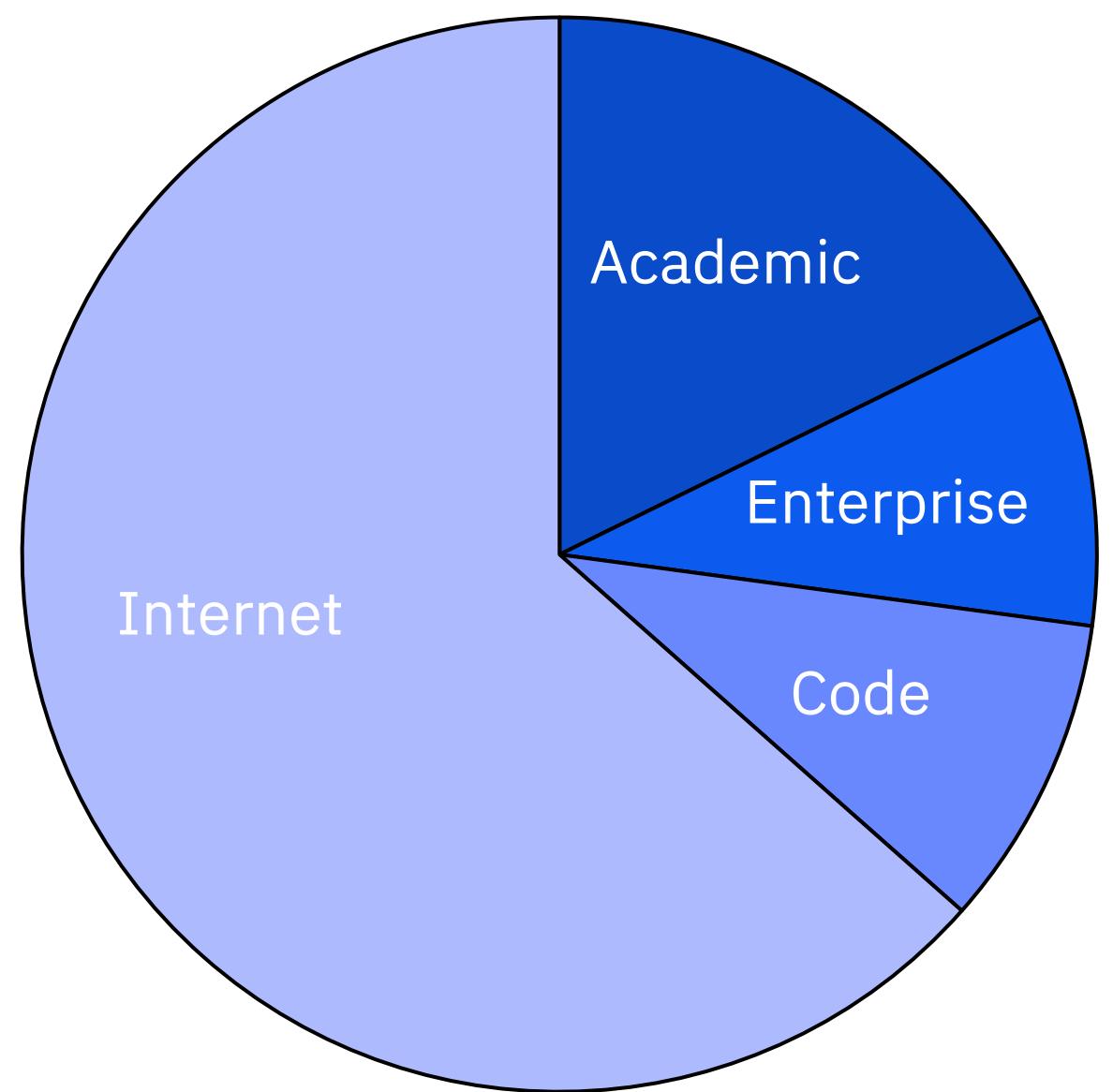
Enterprise



Performance

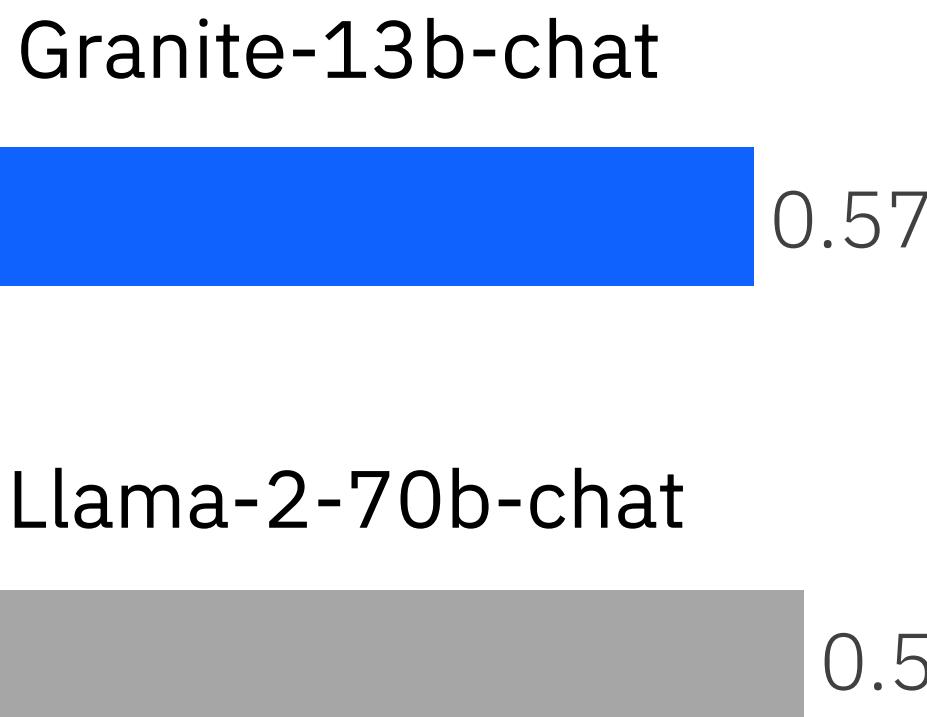


Delivered Faster

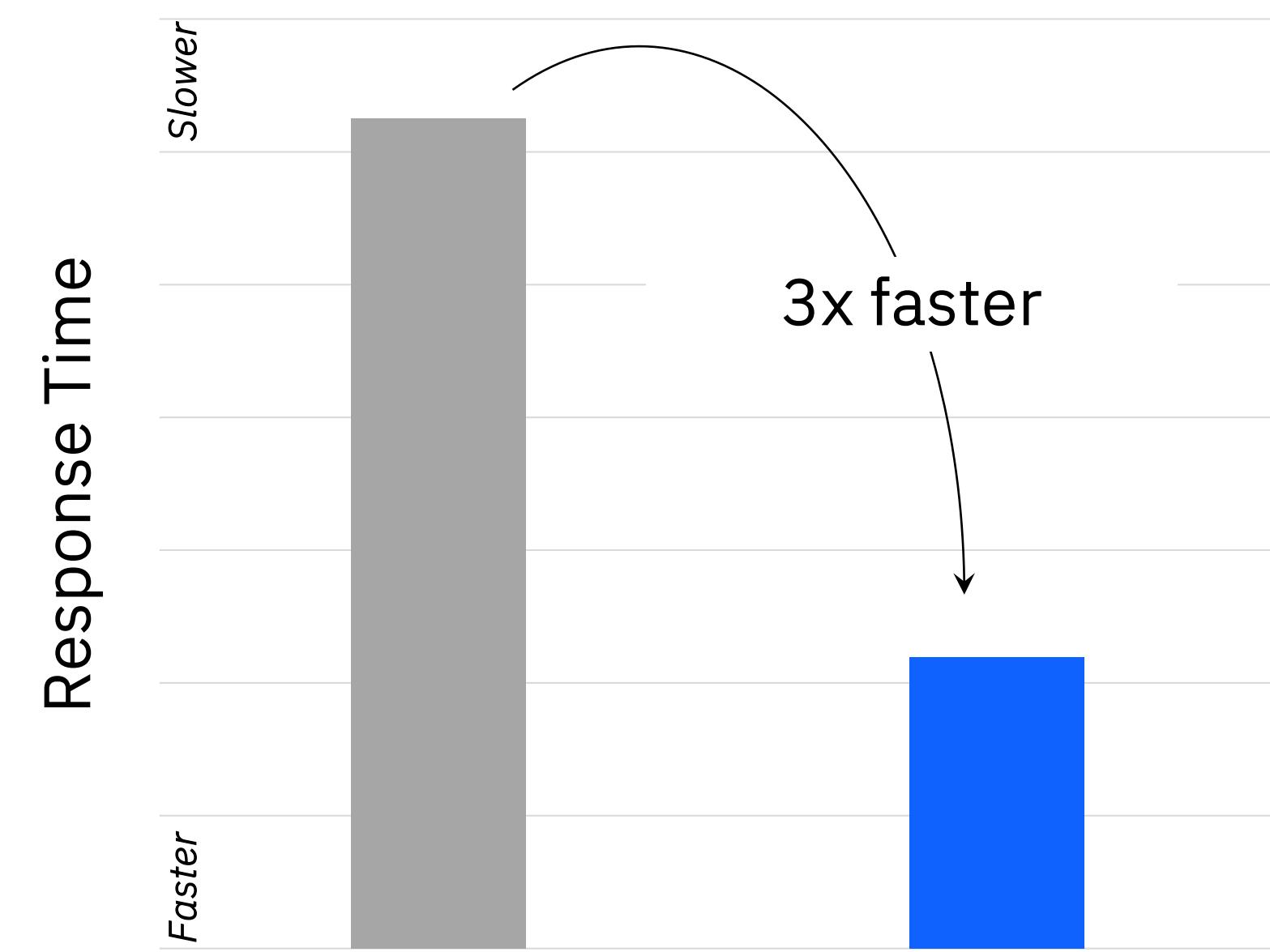


~10% of the data used to train granite is aligned to enterprise tasks including legal and *finance*

Financial Acumen
(Avg of 11 Financial Tasks)



Granite-13b performance is similar to the much larger Llama-2-70b-chat across 11 different *finance* related tasks



Llama-2-70b-chat Granite-13b-chat
In production, Granite-13b can be up to 3x faster than Llama-2-70b for 1000 tokens

(IBM-internal development tests)

IBM Research Finance Evaluation

Finance Panel (11 tasks)

IBM Research Finance Panel

Sentiment analysis

- Financial Phrase Bank
- Stock and Earnings Call Transcripts
- Financial Question Answering

Classification

- News Headline Classification

Entity Extraction

- Edgar SEC Filings (x2)
- Financial Credit Risk Assessment

Question and Answering

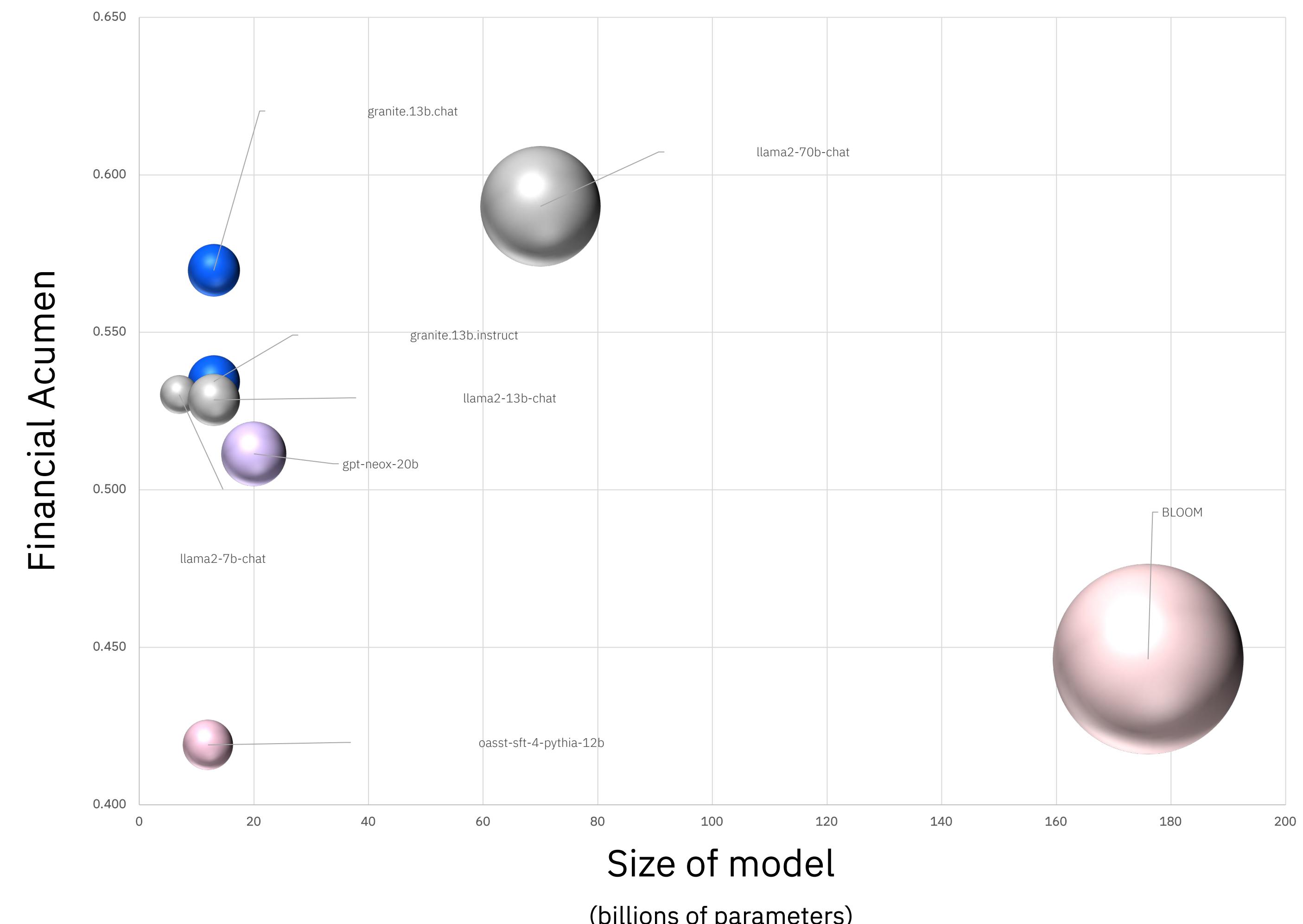
- Financial Question Answering (x2) • Insurance

Summarization

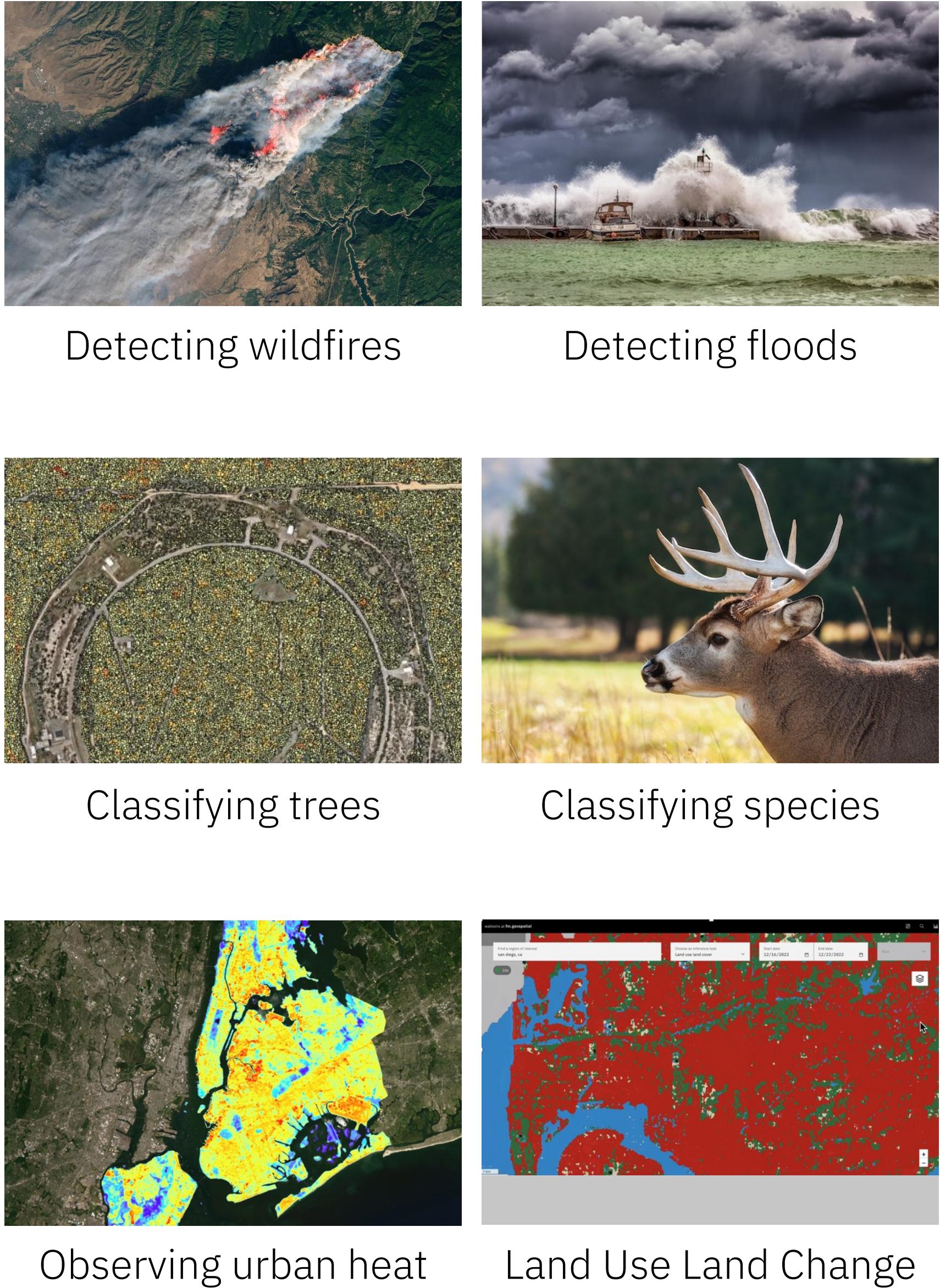
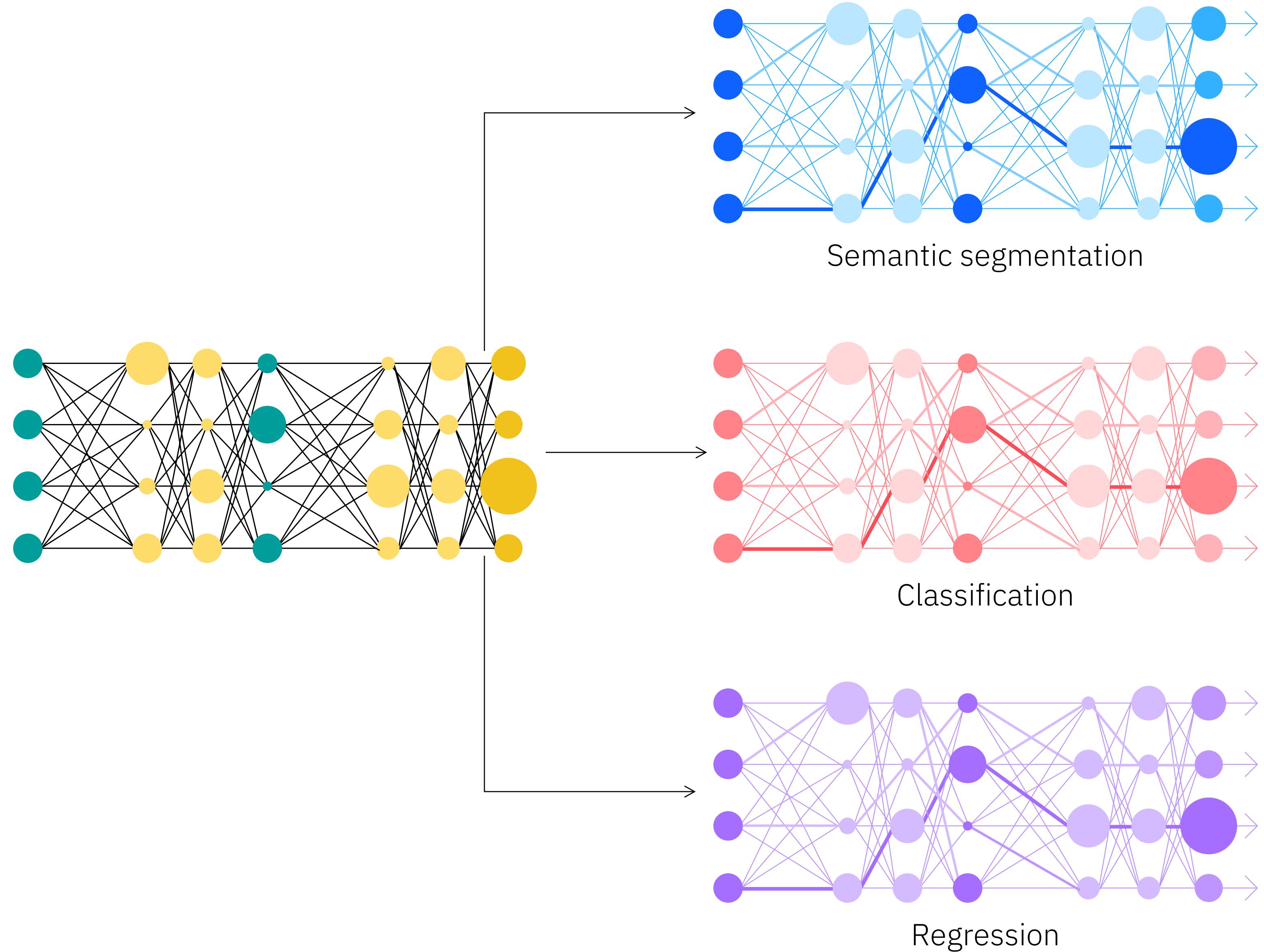
- Corporate Event Detection

Model Performance on Finance Tasks

(average of 11 Financial tasks)



Geospatial Foundation Model for many tasks



Harmonized Landsat Sentinel-2

Geospatial foundation models
focused on **remote sensing** data.

Harmonized Landsat Sentinel-2 (**HLS**)
provides consistent global
observations of the land.

- Data available in **tiles**, aligned with
the Military Grid Reference System
(MGRS).

- 30m** resolution

- Revisited every **2-3 days**

- Each tile has **3660 x 3660 pixels**,
corresponding to ~110 x 110 km.



Model architecture

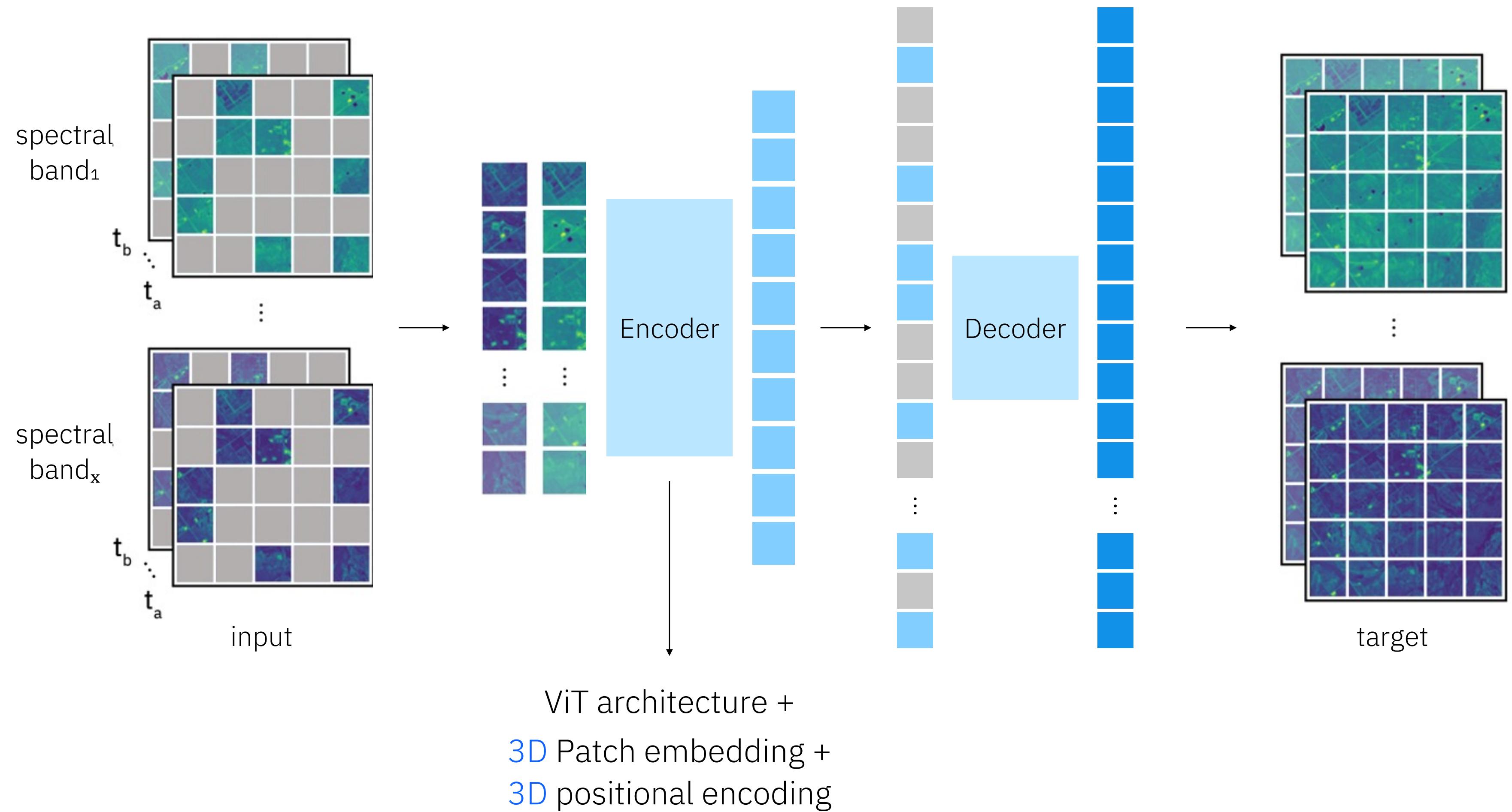
MAE → Masked AutoEncoder

- Pre-training task: reconstruct **masked** patches → target = original data.
- MSE loss on **masked** patches.

Encoder → Vision transformer (**ViT**) for multispectral 3D data.

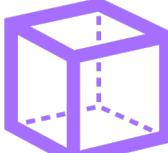
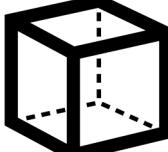
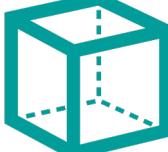
- 3D patch embeddings
- 3D positional encoding

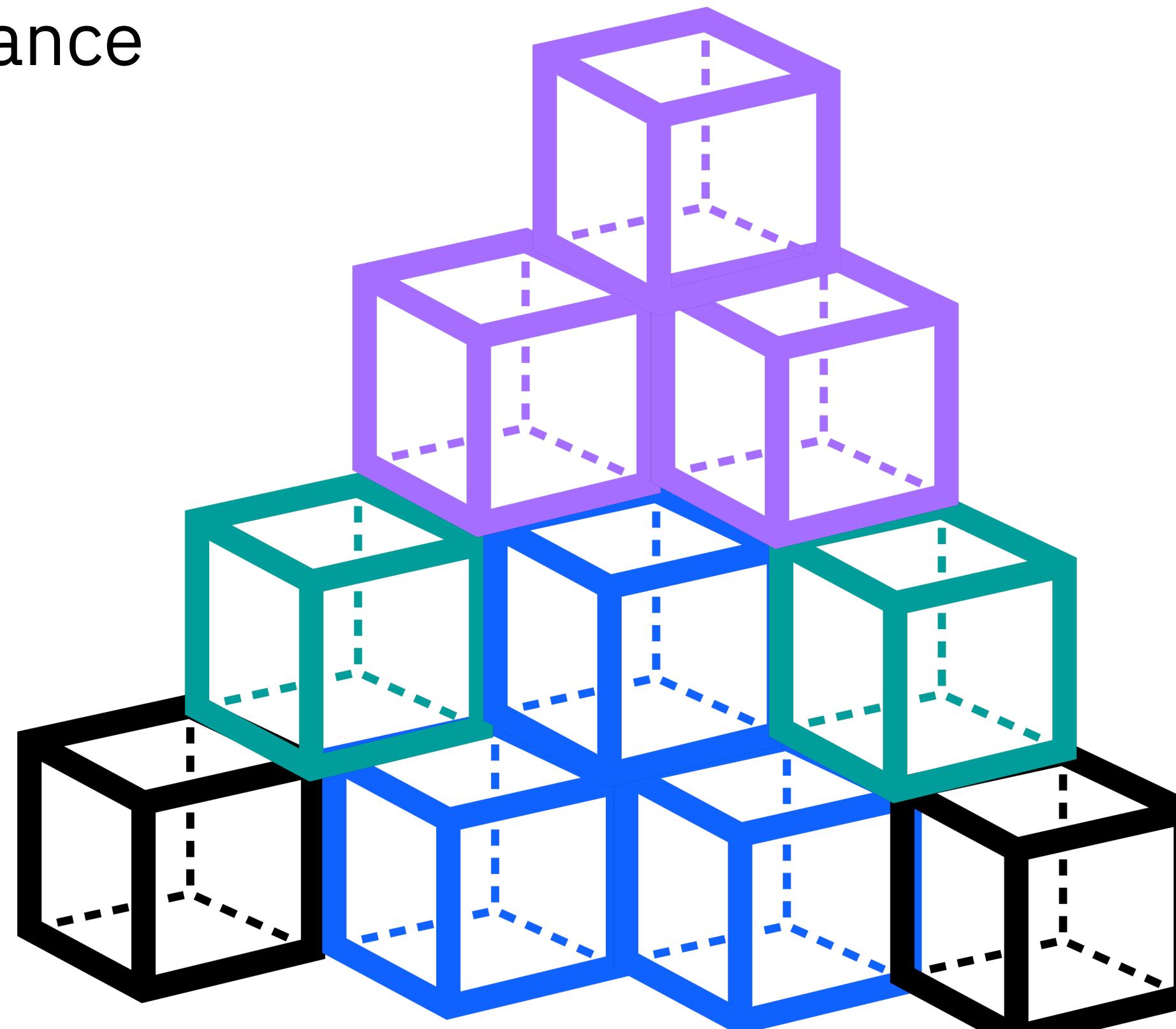
Decoder → Transformer blocks + linear projection layer to match the target patch size.



Our approach to selecting third-party models in watsonX.ai

Technical considerations

-  Model performance
-  Research
-  Ethics
-  Legal and data

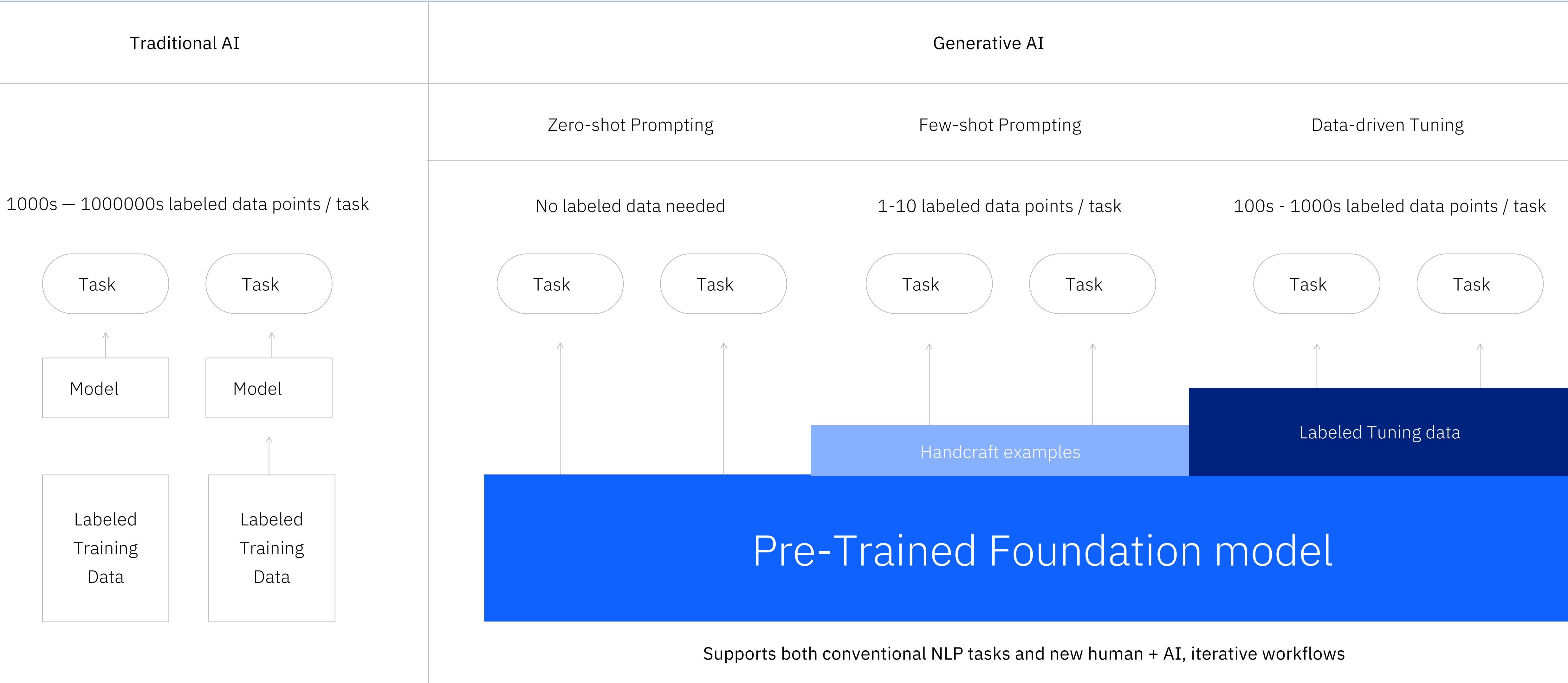


Workflow

-  1 Review technical papers
-  2 Model Information
-  3 Performance Benchmark
-  4 Internal IBM use
-  5 Commercial Applicability
-  6 Licensing
-  7 Reputation
-  8 Use Case Alignment
-  9 Training Data
-  10 Infrastructure

Architecture model

Take advantage of both traditional and Next-Gen AI Models in [watsonx.ai](#)



[**Multitask Prompt Tuning \(MPT\)**](#)

What IBM offers

IBM's approach for AI: Unleash the intelligence in your business

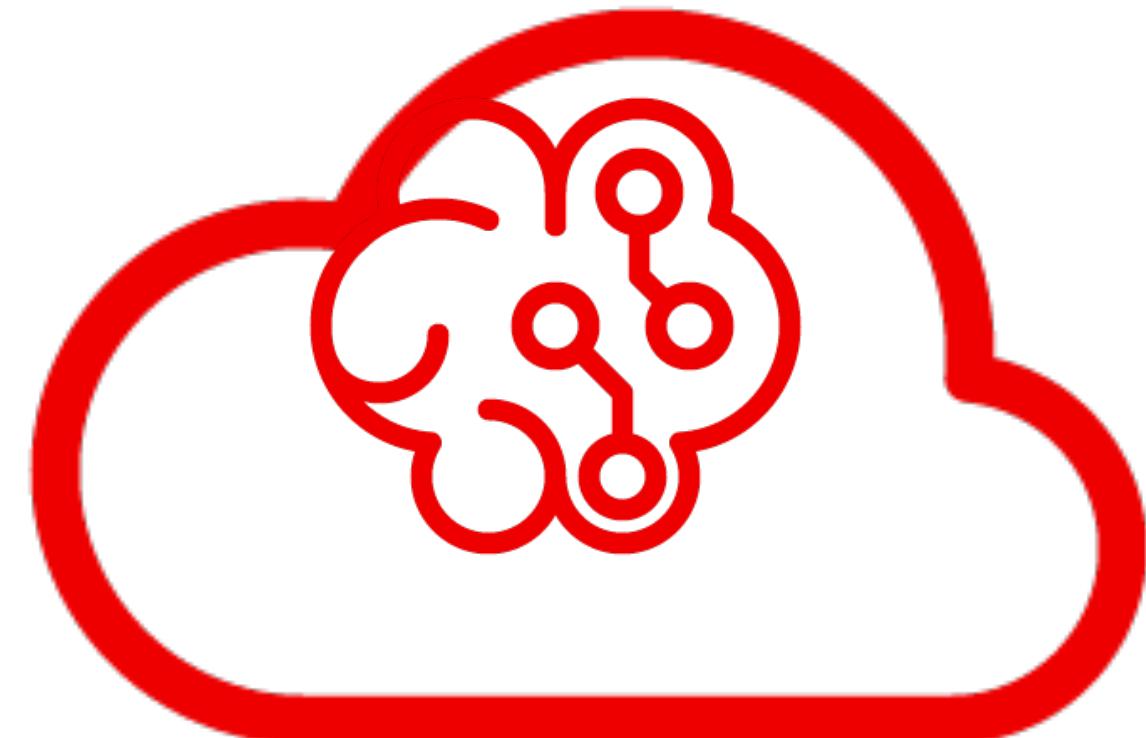


Red Hat OpenShift AI

An AI-focused portfolio that provides tools across the full lifecycle of AI/ML experiments and models and helps build, train, test, and deploy models optimized for hybrid cloud environments.

Red Hat OpenShift AI builds and expands upon the proven capabilities of Red Hat OpenShift and Red Hat OpenShift Data Science, to:

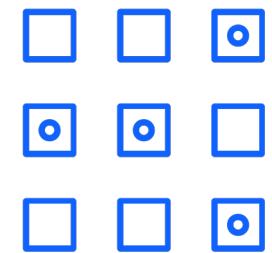
- Provide a unified platform for data scientists and intelligent application developers.
- Scale to handle workload demands of foundation models (volume of data, duration of training run, size of model, acceleration required, and scalability).
- Deliver consistency, ease-of-use, and cloud-to-edge deployment options.



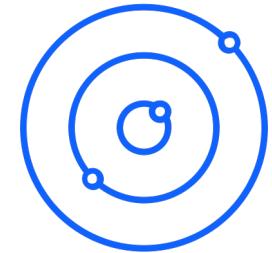
What is watsonx.ai

Train, validate, tune,
and deploy AI models
with confidence

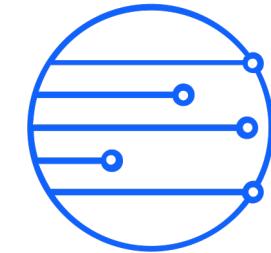
Generative AI capabilities



Foundation
model Libraries

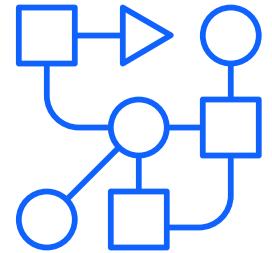


Prompt Lab

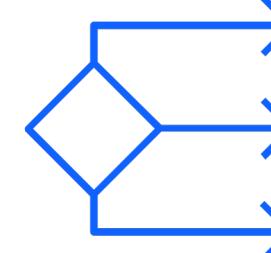


Tuning Studio

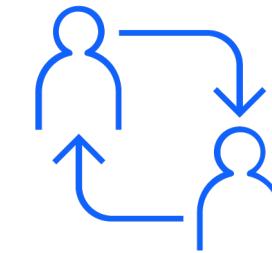
+ a proven studio for Machine Learning



ModelOps



Automated
Development



Team
Collaboration



Decision
Optimization



Leveraged foundation models to train AI to create commentary. [Generated informative and engaging video clip narrations](#) with varied sentence structures and vocabulary.



With Watson Orders, McDonalds uses AI to automate drive thru order taking, enabling employees to increase focus on food delivery and customer service.

Disparate AI Modeling Patterns

- Wizard tooling: AutoAI - no-code
- Drag/drop canvas: Visual - low-code
- Notebooks / IDEs: Programmatic -all-code
- Distributed Compute and DL
- Prompt, Tuning, Foundational Models

Pipeline inspection

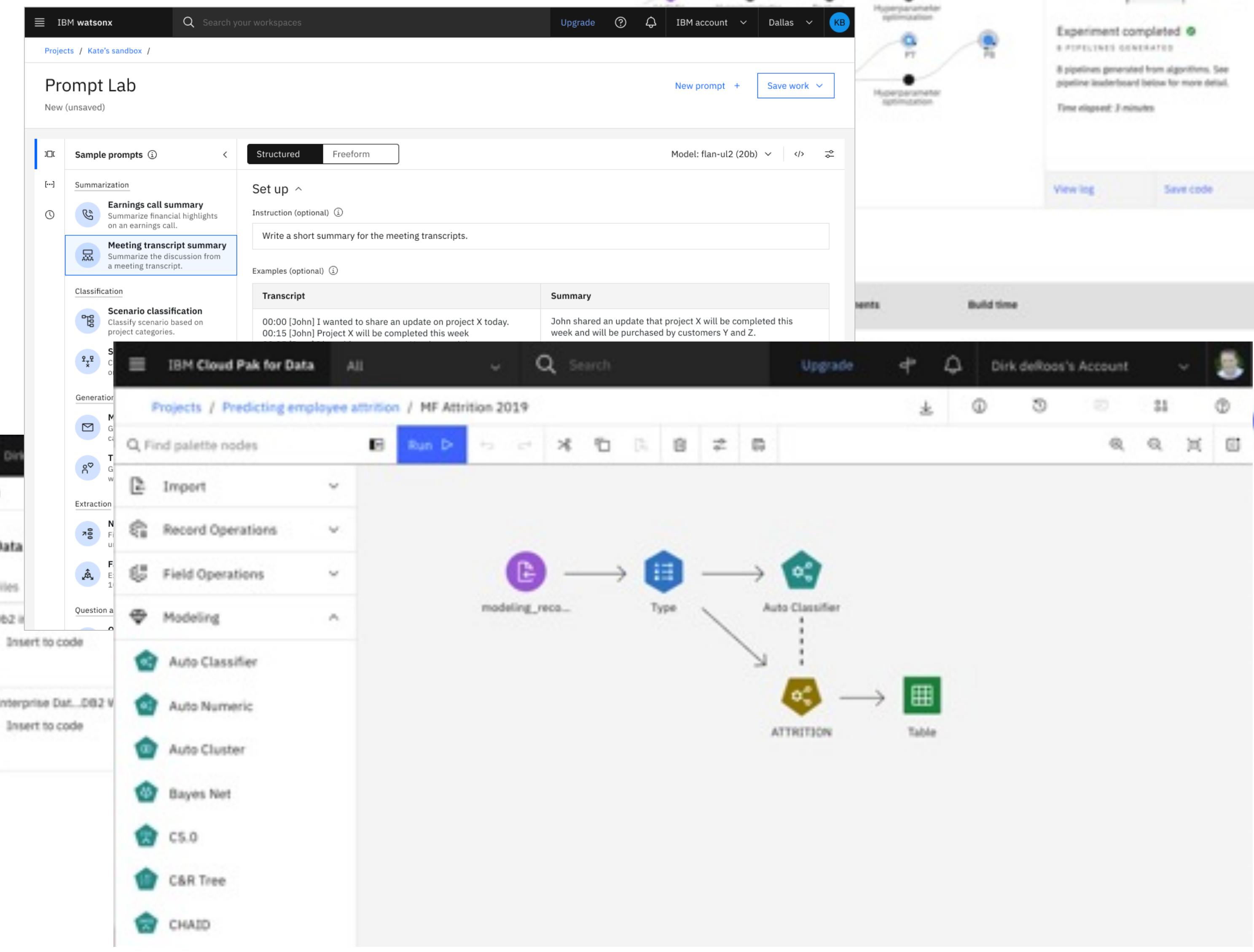
Read training data

Retrieve training dataset from AutoAI experiment as pandas DataFrame. If reading data using Flight Service Connection results with error, please provide data as Pandas DataFrame object e.g. reading .CSV file with `pandas.read_csv()`.

```
In [ ]: df = training_data_reference[0].read(csv_separator=experiment_metadata['csv_separator'])
df.dropna('rows', how='any', subset=[experiment_metadata['prediction_column']], inplace=True)
```

Train and test data split

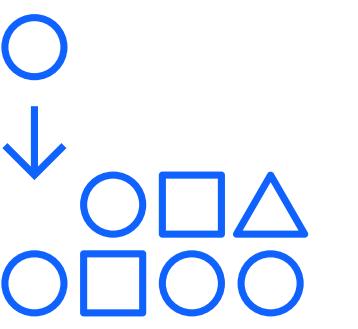
```
In [ ]: from sklearn.model_selection import train_test_split
df.drop_duplicates(inplace=True)
```



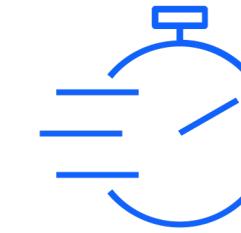
watsonx. governance

Enable responsible,
transparent and
explainable AI workflows

An end-to-end toolkit for
AI governance



Govern across the AI lifecycle by automating and consolidating tools, applications and platforms.



Manage risk and protect reputation by automating workflows to better detect fairness, bias and drift.



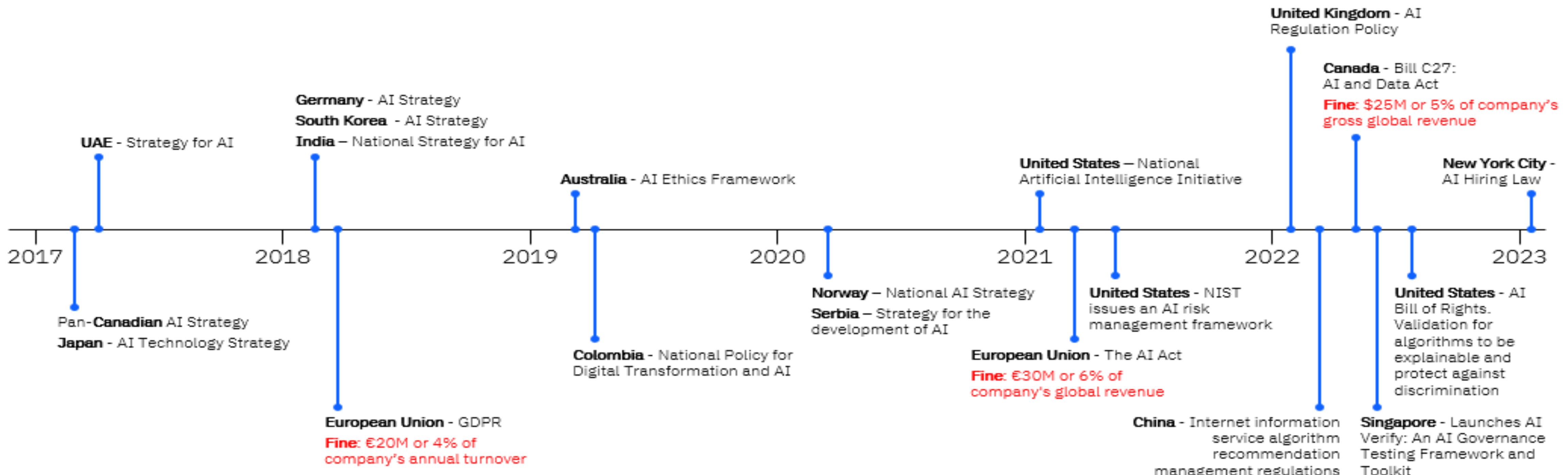
Help adhere to regulations by translating growing regulations into enforceable policies.

Comprehensive
Govern the end-to-end AI lifecycle with metadata capture at each stage

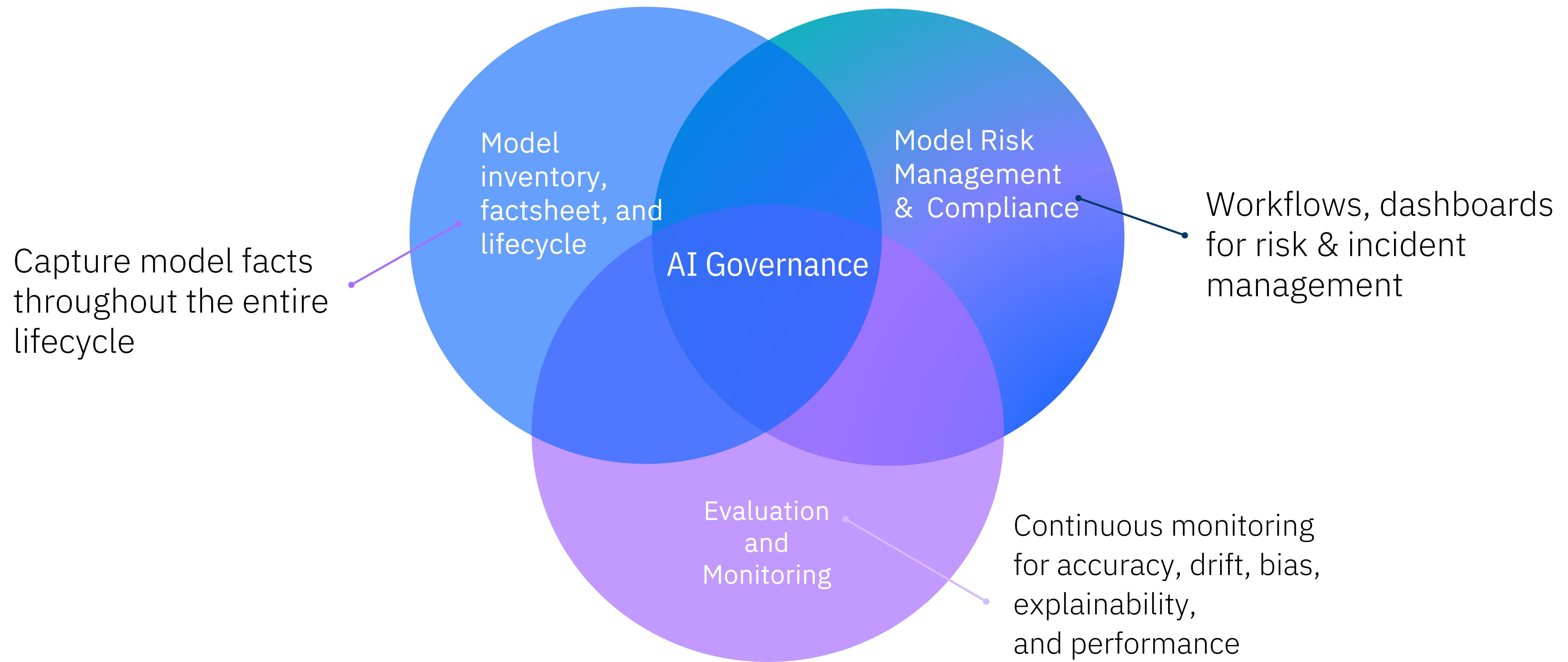
Open
Support governance of models built and deployed in 3rd party tools.

Automatic metadata
and data transformation/lineage capture through Python notebooks.

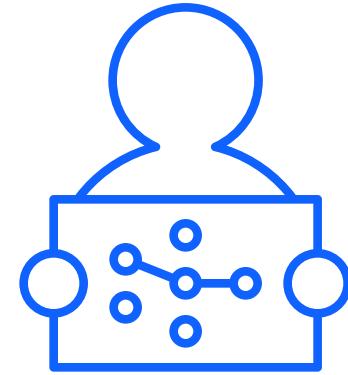
Changing AI regulations



AI Governance is imperative

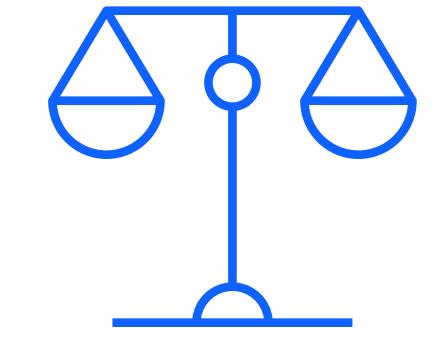


Pillars of Trust



Explainability

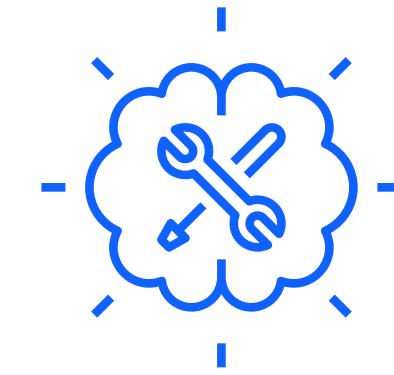
An AI system's ability to provide a human-interpretable explanation for its predictions and insights



Fairness

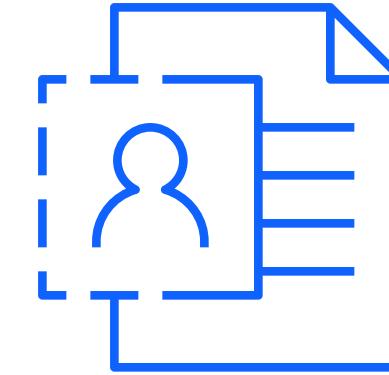
Equitable treatment of individuals or groups by an AI system

Depends on the context in which the AI system is used



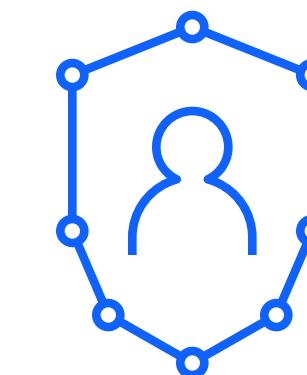
Robustness

An AI system's ability to effectively handle exceptional conditions, such as abnormalities in input



Transparency

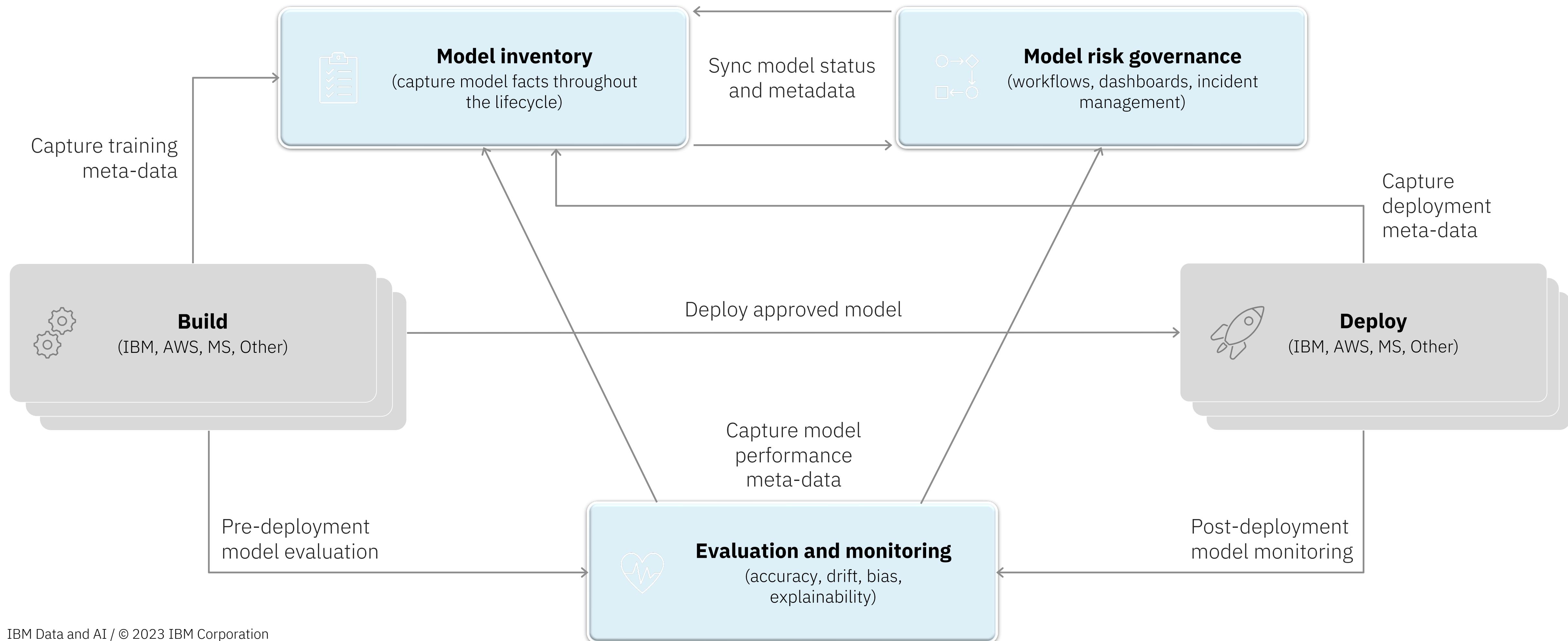
An AI system's ability to include and share information on how it has been designed and developed



Privacy

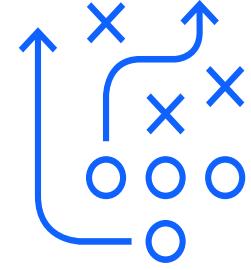
An AI system's ability to prioritize and safeguard consumers' privacy and data rights

IBM watsonx.governance integrates and augments your existing ML development and deployment tools and processes



Three ways to get started with **watsonx.ai** today

IBM's investment in partnering with you



FREE TRIAL

Experience **watsonx.ai** and test out core capabilities yourself with a free trial



[Try our free trial](#)

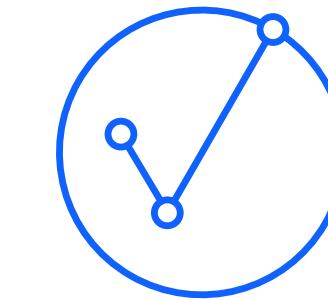


CLIENT BRIEFING

Discussion and custom demonstration of IBM's generative AI **watsonx** point-of-view and capabilities. Understand how **watsonx.ai** can be leveraged in your AI strategy.

2-4 hours

Onsite or virtual



PILOT PROGRAM

watsonx.ai pilot developed with IBM AI engineers. Prove **watsonx.ai** value for the selected use case(s) with a plan for adoption.

1-4 weeks

Workshop - Hands On Lab



<https://github.com/krondor/missouridatasciencesymposium2023>

Before we Begin



Create Project or Import Project



Associate Machine Learning Instance to Project



Define Access Token



Define IAM API Key

Create Access Token

Programmatic Access with Project
Interfacing Libraries

The screenshot shows the WatsonX workshop interface. At the top, the navigation bar includes 'Projects / watsonx workshop' and tabs for 'Overview', 'Assets', 'Jobs', and 'Manage'. The 'Manage' tab is selected, highlighted with a blue underline. Below the tabs, there's a sidebar titled 'Project' with options: 'General', 'Access control' (which is selected and highlighted with a blue border), 'Environments', 'Resource usage', and 'Services & integrations'. The main content area is titled 'Access control' and shows two tabs: 'Collaborators (1)' and 'Access tokens (1)', with 'Access tokens (1)' also having a blue border around it. A search bar says 'Find access tokens'. A table header row has columns for 'Name' and 'Role'. A modal window titled 'New access token' is open, containing instructions: 'Create a token to access assets in this project from a code editor. A token is a unique, multi-character string.' It has fields for 'Name' (containing 'mizzou-watsonx') and 'Access role' (set to 'Editor'). At the bottom of the modal are 'Cancel' and 'Create' buttons, with 'Create' being highlighted with a blue background.

Projects / watsonx workshop

Overview Assets Jobs Manage

Project

General

Access control

Environments

Resource usage

Services & integrations

Access control

Collaborators (1)

Access tokens (1)

Find access tokens

Name Role

New access token

Create a token to access assets in this project from a code editor. A token is a unique, multi-character string.

Name

mizzou-watsonx

Access role ⓘ

Editor

Cancel Create

Associate Machine Learning Instance

Add instance to project for inferencing

Associate service

Choose an existing or add a new service to associate with your project.

The screenshot shows a search interface for selecting a service. At the top, there are dropdown menus for 'Resource Groups' (set to 'Locations'), 'Locations' (set to 'Ryan.Kather'), and a search bar containing '1 x Ryan.Kather'. Below this is a list of service instances:

Service Name	Type
Machine Learning-5f ⓘ	Watson Machine Learning
Machine Learning-di ⓘ	Watson Machine Learning
Machine Learning-kr ⓘ	Watson Machine Learning
Machine Learning-sg ⓘ	Watson Machine Learning
Machine Learning-74 ⓘ	Watson Machine Learning

The 'Machine Learning-di' entry is highlighted with a checked checkbox and a grey background, indicating it is selected for association.

Services & integrations

IBM services (1)

Third-party integrations

Find services

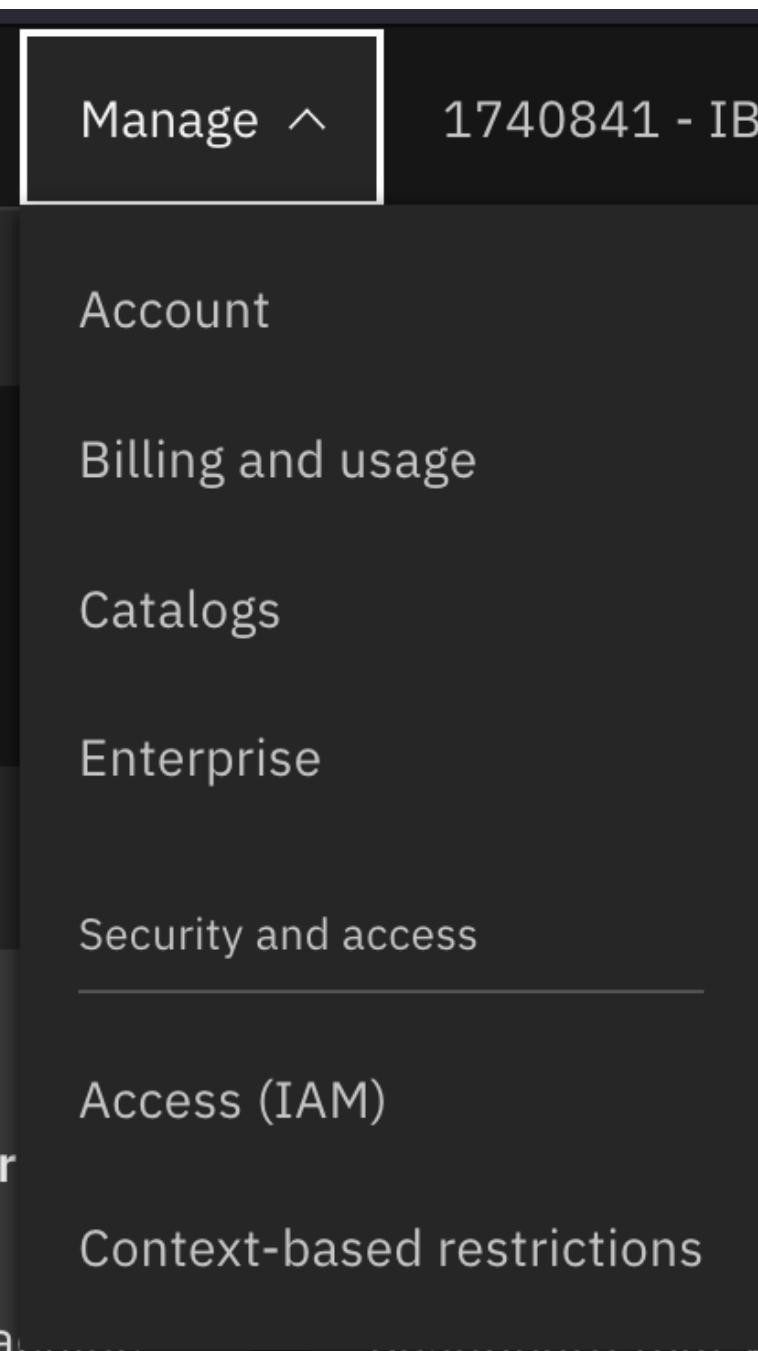
Name

Machine Learning-di

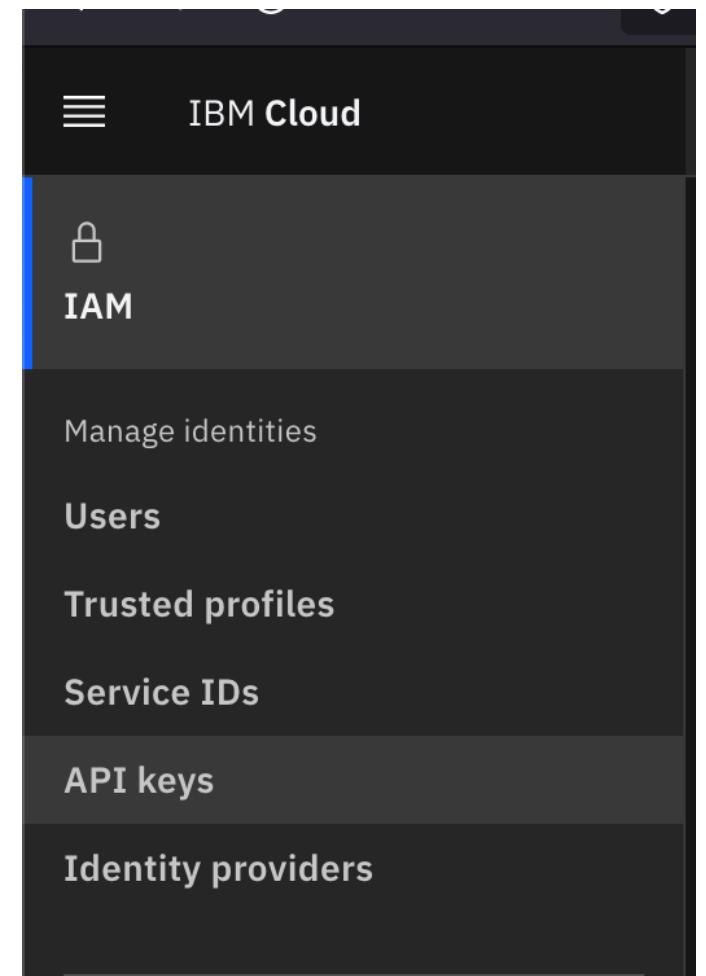
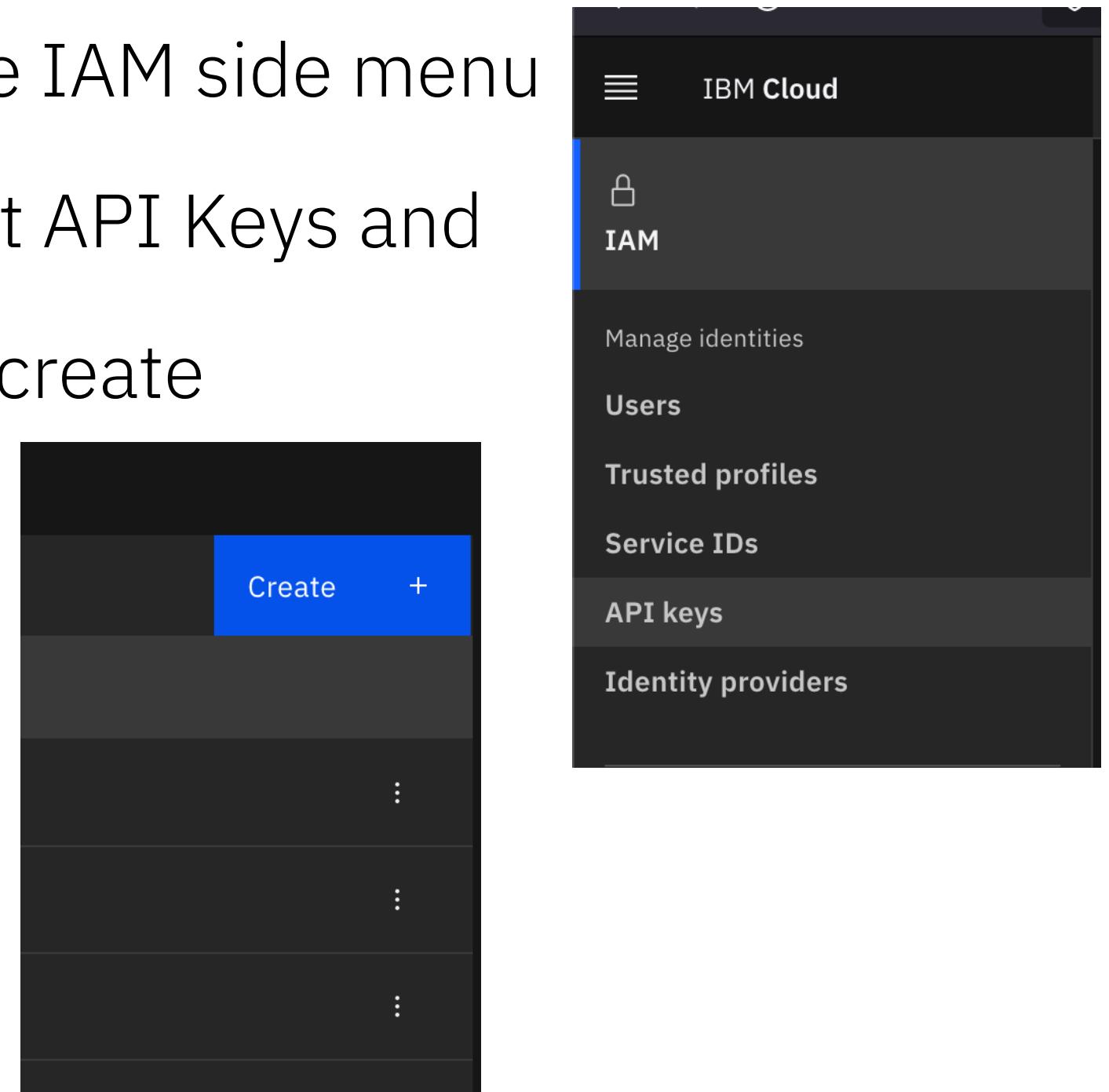
Associate service +

IAM API Key

Login to <https://cloud.ibm.com> and click Manage and then Access (IAM)



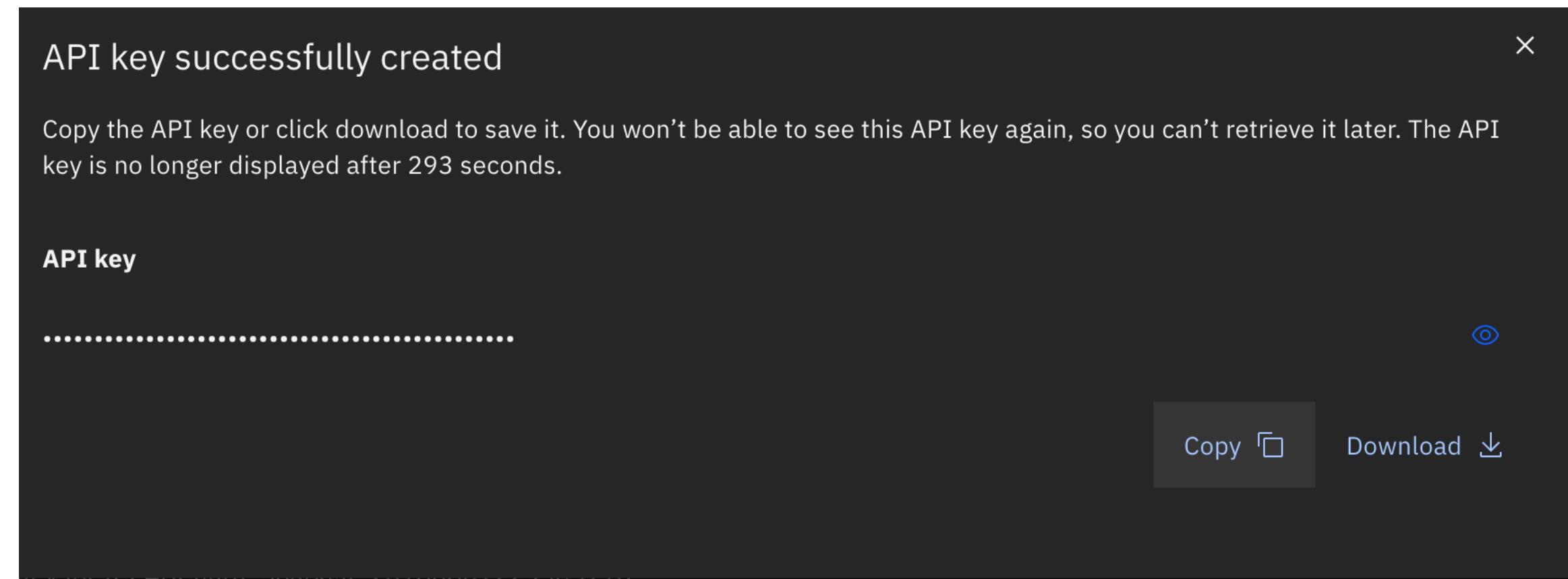
In the IAM side menu select API Keys and click create



– Name the key and provide a description.

A screenshot of the 'Create IBM Cloud API key' dialog. It has two input fields: 'Name' with 'mizzou-workshop' and 'Description' with 'mizzou IAM workshop key'. At the bottom are 'Cancel' and 'Create' buttons, with 'Create' being blue and highlighted.

– Save the Key



Watson Studio & WML APIs and SDK

Related but Different Purposes

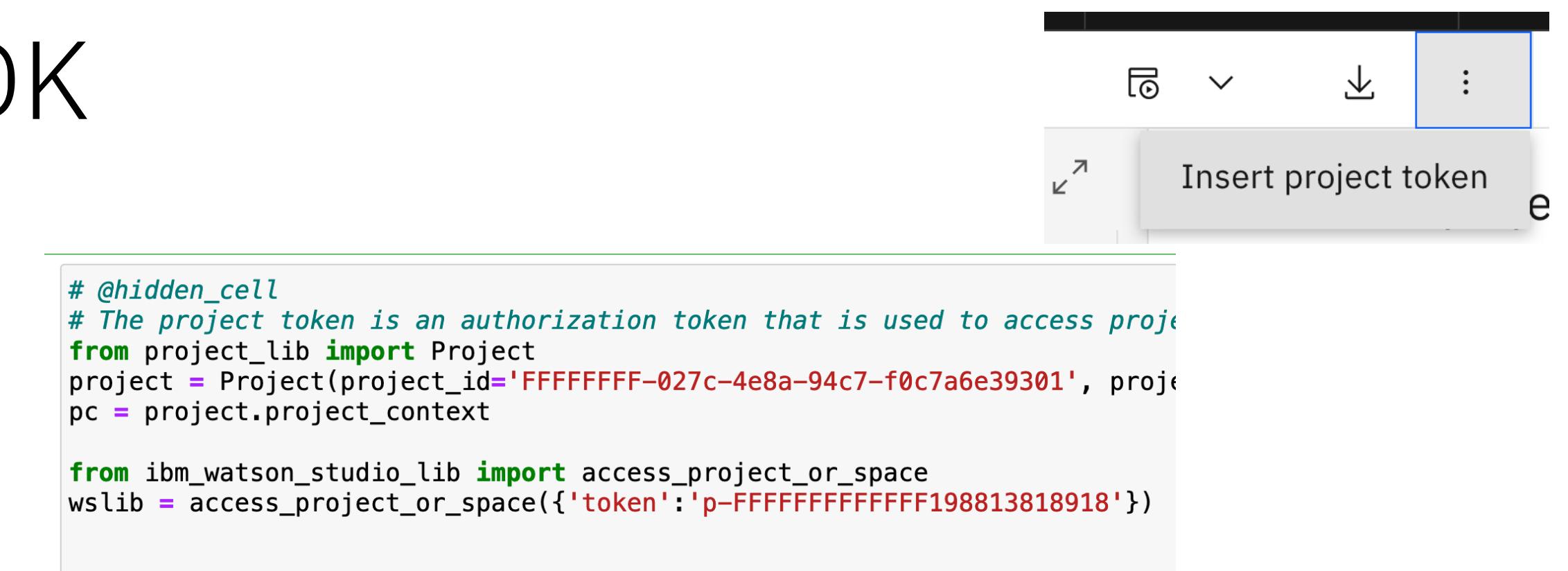
Programmatically interface with projects, data, and assets within project spaces.

Samples Repo:

<https://github.com/IBMDatascience/sample-notebooks/blob/master/CloudPakForData/notebooks/4.0/IPYNB/Working%20with%20ibm-watson-studio-lib%20in%20CPD.ipynb>

Documentation:

<https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/using-ibm-ws-lib.html?context=wx&audience=wdp>



```
# @hidden_cell
# The project token is an authorization token that is used to access projects
from project_lib import Project
project = Project(project_id='FFFFFFFFFF-027c-4e8a-94c7-f0c7a6e39301', project_token='p-FFFFFFFFFF198813818918')
pc = project.project_context

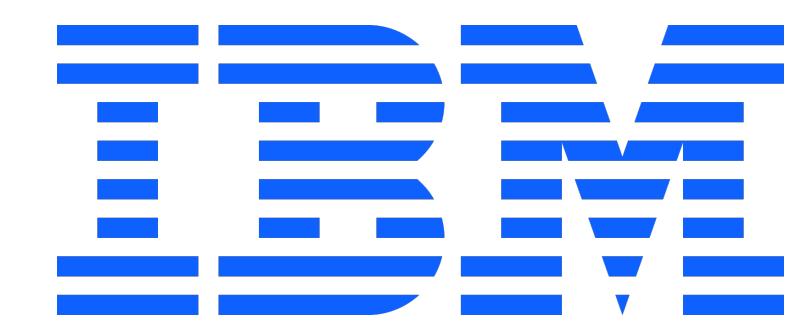
from ibm_watson_studio_lib import access_project_or_space
wslib = access_project_or_space({'token': 'p-FFFFFFFFFF198813818918'})
```

Programmatically instantiate models, package metadata, manage deployments, and deployment spaces.

Samples Repo: <https://github.com/IBM/watson-machine-learning-samples>

Documentation: <https://ibm.github.io/watson-machine-learning-sdk/>

- Persist Data without Exposing in UI: `wslib.storage.save_data()` `wslib.upload_file()` `wslib.save_data()`



Metrics for evaluating Large Language Models

Text Summarization Metrics

- [ROUGE](#)
- [SARI](#)
- [WIKI_SPLIT](#)
- [BLEURT](#)
- [METEOR](#)
- [Sentence Similarity - Jaccard Similarity](#)
- [Sentence Similarity - Cosine Similarity](#)

Content Generation, Q&A Evaluation Metrics

- [BLEU](#)
- [exact_match](#)
- Perplexity**
- [rl_reliability**](#)

Text Classification Metrics

- [Accuracy](#)
- [Precision](#)
- [Recall](#)
- [ROC AUC](#)
- [F1 Score](#)
- [Brier Score](#)
- [GLUE metrics](#)
- [Matthews Correlation Coefficient](#)
- [Label Skew](#)

Watson NLP

- HAP Detection**
- PII Detection**

Entity Extraction Metrics

- [Seq eval](#)
- ### Language Translation Evaluation Metrics
- [Character](#)
- [charcut_mt](#)
- [Chrf](#)
- [google_bleu](#)
- [super_glue](#)
- [TER](#)
- [nist_mt](#)
- [Pöseval](#)
- [sacrebleu](#)
- [XTREME-S](#)

Reference-Free Metrics from Haifa Research Labs

- Levenshtein distance based Diversity metrics
- [Textstat](#) toolkit based flesch metrics to determine readability, complexity, and grade level.
- blanchelp**
- Shannon**
- BartScore - Exploring
- Rquge - Exploring

Under Exploration from SVL Research Labs

- Stigma Detection**
- Social Bias/Values Detection**
- Faithfulness / Hallucination**

** Uses Metrics Computation Model

Reinventing how work gets done | Application modernization solution

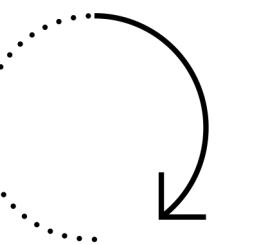
Watson Code Assistant

Write code faster with AI-generated recommendations based on natural language inputs

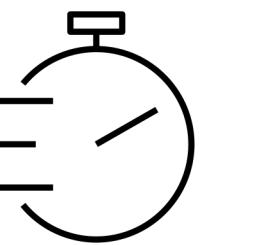
Proven results:

30%
Productivity gain
in application
modernization

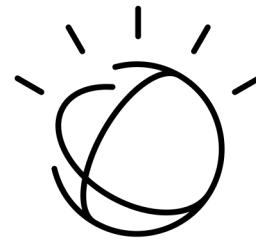
Accelerate code development through AI-powered natural language processing and increase developer productivity



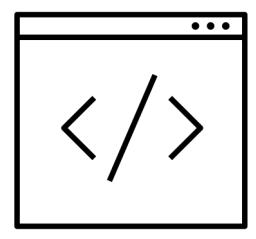
Reduce time to value for automation and improve development cycles with AI code recommendations



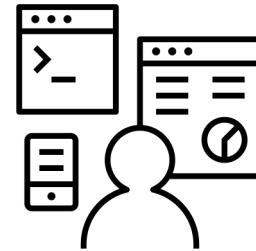
Enhance the quality of code, with greater efficiency and improved accuracy



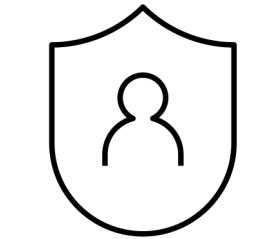
Maintain high levels of accuracy and transparency through data-source attribution



Narrow the IT skills gap by accelerating developer onboarding, making software development more accessible



Tune the **watsonx** foundation model with your own data set to apply organizational best practices



Sample code conversion from Cobol to Java using Generative AI

COBOL

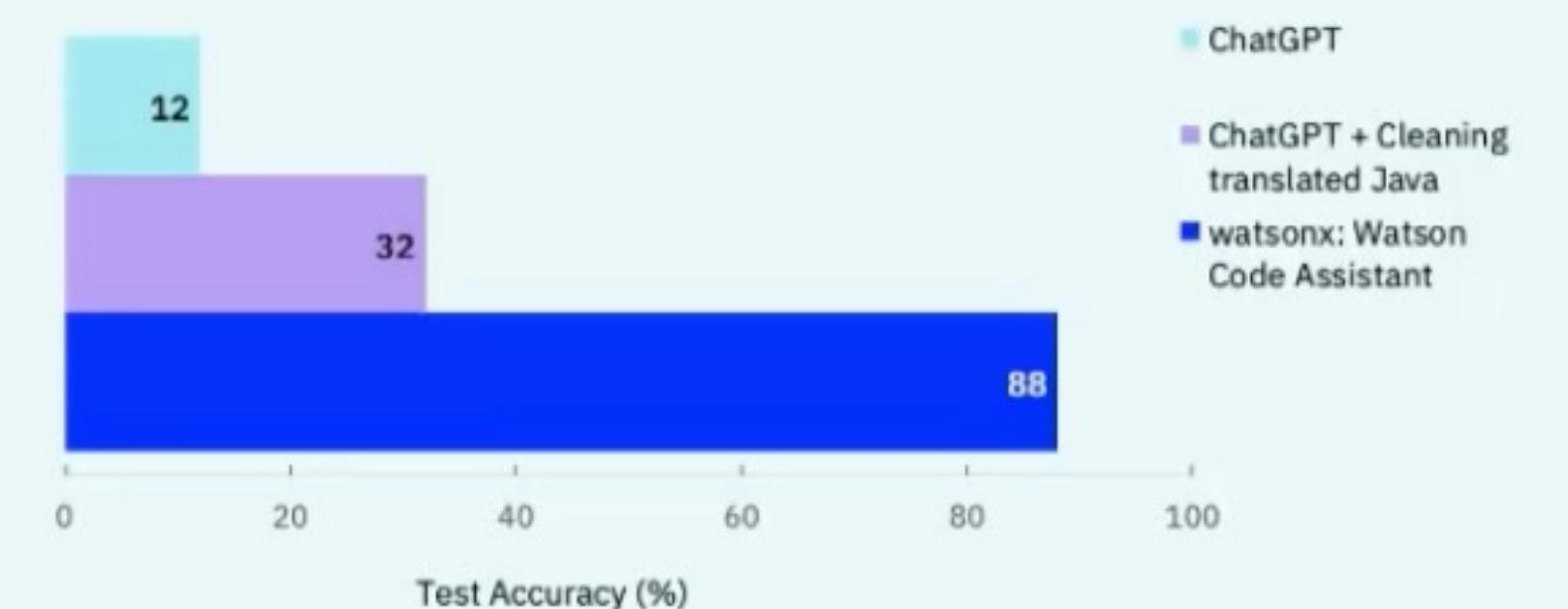
```
1 IDENTIFICATION DIVISION.  
2 PROGRAM-ID. HELLO.  
3  
4 DATA DIVISION.  
5 WORKING-STORAGE SECTION.  
6 01 ARGS_tbl OCCURS 100 TIMES .  
7 05 ARGS PIC X(100).  
8 01 N PIC S9(10).  
9 01 G PIC S9(10).  
10 01 MIN PIC S9(10).  
11 01 COUNT1 PIC S9(10).  
12 01 SUM1 PIC S9(10).  
13 01 I PIC S9(10).  
14 01 X PIC S9(10).  
15 01 TEMP_0 PIC S9(10).  
16 01 DIVCOBOL PIC S9(10).  
17  
18 PROCEDURE DIVISION.  
19 ACCEPT N  
20 ACCEPT G  
21 MOVE G TO MIN  
22 MOVE 0 TO COUNT1  
23 MOVE 0 TO SUM1  
24 PERFORM VARYING I FROM 0 BY +1  
      UNTIL (I >= N)  
25 ACCEPT X  
26 ADD X TO SUM1  
27 IF (MIN > X) THEN  
    MOVE X TO MIN  
28 END-IF  
29 END-PERFORM  
30 SUBTRACT SUM1 FROM G GIVING SUM1  
31 PERFORM UNTIL ((SUM1 - MIN) < 0)  
32     SUBTRACT MIN FROM SUM1  
33     ADD 1 TO COUNT1  
34 END-PERFORM  
35 ADD N COUNT1 GIVING TEMP_0  
36 DISPLAY TEMP_0  
37 STOP RUN.  
38 EXIT PROGRAM.
```

Java

```
1 import java.util.*;  
2  
3 public class Main {  
4     public static void main(String[]  
5         args) {  
6         Scanner sc = new  
7             Scanner(System.in);  
8         int n = sc.nextInt();  
9         int g = sc.nextInt();  
10        int min = g;  
11        int count = 0;  
12        int sum = 0;  
13        for(int i = 0; i < n; i++){  
14            int x = sc.nextInt();  
15            sum += x;  
16            if(min > x){  
17                min = x;  
18            }  
19            sum = g - sum;  
20            while(sum - min >= 0){  
21                sum -= min;  
22                count++;  
23            }  
24            System.out.println(n + count);  
25        }  
26    }
```

Purpose-built foundation models with quality at the core drives better performance and more efficiency

Problem level strategy (zero-shot scenario) for COBOL-Java translation



Reinventing how work gets done | Application modernization solution

Accelerate hybrid cloud automation

Red Hat Ansible Lightspeed with IBM Watson Code Assistant

Watson Code Assistant (commercially available as Ansible Lightspeed from Red Hat) being developed in collaboration with Red Hat and IBM Research, infuses generative AI capabilities into Red Hat Ansible Automation Platform. It simplifies the process of Ansible playbook creation through "Pair Programming" with an AI in the "navigator" seat, providing AI-generated recommendations.

Generative AI + 

=



Open-Source IT Automation Tool

Automates IT infrastructure management and application deployment

Uses simple, **human-readable YAML** language for configuration and automation tasks

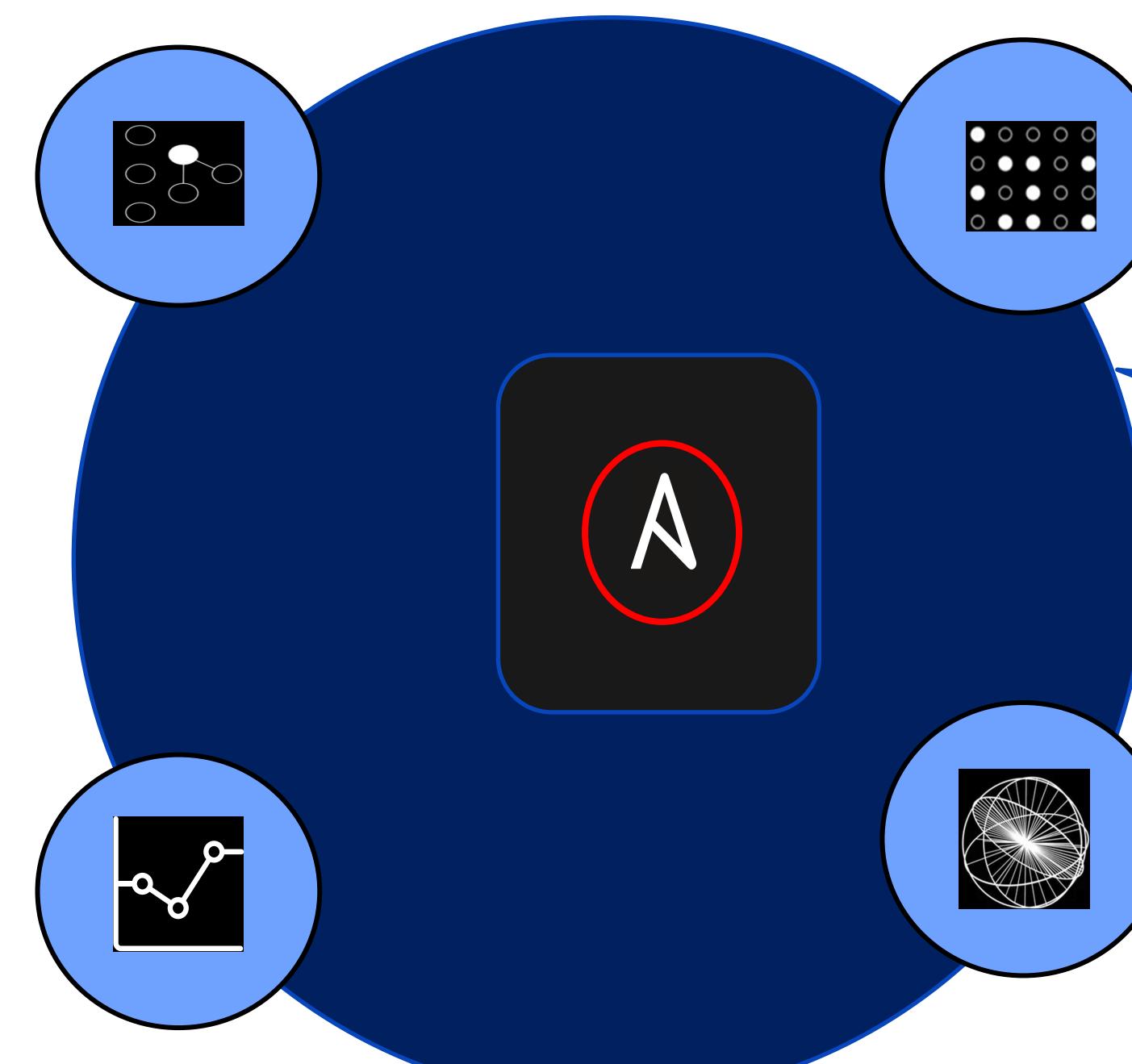
Scalable for managing hundreds or thousands of servers from a single control machine

Orchestrates **complex workflows**, including configuration management and application deployment

Agentless architecture means no software installation on target machines, making setup and use easy.

Playbook Generation
‘Generate a playbook or role’

Code Optimization
‘Review my Playbook and help me make it better’



Content Discovery
‘Find me a playbook or role similar to what I am writing’

Red Hat & IBM curated Ansible content

Code Explanation
‘Tell me what this Playbook is doing and its impact’; ‘Tell me where the generated code came from’

```
1 - name: Add user to z/OS system
2   hosts: all
3   gather_facts: false
4   environment: "{{ environment_vars }}"
5
6 tasks:
```



watsonx.ai drives new IBM innovation and Red Hat OpenShift enhancements for AI workloads

