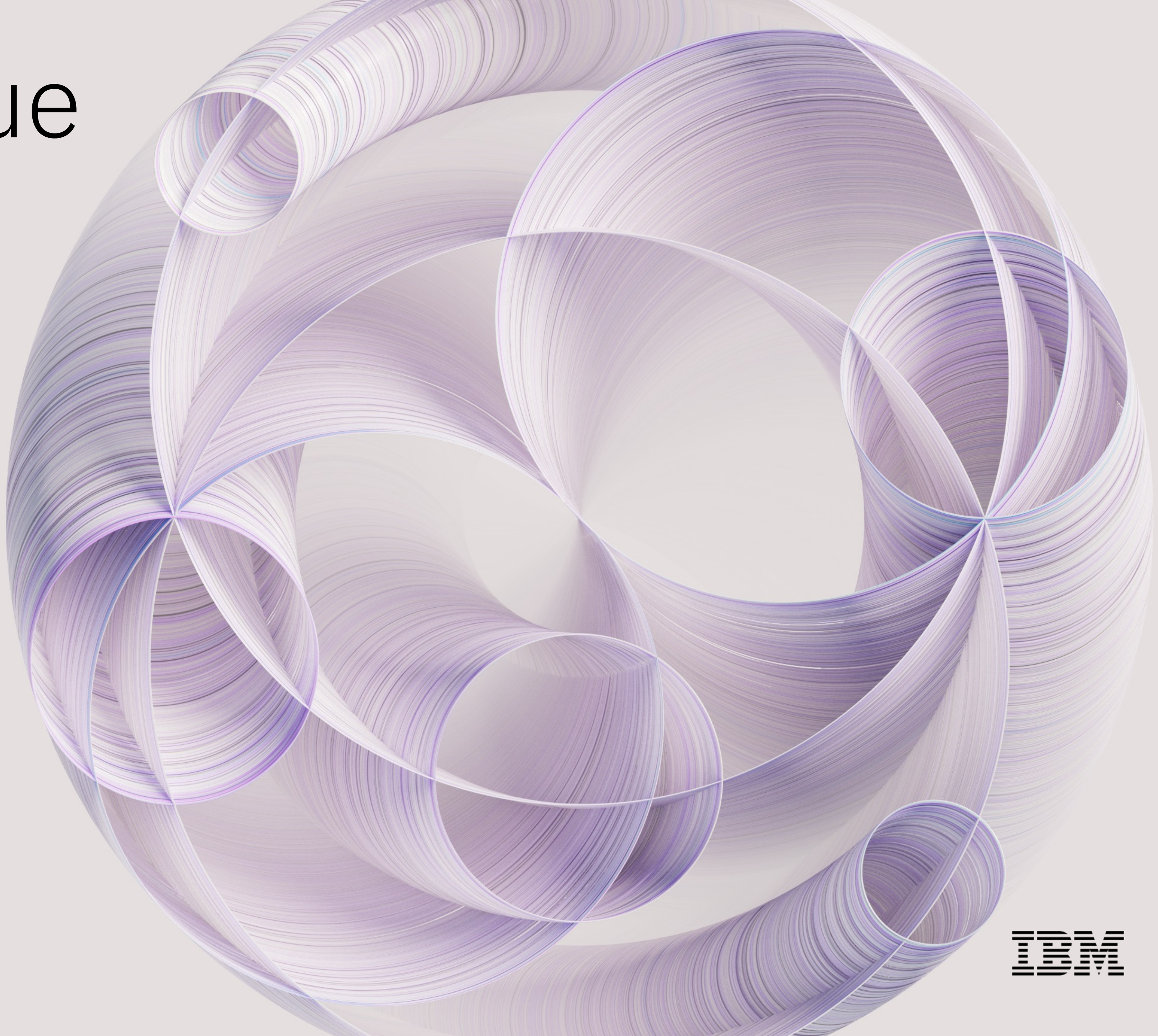
Scale the value from AI with trusted data



Amith Singhee, Ph.D.
Director, IBM Research India
CTO, IBM India/South Asia

watsonx

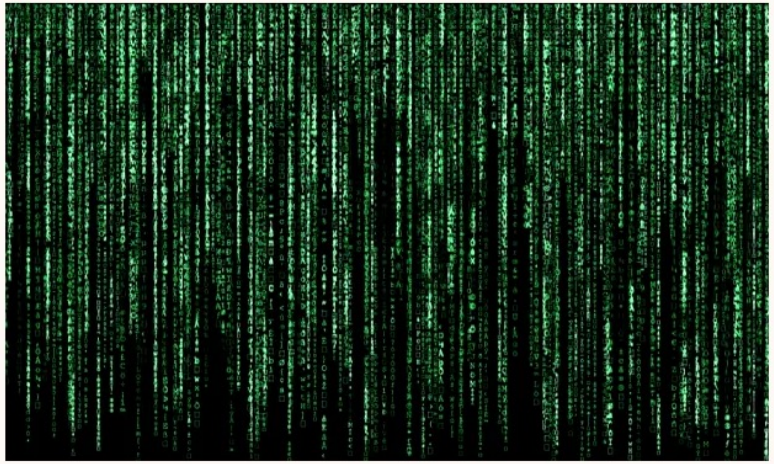
IBM

Foundation models are driving state of the art NLP

A robot wrote this entire article. Are you scared yet, human?

We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below



I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

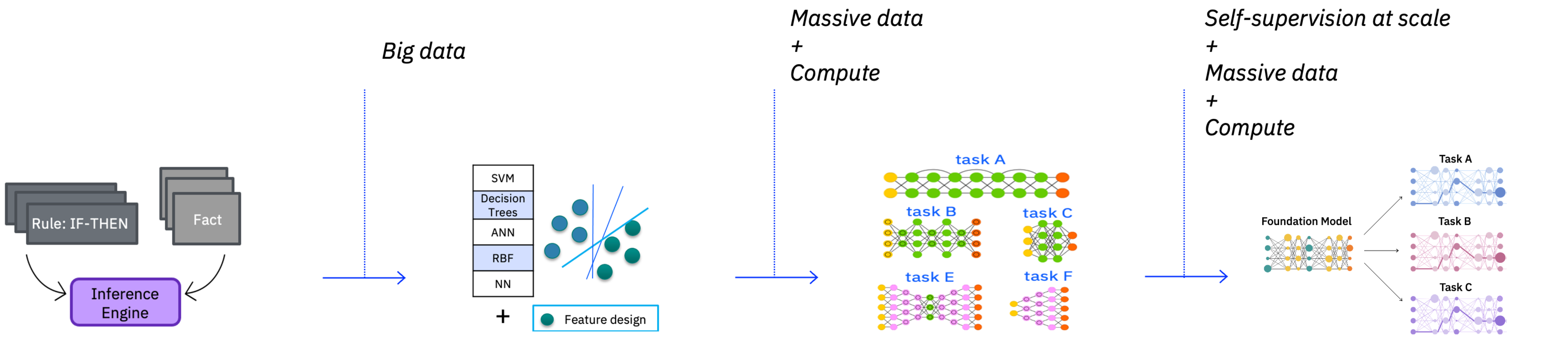
IBM

Foundation models are already part of our commercial capabilities and research tech

- Conversational systems
- Language & Document Understanding
- NLP Leaderboards

#1 in TiDy (multilingual question answering)
#1 in Xor-TIDY (cross lingual QA)
#3 in Natural Language Queries (QA)
#1 in Wizard of Wikipedia/KILT
(content grounded dialog)
#1 in Fact Checking/KILT
#1 in Table Question Answering
#1 in DREAM (multiple choice questions for dialog)
#1 in Switchboard 500 (English Speech to Text)

Foundation models are poised to dramatically accelerate enterprise AI adoption



Expert Systems

- No use of data
- Manually authored rules
- Brittle

Machine Learning

- Less brittle but labor intensive
- Demanding data prep and feature engg.

Deep Learning

- Automatically learn if you have enough labeled data
- Enterprise adoption limited by availability of labeled data

Foundation Models

- Learn from lots of data without requiring labels
- Adapt quickly to many tasks
- Accelerate enterprise adoption

Foundation Models are driving a fundamental change in AI methodology and operations.

Less labeling means less effort and lower upfront costs

Effort mostly on fine tuning and inferencing means faster deployment

Equal or better accuracy than state-of-the-art for multiple use cases

Better performance means incremental revenue

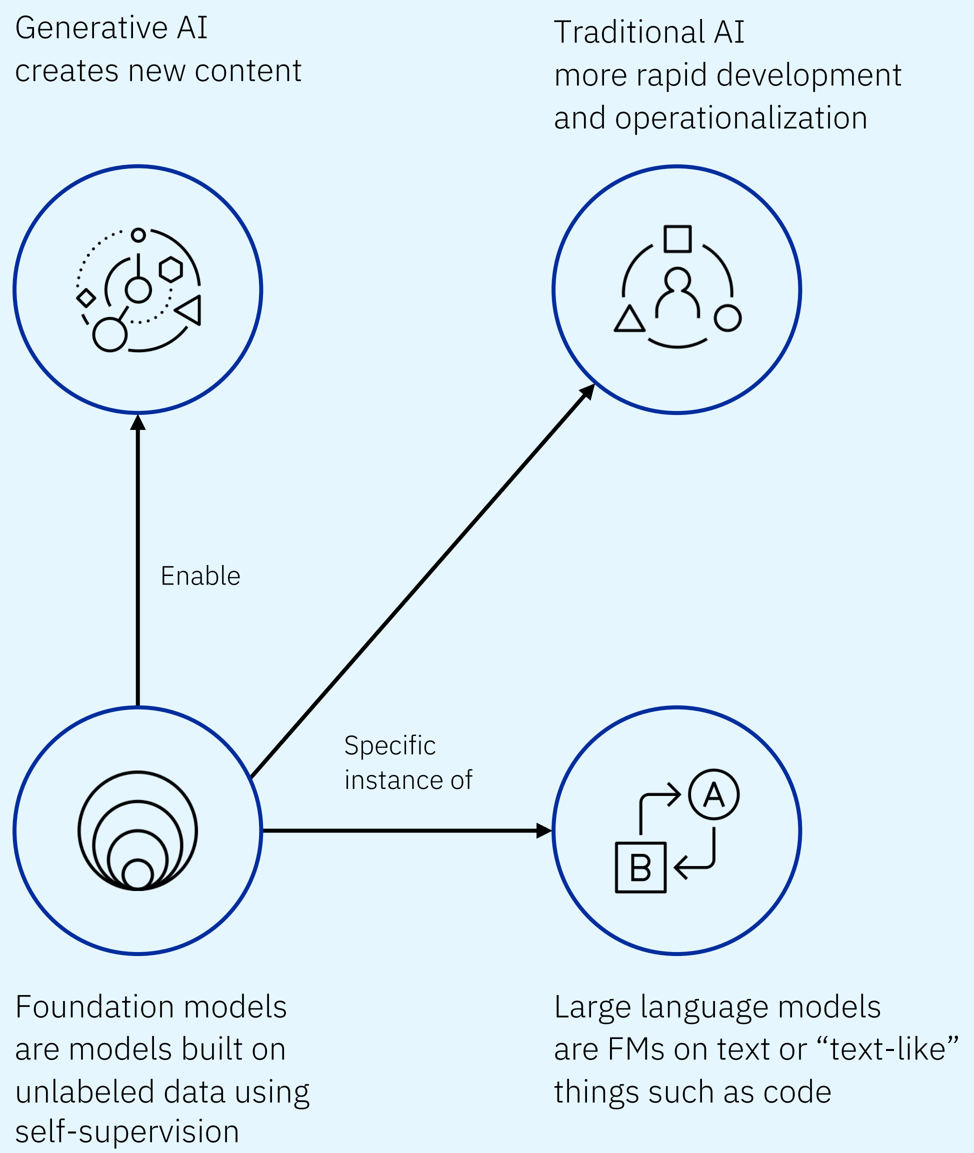
10-100x

decrease in labeling requirements

6X

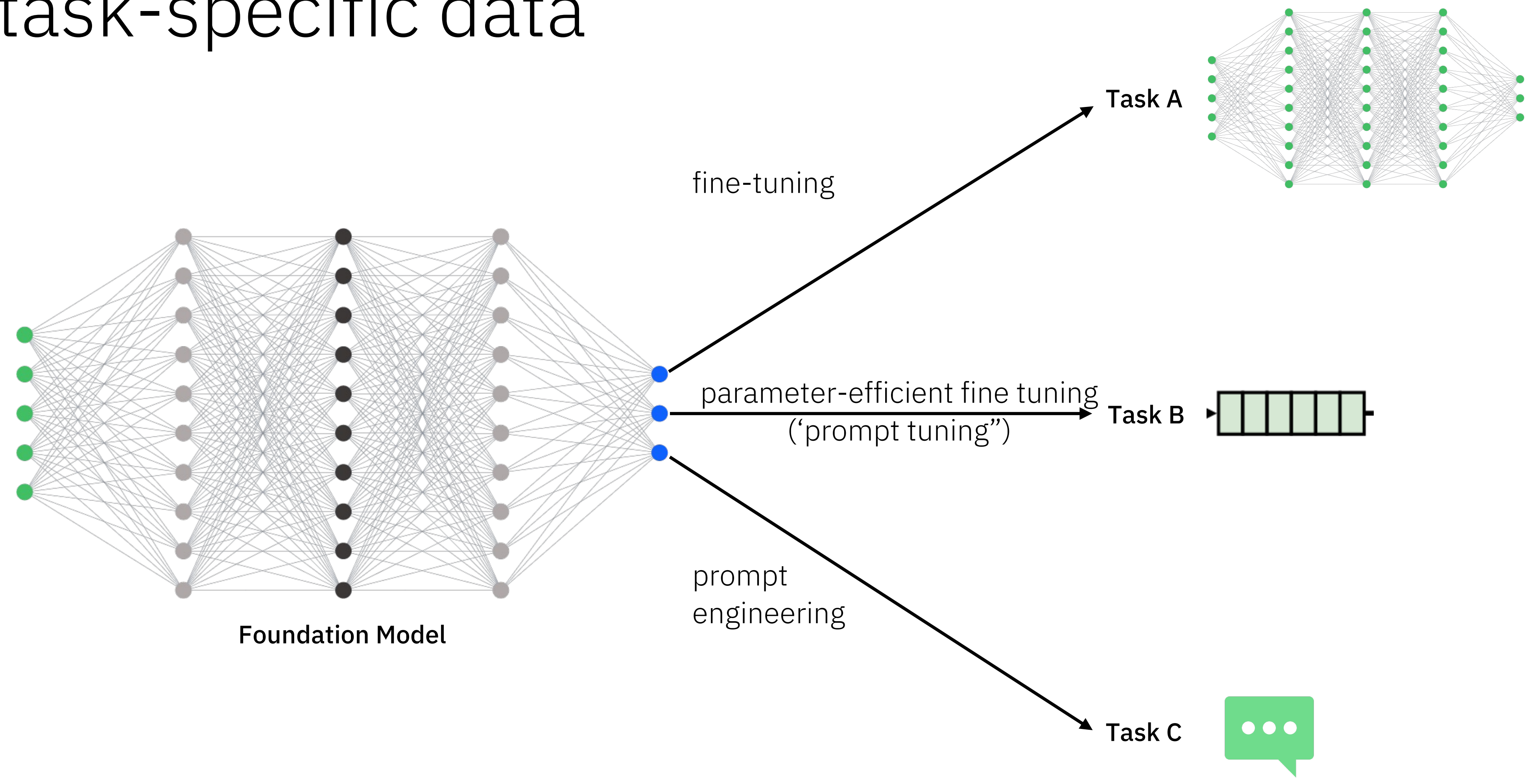
decrease in training time

Foundation models, large language models, and generative AI



Rapid adaptation to multiple tasks with small amounts of task-specific data

A foundation model is an AI model trained on large amounts of unlabeled data that can be adapted easily to new use cases.



Sample HR Policies

DEMO
Conversational AI

Collection

HR Collection 1

Start a chat

HR Collection 1 / Vacation

Sample Benefits Pages

Vacation

IBM offers a competitive vacation plan to all regular full-time employees and regular IBM employees who work alternative work schedules. All regular full-time IBM employees receive a minimum of 15 days paid vacation each year. Your vacation time is based on years of service and type of employment.

Full time employees

- Full-time employees with 0-9 years of service are entitled to a maximum of 15 days of vacation per year.
- Full-time employees with 10-20 years of service are entitled to a maximum of 20 days of vacation per year.
- Full-time employees with more than 20 years service are entitled to a maximum of 25 days days of vacation per year.

Full-time employees are eligible for 20 days of vacation in the year of your 10th anniversary with IBM. For example, if your first date of full-time IBM employment was September 1, 1999, you would be eligible for 20 days of vacation beginning in 2009. You may take your full amount of vacation at any time during this year, but your full amount (that is, 20 days), is not earned until the end of the year.

Supplemental employees

- Supplemental employees with 0-9 years of service are entitled to a maximum of 10 days of vacation per year.
- Supplemental employees with 10-20 years of service are entitled to a maximum of 15 days vacation per year.
- Supplemental employees with more than 20 of service are entitled to a maximum of 20 days vacation per year.

Earning Vacation

Vacation can be taken at any time of the year (based on business requirements and your personal preference), but the **full amount** isn't earned until year-end. You can take vacation time in weeks, days or half days. You may take your accrued vacation before you have earned it. However, vacation taken before it's earned is considered a salary advance.

If you separate from IBM before you earn the vacation you have taken, IBM will seek reimbursement for the value of the unearned days.

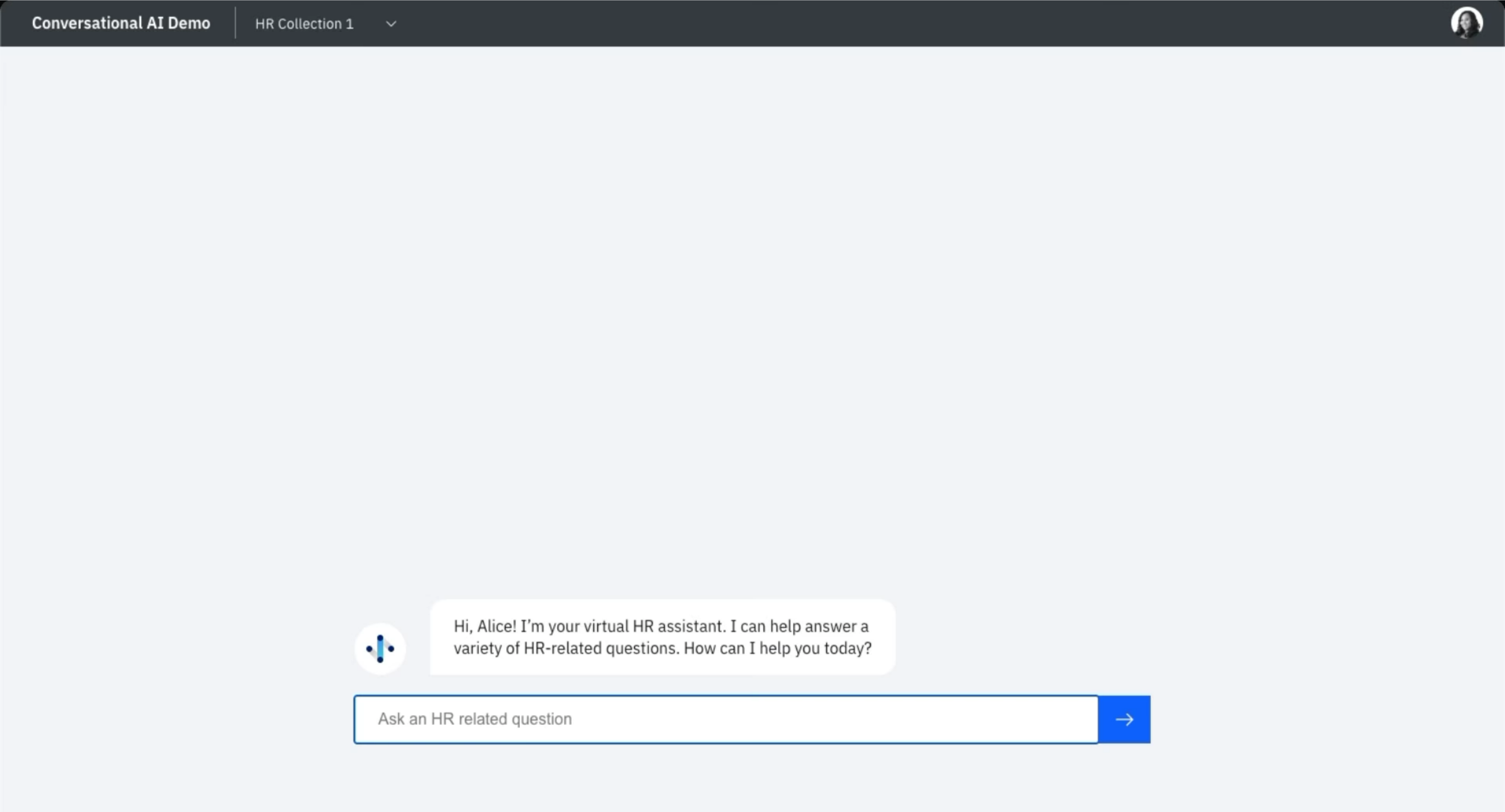
Vacation Pay

Vacation pay is at the same rate as an employee's regular salary or hourly pay. All vacation pay is subject to tax withholding. Employees don't have the option of choosing cash payments over taking vacation time from work.

Eligibility in year of hire: In the year of hire, rehire, or conversion from supplemental to regular employment, vacation is based on the number of days worked during the year.

Flexible work week schedules: **Employees on flexible work week schedules** earn vacation based on hours worked per week and years of service.

1 / 4


Conversational AI Demo
HR Collection 1
Hi, Alice! I’m your virtual HR assistant. I can help answer a variety of HR-related questions. How can I help you today?
Ask an HR related question

Conversational AI Demo | HR Collection 1

Hi, Bob! I'm your virtual HR assistant. I can help answer a variety of HR-related questions. How can I help you today?

Ask an HR related question

Full time employees

- Full-time employees with 0-9 years of service are entitled to a maximum of 15 days of vacation per year.
- Full-time employees with 10-20 years of service are entitled to a maximum of 20 days of vacation per year.
- Full-time employees with more than 20 years service are entitled to a maximum of 25 days days of vacation per year.

Supplemental employees

- Supplemental employees with 0-9 years of service are entitled to a maximum of 10 days of vacation per year.
- Supplemental employees with 10-20 years of service are entitled to a maximum of 15 days vacation per year.
- Supplemental employees with more than 20 of service are entitled to a maximum of 20 days vacation per year.

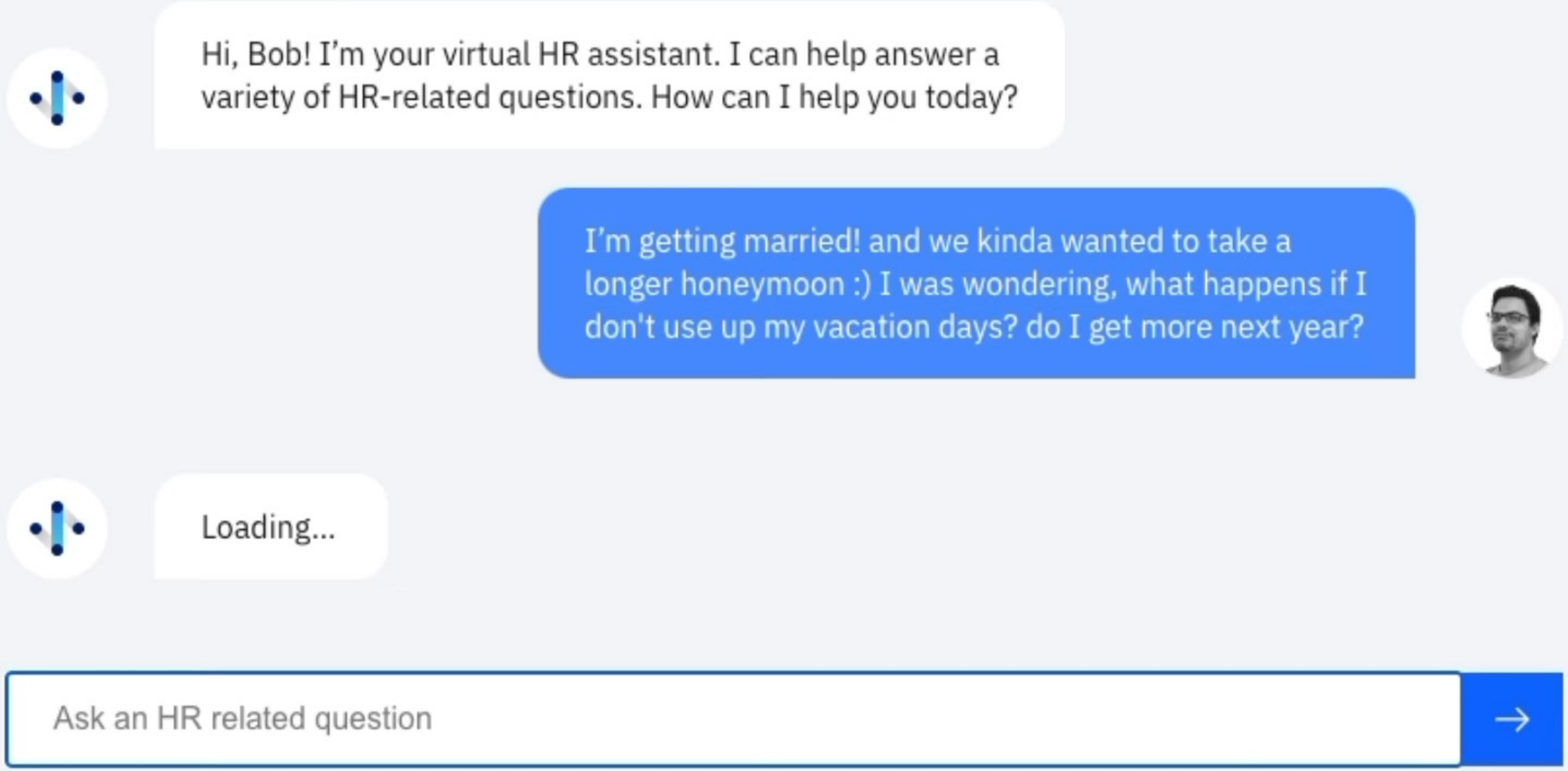
Earning Vacation

Vacation can be taken at any time of the year (based on business requirements and your personal preference), but the **full amount** isn't earned until year-end. You can take vacation time in weeks, days or half days. You may take your accrued vacation before you have earned it. However, vacation taken before it's earned is considered a salary advance.

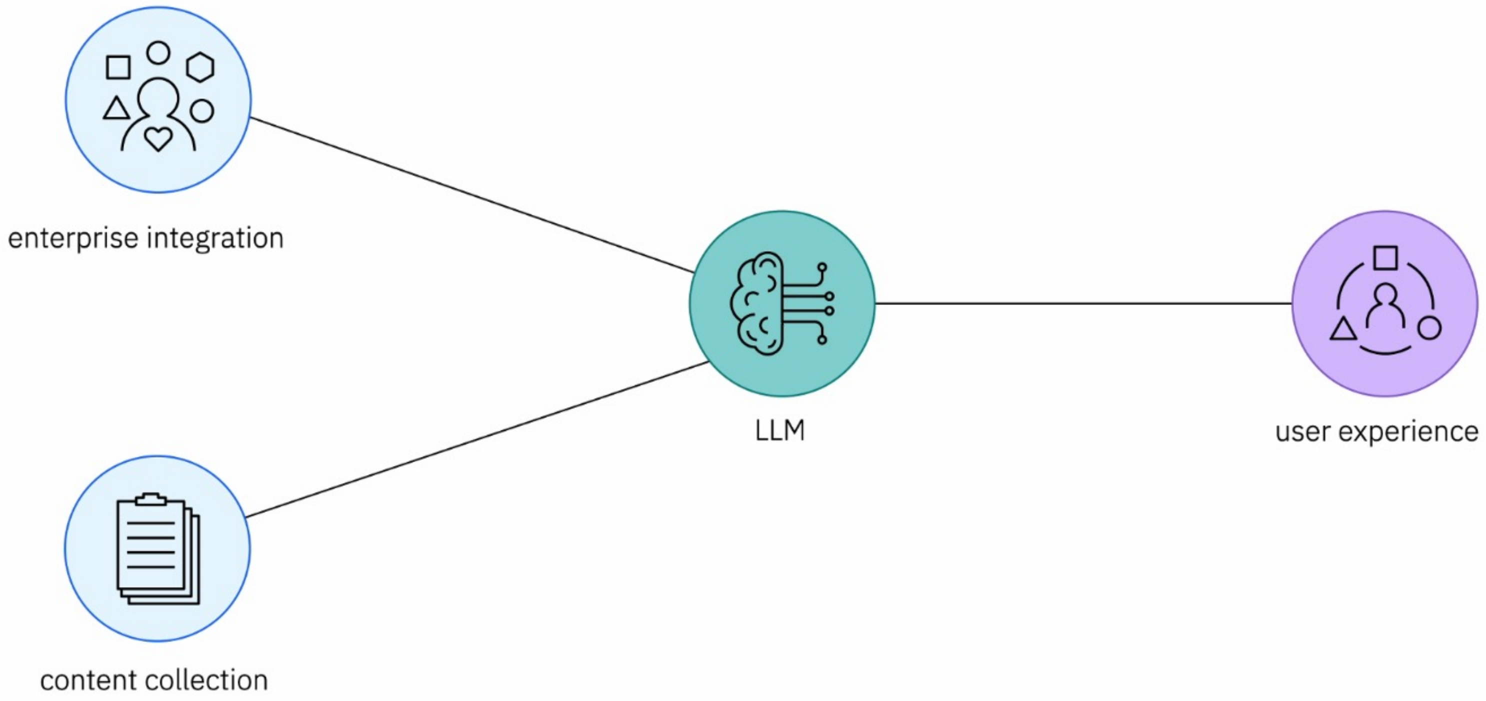
Vacation Pay

Vacation pay is at the same rate as an employee's regular salary or hourly pay. All vacation pay is subject to tax withholding. Employees don't have the option of choosing cash payments over taking vacation time from work.

Eligibility in year of hire: In the year of hire, rehire, or conversion from supplemental to regular employment, vacation is based on the number of days worked during the year.

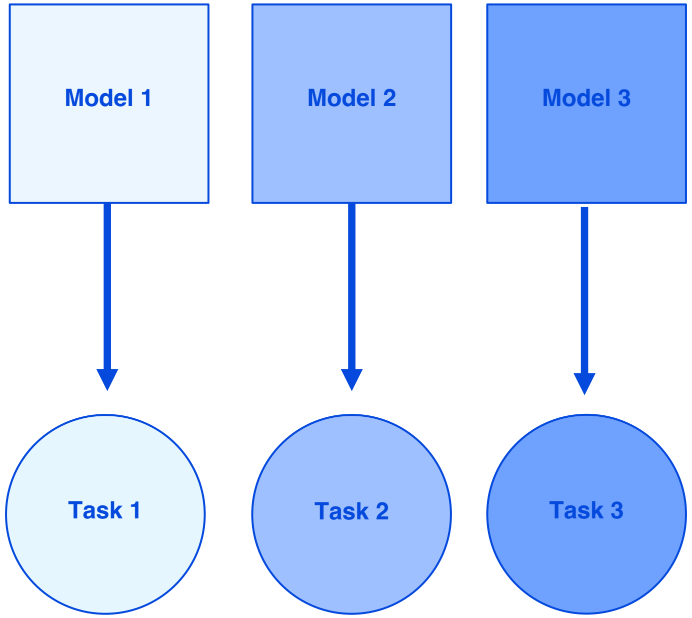

Hi, Bob! I'm your virtual HR assistant. I can help answer a variety of HR-related questions. How can I help you today?
I'm getting married! and we kinda wanted to take a longer honeymoon :) I was wondering, what happens if I don't use up my vacation days? do I get more next year?
Loading...
Ask an HR related question

Content-Grounded Conversational AI



Traditional AI Models

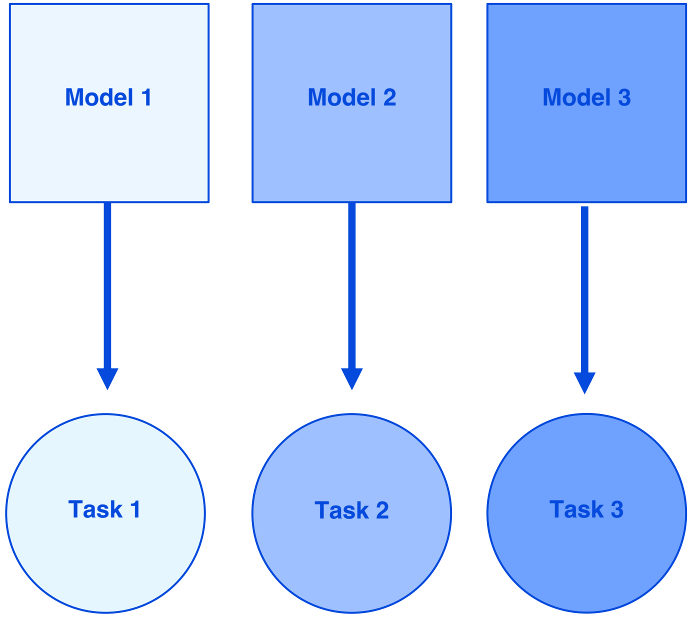
Each model is trained for a specific task



1,000s to 1,000,000s labeled data points per task

Traditional AI Models

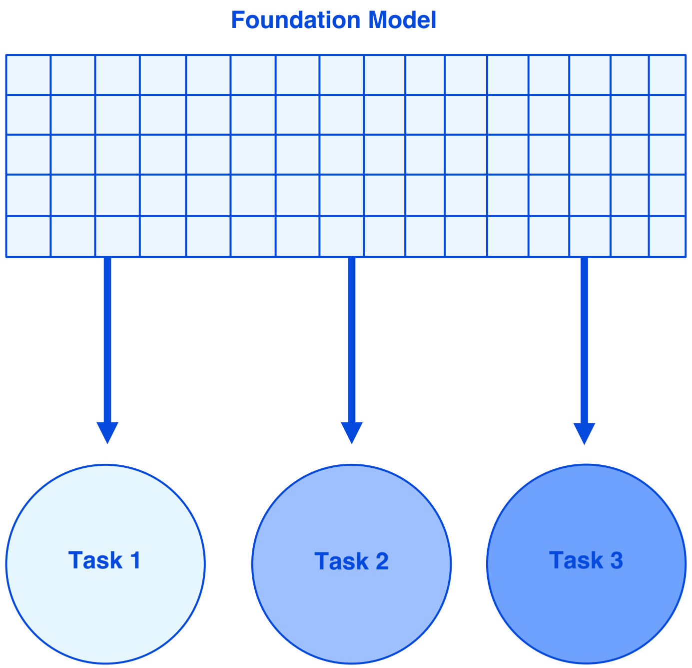
Each model is trained for a specific task



1,000s to 1,000,000s labeled data points per task

Foundation Models

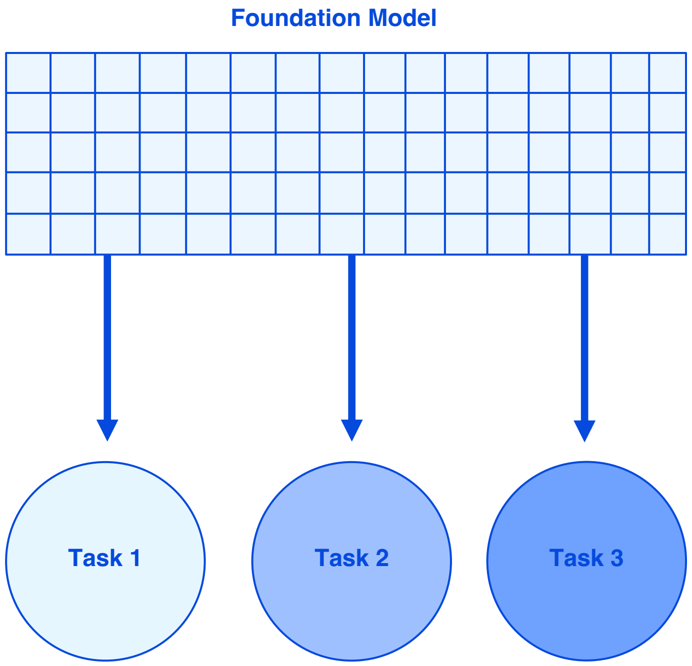
One model that can address many tasks...



0 to 1,000s labeled data points per task

Foundation Models

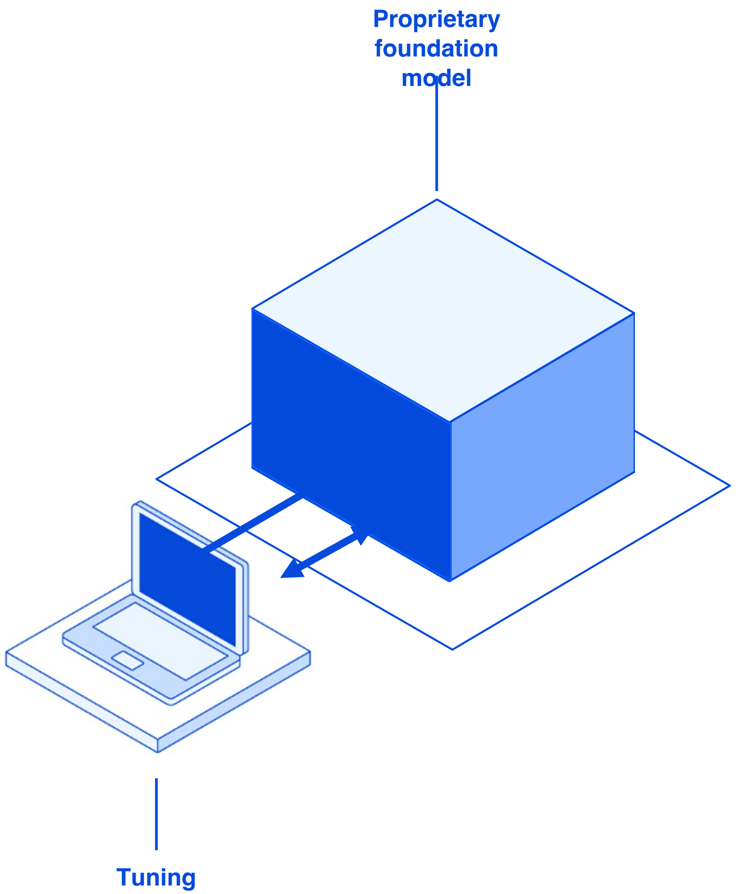
One model that can address many tasks...



0 to 1,000s labeled data points per task

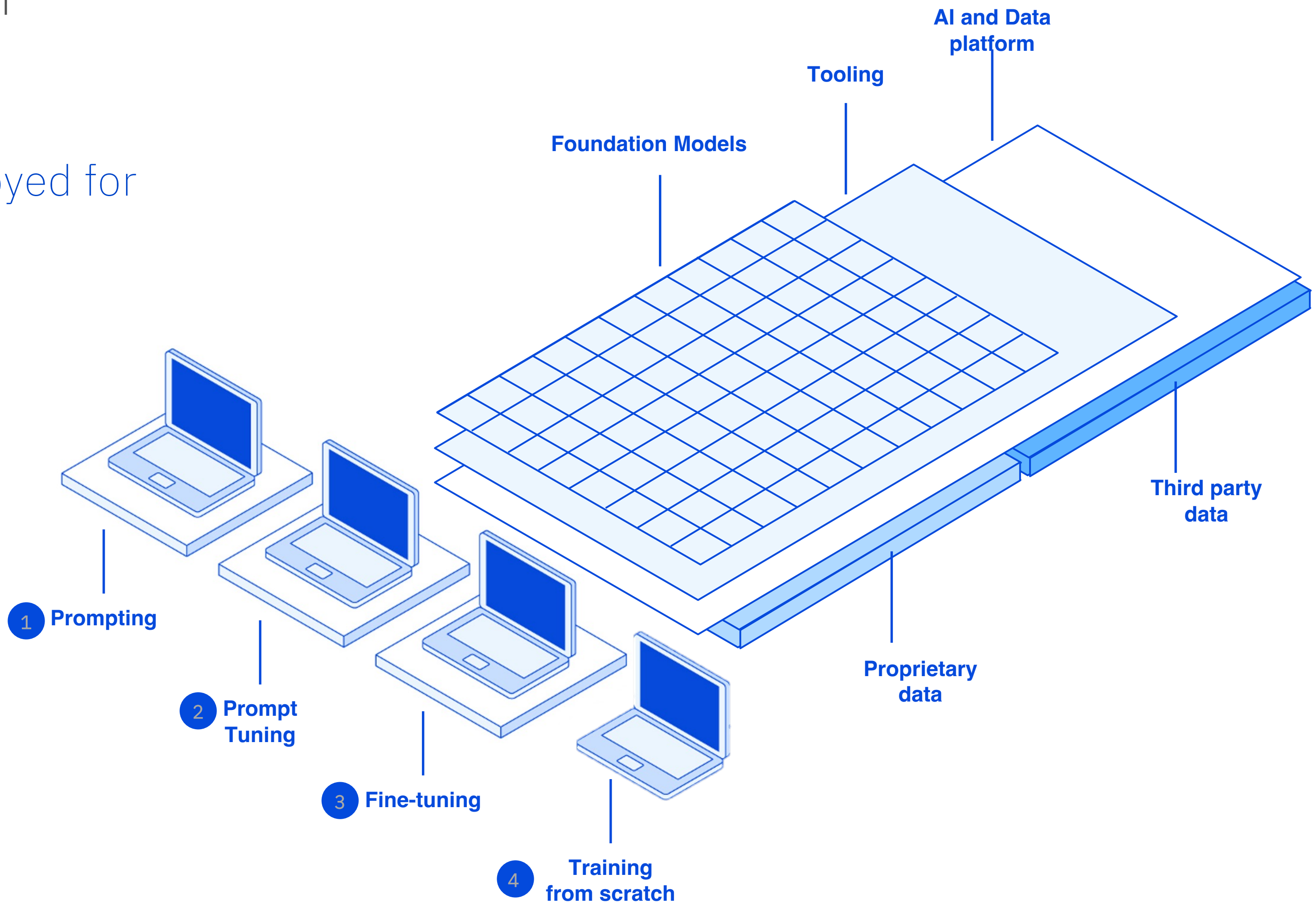
Prompting Foundation Models

...be customized further...



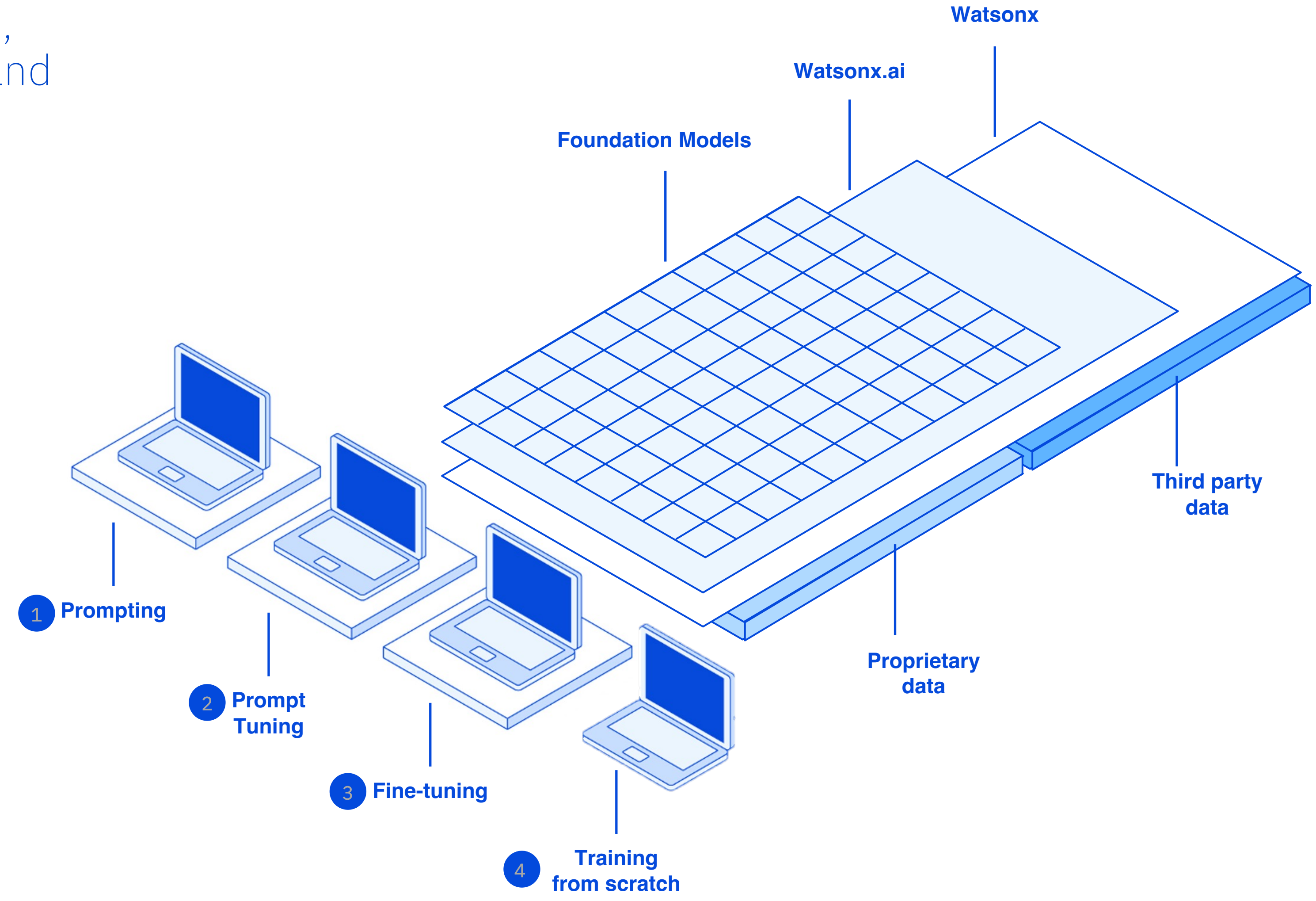
Creating Value with Foundation Models

...and fully deployed for enterprise use



watsonx

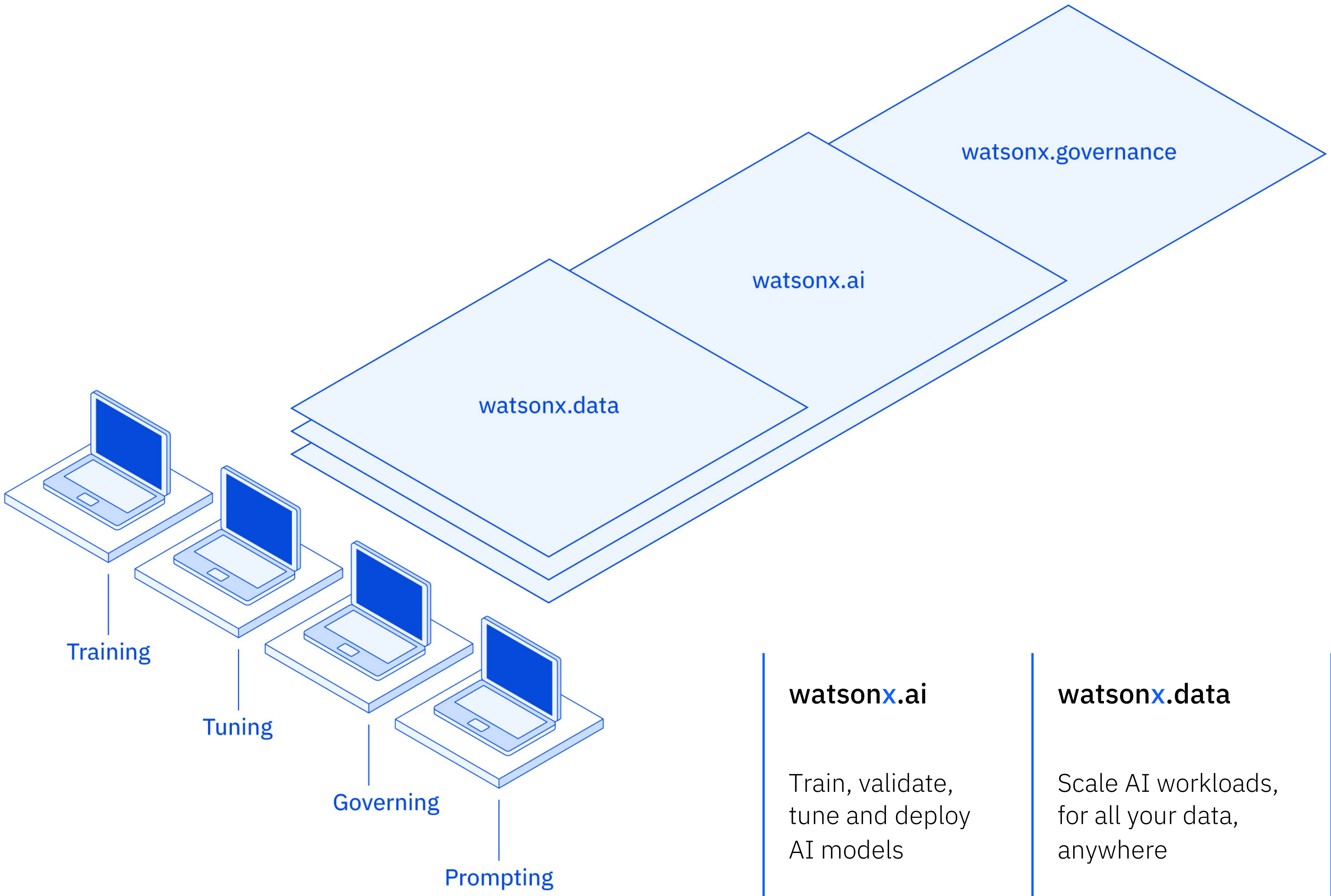
Next generation enterprise studio for AI builders to train, validate, tune, and deploy models



watsonx

The platform for AI and data

Scale and accelerate the impact of AI with trusted data.



watsonx.ai

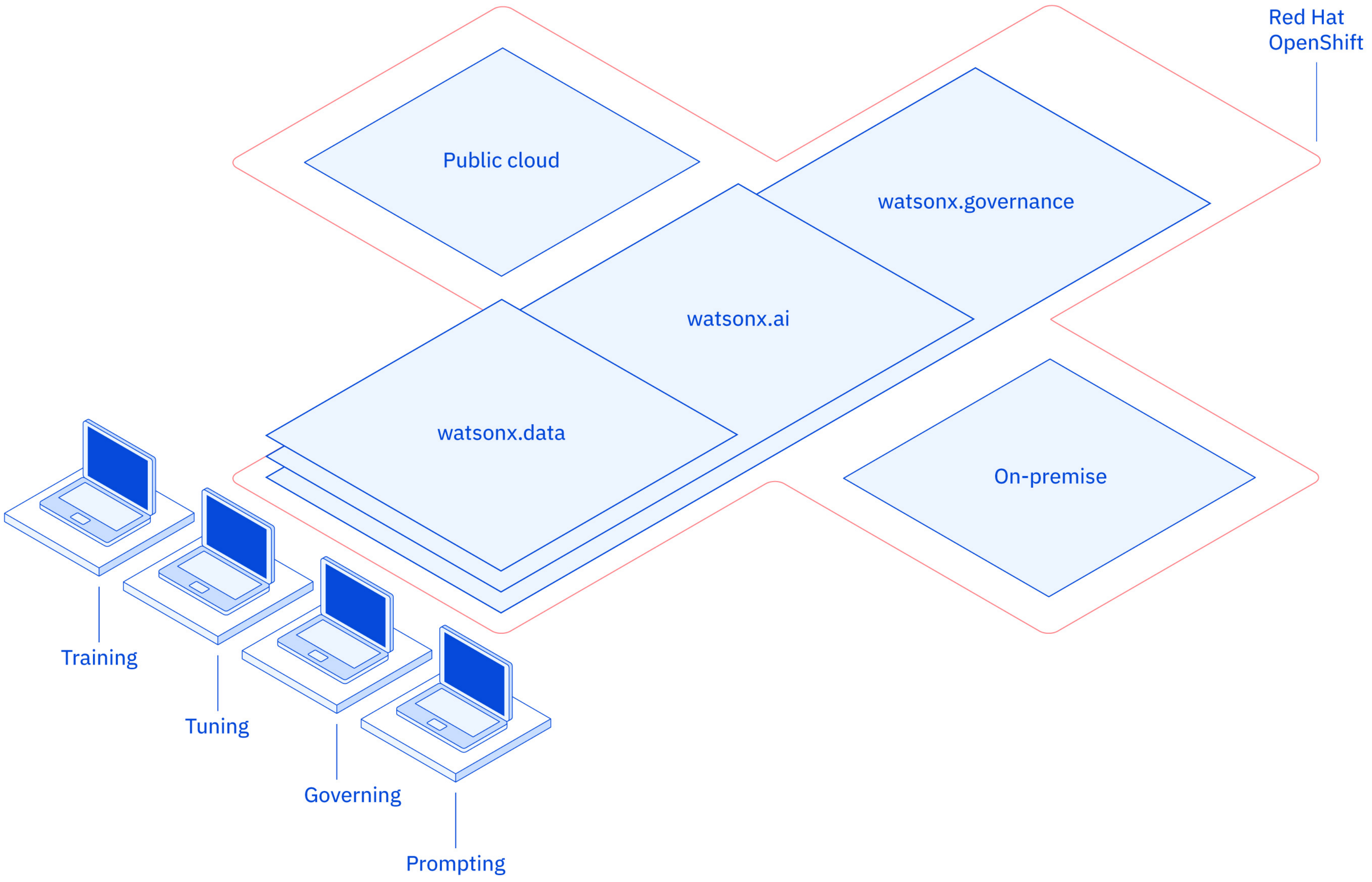
Train, validate, tune and deploy AI models

watsonx.data

Scale AI workloads, for all your data, anywhere

watsonx.governance

Enable responsible, transparent and explainable AI workflows


Red Hat
OpenShift
Public cloud
watsonx.governance
watsonx.ai
watsonx.data
On-premise
Training
Tuning
Governing
Prompting

Example use cases

1 **Summarization**
Transform text with domain-specific content into personalized overviews, capturing key points.
E.g., sales conversation summaries, insurance coverage, meeting transcripts, and contract information

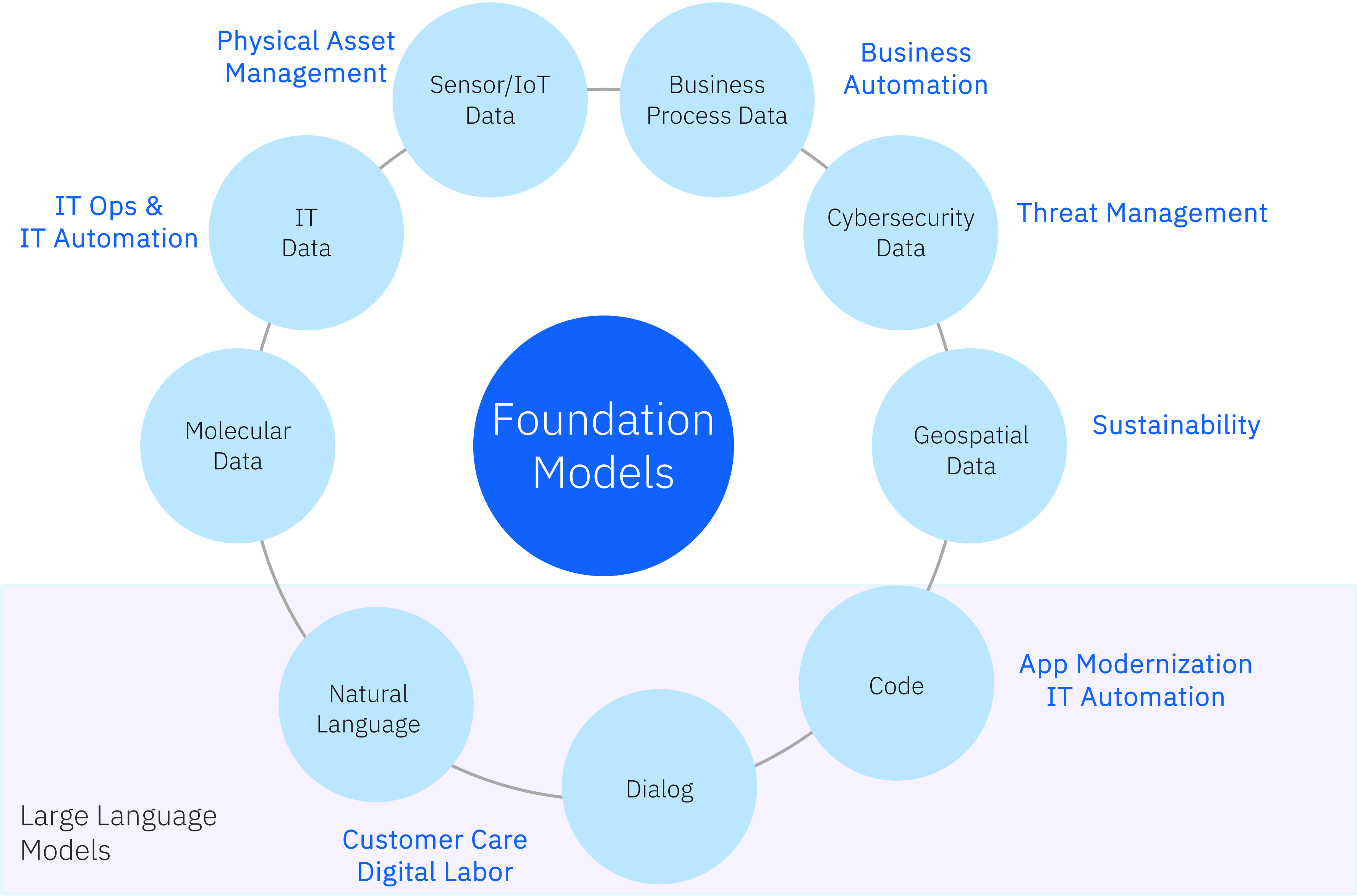
2 **Classification**
Read and classify written input with as few as zero examples
E.g., sorting of customer complaints, threat & vulnerability classification, sentiment analysis, and customer segmentation

3 **Content Generation**
Generate text content for a specific purpose.
E.g., content creation for marketing campaigns, job descriptions, blog posts and articles, and email drafting support

4 **Extraction**
Analyze and extract essential information from unstructured text.
E.g., audit acceleration, SEC 10K fact extraction, user research findings

5 **Question Answering**
Create a question-answering feature grounded on specific content.
E.g., build a product specific Q&A resource for customer service agents.

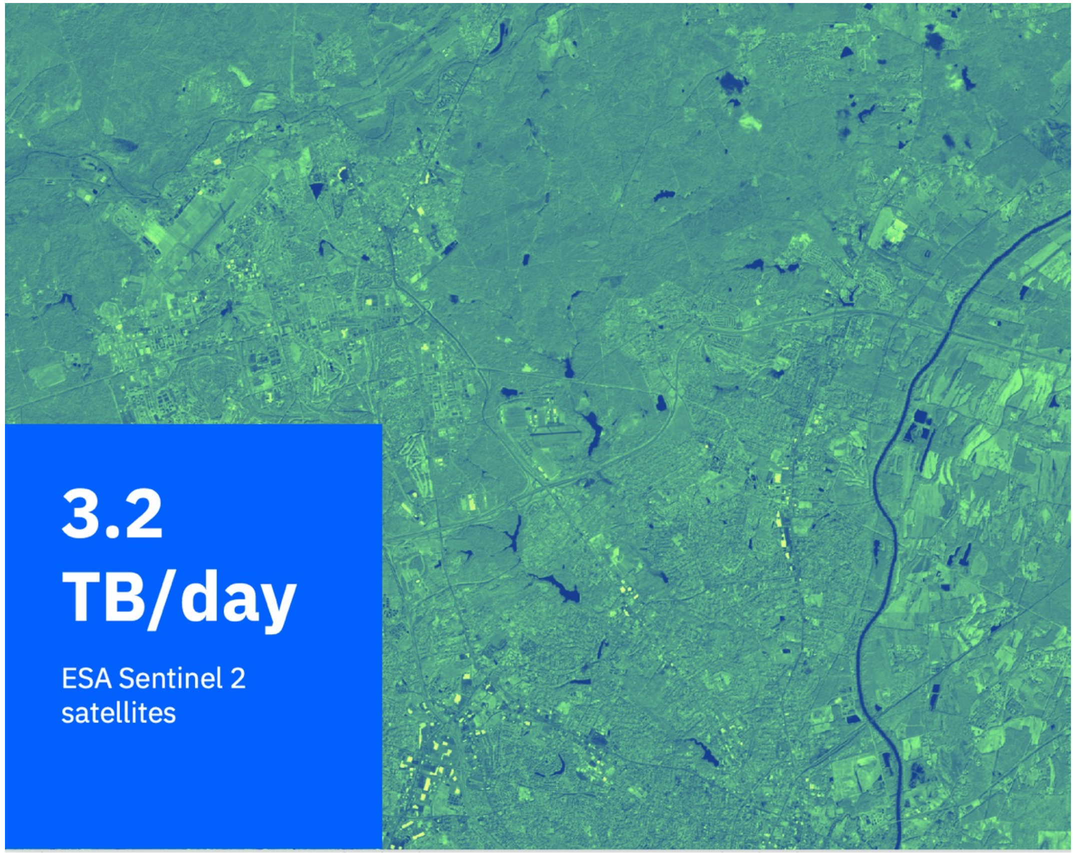
Foundation Models for Business: over a variety of data modalities



Two core types of geospatial data relevant for geospatial and sustainability

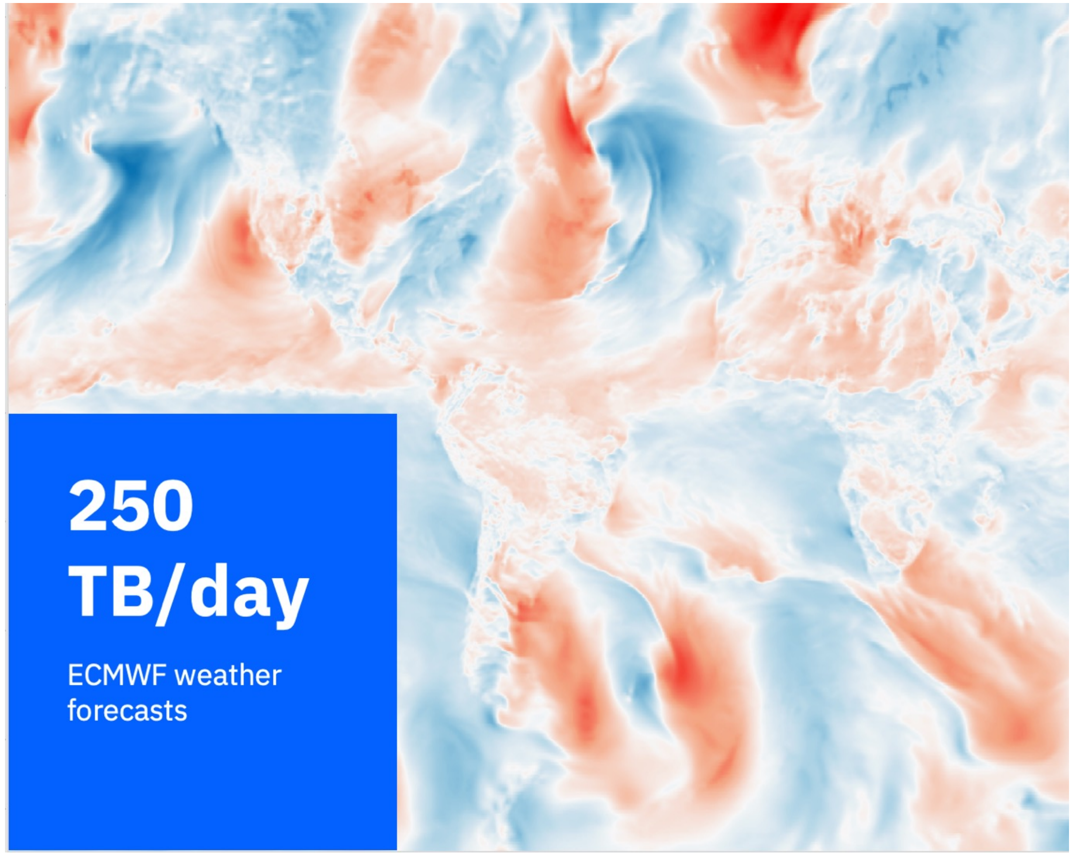
Satellite and aerial imagery

Multimodal – images from multiple satellites (HLS2) representing different spectral bands

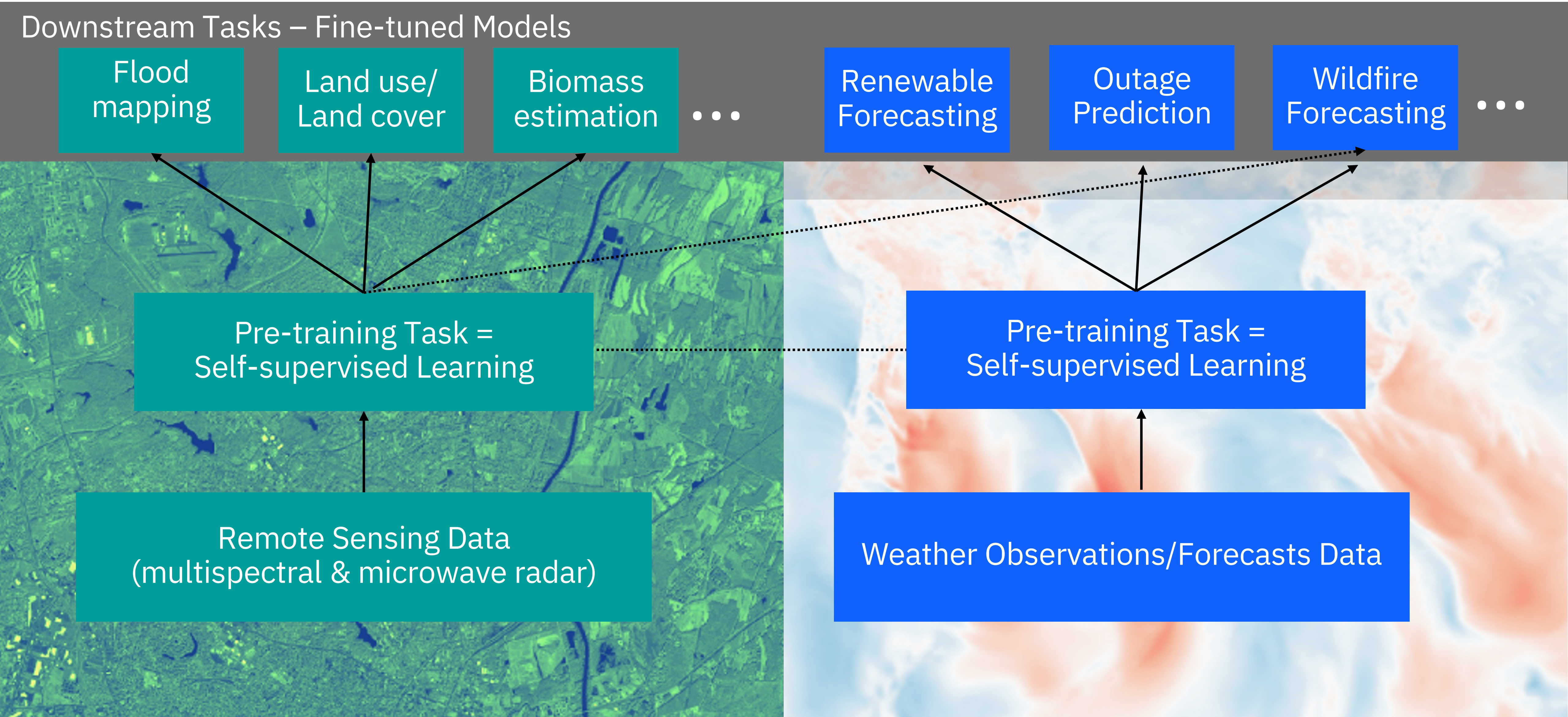


Weather measurements and forecasts

Multimodal – time series from different processes (temperature, precipitation, wind,...)



Geospatial Foundation Models



Model architecture

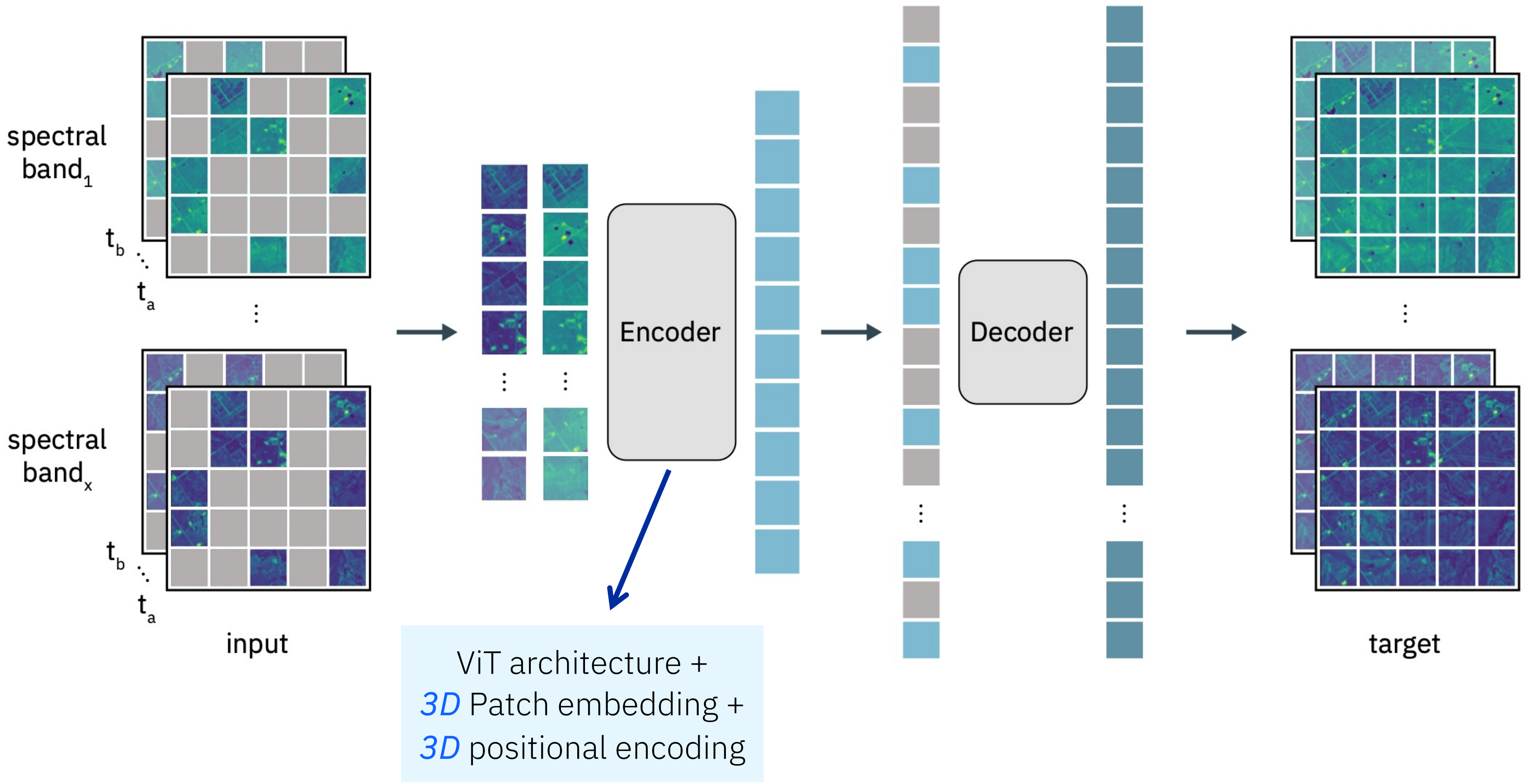
MAE → Masked AutoEncoder

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

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

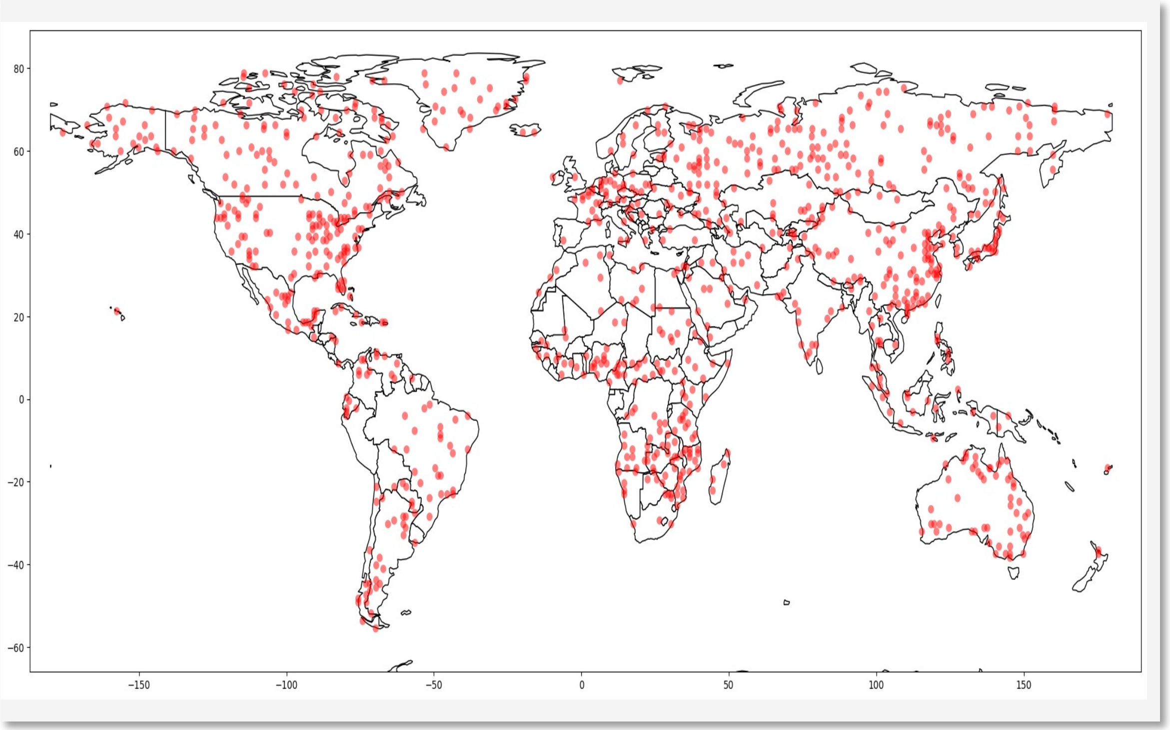
- 3D patch embeddings
- 3D positional encoding

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



Data sampling procedure

Selecting pre-training data



Requirement → *diversified* pre-training dataset.

– For a given region, images can look similar across time.

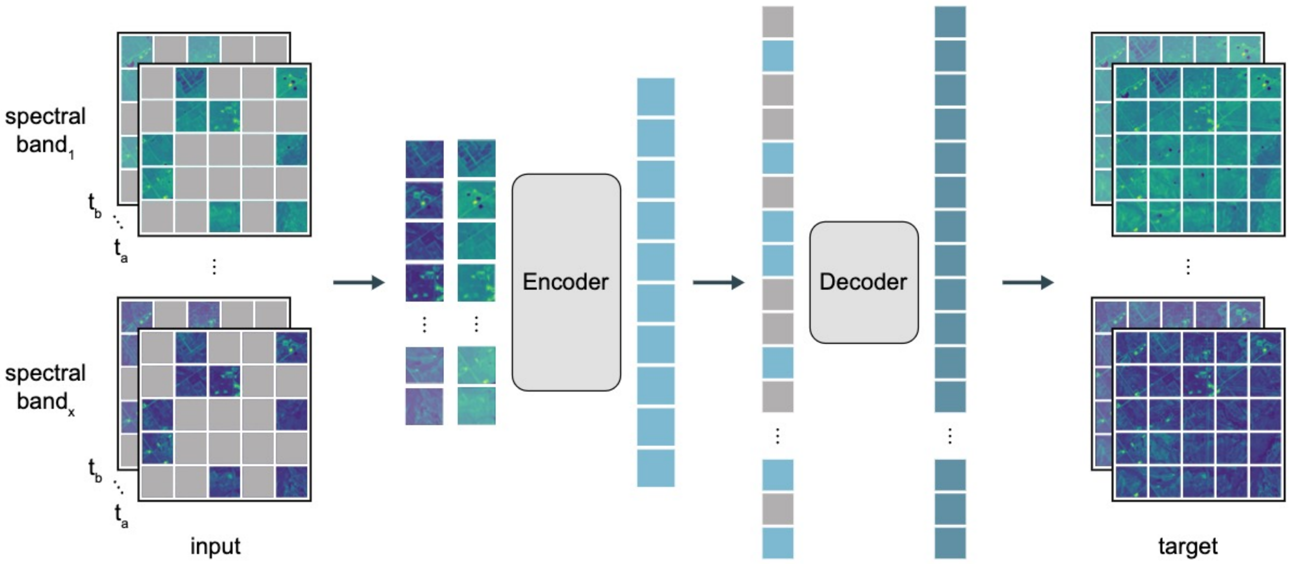
– Random sampling → can bias towards most common landscapes.

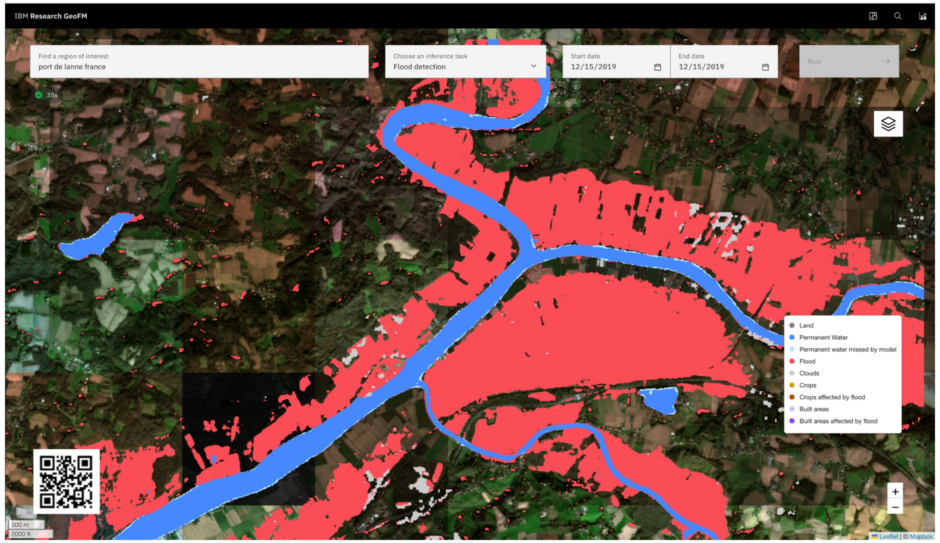
Intelligent sampling scheme based on *geospatial statistics*.

Summary: Geospatial foundation models

Prithvi is a suite of geospatial foundation models that accelerate the development of geospatial applications. Available in IBM's watsonx.ai, it includes pre-trained and fine-tuned models for disaster mapping, environmental change monitoring, and data discovery tasks (such as flood, fire scars, and land use/change), thus enabling geospatial analytics at higher accuracy, lower cost, and faster speed.

- **Pre-trained** using NASA datasets and expertise
- Leverage **self-supervised learning** (i.e., masking imagery or timeseries)
- Able to effectively complete **multiple geospatial and environmental applications** while meeting accuracy baselines (e.g., disaster response, agriculture, and climate change)

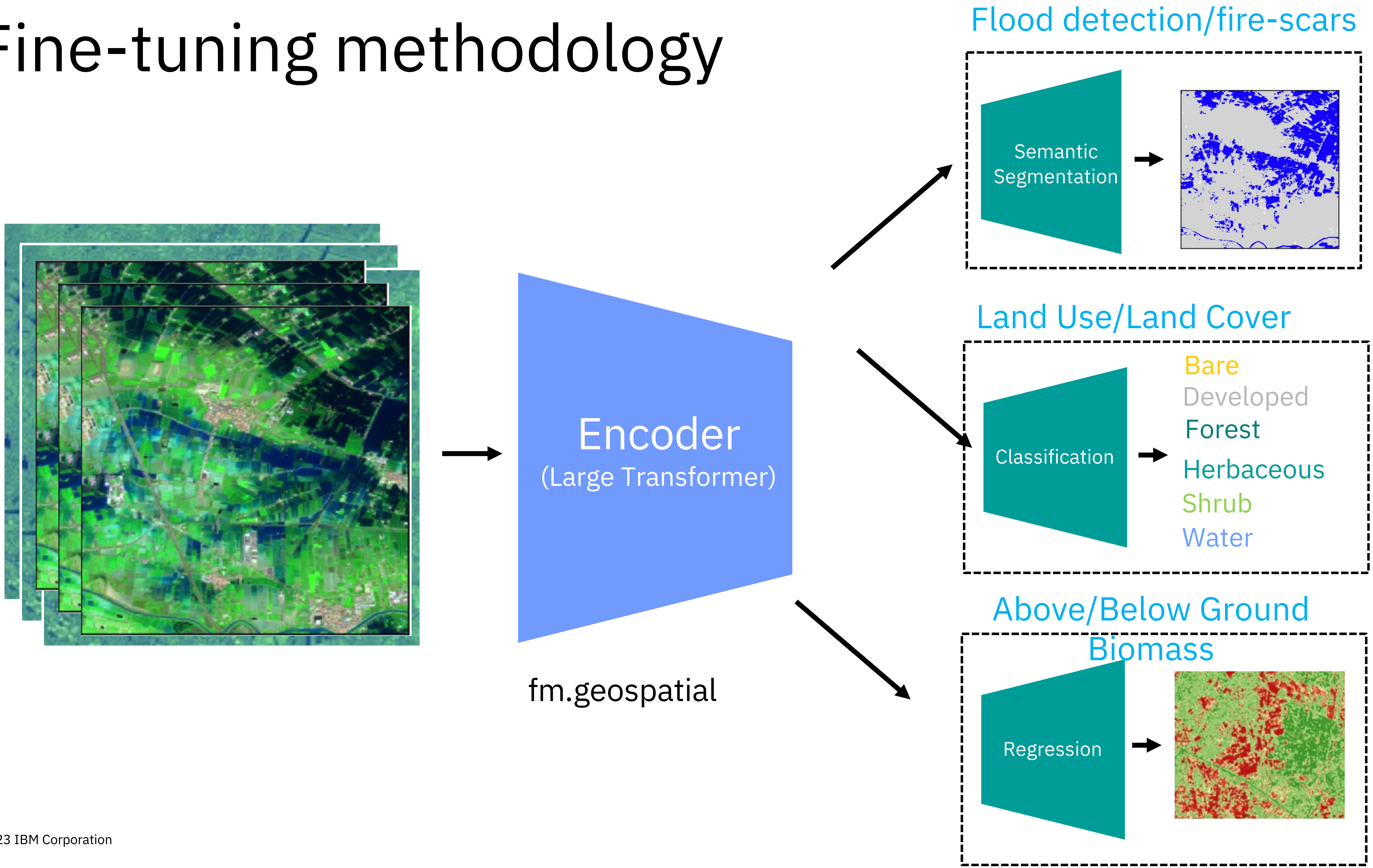




Key differentiations

- Outperforms SoTA by upto 15%
- Needs upto 50% less labeled data
- Generalizes across applications
- Pre-built workflows cut dev time from months to days

Fine-tuning methodology



Inference insights by Prithvi – Image reconstruction

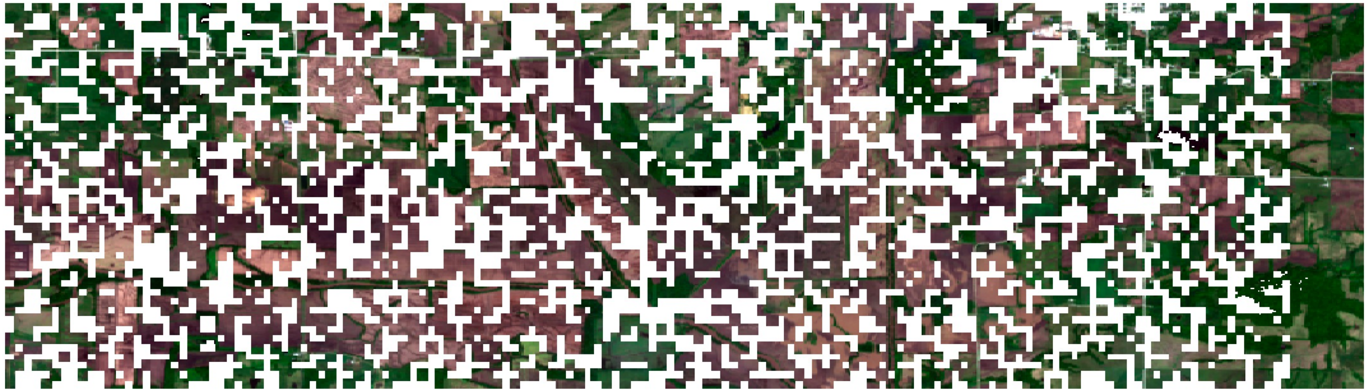
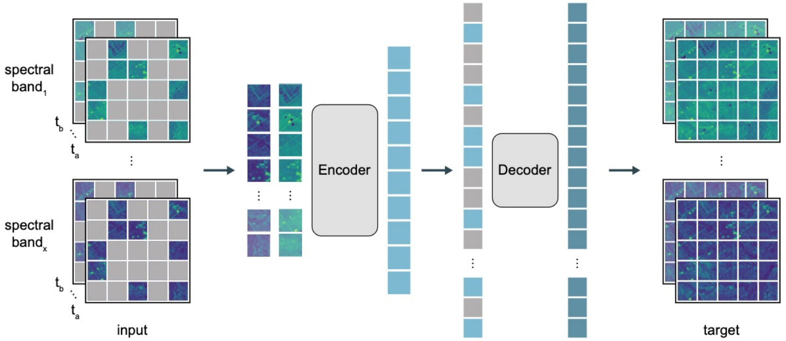


Image masking

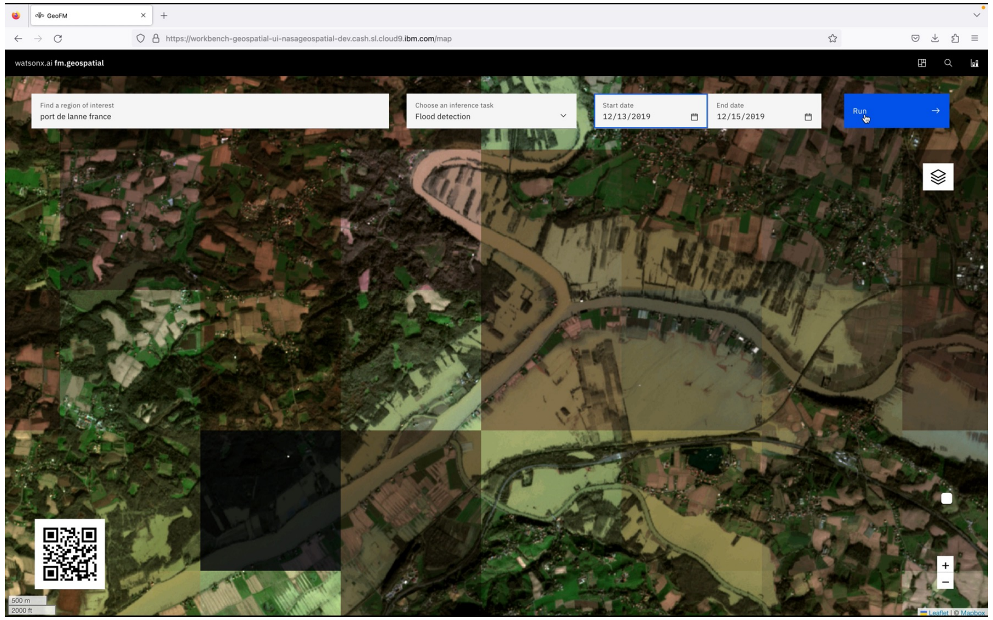
Reconstructed Image

“**Prompt**”: Image(s) (spatial + temporal domains)



GFM pre-training architecture

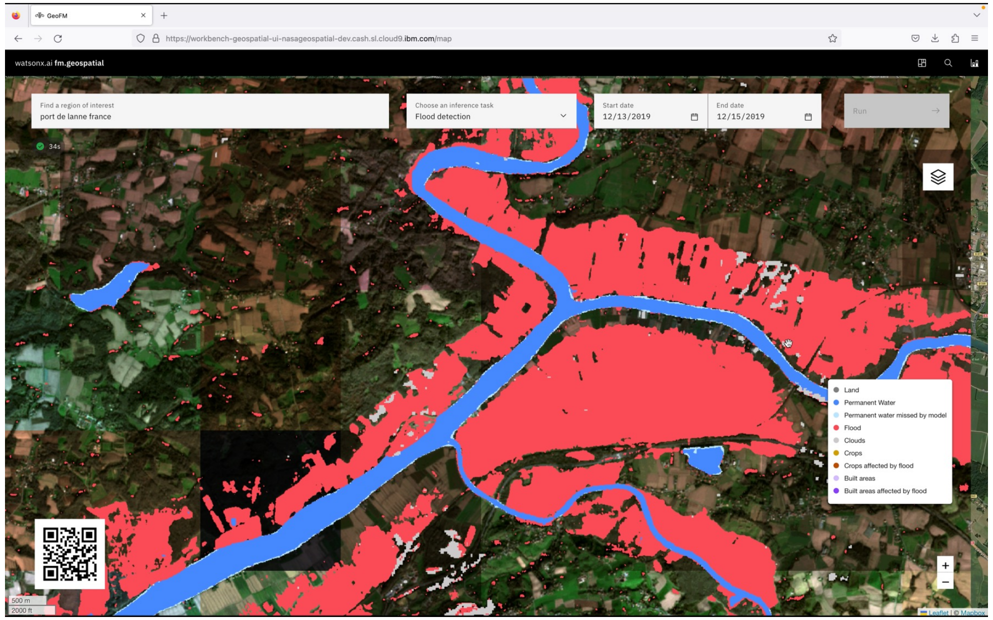
Inference insights by Prithvi – Disaster Mapping



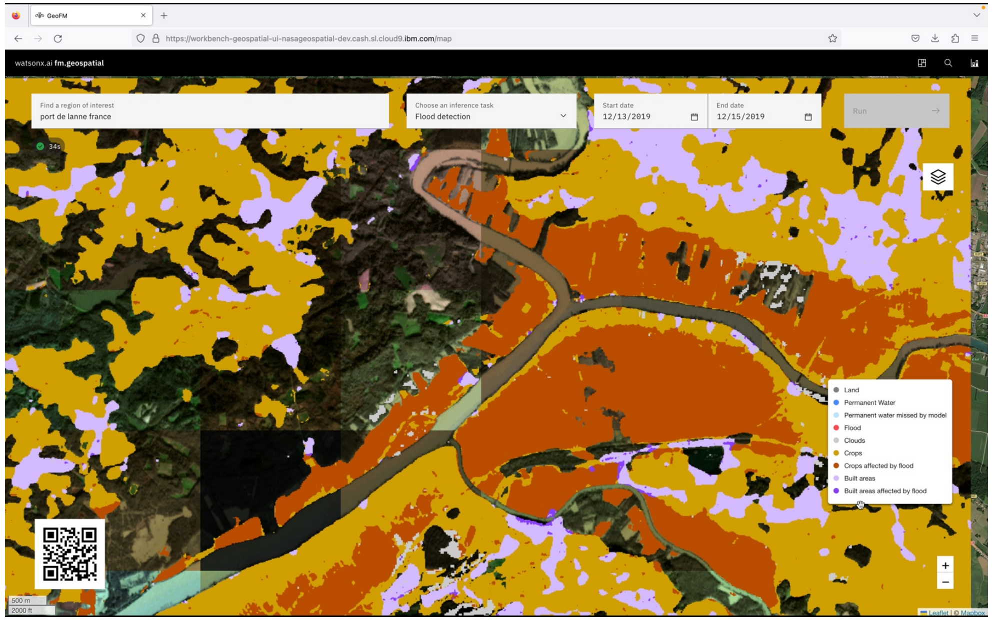
“**Prompt**”: Image(s) (spatial + temporal domains)

<< Inference>>
(e.g., flood task)

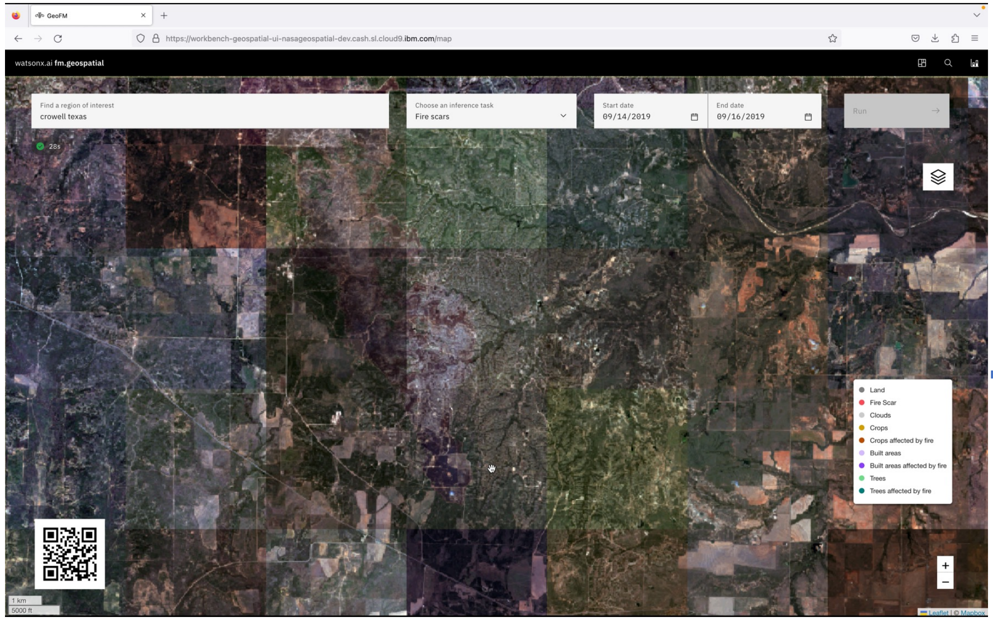
Insights: Flood detection



Insights: Flood impact



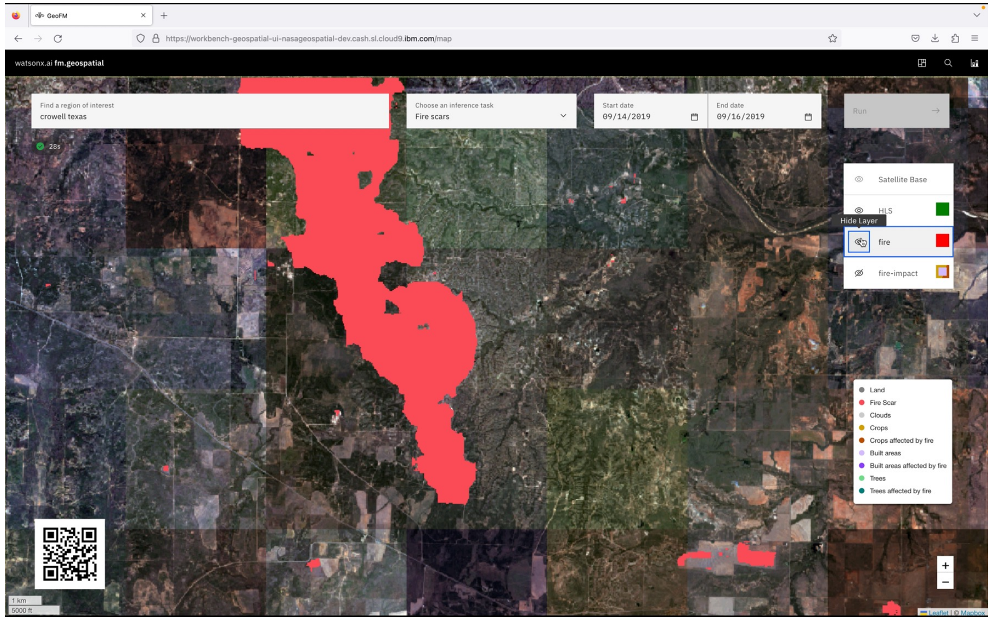
Inference insights by Prithvi – Disaster Mapping



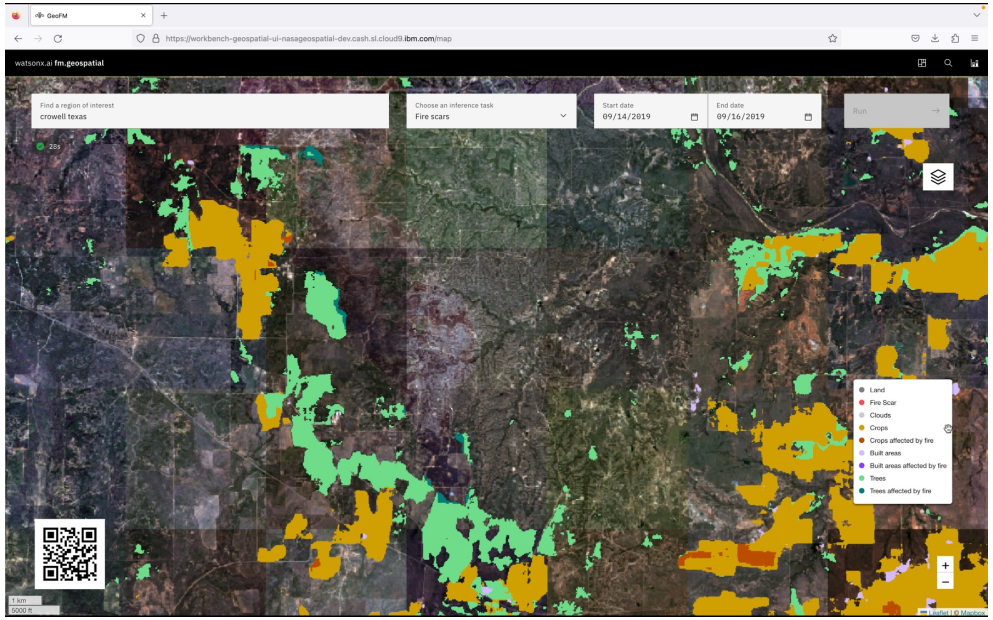
“**Prompt**”: Image(s) (spatial + temporal domains)

<< Inference>>
(e.g., fire-scares task)

Insights: fire scare detection



Insights: Fire impact

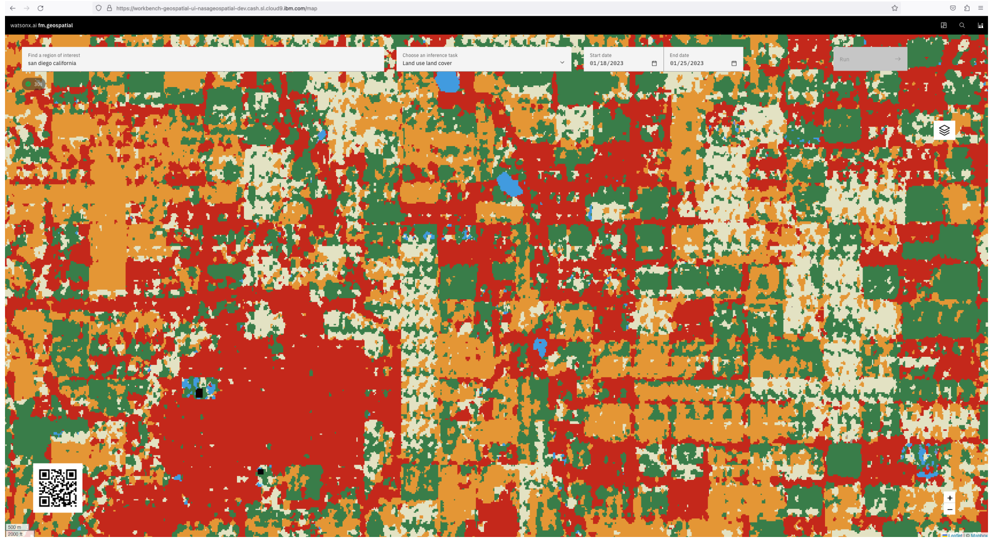


Inference insights by Prithvi – environmental change monitoring

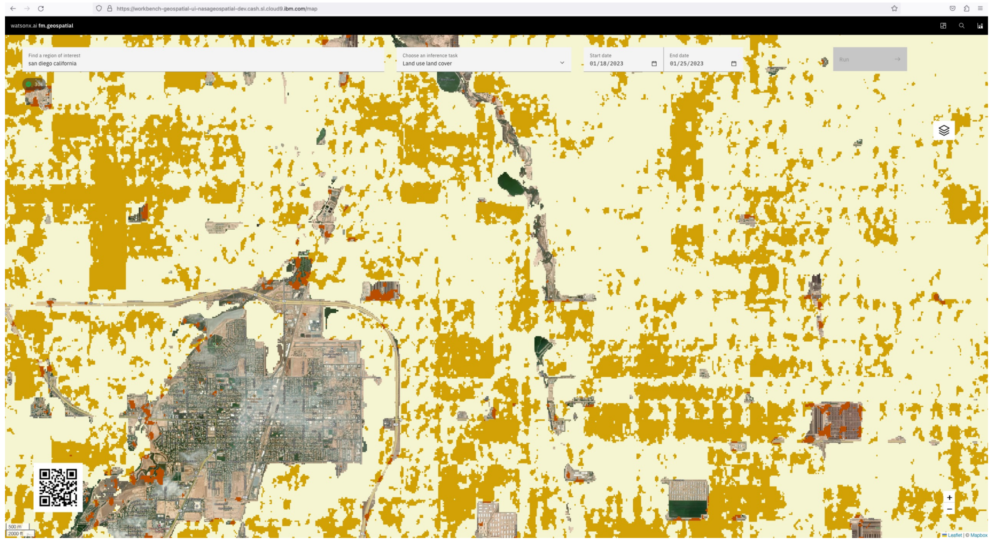
"**Prompt**": Image(s) (spatial + temporal domains)

<< Inference>>
(e.g., land use classification task)

Insights: land use classified



Crop comparison to previous year



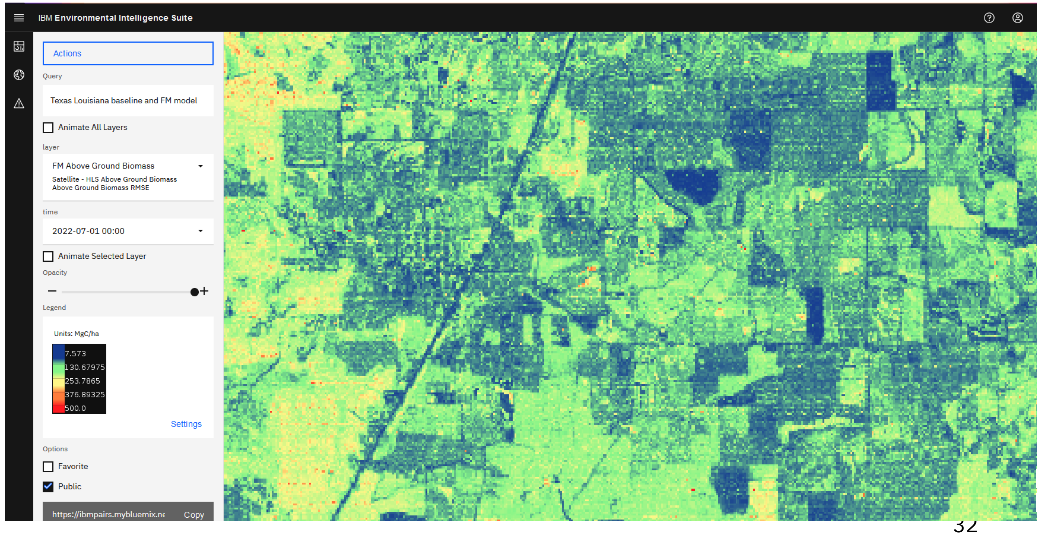
Inference insights by Prithvi – environmental change monitoring

“**Prompt**”: Image(s) (spatial + temporal domains)

<< Inference>>
(e.g., biomass estimation task)

Insights: Biomas estimate June 2022

Insights: Biomass estimate July 2022



Distinction between NLP and Time Series

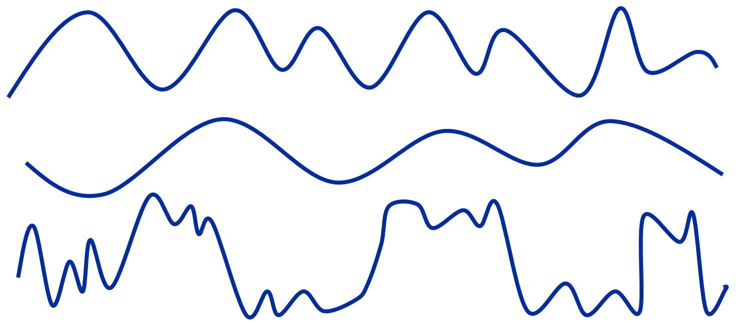
NLP

- Each word contains sematic meaning
- Does not contain spatial information
- Grammatical structures can be similar across different domains.

A cat is walking near the door

Time Series

- Values in a single time step does not contain particular meaning and can be replaced by the neighbors
- Multivariate time series can be constrained in both spatial and temporal dimensions.
- Time series data can be widely different across different industries: resolution, seasonality, drift, ...



NLP has LLMs available but there are yet to have pre-trained general purpose models for TS

Time Series Foundation Models (TSFM)

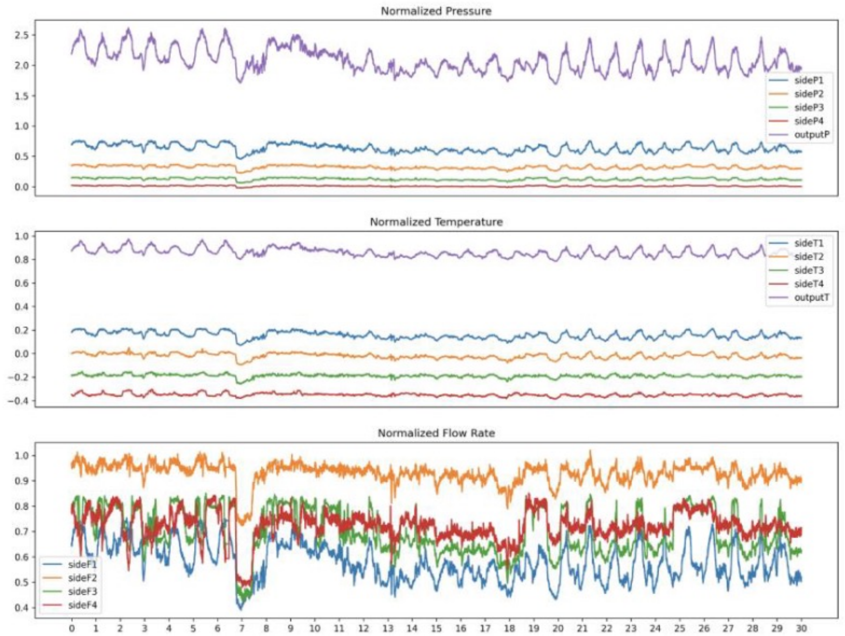
Challenges with Time series FM:

Data quality: Noise and missing data are often encountered in time series. A learned representation that is noise tolerant would be expected.
Potential solution: Smoothing and quantization in the representation space can help.

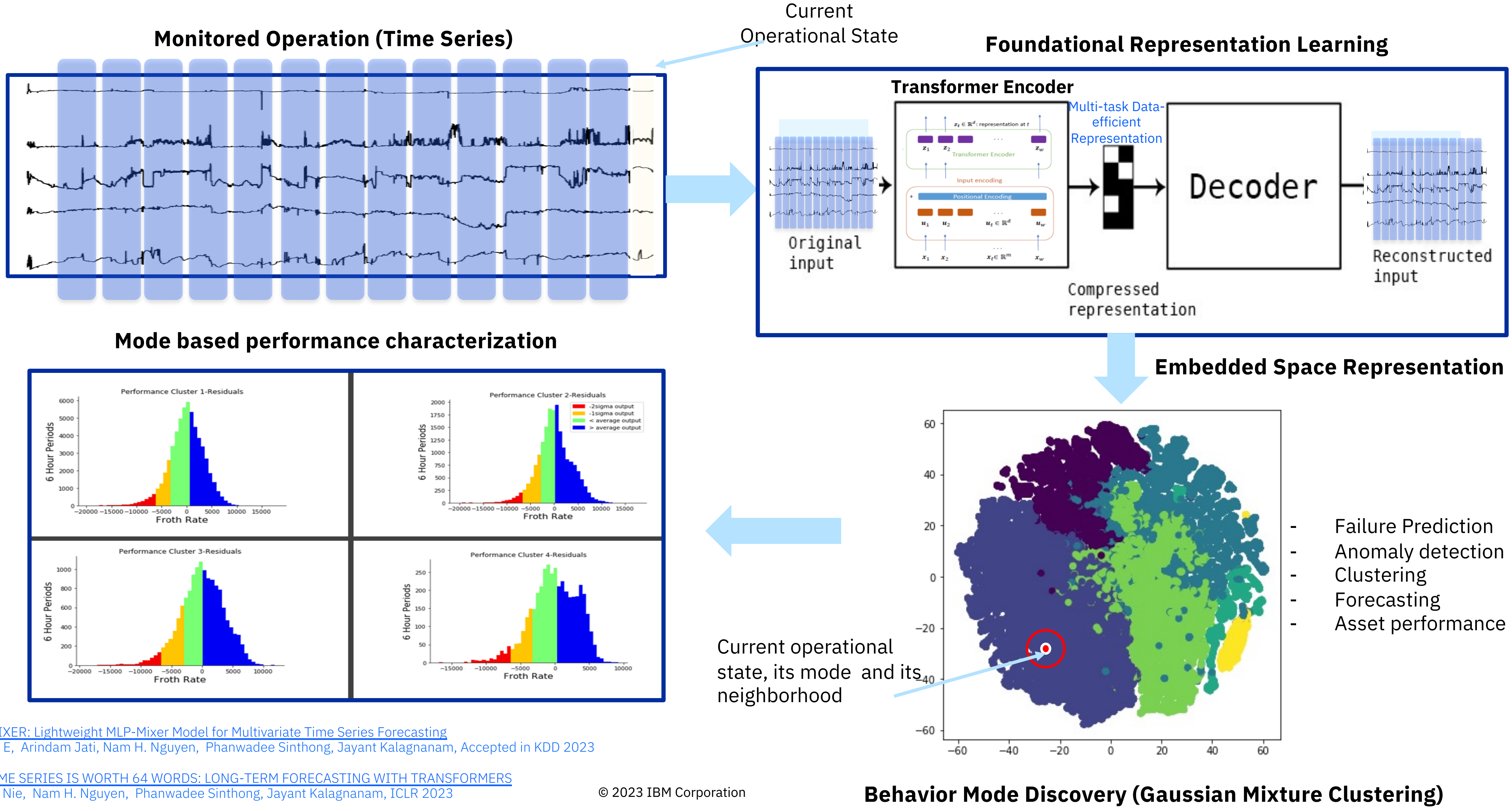
Distribution shift: Non-stationarity can cause distribution shift between the data used for pretraining and data used for downstream tasks.
Potential solution: Pretrain the model on a very large amount of data.

Multivariate data: Leveraging spatial information across variables and temporal information within each series can improve the downstream task substantially.
Potential solution: New design of Transformer to capture special-temporal signal.

Lack of pretraining unlabeled data: As opposed to NLP and vision applications where unlabeled data is unlimited, with time series applications, even unlabeled data is scarce for certain specific manufacturing.
Potential solution: Can we employ simulated data or data from similar domain for pretraining?



Foundational Model Representation for Time Series



[TSMIXER: Lightweight MLP-Mixer Model for Multivariate Time Series Forecasting]
Vijay E, Arindam Jati, Nam H. Nguyen, Phanwadee Sinthong, Jayant Kalagnanam, Accepted in KDD 2023

[A TIME SERIES IS WORTH 64 WORDS: LONG-TERM FORECASTING WITH TRANSFORMERS]
Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, Jayant Kalagnanam, ICLR 2023

Ensuring that AI solutions and deployments are trustworthy is a critical requirement for the adoption of the technology

Credit

Employment

Admission

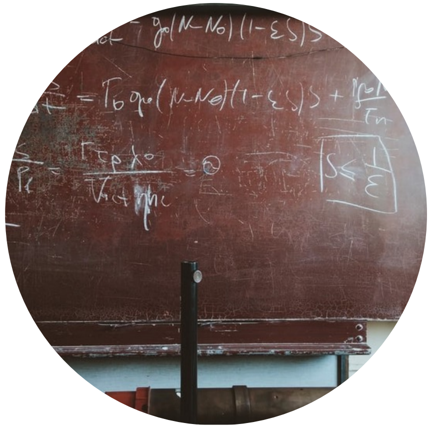
Sentencing

Enterprise Workflows

What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)

Is it fair?



Is it easy to understand?

Did anyone tamper with it?

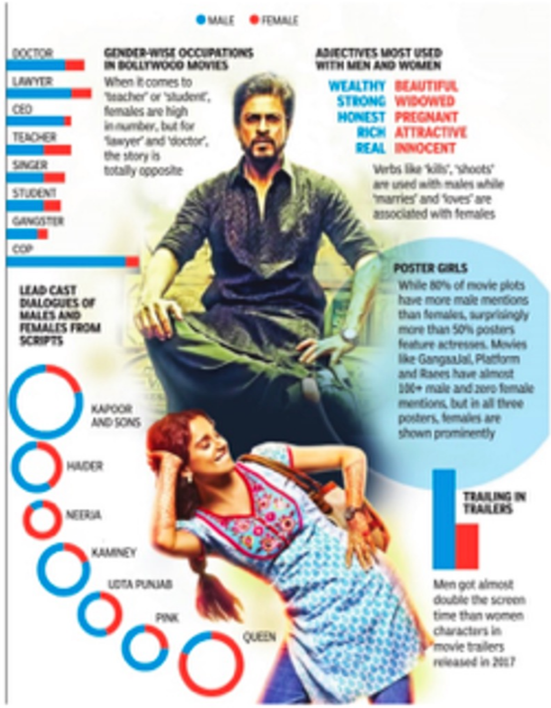
Is it accountable?

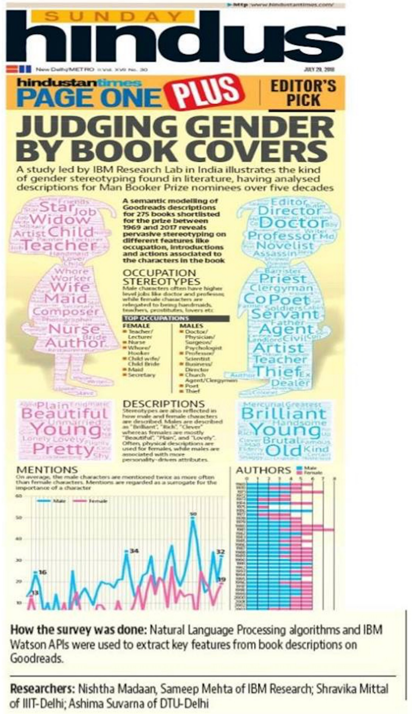
Bollywood and Booker Bias

- Analyzed over 4000 Bollywood movies from the 1970's – 2018.
  - Comparison of Adjectives used for Males and Females
  - Comparison of characters centrality in plot
  - Comparison of screen time of Male and Female Actors.

- Analyzed 275 Bookers shortlists from Goodreads.
  - Comparison on Adjectives, mentions, centrality, verbs.

- IBM Watson APIs and Novel Graph Based Algorithms were developed to computationally quantify the bias





Representation of Male Actors

Representation of female Actors

IBM is using Bollywood movies to identify and neutralize gender bias

Bollywood is 'Crazy Sexist', Says Research

Men are Always Angry and Women Always Happy. What We Learnt From a New Study on Sexism in Bollywood Movies

A Study Of 4,000 Bollywood Films Shows How Exactly Do We Treat Our Women

Bollywood is sexist, researchers claim after analysing 4,000 Indian movies

To investigate such disparities, researchers used an IBM dataset of Wikipedia pages of 4,000 Hindi cinema expelled between 1970 and 2017, extracting titles, expel information, plots, soundtracks, and posters. They also analysed 880 executive trailers of cinema expelled between 2008 and 2017.

THE "WEAKER" SEX

After analysing 4,000 films, researchers confirm that Bollywood movies are still crazy sexist

IBM Uses Bollywood Films To Nullify Gender Bias In Film

"They want to publicise through (the actress) but when it comes to actual story, she has been sidelined," said Nishtha Madaan of IBM India. Madaan co-wrote the paper with Sameep Mehta of IBM and researchers from the Indraprastha Institute of Information Technology, Delhi, and Delhi Technological University.

The quest for “unbiased AI”

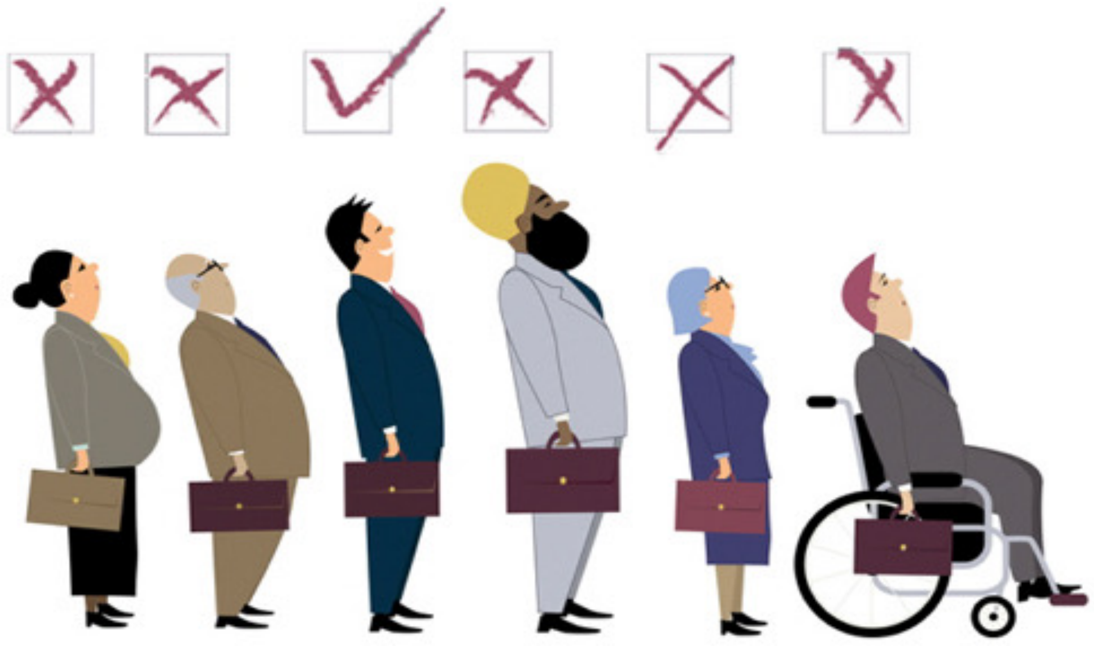
The Guardian US edition
Rise of the racist robots – how AI is learning all our worst impulses

FAST COMPANY
Now Is The Time To Act To End Bias In AI
As decisions made by algorithms come to control more and more aspects of modern life, we need to act swiftly to make sure those decisions are actually fair. As of right now, they’re often not.

Forget Killer Robots—Bias Is the Real AI Danger

Harvard Business Review
TECHNOLOGY
Can We Keep Our Biases from Creeping into AI?
by Kriti Sharma
FEBRUARY 09, 2018

Unwanted bias and algorithmic fairness



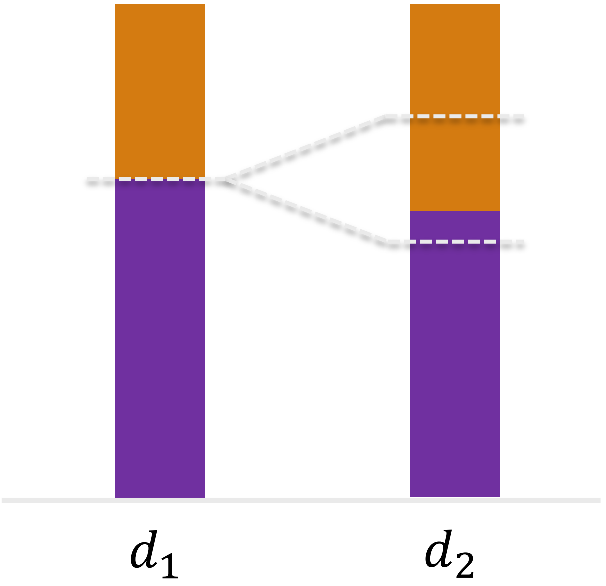
Objectionable when it places certain groups at systematic advantage / disadvantage

Unwanted bias in training data yields models that scale the bias out

Data preprocessing for discrimination prevention

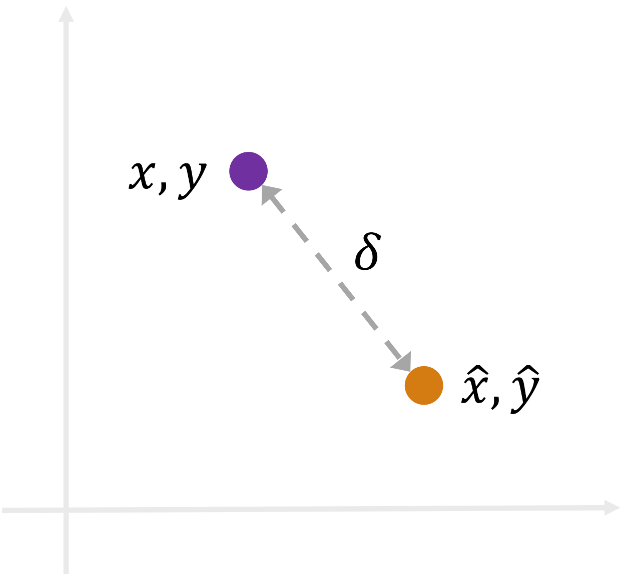
1. Group discrimination

Control dependence $p_{\hat{Y}|D}$ of transformed outcome $\hat{Y}$ on D



2. Individual distortion

Avoid large changes in individual features



3. Utility preservation

Retain joint distribution $p_{X,Y}$ so model can still learn task

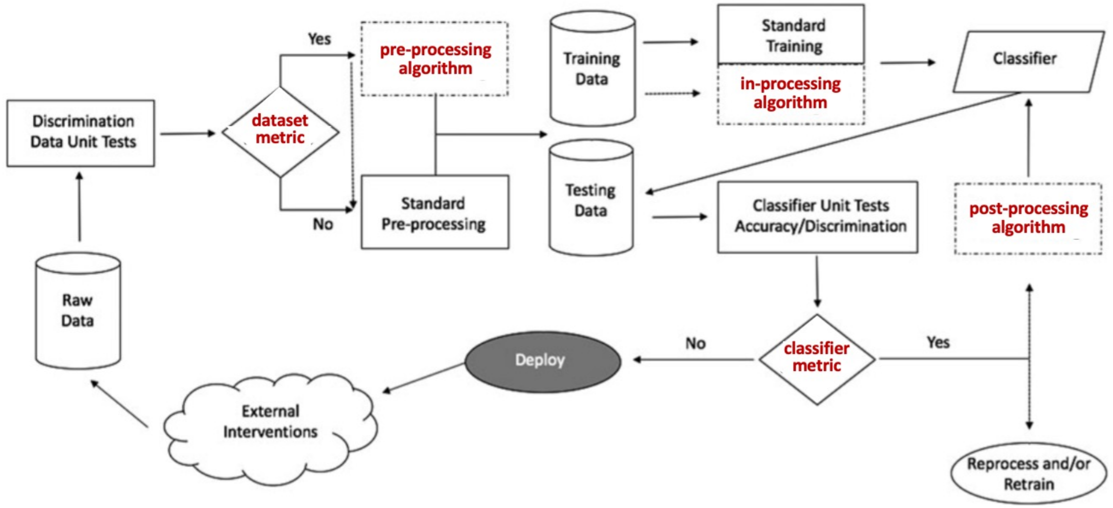
$$\min \quad \Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$$

$$\text{s.t.} \quad J\left(p_{\hat{Y}|D}(\hat{y}|d_1), p_{\hat{Y}|D}(\hat{y}|d_1)\right) \leq \epsilon$$

$$\mathbf{E}\left[\delta\left((x,y),(\hat{X},\hat{Y})\right) \mid d, x, y\right] \leq c$$

Optimized Pre-Processing for Discrimination Prevention. NeurIPS 2017.

Going from theory to practice requires mastery in AI, data science, understanding of algorithmic fairness, diverse and forward thinking, and so much more…



Name	Closest relative	Note
Statistical parity	Independence	Equivalent
Group fairness	Independence	Equivalent
Demographic parity	Independence	Equivalent
Conditional statistical parity	Independence	Relaxation
Equal opportunity	Separation	Relaxation
Equalized odds	Separation	Equivalent
Conditional procedure accuracy equality	Separation	Equivalent
Disparate mistreatment	Separation	Equivalent
Balance for positive class	Separation	Relaxation
Balance for negative class	Separation	Relaxation
Predictive equality	Separation	Relaxation
Conditional use accuracy equality	Sufficiency	Equivalence
Predictive parity	Sufficiency	Relaxation
Calibration	Sufficiency	Equivalence

d'Alessandro, O'Neil, LaGatta. **Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification.** Big Data, June 2017.

Barocas, Hardt, Narayanan. **Fairness and Machine Learning.** https://fairmlbook.org/

AI Fairness 360

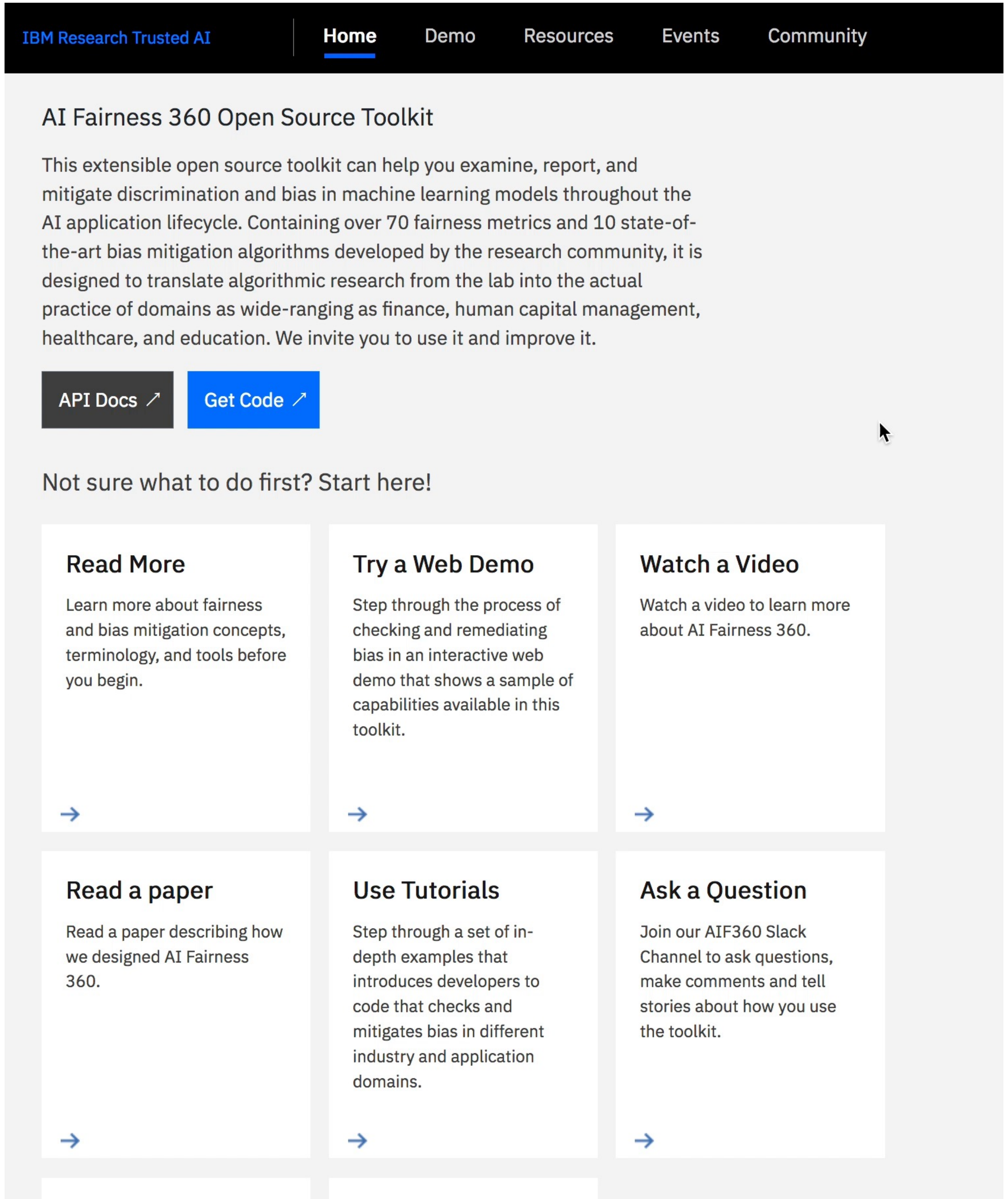
A comprehensive open-source toolkit for handling bias in ML models:

- over 70 fairness metrics
- 10 bias mitigators
- extensive industry tutorials

https://github.com/IBM/AIF360
http://aif360.mybluemix.net/

Think Your Artificial Intelligence Software is Fair? Think Again. IEEE Software, vol. 36, issue 4, p. 76-80, July-August 2019.

AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. IBM Journal of Research and Development, vol. 63, issue 4/5, p. 4, July/September 2019.



The quest for “explainable AI”

CIO JOURNAL.

Companies Grapple With AI’s Opaque Decision-Making Process

THE WALL STREET JOURNAL.

Why Explainable AI Will Be the Next Big Disruptive Trend in Business

AlleyWatch

When a Computer Program Keeps You in Jail

Don't Trust Artificial Intelligence? Time To Open The AI 'Black Box'

One explanation does not fit all

Example: An AI-powered loan approval process

Loan Officer

"Why did you recommend rejection?"

Affected Consumer

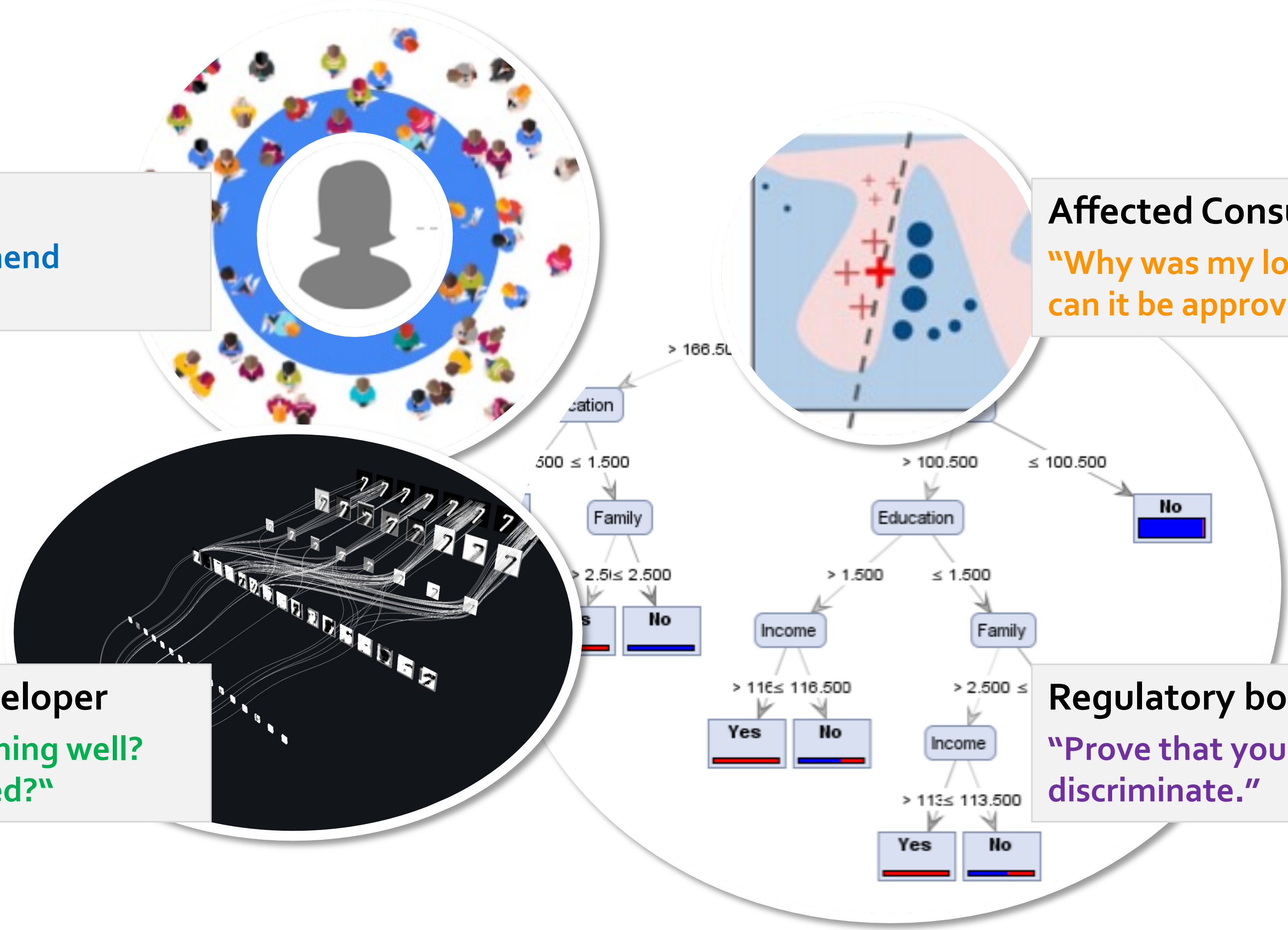
"Why was my loan denied? How can it be approved?"

Data Scientist / Developer

"Is the system performing well? How can it be improved?"

Regulatory bodies

"Prove that your system didn't discriminate."

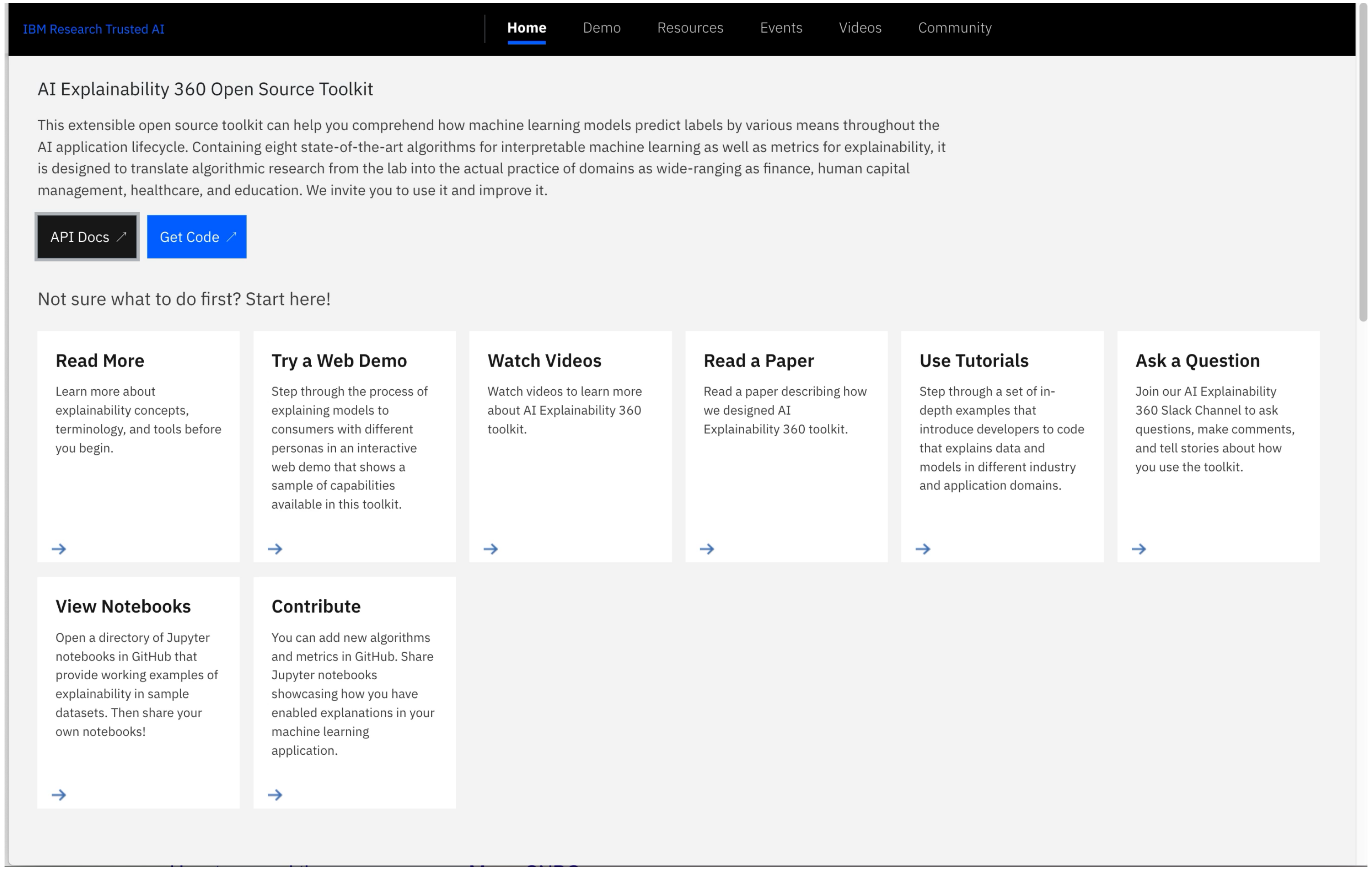


AI Explainability 360

Supporting diverse and rich explanations:

- 8 unique techniques from IBM Research
  - data vs model
  - global vs local
  - directly vs post hoc
- 2 explainability metrics
- extensive industry tutorials to educate users and practitioners

https://github.com/IBM/AIX360/
http://aix360.mybluemix.net/



One Explanation Does Not Fit All:A Toolkit and Taxonomy of AI Explainability Techniques. https://arxiv.org/pdf/1909.03012.pdf

AI Explainability 360: Hands-on Tutorial. FAT* 2020.

The quest for safe and robust AI

INFOSEC INSTITUTE

How Criminals Can Exploit AI

SecurityIntelligence

How Can Companies Defend Against Adversarial Machine Learning Attacks in the Age of AI?

Home > Security

NEWS

Hackers get around AI with flooding, poisoning and social engineering

Many defensive systems need to be tuned, or tune themselves, in order to appropriately respond to possible threats.

OPINION

The rise of artificial intelligence DDoS attacks

The leaves may change color, but the roots are the same. Are you ready for AI-based DDoS attacks?

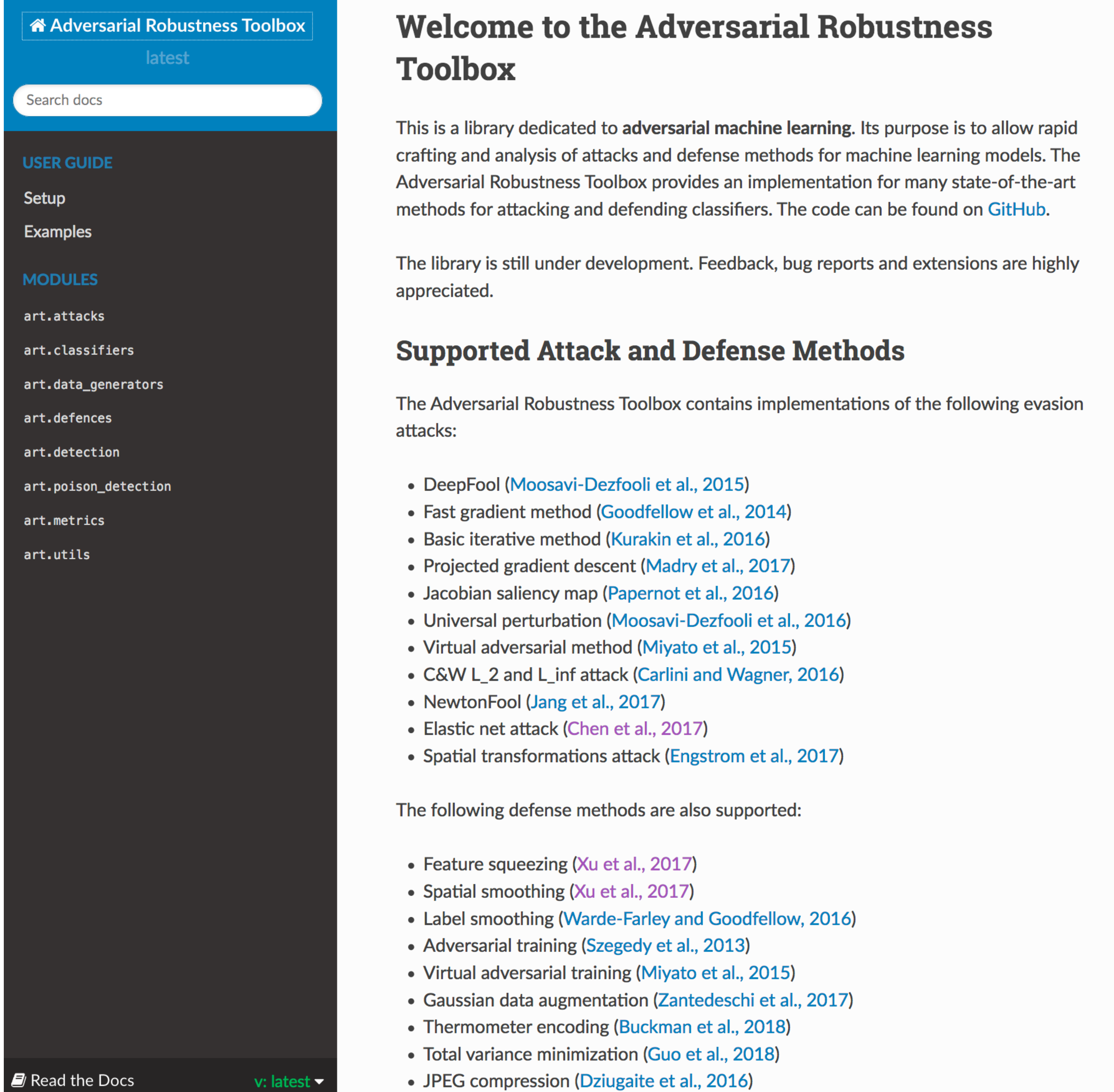
Adversarial Robustness Toolbox

An open source toolkit for “attacking” and defending AI

https://github.com/IBM/adversarial-robustness-toolbox

https://art-demo.mybluemix.net/

Adversarial Robustness Toolboxv1.0.0
https://arxiv.org/pdf/1807.01069.pdf.



Adversarial Robustness Toolbox
latest
Search docs

USER GUIDE
- Setup
- Examples

MODULES
- art.attacks
- art.classifiers
- art.data_generators
- art.defences
- art.detection
- art.poison_detection
- art.metrics
- art.utils

Read the Docs   v: latest

Welcome to the Adversarial Robustness Toolbox

This is a library dedicated to **adversarial machine learning**. Its purpose is to allow rapid crafting and analysis of attacks and defense methods for machine learning models. The Adversarial Robustness Toolbox provides an implementation for many state-of-the-art methods for attacking and defending classifiers. The code can be found on GitHub.

The library is still under development. Feedback, bug reports and extensions are highly appreciated.

Supported Attack and Defense Methods

The Adversarial Robustness Toolbox contains implementations of the following evasion attacks:

- DeepFool (Moosavi-Dezfooli et al., 2015)
- Fast gradient method (Goodfellow et al., 2014)
- Basic iterative method (Kurakin et al., 2016)
- Projected gradient descent (Madry et al., 2017)
- Jacobian saliency map (Papernot et al., 2016)
- Universal perturbation (Moosavi-Dezfooli et al., 2016)
- Virtual adversarial method (Miyato et al., 2015)
- C&W L_2 and L_inf attack (Carlini and Wagner, 2016)
- NewtonFool (Jang et al., 2017)
- Elastic net attack (Chen et al., 2017)
- Spatial transformations attack (Engstrom et al., 2017)

The following defense methods are also supported:

- Feature squeezing (Xu et al., 2017)
- Spatial smoothing (Xu et al., 2017)
- Label smoothing (Warde-Farley and Goodfellow, 2016)
- Adversarial training (Szegedy et al., 2013)
- Virtual adversarial training (Miyato et al., 2015)
- Gaussian data augmentation (Zantedeschi et al., 2017)
- Thermometer encoding (Buckman et al., 2018)
- Total variance minimization (Guo et al., 2018)
- JPEG compression (Dziugaite et al., 2016)

The quest for transparent AI

Trust in AI systems will come from:

1 Applying general safety and reliability engineering methodologies across the entire lifecycle of an AI service.

2 Identifying and addressing new, AI-specific issues and challenges in an ongoing and agile way.

3 Creating standardized tests and **transparent reporting mechanisms on how such a system or service operates and performs.**

Transparent reporting mechanism are basis for trust and safety in many industries and applications

Nutrition Facts
Serving Size 8 oz
Servings Per Container 1.5

Amount Per Serving	
Calories 23	
	% Daily Value*
Total Fat 0g	0%
Saturated Fat 0g	0%
Trans Fat 0g	
Cholesterol 0mg	0%
Sodium 0mg	0%
Total Carbohydrate 5g	2%
Dietary Fiber 0g	0%
Sugars 6g	
Protein 1g	2%

*Percent Daily Values are based on a 2,000 calorie diet.

ENERGY STAR

Moody's		S&P		Fitch		Rating description	
Long-term	Short-term	Long-term	Short-term	Long-term	Short-term		
Aaa	P-1	AAA	A-1+	AAA	F1+	Prime	Investment-grade
Aa1		AA+		AA+		High grade	
Aa2		AA		AA			
Aa3		AA−		AA−			
A1		A+	A-1	A+	F1	Upper medium grade	
A2		A		A			
A3	P-2	A−	A-2	A−	F2		
Baa1		BBB+		BBB+		Lower medium grade	
Baa2	P-3	BBB	A-3	BBB	F3		
Baa3		BBB−		BBB−			
Ba1	Not prime	BB+	B	BB+	B	Non-investment grade speculative	Non-investment grade aka high-yield bonds aka junk bonds
Ba2		BB		BB			
Ba3		BB−		BB−			
B1		B+		B+		Highly speculative	
B2		B		B			
B3		B−		B−			
Caa1		CCC+	C	CCC	C	Substantial risks	
Caa2		CCC				Extremely speculative	
Caa3		CCC−				Default imminent with little prospect for recovery	
Ca		CC					
		C					
C		D	/	DDD	/	In default	
/				DD			
				D			

IEEE
IEEE STANDARDS ASSOCIATION

We have proposed "FactSheets" for AI services

FactSheets: Increasing Trust in AI Services
through Supplier's Declarations of Conformity

M. Arnold,[1] R. K. E. Bellamy,[1] M. Hind,[1] S. Houde,[1] S. Mehta,[2] A. Mojsilović,[1]
R. Nair,[1] K. Natesan Ramamurthy,[1] D. Reimer,[1] A. Olteanu,* D. Piorkowski,[1]
J. Tsay,[1] and K. R. Varshney[1]
IBM Research
[1]Yorktown Heights, New York, [2]Bengaluru, Karnataka

Abstract

Accuracy is an important concern for suppliers of artificial intelligence (AI) services, but considerations beyond accuracy, such as safety (which includes fairness and explainability), security, and provenance, are also critical elements to engender consumers' trust in a service. Many industries use transparent, standardized, but often not legally required documents called supplier's declarations of conformity (SDoCs) to describe the lineage of a product along with the safety and performance testing it has undergone. SDoCs may be considered multi-dimensional fact sheets that capture and quantify various aspects of the product and its development to make it worthy of consumers' trust. Inspired by this practice, we propose FactSheets to help increase trust in AI services. We envision such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers. We suggest a comprehensive set of declaration items tailored to AI and provide examples for two fictitious AI services in the appendix of the paper.

1 Introduction

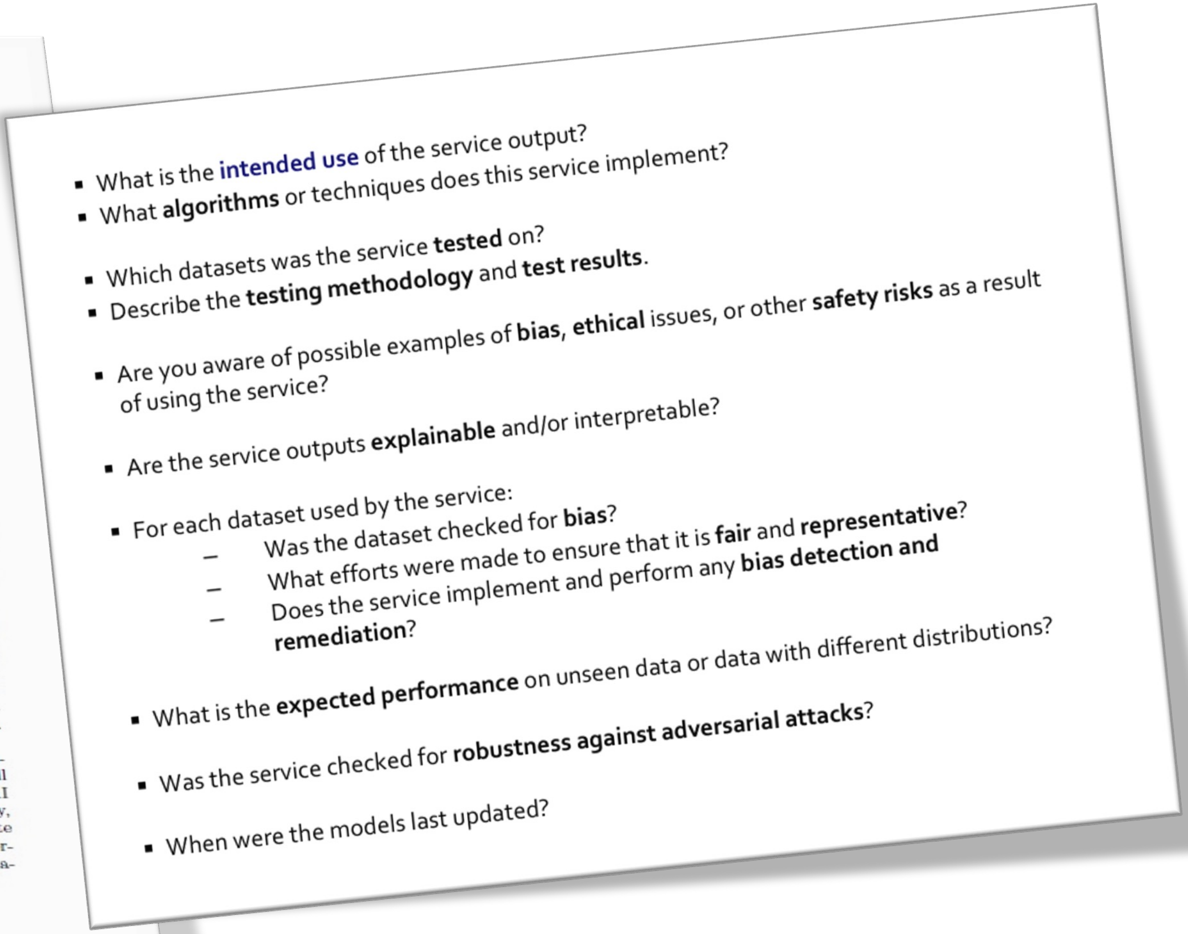
Artificial intelligence (AI) services, such as those containing predictive models trained through machine learning, are increasingly key pieces of products and decision-making workflows. A service is a function or application accessed by a customer via a cloud infrastructure, typically by means of an application programming interface (API). For example, an AI service could take an audio waveform as input and return a transcript of what was spoken as output, with all complexity hidden from the user, all computation done in the cloud, and all models used to produce the output pre-trained by the supplier of the service. A second more complex example would provide an audio waveform translated into a different language as output. The second example illustrates that a service can be made up of many different models (speech recognition, language translation, possibly sentiment or tone analysis, and speech synthesis) and is thus a distinct concept from a single pre-trained machine learning model or library.

In many different application domains today, AI services are achieving impressive accuracy. In certain areas, high accuracy alone may be sufficient, but deployments of AI in high-stakes decisions, such as credit applications, judicial decisions, and medical recommendations, require greater trust in AI services. Although there is no scholarly consensus on the specific traits that imbue trustworthiness in people or algorithms [1, 2], fairness, explainability, general safety, security, and transparency are some of the issues that have raised public concern about trusting AI and threatened the further adoption of AI beyond low-stakes uses [3, 4]. Despite active research and development to address these issues, there is no mechanism yet for the creator of an AI service to communicate how they are addressed in a deployed version. This is a major impediment to broad AI adoption.

Toward transparency for developing trust, we propose a *FactSheet* for AI Services. A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance, safety, and security. Performance will include appropriate accuracy or risk measures along with timing information. Safety, discussed in [5, 3] as the minimiza-

*A. Olteanu's work was done while at IBM Research. Author is currently affiliated with Microsoft Research.

1



- What is the **intended use** of the service output?
- What **algorithms** or techniques does this service implement?
- Which datasets was the service **tested** on?
- Describe the **testing methodology** and **test results**.
- Are you aware of possible examples of **bias**, **ethical** issues, or other **safety risks** as a result of using the service?
- Are the service outputs **explainable** and/or interpretable?
- For each dataset used by the service:
  - Was the dataset checked for **bias**?
  - What efforts were made to ensure that it is **fair** and **representative**?
  - Does the service implement and perform any **bias detection and remediation**?
- What is the **expected performance** on unseen data or data with different distributions?
- Was the service checked for **robustness against adversarial attacks**?
- When were the models last updated?

FactSheets: Increasing trust in AI services through supplier's declarations of conformity.
https://arxiv.org/abs/1808.07261

Companies, organizations, and universities are working towards standardized ways of documenting AI models and services

Recent works (in chronological order)

Datasheets for Datasets
Gebru, Morgenstern, Vecchione, Vaughan, Wallach, Daumeé, and Crawford, 2018.

The Dataset Nutrition Label: A Framework to Drive Higher Quality Data Standards
Holland, Hosny, Newman, Joseph, and Chmielinski, 2018.

Of Oaths and Checklists
Loukides, Mason, and Patil, 2018.

FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity
Arnold, Bellamy, Hind, Houde, Mehta, Mojsilović, Nair, Natesan Ramamurthy, Reimer, Olteanu, Piorkowski, Tsay, and Varshney, 2018.

Model Cards for Model Reporting
Mitchell, Wu, Zaldivar, Barnes, Vasserman, Hutchinson, Spitzer, Raji, and Gebru, 2018.

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science
Bender, and Friedman, 2018.

Experiences with Improving the Transparency of AI Models and Services
Hind, Houde, Martino, Mojsilovic, Piorkowski, Richards, and Varshney, 2019.

Data Readiness Report
Afzal, Rajmohan, Kesarwani, Mehta, Patel 2020

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI
Madaio, Wortman Vaughan, Stark, and Wallach, 2020.

Draft guidelines

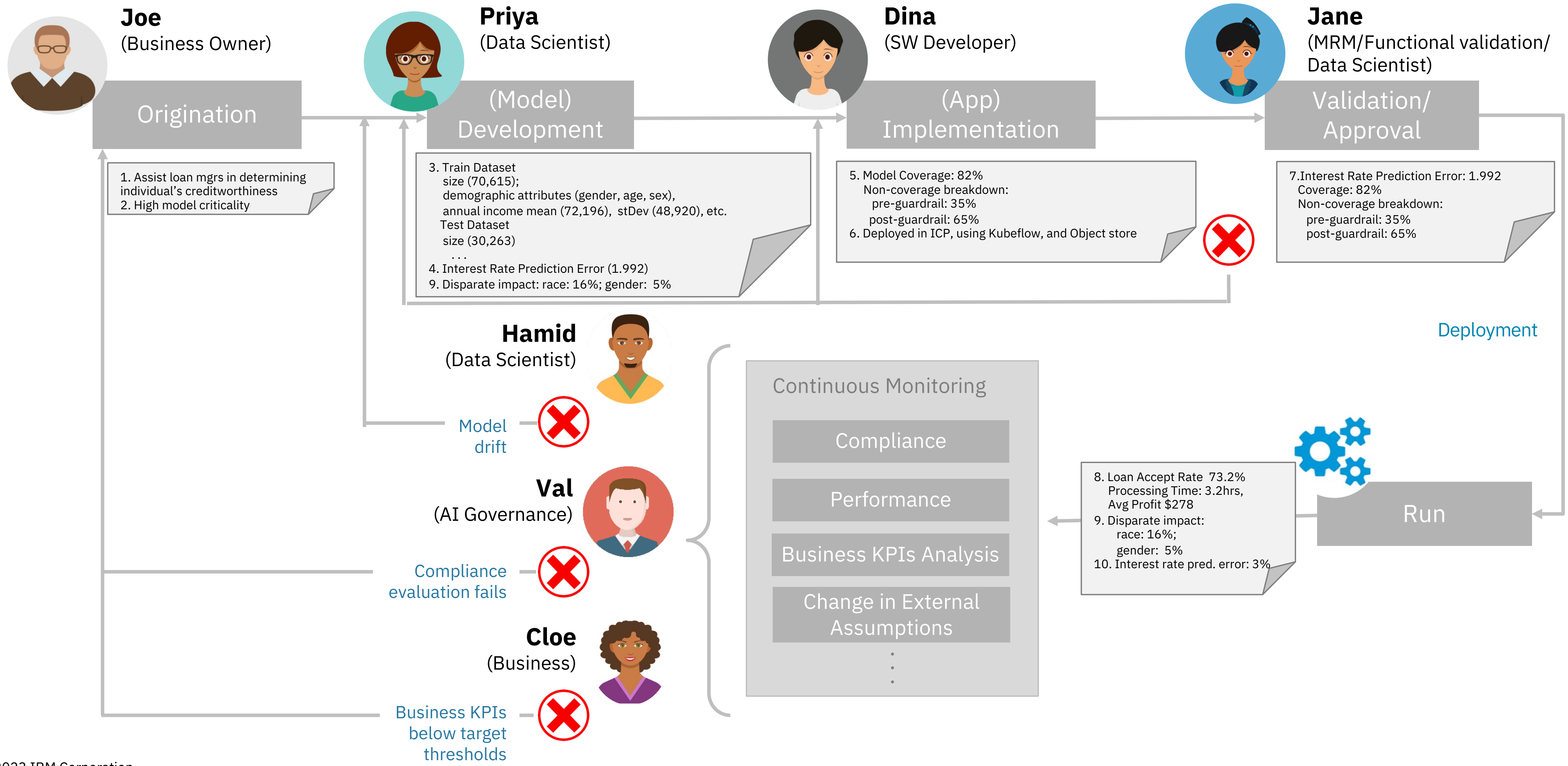
EU Ethics guidelines for trustworthy AI
European Union

TM Forum AI Checklist

Parnership for AI About ML Project

Factsheets and Trusted Lifecycle

Loan processing example



IBM