Primena generativne veštačke inteligencije u poslovnom okruženju

Dane Hinić
Solution Architect
Professional Services, Ingram Micro

What is AI?

AI refers to the ability of computer systems to attempt to mimic the problem-solving and decision-making capabilities of the human mind.

Computer vision

Data science

Natural language processing (NLP)

Robotics

Artificial Intelligence (AI)
Human intelligence exhibited by machines

AI can be defined as a technique that enables machines to mimic cognitive functions associated with human minds – cognitive functions include all aspects of learning, reasoning, perceiving, and problem solving.

Machine Learning (ML)
Systems that learn from historical data

ML-based systems are trained on historical data to uncover patterns. Users provide inputs to the ML system, which then applies these inputs to the discovered patterns and generates corresponding outputs.
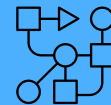
Deep Learning (DL)
ML technique that mimics human brain function

DL is a subset of ML, using multiple layers of neural networks, which are interconnected nodes, which work together to process information. DL is well suited to complex applications, like image and speech recognition.

Foundation Model
Generative AI systems

AI model built using a specific kind of neural network architecture, called a transformer, which is designed to generate sequences of related data elements (for example, like a sentence).

1950's   1980's   2010's   2020's

Traditional AI applications

Speech recognition

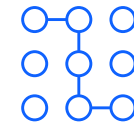Customer service.

Computer vision

Recommendation engines

Automated stock trading

INGRAM MICRO

Generative AI applications
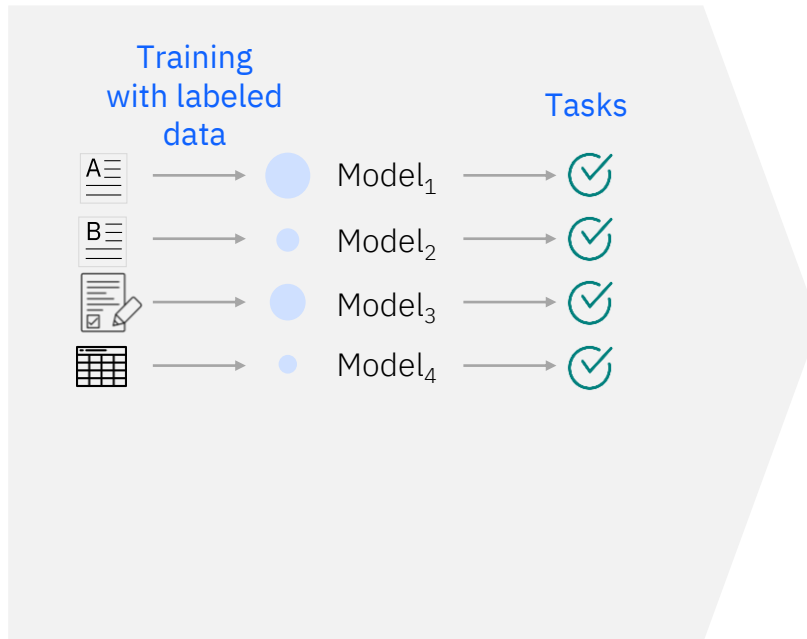
Text generation

Code generation

Music generation

Generate simulated data

Image & video generation

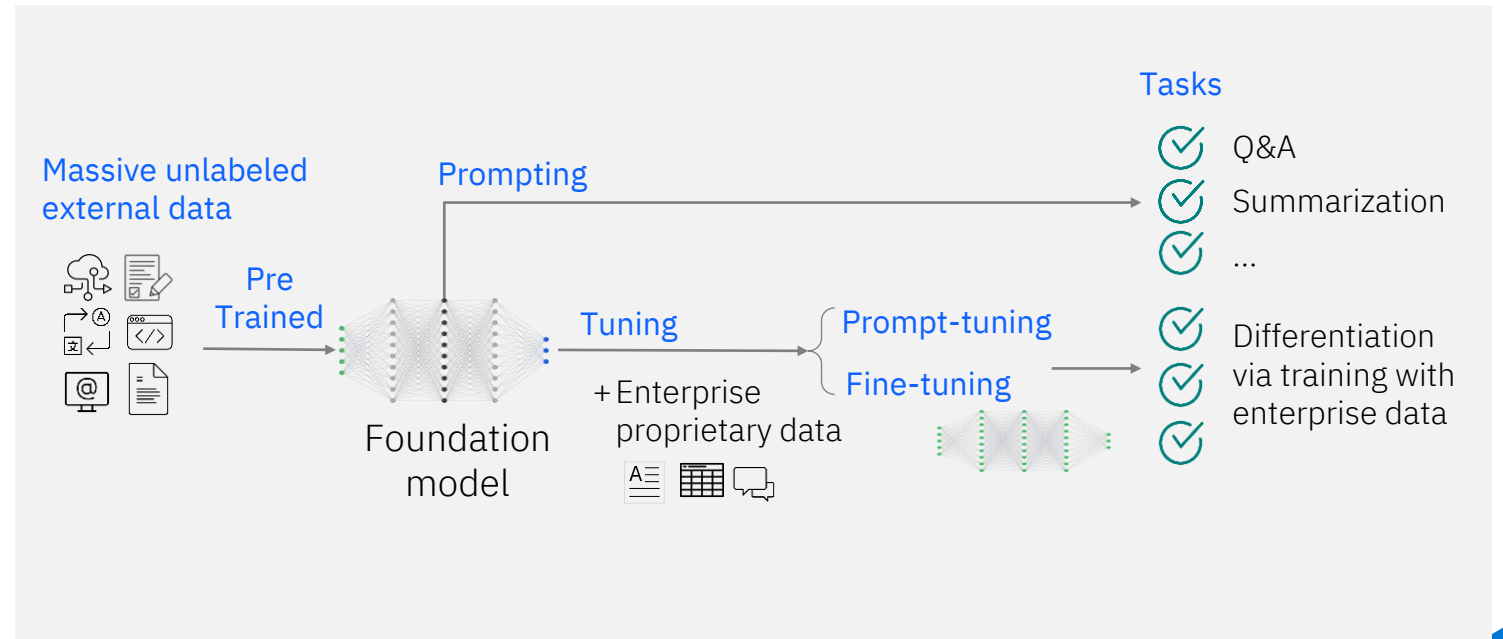**INGRAM** MICRO

Foundational models enable a new paradigm of data-efficient AI development – generative AI

Traditional AI models

Training with labeled data

Tasks

$Model_1$

$Model_2$

$Model_3$

$Model_4$

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

Foundation Models

Massive unlabeled external data

Prompting

Pre Trained

Foundation model

Tuning

+ Enterprise proprietary data

Prompt-tuning

Fine-tuning

Tasks

Q&A

Summarization

...

Differentiation via training with enterprise data

- Rapid adaptation to multiple tasks with small amounts of task-specific data
- Pre-trained unsupervised learning

Generative AI and traditional AI

Both traditional AI and generative AI are useful for enterprises.
Neither replaces the other, generative AI opens new possibilities

Generative AI

- Foundation models trained
  with unlabeled data
- Unsupervised
- Trained on very big data sets
- No specific task
- Transferable
- Works well for general tasks and can
  improve for specific tasks with less training
- Need to monitor bias and drift

Traditional AI

- Traditional Machine learning (ML/AI)
  model trained with "labeled" data
- Training is supervised
- Trained on proper, large data sets
- Trained for a specific task
- Does not transfer well to other tasks
- A tuned model can be very efficient for
  the specific task it was designed for
- Need to monitor bias and drift

Most common generative AI tasks implemented today

Summarization

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

Conversation summaries, insurance coverage, meeting transcripts, contract information

Classification

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

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

Generation

Generate text content for a specific purpose.

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

Extraction

Analyze and extract essential information from unstructured text.

Medical diagnosis support, user research findings

Question-answering

Create a question-answering feature grounded on specific content.

Build a product specific Q&A resource for customer service agents.

INGRAM MICRO

watson**x**

and its 3 components

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

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, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.
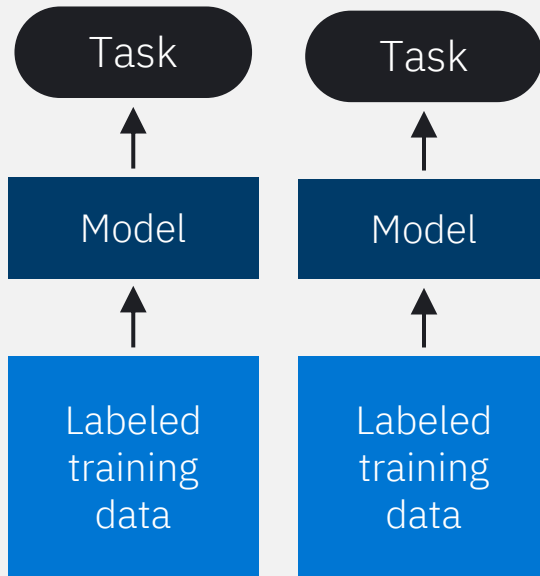
watsonx.governance

Enable responsible, transparent and explainable AI workflows

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

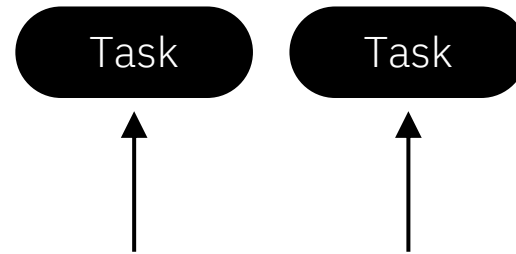**INGRAM** MICRO

What's next with foundation models:

Conventional AI
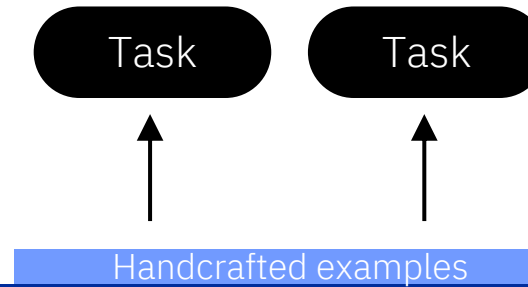
1000s – 100000s labeled data points / task
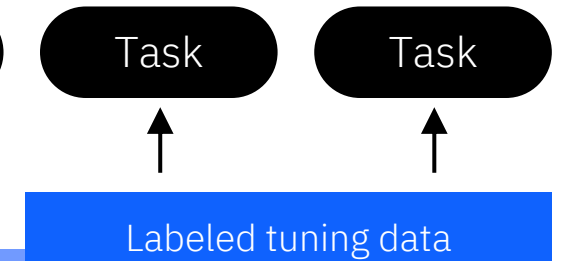
Zero-shot prompting

No labeled data needed

Few-shot prompting
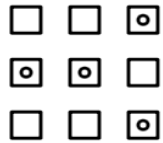
1-10 labeled data points / task

Data-driven tuning

100s – 1000s labeled data points / task

Task
Task

Model
Model

Labeled training data
Labeled training data

Task
Task
Task
Task
Task
Task

Labeled tuning data

Handcrafted examples

Pre-trained generative foundation model

INGRAM MICRO

watsonx.ai

Train, validate, tune, and deploy AI models with confidence
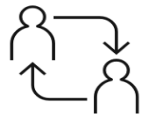
Generative AI capabilities
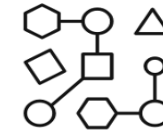
 Foundation model library

 Prompt Lab

 Tuning Studio*
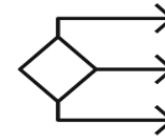
 Team collaboration and data preparation

Plus, a proven studio for machine learning

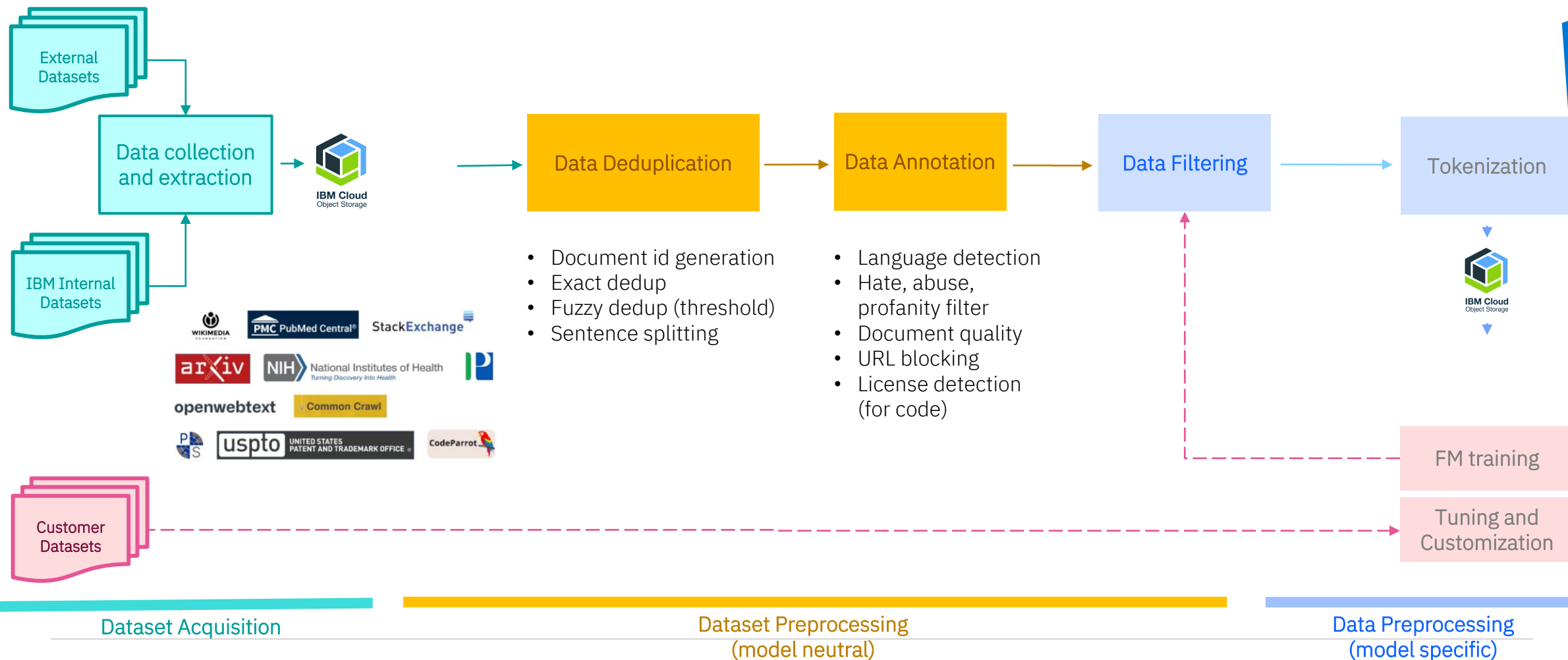 ModelOps

 Automated development

 Decision optimization

INGRAM MICRO

watsonx.data: curated IBM Data Pile

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

External Datasets

IBM Internal Datasets

Data collection and extraction

IBM Cloud Object Storage

Data Deduplication

- Document id generation
- Exact dedup
- Fuzzy dedup (threshold)
- Sentence splitting

Data Annotation

- Language detection
- Hate, abuse, profanity filter
- Document quality
- URL blocking
- License detection (for code)

Data Filtering

Tokenization

IBM Cloud Object Storage

FM training

Tuning and Customization

Customer Datasets

Dataset Acquisition | Dataset Preprocessing (model neutral) | Data Preprocessing (model specific)

INGRAM MICRO

Put AI to work with **watson**x

Scale and accelerate the impact of AI with trusted data on hybrid cloud

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 data and AI workflows

Red Hat OpenShift

INGRAM MICRO

Thanks!

dane.hinic@ingrammicro.com

INGRAM MICRO®