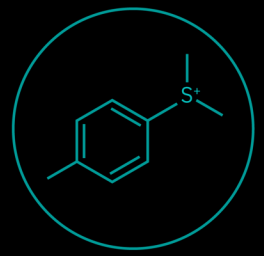
KubeCon | CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022

Training AI to Code using the Largest Code Dataset (CodeNet)

Animesh Singh – CTO and Director

Tommy Li – Senior Software Developer

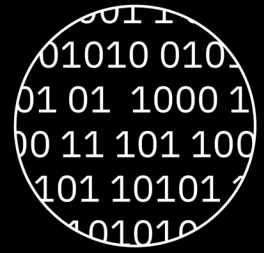
Languages have been the fundamental symbolic systems through which humans have communicated, but there are other languages.



Molecules

Stories

Imagery



Code

AI for Code: *"Teaching machines the language of machines"*

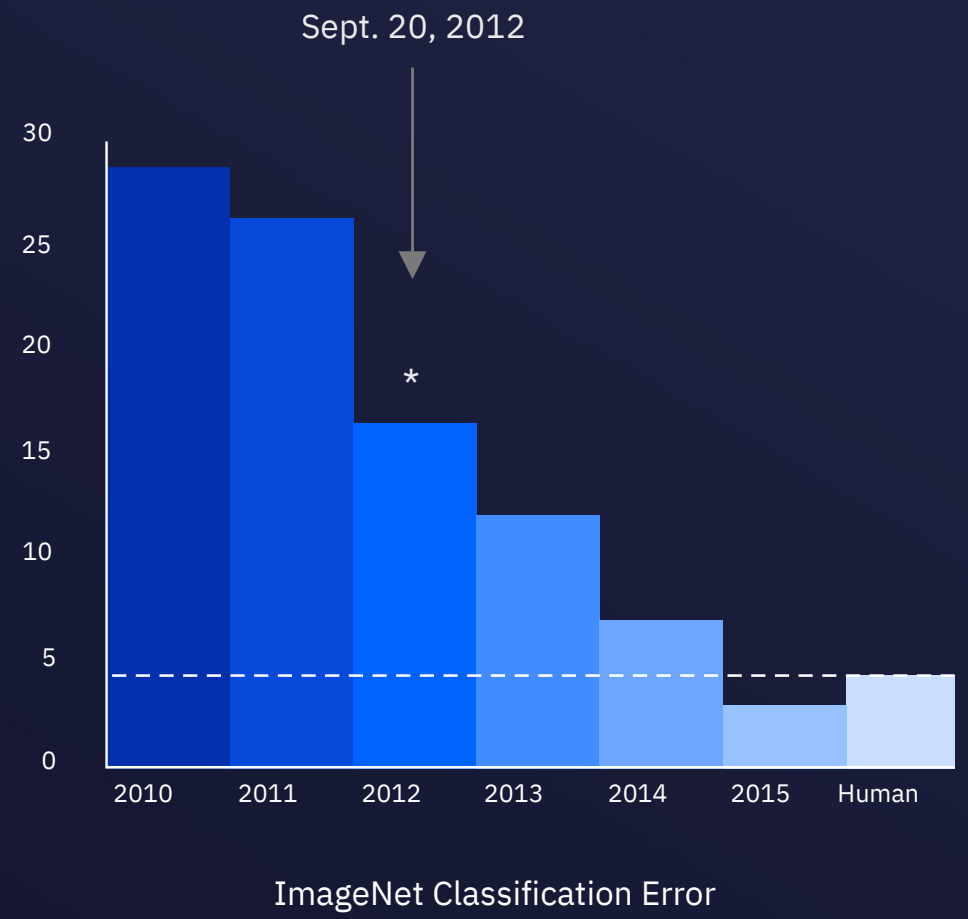
- Relatively new field in computer science
- Leverages technologies like:
  - Natural Language Processing (NLP)
  - Document Understanding
  - Code analysis and compilation techniques
- Goal:
  - Automating the software engineering process
  - Help software developers improve their productivity
  - Perform practical tasks, such as code search, summarization, and completion, as well as code-to-code translation
  - Analyze and modernize legacy software by helping migrate monolithic applications to modern microservices for enterprise applications

The Machine Learning revolution is fueled by data...

2009: 3 million images
5,000 classes

2012: 14 million images
22,000 classes

Resulting in human level ability on narrow tasks.



2011: Watson Jeopardy

2019: Project Debater

The world's software infrastructure urgently needs modernization.

DesignNews

COBOL Coders Needed for Coronavirus Fight

1960: COSADYL Committee creates COBOL

One of the oldest of all programming languages is still in demand but

220 Billion Lines of Legacy COBOL

IEEE SPECTRUM

Tech Talk | Artificial Intelligence | Machine Learning

IBM Watson's Next Challenge: Modernize Legacy Code

15 Oct 2020 | 18:24 GMT

By **Dexter Johnson**

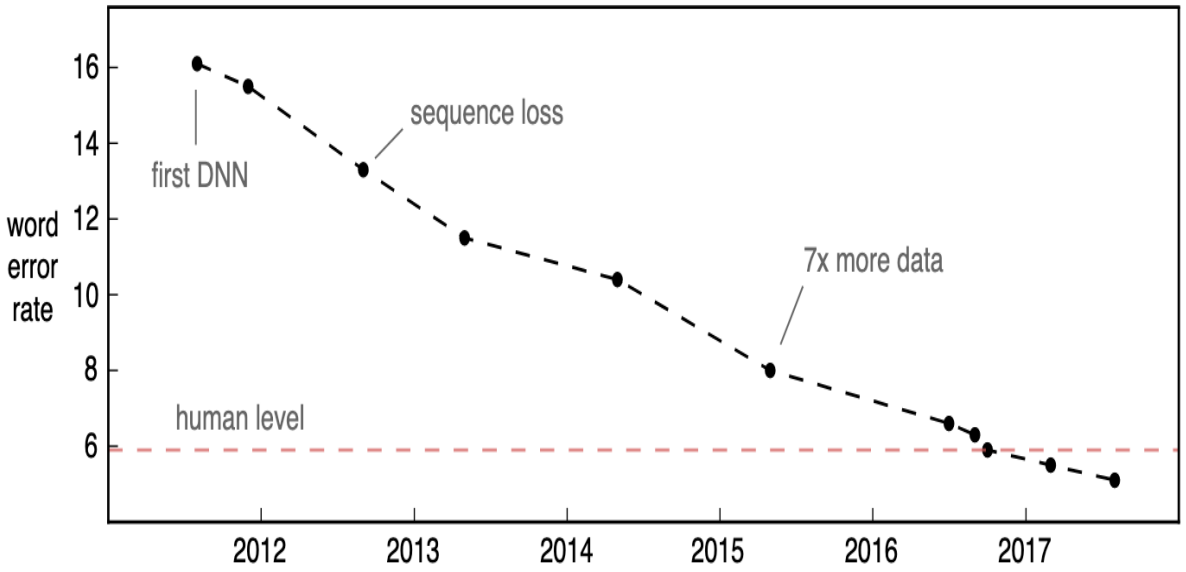
TechChannel

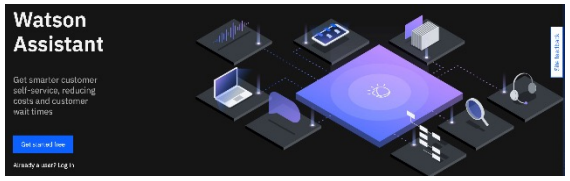
Closing the COBOL Programming Skills Gap

We have seen the power of AI applied to human language.

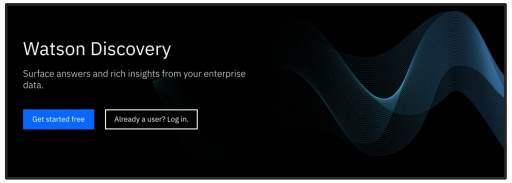
Code is the language of machines. AI will help us master code.

Speech Performance





Voice and conversation



Document understanding

NLP and ML for software artifacts

Automated reasoning and decision-making

Explainable AI

AI for Code needs its “ImageNet” for breakthroughs

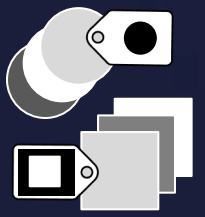
Code language translation

Code search and retrieval

Code similarity

Code performance improvement

Code memory improvement



Code classification

Project CodeNet

A high-quality code dataset for algorithmic innovation and benchmarking

- Large-scale
  - 14 million code samples
  - 4000+ code problems
  - 55 programming languages
  - ½ billion lines of code
- Diverse set of problems and codes
- Codes are well tested
- Test set provided for each problem

```
onst std::size_t KVStore_test::estimated_object_count =
if 0
 pmem_simulated ? estimated_object_count_small : estimated_object_c
else
 0
endif
 ;

onstexpr unsigned KVStore_test::many_key_length;
onstexpr unsigned KVStore_test::many_value_length;

onst std::size_t KVStore_test::many_count_target = pmem_simulated ?
any_count_target_large;
td::size_t KVStore_test::many_count_actual;
td::vector<KVStore_test::kv_t> KVStore_test::kvv;
```

github.com/IBM/Project_CodeNet

Modernizing legacy code

Client Application Statistics

Large automotive client, mission critical application ($200M asset)

3500+ Java files and more than 1 million lines of code

Multi-generation of Java tech over a decade

Over a year of ongoing manual migration effort

Results with AI for code Modernization

Recommendations and validation of code factoring within ***4 weeks vs. a year.***

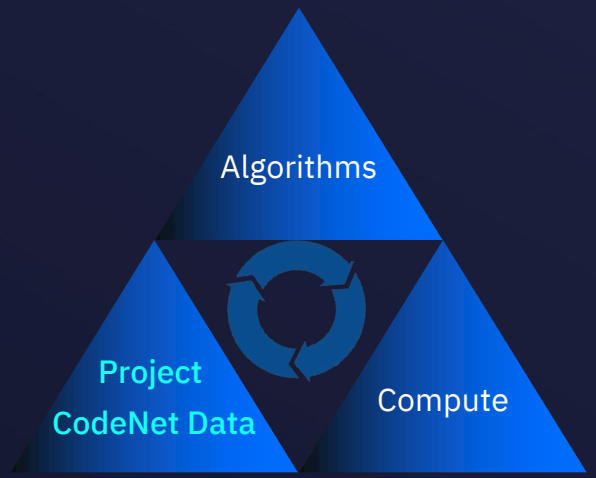
25+ partitions (microservices) consisting of 450+ classes

Leveraging runtime and data dependency analysis

Investigation exposed potential dead code

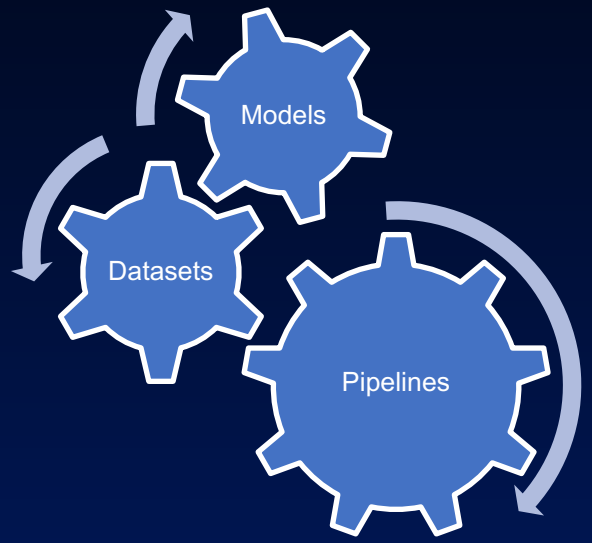
Data

Project CodeNet



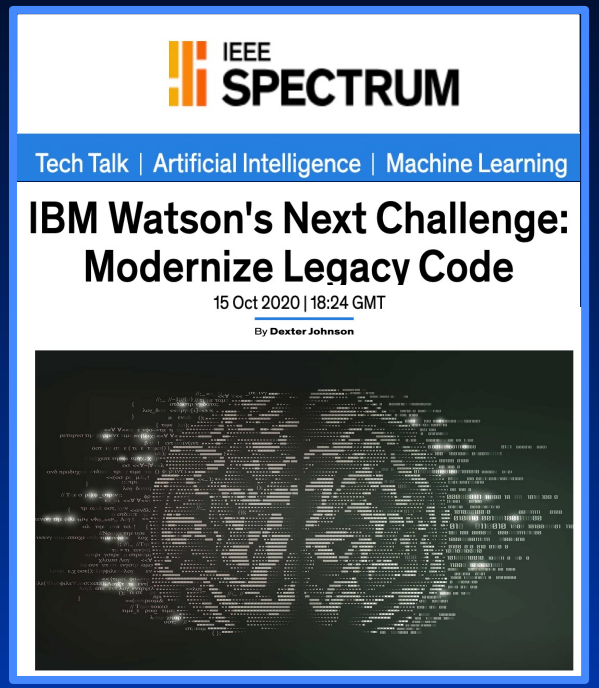
Technology

AI system stack



Business value

Modernize legacy code and boost developer productivity



IEEE SPECTRUM

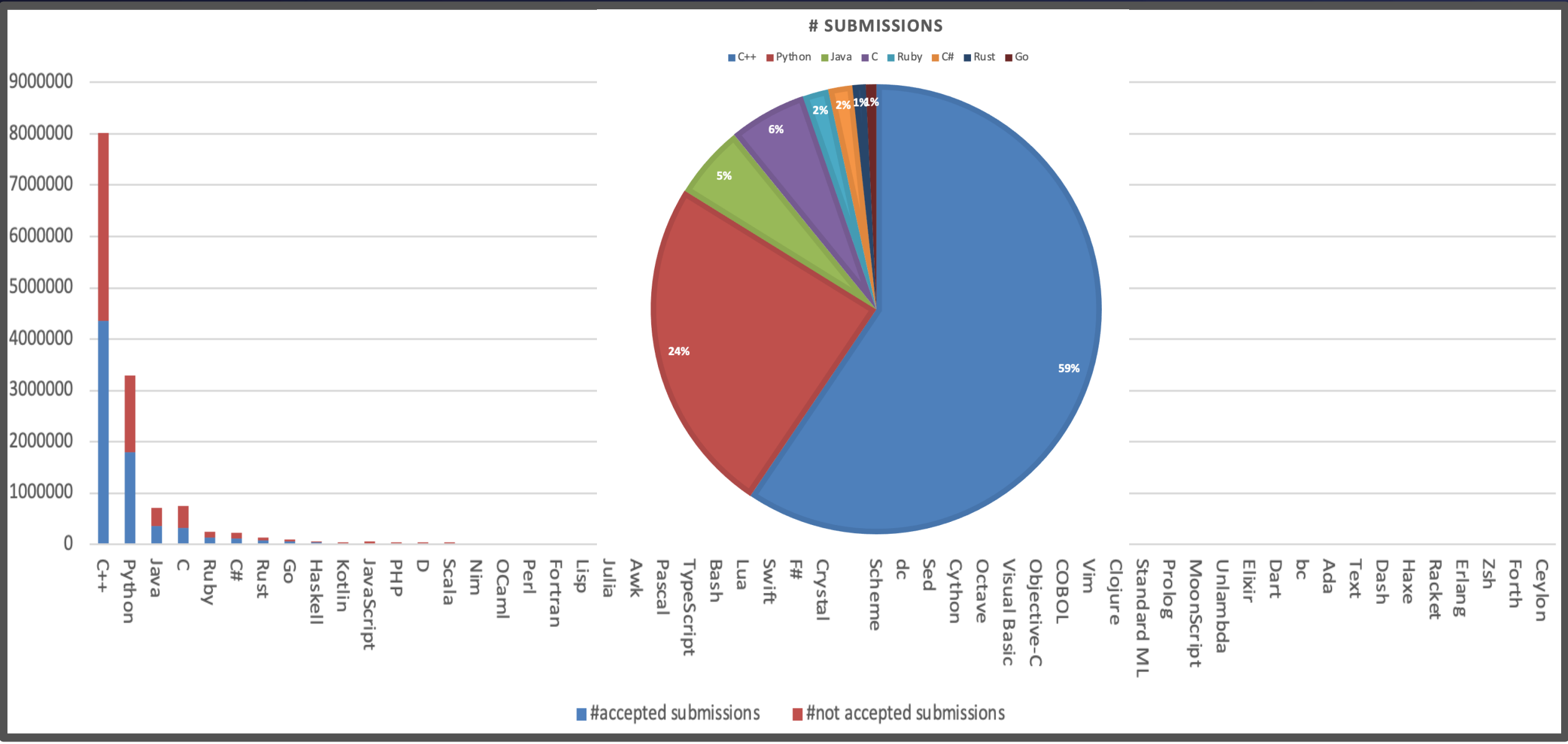
Tech Talk | Artificial Intelligence | Machine Learning

IBM Watson's Next Challenge: Modernize Legacy Code

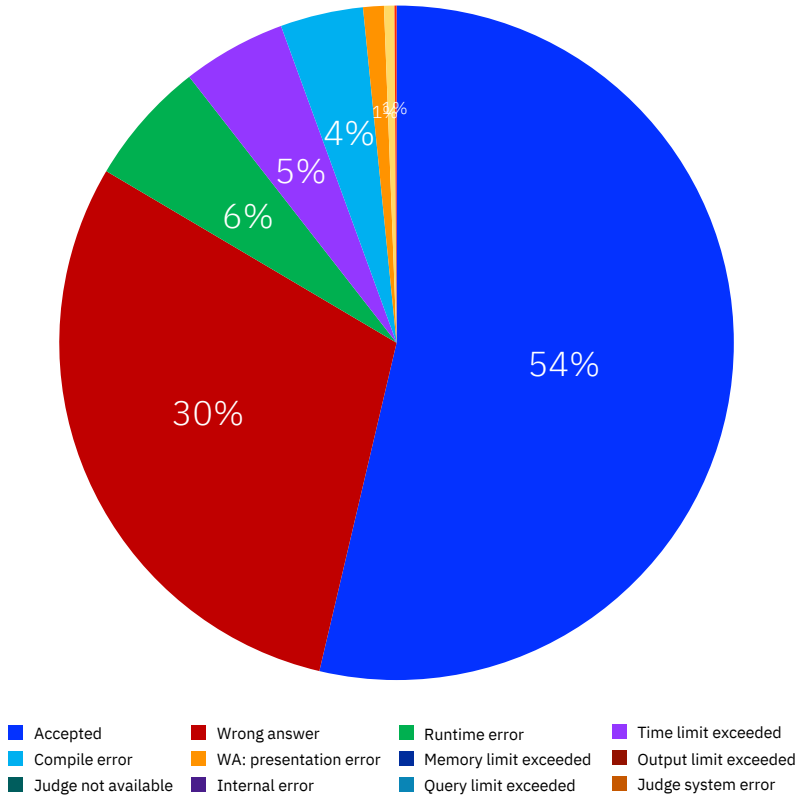
15 Oct 2020 | 18:24 GMT

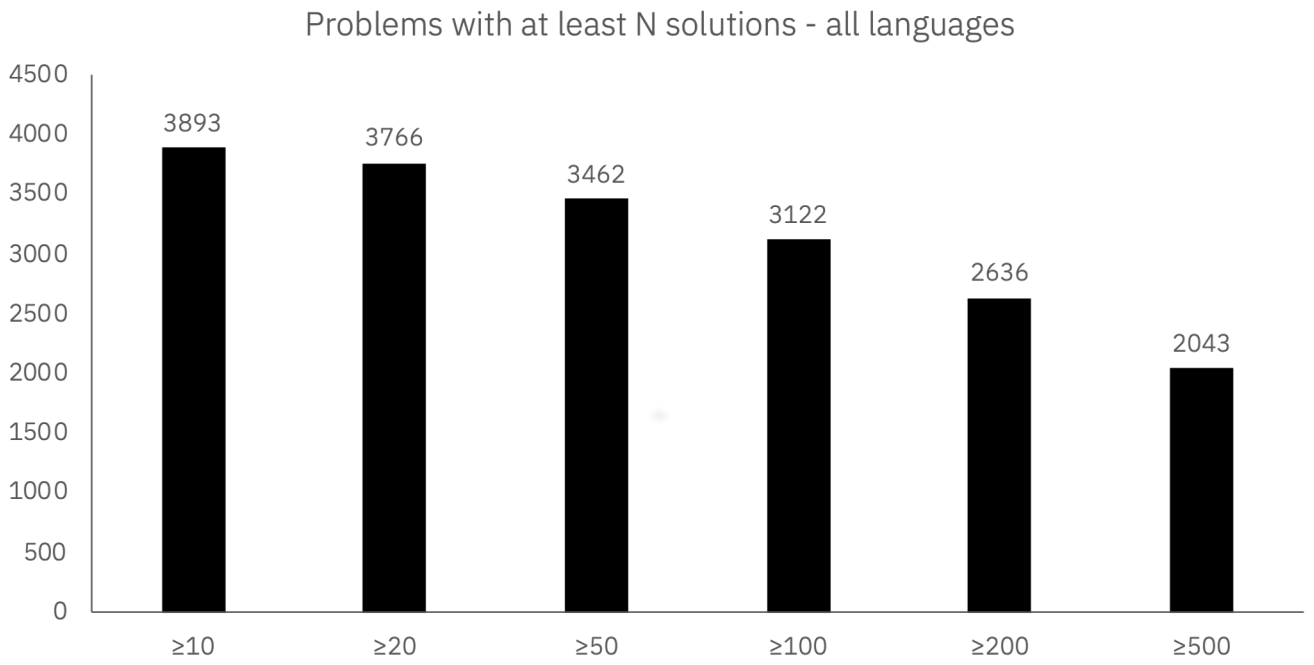
By Dexter Johnson

Project CodeNet: Polyglot of languages



Project CodeNet: diversity of solutions for each problem





80% of the problems have more than 100 solutions each

The Data

Origin:

- Online Judge websites **AIZU Online Judge** and **AtCoder**
- 4053 problems (5 are empty)
- 13,916,868 submissions
  - **53.6% Accepted "*solutions*"**: compilable, executable, produce expected results
  - **29.5% Wrong Answer**
  - **16.7% Rejected** for various causes
- 55 different languages
  - 95% in the six most common languages (C++, Python, Java, C, Ruby, C#)
  - C++ most common with 8,008,527 submissions (57% of the total) of which 4,353,049 are accepted

Data:

- Complete programs in a particular programming language
- Each program is contained in a single file
- Each program attempts to solve a certain programming task or problem
- Each problem might have many solutions in different languages

	CodeNet	GCJ	POJ
Total number of problems	4053	332	104
Number of programming languages	55	20	2
Total number of code samples	13,916,828	2,430,000	52,000
C++/C subset data size (code samples)	8,008,527	280,000	52,000
Percentage of problems with test data	51%	0%	0%
Task: Memory Consumption Prediction	Yes	No	No
Task: Runtime Performance Comparison	Yes	No	No
Task: Error Prediction	Yes	No	No
Task: Near duplicate prediction	Yes	No	No

Metadata – Dataset Level

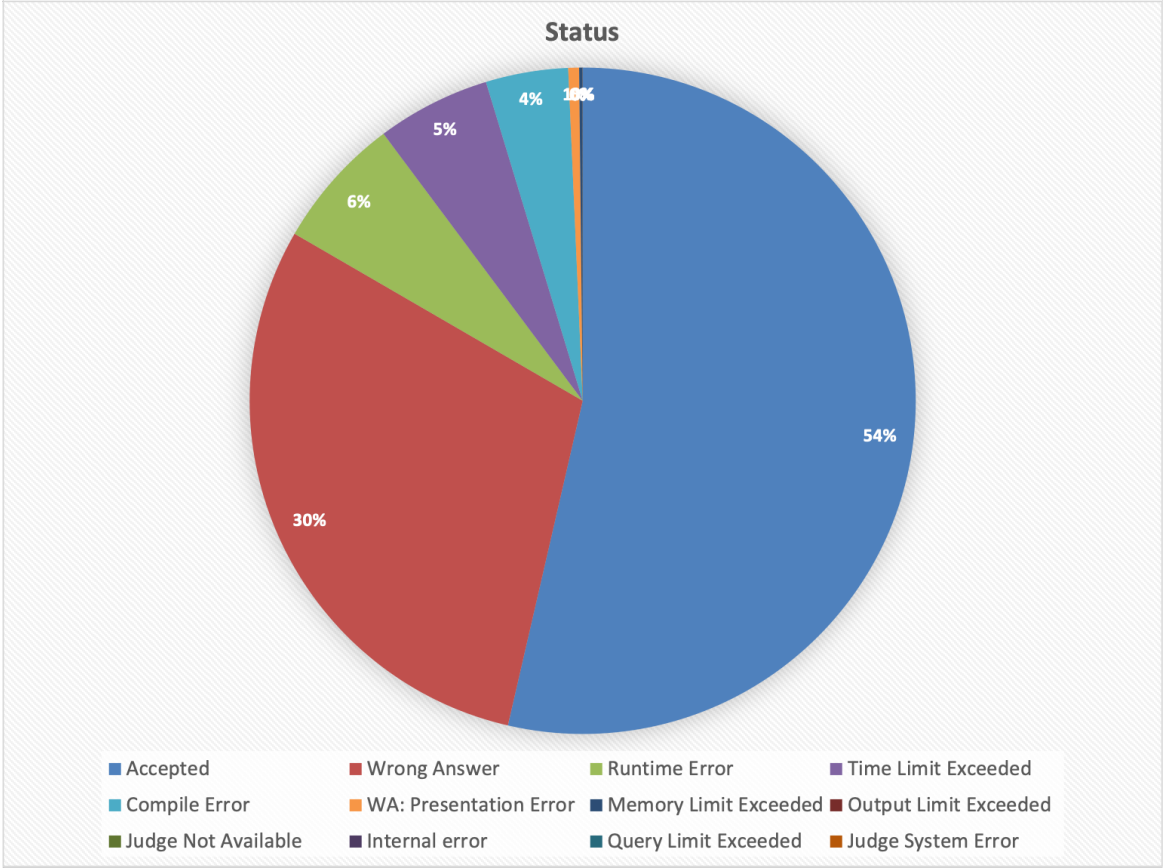
name of column	data type	unit	description
id	string	none	unique anonymized id of the problem
name	string	none	short name of the problem
dataset	string	none	original dataset, AIZU or AtCoder
time_limit	int	millisecond	maximum time allowed for a submission
memory_limit	int	KB	maximum memory allowed for a submission
rating	int	none	rating, i.e., difficulty of the problem
tags	string	none	list of tags separated by "|"; not used
complexity	string	none	degree of difficulty of the problem; not used

Metadata – Problem Level

name of column	data type	unit	description
submission_id	string	none	unique anonymized id of the submission
problem_id	string	none	anonymized id of the problem
user_id	string	none	anonymized user id of the submission
date	int	seconds	date and time of submission in the Unix timestamp format
language	string	none	mapped language of the submission (ex: C++14 -> C++)
original_language	string	none	original language specification
filename_ext	string	none	extension of the filename that indicates the programming language used
status	string	none	acceptance status, or error type
cpu_time	int	millisecond	execution time
memory	int	KB	memory used
code_size	int	bytes	size of the submission source code in bytes
accuracy	string	none	number of tests passed; *Only for AIZU

Metadata – Submission status codes

status	abbreviation	numeric code
Compile Error	CE	0
Wrong Answer	WA	1
Time Limit Exceeded	TLE	2
Memory Limit Exceeded	MLE	3
Accepted	AC	4
Judge Not Available	JNA	5
Output Limit Exceeded	OLE	6
Runtime Error	RE	7
WA: Presentation Error	PE	8



Tools and Examples

Tools:

- derive statistics from the dataset
- access the dataset files to make selections
- convert between popular formats
- preprocess the source files
  - generate stream of tokens tokenizer
  - parsing to tree/abstract syntax tree AST generation
  - control and data flow graph construction code analysis

Experiments:

- Graph neural network (GNN) experiments
- Masked language model
- Token-based similarity classification

Notebooks:

- Masked Language Model
- Language Classification

Real-World Applications

Potential Use Cases:

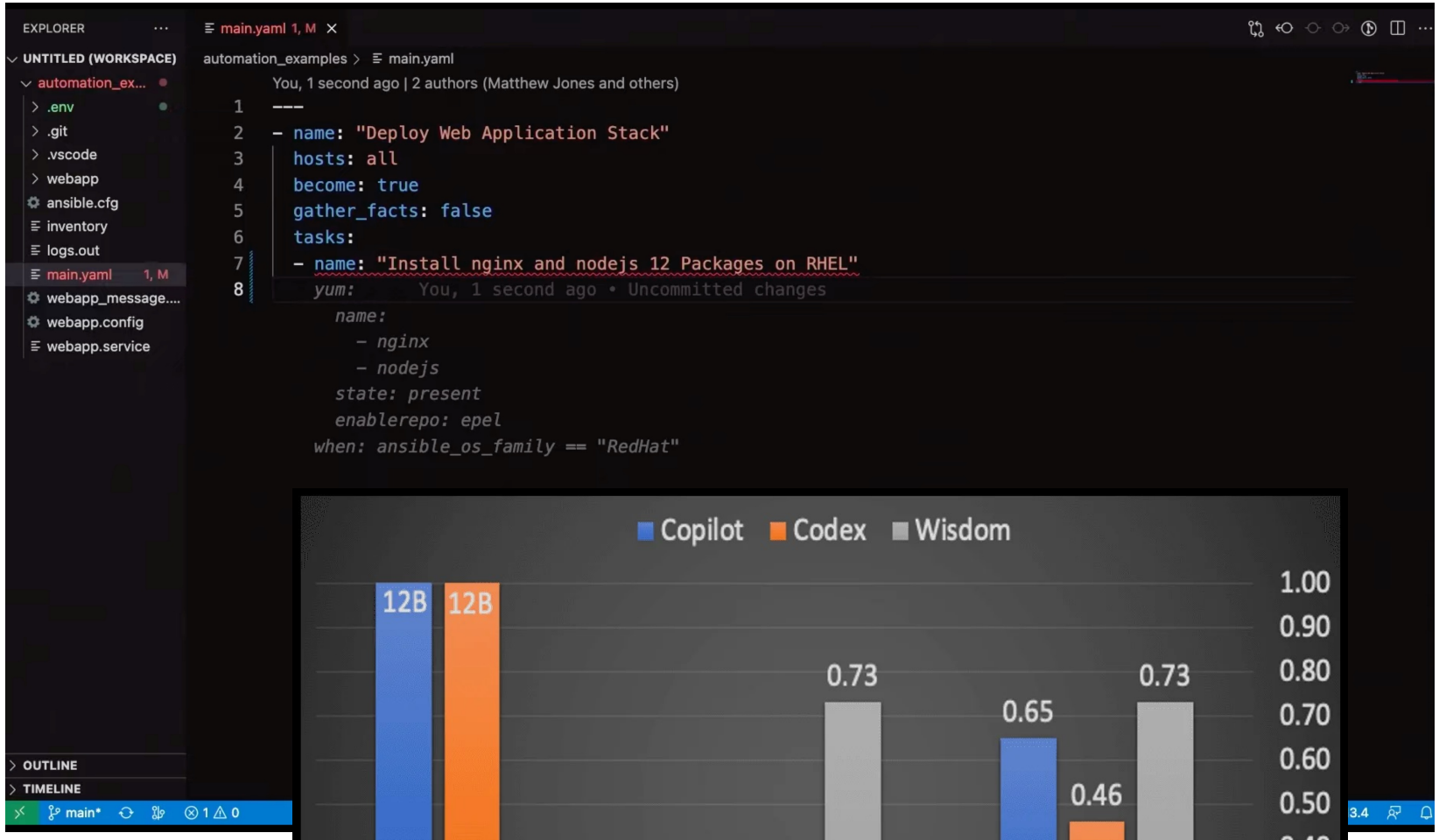
- Code classification
- Code similarity search
- Source-to-source code translation
- Modernizing legacy software systems
  - 220 Billion lines of COBOL used in finance
- Help developers write better code faster
  - Use natural language to generate code
  - Improve existing code performance and memory footprint
  - Find errors and debug code
  - Automatic test generation

Existing Applications:

- IBM AI for Code Stack
- DeepMind AlphaCode
  - machine to compete with human programmers 50-60% accuracy
  - used CodeNet as one of training data sources

IBM Research: Project Wisdom

- Generate Red Hat ansible playbook pipeline using plain English
- Automation and Infrastructure as code
- Aim to develop foundation models that maintain the highest levels of accuracy possible while relying on a smaller computing footprint.



https://research.ibm.com/blog/ai-for-code-project-wisdom-red-hat

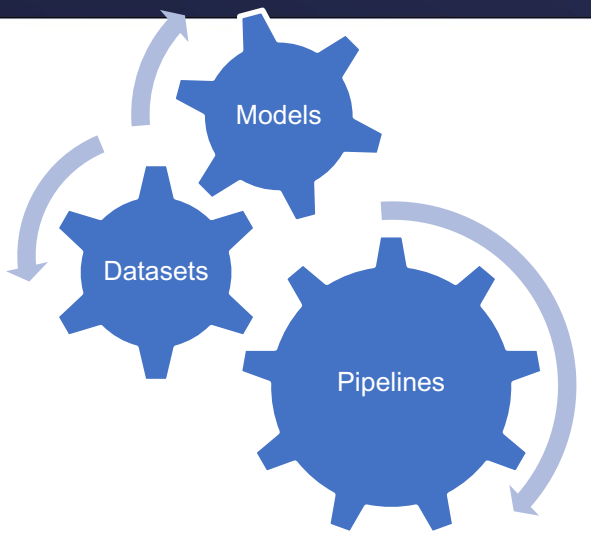
How do we share our findings?

The Machine Learning eXchange (MLX)

IBM

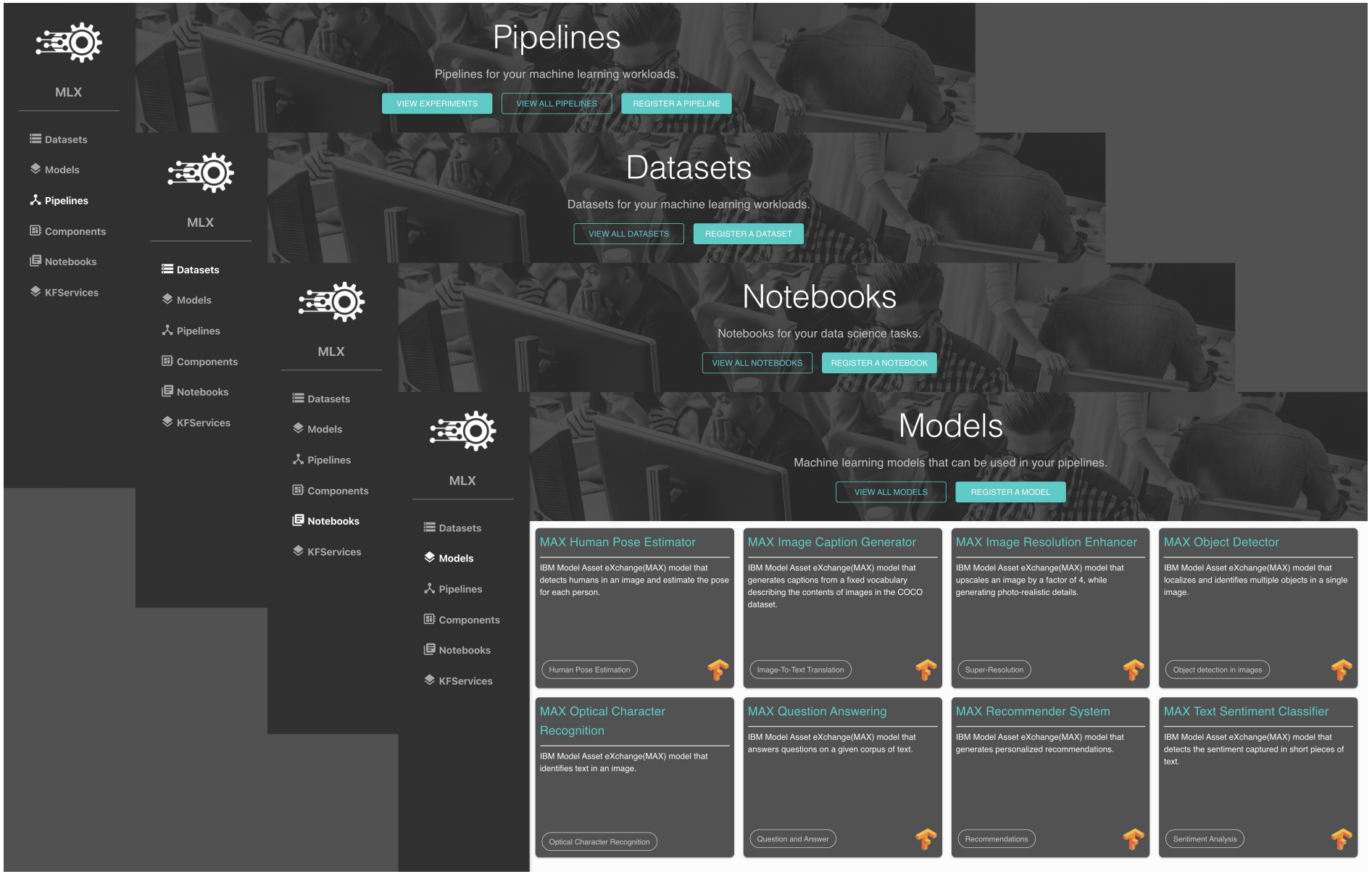
+

LF AI & DATA



Machine Learning eXchange

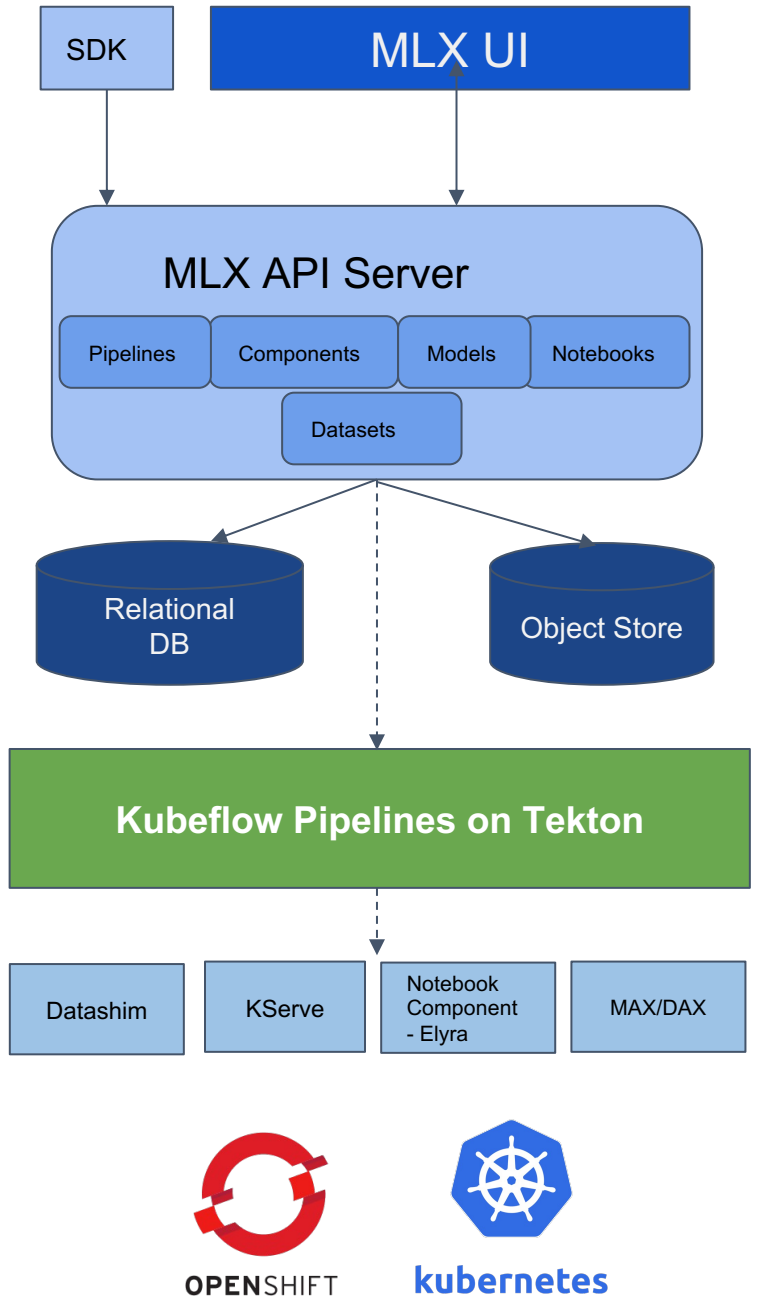
Machine Learning eXchange (MLX): Data and AI Assets Catalog and Execution Engine



Machine Learning eXchange (MLX)

Data and AI Assets Catalog and Execution Engine

- Upload, register, execute, and deploy
  - AI Pipelines and Components
  - Models
  - Datasets
  - Notebooks
- Automated sample pipeline code generation to train, validate, serve your registered models, datasets and notebooks
- Pipelines Engine powered by **Kubeflow Pipelines on Tekton**, core of Watson Studio Pipelines
- Serving engine by **Kserve or Plane Kubernetes deployment**
- Datasets Management by **DataShim**
- Preregistered Datasets from Data Asset Exchange (DAX) and Models from Model Asset Exchange (MAX)
- Model Metadata schema aligned with **MLSpec**



MLX Catalog – High Quality Curated Content

Pipelines

- Trusted AI Pipeline (with AI Fairness 360 and Adversarial Robustness 360)
- Training and Serving Models with Watson Machine Learning
- Lightweight Python Component
- The Flip-Coin Pipeline
- Hyperparameter Tuning using Katib
- A Nested Pipeline
- Pipeline with Nested Loops

Pipeline Components

- Generate Dataset Metadata
- Create Dataset Volume with DataShim
- Create Kubernetes Secret
- Kubernetes Model Deploy
- Create Model Config
- Model Fairness Check
- Adversarial Robustness Evaluation

Models

- Human Pose Estimator
- Image Caption Generator
- Image Resolution Enhancer
- Object Detector
- Optical Character Recognition
- Question Answering
- Recommender System
- Text Sentiment Classifier
- Toxic Comment Classifier
- Weather Forecaster

Datasets

- Finance Proposition Bank
- Groningen Meaning Bank - Modified
- NOAA Weather Data - JFK Airport
- PubLayNet
- PubTabNet
- IBM Debater® Thematic Clustering of Sentences
- Project CodeNet

Notebooks

- AIF360 Bias Detection
- ART Metector Model
- ART Poisoning Attack
- JFK Airport Analysis
- Code Language Classification
- Masked Language Model

MLX Demo

https://ml-exchange.org/

Summary

Data: Project CodeNet

- A high-quality code dataset for algorithmic innovation and benchmarking

Technology: AI system stack

- Open-Source platforms such as Machine Learning Exchange provide tools for developer to exchange AI assets and convert them into use cases.

Business value

- Use cases such as code translation can create new business value to help migrate and modernize legacy code.

KubeCon | CloudNativeCon
North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

Please scan the QR Code above to
leave feedback on this session