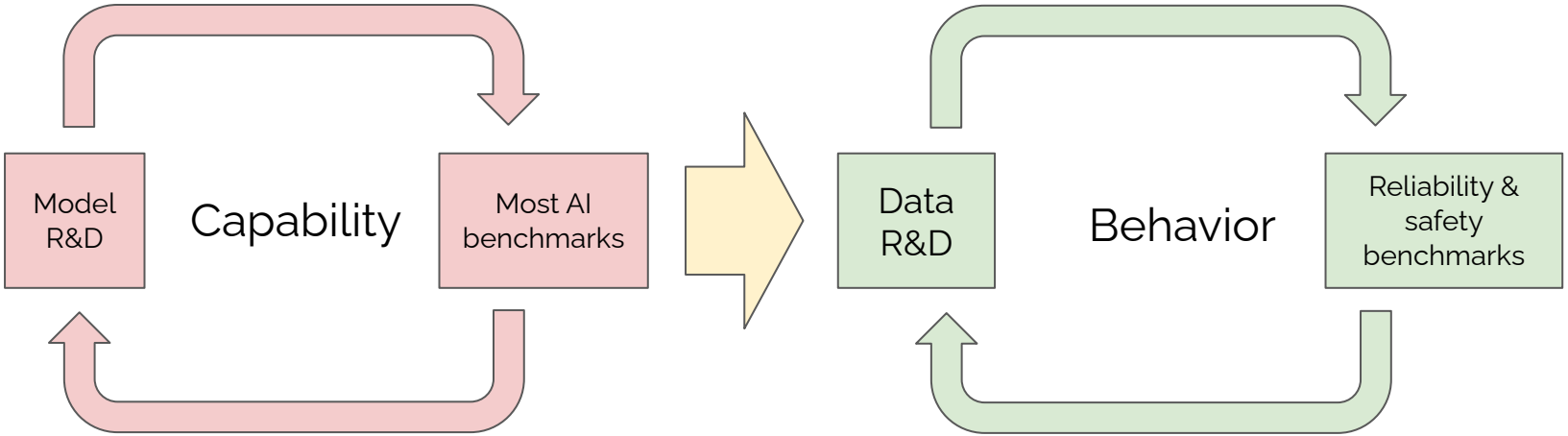
Croissant: A Metadata Format for ML-ready Datasets

MLCommons Croissant Working Group

11/2024

NEURAL INFORMATION PROCESSING SYSTEMS

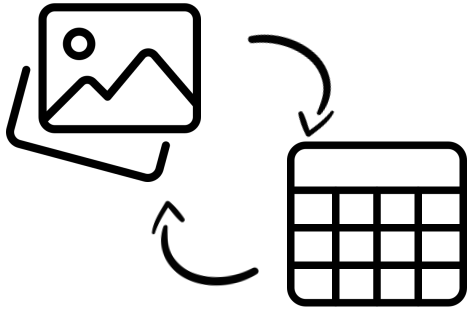
Motivation



How to build reliable and safe AI? Better training/test data.

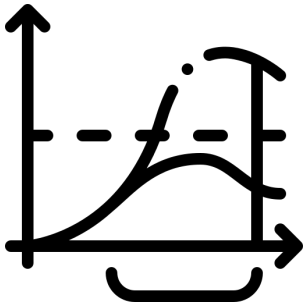
NEURAL INFORMATION PROCESSING SYSTEMS

What makes ML-ready data special?



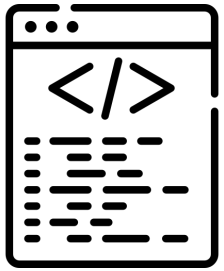
- Often combine **unstructured** (text, image, video) and **structured** (tabular, json) data

What makes ML-ready data special?

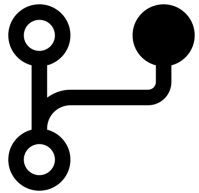


- Often combine **unstructured** (text, image, video) and **structured** (tabular, json) data
- Need to be **"flattened" / denormalized** to be used in ML frameworks and tools

What makes ML-ready data special?



- Often combine **unstructured** (text, image, video) and **structured** (tabular, json) data
- Need to be **"flattened" / denormalized** to be used in ML frameworks and tools
- Need **ML-specific metadata**
  (e.g., Responsible AI info, test/train/validation splits, labels)



What makes ML-ready data special?

- Often combine **unstructured** (text, image, video) and **structured** (tabular, json) data
- Need to be **"flattened" / denormalized** to be used in ML frameworks and tools
- Need **ML-specific metadata**
  (e.g., Responsible AI info, test/train/validation splits, labels)
- Require **versioning / checkpointing** to support model snapshots and reproducibility

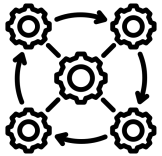
What is Croissant?

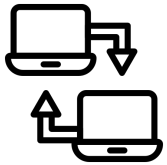
A metadata format designed for ML dataset, based on **schema.org** vocabulary

What Croissant offers!

Discoverability

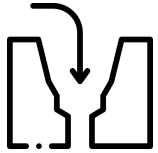
Interoperability





Portability

Gap in Non-Standardization



NEURAL INFORMATION PROCESSING SYSTEMS

What Croissant offers!

A metadata format designed for ML dataset, based on **schema.org** vocabulary

Croissant aims to improve data:

- **Discoverability** => integration in repositories (e.g., HF & Kaggle Datasets) & Google Dataset Search
- Interoperability
- Portability

What Croissant offers!

A metadata format designed for ML dataset, based on **schema.org** vocabulary

Croissant aims to improve data:

- Discoverability
- **Interoperability** => `mlcroissant` library interfacing w/ data loaders (e.g., TensorFlow Datasets & PyTorch DataPipes) allow loading data from any repository
- Portability

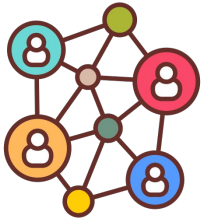
NEURAL INFORMATION PROCESSING SYSTEMS

What Croissant offers!

A metadata format designed for ML dataset, based on **schema.org** vocabulary

Croissant aims to improve data:

- Discoverability
- Interoperability
- **Portability** => across ML frameworks and tools



What Croissant offers!

A metadata format designed for ML dataset, based on **schema.org** vocabulary

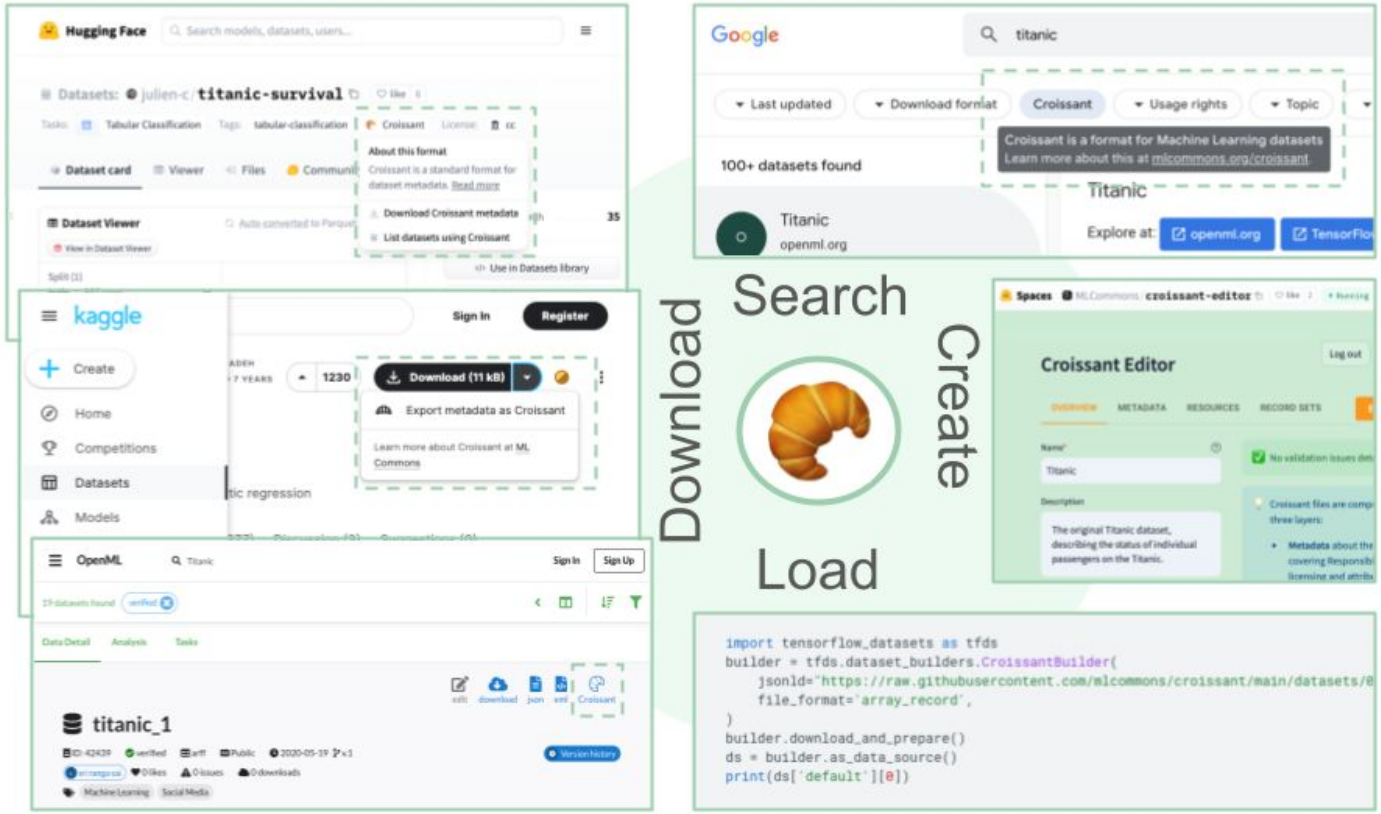
Croissant aims to improve data:

- Discoverability
- Interoperability
- Portability

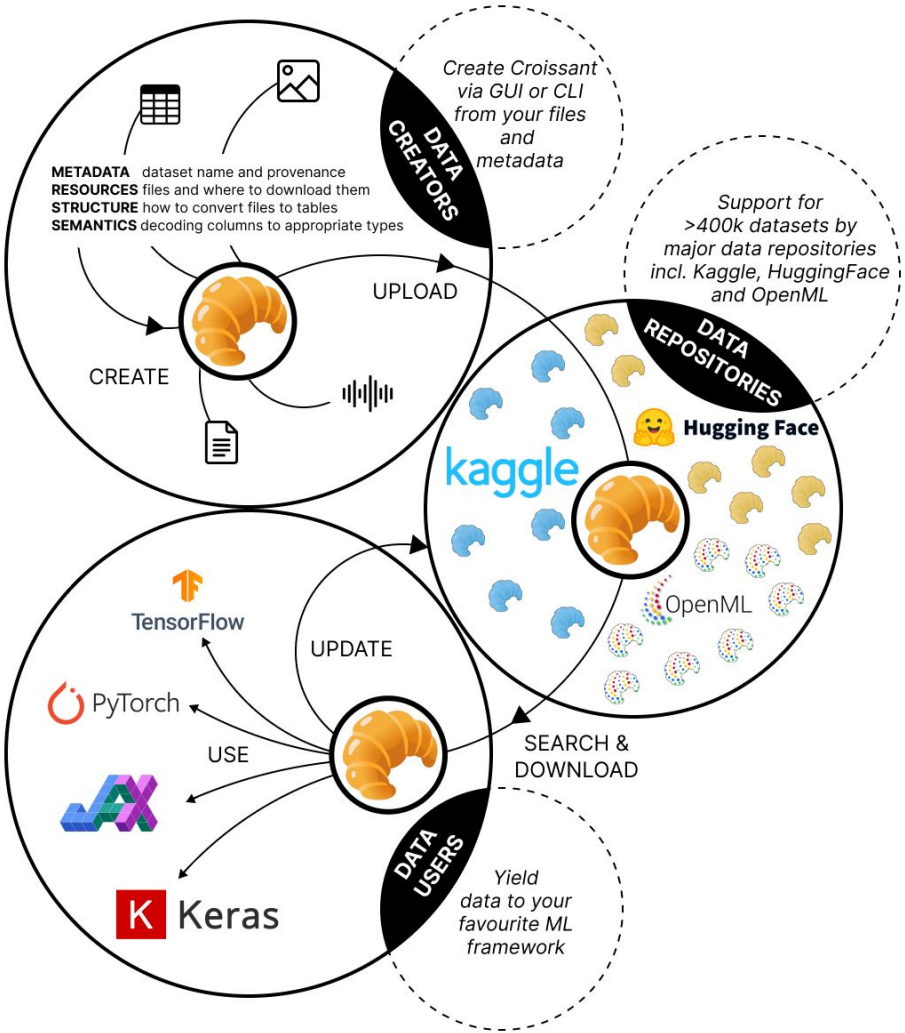
Adoption: A recommended artifact in the NeurIPS Datasets and Benchmarks Track 2024

NEURAL INFORMATION PROCESSING SYSTEMS

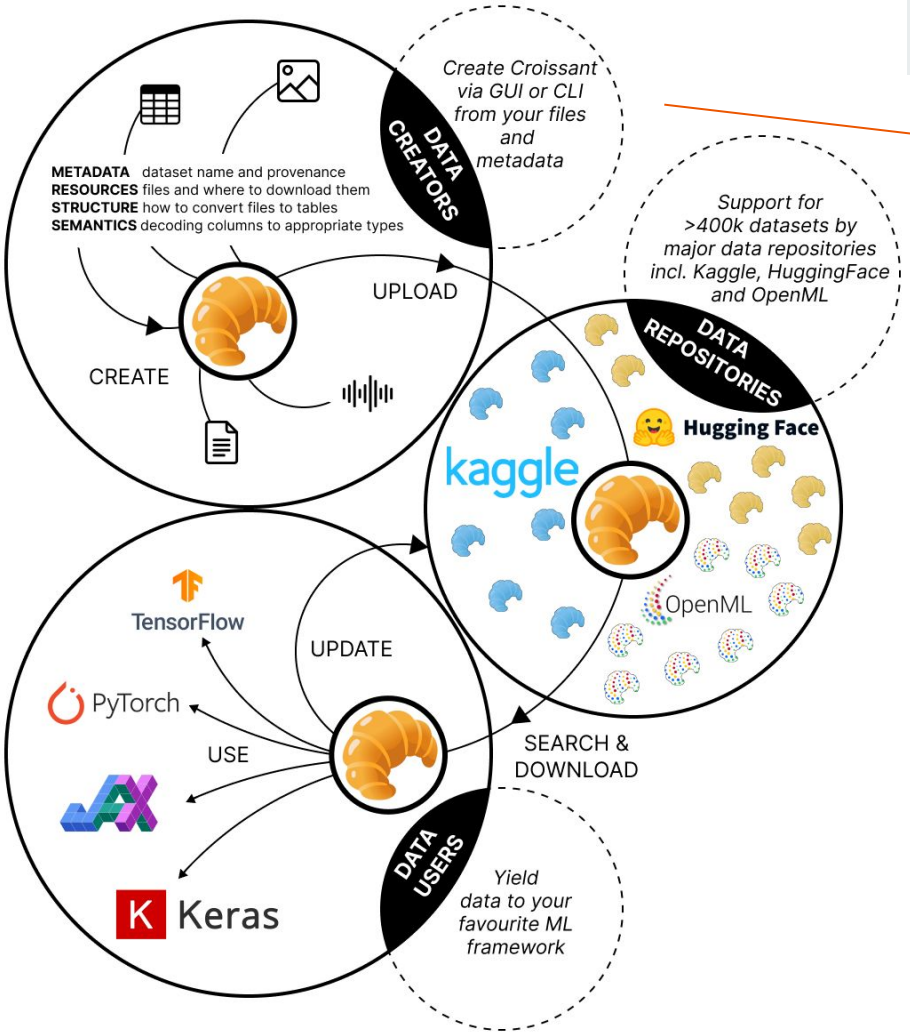
Croissant Features



NEURAL INFORMATION PROCESSING SYSTEMS

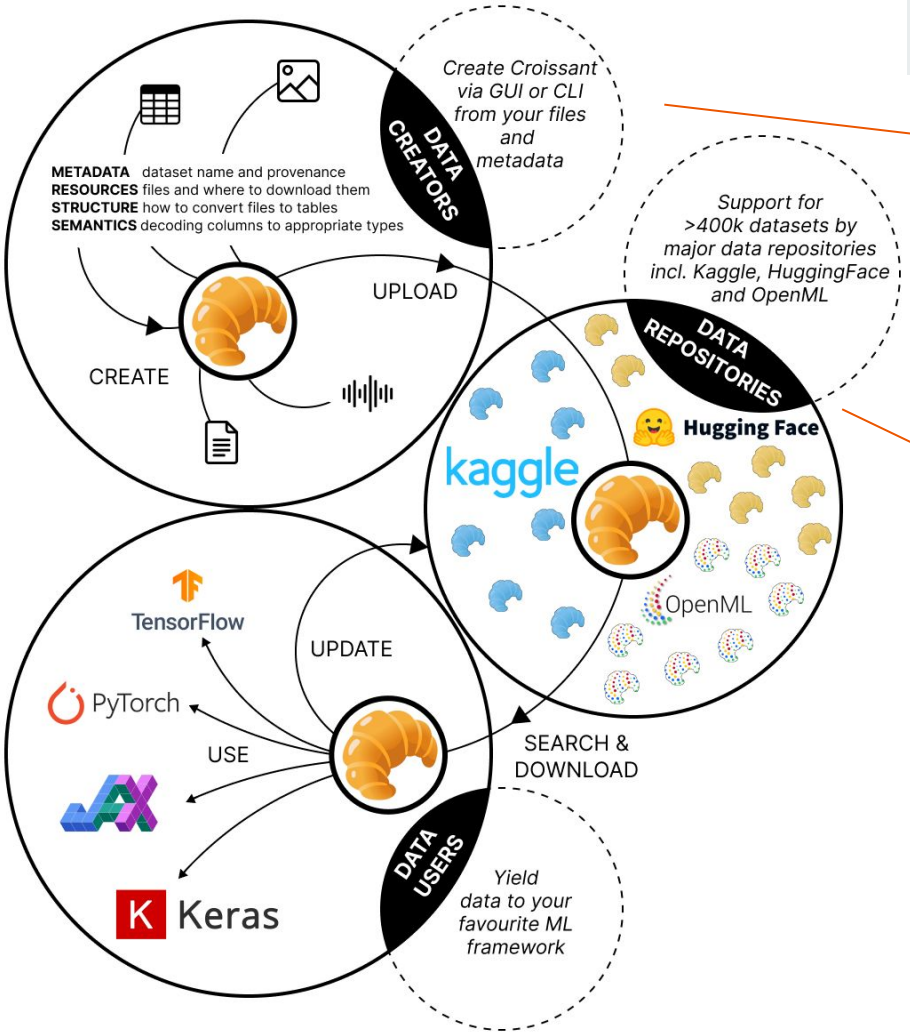


NEURAL INFORMATION PROCESSING SYSTEMS



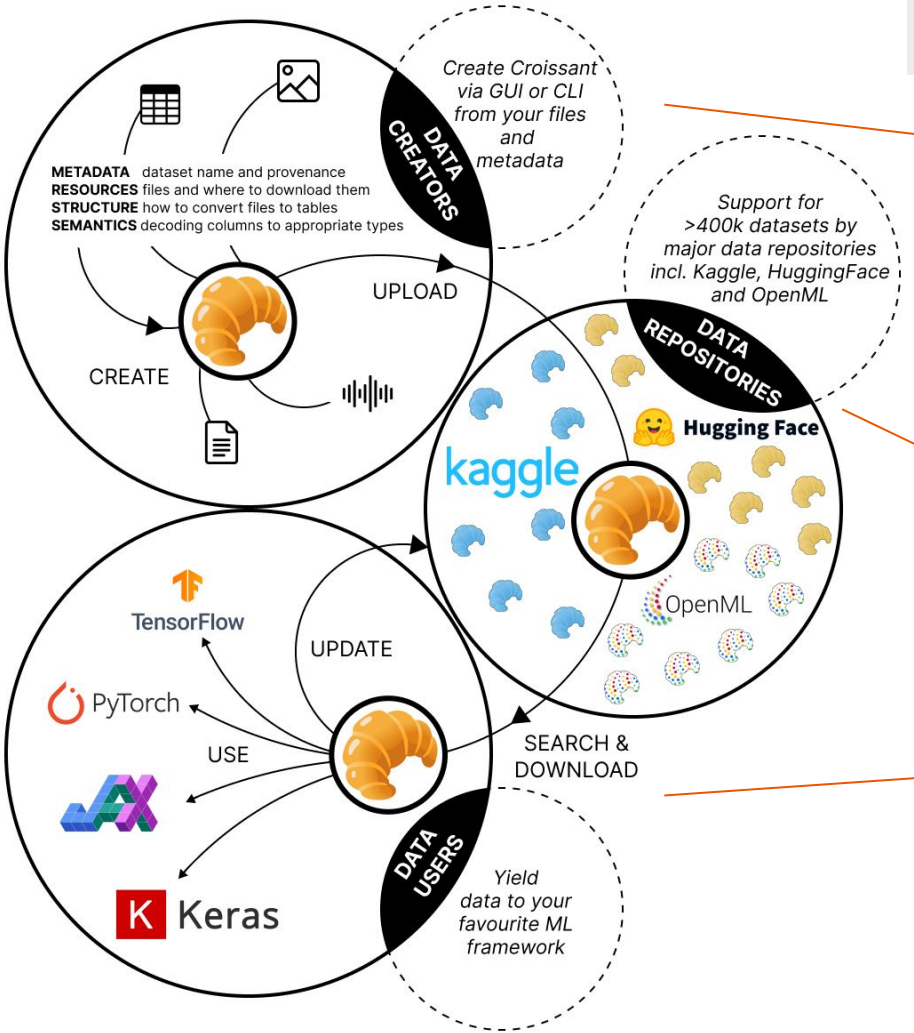
- Create
  - Editor
    - https://huggingface.co/spaces/MLCommons/croissant-editor
  - Platform auto generate
    - https://huggingface.co/datasets
    - https://www.kaggle.com/datasets

NEURAL INFORMATION PROCESSING SYSTEMS



- Create
  - Editor
    - https://huggingface.co/spaces/MLCommons/croissant-editor
  - Platform auto generate
    - https://huggingface.co/datasets
    - https://www.kaggle.com/datasets
- Discover and find
  - Google Dataset Search
    - https://datasetsearch.research.google.com

NEURAL INFORMATION PROCESSING SYSTEMS



- Create
  - Editor
    - https://huggingface.co/spaces/MLCommons/croissant-editor
  - Platform auto generate
    - https://huggingface.co/datasets
    - https://www.kaggle.com/datasets
- Discover and find
  - Google Dataset Search
    - https://datasetsearch.research.google.com
- Use it
  - Colab [https://githubtocolab.com/mlcommons/croissant/blob/main/python/mlcroissant/recipes/tfds_croissant_builder.ipynb]

NEURAL INFORMATION PROCESSING SYSTEMS

Human Evaluation

- Evaluation of Croissant metadata format through a user study with ML practitioners
- Focus on annotating datasets and assessing metadata completeness, conciseness, readability, and understandability
- Annotations collected across ten datasets from diverse modalities (language, vision, audio).
- Few insights gained:
  - High readability and understandability ratings, with noted challenges for specific attribute pairs (e.g., `sc:creator` vs. `sc:publisher`)
  - Higher agreement among annotators for Croissant over Croissant-RAI attributes
  - High confidence in annotations for over **75%** of responses

Resources

Croissant Specification

Croissant Working Group

Croissant RAI Paper

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!

NEURAL INFORMATION PROCESSING SYSTEMS