Carnegie Mellon University

Recitation 0 - Datasets

Introduction to Deep Learning- 11-485/685/785

Mugur Preda, Mengchun Zhang, Bhiksha Raj

December, 2025

Why do we need Datasets and Dataloaders ?

- **Modularization**

  Dataset abstracts **what the data is** and how to access a single sample.

  DataLoader abstracts **how to iterate** over data efficiently. ***How*** the data is fed into the model (batching, shuffling, etc).

- **Efficiency**

  Load data **lazily per batch**, not all at once.

  Enable **multi-process loading** via num_workers.

- **Training Support**

  Handles **batching**, **shuffling**, and **epoch iteration**.

  Essential for **mini-batch gradient descent**.

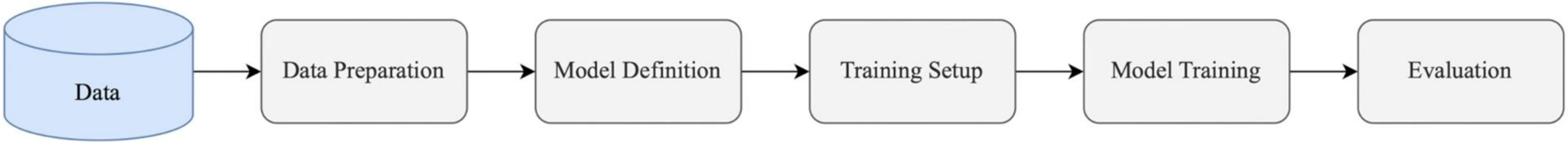
- **Customization**

  Integrates **transformations and preprocessing** (e.g., normalization, augmentation).

  Supports custom datasets (images, text, EEG, etc).

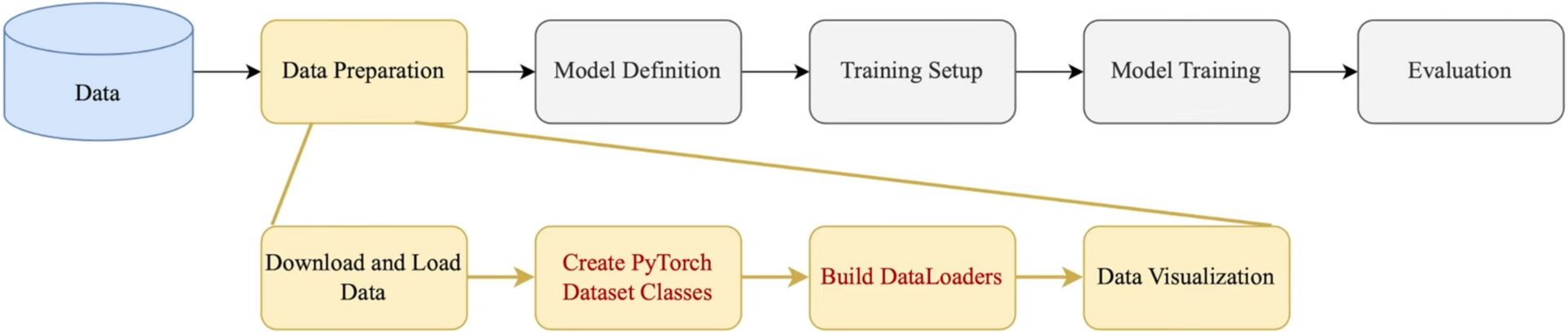
Why do we need Datasets and Dataloaders ?

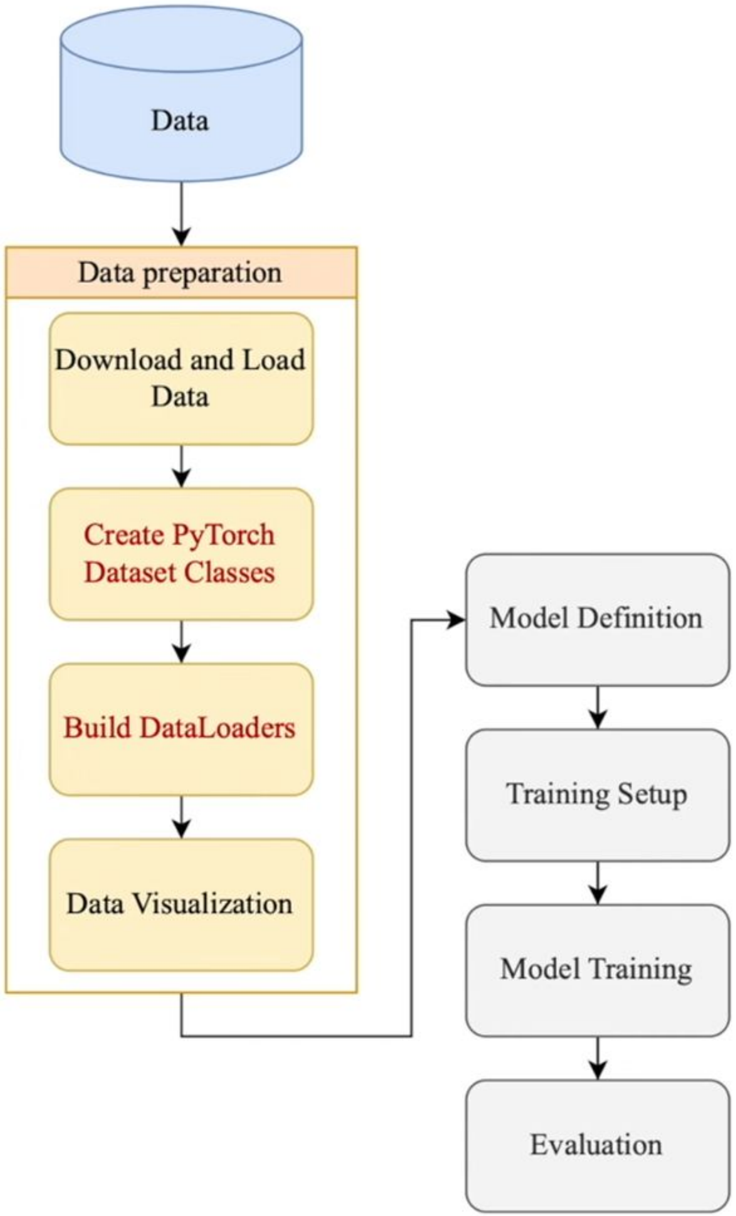
- DL Training Pipeline



Why do we need Datasets and Dataloaders ?

- DL Training Pipeline





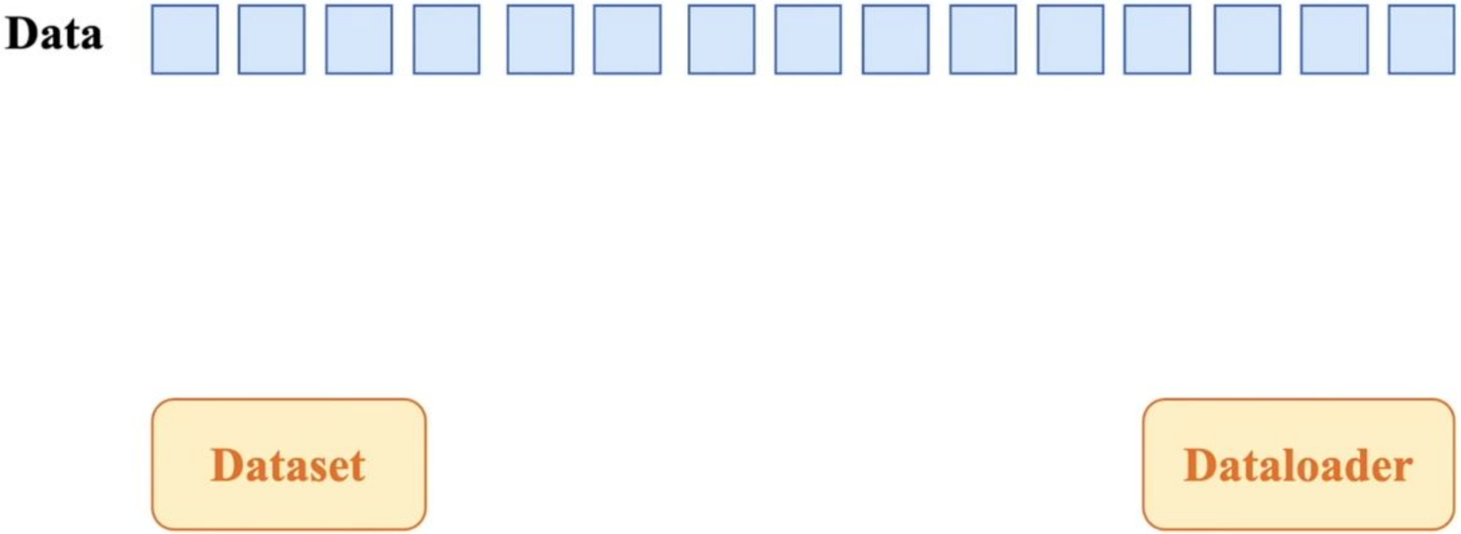
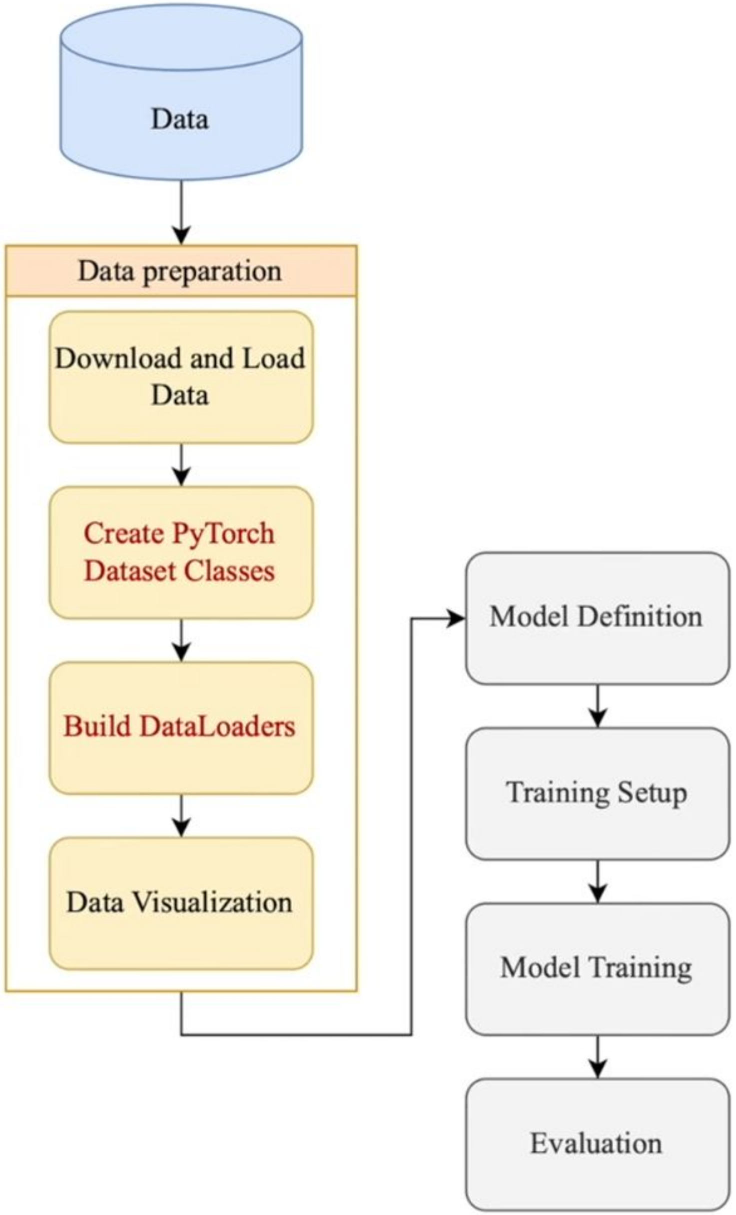


Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



Data

Data preparation

Download and Load Data

Create PyTorch Dataset Classes

Build DataLoaders

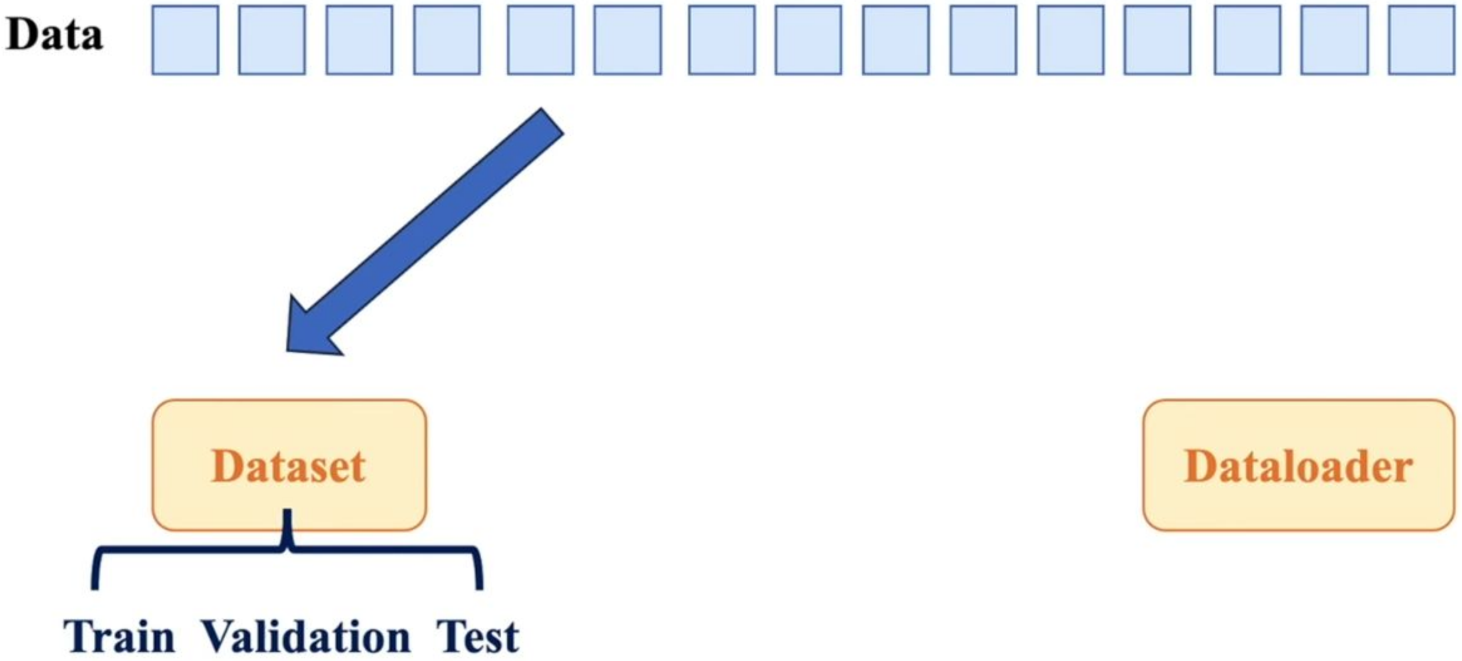
Data Visualization

Model Definition

Training Setup

Model Training

Evaluation



Data

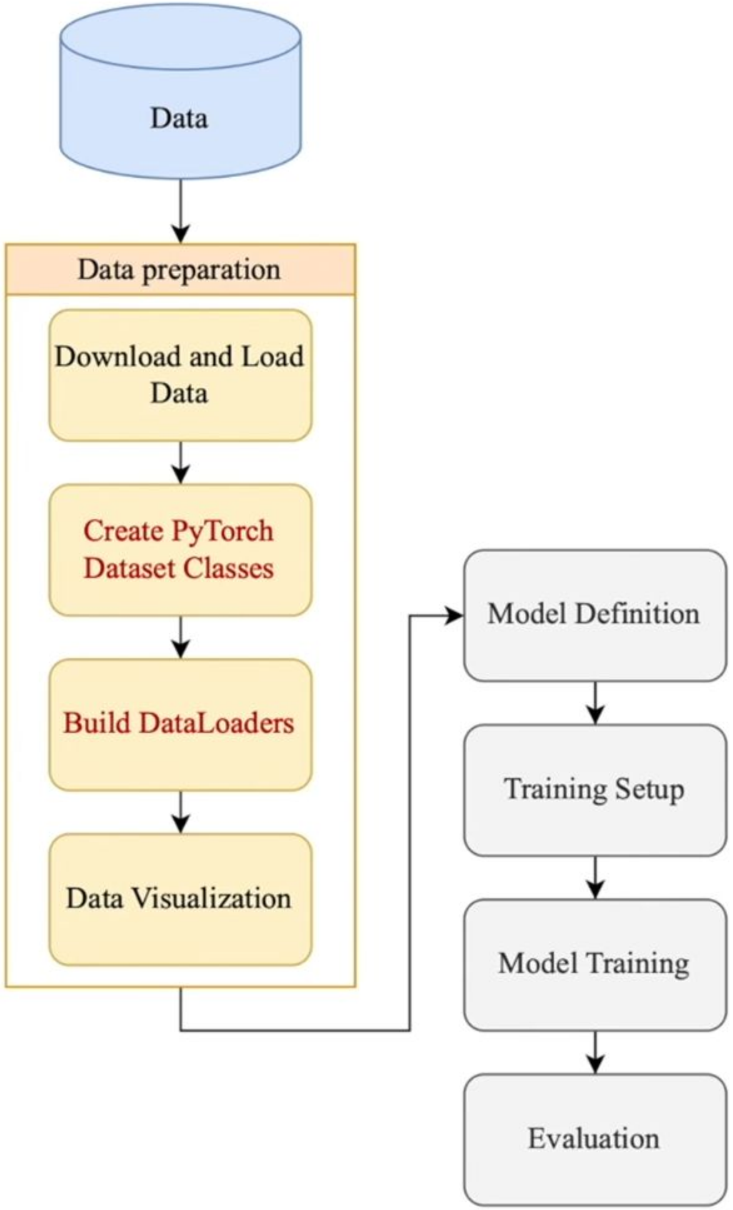
Dataset

Train Validation Test

Dataloader

Carnegie Mellon University

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



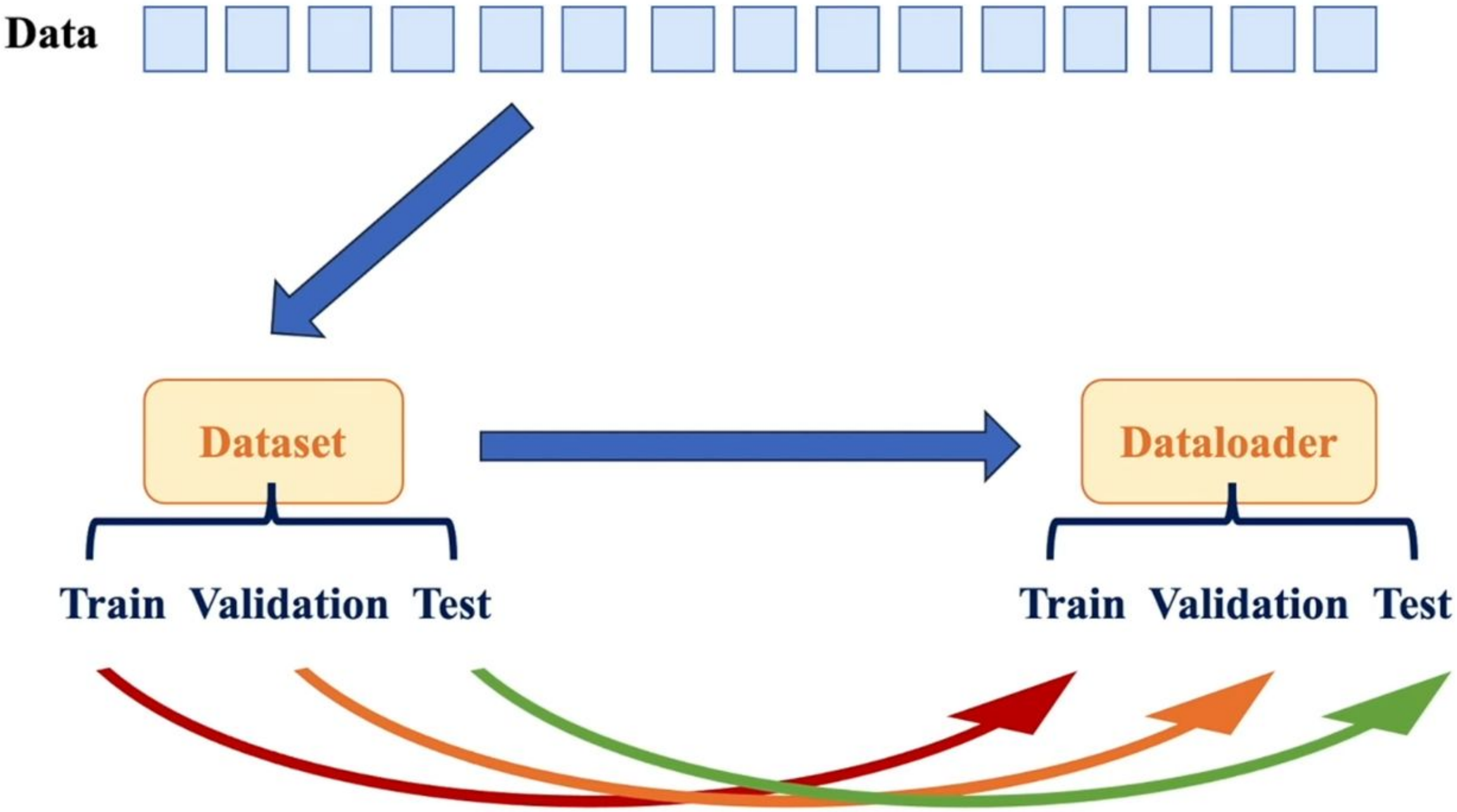
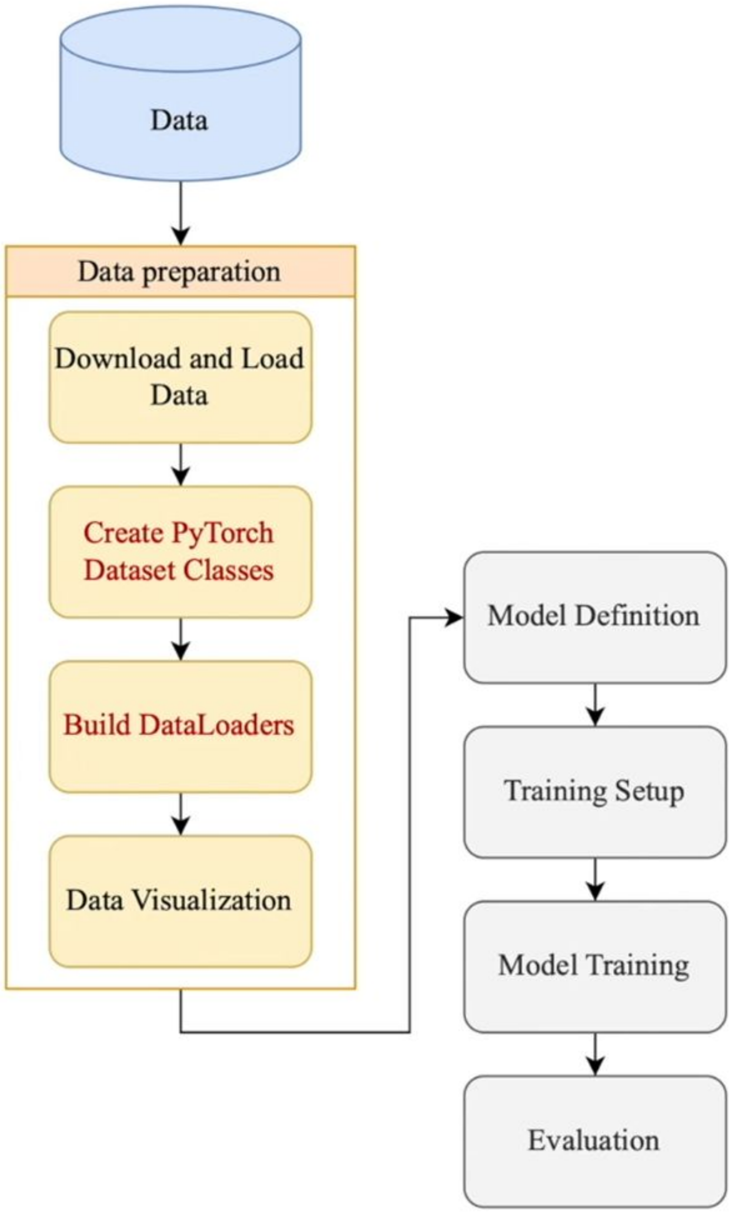


Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



Data

Data preparation

Download and Load Data

Create PyTorch Dataset Classes

Build DataLoaders

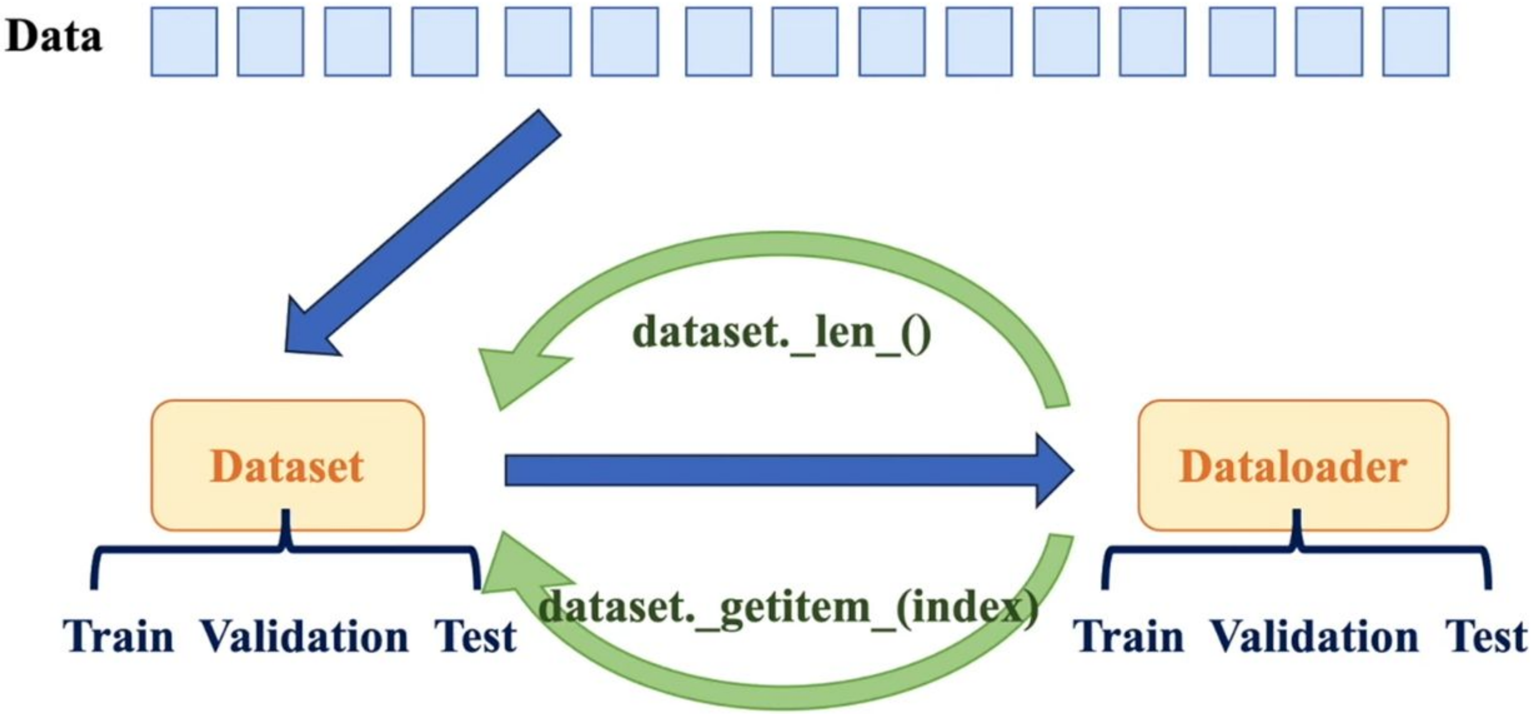
Data Visualization

Model Definition

Training Setup

Model Training

Evaluation



Data

dataset._len_()

Dataset

Dataloader

dataset._getitem_(index)

Train Validation Test

Train Validation Test

Access to the data itself

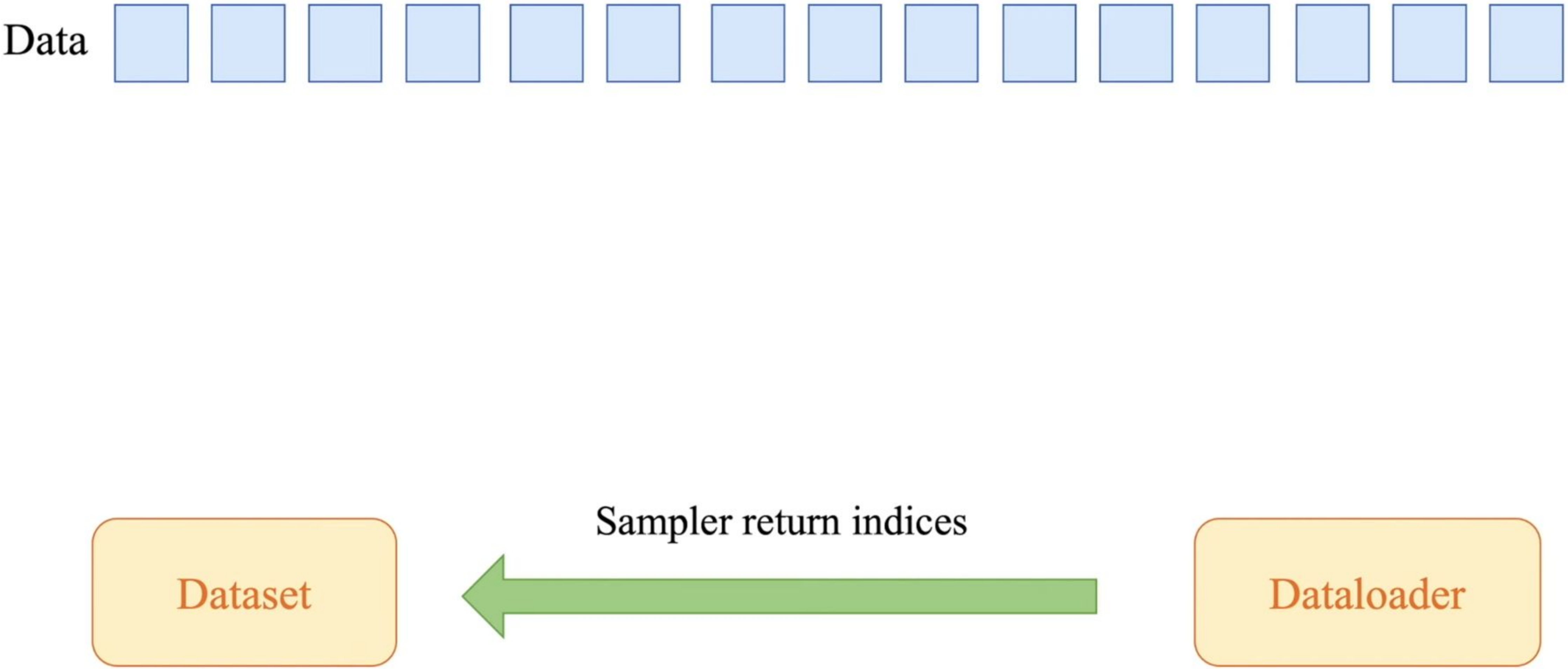
How to load the data (batching, shuffling, multiprocessing, etc)

Data

Dataset

Dataloader

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



Data

Sampler return indices

Dataset

Dataloader

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.

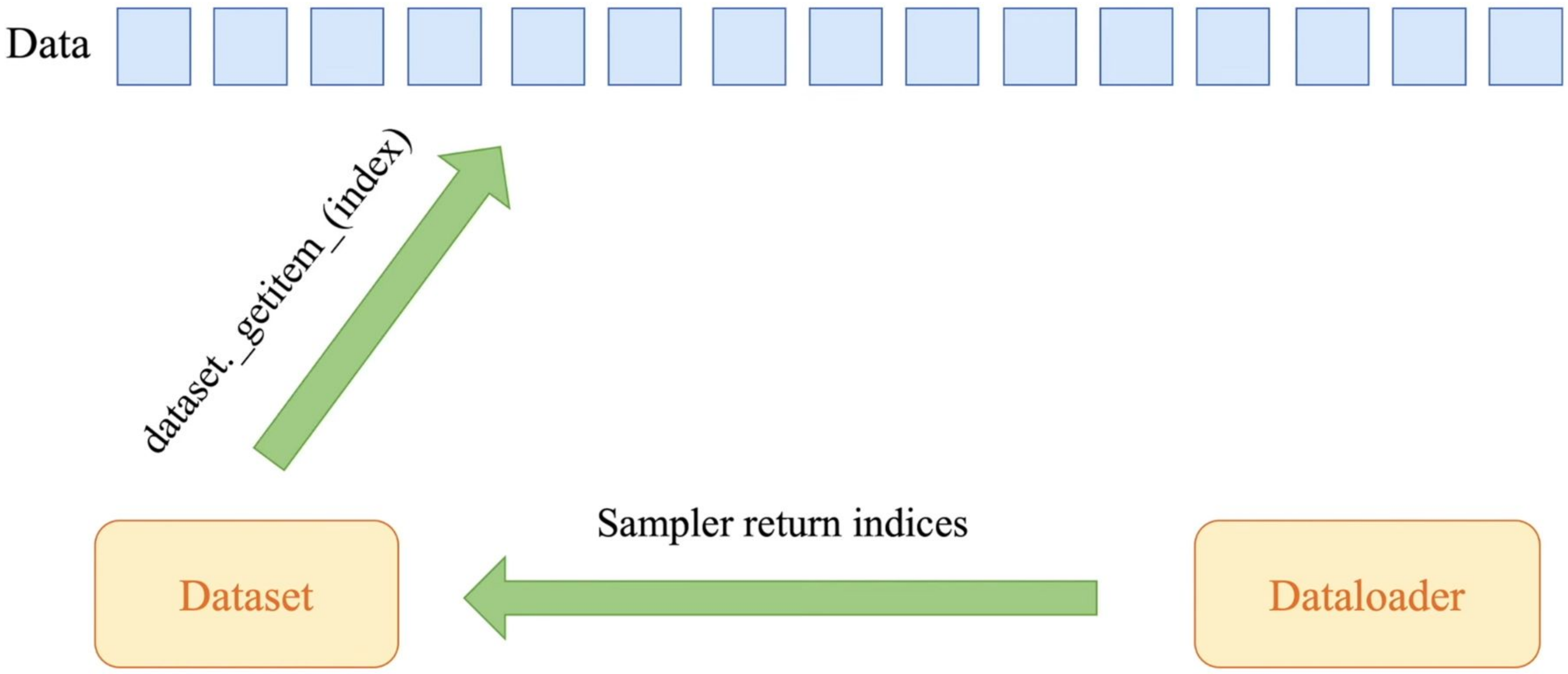
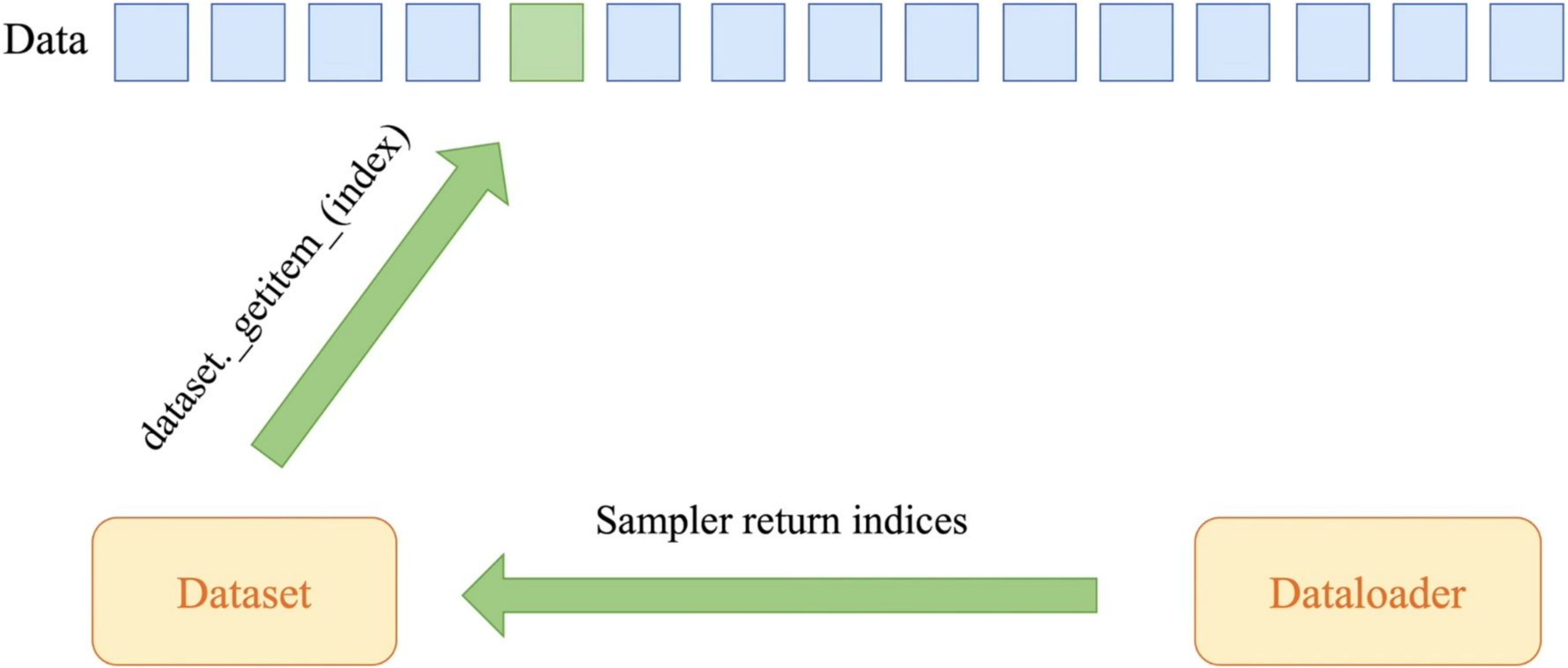


Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



Data

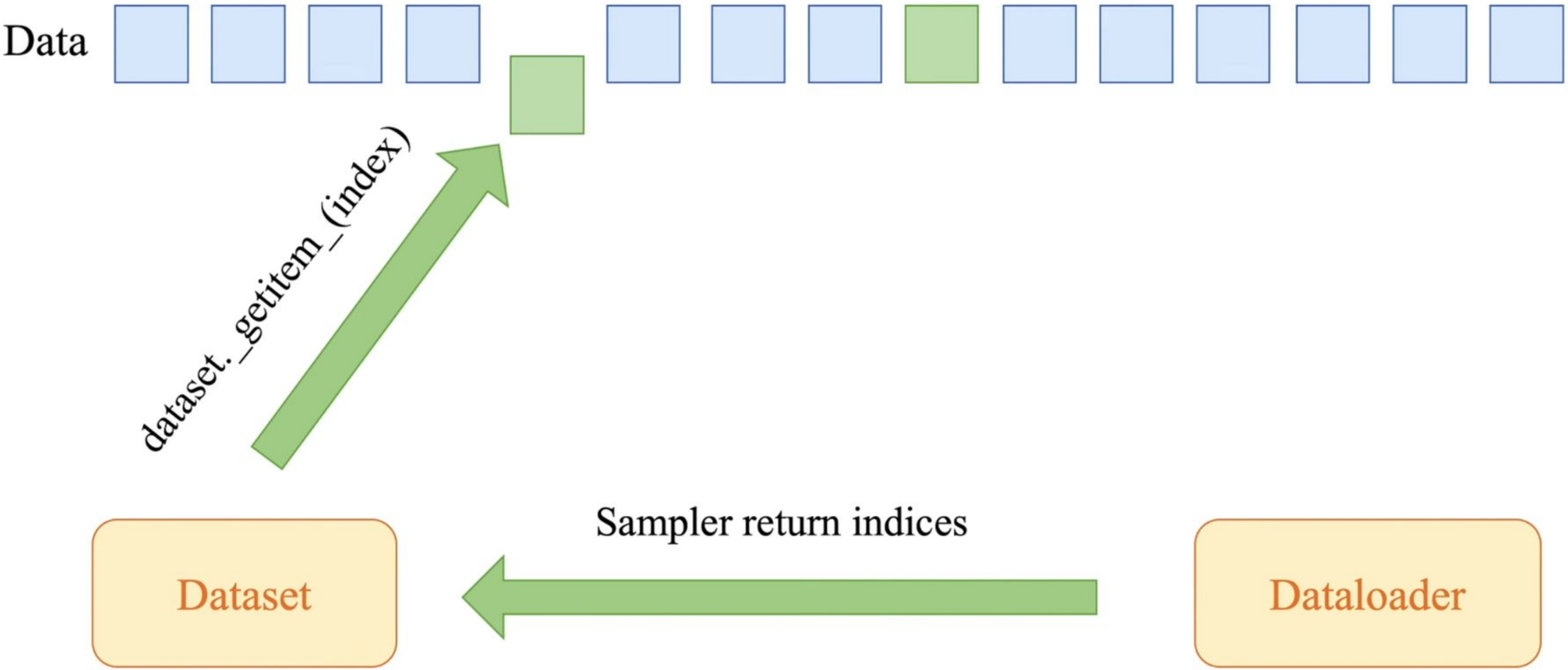
dataset._getitem_(index)

Sampler return indices

Dataset

Dataloader

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



Carnegie Mellon University

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.

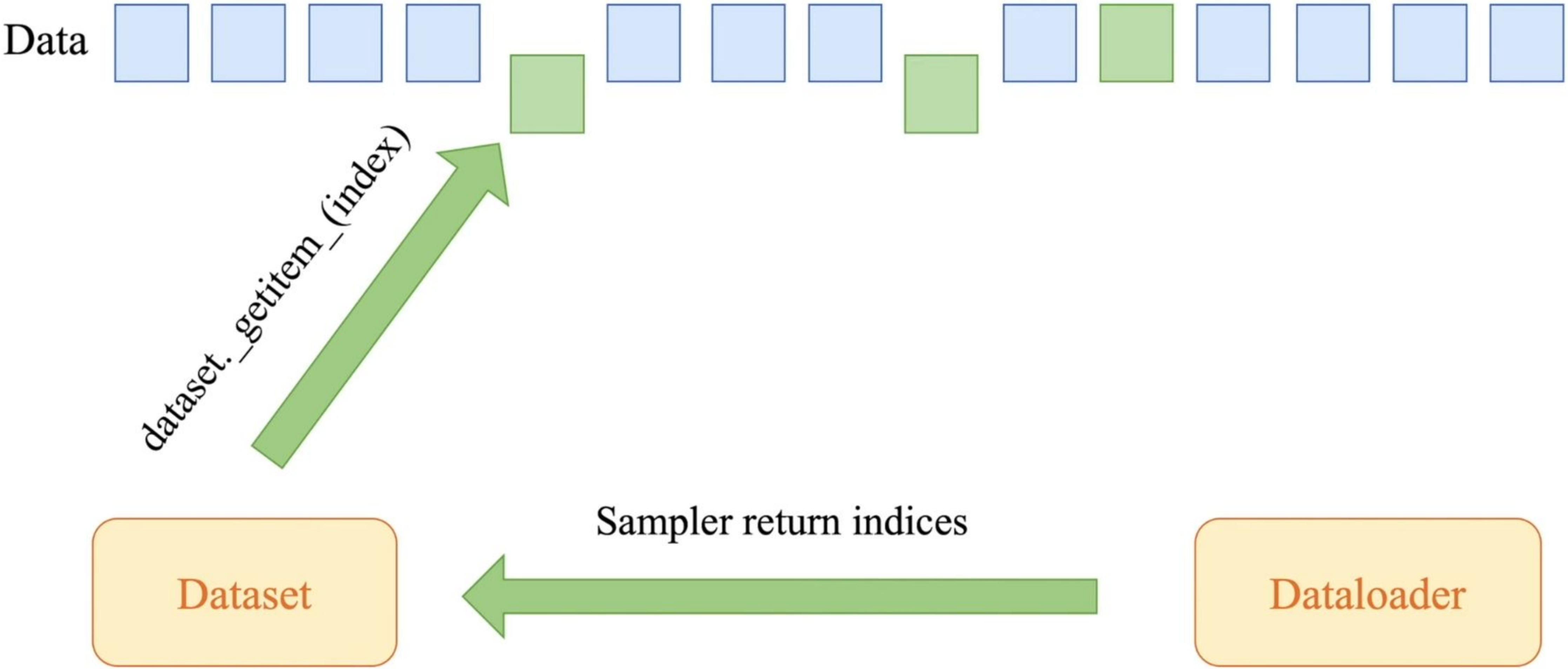
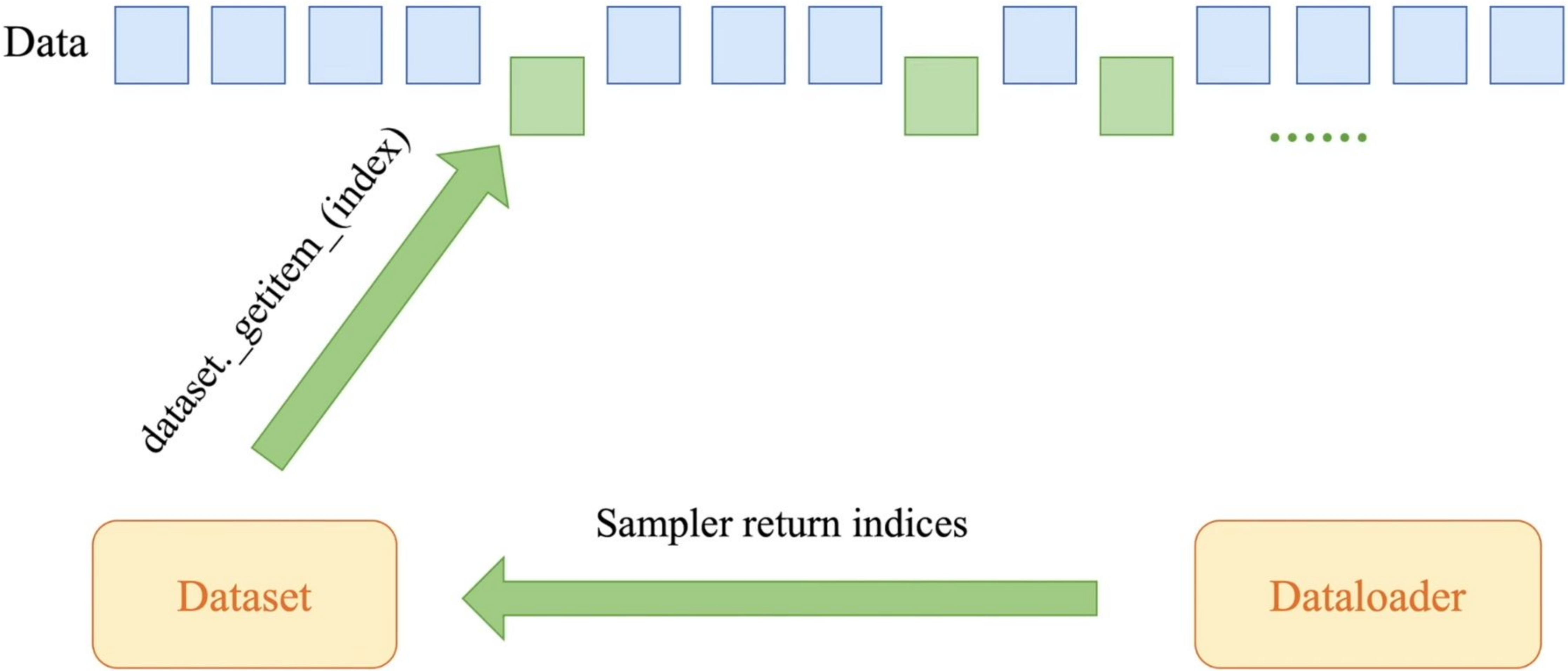


Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



Data

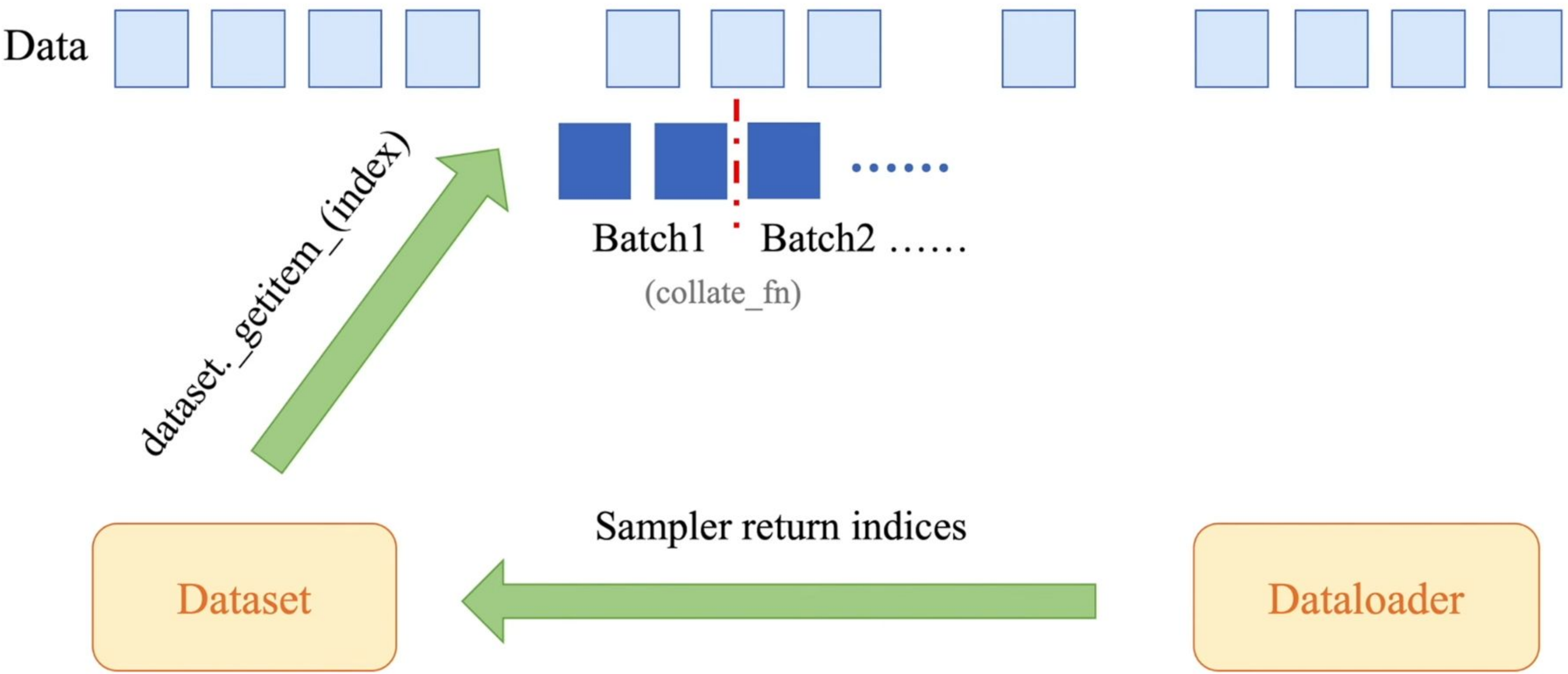
dataset.__getitem__(index)

Sampler return indices

Dataset

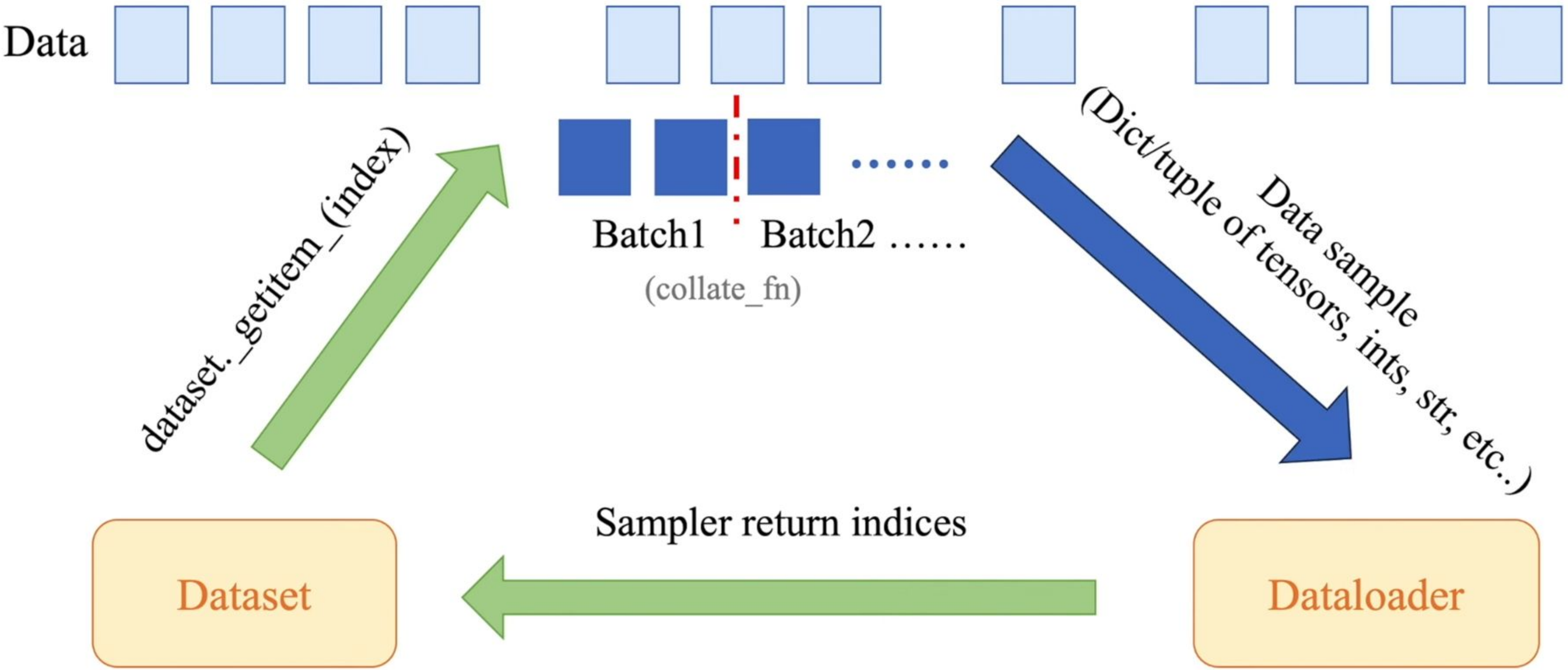
Dataloader

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



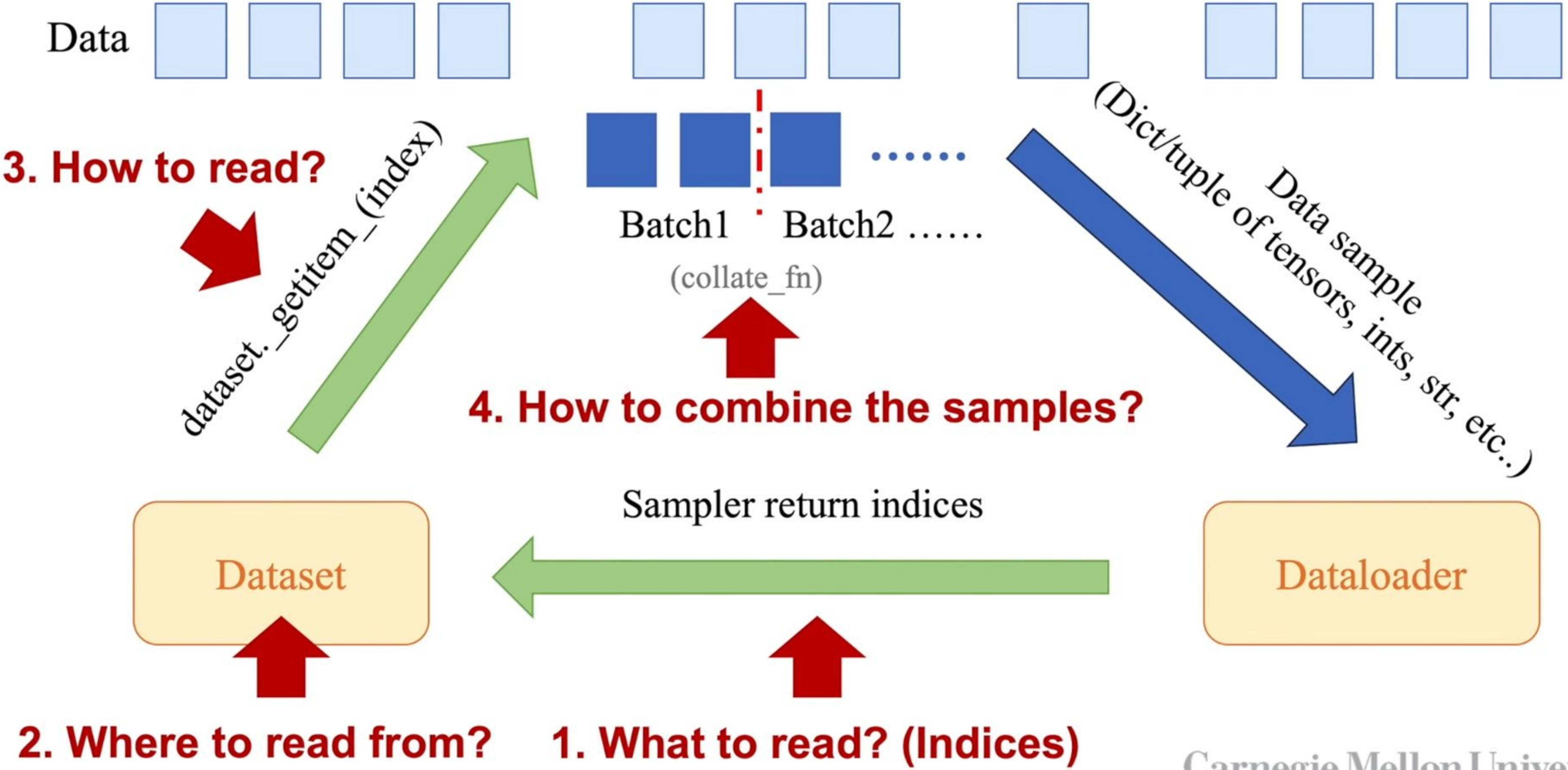
Carnegie Mellon University

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



Carnegie Mellon University

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.



Data

3. How to read?

dataset._getitem_(index)

Batch1

Batch2 ……

……

(collate_fn)

4. How to combine the samples?

Data sample

(Dict/tuple of tensors, ints, str, etc..)

Sampler return indices

Dataset

Dataloader

2. Where to read from?

1. What to read? (Indices)

Diagram adapted from *"Algorithm Researcher explains how PyTorch Datasets and DataLoaders work"* by Practical ML (YouTube). Redrawn and modified by the presenter.

Thank you!

Mugur Preda - mpreda