Quality of Datasets for Network Traffic

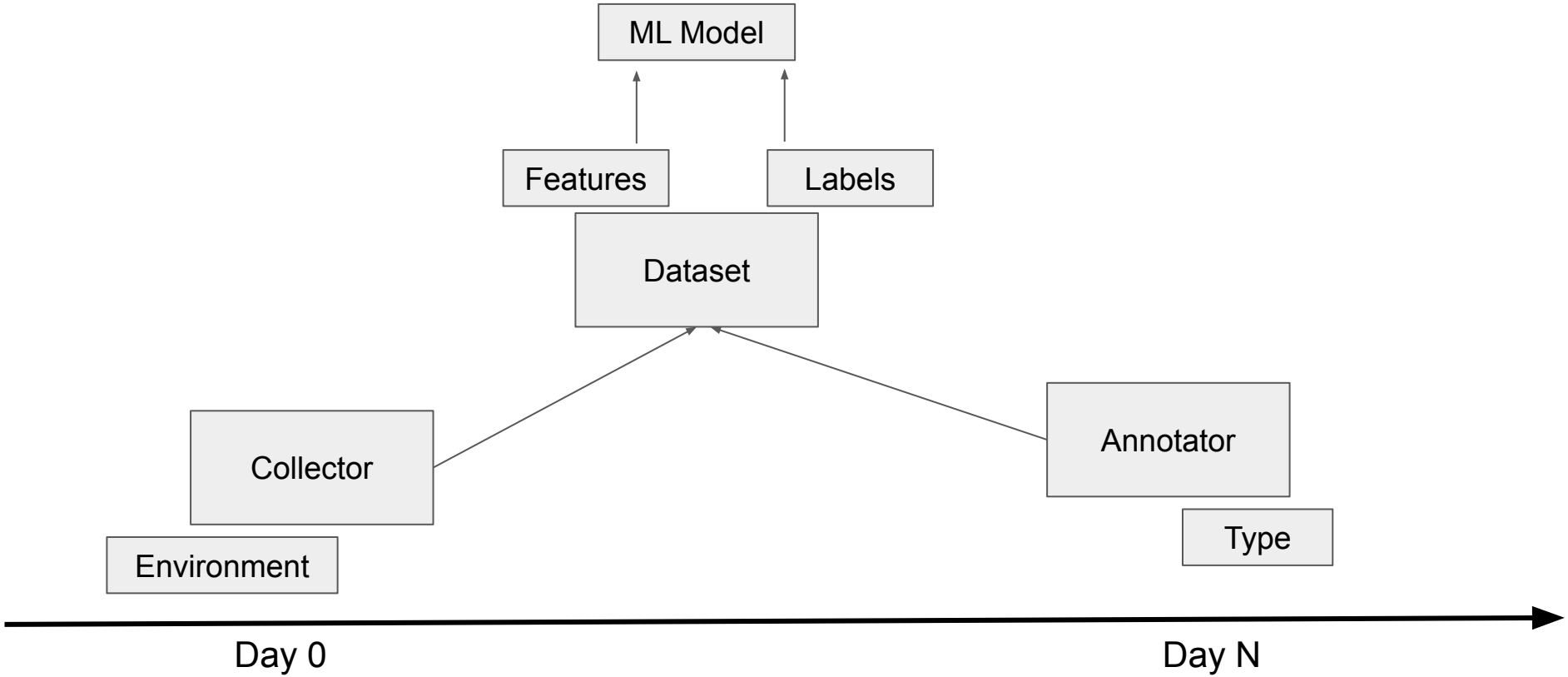
Dominik Soukup

Czech Technical University in Prague

CNSM 2024
28.10.2024

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Dataset Lifecycle



Problems To Solve

- Is public/created dataset good for my use case?
- Is public/created dataset trustful?
- How to use/optimize the dataset?
- How to capture & annotate network traffic dataset?

QoD (Quality of Datasets) Approach [1]

FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

QoD = Dataset + ML + Use Case

Tools For Dataset Operations

(Merge, Analyze, Visualize)

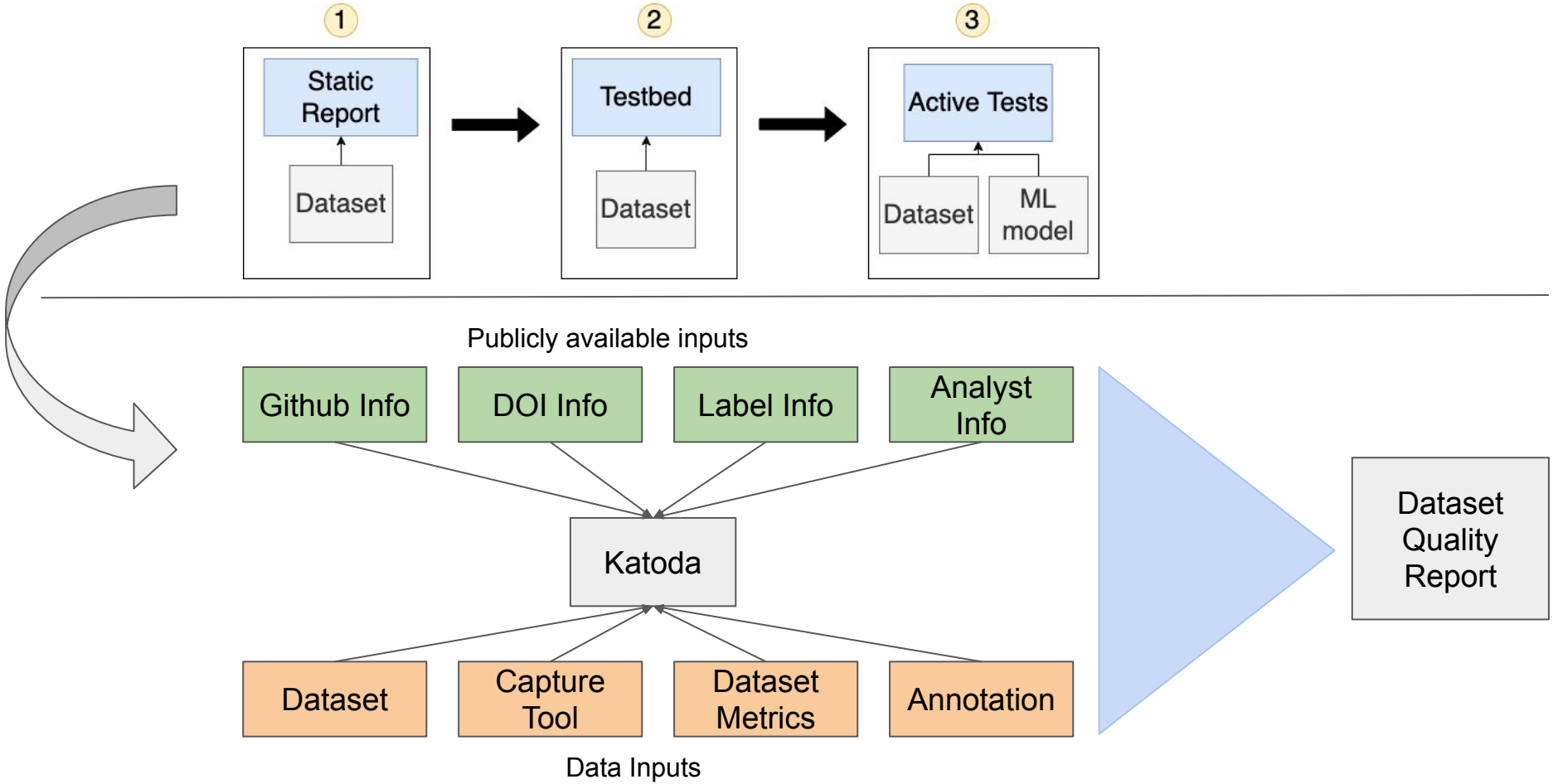
QoD Workflow

(Get Insights)

Tools For Dataset Operations

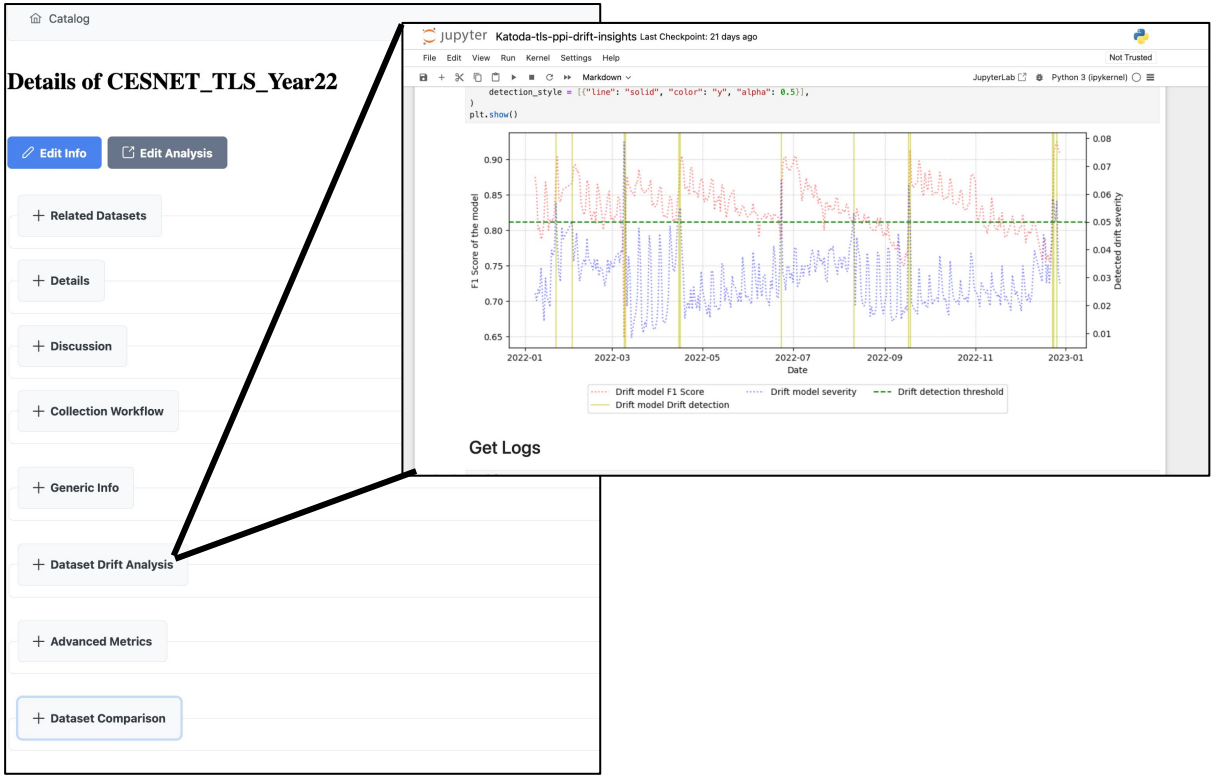
- Dataset Metrics [2], [3], [4]
  - Redundancy
  - Separability
  - Similarity
- Active Learning Framework [5], [6]
  - Dataset merge and updates
- Dataset Drifts [7], [8]
  - Behavior and distribution changes

QoD Workflow



Katoda = Catalogue of Datasets

- Capture inputs for QoD Report
- Calculate evaluation metrics and logs
- GUI for users



Summary

- QoD is wide and open research area
- We describe the area and provide initial tools for dataset quality
- Katoda is unifying knowledge of networking datasets
- Let us know if you want to test your dataset

Thank You!

Resources