Actionable Data Insights for Machine Learning

Nils Braun

EuroMLSys 2023 | Apple Inc. | 08.05.2023

Agenda

Motivation

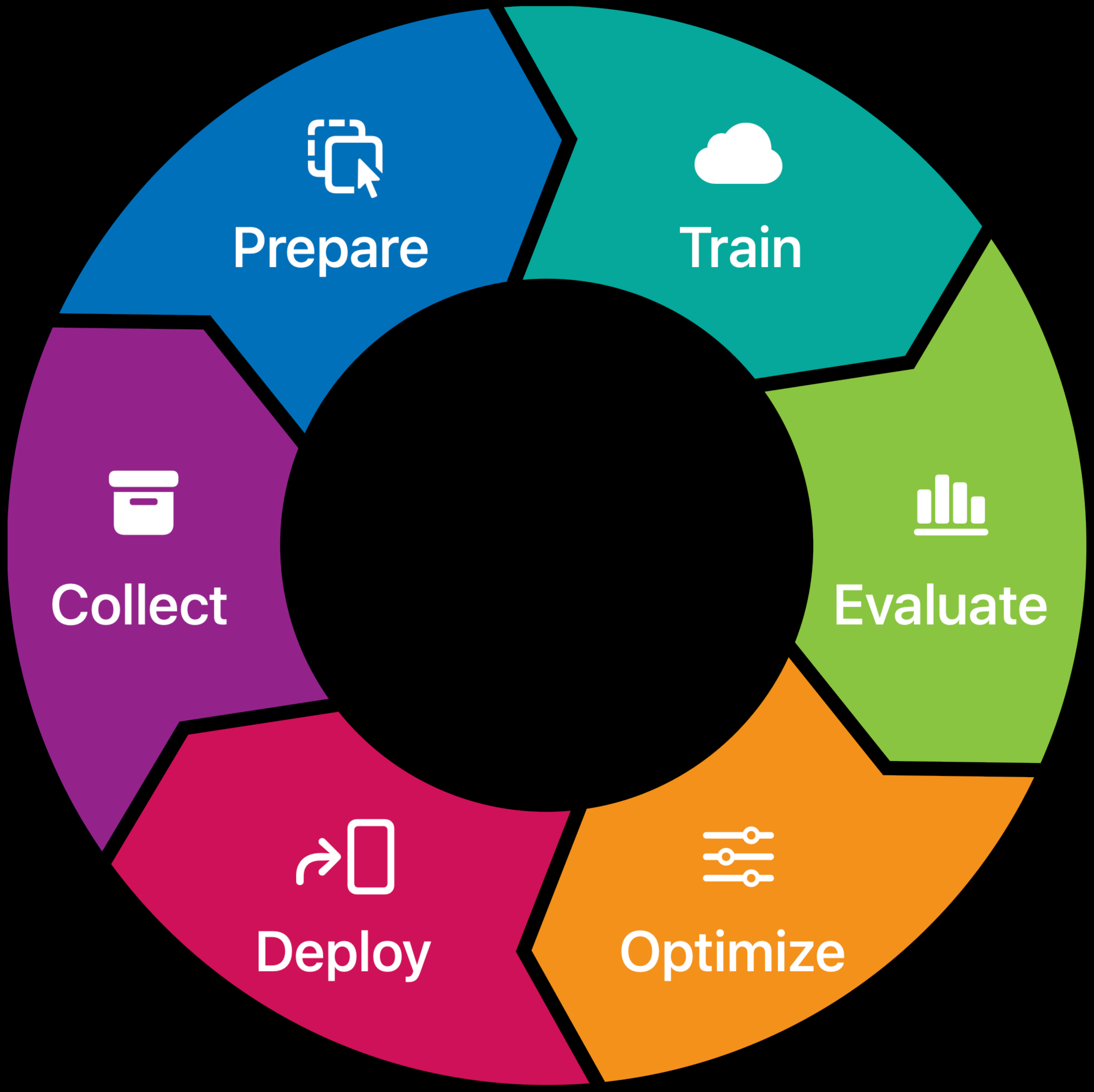
Implementation

Case Studies

Summary

Motivation

Drive model performance with better data understanding



Interview Learnings

Projects included: Image understanding, machine translation, text understanding, speech recognition

Enable rapid data experiments

- High-level declarative interfaces
- User defined functions as first class citizen

Flexibility

- Enable usage across different storage systems
- Enable portability from local machines to cloud to devices

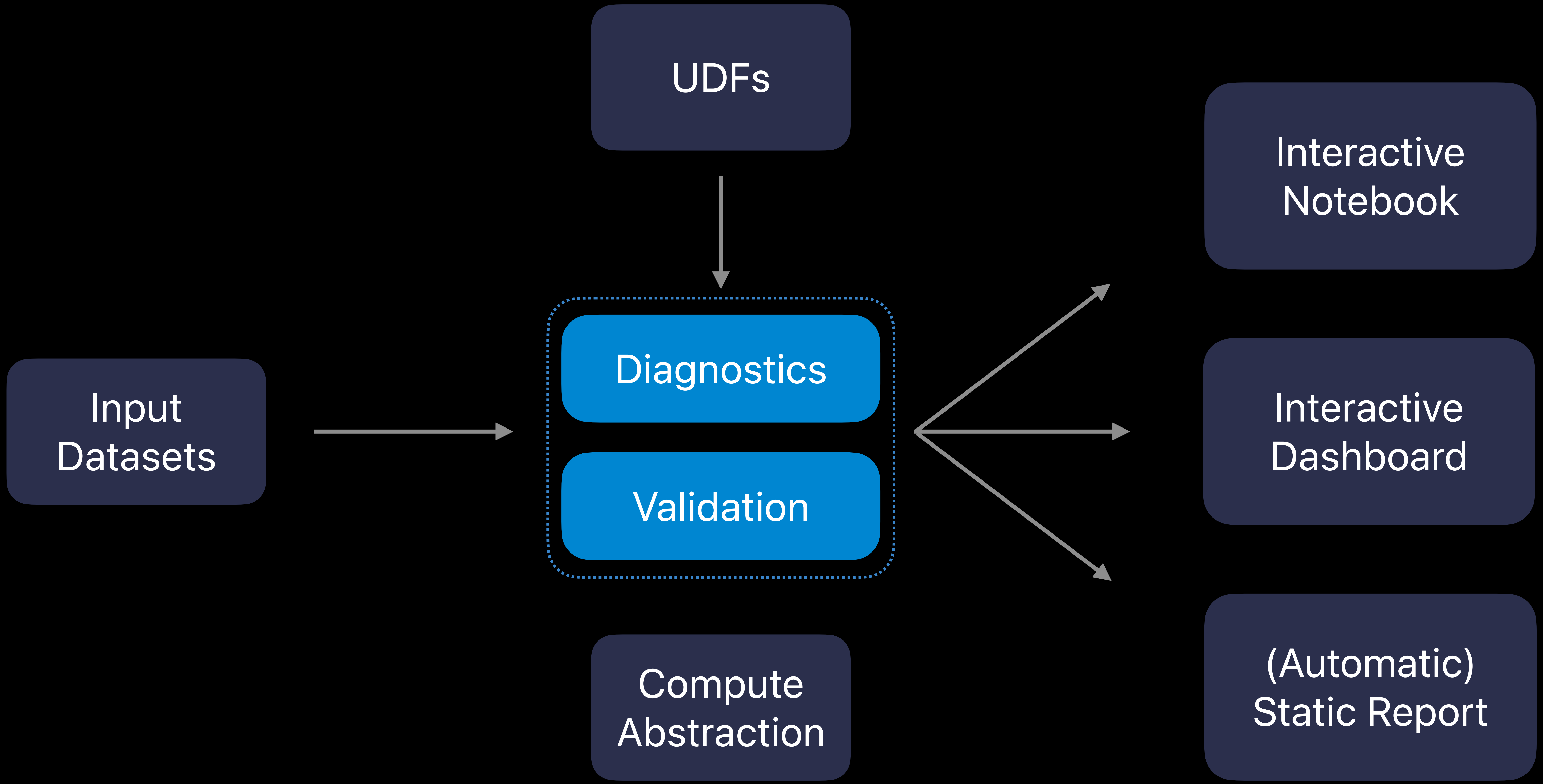
Collaboration & Reproducibility

- Documentation for data and models
- Share findings

Implementation

Data Tooling

High-level overview



Implementation

Required components for data insights

Data Diagnostics

- Interactive roll-up and drill-down
- Raw data preview

Data Validation

- Define constraints on sample or aggregate level

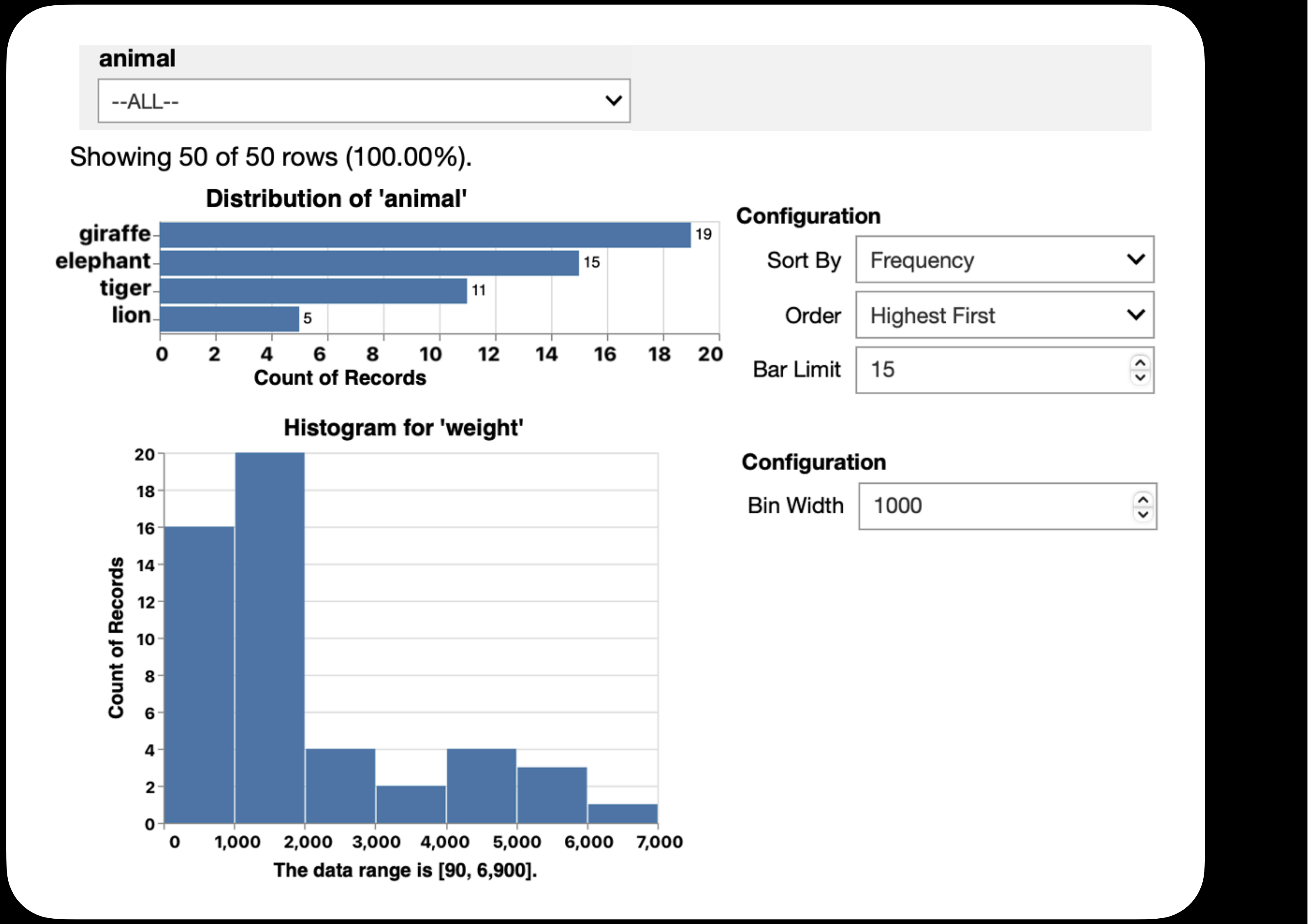
- Data Access abstraction enables loading data from many sources
- Compute abstraction enables portability

Data Diagnostics

```
def show_sample(row, **kwargs):
  # visualise image with bounding box

sample_view = SampleView(
  details [CustomDetail(show_sample)]
)

Explorer(
  filters ["animal"],
  views [
    "animal",
    "weight",
    ("animal", "background"),
    sample_view
  ]
).show(data)
```



animal

--ALL--

Showing 50 of 50 rows (100.00%).

Configuration

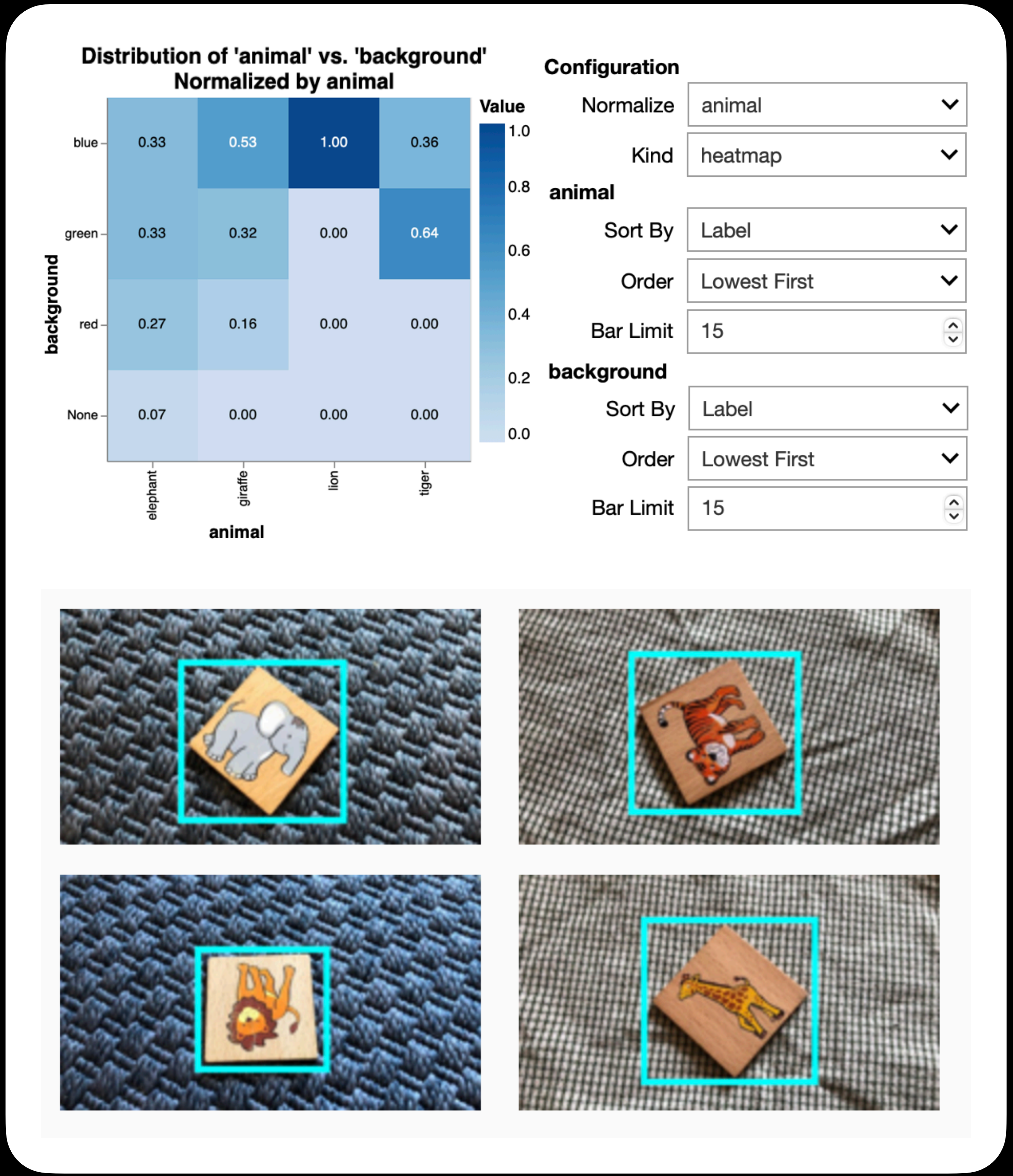
Sort By: Frequency

Order: Highest First

Bar Limit: 15

Configuration

Bin Width: 1000



Distribution of 'animal' vs. 'background'
Normalized by animal

background	elephant	giraffe	lion	tiger
blue	0.33	0.53	1.00	0.36
green	0.33	0.32	0.00	0.64
red	0.27	0.16	0.00	0.00
None	0.07	0.00	0.00	0.00

Configuration

Normalize: animal

Kind: heatmap

animal

Sort By: Label

Order: Lowest First

Bar Limit: 15

background

Sort By: Label

Order: Lowest First

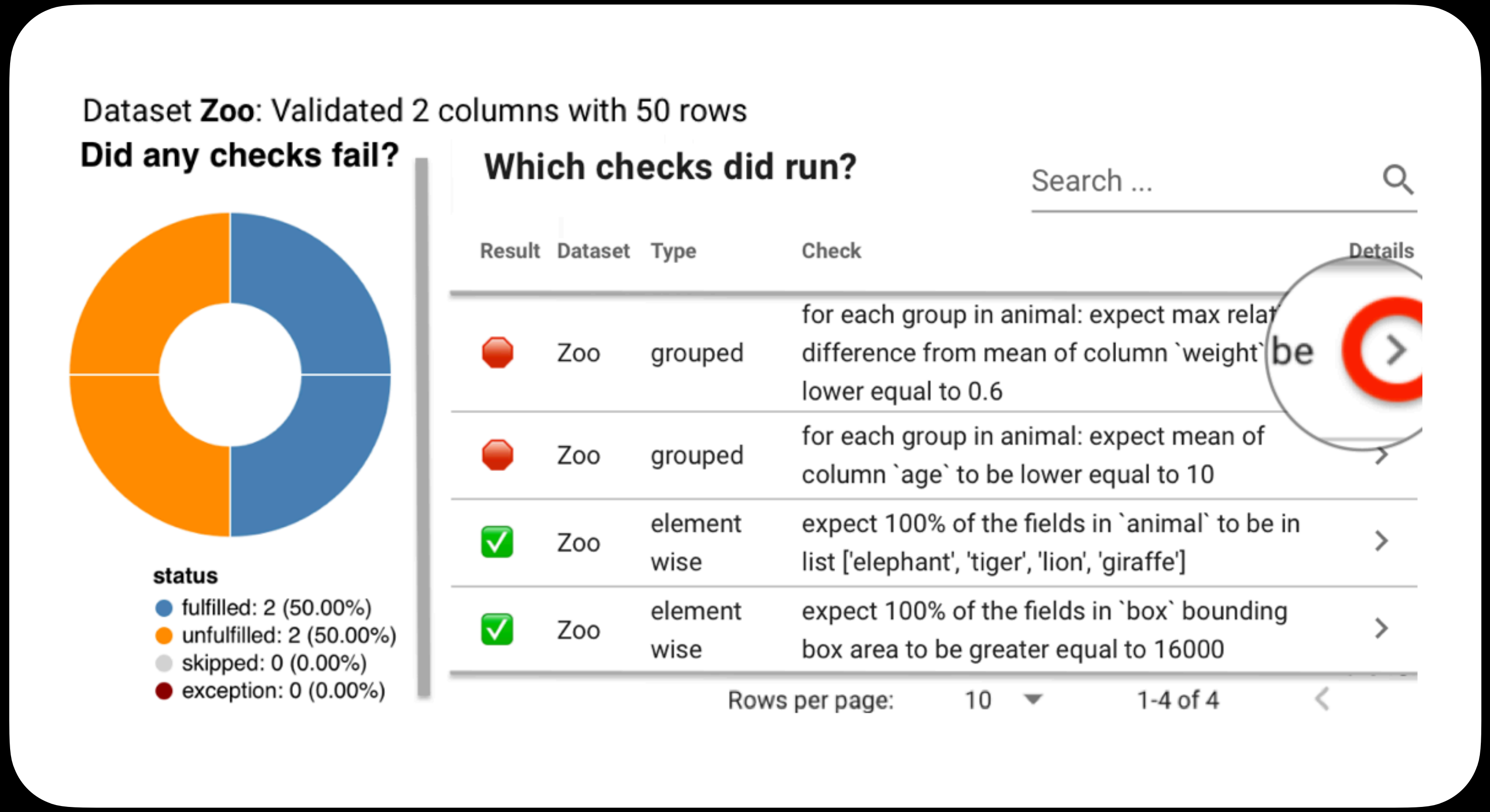
Bar Limit: 15

Data Validation

```
from udf import bbox_area, mean, max_diff_from_mean

checks = [
  Cell("animal") in ["elephant", "tiger", “lion", "giraffe"],
  Cell("box", bbox_area) >= 16000,
  Group("animal", Column("age", mean) < 10),
  Group(“animal",
    Column("weight", max_diff_from_mean) < 0.6
  )
]

Validator(expectations=checks).run(data)
```



Dataset **Zoo**: Validated 2 columns with 50 rows

Did any checks fail?

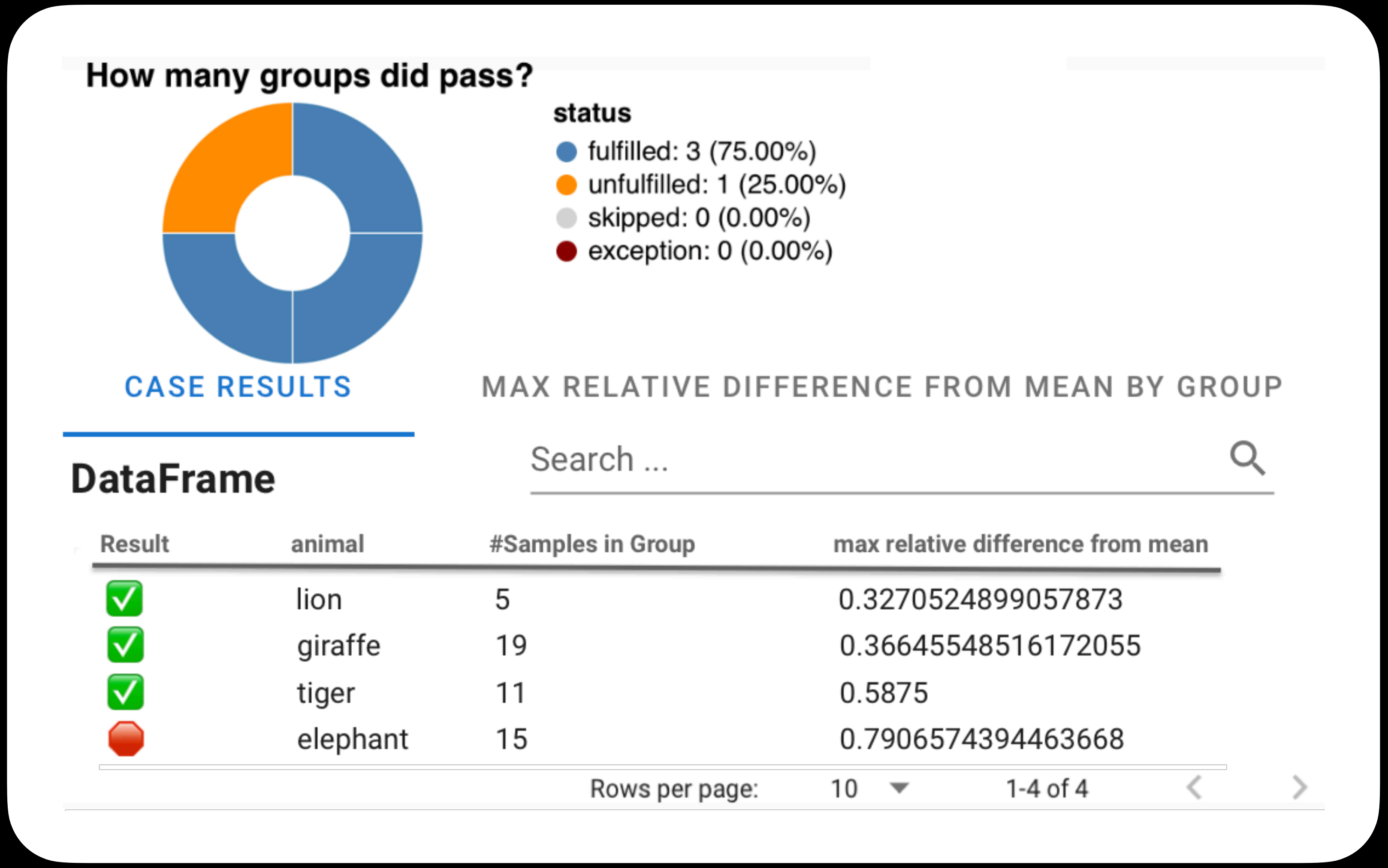
status
- ● fulfilled: 2 (50.00%)
- ● unfulfilled: 2 (50.00%)
- ● skipped: 0 (0.00%)
- ● exception: 0 (0.00%)

Which checks did run?

Search ...

Result	Dataset	Type	Check	Details
🔴	Zoo	grouped	for each group in animal: expect max relat difference from mean of column `weight` be lower equal to 0.6	>
🔴	Zoo	grouped	for each group in animal: expect mean of column `age` to be lower equal to 10	>
✅	Zoo	element wise	expect 100% of the fields in `animal` to be in list ['elephant', 'tiger', 'lion', 'giraffe']	>
✅	Zoo	element wise	expect 100% of the fields in `box` bounding box area to be greater equal to 16000	>

Rows per page: 10 ▾ 1-4 of 4 <



How many groups did pass?

status
- ● fulfilled: 3 (75.00%)
- ● unfulfilled: 1 (25.00%)
- ● skipped: 0 (0.00%)
- ● exception: 0 (0.00%)

DataFrame

Search ...

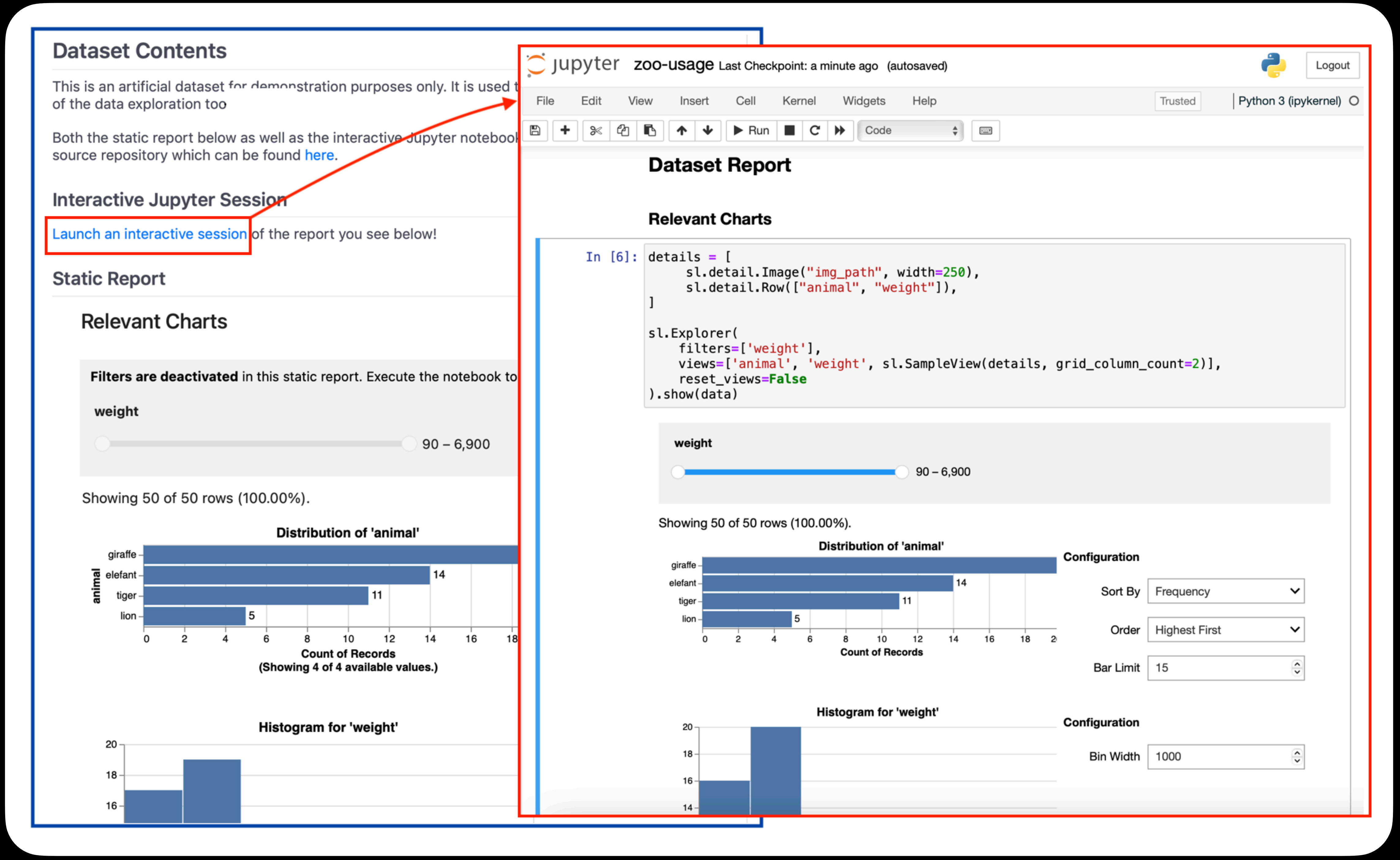
Result	animal	#Samples in Group	max relative difference from mean
✅	lion	5	0.3270524899057873
✅	giraffe	19	0.36645548516172055
✅	tiger	11	0.5875
🔴	elephant	15	0.7906574394463668

Rows per page: 10 ▾ 1-4 of 4 < >

Case Studies

Robust and Reproducible Dataset Documentation

- Capture context about datasets in documentation
- Enrich with summary statistics and raw data preview
- Transition seamlessly to interactive usage



Dataset Contents

This is an artificial dataset for demonstration purposes only. It is used t
of the data exploration too

Both the static report below as well as the interactive Jupyter notebook
source repository which can be found here.

Interactive Jupyter Session

Launch an interactive session of the report you see below!

Static Report

Relevant Charts

Filters are deactivated in this static report. Execute the notebook to

weight 90 – 6,900

Showing 50 of 50 rows (100.00%).

zoo-usage Last Checkpoint: a minute ago (autosaved)

Dataset Report

Relevant Charts

```
In [6]: details = [
             sl.detail.Image("img_path", width=250),
             sl.detail.Row(["animal", "weight"]),
        ]

        sl.Explorer(
            filters=['weight'],
            views=['animal', 'weight', sl.SampleView(details, grid_column_count=2)],
            reset_views=False
        ).show(data)
```

weight 90 – 6,900

Showing 50 of 50 rows (100.00%).

Configuration

Sort By: Frequency

Order: Highest First

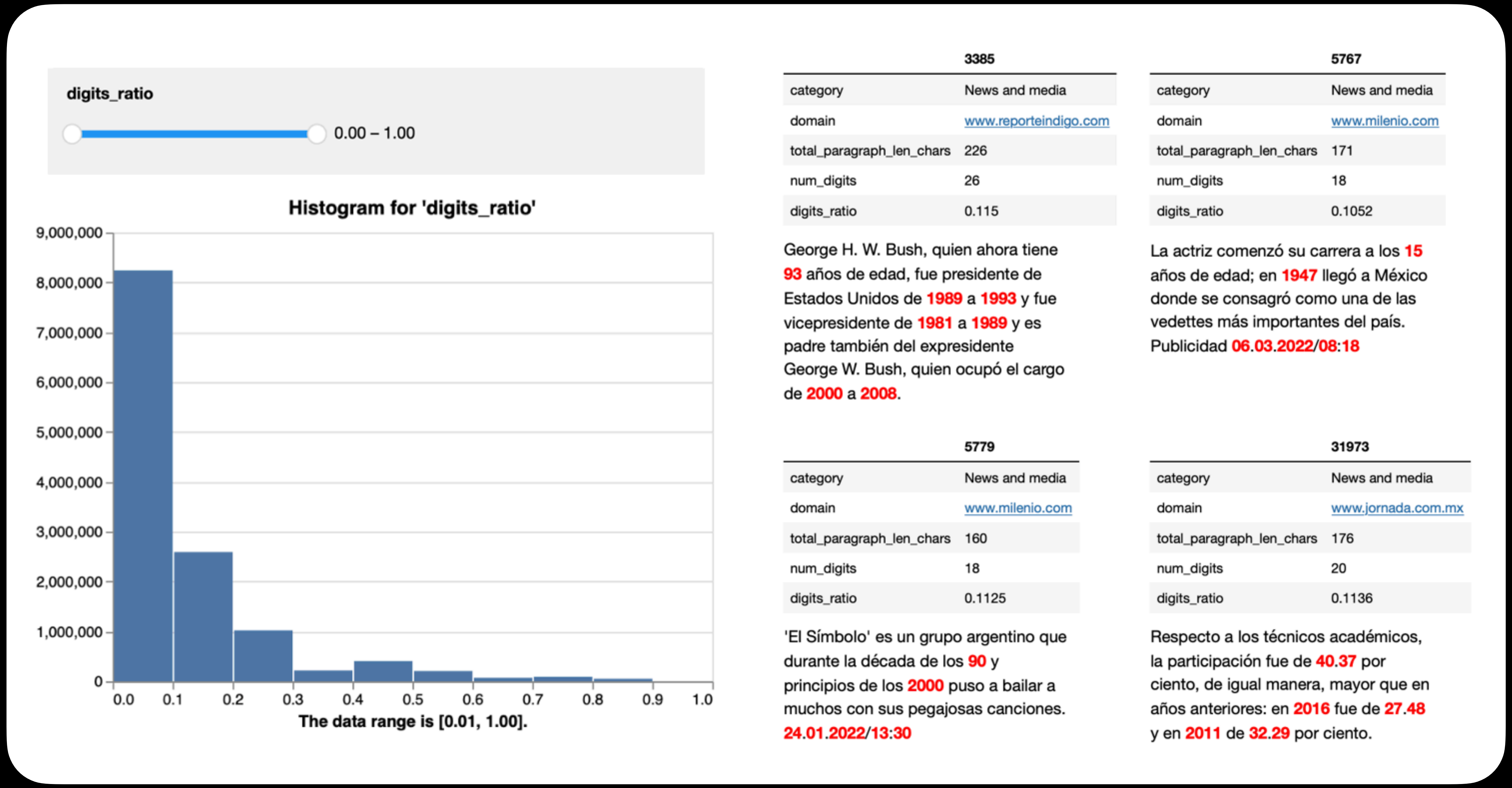
Bar Limit: 15

Configuration

Bin Width: 1000

Large Corpora Curation with Linguists in the Loop

- Empower domain experts to set filters while inspecting raw data
- Enable interactivity for 100M+ paragraphs
- Seamless collaboration between ML engineers and domain experts



	3385
category	News and media
domain	www.reporteindigo.com
total_paragraph_len_chars	226
num_digits	26
digits_ratio	0.115

George H. W. Bush, quien ahora tiene 93 años de edad, fue presidente de Estados Unidos de 1989 a 1993 y fue vicepresidente de 1981 a 1989 y es padre también del expresidente George W. Bush, quien ocupó el cargo de 2000 a 2008.

	5767
category	News and media
domain	www.milenio.com
total_paragraph_len_chars	171
num_digits	18
digits_ratio	0.1052

La actriz comenzó su carrera a los 15 años de edad; en 1947 llegó a México donde se consagró como una de las vedettes más importantes del país. Publicidad 06.03.2022/08:18

	5779
category	News and media
domain	www.milenio.com
total_paragraph_len_chars	160
num_digits	18
digits_ratio	0.1125

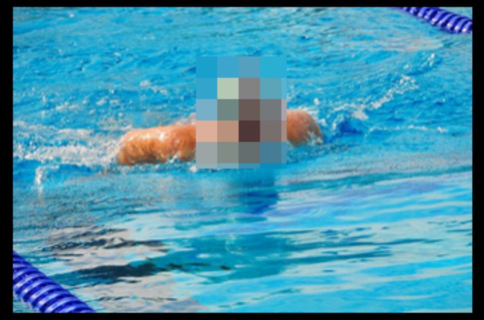
'El Símbolo' es un grupo argentino que durante la década de los 90 y principios de los 2000 puso a bailar a muchos con sus pegajosas canciones. 24.01.2022/13:30

	31973
category	News and media
domain	www.jornada.com.mx
total_paragraph_len_chars	176
num_digits	20
digits_ratio	0.1136

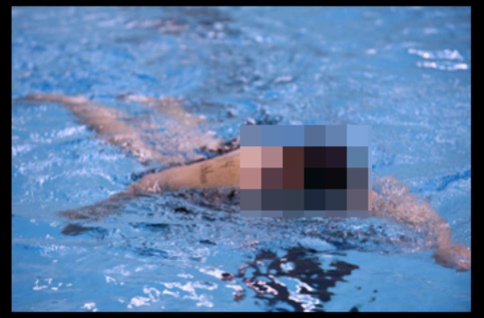
Respecto a los técnicos académicos, la participación fue de 40.37 por ciento, de igual manera, mayor que en años anteriores: en 2016 fue de 27.48 y en 2011 de 32.29 por ciento.

Actionable Failure Analysis of an Image Classifier

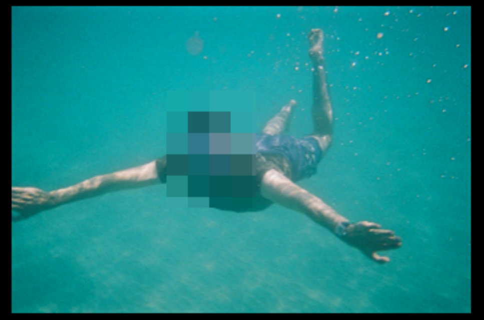
- When available metadata does not explain wrong predictions, visual inspection can reveal clues
- Image-text encoder models can be used to source new data in the same interactive session



Human ✓



Hippo ✗



Dolphin ✗

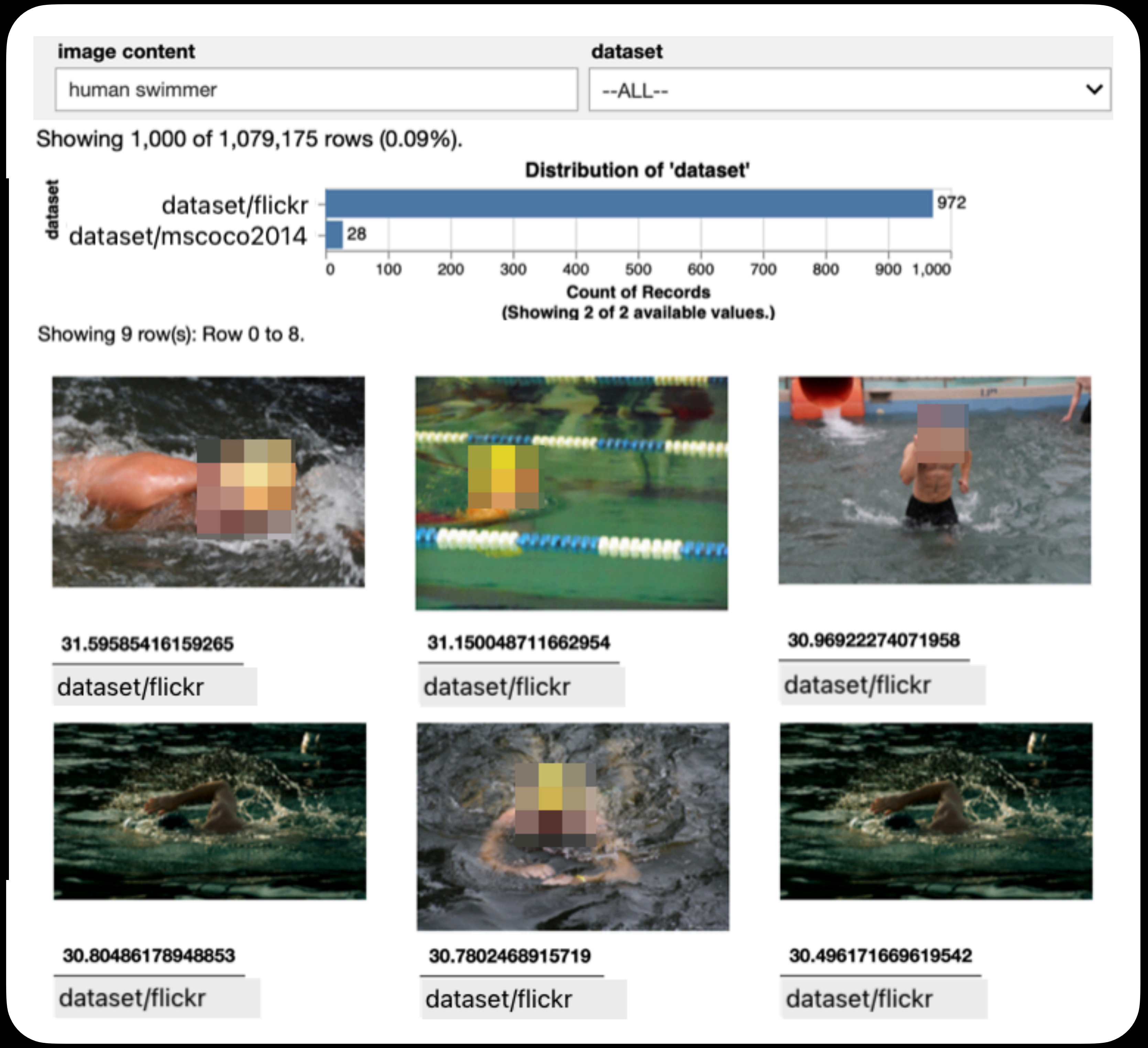


image content: human swimmer

dataset: --ALL--

Showing 1,000 of 1,079,175 rows (0.09%).

Distribution of 'dataset'

Showing 9 row(s): Row 0 to 8.

31.59585416159265
dataset/flickr

31.150048711662954
dataset/flickr

30.96922274071958
dataset/flickr

30.80486178948853
dataset/flickr

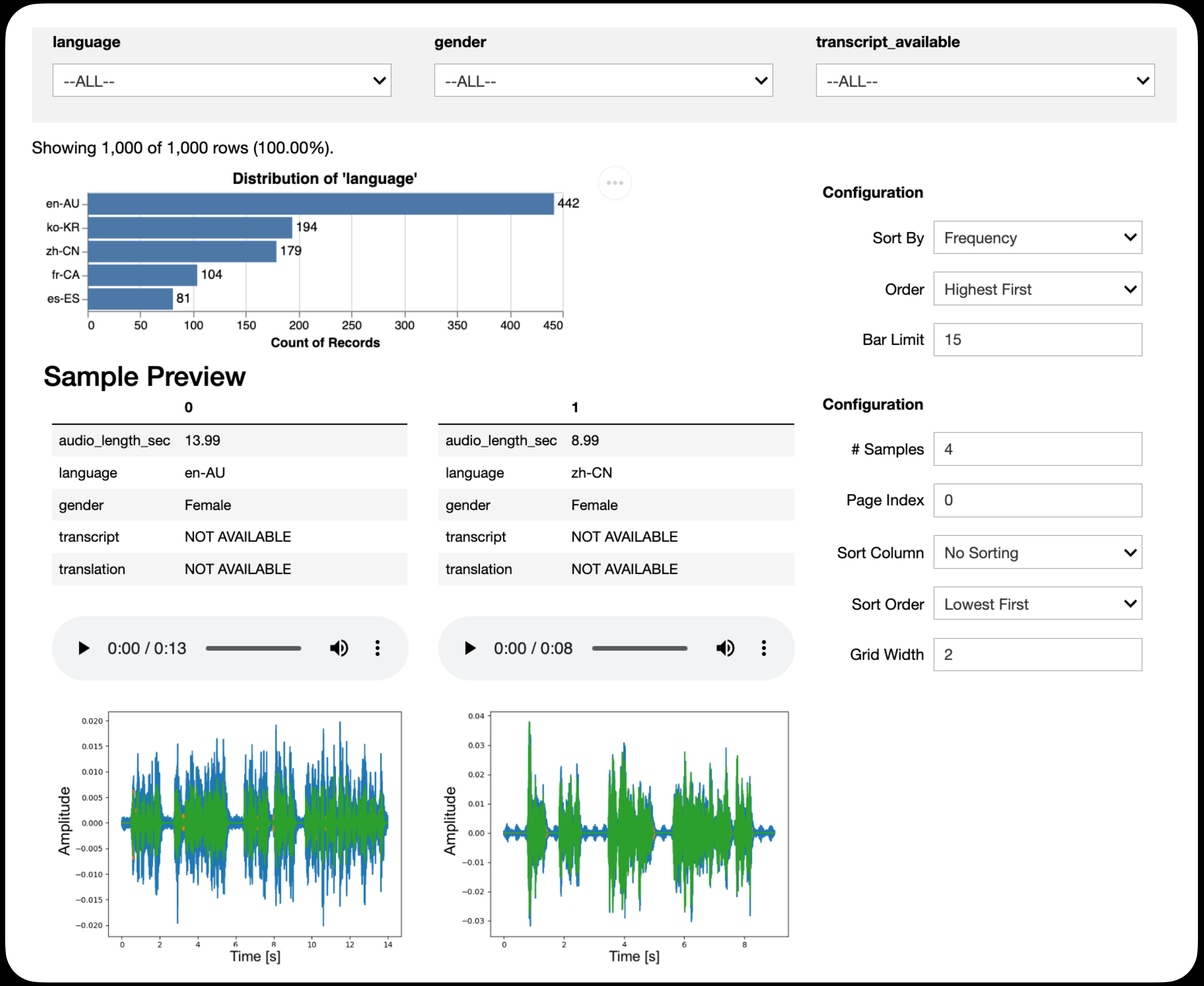
30.7802468915719
dataset/flickr

30.496171669619542
dataset/flickr

Rapid Iteration on Multi-Modal Datasets

from heterogeneous data sources

- Multi-modal machine translation project
- Necessity to blend different data sources and modalities in a unified preview
- Learnings from exploration could be persisted in validation checks



language: --ALL--

gender: --ALL--

transcript_available: --ALL--

Showing 1,000 of 1,000 rows (100.00%).

Configuration

Sort By: Frequency

Order: Highest First

Bar Limit: 15

Sample Preview

	0
audio_length_sec	13.99
language	en-AU
gender	Female
transcript	NOT AVAILABLE
translation	NOT AVAILABLE

	1
audio_length_sec	8.99
language	zh-CN
gender	Female
transcript	NOT AVAILABLE
translation	NOT AVAILABLE

Configuration

Samples: 4

Page Index: 0

Sort Column: No Sorting

Sort Order: Lowest First

Grid Width: 2

Conclusions

- We presented ADIML, a toolset to democratise data technology throughout the ML lifecycle and to enable the data-centric ML approach.
- The design of ADIML is based on the set of challenges and pain points we collected and validated from a wide range of ML teams.
- The case studies showing how easily ADIML can enable ML teams to focus and to improve data quality at scale are testimonies of its values.