NEURAL INFORMATION
PROCESSING SYSTEMS

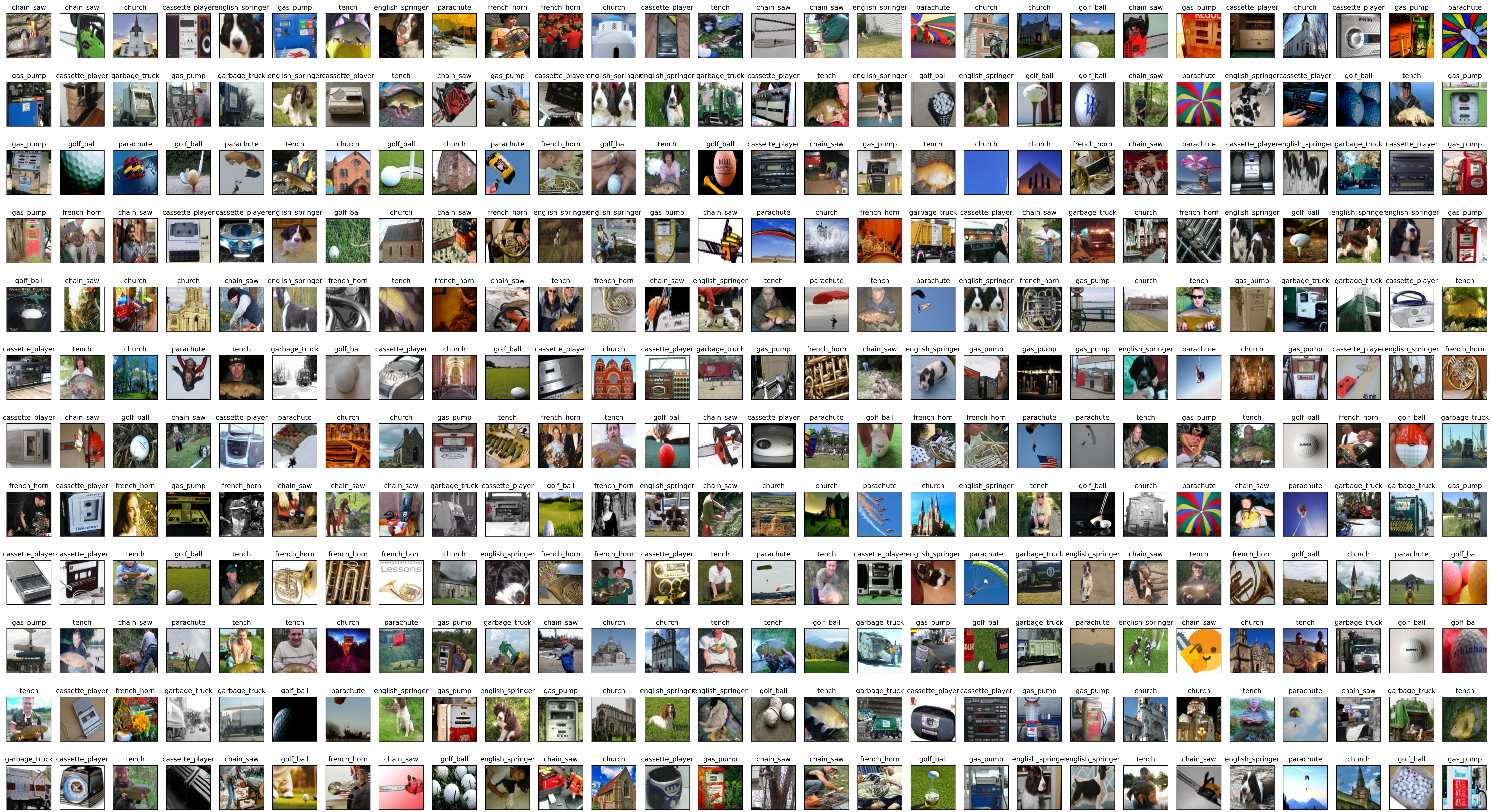
Intrinsic Self-Supervision for Data Quality Audits

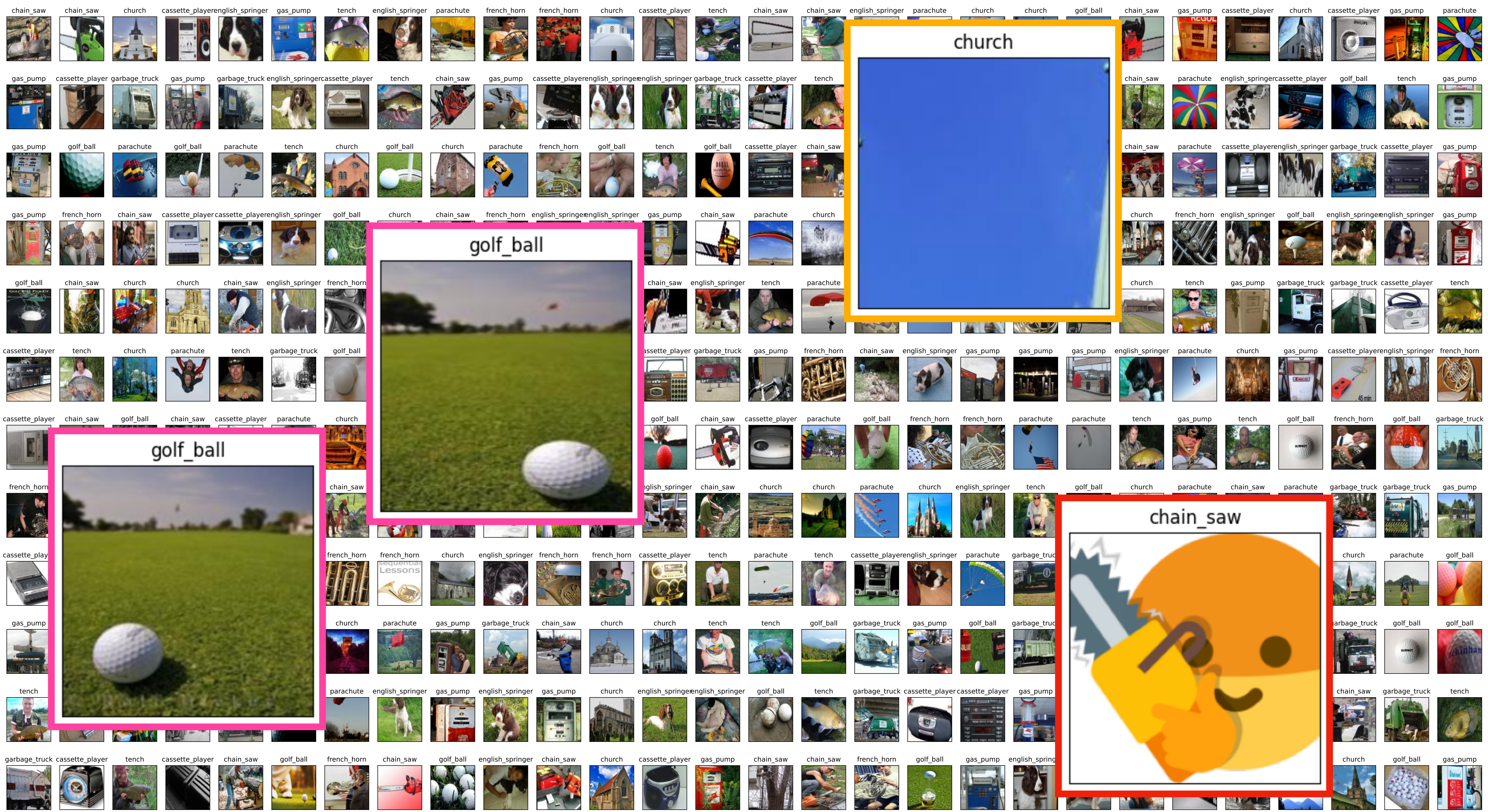
Fabian Gröger, Simone Lionetti, Philippe Gottfrois, Alvaro Gonzalez-Jimenez, Ludovic Amruthalingam, Matthew Groh, Alexander A. Navarini, Marc Pouly

University of Basel

HSLU Lucerne University of Applied Sciences and Arts

Northwestern University


chain_saw
chain_saw
church
cassette_player
english_springer
gas_pump
tench
english_springer
parachute
french_horn
french_horn
church
cassette_player
tench
chain_saw
chain_saw
english_springer
parachute
church
church
golf_ball
chain_saw
gas_pump
cassette_player
church
cassette_player
gas_pump
parachute
gas_pump
cassette_player
garbage_truck
gas_pump
garbage_truck
english_springer
cassette_player
tench
chain_saw
gas_pump
cassette_player
english_springer
english_springer
garbage_truck
cassette_player
tench
english_springer
golf_ball
english_springer
golf_ball
golf_ball
chain_saw
parachute
english_springer
cassette_player
golf_ball
tench
gas_pump
gas_pump
golf_ball
parachute
golf_ball
parachute
tench
church
golf_ball
church
parachute
french_horn
golf_ball
tench
golf_ball
cassette_player
chain_saw
gas_pump
tench
church
church
french_horn
chain_saw
parachute
cassette_player
english_springer
garbage_truck
cassette_player
gas_pump
gas_pump
french_horn
chain_saw
cassette_player
cassette_player
english_springer
golf_ball
church
chain_saw
french_horn
english_springer
english_springer
gas_pump
chain_saw
parachute
church
french_horn
garbage_truck
cassette_player
chain_saw
garbage_truck
church
french_horn
english_springer
golf_ball
english_springer
english_springer
gas_pump
golf_ball
chain_saw
church
church
chain_saw
english_springer
french_horn
tench
french_horn
chain_saw
tench
french_horn
chain_saw
english_springer
tench
parachute
tench
parachute
english_springer
french_horn
gas_pump
church
tench
gas_pump
garbage_truck
garbage_truck
cassette_player
tench
cassette_player
tench
church
parachute
tench
garbage_truck
golf_ball
cassette_player
church
golf_ball
cassette_player
church
cassette_player
garbage_truck
gas_pump
french_horn
chain_saw
english_springer
gas_pump
gas_pump
gas_pump
english_springer
parachute
church
gas_pump
cassette_player
english_springer
french_horn
cassette_player
chain_saw
golf_ball
chain_saw
cassette_player
parachute
church
church
gas_pump
tench
french_horn
tench
golf_ball
chain_saw
cassette_player
parachute
golf_ball
french_horn
french_horn
parachute
parachute
tench
gas_pump
tench
golf_ball
french_horn
golf_ball
garbage_truck
french_horn
cassette_player
french_horn
gas_pump
french_horn
chain_saw
chain_saw
chain_saw
garbage_truck
cassette_player
golf_ball
french_horn
english_springer
chain_saw
church
church
parachute
church
english_springer
tench
golf_ball
church
parachute
chain_saw
parachute
garbage_truck
garbage_truck
gas_pump
cassette_player
cassette_player
tench
golf_ball
tench
french_horn
french_horn
french_horn
church
english_springer
french_horn
french_horn
cassette_player
tench
parachute
tench
cassette_player
english_springer
parachute
garbage_truck
english_springer
chain_saw
tench
french_horn
golf_ball
church
parachute
golf_ball
gas_pump
tench
chain_saw
parachute
tench
tench
church
parachute
gas_pump
garbage_truck
chain_saw
church
church
tench
tench
golf_ball
garbage_truck
gas_pump
golf_ball
garbage_truck
parachute
english_springer
chain_saw
church
tench
garbage_truck
golf_ball
golf_ball
tench
cassette_player
french_horn
garbage_truck
garbage_truck
golf_ball
parachute
english_springer
gas_pump
english_springer
gas_pump
church
english_springer
english_springer
golf_ball
tench
garbage_truck
cassette_player
cassette_player
gas_pump
gas_pump
church
church
tench
parachute
chain_saw
garbage_truck
tench
garbage_truck
cassette_player
tench
cassette_player
chain_saw
golf_ball
french_horn
chain_saw
golf_ball
english_springer
chain_saw
church
cassette_player
gas_pump
chain_saw
chain_saw
french_horn
golf_ball
gas_pump
english_springer
english_springer
tench
chain_saw
english_springer
parachute
church
golf_ball
gas_pump



Off-topic samples

Near duplicates

Label errors

Everyone who is doing ML knows …

… training AND evaluation data can be messy.

… noise during evaluating leads to inconsistent performance estimates.

* Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Forbes, 2016

Everyone who is doing ML knows …

… training AND evaluation data can be messy.

… noise during evaluating leads to inconsistent performance estimates.

BUT, data cleaning …

… can be very time-consuming and labor intensive.

… is the least enjoyable task for many practitioners*.

* Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Forbes, 2016

Goals of this project 🏆

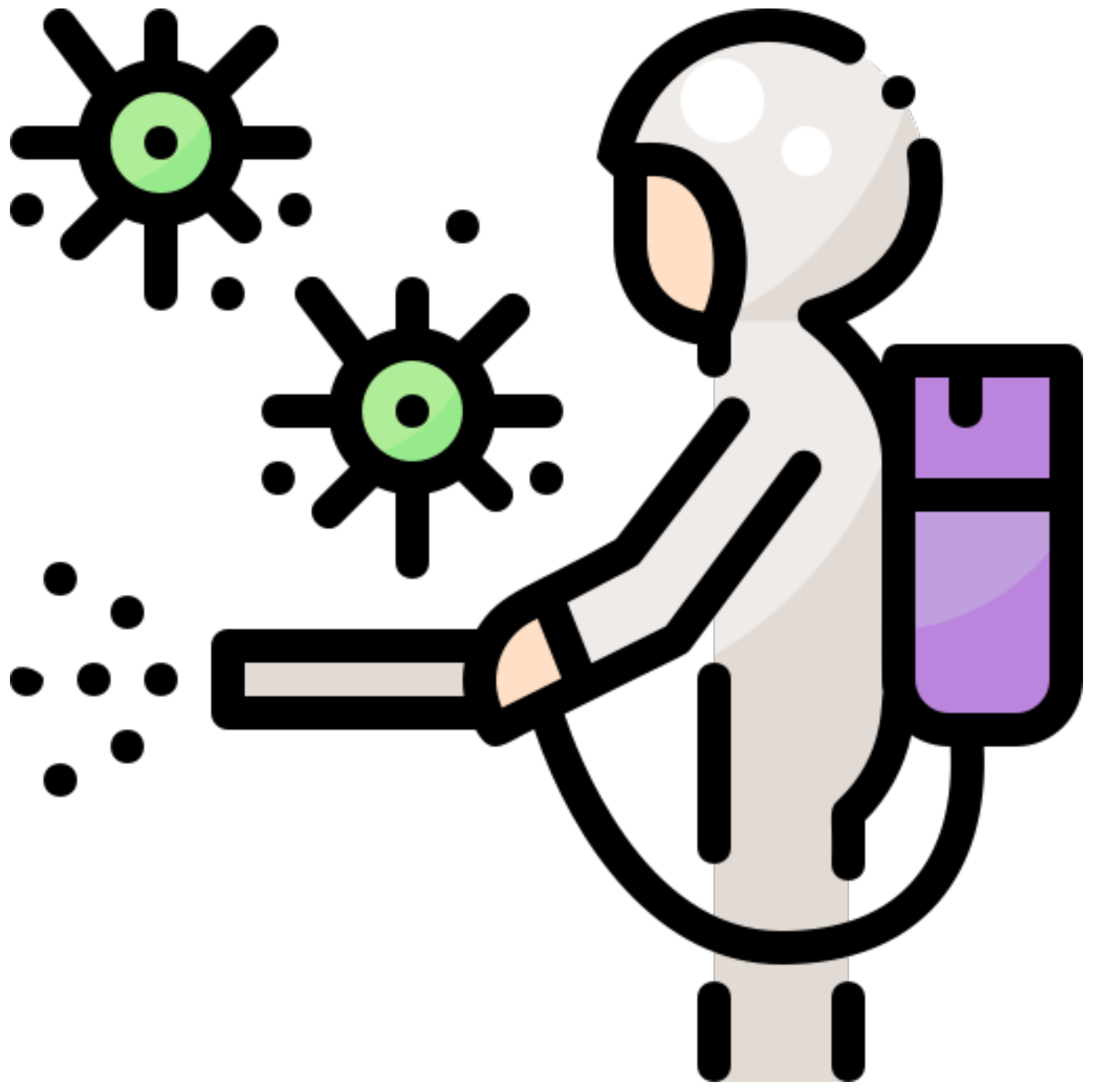
1. Reliably **detect data quality issues**, such as off-topic images, near duplicates, and label errors, in image datasets without introducing significant biases.

2. **Reduce the time needed** for detecting and confirming data quality issues.

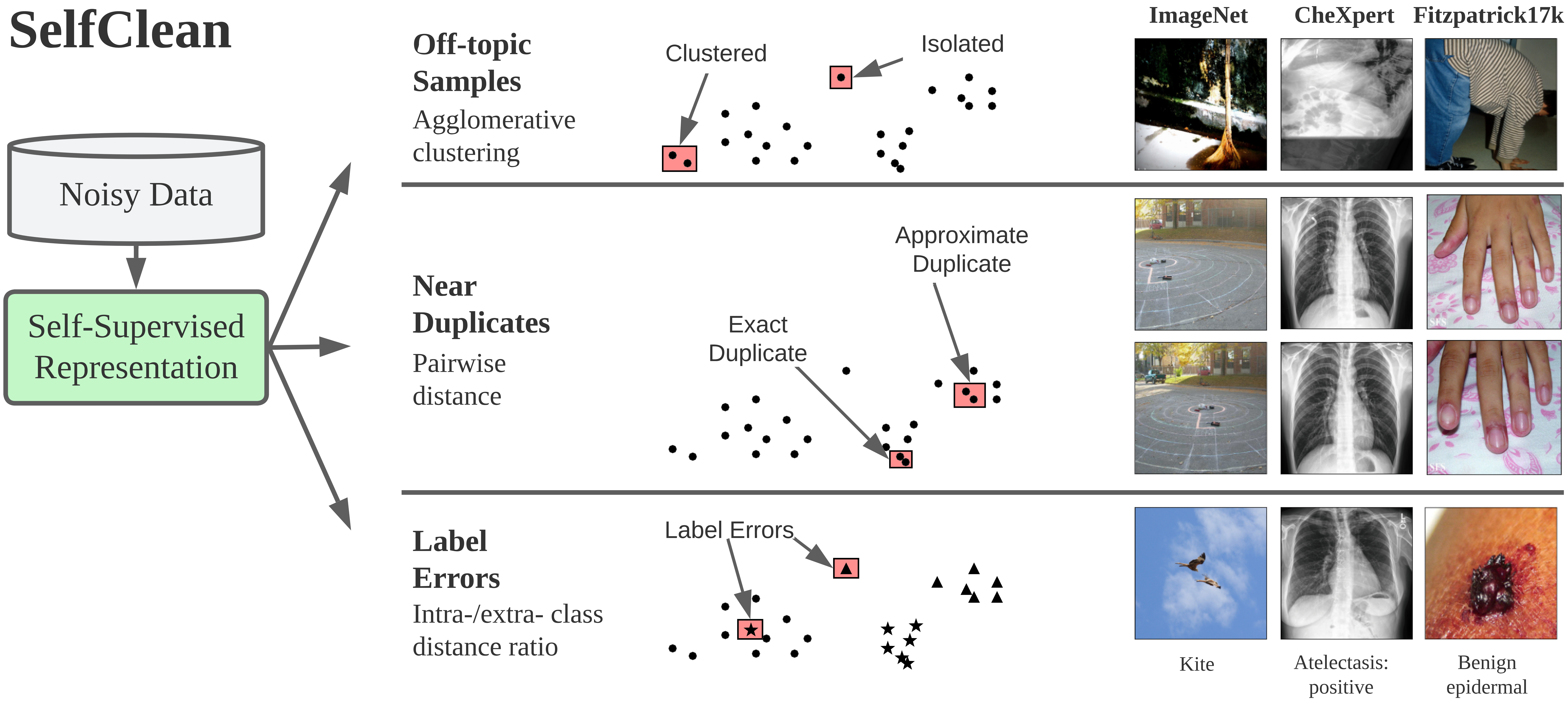
3. Investigate the **influence of data quality issues** on model training and evaluation.

Our findings

- Self-supervised learned (SSL) **representations can be exploited to find data quality issues**.
- **Context-aware SSL** representations can capture the dataset context with minimal bias.
- Combination of SSL **representations and distance-based indicators** effectively finds quality issues.



SelfClean



Noisy Data

Self-Supervised Representation

Off-topic Samples

Agglomerative clustering

Clustered

Isolated

Near Duplicates

Pairwise distance

Exact Duplicate

Approximate Duplicate

Label Errors

Intra-/extra- class distance ratio

Label Errors

ImageNet **CheXpert** **Fitzpatrick17k**

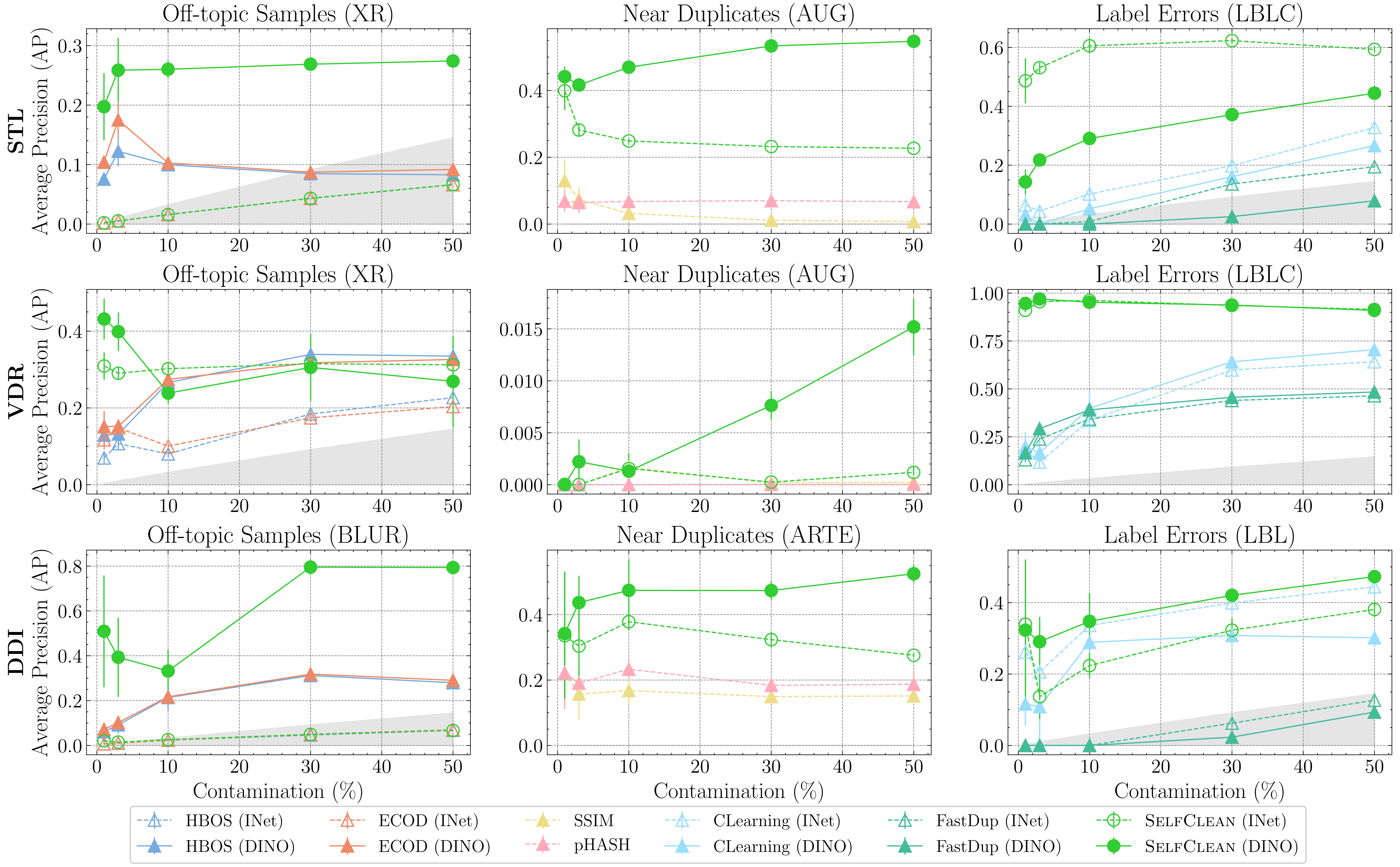
Kite

Atelectasis: positive

Benign epidermal

Results I

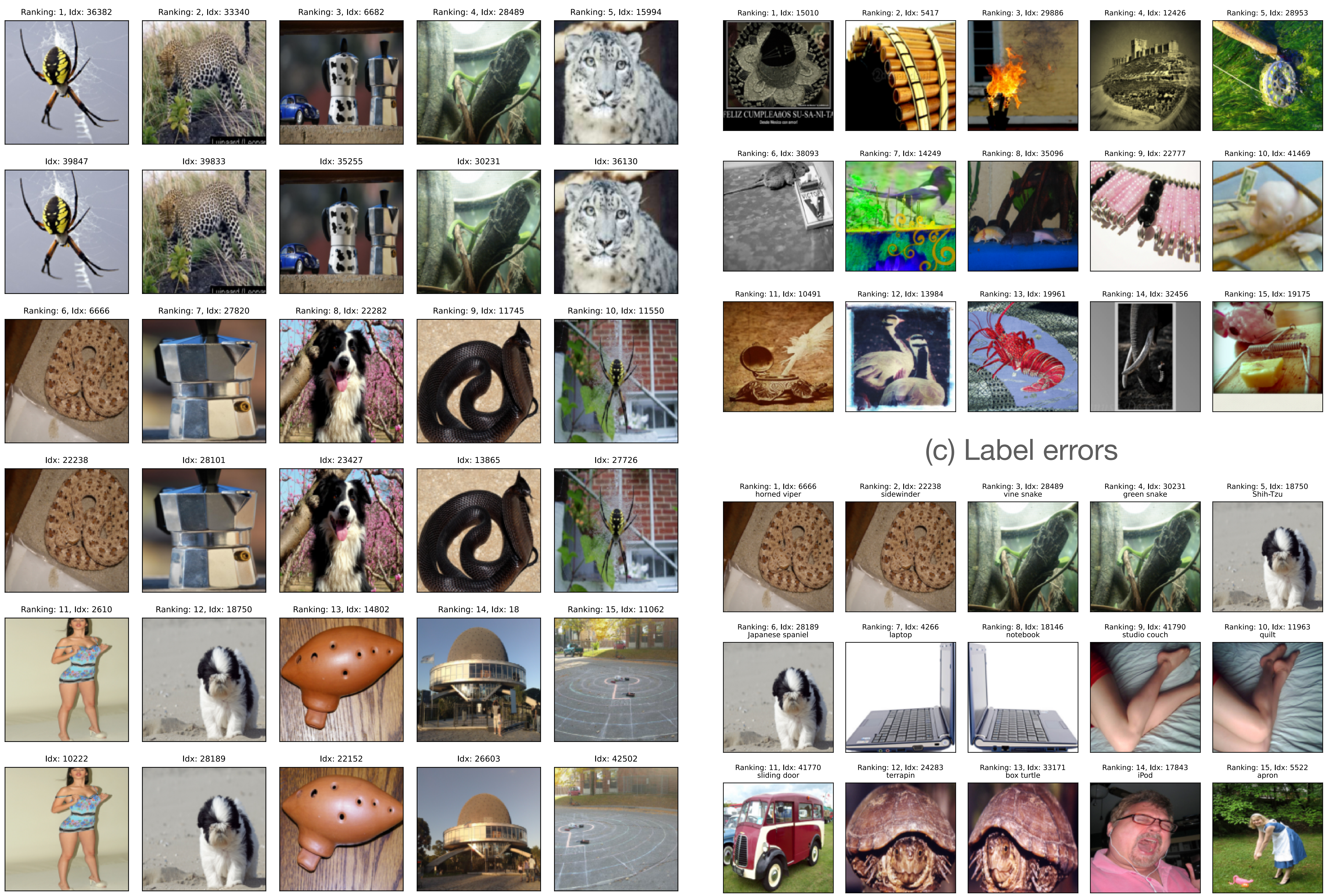
- Evaluation on both synthetic and natural contamination showed a **significant improvement** compared to current solutions.



Synthetic evaluation results.

Results II

- Applied to multiple image benchmarks, we identify up to **16% of issues**, and confirm an **improvement in evaluation reliability** upon cleaning.



(a) Near duplicates

(b) Off-topic samples

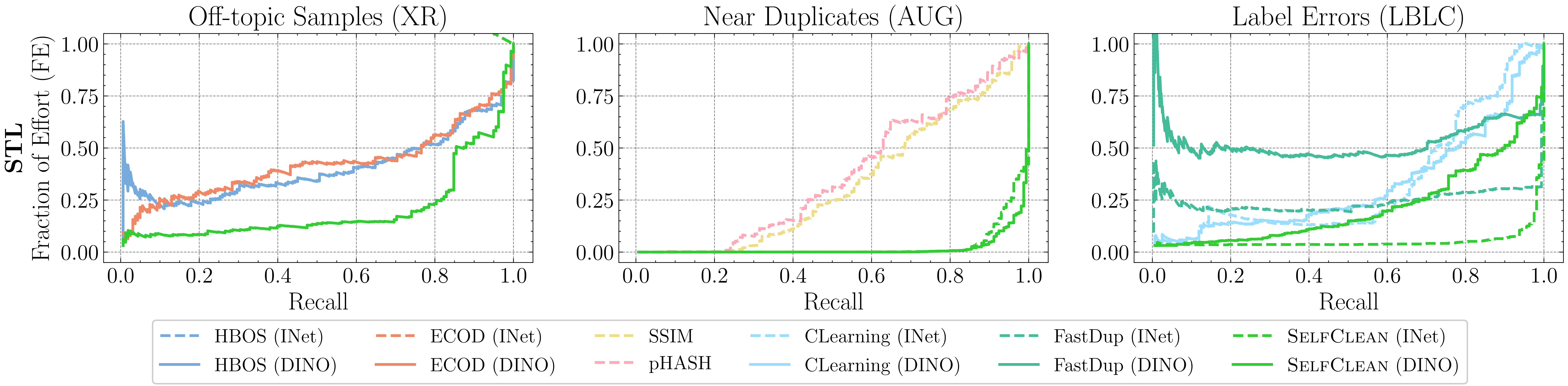
(c) Label errors

Ranking: 1, Idx: 6666 horned viper
Ranking: 2, Idx: 22238 sidewinder
Ranking: 3, Idx: 28489 vine snake
Ranking: 4, Idx: 30231 green snake
Ranking: 5, Idx: 18750 Shih-Tzu
Ranking: 6, Idx: 28189 Japanese spaniel
Ranking: 7, Idx: 4266 laptop
Ranking: 8, Idx: 18146 notebook
Ranking: 9, Idx: 41790 studio couch
Ranking: 10, Idx: 11963 quilt
Ranking: 11, Idx: 41770 sliding door
Ranking: 12, Idx: 24283 terrapin
Ranking: 13, Idx: 33171 box turtle
Ranking: 14, Idx: 17843 iPod
Ranking: 15, Idx: 5522 apron

Analysis of ImageNet-1k.

Results III

- For a typical dataset SelfClean can **reduce the inspection effort by a factor between 5 and 50**.



Analysis of the inspection effort saved.

NEURAL INFORMATION
PROCESSING SYSTEMS

Intrinsic Self-Supervision for Data Quality Audits

Project website: selfclean.github.io/

University of Basel

HSLU Lucerne University of Applied Sciences and Arts

Northwestern University