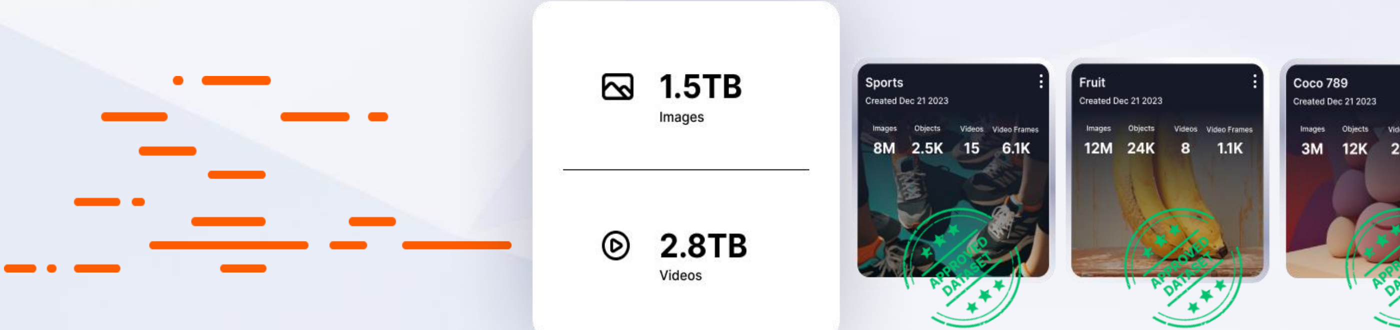
From Raw Data to Refined Datasets

Amir Alush, PhD

CTO @Visual Layer



Visual Layer

Pixels to Products: My Tech Journey

- Ph.D. on Discrete Optimization Problems (in computer vision)
- Deep Learning since 2012
- Previously CTO & Co founder of **Brodmann17**
- Currently CTO & Co founder of **Visual Layer (fastdup)**

Visual Layer

Family Album: 500GB of Chaos (and growing..)

- Married + 2 children + dog
- 500 hours of manually sifting

Generated by DALL·E 3

Deja Vu at Work: 100TB of Chaos

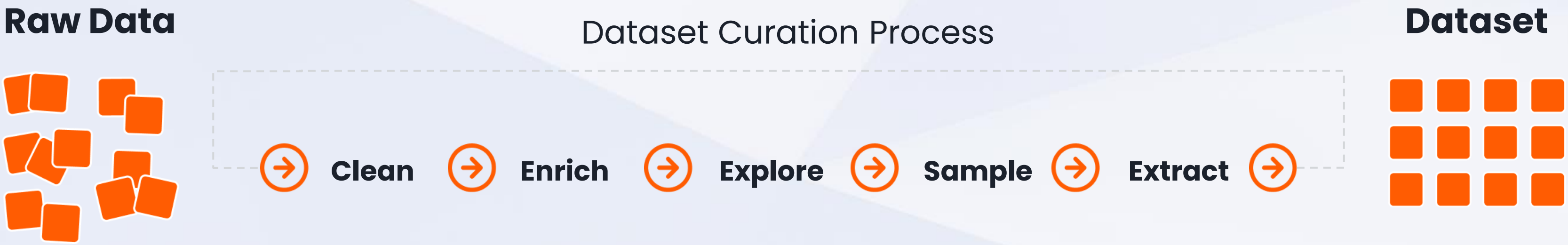
- Automotive, tons of data collected on a daily basis
- All this data has to go through a curation process

Raw Data

Dataset

Deja Vu at Work: 100TB of Chaos

- Automotive, tons of data collected on a daily basis
- All this data has to go through a curation process

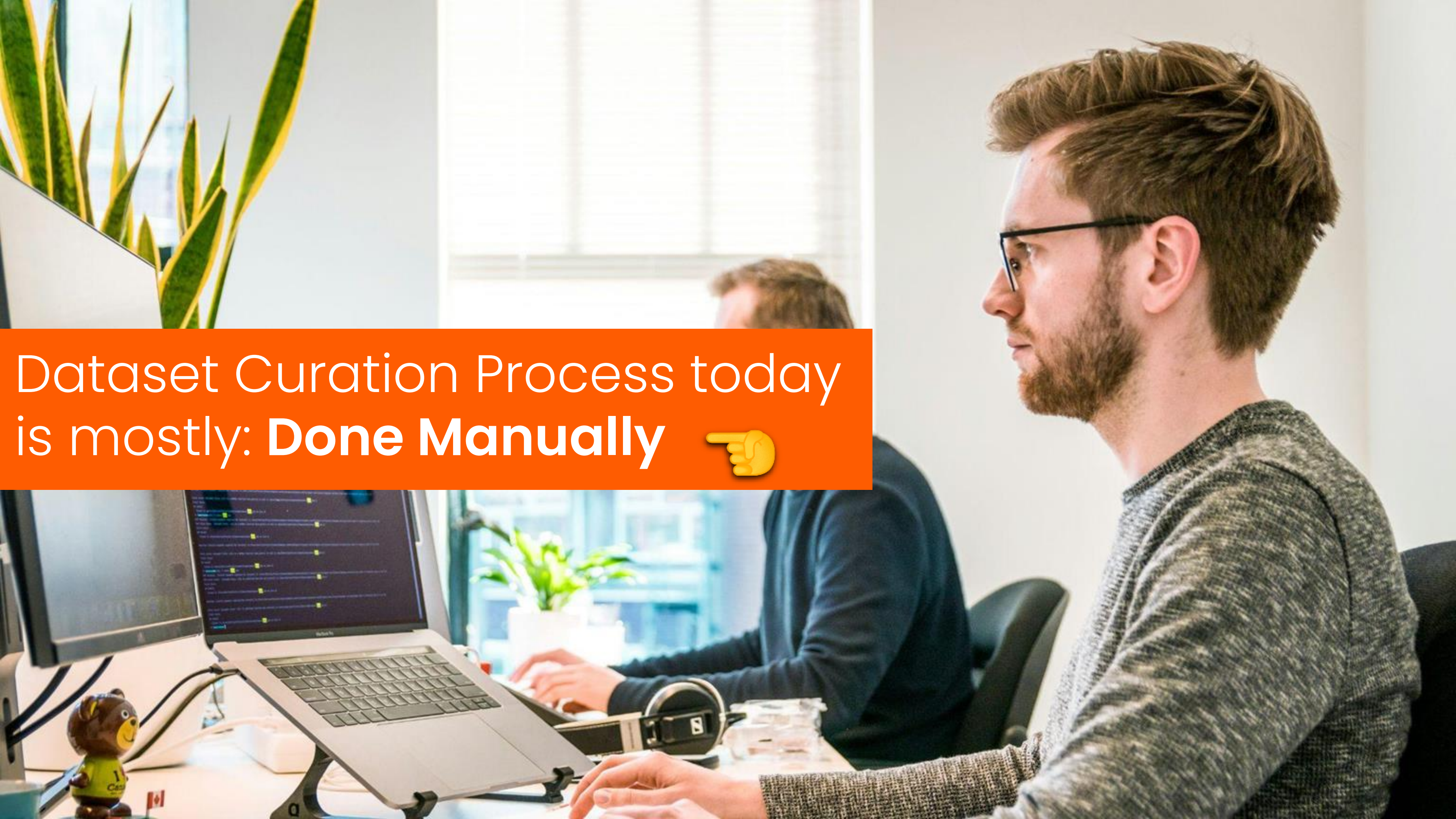


Raw Data

Dataset Curation Process

→ **Clean** → **Enrich** → **Explore** → **Sample** → **Extract** →

Dataset



Dataset Curation Process today is mostly: **Done Manually** 👉

The challenges of: Dataset Curation

1

Speed

Slow and Unscalable

2

Cost

Very expensive

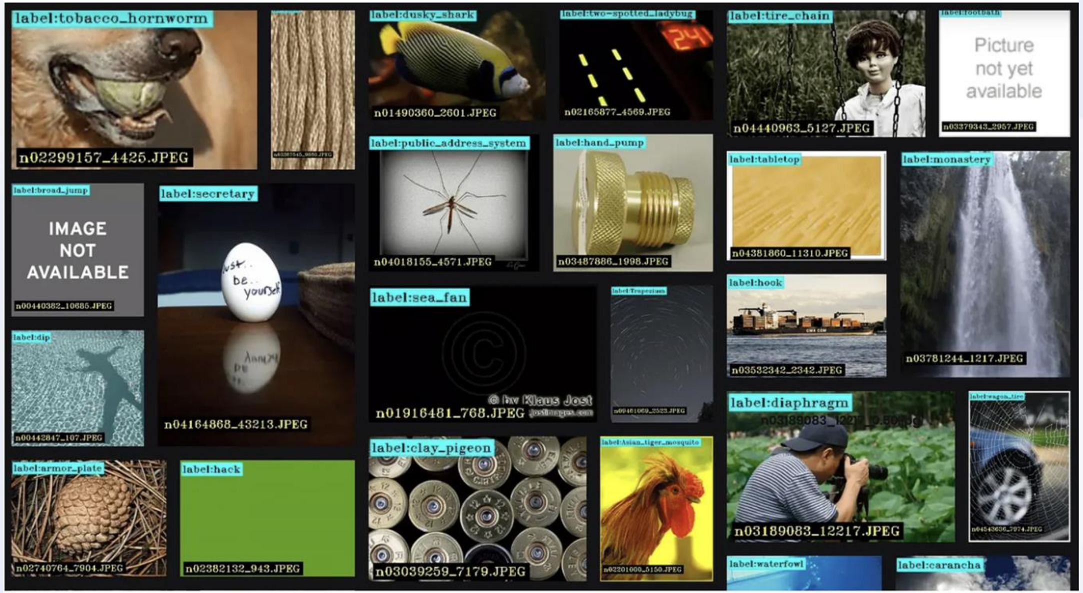
3

Quality

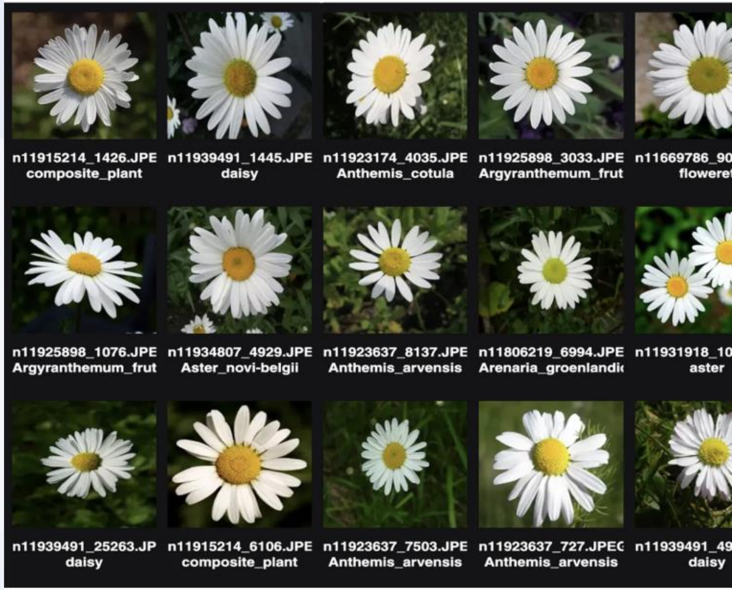
Very poor quality

Visual Layer

When Datasets Go Wrong: Bad Labels



Example wrong labels in the ImageNet-21K dataset. Additional information can be found in our GitHub repo.

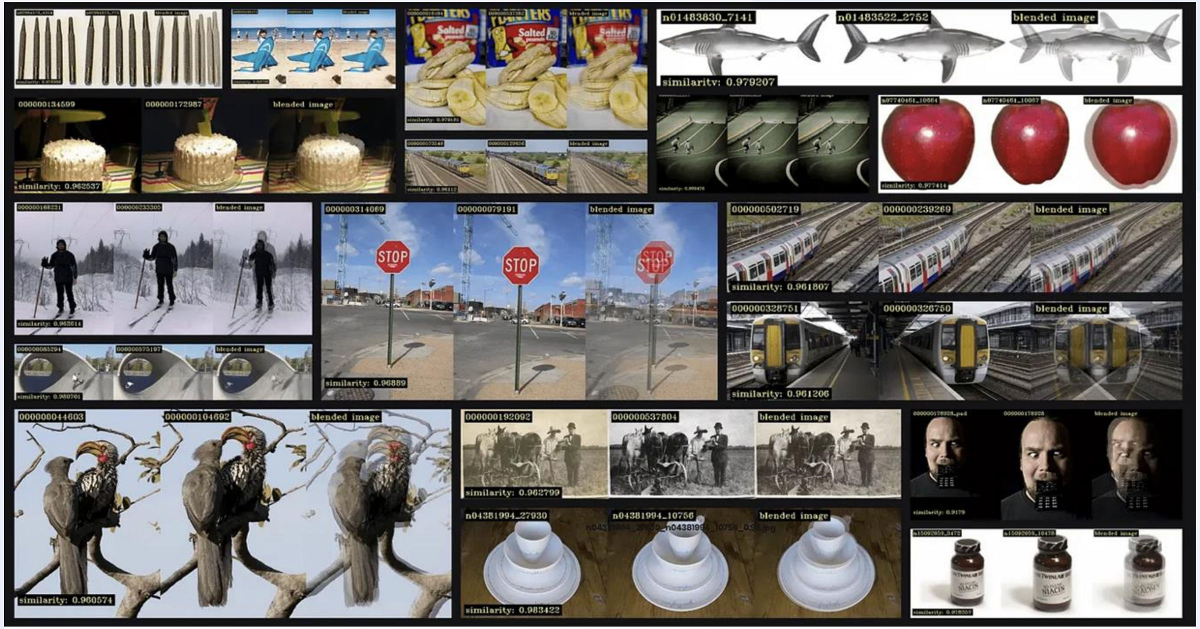


Example of many different labels for the same object. There are thousands of such clusters.

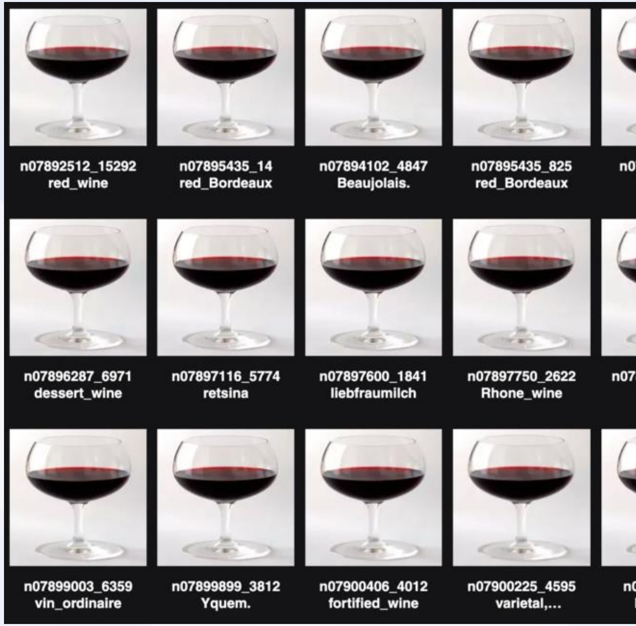
large-image-datasets-today-are-a-mess

Visual Layer

When Datasets Go Wrong: Duplicates



Example near duplicates identified in the MS-COCO (160K images) & ImageNet-21K datasets (11.5M images). A record breaking number of 1.2M duplicates were identified in the ImageNet-21K dataset! Additional



large-image-datasets-today-are-a-mess

Visual Layer

fastdup: Efficient Blocks for Dataset Curation

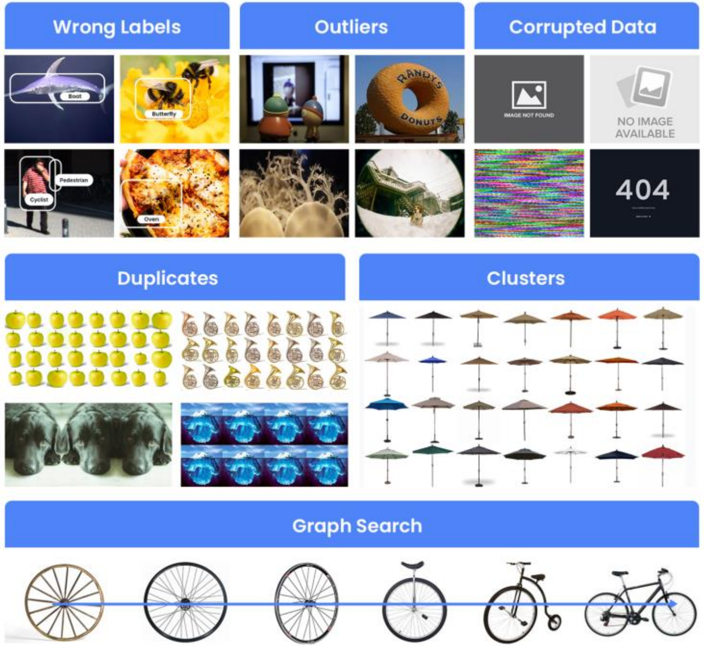
fastdup

1.4K+
GitHub stars

380K+
Downloads

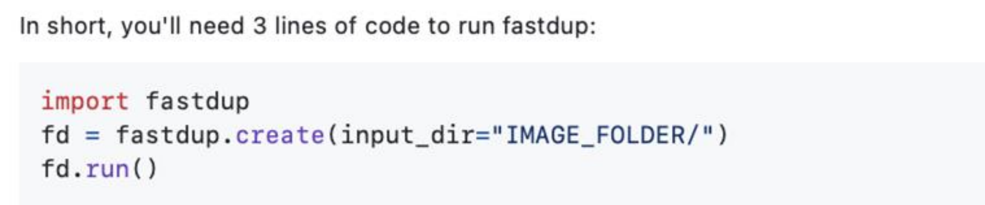
50B+
Images processed

visual-layer/fastdup



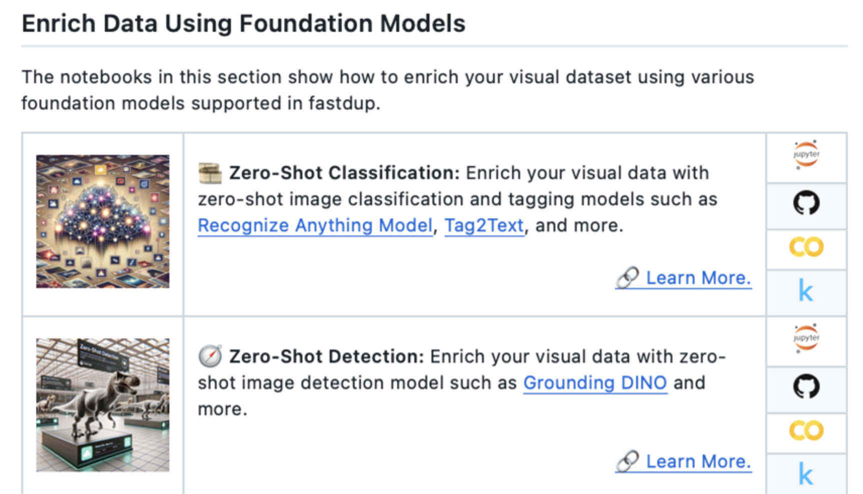
Visual Layer

fastdup: Efficient Blocks for Dataset Curation



In short, you'll need 3 lines of code to run fastdup:

```
import fastdup
fd = fastdup.create(input_dir="IMAGE_FOLDER/")
fd.run()
```



Enrich Data Using Foundation Models

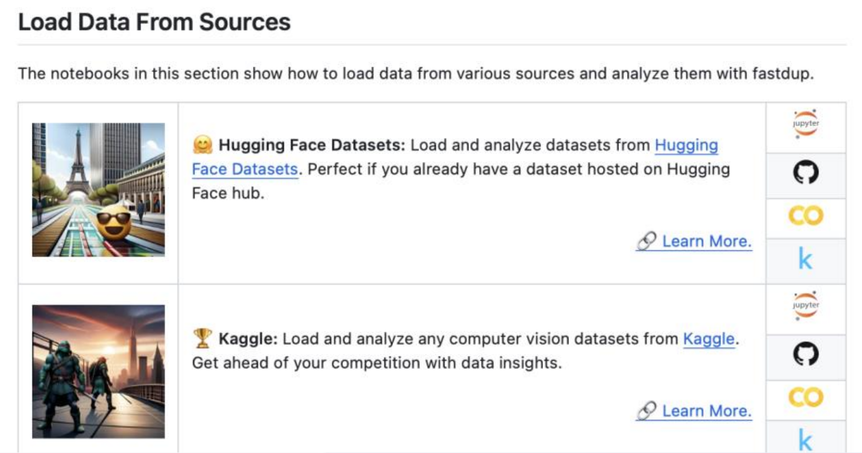
The notebooks in this section show how to enrich your visual dataset using various foundation models supported in fastdup.

🗃️ **Zero-Shot Classification:** Enrich your visual data with zero-shot image classification and tagging models such as Recognize Anything Model, Tag2Text, and more.

🔗 Learn More.

🧭 **Zero-Shot Detection:** Enrich your visual data with zero-shot image detection model such as Grounding DINO and more.

🔗 Learn More.



Load Data From Sources

The notebooks in this section show how to load data from various sources and analyze them with fastdup.

🤗 **Hugging Face Datasets:** Load and analyze datasets from Hugging Face Datasets. Perfect if you already have a dataset hosted on Hugging Face hub.

🔗 Learn More.

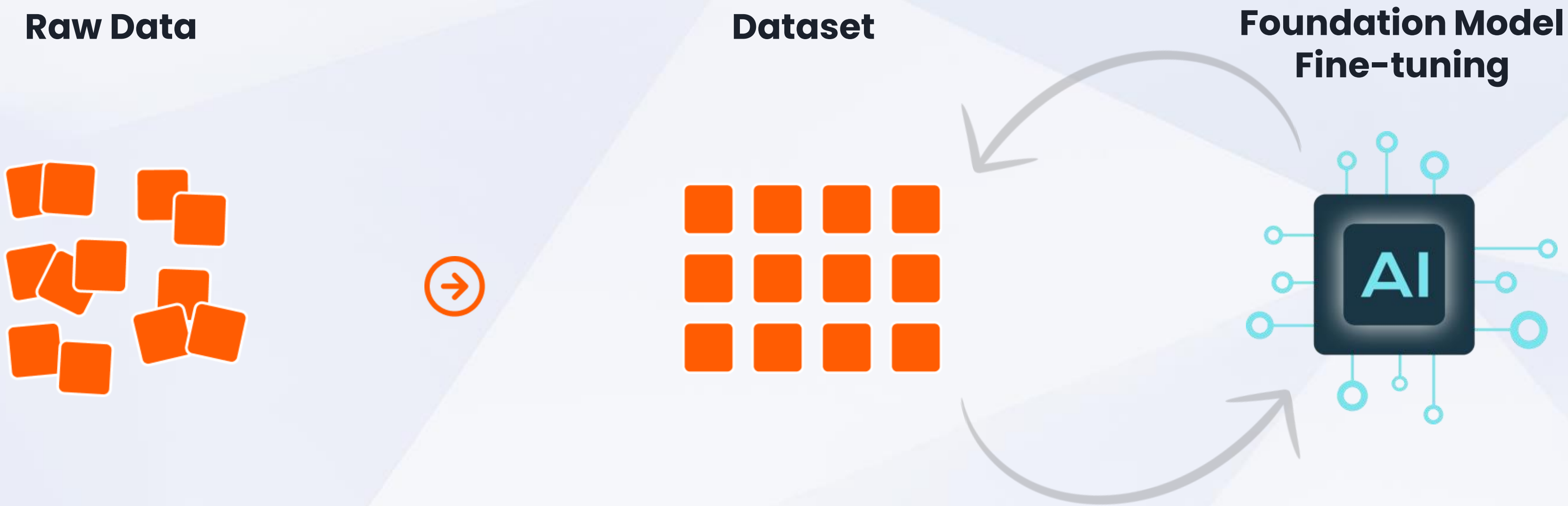
🏆 **Kaggle:** Load and analyze any computer vision datasets from Kaggle. Get ahead of your competition with data insights.

🔗 Learn More.

Visual Layer

Foundation Models: Refocusing AI on Data

Foundation Models: Refocusing AI on Data



Visual Layer

From fastdup to Visual-Layer

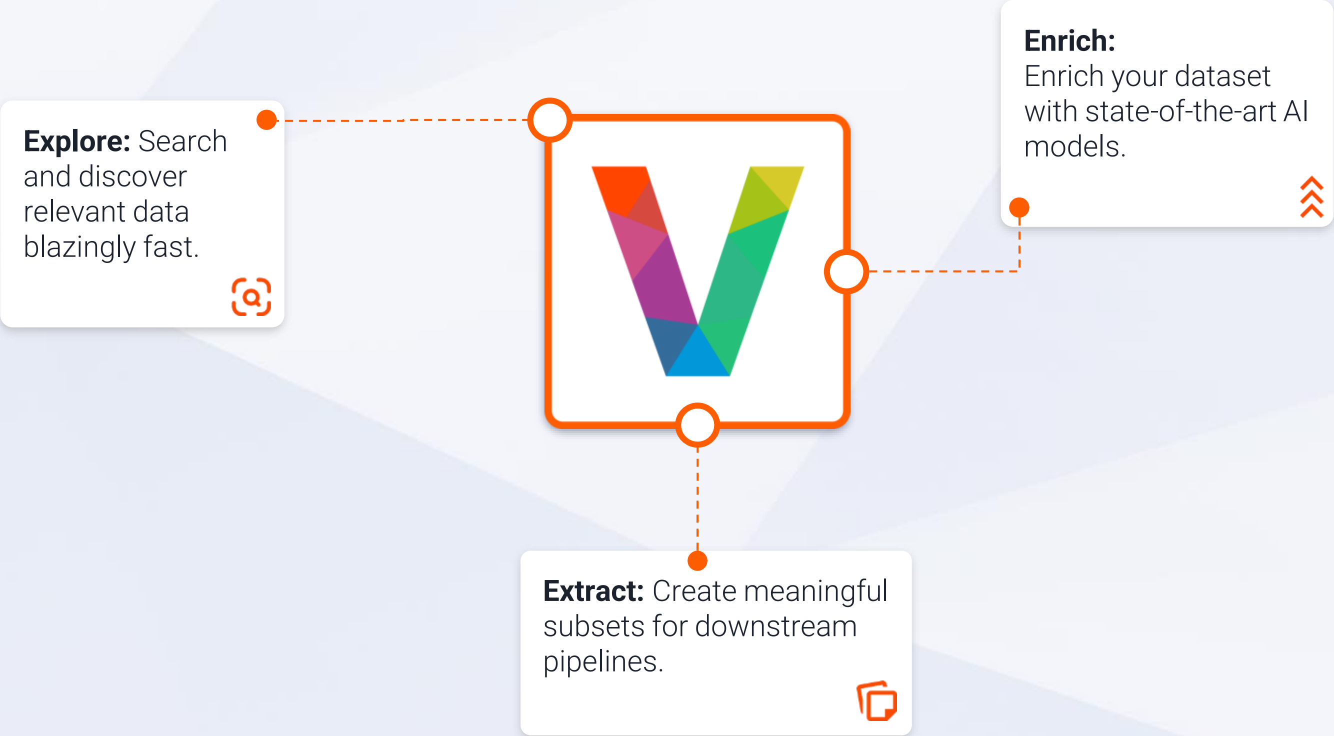
Explore

Enrich

Extract

Visual Layer

Explore – Enrich – Extract

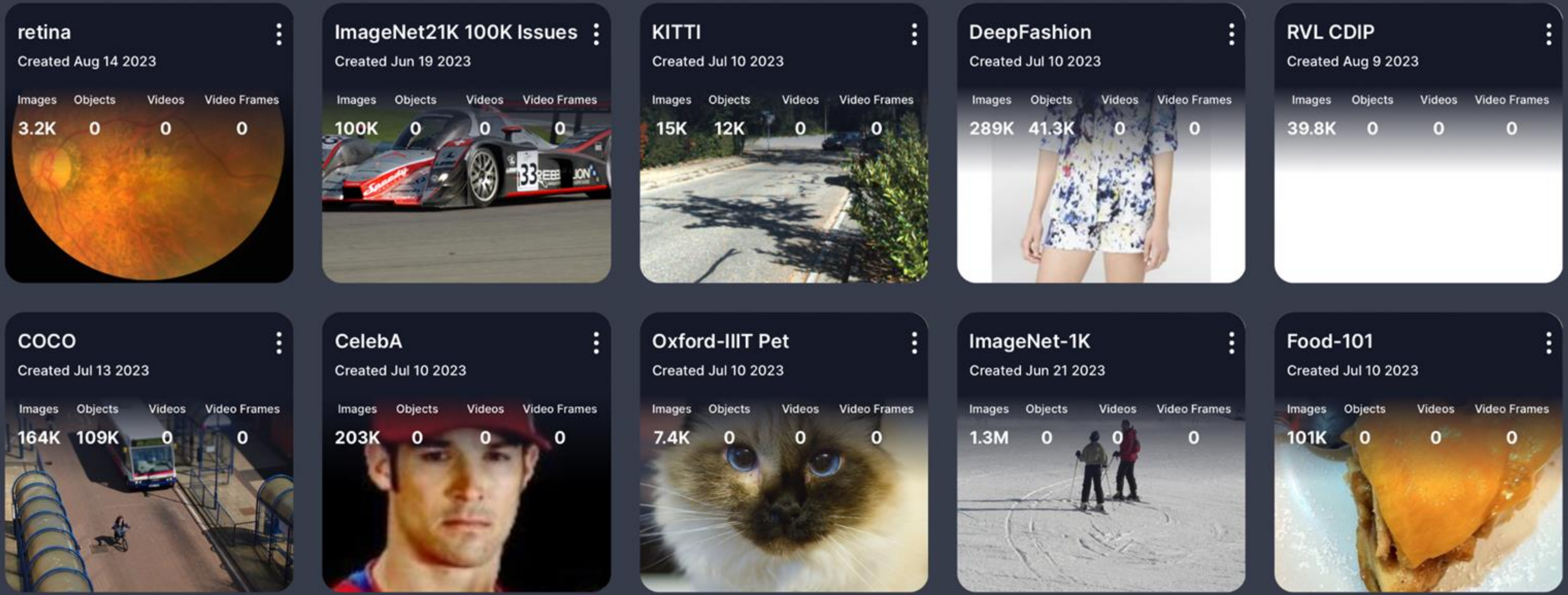


Explore: Search and discover relevant data blazingly fast.

Enrich:
Enrich your dataset with state-of-the-art AI models.

Extract: Create meaningful subsets for downstream pipelines.

Introducing Public VL-Datasets



retina

Created Aug 14 2023

Images	Objects	Videos	Video Frames
3.2K	0	0	0

ImageNet21K 100K Issues

Created Jun 19 2023

Images	Objects	Videos	Video Frames
100K	0	0	0

KITTI

Created Jul 10 2023

Images	Objects	Videos	Video Frames
15K	12K	0	0

DeepFashion

Created Jul 10 2023

Images	Objects	Videos	Video Frames
289K	41.3K	0	0

RVL CDIP

Created Aug 9 2023

Images	Objects	Videos	Video Frames
39.8K	0	0	0

COCO

Created Jul 13 2023

Images	Objects	Videos	Video Frames
164K	109K	0	0

CelebA

Created Jul 10 2023

Images	Objects	Videos	Video Frames
203K	0	0	0

Oxford-IIIT Pet

Created Jul 10 2023

Images	Objects	Videos	Video Frames
7.4K	0	0	0

ImageNet-1K

Created Jun 21 2023

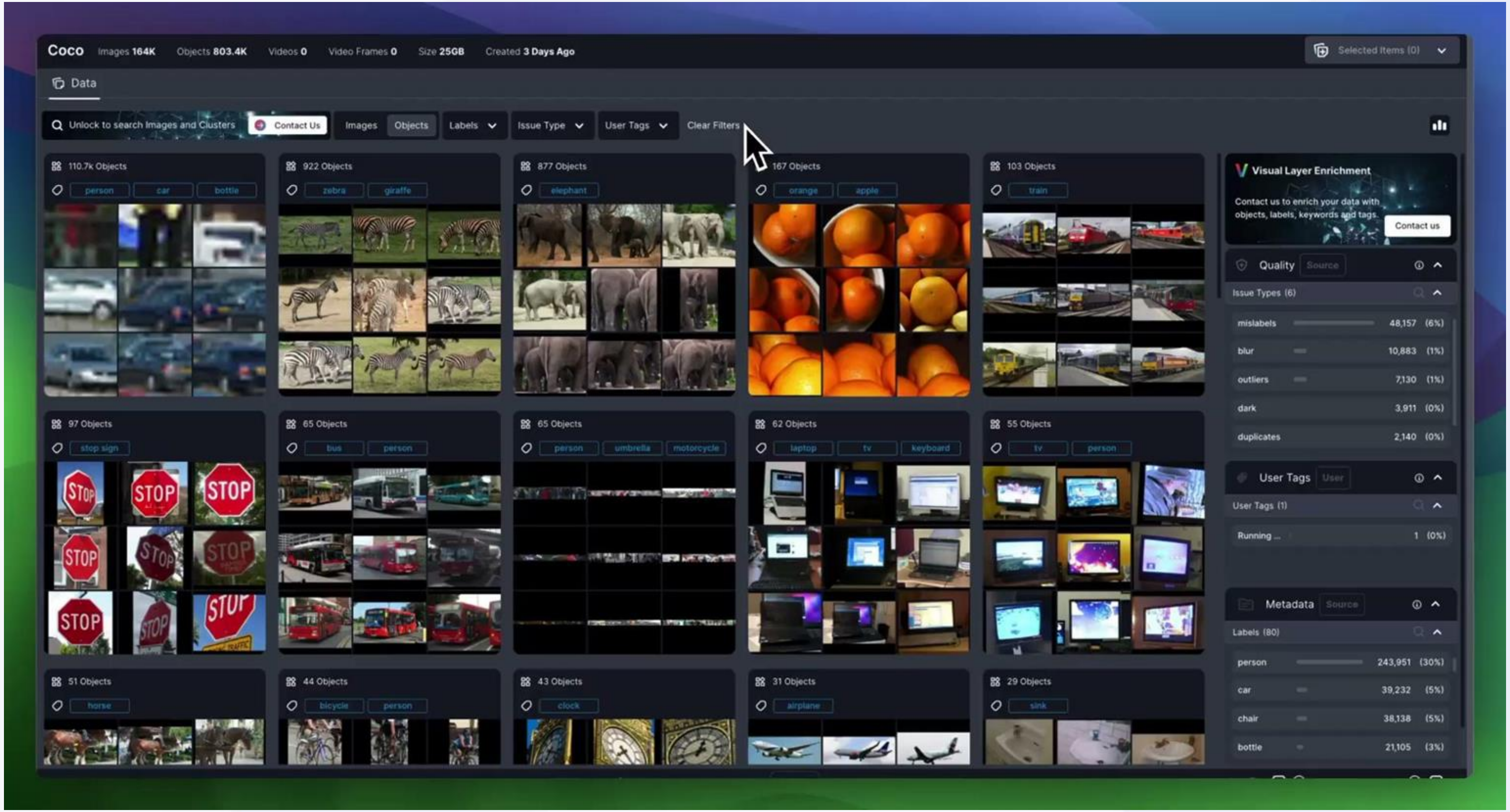
Images	Objects	Videos	Video Frames
1.3M	0	0	0

Food-101

Created Jul 10 2023

Images	Objects	Videos	Video Frames
101K	0	0	0

Public VL-Datasets: Coco Mislabels



Coco Images 164K Objects 803.4K Videos 0 Video Frames 0 Size 25GB Created 3 Days Ago

Visual Layer Enrichment

Contact us to enrich your data with objects, labels, keywords and tags.

Quality

Issue Types (6)

mislabels	48,157	(6%)
blur	10,883	(1%)
outliers	7,130	(1%)
dark	3,911	(0%)
duplicates	2,140	(0%)

User Tags

User Tags (1)

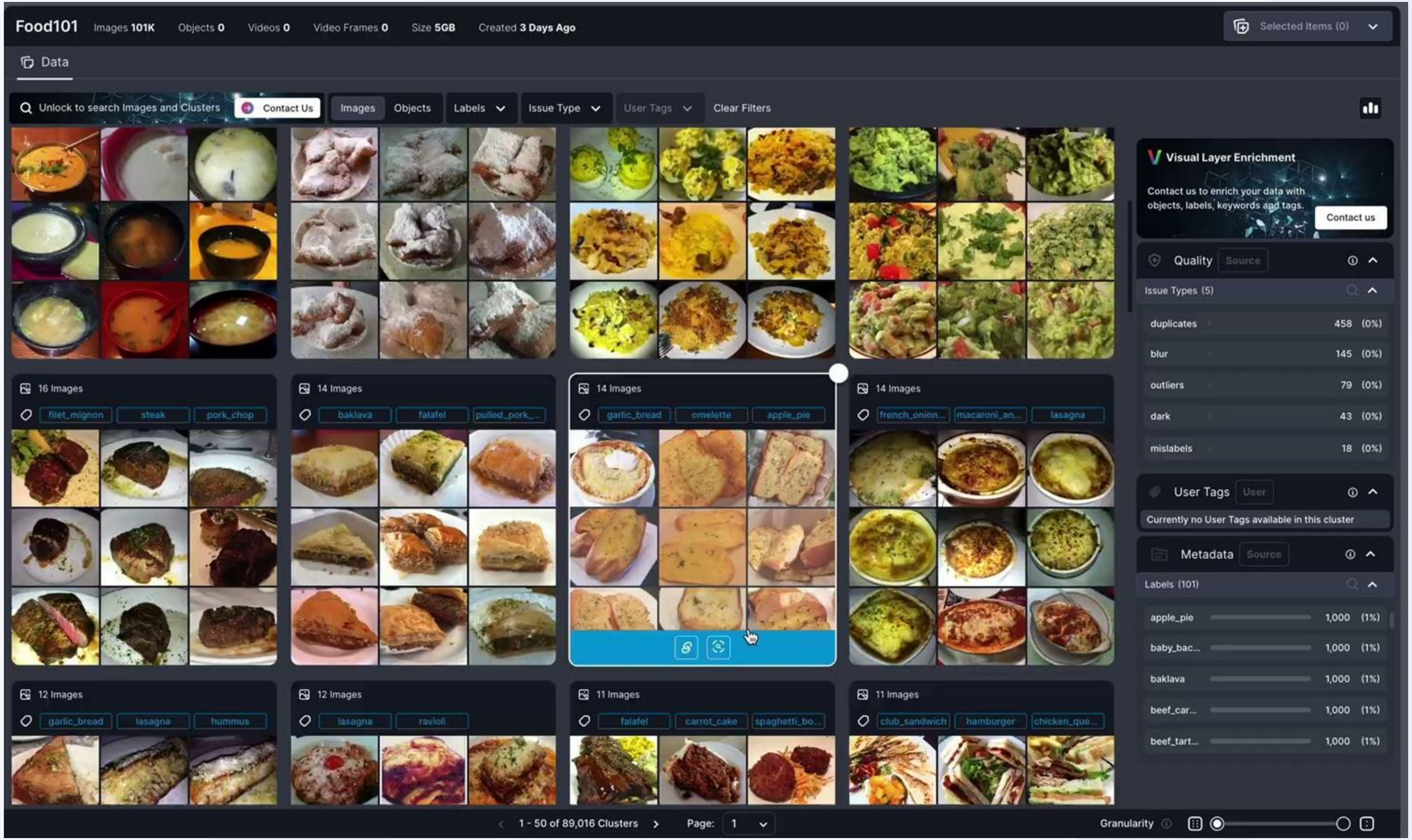
Running ...	1	(0%)

Metadata

Labels (80)

person	243,951	(30%)
car	39,232	(5%)
chair	38,138	(5%)
bottle	21,105	(3%)

Public VL-Datasets: Food-101 Outliers & Batman



Food101 Images 101K Objects 0 Videos 0 Video Frames 0 Size 5GB Created 3 Days Ago

Selected Items (0)

Data

Unlock to search Images and Clusters | Contact Us | Images | Objects | Labels | Issue Type | User Tags | Clear Filters

16 Images: filet_mignon, steak, pork_chop

14 Images: baklava, falafel, pulled_pork_...

14 Images: garlic_bread, omelette, apple_pie

14 Images: french_onion..., macaroni_an..., lasagna

12 Images: garlic_bread, lasagna, hummus

12 Images: lasagna, ravioli

11 Images: falafel, carrot_cake, spaghetti_bo...

11 Images: club_sandwich, hamburger, chicken_que...

1 - 50 of 89,016 Clusters | Page: 1 | Granularity

Visual Layer Enrichment

Contact us to enrich your data with objects, labels, keywords and tags. Contact us

Quality Source

Issue Types (5)

duplicates	458	(0%)
blur	145	(0%)
outliers	79	(0%)
dark	43	(0%)
mislabels	18	(0%)

User Tags User

Currently no User Tags available in this cluster

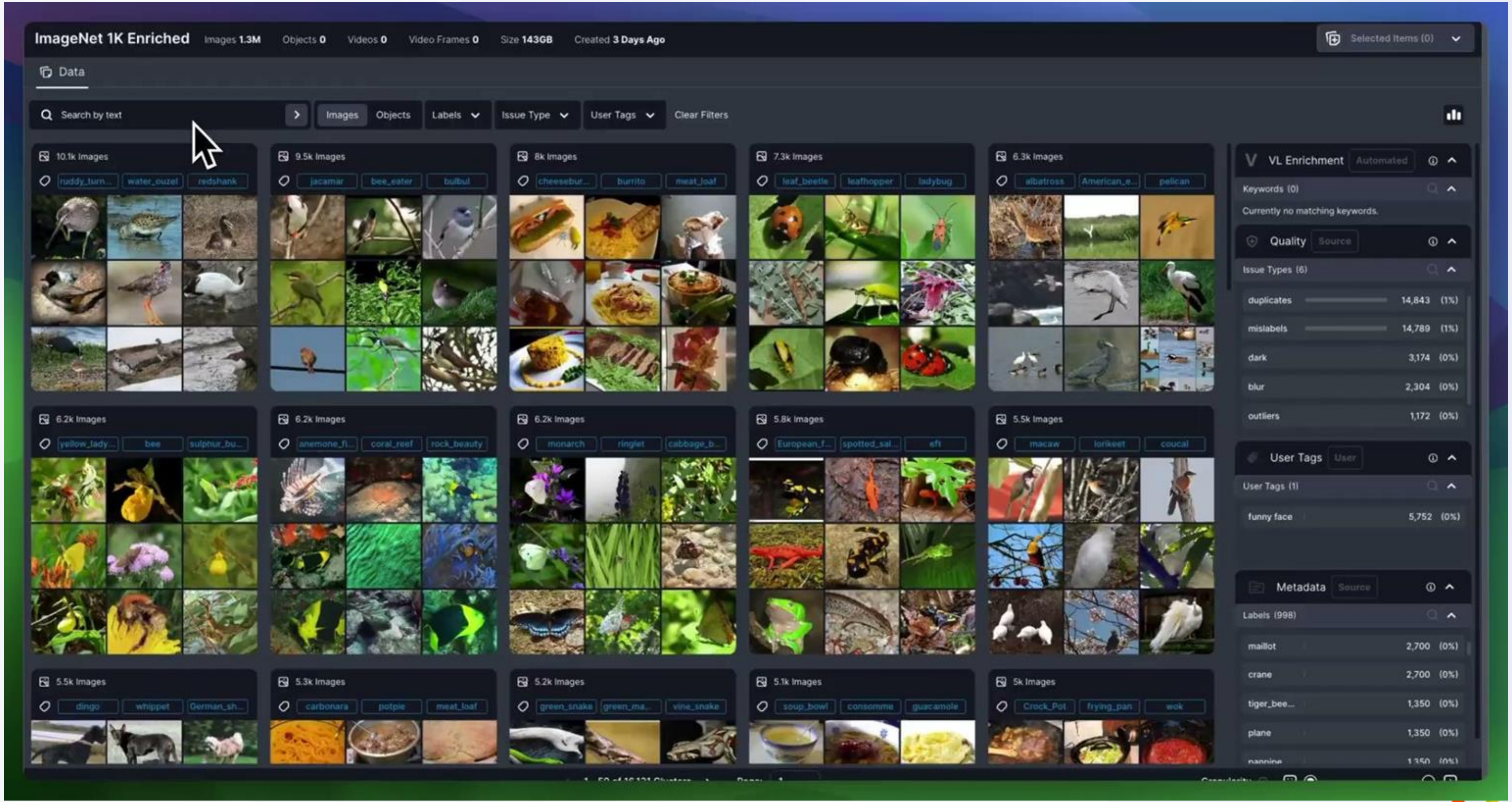
Metadata Source

Labels (101)

apple_pie	1,000	(1%)
baby_bac...	1,000	(1%)
baklava	1,000	(1%)
beef_car...	1,000	(1%)
beef_tart...	1,000	(1%)

Visual Layer

Public VL-Datasets: Imagenet 1K semantic search



ImageNet 1K Enriched   Images 1.3M   Objects 0   Videos 0   Video Frames 0   Size 143GB   Created 3 Days Ago

10.1k Images — ruddy_turn... · water_ouzel · redshank

9.5k Images — jacamar · bee_eater · bulbul

8k Images — cheesebur... · burrito · meat_loaf

7.3k Images — leaf_beetle · leafhopper · ladybug

6.3k Images — albatross · American_e... · pelican

6.2k Images — yellow_lady... · bee · sulphur_bu...

6.2k Images — anemone_fi... · coral_reef · rock_beauty

6.2k Images — monarch · ringlet · cabbage_b...

5.8k Images — European_f... · spotted_sal... · eft

5.5k Images — macaw · lorikeet · coucal

5.5k Images — dingo · whippet · German_sh...

5.3k Images — carbonara · potpie · meat_loaf

5.2k Images — green_snake · green_ma... · vine_snake

5.1k Images — soup_bowl · consomme · guacamole

5k Images — Crock_Pot · frying_pan · wok

VL Enrichment   Automated

Keywords (0)

Currently no matching keywords.

Quality   Source

Issue Types (6)

duplicates	14,843	(1%)
mislabels	14,789	(1%)
dark	3,174	(0%)
blur	2,304	(0%)
outliers	1,172	(0%)

User Tags   User

User Tags (1)

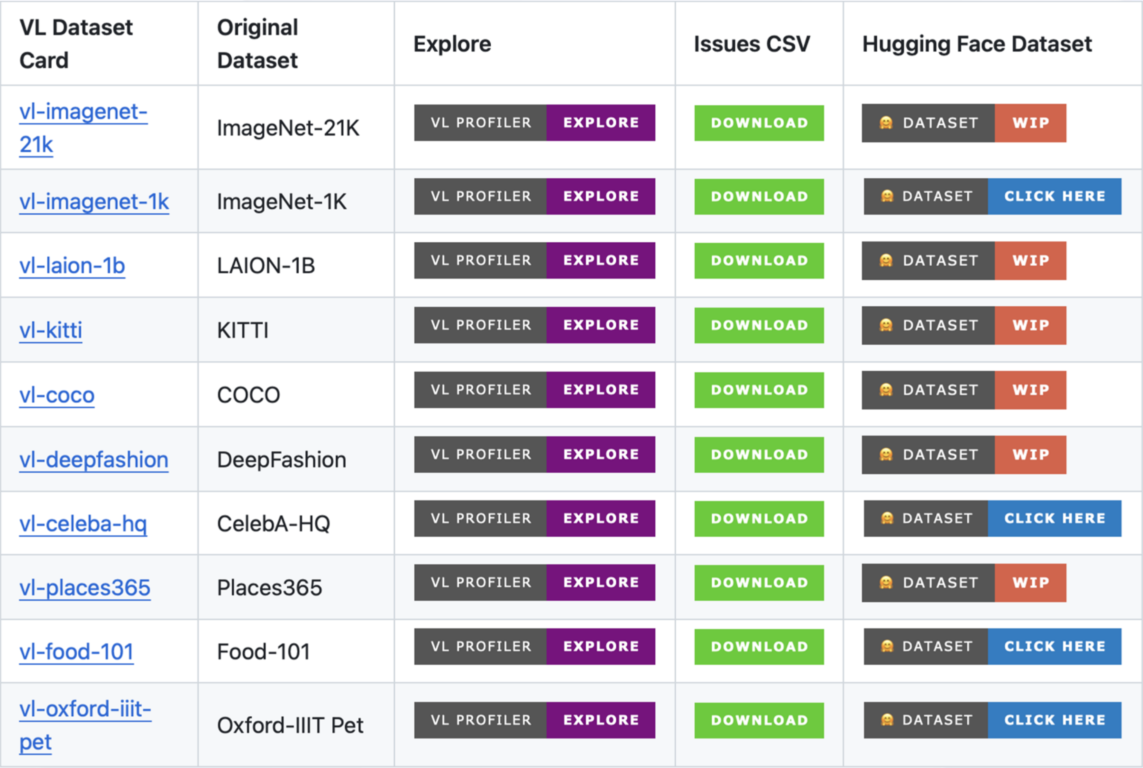
funny face	5,752	(0%)

Metadata   Source

Labels (998)

maillot	2,700	(0%)
crane	2,700	(0%)
tiger_bee...	1,350	(0%)
plane	1,350	(0%)

Download Quality Issues Here



VL Dataset Card	Original Dataset	Explore	Issues CSV	Hugging Face Dataset
vl-imagenet-21k	ImageNet-21K	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET WIP
vl-imagenet-1k	ImageNet-1K	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET CLICK HERE
vl-laion-1b	LAION-1B	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET WIP
vl-kitti	KITTI	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET WIP
vl-coco	COCO	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET WIP
vl-deepfashion	DeepFashion	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET WIP
vl-celeba-hq	CelebA-HQ	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET CLICK HERE
vl-places365	Places365	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET WIP
vl-food-101	Food-101	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET CLICK HERE
vl-oxford-iiit-pet	Oxford-IIIT Pet	VL PROFILER EXPLORE	DOWNLOAD	🤗 DATASET CLICK HERE

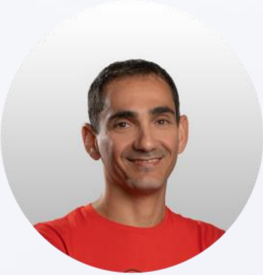
Visual Layer

Our **Team**

Dr. Danny Bickson

Co-founder & CEO

Co-founder and VP EMEA of Turi (acquired by Apple). CMU Researcher. Sr. Mgr at Apple.



Dr. Amir Alush

Co-founder & CTO

Co-founder and CTO of Brodmann17. Highly experienced in building CV/AI Groups and leading into production.

Prof. Carlos Guestrin

Co-founder & CSO

Stanford Prof, Co-founder and CEO of Turi (acquired by Apple). Sr. Dir. at Apple. Deep Learning Infra Team Pioneer.

Our **Team**

Liah Romantsev

Technical Customer support

Sinai Yoktan

Sr. Backend Engineer

Gal Bar Nissan

Staff Engineer

Tom Shani

Machine Learning Engineer

Dima Frid

VP Of Engineering

Elad Yaakov

Director of Product

Daniella Bromkish

Sr. Frontend Engineer

Head of AI

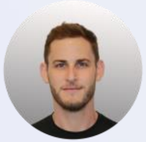
WE ARE HIRING!

Noa Avidar

Backend Engineer

Gagandeep Gambhir

Sr. FrontEnd Engineer



Guy Singer

Machine Learning Engineer

Achiya Jerbi

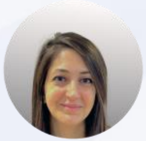
Machine Learning Engineer

Eli Heifetz

Principal Engineer

Ofri Assif

Office Manager



Etti Leibovitz

Head of Product

Nimrod Mozes

Design Lead

Backed **By**

Madrona

INSIGHT PARTNERS

Visual Layer

Visual Layer

🫶

Thank you

Start exploring today