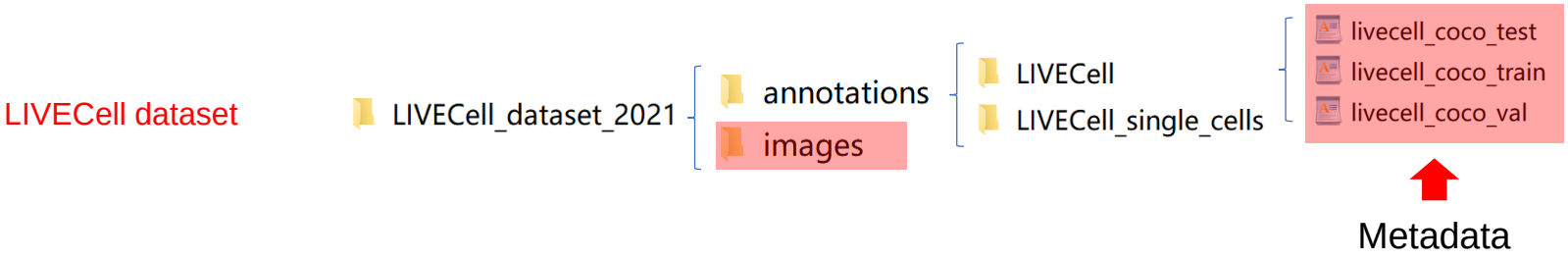
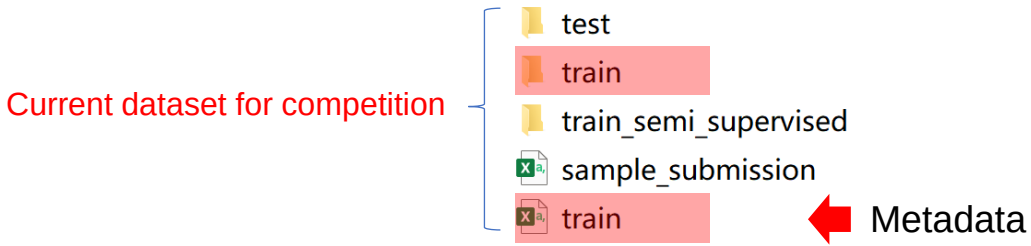
Data Exploration

Datasets


LIVECell dataset
LIVECell_dataset_2021
annotations
images
LIVECell
LIVECell_single_cells
livecell_coco_test
livecell_coco_train
livecell_coco_val
Metadata


Current dataset for competition
test
train
train_semi_supervised
sample_submission
train
Metadata

LIVECell dataset - Metadata

Images

	id	width	height	file_name	original_filename	url
0	1	704	520	BT474_Phase_A3_2_00d04h00m_3.tif	BT474_Phase_A3_2_00d04h00m_3.png	https://darwin.v7labs.com/api/images/870028/or...
1	133	704	520	BT474_Phase_C3_2_02d12h00m_4.tif	BT474_Phase_C3_2_02d12h00m_4.png	https://darwin.v7labs.com/api/images/37512/ori...

Annotations

	id	image_id	category_id	segmentation	area	bbox	iscrowd
2	2	1	1	[[288.02, 305.63, 286.01, 298.87, 286.01, 295....	307.47860	[286.01, 287.73, 19.170000000000016, 20.269999...	0
3	3	1	1	[[271.22, 323.34, 267.93, 322.61, 266.29, 320....	247.47555	[263.0, 304.9, 20.44999999999999, 18.439999999...	0

Categories …

Info …

Licenses …

LIVECell dataset – Metadata

- Conclusions applied to training, validation and testing
  - All the entries in *id* col of **Images** and *image_id* col of **Annotations** can be paired with each other.
  - All the entries in *width* and *height* col of **Images** are of the same values: (x, y) = (704, 502).
  - All the entries in *file_name* of **Images** are valid in terms of the required file name suffix ".tif" and their counterparts(images with the same file names) can be found in the corresponding image directory.

Images

	id	width	height	file_name	original_filename	url
0	1	704	520	BT474_Phase_A3_2_00d04h00m_3.tif	BT474_Phase_A3_2_00d04h00m_3.png	https://darwin.v7labs.com/api/images/870028/or...

Annotations

	id	image_id	category_id	segmentation	area	bbox	iscrowd
2	2	1	1	[[288.02, 305.63, 286.01, 298.87, 286.01, 295....	307.47860	[286.01, 287.73, 19.170000000000016, 20.269999...	0

LIVECell dataset – Metadata

- Conclusions applied to training, validation and testing
  - All the entries in *bbox* col of **Annotations** are valid in terms of the requirement that there must be the coordinate of top-left corner as well as the width and height of bbox available.
  - All the entries in *category_id* col of **Annotations** have only one value: 1 -> all the images are label-free in terms of classification. (But the info for class labeling is also available via some transformation if needed.)
  - All the entries in *iscrowd* col of **Annotations** have only one value: 0 -> all the images are segmented based on polygon.
  - All the entries in *segmentation* of **Annotations** are valid in terms of the requirement that there are at least 3 points to form the polygon of an instance.
  - All the entries in *area* of **Annotations** valid in terms of the requirement that the size must be greater than zero.

Images

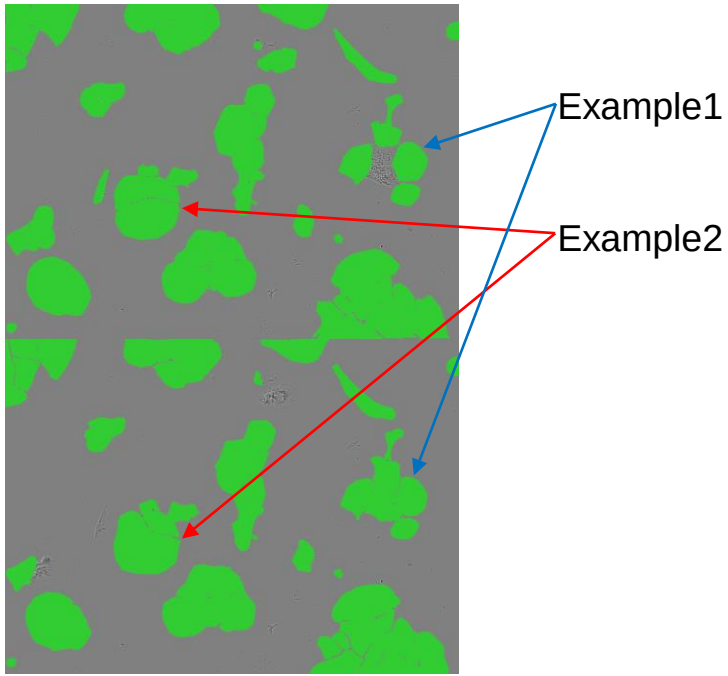
	id	width	height	file_name	original_filename	url
0	1	704	520	BT474_Phase_A3_2_00d04h00m_3.tif	BT474_Phase_A3_2_00d04h00m_3.png	https://darwin.v7labs.com/api/images/870028/or...

Annotations

	id	image_id	category_id	segmentation	area	bbox	iscrowd
2	2	1	1	[[288.02, 305.63, 286.01, 298.87, 286.01, 295....	307.47860	[286.01, 287.73, 19.170000000000016, 20.269999...	0

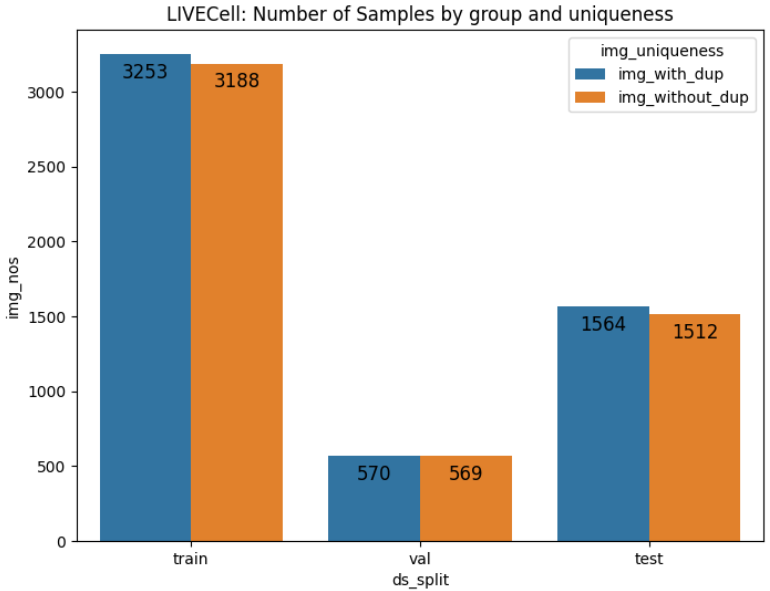
LIVECell dataset – Metadata

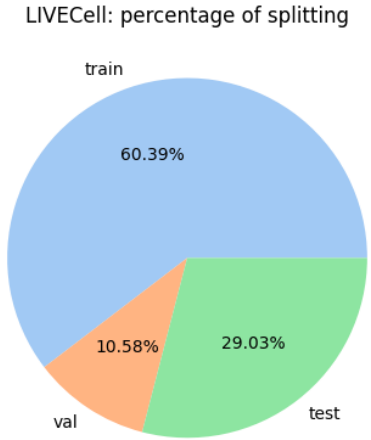
- Conclusions applied to training, validation and testing
  - Even though all the entries in *id* col of **Images** (or *image_id* col of **Annotations**) are different, some of them point to the same images, but with different annotation styles, as illustrated on the right hand.
  - Such phenomon takes place within each splitted dataset and between training and validation dataset (amount to 30).



LIVECell dataset – Metadata

- Conclusions applied to training, validation and testing





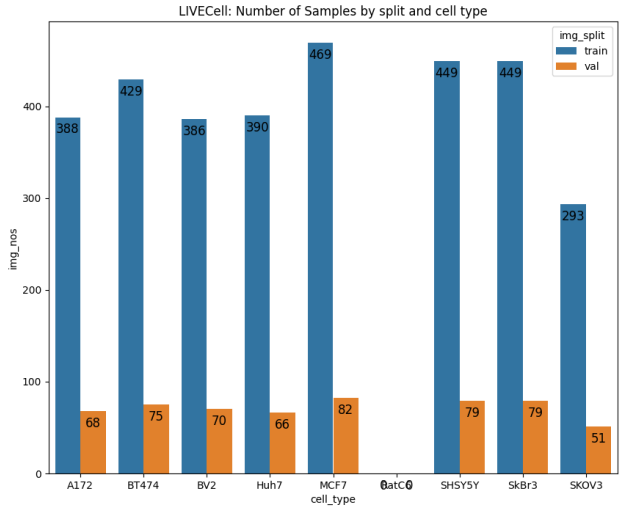
Current dataset for competition – Metadata

- Conclusions applied to training and validation
  - The unique *image ids* of the given metadata count to **606**, and all of their counterparts can be found in the directory *dataset/train*, which also owns 606 valid images.
  - All the images are of the same size **(704, 502)** according to the metadata.
  - Some segmentation annotations for the same image ids are the same, i.e., even for the same images, some segmentation annotations are duplicate to each other. -> To be removed or something else.
  - The split into training and validation dataset are not provided, which should be done manually.
  - The directory *dataset/train_semi_supervised* owns **1972** valid images without any annotations.

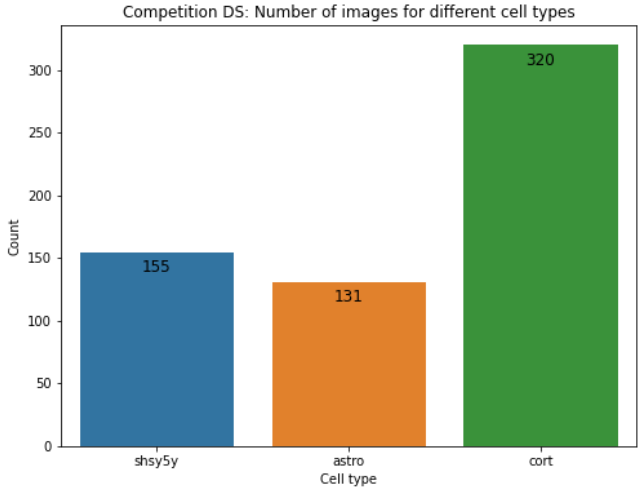
	id	annotation	width	height	cell_type	plate_time	sample_date	sample_id	elapsed_timedelta
0	0030fd0e6378	118145 6 118849 7 119553 8 120257 8 120961 9 1...	704	520	shsy5y	11h30m00s	2019-06-16	shsy5y[diff]_E10-4_Vessel-714_Ph_3	0 days 11:30:00
1	0030fd0e6378	189036 1 189739 3 190441 6 191144 7 191848 8 1...	704	520	shsy5y	11h30m00s	2019-06-16	shsy5y[diff]_E10-4_Vessel-714_Ph_3	0 days 11:30:00

Current dataset for competition – Metadata

- Conclusions applied to training and validation
  - The different cell type categories characterized by the number of images are: **[‘astro’,'cort', 'shsy5y’], the first two being new cell types and the last one being expanded to the original one.**



LIVECell dataset



Current dataset