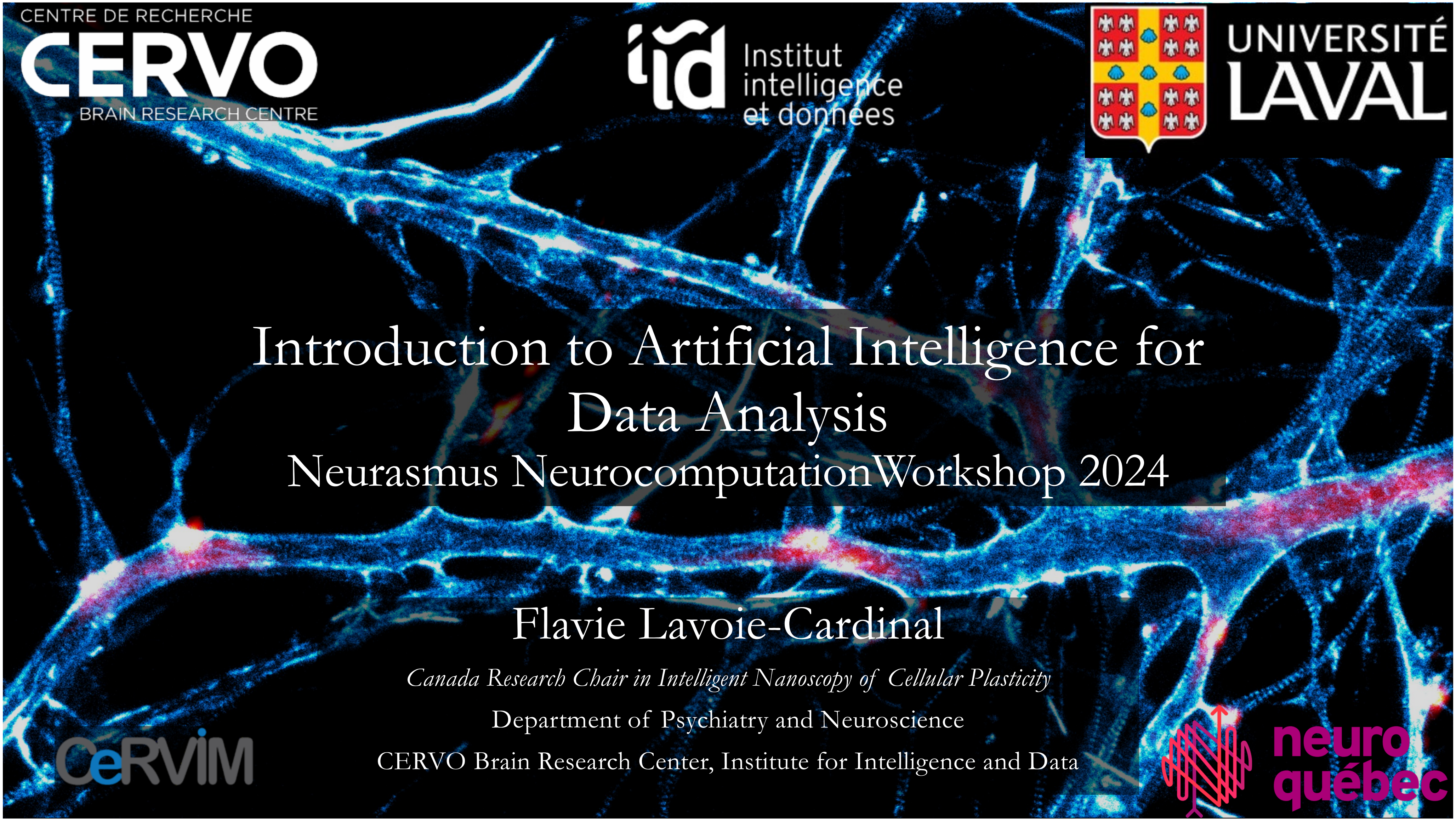


Introduction to Artificial Intelligence for Data Analysis

Neurasmus NeurocomputationWorkshop 2024

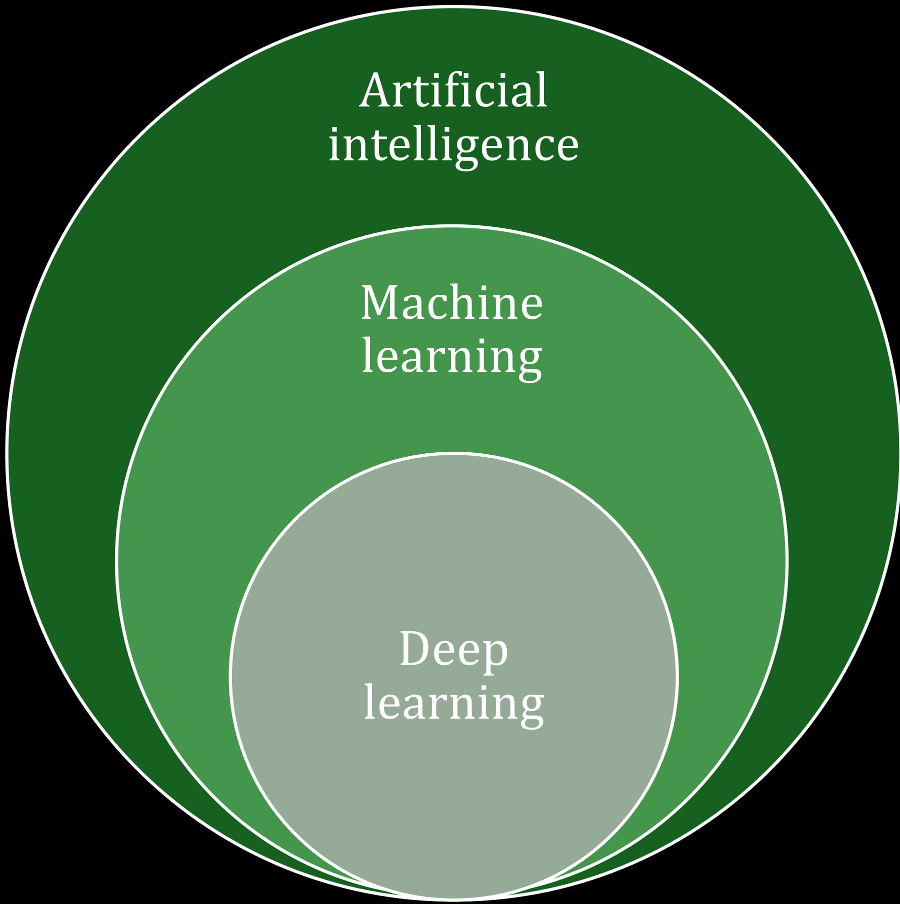
Flavie Lavoie-Cardinal

Canada Research Chair in Intelligent Nanoscopy of Cellular Plasticity

Department of Psychiatry and Neuroscience

CERVO Brain Research Center, Institute for Intelligence and Data


CERVO
neurophotonics.ca
❖ International Summer School
Frontiers in Neurophotonics
❖ Canadian Neurophotonics Center
17th edition
May 28th - June 7th 2024 | Québec City | Canada
Frontiers in Neurophotonics
Summer School
A summer school on advanced optical imaging and photoactivation techniques

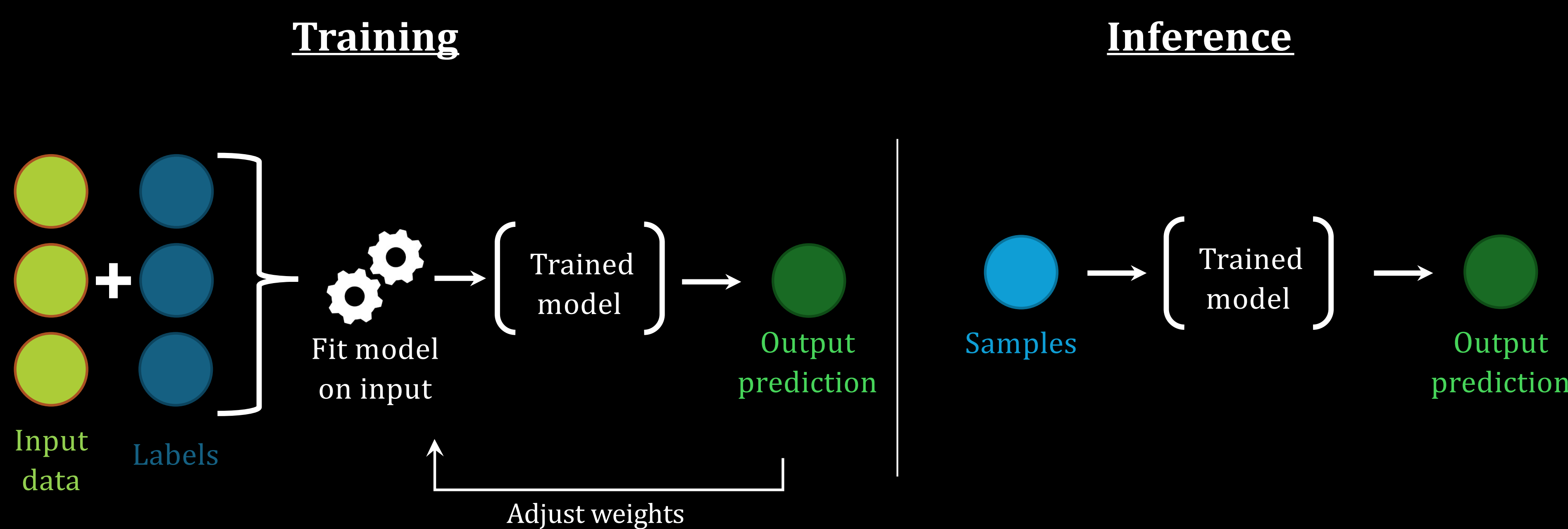


Machine learning is programming computers to optimize a performance criterion using ***example data or past experience***

(Introduction to Machine Learning – Ethem Alpaydin)

What is machine learning?

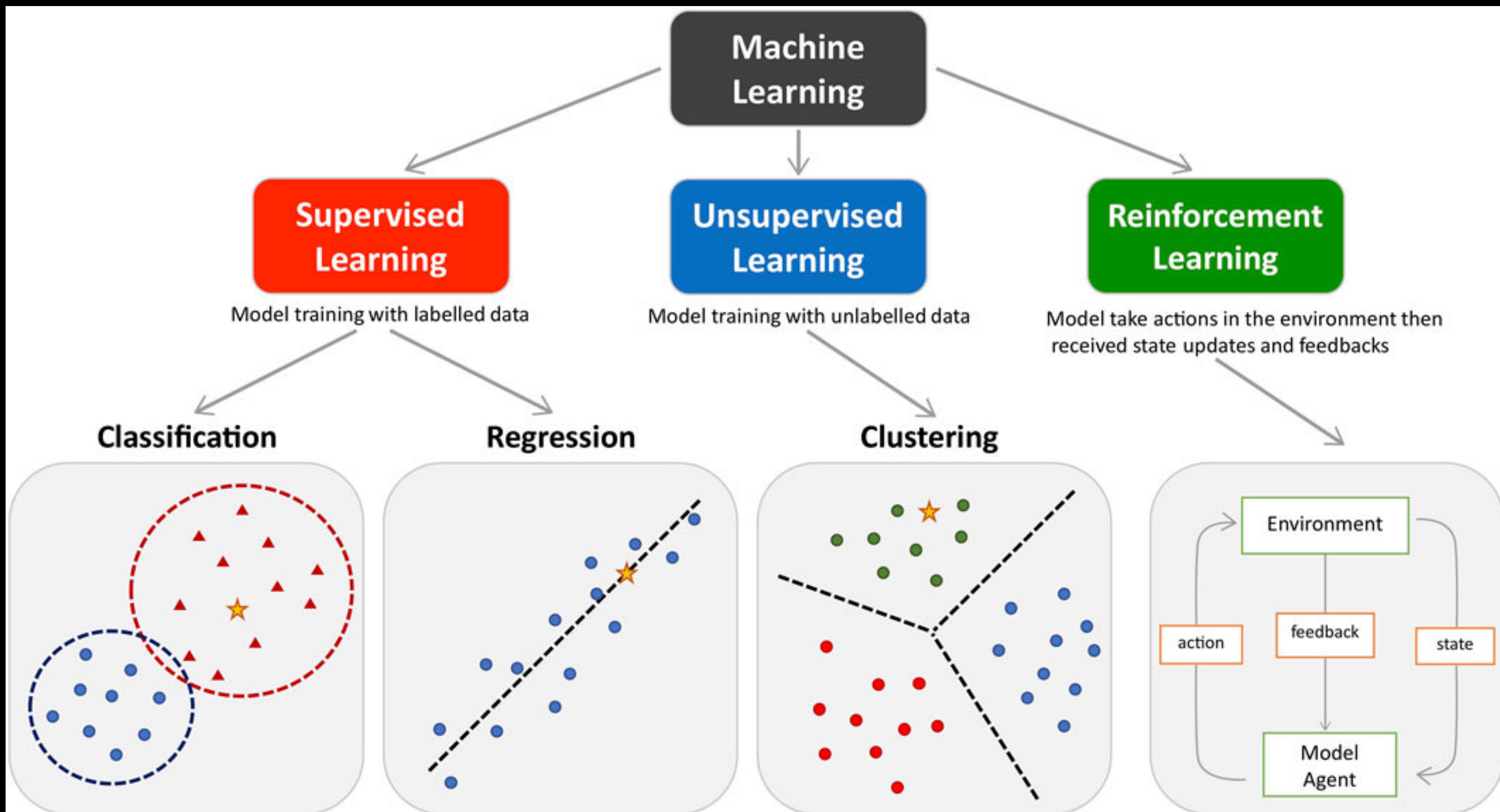
- Algorithms that **learn** from input data to make **predictions** on samples



Terminology of Machine Learning

- **Features :**
  - input variables – the way we represent our data
  - e.g. words in an email text (spam recognition), structure on an image (edges, particular objects)
- **Labels/Annotations :**
  - target that we try to predict
  - e.g. spam or not spam, cat, dog, structure of interest
- **Example :**
  - one piece of data
  - Labeled or not labeled
  - e.g. one image, one e-mail
- **Model :**
  - Algorithm doing the predictions
  - What we want to create by learning from data

Different machine learning strategies



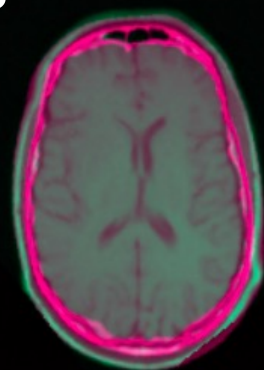
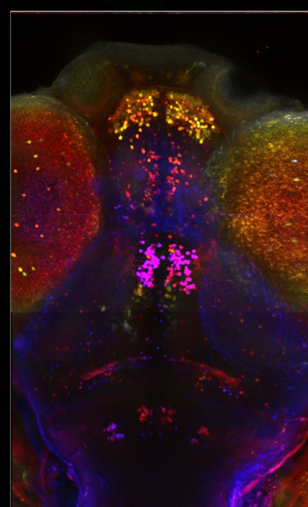
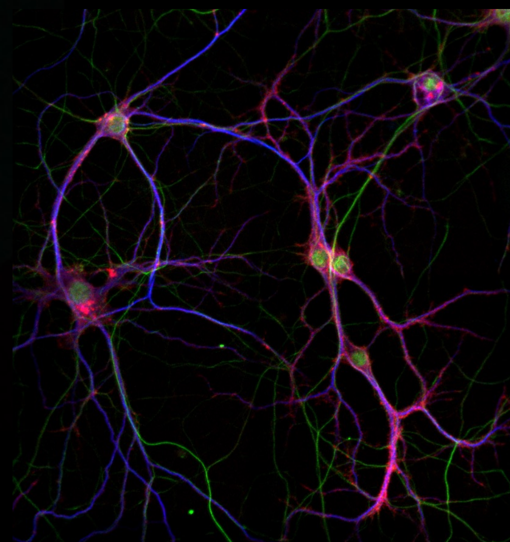
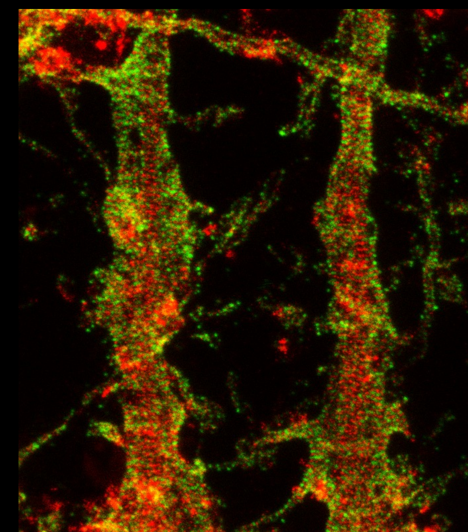
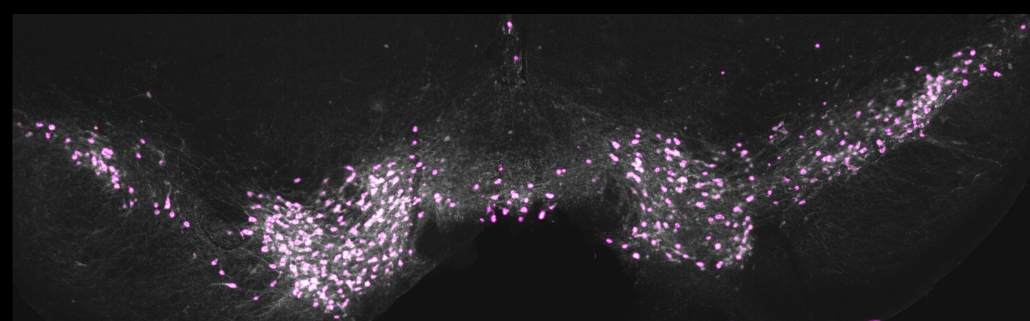
Peng et al. **2021** *Front. Pharmacol. https://doi.org/10.3389/fphar.2021.720694*

Workflow

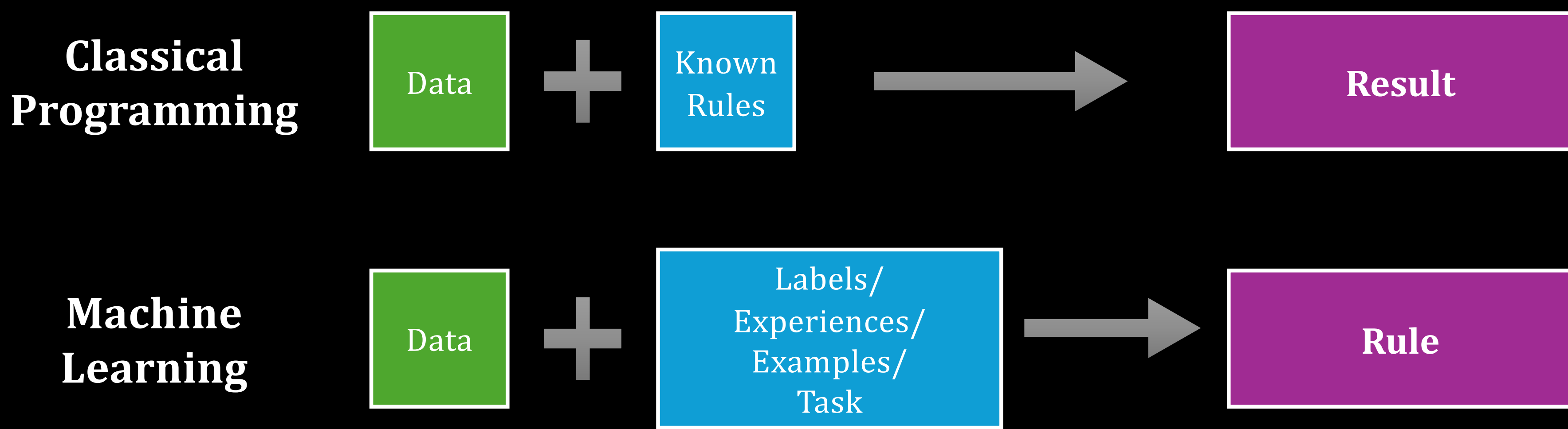
Peng et al. **2021** *Front. Pharmacol. https://doi.org/10.3389/fphar.2021.720694*

Quantitative & High Throughput Biomedical Image Analysis

- Different tasks
- Multiple approaches
- Multiple modalities
- Careful dataset generation

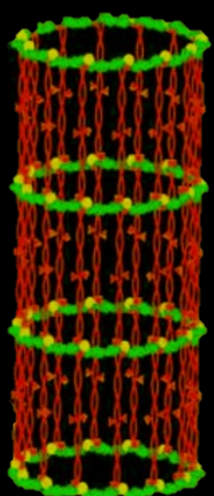






Classical vs ML data analysis approaches

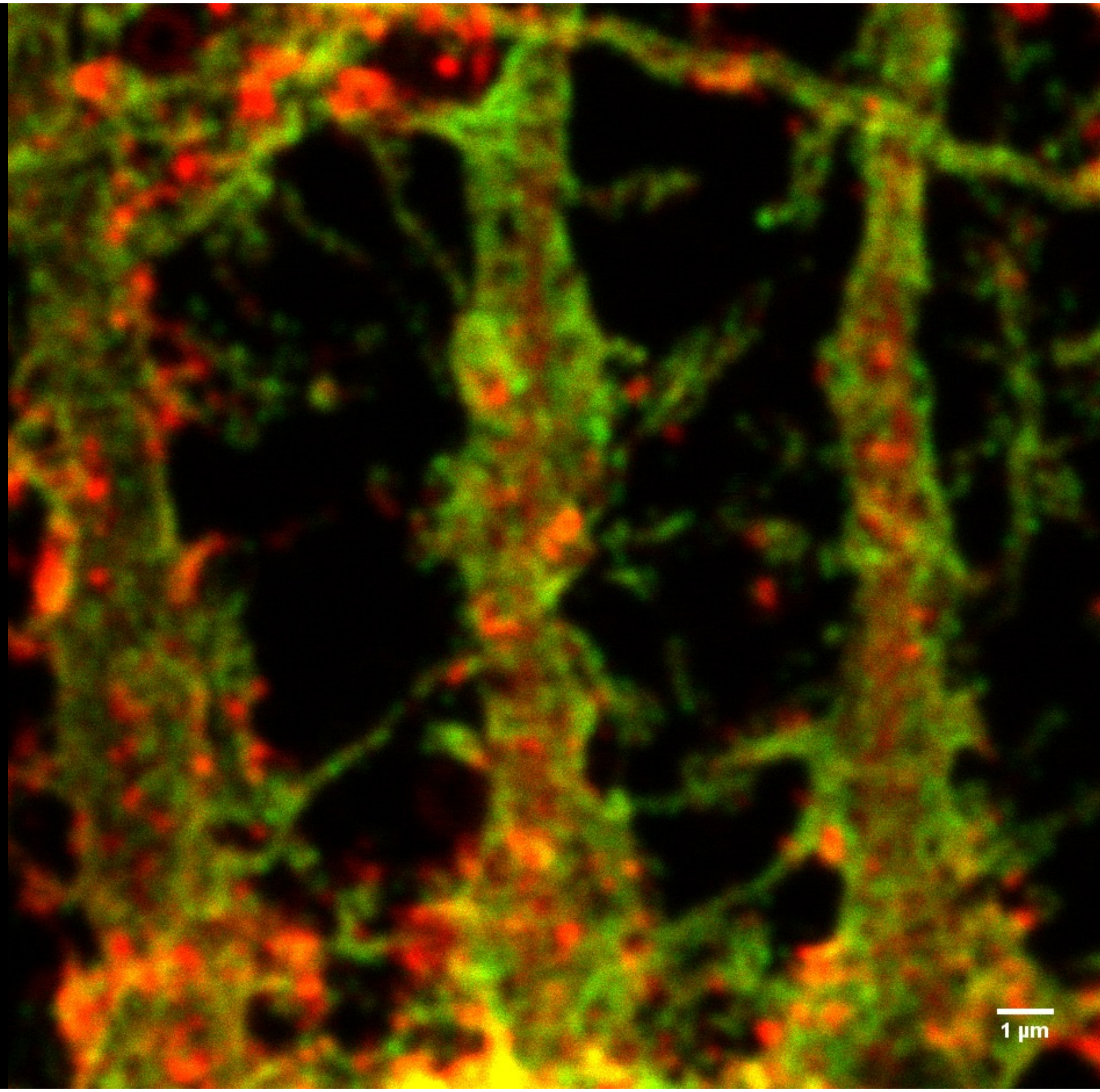



Chihuahua vs Rat

Test case : detecting nanoscopic structures on super-resolution images

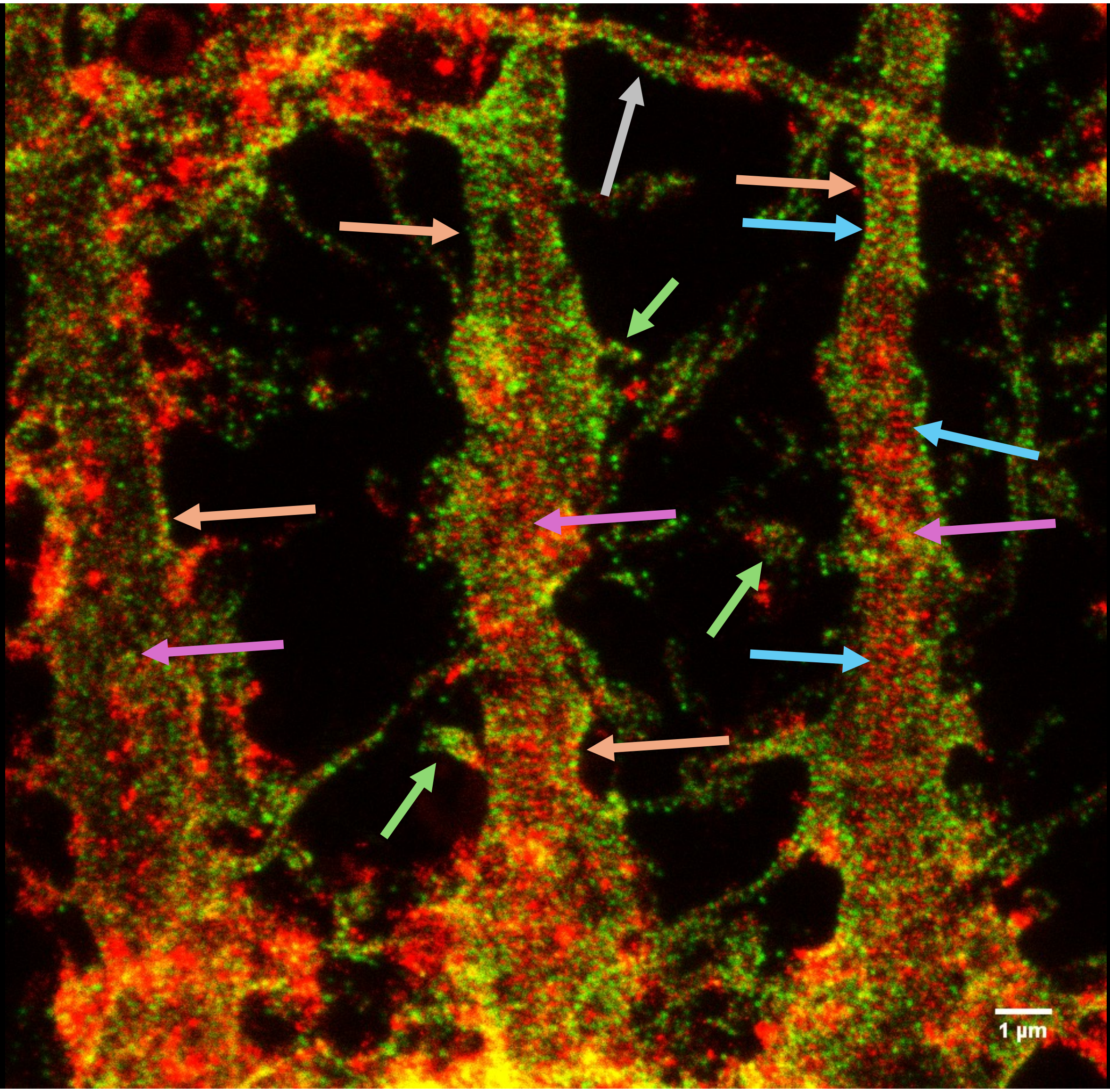


Actin
Spectrin



Features :

- Spine
- Spectrin rings
- Actin rings
- Dendrite
- Axon



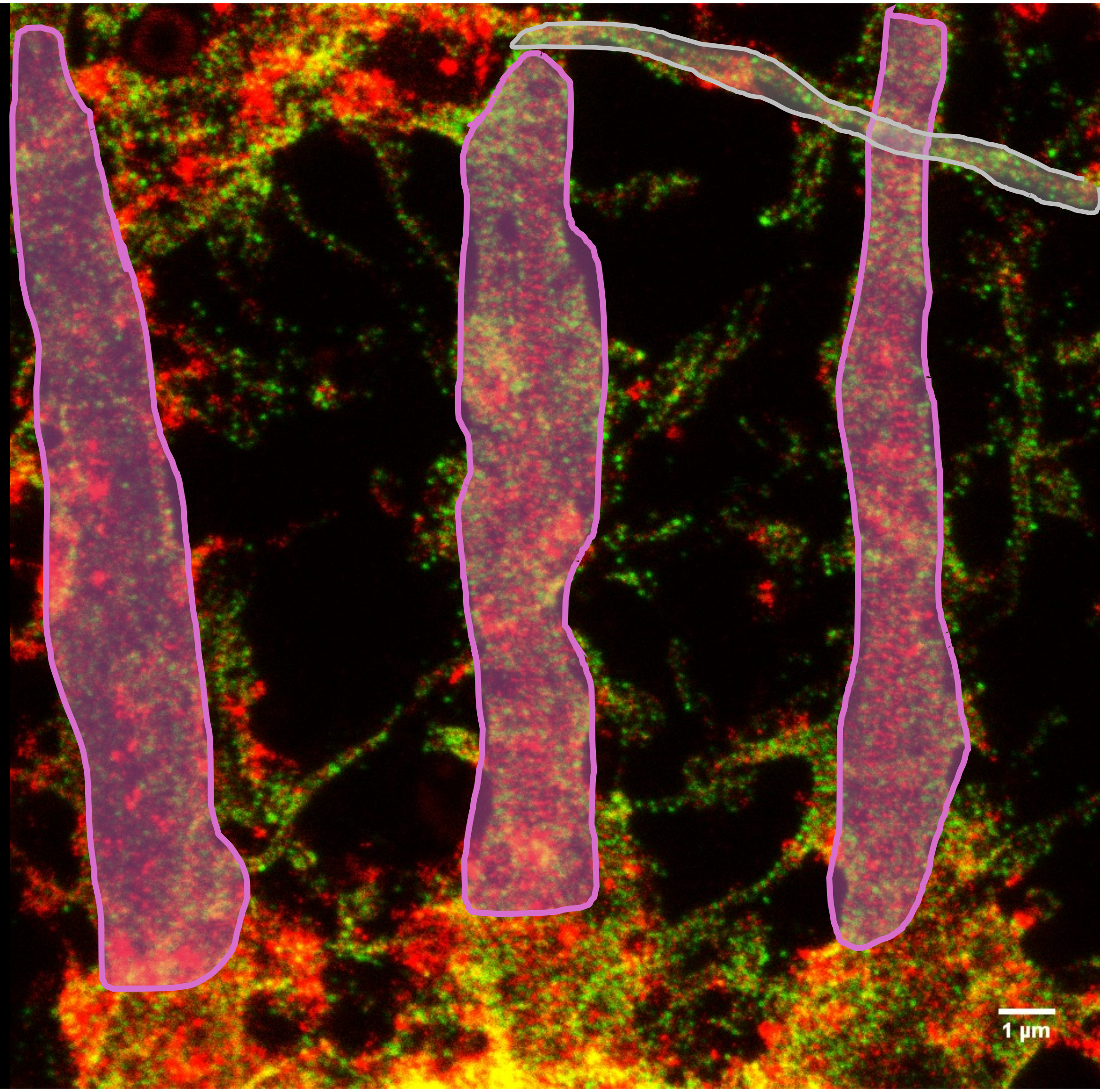
Labels :

A. Image tags

- Spine
- Spectrin rings
- Actin rings
- Dendrite
- Axon

B. Segmentation map

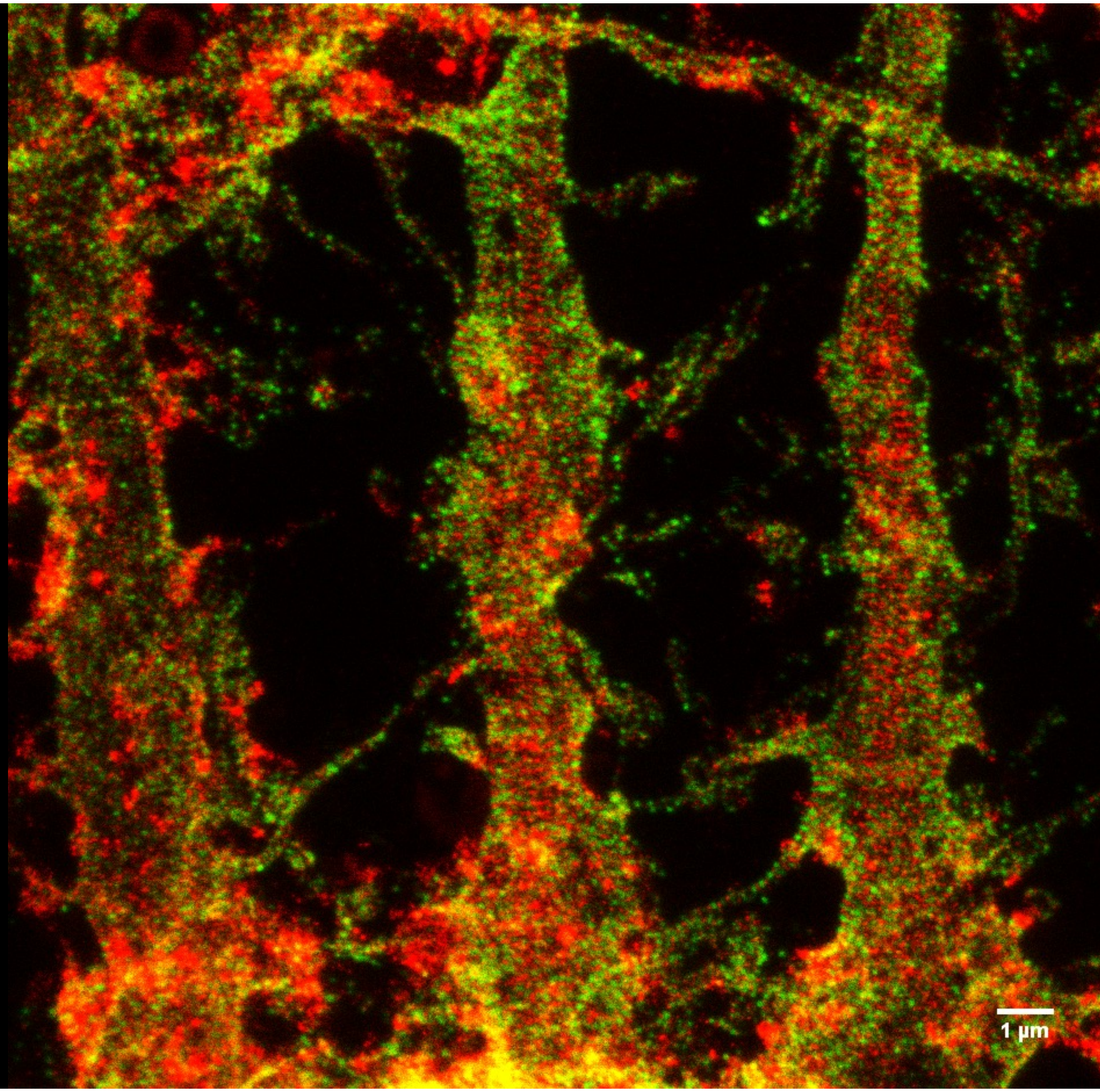
- Dendrite
- Axon



Models :
Defines the relationship between labels and features

1. **Training :**
    - The model gradually learn to associate features (F-actin rings and fibers) and labels (expert annotations) by seeing many labeled examples (images)

2. **Inference :**
    - Use the trained model to make usefull predictions (on ulabeled examples)



What to do **before**, during, and after starting a project involving deep learning with your data?

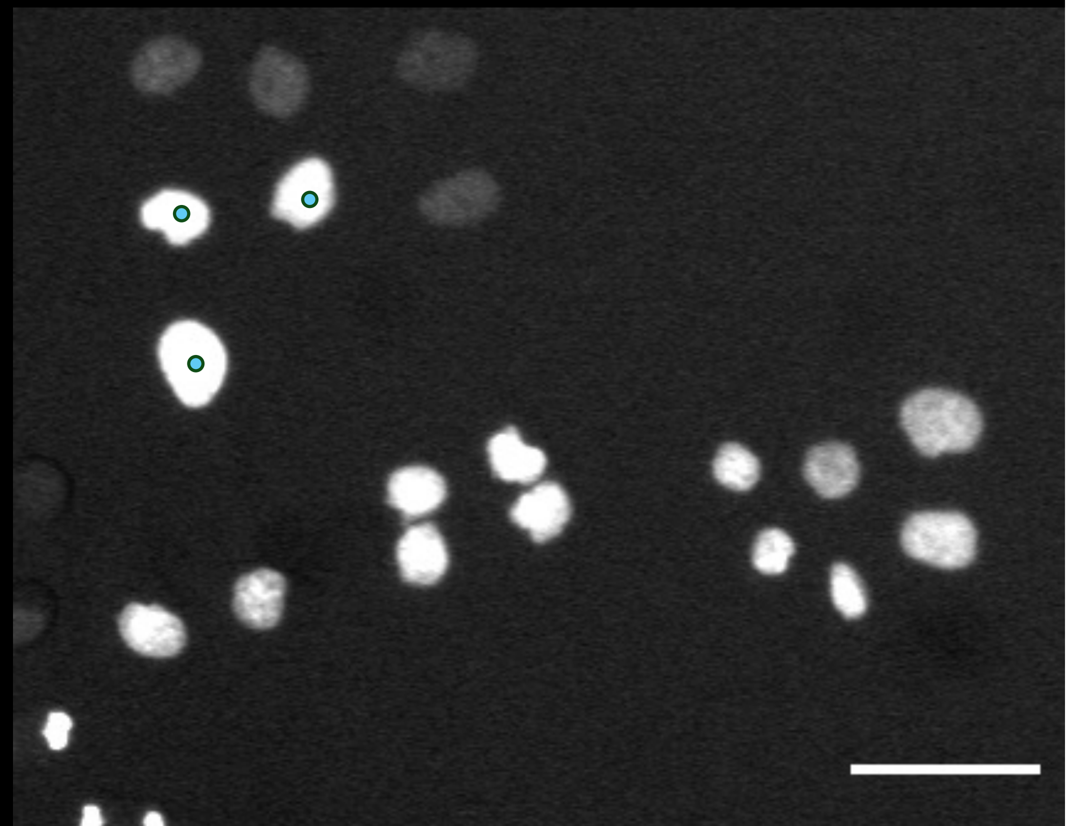
1. Curate your dataset

What to do **before**, during, and after starting a project involving deep learning with your data?

1. Curate your dataset
2. Define the task you want to address

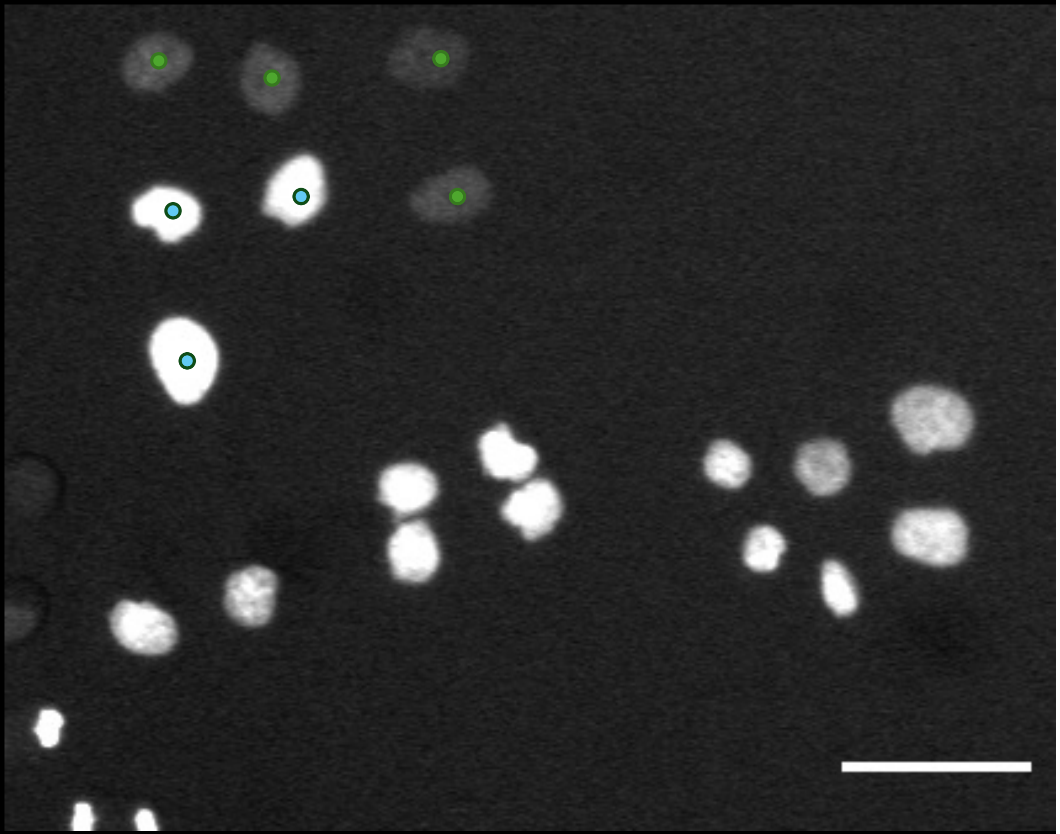
Common tasks

- Detection



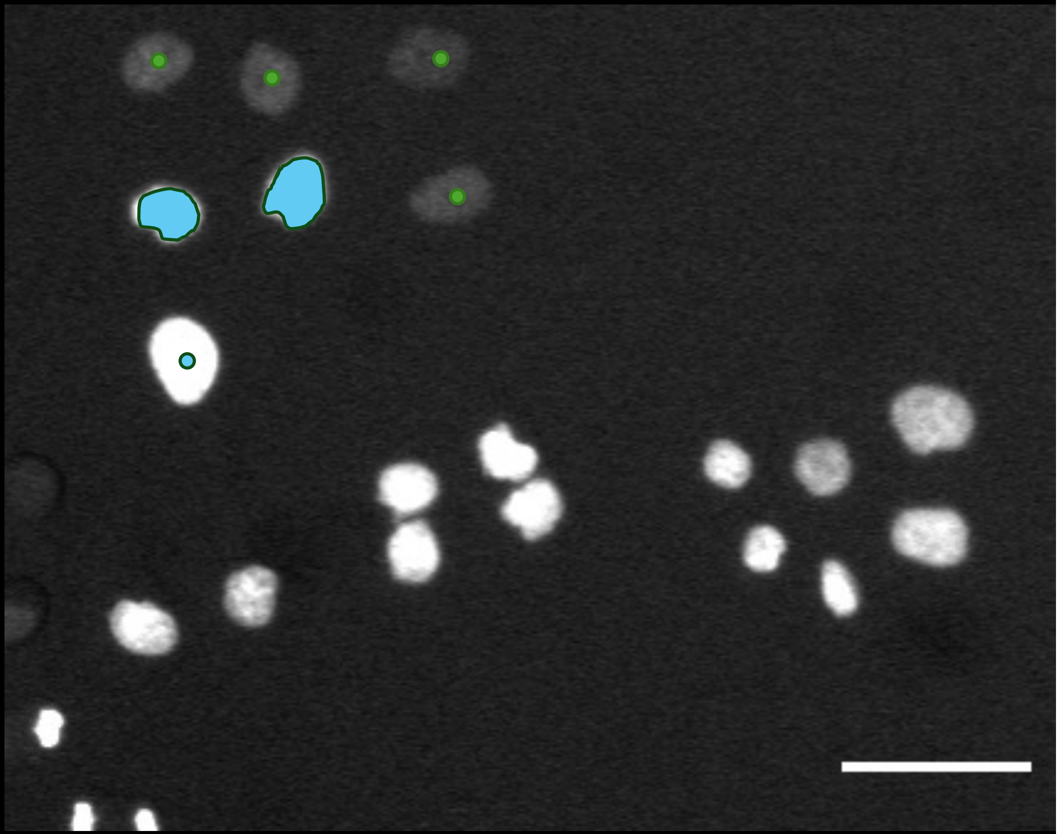
Common tasks

- Detection
- Classification



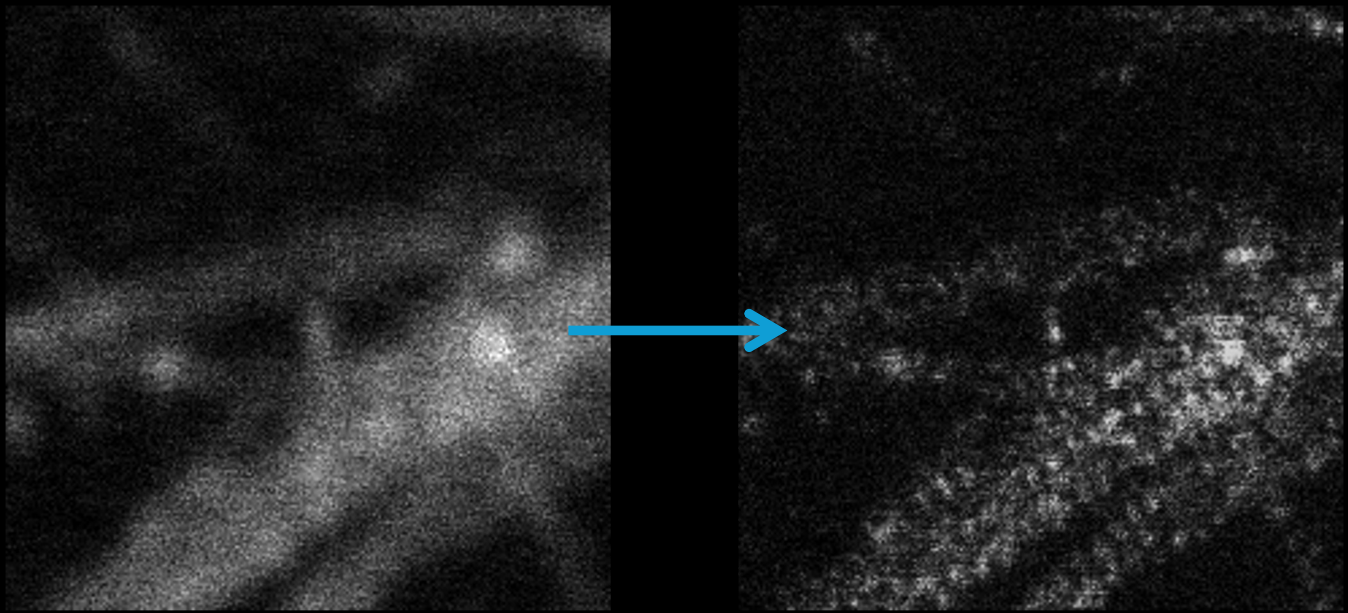
Common tasks

- Detection
- Classification
- Segmentation



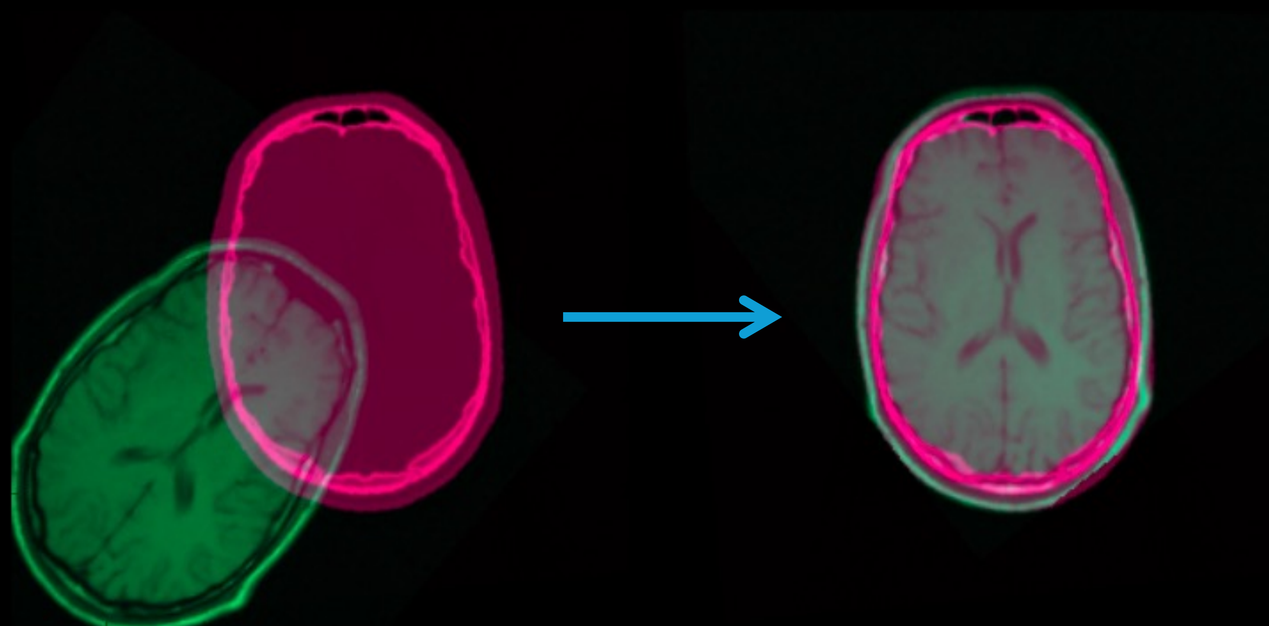
Common tasks

- Detection
- Classification
- Segmentation
- Generation



Common tasks

- Detection
- Classification
- Segmentation
- Generation
- Registration

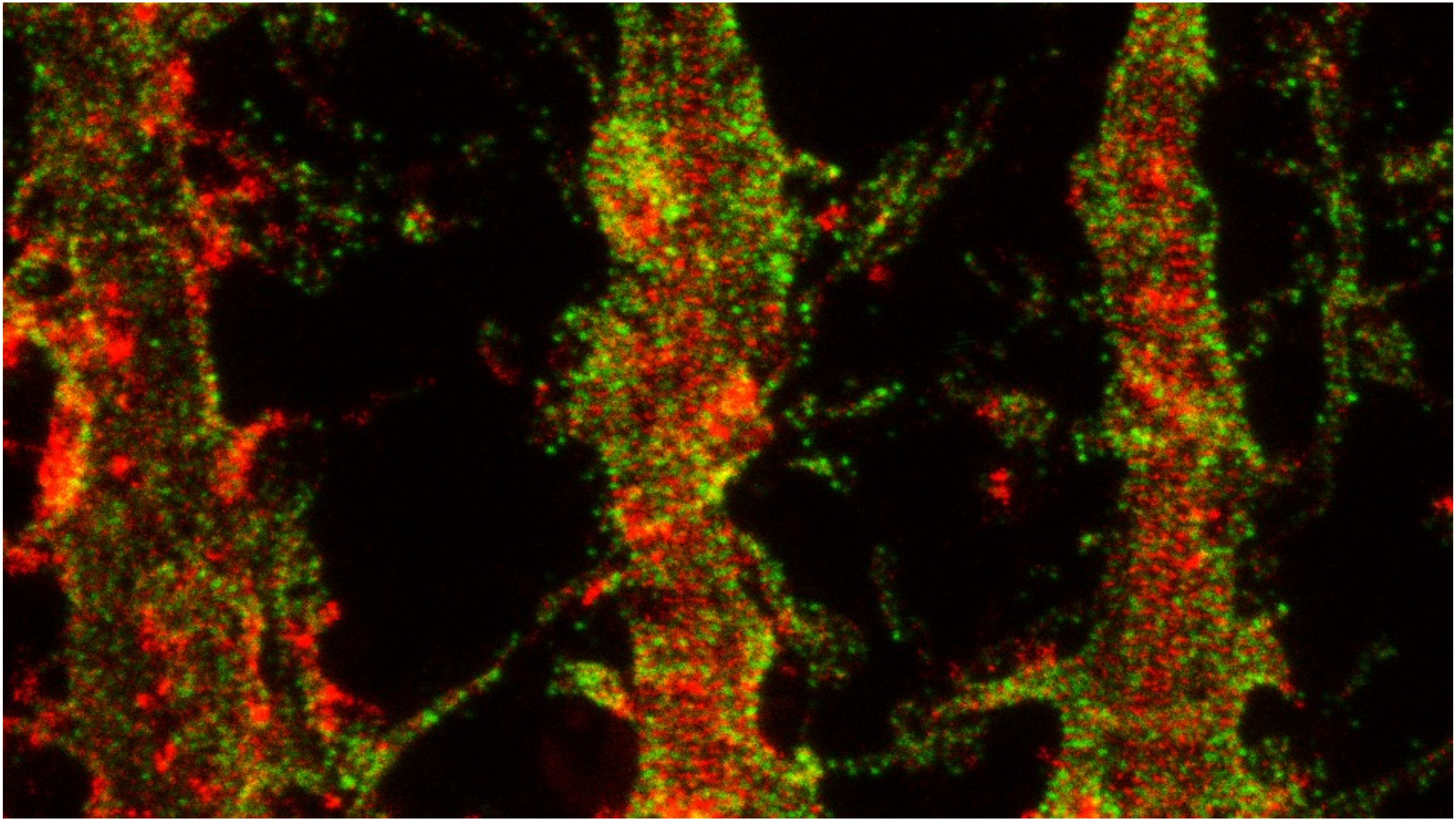


https://nbviewer.jupyter.org/github/ozan-oktay/Medical-Image-Analysis-IPython-Tutorials/blob/master/tutorial2/multimodal_reg.ipynb

Bouchard et al. 2023 Nat. Mach. Intell.

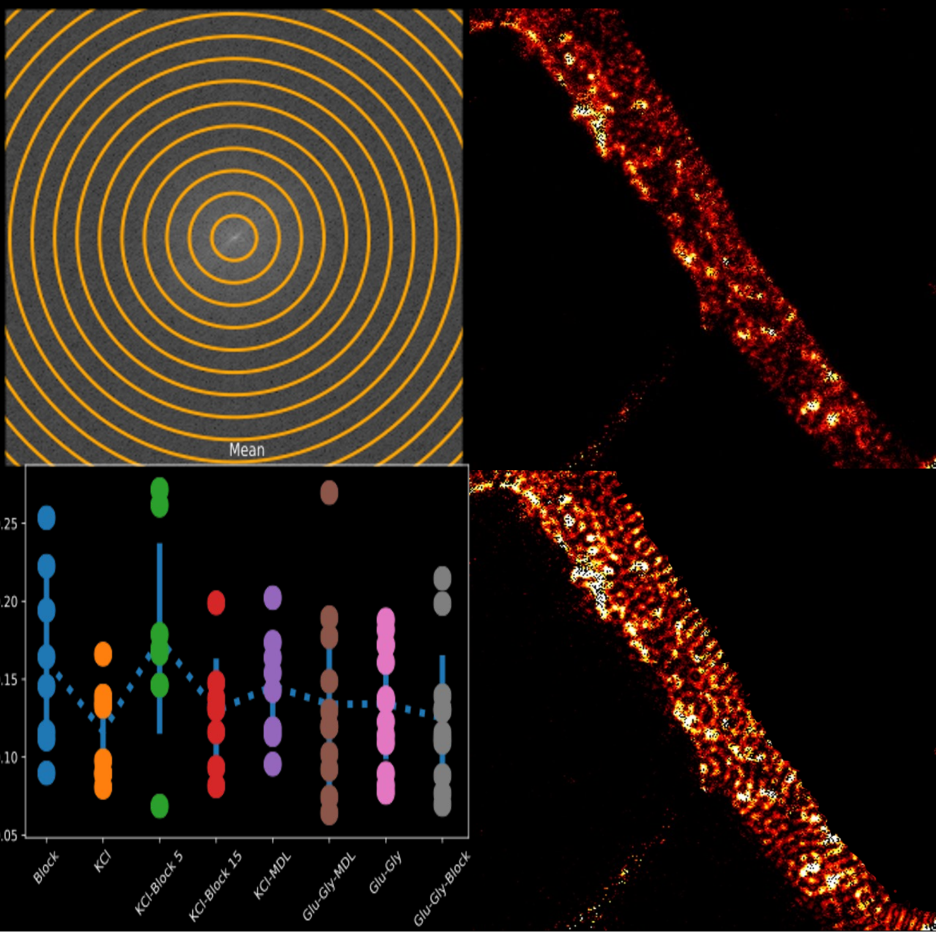
What to do **before**, during, and after starting a project involving deep learning with your data?

1. Curate your dataset
2. Define the task you want to address
3. Try to solve it with a well-established classical approach

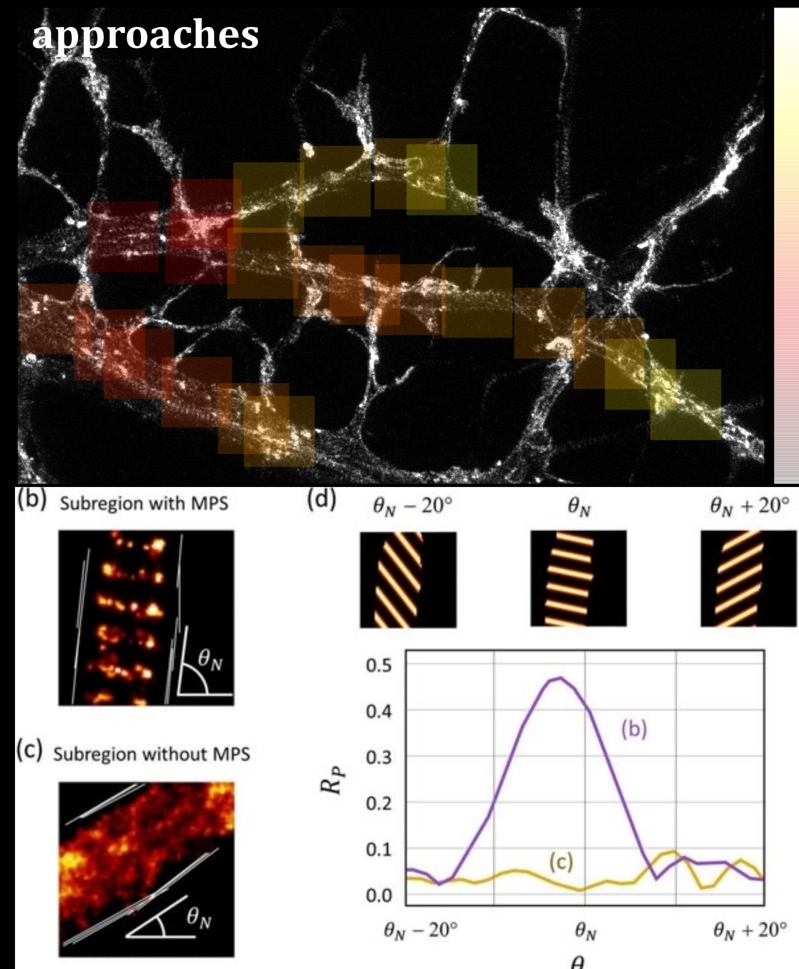


Try other approaches

Fourier analysis



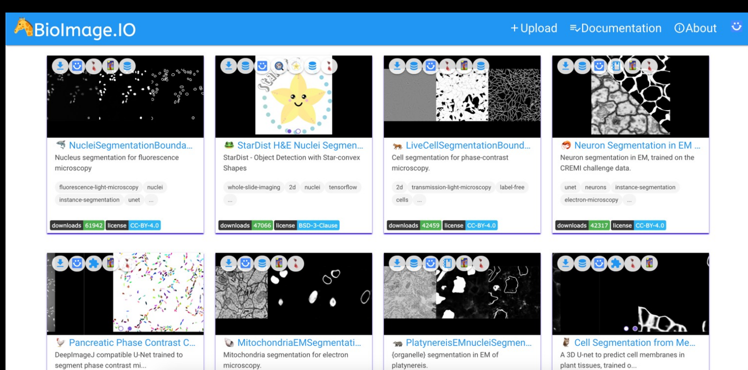
Correlative approaches



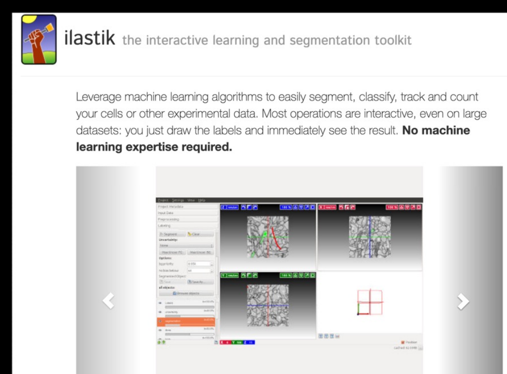
Barabas et al. *Scientific Reports*

What to do before, **during**, and after starting a project involving deep learning with your data?

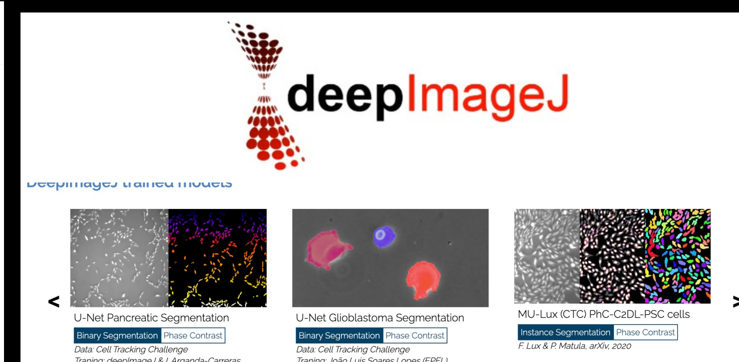
1. Curate your dataset
2. Define the task you want to address
3. Try to solve it with a well-established classical approach
4. Look online for already published user-friendly tools or pre-trained models



Bioimage Model Zoo



Ilastik



Deep ImageJ

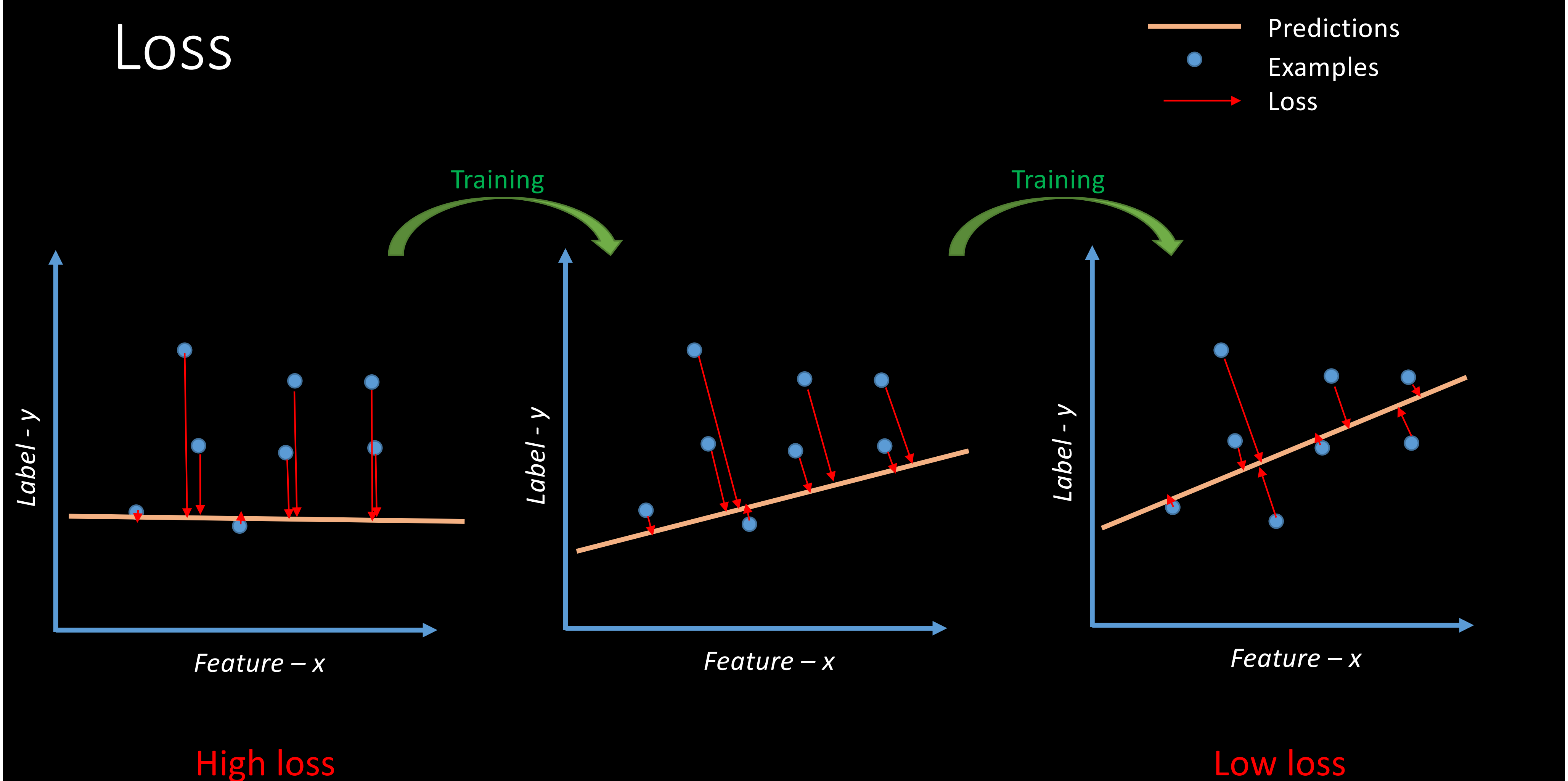
What to do before, **during**, and after starting a project involving deep learning with your data?

1. Curate your dataset
2. Define the task you want to address
3. Try to solve it with a well-established classical approach
4. Look online for already published user-friendly tools or pre-trained models
5. Refine the task and pipeline by evaluating :
   - The size of your dataset
   - The complexity of the task
   - The annotation requirements
   - Which performance metrics you can use

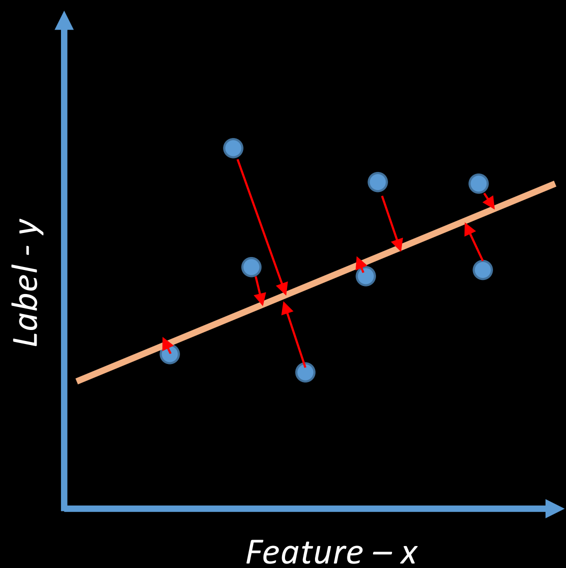
Training a model

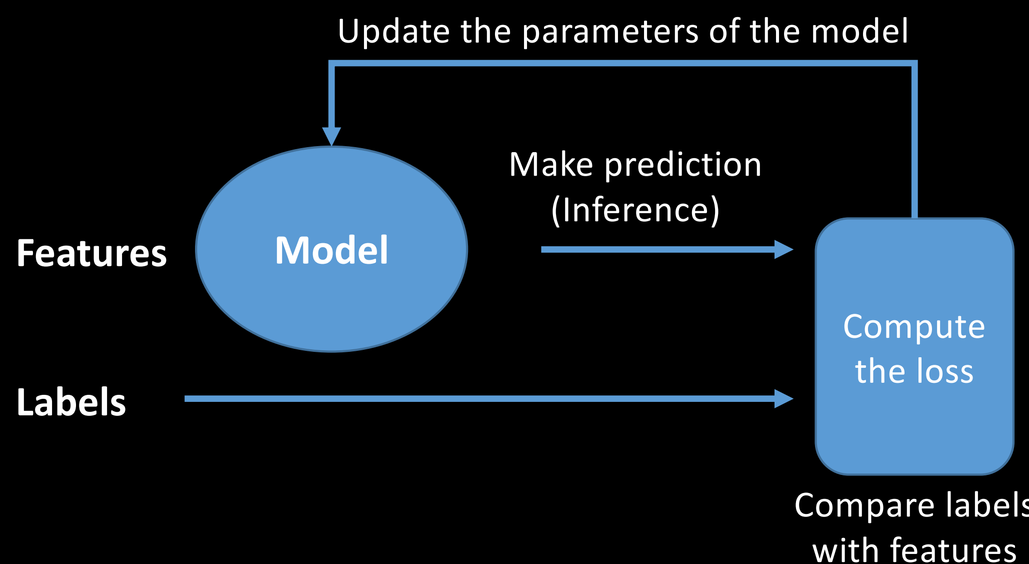
- Examinig many examples
- Build a model that will minimize the ***loss***
- ***Loss :*** How bad was a model to predict the features on a single example
  - Loss of 0 = model is perfect
  - High loss = model is bad
- ***Goal :*** find a set of weigths and biases that have low loss on ***all examples.***

Loss


Predictions
Examples
Loss
Training
Training
Label - y
Feature – x
Label - y
Feature – x
Label - y
Feature – x
High loss
Low loss

Loss





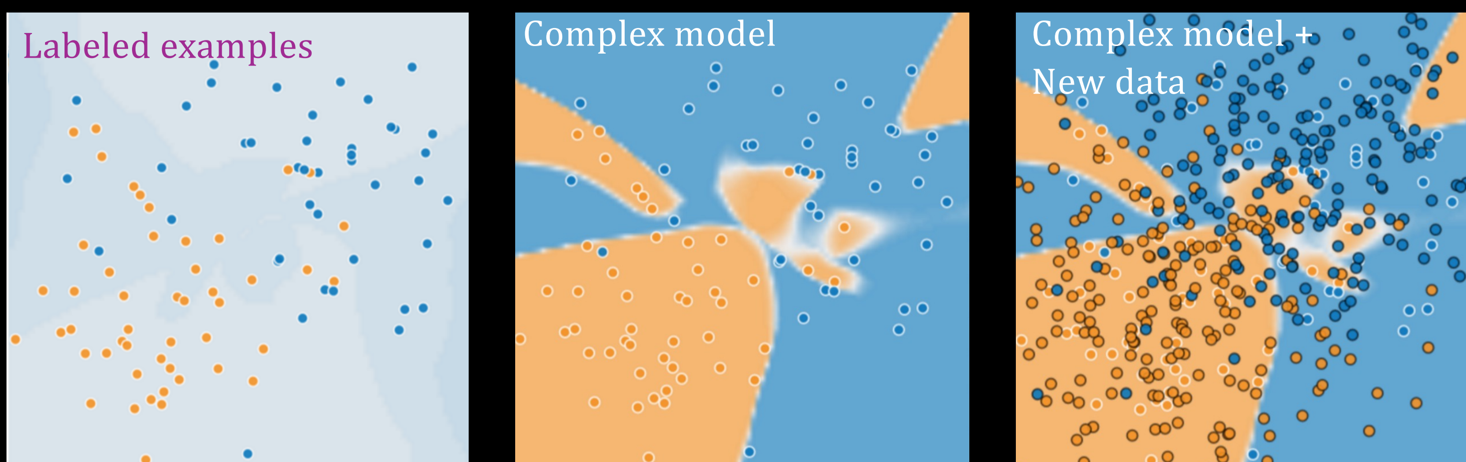
Predictions
Examples
Loss

Loss function examples :

- *Mean square error*
- *Cross entropy*

Generalization

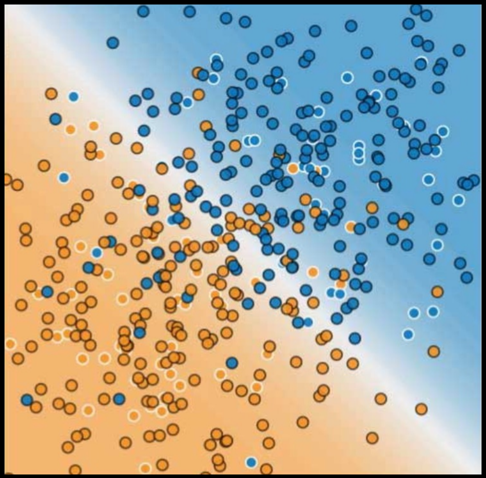
- In Machine Learning the goal is to learn a model that will give good prediction on new examples that were never seen
- A too complex model may perform well with the training data set but gives poor prediction on new examples : **overfitting**

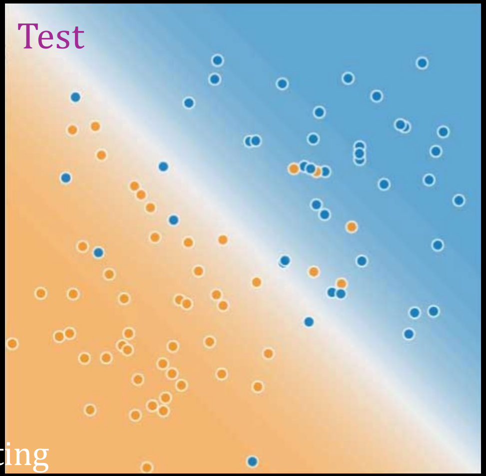


https://developers.google.com/machine-learning/crash-course/generalization/peril-of-overfitting

Generalization

- A model that overfits :
  - Low loss on the training dataset
  - High loss on the test dataset (new examples that were never seen before)
- Test set :
  - Large enough
    - But small partition of the dataset to maximize the number of examples used in training
    - Generally 15-25% of the whole dataset
  - Is not used over and over again
  - Representative from the whole dataset
- To prevent overfiting
  - *Try to have a model that is as simple as possible*

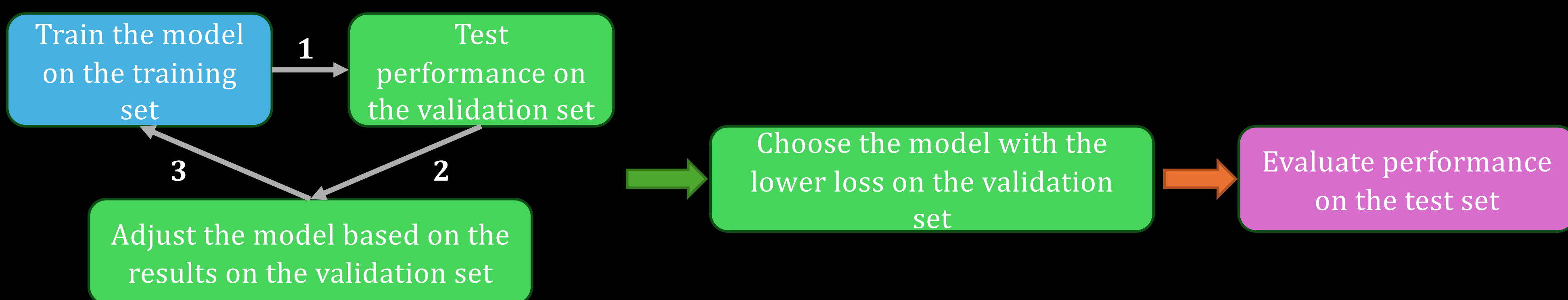




https://developers.google.com/machine-learning/crash-course/generalization/peril-of-overfitting

Generalization

- Adding a third partition : **Validation set**
- Greatly helps to reduce overfitting
  - Tweaking is done on the validation set
  - The test set is kept completely unknown to the model



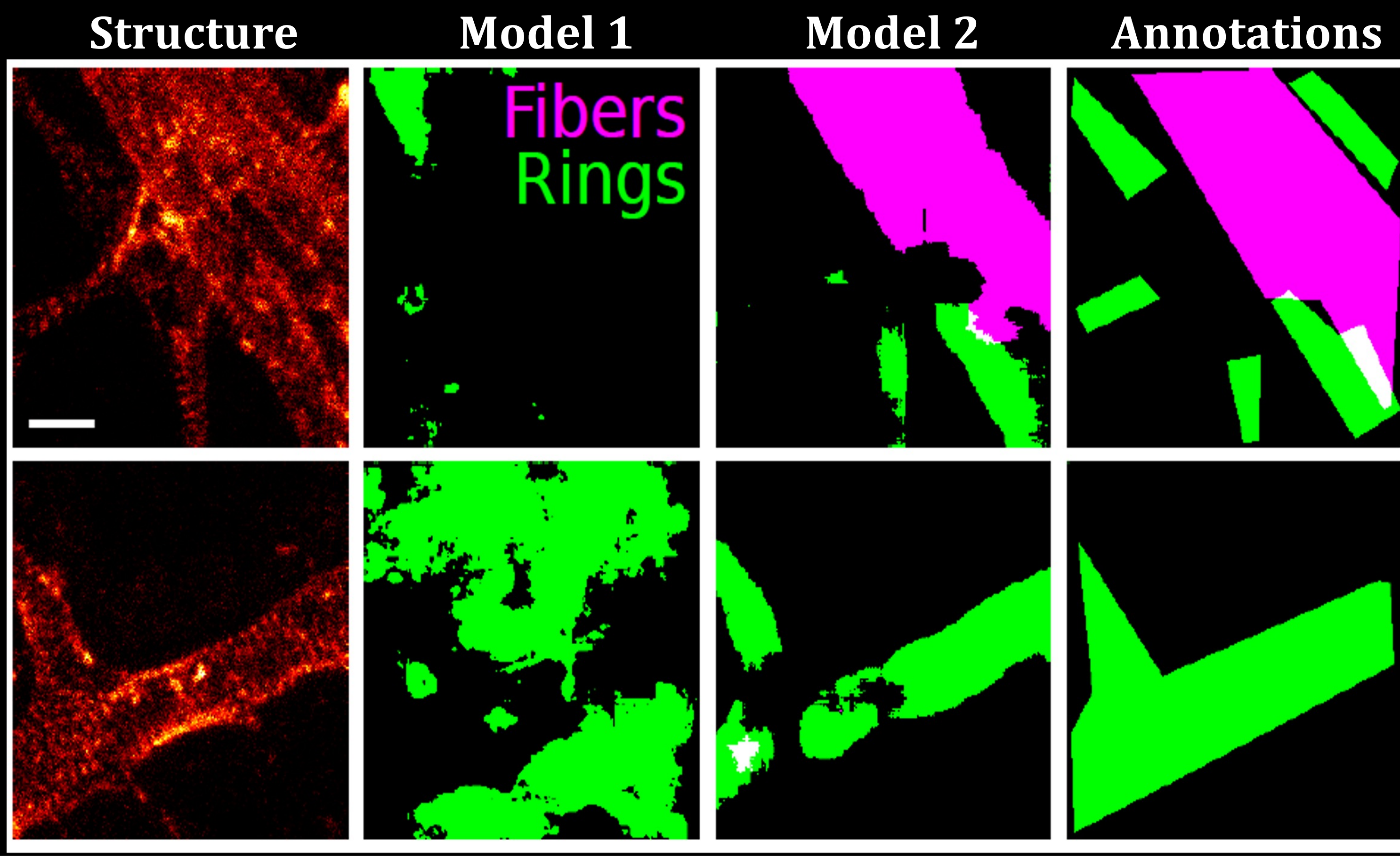
What to do before, **during**, and after starting a project involving deep learning with your data?

1. Curate your dataset
2. Define the task you want to address
3. Try to solve it with a well-established classical approach
4. Look online for already published user-friendly tools or pre-trained models
5. Refine the task and pipeline
6. **Collaborate** with computer scientist or image analysis specialist

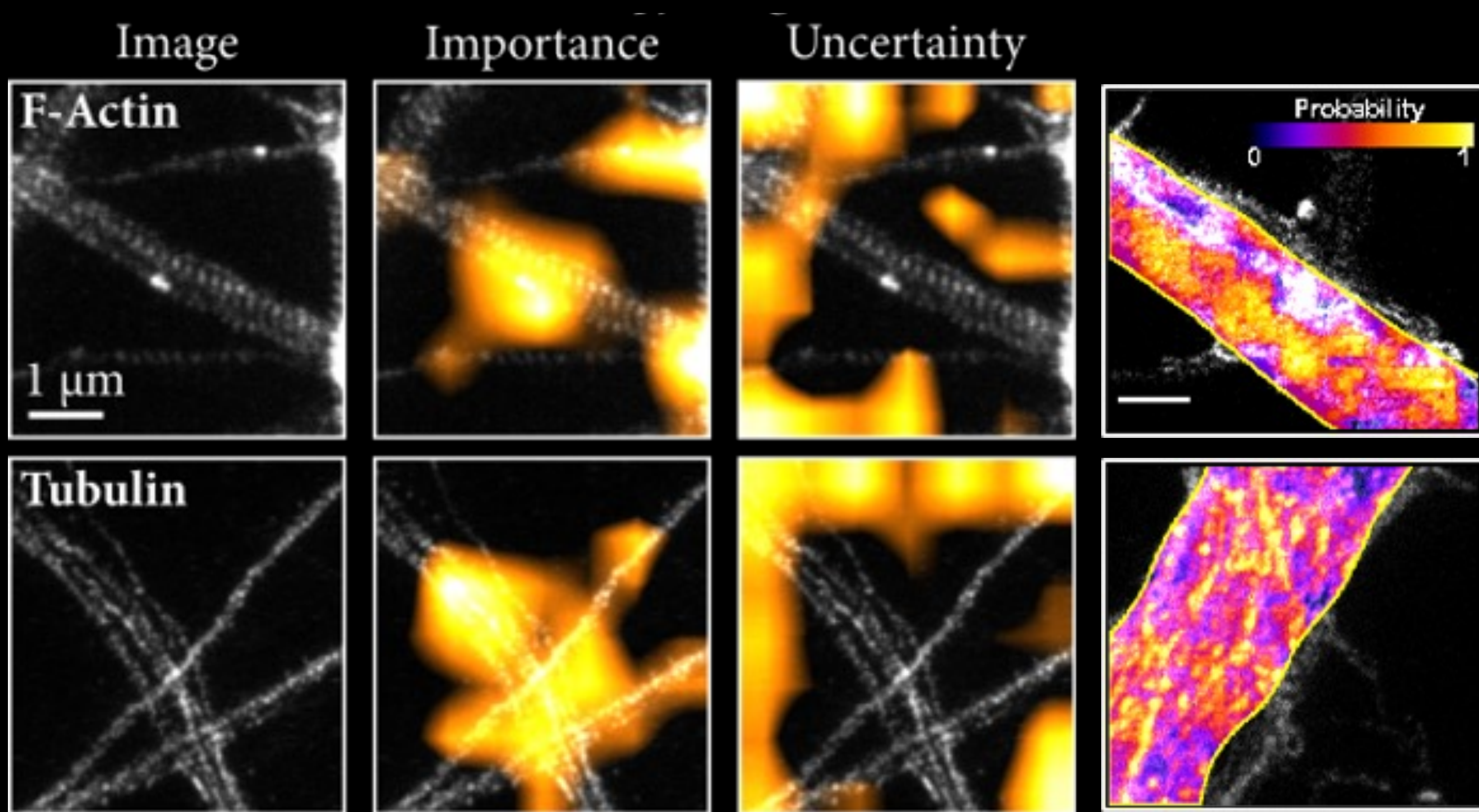
What to do before, during, and after starting a project involving deep learning with your data?

1. Curate your dataset
2. Define the task you want to address
3. Try to solve it with a well-established classical approach
4. Look online for already published user-friendly tools or pre-trained models
5. Refine the task and pipeline
6. **Collaborate** with computer scientist or image analysis specialist
7. **Inspect the results** (look at the images, the segmentation, the classification, etc.)

Inspect results



Inspect the results



Typical pipeline

Data acquisition

- Standardized practices
- Organize the data
- Keep the metada

Preprocessing

- Signal normalization
- Signal unmixing
- Dataset curation
- Alignment and registration
- Batch normalization
- Illumination correction

Processing

- Classification
- Segmentation
- Resolution Enhancement
- Artificial labeling
- Tracking
- Reconstruction

Data analysis

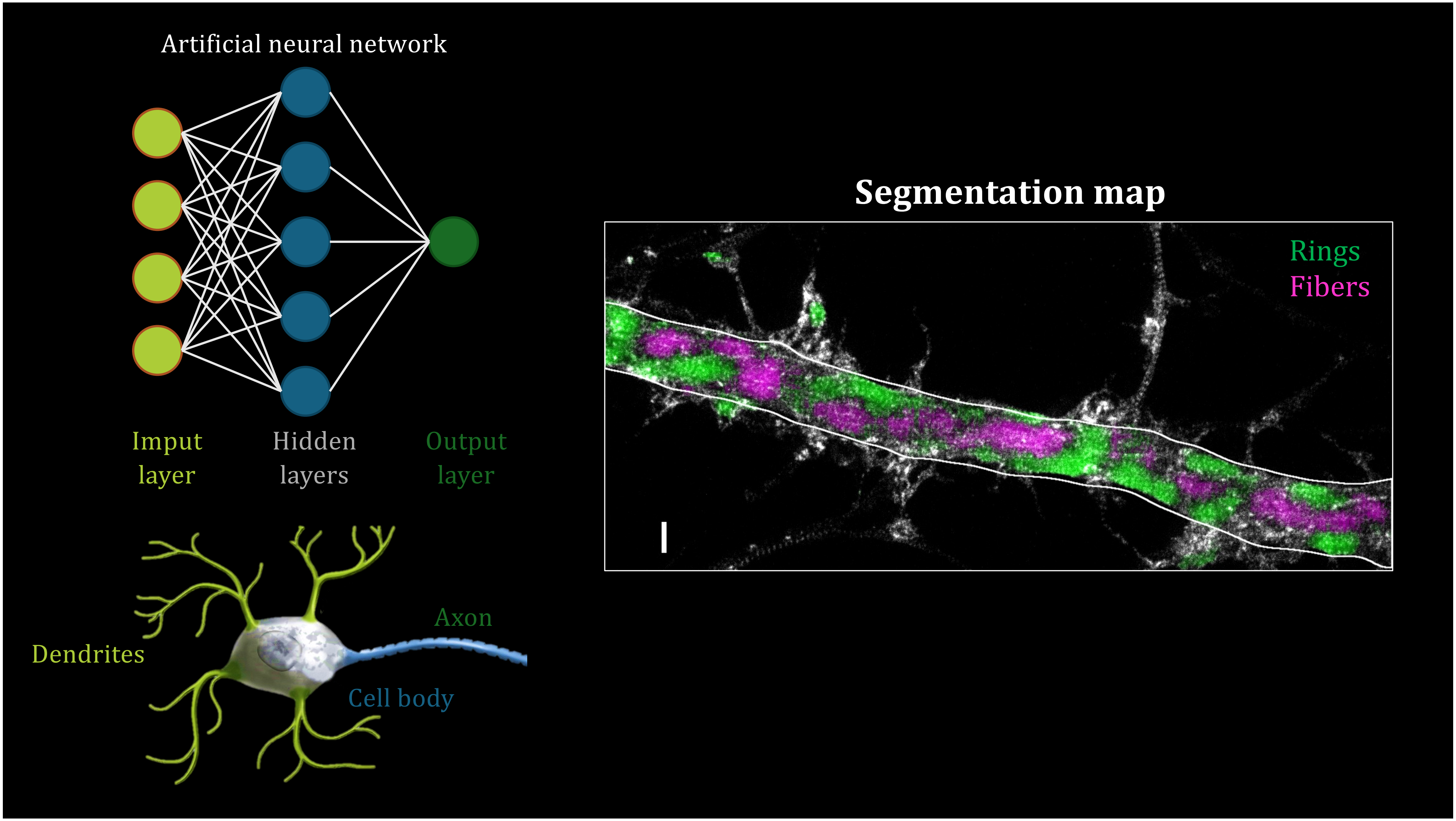
- Distributions
- Multidimensional analysis
- Feature analysis
- Clustering

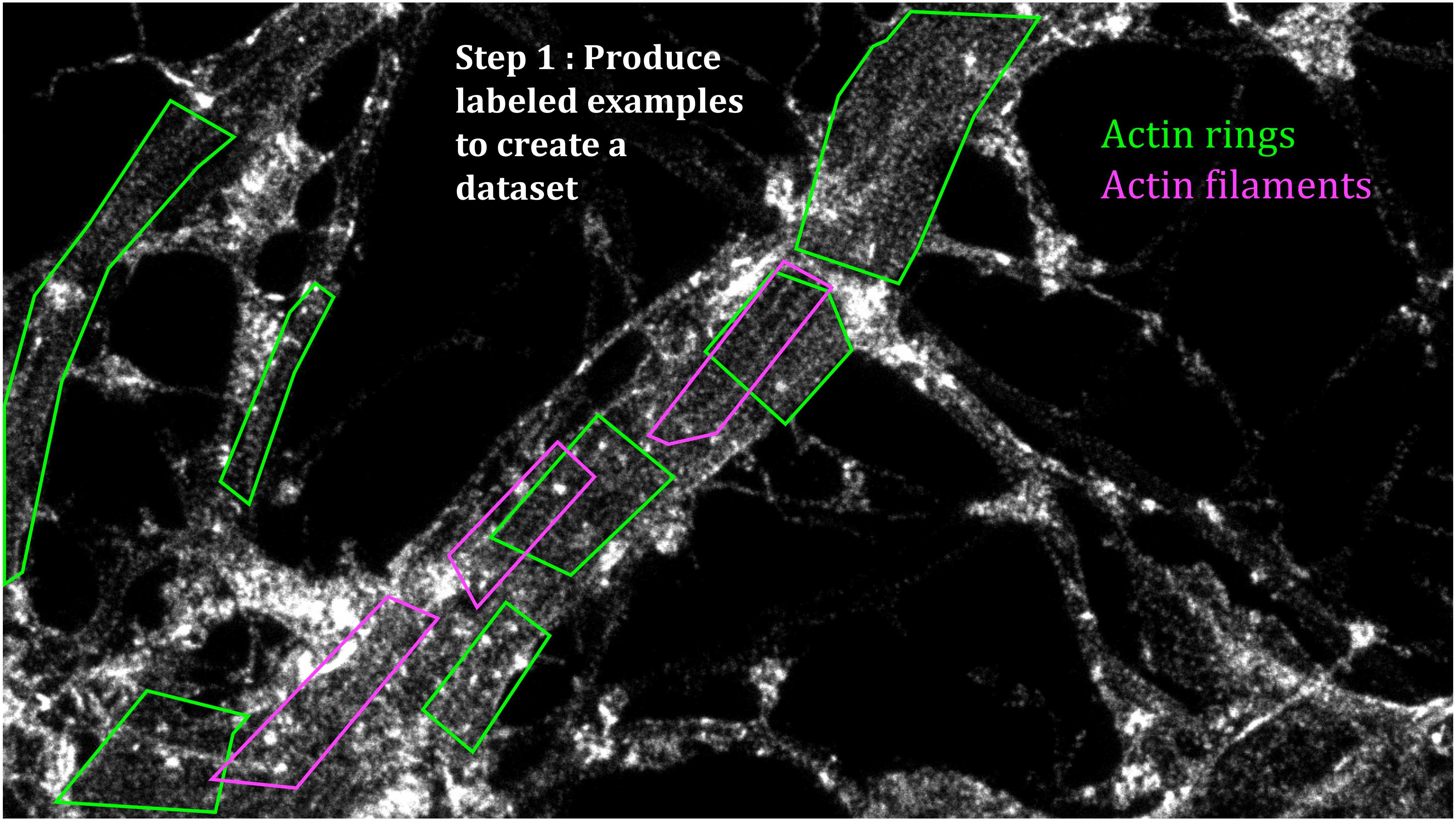
Visualization

- Dimensionality reduction
- Multidimensional representation
- Interactive tools
- Interpretability methods

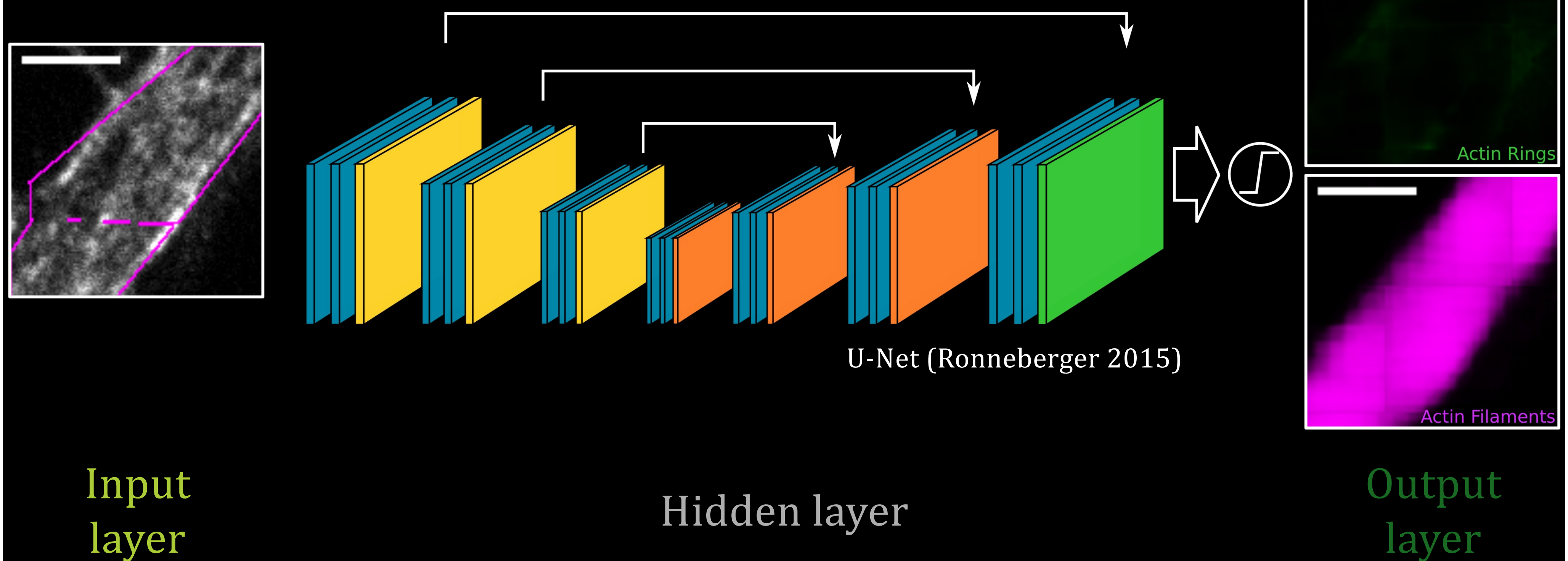
Sharing

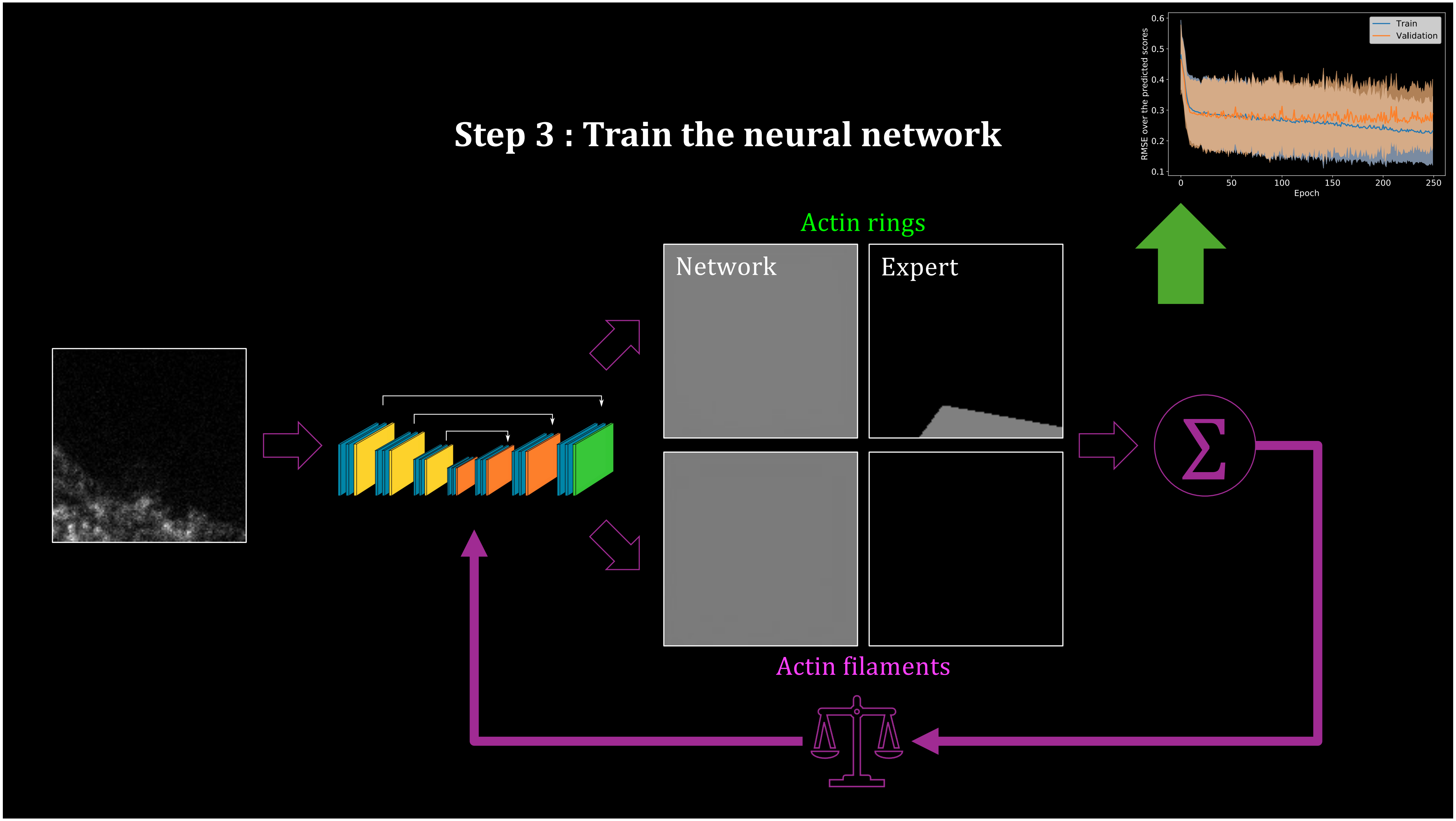
- Publish the dataset **with** metadata and annotations
- Open-access code (Github)
- Accessible model (public repository)
- Notebooks

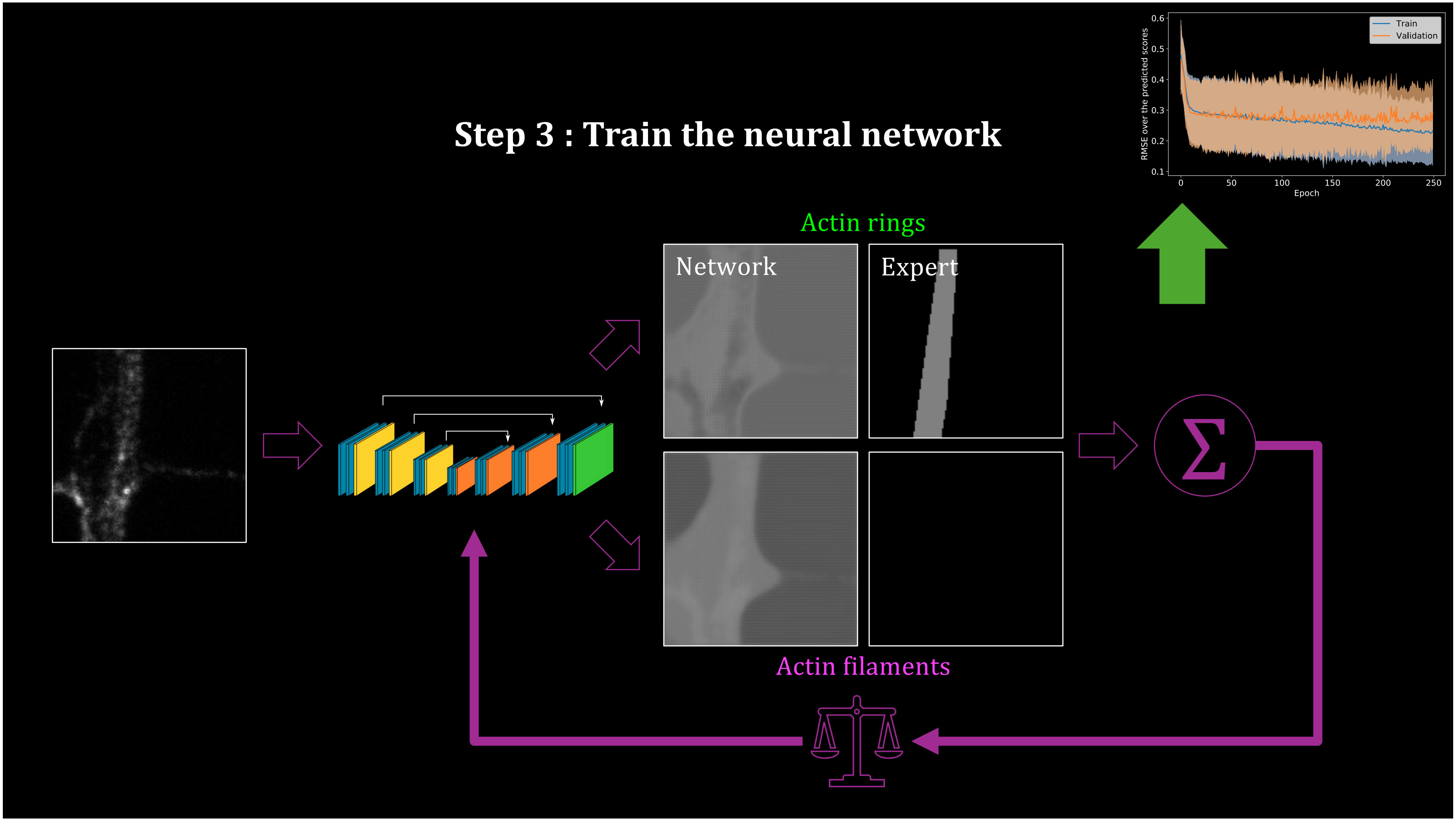

Artificial neural network
Imput layer
Hidden layers
Output layer
Dendrites
Axon
Cell body
Segmentation map
Rings
Fibers

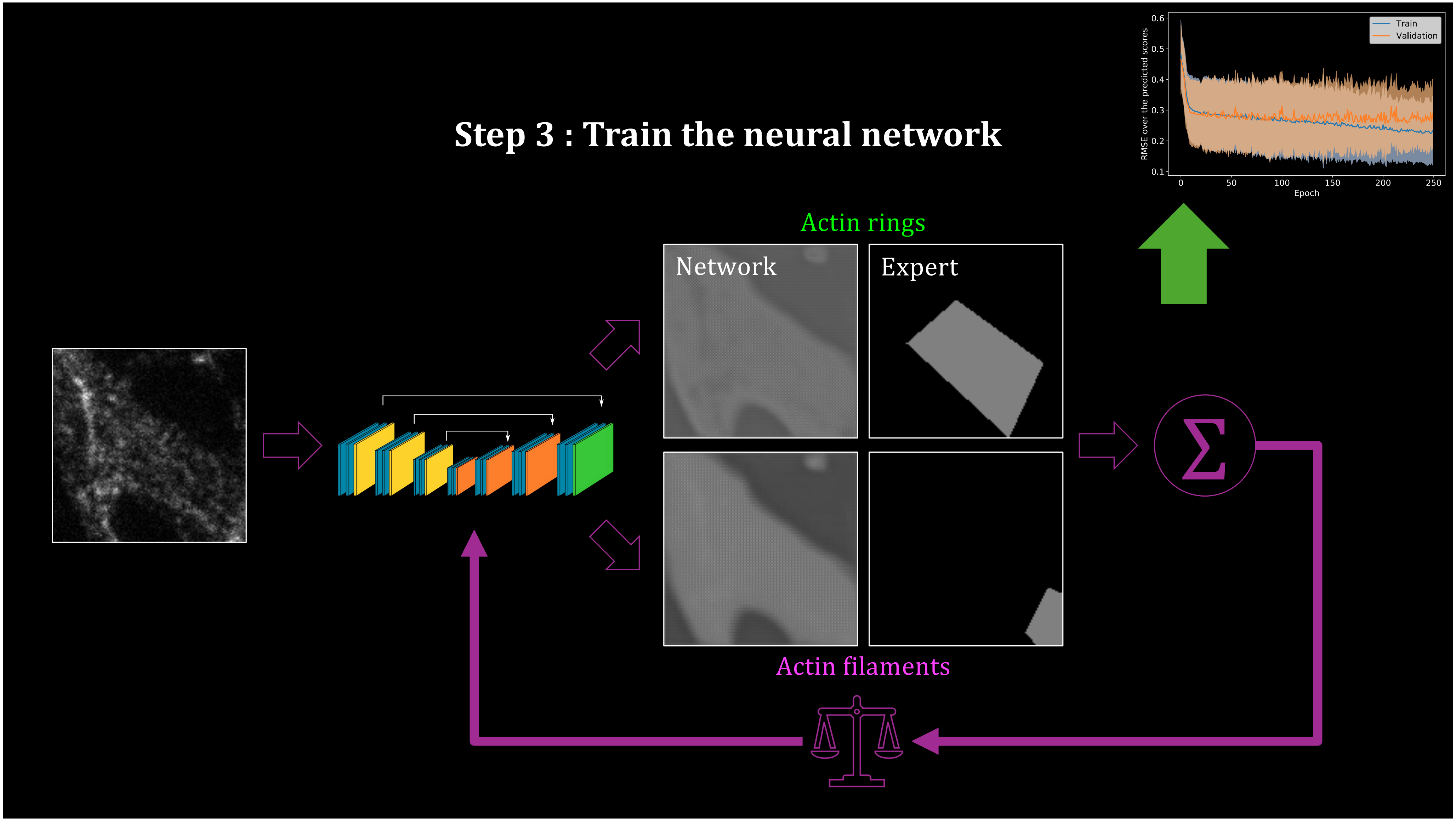

Step 1 : Produce labeled examples to create a dataset
Actin rings
Actin filaments

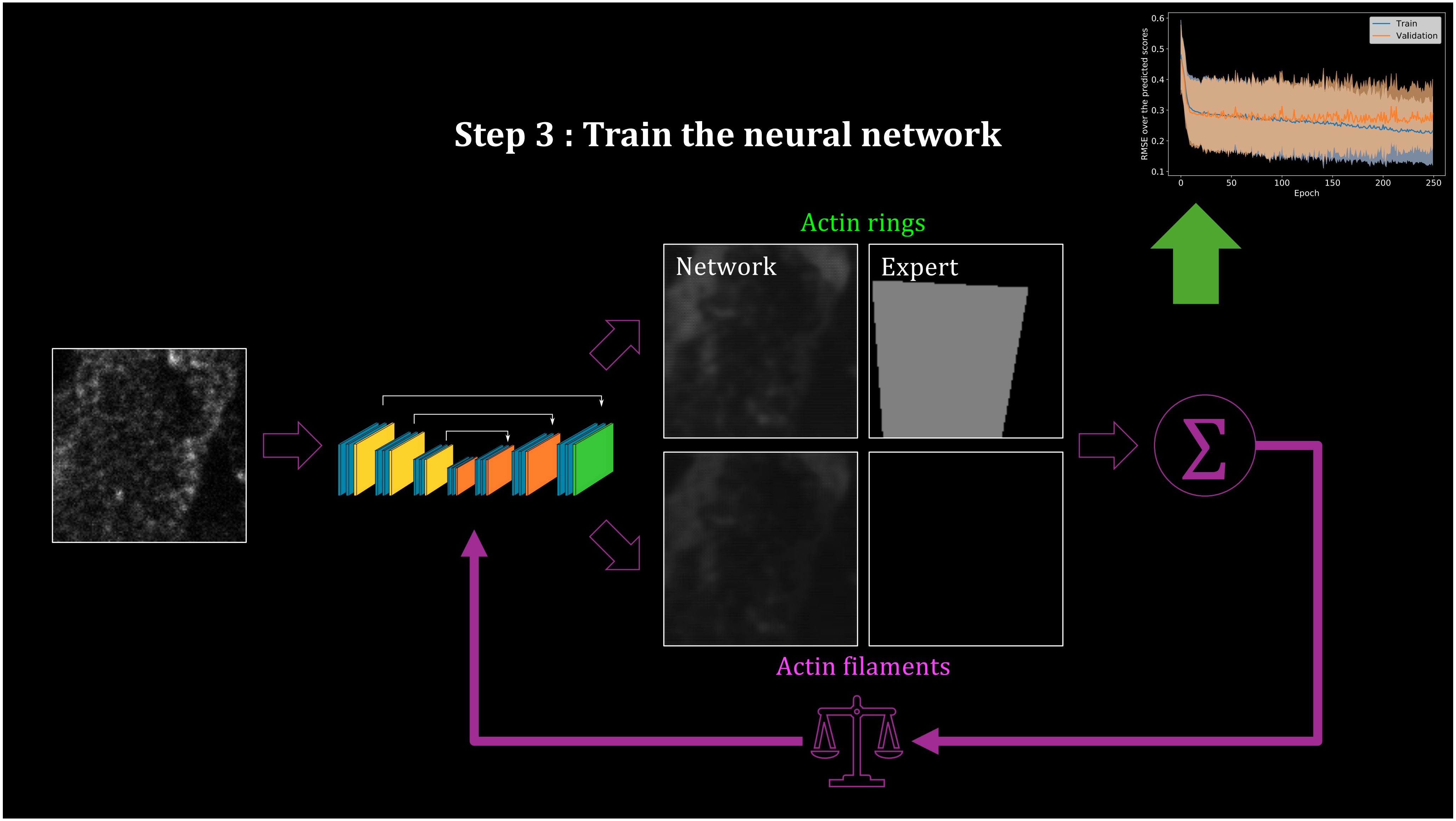
Step 2 : Design a suitable neural network

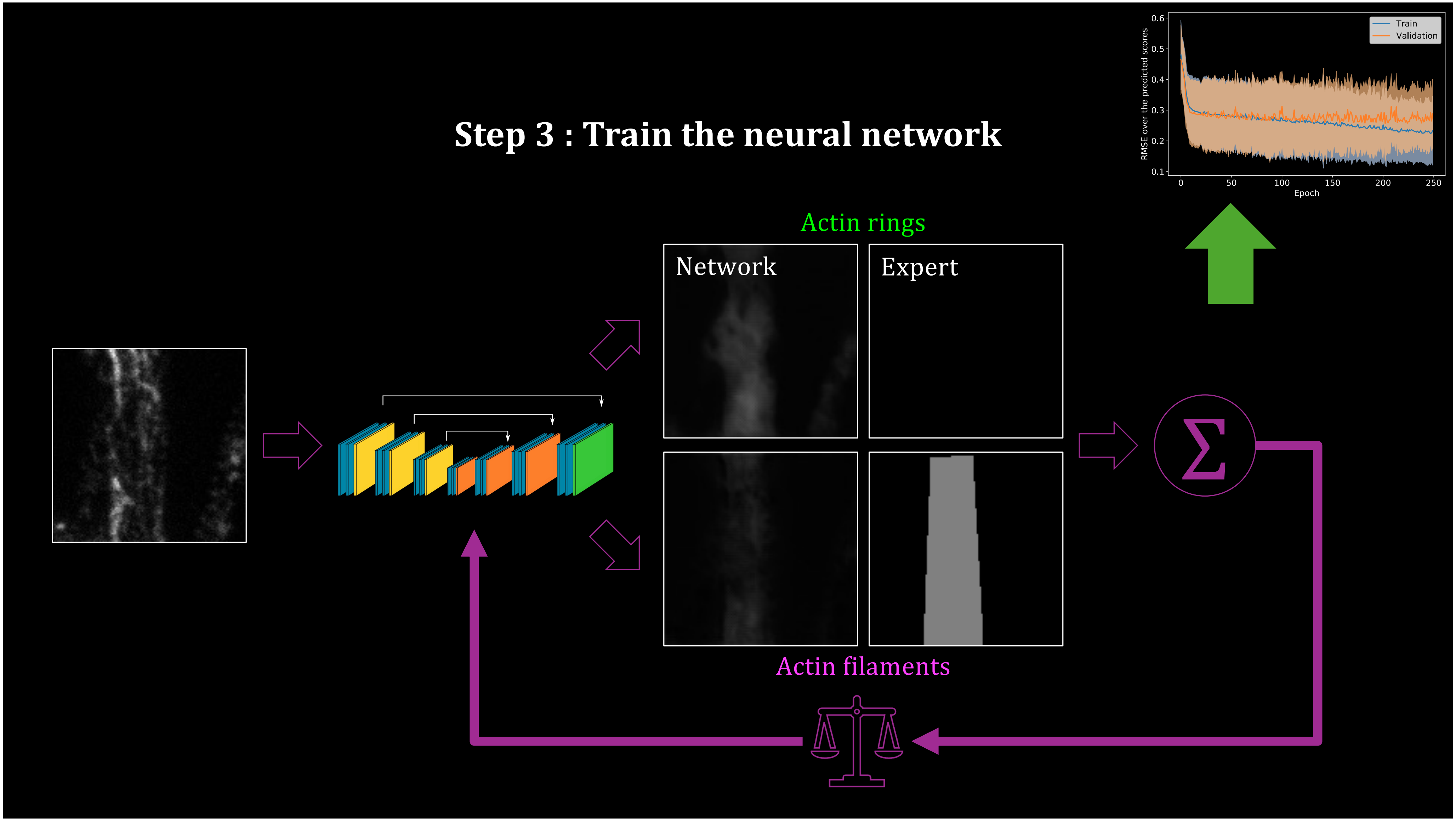


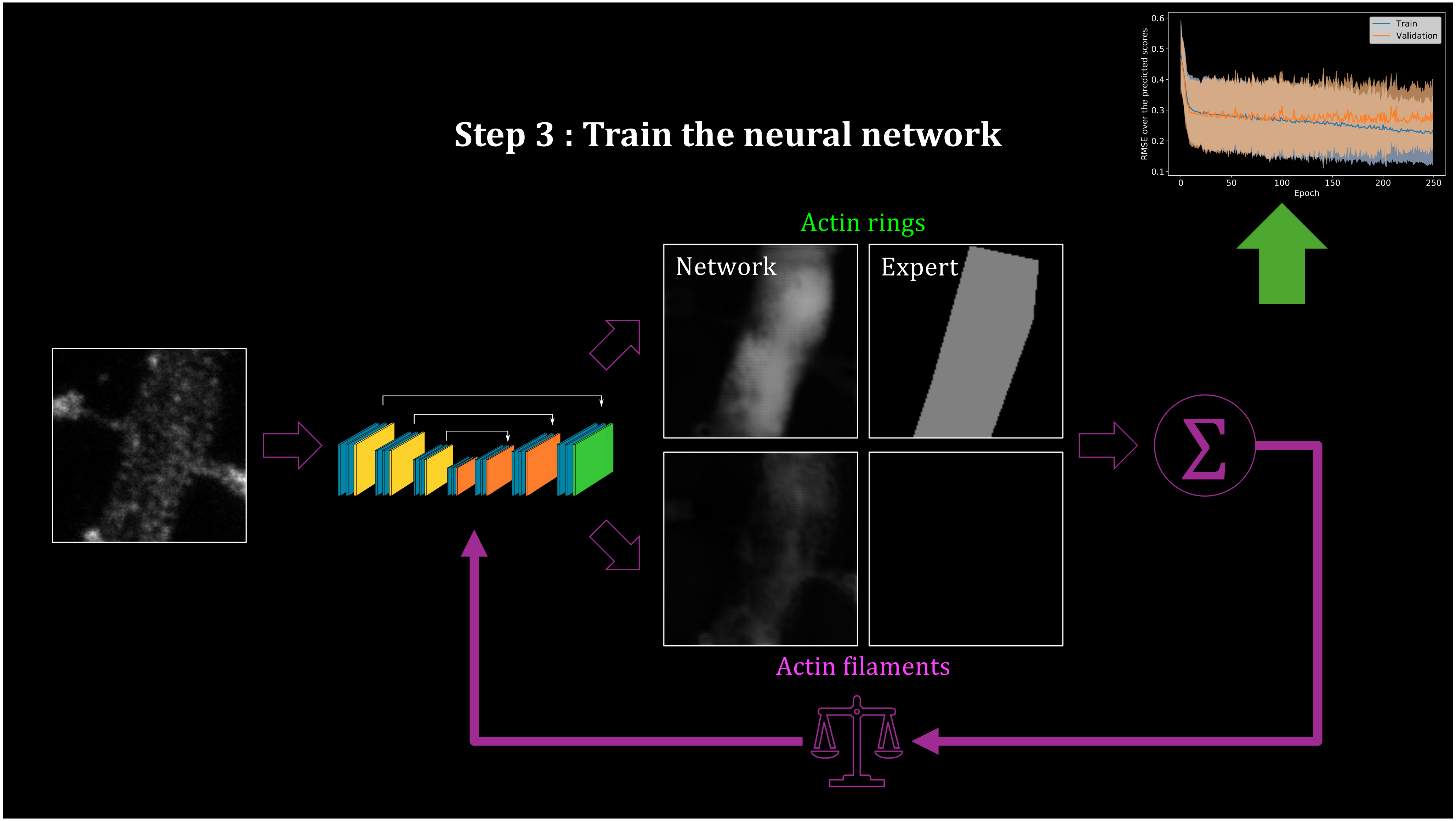

Step 3 : Train the neural network
Actin rings
Network
Expert
Actin filaments
Σ
Train
Validation
RMSE over the predicted scores
Epoch
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250

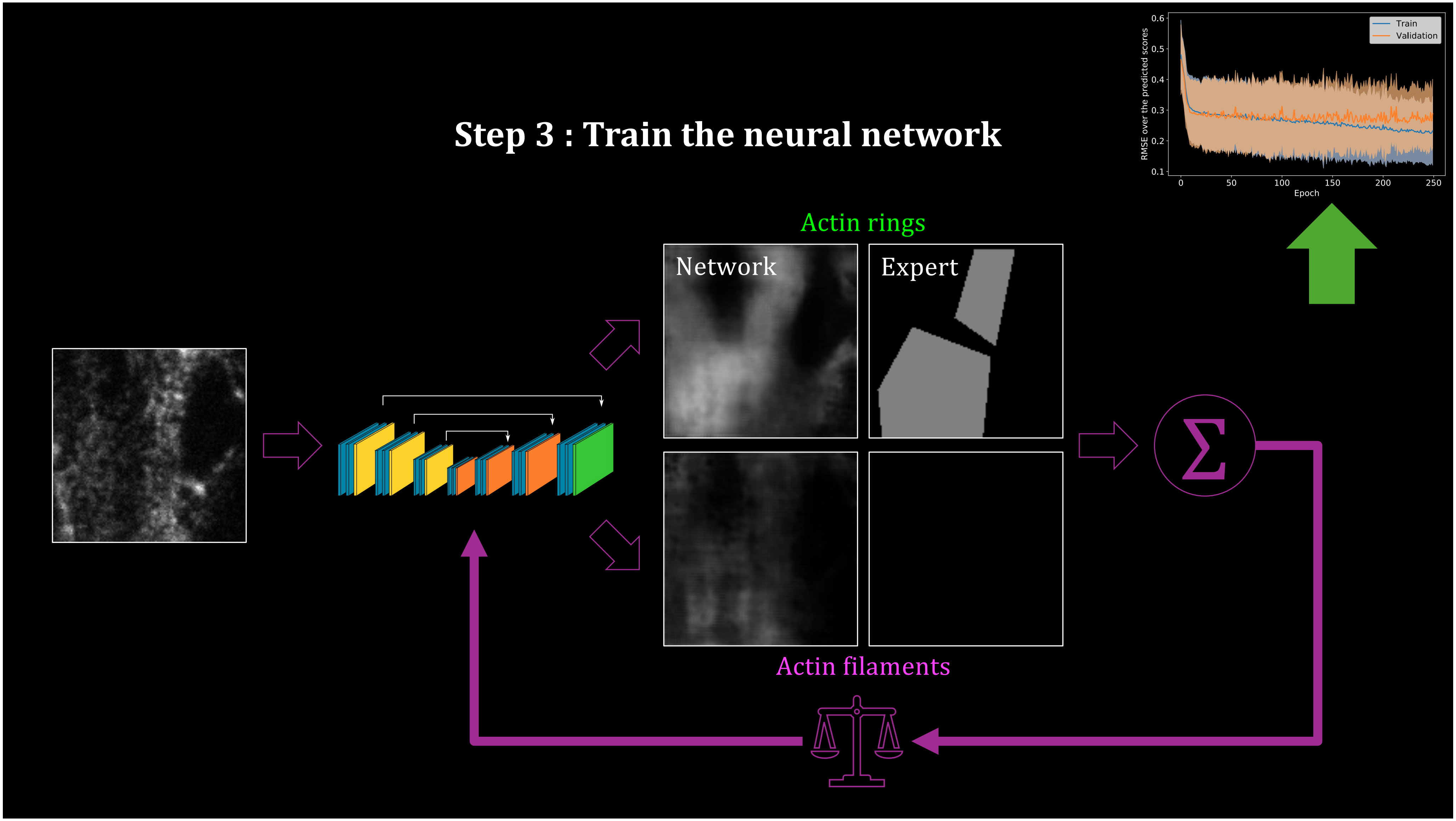

Step 3 : Train the neural network
Train
Validation
RMSE over the predicted scores
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
Epoch
Actin rings
Network
Expert
Σ
Actin filaments


Step 3 : Train the neural network
Actin rings
Network
Expert
Actin filaments
Σ
Train
Validation
RMSE over the predicted scores
Epoch
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250

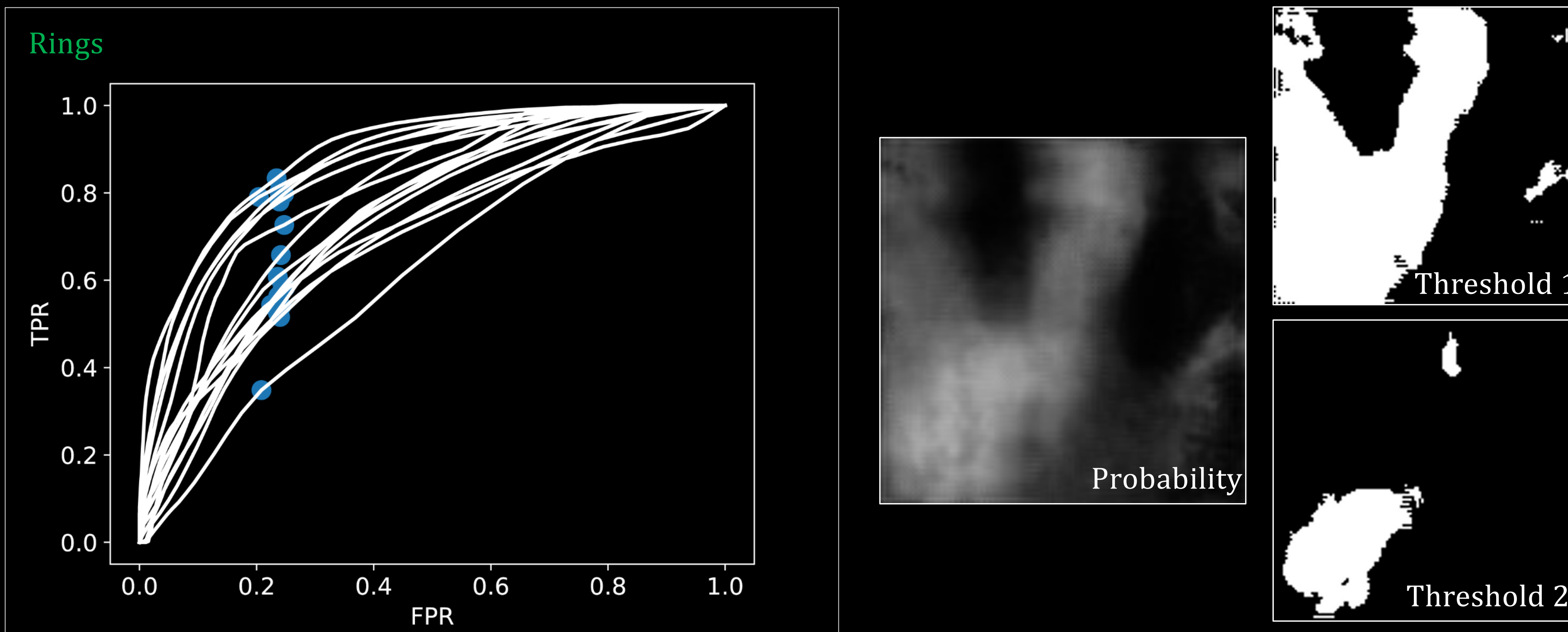

Step 3 : Train the neural network
Actin rings
Network
Expert
Actin filaments
Train
Validation
RMSE over the predicted scores
Epoch
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
Σ


Step 3 : Train the neural network
Actin rings
Network
Expert
Actin filaments
Σ
Train
Validation
RMSE over the predicted scores
Epoch
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250


Step 3 : Train the neural network
Train
Validation
RMSE over the predicted scores
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
Epoch
Actin rings
Network
Expert
Actin filaments
Σ

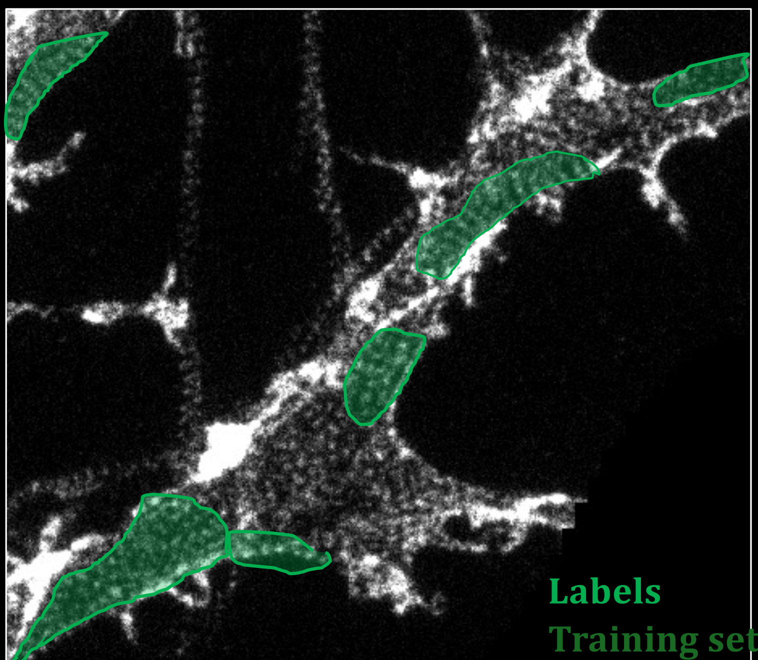

Step 3 : Train the neural network
Actin rings
Network
Expert
Actin filaments
Σ
Train
Validation
RMSE over the predicted scores
Epoch
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250

Step 4 : Determine the thresholds

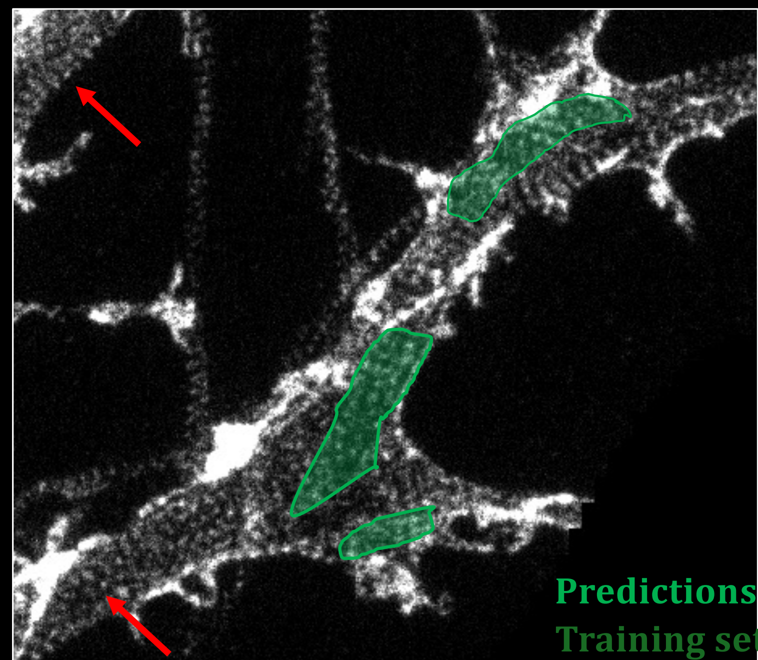


False positive rate of 0.25

Ground truth



Predictions

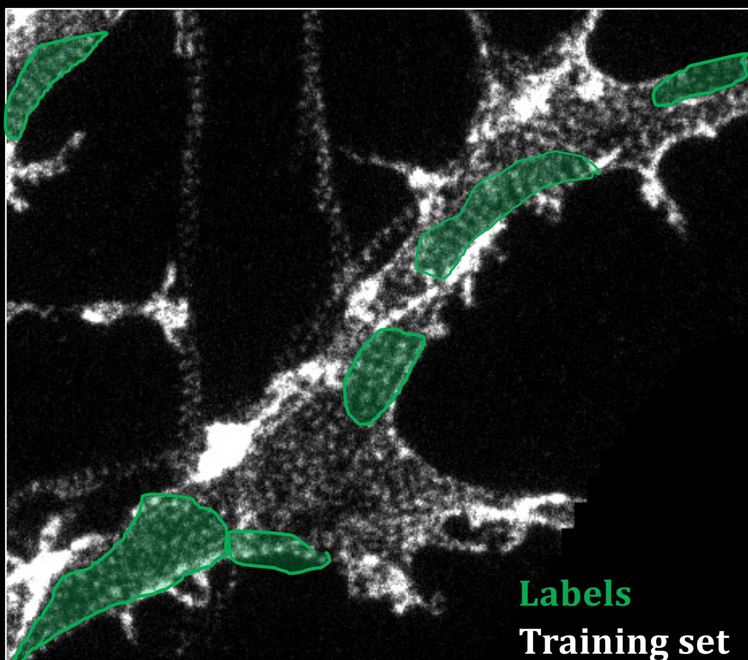


Ground truth and predictions are very different :

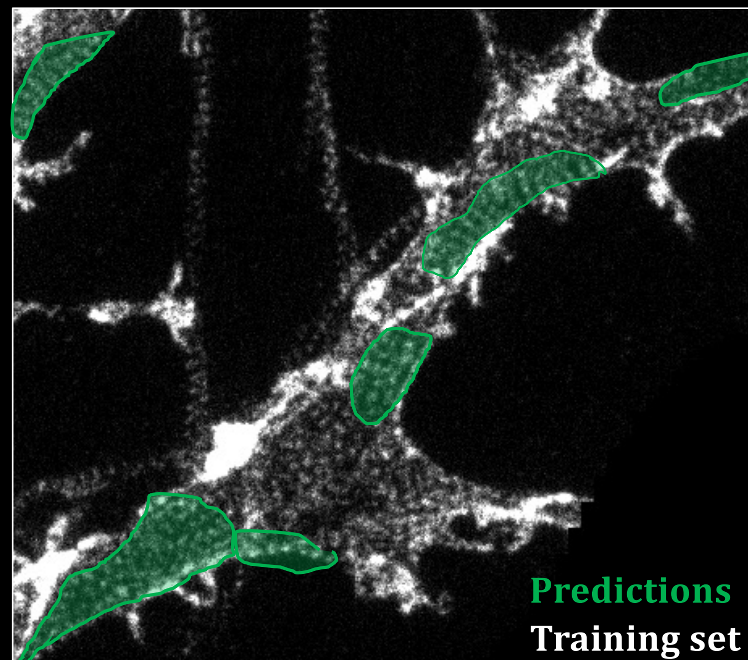
High loss

Model is not optimal

Ground truth



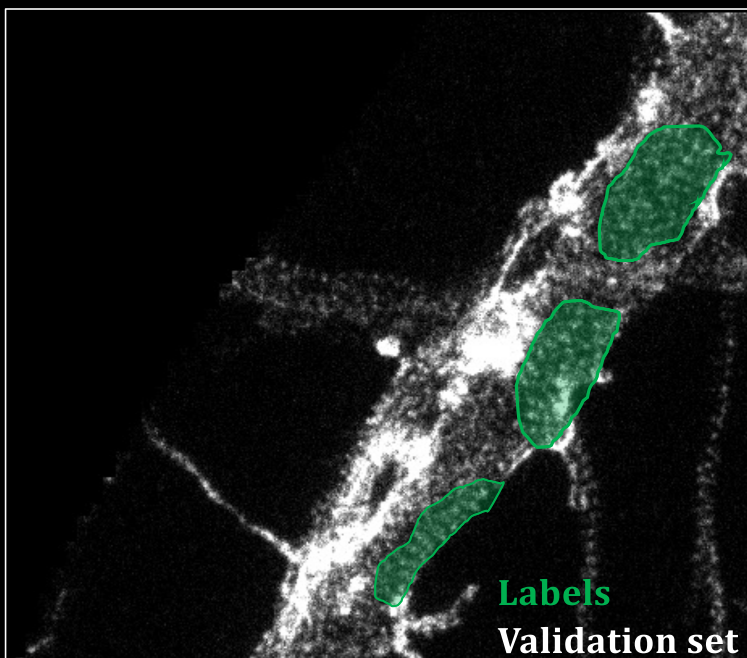
Predictions



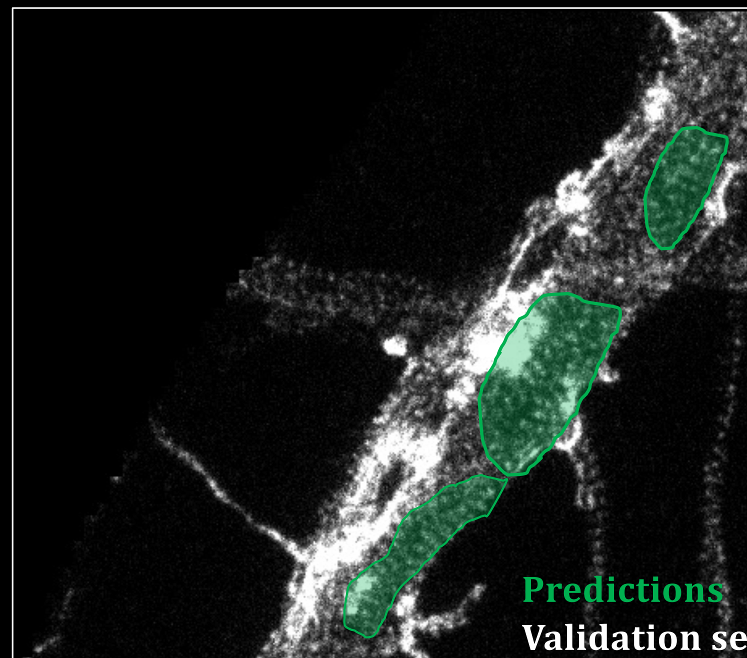
Ground truth and predictions are very almost the same: **Low loss**

Does the model generalizes well?

Ground truth



Predictions

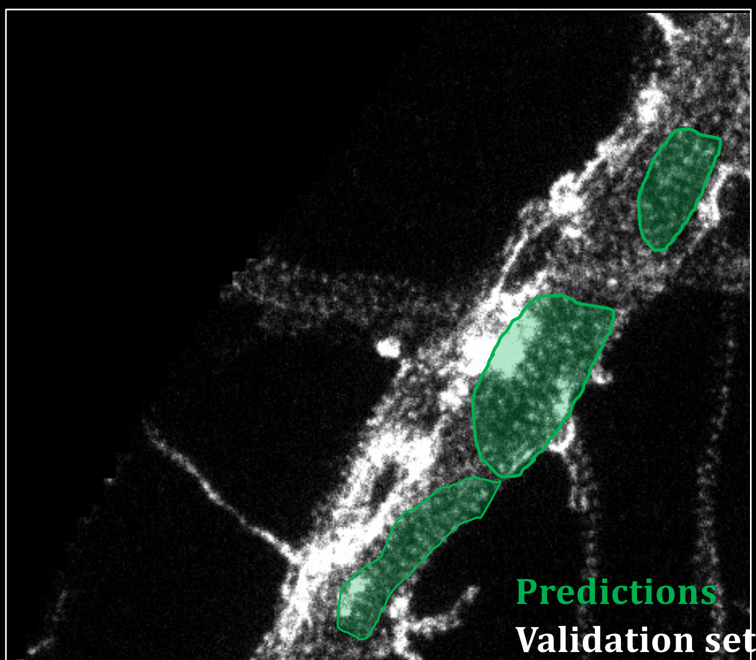


Ground truth and predictions are similar: **Loss was minimized**

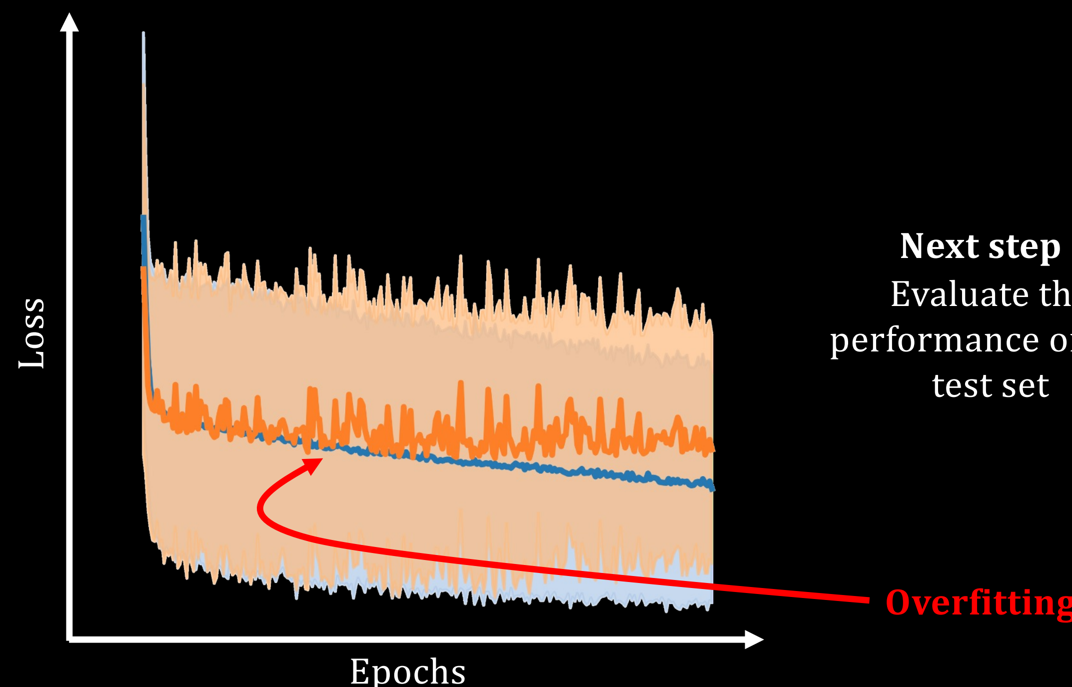
Choose the model based on the validation loss

Decide the model to use

Predictions



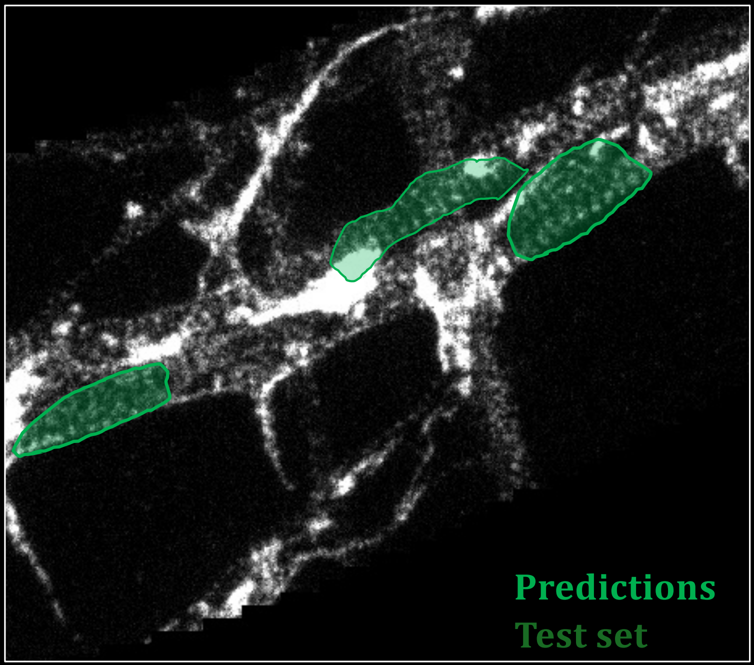
Training curve



Next step :
Evaluate the performance on the test set

Evaluate the performance on the test set

Predictions



The output is often a probability :

- Use it as it is
- Convert into binary (thresholding)

Classification problems :

- Accuracy
- Precision
- Recall

Pixel-wise classification = segmentation

- Decision threshold
  - Map a logistic regression value (*probability*) into a binary value (*yes/no, class1/class2*)
- **True positive :**
  - Label = Actin rings
  - Prediction = Actin rings
- **False positive :**
  - Label = Background
  - Prediction = Actin rings
- **False negative :**
  - Label = Actin rings
  - Prediction = Background
- **True negative :**
  - Label = Background
  - Prediction = Actin rings

Confusion Matrix

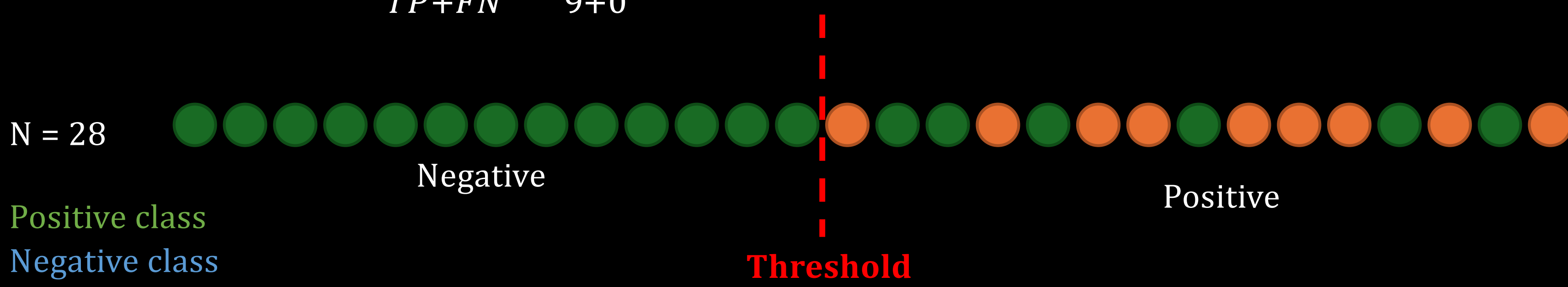
Labels \ Predictions	Background	Actin rings
Background	**True Negative**	**False Positive**
Actin rings	**False Negative**	**True Positive**

Classification

- Accuracy :
  - fraction of right predictions
  - $Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ predictions} = \frac{TP+TN}{TP+TN+FP+FN}$
- Precision :
  - proportion of positive predictions that were right
  - $Precision = \frac{TP}{TP+FP}$
- Recall :
  - proportion of positive labels that were identified correctly
  - $Recall = \frac{TP}{TP+FN}$

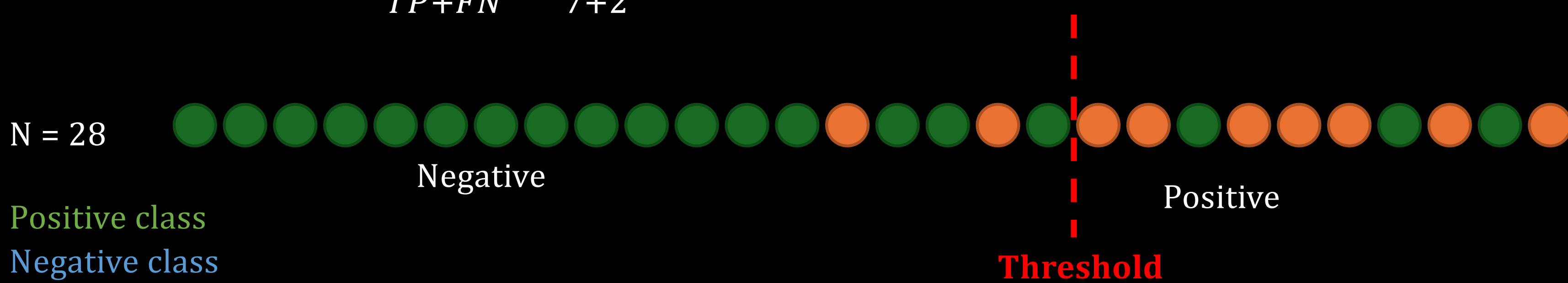
Classification

- Accuracy : fraction of right predictions
  - $Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ predictions} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{9+13}{9+13+6+0} = 0.79$
- Precision : proportion of positive predictions that were right
  - $Precision = \frac{TP}{TP+FP} = \frac{9}{9+6} = 0.6$
- Recall : proportion of positive labels that were identified correctly
  - $Recall = \frac{TP}{TP+FN} = \frac{9}{9+0} = 1.0$



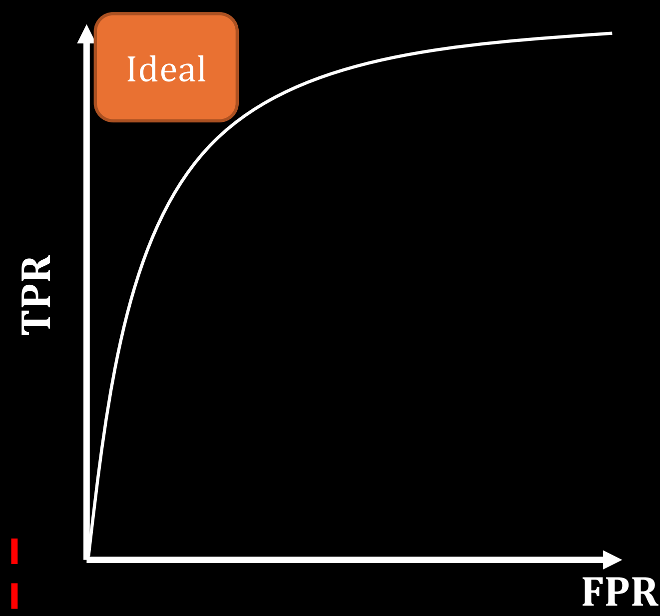
Classification

- Accuracy : fraction of right predictions
  - $Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ predictions} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{7+16}{7+16+3+2} = 0.82$
- Precision : proportion of positive predictions that were right
  - $Precision = \frac{TP}{TP+FP} = \frac{7}{7+3} = 0.7$
- Recall : proportion of positive labels that were identified correctly
  - $Recall = \frac{TP}{TP+FN} = \frac{7}{7+2} = 0.78$

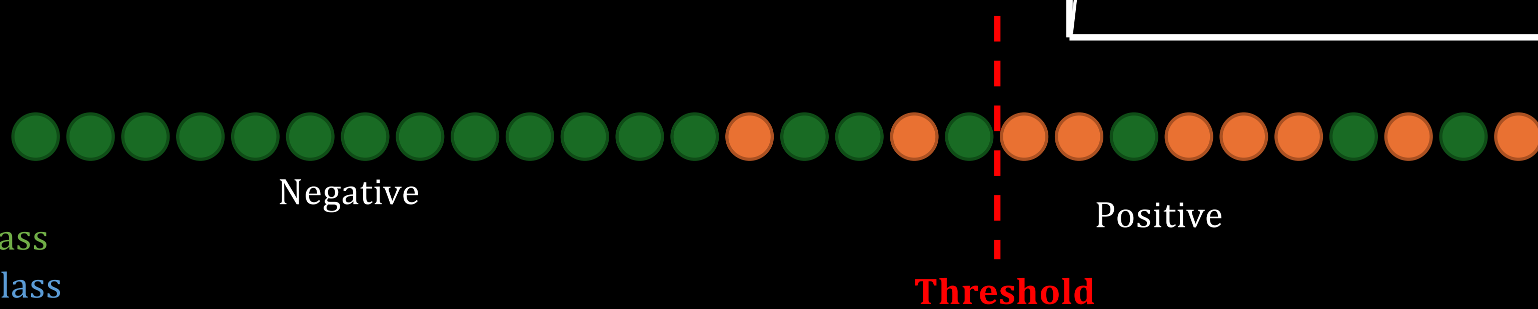


Classification – ROC curve

- ROC : receiver operating curve
  - Plot true positive rate (TPR) against the false positive rate (FPR)
  - $TPR = Recall = \frac{TP}{TP+FN} = \frac{7}{7+2} = 0.78$
  - $FPR = \frac{FP}{FP+TN} = \frac{3}{3+16} = 0.16$

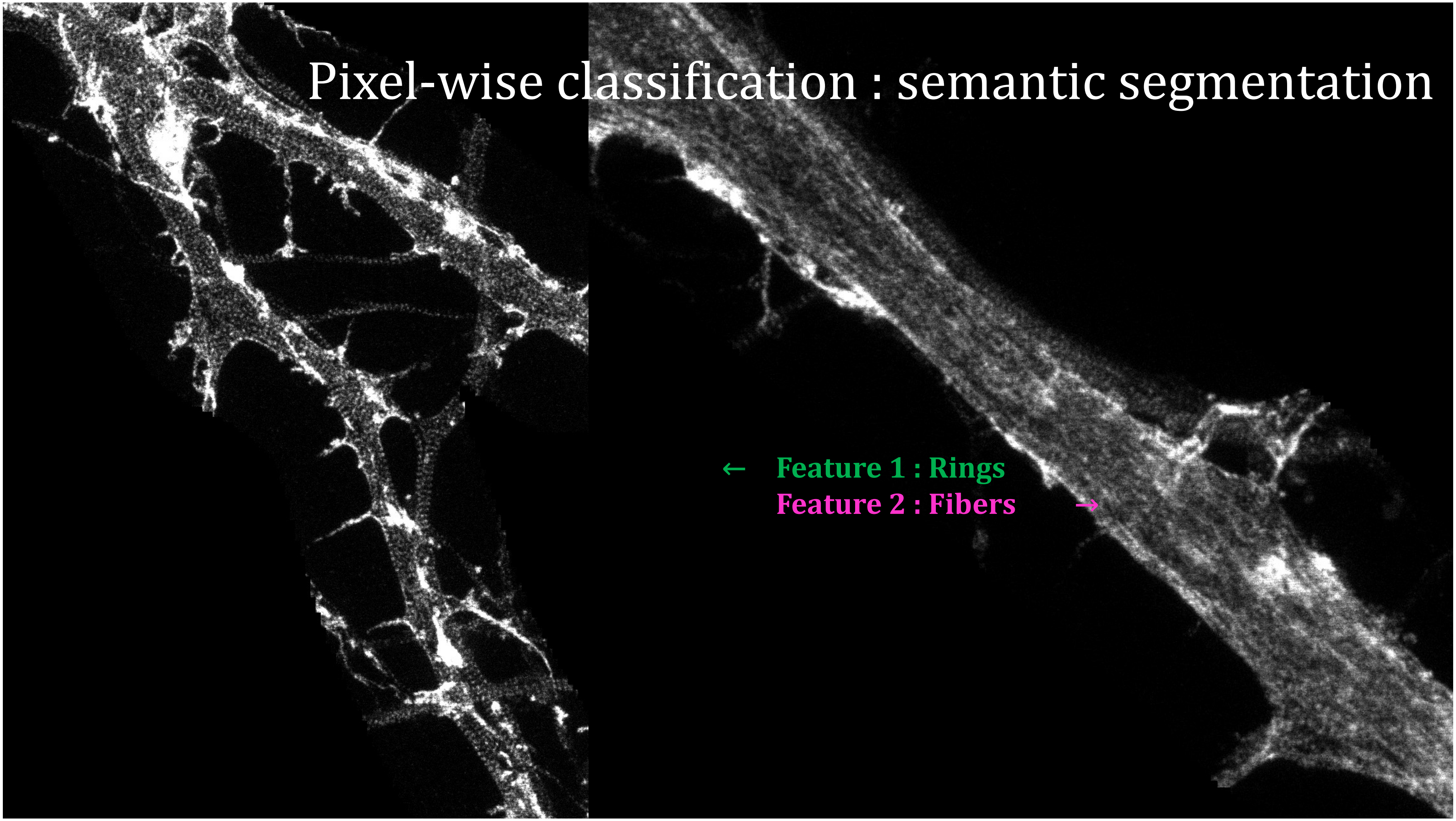


N = 28

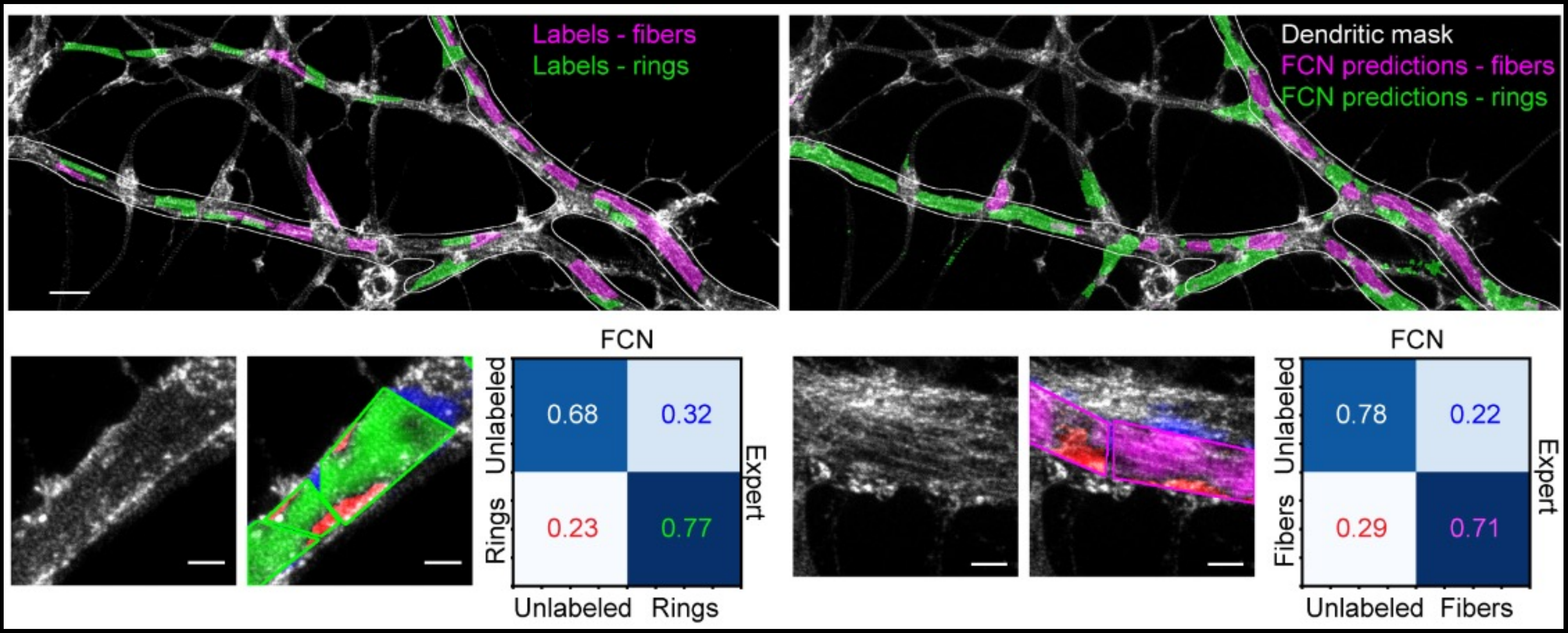


Positive class

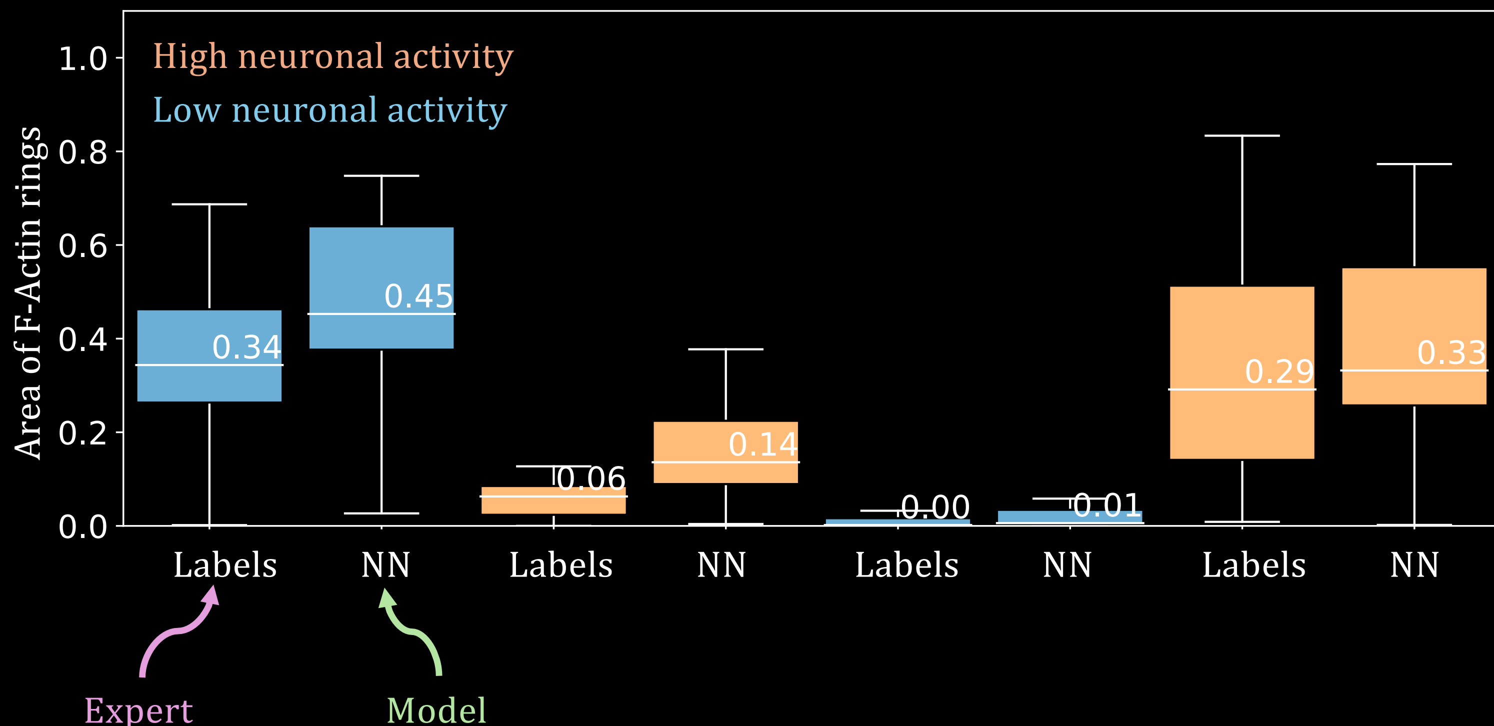
Negative class

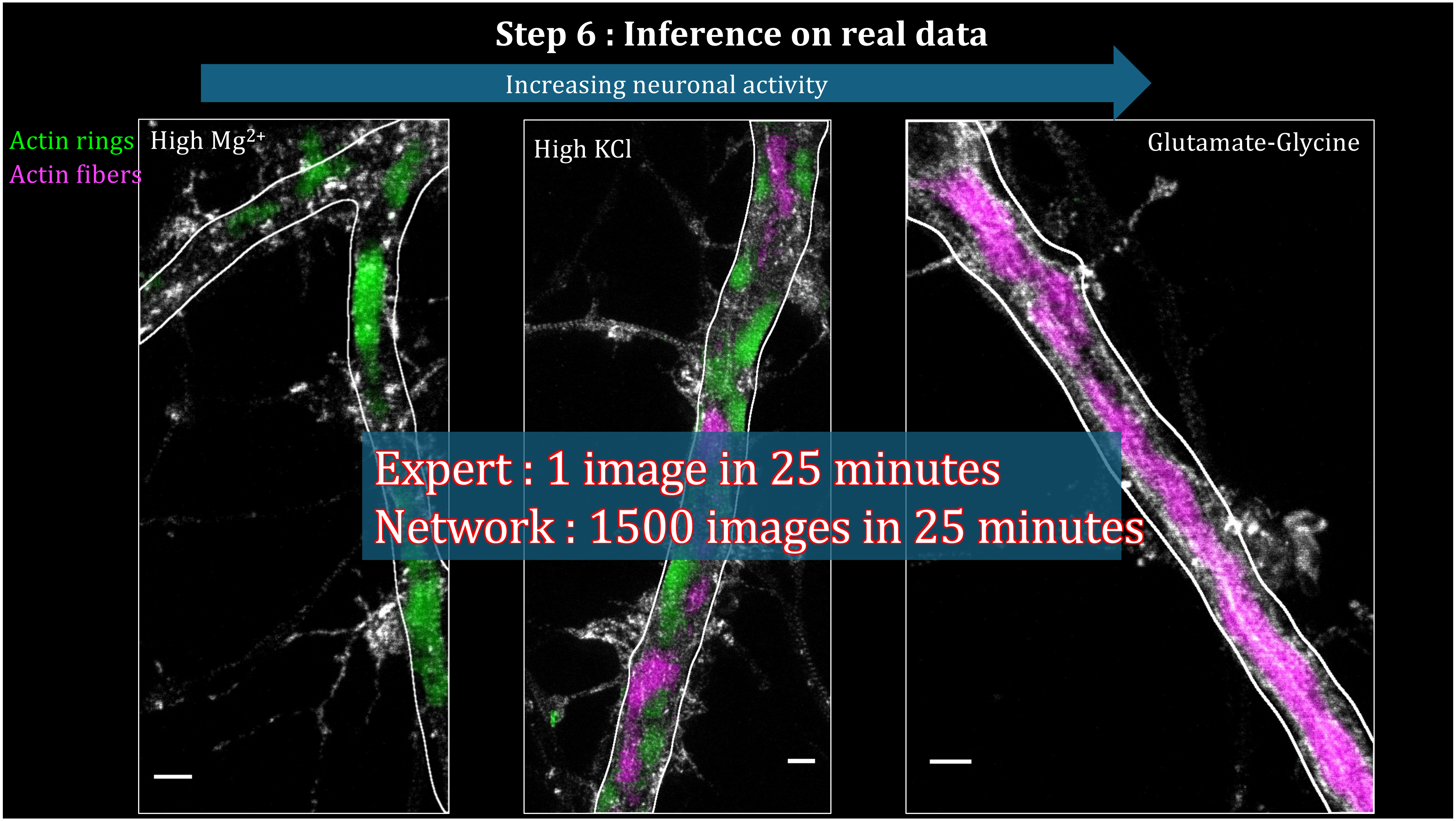

Pixel-wise classification : semantic segmentation
← Feature 1 : Rings
Feature 2 : Fibers →

Step 5 : Evaluate performance on the test set



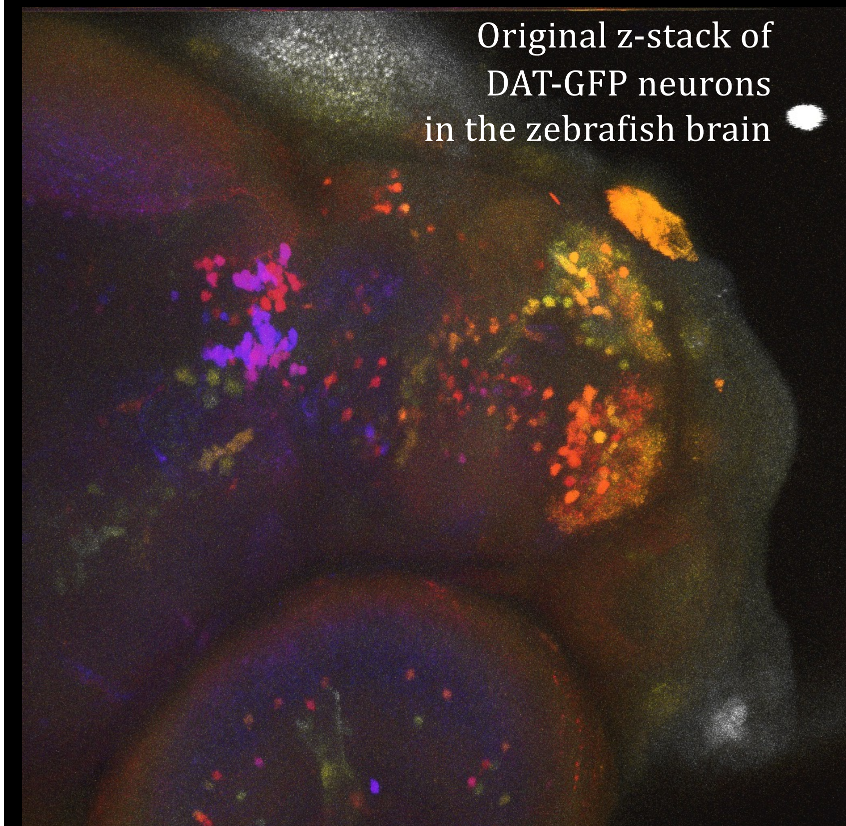
Step 5 : Evaluate performance on the test set



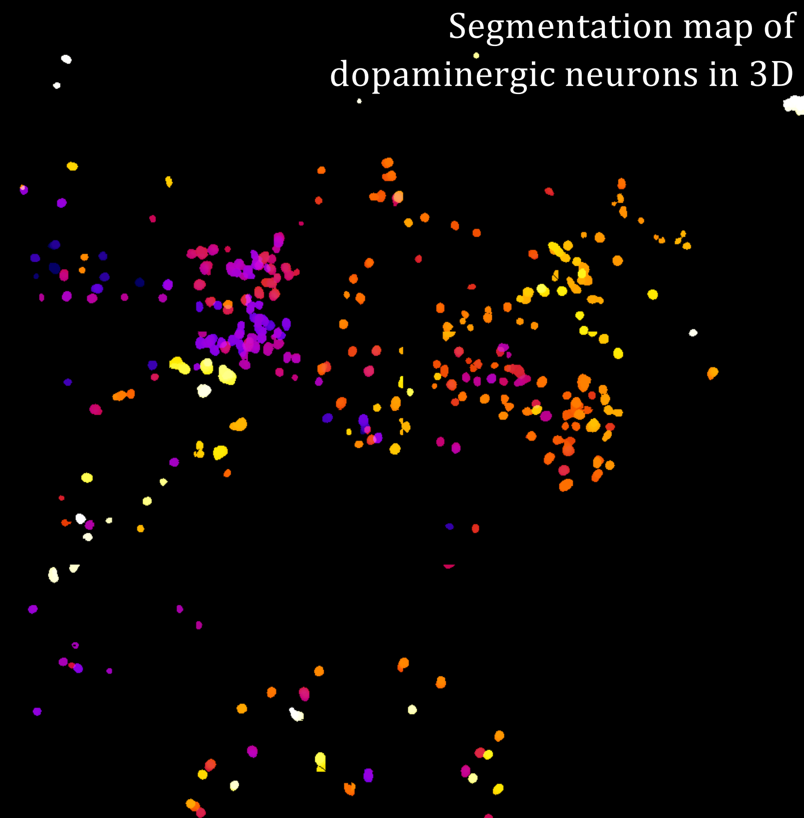


Step 6 : Inference on real data
Increasing neuronal activity
Actin rings
Actin fibers
High Mg2+
High KCl
Glutamate-Glycine
Expert : 1 image in 25 minutes
Network : 1500 images in 25 minutes

Automating the analysis of two-photon microscopy images of the dopaminergic neuronal circuit

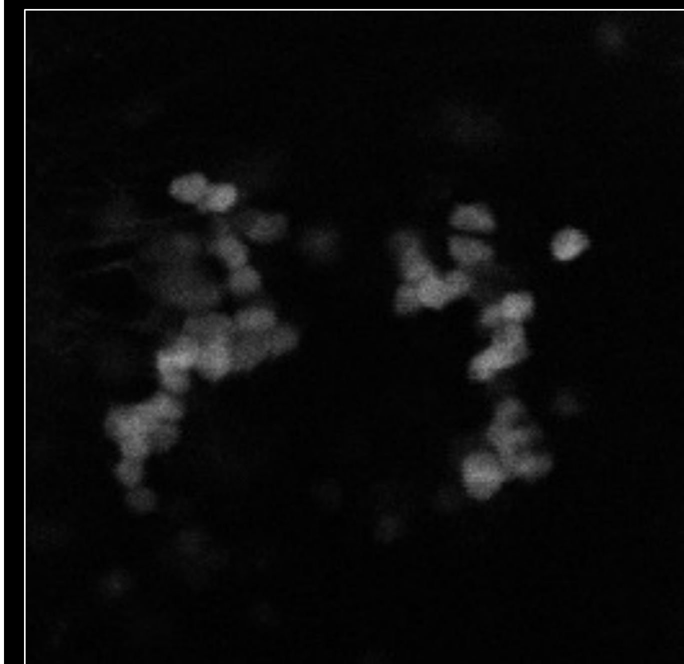


Deep learning



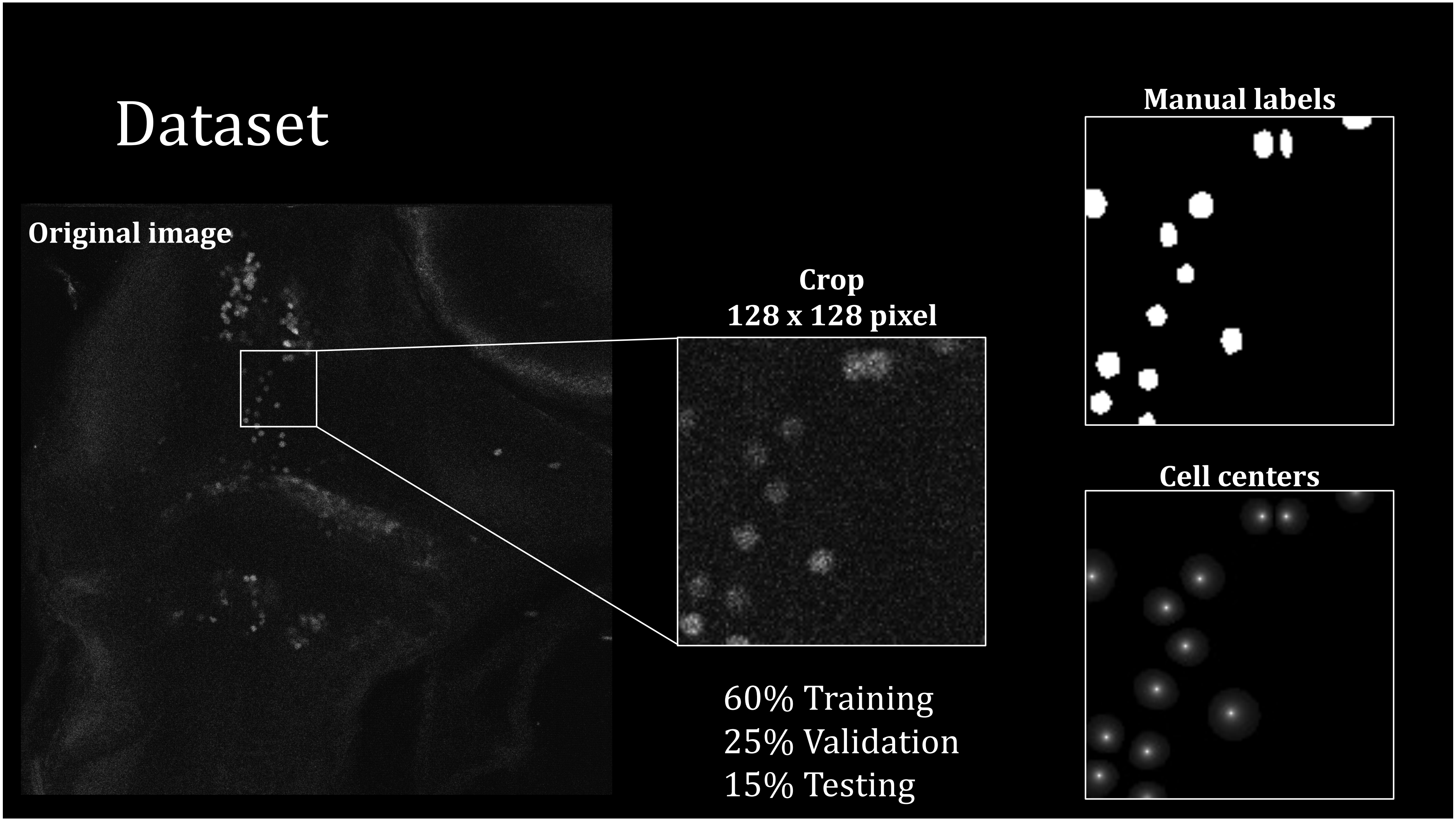
Detection challenges

Dense regions

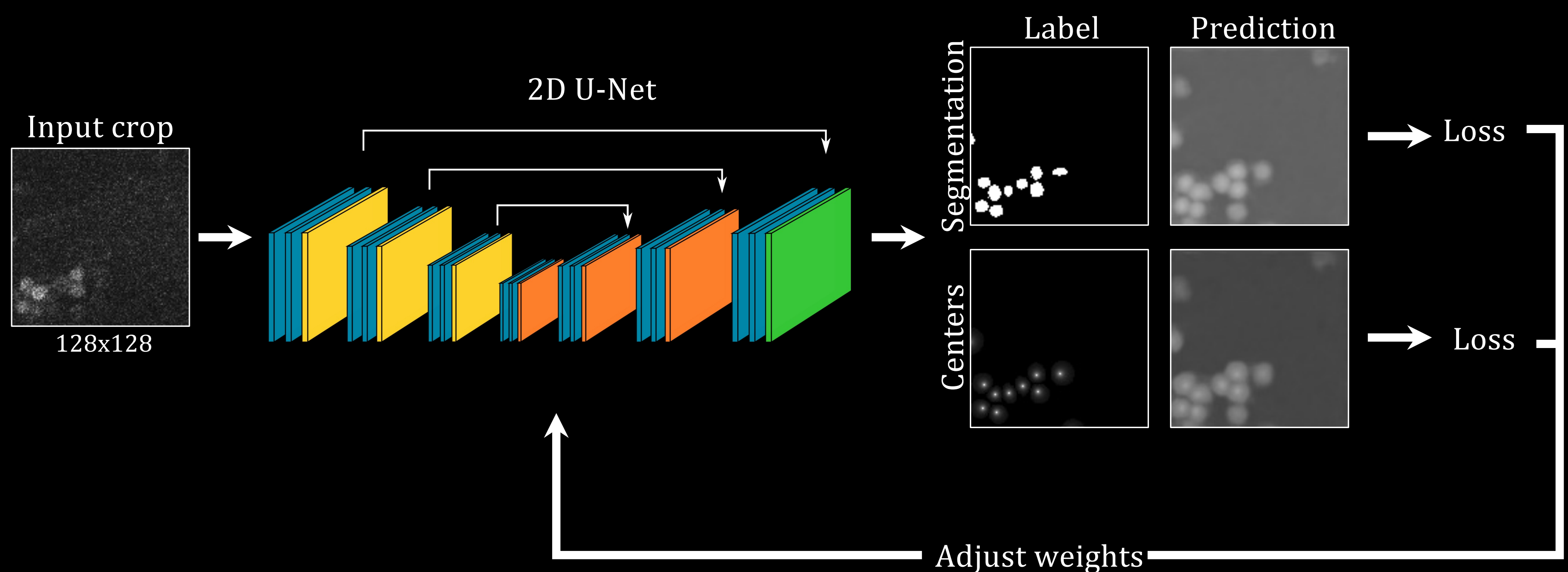


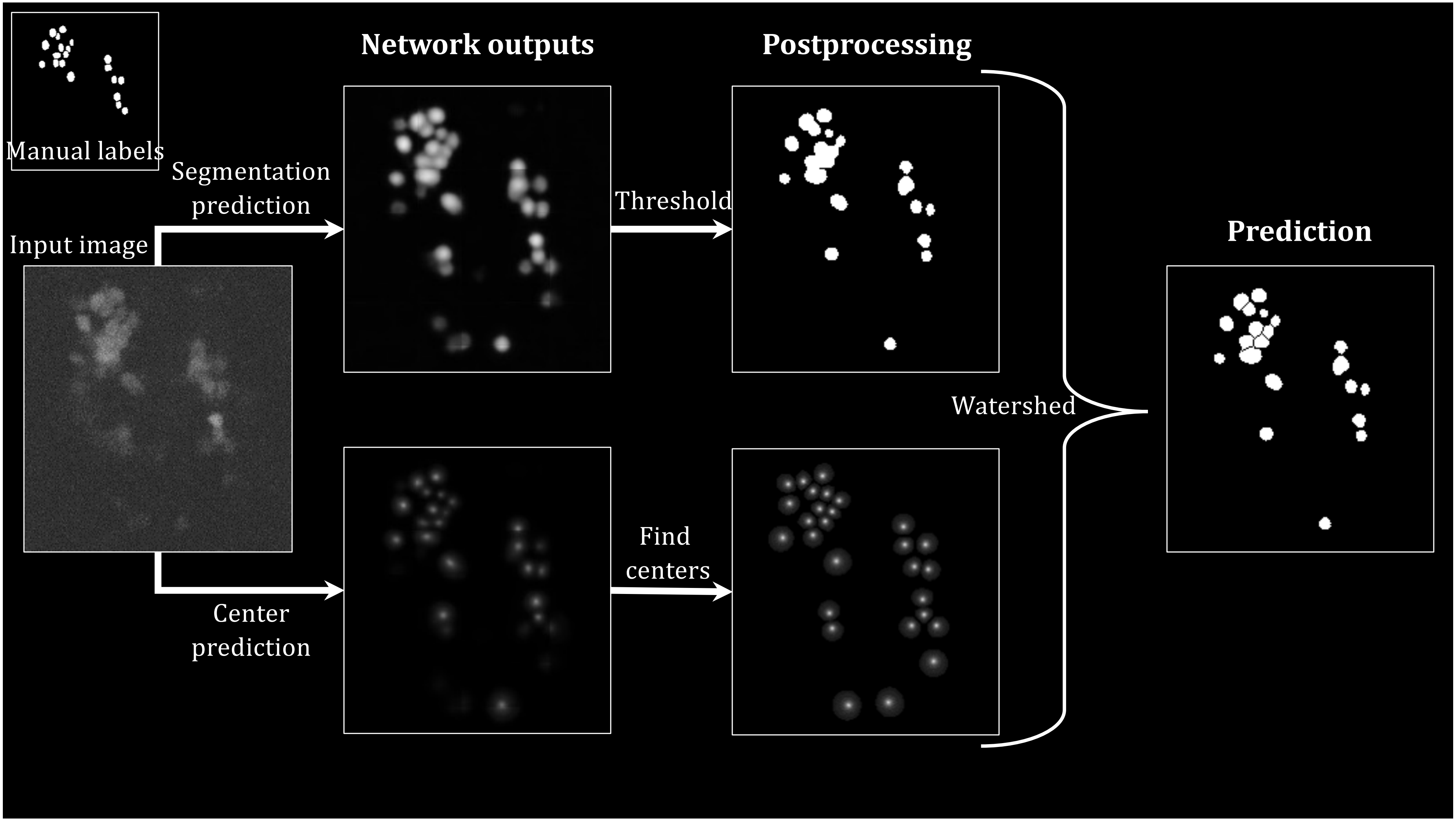
Intensity and contrast variations

Undesirable foreground

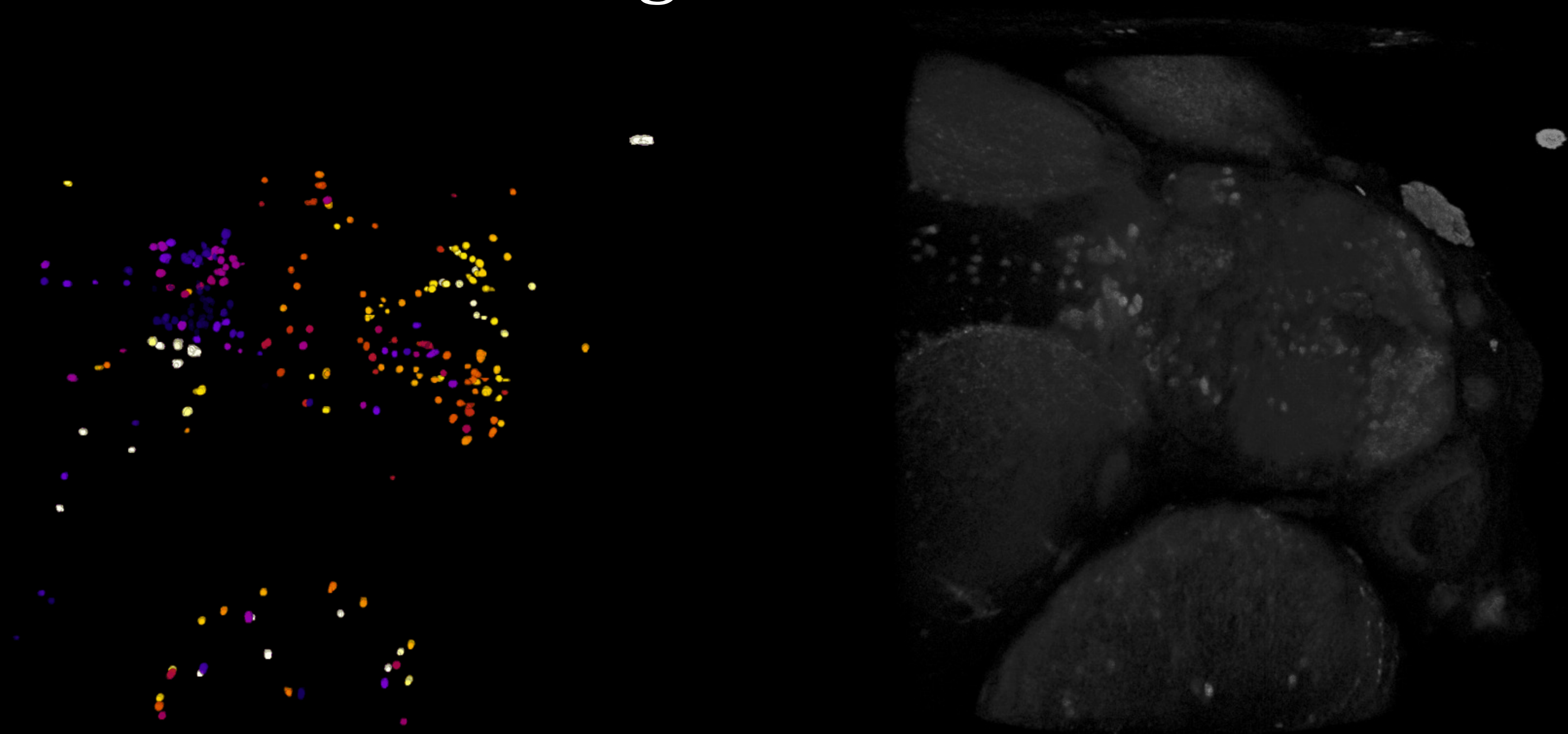

Dataset
Original image
Crop
128 x 128 pixel
60% Training
25% Validation
15% Testing
Manual labels
Cell centers

Fully convolutional network : U-Net




Manual labels
Network outputs
Postprocessing
Segmentation
prediction
Threshold
Input image
Prediction
Watershed
Find
centers
Center
prediction

Automated 3D segmentation

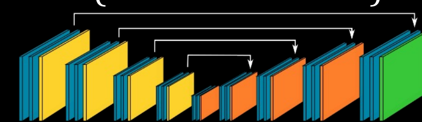


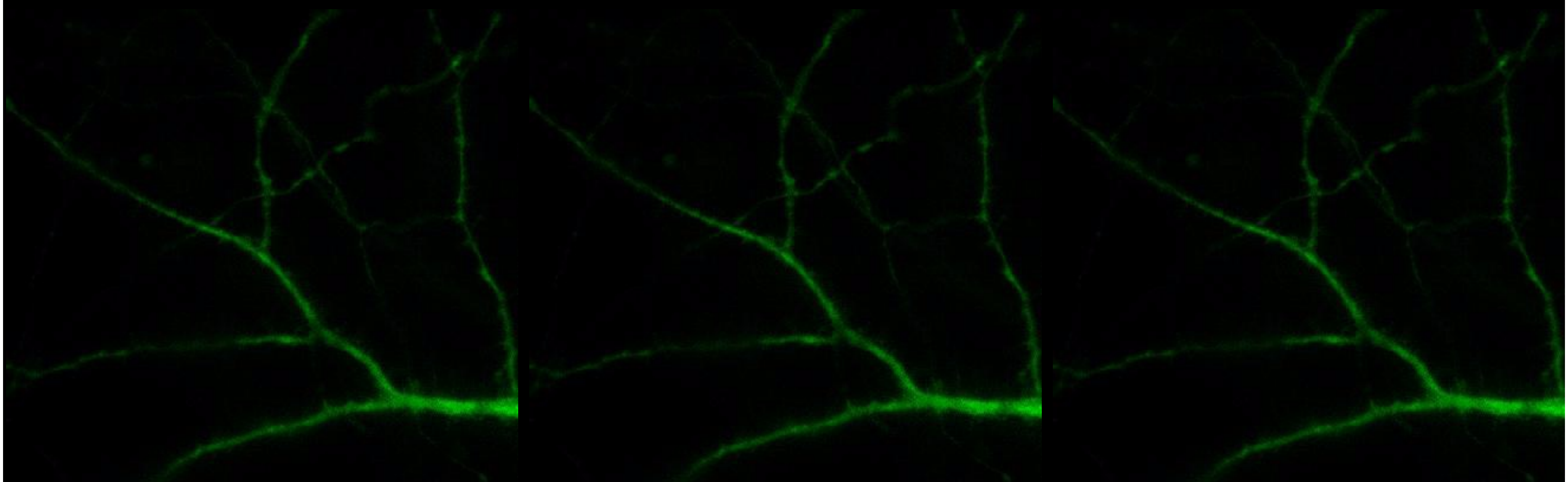
Detection of neuronal activity with a 3D neural network

Widefield imaging – GCaMP6f
spontaneous miniature Ca^{2+} events

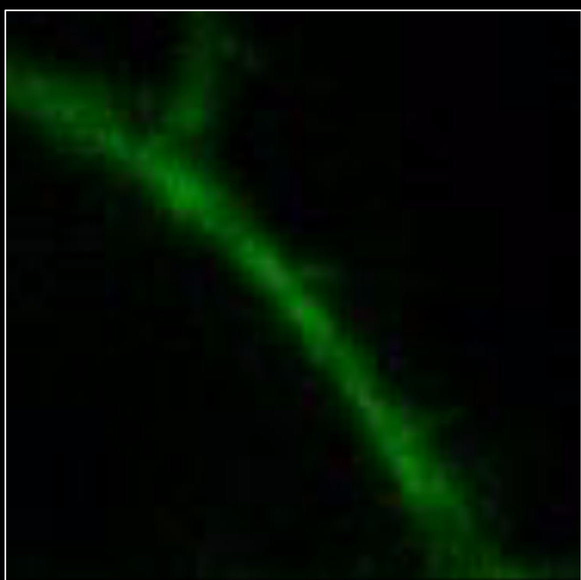
Foreground detection
(classical Matlab analysis)

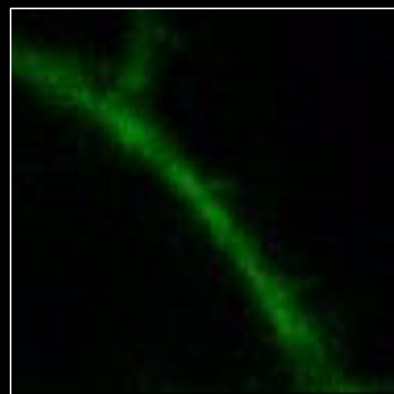
3D-UNet
(Neural Network)



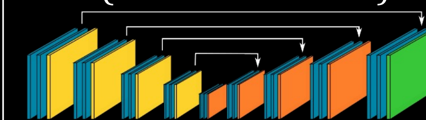


Increased segmentation precision with a 3D neural network





3D-UNet
(Neural Network)

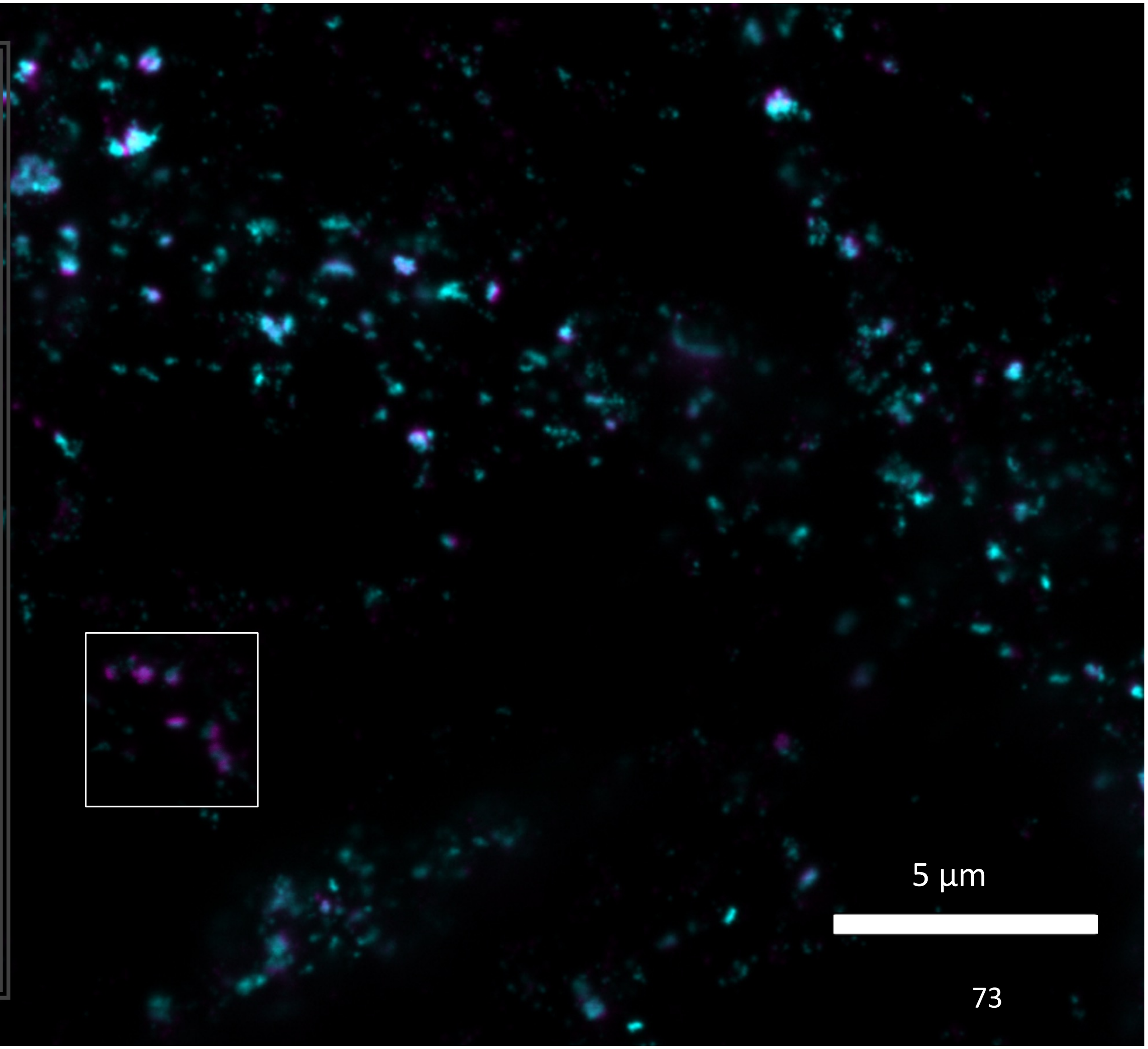


Foreground detection
(classical Matlab analysis)

Challenges

Metric evaluation

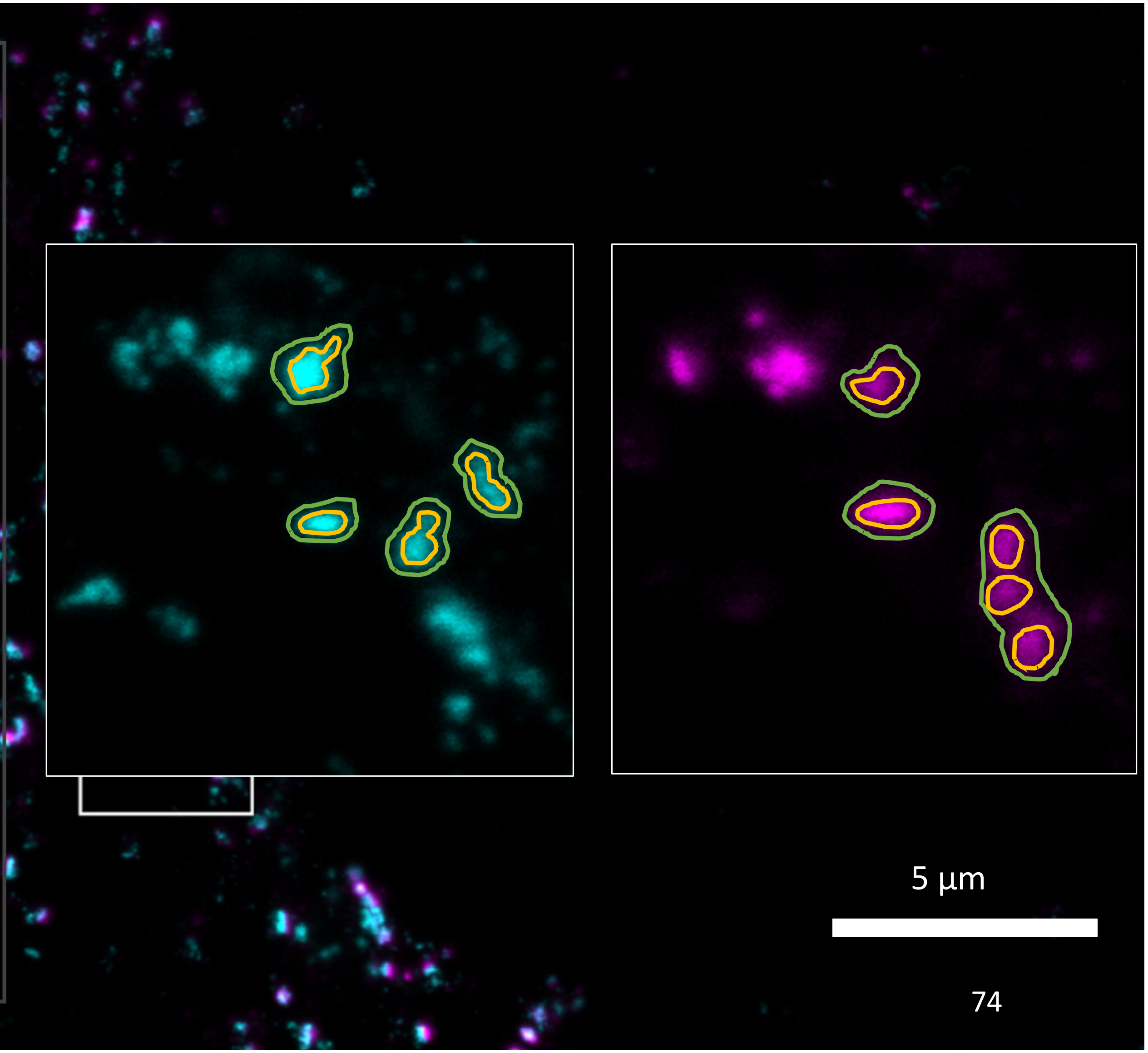
Bakas *et al.*, **Sci Data** (2017); Bakas *et al.*, **arXiv** (2019); Chen *et al.*, **ECCV** (2018); Menze *et al.*, **IEEE Trans Med Imaging** (2019); Taha and Hanbury, **BMC Med Imaging** (2019); Zhou *et al.*, **Array** (2019);



Challenges

Metric evaluation

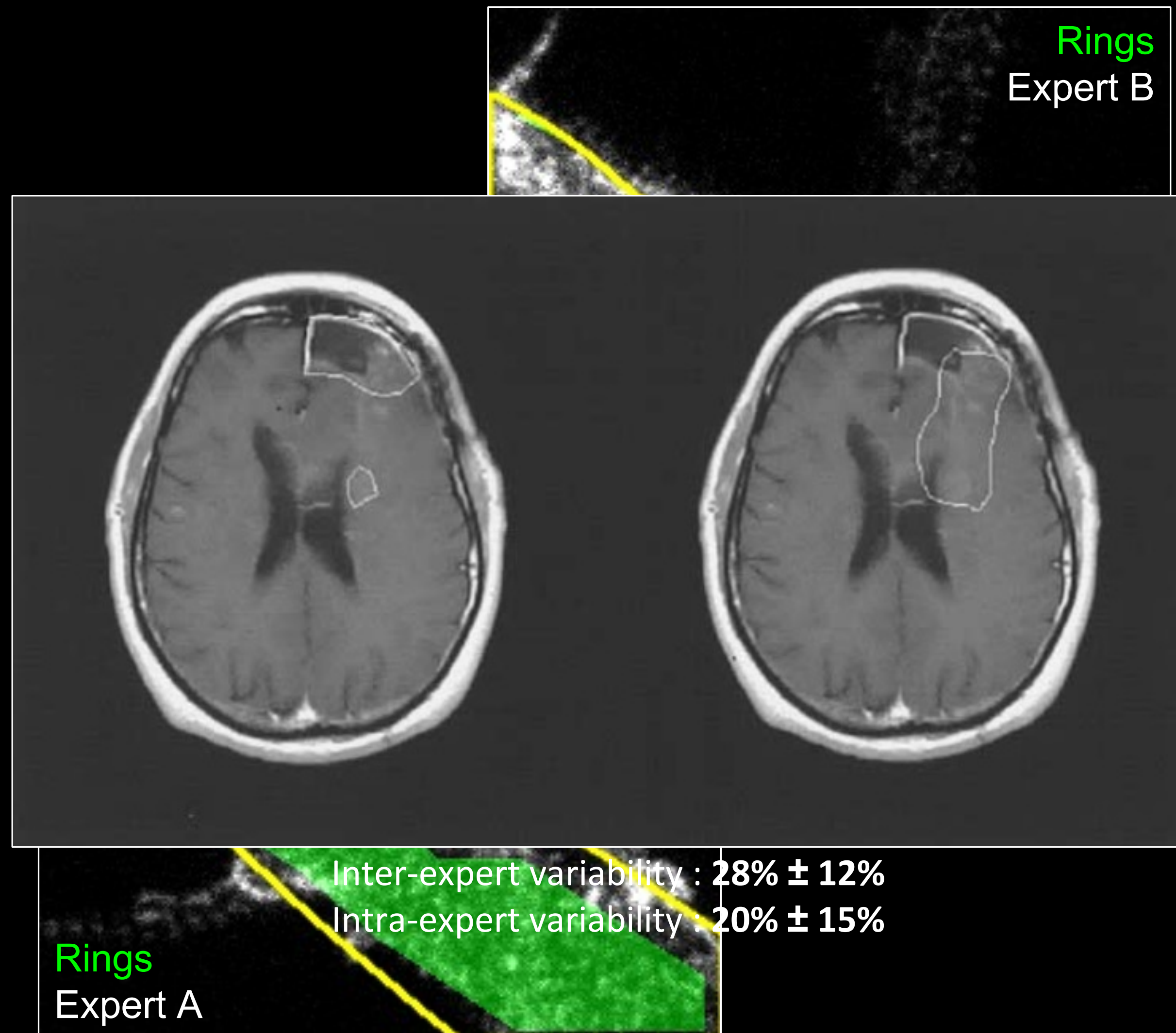
Bakas *et al.*, **Sci Data** (2017); Bakas *et al.*, **arXiv** (2019); Chen *et al.*, **ECCV** (2018); Menze *et al.*, **IEEE Trans Med Imaging** (2019); Taha and Hanbury, **BMC Med Imaging** (2019); Zhou *et al.*, **Array** (2019);



Challenges

Expert variability

Mazzara *et al.*, **J Rad Onco** (2004);

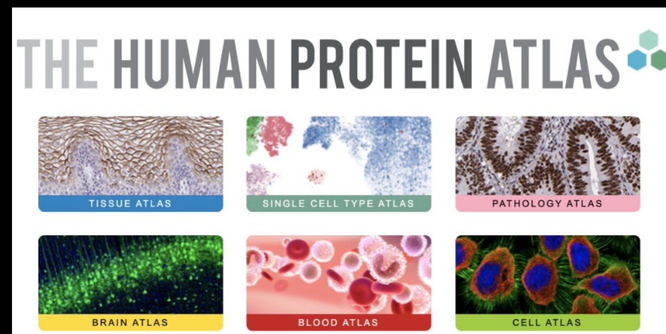


Challenges

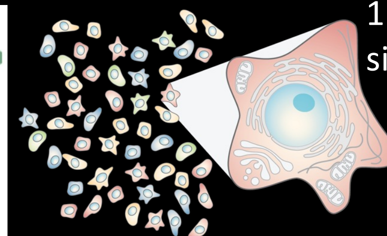
Datasets

IMAGENET

14 000 000 images
1 000 000 images with bbox



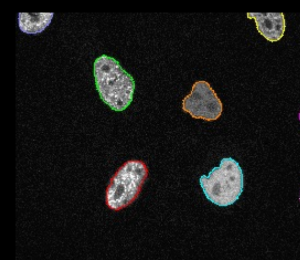
~ 125 000 images – Cell Atlas

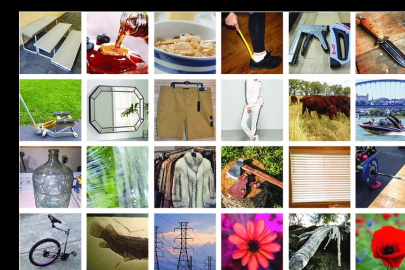


16 billions of single cells

JUMP Cell-Painting Consortium

Cell tracking challenge





THINGS Datasets

- **Hystopatology Astrocyte dataset**
  8730 patches of 500 × 500 pixels
  Olar et al. **2024** *Scientific Data*

- **F-actin dataset**
  ~ 200 large images
  Lavoie-Cardinal et al. **2020** *Scientific Reports*

Challenges

Generalization properties