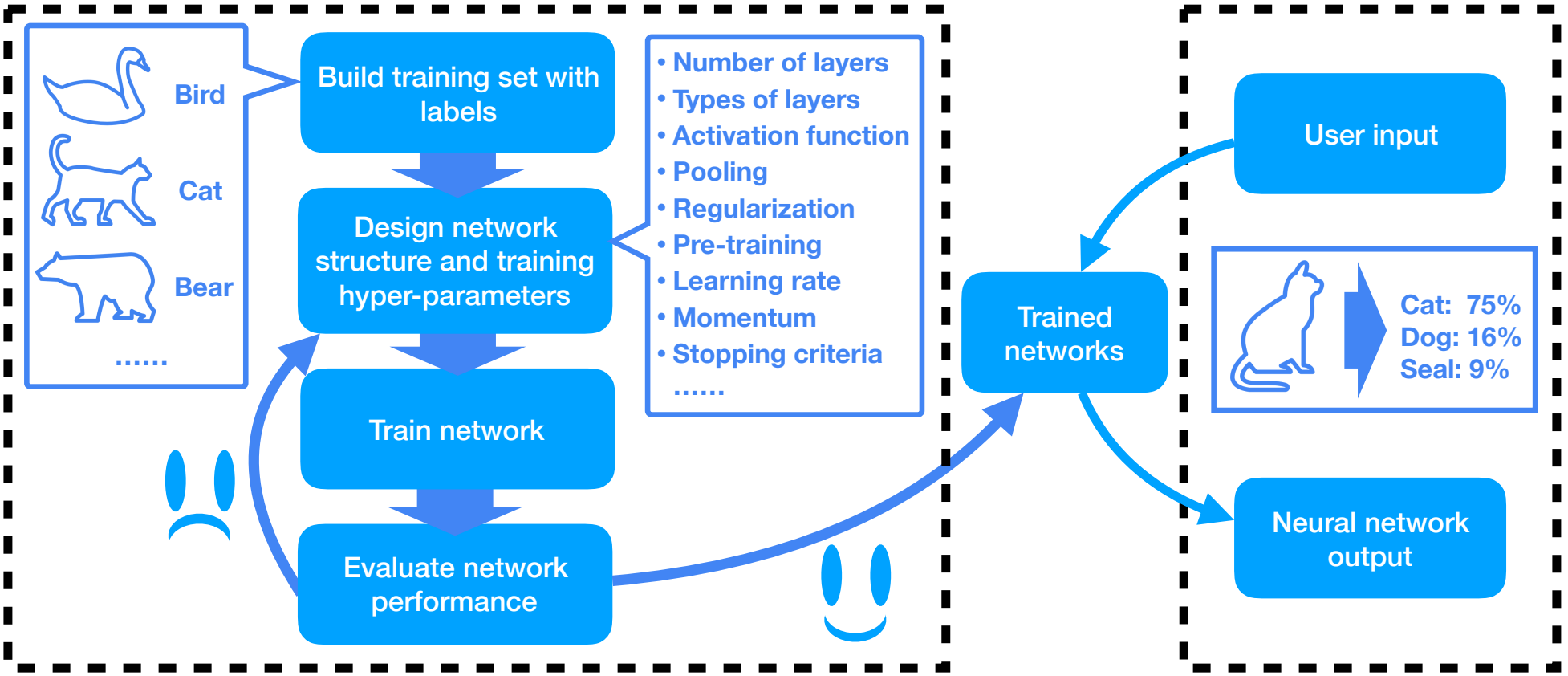
Challenges in deep learning applications to CryoEM data

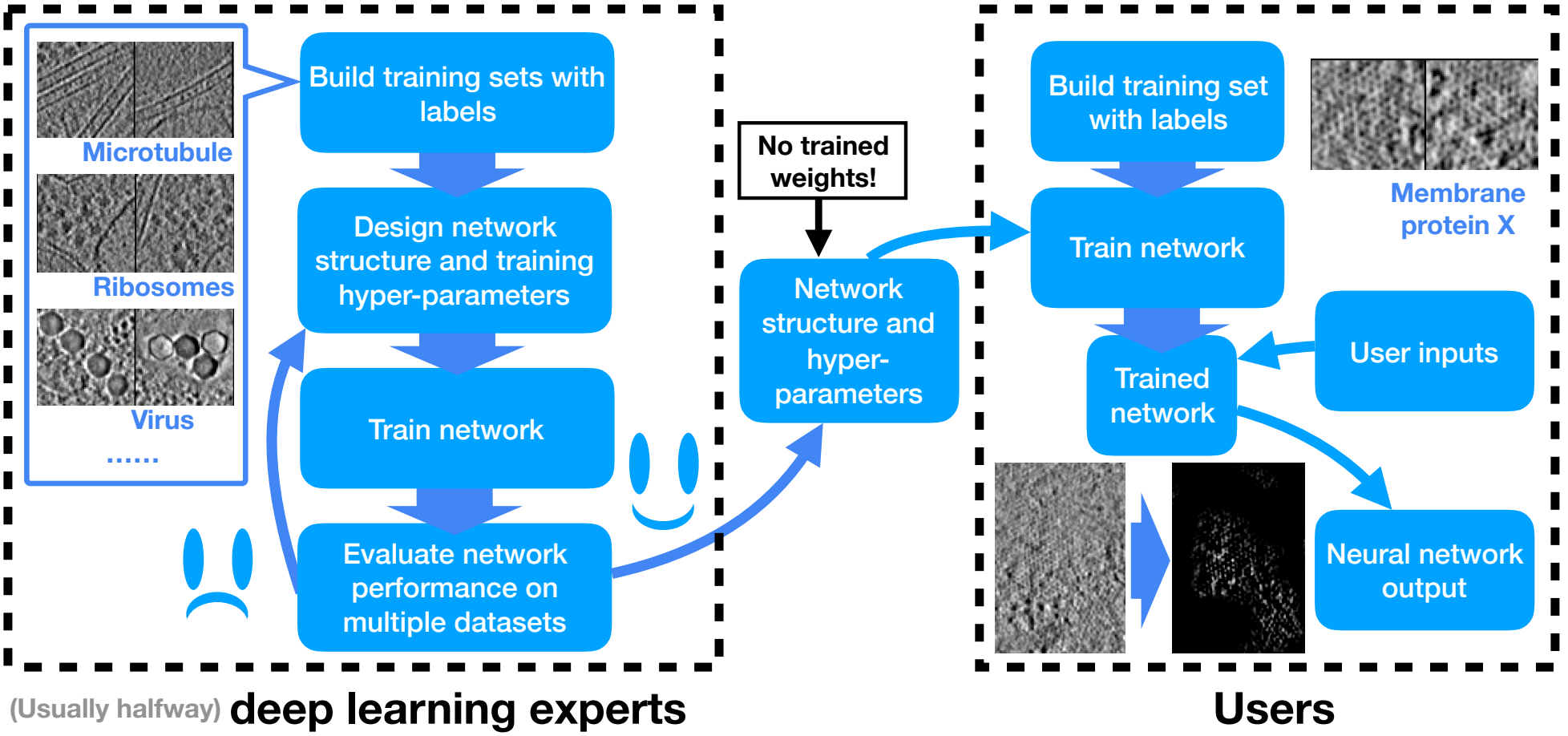
Muyuan Chen
2018-04

Business model

Typical deep learning applications



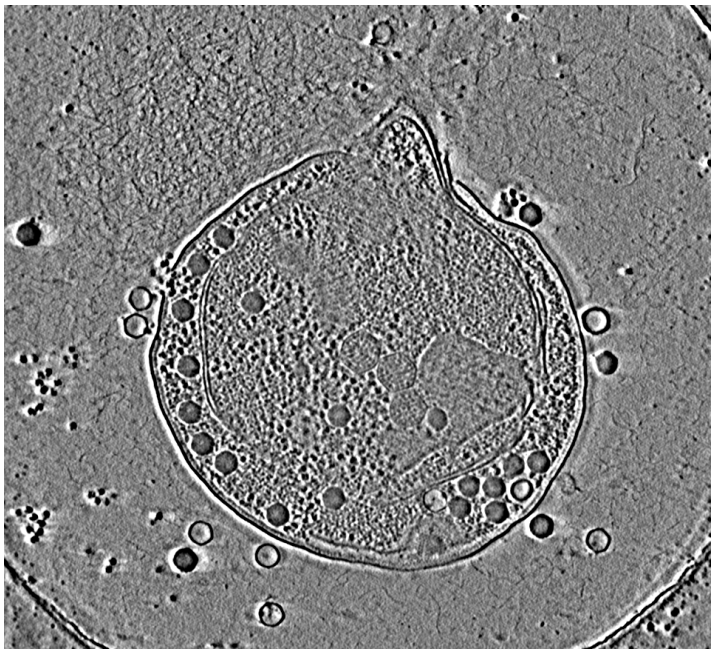
CryoEM applications

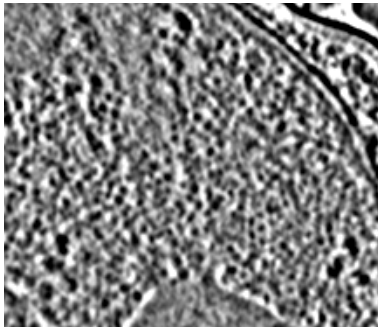


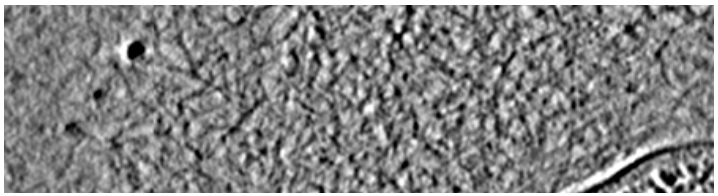
Datasets

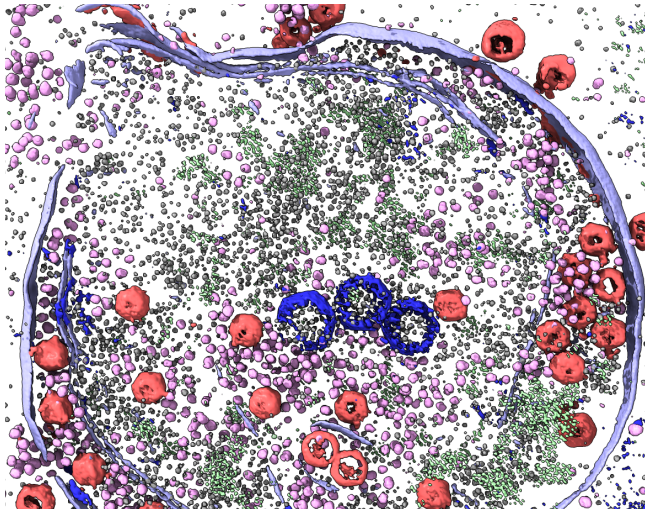
“Noise” & artifacts

- ‘Junk’
- High contrast objects
- Crowding
- CTF





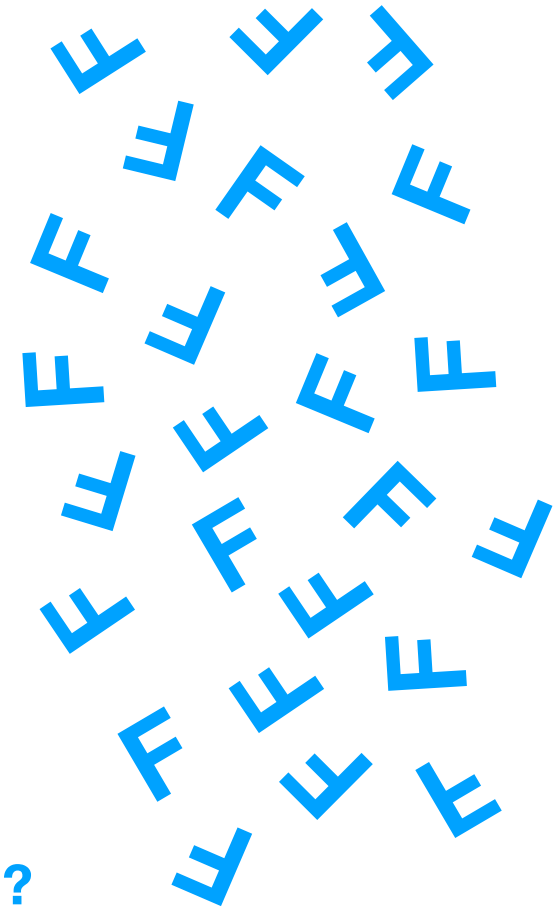




We are only identifying a small fraction of cellular features…

Rotational-translation invariance

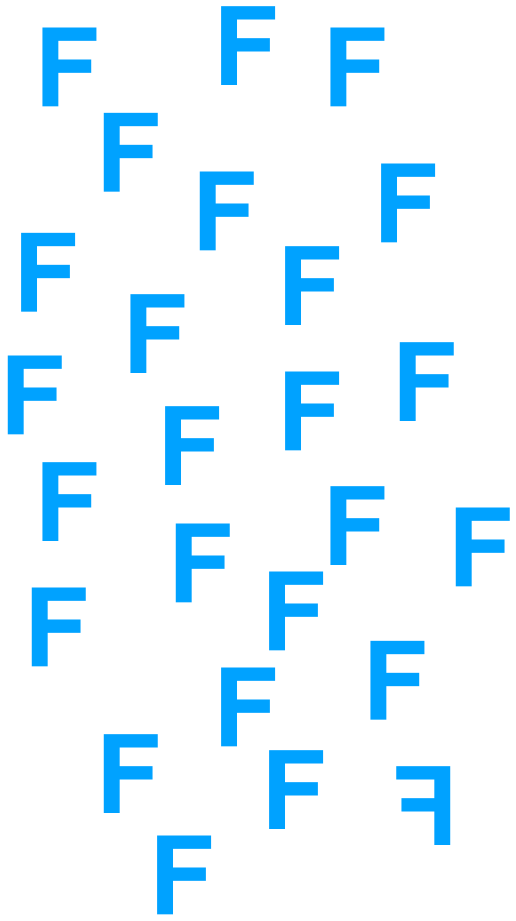
- Most biological features have random rotation-translation in CryoEM data
- Neural networks (biological and computational) are intrinsically bad at dealing with rotations….



Which ‘F’ is flipped?

Rotational-translation invariance

- Most biological features have random rotation-translation in CryoEM data
- Neural networks (biological and computational) are intrinsically bad at dealing with rotations….



This is even harder in 3D….

Rotational-translation invariance

- Inside deep learning framework
  - Pooling
  - Data augmentation
  - Max-out
  - Transforming autoencoder
- External methods
  - Starting from rotational invariants
    (spherical harmonics,
    bi-spectrum etc.)
  - Pre-align images

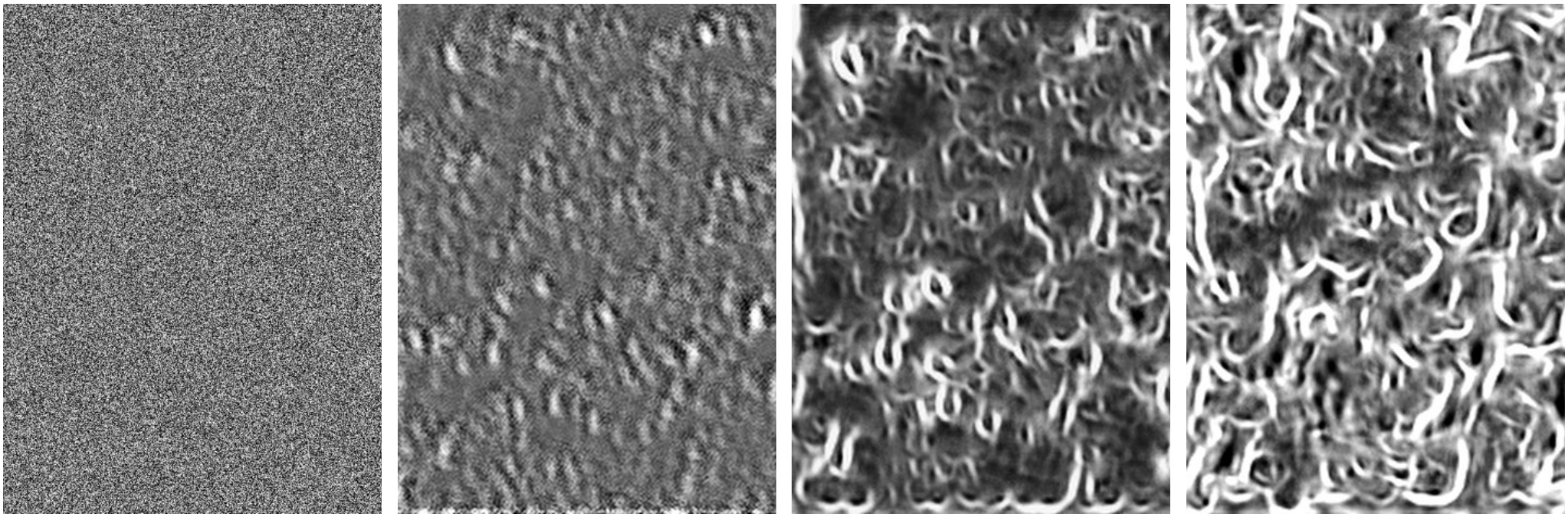
Training

Hyper-parameter selection

- How good are the default parameters?
- Are they robust to image size, dataset size, feature shape, noise level?
- When does training converge and how do we know it?

Overfitting

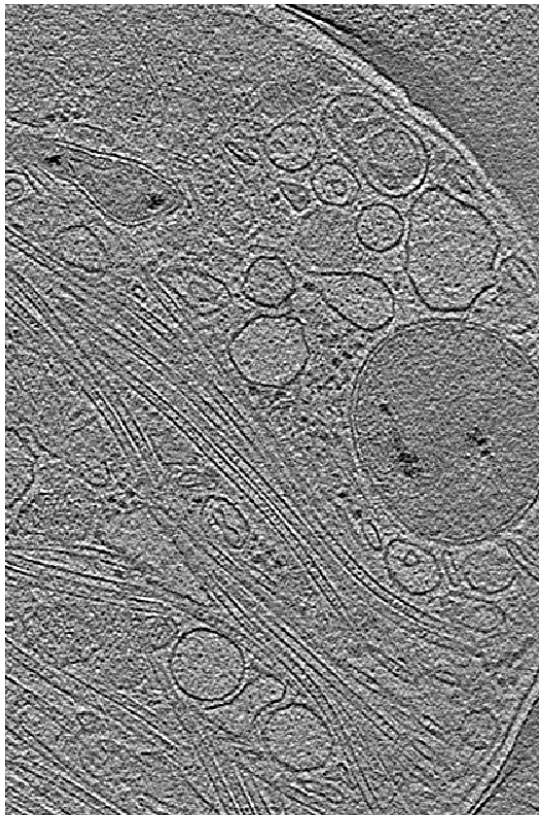
Dream of membranes…

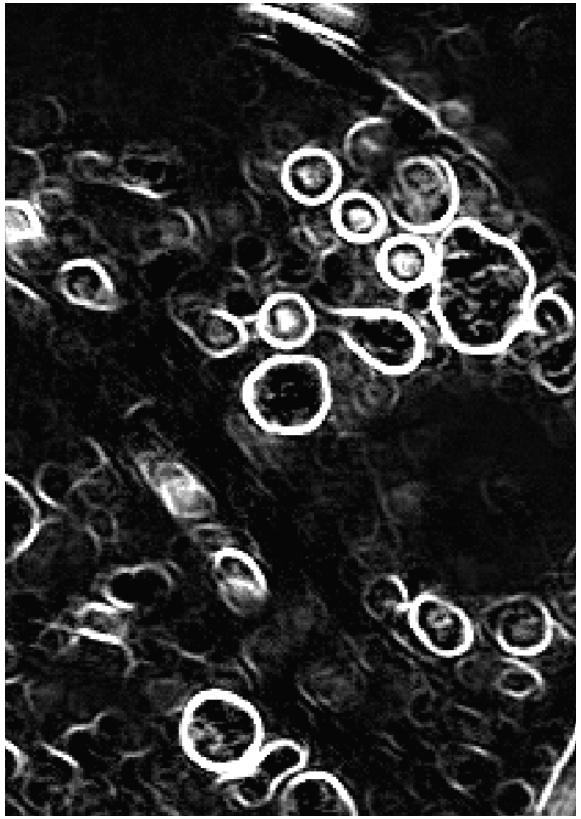


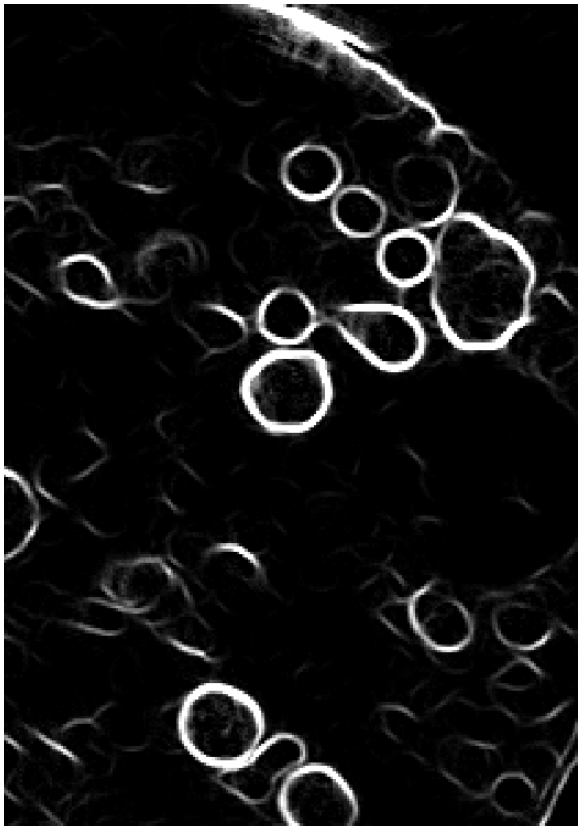
Overfitting is a problem but usually not too bad…

Overfitting

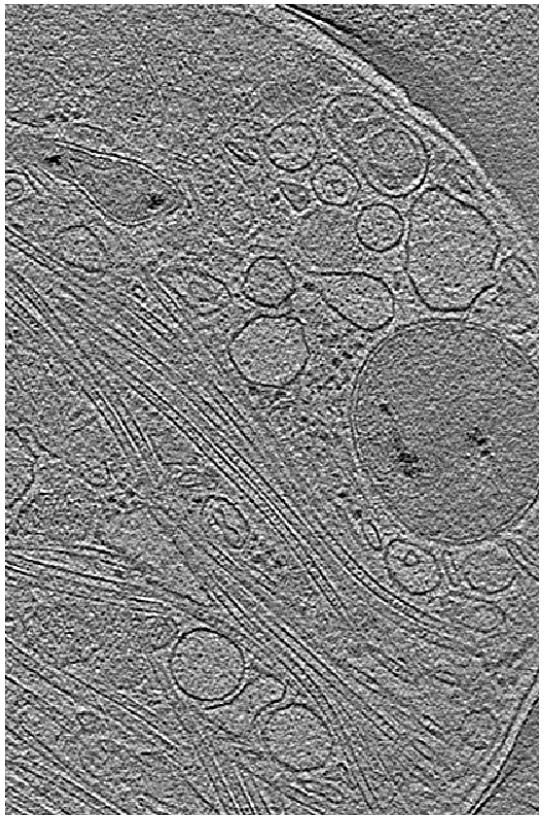
Effect of different training sets

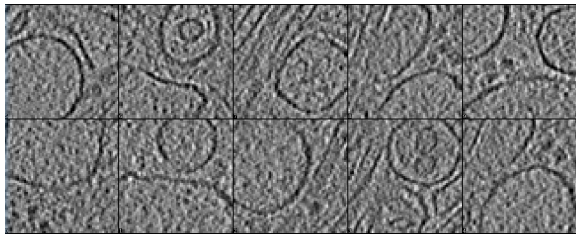


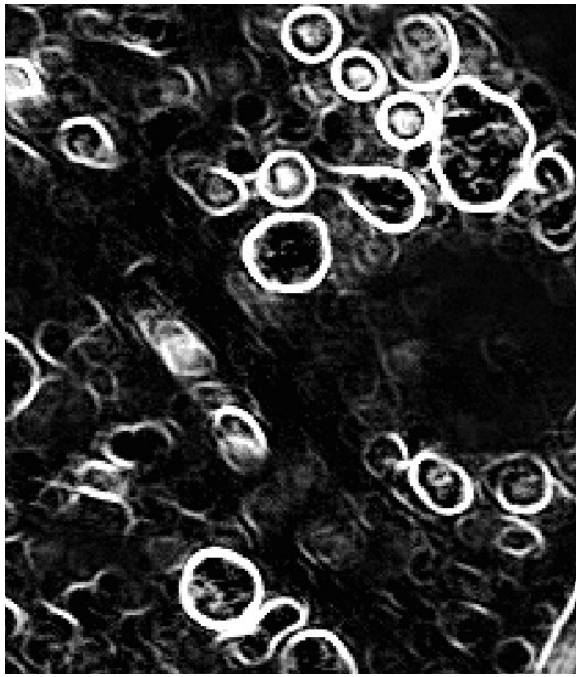


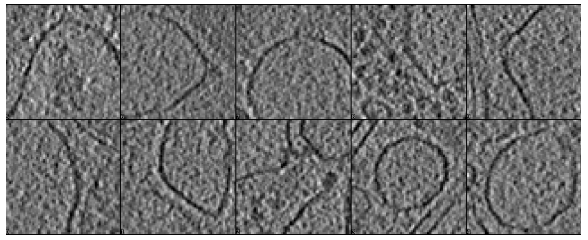


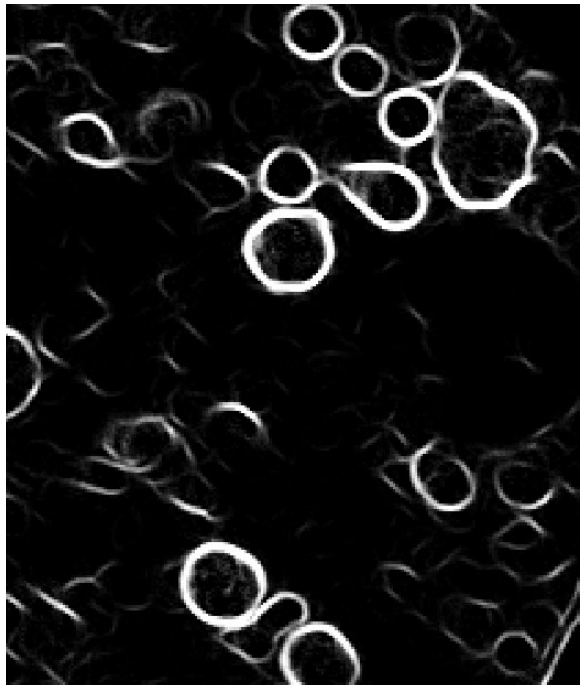
Overfitting







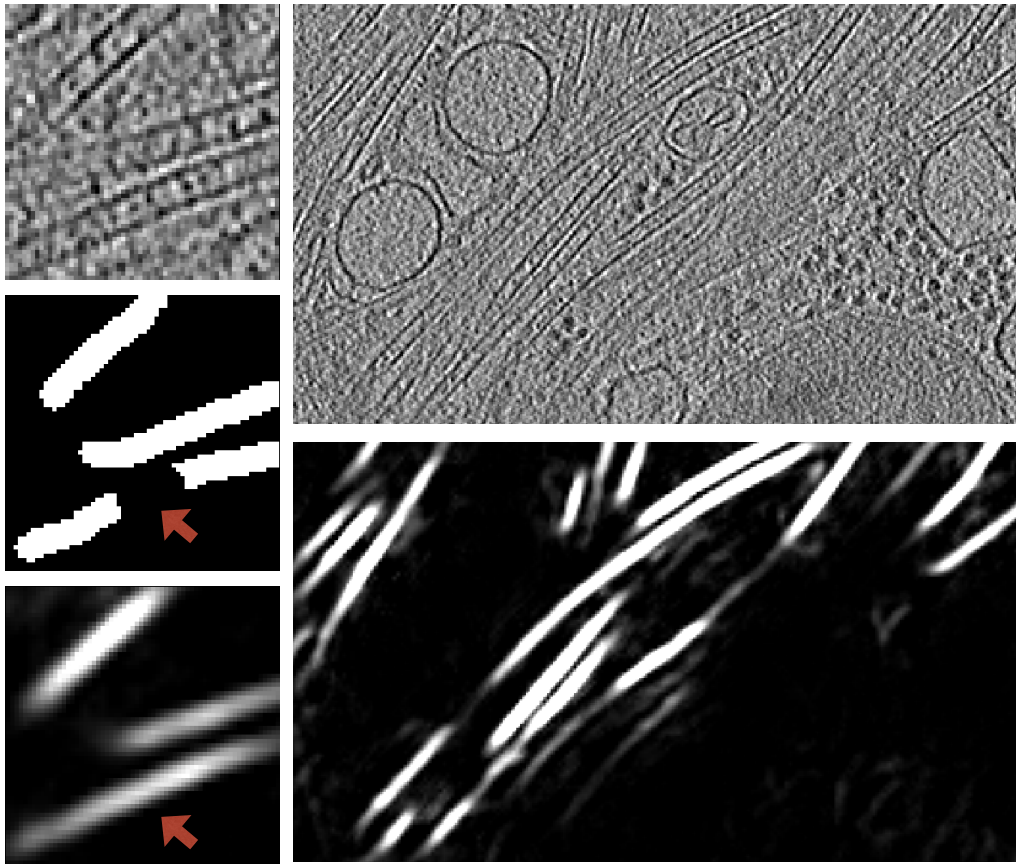




Overfitting and regularization

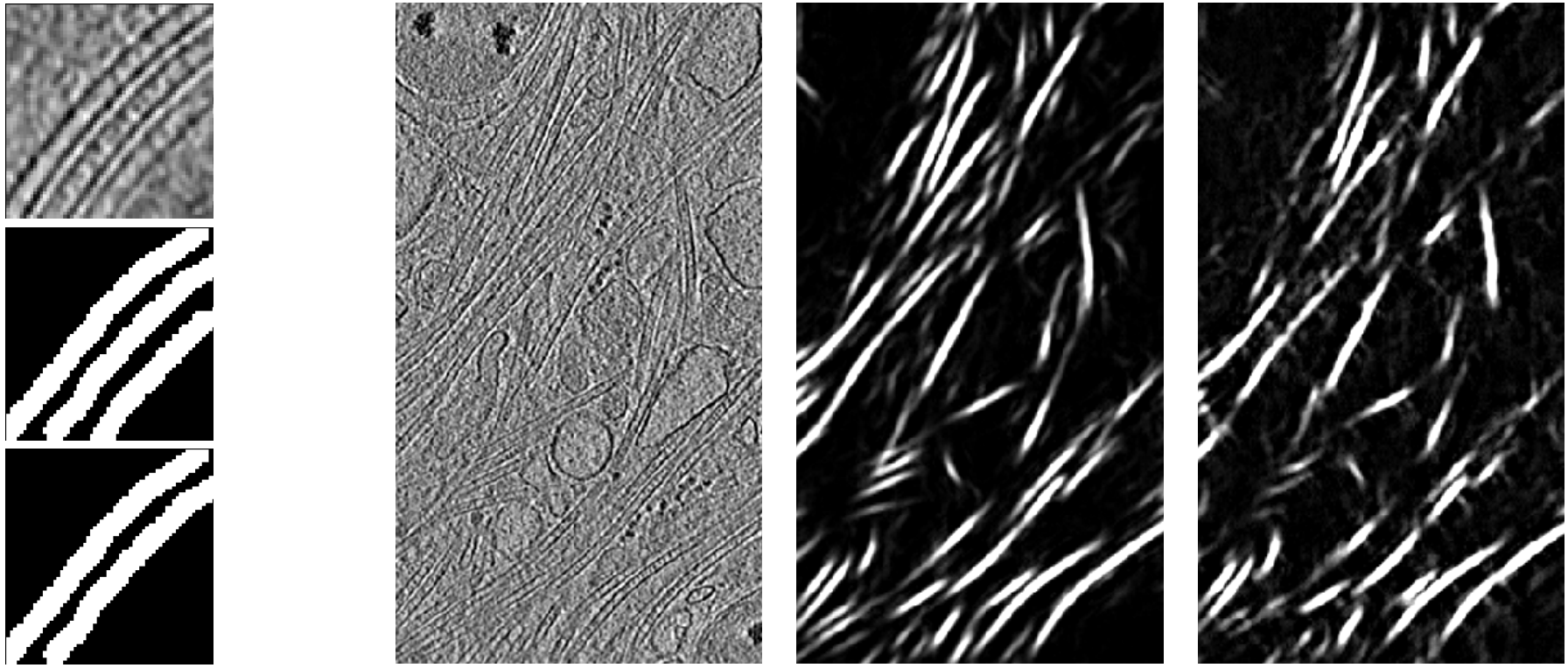
- Noise and rotational invariance helps..
- Selection of positive/negative training set
- Do we need a validation set and how big should it be?
- How does regularization affect convergence?

Imperfect supervision



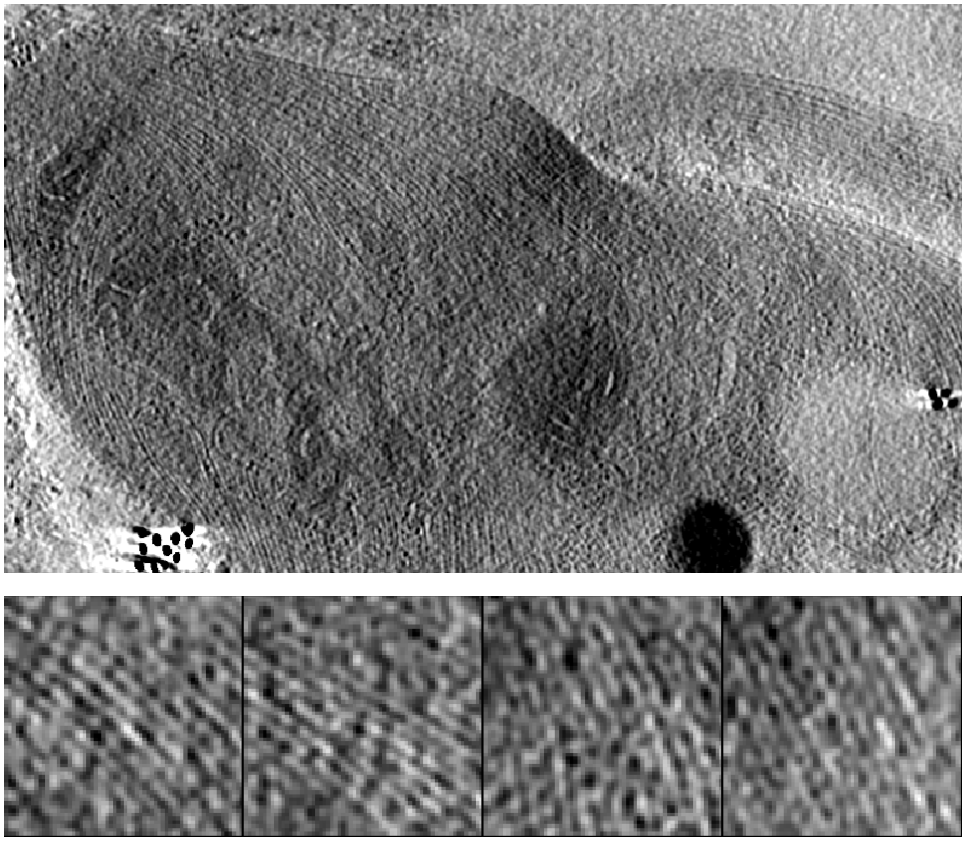
1 error out of 5 particles

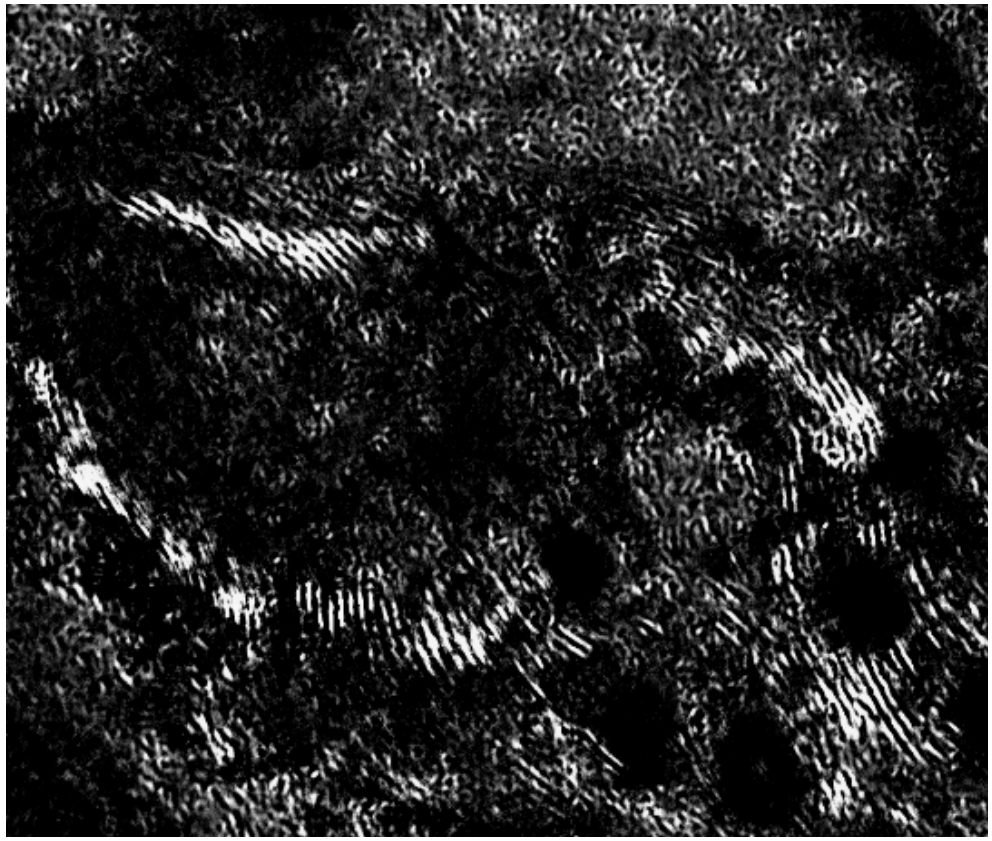
Imperfect supervision



4 errors out of 8 particles

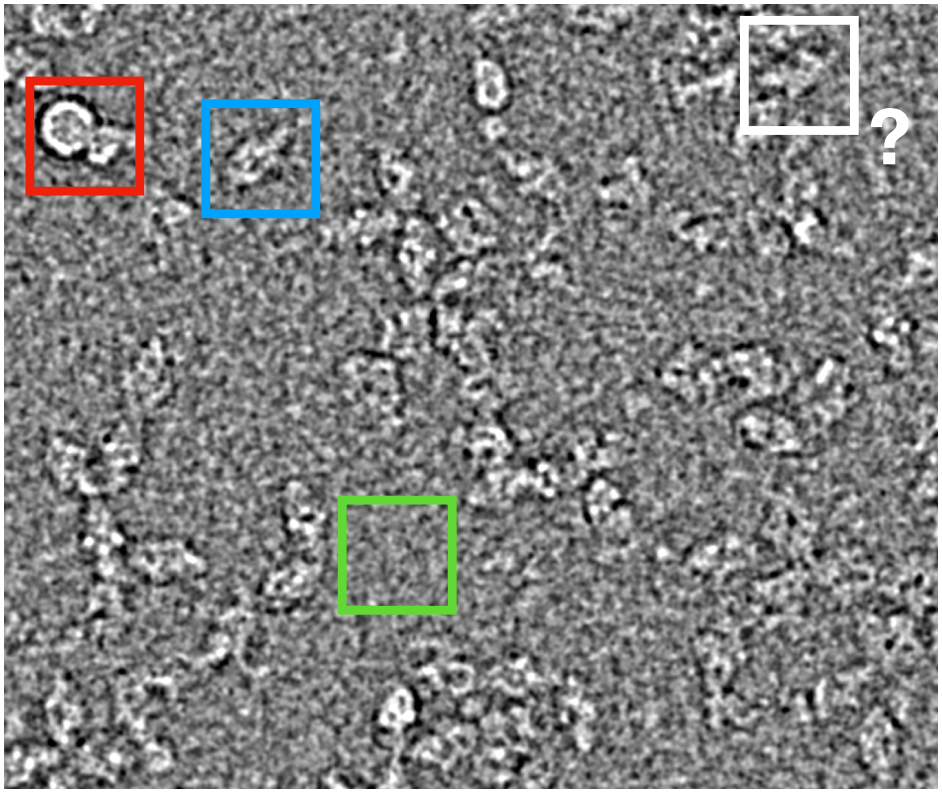
Imperfect supervision





Only use confident regions for training…

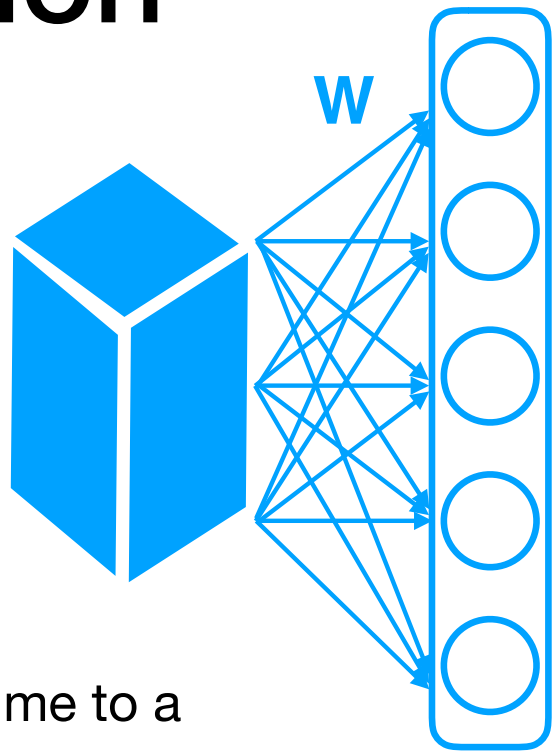
Imperfect supervision



Uncertain is an option..

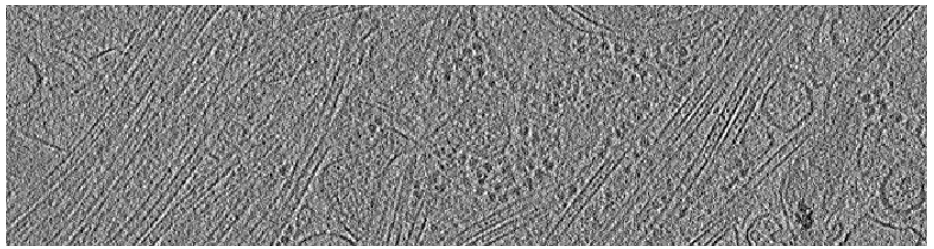
Other problems…

Memory limitation

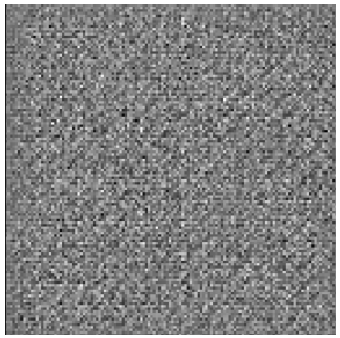


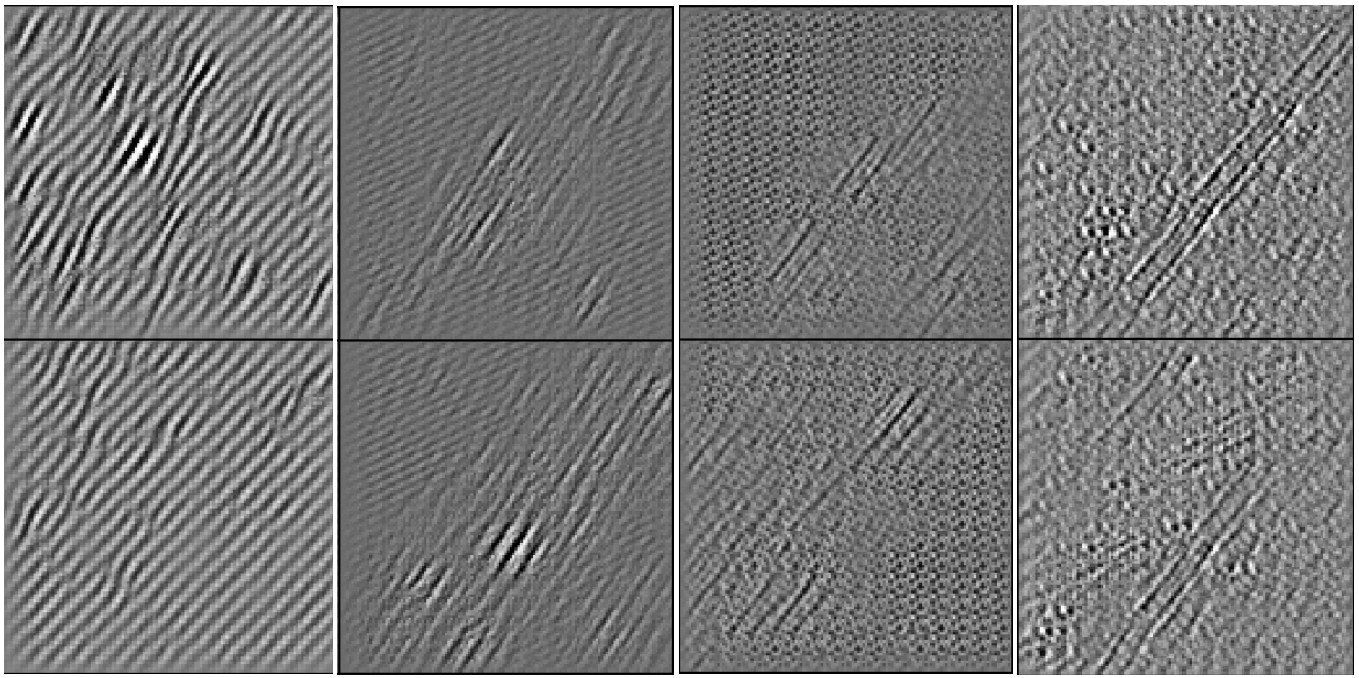
- Heterogeneity problems
  subtomogram, single particle
- 3D feature annotation
- Nvidia GTX 1080 : 8GB memory
- One fully connected network from a 64x64x64 volume to a layer of 2048 units:
  at float32: 64x64x64x2048x4=7.2GB
  The actual cost is much higher due to optimization etc..

Unsupervised methods



- Generative Adversarial Network?
  - Rotational invariant
  - 3D with missing wedge
- Look into the black box…





Solutions?

- What is the biological question?
- Solving constrained and well-defined problems is much easier…
  - Particle picking: two-step solution
  - Tomogram annotation: making figures? SPT? statistical conclusion?
  - Heterogeneity - limited scale, region.. separate from alignment
- Combine with conventional methods
  - Alignment, rotational invariants, PCA …
- It is okay to have some manual intervention….

Future deep learning applications

- Identification and classification of structures in cells
- Heterogeneity analysis in SPR and SPT
- Protein sequence-structure relationship

Acknowledgement

- Data providers:
  - Wei Dai, Center for Integrative Proteomics Research, Rutgers University
  - Stella Ying Sun, Stanford University
- PIs:
  - Steven Ludtke, Baylor College of Medicine
- NIH grants:
  - R01GM080139, P41GM103832

Thank you