Materials Discovery: Understanding Polycrystals from Large-Scale Electron Patterns

3rd Workshop on Advances in Software and Hardware (ASH) at BigData

December 5, 2016

Ruoqian Liu, **Dipendra Jha**, Ankit Agrawal, Wei-keng Liao,
Marc De Graef and Alok Choudhary

Northwestern University

Carnegie Mellon University

Hypothesis:

We can make use of ***big data + deep learning + advanced hardwares and softwares*** to make interesting discoveries - in materials!

Outline

- Deep Learning: the Basics
- Materials Discovery: the Basics
- Deep CNN for EBSD Indexing
- Advanced Software and Hardware in Use

Outline

- **Deep Learning: the Basics**
- Materials Discovery: the Basics
- Deep CNN for EBSD Indexing
- Advanced Software and Hardware in Use

What is Deep Learning

- A revolutionary breakthrough in machine learning
- Caused big companies, such as Microsoft, Facebook, Google, Apple, Baidu Yahoo! and IBM to heavily invest in this technology
- The perfect method to build large-scale recognition systems to exploit the information locked away in Big Data
- Boosted new surge of Artificial Intelligence

What Does Deep Learning Learn

Learning the representation

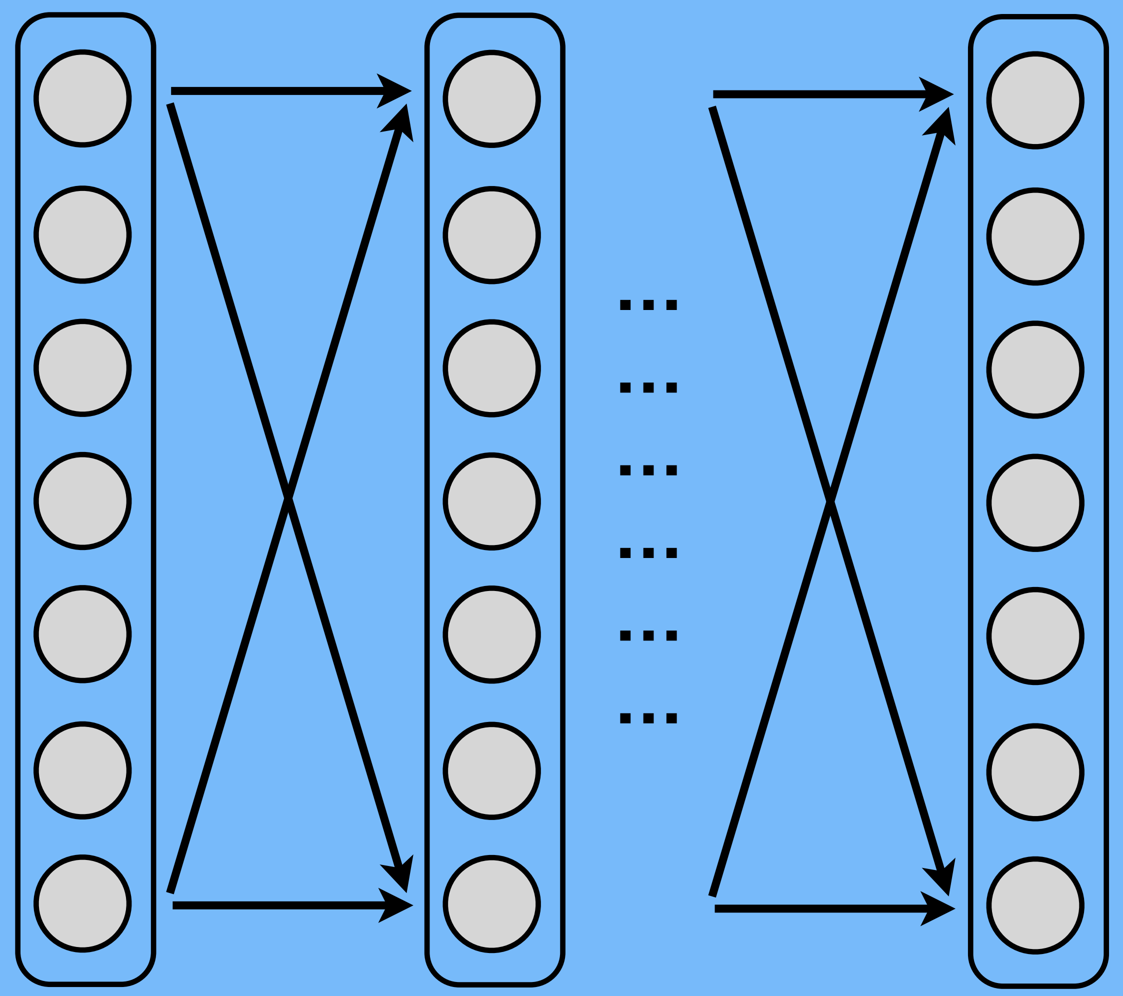
- The way in which data are **represented** can make a huge difference in the success of a learning algorithm.
- Deep learning enables the learning of multiple levels of representation, discovering more abstract features in the higher levels.

Learning as human does

- Because human brains appear deep, AI-tasks require deep circuits
- Because it is natural for humans to represent concepts at multiple levels of abstractions, deep architecture makes sense.
- Because human learn mostly unsupervised, only partially supervised.

Three Keys in Deep Learning

Deep neural networks



Big data

Fast computing

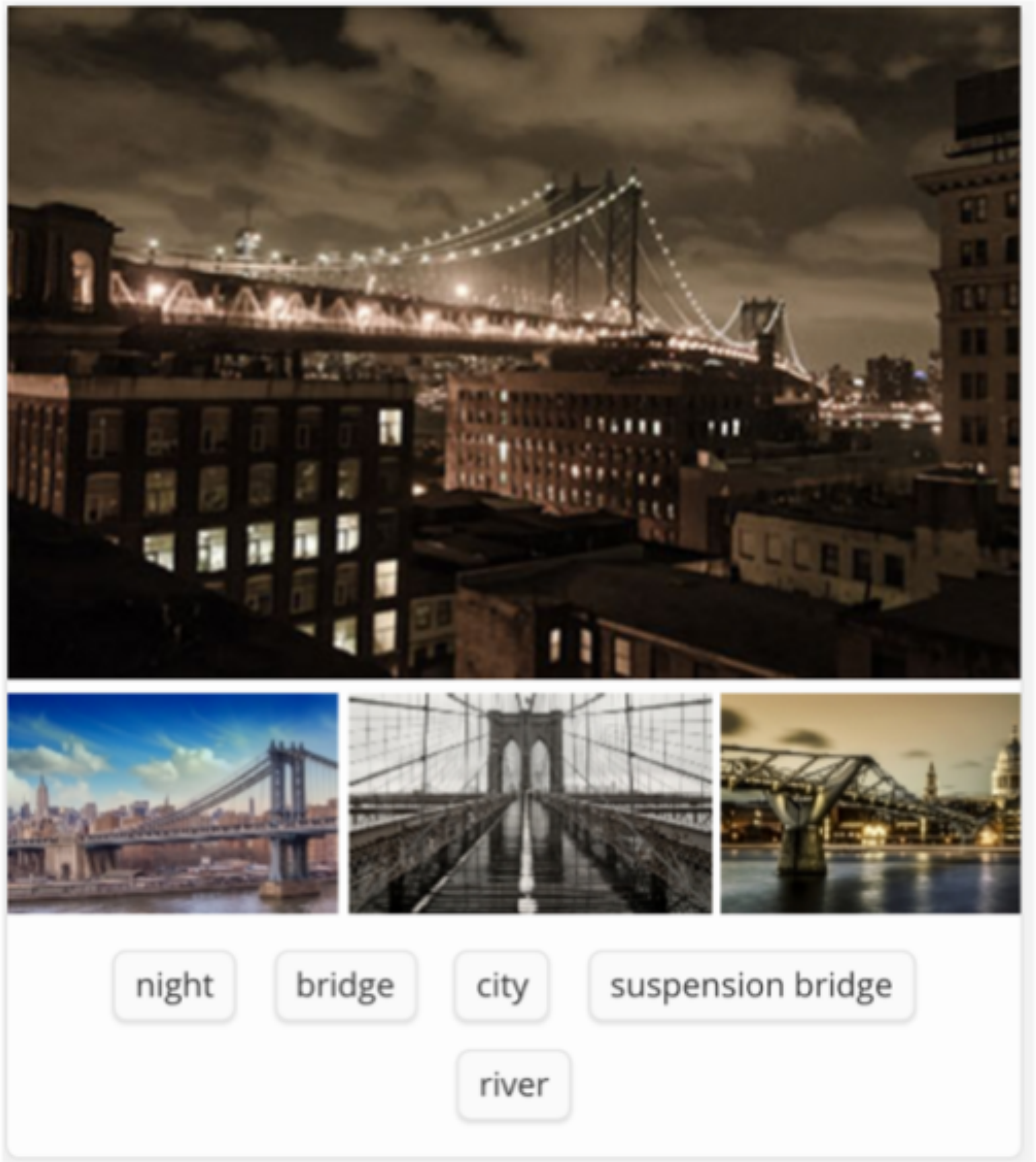
DL Achievements in Various Fields

- **Computer vision**: where DL first showed its power.
- **Speech recognition**: improved by 30% — an earthquake in this field.
- Genetics, drug discovery, health, …

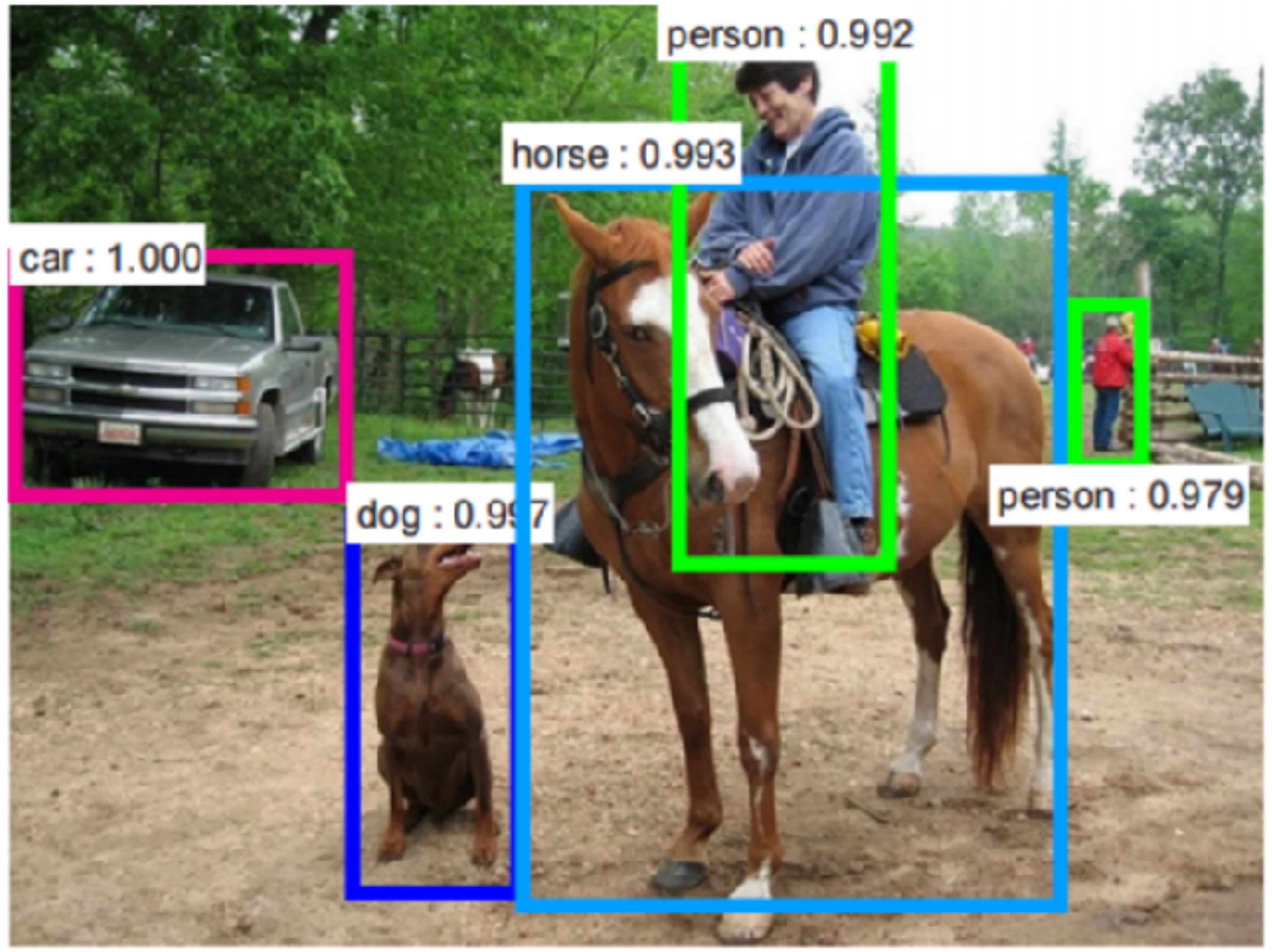
- Materials science 

Convolutional Neural Networks

- Convolutional Neural Networks (**ConvNets** or **CNNs**) are a category of Neural Networks that have proven very effective in learning **image** tasks.



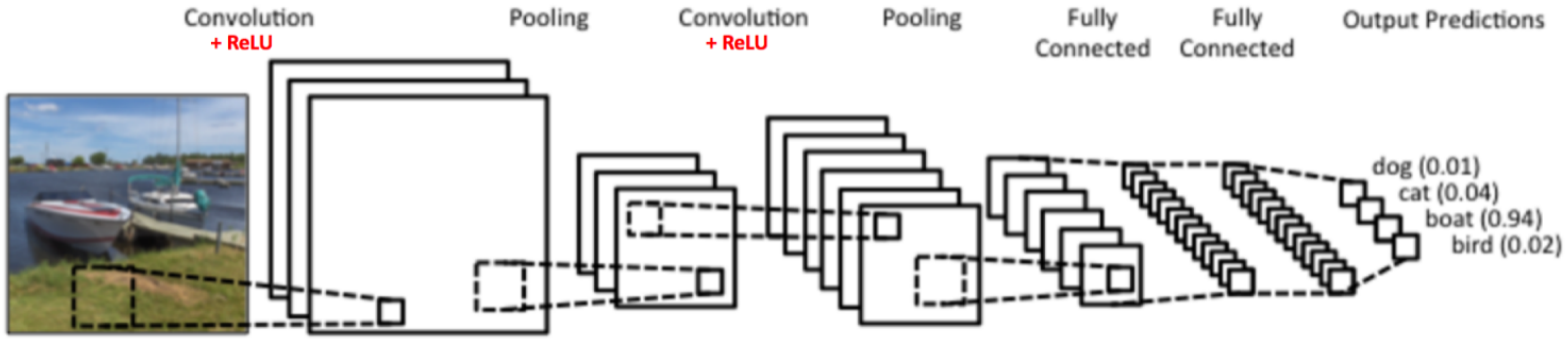
Given images, produce relevant tags



Given images, recognize/locate objects

CNN for Image Classification

- CNNs can come with different architecture. Below is similar to what's known as the LeNet. It classifies an input image into four categories: dog, cat, boat or bird.



On receiving a boat image as input, the network correctly assigns the highest probability for boat (0.94) among all four categories. The sum of all probabilities in the output layer should be one

LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

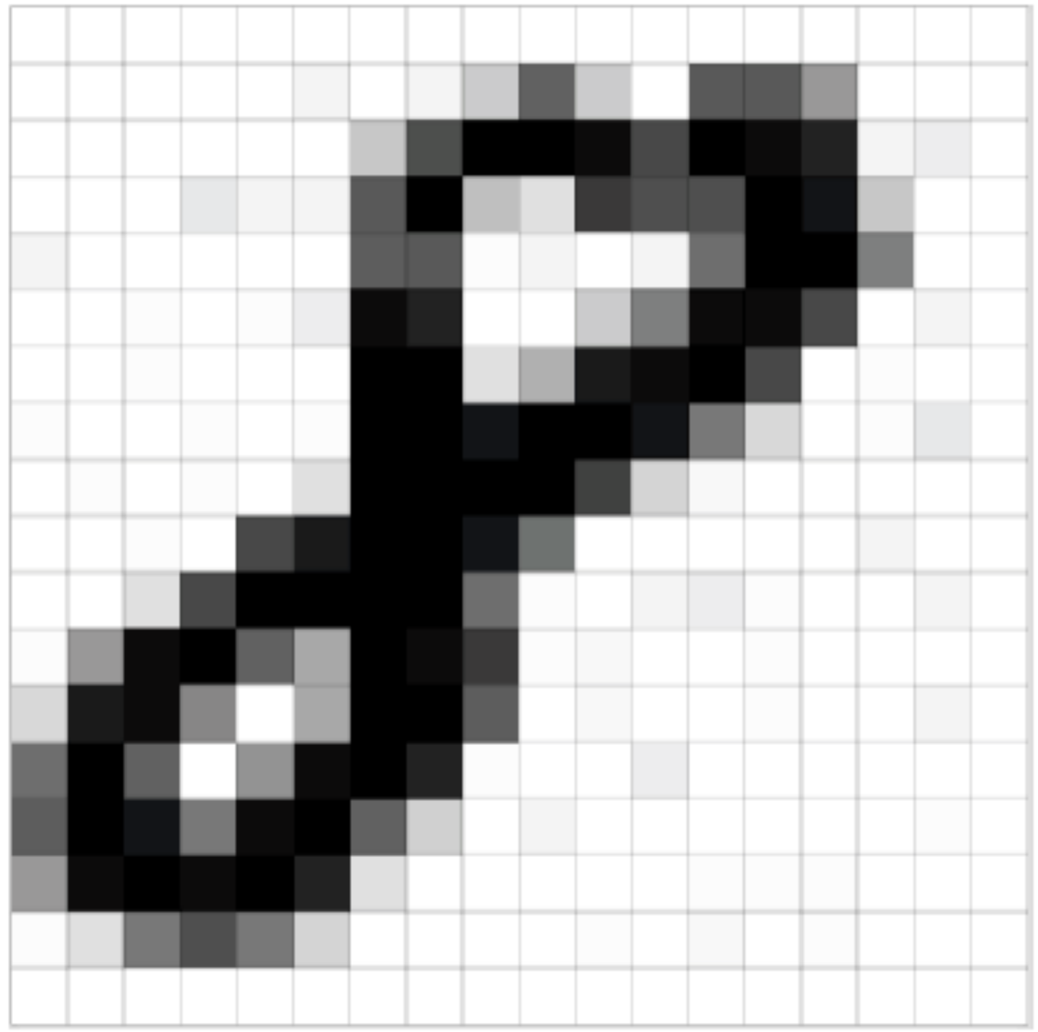
Key Operations in a CNN

There are 4 main operations in the CNN LeNet.

1. Convolution
2. Non Linearity (ReLU)
3. Pooling or Sub Sampling
4. Classification (Fully Connected Layer)

First of all, images are a matrix of pixel values

- A colored image will have three channels – red, green and blue
- A grayscale image has just one channel.

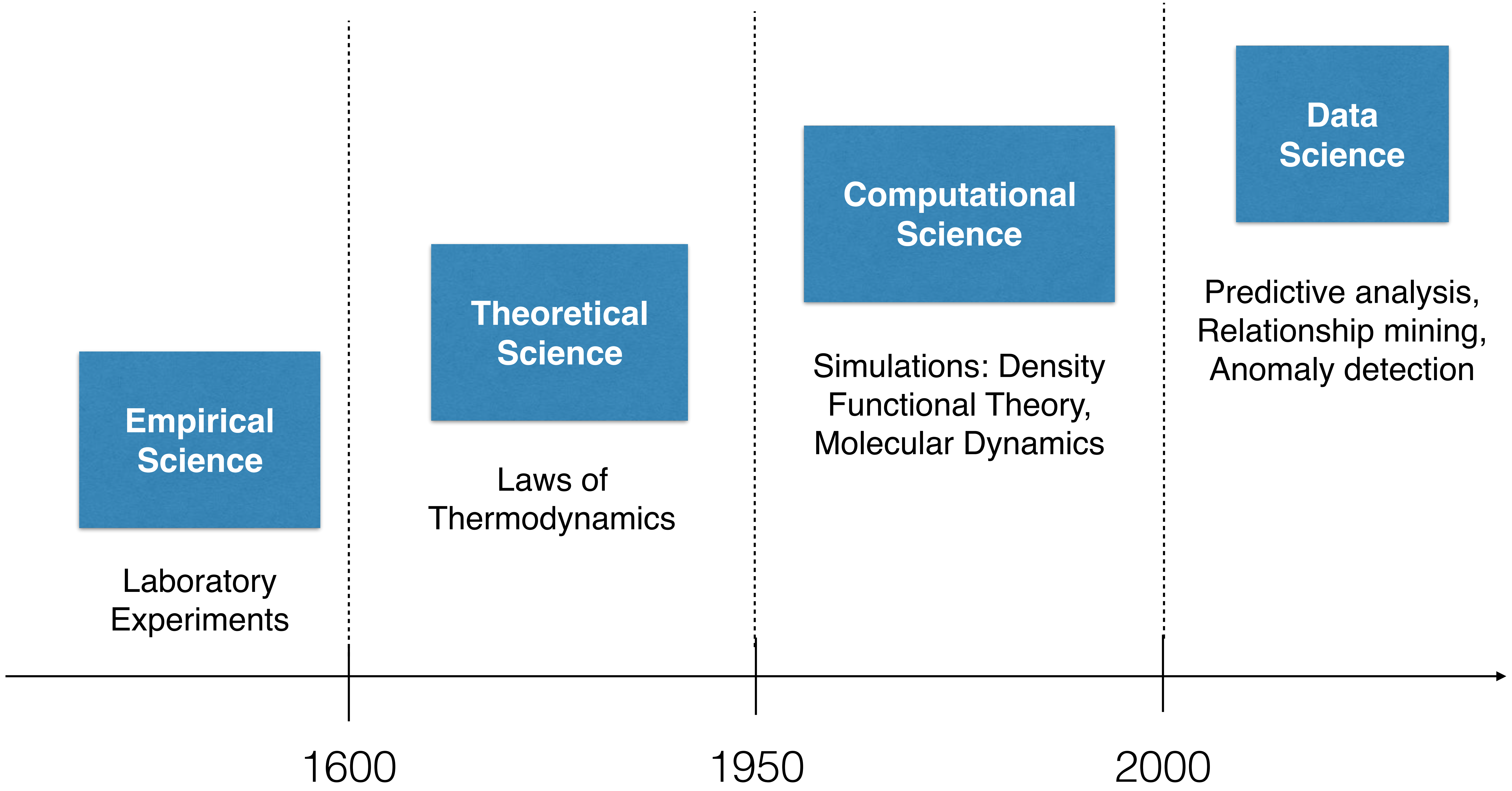


0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	12	0	11	39	137	37	0	152	147	84	0	0	0
0	0	1	0	0	0	41	160	250	255	235	162	255	238	206	11	13	0
0	0	0	16	9	9	150	251	45	21	184	159	154	255	233	40	0	0
10	0	0	0	0	0	145	146	3	10	0	11	124	253	255	107	0	0
0	0	3	0	4	15	236	216	0	0	38	109	247	240	169	0	11	0
1	0	2	0	0	0	253	253	23	62	224	241	255	164	0	5	0	0
6	0	0	4	0	3	252	250	228	255	255	234	112	28	0	2	17	0
0	2	1	4	0	21	255	253	251	255	172	31	8	0	1	0	0	0
0	0	4	0	163	225	251	255	229	120	0	0	0	0	0	11	0	0
0	0	21	162	255	255	254	255	126	6	0	10	14	6	0	0	9	0
3	79	242	255	141	66	255	245	189	7	8	0	0	5	0	0	0	0
26	221	237	98	0	67	251	255	144	0	8	0	0	7	0	0	11	0
125	255	141	0	87	244	255	208	3	0	0	13	0	1	0	1	0	0
145	248	228	116	235	255	141	34	0	11	0	1	0	0	0	1	3	0
85	237	253	246	255	210	21	1	0	1	0	0	6	2	4	0	0	0
6	23	112	157	114	32	0	0	0	0	2	0	8	0	7	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Outline

- Deep Learning: the Basics
- **Materials Discovery: the Basics**
- Deep CNN for EBSD Indexing
- Advanced Software and Hardware in Use

Materials Discovery: A Fourth Paradigm



Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. The fourth paradigm: data-intensive scientific discovery, volume 1. Microsoft research Redmond, WA, 2009.

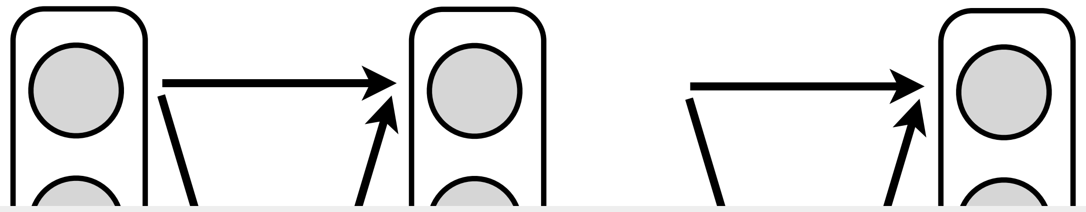
Ankit Agrawal and Alok Choudhary. Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science. *APL Materials*, 4(5):053208, 2016.

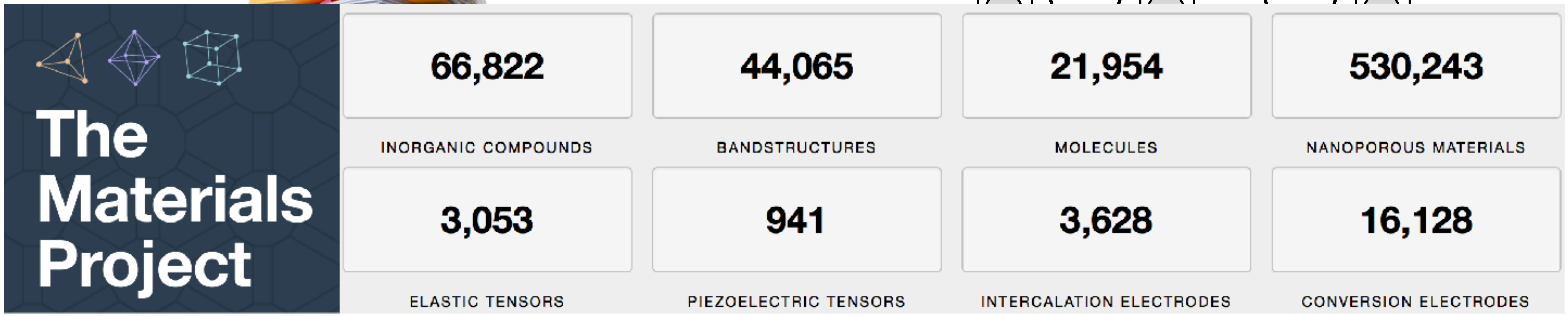
Deep Learning for Materials Science Applications

Fast computing?

Big data?

Deep neural networks?





Yes! (good enough)

Sure!

Materials Project:

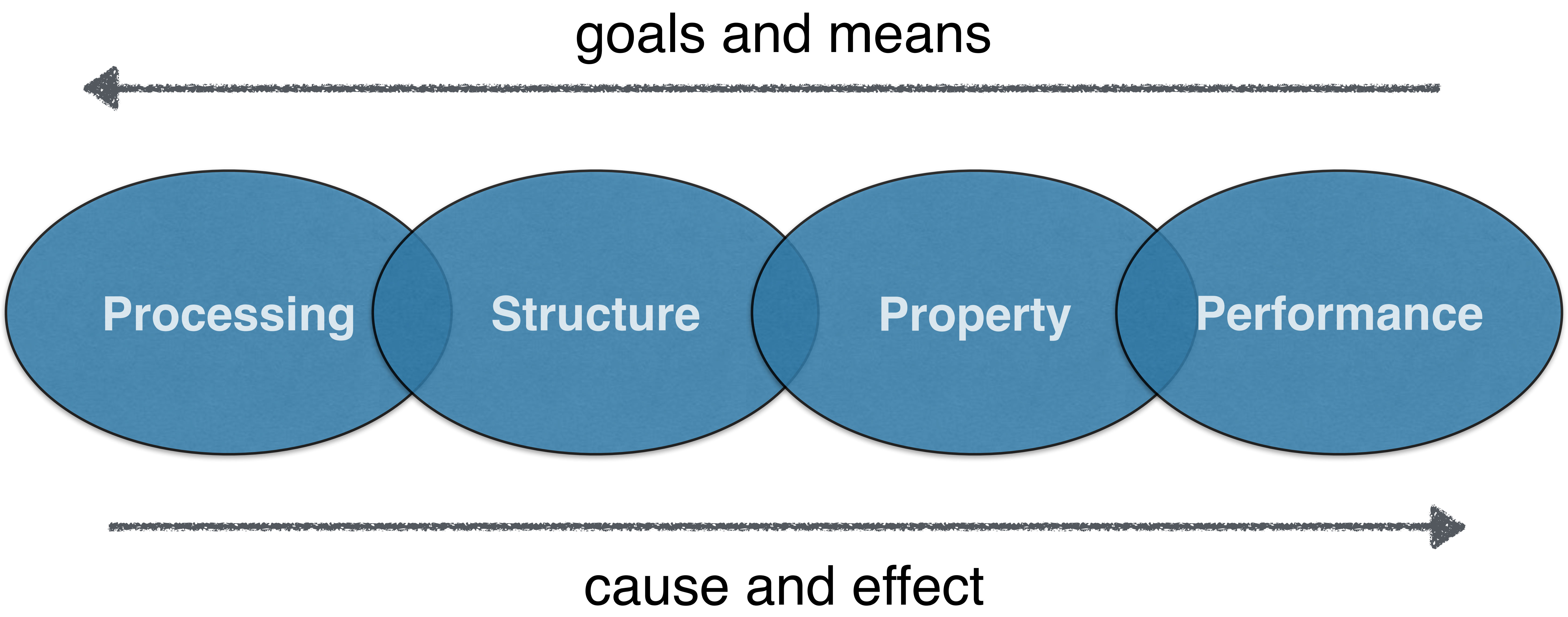
https://www.materialsproject.org

Open Quantum Materials Database:

http://www.oqmd.org

We are the first!

Materials Knowledge Discovery



Our specific problem in materials discovery:

Electron imaging (observation) $\Longrightarrow$ polycrestal constitution (cause)

Outline

- Deep Learning: the Basics
- Materials Discovery: the Basics
- **Deep CNN for EBSD Indexing**
- Advanced Software and Hardware in Use

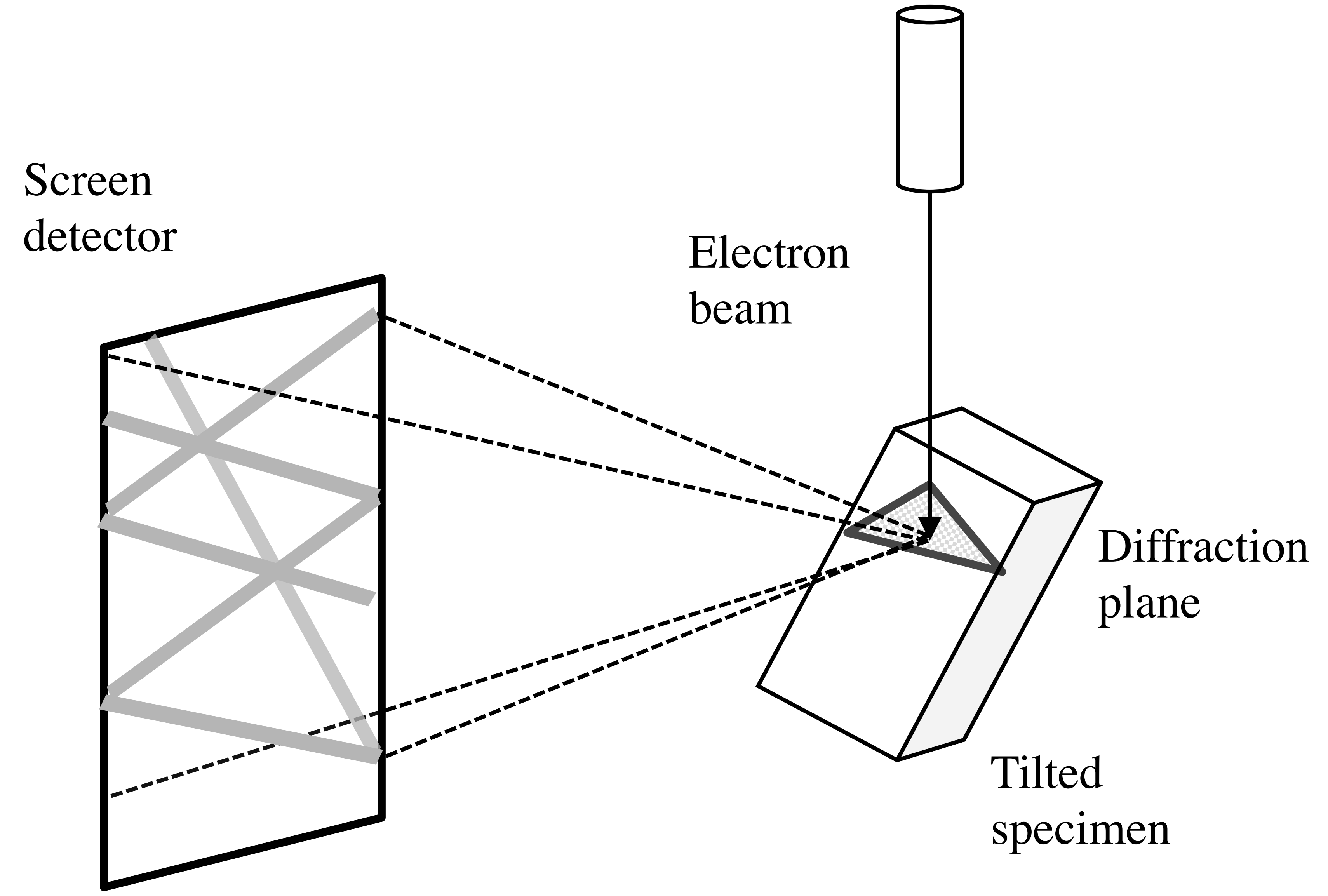
Problem: EBSD Indexing

Electron backscatter diffraction (EBSD) is a standard technique detecting certain microstrcture characteristics on the surfaces of polycrystals.

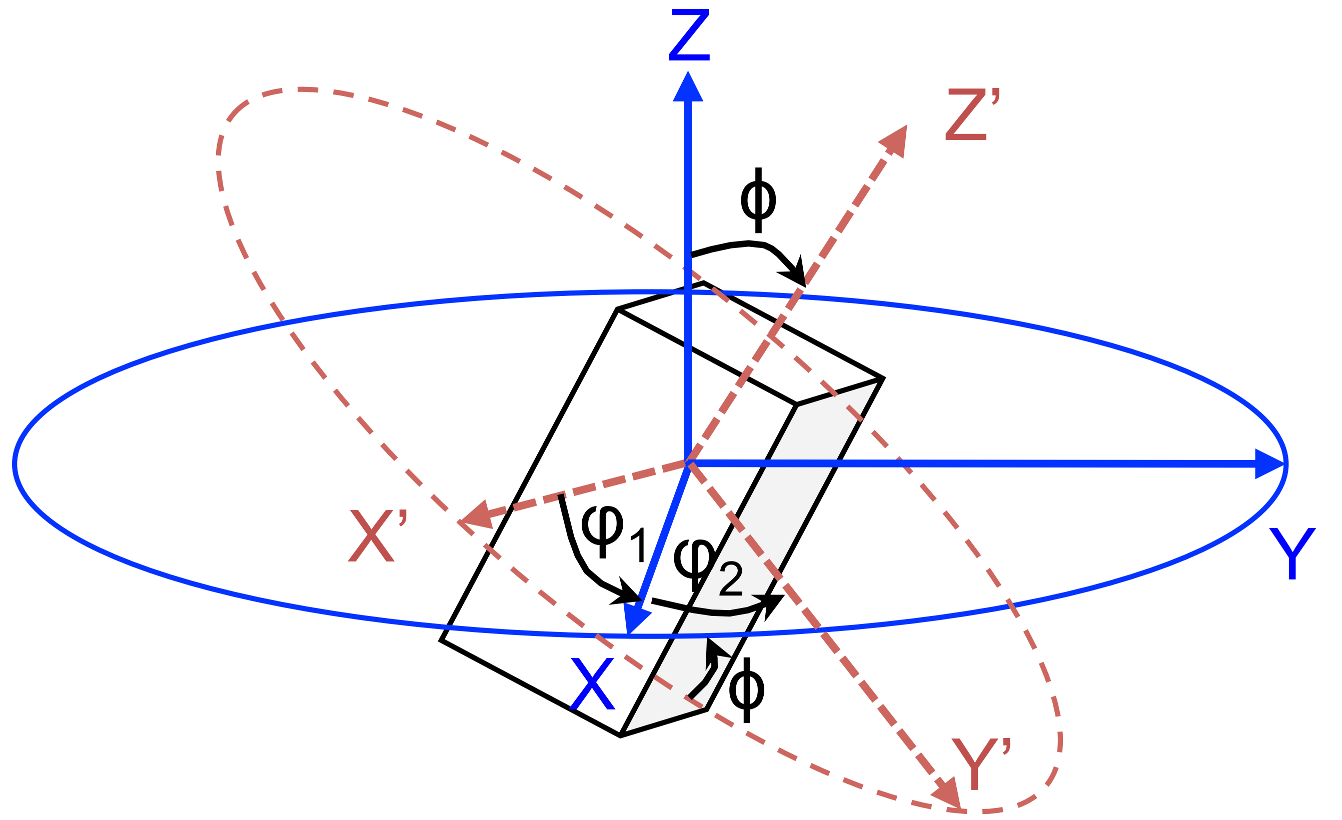
Traditional approach to EBSD indexing is pattern matching.

- Store every distinct pattern-angle pair
- build a pattern-angle dictionary
- When a new pattern is observed, it is looked up in the dictionary and the orientation of its 1-Nearest Neighbor (1-NN) is returned

High cost at prediction time!



(a)



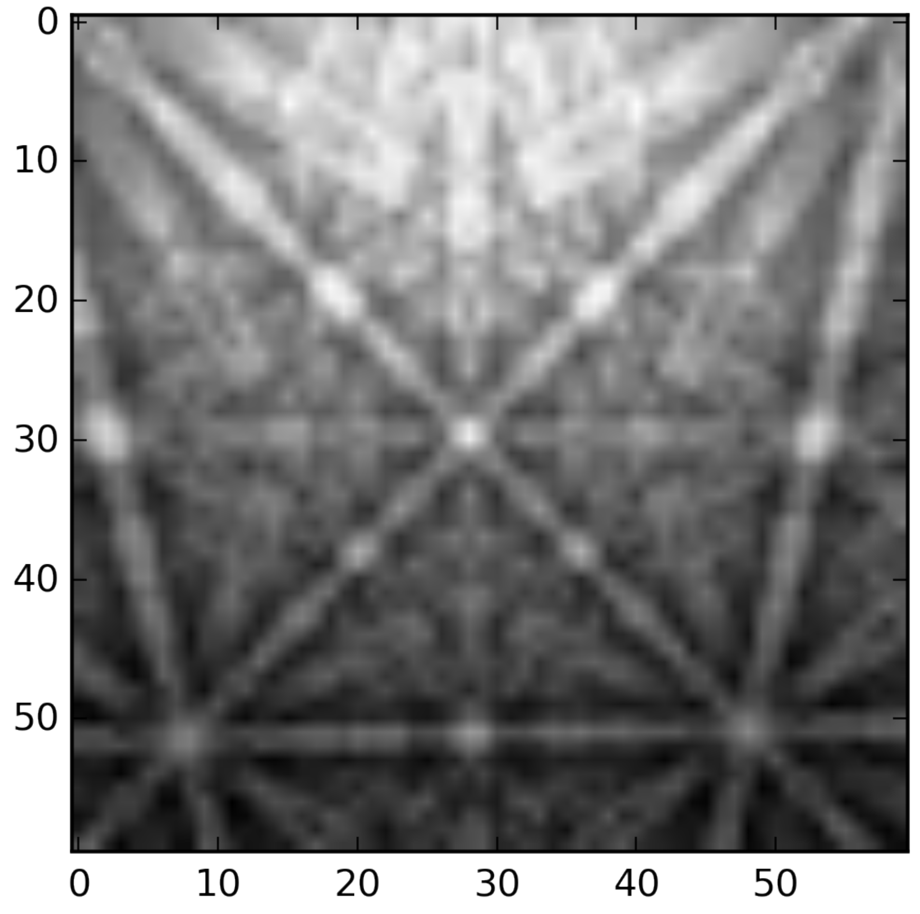
(b)

EBSD Patterns

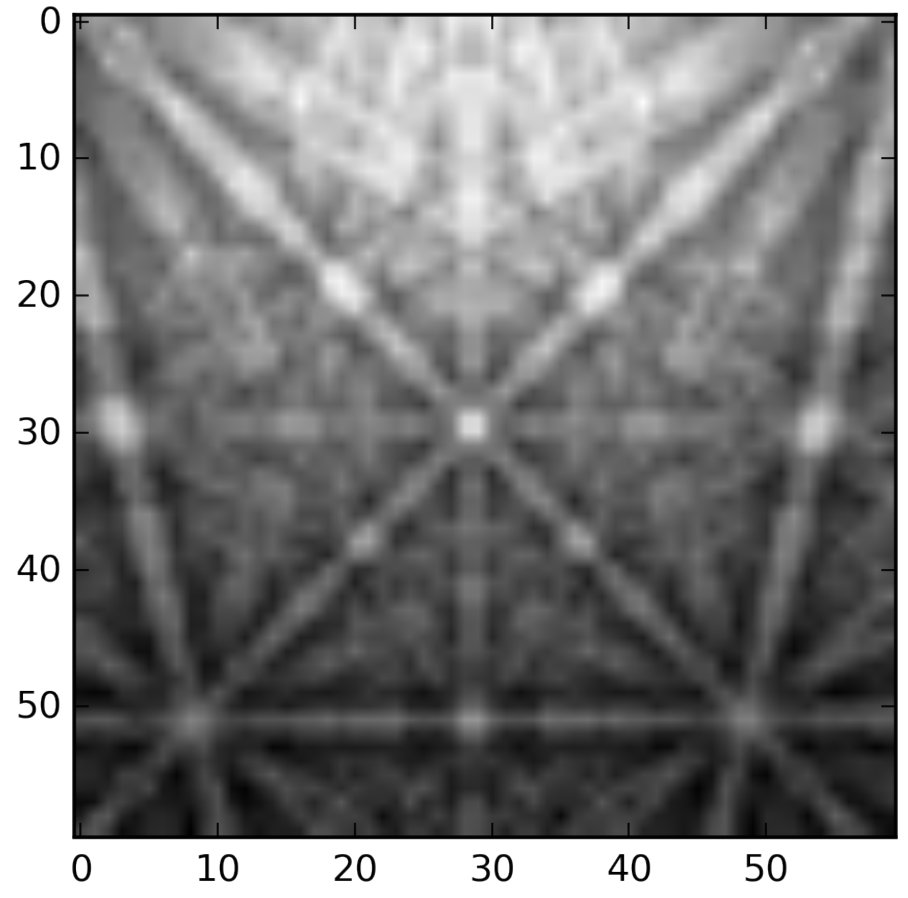
Electron backscatter diffraction (EBSD) is a standard technique detecting certain microstrcture characteristics on the surfaces of polycrystals.

Traditional approach to EBSD indexing is pattern matching.

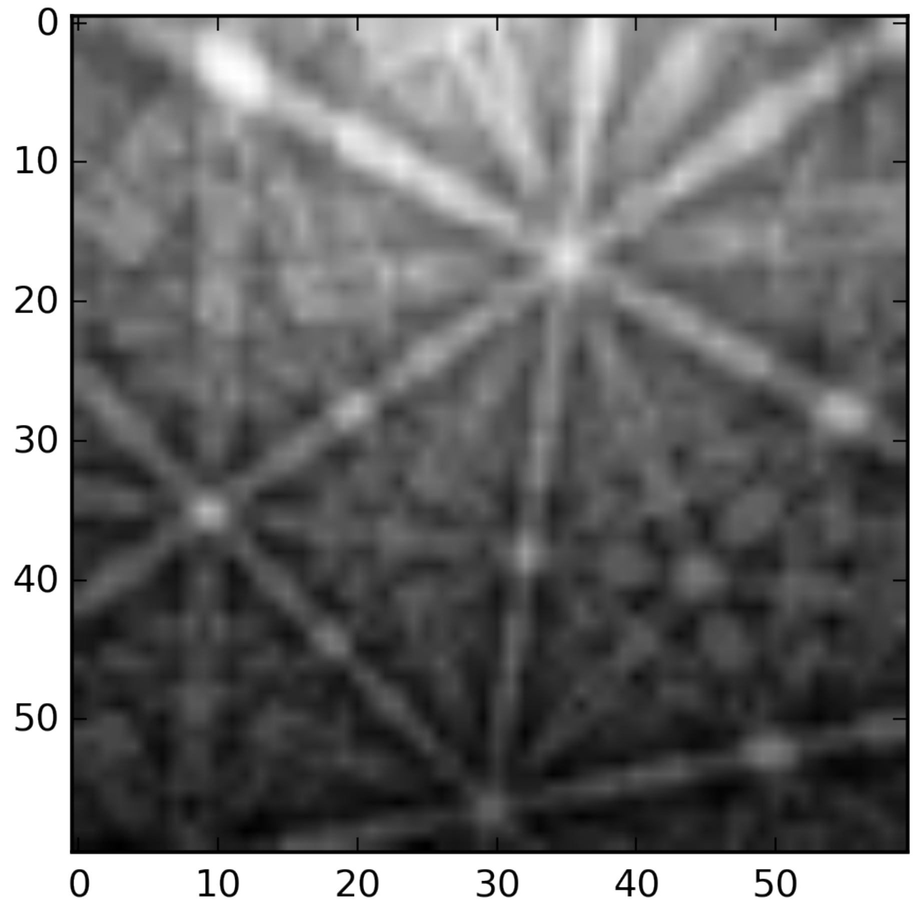
- Store every distinct pattern-angle pair
- build a pattern-angle dictionary
- When a new pattern is observed, it is looked up in the dictionary and the orientation of its 1-Nearest Neighbor (1-NN) is returned



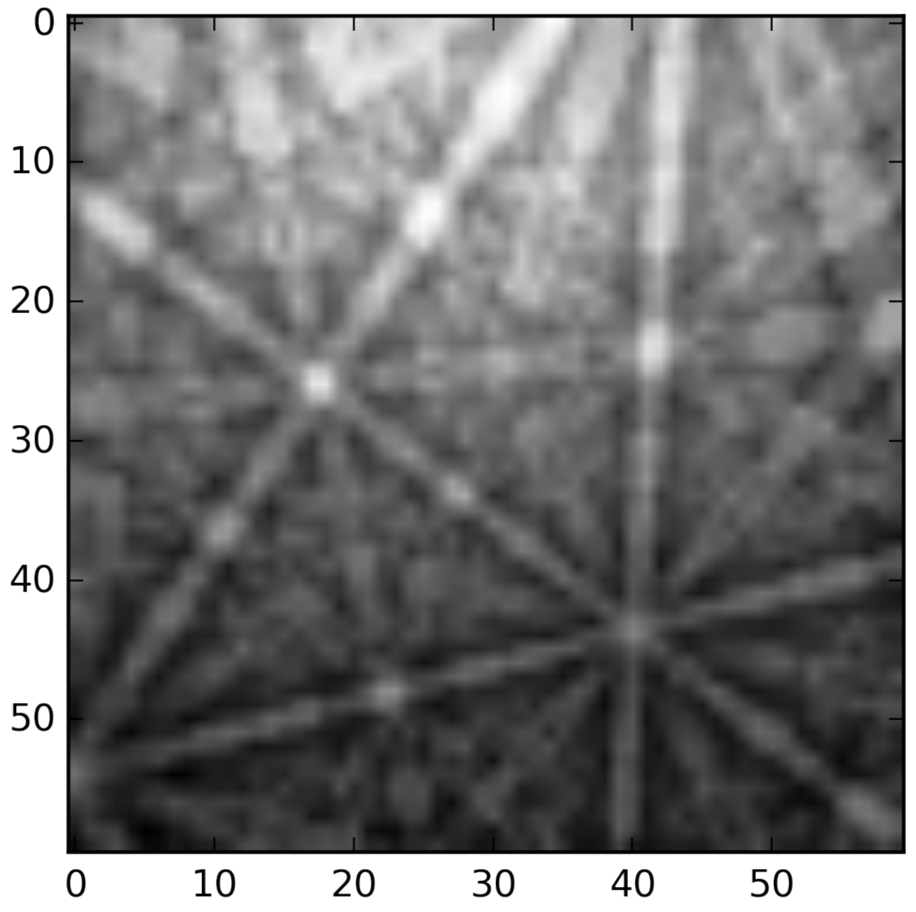
(214.45, 58.86, 124.45)



(214.94, 59.17, 124.94)

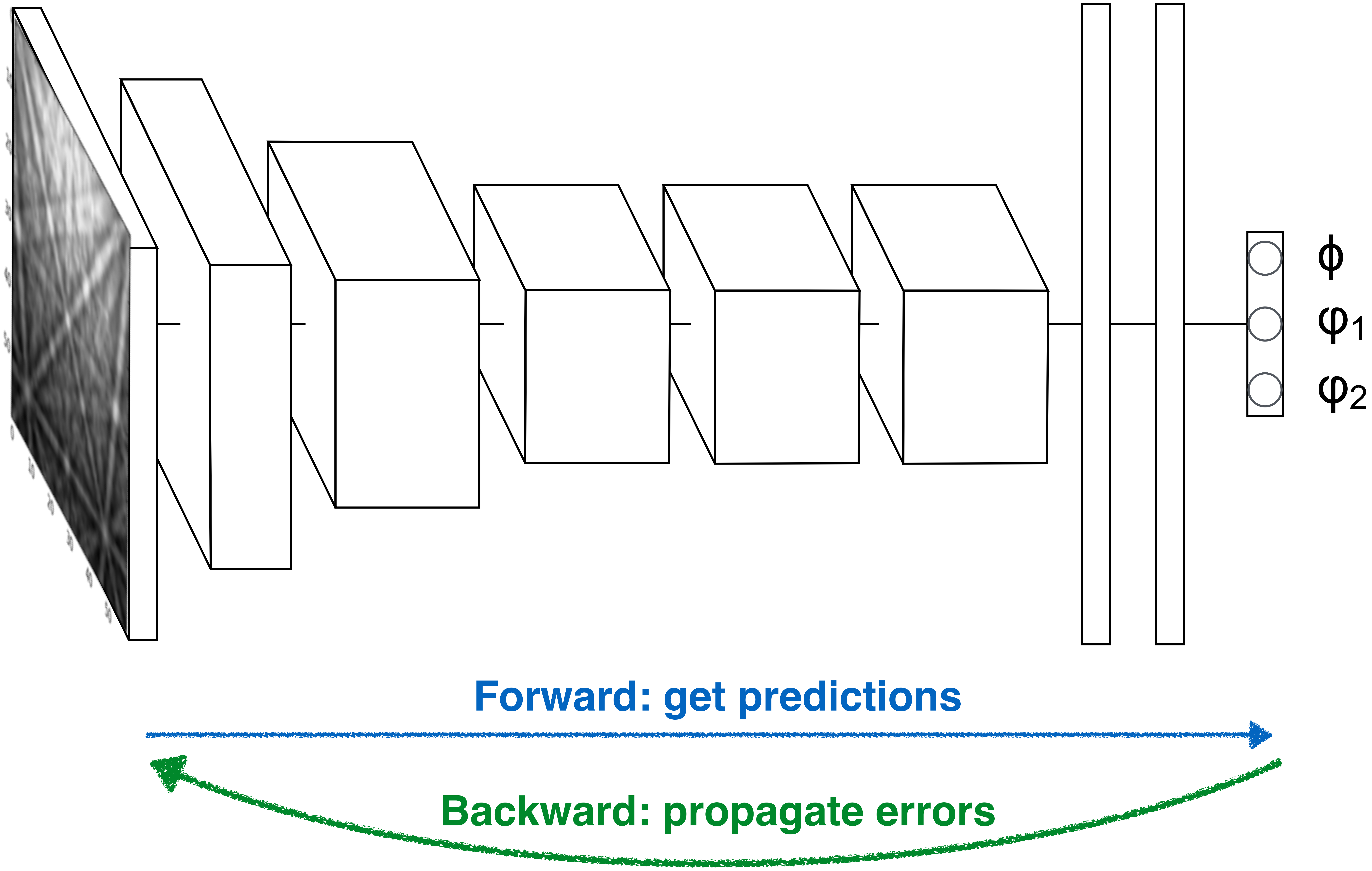


(222.43, 44.23, 182.46)



(197.23, 50.39, 127.96)

Deep Learning Solution



Loss Function

- Usually in a regression problem the loss to be propagated is the Mean Square Error (MSE) between predicted value and target value.

- However in our problem the target variable is an angle; the periodicity of angular data has to be addressed.

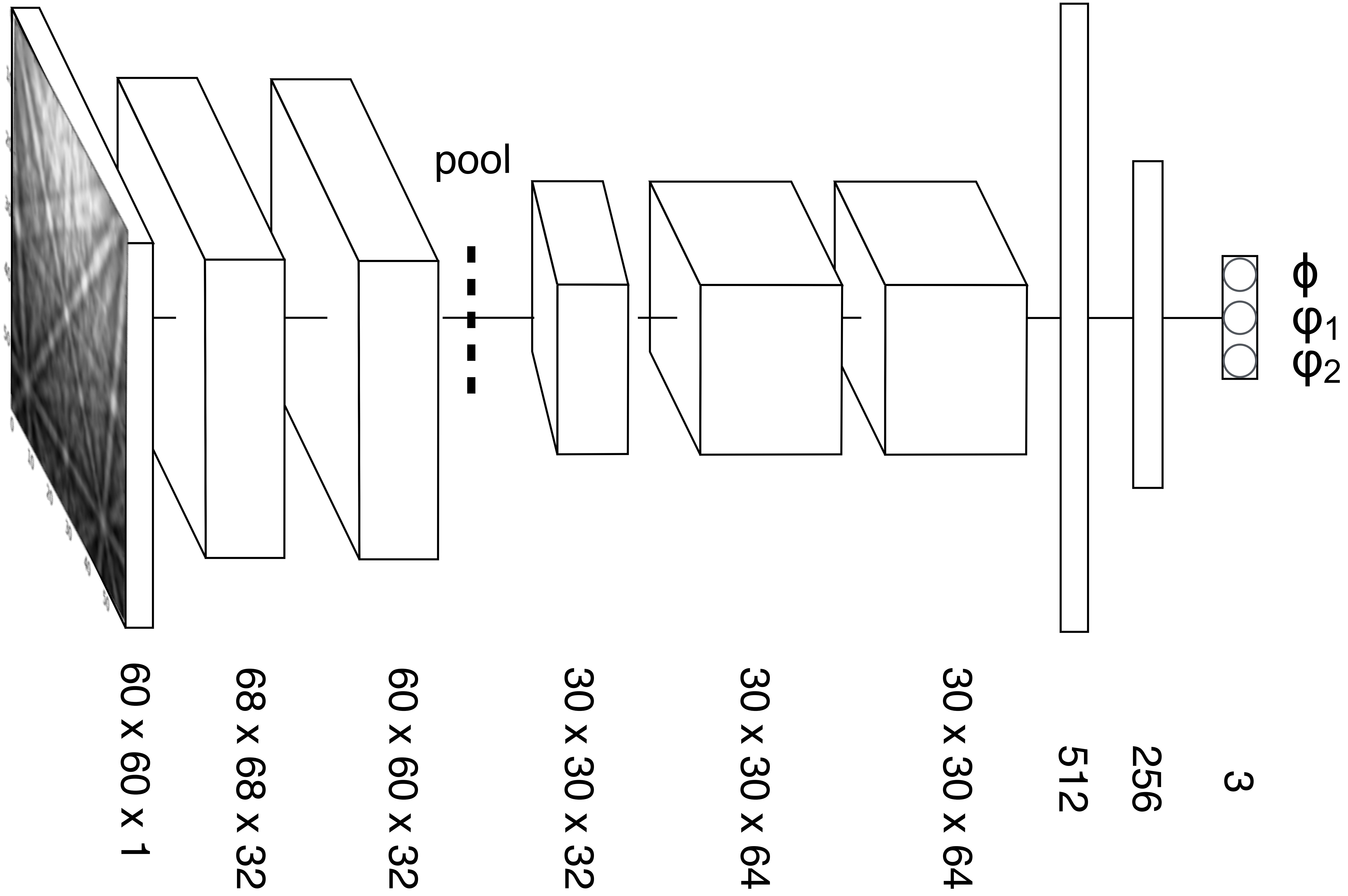
- We designed a special loss function to account for the fact that in angular values 0° is close to 359

$$L_i(y_i, \hat{y}_i) = \arccos(\cos(||y_i - \hat{y}_i||))$$

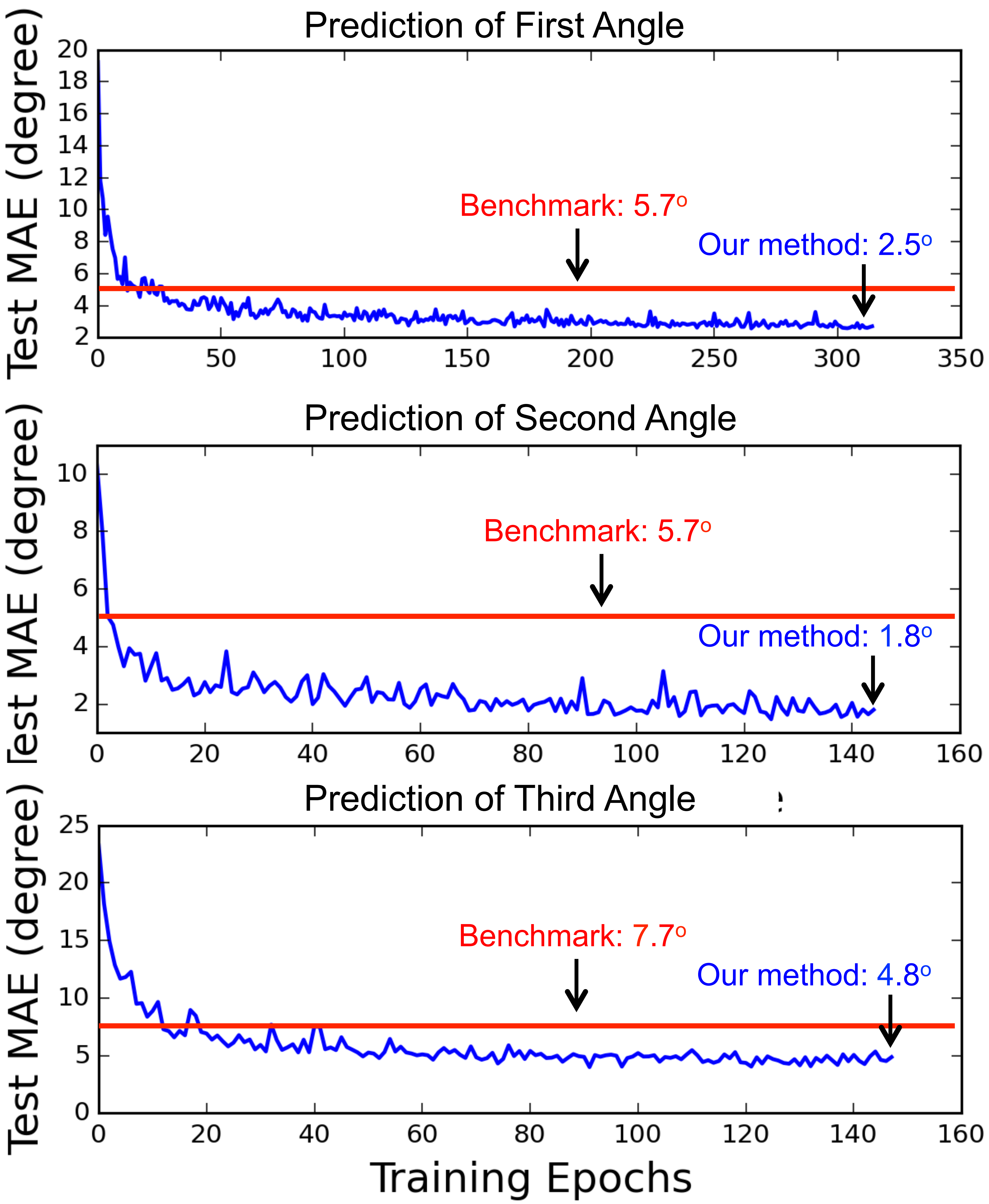
predicted angle

true angle

Network Configuration



Prediction Results



On average we are 54% better than state-of-the-art benchmark*

* Y. H. Chen, S. U. Park, D. Wei, G. Newstadt, M. A. Jackson, J. P. Simmons, M. De Graef, and A. O. Hero, "A dictionary approach to electron backscatter diffraction indexing," Microscopy and Microanalysis, vol. 21, no. 03, pp. 739–752, 2015.

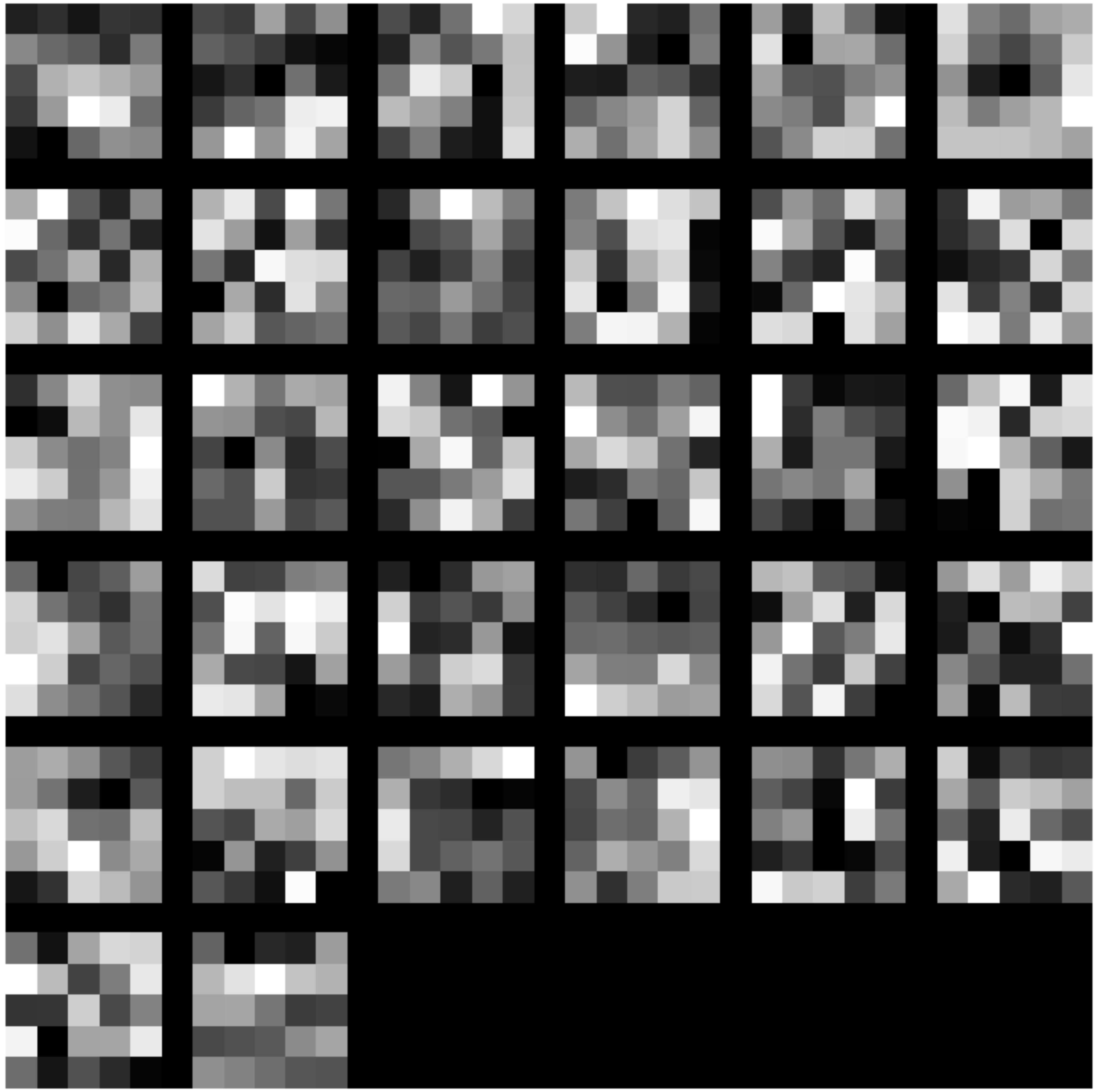
Time Results

Predictor	MAE (eV / atom)	Training time (entire data)	Run time (entire data)
1-NN	1.299	0	375s
DNN	0.072	7 days	50s

The 1-NN runtime at testing is calculated with all test data (30k) processed in one batch, which will produce with the fastest speed with maximum memory consumption.

The DNN runtime at testing is calculated with a batch size of 1k to balance the memory usage and time consumption.

Weight Visualization

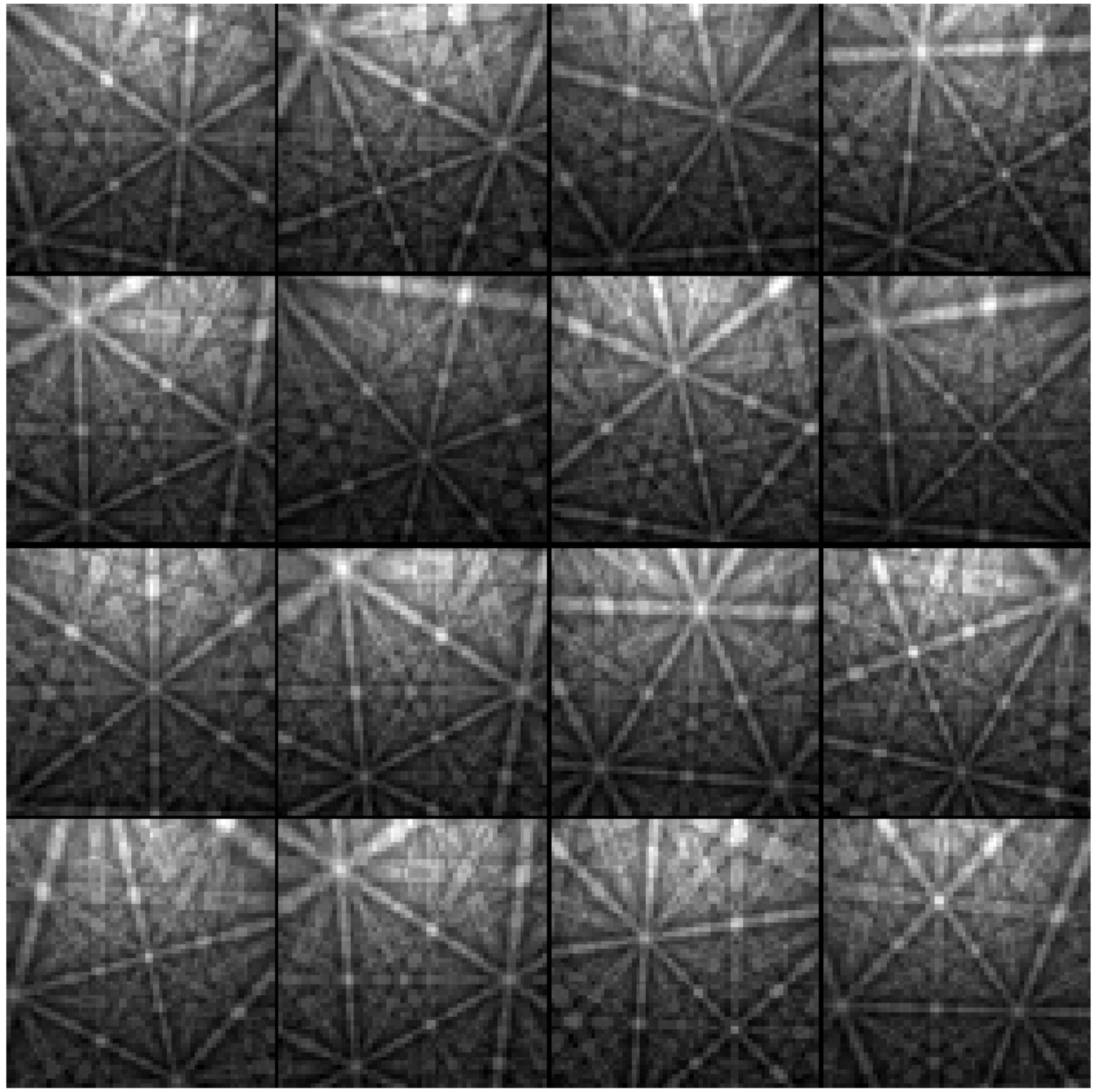


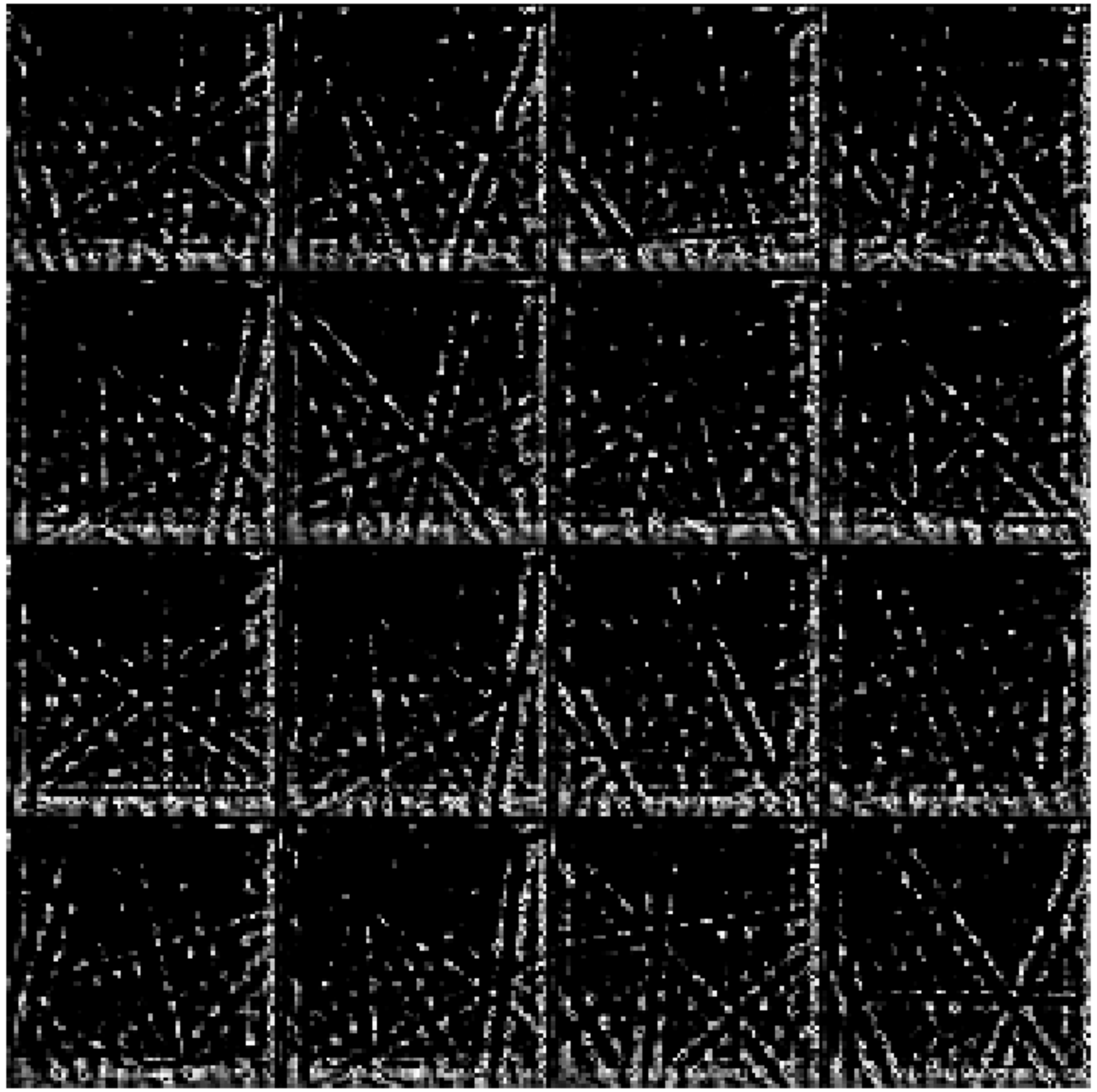
A total of 32 Filters learned by the first Conv layer in our CNN from EBSD images.

Black: weight value towards 0 - not important

White: weight value towards 1 - significant

Activation Visualization





Important information retained while noisy background discarded.

Outline

- Deep Learning: the Basics
- Materials Discovery: the Basics
- Deep CNN for EBSD Indexing
- **Advanced Software and Hardware in Use**

Advanced Software and Hardware

Hardware

- DIGITS DevBox, featuring:
  - 4 TITAN X GPUs with 12GB of memory per GPU
  - 64GB DDR4
  - Core i7-5930K 6 Core 3.5GHz desktop processor
  - 3 x 3TB SATA RAID5
  - 250GB SSD

Software

- CUDA
- cuDNN
- Theano

Conclusion

- We presented the first large-scale, big data application of deep learning in materials discovery
- Deep convolutional neural networks achieved state-of-the-art accuracy in an electron imaging indexing problem.
- Advanced software frameworks and hardware infrastructure are adopted.

Materials Discovery: Understanding Polycrystals from Large-Scale Electron Patterns

3rd Workshop on Advances in Software and Hardware (ASH) at BigData

December 5, 2016

Ruoqian Liu, **Dipendra Jha**, Ankit Agrawal, Wei-keng Liao, Marc De Graef, Alok Choudhary

Northwestern University

Carnegie Mellon University

Thank you!