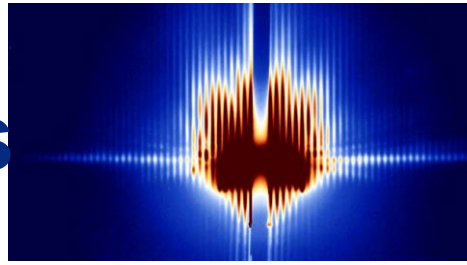
Deep Learning for Analyzing NSLS-II Data Stream

Dantong Yu[1](presenter) BNL Computational Science Initiative
Boyu Wang[2], Ziqiao Guan[2], Stony Brook University
Kevin Yager, Jiliang Liu, BNL Center for Functional Nanomaterials
Ronald Lashley, Bo Sun, Lincoln University

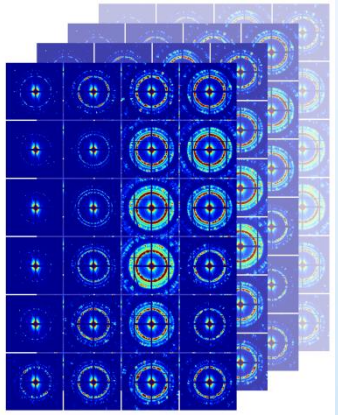
BROOKHAVEN NATIONAL LABORATORY
a passion for discovery

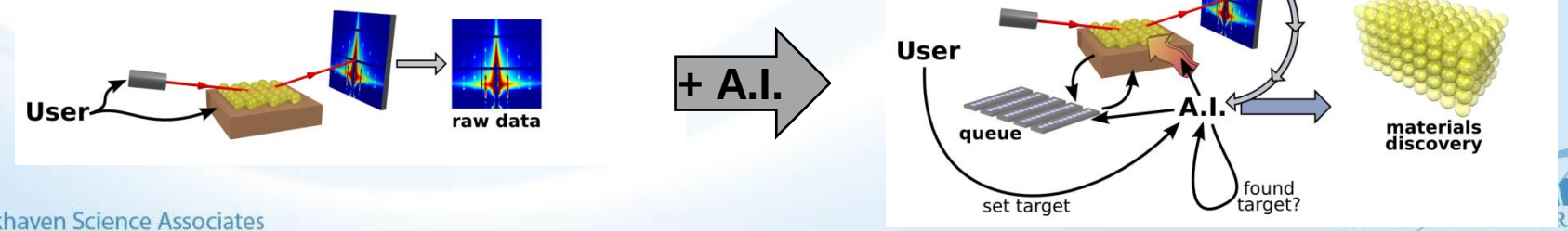
U.S. DEPARTMENT OF ENERGY | Office of Science

Research Goals and Objectives



- Modern scientific experiments (e.g. NSLS-II, CFN) generate massive amounts of data.
- Complex data analysis consumes scientists' precious time, distracting from deep scientific questions.
- We propose a transformative **autonomous experimentation** paradigm, where data acquisition, data analysis, and experimental decision-making are automated. This **liberates the human scientist to focus on science**.
- This work will develop **deep learning** algorithms and methods for extracting hierarchical and physically-meaningful insights from scientific datasets collected at NSLS-II and CFN.



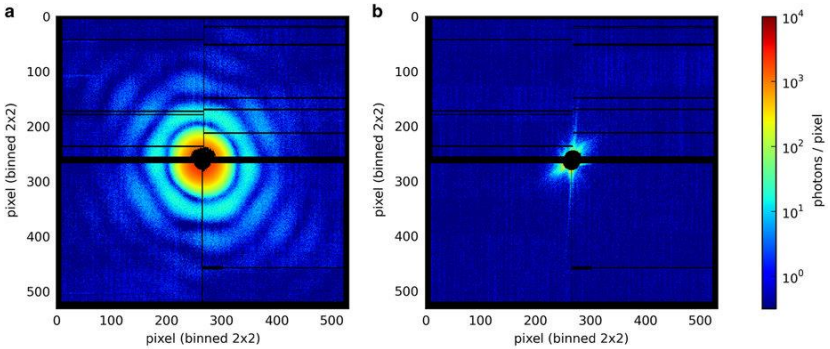


Outline

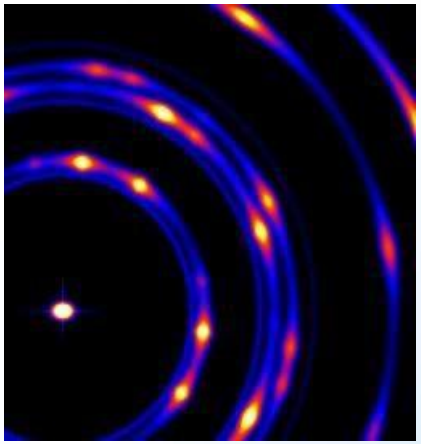
- Overarching Goals and Objectives
- Standard Image Analysis Methods
- Our Proposed Method Based on Deep Learning
- Graphical Interface and Demo of Using Deep Learning to Analysis Real Images
- Conclusions and Future Works

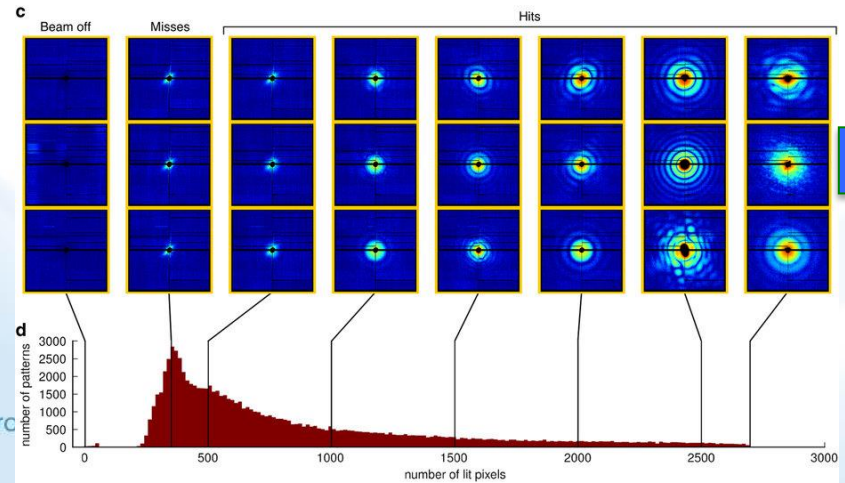
Traditional Approach

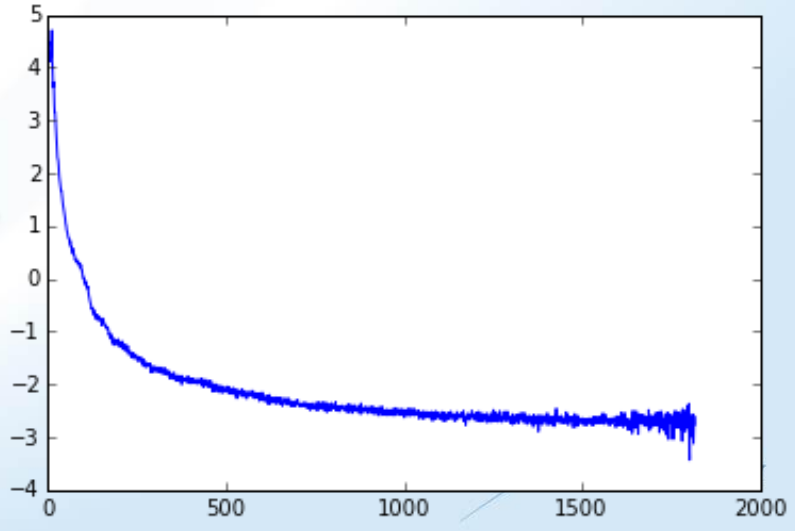
- Data Preprocessing ➔ Raw image ➔ Binning on q-ring, obtaining the average value of q-ring intensity, variations, normalized intensity.



What About

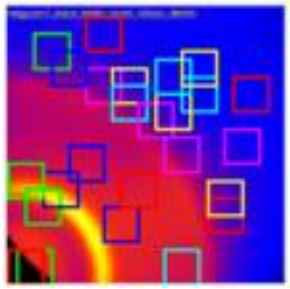




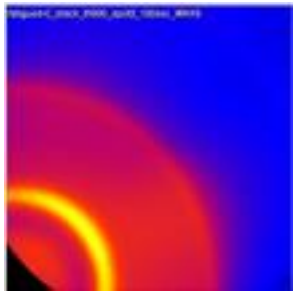


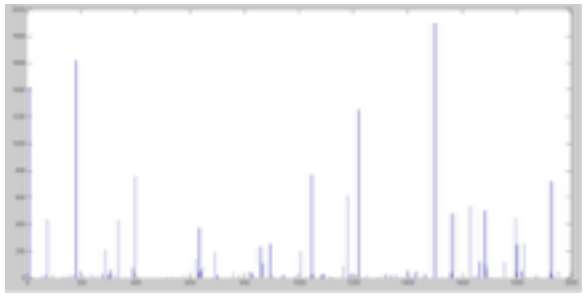
Baseline Algorithm Design with Hand-Crafted Feature Descriptors

- Extract small patches from the scattering images
- Build a dictionary about all those patches
- Perform Spatial Pyramid Matching
- Use SVM for classifier
- Mean Average Precision (mAP) on synthetic data: 67.05%

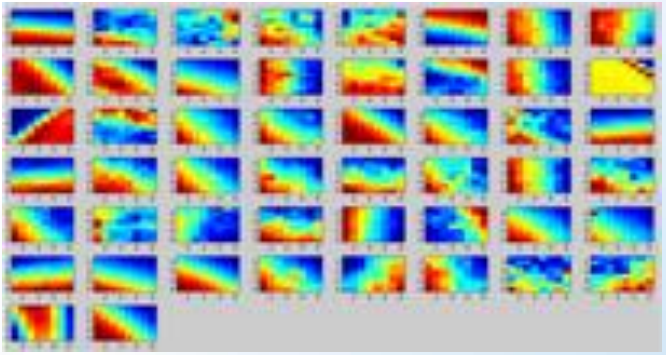


Random small patches

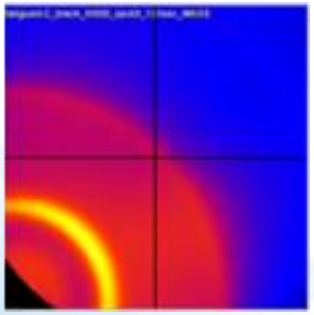


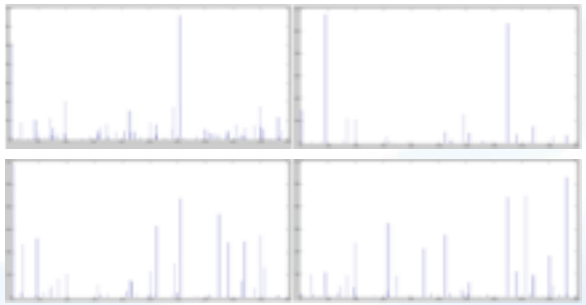


Spatial level 0 histogram



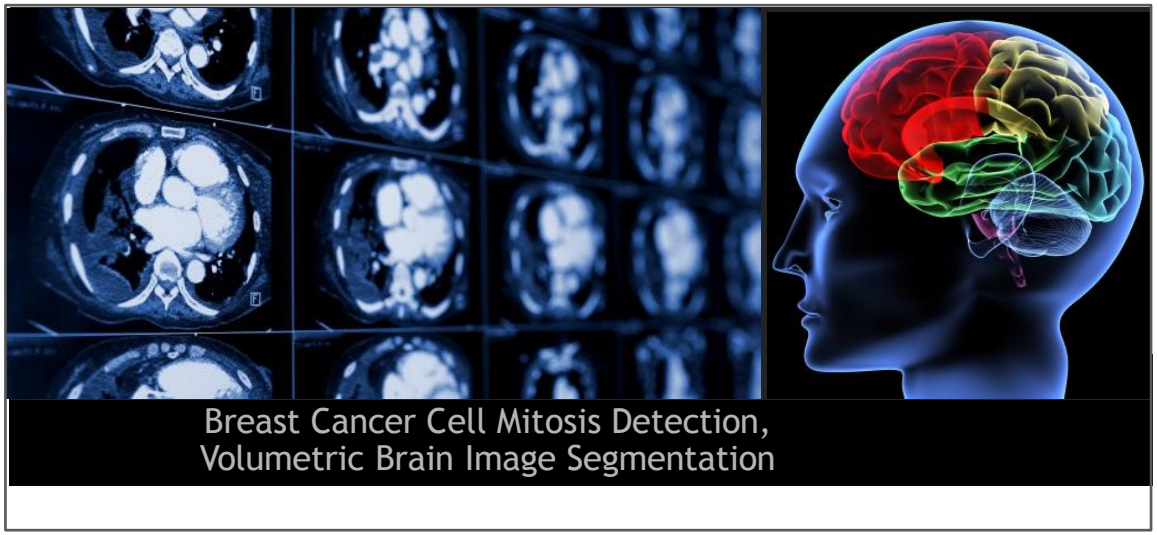
visual dictionary created from those patches





Spatial level 1 histogram

Deep Learning for On-Line Recognition and Detection



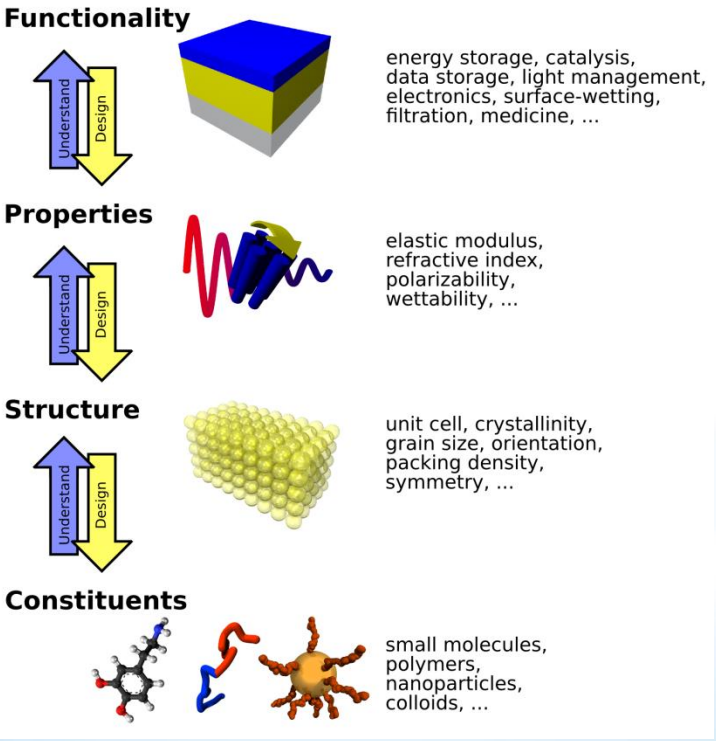
Breast Cancer Cell Mitosis Detection,
Volumetric Brain Image Segmentation

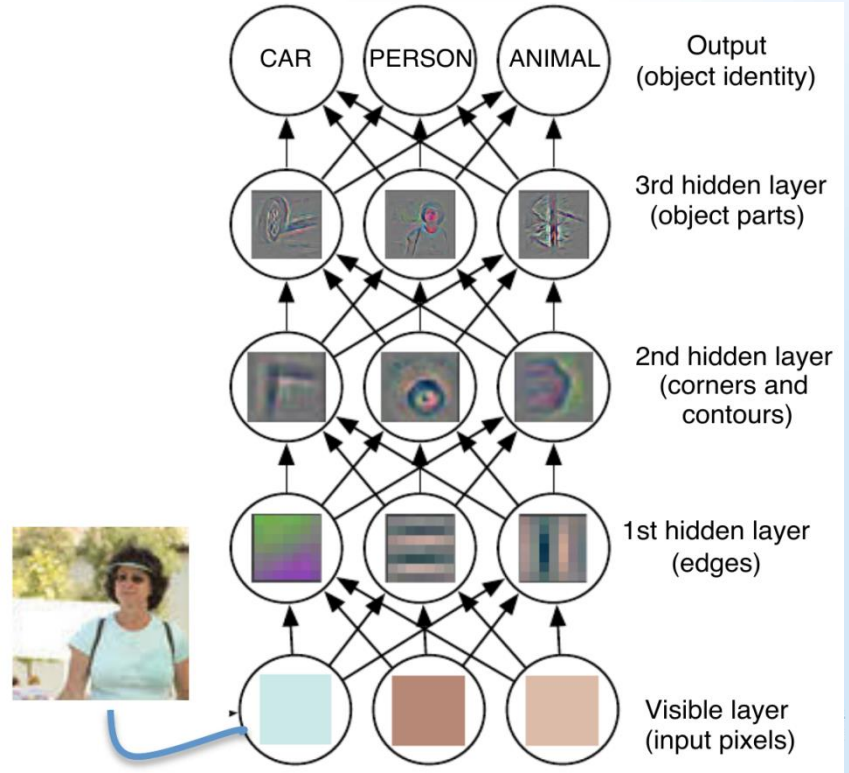
Pedestrian Detection, Traffic Sign Recognition

- Off-line Training, On-line detection ➔ On-line Training, on-line detection
- Incremental Update to Existing Training Model
- On-line optimization
- Fit Squarely with Image Stream Analysis and Experiment Steering

Proposed Method and Approach

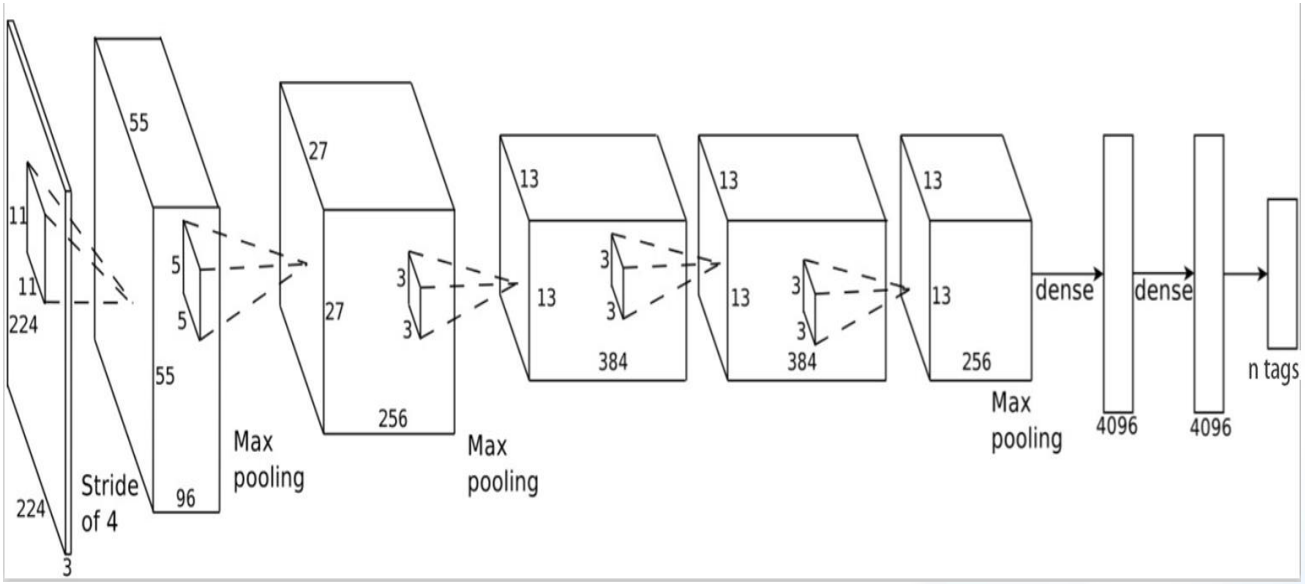
- Develop data analysis pipeline for x-ray scattering data, where outputs are useful to scientists (physically-meaningful) and as inputs to machine-aided experiment steering.
- Develop a deep learning hierarchy that extracts features that have domain-specific relevance.
- Combining ‘physics’ with ‘machine learning’ will yield improved results.





Convolution Neural Networks

- We started with ALEXNET



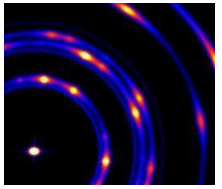
Progress Status

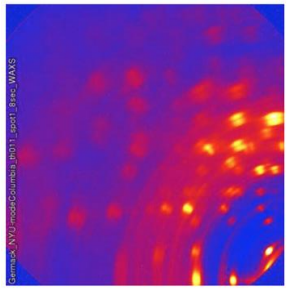
- Developed software to output synthetic x-ray scattering data, to use as train/test set for machine learning.
  - Generated abundant simulation data for Transfer Learning.
- Augmented Alexnet (A deep learning CNN)and Residual Networks in Google TensorFlow on scattering data. Parameter tuning remains outstanding challenge.
- The output will be the probability of image carrying an attribute (tag)

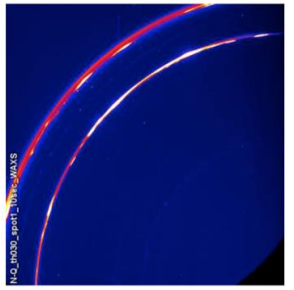
  Multi attribute classification: Each image could have multiple tags: images (ring, halos, diffuse, textured) and physics properties (disordered, single crystal, amorphous)

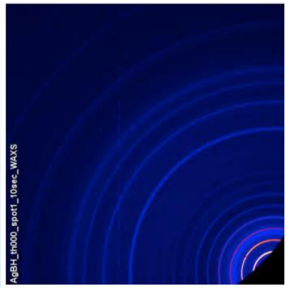
  The output function is different from that of the standard CNN, has to be fine tuned for convergence.

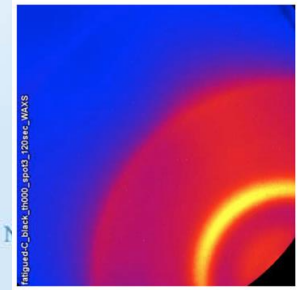
  Robust networks to handle Imbalanced dataset (some tags only have few images)





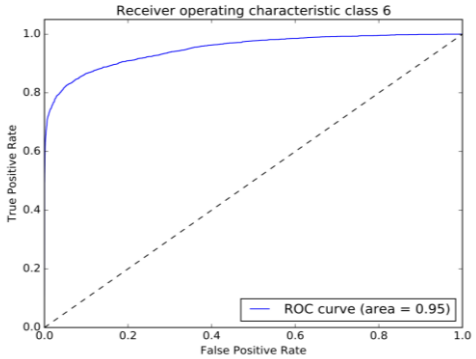




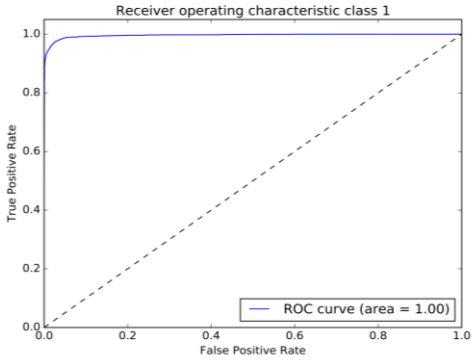


Results of Deep Learning based Approach

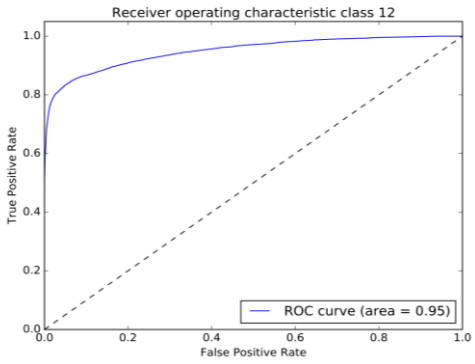
- Train an Alex Net (5 convolutional layers)
- mean Average Precision (mAP) on synthetic data: 77.10%



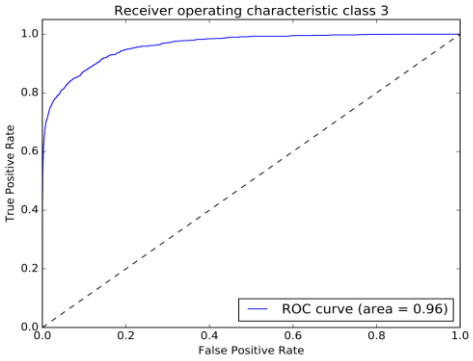
ROC curve on tag: Halo



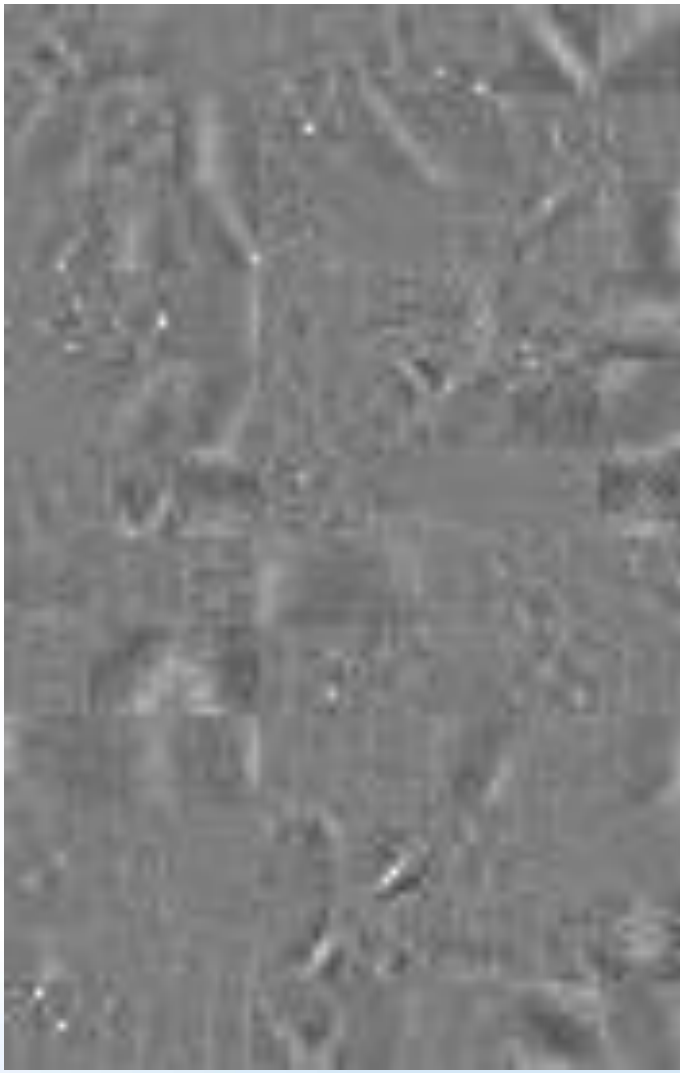
ROC curve on tag: Beam Off Image



ROC curve on tag: Ring



ROC curve on tag: Diffuse high-q

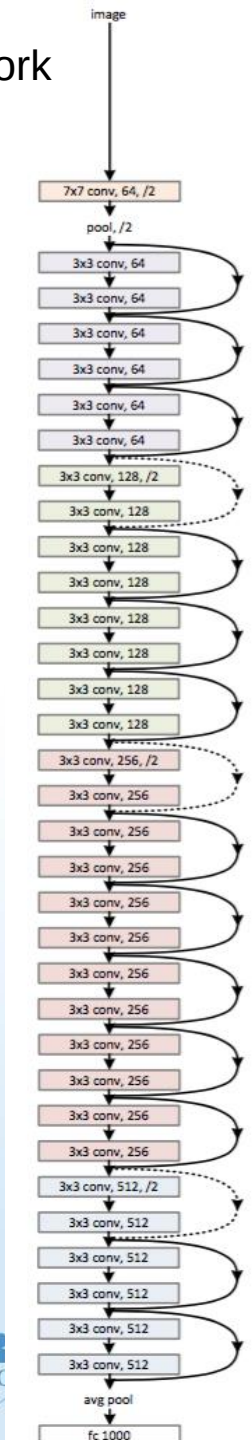


Visualization of first layer filter

Research Topics: Transfer Learning

- Transfer the model learned on synthetic data to real experimental data
- Why not directly train the model on real data?
  - Limited amount of real data
  - Collecting real data are both time and cost intensive
  - Real data need experts labeling
- For each real experimental data, extract the last Fully Connected (FC) layer features from the Residual Network (50 layers), which is trained on synthetic dataset
- Train an SVM on those FC features
- Mean Averaged Performance (mAP) on real dataset: 67.04%

Deep Residual Learning for Image Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, IEEE Conference on Computer Vision and Pattern



Research Topic 2: Incrementally retrain the network

- The network is able to adjust itself with new data stream
- No need to start from scratch. Start from the checkpoint and fine-tune the network with new data stream
- The reason behind this is that the training algorithm of CNN is based on Stochastic Gradient Descent

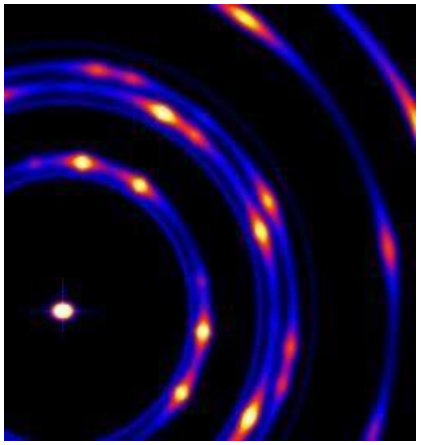
$$\nabla f^{(j)} = \frac{\partial f(w,b)}{\partial w^{(j)}} = w^{(j)} + C\sum_{i=1}^{n} \frac{\partial L(x_i, y_i)}{\partial w^{(j)}} \qquad \nabla f^{(j)}(x_i) = w^{(j)} + C \cdot \frac{\partial L(x_i, y_i)}{\partial w^{(j)}}$$

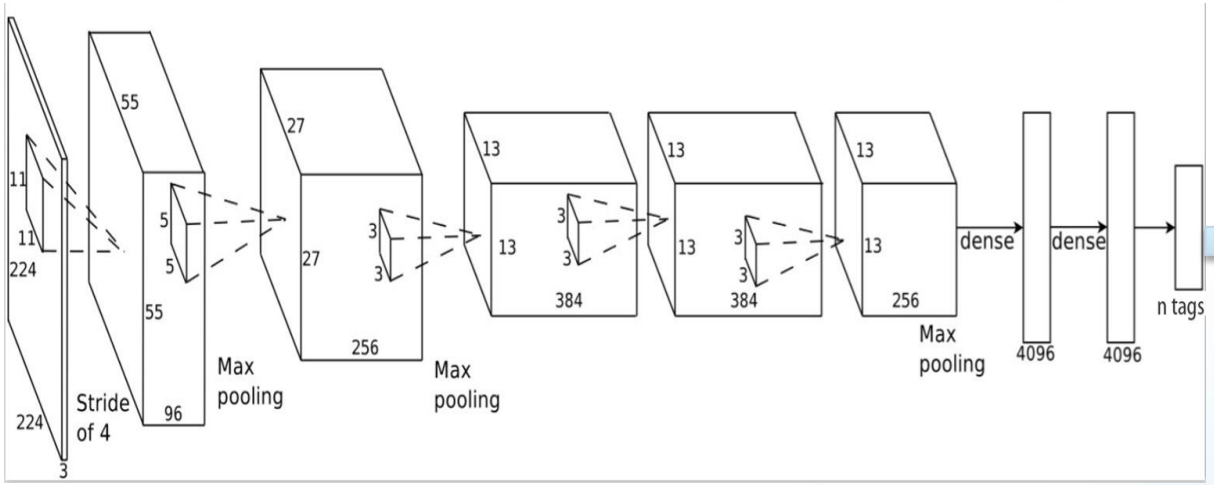
- The network can be updated by just computing the gradient from a mini-batch of data

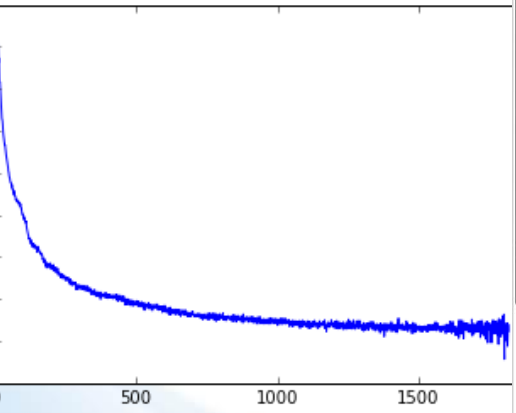
Preliminary results:

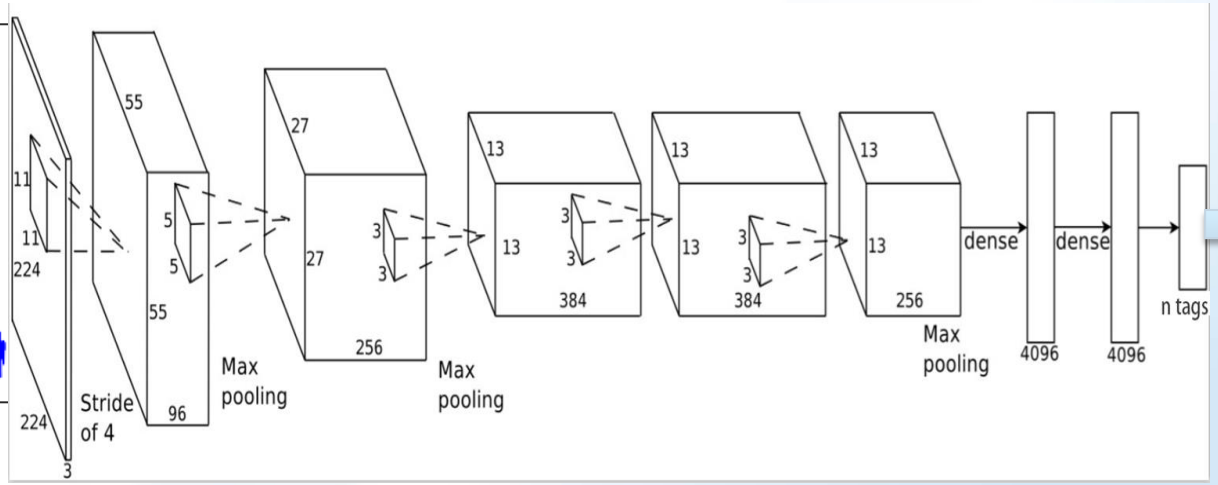
- Network trained purely on synthetic data
- Fine-tune the network on some real experimental data
- mAP (on real experimental data) improves from 50.37% to 52.11%
- Fine-tune can improve the performance

Topic 3: Physics-Aware Deep Learning





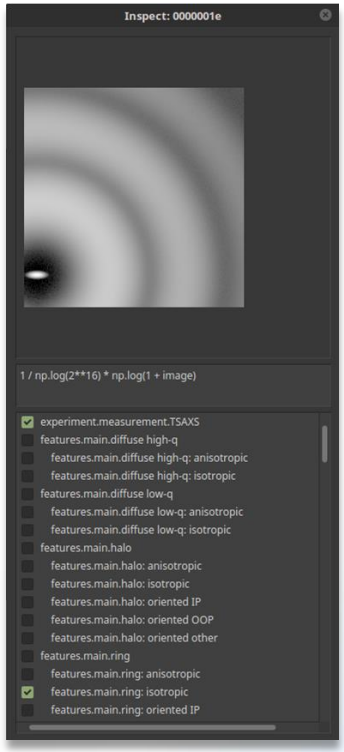


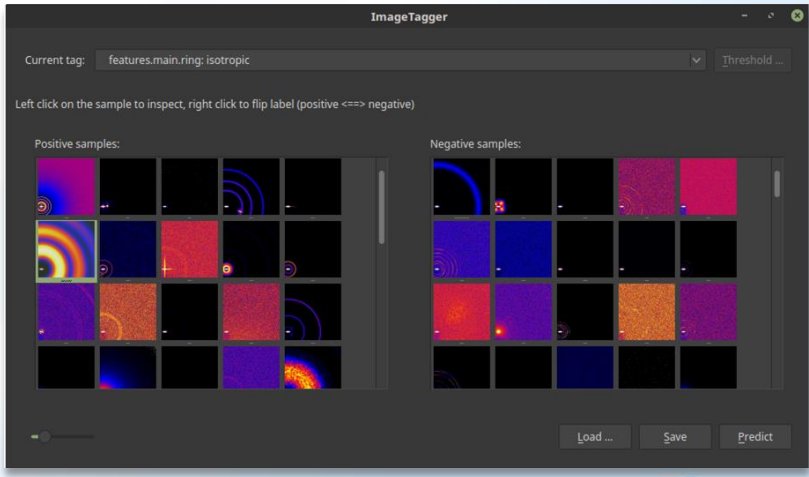


Ensemble

GUI aided image tagging

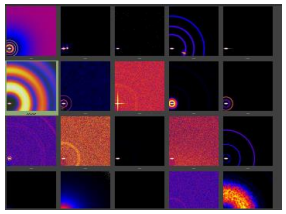
- Developed an intuitive GUI for scattering image dataset exploration, prediction and tagging
- Supports exploration and tag editing by tag names or individual images
- Interacts TensorFlow prediction module for fast batch tagging
- Highly configurable with tag, image manipulation and neural network model settings



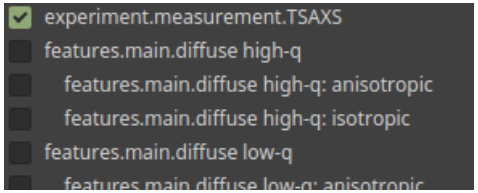


GUI aided image tagging

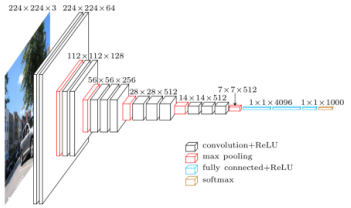
User upload



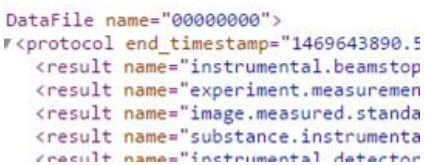
Select tag to inspect



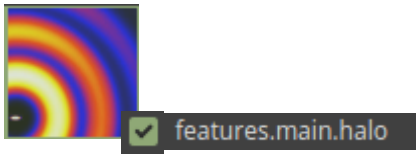
Deep learning prediction



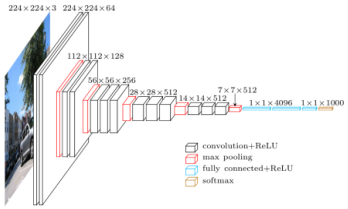
Generate predicted tag



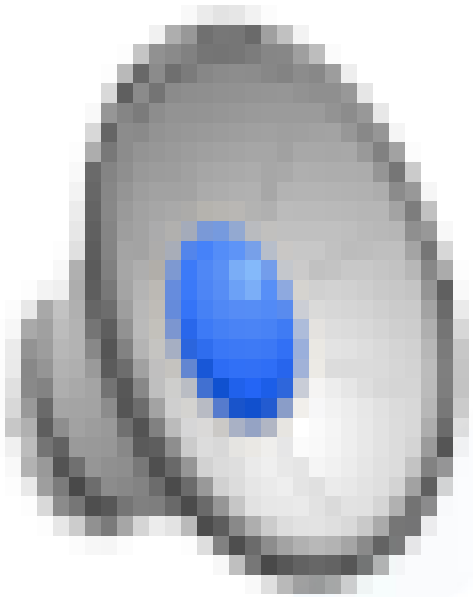
Threshold adjustment and manual correction



Incremental retrain

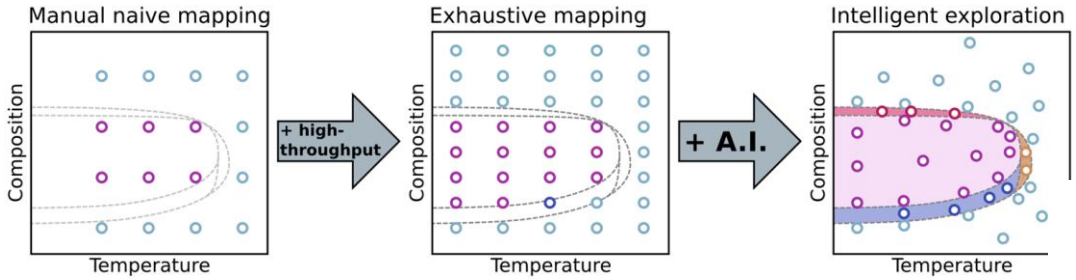


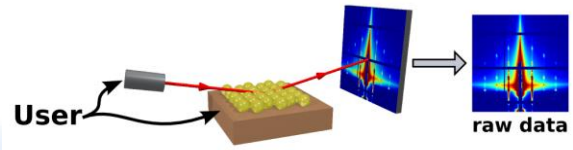
GUI aided image tagging

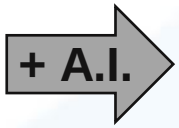


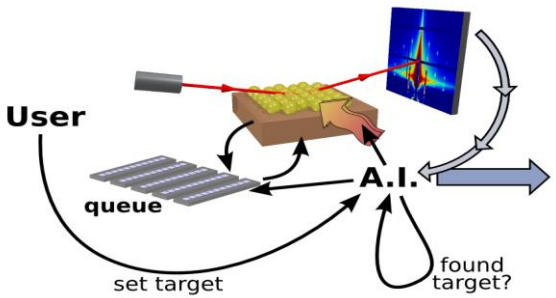
Conclusion

- Machine-learning is a critical component of **automated materials discovery**; a new experimental mode that:
  - Liberates scientists to work on science
  - Enables computer-controlled ‘intelligent’ exploration of materials questions
  - Accelerate scientific discoveries
  - Computer-directed beamline experiments would allow the instrument to explore physical parameter spaces, without human intervention





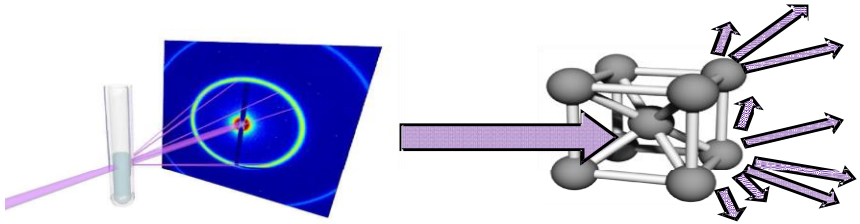




- Deep-learning is an effective tool, allowing the computer to extract physically-relevant meaning from abstract datasets

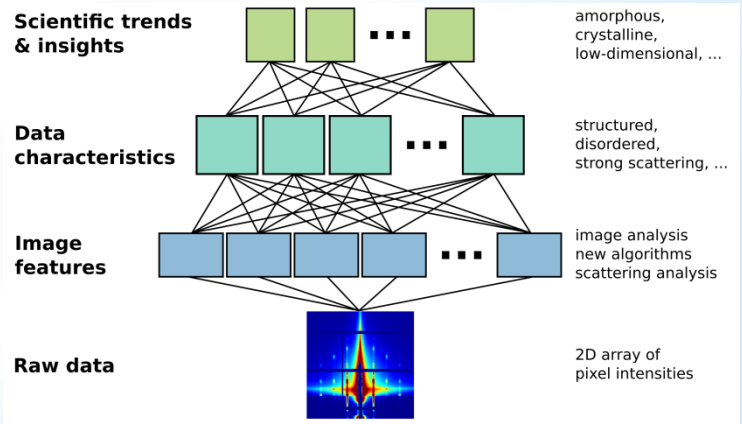
Future Works

- Complete data analysis pipeline and CNN for automated tagging of x-ray scattering data with rich labels:
  - Extensive set of domain-specific feature extractors.
  - Machine learning on combined inputs (domain-specific + deep-learning trained extractors).
  - Minimize computing time (each epoch 15 minutes).
  - Estimate Physics properties from (scattering) images



- Refine code:
  - Port code to Multiple GPU-accelerated context.
  - Augment pipeline to operate in streaming modes (for both recognition, and incremental training).

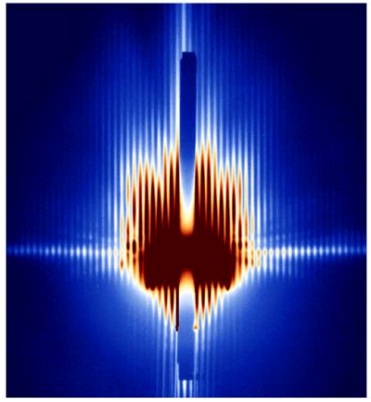
- Run code on CMS beamline at NSLS-II.



Future Machine Learning Aided Material Design

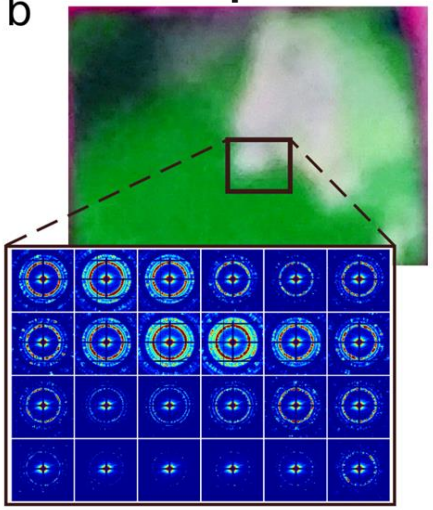
Processed area detector frame

a



Grid of data forms map of sample

b



Physical phase-diagram for experimental system

c