COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

ACCELERATING DEEP NEURAL NETWORKS FOR REAL-TIME DATA SELECTION FOR HIGH-RESOLUTION IMAGING PARTICLE DETECTORS

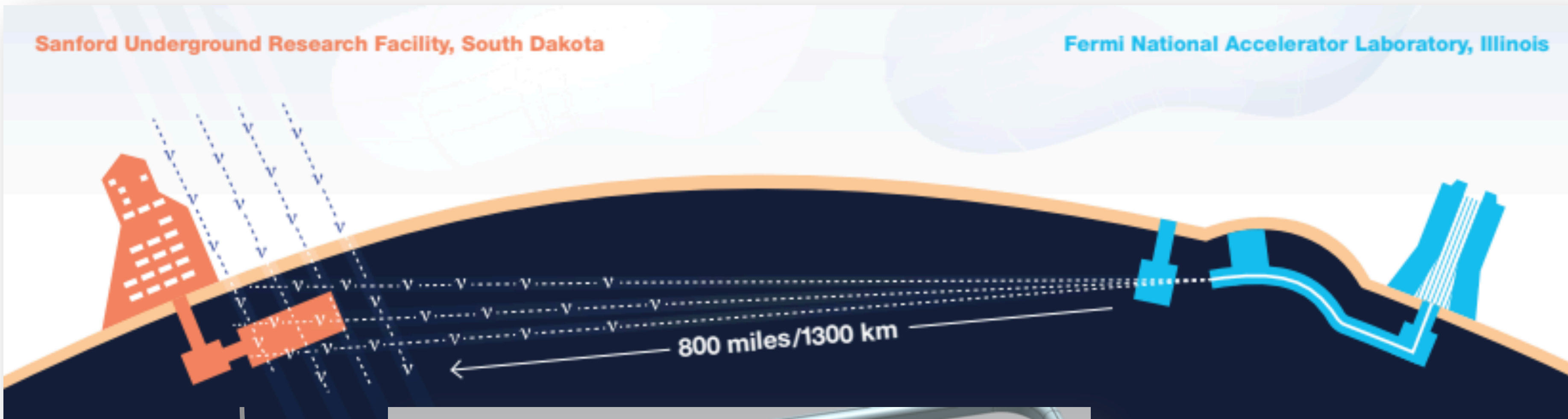
Georgia Karagiorgi[1], with Luca Carloni[2], Giuseppe Di Guglielmo[2], and Yeon-jae Jwa[1]

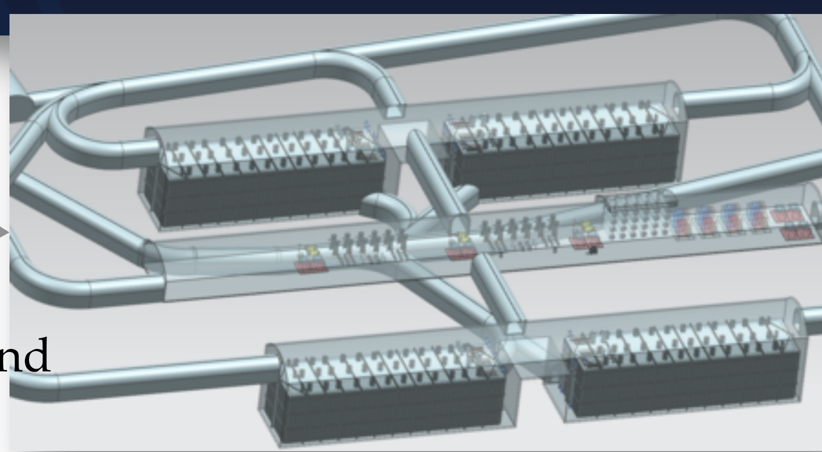
[1]Dept. of Physics, Columbia University

[2]Computer Science Department, Columbia University

High-resolution imaging particle physics detectors

- E.g. Deep Underground Neutrino Experiment (DUNE)





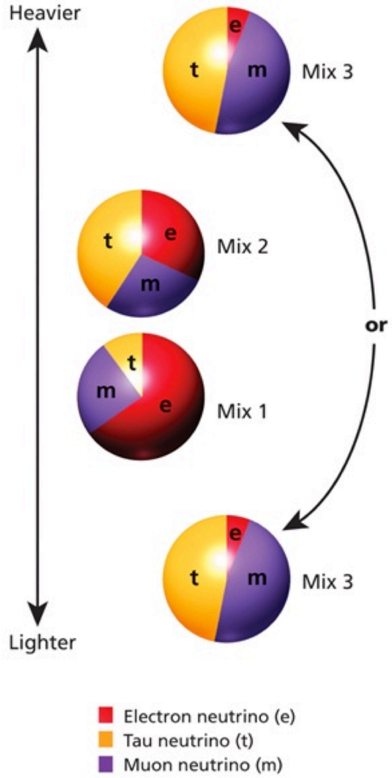
4 neutrino
detector modules
1 mile underground

DUNE
DEEP UNDERGROUND
NEUTRINO EXPERIMENT

[https://www.dunescience.org/]

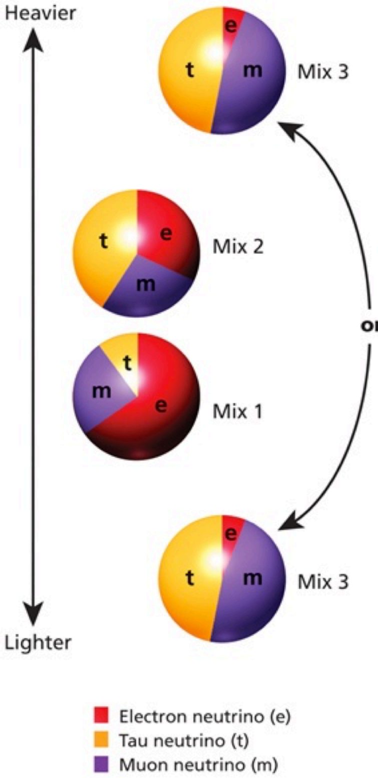
What is DUNE "looking for"?

- Rare interactions of (otherwise) invisible particles:
  - Neutrinos from a beam produced at the Fermi US National Lab (~few hundred per year)
  - Neutrinos produced in cosmic ray air showers in the atmosphere (~few thousand per year)



What is DUNE "looking for"?

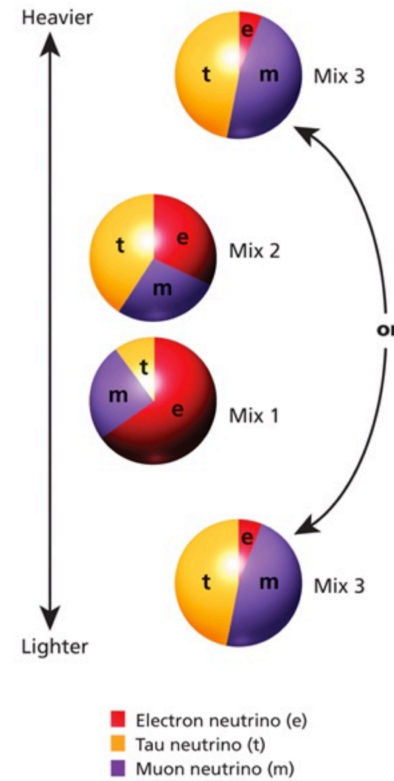
- Rare interactions of (otherwise) invisible particles:
  - Neutrinos from a beam produced at the Fermi US National Lab (~few hundred per year)
  - Neutrinos produced in cosmic ray air showers in the atmosphere (~few thousand per year)
  - Neutrinos produced in a (potential) nearby supernova burst (up to ~few thousand over 10 seconds, but ~once per century)

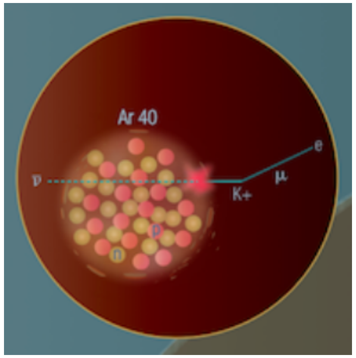


What is DUNE "looking for"?

- Rare interactions of (otherwise) invisible particles:
  - Neutrinos from a beam produced at the Fermi US National Lab (~few hundred per year)
  - Neutrinos produced in cosmic ray air showers in the atmosphere (~few thousand per year)
  - Neutrinos produced in a (potential) nearby supernova burst (up to ~few thousand over 10 seconds, but ~once per century)
  - Protons or neutrons inside the detector volume (liquid argon) spontaneously "decaying" in a way that violates fundamental symmetries of nature (~1 per year)

Rare processes, of fundamental importance in nature!

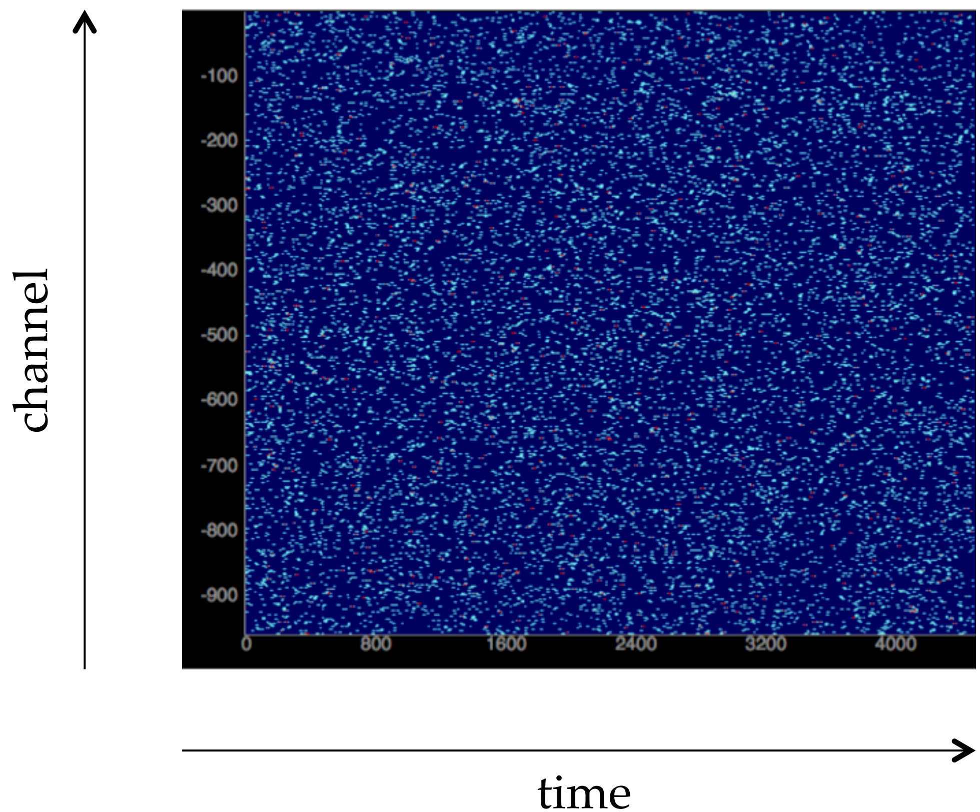




What would DUNE "see"?

What would DUNE "see"?

- For the most part:



Single frame from high-resolution video: One of three 2D views from one of hundreds of cells in the detector

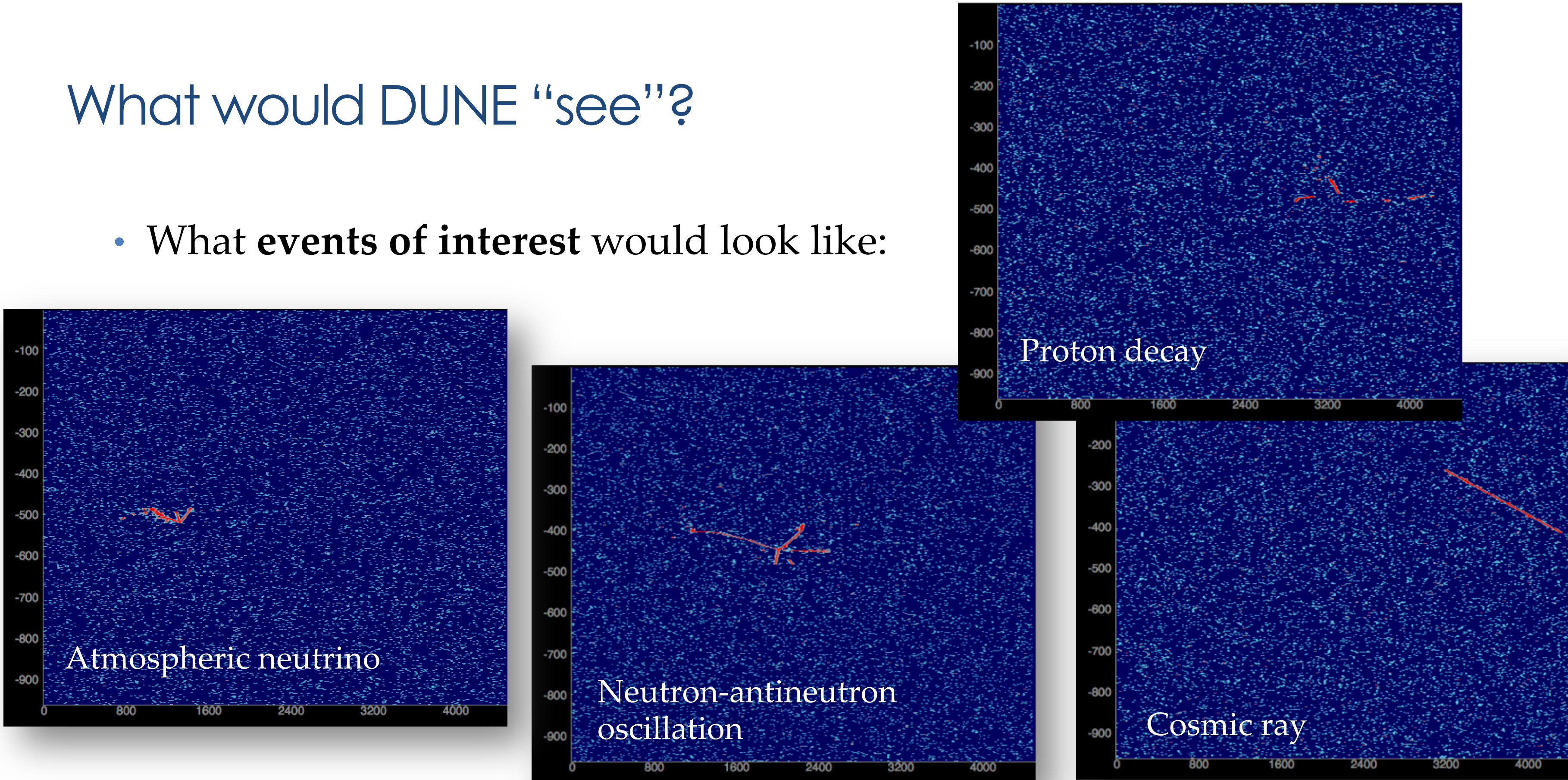
Color scale represents energy deposition (due to ionization) in the detector

"Static" is noise and small energy deposits from radiological impurities in the detector

[simulation]

What would DUNE “see”?

- What **events of interest** would look like:

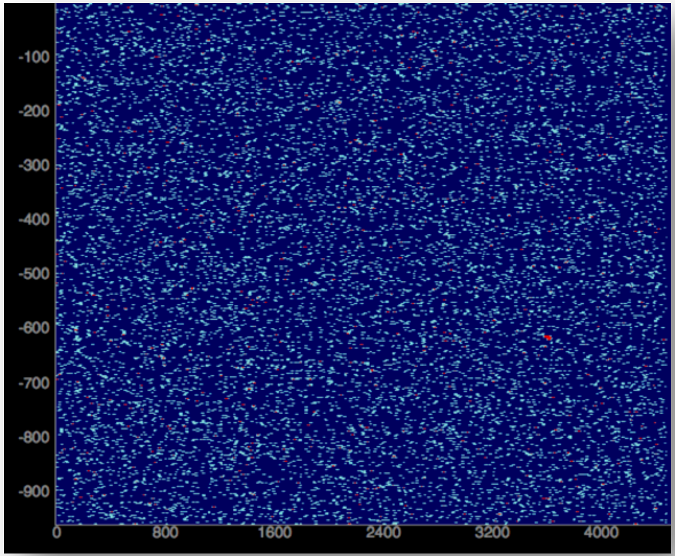


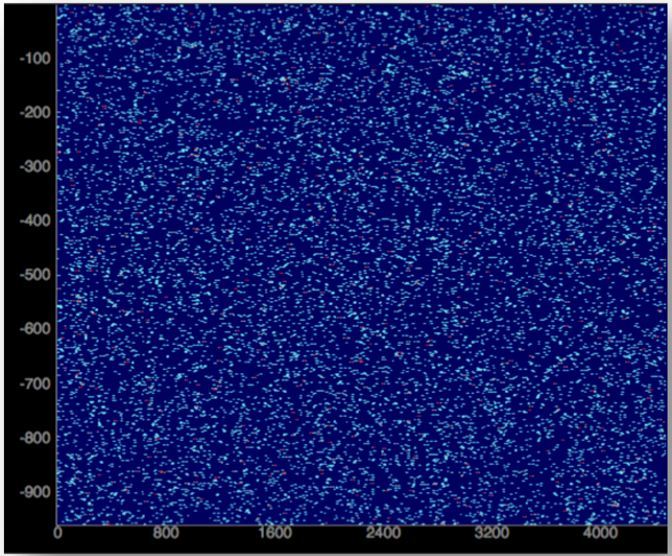
- Easy to pick out from background!
- On an event-by-event basis, difficult to differentiate between them
- On average, events can be differentiated based on their energy (pixel intensity) and topology characteristics (spatial extent, shape, e.g. tracks vs. showers and multiplicity, connected vs. detached, …)

Not all events of interest are as easy to pick out!

See: yesterday's talk by P. Nugent

- Special challenge: **neutrinos from supernova core collapse**
- Very low energy and small (in extent) topology, similar to radiological background activity in the detector



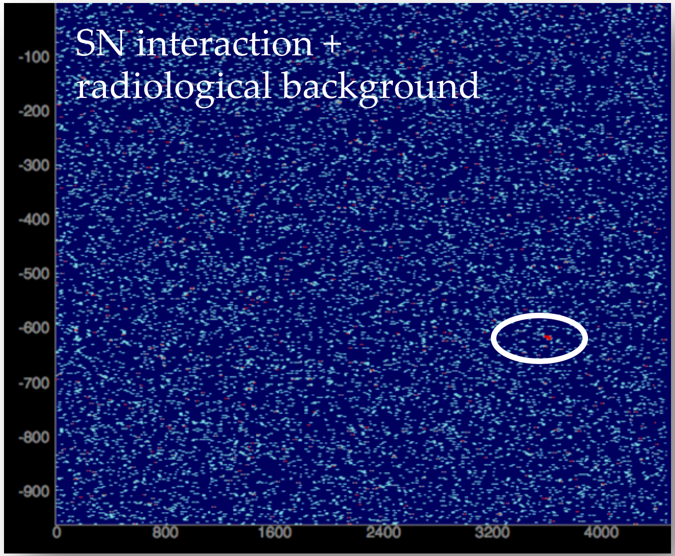


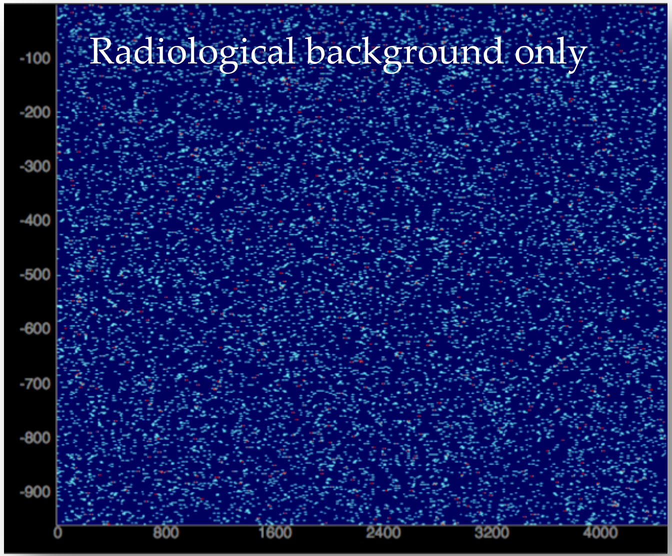
- Need $O(10^4)$ background suppression, while maintaining high efficiency to a frame containing a supernova neutrino interaction

Not all events of interest are as easy to pick out!

See: yesterday's talk by P. Nugent

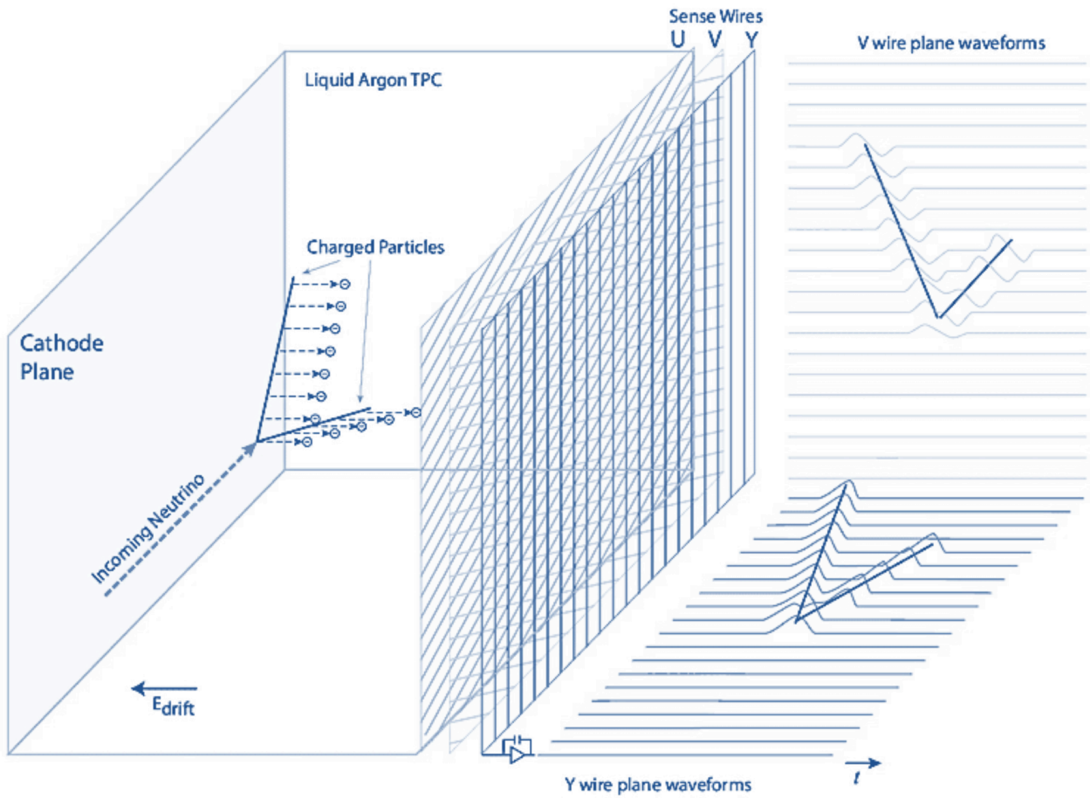
- Special challenge: **neutrinos from supernova core collapse**
- Very low energy and small (in extent) topology, similar to radiological background activity in the detector





- Need $O(10^4)$ background suppression, while maintaining high efficiency to a frame containing a supernova neutrino interaction

DUNE detector: working principle*

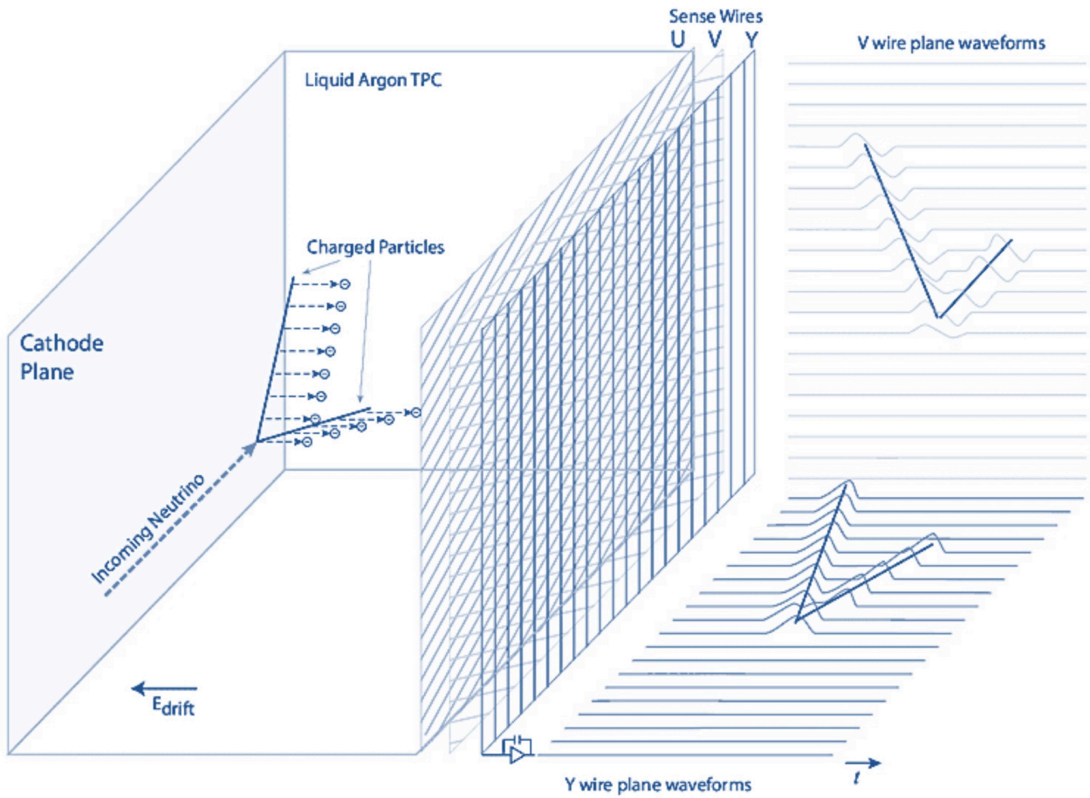


- particle-imaging detector
- stereoscopic "video capture" of activity within detector volume with sub-mm spatial resolution
- high-resolution "video" streams:
  - up to 4x150 cell volumes
  - 11.5 megapixel frames per 2.25ms
  - 12-bit resolution
  - a total of **~40 terabits/s**
- ***continuous operation for more than a decade***

See: Poster #8, Session #2, by J. I. Crespo-Anadon

*shown only for "single-phase" module technology; ~similar "dual-phase" module

DUNE detector: working principle*



- particle-imaging detector
- stereoscopic "video capture" of activity within detector volume with sub-mm spatial resolution
- high-resolution "video" streams:
  - up to 4x150 cell volumes
  - 11.5 megapixel frames per 2.25ms
  - 12-bit resolution
  - a total of **~40 terabits/s**
- **continuous operation for more than a decade**

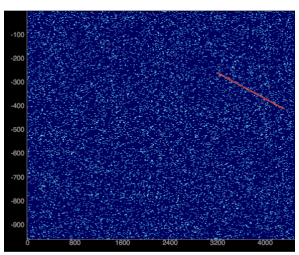
DATA PROCESSING CHALLENGE!

See: Poster #8, Session #2, by J. I. Crespo-Anadon

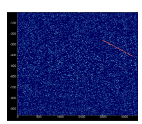
*shown only for "single-phase" module technology; ~similar "dual-phase" module

Promise of imaging techniques for DUNE

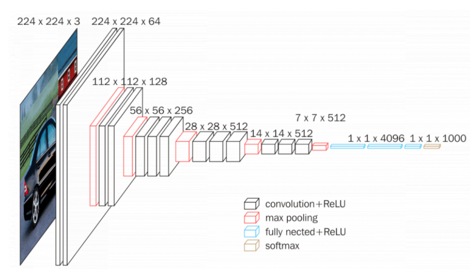
- Raw data format ideally suited for image analysis
- **Convolutional Neural Networks (CNNs)** could be applied for image classification "on the fly"
  - Work with only one projection (2D): 4.3 megapixel
  - Down-sample and resize image to 0.36 megapixel
  - Classify via CNN as one of three cases: **background**/**supernova-like low energy activity**/**high-energy activity**



raw image input (4450x960)



downsampling, resizing (600x600)



CNN classification

○ bkgd
○ SN LE
○ HE

selection (e.g., lowest background class score)

Promise of imaging techniques for DUNE

- Classification studies performed for DUNE simulated frames using CNN vgg16b:

		Background processes: consume most of the total data rate	Treasured physics processes: Possible ground breaking discovery			Physics processes for calibration and physics measurements	
Background CNN score cut	Background frame selection efficiency	Background data rate	Supernova frame sel. efficiency	n-nbar frame sel. efficiency	p-decay frame sel. efficiency	atmo. nu frame sel. efficiency	cosmic frame sel. efficiency
<0.05	0.56% (99.44% rejection)	6.4 GB/s (201 PB/year)	89%	100%	99%	92%	92%
<0.01	0.18% (99.82% rejection)	2.05 GB/s (65 PB/year)	86%	100%	99%	91%	92%
<0.001	0.031% (99.969% rejection)	350 MB/s (11 PB/year)	77%	100%	98%	89%	90%
<0.0002	0.011% (99.989% rejection)	125 MB/s (3.9 PB/year)	69%	100%	97%	87%	88%

- High selection efficiency across all topologies of interest

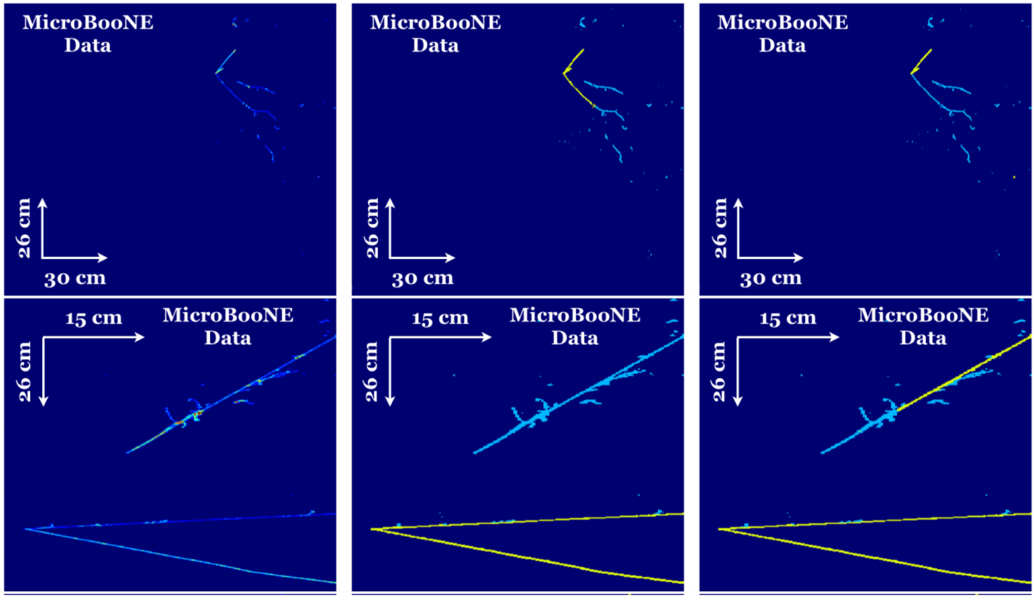
  ✔ **CNN-based selection on unprocessed, raw data**

- Further improvements possible by considering time-coincidence of activity over multiple (sequential) frames

Promise of imaging techniques for DUNE

- Deep Learning techniques already applied successfully in detectors sharing the same technology as DUNE
- E.g. MicroBooNE experiment (1/500th size of DUNE) is pioneering such applications

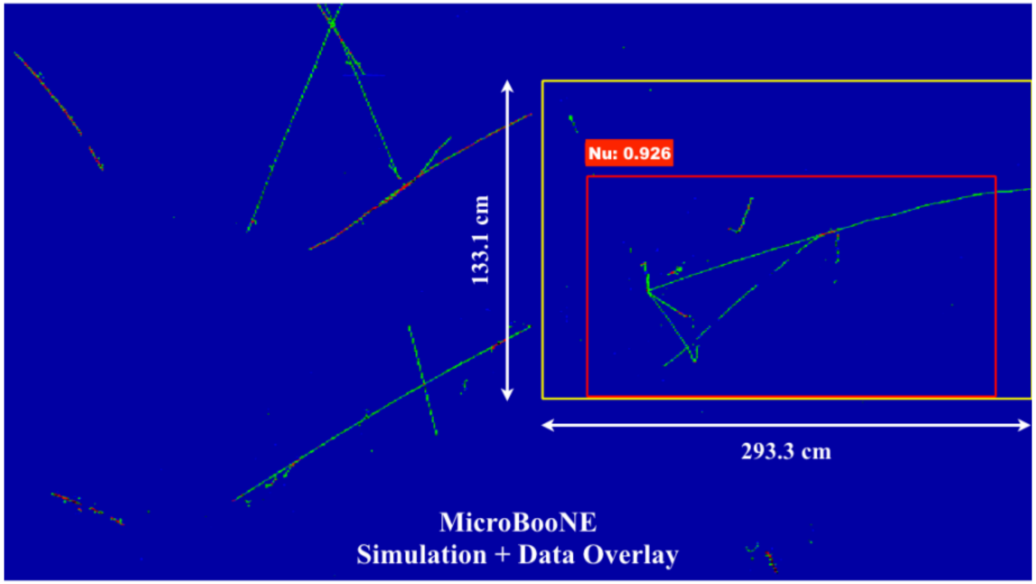
[1]



See, e.g.:
[1] "Deep neural network for pixel-level electromagnetic particle identification in the MicroBooNE liquid argon time projection chamber," Phys. Rev. D99 (2019) No. 9, 092001.
[2] "Convolutional Neural Networks Applied to Neutrino Events in a Liquid Argon Time Projection Chamber," JINST 12 (2017) No. 03, P03011.

Promise of imaging techniques for DUNE

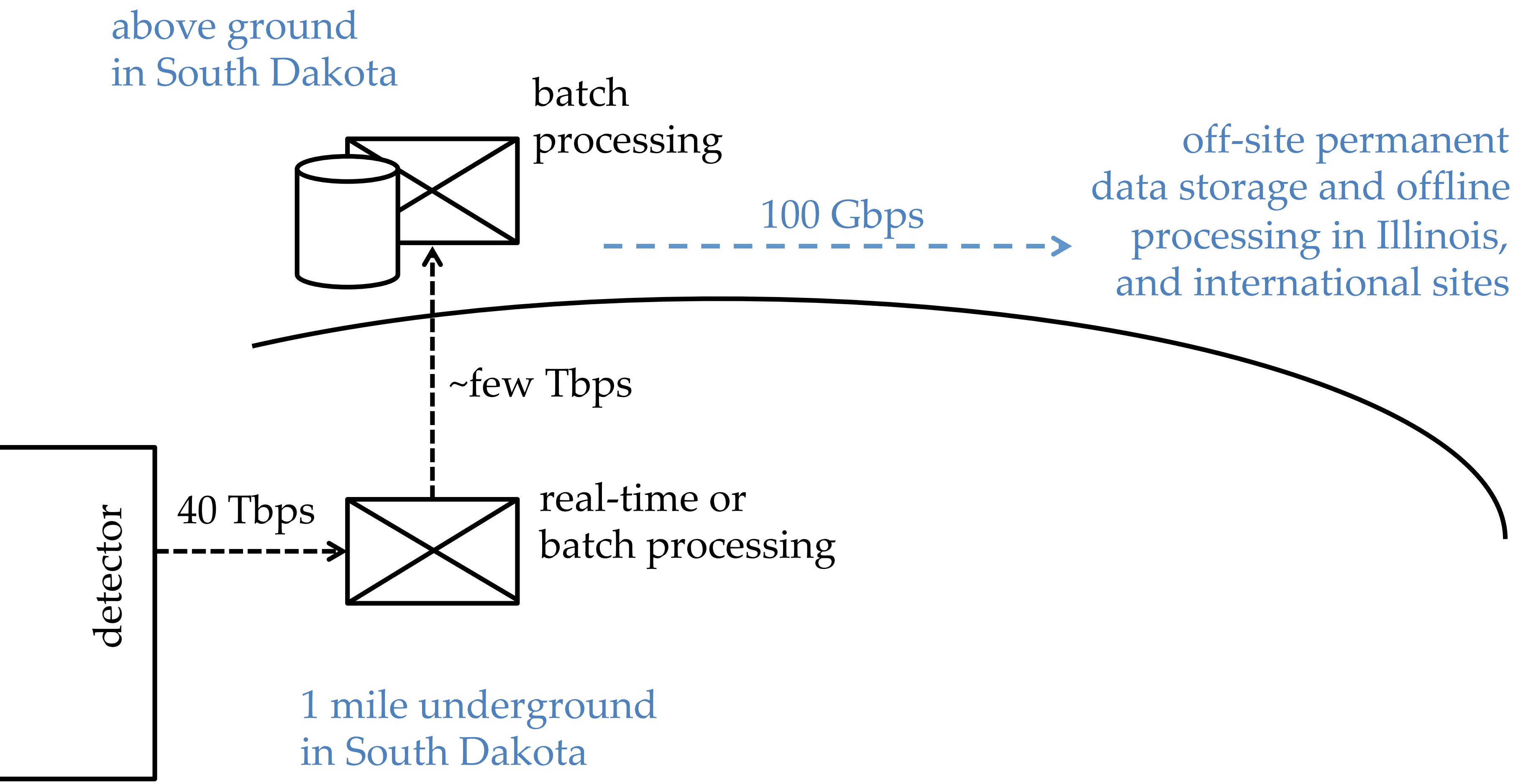
- Deep Learning techniques already applied successfully in detectors sharing the same technology as DUNE
- E.g. MicroBooNE experiment (1/500th size of DUNE) is pioneering such applications



CNNs can be trained to do particle classification, particle and neutrino detection, and neutrino event identification [2].

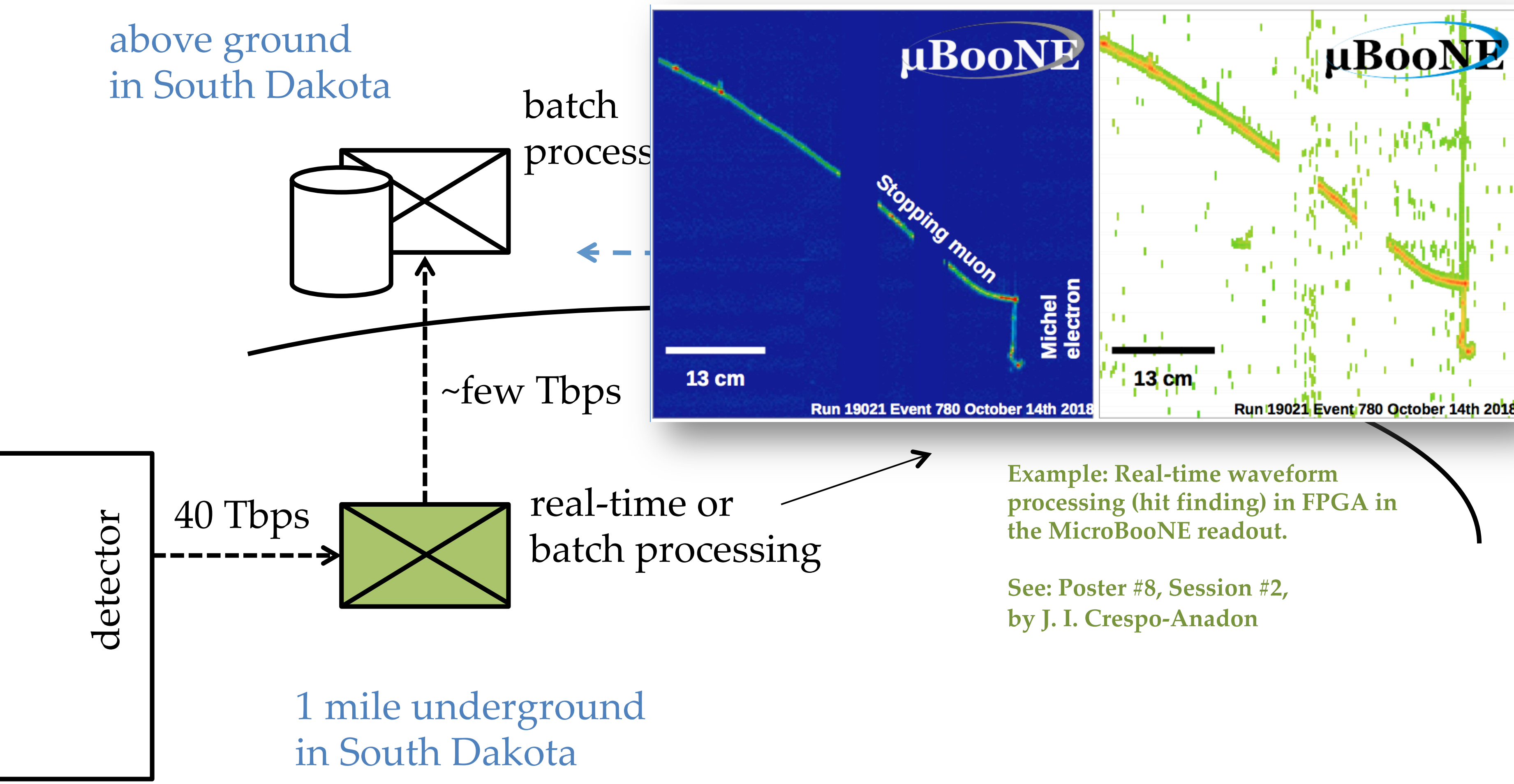
See, e.g.:
"Deep neural network for pixel-level electromagnetic particle identification in the MicroBooNE liquid argon time projection chamber," Phys. Rev. D99 (2019) No. 9, 092001.
"Convolutional Neural Networks Applied to Neutrino Events in a Liquid Argon Time Projection Chamber," JINST 12 (2017) No. 03, P03011.

DUNE readout and data acquisition system design



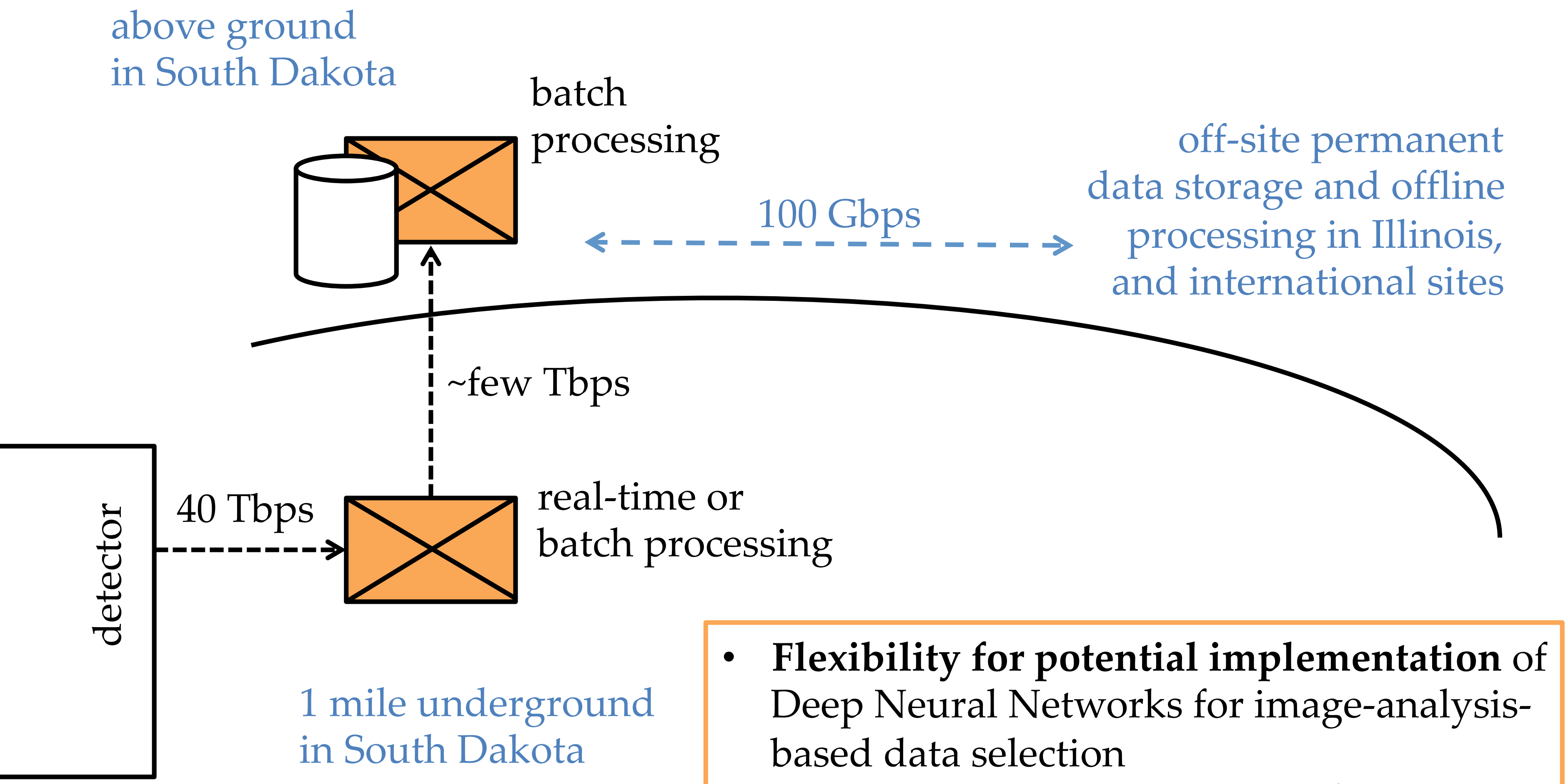
[DUNE Technical Design Report, in preparation.]

DUNE readout and data acquisition system design



[DUNE Technical Design Report, in preparation.]

DUNE readout and data acquisition system design



- **Flexibility for potential implementation** of Deep Neural Networks for image-analysis-based data selection
- Must **keep within cost and performance constraints:** latency, power, cost envelope

Performance for batch processing for data selection with GPU implementation

- **GPU advantages**: High computational density, level of programmability, data-parallelism, flexibility
- **Investigated CNN-based selection performance** (latency) for DUNE simulated frames:
  On single GPU (NVIDIA GeForce GXT 1080 Ti)
  - vgg16b 26 ms/frame (compare to 2.25ms frame)
  - resnet14b 24 ms/frame

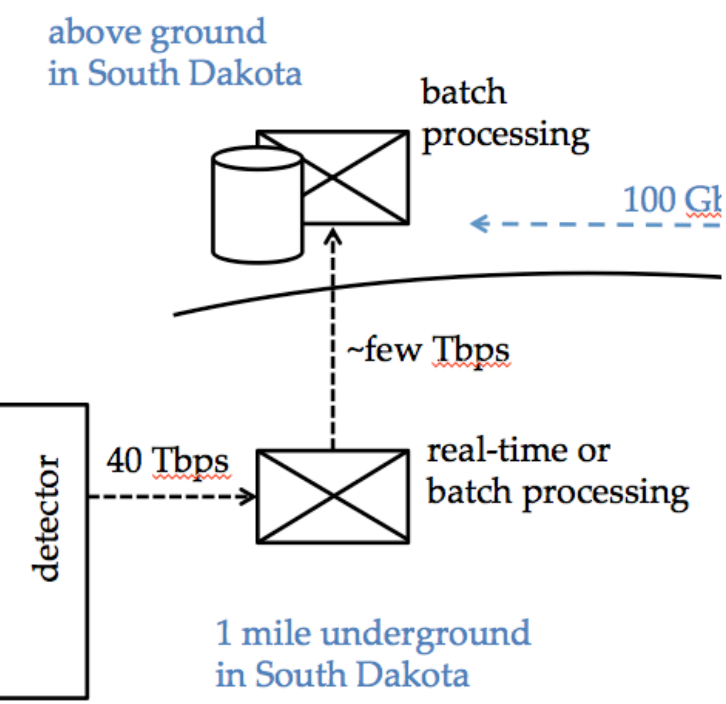
  Includes data i/o and network inference time

- Speed sufficient for downstream implementation; but a factor of 10 speedup needed for upstream implementation (power constraints aside…)
  - Further optimization may be possible: e.g. image size: further down-sampling vs. region-of-interest

Frame size	140×140	280×280	600×600
Measured time (ms)	18.92	22.10	27.58

R&D for real-time processing for data selection with FPGA implementation

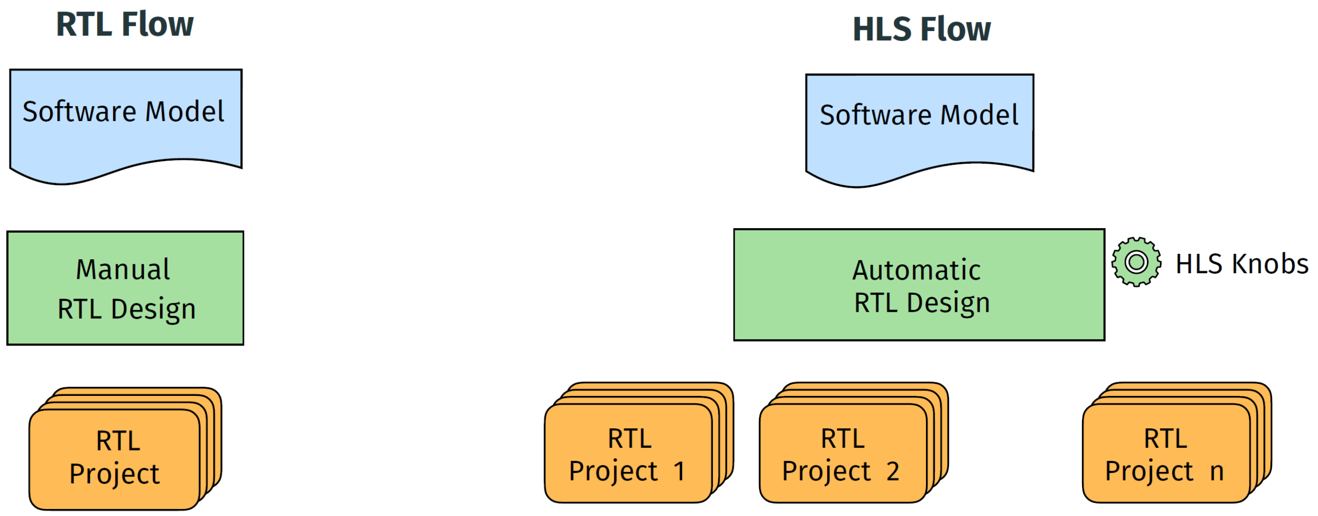
- **Advantages for upstream (FPGA) implementation**: reduction in overall data transfer to above ground, buffering needs, power dissipation
  - FPGA: power-aware platform for CNN acceleration, but resource-constrained
  - Concern: network size (resnet14b, vgg16b) and input image size are large



- **Exploring CNN acceleration** using a customizable and efficient hardware accelerator design for the various layers of CNN, utilizing High Level Synthesis-based design flow

- Flexibility for optimization (processing time, efficiency, resource utilization)

Design Flow for FPGA Accelerators

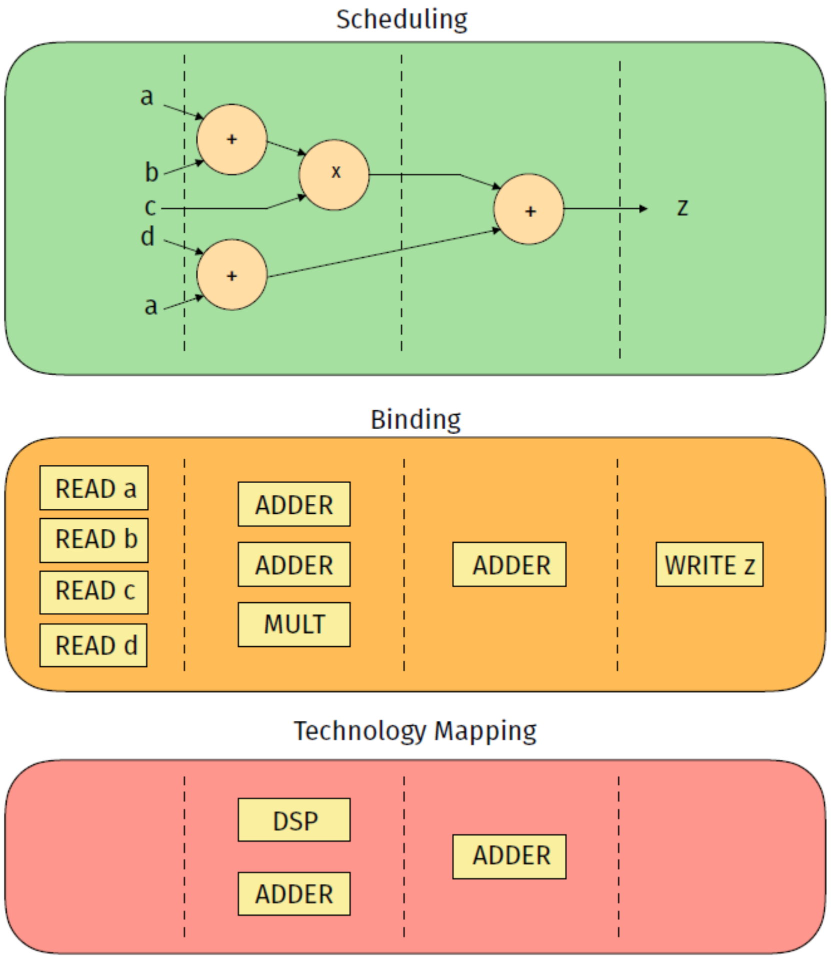


- Register Transfer Level (RTL) is a low level representation of digital circuits and is a *de facto* standard for designing hardware
- High Level Synthesis (HLS) allows hardware designers to take advantage of benefits of working at a **higher level of abstraction**, while creating high-performance hardware
  - HLS allows to efficiently and rapidly perform **Design Space Exploration (DSE)**

High Level Synthesis

- HLS transforms a **behavioral description** into **timed design**
- This is done in three steps: **scheduling, binding** and **technology mapping**

```
int func(int a, int b, int c, int d) {
  int z;
  z = (a + b) * c + d + a;
  return z;
}
```

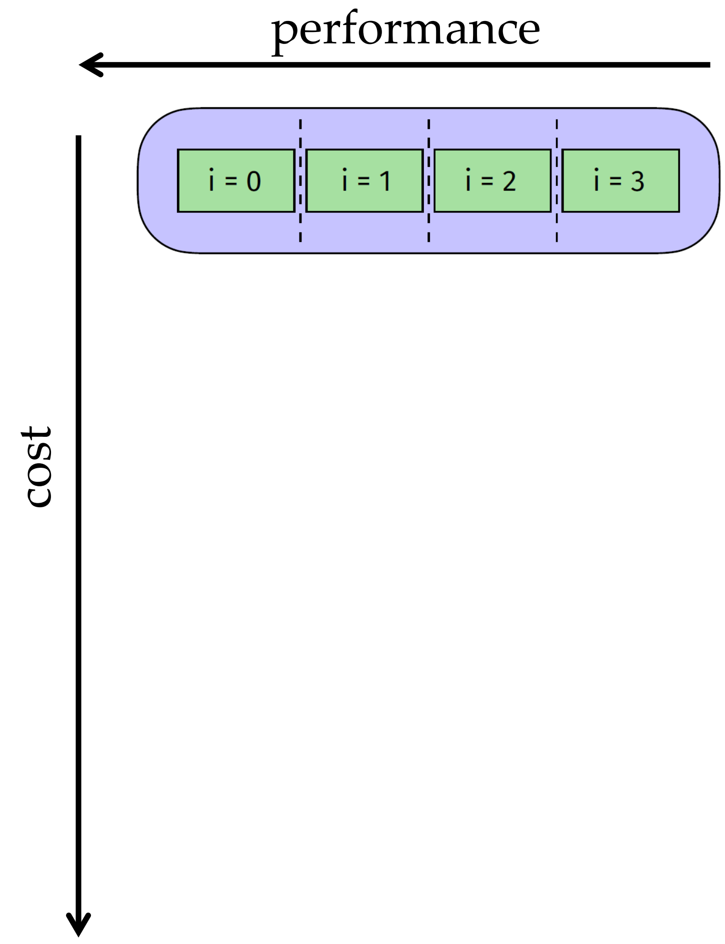


Knobs of High Level Synthesis

- HLS allows to control fine-grain architectural implementation using pre-defined **knobs**
- Allow exploring concurrency in design, e.g.

```
void sum(int a[4], int b[4], int c[4]) {
  for (int i = 0; i < 4; i++) {
#pragma HLS UNROLL factor=1
    c[i] = a[i] + b[i];
}
```

- **Can explore implementations based on desired performance (latency) and cost (area, power)**

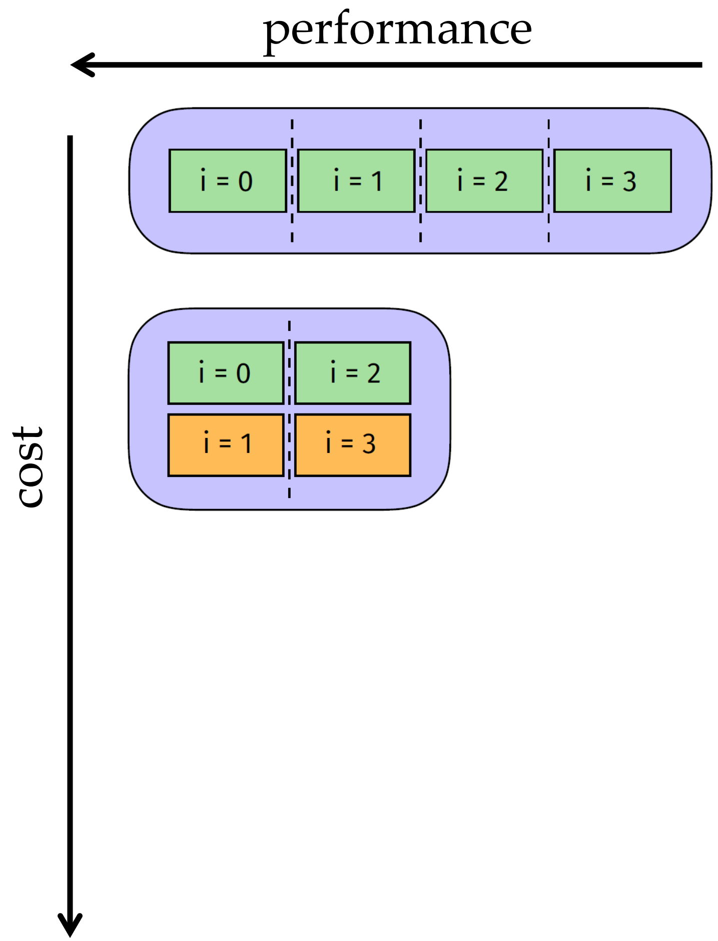


Knobs of High Level Synthesis

- HLS allows to control fine-grain architectural implementation using pre-defined **knobs**
- Allow exploring concurrency in design, e.g.

```
void sum(int a[4], int b[4], int c[4]) {
  for (int i = 0; i < 4; i++) {
#pragma HLS UNROLL factor=2
    c[i] = a[i] + b[i];
}
```

- **Can explore implementations based on desired performance (latency) and cost (area, power)**

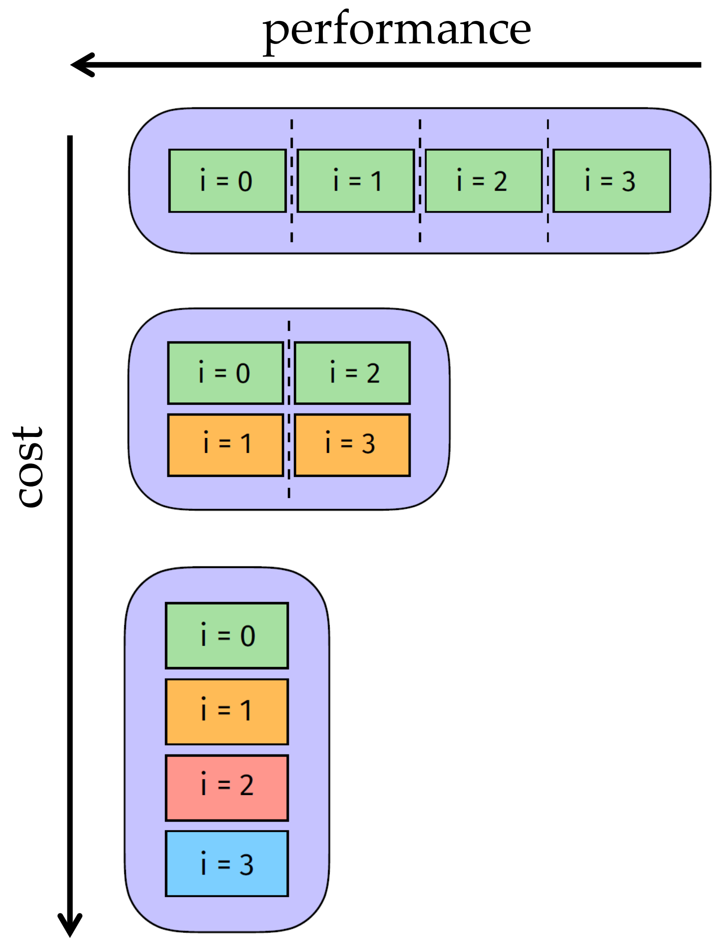


Knobs of High Level Synthesis

- HLS allows to control fine-grain architectural implementation using pre-defined **knobs**
- Allow exploring concurrency in design, e.g.

```
void sum(int a[4], int b[4], int c[4]) {
  for (int i = 0; i < 4; i++) {
#pragma HLS UNROLL factor=4
    c[i] = a[i] + b[i];
}
```

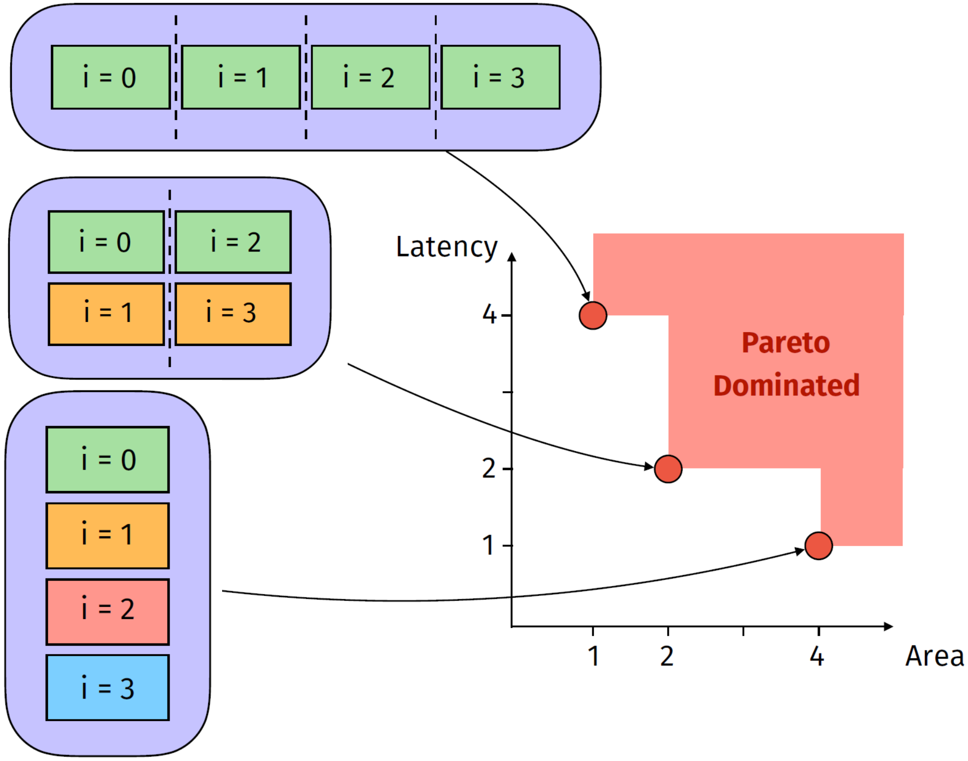
- **Can explore implementations based on desired performance (latency) and cost (area, power)**



Knobs of High Level Synthesis

- All of these implementations are optimal in terms of cost (area) and performance (latency)

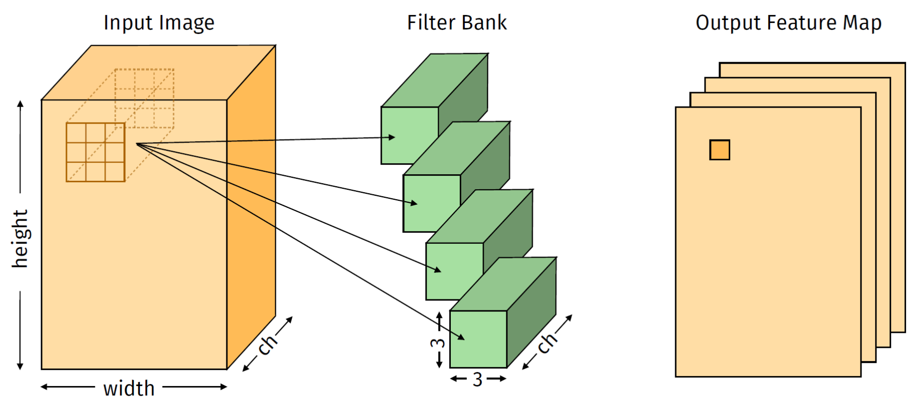
```
void sum(int a[4], int b[4], int c[4]) {
  for (int i = 0; i < 4; i++) {

    c[i] = a[i] + b[i];
}
```

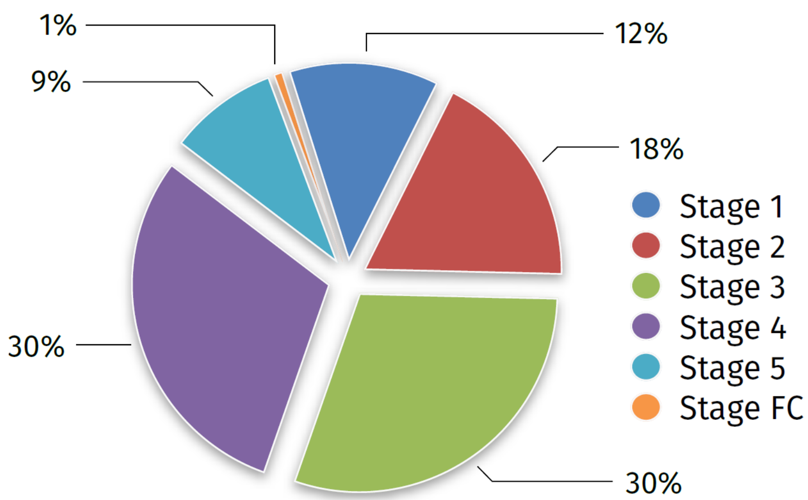


Convolutional Layers

- Convolutional layers are the most computational intensive part in CNNs

$$Y_{k,i,j} = \sum_{c=0}^{C-1} X_c * W_{k,c} + B_k = \left[\sum_{c=0}^{C-1} \sum_{x=0}^{F-1} \sum_{y=0}^{F-1} X_{c,i+x-\frac{F}{2},j+y-\frac{F}{2}} \cdot W_{k,c,x,y} \right] + B_k$$

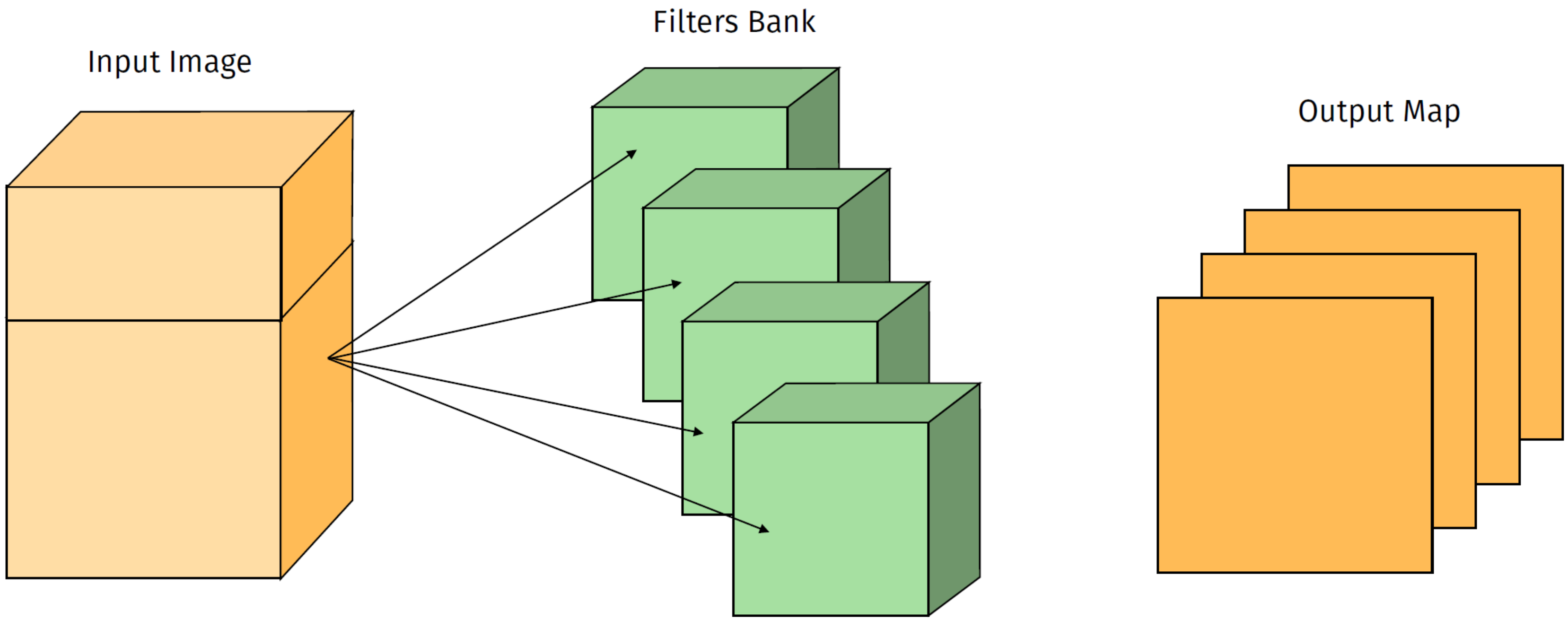




Distribution of floating-point operations per stages in vgg16b

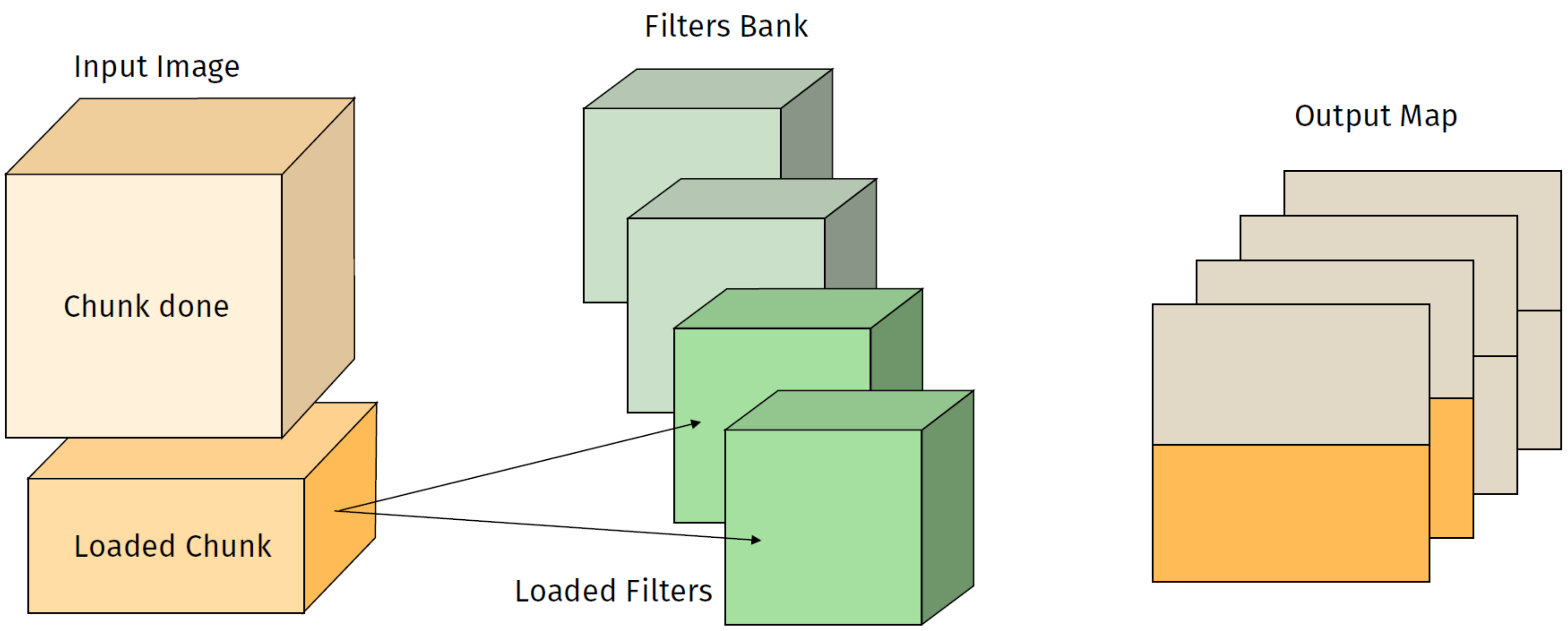
Balance of Computation and Communication

- For hardware accelerator, one should carefully design the algorithm to reuse data as much as possible, thus reducing expensive memory transfers from and to off-chip DRAM



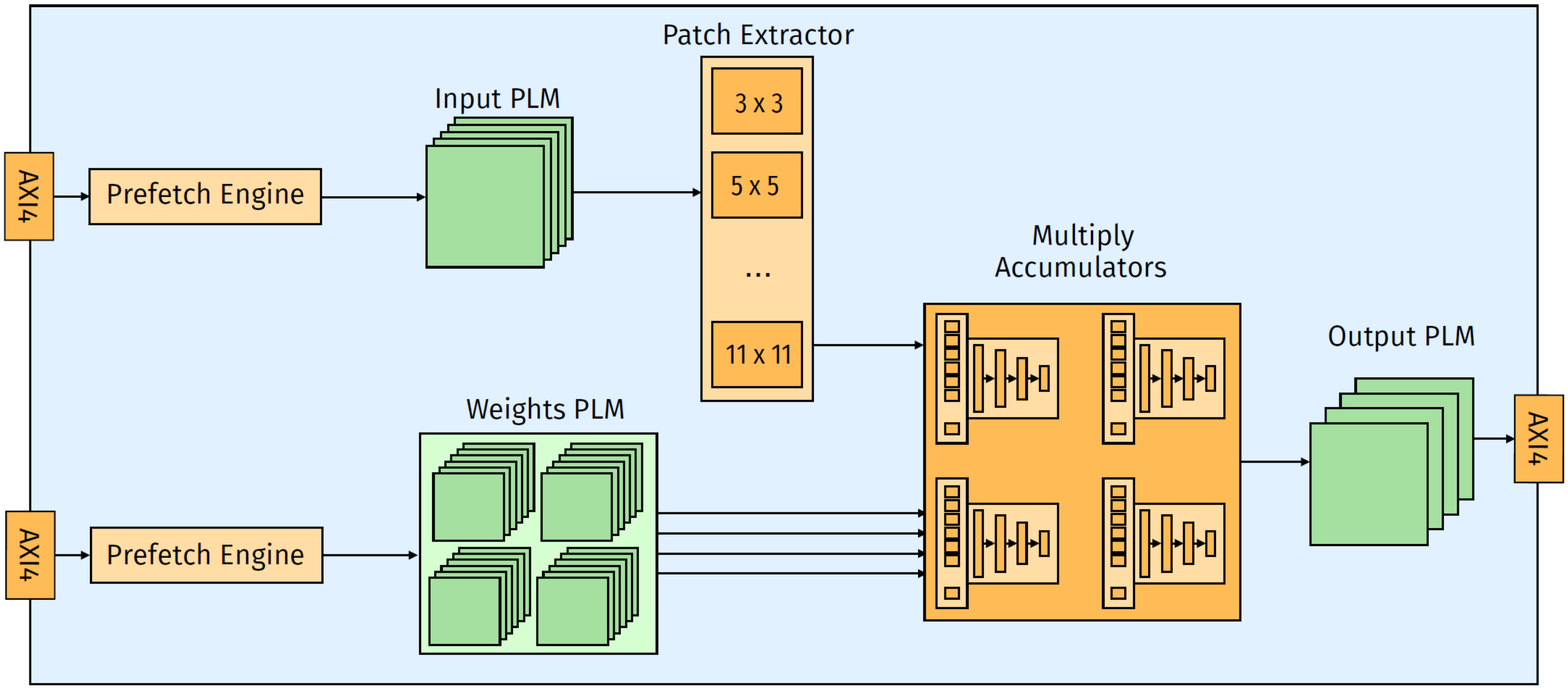
Tailoring Private Local Memory

- Both inputs and weights are divided in chucks and the computation is done only with the on-chip copy of the data



Accelerator Structure Overview

- Highly configurable accelerator



Preliminary Results

	CPU			Accelerator		
	MFLOP	Time	GFLOPS	Time	GFLOPS	Speedup
conv1_1	86.7	2.17	0.04	0.21	0.41	10.31
conv1_2	3699.4	51.05	0.07	3.66	1.01	13.95
conv2_1	1849.7	25.24	0.07	1.82	1.02	13.87
conv2_2	3699.4	51.27	0.07	3.46	1.07	14.82
conv3_1	1849.7	24.84	0.07	1.72	1.08	14.44
conv3_2	3699.4	50.85	0.07	3.37	1.10	15.09
conv3_3	3699.4	51.24	0.07	3.37	1.10	15.20
conv4_1	1849.7	25.23	0.07	1.68	1.10	15.02
conv4_2	3699.4	50.68	0.07	3.34	1.11	15.17
conv4_3	3699.4	50.68	0.07	3.34	1.11	15.17
conv5_1	924.8	12.46	0.07	0.84	1.10	14.83
conv5_2	924.8	12.46	0.07	0.84	1.10	14.83
conv5_3	924.8	12.46	0.07	0.84	1.10	14.83

	Time (s)	**Power** (W)	**PET** (Img/s/W)
ARM A53	420	3.2	0.001
Xilinx XCZU9EG FPGA	**28**	**0.8**	**0.045**

Xilinx ZynqMP UltraScale+ XCZU9EG

15x average speedup
45x more power efficient
w.r.t software implementation on ARM Cortex A53

Summary

- There is an increasing need for real-time processing of high-resolution images from particle detectors

- DUNE is a prime application for image processing using DNNs, and calls for optimizing DNN implementation on power-efficient platforms

- Serves as an ideal case for collaboration between physics and computer science

  - Demonstrated applicability of DNN-based selection
  - In the process of optimizing implementation on power-efficient platform
  - Future plans: demonstration of real-time processing meeting performance and cost requirements

Acknowledgements

- Simone Rossi for initial accelerator studies
- Ashley Koo for ResNet studies
- Yuyang Zhou for initial CNN data selection studies
- Jeremy Hewes for initial CNN physics studies
- Kazu Terao for valuable feedback

Funding support though:

- National Science Foundation
- Columbia Research Initiatives in Science and Engineering (RISE)
- Columbia University Provost's Office

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK