An Overview of Machine Learning at Jefferson Lab

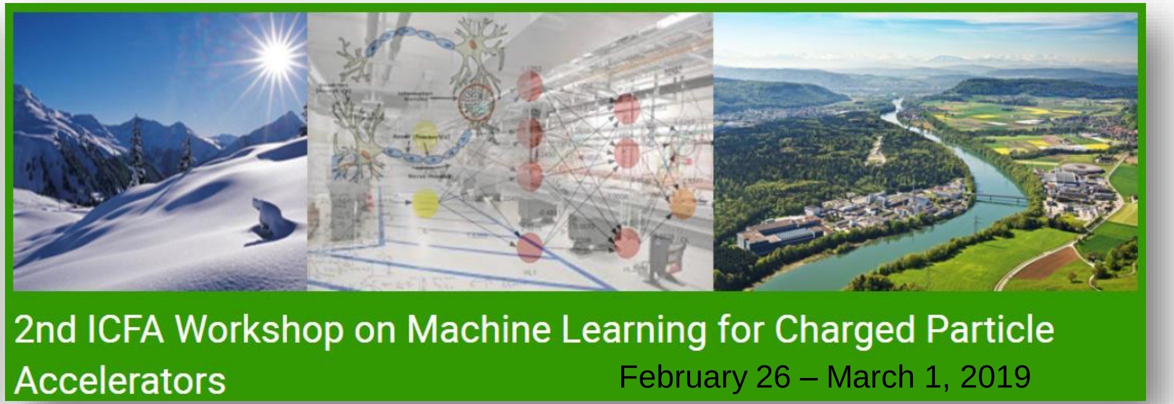
Chris Tennant

October 6, 2019

ICALEPCS 2019

Data Science and Machine Learning Workshop

Motivation



~60 participants

DOE "AI for Science" Town Hall Meetings

350

Chicago AI for Science Town Hall
Argonne National Laboratory
July 22-23, 2019

400+

Berkeley AI for Science Town Hall
Lawrence Berkeley National Laboratory
September 11-12, 2019

Oak Ridge AI for Science Town Hall
Oak Ridge National Laboratory
August 20-21, 2019

350

Washington DC AI for Science Town Hall
October 22-23, 2019

400

Final report will include contributions from 1,000+ participants

One of the Takeaways

- lots of open questions about data
  - ✓ what do you collect?
  - ✓ when do you collect it?
  - ✓ how to handle sparse data sets?
  - ✓ how to handle enormous data sets?
  - ✓ how to deal with disparate/diverse data?

One of the Takeaways

- lots of open questions about data
  - ✓ what do you collect?
  - ✓ when do you collect it?
  - ✓ how to handle sparse data sets?
  - ✓ how to handle enormous data sets?
  - ✓ how to deal with disparate/diverse data?

"we are data rich, but information poor"

Why Is Accelerator Physics Lagging?

- we have lots of data, lots of inputs and outputs… what’s the problem?

Why Is Accelerator Physics Lagging?

- we have lots of data, lots of inputs and outputs… what's the problem?
- we often do <u>not</u> have *the right kind of data, recorded at the right times*

Why Is Accelerator Physics Lagging?

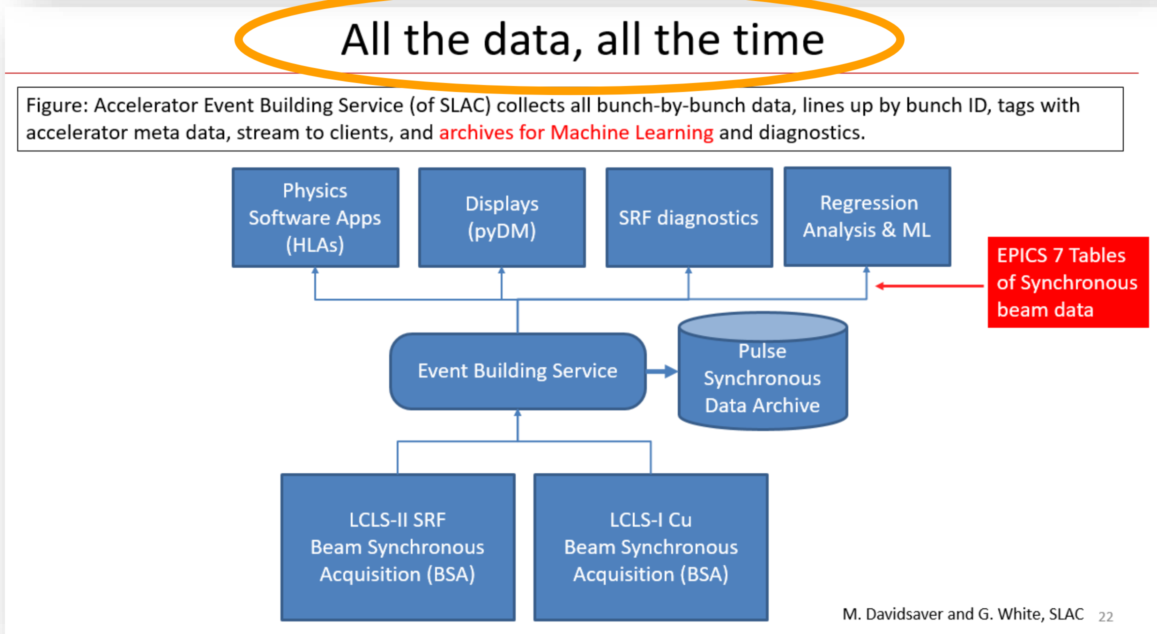
- we have lots of data, lots of inputs and outputs… what's the problem?
- we often do <u>not</u> have *the right kind of data, recorded at the right times*
- need a fundamental shift in the way accelerator side deals with data
- overhauling EPICS

"Recently, EPICS has undergone a major revision, with the aim of better computing supporting for the next generation of machines and analytical tools…The result has been that controls are now being integrated with modelling and simulation, machine learning, enterprise databases, and experiment DAQs."

THE EPICS SOFTWARE FRAMEWORK MOVES FROM CONTROLS TO PHYSICS

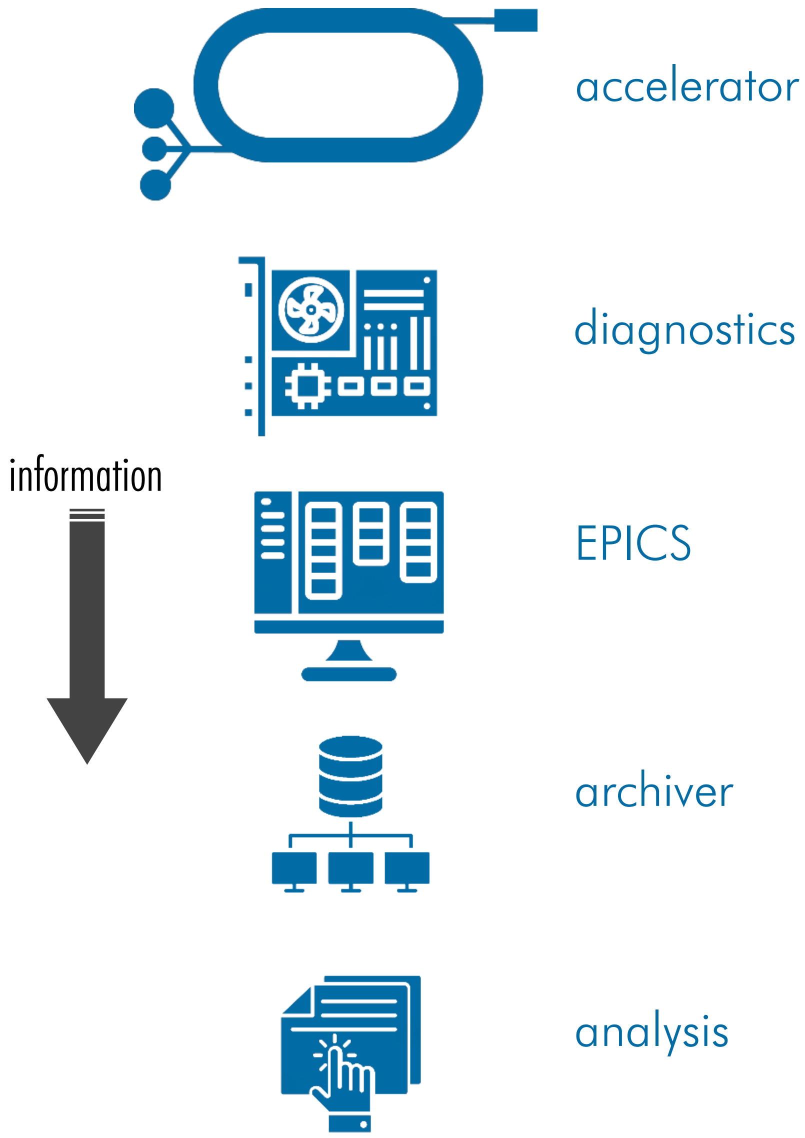
Greg White, for the EPICS Core Working Group

21st May 2018, IPAC 19



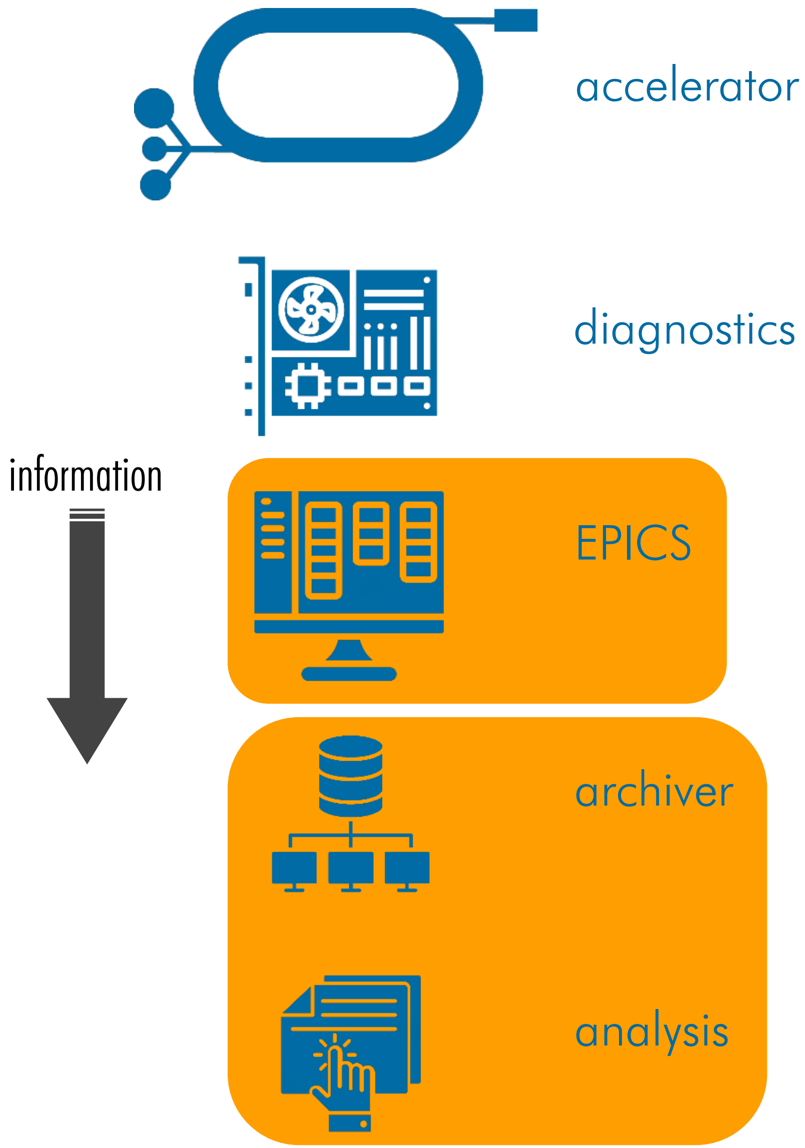
Data's Explosive Growth

- CEBAF archiver represents a potentially data rich resource
  - ✓ 2016: 236K channels
  - ✓ 2019: 354K channels



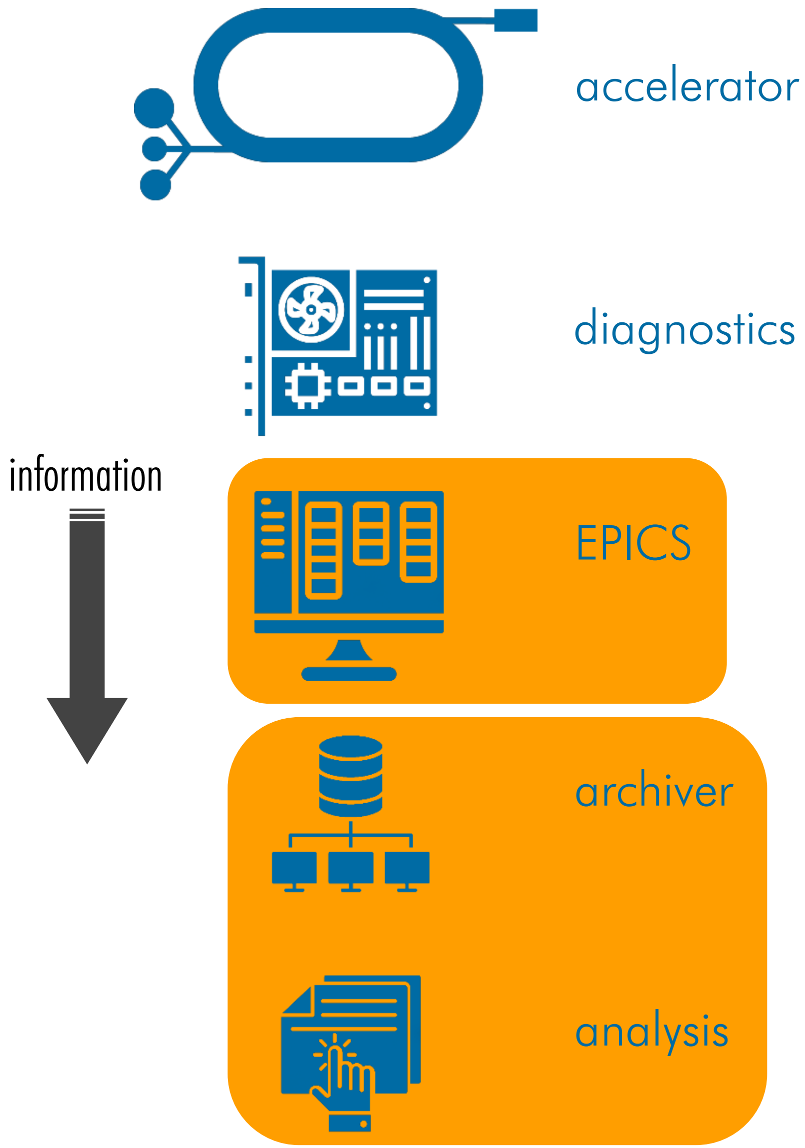
Data's Explosive Growth

- CEBAF archiver represents a potentially data rich resource
    - ✓ 2016: 236K channels
    - ✓ 2019: 354K channels



Data's Explosive Growth

- CEBAF archiver represents a potentially data rich resource
  - ✓ 2016: 236K channels
  - ✓ 2019: 354K channels
- it is possible to record enormous amounts of data, but unless it is the *right kind of data, recorded at the right times*, it will never lead to useful information
- how do we know?



Knowledge Discovery in Databases (KDD)

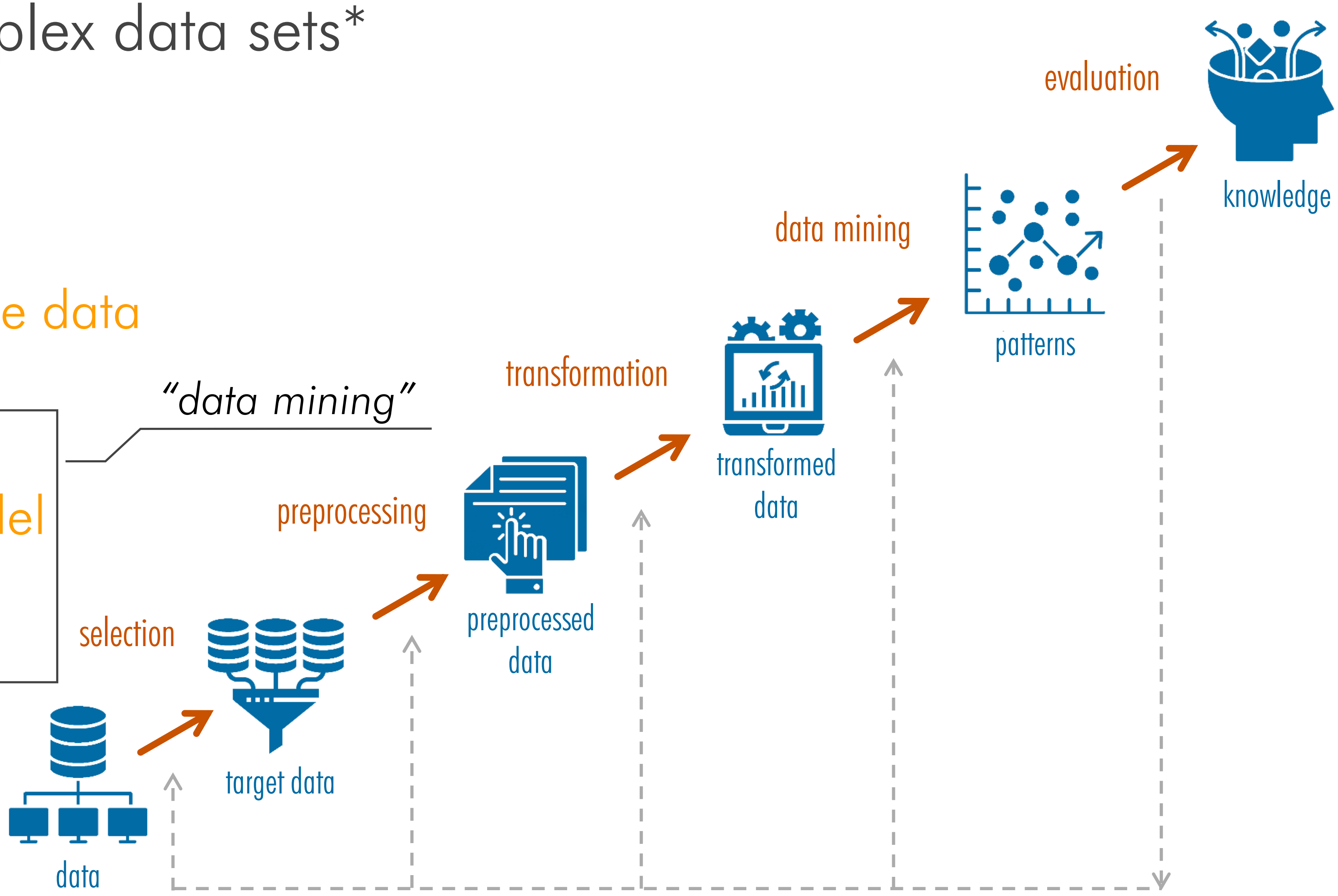
- definition: the process of discovering useful information (knowledge) from large and complex data sets*

*U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Volume 17, Number 3 (1996).

Knowledge Discovery in Databases (KDD)

- definition: the process of discovering useful information (knowledge) from large and complex data sets*
- procedure:
  1. identify the goal
  2. select the data
  3. clean and pre-process the data
  4. data transformation
  5. choose data mining task
  6. choose data mining model
  7. implement model
  8. evaluate model
  9. apply knowledge

"data mining"

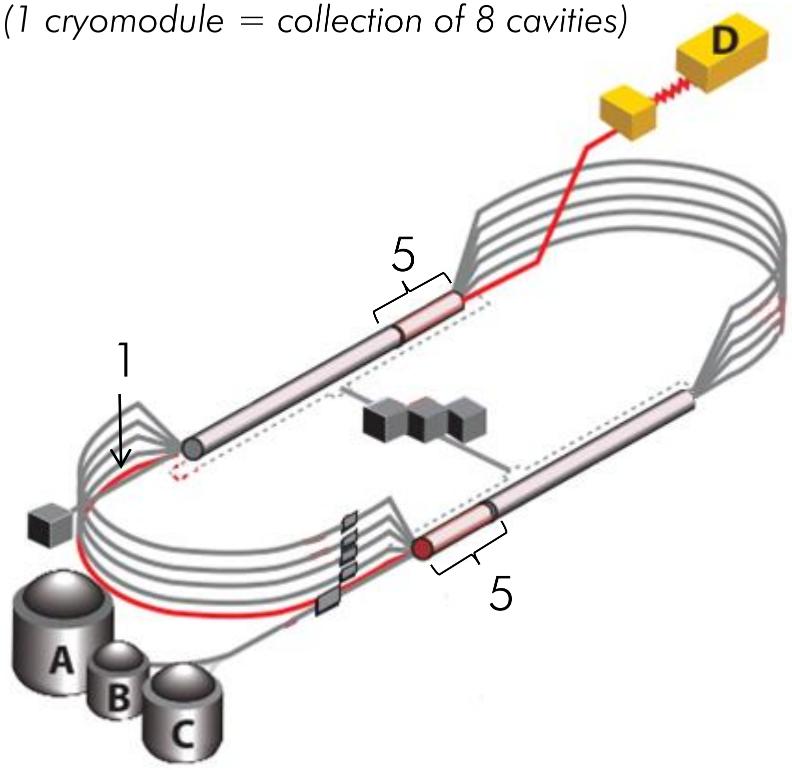


*U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Volume 17, Number 3 (1996).

Defining the Problem

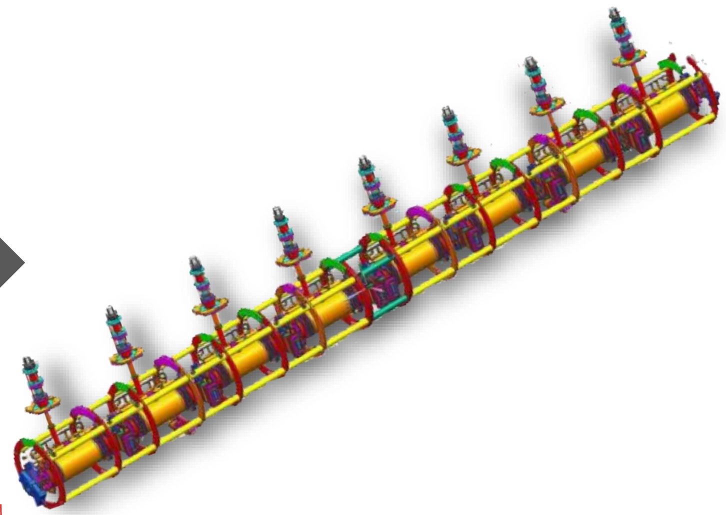
we have the ability to record data from 11 cryomodules in CEBAF

(1 cryomodule = collection of 8 cavities)



Q1: which of the 8 cavities faulted first?

17 signals/cavity × 8 cavities = 136 signals



ML Task #1

Q2: what kind of trip was it?

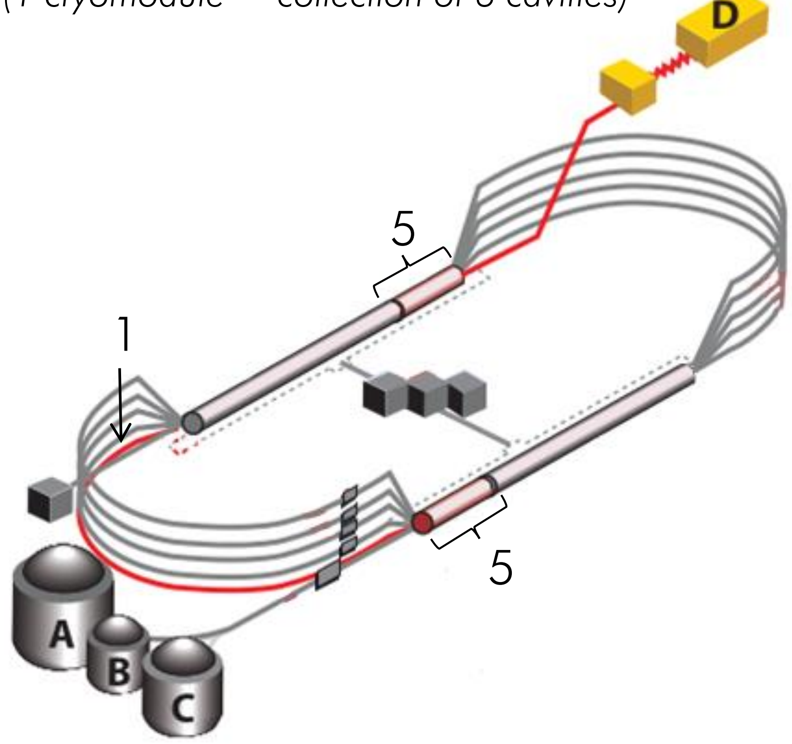
17 signals

ML Task #2

Defining the Problem

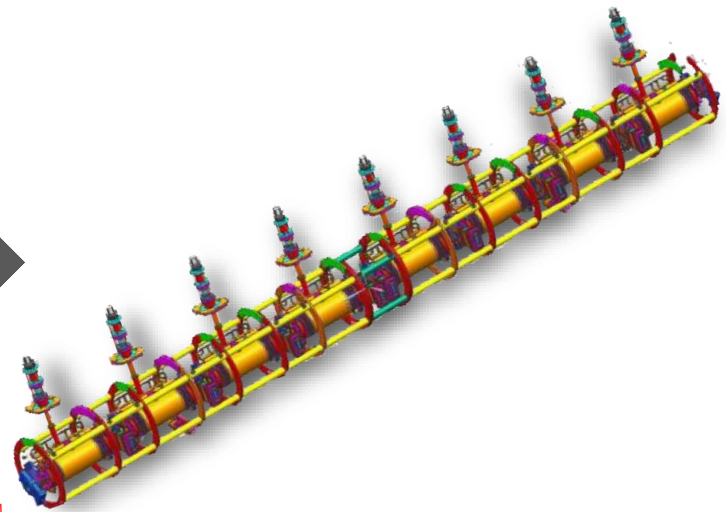
we have the ability to record data from 11 cryomodules in CEBAF

(1 cryomodule = collection of 8 cavities)



Q1: which of the 8 cavities faulted first?

17 signals/cavity × 8 cavities = 136 signals



ML Task #1

Q2: what kind of trip was it?

17 signals

ML Task #2

train a model to correctly classify the cavity and type of RF fault given waveform data

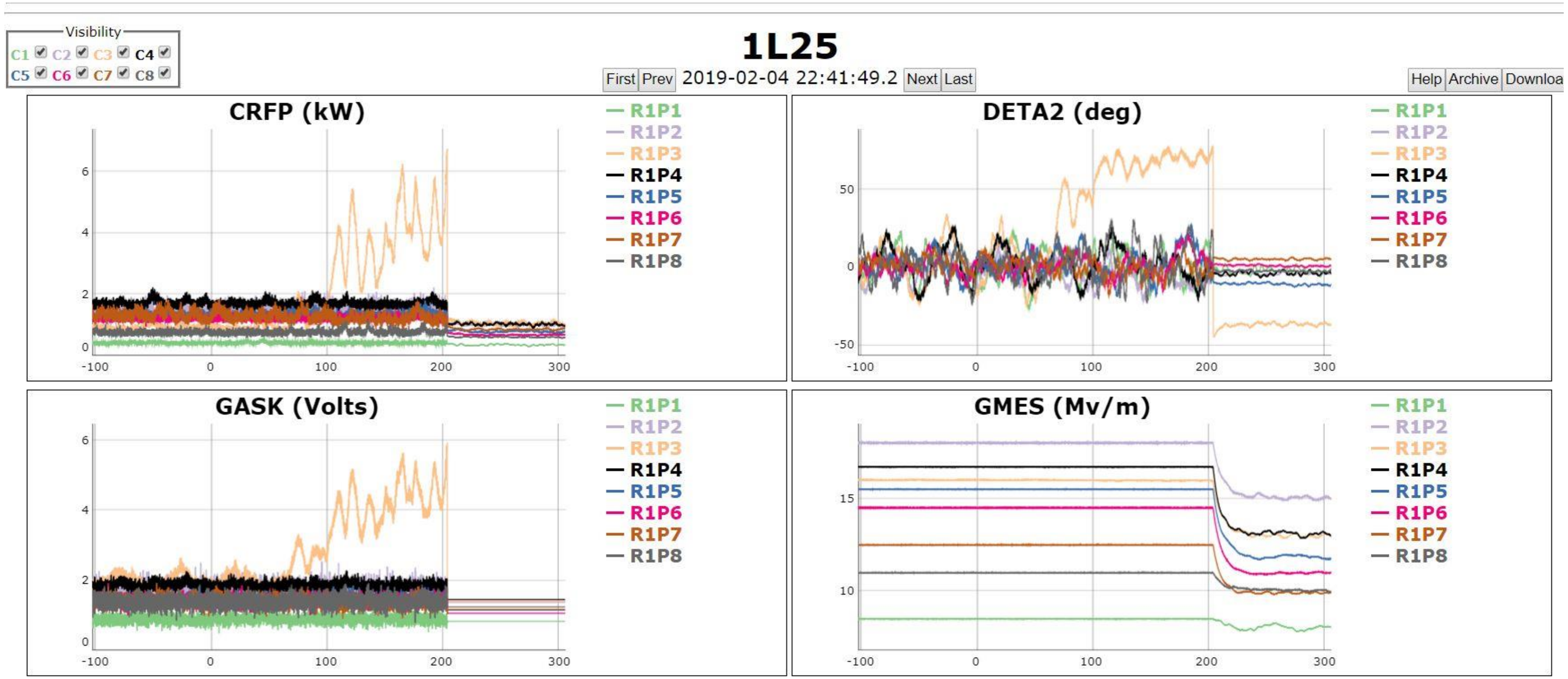
machine learning

multi-class classification

time-series data

Waveform Data for a Single Trip Event

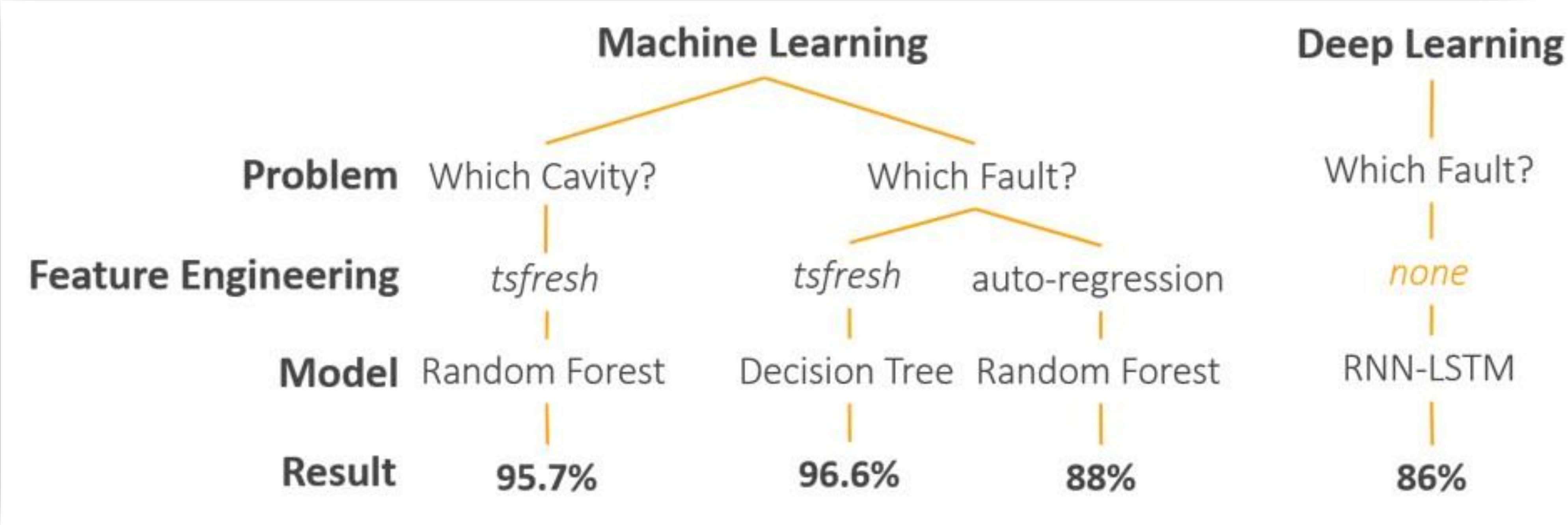
- using machine learning to automate the classification means:
  - ✓ results can be near real-time
  - ✓ frees up valuable subject matter expert time
  - ✓ provides important feedback to control room operators



17 signals/cavity × 8 cavities = 136 traces

Promising Initial Results

- using conventional machine learning tools and also deep learning architectures, we have achieved excellent results for predicting the cavity ID and type of cavity fault

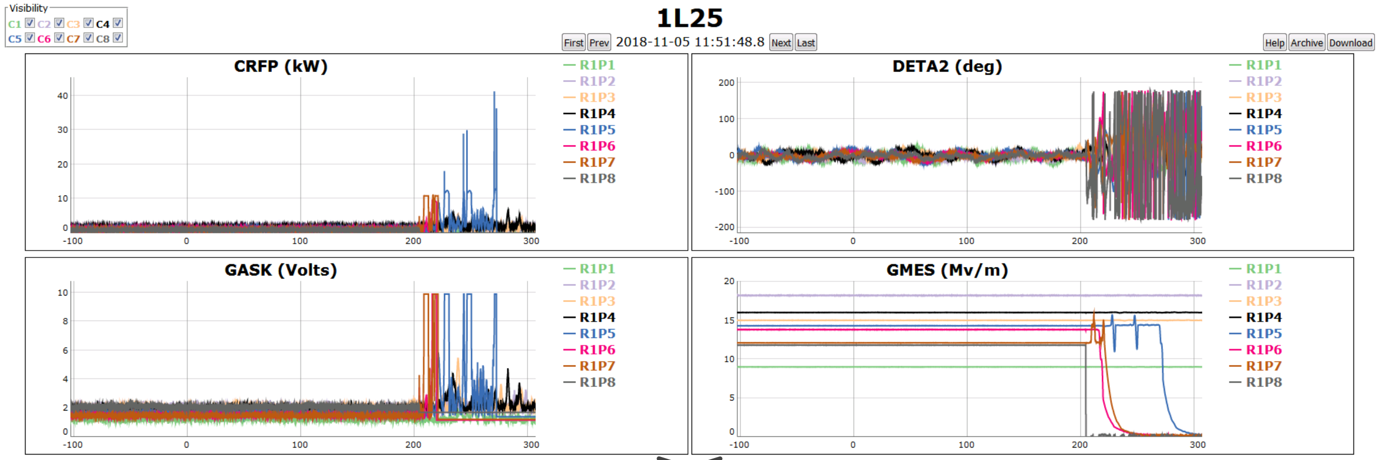


- this work has generated interest from other laboratories that are utilizing, or will in the near future, SRF cavities
- software is currently being development for <u>online deployment</u> of the system for the CEBAF fall 2019 physics run

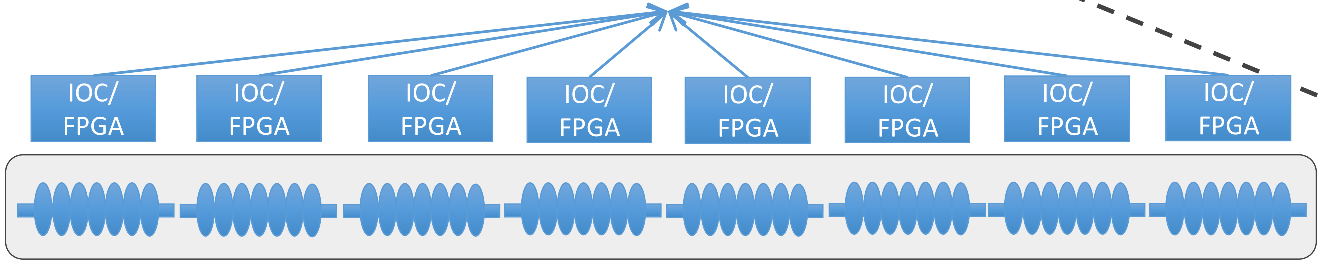
ML Implementation at CEBAF

- includes all 11 C100 zones
- harvester software saves waveform data from faults
- viewer software presents waveforms via web browser
- ML-based classifier labels a fault with responsible cavity and fault type
  - ✓ communicates results to control room operators

Waveform Viewer Software



Waveform Harvester Software

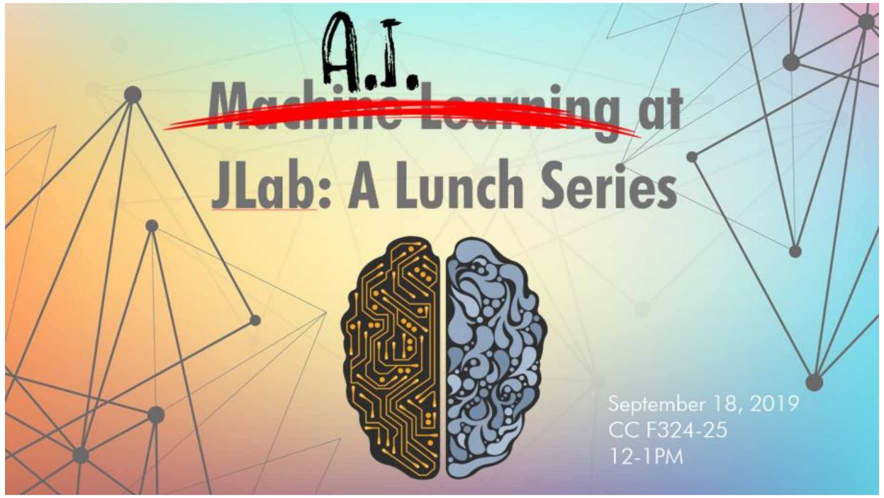


ML Fault Classifier

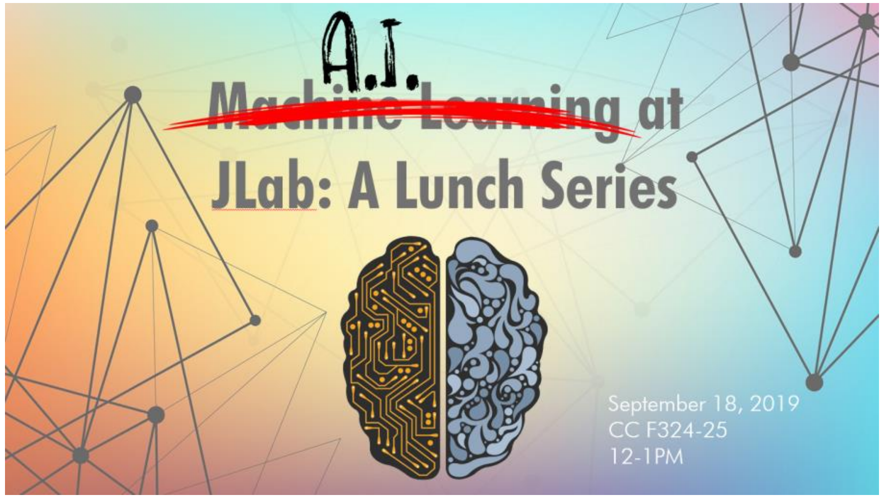

```
[
  {
    "location": "1L26",
    "timestamp": "2018-05-05 18:15:45.5",
    "cavity-label": "1",
    "cavity-confidence": 0.884,
    "fault-label": "E_Quench",
    "fault-confidence": 0.824
  }
]
```


See A. Carpenter's poster WEPHA025 for details!

Community Building at Jefferson Laboratory



Community Building at Jefferson Laboratory



- Jefferson Lab is a single-purpose laboratory
- data comes from
  - ✓ experimental end stations
  - ✓ accelerator

Experimental End Stations

- Particle Tracking
- Particle Identification
- Data Quality Monitoring*
- Efficient Data Reduction
- Detector Design

Accelerator

- SRF Fault Classification*
- Latent Knowledge in Archived Data

Moving Forward

- start building toy model problems with curated data sets among laboratories

AI LUNCH SERIES PROBLEM OF THE QUARTER

View published | New draft | Moderate

Problem 2: State Vector Prediction in a GlueX Tracking Detector

The second "Problem of the Quarter" has been released! In this problem participants will try to train AI to accurately predict a particle's position and momentum as it flies through the GlueX Forward Drift Chamber (FDC). Interested in joining in? More information, as well as training data may be found at https://halldweb.jlab.org/talks/ML_lunch/Sep2019/

Problem 1: Tracking a particle through a drift chamber analog

The inaugural problem is over and the winner is in; congrats to Adam Carpenter. Even though this round is over you may attempt the problem yourself in your own time at https://halldweb.jlab.org/talks/ML_lunch/May2019/

Thank you.