NPPS Intro Talk: Former Activities & Plans for DUNE

with focus on databases

Lino Gerlach, Paul Laycock

Date 2022-02-11

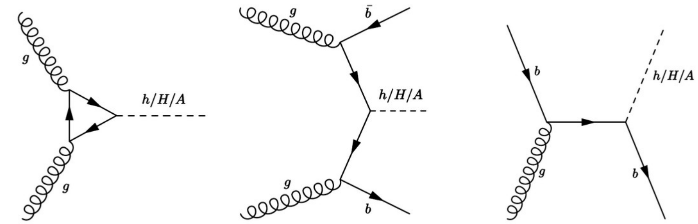
Overview

I’m thrilled to now be part of BNL & NPPS group!

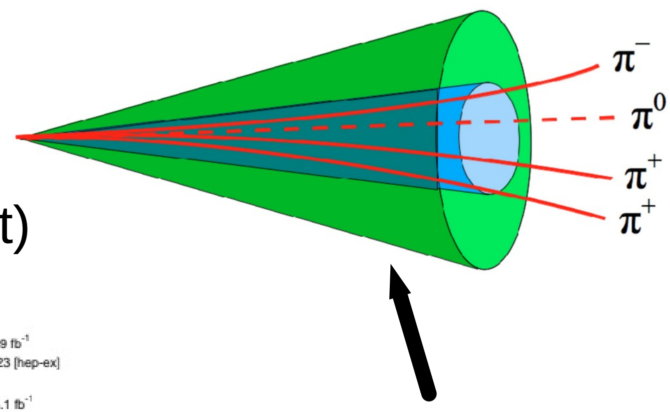
- My former activities:
  - Research at ATLAS
    - Quali task: Evaluate tauID performance
    - PhD thesis: Search for BSM A/H to tautau
  - Other activities
    - Deep Learning applications for LiDAR sensors

- Computing challenges at DUNE
  - My planned involvement

Activities at ATLAS



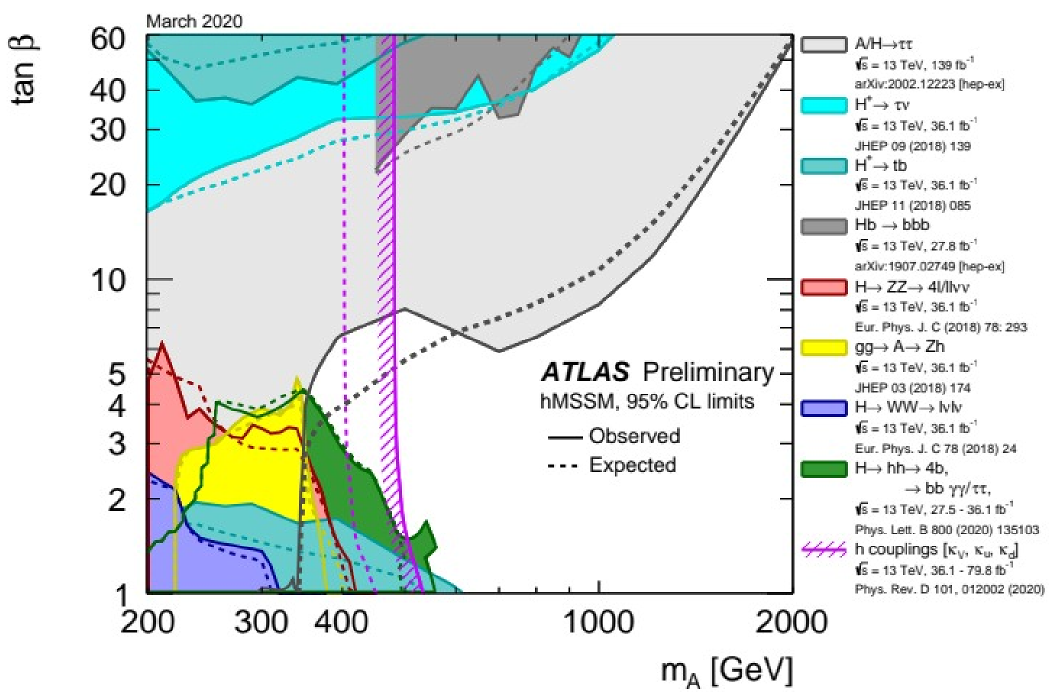
- Search for BSM A/H to tautau in fully hadronic decay channel:
  - Best limits in large part of MSSM parameter space
  - Special challenges:
    - Background from QCD
    - Mass reconstruction (≥ 2ν per event)



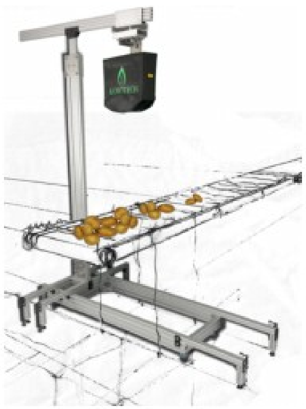
Looks like a QCD jet!

Deploy RNN-based classifier

Derived Scale Factors to qualify as ATLAS author



Machine Learning for LiDARs - I

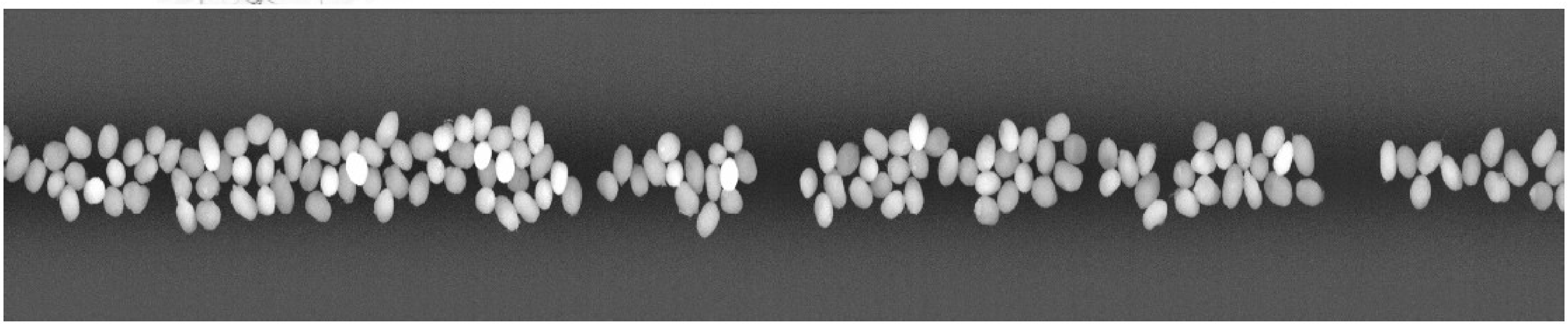


~ 500 kHz

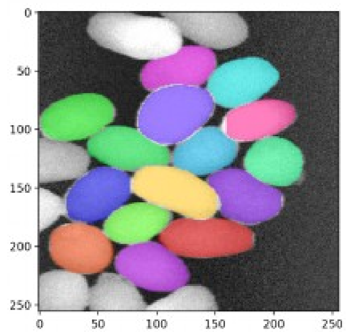
Real Time Output

RAW DATA

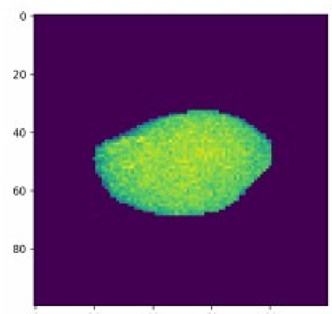
3D Reco



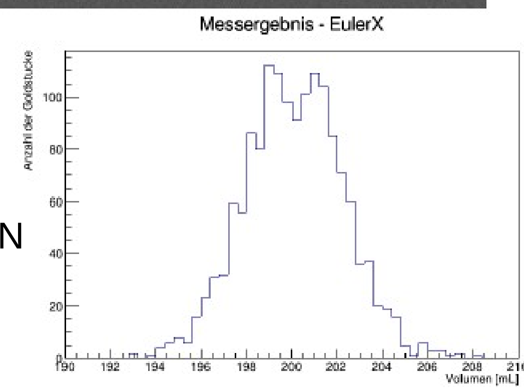
Segm.
mRCNN



Isolation



Regression
custom CNN



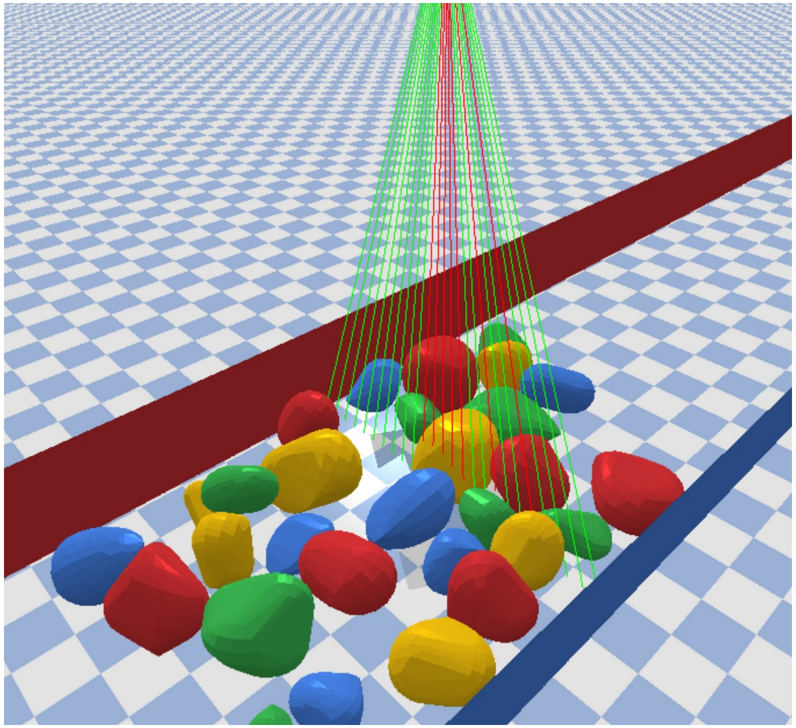
Machine Learning for LiDARs - II

Supervised learning approach

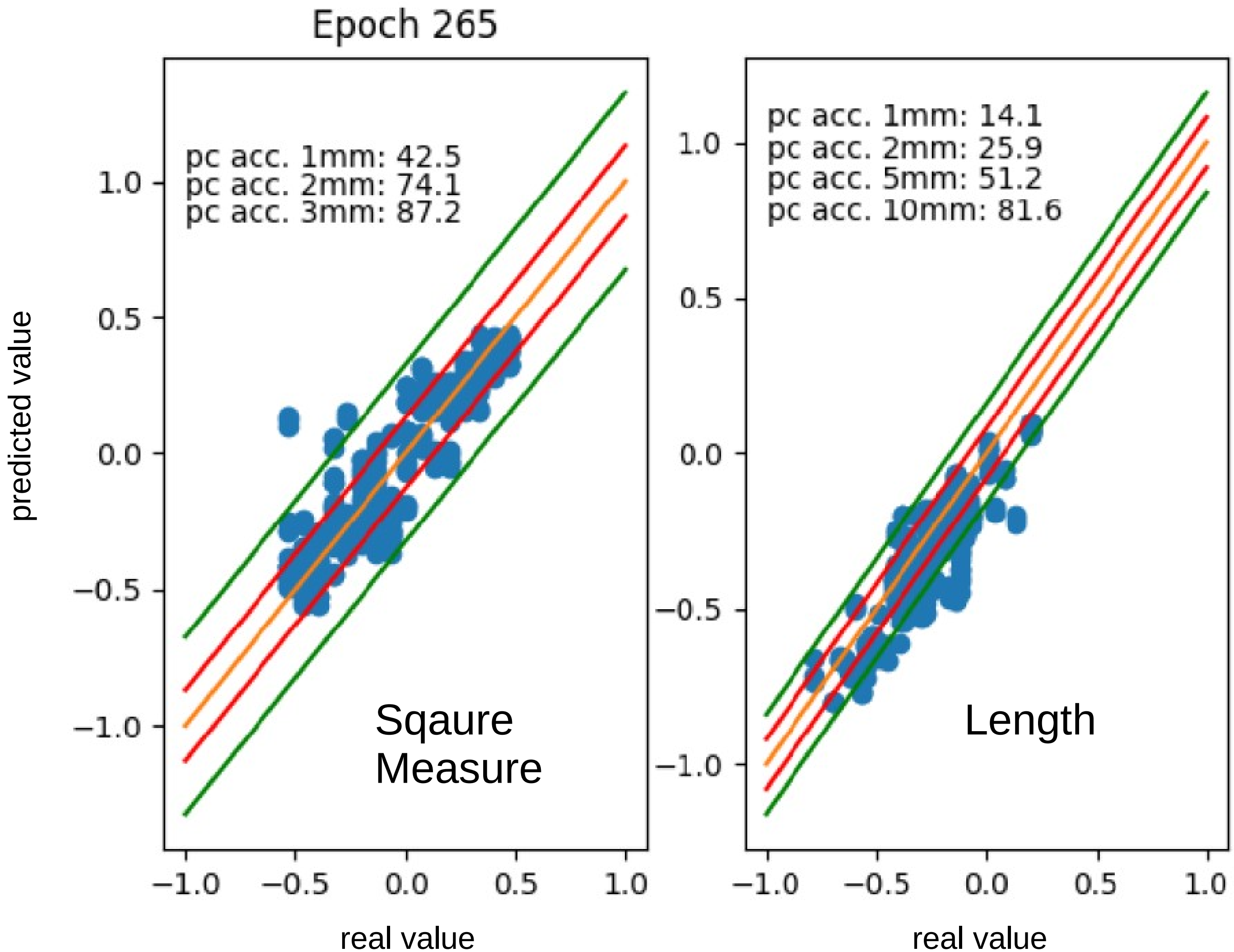
- Need large labelled data set

Developed Simulation

- MC potatoe generator
- Physics engine
- Detector simulation

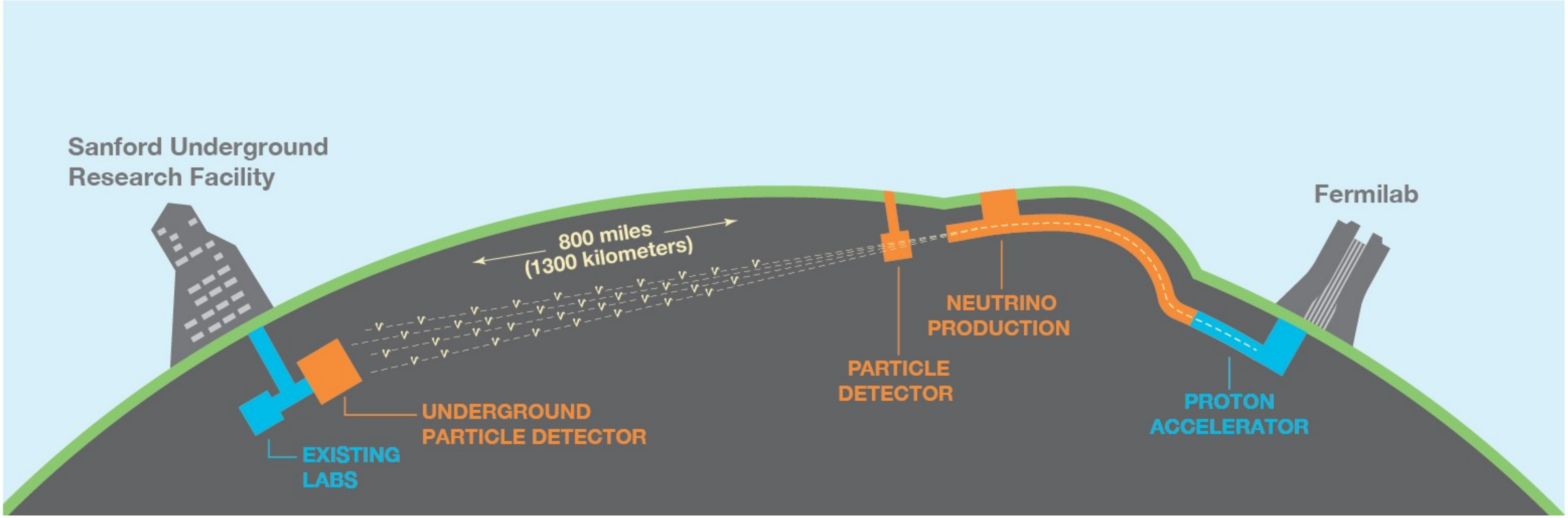


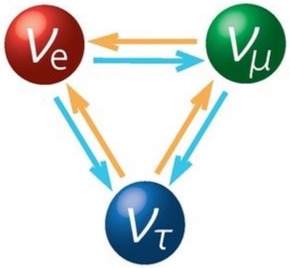
Machine Learning for LiDARs - III



Validated precision close to detector resolution

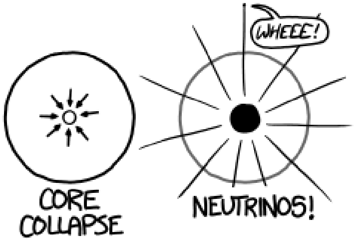
Deep Underground Neutrino Experiment



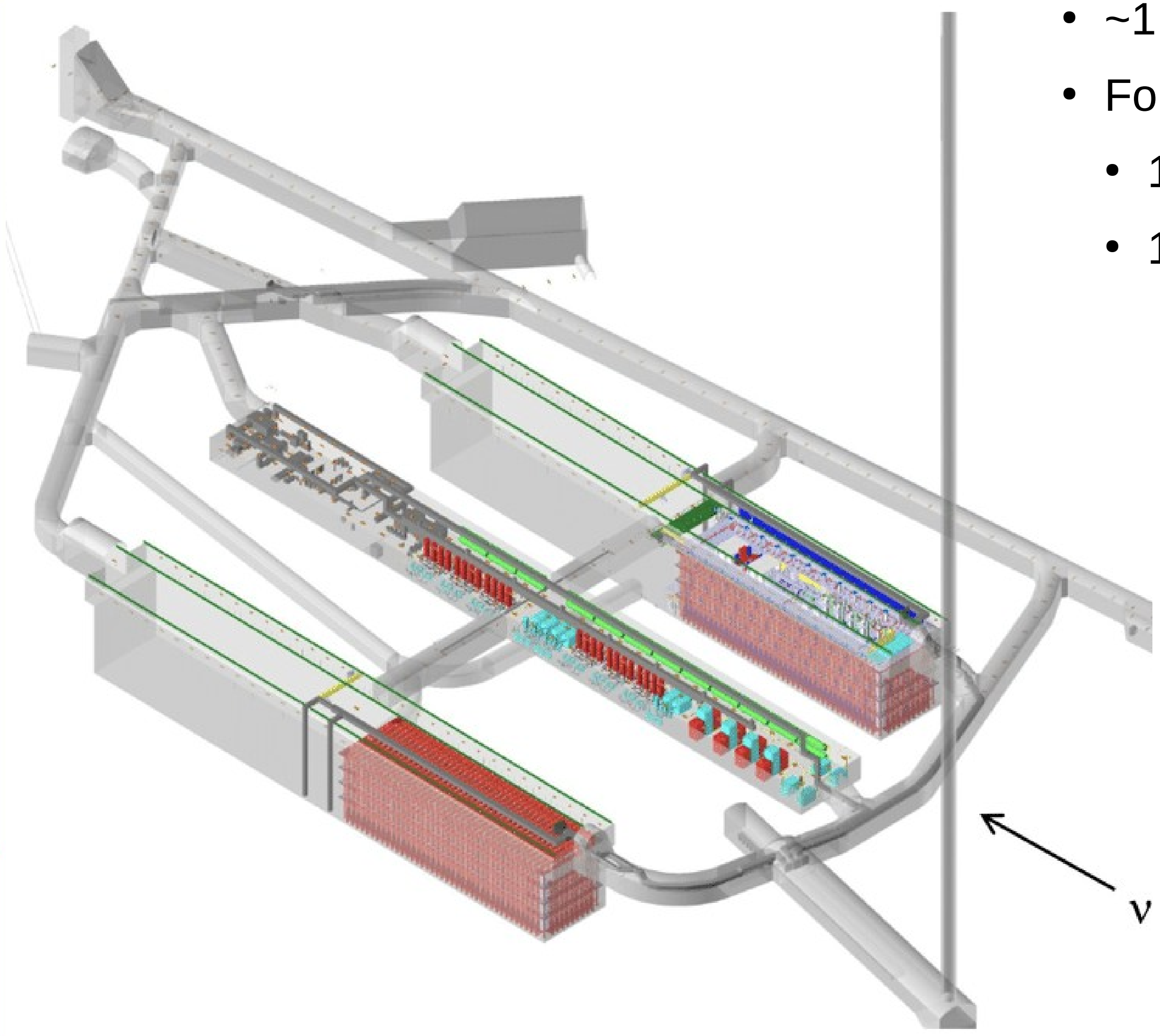


Three primary physics goals:

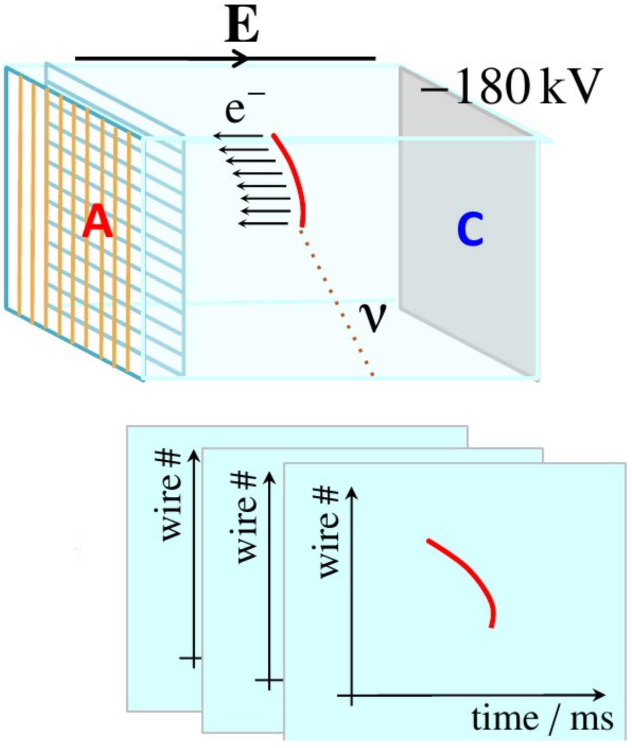
- **Determine CP violation in neutrino sector**
- **Investigate supernovae**
- **Search for proton decay**



DUNE Far Detector



- ~1 mile underground
- Four LAr TPC modules
  - 17 kt LAr each
  - 150 modules (APA) each



Computing Challenges for DUNE – I

DUNE will observe neutrino interactions at highest rate so far

- Overall amount of data is not too high (10-30 PB / year)
- But: DUNE trigger records (events) are large:
  - 150 APAs with 2560 wires each
  - Read-out 12-bit ADCs every 0.5 μs for ~6 ms
  - Roughly 6 GB uncompressed data per module per trigger record
  - Reading one full event into memory not feasible!
  - Sub-event processing necessary

Experiment	RAW event size
ATLAS	3 MB
protoDUNE	200 MB
DUNE	6 GB
DUNE (Supernova)	460 TB

Computing Challenges for DUNE – II

Memory management not the only challenge:

- Future computing infrastructure unknown
- Compatibility with external software packages (AI / ML)
- Proper handling of conditions data

Will focus on this

Goal: leverage expertise from other experiments through HSF

HSF

HEP Software Foundation

Conditions Data - Intro

“Conditions data is any additional data needed to process event data”

- Many different sources of conditions data
- Not yet fully understood, what information is really needed
  - This must be figured out before designing a common approach
  - ProtoDUNE is a good testing ground
    - Currently: no unified approach – patchwork of different solutions

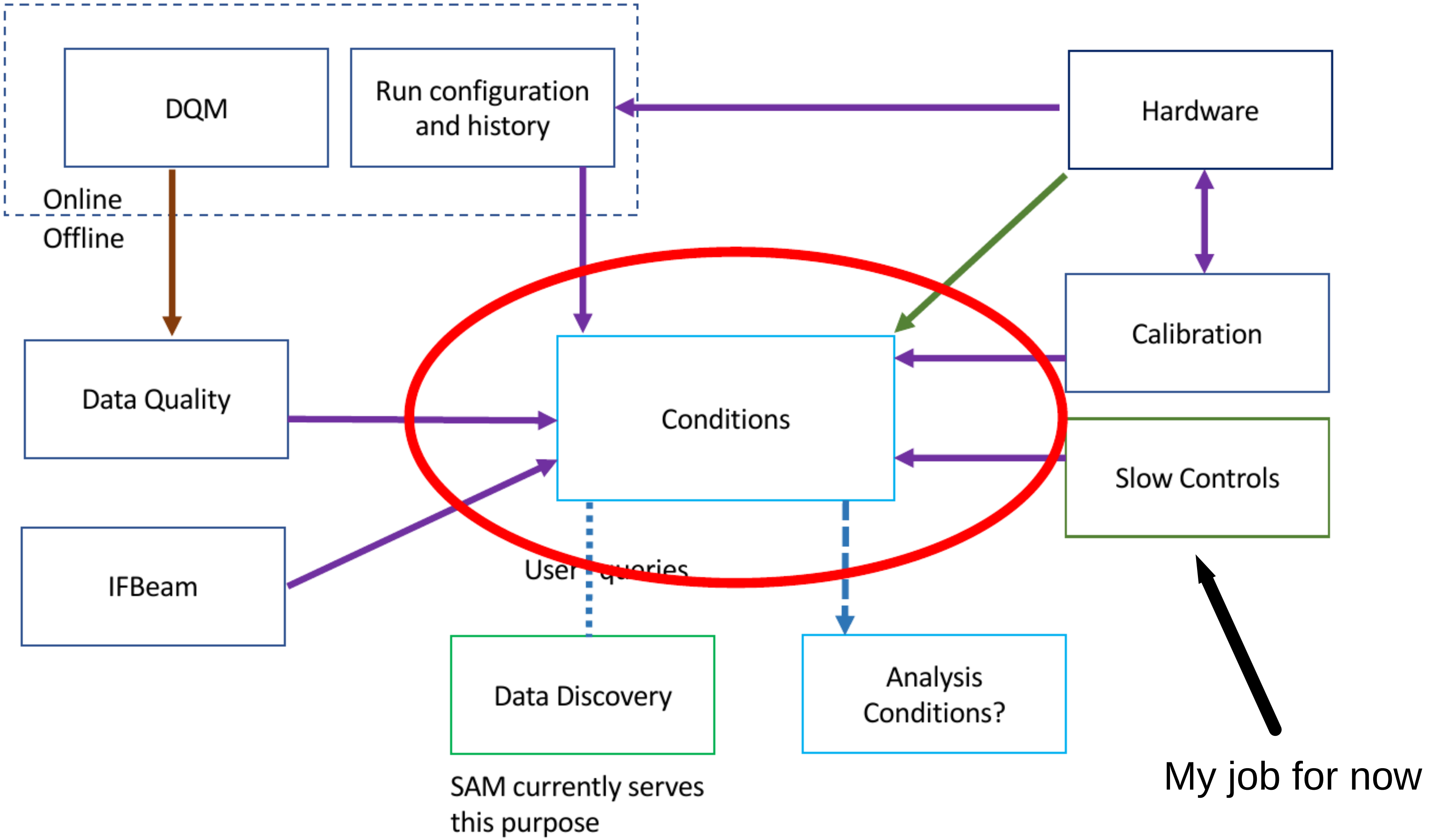
Heterogeneous sources of conditions data

Source (raw data)	Indexed by
Run metadata	Run number
Slow controls	Time stamp
Detector status	APA number
Geometry	Global

Common conditions database:

- Interface as homogeneous as possible

Conditions Data - Sources

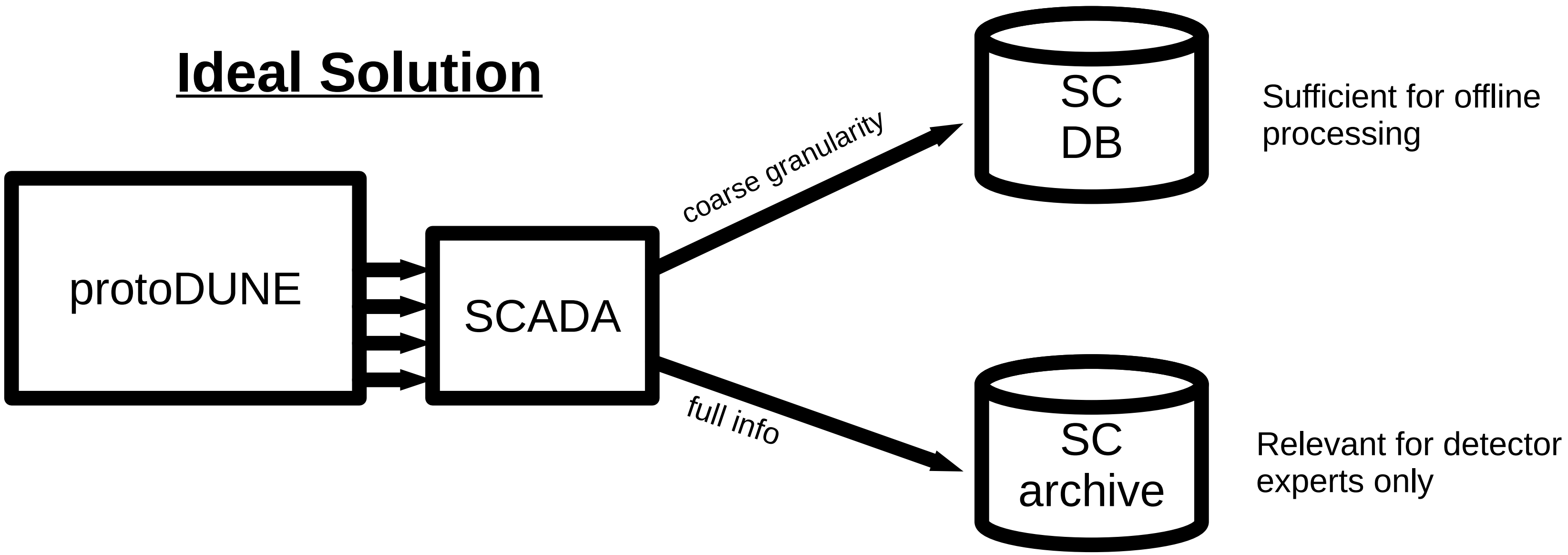


graph made by Paul Laycock / Norm Buchanan

Slow Controls (SC)

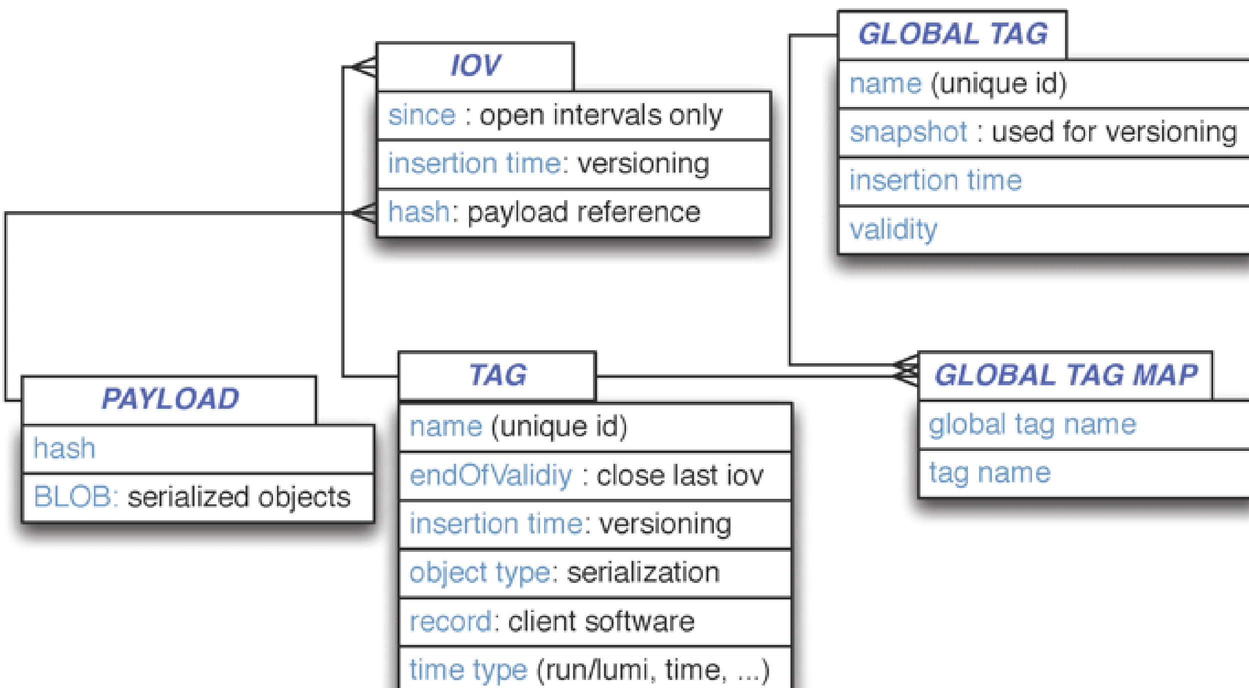
- SCADA system records raw data (indexed by time stamp)
  - Stored in ‘SC archive’
- SC experts operate SCADA, not the the database
  - Data base group (we) have to provide the database + offline access
- Problem: raw data written w/ very high granularity
  - Way more granular than needed for offline processing

Ideal Solution



HSF Recommendations – Cond. Data Model

- Dedicated HSF conditions data activity:
  https://hepsoftwarefoundation.org/activities/conditionsdb.html
  - Loose coupling between client and server using RESTful interfaces
  - The ability to cache queries as well as payloads
  - Separation of payload queries from metadata queries



Marko Bracko et al: EP Software Foundation Community White Paper Working Group – Conditions Data

Current State of conditions database

- Igor Mandrichenko wrote “Unstructured Conditions Database (UconDB)”
  - Command line and API

```
$ ucondb folders
```

```
sp_protodune
test
```

```
$ ucondb objects test
```

```
file.dat
file_sc2.dat
```

```
$ ucondb versions test file.dat
```

```
id         key Tr (UTC)            Tv                 Size
---------- --- ------------------- ------------------ ---------
69             2020-02-06 19:43:36                0.000      6
```

```python
from ucondb.webapi import UConDBClient
client = UConDBClient("https://dbdata0vm.fnal.gov:9443/protodune_ucon_prod/app")
data = client.get_data("sp_protodune", version_id=7183)
print(f"type(data): {type(data)}")
```

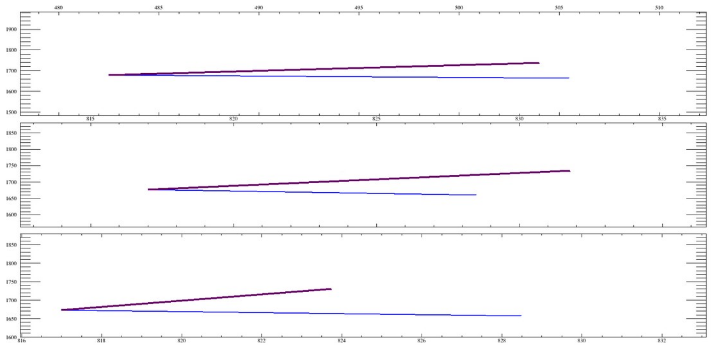
```
type(data): <class 'bytes'>
```

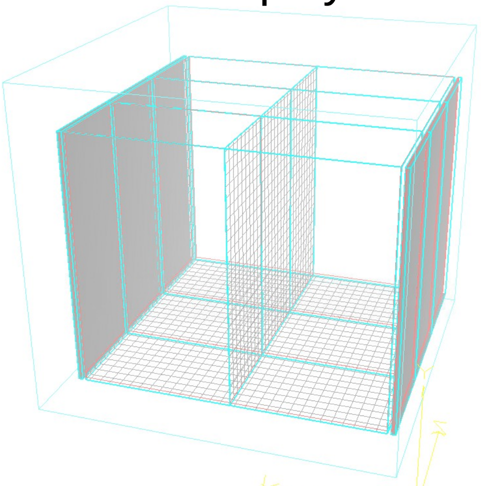
Retrieving the data via the API: type = ‘bytes’

- So far, only Run config meta data is stored in UconDB (incomplete)
- Igor provided login to experiment with uploading data
- Ana Paula (CSU): include DAQ conditions data
- Me: include slow controls conditions data

Finding database accesses in dunesw

- Familiarizing myself with DUNE software stack
  - Simulation (evGen, G4, detSim, reco) looks reasonable
  - Reconstructed real events appear empty in event display





- Identified four obvious DB queries so far: Corrections for lifetime, dQ/dx, X, YZ

DBWeb query: https://dbdata0vm.fnal.gov:9443/dune_con_prod/app/get?table=pdunesp.lifetime_purmon&type=data&tag=v1.1&t0=1539711086&t1=1539883886&columns=center,low,high
Got 3 rows from database
run: 5387 ; subrun: 1 ; event: 3
evttime: 1539797486
fLifetime: 17518.348506 [us]

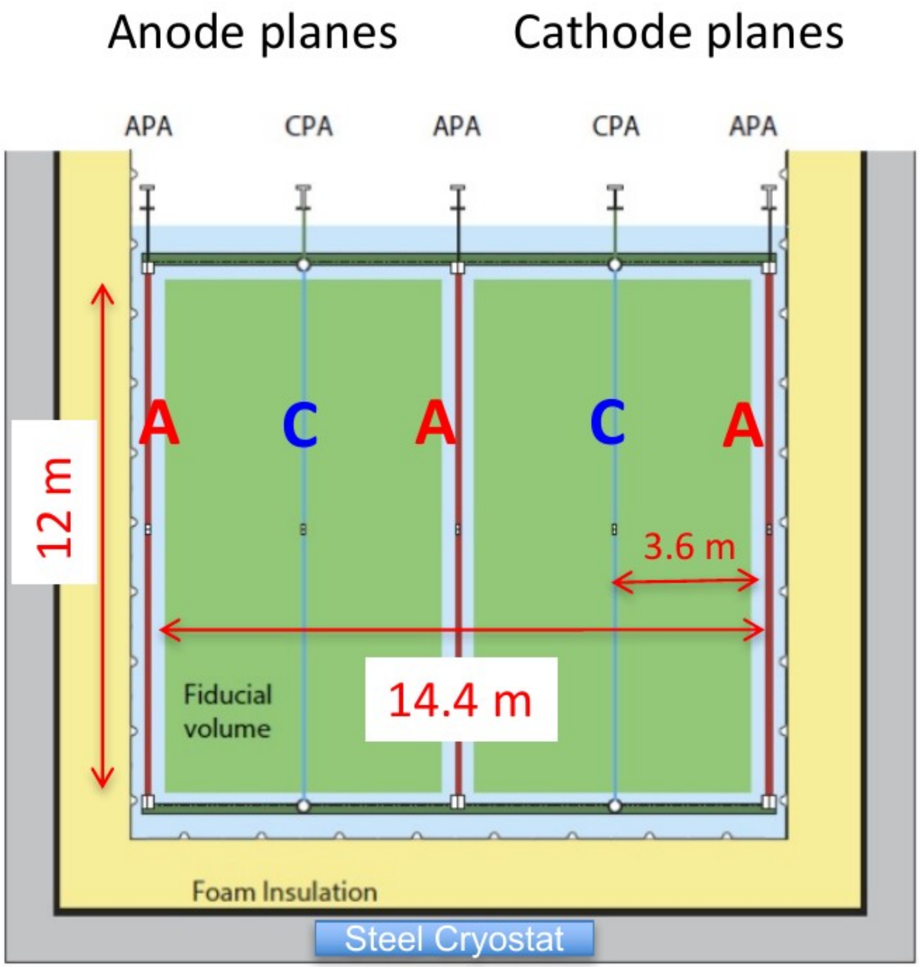
```
##   a) stage 1 with calibration sce, lifetime (protoDUNE_SP_keepup_decoder_reco_stage1.fcl)
##   b) stage 2 with calibration yz,x,t (protoDUNE_SP_keepup_decoder_reco_stage2.fcl)
```

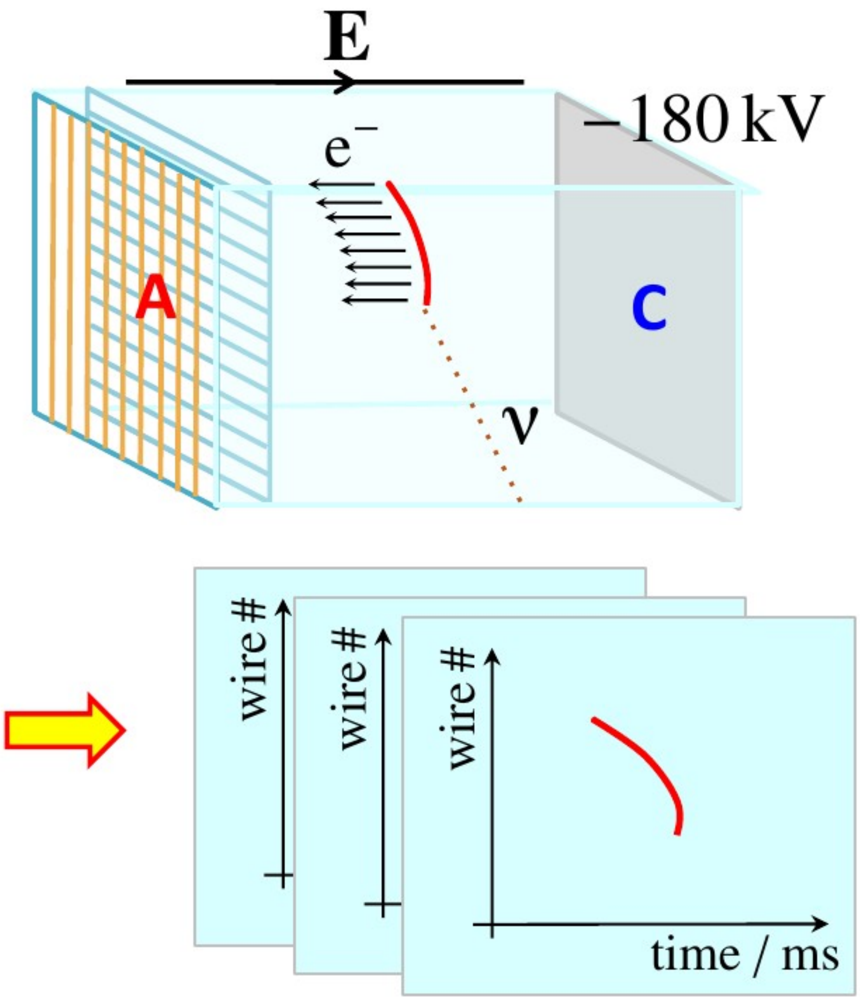
My Next Steps

- Mastering the DUNE software stack
  - Reproduce chain from raw to reconstructed protoDUNE data
  - Conduct example data analysis (typical use case)
- Identify which databases are accessed and why
- Drafting list of requirements for a centralised conditions database
  - ProtoDUNE as test ground for DUNE
- Get in touch with SC team
  - Understand SC conditions data needs

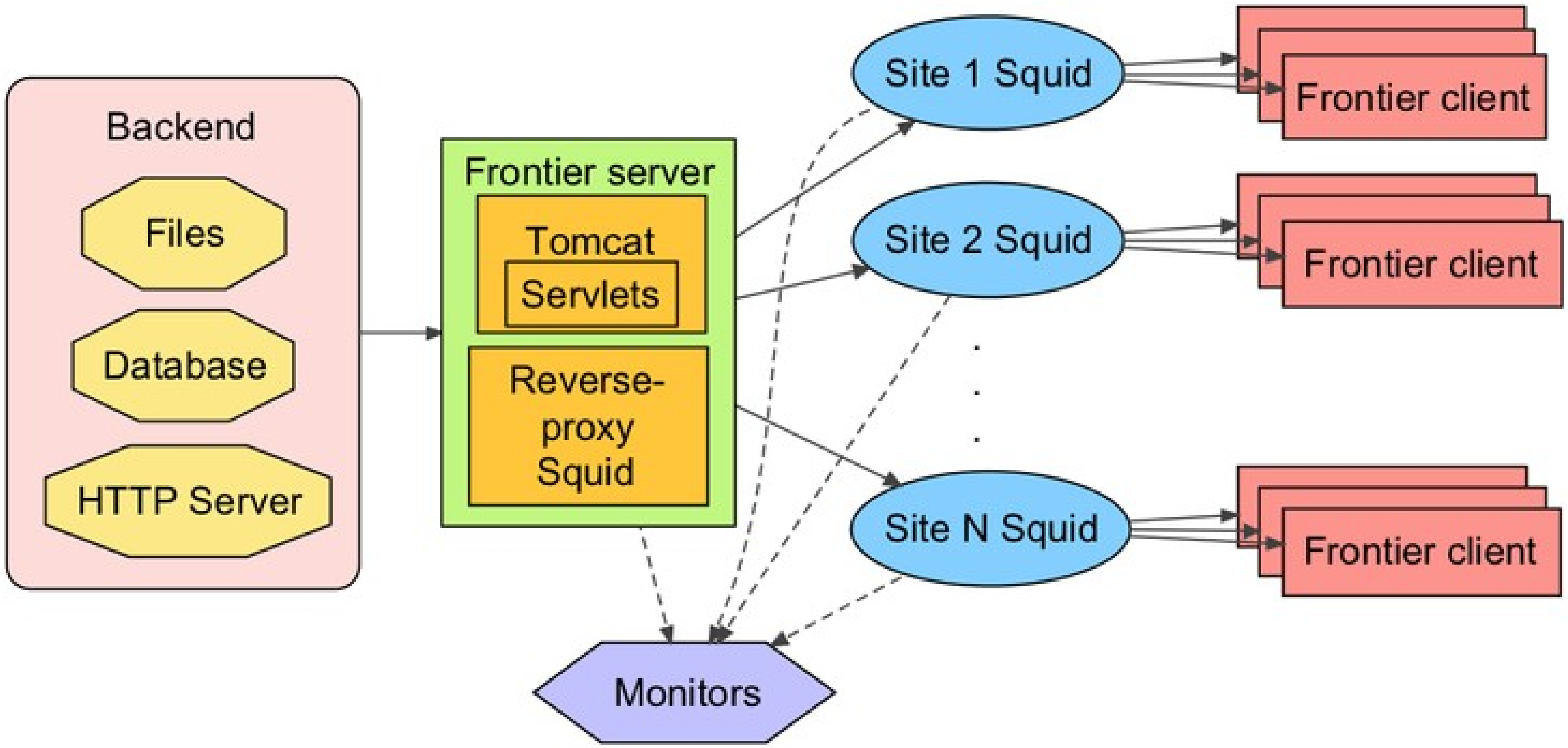
Thank you for your attention

DUNE Far Detector





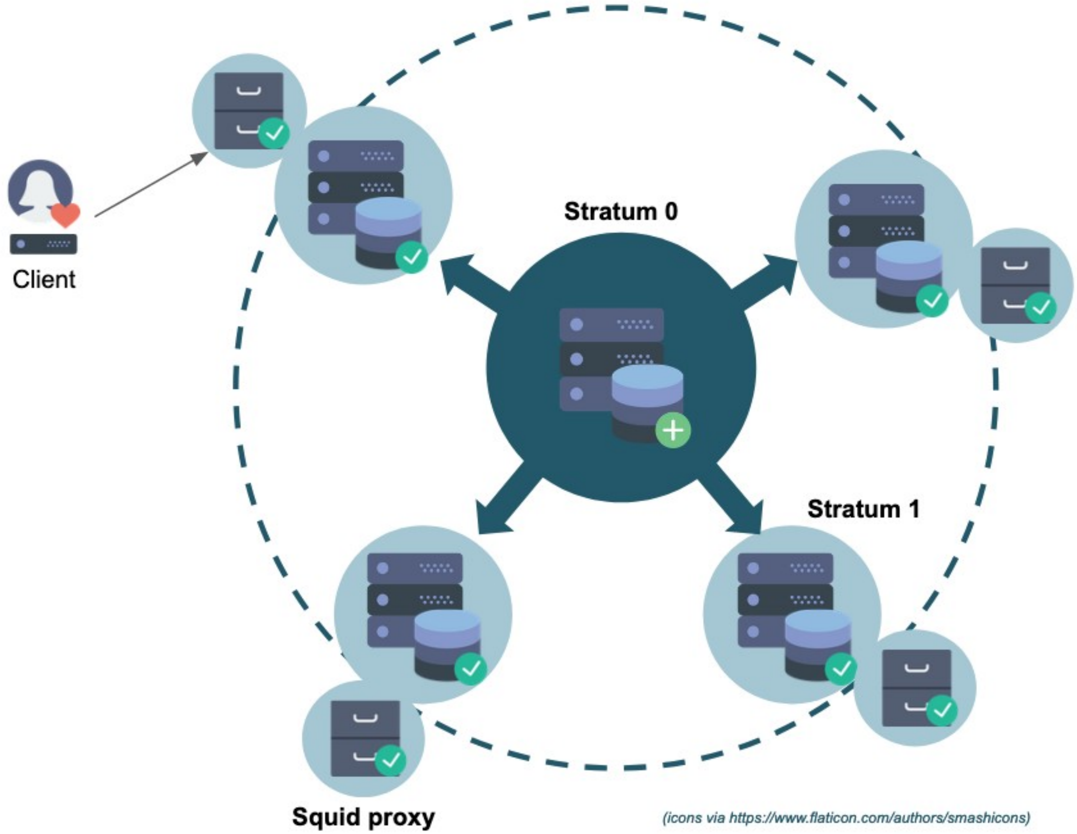
Frontier Architecture



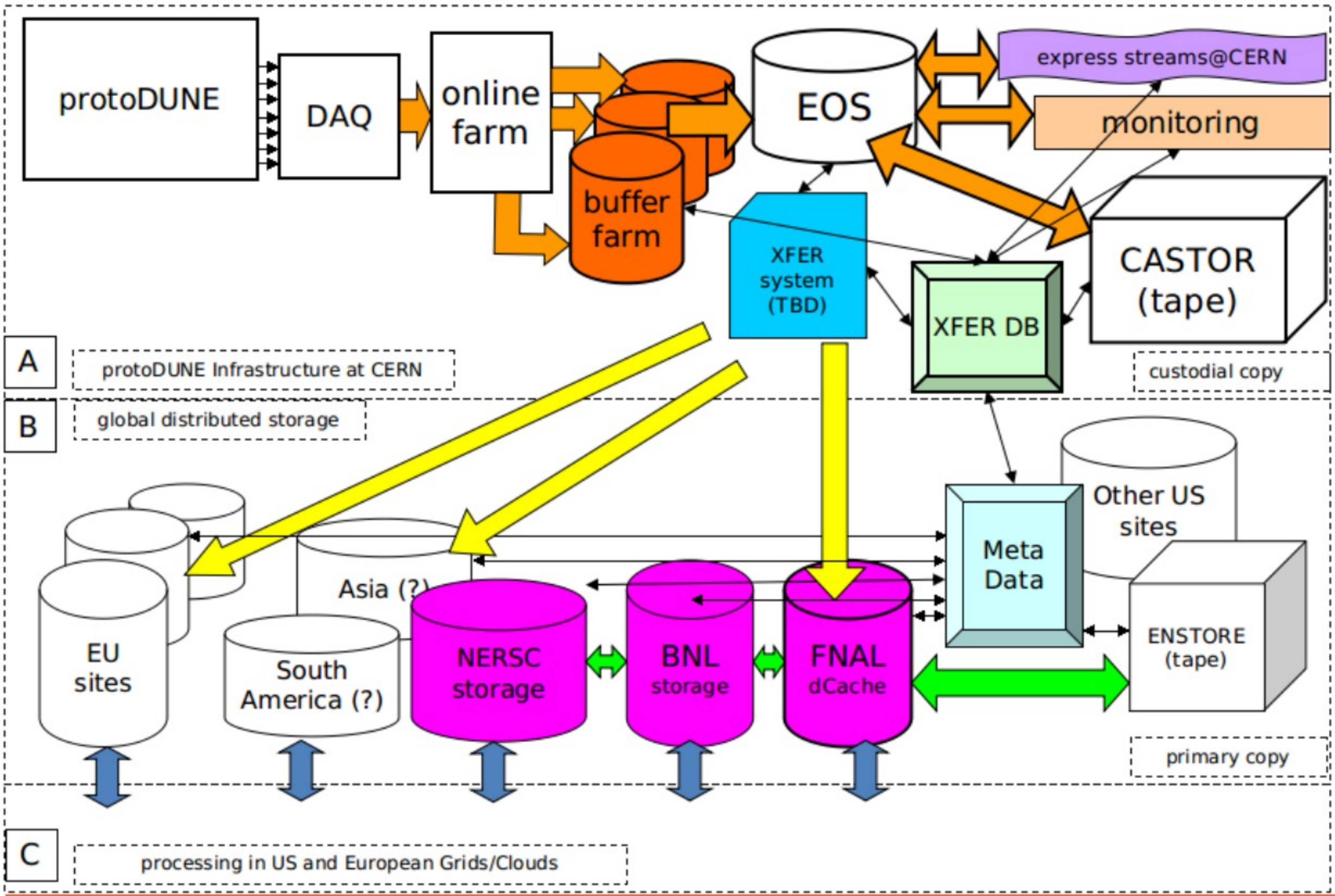
- Clients contact squid → Squid only contacts Frontier server if not already cached
- Frontier decodes request → contacts Backend
- Many powerful features: Queuing, load balancing, data compression...
- 2012 CMS study found 140:1 in requests 1000:1 in data reduction
  - 5 million responses from 3 Frontier servers per day (40 GB)
  - Squid caches served over 700 million (40 TB)

CernVM File System (CVMFS)

- Web based global file versioning system (POSIX, read-only)
- Originally designed & optimized for distribution of software installations
- Cryptographic hashes & signatures allow use of http → cacheable



Raw Data Flow in protoDUNE : the Concept



Fuess, S., Illingworth, R., Mengel, M., Norman, A., Potekhin, M., & Viren, B. Design of the protoDUNE raw data management infrastructure.

BDT vs RNN TauID

<u>BDT TauID</u>

- 12 'high-level' input variables

<u>RNN TauID</u>

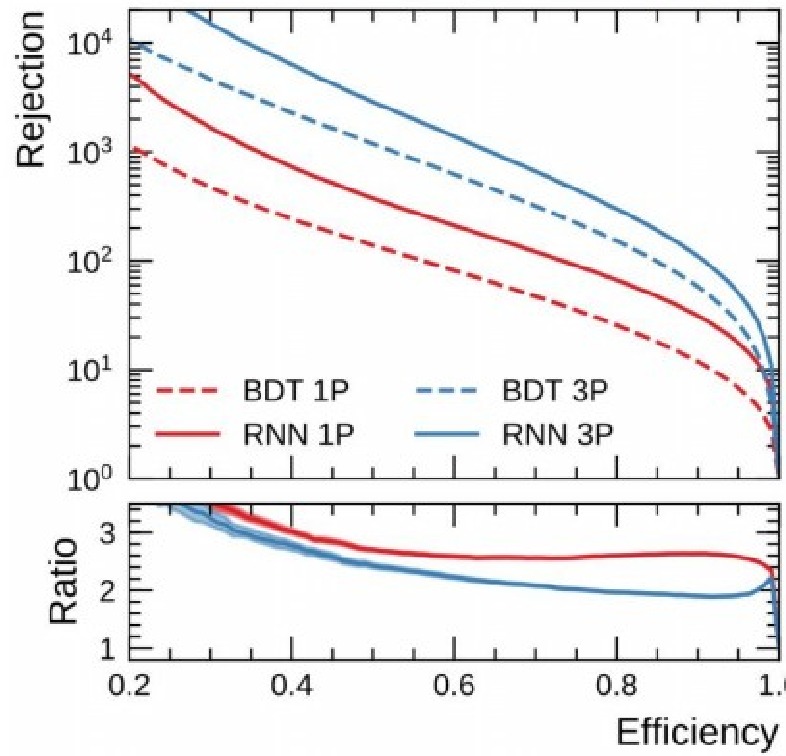
- BDT input variables
- Track-level variables
- Cluster-level variables

<u>RNN clearly outclasses BDT ID</u>

- Expect $\approx$ 30% higher di-Tau yield

But: New Scale Factors were needed for RNN ID by tauWG

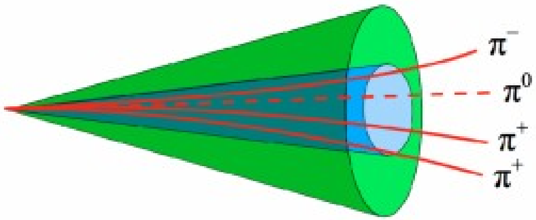
(Also BDT ID SF for full Run-2 dataset)

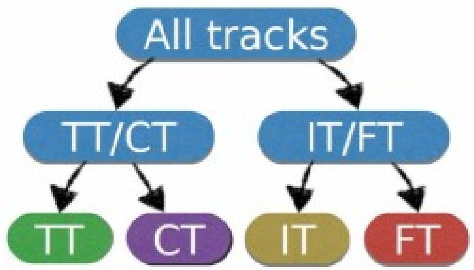


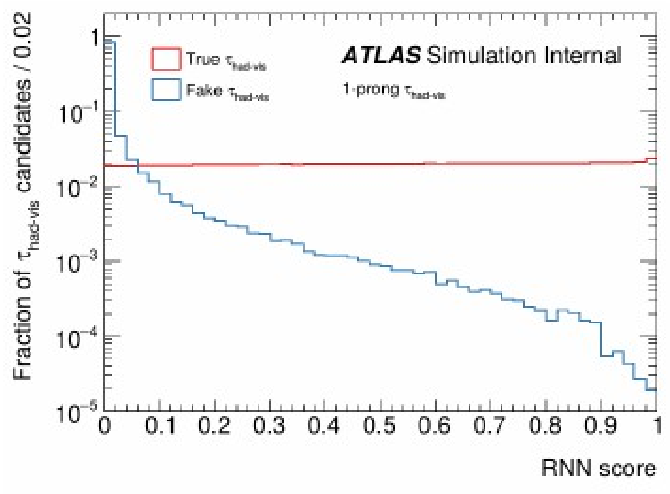
Plot by Chris Deutsch

Hadronically Decaying Tau-Leptons

- τ: only lepton that can decay hadronically
  - Mainly into Pions
- Start from 'jet' in calorimeter
  - Clustering algorithm 'anti-k_T'
- Use BDT to classify tracks into tau tracks conversion tracks isolation tracks and pile-up tracks
- After Reco: $\tau_{\text{had-vis}}$ candidates mostly jets from quarks/gluons
- RNN Identification ('ID') algorithm to discriminate against 'fakes'



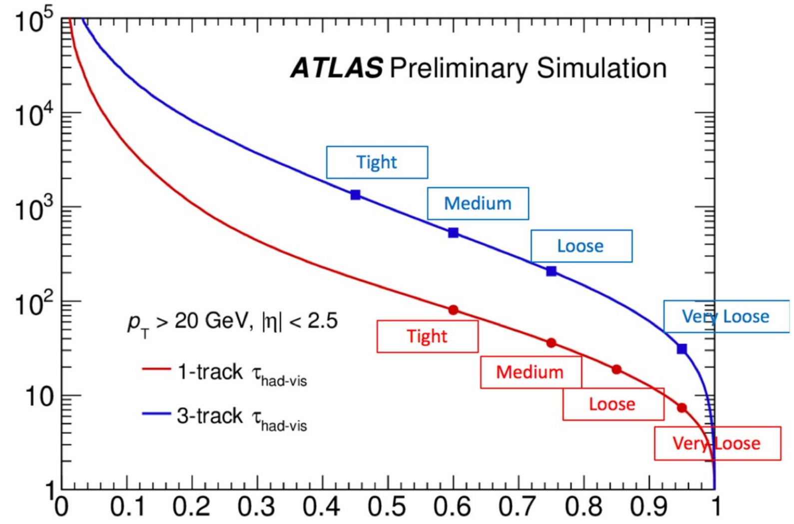


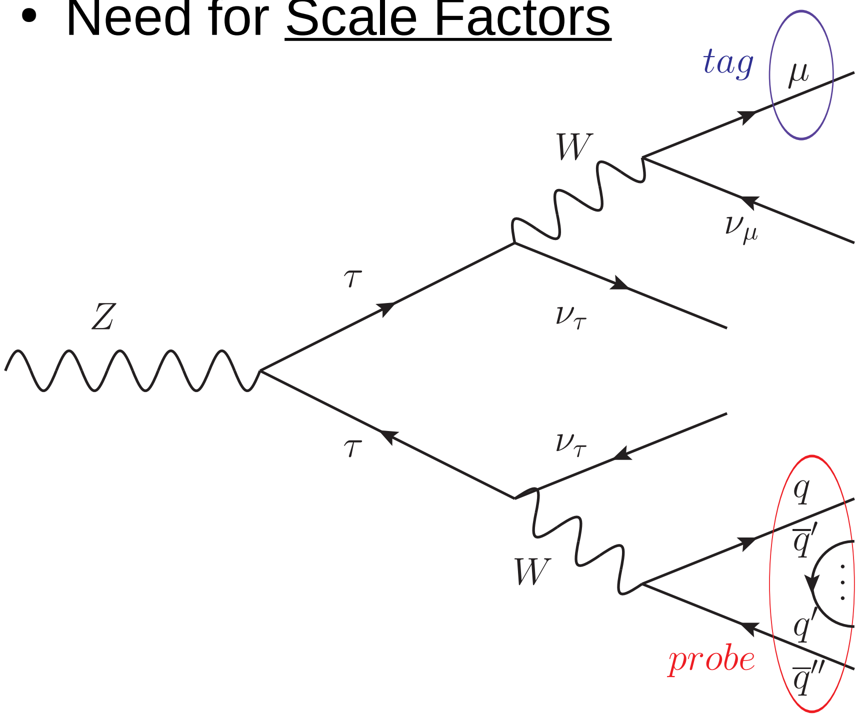


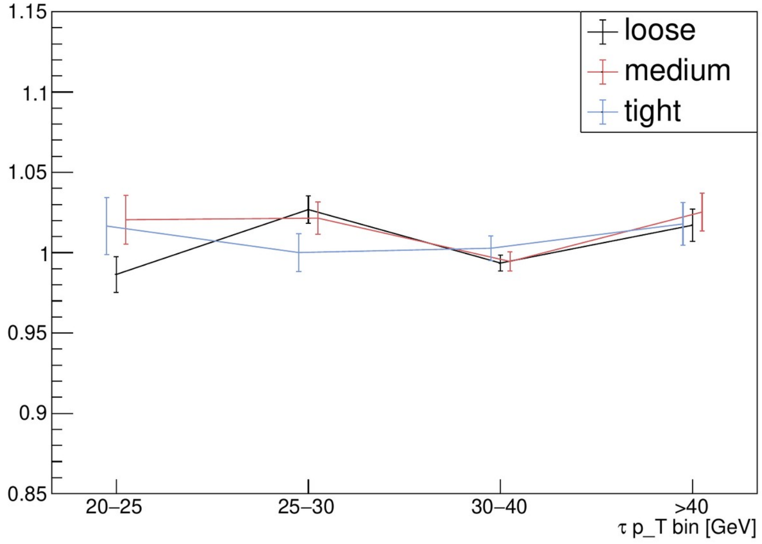
TauID Scale Factors

Analyses apply tauID to MC generated taus

- TauID might perform different on MC and real data
- Need for Scale Factors







Search for BSM A/H to tautau

- Tau-Leptons very promising for searches for BSM Higgs
- Two parameters to describe MSSM Higgs sector at tree level
- Focus on fully-hadronic di-tau final state
- Special challenges:
    - Much background from QCD
    - At least two neutrinos
        - Difficult masss reconstruction