Opportunities and Challenges in DUNE computing/offline

(computing model and status/needs for offline software)

Robert Sulej

European Neutrino Meeting LBNF/DUNE

April 7-8, 2016

Physics driving the computing:

DUNE

- Beam neutrinos
- Nucleon decays
- Cosmogenics

- Supernova neutrinos

- Near detector data

protoDUNE's

- Test-beam particles

- Cosmogenics

Challenges: data volume ● procesing ● reconstruction

Foreword

- A decade until DUNE FD commissioning, and ~2 years to protoDUNE's test-beam data
  - data managment, processing systems, reconstruction and analysis software deployed first for prototypes
  - scaling up and evolving these systems for the far detector
- DUNE is large intenational collaboration → distributed computing infrastructure, following LHC principles
- **Computing model designing just started and evolving.**
- **Reco algorithms still under development, new approaches arise.**
- Frequent MC productions, reconstruction development & tests → induces additional storage/processing
- High CPU requirements from the sophisticated reconstruction algorithms

Data volumes

- LArTPC high-energy event occupies: few thousands of wires × few thousands of ADC samples in 2 or 3 views

  → data size very comparable to a digital photograph

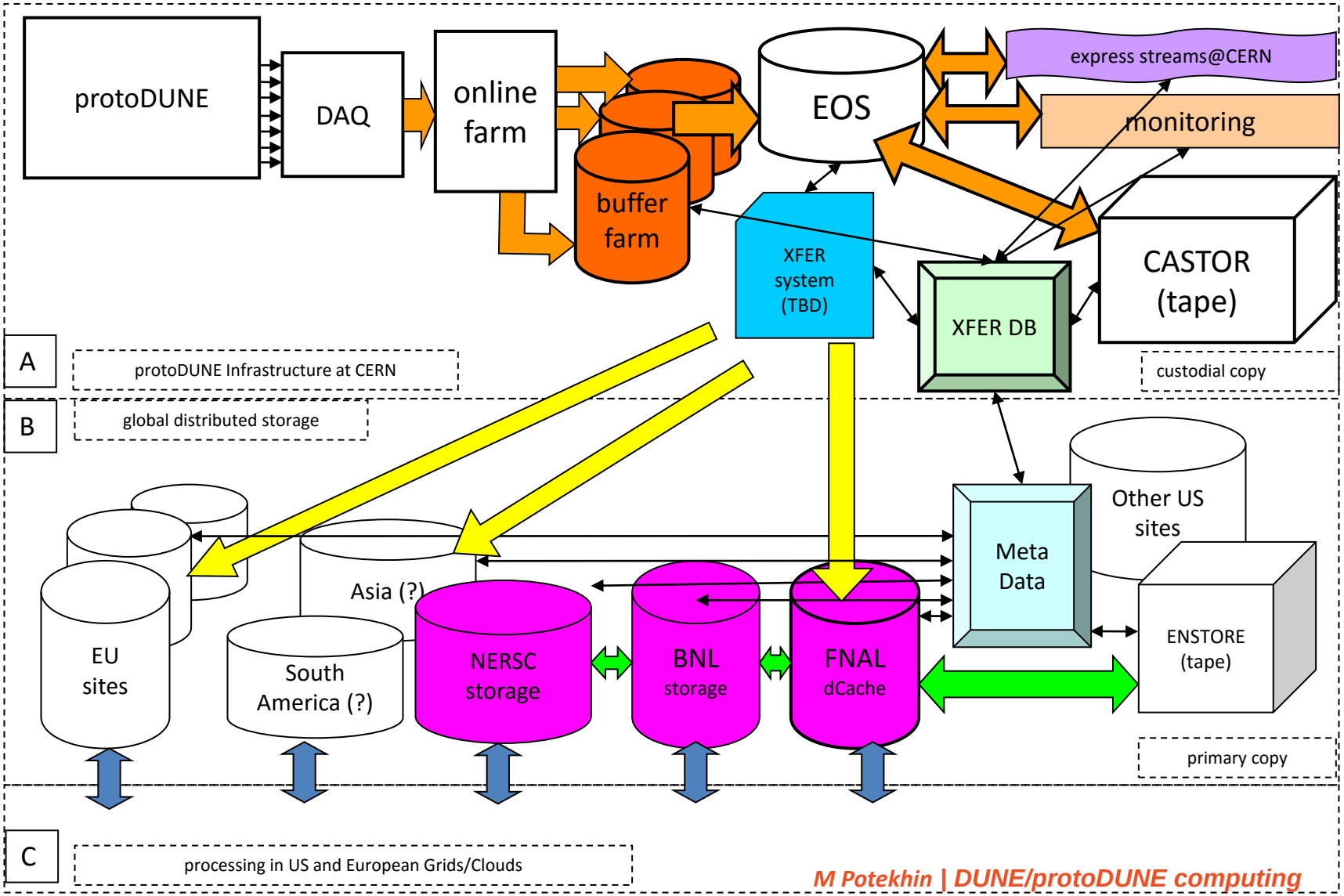
- **ProtoDUNE raw data: 2.5PB** (test beam, cosmics, commissioning)

- FD data/year, compressed: 5TB (cosmic μ), 0.5TB (nucleon decay candidates), 5GB (beam ν)
- SNB candidates data/year, full stream, uncompressed: **500TB**
- ND data/year: ~100TB (not knowing technology choice)
- MC data/year: 15TB (FD), 20TB (ND), 1TB (beamline, target)

+ Derived data (reconstruction, analysis), headroom for processing, software comparisons, … → factor of 8 over the raw data

DUNE total / year: 6.5PB

protoDUNE's data flow



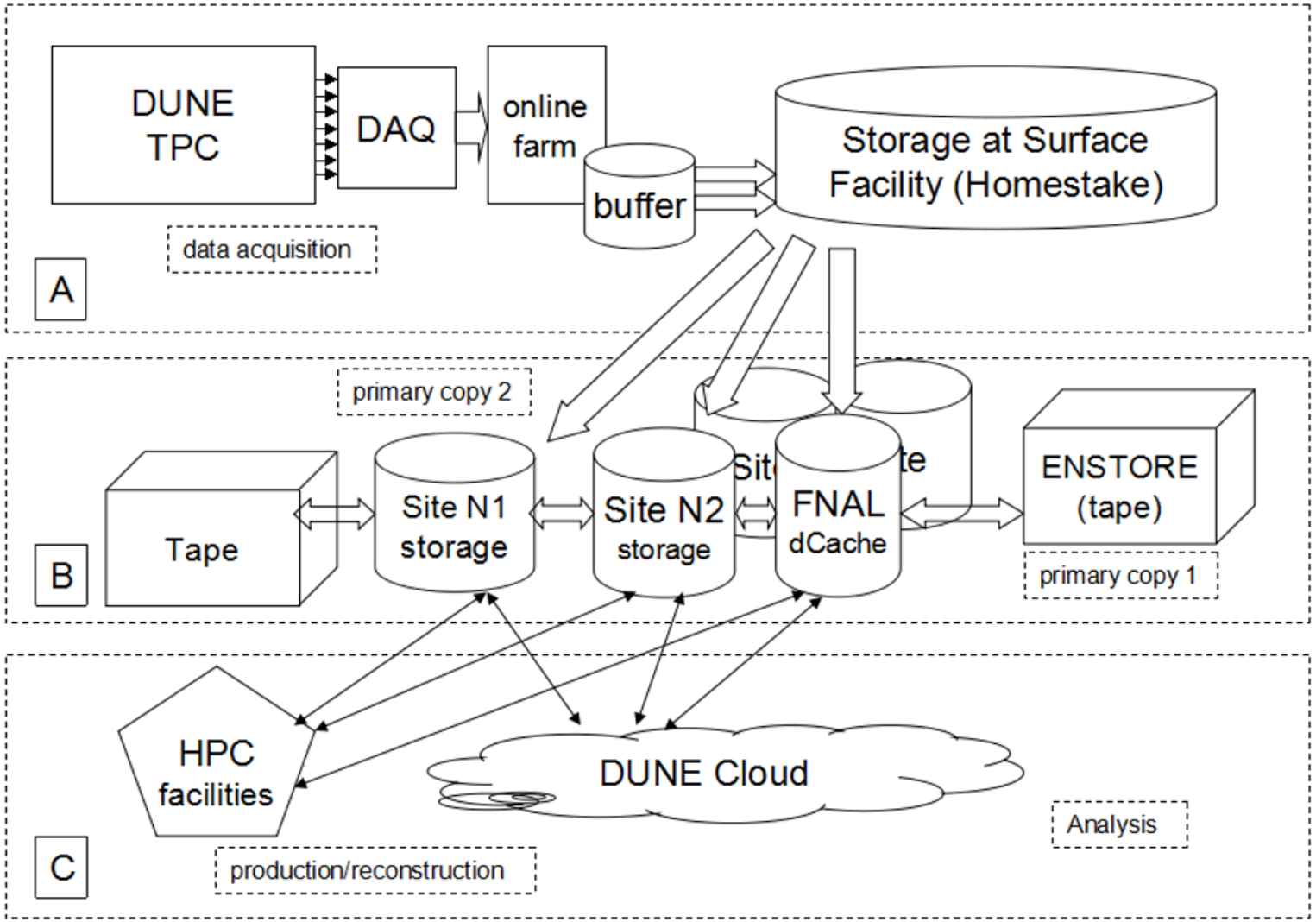
protoDUNE

Design and development of the data handling and processing systems for protoDUNE have started just recently. There are numerous challenges to be met when capturing data at the high rate required by the experiment (~1GB/s) and establishing robust transmission and distribution for production and analysis.

Integration with the CERN infrastructure, transmission of data to FNAL, other US and European sites will require substantial effort and offer plenty of opportunities for participation.

It is clear even now that production and analysis of protoDUNE data will create considerable storage and computing power requirements. Contributions from member institutions in this area may well prove critical to the success of protoDUNE.

DUNE Far Detector data flow



(1, or likely 2 replicas of raw data, at a site capable of storage and bandwidth)

DUNE FD data flow

- Data merging: DAQ stream for each of 4 modules
  - → merged likely in on-surface facilities of FD site
  - → separate stream for SNB, near-time analysis at FD site, fast distribution to Collaboration
- Far Site / Near Site interfaced to FNAL storage infrastructure
  - → dCache (access from outside of FNAL), SAM (Sequential data Access via Metadata), BlueArc (mass, disk-based storage), Enstrore (tape)
- First-pass production at FNAL
  - → provides refined calbration data (e.g. based on muon samples)
- Reprocessing: FNAL and other sites
  - → use refined calibrations, various software solutions

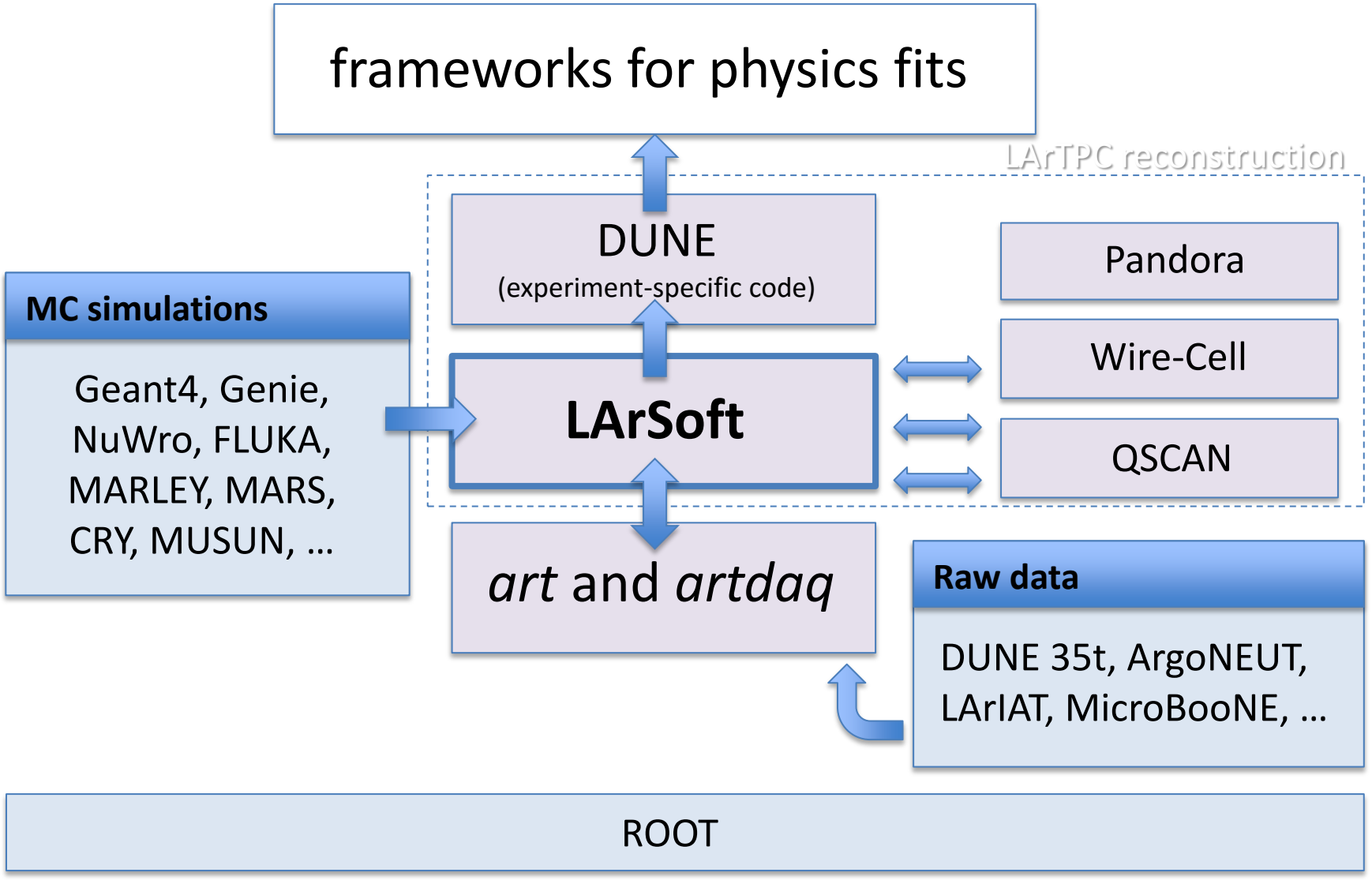
Participation to protoDUNE computing is a great opportunity to gain expertise for future DUNE data handling processes.

DUNE ND data flow

- Precision measurement of momenta and energy
- ND Technology choice no yet ready
  - → if LAr then solutions (and all challenges) similar to FD

  But for the moment:
  - → file formats, reconstruction algorithms not yet known
  - → potentially complex callibration scheme
  - → 2-pass processing similar to FD assumed

Offline Software: ONE framework for any LArTPC
→ shared algorithms, big community – rather unusual!



Simulations

various generators available, or *can be interfaced*

beam sim.

SNB sim.

atmospheric ν

ν flux

neutrino physics — GENIE

cosmogenics — CRY, …

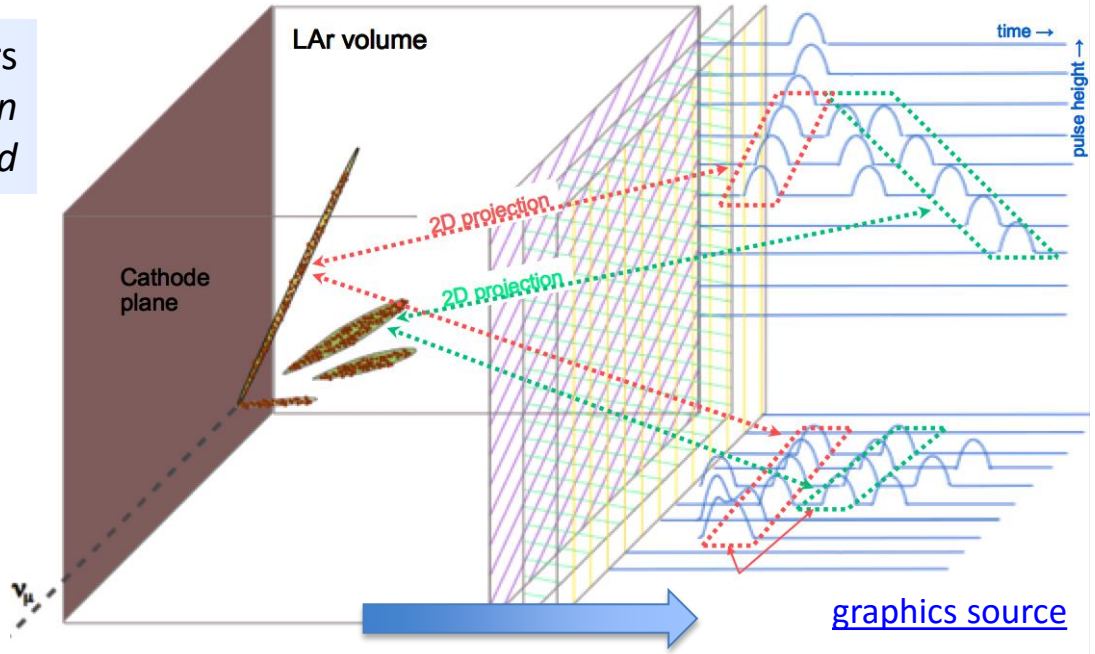
particle list, kinematics

general det.sim. — now Geant4, *FLUKA can enter here*

energy depositions, light

specific det.sim.

raw data like from real detector



graphics source

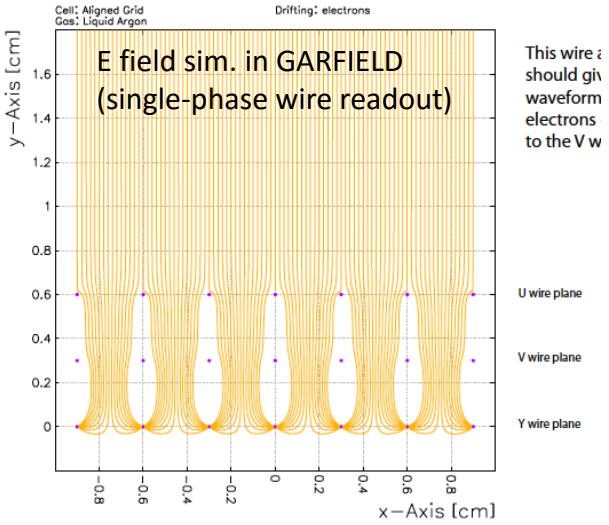
→ use test-beam data to improve models

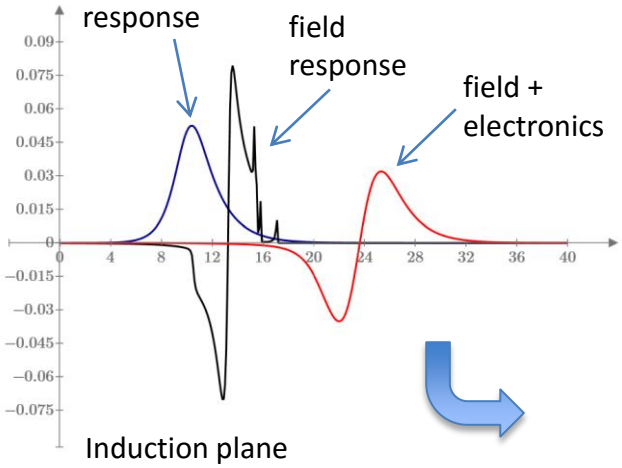
- intelectual effort on models
- …and on tools for extracting knowledge from data

→ place for many improvements, MC is still to simple

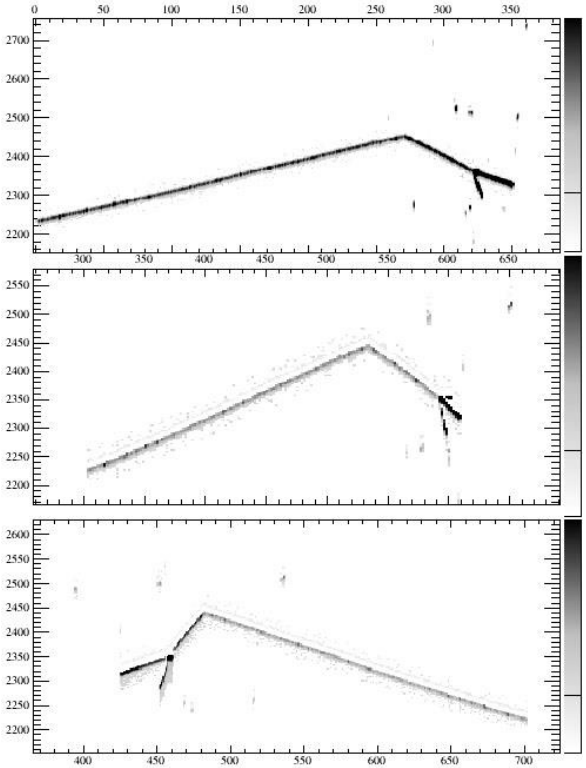
- E field distortions: space charge, det. construction
- signal in induction and collection planes
- electronics noises

Simulations: still a lot of work to include all LArTPC effects

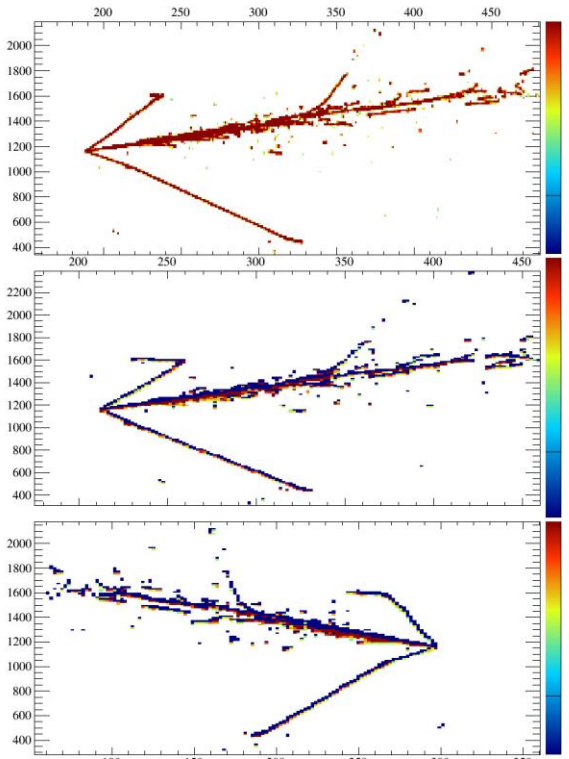


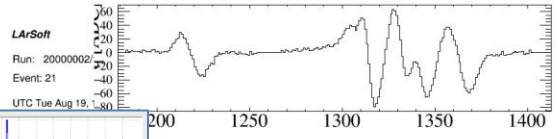


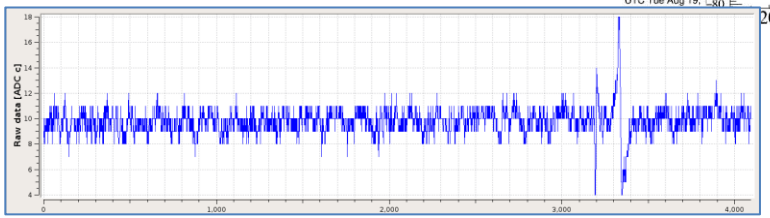
proton in **protoDUNE**



ν_e in DUNE **FD**







LArTPC reconstruction

raw data, sim. or real detector

↓

deconvolution

ionization = f(time)

↓

hit finding / 3D imaging

2D hits / 3D cells

↓

pattern recognition

trajectory fit

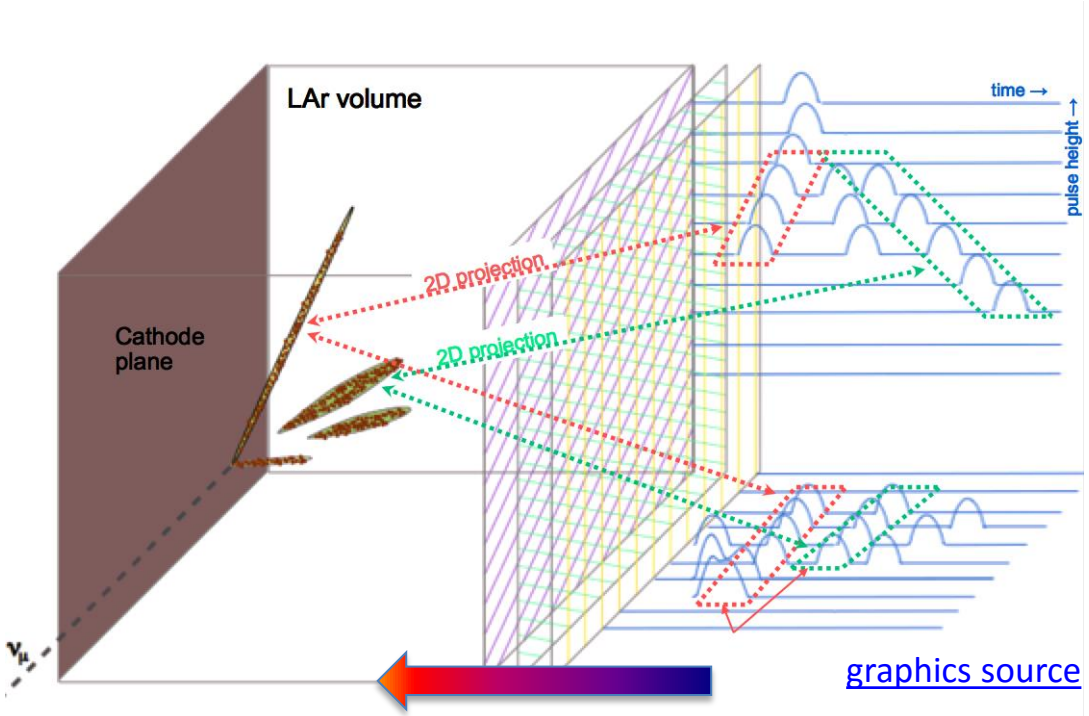
vertex finding

EM showers

hierarchy recognition

↓

full event, ready for PID/kinematics



graphics source

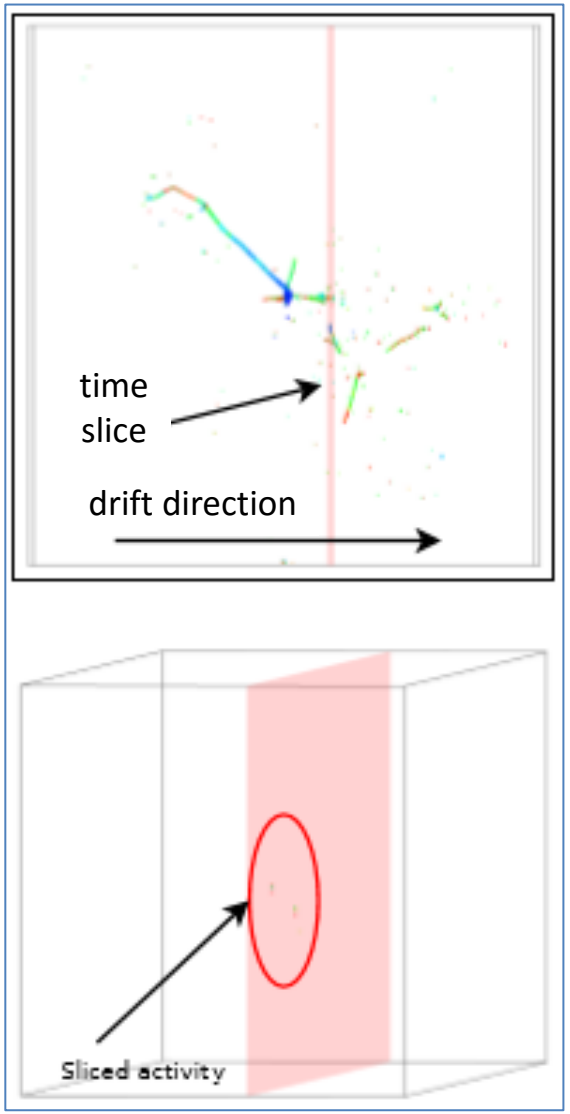
→ various approaches mix the order in this sequence

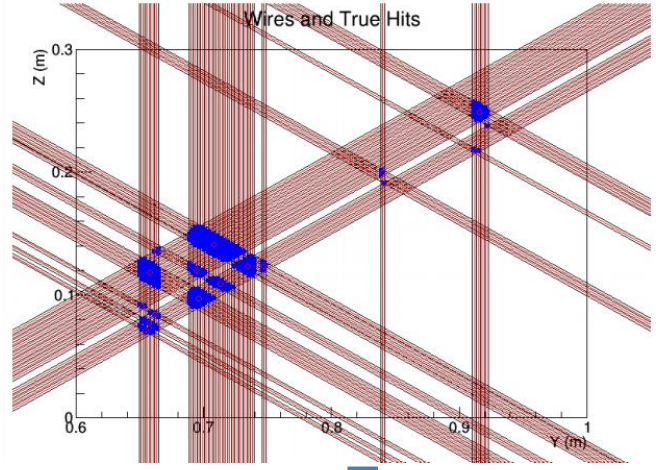
→ modular design in LArSoft is a good ground:

- e.g. blurred clustering imported from T2K
- 3D approach started in ICARUS developed into full-event-topology fit (Projection Matching) and uses input from many now available 2D pattern recognition
- Pandora takes hits and returns reach pattern recognition output
- Wire-cell output also ready to use inside LArSoft

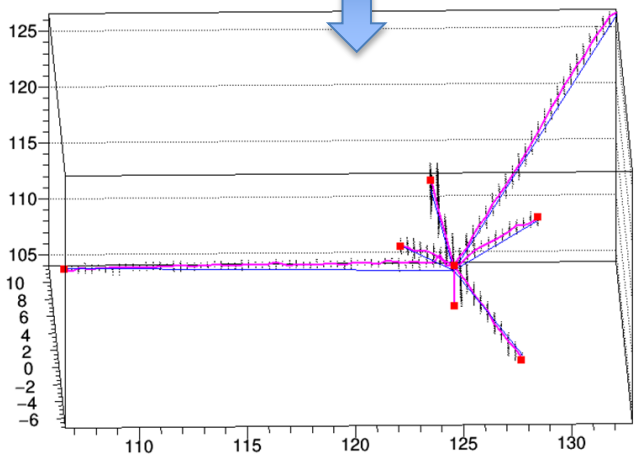
LArTPC reconstruction: Wire-Cell

(X.Qian, B.Viren, C.Zhang)





time slice



3D cells, tagged tracks

- advanced 3D imaging step
- following pattern recogniton, fits → waiting for the work

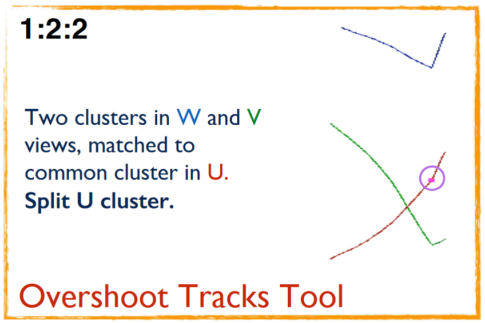
LArTPC pattern recognition: Pandora (A.S.T.Blake, J.S.Marshall, M.A.Thomson)

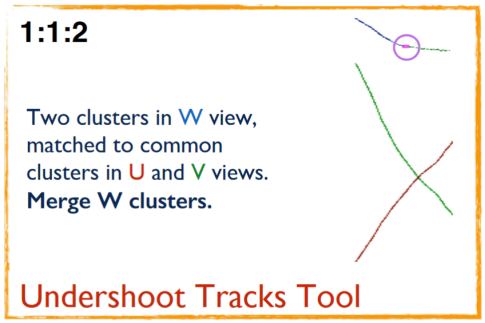
→ does pattern recognition step (a lot in 2D, but buids also 3D results)

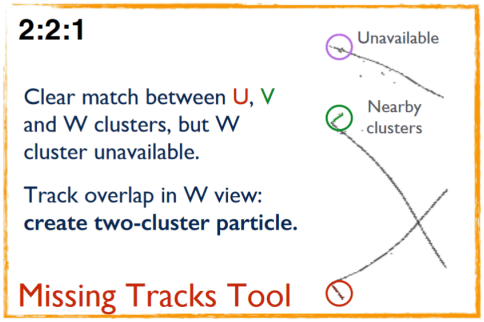
→ designed for combining chains of multiple algorithms (now 150 in the box)

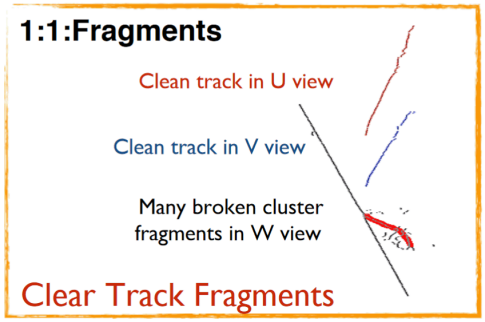
→ can output „particle hierarchy”: resolved full-event topology

→ well interfaced to LArSoft

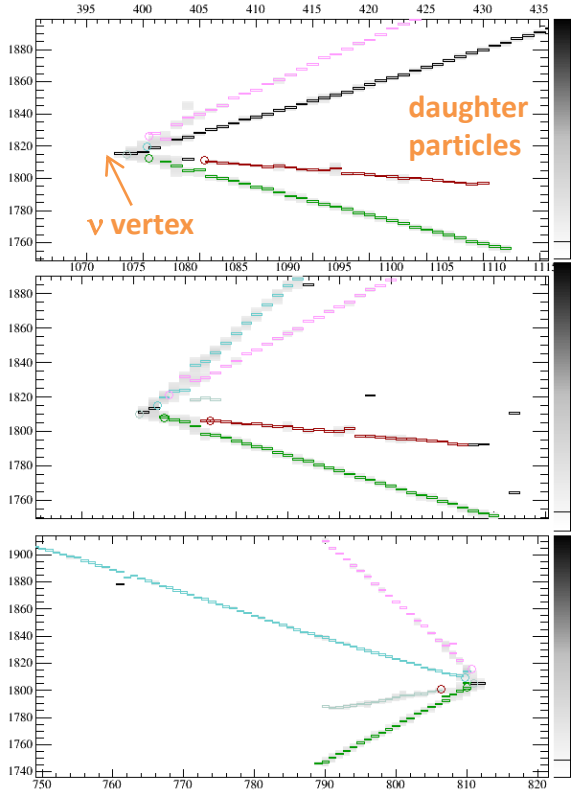








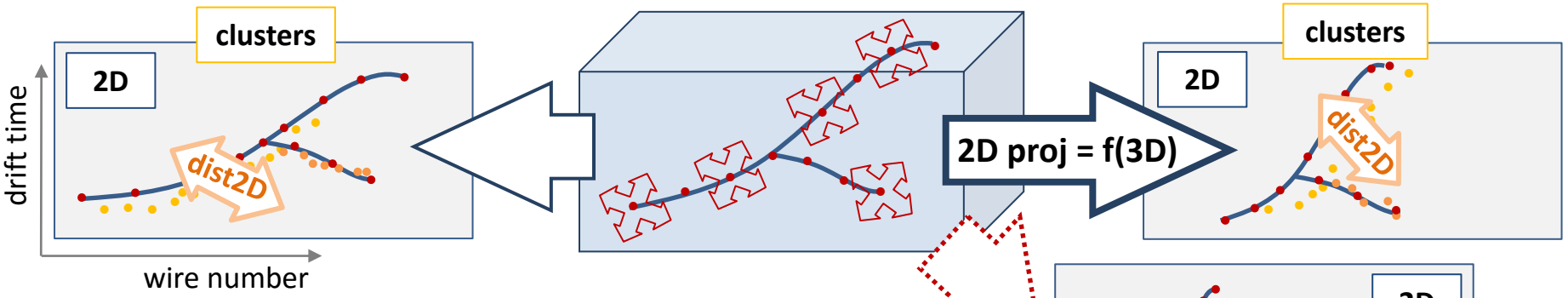
neutrino event back in LArSoft



https://indico.fnal.gov/getFile.py/access?contribId=5&sessionId=0&resId=0&materialId=slides&confId=10394

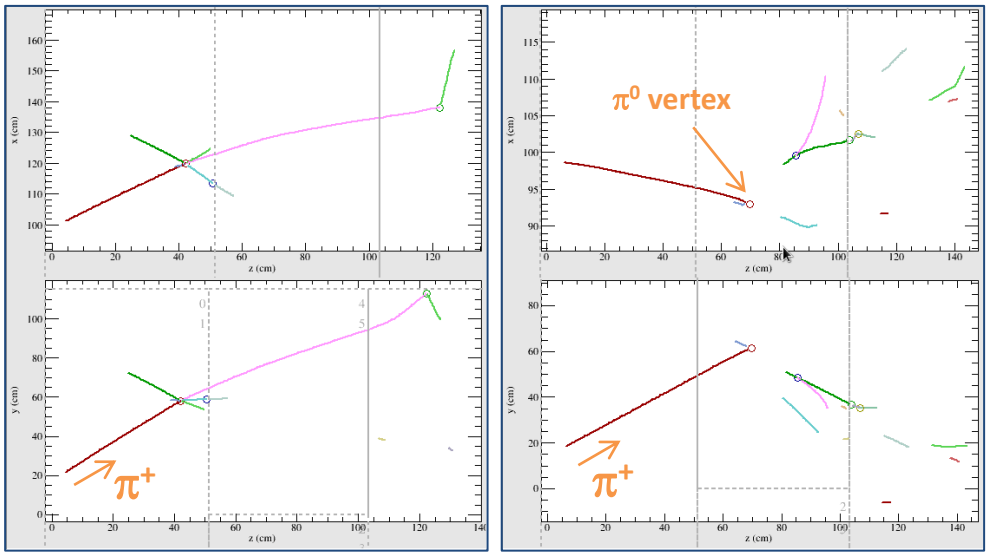
LArTPC 3D topology fit: Projection Matching

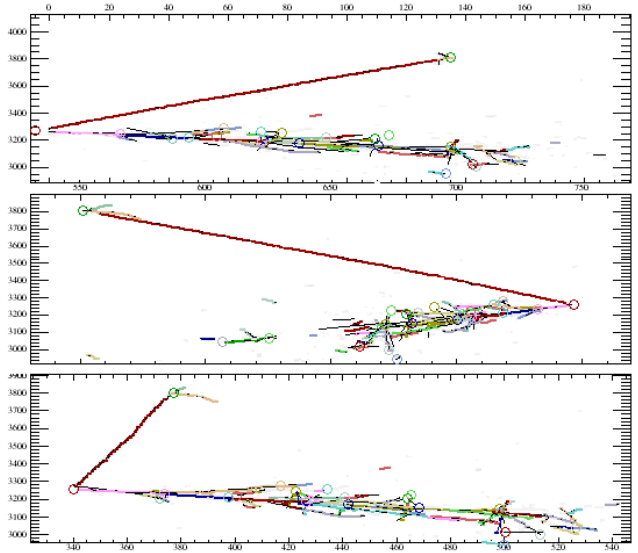
(D.Stefan, R.Sulej)



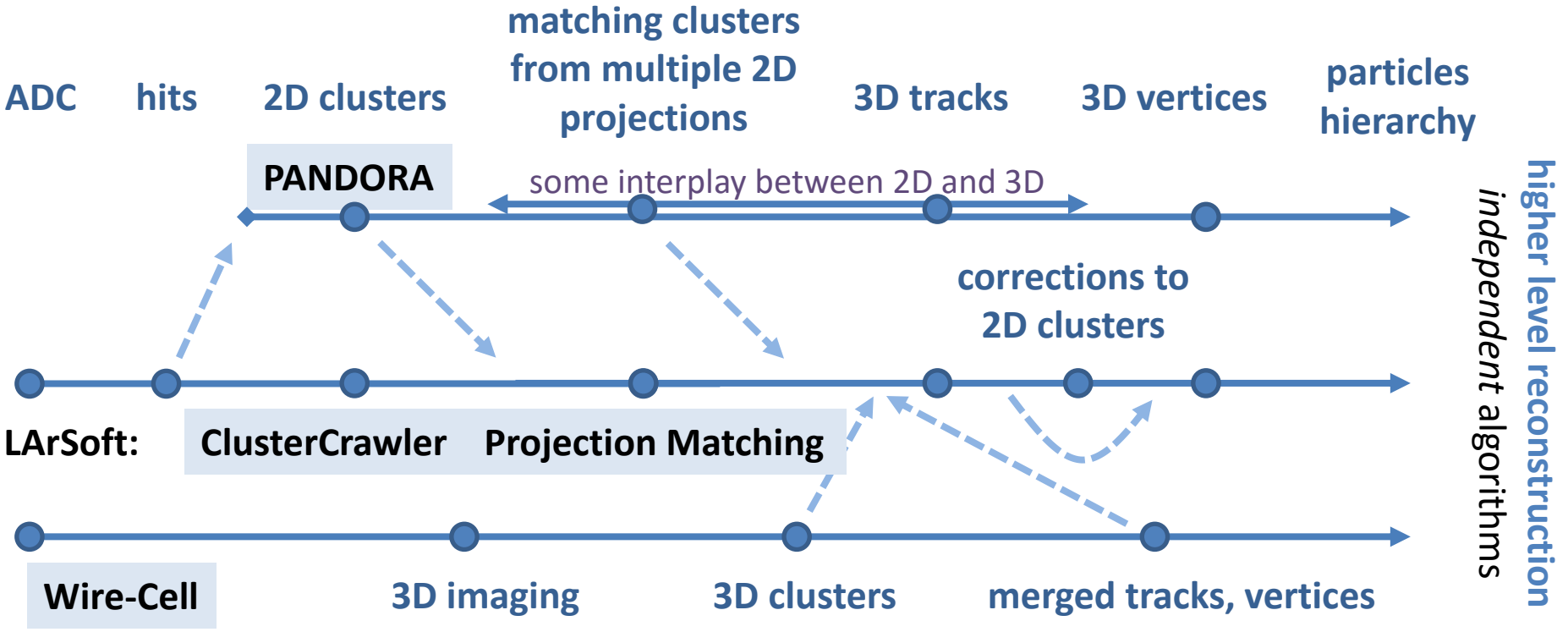
- full 3D topology fitted to multiple 2D projections
- pattern recognition input can be very basic: fit properties used to build event from 2D fragments

3rd plane 2D: robust validation





Reconstruction chain(s)



Each approach is strong at a slightly different stage

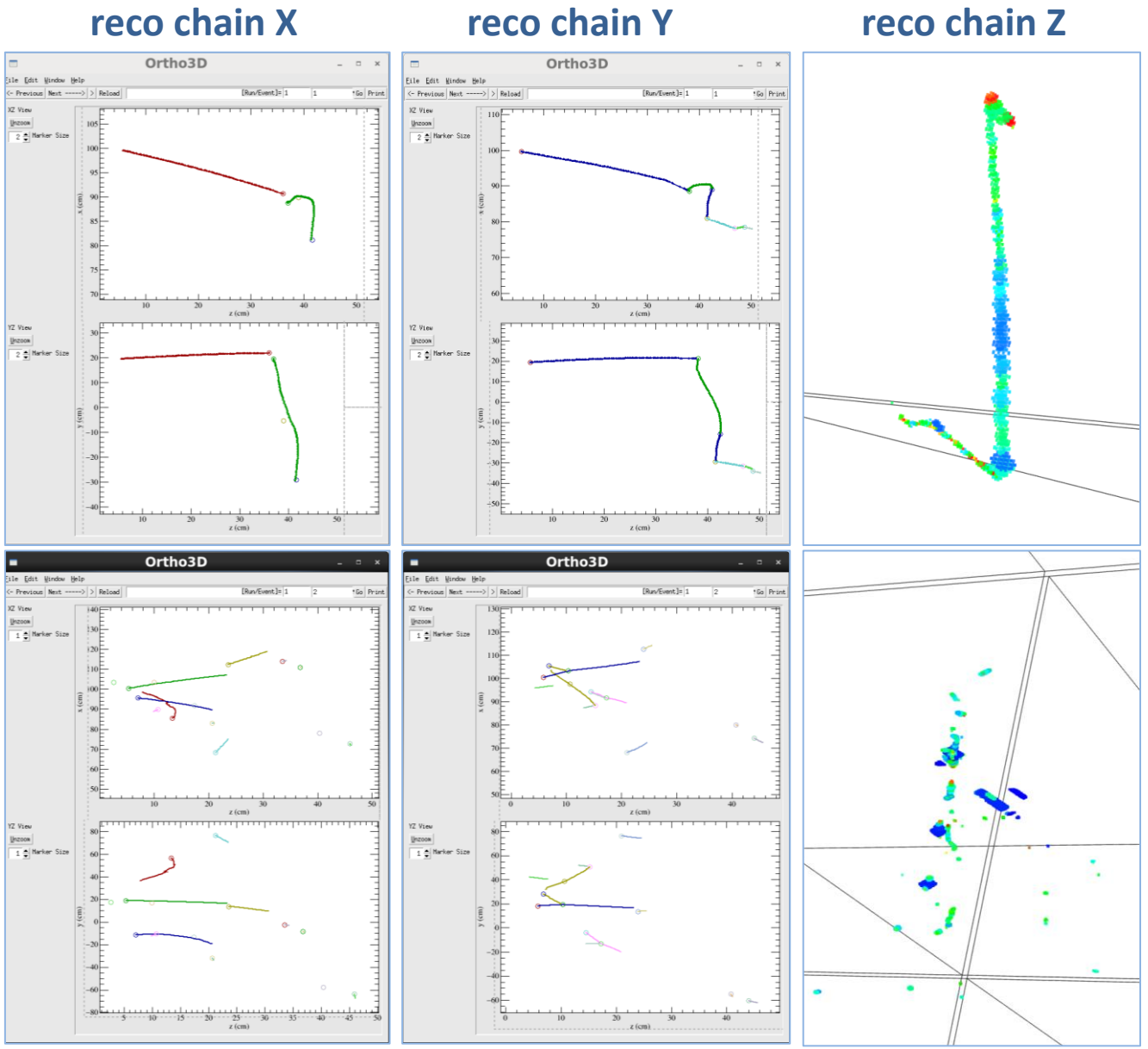
- many possible paths and variants open for selection of the optimal reconstruction
- e.g. try clusters from Pandora instead of ClusterCrawler to see if e.g. PMA can give better track trajectories, dE/dx, and direction fit for cascades with such input
- within shown paths pattern recognition and trajectory & vertex fittig often mixed
- higer level reconstruction / analysis is completely independent issue from these paths

Reconstruction today

decaying kaon
(thought to be easy)

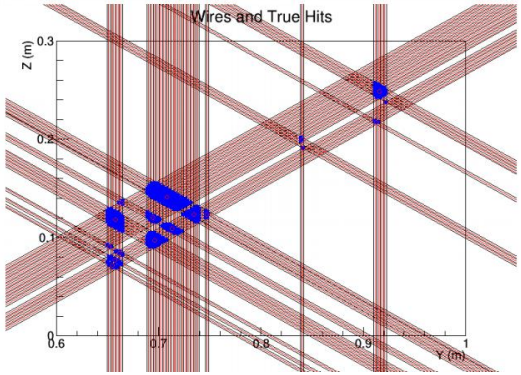
can be OK:

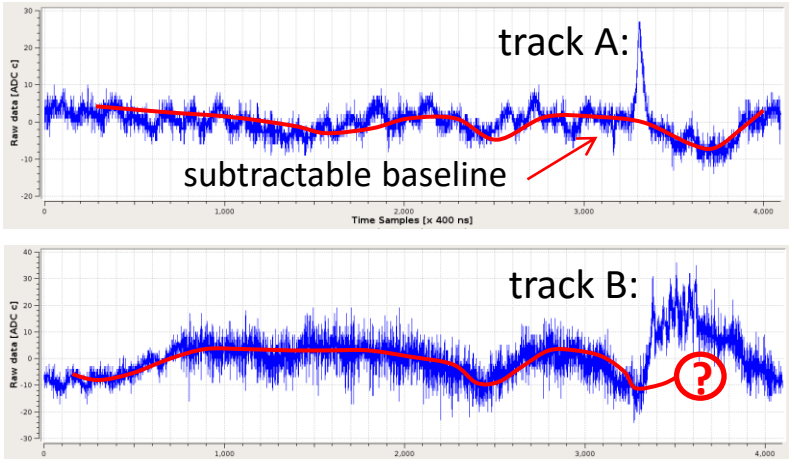
…or easyli can become messy:

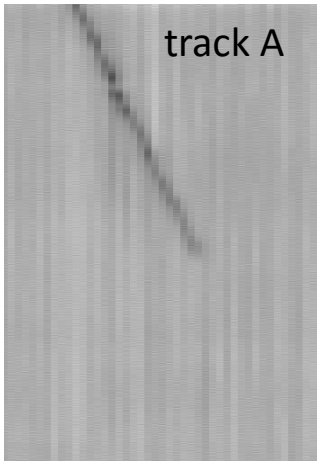


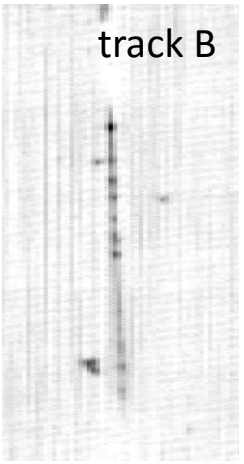
Reconstruction: non-uniformities in LArTPC

- 3D position:
  - → distance-to-wires in drift direction
  - → intersection of wires associated in complemetary views
- what if track (or EM shower) direction is parallel to wire planes:
  - → ambiguity in wire associations if based only on drift time of hits
  - → context of such objects is used, amplitude-based correlations can help: this is a place for new ideas
- what if track / shower direction is drift-parallel:
  - → deconvolution and denoising technique is a key
  - → note: *any* EM shower can have signals long in drift dir.



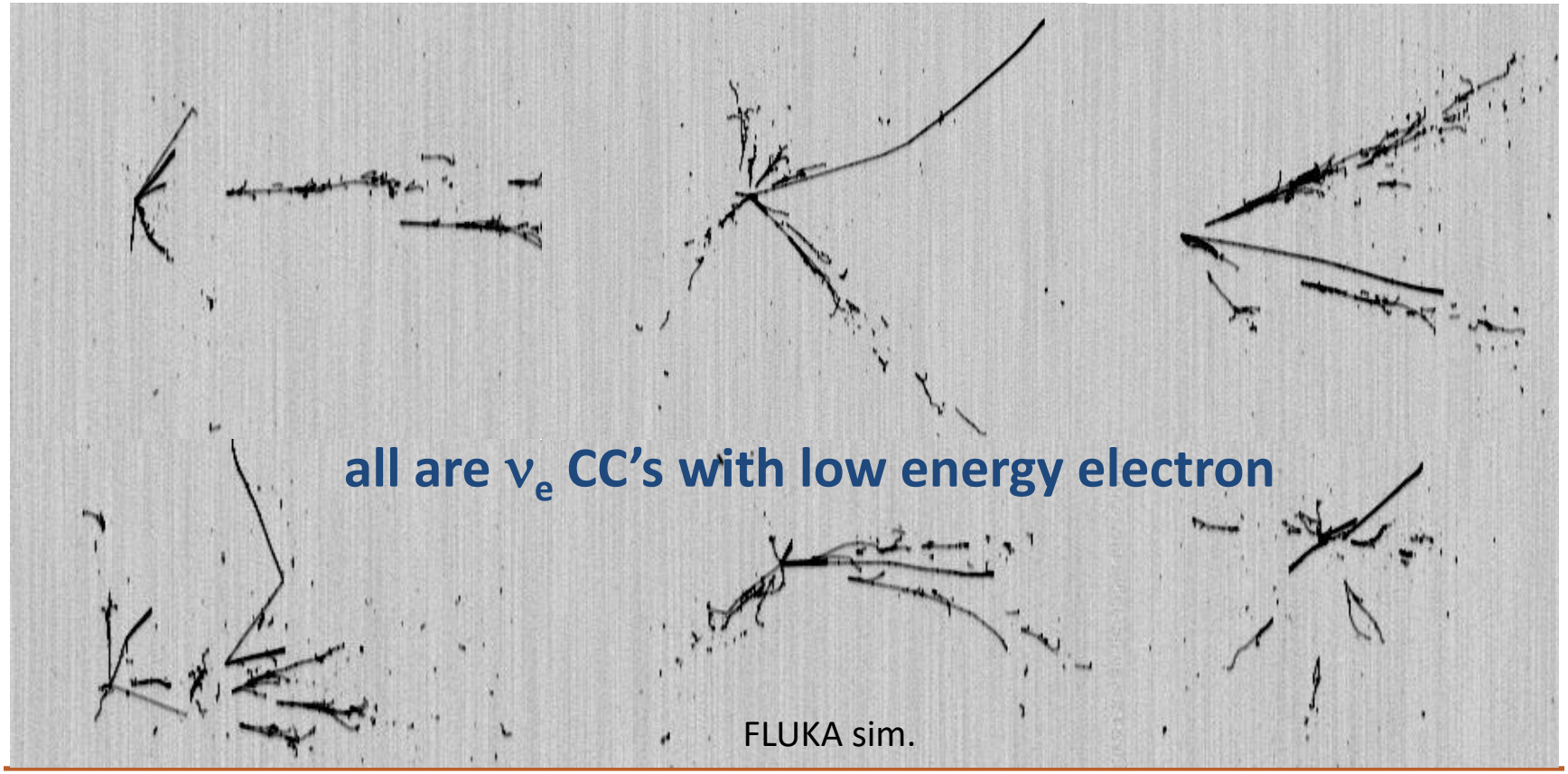




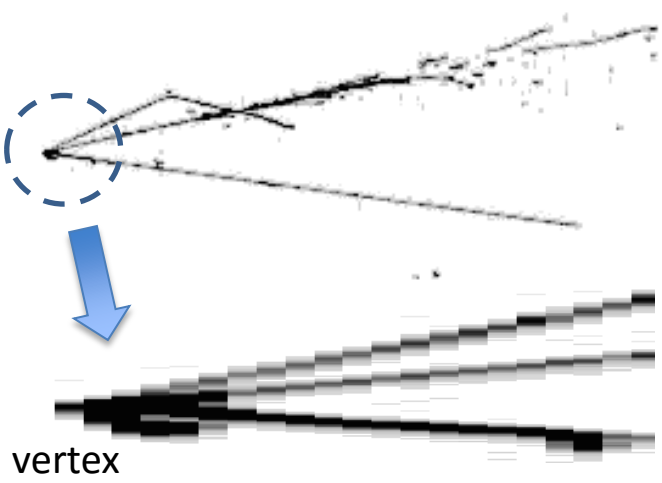


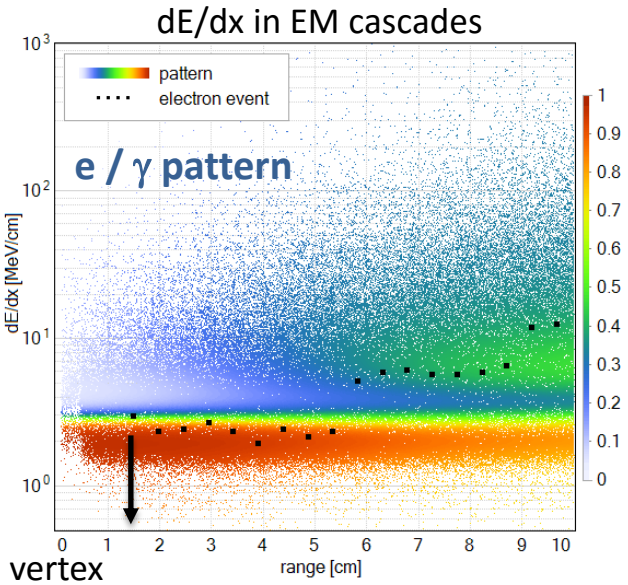
Reconstruction: pattern recognition continued…

- technology allows for much more complex approaches
- machine vision for understanding the LAr „images”
  - at very low level of reconstruction can add context for present algorithms
  - from the other end: simply identify ν or reconstruct topology with „black box”
  - use your imagination how to put new techniques to work for LAr



High level reconstruction – the last step beffore physics





- identify electron neutrino candidates
- measure the neutrino energy

Both goals rely on the complete reconstruction.

Each stage of the reconstruction already now is quite efficient.

But the final selection efficiency is far from DUNE requirements.

- not all capabilities of algorithms were explored
- energy reconstruction is waiting for efforts
- protoDUNE's are perfect testbed for algorithms development and validation in real data conditions

Opportunities

- computing infrastructure: integration with CERN, data distribution to the processing sites in US and Europe

- particle simulation models, challenging ways of using real data for model studies
- realistic simulation of detector signals

- signal processing and „hit" reconstruction
- pattern recognition, trajectory fits - based on Wire-Cell technology
- again and again: fresh ideas for the pattern recognition
- put double-phase in common reconstruction efforts
- visualisation of multi-TPC data is not trivial

- tools/procedures for test-beam measurements (reconstruction-based, calibration-oriented)
- neutrino energy reconstruction methods – really unexplored field

- Neutrino Platform team can provide training and collaboration in most of the above!

- sim/reco group model like we have seen in FNAL: this is a real lifetime opportunity