

A kaggle Charged Particle Tracking with Machine Learning Challenge

iML Forum
Feb 28, 2018

Jean-Roch Vlimant (CalTech), David Rousseau (LAL-Orsay)
with

Markus Elsing, Vincenzo Innocente, Andreas Salzburger (CERN),
Cécile Germain, Balazs Kegl, Yetkin Yilnaz (LAL/LRI-Orsay) , Isabelle
Guyon (Chalearn/LRI-Orsay) Paolo Calafiura, Steve Farrell (LBNL),
Michael Kagan (SLAC), Davide Costanzo (UCL), Tobias Golling,
Moritz Kiehn, Sabrina Amrouche (U Geneva), Amir Farbin (UTA),
Mikhail Hushchyn, Andrey Ustyuzhanin (YandexDSA)



Partners

kaggle™



02/28/18



IML

TrackML Challenge, iML, J.-R. Vlimant

Outline

- The Issue with Charged Particle Tracking
- Pattern Recognition in Data Science
- A Kaggle Tracking with Machine Learning Challenge
- IML Hackathon

Charged Particle Tracking

02/28/18

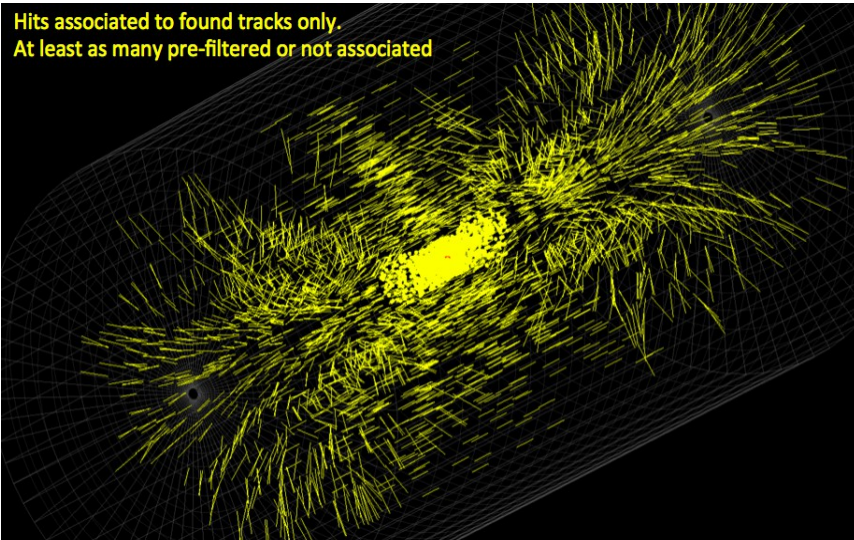


IML

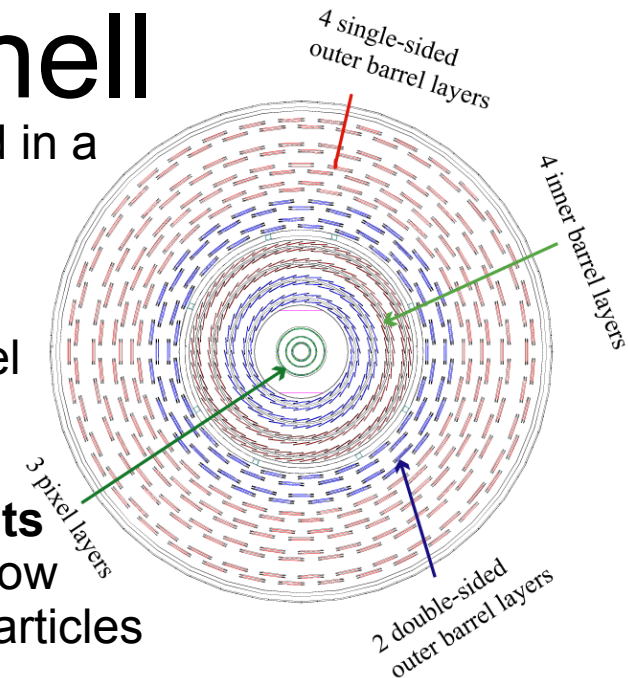
TrackML Challenge, iML, J.-R. Vlimant

Tracking in a Nutshell

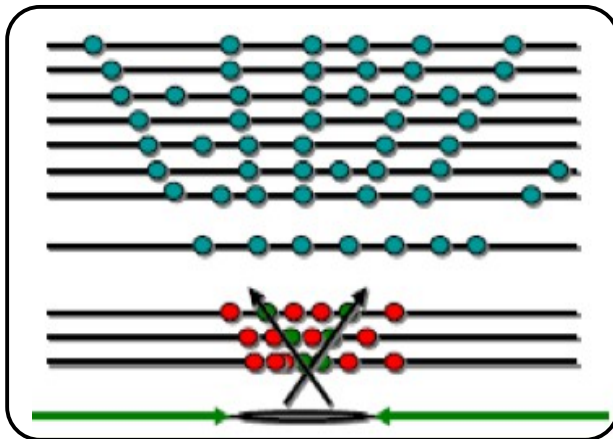
Hits associated to found tracks only.
At least as many pre-filtered or not associated



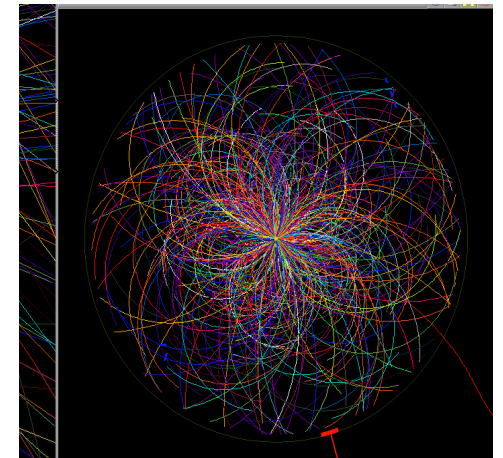
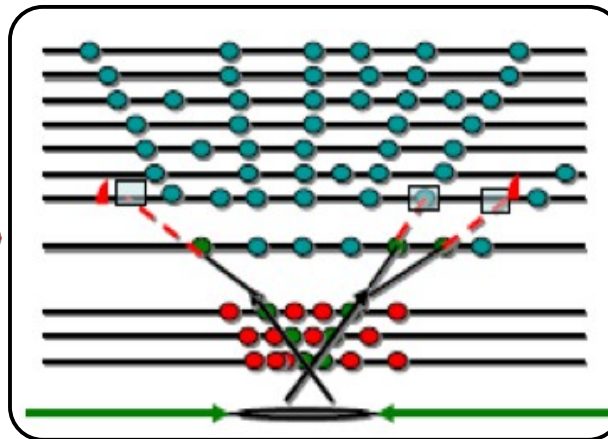
- Particle trajectory bended in a solenoid magnetic field
- Curvature is a proxy to momentum
- Particle ionize silicon pixel and strip throughout several concentric layers
- **Thousands of sparse hits**
- Lots of hit pollution from low momentum, secondary particles



Seeding

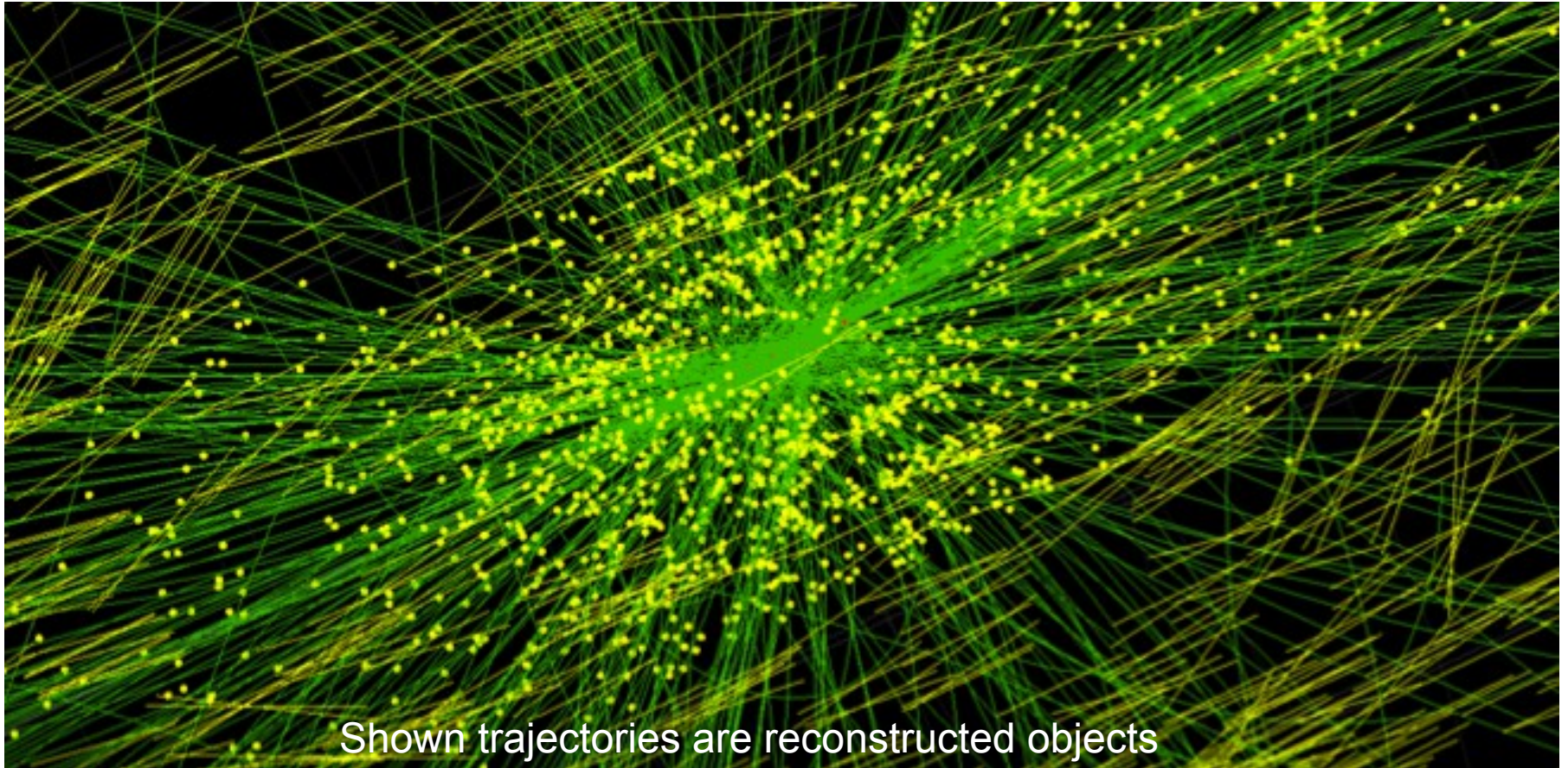


Kalman Filter



- **Explosion in hit combinatorics** in both seeding and stepping pattern recognition
- **Highly computing consuming task** in extracting physics content from LHC data

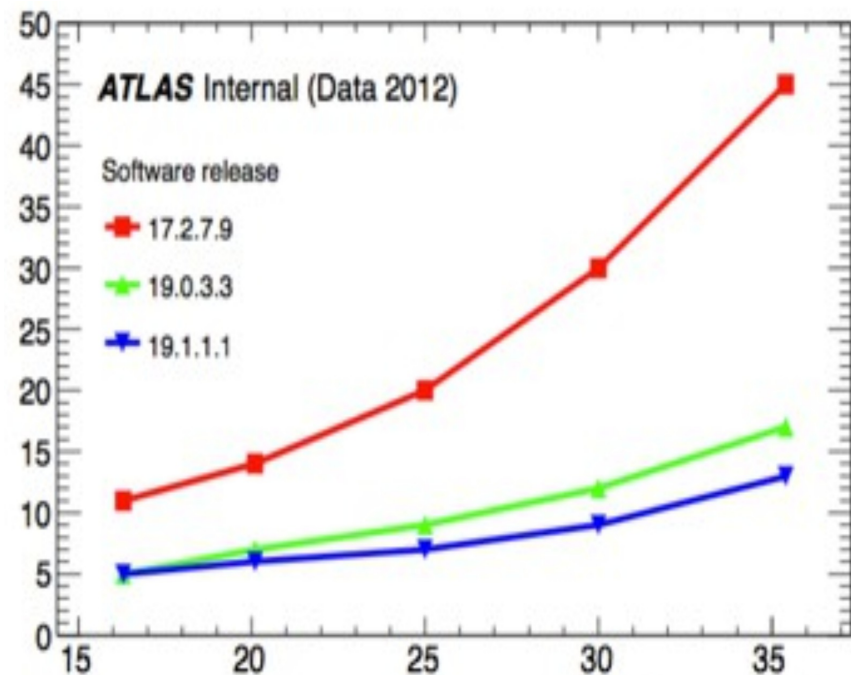
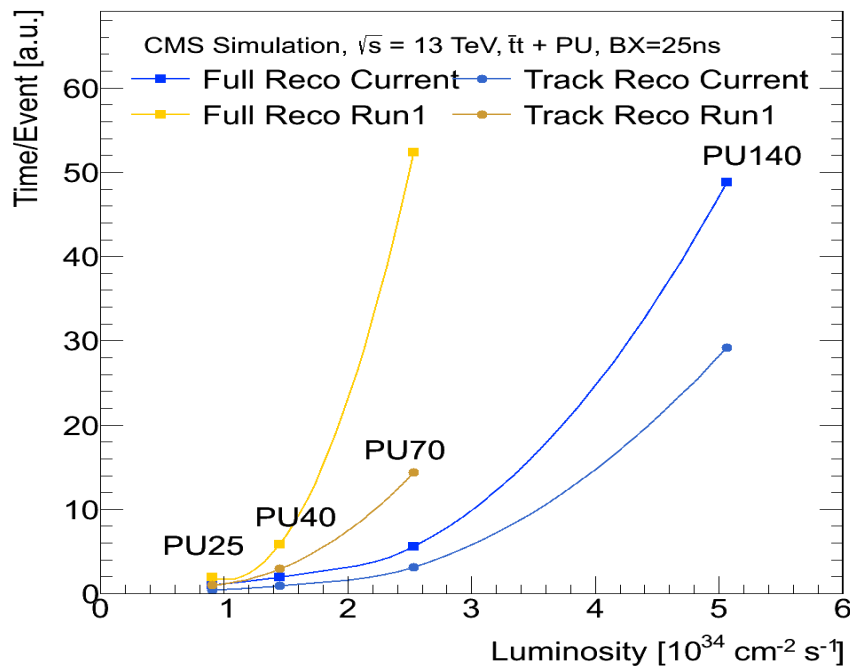
Complexity and Ambiguity



The future is with **x10 more hits**

Cost of Tracking

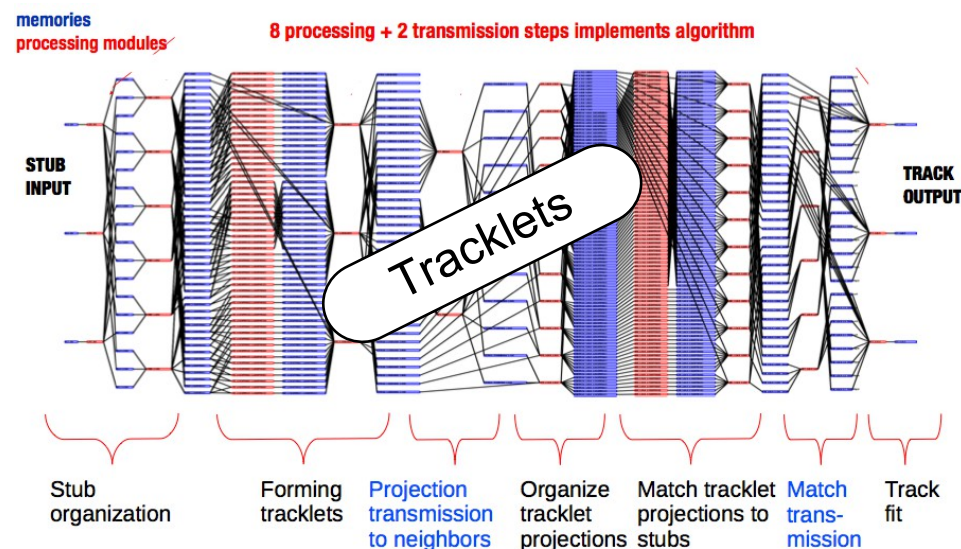
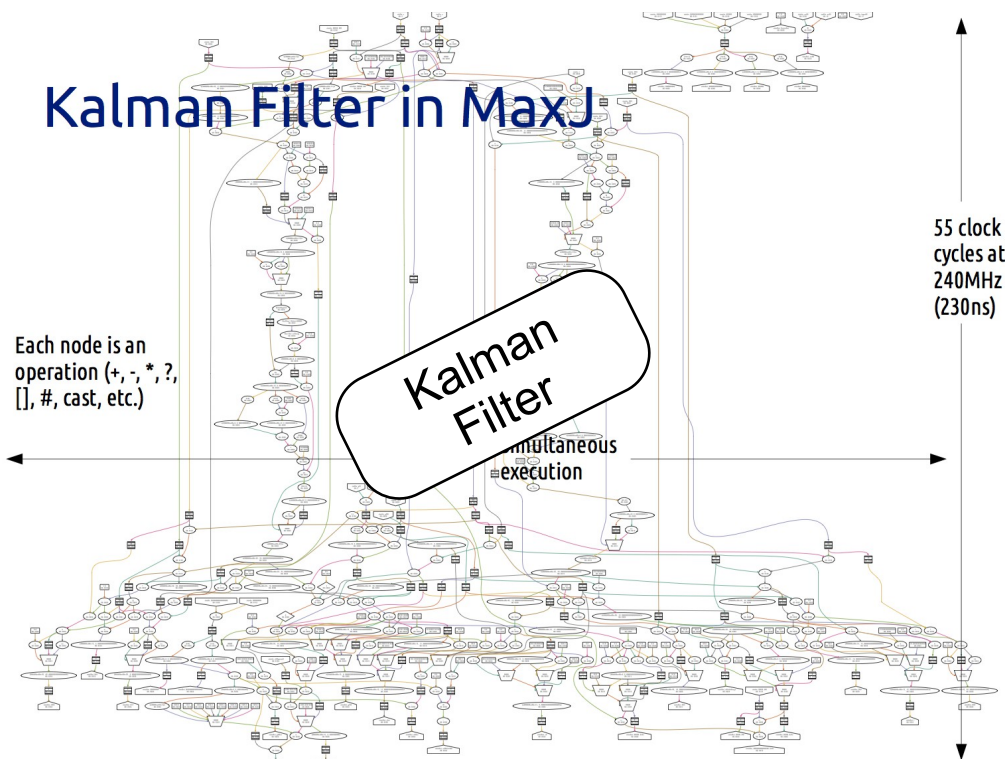
- Charged particle track reconstruction is one of the **most CPU consuming task** in event reconstruction
- Future computing budget flat at best
- Optimizations (to fit in computational budgets) **mostly saturated** and **long way to go** for HL-LHC
- Need factor **10-100 speed-up**



Fast Hardware Tracking

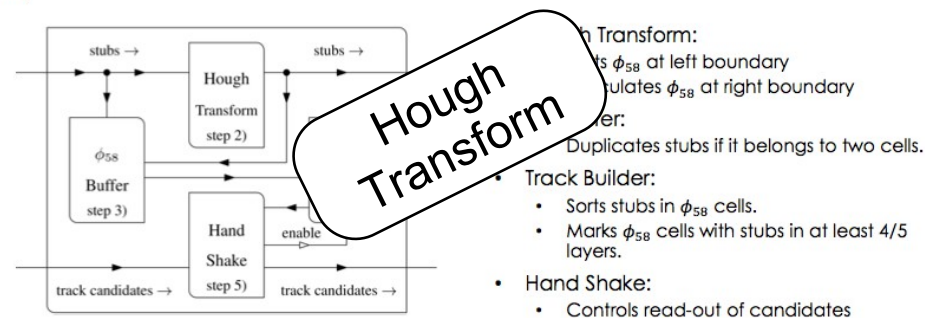
- Track trigger implementation for Trigger upgrades development on-going
- Several approaches investigated
- **Dedicated hardware is the key** to fast computation.
- **Not applicable for offline** processing unless by adopting heterogeneous hardware.

Kalman Filter in MaxJ



Firmware Implementation - Bin

- Each bin represents a q/p_T column in the HT array



See <https://ctdwit2017.lal.in2p3.fr/>

Bottom Lines

Current algorithms for tracking are highly performant physics-wise and scale badly computation-wise

Faster implementations are possible with dedicated hardware

Reach out to machine learning community for new methods and possible solution, applicable on commodity hardware

Pattern Recognition With Deep Learning



Scene Labeling



Farabet et al. ICML 2012, PAMI 2013

Scene Labeling



LeCun Seminar at CERN

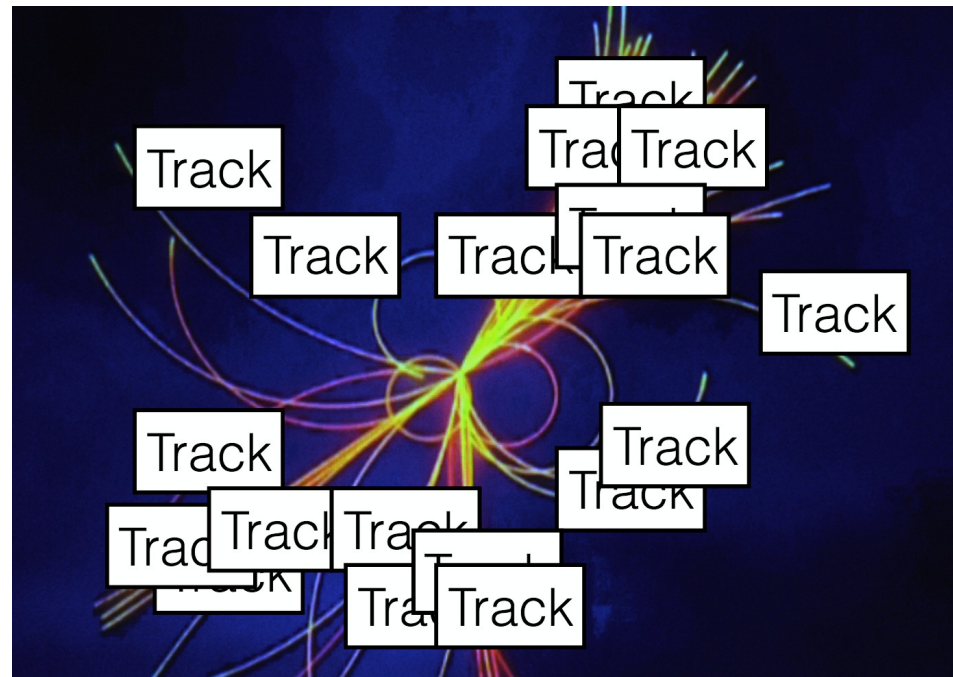


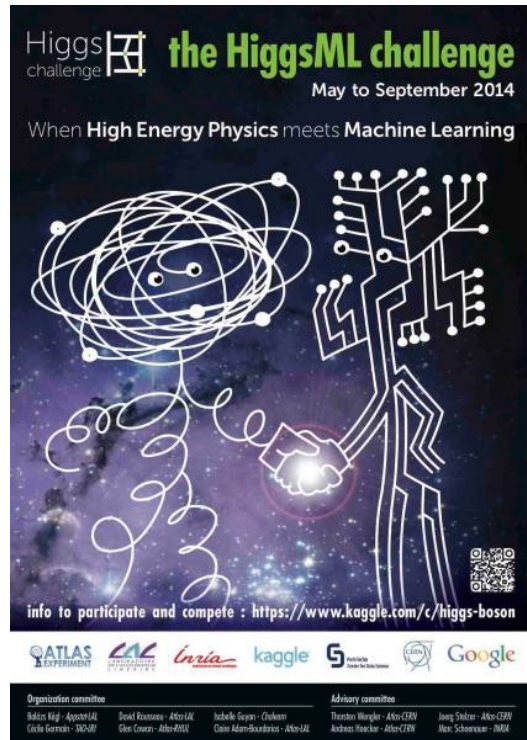
Photo by Pier Marco Tacca/Getty Images

- There exists recent work on applying machine learning and deep learning as we know it to the challenge of particle tracking :
 - Hopefield network : <http://inspirehep.net/record/300646/>
 - CNN in NOVA : <https://arxiv.org/abs/1604.01444>
 - HEP.TrkX : <https://heptrkx.github.io/>
 - TrackML RAMP : <https://tinyurl.com/y84yd5hn>
 - ... no golden solution yet

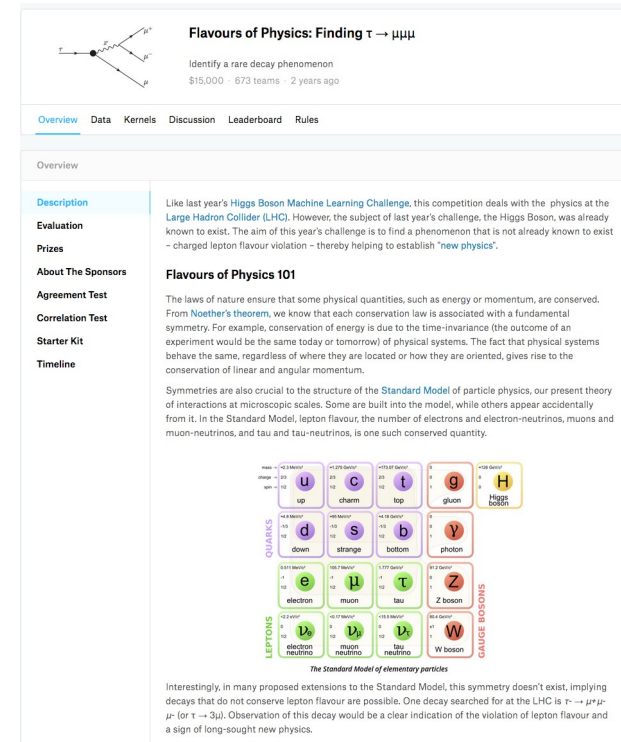
A kaggle Tracking with Machine Learning Challenge



Previous Challenges



- 2000 teams. Largest competition at the time
 - Winners went to DeepMind and OpenAI
- <https://www.kaggle.com/c/higgs-boson>



- 700 teams.
 - Experienced data exploitation
 - Some methods learned and re-applied later
- <https://www.kaggle.com/c/flavours-of-physics>

The organizing team has participated in the organization of both events

02/28/18



IML

TrackML Challenge, iML, J.-R. Vlimant

14

Challenge Datasets

- Accurate simulation engine (ACTS) to produce realistic dataset
 - One file with list 3D points
 - Ground truth : one file with point to particle association
 - Ground truth auxiliary : true particle parameter (origin, direction, curvature)
 - Typical events with ~200 parasitic collisions (~10.000 tracks/event)
- Large training sample 100k events, 10 billion tracks ~100 GByte
- Participants train on the training sample
- Participants are given the test sample
- A private sample is retained for evaluation on the leaderboard

A 2-Phases Challenge

- First Phase : **Accuracy Phase**
 - Contestants will upload **lists of points belonging together**
 - Score : **fraction of points correctly assigned** (the estimation of the particle parameters is not a criteria)
 - Evaluation on test sample with per-mille precision on 100 event
 - Prizes to top-3 on the leaderboard
 - Special prize from a jury scrutinizing algorithms
- Second Phase : **Throughput Phase**
 - Contestants will **upload their program** for pattern recognition
 - Score : **time for inference**, with a penalty on loss of accuracy score
 - Prizes to top-3 on the leaderboard
 - Special prize from a jury scrutinizing algorithms

Schedule

- Challenge Schedules
 - February to May : Run challenge Accuracy phase
 - June to October : Run challenge Throughput phase
- Conference/workshops
 - July 2018 : accepted as an official competition for the IEEE World Congress on Computational Intelligence at Rio de Janeiro
<http://www.ecomp.poli.br/~wcci2018/competitions/>
 - July 2018 : (to be submitted) as a talk at Computing in High Energy Physics computing at Sofia
 - December 2018 : (to be submitted) as a NIPS 2018 competition and workshop
 - Spring 2019 : grand finale workshop at CERN with prize delivery

iML Workshop Hackathon

02/28/18



IML

TrackML Challenge, iML, J.-R. Vlimant

18

Tracking Hackathon

Last day (April 12) of the April workshop

<https://indico.cern.ch/event/668017/>

Walkthrough of the starting kit.

Answer questions on how to proceed.

Register and come work on a solution to the challenge !

Conclusions

- Building up on successful challenges in HEP
- Unique pattern recognition challenge
- Seeking sponsors to cover operation costs
- Seeking sponsors for in-kind computing resource
- Contacts :
 - jvlimant@caltech.edu, rousseau@lal.in2p3.fr,
trackml.contact@gmail.com
- More details, news, etc, ... :
<https://sites.google.com/site/trackmlparticle>

Partners

kaggle™

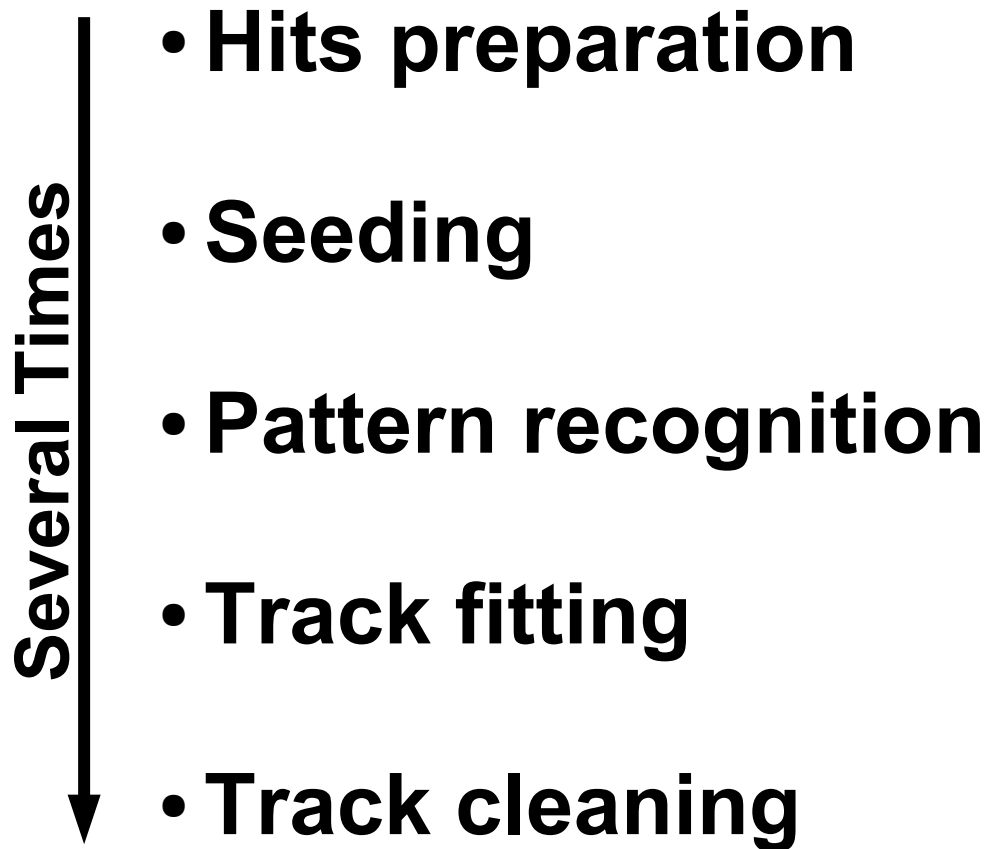


Backup



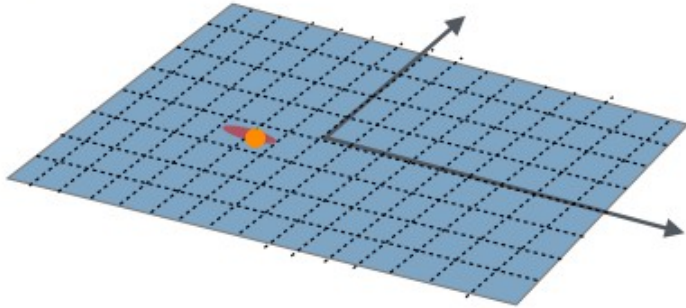
Sponsor Type	Price	Slots	Advantages
Diamond	\$20,000	2 slots	<ul style="list-style-type: none"> • Allowed a talk at NIPS and/or CERN workshops. • Allowed a member in the jury. • Big logo on presentation front page at conferences • Big logo and text on challenge Kaggle website • Big logo and text on the challenge website • Name and logo on all communication media
Platinum	\$10,000	5 slots	<ul style="list-style-type: none"> • Big logo on presentation front page at conferences • Big logo and text on challenge Kaggle website • Big logo and text on the challenge website • Name and logo on all communication media
Gold	\$5,000	10 slots	<ul style="list-style-type: none"> • Logo on presentation at conferences • Big logo on challenge Kaggle website • Logo on the challenge website
Silver	\$2,000	-	<ul style="list-style-type: none"> • Logo on presentation at conferences • Mention on challenge Kaggle website • Logo on the challenge website

Tracking **Not** In a Nutshell

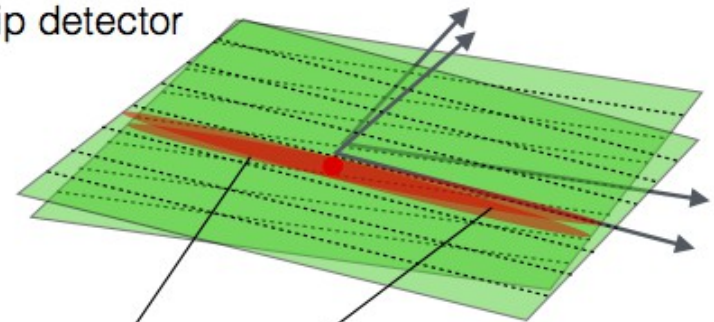


Hit Preparation

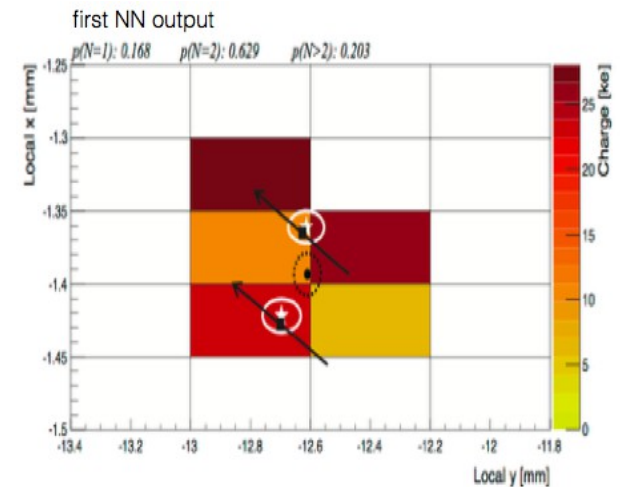
pixel detector



strip detector

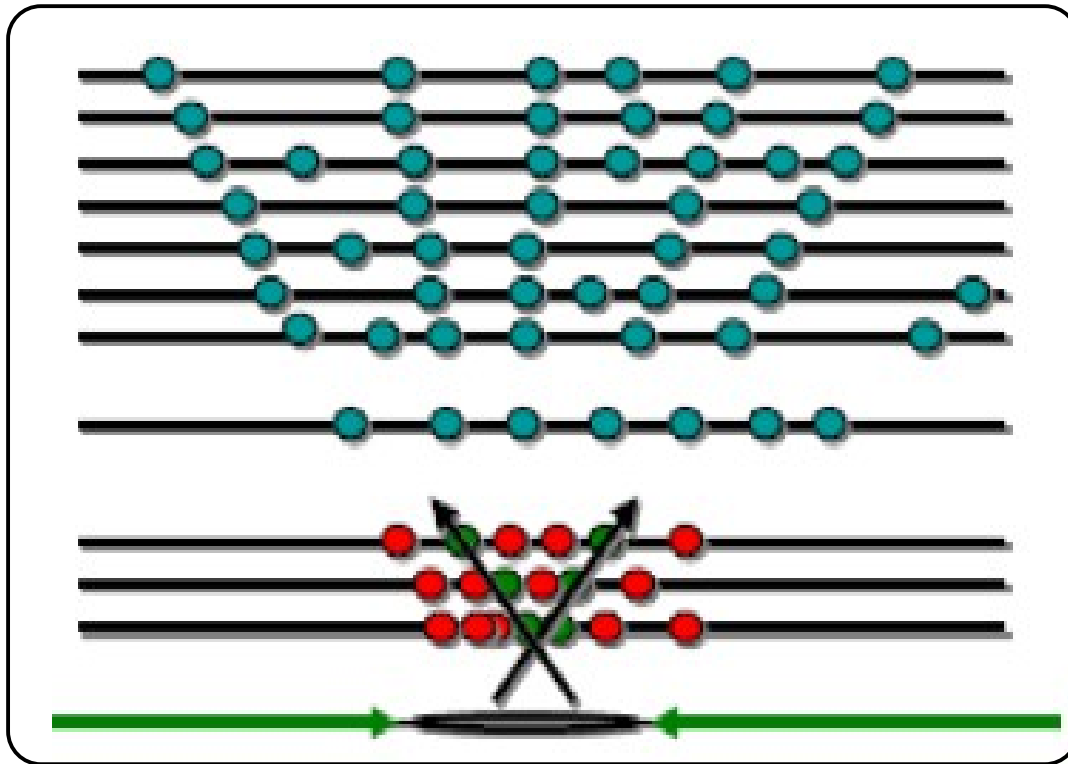


- Calculate the hit position from barycenter of charge deposits
- Use of neural net classifier to split cluster in ATLAS
- Access to trajectory local parameter from cluster shape
- Remove hits from previous tracking iterations
- HL-LHC design include double layers giving more constraints on the local trajectory parameters



Example of cluster split

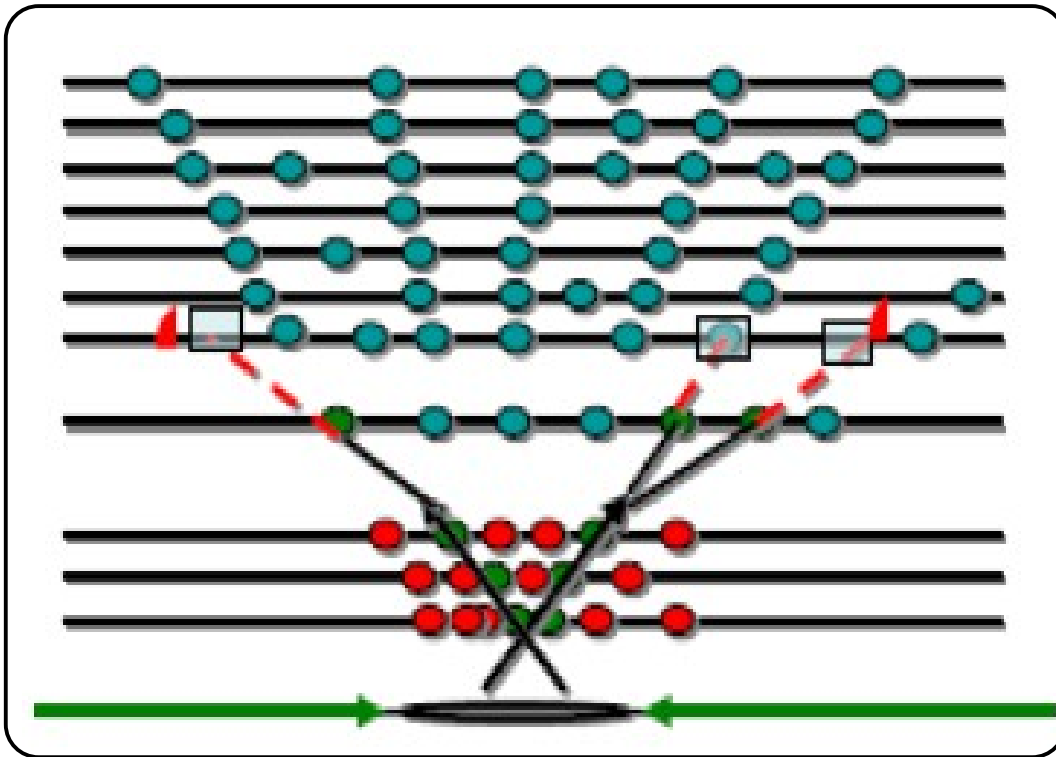
Seeding



- Combinatorics of 2 or 3 hits with tight/loose constraints to the beam spot or vertex
- Seed cleaning/purity plays an important role in reducing the CPU requirements of sub-sequent steps
 - Consider pixel cluster shape and charge to remove incompatible seeds
- Initial track parameters from helix fit

Pattern Recognition

- Use of the Kalman filter formalism with weight matrix
- Identify possible next layers from geometrical considerations
- Combinatorics with compatibles hits, retain N best candidates
- No smoothing procedure
- Resilient to missing modules
- Hits are mostly belonging to one track and one track only
- Hit sharing can happen in dense events, in the innermost part



- Lots of hits from low momentum particles

Kalman Filter

$$K_k = C_{k|k-1} H_k^\top (V_k + H_k C_{k|k-1} H_k^\top)^{-1}$$

$$p_{k|k} = p_{k|k-1} + K_k (m_k - H_k p_{k|k-1})$$

$$C_{k|k-1} = (I - K_k H_k) C_{k|k-1}$$

H_k is the projection matrix

V_k is the hit covariance matrix

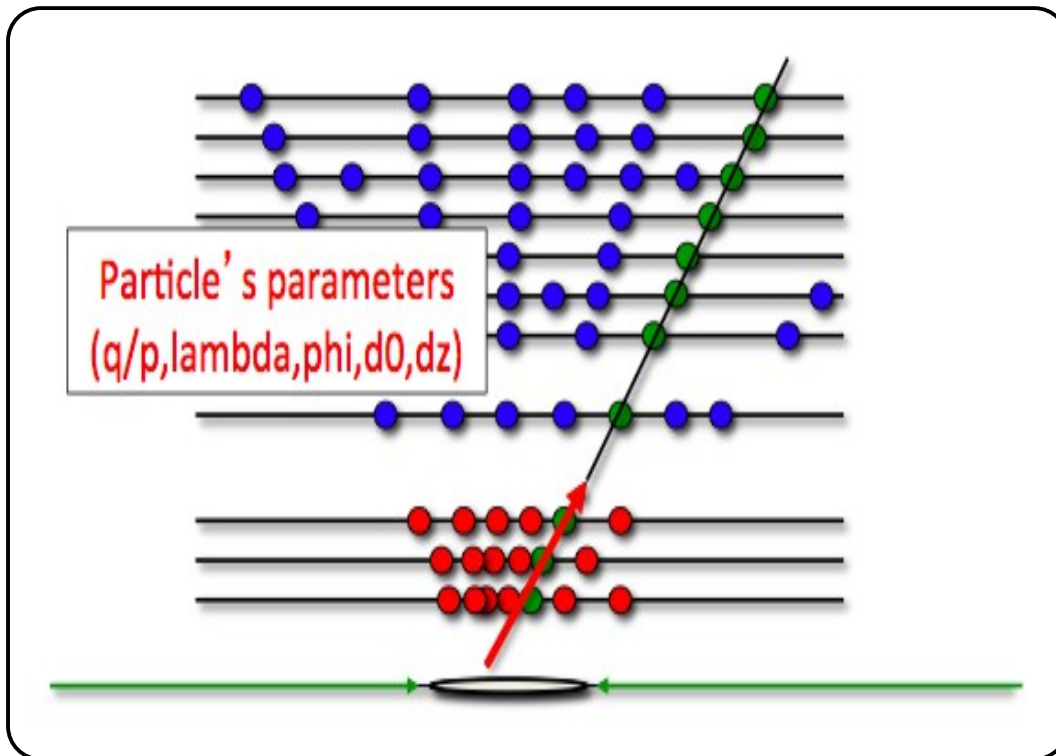
$p_{i|j}$ is the trajectory state at i given j

$C_{i|j}$ is the trajectory state covariance matrix at i given j

- Trajectory state propagation done either
 - ✓ Analytical (helix, fastest)
 - ✓ Stepping helix (fast)
 - ✓ Runge-Kutta (slow)
- Material effect added to trajectory state covariance
- Projection matrix of local helix parameters onto module surface
 - ➔ Trivial expression due to local helix parametrisation
- Hits covariance matrix for pixel and stereo hits properly formed
 - ✗ Issue with strip hits and longitudinal error being non gaussian (square)

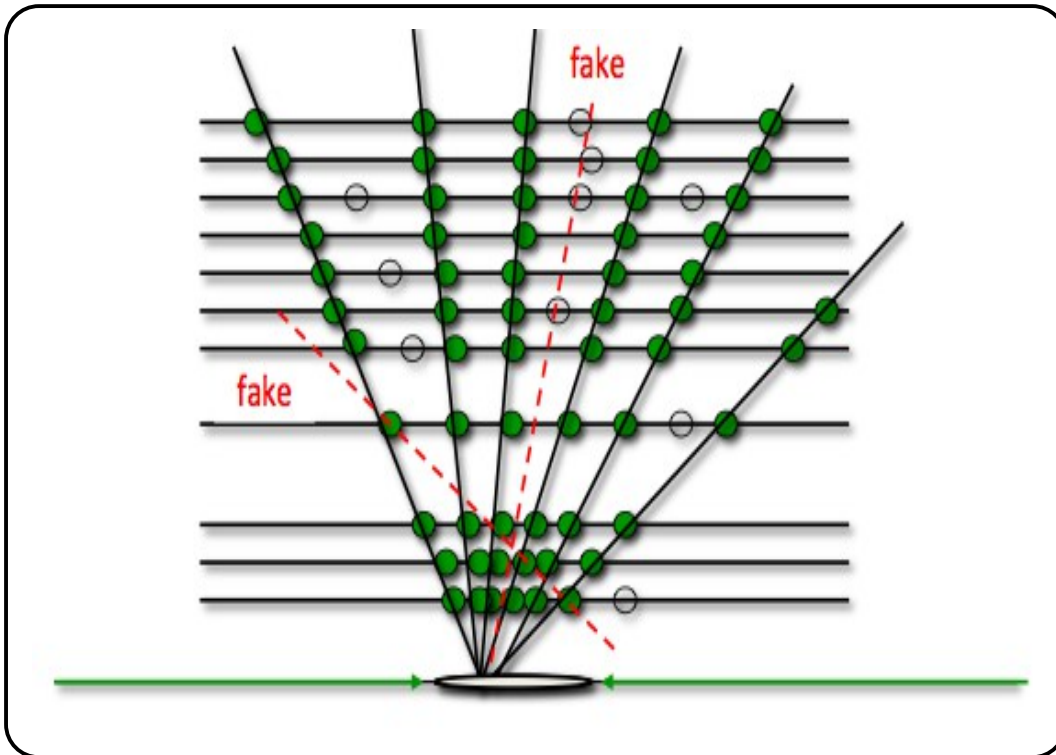
Track Fitting

- Use of the Kalman filter formalism with weight matrix
- Use of smoothing procedure to identify outliers
- Field non uniformity are taken into account
- Detector alignment taken into account



Cleaning, Selection

- Track quality estimated using ranking or classification method
→ Use of MVA
- Hits from high quality tracks are removed for the next iterations where applicable



A Charged Particle Journey

02/28/18



IML

TrackML Challenge, iML, J.-R. Vlimant

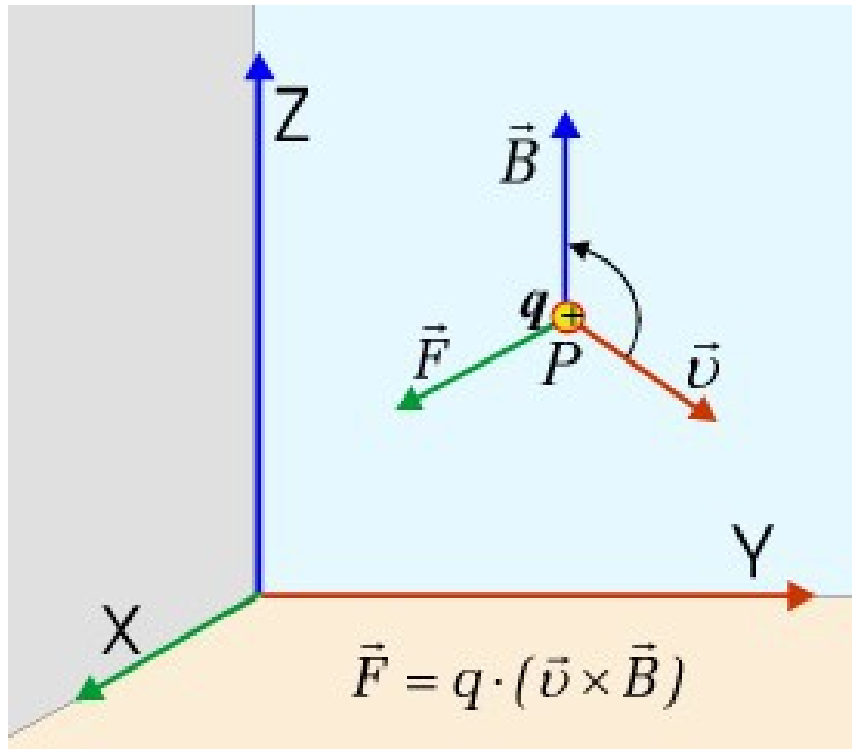
31

First order effect : electromagnetic elastic interaction of the charge particle with nuclei (heavy and multiply charged) and electrons (light and single charged)

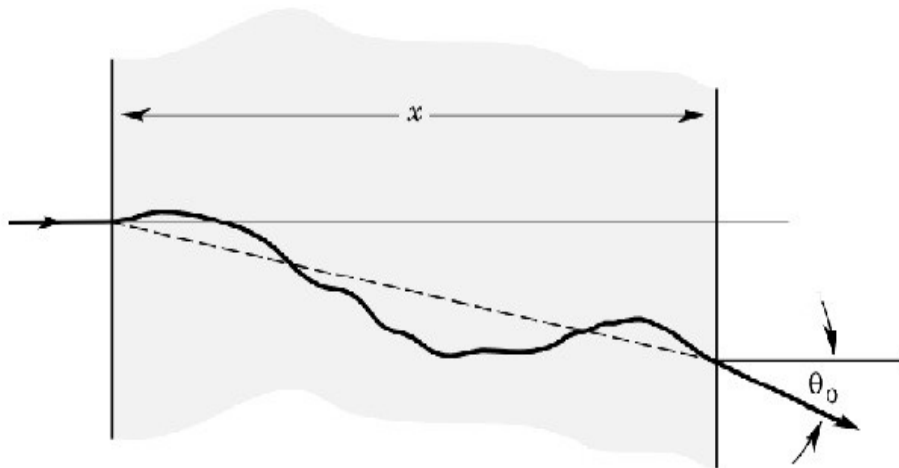
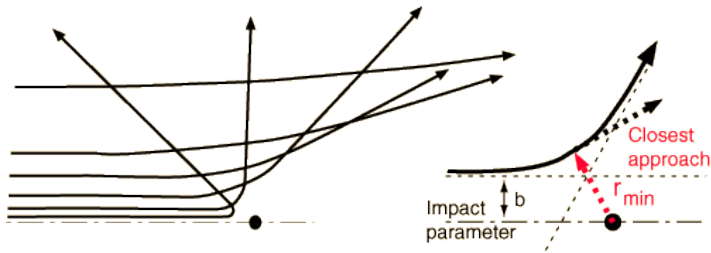
Second order effect : inelastic interaction with nuclei.

Magnetic Field

- Magnetic field \vec{B} acts on charged particles in motion : Lorentz Force
- The solution in uniform magnetic field is an helix along the field : 5 parameters
- Helix radius proportional to the component of momentum perpendicular to \vec{B}
- Separate particles in dense environment
- Bending induces radiation : bremsstrahlung
- The magnetic field has to be known to a good precision for accurate tracking of particle



Multiple Scattering



- **Deflection on nuclei** (effect from electron are negligible)
- Addition of scattering processes
- Gaussian approximation valid for substantial material traversed

Gaussian Approximation

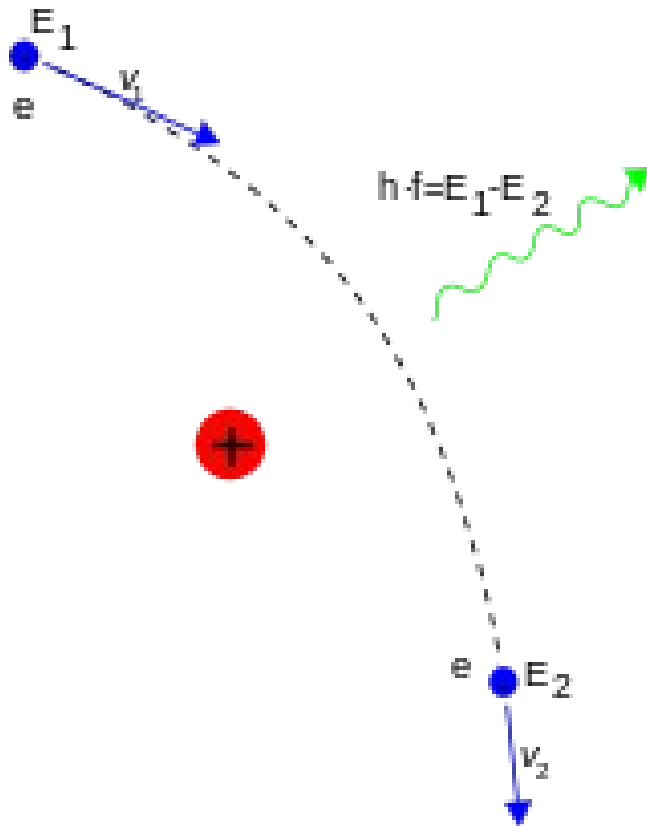
$$\theta^2 = \left(\frac{13.6 \text{ MeV}}{\beta c p} \right)^2 * \frac{x}{X_0}$$

β - particle velocity

ρ – material density

P - particle momenta

Bremsstrahlung



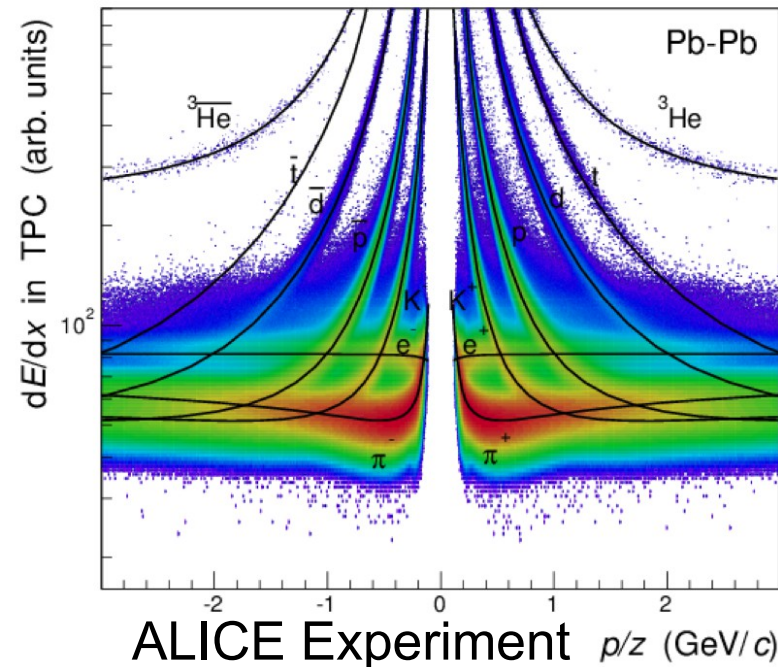
- Electromagnetic radiation of charged particles under acceleration due to nuclei charge
- Significant at low mass or high energy
- Discontinuity in energy loss spectrum due to photon emission and track curvature
- Can be observed as kink in the trajectory or presence of collinear energetic photons

Energy Loss

- Momentum transfer to electrons when traversing material (effect of nuclei is negligible)
- Energy loss at low momentum depends on mass : can be used as mass spectrometer

$$dE / dx = k_1 \frac{Z}{A} \frac{1}{\beta^2} \rho \left(\ln \left(\frac{2m_e c^2 \beta^2}{I(1-\beta^2)} \right) - \beta^2 - \frac{\delta}{2} \right)$$

β - particle velocity
 ρ - material density
 Z - atomic number of absorber
 A - mass number of absorber
 I - mean excitation energy
 δ - density effect correction factor - material dependent and β dependent



Scene Labeling



LeCunn Seminar at CERN

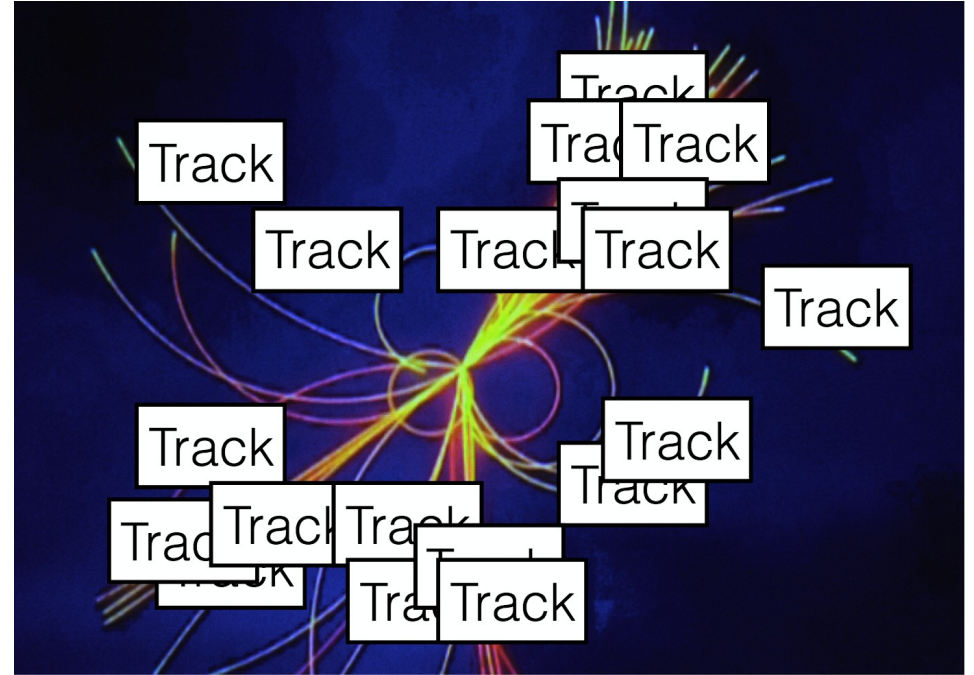


Photo by Pier Marco Tacca/Getty Images

- Partners
- Outline
- Tracking in a Nutshell
- Complexity and Ambiguity
- Cost of Tracking
- Fast Hardware Tracking
- Bottom Lines
- Scene Labeling

02/28/14 • Scene Labeling



IML