Clustering with adaptive similarity measure for track reconstruction

Connecting The Dots 2018

Sabrina Amrouche c.amrouche@cern.ch

ATLAS EXPERIMENT

universität wien

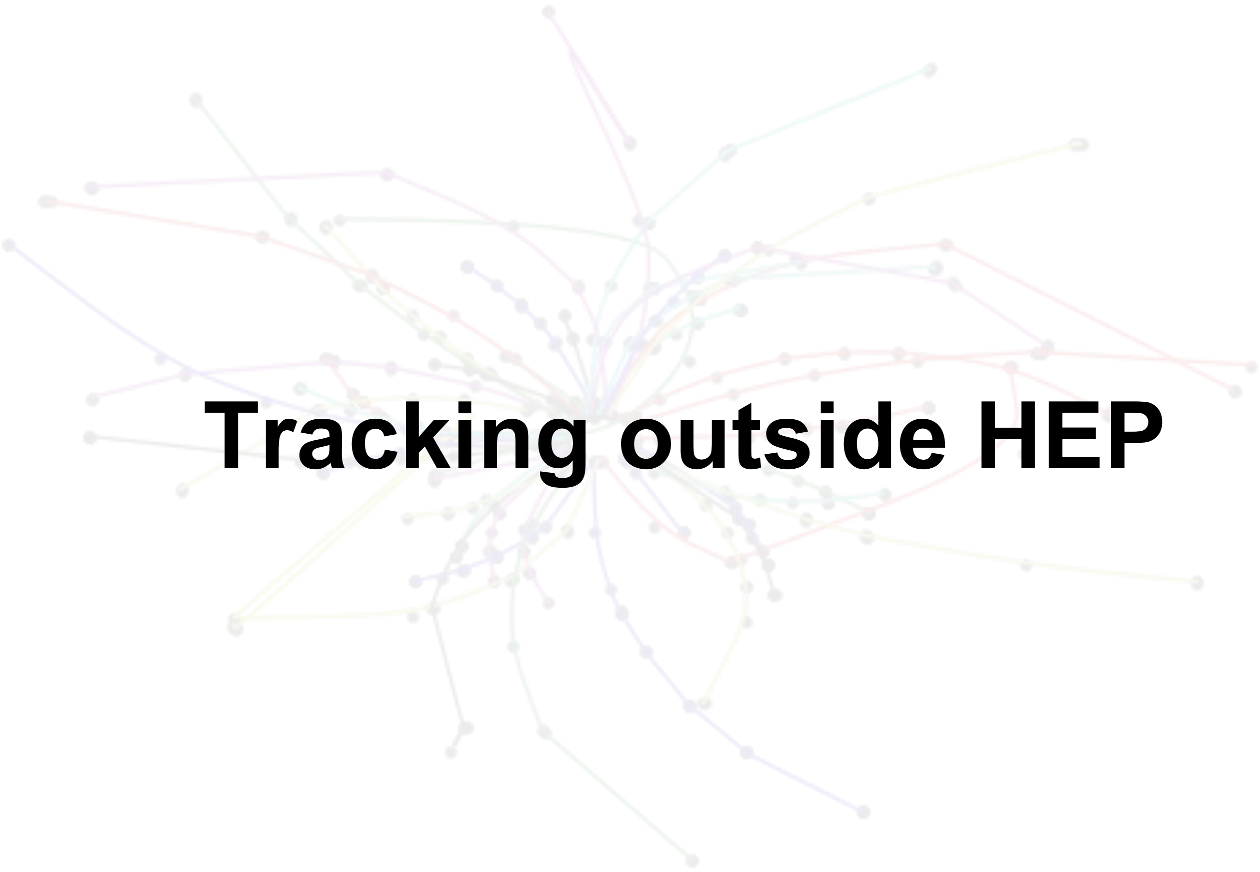
Learning to Track

Goal and Motivation

- HL-LHC will dramatically increase the reconstruction difficulty.
- Average 200 proton-proton collisions for HL-LHC.
- Combinatorics becomes very challenging (impossible).
- Need for generalization and adaptive behavior.
- Grouping sets of hits that originated from same particles.

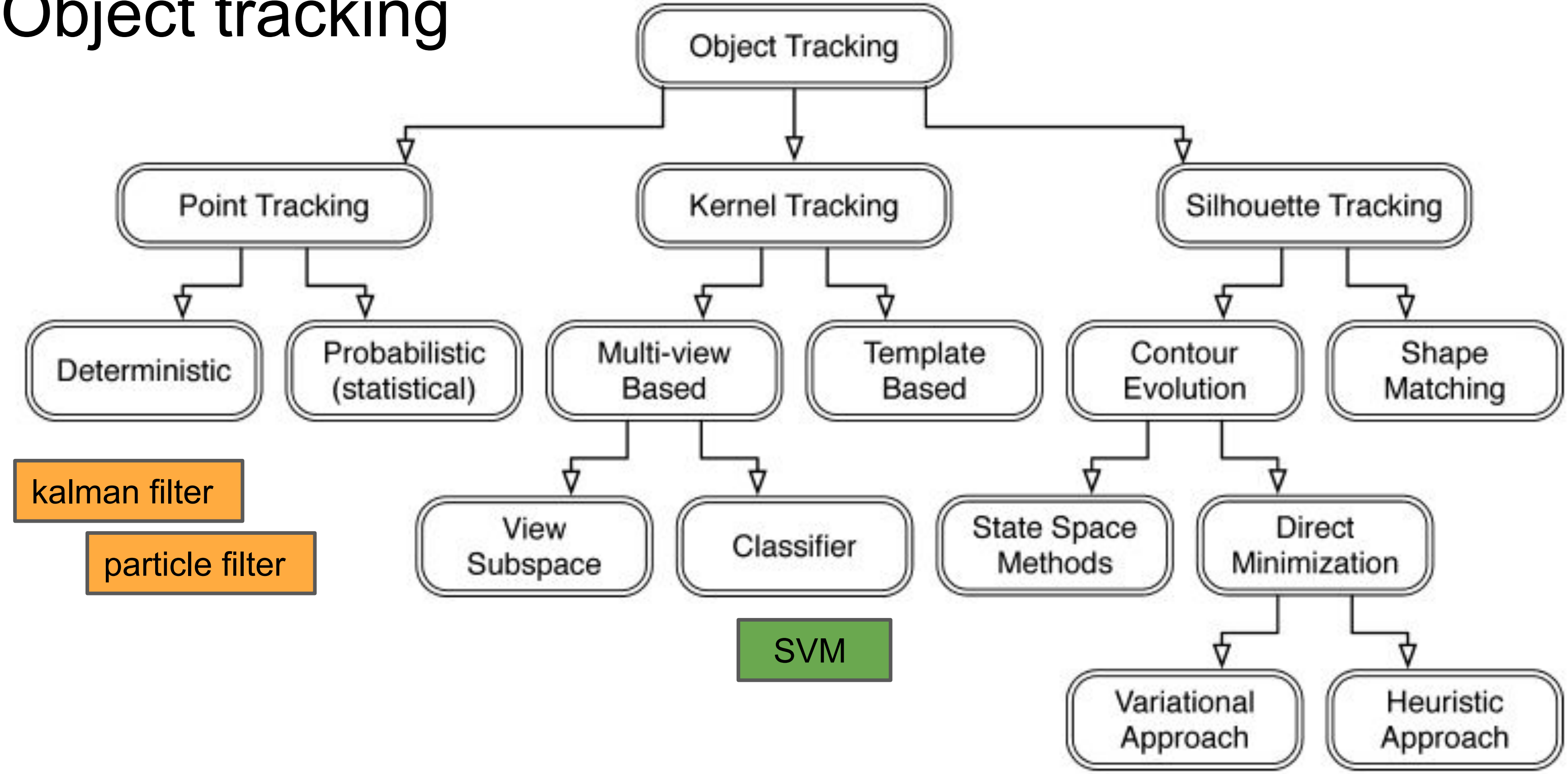
Outline

- ML and tracking outside HEP
- The (many) Challenges
- The dataset
- The adaptive clustering
- Next steps



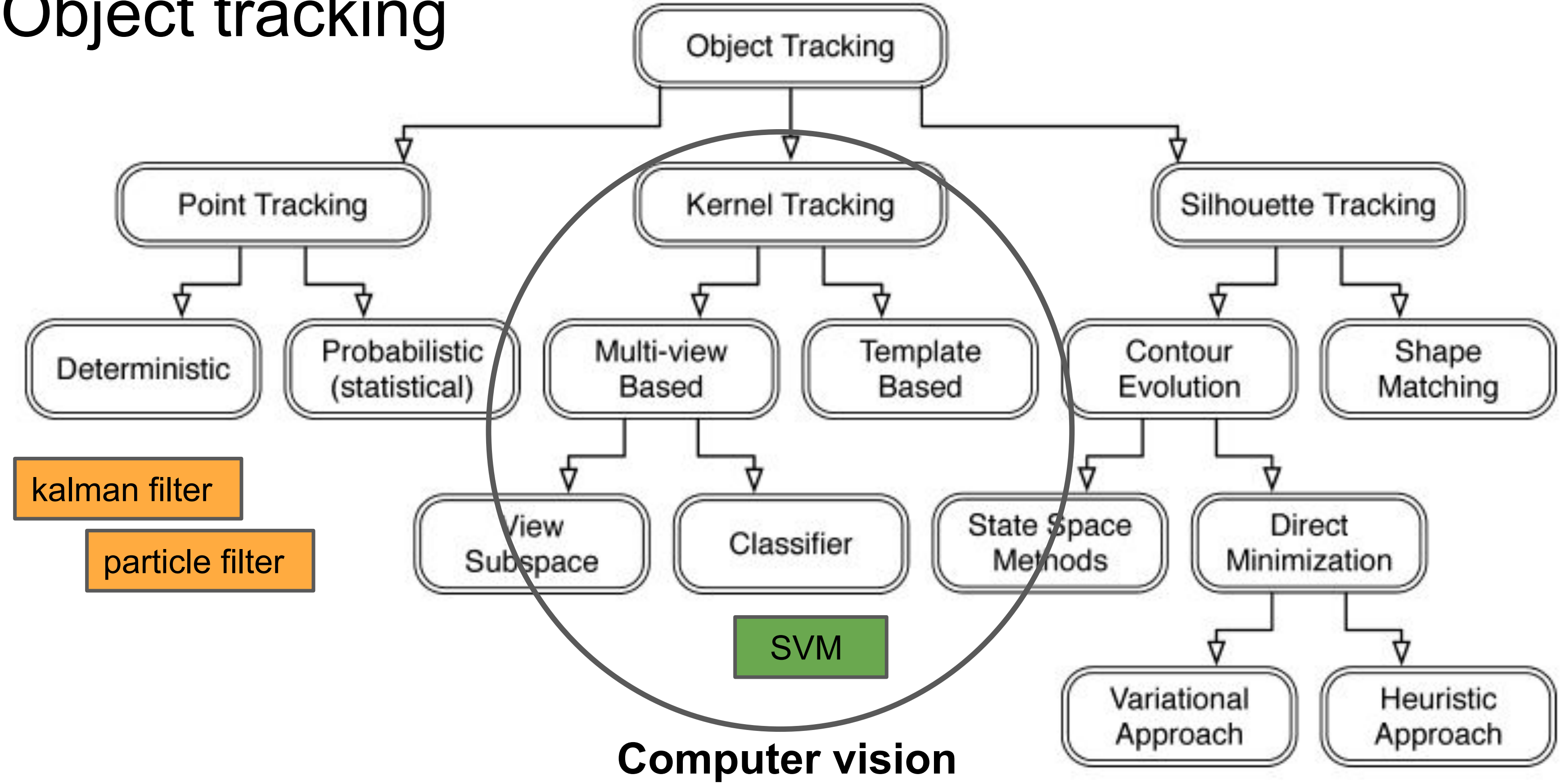
Tracking outside HEP

Object tracking



©Rout, Rupesh Kumar

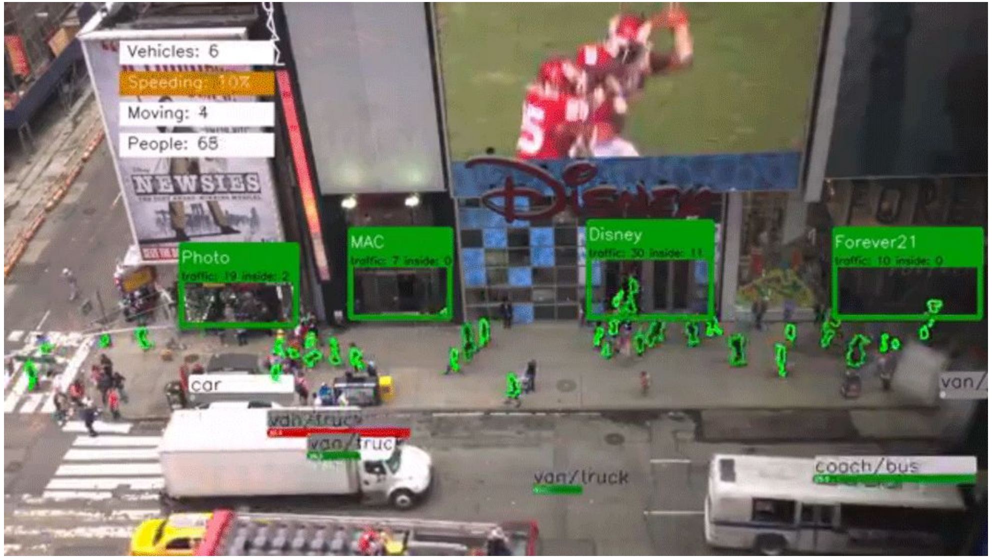
Object tracking



©Rout, Rupesh Kumar

Hunting patterns in … Computer Vision

- Object identification
- Trajectories detection
- Deep learning

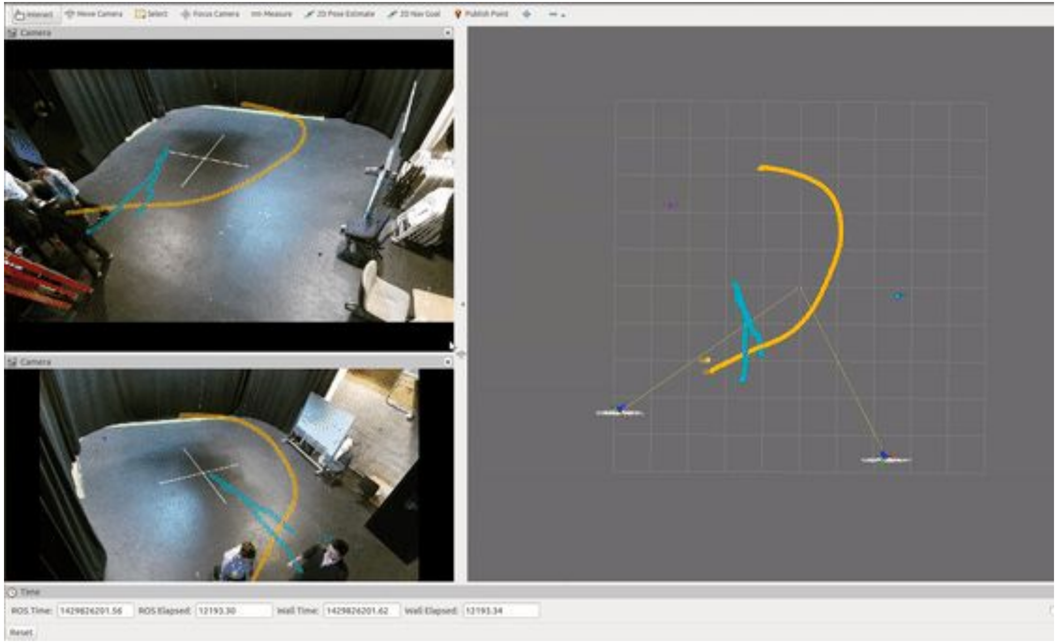


© wired.com

Hunting patterns in … Computer Vision

Key differences

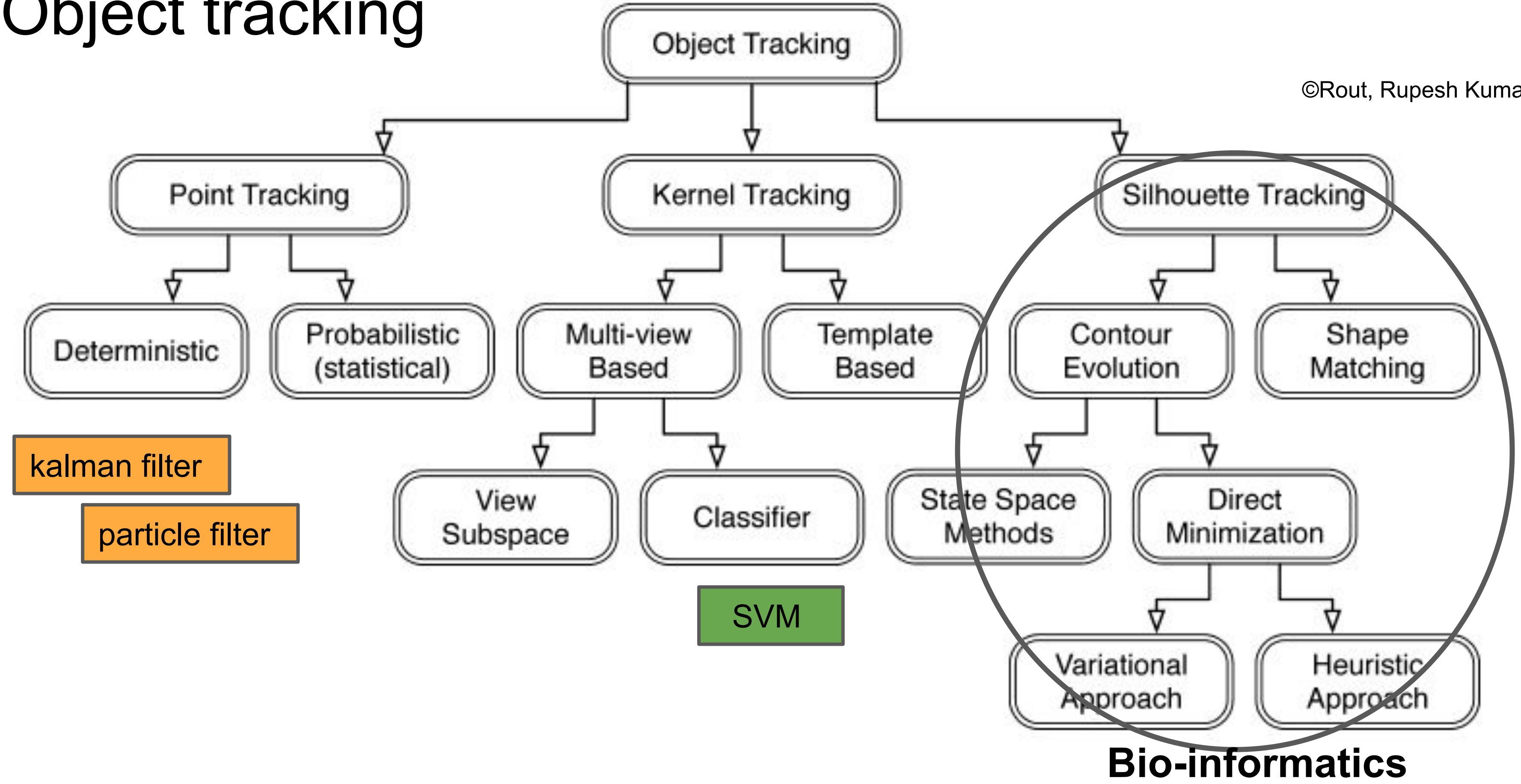
- Online
- Human density<< HEP pileup
- Individual characteristics



Objects

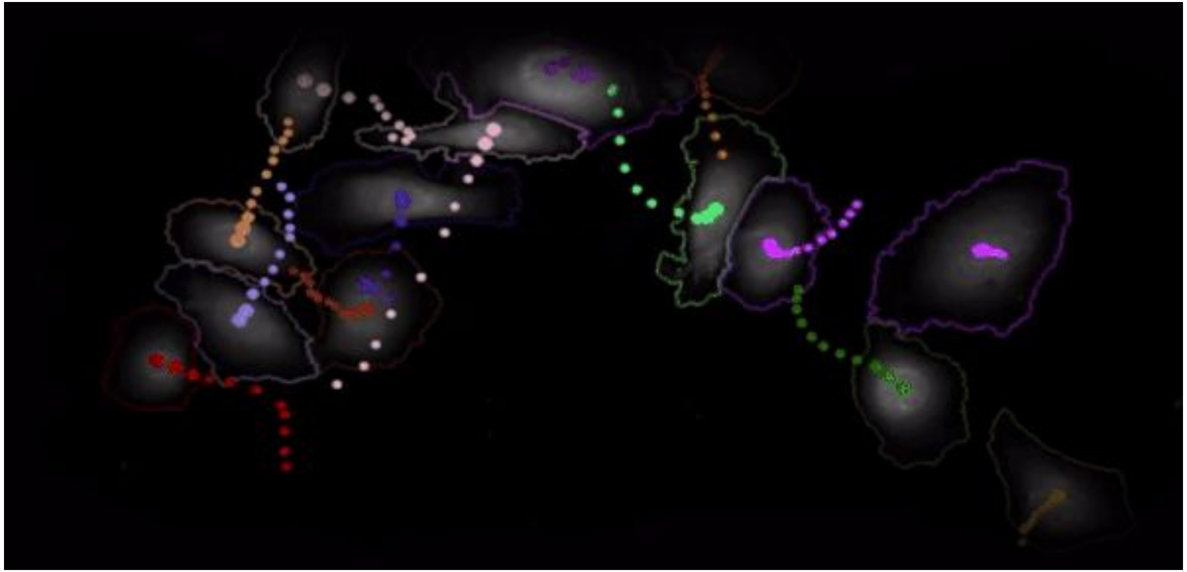
Object tracking

©Rout, Rupesh Kumar



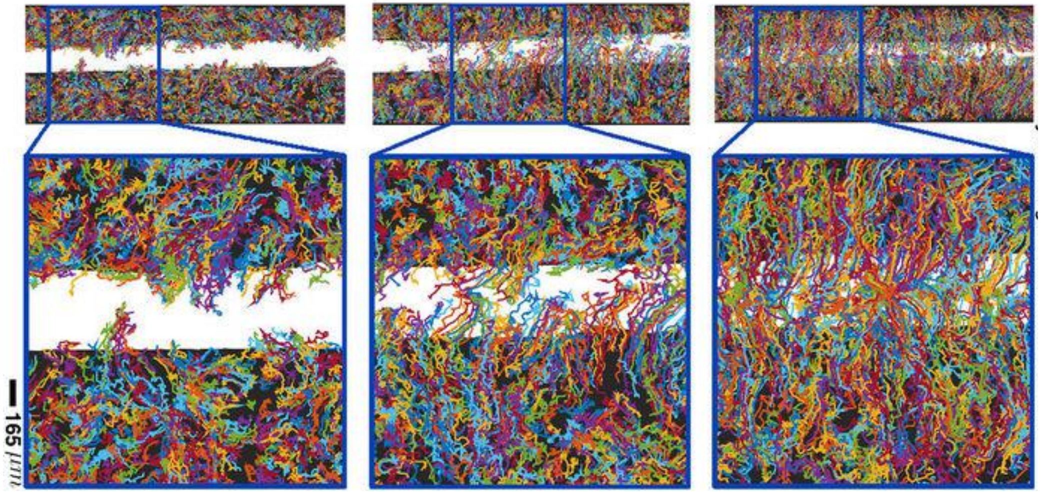
Hunting patterns in … Bioinformatics

- Cells tracking
- Extremely complex tracks
- Random behaviors



Hunting patterns in … Bioinformatics

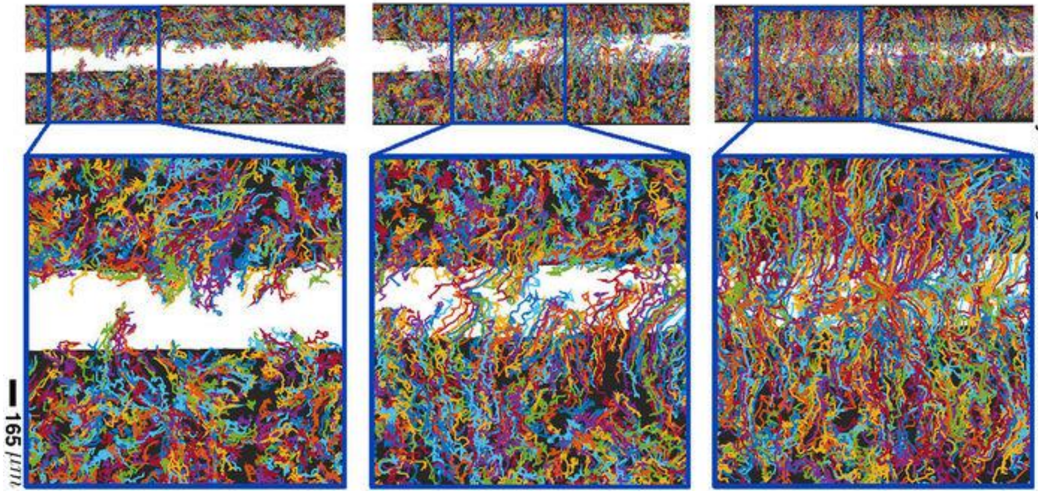
- Cells tracking
- Extremely complex trajectories
- Random behaviours
- Dense environments over long periods



Pseudo-trajectories generated from PIV data
(Baker, Richard M., et al)

Hunting patterns in … Bioinformatics

- Cells tracking
- Extremely complex trajectories
- Random behaviours
- Dense environments over long periods



Pseudo-trajectories generated from PIV data
(Baker, Richard M., et al)

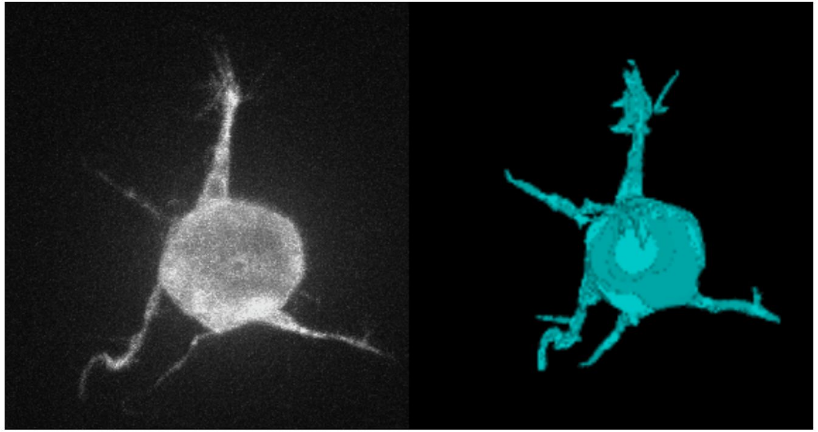
Colors based identification + **Time component**

Cell tracking challenge

- IEEE International Symposium on Biomedical Imaging 2013, 2014, 2015
- Open since 2017 for online submissions and **results are public**
- celltrackingchallenge.net

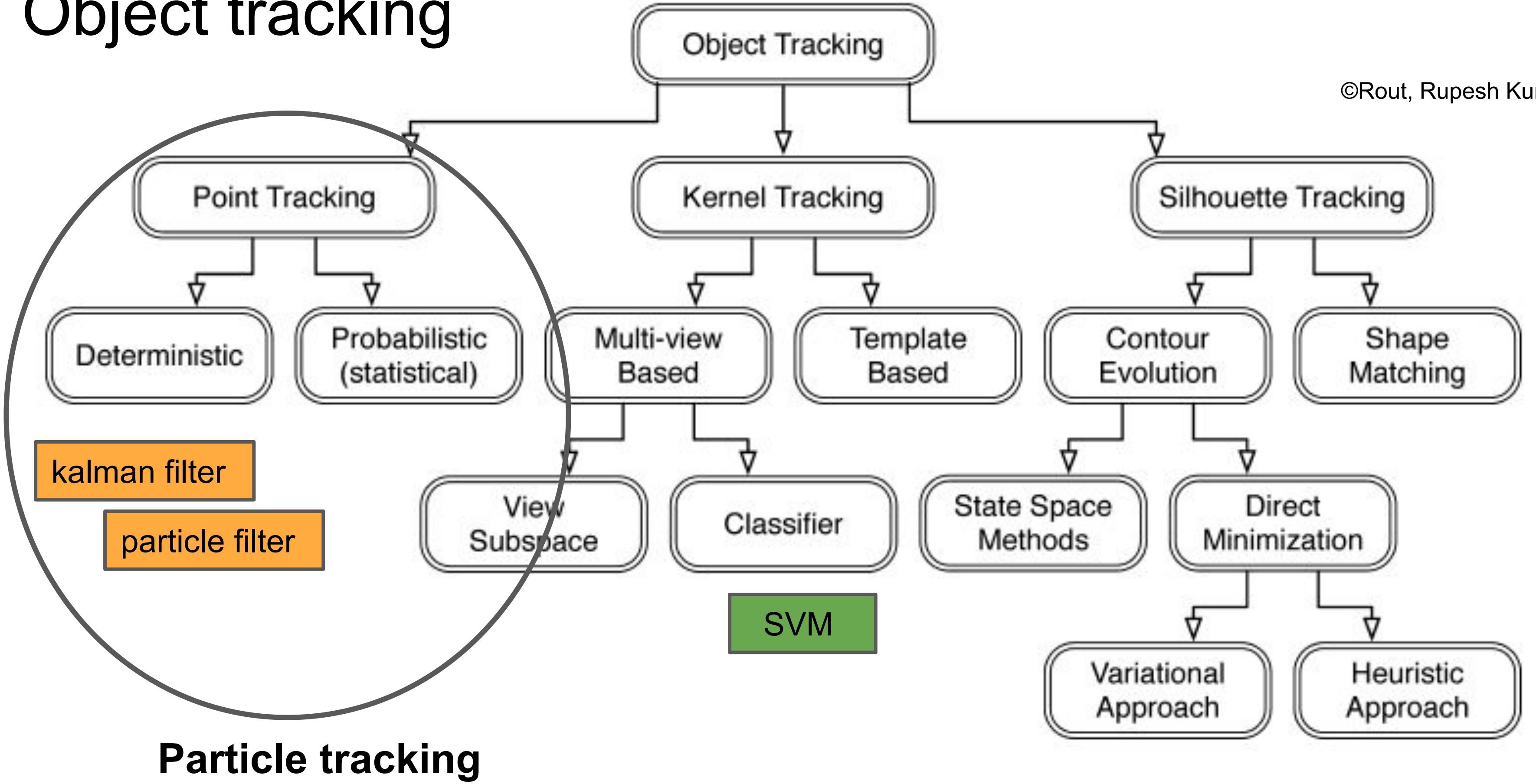
Welcome Organizers Description Datasets Registration
Evaluation Submission of Results Participants Latest Results

Cell Tracking Challenge



Tracking moving cells in time-lapse video sequences is a challenging task, required for many applications in both scientific and industrial settings. Properly characterizing how cells move as they interact with their surrounding environment is key to understanding the mechanobiology of cell migration and its multiple implications in both normal tissue development and many diseases. In this challenge we objectively compare and evaluate state-of-the-art whole-cell and nucleus tracking methods using both real (2D and 3D) time-lapse microscopy videos of labeled cells and nuclei, along with computer generated video sequences simulating nuclei moving in realistic environments.

Object tracking



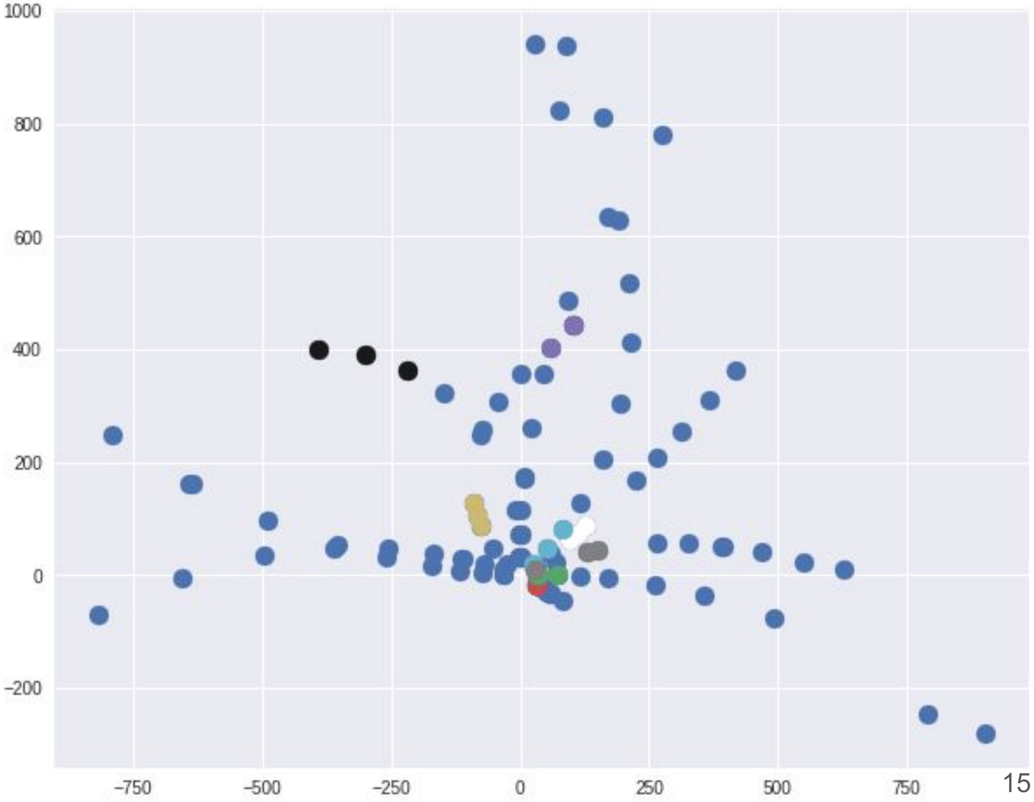
Unsupervised Learning : Clustering

- Grouping hits into the *right* particles.
- Bottom-up approach : until inconsistency reached.
- Distance measure such as:

$$d(h_{1i}, h_{1j}) < d(h_{1i}, h_{2j})$$

$\forall$ i,j $\in$ [0,N] where N : track size

- Incorporate the **domain knowledge** into the distance



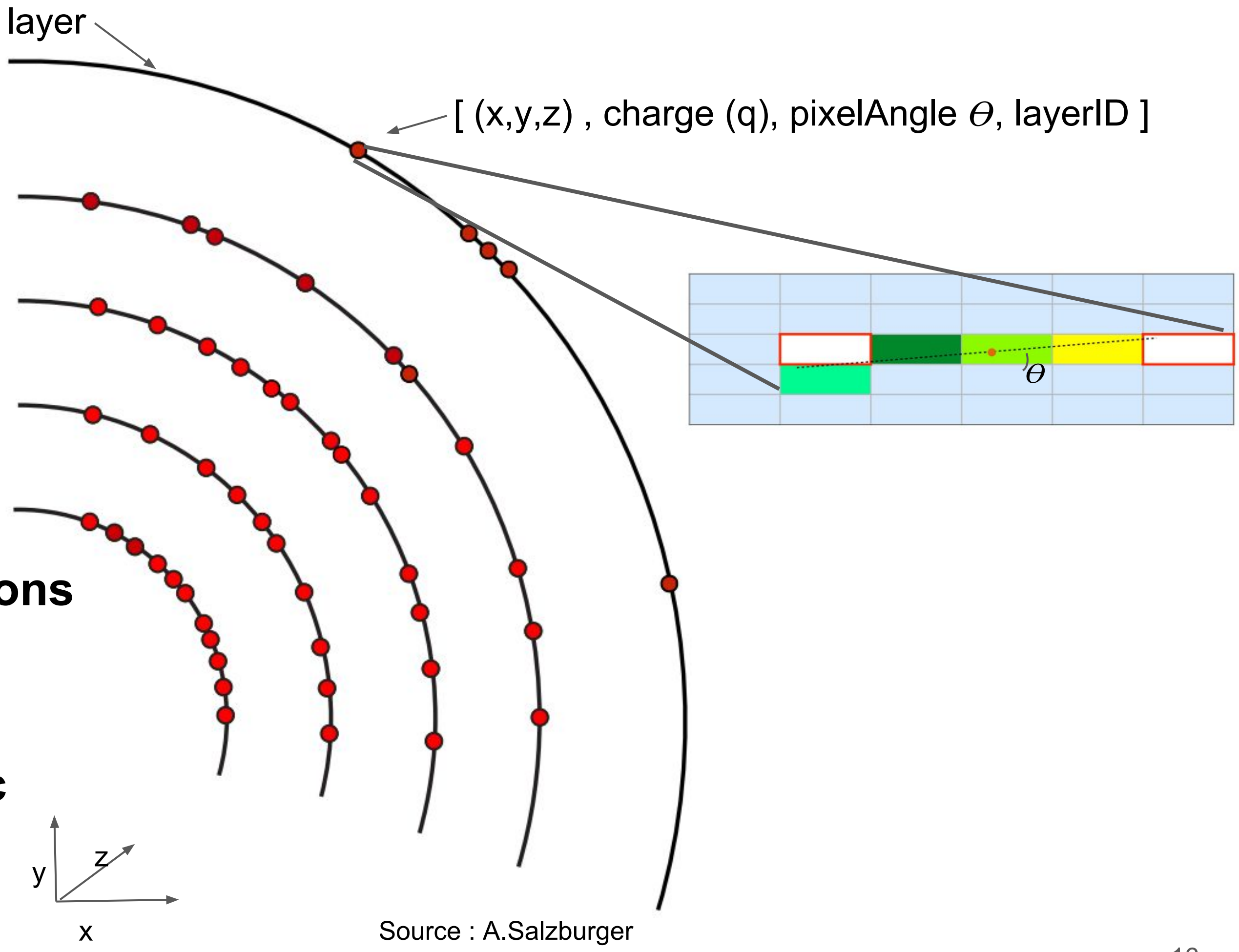
The dataset

- Simulation with **ACTS**
- Example stat
  - 110.961 points
  - 10.080 trajectories
  - ~ 11 point per track

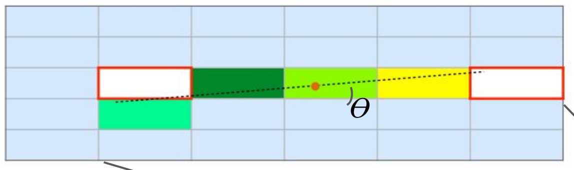
We have the truth associations

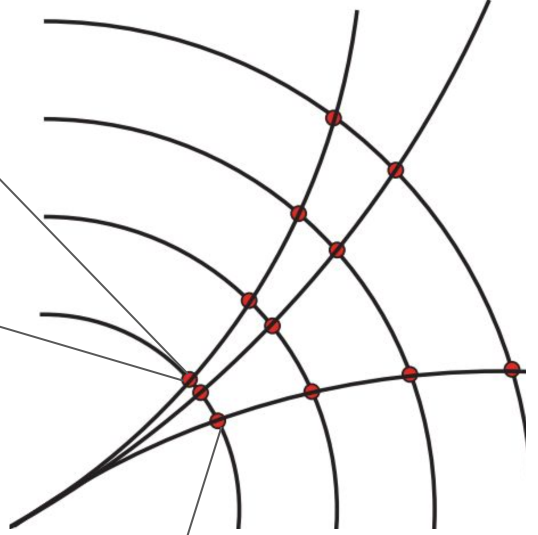
→ labels

Therefore : Accuracy metric



Source : A.Salzburger





Particles.csv

particle_id	vx	vy	vz	px	py	pz	q

Hits.csv

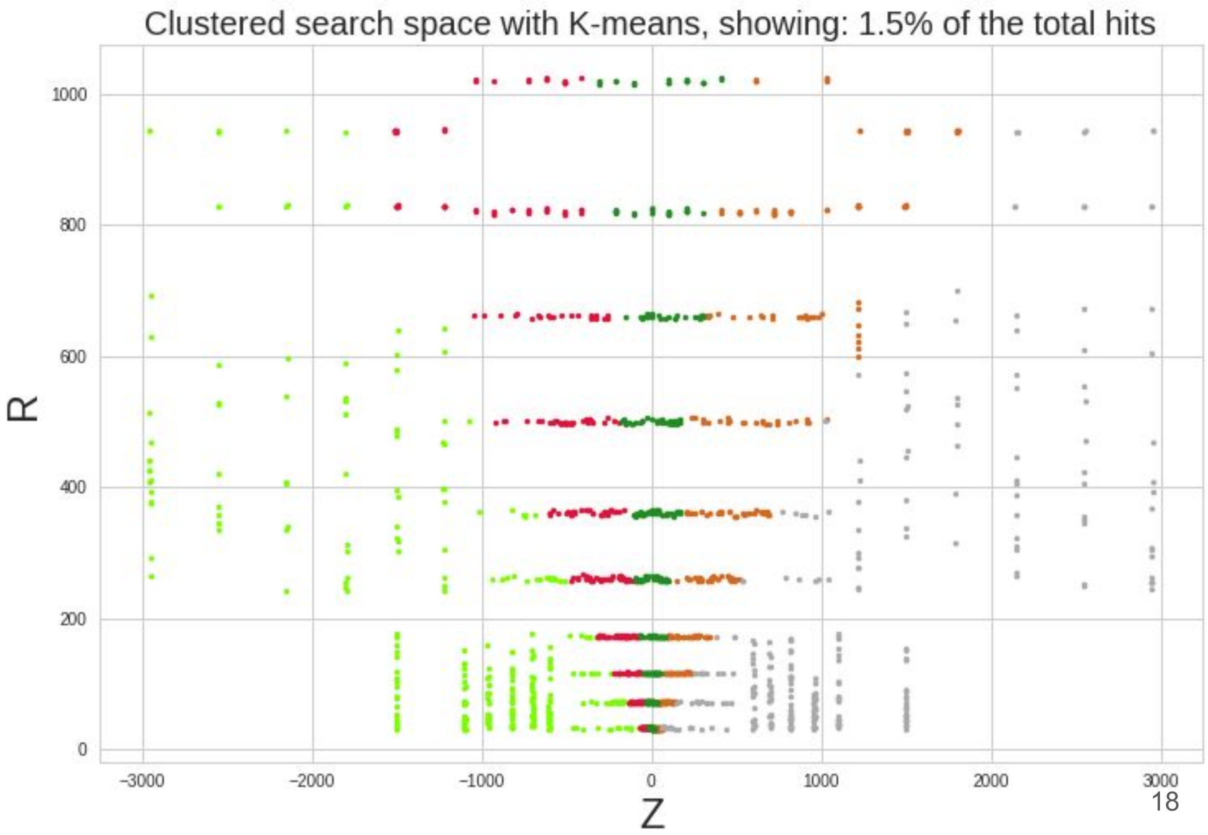
hit_id	volume_id	layer_id	module_id	x	y	z	ncells	ch0	ch1	value

Truth.csv

hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight

Creating initial sub-spaces

- Colors identify subspaces
- Reducing the search space by a factor of ~5.
- Favors parallelized processing : subspace per core.



Clustering v0

Metric 1 : Cosine on xyz

- Pairwise distance
- distance= cos(θ)= $A.B / ||A||_2.||B||_2$
- Hits to triplets
- **86%** true triplets on 1 event (2668 particles)
- **0.98%** true pairs

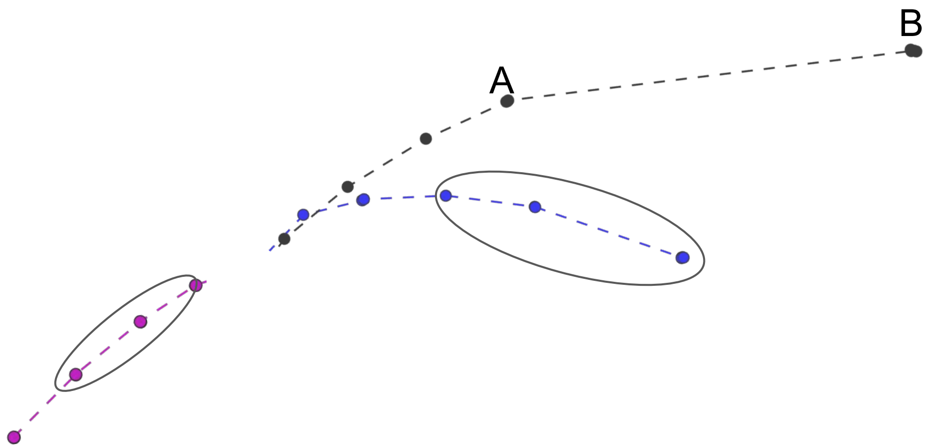


Fig 1 Hits into triplets

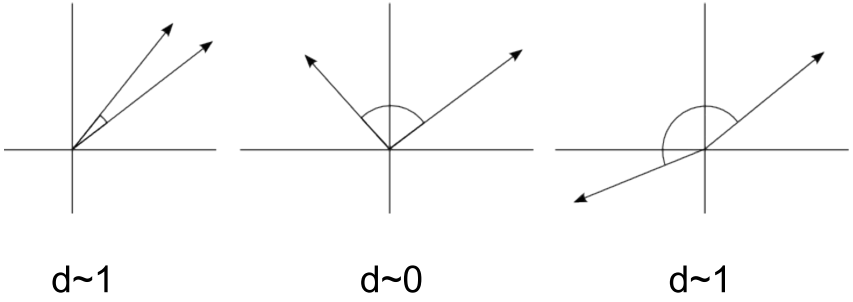


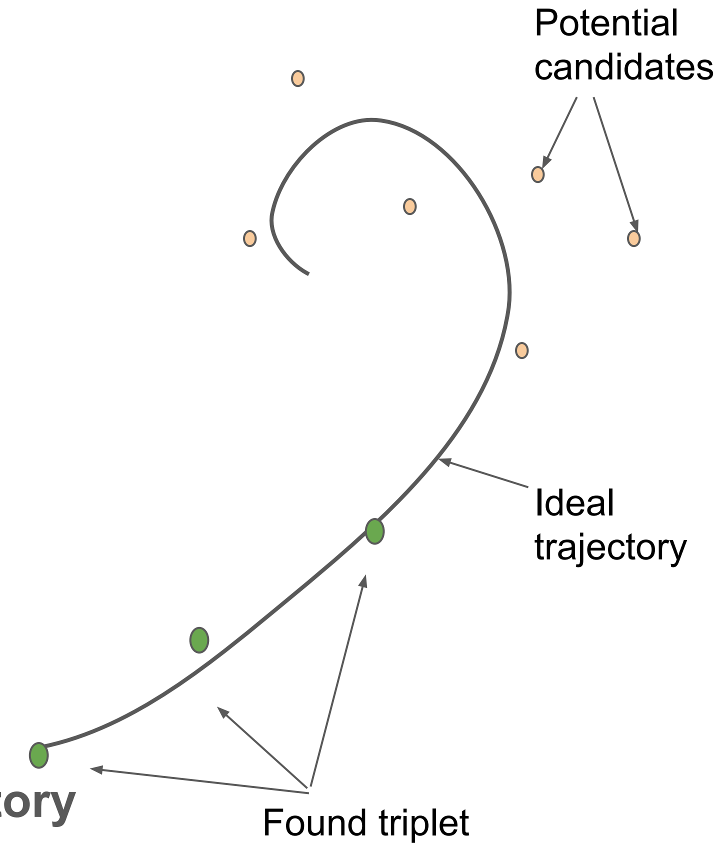
Fig 2 Cosine definition

Triplets to tracks

- Particles trajectories are helixes
- Helix = circle (xy) + line (z)
- A circle is formed by at least 3 points

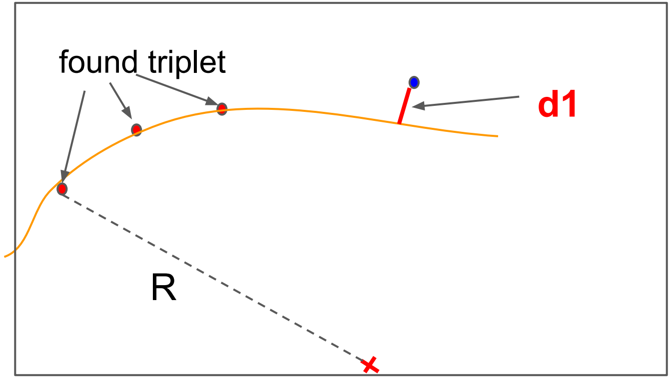
Triplets ⟶ circle fit (xy) + line fit (z)

⟶ distance to trajectory

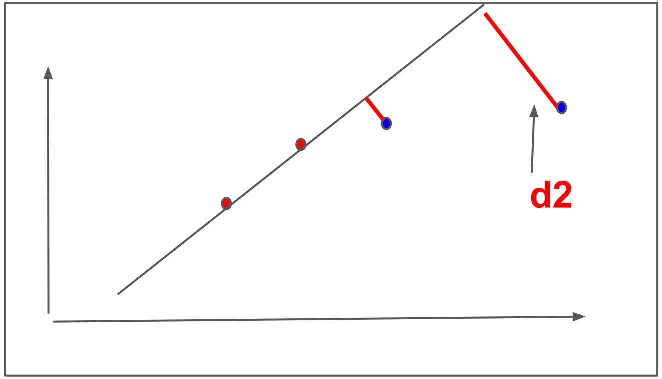


The linkage metric

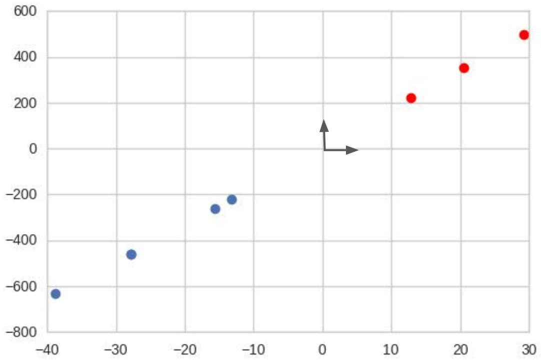
X Y plane



R Z plane
sqrt(x^2+y^2)



- Adaptive distance = min(d1 , d2)
- d1
  - distance to fitted circle
  - circle is updated after each merge (**adaptive**)
- d2
  - avoid grouping symmetric particles
  - Adds z coordinate constraint

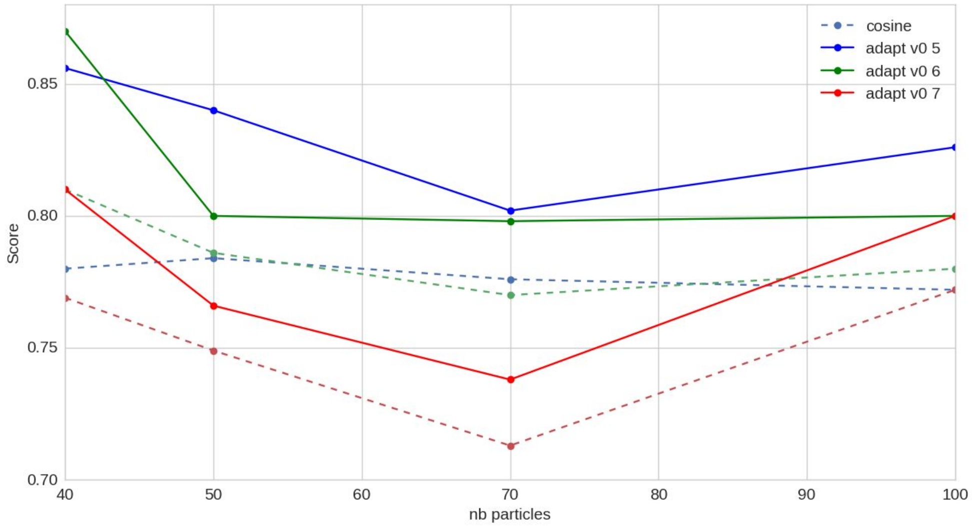


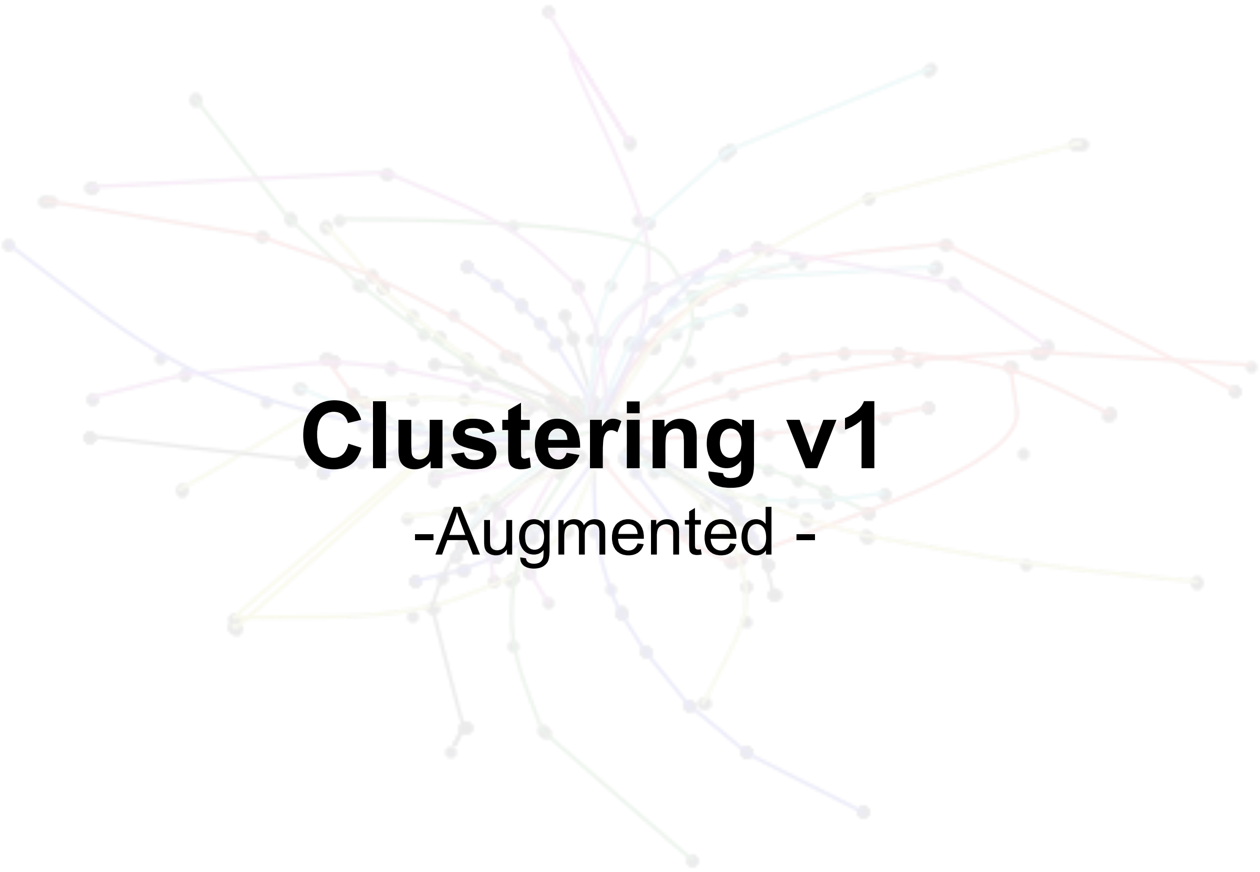
Symmetric particles

Initial trial

- Using a single metric on xyz (cosine) and add hits by distance to ideal track
  - Relying only on x.y.z
  - initial clusters (seeds) are only triplets (sometimes less)
  - Search space for each found seed is too big

Nb particles	Method	Score
120	cosine (**pre-clustering**) Adaptive	0.78 0.86
1000	cosine adaptive	0.62 0.68
1 event (2668 particles)	cosine Adaptive	0.48 0.55





Clustering v1

-Augmented -

Metric 1 : Cosine on xyz

- Pairwise distance
- distance= $\cos(\theta)= A.B/||A||_2.||B||_2$

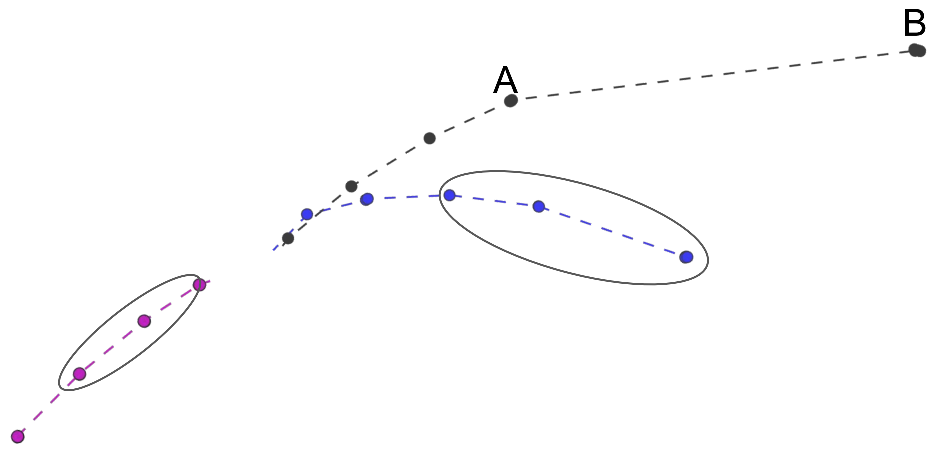


Fig 1 *Hits into triplets*

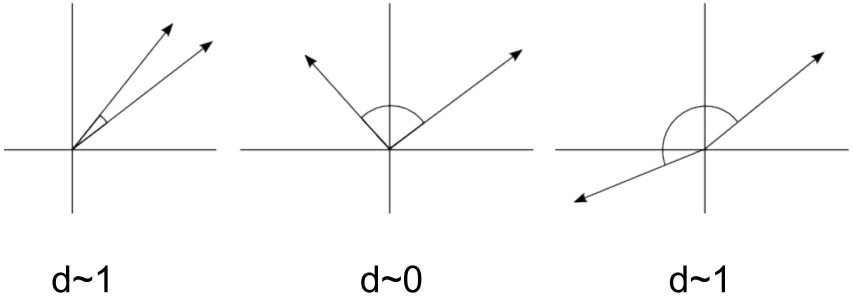
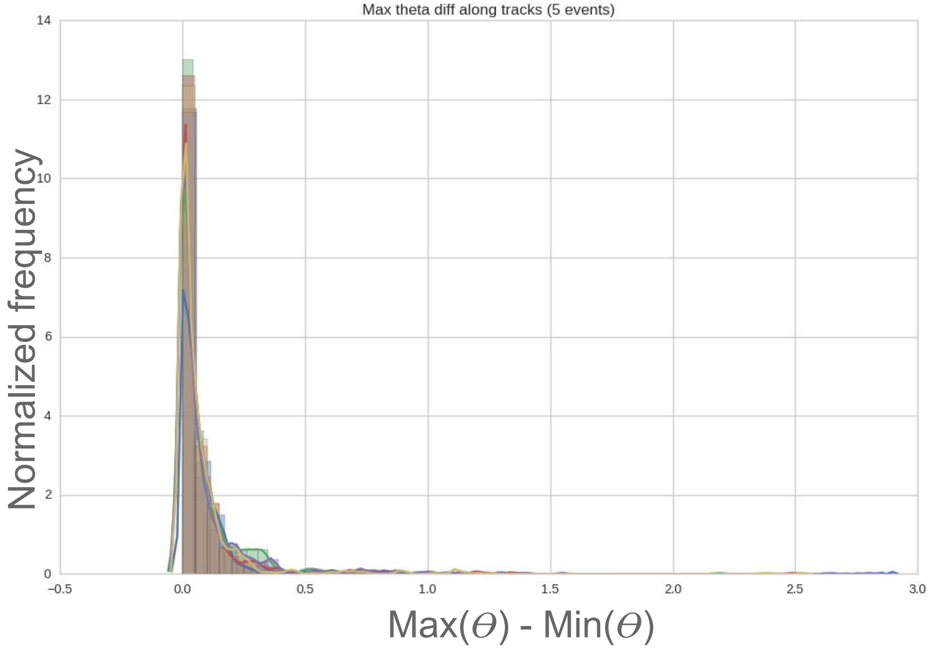


Fig 2 *Cosine definition*

B

Metric 2 : Euclidean on θ, ϕ (pixel angles)

- Pairwise distance
- distance= $||(\theta_1,\phi_1)-(\theta_2-\phi_2)||$



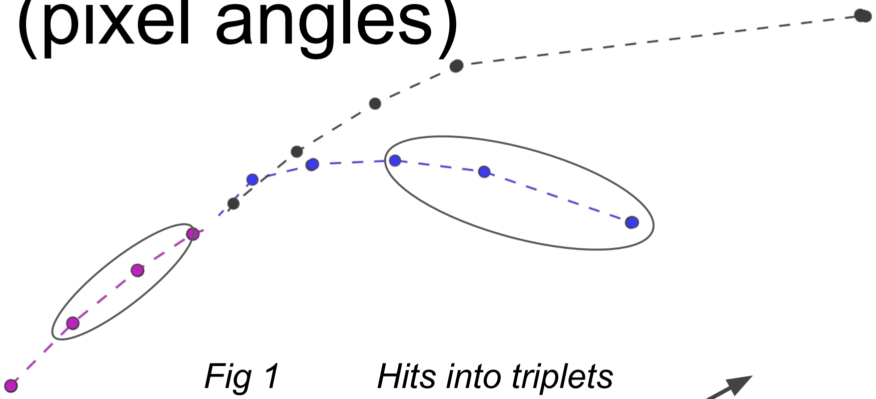
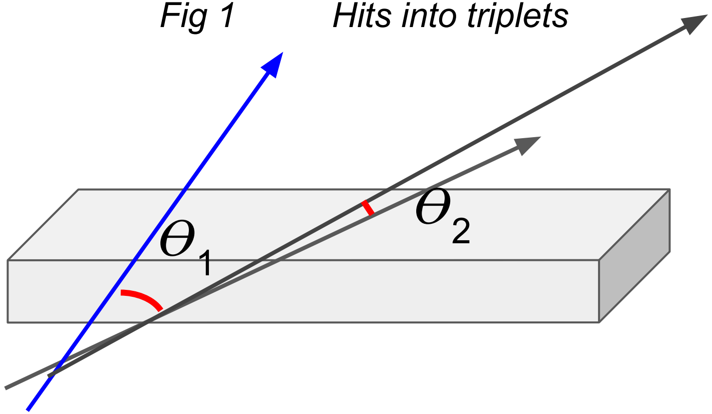
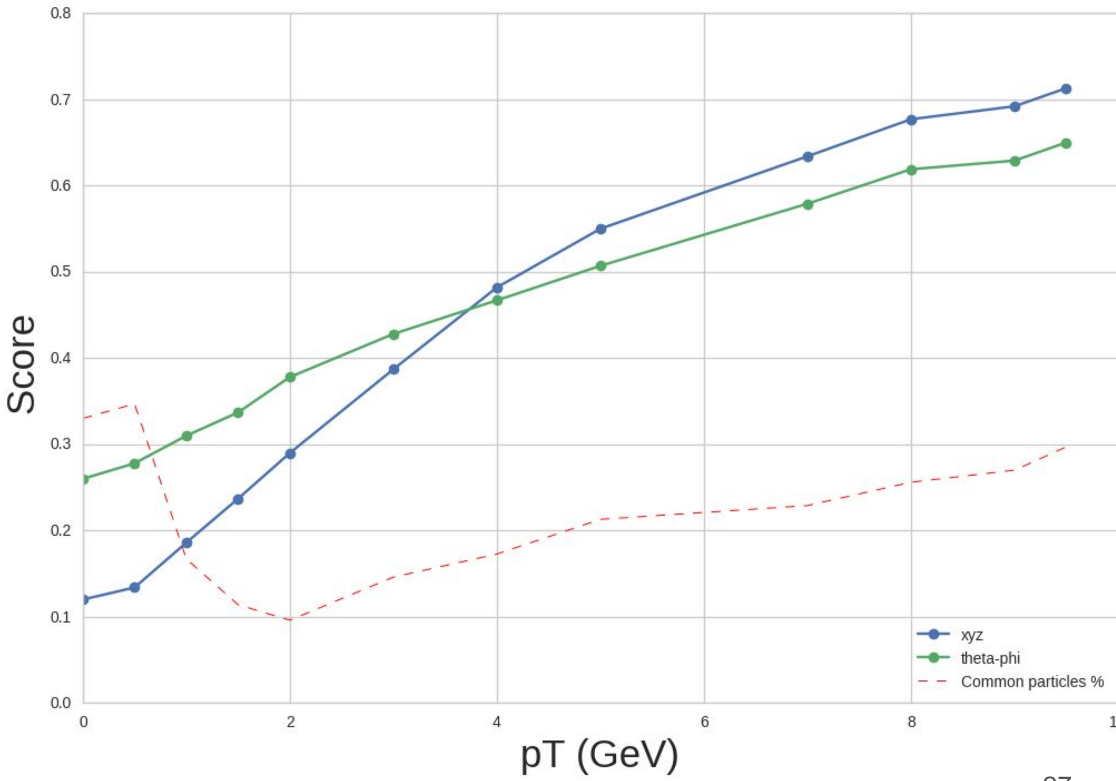


Fig 1 *Hits into triplets*



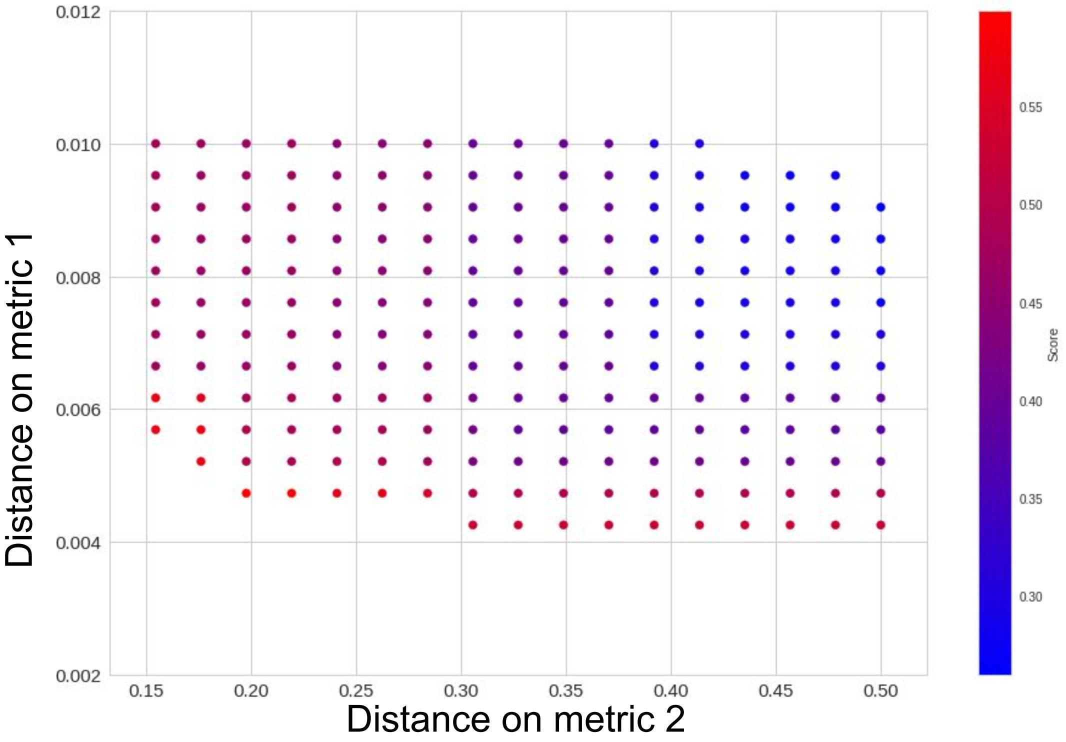
Merging distance by pT ranges

- Considering 1000 particles
- **Alternating** performances with regions
- Two metrics address different particles (types)
- Adding metrics increase coverage

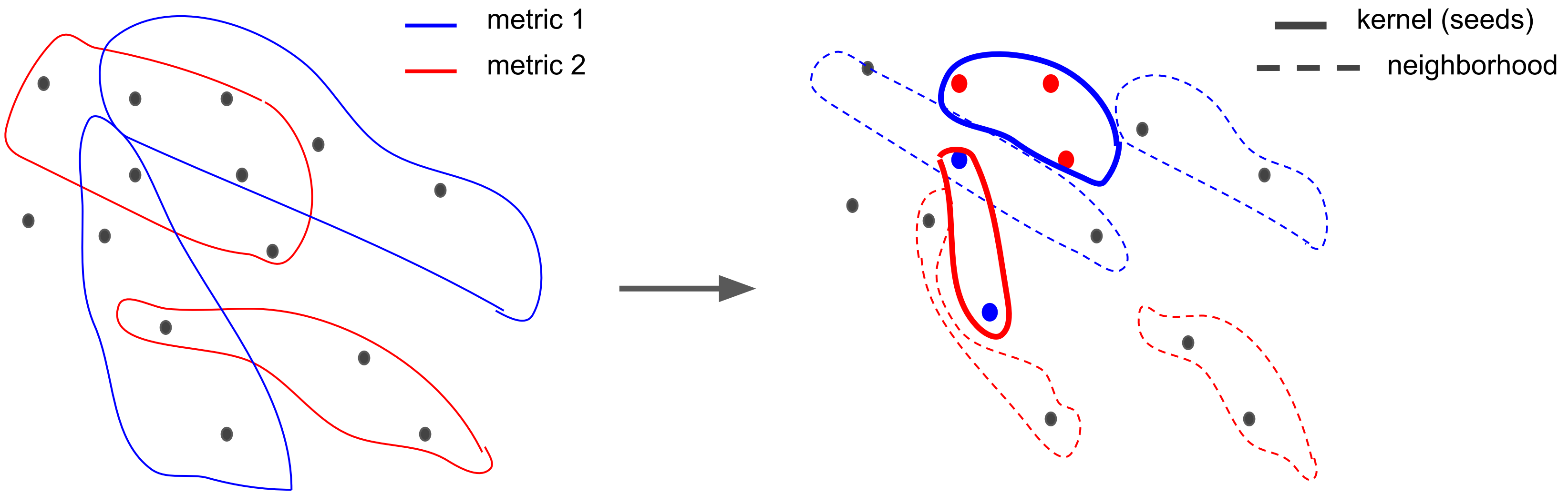


Finding optimal collaboration regions

- Looking for a space where both distances are meaningful
- Found clusters have higher confidences : Compact in both metrics
- Seeds and search space



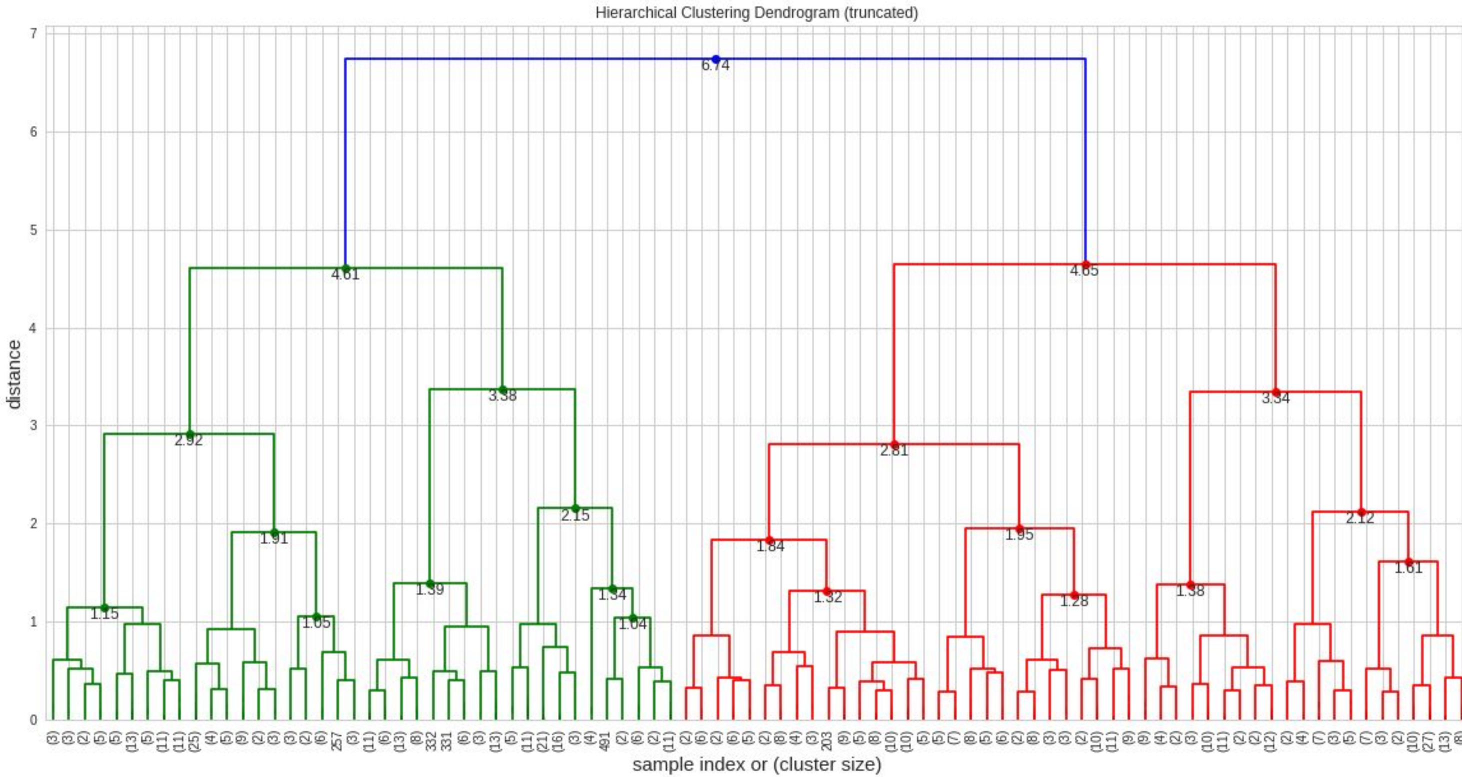
The kernel and the neighborhood



- Seeds : Clusters found by both metrics
- Neighborhood: Union of disjoint labeling

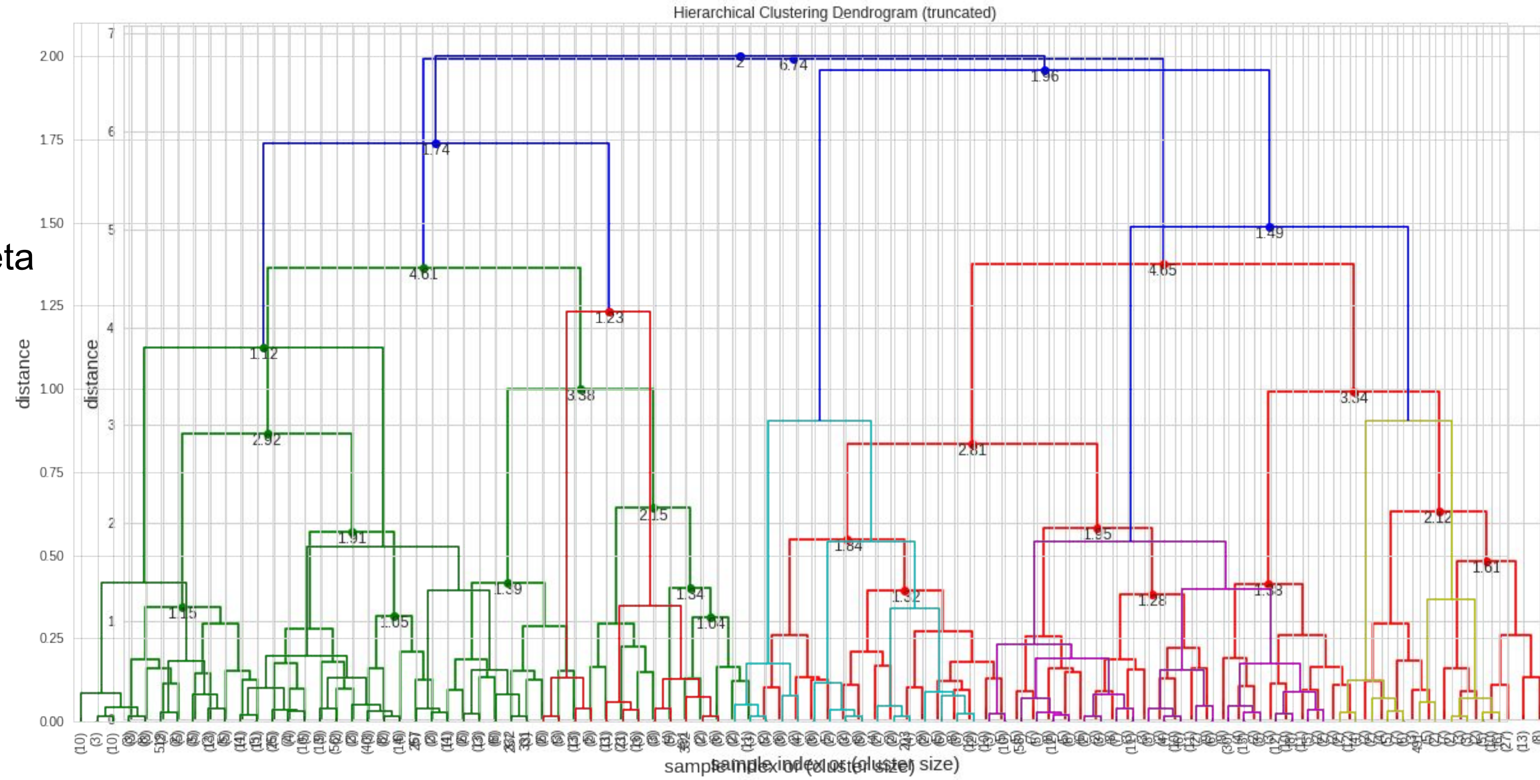
In reality...

- Metric 1 : on xyz

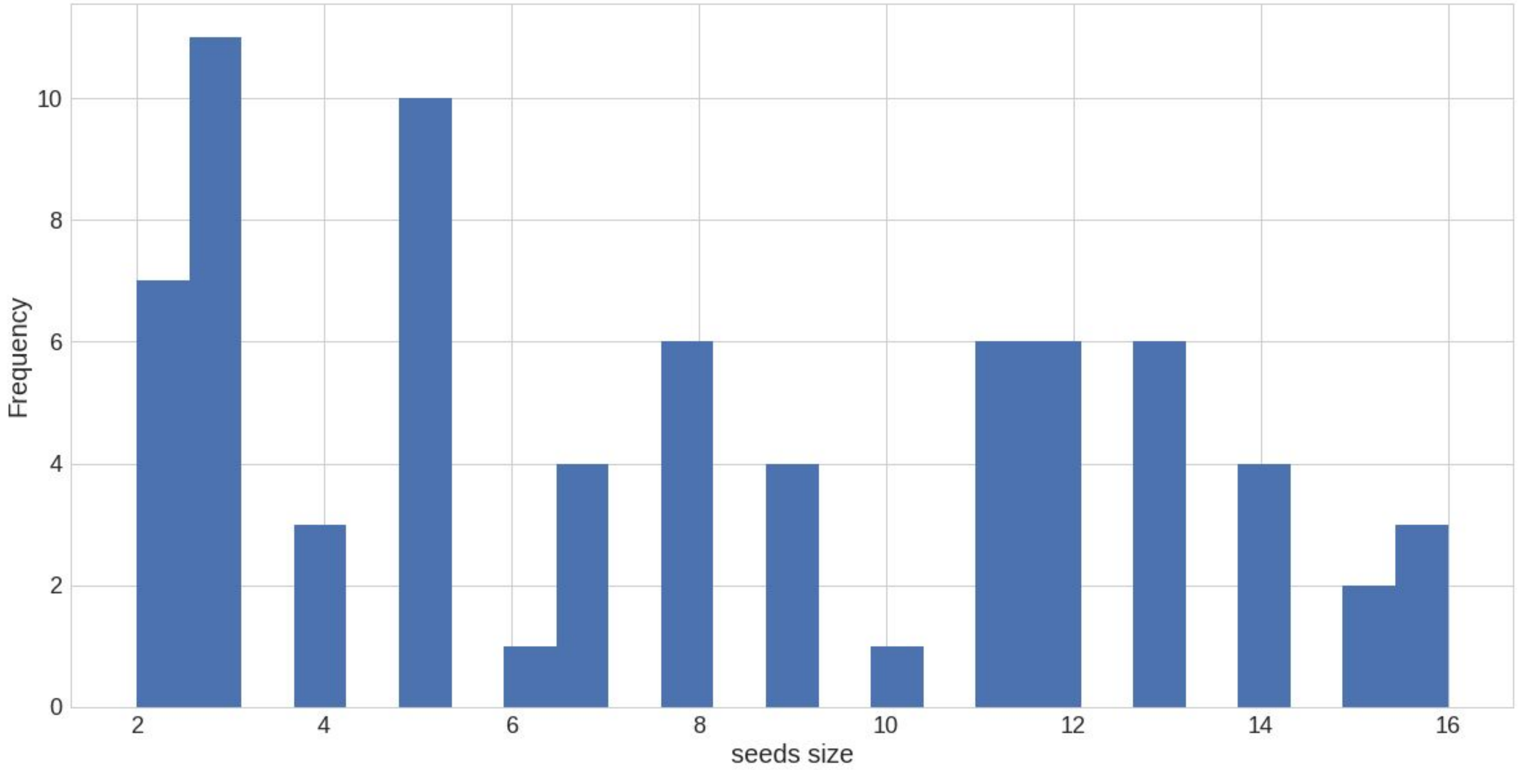


In reality...

- Metric 1 : on xyz
- +Metric 2 : on Theta

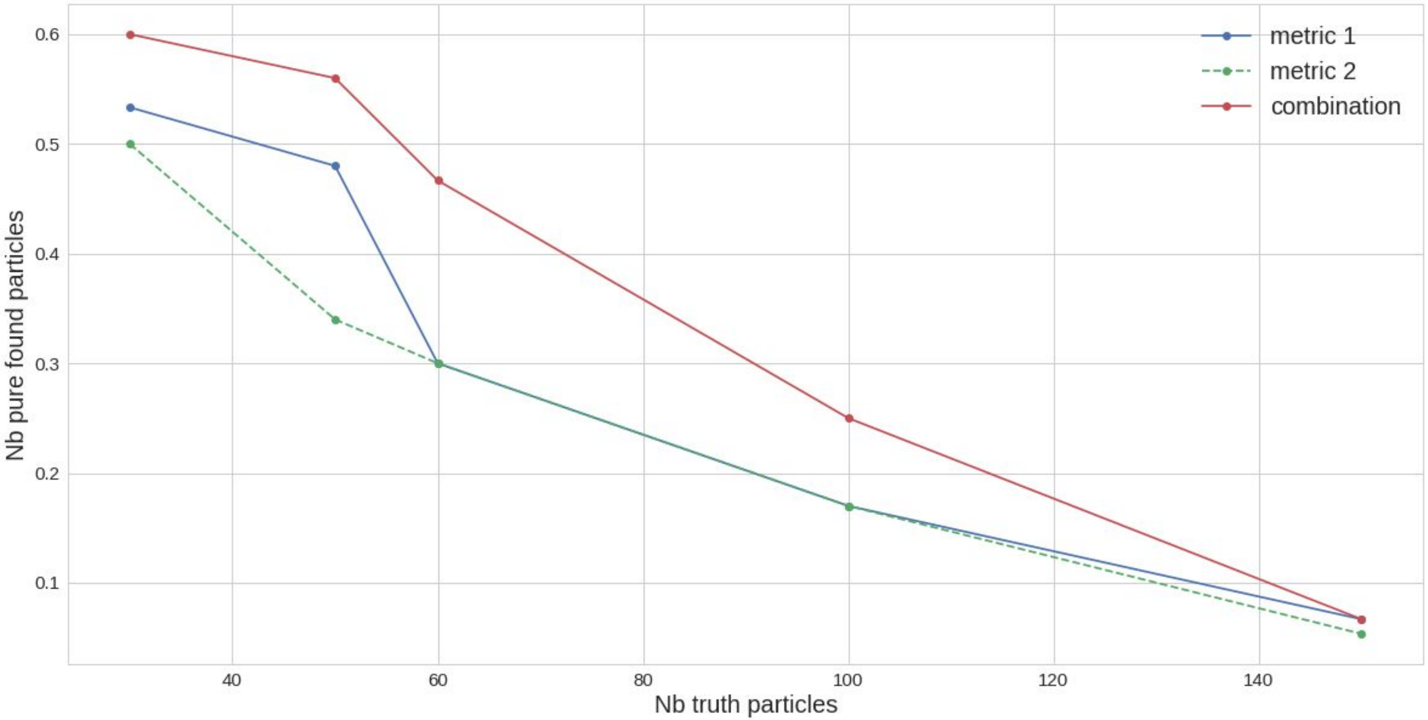


Combining the distances



- 50 particles example, 76 kernels
- Cut on kernel size

Combining the distances



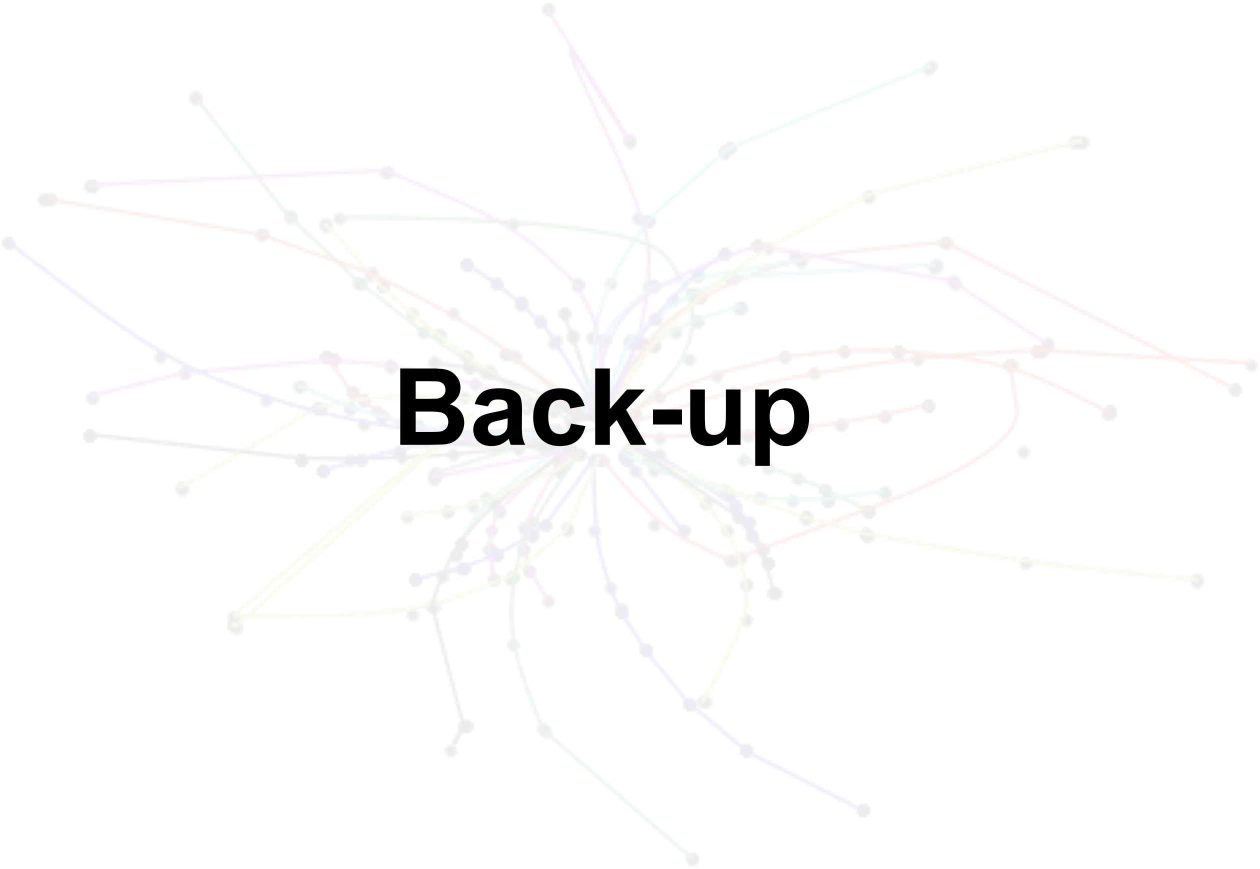
- Some of the seeds are already found/complete particles
- Final seeds are the intersection of the two clusterings

Outlook and next steps

- Major difference to current Atlas impl
  - No combinatorics, no consistency checks/matches.
  - Larger seed dimensions (full particles retrieved).
  - Search space reduced by adding more features.
- Any additional (engineered) feature will refine the neighborhood for speed and accuracy.
  - Time information will be added as an euclidean distance measure (ACTS)
- All clustering can be parallelised

Thank you

Questions?

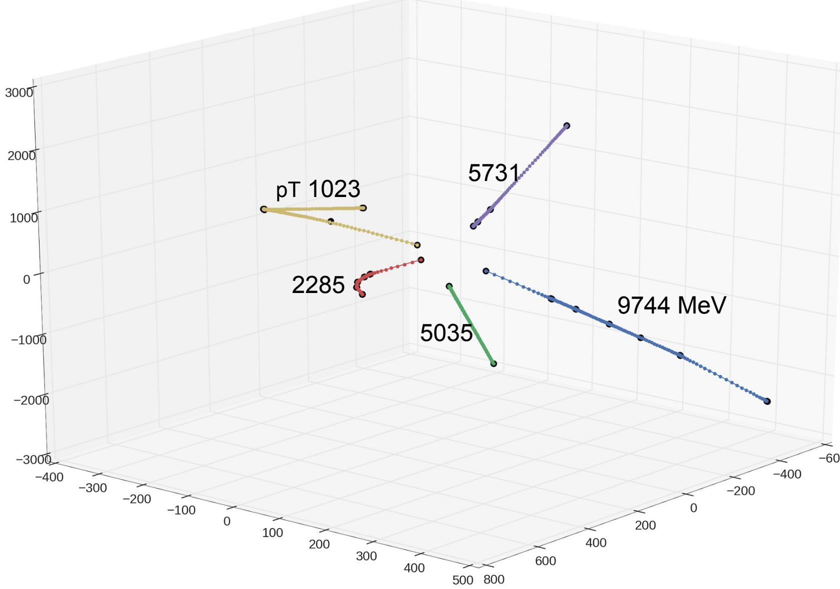


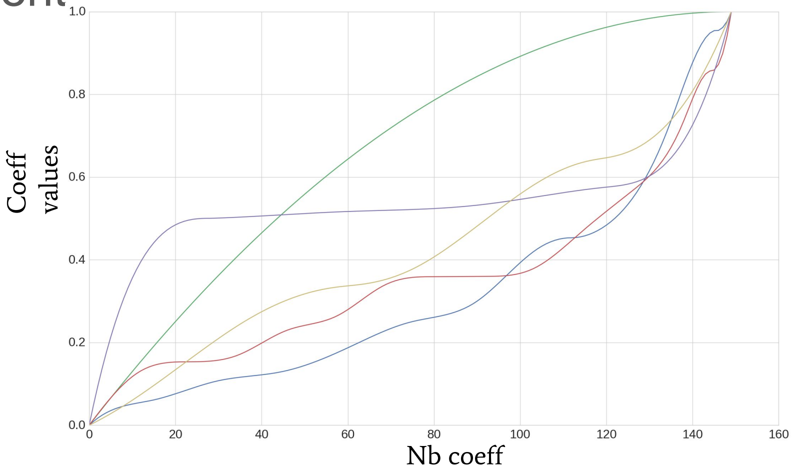
Back-up

Tracks as functions

Functional data analysis

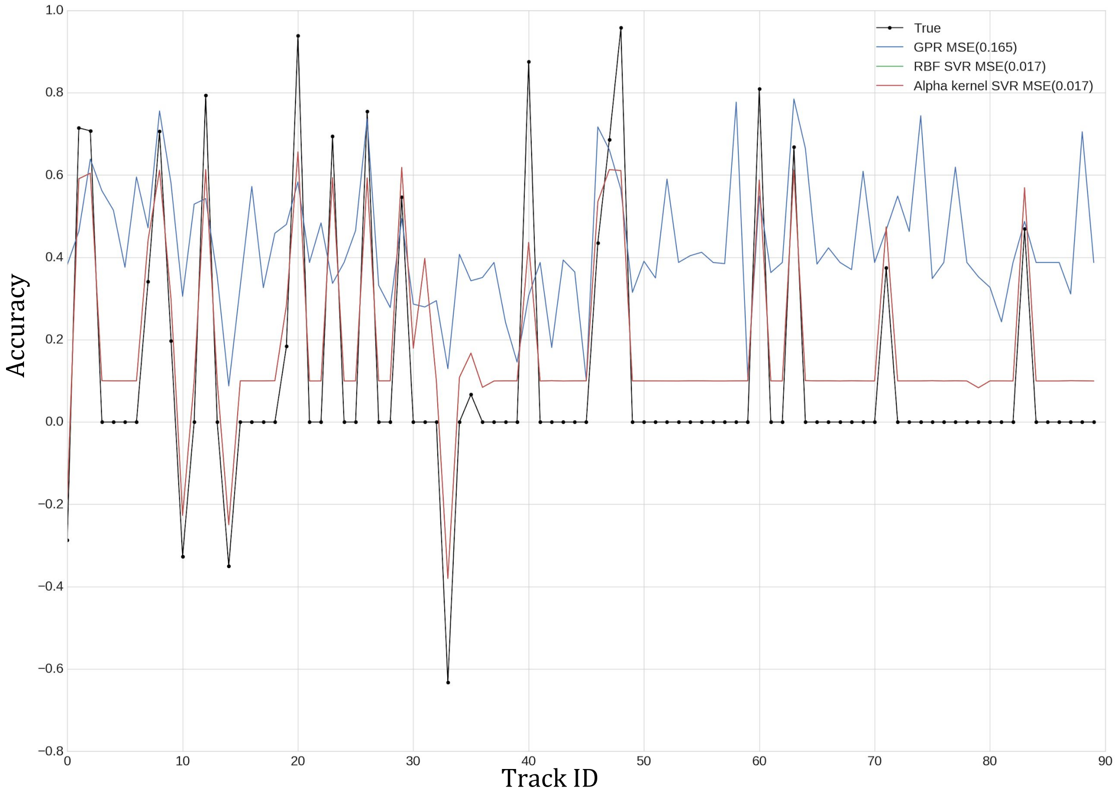
- Core idea: Turn sequence of individual observations (hits) into a continuum (functions).
- Study of the derived curves shape and homogenous representation of different size tracks.
- Algorithm :
    - Splines fitting on 7D features
    - Interpolation to extract coefficients
    - Use coefficient to learn information





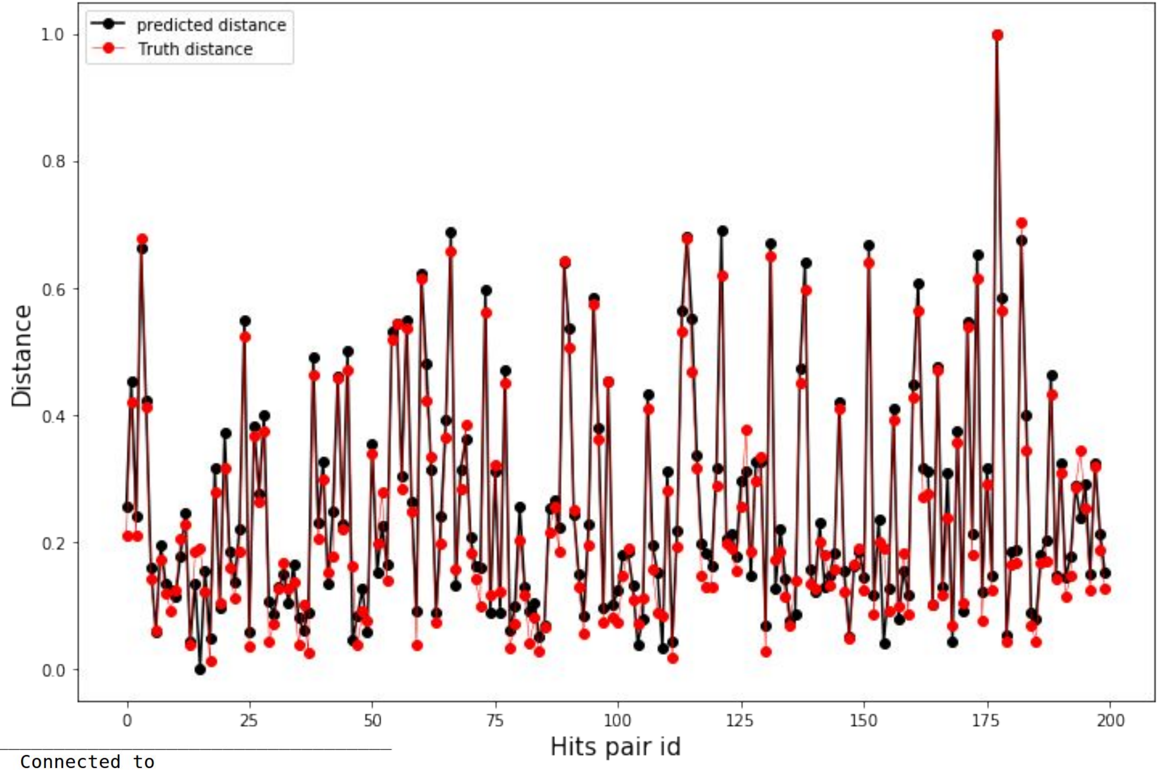
Predict track quality from coefficients

- Train a regression model on good/bad tracks with their cost as target.
- Good prediction reflects FD value.
- The model learns the quality of its output.



Cost prediction vs truth based on coefficients

Deep learning




```
____________________________________________________________________________________________________
Layer (type)                     Output Shape          Param #     Connected to
====================================================================================================
input_9 (InputLayer)             (None, 3)             0
____________________________________________________________________________________________________
input_10 (InputLayer)            (None, 3)             0
____________________________________________________________________________________________________
merge_5 (Merge)                  (None, 6)             0           input_9[0][0]
                                                                   input_10[0][0]
____________________________________________________________________________________________________
dense_9 (Dense)                  (None, 6)             42          merge_5[0][0]
____________________________________________________________________________________________________
dense_10 (Dense)                 (None, 1)             7           dense_9[0][0]
====================================================================================================
```