The High Energy Physics Tracking Machine Learning challenge

David Rousseau (LAL) (rousseau@lal.in2p3.fr),
Cécile Germain (LAL/LRI) , Isabelle Guyon (Chalearn/LRI)

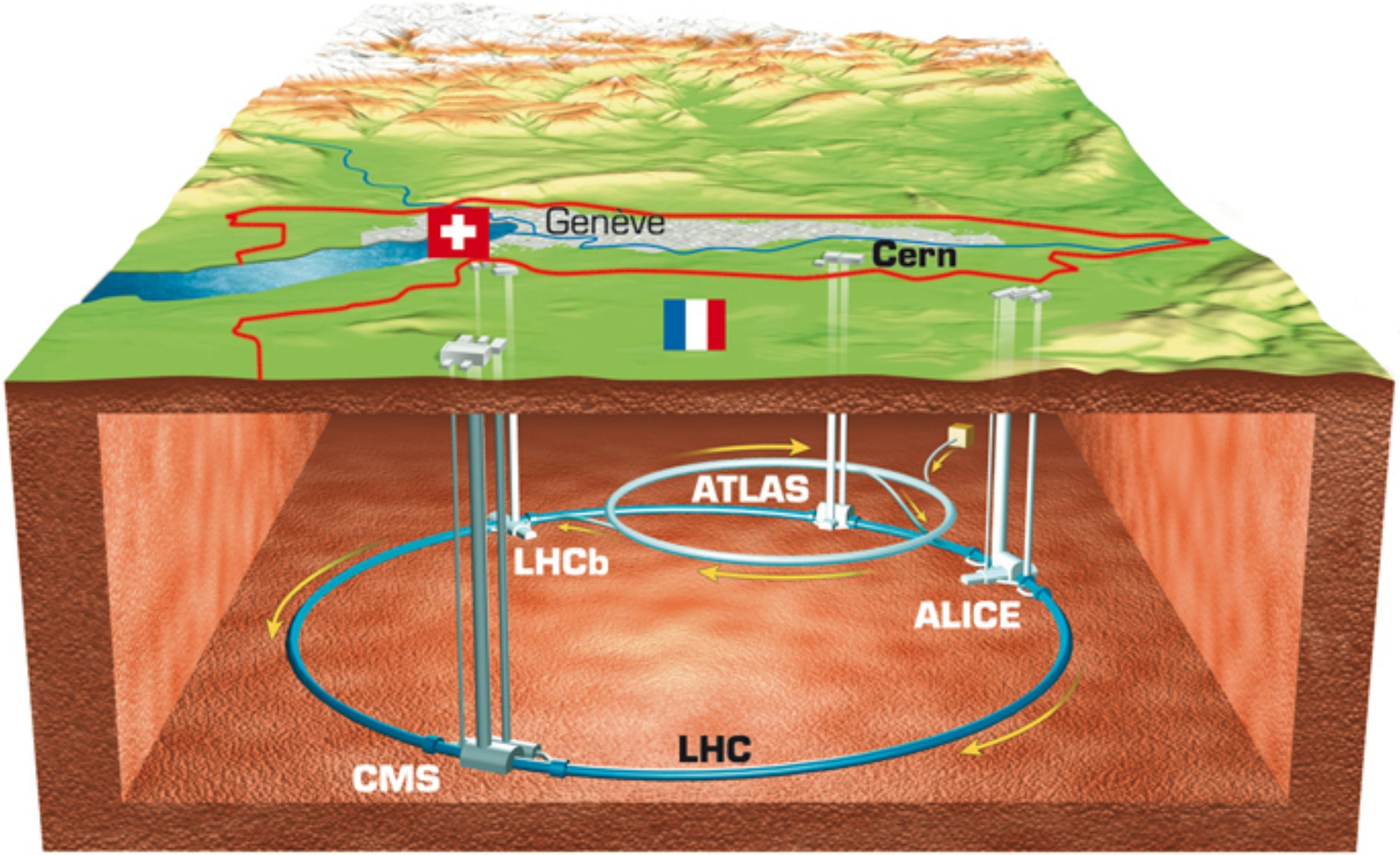
with Paolo Calafiura, David Clark (LBNL), Davide Costanzo (UCL), Armin Farbin (UTA), Tobias Golling, Moritz Kiehn, Sabrina Amrouche (U Geneva), Markus Elsing, Vincenzo Innocente, Andreas Salzburger (CERN), Peter Sadowski (UCI), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex) Jean-Roch Vlimant (CalTech)
...

CDS pitching day 9th Nov 2016

LHC purpose in a nutshell

CDS pitching day 9th Nov 2016

Le LHC



Collision de protons

$E=mc^2$

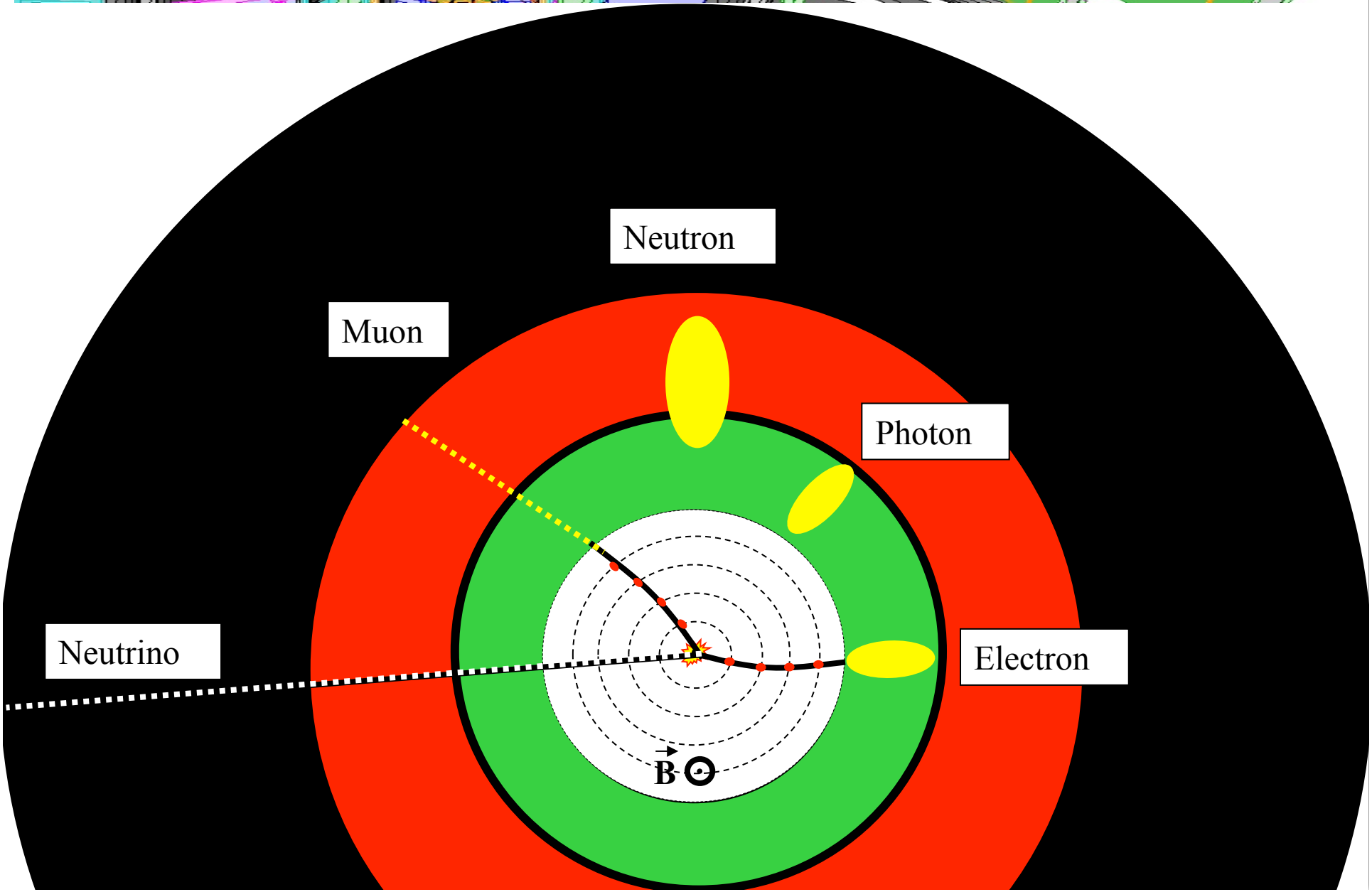
Conversion de l'énergie cinétique en masse.

Création de nouvelles particules, d'une centaine de sortes

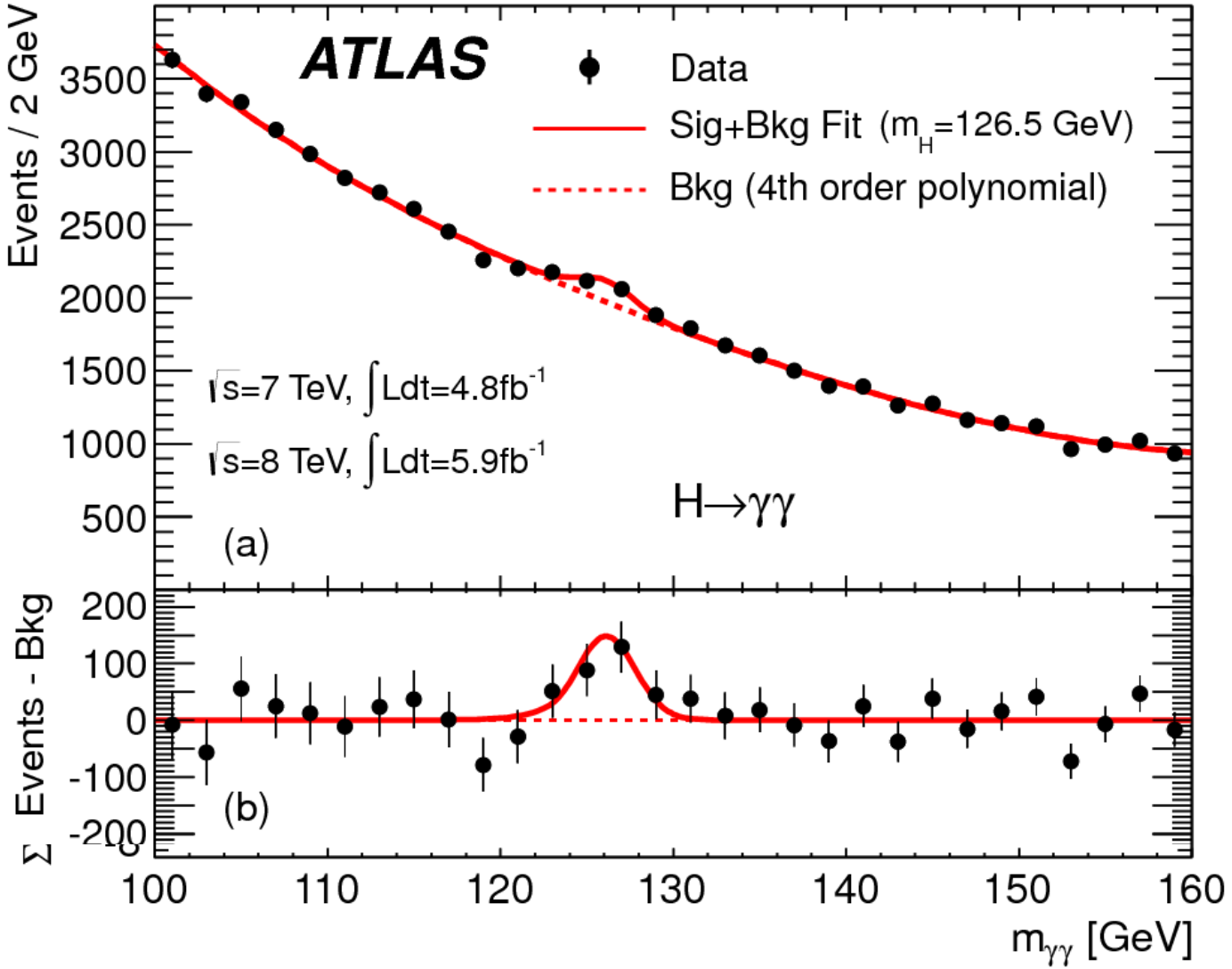
La plupart se désintègrent immédiatement
⇨Il n'en reste que de ~6 sortes,
qui vont traverser le détecteur.

4

Détection des particules



Higgs evidence



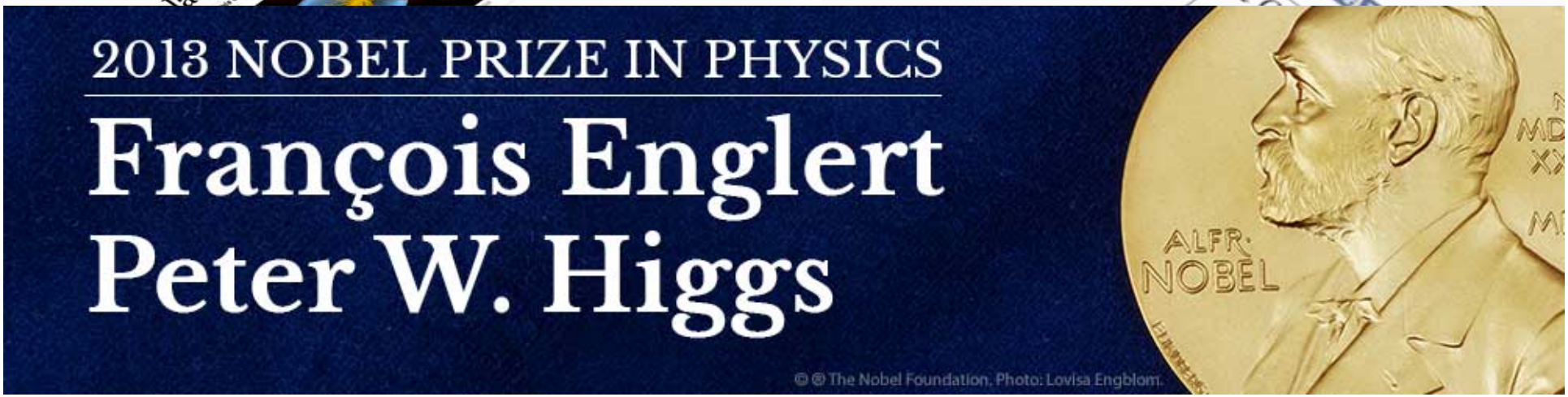
6


Le Monde
Science : la matière dévoilée


Libération
Physique des particules
La masse est dite
Le Cern a réussi à mettre en évidence le boson de Higgs qui résout une énigme fondamentale et ouvre une nouvelle étape scientifique.
Les derniers feux des pharaons
Suicides chez France Télécom : l'ancien patron mis en examen
A nos lecteurs


U.S. Edition
The New York Times
Wednesday, July 4, 2012
Pool photo by Denis Balibouse
New Particle Could Be Physics' Holy Grail


EL PAÍS
Hallada "la más sólida evidencia" de la existencia del bosón de Higgs

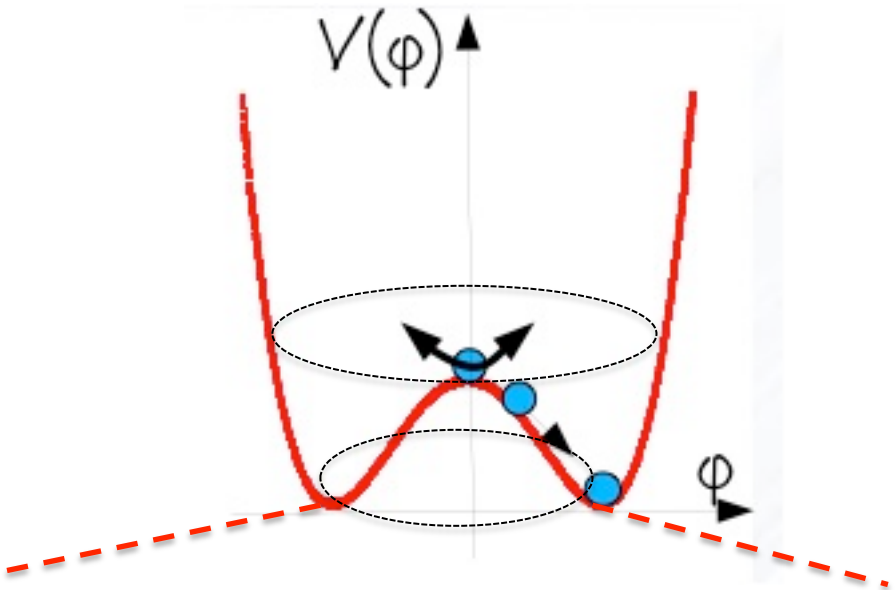

2013 NOBEL PRIZE IN PHYSICS
François Englert
Peter W. Higgs
© The Nobel Foundation. Photo: Lovisa Engblom.
ALFR. NOBEL

Future of LHC beyond Higgs boson discovery

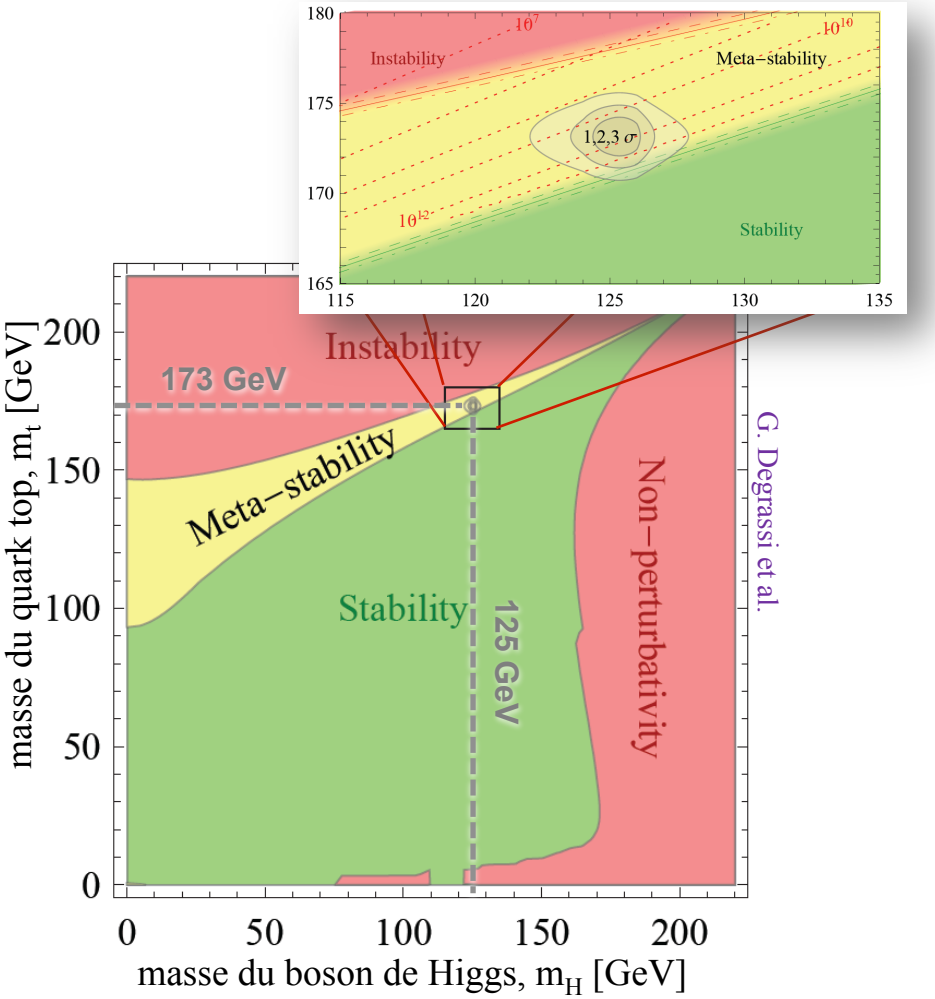
CDS pitching day 9th Nov 2016

L'Univers est-il stable ?

La stabilité du vide dépend des masses du boson de Higgs et du quark top



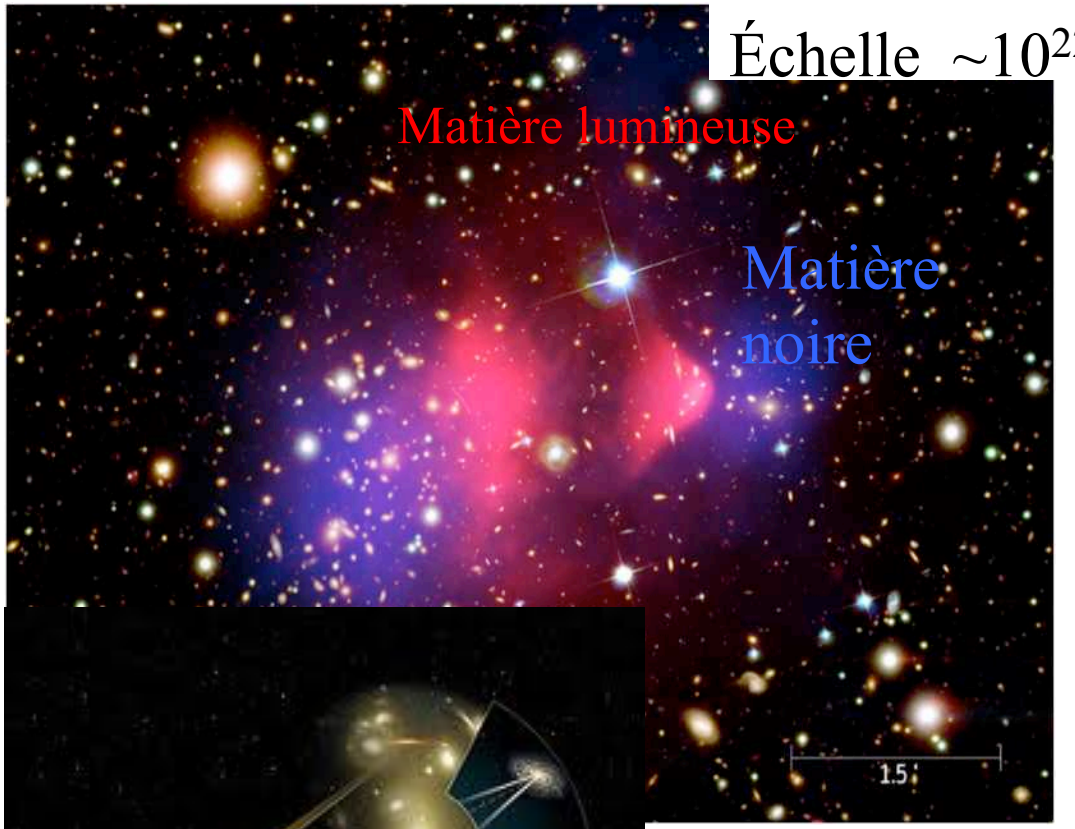
Notre Univers vit au bord du pre



“Physique des deux infinis

Les deux infinis

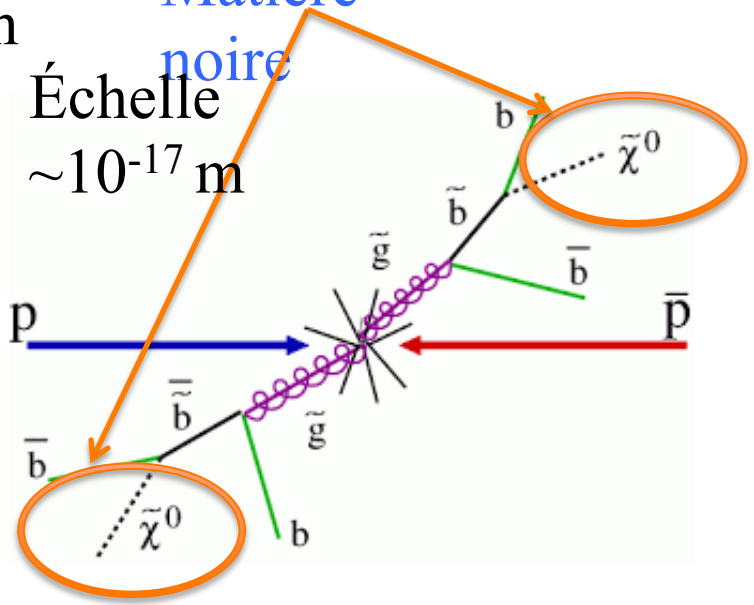
Échelle $\sim 10^{22}$ m



Lentille gravitationnelle

Matière noire

Échelle $\sim 10^{-17}$ m



10

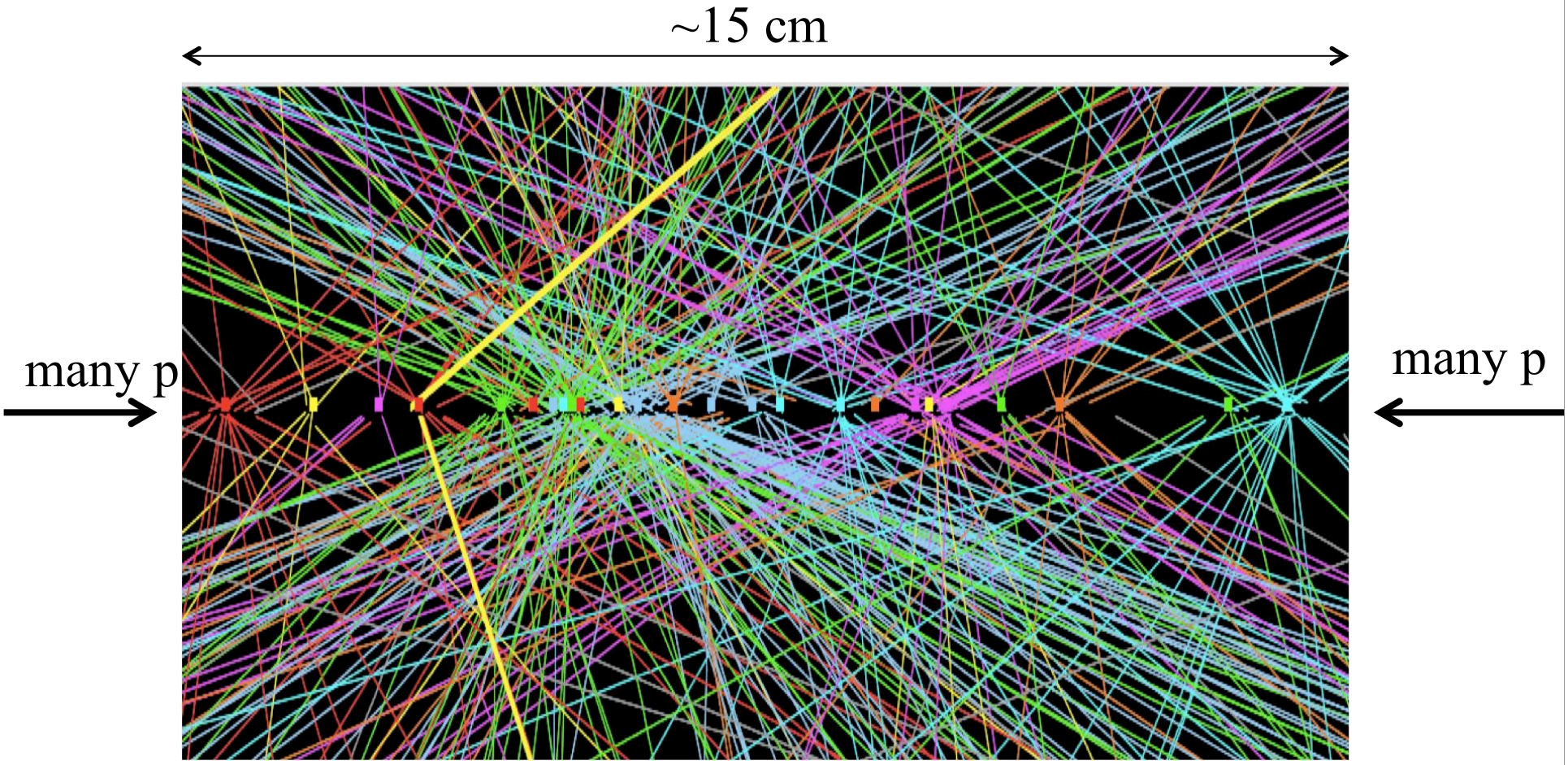
HighLumi-LHC

- Physics case : precision measurements and searches for new particles
- Wide scientific and political consensus that the community should get the most of the LHC
- Energy can’t be increased
- ➔increase the “luminosity” == number of proton collisions
- Number of proton bunches can’t be increased
- ➔increase the number of proton collision in one proton bunch collision (one “event”)➔parasitic collisions
- Progressive increase from now to 2025

Tracking challenge

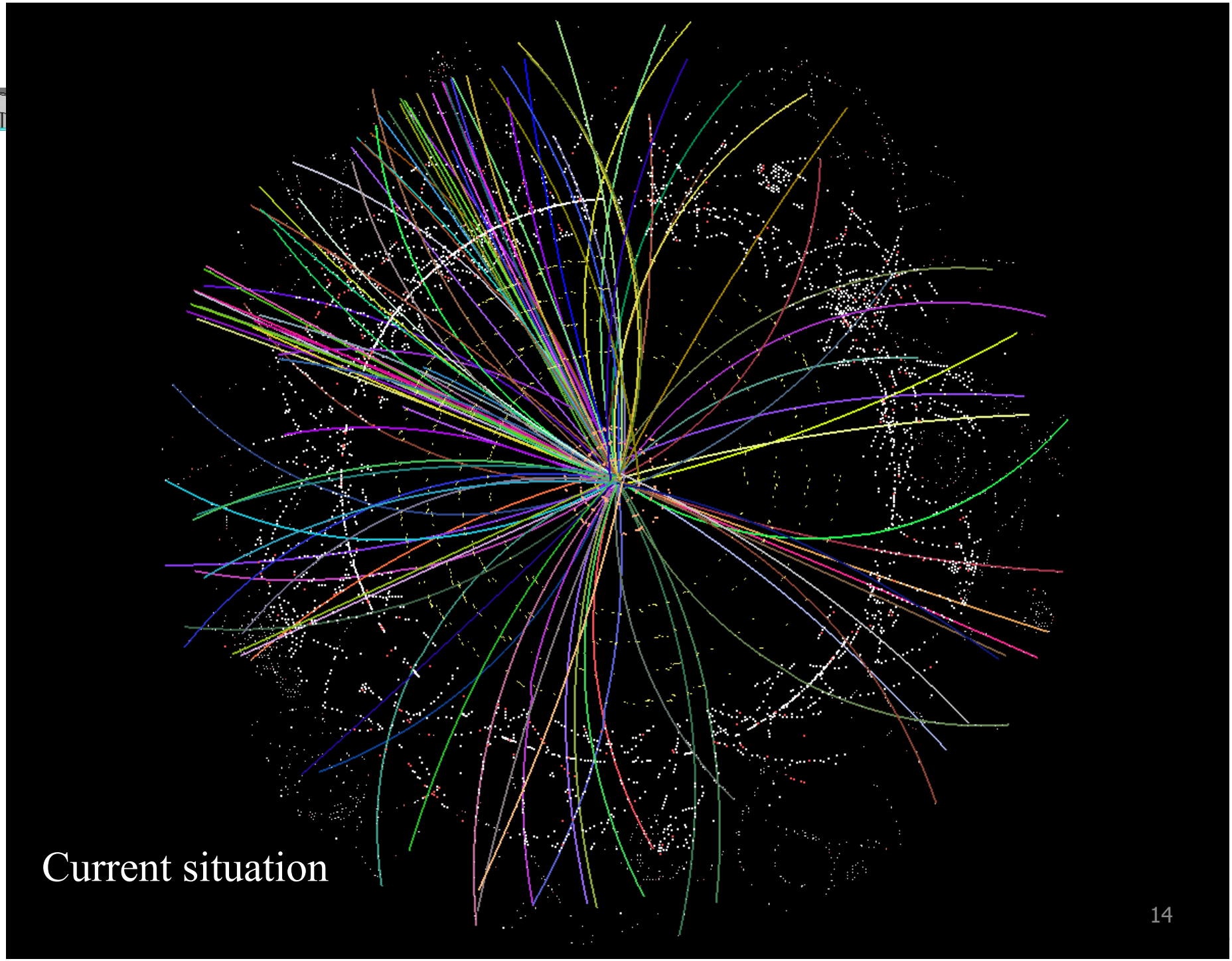
CDS pitching day 9th Nov 2016

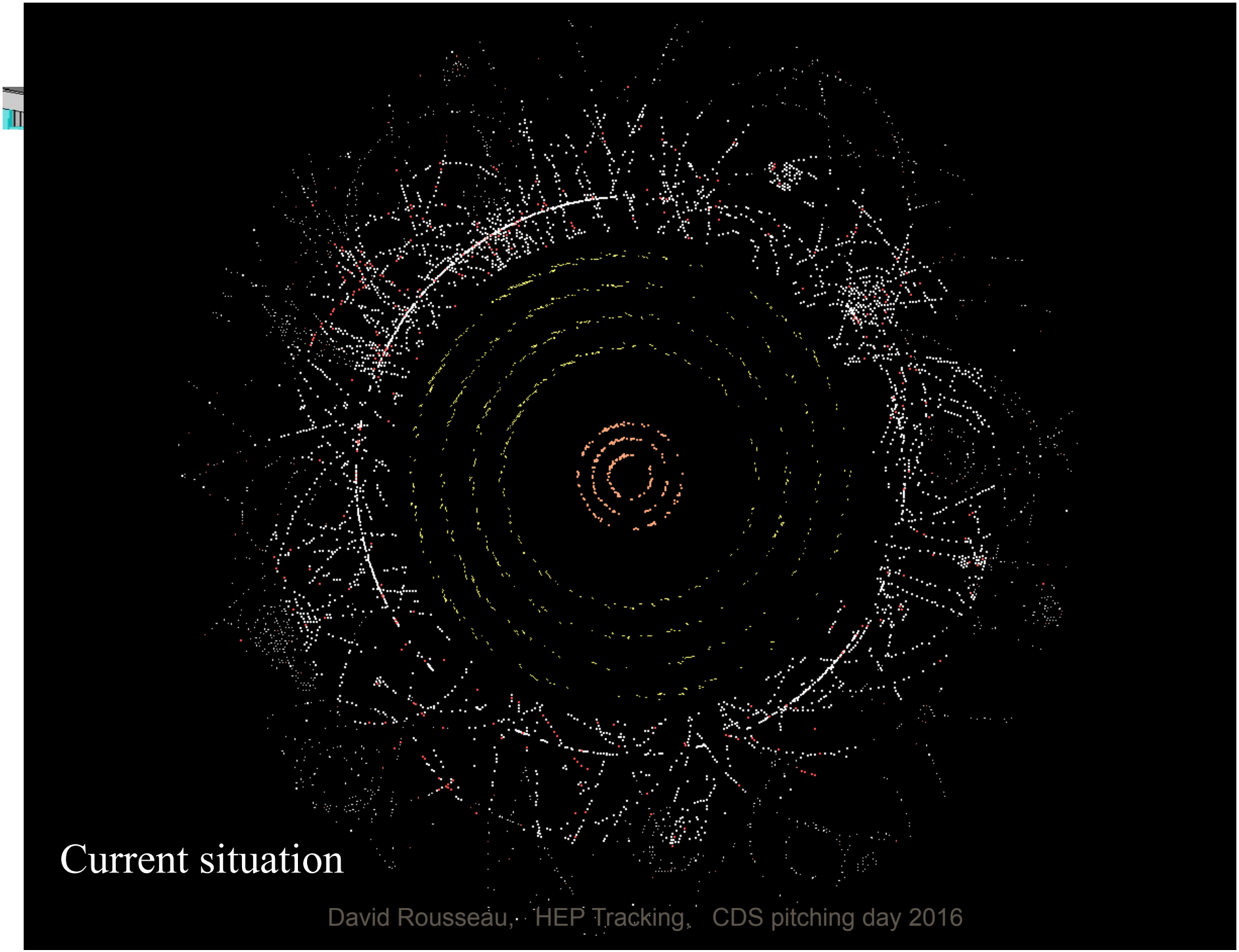
Bunch collision



Situation actuelle : 20aine de collision parasites
HL-LHC : facteur 10

13

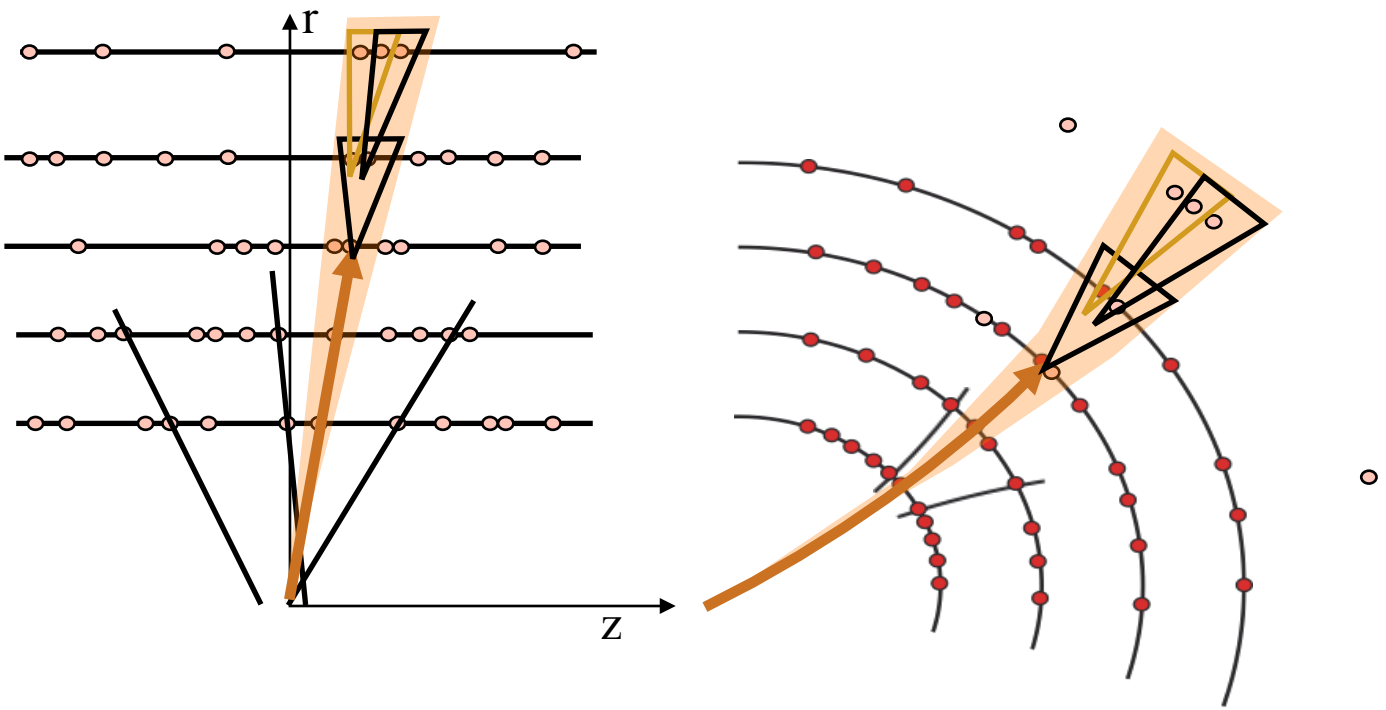

Current situation



Current situation

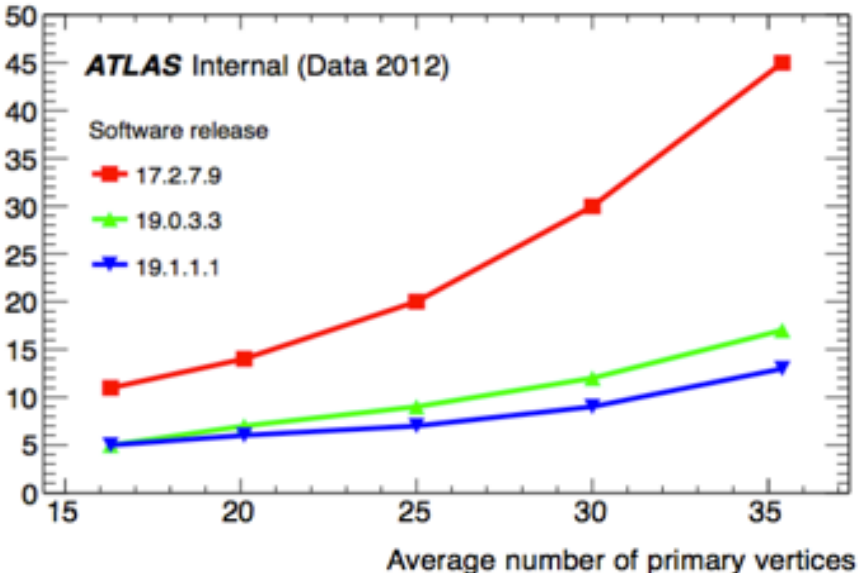
Curent technique

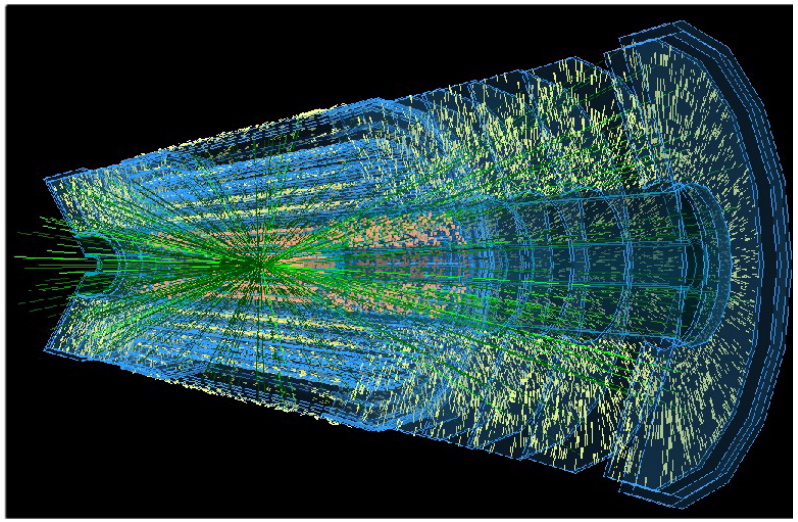
- Pattern : connect 3D points into tracks
- Essentially combinatorial approach
- Tracks are (not perfect) helices pointing (approximately) to the origin
- Challenge : explore completely new approaches
- (not part of the challenge : given the points, estimate the track parameters)

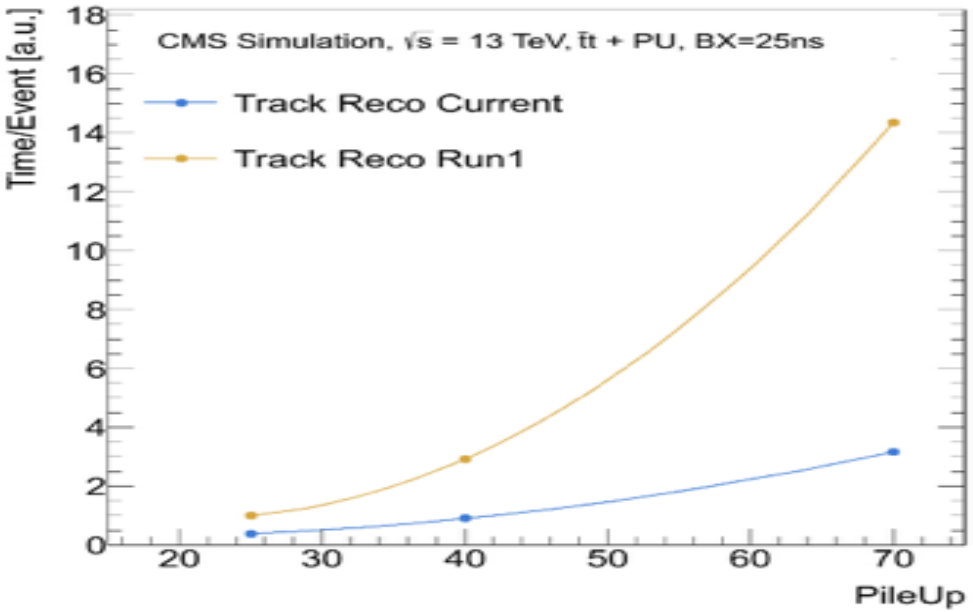


Motivation 1

- Tracking (in particular pattern recognition) dominates reconstruction CPU time at LHC
- HighLumi-LHC perspective : increased rate of parasitic collisions
  - Run 1 (2010-2012): <>~20
  - Run 2 (2015-2018): <>~30
  - Phase 2 (2025): <>~150
- CPU time of current software quadratic/ exponential extrapolation (difficult to quote any number)
- (but current software give reasonably good results)





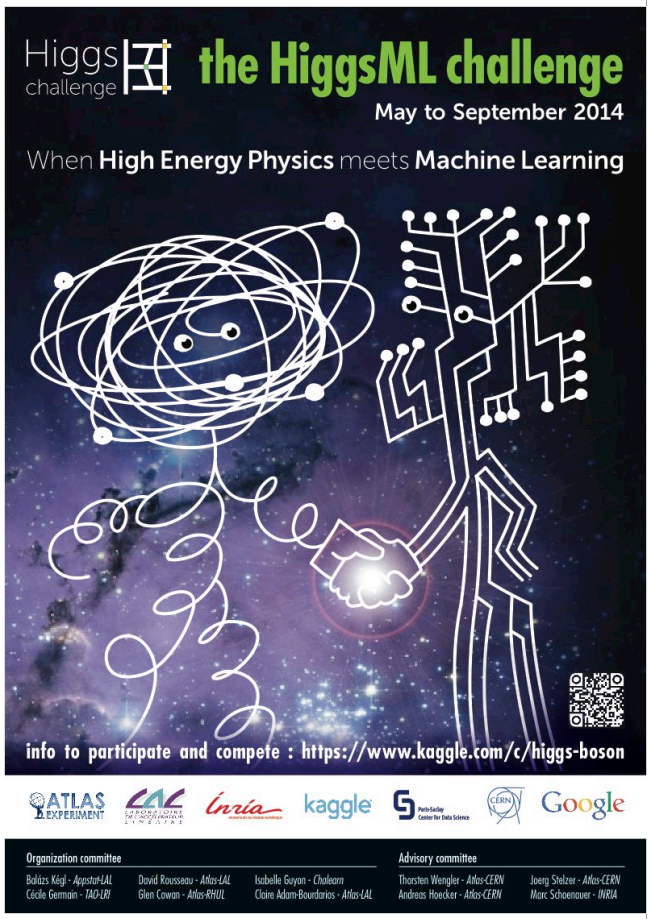


Motivation 2

- LHC experiments future computing budget flat (at best) (LHC experiments use 300.000 CPU cores worldwide)
- Installed CPU power per $==€==CHF expected increase factor ~10 in 10 years
- Experiments plan on increase of data taking rate ~10 as well (~1kHz to 10kHz)
- ➔HighLumi reconstruction to be as fast as current reconstruction despite factor 10 in complexity
- ➔requires very significant software CPU improvement, factor 10-100
- Even today (Nov 2016) we're not sure how we'll do in 2017
- Large effort within HEP to optimise software and tackle micro and macro parallelism, likely not enough
- >20 years of LHC tracking development. Everything has been tried!
  - Maybe yes, but maybe algorithm slower at low lumi but with a better scaling have been dismissed ?
  - Maybe no, brand new ideas from ML (i.e. Convolutional NN)
- Need to engage a wide community to tackle this problem

Higgs Machine learning challenge

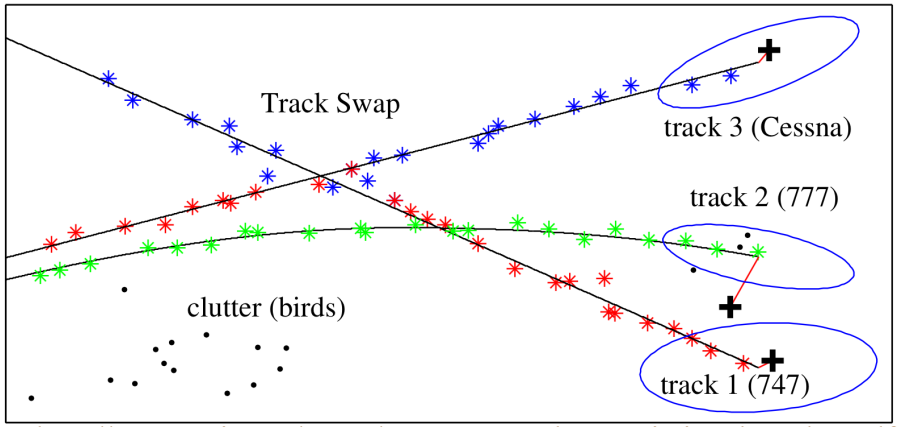
- The team builds on the HiggsML experience funded by CDS1
- An ATLAS Higgs signal vs background classification problem, optimising statistical significance
- Ran in summer 2014
- 2000 participants (largest on Kaggle at that time)
- Outcome
  - Best significance 20% than with TMVA
  - BDT algorithm of choice in this case where number variables and number of training events limited (NN very slightly better but much more difficult to tune)
  - XGBoost best BDT on the market (quite wide spread nowadays)
  - Wealth of ideas, documented in JMLR proceedings v42
  - Still working on what works in real life what does not
  - Raised awareness about ML in HEP



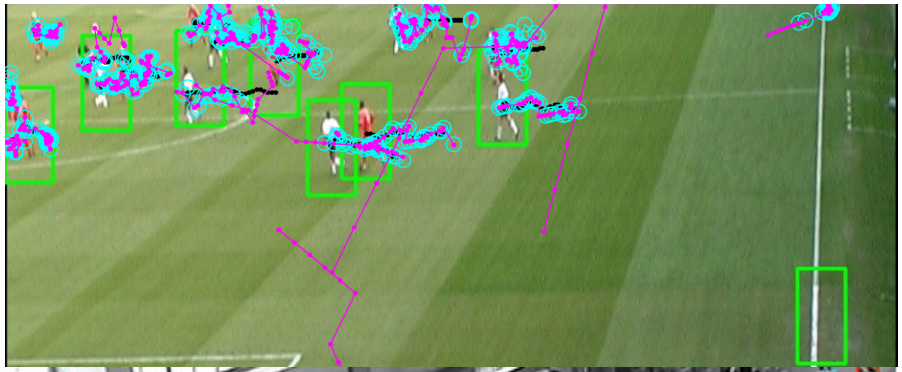
- ➔many HEP-ML projects kick-started in the last 1-2 years

Pattern recognition

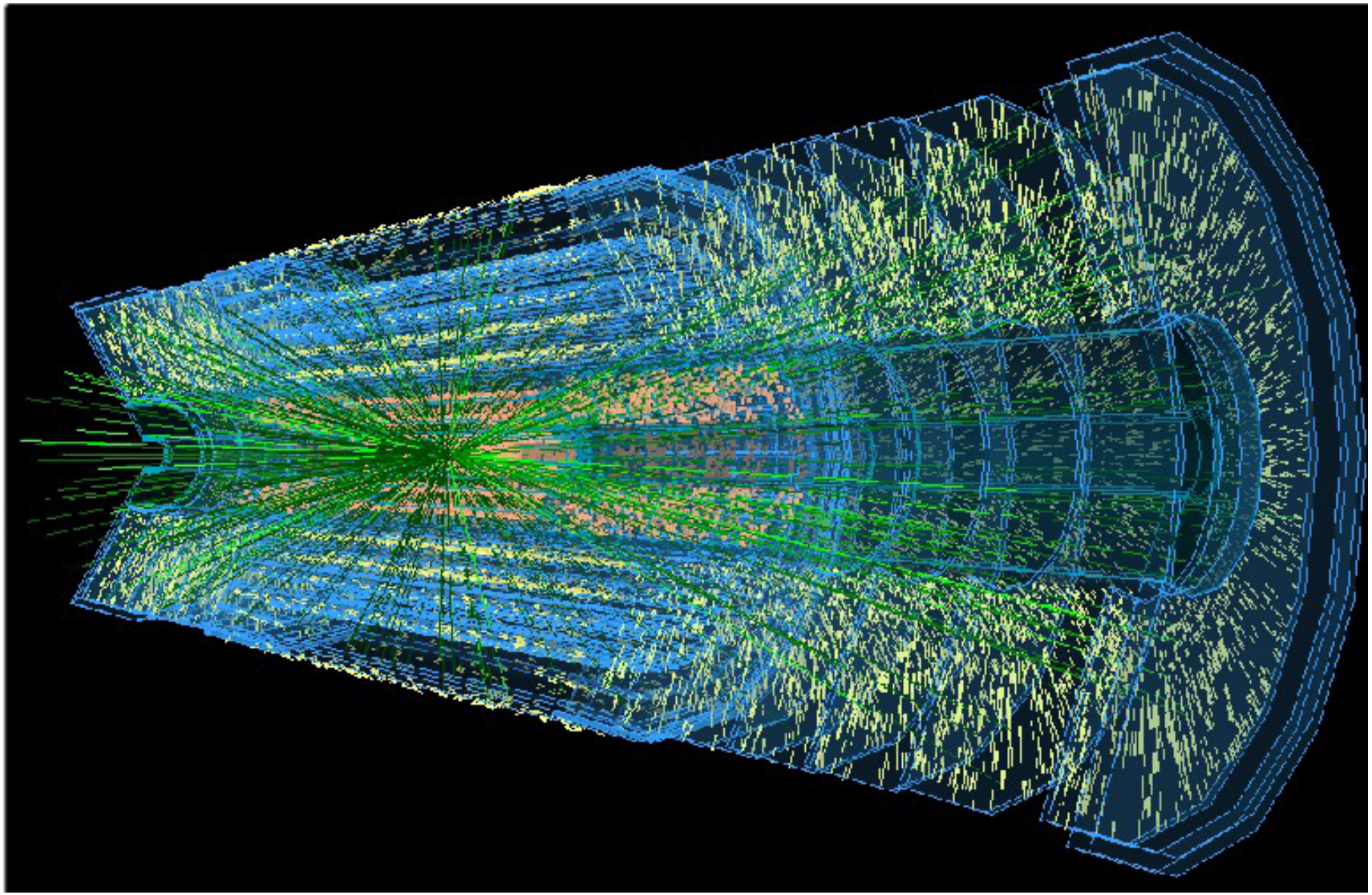
- HiggsML was "just" a classification problem, with many on-the-shelf algorithms
- Pattern recognition, tracking, is a very old, very hot topic in Artificial Intelligence : examples➔



http://papers.nips.cc/paper/5572-a-complete-variational-tracker.pdf

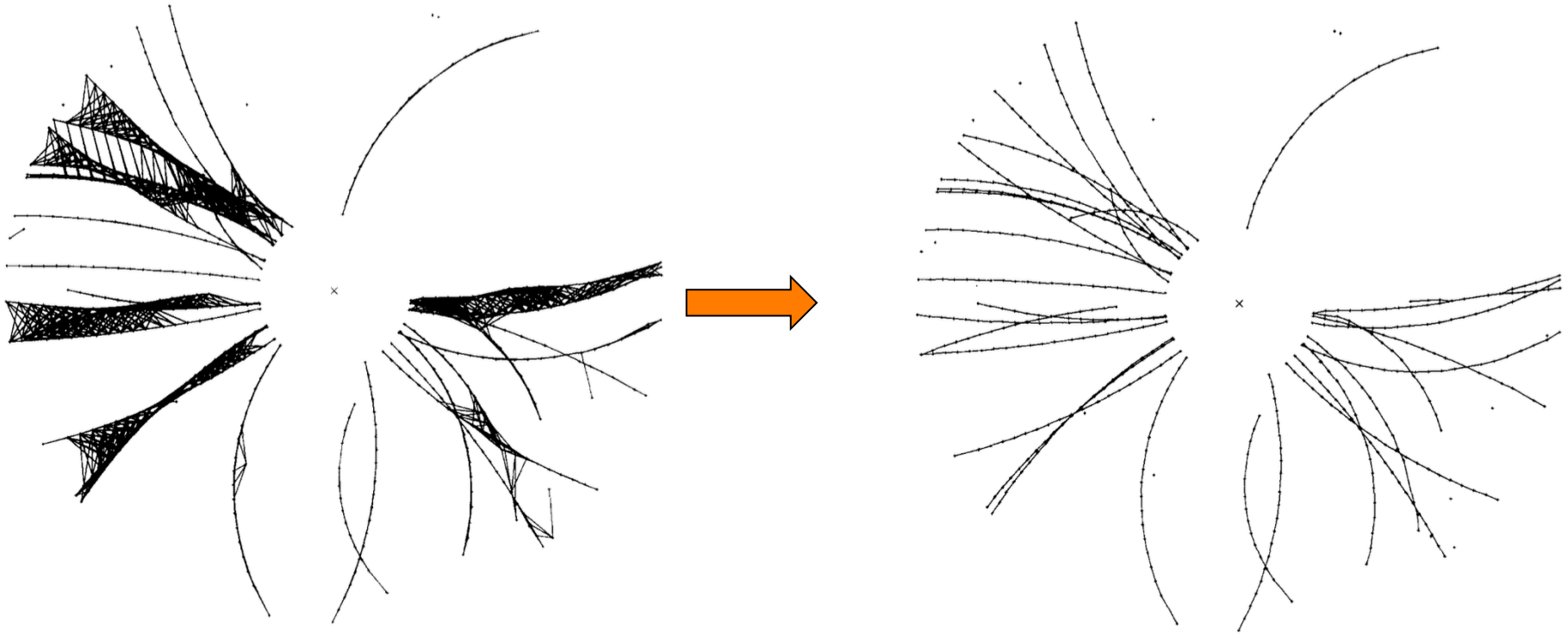


- Note that these are real-time applications, with CPU constraints
- Worry about efficiency, "track swap",...



21

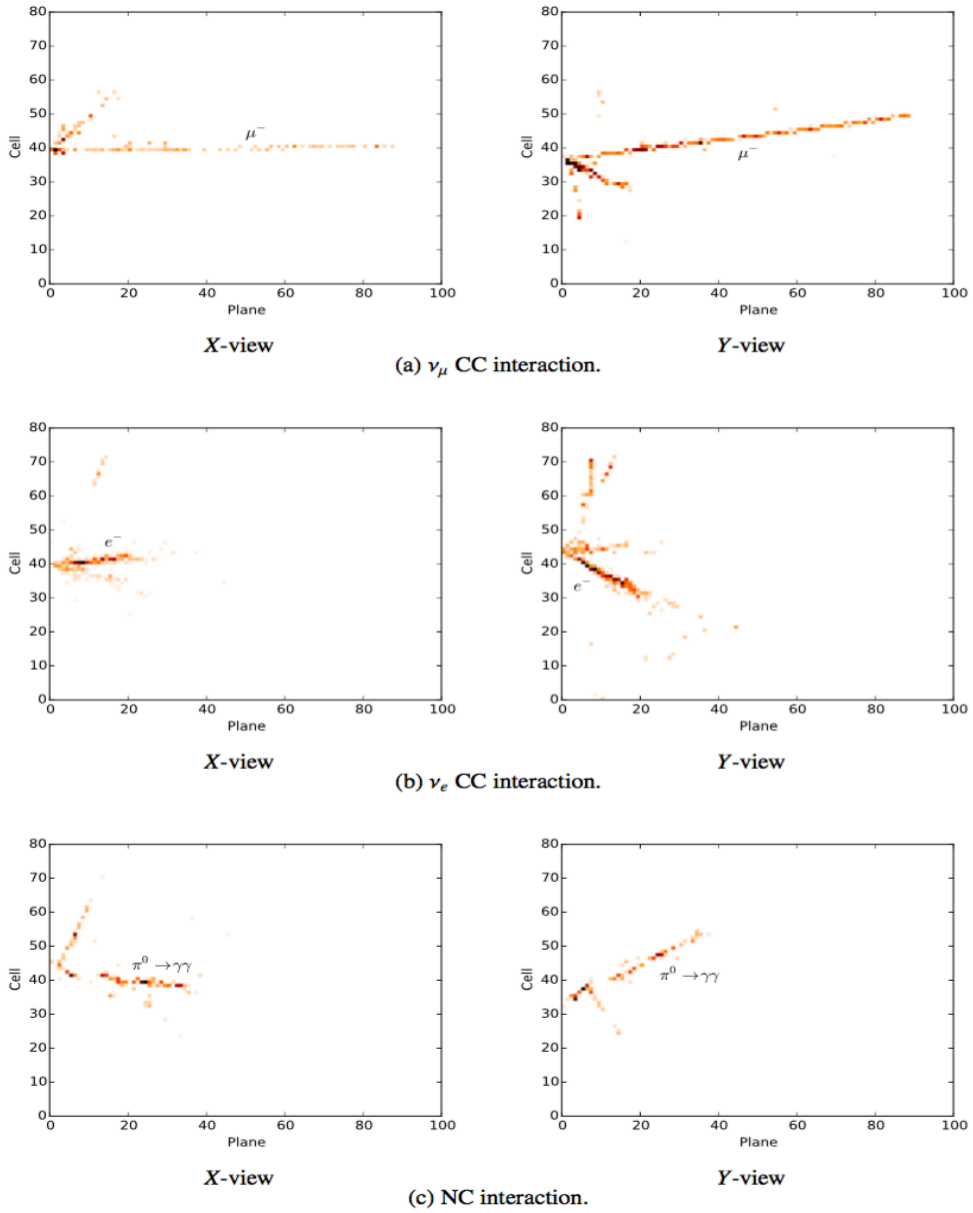
An early attempt



- Stimpfl-Abele and Garrido (1990) (ALEPH)
- All posssible neighbor connections are built, the correct ones selected by the NN
- Reasonable performance, not used for real

A recent attempt : NOVA

arXiv 1604.01444 Aurisano et al

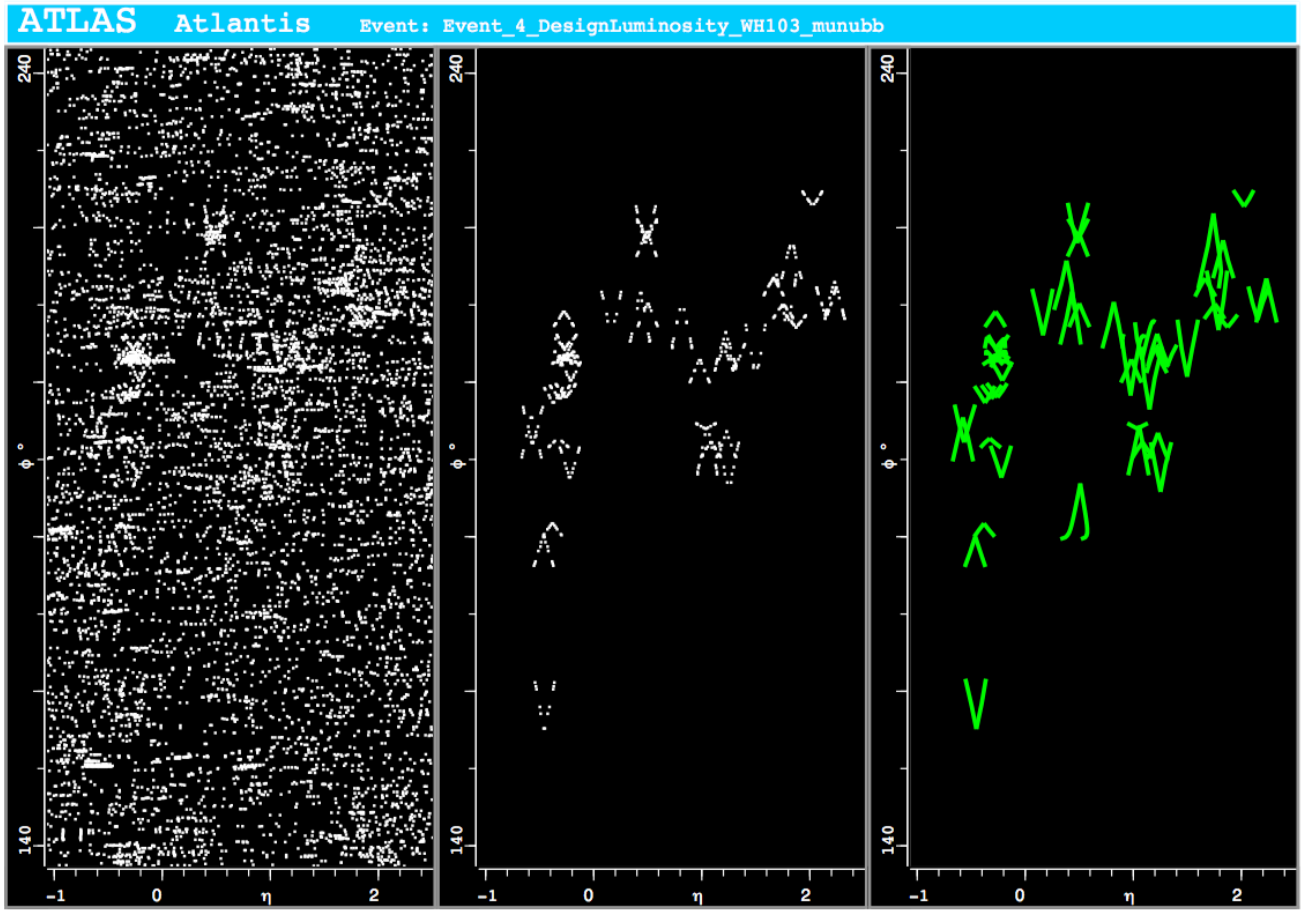


Neutrino interaction classification
Using Convolutionnal Neural Network

V plots: connection to computer vision ?

- Computer vision : try to do as well as human
- Tracking : tracks are not visible by eye!

CHEP04



- Hans Drevermann, ALEPH/DALI then ATLAS/ATLANTIS event display
- Eta phi projection with $\delta\eta = +/- \varepsilon(r_{max} - r)$

TrackML : current thinking

- First idea Feb 2015, aiming at summer 2017
- Use ACTS (A Common Tracking Software) to generate fast simulation of a generic Silicon detector at HL-LHC (cylinder and disks)
  - battlefield tested ATLAS software moved to public gitlab@cern, will be moved to github
  - ➔simplified simulation but not too simple (otherwise a simple Hough transform would probably work)
  - "cheap" but realistic events which do not "belong" to any collaboration (ATLAS, CMS,…)
- Dataset:
  - 3D points and truth track parameters for n events
  - Typical events with ~200 parasitic collisions (~10.000 tracks/event)
  - Large training sample 1 million events, 100 billion tracks ~1TeraByte
  - Also thinking of allowing participants to generate their dataset
- Participants are given the test sample. They should upload the tracks they have found
  - A track is a list of points belonging to it
  - We don't ask for track parameters, nothing will beat Kalman filter
  - Figure of merit built from efficiency, fake rate, CPU time
- Also thinking to split the algorithmic problem from the CPU optimisation problem

What with CDS 2.0?

- We're not looking really for new collaboration on preparing the challenge itself but still:
  - 2-months of an engineer (preferably from CDS core) to finalise the challenge operation
- Put more emphasis on post challenge analysis and mid/long term collaboration
  - use the CDS channels to advertise the challenge
  - master internship for post-challenge analysis
  - Build in/post-challenge collaboration with CDS scientists on innovative approaches to the tracking problem as revealed by the challenge.
  - possibly collaboration on the visualization