Big data & machine learning en physique des particules

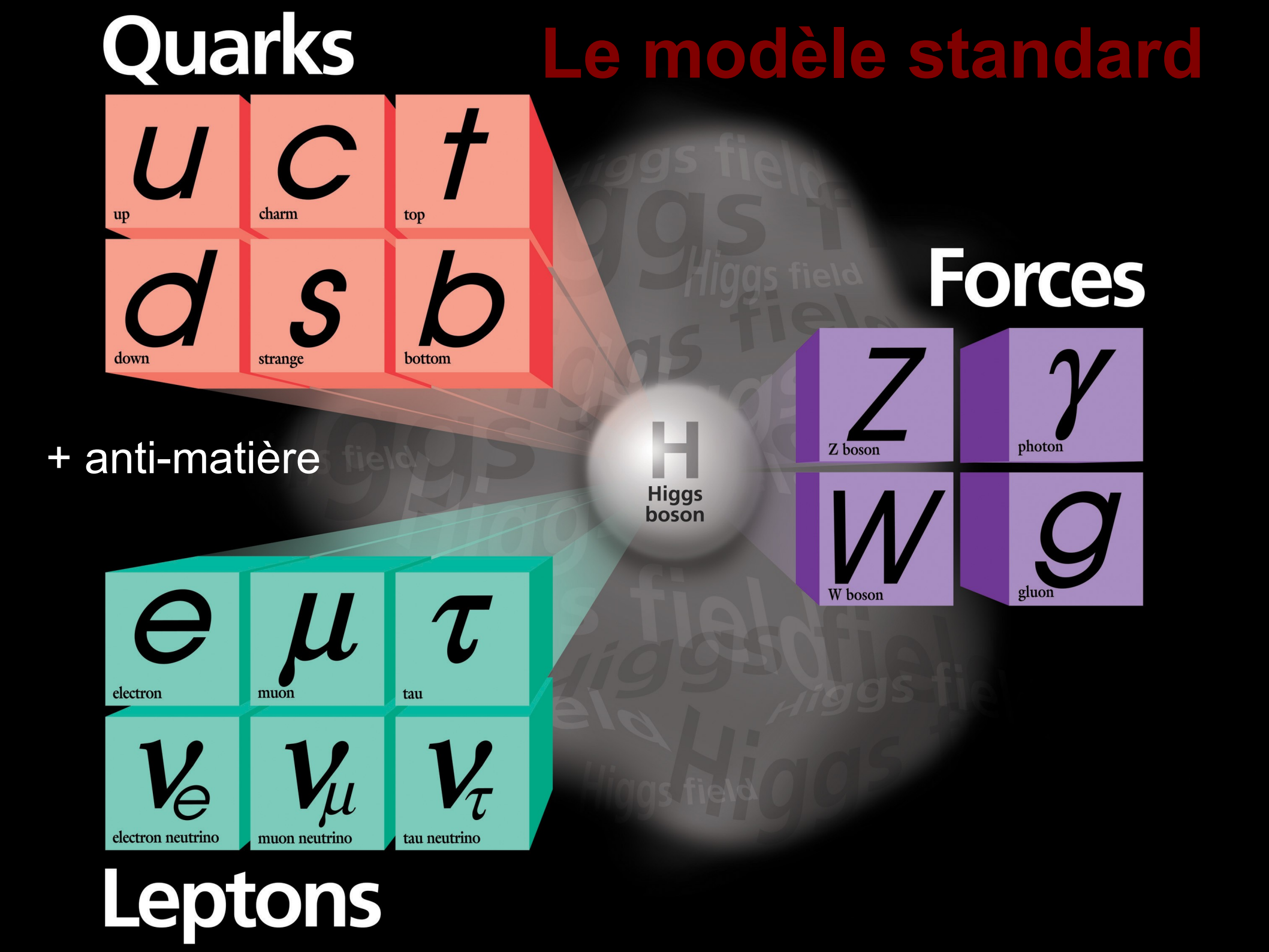
Yann Coadou

Centre de physique des particules de Marseille

Méthodes IA / Data Science pour la physique
29 janvier 2018

cnrs dépasser les frontières
IN2P3 Les deux infinis
Aix*Marseille université Initiative d'excellence
CENTRE DE PHYSIQUE DES PARTICULES DE MARSEILLE CPPM

Le modèle standard




CERN

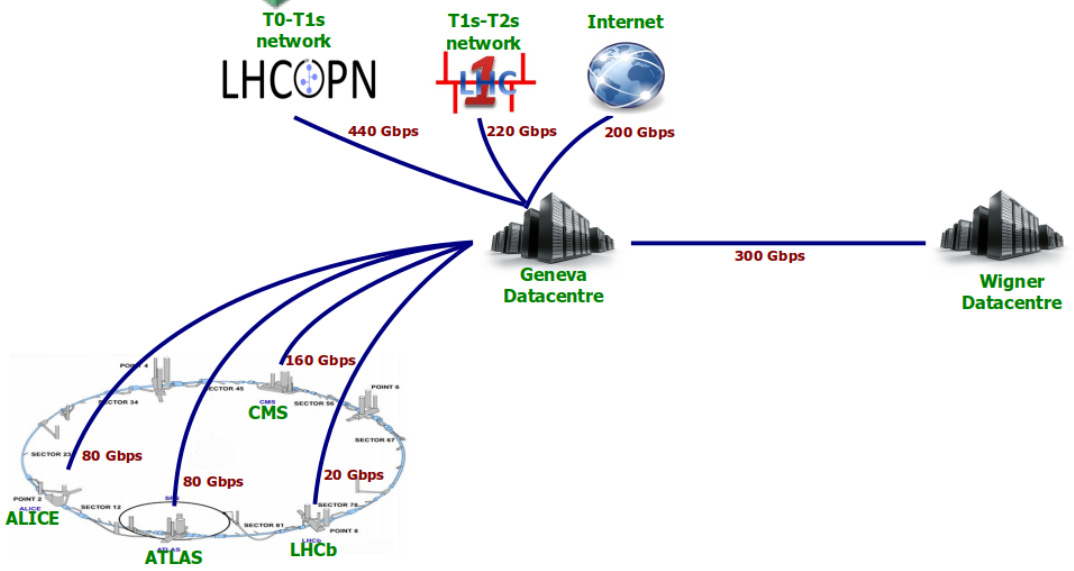
CERN

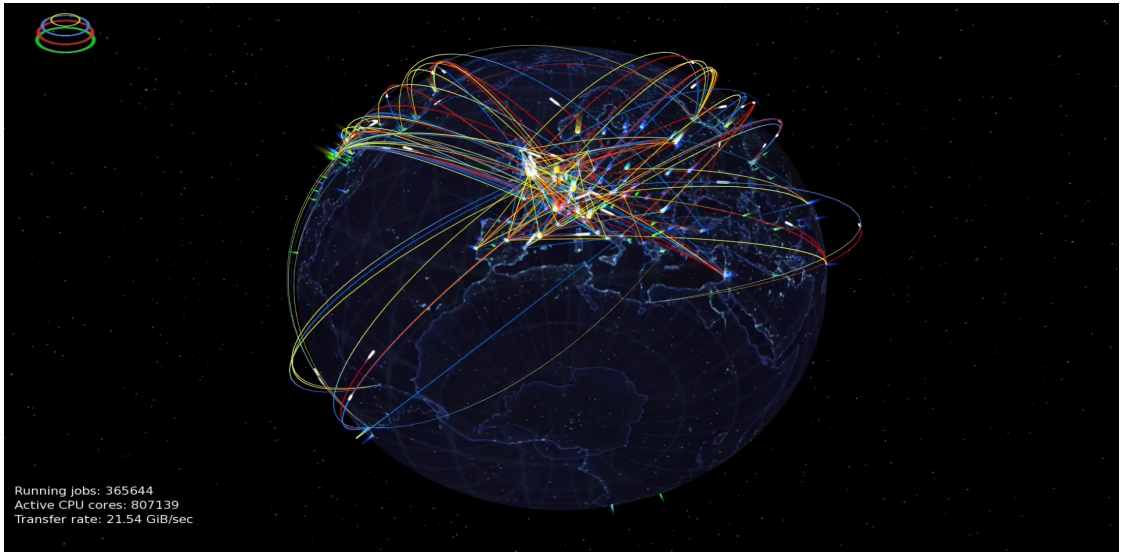
- 2500 employés
- 13 000 utilisateurs dans 500 instituts de 80 pays
- « Where the Web was born »
- CERN data centre : 230 000 cœurs sur 15 000 serveurs, 250 PB

La grille de calcul

WLCG
Worldwide LHC Computing Grid

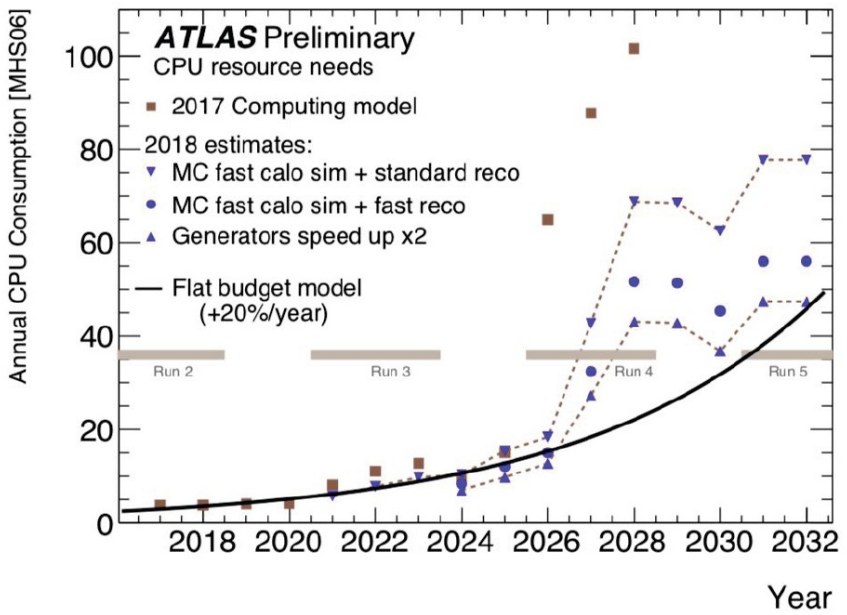
- Plus de 800 000 cœurs
- 170 sites dans 42 pays
- LHC : 50-70 petabytes/an
  CERN : +25 PB
- 2 milliards de fichiers
- > 250 000 tâches simultanées
- 2 millions de tâches/jour
- Typiquement > 2 PB accédés chaque jour
- Taux de transfert typiques 35 GB/s
- Stockage total :
  ~ exabyte !



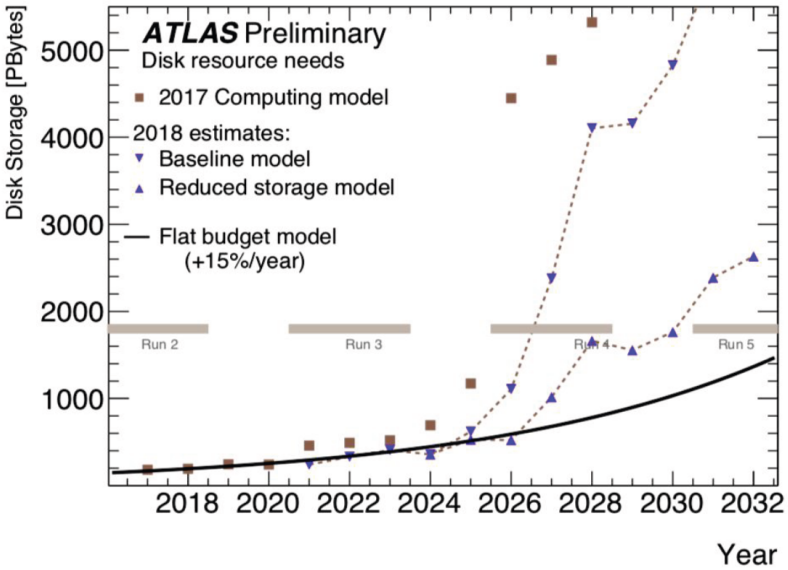


Le futur proche : HL-LHC

CPU projections for HL-LHC

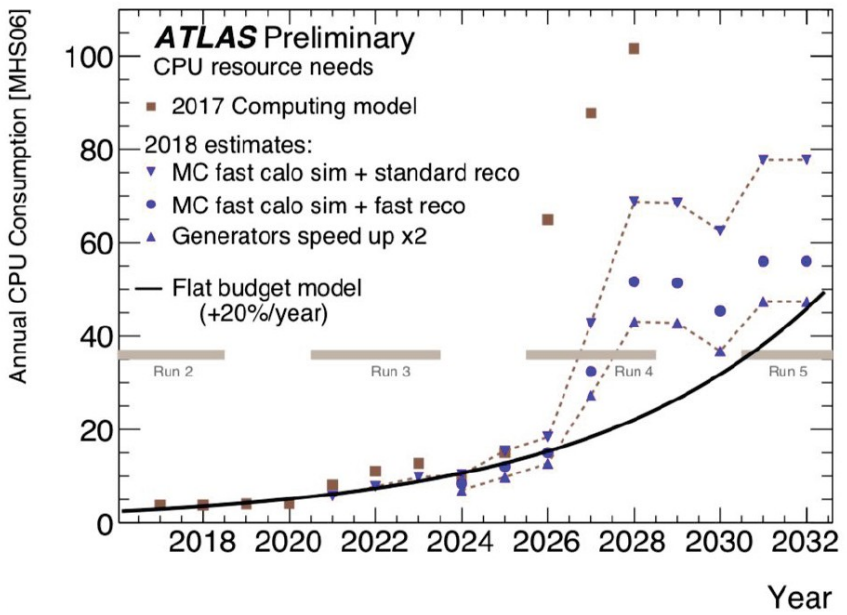


Disk storage projections for HL-LHC

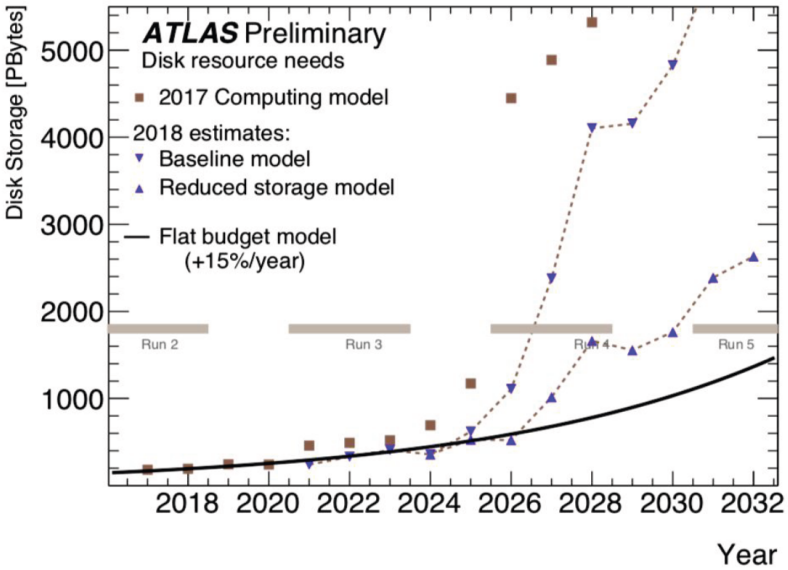


Le futur proche : HL-LHC

CPU projections for HL-LHC

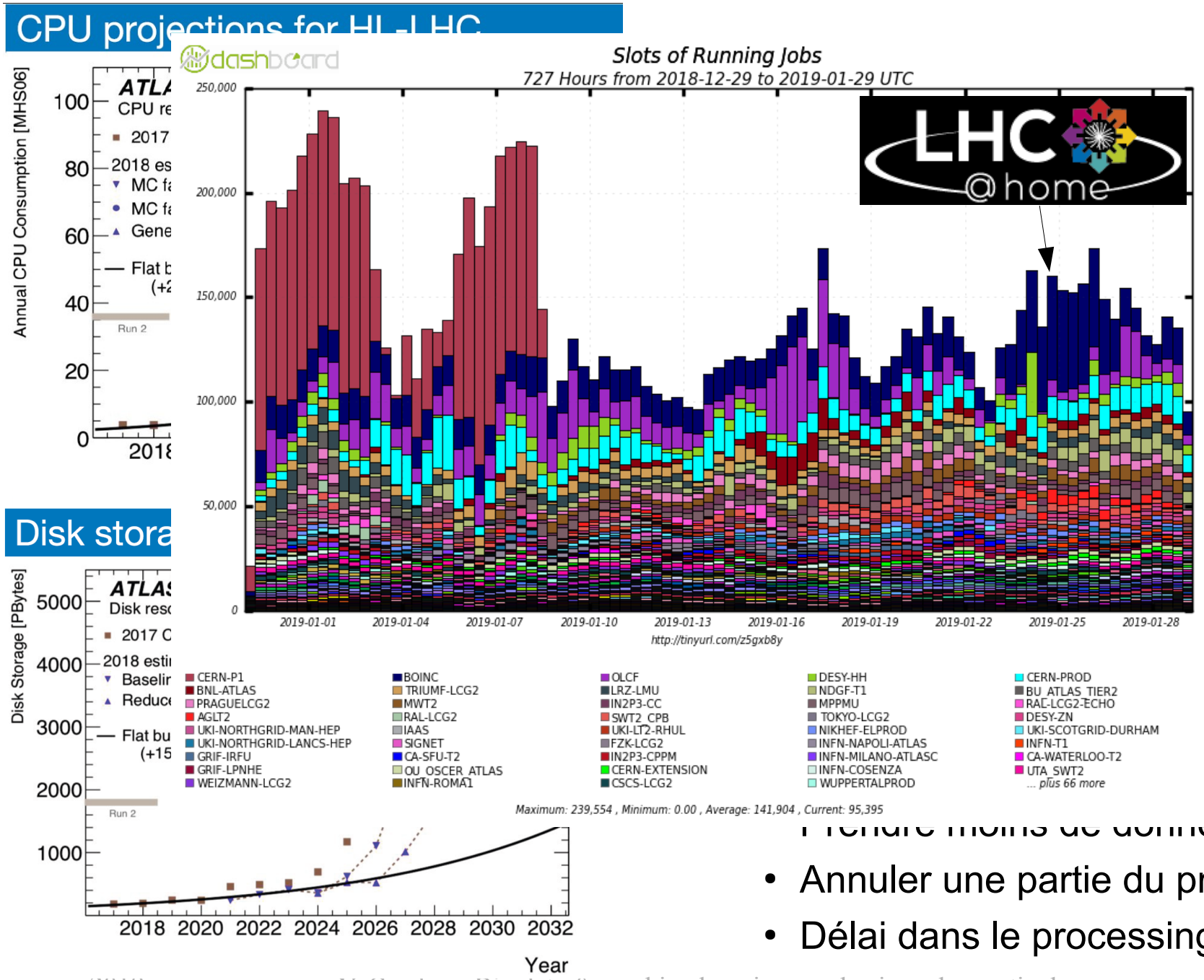


Disk storage projections for HL-LHC



- Solutions possibles
  - Techniques :
    - Machines plus performantes (GPU, FPGA, etc.)
    - Meilleur software (vectorisation, etc.)
  - Opérationnelles
    - Stocker moins d'informations
    - Éviter les « reprocessings »
  - Politique
    - Obtenir plus d'argent
    - Accéder à plus de ressources (HPC, volontaire)
  - Physique
    - Prendre moins de données
    - Annuler une partie du programme
    - Délai dans le processing

Le futur proche : HL-LHC



tes (GPU,

sation, etc.)

ons

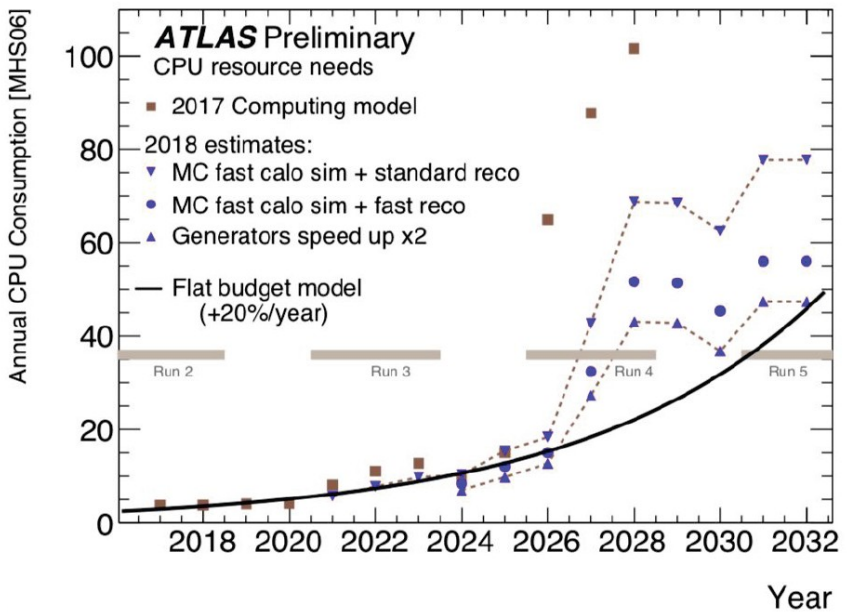
s »

rces (HPC,

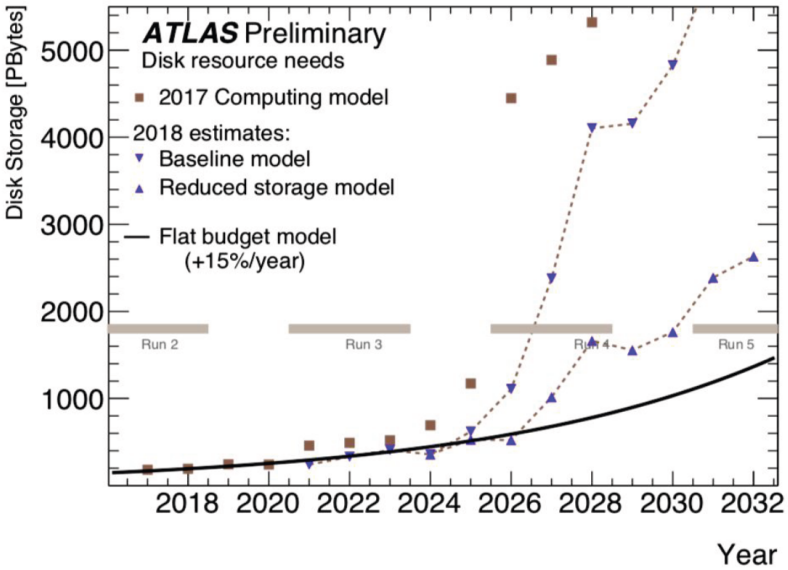
- Annuler une partie du programme
- Délai dans le processing

Le futur proche : HL-LHC

CPU projections for HL-LHC



Disk storage projections for HL-LHC



- Solutions possibles
  - Techniques :
    - Machines plus performantes (GPU, FPGA, etc.)
    - Meilleur software (vectorisation, etc.)
  - Opérationnelles
    - Stocker moins d'informations
    - Éviter les « reprocessings »
  - Politique
    - Obtenir plus d'argent
    - Accéder à plus de ressources (HPC, volontaire)
  - Physique
    - Prendre moins de données
    - Annuler une partie du programme
    - Délai dans le processing

Le LHC
(grand collisionneur de hadrons)

Le LHC
(grand collisionneur de hadrons)

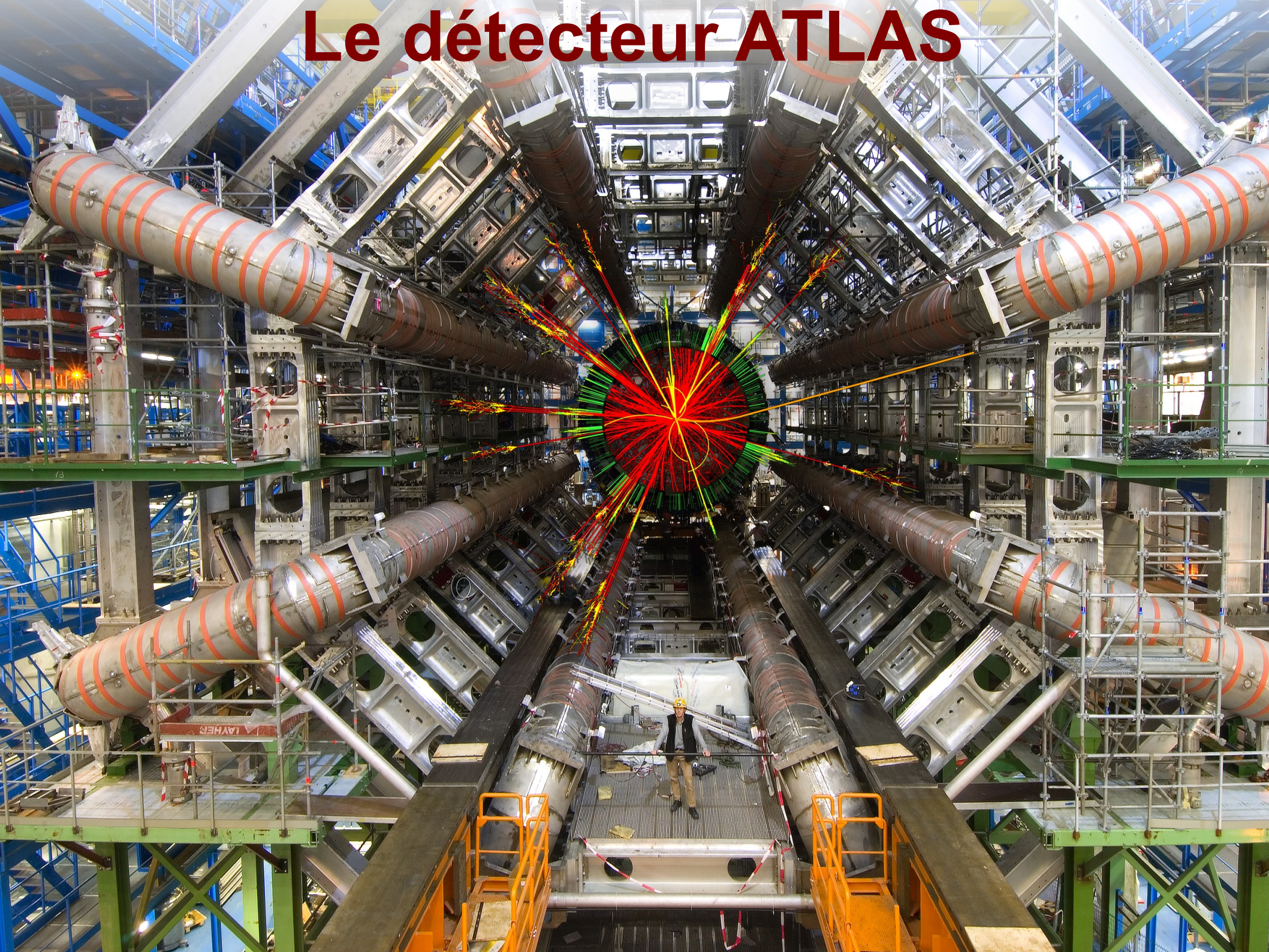
Collisions toutes les 25 ns

Le LHC
(grand collisionneur de hadrons)

Le LHC
(grand collisionneur de hadrons)

Le LHC
(grand collisionneur de hadrons)

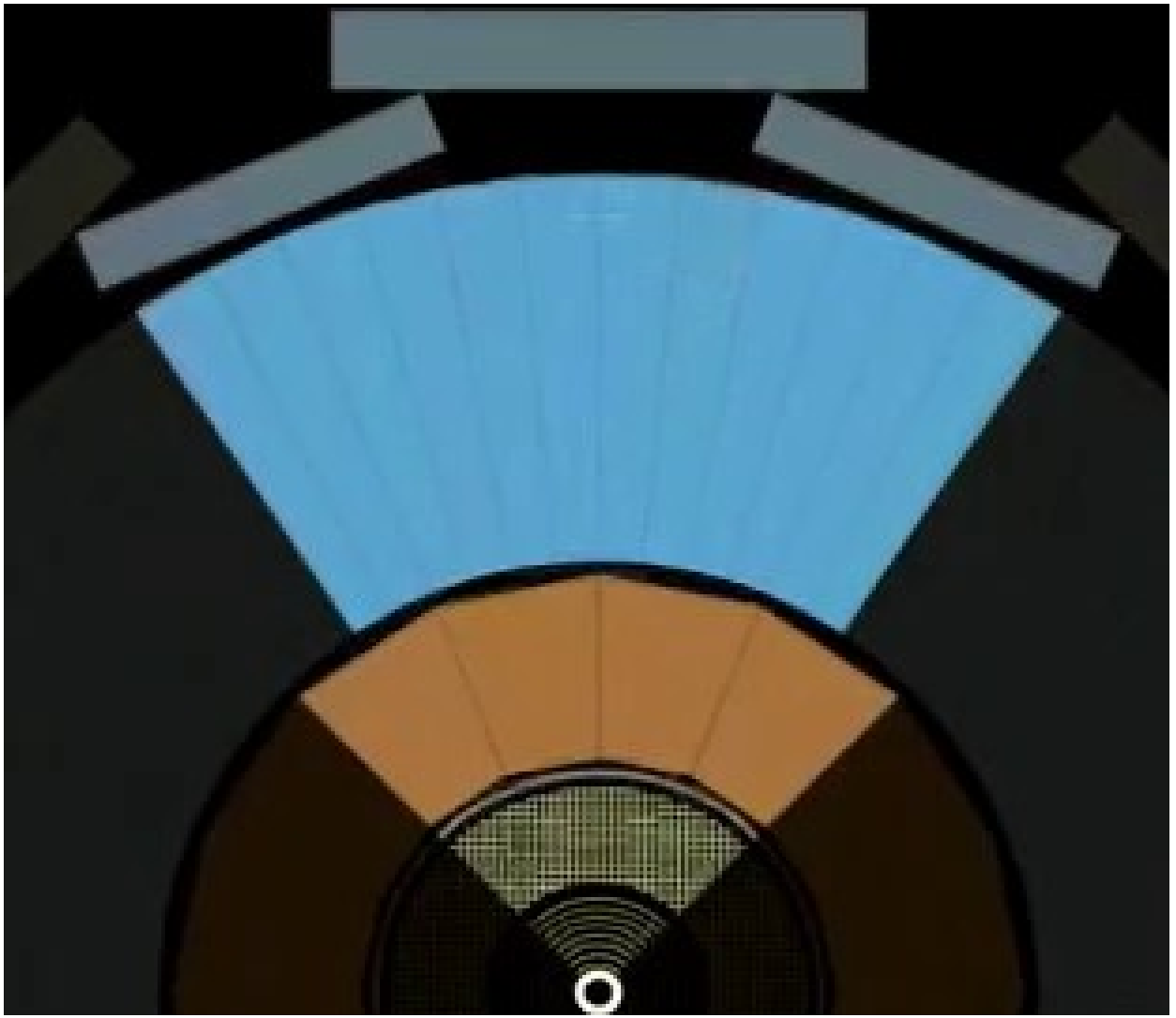
Le détecteur ATLAS



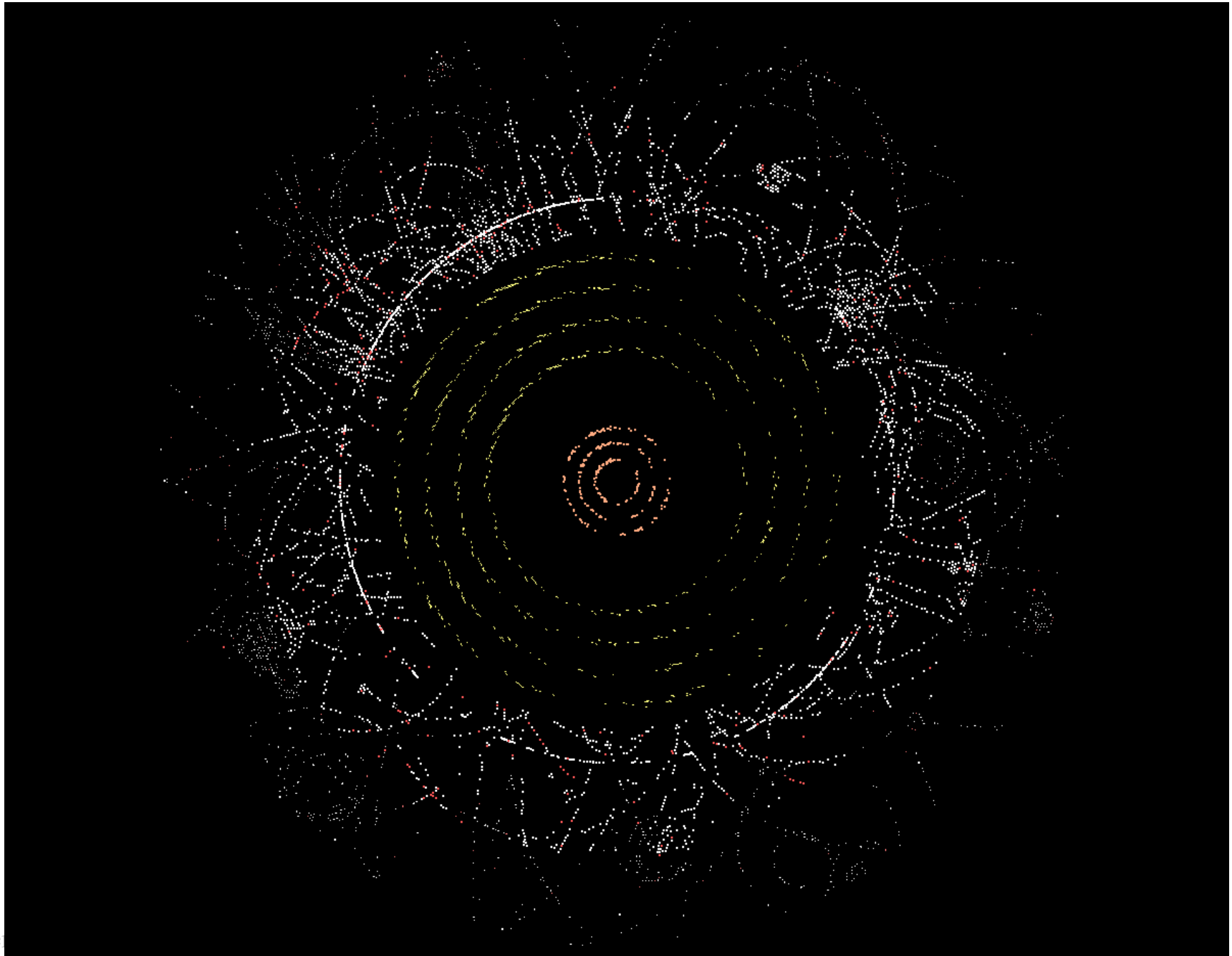
Le détecteur ATLAS

- Typiquement 35 collisions simultanées
  → > 1 milliard de collisions par seconde
- ~ 100 millions de canaux de mesure
- Rejet de 99,998 % des événements « en ligne »

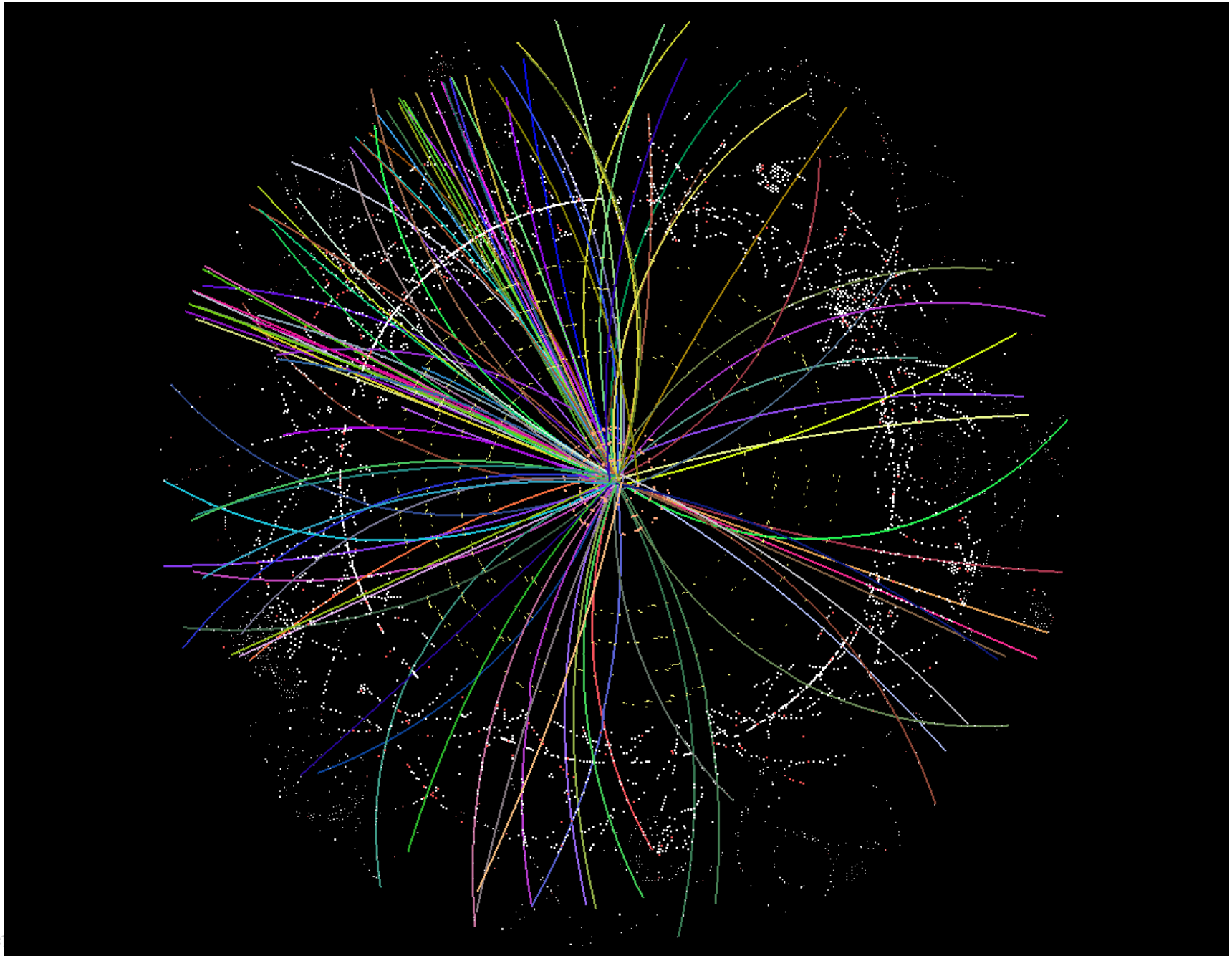
Interaction des particules avec le détecteur



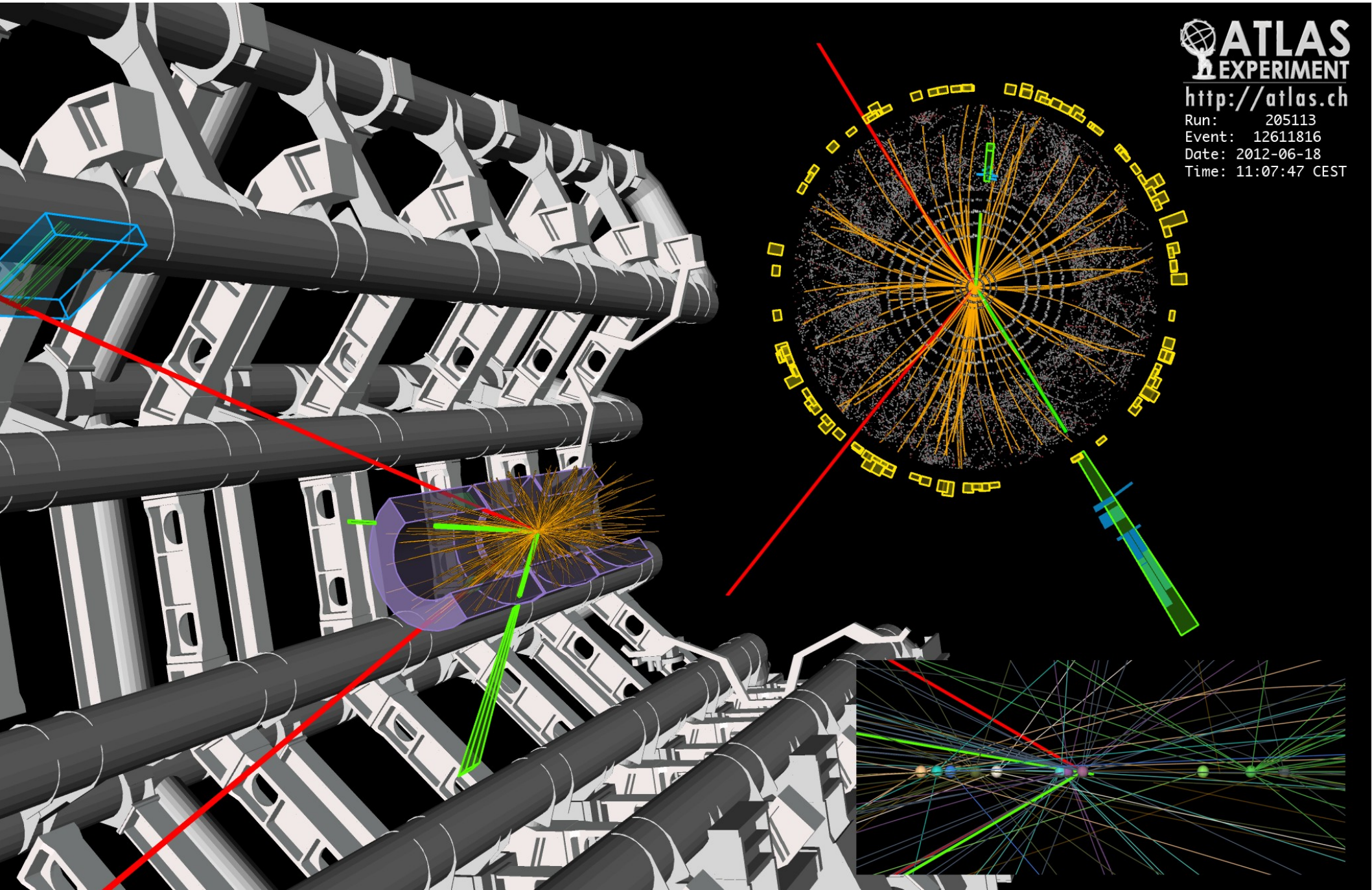
Passage des particules



Reconstruction des trajectoires



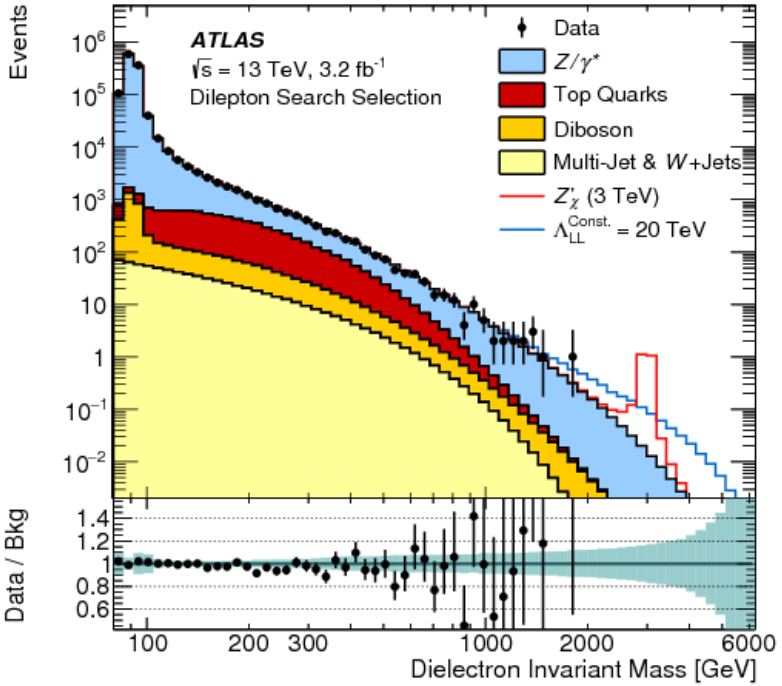
Candidat $H \rightarrow ZZ^* \rightarrow ee\mu\mu$

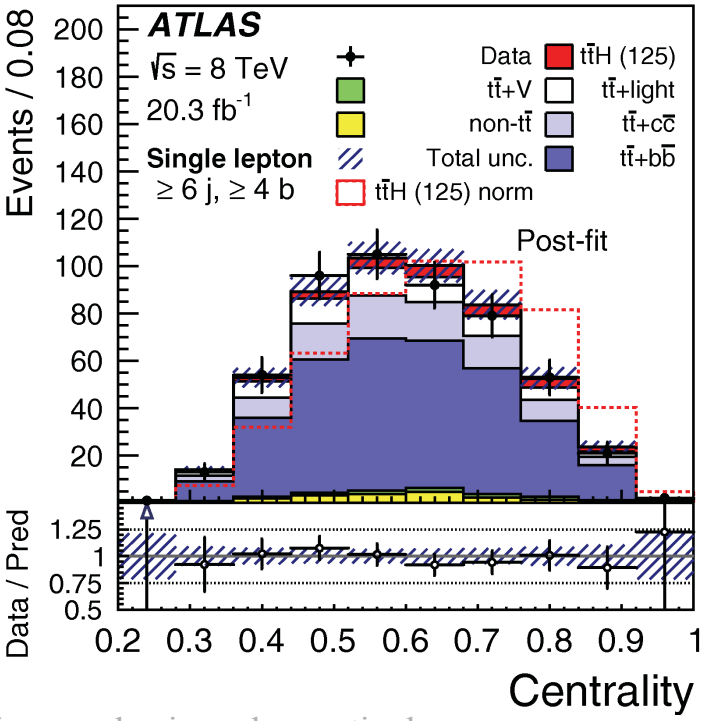


Le LHC en action

Mesure

- Analyse typique : sélection d'événements en coupant sur quelques variables (basées sur la physique), en gardant le plus de signal possible et en rejetant le plus de bruit de fond possible
- Apparition d'un pic (idéal) ou d'un petit excès distribué un peu partout (habituel...)



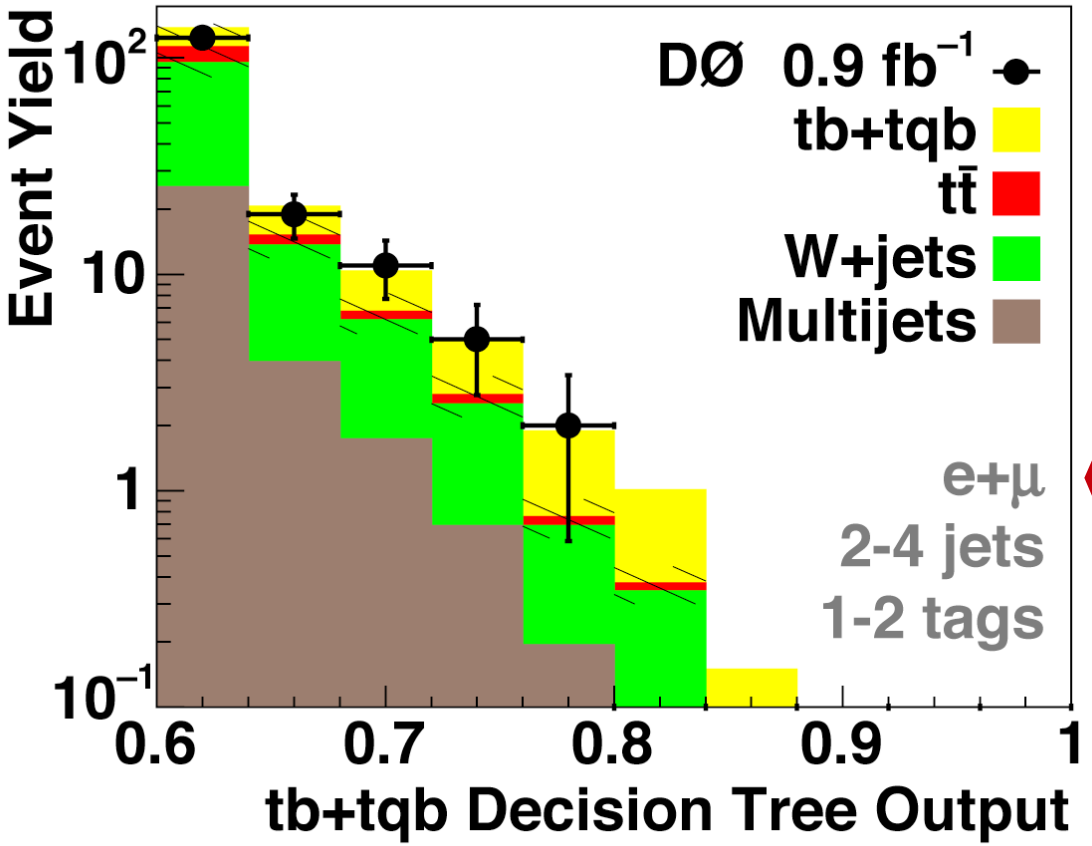


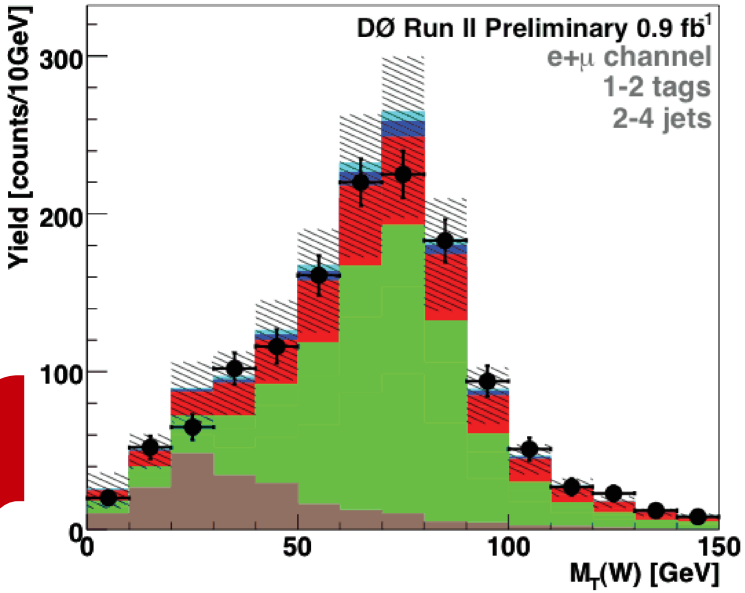
Améliorer le potentiel de découverte avec le ML

- Début des années 2000 : apparition de quelques analyses avec des réseaux de neurones
- Énormément de réticences dans la communauté (black box)

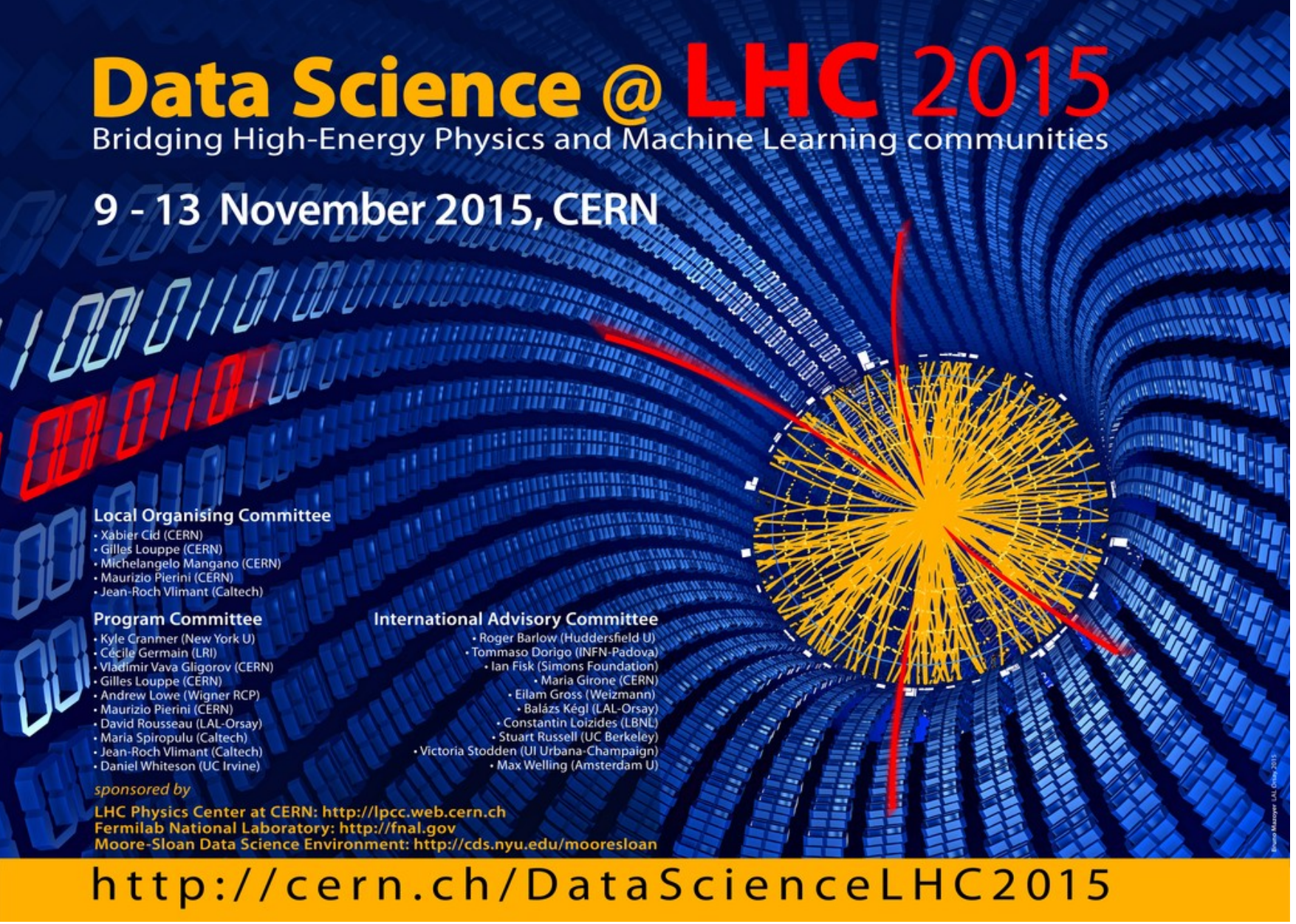
Améliorer le potentiel de découverte avec le ML

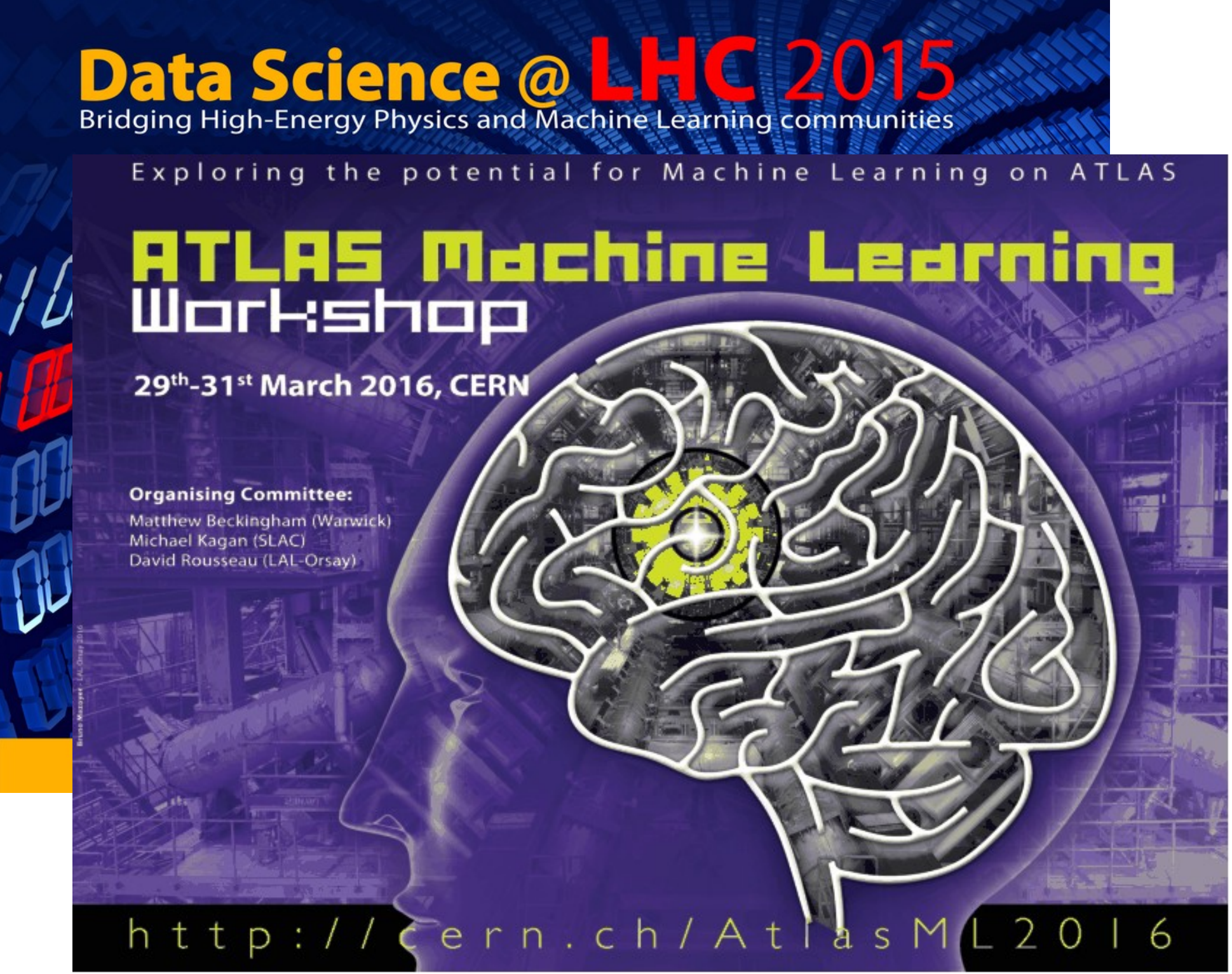
- Début des années 2000 : apparition de quelques analyses avec des réseaux de neurones
- Énormément de réticences dans la communauté (black box)

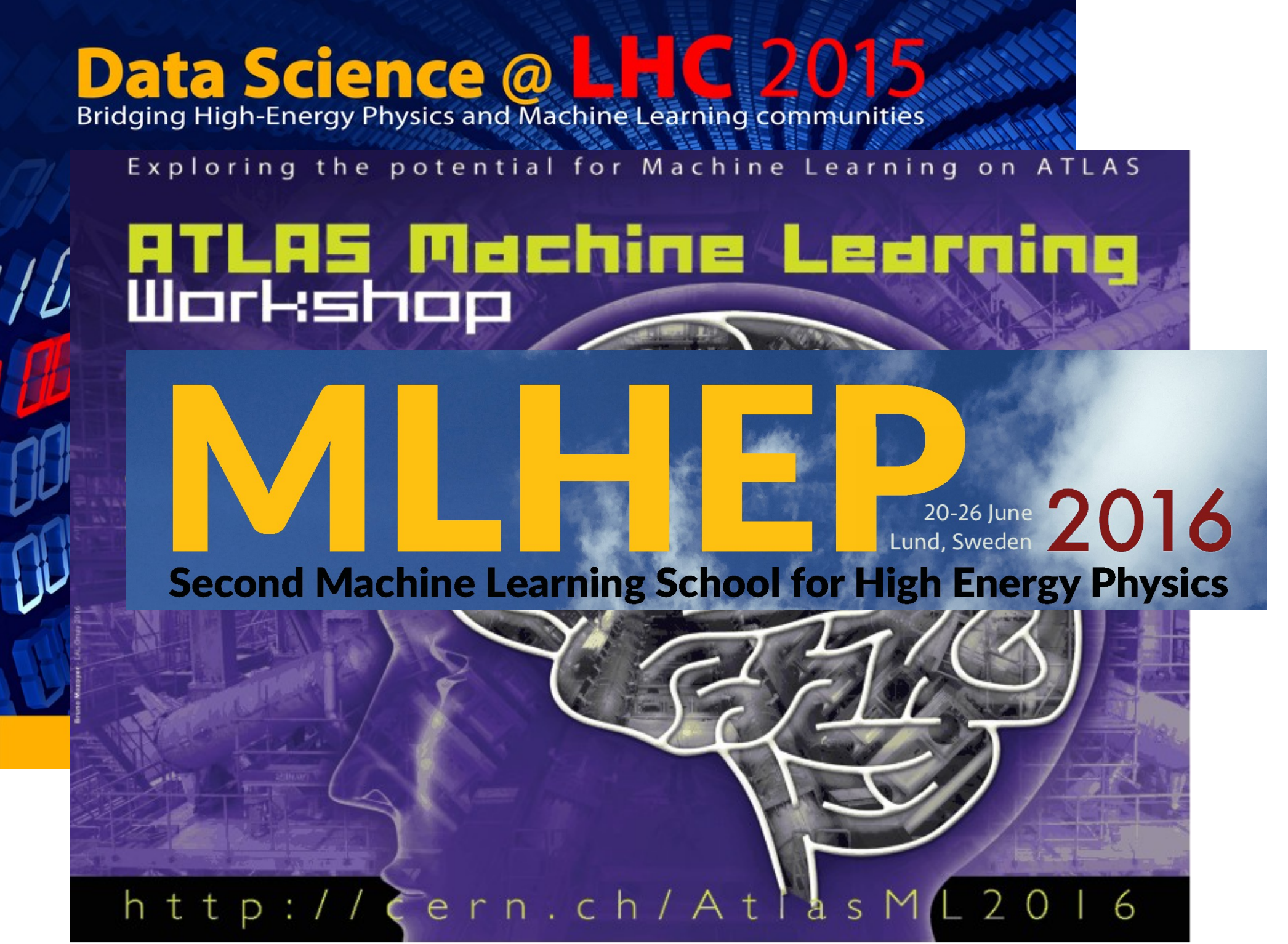


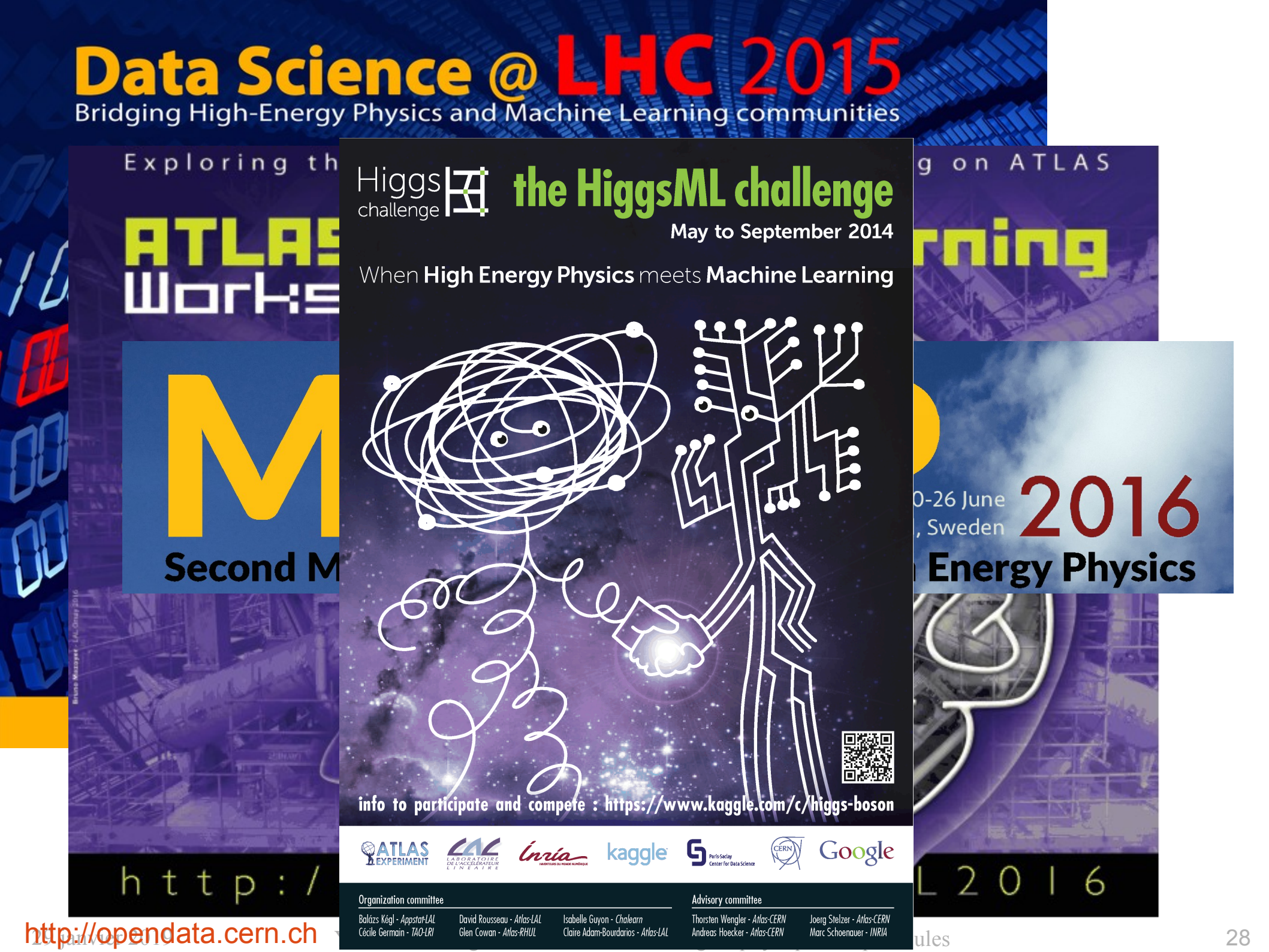


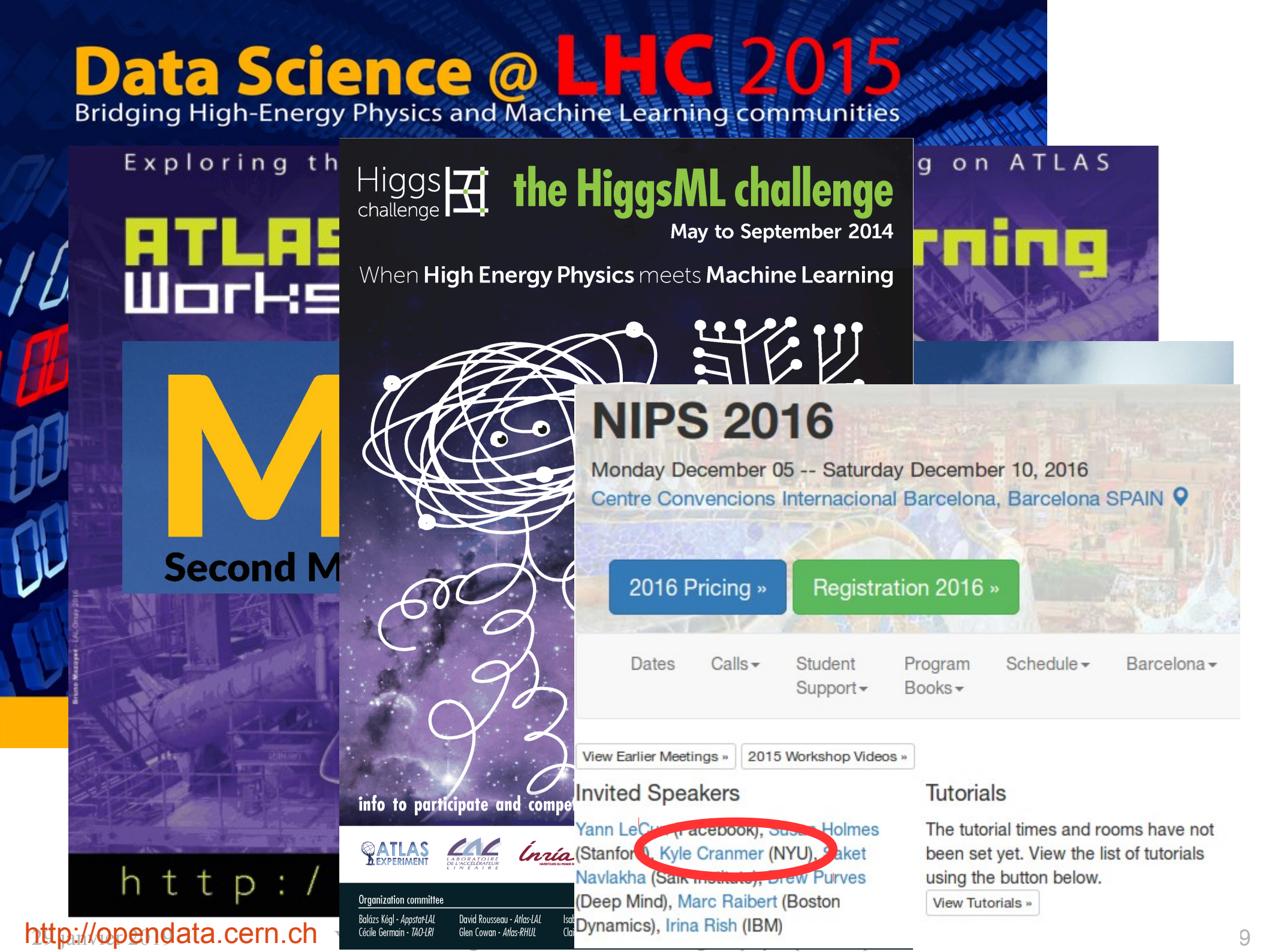
- 2006 : première utilisation des Boosted decision trees dans une analyse de physique des particules
- Très populaires depuis, car « faciles » à utiliser, bons résultats « out-of-the-box »
- Nombreux résultats du LHC avec des BDT (classification et régression)


Data Science @ LHC 2015
Bridging High-Energy Physics and Machine Learning communities
9 - 13 November 2015, CERN
Local Organising Committee
• Xabier Cid (CERN)
• Gilles Louppe (CERN)
• Michelangelo Mangano (CERN)
• Maurizio Pierini (CERN)
• Jean-Roch Vlimant (Caltech)
Program Committee
• Kyle Cranmer (New York U)
• Cécile Germain (LRI)
• Vladimir Vava Gligorov (CERN)
• Gilles Louppe (CERN)
• Andrew Lowe (Wigner RCP)
• Maurizio Pierini (CERN)
• David Rousseau (LAL-Orsay)
• Maria Spiropulu (Caltech)
• Jean-Roch Vlimant (Caltech)
• Daniel Whiteson (UC Irvine)
International Advisory Committee
• Roger Barlow (Huddersfield U)
• Tommaso Dorigo (INFN-Padova)
• Ian Fisk (Simons Foundation)
• Maria Girone (CERN)
• Eilam Gross (Weizmann)
• Balázs Kégl (LAL-Orsay)
• Constantin Loizides (LBNL)
• Stuart Russell (UC Berkeley)
• Victoria Stodden (UI Urbana-Champaign)
• Max Welling (Amsterdam U)
sponsored by
LHC Physics Center at CERN: http://lpcc.web.cern.ch
Fermilab National Laboratory: http://fnal.gov
Moore-Sloan Data Science Environment: http://cds.nyu.edu/mooresloan
http://cern.ch/DataScienceLHC2015


Data Science @ LHC 2015
Bridging High-Energy Physics and Machine Learning communities
Exploring the potential for Machine Learning on ATLAS
ATLAS Machine Learning Workshop
29th-31st March 2016, CERN
Organising Committee:
Matthew Beckingham (Warwick)
Michael Kagan (SLAC)
David Rousseau (LAL-Orsay)
http://cern.ch/AtlasML2016

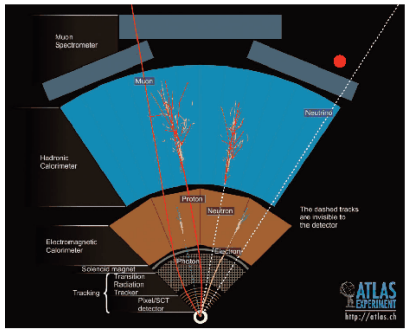

Data Science @ LHC 2015
Bridging High-Energy Physics and Machine Learning communities
Exploring the potential for Machine Learning on ATLAS
ATLAS Machine Learning
Workshop
MLHEP
20-26 June
Lund, Sweden
2016
Second Machine Learning School for High Energy Physics
http://cern.ch/AtlasML2016

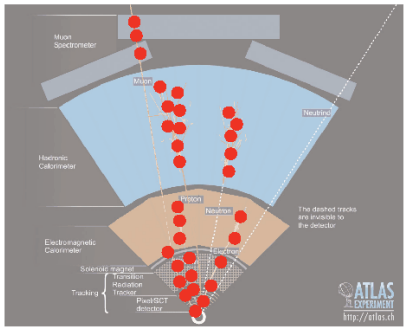

Data Science @ LHC 2015
Bridging High-Energy Physics and Machine Learning communities
Exploring th
g on ATLAS
ATLAS
Works
rning
M
Second M
0-26 June
, Sweden
2016
Energy Physics
http:/
L 2016
Higgs challenge
the HiggsML challenge
May to September 2014
When High Energy Physics meets Machine Learning
info to participate and compete : https://www.kaggle.com/c/higgs-boson
ATLAS EXPERIMENT
LAL LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE
Inria INVENTEURS DU MONDE NUMÉRIQUE
kaggle
Paris Saclay Center for Data Science
CERN
Google
Organization committee
Balázs Kégl - Appstat-LAL
Cécile Germain - TAO-LRI
David Rousseau - Atlas-LAL
Glen Cowan - Atlas-RHUL
Isabelle Guyon - Chalearn
Claire Adam-Bourdarios - Atlas-LAL
Advisory committee
Thorsten Wengler - Atlas-CERN
Andreas Hoecker - Atlas-CERN
Joerg Stelzer - Atlas-CERN
Marc Schoenauer - INRIA

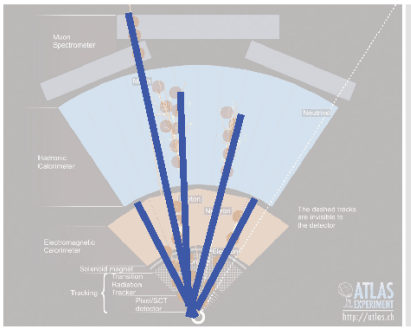

Data Science @ LHC 2015
Bridging High-Energy Physics and Machine Learning communities
Exploring th
g on ATLAS
ATLAS
Works
rning
Higgs challenge
the HiggsML challenge
May to September 2014
When High Energy Physics meets Machine Learning
Second M
info to participate and compe
NIPS 2016
Monday December 05 -- Saturday December 10, 2016
Centre Convencions Internacional Barcelona, Barcelona SPAIN
2016 Pricing »
Registration 2016 »
Dates
Calls
Student Support
Program Books
Schedule
Barcelona
View Earlier Meetings »
2015 Workshop Videos »
Invited Speakers
Yann LeCun (Facebook), Susan Holmes (Stanford), Kyle Cranmer (NYU), Blaket Navlakha (Salk Institute), Drew Purves (Deep Mind), Marc Raibert (Boston Dynamics), Irina Rish (IBM)
Tutorials
The tutorial times and rooms have not been set yet. View the list of tutorials using the button below.
View Tutorials »
ATLAS EXPERIMENT
LAL LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE
Inria
Organization committee
Balázs Kégl - Appstat-LAL
Cécile Germain - TAO-LRI
David Rousseau - Atlas-LAL
Glen Cowan - Atlas-RHUL
http:/

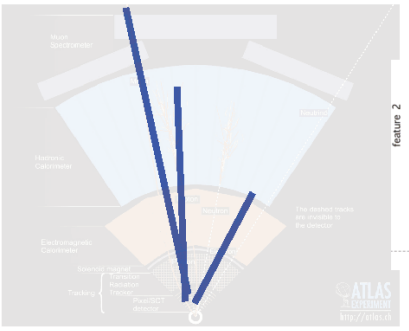
Dimensionnalité

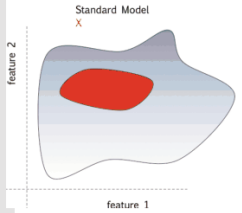
Raw	Sparsified	Reco	Select	Physics	Ana
1e7	1e4	100-ish*	50	10	1

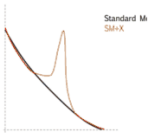






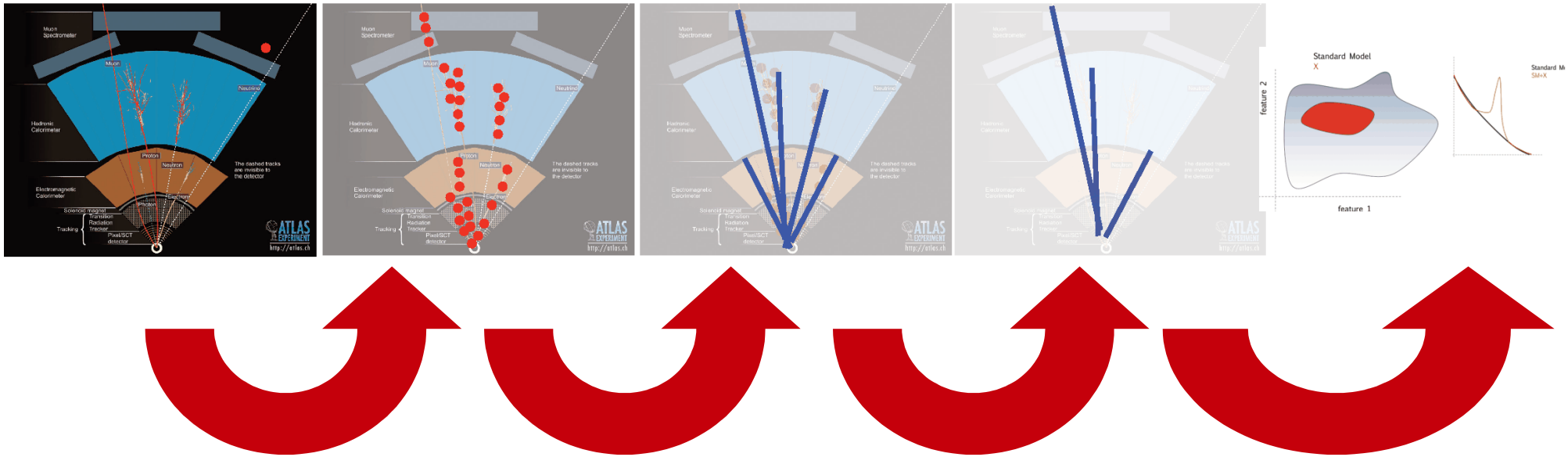






Dimensionnalité

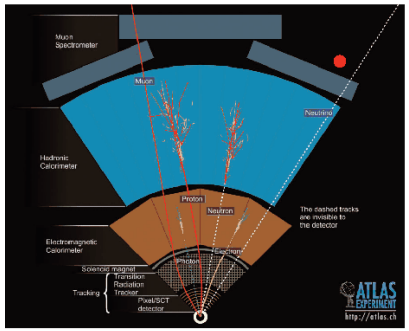
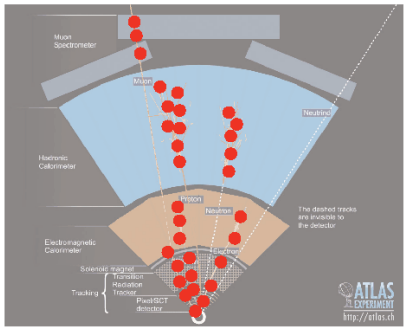
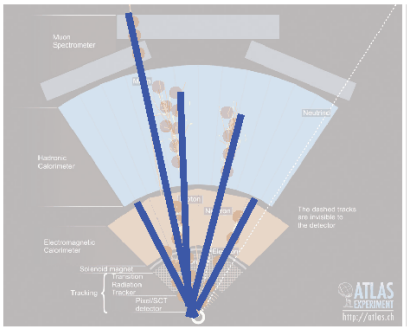
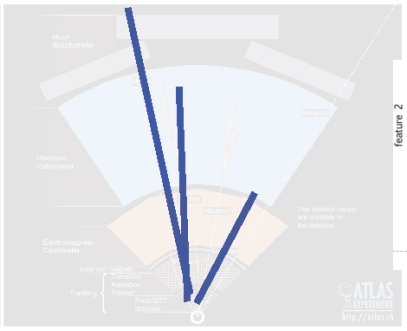
Raw	Sparsified	Reco	Select	Physics	Ana
1e7	1e4	100-ish*	50	10	1



Améliorer chaque étape avec le ML ?

Dimensionnalité

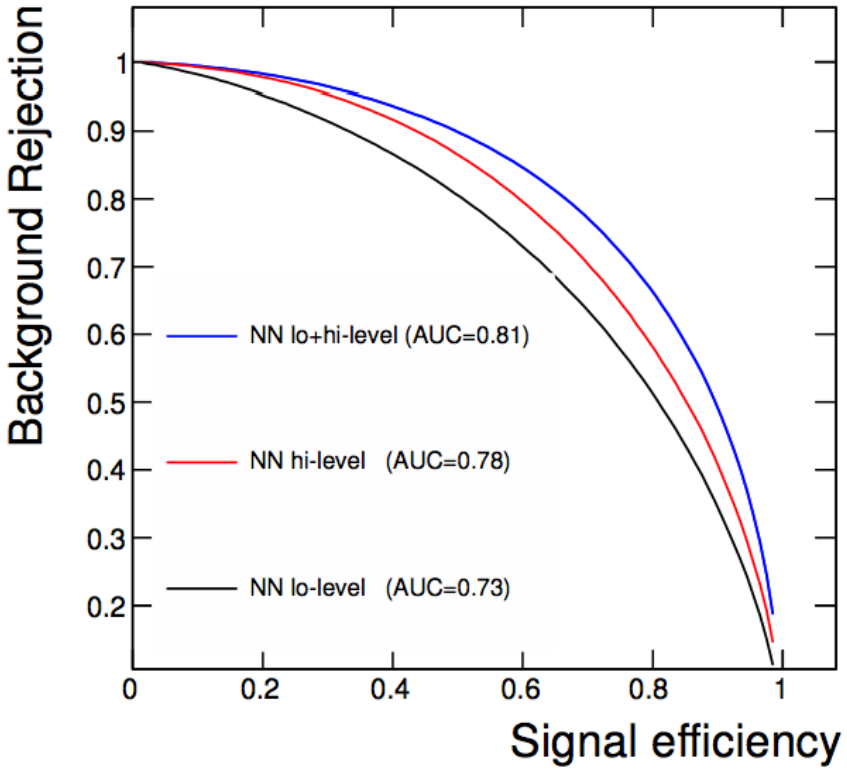
Raw	Sparsified	Reco	Select	Physics	Ana
1e7	1e4	100-ish*	50	10	1

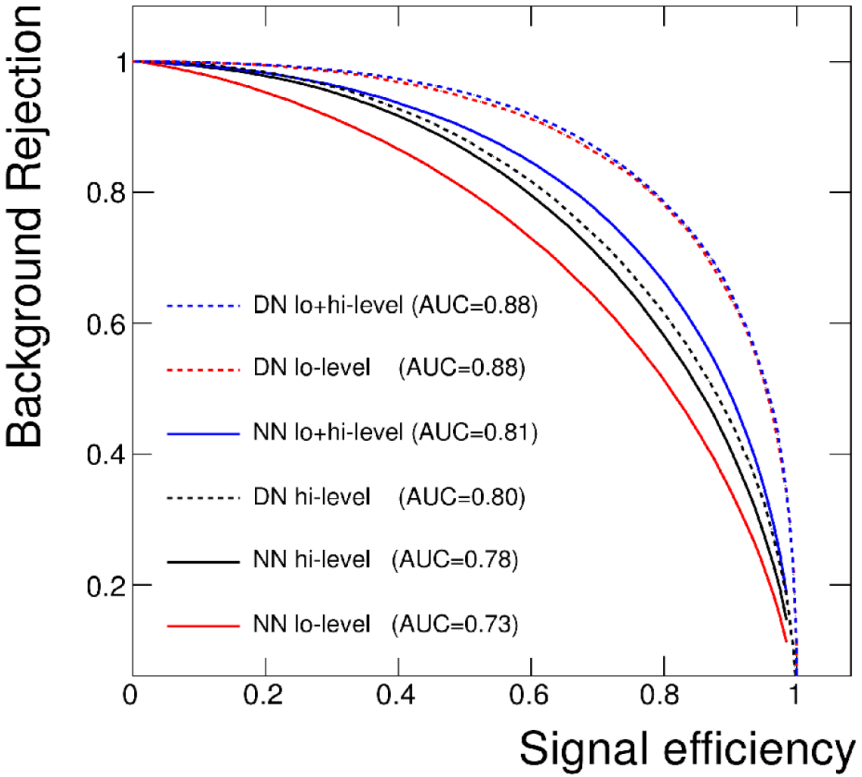





Sauter une ou des étapes avec le ML ?

Supprimer les variables physiques ?

http://arxiv.org/abs/1410.3469

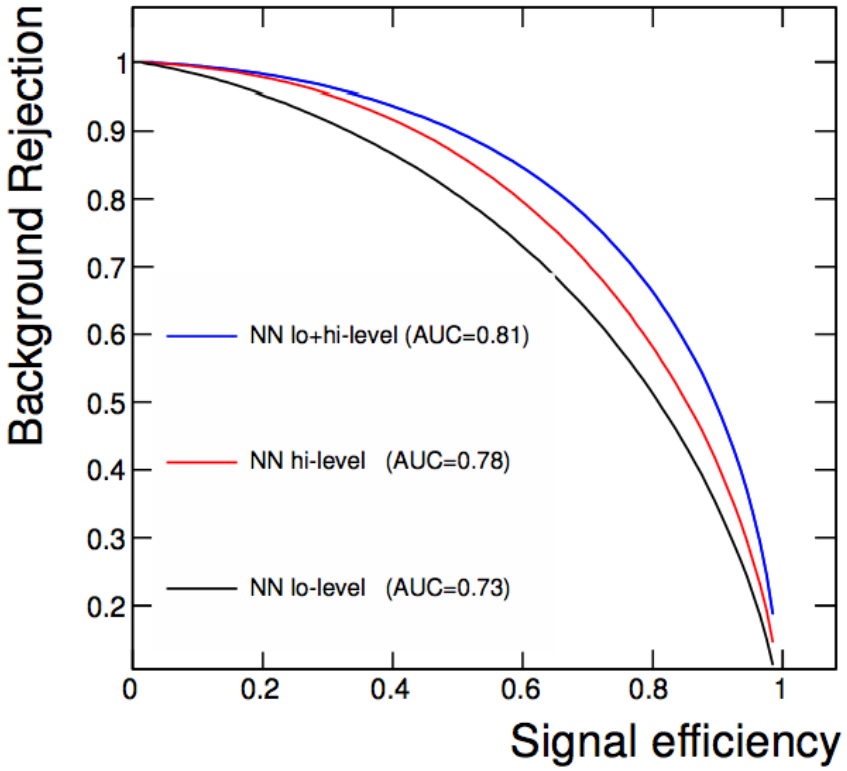


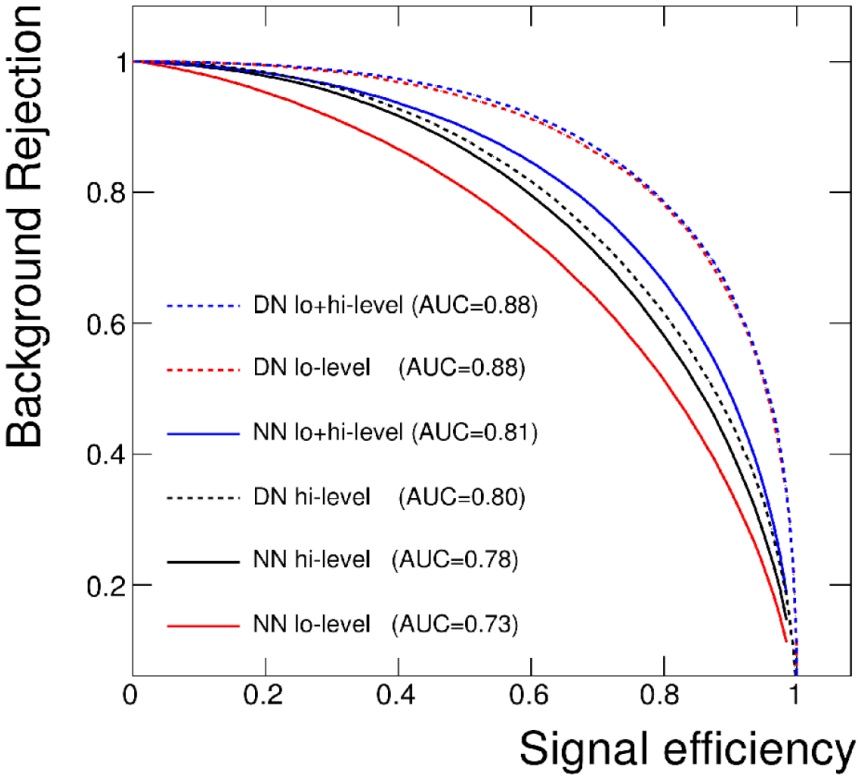


- Lo-level : propriétés de base des objets
- Hi-level : variables construites par les physiciens
- NN (shallow network classique) : n'y arrive pas avec low seul
- DNN : meilleur avec low que high (!)

Supprimer les variables physiques ?

http://arxiv.org/abs/1410.3469



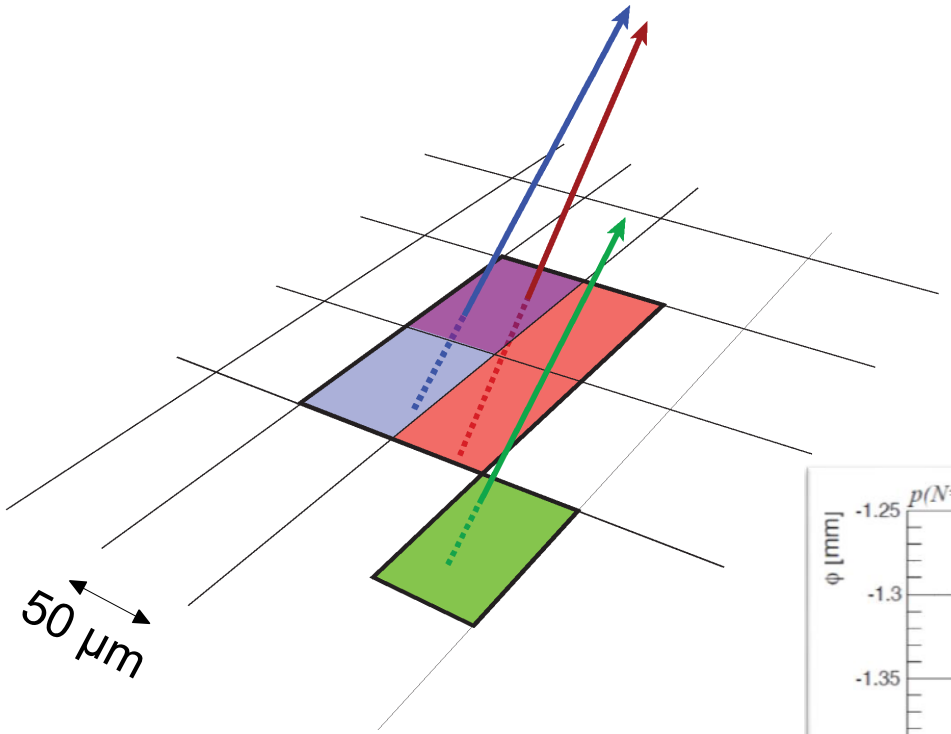


- Lo-level : propriétés de base des objets
- Hi-level : variables construites par les physiciens
- NN (shallow network classique) : n'y arrive pas avec low seul
- DNN : meilleur avec low que high (!)

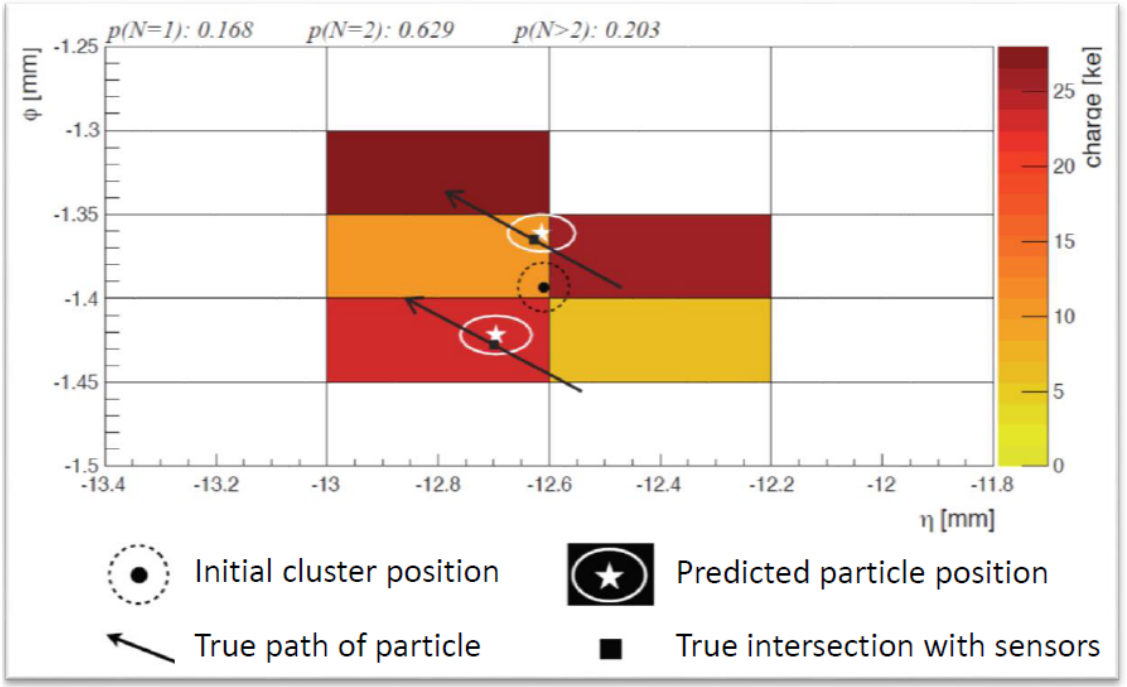
Recherche en cours, pourrait révolutionner notre façon de faire des analyses

Améliorer les inputs

- Mieux mesurer les propriétés des traces

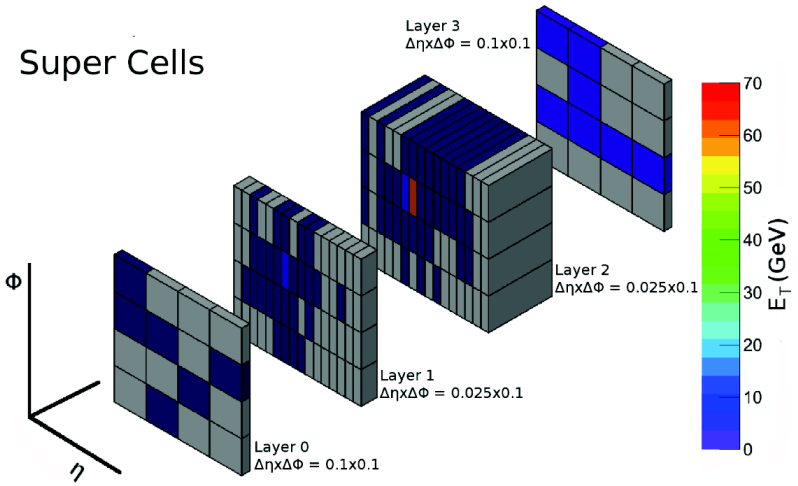


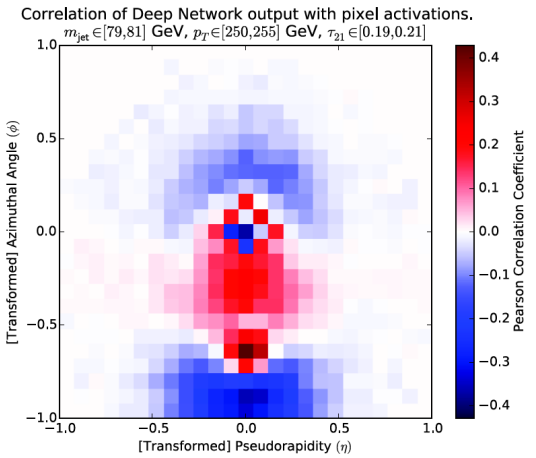
- 10 NN pour décider
  - du nombre de traces
  - du point d'impact
  - des erreurs associées

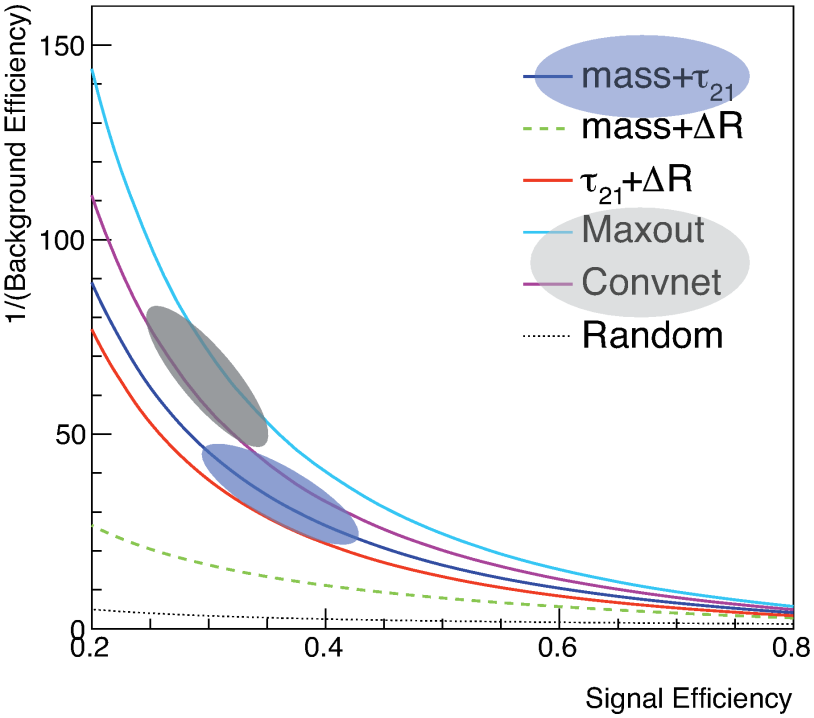


Transformer nos données en images

- Tentative d'utiliser les algorithmes d'analyse d'image (CNN, etc.) pour classer les jets de particules







http://arxiv.org/abs/1511.05190

Détection d'anomalies

- Pas encore actif, mais études en cours
- Potentiel :
  - découvrir plus rapidement des problèmes pendant la prise de données
  - validation de notre software de physique et de la qualité des données
    - actuellement : comparaison d'histogrammes
    - en considération : entraîner un classifieur pour séparer versions A et B, ou entraîner un auto-encoder sur A et tester sur B

Détection d'anomalies

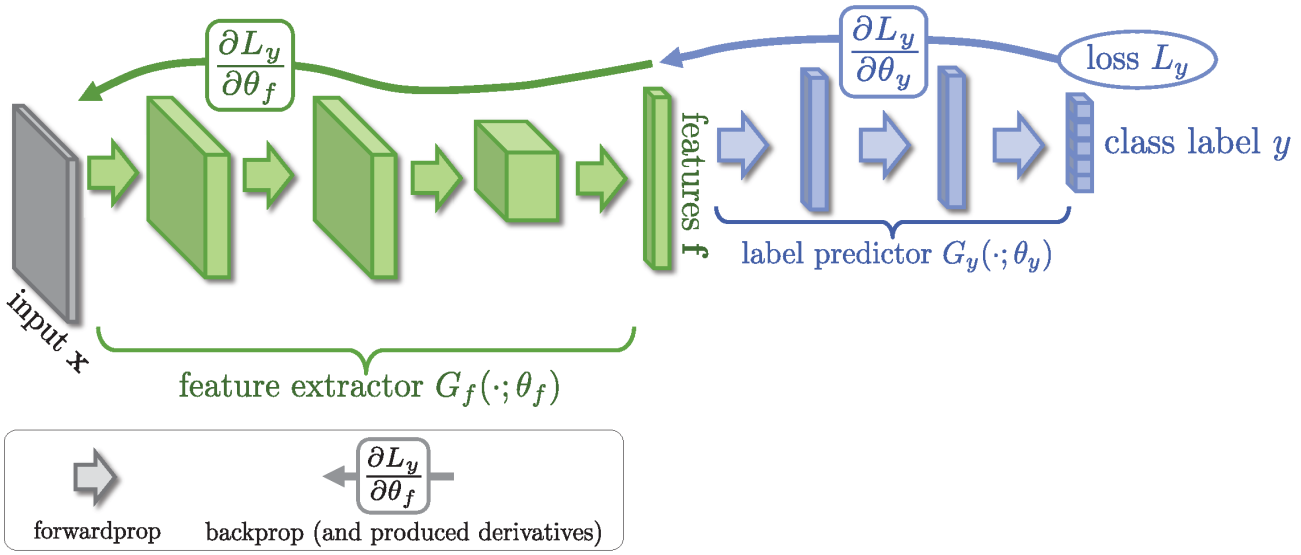
- Pas encore actif, mais études en cours
- Potentiel :
  - découvrir plus rapidement des problèmes pendant la prise de données
  - validation de notre software de physique et de la qualité des données
    - actuellement : comparaison d'histogrammes
    - en considération : entraîner un classifieur pour séparer versions A et B, ou entraîner un auto-encoder sur A et tester sur B

En cours de développement

Encore prospectif

http://arxiv.org/abs/1409.7495
http://arxiv.org/abs/1505.07818

- Entraînement typique :
  - signal et bruit de fond viennent de simulation
  - résultats comparés aux données réelles
- Impose un bon accord données/simulation

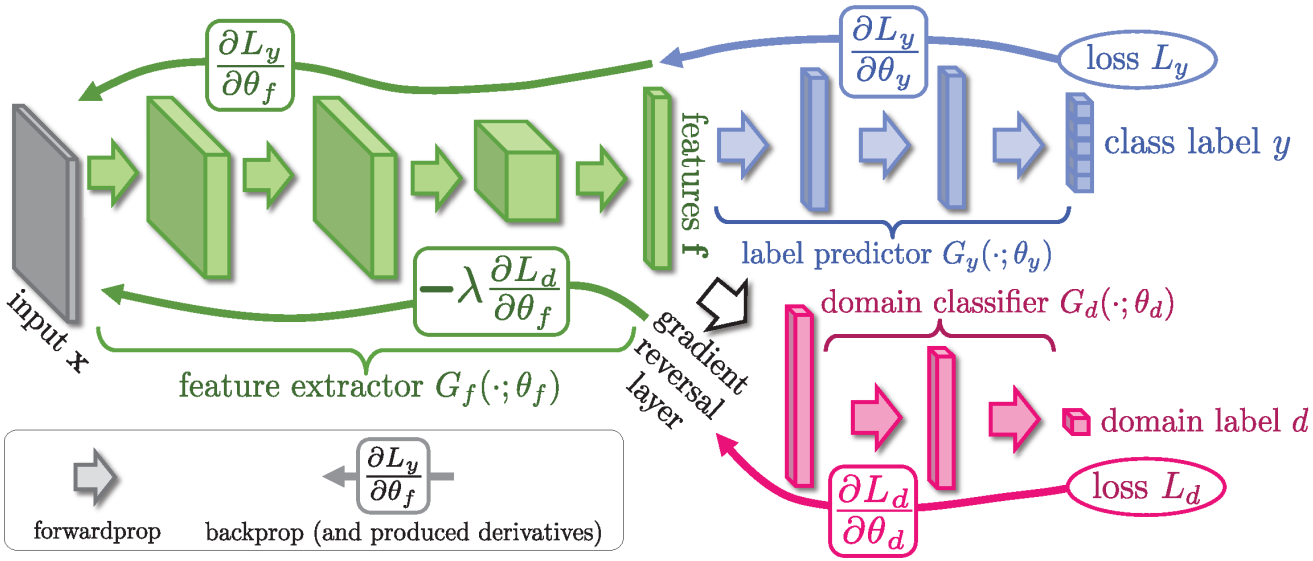


Encore prospectif

http://arxiv.org/abs/1409.7495
http://arxiv.org/abs/1505.07818

- Entraînement typique :
  - signal et bruit de fond viennent de simulation
  - résultats comparés aux données réelles
- Impose un bon accord données/simulation

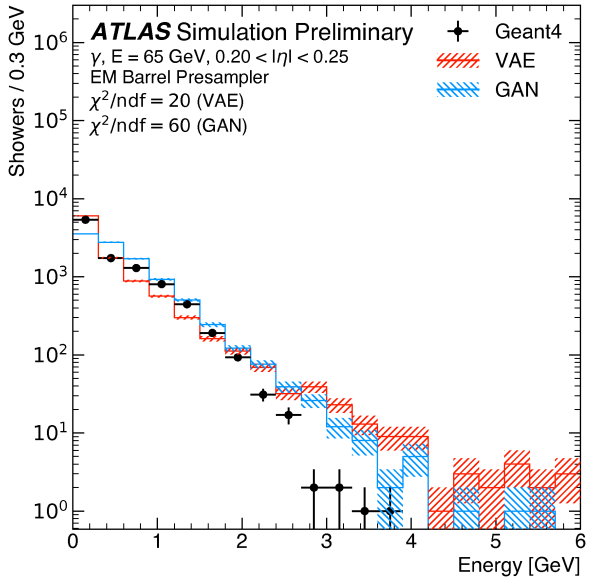
- Possible utilisation de l'adversarial training/domain adaptation

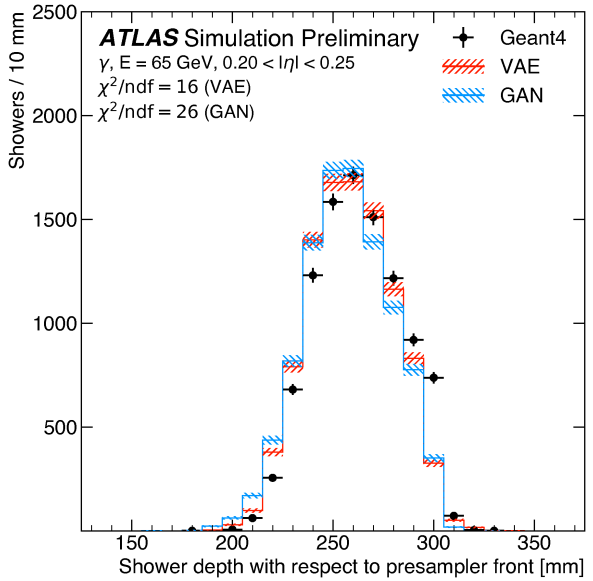


Modèles génératifs

ATLAS PUB note ATL-SOFT-PUB-2018-001

- Simulation : coût CPU très élevé
  - facteur de plus en plus limitant pour les analyses
  - pas assez d'événements simulés
- Remplacer « full simulation » par approximation
- Generative adversarial networks (GAN) et variational auto-encoders (VAE)

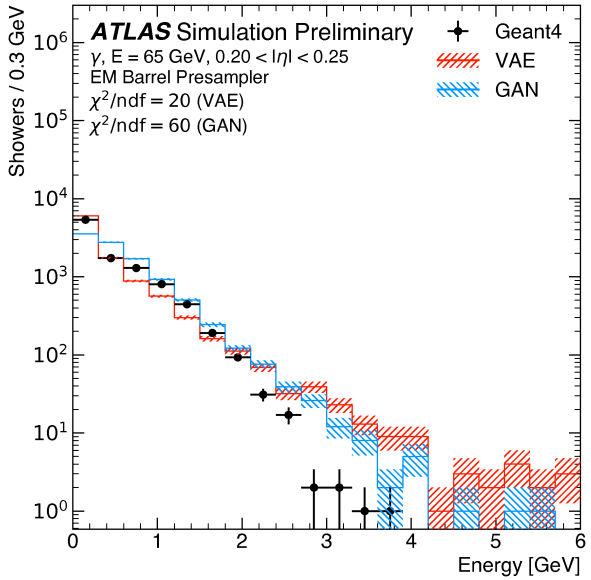


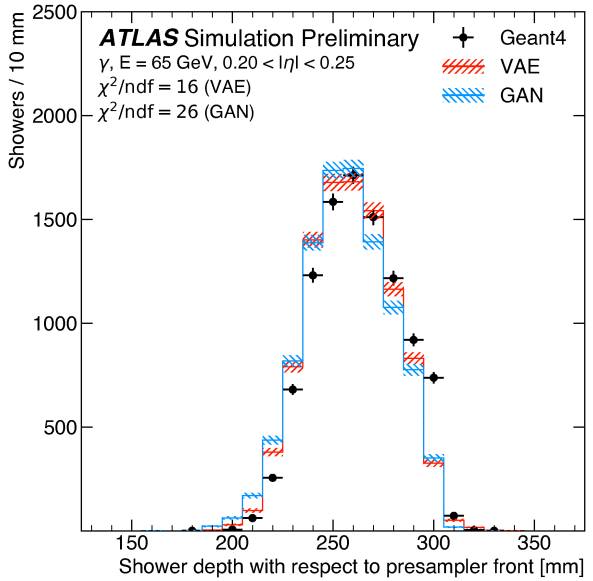


Modèles génératifs

ATLAS PUB note ATL-SOFT-PUB-2018-001

- Simulation : coût CPU très élevé
  - facteur de plus en plus limitant pour les analyses
  - pas assez d'événements simulés
- Remplacer « full simulation » par approximation
- Generative adversarial networks (GAN) et variational auto-encoders (VAE)





Encore beaucoup de chemin à parcourir

Au CPPM

- Infrastructure
  - T2 grille LHC
    - 1920 cœurs, 1.4 PB, réseau 10 Gb/s
- Middleware
  - DIRAC
- Applications
  - LHC
    - BDT dans ATLAS et LHCb
    - ATLAS : thèse en cours en codirection inter-écoles doctorales avec le LIS pour évaluer le potentiel en analyse (low level inputs, RNN, adversarial training, parse trees, etc.)
  - Traitement d'images
  - LSST (cf. Johanna)
- Mise en place d'un groupe de discussions inter-expériences

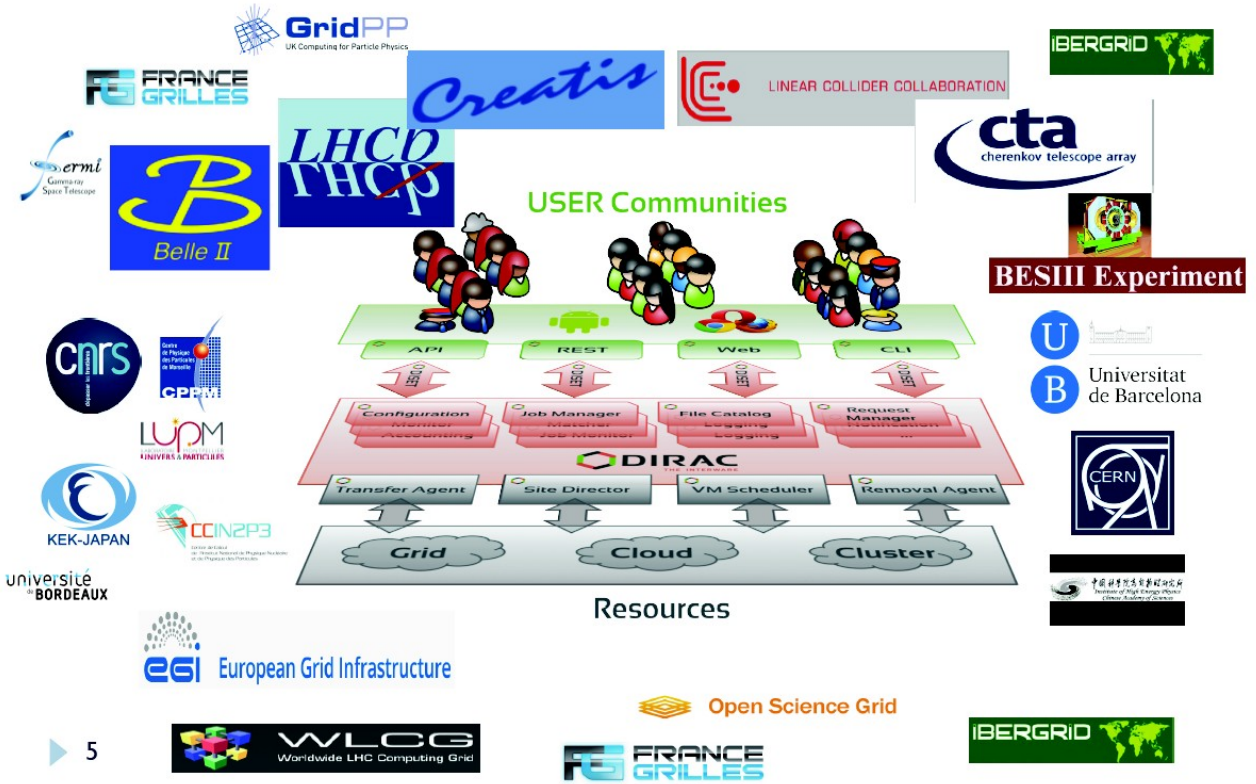
Au CPPM

http://diracgrid.org

- Inf
  - .
- Mi
  - |
- Ap
  - |
  - .
  - |
- Mi
  - .

DIRAC
THE INTERWARE

- DIRAC provides all the necessary components to build ad-hoc grid infrastructures **interconnecting** computing resources of different types, allowing **interoperability** and simplifying **interfaces**. This allows to speak about the DIRAC ***interware***.



pour

2

5

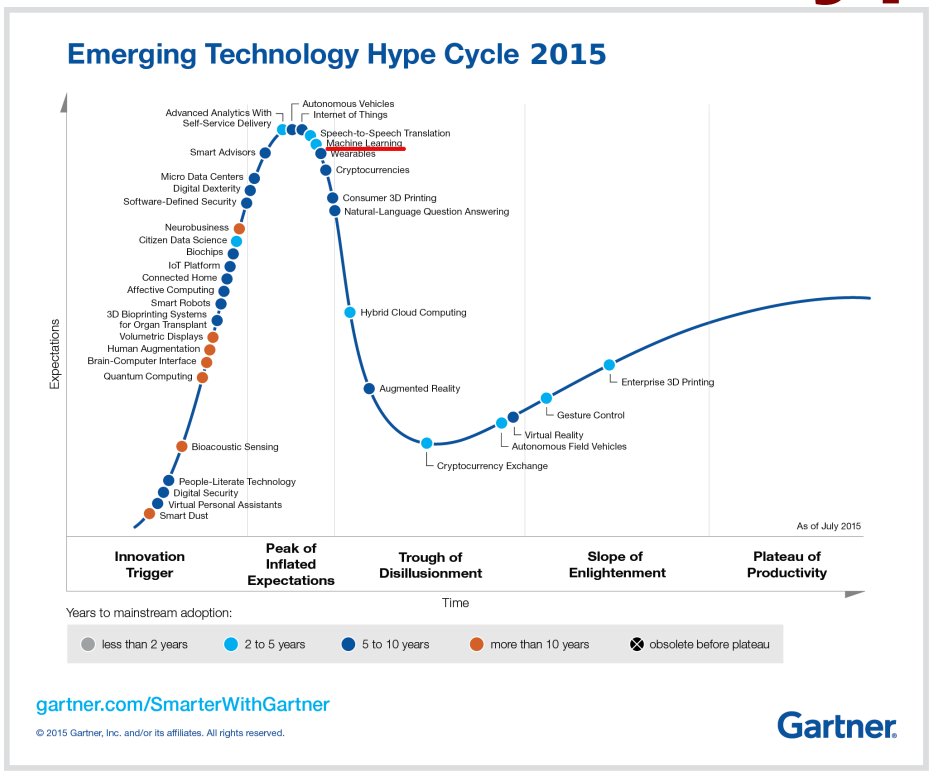
Au CPPM

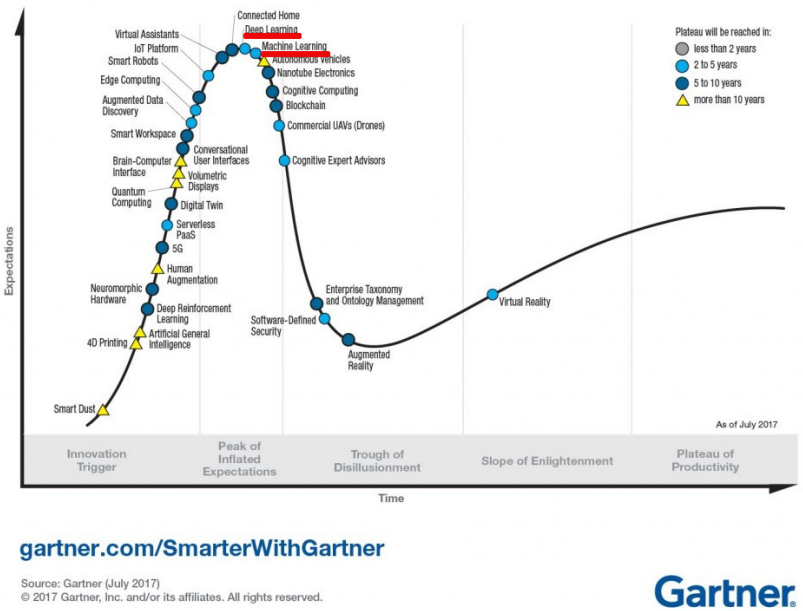
- Infrastructure
  - T2 grille LHC
    - 1920 cœurs, 1.4 PB, réseau 10 Gb/s
- Middleware
  - DIRAC
- Applications
  - LHC
    - BDT dans ATLAS et LHCb
    - ATLAS : thèse en cours en codirection inter-écoles doctorales avec le LIS pour évaluer le potentiel en analyse (low level inputs, RNN, adversarial training, parse trees, etc.)
  - Traitement d'images
  - LSST (cf. Johanna)
- Mise en place d'un groupe de discussions inter-expériences

Conclusion

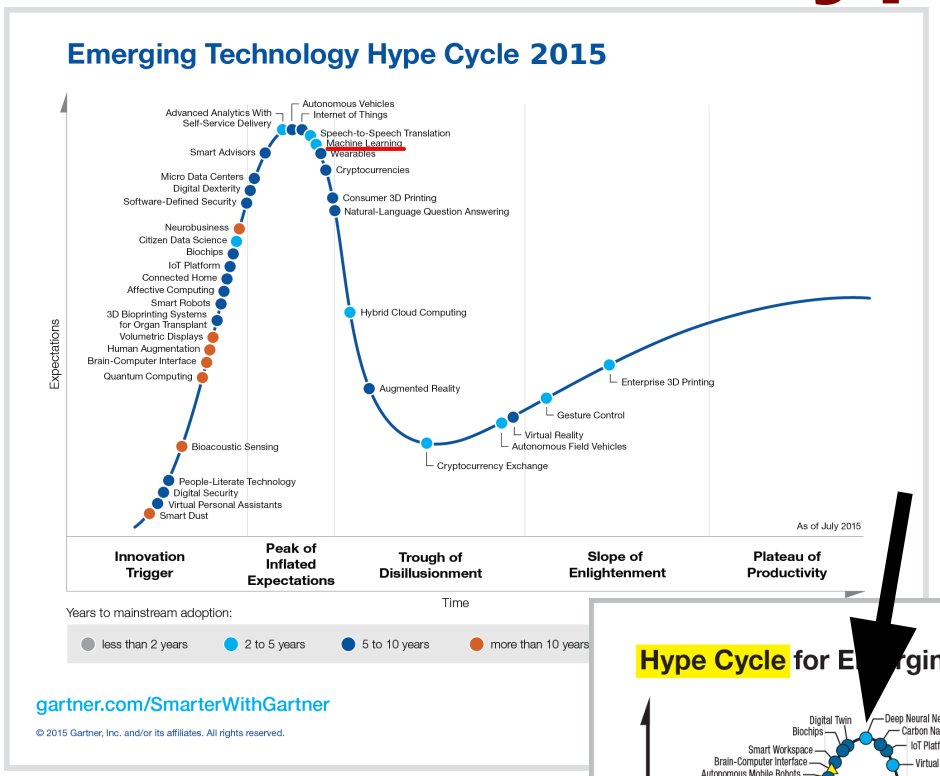
- LHC = big data (exabyte)
- Grille de calcul pour le rendre possible
  - Challenge du HL-LHC
- Nouveautés en Machine Learning : tendance à prendre un peu de temps à arriver en physique des particules (par ex 10 ans pour les BDT)
- La communauté a encore un peu de mal à accepter les black boxes, mais les performances parfois très supérieures aident à convaincre
- Une grande partie des résultats du LHC dépendent maintenant du ML
  - Sans le ML : besoin d'encore plus de données, certains résultats impossibles
- Génère un « coût » computing de plus en plus élevé
- La tendance à l'adoption s'accélère : le Deep learning a le vent en poupe, les RNN, adversarial training, etc.

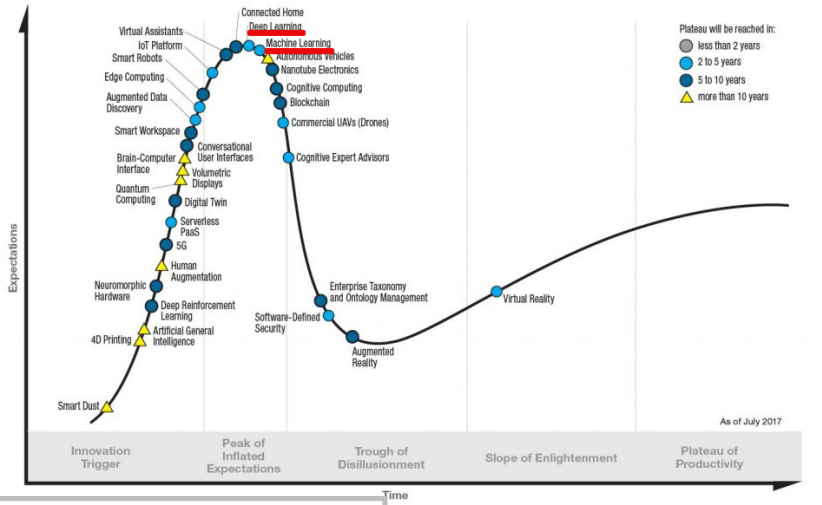
Hype cycle

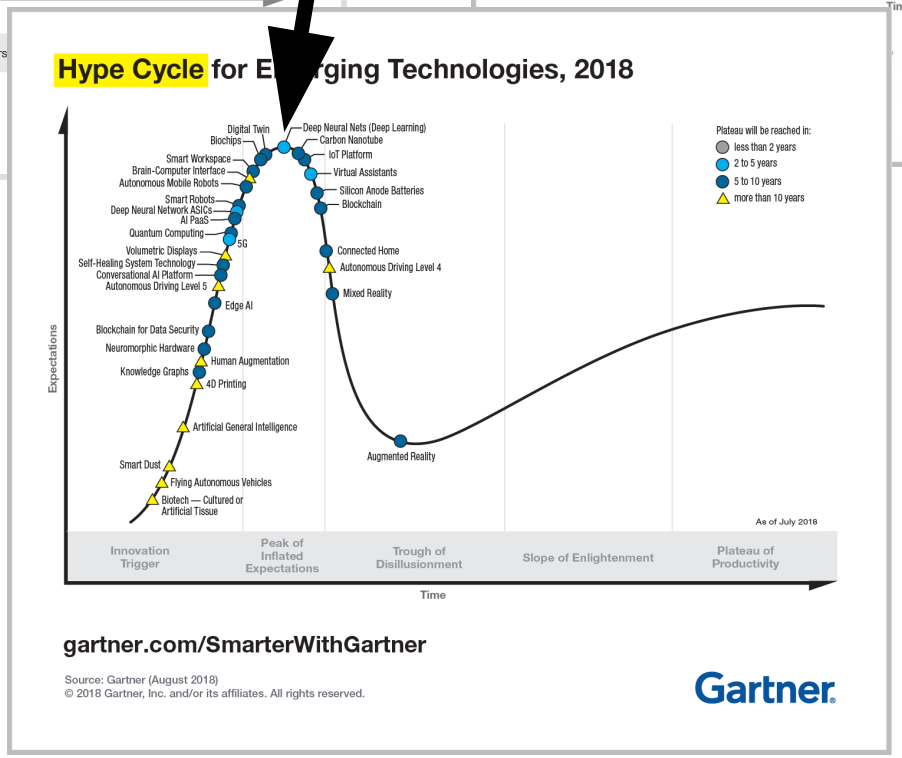

Emerging Technology Hype Cycle 2015
Expectations
Advanced Analytics With Self-Service Delivery
Autonomous Vehicles
Internet of Things
Speech-to-Speech Translation
Machine Learning
Wearables
Smart Advisors
Cryptocurrencies
Micro Data Centers
Digital Dexterity
Software-Defined Security
Consumer 3D Printing
Natural-Language Question Answering
Neurobusiness
Citizen Data Science
Biochips
IoT Platform
Connected Home
Affective Computing
Smart Robots
3D Bioprinting Systems for Organ Transplant
Volumetric Displays
Human Augmentation
Brain-Computer Interface
Quantum Computing
Hybrid Cloud Computing
Augmented Reality
Enterprise 3D Printing
Gesture Control
Virtual Reality
Autonomous Field Vehicles
Cryptocurrency Exchange
Bioacoustic Sensing
People-Literate Technology
Digital Security
Virtual Personal Assistants
Smart Dust
As of July 2015
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Years to mainstream adoption:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
gartner.com/SmarterWithGartner
© 2015 Gartner, Inc. and/or its affiliates. All rights reserved.
Gartner


Gartner Hype Cycle for Emerging Technologies, 2017
Expectations
Connected Home
Deep Learning
Machine Learning
Virtual Assistants
IoT Platform
Smart Robots
Autonomous Vehicles
Nanotube Electronics
Edge Computing
Cognitive Computing
Augmented Data Discovery
Blockchain
Commercial UAVs (Drones)
Smart Workspace
Conversational User Interfaces
Cognitive Expert Advisors
Brain-Computer Interface
Volumetric Displays
Quantum Computing
Digital Twin
Serverless PaaS
5G
Human Augmentation
Neuromorphic Hardware
Deep Reinforcement Learning
Artificial General Intelligence
4D Printing
Smart Dust
Enterprise Taxonomy and Ontology Management
Software-Defined Security
Augmented Reality
Virtual Reality
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
As of July 2017
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
gartner.com/SmarterWithGartner
Source: Gartner (July 2017)
© 2017 Gartner, Inc. and/or its affiliates. All rights reserved.
Gartner

Hype cycle


Emerging Technology Hype Cycle 2015
Expectations
Time
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Autonomous Vehicles
Internet of Things
Advanced Analytics With Self-Service Delivery
Speech-to-Speech Translation
Machine Learning
Wearables
Smart Advisors
Cryptocurrencies
Micro Data Centers
Digital Dexterity
Software-Defined Security
Consumer 3D Printing
Natural-Language Question Answering
Neurobusiness
Citizen Data Science
Biochips
IoT Platform
Connected Home
Affective Computing
Smart Robots
3D Bioprinting Systems for Organ Transplant
Volumetric Displays
Human Augmentation
Brain-Computer Interface
Quantum Computing
Hybrid Cloud Computing
Augmented Reality
Enterprise 3D Printing
Gesture Control
Virtual Reality
Autonomous Field Vehicles
Cryptocurrency Exchange
Bioacoustic Sensing
People-Literate Technology
Digital Security
Virtual Personal Assistants
Smart Dust
As of July 2015
Years to mainstream adoption:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
gartner.com/SmarterWithGartner
© 2015 Gartner, Inc. and/or its affiliates. All rights reserved.


Gartner Hype Cycle for Emerging Technologies, 2017
Expectations
Time
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Connected Home
Deep Learning
Machine Learning
Virtual Assistants
IoT Platform
Smart Robots
Autonomous Vehicles
Nanotube Electronics
Edge Computing
Cognitive Computing
Augmented Data Discovery
Blockchain
Commercial UAVs (Drones)
Smart Workspace
Conversational User Interfaces
Cognitive Expert Advisors
Brain-Computer Interface
Volumetric Displays
Quantum Computing
Digital Twin
Serverless PaaS
5G
Human Augmentation
Neuromorphic Hardware
Deep Reinforcement Learning
Enterprise Taxonomy and Ontology Management
Virtual Reality
Software-Defined Security
4D Printing
Artificial General Intelligence
Augmented Reality
Smart Dust
As of July 2017
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
Gartner


Hype Cycle for E rging Technologies, 2018
Expectations
Time
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Digital Twin
Deep Neural Nets (Deep Learning)
Biochips
Carbon Nanotube
Smart Workspace
IoT Platform
Brain-Computer Interface
Virtual Assistants
Autonomous Mobile Robots
Silicon Anode Batteries
Smart Robots
Blockchain
Deep Neural Network ASICs
AI PaaS
Quantum Computing
5G
Connected Home
Volumetric Displays
Autonomous Driving Level 4
Self-Healing System Technology
Conversational AI Platform
Mixed Reality
Autonomous Driving Level 5
Edge AI
Blockchain for Data Security
Neuromorphic Hardware
Human Augmentation
Knowledge Graphs
4D Printing
Artificial General Intelligence
Augmented Reality
Smart Dust
Flying Autonomous Vehicles
Biotech — Cultured or Artificial Tissue
As of July 2018
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
gartner.com/SmarterWithGartner
Source: Gartner (August 2018)
© 2018 Gartner, Inc. and/or its affiliates. All rights reserved.
Gartner

Liens

ATLAS@home atlasathome.cern.ch

ATLAS grand public atlas.cern

ATLAS en direct atlas-live.cern.ch

ATLAS sur twitter.com/ATLASexperiment

ATLAS sur www.facebook.com/ATLASexperiment

ATLAS sur www.instagram.com/atlasexperiment

ATLAS sur www.youtube.com/theATLASExperiment

Site français du www.lhc-france.fr

Le CPPM www.cppm.in2p3.fr twitter.com/cppmluminy

Le CERN home.cern twitter.com/cern

Le CERN sur www.facebook.com/cern