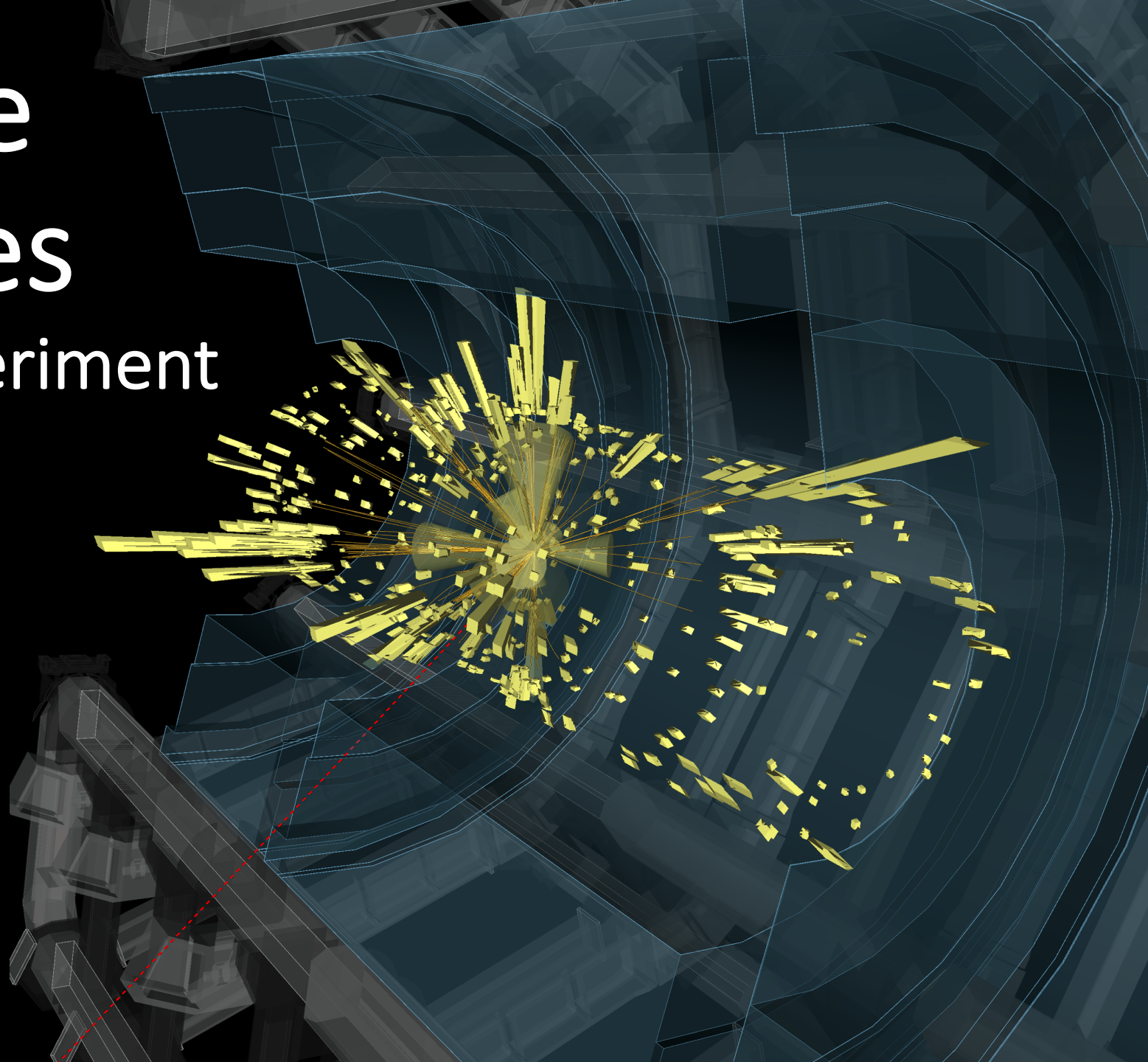
Collaborative data practices

at the ATLAS experiment of the CERN LHC



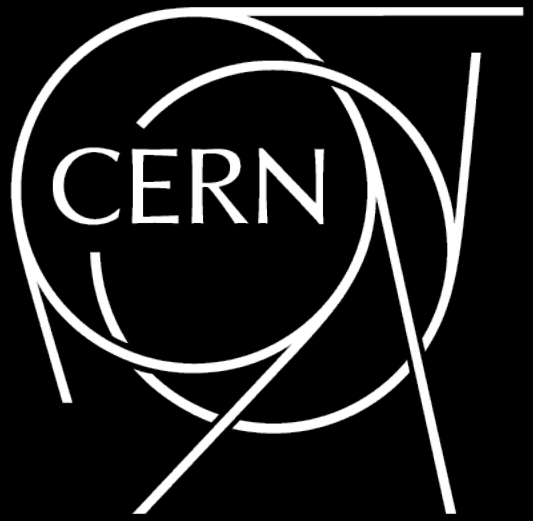
Anna Sfyrla

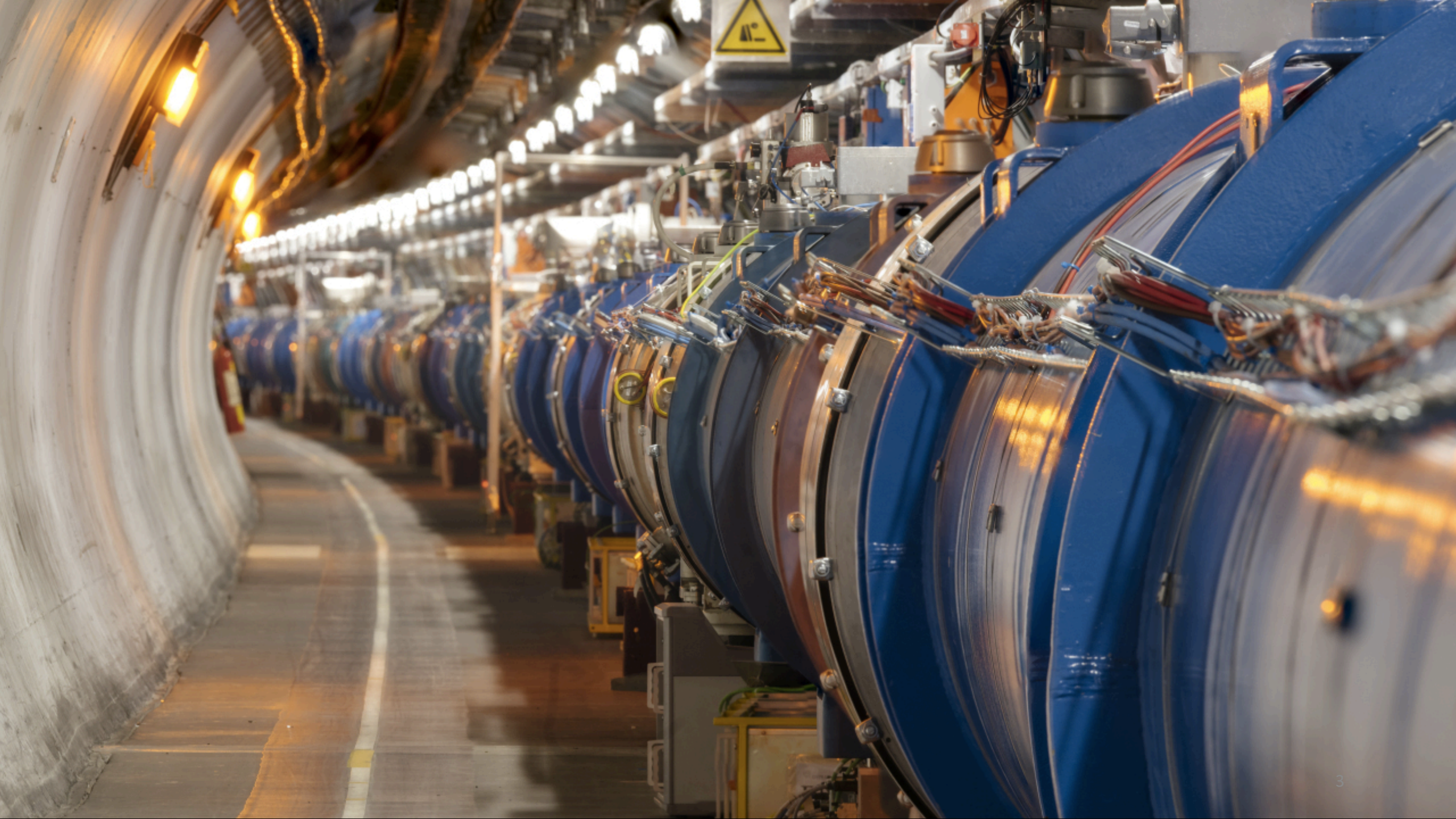
Data Science Seminar – 11 January 2021

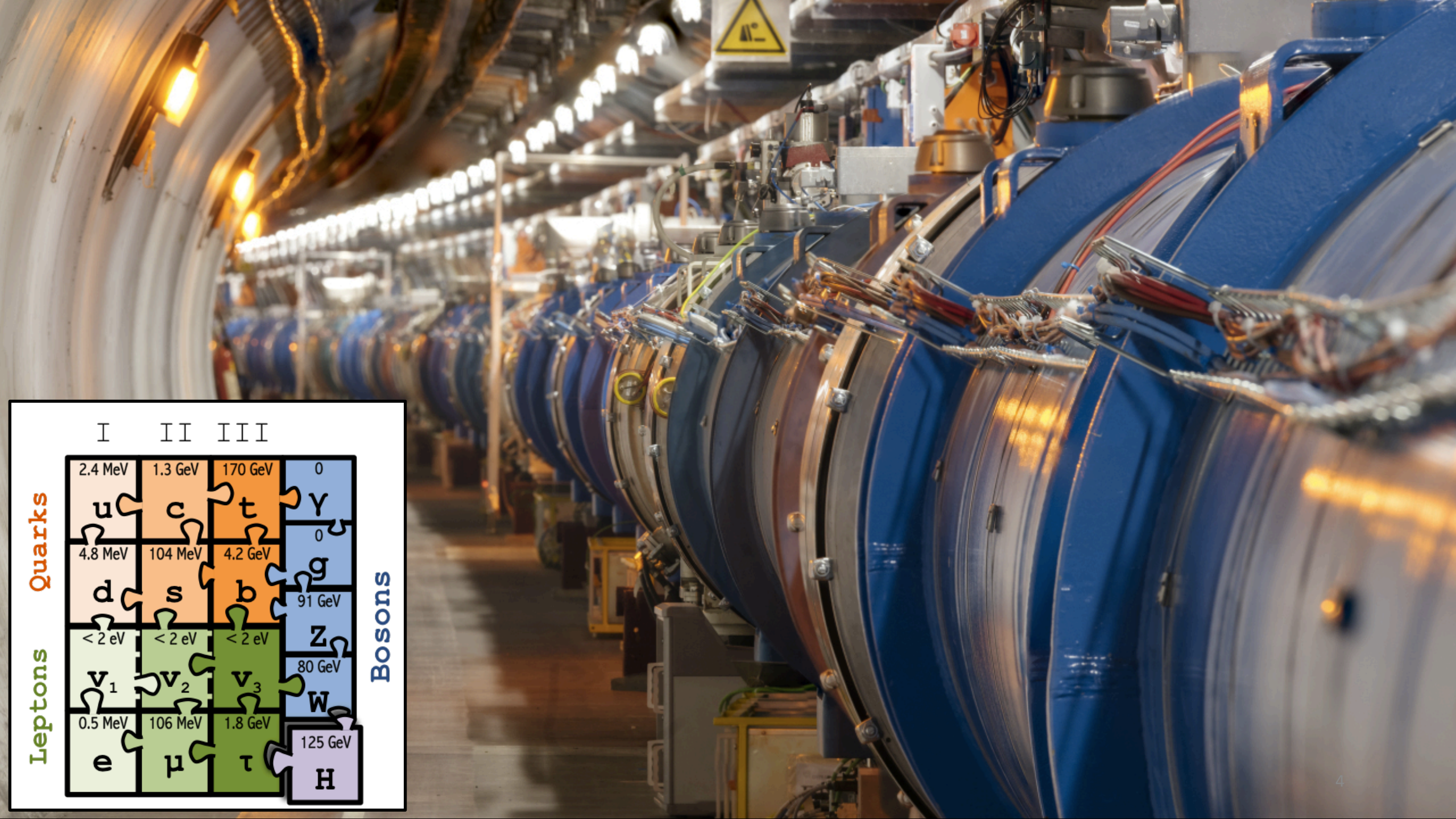
UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES
Département de physique nucléaire et corpusculaire

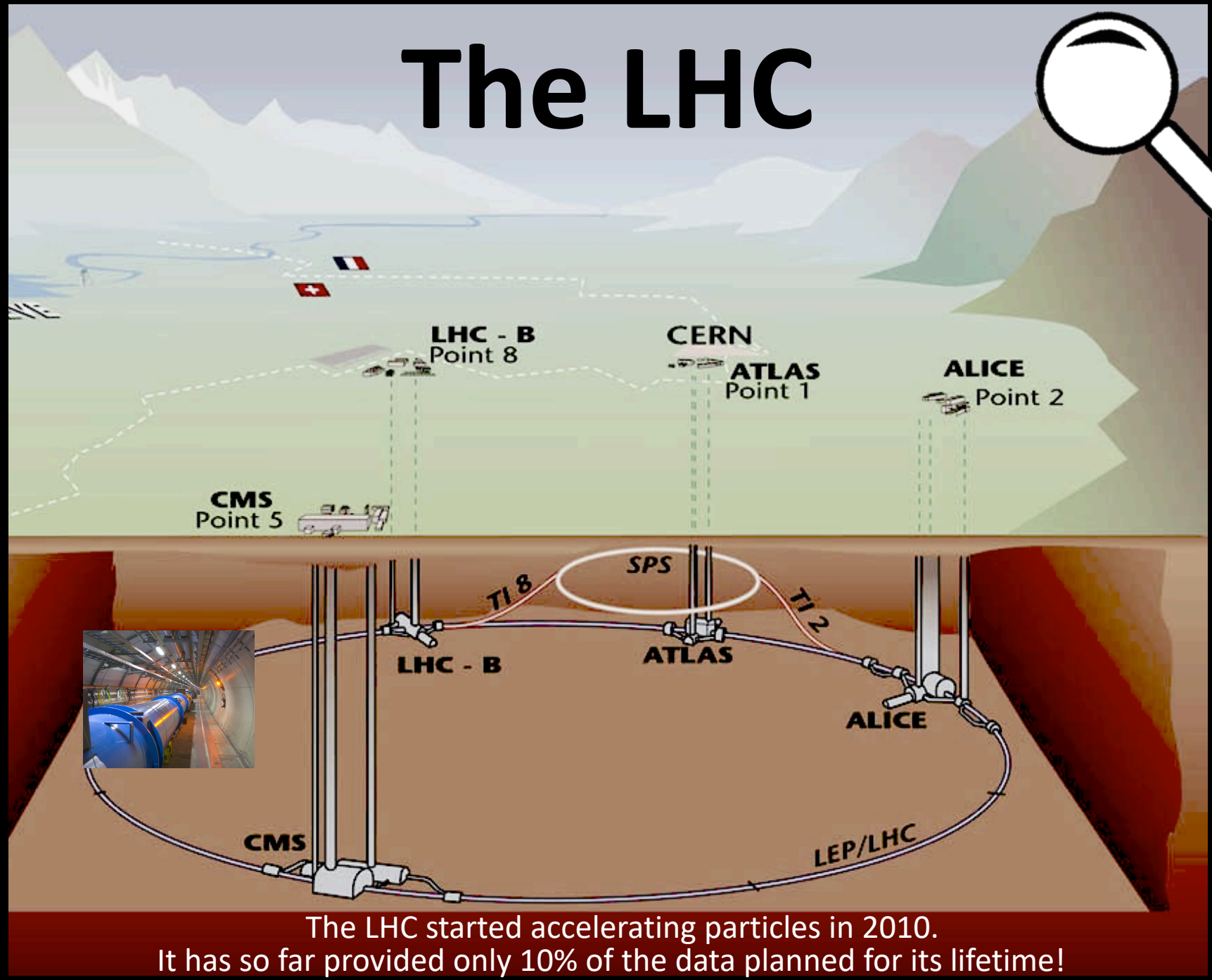
The  experiment

at the  LHC



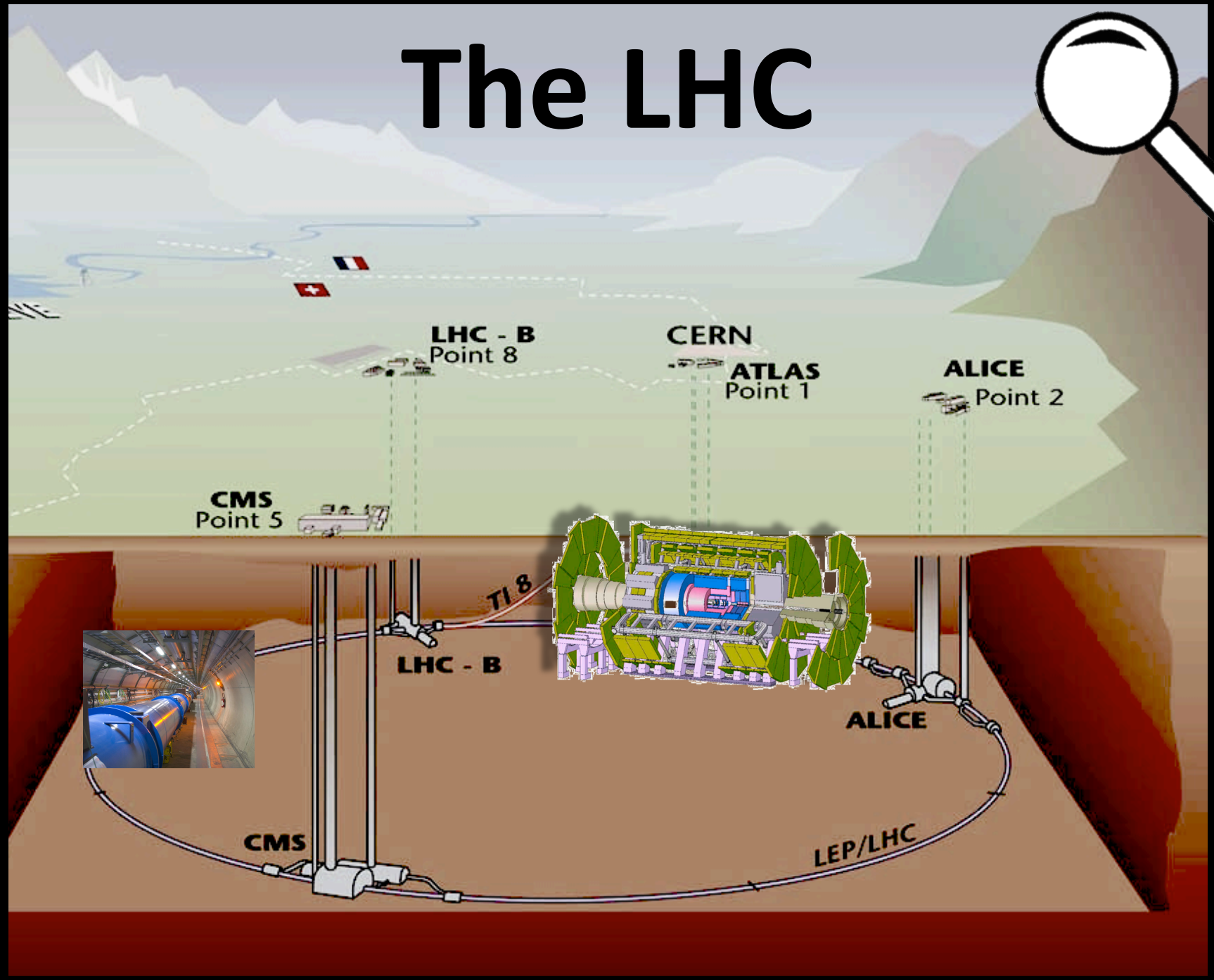

I II III
Quarks
Leptons
Bosons
2.4 MeV
u
1.3 GeV
c
170 GeV
t
0
γ
4.8 MeV
d
104 MeV
s
4.2 GeV
b
0
g
< 2 eV
ν1
< 2 eV
ν2
< 2 eV
ν3
91 GeV
Z
80 GeV
W
0.5 MeV
e
106 MeV
μ
1.8 GeV
τ
125 GeV
H

The LHC



The LHC started accelerating particles in 2010.
It has so far provided only 10% of the data planned for its lifetime!

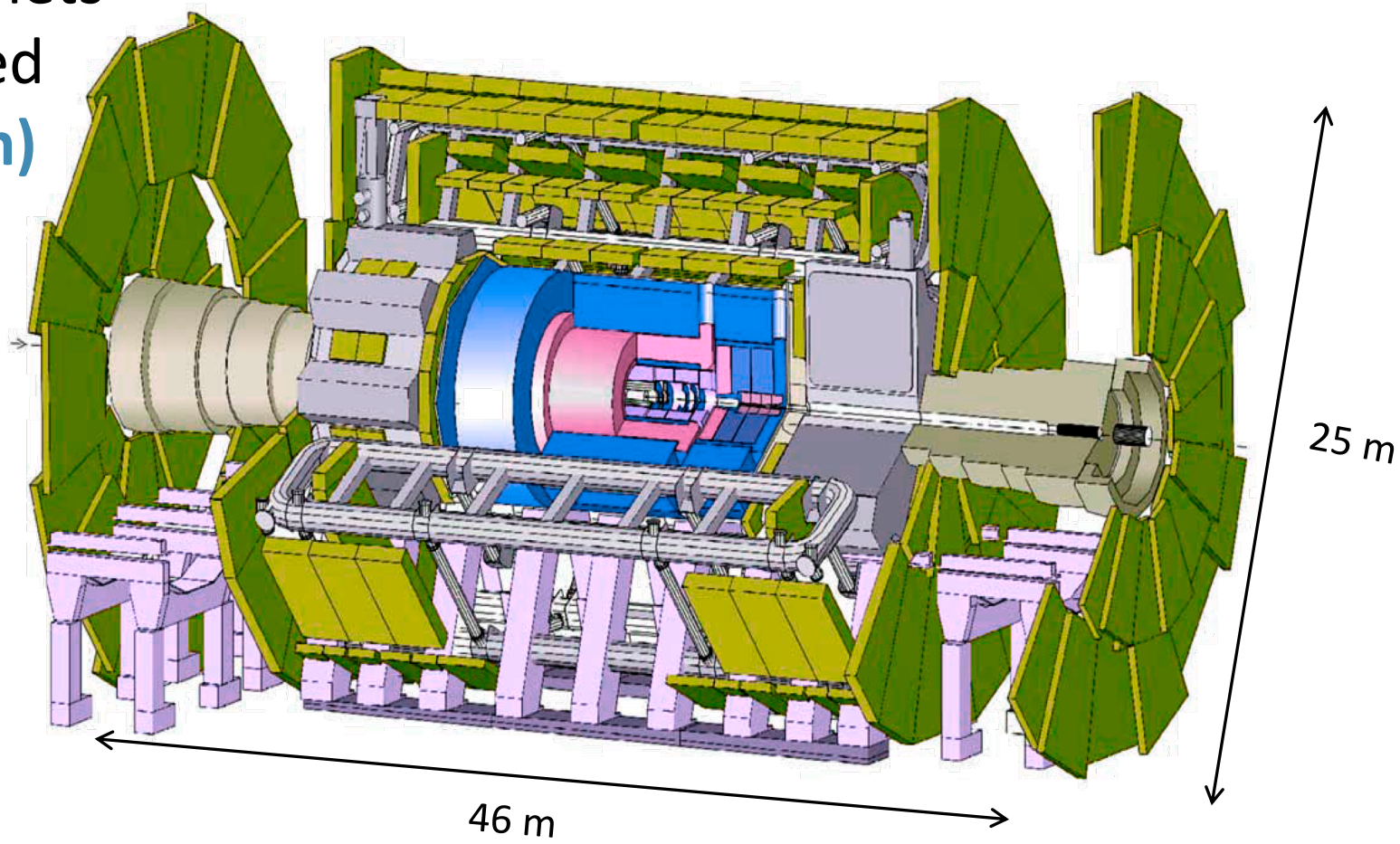
The LHC

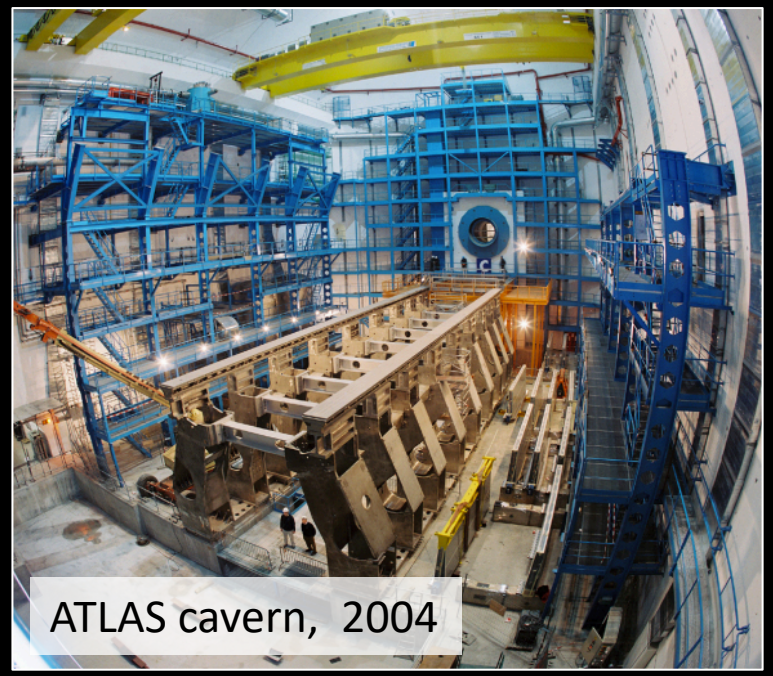

LHC - B
Point 8
CERN
ATLAS
Point 1
ALICE
Point 2
CMS
Point 5
TI 8
LHC - B
ALICE
CMS
LEP/LHC

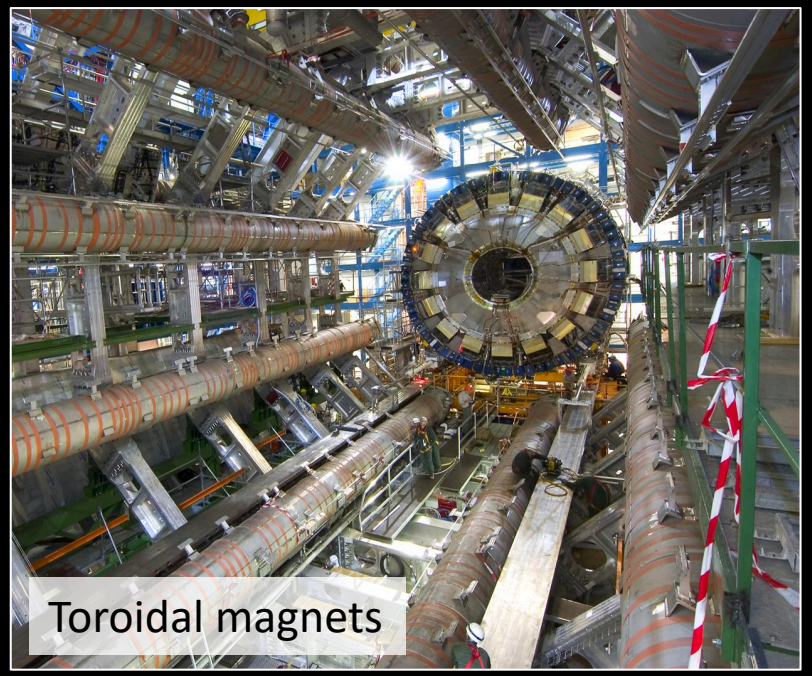
The ATLAS detector
in numbers

- ✓ Weights **7 ktonnes** ( )
- ✓ **2-4 T** superconducting magnets
- ✓ Position of particles recorded with an accuracy of **O(10µm)**
- ✓ **100 M** channels

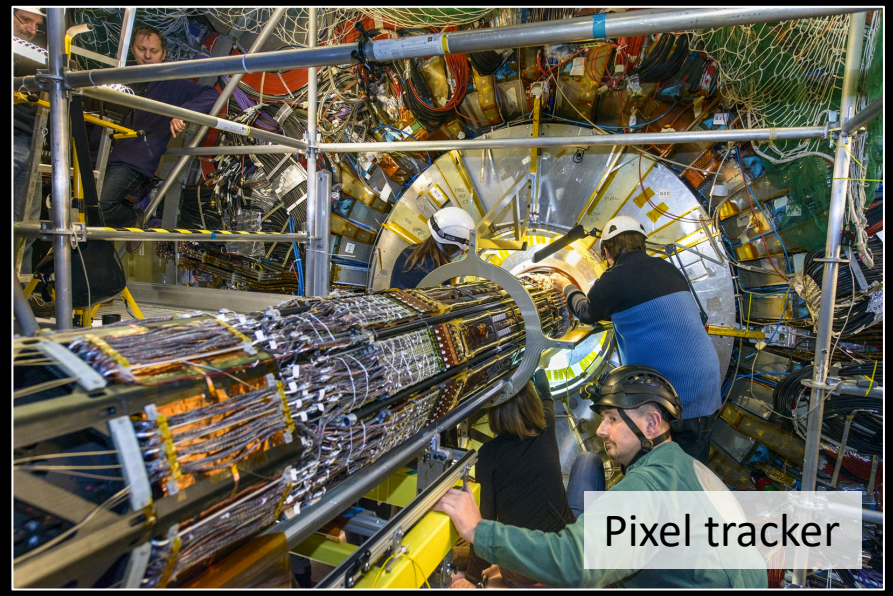
- ✓ **1 Giga** collisions/second
- ✓ **1000** events/second stored
- ✓ **500 PB** data on disk & tape
- ✓ **0.5 M** CPU cores used 24/7

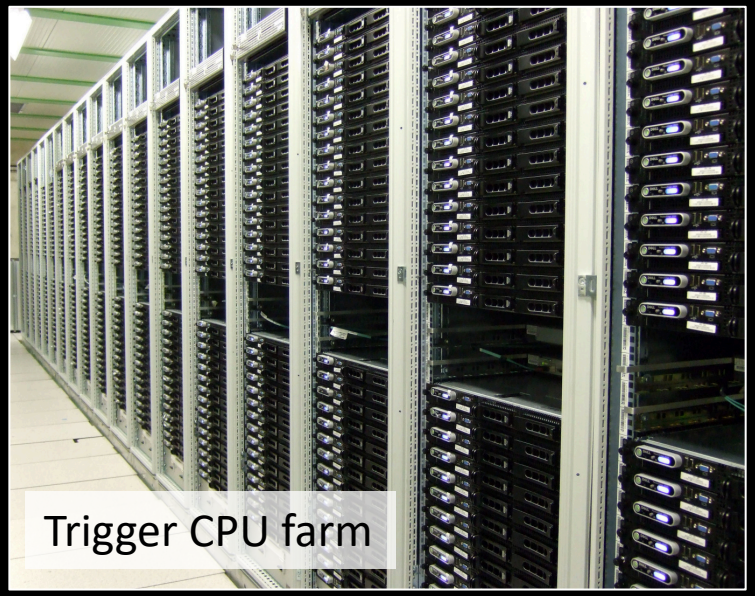



ATLAS cavern, 2004


Toroidal magnets

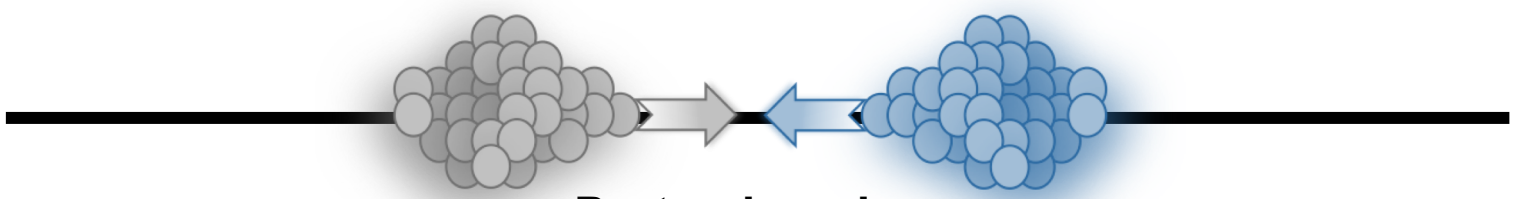

Calorimeter


Pixel tracker

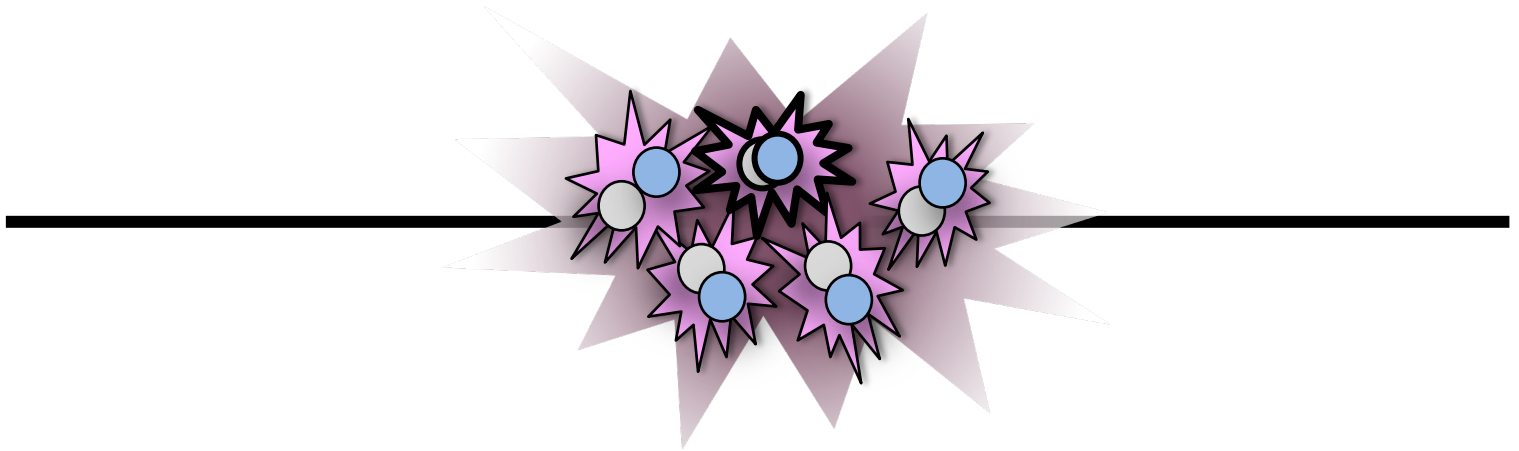

Trigger CPU farm


CERN computing center

The ATLAS experiment: collecting & using data

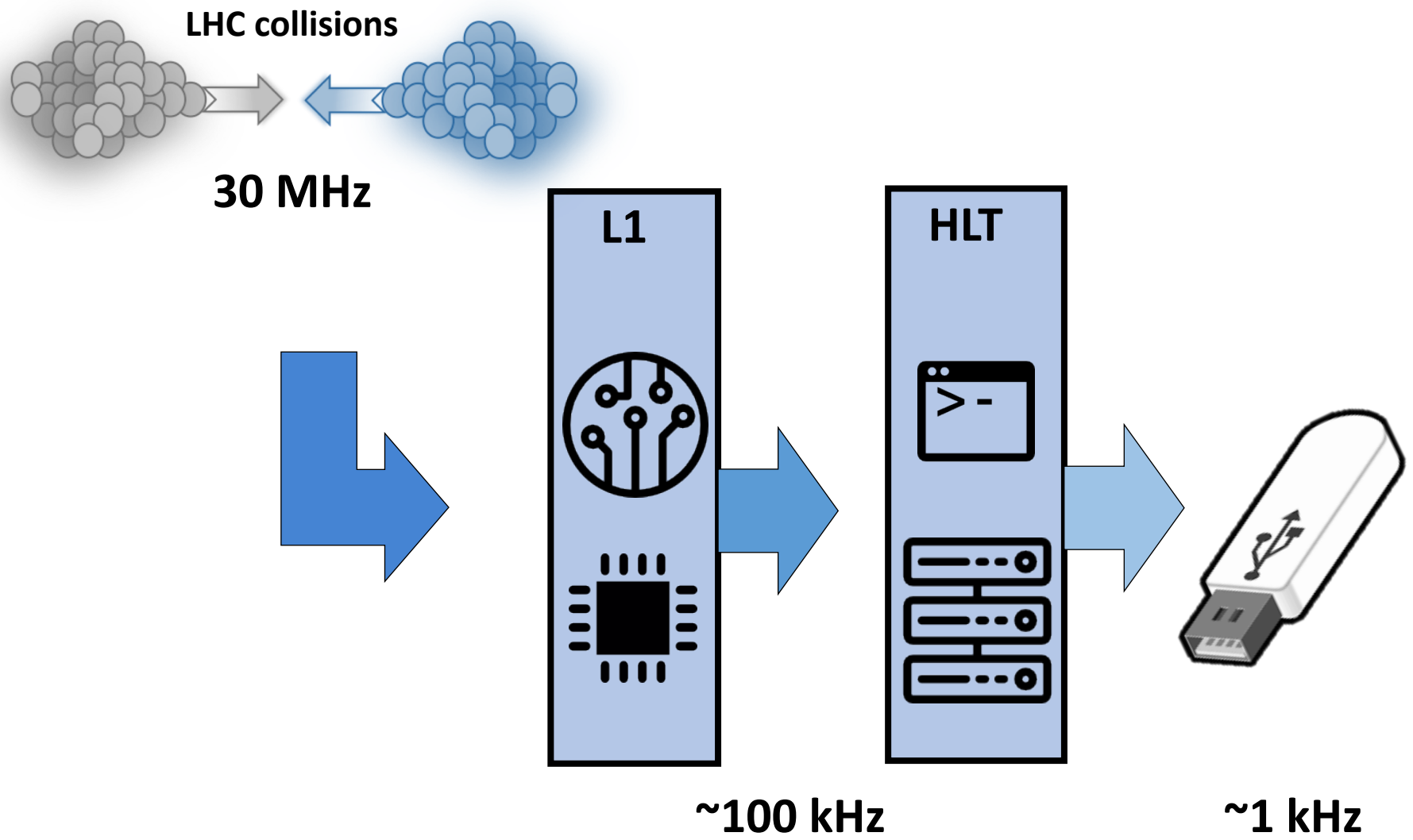


Proton bunches
>10^{11} protons/bunch
(colliding at **~30MHz in Run2)**

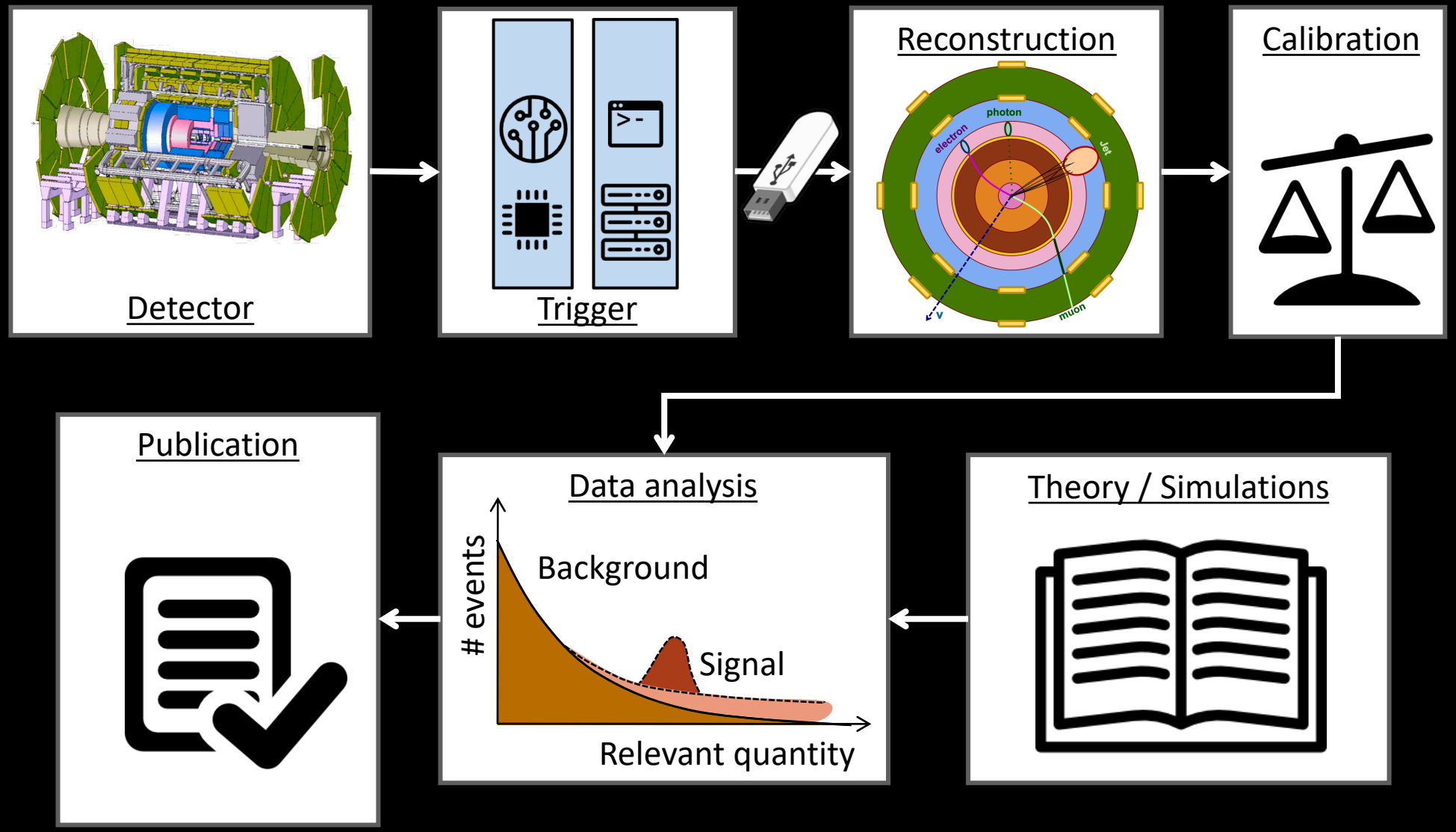


Up to 60 p-p collisions / bunch crossing

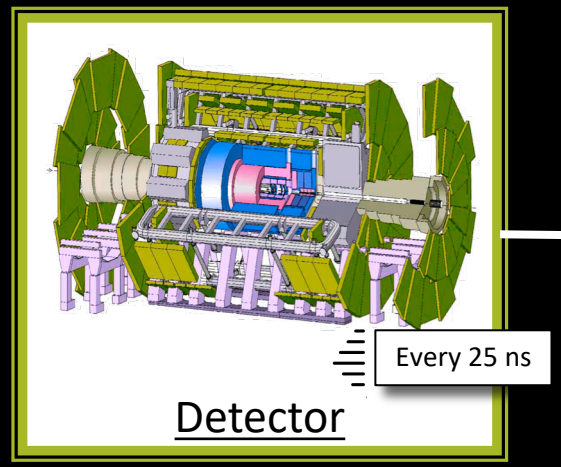
Triggering on physics

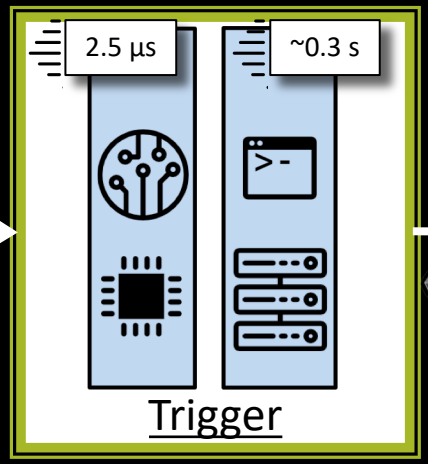


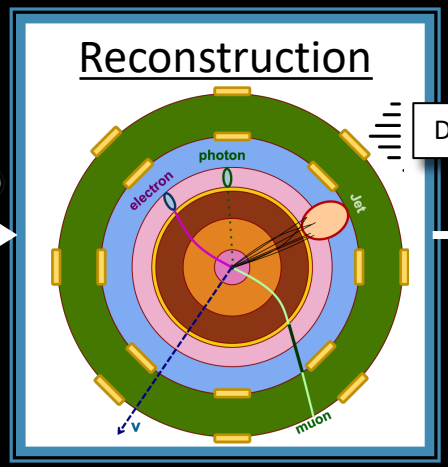
An event's lifetime

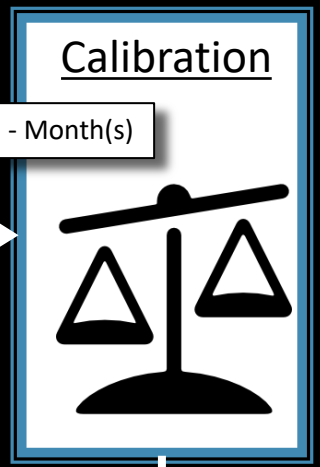


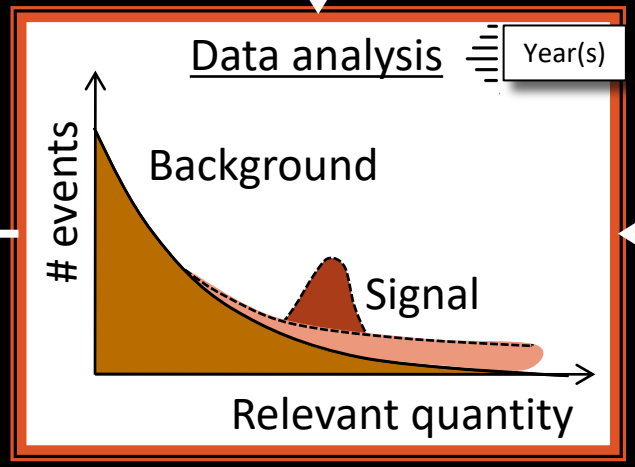
An event's lifetime

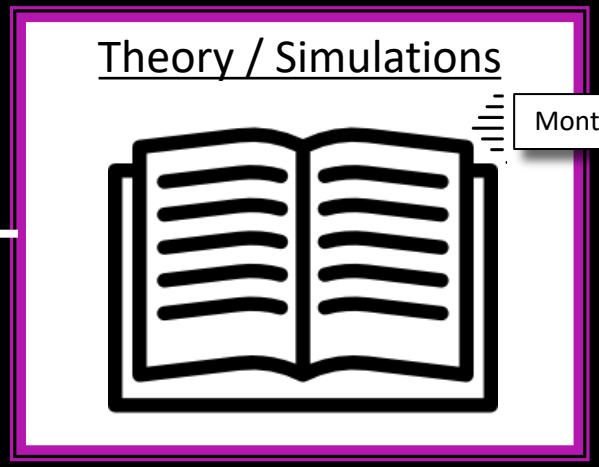


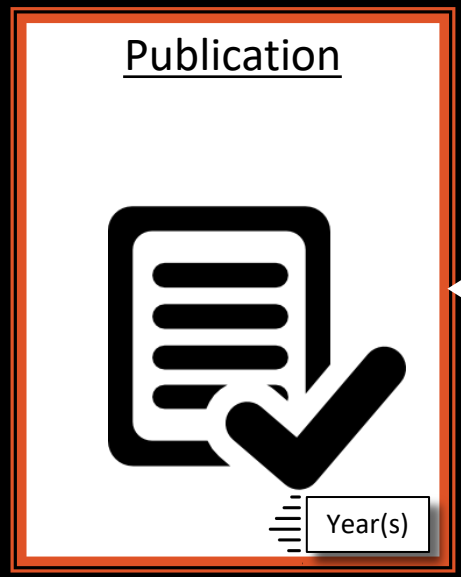


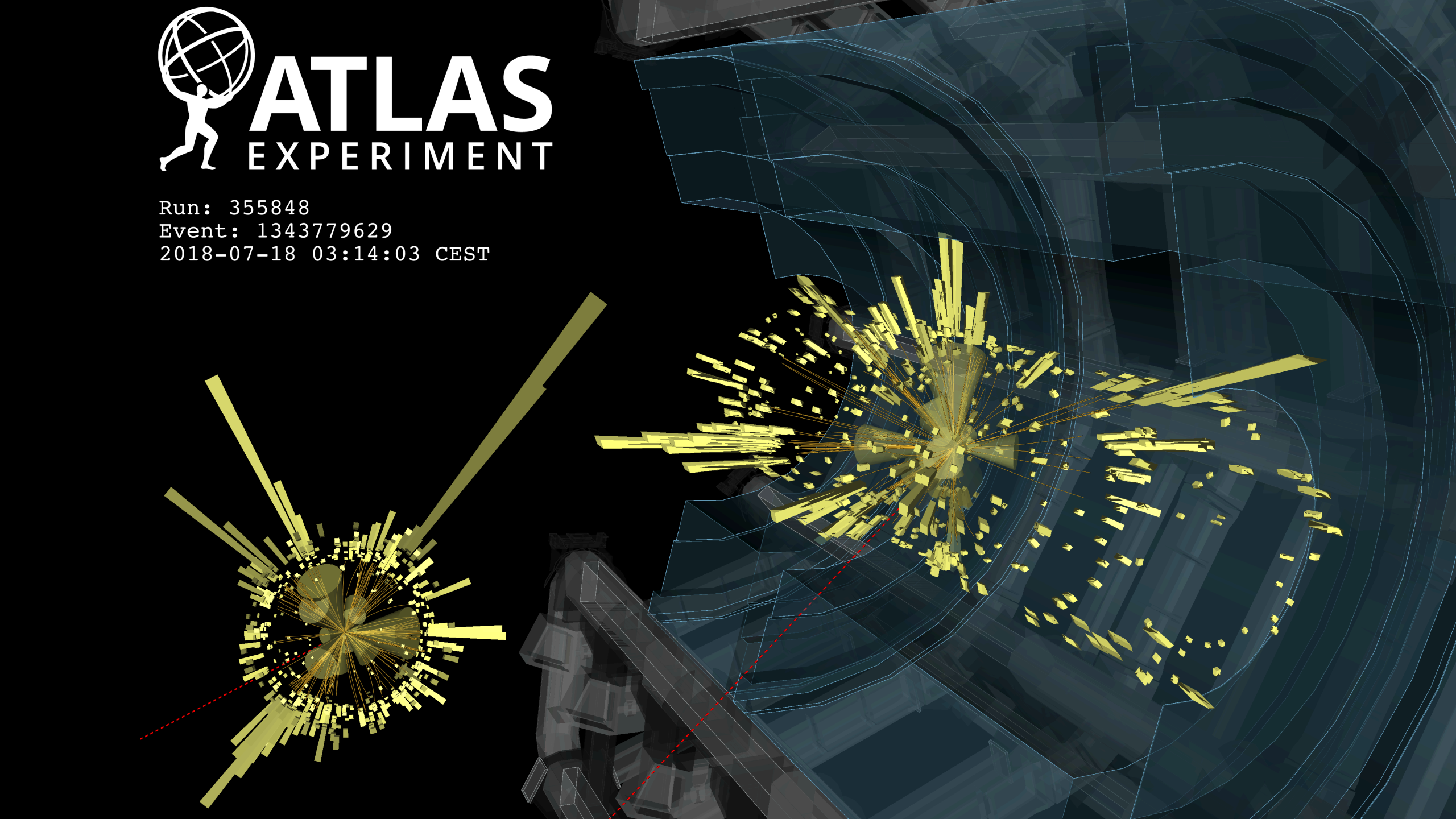


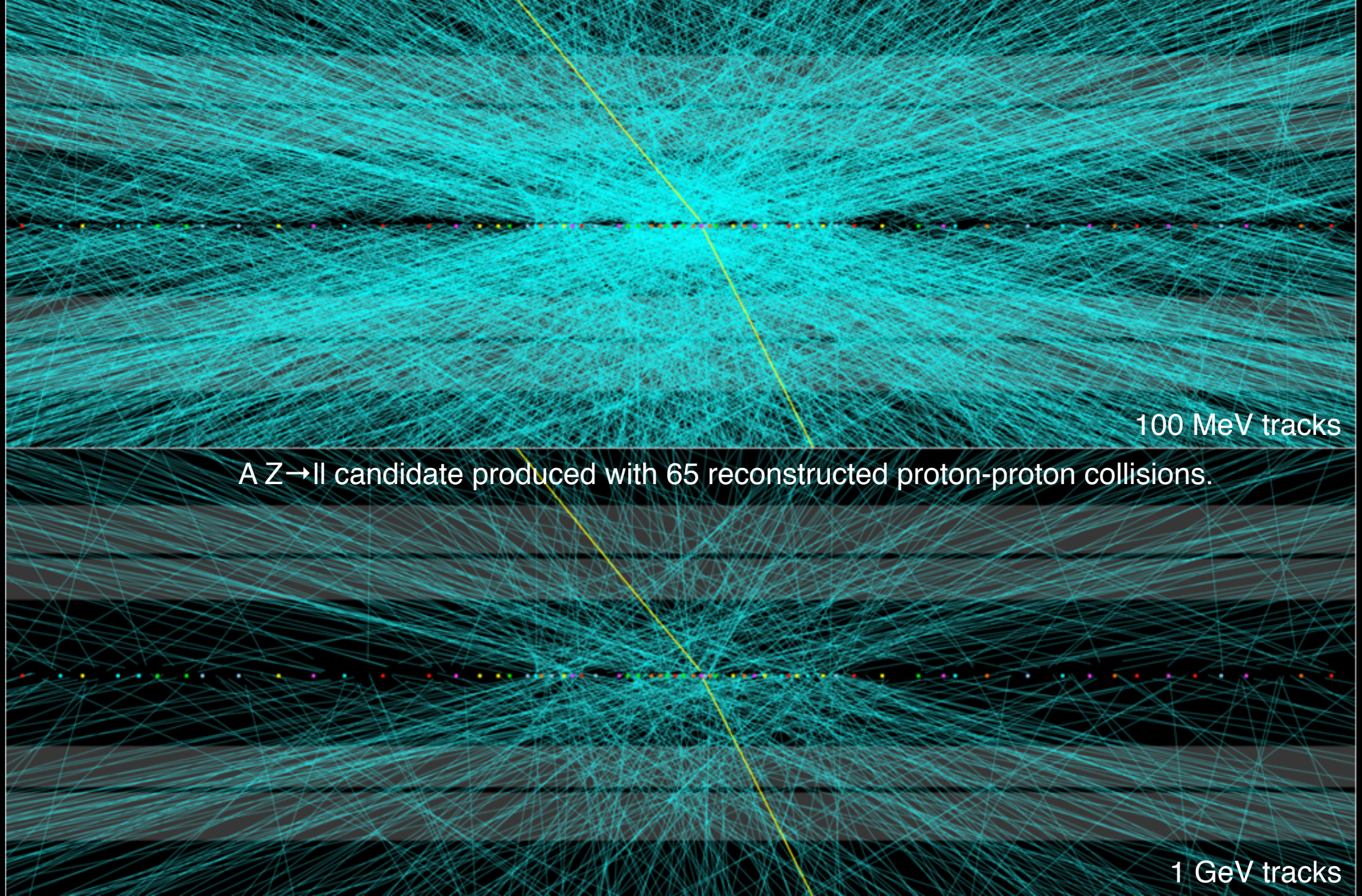






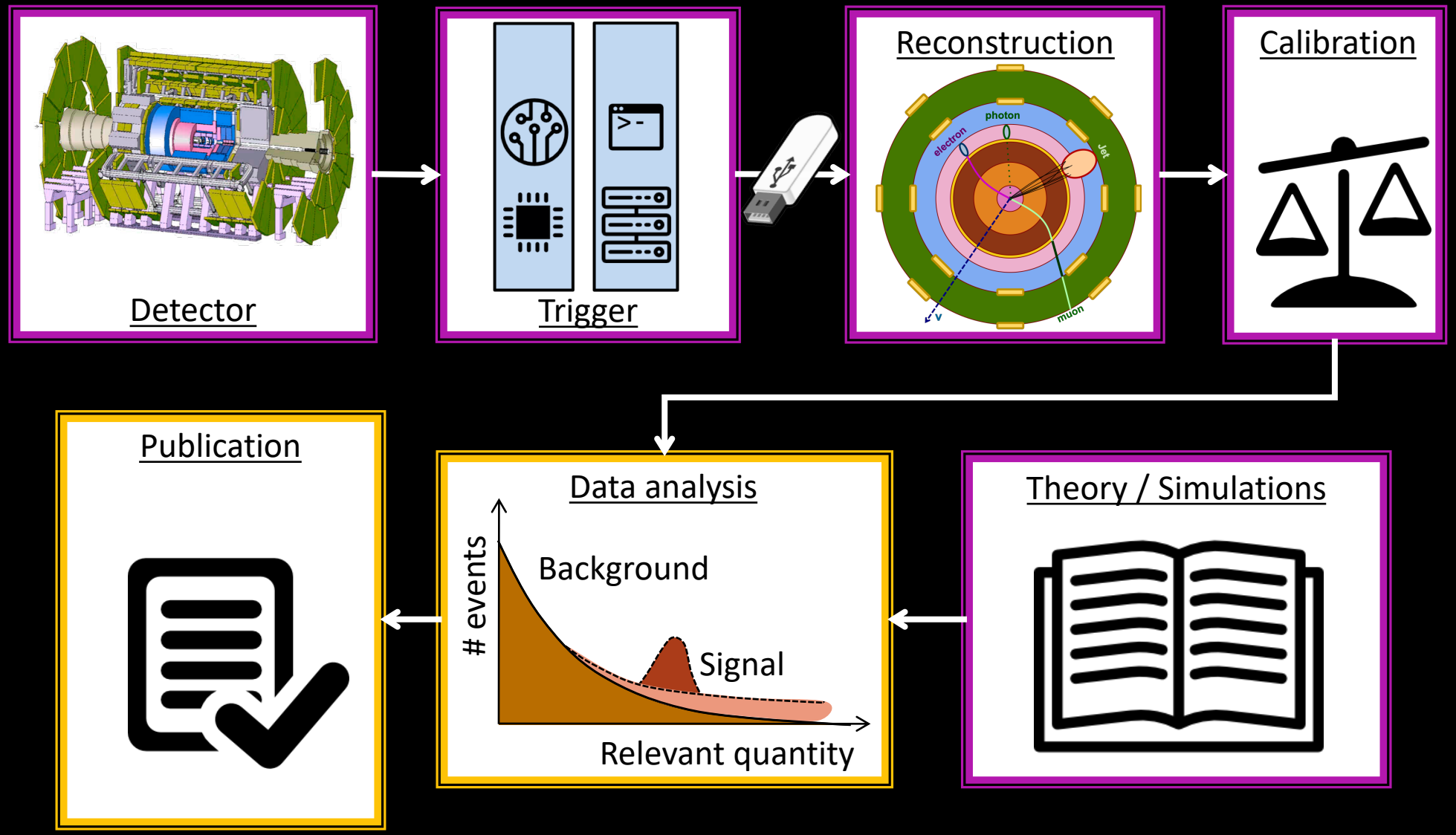



ATLAS
EXPERIMENT
Run: 355848
Event: 1343779629
2018-07-18 03:14:03 CEST



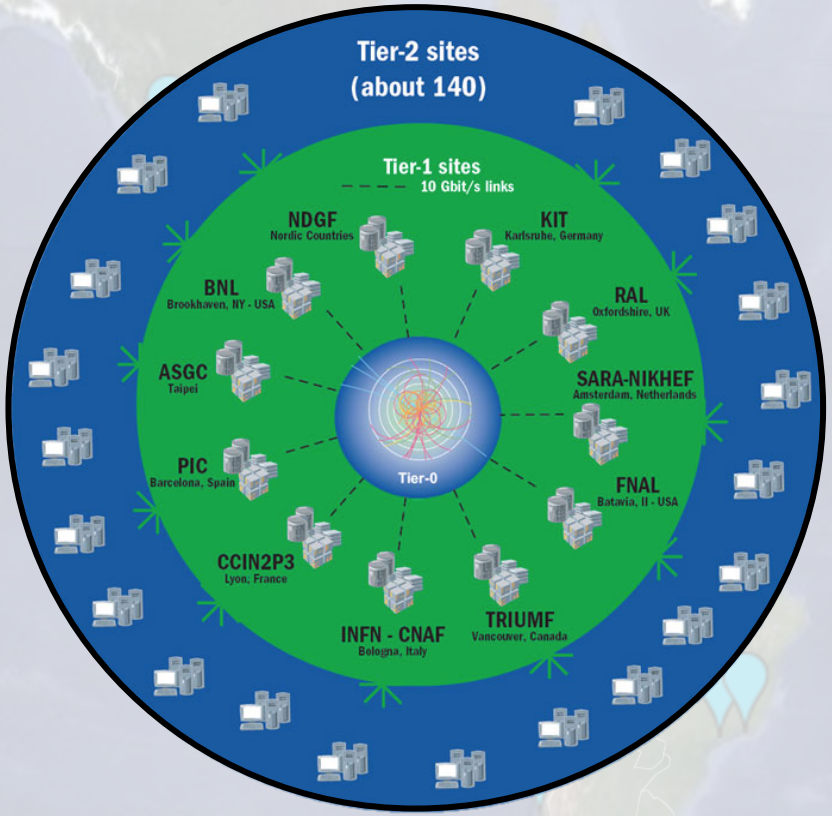
A Z→ll candidate produced with 65 reconstructed proton-proton collisions.

An event's lifetime



Worldwide LHC Computing Grid
an international collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists.



- 161 sites, 42 countries
- 1 M CPU cores
- 1 EB of storage
- > 2 M jobs/day
- > 100 PB moved/month
- accessed by 10k users
- 10-100 Gb links

Network proved better than anyone imagined: Any job can run anywhere

ATLAS data management

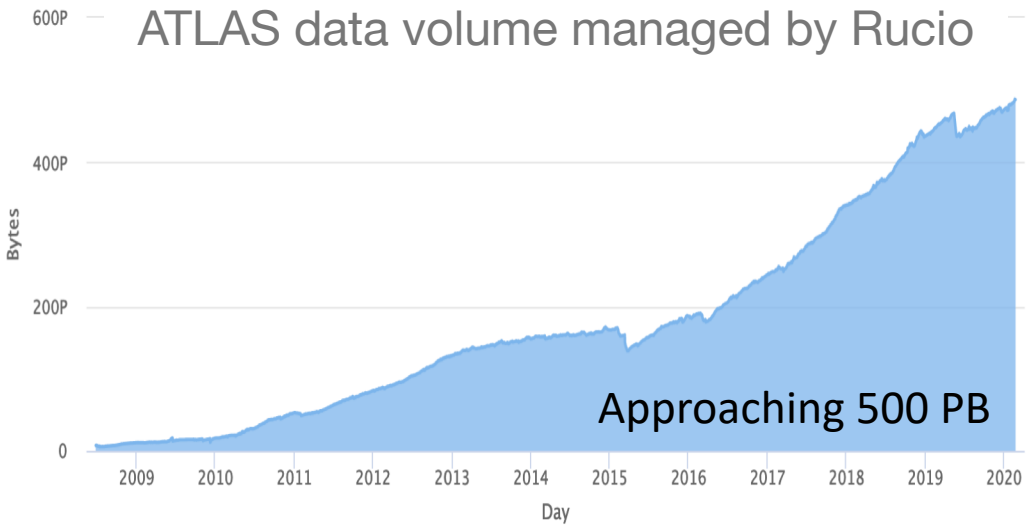
RUCIO

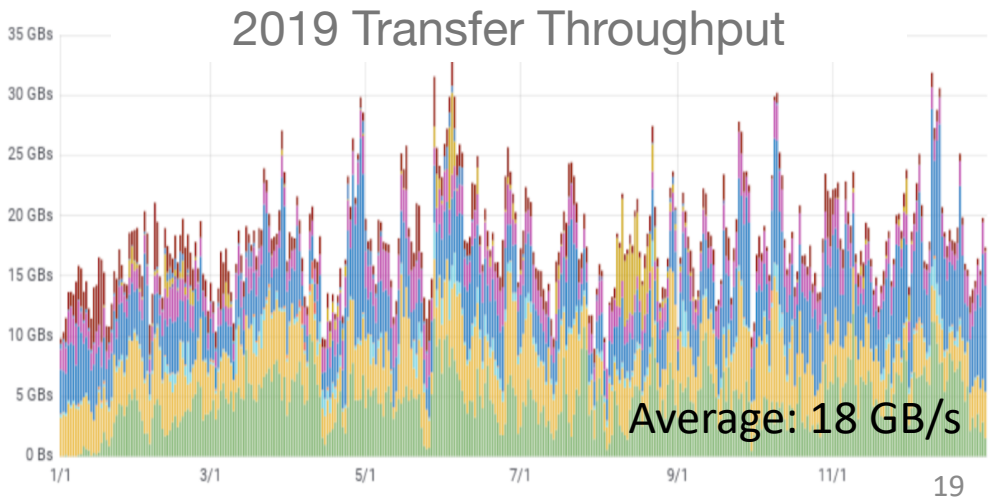
Data storage
Access
Replication
Deletion

Scalable
Policy-driven
Monitorable
Supporting “FAIR” data principles

Findable
Accessible
Interoperable
Reusable

Now established in the HEP community and beyond





Hardware


Magnetic tapes, retrieved by robotic arms, are used for long-term storage

Storage

Tape (at CERN) about 270 PB	• Most reliable and cost-effective technology for large-scale archiving • Data stored there infinitely
Disk about 200 PB	• Data for initial processing • Copies for further processing / user analysis • Data in disks gets staged from tape, on demand

Processing power

CPUs	• Mainly GRID • About 400k cores
GPUs	• Mostly for RnD • Few 10s *Also considering for the future: FPGA accelerators*
Opportunistic resources	• Online farm, 100k cores • High Performance Computers, primarily in the US • Volunteer computing (see later talk)


Nvidia GeForce

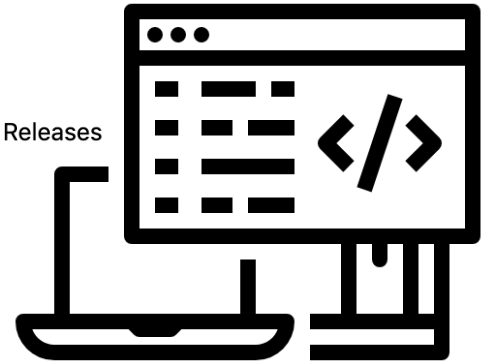
LHC @home Volunteer computing for the LHC

Software

athena
Project ID: 53790

70,356 Commits | 34 Branches | 1,374 Tags | 2.6 GB Files | 2.6 GB Storage | 124 Releases

The ATLAS Experiment's main offline software repository



- **All software organized in packages in Git.** For example:
  https://gitlab.cern.ch/atlas/athena
- **All software open source, copyrighted and licenced (Apache 2)**
  - "Copyright (C) 2002-2020 CERN for the benefit of the ATLAS collaboration"
  - For open use – but also for crediting developers who move out of academia
- **Thorough tracking of software developments a key of success**
  - Via the Jira software, supported by CERN IT Jira Software
  - Multiple releases exist for merging of new code with existing one
  - Automated tools run nightly to verify code sanity & performance
  - Globally the software projects are coordinated with careful planning
- **Software Tools**
  - Databases
  - Analysis tools: ROOT is the workhorse!

ROOT
Data Analysis Framework

- **Analysis-specific software developed by teams available to whole collaboration!**

The ATLAS collaboration

The ATLAS Collaboration

3000
Scientific authors

38
Countries

180
Institutions

1200
Doctoral students

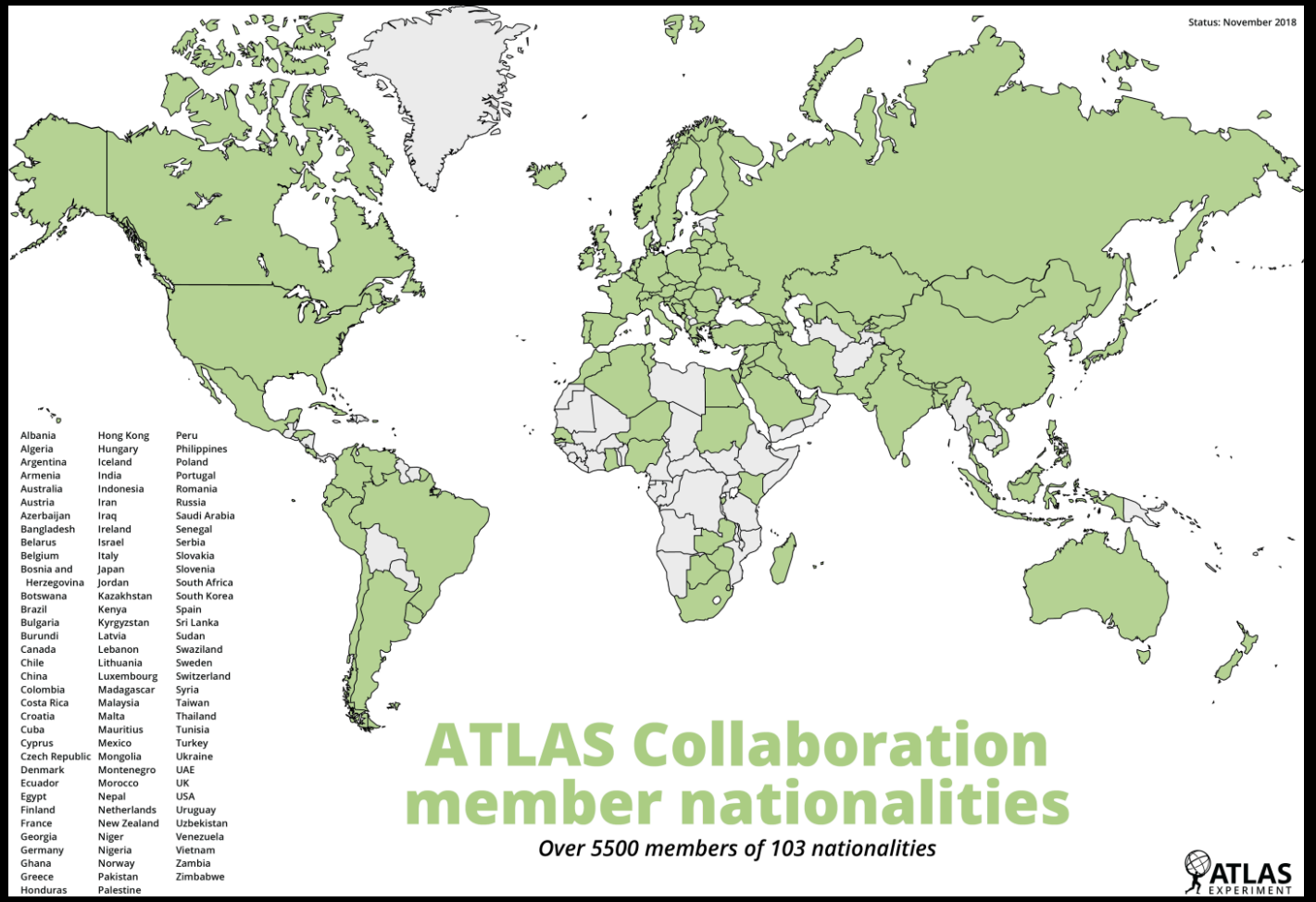
The ATLAS Collaboration

3000
Scientific authors

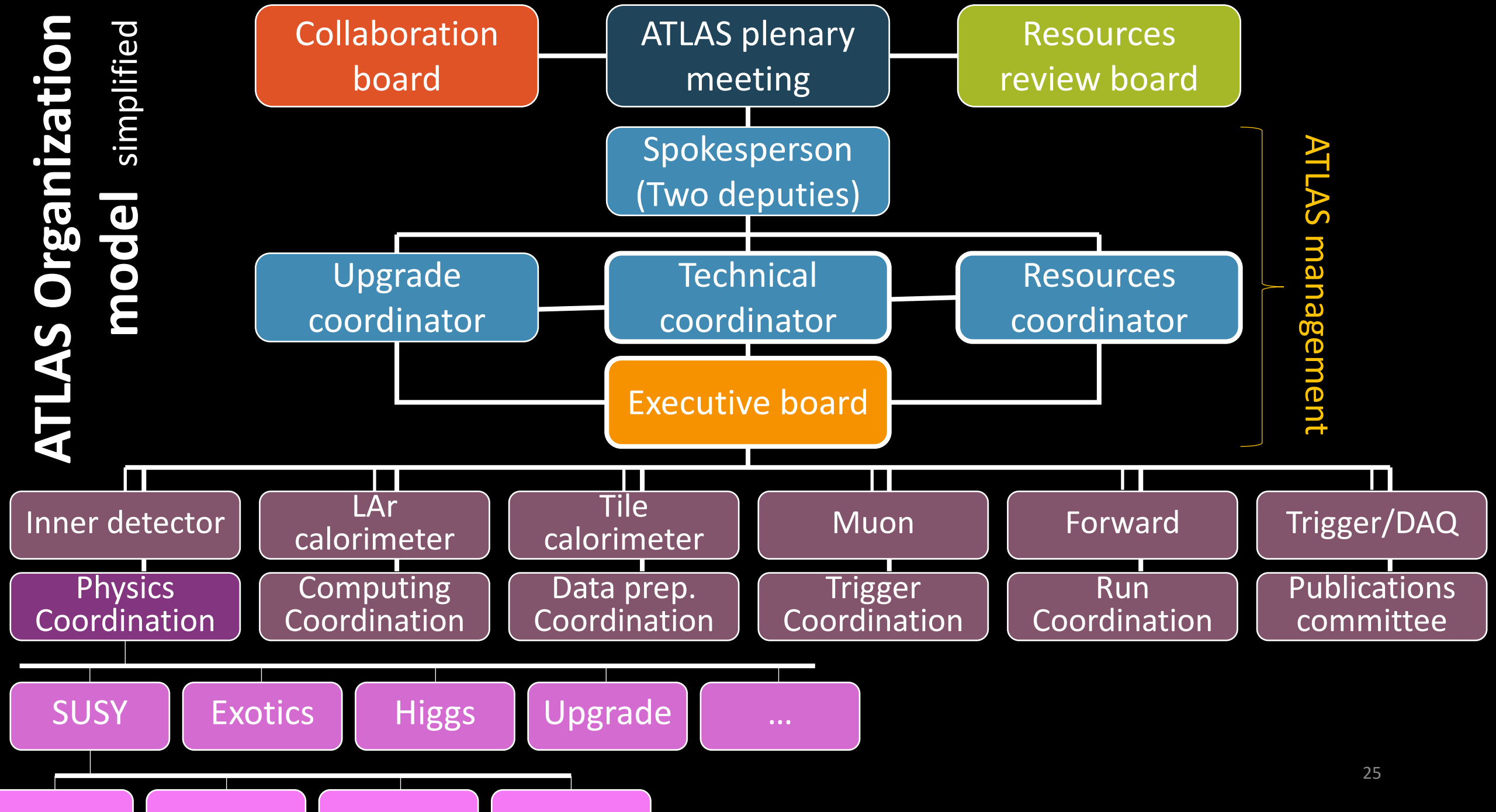
38
Countries

180
Institutions

1200
Doctoral students

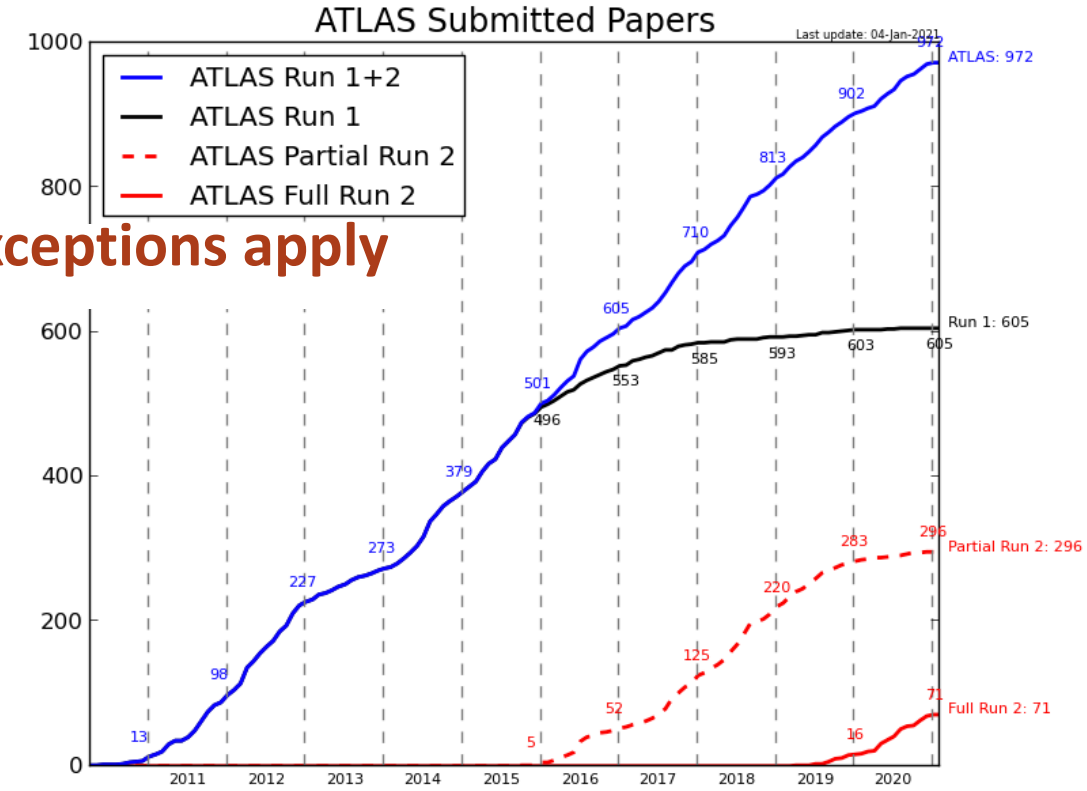


ATLAS Organization model simplified



ATLAS author-ship/list

- **Only ATLAS authors sign ATLAS papers; Exceptions apply**
- **All authors sign all papers**



[Submitted on 13 Aug 2020 (v1), last revised 20 Nov 2020 (this version, v2)]

Search for new phenomena in final states with large jet multiplicities and missing transverse momentum using $\sqrt{(s)} = 13$ TeV proton−proton collisions recorded by ATLAS in Run 2 of the LHC

ATLAS Collaboration

Results of a search for new particles decaying into eight or more jets and moderate missing transverse momentum are presented. The analysis uses 139 fb^{-1} of proton−

The ATLAS Collaboration

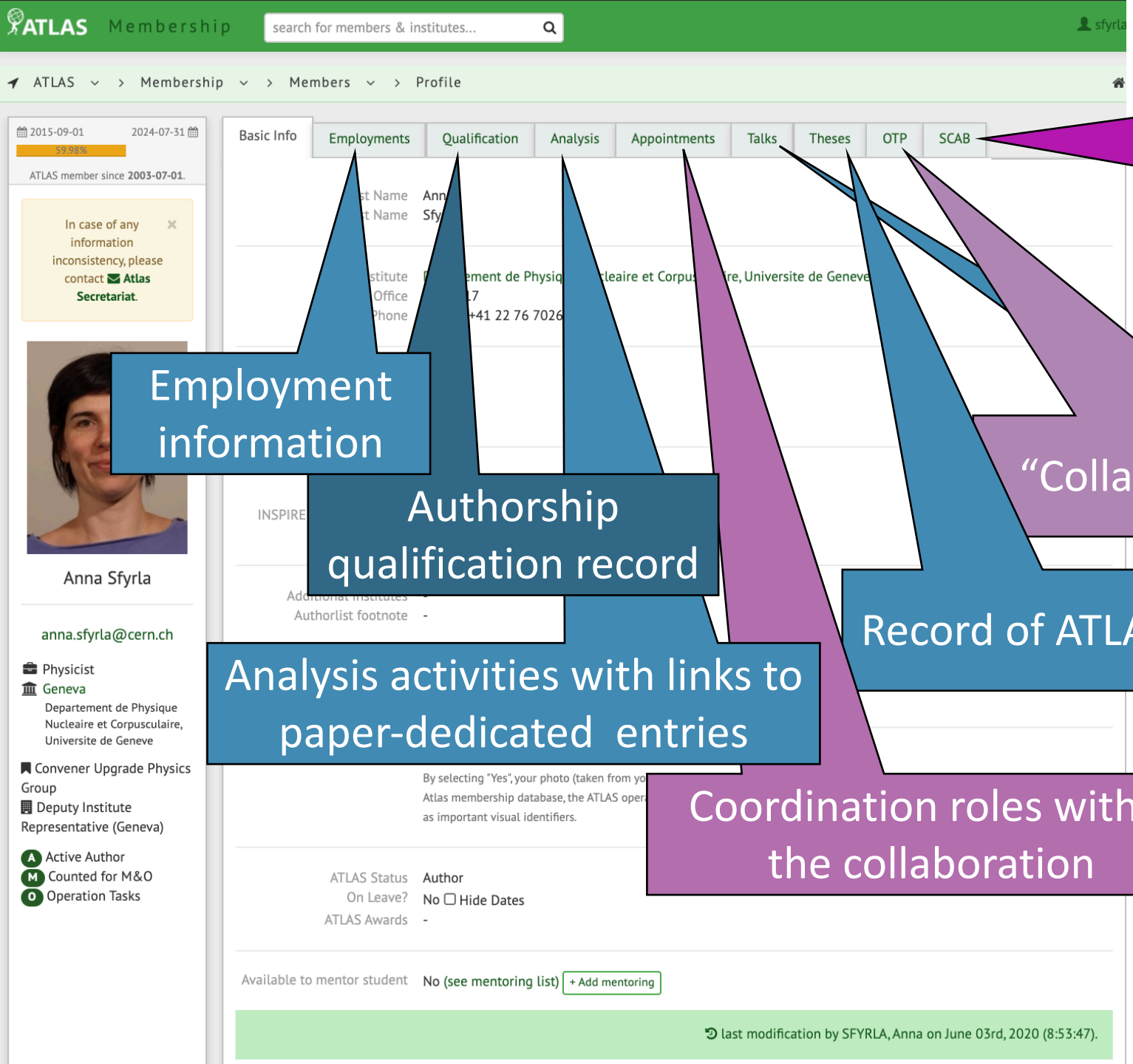
G. Aad[102], B. Abbott[128], D.C. Abbott[103], A. Abed Abud[36], K. Abeling[53], D.K. Abhayasinghe[94], S.H. Abidi[166], O.S. AbouZeid[40], N.L. Abraham[155], H. Abramowicz[160], H. Abreu[159], Y. Abulaiti[6], B.S. Acharya[67a,67b,n], B. Achkar[53], L. Adam[100], C. Adam Bourdarios[5], L. Adamczyk[84a], L. Adamek[166], J. Adelman[121], M. Adersberger[114], A. Adiguzel[12c], S. Adorni[54], T. Adye[143], A.A. Affolder[145], Y. Afik[159], C. Agapopoulou[65], M.N. Agaras[38], A. Aggarwal[119], C. Agheorghiesei[27c], J.A. Aguilar-Saavedra[139f,139a,ad],

… 10 pages later …

D. Zhong[172], B. Zhou[106], C. Zhou[180], H. Zhou[7], M.S. Zhou[15a,15d], M. Zhou[154], N. Zhou[60c], Y. Zhou[7], C.G. Zhu[60b], C. Zhu[15a,15d], H.L. Zhu[60a], H. Zhu[15a], J. Zhu[106], Y. Zhu[60a], X. Zhuang[15a], K. Zhukov[111], V. Zhulanov[122b,122a], D. Zieminska[66], N.I. Zimine[80], S. Zimmermann[52], Z. Zinonos[115], M. Ziolkowski[150], L. Živković[16], G. Zobernig[180], A. Zoccoli[23b,23a], K. Zoch[53], T.G. Zorbas[148], R. Zou[37], L. Zwalinski[36].

A new collaborator becomes an author if:

- Have been a *qualifying ATLAS member* for *at least one year*.
- Not be an author of another major LHC collaboration at the time of application.
- Have spent *at least 80 working days* doing **pre-agreed** ATLAS technical work.



Nominations and prioritization for ATLAS talks
NB: Each collaborator gives about 1 talk every 3 years !

Employment information

Authorship qualification record

“Collaborative work” credit

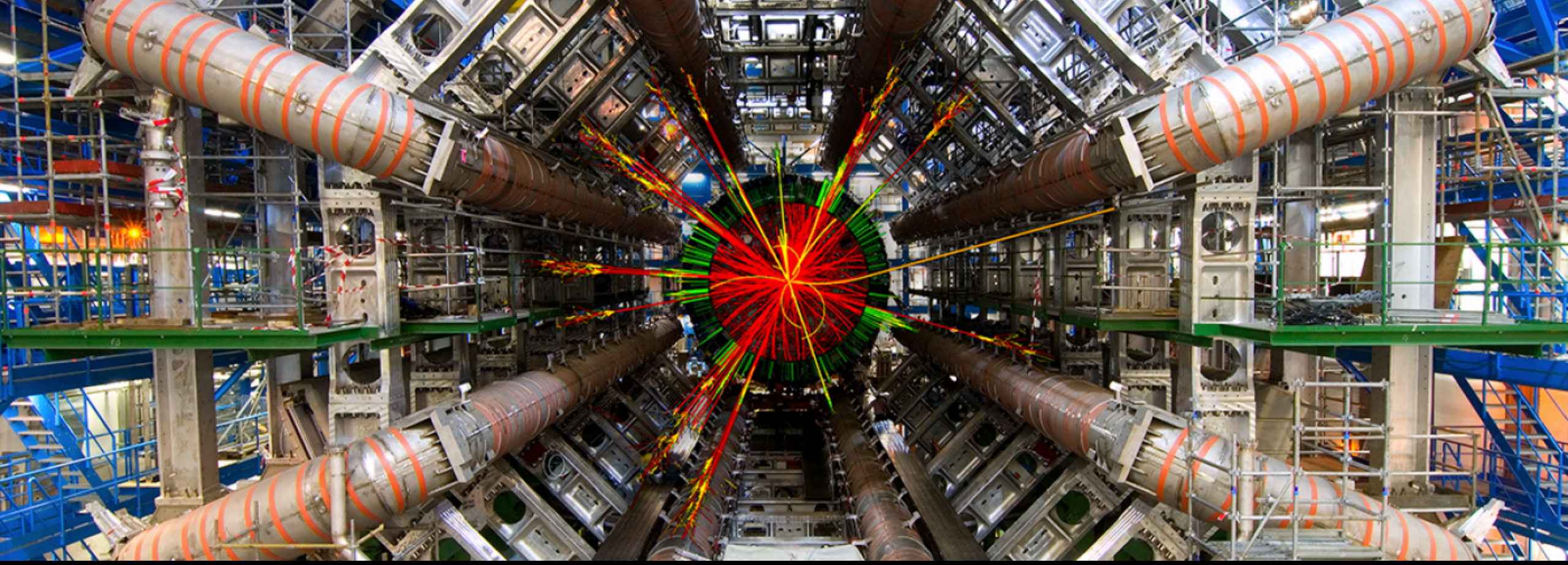
Record of ATLAS thesis

Talks given on behalf of the collaboration

Analysis activities with links to paper-dedicated entries

Coordination roles within the collaboration

Recognition is a tricky matter in a large collaboration!



The ATLAS Open Data

Why? Guarantee openness and preservation of experimental data

New open data policy in support of open science from CERN & the LHC experiments

Peer-reviewed publications

- Open Access
- Followed by detailed data related to the results, available at hepdata.net

Purpose: Communicate results and maximize their scientific value

Reconstructed & calibrated data

- Followed by related metadata
- Accompanied by appropriate simulated data samples

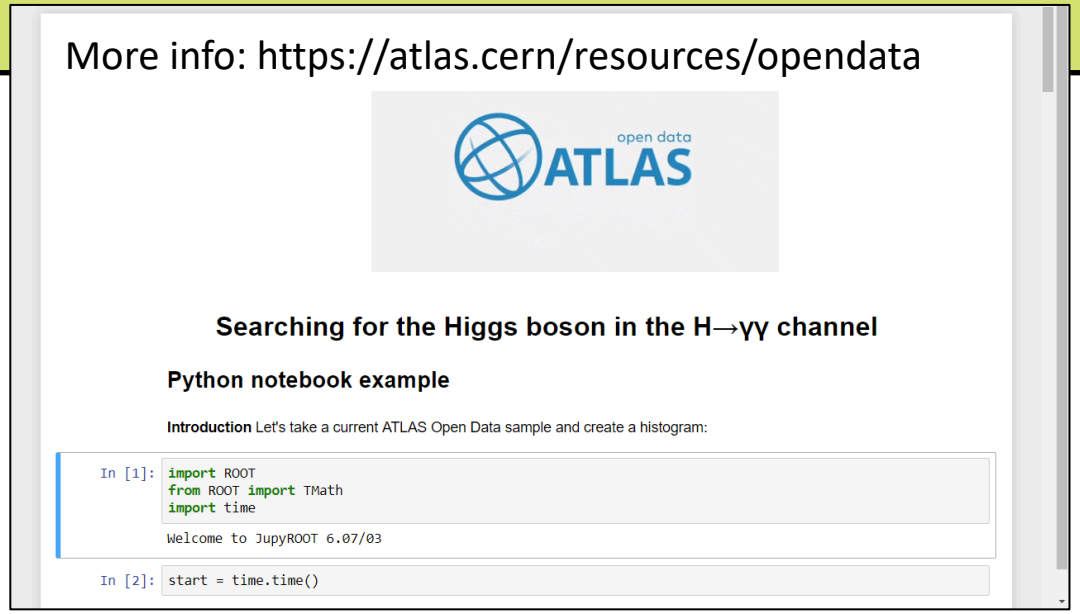
Purpose: Algorithmic, performance and physics studies

Data for outreach and education

- Selected and formatted ("light") datasets
- Examples available in Jupyter notebooks
- Used in university classes, in growing numbers

Purpose: Maximize educational impact

More info: https://atlas.cern/resources/opendata



In brief

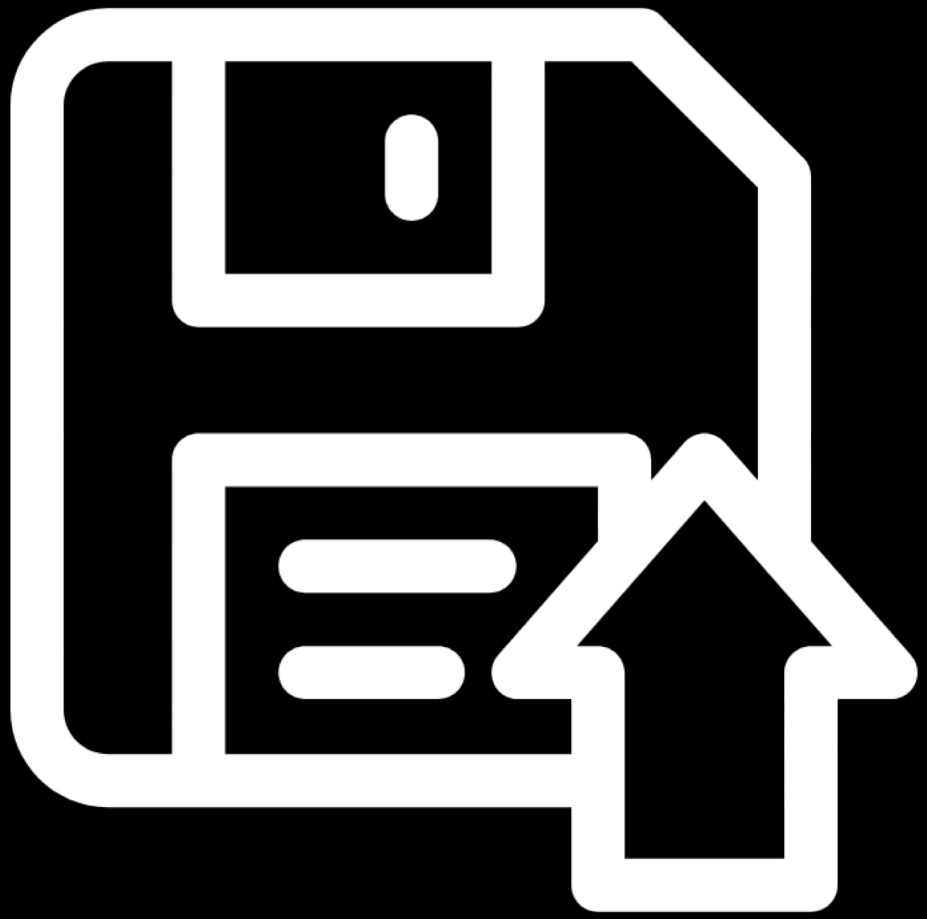
- The ATLAS experiment at the CERN LHC handles about **500 PB of data**
- The data are analysed by an **international collaboration** of 3000 scientific authors
- **Collaborative tools** allow for smooth access and use of these data, as well as for handling of collaborative matters
- ATLAS Open Data is available for outreach, education and **(open) science**

THANK YOU

for your attention

and my ATLAS colleagues for material

In this presentation, many icons come from flaticon.com



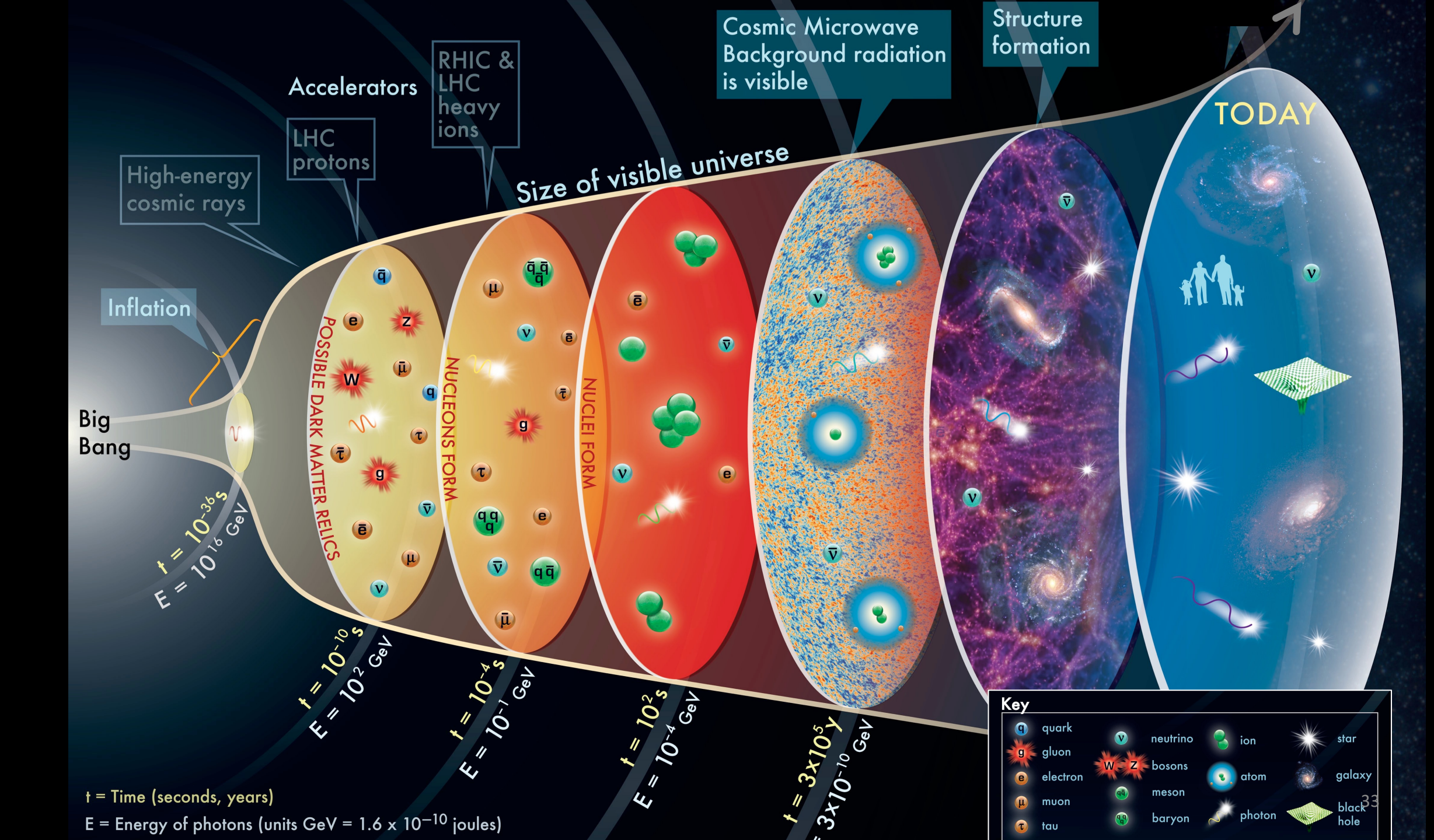
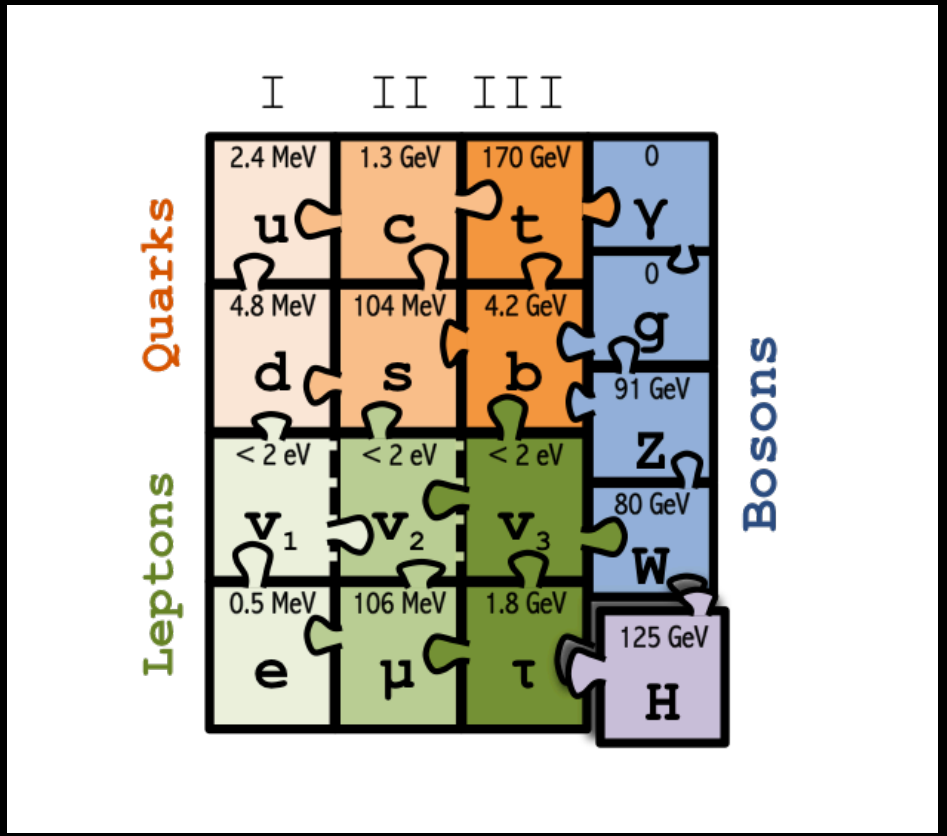
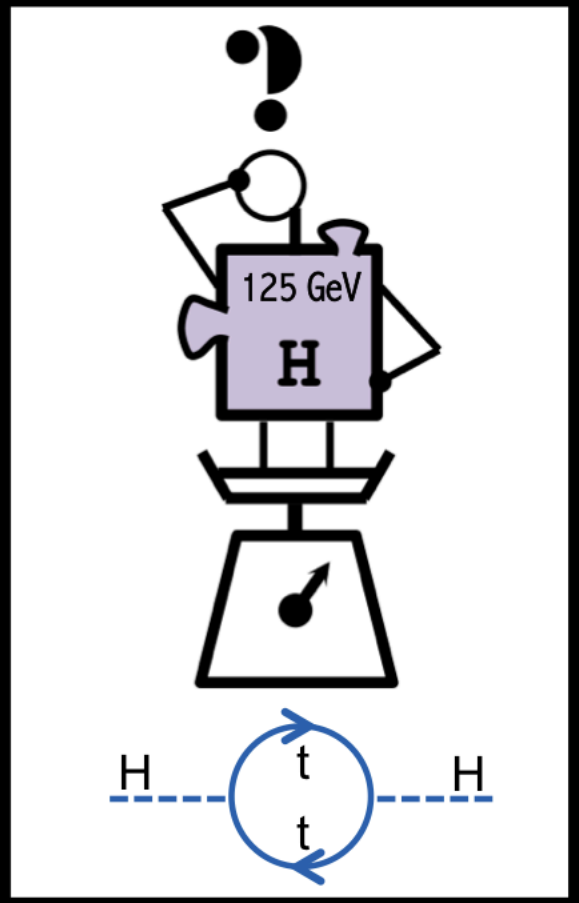
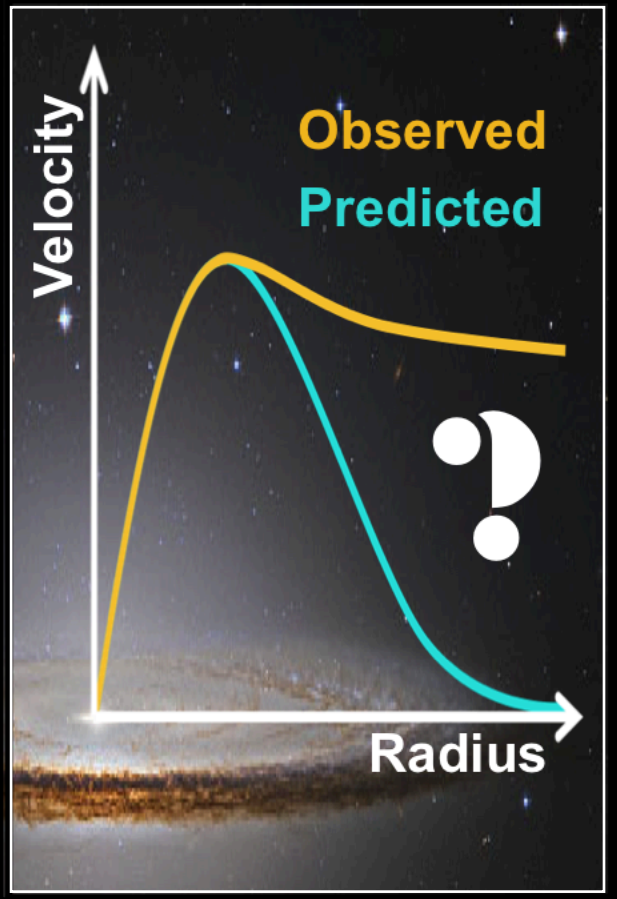
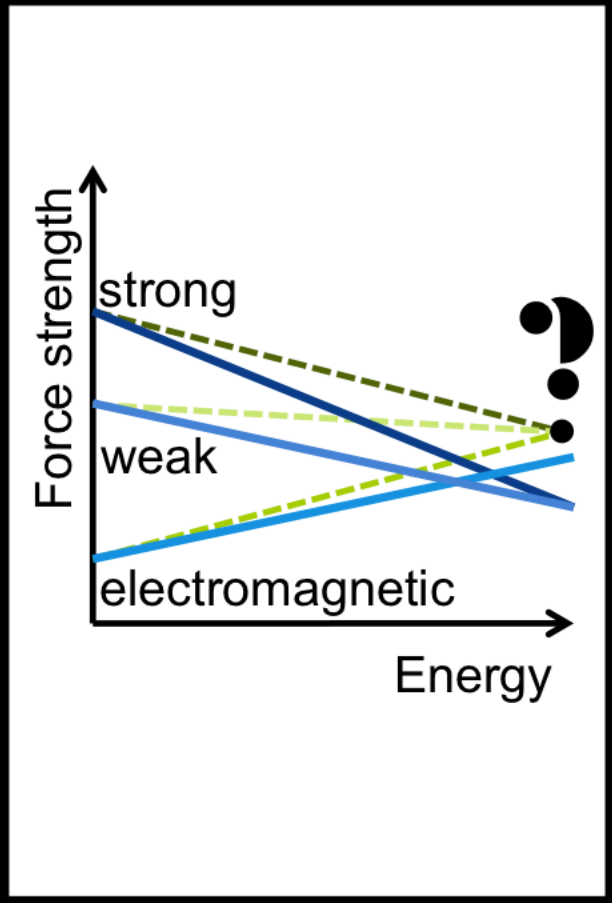


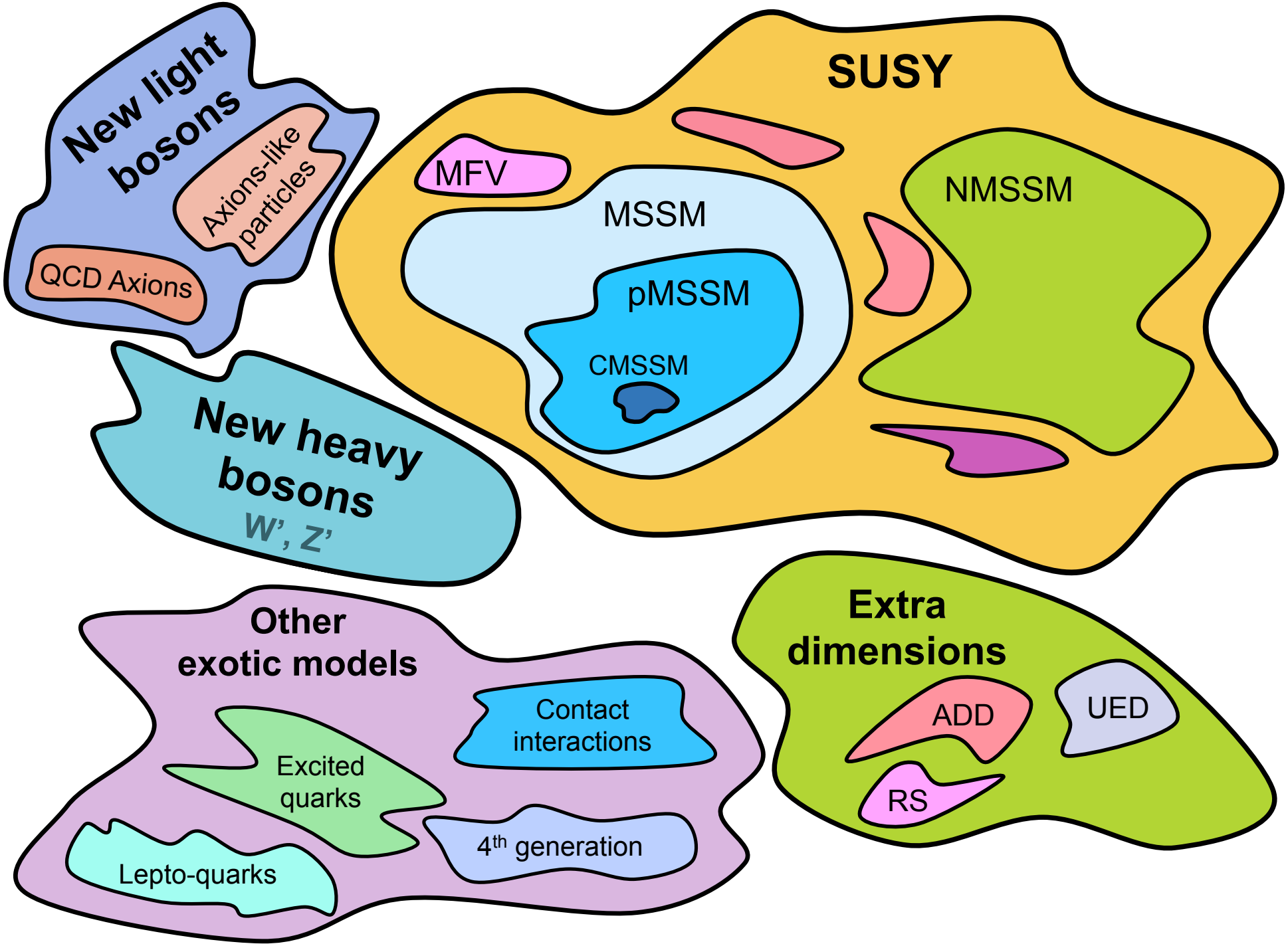
Figure from Particle Data Group, LBNL


I II III
Quarks
Leptons
Bosons
2.4 MeV
u
1.3 GeV
c
170 GeV
t
0
γ
4.8 MeV
d
104 MeV
s
4.2 GeV
b
0
g
< 2 eV
ν1
< 2 eV
ν2
< 2 eV
ν3
91 GeV
Z
80 GeV
W
0.5 MeV
e
106 MeV
μ
1.8 GeV
τ
125 GeV
H

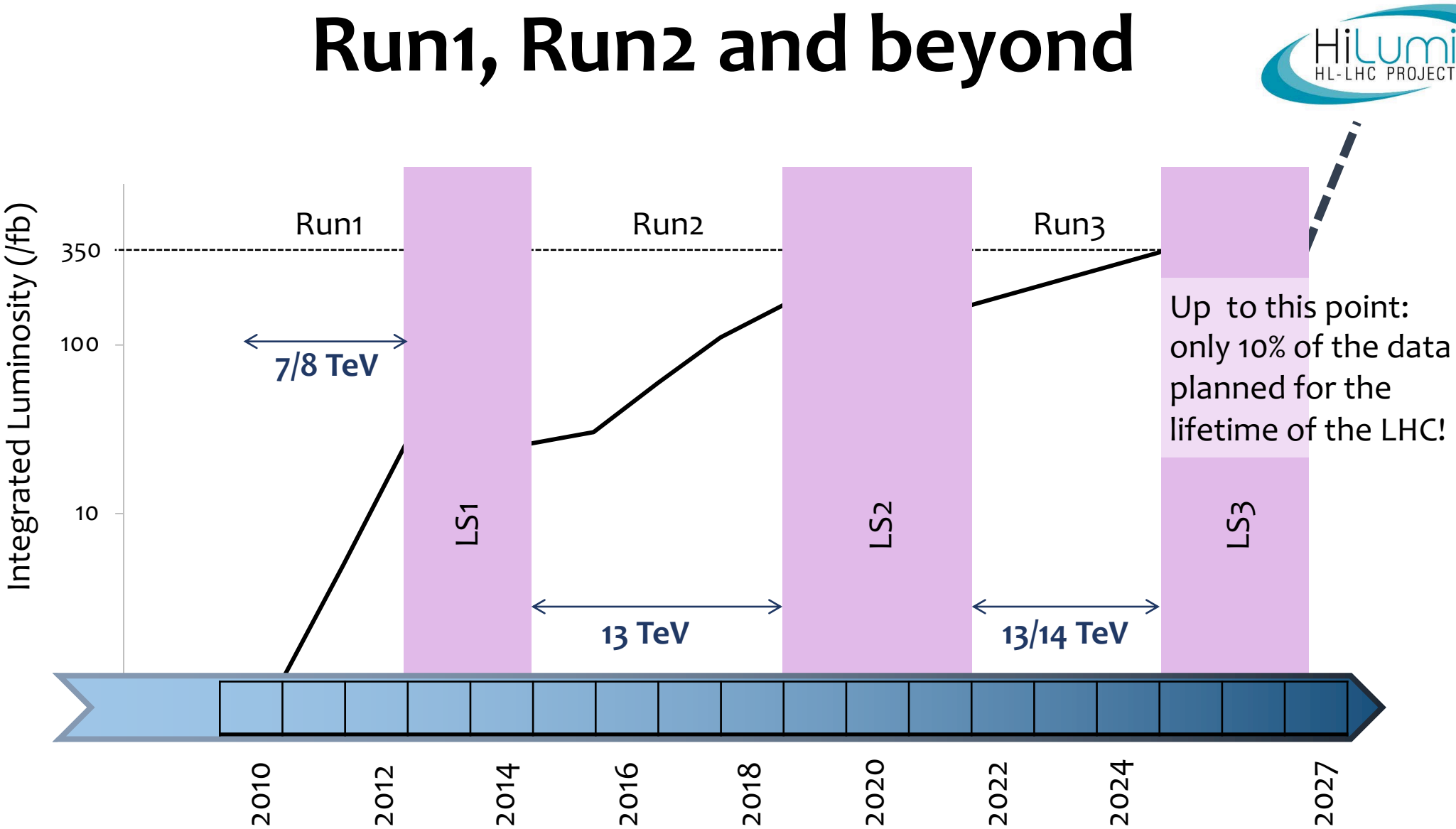

125 GeV
H
H
t
t
H


Velocity
Observed
Predicted
Radius

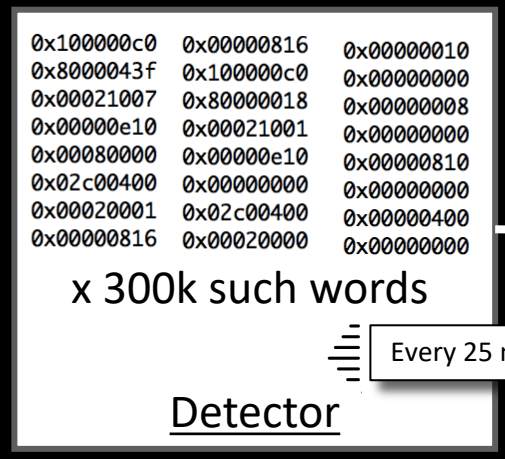

Force strength
strong
weak
electromagnetic
Energy

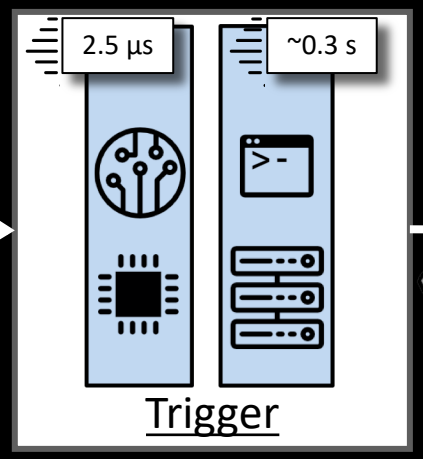

New light bosons
Axions-like particles
QCD Axions
SUSY
MFV
MSSM
pMSSM
CMSSM
NMSSM
New heavy bosons
W', Z'
Other exotic models
Contact interactions
Excited quarks
4th generation
Lepto-quarks
Extra dimensions
ADD
UED
RS

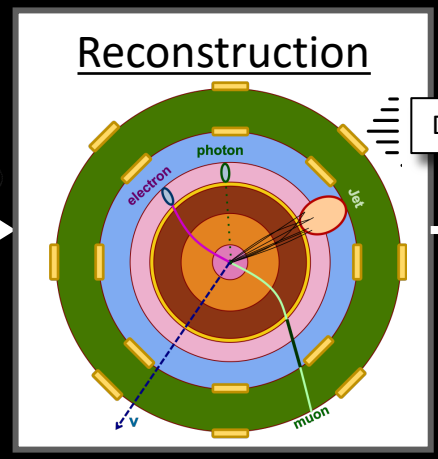
Run1, Run2 and beyond

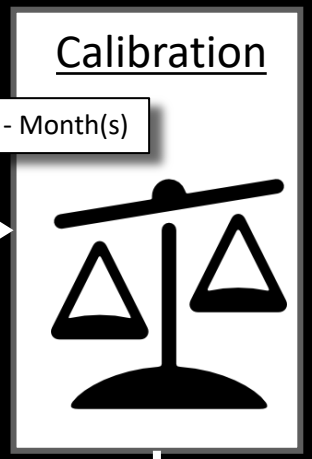


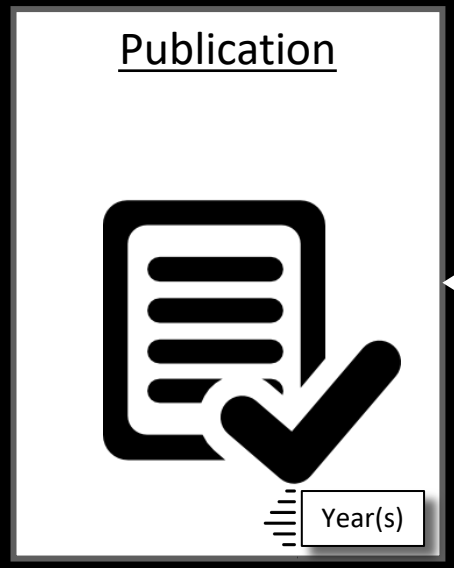
An event's lifetime

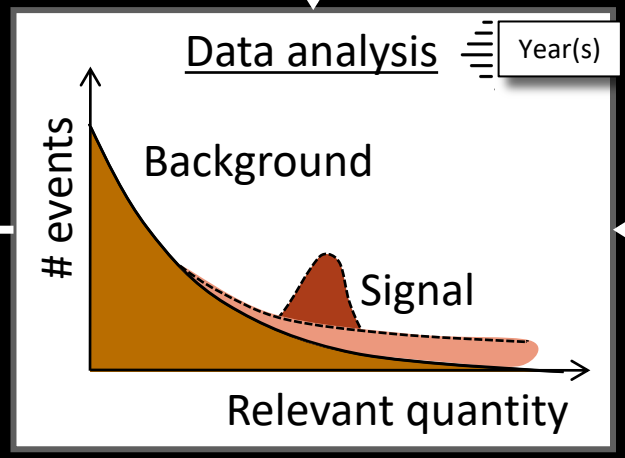


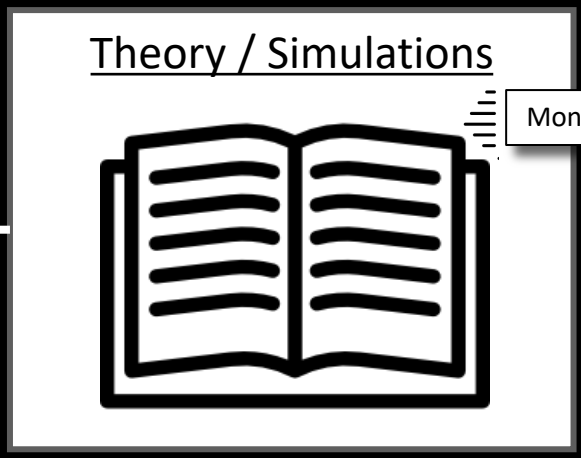




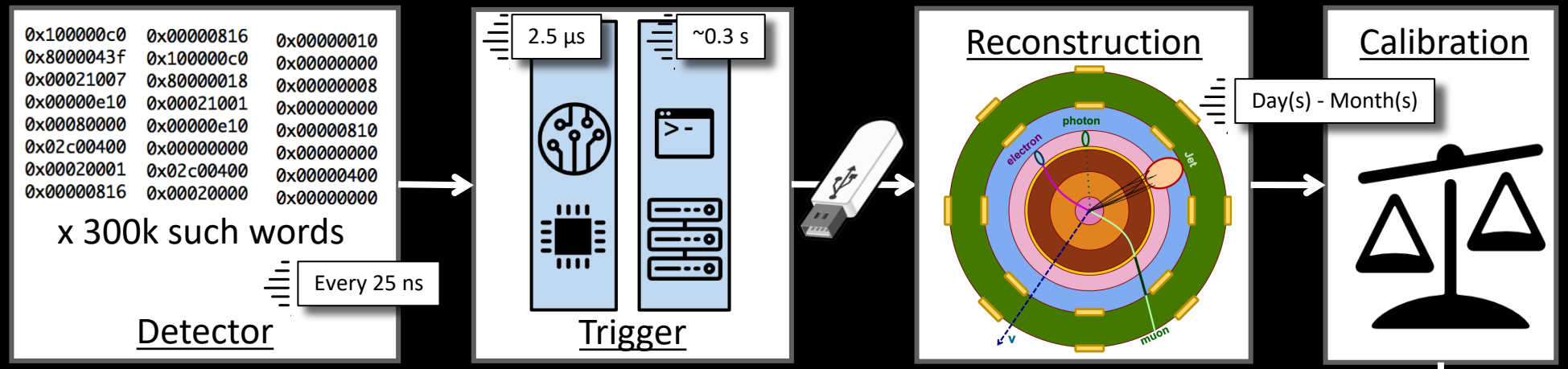






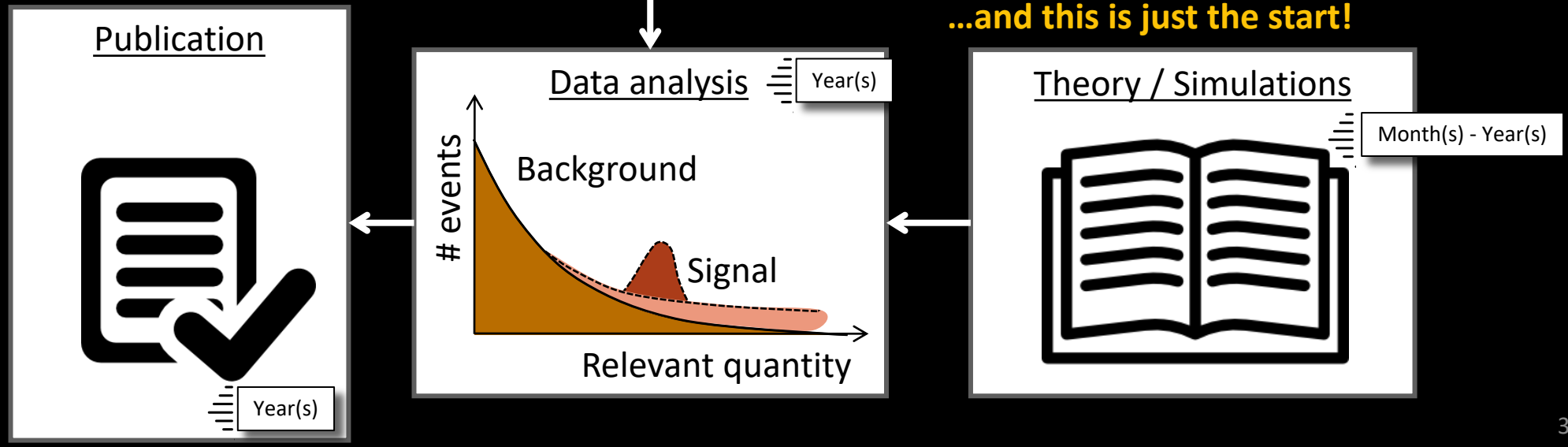


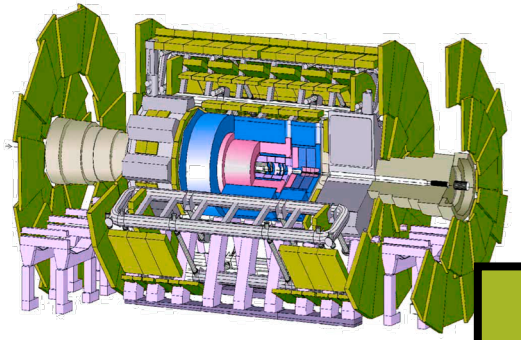
An event's lifetime



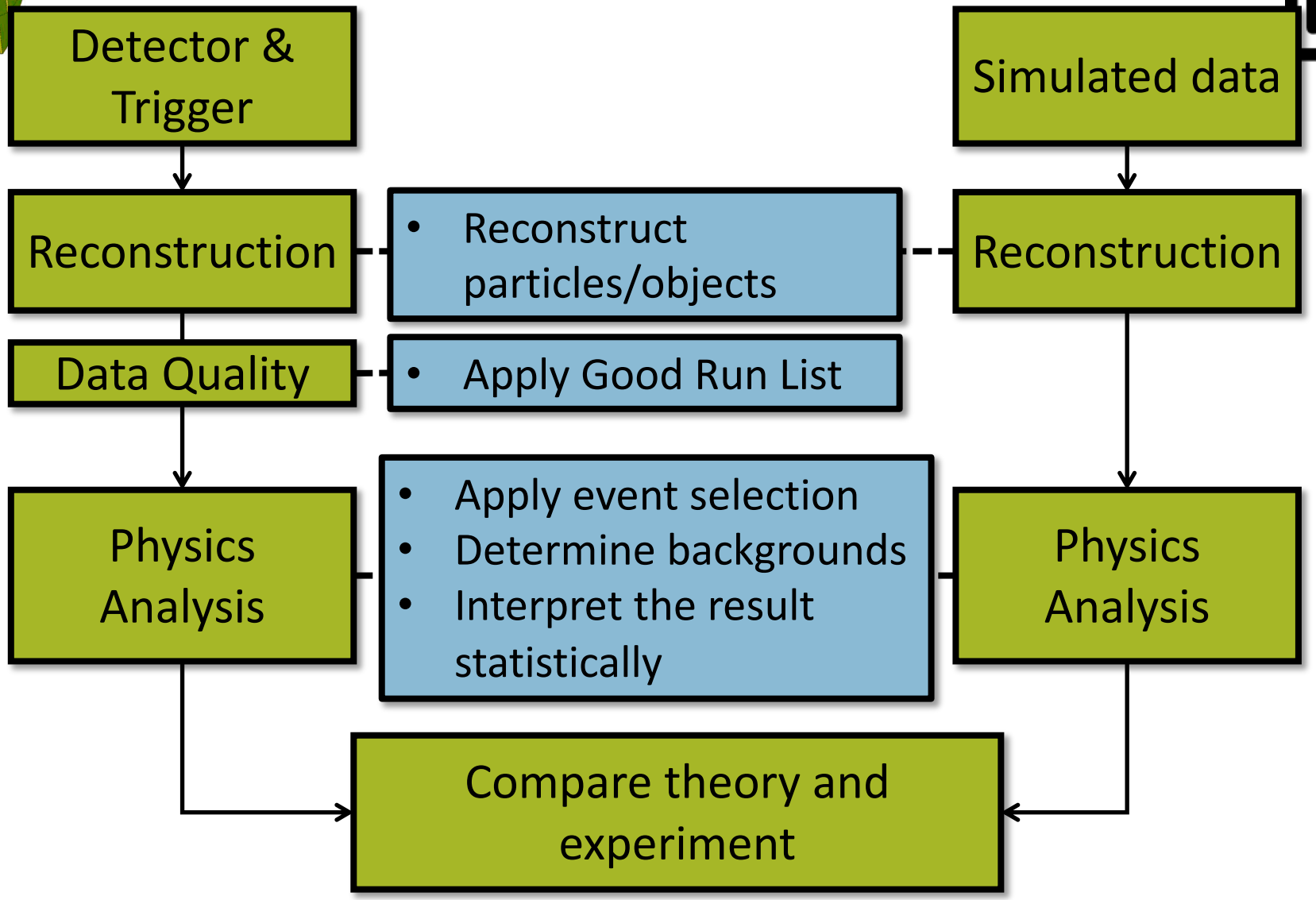
ATLAS has collected about 2 billion events, corresponding to **20 PB raw data**

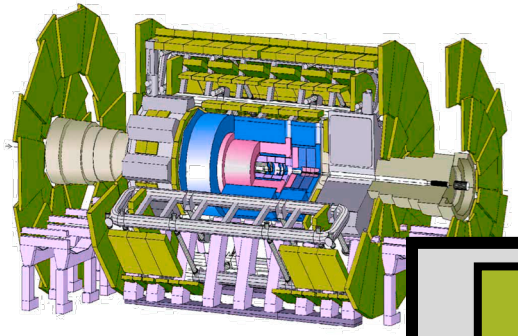
…and this is just the start!



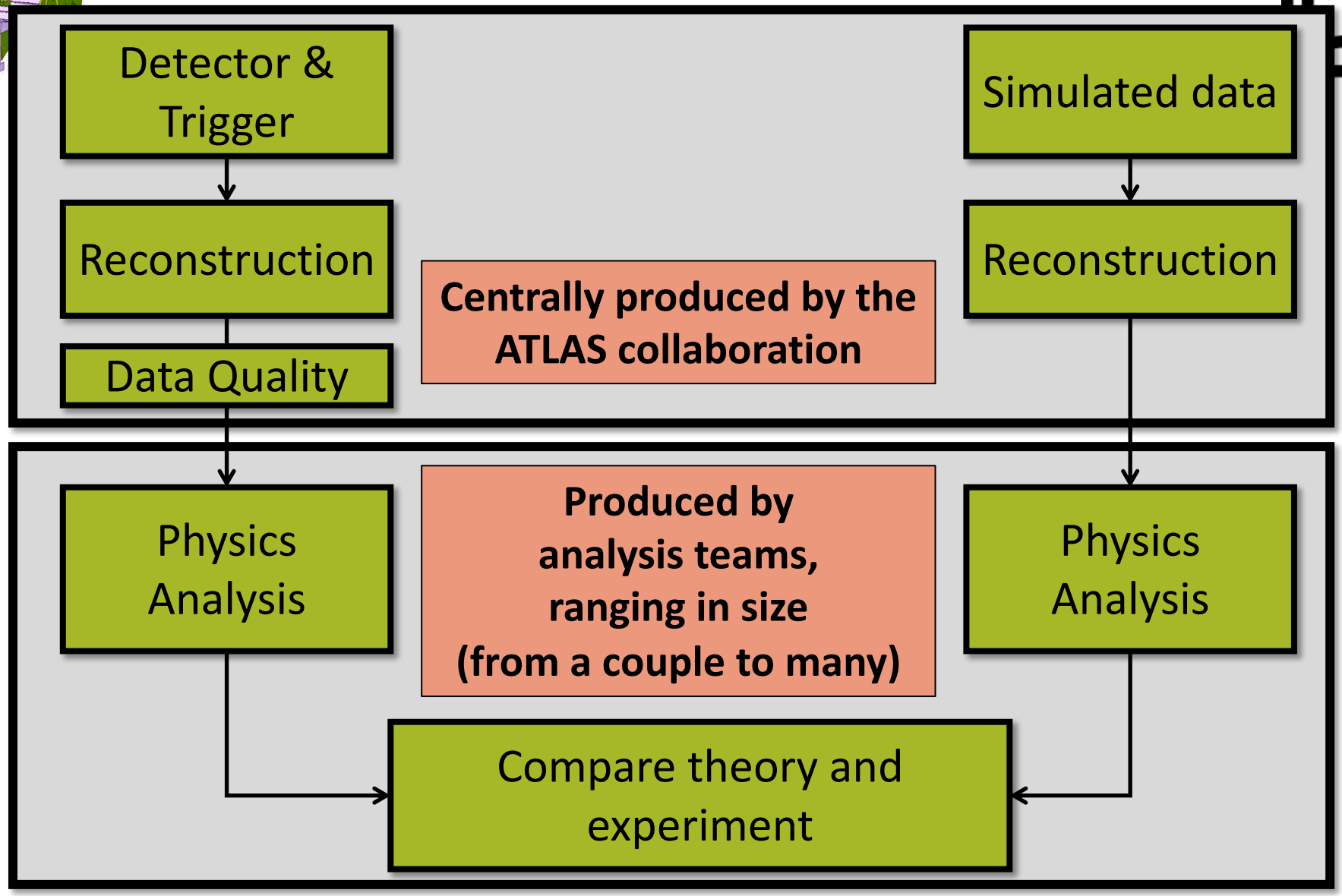


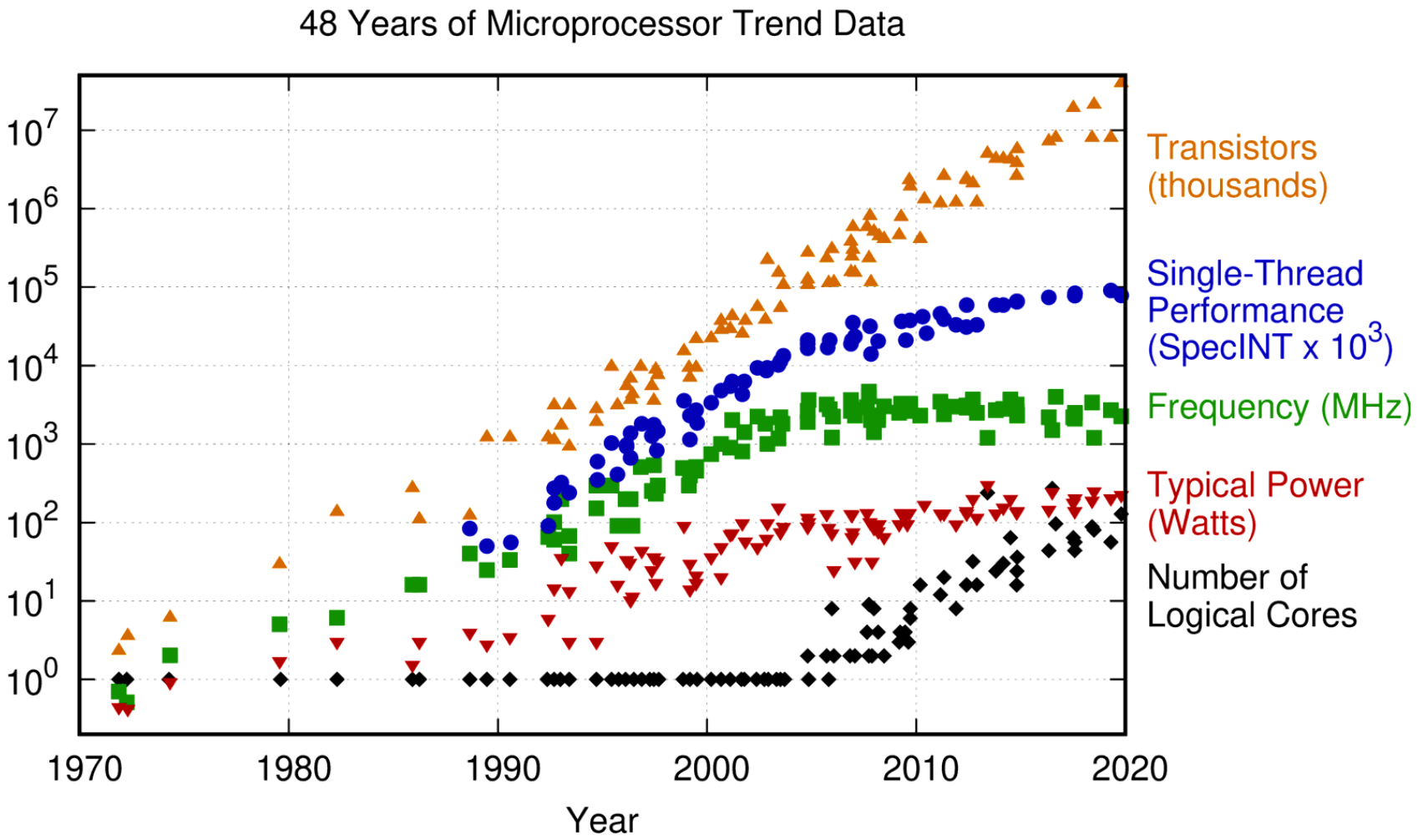
Simplified analysis flow





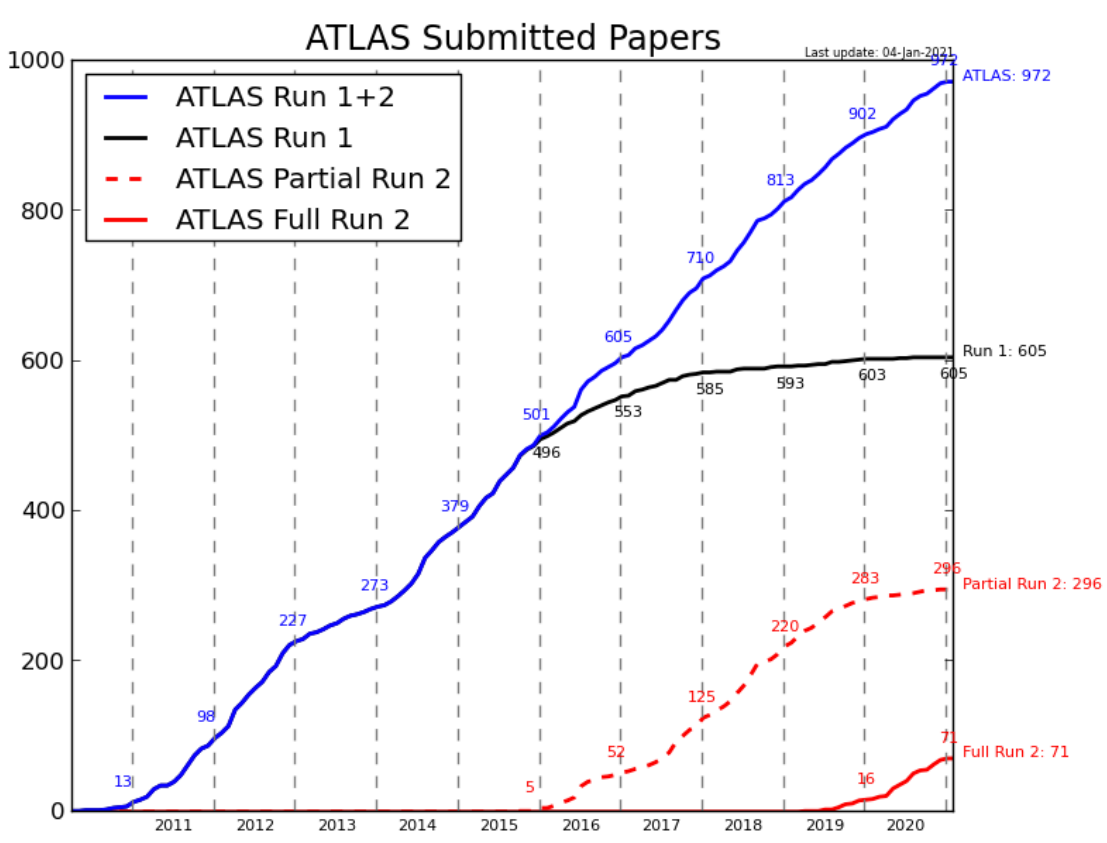
Simplified analysis flow

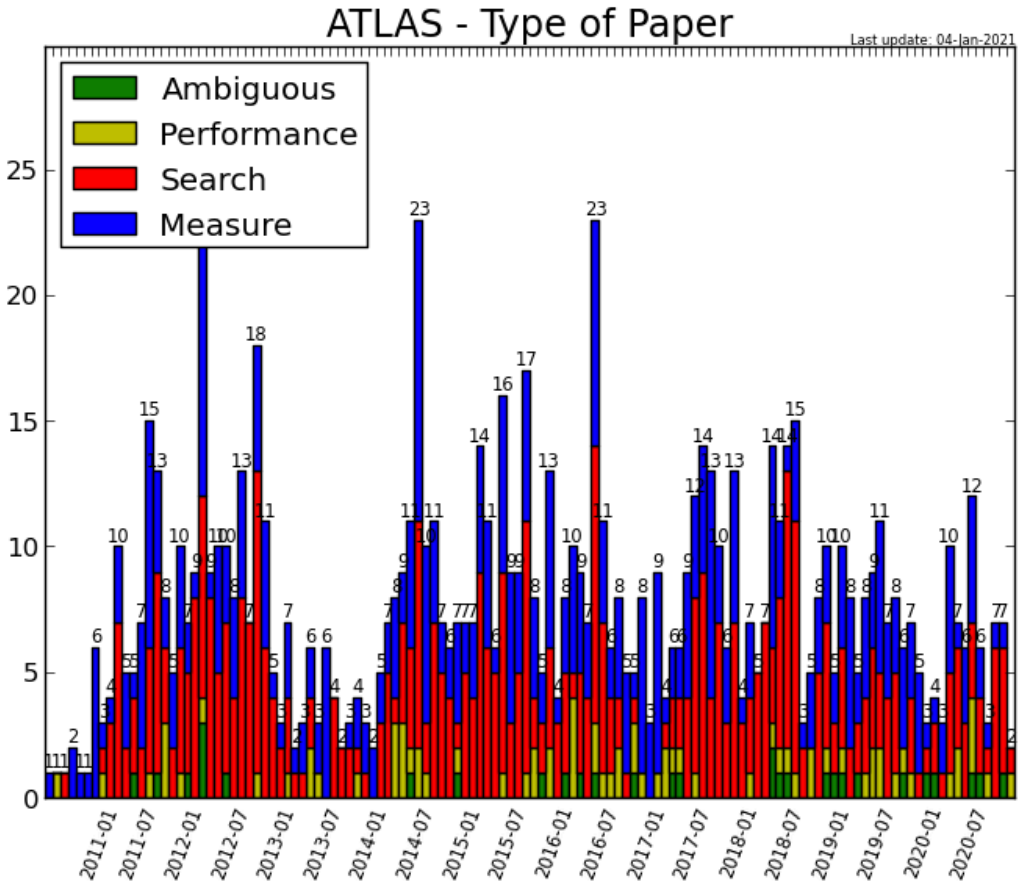


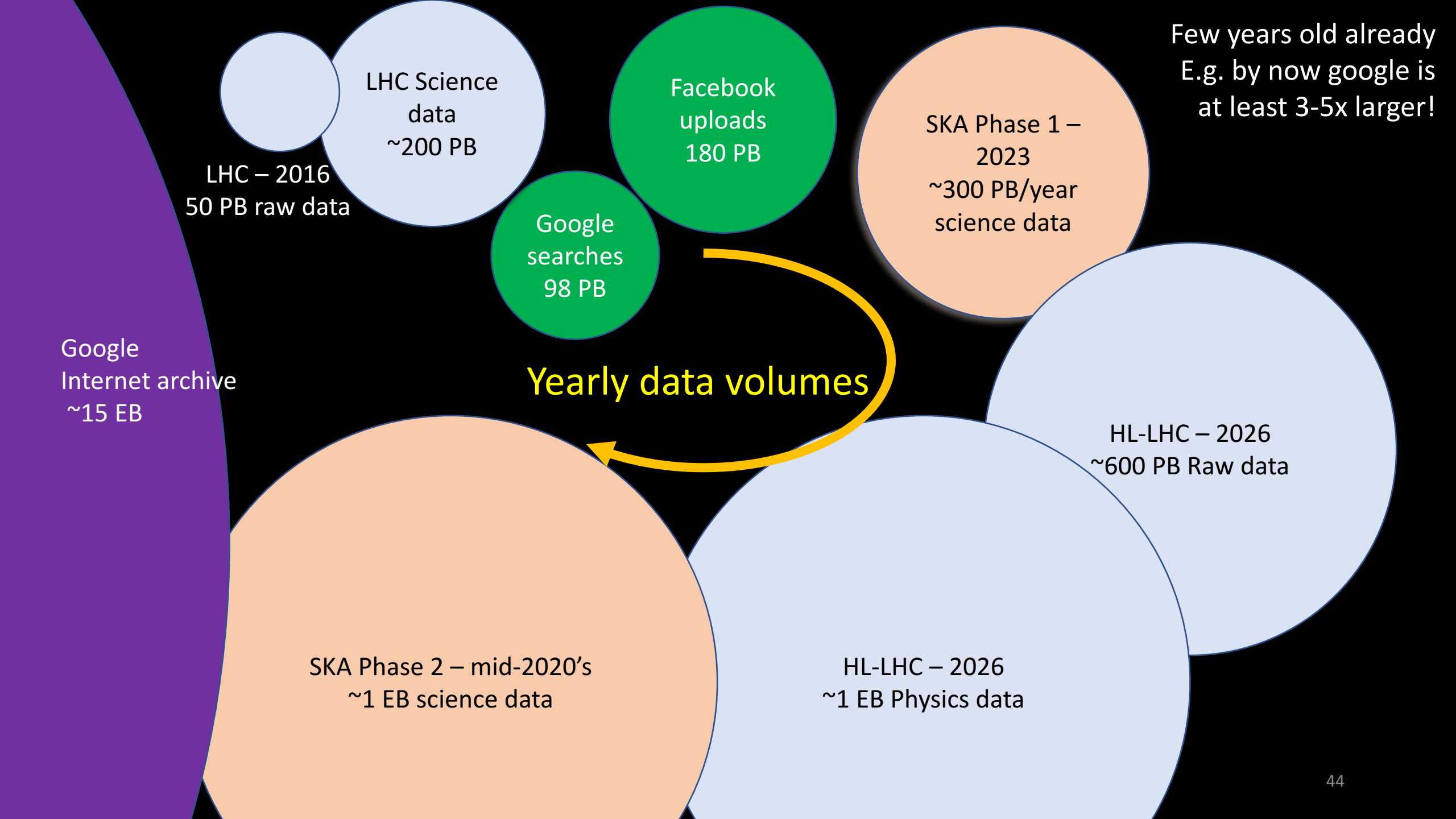


https://github.com/karlrupp/microprocessor-trend-data

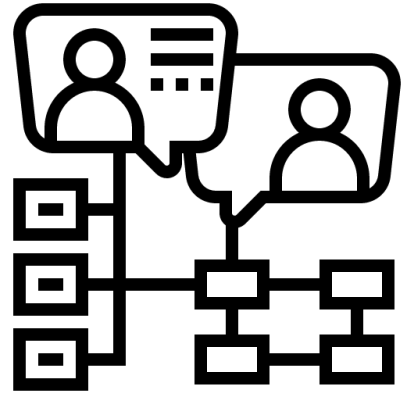
ATLAS Publications





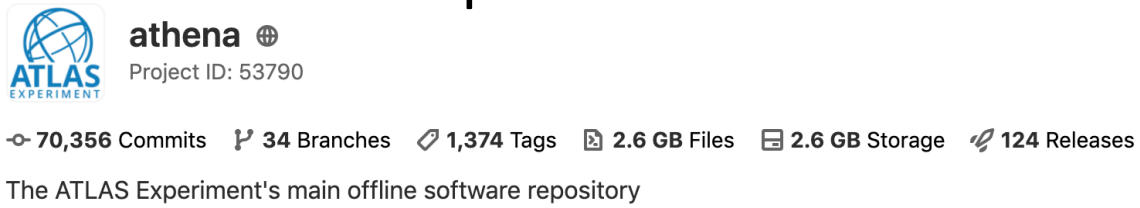

Few years old already
E.g. by now google is
at least 3-5x larger!
LHC – 2016
50 PB raw data
LHC Science
data
~200 PB
Facebook
uploads
180 PB
Google
searches
98 PB
SKA Phase 1 –
2023
~300 PB/year
science data
Google
Internet archive
~15 EB
Yearly data volumes
HL-LHC – 2026
~600 PB Raw data
SKA Phase 2 – mid-2020's
~1 EB science data
HL-LHC – 2026
~1 EB Physics data

Software - organisation



- All software organized in packages in Git. For example:

https://gitlab.cern.ch/atlas/athena



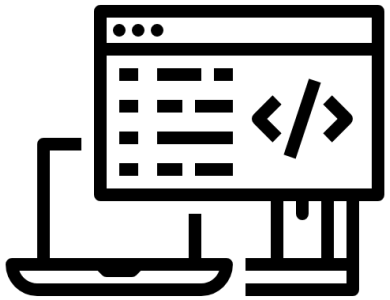
- Central software, not including user analysis code, counts about 6.6M lines of c++ and python (primarily)

- All software open source, copyrighted and licenced (Apache 2)
  - "Copyright (C) 2002-2020 CERN for the benefit of the ATLAS collaboration"
  - For open use – but also for crediting developers who move out of academia

- "HEP software foundation" assists experiments to set common practices and use common tools

HSF
HEP Software Foundation

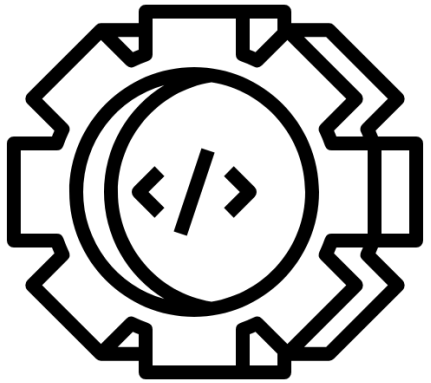
Software – development



- **Continuous development**
  - **To improve performance**, both scientific and computational
  - **To adapt to technology evolution** (e.g. HPCs, GPUs)
  - **To accommodate scientific needs** (e.g. more challenging data taking conditions)

- **Thorough tracking of software developments**
  - Via the Jira software, supported by CERN IT Jira Software
  - Multiple releases exist for merging of new code with existing one
  - Automated tools run nightly to verify code sanity & performance
  - Globally the software projects are coordinated with careful planning

Other tools…



Primarily, databases

- Example use cases:
  - data taking conditions and run details
  - simulated samples and how they were produced
  - jobs running on grid
  - status of analyses / papers
  - ATLAS membership information
  - …
- Either commercial products, supported by CERN (primarily so far: Oracle), or custom made for ATLAS needs
- By now, they all matched with attractive user interfaces

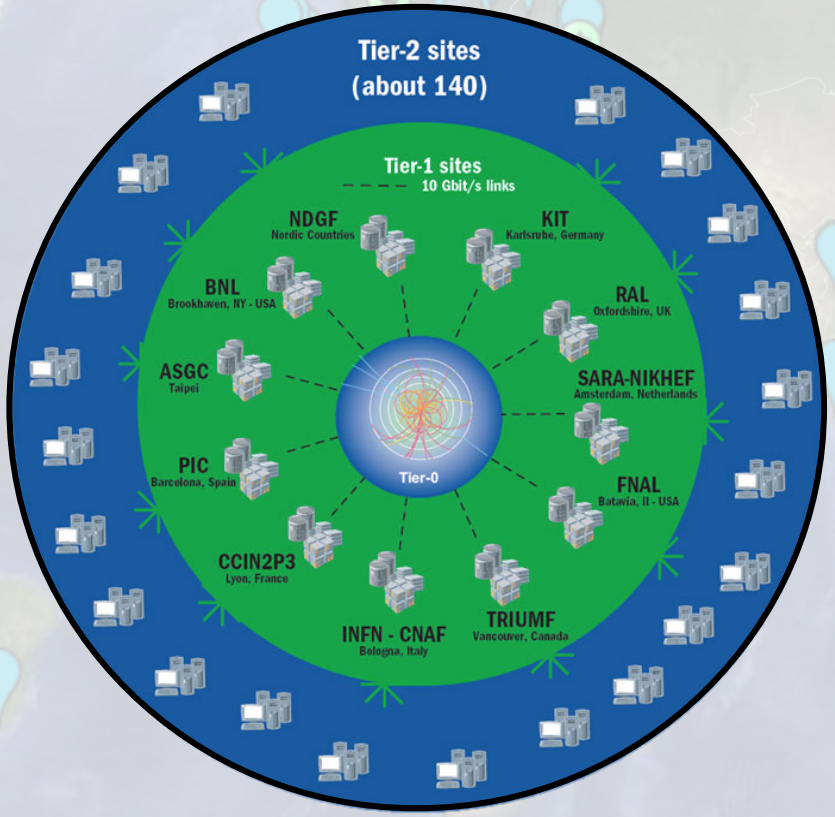
But also, analysis tools

- Workhorse: root.cern.ch (open source!)
- ATLAS software largely built around it
- Analysis-specific software developed by teams available to whole collaboration

Worldwide LHC Computing Grid
an international collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists.

- **Tier-0 (CERN):**
  - Data recording, reconstruction and distribution
- **Tier-1:**
  - Permanent storage, re-processing, analysis
- **Tier-2:**
  - Simulation, end-user analysis



- **161 sites, 42 countries**
- **1 M CPU cores**
- **1 EB of storage**
- **> 2 M jobs/day**
- **> 100 PB moved/month**
- **accessed by 10k users**
- **10-100 Gb links**

WLCG Worldwide LHC Computing Grid

Worldwide LHC Computing Grid
an international collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists.

- **Tier-0 (CERN):**
  - Data recording, reconstruction and distribution
- **Tier-1:**
  - Permanent storage, re-processing, analysis
- **Tier-2:**
  - Simulation, end-user analysis

- **161 sites, 42 countries**
- **1 M CPU cores**
- **1 EB of storage**
- **> 2 M jobs/day**
- **> 100 PB moved/month**
- **accessed by 10k users**
- **10-100 Gb links**

Network proved better than anyone imagined: Any job can run anywhere

WLCG Worldwide LHC Computing Grid