LUNDS UNIVERSITET

Real-time analysis at the LHC

HOW TO MAKE THE MOST OF LHC DATA

CATERINA DOGLIONI - LUND UNIVERSITY

@CATDOGLUND, SHE/HER

http://www.hep.lu.se/staff/doglioni/

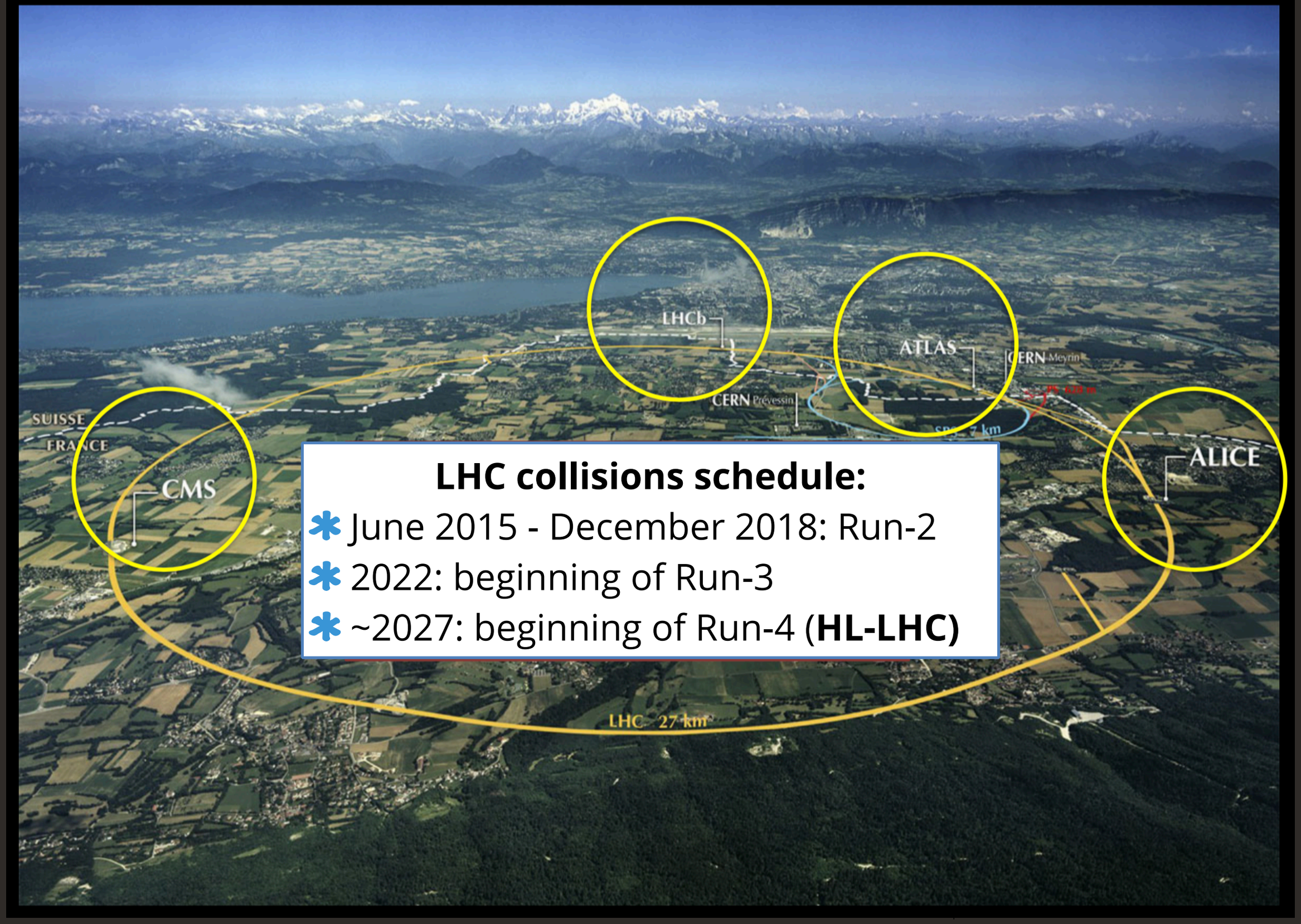
European Research Council
Established by the European Commission

ATLAS EXPERIMENT

Outline

- The Large Hadron Collider and its experiments
- How/why making the most of the data
- Dark matter searches with *real-time analysis*
- Real-time analysis beyond LHC/high energy physics

Introduction

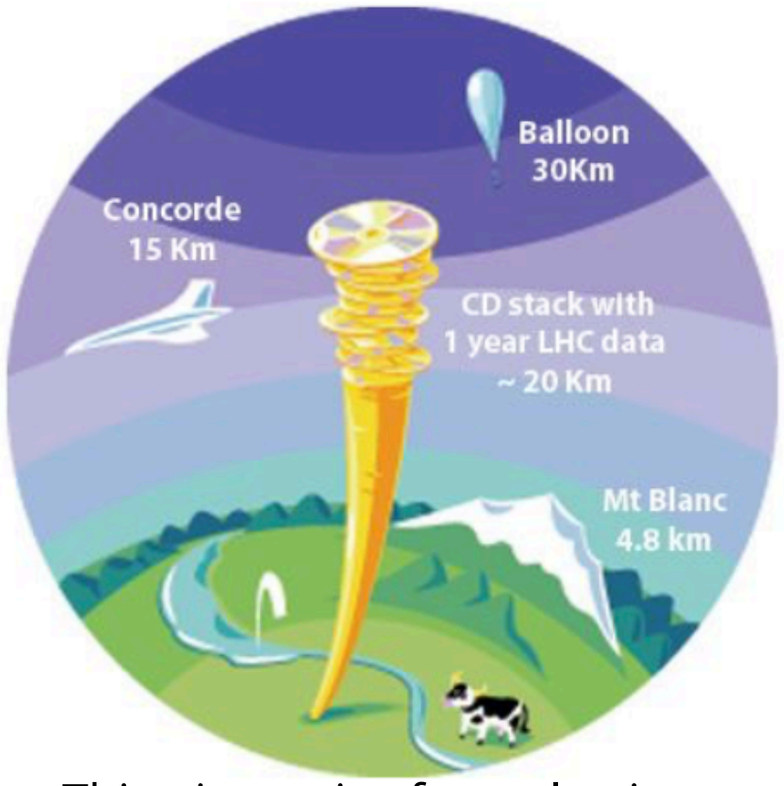


LHC collisions schedule:

- June 2015 - December 2018: Run-2
- 2022: beginning of Run-3
- ~2027: beginning of Run-4 (**HL-LHC)**

Challenge: selecting interesting data at the LHC

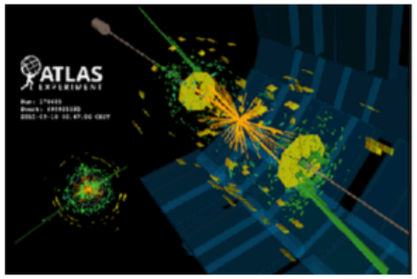
- If everything was recorded at the LHC:
  - up to 30 million collisions/second (MHz)
  - 1-1.5 MB/data per collision
  - 30 MHz * 1 MB = 30 TB/s
  - 30 TB/s * 10e+6 s/year (day & night) ~ 0.05 ZB/year
- facebook
  - 600 TB/day ~ 200 PB/year [Facebook 2014]
    - **"There's always a bigger fish"**
      [C. Tully's talk @ siRTDM18]
- But bigger fish also have bigger money…
  cost-effectiveness important for scientific instruments!

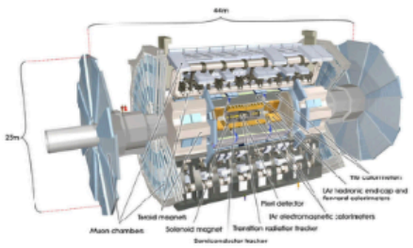


This picture is after selection of "interesting" data: with all data, the stack of CDs reaches to the moon

LHC experiments need to select "interesting" events **in real-time** (milli/microseconds)

Trigger and data acquisition chain





Online | Offline

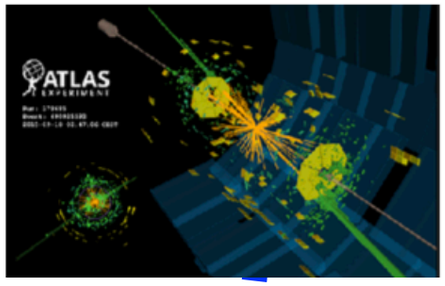
Event selection (trigger) → Object reconstruction and calibration → Data analysis

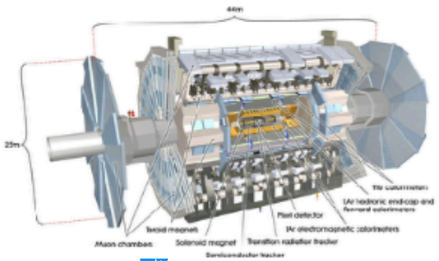
Computing resources are essential for the full data taking chain

Trigger and data acquisition: select interesting events

First step: **fast hardware selection (Level 1)**
ATLAS/CMS data taking rate: 100 kHz
Second step: **computer farm (High-Level Trigger)**
ATLAS/CMS data taking rate: 1 kHz

Online | Offline





Event selection (trigger) → Object reconstruction and calibration → Data analysis

Data produced by the LHC
(multiplied by large number)

Data I like

...maybe discoveries?

Data I would have used to like if I had been a data hipster

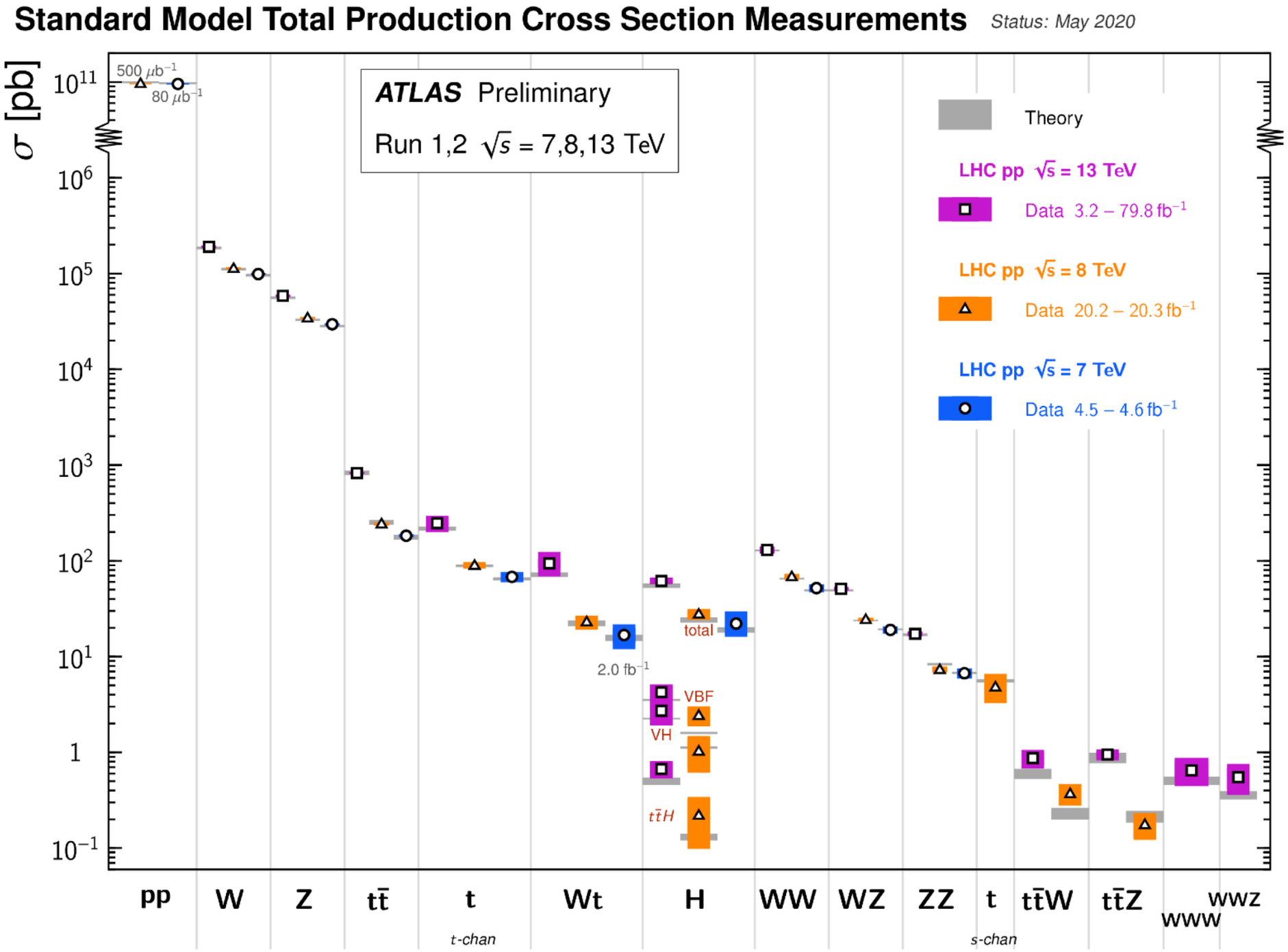
Data you like

What we can record in full after the trigger

Data ~nobody likes

This works for a number of LHC measurements (& searches…)

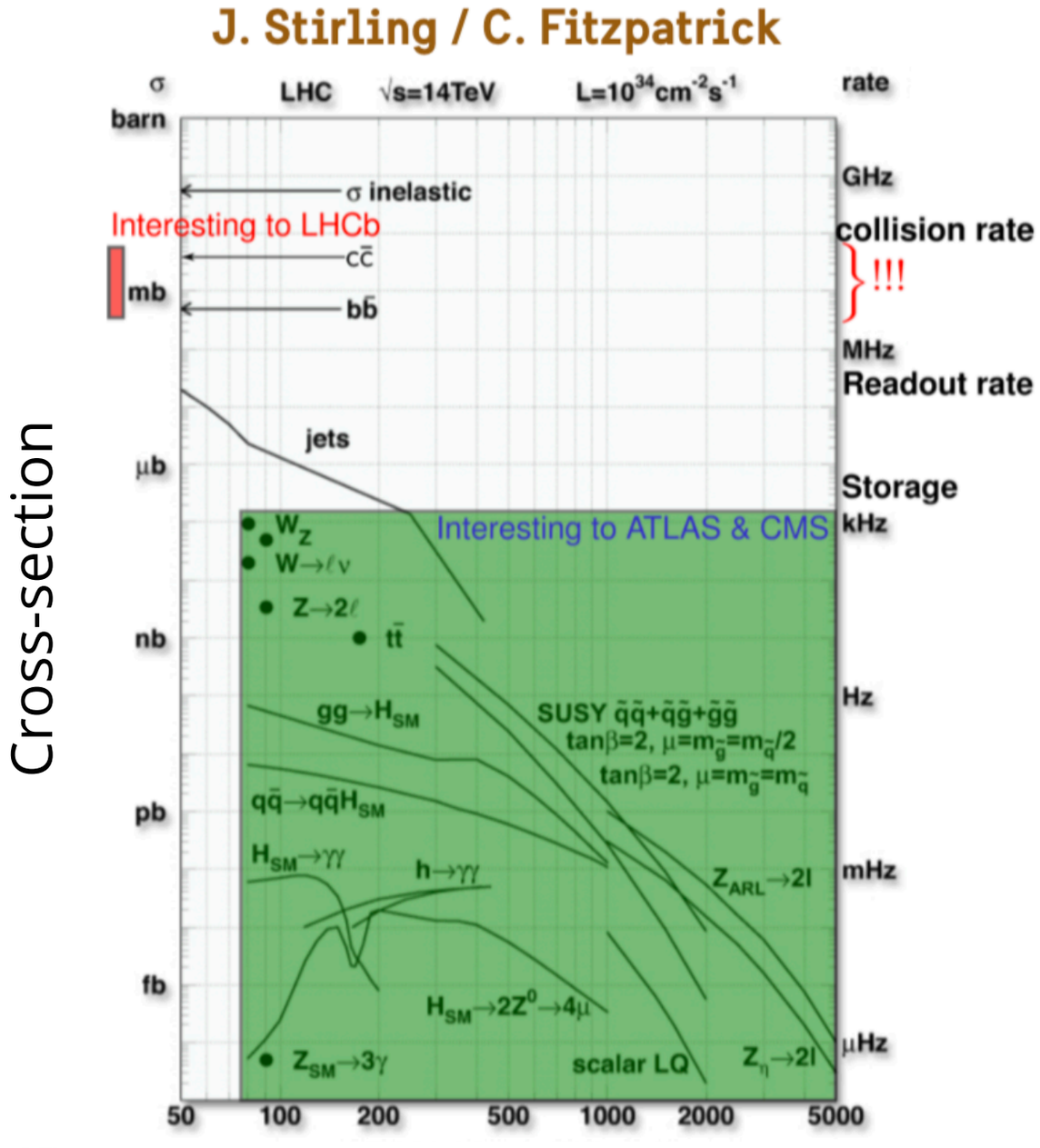
https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2020-010/



...but are we missing something?

What we can trigger on

What is interesting at the LHC?



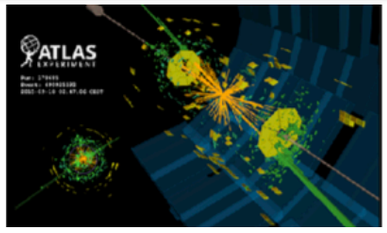
Cross-section * Luminosity
= number of events produced

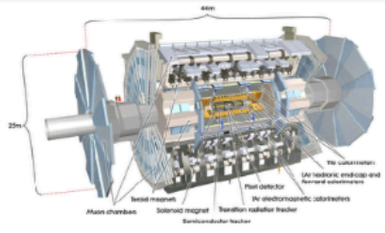
Challenges:

The **definition of "interesting"** changes experiment by experiment

Rare signal processes that are buried in **high-rate backgrounds** have to be discarded

Where are the limitations to record more data?





Event selection (trigger) → Object reconstruction and calibration → Data analysis

Detector readout to hardware trigger

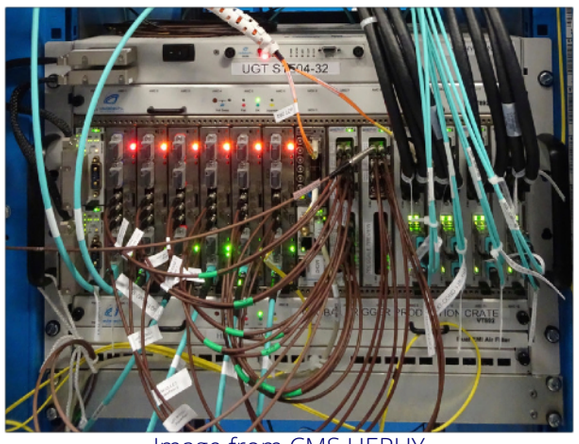


Image from CMS HEPHY

CPU for processing events
(within the software trigger, and in the offline farm)


Image from C. Bernius's talk

Disk/tape to store events


Image from CERN

This talk: how/why overcome them

Alternative data taking, trigger and analysis workflows

Low-mass resonance decays (motivated by dark matter & dark sectors)

How LHC collaborations can make the most of the data

Interesting time for high energy collider physics:
we don't know what to expect from dark matter / new physics
(but we have a prior: it should be *somewhere*)
we have the **LHC running now**,
and the data we discard is **gone forever**

How LHC collaborations can make the most of the data

Interesting time for high energy collider physics:
we don't know what to expect from dark matter / new physics
(but we have a prior: it should be *somewhere*)
we have the **LHC running now**,
and the data we discard is **gone forever**

1. Build **detectors** (+services) that can read-out more
   - Some LHC experiments aiming for 40 MHz readout for Run-3

Overcome detector limitations

2. Save data for **further reconstruction, later**
   - Periods between data taking can be long…keep as much as possible
   - Byproduct: make data & workflows FAIR and sustainable!
3. Analyze as much data as possible, **as fast as possible**
   - Requires making **hard choices** on what information to keep for further analysis
4. Make **software/hardware faster**
   - Optimize software or use heterogeneous architectures, e.g. FPGAs/GPUs

Overcome computing limitations

5. Implement more refined algorithms to **look for the unexpected**
   - Including **unsupervised searches / novelty detection**

How LHC collaborations can make the most of the data

Interesting time for high energy collider physics:
we don't know what to expect from dark matter / new physics
(but we have a prior: it should be *somewhere*)
we have the **LHC running now**,
and the data we discard is **gone forever**

1. Build **detectors** (+services) that can read out more
   - Some LHC experiments aiming for 40 MHz readout for Run-3

2. Save data for **further reconstruction, later**
   - Periods between data taking can be long…keep as much as possible
   - Byproduct: make data & workflows FAIR and sustainable!

3. Analyze as much data as possible, **as fast as possible**
   - Requires making **hard choices** on what information to keep for further analysis

4. Make **software/hardware faster**
   - Optimize software or use heterogeneous architectures, e.g. FPGAs/GPUs

5. Implement more refined algorithms to **look for the unexpected**
   - Including **unsupervised searches / novelty detection** see Maurizio Pierini's talk

data parking / selective persistency (not covered in this talk…but worth mentioning)

real-time analysis

Real-Time Analysis (RTA)

A paradigm change

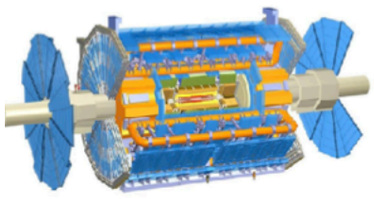
Asynchronous data analysis

First record data, then reconstruct/analyze it

Real-time data analysis

Reconstruct/analyze data as soon as it is read out
so that only (**smaller**) final-state objects
or histograms need to be stored

Regular trigger & data analysis path



Trigger event reconstruction

Decision

only if trigger passed

Full event reconstruction

Main stream

Standard data analysis

Turbo/Data Scouting/TLA path

Turbo stream (LHCb),
Data Scouting (CMS),
Trigger-level Analysis (ATLAS),

Trigger event reconstruction

Decision

only if trigger passed

Full event reconstruction

Main stream

Standard data analysis

In any case

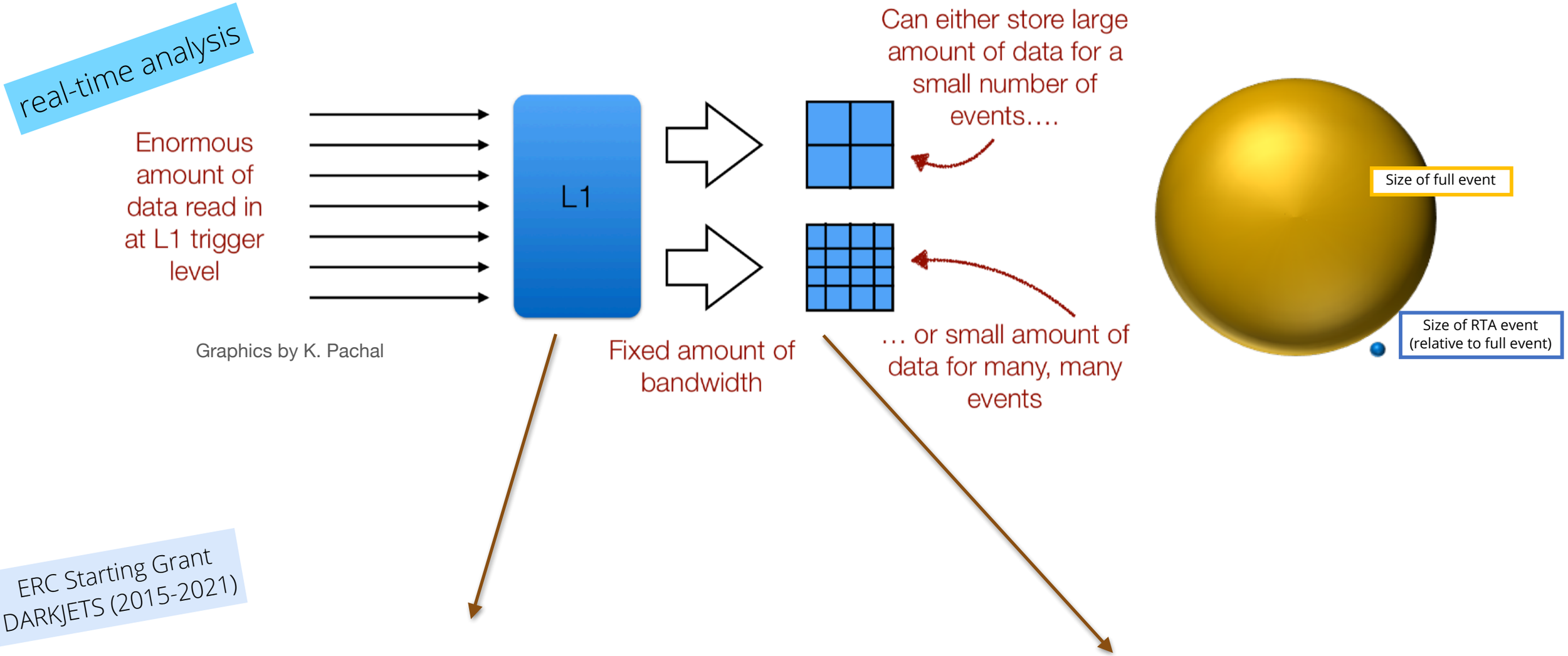
Partial (HLT-only) event reconstruction

Special trigger analysis stream

Analysis with trigger objects

(Near-)real-time analysis of LHC data

real-time analysis



Graphics by K. Pachal

ERC Starting Grant
DARKJETS (2015-2021)

Perform as much "analysis" as possible @ HLT

- Reconstruction & calibration
- First preselection to skim "backgrounds"

Reduced data formats:

- Only keep final trigger objects (drop raw data)
- Save only "interesting" parts of the detector
- A combination of the two

ERC Consolidator Grant
REALDARK (2021-2026)

More with less: Selective persistency/Partial Event Building

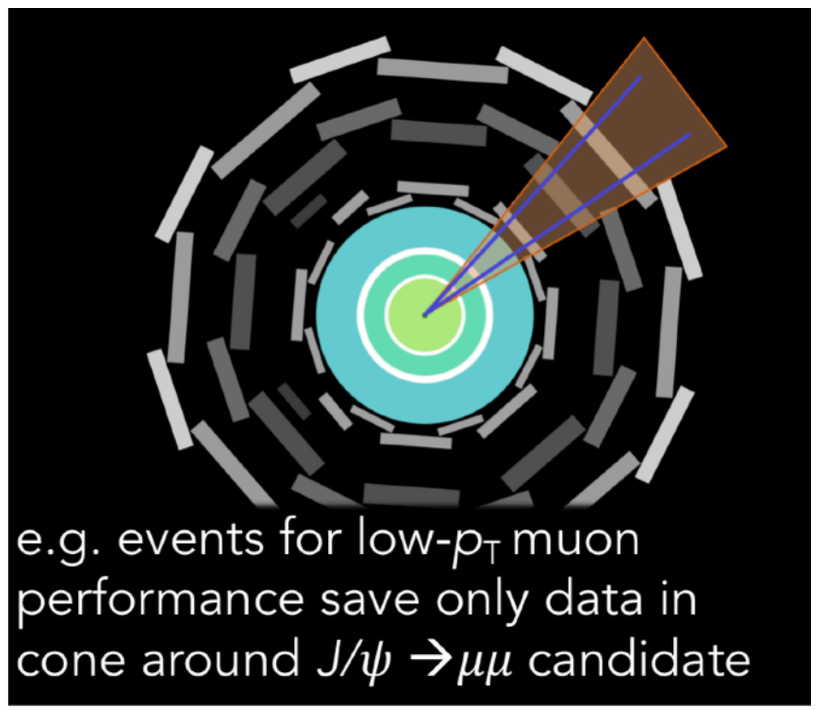
Real-time analysis is necessary for searches
that would otherwise have been impossible due to trigger constraints

Traditional offline analysis still required for a number of searches/final states
where all raw information is needed (but we could do better)

Partial Event Building / Selective Persistency as a middle way:
save raw data && trigger objects only in the regions of interest, re-reconstruct later

H. Russell, EPS-HEP 2019,

Customizable output data @ LHCb:
- keep trigger objects only (7 kB)
- **keep trigger objects + "on-demand" raw and/or reco in selected regions (< 200 kB)**
- keep everything (200 kB)



HSF Trigger & Reco / Institut Pascal discussion, July 2016:
https://indico.cern.ch/event/835074/

data parking / **selective persistency**

Physics use cases for real time analysis

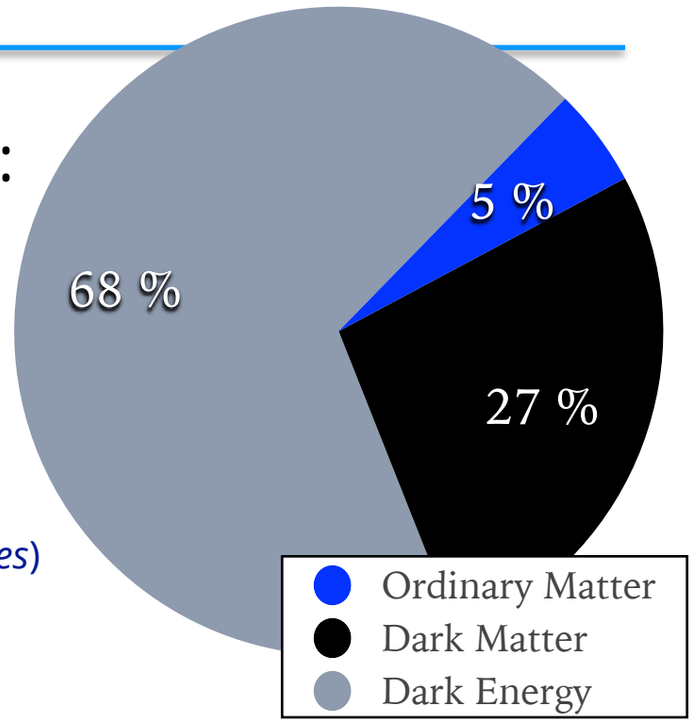
The need for dark matter

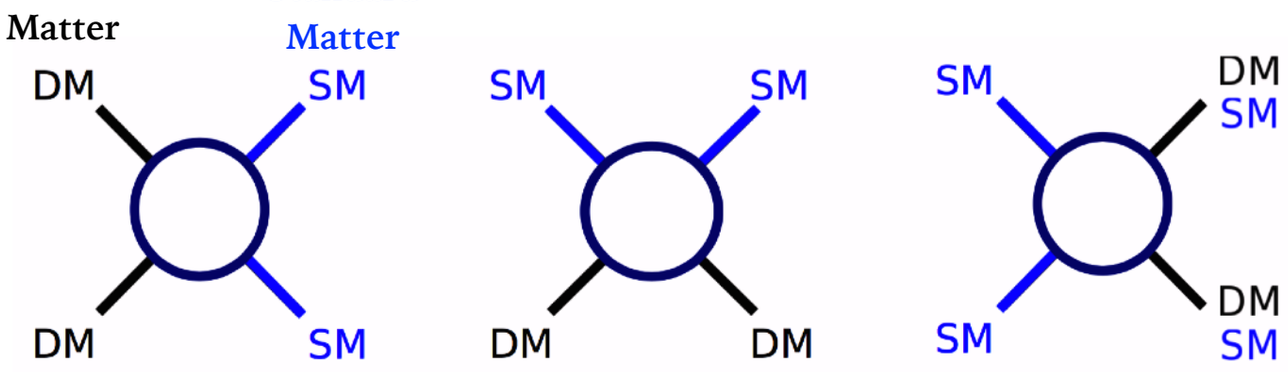
Empirical **problem** in the Standard Model of Particle Physics:
no explanation for **Dark Matter**

arXIv:0704.2276v1

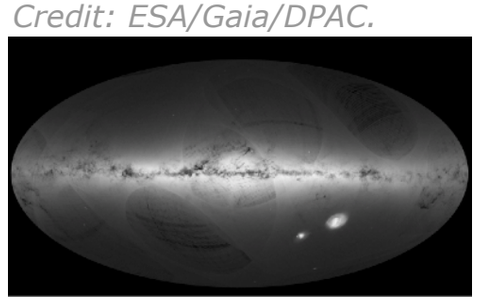
A possible **solution,** guided by **relic density**:
invisible **Dark Matter particles** in the ~**GeV-TeV scale**

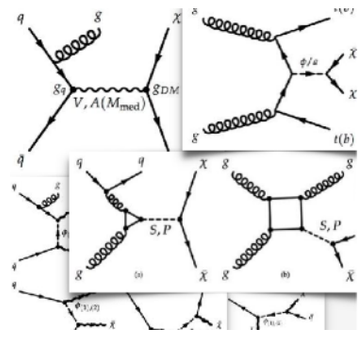
(including but not limited to one of the most studied DM hypotheses: *Weakly Interacting Massive Particles*)





Credit: ESA/Gaia/DPAC.

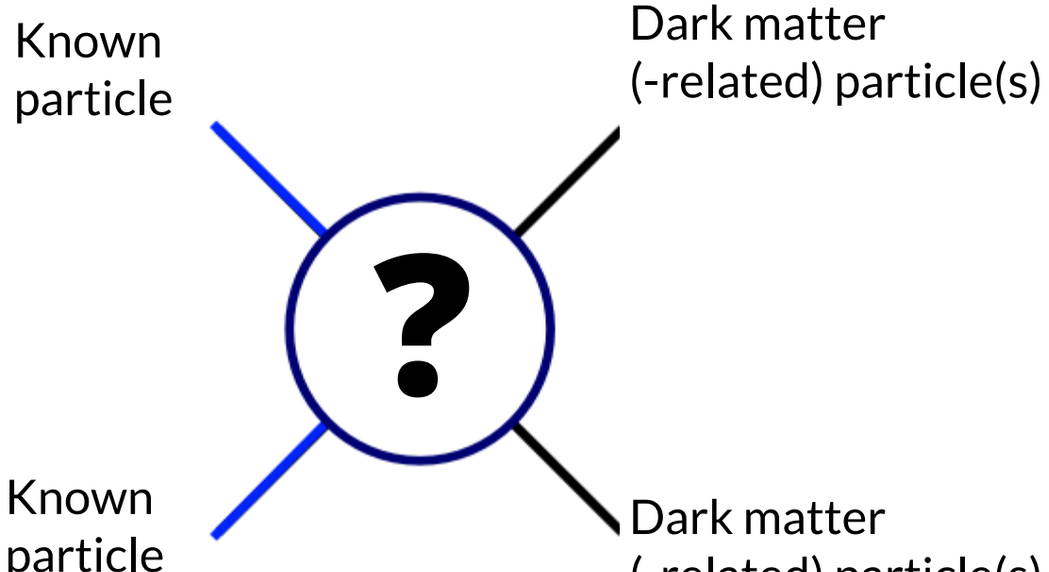




Indirect Detection | Direct Detection | Colliders & accelerators | Astrophysics | Theory

Complementary experimental strategies & inputs

Recreating dark matter/dark sectors in the lab: challenges



Trying to stay
as **model-agnostic** as possible,
while exploiting what the **LHC** is good at:
focus on the presence of a **resonance**
(alongside EFTs/more complete theories)

added bonus: resonance searches are bread&butter at colliders → robust analysis toolkit available

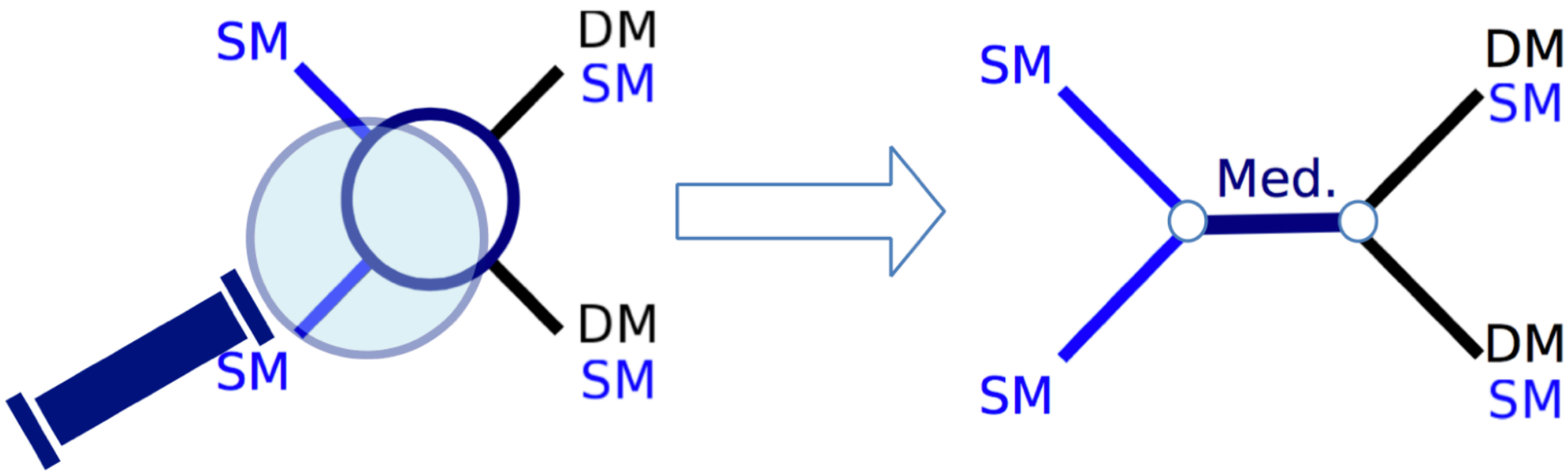
Challenges:

1. This kinds of processes are very **rare**
2. Many other processes may look the same (→large **backgrounds**)
3. Often **we don't know** how the resonance decays look like

These challenges can be met with non-standard workflows!

Dark matter mediators at the LHC

If there's a force there's a mediator:
(in this case, with axial vector couplings)

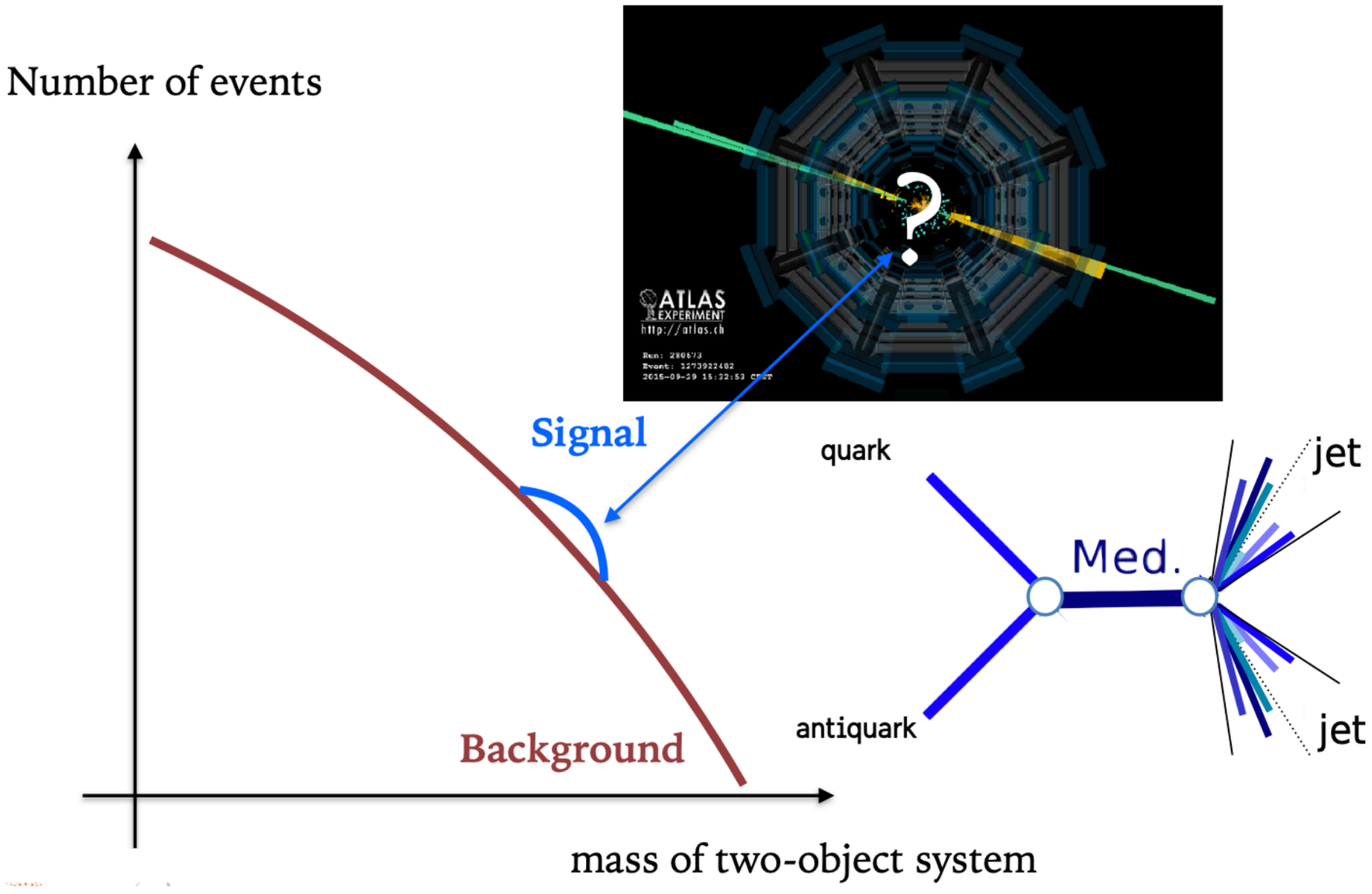


For this kind of models to be consistent with relic density: mediators should have low masses

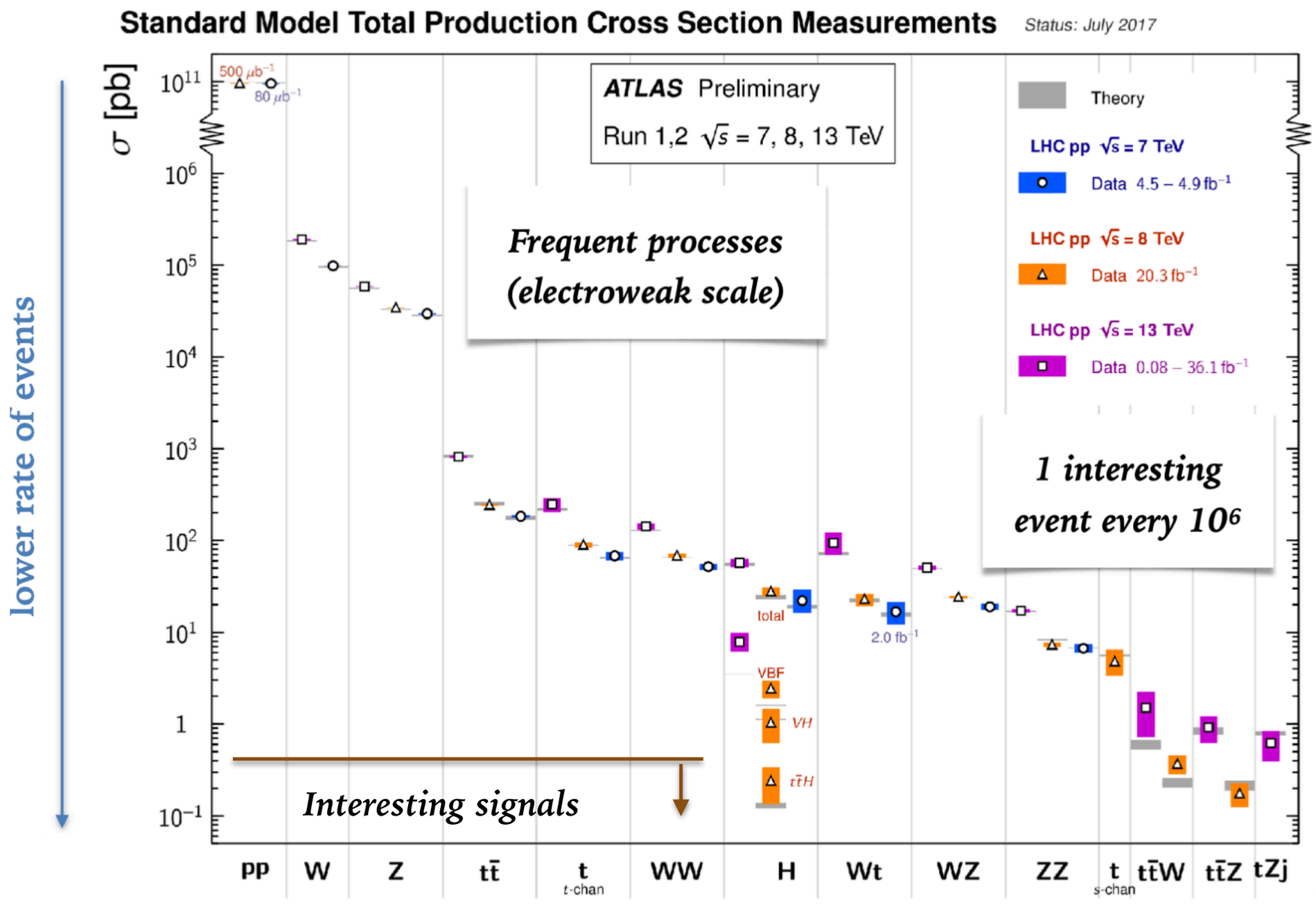
Caveat: very simple picture
See e.g. DESY's "Puzzle of Dark Matter" workshop talks for more

How would new particles manifest?

New particles: resonant excess (bump) over Standard Model background

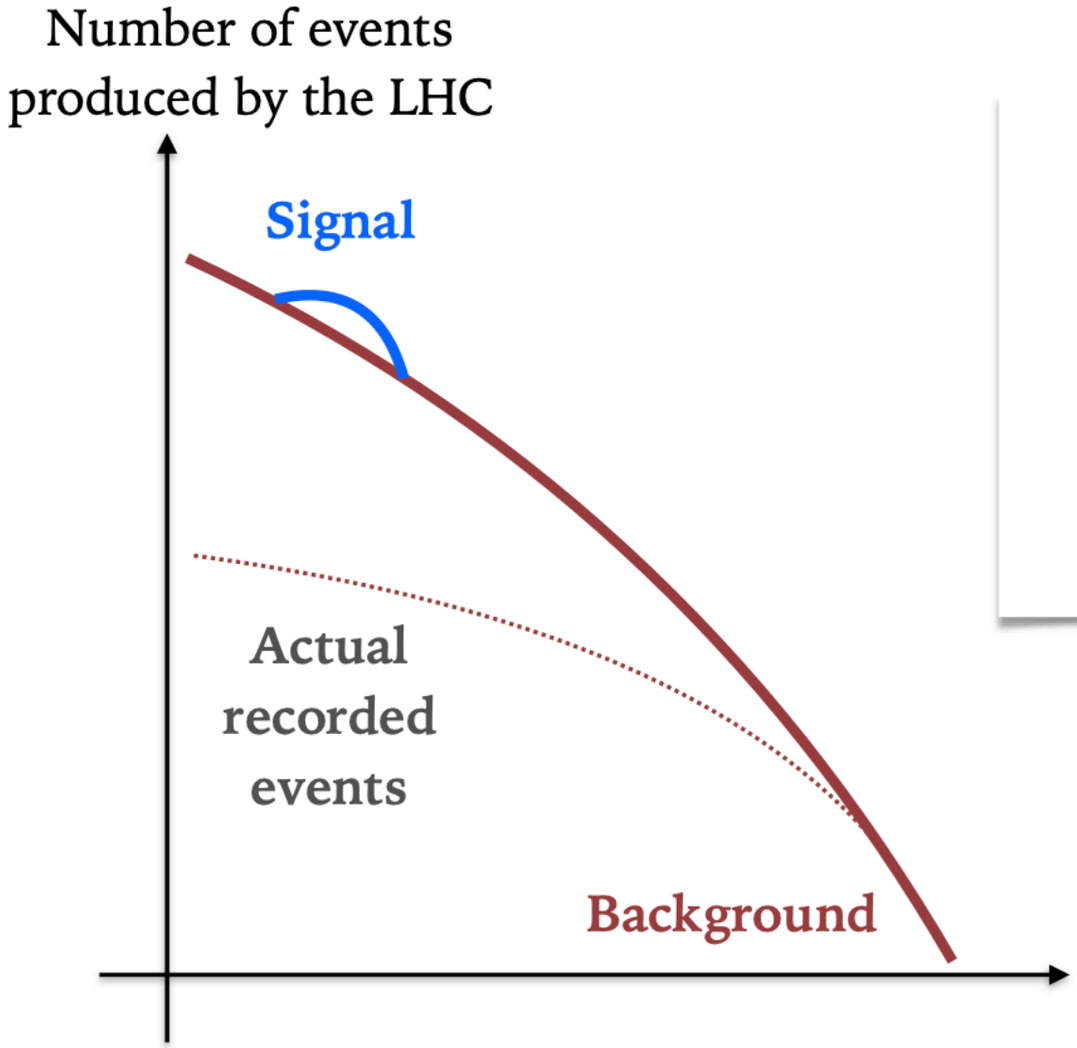


Another look at the Standard Model



Challenges for new particle searches

Main challenge for resonance searches: large backgrounds
and signal that looks very much like background

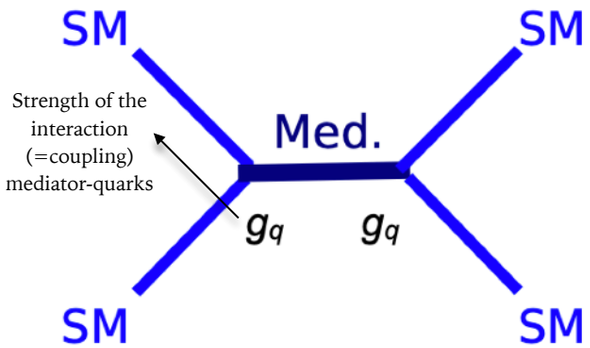


Impossible to record all events in full
→statistical error harms sensitivity!

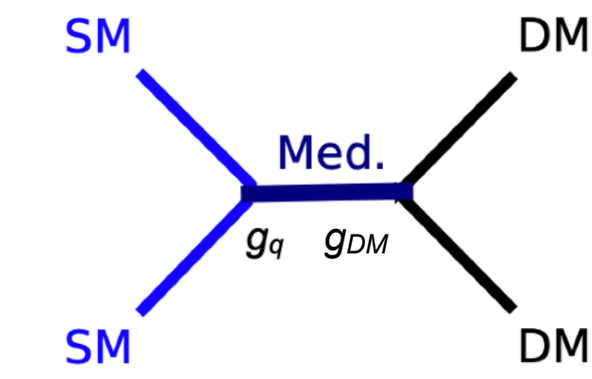
Example: dijet decays of DM mediators

Selecting interesting events works for most of the LHC physics program...
...but it is not optimal for rare processes with high-rate backgrounds:
we cannot record and store all data, and trigger discards both background and signal

This prevented us from being sensitive to low-mass DM mediators decaying into jets

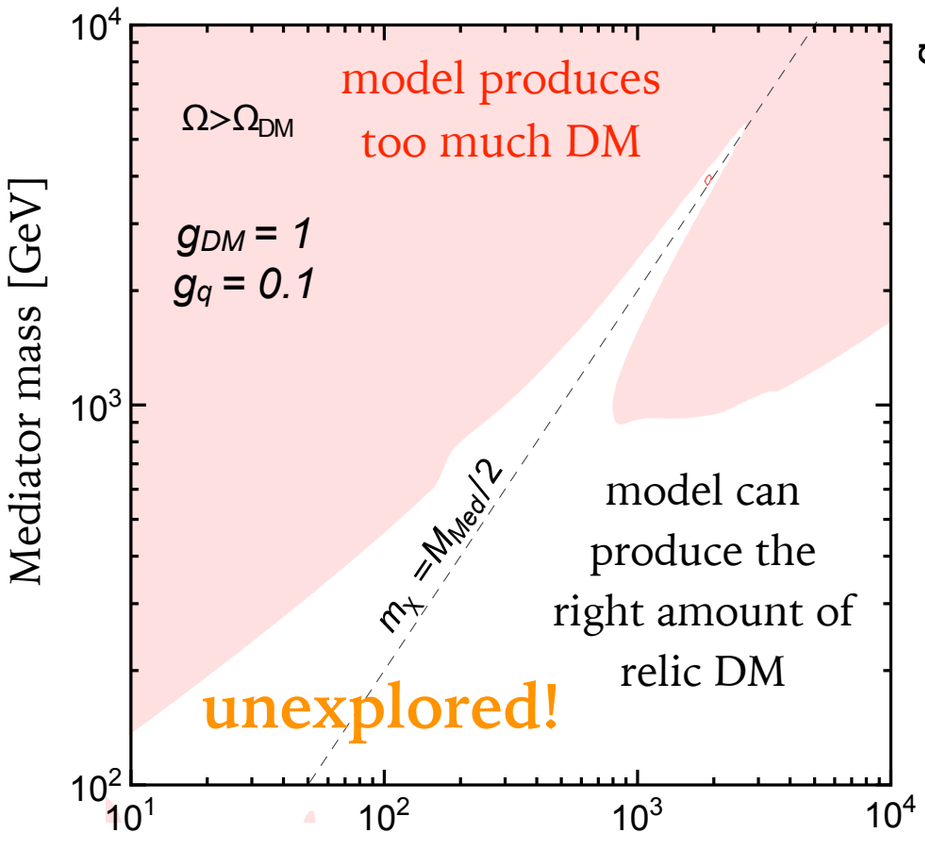


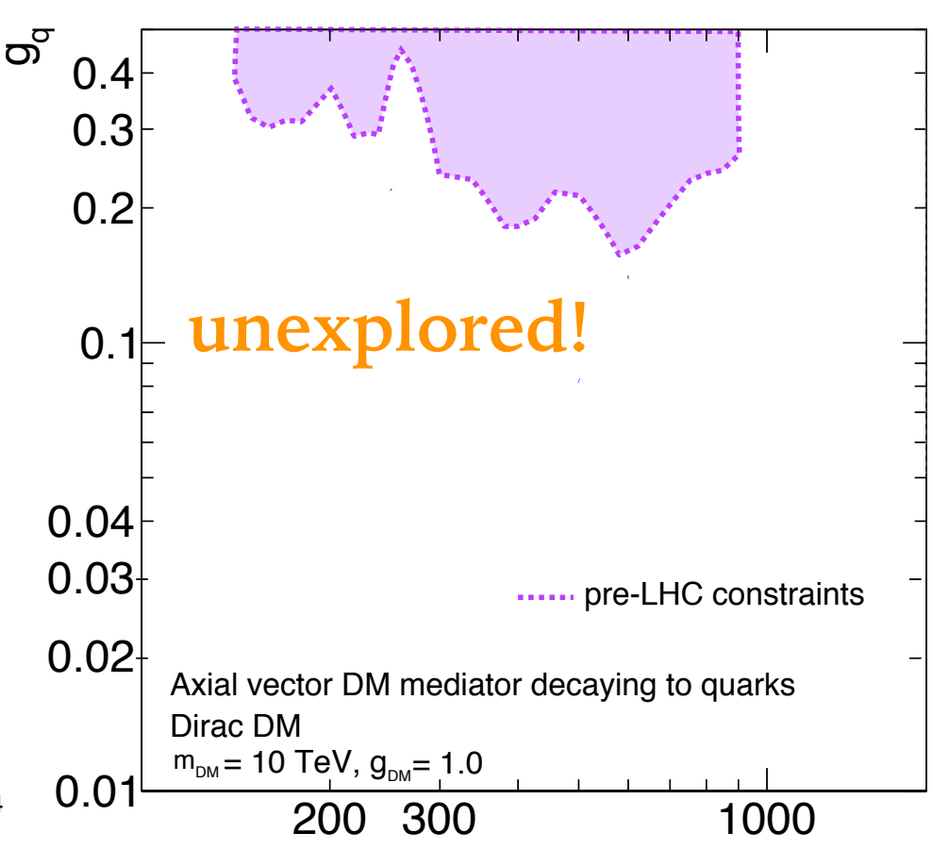
Visible mediator decays



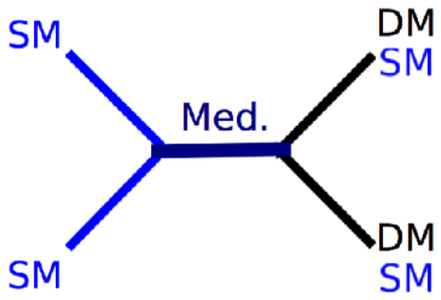
Invisible mediator decays

Adapted from JHEP 1507 (2015) 089

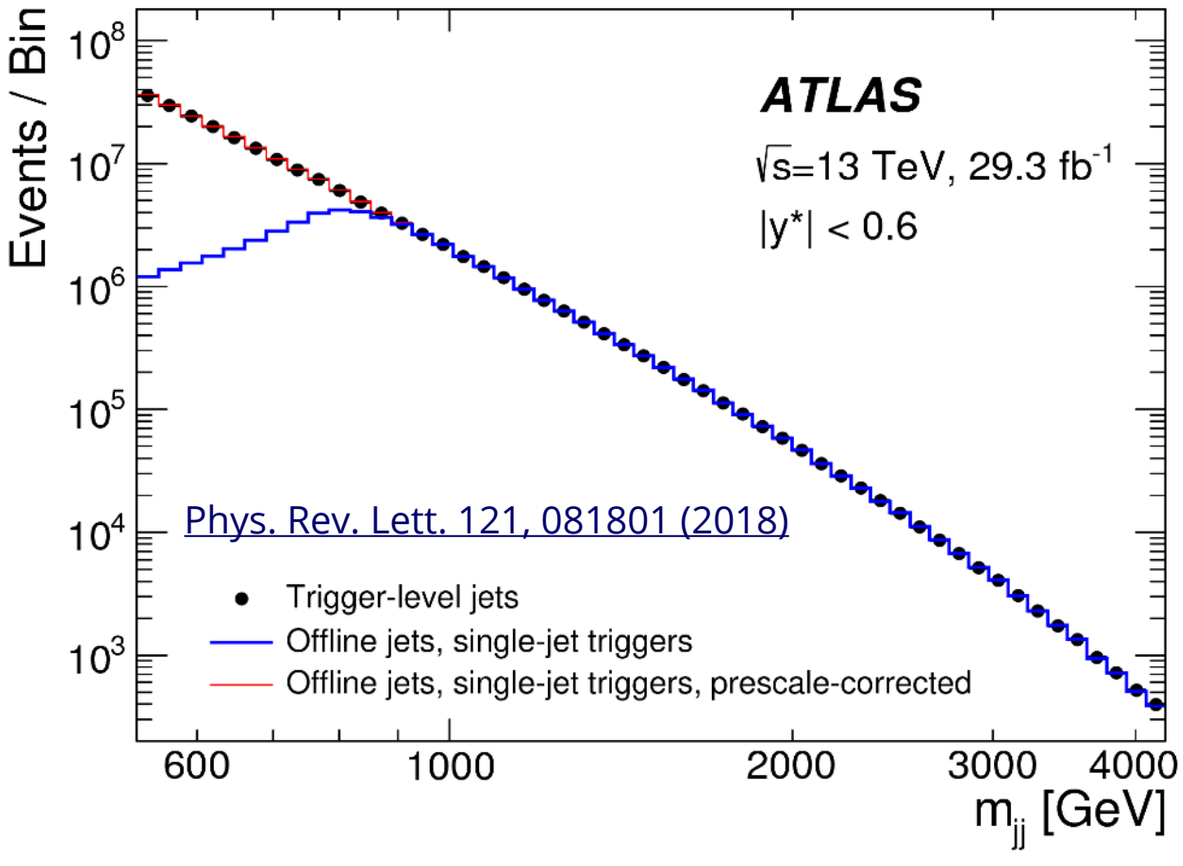




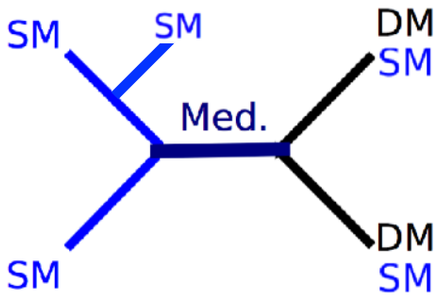
ATLAS/CMS results on DM mediator searches



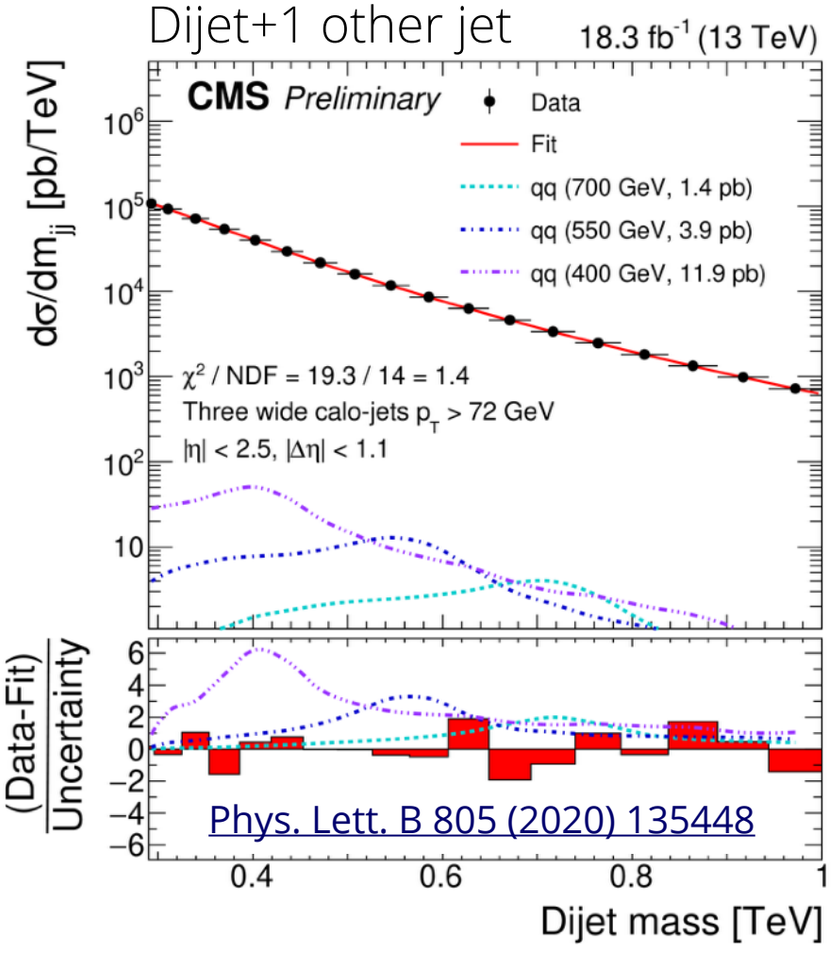
Dijet+anything



Mass reach limitation by hardware trigger…



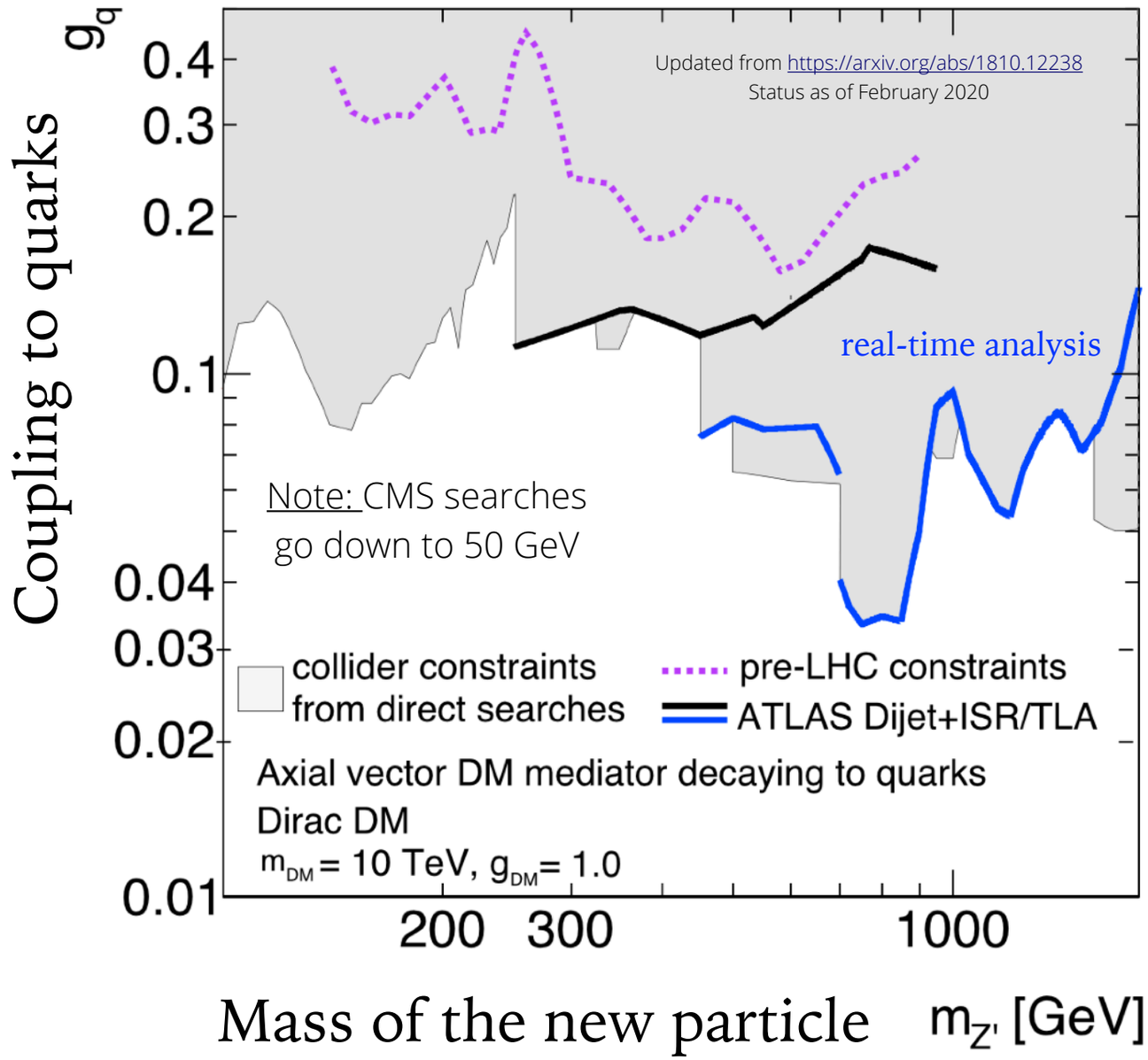
Dijet+1 other jet



…can also be overcome by choosing other signatures!

CMS and ATLAS are closing the gap at low mediator masses

Summary of public material from ATLAS and CMS, plot by C. Doglioni / W. Kalderon
Apologies for ATLAS-centricity! CMS results are equally sensitive

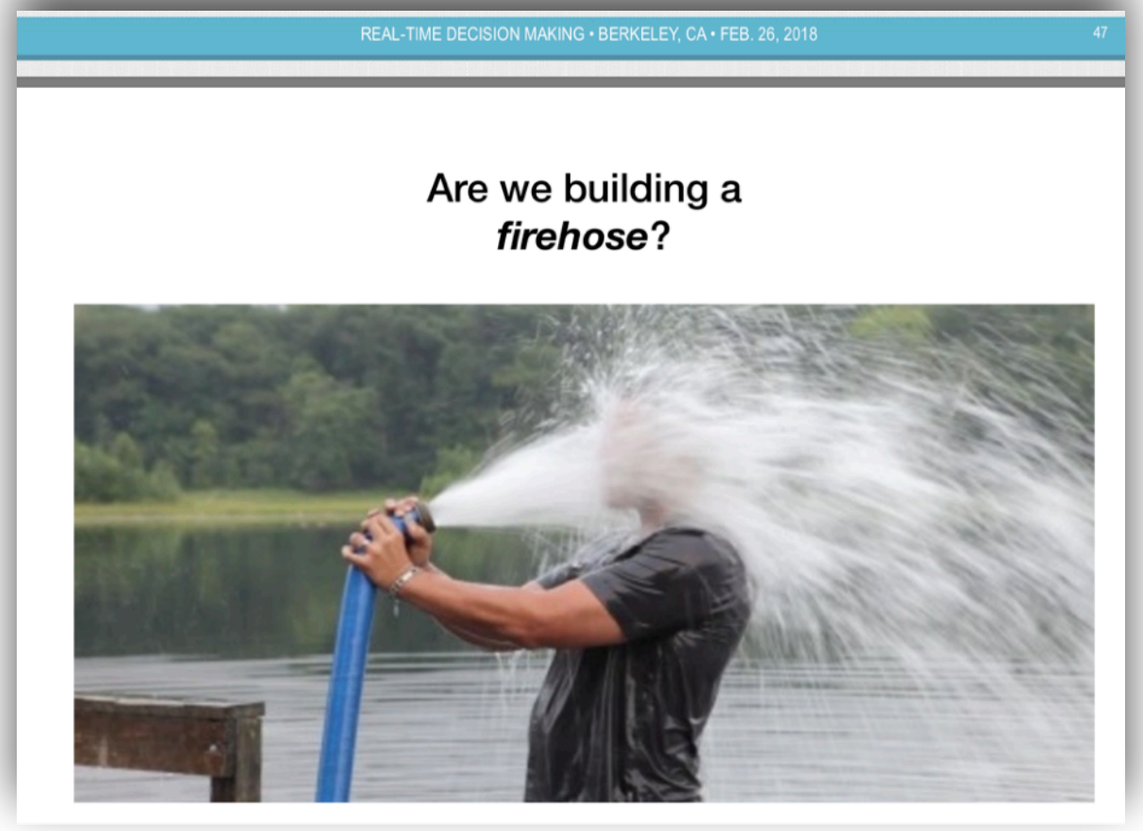


Real-time analysis beyond HEP

Parallels with astrophysics?

C. Fitzpatrick, LHCb

E. Bellm, LSST



The LHC and modern surveys are **data firehoses**

Example: common "big data" problem in large infrastructures

REALTIME advanced study group, Pufendorf Institute for Advanced Studies (2018)

	LHC (ATLAS, 2018)	LHC (ATLAS, 2026)	EuXFEL (2017)
Raw Data Size per event	1 MB	5 MB	8 MB
Events/pulses per second	40 million evts/s	40 million evts/s	27000 evts/s
Potential data generated (before selection)	o(100) GB/s	o(100) GB/s	216 GB/s
Event/pulse rate after real-time data level reduction	1000	10000	Not yet implemented
Real-time analysis event rate	Up to 20000 events/s	Not yet implemented	Not yet implemented

Table 1: Data challenges at LHC[7] and EuXFEL

about HELIOS

Helmholz-Lund International Graduate School

https://www.heliosgraduateschool.org

Real-time analysis in HEP and industry

- **Big data analysis / real-time analysis**
  - Shared tools: machine learning, hybrid computing architectures
  - Connections: IT, industry (e.g. Internet-of-Things)
    - data is mostly cheap to record: large datasets in industry as well
    - **time-to-insight** is the key metric

SMARTHEP:

International network of high energy physicists and companies for **real-time data analysis**

Main challenge tackled: **how to take decisions, fast and efficiently**

Tools:

- Hardware (FPGA, GPU) & software
- Machine learning to enable fast decisions

Conclusions

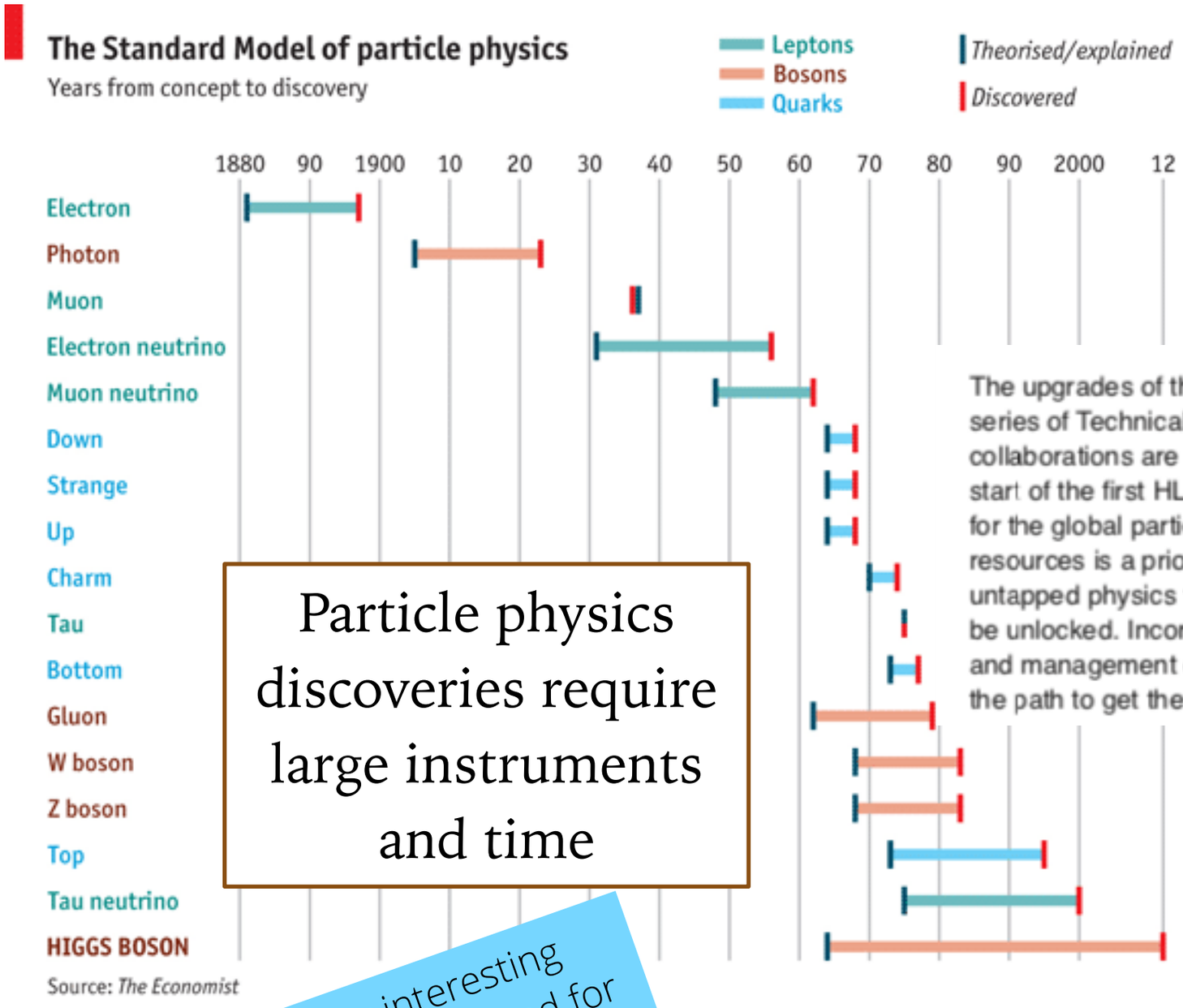
Motivating real-time analysis in HEP

- **Many different theories** can explain the shortcomings of the SM including **low-mass resonances**
  - None of these theories is yet favored by data
  - Very different detector signatures
    - Some of them buried in **high-rate backgrounds**
    - Some of them **rare but very unusual**

Probing for new physics below/at the electroweak scale where many (rare) SM particles are located requires **efficient trigger systems** and **novel data taking** techniques

Making the most of the data: **enabling discoveries** ensuring these events are recorded and analyzed

What does it take for a discovery? ~~Real~~ Time



Particle physics discoveries require large instruments and time

Many interesting upgrades planned for HL-LHC (and beyond)!

The Economist

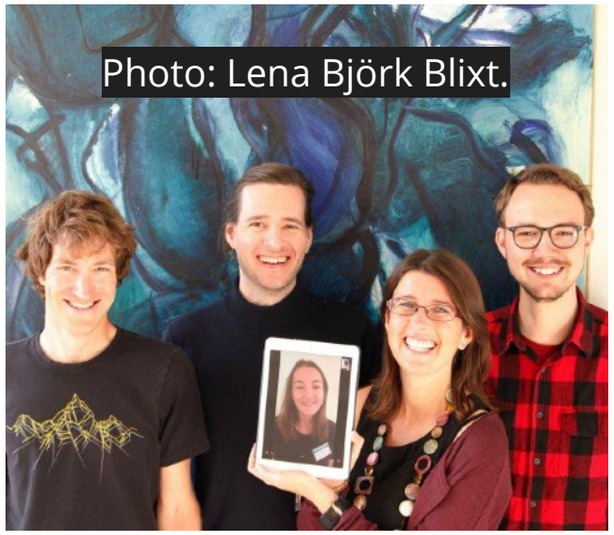
- **We aren't done taking LHC data (10x more expected)** "low-hanging fruit" checked first, expect surprises

European Strategy Update, deliberation document

> The upgrades of the ATLAS and CMS experiments have been documented in a series of Technical Design Reports and have been approved, and the international collaborations are gearing up to commission these detectors by 2027, the scheduled start of the first HL-LHC run. The timely delivery of these upgrades is a milestone for the global particle physics community, and the continued allocation of adequate resources is a priority. Based on continued innovations in experimental techniques, the untapped physics that is surely awaiting in the third LHC run and the HL-LHC era can be unlocked. Incorporating emerging new technologies into trigger systems, computing and management of big data, reconstruction algorithms and analysis methods is the path to get the best out of these upcoming datasets.

- **Real-time analysis and decision making** cross fields: let's think together and collaborate on tools & infrastructure

Thanks for your attention from the DARKJETS/TLA team



In the picture from the left:

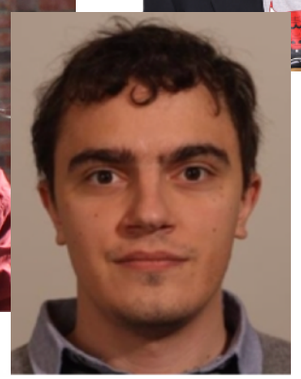
- Postdocs **William Kalderon (now BNL), Jannik Geisen** (not pictured)
- PhD students **Eric Corrigan, Eva Hansen** (remote on that day, now @ LHCb Manchester)
- Master's (now PhD) student **Alexander Ekman**

+ Lund Master's and Bachelor's students

Heidelberg
Geneva
Buenos Aires

Ohio State
Oregon
CERN

DM Complementarity plots:
Isabelle John (Lund, now Stockholm)
Boyu Gao (OSU)
Marco Rimoldi (DESY)



Collaborators & Contributors for this talk: Jannik Geisen, Alexander Ekman, Eva Hansen, Eric Corrigan, Will Kalderon, Antonio Boveia, Monica Dunford, Steven Schramm, David Strom, Suchita Kulkarni, Marie-Helene Genest, Dilia Portillo, Nathal Lalloué, Deepak Kar, Sukanya Sinha, Conor Fitzpatrick, the SMARTHEP network & many others

Backup slides

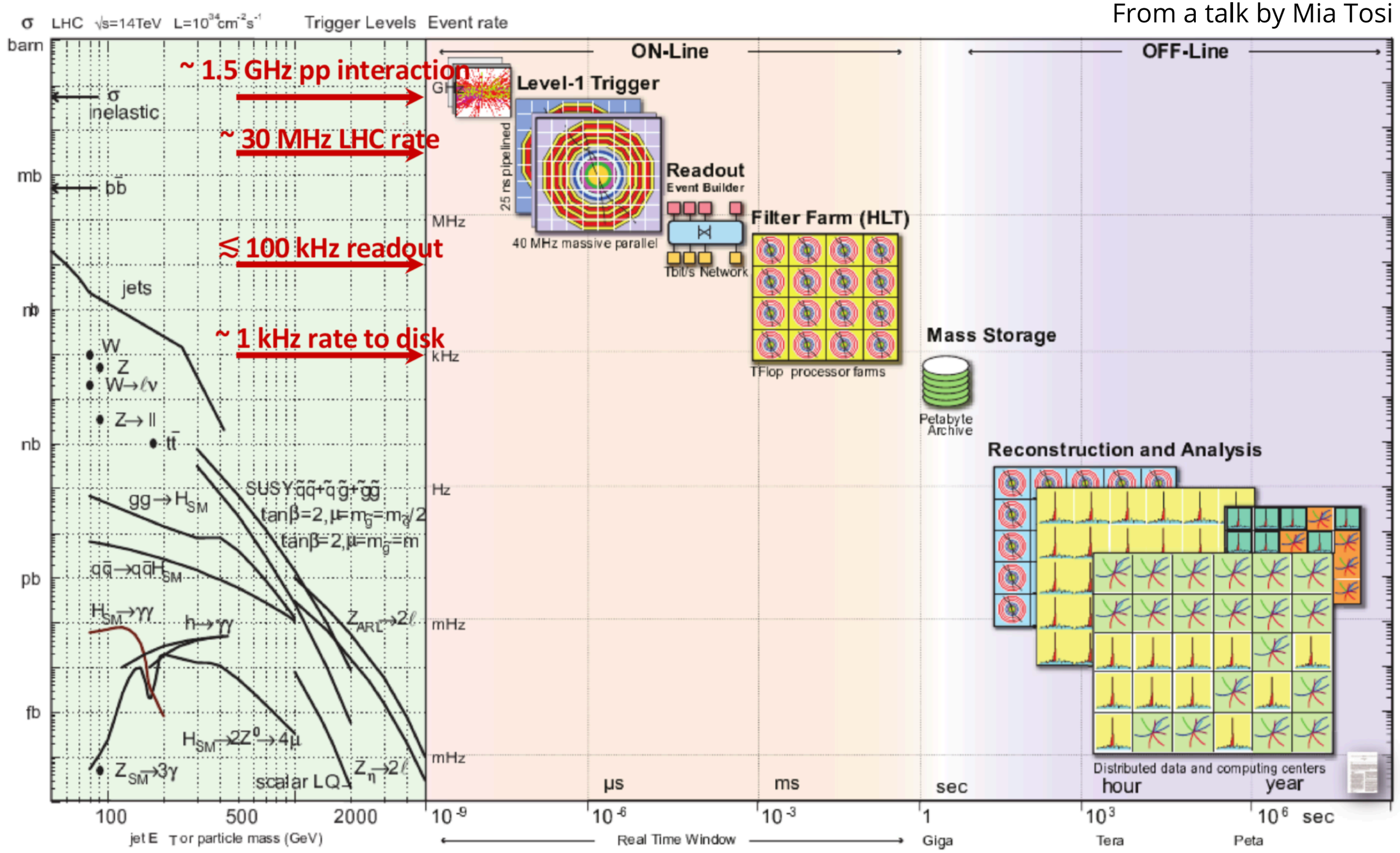
LUNDS
UNIVERSITET

Video: LHC collisions and data selection

CERN-MOVIE-2013-041-001

Real-time analysis, in the CMS trigger

From a talk by Mia Tosi



Which workflow to choose? (a rough guide)

My analysis is limited by HLT and I have *relatively* simple objects/backgrounds
- use **real-time analysis**

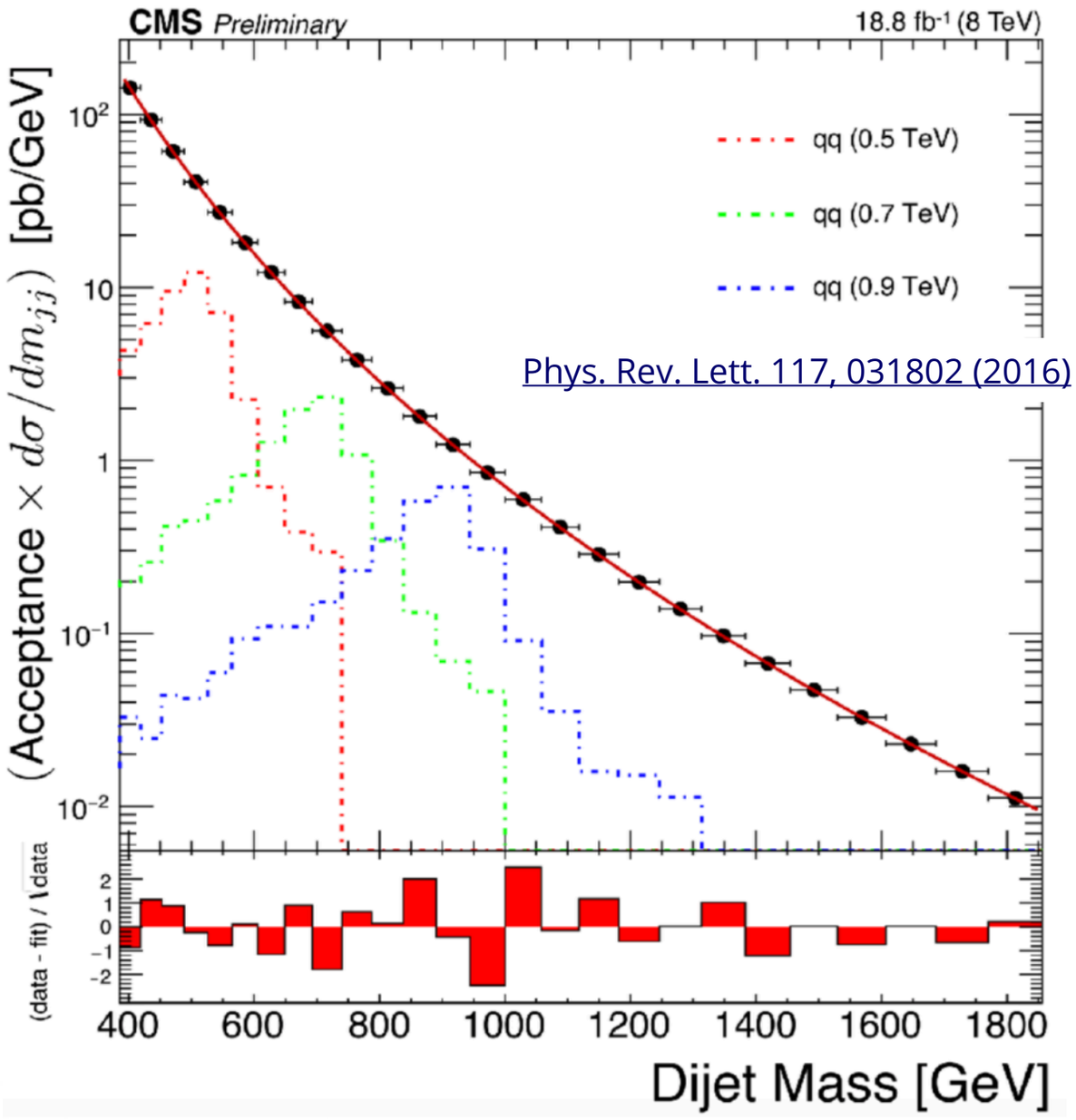
I have a real-time analysis going and I am sure I'll discover something
- use partial event building to keep more raw data behind the objects
- use **delayed stream** as a safety net (CMS)

My analysis is limited by HLT and I have more complicated objects
- use r**eal-time analysis** and **partial event building** to look into region of interests

My analysis is limited by HLT but I still need the full event
(and I don't mind when I get it)
- use **delayed stream**

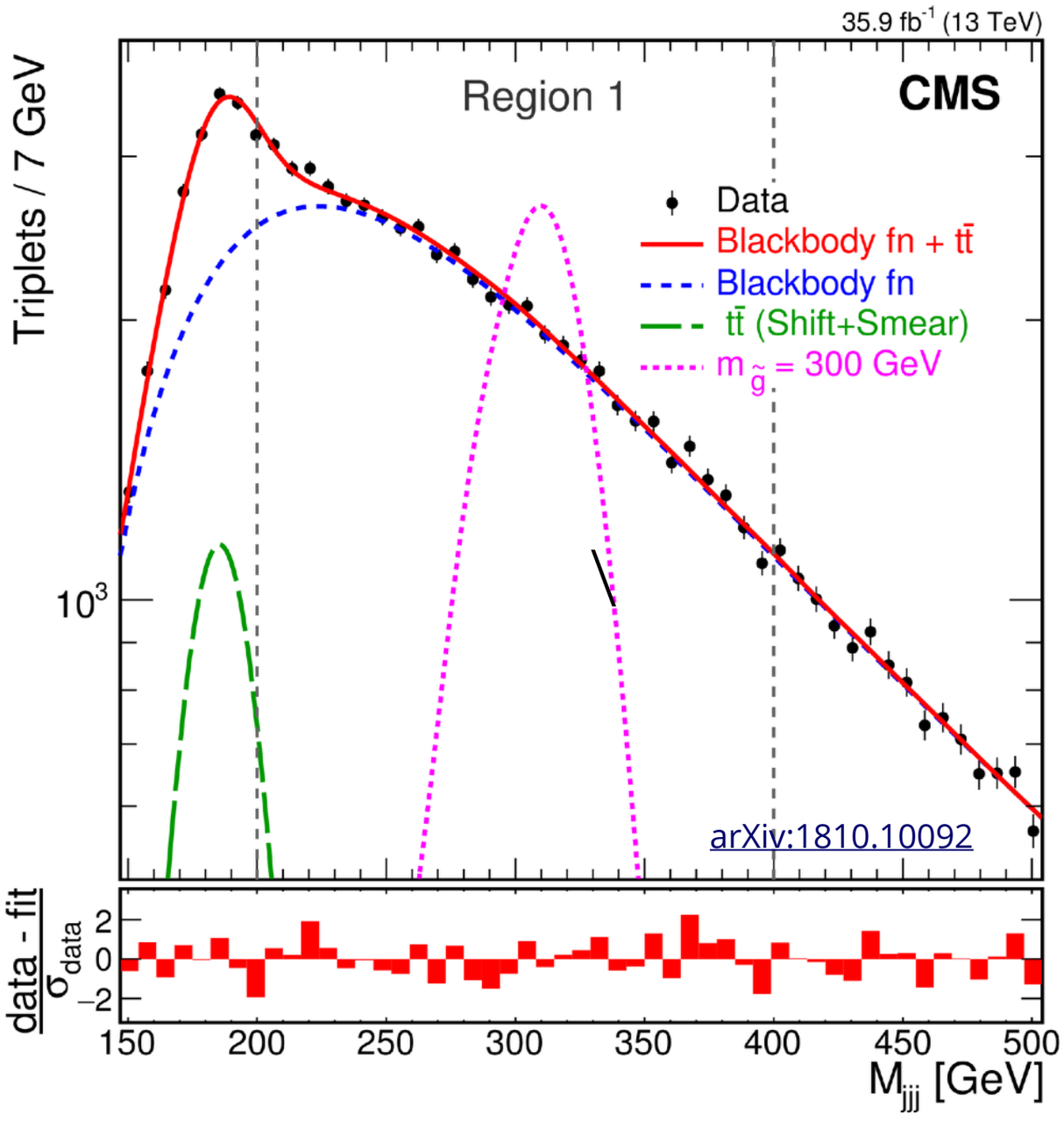
More CMS results from jet searches

First results of this technique for dijets at 8 TeV...



Phys. Rev. Lett. 117, 031802 (2016)

...now also extended to three-jet searches



arXiv:1810.10092

Turbo/Data Scouting/TLA path

Turbo stream (LHCb),
Data Scouting (CMS),
Trigger-level Analysis (ATLAS),

Customizable output data @ LHCb:

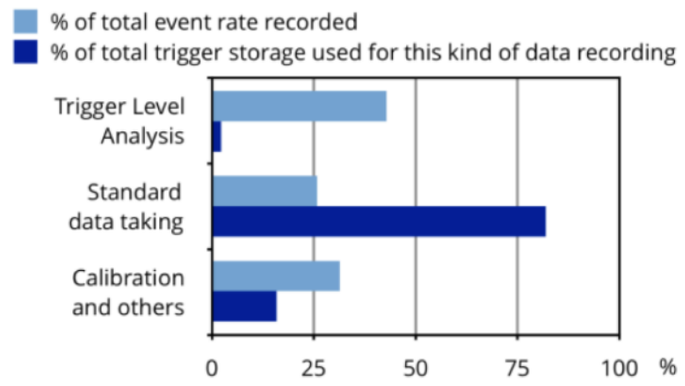
- keep trigger objects only (7 kB)
- keep trigger objects + ”on-demand” raw and/or reco in selected regions (< 200 kB)
- keep everything (200 kB)

Objects and data sizes @ CMS:

Stream	Rate (Hz)	Event Size	Bandwidth (MB/s)
PhysicsMuons	420	0.86 MB	360
PhysicsHadronsTaus	345	0.87 MB	300
ScoutingCaloMuon	4580	8.9 KB	40
ScoutingPF	1380	14.8 KB	20

Selected CMS stream rate, event size, and bandwidth at the beginning of LHC Fill 7334 (23 Oct. 2018, L≈1.5×10^{34}cm^{-2}s^{-1})

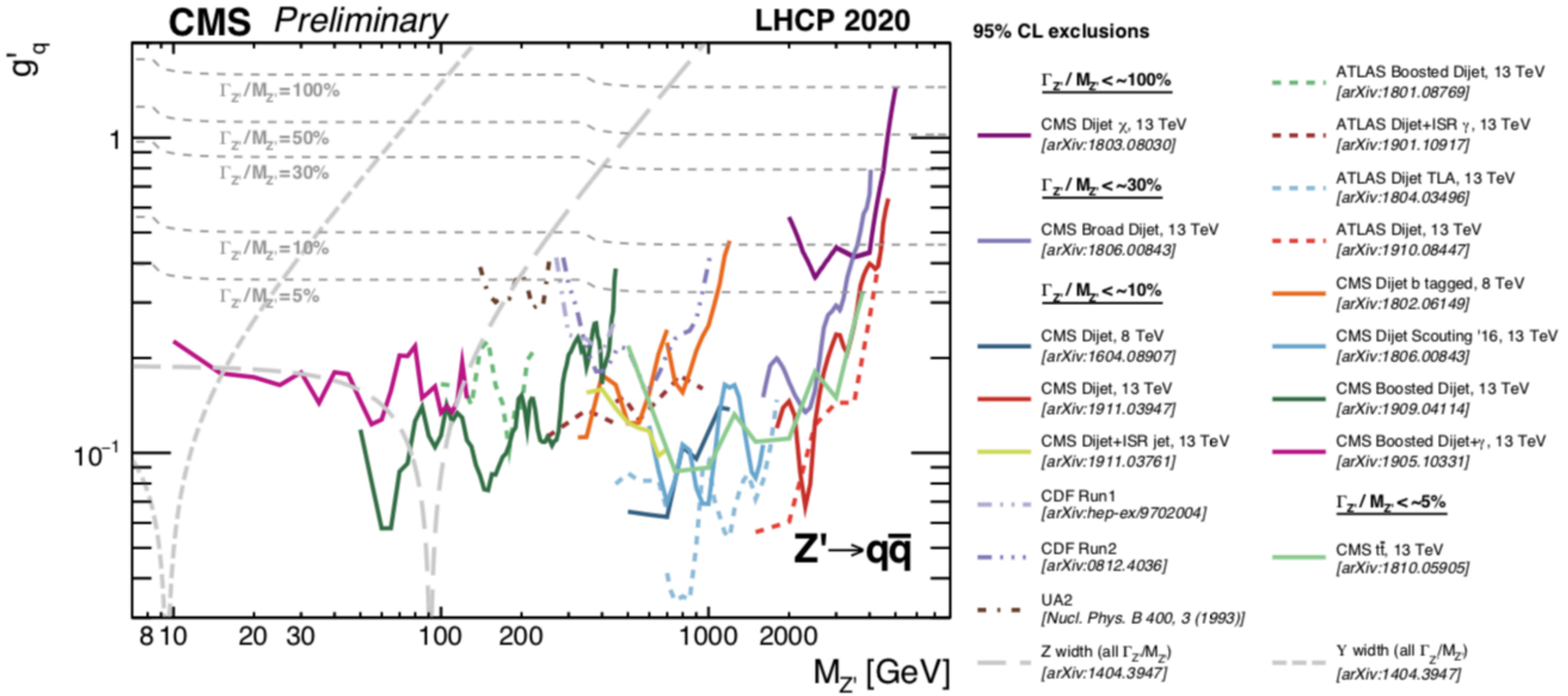
Data sizes @ ATLAS:



Information from ATLAS Trigger Operation plots, 2017

JLab HSF workshop session on RTA: LHCb, ALICE, ATLAS, CMS

Mediator mass-coupling summary plot

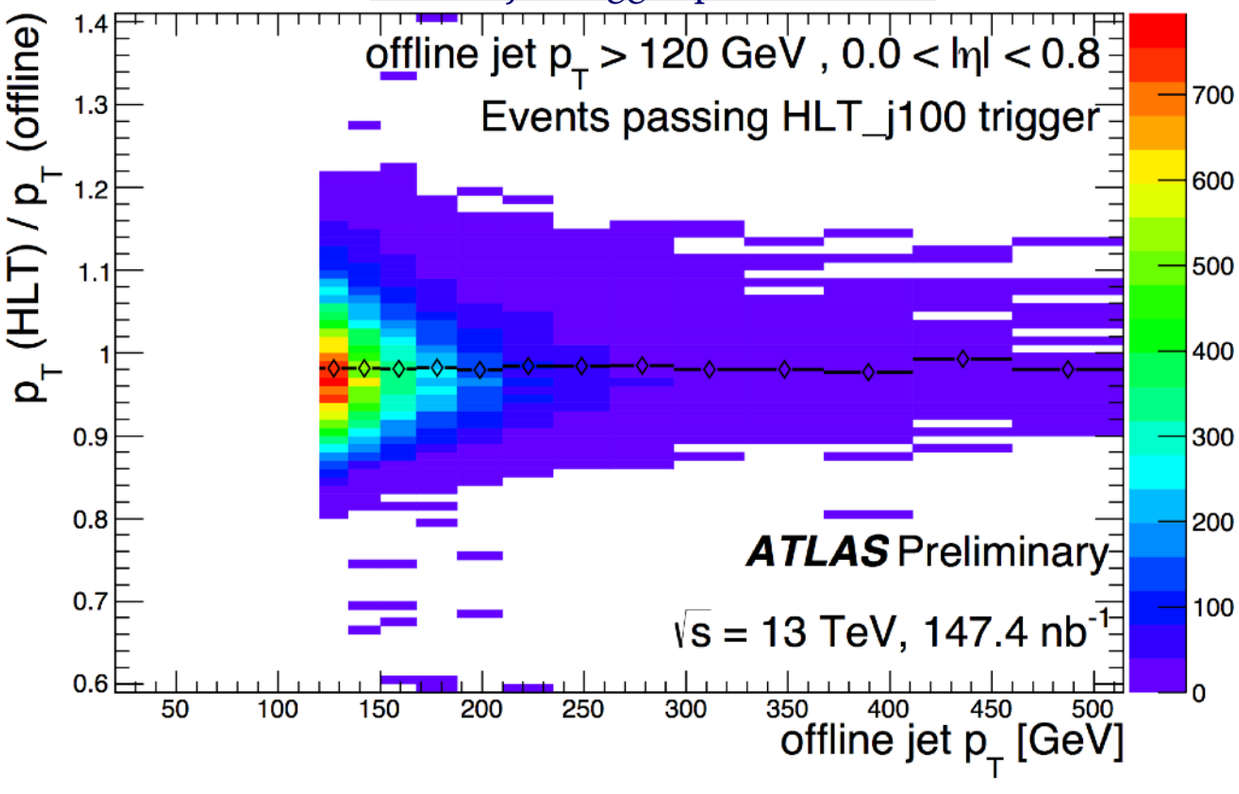


Are trigger jets good enough?

…they have to be, to be able to detect very small signals!

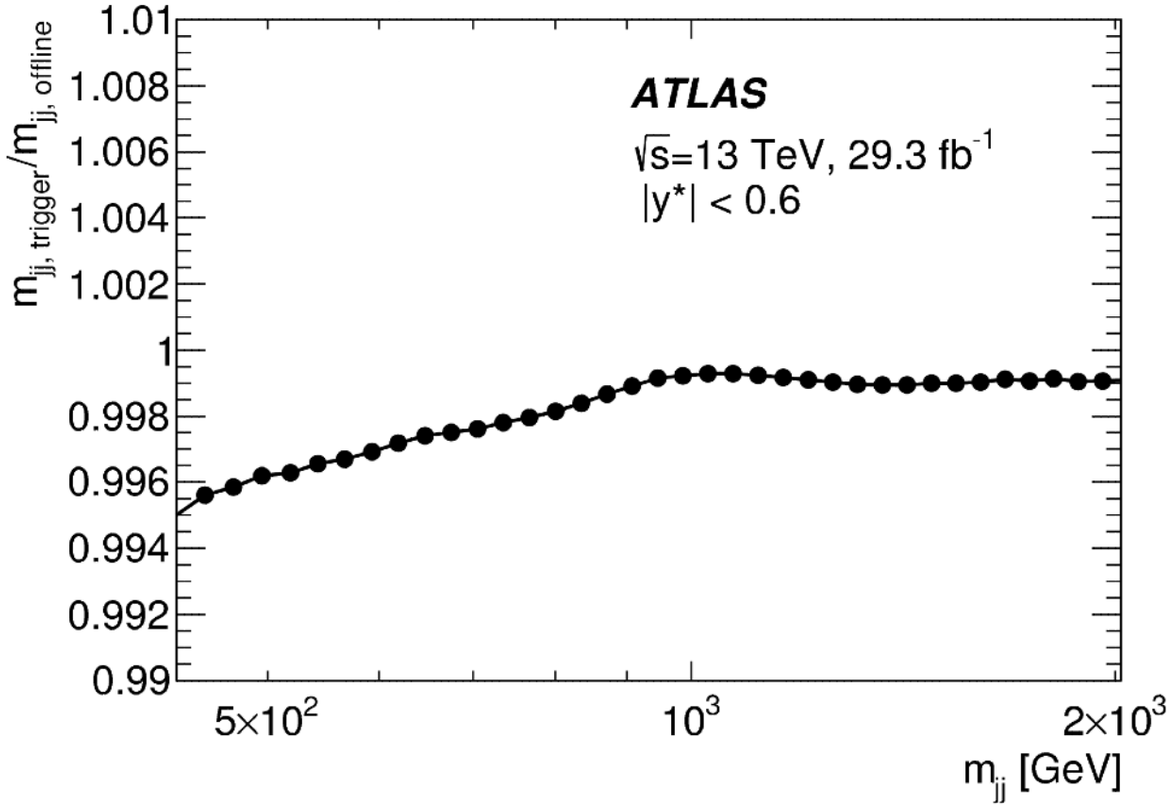
September 2015 (interview)

ATLAS jet trigger public results



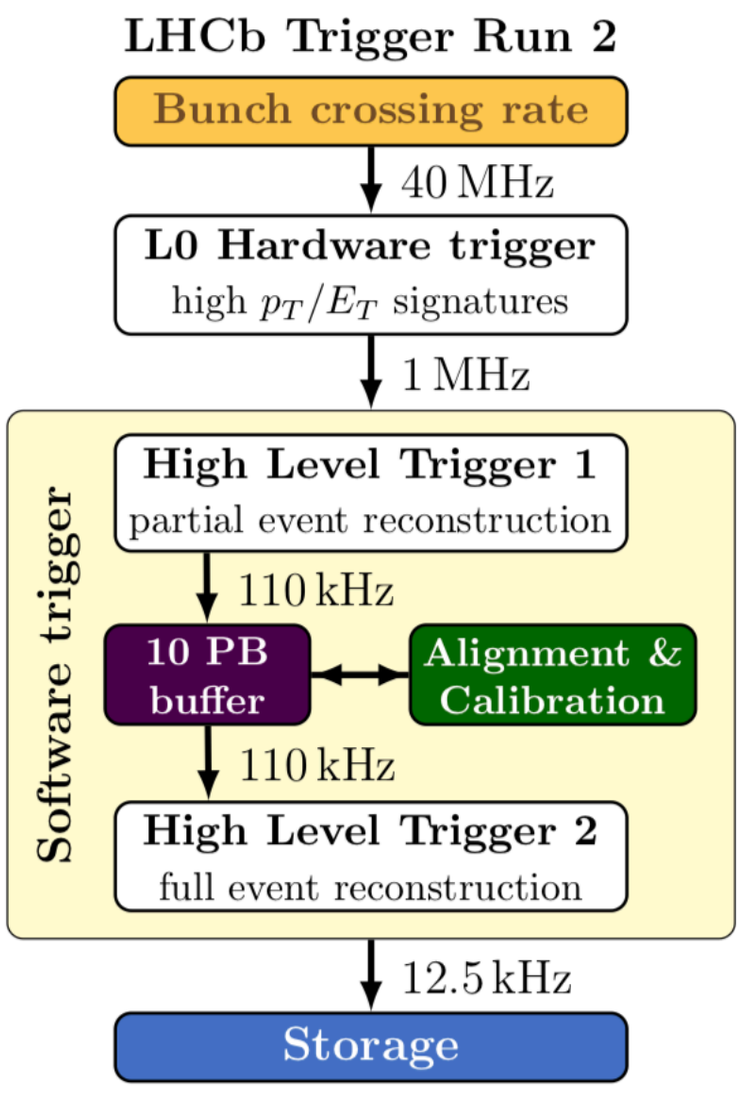
May 2017 (paper submission)

Phys. Rev. Lett. 121, 081801 (2018)

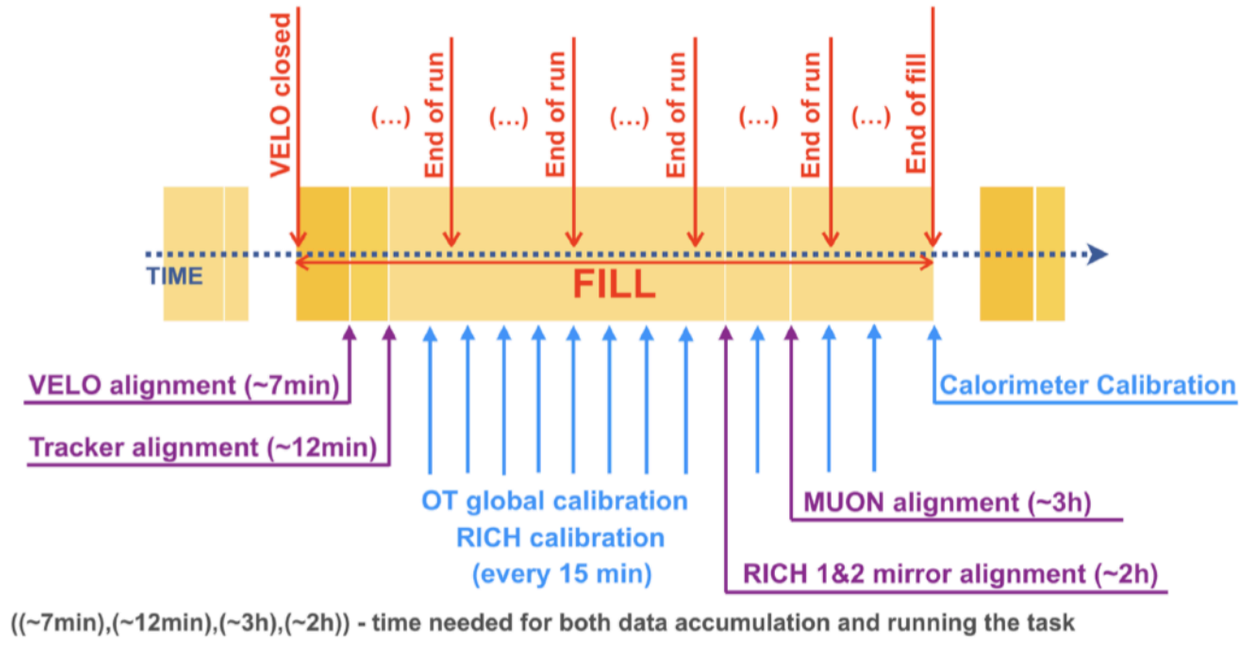


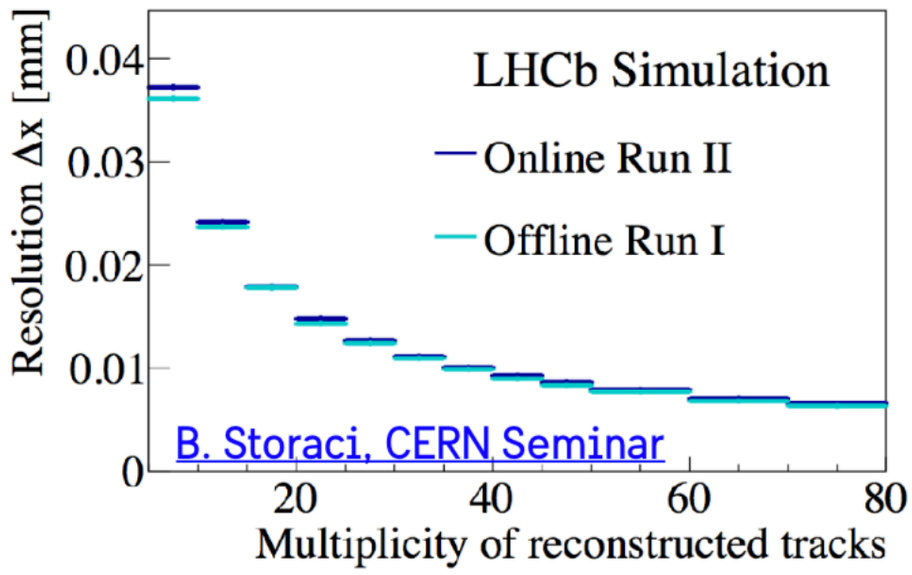
(Note the change in y-axis scale)

LHCb online vs offline reconstruction



arXiv:1812.10790





Real-time data analysis requires real-time detector alignment and calibration, in computing farm

This is also done in ATLAS and CMS

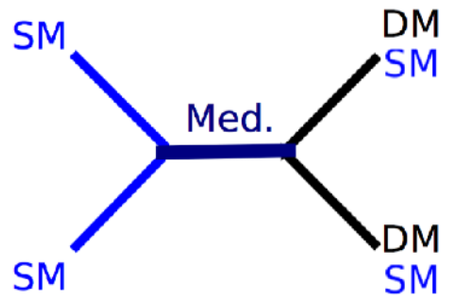
Choice of benchmarks

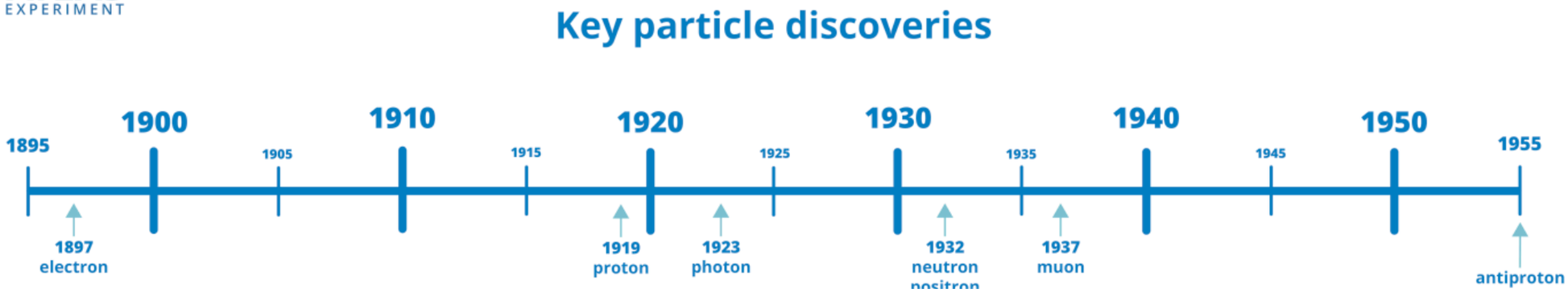
https://abstrusegoose.com/406

"Why should we choose/believe the simplest models?"

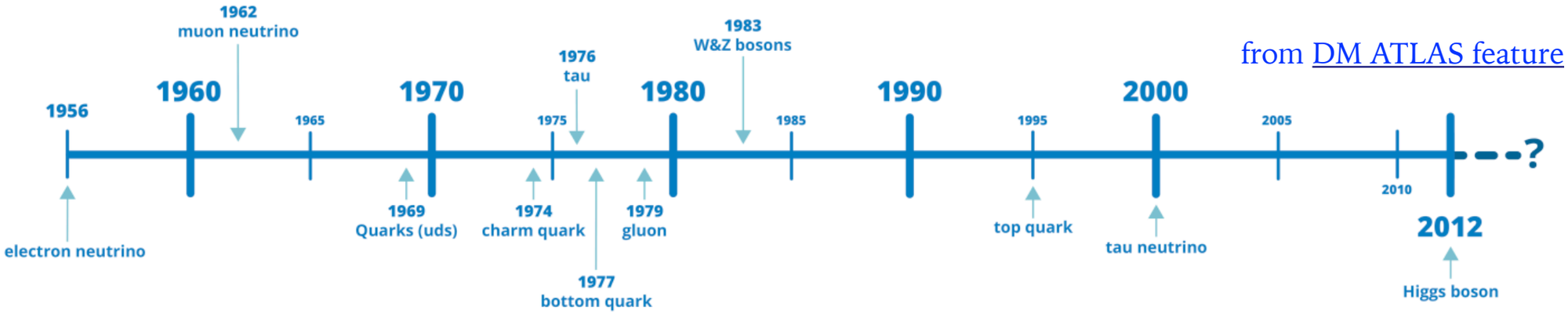
"Do we think DM is all made of a single WIMP model?"

(not really…see dark sectors!)





- **Lesson from SM:** most common particles discovered first



from DM ATLAS feature

- Even simple models can encapsulate **relevant experimental characteristics** representing wider classes of theories

as long as we are aware that they can be more rare than what we choose as example

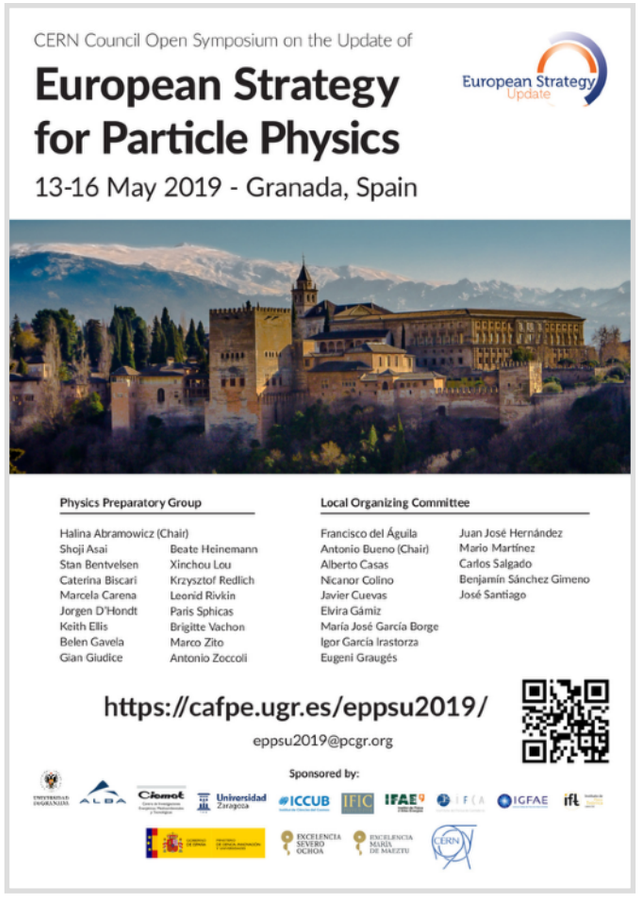
Possibilities for cross-talk in collaborative environments

- **High Energy Physics Software Foundation**
  - Not only LHC experiments
  - Forum for physicists with interest in software for HEP
  - Instrumental for creation of IRIS-HEP NSF effort
  - Working groups including trigger & reconstruction
    - Website
    - Mailing list: hsf-forum@googlegroups.com (google group)
- **DarkMachines**
  - Collective of astro/particle physicists interested in machine learning tools for dark matter
  - Various (self-organized) efforts, including unsupervised searches
    - Website
    - Subscribe to main mailing list

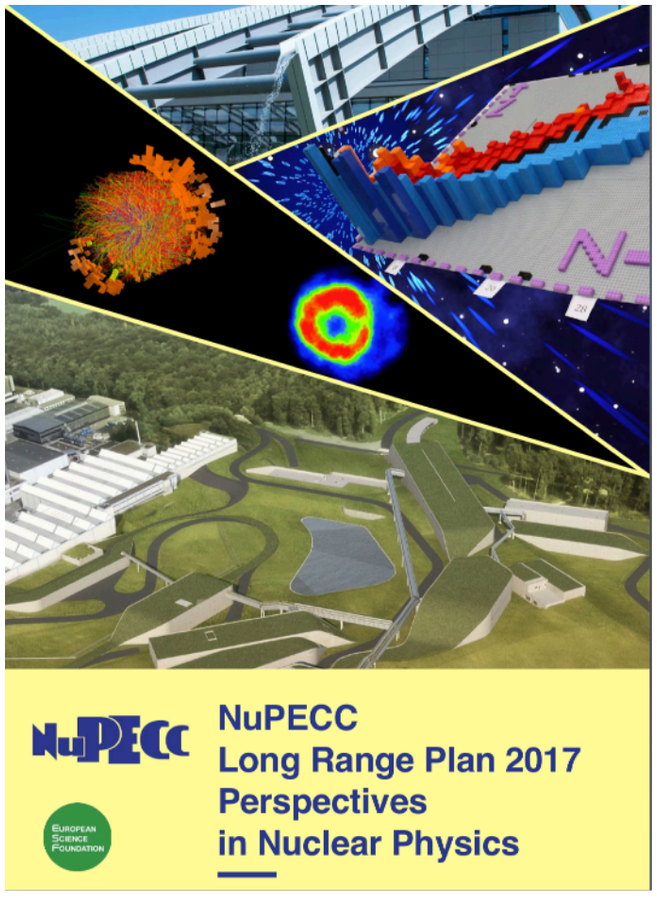
Finding synergies

Astroparticle (APPEC)

Particle (ECFA)



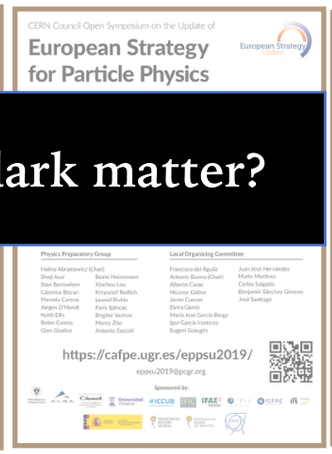
Nuclear physics (NuPECC)

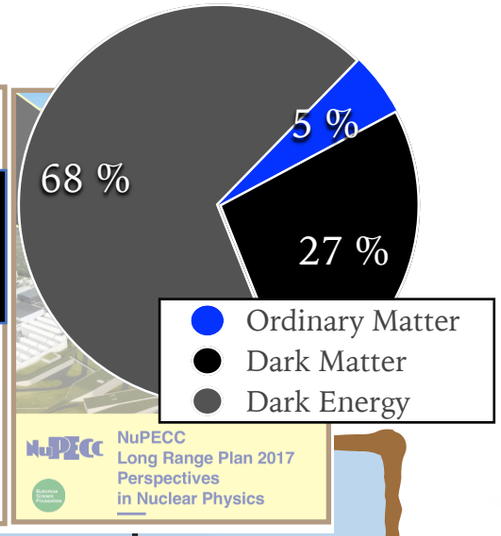


Astroparticle, particle and nuclear physics in Europe have **strategies and plans** that **recognize the importance of synergies** between the different fields

US: ***Snowmass*** effort has started, (for *DM at Colliders* WG:

Foundations needed to exploit synergies





Common theory ground

instrumentation (accelerators, beams, detectors, vacuum & cryogenics, control & automation...)

data acquisition, computing, data sharing & open science

HSF HEP Software Foundation

ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

iris hep

& more...

EuCAPT

[Talk at EPS-HEP / ECFA session 2019](), [CERN EP Newsletter]()

[Slides from European Strategy Update release, 19/06/2020]()

European Strategy Update

2020 Strategy Statements

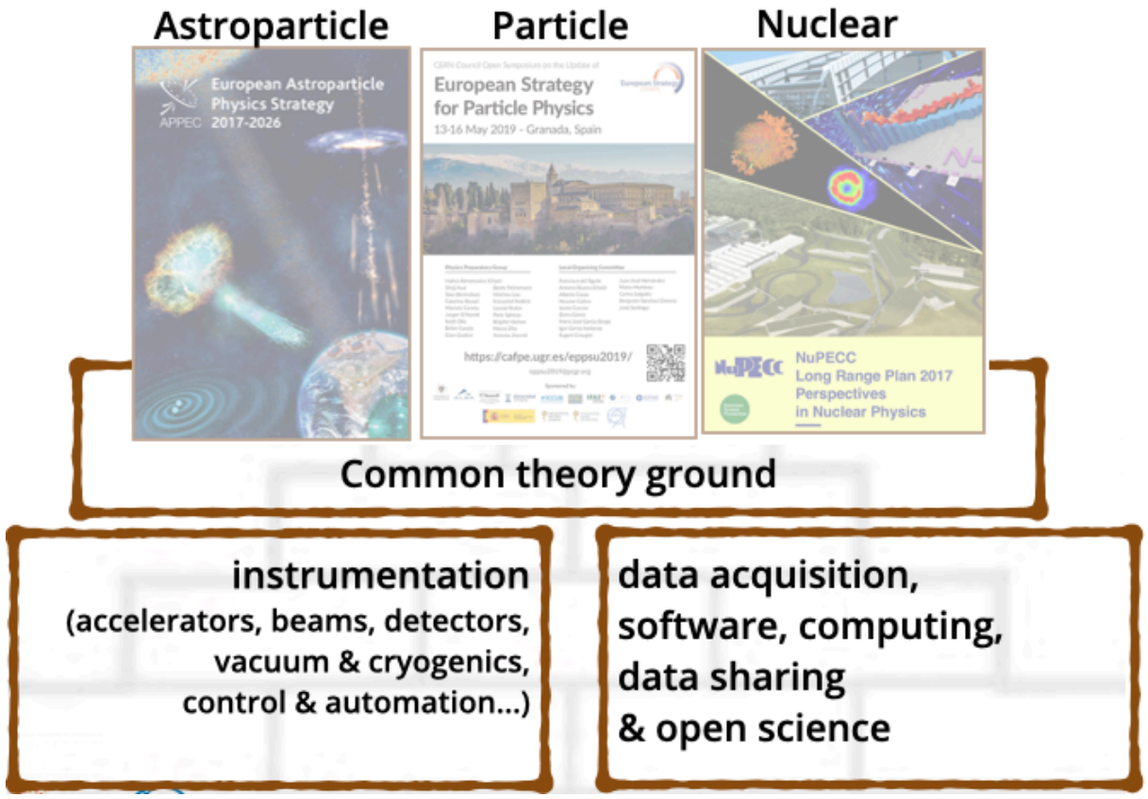
5. Synergies with neighbouring fields

Particle and Astroparticle Physics

- Synergies exist at the level of infrastructure, detectors, computing, interaction models and physics goals (ex.: neutrinos, dark matter, cosmic rays and gravitational waves)
- The need to foster these synergies has been clearly identified in the national inputs

Two ongoing projects focused on Dark Matter

searches & interpretation



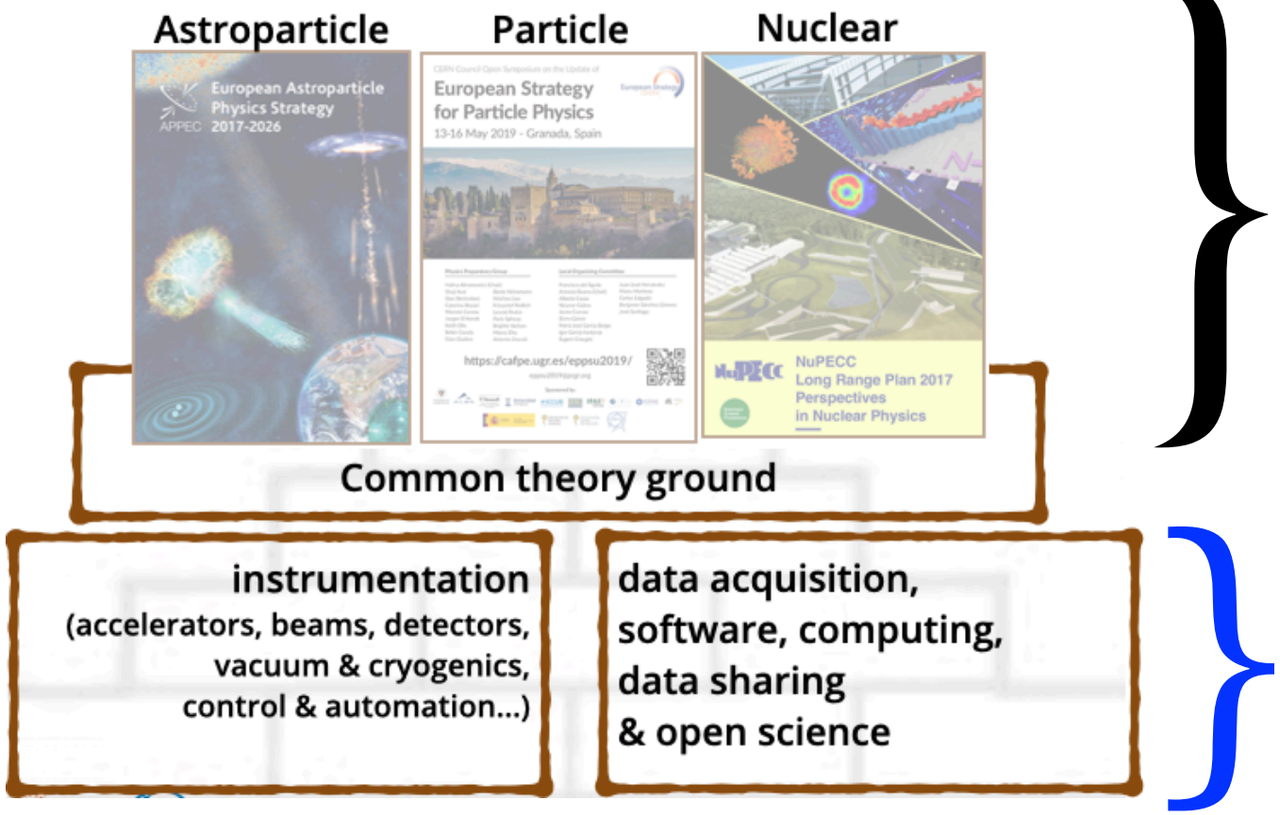
JENAS EoI: Initiative for Dark Matter in Europe and beyond: Towards facilitating communication and result sharing in the Dark Matter community (iDMEu)

https://indico.cern.ch/event/869195/
ESCAPE newsletter APPEC newsletter

build a discussion platform and tools to facilitate collaboration of existing groups/ efforts on **dark matter searches** and **interpretation**

Two ongoing projects focused on Dark Matter

searches & interpretation



JENAS EoI: Initiative for Dark Matter in Europe and beyond: Towards facilitating communication and result sharing in the Dark Matter community (iDMEu)

https://indico.cern.ch/event/869195/
ESCAPE newsletter APPEC newsletter

build a discussion platform and tools to facilitate collaboration of existing groups/ efforts on **dark matter searches** and **interpretation**

software & data

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

HSF
HEP Software Foundation

Towards a Dark Matter Test Science Project

ESCAPE Progress Meeting, 2020

compare **end-to-end analysis workflows** for WIMP searches, towards their implementation in a common **Software Catalogue** and as input to the design of the **European Open Science Cloud**

European Research Council
Established by the European Commission

ATLAS EXPERIMENT

LUND UNIVERSITY

Real-time analysis (well) beyond HEP

Welcome to our REALTIME ASG blog! We're a group of researchers interested in real-time data acquisition and decision making as well as open data working as an Advanced Study Group 2019-2020 at the Pufendorf Institute for Advanced Studies.

- **Connections to astrophysics**
  - Real-time alerts for interesting events
- **Connections to accelerator beams / physics / engineering**
  - Beam steering, "triggering" for synchrotron and laser experiments
  - Sharing of technology, hybrid architectures (FPGA/GPU)
- **Connections to social sciences & law**
  - Real-time data deposition promotes open data and credibility of science
    - Implications for treatment and ownership of data treatment
    - Sharing of tools (versioning / data persistency)

- Ask Me! about the REALTIME Advanced Study Group

open to discuss further collaborations!