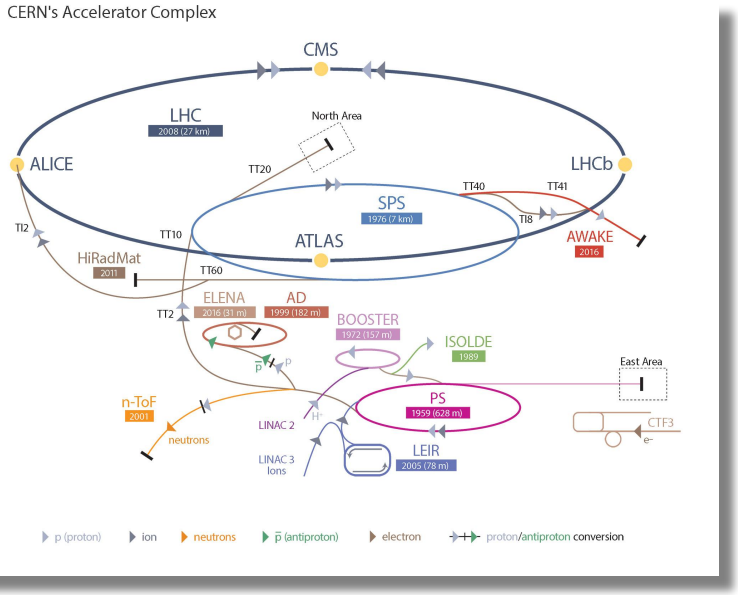
Adatelemzés és számítástechnika

Krasznahorkay Attila

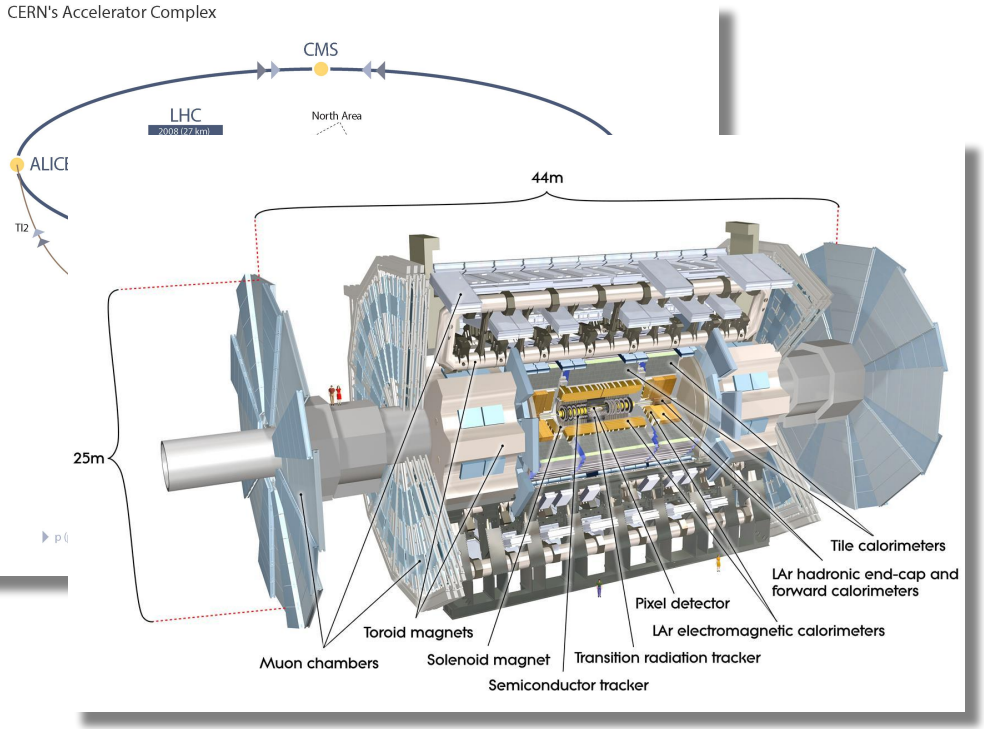
Emlékeztető

A madártávlat



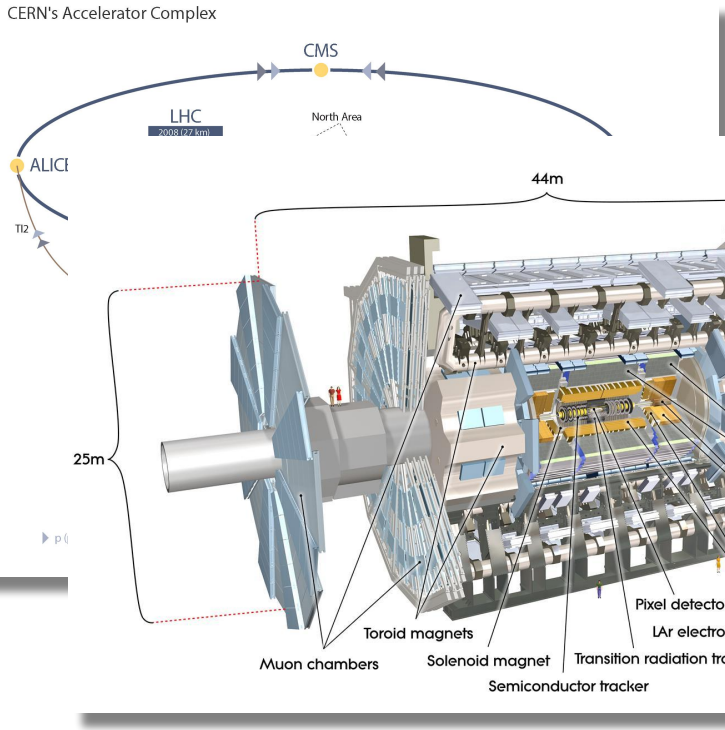
Létrehozzuk az “érdekes” reakciókat
(Varga Dezső előadása)

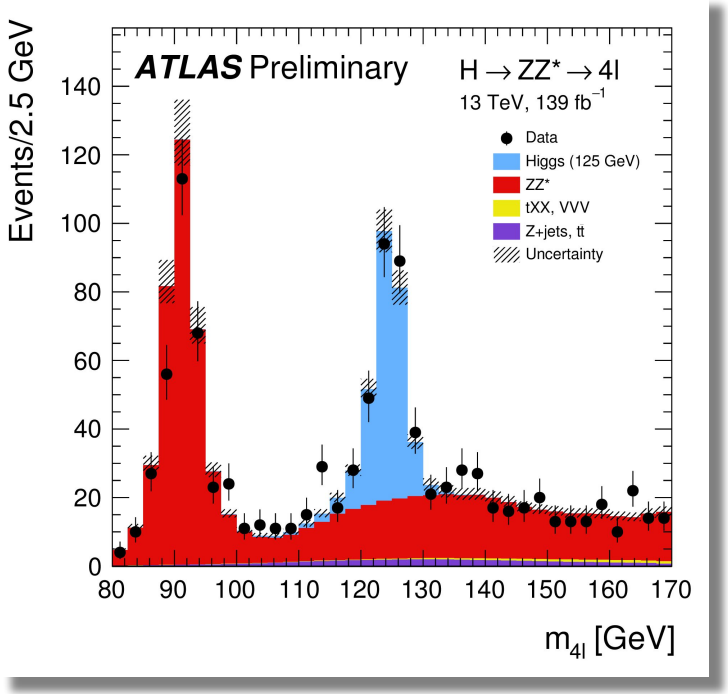
A madártávlat



Érzékeljük a kijövő részecskéket
(Varga Dezső előadása)

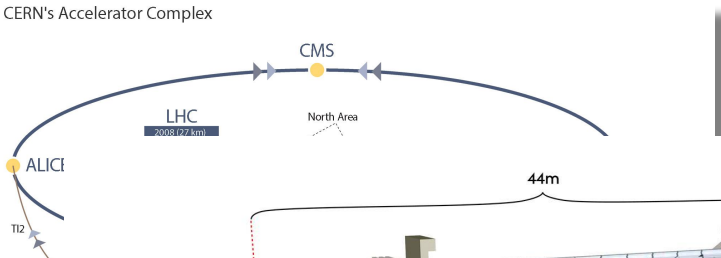
A madártávlat

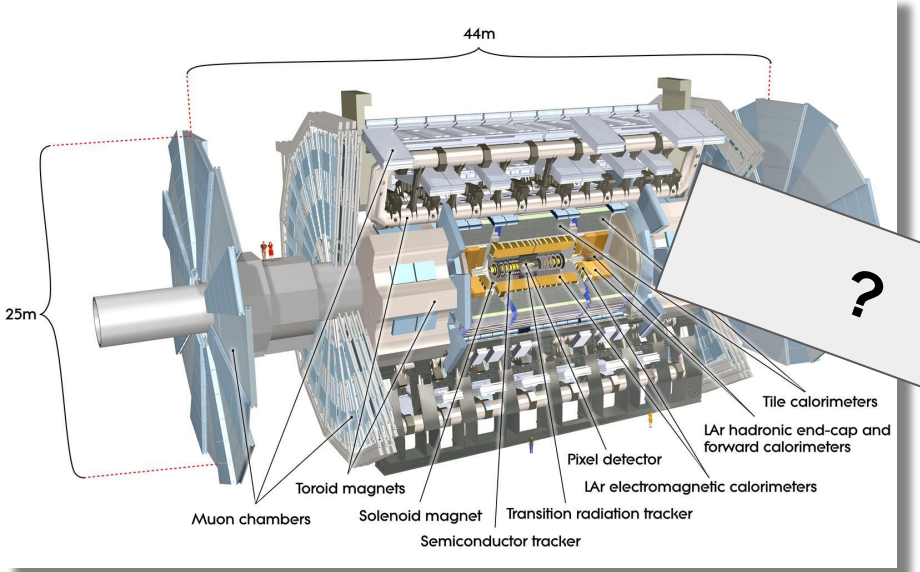




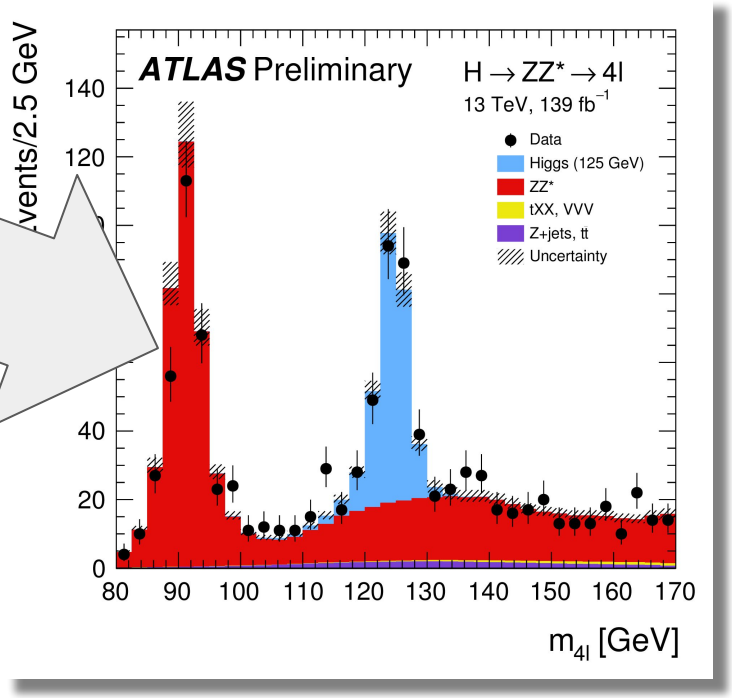
Levonjuk a fizikára vonatkozó következtetéseket
(Újvári Balázs előadásai)

A madártávlat

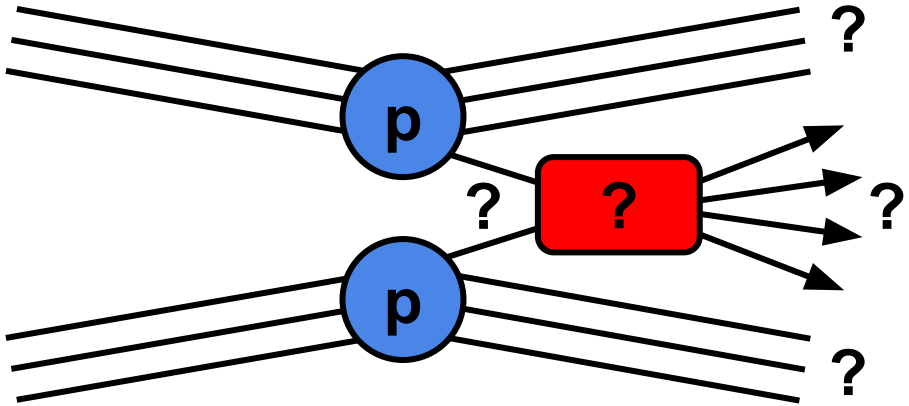




?



A kísérlet



- Amire igazán kíváncsiak vagyunk az LHC-nál az a piros dobozban végbemenő folyamat
- Az alapján amit a reakcióból kijönni látunk, következtetünk vissza, hogy:
  - Valóban proton-proton ütközést figyeltünk-e meg
  - Milyen reakció zajlott le pontosan

Az adatgyűjtés

- A detektorokkal egymástól független “eseményeket” rögzítünk
  - Ahogyan Dezső bemutatta, minden detektor típus rá jellemző módon tudja érzékelni azt ha különböző részecskék kölcsönhatásba lépnek vele
  - A rögzített információt, mint egy fényképezőgép CCD kameráját, kiolvassuk
- Az LHC másodpercenként >1 milliárd proton-proton ütközést hoz létre, amit mindet nem rögzíthetünk
  - Kb. 40 millió proton csomag ütközés jön létre másodpercenként, mindegyikben 40-60 proton-proton ütközéssel
  - Egy válogató (trigger) rendszerrel választjuk ki a fizika szempontjából legérdekesebb eseményeket, kb. 1000 darabot másodpercenként. Ez néhány GB per másodperces adatgyűjtéshez vezet.
- A “nyers” adatokat a CERN számítóközpontjába küldjük, ahonnan biztonsági másolatok készülnek róla a világ több pontjára
  - Évente több milliárd eseményt rögzítünk így

Az adatfeldolgozás

- A nyers adatokat erre fejlesztett célprogramokkal “rekonstruáljuk”. Így állapítva meg, hogy milyen részecskék milyen tulajdonságokkal repültek át (és a legtöbb esetben semmisültek meg) a detektoron
  - Megkapjuk, hogy mennyi elektron, müon, hadronzápor, stb. keletkezett az egyes eseményekben, és milyen tulajdonságokkal
- Az egységesen rekonstruált adatokat több lépésen keresztül analizáljuk, az egyes analízisek kívánalmainak megfelelően
  - Például kereshetünk két leptonos eseményeket kevés “hiányzó energiával” ha Z bozonos eseményeket akarunk vizsgálni
  - A kiválasztott eseményekben egyenként kiszámíthatunk különböző komplex paramétereket (például az eseményben feltételezett Higgs bozon tömegét), amiknek az eloszlásából statisztikai analízissel a háttérben zajló fizikára tudunk visszakövetkeztetni.

Történelmi kontextus

Adat-analízis a múltban

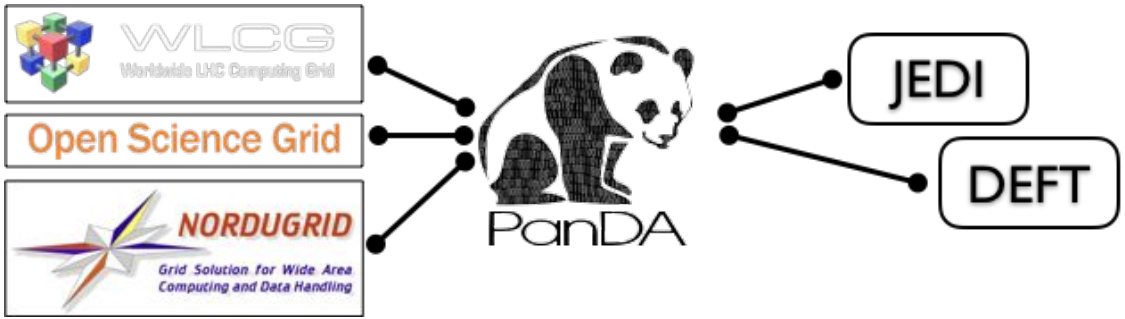
- Érdemes látni, hogy hogyan is jutottunk idáig
  - A jelenlegi komplex adatfeldolgozás nem egyik percről a másikra lett kitalálva…
- Milyenek voltak a kísérletek a múltban?
  - **Legkorábban:** Fényképek és besugárzott minták szemmel tanulmányozása
    - Gáz és buborékkamra képeket ellenőriztek egyenként, betanított munkával
  - **Később:** Kevés paraméter (pl. energialeadás, pozíció) mérése, és közvetlen analizálása
    - Képesek voltak a különböző detektorokból kijövő elektromos jeleket olyan formára hozni, hogy azokból a detektor által érzékelt részecskék milyenségére tudjanak következtetni
    - De a jeleket lehetetlen volt egyesével rögzíteni, a kor számítástechnikája ezt nem engedte meg
    - Sokcsatornás analizátorokkal elektronikusan készítettek eloszlásokat a mért paraméterekből
  - **Végül:** Amint a mért elektronikus paramétereket egyenként rögzíteni lehetett, az adatok feldolgozása a kísérlet elvégzésétől későbbre tolódhatott
    - Lehetővé téve, hogy kifinomult adatfeldolgozással pontosabb méréseket lehessen elvégezni

Számítástechnika a múltban

- A CERN első “valódi” számítógépe 1958-ban állt Üzembe
  - Programokat és adatokat lyukszalagon tudott kapni, és természetesen mai szemmel nagyon lassú volt
- Az évek folyamán egyre nagyobb és nagyobb teljesítményű szinguláris gépeket helyeztek üzembe
  - https://cerncourier.com/a/computing-at-cern-the-mainframe-era
- A nagy változást a személyi számítógépek megjelenése hozta
  - Ezeket használni lehetett az adatgyűjtés irányításától egészen az adatfeldolgozás végső lépéséig
  - Azóta is a személyi számítógépek uralják a fizikai kísérleteket...

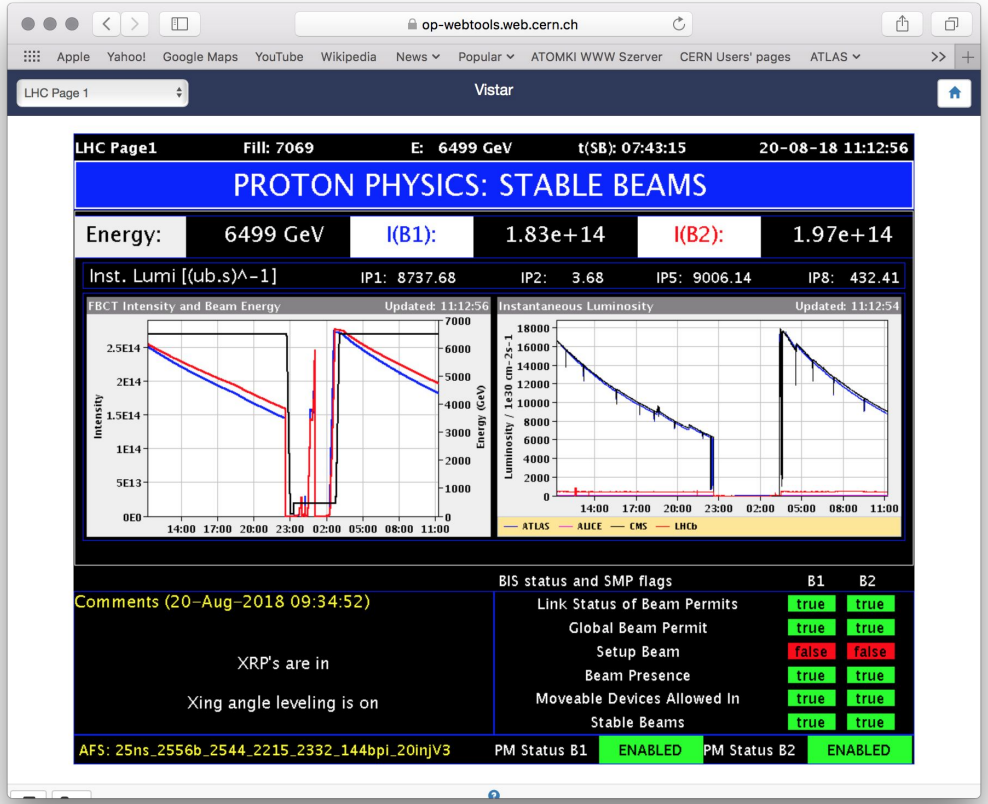
A grid létrejötte

- A személyi számítógépek előretörésével hamar megjelentek az ezekből épített klaszterek
  - Ezek arra lettek kitalálva, hogy ugyanazt a feladatot, egyenként kis változtatásokkal, sok gépen/processzoron tudják egyszerre futtatni
  - A LEP adatait is ilyen klasztereken dolgozták fel annak idején
- Az LHC tervezésénél hamar világossá vált, hogy nem célszerű egy nagy klaszterre bízni az adatfeldolgozását → A világ sok egyéni klaszterét kell inkább összekötni egy egységes rendszerré
  - Amire úgy lehet feladatokat küldeni, hogy nem kell tudjuk az adatok hol vannak, és a programunk pontosan hol fog lefutni

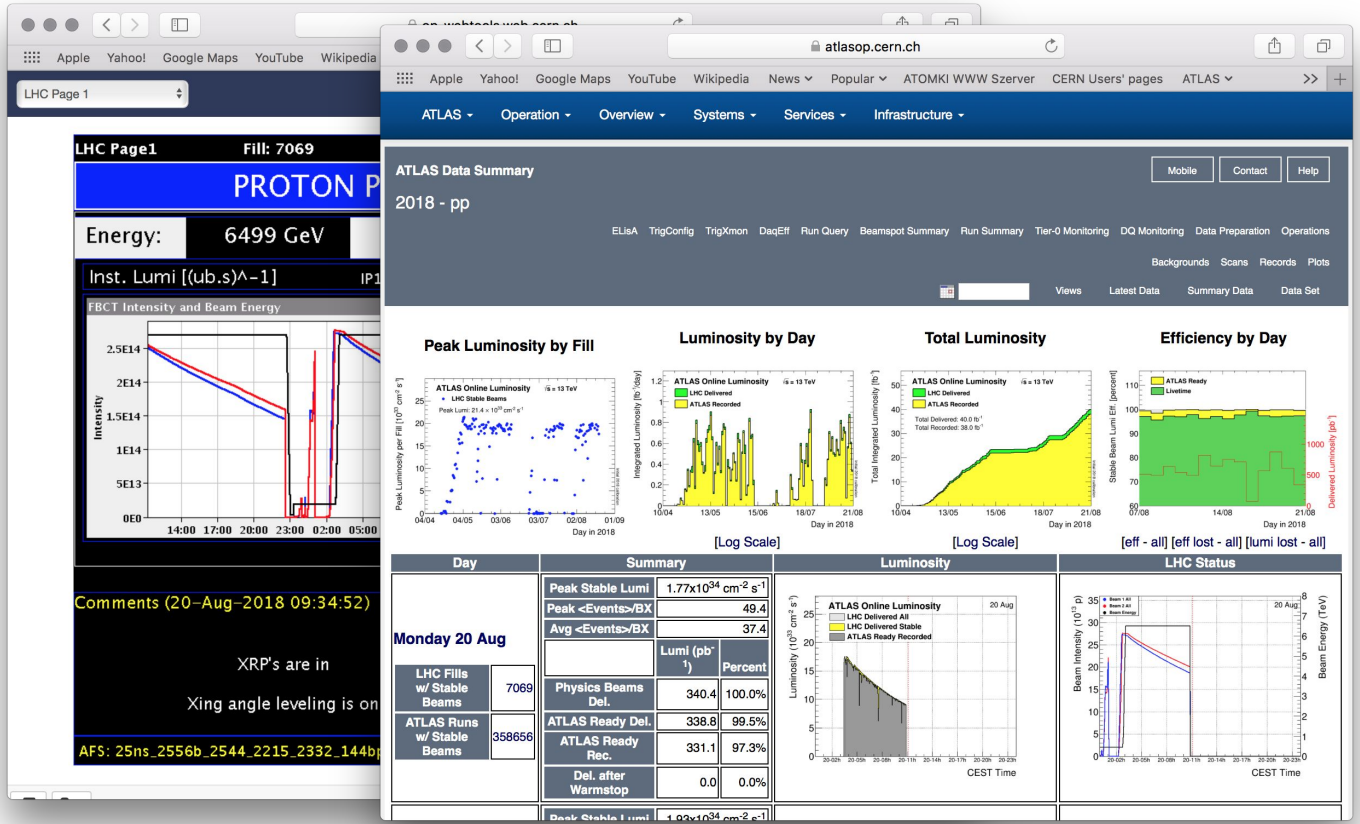


Modern adatfeldolgozás

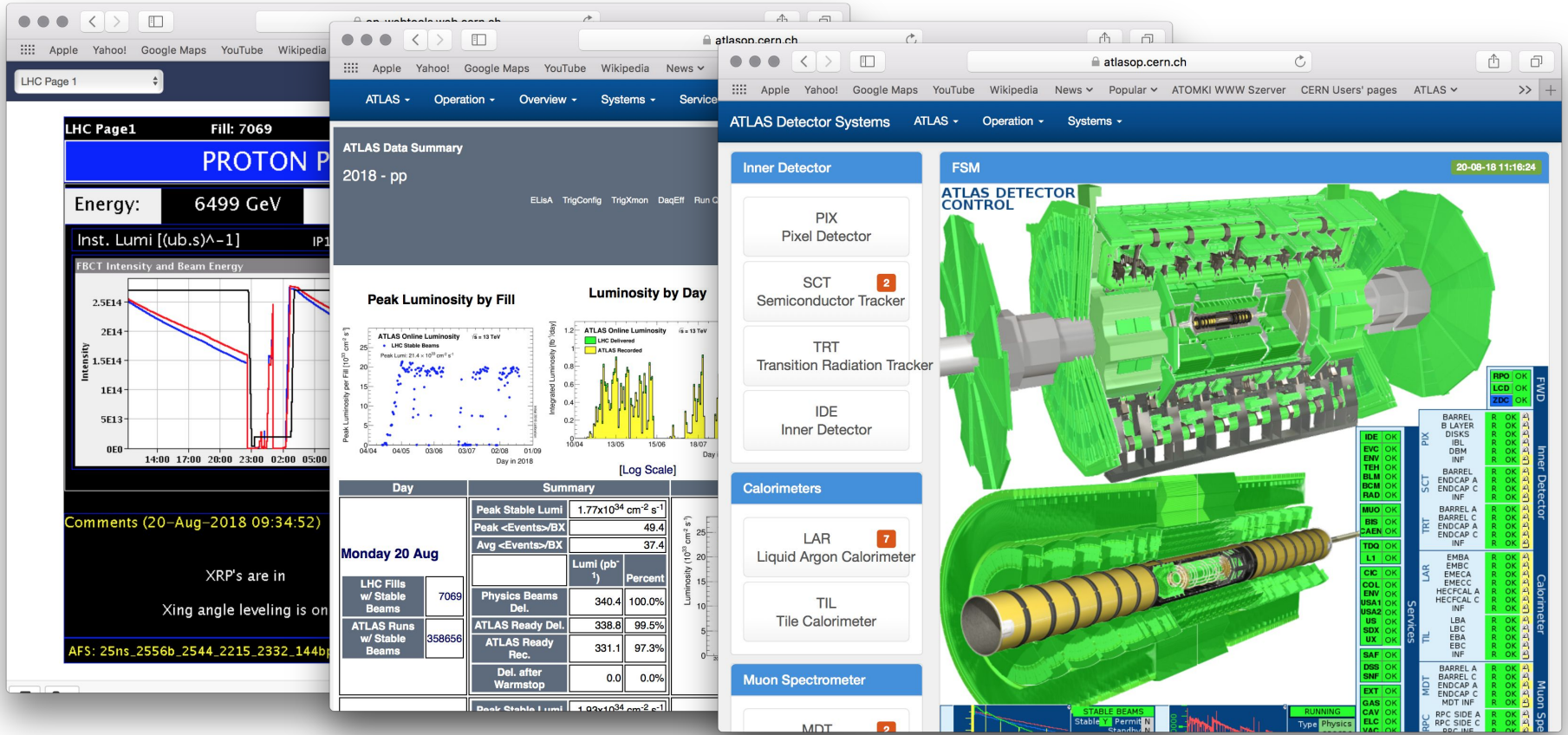
Az ATLAS adatgyűjtése



Az ATLAS adatgyűjtése

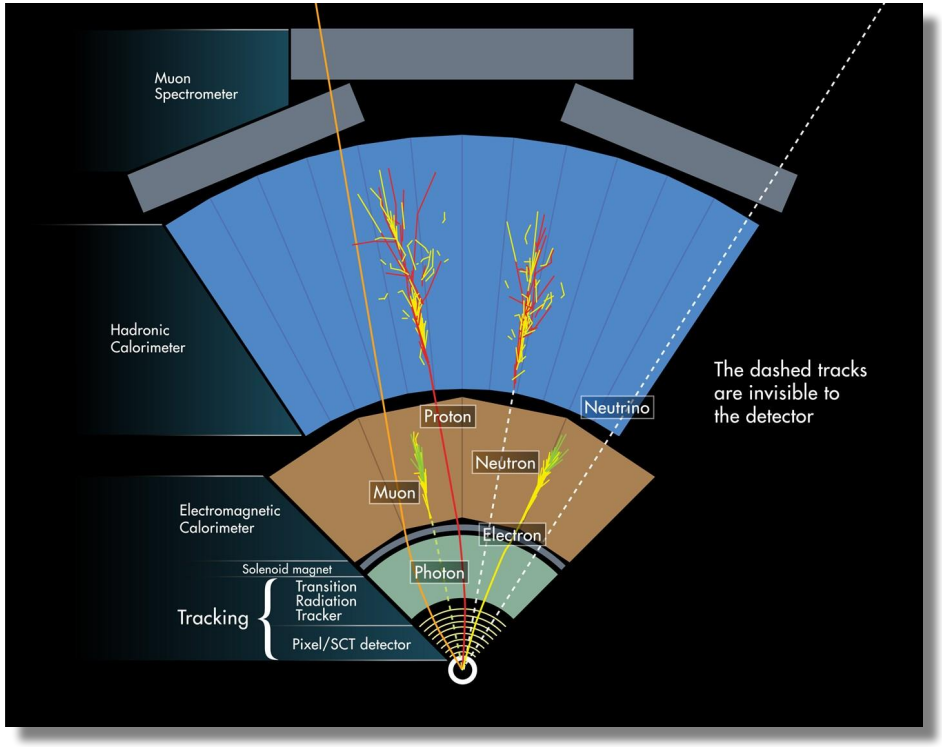


Az ATLAS adatgyűjtése



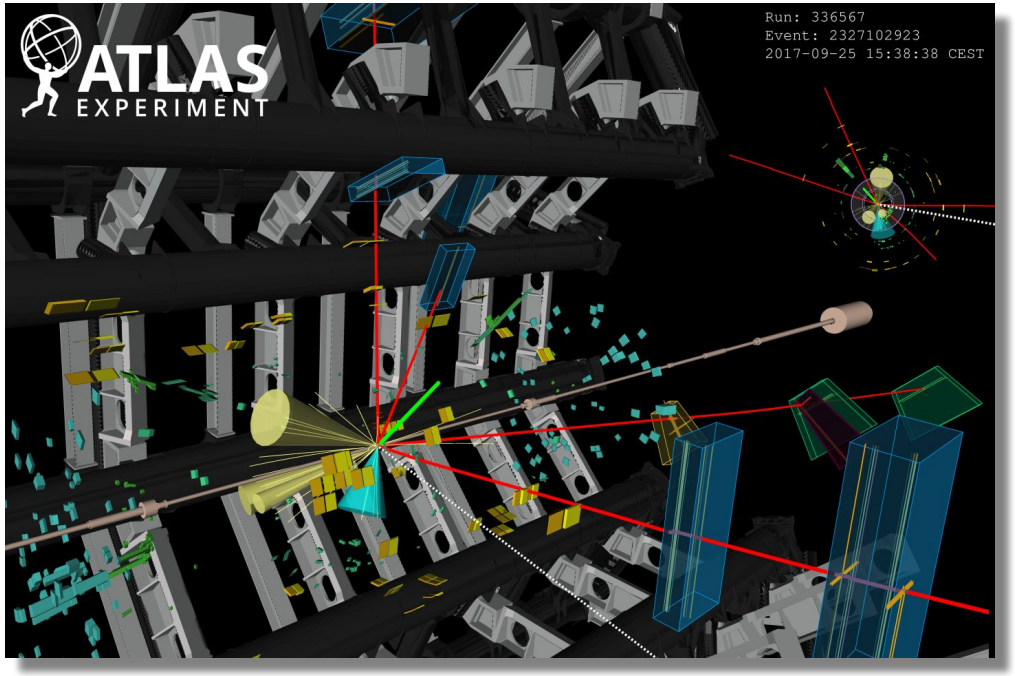
Rekonstrukció

- A detektorokat úgy építjük meg, hogy a nekünk érdekes részecskék egymástól különböző jeleket hagyjanak bennük
- Pl. egy elektron rekonstrukciójához:
  - Keresnünk kell egy töltött részecske nyomát a nyomkövető detektorban
  - Ehhez hozzá kell tudnunk rendelni egy energia-klasztert az elektromágneses kaloriméterben
    - A klaszter és töltött nyom tulajdonságaira sok technikai feltételt is szabunk
  - Nem szabad “aktivitást” találnunk a hadron kaloriméterben az elektron “mögött”



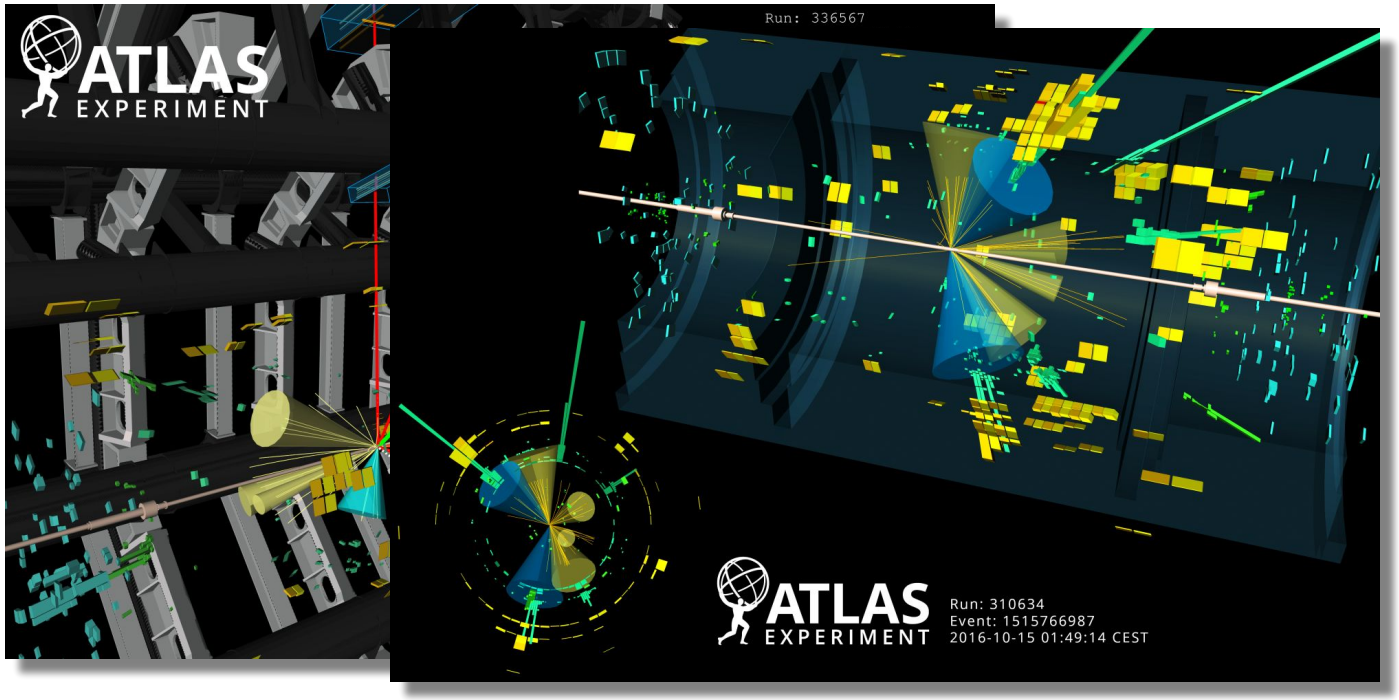
Trigger

- Ez történik legelőször, de egyszerűbb a rekonstrukció után megemlíteni
- Az érdekes eseményeket azok gyors rekonstrukciójával válogatjuk ki



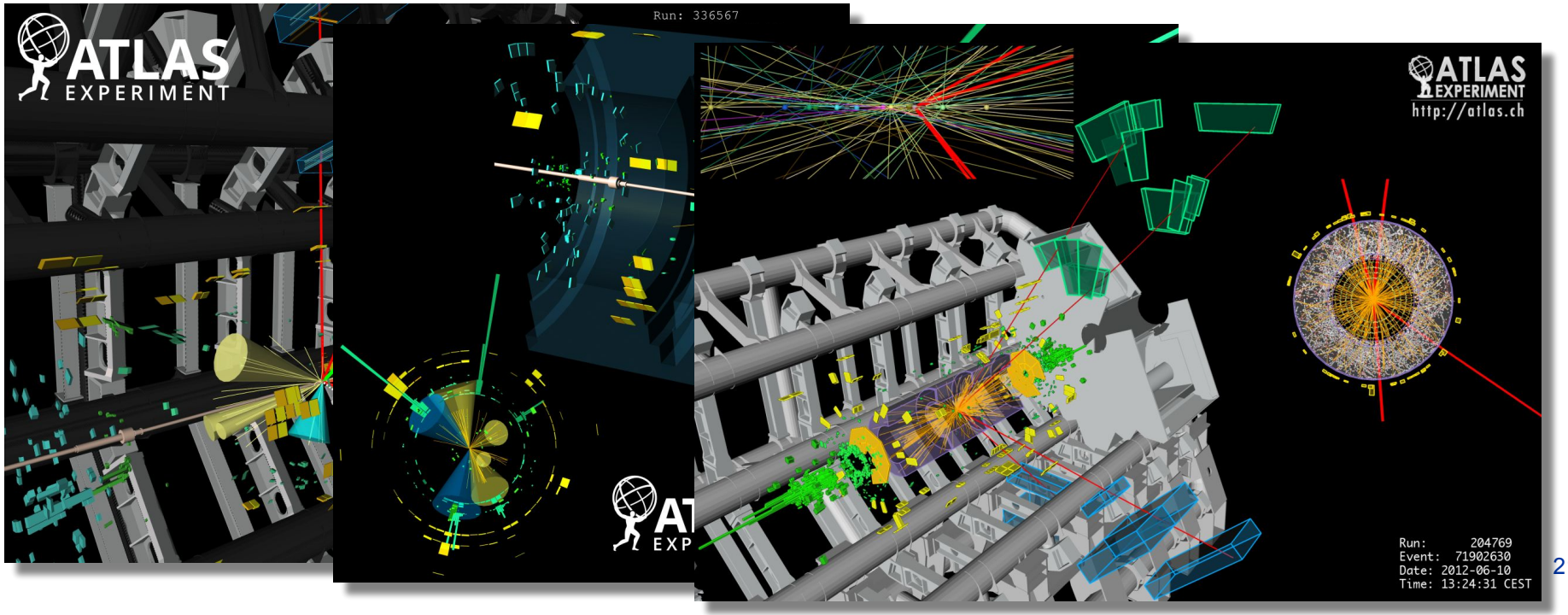
Trigger

- Ez történik legelőször, de egyszerűbb a rekonstrukció után megemlíteni
- Az érdekes eseményeket azok gyors rekonstrukciójával válogatjuk ki



Trigger

- Ez történik legelőször, de egyszerűbb a rekonstrukció után megemlíteni
- Az érdekes eseményeket azok gyors rekonstrukciójával válogatjuk ki

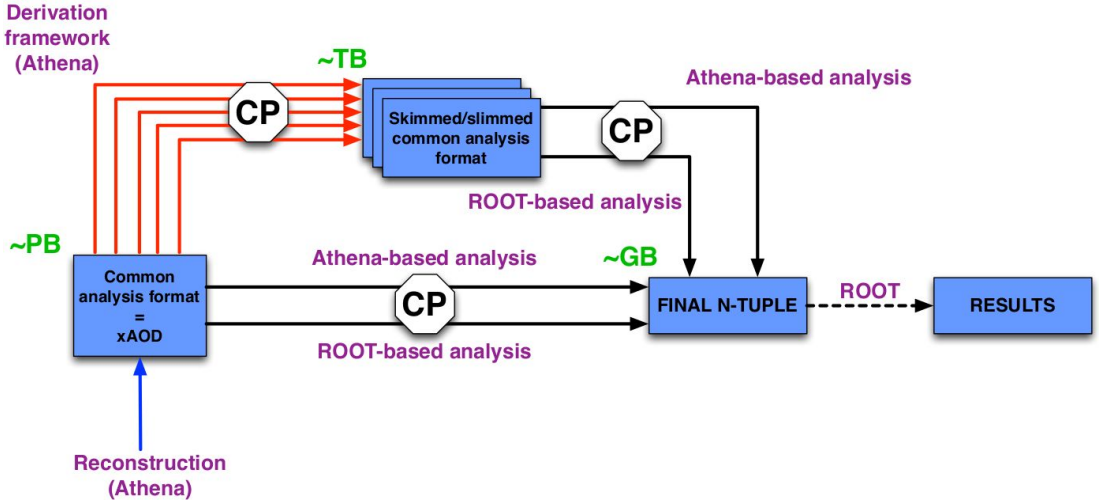


Szimuláció

- A rögzített adatokat a legtöbb esetben csak szimulációk segítségével tudjuk értelmezni
  - Sokszor a legegyszerűbb fizikai folyamatok is annyira bonyolult módon hatnak kölcsön a detektorokkal, hogy csak így tudjuk megbecsülni, mit is mértünk pontosan
- Sok lépésből áll. Az érdekesek ezekből:
  - **Esemény generálás:** Elméleti fizikus kollégák írnak rá szoftvert, hogy véletlenszerű statisztikával valamilyen fizikai folyamathoz “eseményeket” gyártsunk.
  - **Detektor szimuláció:** A generált események “stabil” részecskéinek a kölcsönhatását modellezzük a teljes detektorral. Szimulálva a detektor minden hibáját, és az LHC beállításait is.
    - Itt vesszük figyelembe, hogy mostanra 40-60 proton-proton ütközés is történhet egyszerre egy-egy eseményben
  - **Trigger és rekonstrukció:** A szoftverünket ugyanúgy futtatjuk a szimuláció kimenetén mint a valódi, a detektorból gyűjtött adatokon

Analízis modell

- Nem engedhetjük meg, hogy minden fizikus O(100) PB adatot dolgozzon fel minden alkalommal amikor változtat valamit az analízisén
  - Ehhez a Föld összes számítási kapacitása sem lenne elég
- Az analízis “elejét” szervezett formában végezzük a legtöbb kísérletben
  - Az eredetinél sokkal kisebb adatmennyiséget adva az egyes analizátoroknak, amiket már néhány nap alatt feldolgozhatnak a grid-en

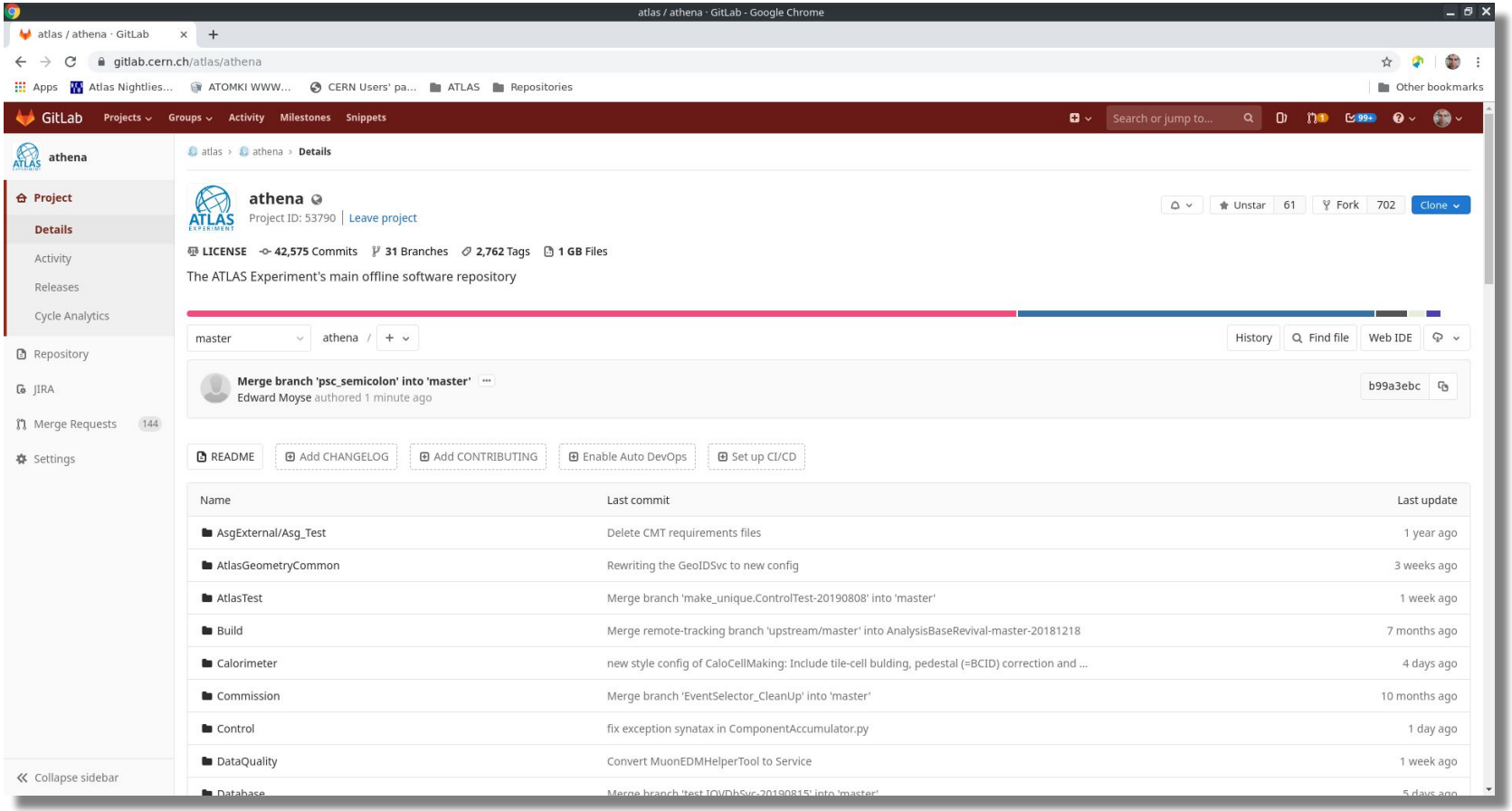


Mit csinál a fizikus?

- Kiválasztja, hogy milyen adatszettet fog feldolgozni
  - Egyúttal azt is, hogy milyen triggert fog használni, ami hatékonyan őrizte meg a számára érdekes fizikai folyamat eseményeit
- Kiválasztja, hogy milyen szimulációkat kell figyelembe vennie az analízisében
  - Ez magában foglalja az analízis által keresett fizikai folyamat szimulációját, és minden más “háttér” folyamatot amire érzékeny lehet az adatfeldolgozás
- Lefuttatja az analízis szoftverét ugyanolyan beállításokkal minden adatra és szimulációra, (relatíve) apró adatfájlokat létrehozva
  - Sokszor ez több lépésben történik, de végül különböző mennyiségek eloszlásait határozzák meg az analízisek a legtöbb esetben
- “Statisztikai analízist” végez a kapott eloszlásokon
  - Megmérendő valamilyen paramétert, vagy felfedezési/kizárási valószínűséget meghatározva
- Elvégzi az egészet még N alkalommal, az analízis minden szisztematikus bizonytalanságát figyelembe véve

Modern számítástechnika

Az ATLAS offline szoftvere



Az ATLAS “offline” szoftvere

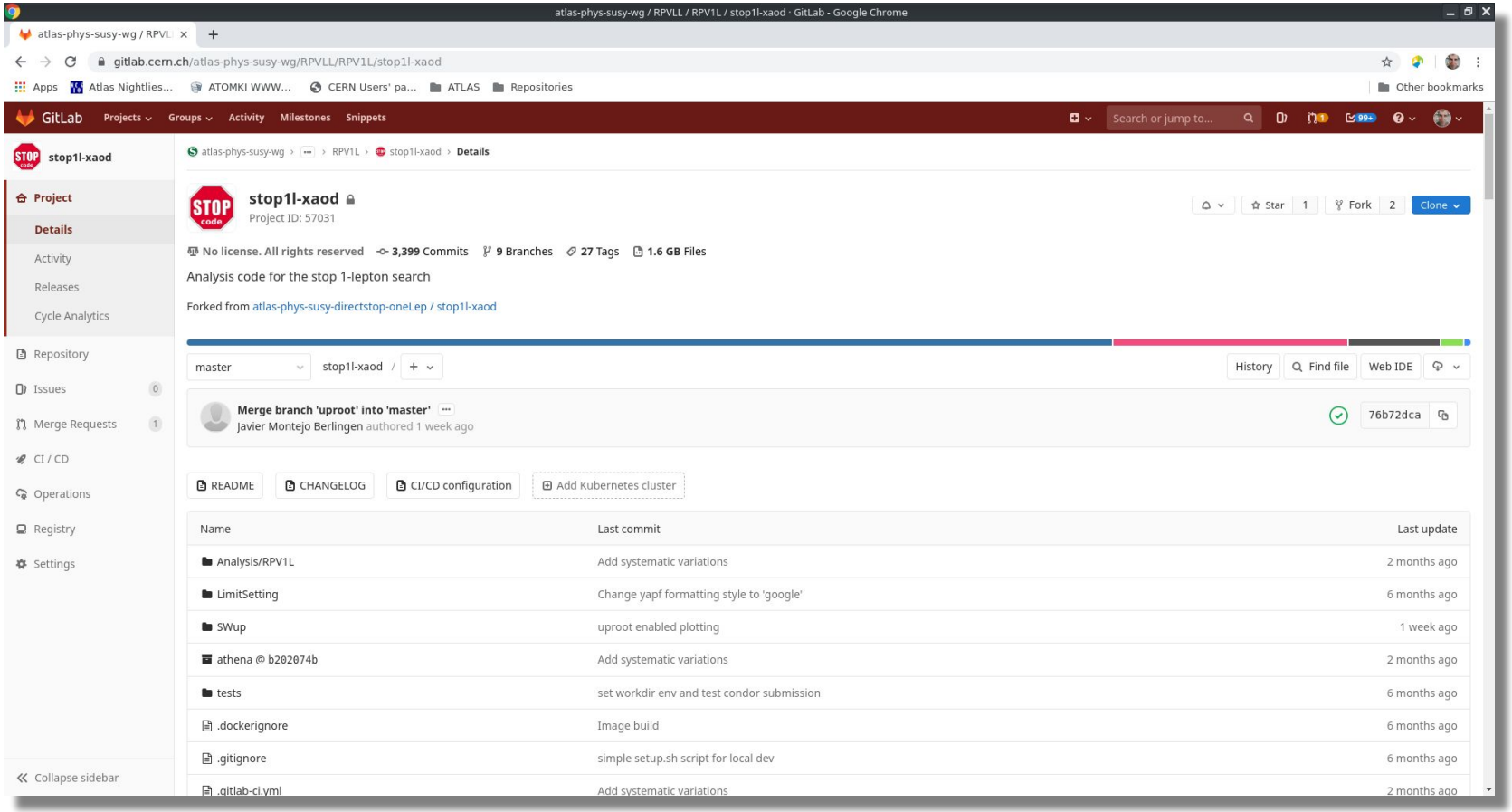
```
Totals grouped by language (dominant language first):
cpp:         3919711 (69.71%)
python:      1178662 (20.96%)
xml:          341767 (6.08%)
sh:            64815 (1.15%)
fortran:       58333 (1.04%)
ansic:         29040 (0.52%)
f90:           12340 (0.22%)
javascript:     6512 (0.12%)
php:            5487 (0.10%)
perl:           3571 (0.06%)
csh:             759 (0.01%)
pascal:          674 (0.01%)
java:            493 (0.01%)
awk:             343 (0.01%)
vhdl:            103 (0.00%)



Total Physical Source Lines of Code (SLOC)                = 5,622,610
Development Effort Estimate, Person-Years (Person-Months) = 1,731.67 (20,780.03)
 (Basic COCOMO model, Person-Months = 2.4 * (KSLOC**1.05))
Schedule Estimate, Years (Months)                         = 9.11 (109.31)
 (Basic COCOMO model, Months = 2.5 * (person-months**0.38))
Estimated Average Number of Developers (Effort/Schedule)  = 190.11
Total Estimated Cost to Develop                           = $ 233,924,916
 (average salary = $56,286/year, overhead = 2.40).
SLOCCount, Copyright (C) 2001-2004 David A. Wheeler
SLOCCount is Open Source Software/Free Software, licensed under the GNU GPL.
SLOCCount comes with ABSOLUTELY NO WARRANTY, and you are welcome to
redistribute it under certain conditions as specified by the GNU GPL license;
see the documentation for details.
Please credit this data as "generated using David A. Wheeler's 'SLOCCount'."
```

- Kb. 4 millió sor C++ és 1 millió sor Python
  - Óvatos becsléssel is >1000 emberi munkaév van benne
- Folyamatosan fejlesztjük, nekem is ez a fő feladatom az ATLAS-ban
- A teljes szoftvert több mint 5 óra megépíteni egy 16 processzoros gépen
  - Így a szoftverfejlesztést is egyedi módon kell végeznünk…
- Egy világméretű elosztott fájlrendszeren tároljuk
  - `/cvmfs/atlas.cern.ch`

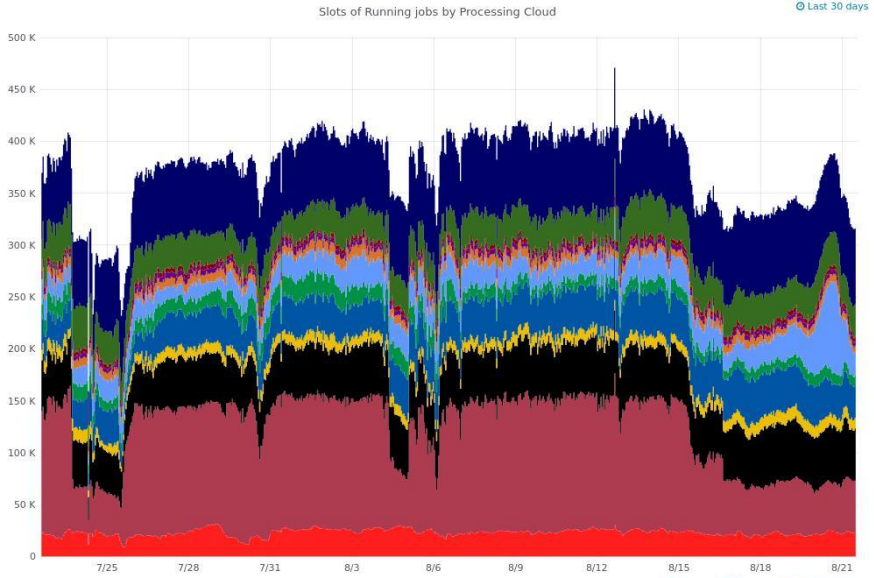
Analízis program-kód



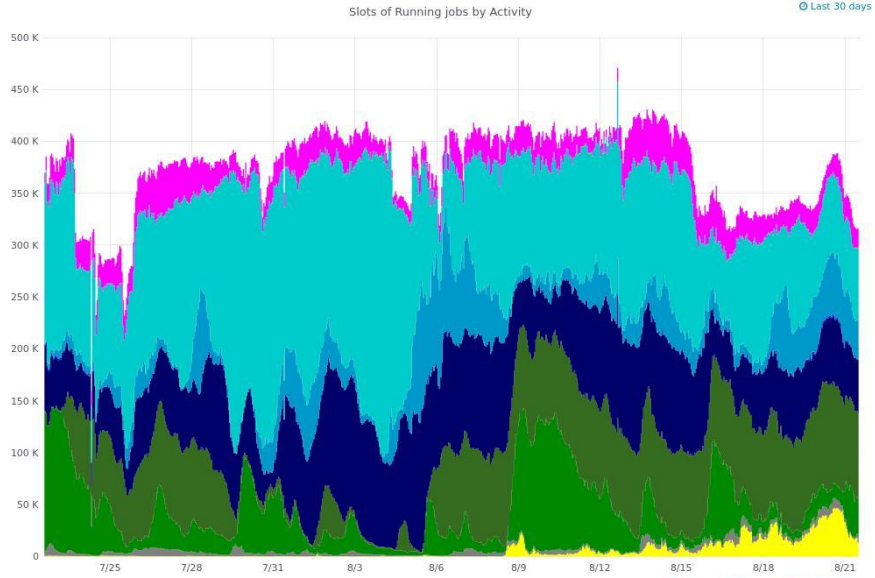
Gépi tanulás

- Nagyon kevés fizika analízis létezik manapság nélküle…
- Részecskefizikában 2 dologra használjuk a legtöbb esetben
  - **Klasszifikáció:** Egy-egy esemény tulajdonságai alapján azt döntjük el, hogy az olyan-e mint amit keresünk
    - A rekonstruált részecskék fajtáját is sokszor így határozzuk meg
  - **Regresszió:** Részecskék / események tulajdonságait számoljuk ki “alacsonyabb rendű” tulajdonságai alapján. Pl. meghatározzuk egy hadronzápor pontos energiáját a kaloriméterekben érzékelt jelekből.
- Ilyen technikákat már nagyon régóta használunk a fizikában.
  - A múltban ez főleg döntési fákat és neurális hálókat jelentett
- Az utóbbi 5-10 évben nagy fejlődésen ment keresztül
  - Mind elméleti, mind gyakorlati szempontból
  - Mostanra “az ipar” által készített kódokat szeretjük a leginkább használni
    - A Google által fényképek azonosítására használt kód nagyon jól használható az LHC eseményeinek megértésére is 😛

Mekkora is az LHC/ATLAS grid?



	min	max	avg	current
CA	9.2 K	31.2 K	23.1 K	19.5 K
CERN	23.4 K	145.9 K	101.4 K	42.4 K
DE	22.5 K	60.5 K	49.5 K	40.6 K
ES	4.4 K	15.2 K	9.8 K	8.2 K
FR	9.8 K	46.6 K	35.8 K	26.0 K
IT	5.7 K	22.1 K	13.2 K	8.9 K
ND	8.5 K	88.1 K	22.3 K	18.1 K
NL	2.4 K	11.7 K	8.0 K	6.1 K
RU	2.2 K	6.3 K	4.6 K	3.7 K
TW	1.6 K	5.1 K	4.1 K	3.5 K
UK	17.6 K	43.4 K	30.7 K	26.7 K
UNKNOWN	0	2	1	1
US	32.5 K	87.9 K	71.9 K	59.0 K



	min	max	avg	current
Data Processing	96	46.9 K	5.9 K	8.9 K
Event Index	2	986	80	5
Group Analysis	428	15.1 K	3.7 K	3.8 K
Group Production	1.8 K	138.8 K	38.0 K	32.9 K
MC Event Generation	512	104.7 K	60.1 K	68.7 K
MC Reconstruction	11.8 K	136.3 K	76.3 K	41.8 K
MC Simulation	3	3.7 K	197	36
MC Simulation Fast	486	124.8 K	29.9 K	27.8 K
MC Simulation Full	45.4 K	289.8 K	135.0 K	61.3 K
Others	29	205	94	94
Testing	156	967	455	252
User Analysis	7.0 K	49.8 K	24.8 K	17.3 K
t0_caf	3	843	19	3

Összefoglalás

- Mostanra a szoftverfejlesztés egy nagyon lényeges elemévé vált a fizikának
  - Gyakorlatilag semmilyen kísérlet nem létezhet saját adatgyűjtő és feldolgozó szoftver nélkül manapság
  - Az adatok feldolgozását speciális programok írásával végezzük
    - Minden LHC fizikusnak muszály legalább “elfogadhatóan” programozni tudnia
- Az LHC több száz PB adatát a világméretű grid sok-százezer processzorán dolgozzuk fel

http://home.cern