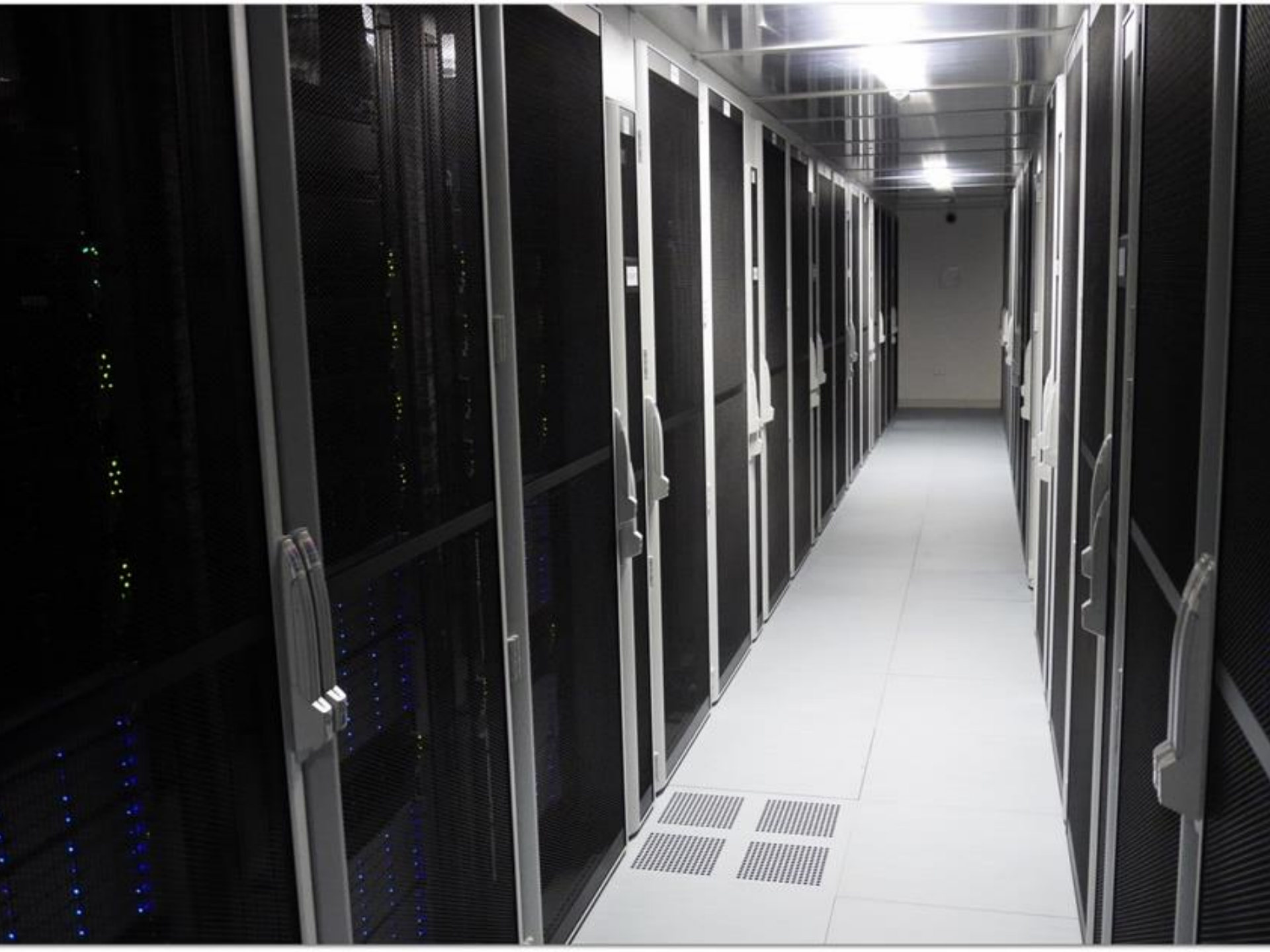


Big Data:
Paradigmaváltás a tudományban?
Vagy annál is több?

Lévai Péter
MTA WIGNER Fizikai Kutatóközpont

2013 szeptember 12., Wigner Adatközpont

A Tudomány fejlődése az ókortól napjainkig - 1

Ókor

Empirikus tudomány

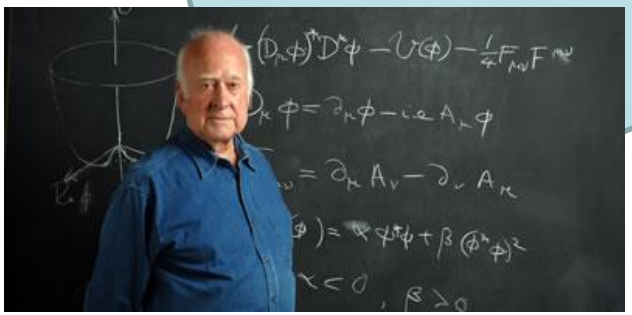
Középkor

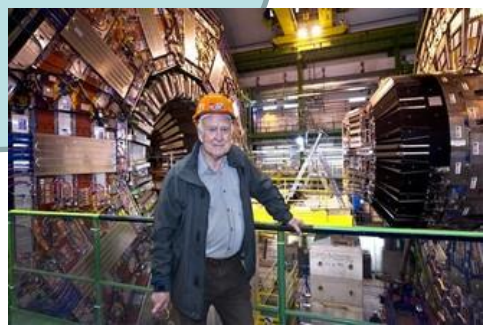
Fenomenológia

Újkor

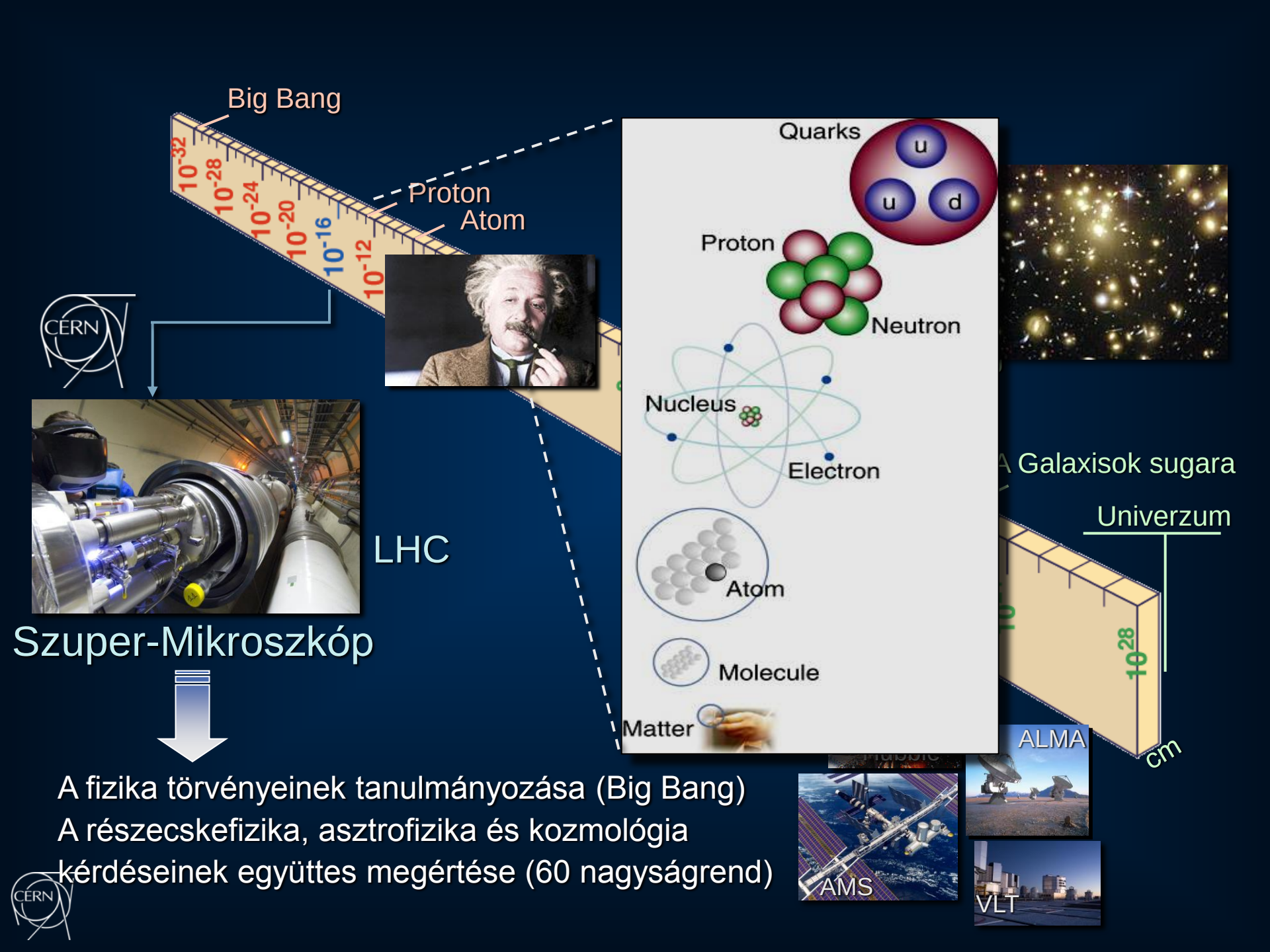
Elmélet

Kísérlet

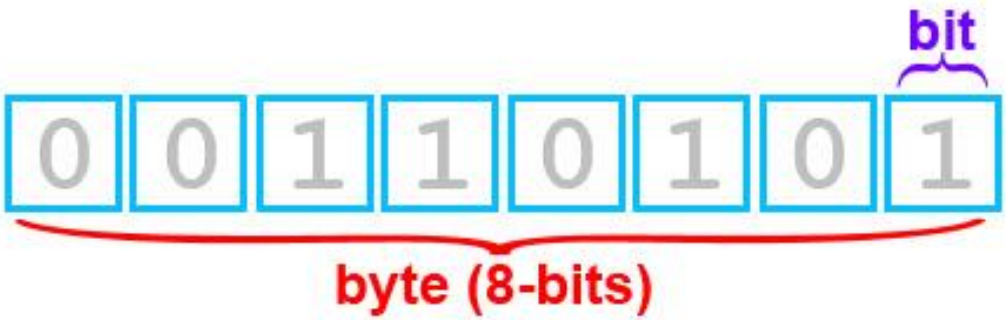




Eredmény → 10^{60} nagyságrenden belül ismerjük Világegyetemünket


Big Bang
Proton
Atom
Quarks
u
u
d
Proton
Neutron
Nucleus
Electron
Atom
Molecule
Matter
Galaxisok sugara
Univerzum
cm
LHC
Szuper-Mikroszkóp
ALMA
AMS
VLT
A fizika törvényeinek tanulmányozása (Big Bang)
A részecskefizika, asztrofizika és kozmológia
kérdéseinek együttes megértése (60 nagyságrend)

Tudomány ↔ Információgyűjtés → 1 byte

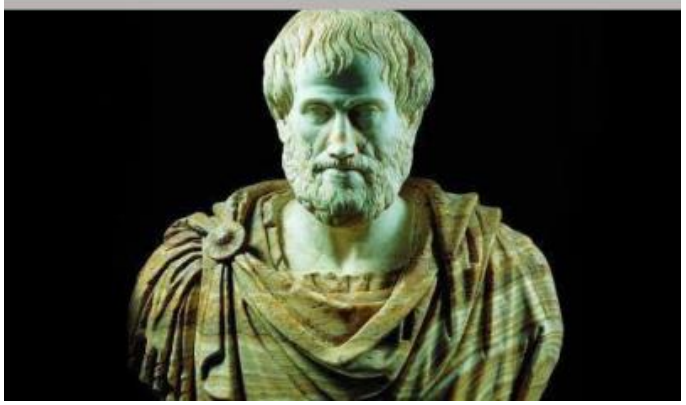


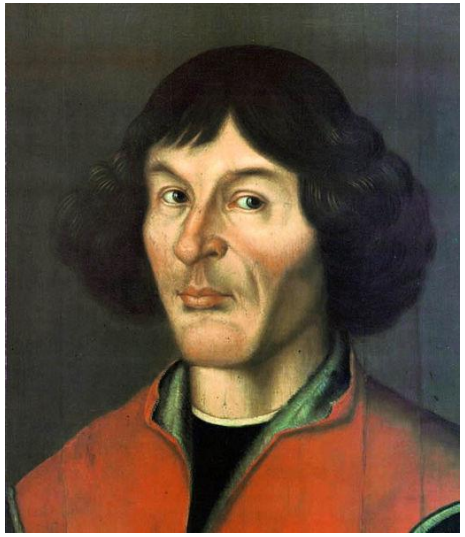
Byte < KiloB < MegaB < GigaB < TeraB < PetaB < ExaB < ZettaB

$1 \quad 10^3 \quad 10^6 \quad 10^9 \quad 10^{12} \quad 10^{15} \quad 10^{18} \quad 10^{21}$

Avogadro-szám: 6×10^{23}

Arisztotelész (Kr. e. 384–322)

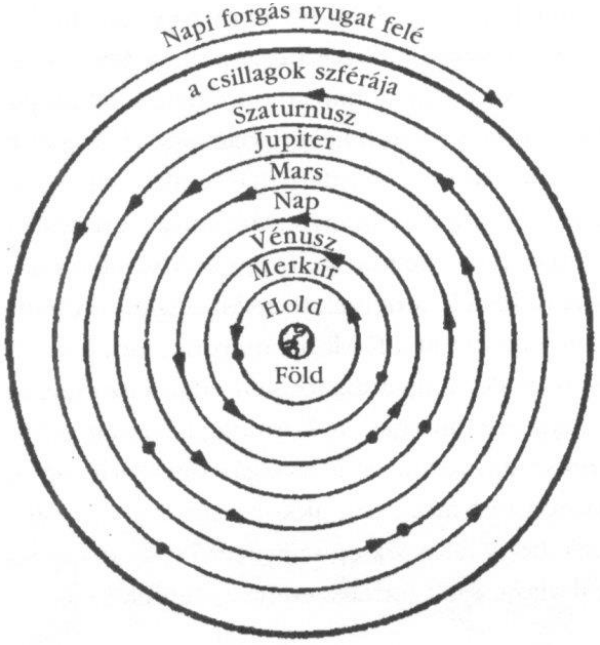


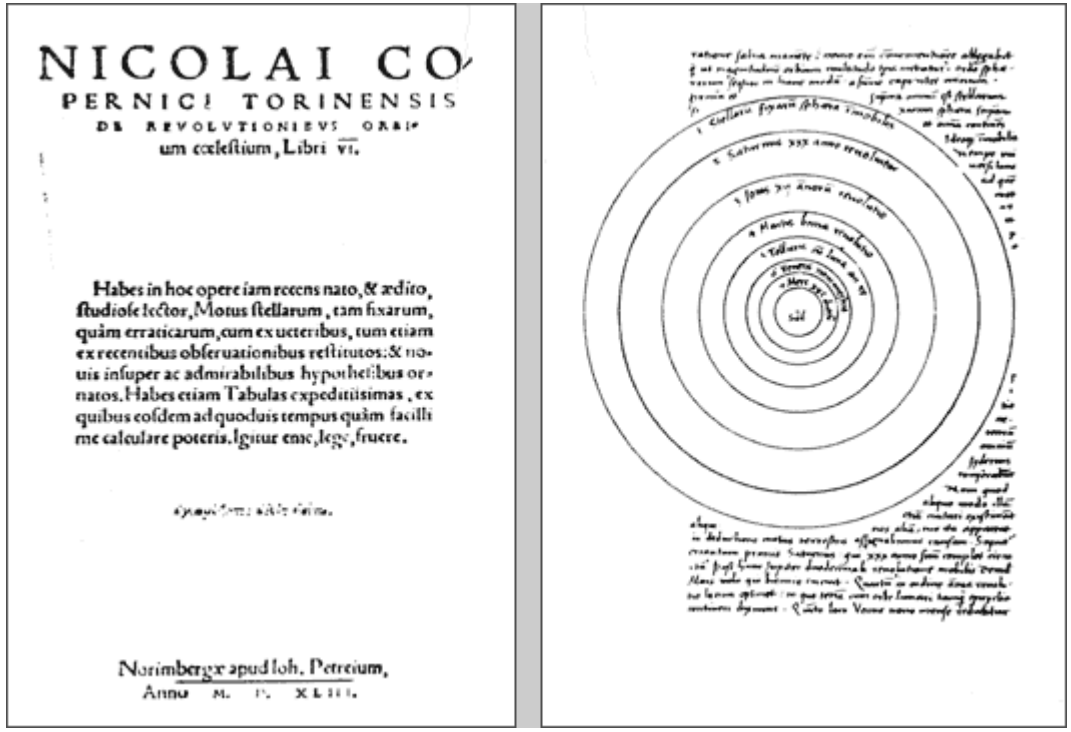


Nikolaus Kopernikusz (1473-1543)

1510: nap-középpontú világkép

1543: De Revolutionibu Orbium Coelestiu





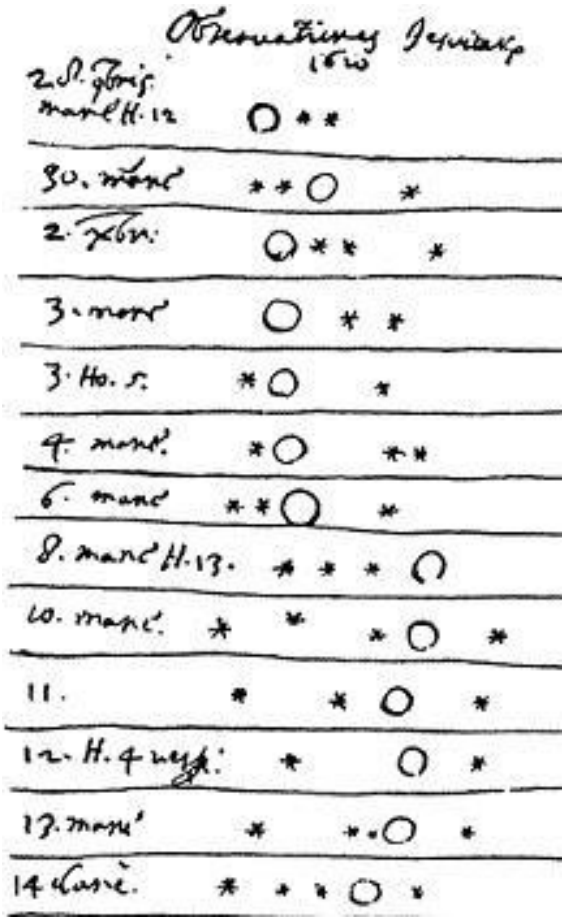
NICOLAI CO
PERNICI TORINENSIS
DE REVOLVTIONIBVS ORBI
um coelestium, Libri VI.

Habes in hoc opere iam recens nato, & ædito, studiose lector, Motus stellarum, tam fixarum, quàm erraticarum, cum ex ueteribus, tum etiam ex recentibus obseruationibus restitutos: & nouis insuper ac admirabilibus hypothesibus ornatos. Habes etiam Tabulas expeditissimas, ex quibus eosdem ad quoduis tempus quàm facilli me calculare poteris. Igitur emc, lege, fruere.

Norimbergæ apud Ioh. Petreium,
Anno M. D. XLIII.

Galilei (1564-1642)
Fiatalkori kép

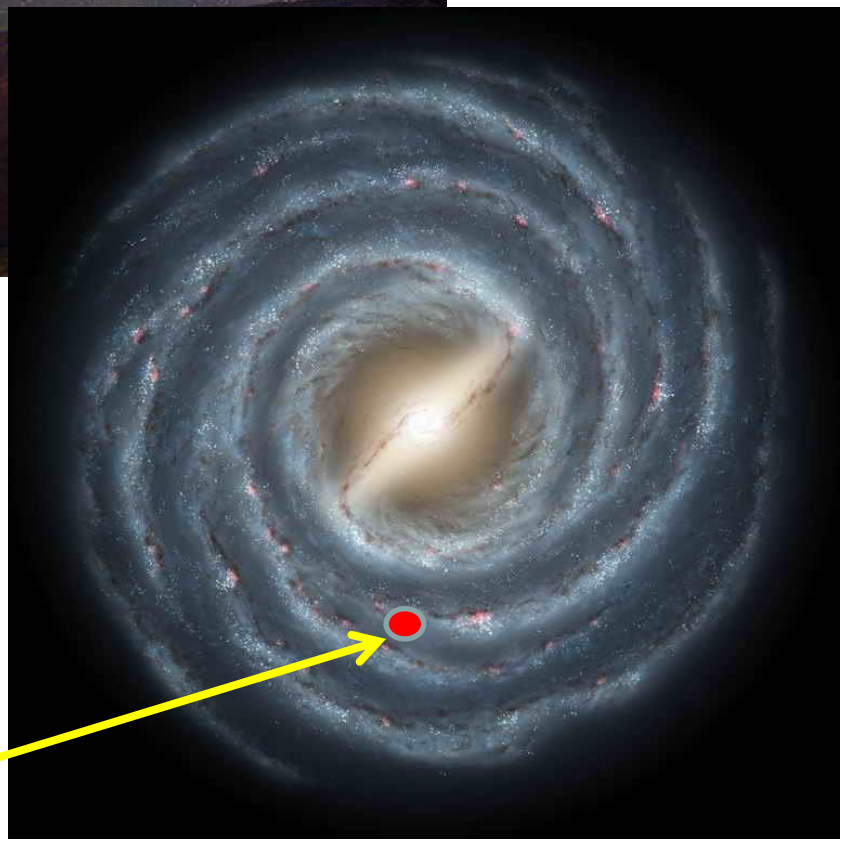
Galilei távcső bemutatója
1609. augusztus 25.
2009: 400 éves a
megfigyelő csillagászat



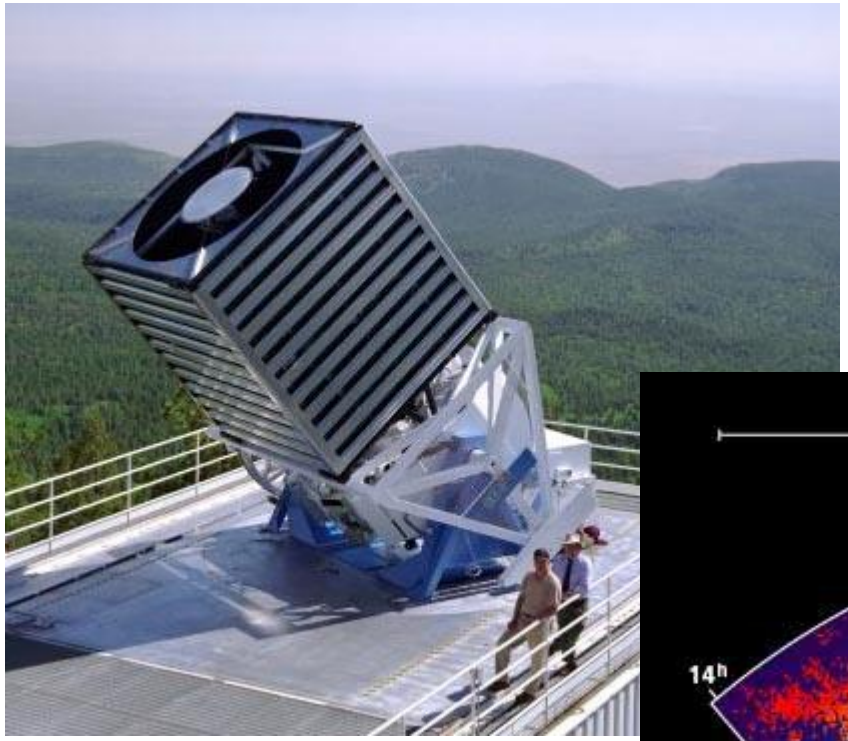
1610. január 7 :
a Jupiter 4 holdjának
felfedezése (Europe, Io,
Kallisztó, Ganümédesz)

1 év alatt 1 kB adat: 3 Byte/éjszaka
→ a Jupiter egy kis Naprendszer,
Kopernikusznak van igaza !

A Tejútrendszer,
a mi galaxisunk:
a Földről nézve
és „messziről” tekintve

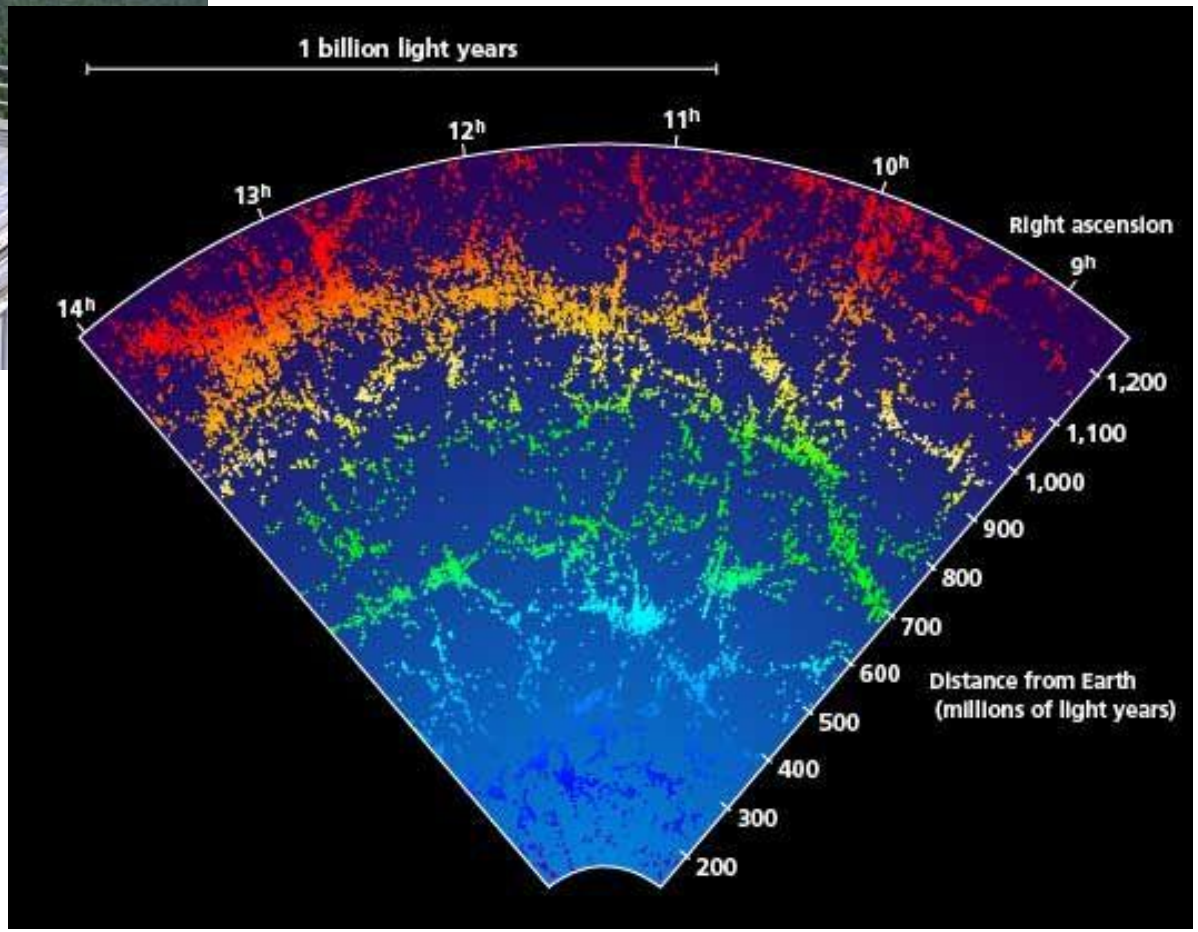


Naprendszer



Sloan Digital Sky Survay
New Mexico, Apache Point, 2000/2005/2014
120 Mpixel CCD kamera

Szalai Sándor John Hopkins Egyetem
Csabai István és csoportja, ELTE
Szapudi István és csoportja, Hawai

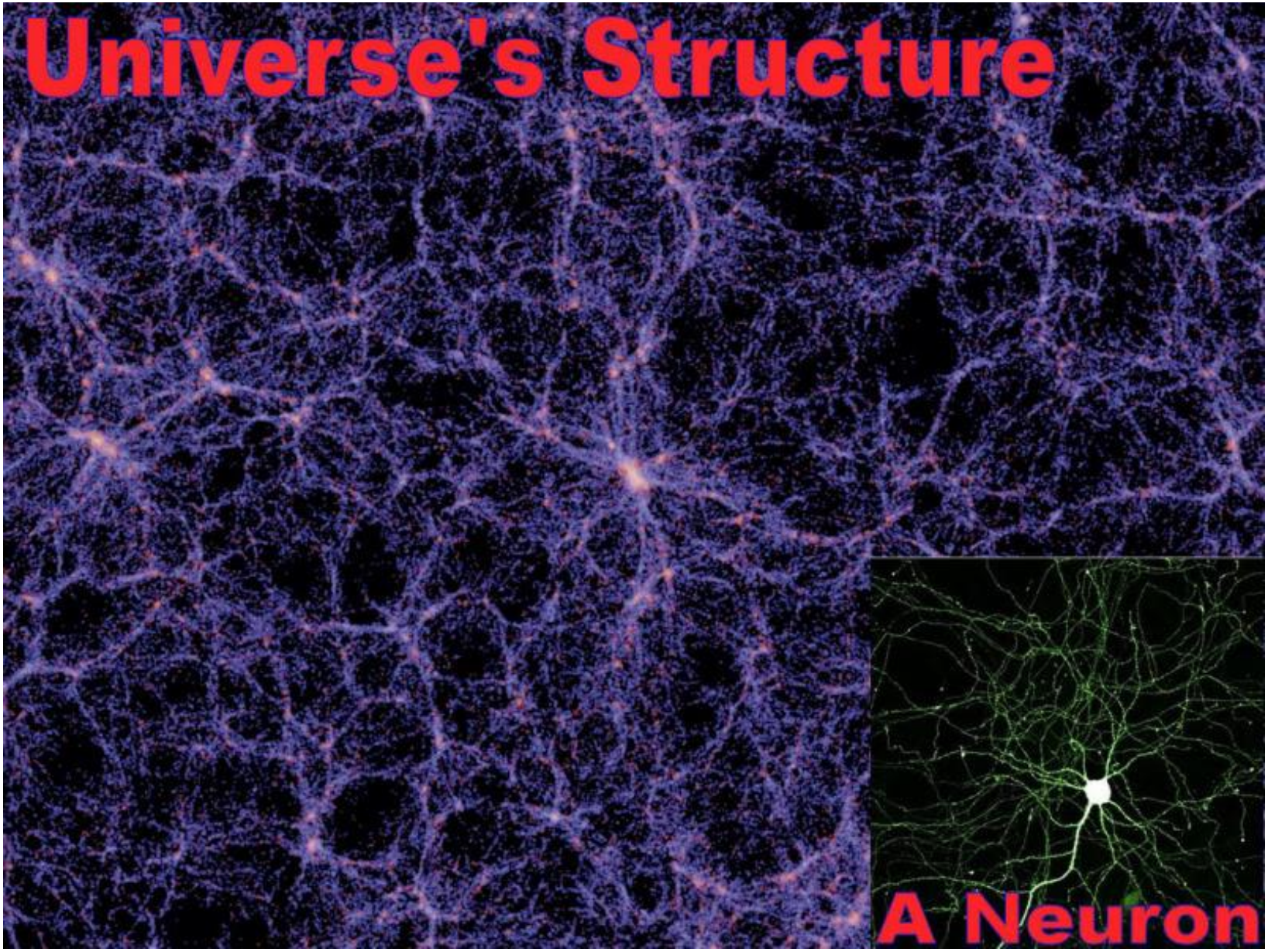


30 db 2048x2048 CCD chip
5 színszűrő:
354, 476, 628, 769, 925 nm

200 GB adat éjszakánként →
2000-2014: 1100 TB adat

Petabájt skála

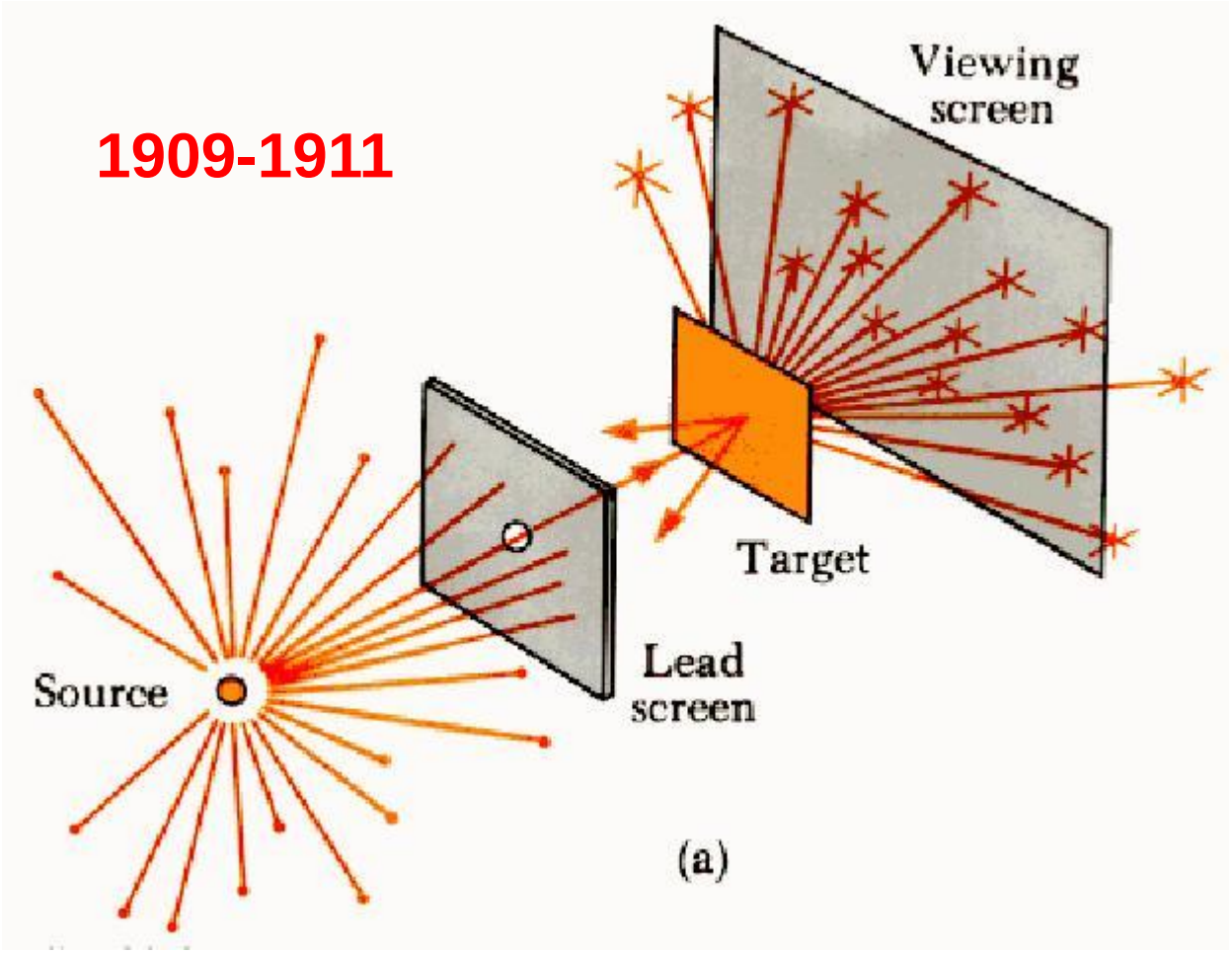
Az Univerzum struktúrája --- A neuronok hálózata



Az Internet térképe

Ernest Rutherford
(1871-1937)

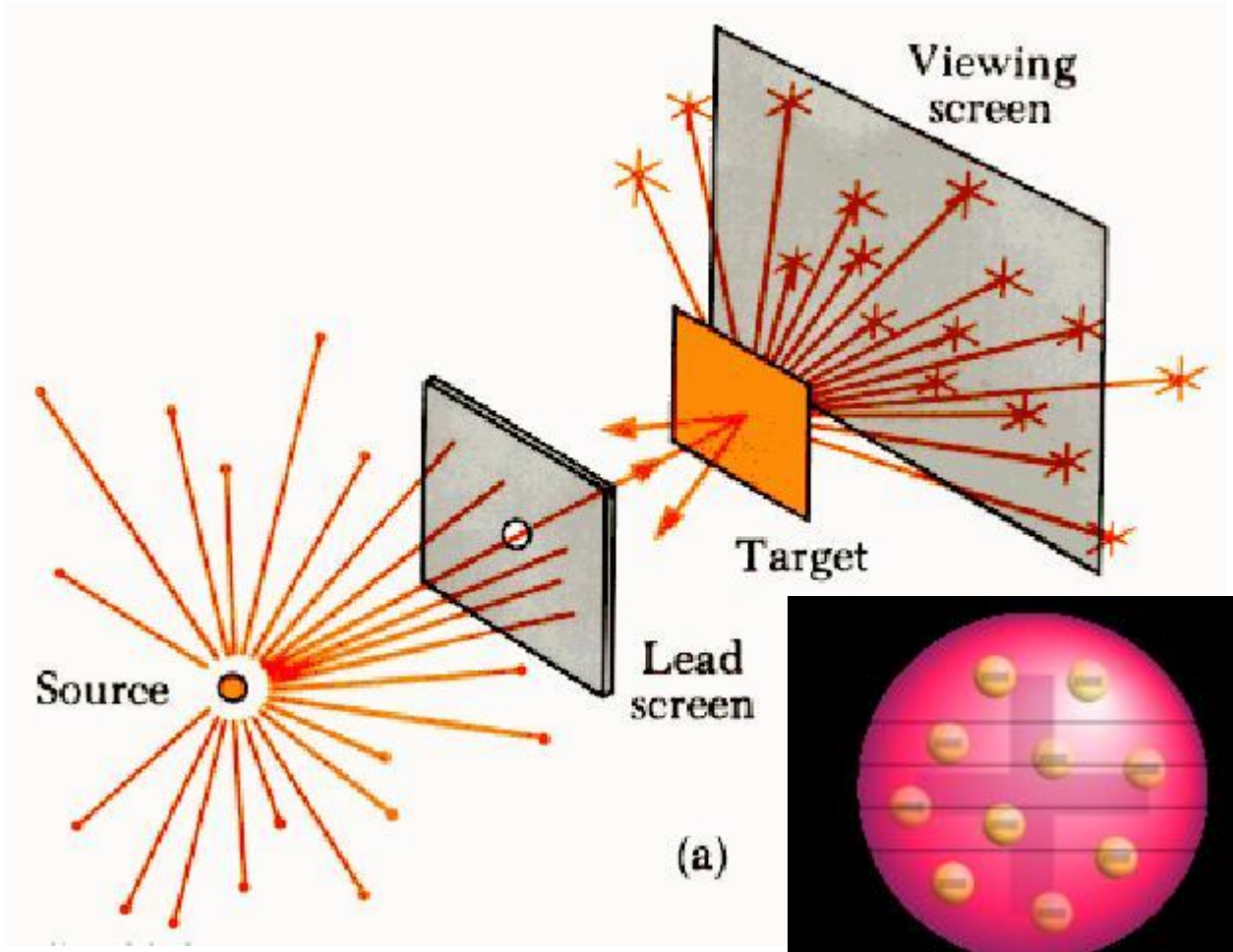
1909-1911

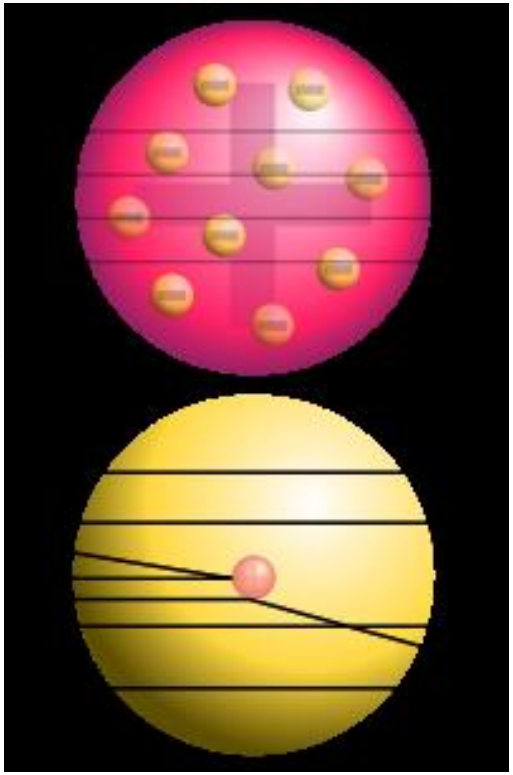


Magfizika (korai részecskefizika)

A Rutherford kísérlet:
az α-részecskék visszaszóródnak az arany fóliáról

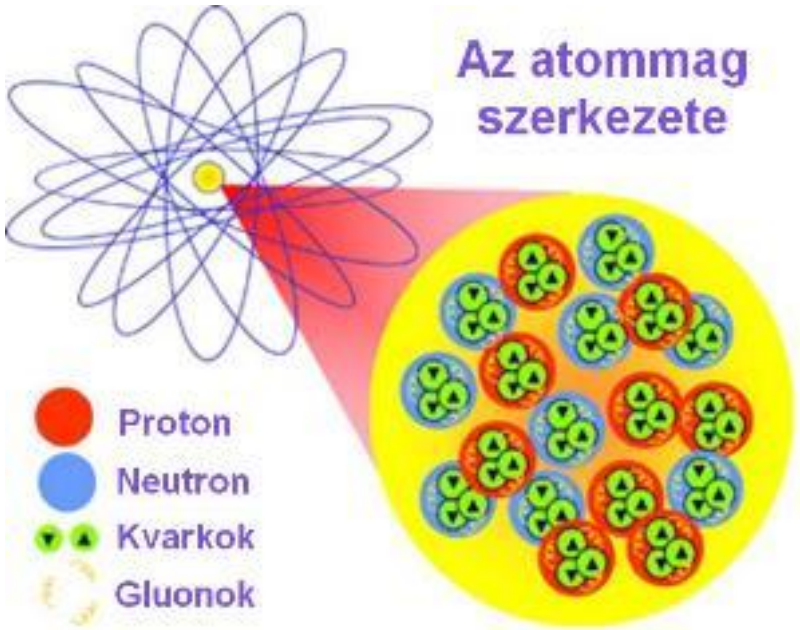
Ernest Rutherford
(1871-1937)



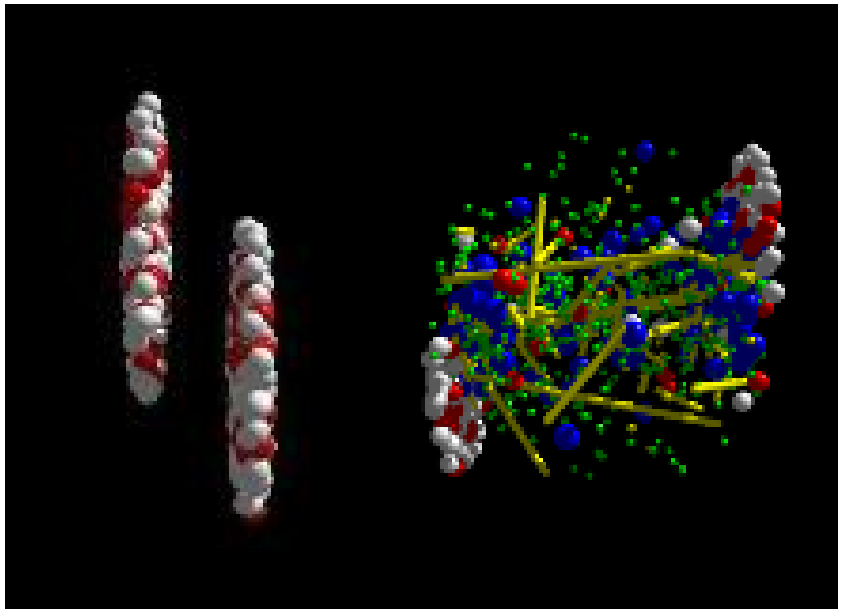


A Rutherford kísérlet:
az α-részecskék visszaszóródnak az arany fóliáról
→ **az arany atomoknak pozitív töltésű magjuk van**
amit negatív töltésű elektron felhő vesz körbe

1 nap 1 KB adat



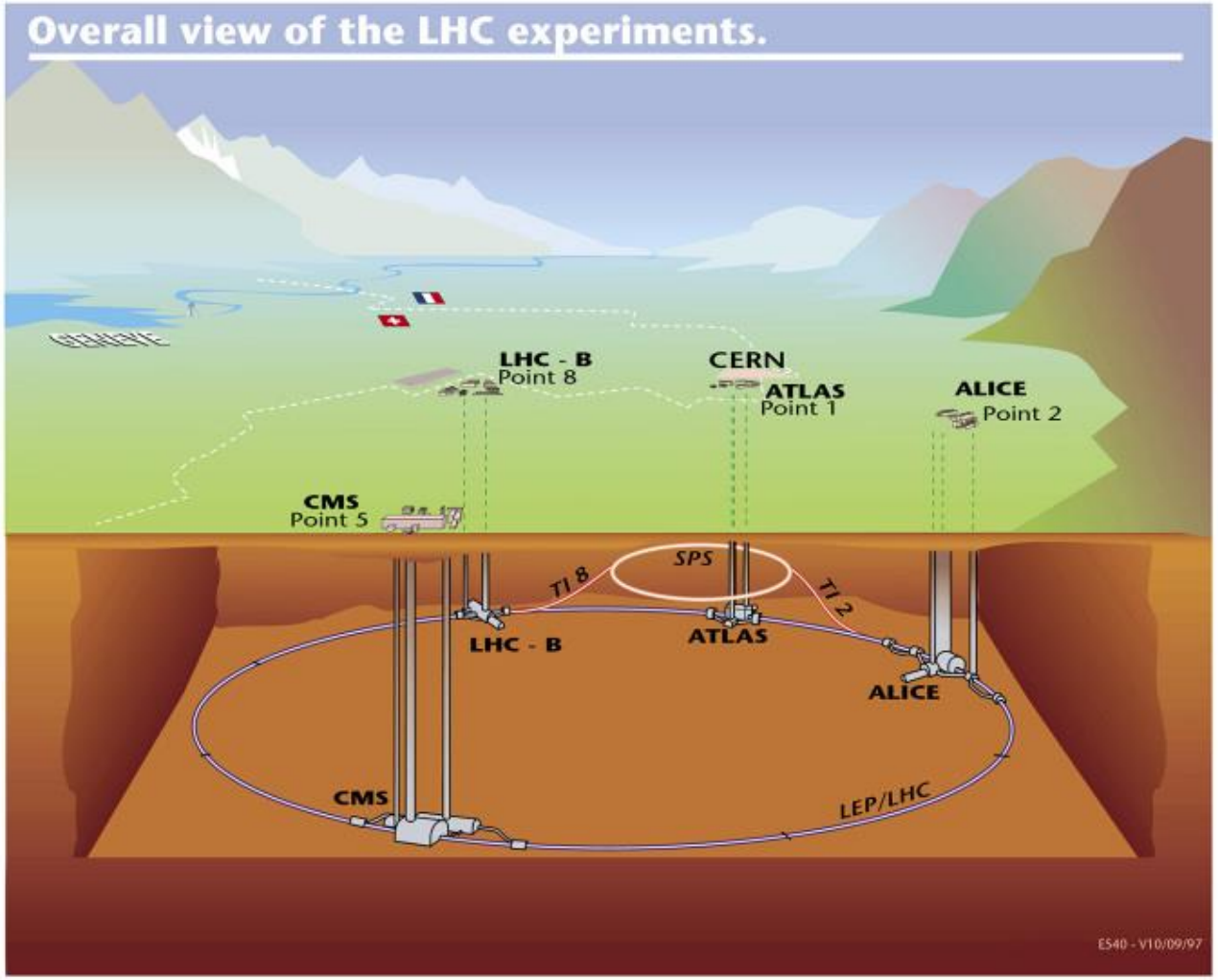
Az atommagok ütközése



Gyorsítók szerepe

CERN LHC: a Föld legnagyobb berendezése

Magyarország 1992 óta teljes jogú tagja a CERN-nek
~1 %-ban vagyunk „tulajdonosok”

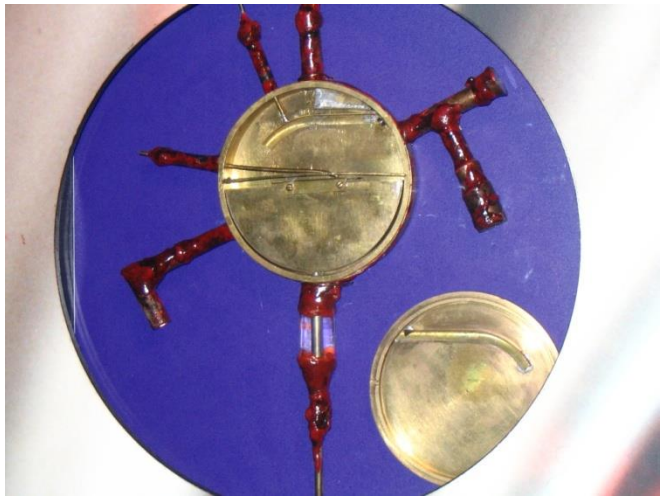


L3
OPAL
NA49
ASACUSA

ALICE
CMS

ATLAS
TOTEM

Gyorsítók: 1930 ⇒⇒⇒ 2010 CERN LHC

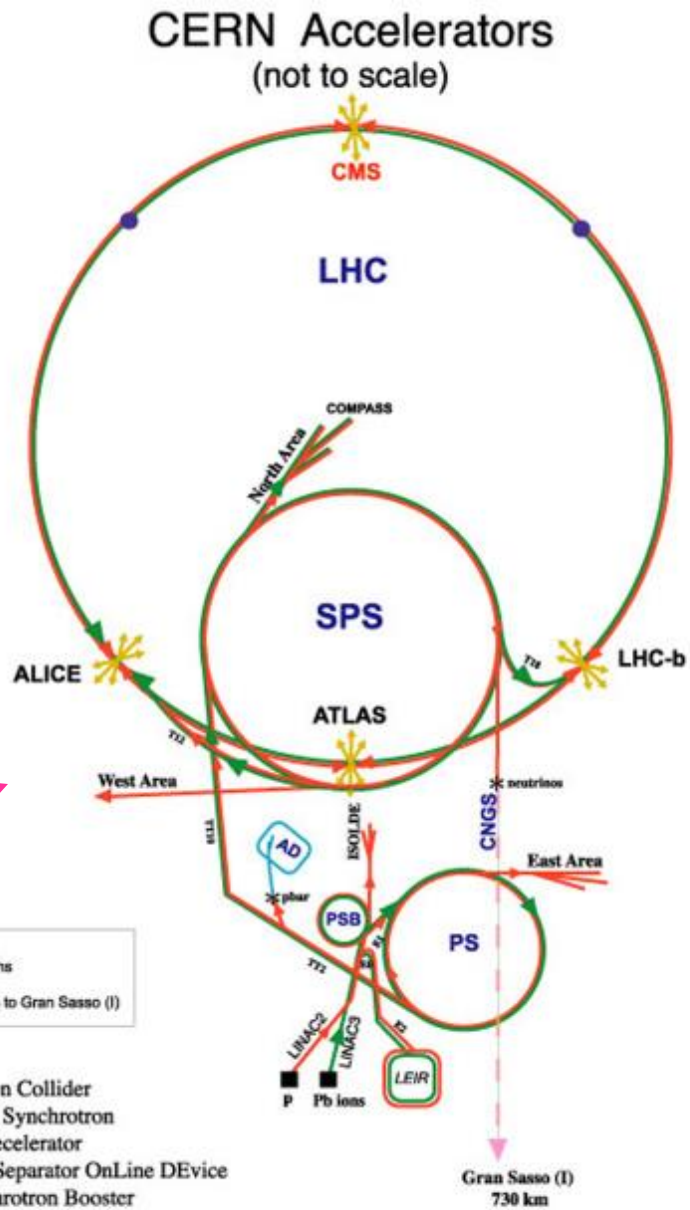


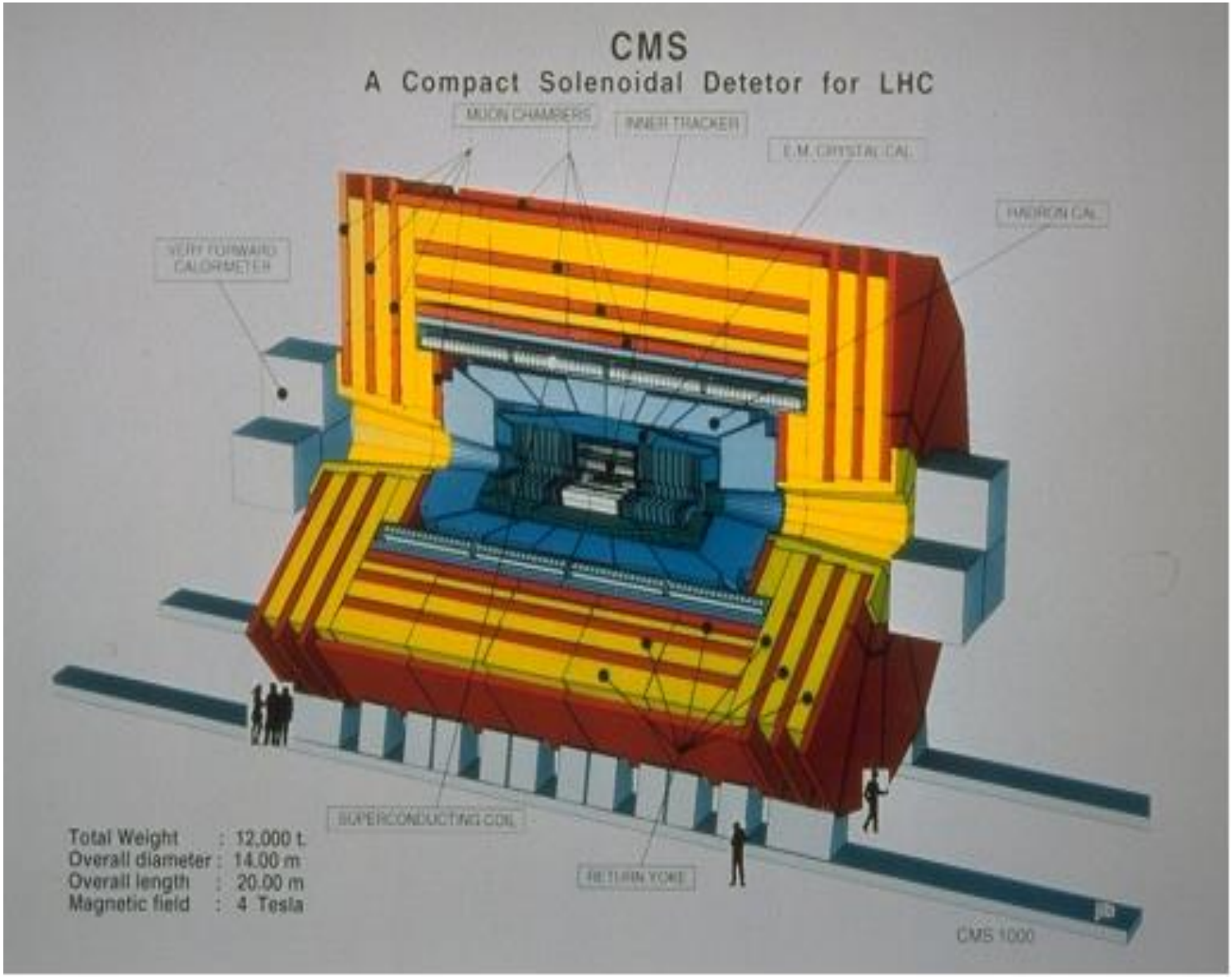
Az első ciklotron, 1930, Lawrence

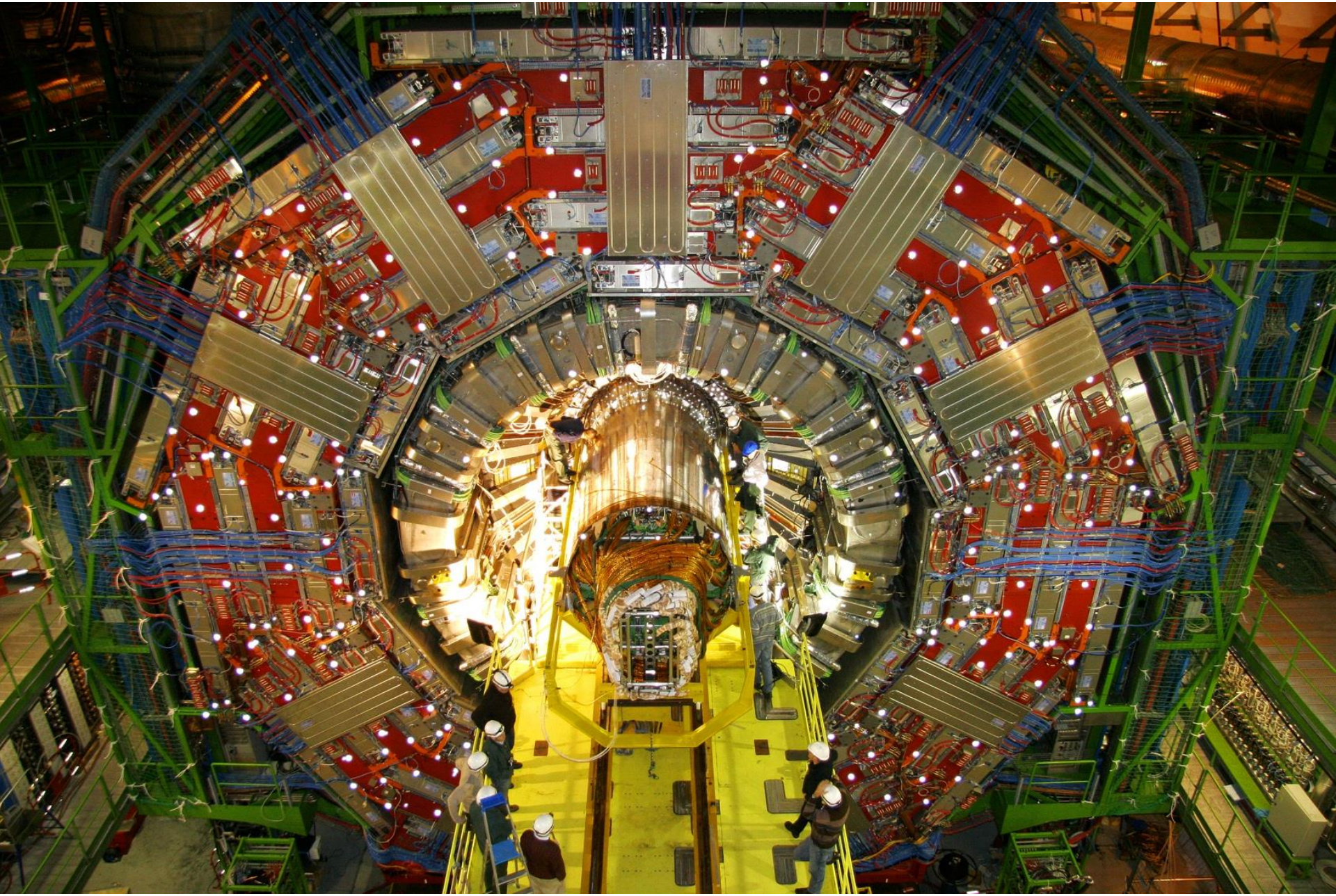
Átmérő:	**12 cm**	**($1.2*10^1$ cm)**
Energia:	**80 ezer eV**	**($8*10^4$ eV)**
Stáb:	**1+1 ember**	**($2*10^0$ fő)**
Mai ár:	**150 euro**	**($1.5*10^2$ €)**

A CERN LHC komplexum, 2010

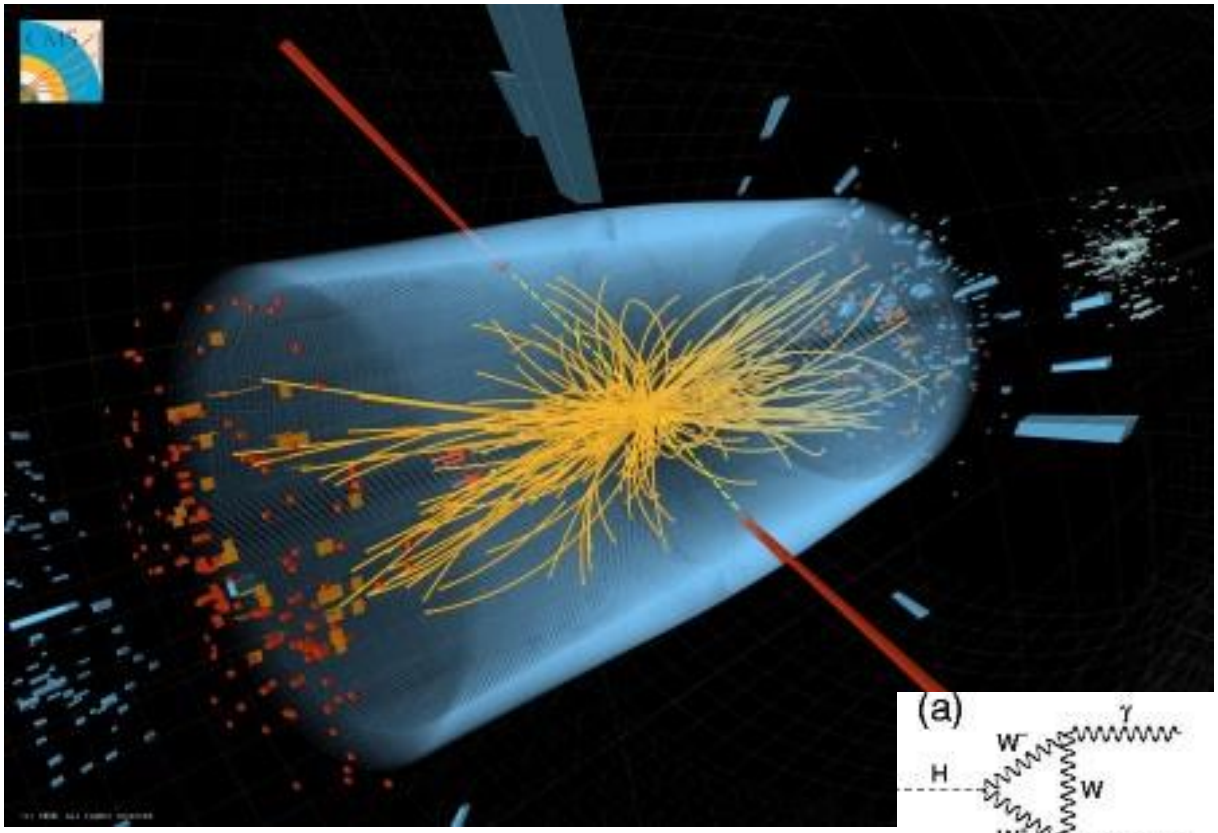
Átmérő:	**8.6 km**	**($8.6*10^5$ cm)**
Energia:	**7 TeV**	**($7*10^{12}$ eV)**
Stáb:	**2500 + 7500 fő**	**($1*10^4$ fő)**
Mai ár:	**~15 mrd euro**	**($1.5*10^{10}$ €)**




CMS
A Compact Solenoidal Detetor for LHC
MUON CHAMBERS
INNER TRACKER
E.M. CRYSTAL CAL.
HADRON CAL.
VERY FORWARD CALORIMETER
SUPERCONDUCTING COIL
RETURN YOKE
Total Weight : 12,000 t.
Overall diameter : 14.00 m
Overall length : 20.00 m
Magnetic field : 4 Tesla
CMS 1000

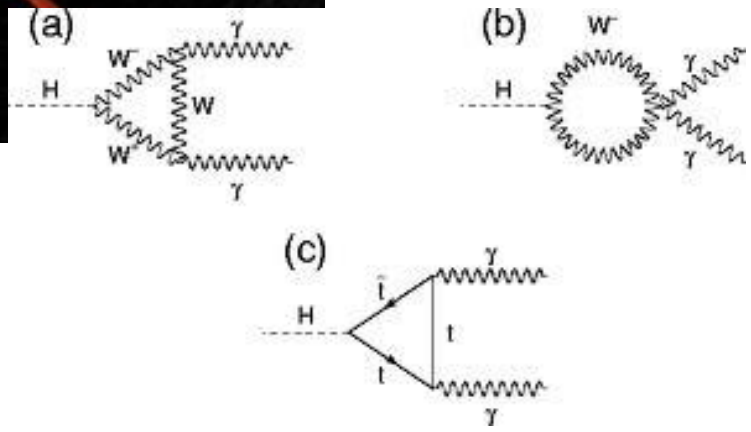


Tiszta jel, ha m_H = 110-130 GeV: p+p → H → γγ



Kísérlet (CMS)

Elmélet:



CERN tudományos programja: 2008 – 2030

2013-18: Emelni a luminozitást 800 M p+p ütközés / sec
Emelni az energiát 7 + 7 TeV-re

2013-30: Új részecskék felfedezése

Csillagászat: sötét anyag
sötét energia

Elmélet: szuperszimmetrikus (árnyék) részecskék

2030 + Új típusú gyorsító kifejlesztése
vagy
VLHC (100 TeV, 80 km gyűrű)

Informatikai kihívások !!!!!!!!

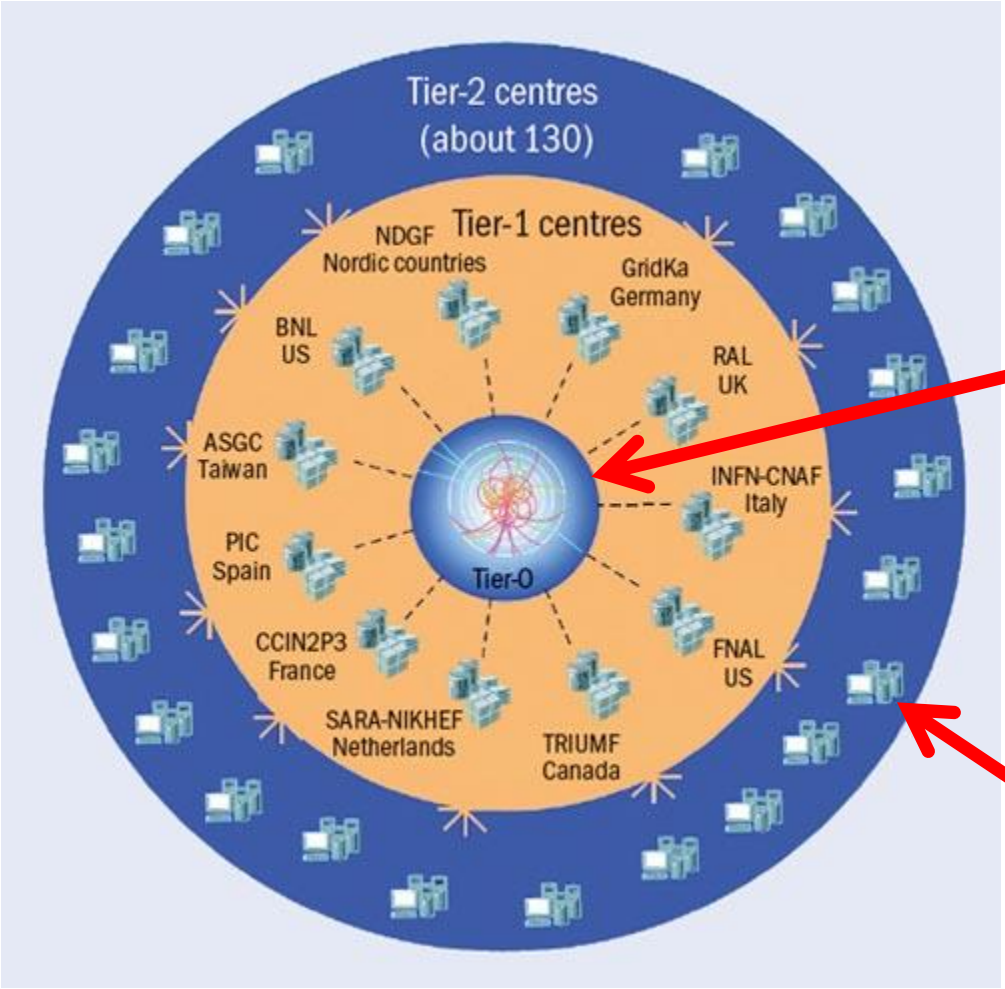
1 fb^{-1}: 70 milliószor millió ütközés

$7*10^{13}$ * 2 kbyte = 140 000 TeraB

Higgs felfedezés: 25 fb^{-1} adat = 3500 PetaB → 25 PetaB

A CERN TIER-0 számítógép központja

Az adatfeldolgozó piramis: TIER-0 → TIER-1 → TIER-2 → TIER-3



Jelenlegi állapot:

Wigner Adatközpont
Tier-0 kiterjesztés
5000 mag, 6 PB

Budapest Wigner FK
600 mag, 300 TB

WLCG, amint behálózza a Földet

WIGNER Adatközpont -- MTA WIGNER FK

2013. január 1: CERN TIER-0 kiterjesztés

1300 km 100 Gbit/s (400 Gbit/s)

Nagybiztonságú adattovábbítás, adatkezelés, adatbányászás

Misszió: Tudásközpont, know-how transzfer

Adatgyűjtés → adatrobbanás

Részecskefizika:

A jelenlegi kutatás fókuszpontja a Higgs-részecske megtalálása, megismerése, újabb részecskék megtalálása (sötét anyag):

CERN: 2 év alatt 25 PetaByte adat begyűjtése és átrostálása

kb. 350 db Higgs-bozon kiválogatása/megtalálása

Csillagászat:

A csillagos ég folyamatos megfigyelése, a történések pontos rögzítése, Új, nagyléptékű folyamatok, jelenségek megfigyelése

SLOAN Digital: 1 PB adat 10 év alatt

Pan-STARRS: 1.5 PB adat

Genetika:

Genetikai kód feltérképezése, genom-program, DNS-szekvenciálás, korrelációk meghatározása, személyre szabott gyógyászat megalapozása

10K Genom program: 5 PB

Kutatók szeretik azt hinni, hogy övék a legnagyobb adatbázisok !!

A Tudomány fejlődése az ókortól napjainkig - 2

Ókor

Empirikus tudomány

Középkor

Fenomenológia

Újkor

Elmélet

Kísérlet

Jelen - Jövő

Hatalmas Adatbázisok

Új paradigma:
Adatbányászás
Statisztikai elemzés
→ Informatika

Tényleg sok tudományos adatunk van/lesz?

2013: kb. 2000 PB (becslés)

Ebből: 1000 PB a személyi számítógépeken
1000 PB a szervereken, adatbankokban

2020: kb. 40 000 PB

Ebből: 1000 PB a személyi számítógépeken
39000 PB a szervereken, adatbankokban

WIGNER Adatközpont: 2013 → kb. 400 PB hely
2020 → kb. 2000 PB hely

Új technológiák adattárolásra
vagy újabb hatalmas tudományos adatközpontok

+ adatelemzők !!!! → → „Big Data Scientists”

És mi van az ÉLET-ben ???

Magyarország: 10 millió állampolgár (+ 5 millió turista)
365 nap

1 Petabájt adat → 0.25 MB / nap / fő

De:

1 PDF file: 0.1 – 0.2 MB
1 film: 3000-4000 MB
1 röntgenfelvétel: 1000-2000 MB
1 fényképezés: 1000-8000 MB → ???? PB HU-adat

Napjainkban óriási adatmennyiség keletkezik és továbbítódik.
Ezek nagy része elveszik, kisebb része tárolódik, minimális feldolgozás

Mi lesz ezzel a rengeteg adattal? → adatbányászat, adatfeldolgozás
„Big Data Technikus”

Hol keletkezik a legtöbb adat?

Adatok keletkezése:

--- orvosi ellátás
--- államigazgatás (törvények betartása, megszegése)
--- hírek, újságok, média
--- személyi szórakozás (fényképek, filmek, blogok)
--- háztartás
--- közlekedés
--- kiszolgáló infrastruktúra

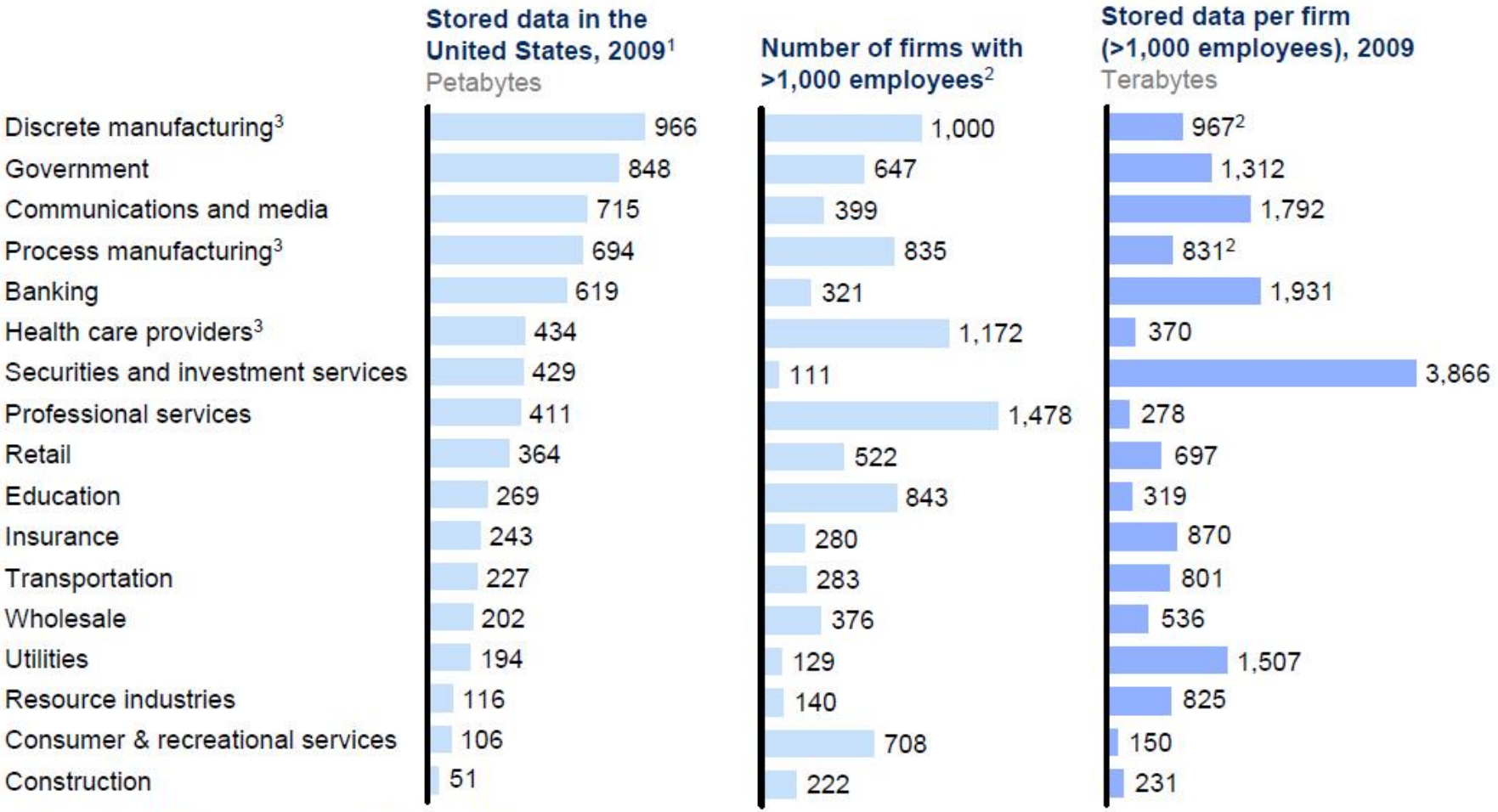
Adatok mozgatása:

--- gyors internet hálózatok
--- WiFi
--- Mobil-szolgáltatók

Adatok feldolgozása:

--- szerverparkok
--- számítási felhők

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte



1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

SO WHAT IS A PETABYTE ANYWAY?

Source – www.mozy.com

WHAT IS A PETABYTE?

TO UNDERSTAND A PETABYTE WE MUST FIRST UNDERSTAND A GIGABYTE.

- 1 GIGABYTE — 7 MINUTES OF HD-TV VIDEO
- 2 GIGABYTES — 20 YARDS OF BOOKS ON A SHELF
- 4.7 GIGABYTES — SIZE OF A STANDARD DVD-R

THERE ARE A MILLION GIGABYTES IN A PETABYTE

"Let me repeat that: we create as much information in two days now as we did from the dawn of man through 2003." *(That's something like 5 Exabytes of Data). - Eric Schmidt – Google 8/10*

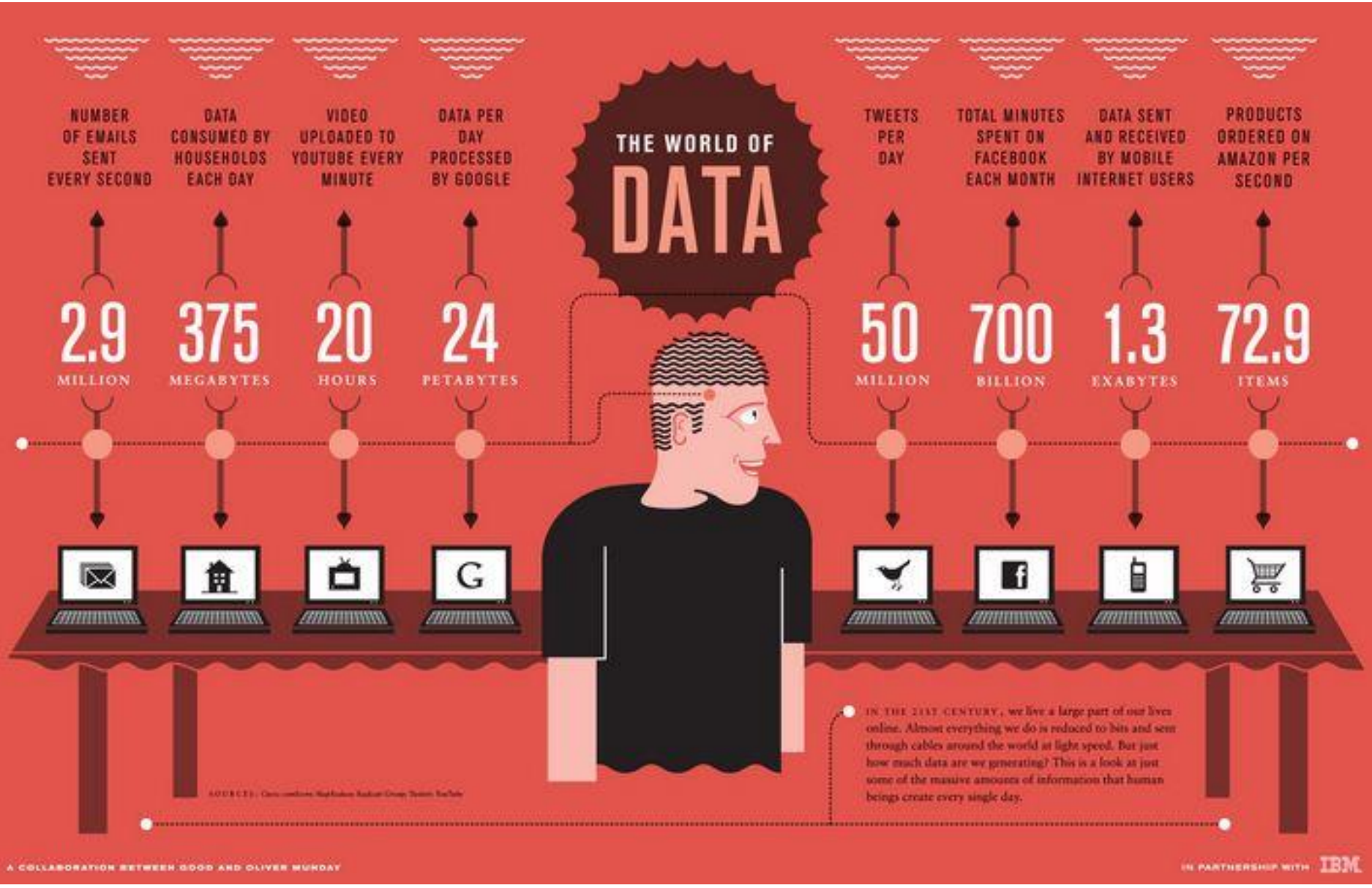
A PETABYTE IS A LOT OF DATA

- 1 PETABYTE — 20 MILLION FOUR-DRAWER FILING CABINETS FILLED WITH TEXT
- 1 PETABYTE — 13.3 YEARS OF HD-TV VIDEO
- 1.5 PETABYTES — SIZE OF THE 10 BILLION PHOTOS ON → FACEBOOK
- 15+ PETABYTES — INTERNET USER'S DATA BACKED UP ON MOZY.COM
- 20 PETABYTES — THE AMOUNT OF DATA PROCESSED BY GOOGLE PER DAY
- 20 PETABYTES — TOTAL HARD DRIVE SPACE MANUFACTURED IN 1995
- 50 PETABYTES — THE ENTIRE WRITTEN WORKS OF MANKIND, FROM THE BEGINNING OF RECORDED HISTORY, IN ALL LANGUAGES

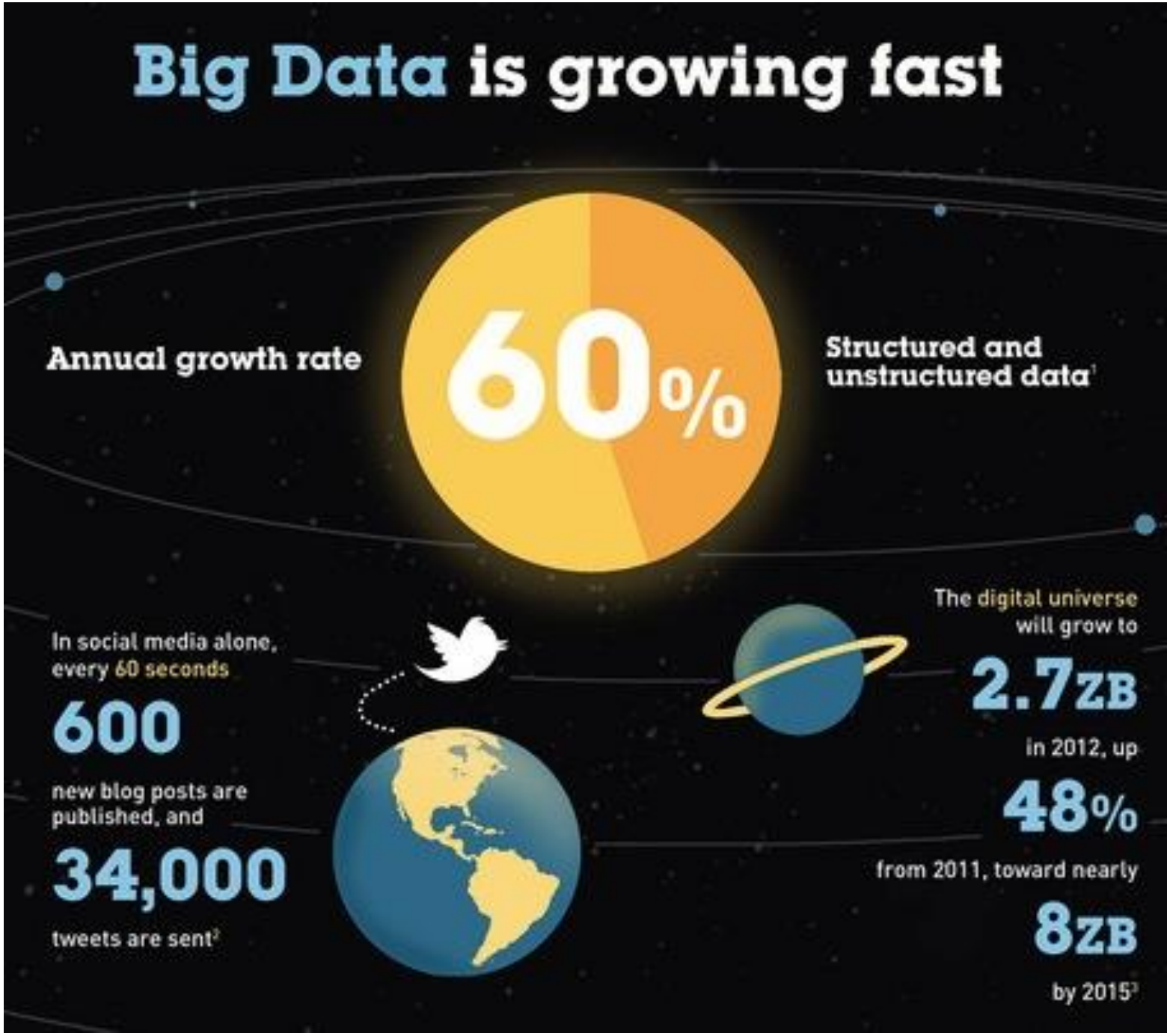
Twitter:
Over 7TB a Day in Tweets.

A ZETABYTE IS ONE MILLION PETABYTES!

Facebook:
More that 750 Million Users.
Average user creates 90 Pieces of content each month.
More than 30B pieces of content shared each month.


THE WORLD OF
DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9
MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375
MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20
HOURS
DATA PER DAY PROCESSED BY GOOGLE
24
PETABYTES
TWEETS PER DAY
50
MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700
BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3
EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9
ITEMS
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

Big Data is growing fast



Közös kezdeményezésünk:

VLAB4BIGD – Virtuális Laboratórium Big Data feladatokhoz

Résztvevők a konzorciumban:

Silicon Computers (vezető)
MTA Wigner Fizikai Kutatóközpont & ELTE
MTA Közgazdaságtudományi és Regionális Tud. Kutatóközpont
MTA Társadalomtudományi Kutatóközpont
Budapesti Műszaki és Gazd. Tudományi Egyetem
Semmelweis Egyetem
Szegedi Tudományegyetem
Országos Meteorológiai Szolgálat
Csertex Kft.

Stratégiai partnerek:

MICROSOFT
Morgan Stanley
MAVIR Zrt.

Nyitott együttműködés: → → Big Data Club

Zárszó:

A jövő félelmetesen nagy kihívásokat tartogat számunkra.

Meg is ijedhetünk, senki nem fog csodálkozni ezen.

De fel is vehetjük a kesztyűt !

Sikeres Big Data Day-t kívánok minden kedves résztvevőnek!

WIGNER Adatközpont -- MTA WIGNER FK

Szept. 28: „CERN OPEN Day a WIGNERben”
Regisztráció: http://WIGNER.MTA.HU