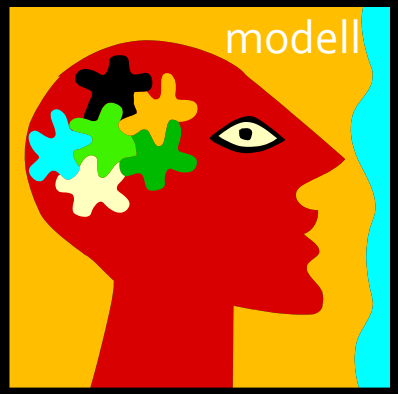
Csabai István
ELTE Komplex Rendszerek Fizikája Tanszék

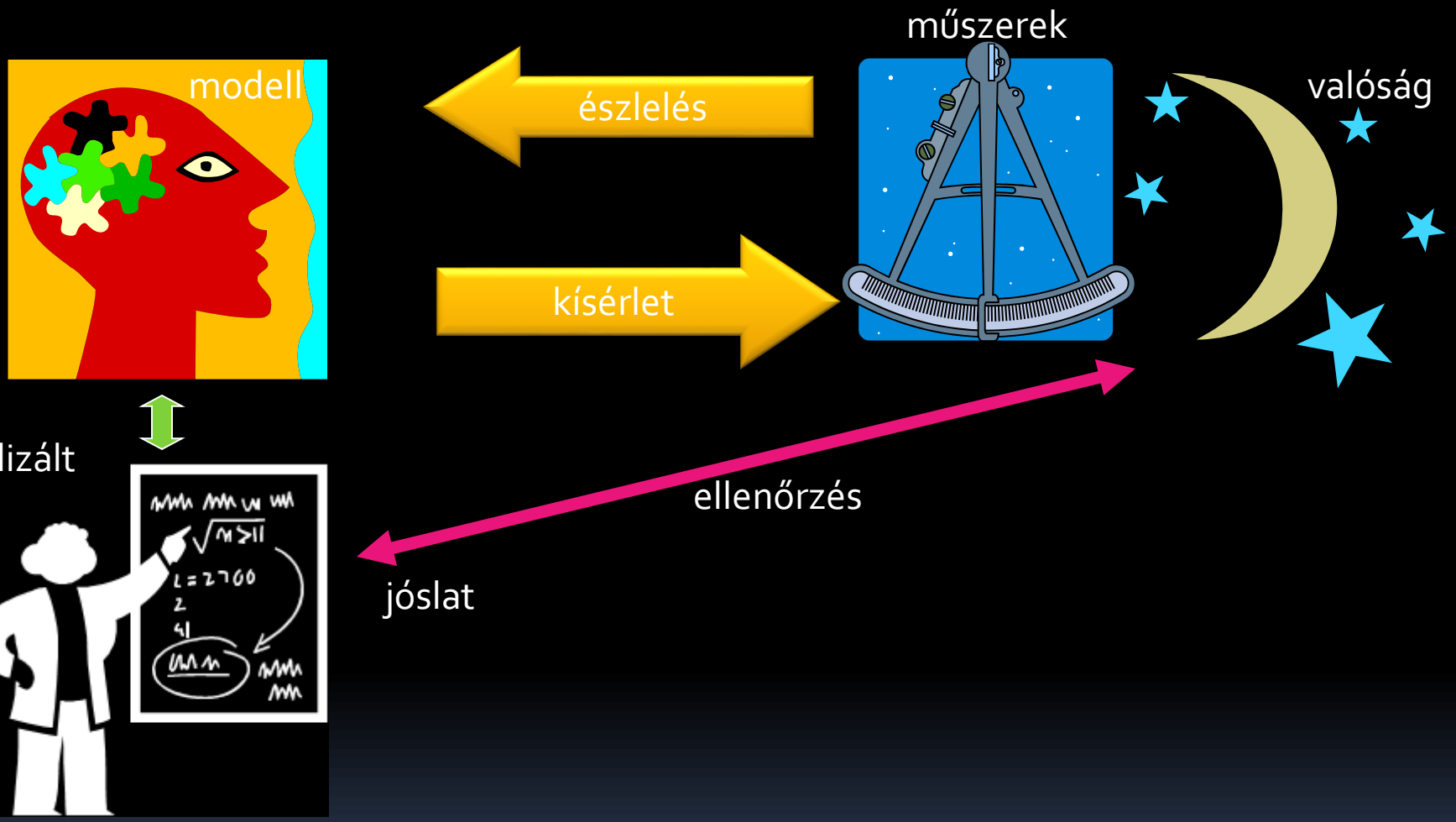
ADAT-INTENZÍV MEGKÖZELÍTÉS A MODERN TERMÉSZETTUDOMÁNYOKBAN

2015 Big Data Mafihe Téli Iskola –ELTE 2015.02.02

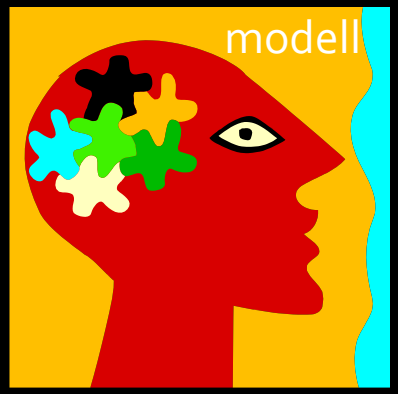
A természettudományok fejlődése: korai idők

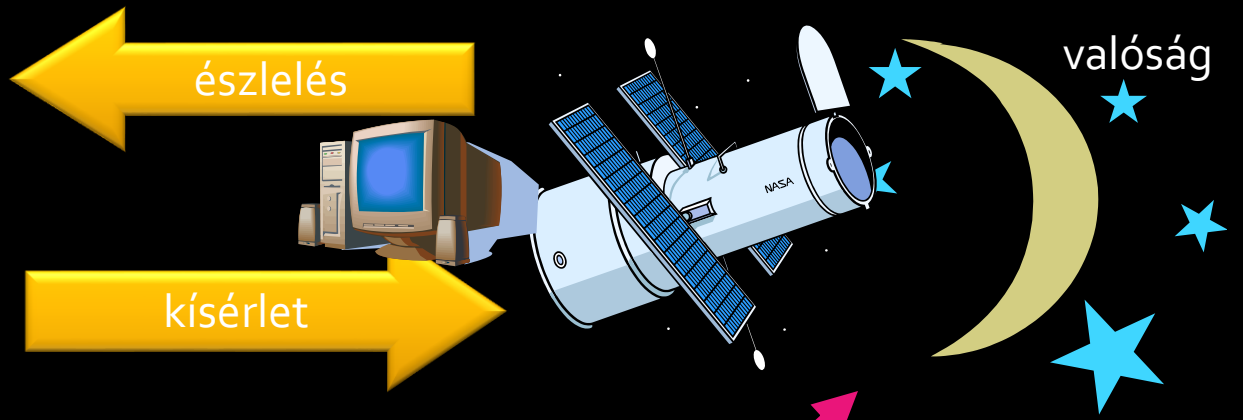


A természettudományok fejlődése: a múlt

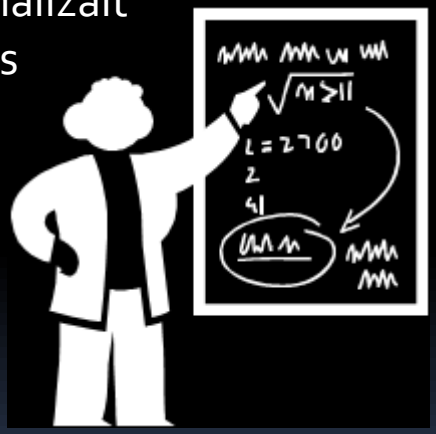


A természettudományok fejlődése: a jelen





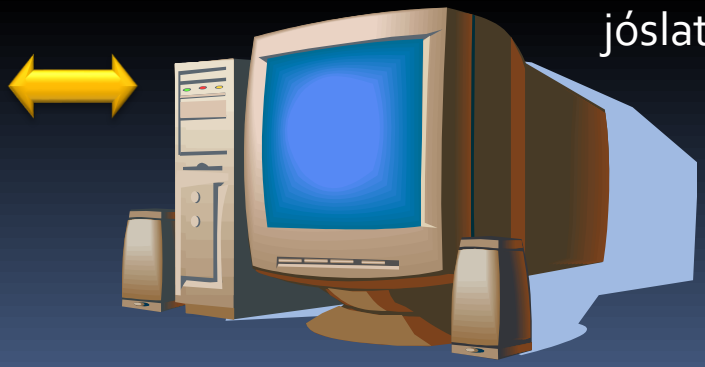
formalizált leírás



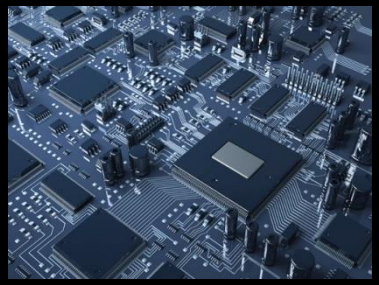
ellenőrzés

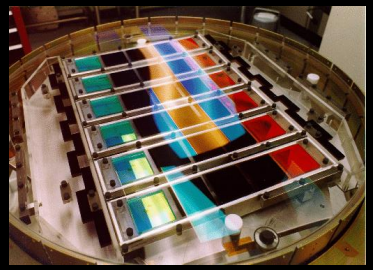
virtuális valóság

jóslat

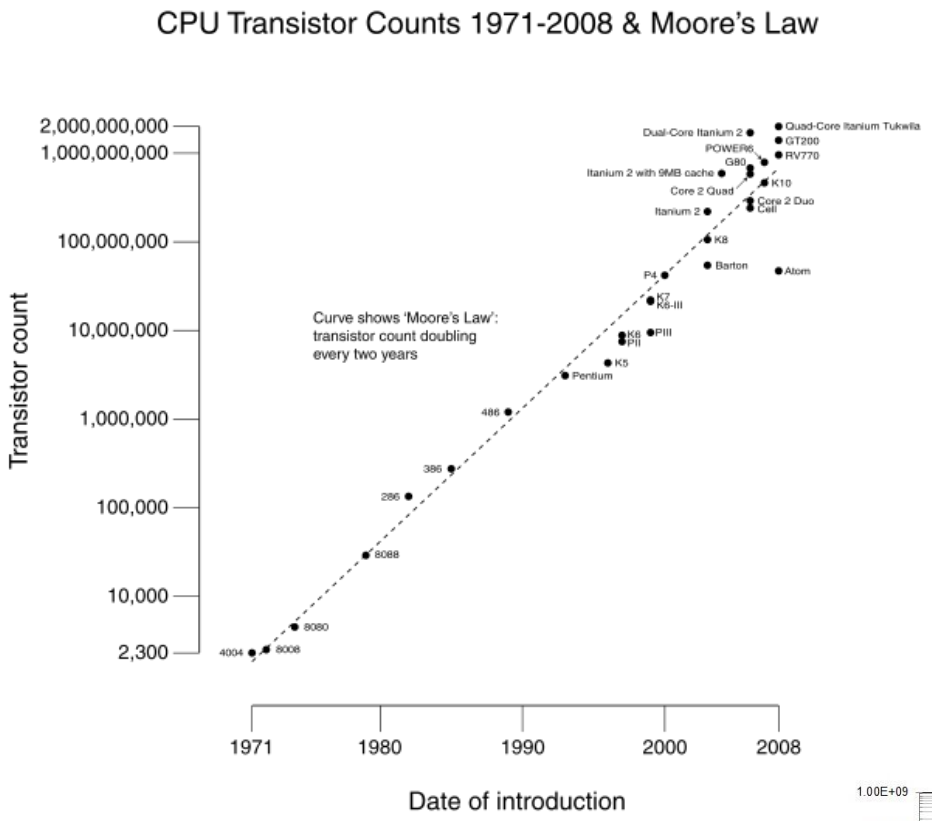


Exponenciális növekedés

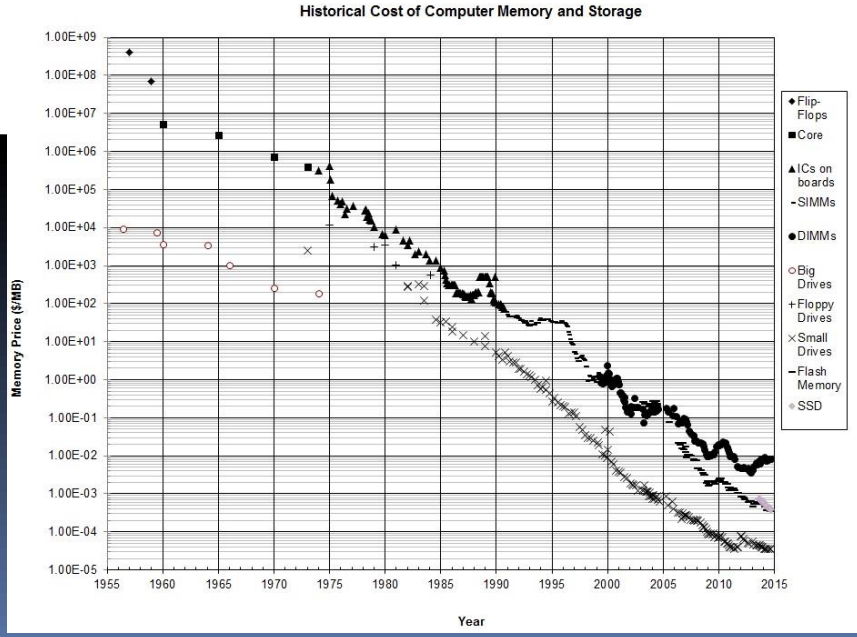




Elektronika → Detektorok → Adatok



Gordon E. Moore, Intel Chairman, 1965

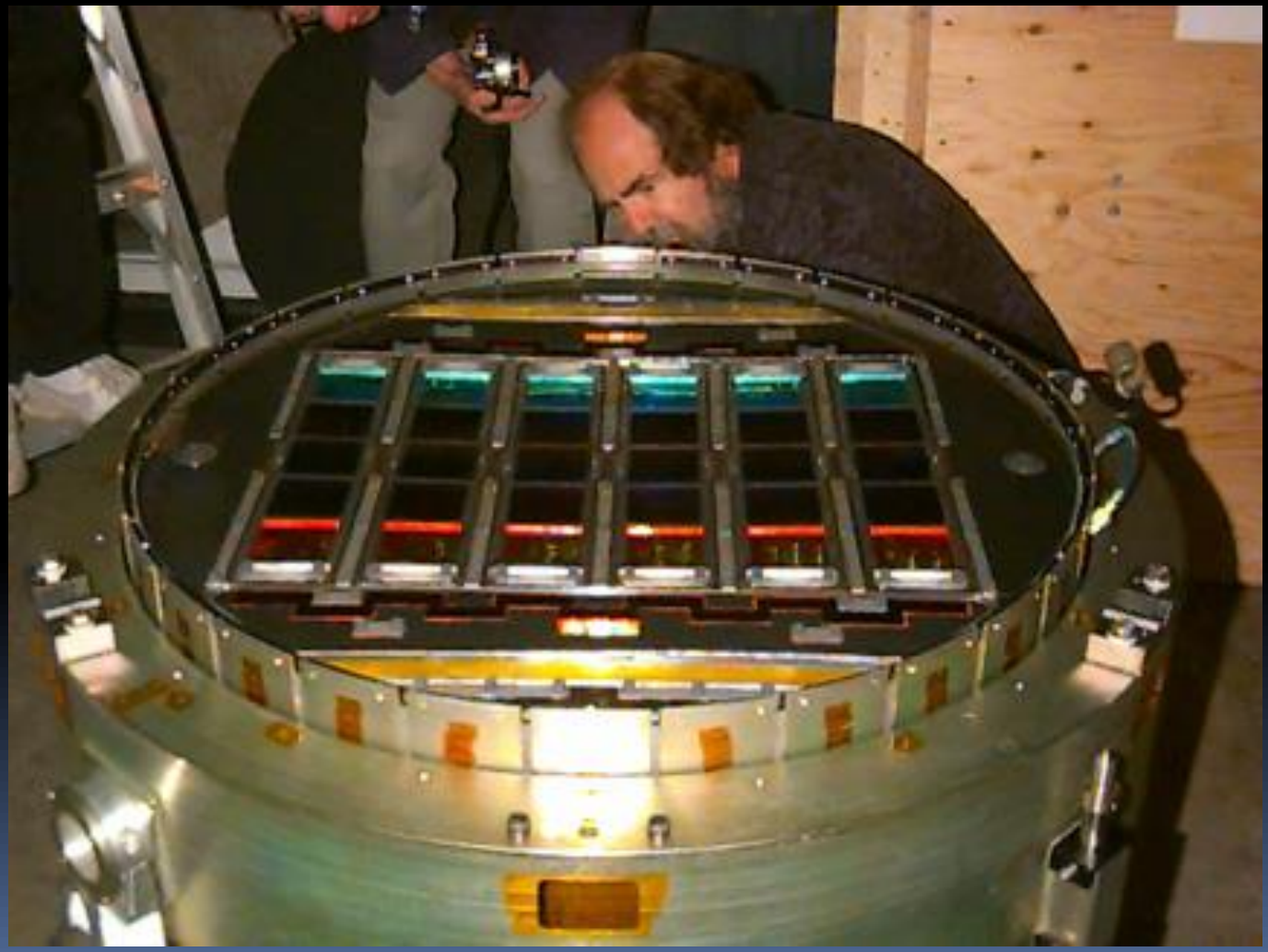


Nagy tudományos felmérés mintapélda:

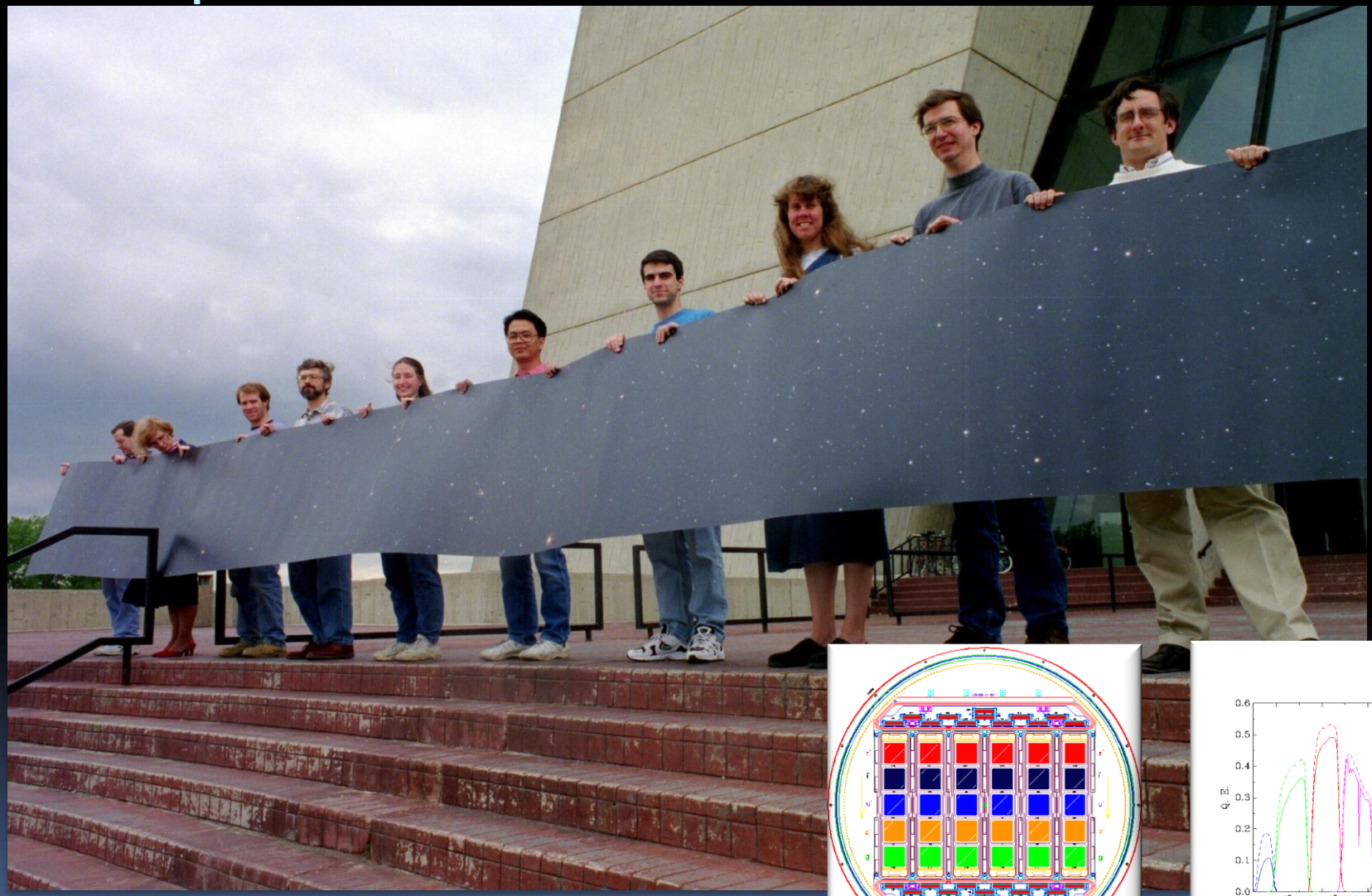
SDSS: AZ UNIVERZUM 3D TÉRKÉPE

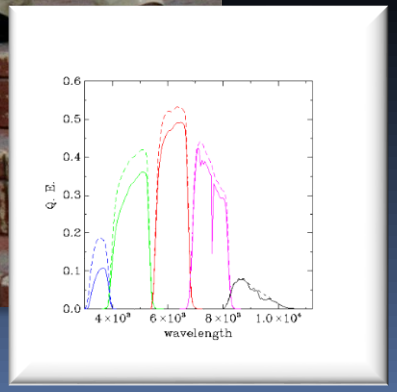
2.5m tükör – 3 fok látószög

120 megapixeles kamera

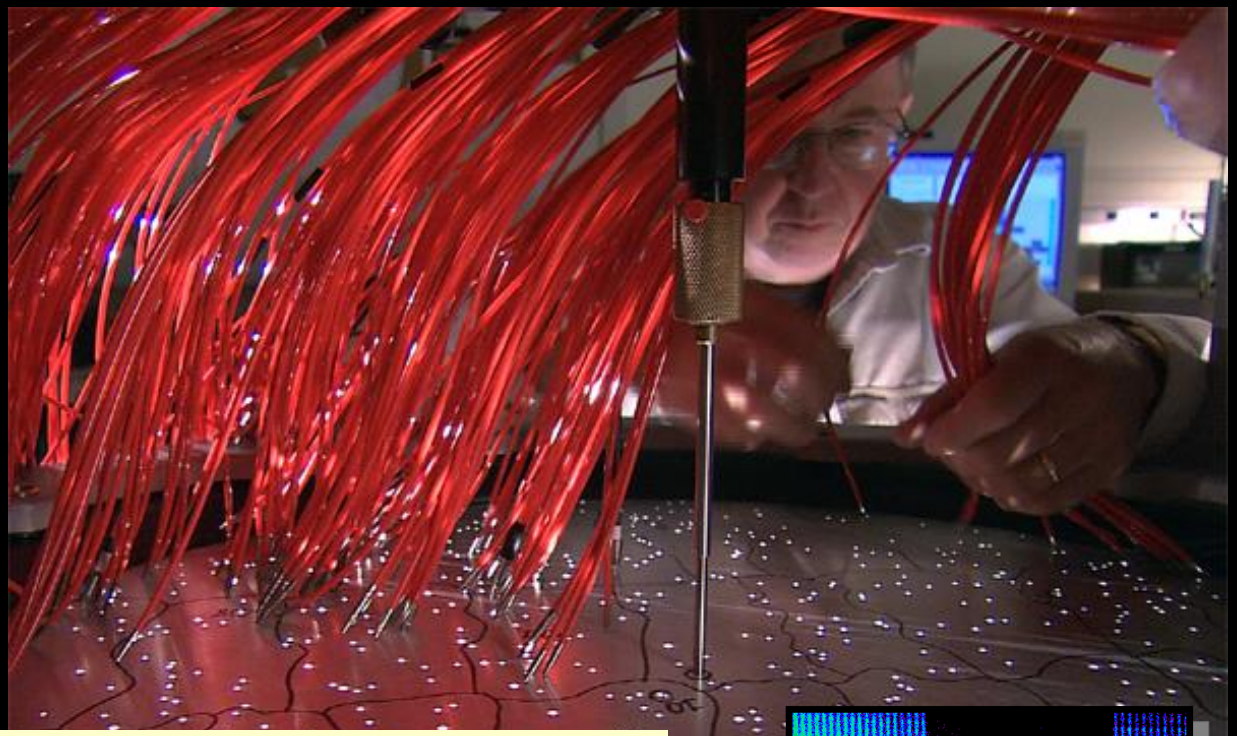


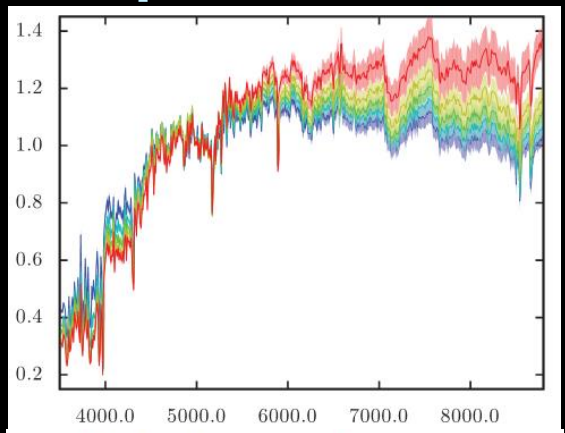
2.5 terapixeles kép – 300 millió galaxis – 5 optikai sáv

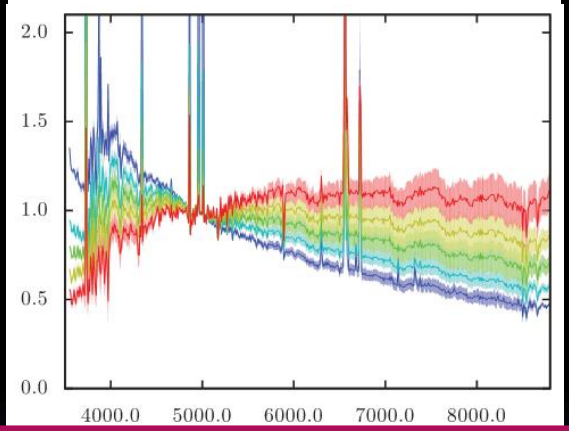




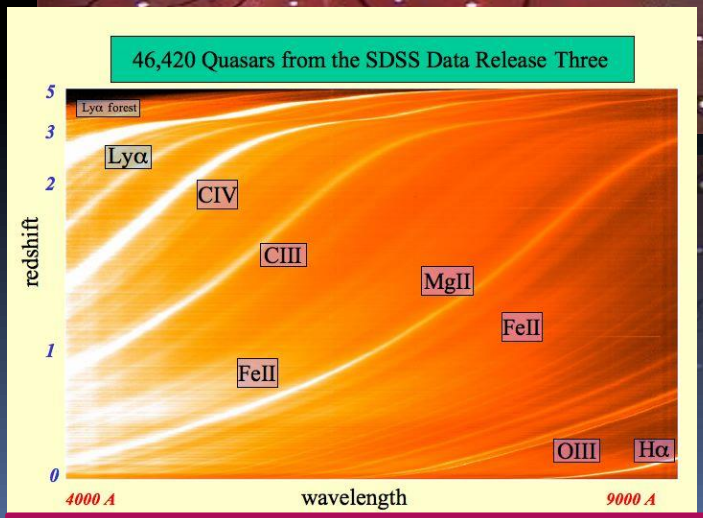
640 üvegszál– 1 millió spektrum



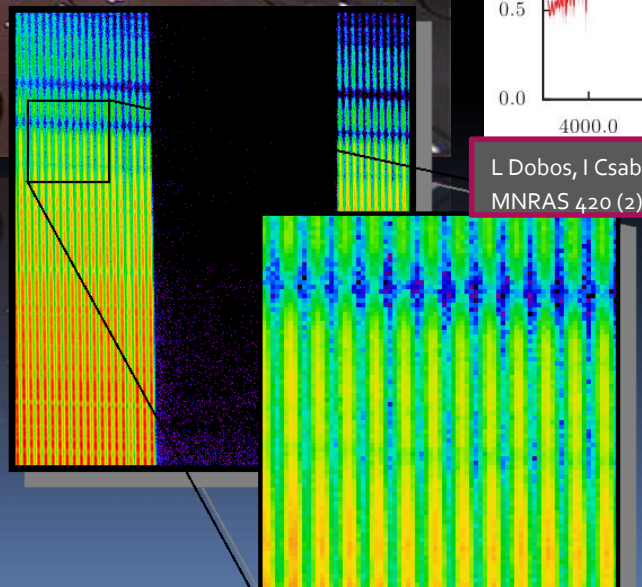




L Dobos, I Csabai, CW Yip, T Budavári, V Wild, AS Szalay
MNRAS 420 (2), 1217-1238(2012)

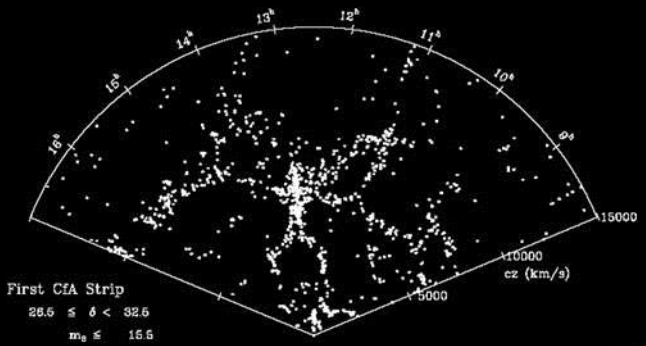


DP Schneider + SDSS collab. Astr. J. 130 (2), 367 (2005)



KOZMIKUS HÁLÓ

CfA: 1100 galaxies



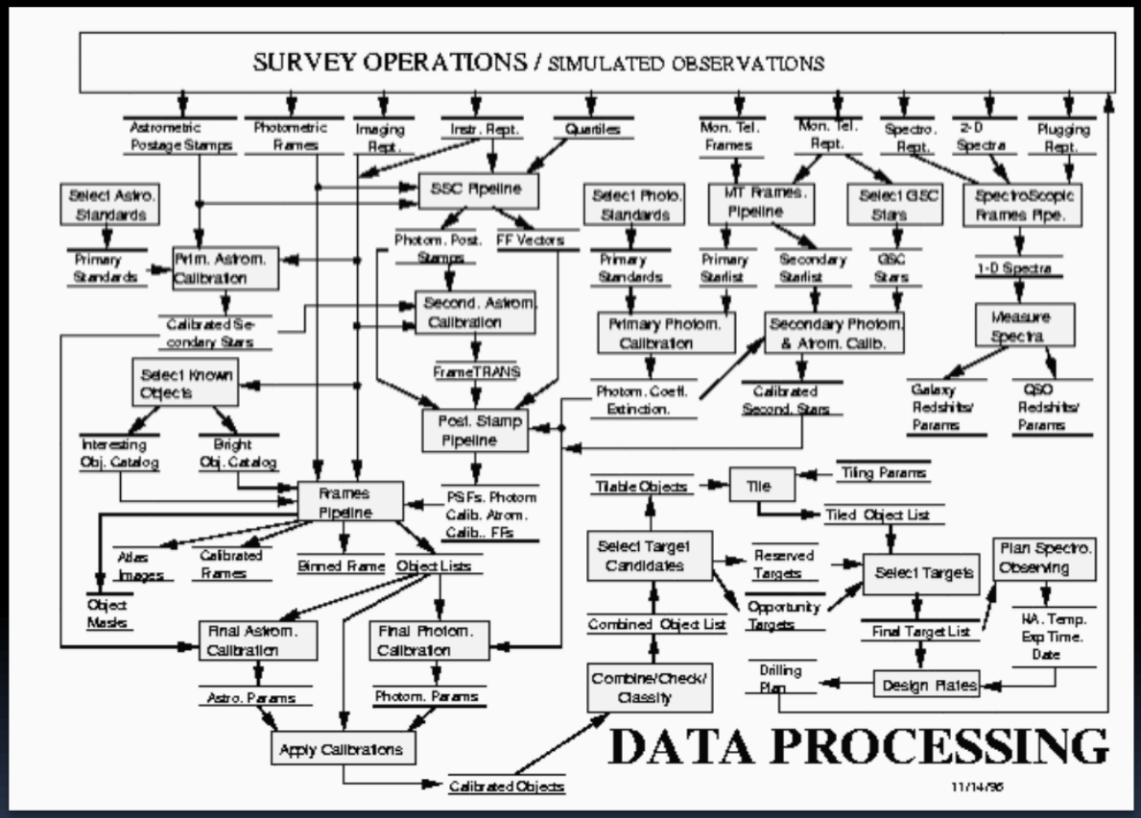
MJ Geller, JP Huchra, Science 246, 897 (1989)

SDSS: 1M galaxies

KN Abazajian + SDSS collab. Astr. J.S. 182(2) 543(2009)

Adatfeldolgozási kihívás

- Automatikus “pipeline”
  - *Több mint 150 ember-évnyi fejlesztés*
  - *Első nagyprojekt ahol a szoftver fejlesztésre fordítódott az erőforrások nagyobbik hányada*
- Nagy adattömeg
  - *Több mint 300 millió objektum, egyenként 300+ paraméter*
  - *Közel 100 TB nyers adat, 10 TB katalógus, 2.5 terapixel*
  - *PUBLIKUS ADATBÁZIS (VO)*



The sloan digital sky survey: Technical summary
DG York + SDSS collab.The Astron. J.l 120 (3), 1579 (2000)

PZ Kunszt, AS Szalay, I Csabai, AR Thakar;
ADASS IX 216, 141(2007)

Csillagászati kérdések – csillagászati adathalmazon

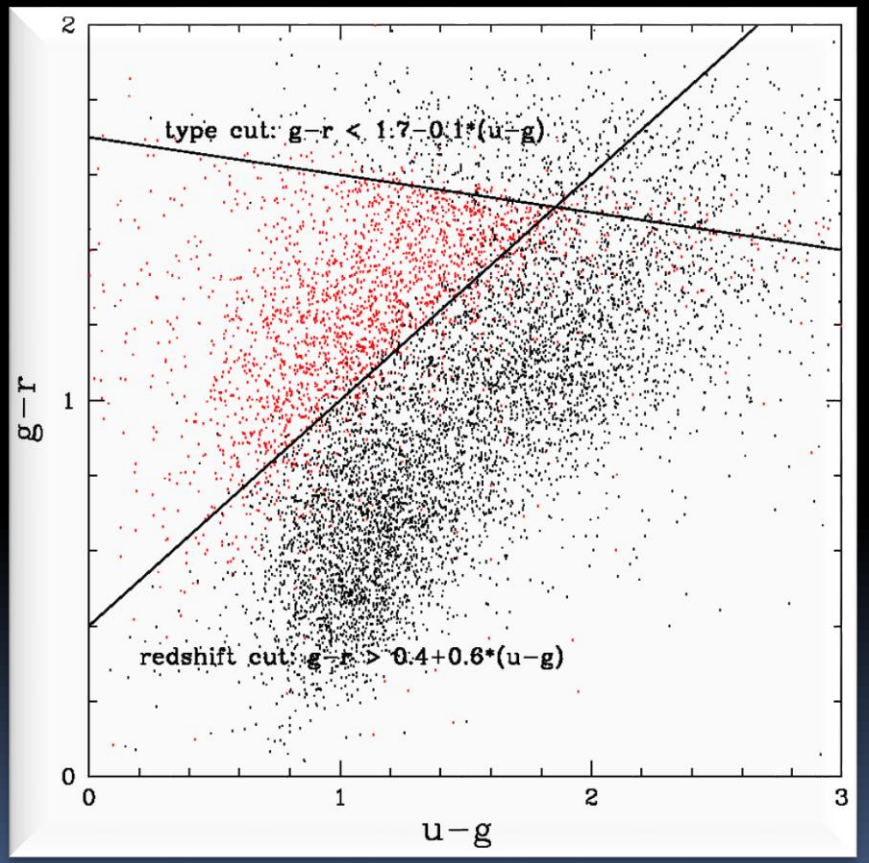
Csillag/galaxis szeparáció
Kvazár target kiválasztás

→ "vágások"

→ **Sok dimenziós poliéderek**

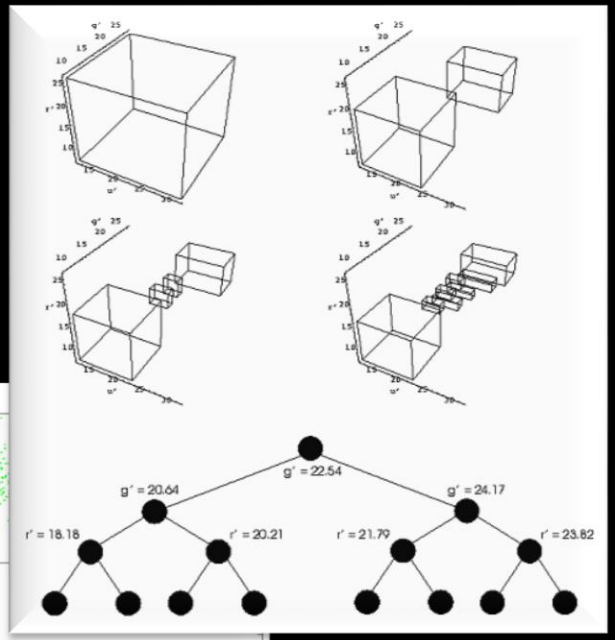
Skyserver: évente több mint 1 millió lekérdezés

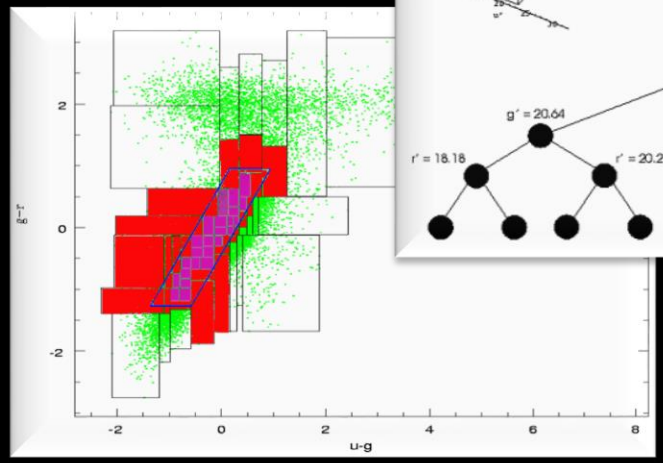
```
petroMag_i > 17.5 and (petroMag_r > 15.5 or petroR50_r > 2)
and (petroMag_r > 0 and g > 0 and r > 0 and i > 0) and (
(petroMag_r-extinction_r) < 19.2 and (petroMag_r -
extinction_r < (13.1 + (7/3) * (dered_g - dered_r) + 4 * (dered_r
- dered_i) - 4 * 0.18) ) and ( (dered_r - dered_i - (dered_g -
dered_r)/4 - 0.18) < 0.2) and ( (dered_r - dered_i - (dered_g -
dered_r)/4 - 0.18) > -0.2) and ( (petroMag_r - extinction_r +
2.5 * LOG10(2 * 3.1415 * petroR50_r * petroR50_r)) < 24.2) )
or ( (petroMag_r - extinction_r < 19.5)
and ( (dered_r - dered_i - (dered_g - dered_r)/4 - 0.18) > (0.45 -
4 * (dered_g - dered_r)) ) and ( (dered_g - dered_r) > (1.35 +
0.25 * (dered_r - dered_i)) ) ) and ( (petroMag_r - extinction_r
+ 2.5 * LOG10(2 * 3.1415 * petroR50_r * petroR50_r) ) < 23.3 )
)
```

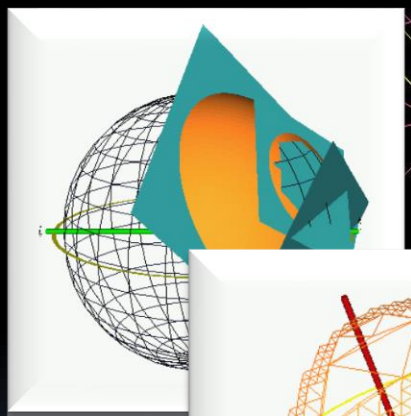


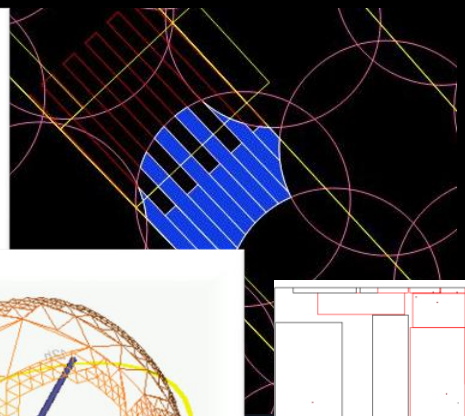
Indexelés, adatbázisok

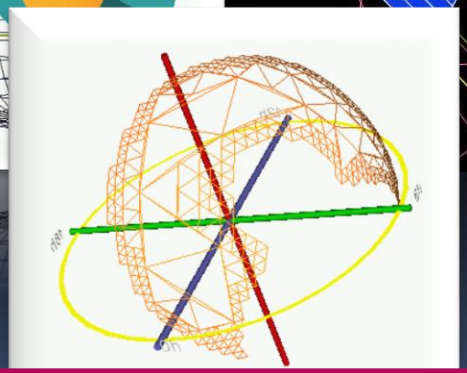
- Az adat nem fér el a memóriában
- A háttértár elérése nagyságrendekkel lassabb
  - Pl. SDSS adatok átolvasása ~1 nap

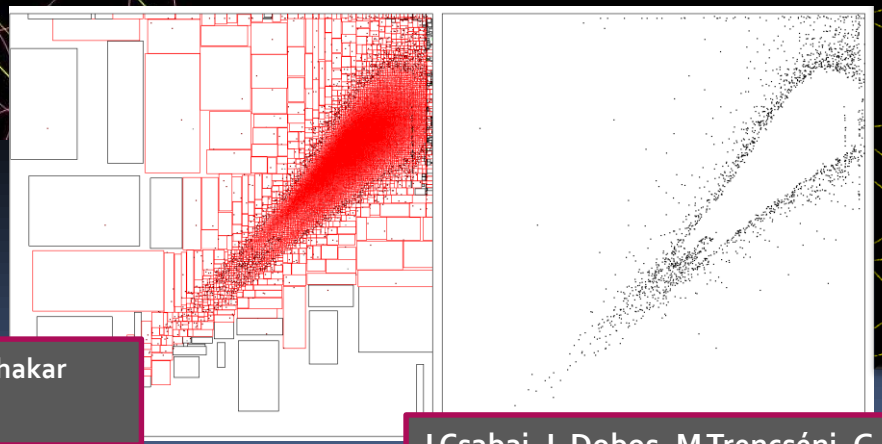


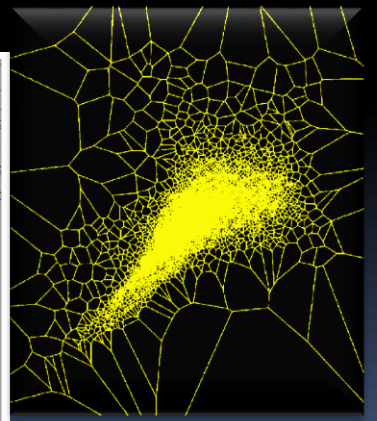












AS Szalay, J Gray, G Fekete,P Kunszt, P Kukol, A Thakar MSR -TR 123 (2005)

T Budavari, L Dobos, AS Szalay, G Greene, J Gray, AH Rots ASP Conf . Ser. 376, 559 (2007)

I Csabai, L Dobos, M Trencséni, G Herczegh, P Józsa, N Purger, T Budavári, AS Szalay Astr. N. 328 (8), 852 (2007)

Adatbázisok, irányzatok

- RDBMS
  - +Üzleti célra fejlesztve, optimalizált IO/memória elérés,deklaratív programozhatóság (SQL), párhuzamos queryk, standard API (ODBC,JDBC)
  - -Relációs adatmodell gyakran nem elég (mátrixok, gráfok, [arrayLib]), nem elosztott [skyQuery]
  - ACID: Atomic,consistent,isolated,durable
- NoSQL,BigTable, Hadoop/MapReduce, column store
  - Akár sok ezer (olcsó) szerver, elosztottság a fő szempont
  - Rendelkezésre állás vs. konzisztencia
  - BASE: Basically available,soft state,eventually consistent)
  - SciDB (array, grid), MonetDB, Vertica

L Dobos, AS Szalay, J Blakeley, B Falck, T Budavári, I Csabai
Astronomical Data Analysis Software and Systems XXI 461, 323 (2012)

L Dobos, T Budavári, N Li, AS Szalay, I Csabai
Scientific and Statistical Database Management, 159-167 (2012)

	RDBMS	BigTable	Hadoop
Deklaratív nyelv, optimalizáló	✓		
Optimalizált scan	✓		✓
Elsődleges kulcs szerinti elérés	✓	✓	
Indexek szerinti keresés	✓	✓	
Join műveletek	✓		
Párhuzamos végrehajtás	✓		✓
Egyszerű sharding		✓	✓
Tranzakciók	**ACID**	**BASE**	
Többlépéses tranzakciók	✓		
Durability	back-up/log	replikáció	replikáció
Egyszerű load balancing		✓	✓
Nem strukturált adat		✓	
Szabványos API	✓		

(Dobos. L. 2014.)

Párhuzamosítás, CPU/GPU vs. IO

- Amdahl törvények kiegyensúlyozott rendszerek esetére (1965)
- Teljes probléma: $1 = P + S$
- Max gyorsulás: $a = 1 / (S + P / N)$
- Amdahl-szám: $\frac{\text{1 bit IO / s}}{\text{1 utasítás / s}}$
- Memória: $\frac{\text{1 bájt memória}}{\text{1 utasítás / s}}$
- Energia?

- Blue Gene: $A_{IO} = 0{,}013$
- Graywulf: $A_{IO} = 0{,}5$
- Amdahl: $A_{IO} = 1{,}25$

L Dobos, I Csabai, AS Szalay, T Budavári, N Li
Proceedings of the 25th International Conference on Scientific and Statistical Database Management, ACM, (2013)

(Dobos. L. 2014.)

Virtual Observatory / cloud

- „Ha az adathegy nem megy …”
- Amazon WebServices, MS Azure, Google Cloud,…
  - Software as a service (SaaS)
  - Development as a service (DaaS)
  - Platform as a service (PaaS)
  - Infrastructure as a service (IaaS)

SDSS Data Access

- SkyServer
  - Web browser-based synchronous access
  - Meant to support several levels of users
    - From casual to moderately advanced queries
    - From simple form-based to direct SQL queries
    - From cone (radial) search to crossid type searches
  - Visual tools to browse image and catalog data
  - Stored procedures
  - API access, e.g. emacs interface, sqlcl (command-line)
  - Strict limits on execution time and output size
    - Fair use for everyone, robots/crawlers discouraged
- ImgCutout
  - Finding Chart and JPEG image browser
  - Accessible from SkyServer (Visual Tools)

CasJobs

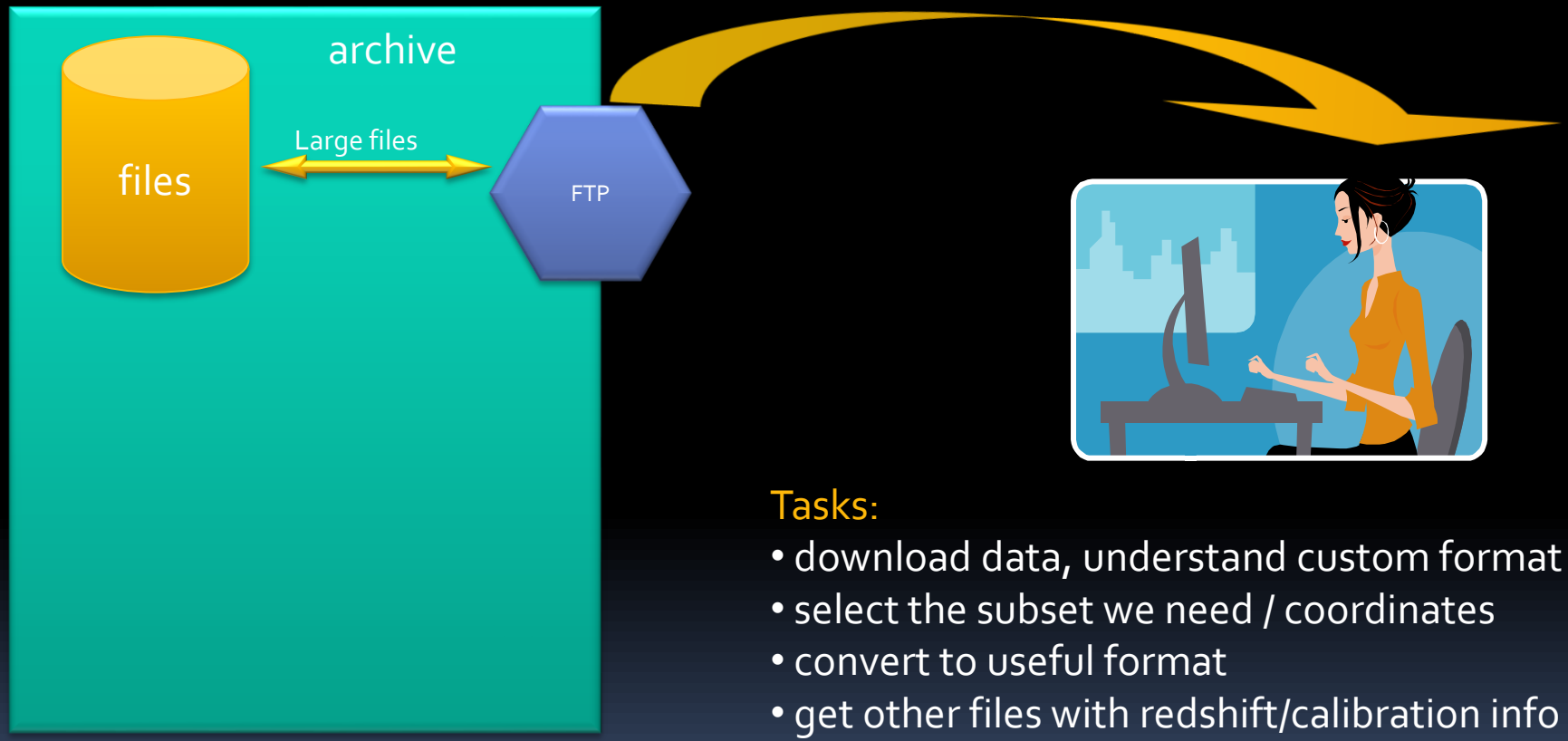
- Link in SkyServer (http://cas.sdss.org/casjobs)
- Batch Query Workbench, personal user DB (MyDB)
  - **Quick** mode: 1 minute cutoff
  - **Submit** mode: up to 8 hours in "long" queue
    - 24-hr queue for collab members
- Preferred method for serious queries
- **MyDB** database to save results of your queries
  - Define your own functions, procedures too
  - Share your tables with collaborators (groups)
- Job history, plotting, FITS/CSV/VOTable output
- Restricted (collab-only databases)
- Table **Import** (upload) for your own data
- **Groups** to share your results with collaborators
- Command-line access Java tool also downloadable
- SOAP/Web Services access

SkyServer Help Resources

- **Help** menu option on top right of SkyServer
- Start with **Archive Intro**
- Next look at **Query Limits** and **How To** pages
- Then **Introduction to SQL** and **Sample Queries**
  - Look at **Optimizing Queries** page (esp. bookmark bug)
  - Try out some of the sample queries
  - Cut and paste to SQL search page (Tools→Search→SQL)
- Browse **FAQ** and **Schema Browser**

- **Glossary** , **Table Descriptions** and **Algorithms**
  - Searchable, dynamically loaded from DB, interlinked
  - The G, A, T symbols are links to Glossary, Algorithm and Table Description entries
- **Data release and technical papers**

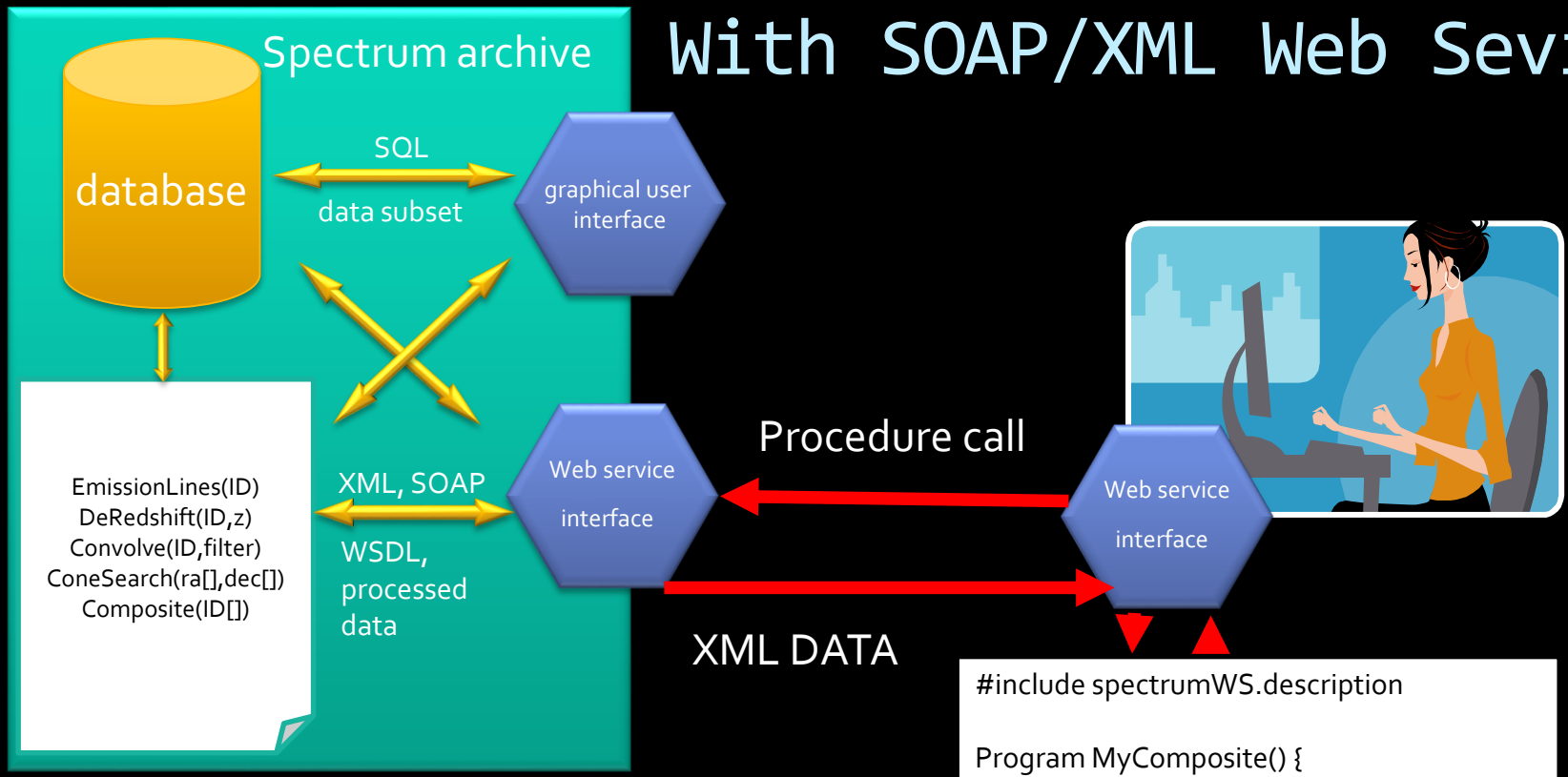
Example: my PhD student has a list of coordinates of galaxies and wants composite spectrum



Tasks:

- download data, understand custom format
- select the subset we need / coordinates
- convert to useful format
- get other files with redshift/calibration info
- rebin/rescale spectrum, shift to restframe
- write processing code
- repeat with other archives

With SOAP/XML Web Sevices




```
#include spectrumWS.description

Program MyComposite() {
 ...

 for(i=1; i<N; i++){
   ra[i] = ReadRA(i);
   dec[i] = ReadDEC(i);
 }

 idStruct ID[] = new idStruct[N];
 ID = ConeSearch(ra,dec);
 sedStruct SED = new sedStruct();
 SED = Composite(ID);
 Plot(SED);
 ...
}
```

optical archive

database

SQL

data subset

graphical user interface

Processing functions
Stored procedures
Search tools
Visualization tools

XML, SOAP

WSDL, processed

Web service interface

Web service interface

IR archive

database

SQL

data subset

graphical user interface

Processing functions
Stored procedures
Search tools
Visualization tools

XML, SOAP

WSDL, processed data

Web service interface

UV archive

database

SQL

data subset

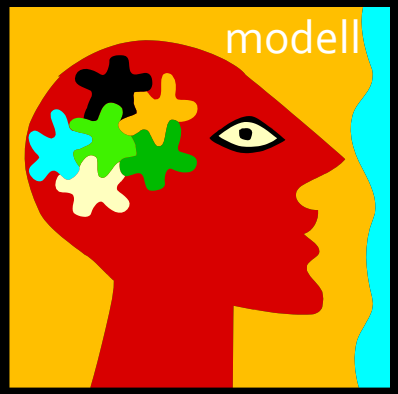
graphical user interface

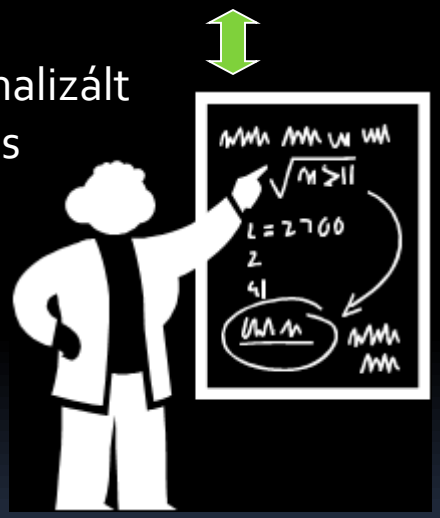
Processing functions
Stored procedures
Search tools
Visualization tools

XML, SOAP

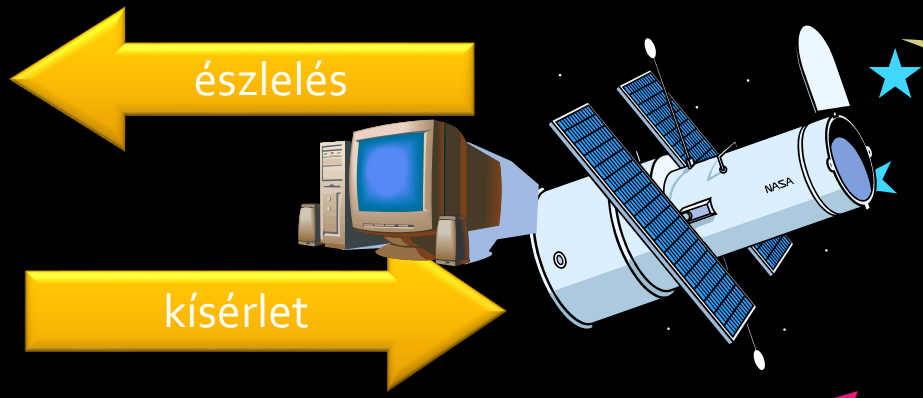
WSDL, processed data

Web service interface



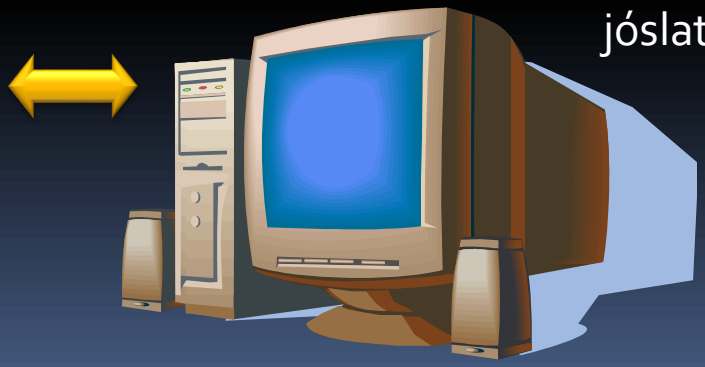


formalizált leírás



elenőrzés

virtuális valóság



jóslat

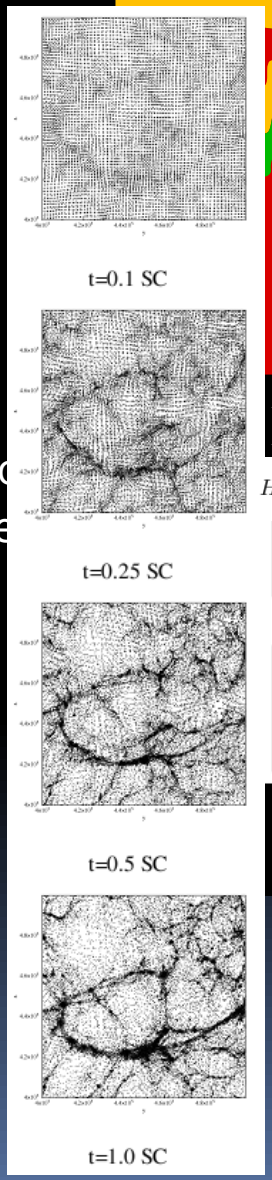
valóság

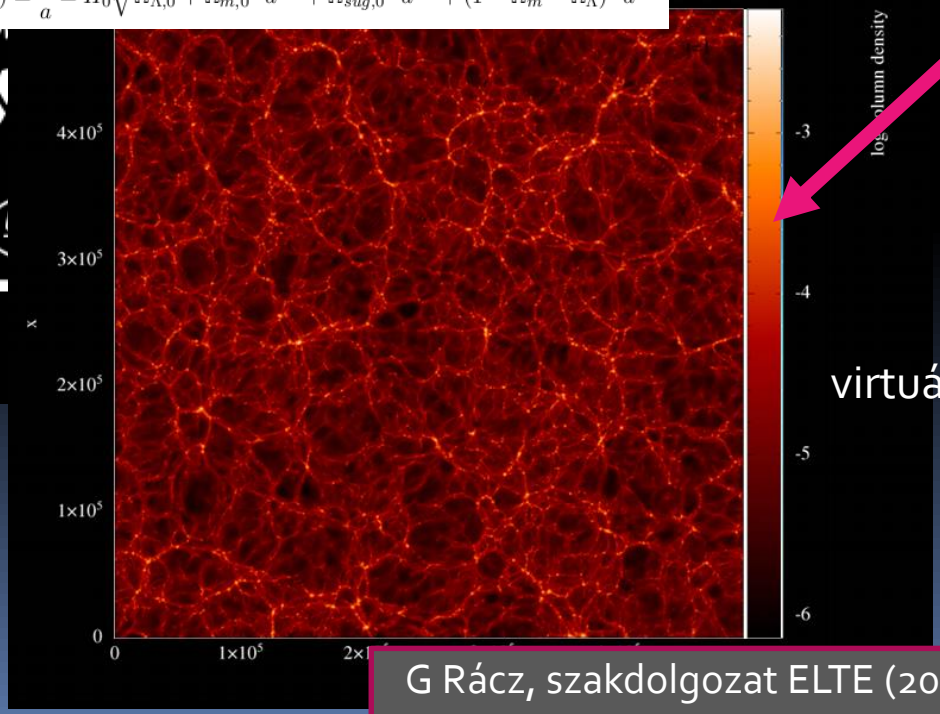
modell

$$H(a) = \frac{\dot{a}}{a} = H_0\sqrt{\Omega_{\Lambda,0} + \Omega_{m,0}\cdot a^{-3} + \Omega_{sug,0}\cdot a^{-4} + (1-\Omega_m-\Omega_\Lambda)\cdot a^{-2}}$$

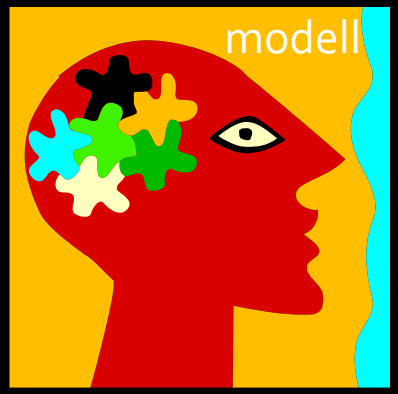
ellenőrzés

virtuális valóság



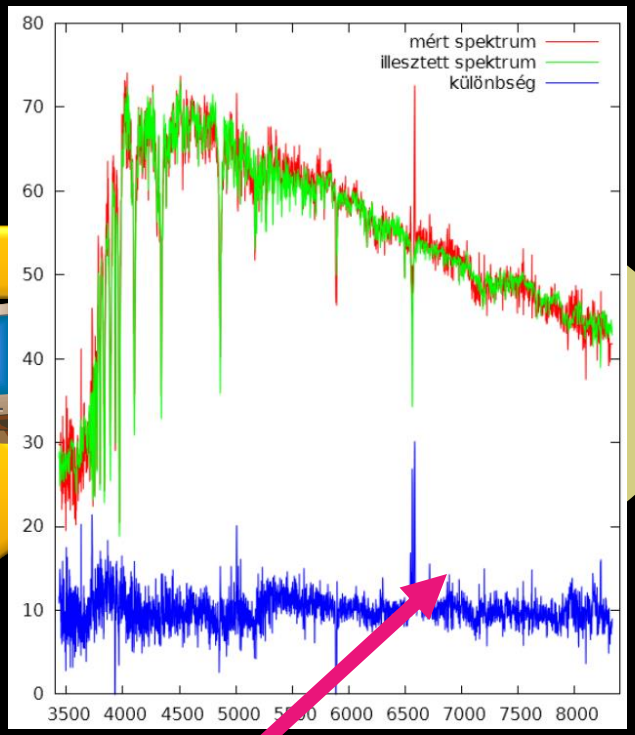


G Rácz, szakdolgozat ELTE (2014)



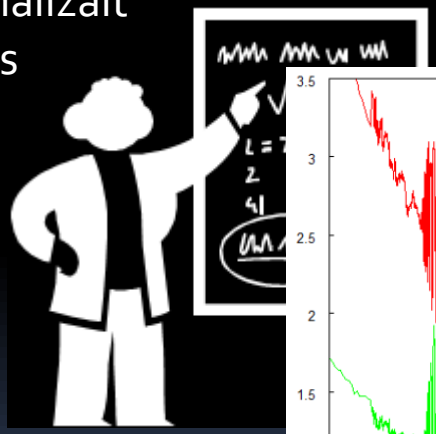
észlelés

kísérlet

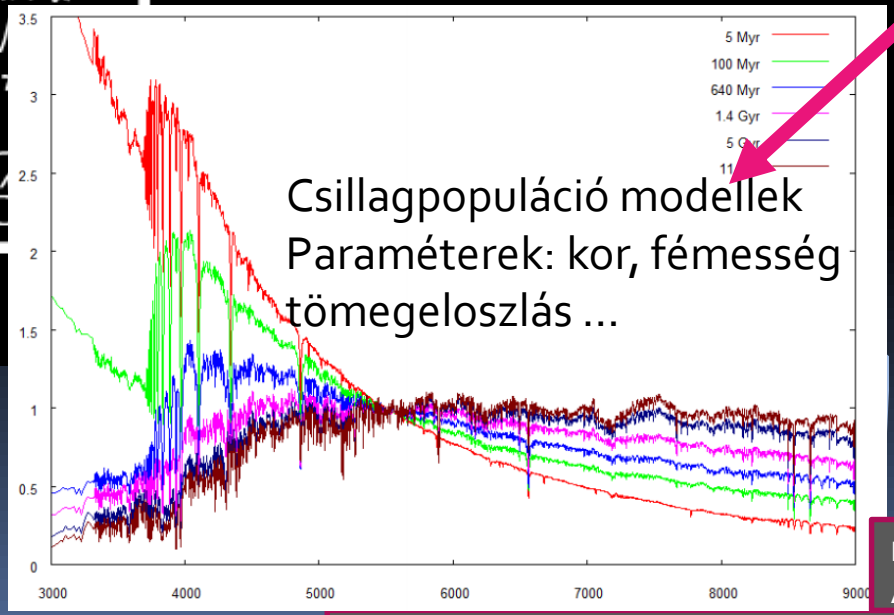


valóság

formalizált leírás



ellenőrzés



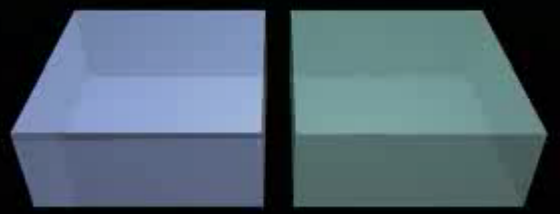
Csillagpopuláció modellek
Paraméterek: kor, fémesség
tömegeloszlás …

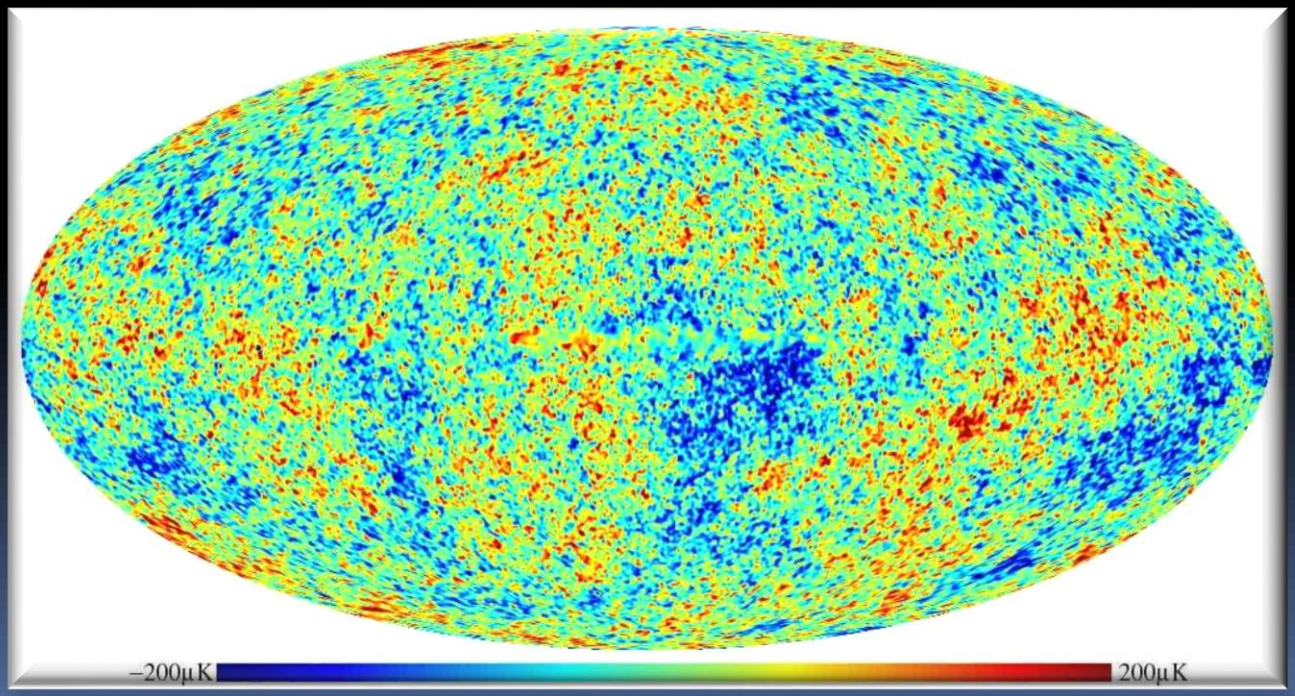
virtuális valóság

L Dobos, I Csabai, CW Yip, T Budavári, V Wild, AS Szalay; MNRAS 420 (2), 1217-1238(2012)

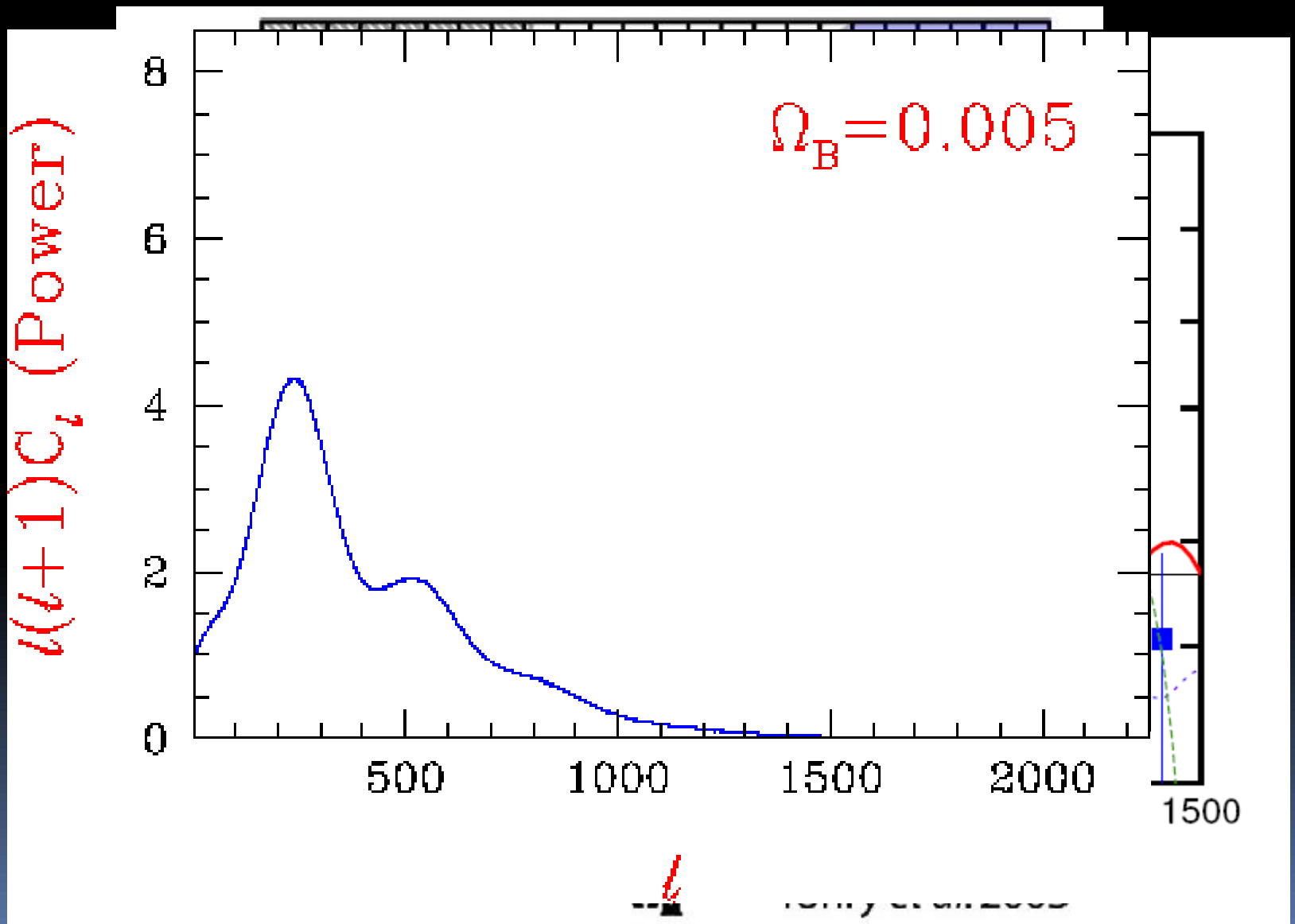
D Ribli, szakdolgozat ELTE (2014)

Az Univerzum paraméterei

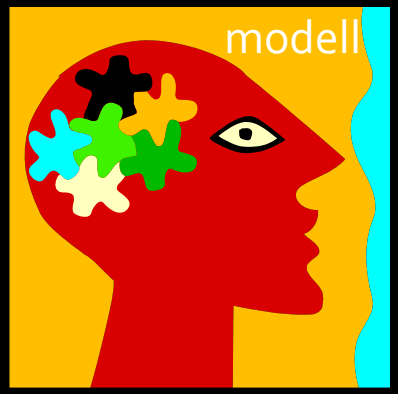


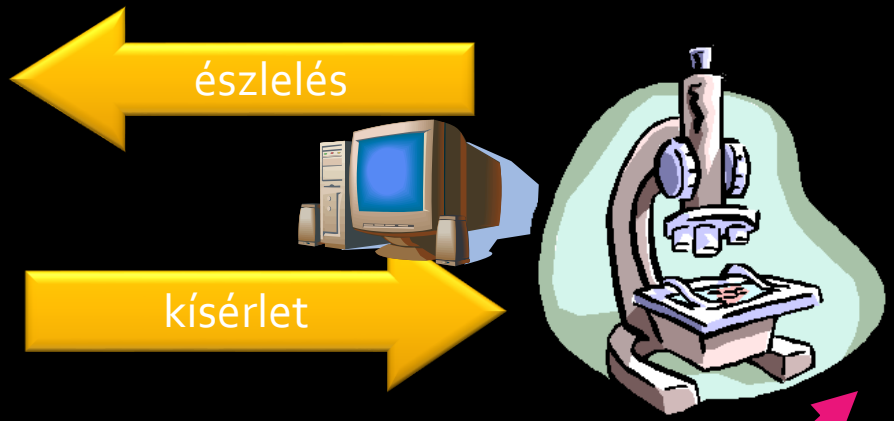


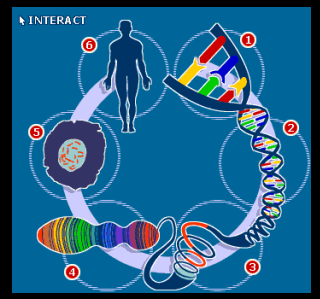
Kozmológiai állandók



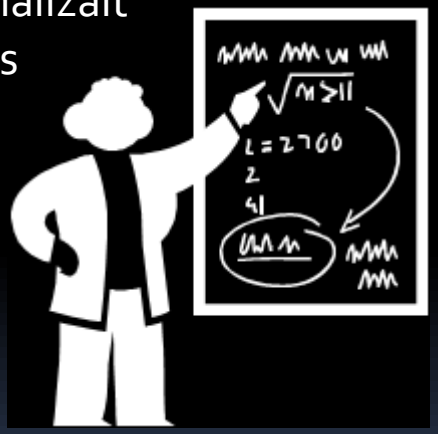
Nem csak a csillagászat: genomika, környezettudományok, társadalomtudományok … Egyre komplexebb kérdések







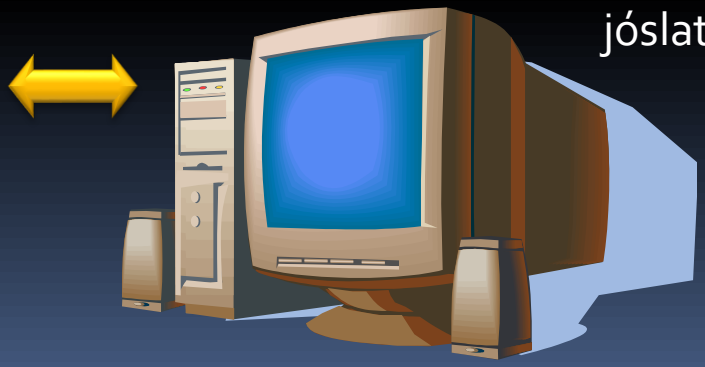
formalizált leírás



ellenőrzés

virtuális valóság

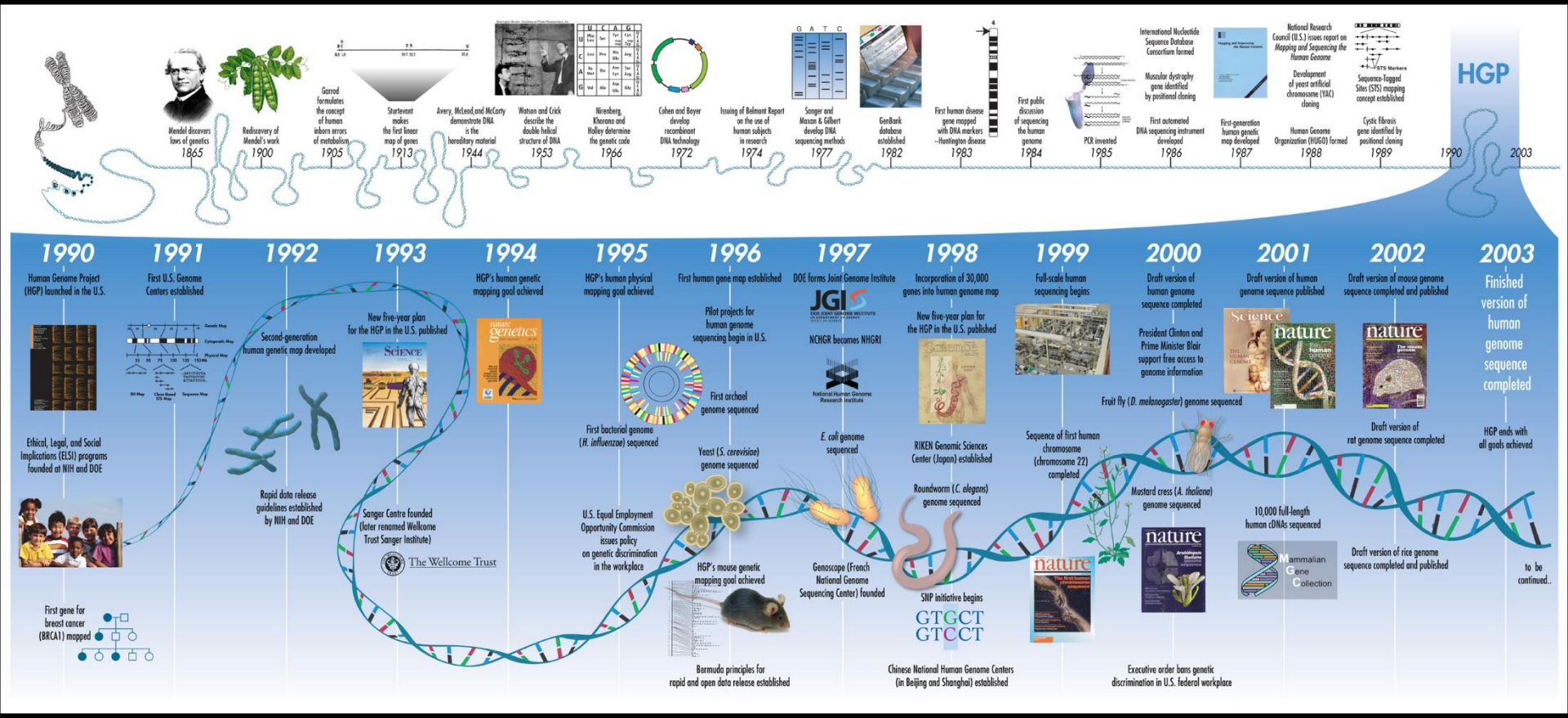
jóslat



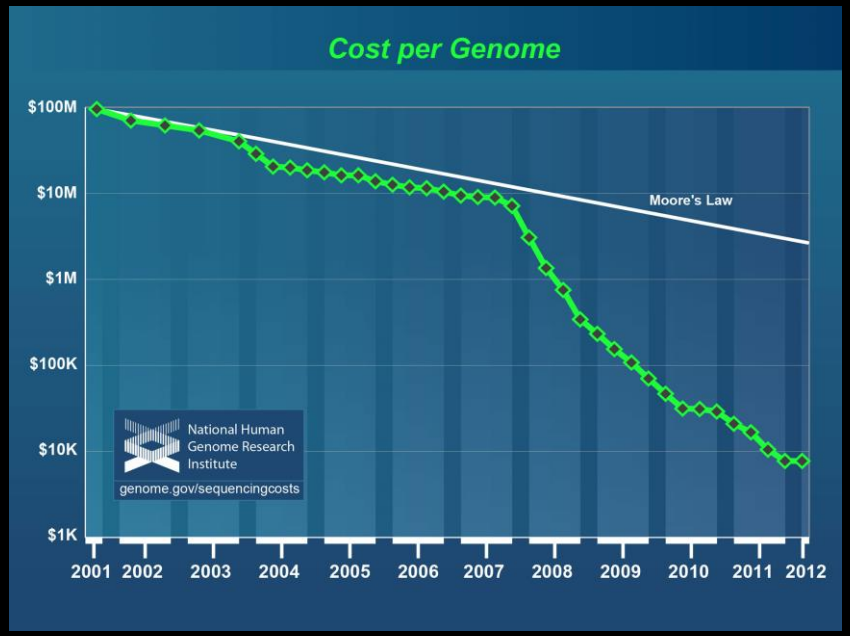
Más tudományágak: hasonló trendek

A GENOM

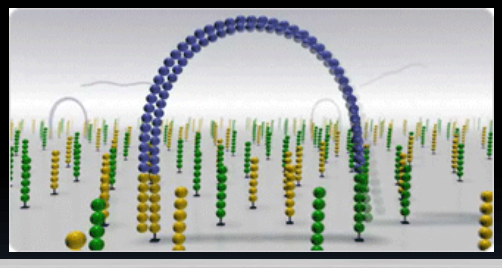
A genetika rövid története

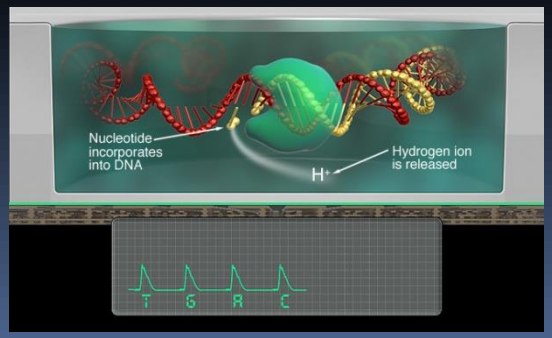


A genomika Moore-törvénye

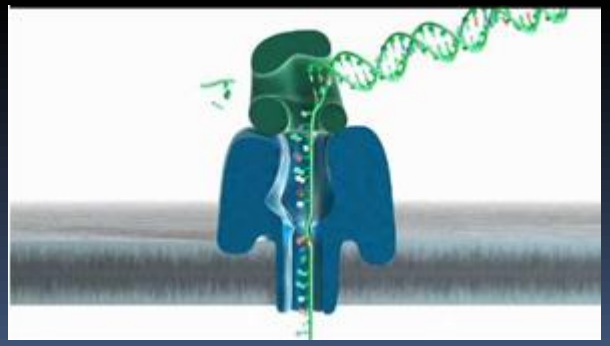


CCD!
- X Prize, 100 genom, 30 nap, $10k - törölve
- Microarray
- Tömegspektrográfia
- Digitális mikroszkópia
- ...





Oxford Nanopore
100Mb,$900



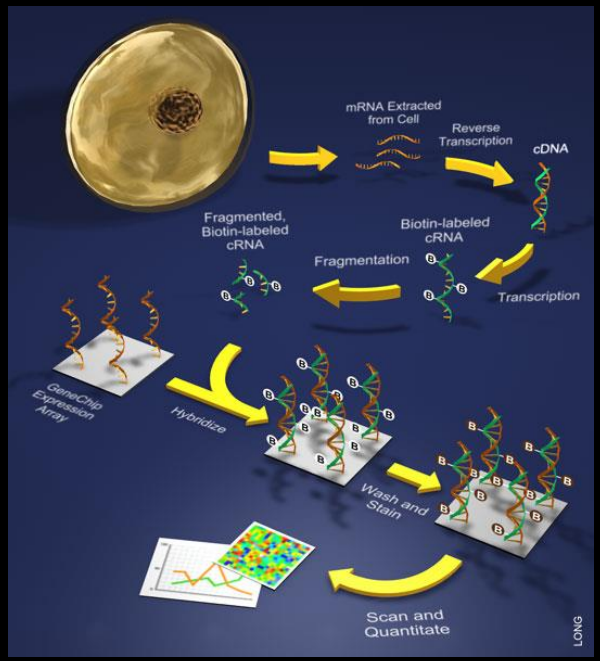
HGP 1990-2003: 2.7 milliárd USD

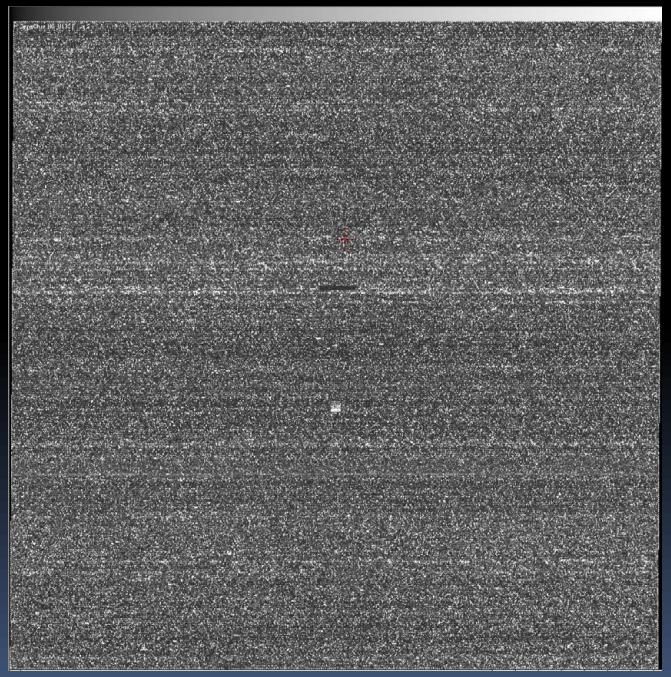
Ma: 10 ezer USD

J Molnár et al. Genes Genomes Genetics, g3.114.013482(2014)

Expresszió microarray

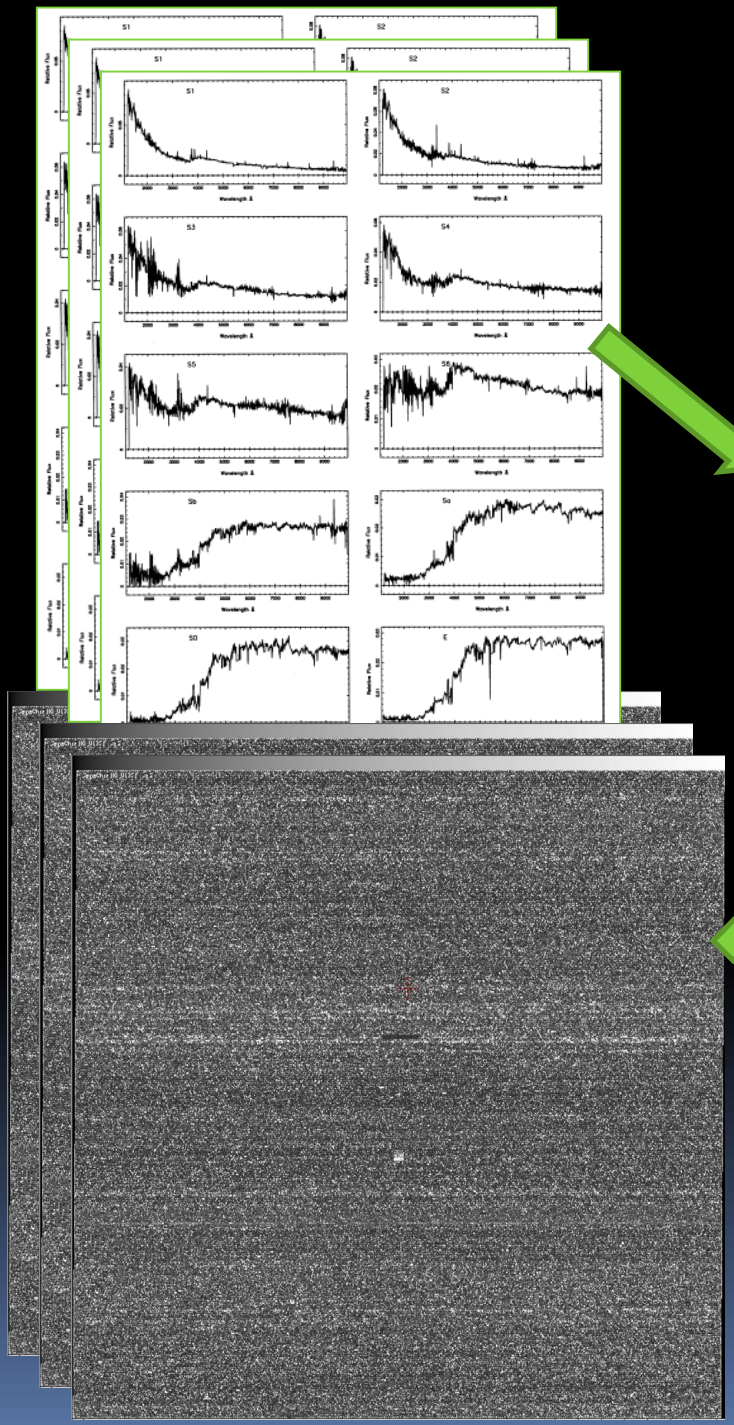
- Affymetrix HG U133 Plus2
  - Nyers adat 67Mpix (fotometria!)
  - 604258 probe
  - 54675 probe set (~gén)
  - 207 minta (colorectalis daganat)





- Hasonló metodika
  - Nagy adatbázisok (saját + publikus)
  - Komputer –intenzív elemő módszerek

Kulcsmarker azonosítás bioinformatikai analízissel
NKTH 3dhisto8 : TECH_08-A1/2-2008-0114, NFÜ-KMR 12-1-2012-0216 (Molnár B., SOTE)



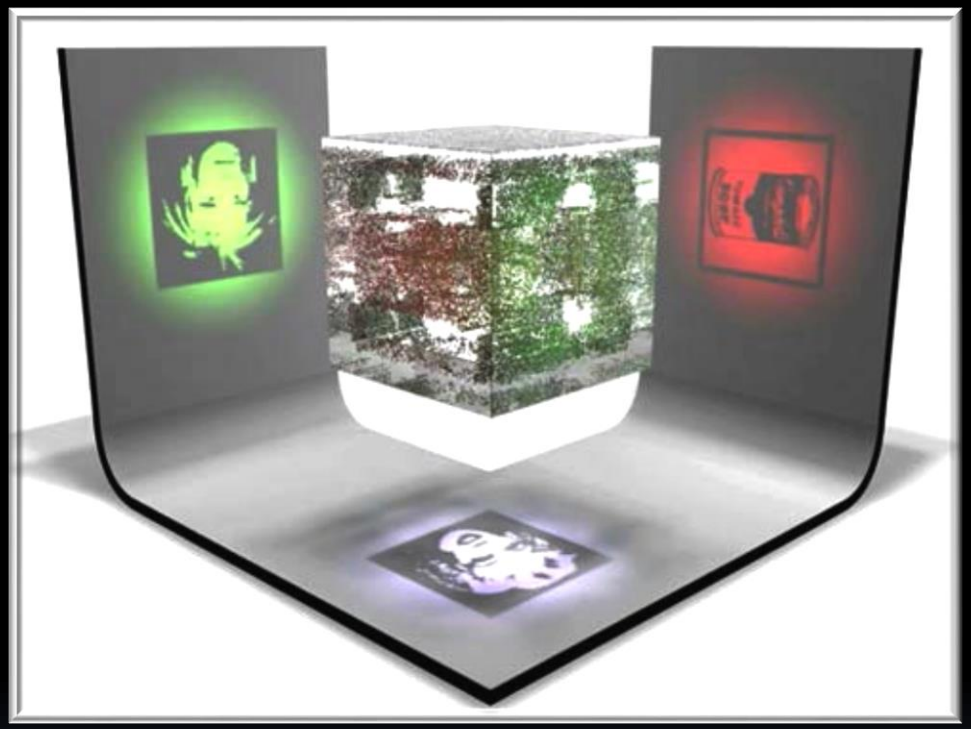
- SDSS spektrumok: 1 millió darab 3000 dimenziós vektor
- Microarray viszgálatunk: 207 darab 54675 dimenziós vektor

Dimenzióredukció, tömörítés

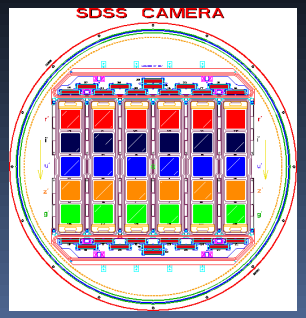
Remény: a háttérben lévő fizikai törvények miatt az adatok nem töltik ki egyenletesen a teret. Alterekre/hiperfelületekre korlátozódnak. Ezért érthetjük meg egyáltalán a világot!

Dimenzió redukció - PCA

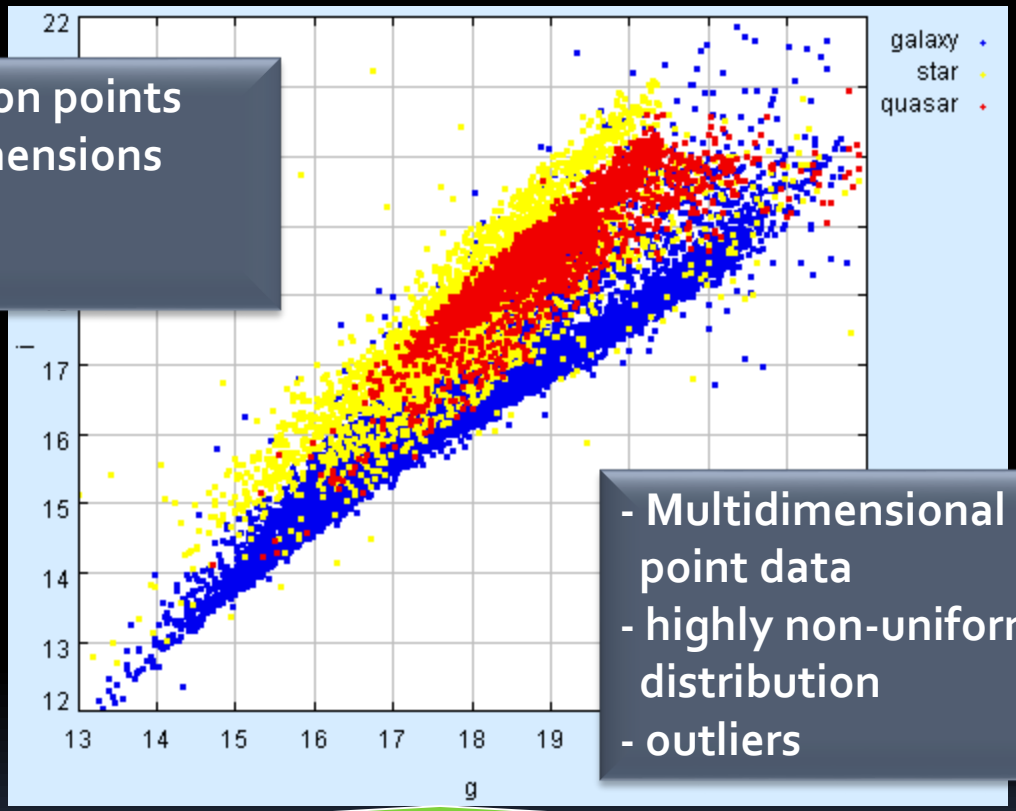
Forgatás 54675D-ban – "optimális" vetítés 2D-ra



„Természetes vetületek”

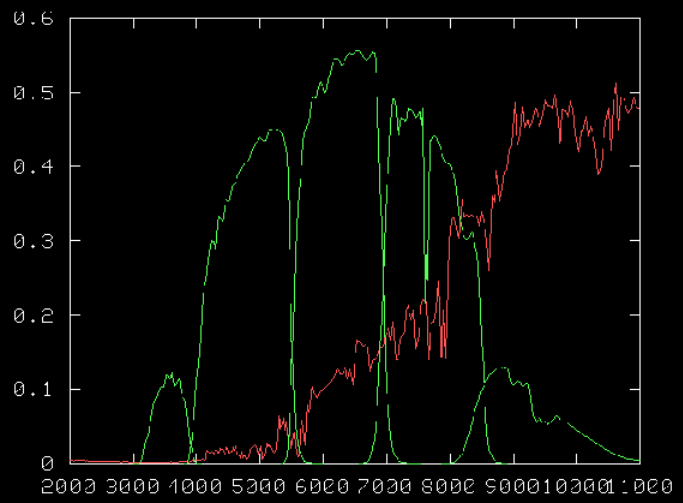


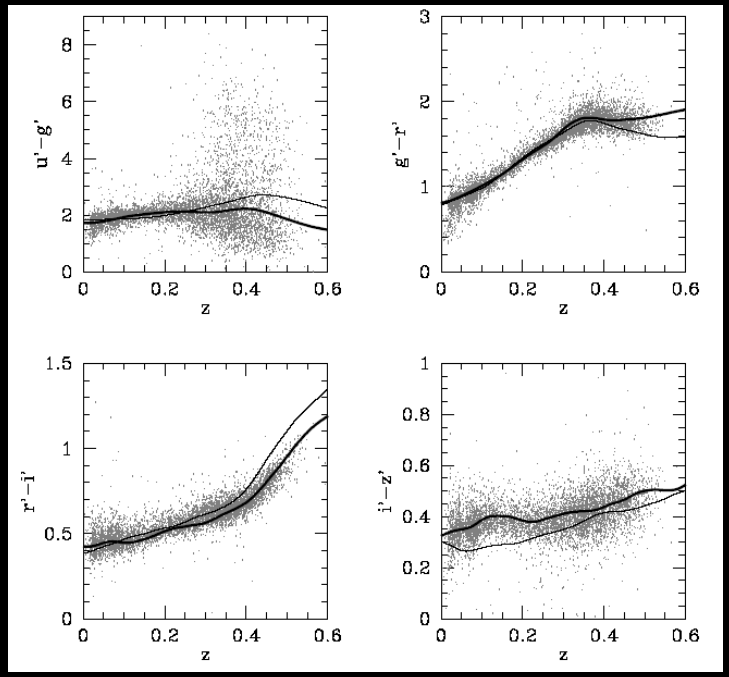
270 million points
in 5+ dimensions
+images
+spectra



- Multidimensional point data
- highly non-uniform distribution
- outliers

u g r i z



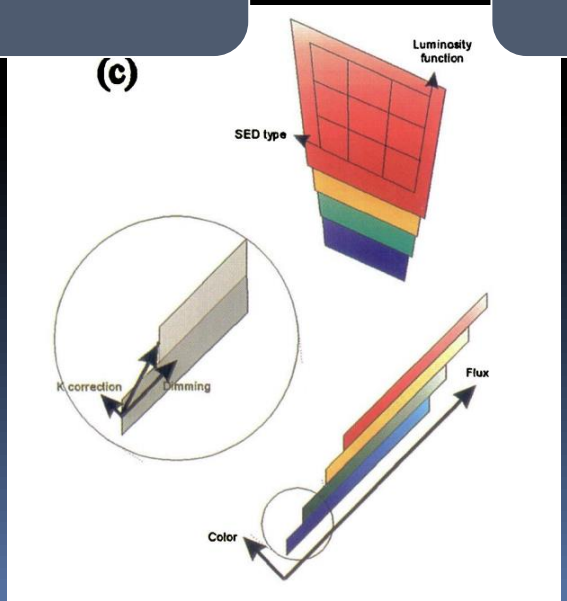


GALAXIS SPEKTRUM

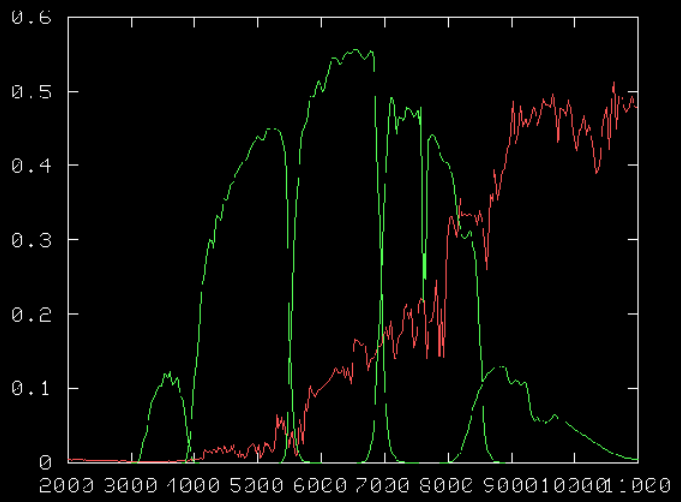
VÖRÖSELTOLÓDÁS

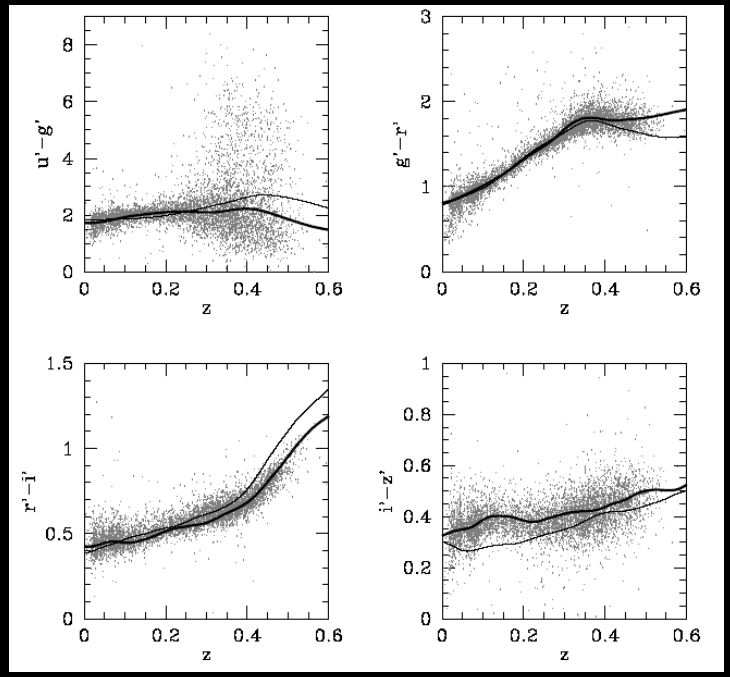
SZÍNSZŰRŐK

MAGNITÚDÓK, SZÍNEK



Fotometrikus vöröseltolódás
Komplex inverz probléma
Mesterséges neuronhálózatok /
Adatbányászati technikák



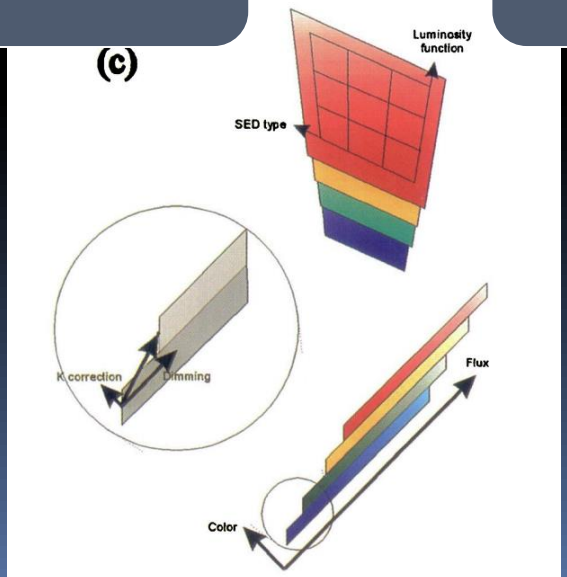


GALAXIS SPEKTRUM

VÖRÖSELTOLÓDÁS

MAGNITÚDÓK, SZÍNEK

SZÍNSZŰRŐK



AJ Connolly, RG Kron, I Csabai, DC Koo, JA Munn, AS Szalay; Astr. J, 110, 2655 (1995)

I Csabai, AJ Connolly, AS Szalay, T Budavári Astr. J. 119 (1), 69(2000)

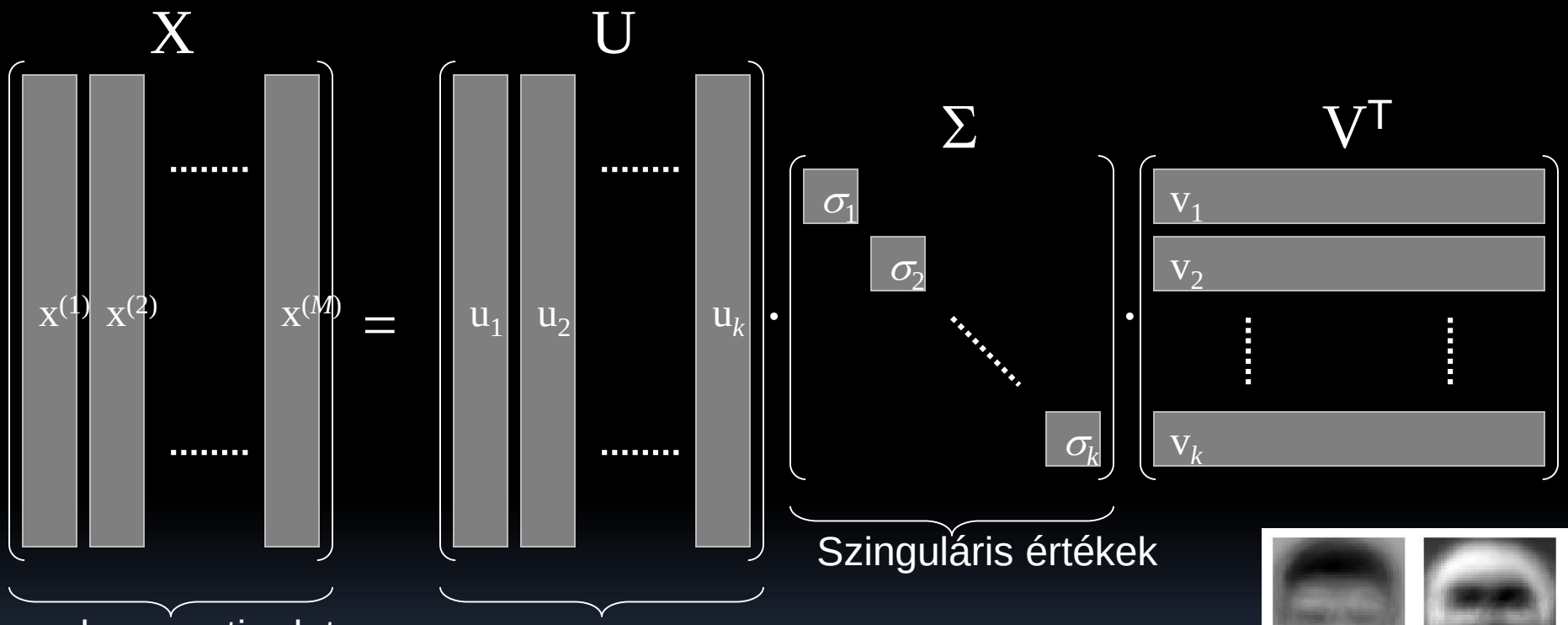
I Csabai + SDSS collab. Astr. J. 125 (2), 580 (2003); SDSS DR1-DR12 (2002-2015)

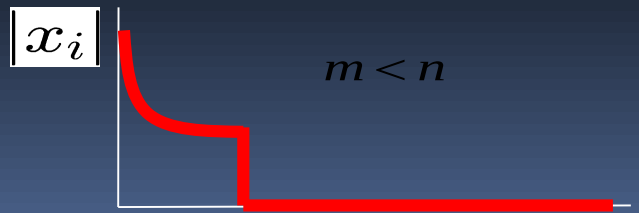
Csillagok metallicitásának becslése:

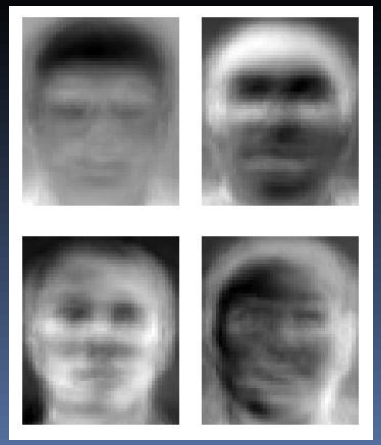
G Kerekes, I Csabai, L Dobos, M Trencséni Astr. N. 334 (9), 1012-1015(2013)

„Okos” vetítés: PCA - SVD

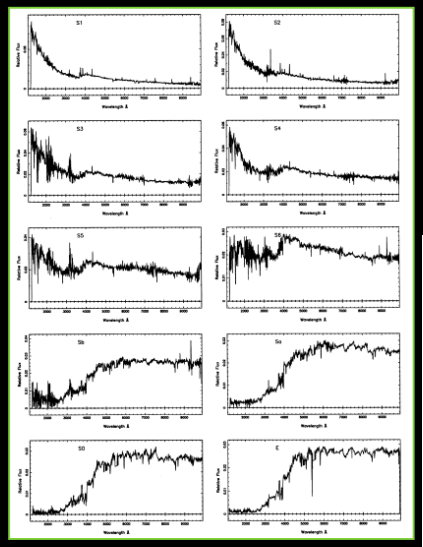
$$X = U\Sigma V^T$$

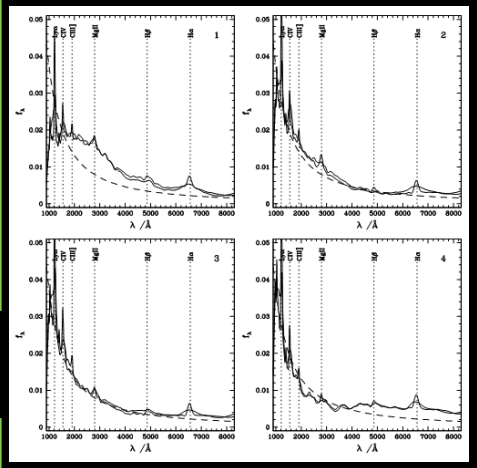


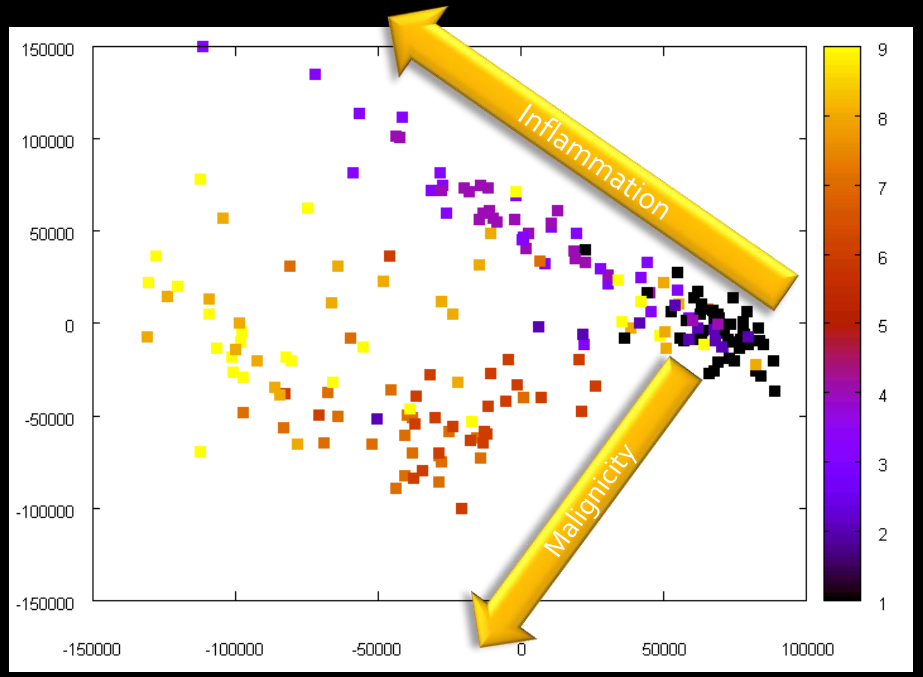


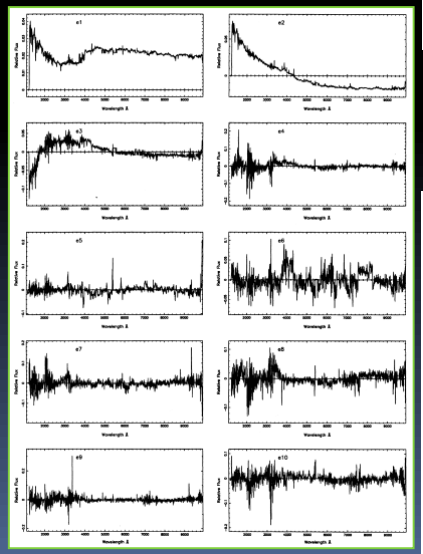


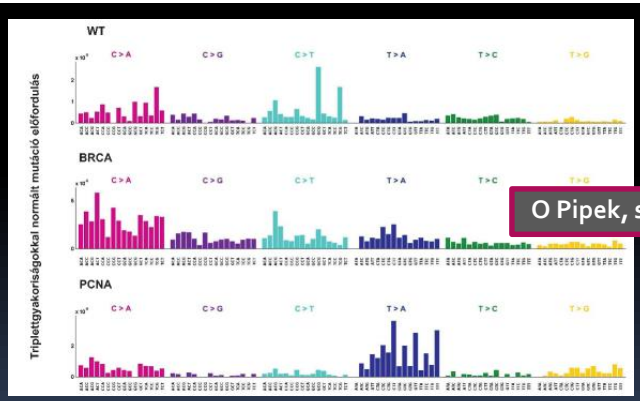
Dimenzió redukció: alkalmazások



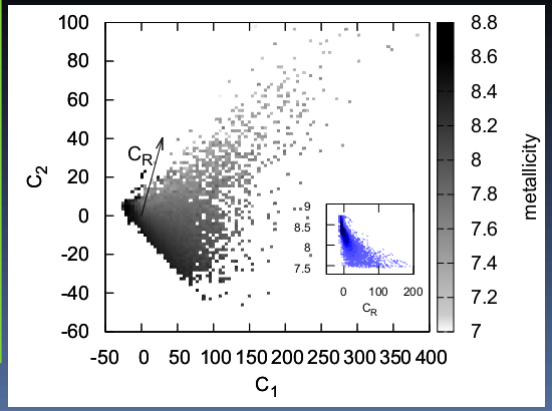








O Pipek, szakdolgozat ELTE (2014)



T Budavári, I Csabai + SDSS collab.; Astr. J. 122 (3) 1163(2001)

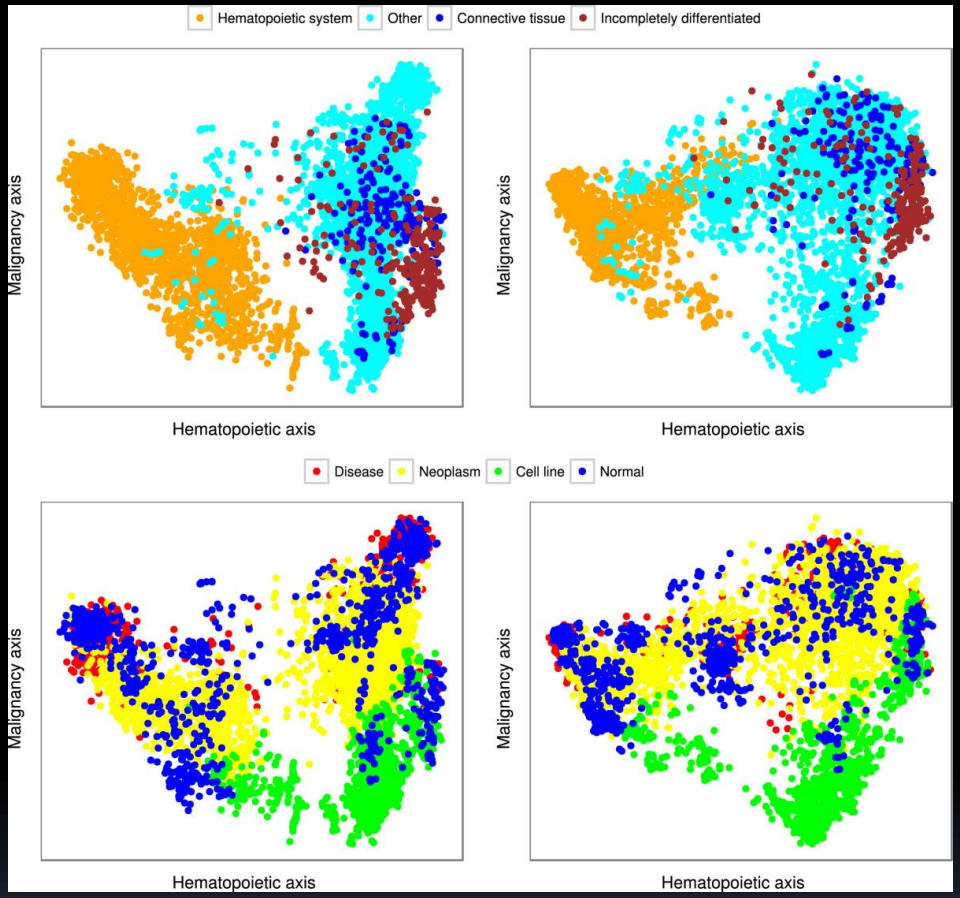
I Csabai, AJ Connolly, AS Szalay, T Budavári; Astr. J. 119 (1), 69 (2000)

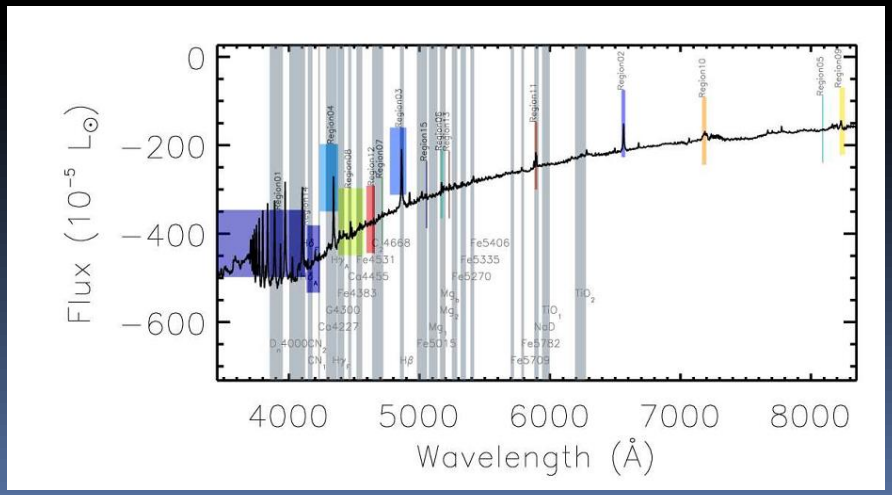
Z Győry, AS Szalay, T Budavári, I Csabai, S Charlot; Astron. J. 141 (4) 133 (2011)

S Spisák, A Kalmár, O Galamb, B Wichmann, F Sipos, B Péterfia, I Csabai, I Kovalszky, S Semsey, Z Tulassay, B Molnár; PloS one 7 (10), e46215(2012)

Dimenzió redukció - PCA

- Fejlesztések
  - Kilógó adatpontok: robusztusság
  - Sok adat: "streaming" / DB
  - Ritka adat-mátrixok (compressive sensing)
  - CUR dekompozíció
  - Nem negatív mátrix faktorizáció
  - Gráf főkomponensek
  - ? Szöveg PCA, Genom PCA





A Bodor, I Csabai, MW Mahoney, N Solymosi; BMC bioinformatics 13 (1), 103 (2012)

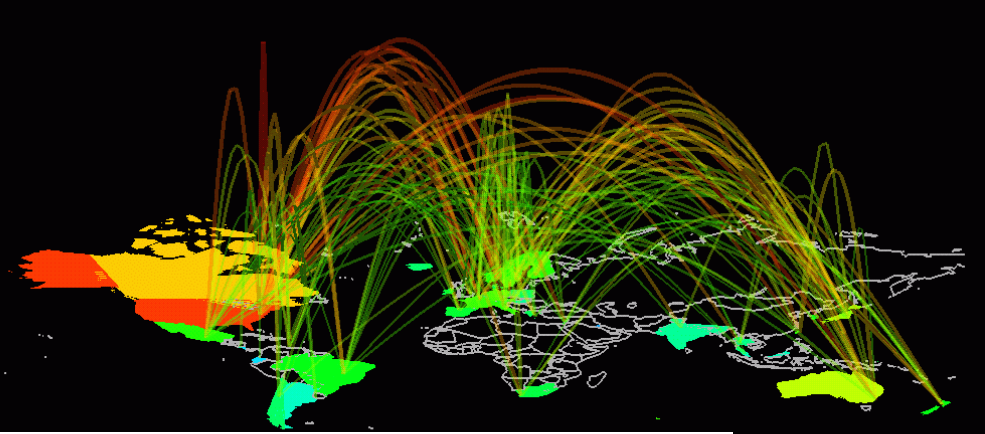
CW Yip, MW Mahoney, AS Szalay, I Csabai, T Budavári, RFG Wyse, L Dobos Astr. J. 147 (5), 110 (2014)

Ember alkotta komplex rendszerek

KOMMUNIKÁCIÓS, SZOCIÁLIS ÉS PÉNZÜGYI HÁLÓZATOK

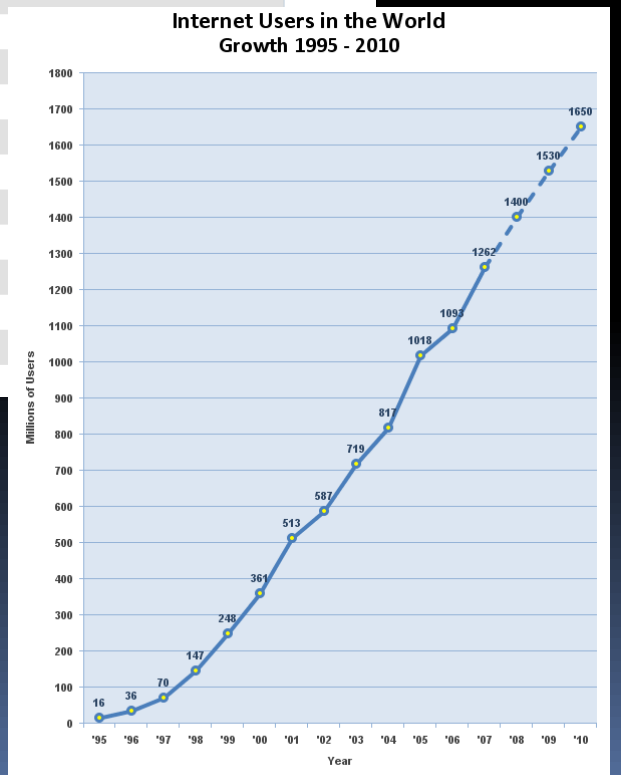
Internet

- Ember alkotta, de nincs meg a “tervrajz”
- “Csillagászati” számú komplex nem-lineárisan kölcsönható elem
- Természettudományos módszerek kellenek
  - Észlelés/kísérlet
  - Modell
  - Jobbat tervezni
- Jövő internet: self-aware, self-managing, self-healing …



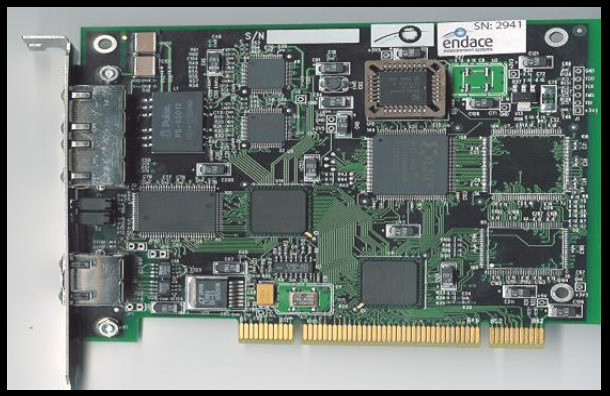
Countries	Total IPs	Last Updated
240	3,030,462,896	2009-07-31 15:20:58

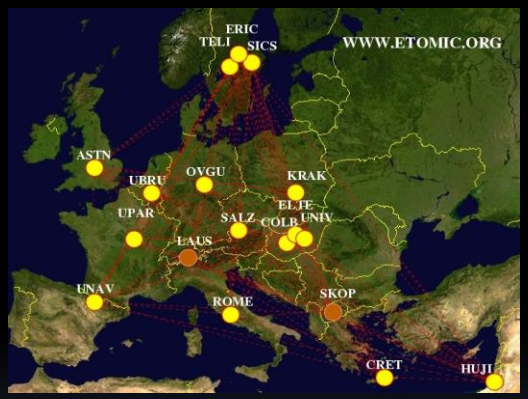
Total IPs	Country
1,478,842,181	UNITED STATES
234,262,365	UNITED KINGDOM
207,440,422	CHINA
176,463,582	JAPAN
108,273,367	GERMANY
81,241,704	FRANCE
77,261,386	CANADA
72,907,648	KOREA, REPUBLIC OF
43,472,356	ITALY
42,894,254	NETHERLANDS
38,451,977	AUSTRALIA
31,906,592	BRAZIL
26,969,719	RUSSIAN FEDERATION



Network Measurement VO ETOMIC/Sonoma/Spotter

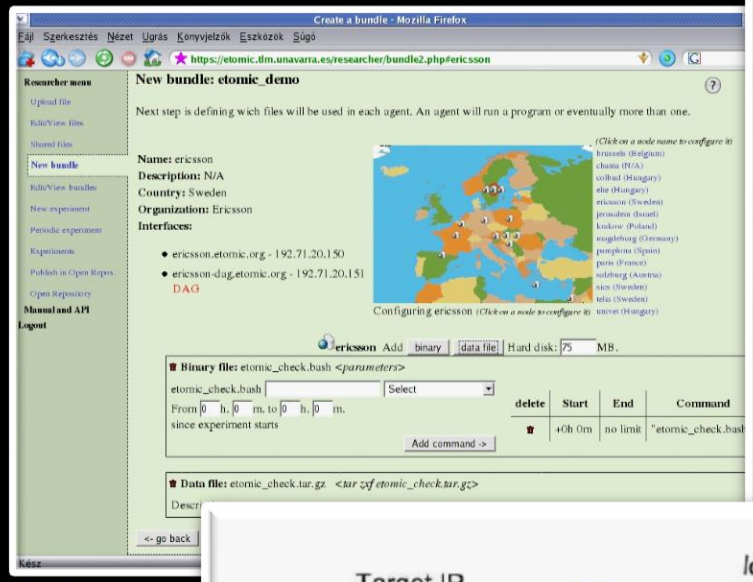
- GPS szinkronizáció
- Speciális precíz időmérés (~60 ns , Endace DAG 3.6 GE card / ARGOS FPGA measurement card)
- Mérésvezérlő rendszer
- Központi adatbázis
- Nemzetközi nagykollaborációk: Evergrow, Planetlab, Onelab
- Ipari partnerek

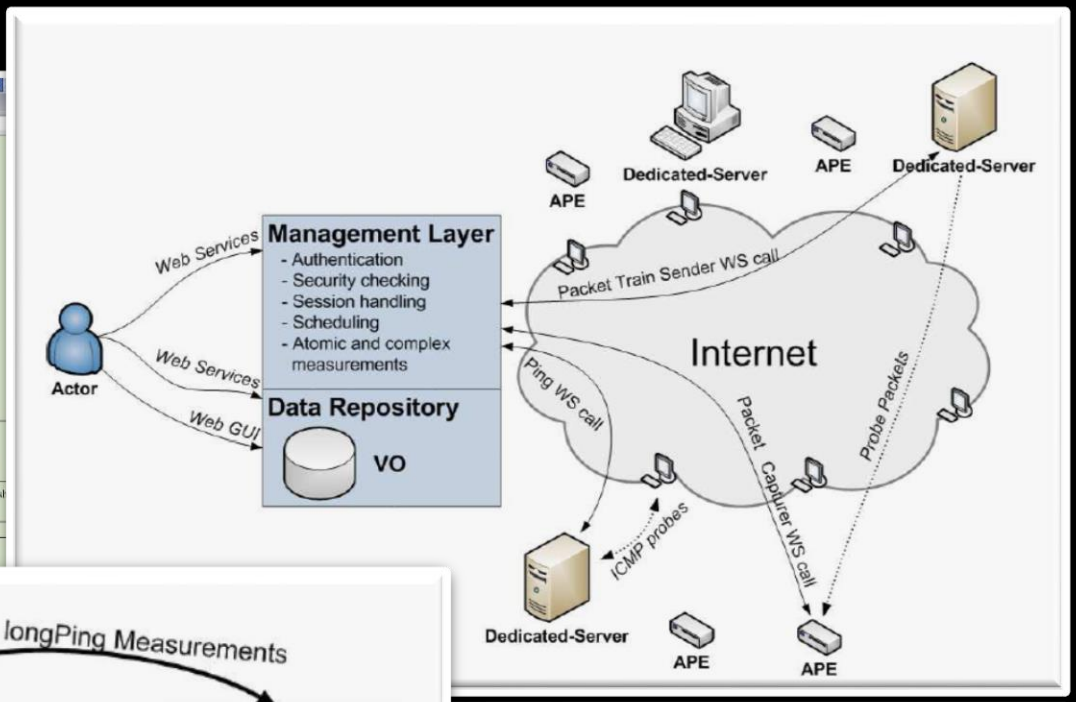


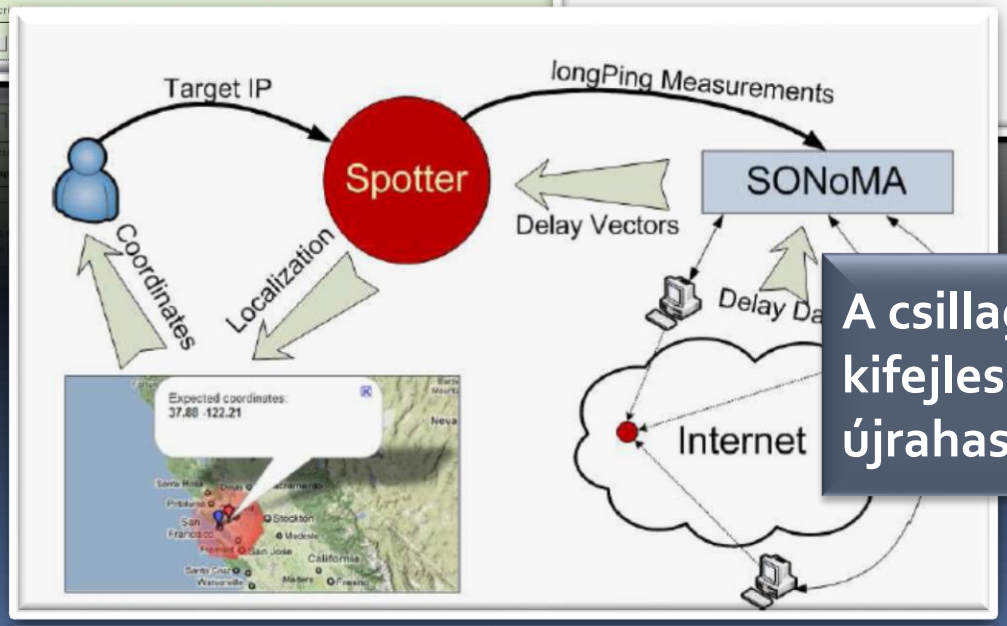


- Spotter: A model based active geolocation service; S Laki, P Mátray, P Hága, T Sebők, I Csabai, G Vattay; INFOCOM, 2011 Proceedings IEEE, 3173-3181 (2011)
- Building a prototype for network measurement virtual observatory; P Matray, I Csabai, P Haga, J Steger, L Dobos, G Vattay;Proc. ACM workshop on Mining network data, 23-28 (2007)
- The European Traffic Observatory Measurement Infrastructure (ETOMIC): A Testbed for Universal Active and Passive Measurements.;D Morato, E Magana, M Izal, J Aracil, FJ Naranjo, P Astiz, U Alonso, I Csabai, P Hága, G Simon, J Stéger, G Vattay; TRIDENTCOM, 283-289 (2005)
- A model based approach for improving router geolocation; S Laki, P Mátray, P Hága, I Csabai, G Vattay; Computer Networks 54 (9), 1490-1501 (2010)

ETOMIC/SONOMA/SPOTTER/NMVO

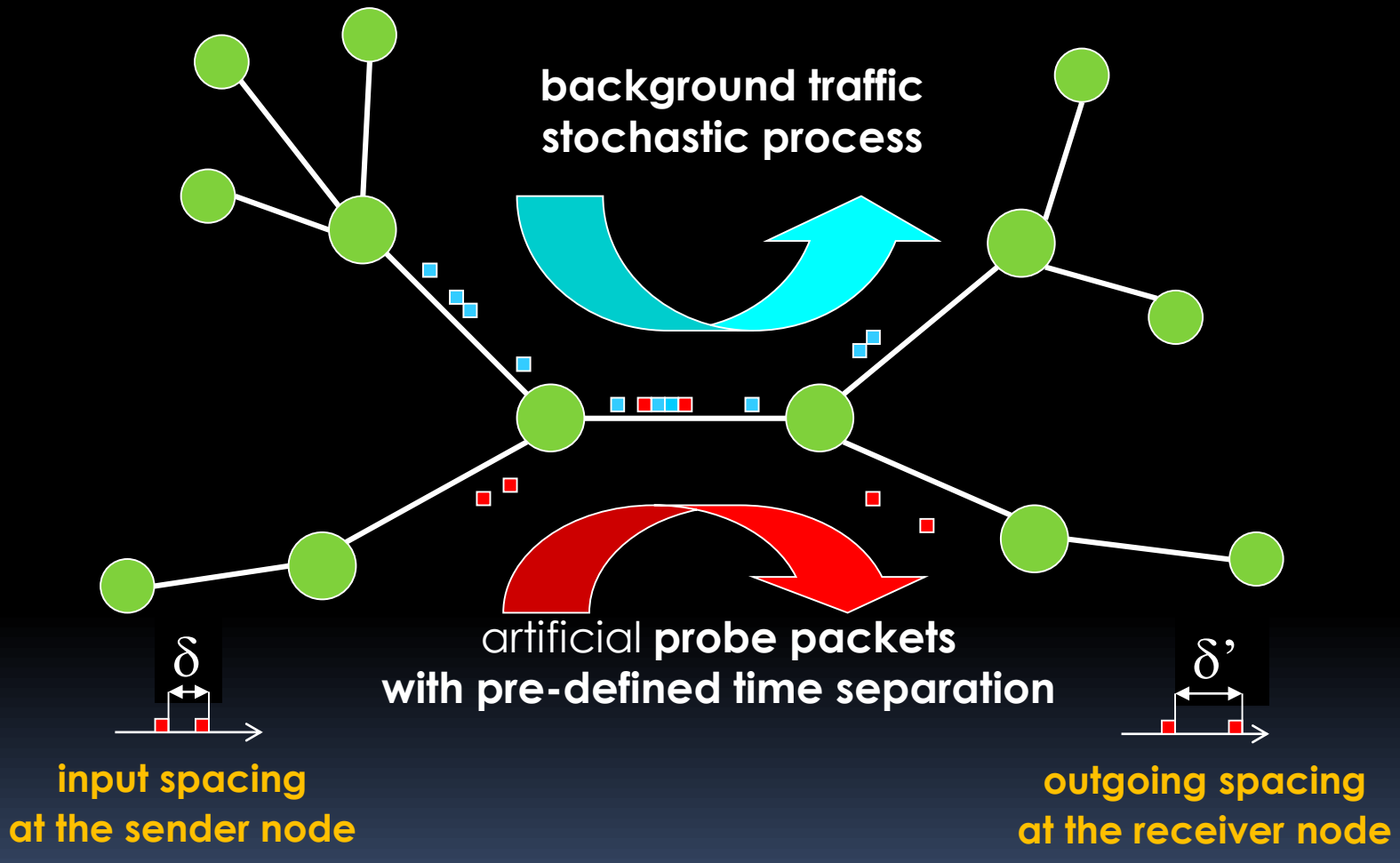




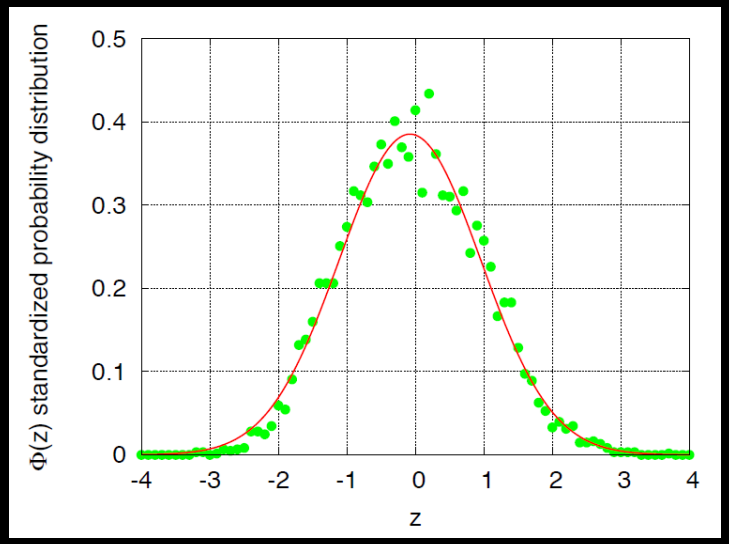


A csillagászati nagyfelméréshez kifejlesztett eszközök újrahasznosítása

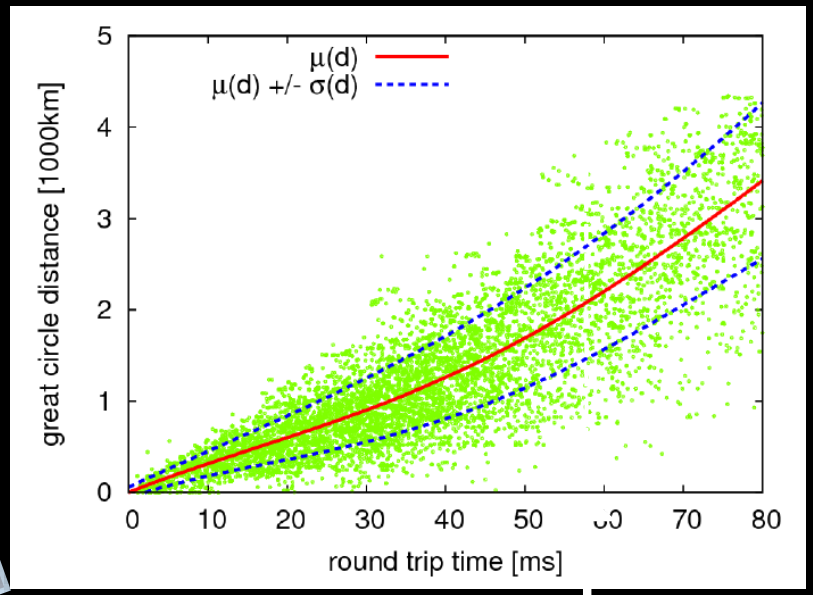
Hálózat tomográfia – végpont-végpont mérések



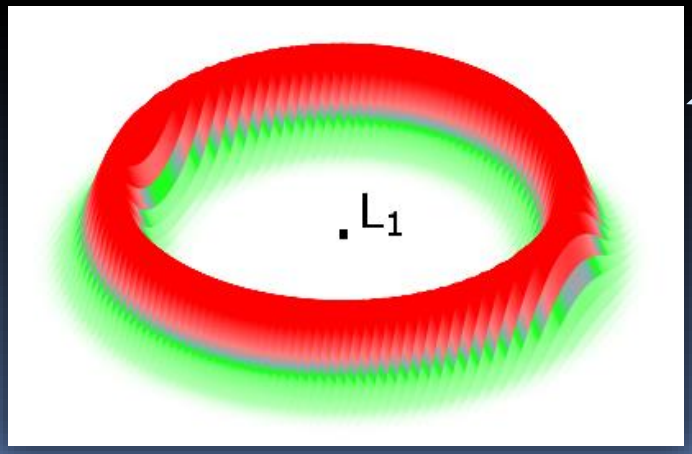
Geolokalizáció

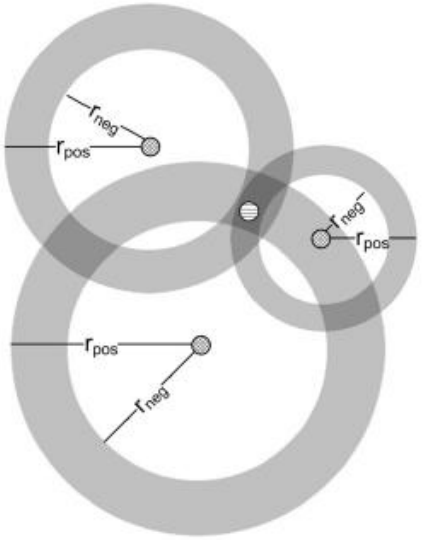


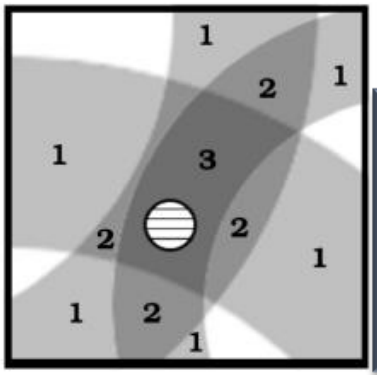
Csomagkésleltetési idő eloszlás



Késleltetési idő – távolság modell

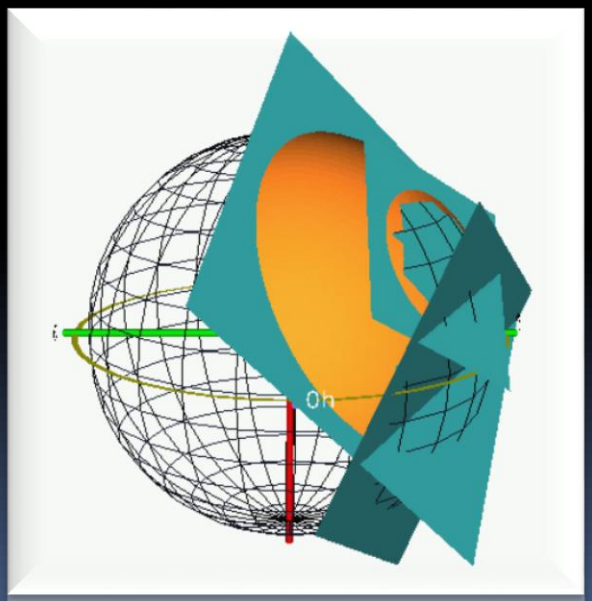


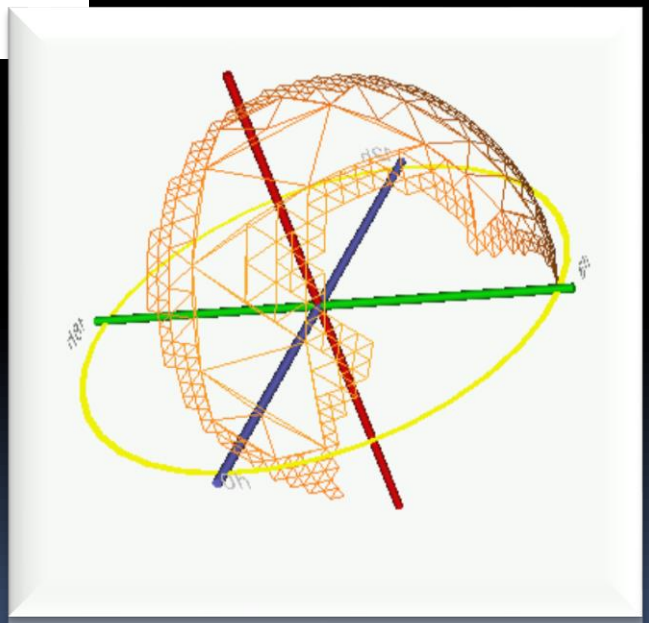




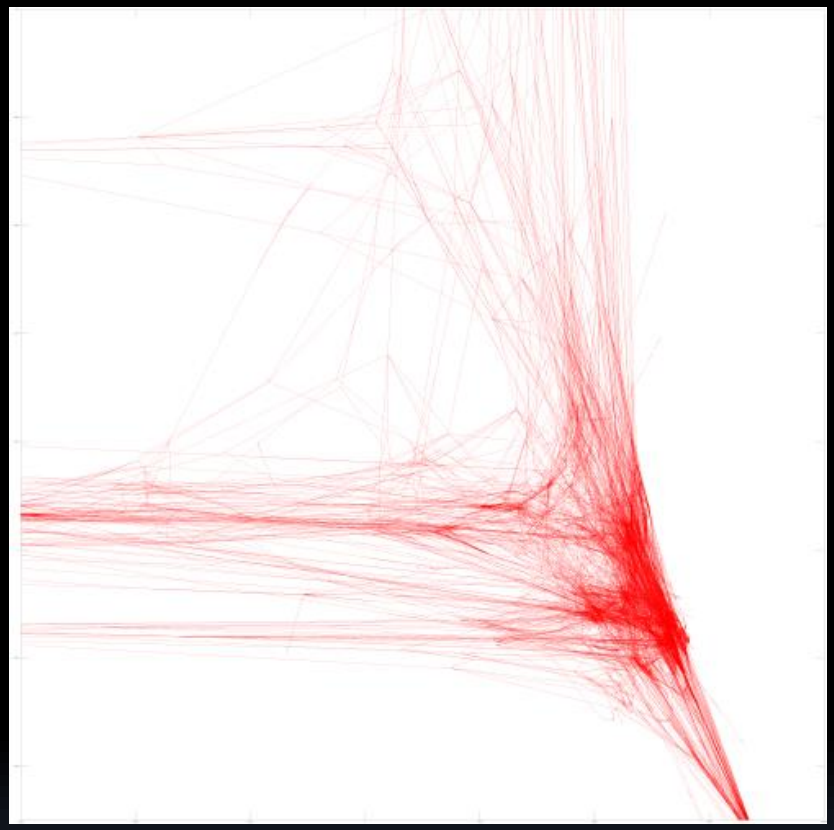
Target node
Landmark node

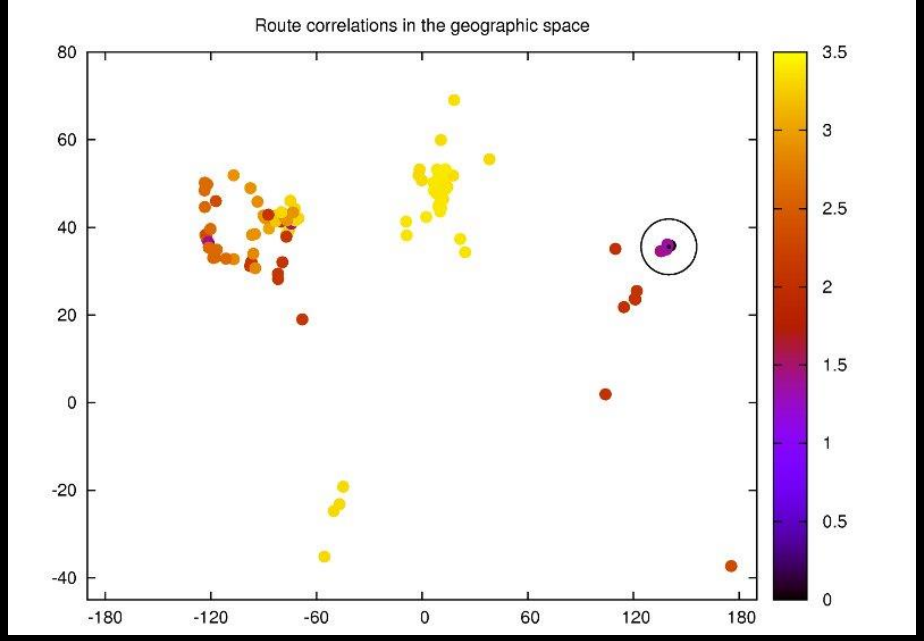
Csillagászati indexelés újrahasznosítás:
HTM index library + SQL Server integráció
Gysors gömbi poliéder manipulációk:
gyorsabb geolokalizáció

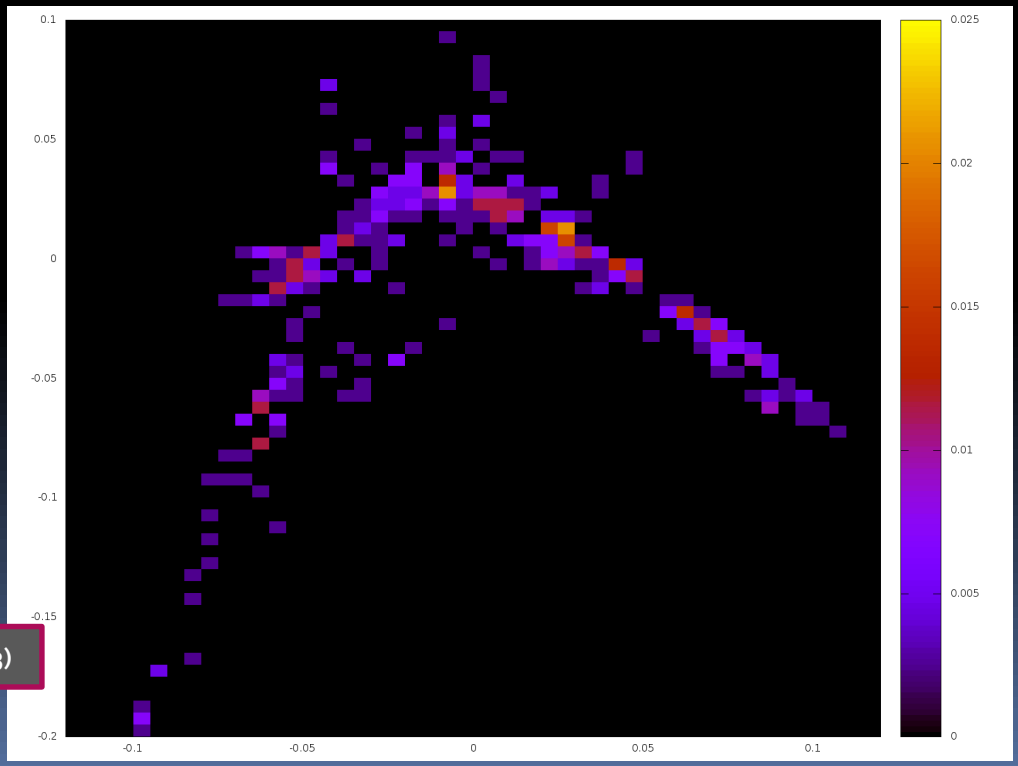




Hálózat PCA

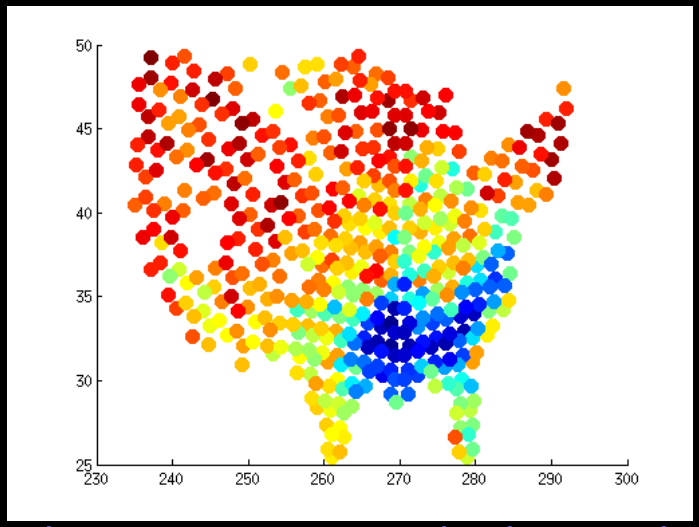




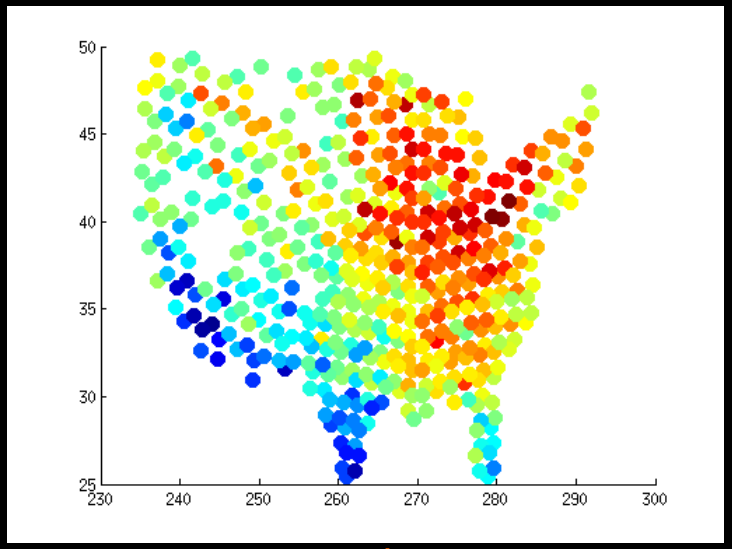


D Kondor, P Mátray, I Csabai, G Vattay; Physica A 392 (18), 4160-4171 (2013)

Szociális hálózatok: TwitterDB



goodmorning, wit, everybody,somebody
lil, jus, hoes, yall, sleepy,tryna,af
imma,smh,yu,kno,bout,gon,hoe
ima,wat,swear,dnt,aint,nobody
nigga,niggas



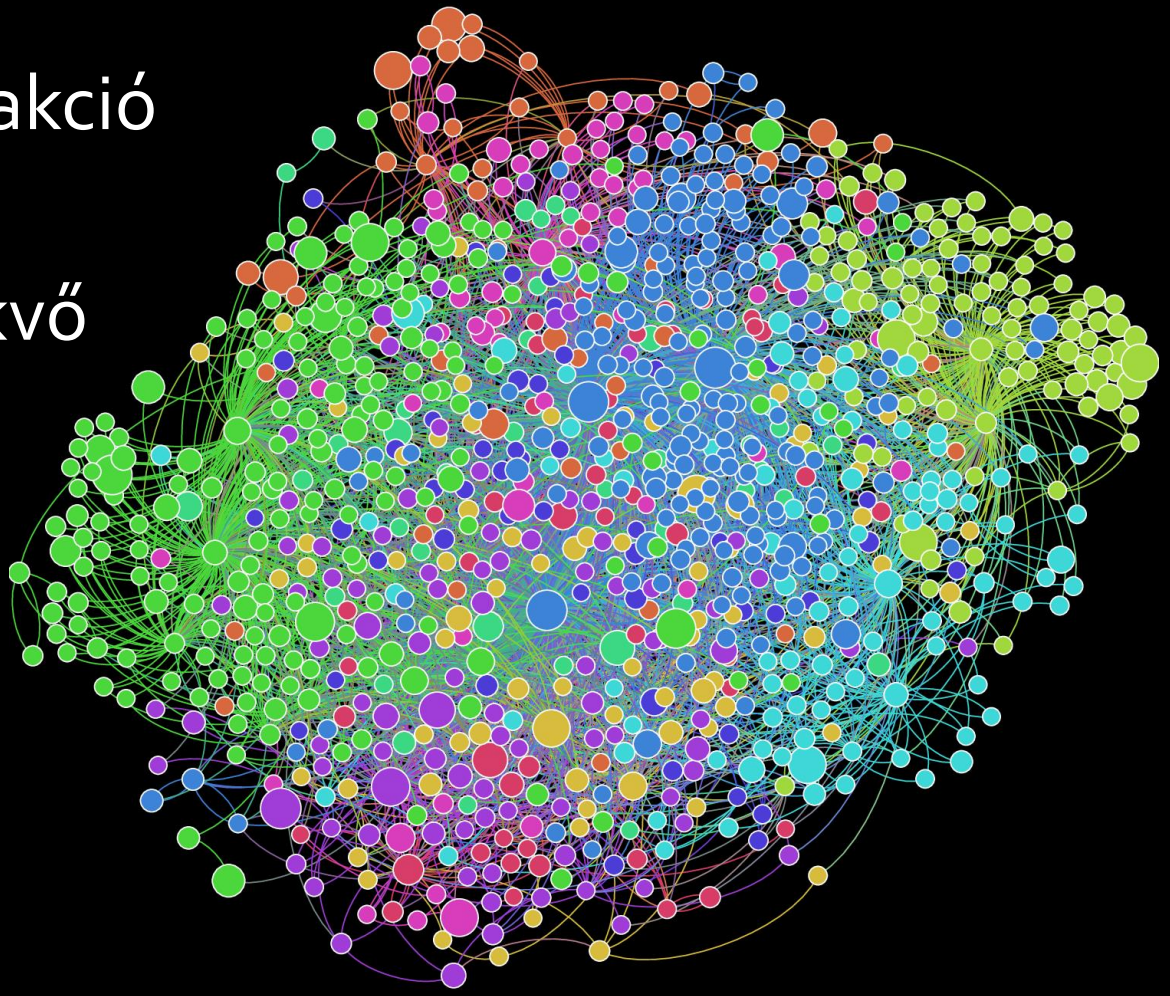
amazing,excited, awesome, nice
idea,favorite,holy,snow,guys
looks,sounds,makes,ever,actually
does,hours,such,into,were,doesn,
little

Efficient classification of billions of points into complex geographic regions using hierarchical triangular mesh; D Kondor, L Dobos, I Csabai, A Bodor, G Vattay, T Budavári, AS Szalay 26th International Conference on Scientific and Statistical Database Management, Aalborg, Denmark, (2014)

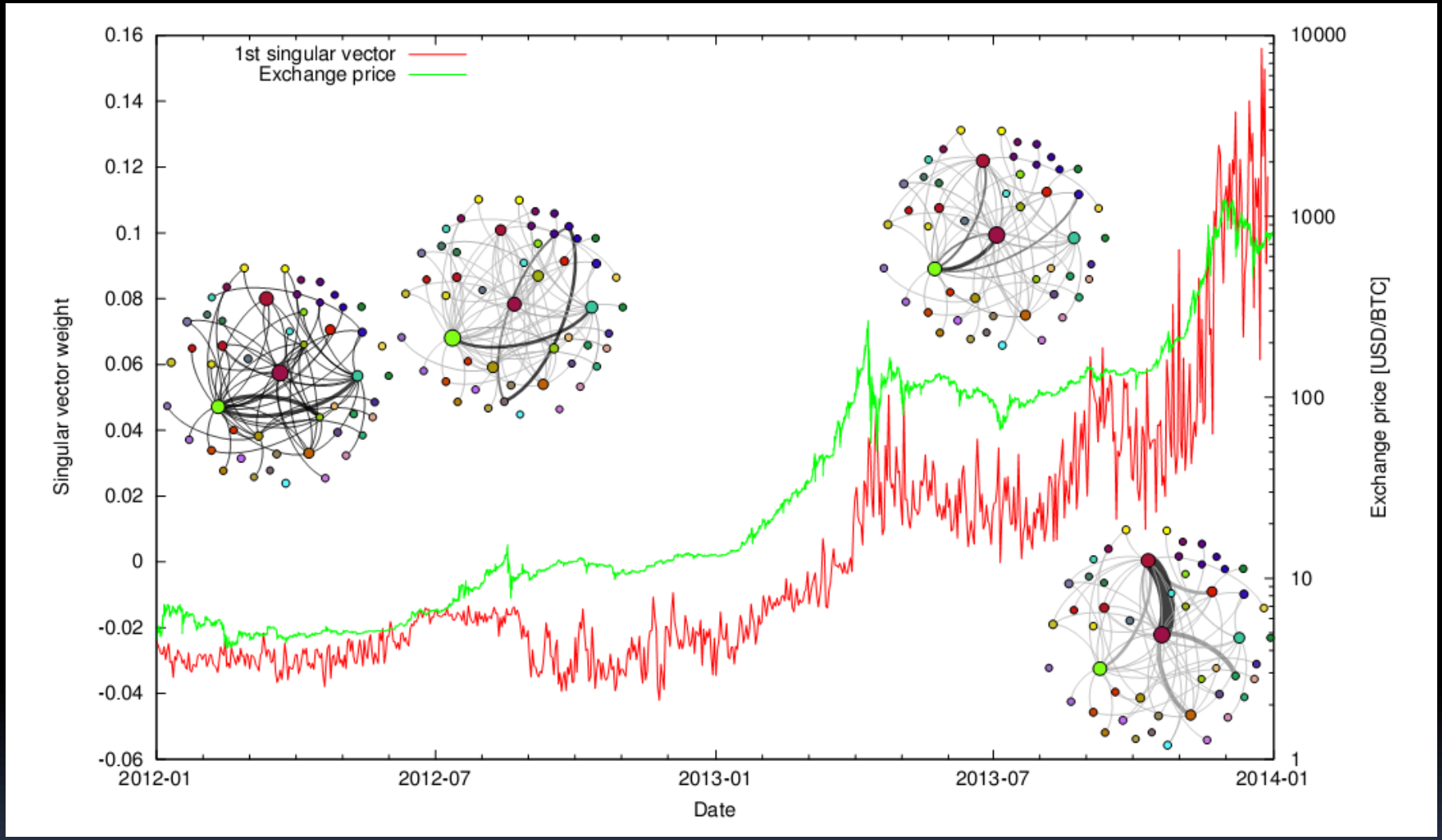
Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages; D Kondor, I Csabai, L Dobos, J Szule, N Barankai, T Hanyecz, T Sebok, Z Kallus, G Vattay; IEEE CogInfoCom) (2013)

Bitcoin pénzügyi hálózat

- Összes (50M) tranzakció ismert
- Dinamikusan növekvő irányított hálózat
- Adatbázis
- Dimenzióredukció



Do the rich get richer? An empirical analysis of the BitCoin transaction network; D Kondor, M Pósfai, I Csabai, G Vattay; PloS one 9 (2), e86197 (2014)



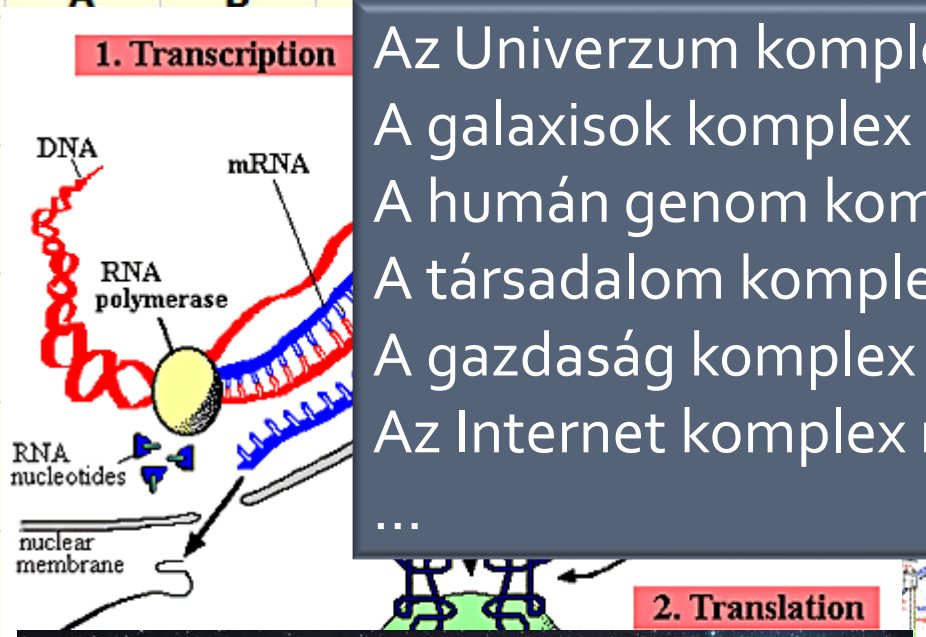
Inferring the interplay of network structure and market effects in Bitcoin; D Kondor, I Csabai, J Szüle, M Pósfai, G Vattay; New Journal of Physics, accepted (2014)

Munkatársak, támogatás

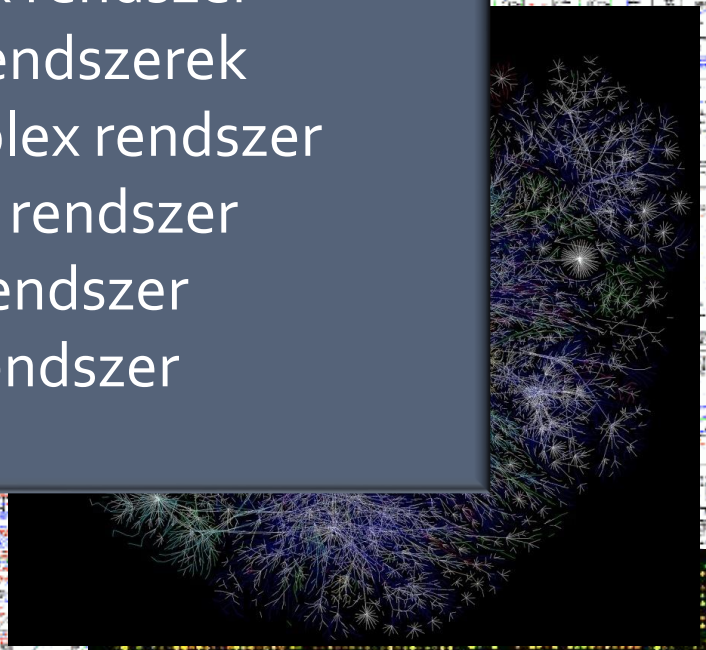
Diákok
PhD hallgatók
Kollégák

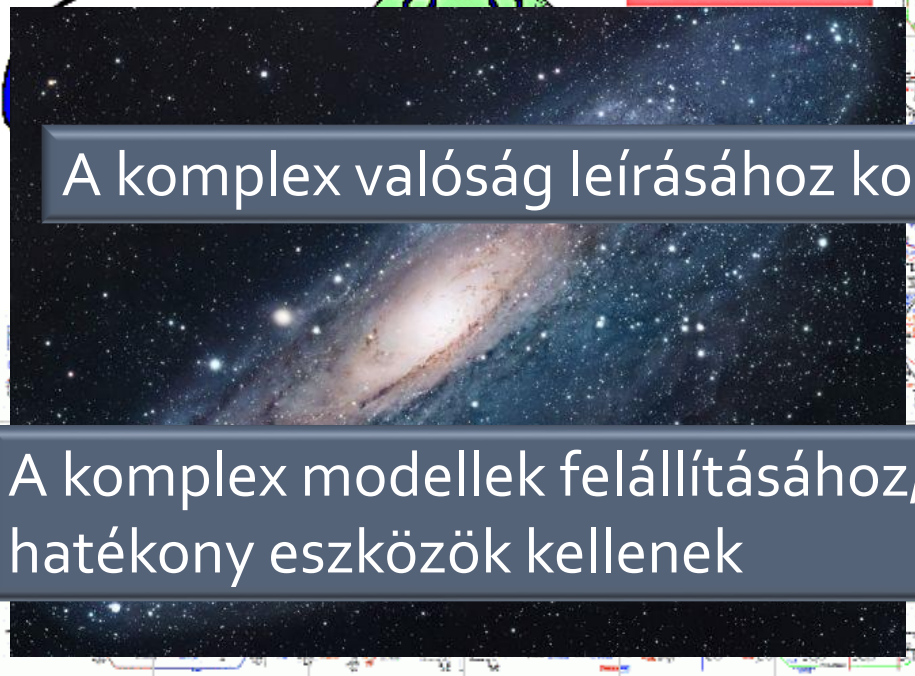
SOTE, MTA TTK, MTA Wigner
Universidad Autonoma de Madrid
Universidad Publica de Navarra
Ericsson Research
Tel Aviv University
Johns Hopkins University
Ericsson, 3Dhistech
Harvard Children's Hosp., Stanford Univ., DTU

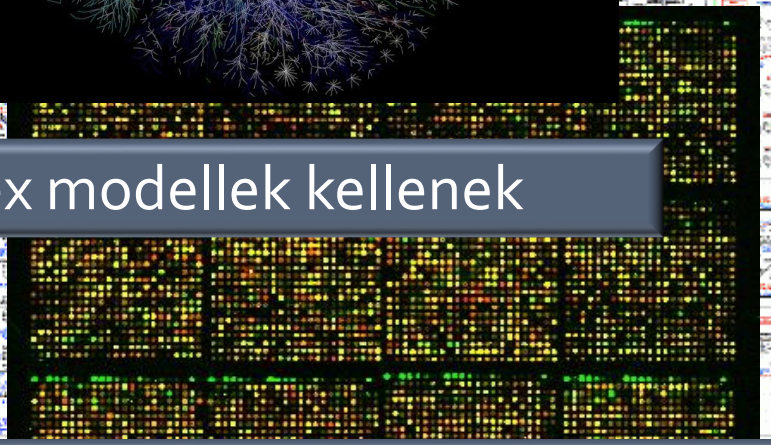
NKTH TECH08:3dhisto8,
KMR_12-1-2012-0216
NAP 2005/ KCKHA005, Polányi
TAMOP: Futurlct
OTKA-103244
OTKA 7779
EU ICT OneLab2 IP #224263
EU FIRE NOVI #257867
EIT KIC



Az Univerzum komplex rendszer
A galaxisok komplex rendszerek
A humán genom komplex rendszer
A társadalom komplex rendszer
A gazdaság komplex rendszer
Az Internet komplex rendszer
…







A komplex valóság leírásához komplex modellek kellenek

A komplex modellek felállításához/validációjához sok-sok adat és hatékony eszközök kellenek “Datascope ”

"Lehetőségek és kihívások"

ITP