Statisztika a csillagászatban

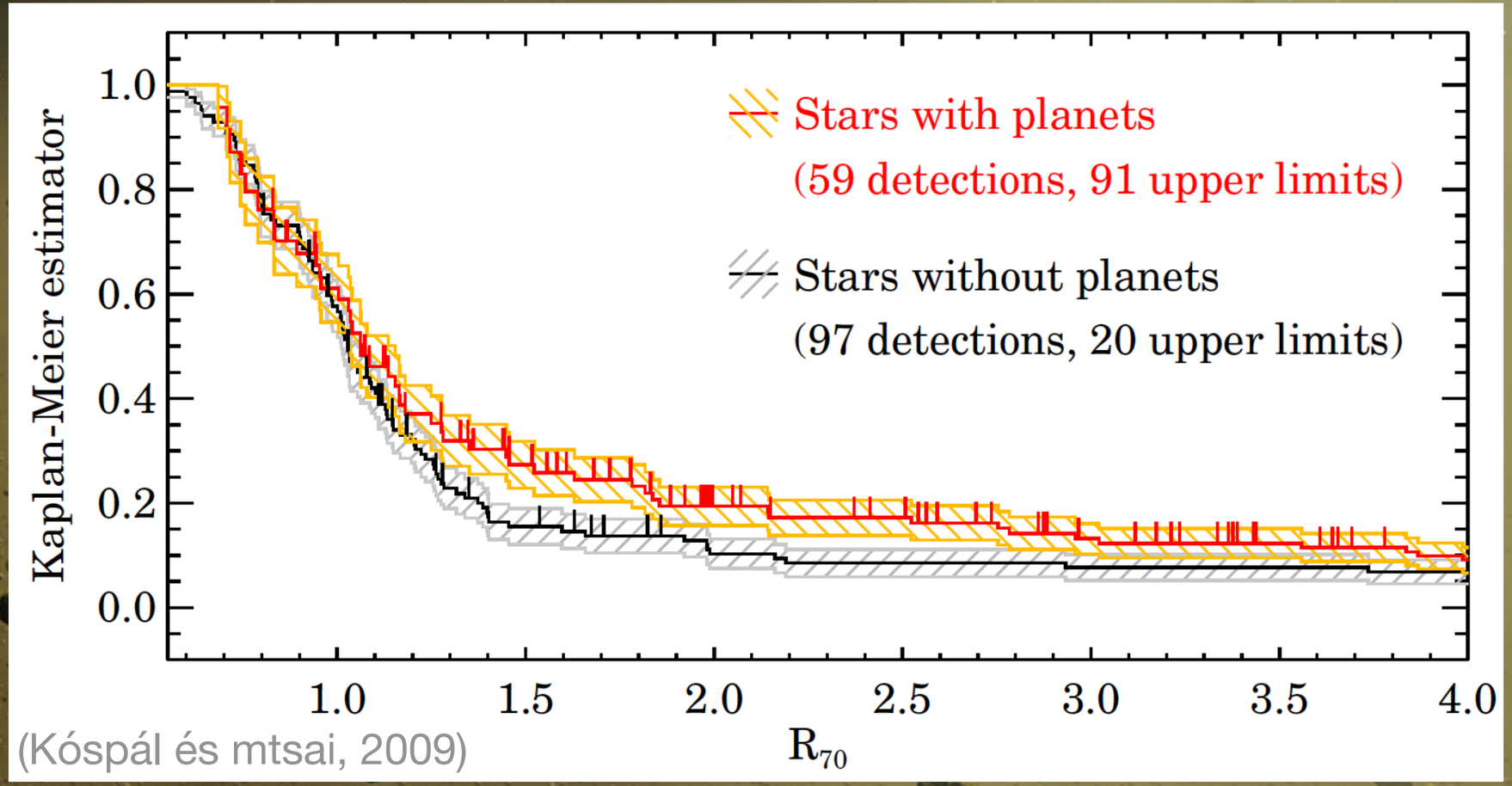
Kóspál Ágnes

MTA Csillagászati és Földtudományi Kutatóközpont
Konkoly-Thege Miklós Csillagászati Intézet

Földi sokaságok, égi tünemények – A statisztika a tudományok világában

2017. október 18.

(T. Pyle/NASA)



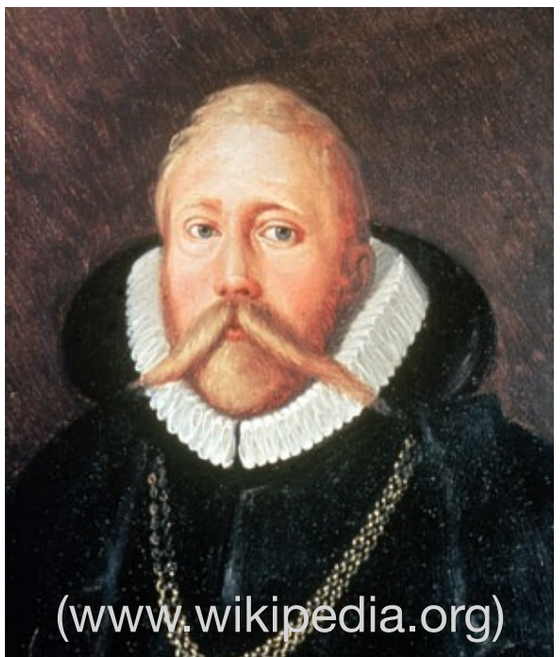
(Kóspál és mtsai, 2009)

(T. Pyle/NASA)

Már az ókori görögök is?

- A modern matematikai statisztika alapjait a 18. században fektették le
- A természettudósok azonban ennél jóval régebb óta alkalmaznak statisztikai módszereket a megfigyelt jelenségek értelmezésére
- A csillagászat az egyik legősibb megfigyelő tudomány

Már az ókori görögök is?



- Hipparkosz: az év hosszára vonatkozó babiloni megfigyelések szórnak; átlag vagy medián helyett az értéktartomány közepét vette, mint a legpontosabb érték

- Al-Biruni: a pontatlan műszerek vagy figyelmetlen megfigyelők miatti hibák terjedése

- Tycho Brahe: megismételt mérések növelik a pontosságot

Csillagász? Matematikus?

A 19. században ezek még ugyanazok az emberek voltak!

- Legendre, Laplace, Gauss: statisztikai módszerek az égi mechanikai jelenségek leírására (pl. az üstökösök pályája)


(Philipp Salzgeber)

- Huygens, Newton, Halley, Bessel, Airy: statisztikai módszerek pl. szerencsejátékokban, pénzügyi kockázatok elemzésében, vagy a társadalomtudományokban

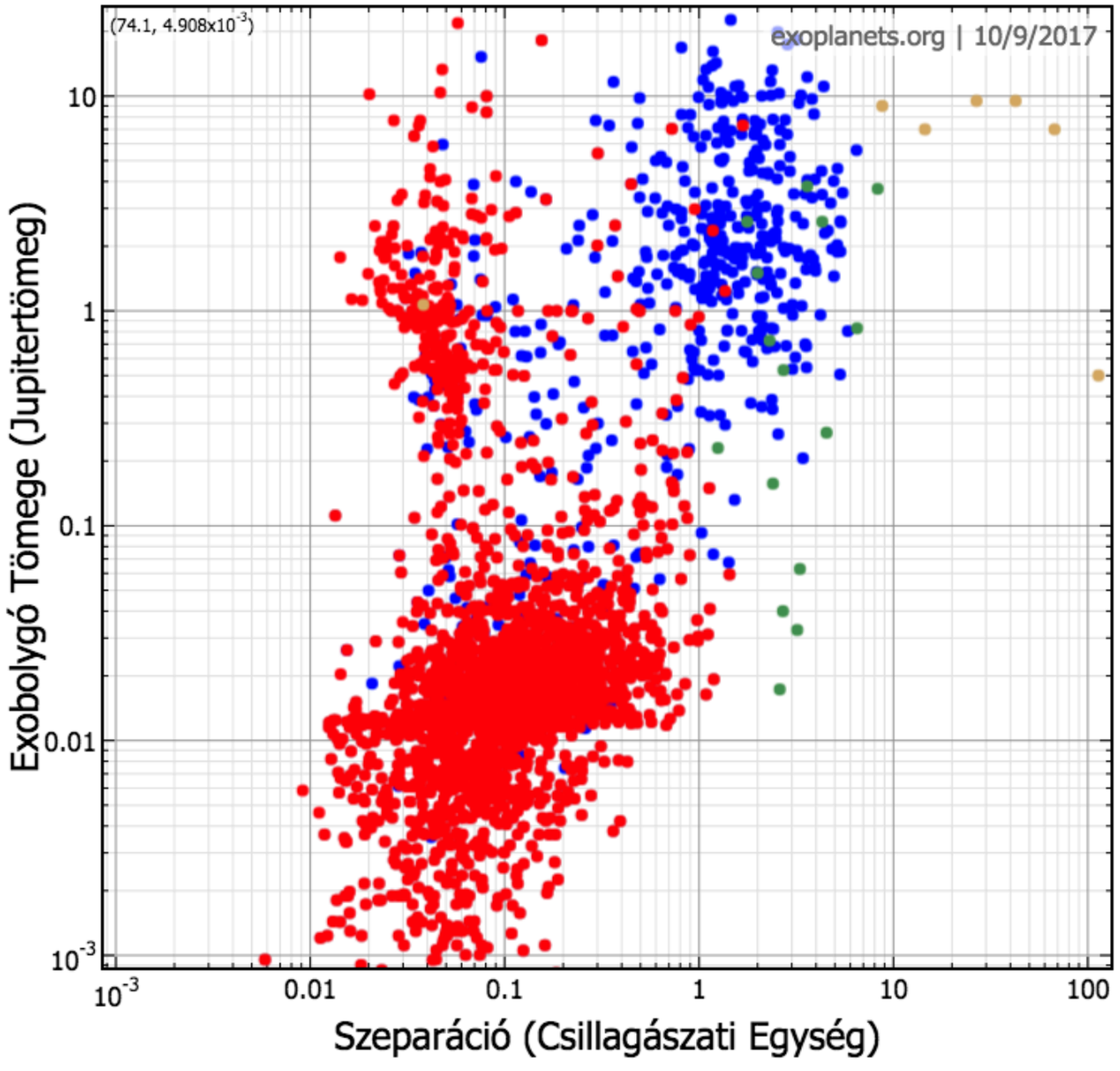
Csillagász? Matematikus?

A 20. században a csillagászat és a statisztika szétvált:

- csillagászat: elektromágnesesség, termodinamika, kvantummechanika és relativitáselmélet
- statisztikus csillagászat egy szűk terület maradt
- statisztika: elsősorban a társadalomtudományokban és élettudományokban (orvostudomány, környezettudomány, mezőgazdaság) alkalmazták

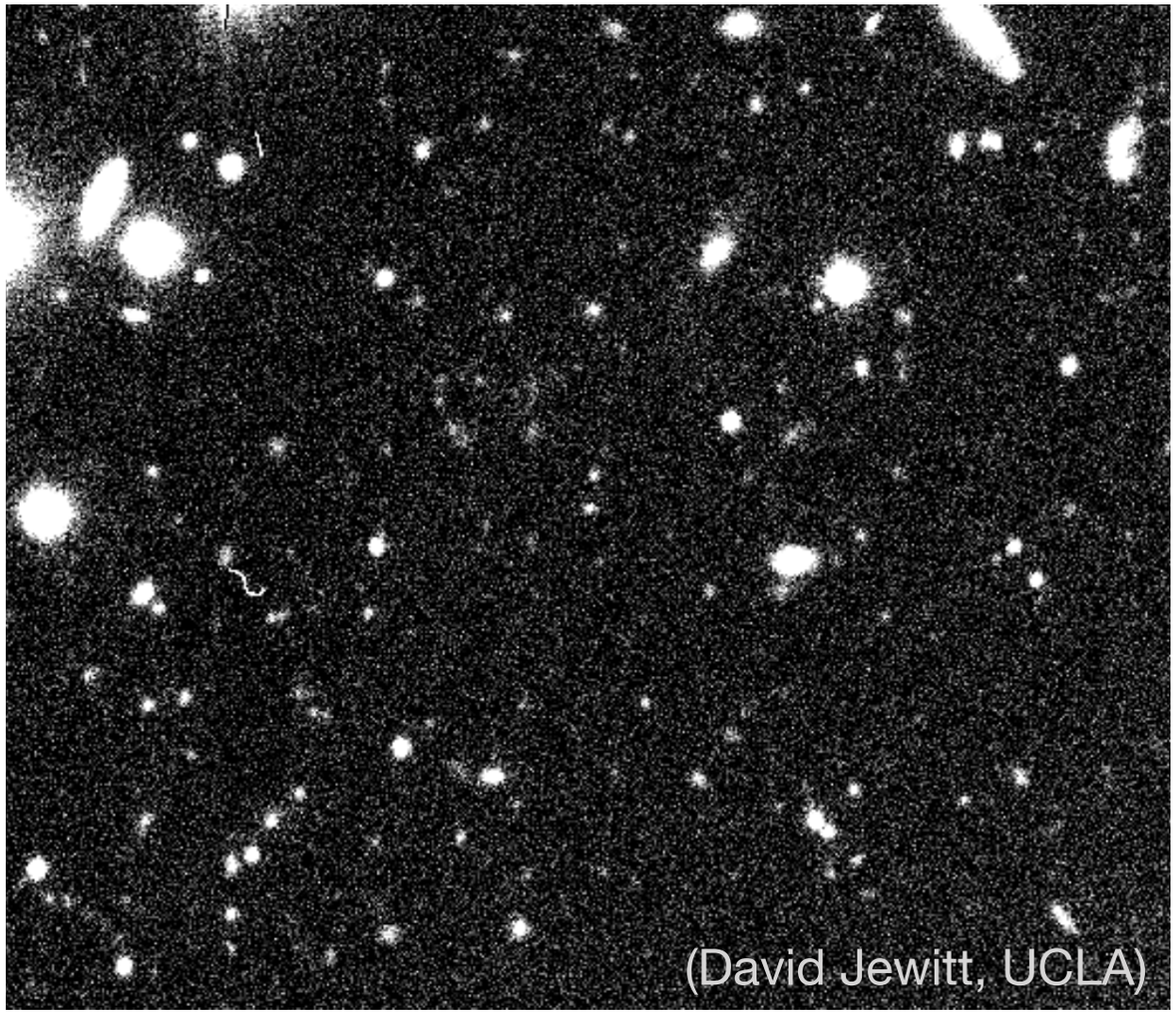
Statisztikára szükség van a csillagászatban!

- A megfigyelt csillagok/ galaxisok/molekulafelhők/ gamma-források vajon egy tipikus, torzítatlan mintáját alkotják a teljes populációnak?
- Mi a fizikai kapcsolat csillagászati objektumok egy csoportjának több különböző tulajdonsága közt, különösen, ha mérési kiválasztási effektusok is jelen vannak?



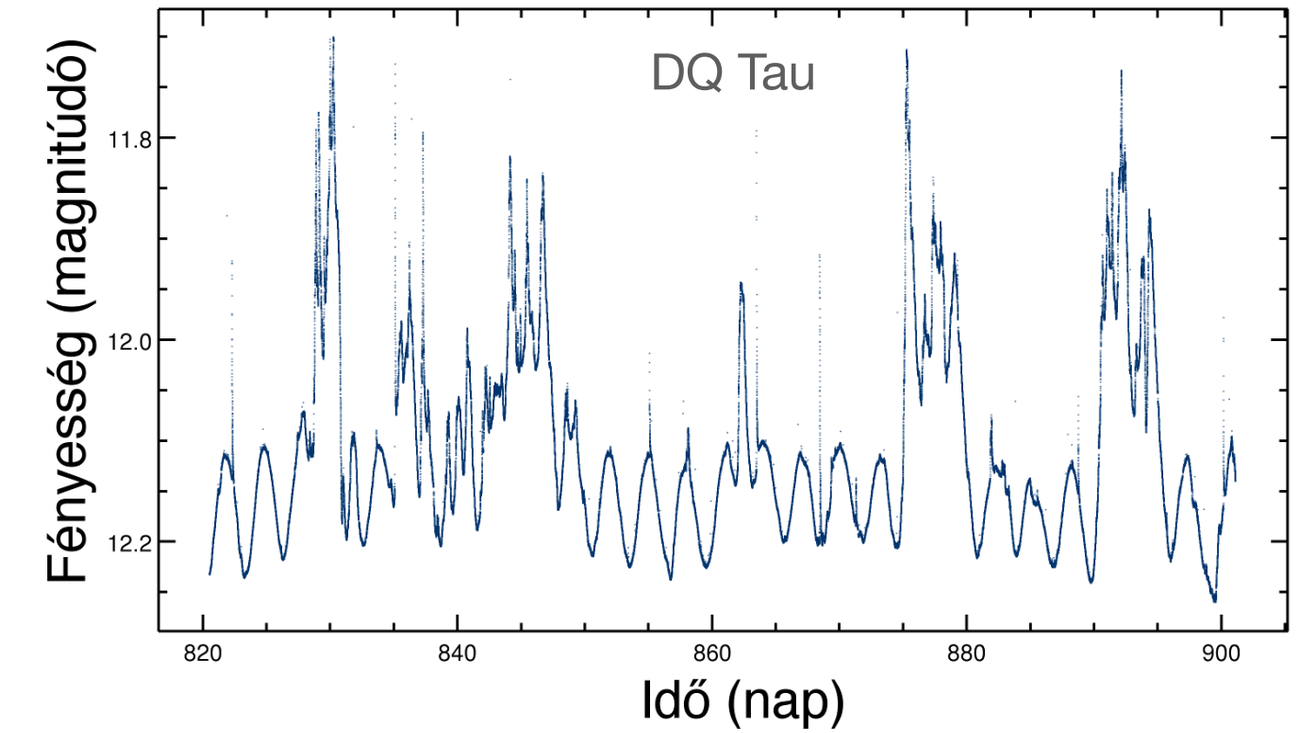
Statisztikára szükség van a csillagászatban!

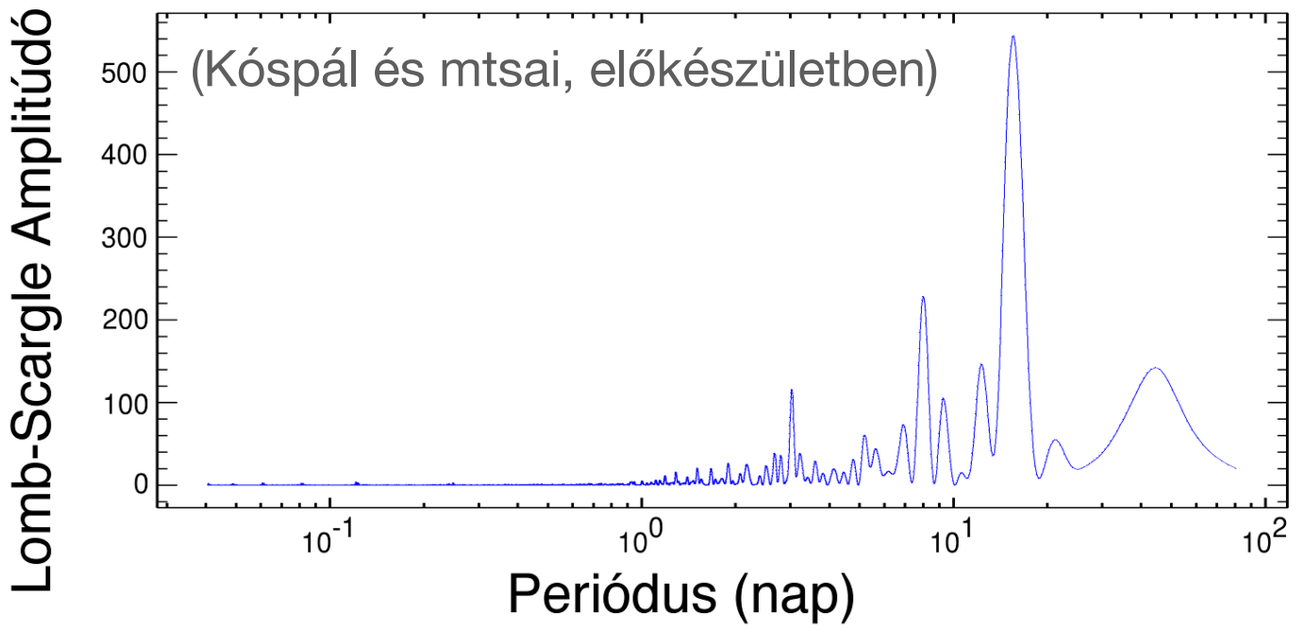
- Hogyan vonhatunk le következtetéseket, ha a mért tulajdonságok mérési hibával terheltek vagy felső/alsó határokat is tartalmaznak?
- A csillagászati képen/ színképben megfigyelt jel mikor valós és mikor lehet a zaj miatti random esemény?



Statisztikára szükség van a csillagászatban!

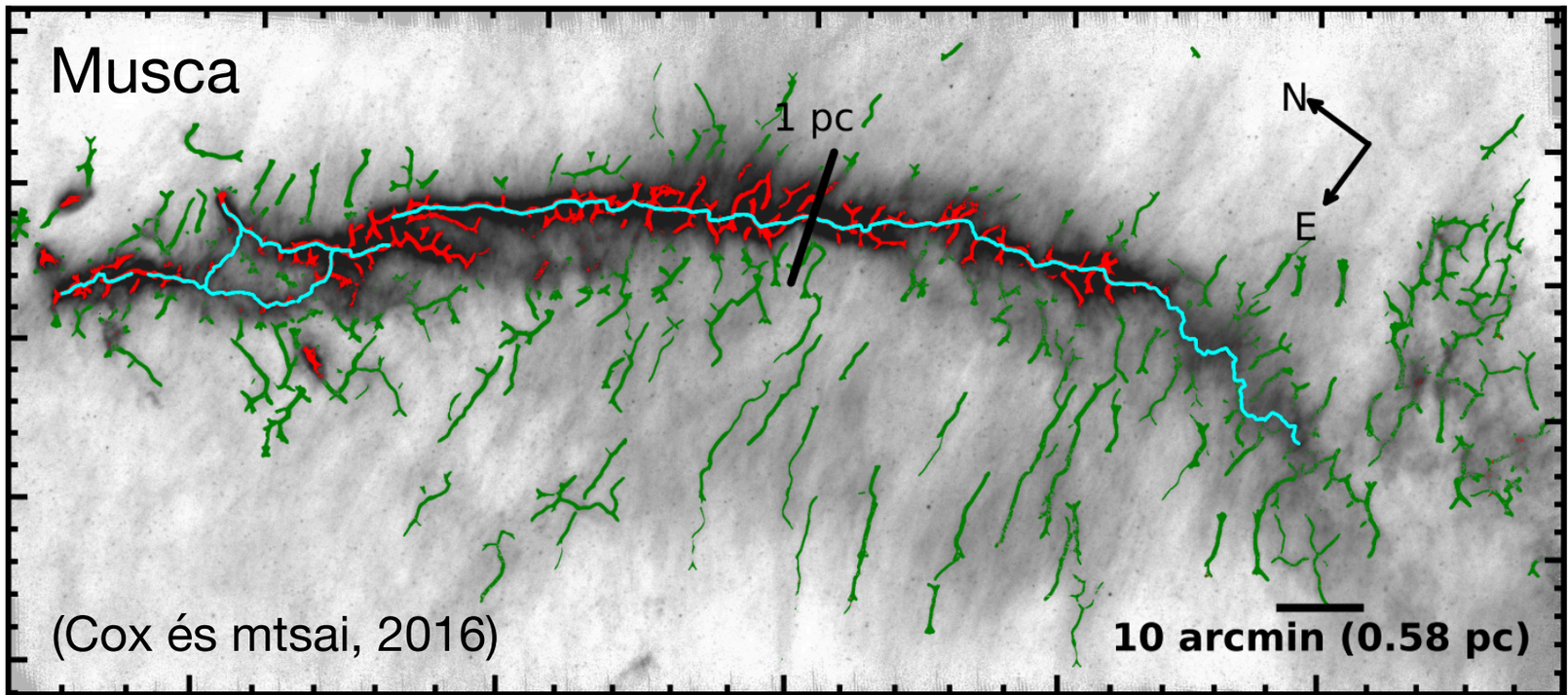
- Hogyan interpretáljuk a csillagászati objektumokról jövő időben változó, periodikus vagy sztochasztikus jeleket?
- Hogyan modellezzünk pontokat 4, 5, 6, … n-dimenziós fázistérben?





Statisztikára szükség van a csillagászatban!

- Hogyan kezeljük a folytonos struktúrákat, mint a csillagközi anyag vagy a kozmikus mikrohullámú háttérsugárzás?
- Hogyan illesszünk csillagászati színképekre nemlineáris asztrofizikai modelleket, és mit mondhatunk a legjobban illeszkedő paraméterek konfidencia-intervallumairól?



Asztrostatisztika

- **Leggyakrabban használt eszközök:**
    - Fourier-transzformáció idősor-analízisben (Fourier 1807)
    - legkisebb négyzetek módszere (Legendre 1805)
    - Kolmogorov-Smirnov teszt (Kolmogorov 1933)
    - főkomponens-analízis (Hotelling 1936)

- **Szükséges modern módszerek:** Hipotézis-vizsgálat, becsléselmélet, Bayes-elmélet, mintavételezési elmélet, túlélés-analízis (hiányzó adatok problémája), mérési hiba-modellek, többváltozós analízis, harmonikus és autoregresszív idősor-elemzés, wavelet analízis, valószínűség-sűrűség becslése, lineáris és nemlineáris regresszió, … **+ ezek kombinációja**

Két hazai példa

- **Balázs és mtsai:**
    - pont-eloszlások statisztikai vizsgálata
- **Marton és mtsai:**
    - gépi tanulási modellek klasszifikáció és regressziós analízis céljából

Az univerzum legnagyobb szabályos alakzata

- Kozmológiai elv: az univerzumban nagy távolságskálán vizsgálva nincsenek kitüntetett irányok és helyek (homogén és izotróp)

- Homogenitási skála:
  - 0.3 milliárd fényév (Ntelis, 2016)
  - 0.3 – 0.4 milliárd fényév (Scrimgeour és mtsai, 2012)
  - > 0.65 milliárd fényév (Silos Labini, 2012)
  - 1.2 milliárd fényév (Yadav és mtsai, 2010)

Az univerzum legnagyobb szabályos alakzata

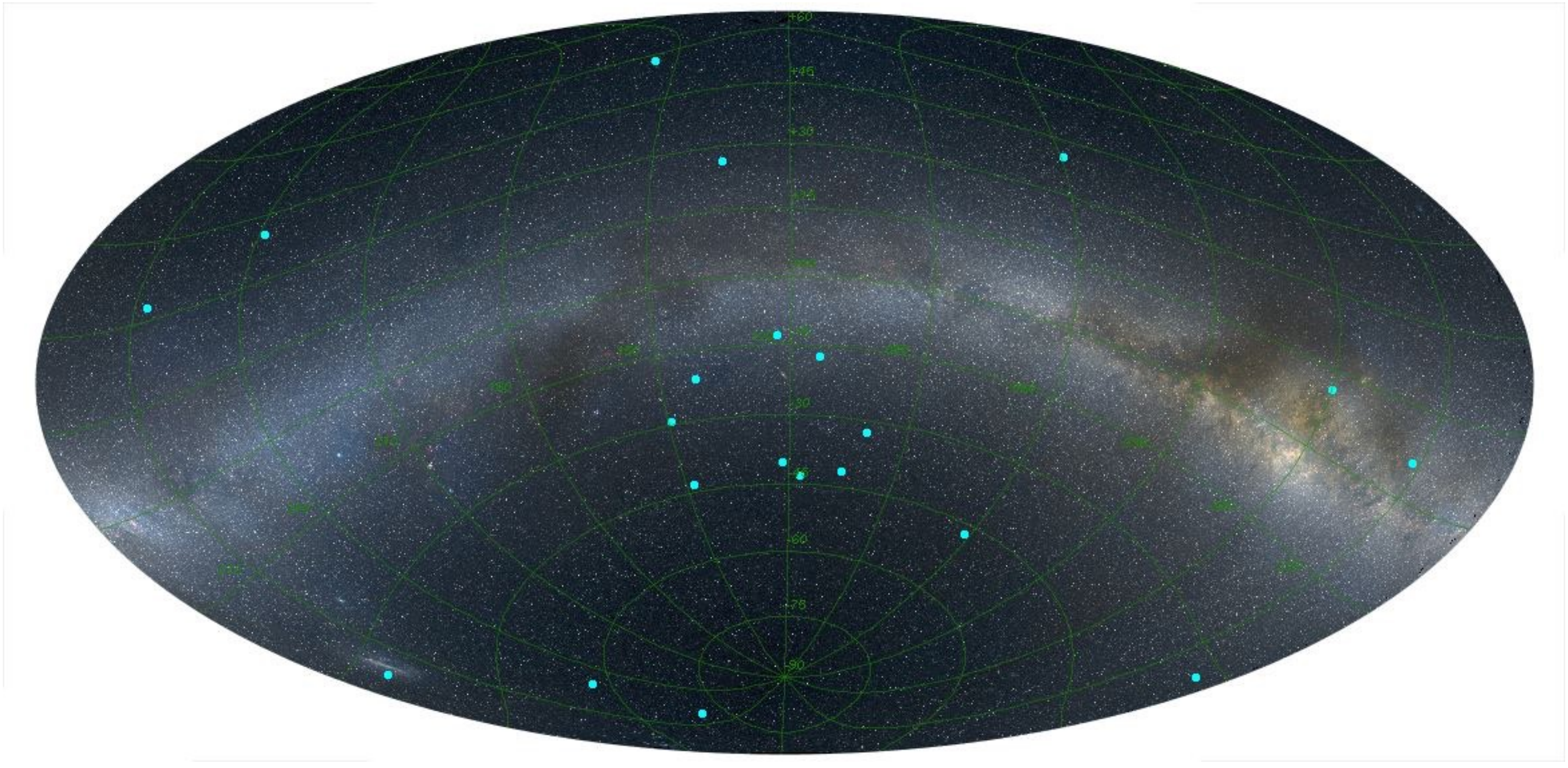
- **Balázs és mtsai (2015, 2017) felfedezése: egy óriási gyűrű, melynek átmérője 5.6 milliárd fényév**

- A gyűrűt olyan gammafelvillanások alkotják, amelyek 6.8 – 7.2 milliárd fényév távolságban történtek

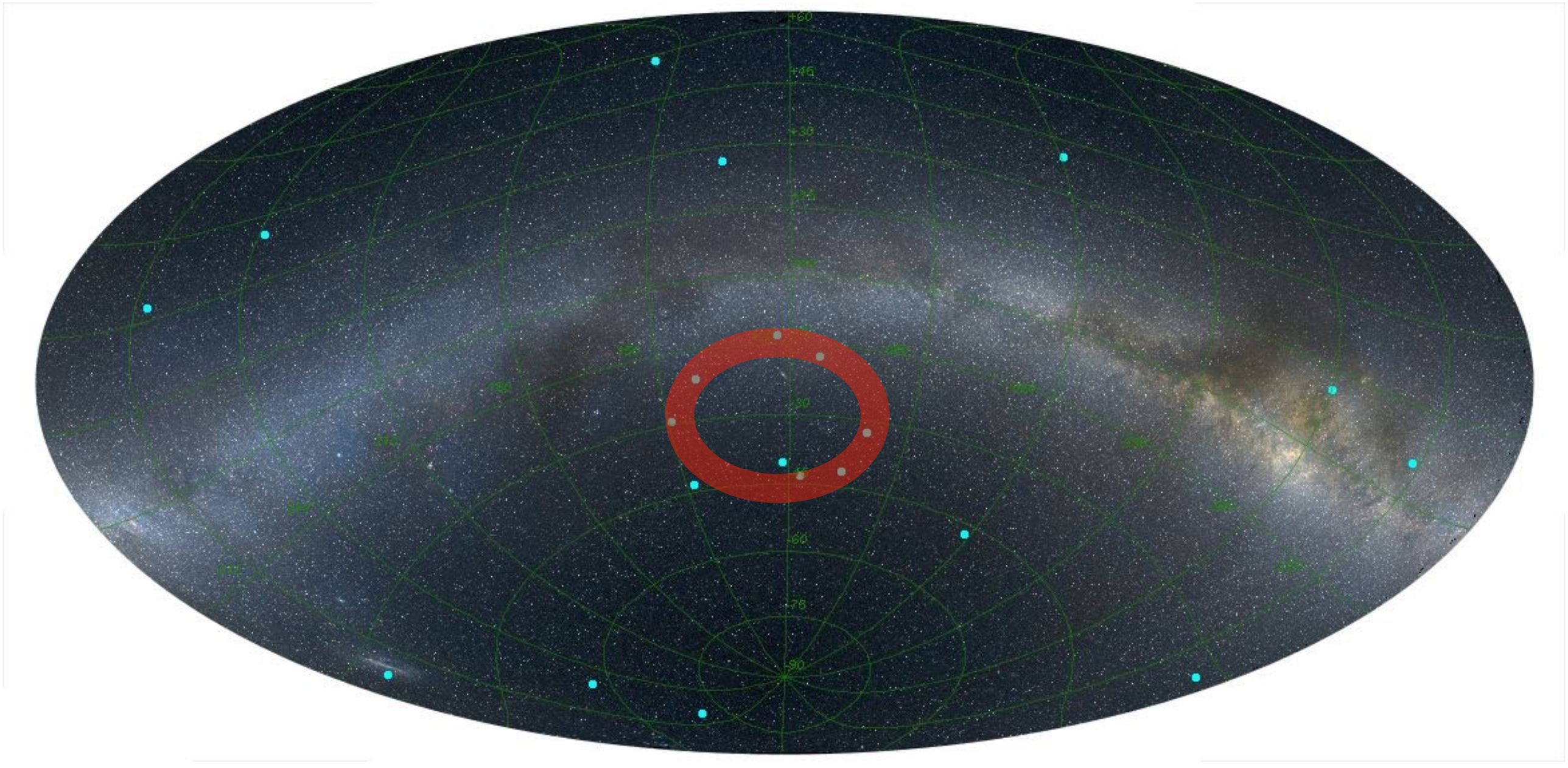
- Gammafelvillanás: univerzum legfényesebb jelenségei, nagy kozmológiai távolságokból is megfigyelhetők

- Hogyan létezhet ilyen óriási struktúra a világegyetemben? Ellentmond a kozmológiai elvnek!

Az univerzum legnagyobb szabályos alakzata

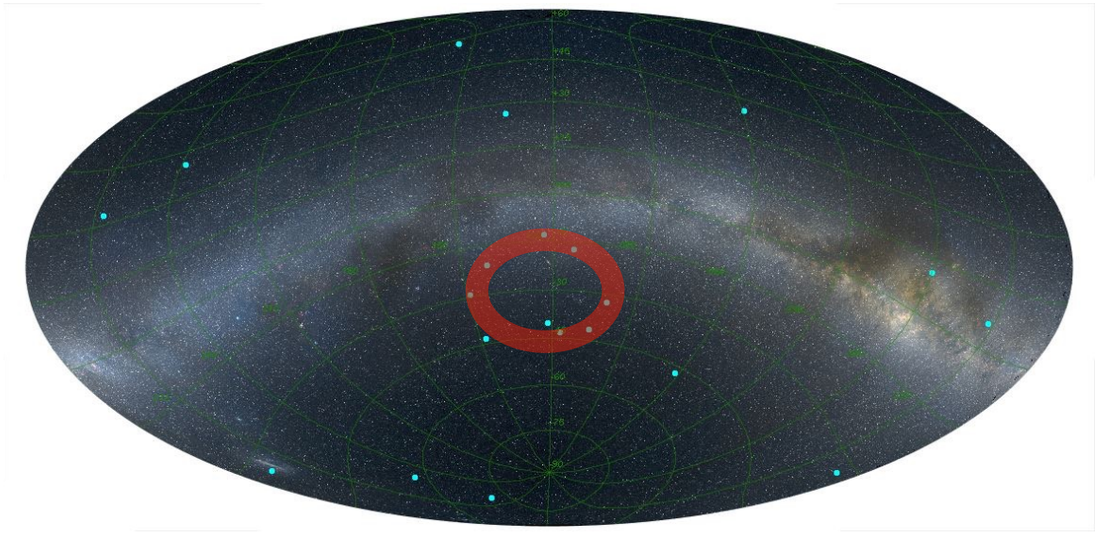


Az univerzum legnagyobb szabályos alakzata



Az univerzum legnagyobb szabályos alakzata

- Biztos, hogy nem a gamma-felvillanások egy random fluktuációja?



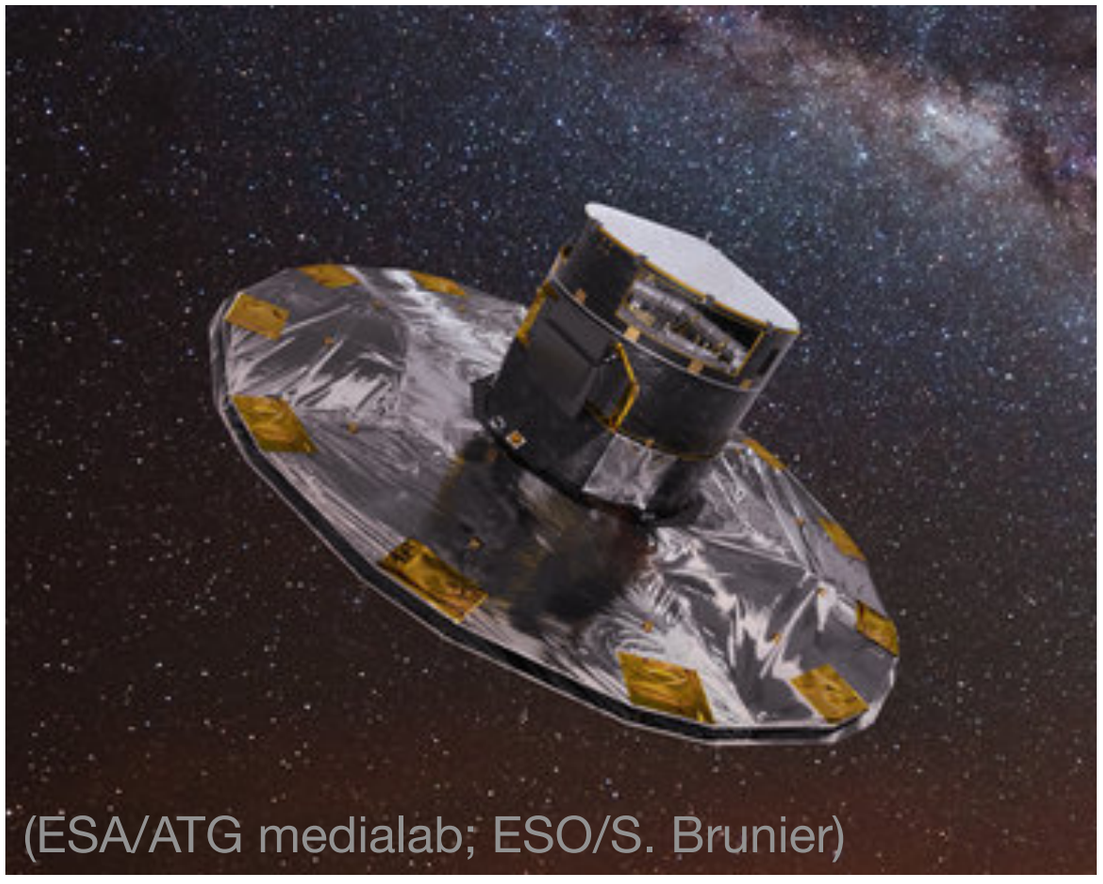
- Nem, ennek a valószínűsége mindössze 2×10^{-6}

- Módszer: Gamma-felvillanások pont-eloszlásának statisztikai vizsgálata

Fiatal csillagok a Gaia-val

Gaia asztrometriai űrmisszió (2014 – 2019):

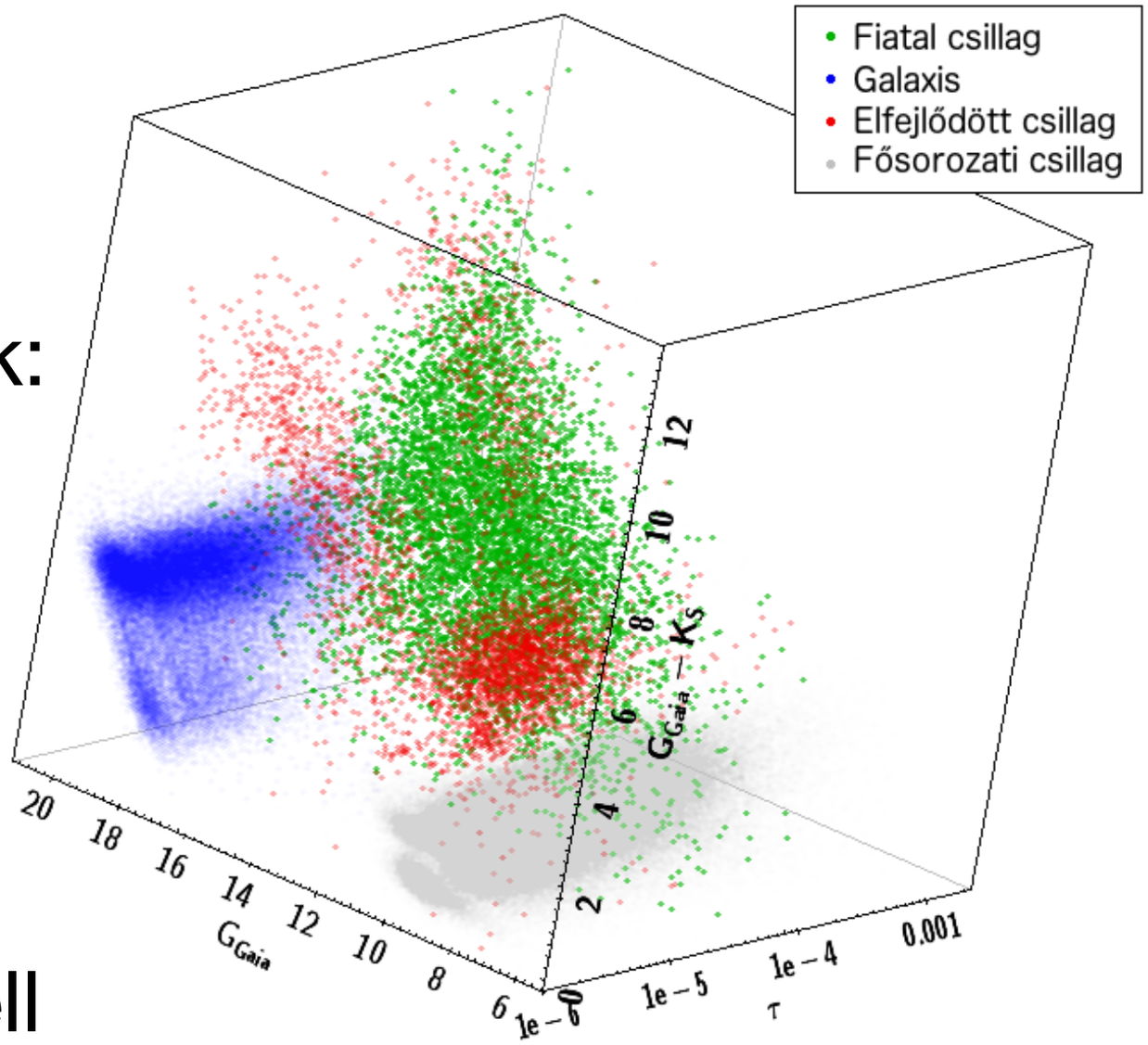
- 1 milliárd csillag pontos pozícióját, távolságát, sebességét és fényességét méri
- naponta 70 millió objektum
- 40 GigaByte adat naponta
- 73 TeraByte teljes adatmennyiség 5 év alatt
- 1 PetaByte lesz a teljes katalógus


(ESA/ATG medialab; ESO/S. Brunier)

Fiatal csillagok a Gaia-val

Hogyan dönthetjük el egy objektumról, hogy fiatal csillag? Gépi tanulással!

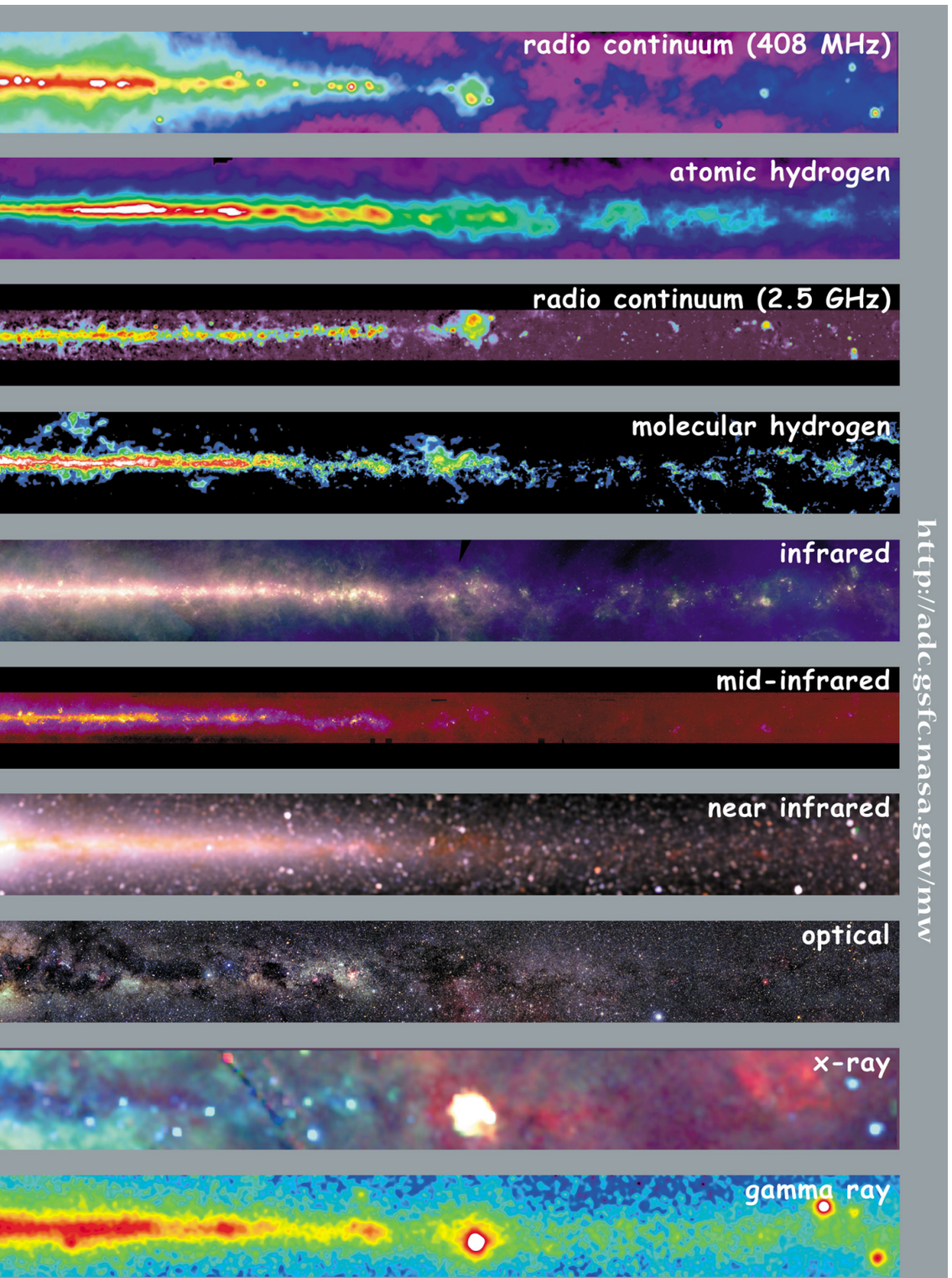
- Rendelkezésre álló adatok: Gaia fényességadatai, Gaia és más távcsövek mérésein alapuló szín, Planck űrtávcső adatain alapuló poropacitás



(Marton és mtsai, előkészületben)

- Tanuló adatbázis → modell
- Adat → modell → információ az adatokról
  pl. mi annak a valószínűsége, hogy egy adatpont fiatal csillag

A jelen kihívásai: VO



- **Virtuális obszervatórium:** heterogén adatbázisok
- Fizikailag eltérő helyen vannak
- Különböző adatbázis-architektúrákat alkalmaznak
- Különböző hullámhosszakon készültek (gamma-röntgen-UV-optikai-infravörös-szubmilliméteres-rádió)
- Különböző térbeli felbontással
- Különböző lefedettséggel

A jövő kihívásai: big data

LSST: Large Synoptic Survey Telescope (2023 – 2033):

- háromnaponta végigméri a teljes látható eget
- naponta 15 TeraByte adat
- 60 PetaByte teljes adatmennyiség 10 év alatt

A jövő kihívásai: big data

Adatbányászat LSST adatokban

Clustering (Segmentation): *Group data items according to tight relationships or greatest similarity, and separate the items that are most different.*

Principal Component Analysis: *Reduce the dimension of the input vectors (multi-attribute data records) by eliminating redundant components – captures the directions of greatest variance in the data.*

Independent Component Analysis (ICA): *Identify the mutual statistically independent components in multi-attribute data records.*

Outlier (Anomaly, Glitch, Deviation) Detection: *Find data items that fall outside the bounds of known or statistically robust clusters.*

Classification: *Assign data items to predetermined groups (classes, clusters).*

Bayesian Analysis: *Assess the probability of a hypothesis being correct (for example, whether a classification is valid) by incorporating the prior probability of the hypothesis and the experimental data supporting the hypothesis.*

Support Vector Machines (SVM): *Map input vectors to a higher dimensional space where the classes are divided by a maximal separating hyperplane.*

Nearest Neighbor Method: *Classify a data item according to its nearest neighbors (i.e., records that are most similar).*

Association Mining (Market Basket Analysis): *Associate data with higher than expected co-occurrence of attribute-value combinations.*

Link Analysis: *Associate data in a graph network that are connected through shared attribute values or semantic relationships.*

Artificial Neural Networks (ANN): *Non-linear predictive models that learn through training and resemble biological neural networks in structure.*

Rule induction: *Extract useful if-then rules from data based upon statistical significance and information gain.*

Decision Trees: *Hierarchical sets of decisions, based upon rules, for rapid classification of a data collection.*

Genetic Algorithms: *Rapid optimization techniques that are based on the concepts of natural evolution.*

Data visualization: *The illustration and visual interpretation of complex relationships in multidimensional data using graphics tools.*

Self-Organizing Map (SOM): *Graphically organizes (in a 2-dimensional map) the information stored within a database based upon similarities and links between concepts. It can be used to find hidden relationships and patterns in more complex data collections.*

(Borne és mtsai, www.lsst.org)

További olvasnivaló

- Wall & Jenkins: Practical Statistics for Astronomers, Cambridge University Press, 2012
- Feigelson & Babu: Statistical Challenges in Modern Astronomy, PHYSTAT2003, Stanford, 2004