

Adatintenzív megközelítés a tudományban és az orvostudományban

CSABAI ISTVÁN
ELTE, KOMPLEX RENDSZEREK FIZIKÁJA TANSZÉK

A tudomány kialakulása – köör

Modell



MEGFIGYELÉS

Valóság



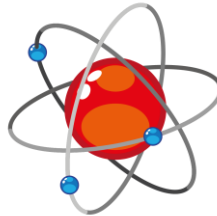
A tudomány kialakulása – köör

Modell



MEGFIGYELÉS

Valóság



A tudomány kialakulása – kőkor

Modell

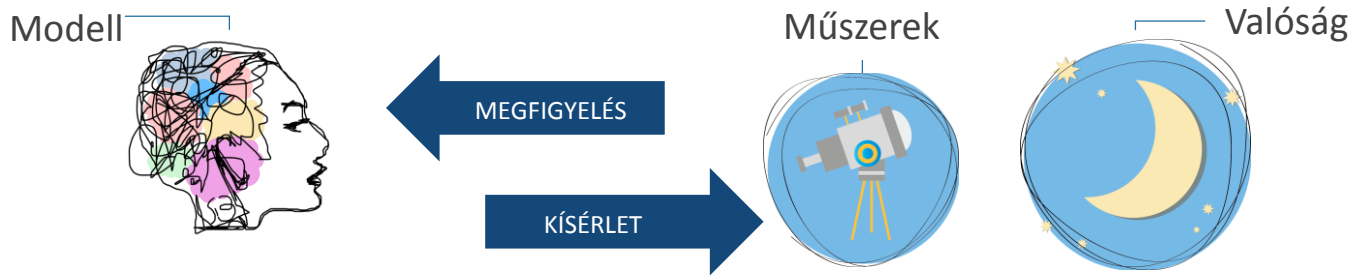


MEGFIGYELÉS

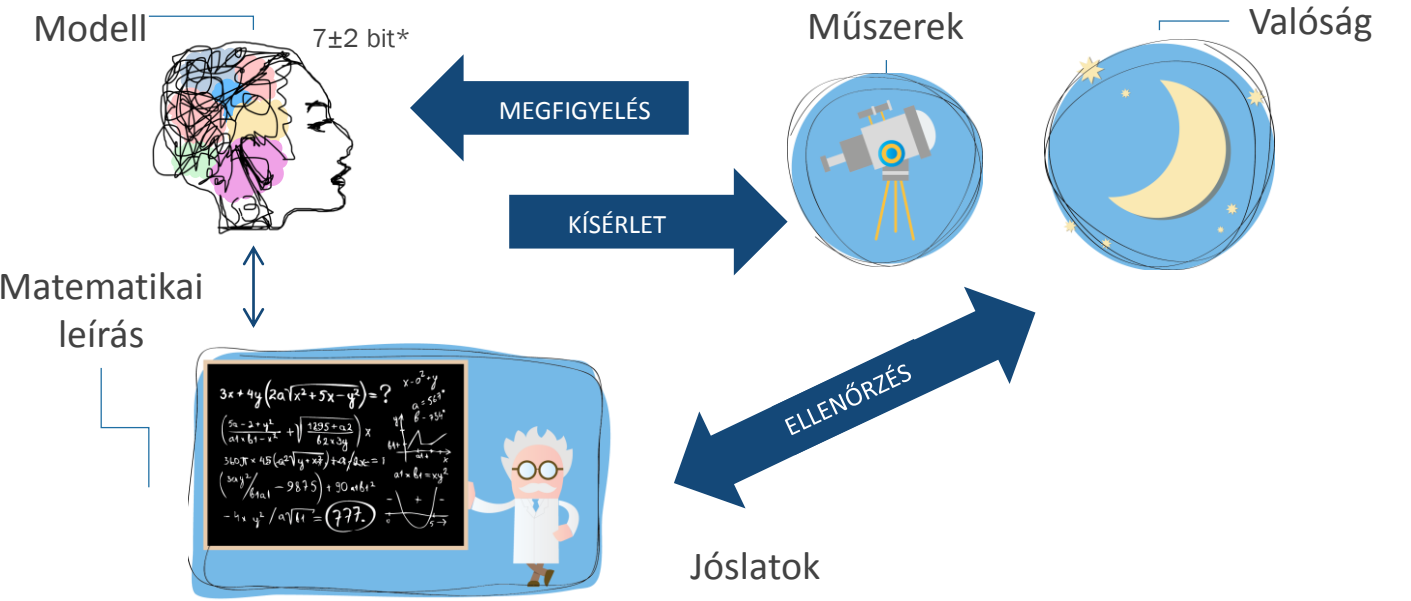
Valóság



A tudomány kialakulása – iparosodás előtt



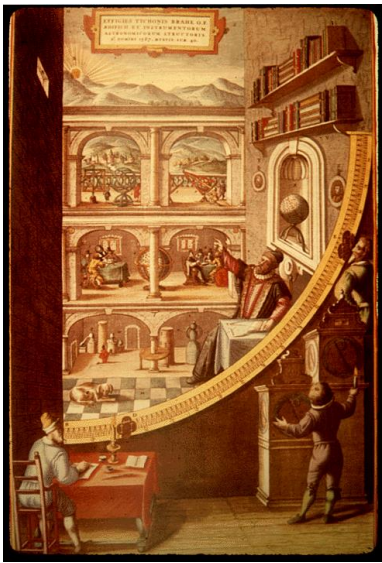
A tudomány kialakulása – iparosodás előtt



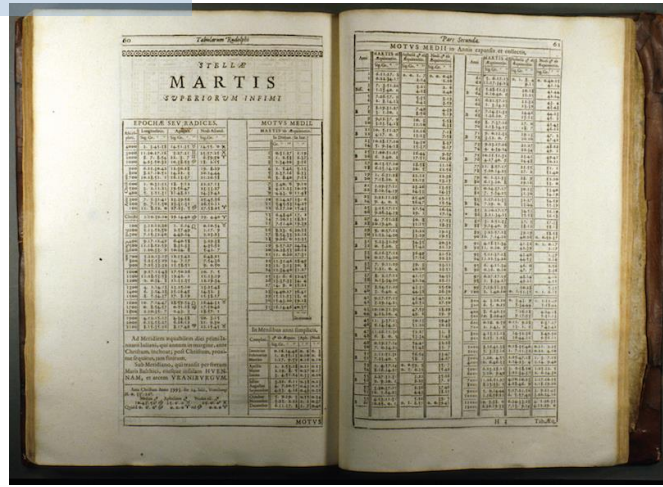
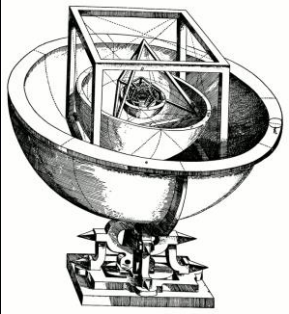
* Miller, G. A. Psychological Review. 63 (2): 81–97 (1956)

Adattudomány a kezdetekben: csillagászat

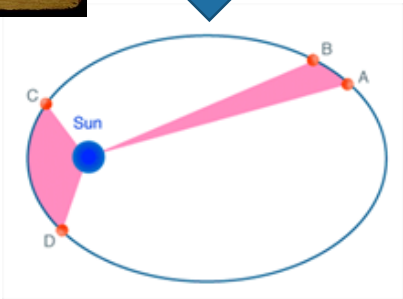
Tycho Brahe, 1572



Tycho Brahe: **adatok**
J. Kepler: **modell**
I. Newton: **elmélet**



Johann Kepler, Tabulae Rudolphinae (1627)

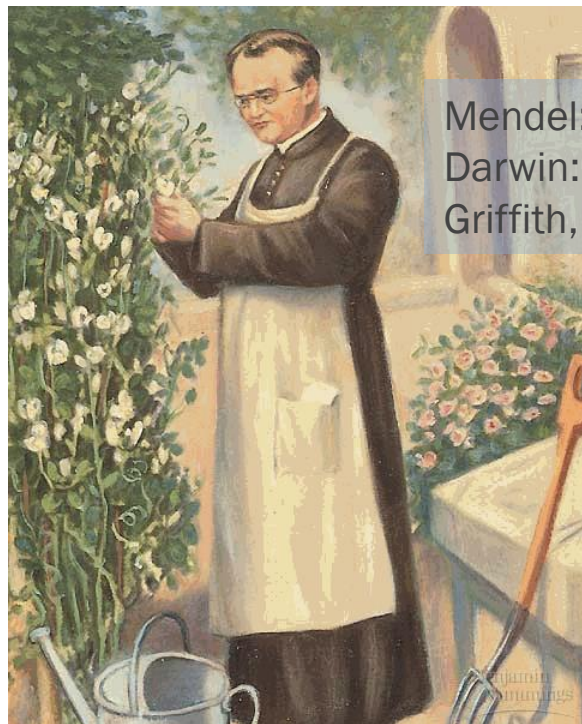


$$F = G \frac{m_1 m_2}{r^2}$$

Adattudomány a kezdetekben: genetika

8 év, ~28.000 borsó növény

Gregor Mendel, 1865



Mendel: data

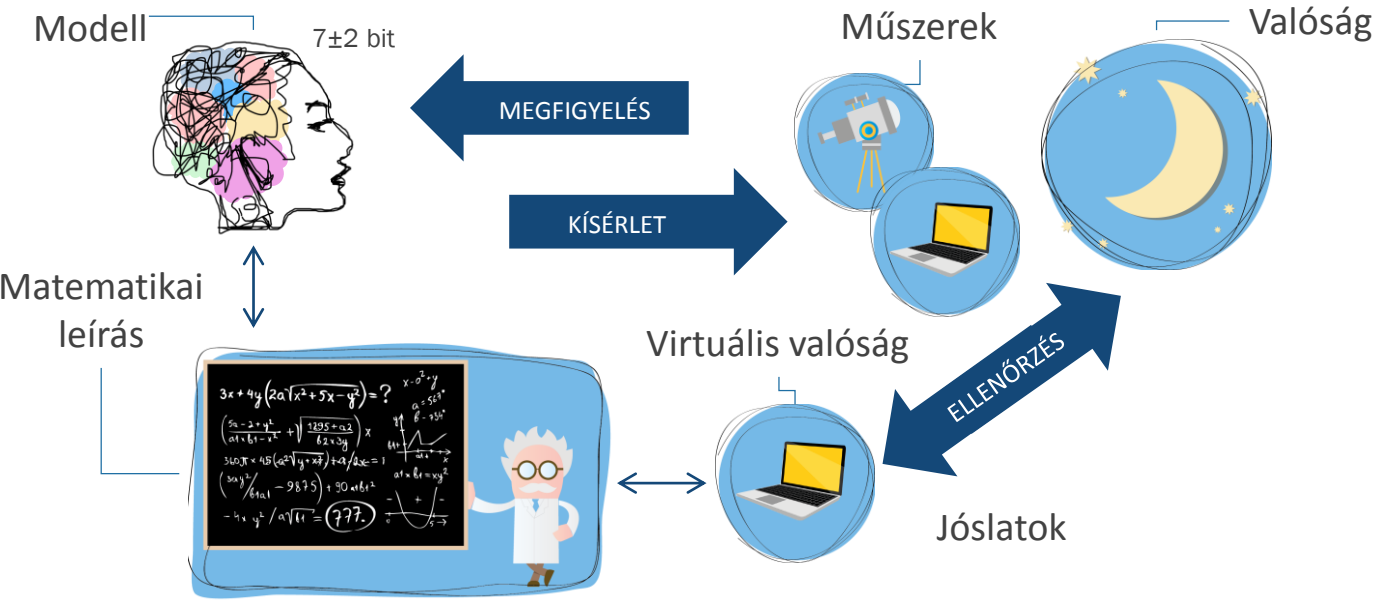
Darwin: "effective" model

Griffith, Avery: natural law

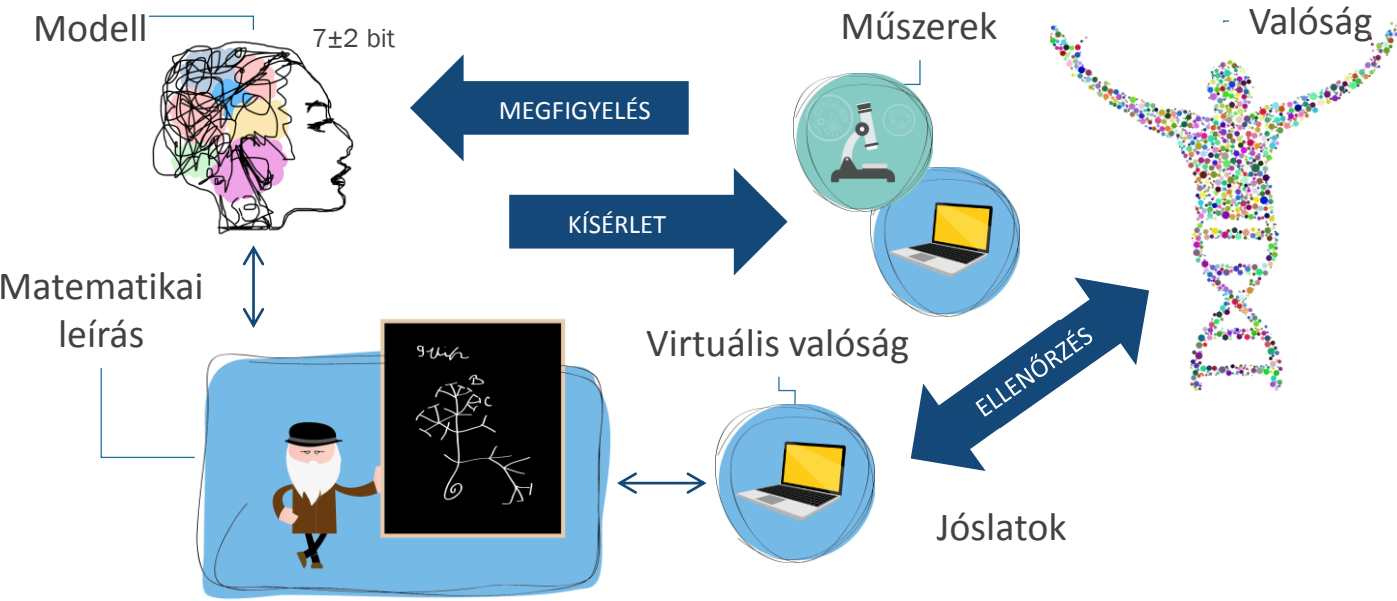
[illegible]

Courtesy of the Mendelianum, Moravian Museum, Brno.
Noncommercial, educational use only

A tudomány kialakulása – napjainkban



Hasonlóan a biológiában és orvostudományban



Egyre olcsóbb eszközök – egyre gyorsabb gépek

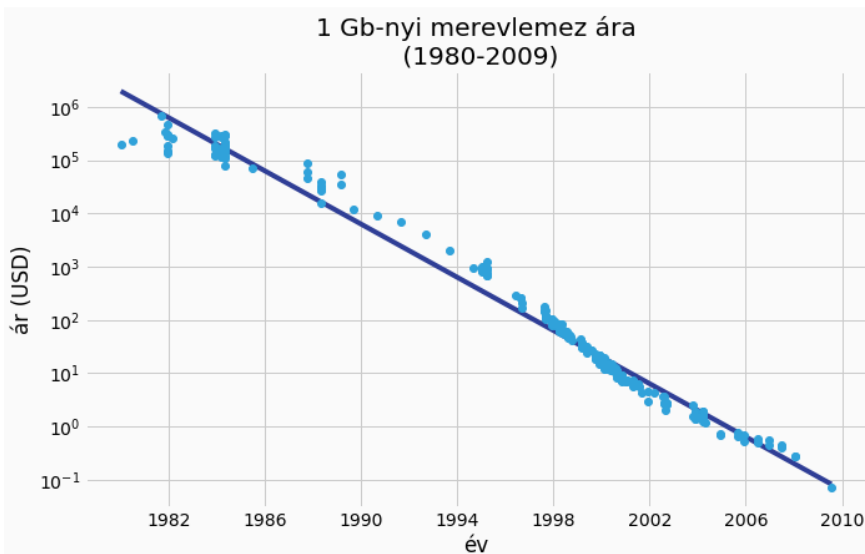
Kvantummechanika



Szilárdtestfizika



Tranzisztor/
mikroelektronika



Moore-törvény:
exponenciális kapacitás növekedés



Elektronika



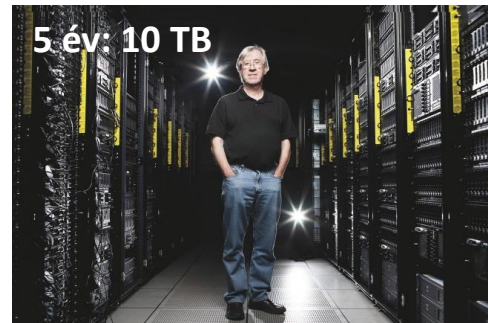
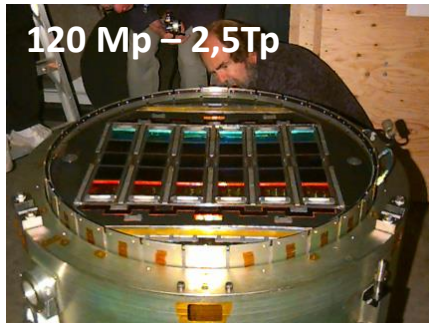
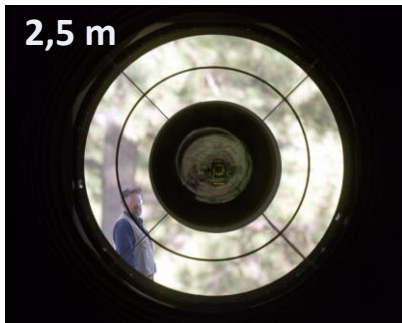
Szenzor



Adat

Tudományos „Big Data” példák

Az univerzum 3D térképe – SDSS



Automatikus „pipeline”

Több mint 150 emberévnyi fejlesztés

Az első nagyprojekt ahol a szoftverfejlesztésre fordítódott az erőforrások nagyobbik hányada

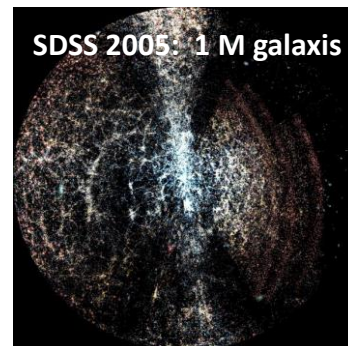
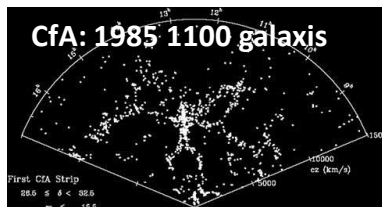
Nagy adattömeg

Több mint 300 millió objektum, 1 M spektrum

Egyenként 300+ paraméter

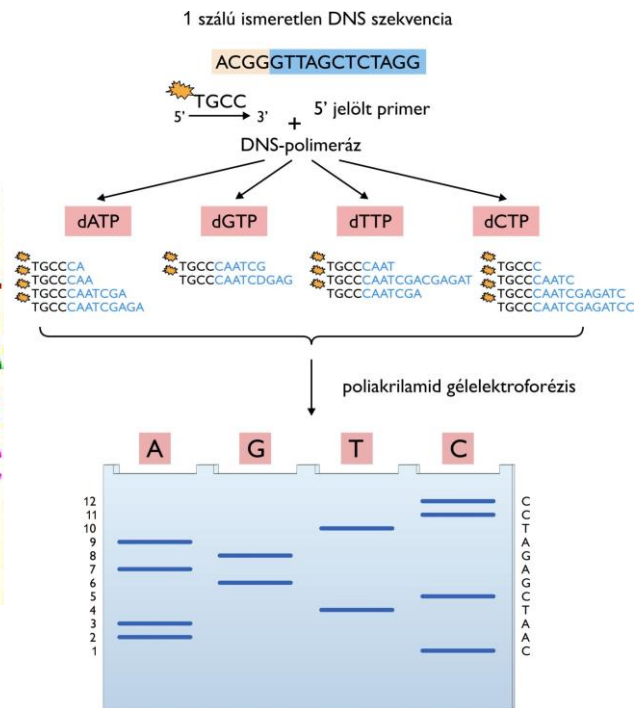
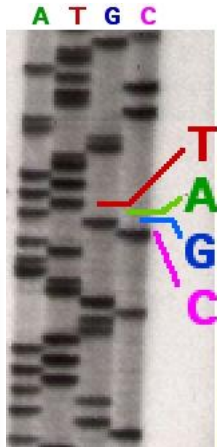
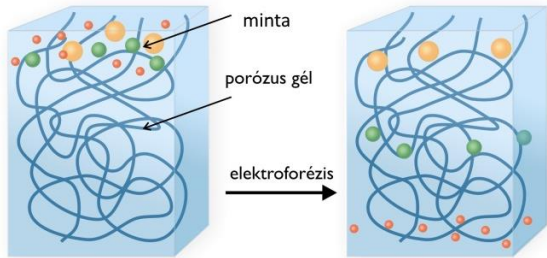
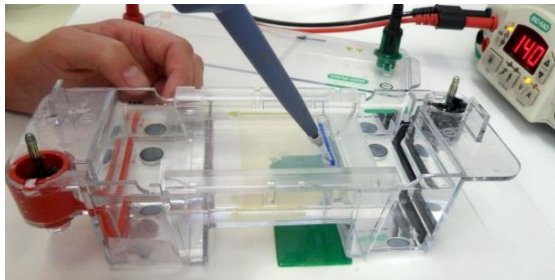
Közel 100 TB nyers adat,
10 TB katalógus, 2,5 terapixel
PUBLIKUS ADATBÁZIS (VO)

~1929: 1 galaxis

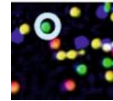
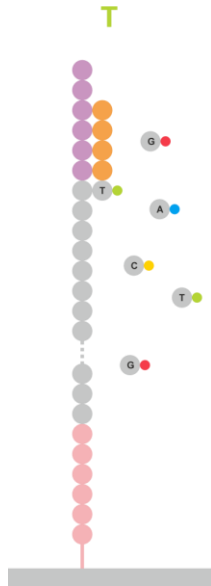


Szekvenálási technikák:

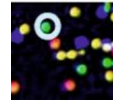
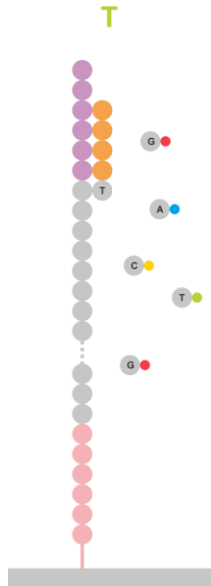
Sanger-szekvenálás (1977)



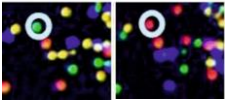
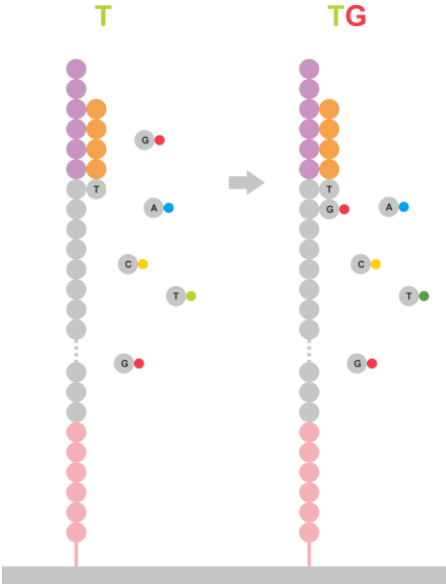
Szekvenálási technikák: NGS



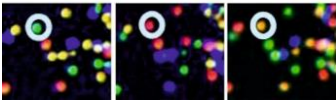
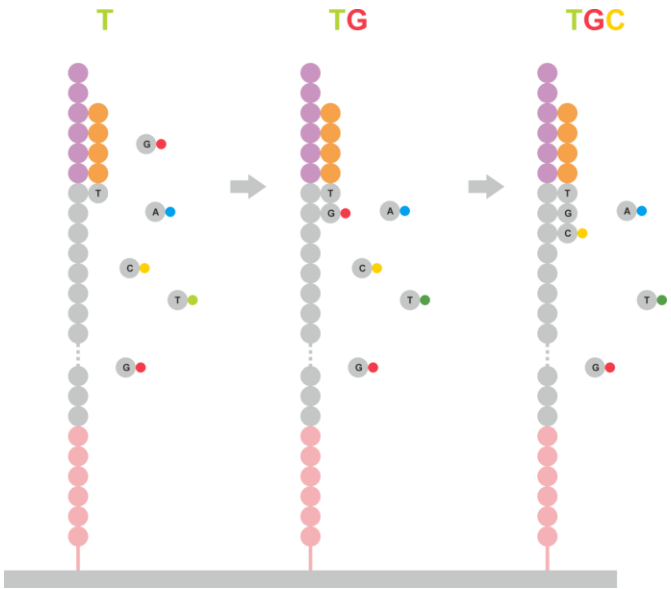
Szekvenálási technikák: NGS



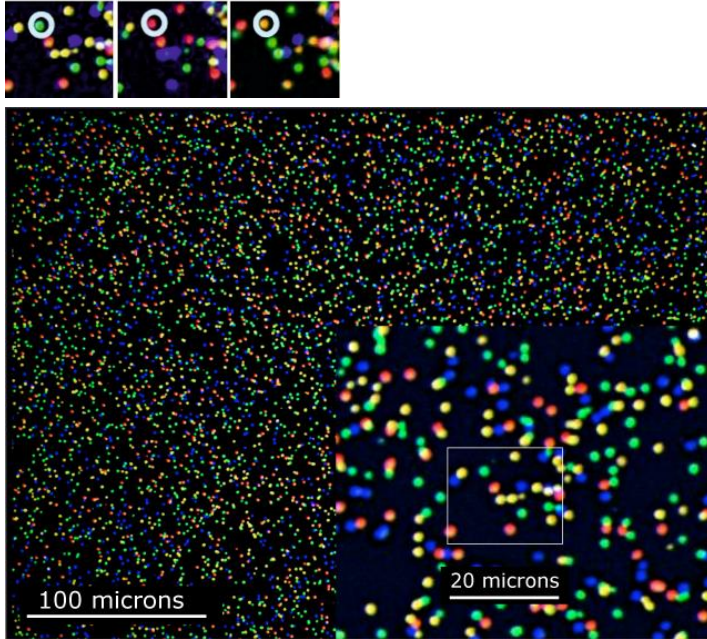
Szekvenálási technikák: NGS



Szekvenálási technikák: NGS



Szekvenálási technikák: NGS

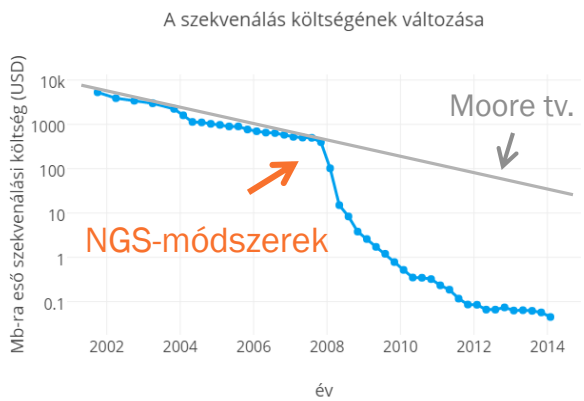


D. Mertens, K. Rippe, German Cancer research Center

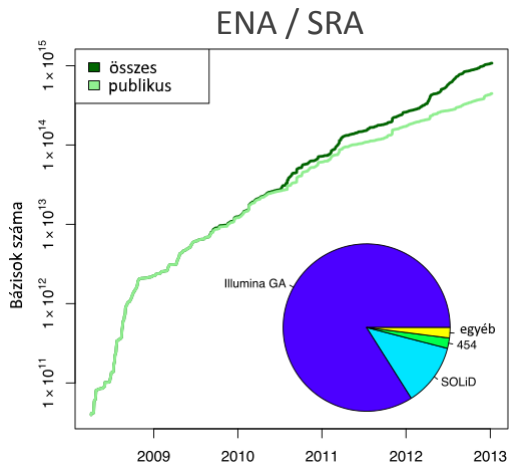


BGI Hong Kong, Scottd400, CC-BY-3.0

A genomika Moore törvénye



Egyre olcsóbb szekvenálni.



Egyre több adat publikus online.

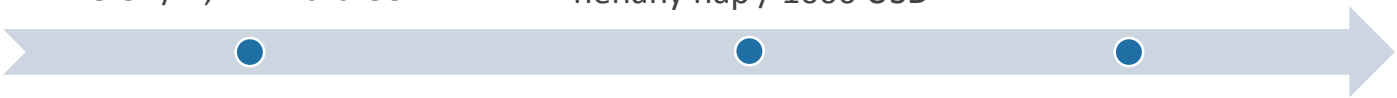
1990–2003

13 év / 2,7 milliárd USD

2016

néhány nap / 1000 USD

2020?





National Cancer Institute
National Human Genome Research Institute

The Cancer Genome Atlas Data: Navigating the Data Portal and the Cancer Genomics Hub

The Cancer Genome Atlas

<http://cancergenome.nih.gov/>

3.2Bn nucleotides / human genome

The Cancer Genome Atlas (TCGA) is a large-scale collaborative effort led by the National Institutes of Health to map the genomic changes that occur in over 30 types of human cancer, including nine rare tumors. Its goal is to support new discoveries and accelerate the pace of research aimed at improving the diagnosis, treatment, and prevention of cancer.

TCGA is a community resource project. The information generated by TCGA is centrally managed and entered into databases as it becomes available, making the data rapidly accessible to the entire research community. By January 2014, TCGA had generated one petabyte of data for about 10,000 cases of tumor and matching normal tissue samples.

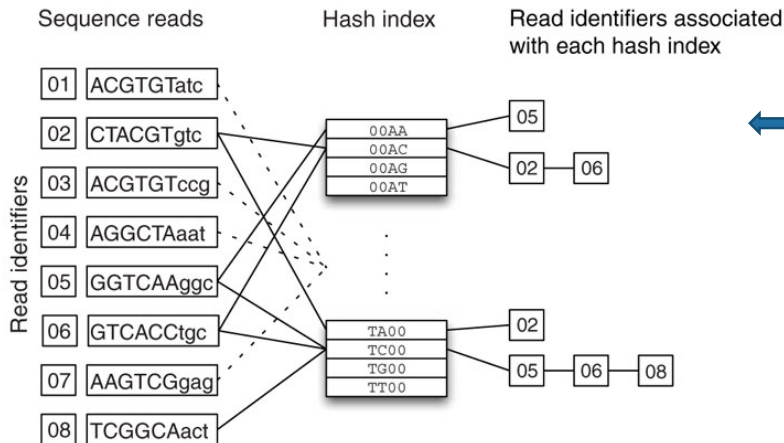
TCGA data are available in two data repositories: the TCGA Data Portal and the Cancer Genomics Hub. All data can be accessed directly from the TCGA Data Portal regardless of which repository houses the data file.

+ teljes populáció
+ „single cell”

NGS illesztés/assembly: nagy számok

- Számítási sebesség: $\sim 10^9$ művelet/sec
- Genom: $\sim 10^9$ nt
- NGS: $\sim 10^9$ short reads, egyenként ~ 100 nt
- Nyers erővel: $\sim 10^{20}$ összehasonlítás/minta $\rightarrow 10^{11}$ sec ~ 3000 év
- $\rightarrow$ **Ügyes indexelő/kereső algoritmusok kellenek!**





Számítástudományi trükkök:
Indexelés: HashTable
Eredetileg kriptográfiához illetve hibajavító
checksum célokra használták.

1. All possible rotations

^TAGTCGAGGCTTTA\$

^TAGTCGAGGCTTTA\$
TAGTCGAGGCTTTA\$^
AGTCGAGGCTTTA\$^T
GTCGAGGCTTTA\$^TA
TCGAGGCTTTA\$^TAG
CGAGGCTTTA\$^TAGT
GAGGCTTTA\$^TAGTC
AGGCTTTA\$^TAGTCG
GGCTTTA\$^TAGTCGA
GCTTTA\$^TAGTCGAG
CTTTA\$^TAGTCGAGG
TTTA\$^TAGTCGAGGC
TTA\$^TAGTCGAGGCT
TA\$^TAGTCGAGGCTT
A\$^TAGTCGAGGCTTT
\$^TAGTCGAGGCTTTA

2. Sort

AGGCTTTA\$^TAGTCG
AGTCGAGGCTTTA\$^T
A\$^TAGTCGAGGCTTT
CGAGGCTTTA\$^TAGT
CTTTA\$^TAGTCGAGG
GAGGCTTTA\$^TAGTC
GCTTTA\$^TAGTCGAG
GGCTTTA\$^TAGTCGA
GTCGAGGCTTTA\$^TA
TAGTCGAGGCTTTA\$^
TAS\$^TAGTCGAGGCTT
TCGAGGCTTTA\$^TAG
TTA\$^TAGTCGAGGCT
TTTA\$^TAGTCGAGGC
^TAGTCGAGGCTTTA\$
\$^TAGTCGAGGCTTTA

3. Select final
column

GTTTGCAGAA^TGTC\$A

Számítástudományi trükkök: Burrows-
Wheeler transzformáció (DEC
Technical Report, 1994)
Adattömörítéshez találták ki, bzip

^TAGTCGAGGCTTTAGATCCGATGAGGCTTAGAGACAG\$

Genomic sequence

GGTTGGTCGATTGGAATCACGGAATAATT^AGATCC\$G

Transform

A biológia a században

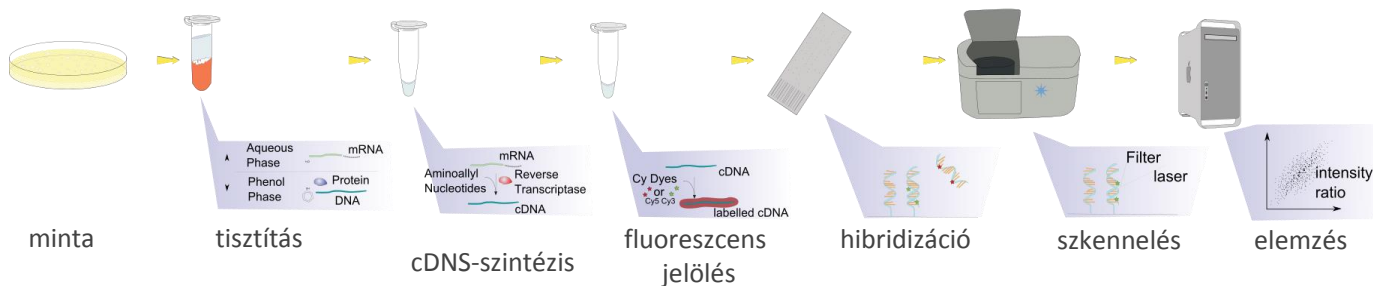
20.



21.

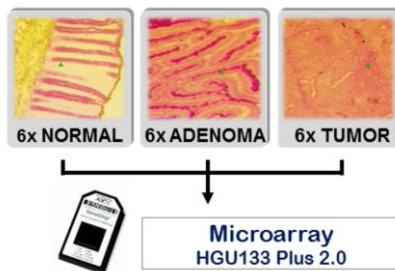


„Big Data” példa (2009)



ELTE-SOTE-3DHISTECH

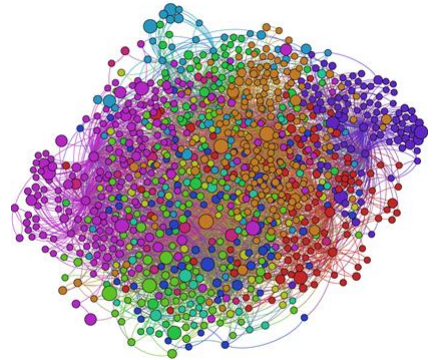
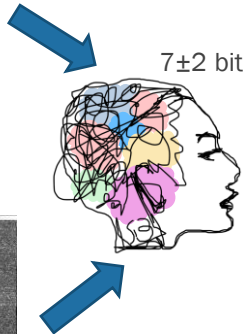
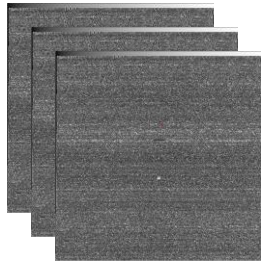
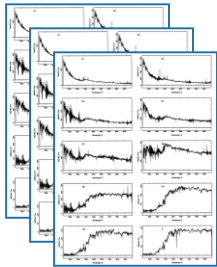
- Affymetrix HG U133 Plus2
- Nyers adat: 67 Mpix-es kép (fotometria!)
- 604 258 probe
- 54 675 probe halmaz (~gén)
- 207 minta (végbélrák)
- Feladat: rák “marker gének”



Feldolgozott adatok:
54675D vektorok + metaadatok

Hasonló problémák ... hasonló elemzés

- SDSS spektrumok: 1 millió db **3000 dimenziós** vektor
- Microarrayek: 207 db **54 675 dimenziós** vektor
- Humán genom: 3,2 Gb



Tömörítés: dimenzióredukció,
mátrixfaktorizálás, gépi tanulás

Dimenziócsökkentés – nem felügyelt tanulás

$$pV = NkT$$

$$6 \cdot 10^{23} \rightarrow 5$$

A háttérben meghúzódó fizikai törvényszerűségek miatt az adatvektorok nem töltik ki a teljes teret, hanem alacsonyabb dimenziójú felületeken/altéren helyezkednek el. Ennek köszönhető, hogy esélyünk lehet megérteni a világot!

vetítés

tömörítés

modell



SIGGRAPH ASIA 2009

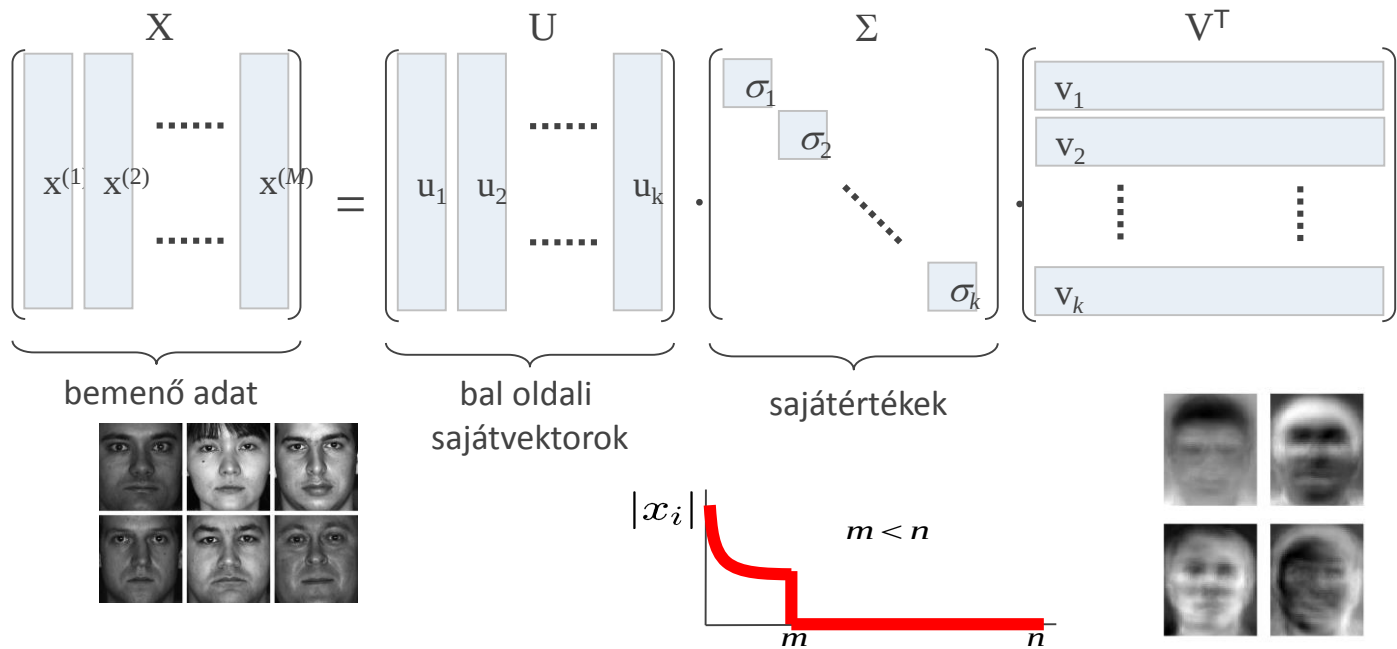
Shadow Art

Niloy J. Mitra
IIT Delhi / KAUST

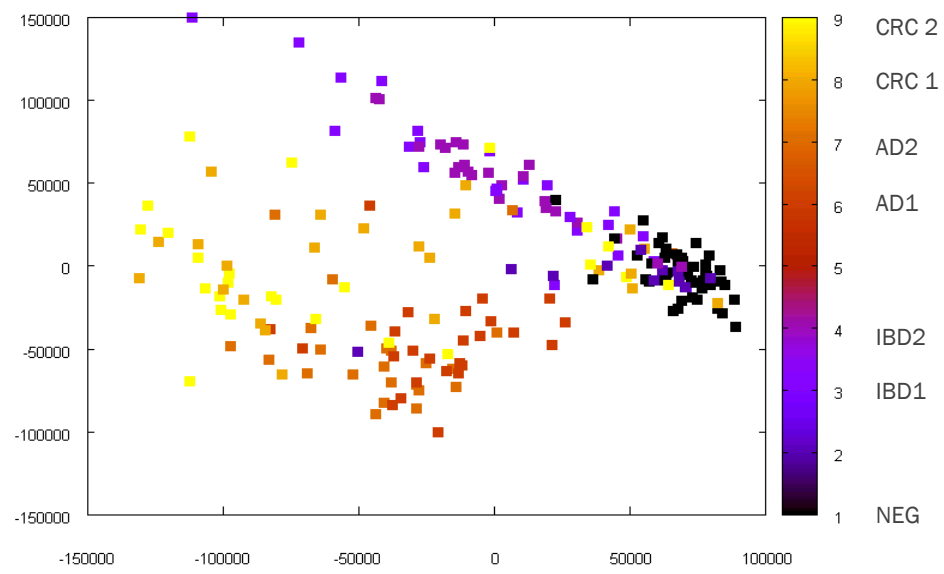
Mark Pauly
ETH Zurich

Lineáris vetítés – PCA, SVD

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

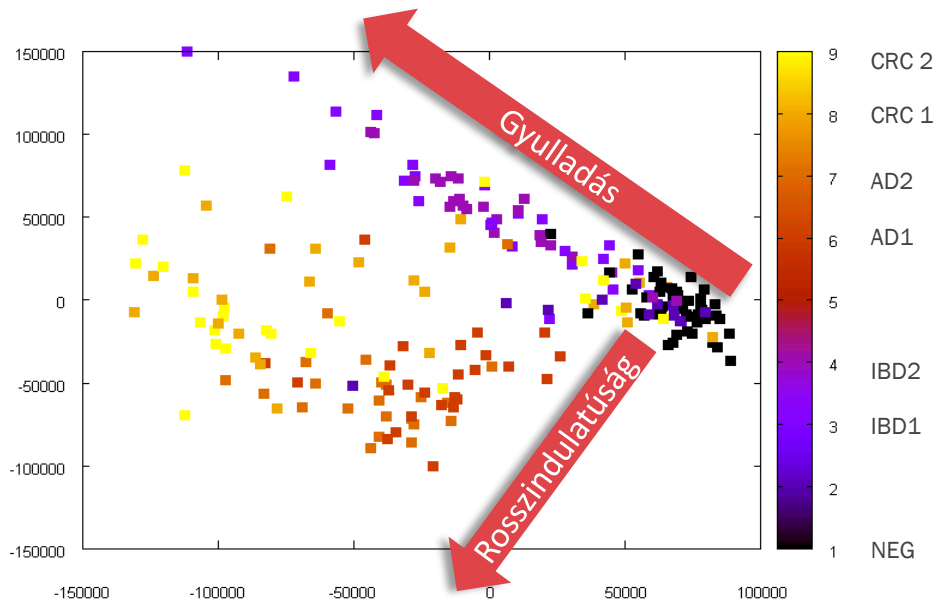


Génexpresszió microarray: 54675D → 2D

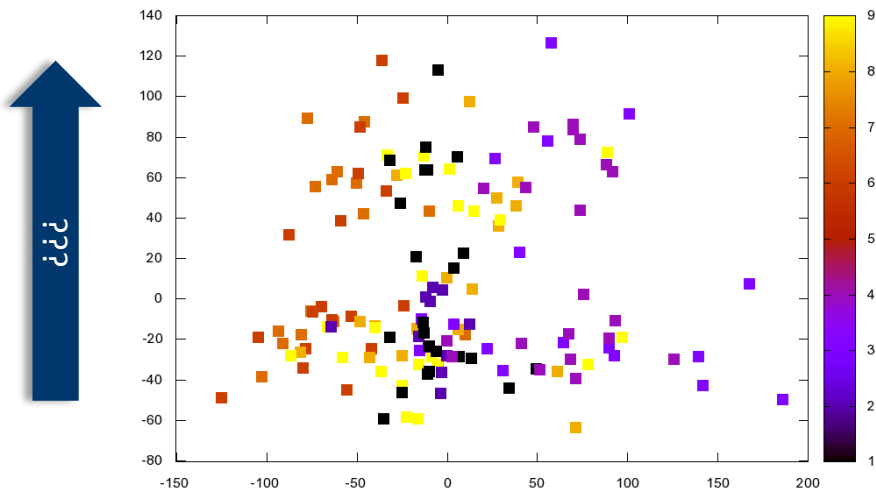


Génexpresszió microarray: 54675D → 2D

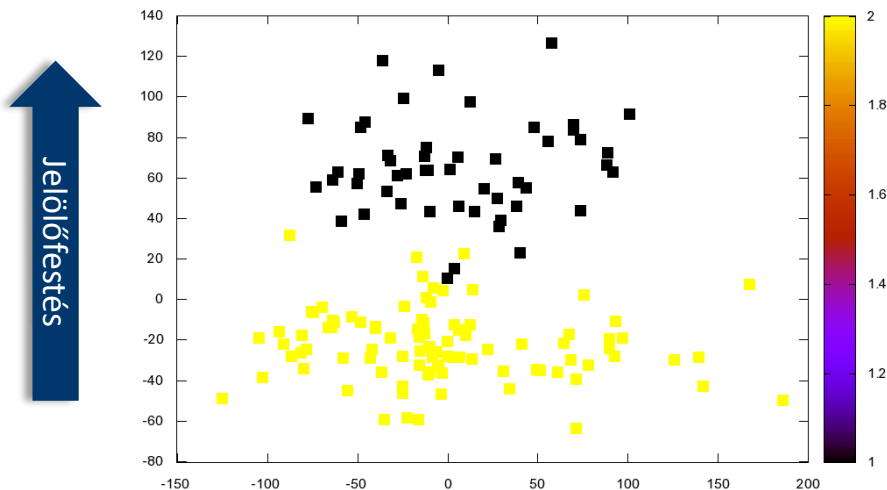
Önállóan „fedezett” fel összefüggéseket az adatokban



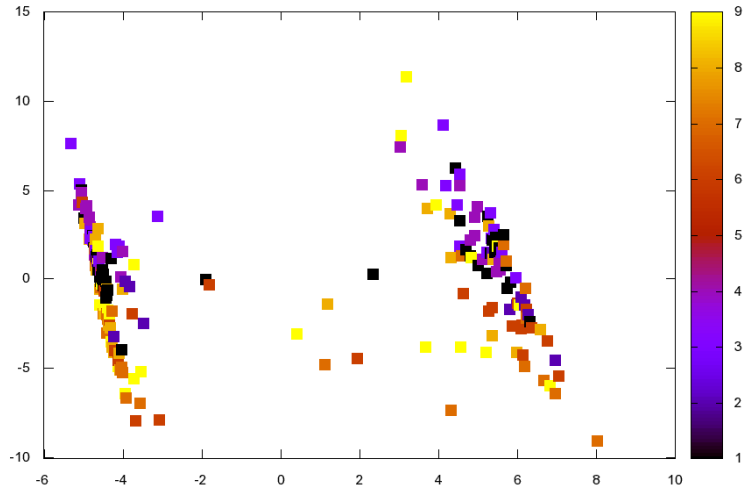
Génexpresszió microarray: PCA₂, PCA₃ klaszterek?



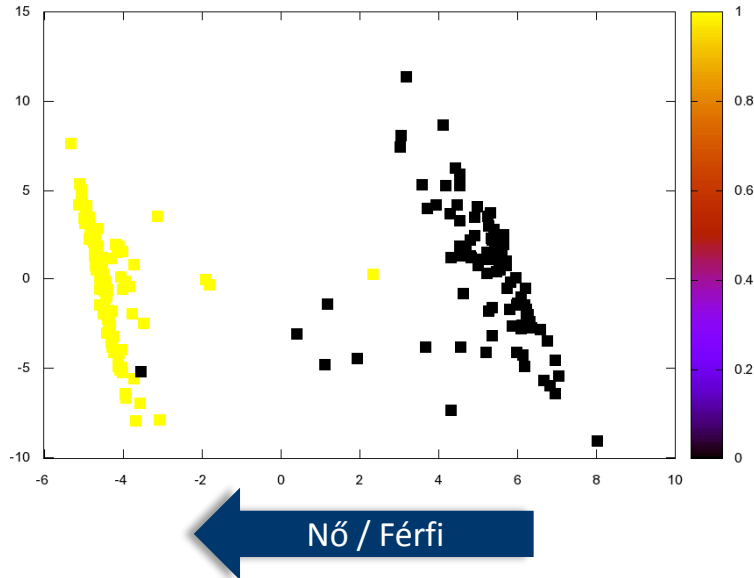
Génexpresszió microarray: PCA₂, PCA₃ klaszterek?



Génexpresszió microarray: KEGG „riboszóma alhálózat”

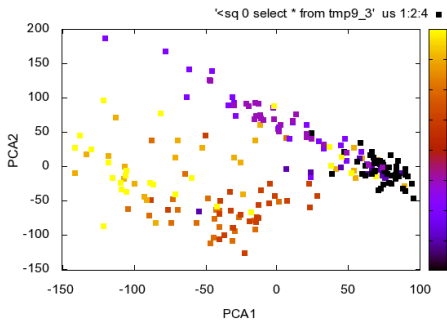


Génexpresszió microarray: KEGG „riboszóma alhálózat”

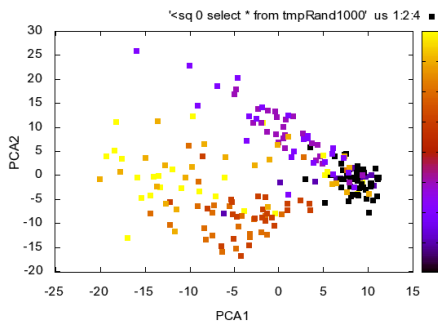


Komplex rendszerek

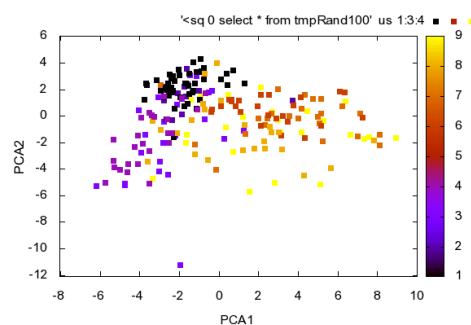
„Vegyük észre, hogy minden mindennel összefügg” – Leonardo da Vinci



minden probe (54 675)



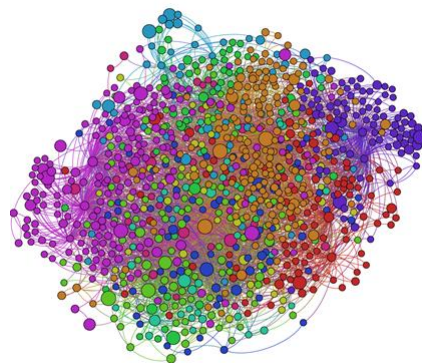
véletlen 1000



véletlen 100

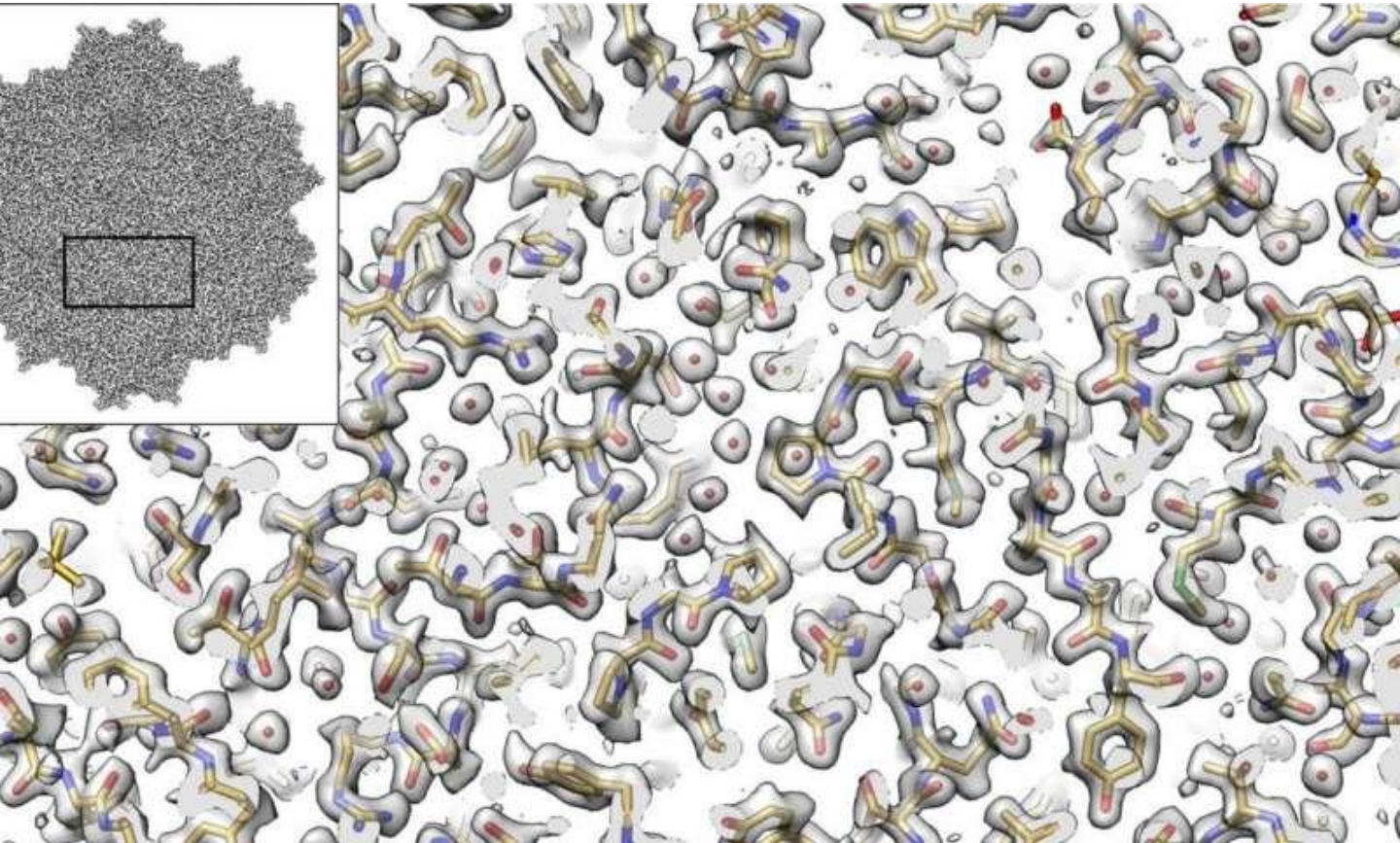
„Marker gének”?

Komplex, nemlineárisan kölcsönható hálózat!



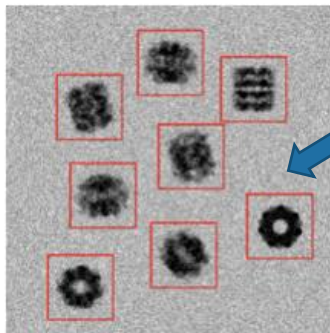
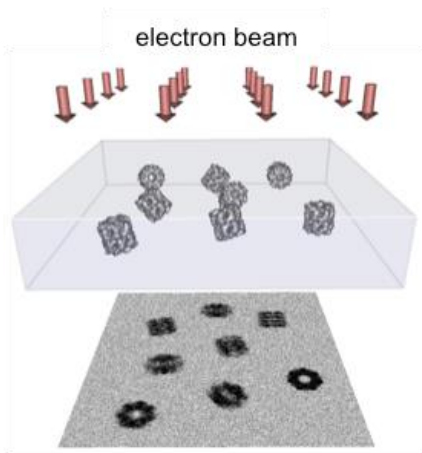
Képi információ

Cryo-elektronmikroszkópia: molekula felbontás 3D



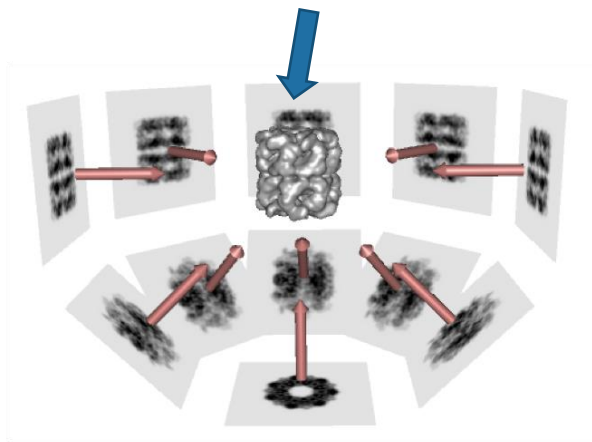
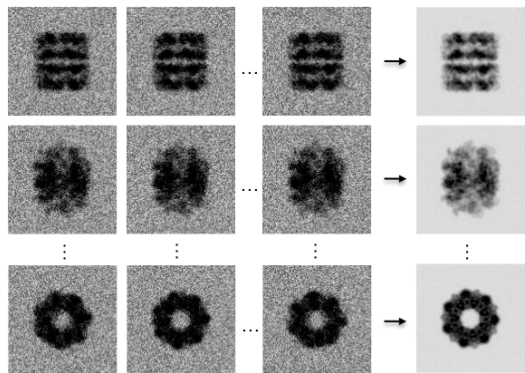
Y.Z Tan et al. Sub-2 Å Ewald curvature corrected structure of an AAV2 capsid variant, Nature Communications; 9:3628 (2018)

„Big data” elemzés a háttérben

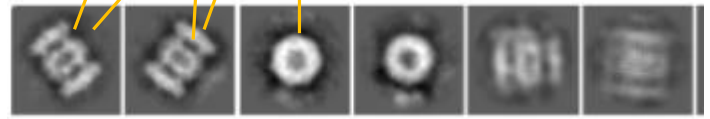
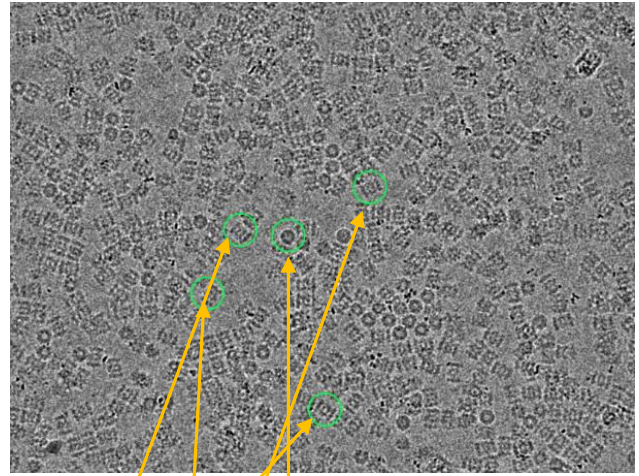
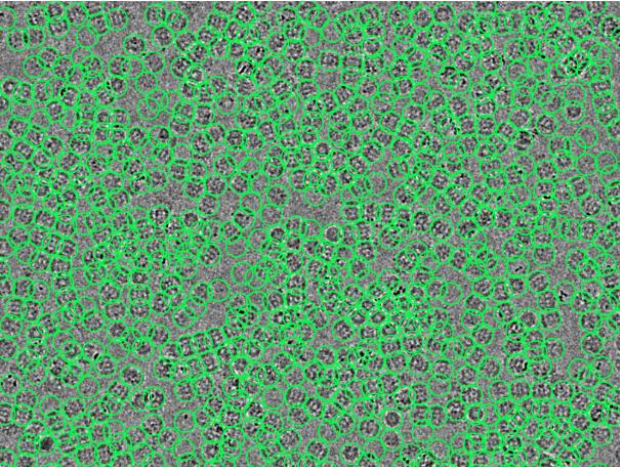
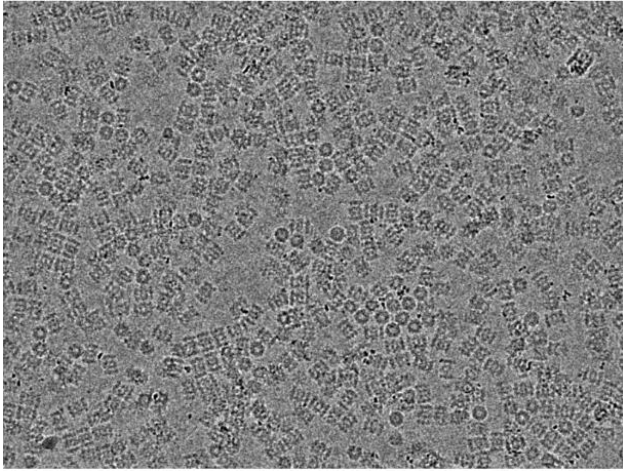


Sok ezer felvétel:
képfelismerés, szegmentáció

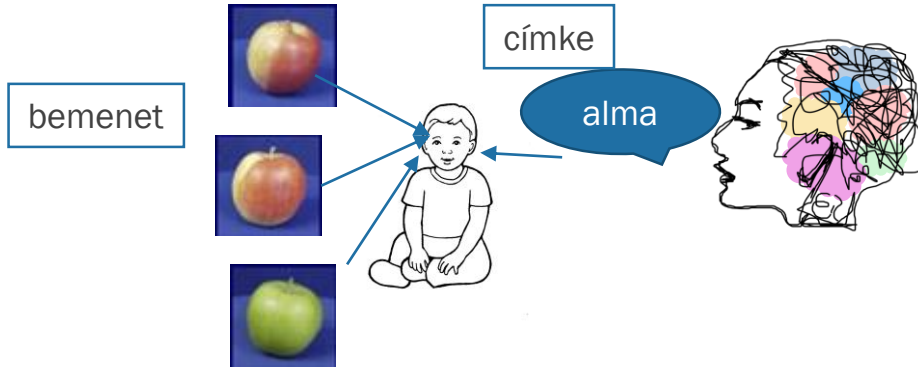
2D vetületekből 3D rekonstrukció
+ molekuladinamikai modellezés



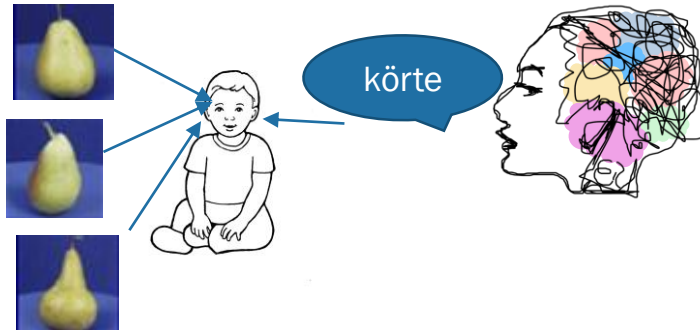
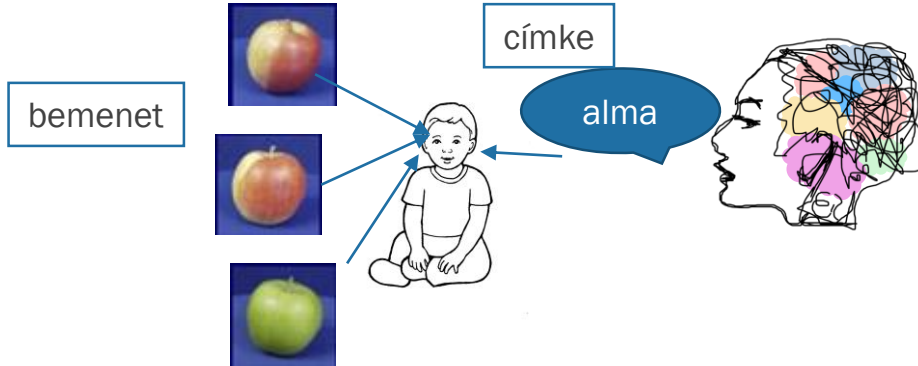
Kézi módszerekkel kezelhetetlen: gépi tanulás kell



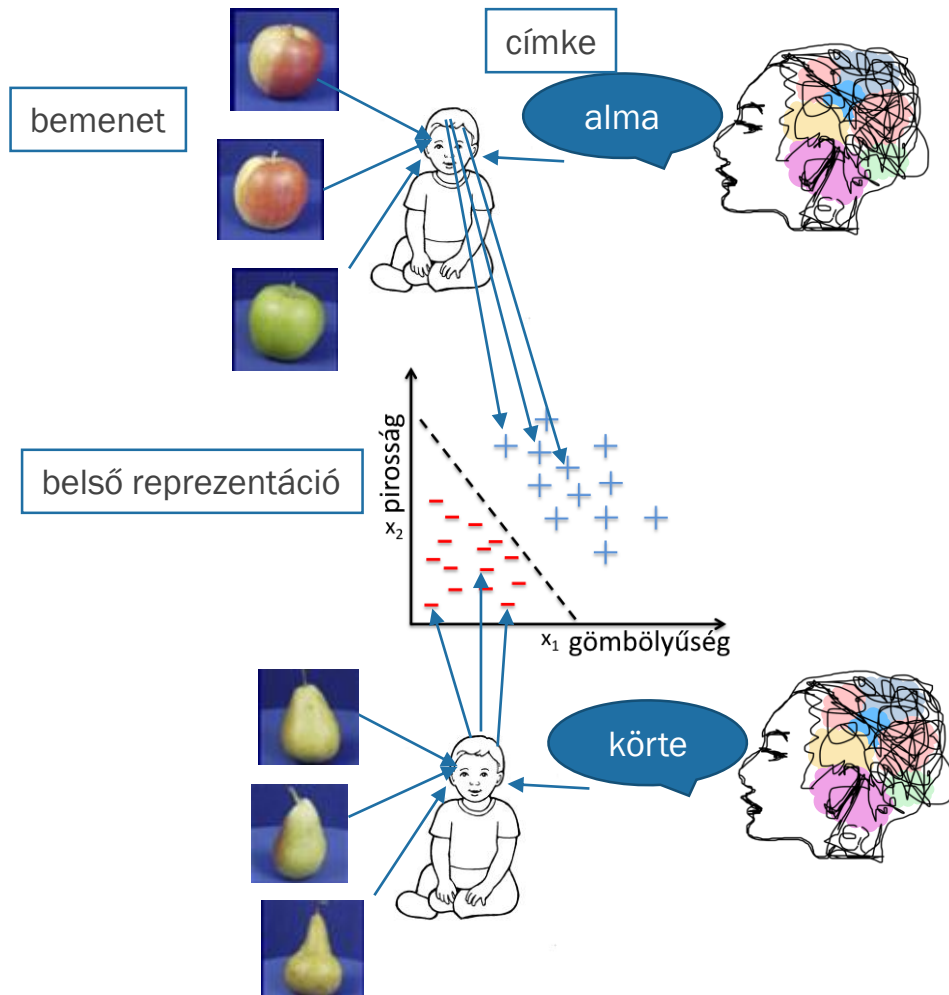
Felügyelt tanulás (supervised learning)



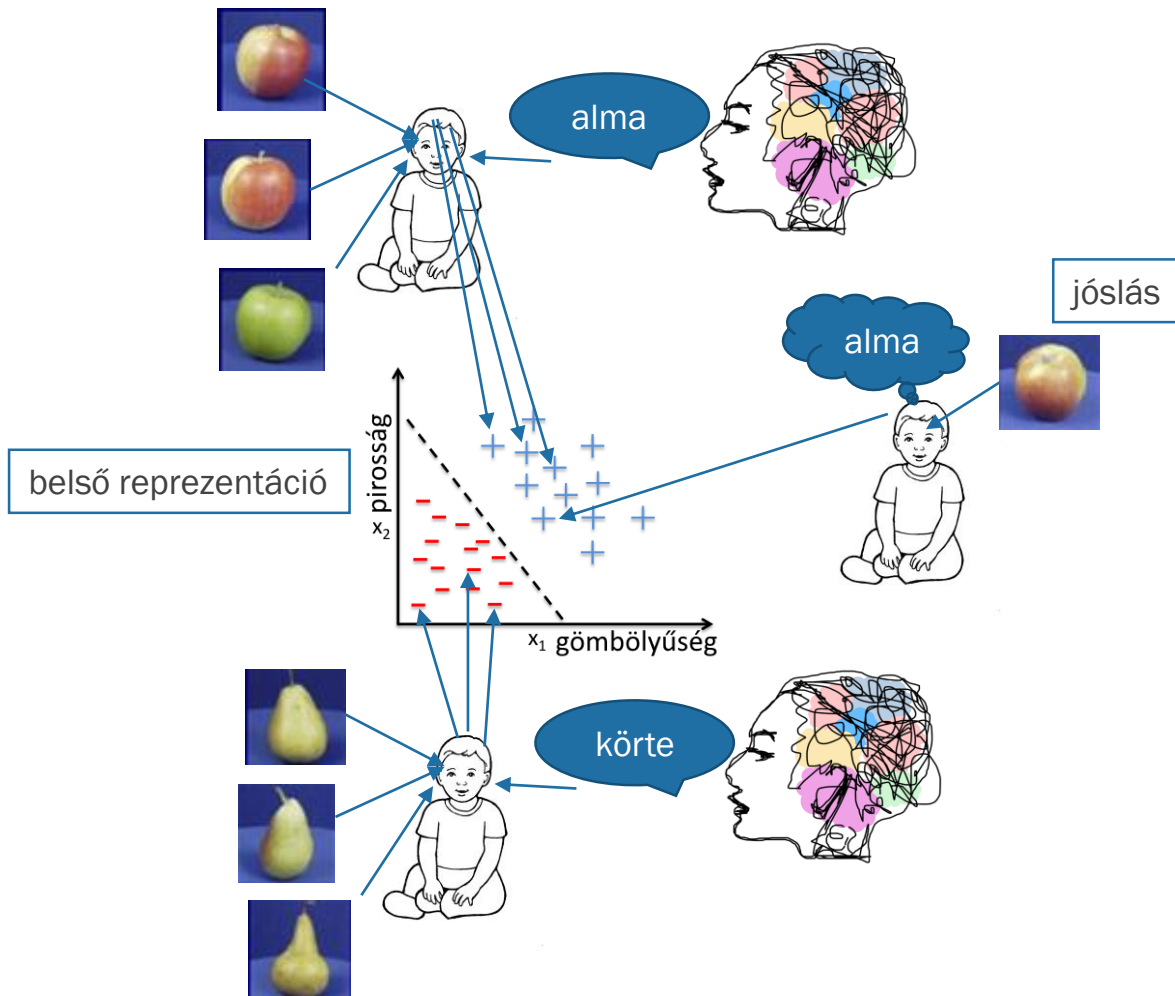
Felügyelt tanulás (supervised learning)



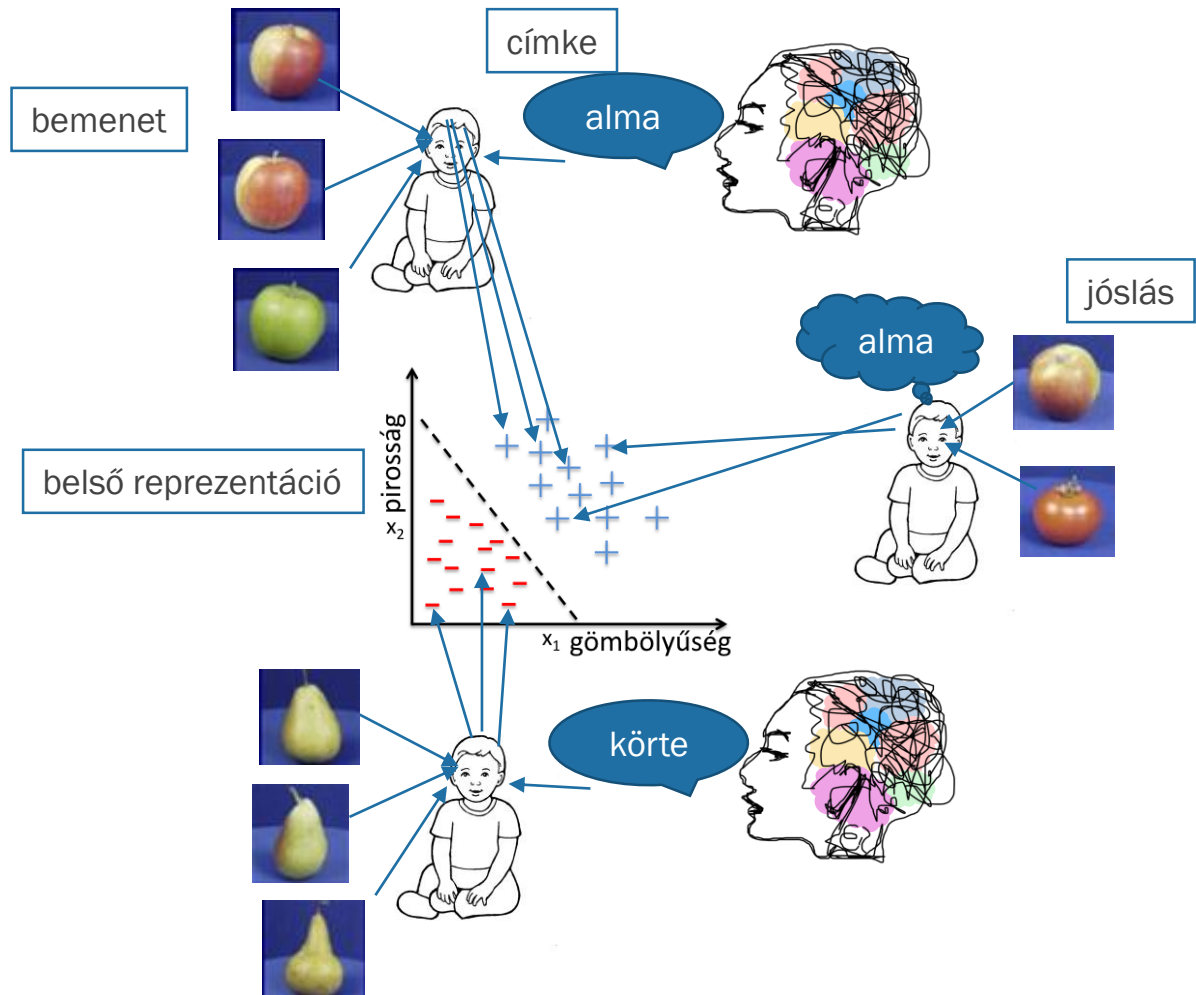
Felügyelt tanulás (supervised learning)



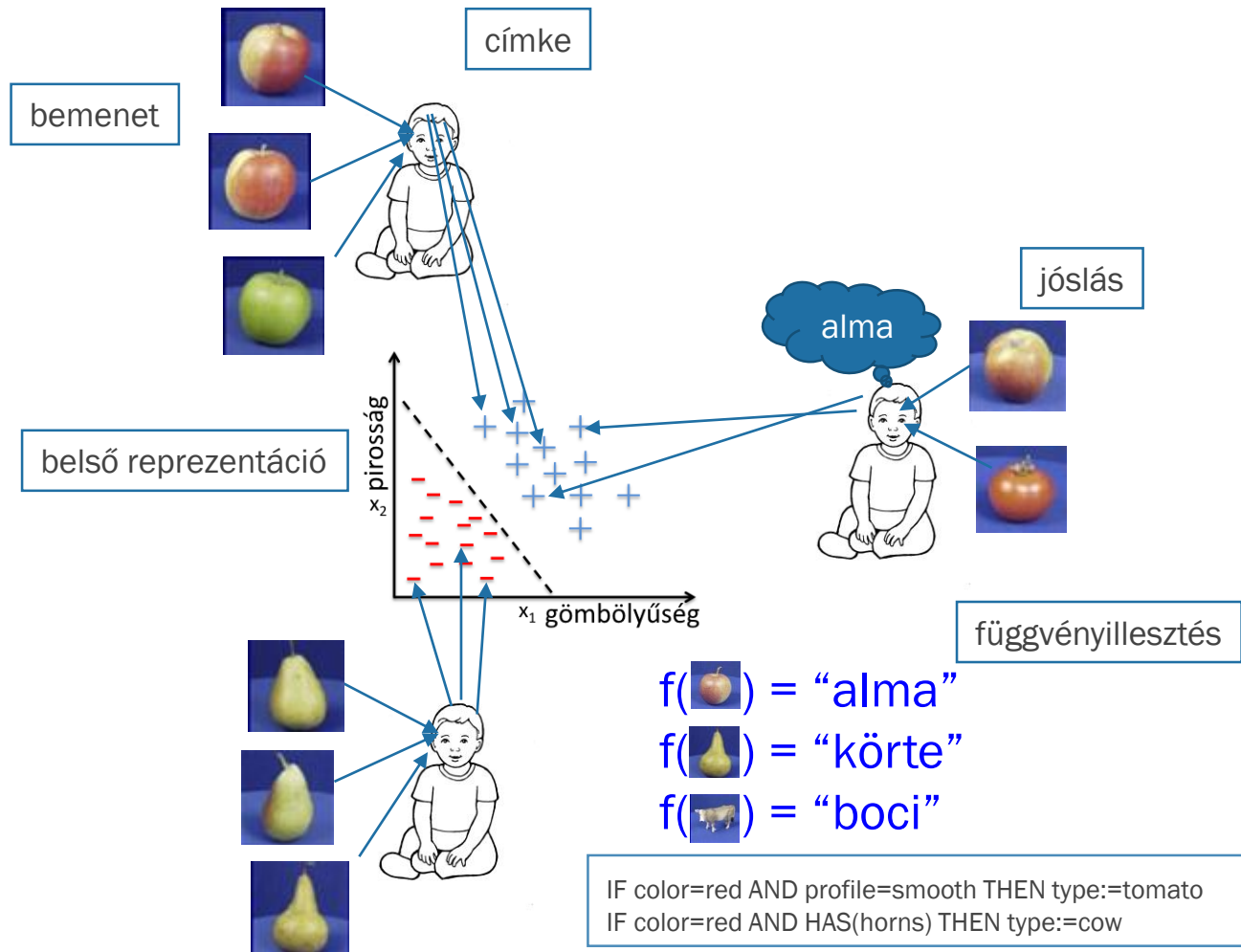
Felügyelt tanulás (supervised learning)



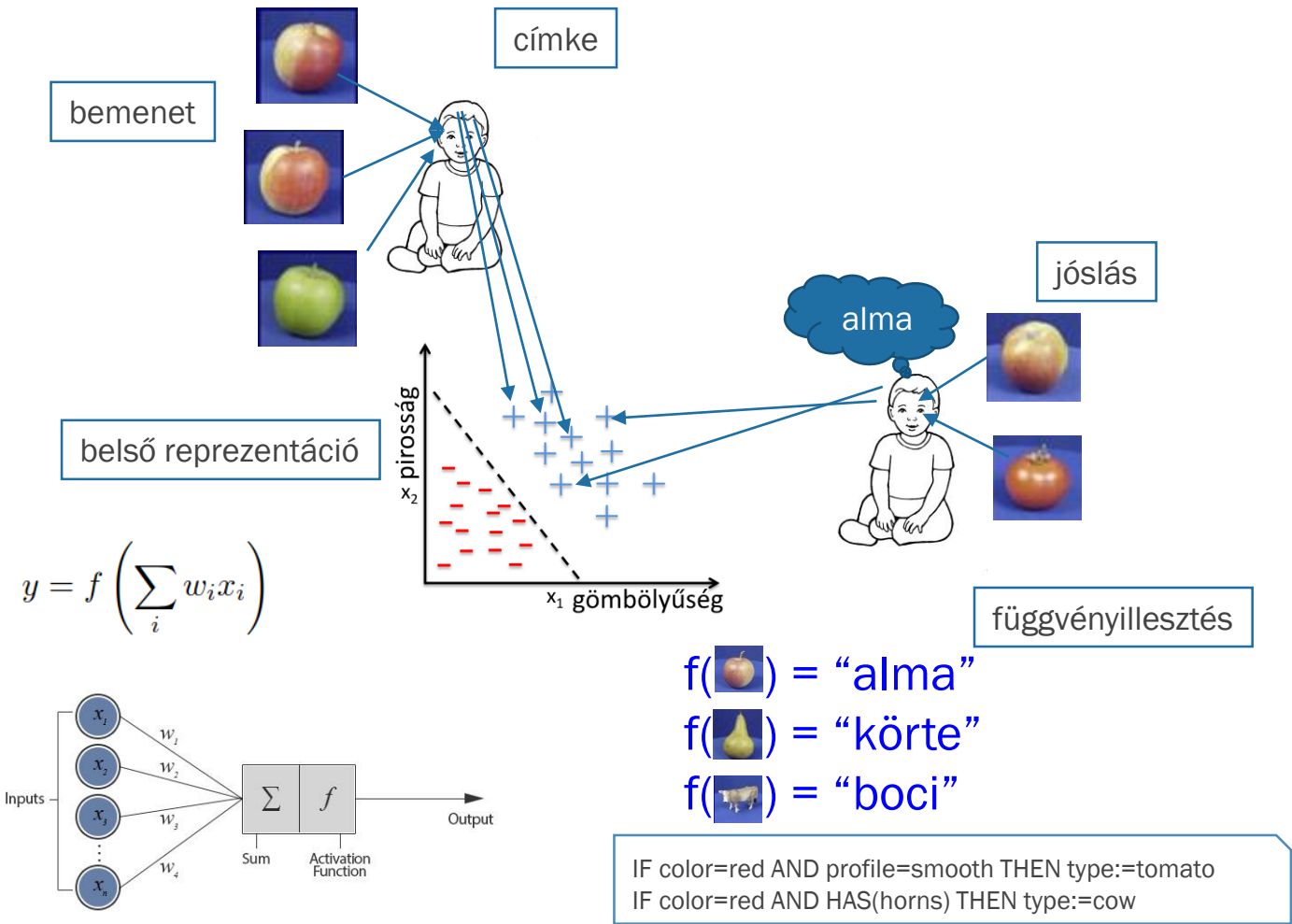
Felügyelt tanulás (supervised learning)



Felügyelt tanulás (supervised learning)



Felügyelt tanulás (supervised learning)



Training

Training
Images



Image
Features



Training
Labels



Training



Learned
model

Testing



Test Image



Image
Features



Learned
model



Prediction

Slide credit: D. Hoiem and L.
Lazebnik

Training

Training Images



Image Features

Training Labels

Training

Learned model

pixel colors

Multidimensional vector space

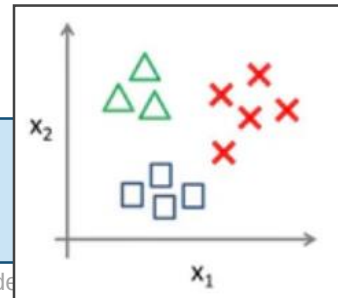
Testing



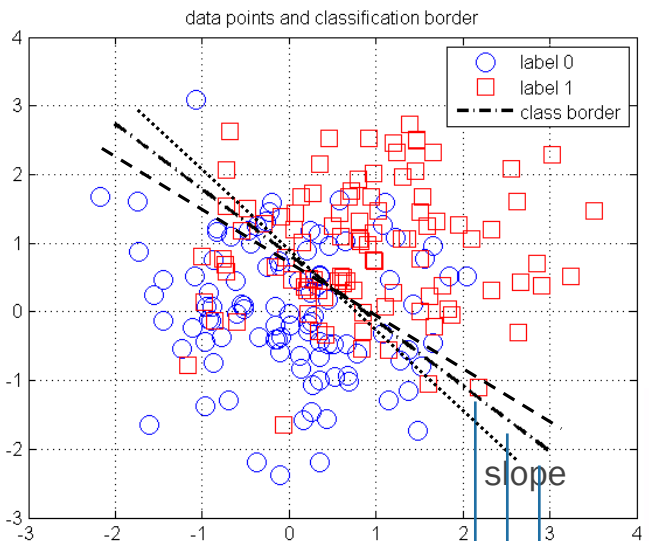
Test Image

Image Features

Learned model

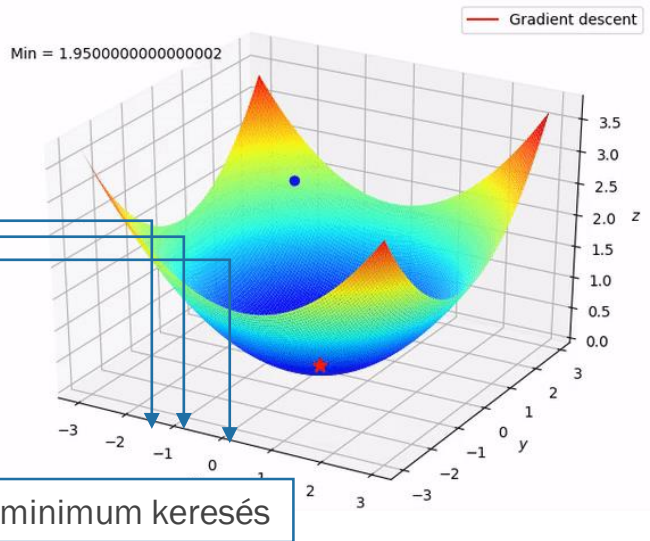


Költség-függvény optimalizáció



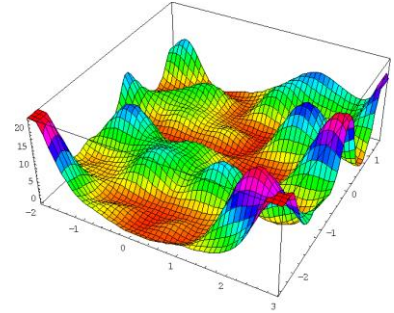
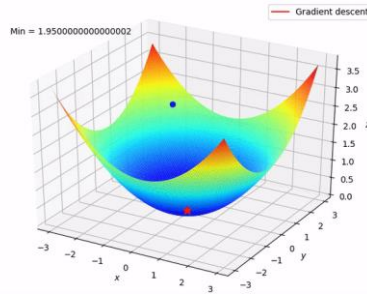
Költség = hibás osztályozások száma

Képek -> pontok
N dimenziós térben



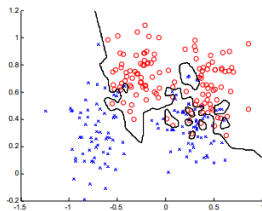
Nehézségek

- Megfelelő méretű, minőségű tanító halmaz
- Jó reprezentáció (képek, szavak, ... -> vektortér)
- Nemlineáris optimalizáció
- Modell komplexitás
 - Pontosság
 - Általánosítási képesség
- ...

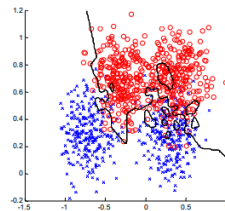


Training data

Testing data



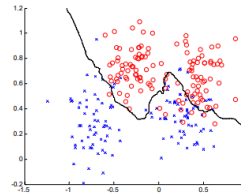
error = 0.0



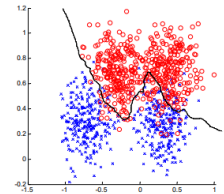
error = 0.15

Training data

Testing data



error = 0.1120



error = 0.0920

Bármely kellően fejlett technológia megkülönböztethetetlen a mágiától.

(Arthur C. Clarke)

Valóban, a mechanika törvényeinek megismerése lehetővé tette, hogy egyszerű gépekkel olyan tárgyakat emeljünk fel, amelyeket emberi erővel lehetetlen. A termodinamikai ismeretekre épülve képesek lettünk kontinenseket és óceánokat átszelni, és mindenki garázsában ott van a „hétmérföldes csizma”. Az elektromosság és kvantummechanika törvényei elhozták a villamosítást és az internetet, a mobiltelefon „varázstükrével” pedig távolba láthatunk és hallhatunk. A XXI. század hajnalára sikerült a humán genom 3 milliárd nukleotidból álló szekvenciáját leolvasni.

Ma már egy teljes emberi genom szekvenálása egy hét alatt elkészülhet, kevesebb mint ezer dollárért.

Pár év óta rendelkezésre áll olyan módszer, amellyel akár nukleotidonként módosítható a genom.

Milyen csodákra lehet képes ezekkel a technológiákkal felvértezve a biológia és az orvostudomány?

Megérthetjük-e az élet működését, eltörölhetjük-e a betegségeket, kitolhatjuk-e az egészséges emberi kor határát?

Milyen kihívásokat hoznak az új adatok és az új ismeretek?



Ki fogja megfejteni az **Univerzum** titkait, megtalálni az **Élet** eredetét?
Ki fogja meggyógyítani a rákot, ki fog segíteni **Mindent** megérteni?

♦ **A:** Fizikusok.

♦ **B:** Informatikusok.

♦ **C:** Orvosok.

♦ **D:** Mi, közösen!

**ÚJ TUDOMÁNYOS MÓDSZERTAN:
ÚJ TUDÓSOK KELLENEK,
AKIK ÉRTIK A SZAKTUDOMÁNYOKAT, ÉS
PROFESSZIONÁLISAN KEZELIK A MATEMATIKAI
MELLETT AZ INFORMATIKAI ESZKÖZTÁRAT IS.**



Csabai István

ELTE Komplex Rendszerek Fizikája Tanszék

csabai@elte.hu

<http://complex.elte.hu/~csabai/>