Data-intensive genomics

ISTVAN CSABAI

PROFESSOR OF PHYSICS

ELTE EÖTVÖS LORÁND UNIVERSITY

DEPT. OF PHYSICS OF COMPLEX SYSTEMS

DATA INTENSIVE SCIENCES AND MACHINE LEARNING GROUP

Acknowledgement: ELTE FIEK, SOTE TKP,
NKFIH NVKP, H2020 VEO

History of (machine) intelligence / data science

Model

OBSERVATION

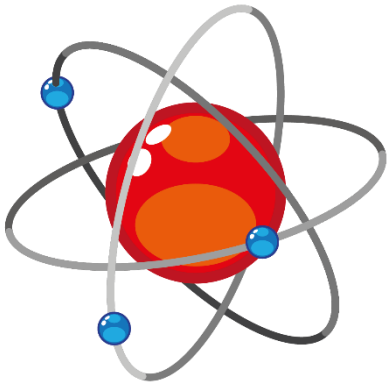
World

History of (machine) intelligence / data science

Model

OBSERVATION

World



History of (machine) intelligence / data science

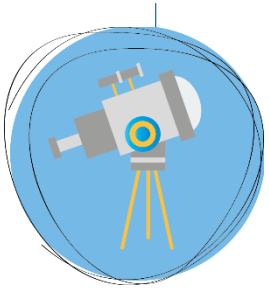
crutch for senses

Model

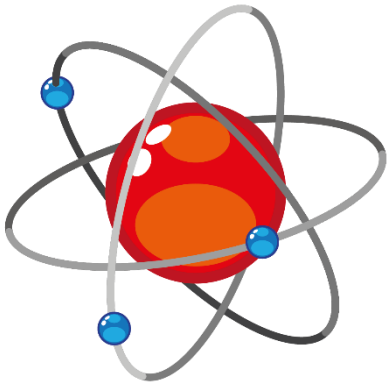
OBSERVATION

DATA !

Instruments



World



Natural intelligence

7±2 bit

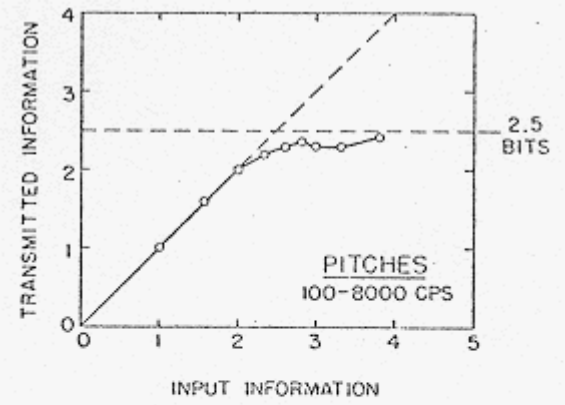


FIG. 1. Data from Pollack (17, 18) on the amount of information that is transmitted by listeners who make absolute judgments of auditory pitch. As the amount of input information is increased by increasing from 2 to 14 the number of different pitches to be judged, the amount of transmitted information approaches as its upper limit a channel capacity of about 2.5 bits per judgment.

G.A. Miller *The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information*, Psychological Review, 63, 81-97. (1956)

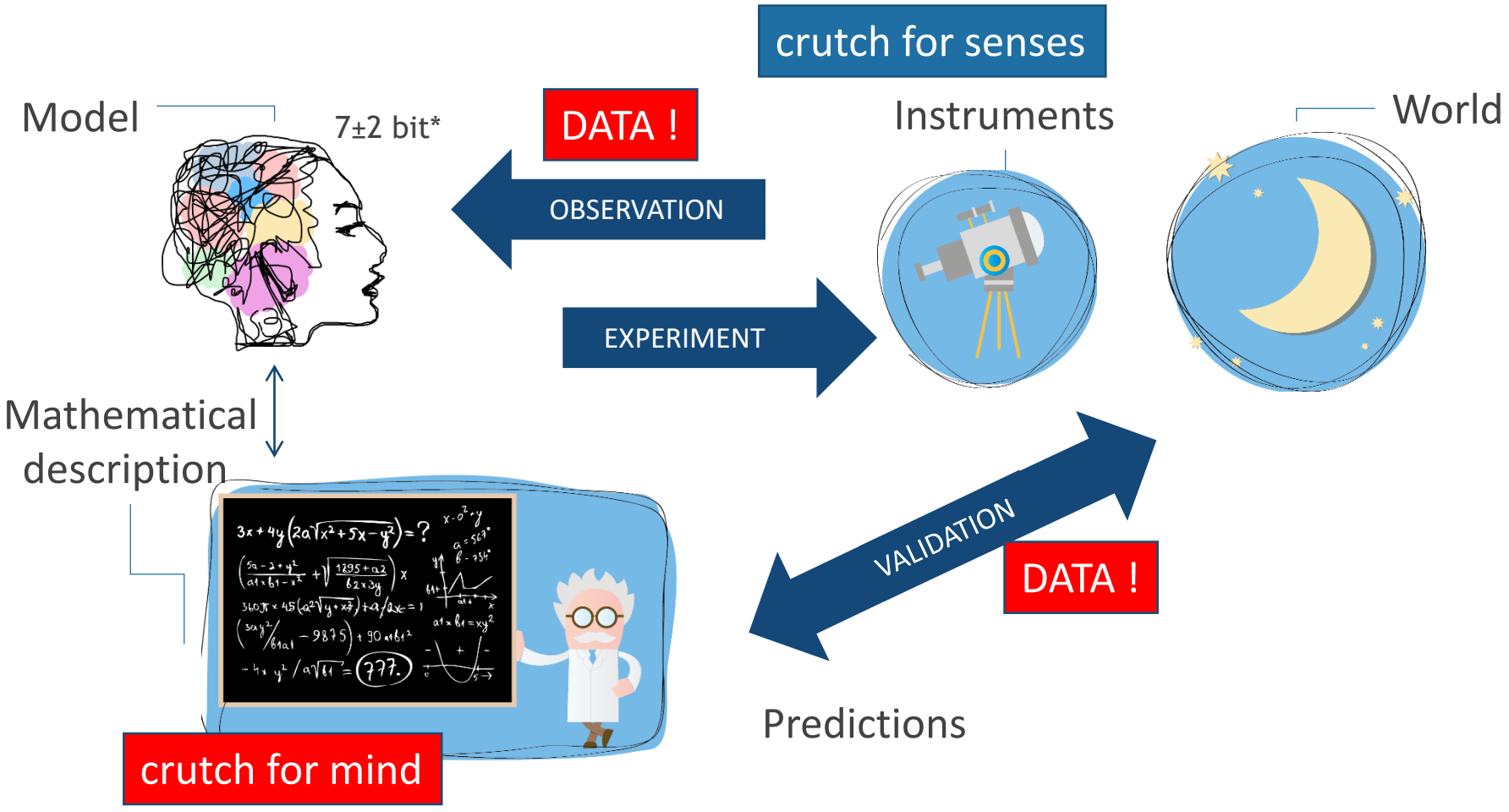
Pollack, I. *The information of elementary auditory displays.*
J. Acoust. Soc. Amer., 1952, 24, 745-749.

Homo Sapiens: Technical Specifications	
CPU	100 GN (giga-neurons)
Clock frequency	**4-32 Hz**
CPU cores	1 (male version), 2+ (female v.)
CPU speed	0.1 Flops (floating point op. / sec)
Memory (short term)	**7 +/-2 bits**
Storage	1TB-2.5PB
Power	20 W
Camera	576Mpix, 24Hz
Touch	Yes
Display	No
Speakers	Mono
GPS	No
WIFI	No
Bluetooth	No
2G/3G/4G/5G	No/No/No/No
Latest version update	100 000 BC

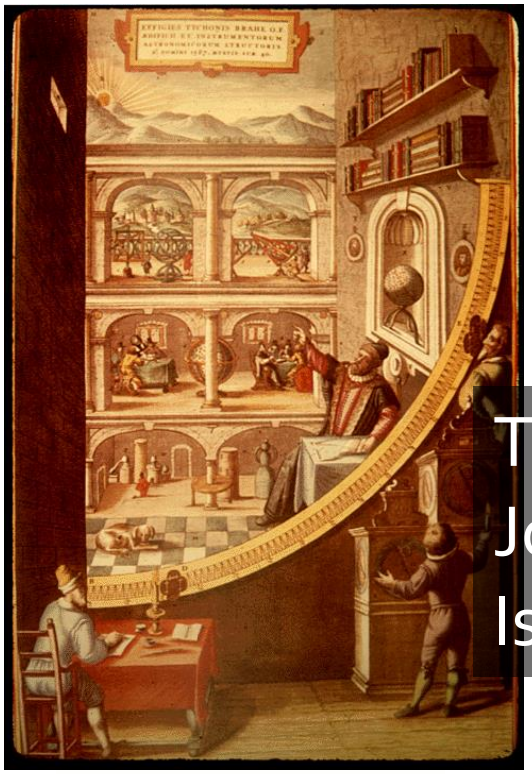
Main Features :

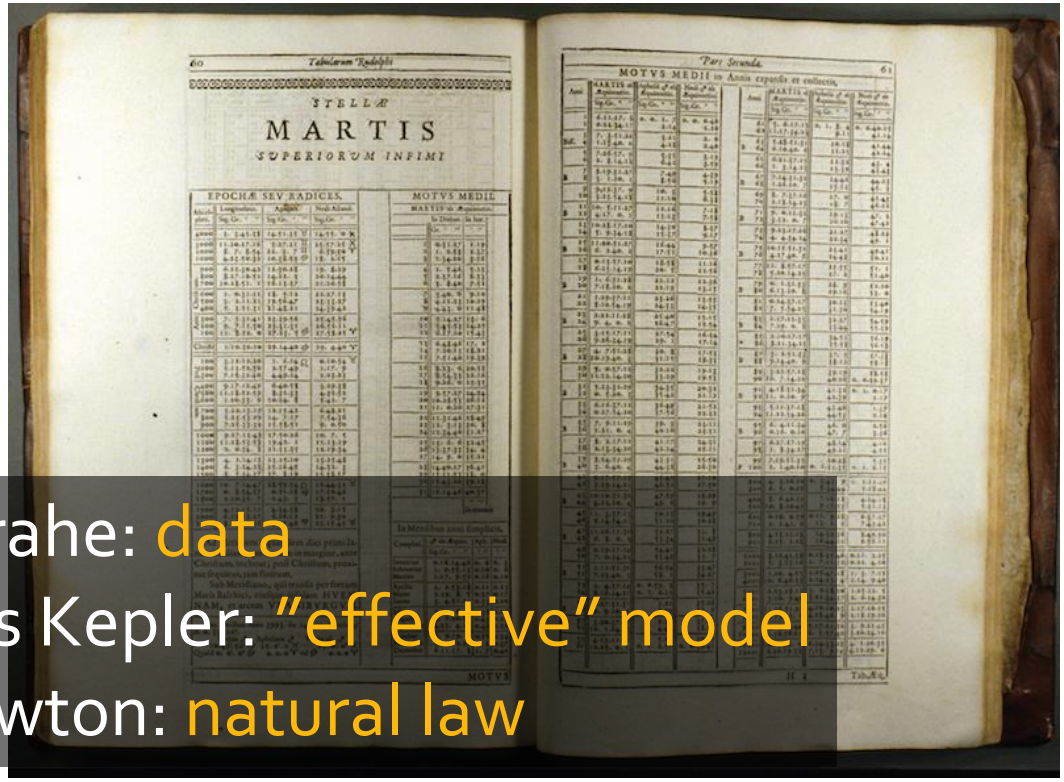
- Find food
- Escape predators
- Kill enemies
- Find mate and reproduce

History of (machine) intelligence / data science



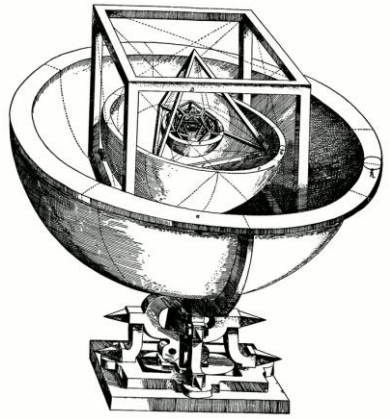
First "Data Science"

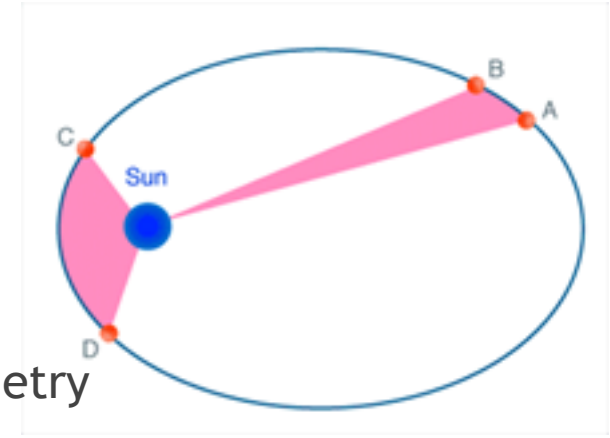




Tycho Brahe: data
Johannes Kepler: "effective" model
Isaac Newton: natural law

Tabulae Rudolphinae (1627), 23 years, position of 1405 stars + planets

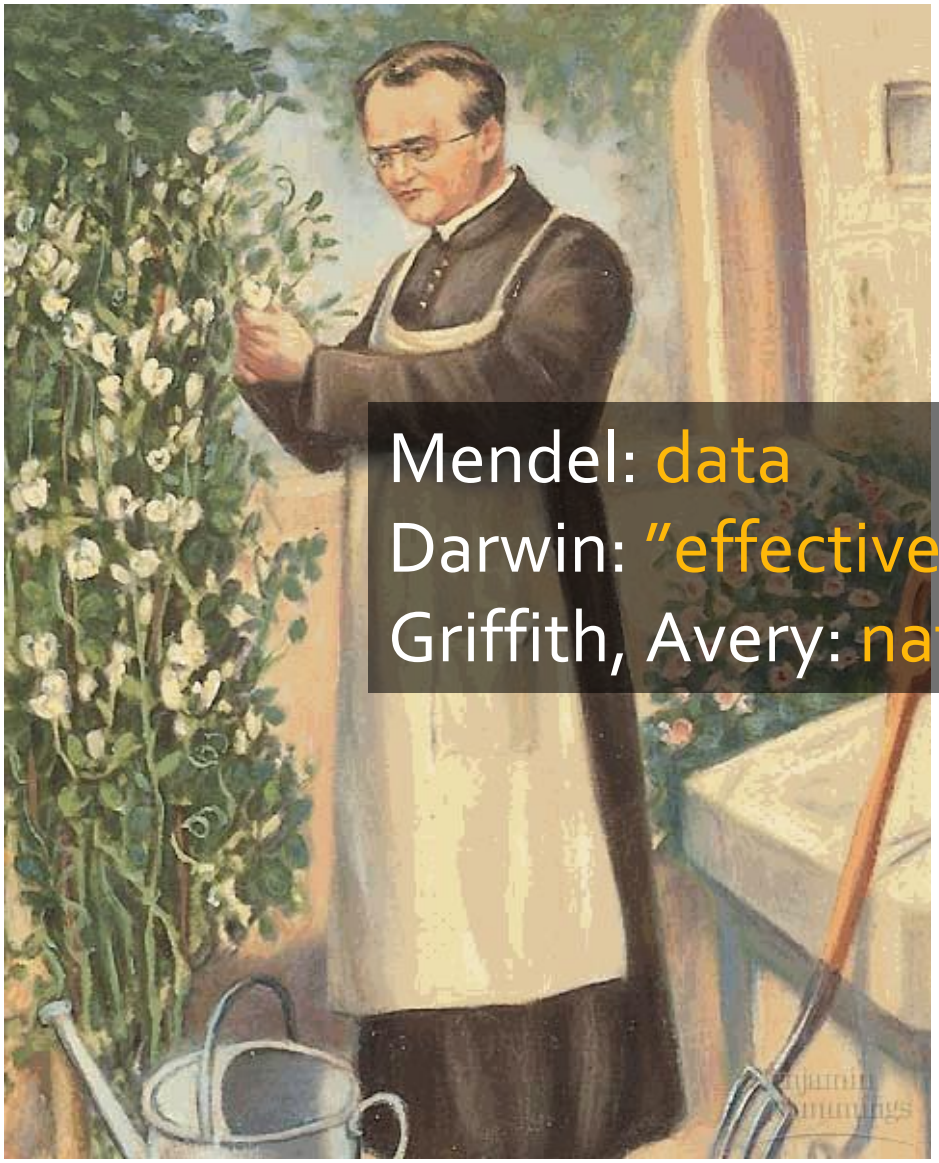




$$F = G\frac{m_1 m_2}{r^2}$$

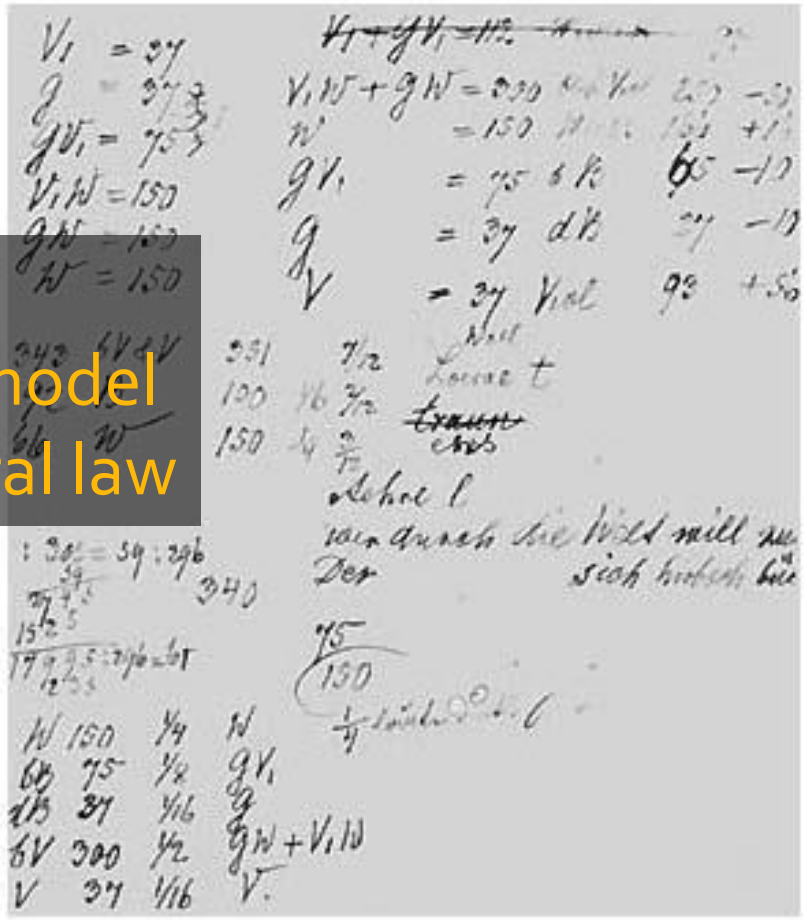
Perfect beauty and symmetry

First "Data Science" in genetics



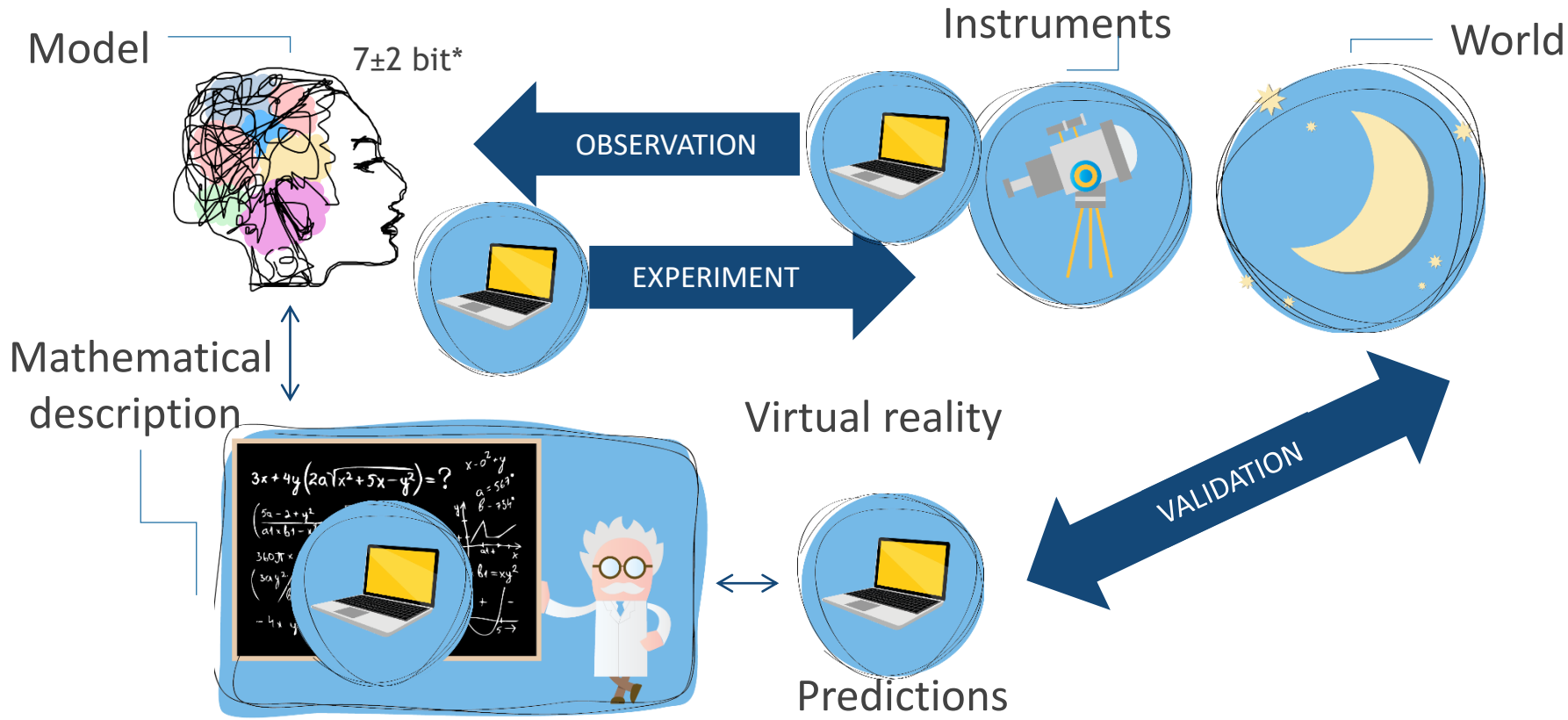
Mendel: data
Darwin: "effective" model
Griffith, Avery: natural law

Gregor Mendel, 1865
8 years, ~28.000 pea plants



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

History of (machine) intelligence / data science



Initial values

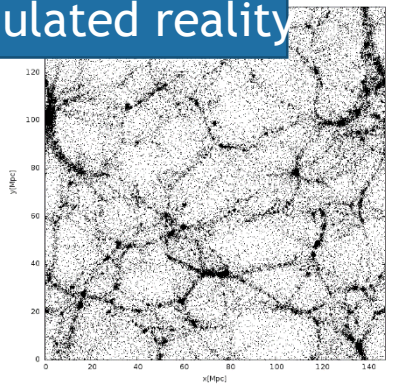
“laws”, equations

Simulated reality

Λ=0.7
Ω_m=0.3

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

$$R_{\mu\nu} - \frac{1}{2}R\,g_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}$$



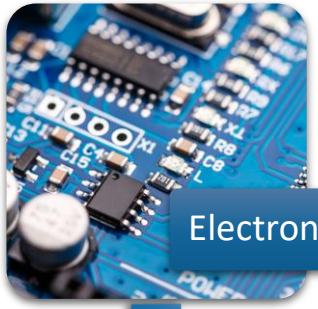
Science – technology – science – technology …

Astronomy → Mechanics → Quantummechanics →

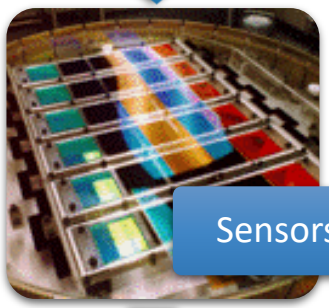
→ Solid state physics → Transistor/ microelectronics

Moore's-law → Better computers

→ Better sensors more data

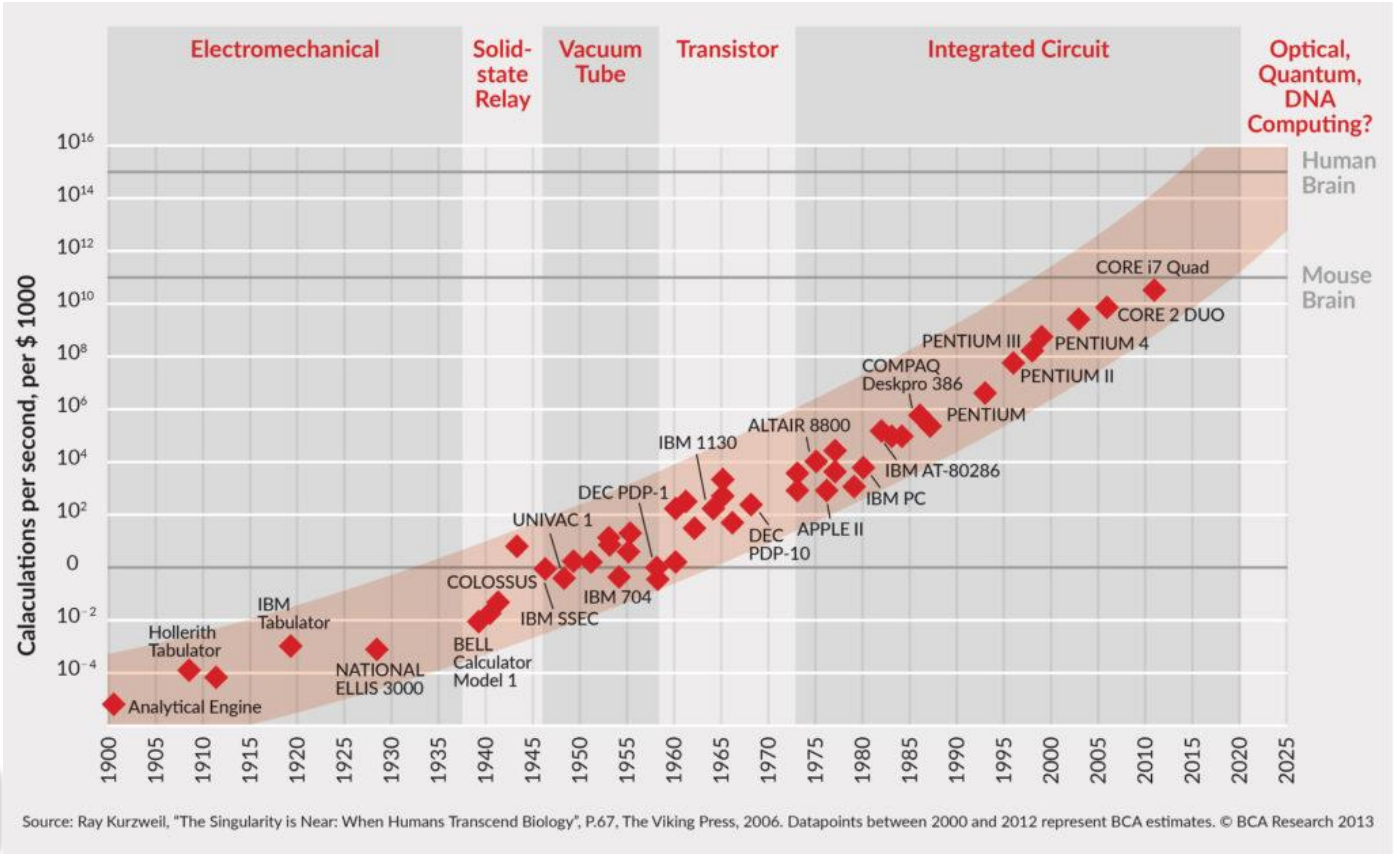


Electronics



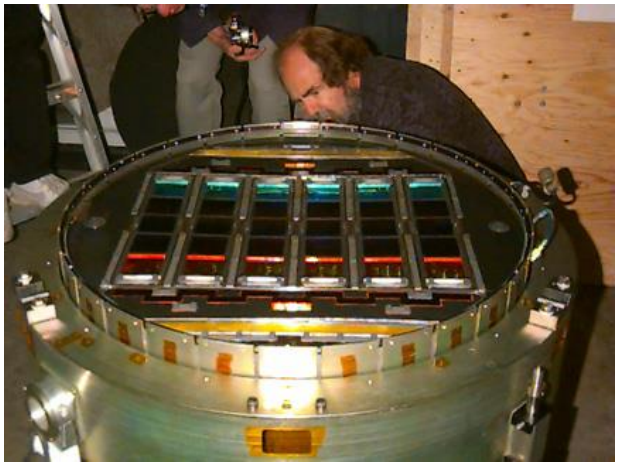
Sensors

Data



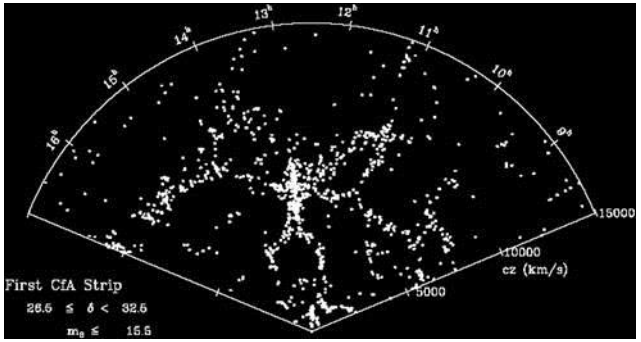
2.5m

120Mp – 2.5Tp

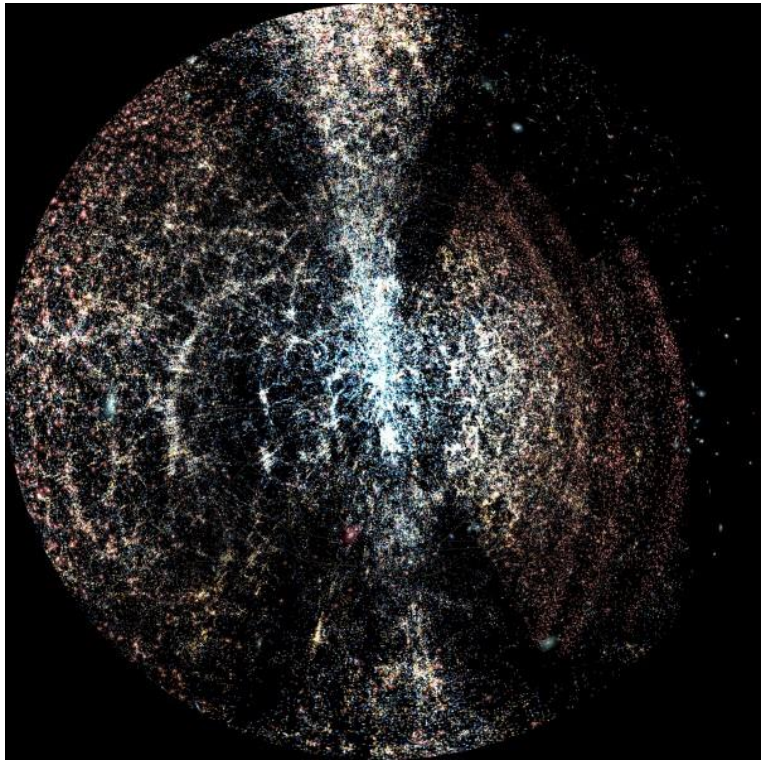


5 years:10TB

CfA 1989: 1100 galaxies



SDSS 2005: 1M galaxies

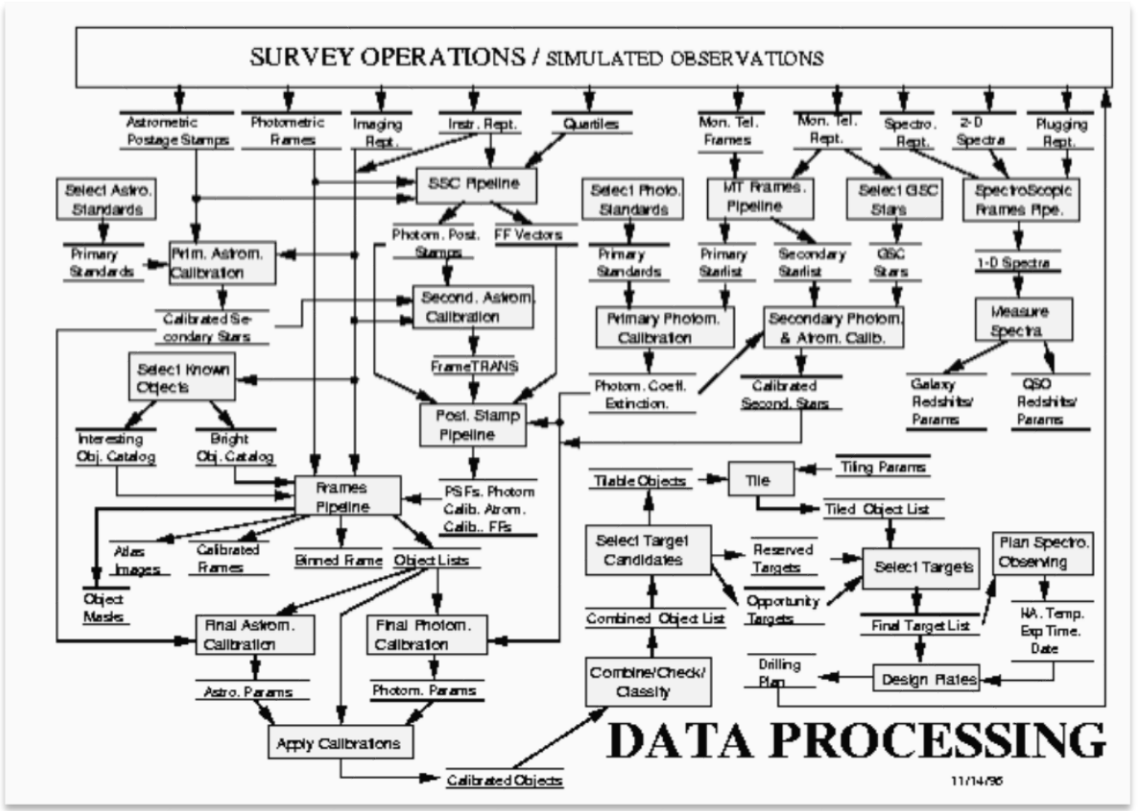


Prototype of modern data science

SDSS: 3D map of the universe

Data processing challenge

- Automatic pipeline
  - *More than 150 man year development*
  - *First astro project where most of the money is spent on software rather on the telescope*
- "Big Data"
  - *More than 300 million objects, 300+ parameters each*
  - *100 TB raw data, 10 TB catalogues, 2.5 terapixels*
  - *PUBLIC (SQL) DATABASE ("Virtual Observatory")*



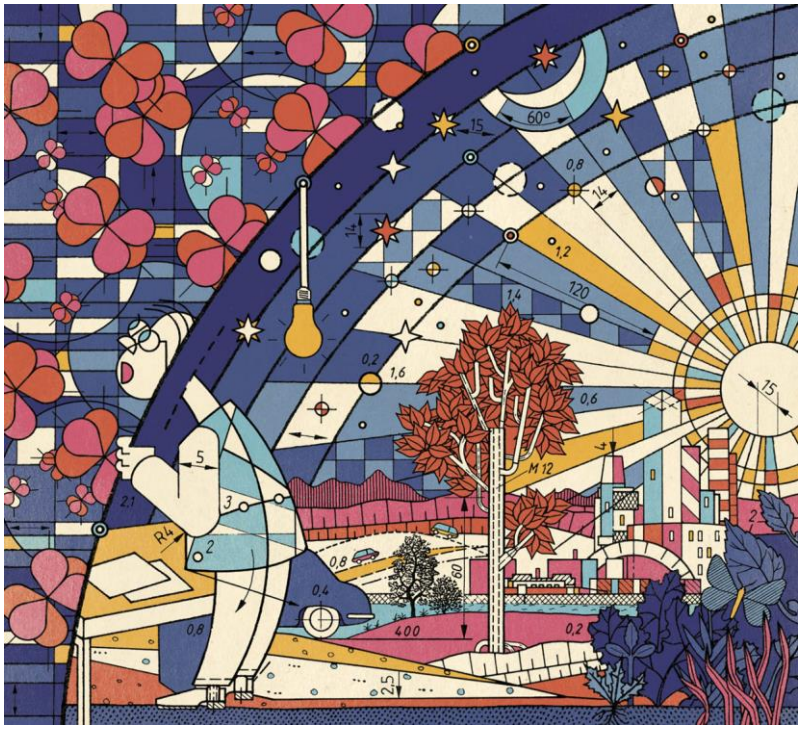
Huge data tables

ra	dec	u	g	r	i	z	deVRad_r	deVPhi_r	redshift	class
348.90253	1.2718862	19.38905	18.24496	17.58728	17.20807	16.90905	3.295783	28.87819	0.03212454	GALAXY
51.443695	1.2700727	19.52808	17.96541	17.03493	16.53754	16.14154	7.599091	63.68505	0.1213151	GALAXY
51.483584	1.2720127	18.72268	17.3852	16.81134	16.51803	16.29502	1.676276	132.2497	0.04876465	GALAXY
49.627485	-1.0417691	17.65612	16.17133	15.5894	15.3785	15.26744	0.0636351	163.8111	-9.77E-05	STAR
40.28569	-0.7149566	17.54884	15.75164	15.031	14.66728	14.36099	9.327478	71.73198	0.04028672	GALAXY
40.272105	-0.6425103	19.23401	17.5333	16.8743	16.63157	16.49762	0.0034072	67.50085	-5.22E-05	STAR
40.582032	0.1347701	18.64558	16.44336	15.52452	15.18185	14.98858	0.0129546	106.2289	0.00017717	STAR
57.025337	0.208845	17.61444	16.17125	15.52131	15.15564	14.86996	10.81576	149.0323	0.0254747	GALAXY
57.047052	0.0843043	19.46874	18.18264	17.59063	17.26436	16.95295	18.96355	31.14236	0.03616738	GALAXY
57.281615	0.0187679	16.4848	14.92993	14.56054	14.53054	14.19394	0.4085672	77.8435	-0.00014215	STAR
57.512104	0.0848866	18.83897	17.63091	17.09078	16.84627	16.71464	0.0103326	106.4699	8.89E-05	STAR
57.605375	0.0272751	18.21801	15.95427	14.95673	14.59481	14.36269	0.000253	73.22543	-2.62E-05	STAR
57.824999	0.215609	17.68076	17.32501	17.1707	17.08611	17.03252	0.0162654	72.24319	0.6822563	QSO
57.943458	0.0596778	16.93403	15.38486	14.69913	14.44319	14.33092	0.0153492	73.84164	0.00011661	STAR
58.175459	0.2186933	19.33956	19.10073	18.66402	18.58816	18.6467	0.0417285	75.5094	1.161747	QSO
58.304024	0.0138137	18.53223	17.24661	16.77493	16.59758	16.50323	0.0204817	106.2418	4.66E-05	STAR
58.395736	0.2097659	17.0049	15.36086	14.49837	14.39811	13.7894	0.021017	105.7351	0.00061353	STAR
36.653674	0.6311025	19.4573	18.126	17.62662	17.45301	17.32834	0.0311647	48.93041	3.63E-06	STAR
37.690126	0.6303724	19.25001	18.32965	17.98234	17.86072	17.78243	0.0071562	73.79427	0.00012205	STAR
40.279741	0.5635092	18.41061	17.24516	17.35439	17.45092	17.5481	0.0150468	105.639	0.00043629	STAR
40.35652	0.5867079	19.15436	18.23266	17.97747	17.89799	17.85765	0.0686916	103.8736	0.00078479	STAR
40.365912	0.4821568	18.40755	16.80093	16.25361	16.07363	15.99621	0.0270869	71.27299	-1.19E-07	STAR
44.223179	1.0513825	17.91608	16.9998	16.61383	16.46706	16.39825	0.0096769	72.74297	-0.00043547	STAR

Photometry table: 300+ columns, 1Bn+ rows

100+ other tables

Scientific observations often result data as multidimensional vector space



Scientific goals

and

researcher's perspective

Queries in “phase space”

Star/galaxy separation
Quasar target selection

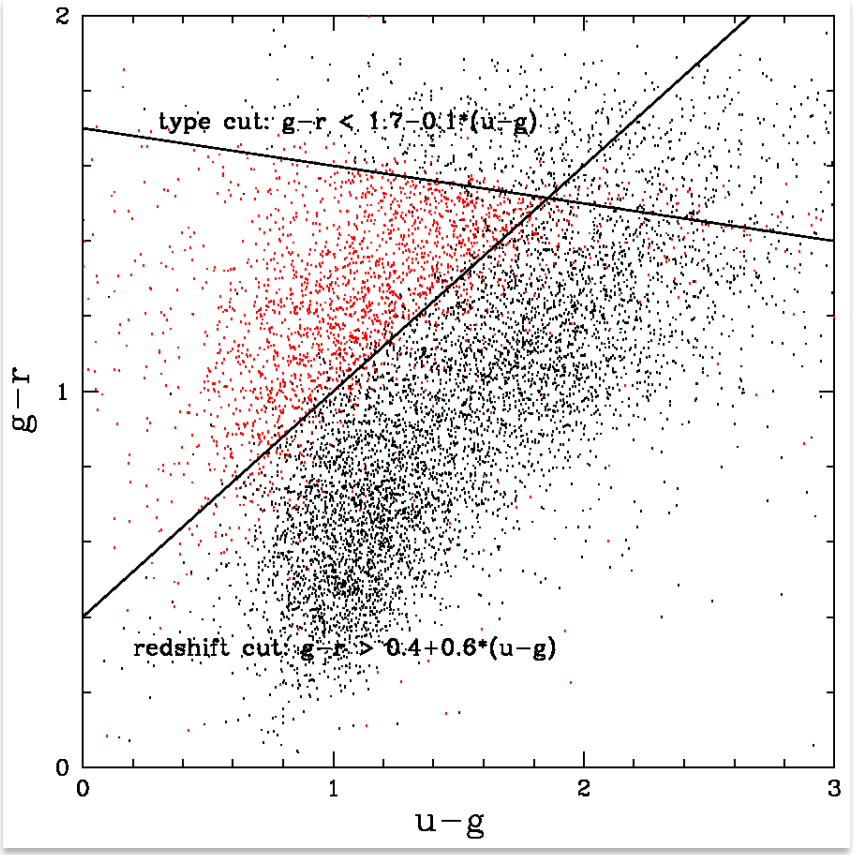
→

“cuts”

→

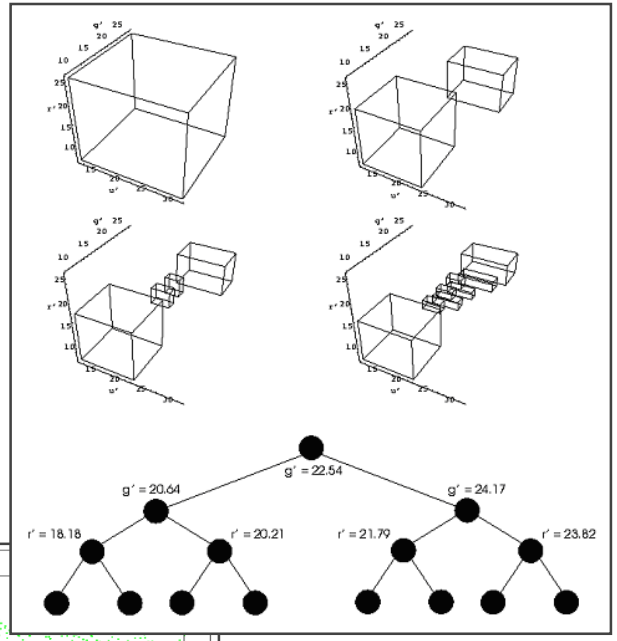
Multidimensional polyhedra

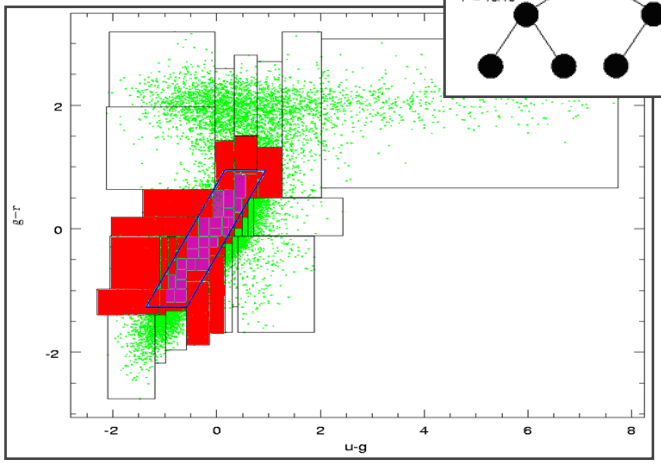
petroMag_i > 17.5 and (petroMag_r > 15.5 or petroR50_r > 2) and (petroMag_r > 0 and g > 0 and r > 0 and i > 0) and ((petroMag_r-extinction_r) < 19.2 and (petroMag_r - extinction_r < (13.1 + (7/3) * (dered_g - dered_r) + 4 * (dered_r - dered_i) - 4 * 0.18)) and ((dered_r - dered_i - (dered_g - dered_r)/4 - 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))

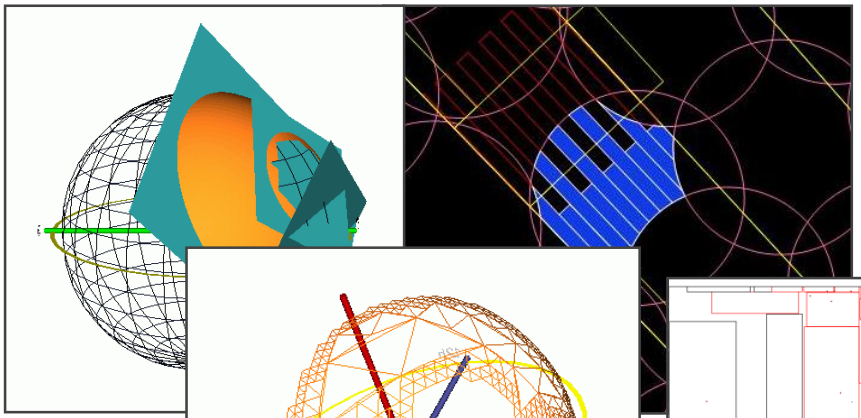


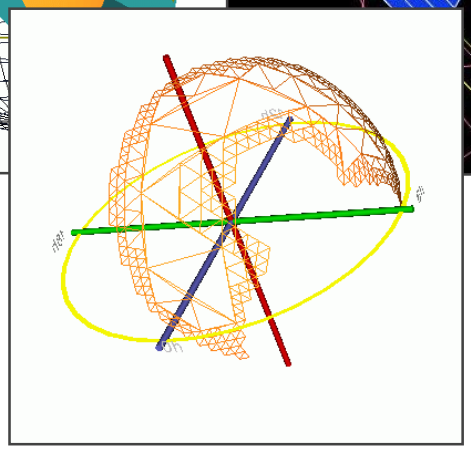
New skills: Indexing, databases

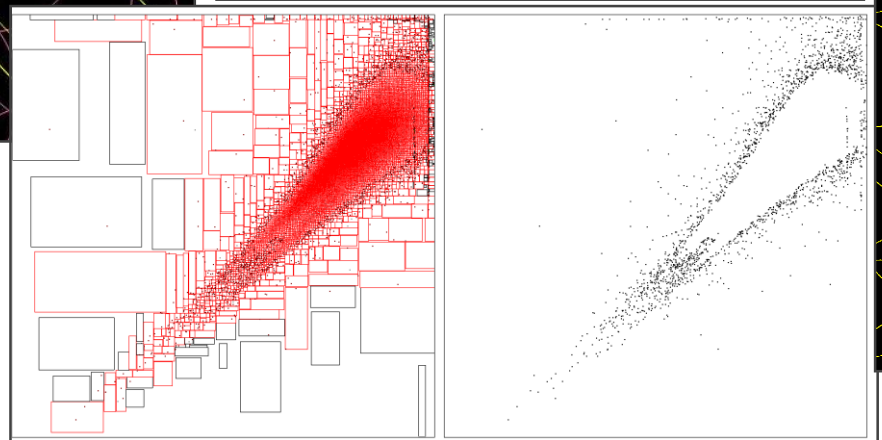
- SDSS data "read through"~1 day
- Astronomers should learn:
  Database programming, computer geometry, search trees, …
- Multidimensional- and spherical indexing

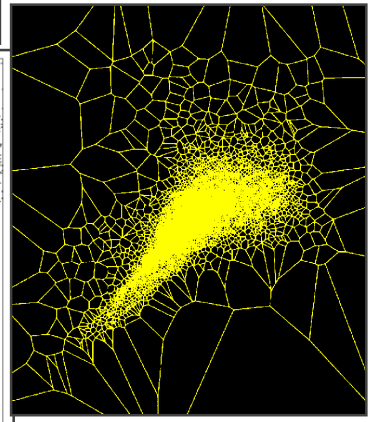




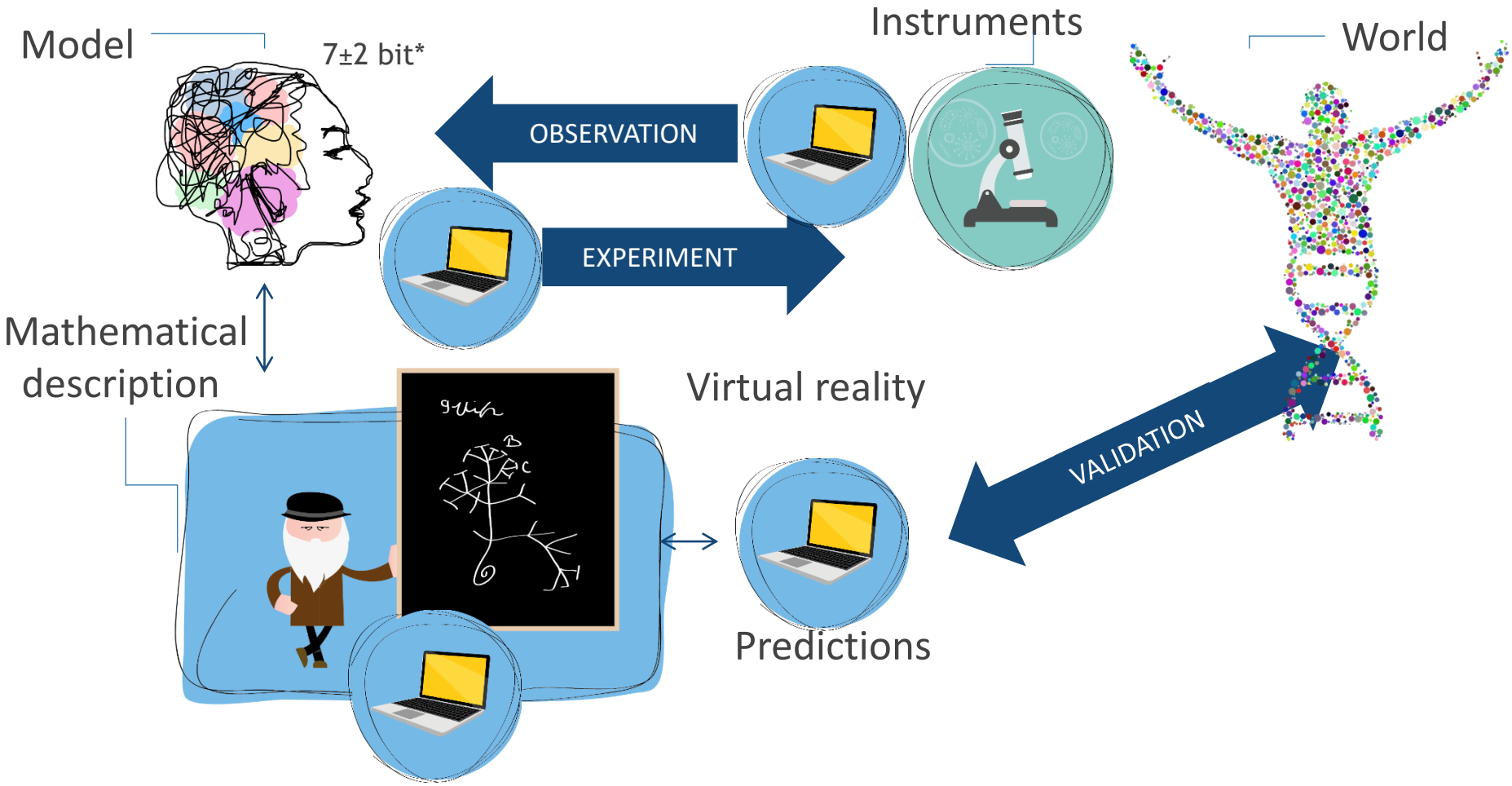




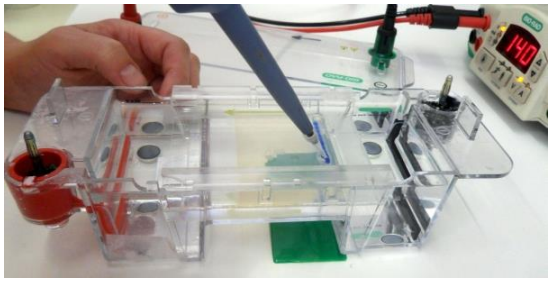


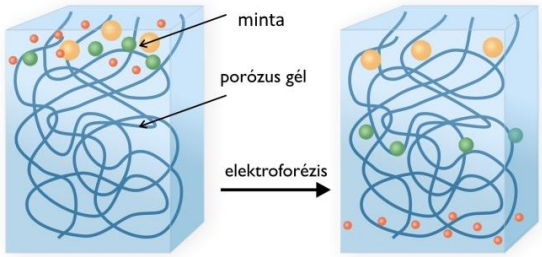


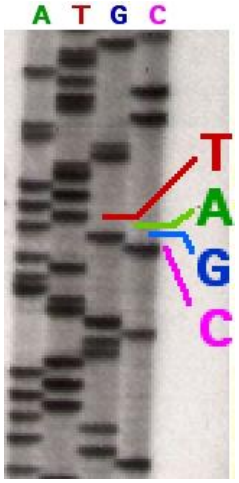
Modern data science: same trends in biology, environmental sciences, social sciences, ….

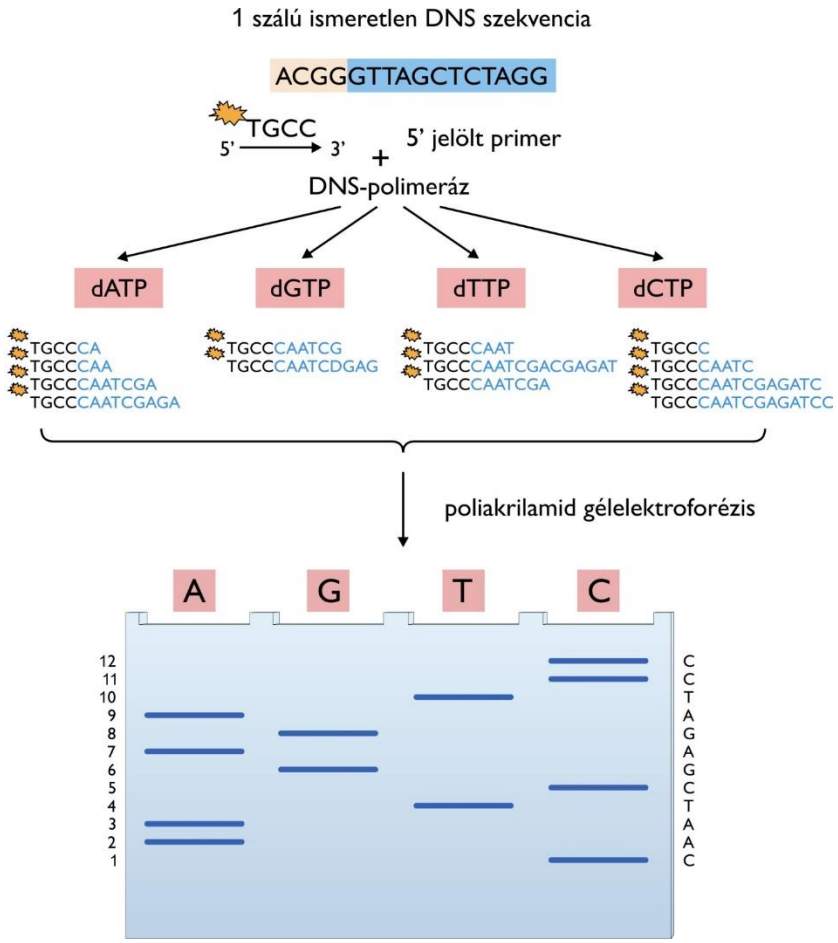


Szekvenálási technikák: Sanger-szekvenálás (1977)



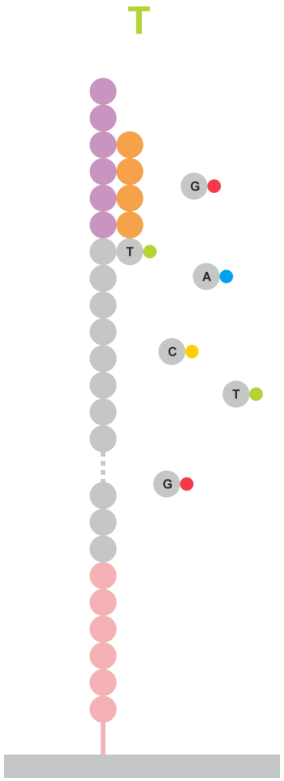


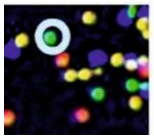




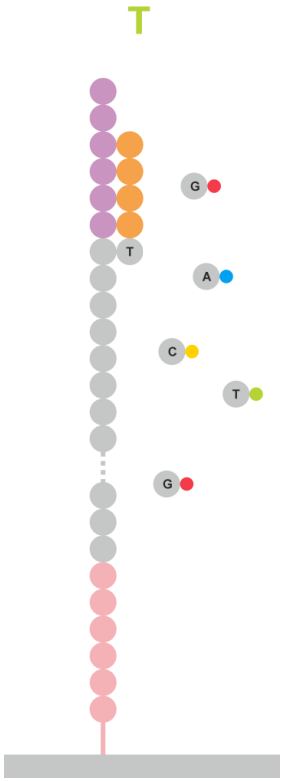
Nyitray László, Pál Gábor: A biokémia és molekuláris biológia alapjai (2013)

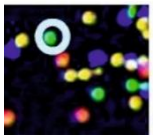
Szekvenálási technikák: NGS



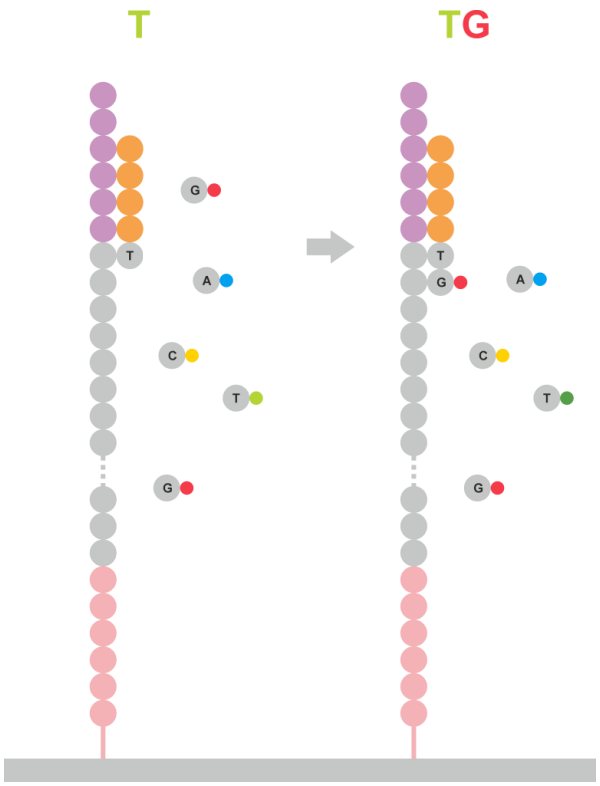


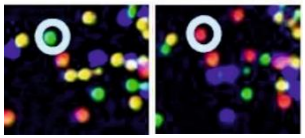
Szekvenálási technikák: NGS



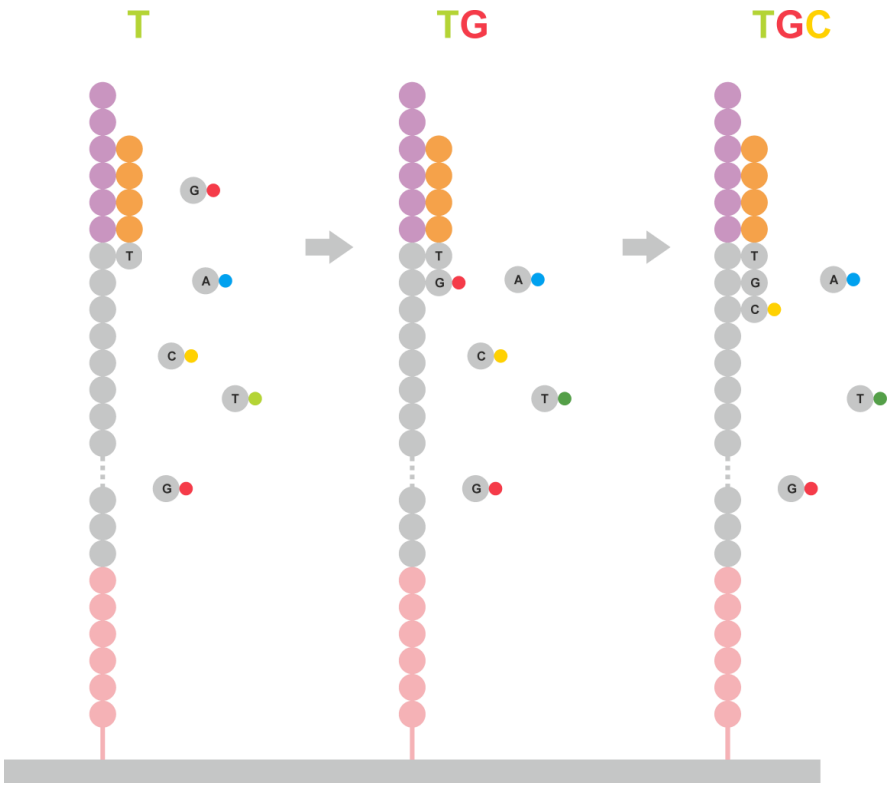


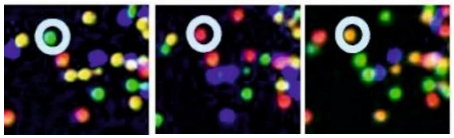
Szekvenálási technikák: NGS



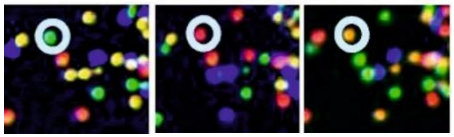


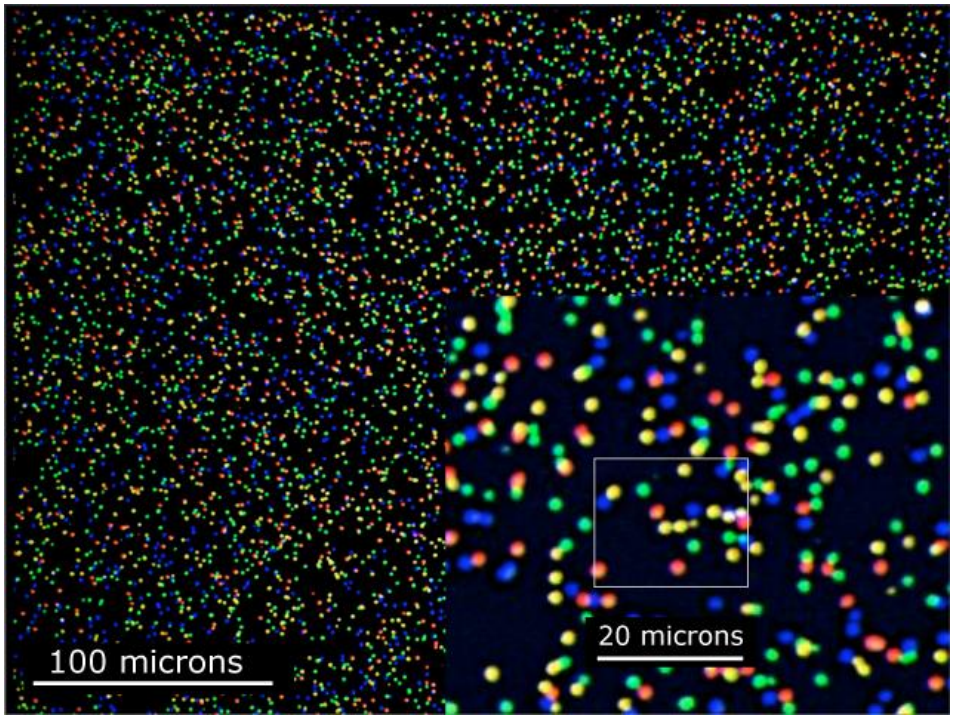
Szekvenálási technikák: NGS





Szekvenálási technikák: NGS

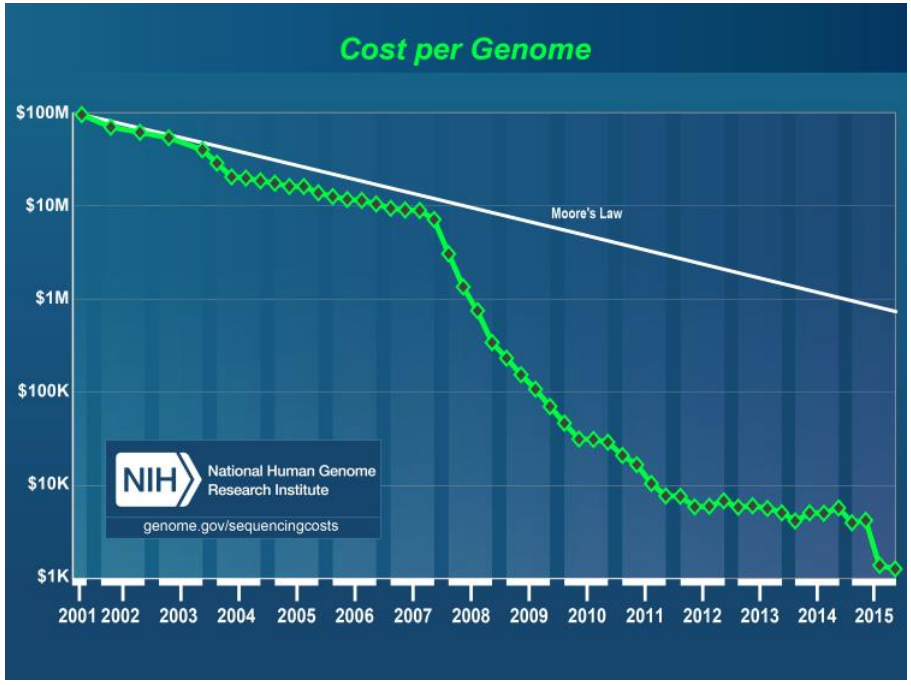




D. Mertens, K. Rippe, German Cancer research Center

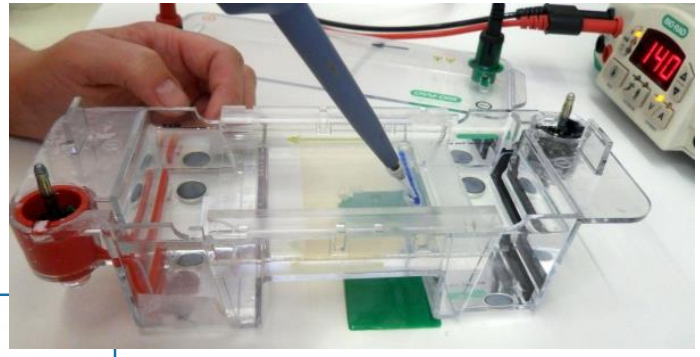
BGI Hong Kong, Scotted400, CC-BY-3.0

Moore's law in gene sequencing



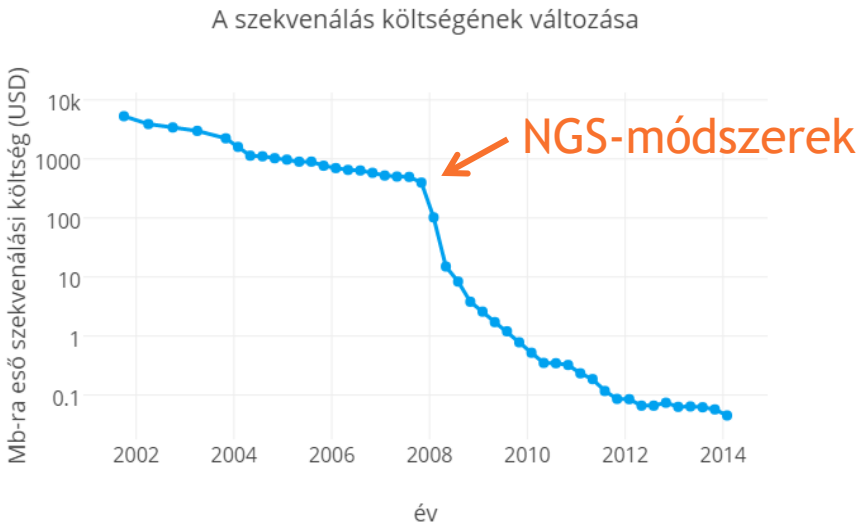
Human genome sequencing
1990-2003: 13yrs /2.7 Bn USD
2016: ~days/1000 USD
2020: ?????

- X Prize $10M, 2006, 100 genom, 30 days, $10k – cancelled (2006)
- Microarray, CCD!
- Mass spectroscopy
- Digital microscopy
- …



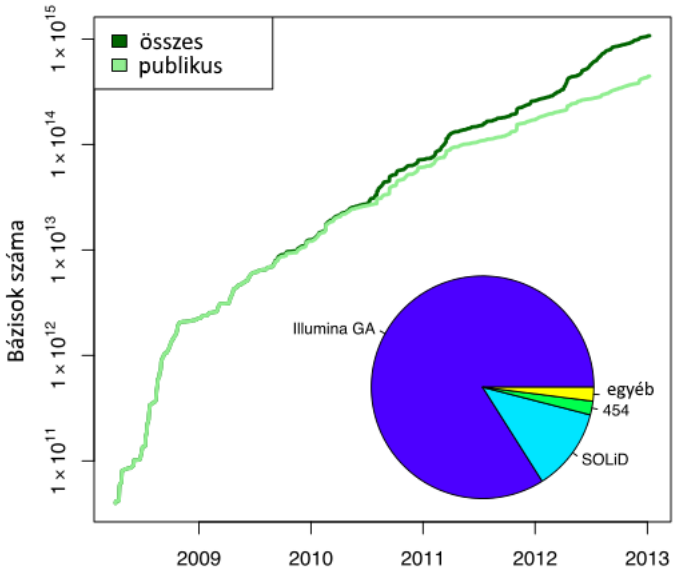
Oxford Nanopore 100Mb,$900

Szekvenálási technikák: NGS



Egyre olcsóbb szekvenálni.

ENA / SRA



Egyre több adat publikus online.

1990–2003
13 év / 2,7 milliárd USD

2016
néhány nap / 1000 USD

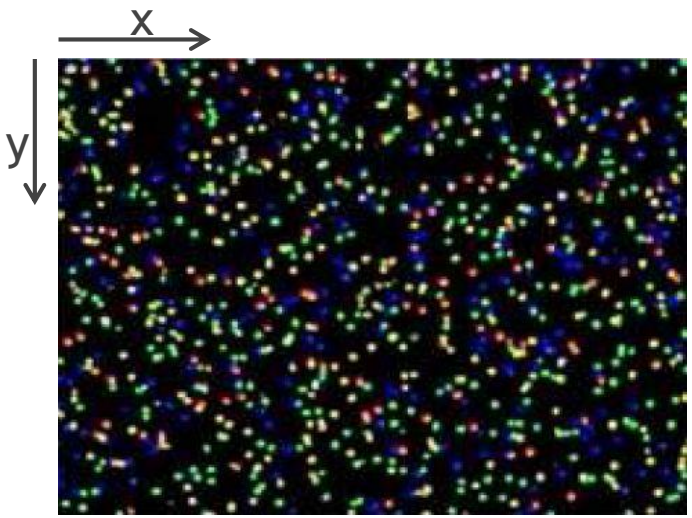
2020?

Biology in the 20th century

21st

NGS – adatfeldolgozás

1. Képekből szöveges *short read*



koordináta		színintenzitás			
x	y	A	C	G	T
…	…	…	…	…	…
17	20	4	13	76	3
17	25	2	45	41	10
…	…	…	…	…	…
1001	1253	8	1	2	97
…	…	…	…	…	…

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

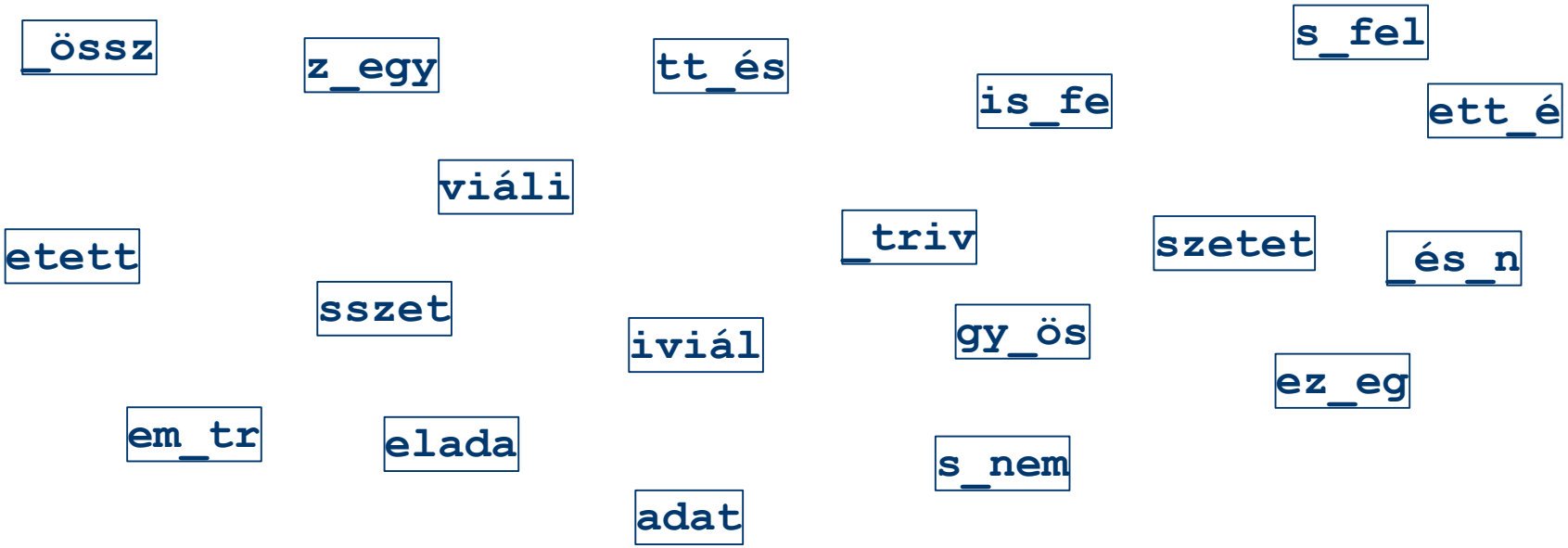
FASTQ formátum

NGS – adatfeldolgozás

„Sok széttépett szakácskönyv”

2. Short readből teljes genom: összeillesztés

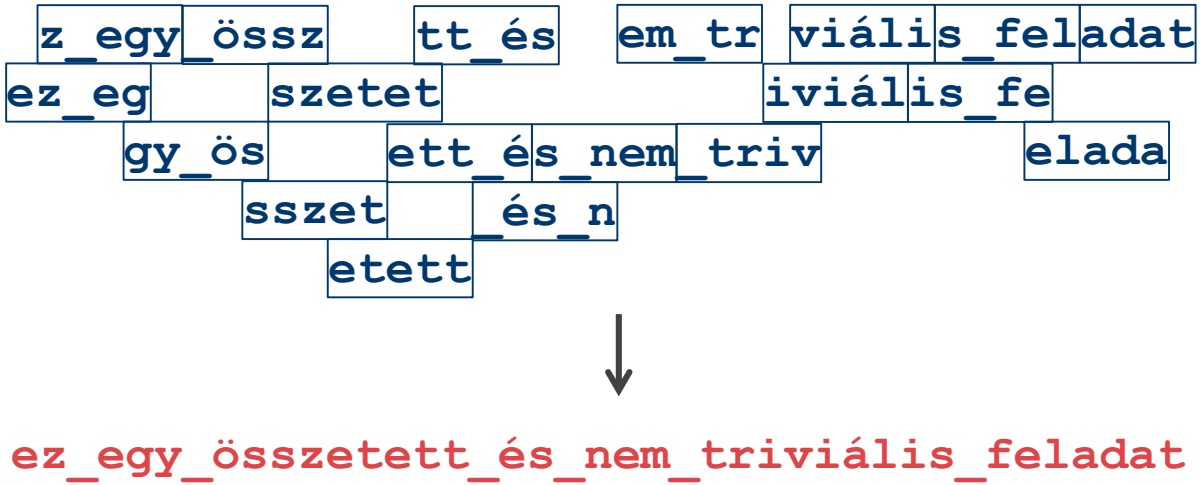
A) referenciagenom (minta) nélkül: *de novo* összerakás



NGS – adatfeldolgozás

2. Short readből teljes genom: összeillesztés

A) referenciagenom (minta) nélkül: *de novo* összerakás

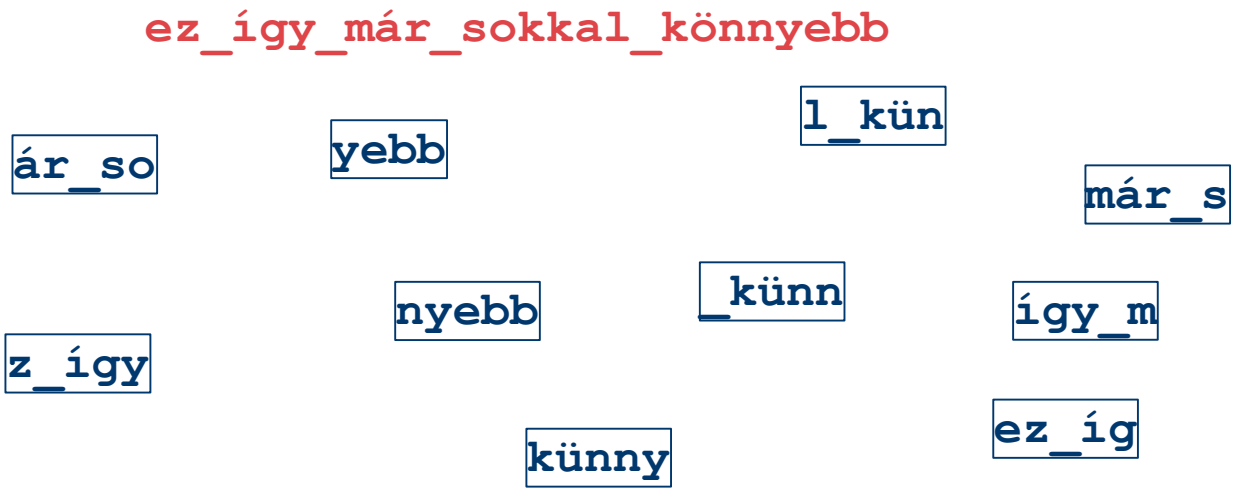


NGS – adatfeldolgozás

2. Short readből teljes genom: összeillesztés

A) referenciagenom (minta) nélkül: *de novo* összerakás

B) referenciagenom segítségével



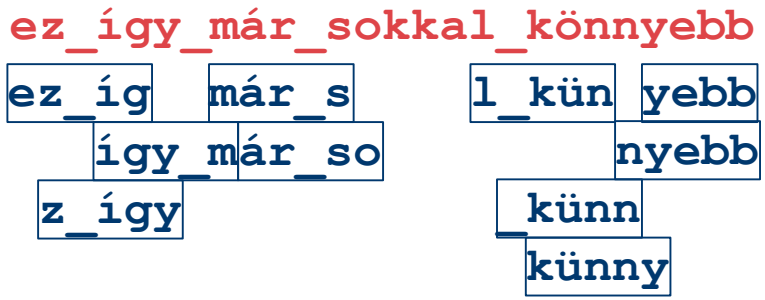
NGS – adatfeldolgozás

„Sok széttépett **hibás** szakácskönyv”

2. Short readből teljes genom: összeillesztés

A) referenciagenom (minta) nélkül: *de novo* összerakás

B) referenciagenom segítségével



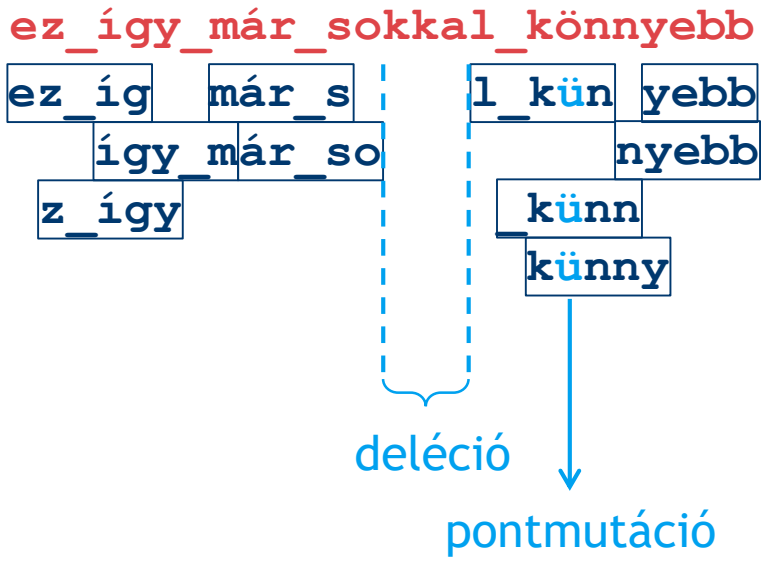
NGS – adatfeldolgozás

„Sok széttépett **hibás** szakácskönyv”

2. Short readből teljes genom: összeillesztés

A) referenciagenom (minta) nélkül: *de novo* összerakás

B) referenciagenom segítségével



NGS – adatfeldolgozás

2. Short readből teljes genom: összeillesztés

A) referenciagenom (minta) nélkül: *de novo* összerakás

B) referenciagenom segítségével

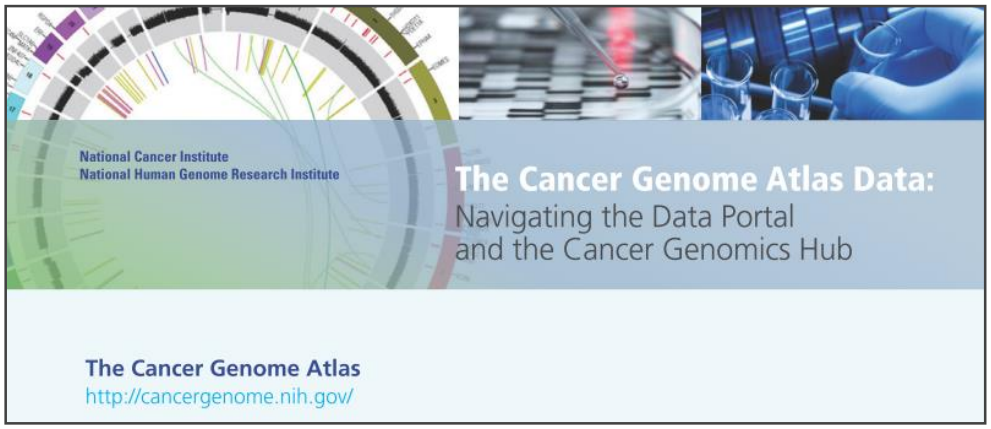
Processzorsebesség:	$\sim 10^9$ utasítás/sec
Humán genom:	$\sim 10^9$ nukleotid
NGS:	$\sim 10^9$ short read
"nyers erővel":	$\sim 10^{18}$ összehasonlítás,
vagyis	$\sim 10^9$ sec ≈ 32 *év*

Kreatív indexelő és kereső algoritmusok kellenek!

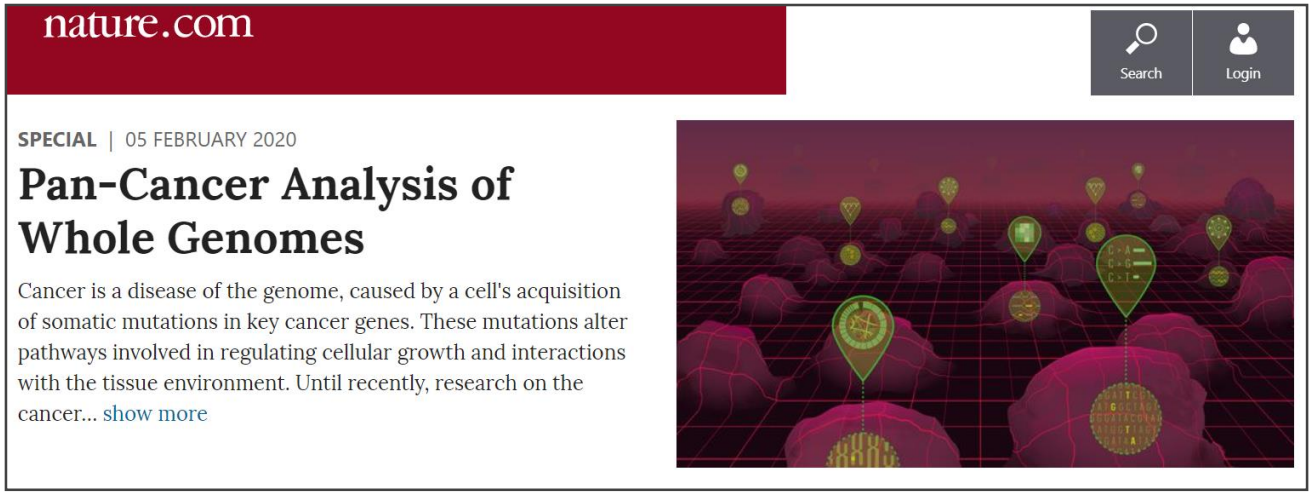
NGS – adatfeldolgozás

3. Illesztett adatok + metaadatok

- 38 ráktípus
- 2600 eset
- 3,2 milliárd nukleotid/genom

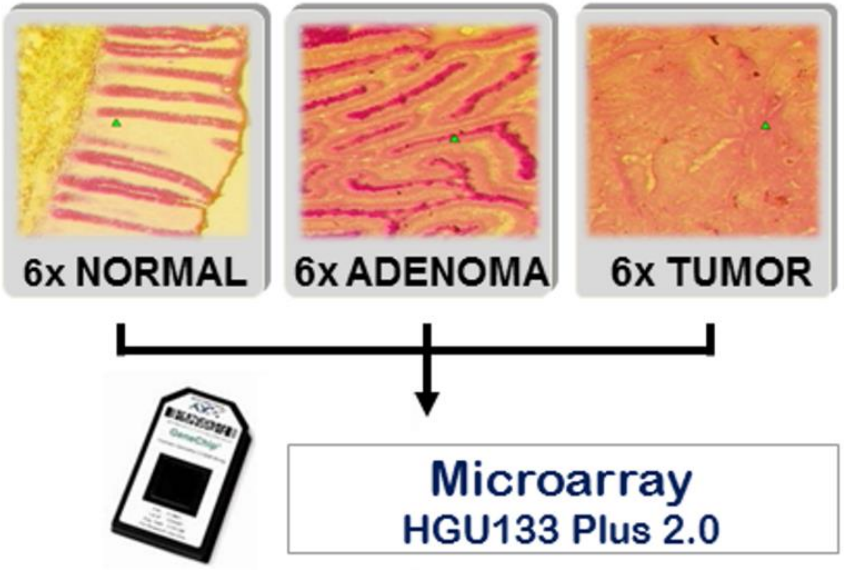


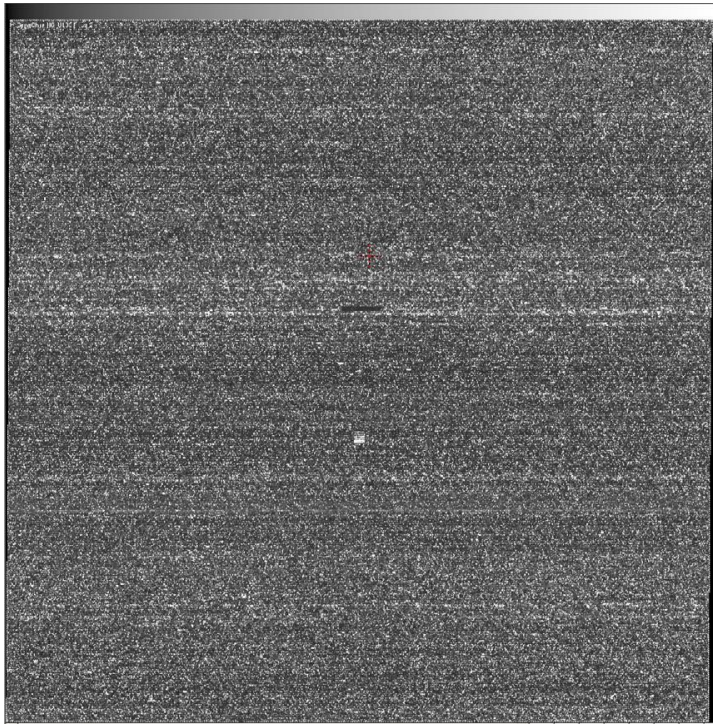
Különböző formátumú nyers adatfájlok és komplex metainformációk egyvelege



The Pan-Cancer Analysis of Whole Genomes Consortium brought together researchers with nearly 750 affiliations across 4 continents. Between them, they sequenced full genomes from more than 2,600 samples representing 38 different types of cancer.

Gene expression “Big Data” (2009)





Gene expression values extracted from images:
54675D vectors + metadata

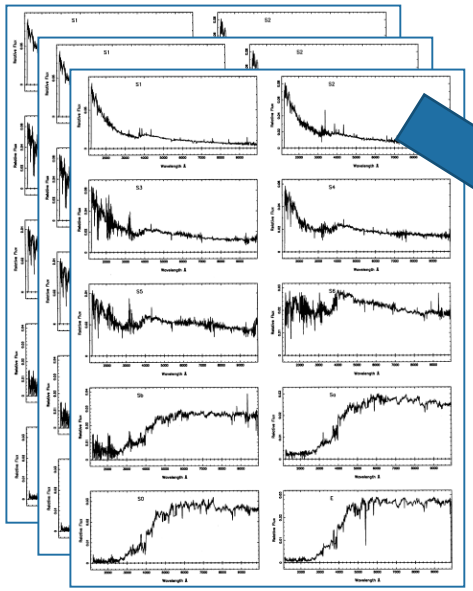
ELTE-SOTE-3DHISTECH

- Affymetrix HG U133 Plus2
- Raw datat: 67 Mpix image (photmetry!)
- -> 604 258 probe
- -> 54 675 probe set (~gene)
- 207 samples (colorectal cancer)
- Goal: “marker genes” of cancer

S Spisák, A Kalmár, O Galamb … I Csabai, I Kovalszky, S Semsey, Z Tulassay, **B Molnar;** PLoS One dx.doi.org/10.1371/journal.pone.0046215 (2012)

Similar challenges

- Galaxy spectra: 1 million times 3000 dim vectors
- Microarray study: 207 times 54675 dim vectors
- 30 million bitcoin users, 3 billion tweets



7±2 bit

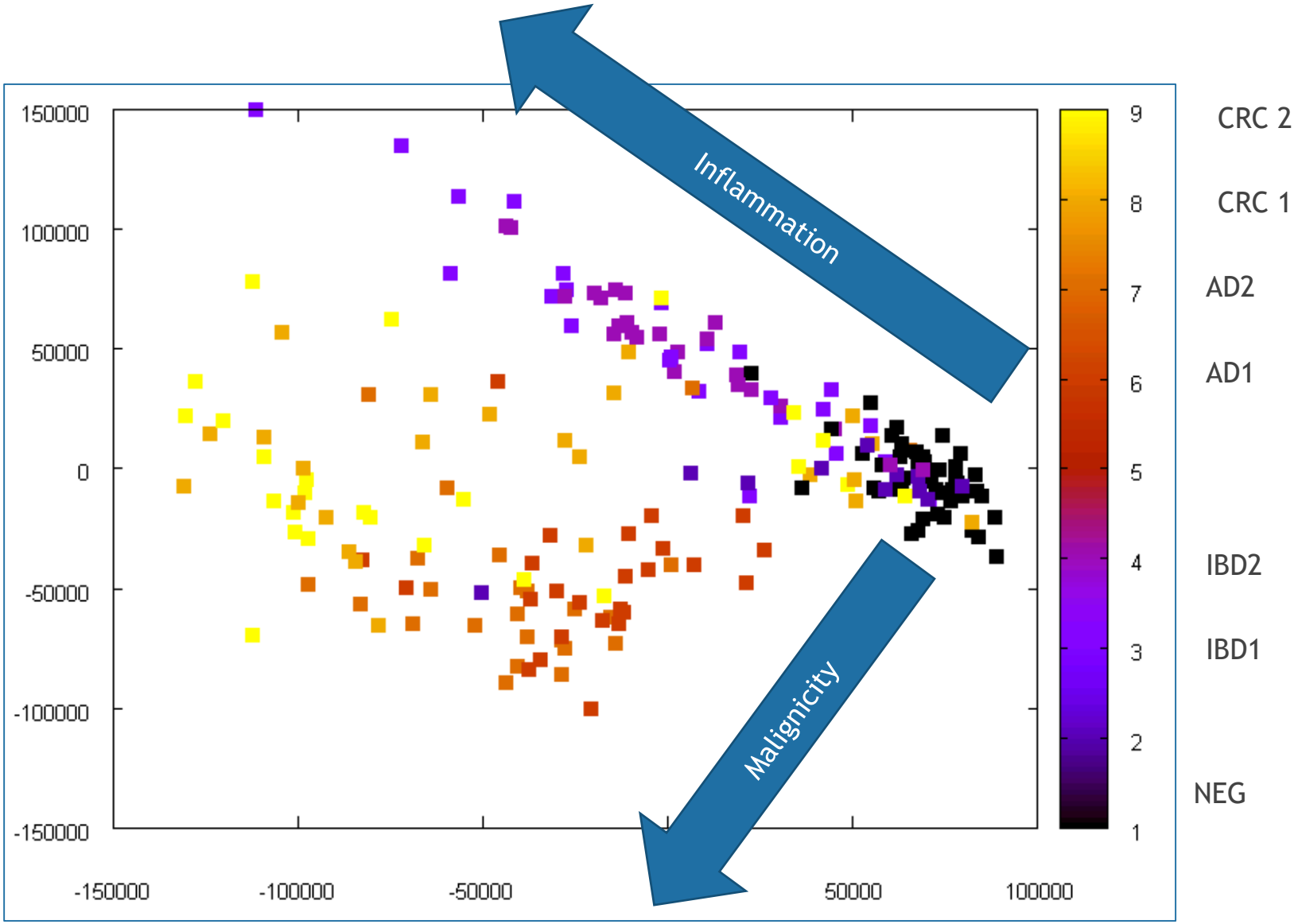
Due to the underlying physical laws, data vectors does not fill the whole space, rather lie on lower dimensional surface/subspace (this is why we can understand the word!)

$pV = NkT$

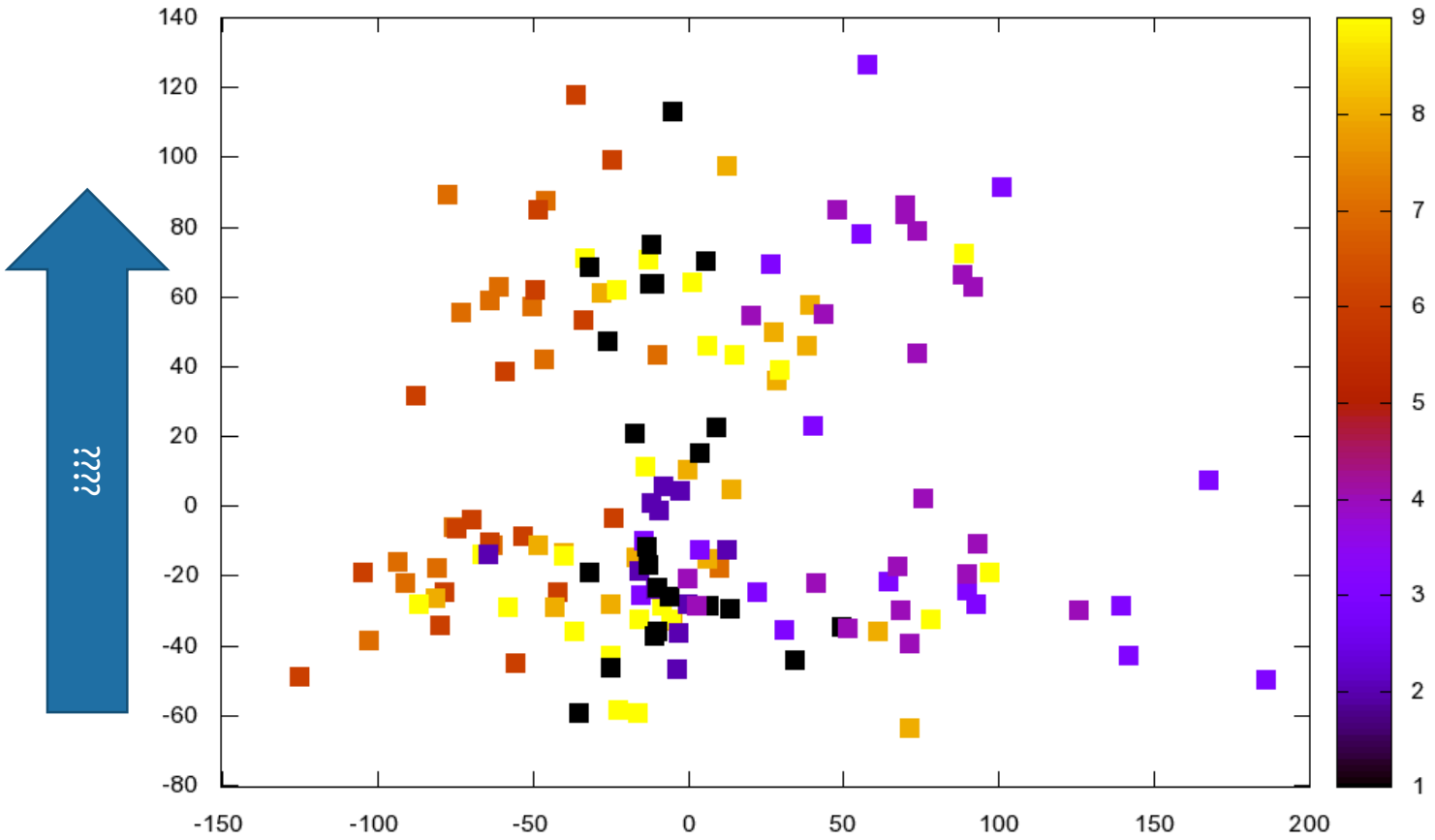
$6 \cdot 10^{23} \rightarrow 5$

Compression : dimension reduction, matrix factorization, machine learning

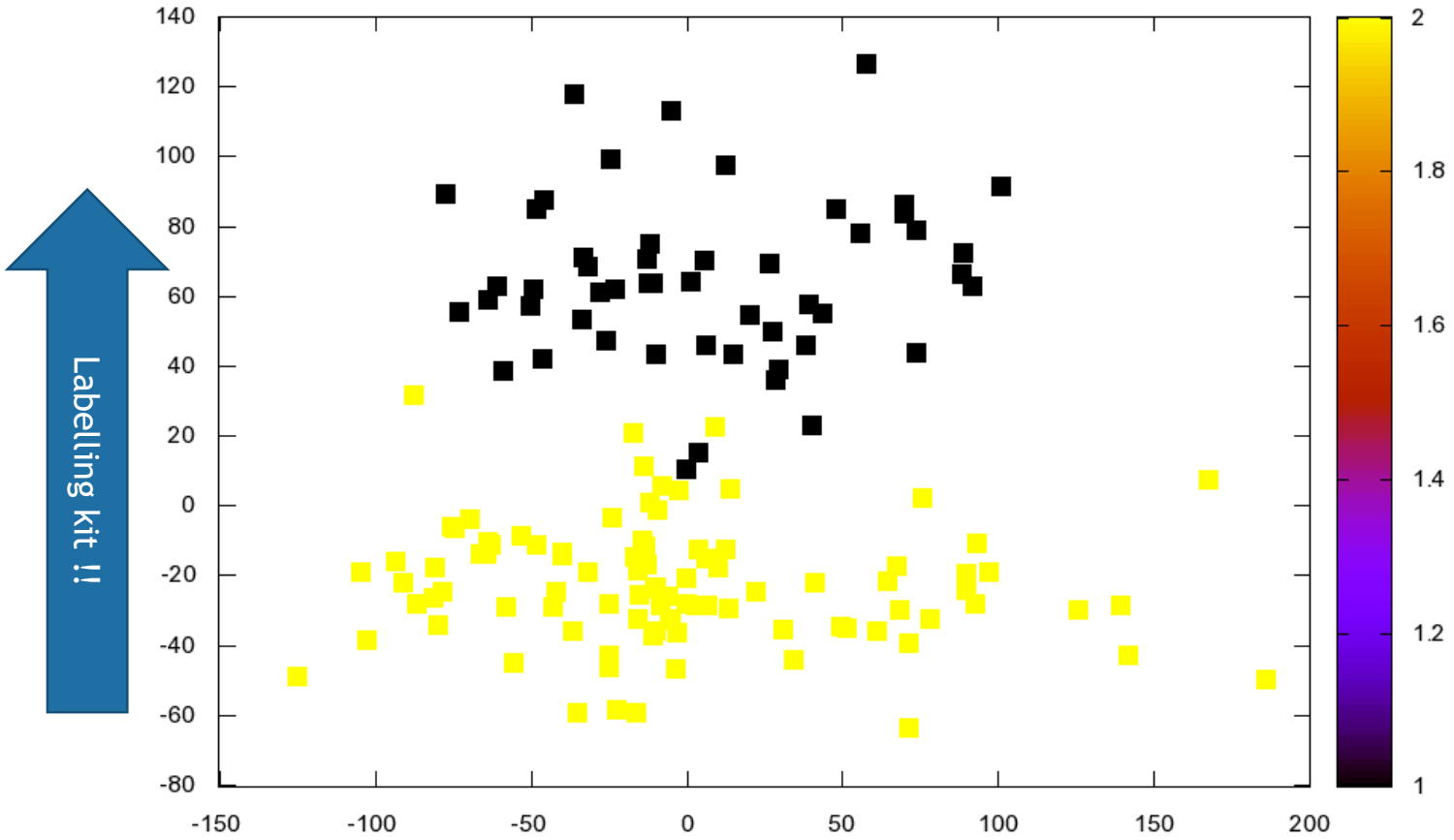
Gene expression microarray: 54675D -> 2D



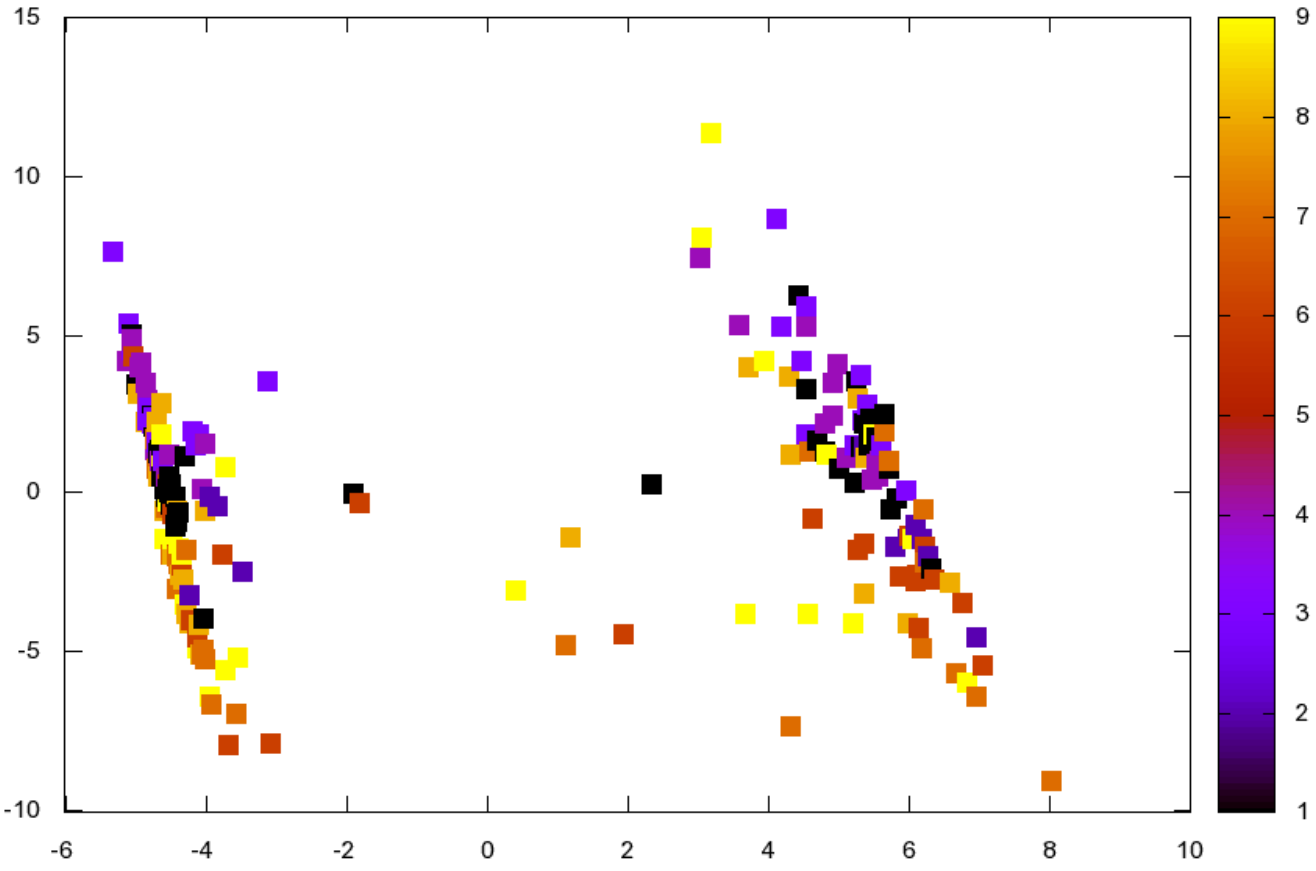
PCA2, PCA3 clusters?



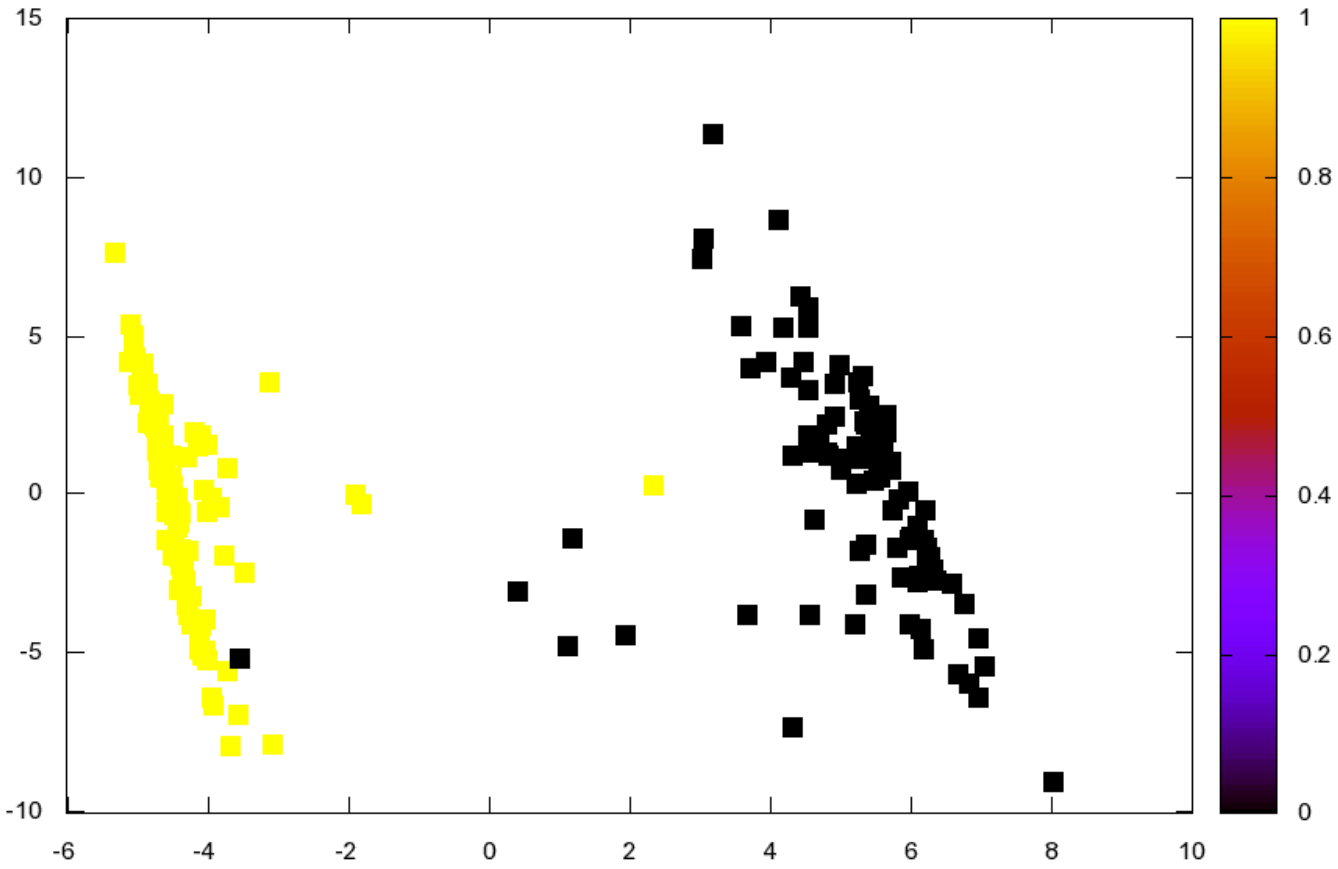
PCA2, PCA3 clusters



PCA – KEGG pathways (ribosome)



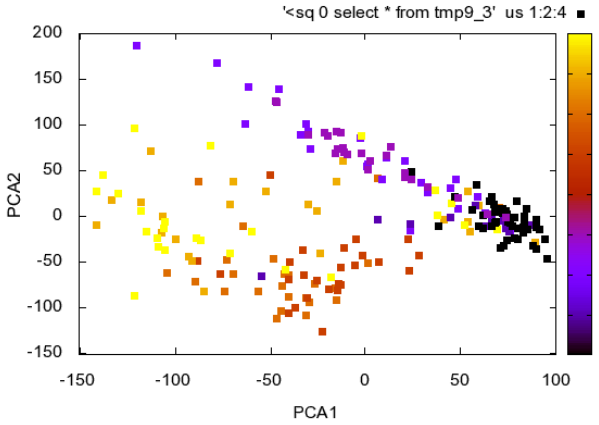
PCA – KEGG pathways (ribosome)



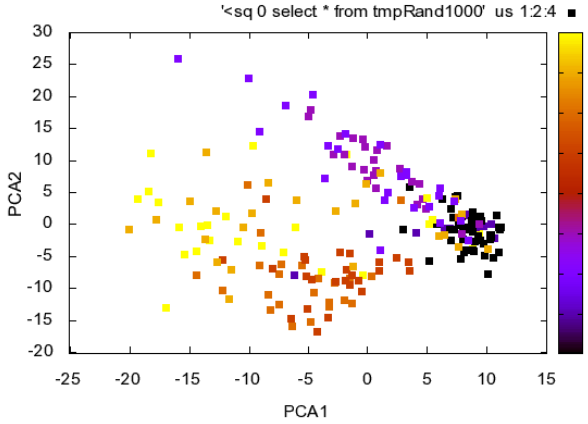
Male - Female

Complex systems

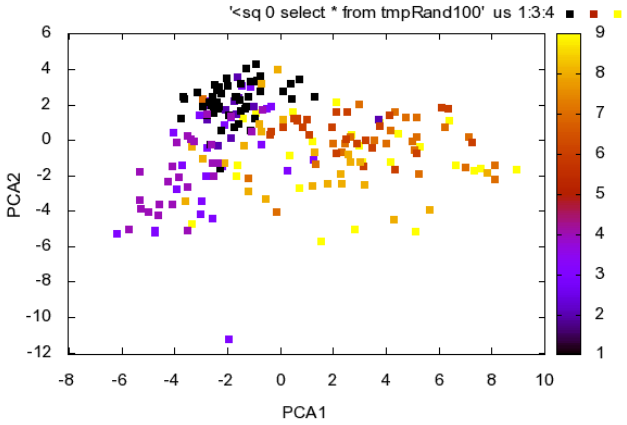
„Realize that everything connects to everything else”
/ Leonardo da Vinci/



all probes (54 675)



random 1000

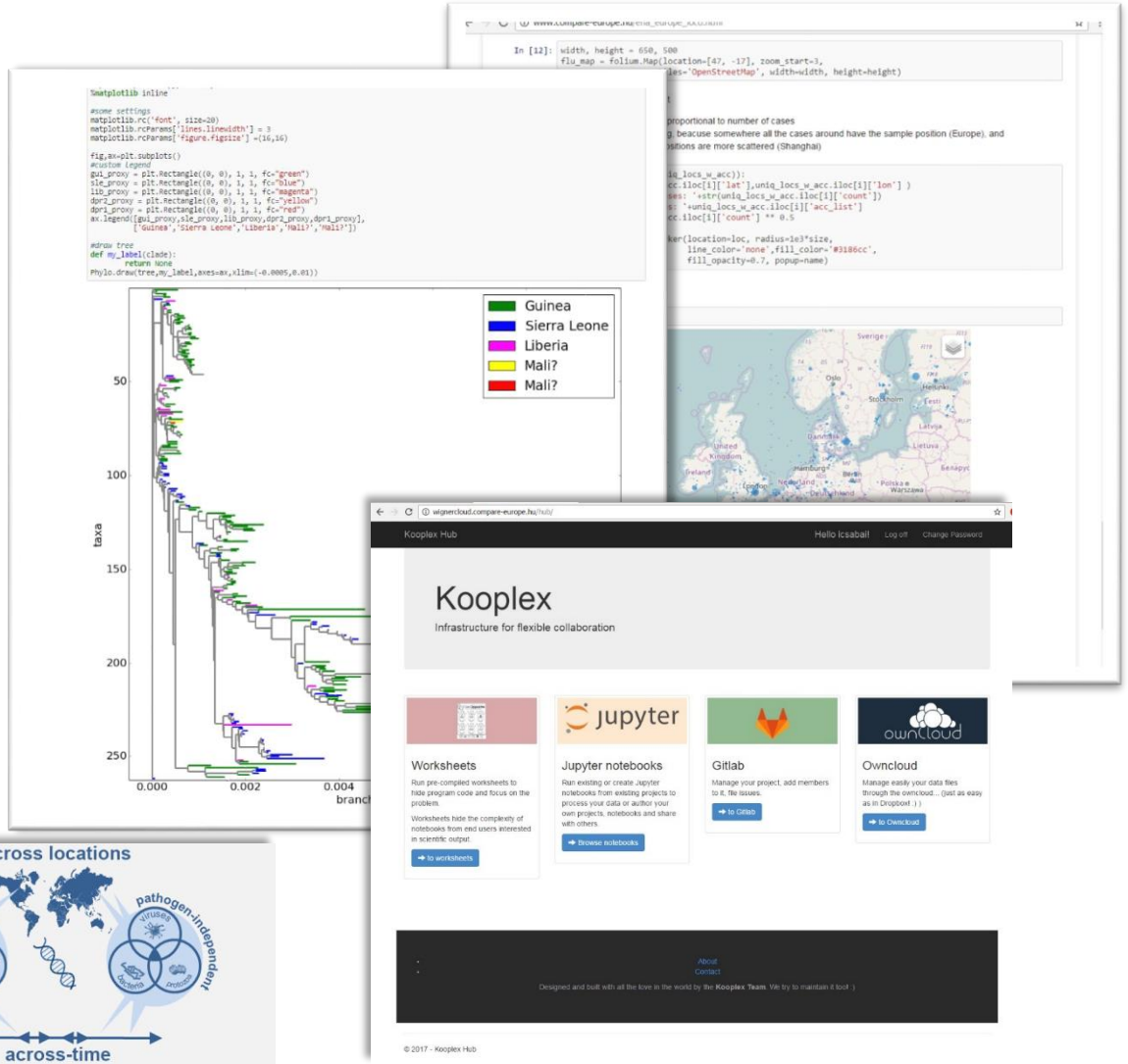


random 100

„Marker genes”?
Complex, nonlinear interacting network!

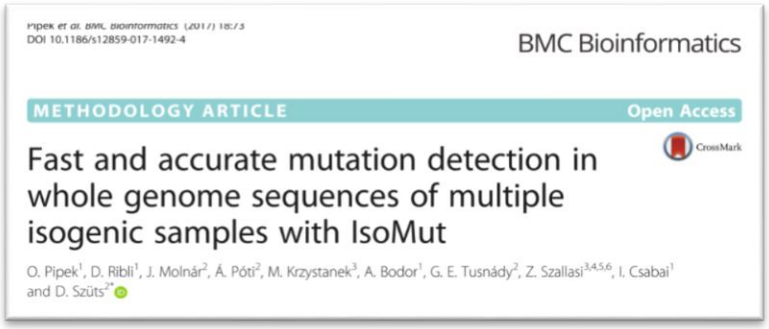
Multidiszciplináris hazai és nemzetközi együttműködések

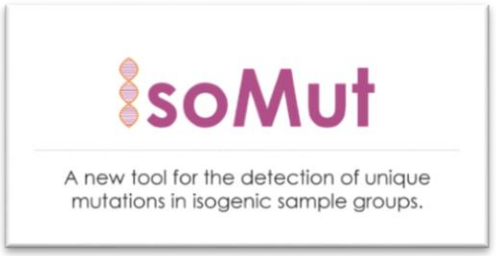
- FIEK_16-1-2016-0005: Biomarkerek (ELTE-MTA TTK-CRU-SERVIER)
- NVKP_16-1-2016-0004: Magyar onkogenom, folyadékbiopszia (SOTE-3DHISTECH-ELTE)
- NKFI OTKA 124881: DNS-javító mechanizmusok (MTA TTK-ELTE)
- Novo Nordisk Multidisciplinary Synergy (Danish Cancer Society Research Center-DTU-Francis Crick Institute-ELTE)
- COMPARE EU H2020: Fertőző betegségek, vírusok, baktériumok, metagenomika (~15 nemzetközi partner, MTA Wigner FK Adatközpont)
- VEO H2020: Fertőző betegségek, vírusok, baktériumok, metagenomika (~15 nemzetközi partner, ELTE)



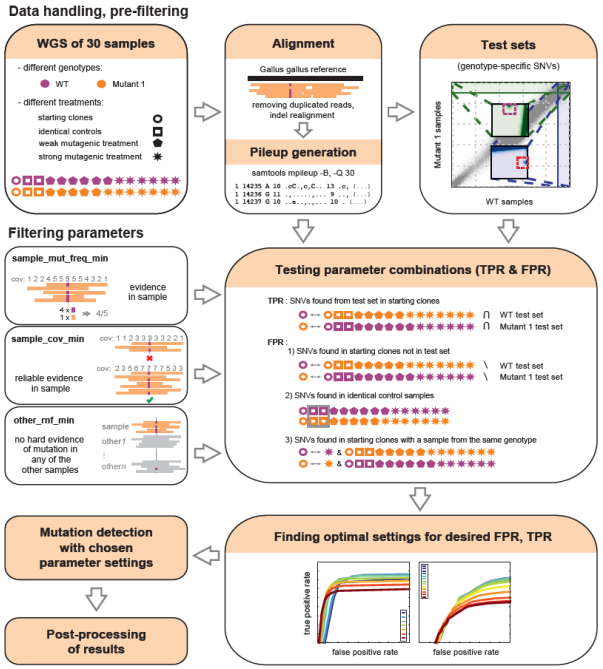
IsoMut

D. Szűts MTA TTK, Z. Szállási Harvard/DTU



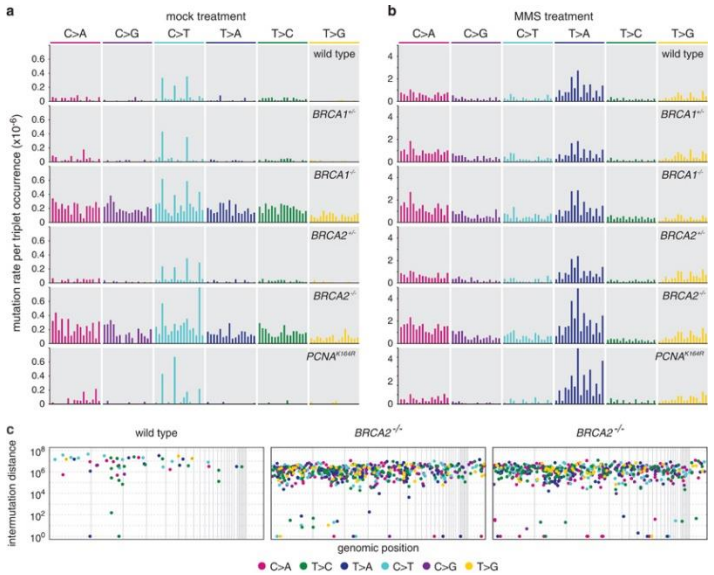


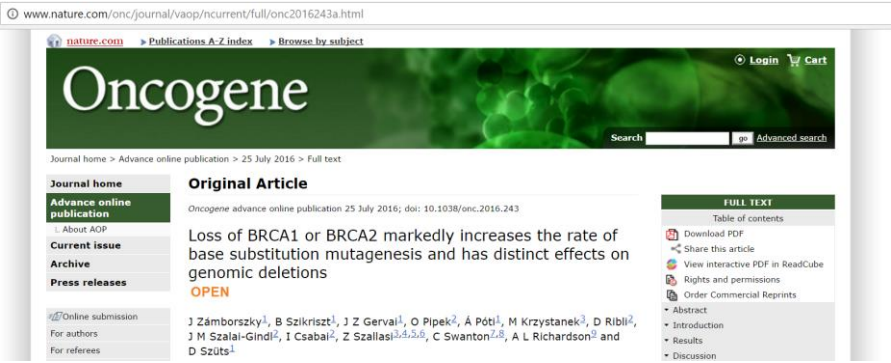
- Sok izogenikus minta esetén
- A referenciagenomtól való eltérések, illesztési hibák korrigálódnak
- Egyedi mutációk detektálása (például kezelés hatása)
- Gyors, pontos (nagyon kevés fals pozitív eredmény)
- Felhasználási példa: kemoterápiás szerek mutagén hatásának vizsgálata (B Szikriszt et al., Genome biology 17 (1), 99 (2016))



D. Szűts MTA TTK, Z. Szállási Harvard/DTU, C. Swanton Francis Crick Institute

DNS-javító mechanizmusok



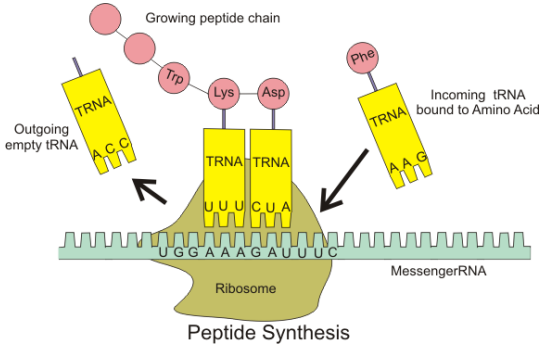


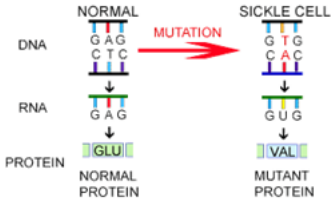
- Génkiütéses sejtvonalak, mutagén kezelések
- Mutációs szignatúrák (NNMF)
- Mutációs spektrumok összevetése a TCGA-eredményekkel

Nem kódoló mutációk szerepe

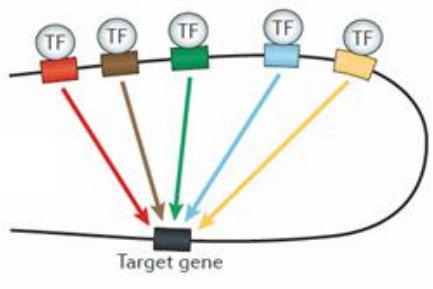
M. Freedman, S. Spisák: Harvard

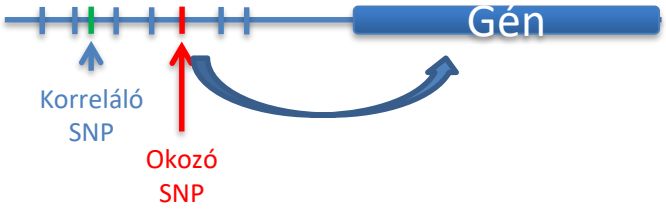
Kódoló régiók (~2%):





„Nem kódoló" régiók (~98%), „GWAS" statisztikai asszociációk:





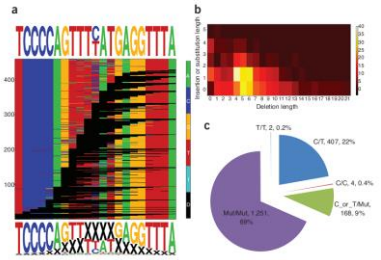
M. Freedman, S. Spisák: Harvard

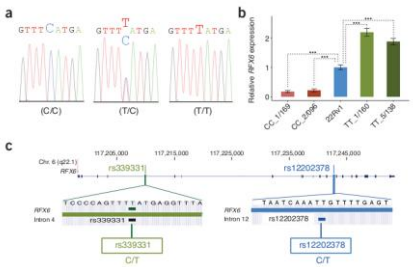
Nem kódoló mutációk : TALEN genomszerkesztés

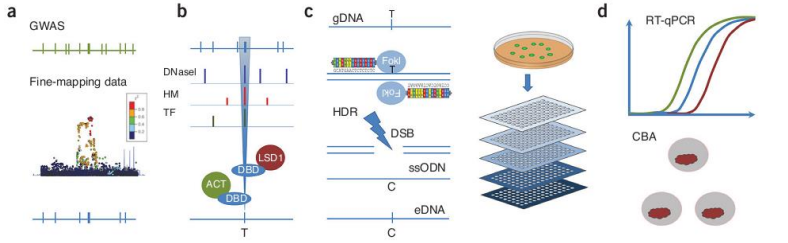
nature medicine

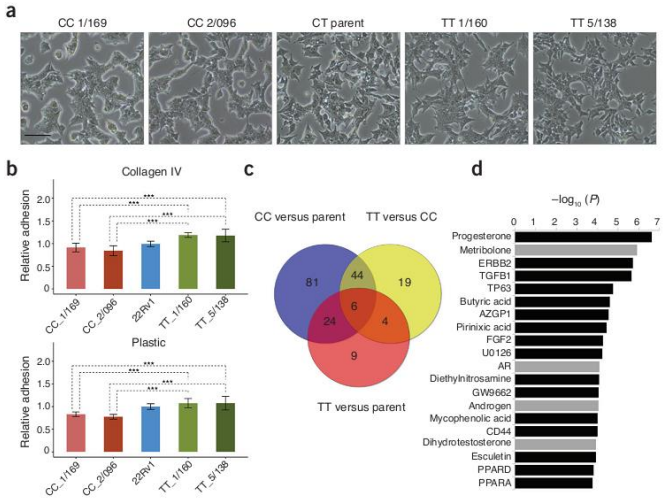
CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants

Sándor Spisák[1,2,20], Kate Lawrenson[3,20,21], Yanfang Fu[4–7,20,21], István Csabai[8], Rebecca T Cottman[4–6,9], Ji-Heui Seo[1,2], Christopher Haiman[3,10], Ying Han[3], Romina Lenci[1,2], Qiyuan Li[1,2,11], Viktória Tisza[1,12], Zoltán Szállási[12–14], Zachery T Herbert[15], Matthew Chabot[1], Mark Pomerantz[1], Norbert Solymosi[16], The GAME-ON/ELLIPSE Consortium[17], Simon A Gayther[3,18], J Keith Joung[4–7,9] & Matthew L Freedman[1,2,19]







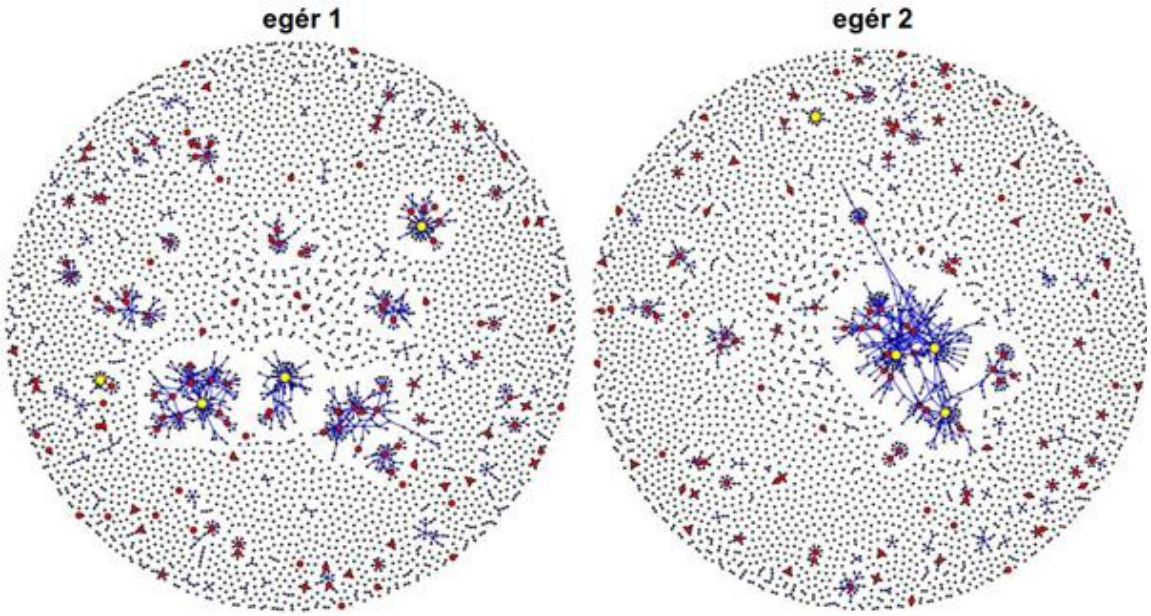


„Egy szög miatt a patkó elveszett.
A patkó miatt a ló elveszett.
A ló miatt a lovas elveszett.
A lovas miatt a csata elveszett.
A csata miatt az ország elveszett. ”
/Lúdanyó meséi/

„A kiátlagolódás hiánya”:
erős csatolás a mikro- és makroskálák között.
Egyetlen, nem kódoló nukleotid megváltoztatása megváltoztathatja a fenotípust.

Immunrendszer-szekvenálás

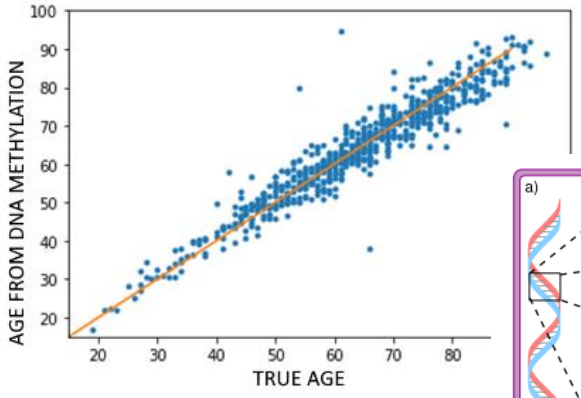
I. Kacskovics, B. Szikora, O. Pipek, ELTE

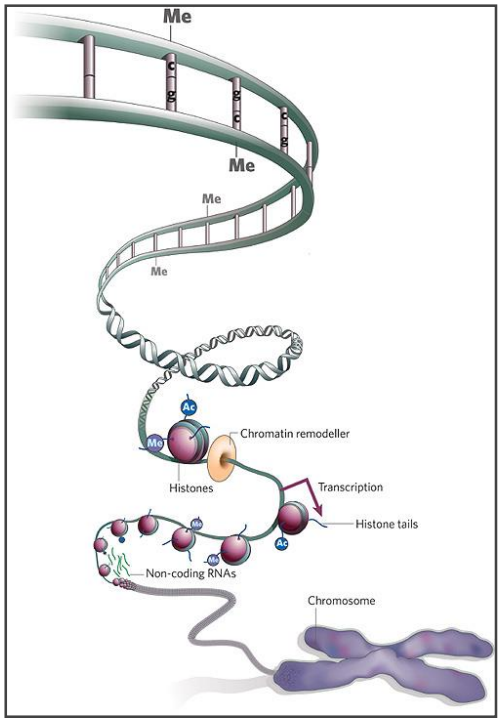


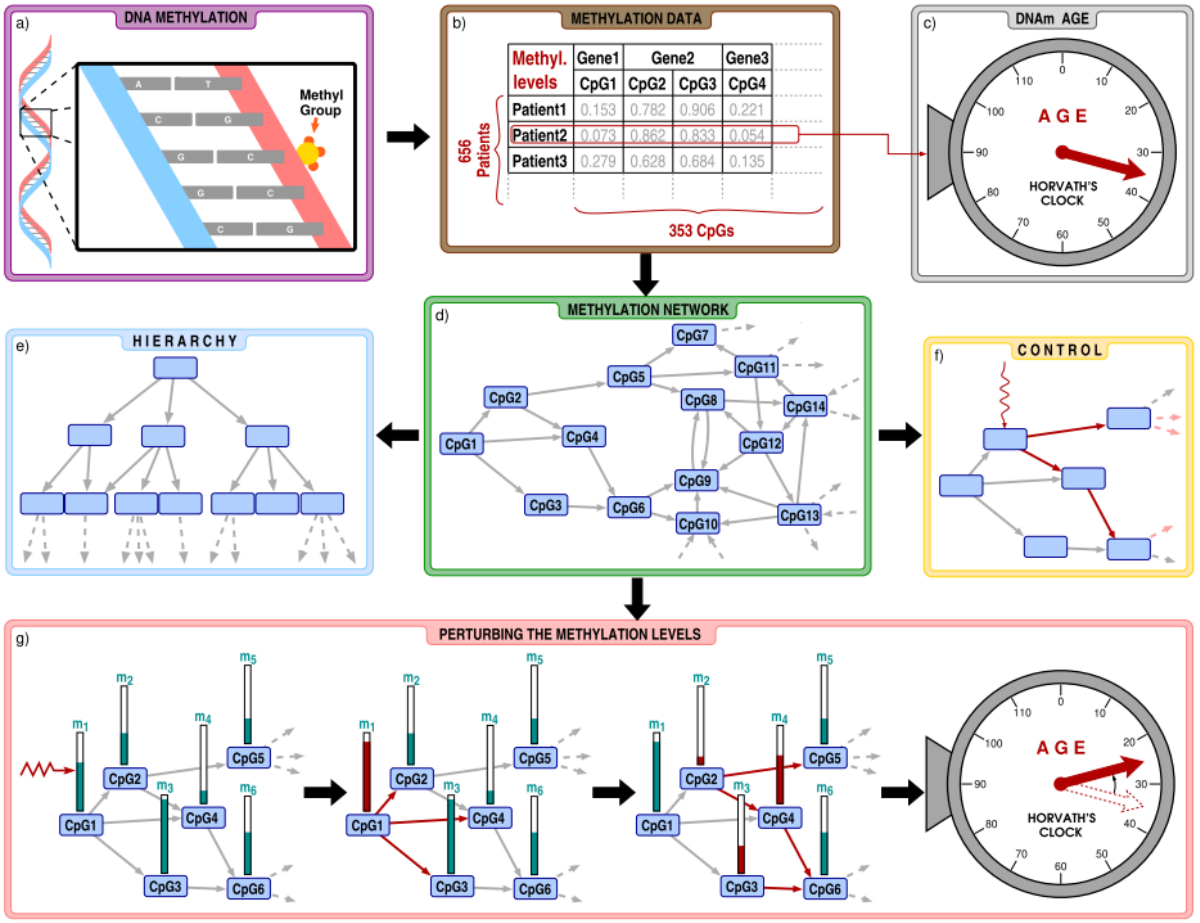
Immunizált egerek lépéből izolált (FACS: CD138+) plazmasejtek immunglobulin-szekvenciáinak (CDR3) elemzése NGS módszerrel

Nagy genetikai variabilitás – szomatikus hipermutáció – szelekció

Epigenetics, DNA methylation, cancer, ageing (ELTE-SOTE)



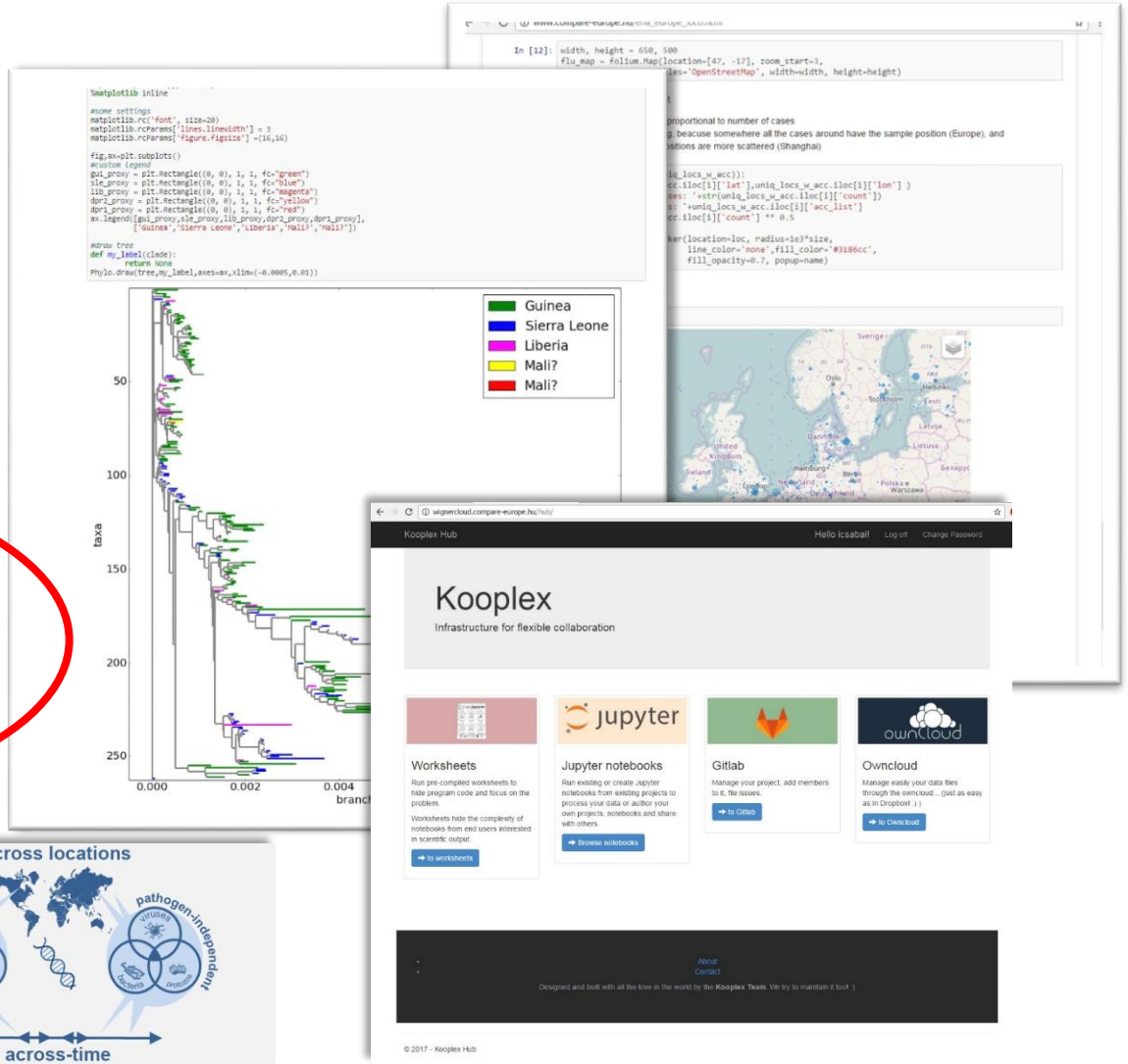


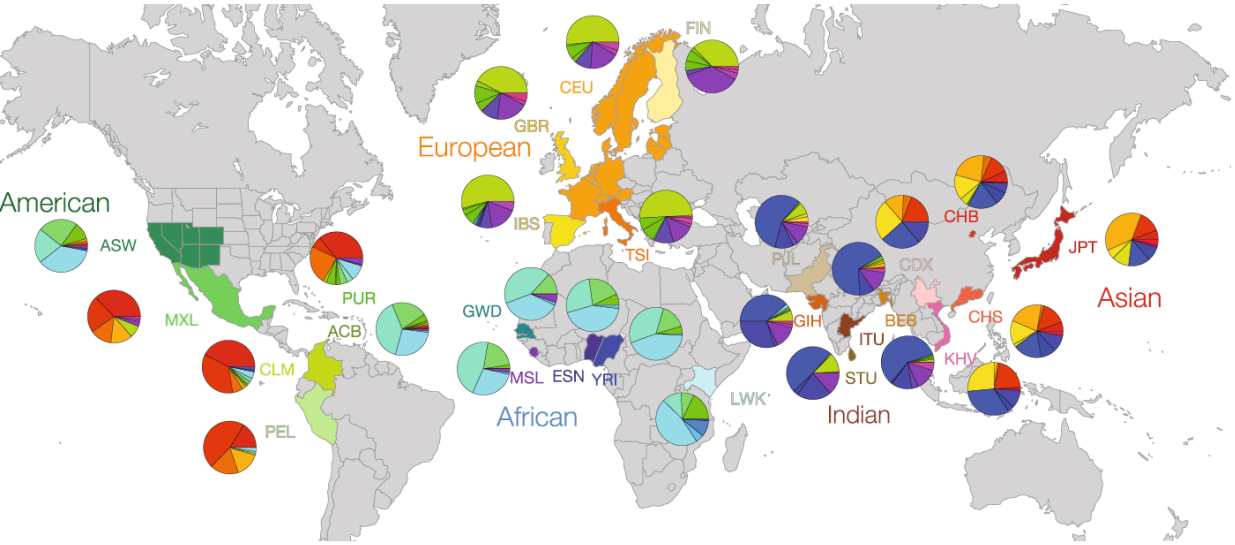


Palla et al. submitted 2020

Multidiszciplináris hazai és nemzetközi együttműködések

- FIEK_16-1-2016-0005: Biomarkerek (ELTE-MTA TTK-CRU-SERVIER)
- NVKP_16-1-2016-0004: Magyar onkogenom, folyadékbiopszia (SOTE-3DHISTECH-ELTE)
- NKFI OTKA 124881: DNS-javító mechanizmusok (MTA TTK-ELTE)
- Novo Nordisk Multidisciplinary Synergy (Danish Cancer Society Research Center-DTU-Francis Crick Institute-ELTE)
- COMPARE EU H2020: Fertőző betegségek, vírusok, baktériumok, metagenomika (~15 nemzetközi partner, MTA Wigner FK Adatközpont)
- VEO H2020: Fertőző betegségek, vírusok, baktériumok, metagenomika (~15 nemzetközi partner, ELTE)

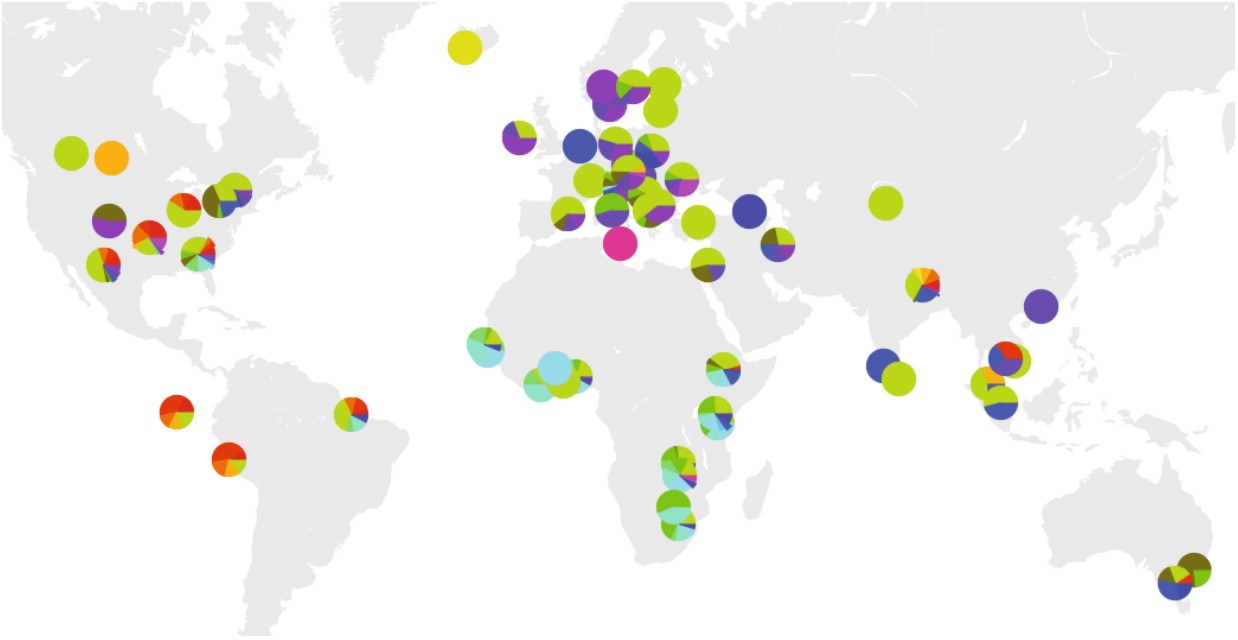


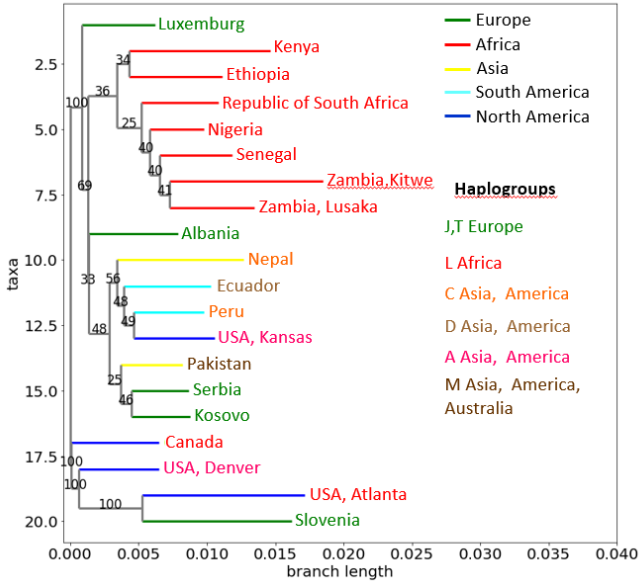


Sewage sequencing from 81 cities

Monitoring diseases, antimicrobiotic resistance … and human phylogeny

(Metagenome!)

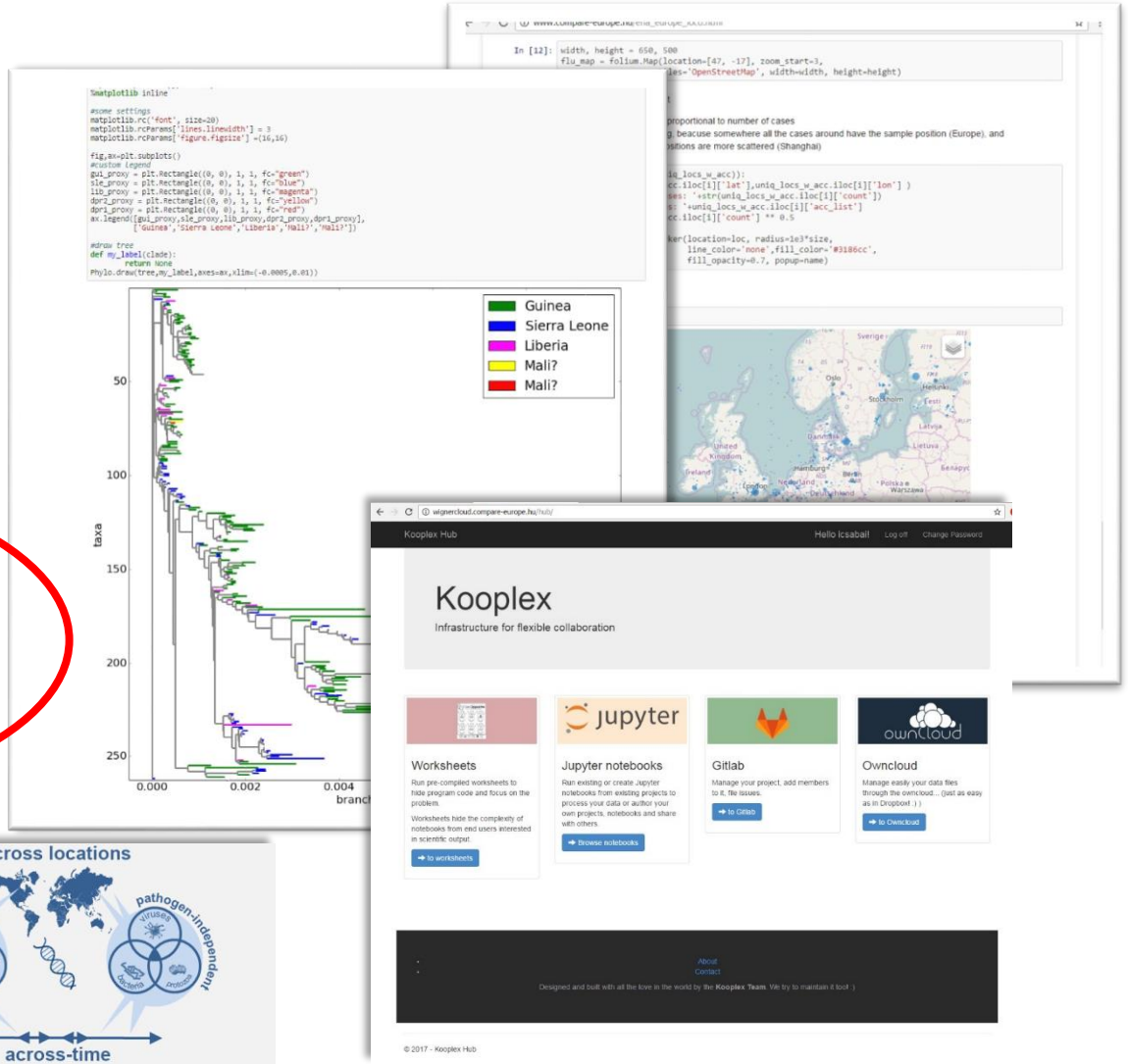




Multidiszciplináris hazai és nemzetközi együttműködések

- FIEK_16-1-2016-0005: Biomarkerek (ELTE-MTA TTK-CRU-SERVIER)
- NVKP_16-1-2016-0004: Magyar onkogenom, folyadékbiopszia (SOTE-3DHISTECH-ELTE)
- NKFI OTKA 124881: DNS-javító mechanizmusok (MTA TTK-ELTE)
- Novo Nordisk Multidisciplinary Synergy (Danish Cancer Society Research Center-DTU-Francis Crick Institute-ELTE)
- COMPARE EU H2020: Fertőző betegségek, vírusok, baktériumok, metagenomika (~15 nemzetközi partner, MTA Wigner FK Adatközpont)
- VEO H2020: Fertőző betegségek, vírusok, baktériumok, metagenomika (~15 nemzetközi partner, ELTE)

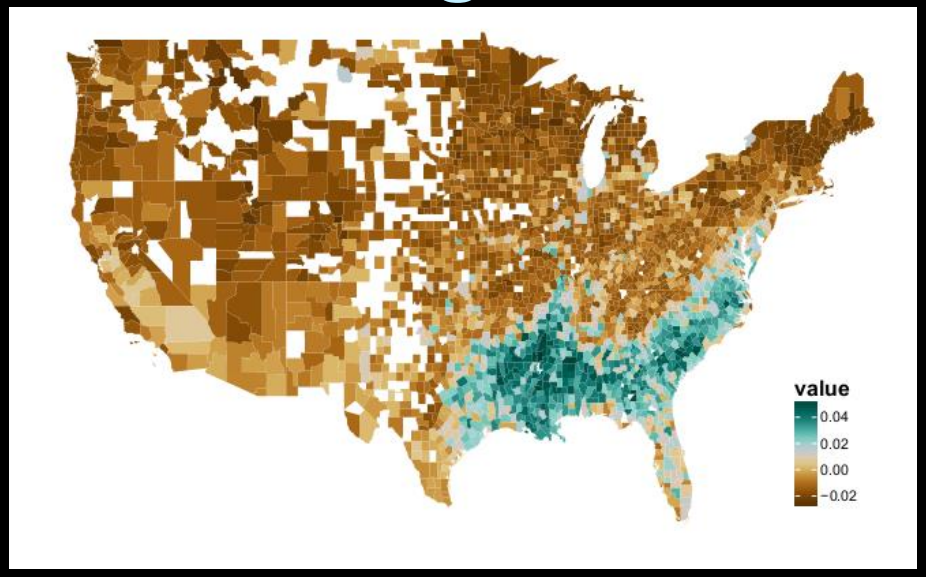
#nCoV,
#Wuhan,

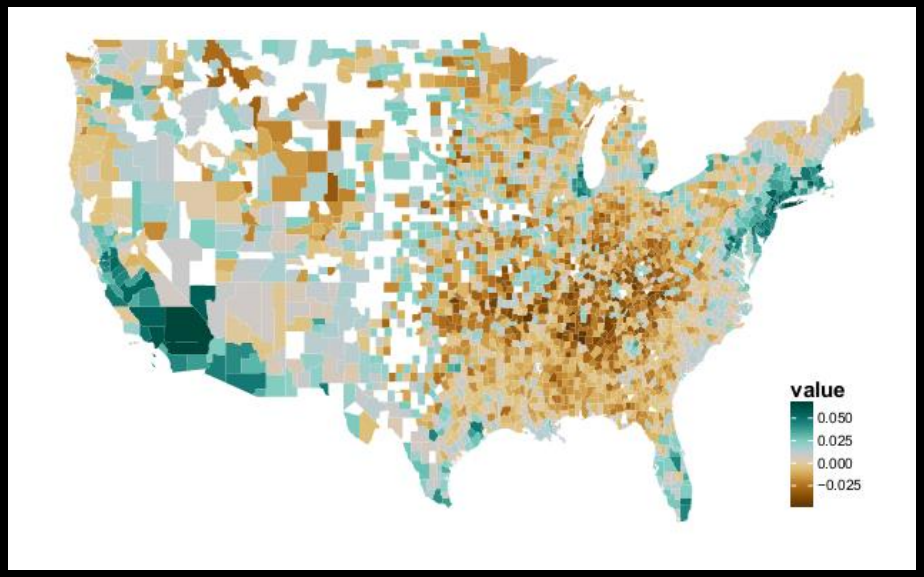


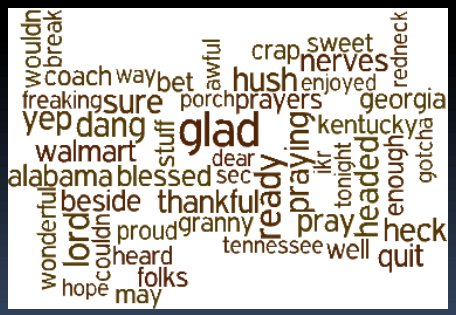
Social networks: TwitterDB

Principal dimensions: race, religion, urbanization

Data type:
Graph + text + geo



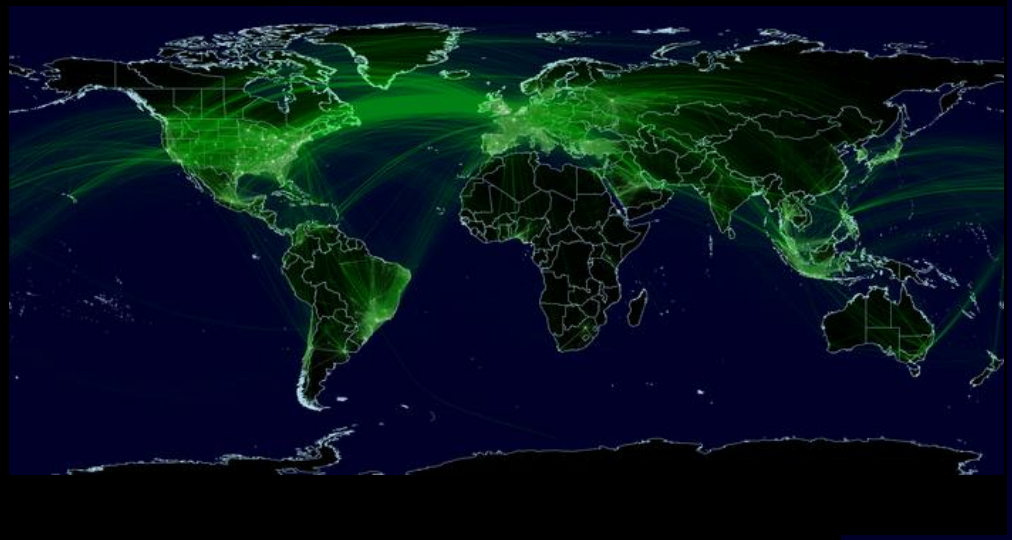


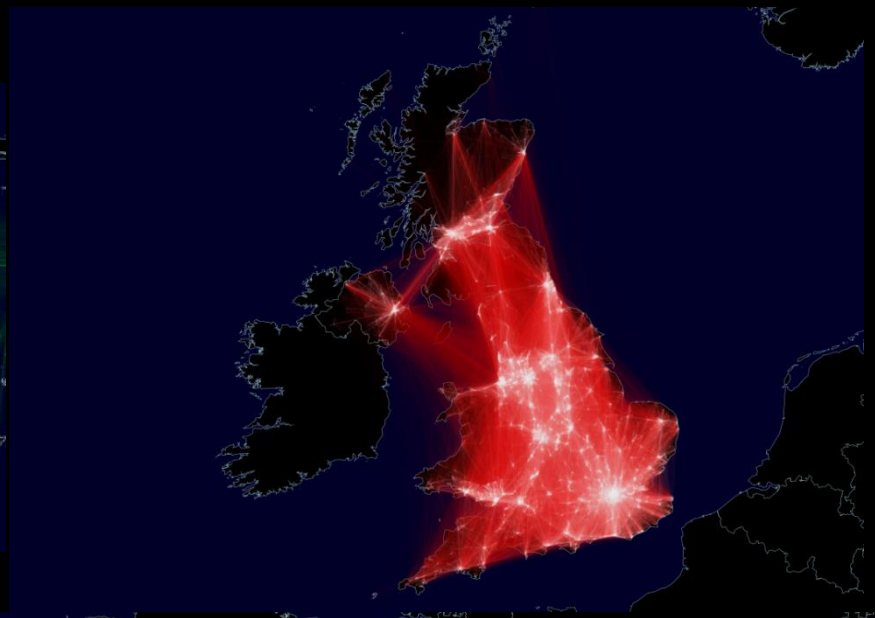


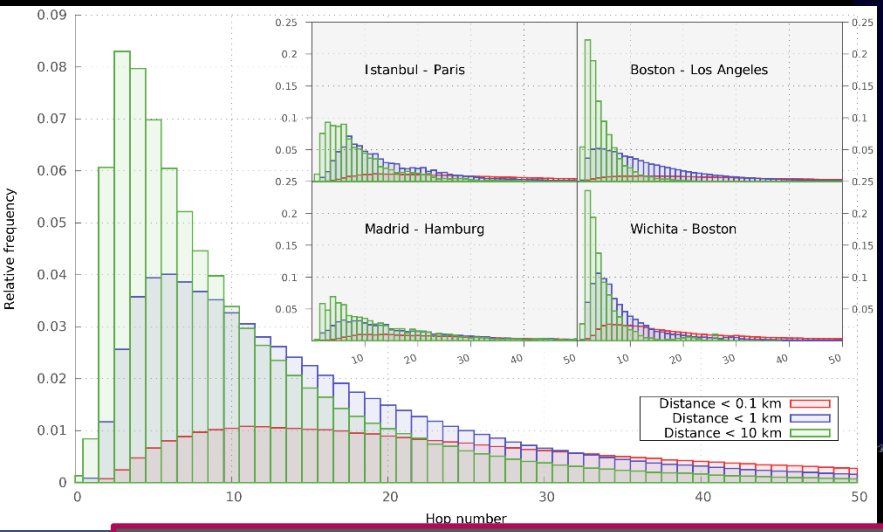
Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages; D Kondor, I Csabai, L Dobos, J Szule, N Barankai, T Hanyecz, T Sebok, Z Kallus, G Vattay; IEEE CogInfoCom) (2013)

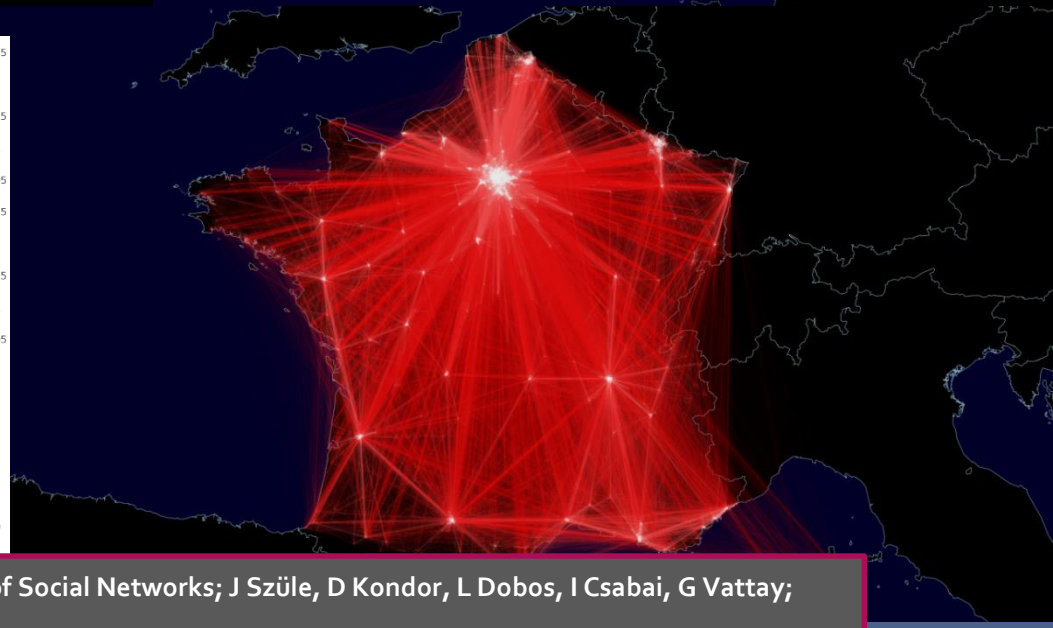
Bokányi Eszter, MSc thesis, ELTE TTK (2015), Bokanyi et al. submitted

Test Milgram's „6 degree" on Twitter









Lost in the City: Revisiting Milgram's Experiment in the Age of Social Networks; J Szüle, D Kondor, L Dobos, I Csabai, G Vattay; PloS one 9 (11), e111973 (2014)

A "GANGNAM-JÁRVÁNY": VÍRUSVIDEÓK A VILÁGHÁLÓN

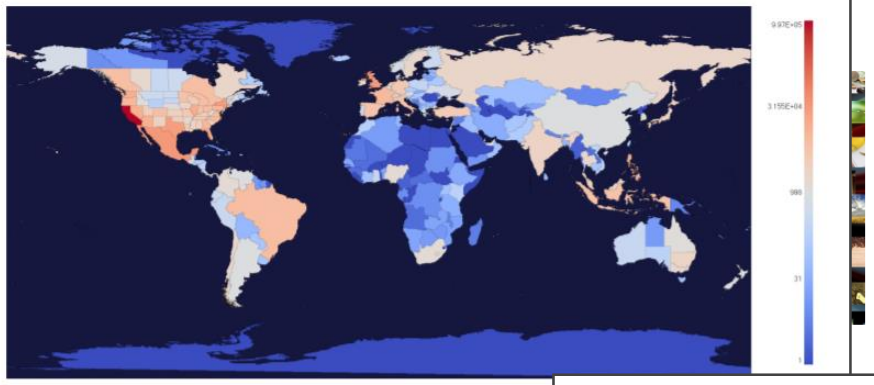


Fig. 2. Social connection weights between large ge
World. The map shows our 261 geo-political regions an
(mutual Twitter followers) between users in California
Colour codes the number of friendships with users in Cali
red means that Californians have $\sim 10^5$ friendships with
indicates that $\sim 10^0$ friendship connects them for example

Az ELTE Komplex Rendszerek Fizikája Tanszék kutatóinak – Kallus Zsófia, Kondor Dániel, Stéger József, Csabai István, Bokányi Eszter és Vattay Gábor – *How the 'Gangnam Style' Video Became a Global Pandemic* címmű tanulmányáról az MIT Technological Review közölt ismertetőt. A cikk a modernkori hírterjedés, a geoszociális és az online szociális hálózatok összefüggéseit vizsgálja.

A modernkori, fizikai és virtuális világunkat átszövő összekötöttség alapjaiban változtatta meg utazási és kommunikációs szokásainkat. Ennek megfelelően a földrajzi távolság már nem feltétlenül a legmegfelelőbb mértéke annak, hogy milyen messze van két város

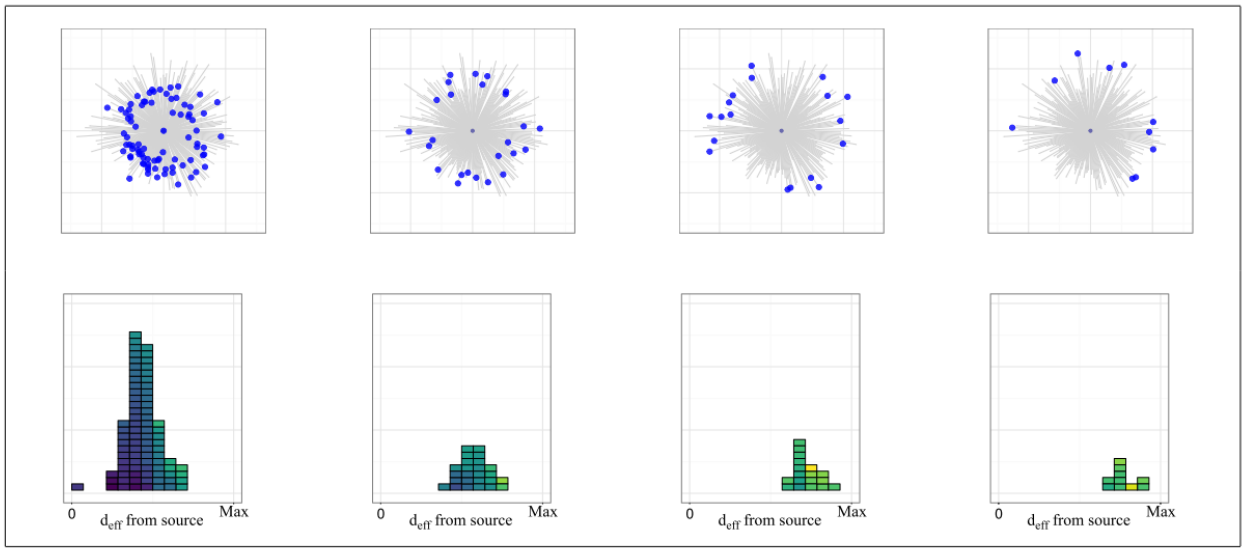


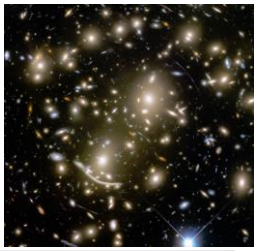
Fig. 4. Progressive stages of the pandemic. The spreading of the wave is shown in four progressive stages of the propagation. Each stage is defined by separate time slice of equal length. The nodes where the news has just arrived in that slice are first shown on the shortest path tree. Second, a corresponding histogram is created based on effective distances. Each rectangle represents one of the regional nodes and a common logarithmic color scale represents the number of users of the nodes (color scale of Fig. 5 is used).

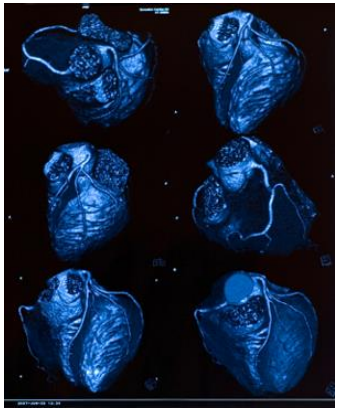
Big data, big simulations, machine learning

20th century

manual observations

21st century





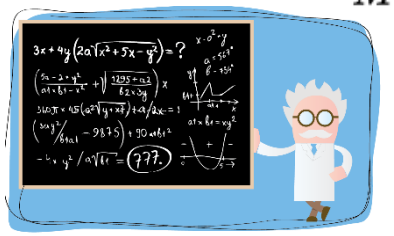
high throughput instruments

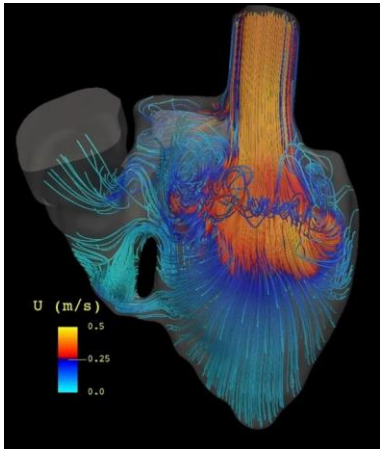
small data → big data

300 million galaxies
2.5 terapixels
3.2 gigabases,
37 trillion cells

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}.$$

$$MAP = (CO \cdot SVR) + CVP$$





simple equations

complex simulations

AI: Why inevitable? Why now?
Key challenges: amount of data and complexity of models

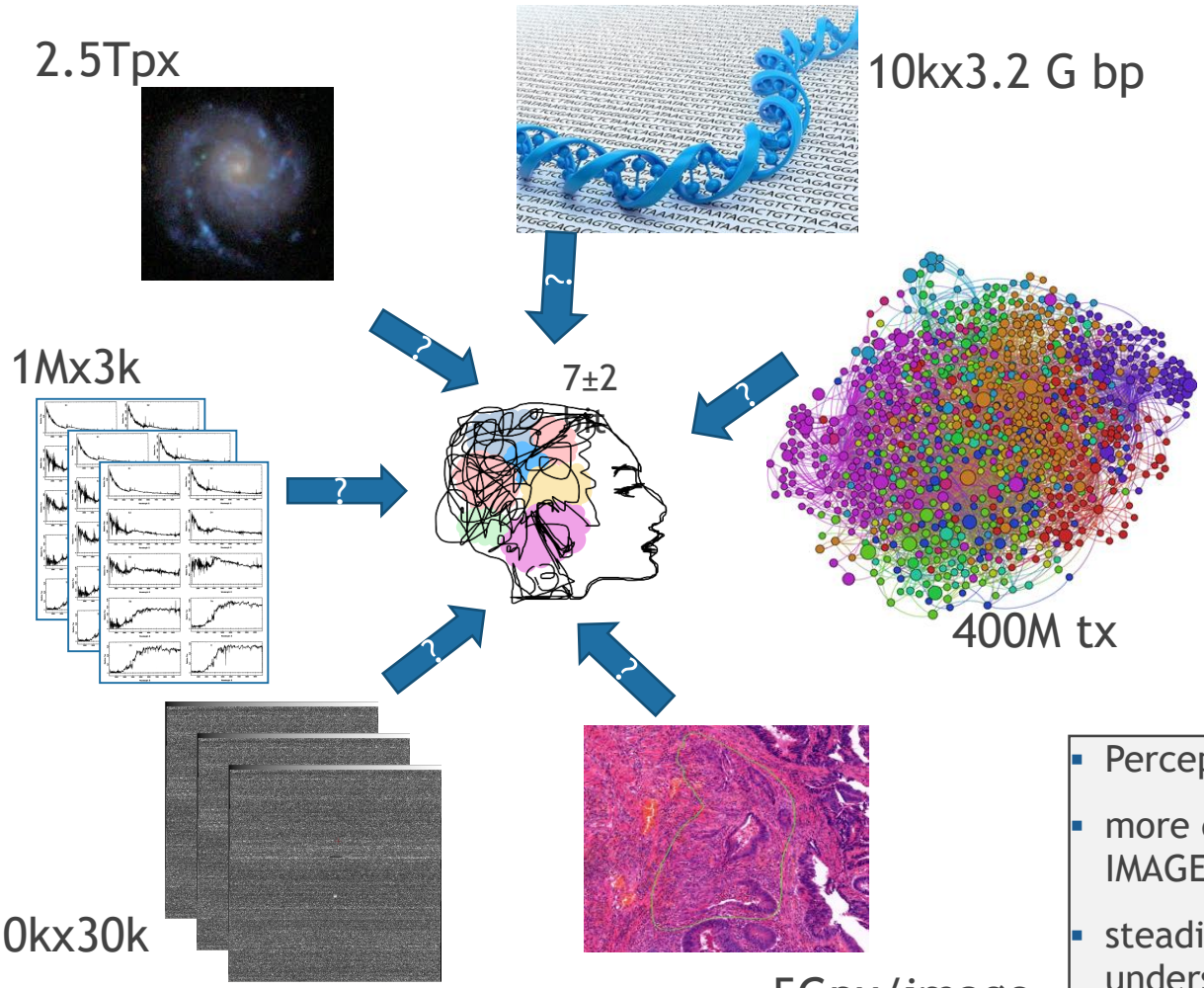
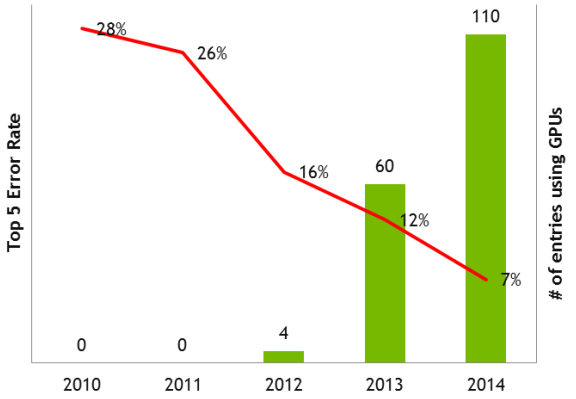


Image recognition progress



- Perceptron ‘57, Hopfield ’82, Backprop ‘86
- more data (MNIST’98 60k, CIFAR’10 60k, IMAGENET’10 14M)
- steadily improving models, deeper understanding of statistics/data/models
- **more compute power. GPU!!**
  - **V100 GPU: 100 TFlops**

AI: paradigm shift

Data →
Model →

Computer → Prediction

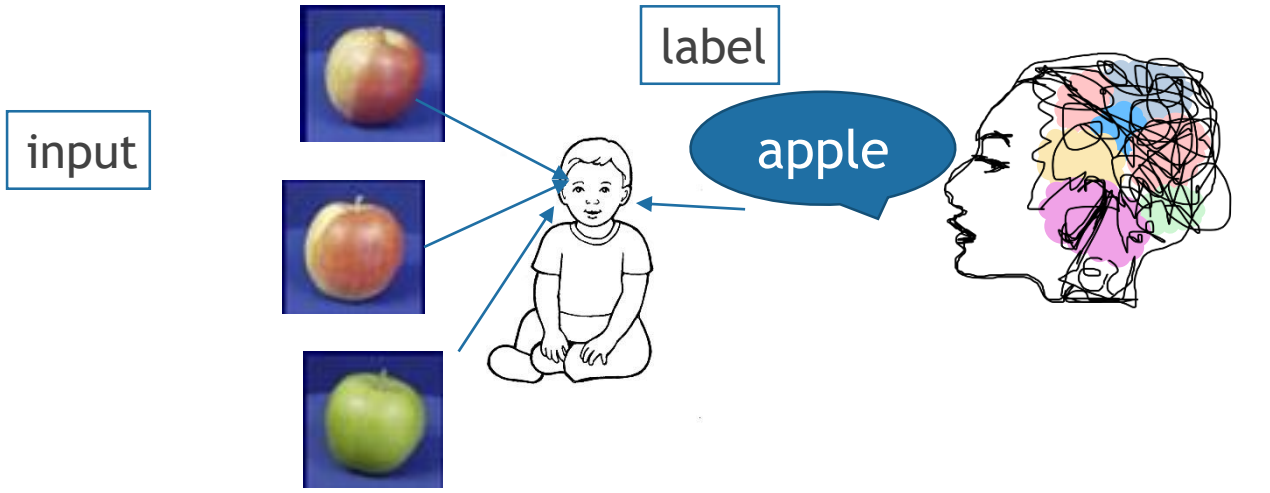
Example: Image recognition
Method: hand crafted features

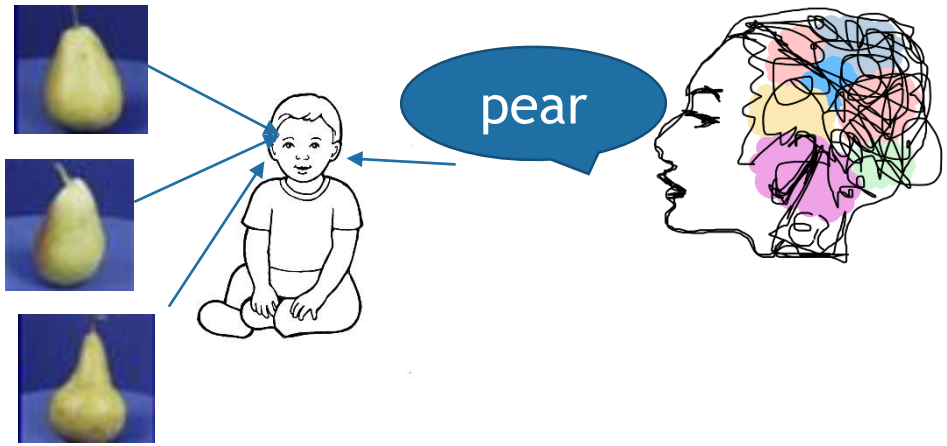
f( ) = “apple”
f( ) = “tomato”
f( ) = “cow”

IF color=red AND profile=smooth THEN type:=tomato
IF color=red AND HAS(horns) THEN type:=cow

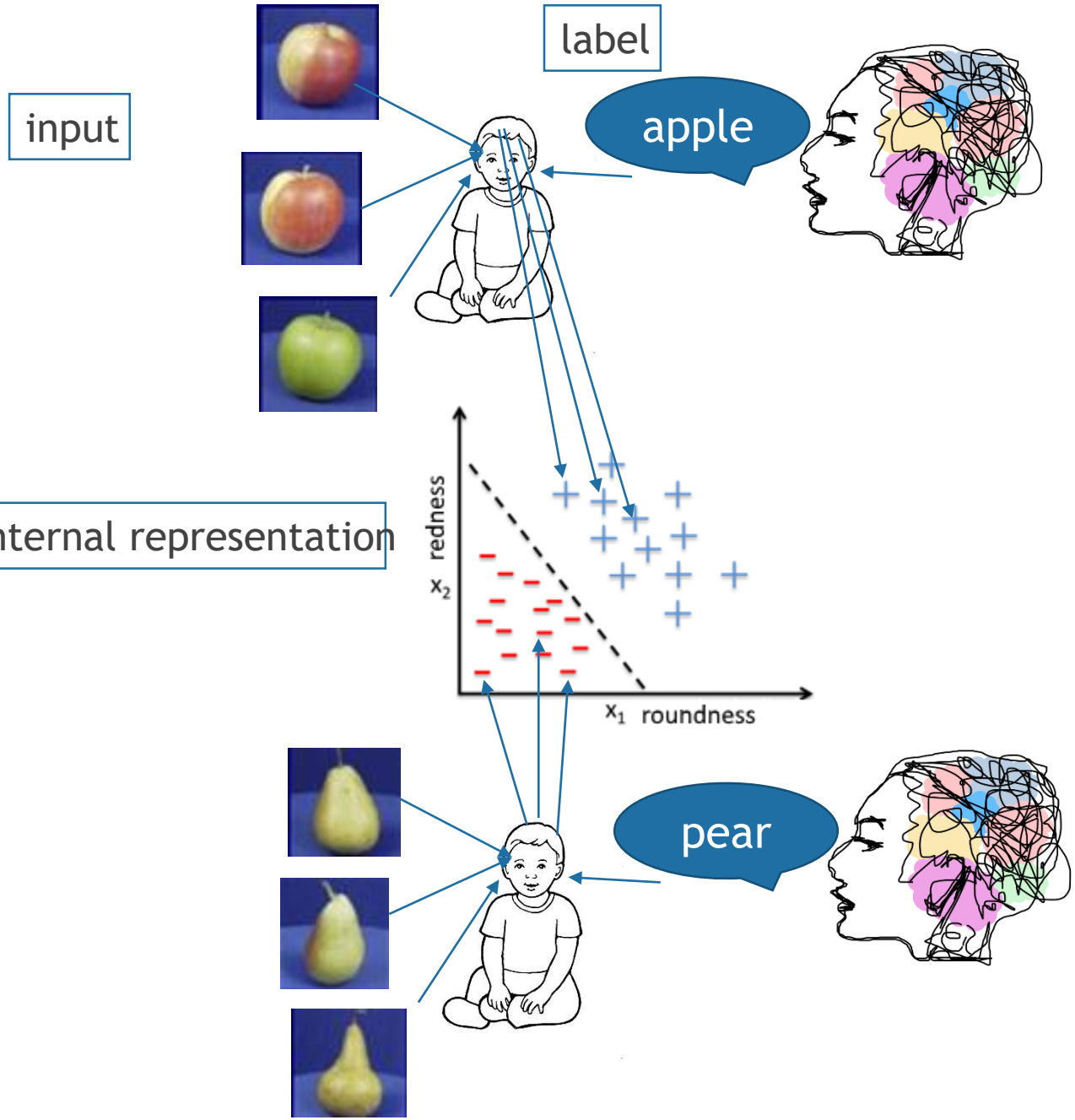
Data → Computer → Model → Prediction

Supervised learning

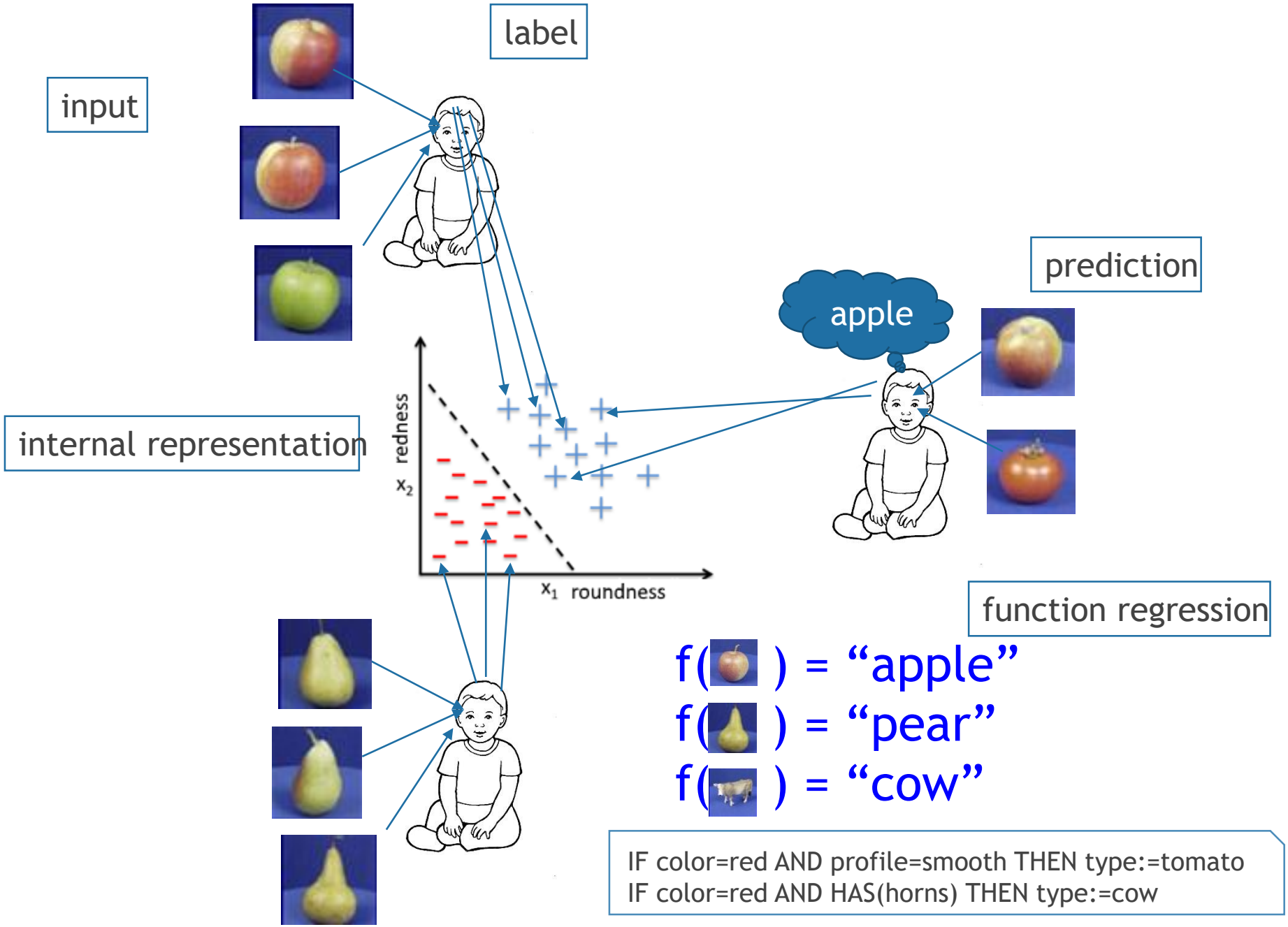




Supervised learning



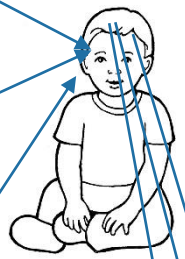
Supervised learning

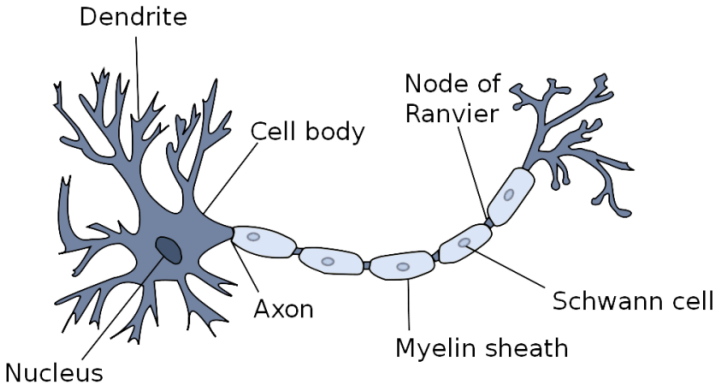


Supervised learning: neural net

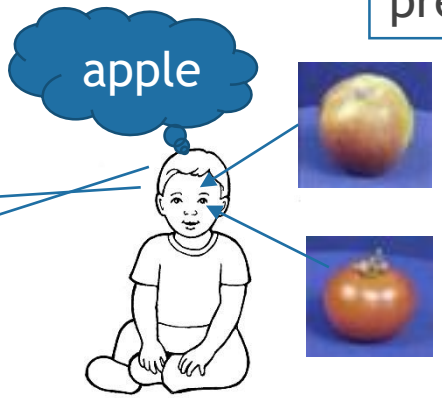
input

label

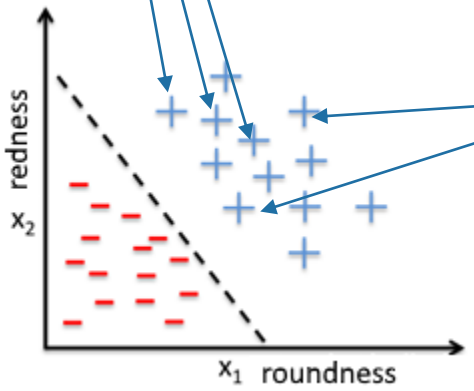




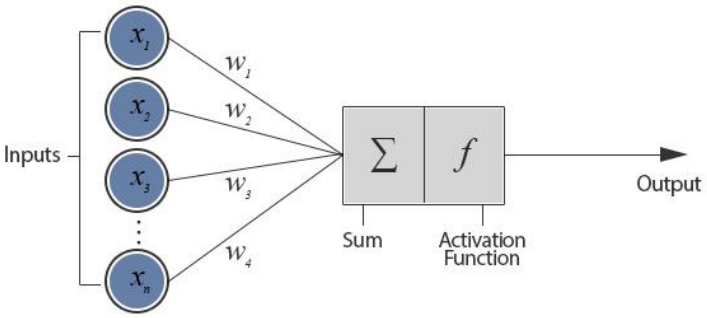
prediction



internal representation



$$y = f\left(\sum_i w_i x_i\right)$$



function regression

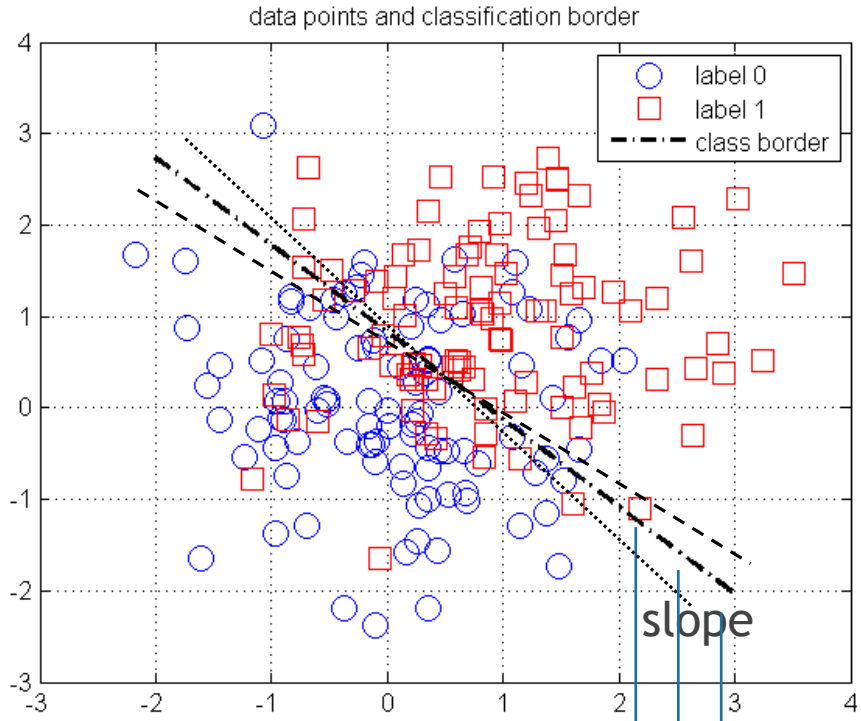
f( ) = “apple”

f( ) = “pear”

f( ) = “cow”

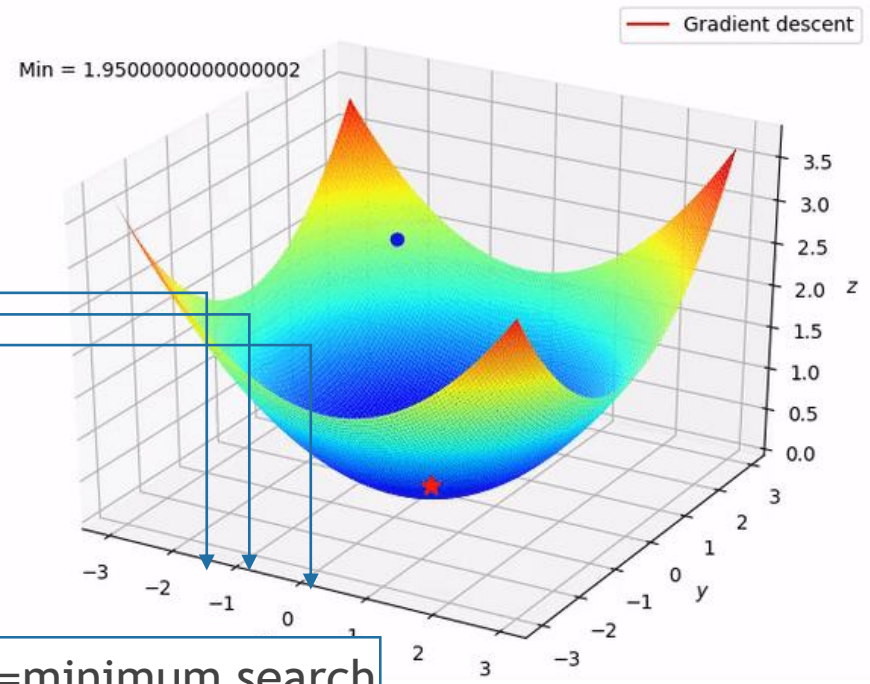
IF color=red AND profile=smooth THEN type:=tomato
IF color=red AND HAS(horns) THEN type:=cow

Learning -> loss function optimization



images -> points in N dim space

Loss = number of wrong categorizations

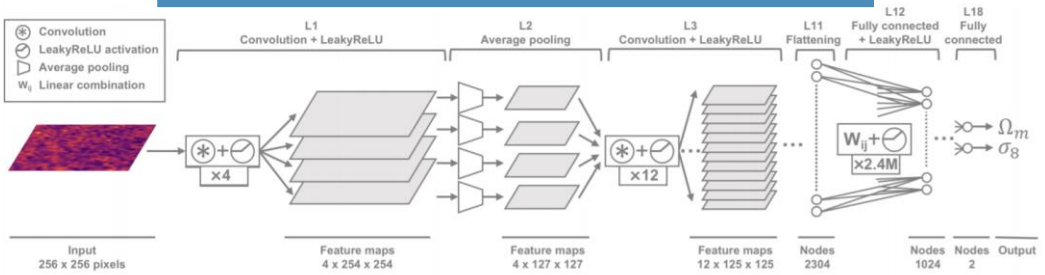


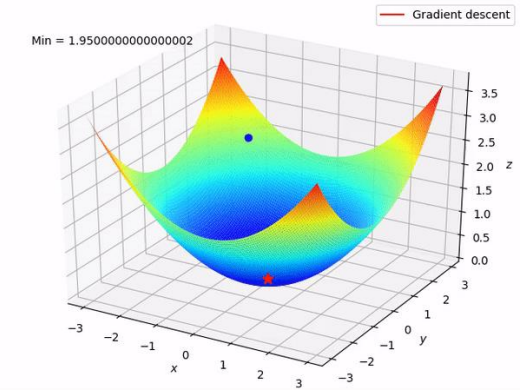
Learning=minimum search

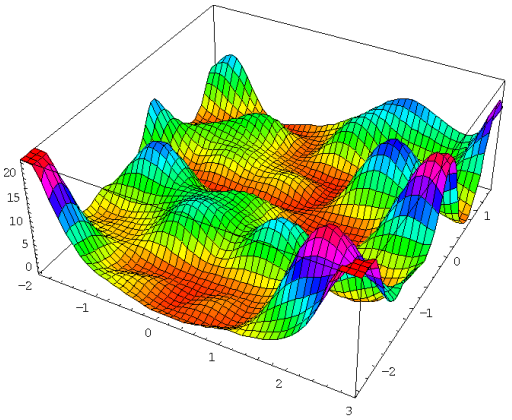
Challenges

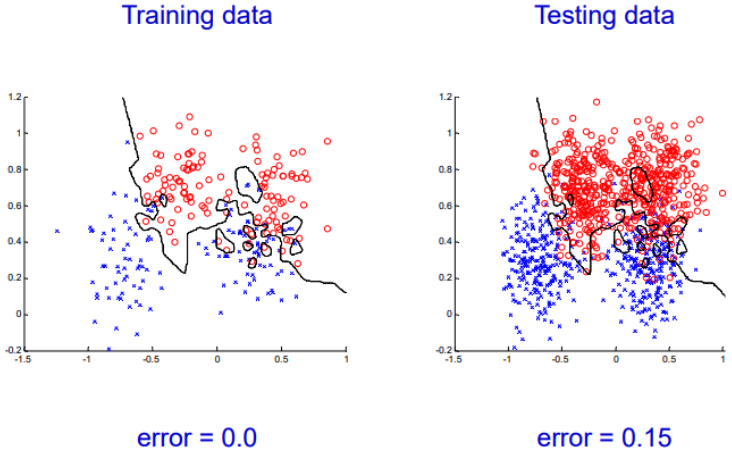
- Proper, **big enough training set**
- Representation of data
  (images, words, … -> vector space)
- Nonlinear optimization
- Model complexity
  - Accuracy
  - Generalization
- "**Black box**", trust
- …

Typical network: 2M adjustable parameter









AI Research, Education and Applications @ Eötvös University Dept. of Physics of Complex Systems

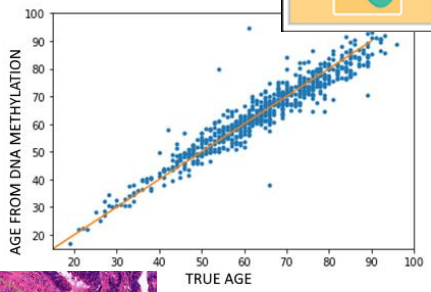
- Mutations -> **antibiotics resistance**
  Matamoros et al., Pataki et al. subm.
- Mobile sensors -> **Parkinson**
  Pataki @DREAM, Laki et al. 2016
- Quantum wave func.-> drug **toxicity**
  Biricz et al. in prep.
- **Medical imaging** -> breast cancer
  Ribli et al. @DREAM, Sci. Rep. 2018
- Weak lensing map -> **cosmology** parameters
  Ribli et al. Nature Astro. 2018, MNRAS 2019
- **Explainable AI**
  Ribli et al. in prep , Patent subm. 2019
- **Control of aging related methylation networks**
- **Pathology** images
  SOTE TKP collab.
- **Quantum neural computing**
- **MSc, PhD courses**
  http://datascience.elte.hu

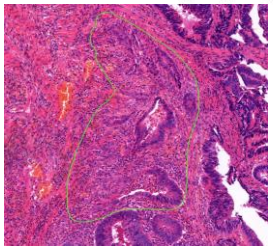
Gergely Palla lecture tomorrow 14:50

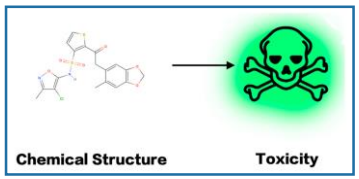
nature astronomy

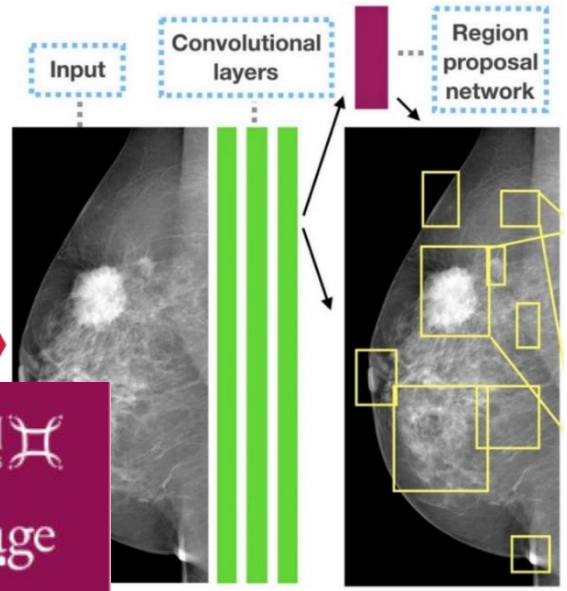
An improved cosmological parameter inference scheme motivated by deep learning

Dezső Ribli, Bálint Ármin Pataki & István Csabai
Nature Astronomy **3**, 93–98 (2019) | Download Citation









Solving analytically untraceable hard inverse problems

NIH DREAM CHALLENGES Sage Winner

nature.com > scientific reports > articles > article

SCIENTIFIC REPORTS

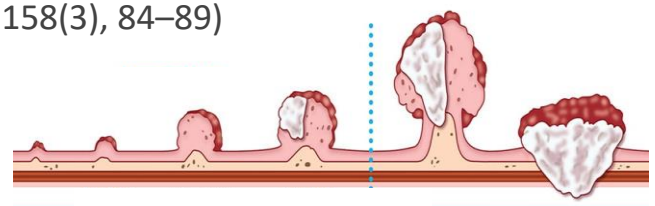
Detecting and classifying lesions in mammograms with Deep Lea...

Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner & István ...

MOST POPULAR

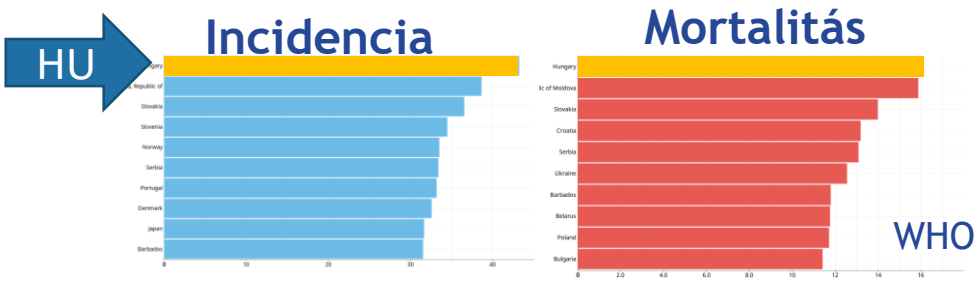
Colorectalis daganat patológia deep learning

- Világviszonylatban Magyarországon a leggyakoribb
- >10,000 új eset, >5,000 halál/év (Orv. Hetil., 2017, 158(3), 84–89)
- 2-10 év alatt alakul ki. Korai detekció!



https://fightcolorectalcancer.org/prevent/colon-polyps/

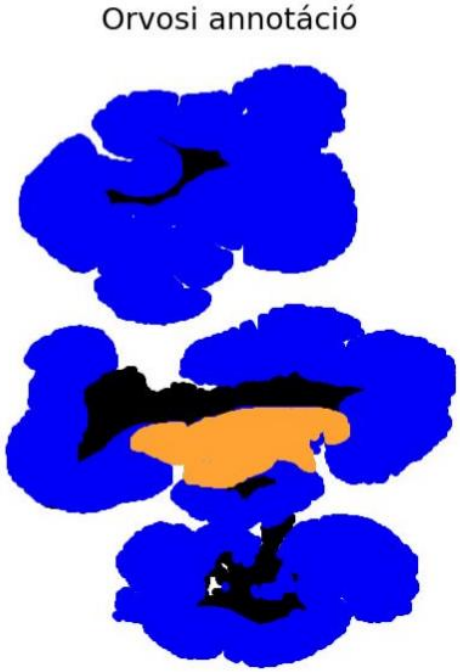
HU

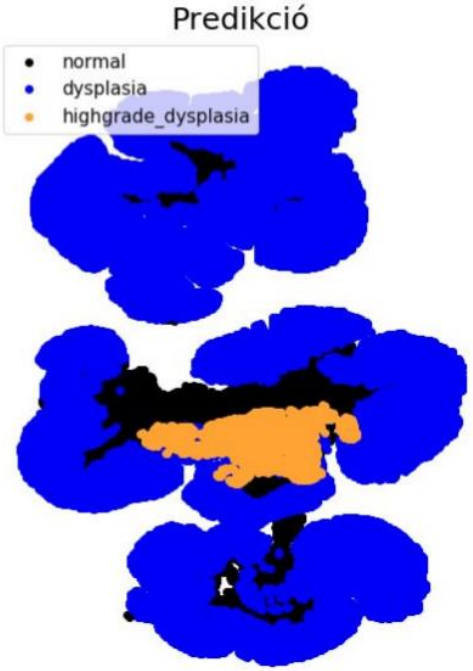


WHO Globocan 2018

Kellően nagy, jól annotált tanító halmaz
a szűk keresztmetszet a gépi tanításhoz!

>2,000 whole slide kép
80,000 x 60,000 pixel, 15GB
Részletes annotáció

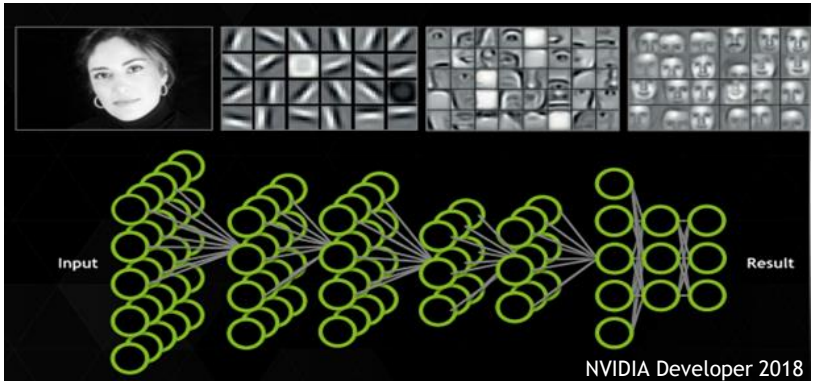


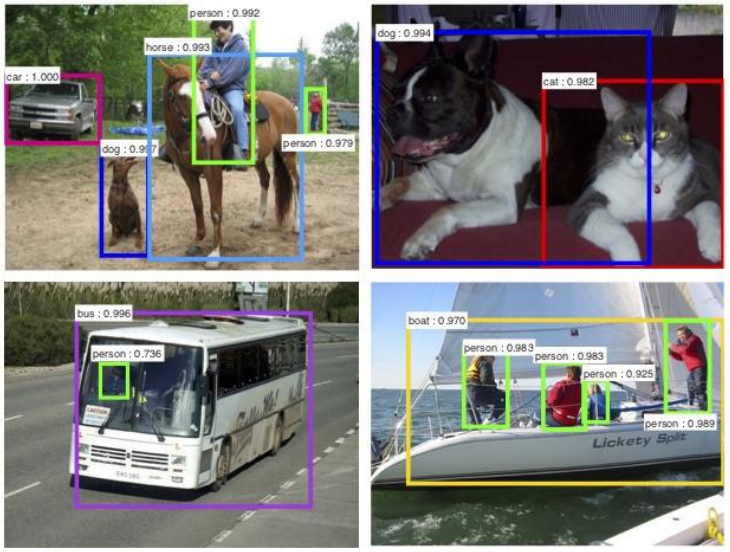


SOTE-ELTE-3DHISTECH, Szócska M, Kiss A, Kontsek E, Gyöngyösi B, Pollner P, Pataki B, Olar A, …

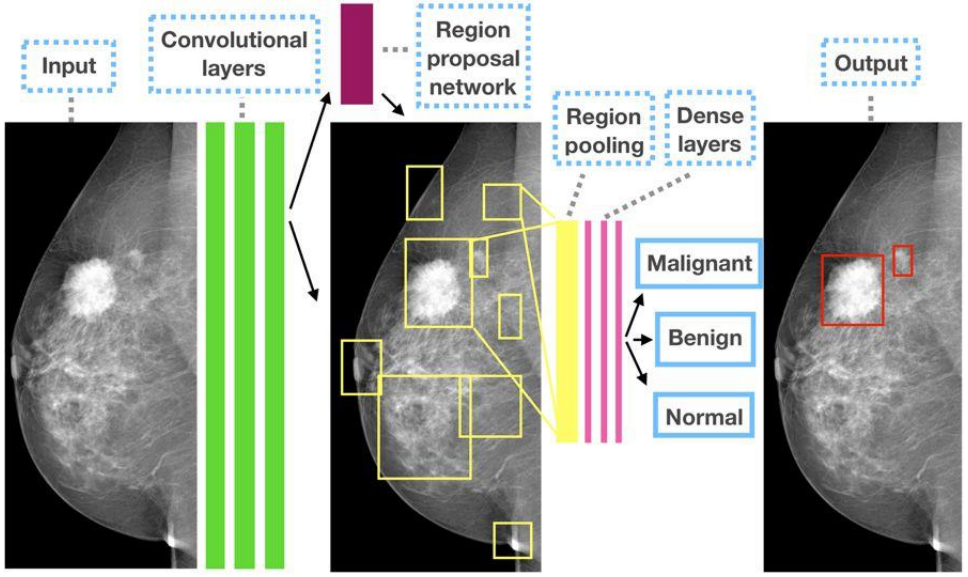
Mammográfia deep learning (Faster R-CNN)

- The Digital Mammography DREAM challenge
  - 1200 versenyző
  - Ribli Dezső, legjobb végső eredmény
  - egyetlen lokalizációt végző módszer
  - AUC = 0.95
- Nature Scientific Reports (2018)
  - 125 citáció
  - Sci. Rep. 17000 cikkből 30. legolv.
  - hazai kórházak, még több adat
  - engedélyeztetés, bevezetés ….





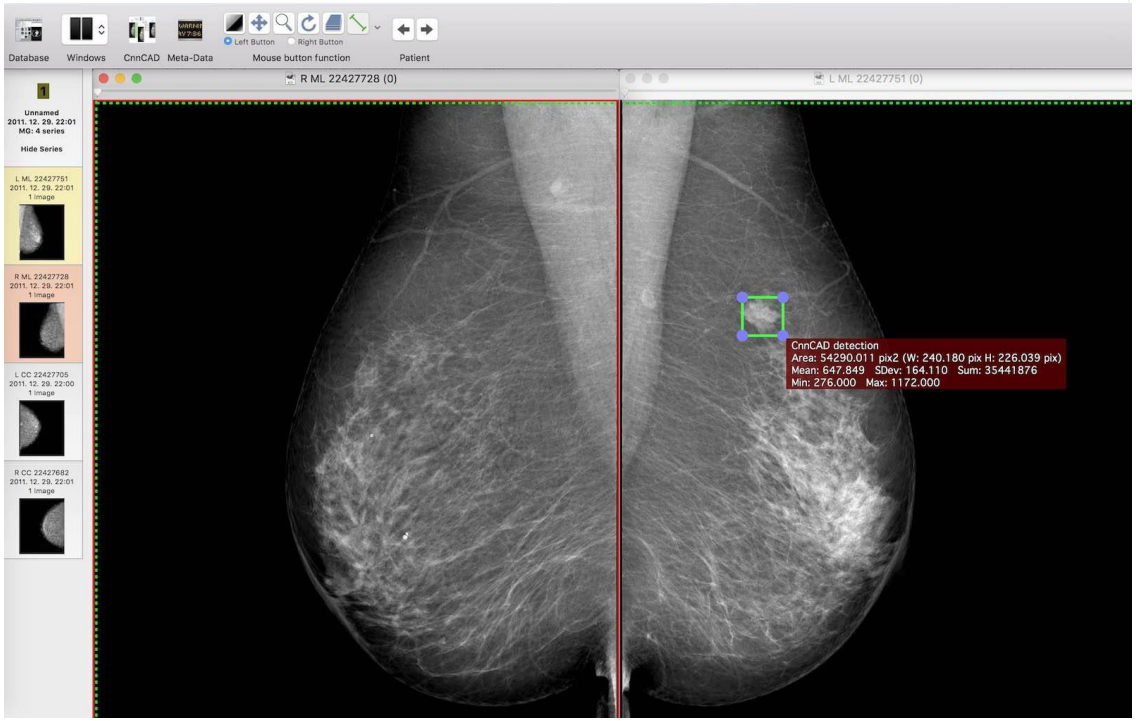
Ren et al. 2016



D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai. "Detecting and classifying lesions in mammograms with deep learning." Scientific reports (2018)

Mammography with deep learning (Faster R-CNN)

- Open source software
  - https://riblidezso.github.io/frcnn_cad/
  - 50 visitors/week
  - runs on a regular laptop with GPU
  - plugin for OsiriX/Horos medical image viewer



riblidezso / frcnn_cad

Unwatch 1 | Star 1 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Computer aided detection using Faster-RCNN

Add topics

29 commits | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

riblidezso Update README.md — Latest commit 7ff9c6b 2 days ago

data	Update README.md	6 days ago
results	Update README.md	6 days ago
README.md	Update README.md	2 days ago
demo.ipynb	demo notebook	18 days ago
vgg16_frcnn_cad_test.prototxt	uploaded prototxt	18 days ago

README.md

Computer aided detection with Faster-RCNN

A VGG16 based Faster-RCNN detector trained to detect breast cancer lesions on mammogram images.

For more details and results please see the article on arXiv: https://arxiv.org/abs/1707.08401

Model weights

Model weights can de downloaded from here, the models is licensed with Creative Commons Attribution + Noncommercial + ShareAlike license License CC BY-NC-SA 4.0

- http://dkrib.web.elte.hu/cad_faster_rcnn/vgg16_frcnn_cad.caffemodel

OsiriX/Horos plugin

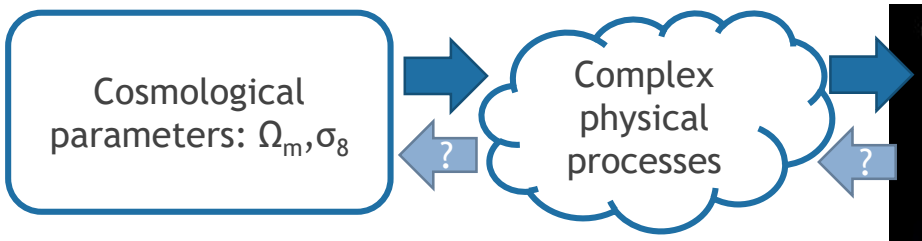
I wrote a plugin for the popular medical image viewer: OsiriX/Horos. This is a simplified version of the model. It is based on CoreML. In order to run it you will need:

- A mac
- High Sierra (10.13) operating system (becase of CoreML)
- Horos (free) or OsiriX medical image viewer.

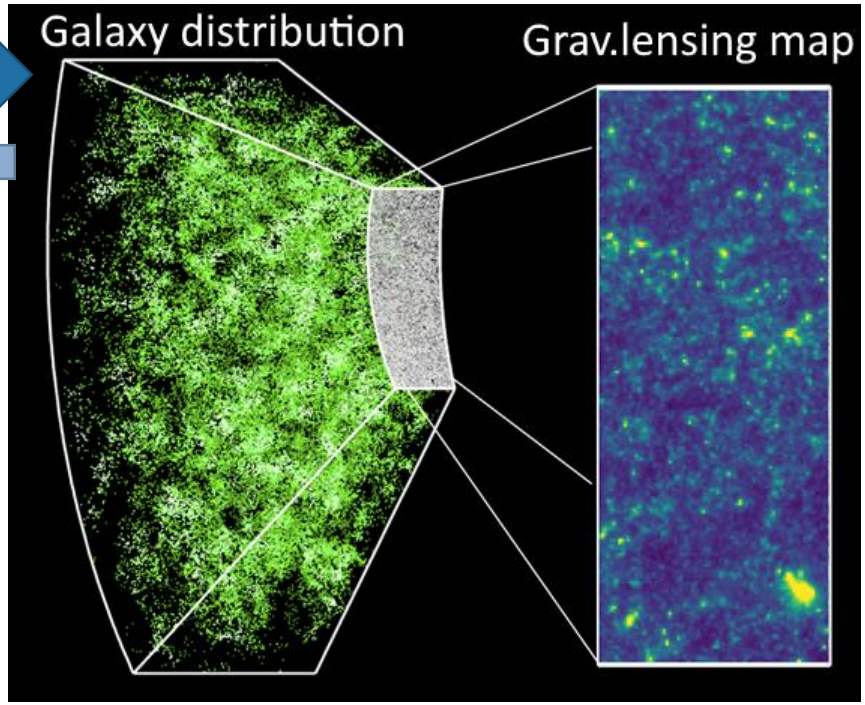
Installation: Download the plugin, unzip it, click on the installer, and approve it. Restart OsiriX/Horos. The plugin will appear in the menu of OsiriX/Horos: Plugins/Image Filters/CnnCAD. You can also add it to the toolbar of the 2D viewer.

Usage: Open the images in the 2D viewer and click on the menu item: Plugins/Image Filters/CnnCAD. A pop-up will appear

Cosmological parameters from gravitational lensing
Learning new tricks from deep learning

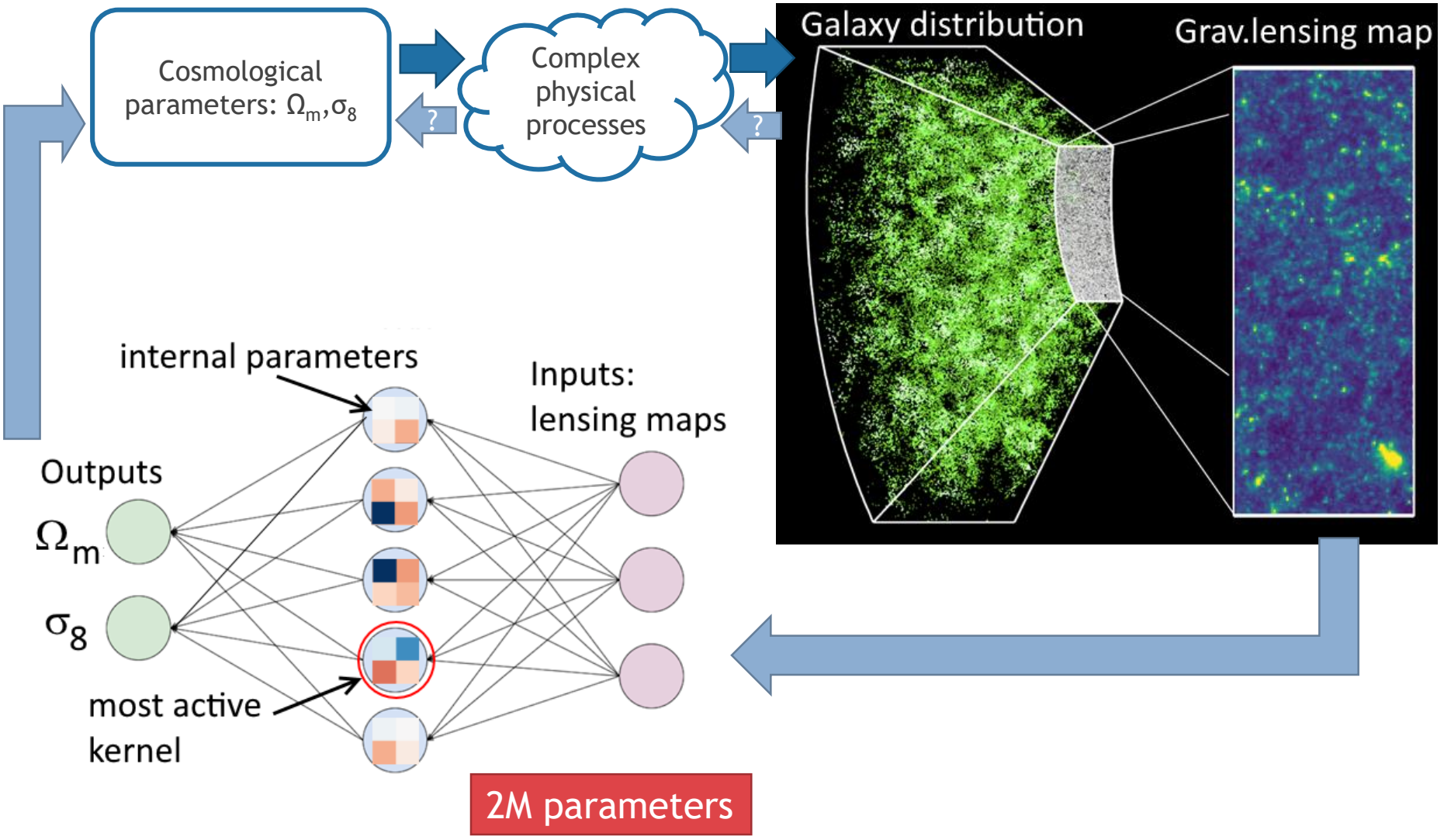


Hard inverse problem!

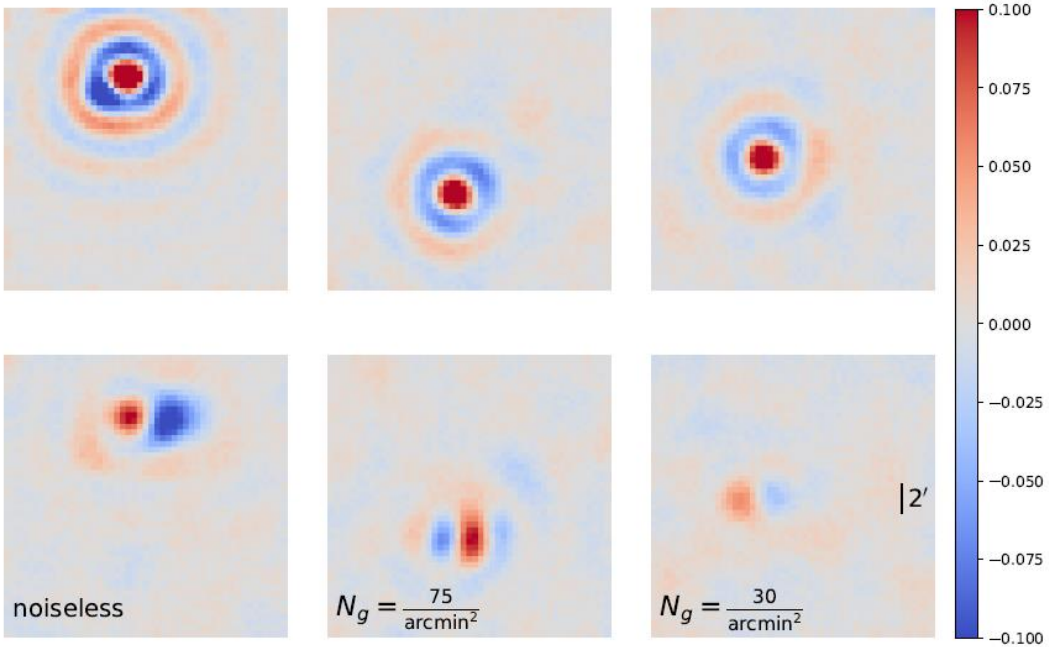


Ribli et al. MNRAS 2019
Ribli et al. Nature Astr. 2019

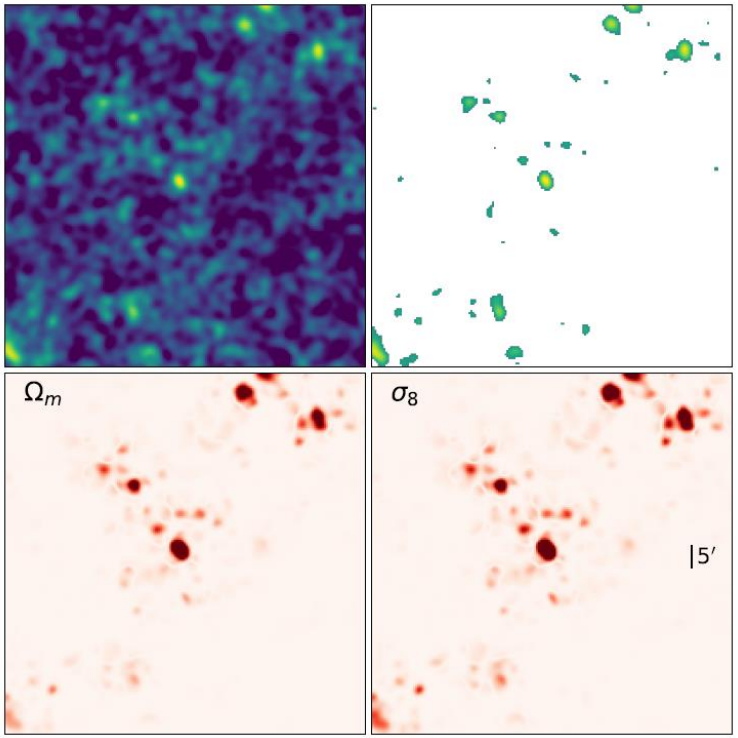
Cosmological parameters from gravitational lensing
Learning new tricks from deep learning



Learned kernels: dark matter halo profile expansion

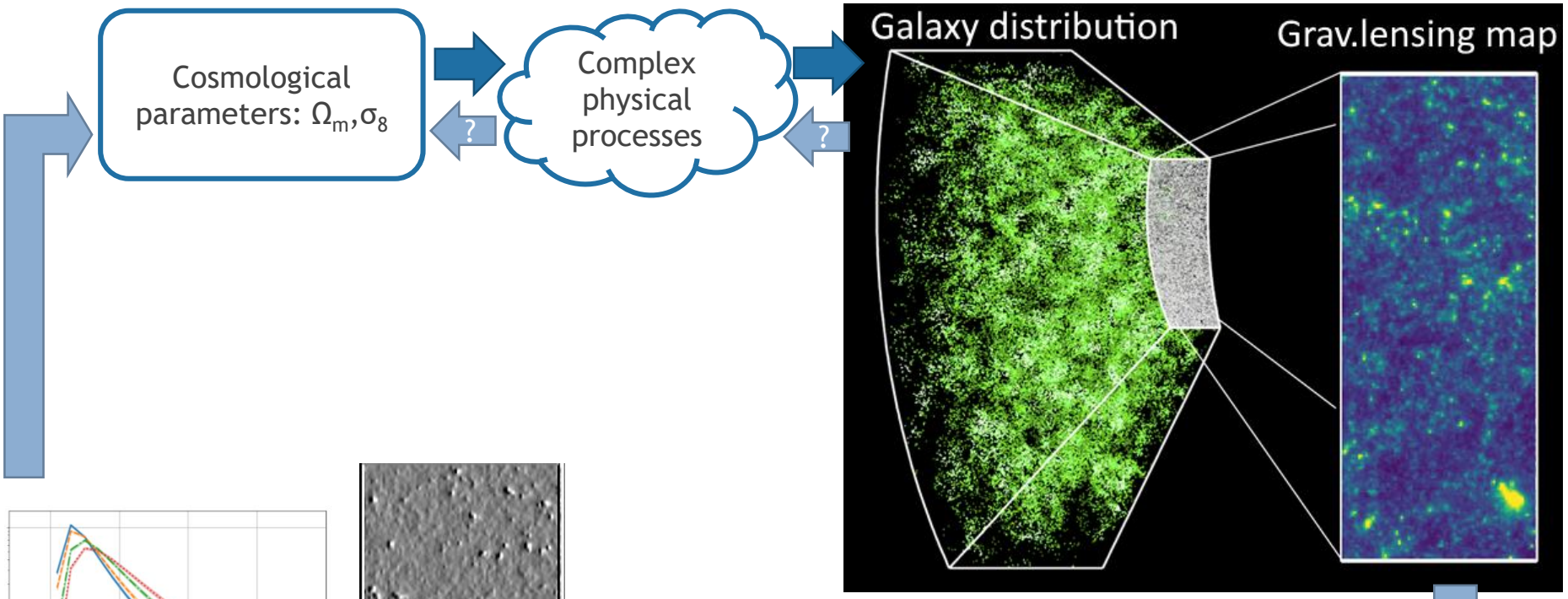


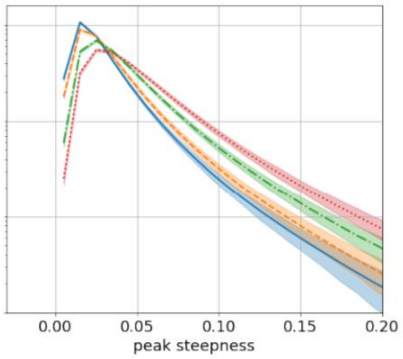
Instead of Fourier power spectrum:
information from halo profiles

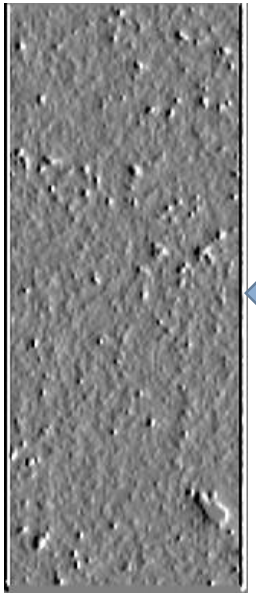


Attention focus of the network with Layer-wise Relevance Propagation

Cosmological parameters from gravitational lensing
Learning new tricks from deep learning

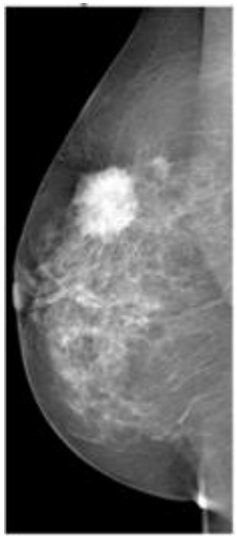


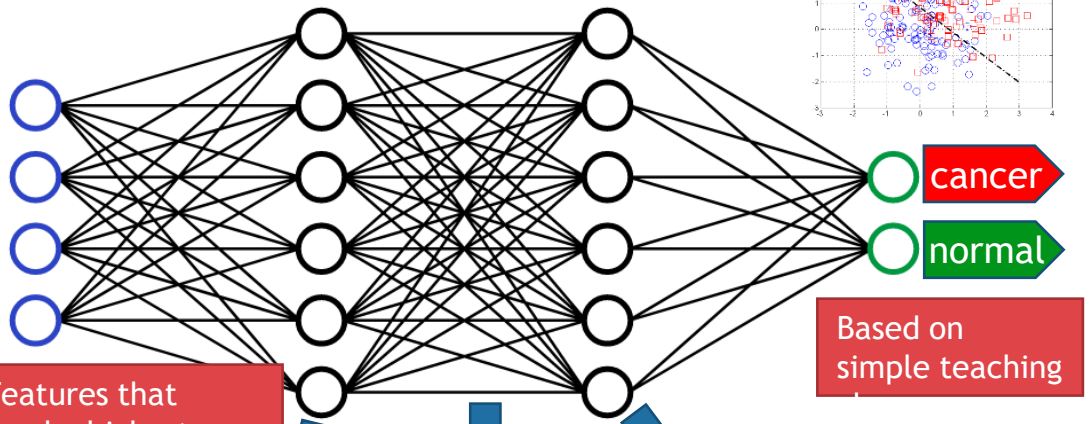




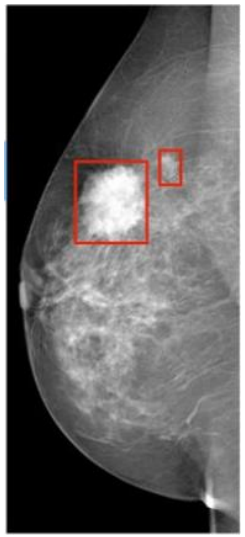
Learning from deep learning:
New simple method
Few parameters, robust, fast, understandable.

Explainable AI: automatic classification enhancement

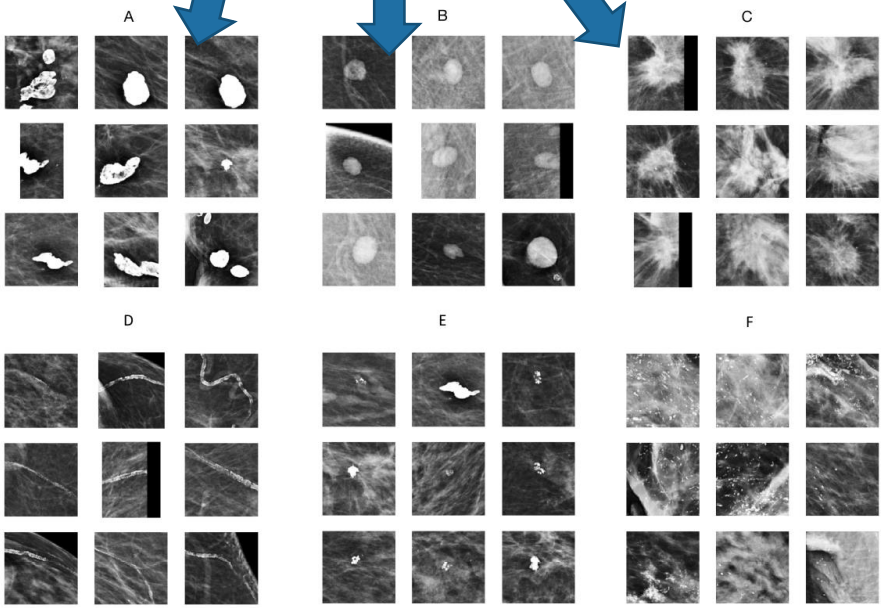




Based on simple teaching

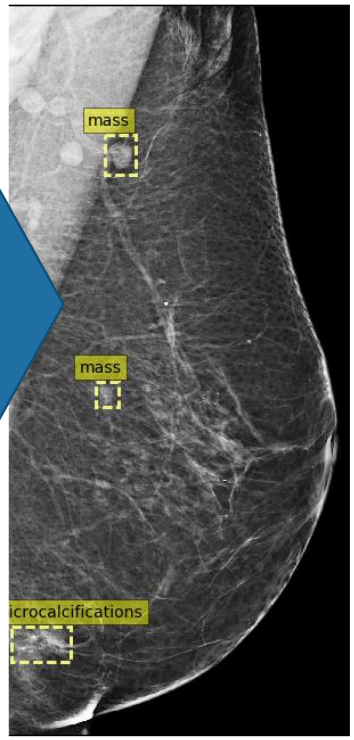


Features that invoke highest



A. Large calcification
B. Oval mass
C. Spiculated mass
D. Calcified vessel
E. Calcification
F. Clusetered micro-calcifications

Automatic labels „discovered ” by the network



Interpretable, trustworthy, for

Ribli et al. in prep , Patent subm. 2019

Any sufficiently advanced technology is indistinguishable from magic.

(Arthur C. Clarke)

Indeed, understanding the laws of **mechanics** made us able to build **pyramids and cathedrals**, based on the laws of **thermodynamics** the invention of the steam engine empowered us to cross oceans and continents and today we all have „**seven-league boots**" in our garages. Understanding **electrodynamics and quantum mechanics** brought us the transistor that is at the heart of the Internet and the modern „**magic mirrors**", the mobile phones.

What miracles will the advancements of **genomics** together with **machine learning** bring? And what kind of challenges?

NEW PARADIGMS

EDUCATION: WE NEED NEW SCIENTIST WHO HAVE PROFESSIONAL SKILLS BOTH IN THEIR DISCIPLINES AND IN MODERN INFORMATION TECHNOLOGIES.

HEALTH DATA: COMMON GOOD. GREAT OPPORTUNITIES, GREAT RESPONSIBILITY.

Fizikus:
Tudományos adatanalitika MSc spec,
BSc, MSc, PhD thesis

BIOINFORMATIKA MSC Spec !!!

István Csabai
ELTE Dept. of Physics of Complex Systems
csabai@elte.hu
http://complex.elte.hu/~csabai/