Mesterséges intelligencia és adat-intenzív tudományok: a galaxisoktól a koronavírusig

Data-intensive approach and AI in sciences: from galaxies to COVID

ISTVAN CSABAI

DEPARTMENT OF PHYSICS OF COMPLEX SYSTEMS

ELTE EÖTVÖS LORÁND UNIVERSITY, BUDAPEST

Acknowledgement:
Ministry of Innovation and Technology NRDI Office, OTKA NN 129148, FIEK_16-1-2016-0005, 2020-4.1.1.-TKP2020, , 2020-1.1.2-PIACI-KFI-2021-00298, H2020 VEO, Horizon EU BY-COVID, Health Security / Data-driven Health & MILAB Hungarian AI National Laboratory

MAFIHE Téli iskola, 2023.02.02.

History of intelligence / (data) science

Model

OBSERVATION

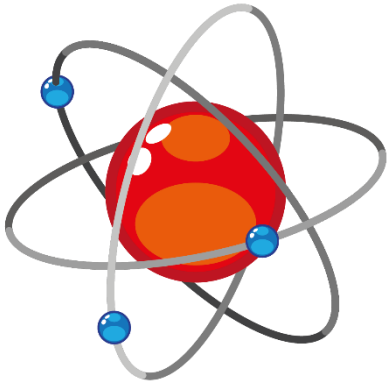
World

History of intelligence / (data) science

Model

OBSERVATION

World



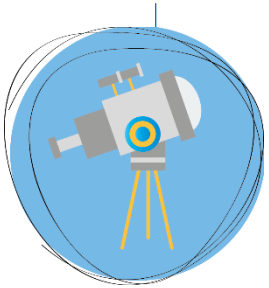
History of intelligence / (data) science

crutch for senses

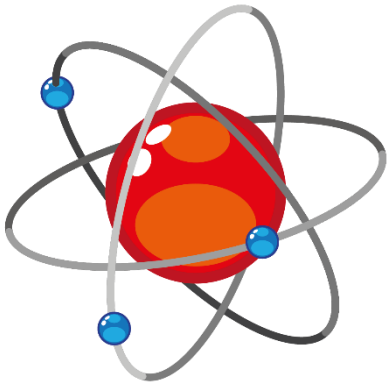
Model

OBSERVATION

Instruments



World



History of intelligence / (data) science

Model

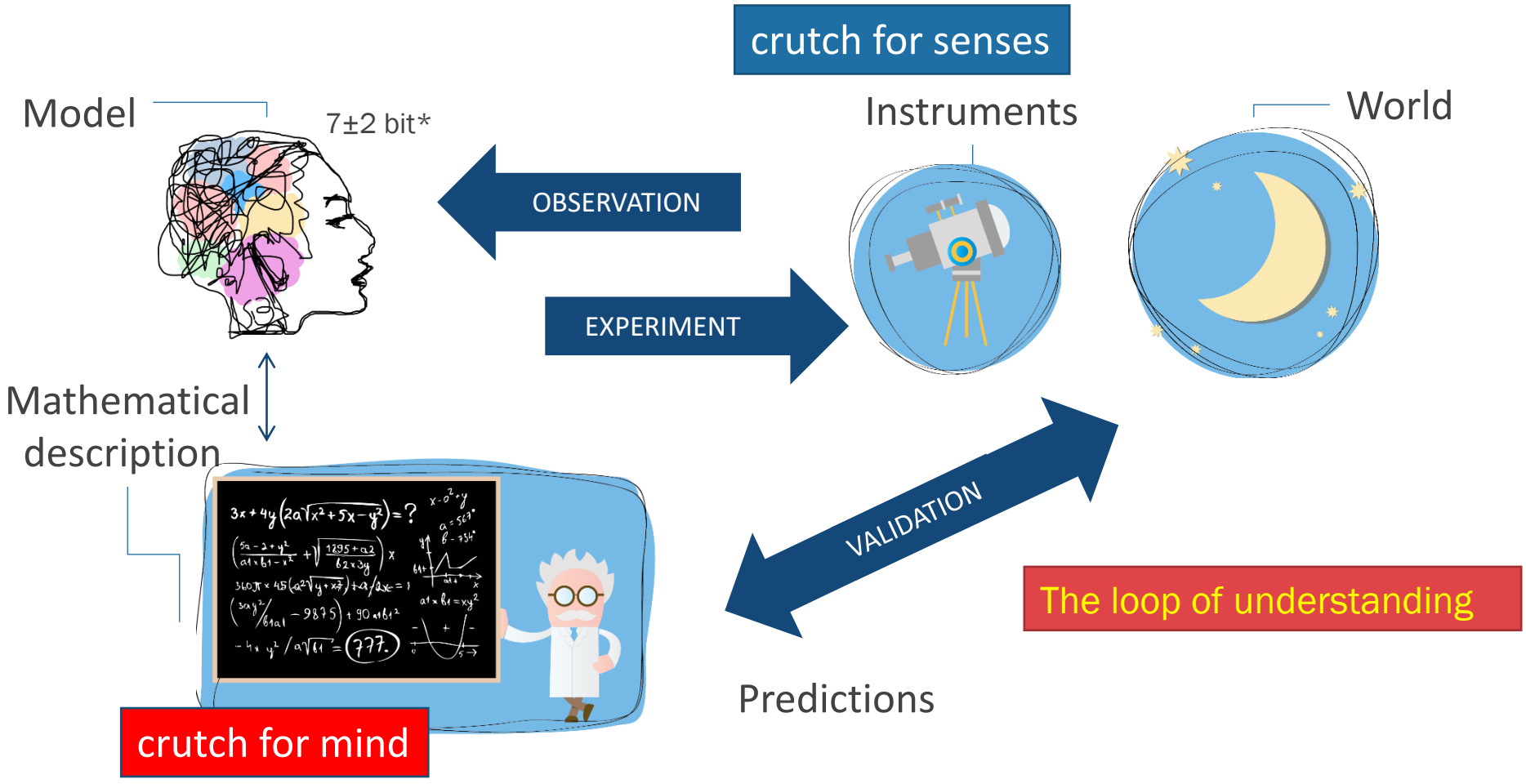
7±2 bit*

OBSERVATION

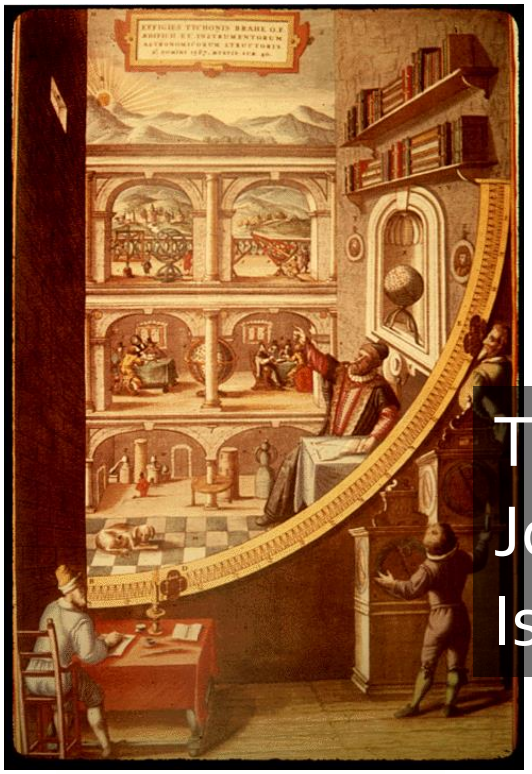
World

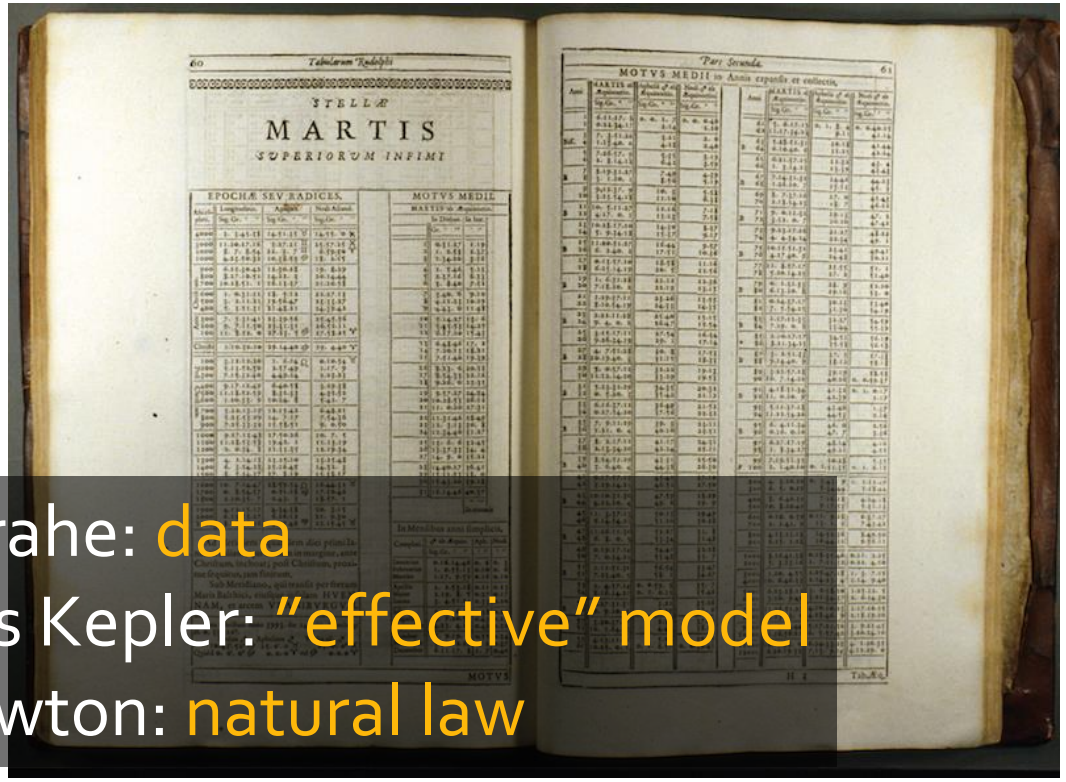
* Miller, G. A. Psychological Review. 63 (2): 81–97 (1956)

History of intelligence / (data) science



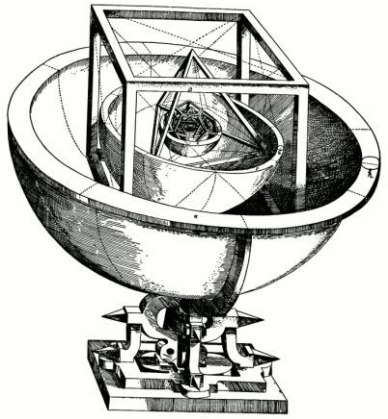
First "Data Science"



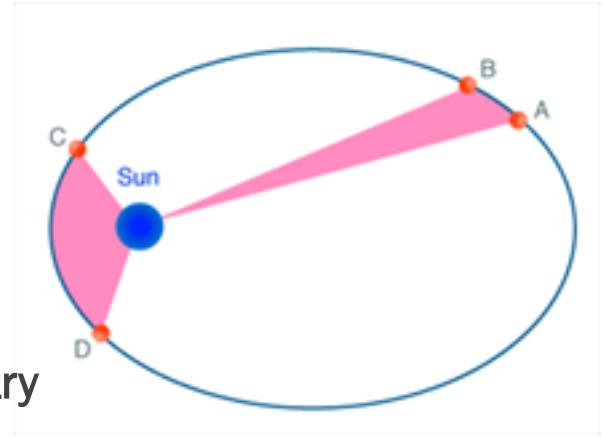


Tycho Brahe: data
Johannes Kepler: "effective" model
Isaac Newton: natural law

Tabulae Rudolphinae (1627), **23 years**, position of 1405 stars + planets + **30 years data curation**

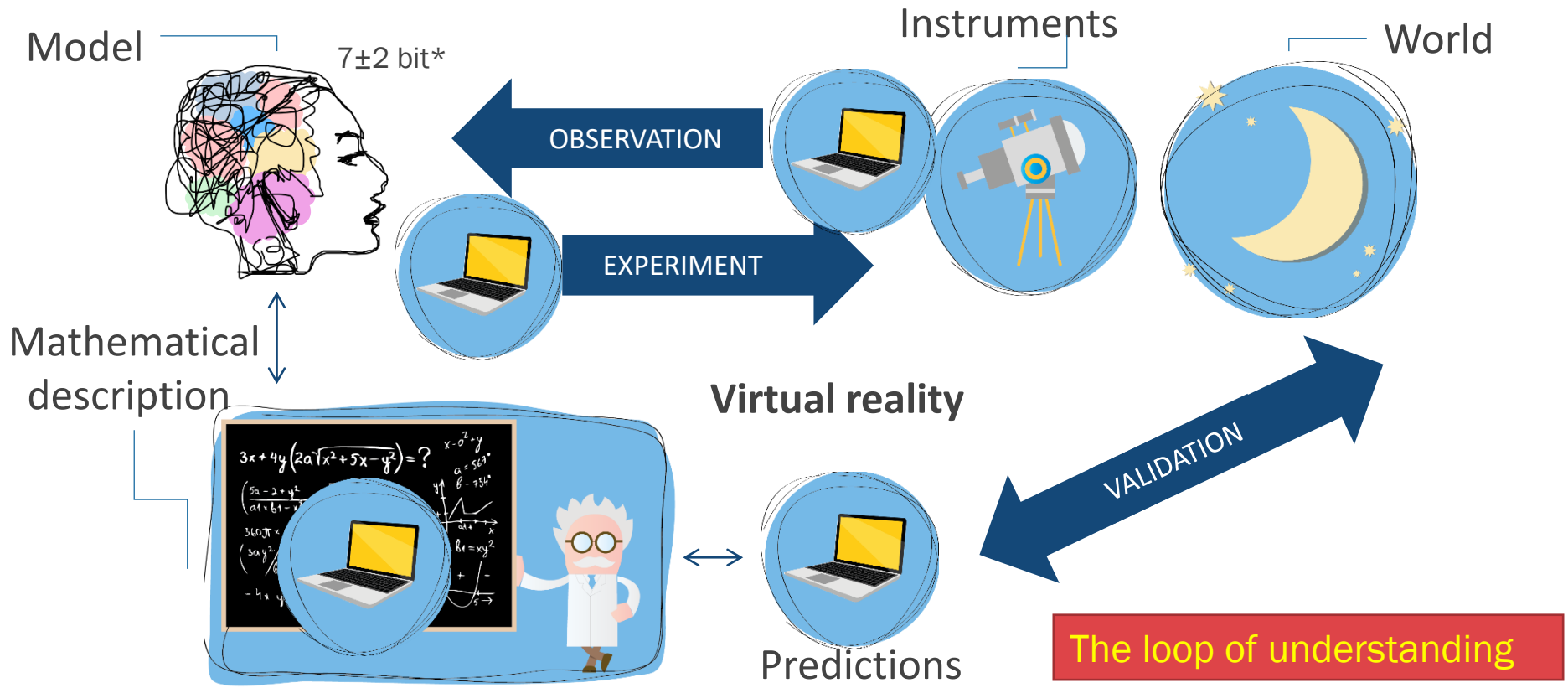


Perfect beauty and symmetry



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

History of intelligence / (data) science



Initial values

Λ=0.7
Ω_m=0.3

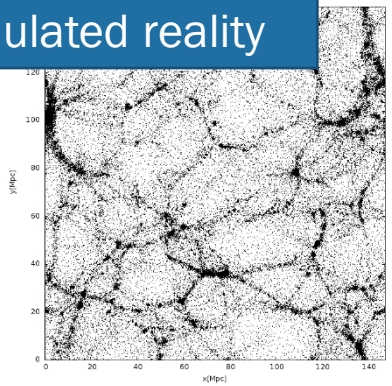
“laws”, equations

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

$$F = ma$$

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

Simulated reality



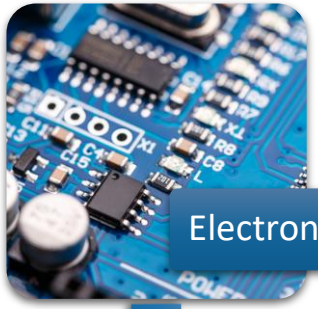
Science – technology – science – technology …

Astronomy → Mechanics → Quantummechanics → Solid state physics → Transistor/ microelectronics → Astronomy

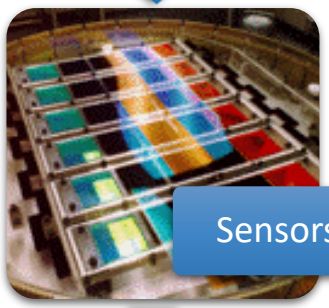
Moore's-law

Better computers

Better sensors more data

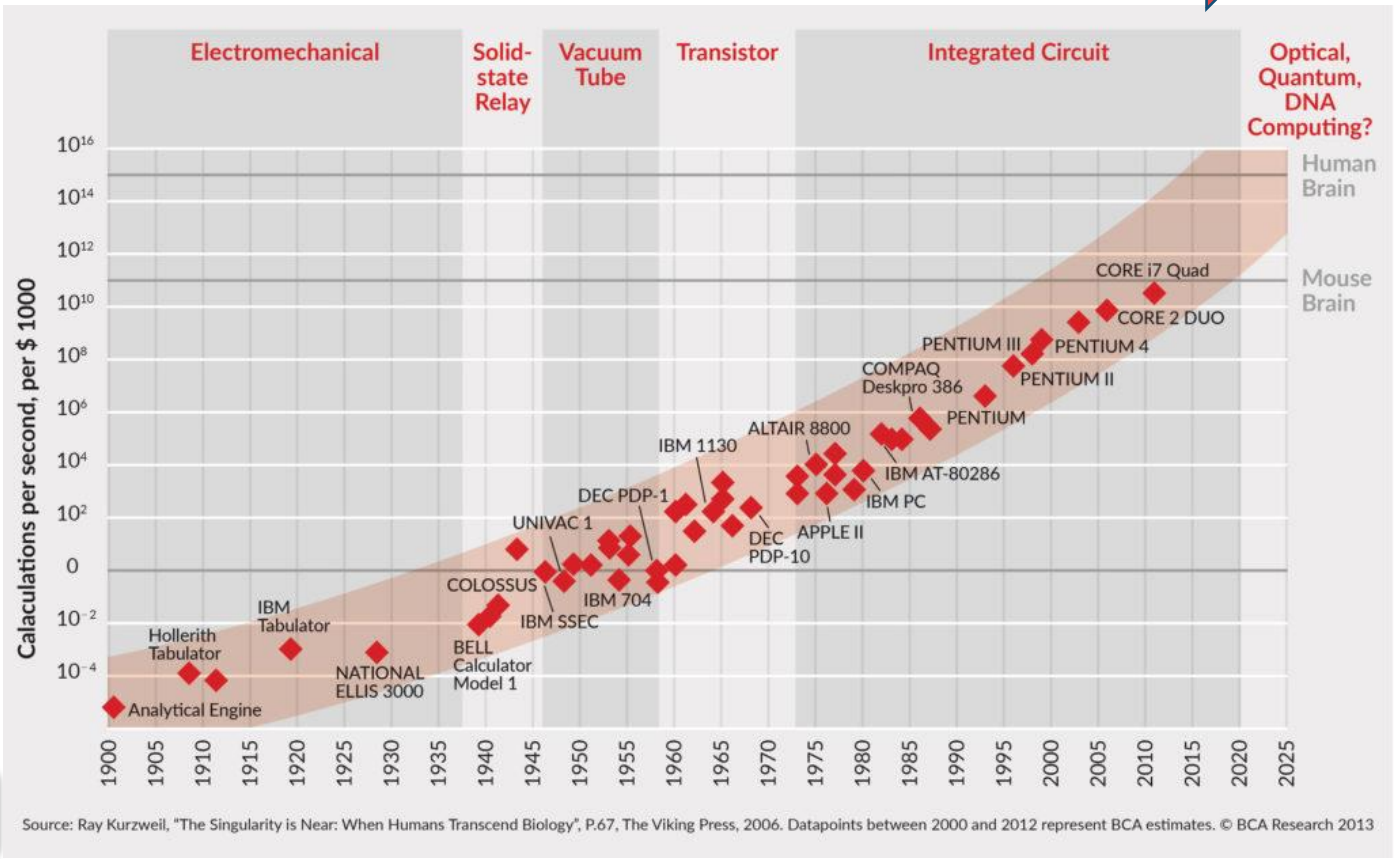


Electronics



Sensors

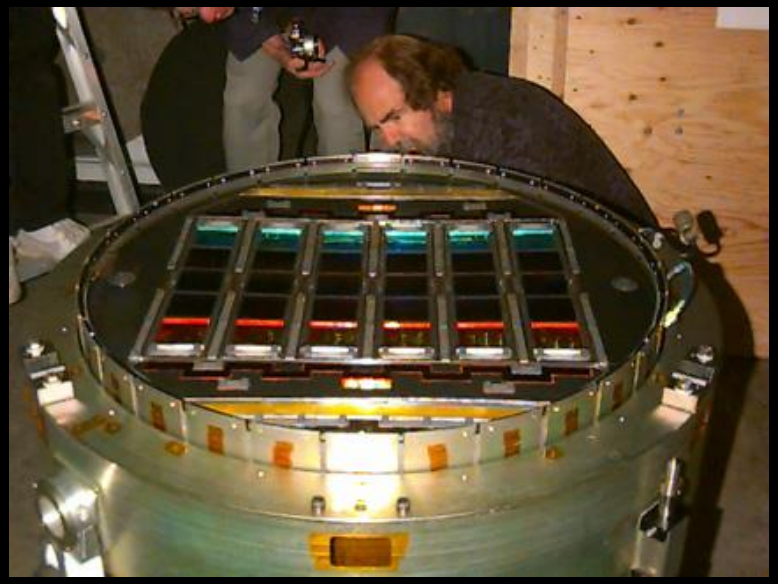
Data

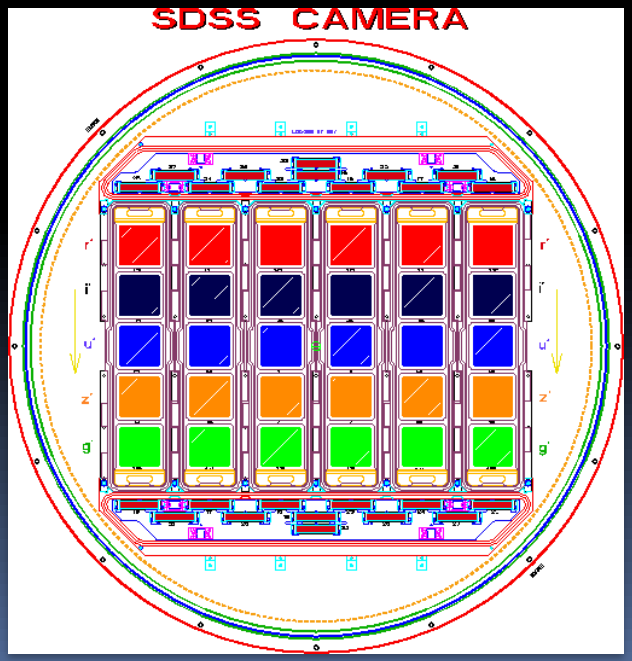


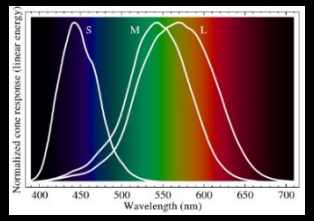
Prototype of modern "data science"

SLOAN DIGITAL SKY SURVEY: 3D MAP OF THE UNIVERSE

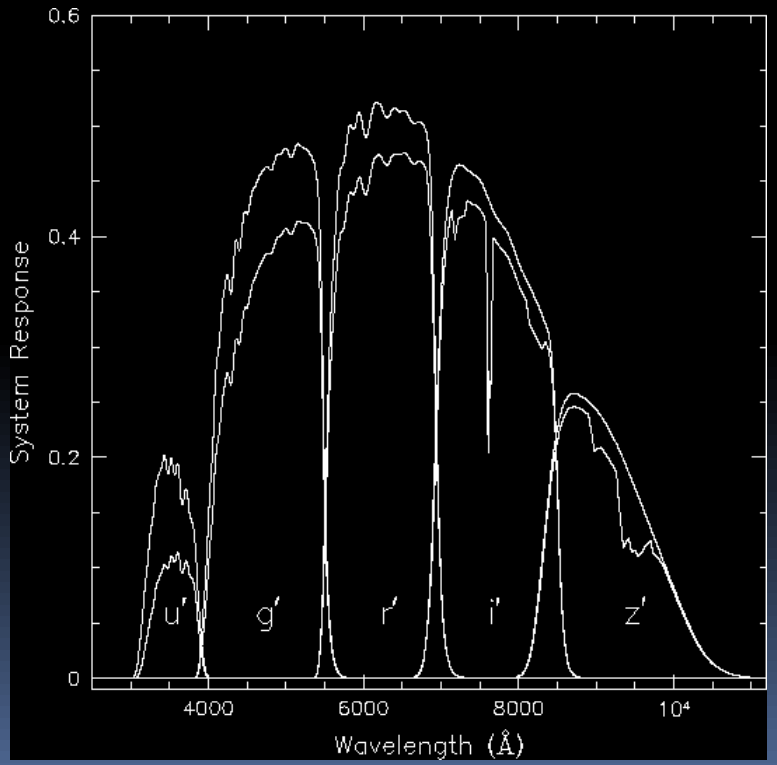
120 megapixel „color” camera (mid 90’s)







Human eye



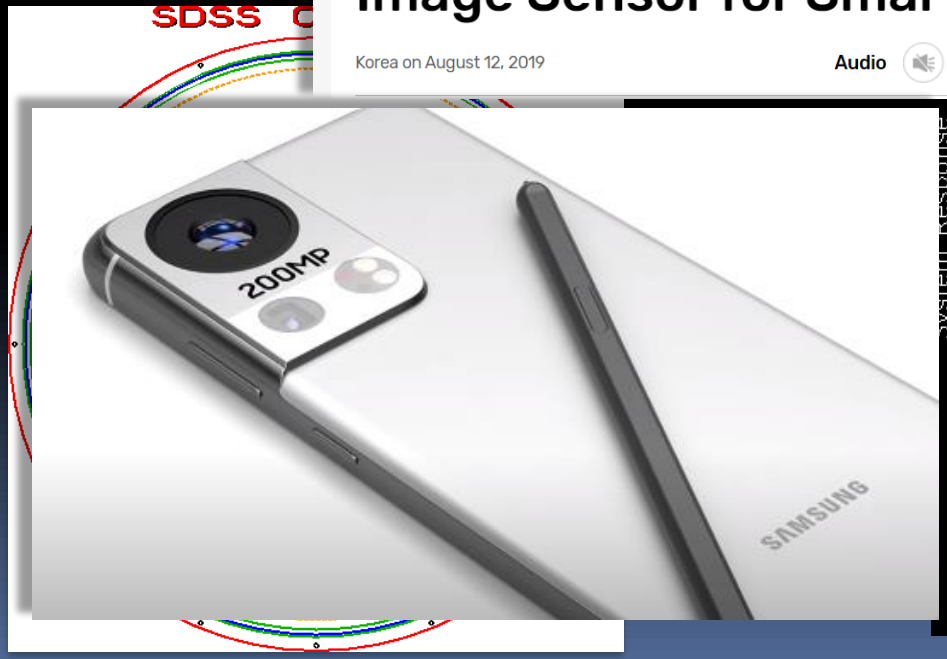
SDSS camera

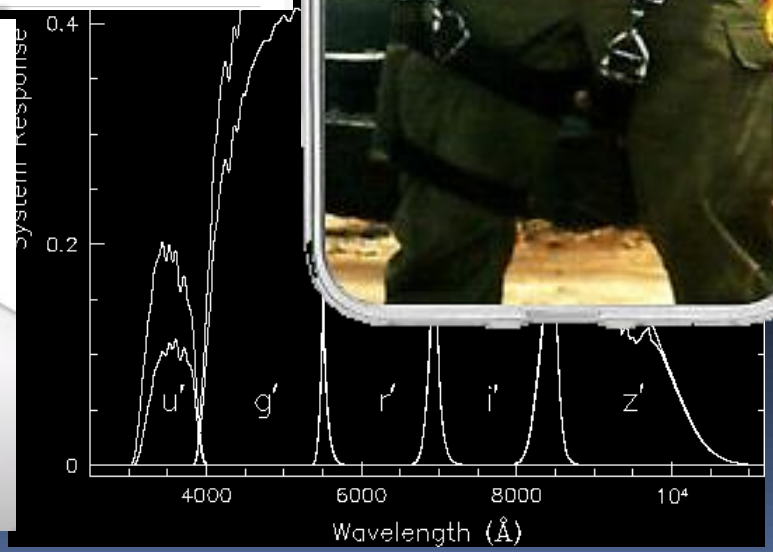
120 megapixel „color” camera (mid 90’s)

Samsung Newsroom

Samsung Takes Mobile Photography to the Next Level with Industry’s First 108Mp Image Sensor for Smartphones

Korea on August 12, 2019 Audio Share

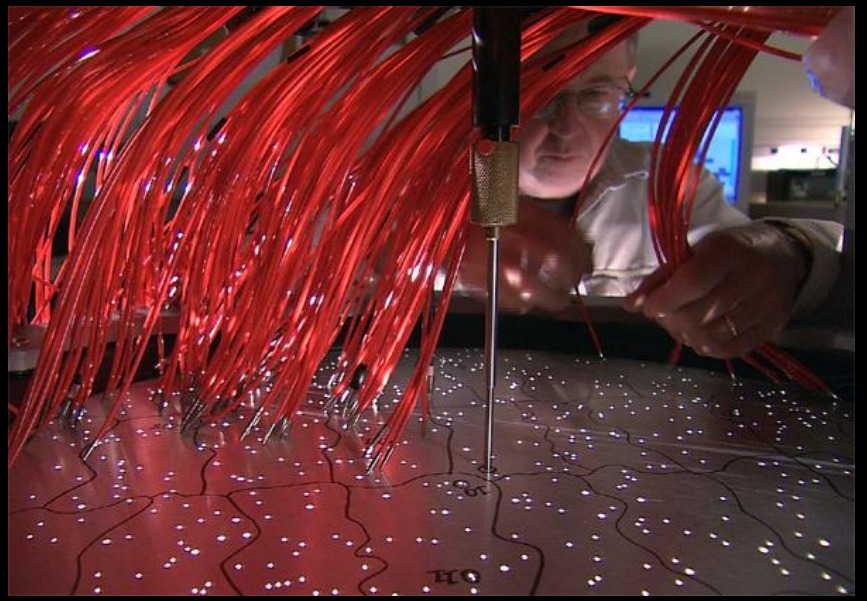


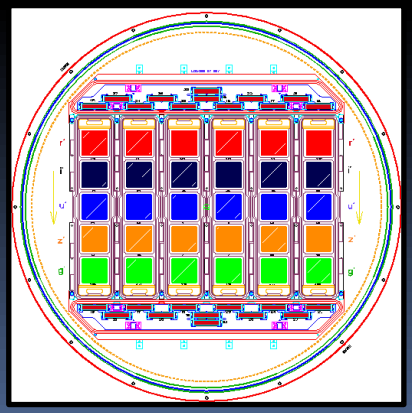


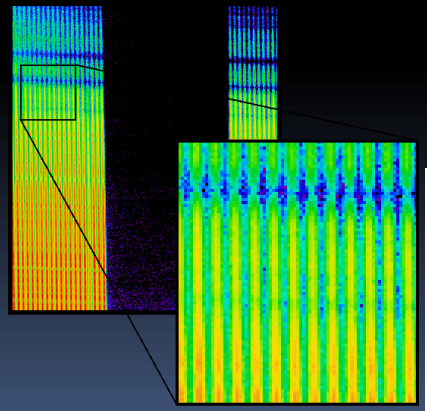
SDSS camera

2.5 terapixel image – 300 million galaxies – 5 optical bands

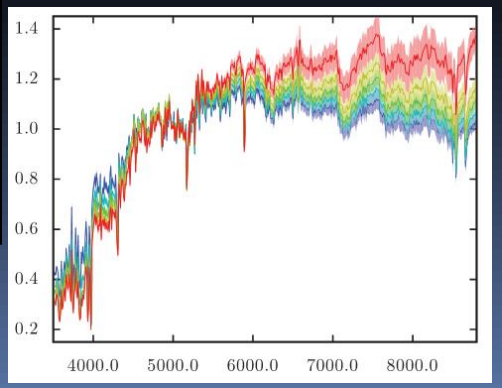
640 fibers–
1 million spectra





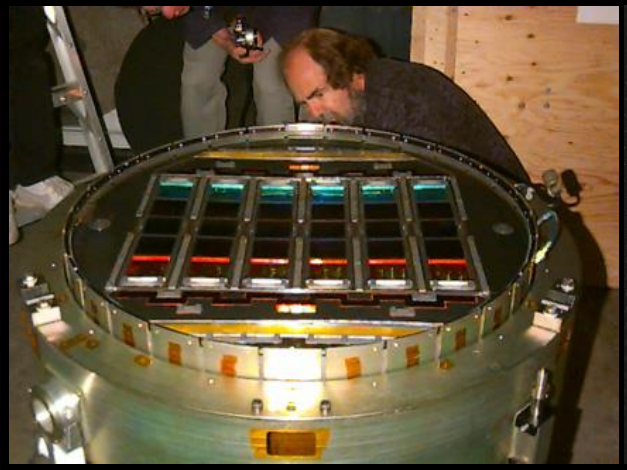


3000D vectors

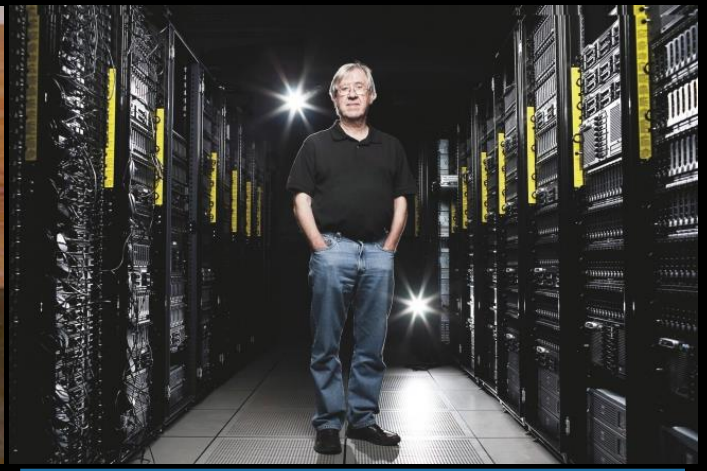


2.5m

120Mp –> 2.5Tp

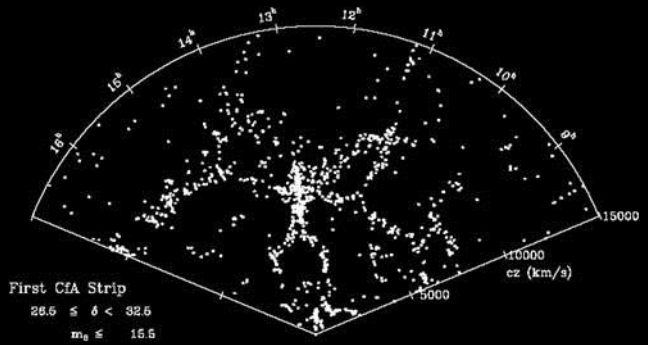


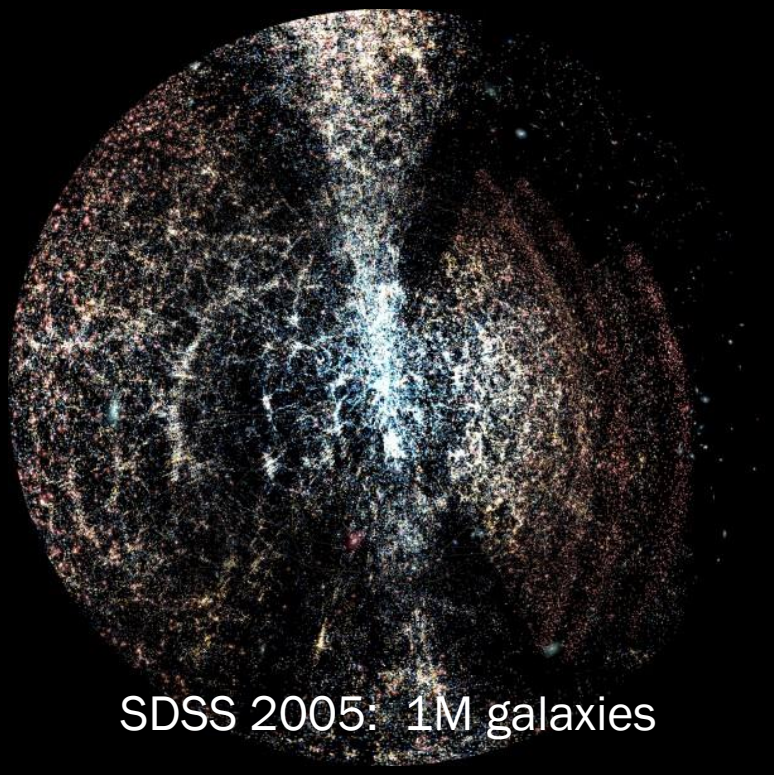
5 years:10TB



150 man-years software dev.

New issue:
BIG DATA !!!





1929: 1 galaxy

CfA 1989: 1100 galaxies

SDSS 2005: 1M galaxies

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					
40.365912	0.4821568	18.40755	16.80093	16.25361	16.07363					
44.223179	1.0513825	17.91608	16.9998	16.61383	16.46706					

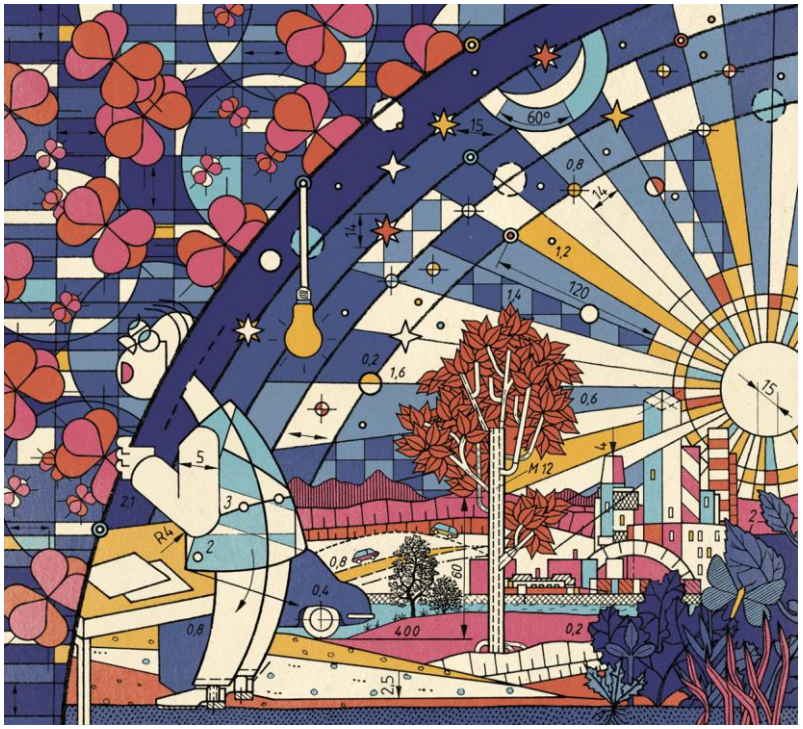
Photometry table: 300+ columns, 1Bn+ rows

Spectra: 1 million 3000 dim vectors

100+ other tables

2.5 Terapix image

Scientific observations often result data as multidimensional vector space



Scientific goals

and

researcher's perspective

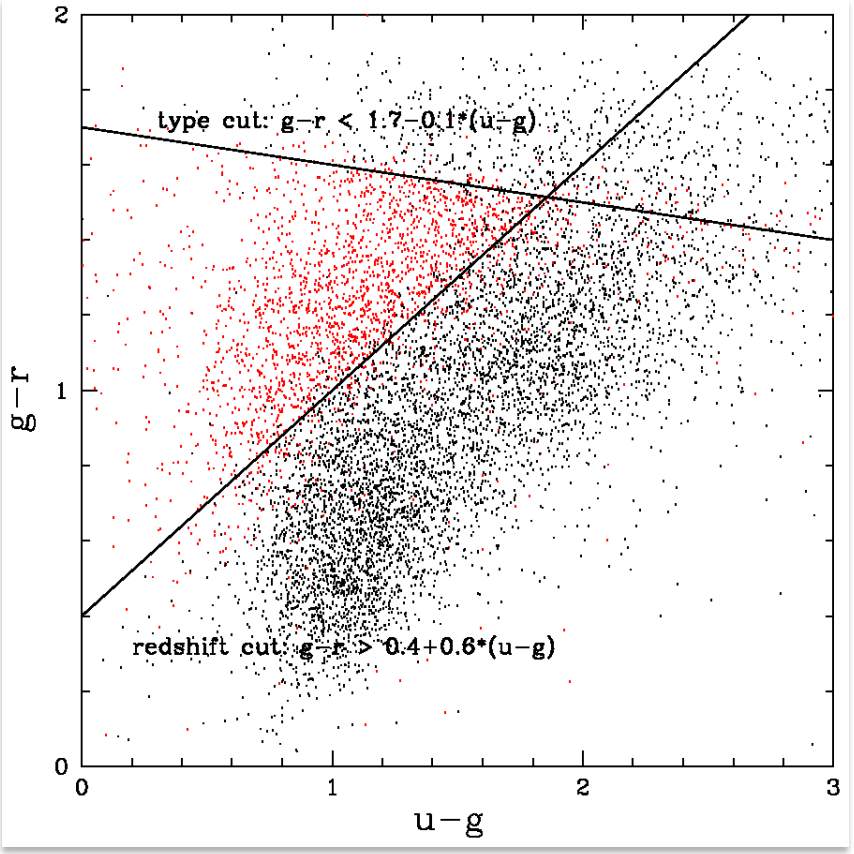
Queries in data space: e.g. separate stars and galaxies

Star/galaxy separation
Quasar target selection

"cuts"

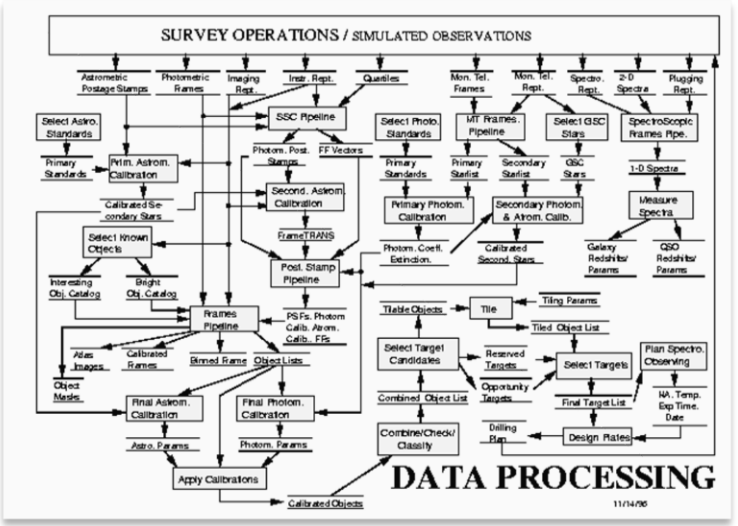
Multidimensional polyhedra

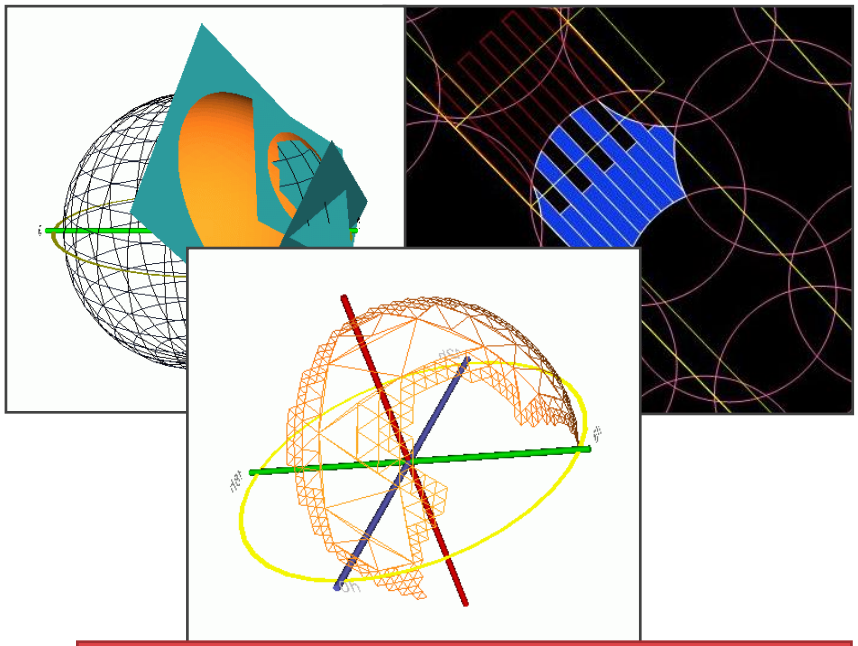
petroMag_i > 17.5 and (petroMag_r > 15.5 or petroR50_r > 2) and (petroMag_r > 0 and g > 0 and r > 0 and i > 0) and ((petroMag_r-extinction_r) < 19.2 and (petroMag_r - extinction_r < (13.1 + (7/3) * (dered_g - dered_r) + 4 * (dered_r - dered_i) - 4 * 0.18)) and ((dered_r - dered_i - (dered_g - dered_r)/4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

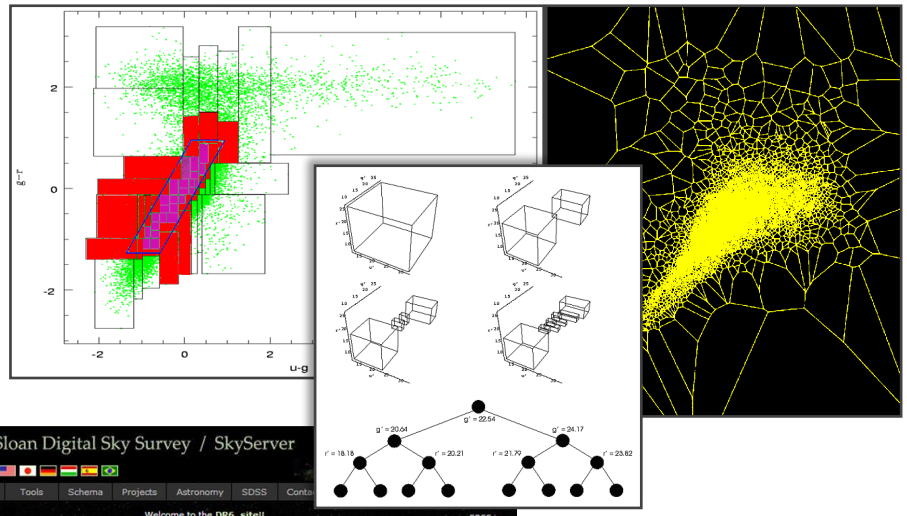


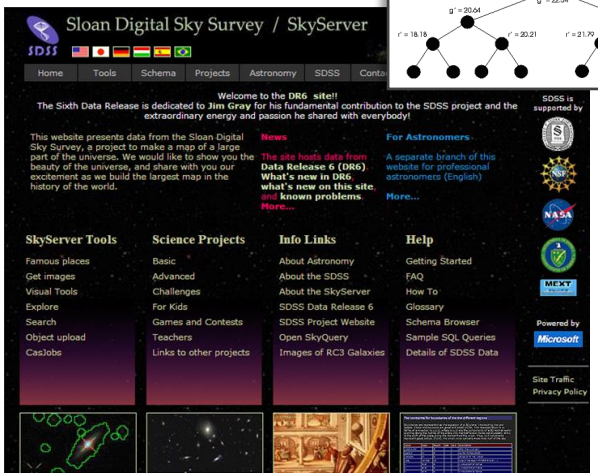
New skills: Indexing, databases

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









New surveys will collect
5 years of SDSS data in 5 days!

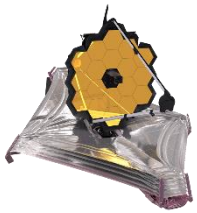
Public "cloud" data access

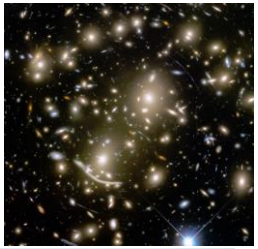
Same trends, similar challenges in all sciences

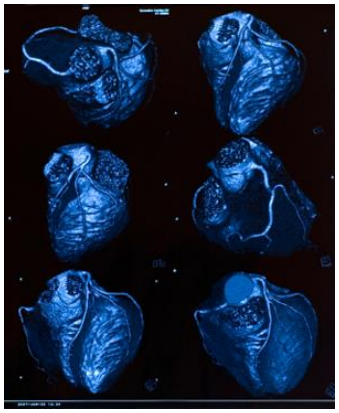
20th century

manual observations

21st century







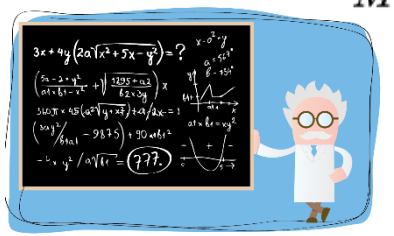
high throughput instruments

small data → big data

300 million galaxies
2.5 terapixels
3.2 gigabases,
37 trillion cells

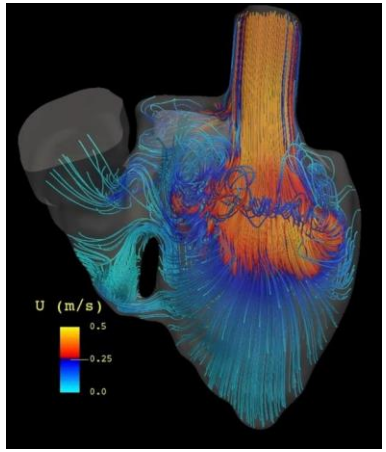
$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}.$$

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



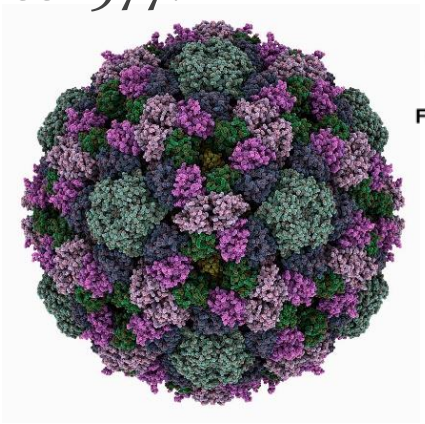
simple equations

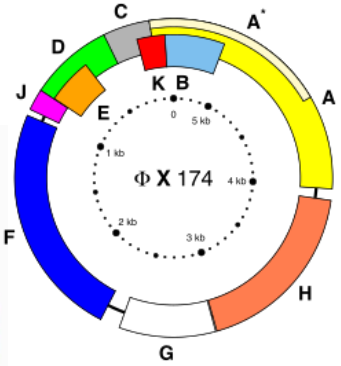
complex simulations

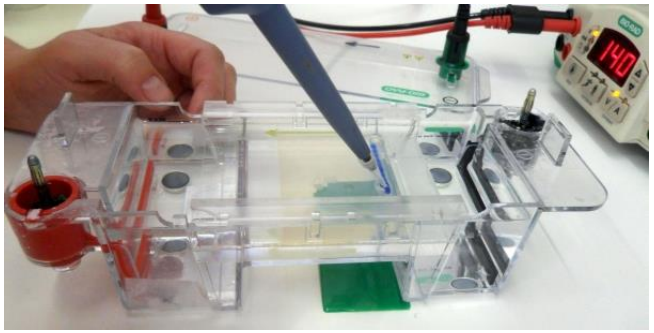


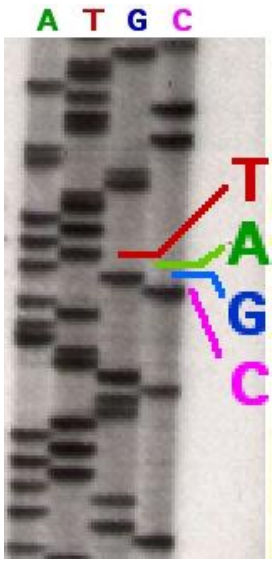
Not only astronomy: genomics

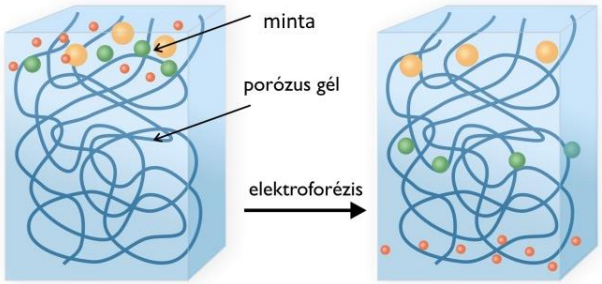
Sanger-sequencing
First virus sequence 1977:
ϕX174, 5386nt

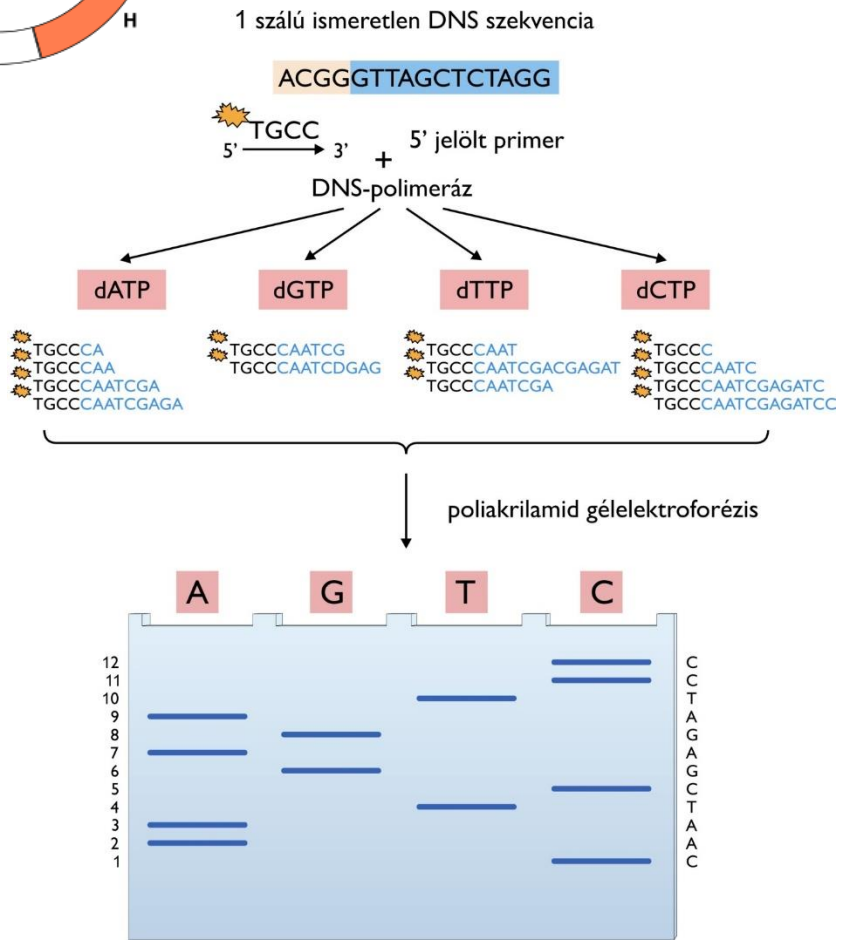






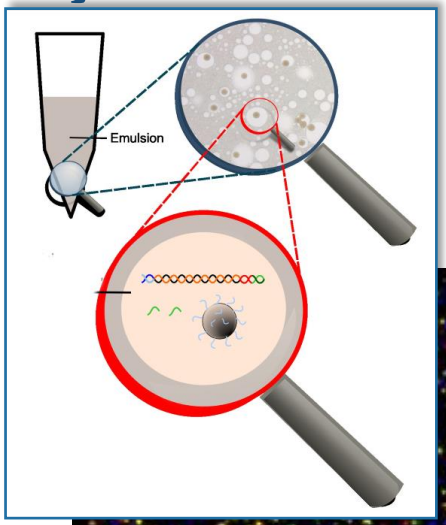


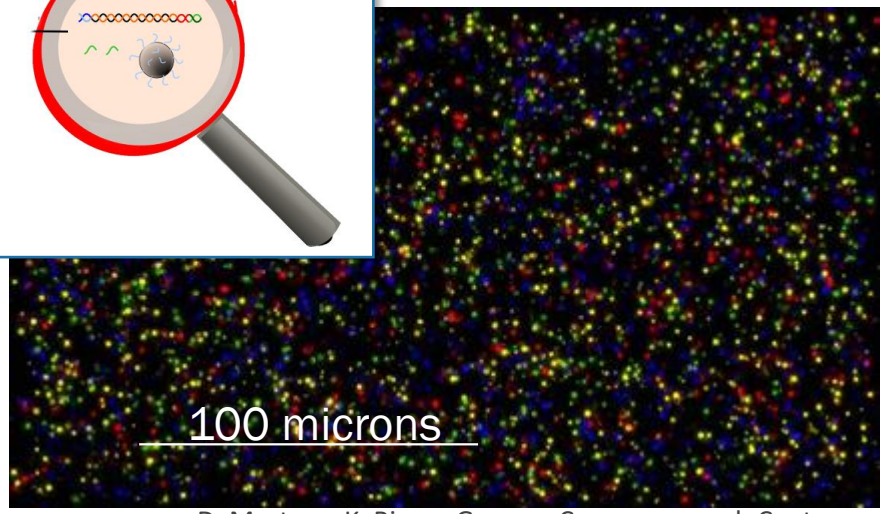




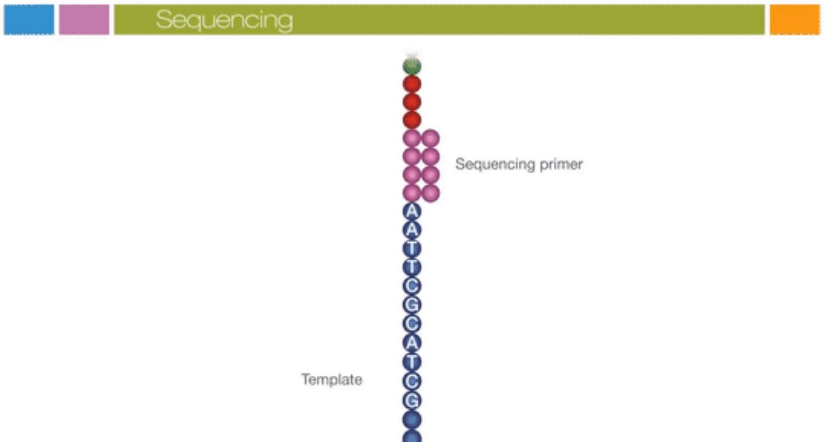
Nyitray László, Pál Gábor: A biokémia és molekuláris biológia alapjai (2013)

30 years later: NGS, nanopore





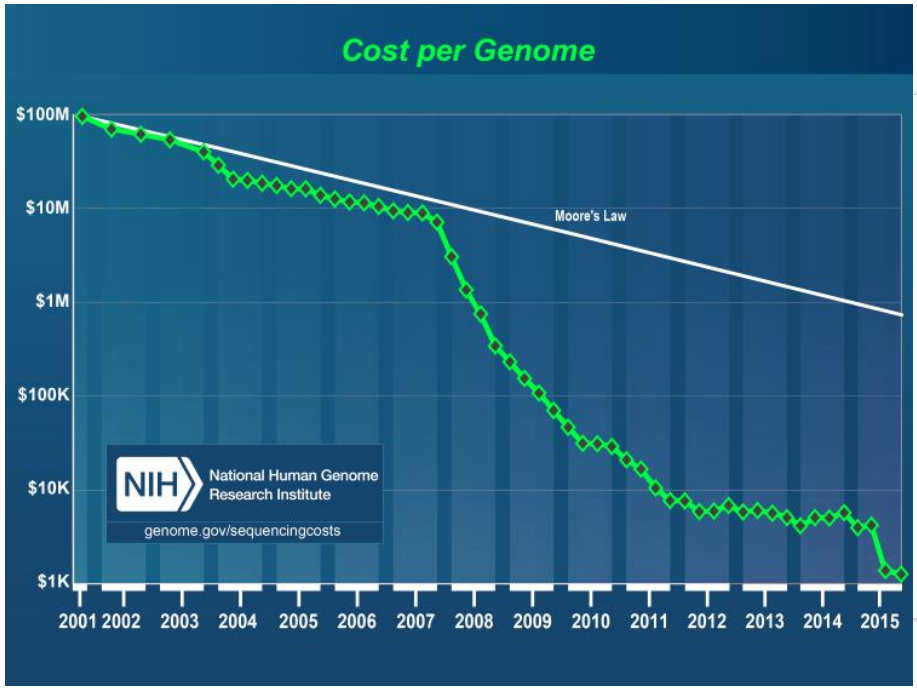
D. Mertens, K. Rippe, German Cancer research Center

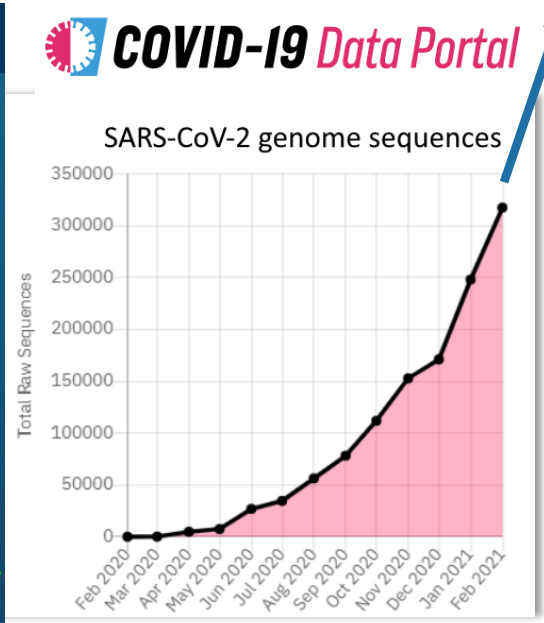


BGI Hong Kong, Scotted400, CC-BY-3.0

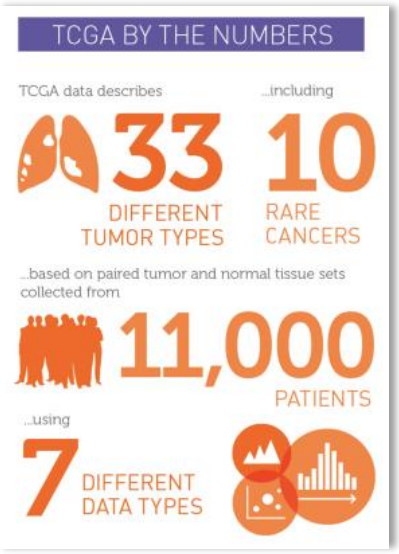
We have one, too!

Moore's law – international data sharing





April. 2022:
10+ million



Sequencing is getting cheaper.

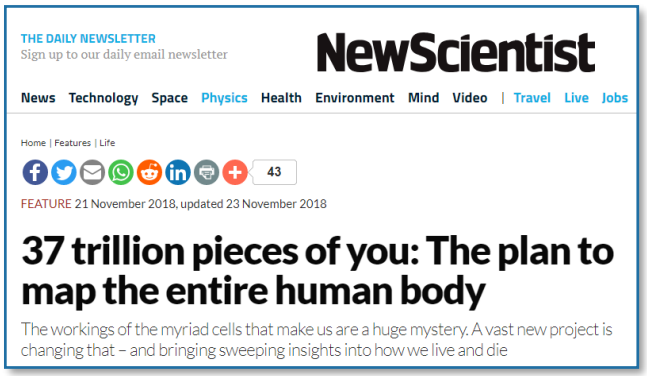
More (public) data available.

(HGP) 1990–2003
13 years / 2,7 billion USD

2020
Few days / <500 USD

2030?

+ public archives
+ whole population
+ single cell


THE DAILY NEWSLETTER
Sign up to our daily email newsletter

NewScientist

News Technology Space Physics Health Environment Mind Video | Travel Live Jobs

Home | Features | Life

43

FEATURE 21 November 2018, updated 23 November 2018

37 trillion pieces of you: The plan to map the entire human body

The workings of the myriad cells that make us are a huge mystery. A vast new project is changing that – and bringing sweeping insights into how we live and die

OTHER DATA:
-Microarray
- Mass spectrography
- Digital microscopy, ...

Biology in the 20th 21st century

Pandemics: not just a virus!

- Infectious diseases are results of **complex interactions** of several domains
- Without **global monitoring** of the drivers we cannot handle or prevent outbreaks
- Need: **collection, integration, organization, sharing and analyzing complex large data sets**
- **Barriers**:
  practical + legal and ethical issues

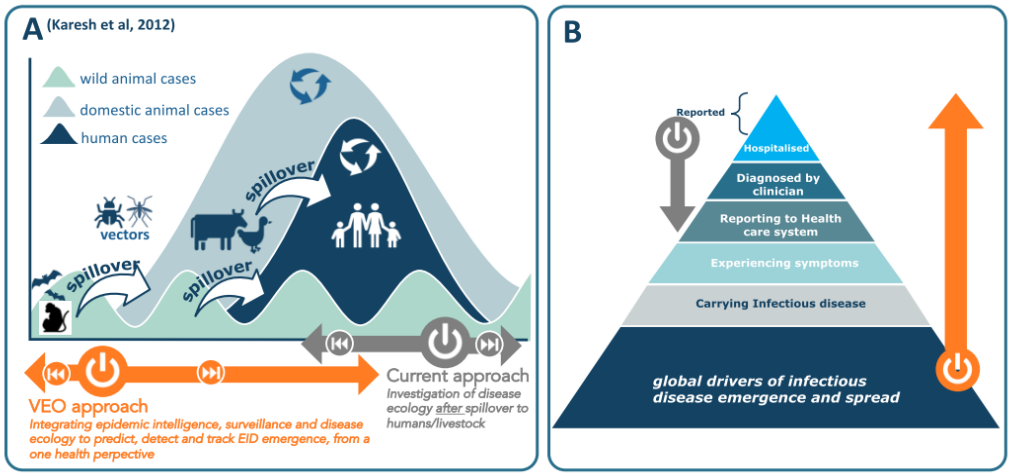
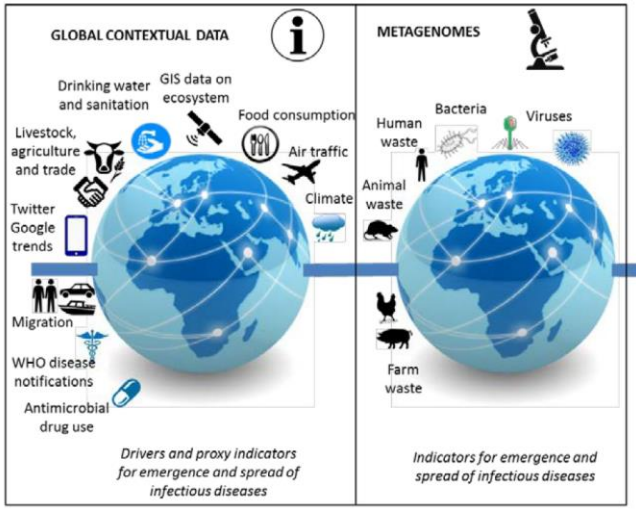
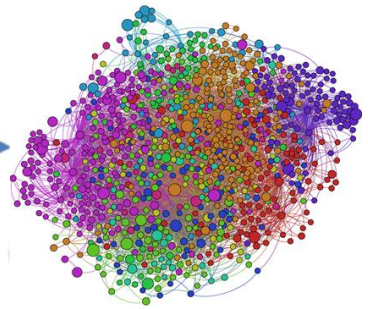
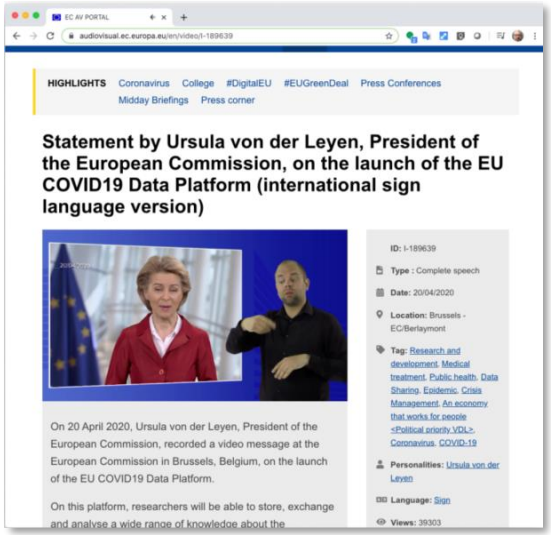
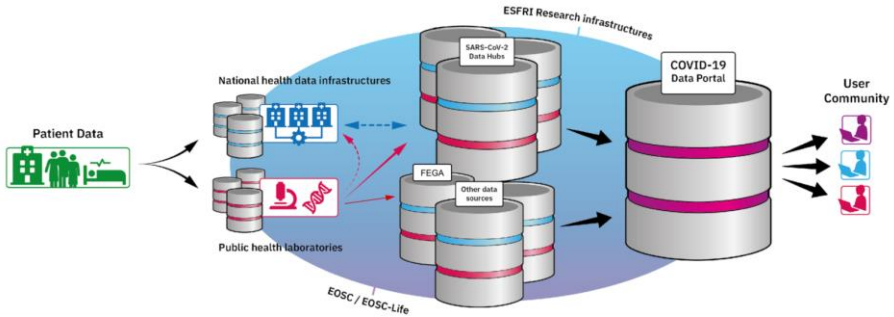


Figure. 1: A. VEO's pro-active, forward looking approach versus the current, reactive approach in EID preparedness and response research (A) and in terms of focusing on drivers of disease emergence and spread instead of taking actions once disease mergence is reported to the healthcare system (B).



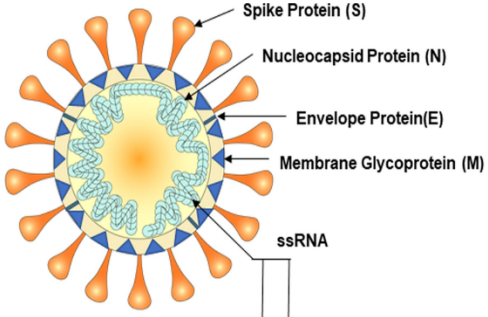


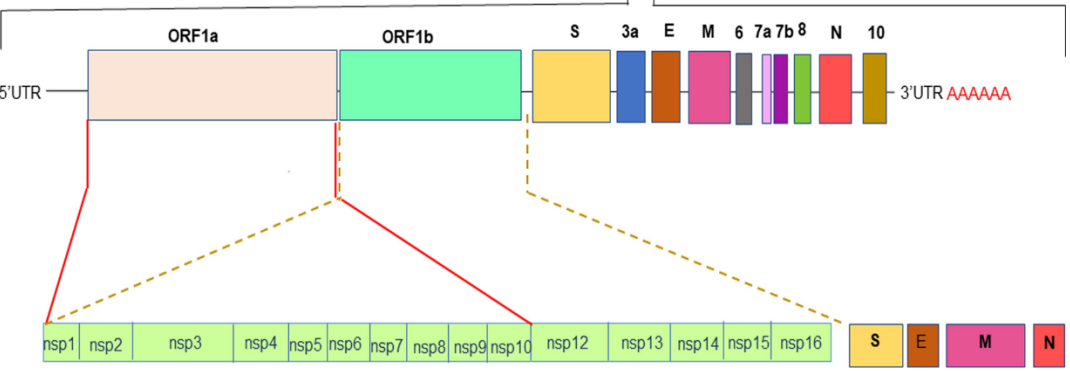


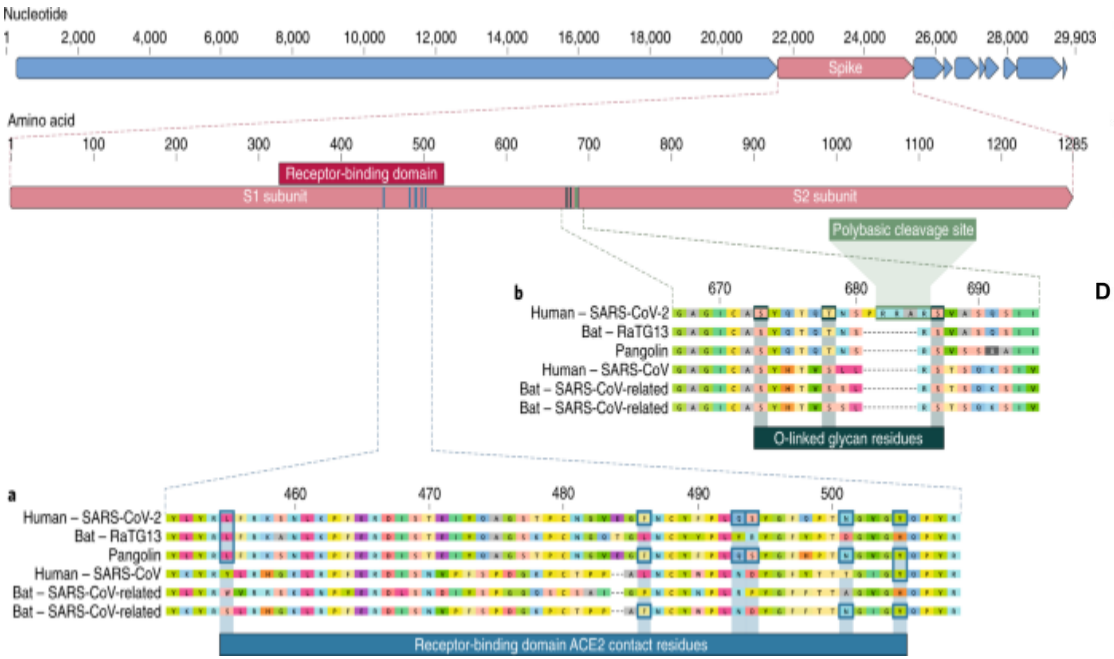


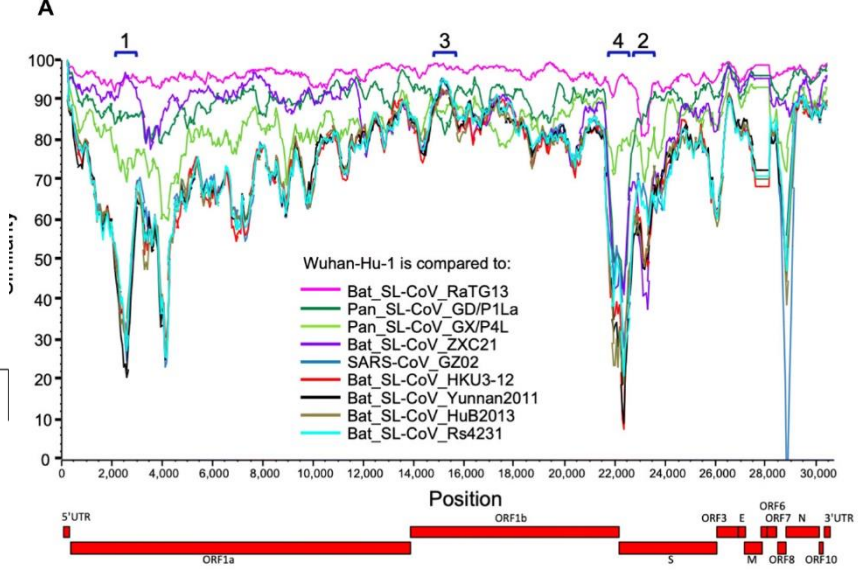
Versatile Emerging infectious disease Observatory
EU H2020 2020.01.01-2025.12.31

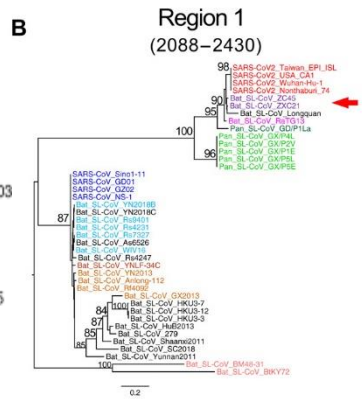
SARS-CoV-2 genome

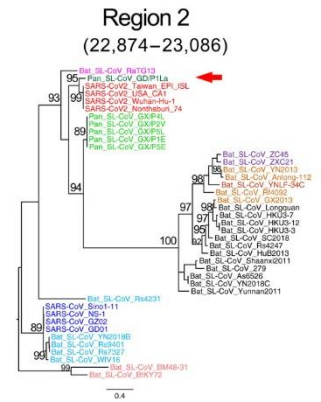


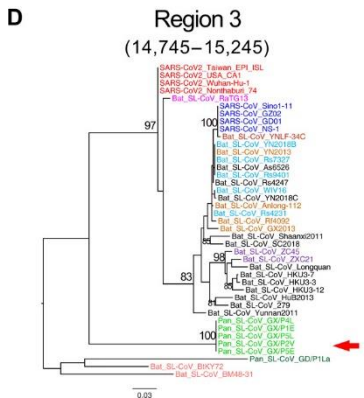


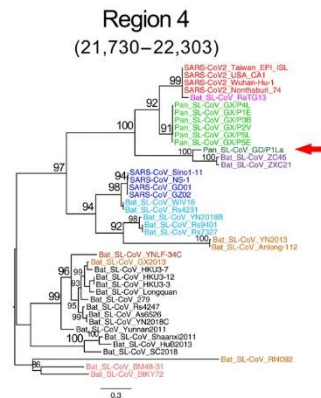






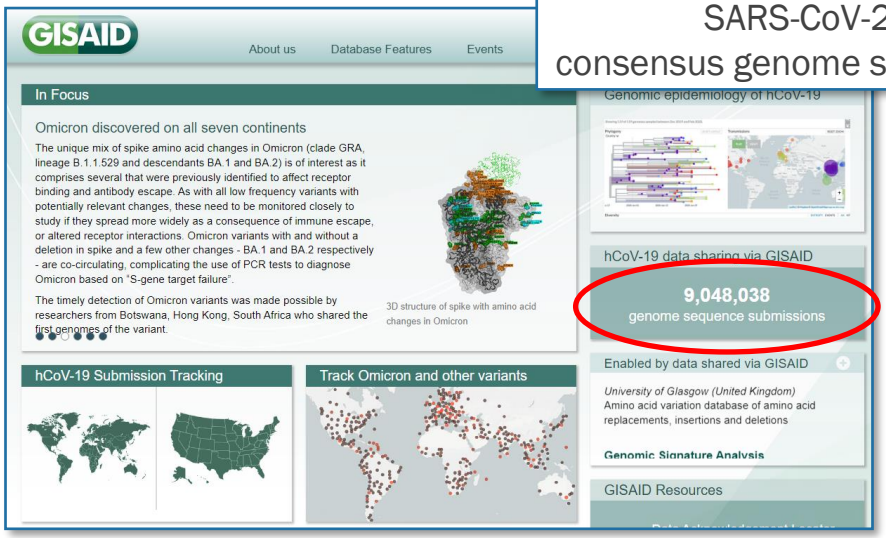




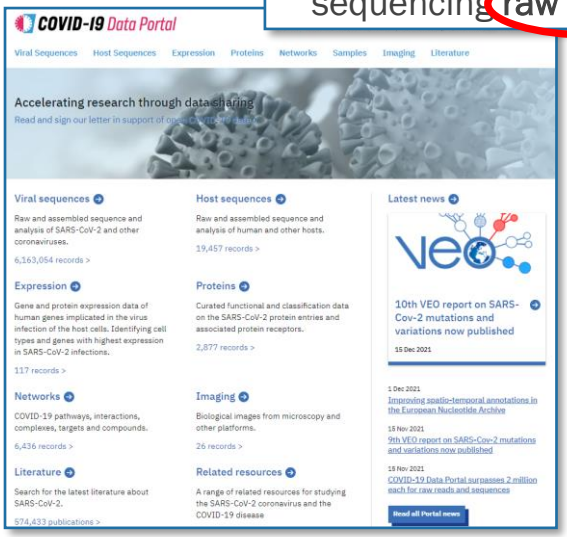


Sequence Archives

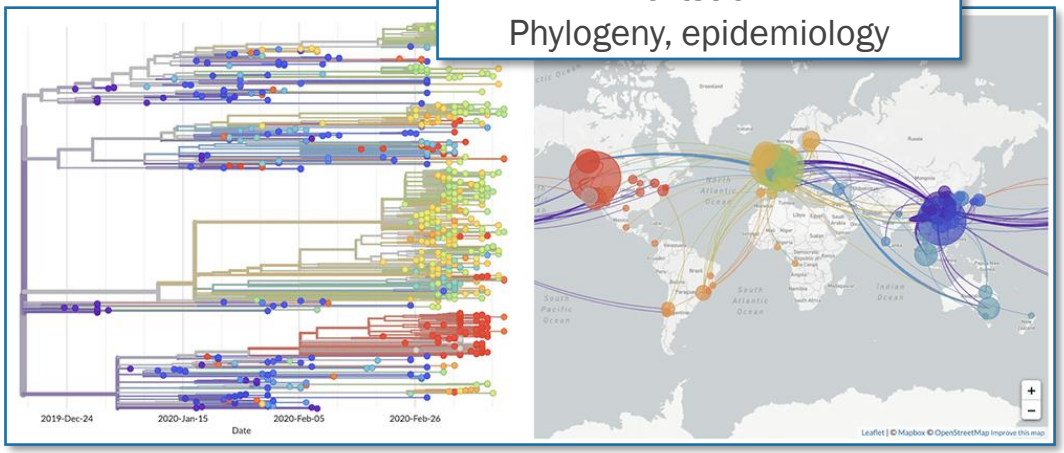
SARS-CoV-2
consensus genome sequences



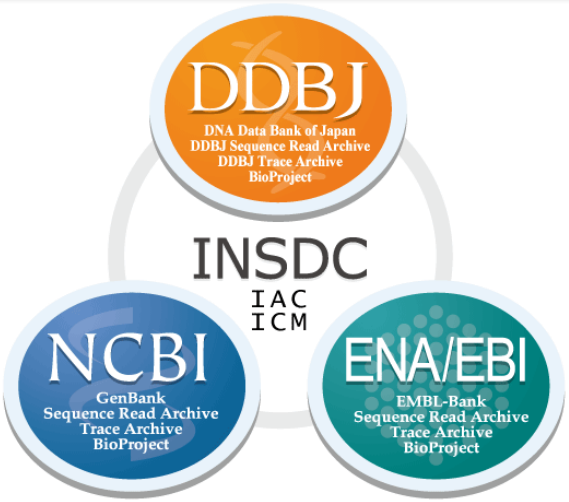
SARS-CoV-2
sequencing **raw data**



Nextstrain
Phylogeny, epidemiology



Human, animal, plant, metagenomic,
... sequencing raw data



"**25.6 Petabase** pairs originating from over **14.8 million** publicly available runs averaging 1.7 Gbp per run, 0.83 GB per run, 9.6 million spots per run and 187 bp per spot."
/INSDC Public data as of Sept. 2021./

Importance of keeping raw and metadata

"One man's trash is another man's treasure."

- **Sequencing archives:** most of genetic sequencing research data is stored at public archives
- They are **indexed** by metadata and there are sequence-based search tools
- Possible "Google" query:

```
SELECT sampleID WHERE collection_date < '2019-12-25'
AND matched_taxon='SARS-CoV-2'
```

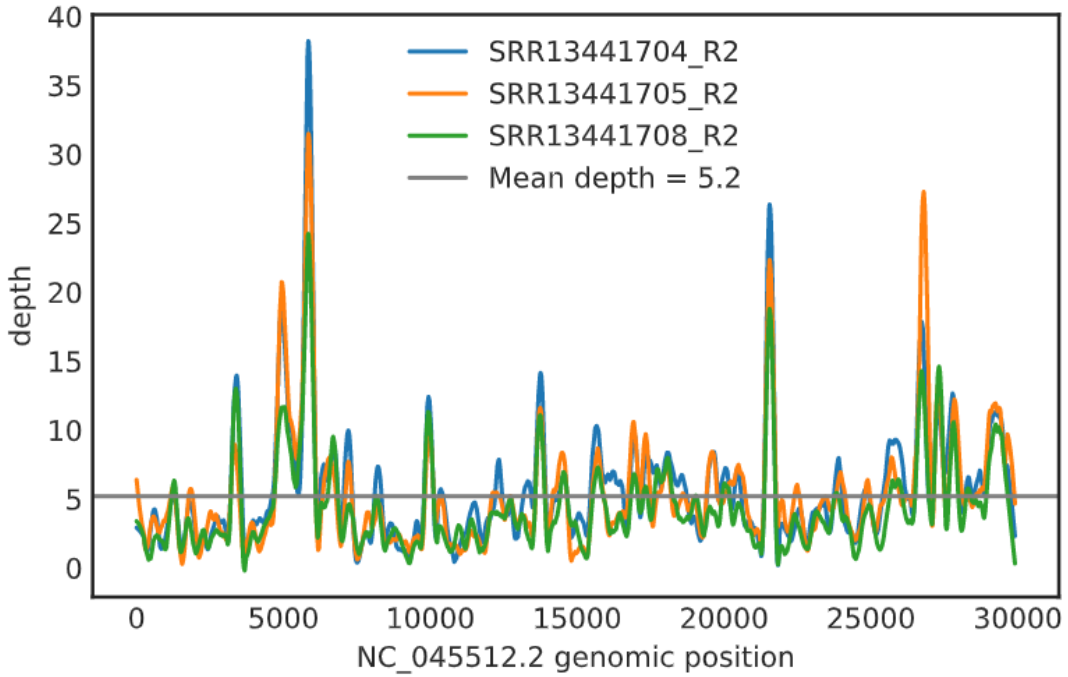
Run	Library Name	Collection date	Isolation source	lat lon	R1	R2
SRR13441700	AKGI_BS1_2018_12_24	2018-12-24	Antarctic soil	62.13 S 58.95 W	112	65
SRR13441701	AKGI_PL3_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.93 W	0	0
SRR13441702	AKGI_PL2_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.93 W	0	0
SRR13441703	AKGI_PL1_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.93 W	0	0
SRR13441704	AKGI_PS3_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.92 W	387	4485
SRR13441705	AKGI_PS2_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.92 W	242	3800
SRR13441706	AKGI_PS1_2019_01_13	2019-01-13	Antarctic soil	62.21 S 58.92 W	0	0
SRR13441707	AKGI_SS3_2019_01_05	2019-01-05	Antarctic soil	62.21 S 59.01 W	0	0
SRR13441708	AKGI_BS3_2018_12_24	2018-12-24	Antarctic soil	62.13 S 58.95 W	349	3537
SRR13441709	AKGI_BS2_2018_12_24	2018-12-24	Antarctic soil	62.13 S 58.95 W	112	161
SRR13441710	AKGI_SS2_2019_01_05	2019-01-05	Antarctic soil	62.21 S 59.01 W	113	106
SRR13441711	AKGI_SS1_2019_01_05	2019-01-05	Antarctic soil	62.21 S 59.01 W	0	0

# of SARS-Cov-2 sequence reads

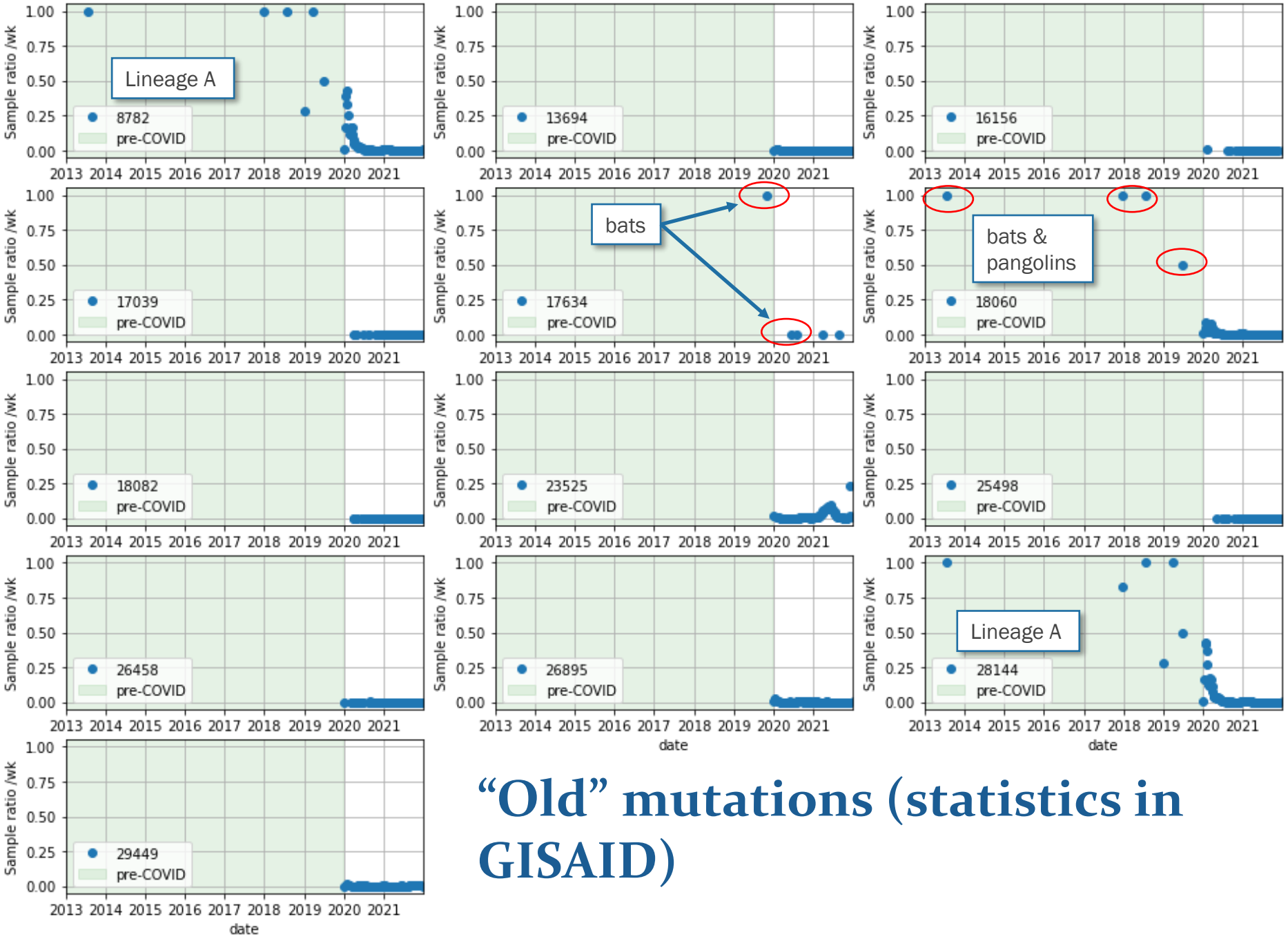
Csabai, I., Papp, K., Visontai, D., Stéger, J. and Solymosi, N., Unique SARS-CoV-2 variant found in public sequence data of Antarctic soil samples collected in 2018-2019. Submitted 2022. https://www.researchsquare.com/article/rs-1177047/v1 , https://www.researchsquare.com/article/rs-1330800/v1

SARS-CoV-2 genome

- Low, but almost **complete coverage**
- Similar for all 3 pivot samples
- Early type **lineage "A"** mutations C8782T, T28144C + C18060T
  - Found in bats but not in the "Wuhan Market" samples (Lineage B)
- **Very rare mutation** C17634G
  - 2 / 6,000,000
- **Rare 27nt deletion** @21761
  - GISAID: 38/6,000,000
  - No lineage A before 2021.04
- Close-by mutations in 150nt reads: mixed (at least 3) genotypes
  - Quasispecies, multiple samples?



POS	REF	ALT	Annotation	Gene	AA change	AF	DP	RUN
8782	C	T	synonymous_variant	ORF1ab	Ser2839Ser	[0.66]	[9]	[08]
13694	C	T	missense_variant	ORF1ab	Thr4482Ile	[0.38, 0.29, 0.22]	[47, 34, 35]	[04, 05, 08]
16156	A	G	missense_variant	ORF1ab	Met5303Val	[0.36, 0.54, 0.5]	[25, 11, 10]	[04, 05, 08]
17039	A	G	missense_variant	ORF1ab	Asn5597Ser	[0.53]	[13]	[05]
17634	C	G	missense_variant	ORF1ab	Asp5795Glu	[0.25]	[20]	[08]
18060	C	T	synonymous_variant	ORF1ab	Leu5937Leu	[0.37, 0.38, 0.46]	[27, 26, 26]	[04, 05, 08]
18082	A	G	missense_variant	ORF1ab	Ile5945Val	[0.46]	[28]	[04]
21761	G	del27	disrupt.inframe.del	S	Ile68_Thr76del	[0.37, 0.33, 1.0]	[8, 9, 5]	[04, 05, 08]
23525	C	T	missense_variant	S	His655Tyr	[0.69, 0.61, 0.66]	[13, 13, 6]	[04, 05, 08]
25498	C	T	missense_variant	ORF3a	Pro36Ser	[0.45]	[11]	[05]
26458	G	T	missense_variant	E	Asp72Tyr	[0.5]	[14]	[05]
26895	C	T	missense_variant	M	His125Tyr	[0.51]	[60]	[05]
28144	T	C	missense_variant	ORF8	Leu84Ser	[0.64, 0.71, 0.66]	[17, 14, 15]	[04, 05, 08]
29449	G	T	synonymous_variant	N	Val392Val	[0.43, 0.24]	[32, 29]	[04, 08]



"Old" mutations (statistics in GISAID)

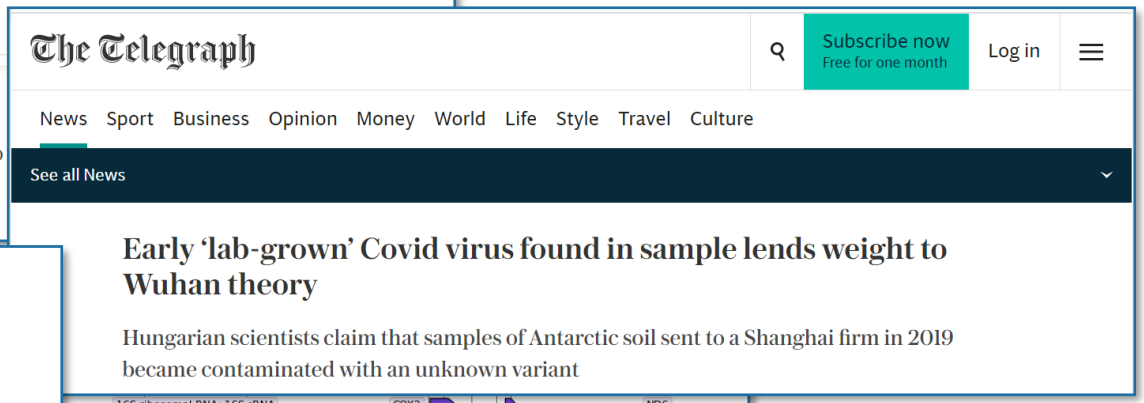
Host genomes

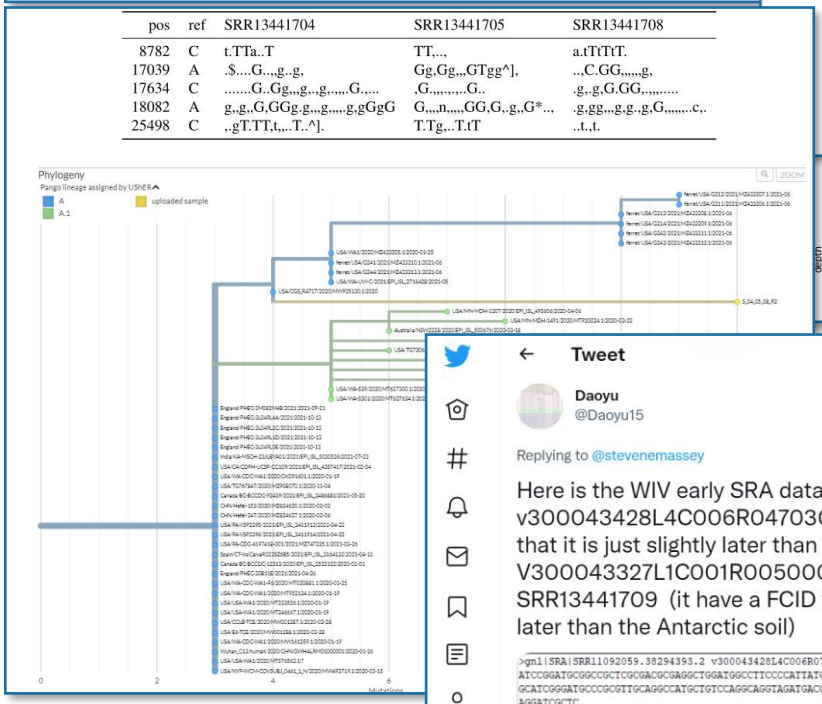
- Align to all available vertebrate MT genomes
- No penguins or seals ☹
- But:
  - Human
  - Chinese hamster
  - Green monkey
  - -> known cell lines: Vero E6, CHO, BHK21
- Most abundant
  - In “pivot” samples 04,05,08
  - Same pattern as with SARS-CoV-2, R2 asym.
- Origin: contamination
  - HiSeq 4000 barcode misassignment

Run	Genome ID	R1 coverage	R1 depth	R2 coverage	R2 depth	Species
SRR13441701	NC_018784.1	0.0	0.00	4.28	0.31	Mobula japanica
SRR13441701	NC_020001.1	3.78	0.26	3.85	0.24	Sooglossus thomasseti
SRR13441701	NC_029404.1	1.09	0.03	2.92	5.10	Geotria australis
SRR13441701	NC_057519.1	1.96	0.28	2.36	0.39	Didelphis pernigra
SRR13441701	NC_020599.1	1.02	0.01	2.24	0.58	Tachycineta cyaneoviridis
SRR13441701	NC_012769.1	2.25	0.07	2.2	0.08	Eulemur fulvus mayottensis
SRR13441701	NC_020605.1	0.0	0.00	2.17	0.10	Progne chalybea
SRR13441701	NC_011944.1	1.59	0.41	1.91	44.77	Thylacinus cynocephalus
SRR13441701	NC_027840.1	0.96	0.08	1.9	0.12	Amazona ochrocephala
SRR13441701	NC_012766.1	2.31	0.07	1.86	0.04	Eulemur fulvus fulvus
SRR13441702	NC_029404.1	0.91	0.01	2.68	0.31	Geotria australis
SRR13441702	NC_020001.1	3.66	0.09	2.65	0.09	Sooglossus thomasseti
SRR13441702	NC_030333.1	1.0	0.05	2.0	0.08	Telmatobius chusmisensis
SRR13441702	NC_018784.1	0.83	0.01	1.95	0.03	Mobula japanica
SRR13441702	NC_021952.1	1.51	0.03	1.69	0.03	Leontopithecus rosalia
SRR13441702	NC_035810.1	1.81	0.79	1.6	0.66	Lutra sumatrana isolate
SRR13441702	NC_057520.1	1.27	0.03	1.59	0.04	Lutreolina crassicaudata
SRR13441702	NC_057519.1	2.06	0.05	1.57	0.07	Didelphis pernigra
SRR13441702	NC_050286.1	0.81	0.01	1.56	0.12	Hirundo smithii
SRR13441702	NC_041147.1	0.62	0.02	1.53	0.03	Isopachys gyldenstolpei
SRR13441704	NC_012920.1	**61.55**	2.05	**89.17**	18.45	Homo sapiens ★
SRR13441704	NC_007936.1	20.89	0.26	**78.01**	4.78	Cricetulus griseus ★
SRR13441704	NC_011137.1	17.27	0.30	**62.31**	3.48	Homo sapiens neanderthalensis
SRR13441704	NC_008066.1	15.01	0.50	**48.03**	4.68	Chlorocebus sabaeus ★
SRR13441704	NC_013993.1	5.21	0.07	**35.17**	0.96	Homo sp. Altai
SRR13441704	NC_023100.1	4.61	0.05	**30.75**	0.76	Homo heidelbergensis
SRR13441704	NC_053822.1	0.18	0.00	19.58	0.22	Cricetulus barabensis
SRR13441704	NC_025654.1	3.11	0.08	8.51	0.16	Anser indicus
SRR13441704	NC_009748.1	0.0	0.00	7.47	0.23	Chlorocebus tantalus
SRR13441704	NC_034277.1	1.99	0.04	7.28	0.09	Chlorocebus djamdjamensis isolate
SRR13441705	NC_012920.1	**47.98**	2.24	**89.03**	17.27	Homo sapiens
SRR13441705	NC_007936.1	10.67	0.12	**73.07**	4.18	Cricetulus griseus
SRR13441705	NC_011137.1	16.76	0.66	**59.51**	3.24	Homo sapiens neanderthalensis
SRR13441705	NC_008066.1	11.34	0.33	**43.28**	4.33	Chlorocebus sabaeus
SRR13441705	NC_023100.1	4.91	0.14	29.8	0.84	Homo heidelbergensis
SRR13441705	NC_013993.1	4.66	0.12	25.95	0.83	Homo sp. Altai
SRR13441705	NC_023832.1	4.98	0.33	11.69	0.49	Anser cygnoides
SRR13441705	NC_053822.1	1.14	0.01	9.08	0.14	Cricetulus barabensis
SRR13441705	NC_011196.1	7.43	0.54	8.8	0.31	Anser anser
SRR13441705	NC_024933.1	1.76	0.03	7.68	0.23	Chlorocebus cynosuros
SRR13441708	NC_012920.1	**68.27**	3.04	**89.15**	17.16	Homo sapiens
SRR13441708	NC_007936.1	15.85	0.29	**74.71**	3.91	Cricetulus griseus
SRR13441708	NC_011137.1	24.05	0.60	**65.2**	3.40	Homo sapiens neanderthalensis
SRR13441708	NC_008066.1	15.08	0.52	**43.94**	4.06	Chlorocebus sabaeus
SRR13441708	NC_013993.1	10.54	0.16	**33.02**	1.03	Homo sp. Altai
SRR13441708	NC_023100.1	8.23	0.12	29.15	0.76	Homo heidelbergensis
SRR13441708	NC_053822.1	1.14	0.01	11.73	0.14	Cricetulus barabensis
SRR13441708	NC_009747.1	3.46	0.05	8.76	0.29	Chlorocebus pygerythrus
SRR13441708	NC_024933.1	2.06	0.02	6.28	0.17	Chlorocebus cynosuros
SRR13441708	NC_034276.1	1.36	0.01	5.57	0.12	Chlorocebus aethiops x

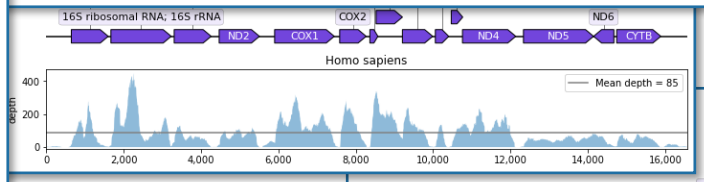
Importance of keeping raw and metadata

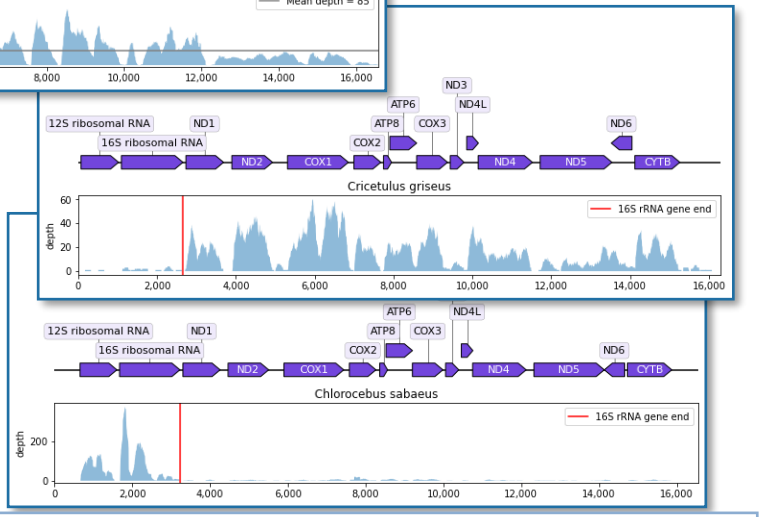
- Mutations, phylogeny, host genomes, cell culture, flow cell id ….





pos	ref	SRR13441704	SRR13441705	SRR13441708
8782	C	t.TTa..T	TT,..,	a.tTtTtT.
17039	A	.$....G...,g..g,	Gg,Gg,,,GTgg^],	..,C.GG,,,,,,g,
17634	C	.....G..Gg,,,g,..g,,,,,.G.,...	,G.,,,.,.,..G..	.g..g,G.GG,,,,.....
18082	A	g.,g.,G,GGg.g.,,g,,,,.g,gGgG	G,,,,n,,,,,GG,G..g.,G*..,	.g,gg,,,g,g..g,G,,,,,,,.c,.
25498	C	.,gT.TT,t,,,.T..^].	T.Tg,..T.tT	..t.,t.





Csabai, I., Papp, K., Visontai, D., Stéger, J. and Solymosi, N., Unique SARS-CoV-2 variant found in public sequence data of Antarctic soil samples collected in 2018-2019. Submitted 2022. https://www.researchsquare.com/article/rs-1177047/v1 , https://www.researchsquare.com/article/rs-1330800/v1

Key challenges: amount of data and complexity of models

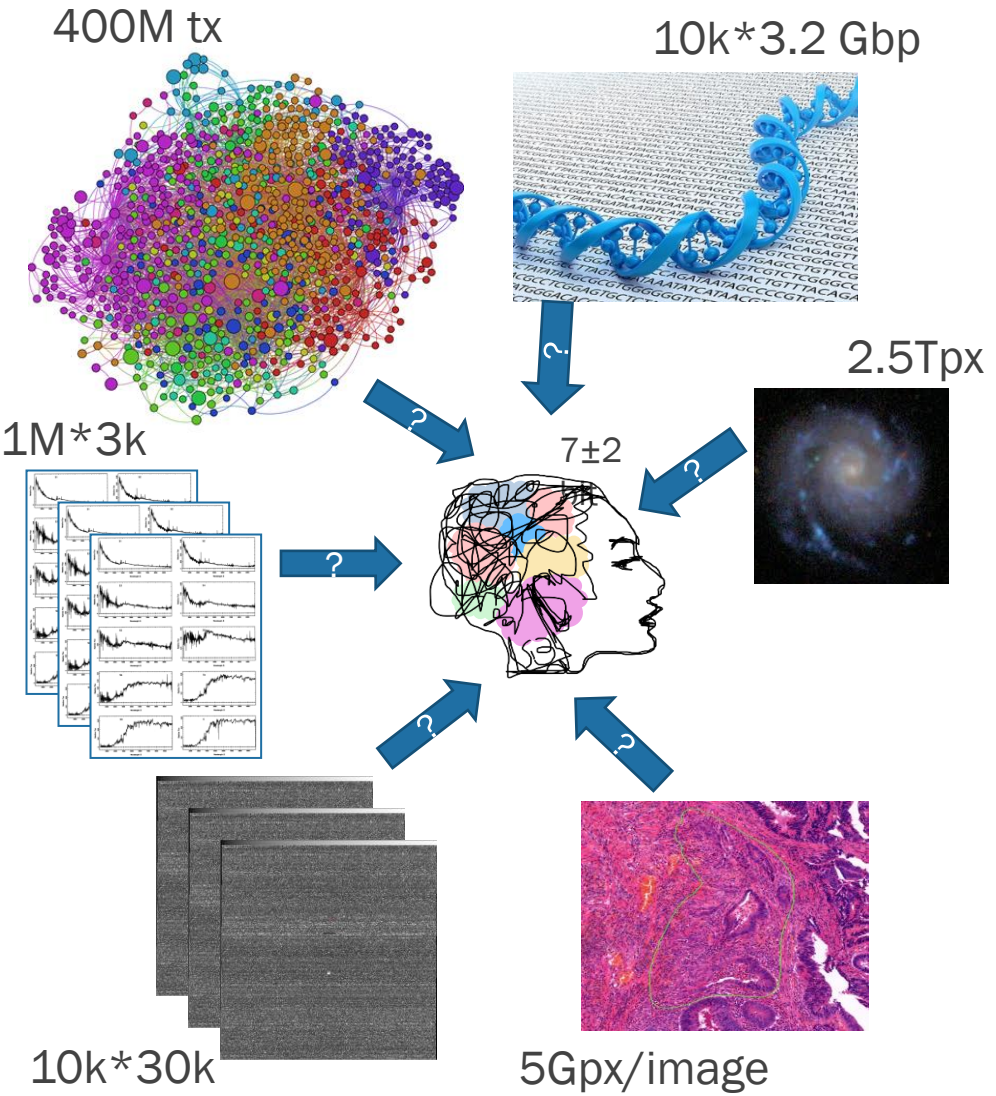
ra	dec	u	g	r	i	z	deVRad_r	deVPhi_r	redshift	class
348.90253	1.2718862	19.38905	18.24496	17.58728	17.20807	16.90905	3.295783	28.87819	0.03212454	GALAXY
51.443695	1.2700727	19.52808	17.96541	17.03493	16.53754	16.14154	7.599091	63.68505	0.1213151	GALAXY
51.483584	1.2720127	18.72268	17.3852	16.81134	16.51803	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

Spectra: 1 million 3000 dim vectors

2.5 Terapix image

Scientific observations often result data as multidimensional vector space



Natural intelligence
Artificial intelligence

7±2 bit

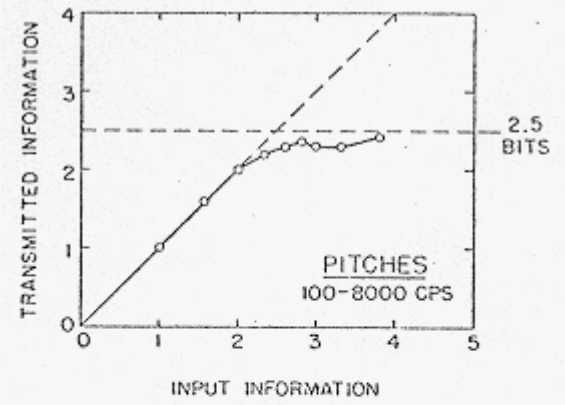


FIG. 1. Data from Pollack (17, 18) on the amount of information that is transmitted by listeners who make absolute judgments of auditory pitch. As the amount of input information is increased by increasing from 2 to 14 the number of different pitches to be judged, the amount of transmitted information 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)

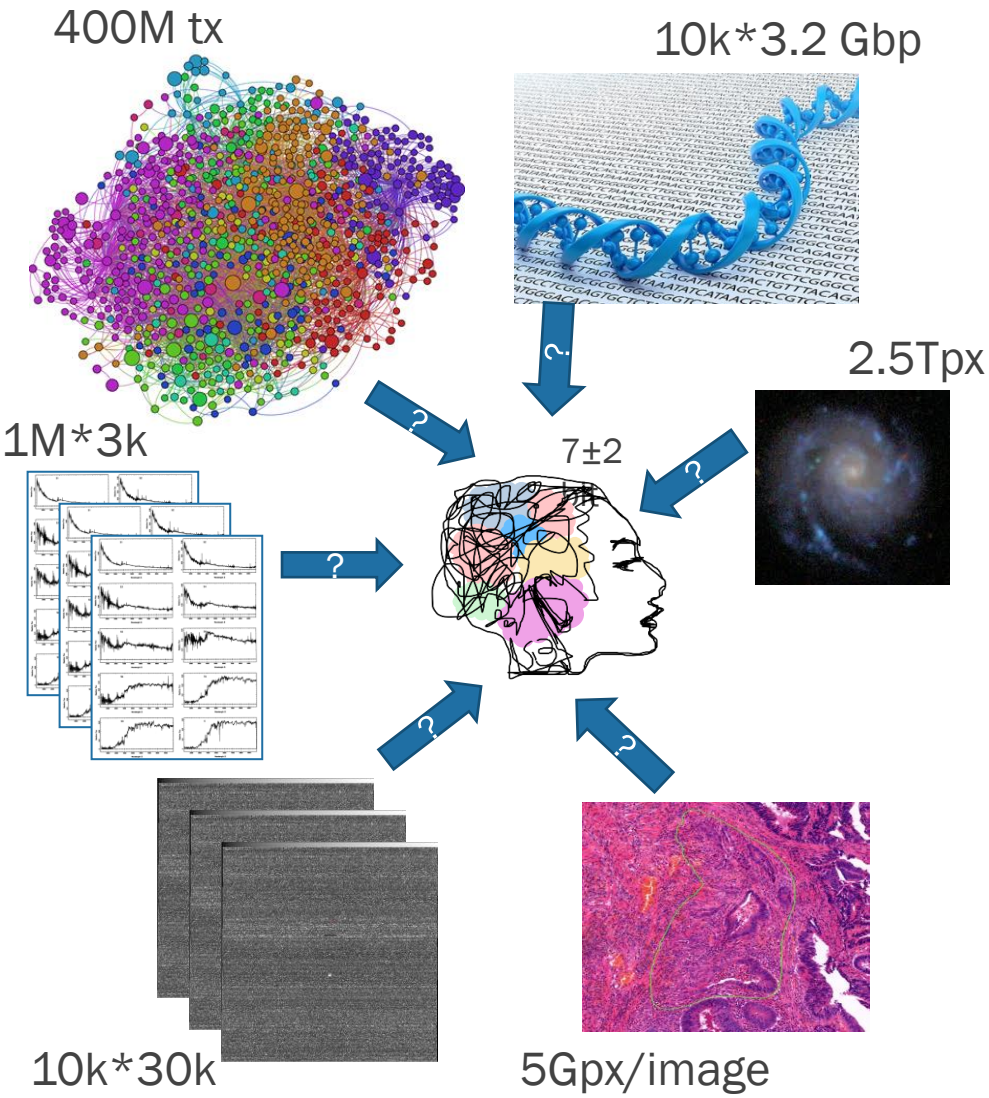
Pollack, I. *The information of elementary auditory displays.* J. Acoust. Soc. Amer., 1952, 24, 745-749.

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

Key challenges: amount of data and complexity of models



SIGGRAPHASIA2009

Shadow Art

Niloy J. Mitra
IIT Delhi / KAUST

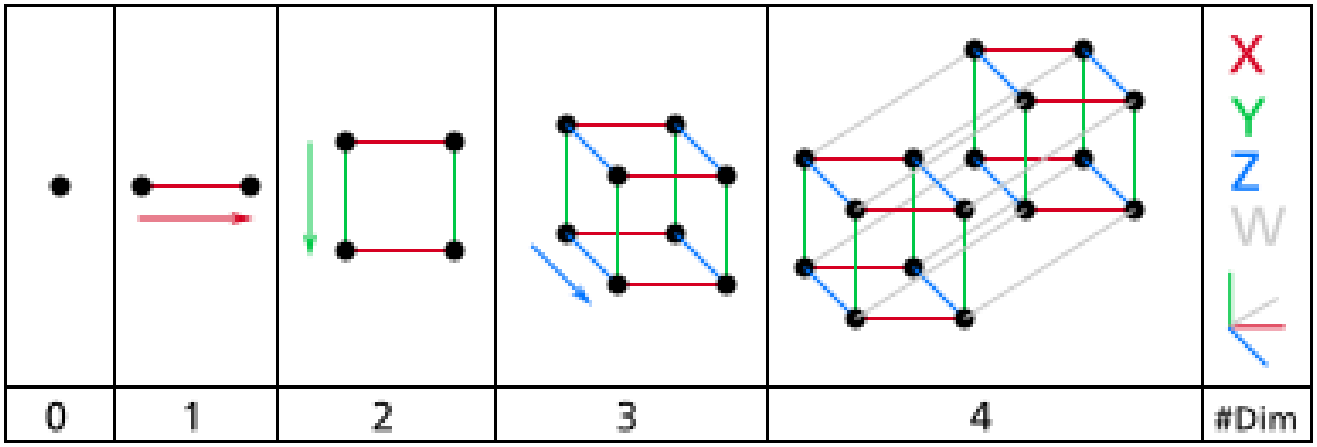
Mark Pauly
ETH Zurich

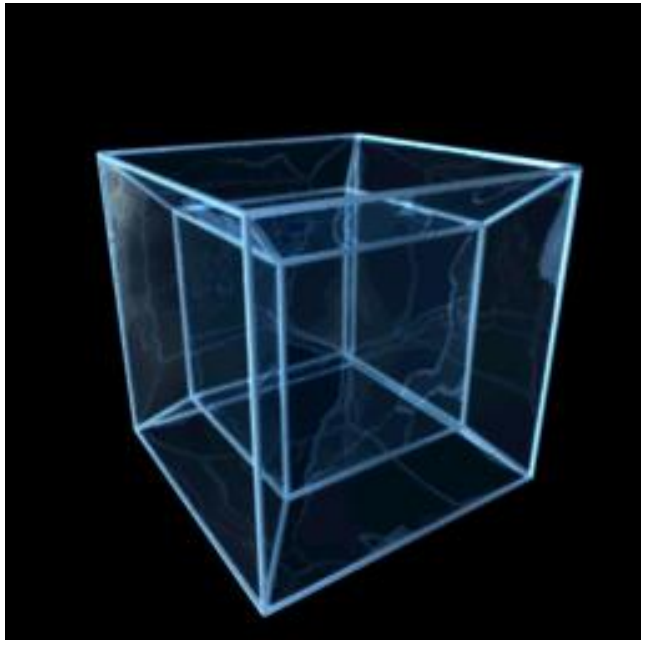
Due to the underlying physical laws, data vectors do not fill the whole space, rather lie on lower dimensional surface/subspace (This is why we can understand the word!)

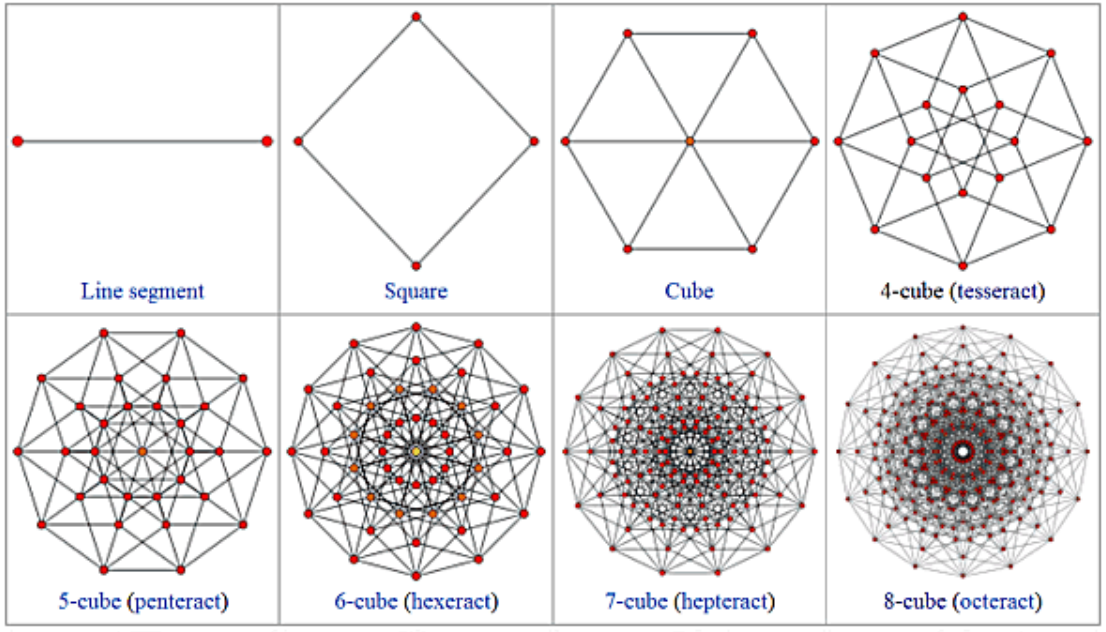
$pV = NkT$

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

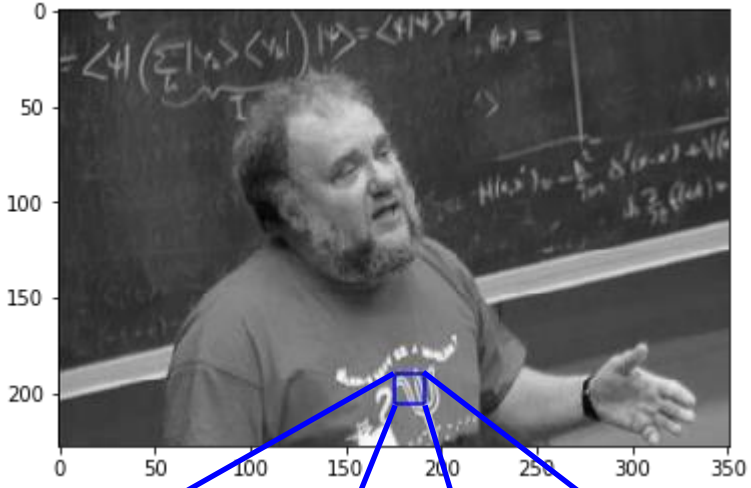
Data hyperspace – dimensions - projections





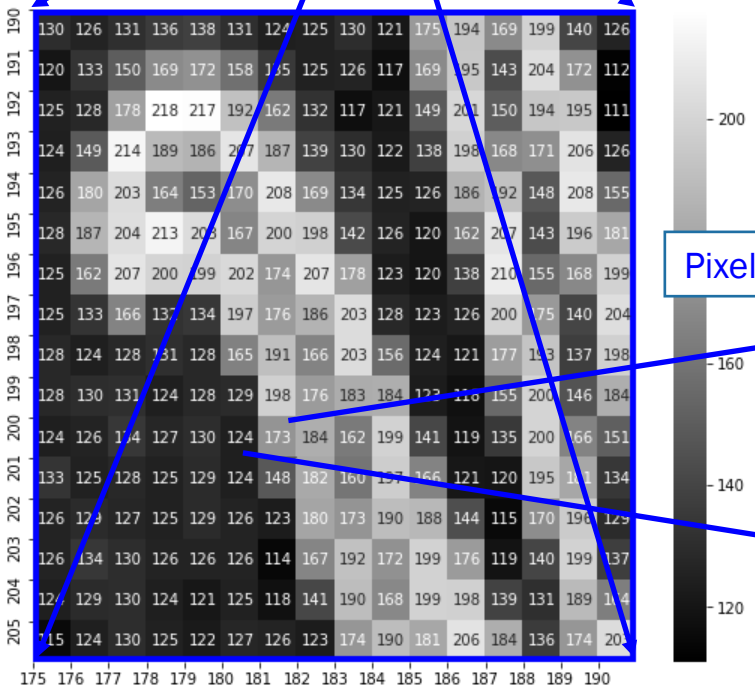


Example: each picture is a point in the hyperspace



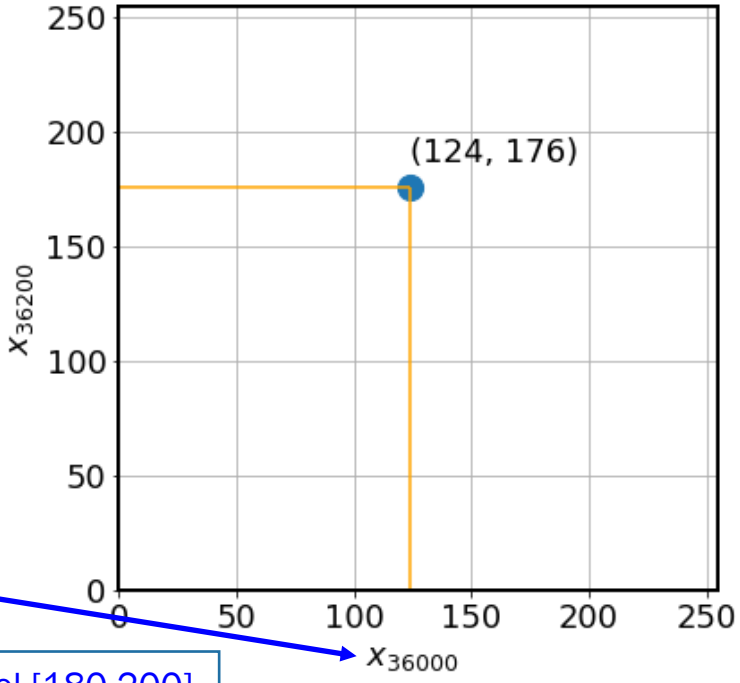
One 228x351 pixel image =
80028 numbers =
x_1, x_2, x_3, ... x_{80028} coordinates in 80028D

http://atomcsill.elte.hu/NEW/



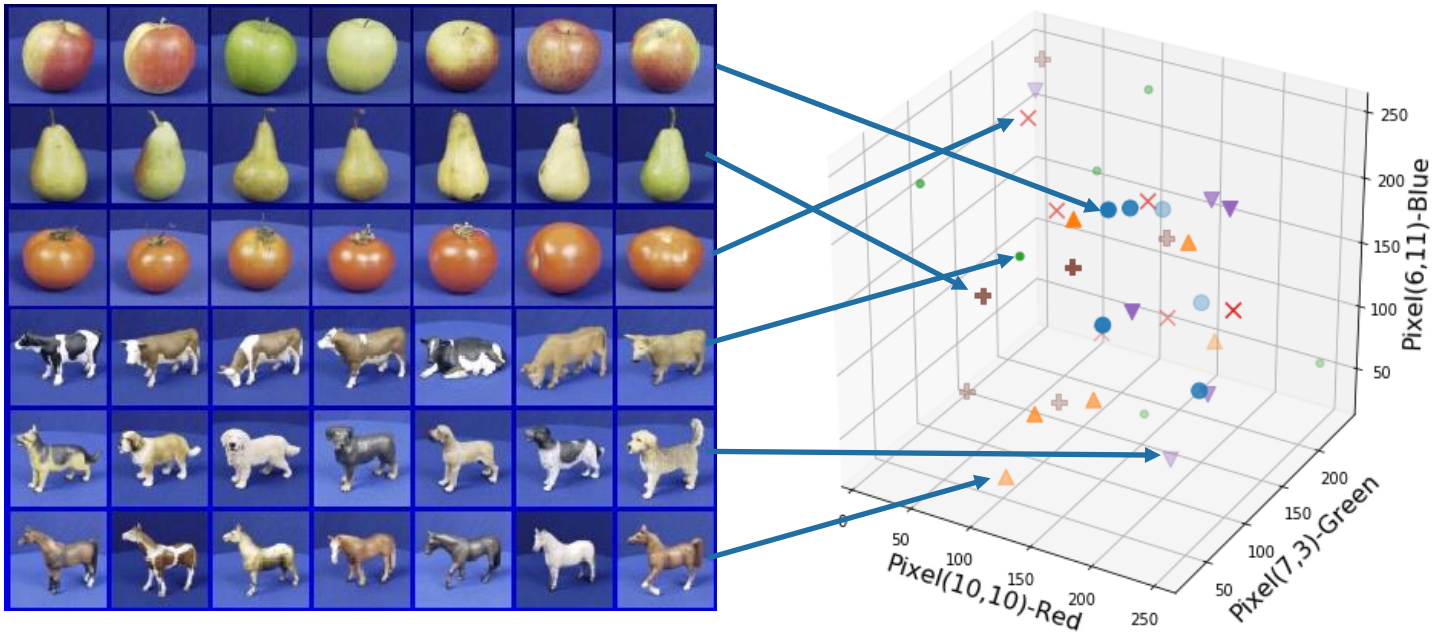
Pixel [181,200]

A 2D subspace of the 80028D hyperspace



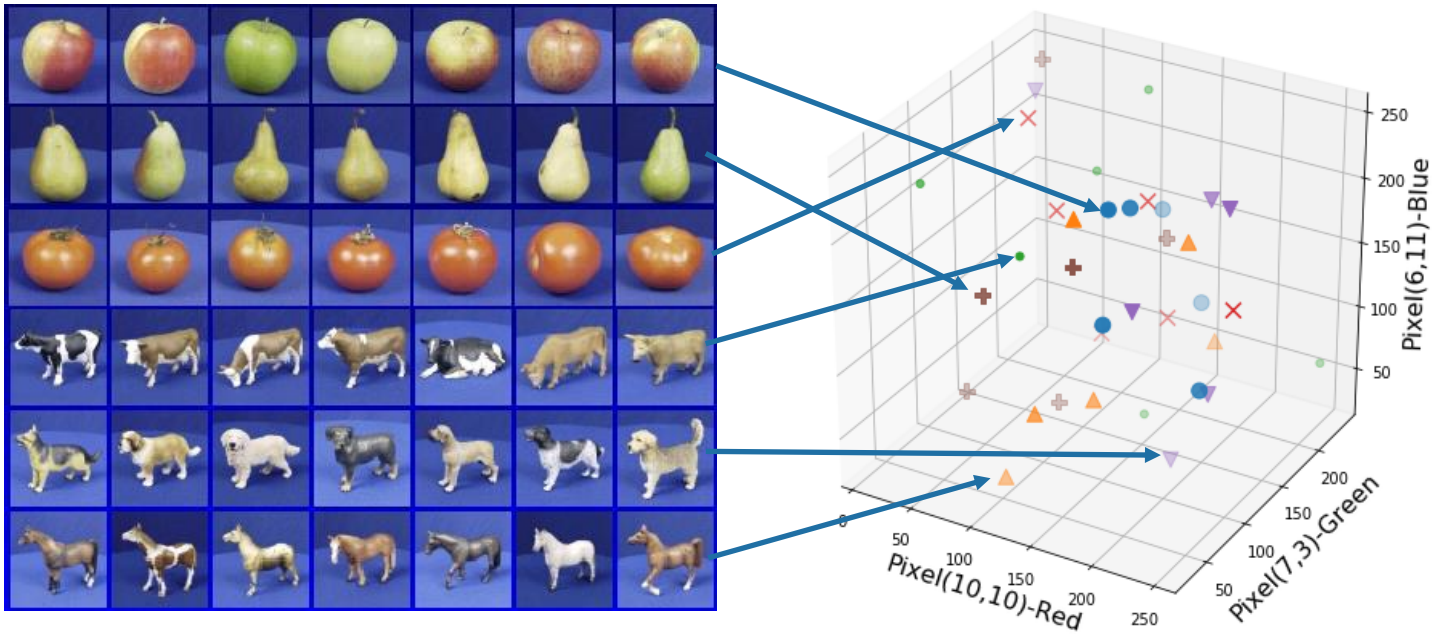
Pixel [180,200]

Many pictures: many points



Video: even more points

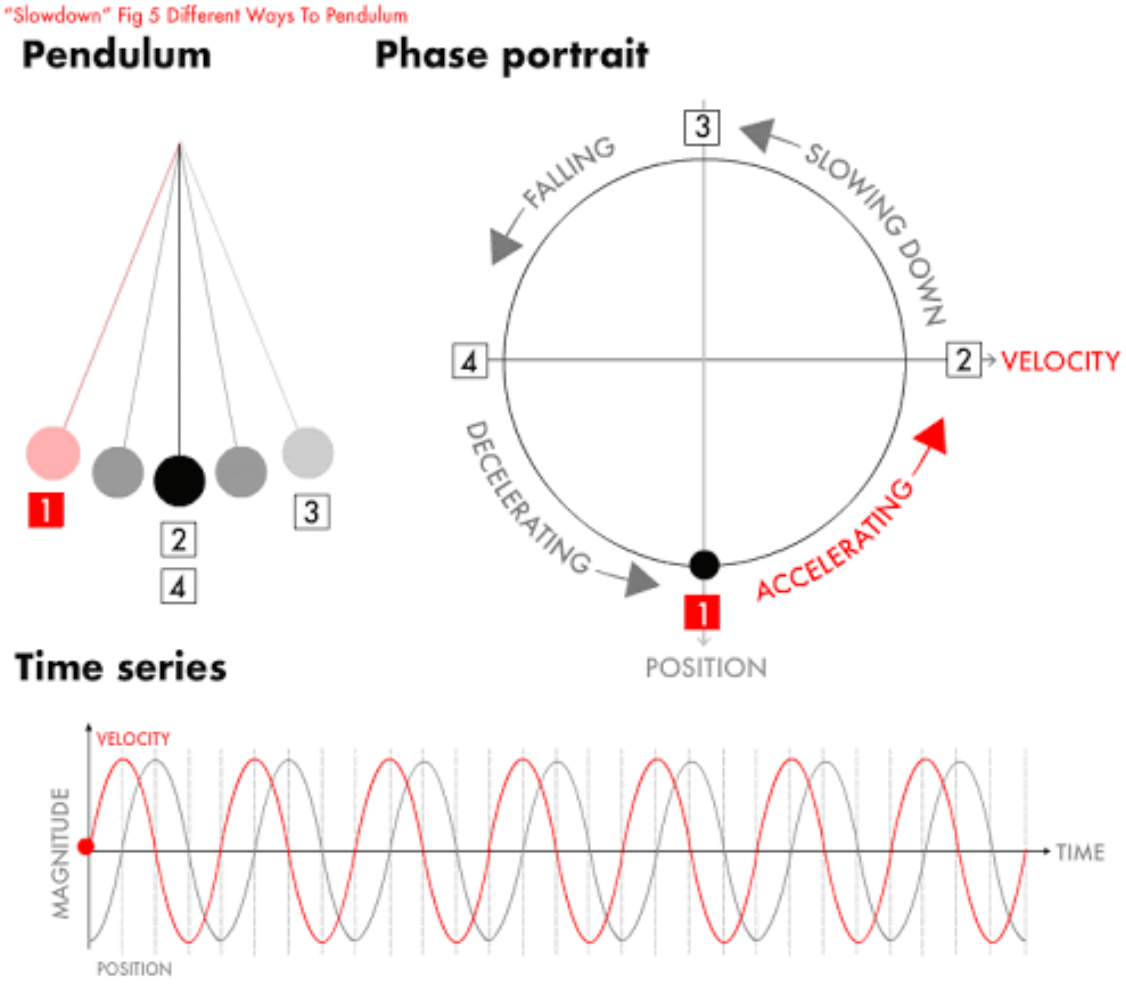
Many pictures: many points



Video: even more points

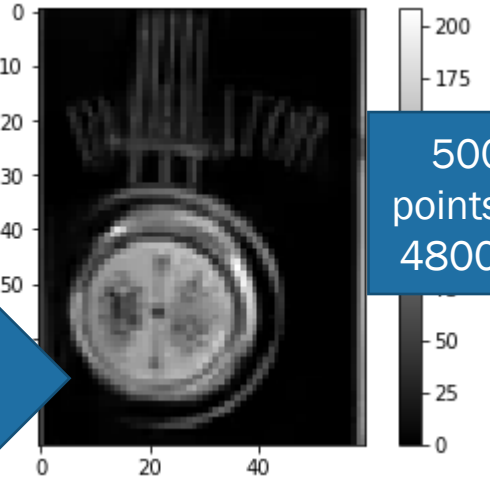
But knowledge of physics can help!

Phase space of a pendulum

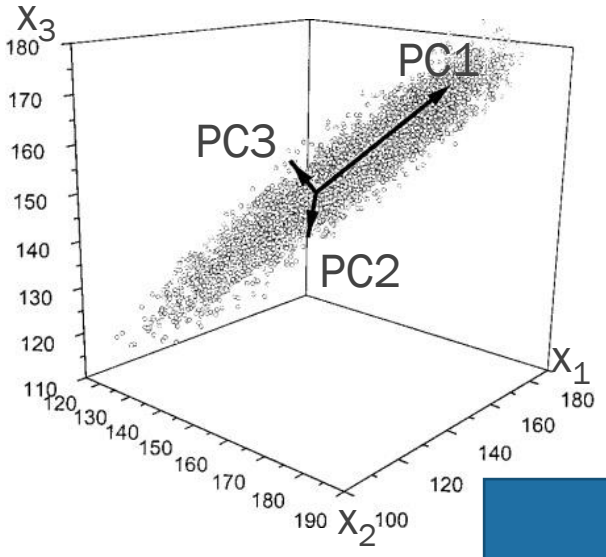


Pendulum video principal component projection to 2D

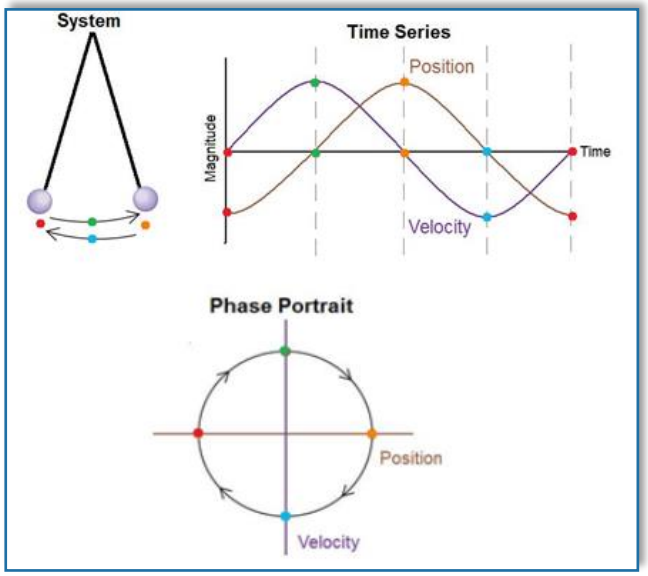
500 frames 60x80 cutouts

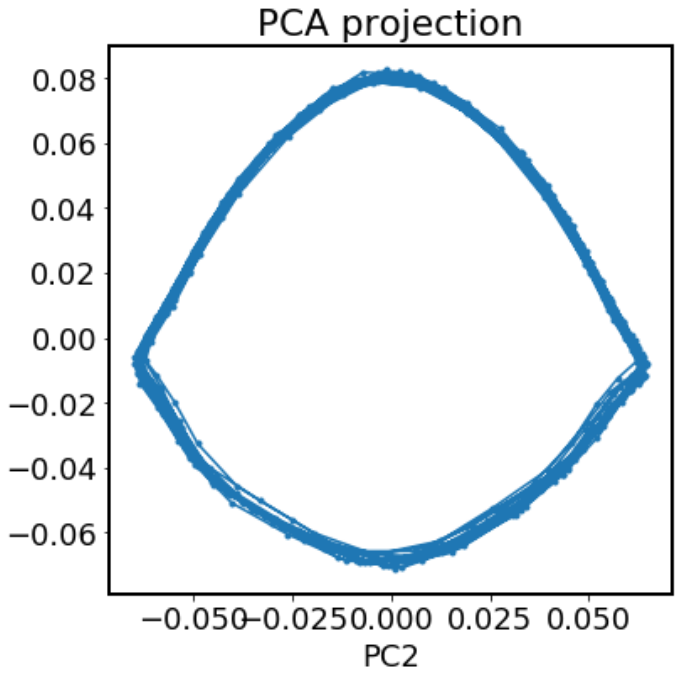


500 points in 4800 D

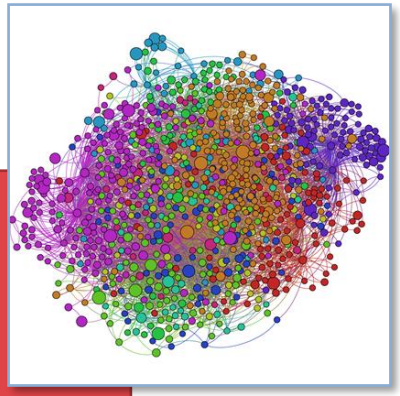


Rotation+ projection





Complex systems – complex models

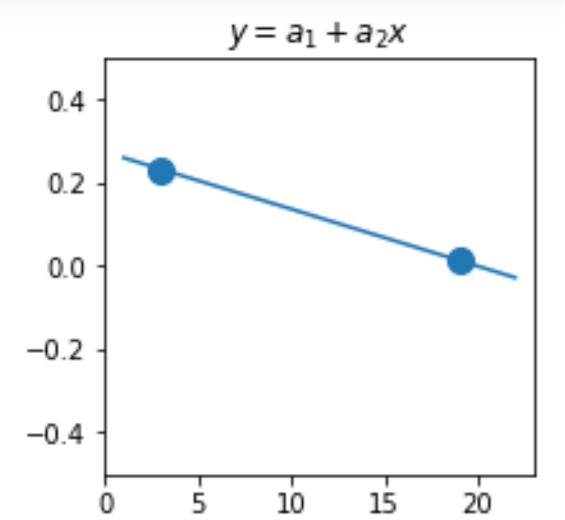


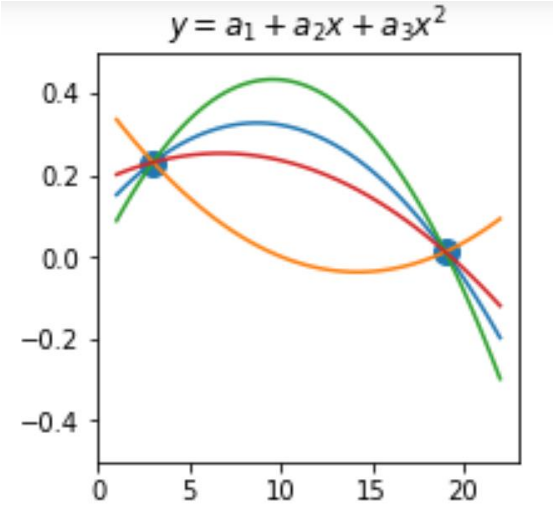
To understand complex systems we need complex models

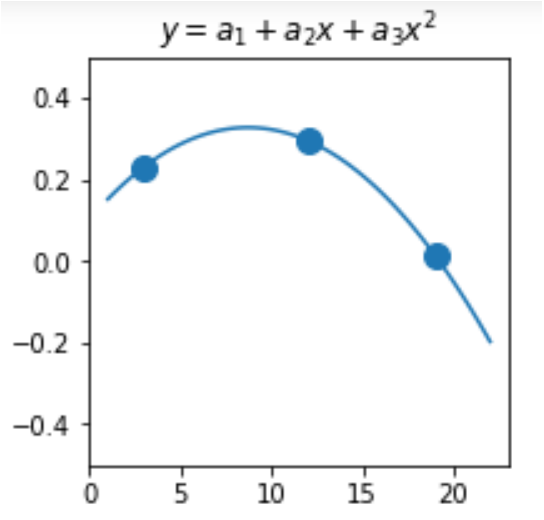
Complex models have many details, parameters, …

We need

- Huge amount of data to set up, constrain, parametrize the models
- Powerful computers and clever algorithms

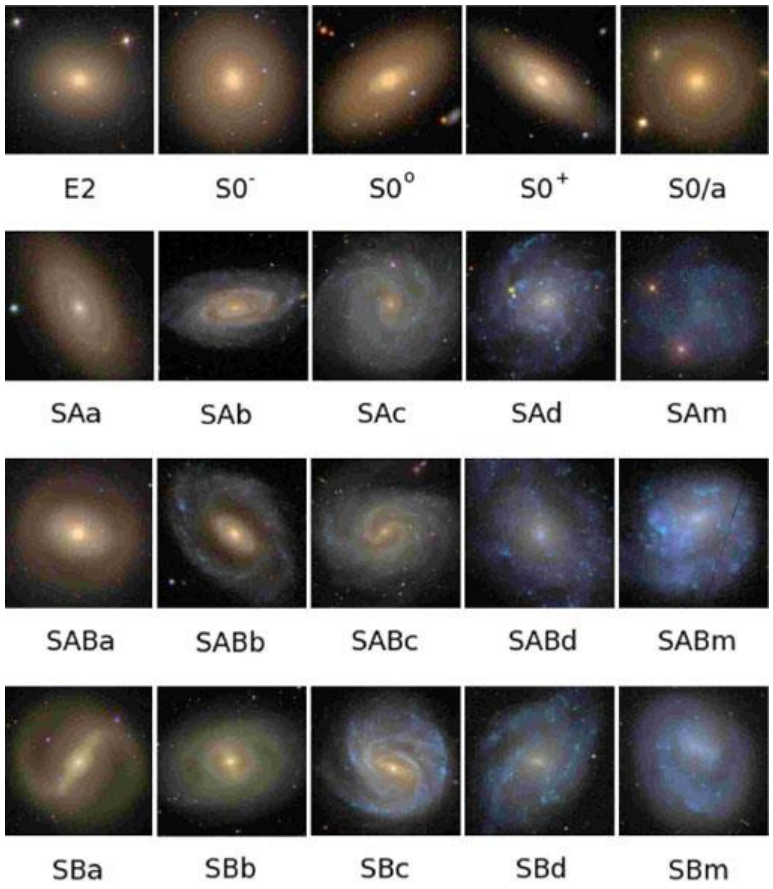






Complex function regression: machine learning!

Image classification: traditional hand crafted code



f( ) = “apple”

f( ) = “tomato”

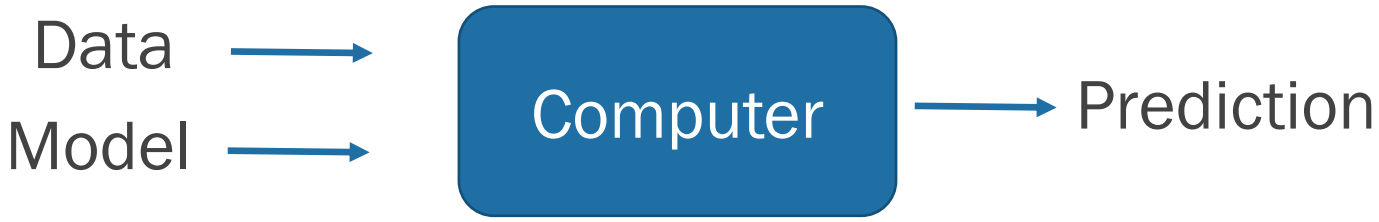
f( ) = “cow”

IF color=red AND profile=smooth THEN type:=elliptical
IF color=blue AND HAS(arms) THEN type:=spiral

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

Slide credit: L. Lazebnik

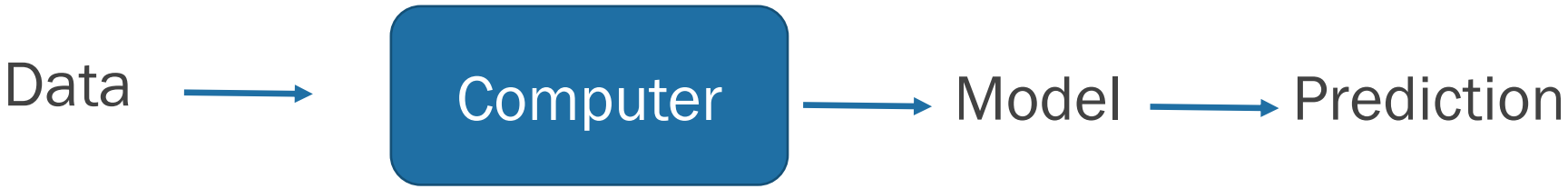
AI: paradigm shift



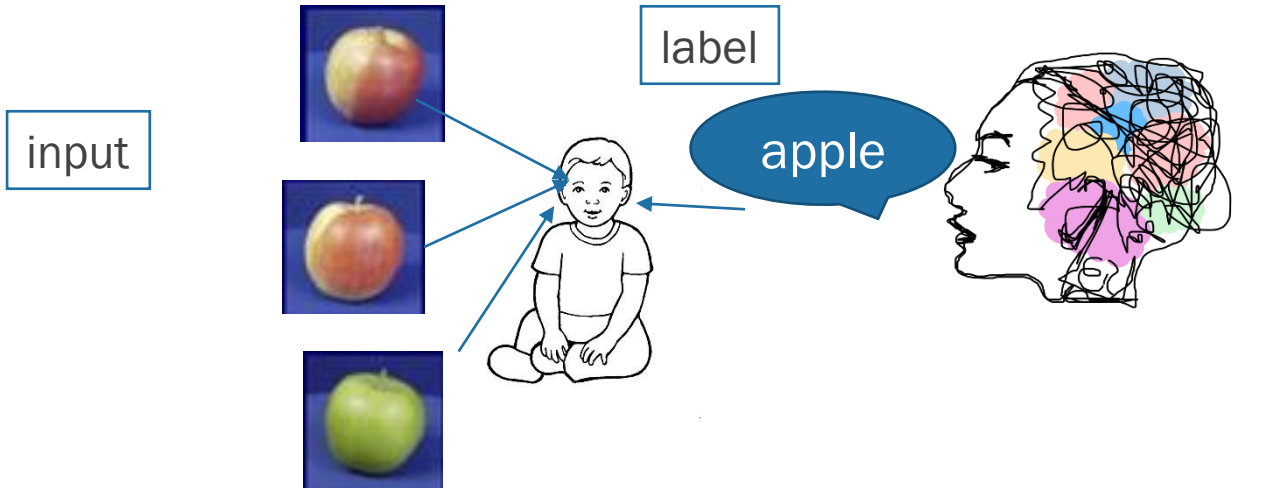
Example: Image recognition
Method: hand crafted features

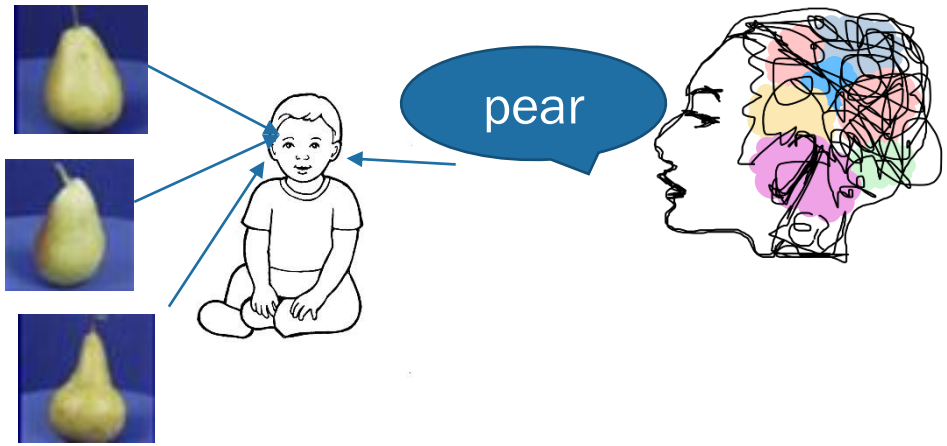
f( ) = “apple”
f( ) = “tomato”
f( ) = “cow”

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

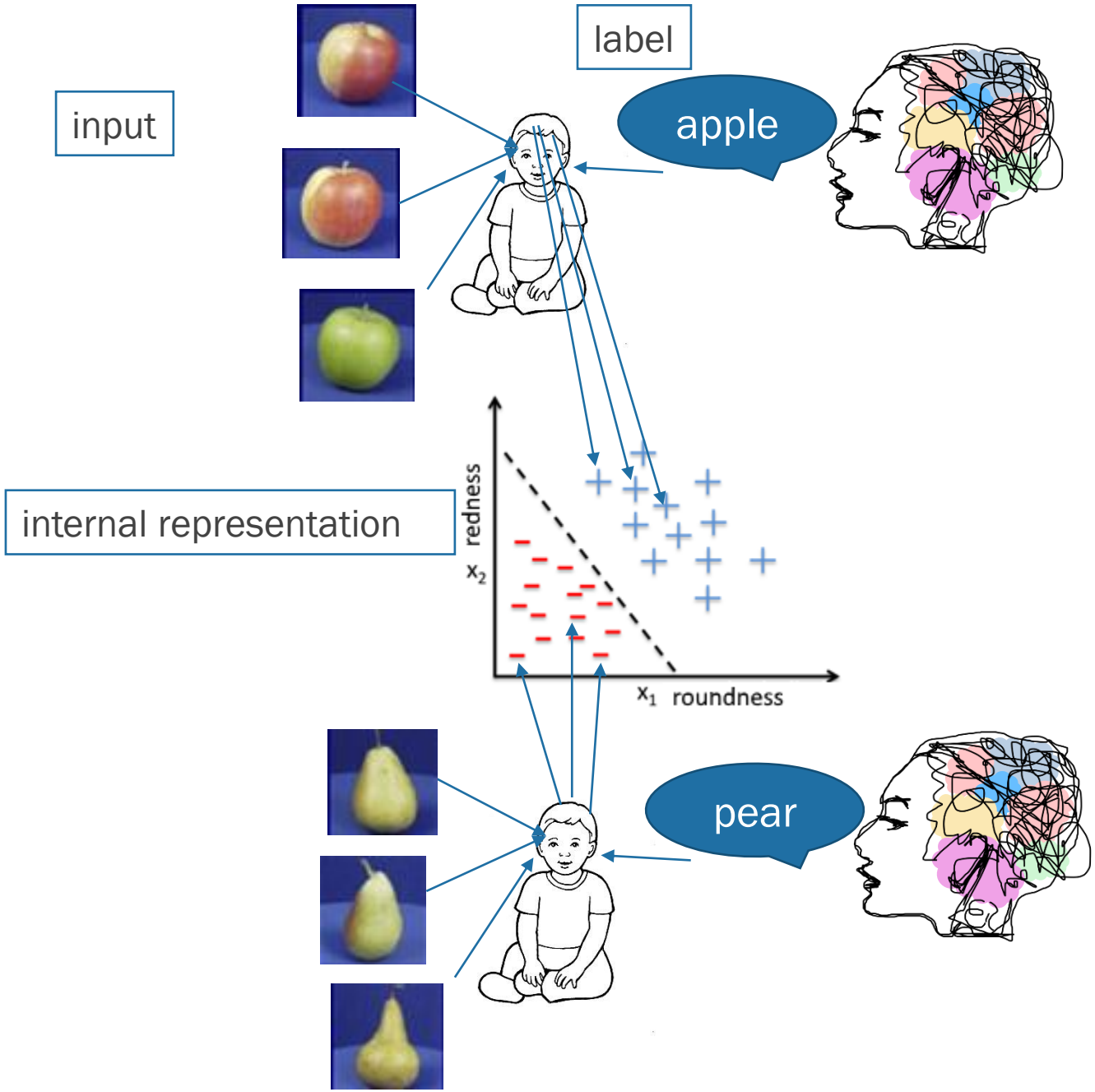


Supervised learning

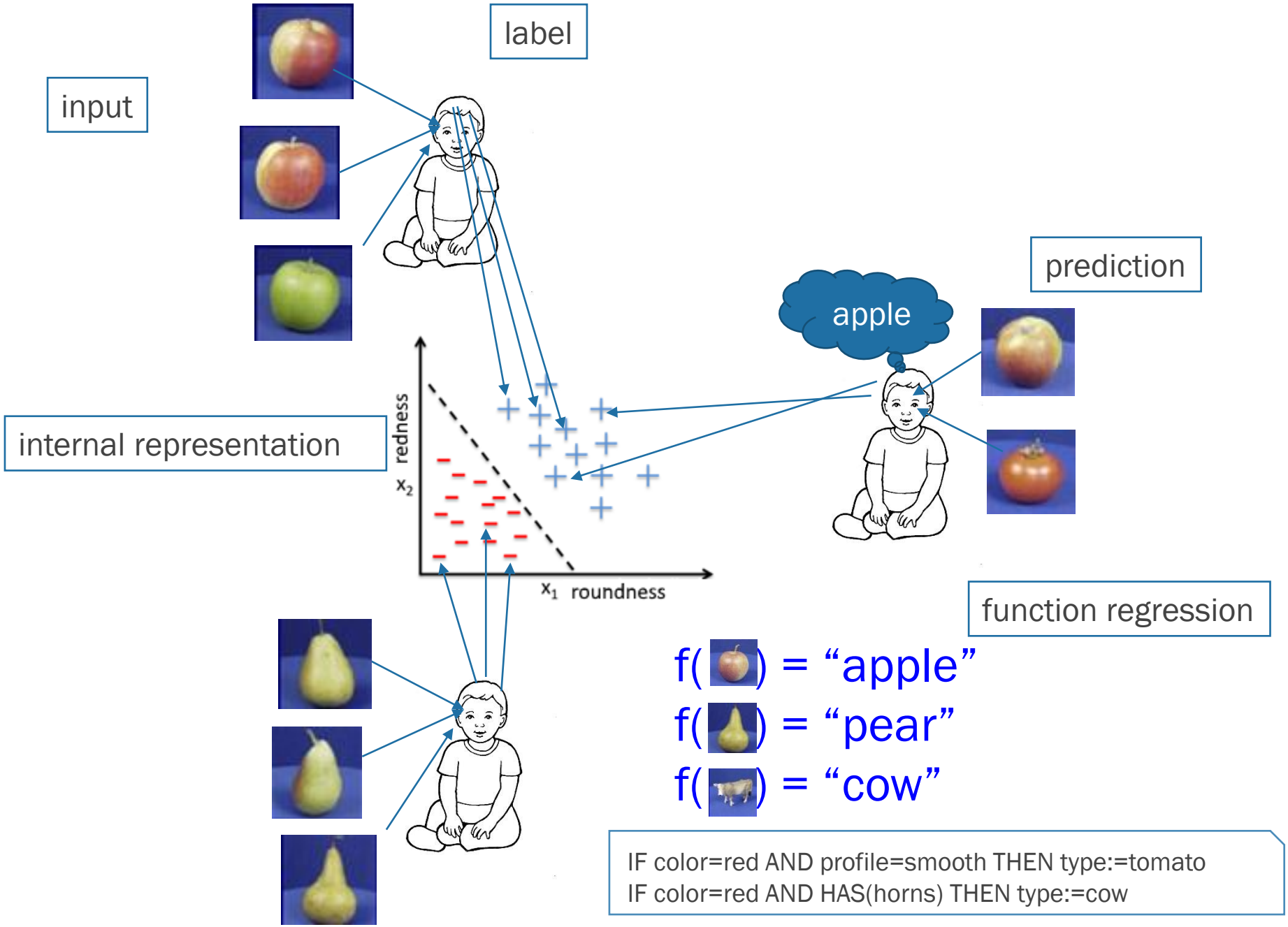




Supervised learning



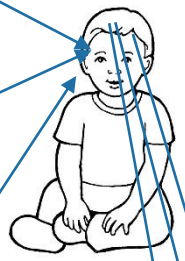
Supervised learning

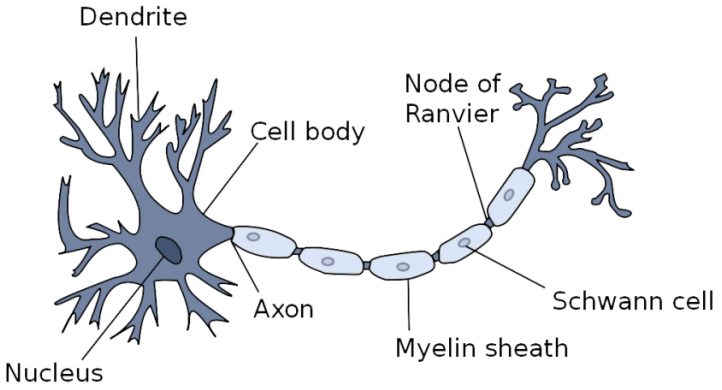


Supervised learning: neural net

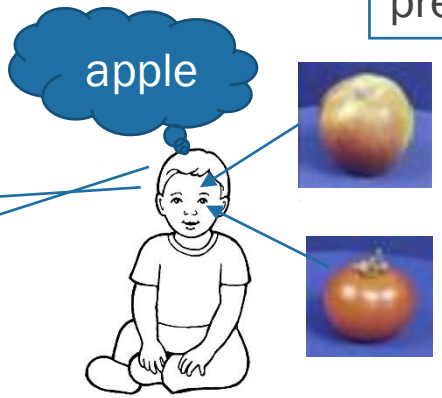
label

input

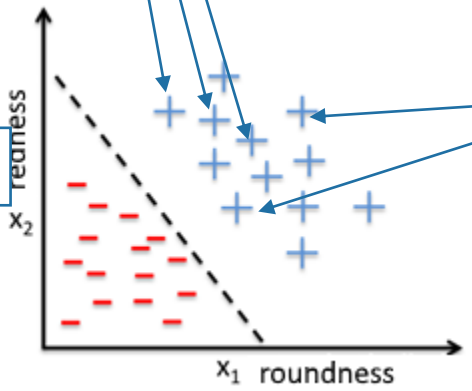




prediction



internal representation



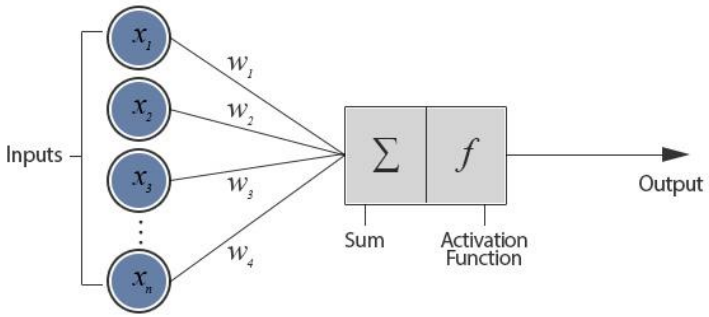
$$y = f\left(\sum_i w_i x_i\right)$$

function regression

f( ) = “apple”

f( ) = “pear”

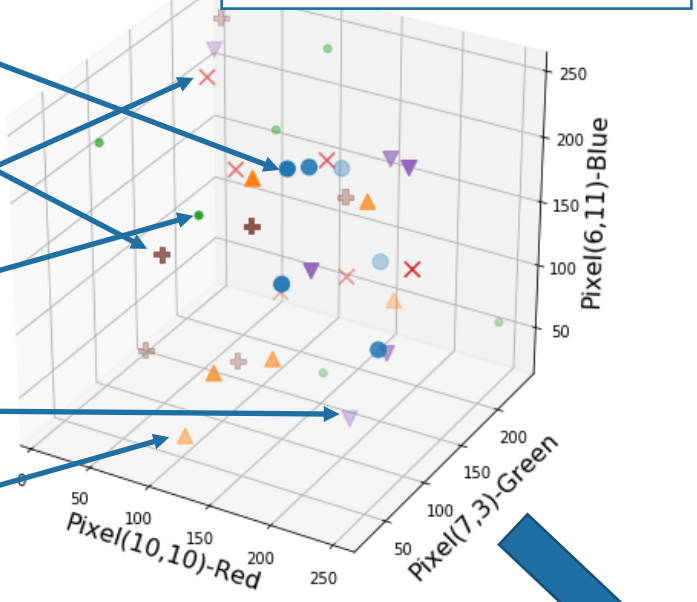
f( ) = “cow”



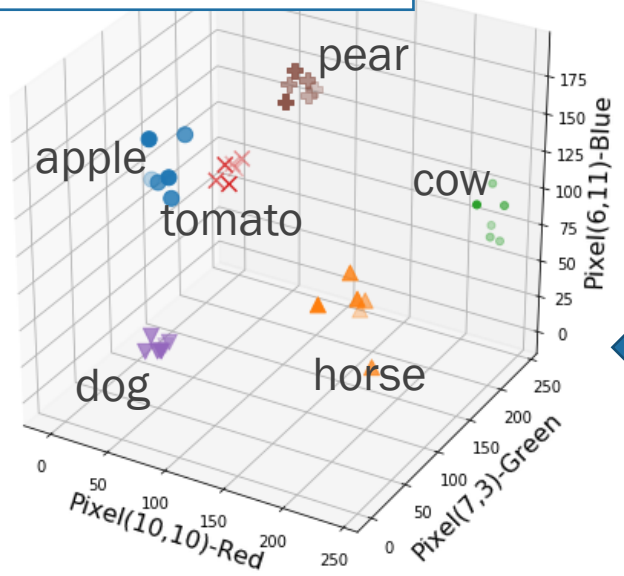
IF color=red AND profile=smooth THEN type:=tomato
IF color=red AND HAS(horns) THEN type:=cow

Latent space

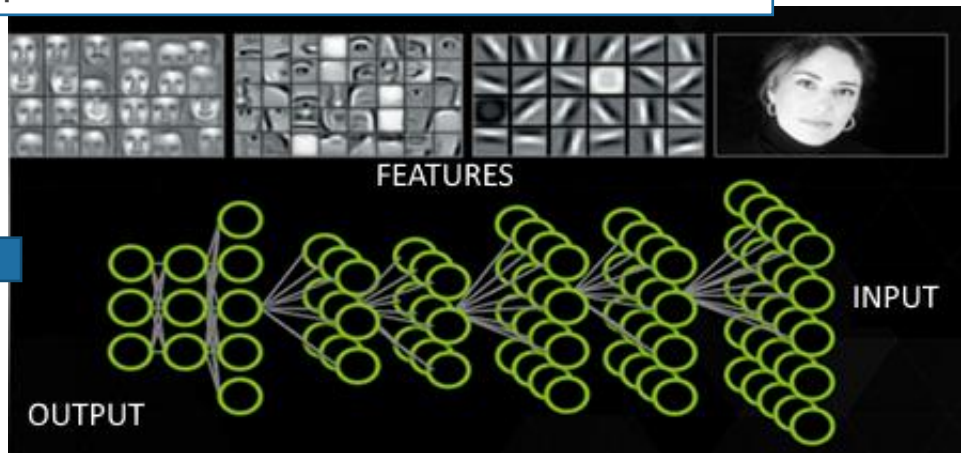
Raw modality space



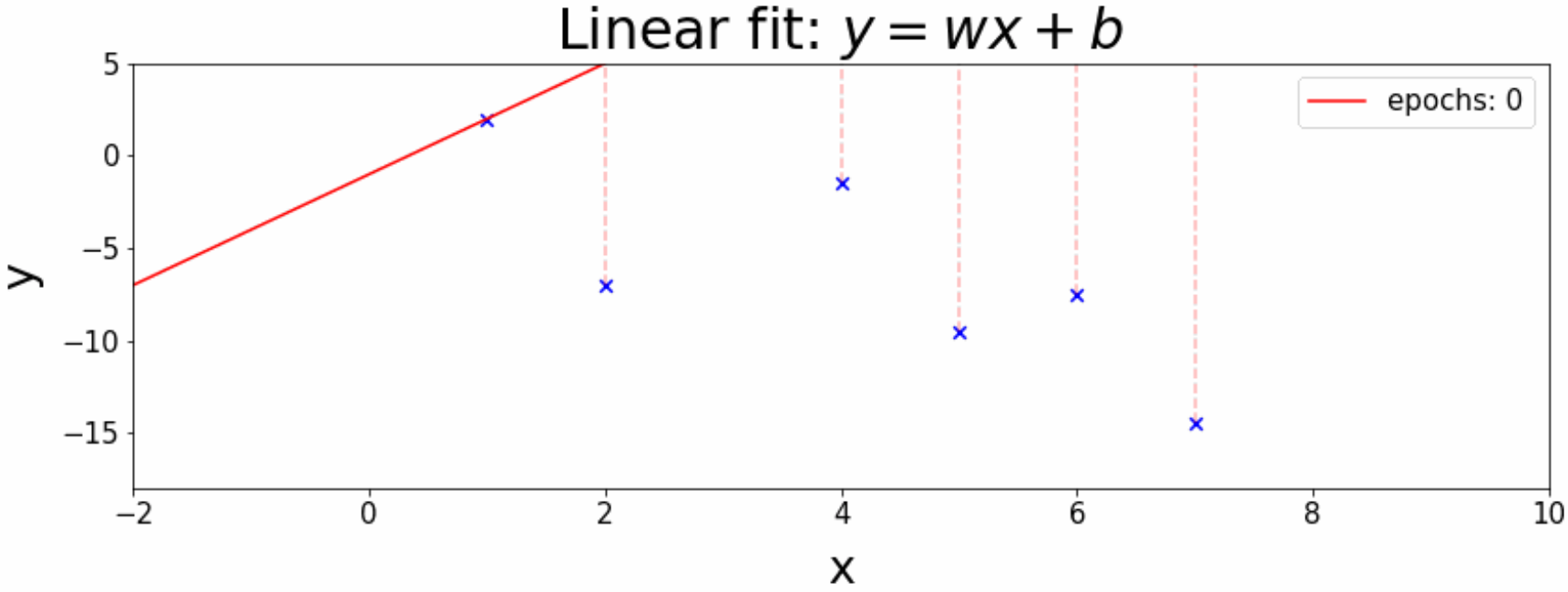
Latent representation

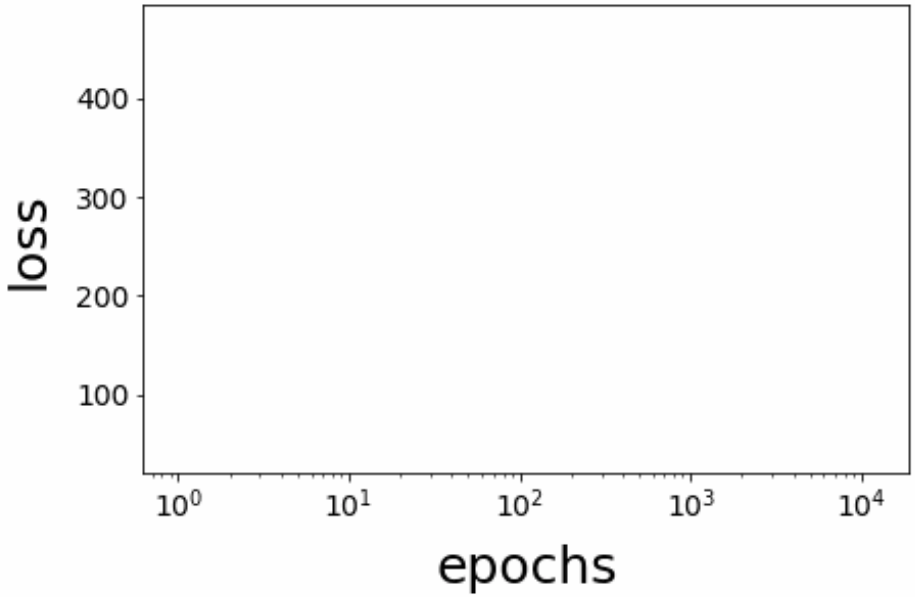


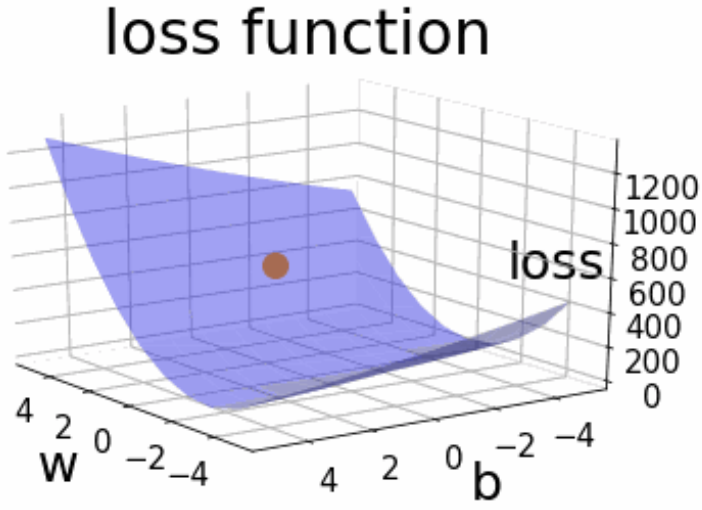
Deep convolution net: **transformation**



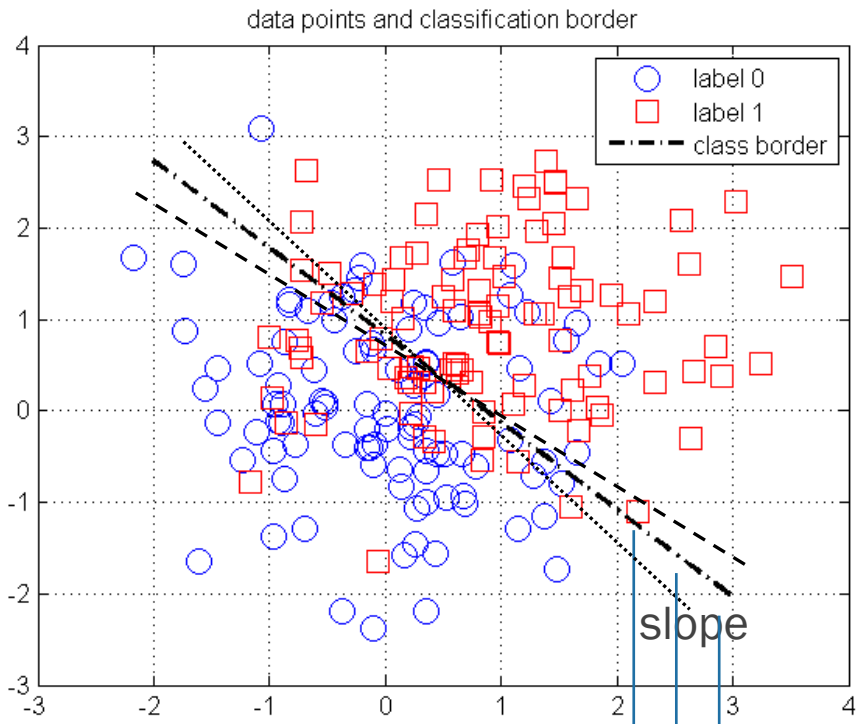
The machinery behind: linear algebra





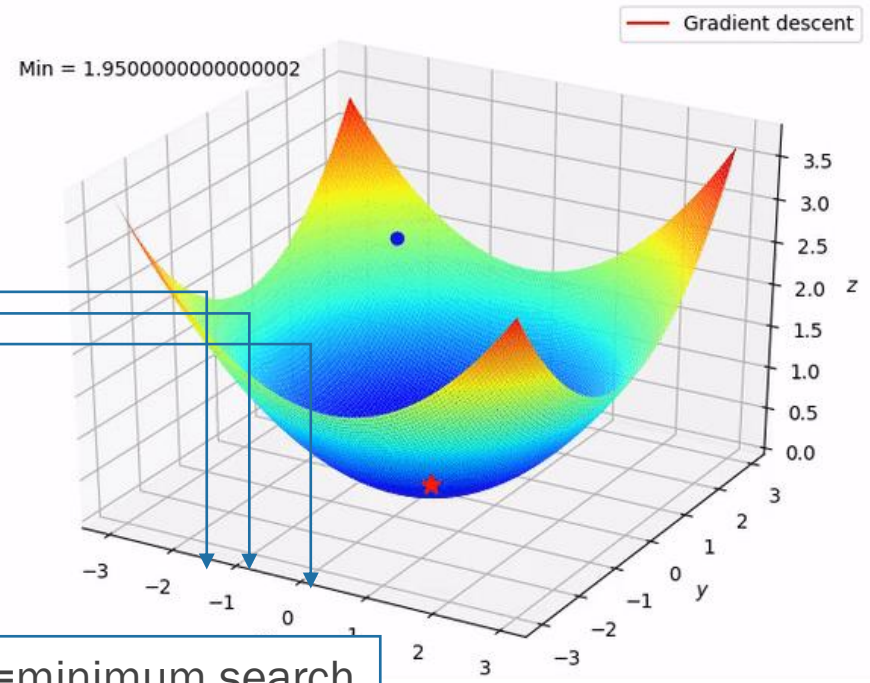


Learning -> loss function optimization



images -> points
in N dim space

Loss = number of wrong categorizations (error)



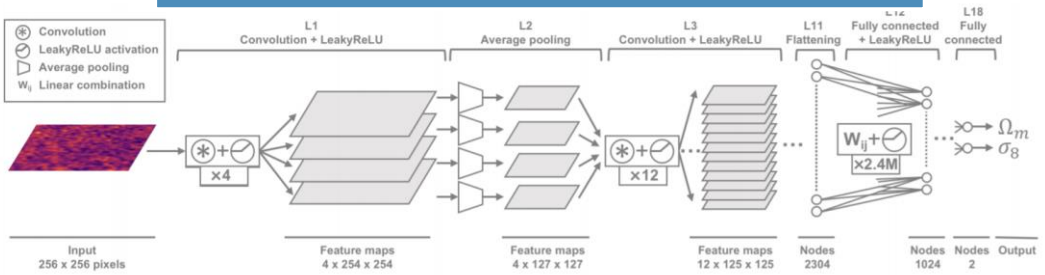
$$E(w_{ij}) = \sum_i \left(f_i(w_{ij}, x_j) - y_i\right)^2$$

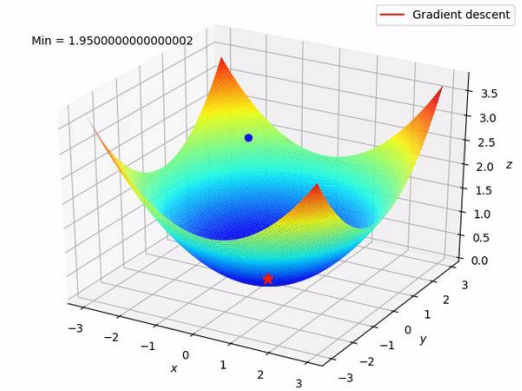
Learning=minimum search

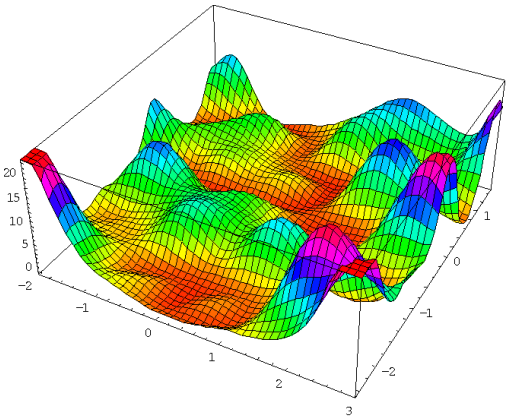
Challenges

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

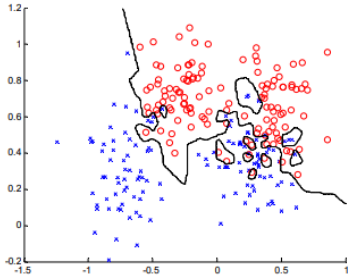
Typical network: 2M adjustable parameters





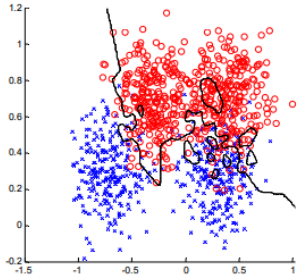


Training data



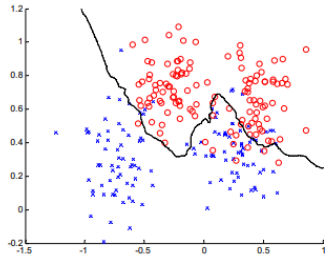
error = 0.0

Testing data



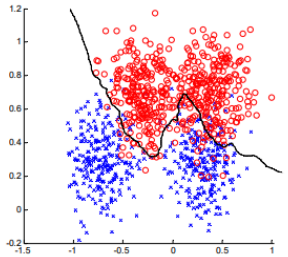
error = 0.15

Training data



error = 0.1120

Testing data

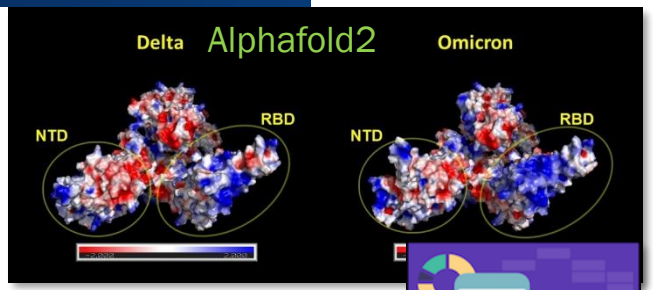


error = 0.0920

Machine Learning Research, Education, Applications @ Dept. of Physics of Complex Systems

- Mutations -> **antibiotics resistance**
  Matamoros et al., Pataki et al. 2020.
- Mobile sensors -> **Parkinson**
  Pataki @DREAM, Laki et al. 2016
- Mosquito images -> vector borne **diseases**
  Pataki et al. Sci.Rep. 2021
- **Medical imaging** -> breast cancer
  Ribli et al. @DREAM, Sci. Rep. 2018
- Weak lensing map -> **cosmology** parameters
  Ribli et al. Nature Astro. 2018, MNRAS 2019
- Lightning whistlers - > **space weather**
  Pataki et al. Space Weather 2022
- **Control of aging related methylation networks**
  Palla et al. Plos Biol. 2021
- **Pathology** images
  SOTE TKP collab.
- **Quantum neural computing**
- **MSc, PhD courses**

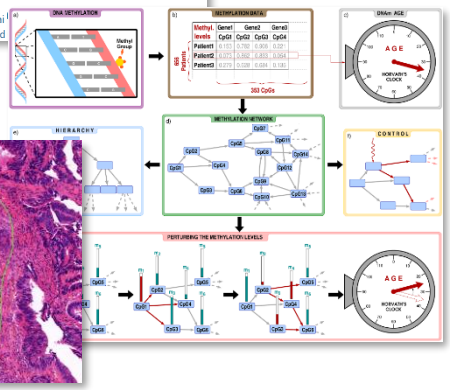
http://datascience.elte.hu

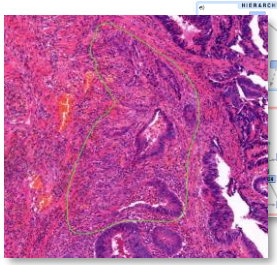


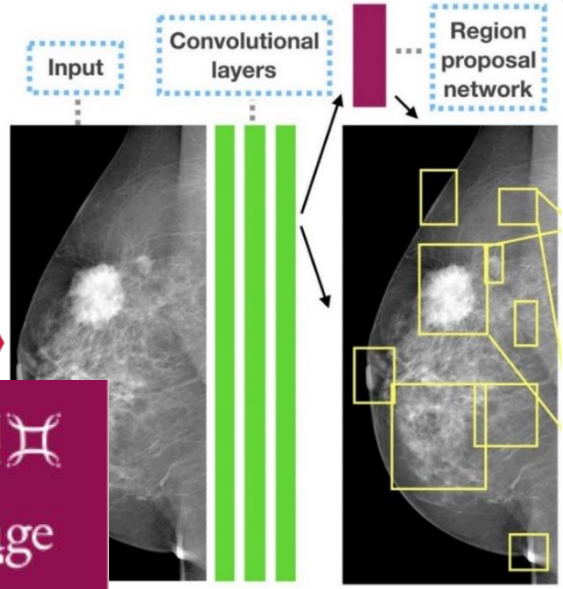
nature astronomy

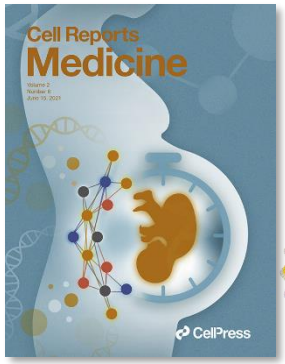
An improved cosmological parameter inference scheme motivated by deep learning

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









nature.com > scientific reports > articles > article

SCIENTIFIC REPORTS

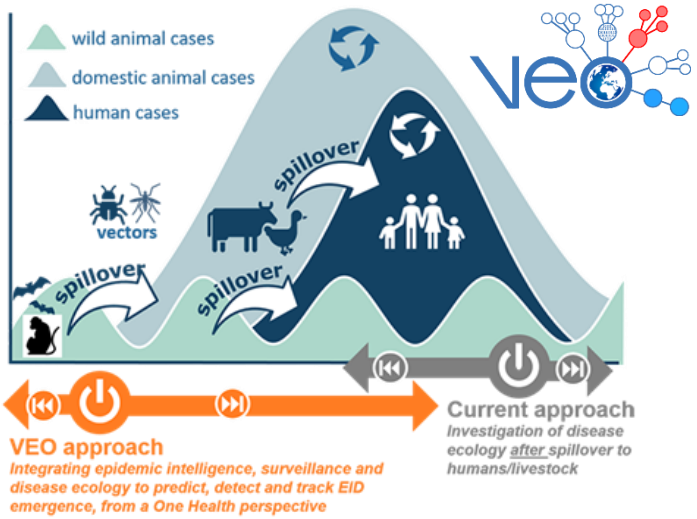
Detecting and classifying lesions in mammograms with Deep Lea...

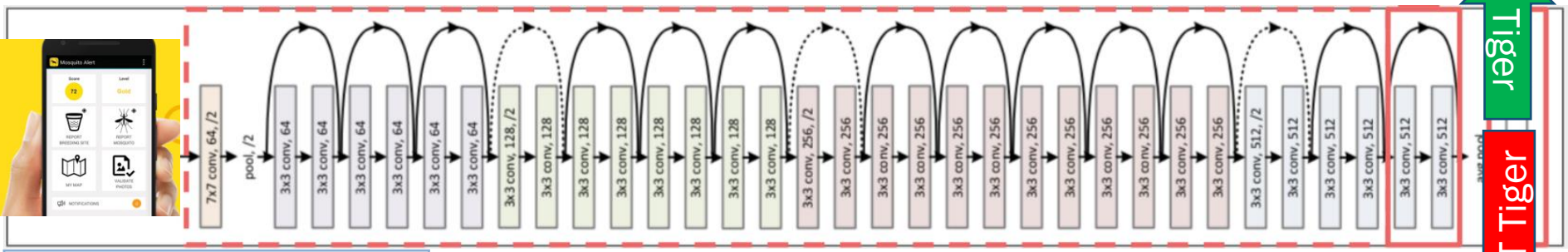
Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner & István ...

MOST POPULAR

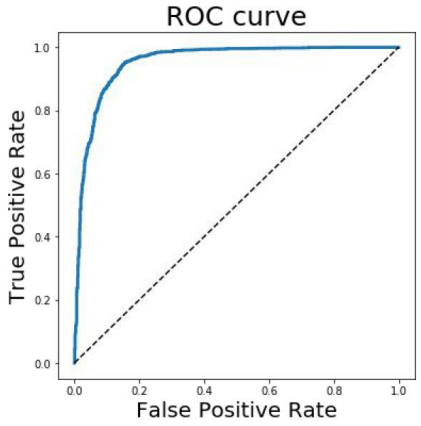
Vector borne diseases: MosquitoAlert image deep learning

"**Zika**, dengue, chikungunya, and yellow fever are all transmitted to humans by Ae. aegypti and Ae. Albopictus."





F. Bartumeus et al.
http://www.mosquitoalert.com/

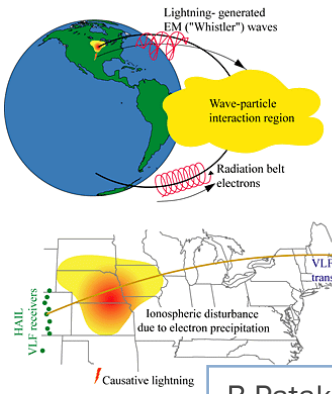


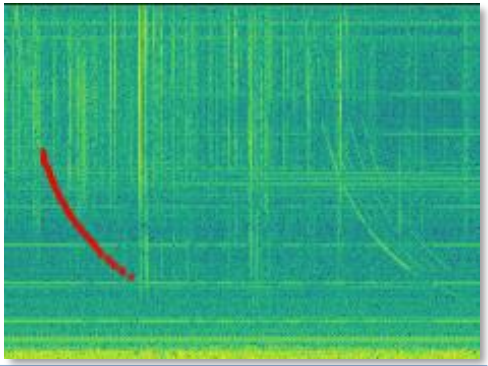
Pataki et al. Sci. Rep. 2021.

False(?) negatives:

False(?) positives:

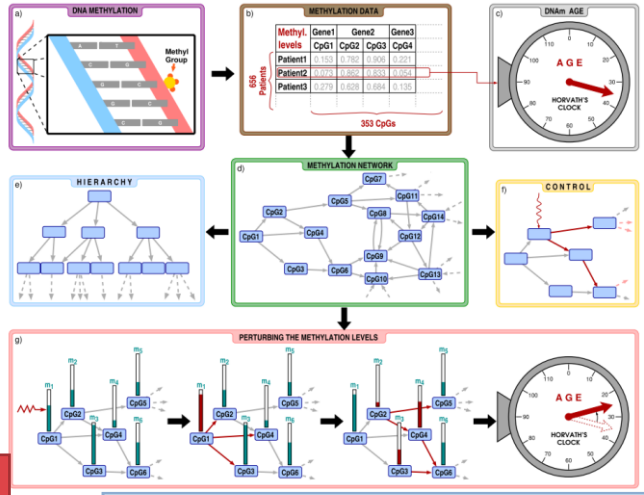
Space weather : whistler detection





B.Pataki, J.Lichtenberger et al. 2021

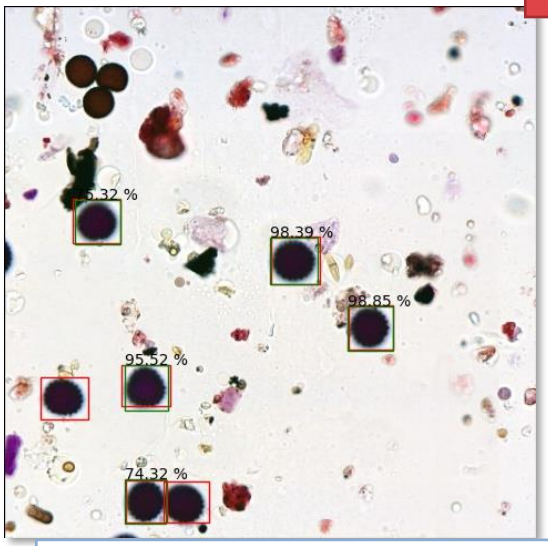
Understand (slow down?) aging



G. Palla et al. 2021.

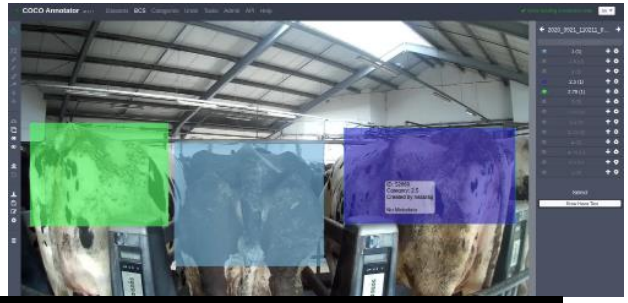
Empowering Sciences
Solving analytically
untraceable
hard inverse problems

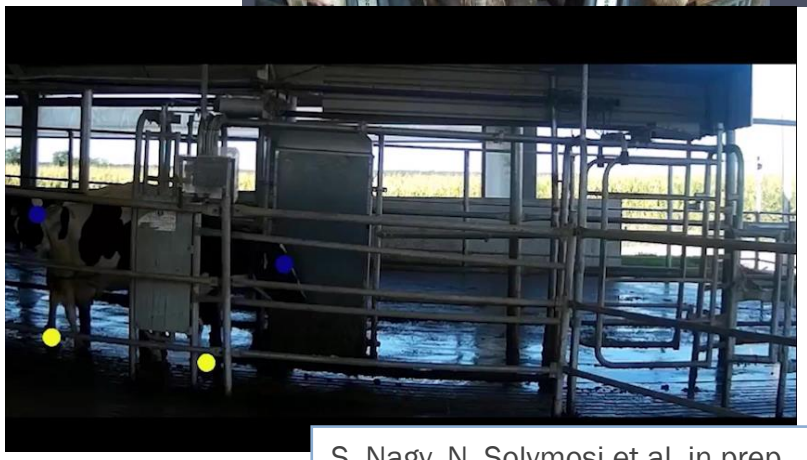
Pollen monitoring



A. Biricz et al. in prep.

Animal health

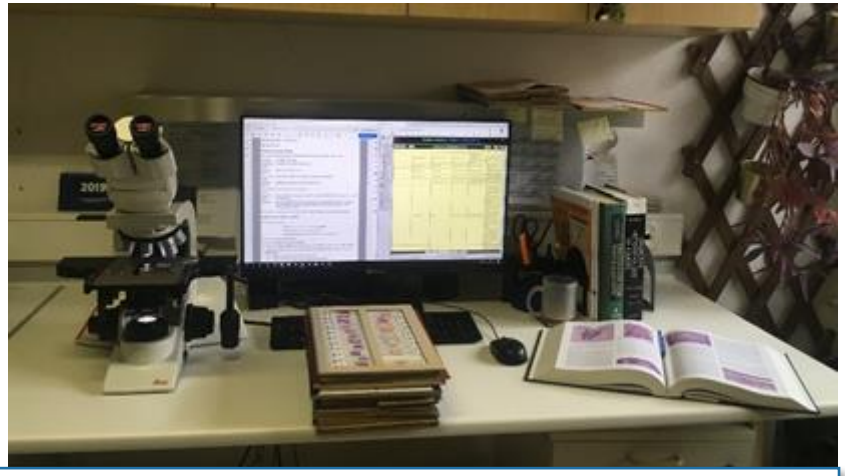




S. Nagy, N. Solymosi et al. in prep.

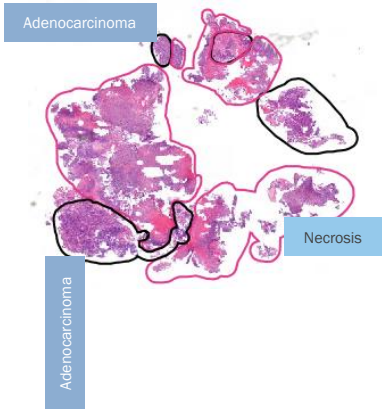
Deep learning for colorectal cancer pathology

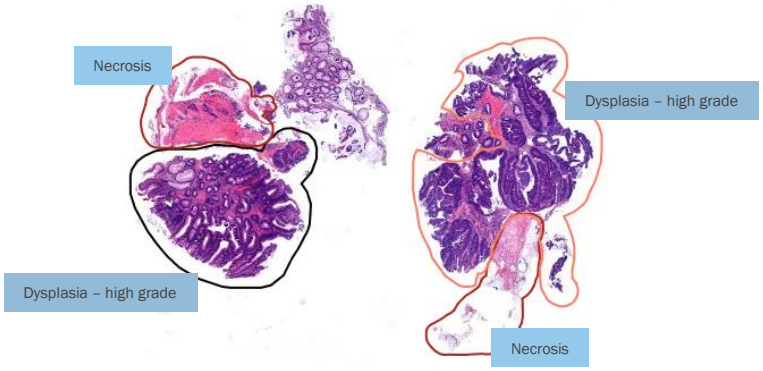
Large, well annotated training set is the key bottleneck for most machine learning tasks!



Collaboration: Semmelweis Univ. II. Path. Inst.

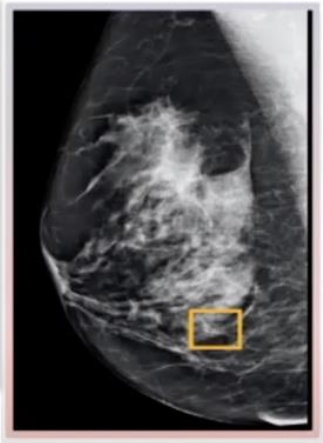
- Hungary has the highest colorectal cancer rate
- >10,000 new cases, **>5,000 death/yr** (Orv. Hetil., 2017)
- **Early detection**!
- Samples from colonoscopy, biopsy
- **Traditional**: by-eye microscope
- **New**: digital scanner
- **>2,000 whole slide images**
  - **80,000 x 60,000 pixels**, 15GB
- **Detailed annotation** by trained pathologists:
  - Dysplasia_low grade
  - Dysplasia_high grade
  - Adenocarcinoma
  - Suspicious for adenocarcinoma
  - Lymphovascular invasion
  - Tumornecrosis
  - Inflammation
  - Artifact

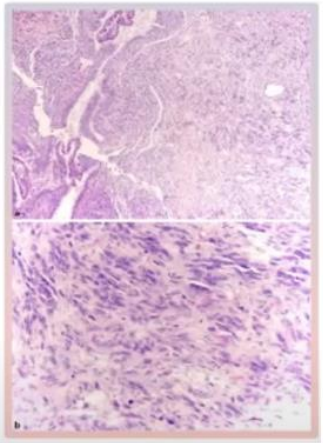




NIGHTINGALE OPEN SCIENCE HOME ABOUT LAUNCH OPEN LABS CONTACT EXPLORE OUR PLATFORM

Winners of the High Risk Breast Cancer Prediction Contest 1





We are pleased to announce the top 3 teams that submitted the best solutions in the contest and have won cash prizes totaling $10,000.

Winner, $5,000 cash prize

Team name: csabAIbio
Team members: András M. Biricz[1], Zsolt Bedőházi[1,2], Oz Kilim[1], István Csabai[1]
Organization: Eötvös Loránd University (ELTE), Budapest, 1117, Hungary
(1) Eötvös Loránd University (ELTE), Department of Complex Systems in Physics, Budapest, 1117, Hungary
(2) Eötvös Loránd University (ELTE), Doctoral School of Informatics, Budapest, 1117, Hungary

Second place, $3,000 cash prize

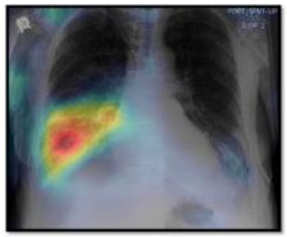
Team name/member: Bonaventure Dossou
Organization: McGill University, Mila Quebec AI Institute, Montreal, Quebec, Canada

Third place, $2,000 cash prize

Team name: PKU-Edinburgh
Team members: Yinghao Zhu[1], Junyi Gao[2,3], Xinze Li[1], Yifan He[1], Wenqing Wang[1], Liantao Ma[1]
Organizations: (1) Peking University, Beijing, China, (2) University of Edinburgh, Edinburgh, UK, (3) Health Data Research UK, UK

Covid CXR Hackathon

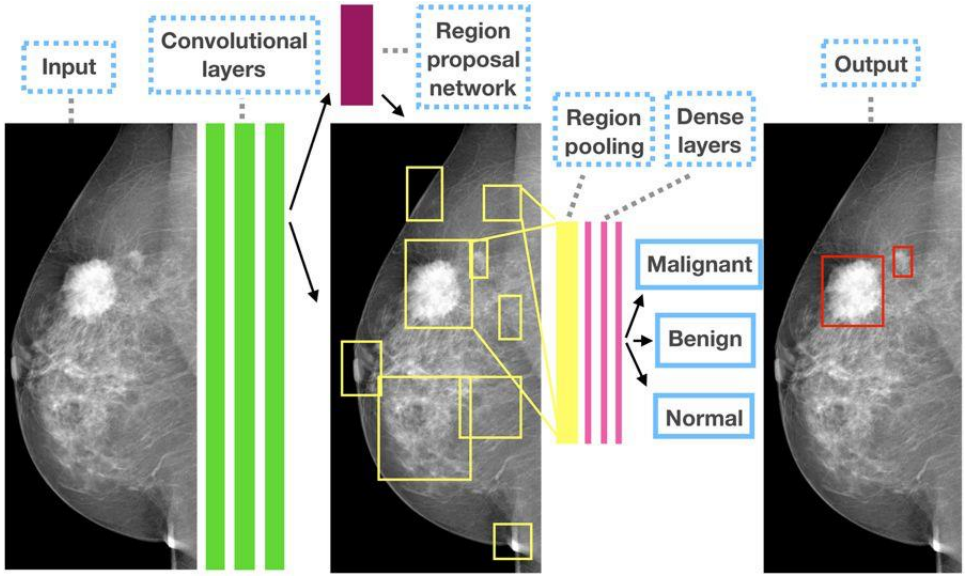
Artificial Intelligence for Covid-19 prognosis: aiming at accuracy and explainability



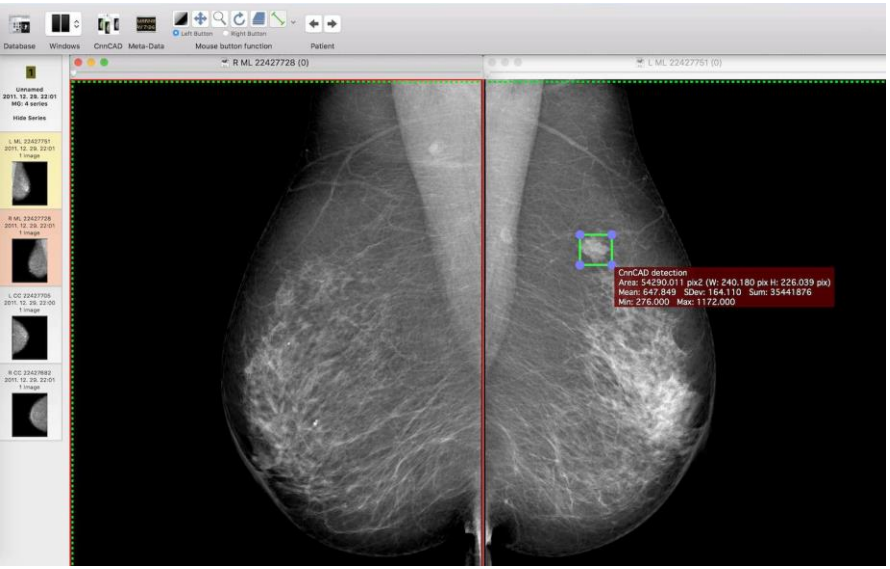
Thus **the Prize of 5.000,00€** kindly **offered by Bracco Imaging** for best result **goes to team csabAIbio from Hungary**. Congratulations to Alex Olar, András Biricz, Bendegúz Sulyok, and Zsolt Bedőházi! They also provide their software openly (code, checkpoints, report).

Mammography with deep learning (Faster R-CNN)

- Digital Mammography **DREAM challenge**
  - 1200 participants
  - Dezső Ribli, **best** final result
  - the only solution with localization
  - AUC = 0.95
- Publication: Nature Scientific Reports (2018)
  - 30-th most popular from 17000 articles
- New **collaborations with hospitals**, clinics
  - more training data
  - **open source** plugin
  - steps towards **licensing**



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



nature.com > scientific reports > articles > article

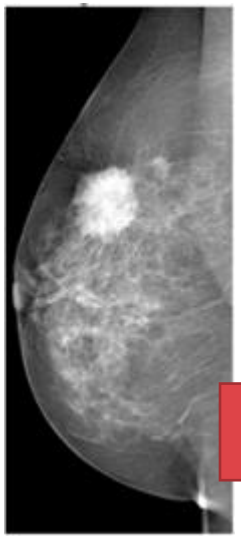
SCIENTIFIC REPORTS

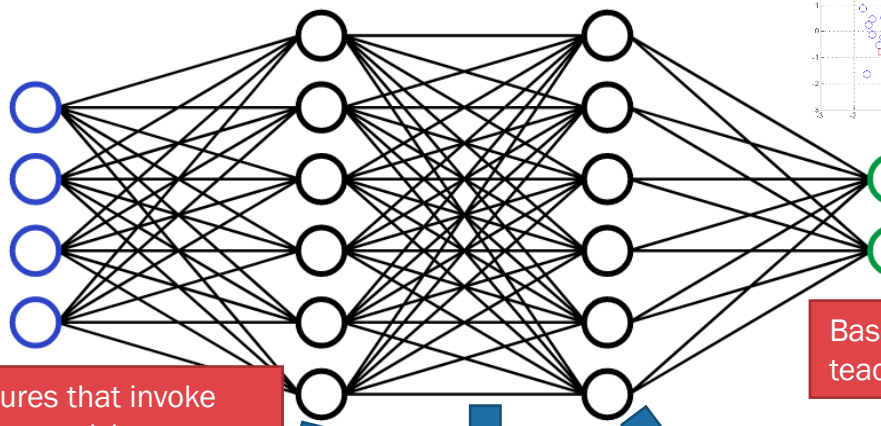
Detecting and classifying lesions in mammograms with Deep Learning

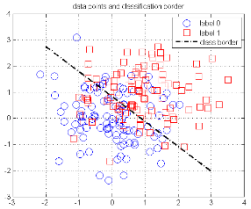
Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner & István Csabai

TOP 100 - READ ARTICLES - 2018 OFFICIAL AUTHOR SCIENTIFIC REPORTS

Explainable AI: automatic classification enhancement



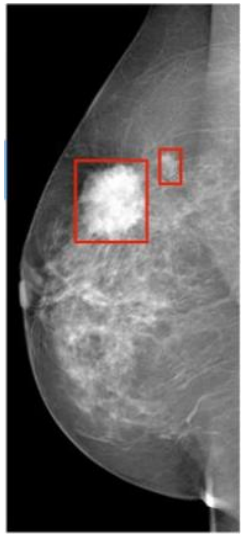




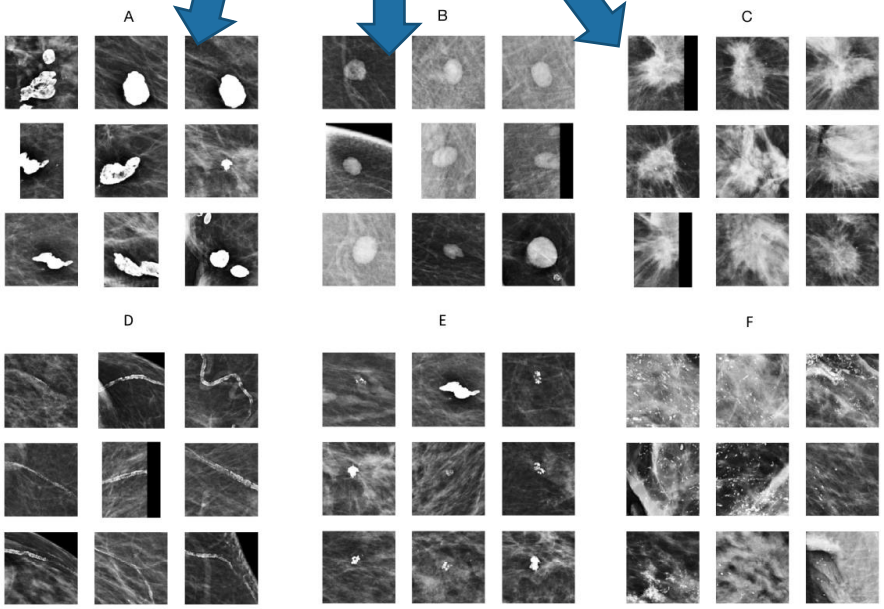
cancer

normal

Based on simple teaching classes

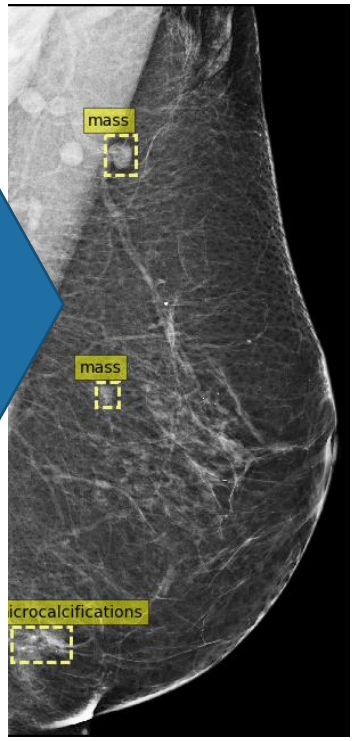


Features that invoke highest activity



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

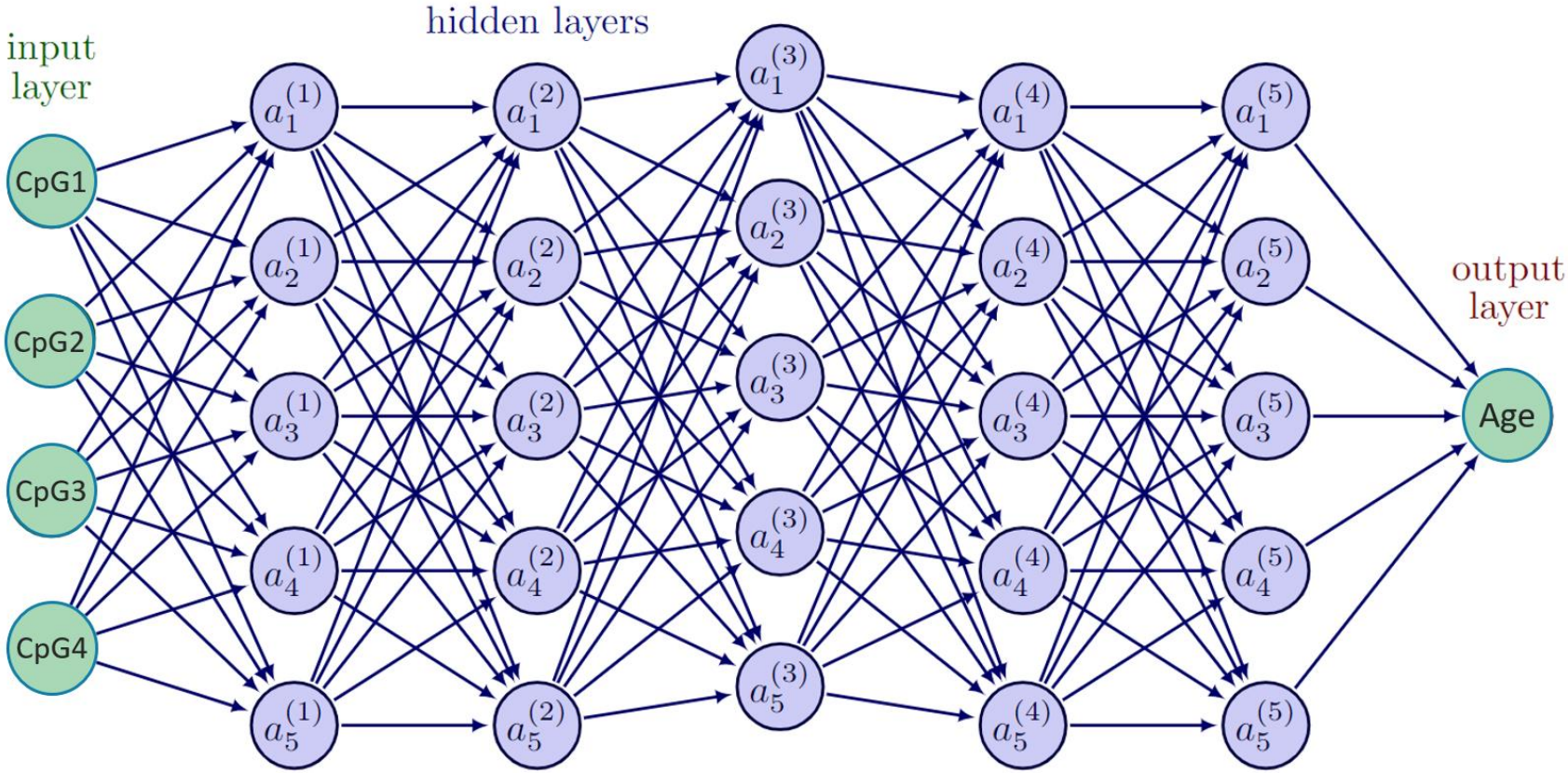
Automatic labels „discovered" by the network



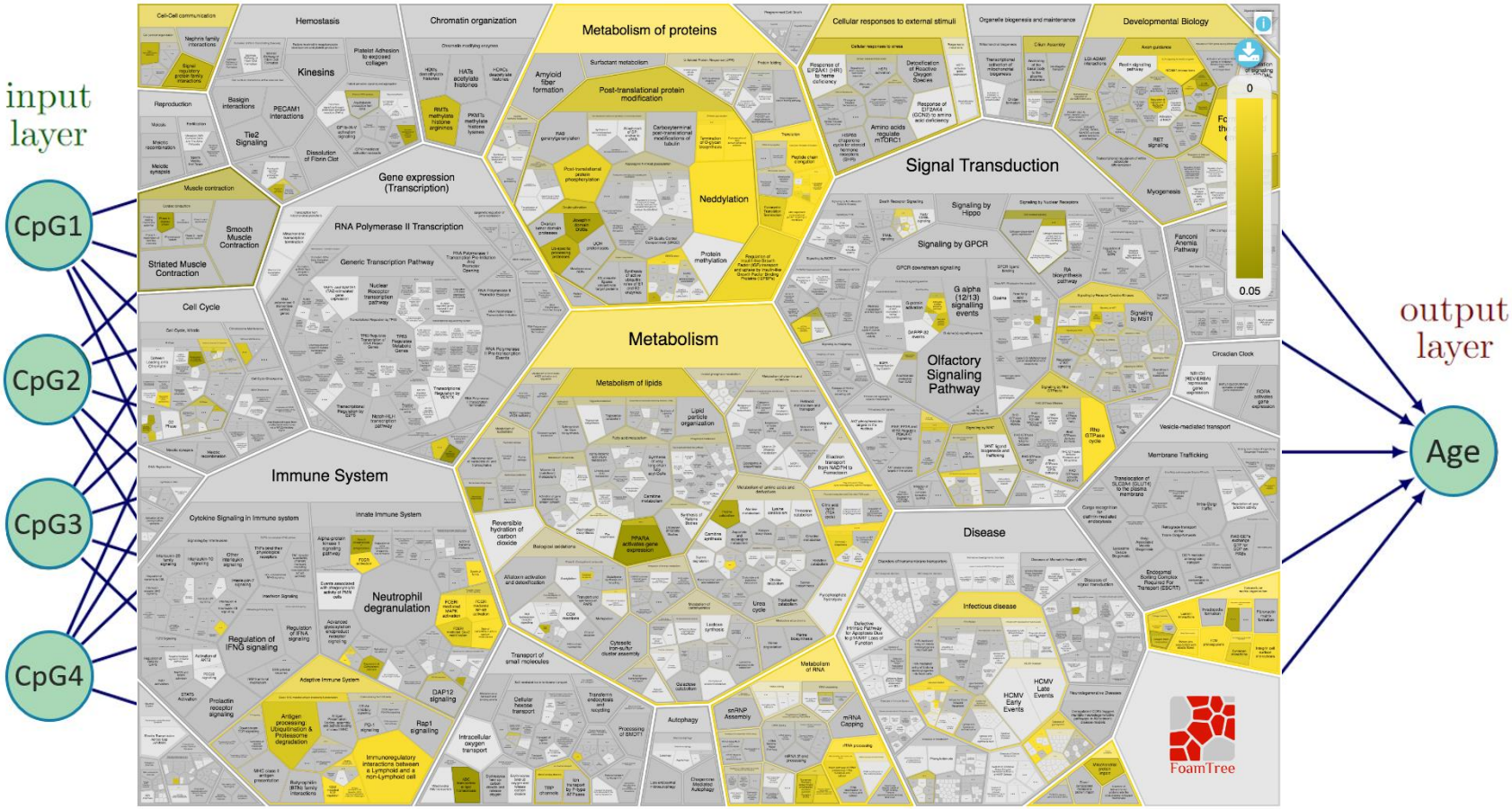
Interpretable, trustworthy, for radiologists

Ribli et al. in prep.

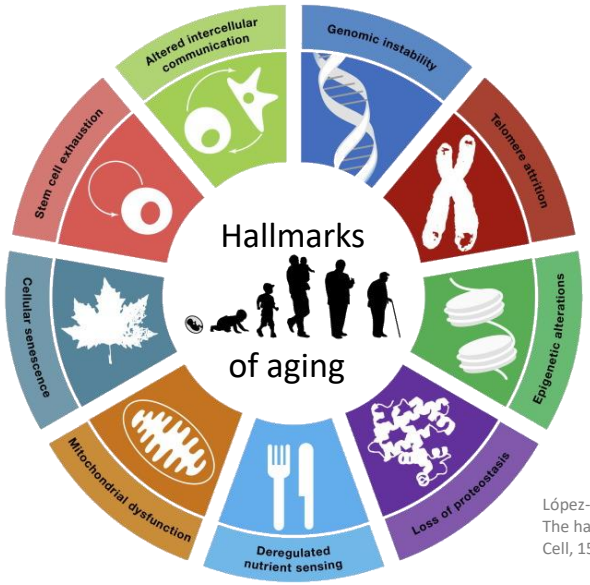
neural network



Biologically informed neural network



Aging: the number one risk factor for all diseases



López-Otín, C., et al., 2013.
The hallmarks of aging.
Cell, 153(6), pp.1194-1217.

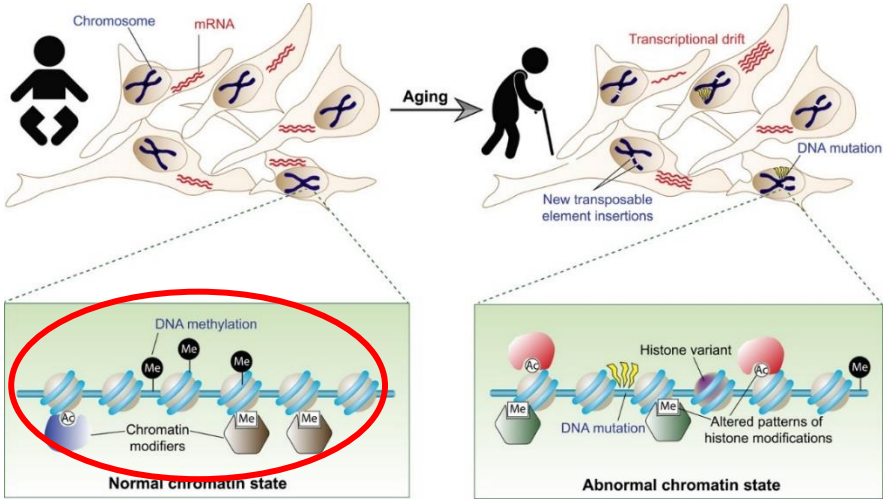
- genomic instability
- telomere attrition
- **epigenetic alterations**
- loss of proteostasis
- deregulated nutrient sensing
- mitochondrial dysfunction
- cellular senescence
- stem cell exhaustion
- altered intercellular communication

EDITORIAL | VOLUME 3, ISSUE 7, E448, JULY 01, 2022

Is ageing a disease?

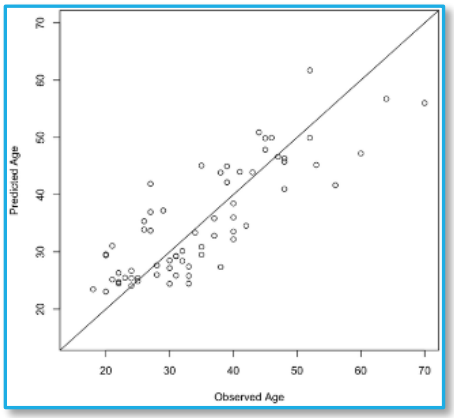
The Lancet Healthy Longevity

Aging research has experienced an unprecedented advance over recent years, particularly with the discovery that **the rate of aging is controlled**, at least to some extent, **by genetic pathways** and biochemical processes conserved in evolution. … **A major challenge is to dissect the interconnectedness between the candidate hallmarks and their relative contributions to aging**, with the final goal of identifying pharmaceutical targets to improve human health during aging, with minimal side effects.

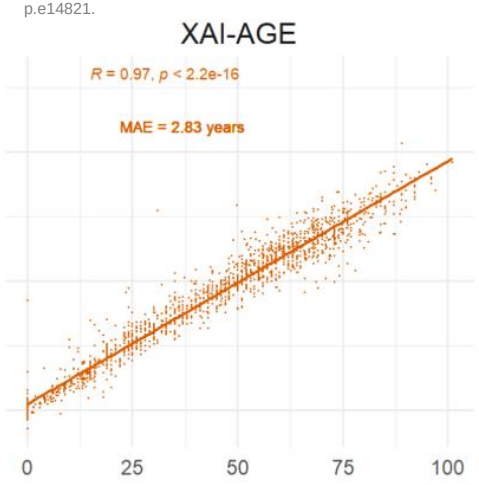


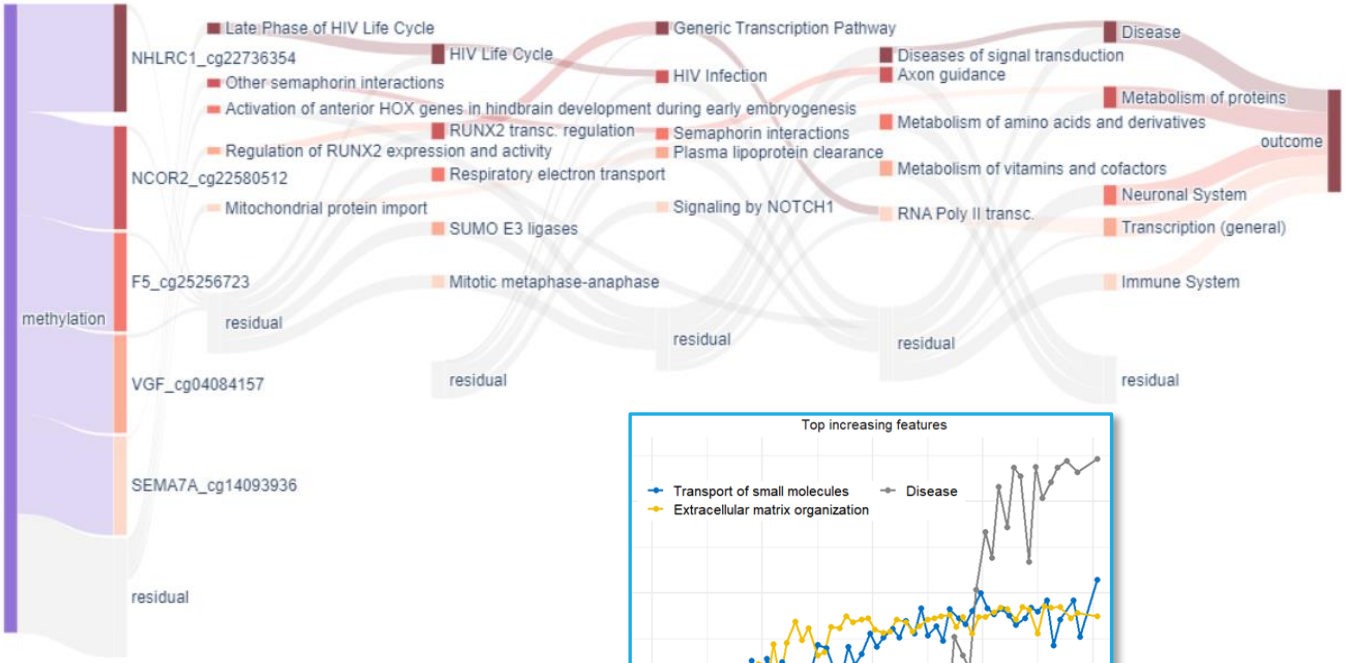
Pal, S. and Tyler, J.K., 2016. Epigenetics and aging. *Science advances*, *2*(7), p.e1600584.

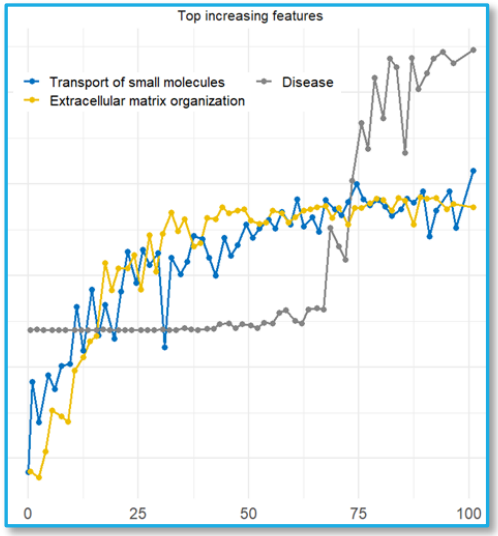
eXplainable AI for molecular mechanisms of aging



Bocklandt, S., …, J.S., Horvath, et al.., 2011. Epigenetic predictor of age. PloS one, 6(6), p.e14821.

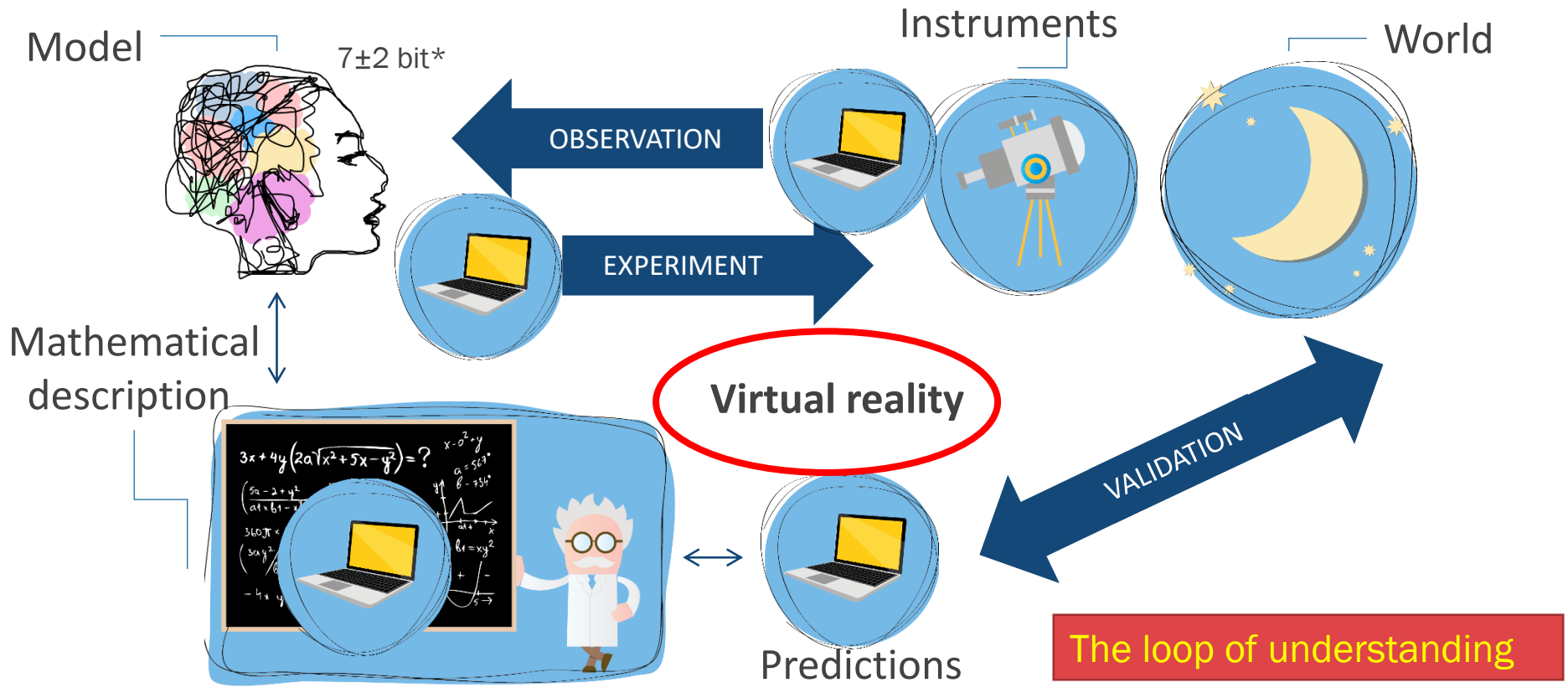






A. Prósz et al., *Biologically informed deep learning for **explainable** epigenetic clocks*, to be submitted, 2022

History of (machine) intelligence / (data) science



Initial values

"laws", equations

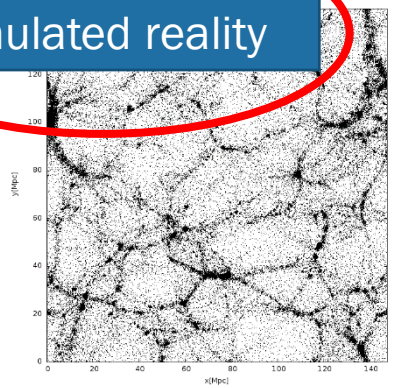
Simulated reality

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

$$F = ma$$

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

Λ=0.7
Ω_m=0.3

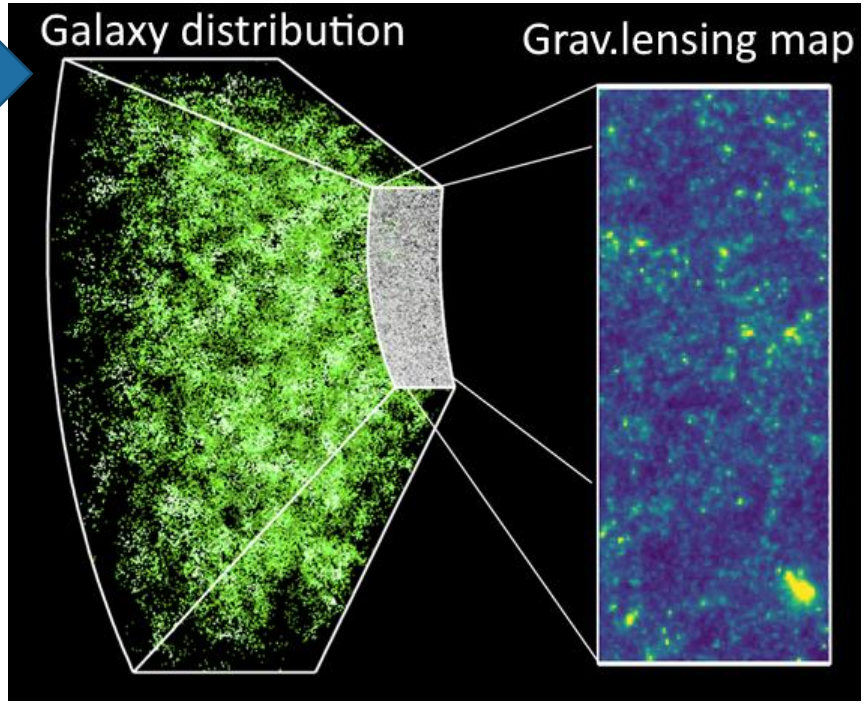


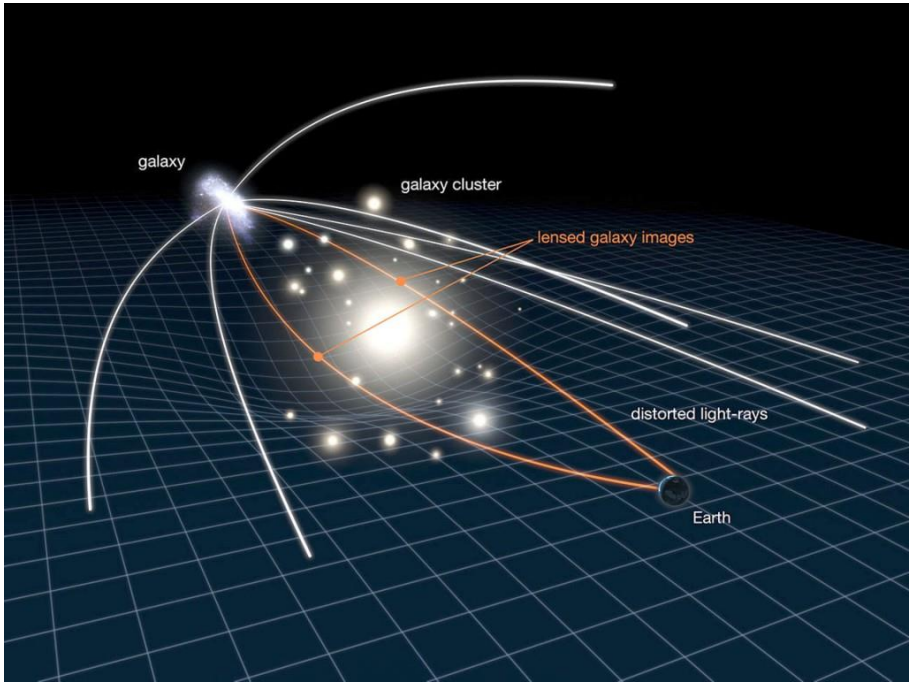
Learning from deep learning

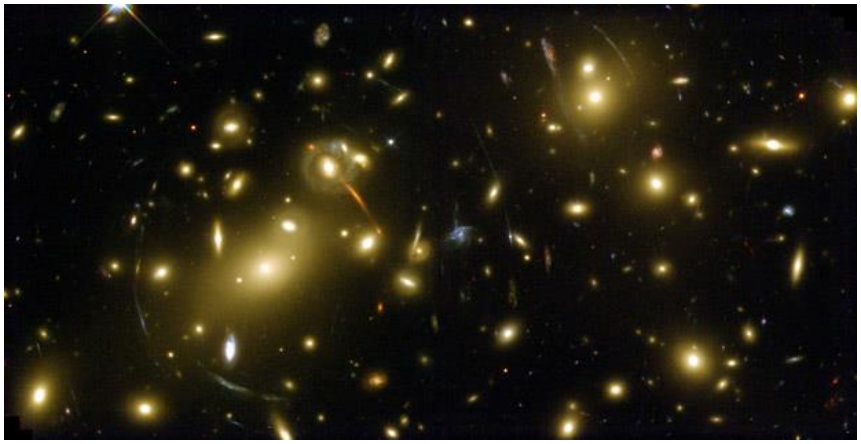
Cosmological parameters from gravitational lensing

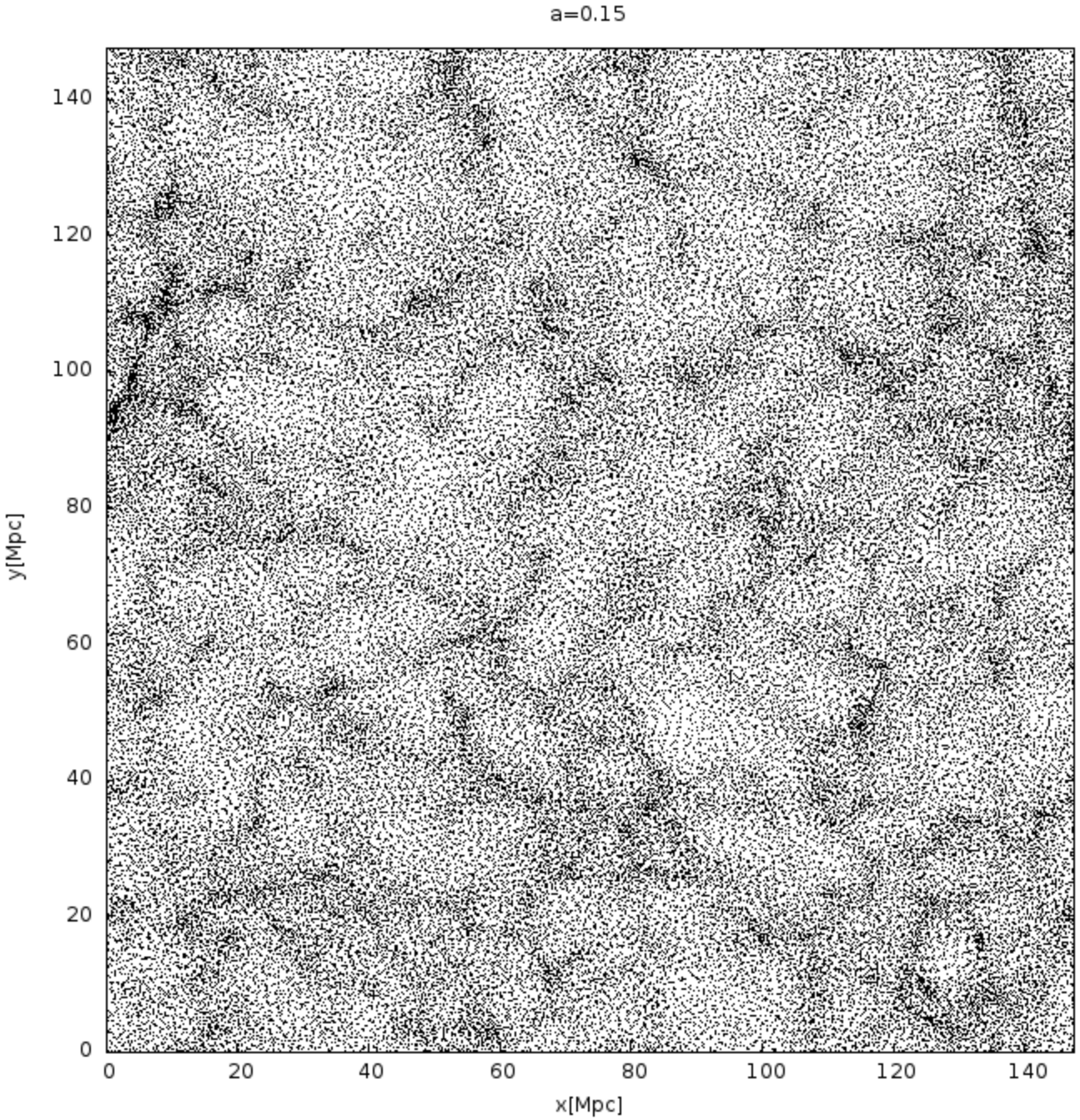
Cosmological parameters: Ω_m,σ_8

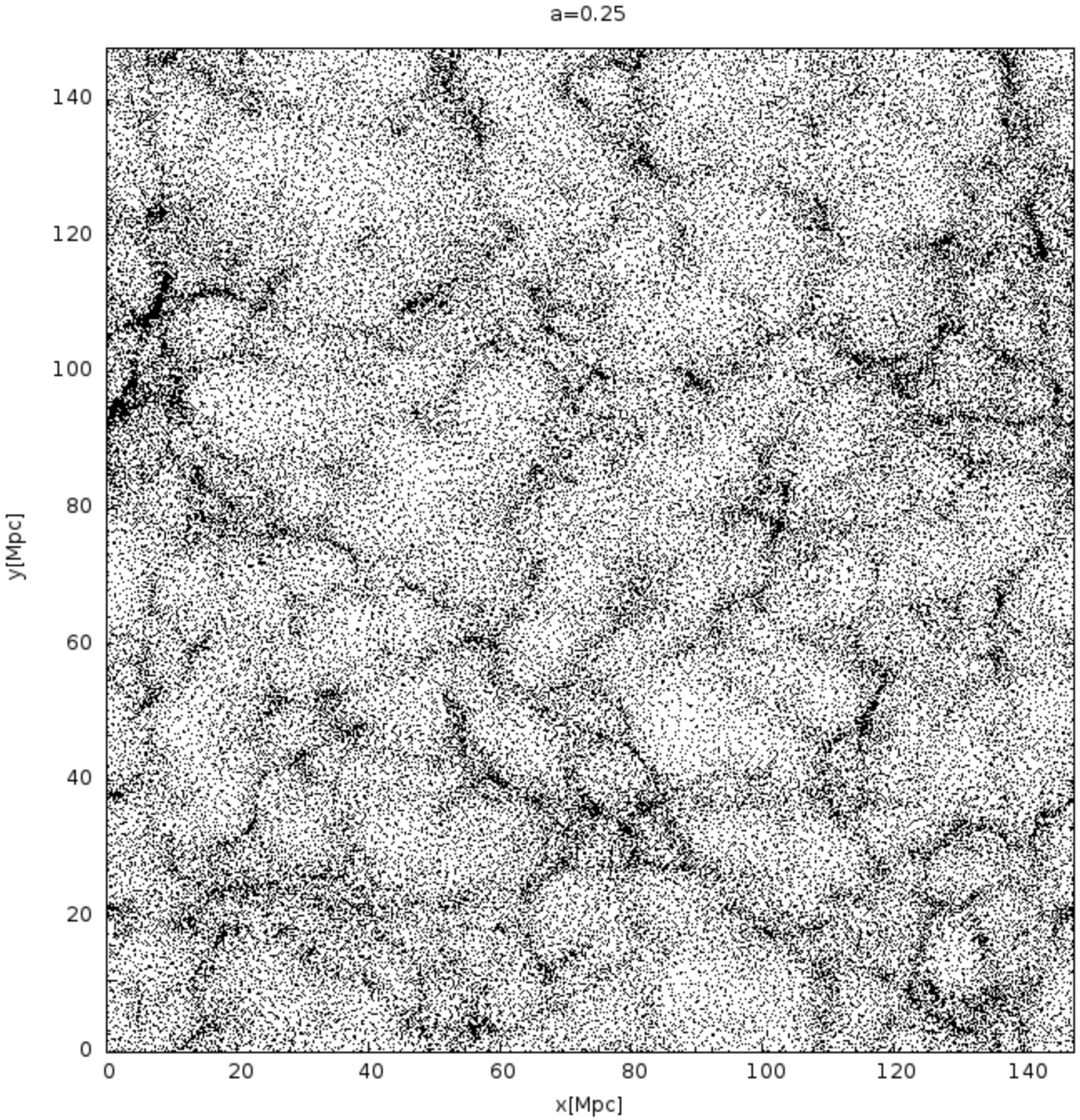
Complex physical processes

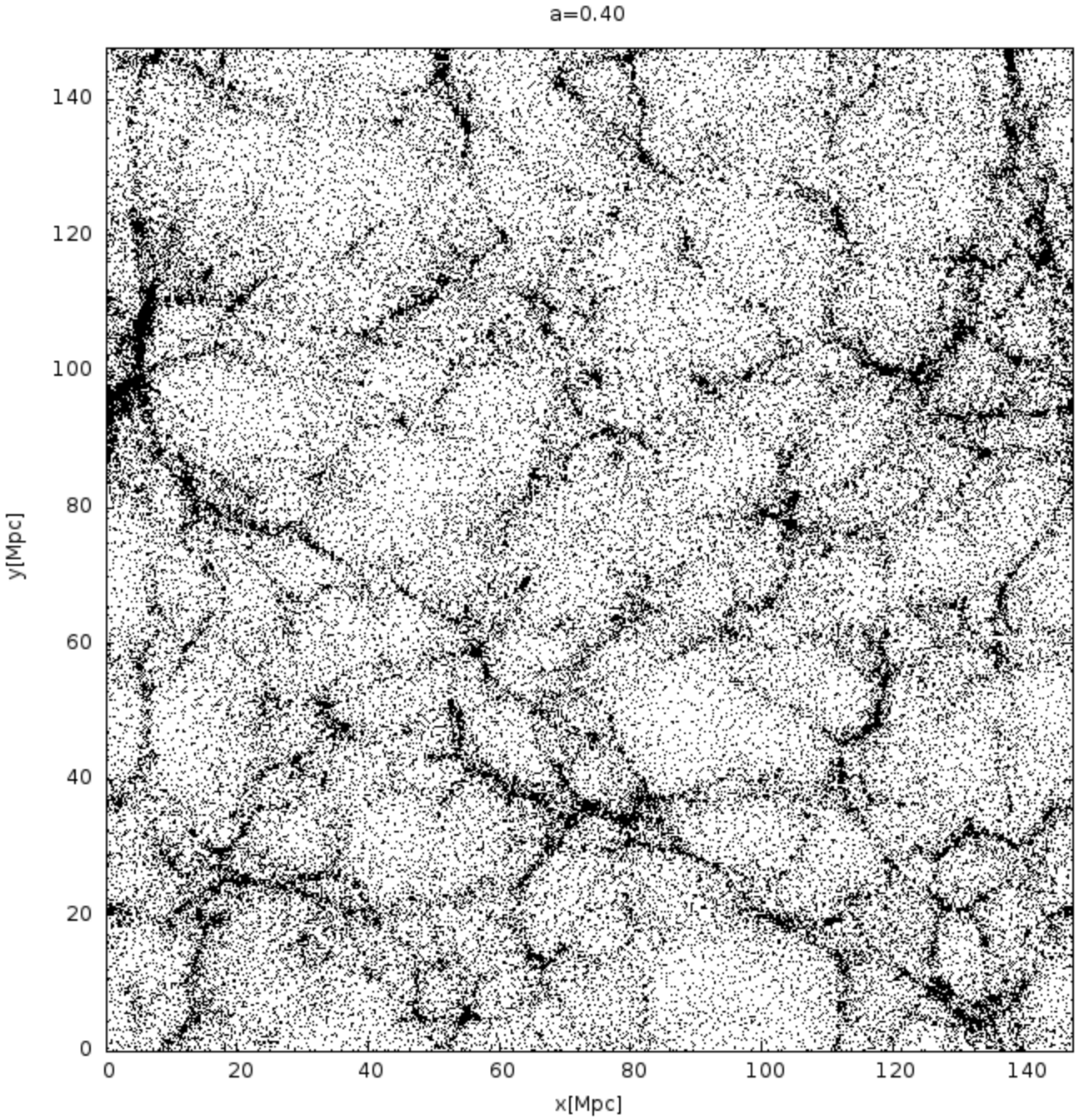


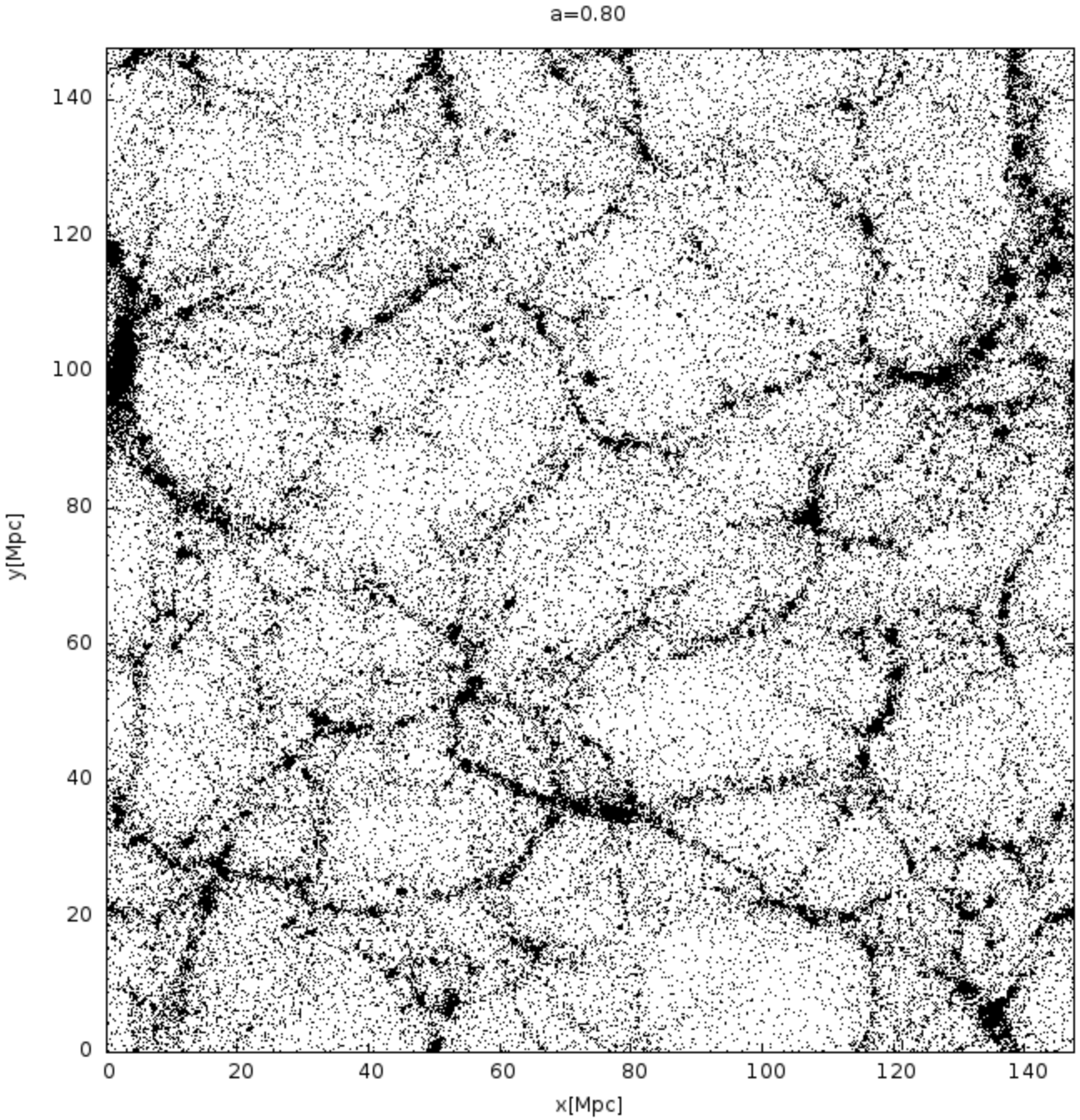


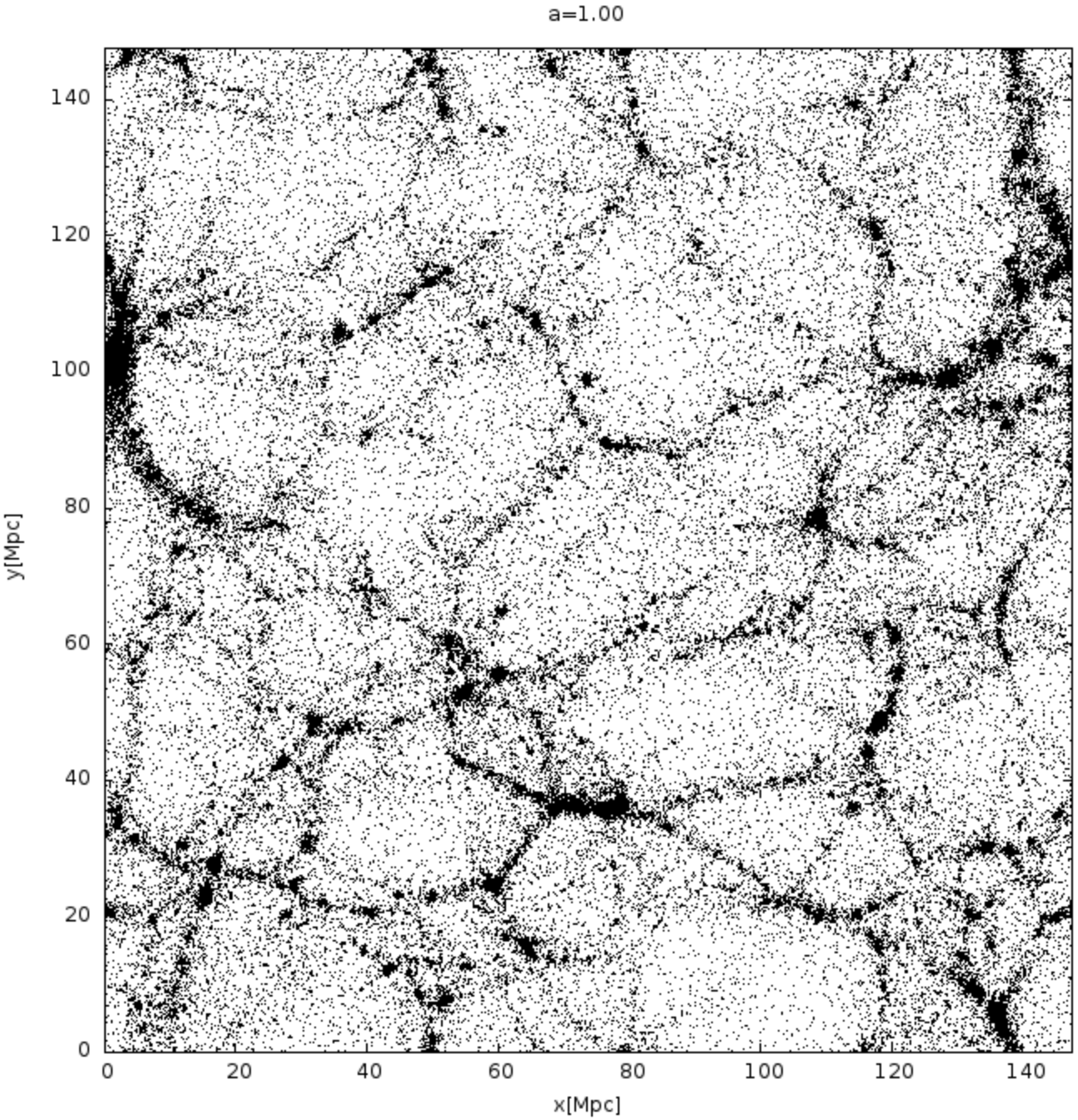



a=0.15
y[Mpc]
x[Mpc]
0
20
40
60
80
100
120
140


a=0.25
y[Mpc]
x[Mpc]
0
20
40
60
80
100
120
140

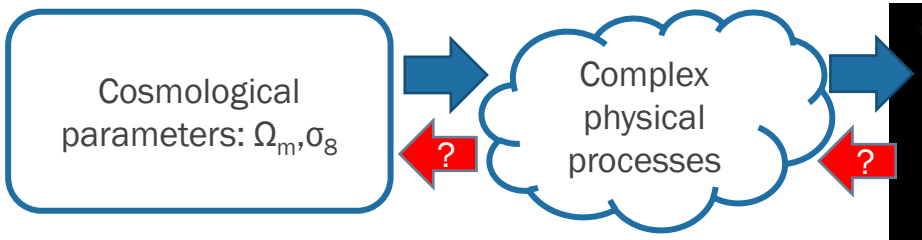

a=0.40
y[Mpc]
x[Mpc]
0
20
40
60
80
100
120
140


a=0.80
y[Mpc]
x[Mpc]
0
20
40
60
80
100
120
140

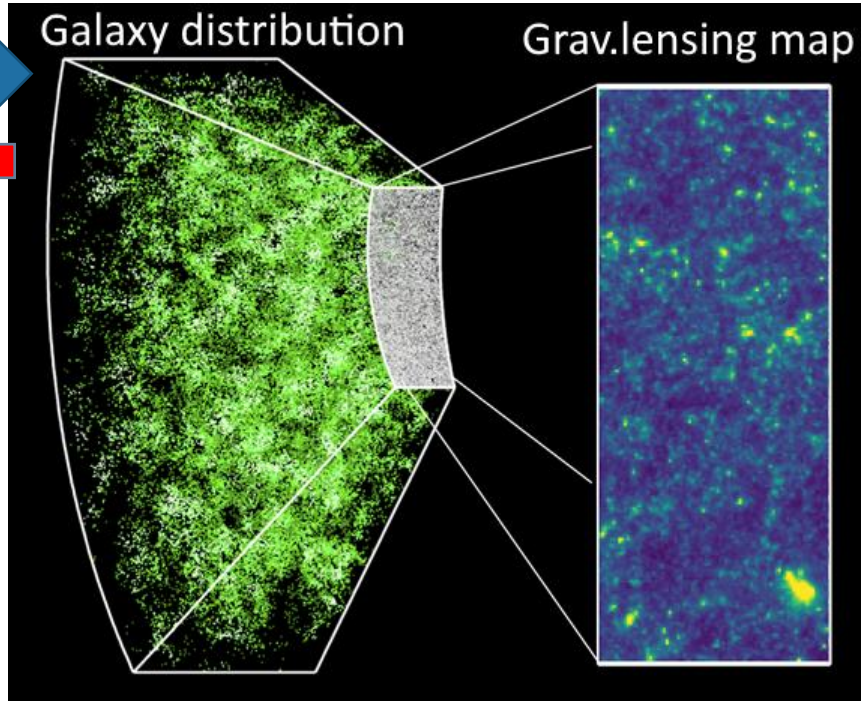

a=1.00
y[Mpc]
x[Mpc]
0
20
40
60
80
100
120
140

Learning from deep learning

Cosmological parameters from gravitational lensing



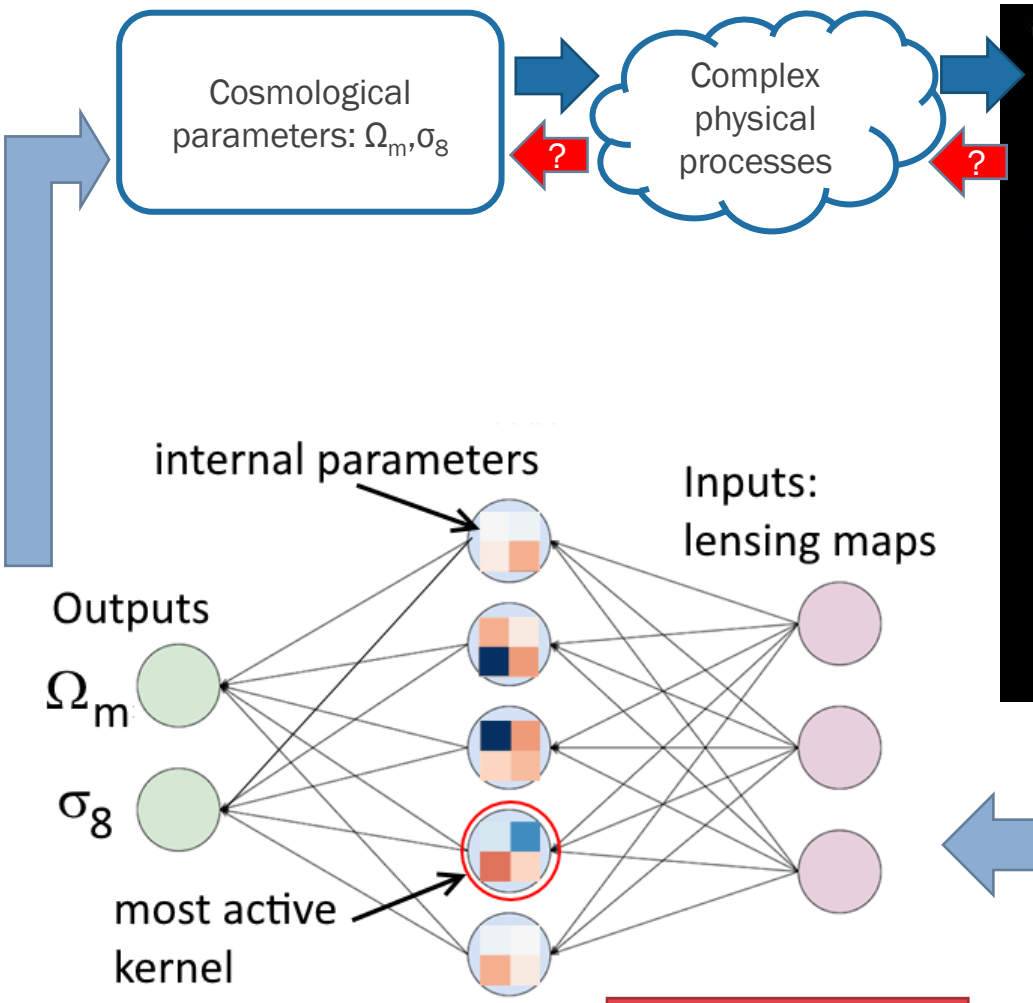
Hard inverse problem!

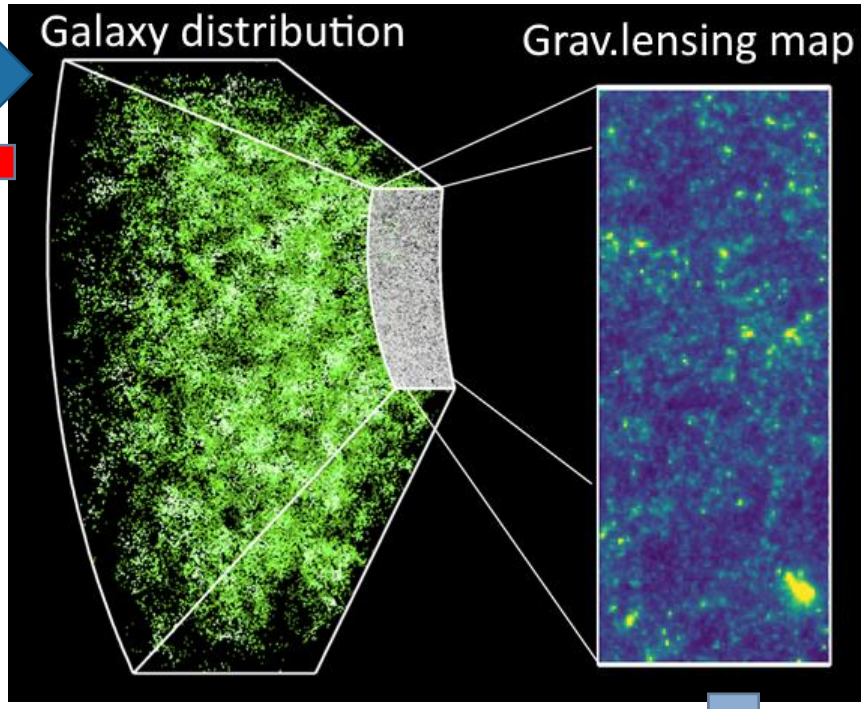


Ribli et al. MNRAS 2019
Ribli et al. Nature Astr. 2019

Learning from deep learning

Cosmological parameters from gravitational lensing





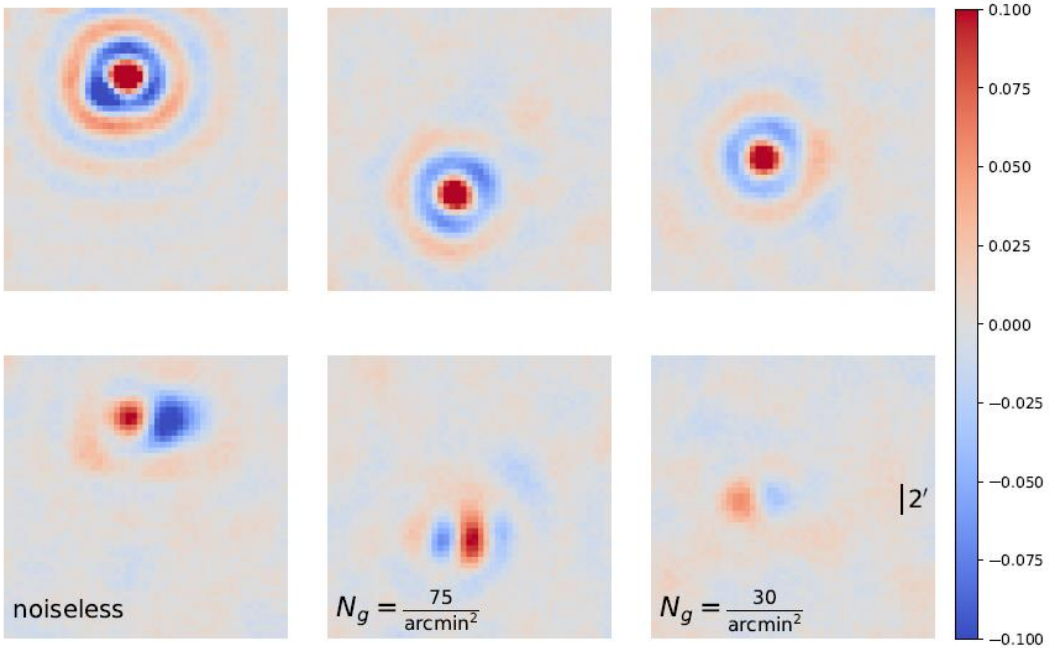
nature astronomy

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

Learned kernels: dark matter halo profile expansion



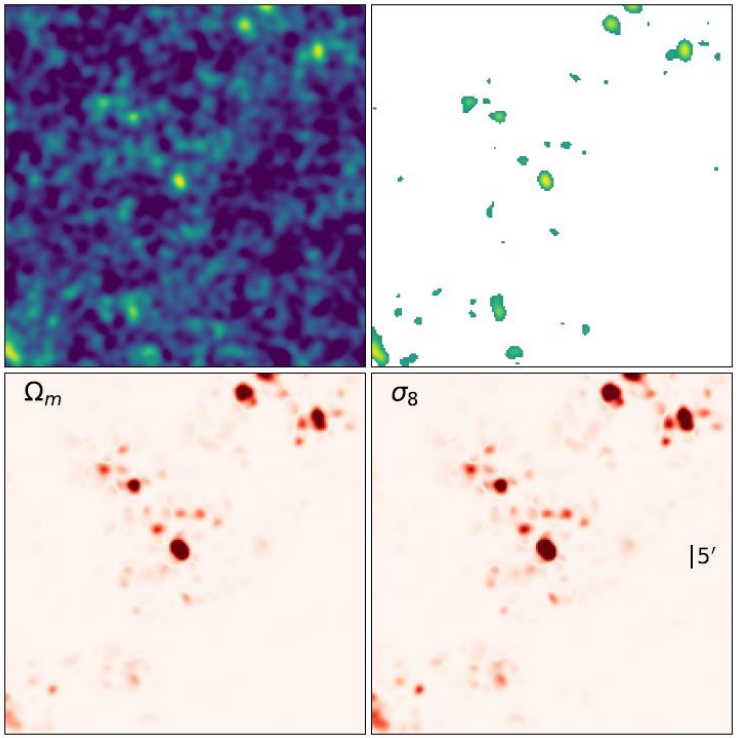
Information extraction:

Cosmologists:

Fourier power spectrum

Neural net:

Halo profiles



Roberts cross kernel

Attention focus of the network with Layer-wise Relevance Propagation method

Learning from deep learning

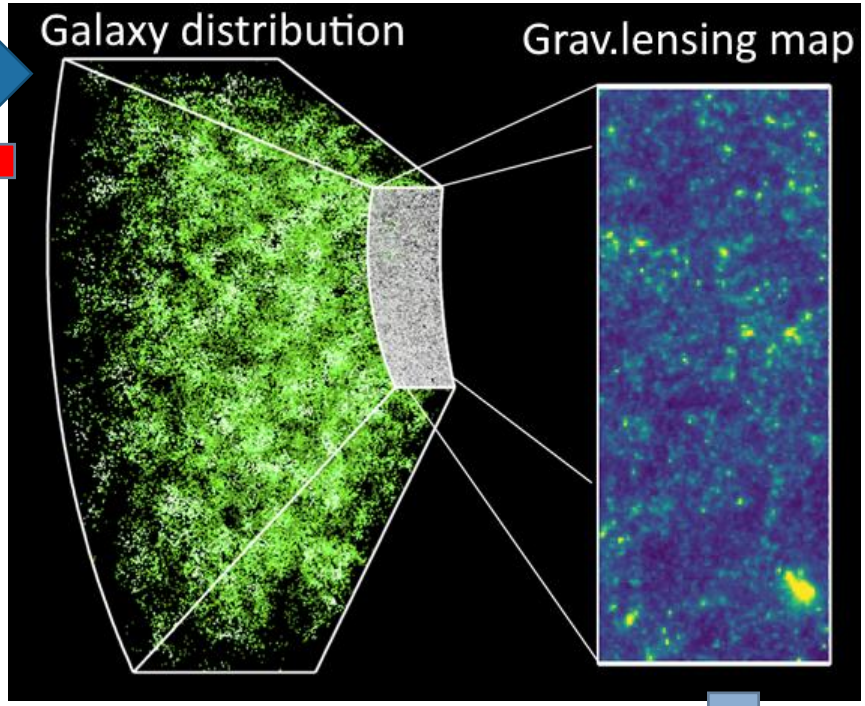
Cosmological parameters from gravitational lensing

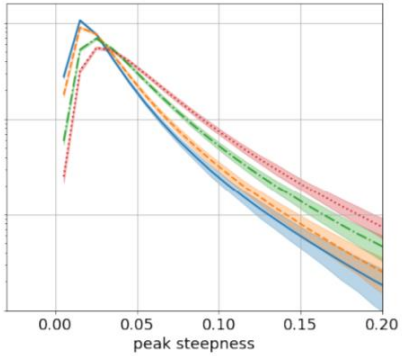
Cosmological parameters: Ω_m,σ_8

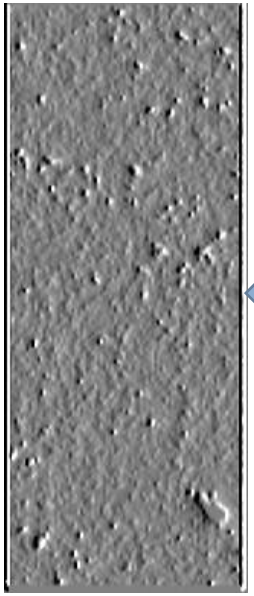
?

Complex physical processes

?



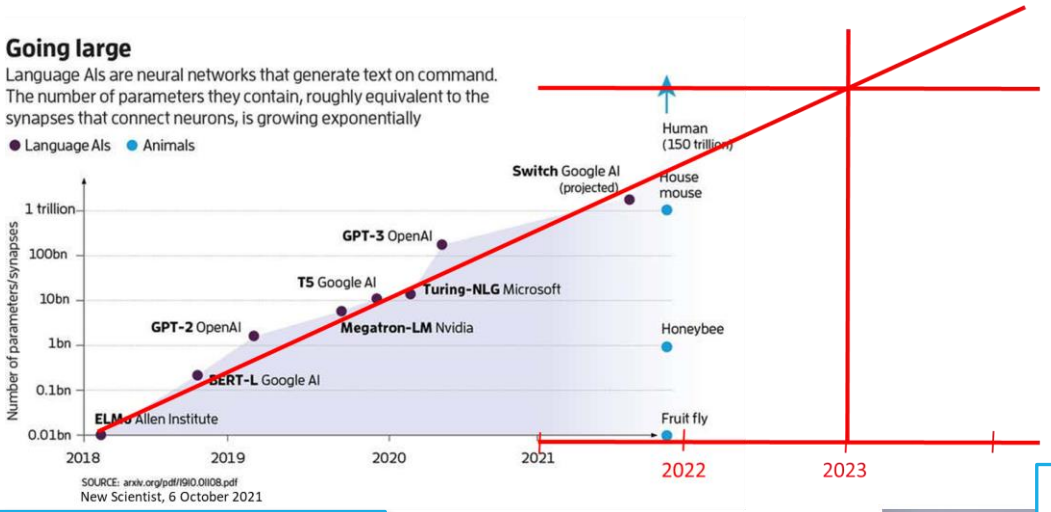




Roberts cross kernel

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

State of the Art



Question: A needle 35 mm long rests on a water surface at 20◦C. What force over and above the needle's weight is required to lift the needle from contact with the water surface?

Facebook Meta AI
Galactica

<work>

$$\sigma = 0.0728\ \mathrm{N/m}$$
$$\sigma = F/L$$
$$0.0728 = F/(2 \times 0.035)$$
$$F = 0.0728(2 \times 0.035)$$

calculate.py
```
f = 0.0728*(2*0.035)
with open("output.txt", "w") as file:
    file.write(str(round(f, 5)))
```
«run: "calculate.py">
</work>
Answer: $F = 0.0051$ N

ChatGPT, …

An astronaut Teddy bears A bowl of soup

mixing sparkling chemicals as mad scientists
shopping for groceries
working on new AI research

as a 1990s Saturday morning cartoon as digital art in a steampunk style

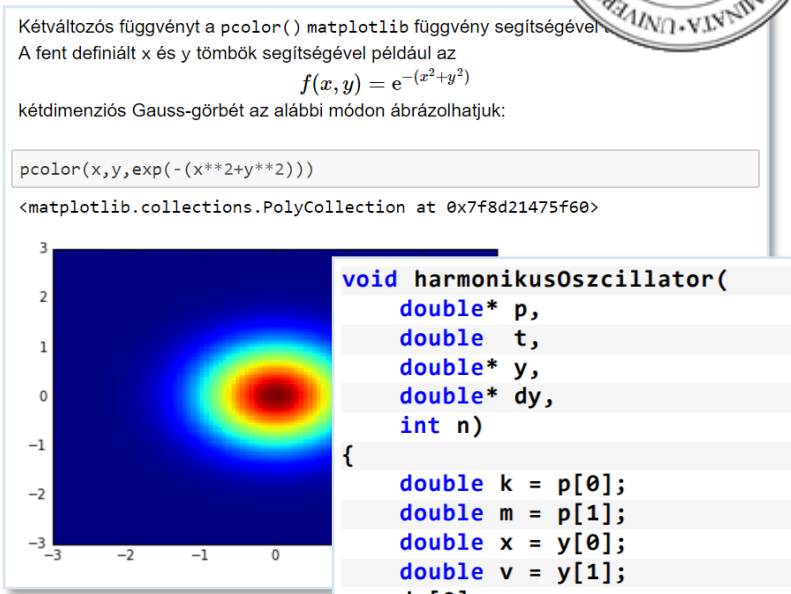
OpenAI **DALL-E 2**

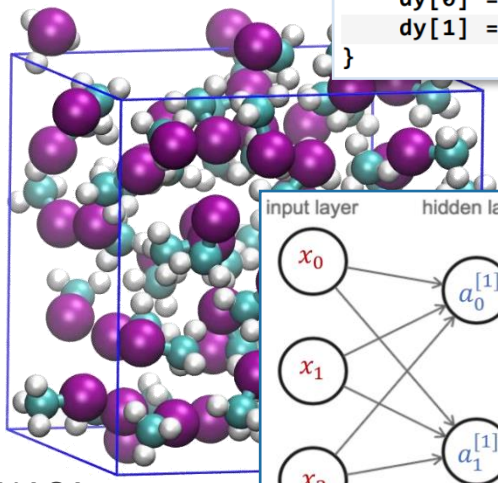
Google
Imagen

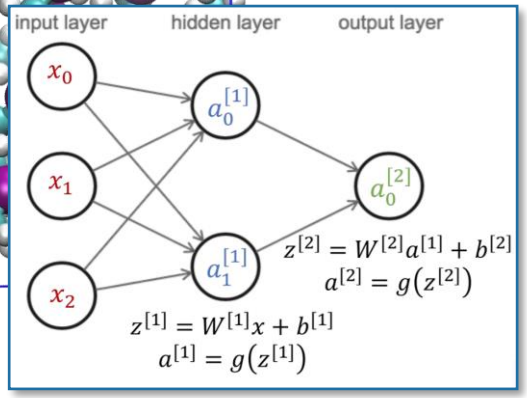
A cute corgi lives in a house made out of sushi.

Nem csak fizikát tanítunk!

- **BSc:**
  - Számítógépes alapismeretek
    (Linux, Python alapok, professzionális tudományos szövegszerkesztés, grafikonok)
    [http://oroszl.web.elte.hu/#teaching]
  - Programozási alapismeretek a fizikában
    (C programozás, adatkezelés, matematikai és fizikai feladatok megoldása programokkal)
  - Fizika numerikus módszerei I-II.
  - Számítógépes szimulációk, …
- **MSc:** Tudományos adatanalitika és modellezés specializáció
  [http://datascience.elte.hu]
  - Adatexploráció és vizualizáció
  - Haladó statisztika és modellezés
  - Adatmodellek és adatbázisok a tudományban
  - Adatbányászat és gépi tanulás
  - Számítógépes laboratórium
  - Adattudomány számítógépes laboratórium
  - Tudományos modellezés számítógépes laboratórium
  - **Deep learning a tudományokban (spec.)**
    [https://csabaibio.github.io/physdl/]
- Volt diákjaink **karrier**je
  - hazai és nemzetközi kutatóintézetek: Univ. Hawaii, JHU, NASA…
  - cégek: Ericsson, Semilab, Aimotive, Morgan-Stanley, …

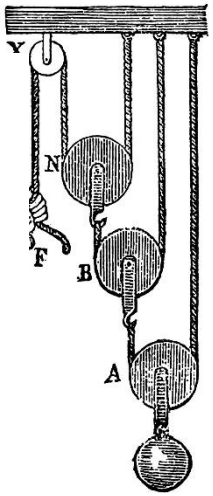






Any sufficiently advanced technology is indistinguishable from magic. /Arthur C. Clarke/

Astronomy → Mechanics → Quantummechanics → Solid state physics → Transistor/μelectronics



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. With the advancements of high throughput techniques we may be ready to tackle another frontier: **life** and **intelligence** at last, because it is the most sophisticated and complex. End of diseases, much **longer healthy life**, … ?

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

NEW PARADIGMS NEED NEW RESEARCHERS

EDUCATION: We need new scientist who have professional skills both in their disciplines and in modern information technologies.

István Csabai

ELTE Dept. of Physics of Complex Systems

csabai@elte.hu

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