HUNPROTEXC

DATA-INTENSIVE APPROACH IN SCIENCES

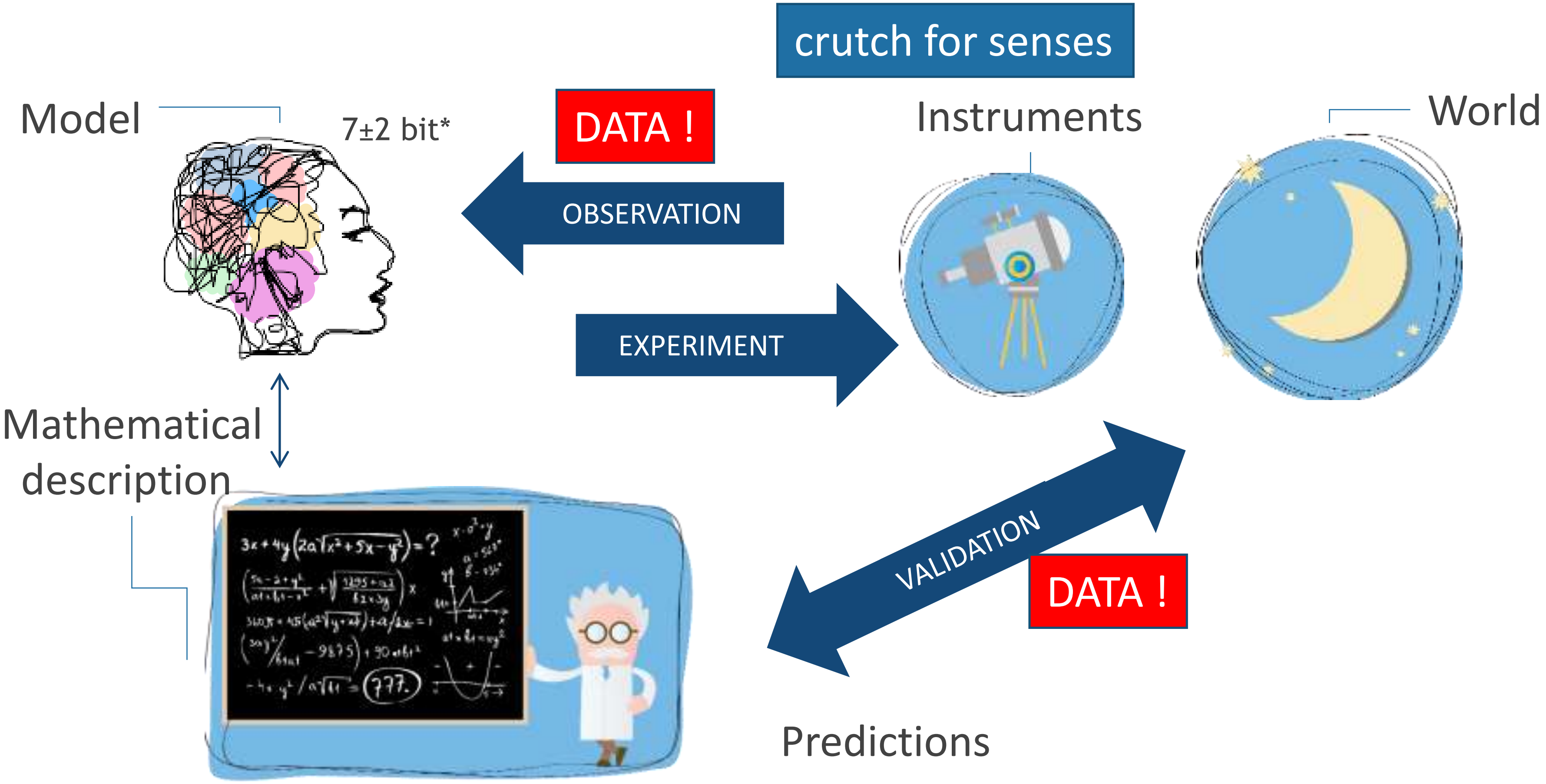
ISTVAN CSABAI
ELTE EÖTVÖS LORÁND UNIVERSITY
DEPT. OF PHYSICS OF COMPLEX SYSTEMS

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

Nemzeti Kutatási, Fejlesztési És Innovációs Hivatal

AZ NKFI ALAPBÓL MEGVALÓSULÓ PROJEKT

(Data) science



crutch for senses

Model

7±2 bit*

DATA !

Instruments

World

OBSERVATION

EXPERIMENT

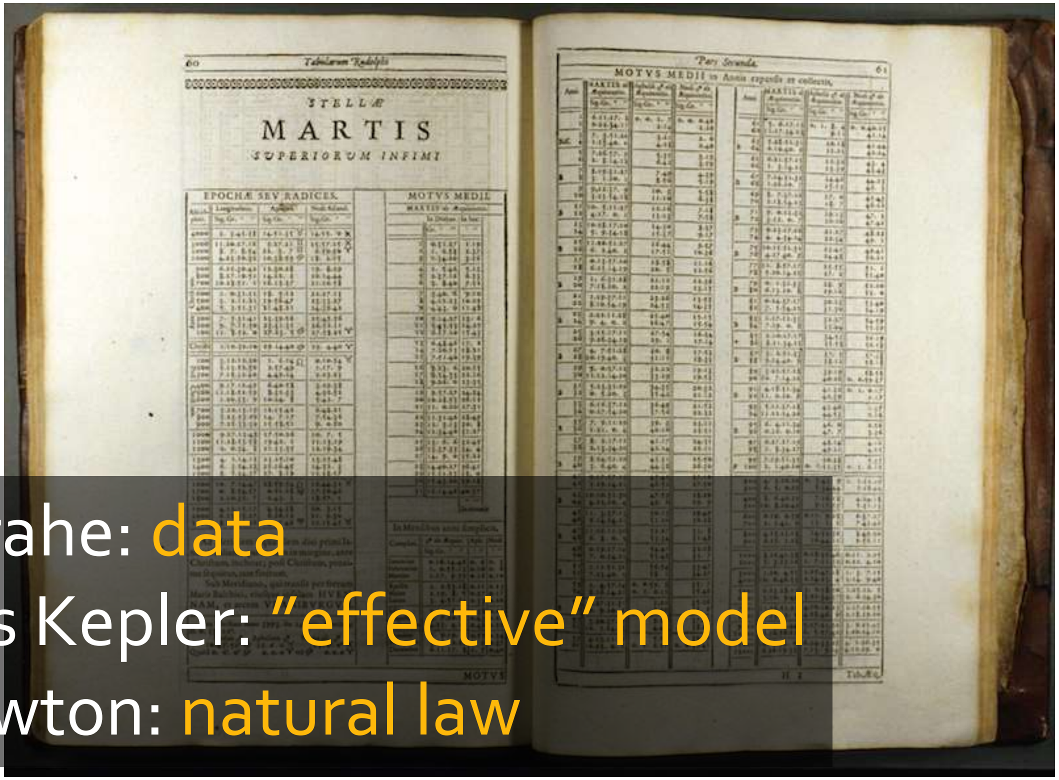
Mathematical description

VALIDATION

DATA !

Predictions

First "Data Science"



Tabularum Rudolphi

STELLÆ

MARTIS

SUPERIORUM INFIMI

EPOCHÆ SEV RADICES.

MOTVS MEDII

Pars Secunda.

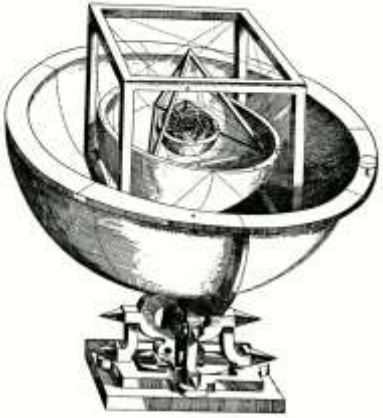
MOTVS MEDII

Tycho Brahe: data

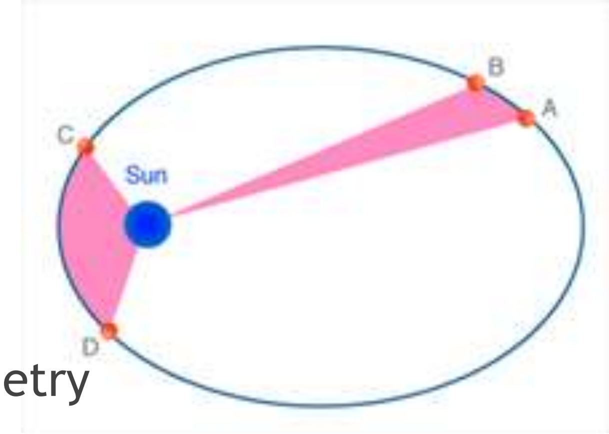
Johannes Kepler: "effective" model

Isaac Newton: natural law

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



Perfect beauty and symmetry



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

First "Data Science" in genetics

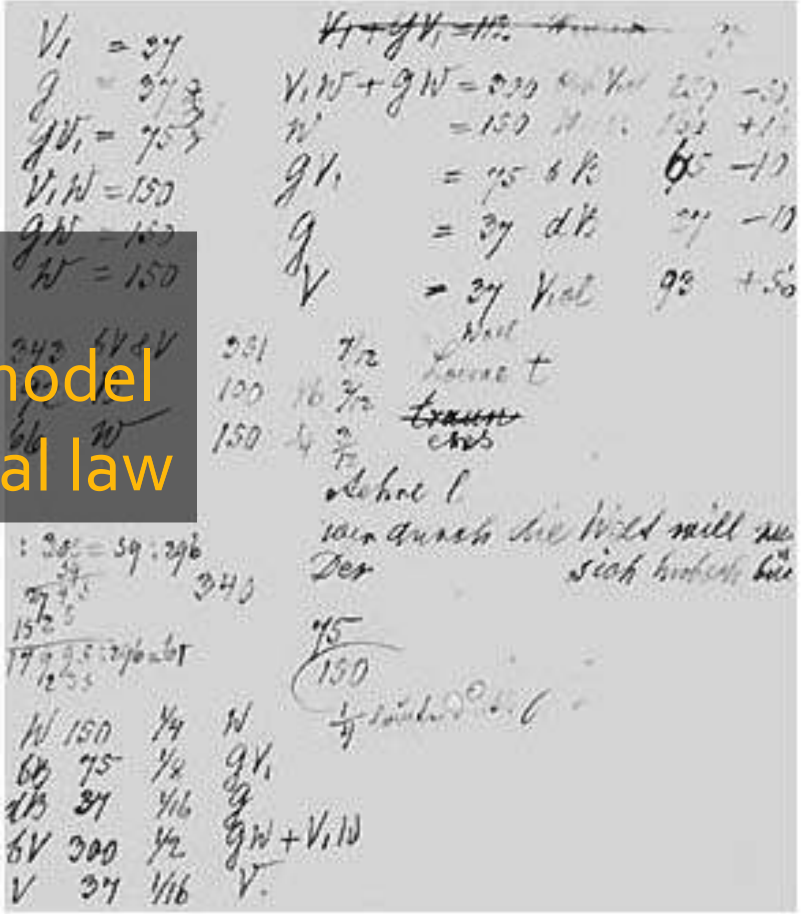
Mendel: data

Darwin: "effective" model

Griffith, Avery: natural law

Gregor Mendel, 1865

8 years, ~28.000 pea plants



$V_1 = 37$

$g = 37$

$gV_1 = 75$

$V_1W = 150$

$gW = 150$

$W = 150$

~~$V_1 + gV_1 = 112$~~

$V_1W + gW = 300$

$W = 150$

$gV_1 = 75$

$g = 37$

$V = 37$

Der

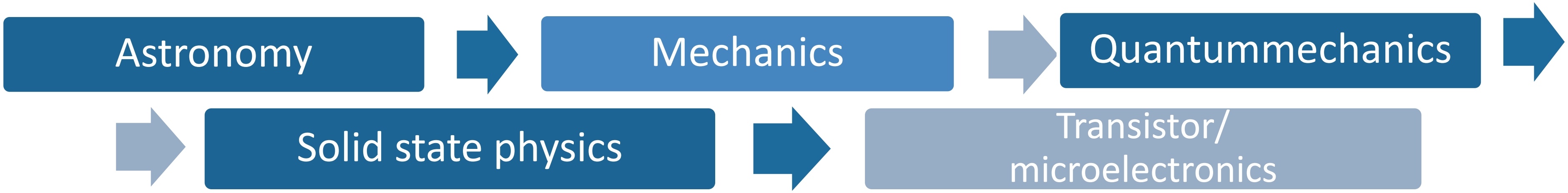
sich

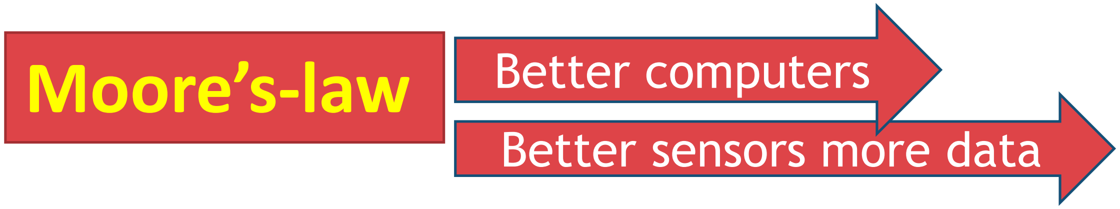
75/150

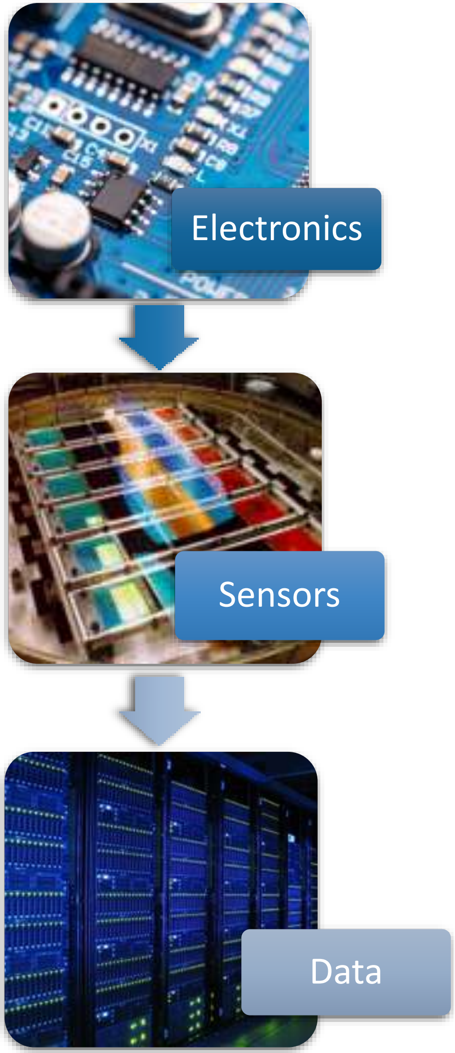
W	150	1/4	W
6b	75	1/4	gV_1
dB	37	1/16	g
bV	300	1/2	$gW + V_1W$
V	37	1/16	V.

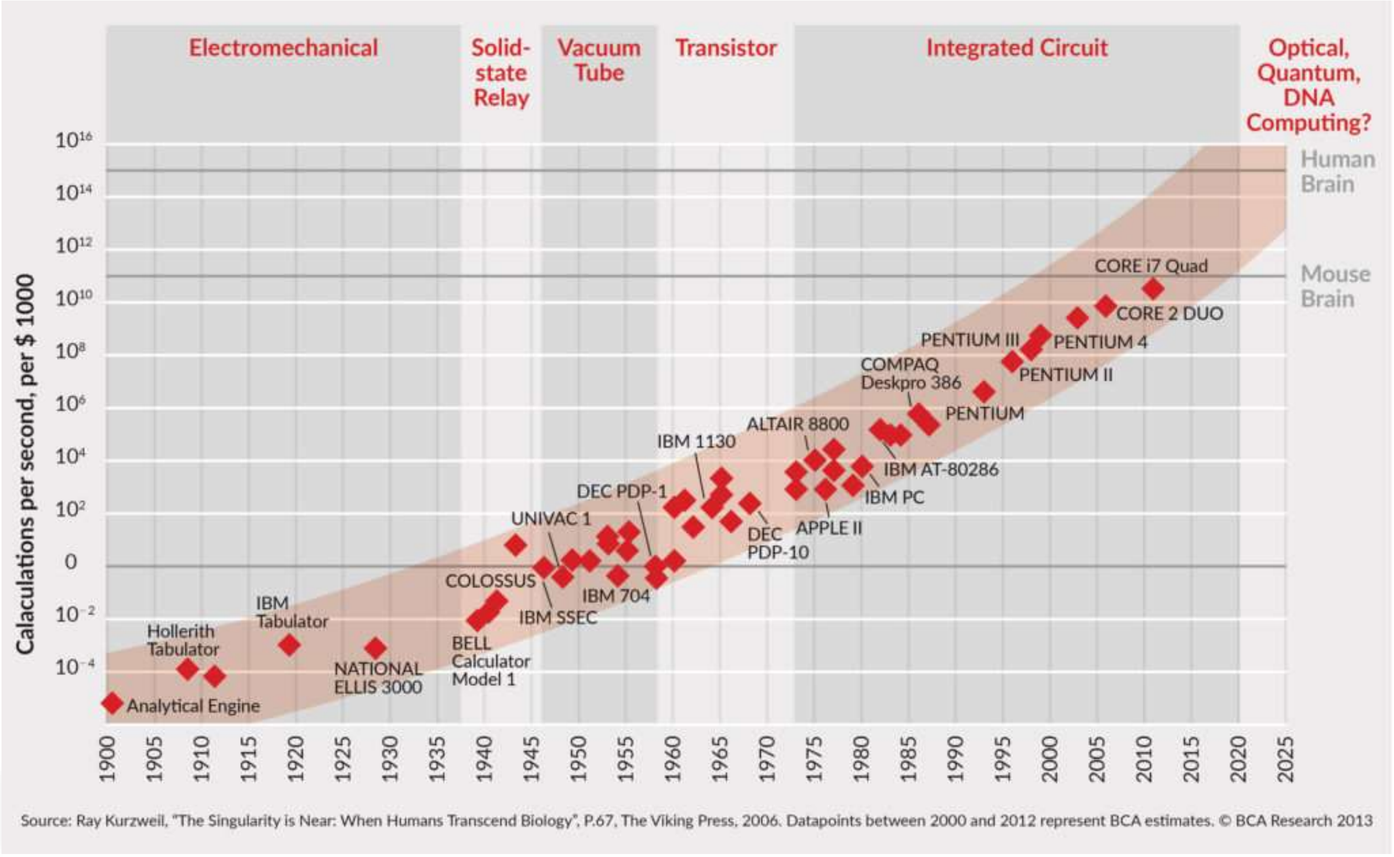
Courtesy of the Mendelianum, Moravian Museum, Brno.
Noncommercial, educational use only

Science – technology – science – technology …









Source: Ray Kurzweil, "The Singularity is Near: When Humans Transcend Biology", P.67, The Viking Press, 2006. Datapoints between 2000 and 2012 represent BCA estimates. © BCA Research 2013

Natural intelligence

7±2 bit

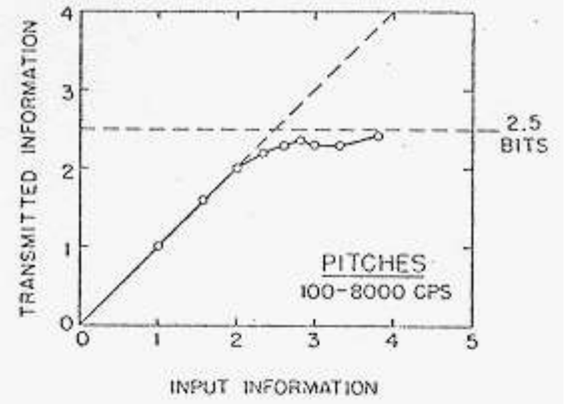


Fig. 1. Data from Pollack (17, 18) on the amount of information that is transmitted by listeners who make absolute judgments of auditory pitch. As the amount of input information is increased by increasing from 2 to 14 the number of different pitches to be judged, the amount of transmitted information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

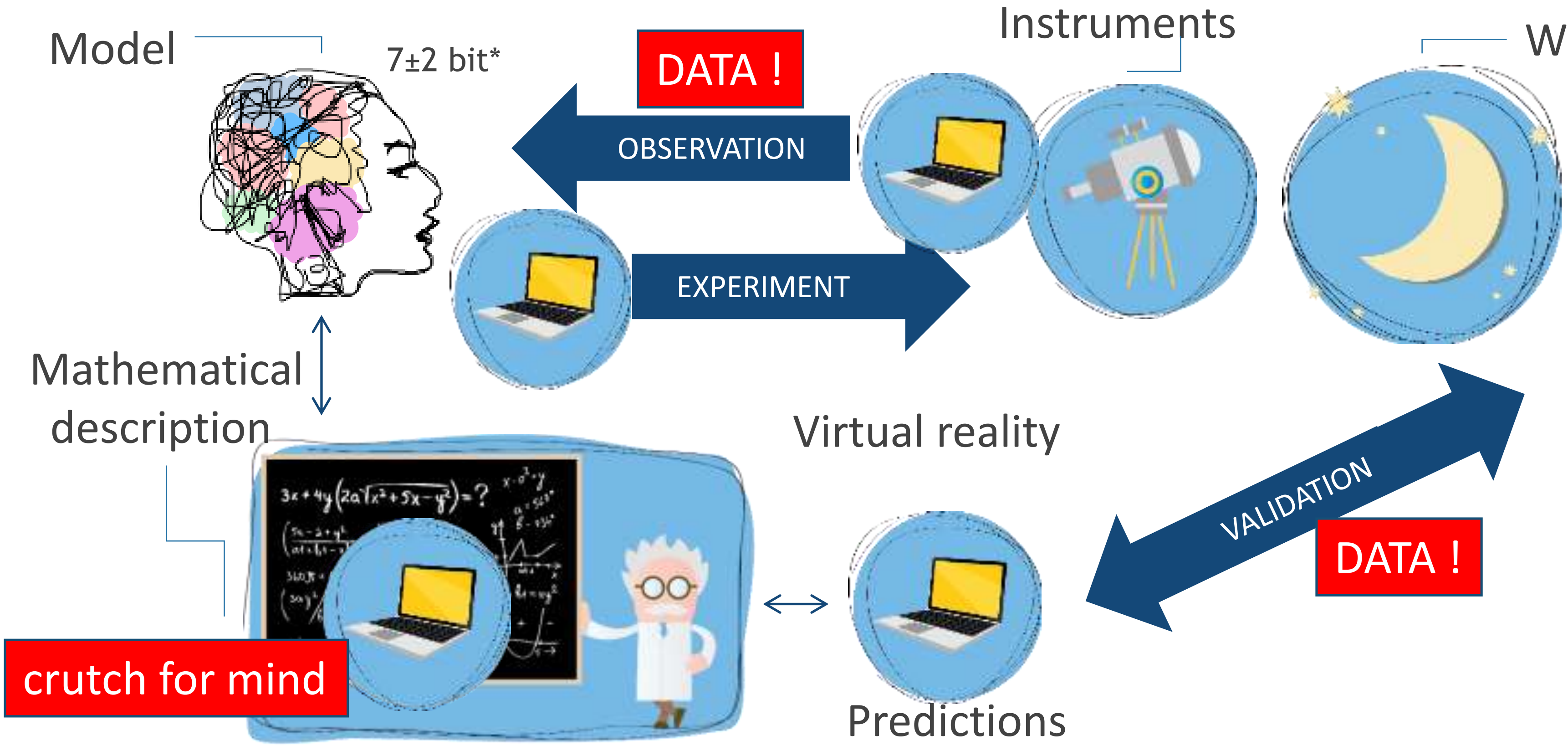
CPU	100 GN (giga-neurons)
Clock frequency	**4-32 Hz**
CPU cores	1 (male version), 2+ (female v.)
CPU speed	0.1 Flops (floating point op. / sec)
Memory (short term)	**7 +/-2 bits**
Storage	1TB-2.5PB
Power	20 W
Camera	576Mpix, 24Hz
Touch	Yes
Display	No
Speakers	Mono
GPS	No
WIFI	No
Bluetooth	No
2G/3G/4G/5G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

Main Features :

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

(Data) science

crutch for senses



Model

7±2 bit*

DATA !

OBSERVATION

Instruments

World

EXPERIMENT

Mathematical description

Virtual reality

VALIDATION

DATA !

crutch for mind

Predictions

Initial values

“laws”, equations

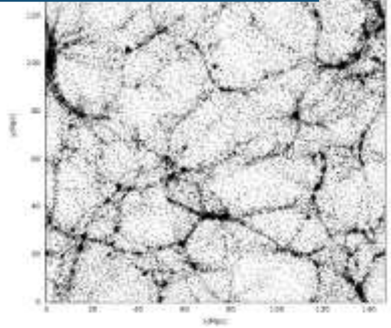
Simulated reality

Λ=0.7

Ω_m=0.3

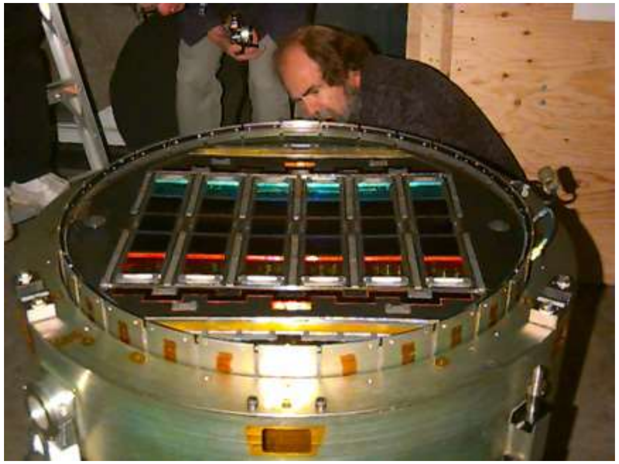
$$F = G\frac{m_1 m_2}{r^2}$$

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



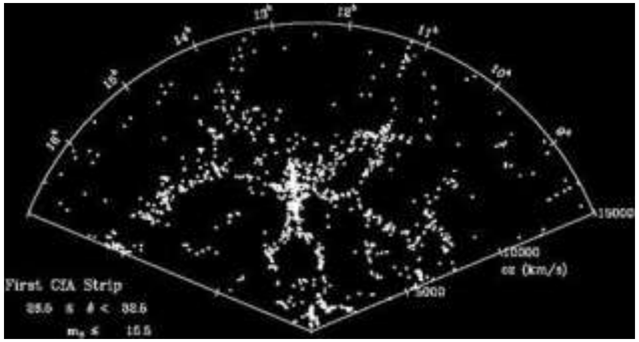
2.5m

120Mp – 2.5Tp



5 years:10TB

CfA 1989: 1100 galaxies

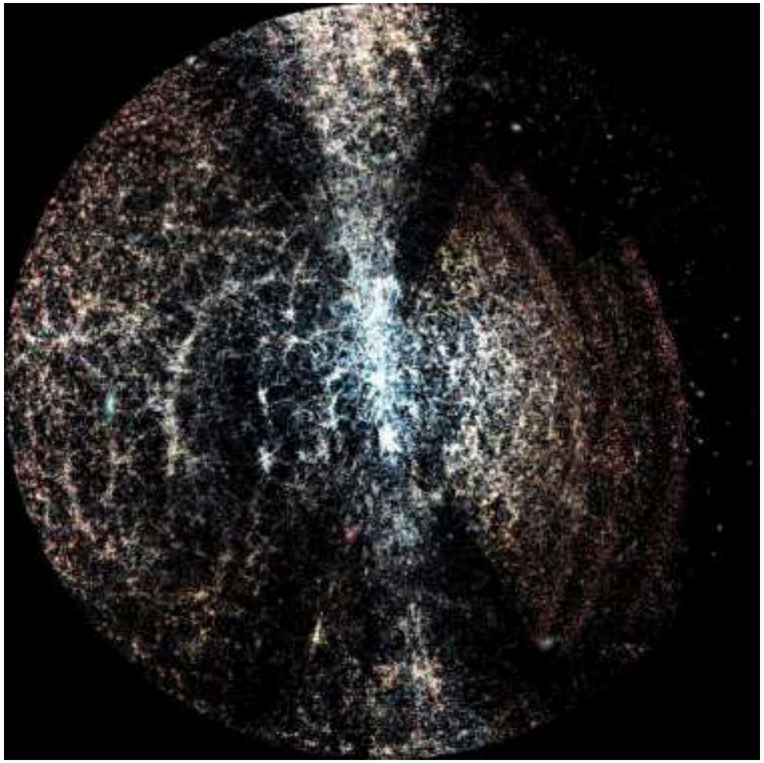


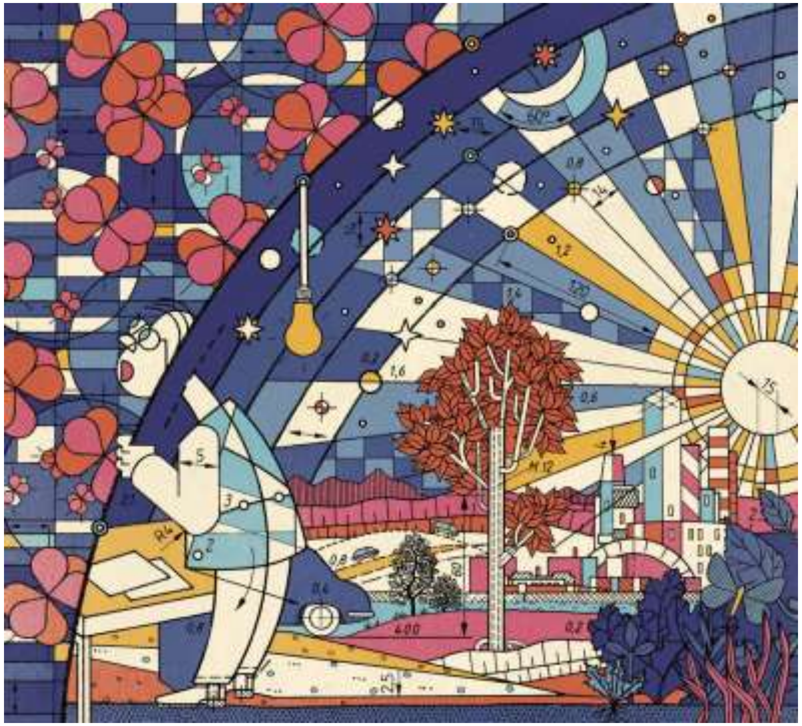
Prototype of modern data science

SDSS: 3D map of the universe

1995 - ...

SDSS 2005: 1M galaxies



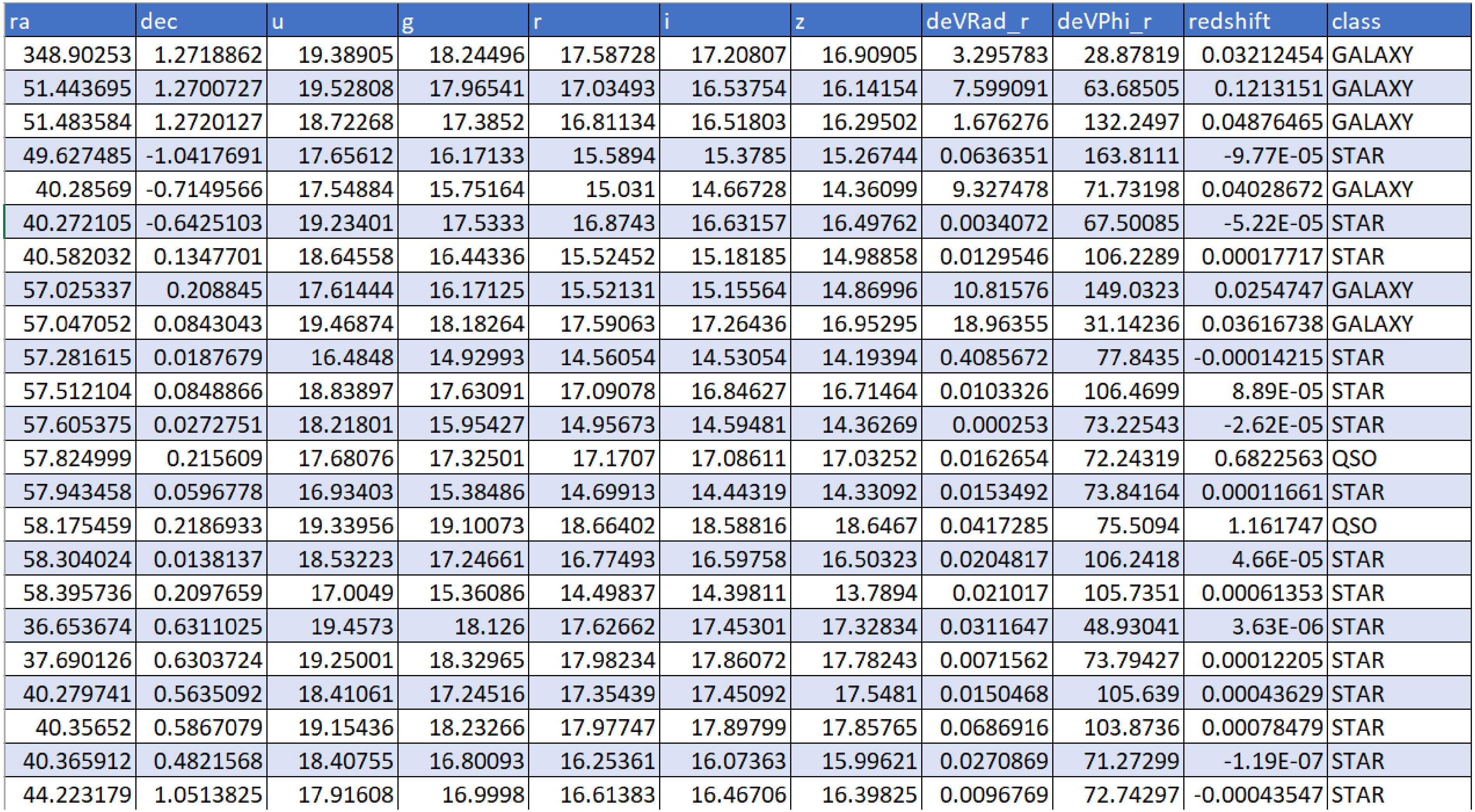


Scientific goals

and

researcher’s perspective

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

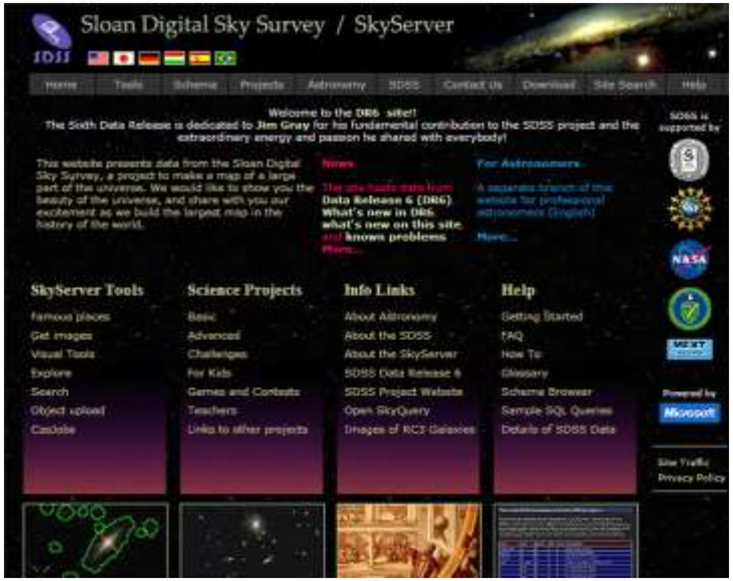
Photometry table: 300+ columns, 1Bn+ rows

100+ other tables

Scientific observations often result data as multidimensional vector space

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")*



Sloan Digital Sky Survey / SkySever

SDSS

Home | Tools | Schema | Projects | Astronomy | SDSS | Contact Us | Download | Site Search | Help

Welcome to the DR6 site!!
The Sixth Data Release is dedicated to Jim Gray for his fundamental contribution to the SDSS project and the extraordinary energy and passion he shared with everybody!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

News

The site hosts data from Data Release 6 (DR6)
What's new in DR6
what's new on this site
and known problems
More...

For Astronomers

A separate branch of this website for professional astronomers (English)
More...

SDSS is supported by

SkyServer Tools
Famous places
Get images
Visual Tools
Explore
Search
Object upload
CasJobs

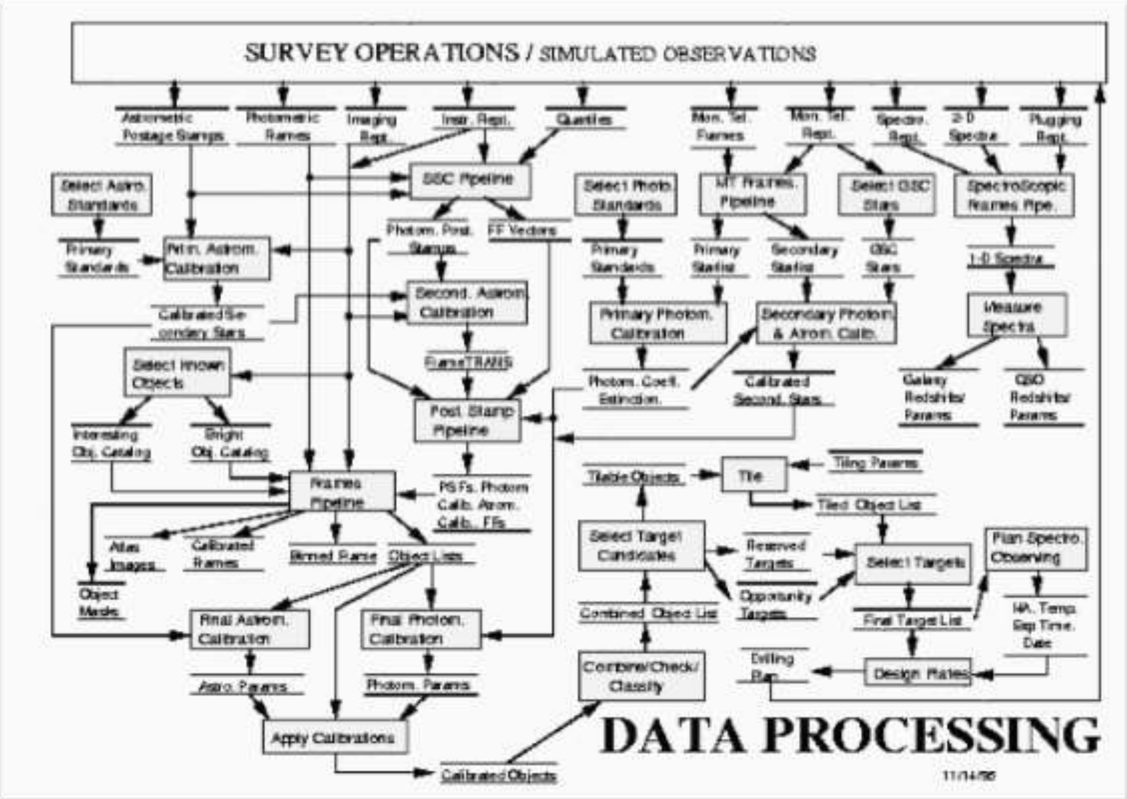
Science Projects
Basic
Advanced
Challenges
For Kids
Games and Contests
Teachers
Links to other projects

Info Links
About Astronomy
About the SDSS
About the SkyServer
SDSS Data Release 6
SDSS Project Website
Open SkyQuery
Images of RC3 Galaxies

Help
Getting Started
FAQ
How To
Glossary
Schema Browser
Sample SQL Queries
Details of SDSS Data

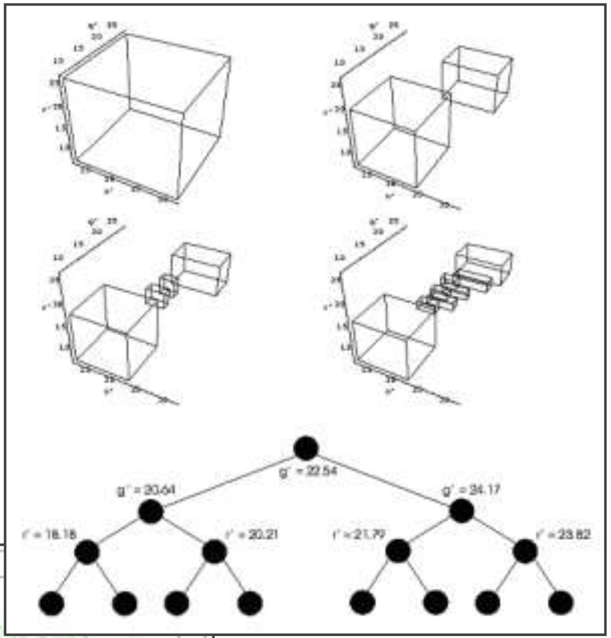
Powered by Microsoft

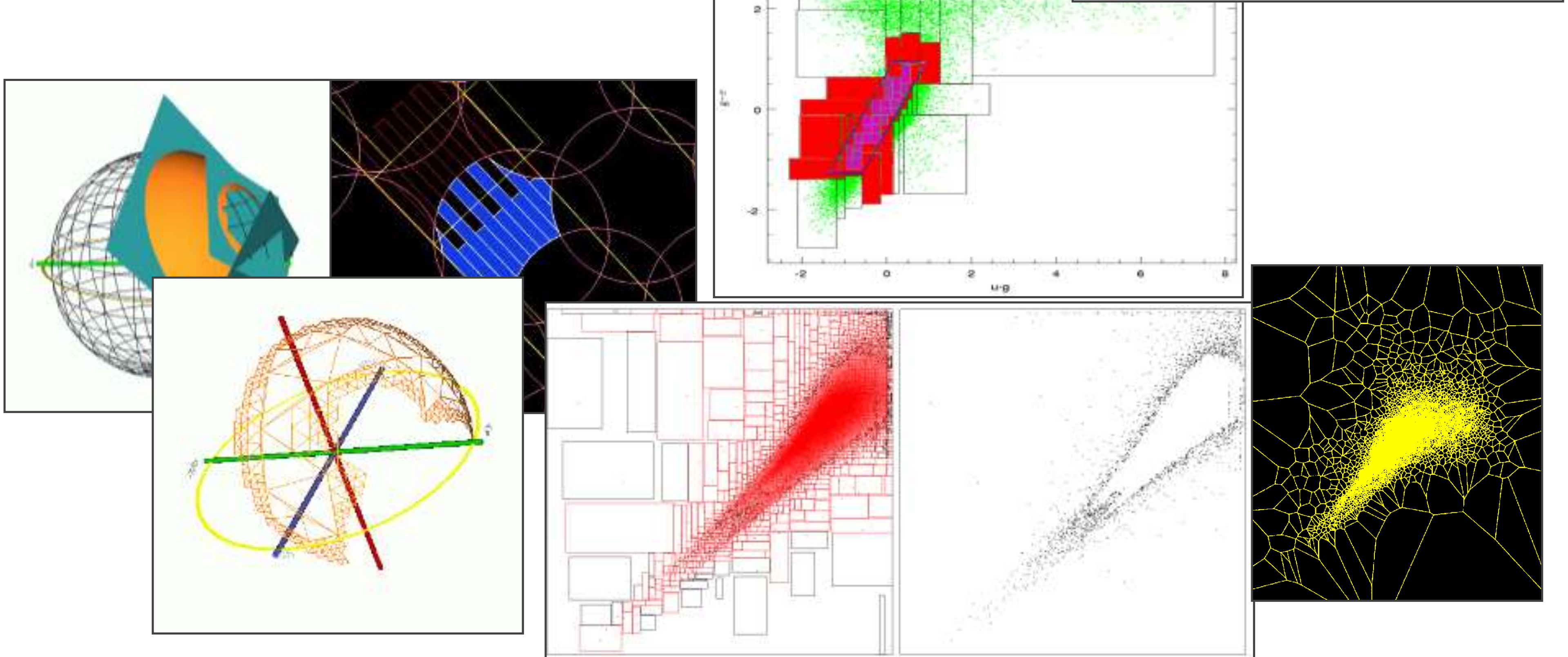
Site Traffic
Privacy Policy



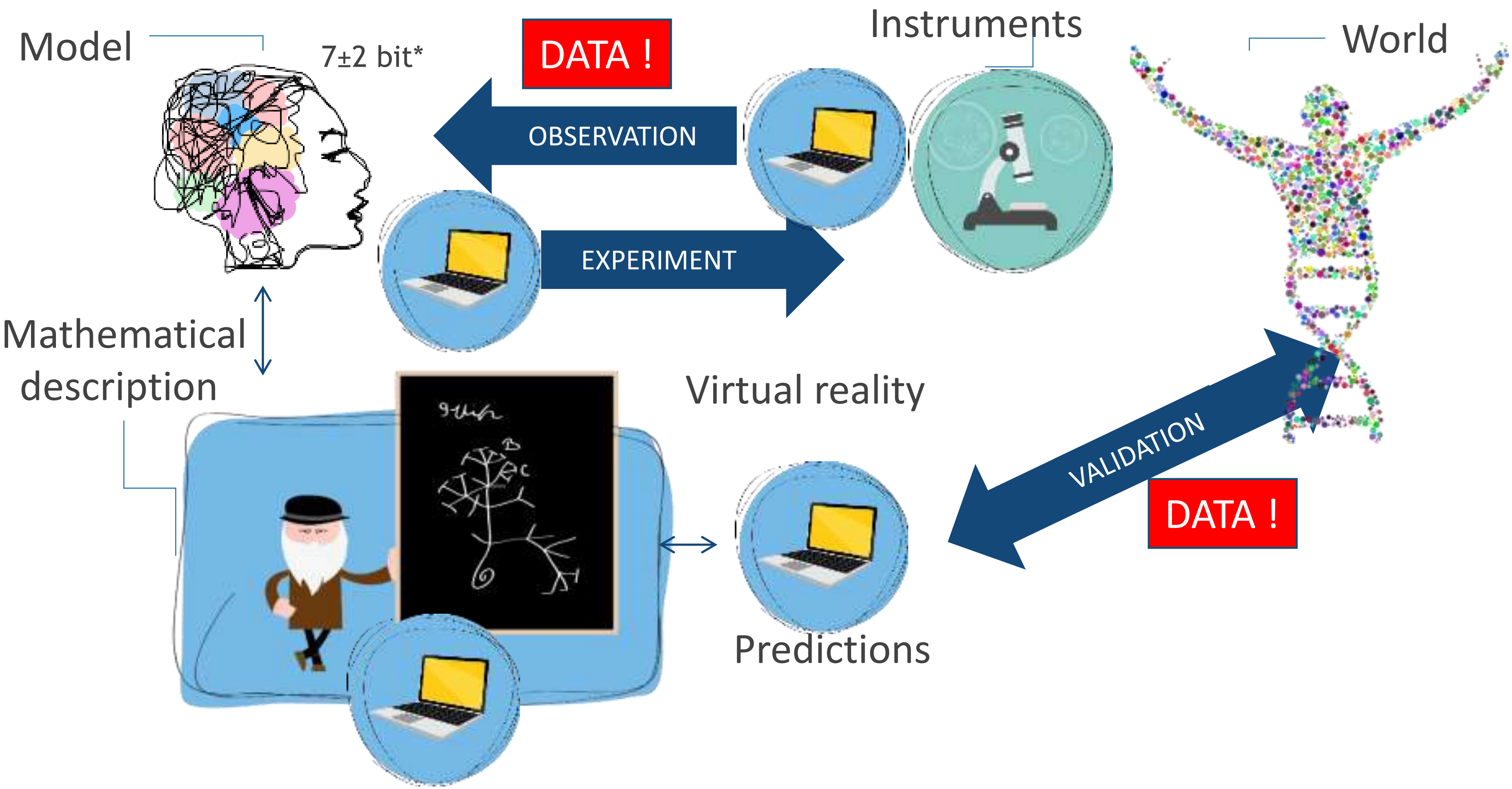
New skills: Indexing, databases

- 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, ...



Model

7±2 bit*

DATA !

OBSERVATION

EXPERIMENT

Instruments

World

Mathematical description

Virtual reality

VALIDATION

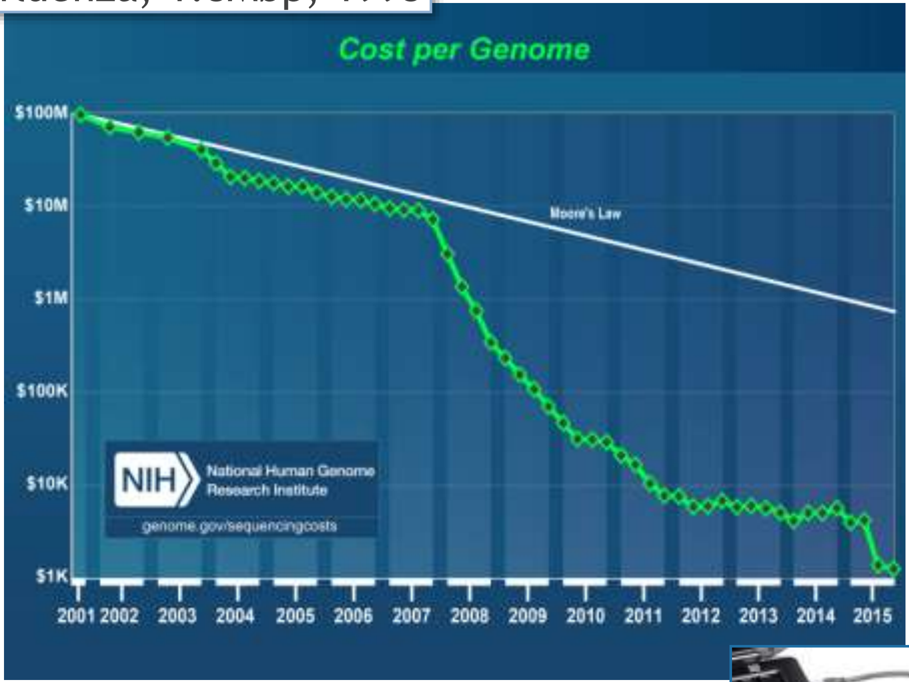
DATA !

Predictions

Moore's law in gene sequencing

phiX 174, 5.4kbp, 1977
H. influenza, 1.8Mbp, 1995

Human genome sequencing
1990-2003: 13yrs /2.7 Bn USD
2016: ~days/1000 USD
2025: ?????



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

-…

2016:$9000/Gb, 2020:$20/Gb, Flongle cell: $90

NGS – data analysis example: genome alignment

Processor speed:	$\sim 10^9$ op/sec
Human genome:	$\sim 10^9$ nt
NGS:	$\sim 10^9$ short reads
“brute force”	$\sim 10^{18}$ comparisions
would take:	$\sim 10^9$ sec $\approx$ $32\ yr$

De novo assembly even more complex task!

Impossible without creative indexing algorithms!

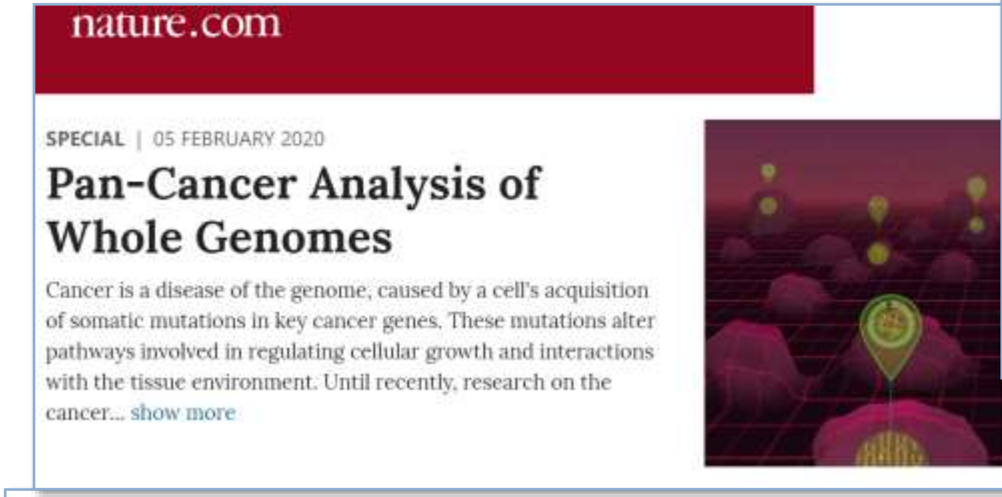
Biology in the

20th 21st

century

Public NGS Data + metadata

ENA
European Nucleotide Archive

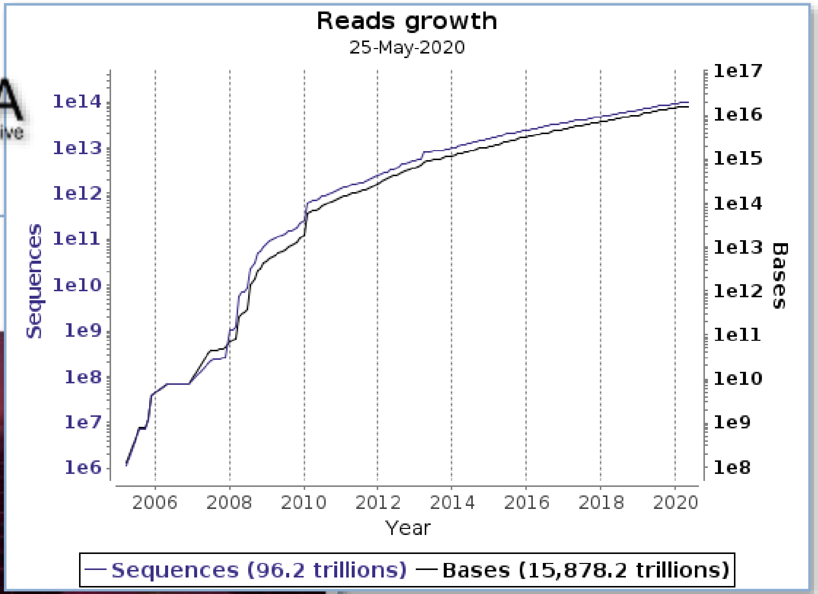


nature.com

SPECIAL | 05 FEBRUARY 2020

Pan-Cancer Analysis of Whole Genomes

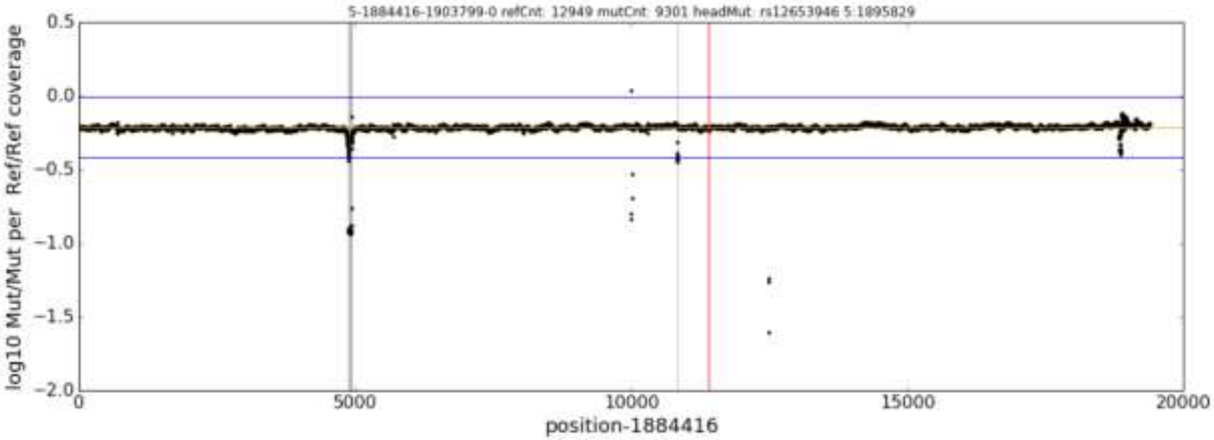
Cancer is a disease of the genome, caused by a cell's acquisition of somatic mutations in key cancer genes. These mutations alter pathways involved in regulating cellular growth and interactions with the tissue environment. Until recently, research on the cancer... show more



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**.

- 3.2 Gb/genome
- + expression, methylation
- + clinical metadata

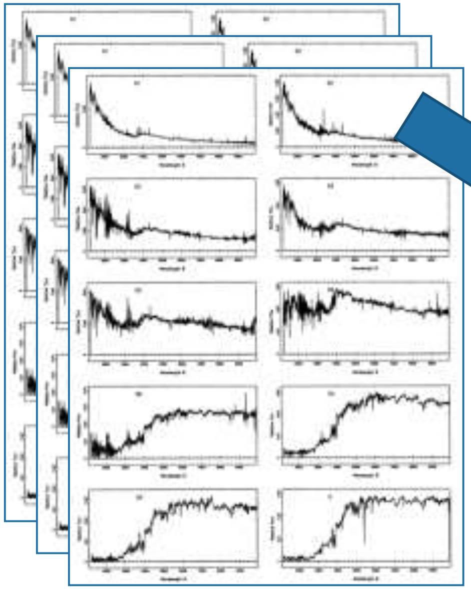
- Example analysis:
  - Role of non-coding variants
  - Indel coverage pileup from ~2000 WGS files



Spisák et al. in prep

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

SIGGRAPHASIA2009

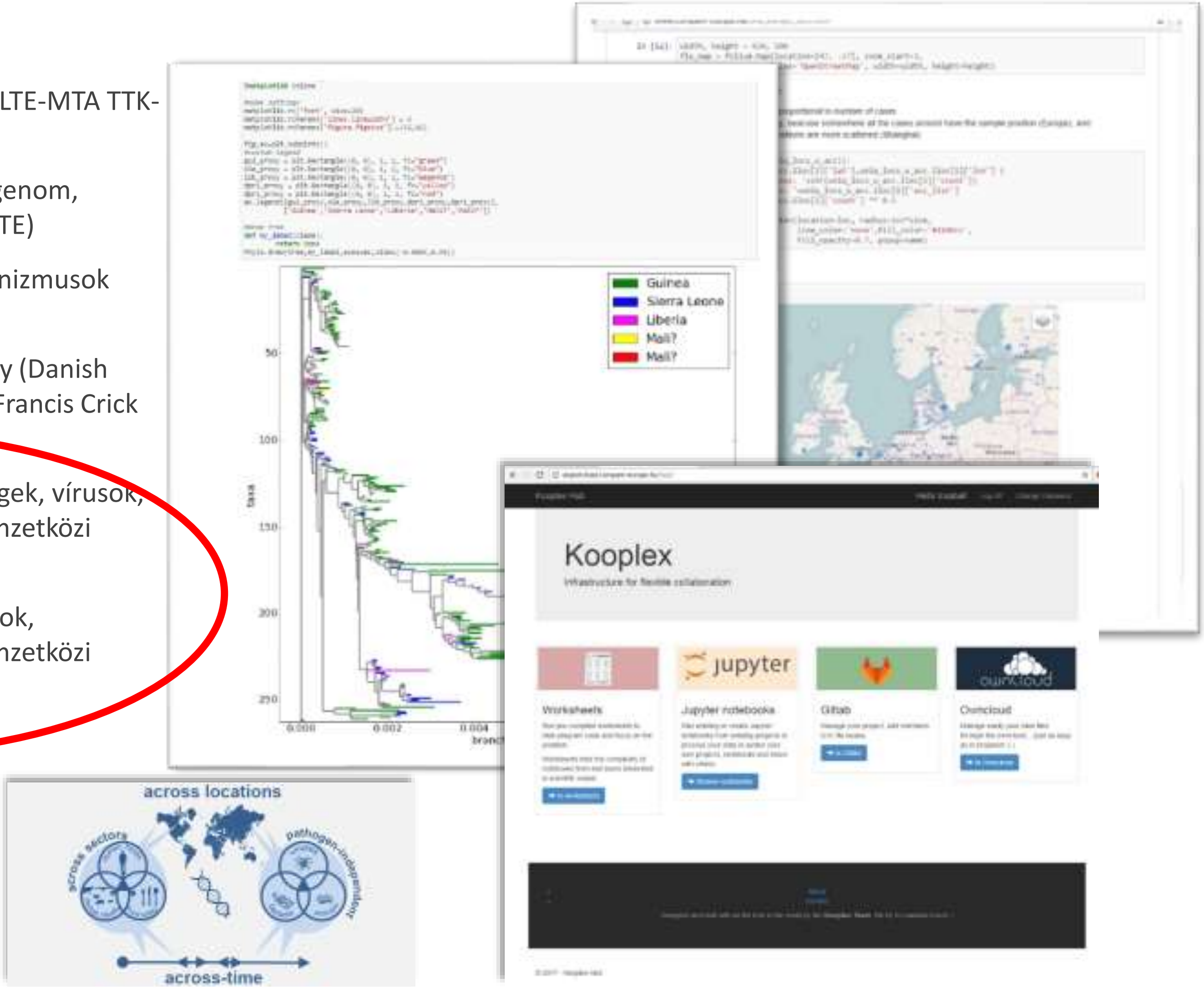
Shadow Art

Niloy J. Mitra
IIT Delhi / KAUST

Mark Pauly
ETH Zurich

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)
- + National and EU COVID projects



Kooplex

Infrastructure for flexible collaboration

Versatile Emerging infectious disease Observatory

2020.01.01-2025.12.31

- 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
- +WP, COVID19

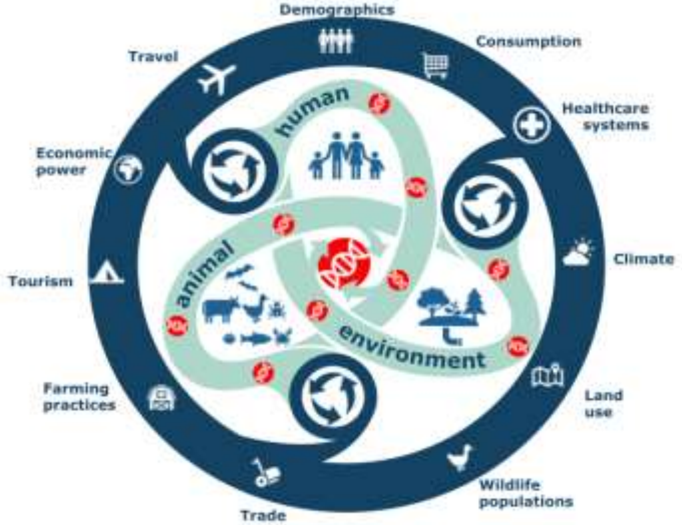


Figure 4: Global changes in global trends acting as drivers of infectious disease emergence and spread in the one health domains

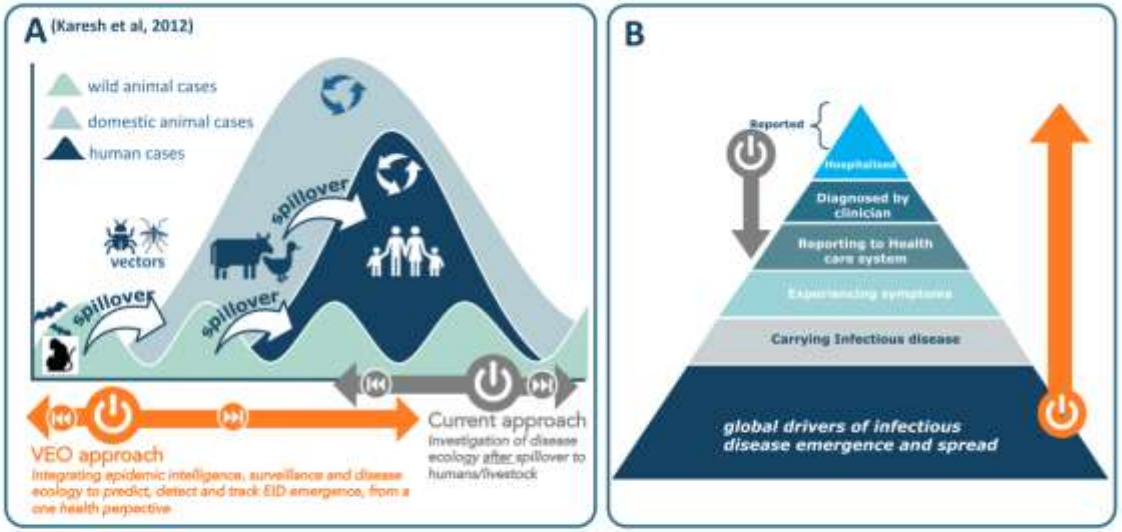
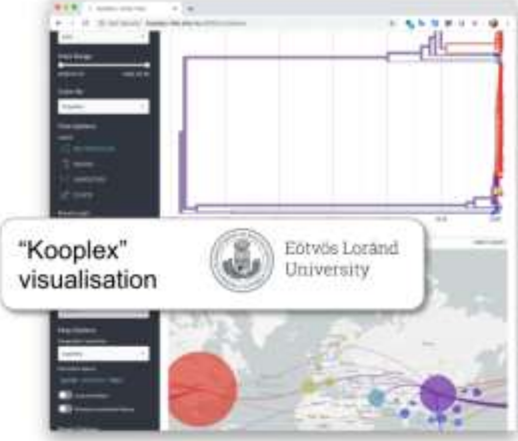

A (Karesh et al, 2012)
wild animal cases
domestic animal cases
human cases
vectors
spillover
spillover
spillover
Current approach
Investigation of disease ecology after spillover to humans/livestock
VEO approach
Integrating epidemic intelligence, surveillance and disease ecology to predict, detect and track EID emergence, from a one health perpective
B
Reported
Hospitalised
Diagnosed by clinician
Reporting to Health care system
Experiencing symptoms
Carrying Infectious disease
global drivers of infectious disease emergence and spread

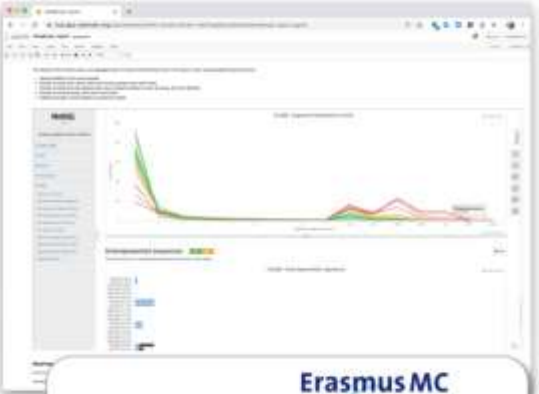
Figure. 1: A. VEO's pro-active, forward looking approach versus the current, reactive approach in EID preparedness and response research (A) and in terms of focusing on drivers of disease emergence and spread instead of taking actions once disease mergence is reported to the healthcare system (B).

SARS-CoV-2 Data Hubs next steps

G. Cochrane, EMBL-EBI 2020.04.28

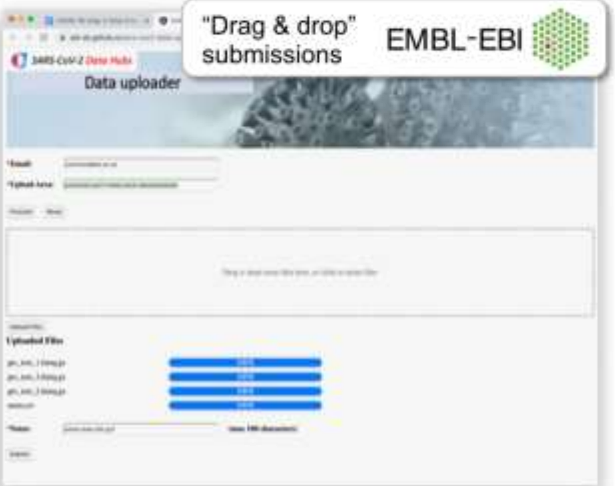

"Kooplex" visualisation

Eötvös Loránd University


"Jovian" analysis

Erasmus MC

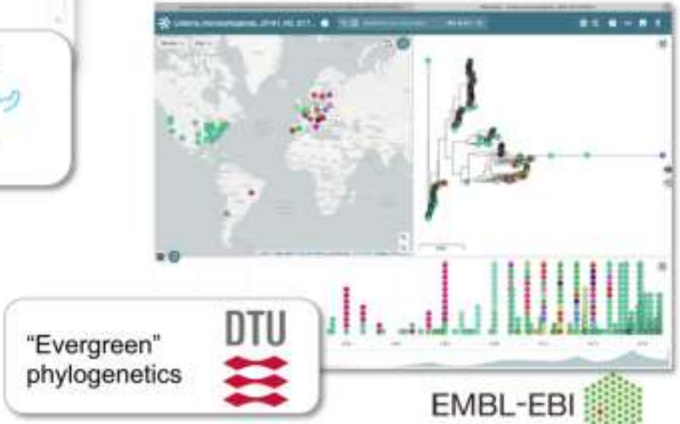
National Institute for Public Health and the Environment
Ministry of Health, Welfare and Sport


"Drag & drop" submissions

EMBL-EBI

"Cohort browser"

UNIVERSITÄTS KLINIKUM HEIDELBERG


"Evergreen" phylogenetics

DTU

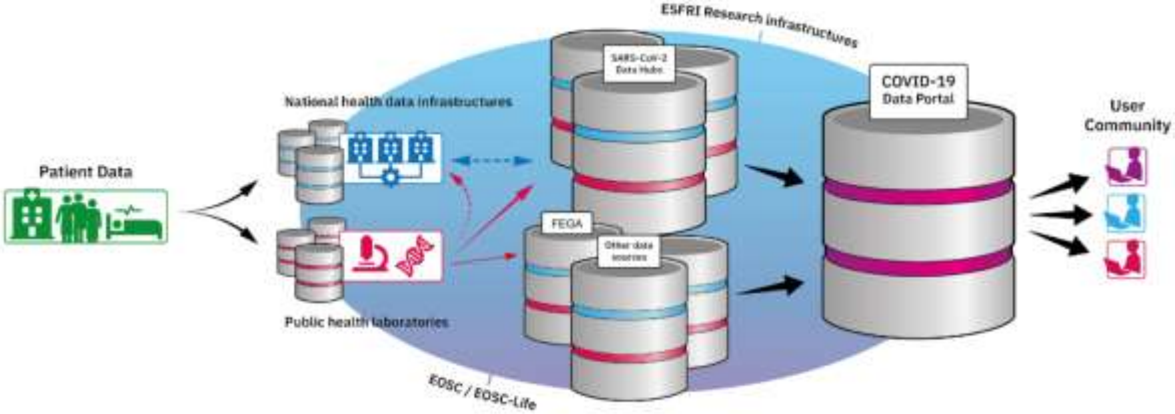
EMBL-EBI

HIGHLIGHTS Coronavirus College #DigitalEU #EUGreenDeal Press Conferences
Midday Briefings Press corner

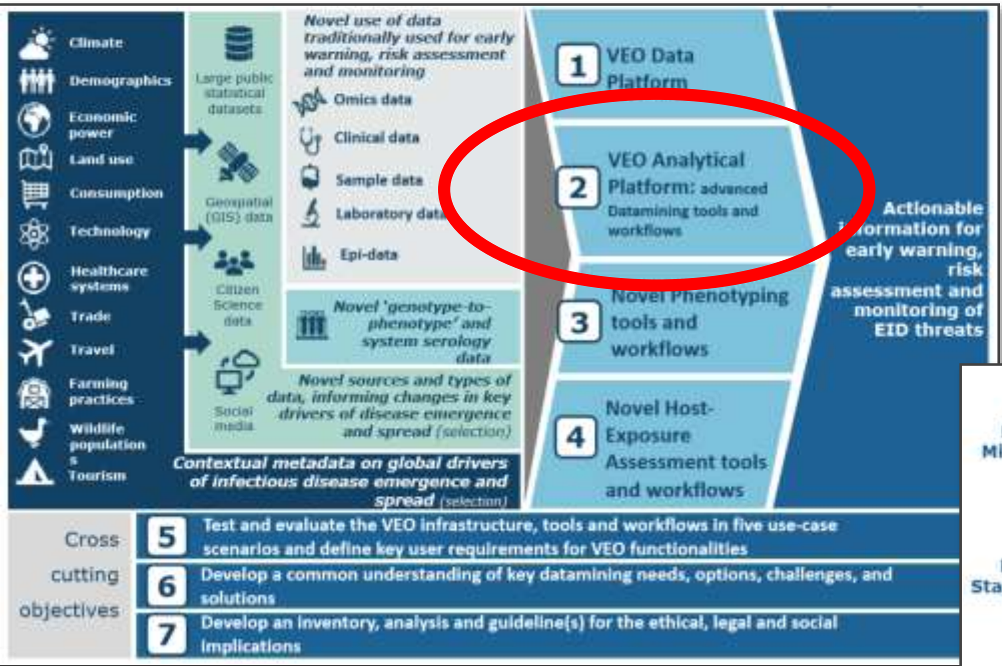
Statement by Ursula von der Leyen, President of the European Commission, on the launch of the EU COVID19 Data Platform (international sign language version)

On 20 April 2020, Ursula von der Leyen, President of the European Commission, recorded a video message at the European Commission in Brussels, Belgium, on the launch of the EU COVID19 Data Platform.

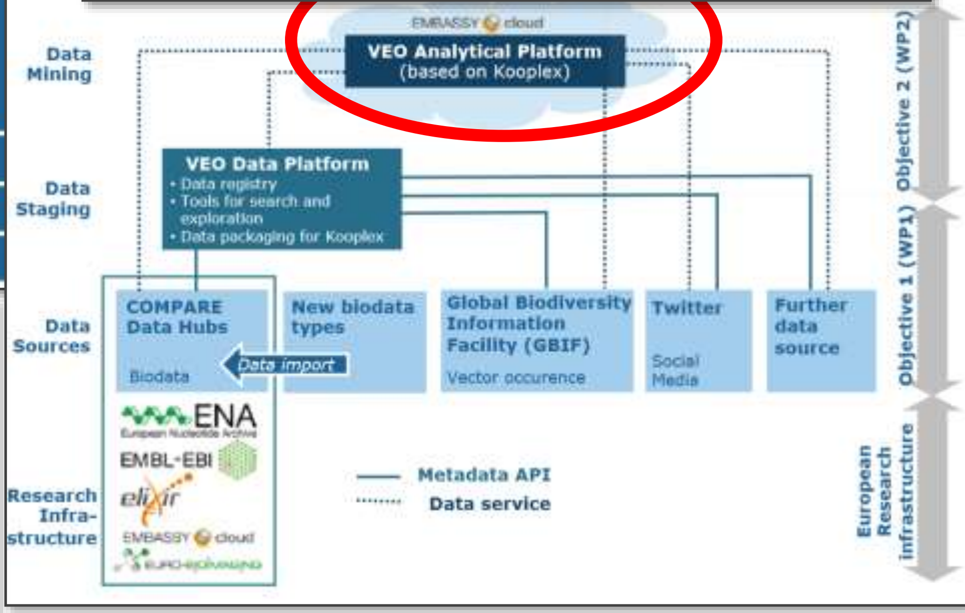
On this platform, researchers will be able to store, exchange and analyse a wide range of knowledge about the



WP2 Analytical Platform: Advanced Datamining tools


Climate
Demographics
Economic power
Land use
Consumption
Technology
Healthcare systems
Trade
Travel
Farming practices
Wildlife populations
Tourism
Large public statistical datasets
Geospatial (GIS) data
Citizen Science data
Social media
Novel use of data traditionally used for early warning, risk assessment and monitoring
Omics data
Clinical data
Sample data
Laboratory data
Epi-data
Novel 'genotype-to-phenotype' and system serology data
Novel sources and types of data, informing changes in key drivers of disease emergence and spread (selection)
Contextual metadata on global drivers of infectious disease emergence and spread (selection)
1 VEO Data Platform
2 VEO Analytical Platform: advanced Datamining tools and workflows
3 Novel Phenotyping tools and workflows
4 Novel Host-Exposure Assessment tools and workflows
Actionable information for early warning, risk assessment and monitoring of EID threats
Cross cutting objectives
5 Test and evaluate the VEO infrastructure, tools and workflows in five use-case scenarios and define key user requirements for VEO functionalities
6 Develop a common understanding of key datamining needs, options, challenges, and solutions
7 Develop an inventory, analysis and guideline(s) for the ethical, legal and social implications

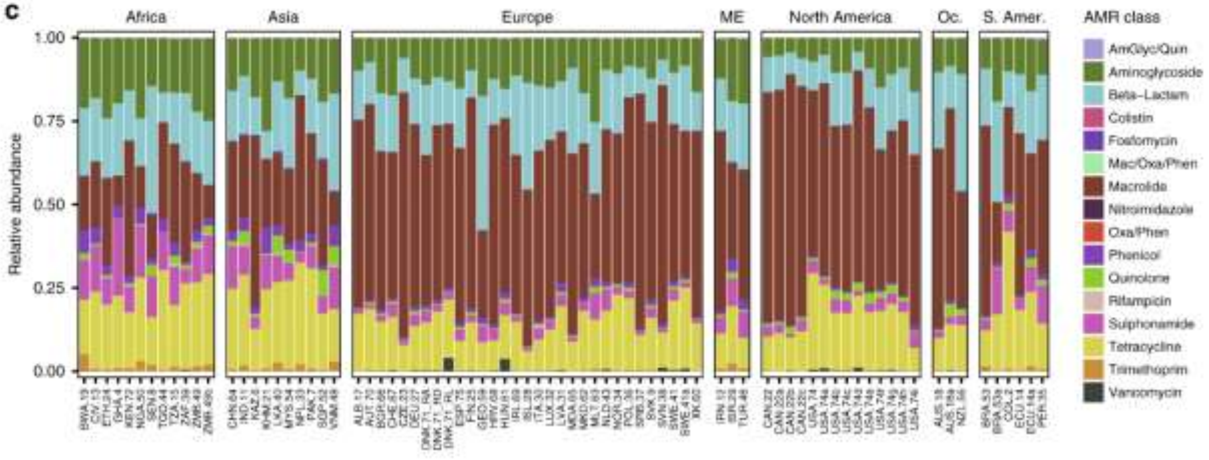
Objective: "develop novel cloud-based datamining tools and services, supporting data-intensive interdisciplinary collaboration of geographically distributed international teams"

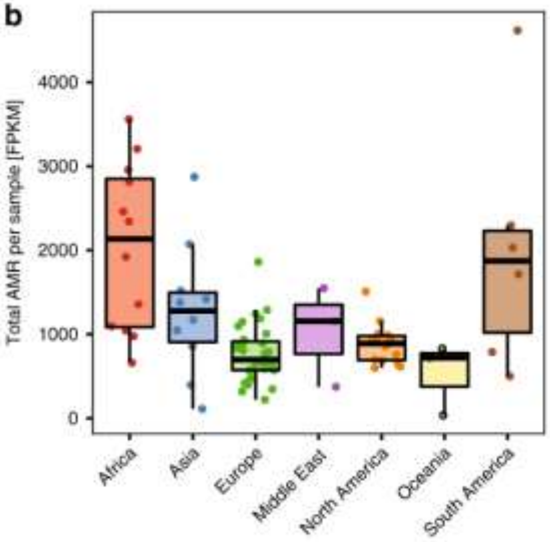


Links:

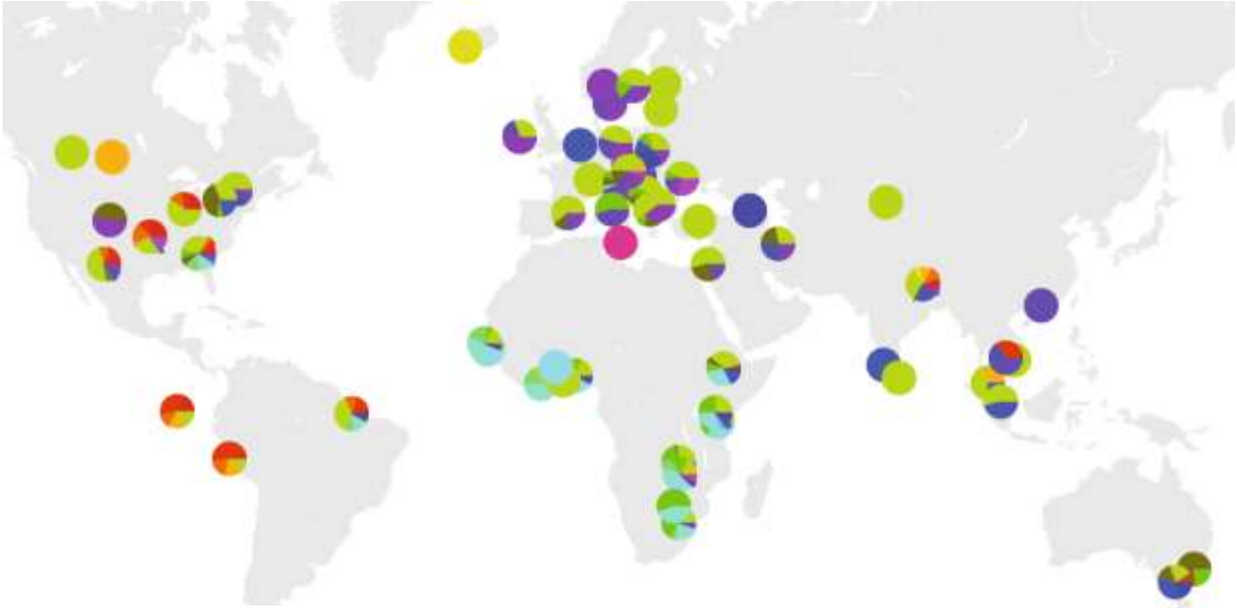
- **WP1** provides data and deployment platform
- **Other WPs** provide data, collaborate on analysis tools, test/use analytical platform

Global Sewage sequencing from 81 cities
Monitoring diseases, antimicrobiotic resistance
… and human phylogeny (Metagenome!)

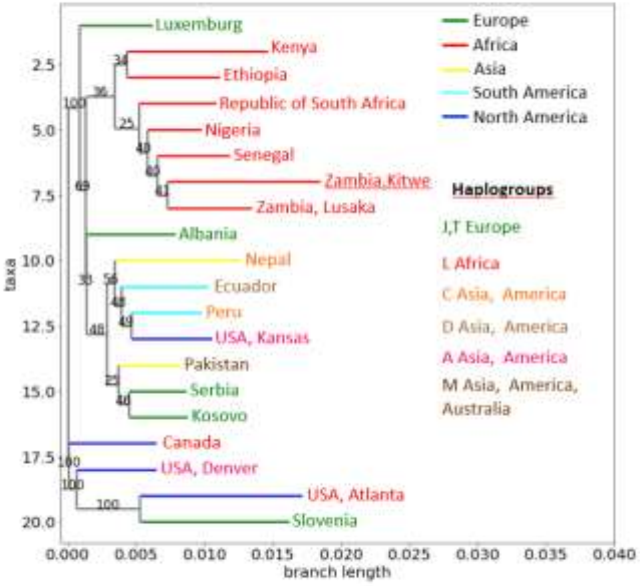




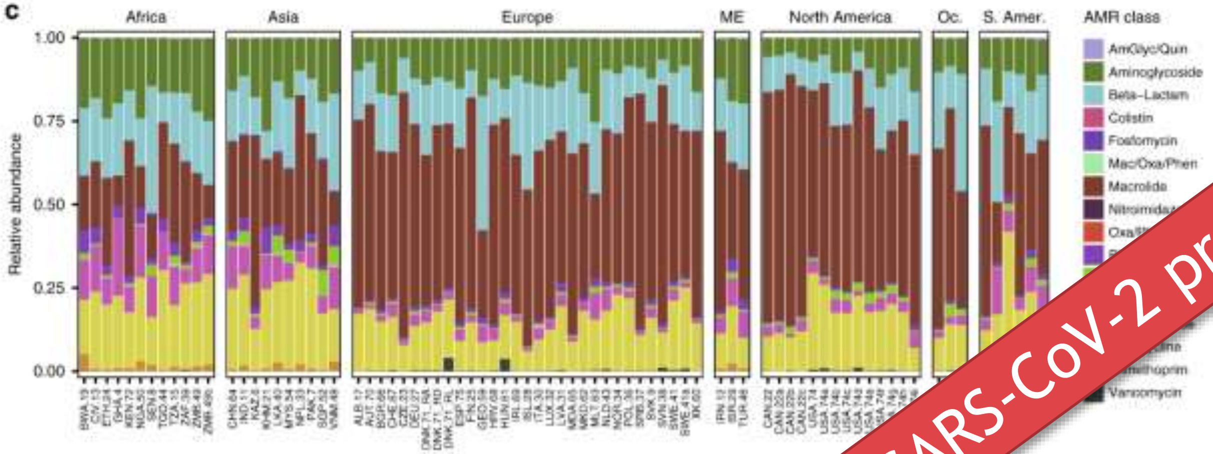
Hendriksen et al. Nat. Comm. 2019

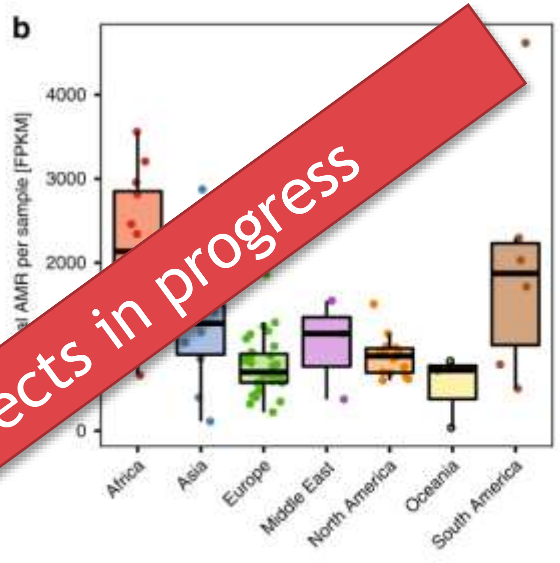


Pipek et al. Sci. Rep. 2019



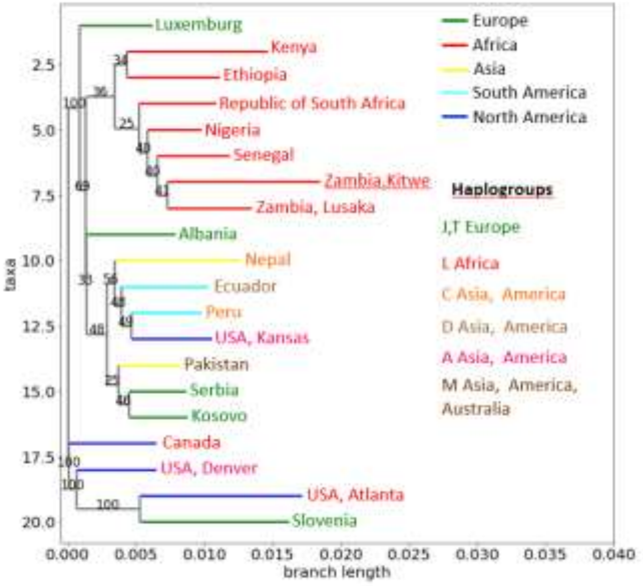
Global Sewage sequencing from 81 cities
Monitoring diseases, antimicrobiotic resistance
… and human phylogeny (Metagenome!)





Hendriksen et al. Nat. Comm. 2019

Global and national sewage based SARS-CoV-2 projects in progress

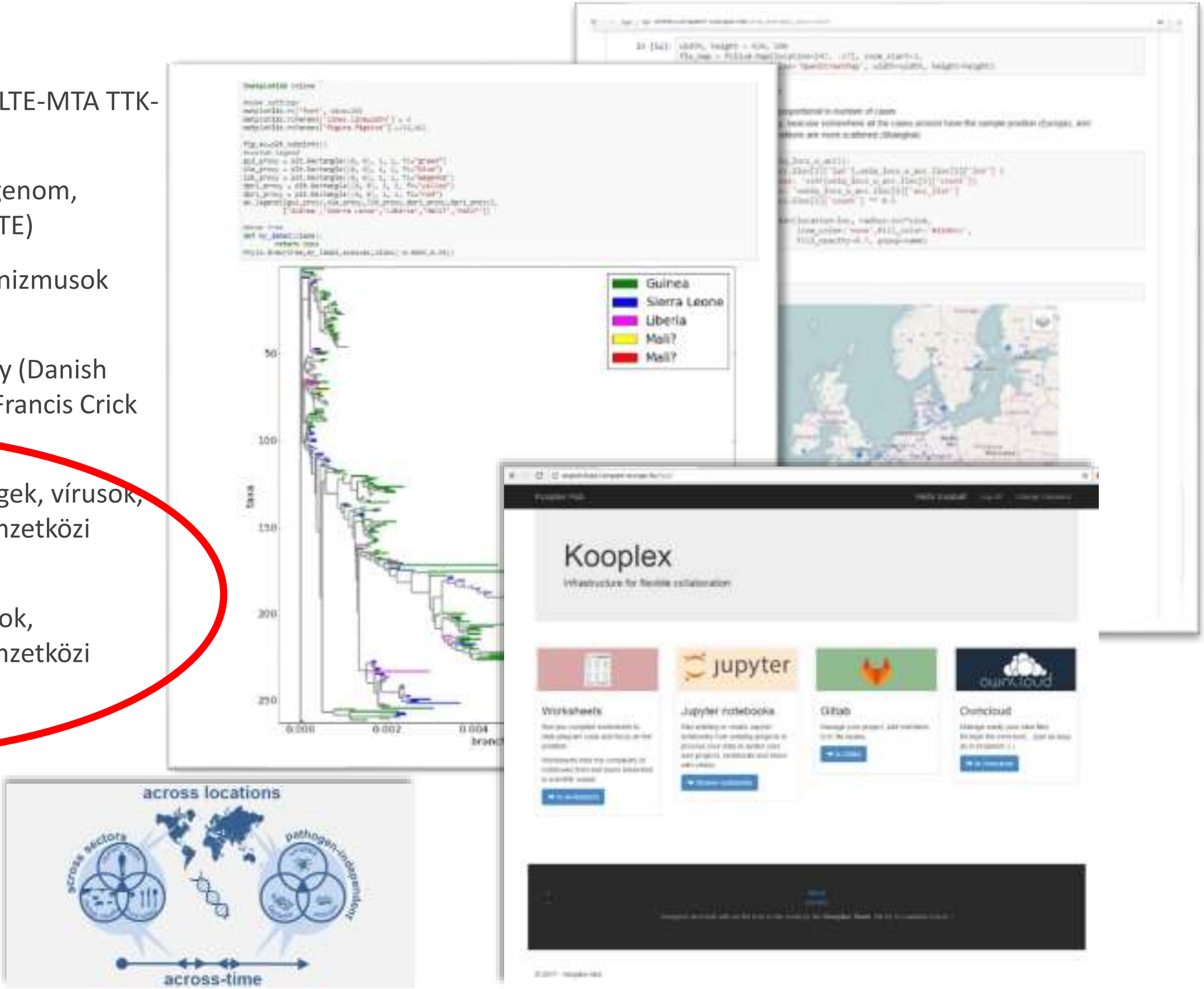


Pipek et al. Sci. Rep. 2019

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)
- + National and EU COVID projects

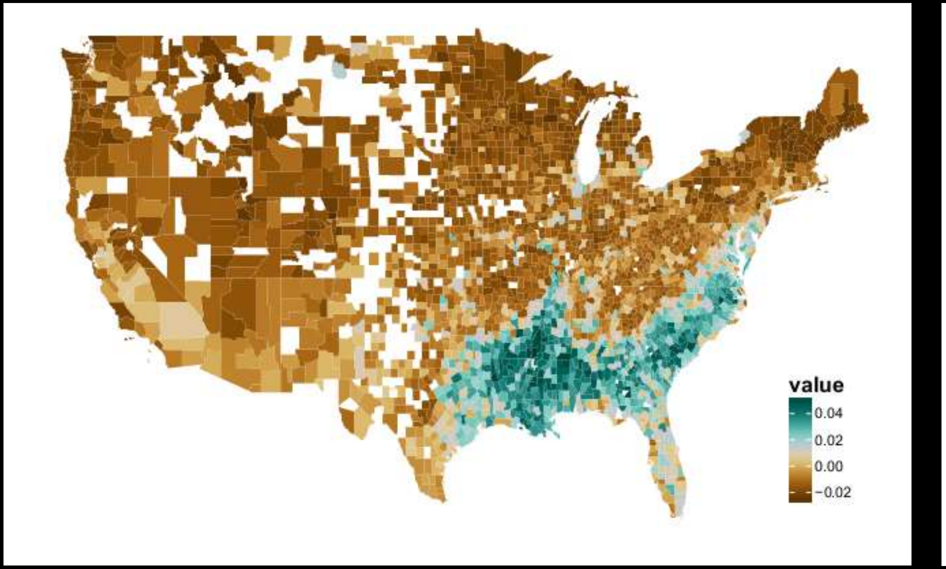
#nCoV,
#Wuhan,

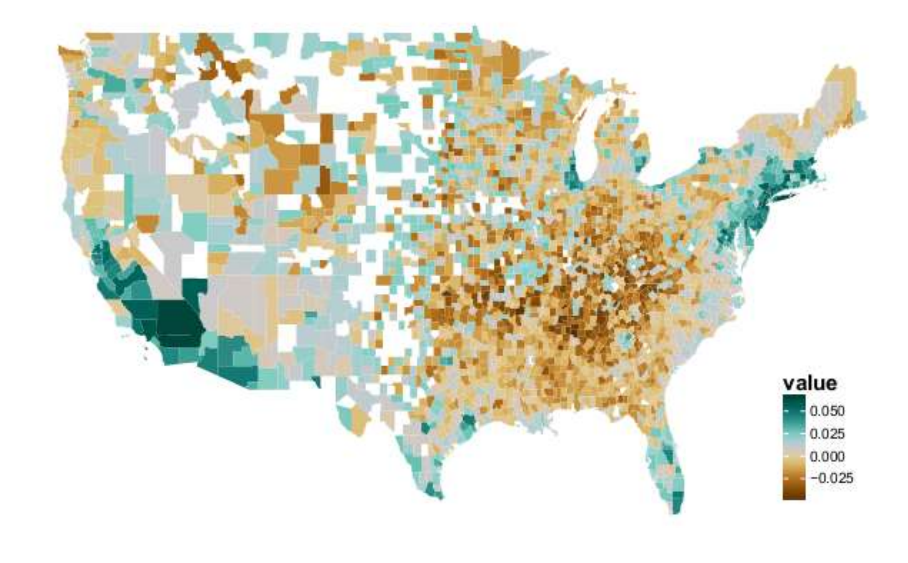


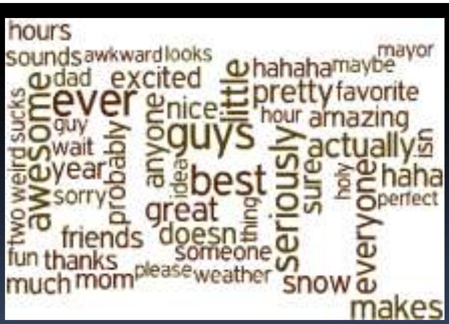
Social networks: TwitterDB

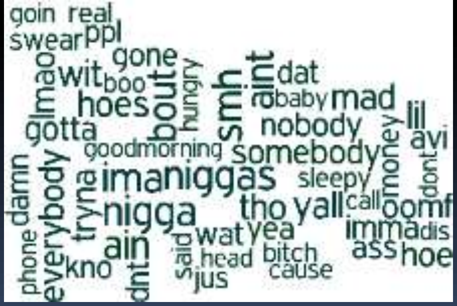
Principal dimensions: race, religion, urbanization

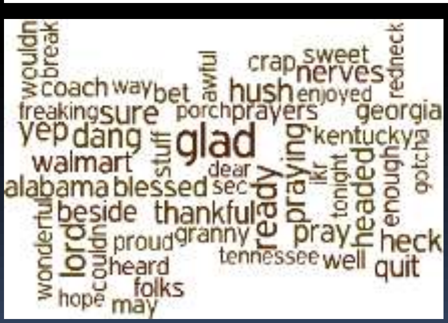
Data type:
Graph + text + geo

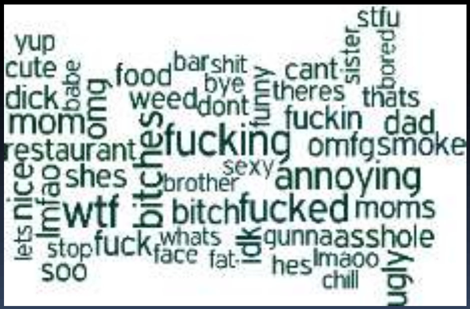








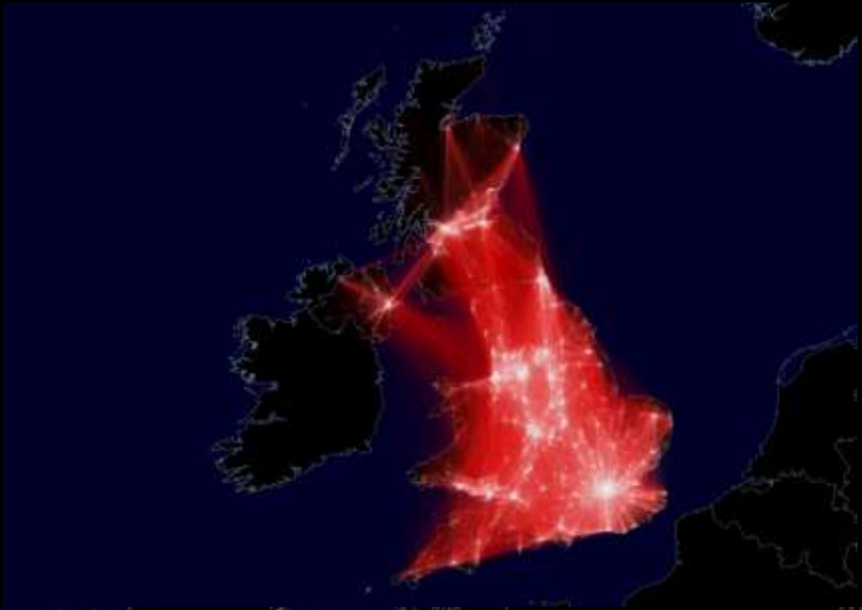


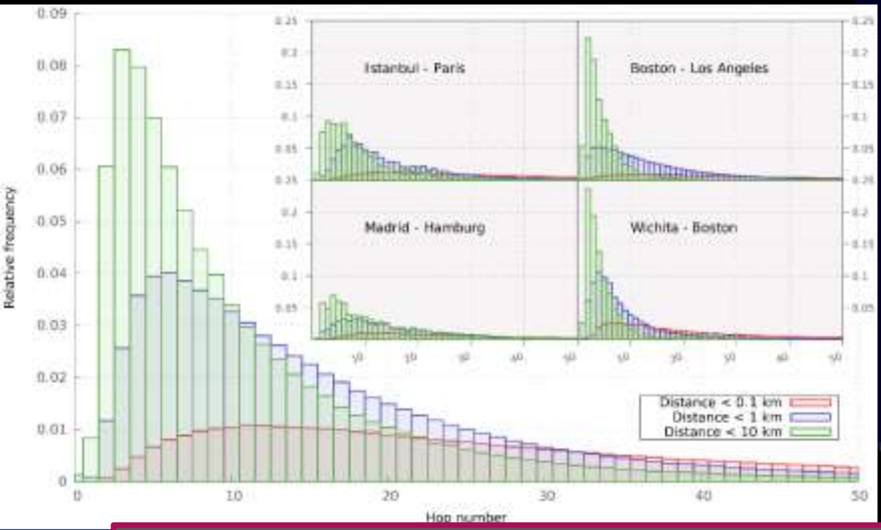


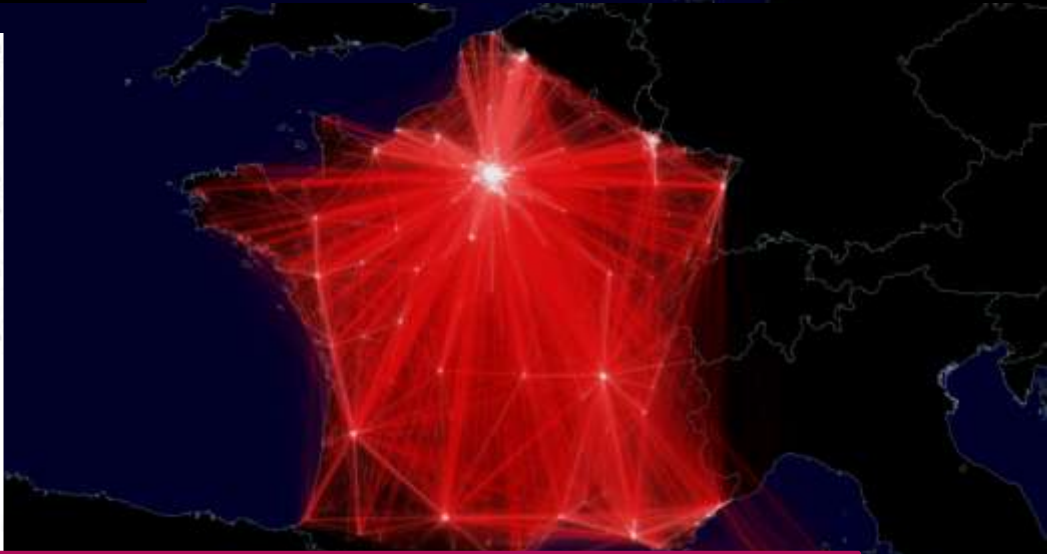
Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages; D Kondor, I Csabai, L Dobos, J Szule, N Barankai, T Hanyecz, T Sebok, Z Kallus, G Vattay; IEEE CogInfoCom) (2013)

Bokányi Eszter, MSc thesis, ELTE TTK (2015), Bokanyi et al. 2016

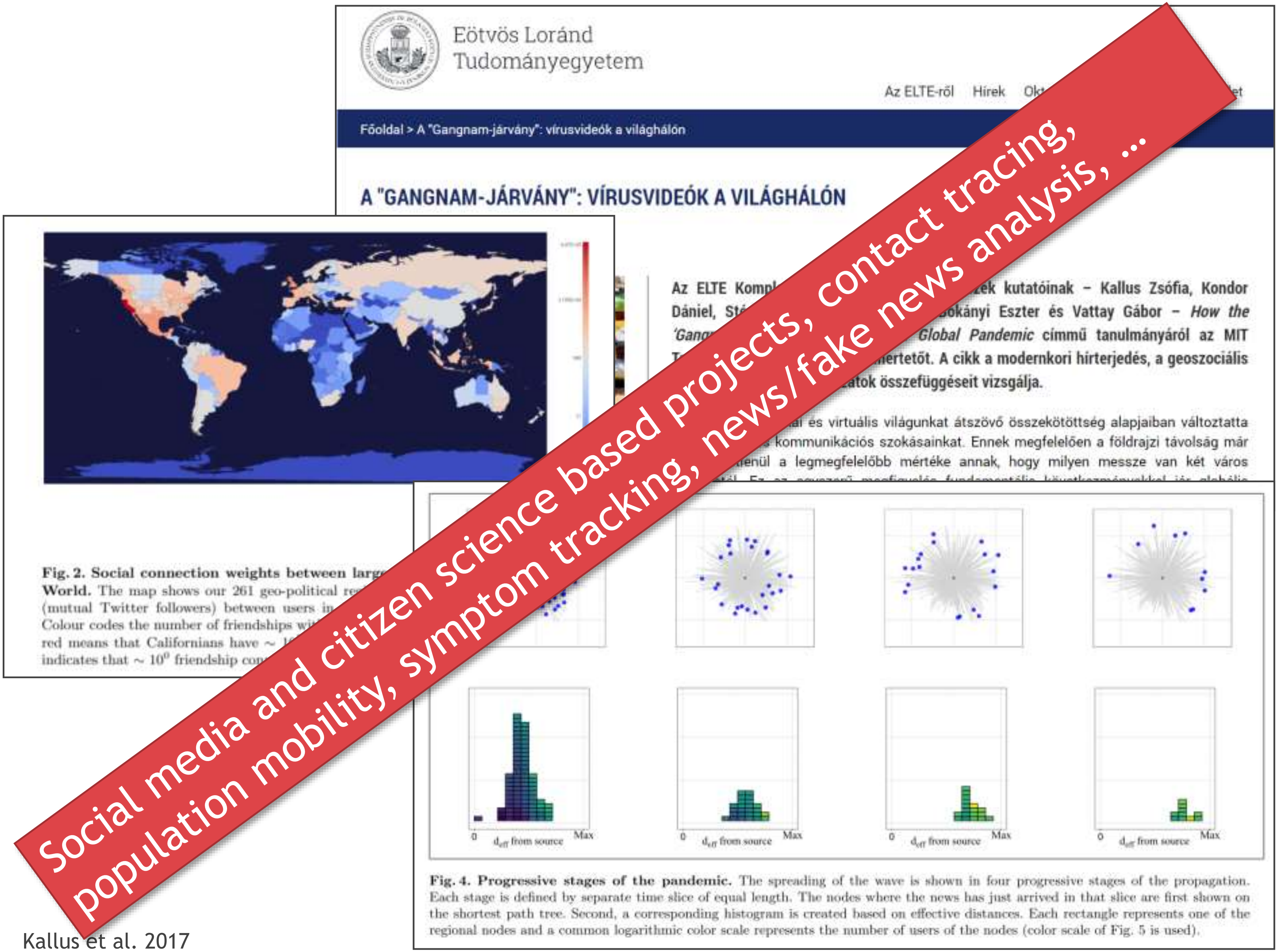
Test Milgram's „6 degree" on Twitter







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)



Social media and citizen science based projects, contact tracing, population mobility, symptom tracking, news/fake news analysis, ...

Eötvös Loránd Tudományegyetem

Az ELTE-ről Hírek Okt

Főoldal > A "Gangnam-járvány": vírusvideók a világhálón

A "GANGNAM-JÁRVÁNY": VÍRUSVIDEÓK A VILÁGHÁLÓN

Az ELTE Kompl zék kutatóinak – Kallus Zsófia, Kondor Dániel, Sté Bokányi Eszter és Vattay Gábor – *How the 'Gang Global Pandemic* címmű tanulmányáról az MIT T mertetőt. A cikk a modernkori hírterjedés, a geoszociális zatok összefüggéseit vizsgálja.

ai és virtuális világunkat átszövő összekötöttség alapjaiban változtatta kommunikációs szokásainkat. Ennek megfelelően a földrajzi távolság már tlenül a legmegfelelőbb mértéke annak, hogy milyen messze van két város

Fig. 2. Social connection weights between large World. The map shows our 261 geo-political re (mutual Twitter followers) between users in Colour codes the number of friendships wi red means that Californians have ~ 1 indicates that ~ 10^0 friendship con

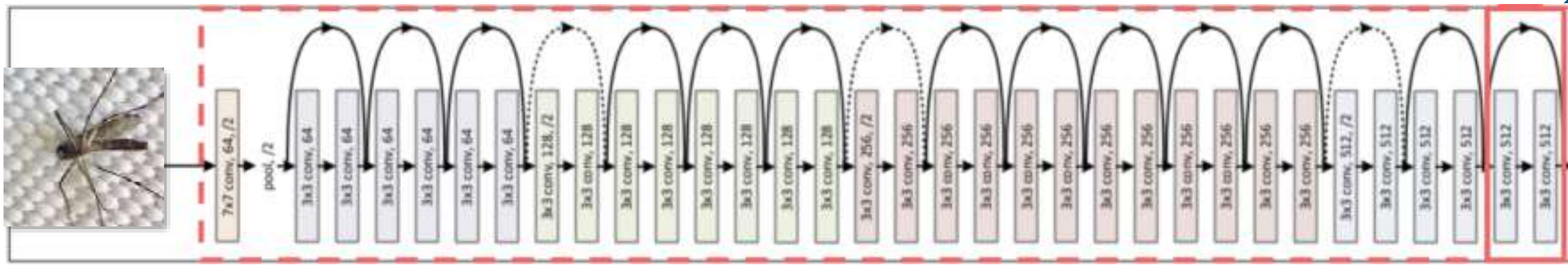
Fig. 4. Progressive stages of the pandemic. The spreading of the wave is shown in four progressive stages of the propagation. Each stage is defined by separate time slice of equal length. The nodes where the news has just arrived in that slice are first shown on the shortest path tree. Second, a corresponding 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).

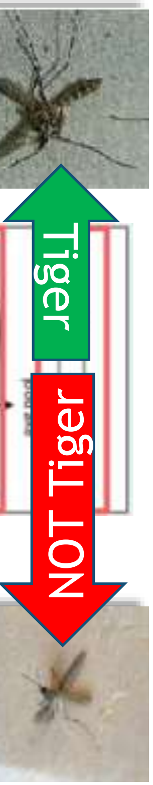
Kallus et al. 2017

Task 2.3 Machine learning tools: MosquitoAlert image deep learning

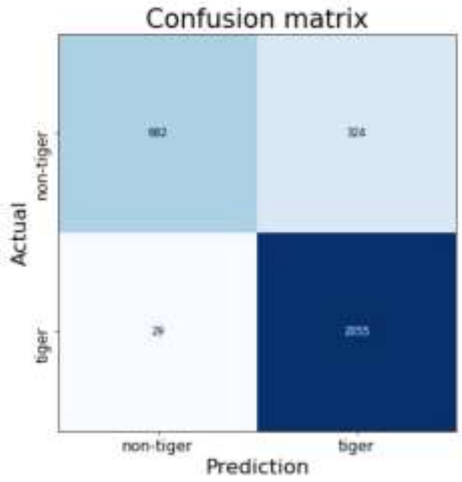
F. Bartumeus et al.
http://www.mosquitoalert.com/

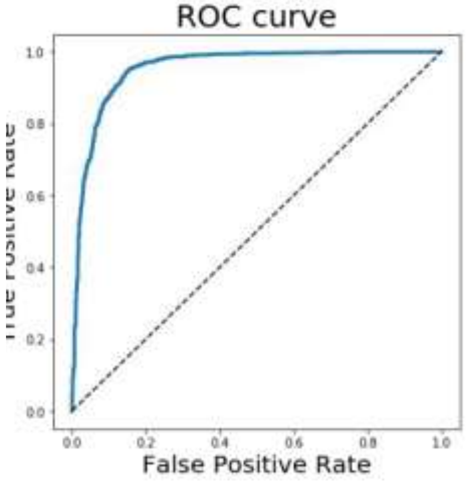
Pataki et al. in prep.





Cropped, AUC=0.944

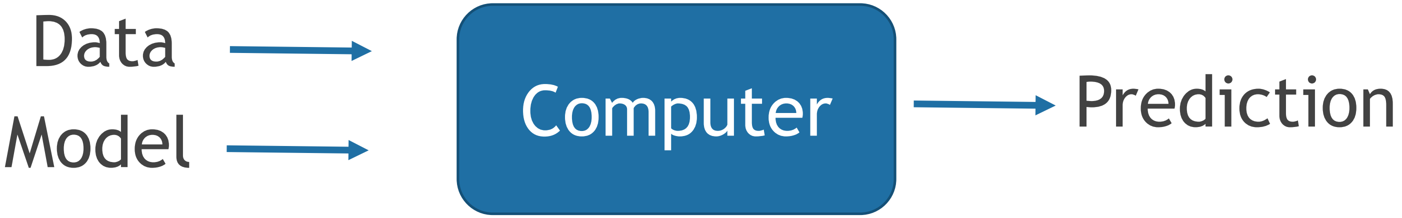




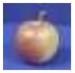
False(?) negatives:

False(?) positives:

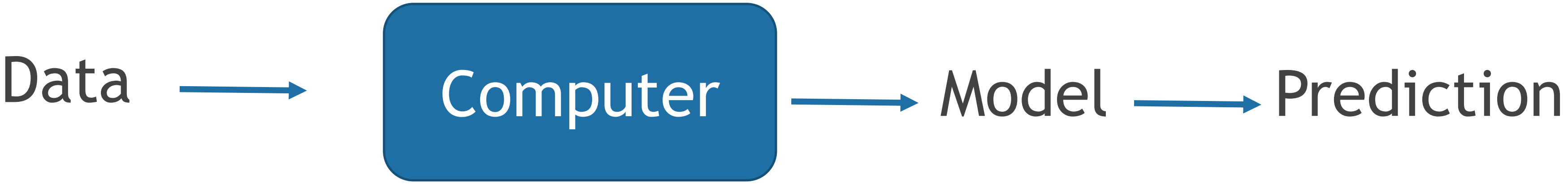
AI: paradigm shift



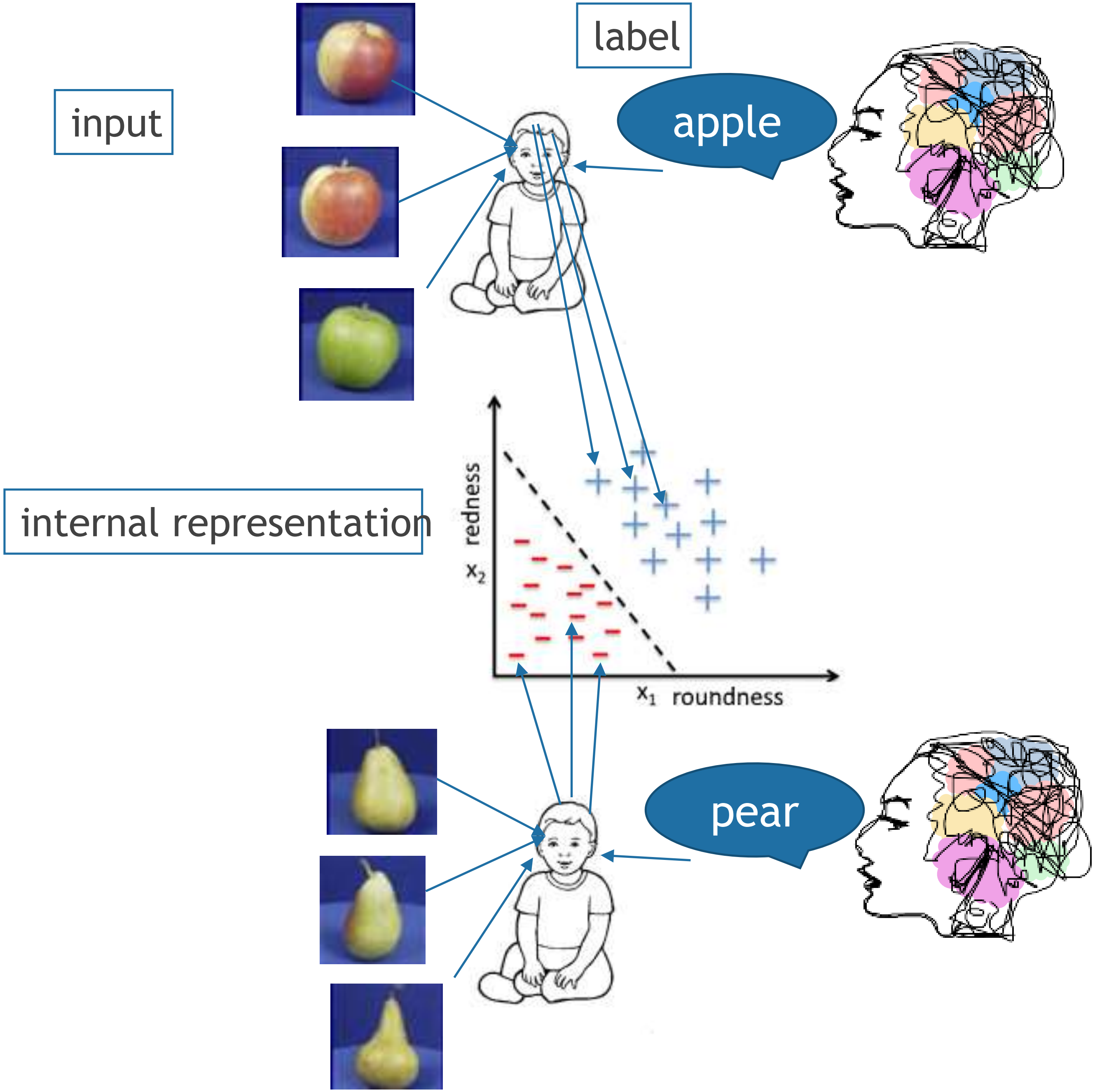
Example: Image recognition
Method: hand crafted features

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



label

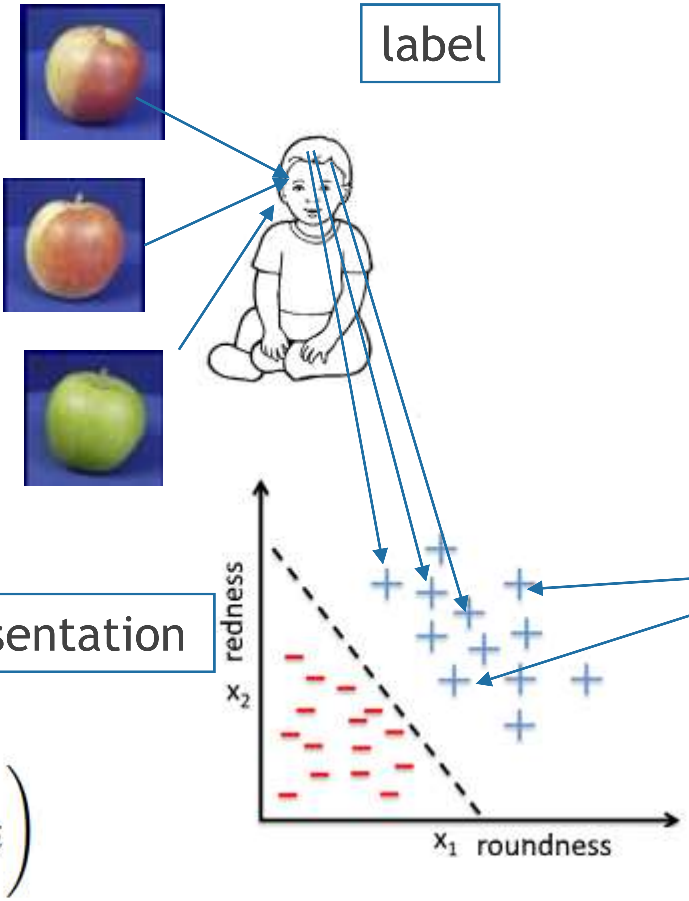
input

internal representation

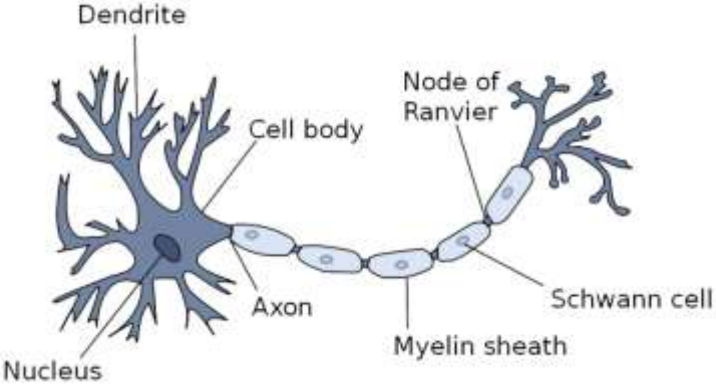
Supervised learning: neural net

input

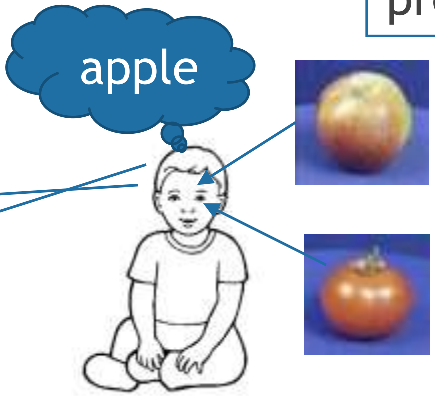
label



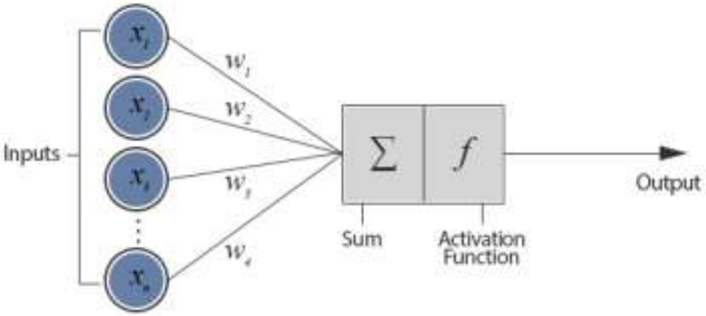
internal representation



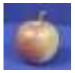
prediction



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



function regression

f() = “apple”

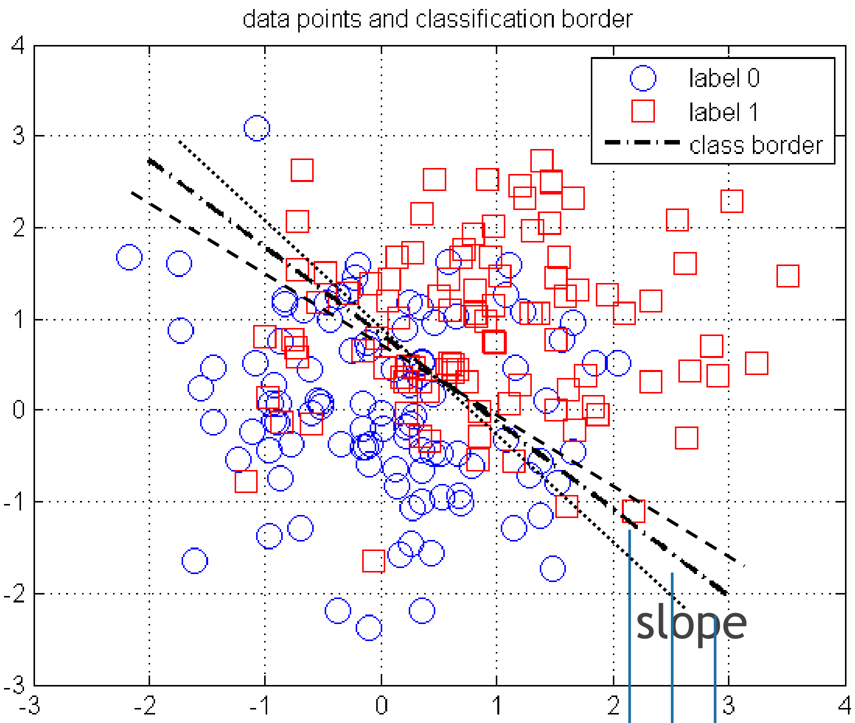
f( ) = “pear”

f( ) = “cow”

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

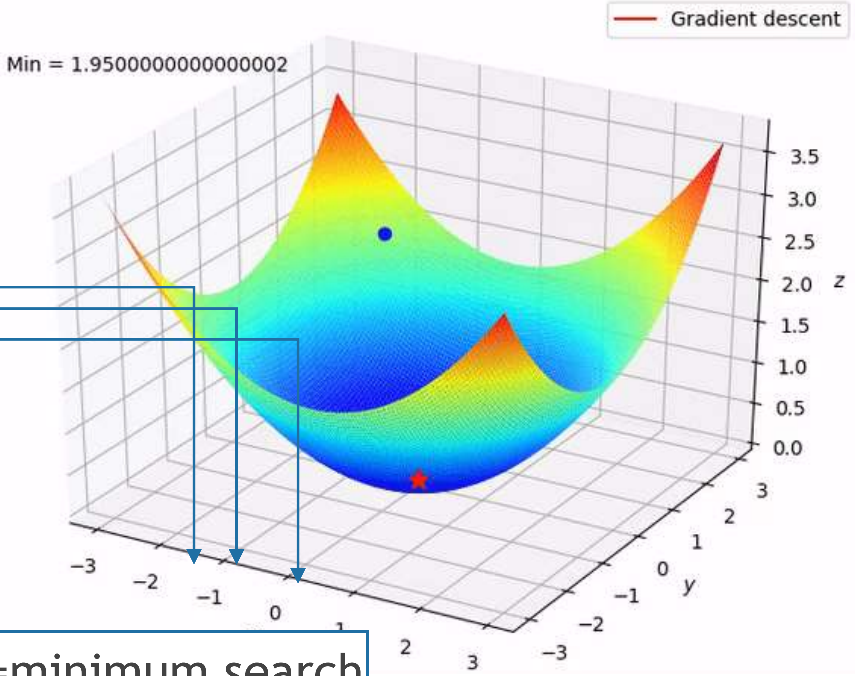
Learning -> loss function optimization

images -> points in N dim space



slope

Loss = number of wrong categorizations

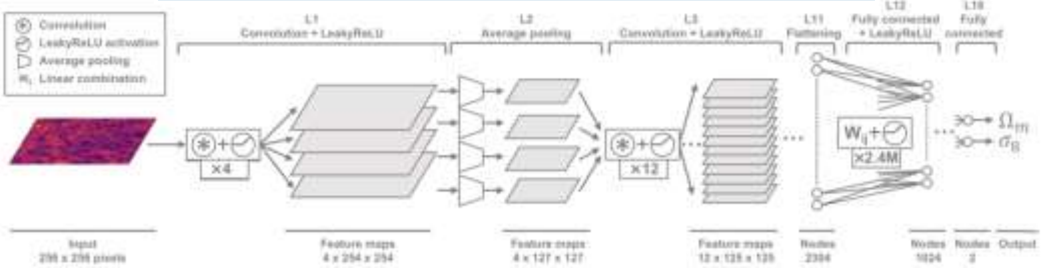


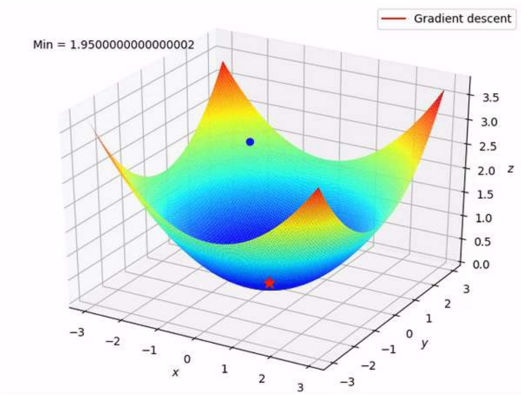
Learning=minimum search

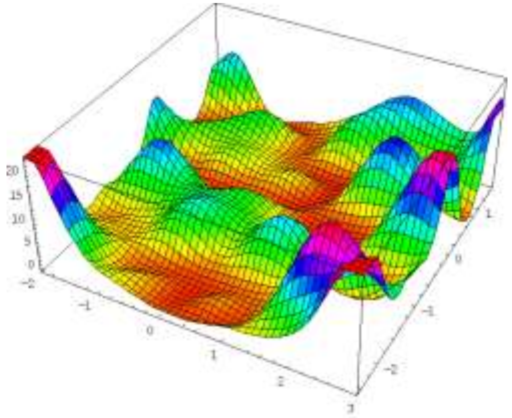
Challenges

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

Typical network: 2M adjustable parameter

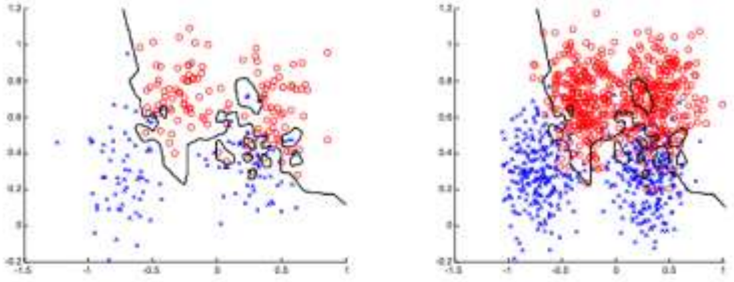






Training data

Testing data

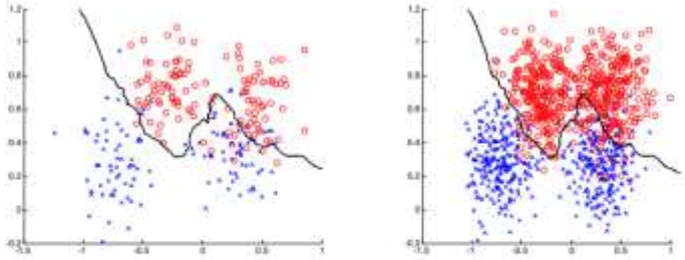


error = 0.0

error = 0.15

Training data

Testing data

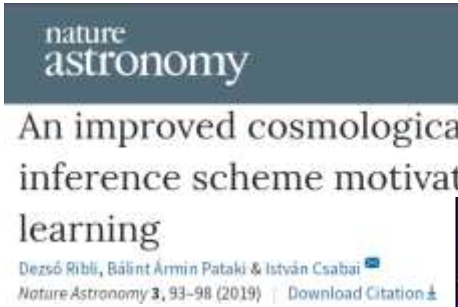


error = 0.1120

error = 0.0920

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**
  Palla et al. subm.
- **Pathology** images
  SOTE TKP collab.
- **Quantum neural computing**
- **MSc, PhD courses**
  http://datascience.elte.hu

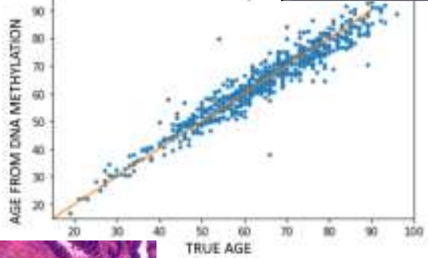


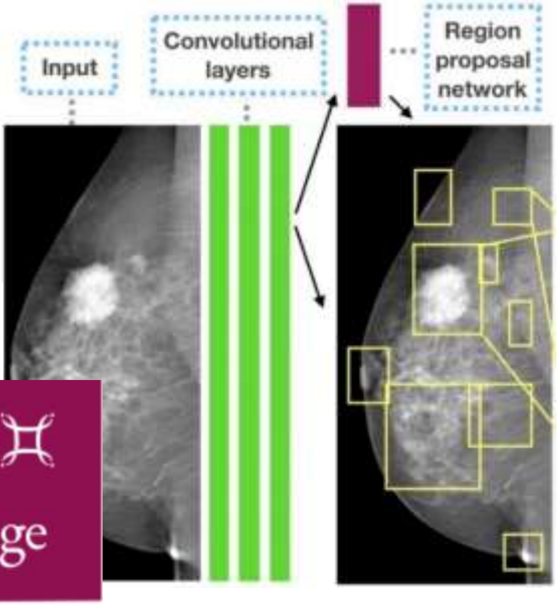
nature astronomy

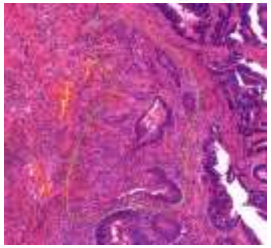
An improved cosmological parameter inference scheme motivated by deep learning

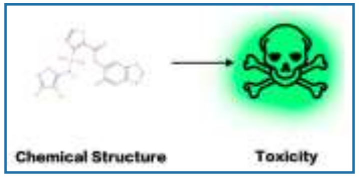
Dezső Ribli, Bálint Ármin Pataki & István Csabai

Nature Astronomy 3, 93–98 (2019) | Download Citation

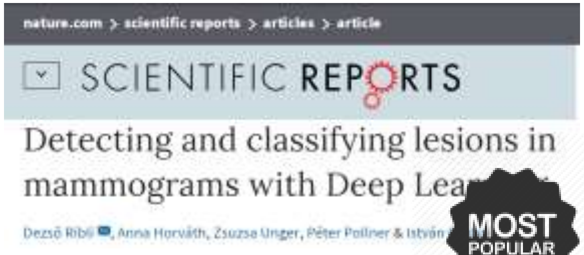








Solving analytically untraceable hard inverse problems



nature.com > scientific reports > articles > article

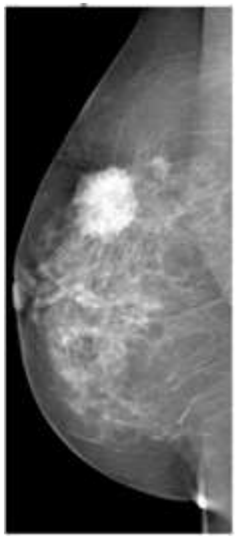
SCIENTIFIC REPORTS

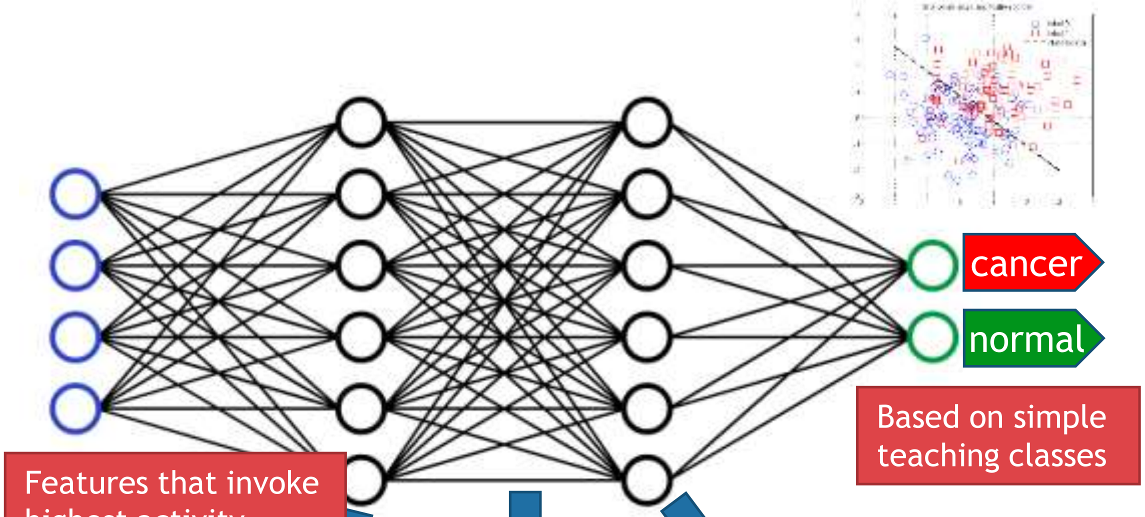
Detecting and classifying lesions in mammograms with Deep Lea

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

MOST POPULAR

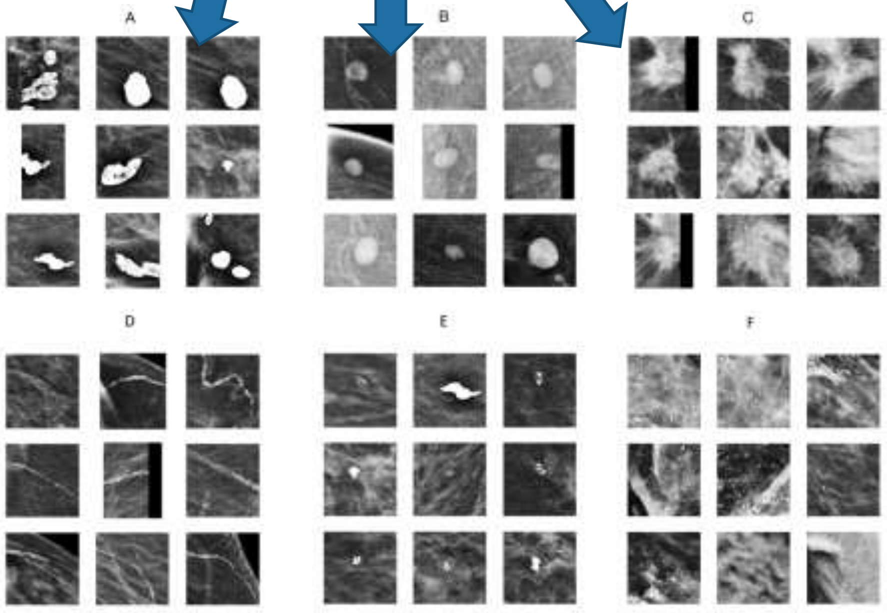
Explainable AI: automatic classification enhancement





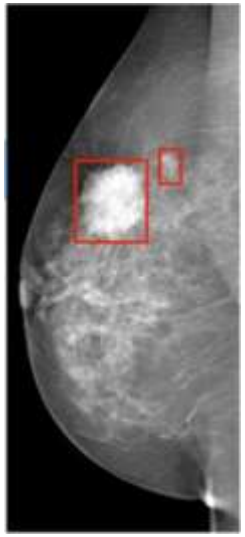
Based on simple teaching classes

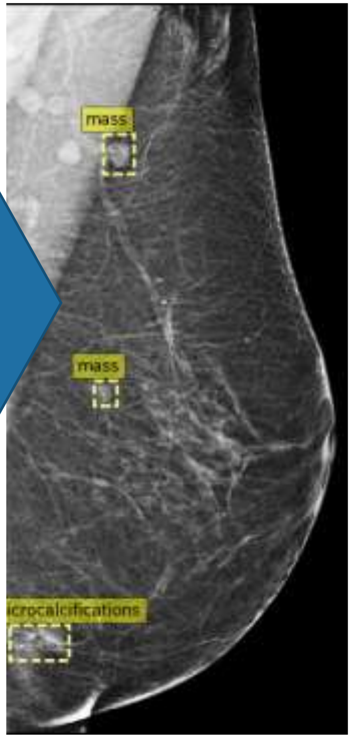
Features that invoke highest activity



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 , 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 **high-throughput equipment** 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.

ITN

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/

Nemzeti Kutatási, Fejlesztési
És Innovációs Hivatal

AZ NKFI ALAPBÓL
MEGVALÓSULÓ
PROJEKT