MACHINE LEARNING IN SCIENCES

István Csabai

ELTE Dept. of Physics of Complex Systems

July 12, 2019

EFOP-3.6.3-VEKOP-16-2017-00001

History of (machine) intelligence / data science

Model

OBSERVATION

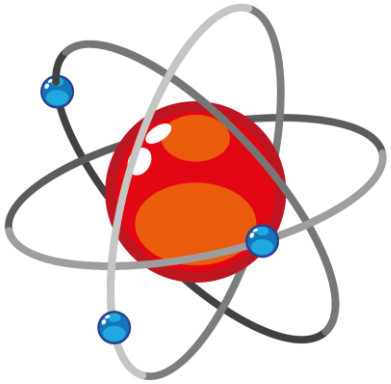
World

History of (machine) intelligence / data science

Model

OBSERVATION

World



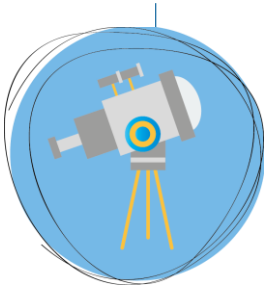
History of (machine) intelligence / data science

crutch for senses

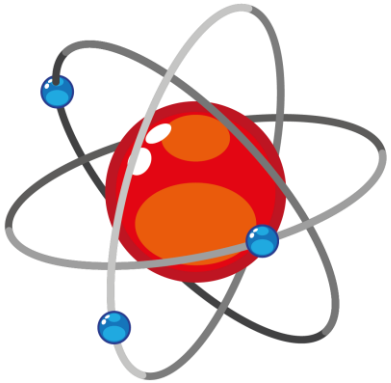
Model

OBSERVATION

Instruments



World



History of (machine) intelligence / data science

Model

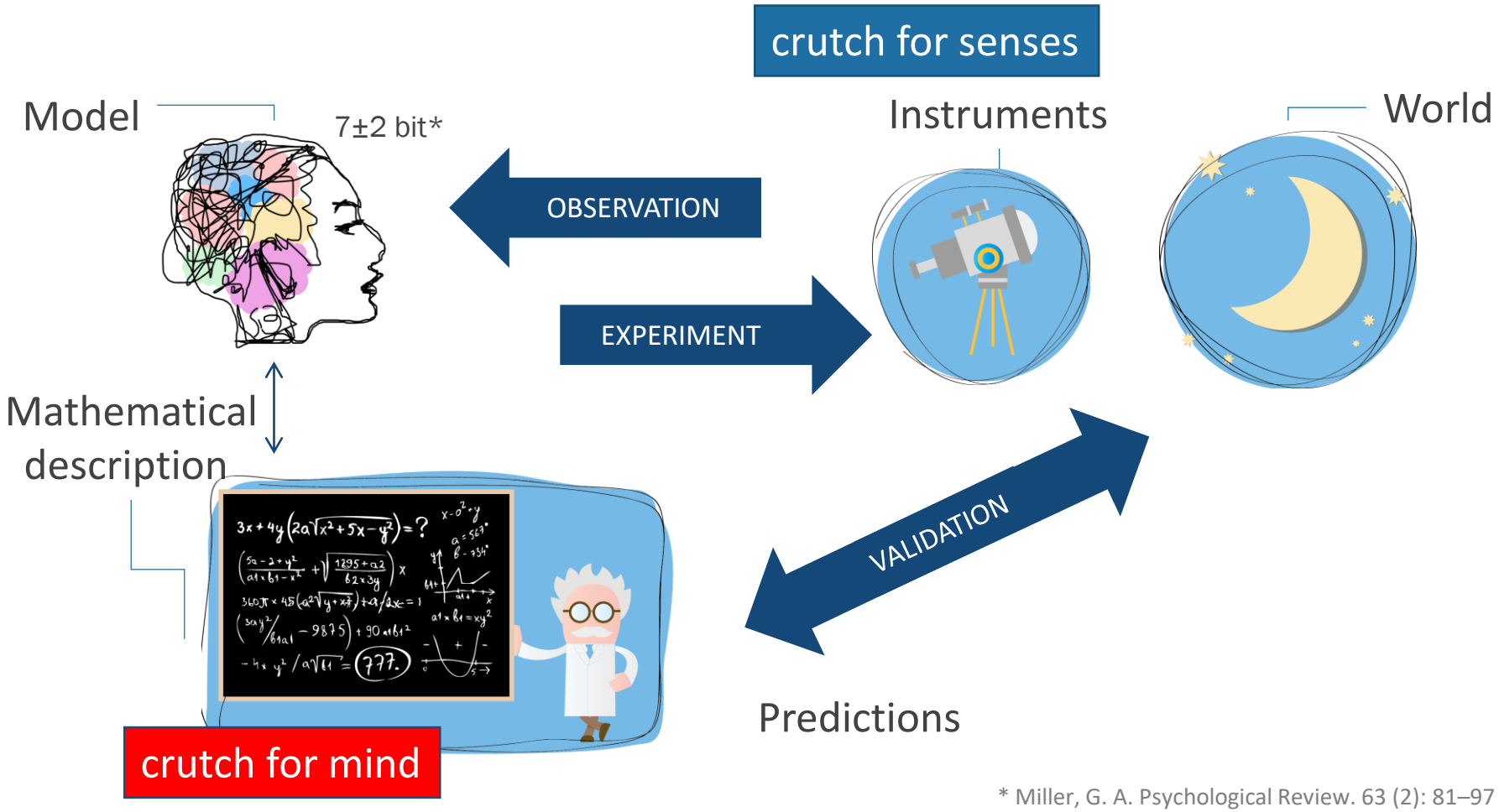
7±2 bit*

OBSERVATION

World

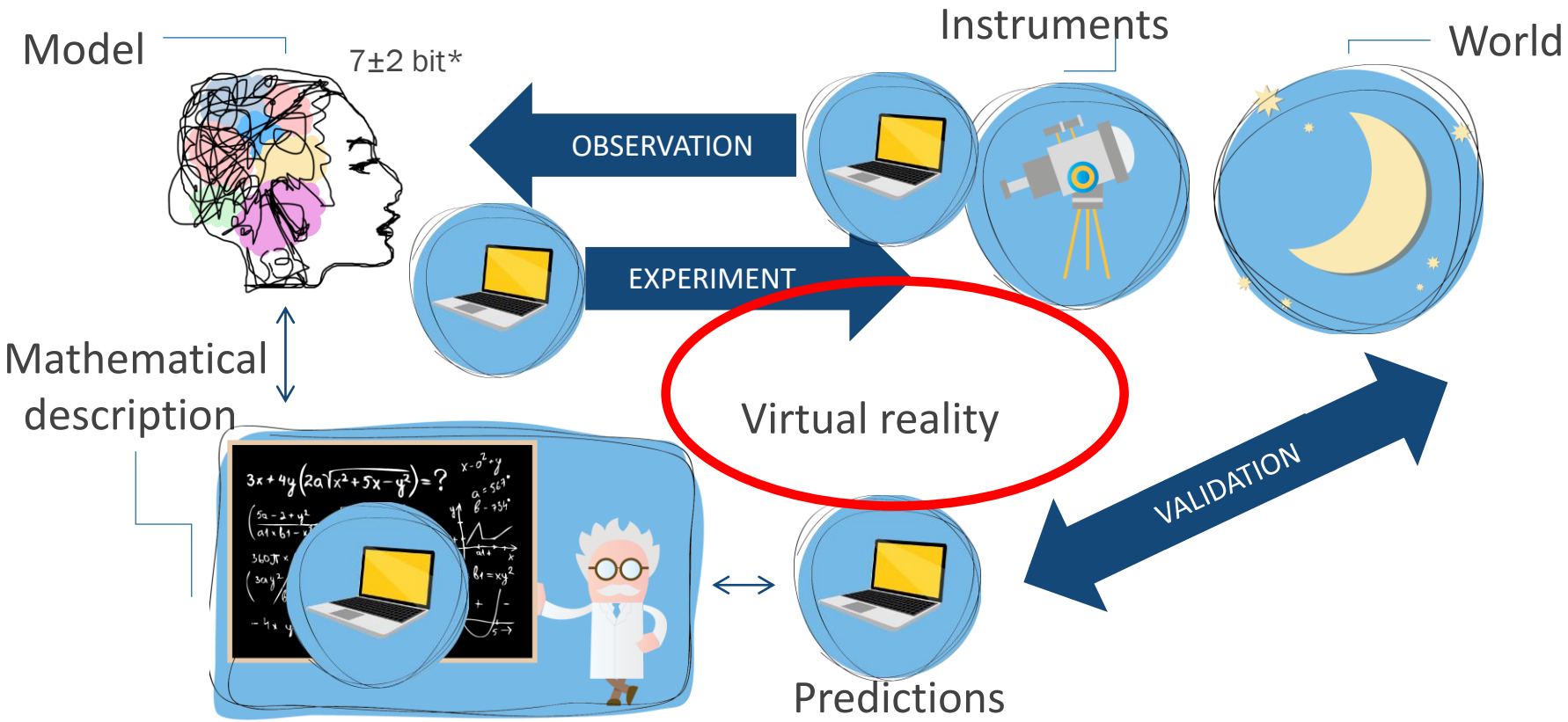
* Miller, G. A. Psychological Review. 63 (2): 81–97 (1956)

History of (machine) intelligence / data science



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

Modern data science



Initial values

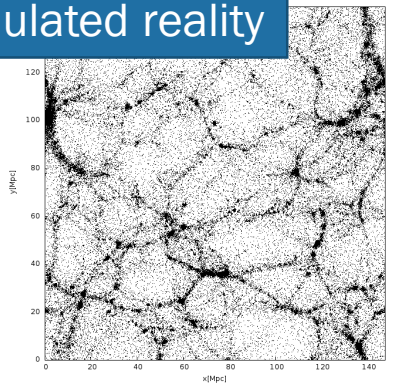
Λ=0.7
Ω_m=0.3

"laws", equations

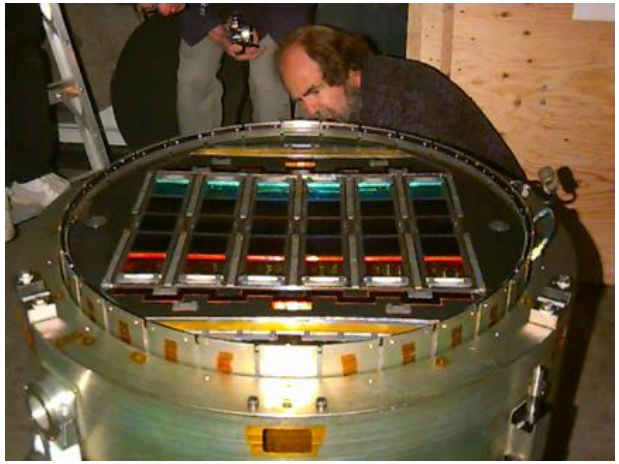
$$F = G\frac{m_1 m_2}{r^2}$$

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

Simulated reality



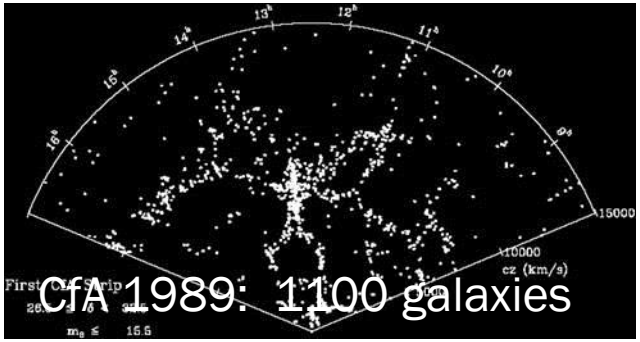
2.5m 120Mp – 2.5Tp 5 years:10TB



150 man-years software dev.

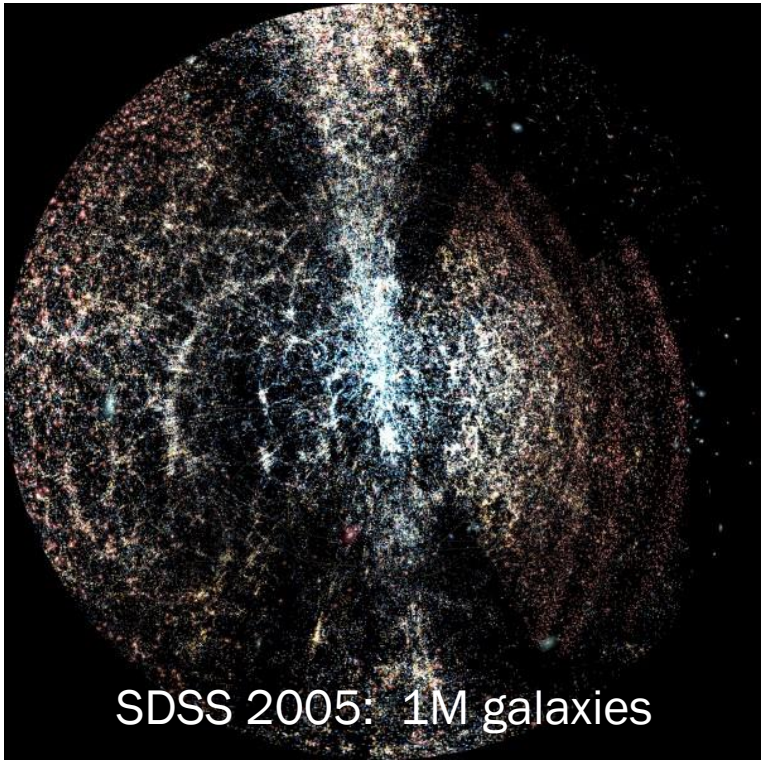
1995-
First "big data" science:
3D MAP OF THE UNIVERSE

8Mhz CPU, 640KB mem, 10GB HDD



1929: 1 galaxy

CfA 1989: 1100 galaxies



SDSS 2005: 1M galaxies

LSST: By the numbers

8 meter
wide-field survey
telescope

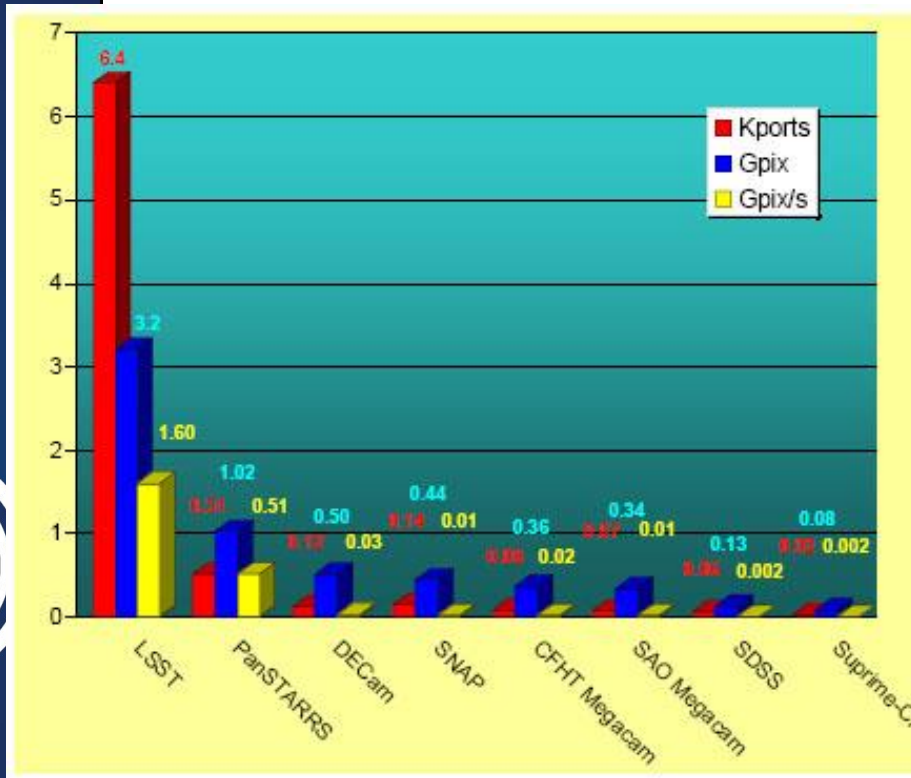
3 billion pixel
digital camera
(largest in the world)

3
mirror construction

30
terabytes
of data per night

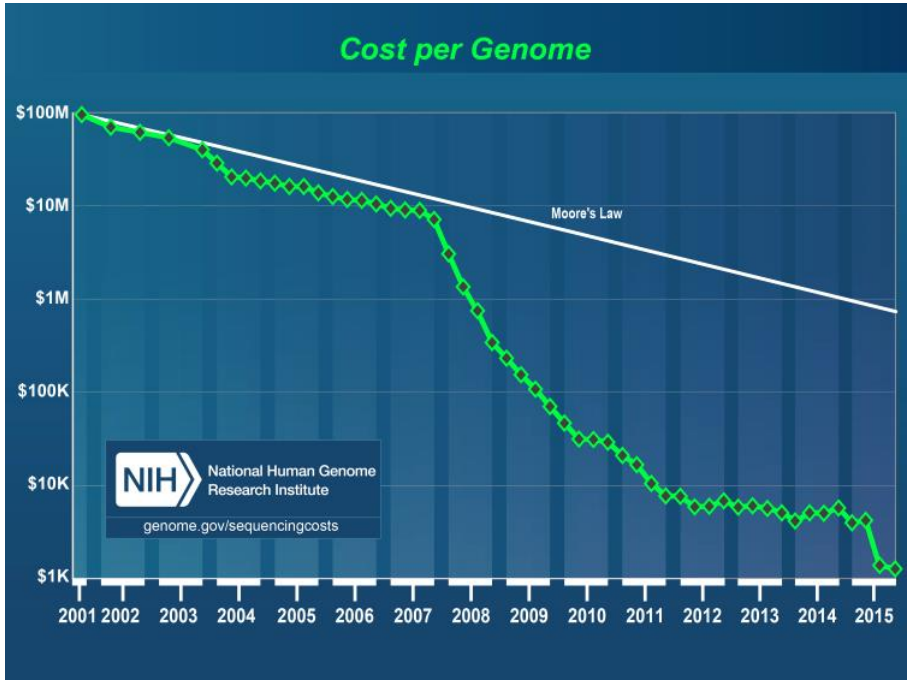
NSF's Large Synoptic Survey Telescope will image the entire visible sky a few times each week for 10 years and is expected to see first light in 2019.

NSF NATIONAL SCIENCE FOUNDATION



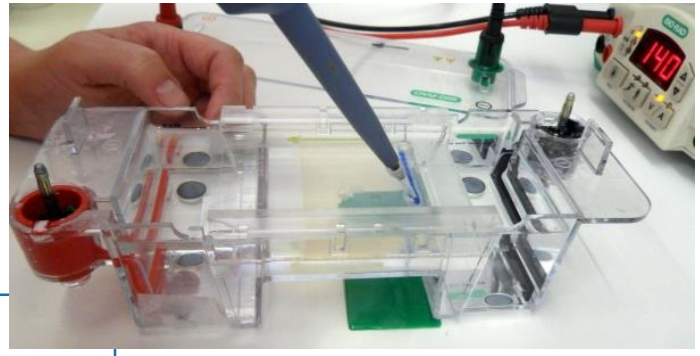
SDSS 5 years =
LSST 2 days
2020

Moore's law in gene sequencing



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

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

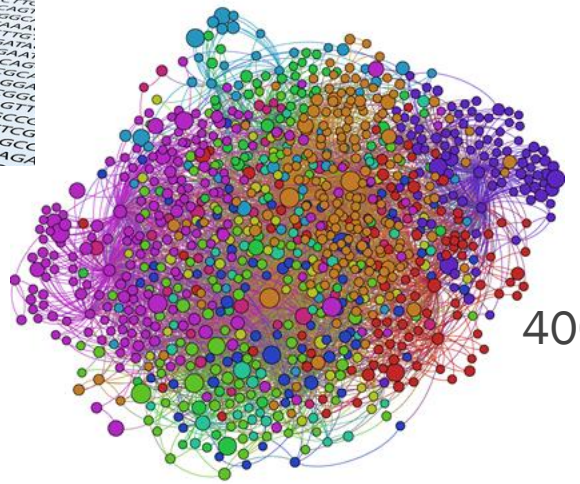


Oxford Nanopore 100Mb,$900

Key challenges: amount of data and complexity of models

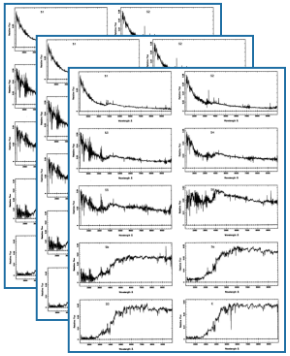
2.5Tpx

10kx3.2 G bp

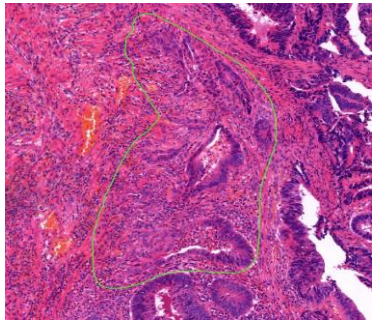


400M tx

1Mx3k

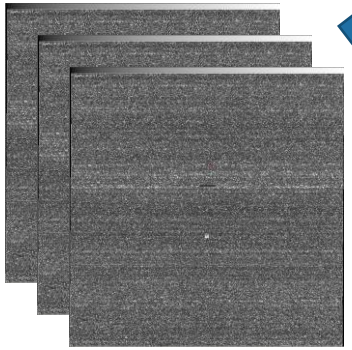


7±2 bit



0.5M x 0.5M

10kx30k

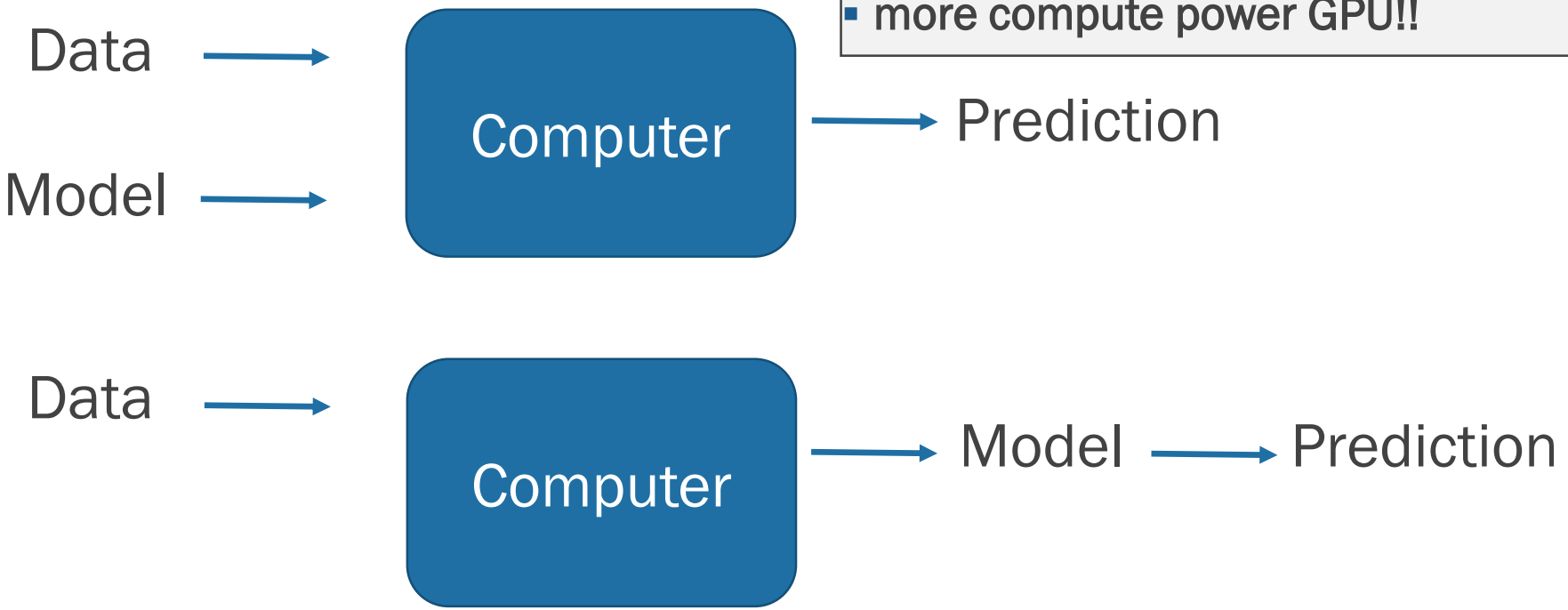


Machine learning to rescue

Machine learning paradigm shift

Why now?

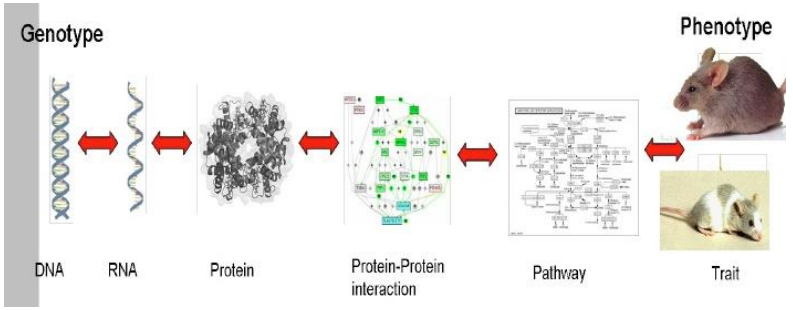
- more data (MNIST'98 60k, CIFAR'10 60k, IMAGENET'10 14M)
- steadily improving models, deeper understanding of statistics/data/models
- **more compute power GPU!!**



Looks like a magic black box, but you need to understand the details and the limitations!

Recent ML projects @ Dept. of Complex Systems, Eotvos Univ.

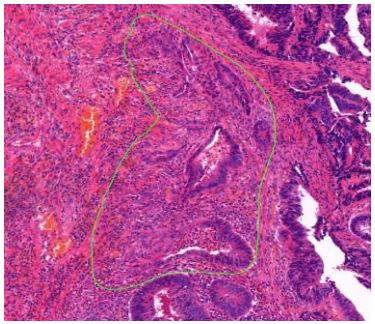
- Mutations -> antibiotics resistance
  Matamoros et al. in prep.
- Mobile sensors -> Parkinson
  Pataki @DREAM, Laki et al. 2016
- Quantum wave func.-> drug toxicity
  Biricz et al. in prep.
- X-ray image -> breast cancer
  Ribli et al. @DREAM, Sci. Rep. 2018
- Weak lensing map -> cosmology params
  Ribli et al. Nature Astro. 2018
- Explainable AI
  Ribli et al. in prep , Patent subm. 2018
- Pathology image analysis



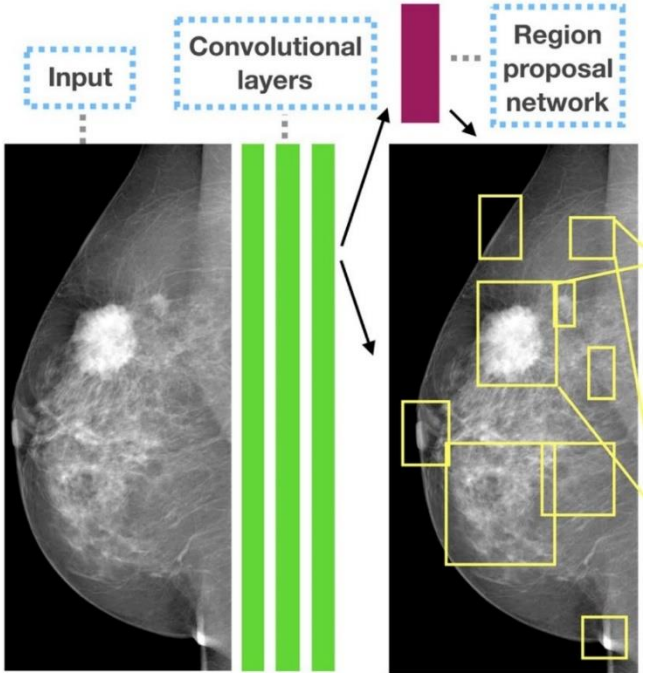
NIH

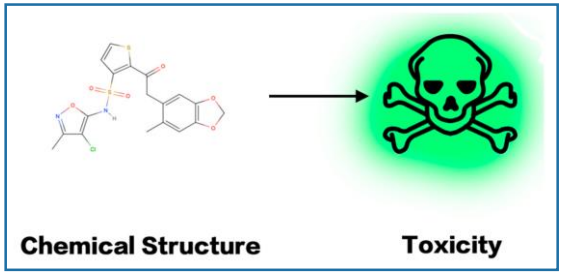
DREAM CHALLENGES

Sage



Analytically untraceable hard inverse problems





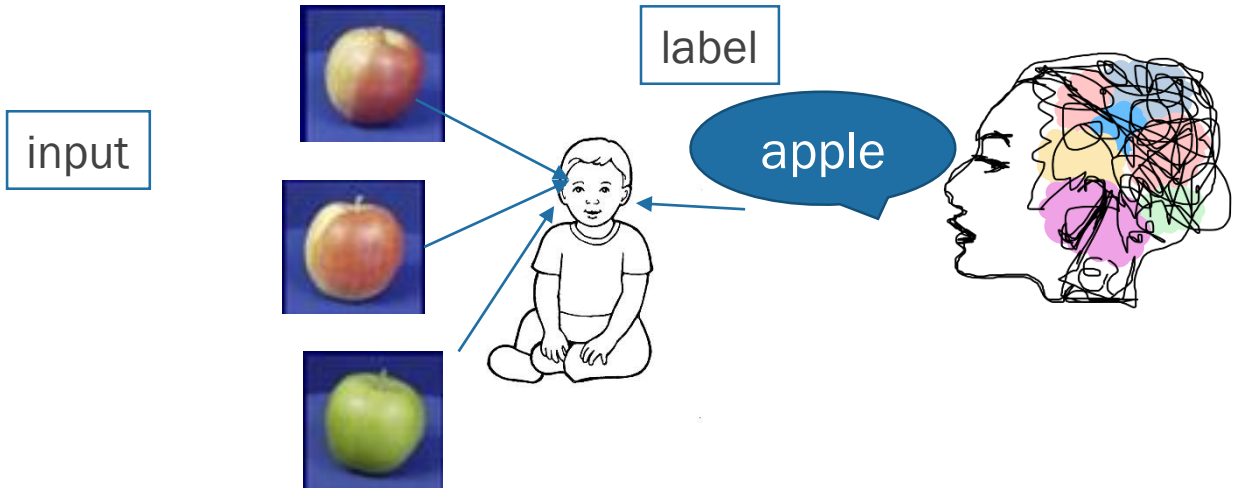
nature.com > scientific reports > articles > article

SCIENTIFIC REPORTS

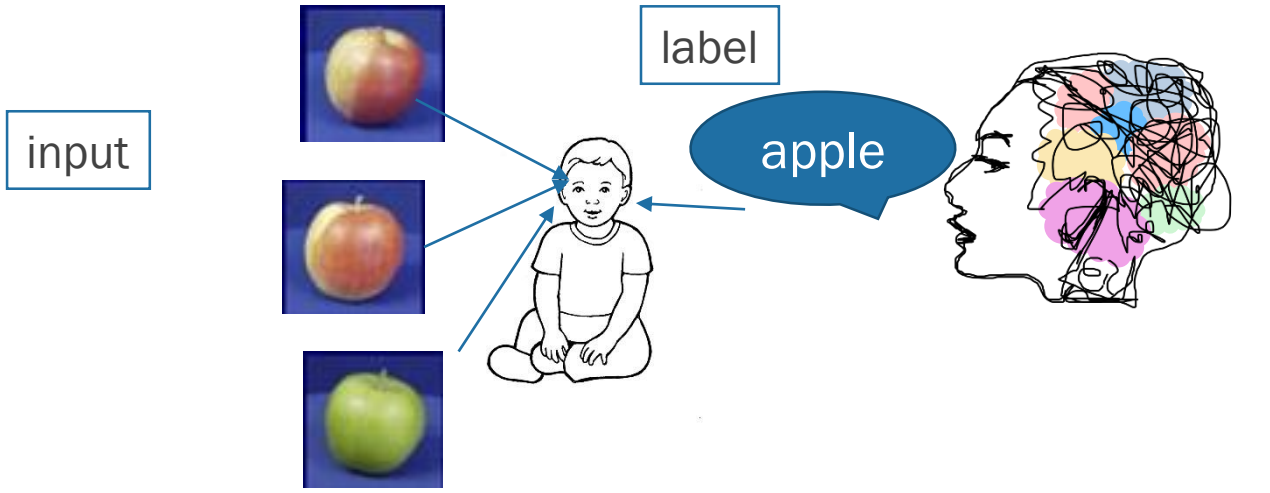
Detecting and classifying lesions in mammograms with Deep Learning

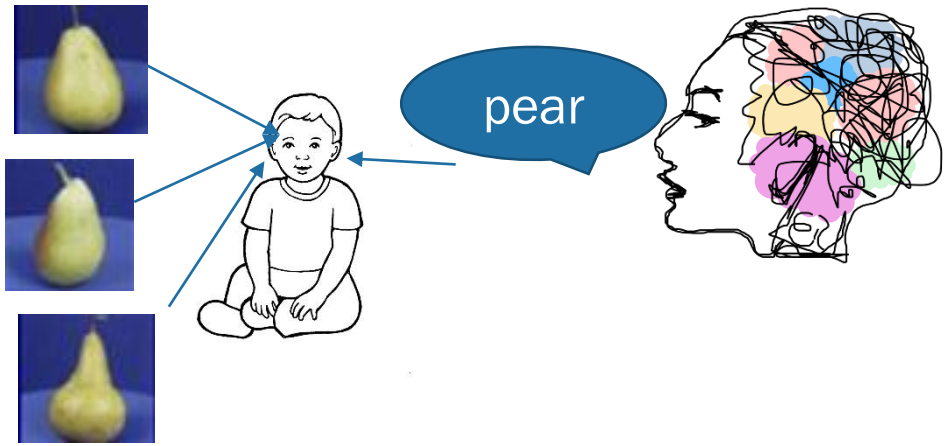
Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner & István Csabai

Supervised learning – quick introduction

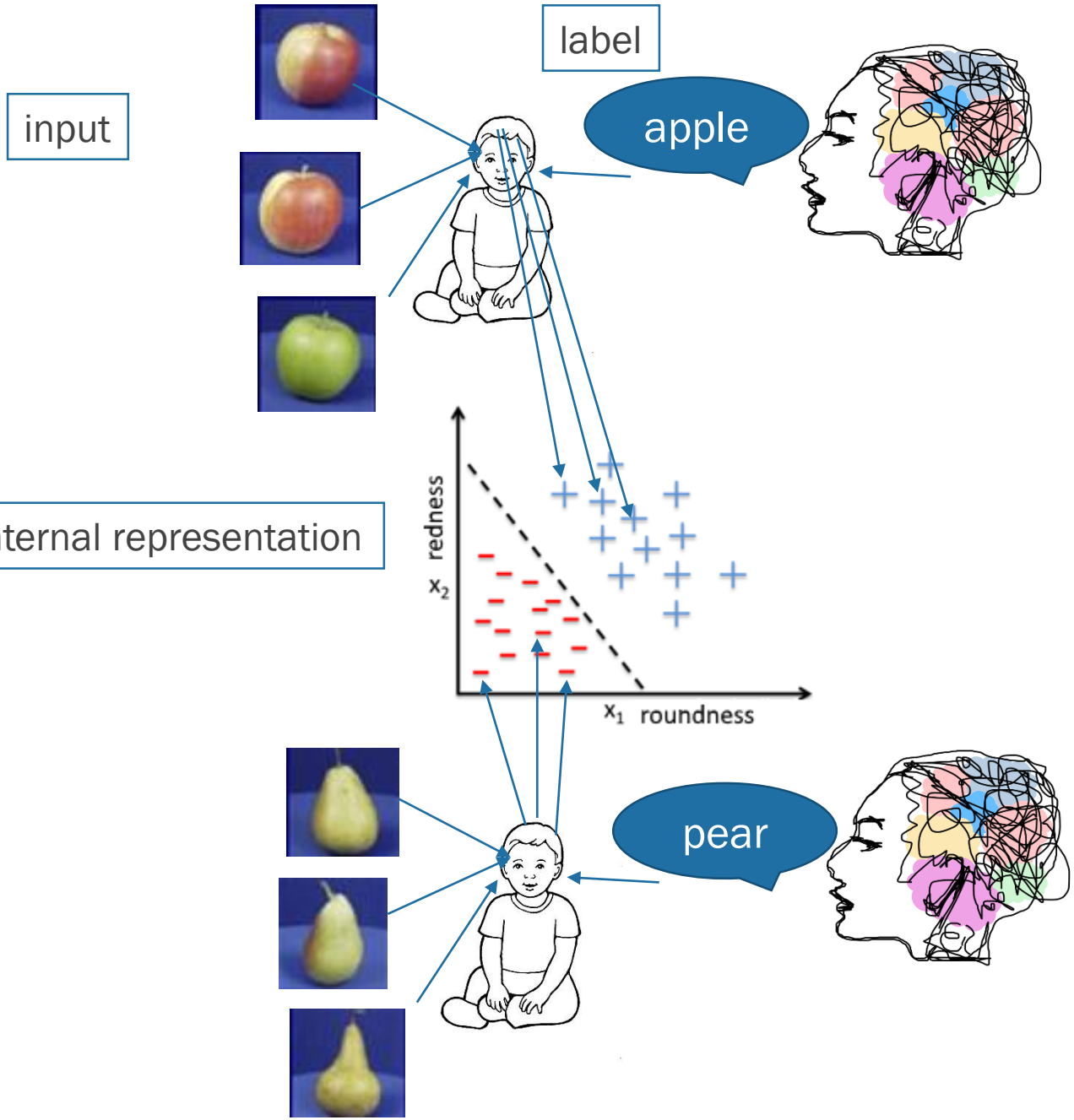


Supervised learning

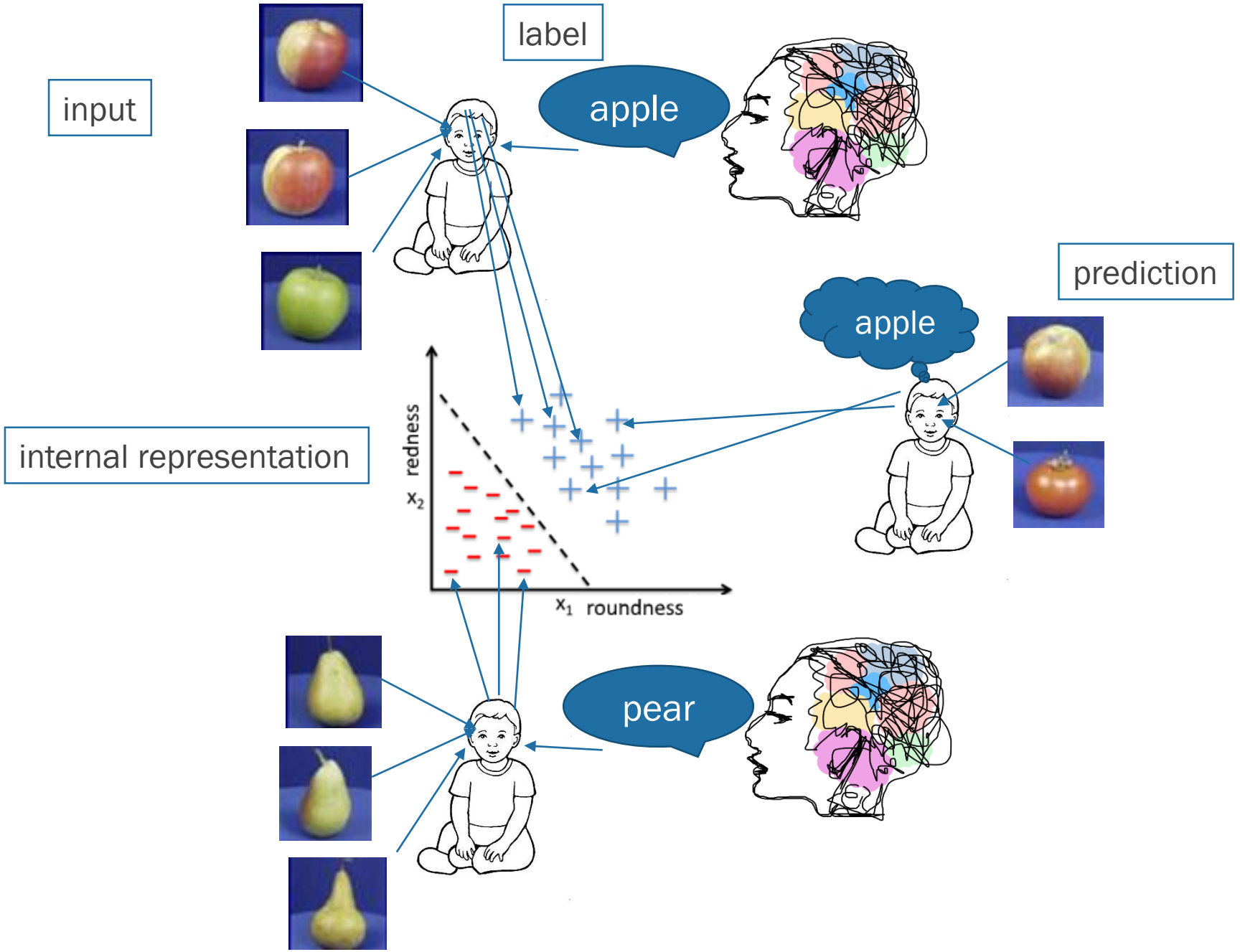




Supervised learning



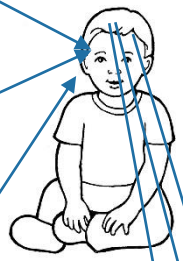
Supervised learning



Supervised learning

label

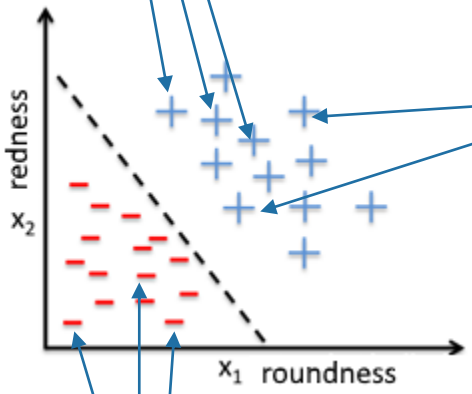
input



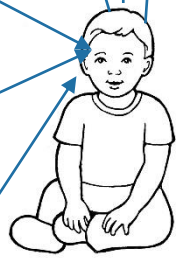
prediction

apple

internal representation



function regression



f( ) = “apple”

f( ) = “pear”

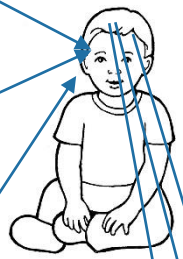
f( ) = “cow”

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

Supervised learning

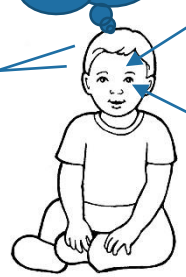
label

input

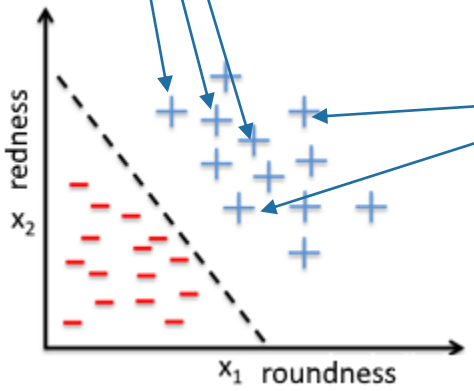


prediction

apple



internal representation



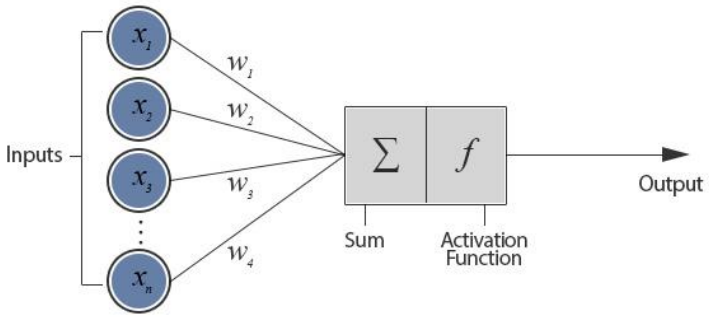
$$y = f\left(\sum_i w_i x_i\right)$$

function regression

f( ) = “apple”

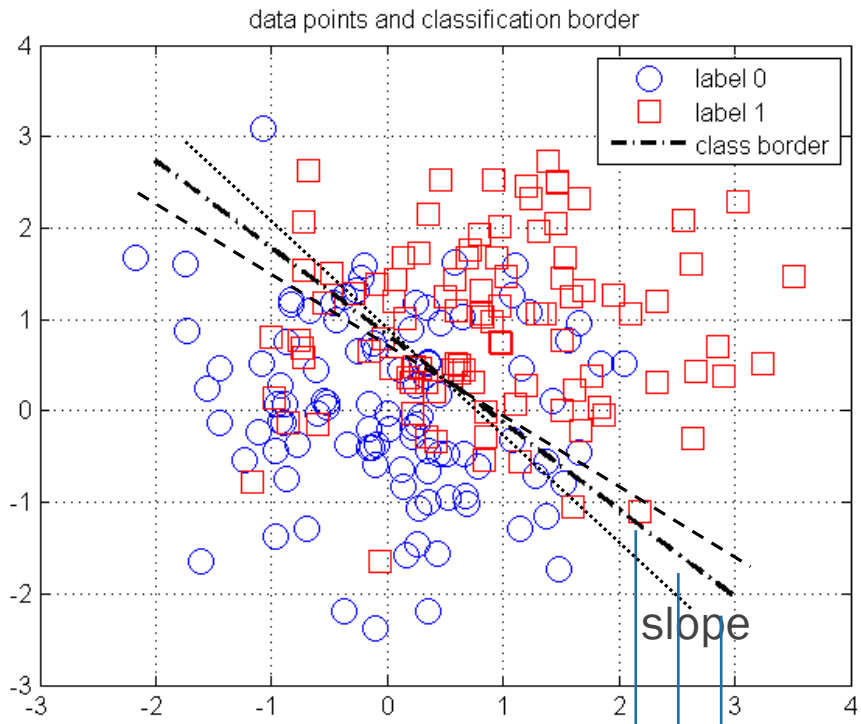
f( ) = “pear”

f( ) = “cow”



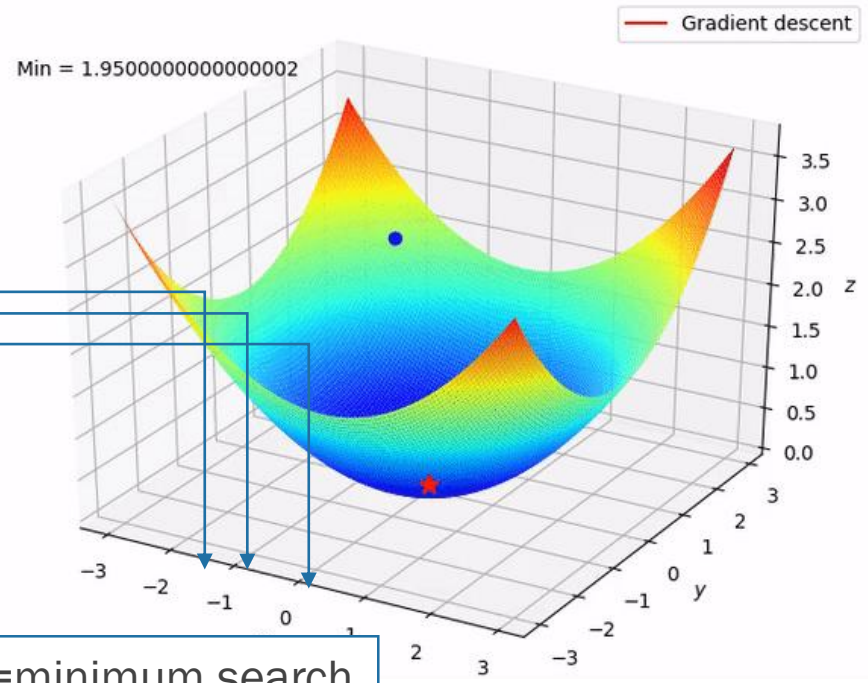
IF color=red AND profile=smooth THEN type:=tomato
IF color=red AND HAS(horns) THEN type:=cow

Learning -> loss function optimization



images -> points
in N dim space

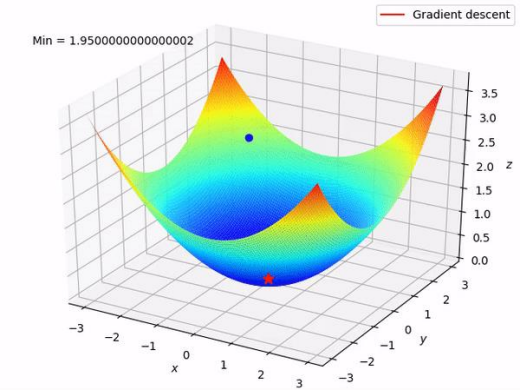
Loss = number of wrong categorizations

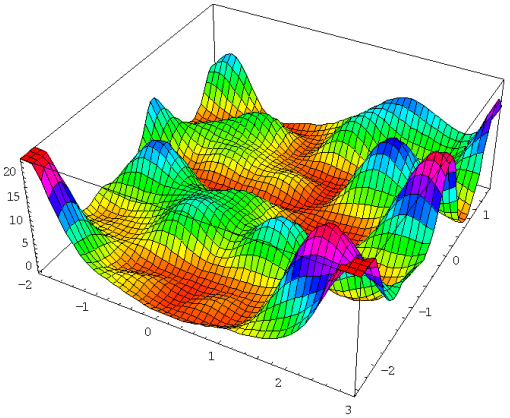


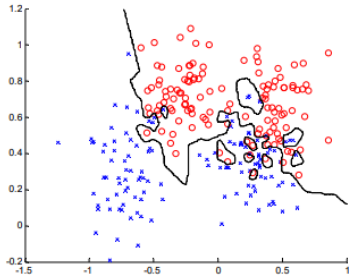
Learning=minimum search

Challenges

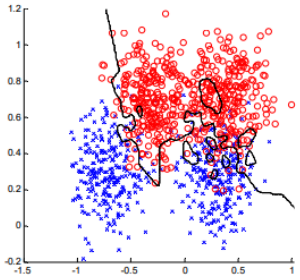
- Proper training set
- Representation of data (images, words, … -> vector space)
- Nonlinear optimization
- Model complexity
  - Accuracy
  - Generalization
- “Black box”, trust
- …



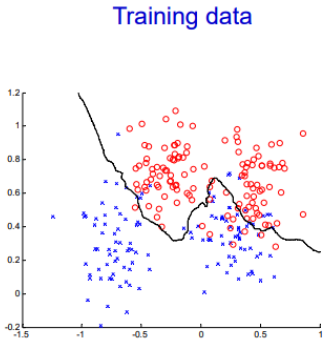




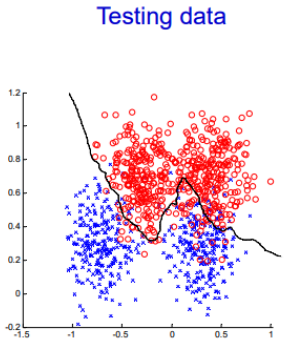
error = 0.0



error = 0.15

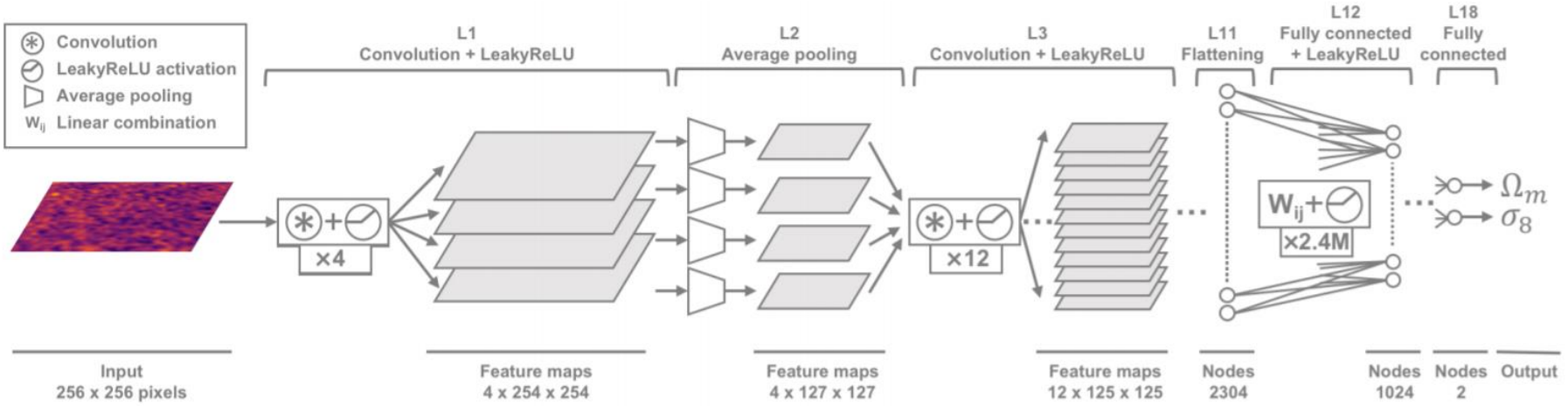


error = 0.1120



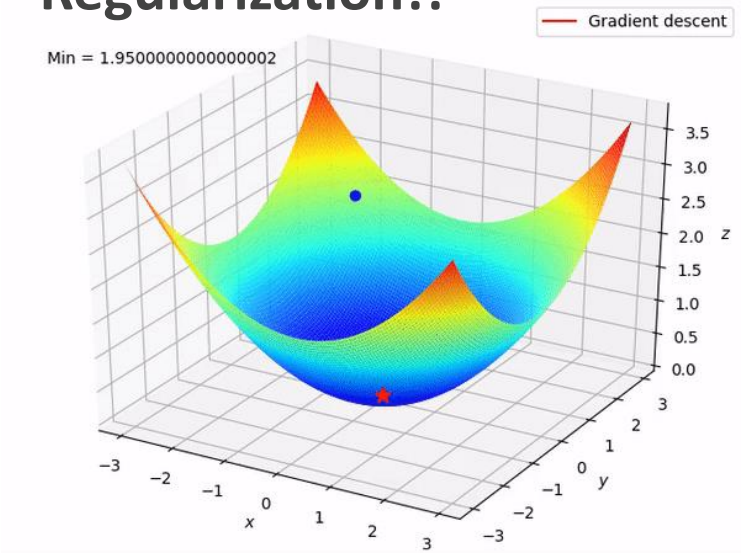
error = 0.0920

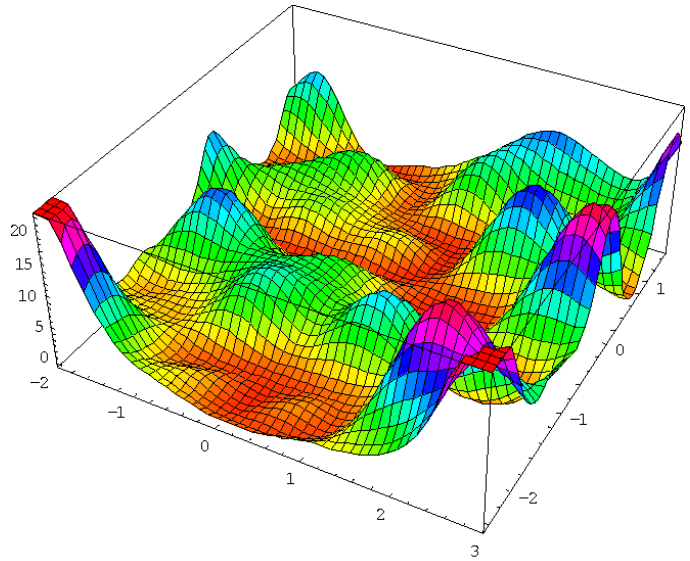
Typical complex multilayer network example



- Number of optimization parameters (dimensions): > 2 million

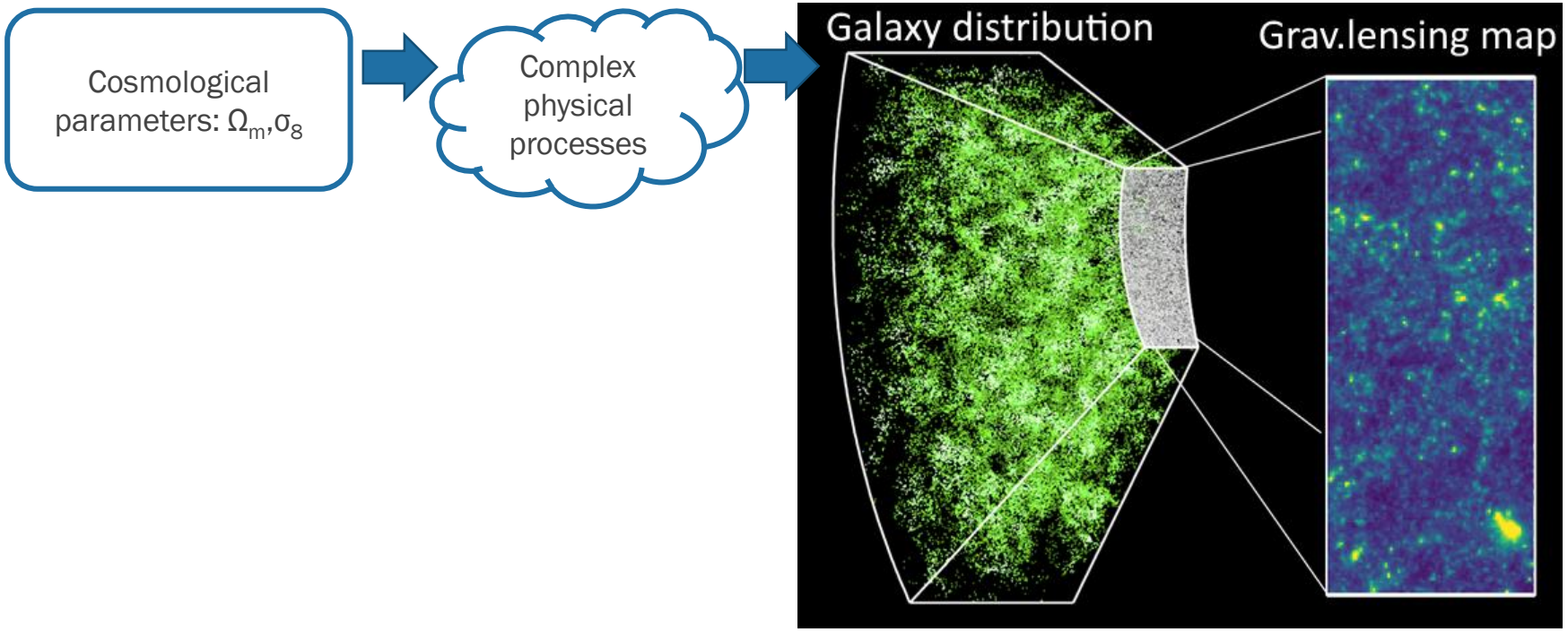
Regularization!!



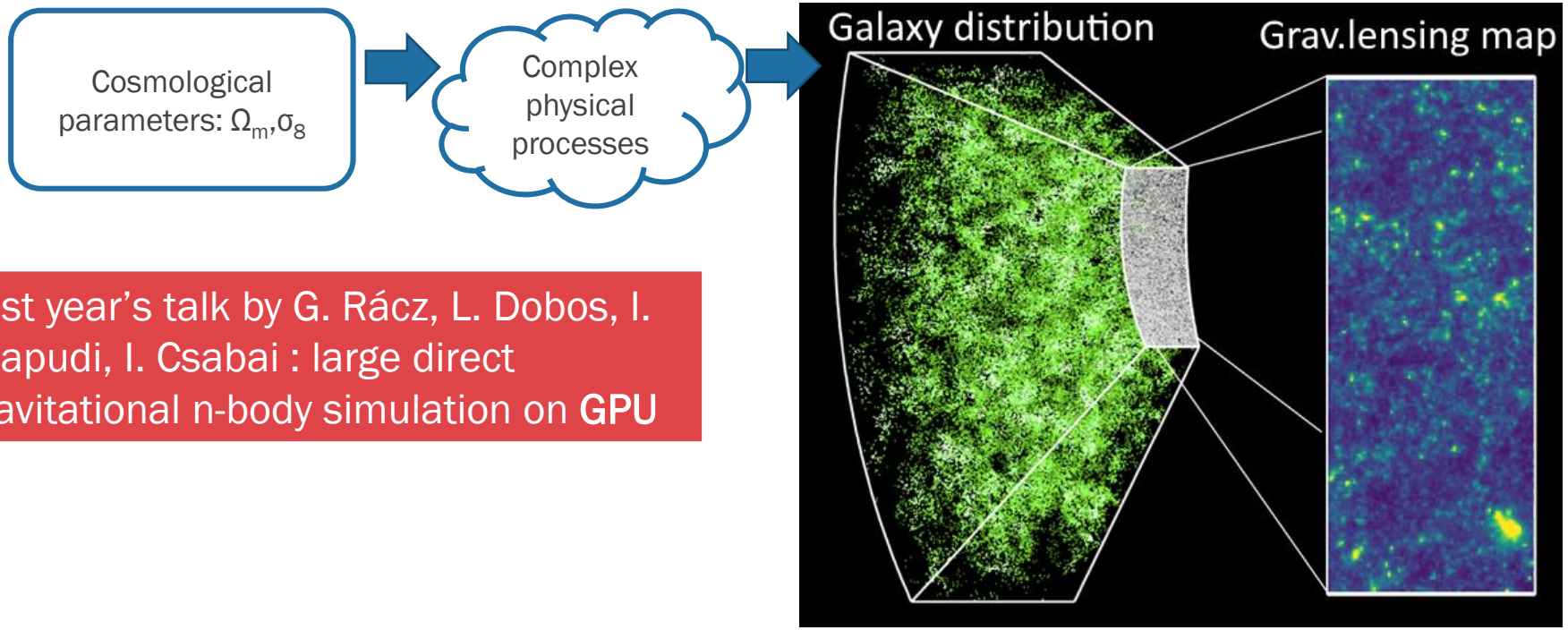


Imagine this with 2 000 000 dimensions!

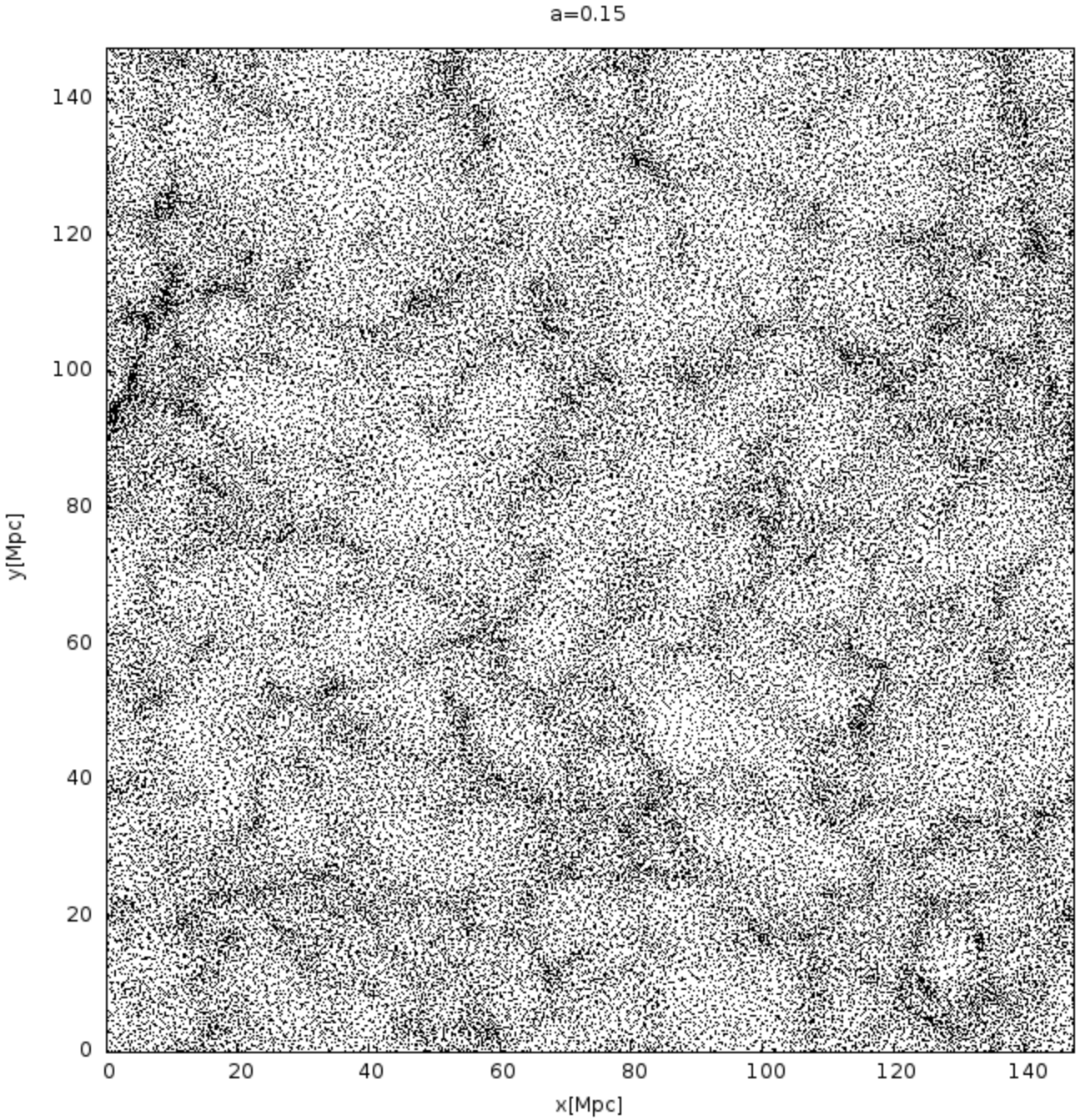
Cosmological parameters from gravitational lensing
Learning new tricks from deep learning

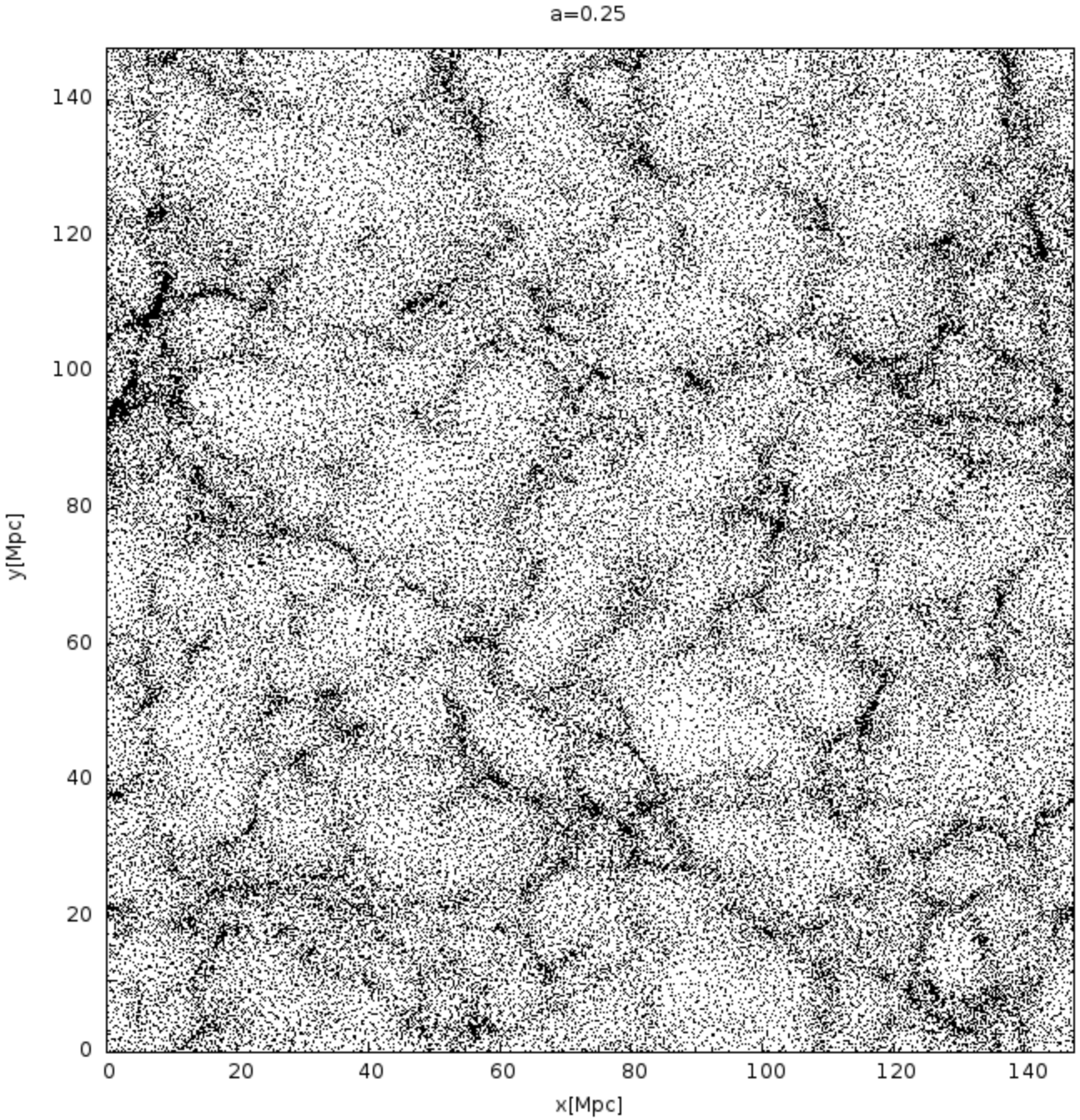


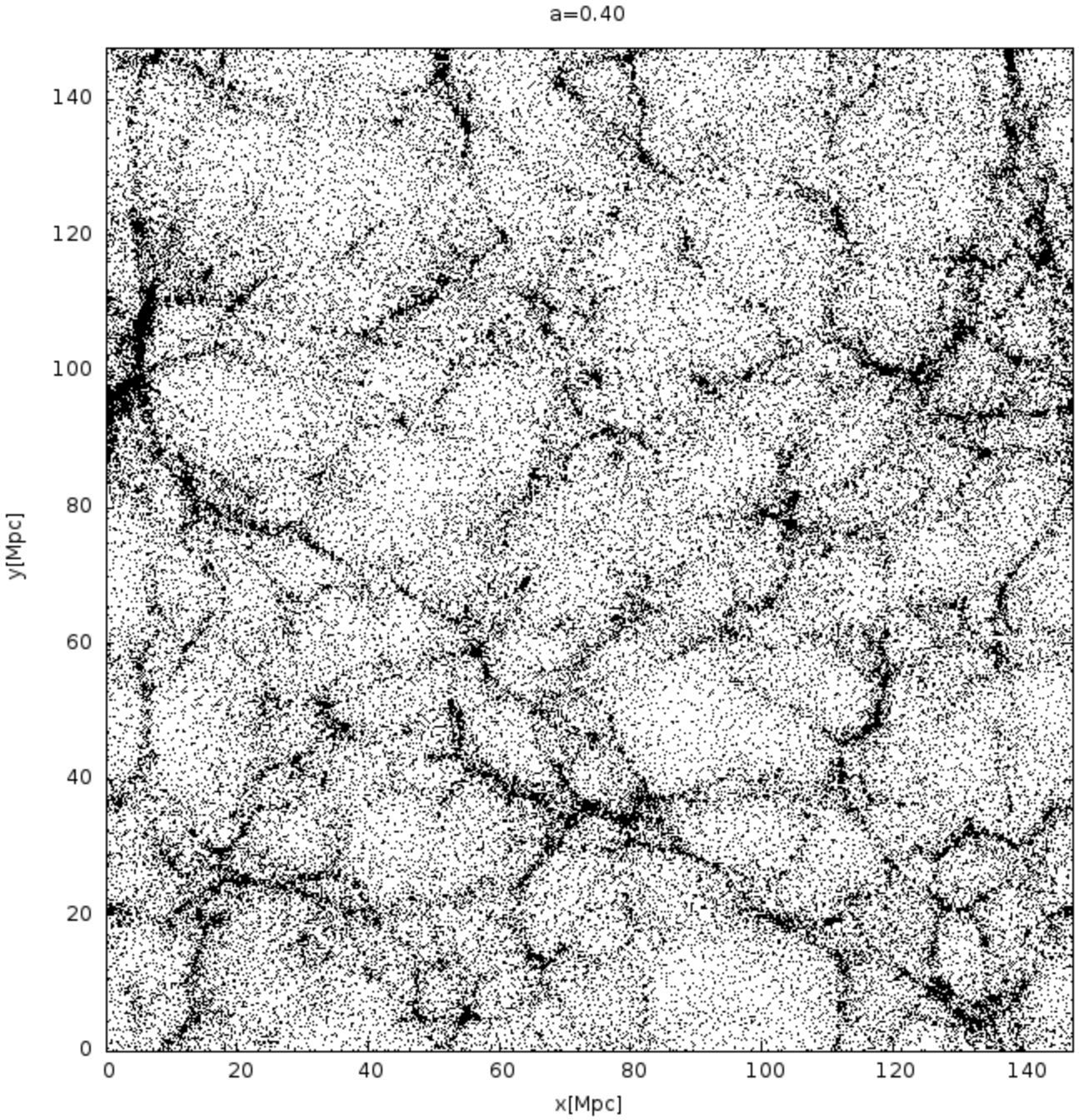
Cosmological parameters from gravitational lensing Learning new tricks from deep learning

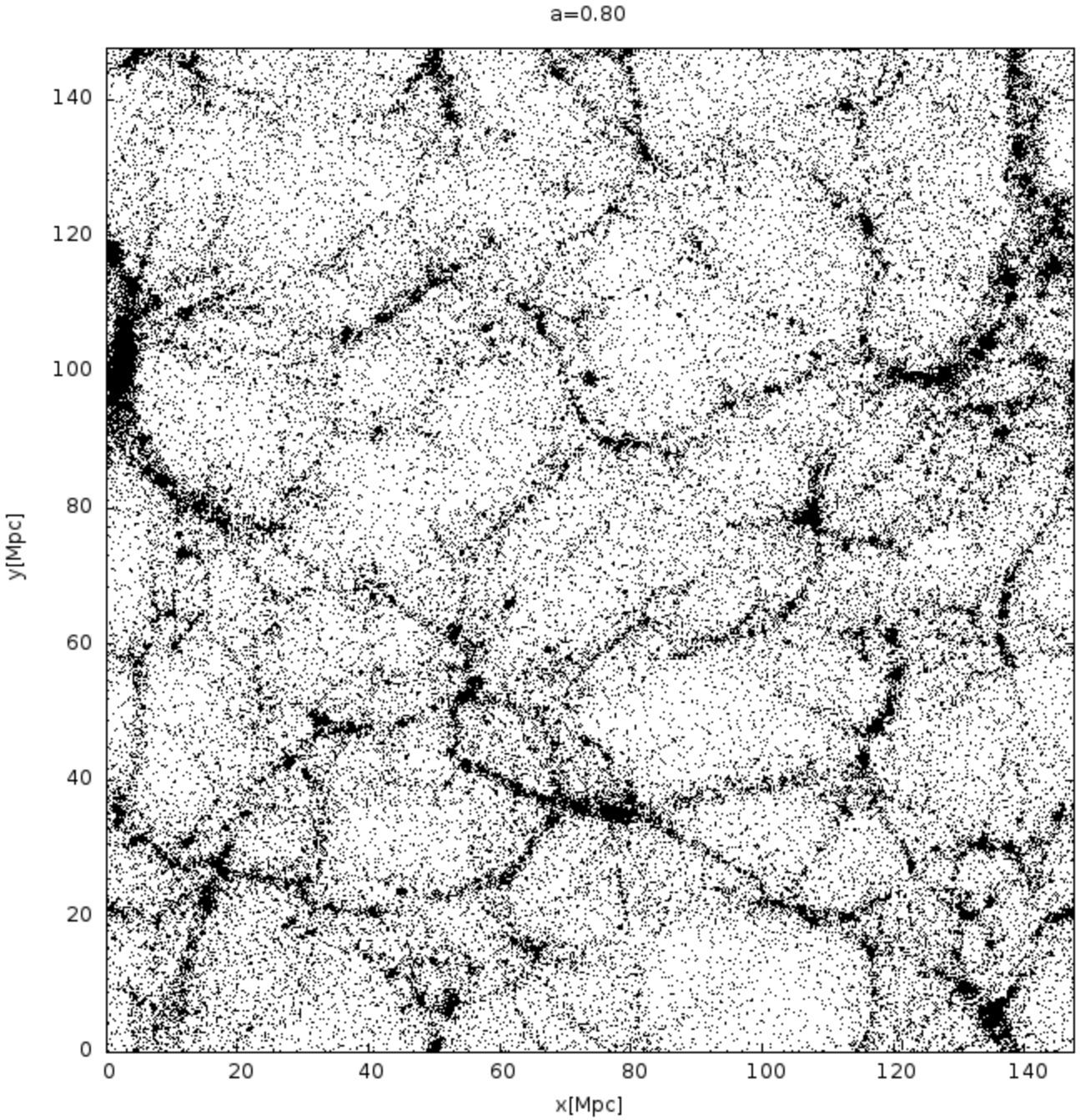


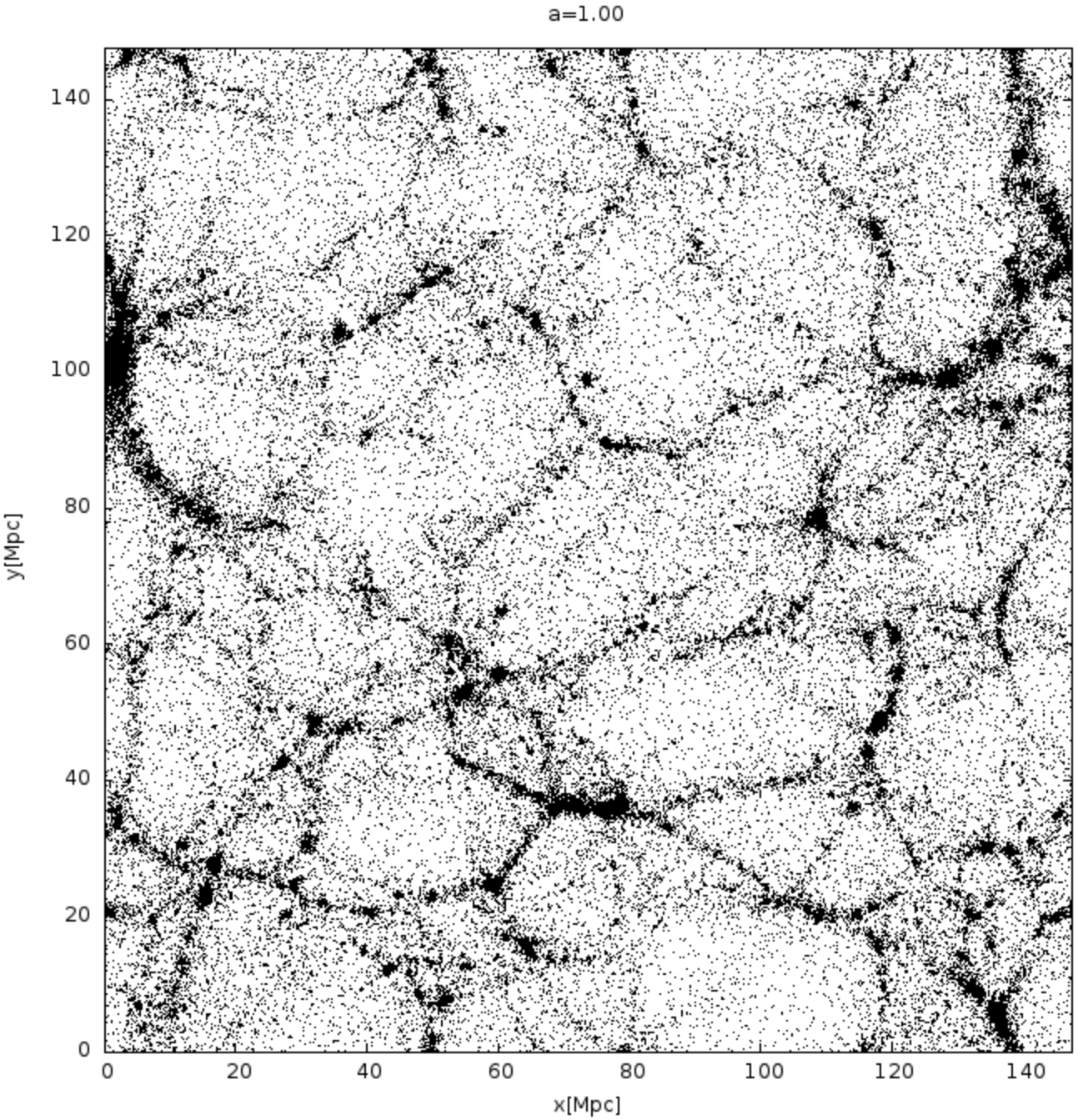
Last year's talk by G. Rácz, L. Dobos, I. Szapudi, I. Csabai : large direct gravitational n-body simulation on **GPU**


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

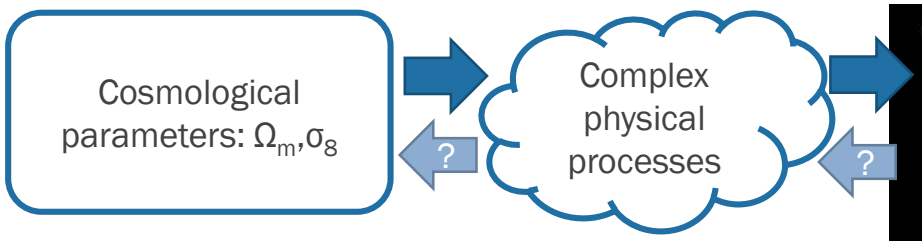

a=0.25
y[Mpc]
x[Mpc]
0
20
40
60
80
100
120
140


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

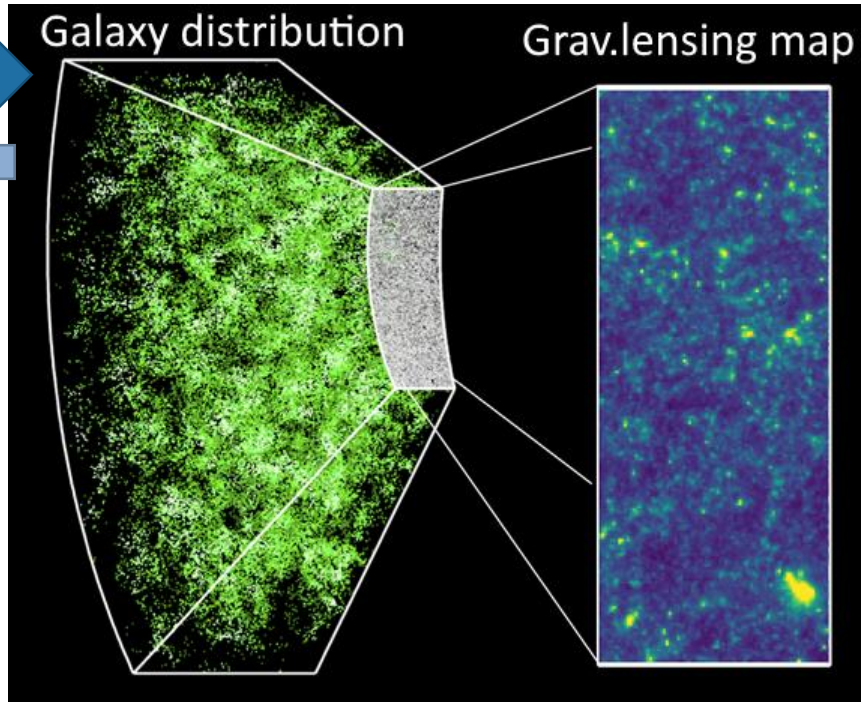

a=0.80
y[Mpc]
x[Mpc]
0
20
40
60
80
100
120
140


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

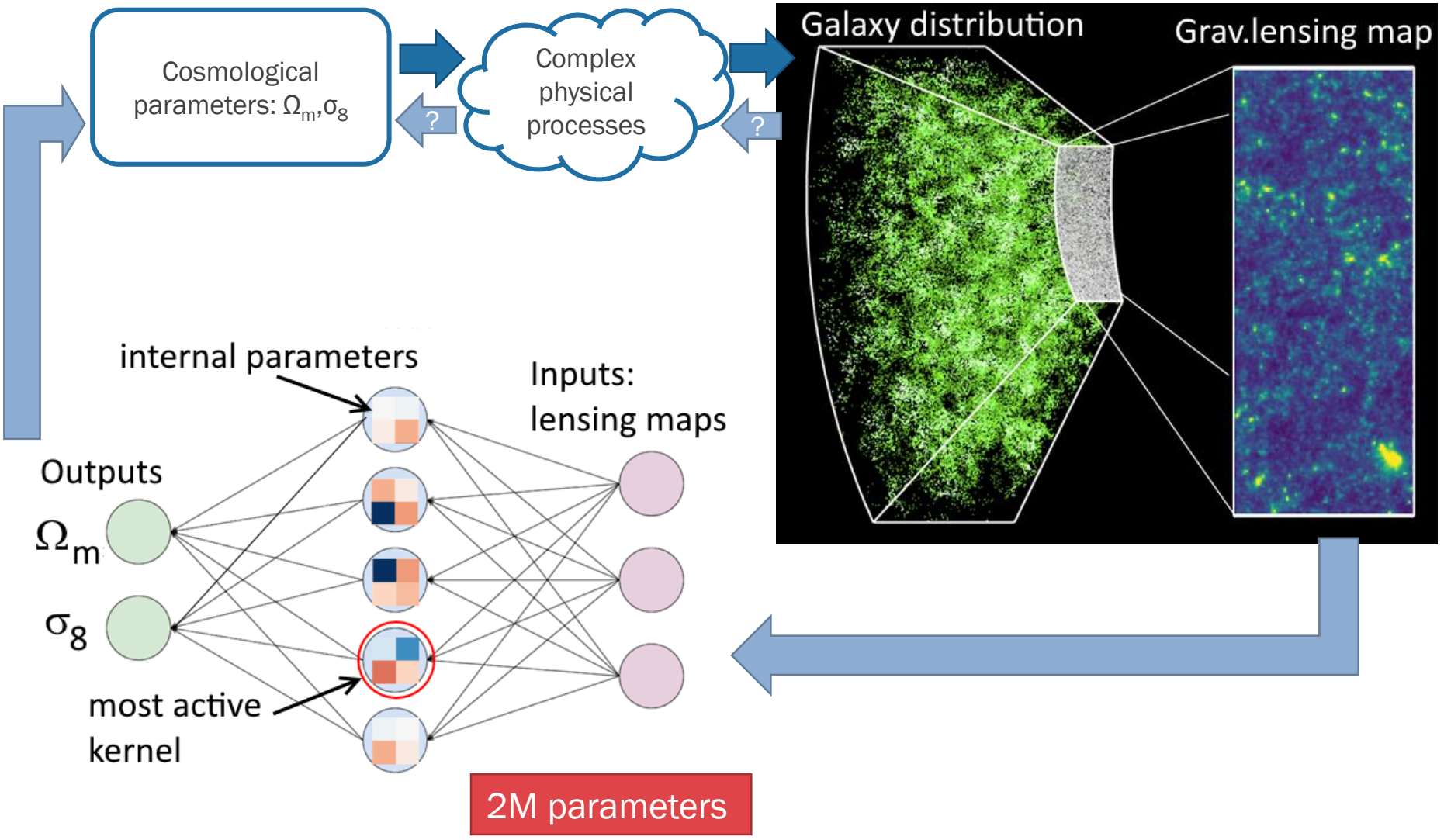
Cosmological parameters from gravitational lensing
Learning new tricks from deep learning



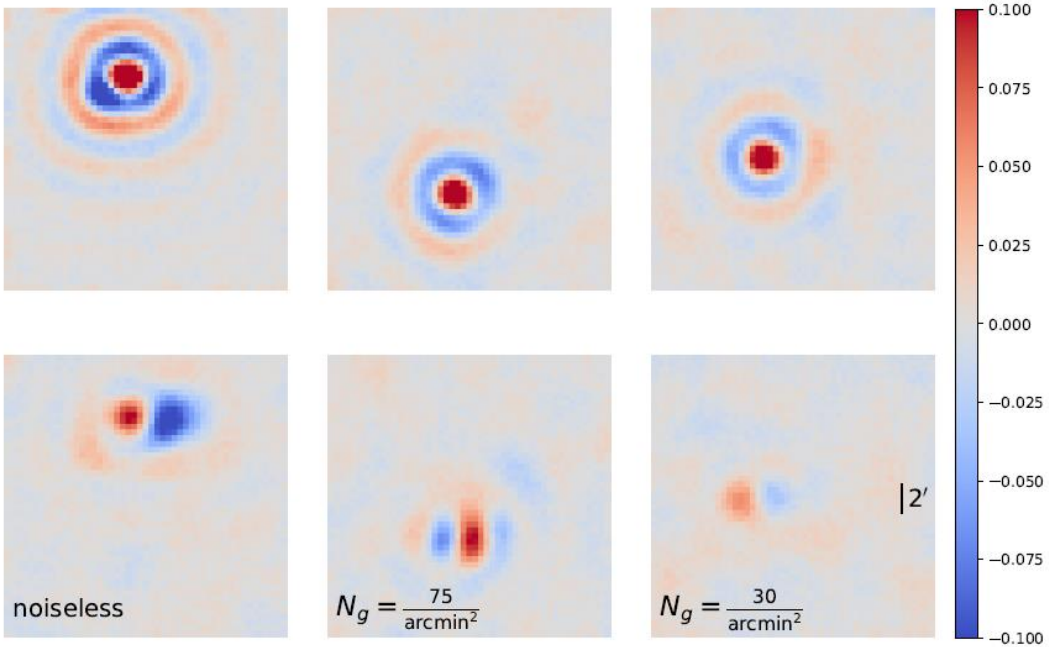
Hard inverse problem!



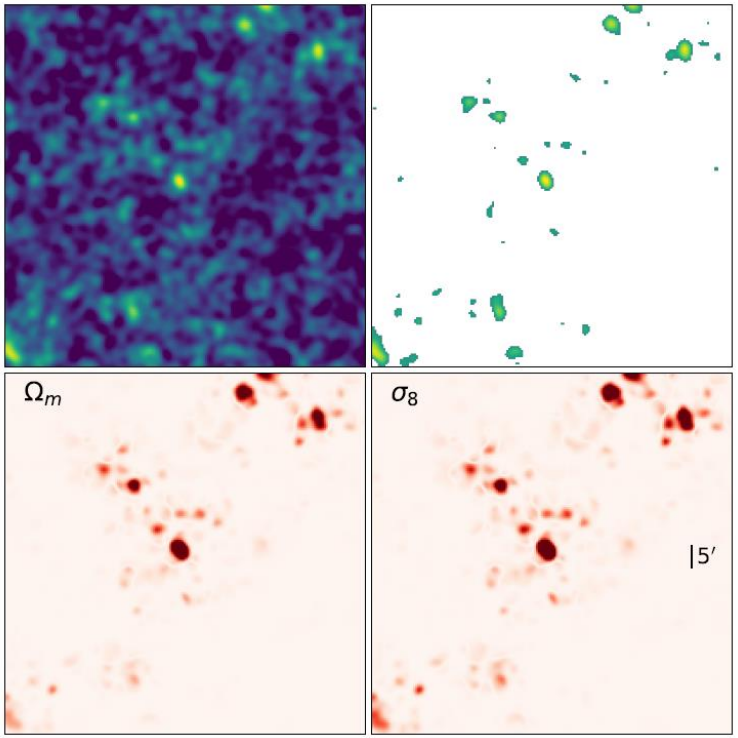
Cosmological parameters from gravitational lensing
Learning new tricks from deep learning



Learned kernels: dark matter halo profile expansion

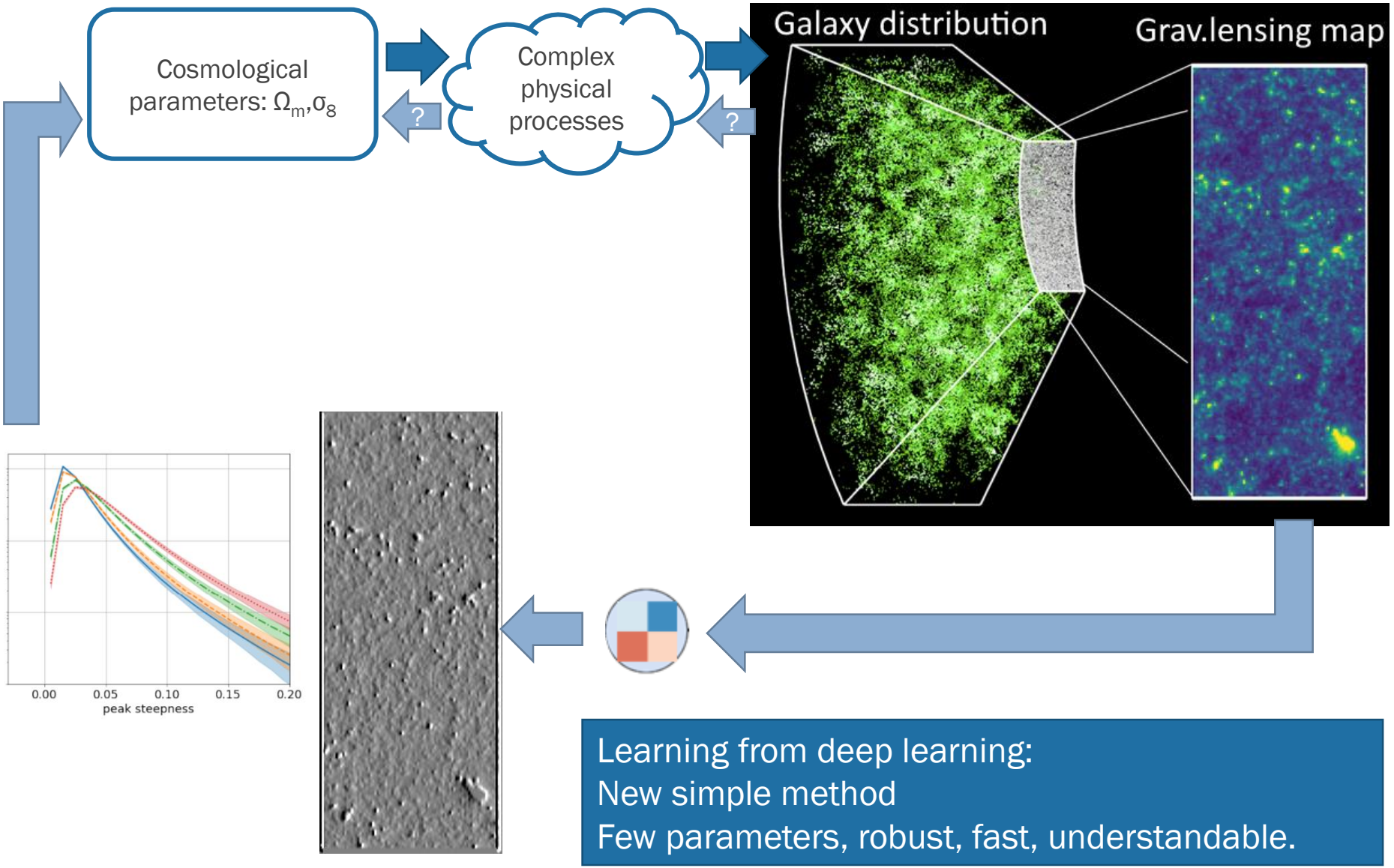


Instead of Fourier power spectrum: information from halo profiles



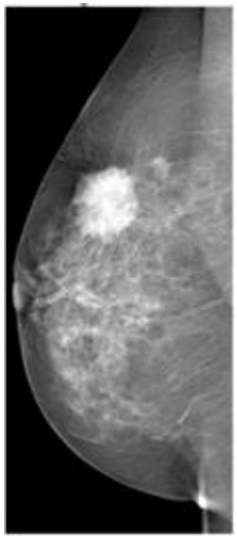
Attention focus of the network with Layer-wise Relevance Propagation

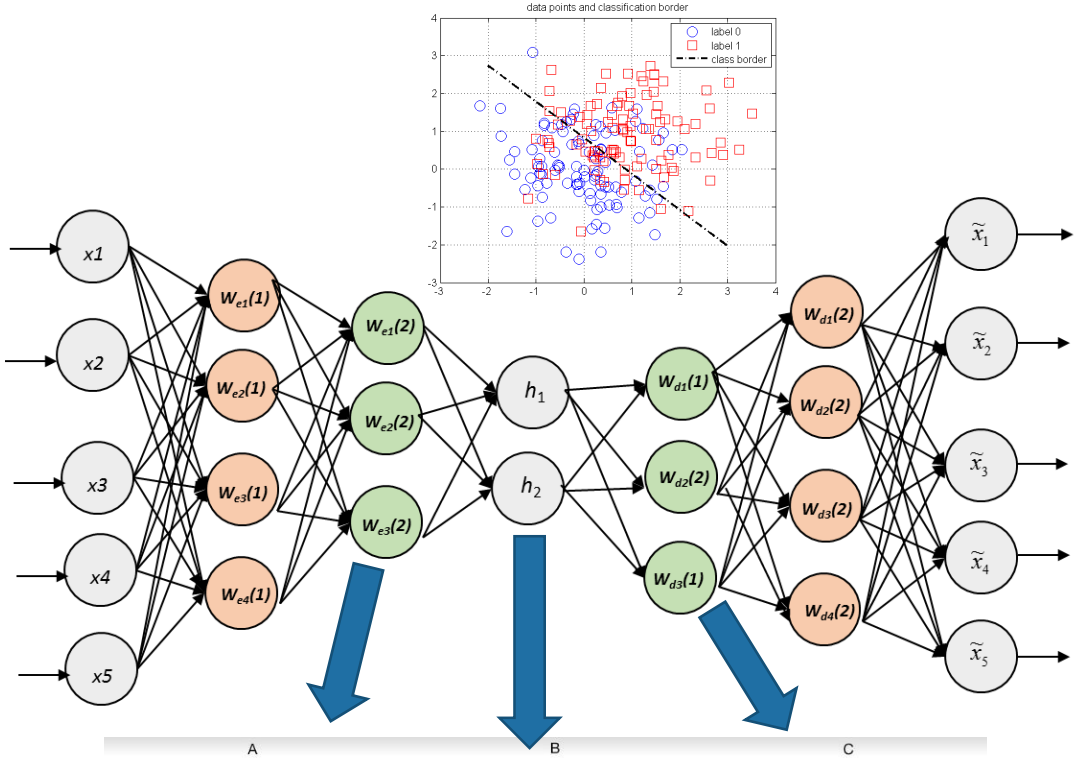
Cosmological parameters from gravitational lensing
Learning new tricks from deep learning

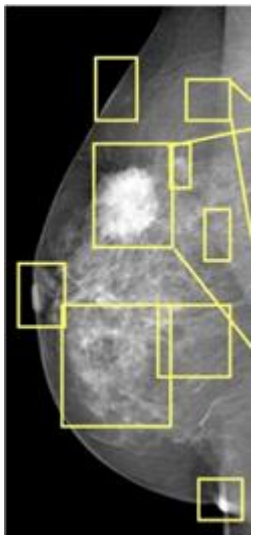


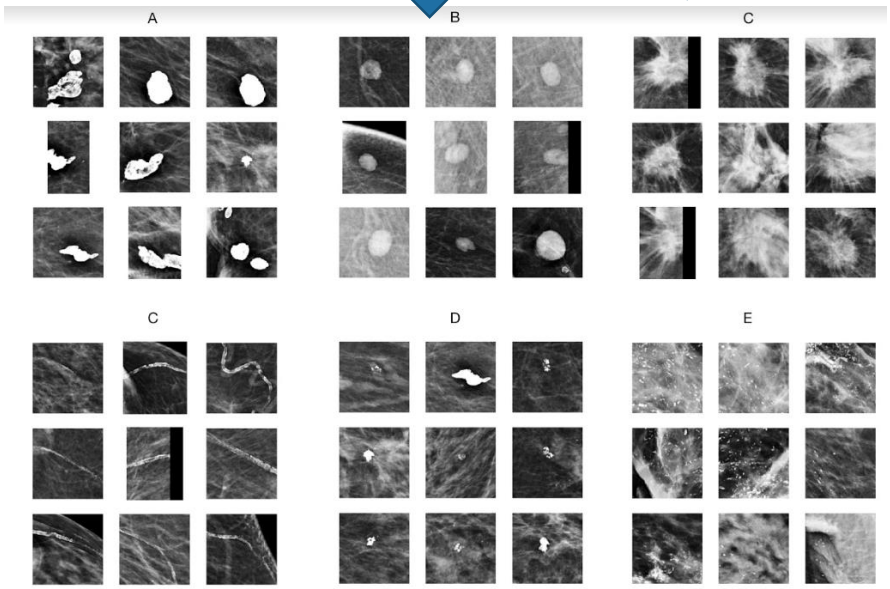
Learning from deep learning:
New simple method
Few parameters, robust, fast, understandable.

Understanding the internal lower dimensional representation





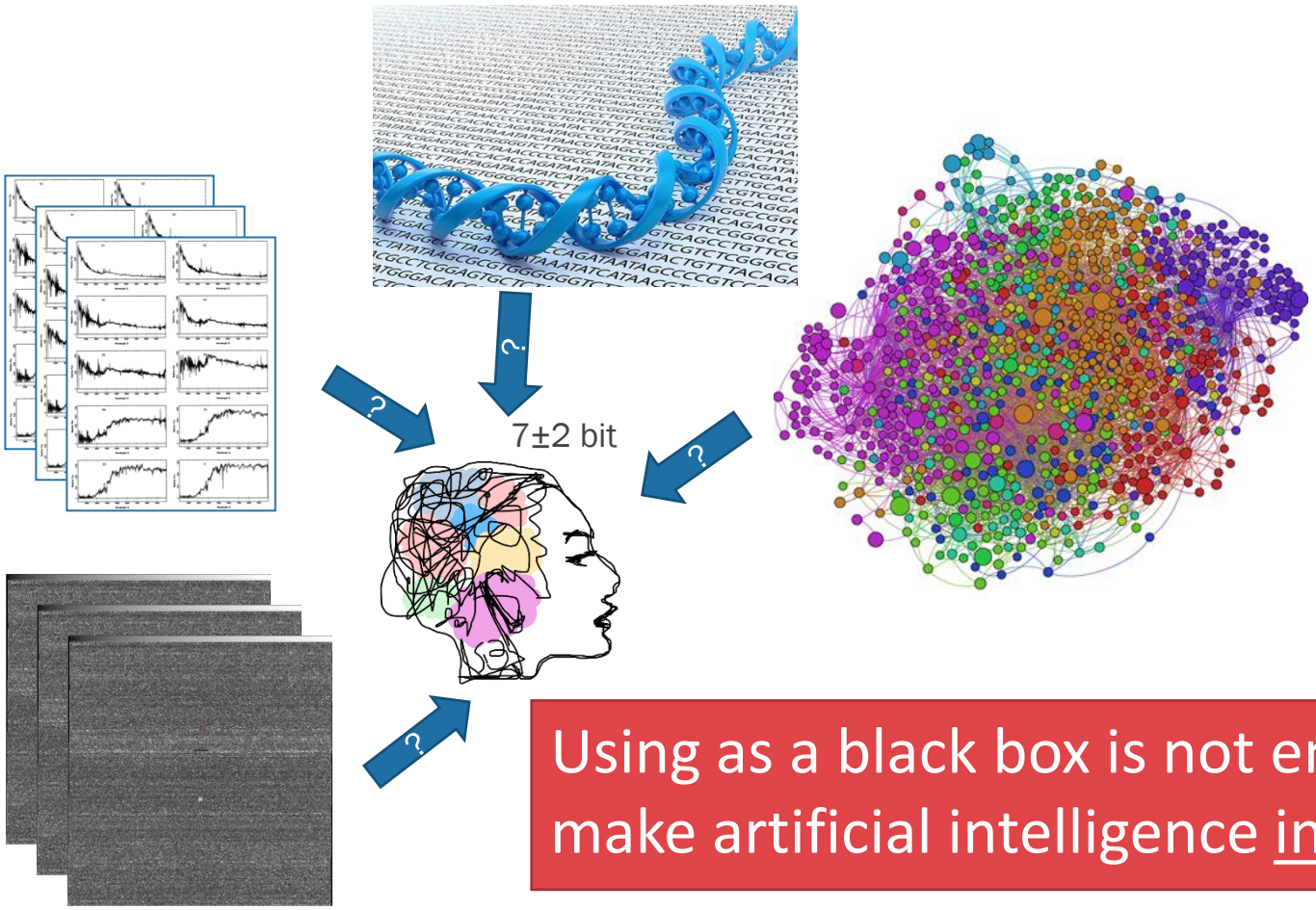




- Features at various levels of hierarchy
- Interpretable, trustworthy, for radiologists

Take home message:

Taming complexity is the key question for most sciences!



Using as a black box is not enough:
make artificial intelligence intelligible!

Istvan Csabai csabai@elte.hu
Dept. of Physics of Complex Systems

Acknowledgements

O. Pipek, D. Ribli, A. Medgyes-Horváth, K. Papp, D. Szuts, Z. Szallasi, S. Spisak, B. Molnar, M. Freedman, C. Swanton, M. Koopmans, F. Aarestrup, G. Cochrane, Z. Haiman, N. Solymosi, I. Kacskovics, C. Bödör, L. Dobos, J. Szalai-Gindl, L. Oroszlany, D. Visontai, J. Steger, A. Bodor, G. Vattay, P. Pollner, A. Major, G.Palla, G Rácz, I Szapudi, J. Börcsök, B. Pataki, Z. Udvarnoki, A. Biricz, A. Olár, M. Krzystanek, …

EU H2020 COMPARE #643476
Otto Monsteds Fond
Novo Nordisk Foundation
FIEK_16-1-2016-0005
NVKP_16-1-2016-0004
NKFI OTKA 124881
National Quantum Technologies Program
2017-1.2.1-NKP-2017-00001

ELTE
SOTE
MTA TTK
MTA Wigner FK
3DHISTECH
Harvard Children's Hosp.
Francis Crick Institute
DTU
Johns Hopkins University

THANK YOU FOR YOUR ATTENTION!

SZÉCHENYI 2020

HUNGARIAN GOVERNMENT

European Union
European Social Fund

INVESTING IN YOUR FUTURE

SLIDE HEADING

- **Bullet Point**

  Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur nec nisi vestibulum, interdum leo vitae, consequat ligula. Mauris ultrices elit vitae metus pellentesque, sit amet vulputate nisl commodo.

- **Bullet Point 2**

  Quat ligula. Mauris ultrices elit vitae metus pellentesque

- **Bullet Point 3**

  QUAT LIGULA. MAURIS ULTRICES ELIT VITAE METUS PELLENTESQUE

  Mauris ultrices elit vitae metus pellentesque, sit amet