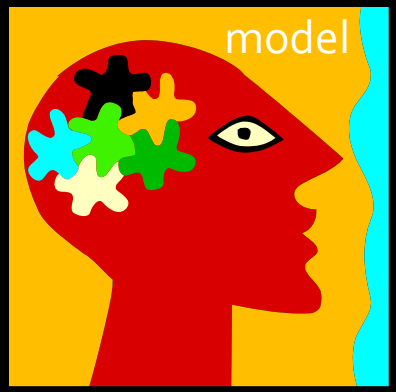
István Csabai
Eötvös University, Budapest
Department of Physics of Complex Systems

DATA-INTENSIVE APPROACH SCIENCES

BIG DATA IN BIOLOGY

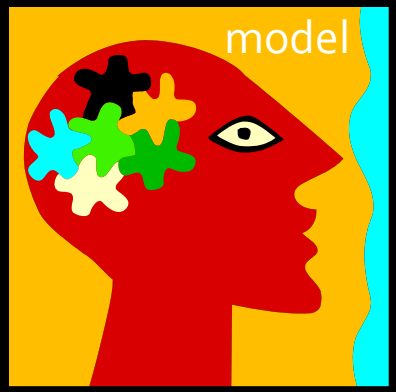
Élő adás - ELTE 2017. 04.19.

Evolution of (data) sciences: stone age

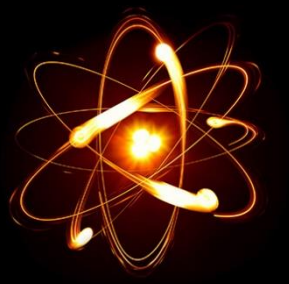


observation

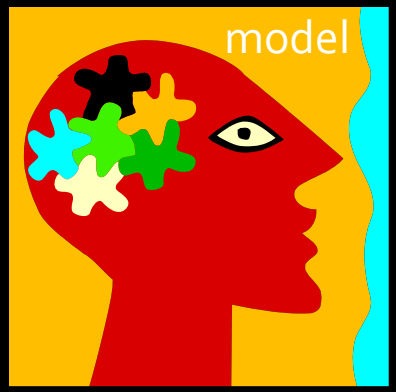
Evolution of (data) sciences: stone age



observation

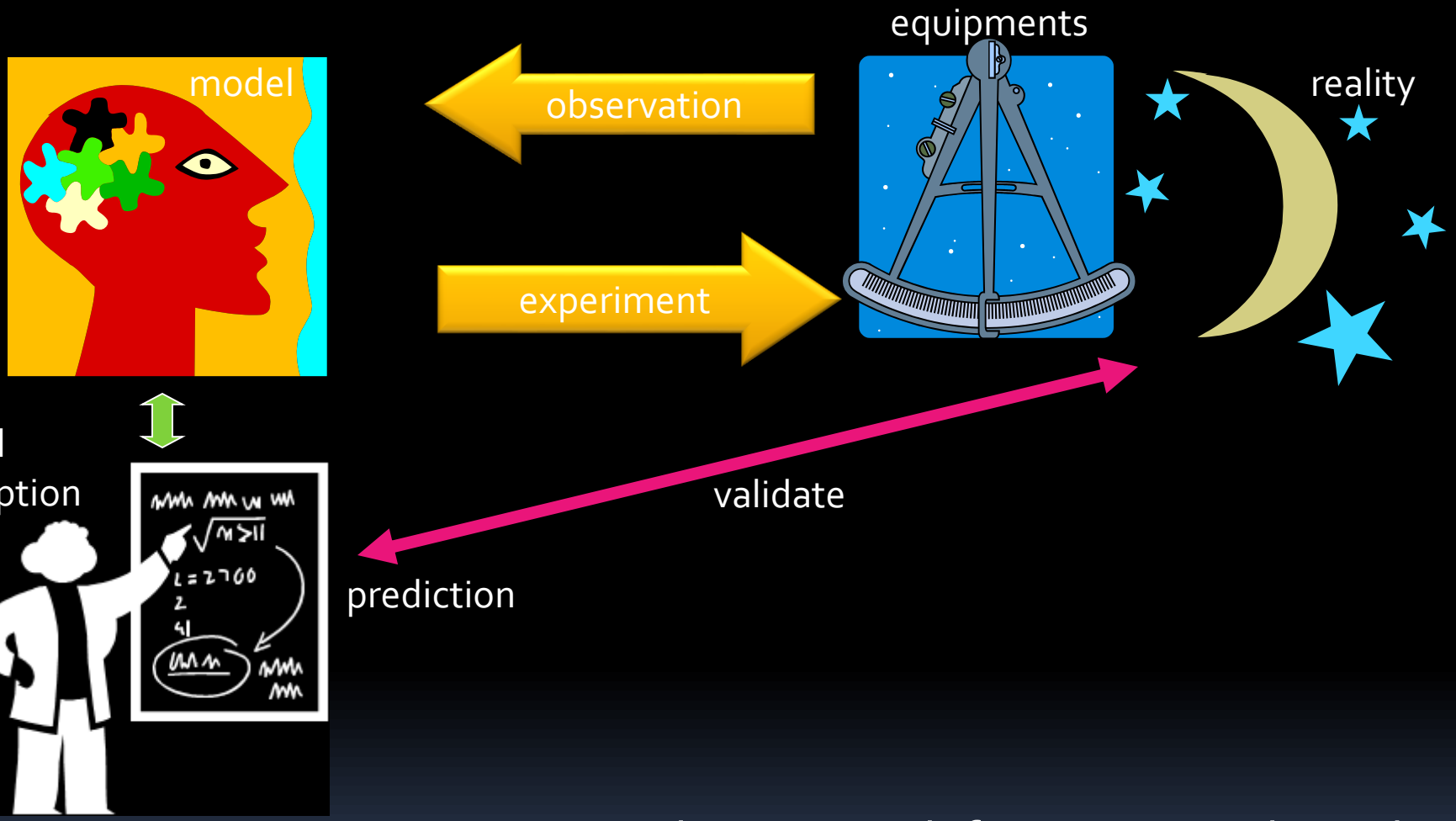


Evolution of (data) sciences: stone age



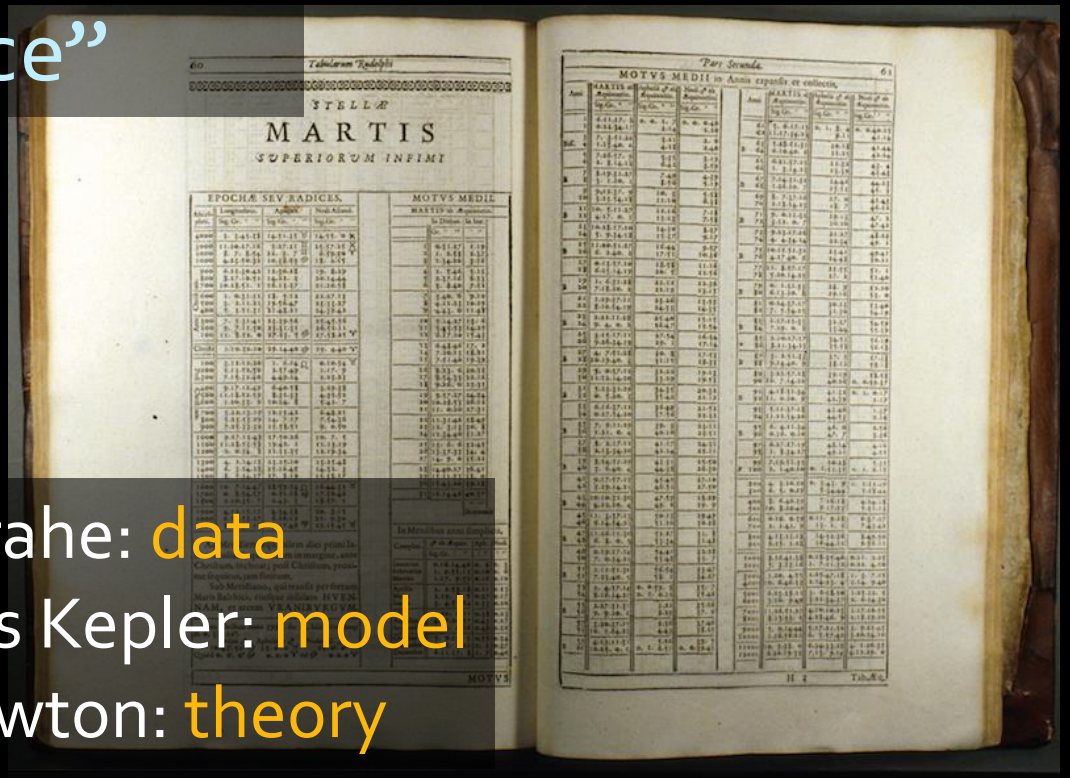
observation

Evolution of (data) sciences: pre-industrial age



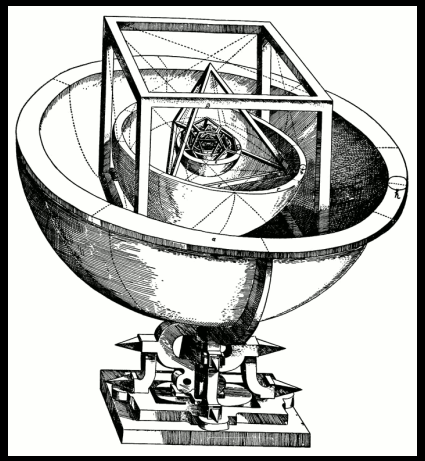
Prosthesis, crutch for senses and mind

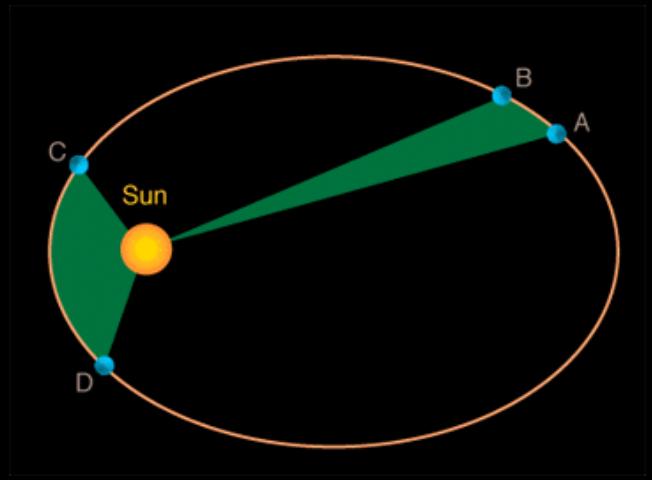
First “Data Science”



Tycho Brahe: data
Johannes Kepler: model
Isaac Newton: theory

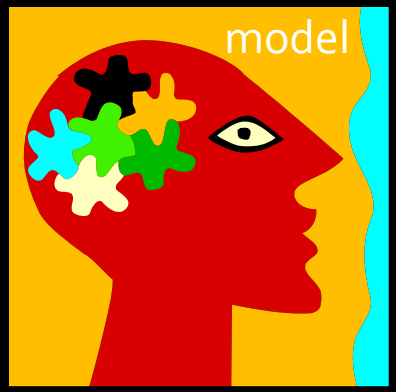
Johann Kepler , *Tabulae Rudolphinae* (1627)

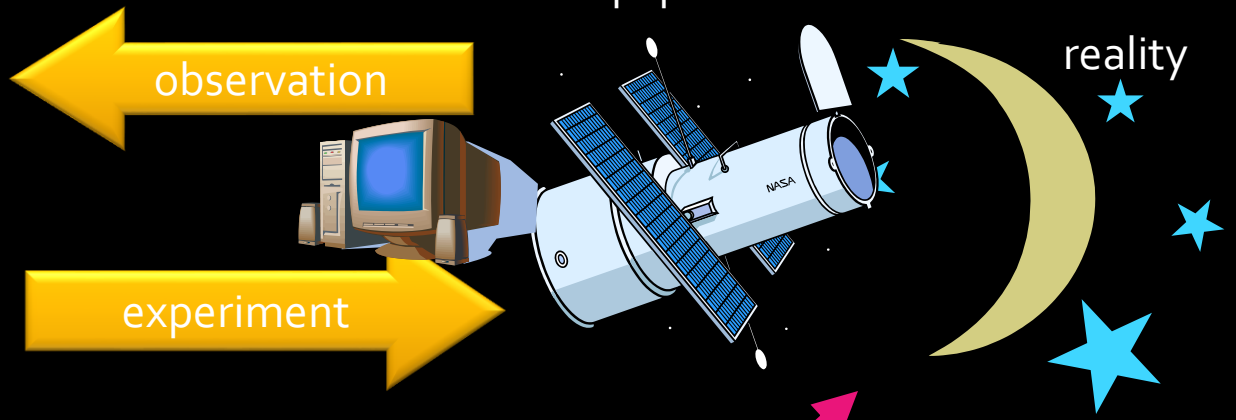




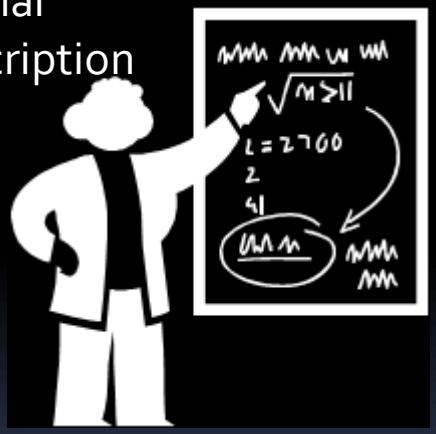
$$F = G\frac{m_1 m_2}{r^2}$$

Evolution of (data) sciences: present





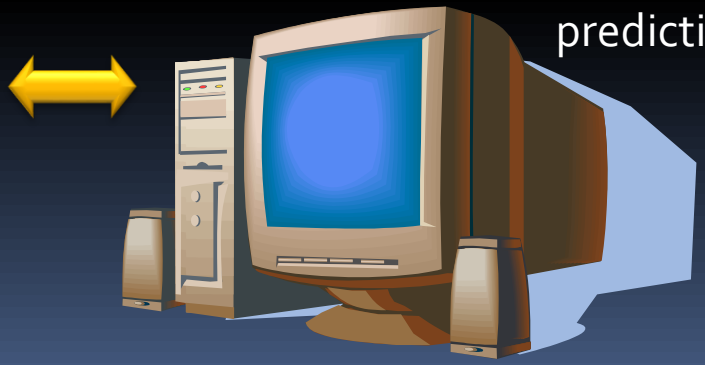
formal description



validation

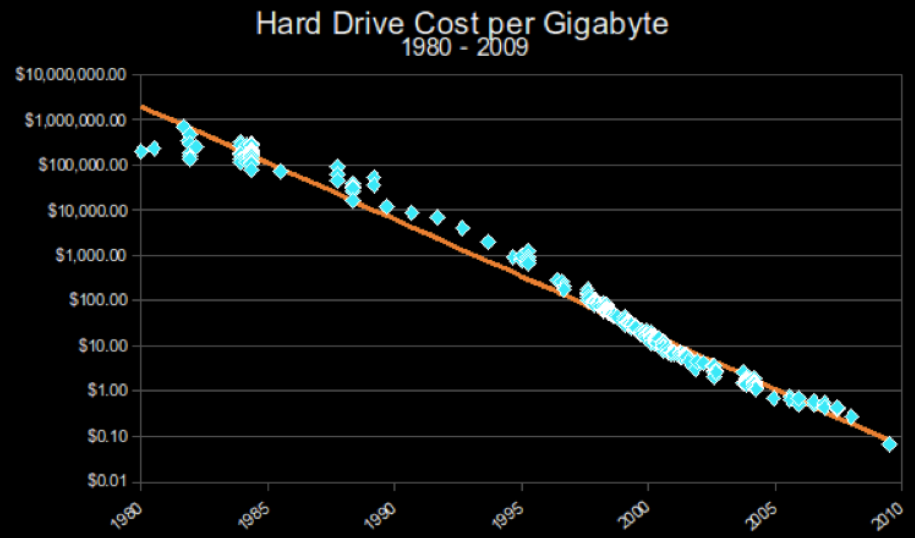
"virtual reality"

prediction



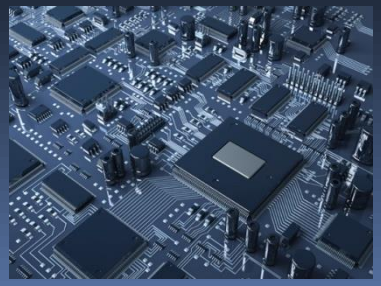
Exponentially cheaper devices – more data

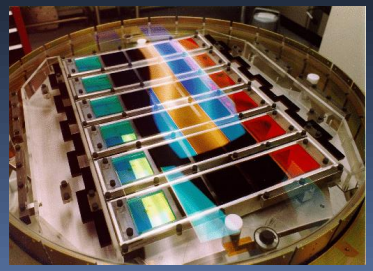
Quantum mechanics → Solid-state physics → Transistor/ Microelectronics



Moore's law

Elektronics → Sensors → Data



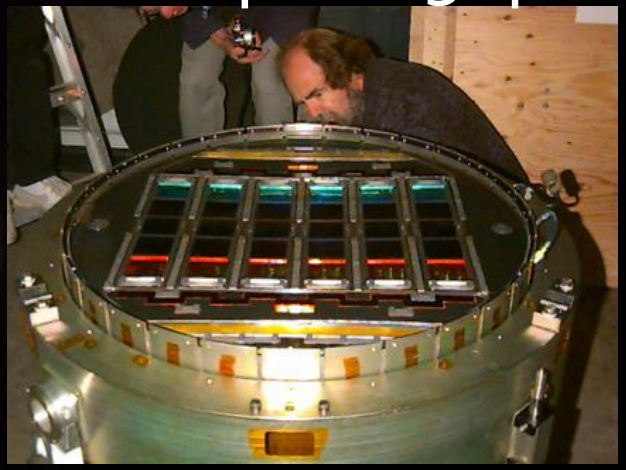


Prototype of data-intensive science project:

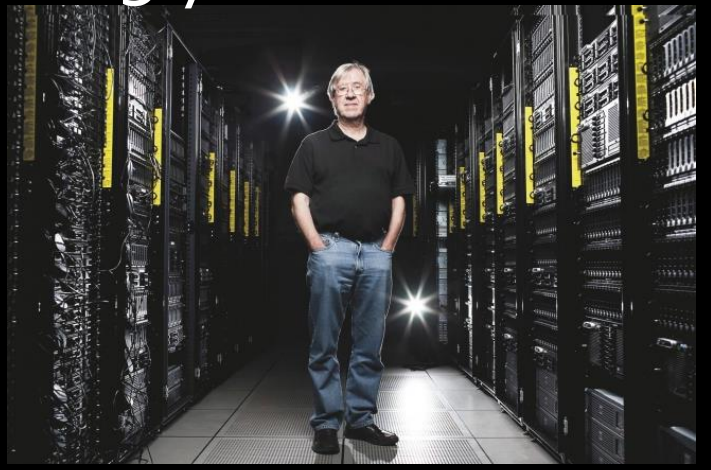
SLOAN DIGITAL SKY SURVEY (SDSS): THE 3D MAP OF THE UNIVERSE 1995-2005…

2.5m

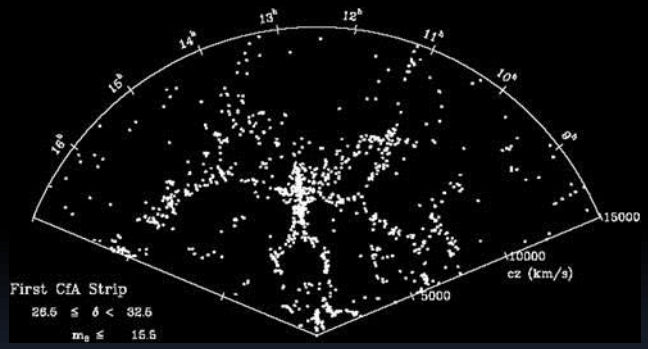
120Mp – 2.5Tp



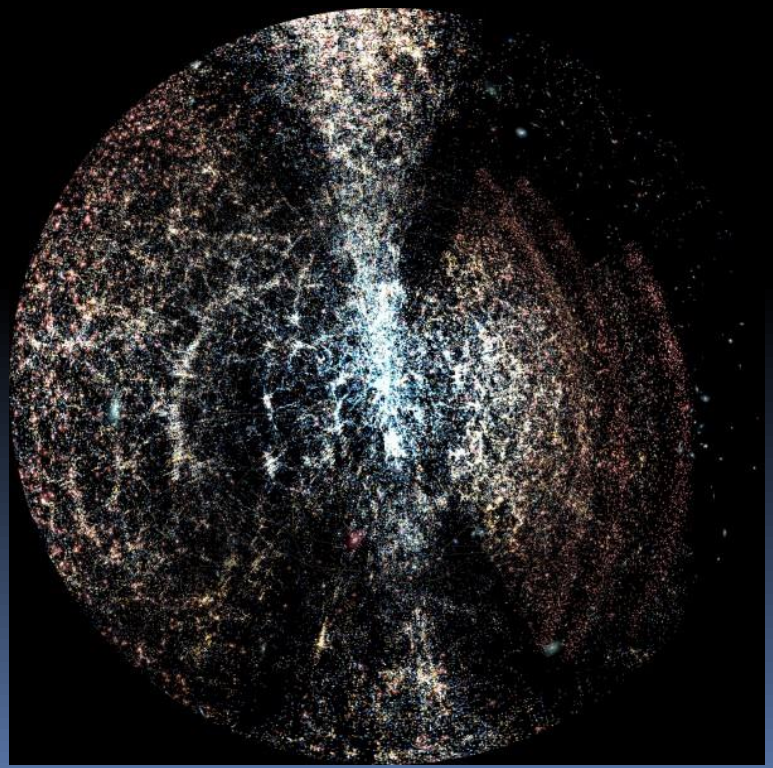
5 years:10TB



CfA 1989: 1100 galaxies



SDSS 2005: 1M galaxies

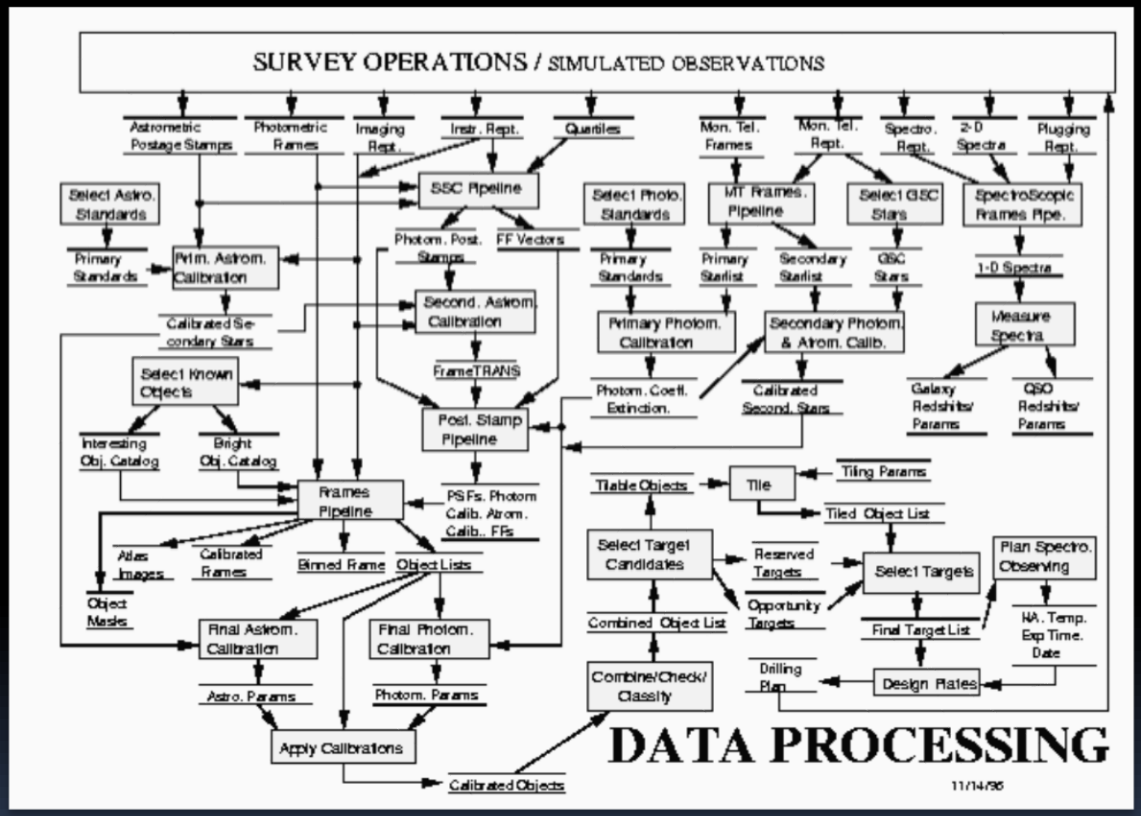


3D MAP OF THE UNIVERSE

Data processing challenge

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

Tables (SQL) + Raw data (files)



The sloan digital sky survey: Technical summary
DG York + SDSS collab.The Astron. J.l 120 (3), 1579 (2000)

PZ Kunszt, AS Szalay, I Csabai, AR Thakar;
ADASS IX 216, 141(2007)

Astronomical data sets–astronomical queries

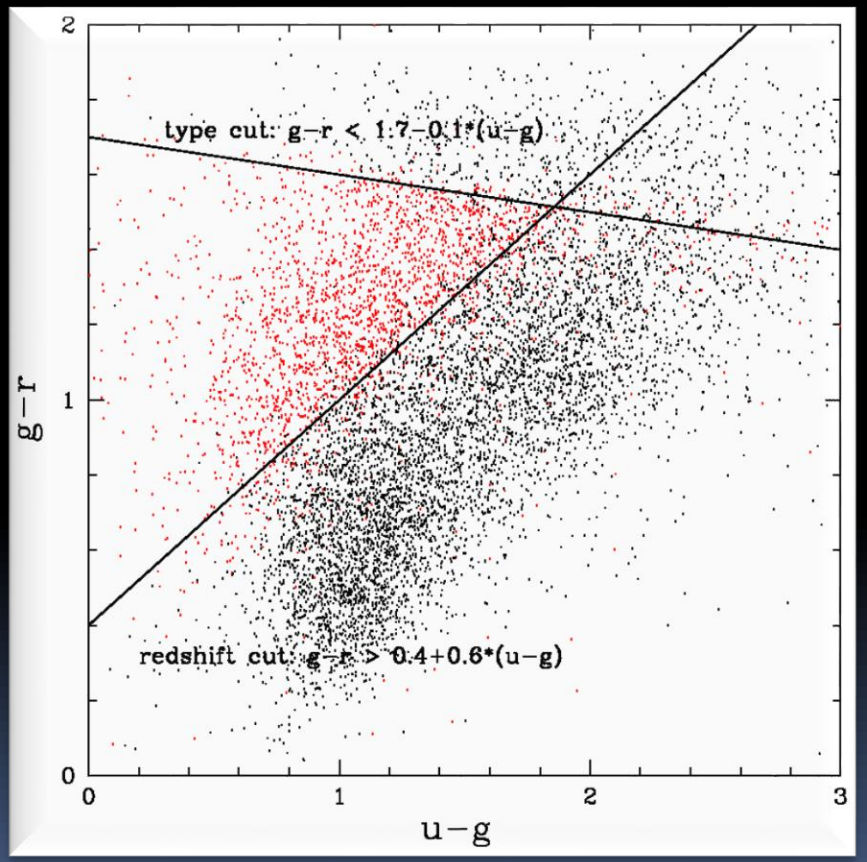
Star/galaxy separation
Quasar target selection

→ "cuts"

→ **Multi-dimensional polyhedra**

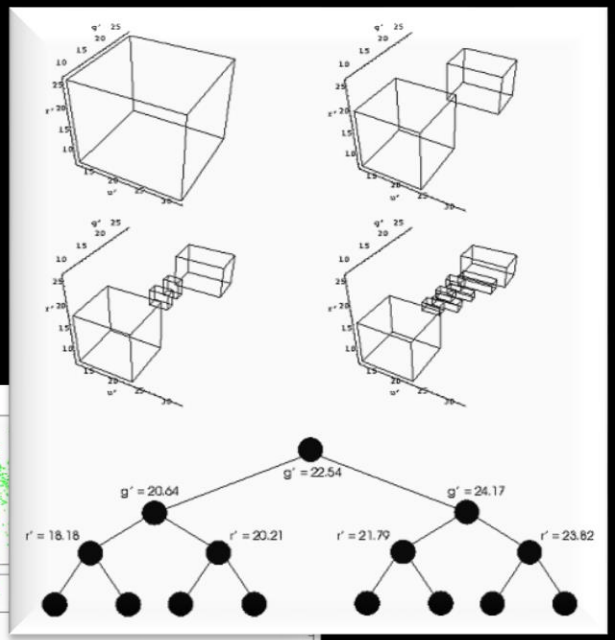
Skyserver: several million queries per year

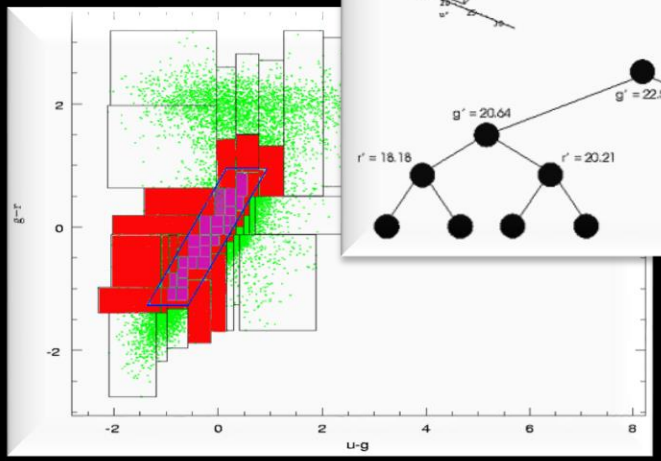
```
petroMag_i > 17.5 and (petroMag_r > 15.5 or petroR50_r > 2)
and (petroMag_r > 0 and g > 0 and r > 0 and i > 0) and (
(petroMag_r-extinction_r) < 19.2 and (petroMag_r -
extinction_r < (13.1 + (7/3) * (dered_g - dered_r) + 4 * (dered_r
- dered_i) - 4 * 0.18) ) and ( (dered_r - dered_i - (dered_g -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

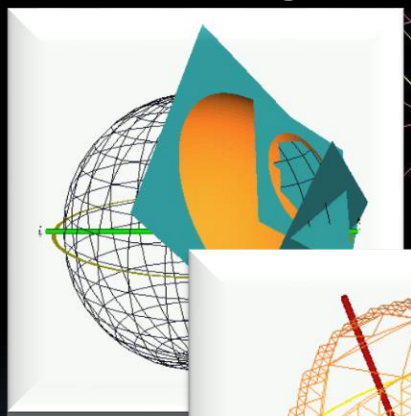


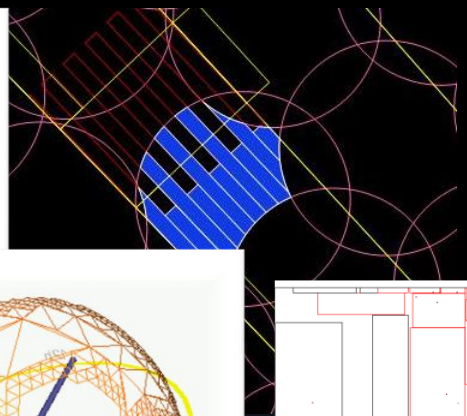
New skills: Indexing, databases

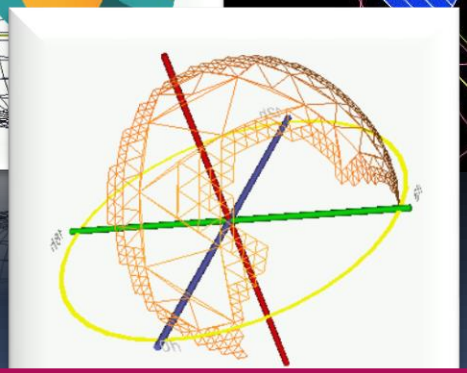
- SDSS data "read through"~1 day
- Astronomers should learn: Database programming, computer geometry, search trees, …
- Multidimensional- and spherical indexing

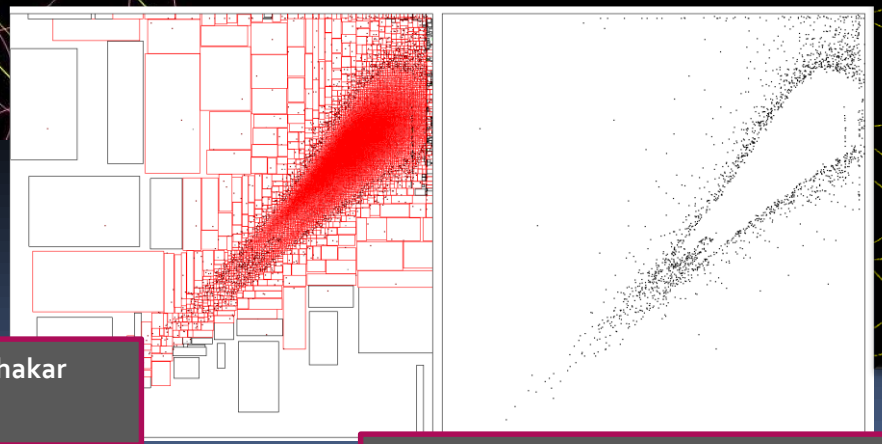


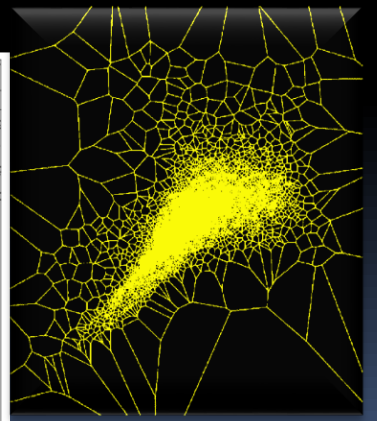












AS Szalay, J Gray, G Fekete,P Kunszt, P Kukol, A Thakar MSR -TR 123 (2005)

T Budavari, L Dobos, AS Szalay, G Greene, J Gray, AH Rots ASP Conf . Ser. 376, 559 (2007)

I Csabai, L Dobos, M Trencséni, G Herczegh, P Józsa, N Purger, T Budavári, AS Szalay Astr. N. 328 (8), 852 (2007)

New skills: Database management systems, virtualization

- RDBMS
  - +Developed for business purposes, optimised IO/memory accsess , declarative language (SQL), parallel queries, standard API (ODBC,JDBC)
  - -Relation data model is often not enough (matrices, graphs, [arrayLib]), not distributed [skyQuery, Graywulf]
  - New technologies: NoSQL,BigTable, Hadoop/MapReduce, column store, -> distributed servers
- Virtual Observatory (now: „cloud")
  - „If the data mountain does not go to …"
  - OpenStack, Docker, Jupyter
  - SciServer

- SkyServer
  - Web browser-based synchronous access
  - Meant to support several levels of users
    - From casual to moderately advanced queries
    - From simple form-based to direct SQL queries
    - From cone (radial) search to crossid type searches
  - Visual tools to browse image and catalog data
  - Stored procedures
  - API access, e.g. emacs interface, sqlcl (command-line)
  - Strict limits on execution time and output size
    - Fair use for everyone, robots/crawlers discouraged
  - **Introduction to SQL** and **Sample Queries**
- CasJobs
  - Batch Query Workbench, personal user DB (MyDB)
    - **Quick** mode: 1 minute cutoff
    - **Submit** mode: up to 8 hours in "long" queue
      - 24-hr queue for collab members
  - **MyDB** database to save results of your queries
    - Define your own functions, procedures too
    - Share your tables with collaborators (groups)
  - Job history, plotting, FITS/CSV/VOTable output
  - Restricted (collab-only databases)
  - Table **Import** (upload) for your own data
  - **Groups** to share your results with collaborators
  - Command-line access Java tool also downloadable
  - SOAP/Web Services access

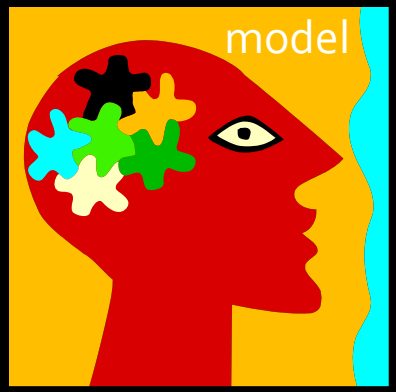
L Dobos, AS Szalay, J Blakeley, B Falck, T Budavári, I Csabai
Astronomical Data Analysis Software and Systems XXI 461, 323 (2012)

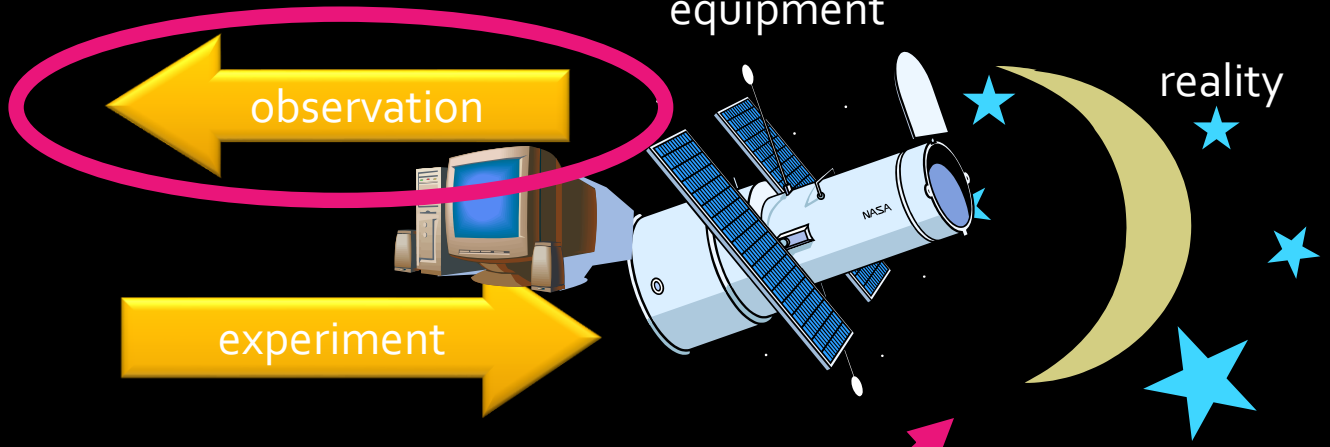
L Dobos, I Csabai, AS Szalay, T Budavári, N Li
Proceedings of the 25th International Conference on Scientific and Statistical Database Management, ACM, (2013)

L Dobos, T Budavári, N Li, AS Szalay, I Csabai
Scientific and Statistical Database Management, 159-167 (2012)

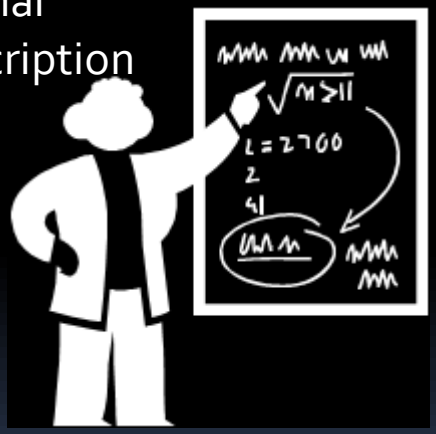
L Dobos, T Budavári, I Csabai, AS Szalay
Astronomical Data Analysis Software and Systems (ADASS) XIII 314, 185 (2007)

T Budavári, L Dobos, AS Szalay, G Greene, J Gray, AH Rots
Astronomical Society of the Pacific Conference Series 376, 559 (2007)





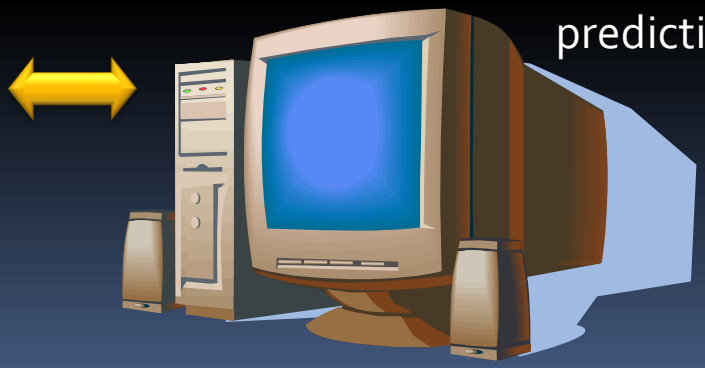
formal description



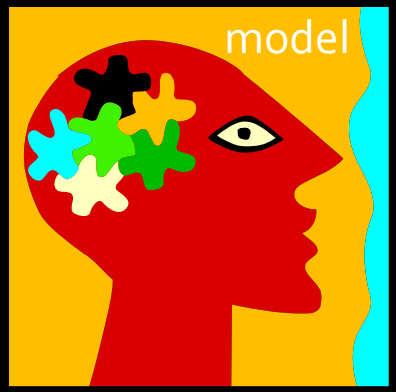
validation

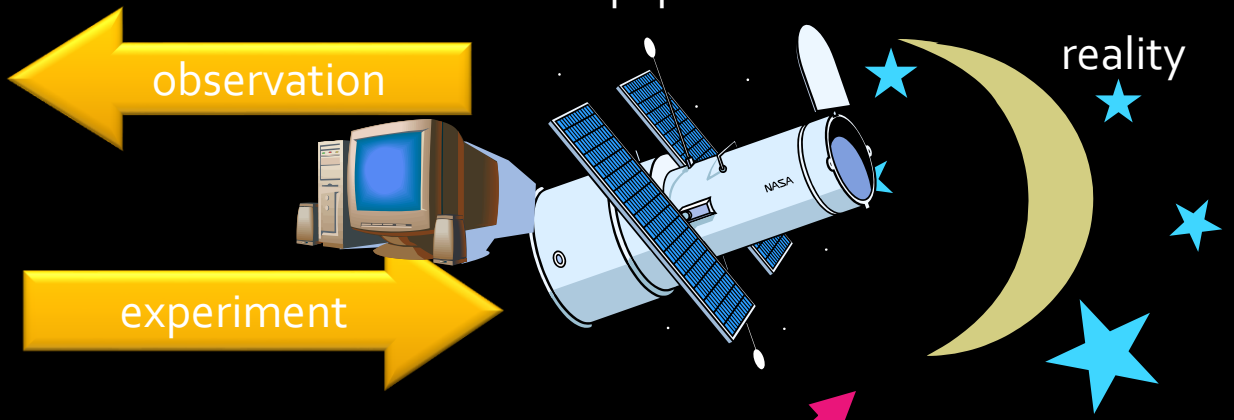
"virtual reality"

prediction



“Virtual reality”

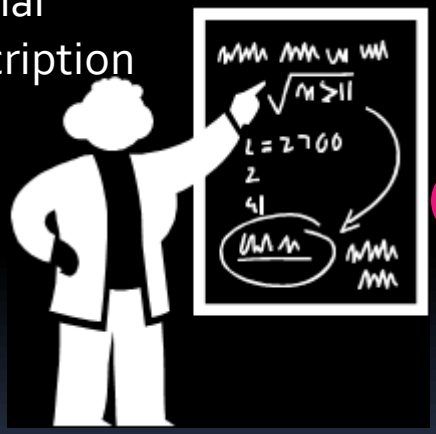




equipment

reality

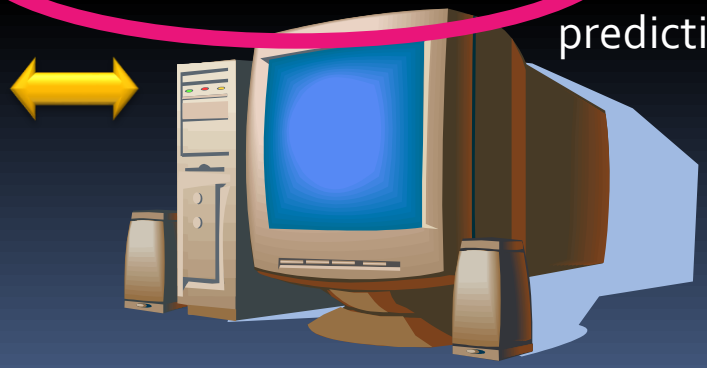
formal description



validation

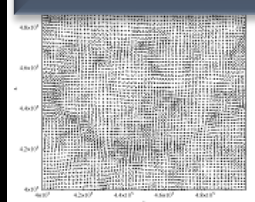
“virtual reality”

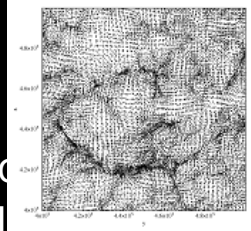
prediction

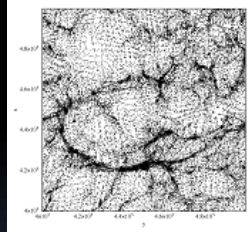


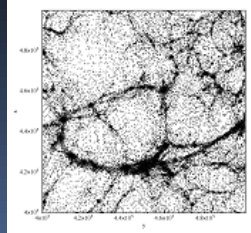
Models: N-body simulations

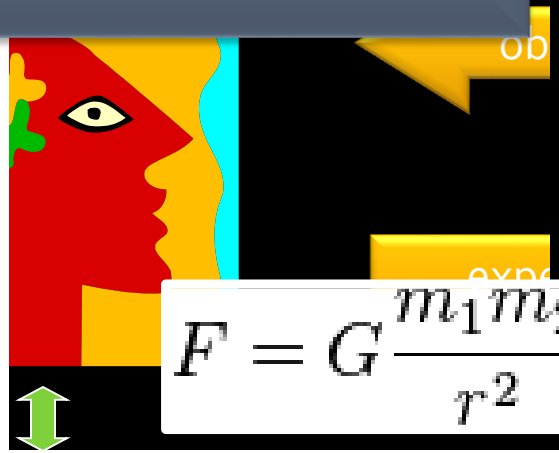
Simulation data can be as big and complex as observed data!





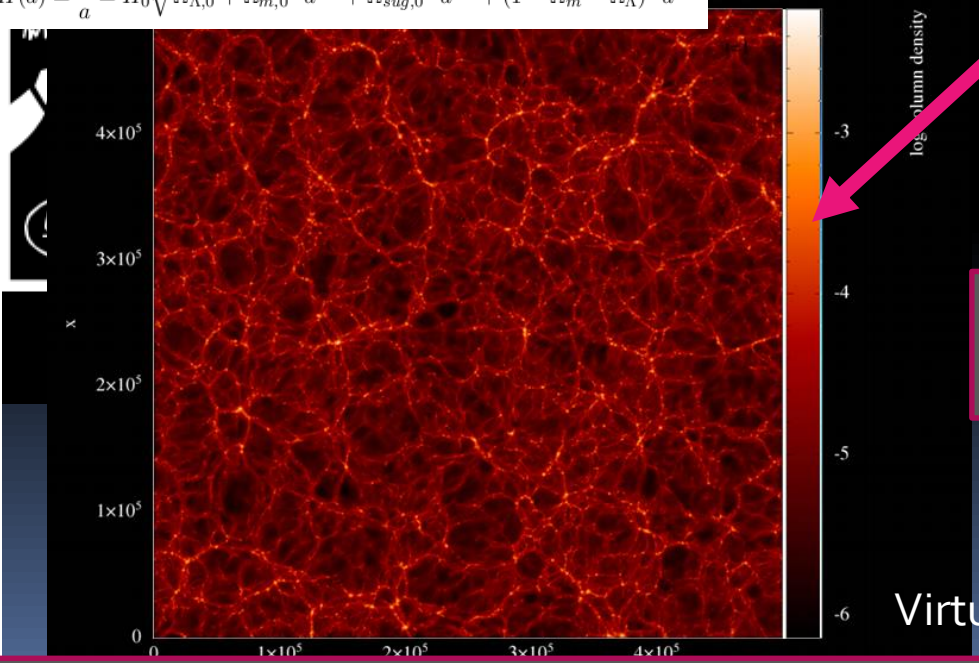




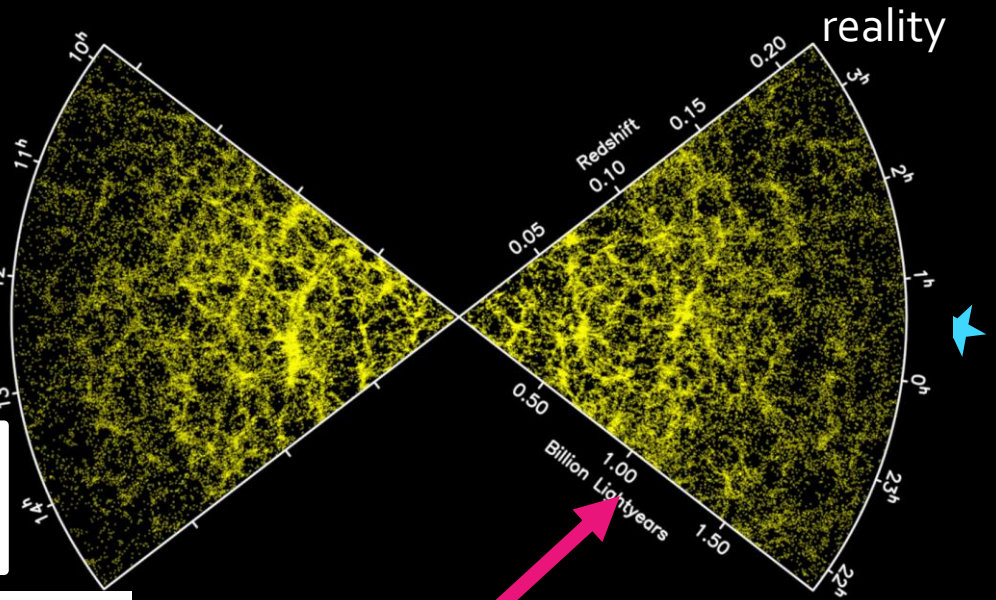


$$F = G\frac{m_1 m_2}{r^2}$$

$$H(a) = \frac{\dot{a}}{a} = H_0\sqrt{\Omega_{\Lambda,0} + \Omega_{m,0}\cdot a^{-3} + \Omega_{sug,0}\cdot a^{-4} + (1-\Omega_m-\Omega_\Lambda)\cdot a^{-2}}$$



Virtual reality



reality

test

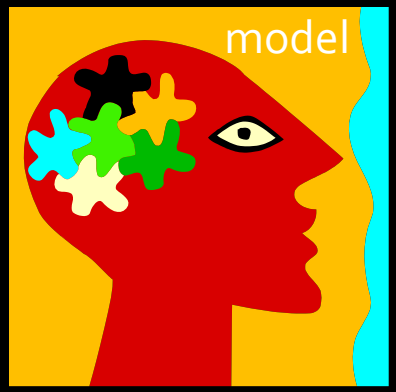
Data type:
7+ D point cloud

L Dobos, I Csabai, JM Szalai-Gindl, T Budavári, AS Szalay: Point cloud databases, SSDBM 2014

G Rácz et al., Concordance cosmology without dark energy, Mon Not R Astron Soc Lett (2017) 469 (1)

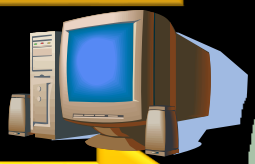
Real Universe – Virtual Universe

Not only astronomy: genomics, environmental sciences, social sciences… Complex questions – large datasets

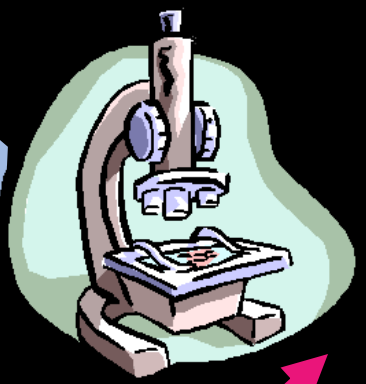


observation

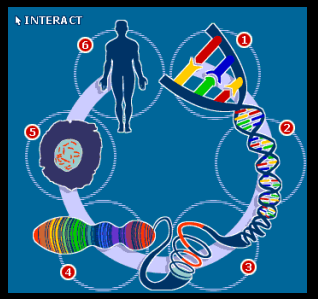
instruments



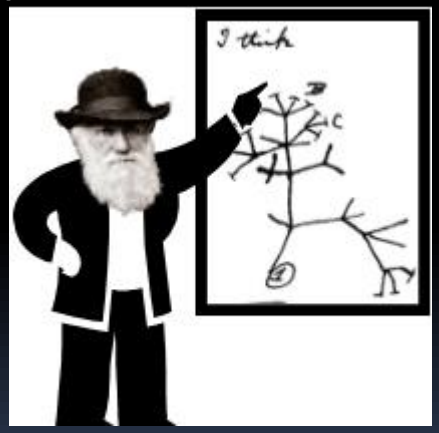
experiment



reality



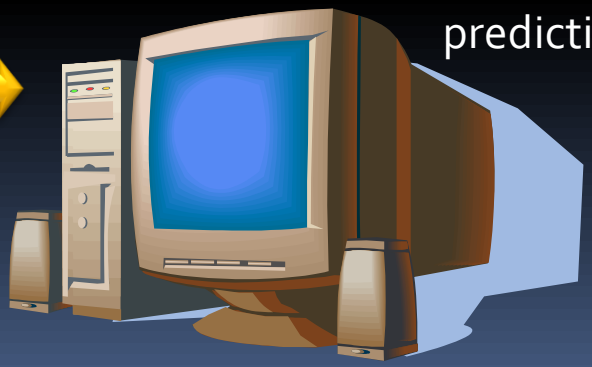
formal description



test

Virtual reality

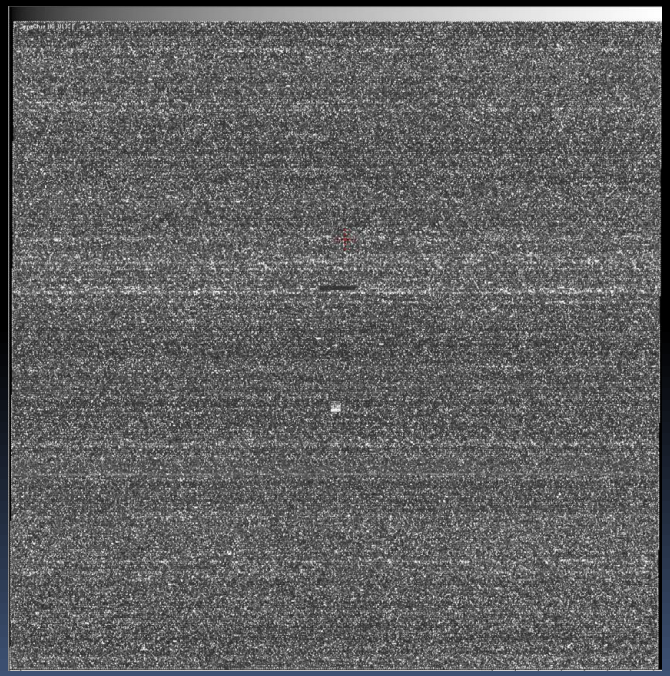
prediction

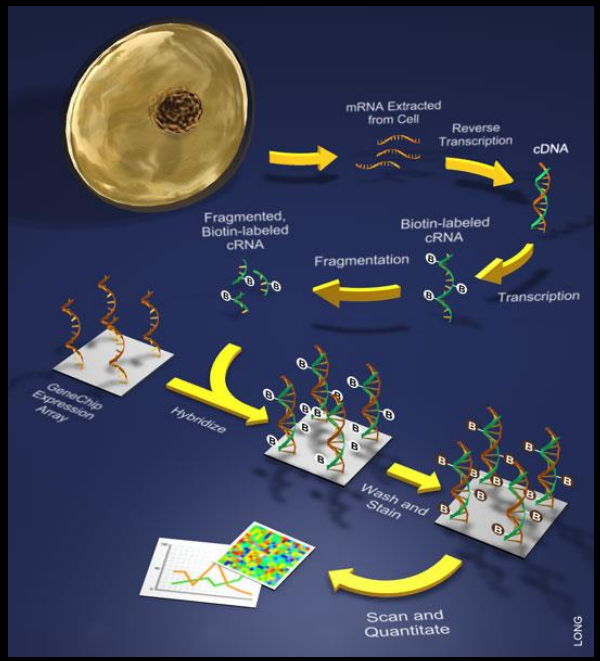


DATA-INTENSIVE GENETICS
NEXT GENERATION SEQUENCING

Expression microarray study (2010)

- Affymetrix HG U133 Plus2
  - Raw data 67Mpix (photometry!)
  - 604258 probes
  - 54675 probe set (~gene)
  - 207 samples (colorectal cancer)





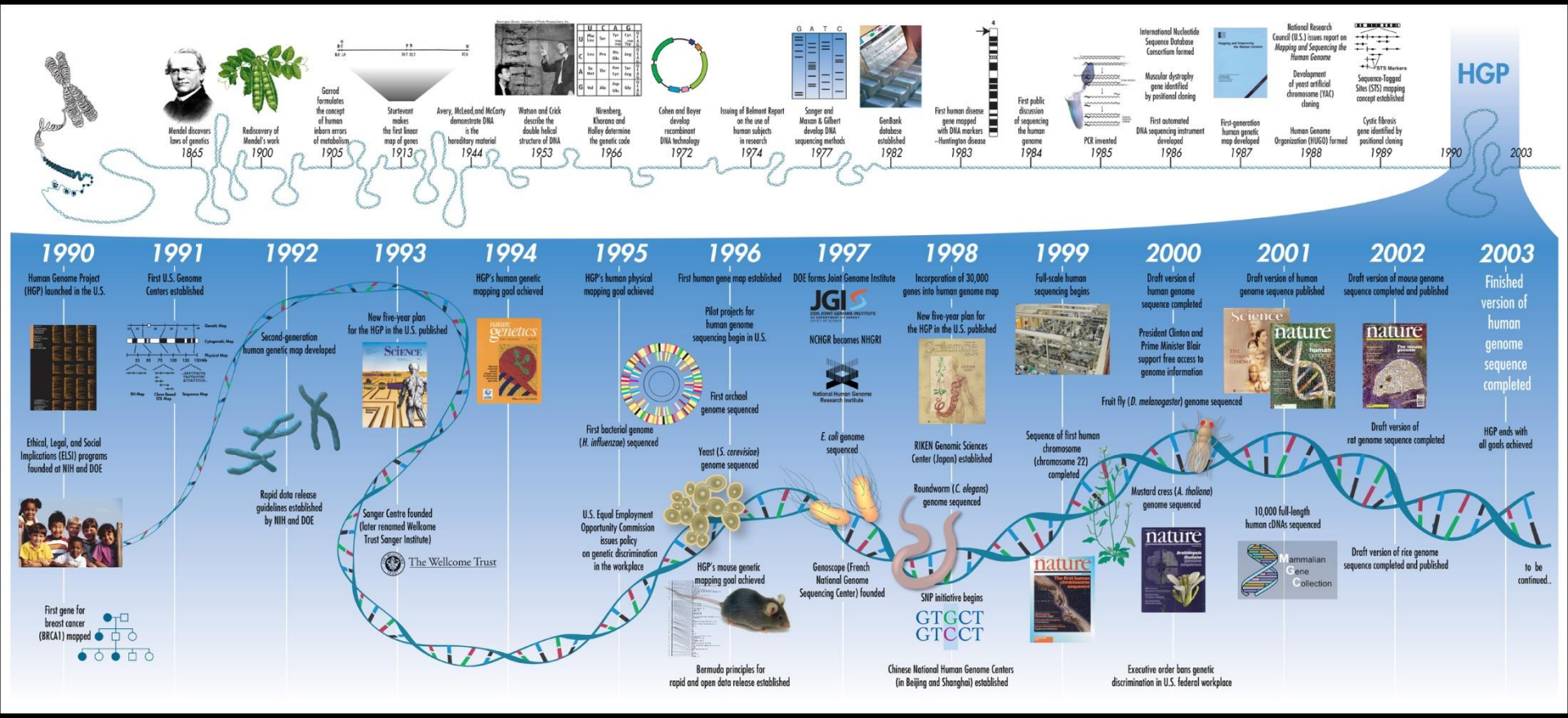
Raw data:
Image
Processed data:
54675D vectors + metadata

"similar" to astronomy

- Large databases (own + public)
- Computer-intensive data analysis

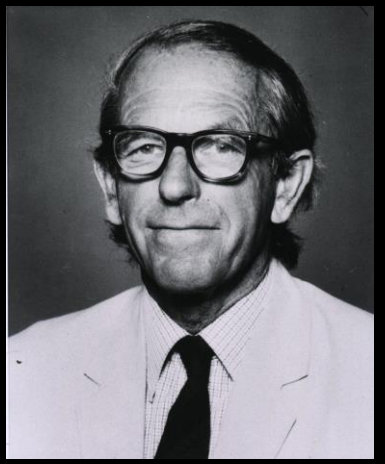
S Spisák, A Kalmár, O Galamb ... I Csabai, I Kovalszky, S Semsey, Z Tulassay, B Molnar. PLoS One dx.doi.org/10.1371/journal.pone.0046215 (2012)

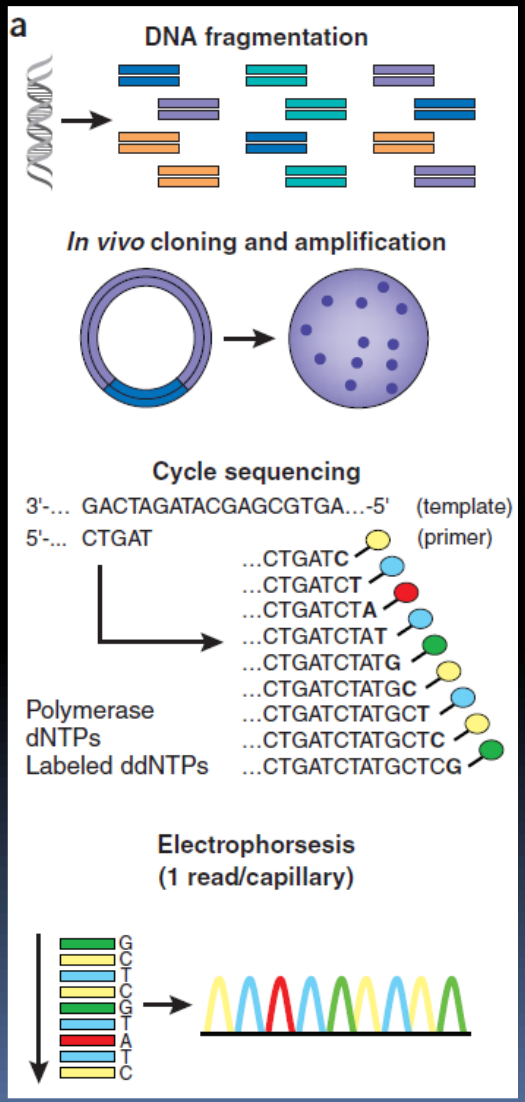
Map of the genome

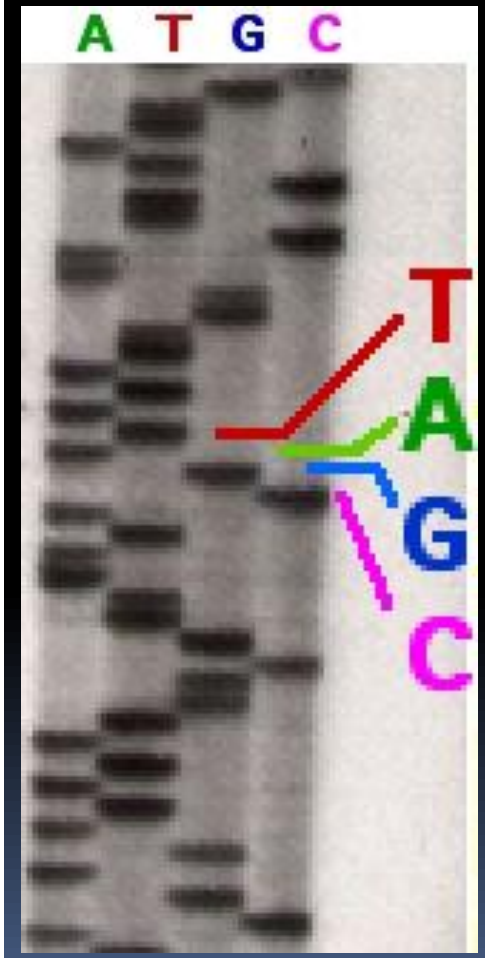


High througput sequencing history: Sanger

1977 Frederick_Sanger

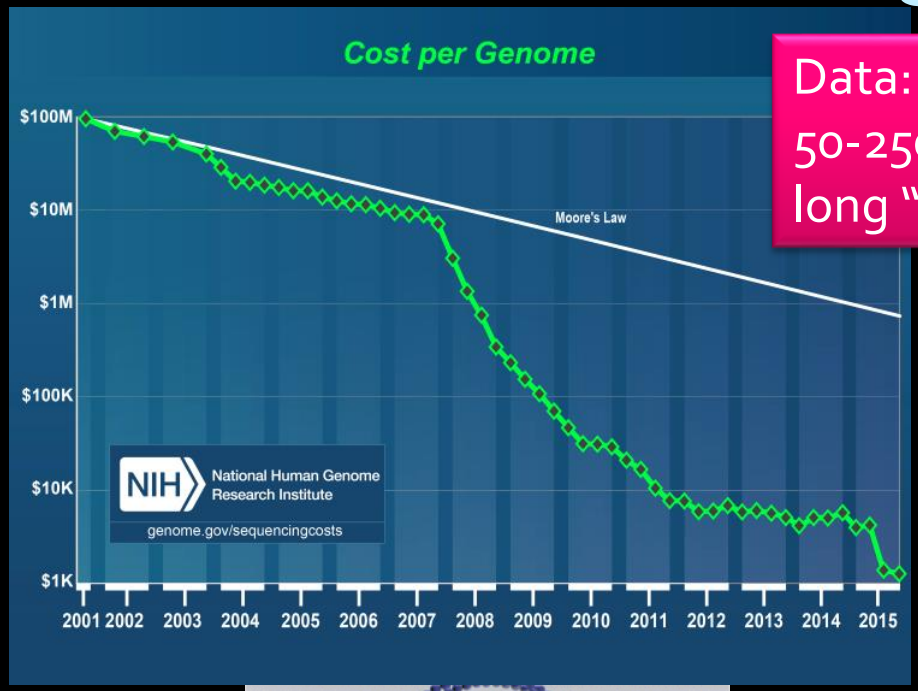






- DNA is fragmented
- Cloned to a plasmid vector
- Cyclic sequencing reaction
- Separation by electrophoresis
- Readout with fluorescent tags

Moore's law in gene sequencing

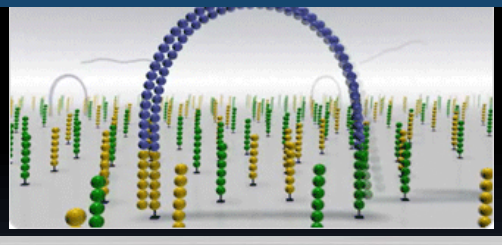


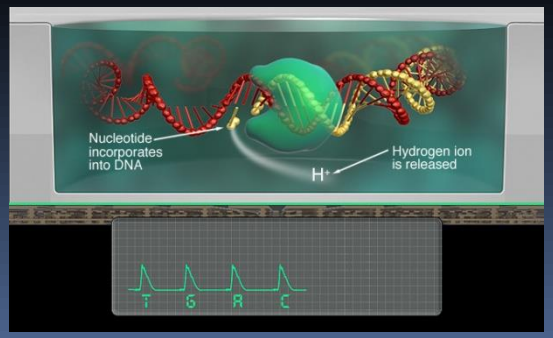
Data:
50-250 or 3Bn letter long "texts"

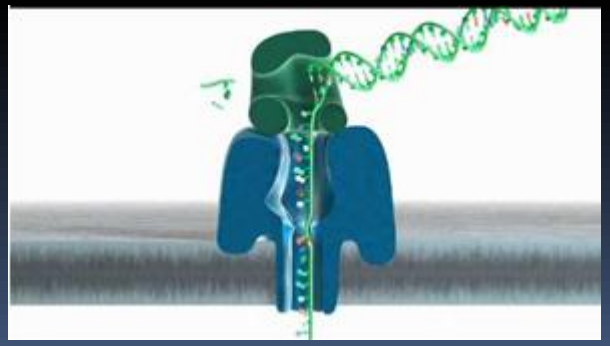
Human genome sequencing
1990-2003: 13yrs /2.7 Bn USD
2016: ~days/1000 USD
2020: ?????

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

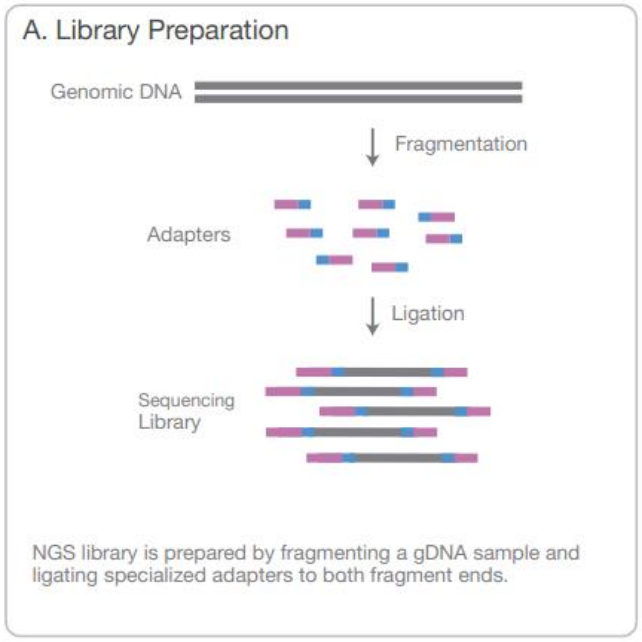
Oxford Nanopore
100Mb,$900

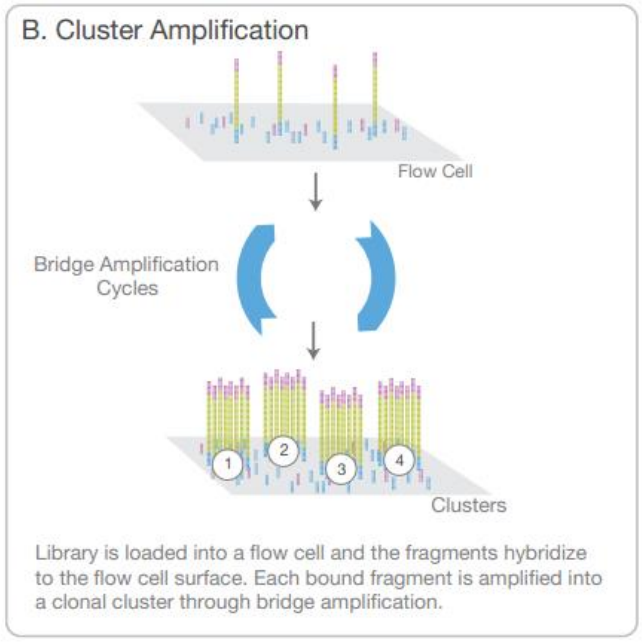


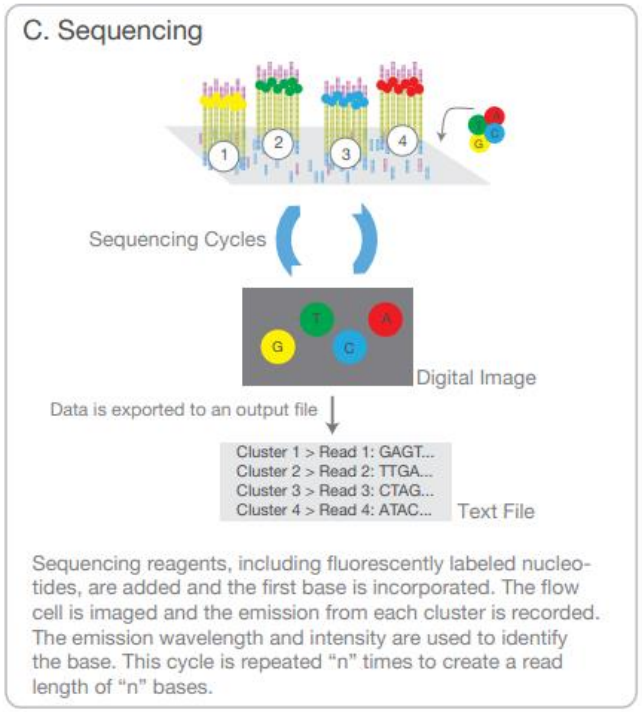




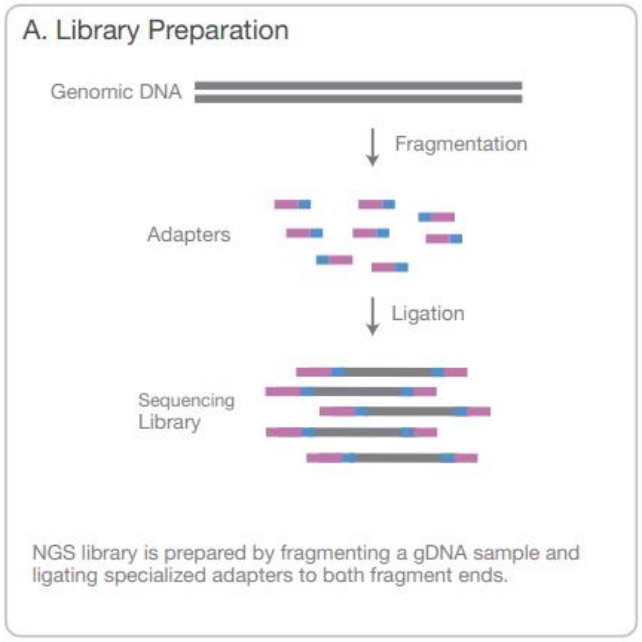
NGS – from samples to data

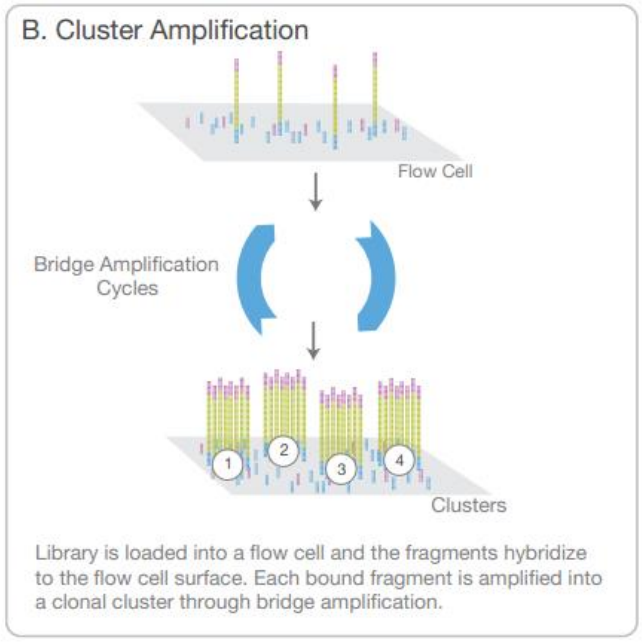


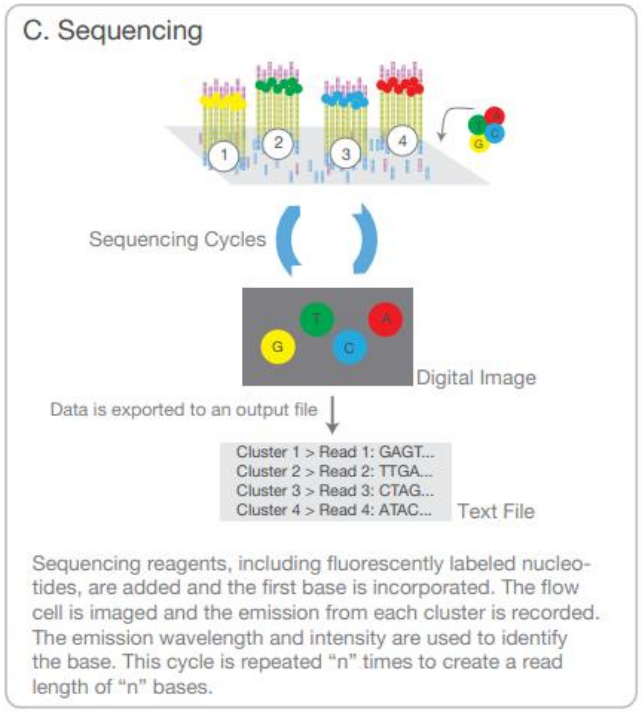




NGS – from samples to data

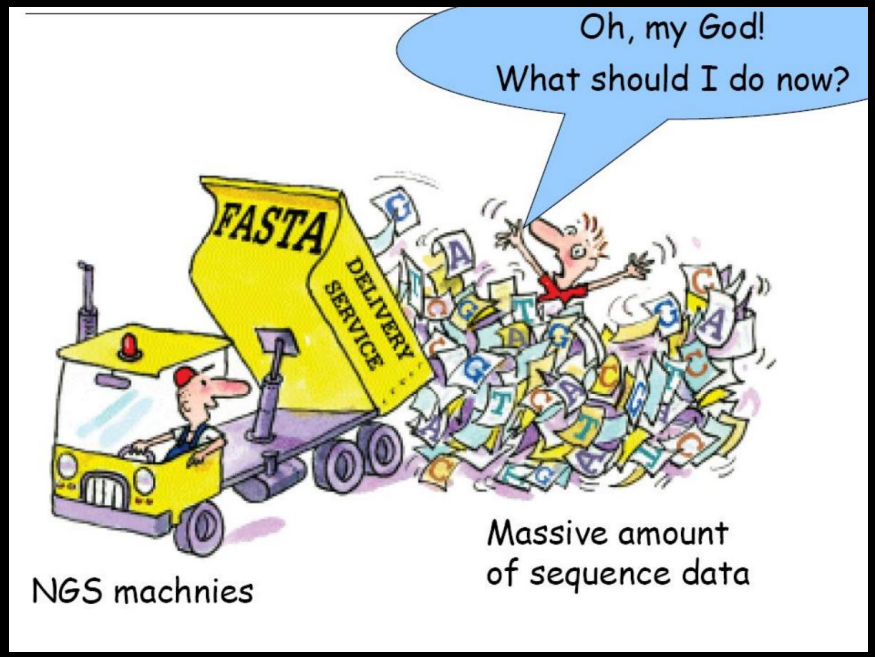




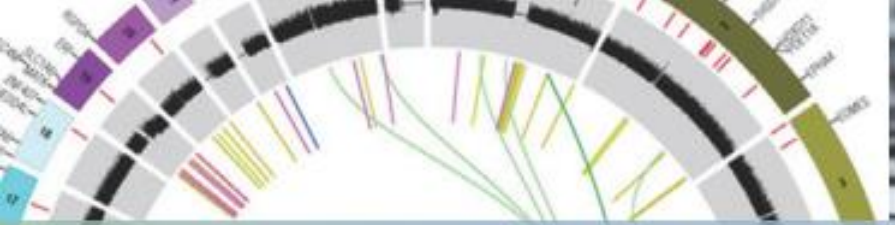


Large part of research starts here

NGS alignment/assembly
Big numbers



- Computer speed: ~10^9 instr./sec
- Genome: ~ 10^9 nt
- NGS: ~10^9 short reads
- Brute force: ~10^{18} comparisions
- -> Need for clever indexing/search algorithms!



National Cancer Institute
National Human Genome Research Institute

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

The Cancer Genome Atlas
http://cancergenome.nih.gov/

3.2Bn nucleotides / human genome

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

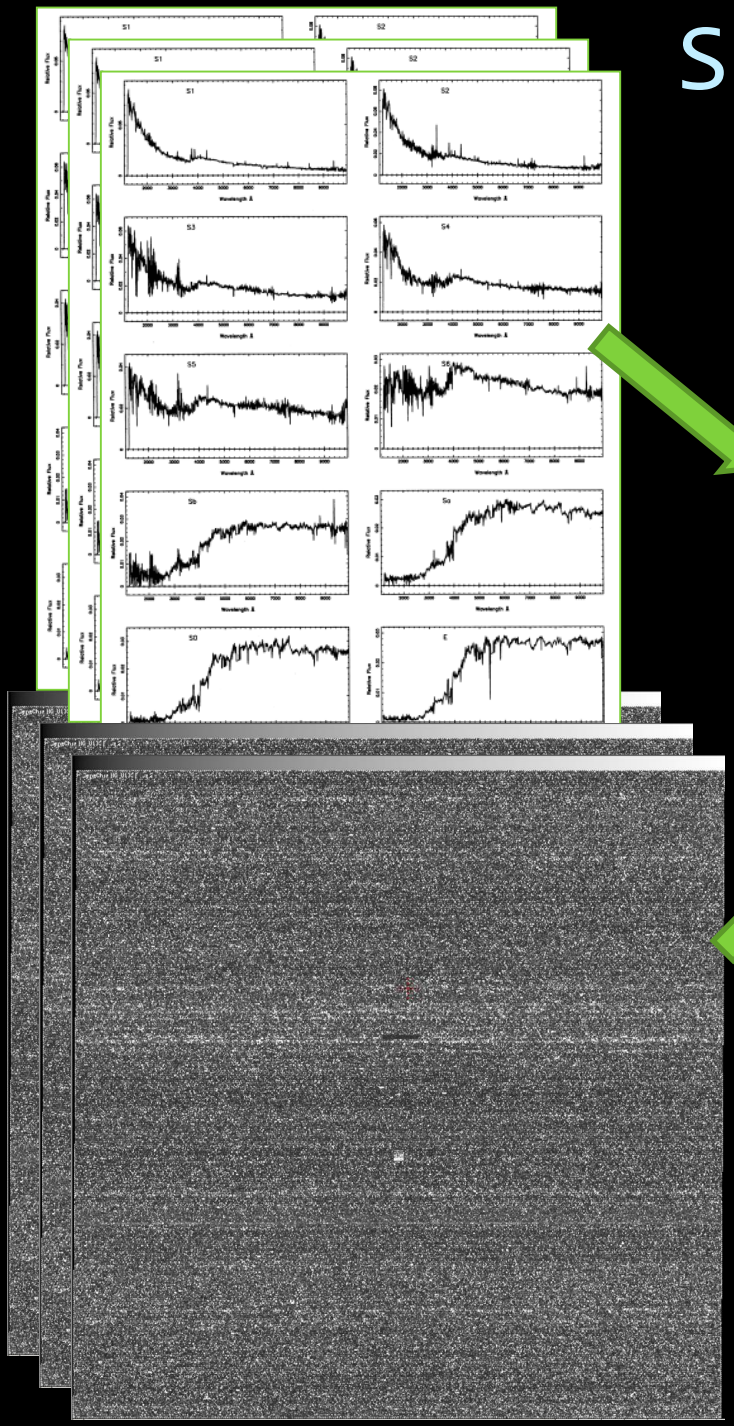
TCGA is a community resource project. The information generated by TCGA is centrally managed and entered into databases as it becomes available, making the data rapidly accessible to the entire research community. By January 2014, TCGA had generated one petabyte of data for about 10,000 cases of tumor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

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

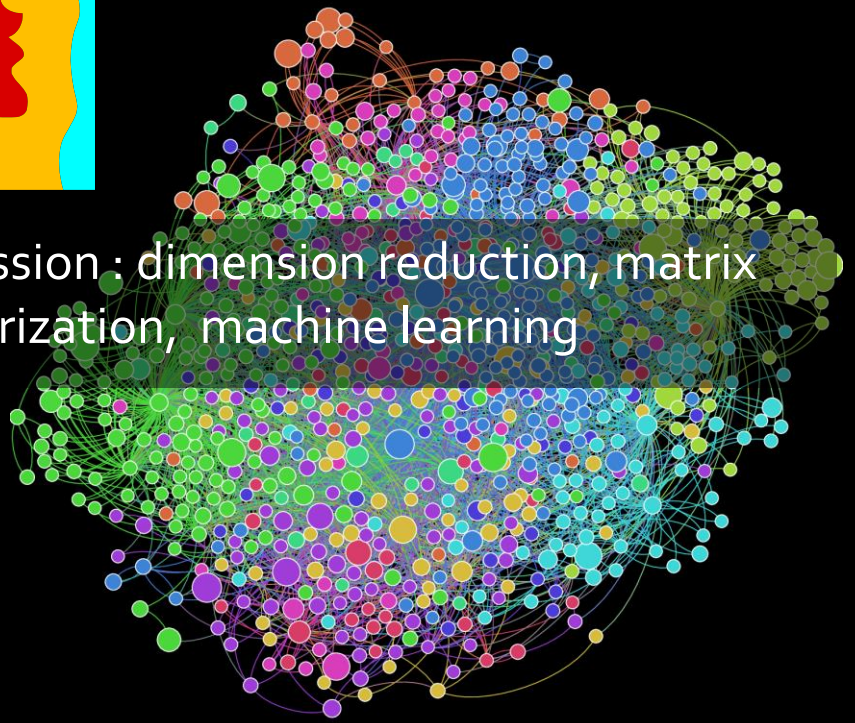
Baroque combination of complex metadata and various raw data file formats

Similar challenges

- SDSS spectra: 1 million times 3000 dimensional vectors
- Microarray study: 207 times 54675 dimensional vectors
- Human genome: 3.2Gb



Compression : dimension reduction, matrix factorization, machine learning

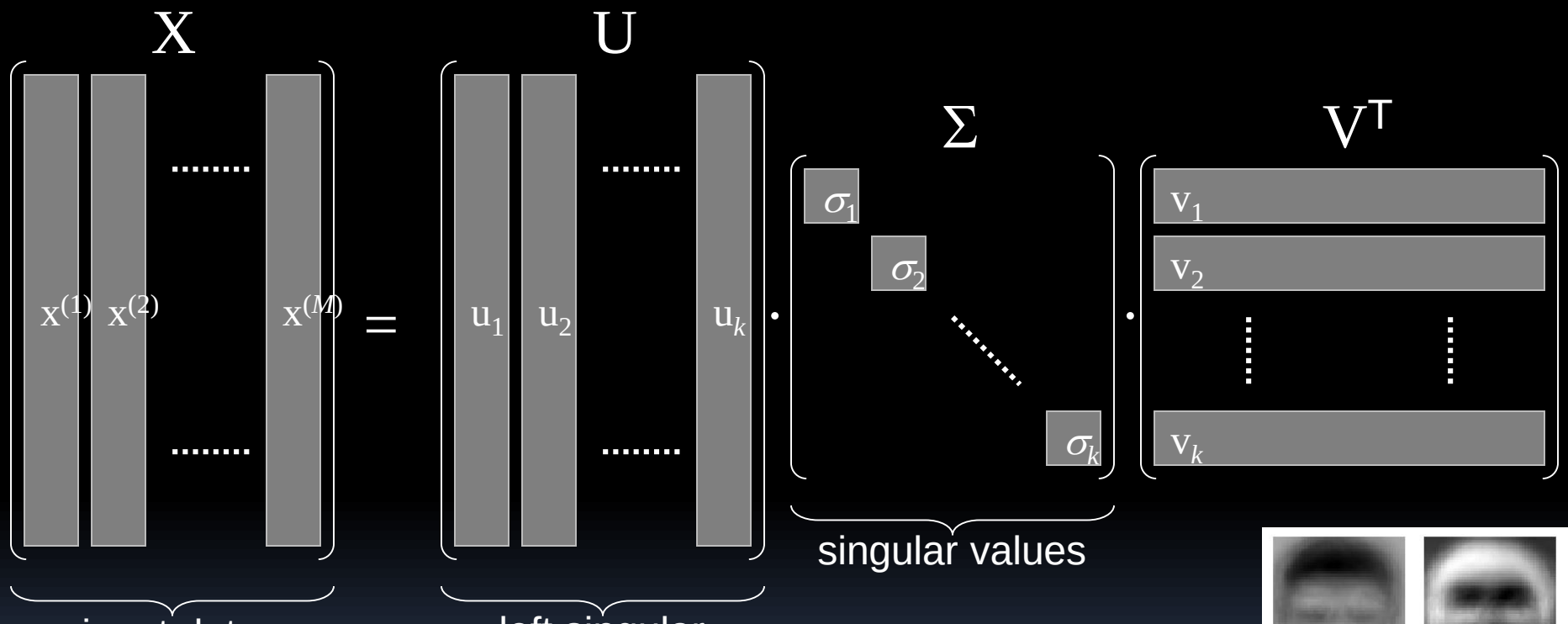


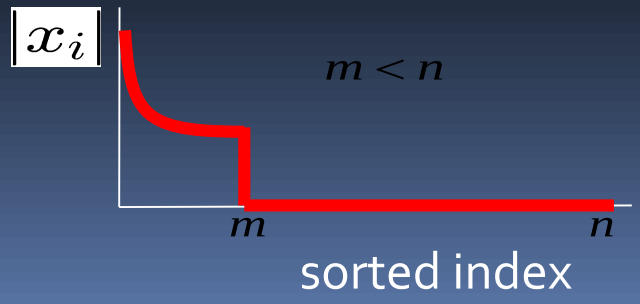
Due to the underlying physical laws, data vectors does not fill the whole space, rather lie on lower dimensional surface/subspace
(this is why we can understand the word!)

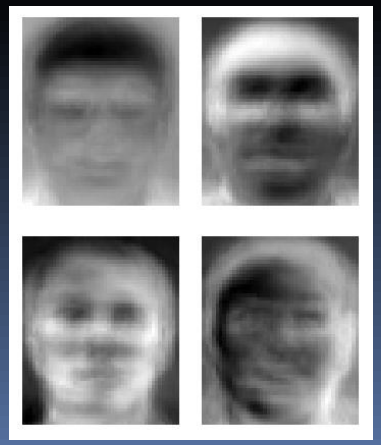
Projection ~ compression ~ model

Linear projection: PCA - SVD

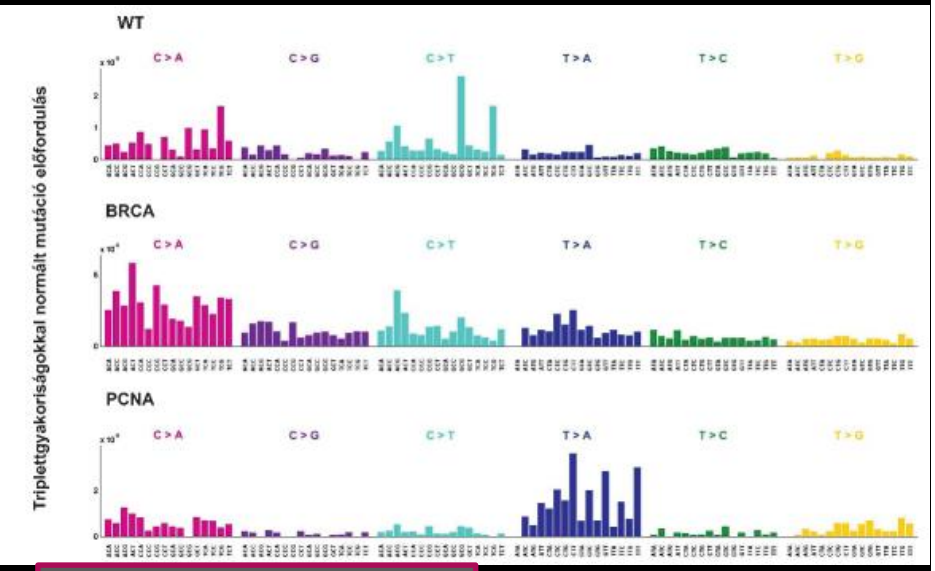
$$X = U\Sigma V^T$$



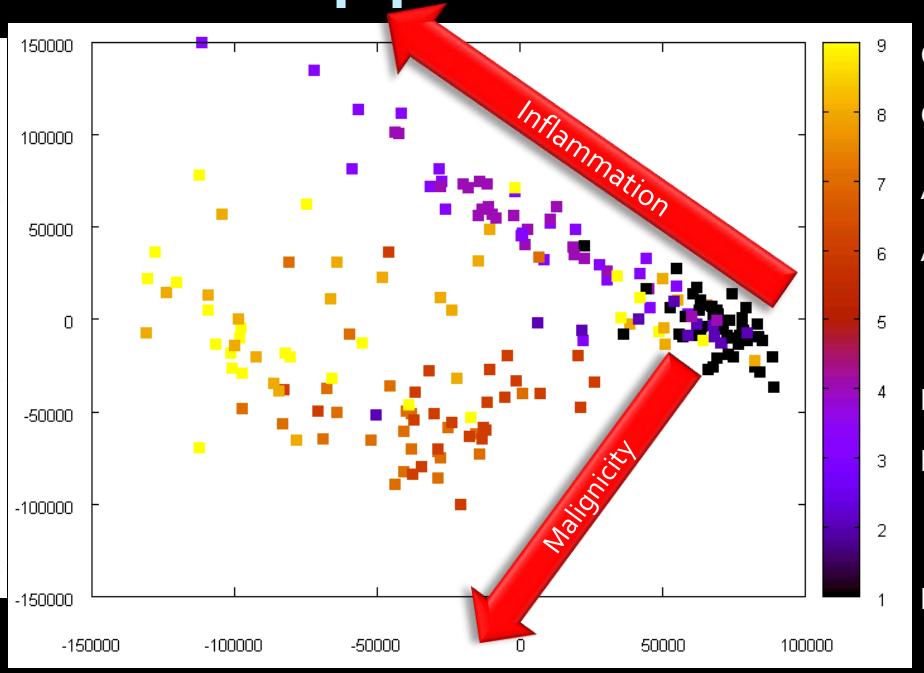


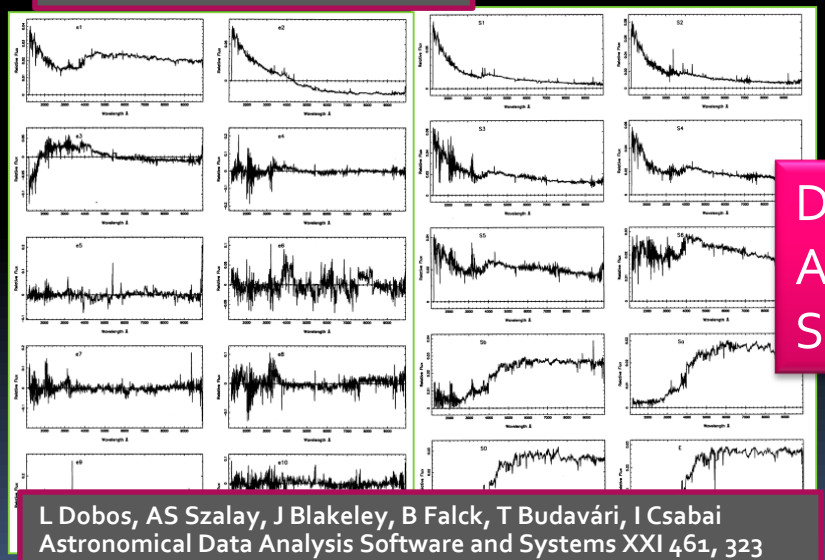


Dimension reduction: applications



O Pipek, MSc thesis ELTE (2014)





Data processing: Array library for SQL Server

L Dobos, AS Szalay, J Blakeley, B Falck, T Budavári, I Csabai Astronomical Data Analysis Software and Systems XXI 461, 323 (2012)

T Budavári, I Csabai + SDSS collab.; Astr. J. 122 (3) 1163(2001)

R. Beck, L. Dobos, C. Yip, A. Szalay, I. Csabai; MNRAS 457 (1), 362-374 (2016)

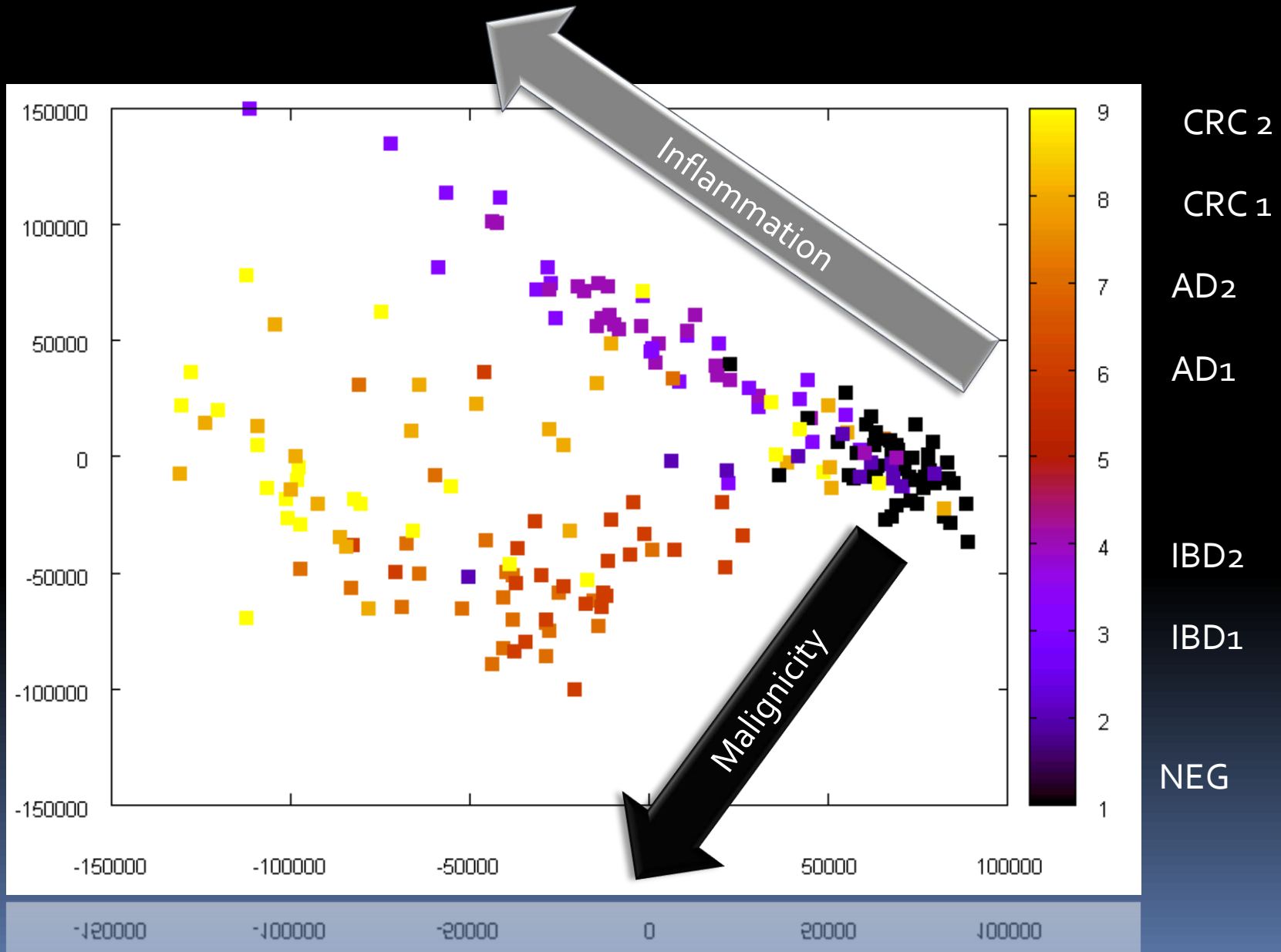
- Algorithm developments
  - Handling outliers: robust PCA
  - Big data: "streaming" / DB
  - Sparse data , compressive sensing
  - CUR decomposition
  - Graph PCA/NNMF
  - Text mining

I Csabai, AJ Connolly, AS Szalay, T Budavári; Astr. J. 119 (1), 69 (2000)

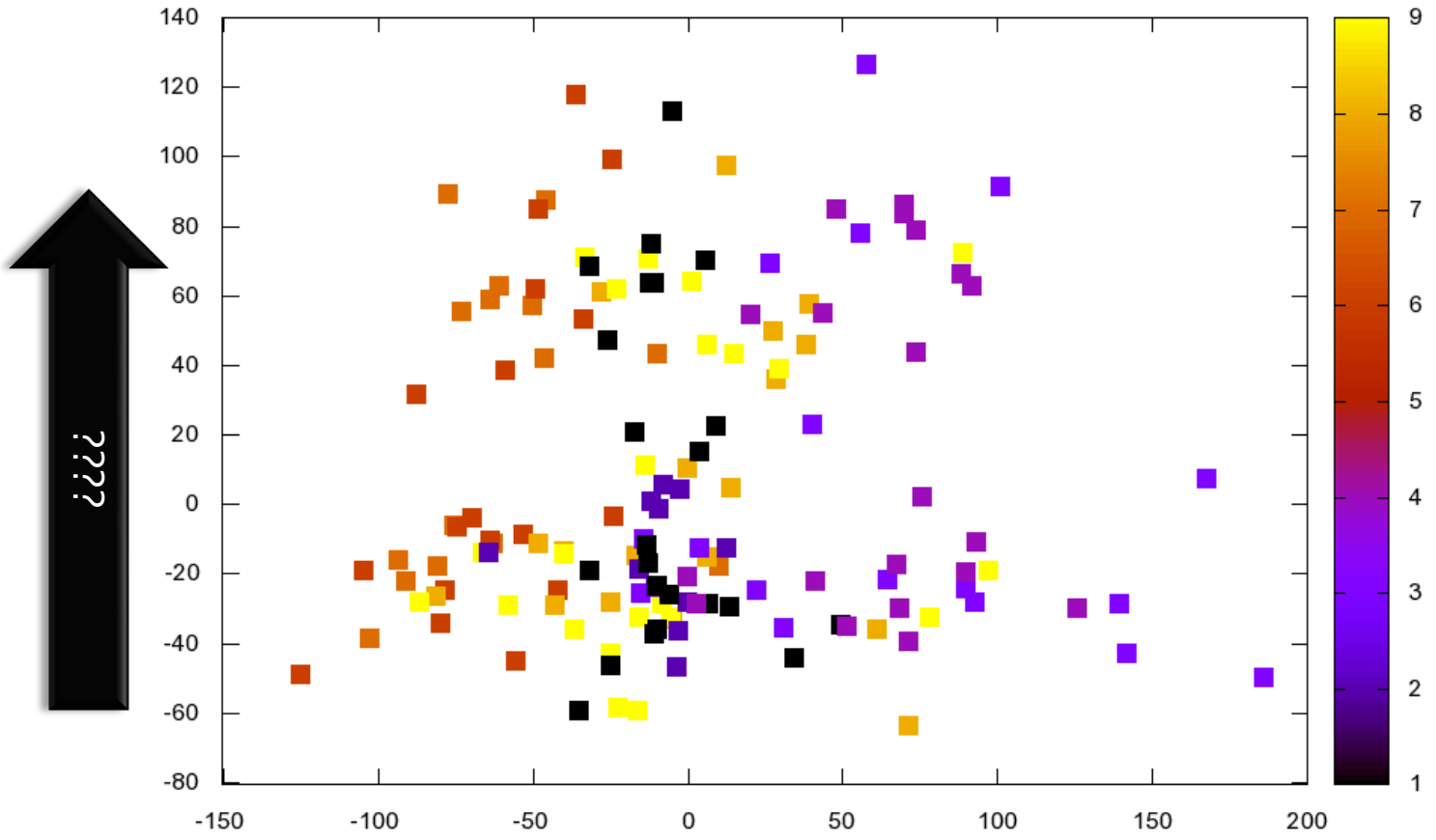
Z Győry, AS Szalay, T Budavári, I Csabai, S Charlot; Astron. J. 141 (4) 133 (2011)

S Spisák, A Kalmár, O Galamb, B Wichmann, F Sipos, B Péterfia, I Csabai, I Kovalszky, S Semsey, Z Tulassay, B Molnár; PloS one 7 (10), e46215(2012)

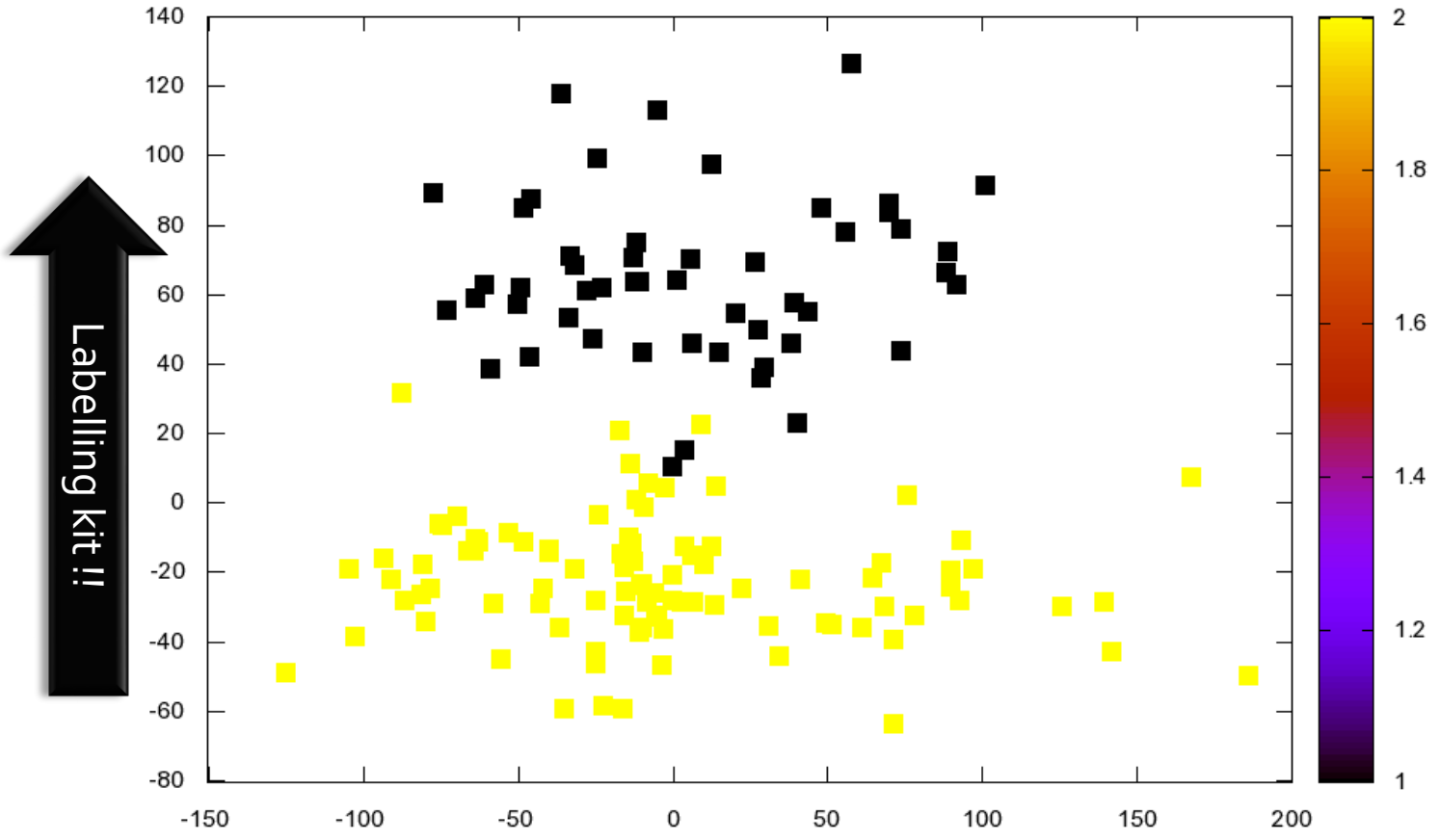
Expression microarray: 54675D -> 2D



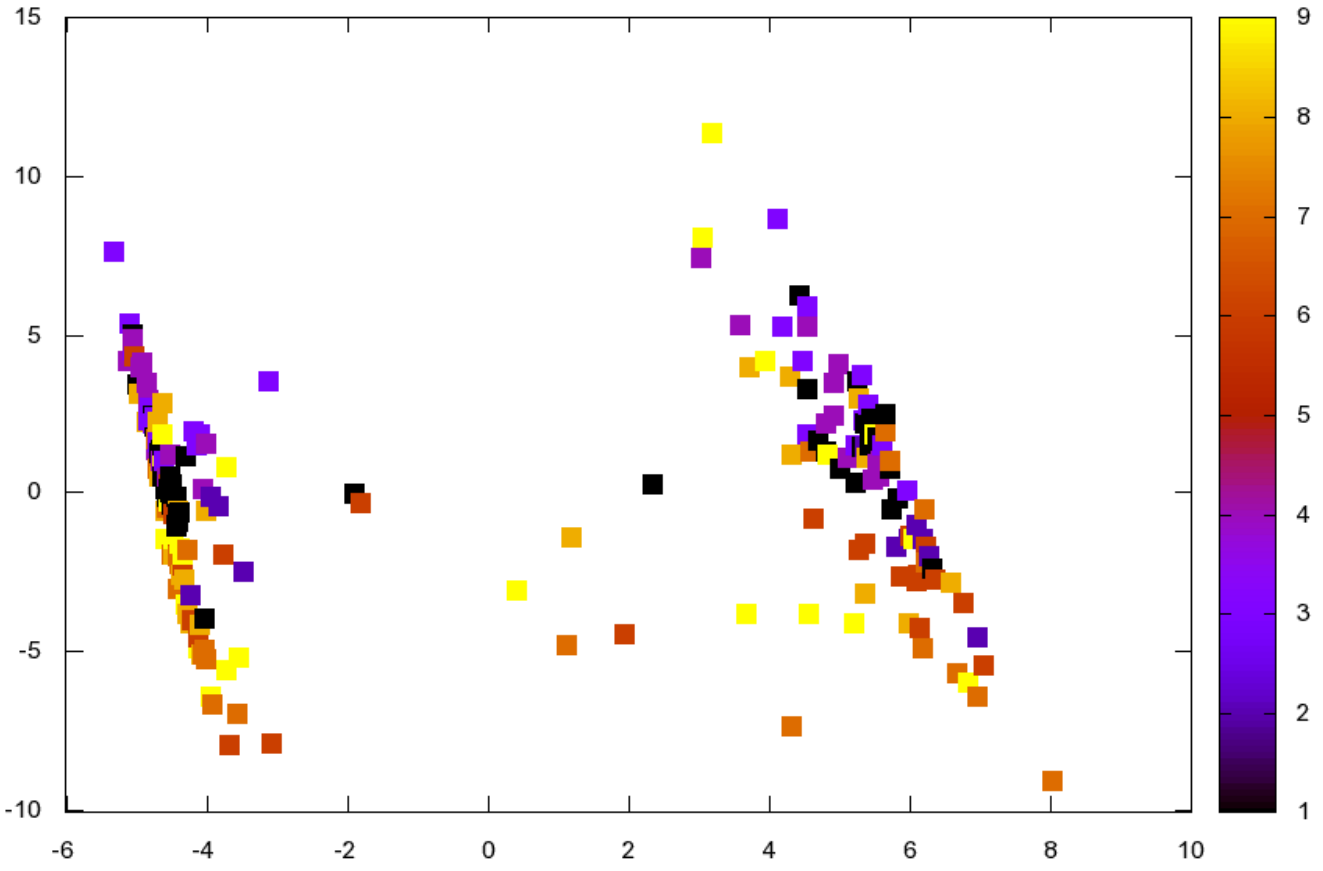
PCA2, PCA3 clusters?



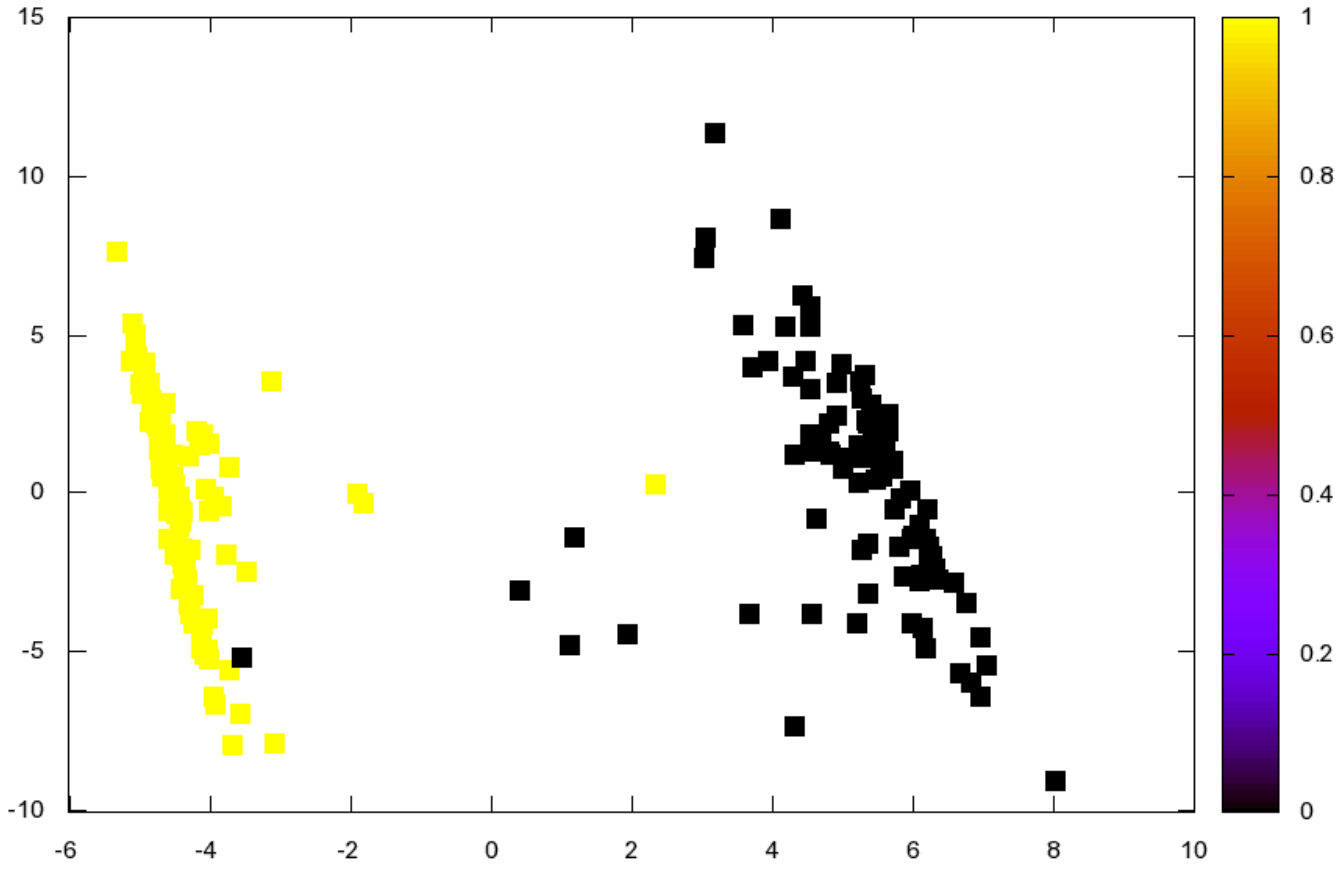
PCA2, PCA3 clusters



PCA – KEGG pathways (ribosome)



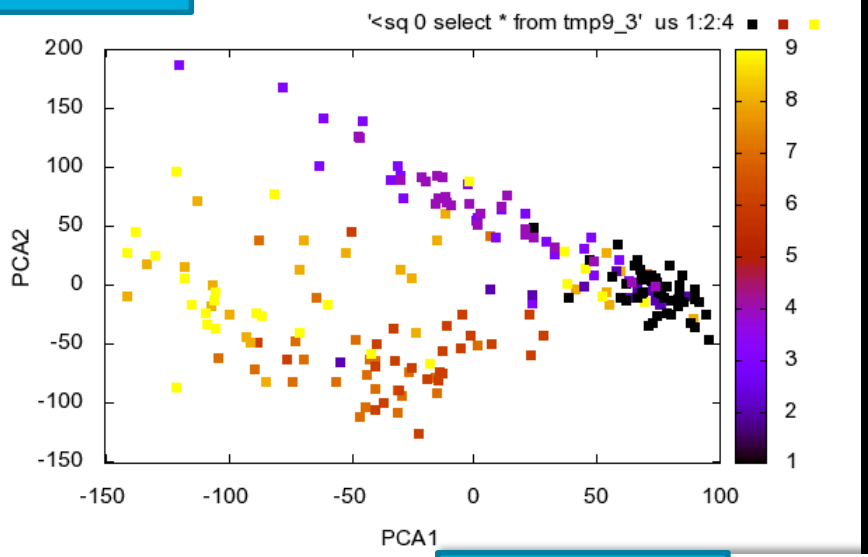
PCA – KEGG pathways (ribosome)



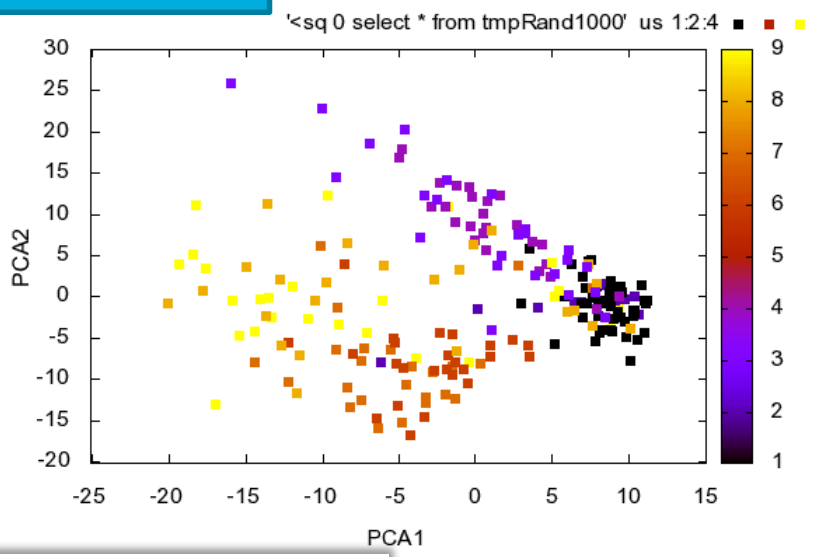
“Realize that everything connects to everything else.”

/Leonardo da Vinci/

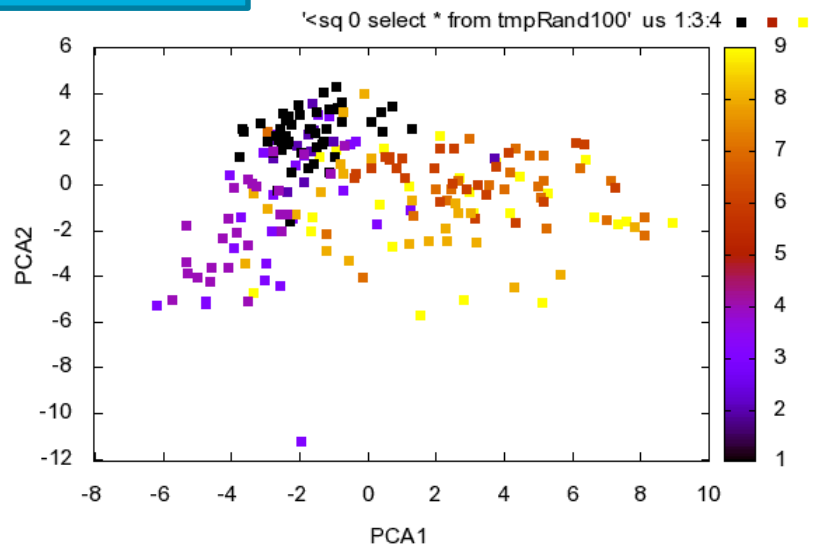
All 54265



Random 1000



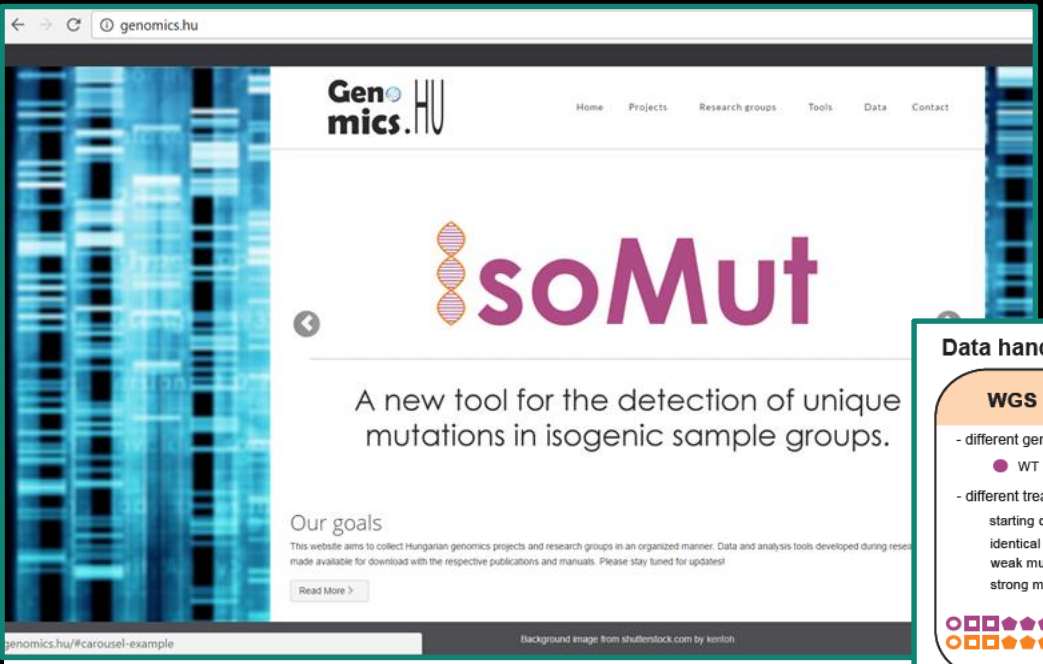
Random 100



Marker genes?
Complex nonlinearly interacting network!

RECENT PROJECTS

Variant detection in isogenic NGS samples



Pipek *et al. BMC Bioinformatics* (2017) 18:73
DOI 10.1186/s12859-017-1492-4

BMC Bioinforma

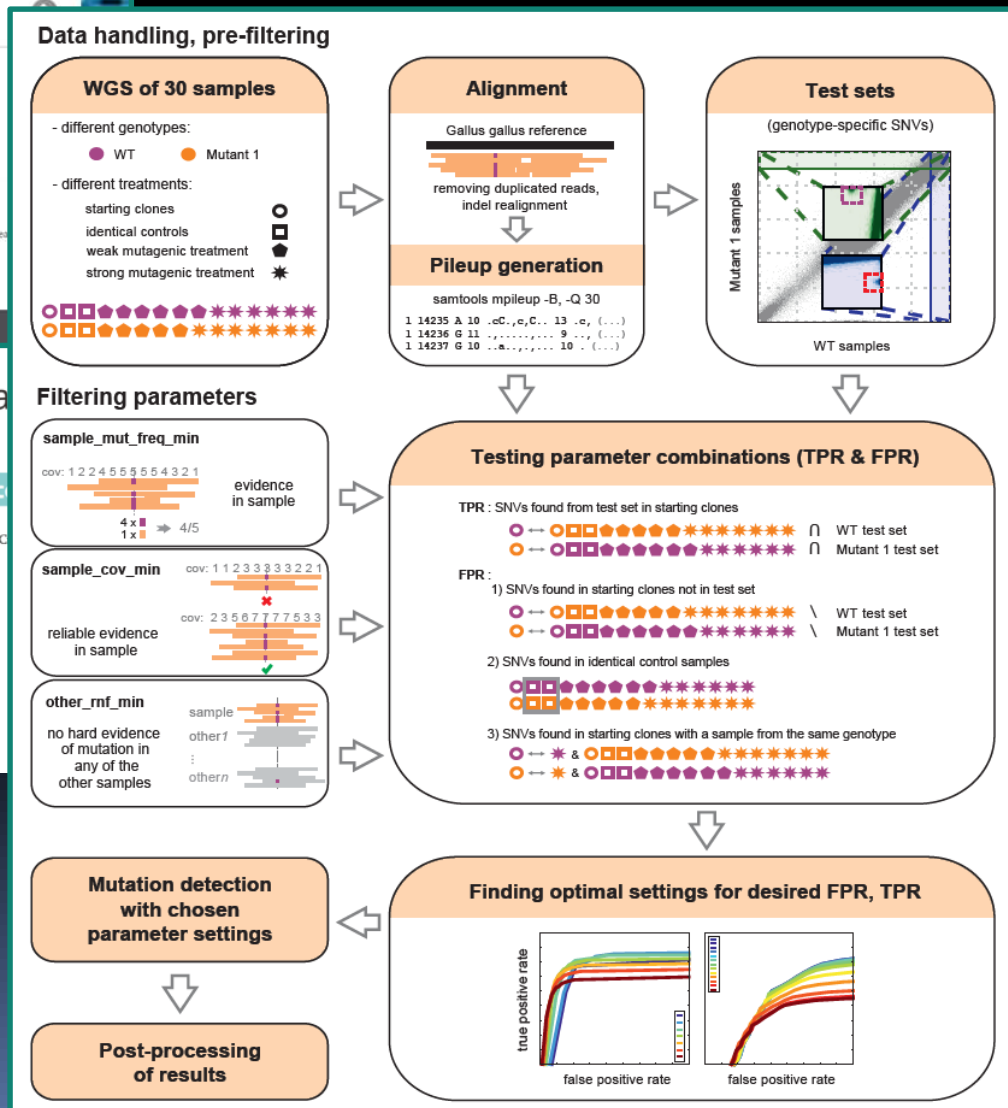
METHODOLOGY ARTICLE

Open Ac

Fast and accurate mutation detection in whole genome sequences of multiple isogenic samples with IsoMut

O. Pipek[1], D. Ribli[1], J. Molnár[2], Á. Póti[2], M. Krzystanek[3], A. Bodor[1], G. E. Tusnády[2], Z. Szallasi[3,4,5,6], I. Csabai[1] and D. Szüts[2*]

Tool to separate germline variation from germline variants and sequencing/alignment noise



www.nature.com/onc/journal/vaop/ncurrent/full/onc2016243a.html

nature.com ▸ Publications A-Z index ▸ Browse by subject

Login Cart

Oncogene

Search go Advanced search

Journal home > Advance online publication > 25 July 2016 > Full text

Journal home
Advance online publication
About AOP
Current issue
Archive
Press releases

Online submission
For authors
For referees

Original Article

Oncogene advance online publication 25 July 2016; doi: 10.1038/onc.2016.243

Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions

OPEN

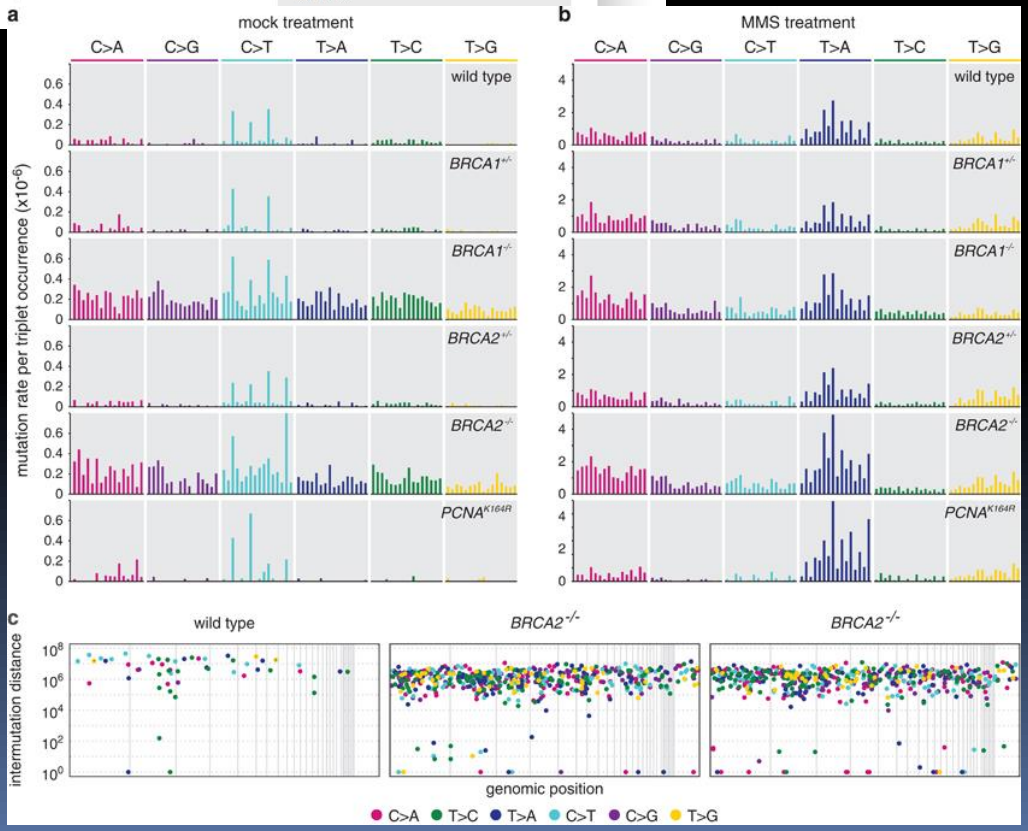
J Zámborszky[1], B Szikriszt[1], J Z Gervai[1], O Pipek[2], Á Póti[1], M Krzystanek[3], D Ribli[2], J M Szalai-Gindl[2], I Csabai[2], Z Szallasi[3,4,5,6], C Swanton[7,8], A L Richardson[9] and D Szüts[1]

FULL TEXT
Table of contents
Download PDF
Share this article
View interactive PDF in ReadCube
Rights and permissions
Order Commercial Reprints
Abstract
Introduction
Results

novo nordisk®

D. Szűts,
Z. Szállási

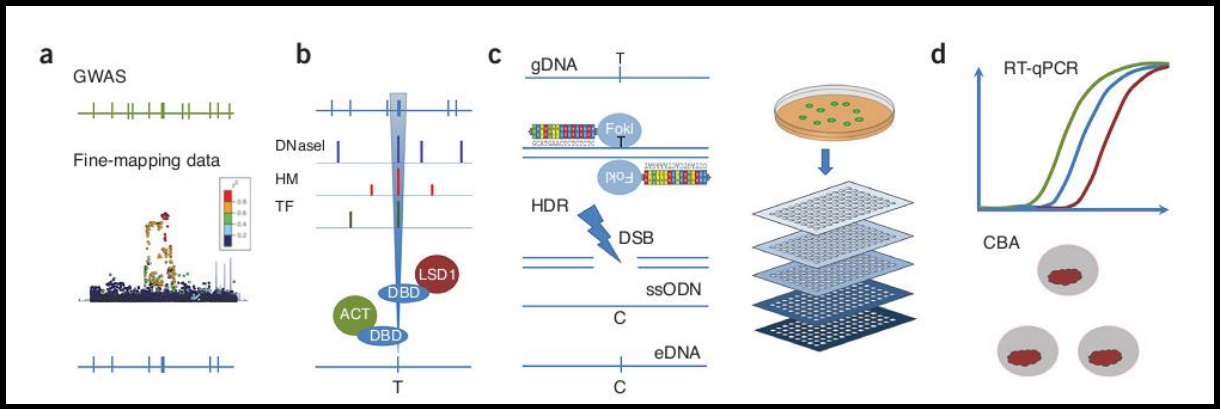
- understanding DNA-repair mechanisms
- gene knock-out cell lines, mutagene treatments
- mutational signatures with non-negative matrix factorization
- mutation spectra – compare to TCGA

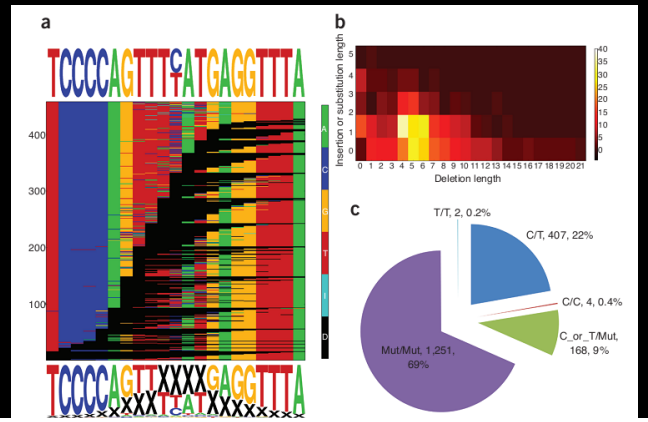


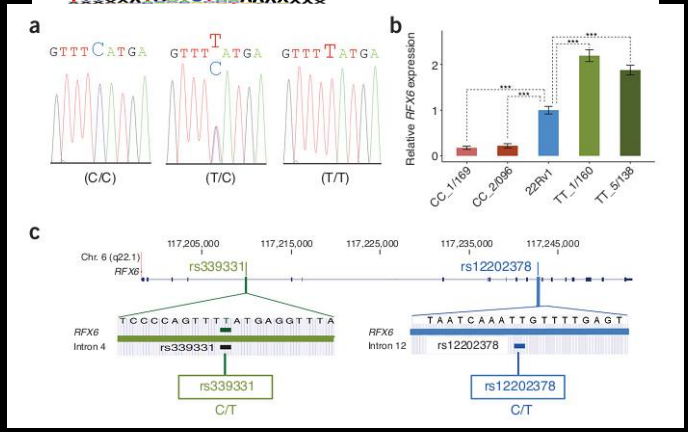
nature medicine

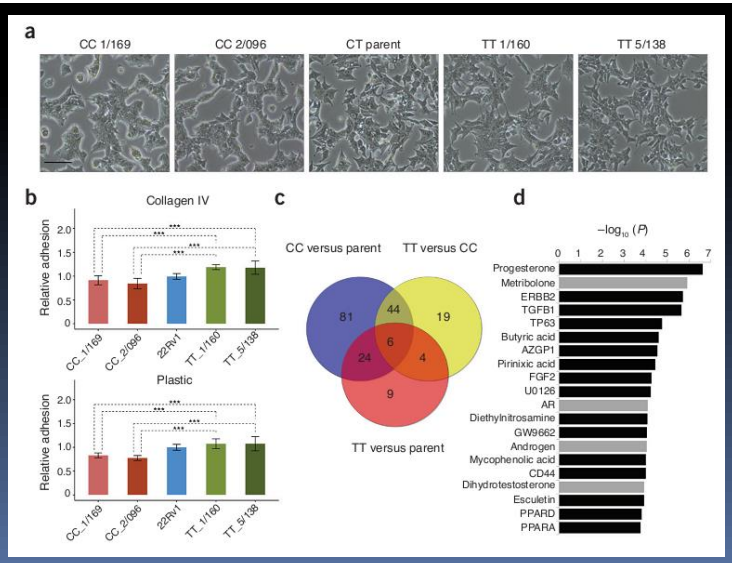
CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants

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



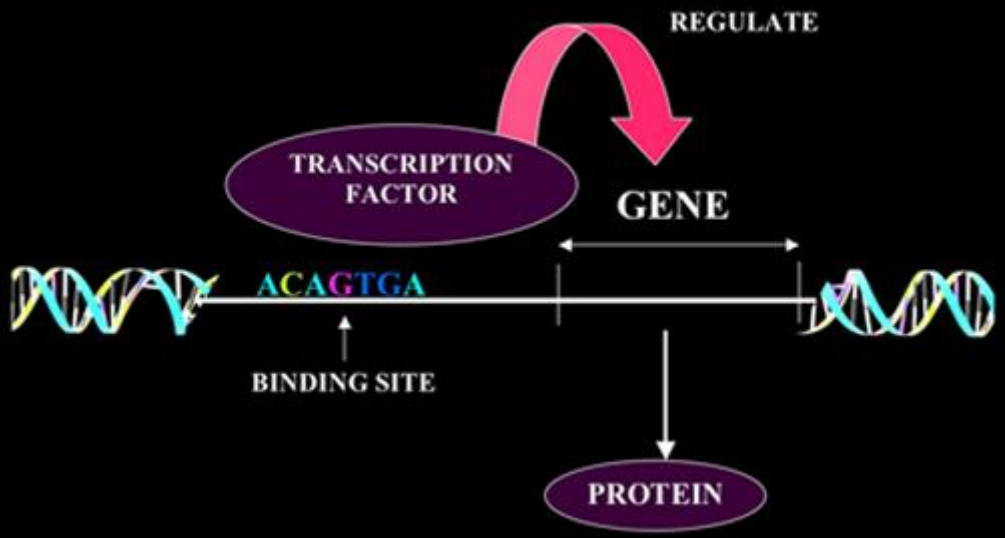




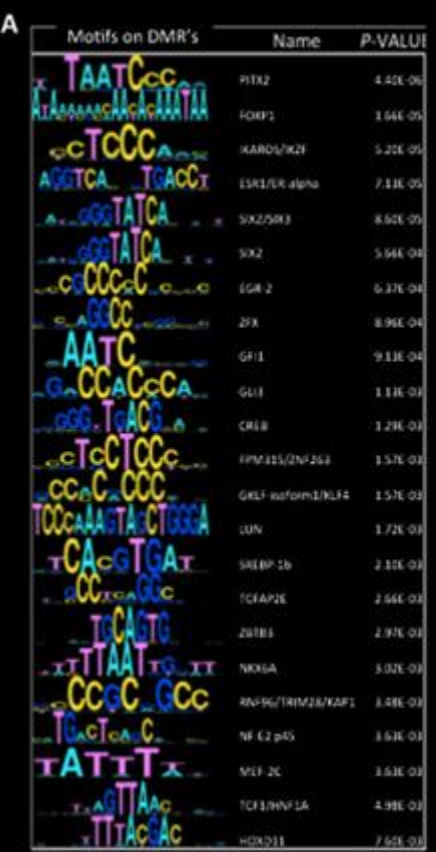


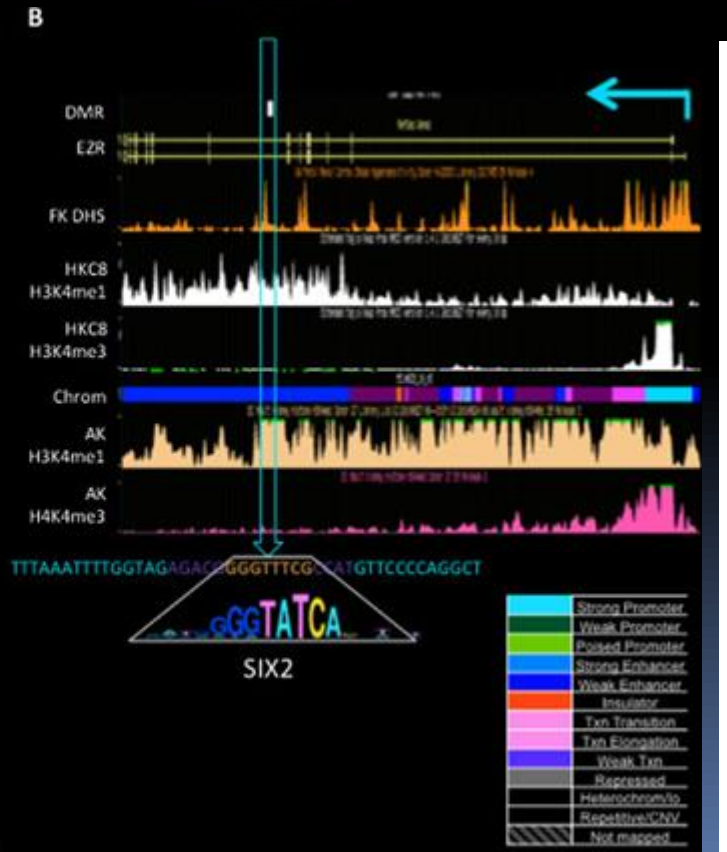
- Genome editing techniques: TALEN, CRISPR-CAS9
- "no averaging" phenomenon:

Strong coupling of micro and macro scales: a single nucleotid in non-coding region can cause phenotypic change

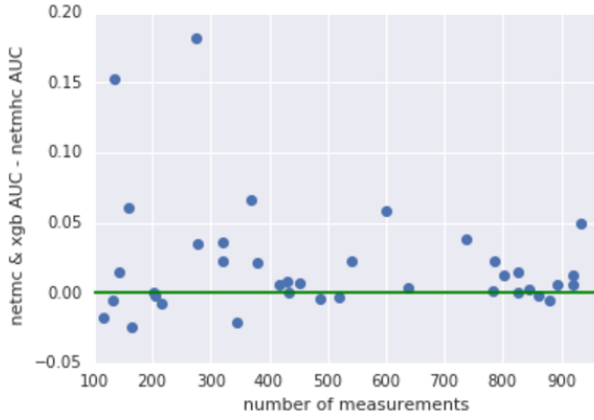


Transcription factor & MHC binding with machine learning



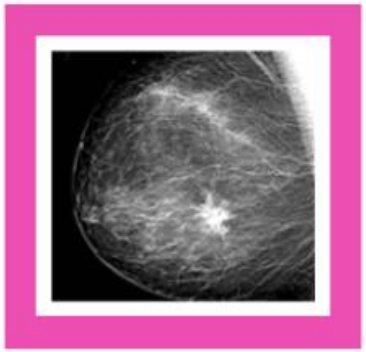


In [240]:
```
plt.plot(all_auc_df['counts'],all_auc_df['netmhc_
plt.axhline(0,c='g')
plt.xlabel('number of measurements')
plt.ylabel(' netmc & xgb AUC - netmhc AUC')
_=plt.xlim(100,1000)
```



In [242]:

```
NetMHC AUC: 0.462322635593
my nn mean AUC: 0.466039091357
my xgb mean AUC: 0.465527709089
my nn+xgb mean AUC: 0.47319931733
NetMHC + xgb mean AUC: 0.476159451622
```



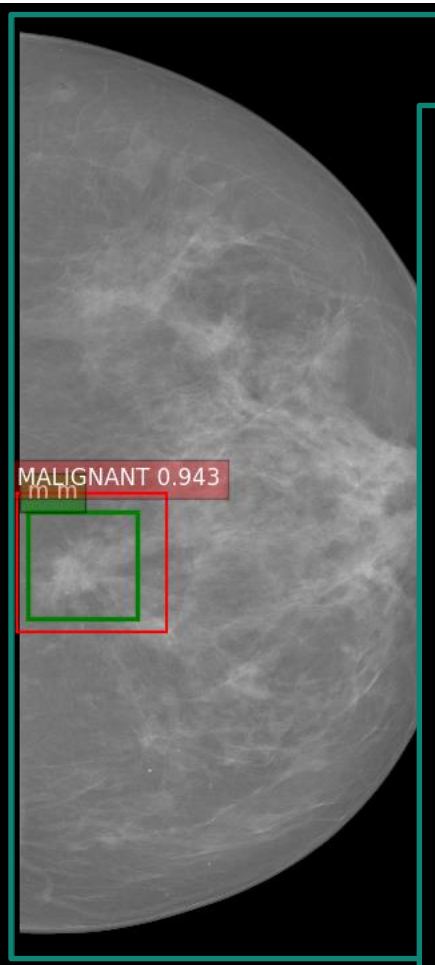
The Digital Mammography DREAM Challenge

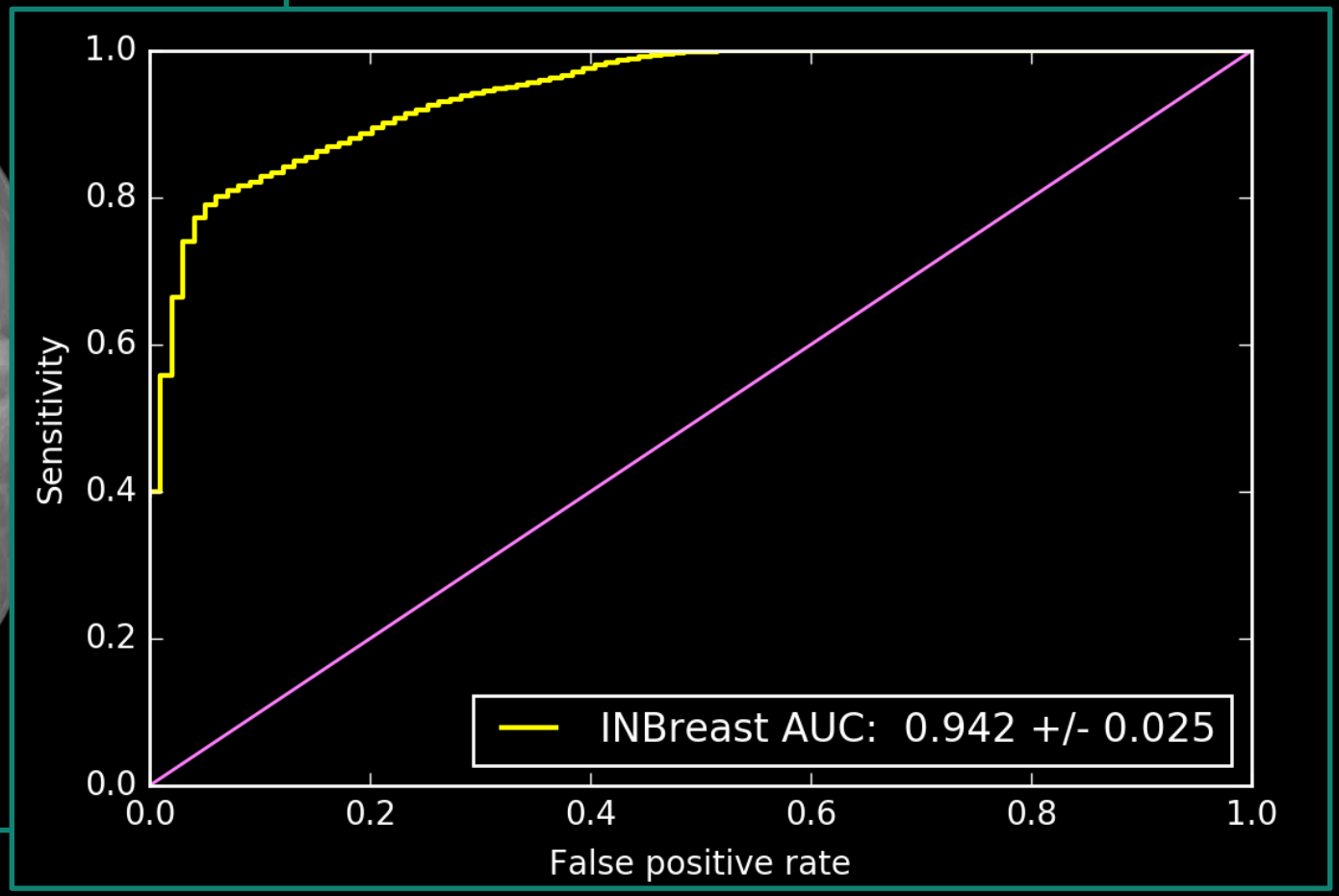
Build a model to help reduce the recall rate for breast cancer screening

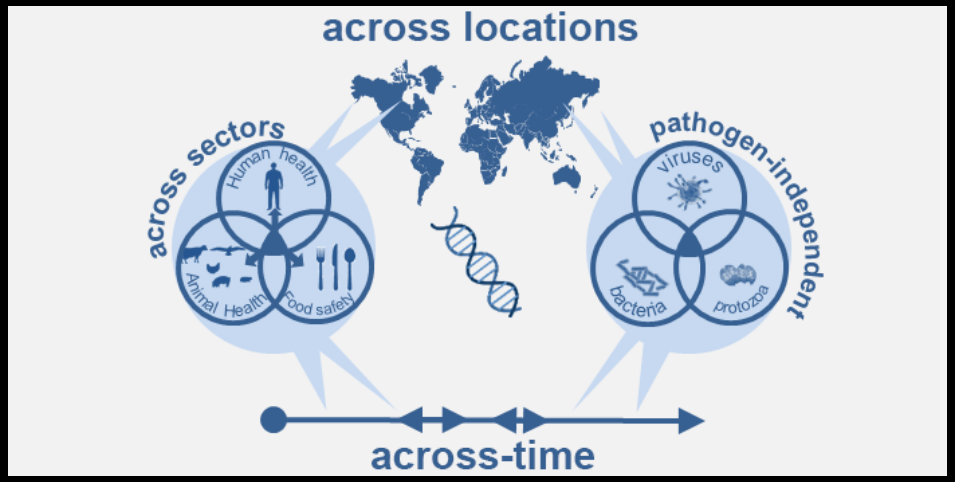
Learn more & register to participate here: www.synapse.org/Digital_Mammography_DREAM_Challenge

Funded by: ljaf

Enabled by: DREAM Challenges, Sage, Coding4Cancer, GroupHealth Research Institute, IBM, BCSC, amazon webservices, Icahn School of Medicine at Mount Sinai, Seattle Cancer Care Alliance, RADISH

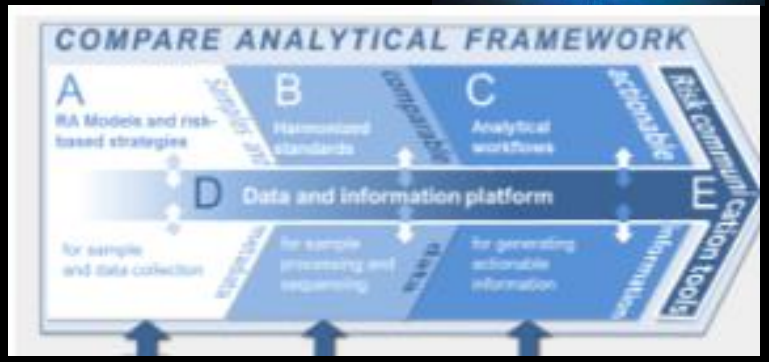






compare

HORIZON 2020



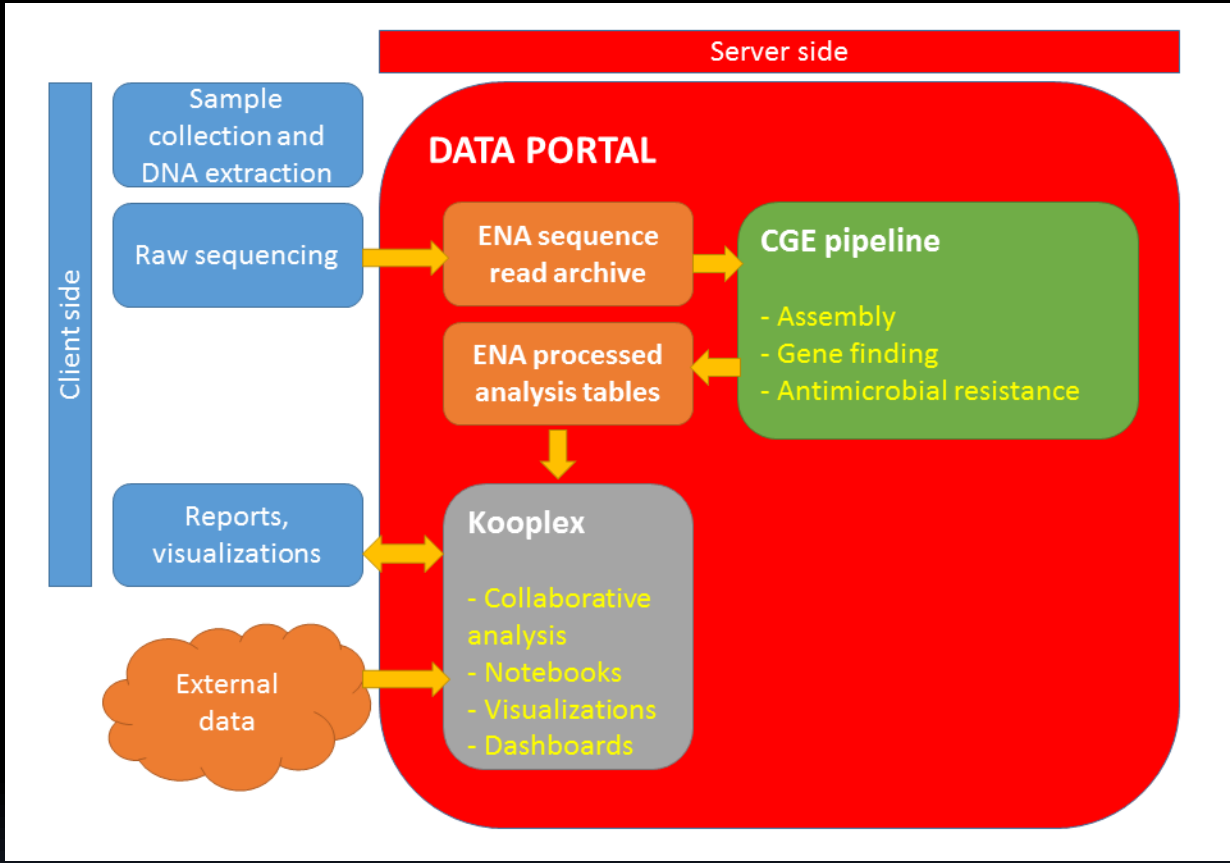
- improve rapid identification, containment and mitigation of infectious diseases and foodborne outbreaks
- cross-sector and cross-pathogen analytical framework and globally linked data- and information-sharing platform
- integrate state-of-the-art strategies, tools, technologies and methods for collecting, processing and analysing sequence-based pathogen data in combination with associated data (clinical, epidemiological, and other), and
- generate actionable information for relevant authorities and other users in human health, animal health and food safety domains.

Examples

Metagenomics - sewage samples from ~100 countries

Antimicrobiotic resistance prediction with machine learning

Cloud based Data Portal



Big Data (EBI ENA 5PB!) – downloading data is not optimal/possible

Data sharing is not enough

– share data + complete processing pipeline + result figures, tables, …

-> reproducible science

Collaborative data analytics

Kooplex

Infrastructure for flexible collaboration

Worksheets

Run pre-compiled worksheets to hide program code and focus on the problem.

Worksheets hide the complexity of notebooks from end users interested in scientific output.

→ to worksheets

Jupyter notebooks

Run existing or create Jupyter notebooks from existing projects to process your data or author your own projects, notebooks and share with others.

→ Browse notebooks

Gitlab

Manage your project, add members to it, file issues.

→ to Gitlab

Owncloud

Manage easily your data files through the owncloud... (just as easy as in Dropbox! :))

→ to Owncloud

www.compare-europe.hu/ena_europe_loco.html

```
In [12]: width, height = 650, 500
         flu_map = folium.Map(location=[47, -17], zoom_start=3,
                              tiles='OpenStreetMap', width=width, height=height)
```

Add point to the map object

- Let's make point area proportional to number of cases
  - This is miseleading, beacuse somewhere all the cases around have the sample position (Europe), and somewhere the positions are more scattered (Shanghai)

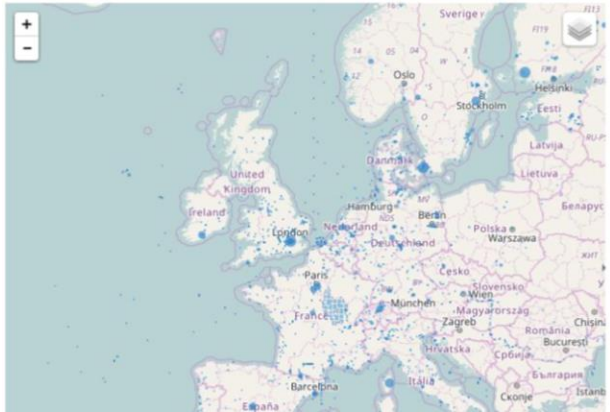
```
In [13]: for i in xrange(len(uniq_locs_w_acc)):
             loc=(uniq_locs_w_acc.iloc[i]['lat'],uniq_locs_w_acc.iloc[i]['lon'] )
             name='Number of cases: '+str(uniq_locs_w_acc.iloc[i]['count'])
             name+='   Accesions: '+uniq_locs_w_acc.iloc[i]['acc_list']
             size=uniq_locs_w_acc.iloc[i]['count'] ** 0.5

             flu_map.circle_marker(location=loc, radius=1e3*size,
                                   line_color='none',fill_color='#3186cc',
                                   fill_opacity=0.7, popup=name)
```

And finally draw the map

```
In [15]: inline_map(flu_map)
```

Out[15]:

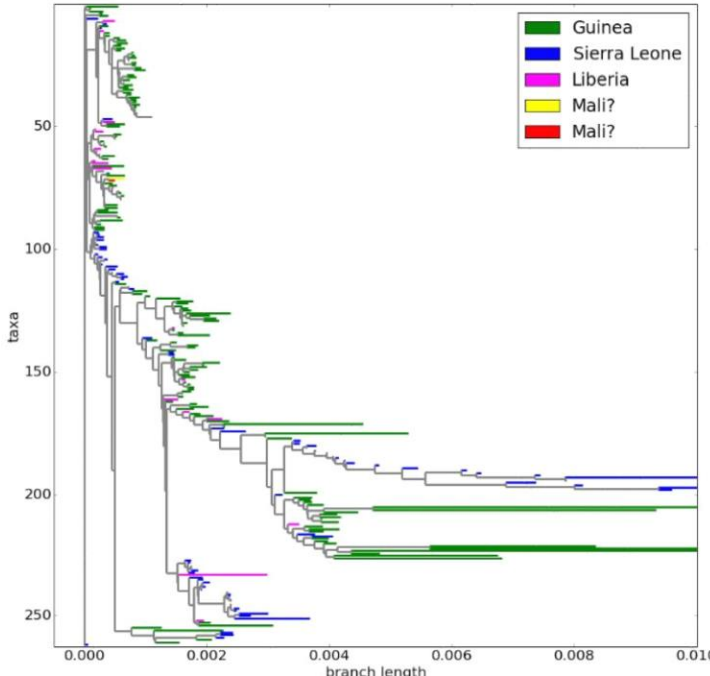



```
%matplotlib inline

#some settings
matplotlib.rc('font', size=20)
matplotlib.rcParams['lines.linewidth'] = 3
matplotlib.rcParams['figure.figsize'] =(16,16)

fig,ax=plt.subplots()
#custom legend
gui_proxy = plt.Rectangle((0, 0), 1, 1, fc="green")
sle_proxy = plt.Rectangle((0, 0), 1, 1, fc="blue")
lib_proxy = plt.Rectangle((0, 0), 1, 1, fc="magenta")
dpr2_proxy = plt.Rectangle((0, 0), 1, 1, fc="yellow")
dpr1_proxy = plt.Rectangle((0, 0), 1, 1, fc="red")
ax.legend([gui_proxy,sle_proxy,lib_proxy,dpr2_proxy,dpr1_proxy],
          ['Guinea','Sierra Leone','Liberia','Mali?','Mali?'])

#draw tree
def my_label(clade):
    return None
Phylo.draw(tree,my_label,axes=ax,xlim=(-0.0005,0.01))
```



New national R&D projects

Biomarker research
ELTE, MTA TTK, Servier, CRU

NEMZETI KUTATÁSI, FEJLESZTÉSI ÉS INNOVÁCIÓS HIVATAL

Cancer genomics, liquid biopsy
SOTE, ELTE, 3DHISTECH

semmelweis.hu/patologia1/2017/02/19/magyar-onkogenom-program-indul-intez

Magyar Onkogenom és Személyre Szabott Diagnosztika és Terápia program indul Intézetünk irányításával

Közzétéve: 2017. február 19. vasárnap

A Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (NKFIH) 1.5 milliárd forintos támogatásával indul útjára Intézetünk koordinálásával a **Magyar Onkogenom és Személyre Szabott Diagnosztika és Terápia Program.**

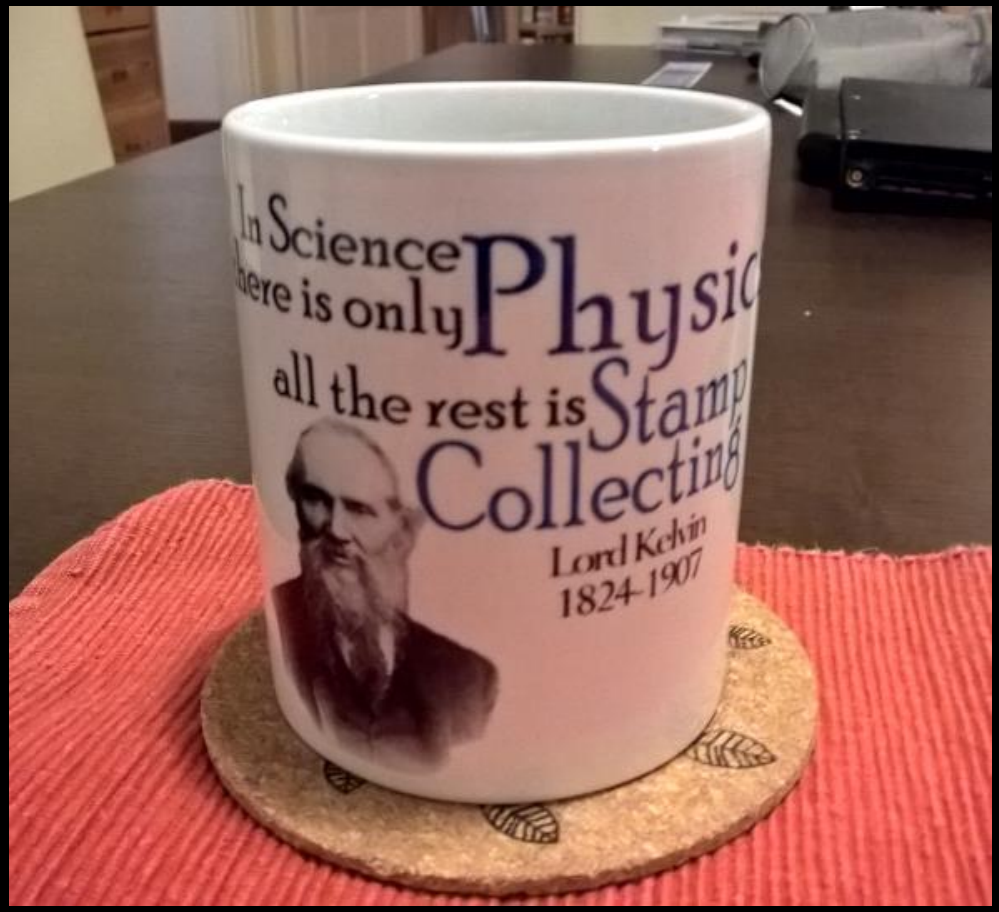
Not only DNA sequencing, but

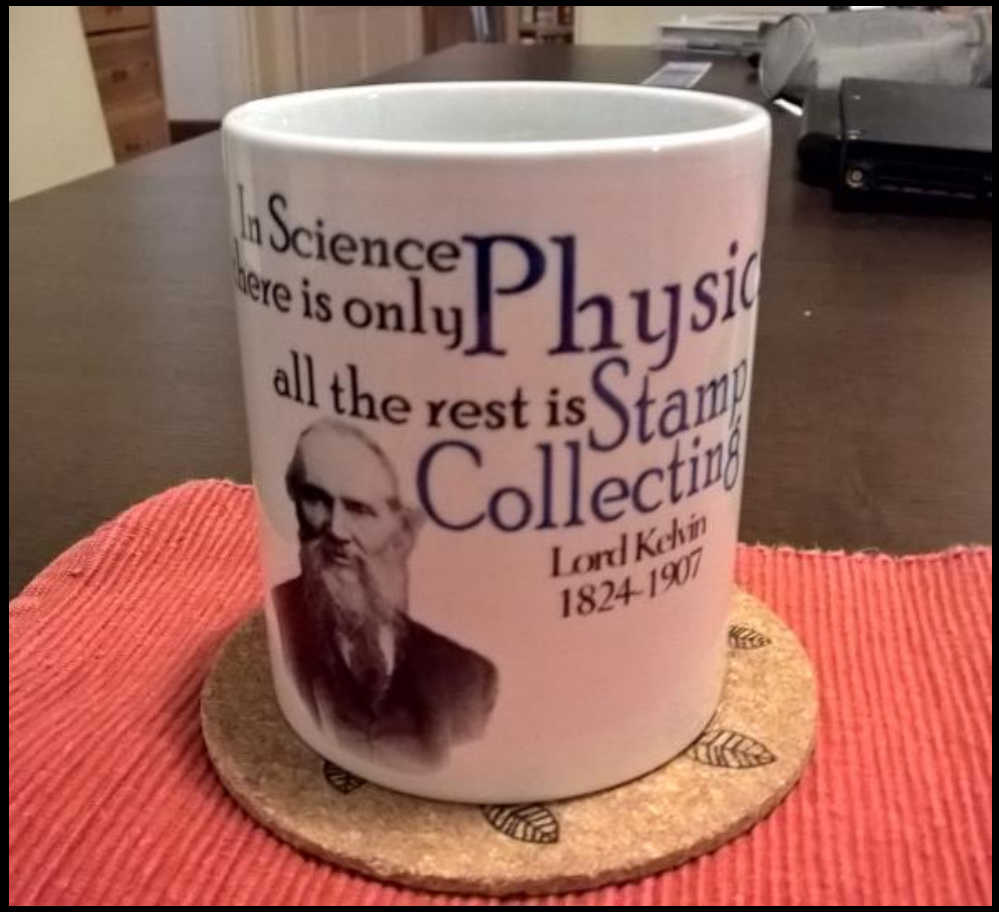
- Methylation, ncRNA, 3D structure, …
- Proteomics, lipidomics, …
- Digital microscopy, medical imaging, …

Not just sciences, but Everything

- Smart watches, wearable EEG, personal genome sequencers -> better health
- Sensors for cars -> less death on roads
- Sensors for sports -> more enjoyment
- …

More “sensors” – more data – better “models” – better life


In Science
there is only Physic
all the rest is Stamp
Collecting
Lord Kelvin
1824-1907



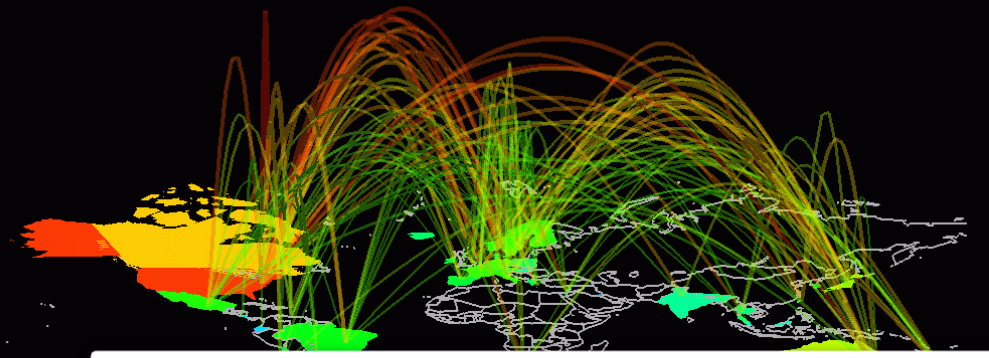
Sorry Dr. Thomson,
everything is science,
maybe even stamp collecting,
too.

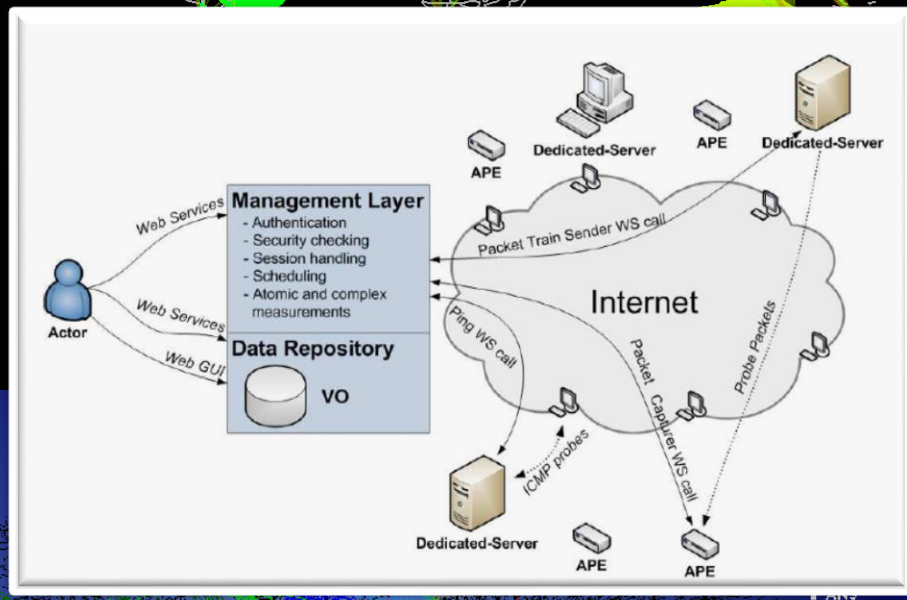
Manmade complex systems

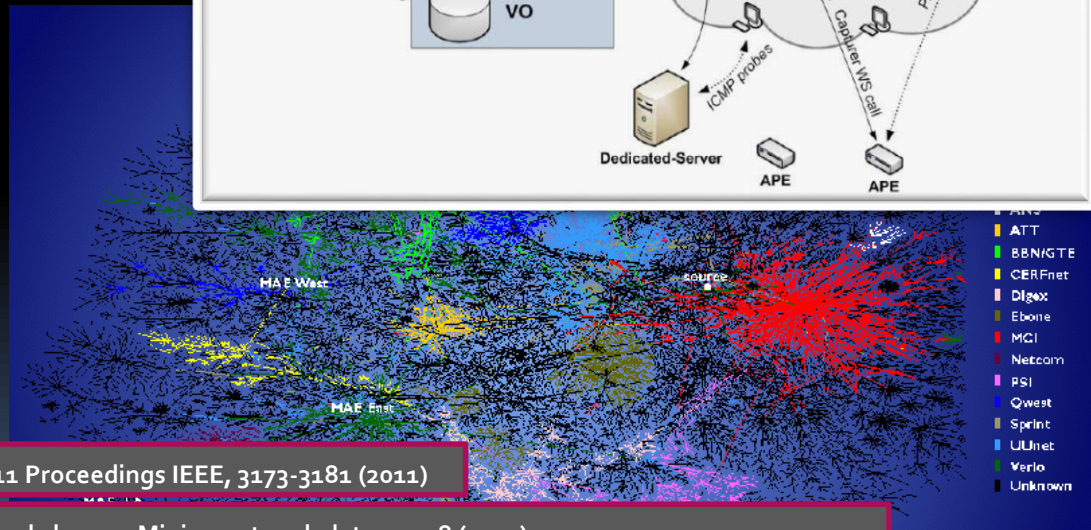
COMMUNICATION- SOCIAL- AND FINANCIAL NETWORKS

Map of Internet

- Manmade, but there is no "blueprint"
- "Astronomical" number of non-linearly interacting complex elements
- Scientific approach is required
  - Observation/experiment
  - Modeling
  - -> plan better
- Future internet: self-aware, self-managing, self-healing …







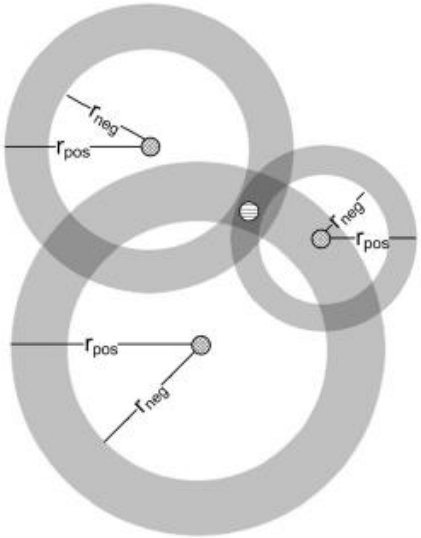
S Laki, P Mátray, P Hága, T Sebők, I Csabai, G Vattay; INFOCOM, 2011 Proceedings IEEE, 3173-3181 (2011)

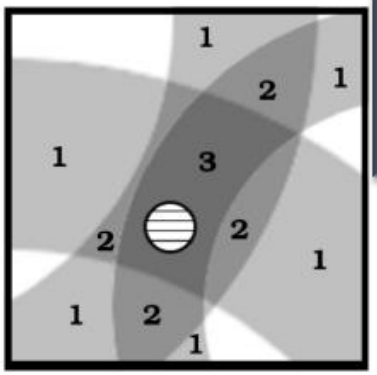
P Matray, I Csabai, P Haga, J Steger, L Dobos, G Vattay; Proc. ACM workshop on Mining network data, 23-28 (2007)

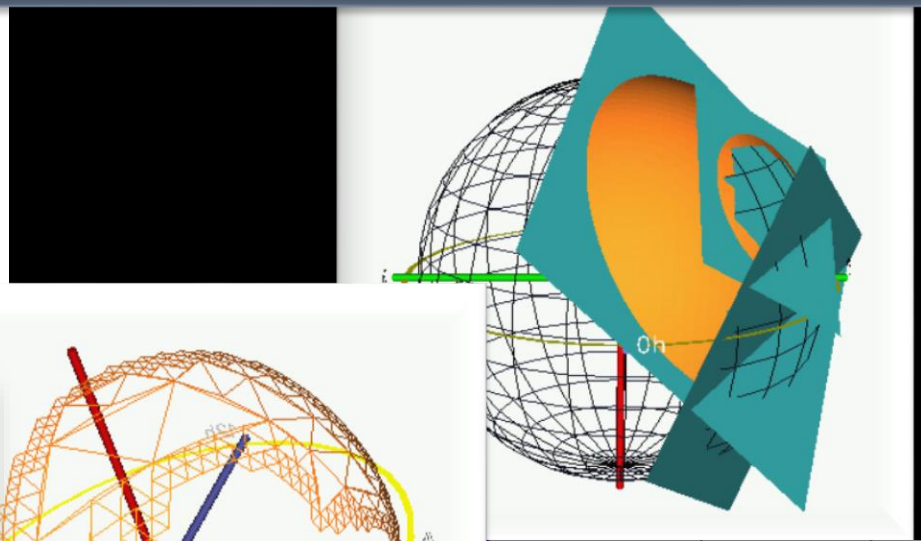
D Morato, E Magana, M Izal, J Aracil, FJ Naranjo, P Astiz, U Alonso, I Csabai, P Hága, G Simon, J Stéger, G Vattay; TRIDENTCOM, 283-289 (2005)

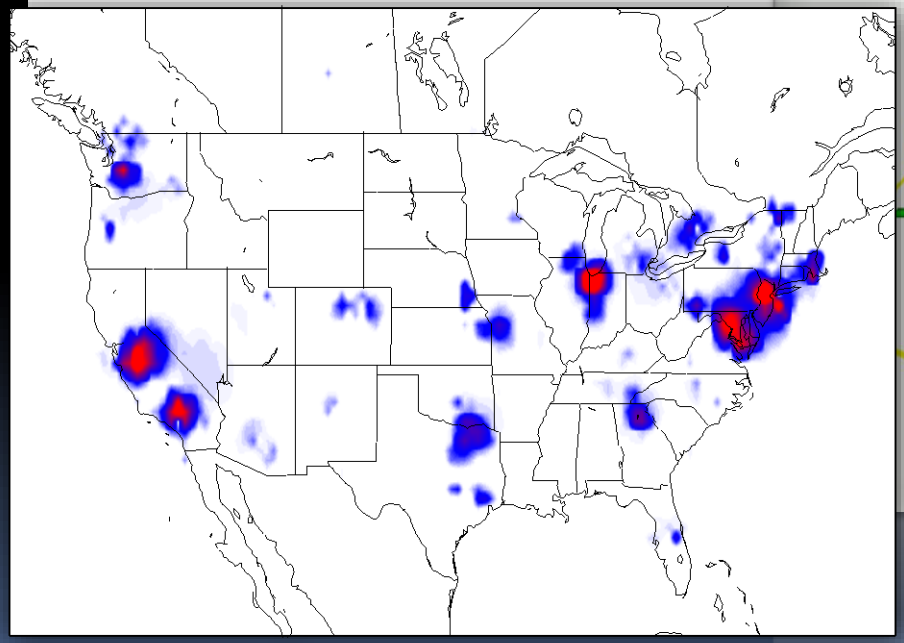
J Szüle, L Dobos, I Csabai, G Vattay; TRIDENTCOM, 137, 65 (2014)

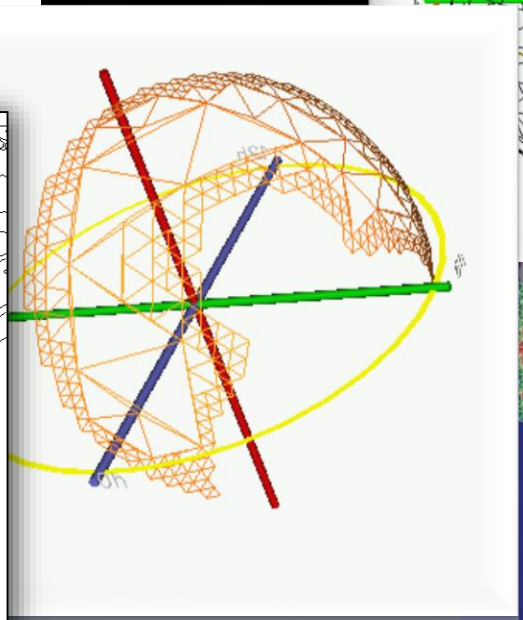
Reuse of celestial indices:
HTM index library + SQL Server integration
Fast spherical polyhedron manipulation:
faster geol-ocalization





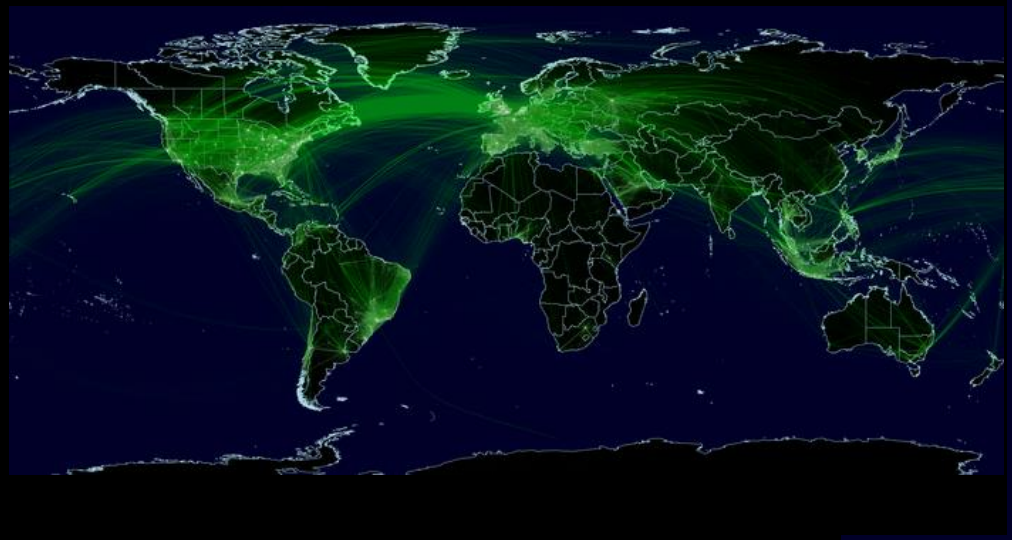


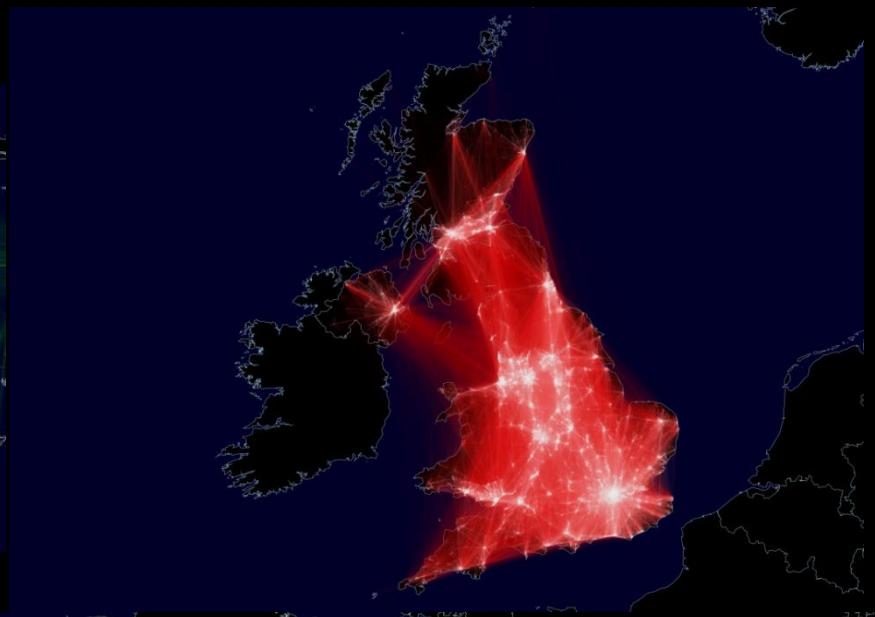


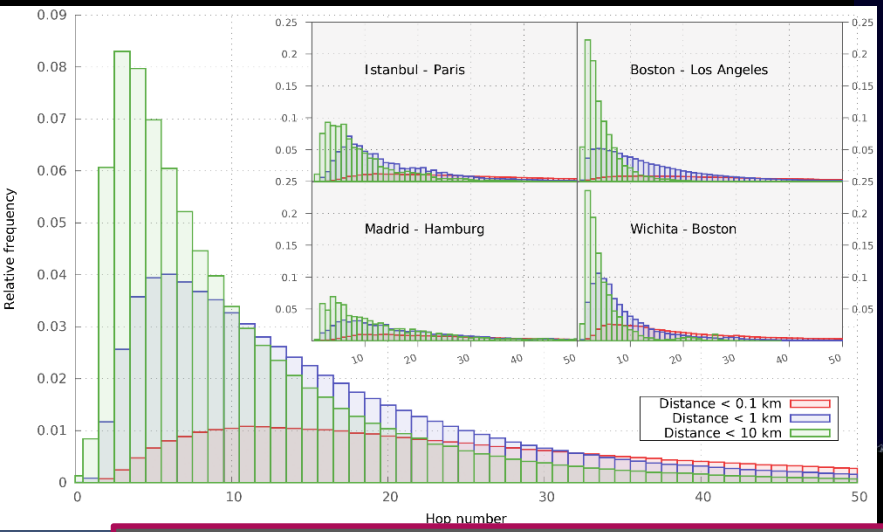


Efficient classification of billions of points into complex geographic regions using hierarchical triangular mesh; D Kondor, L Dobos, I Csabai, A Bodor, G Vattay, T Budavári, AS Szalay; Proc. of the 26th Int. Conf. on Scientific and Statistical Database Management, ACM (2014)

Test Milgram's „6 degree" on Twitter



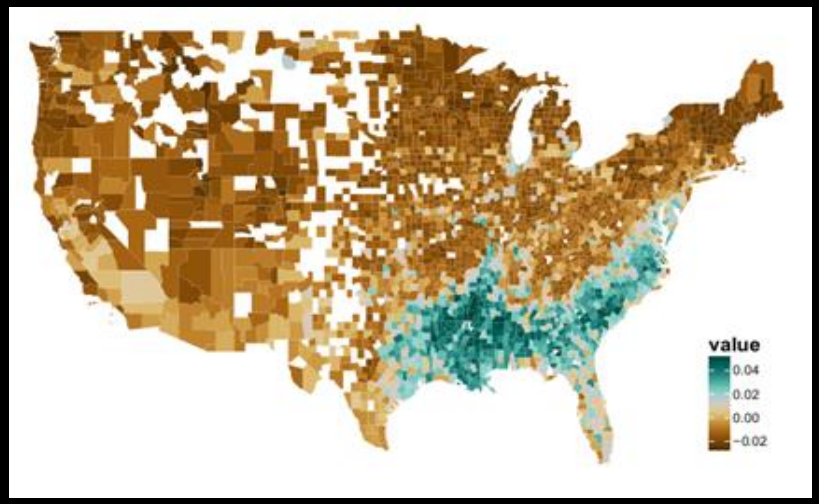


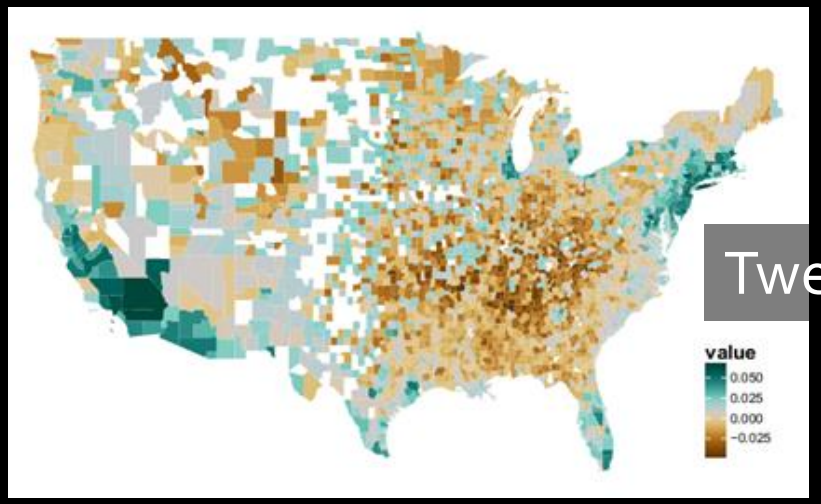


Lost in the City: Revisiting Milgram's Experiment in the Age of Social Networks; J Szüle, D Kondor, L Dobos, I Csabai, G Vattay; PloS one 9 (11), e111973 (2014)

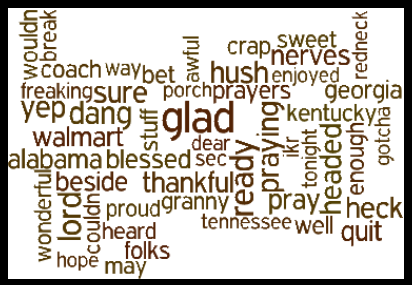
Map of society: TwitterDB

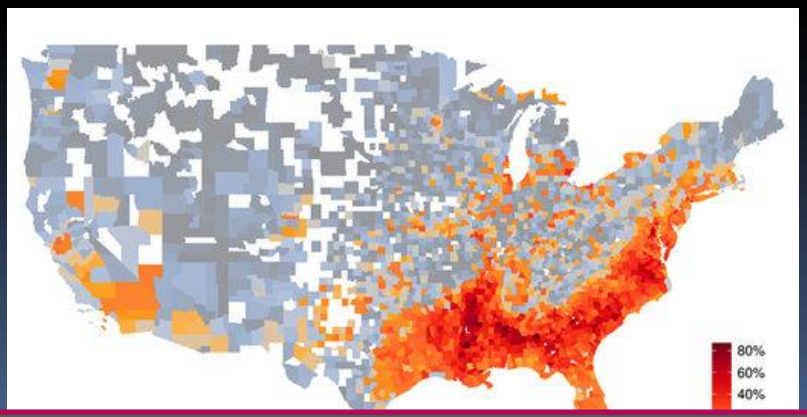
Principal dimensions: race, religion, urbanization

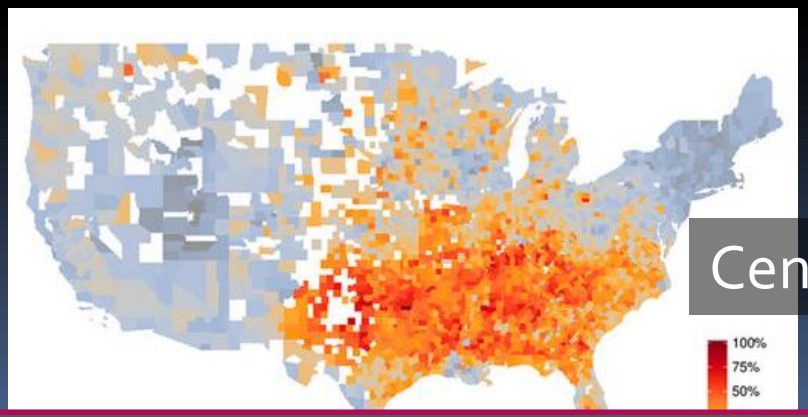




Tweets





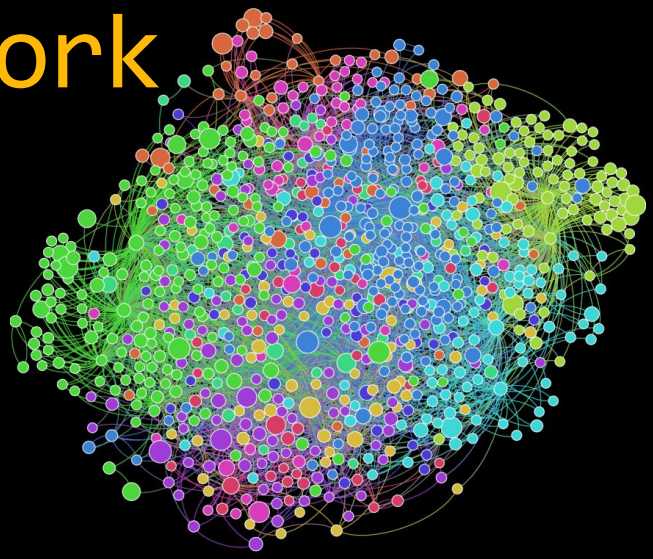


Census

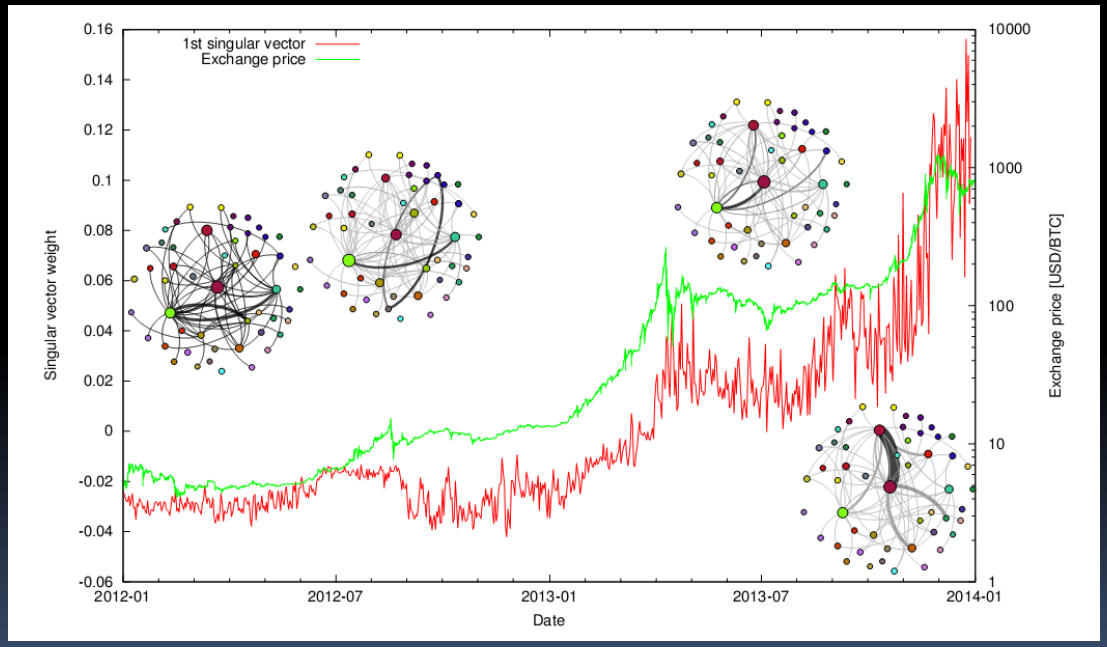
E Bokányi , D Kondor, L Dobos, T Sebők, J Stéger I Csabai, G Vattay Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States , Palgrave Communications 2, 16010 (2016)

Bitcoin financial network

Map of economy



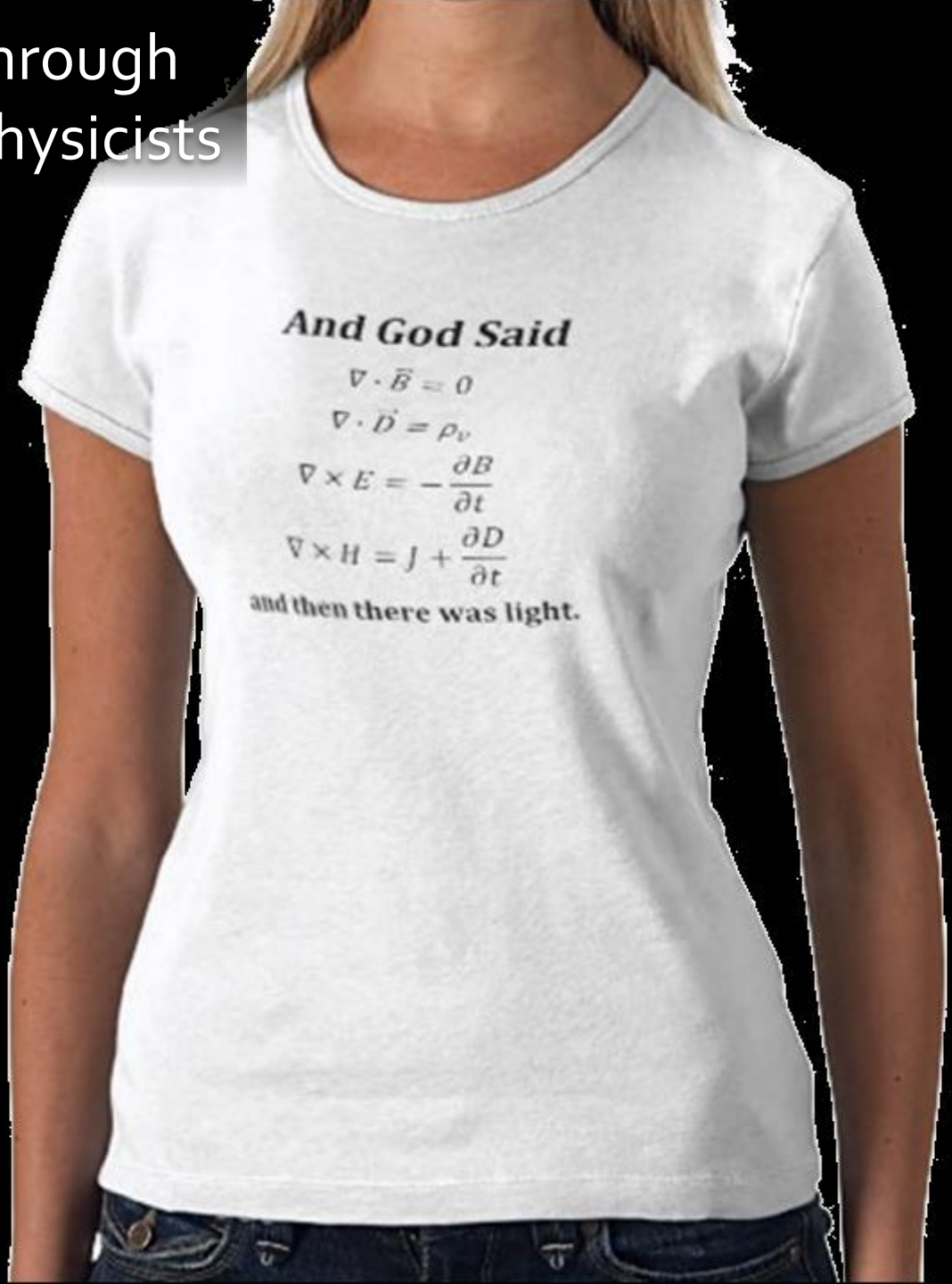
- All (50M) transactions are logged, public
- Dynamic evolving network
- Database
- Dimension reduction (graph non-negative factorization)

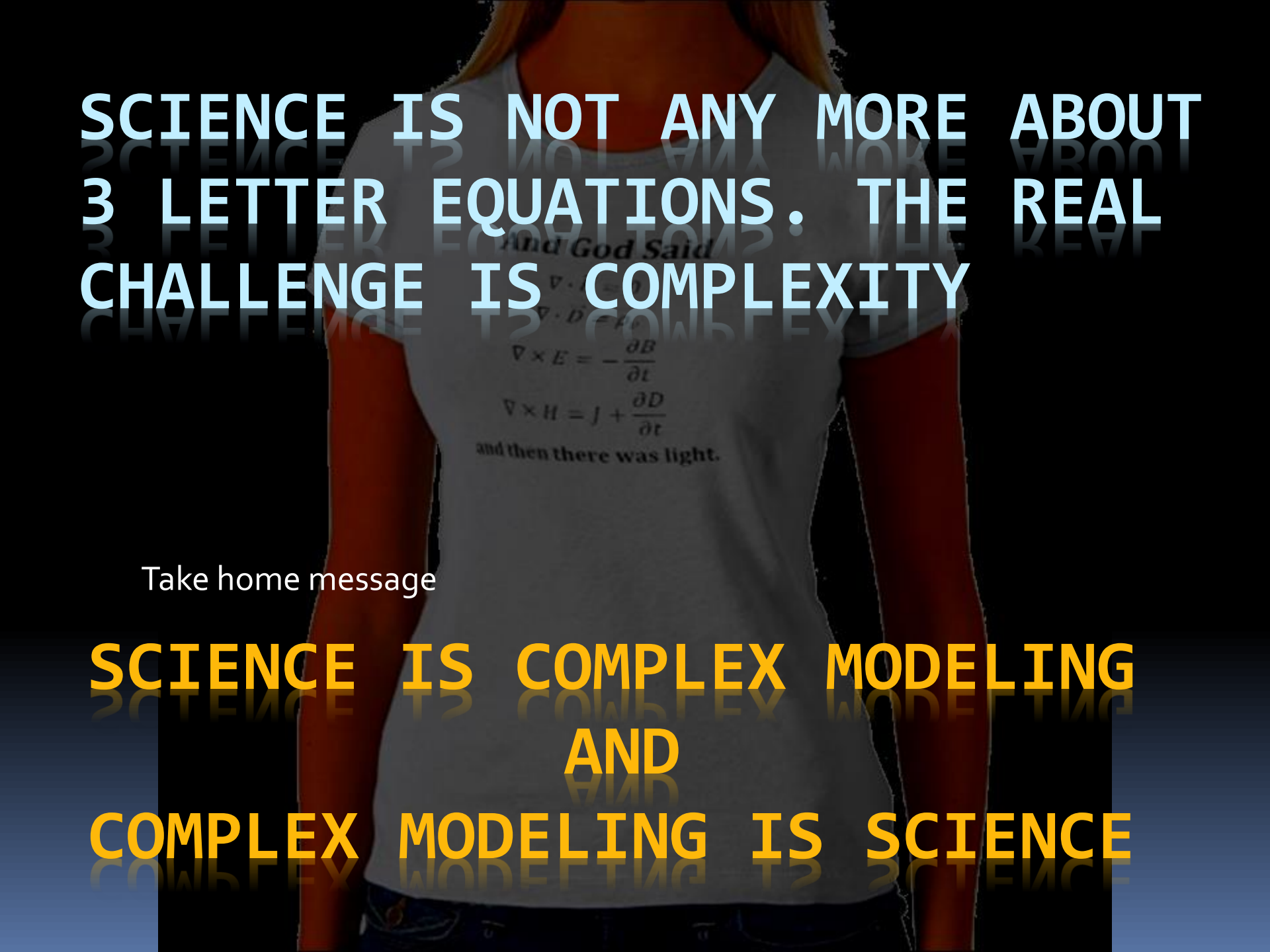


Strong random correlations in networks of heterogeneous agents; I Kondor, I Csabai, G Papp, E Mones, G Czimbalmos, MC Sándor Journal of Economic Interaction and Coordination 9 (2), 203-232 (2014)

Do the rich get richer? An empirical analysis of the BitCoin transaction network; D Kondor, M Pósfai, I Csabai, G Vattay; PloS one 9 (2), e86197 (2014)

The world through the eye of physicists

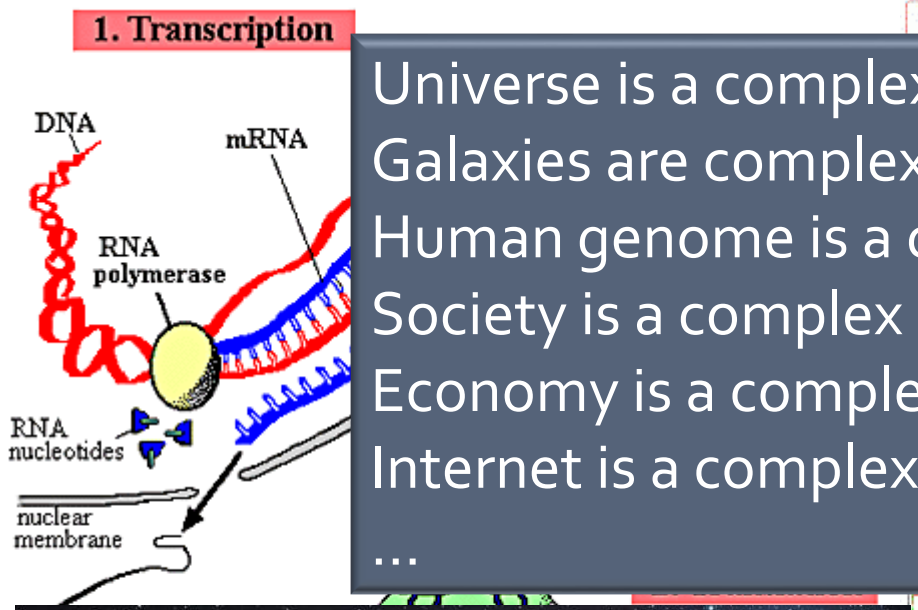




SCIENCE IS NOT ANY MORE ABOUT 3 LETTER EQUATIONS. THE REAL CHALLENGE IS COMPLEXITY

Take home message

SCIENCE IS COMPLEX MODELING AND COMPLEX MODELING IS SCIENCE



Universe is a complex system
Galaxies are complex systems
Human genome is a complex system
Society is a complex system
Economy is a complex system
Internet is a complex system
…

Only complex models can describe complex systems

To build/validate complex models we need "big data" and efficient computational tools (prosthesis): "Datascope " (© Alex Szalay)

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

"If you think of the phases of the golden age of humanity,
once we started to understand the basic laws of physics
and made engineering rules to apply them, we saw exponential
growth of energy and power use"
Last century we saw a similar breakthrough in our understanding
of electricity.
Now, he says, we are ready to tackle another frontier.
"Medicine is the last, because it is the most sophisticated and complex."

/ Dean Kamen, White House Frontiers Conference 2016 October /

Mechanics -> simple machines
Thermodynamics -> steam and internal combustion engines
Electrodynamics -> electricity
+ Quantum mechanics -> microelectronics
? Biology -> end of diseases, longer healthy life, ...

Who will unlock the secrets of the Universe, find the origin of Life? Who will cure cancer? Who will help to understand Everything?

- **A:** Nobody
- **B:** Sorcerers
- **C:** Superintelligent aliens
- **D:** We! Together.

Új tudományos módszertan: Új tudósok kellenek

- akik értik a szaktudományokat
- professzionálisan kezelik a matematikai mellett az informatikai eszköztárat is

Csabai István
ELTE Komplex Rendszerek Fizikája Tanszék
csabai@elte.hu
http://complex.elte.hu/~csabai/