Csabai István
ELTE Komplex Rendszerek Fizikája Tanszék

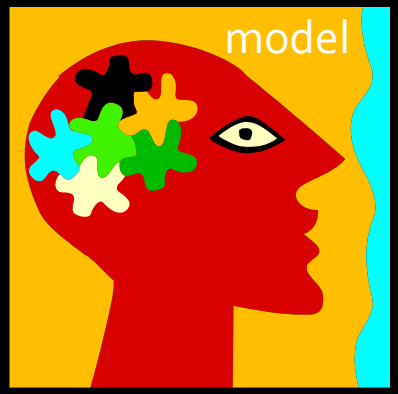
DATA-INTENSIVE GENOMICS
(AND DATA-INTENSIVE BIOLOGY AND DATA-INTENSIVE SCIENCES)

Neuroinformatika Doktori Kurzus– Budapest 2016.09.21

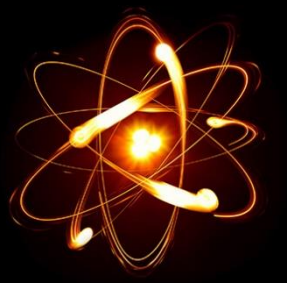
Evolution of the data intensive scientist: early times

observation

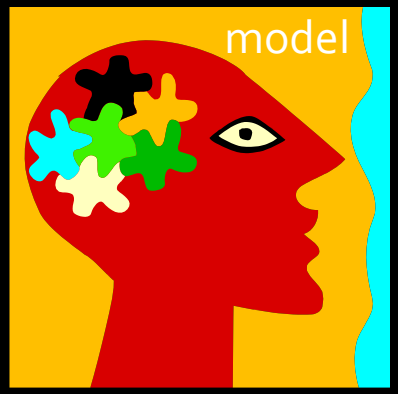
Evolution of the data intensive scientist: early times



observation

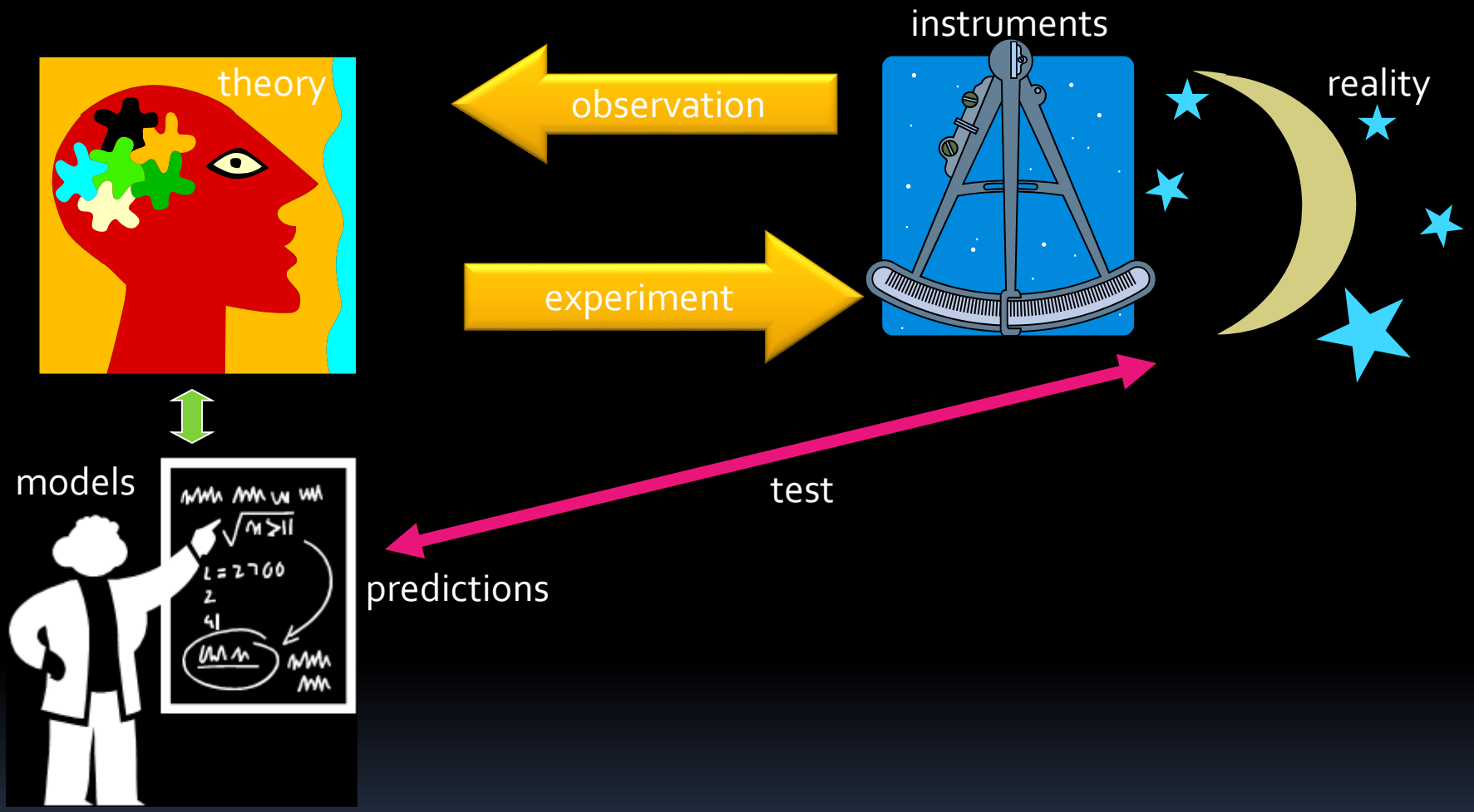


Evolution of the data intensive scientist: early times

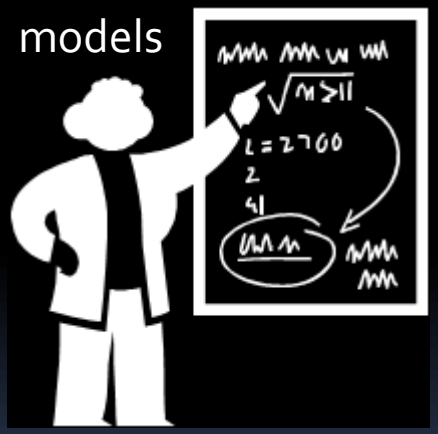


observation

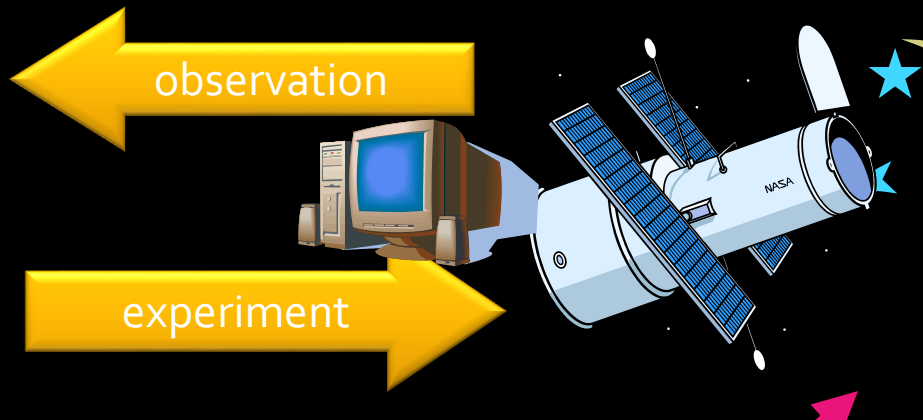
Evolution of the data intensive scientist: past



Evolution of the data intensive scientist: present



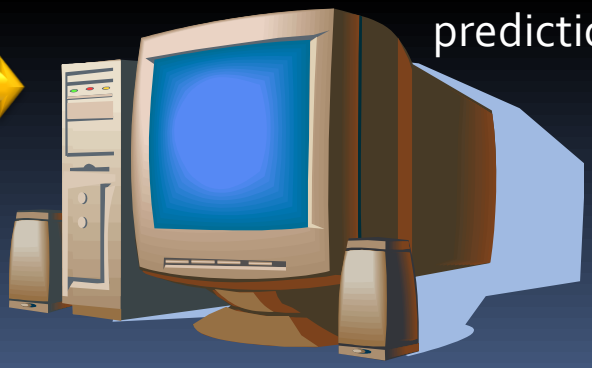
instruments



reality

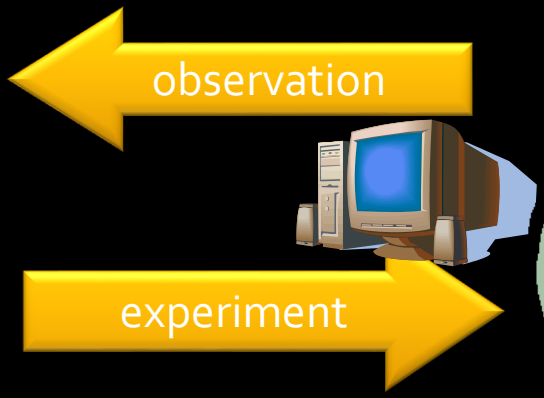
test

virtual reality

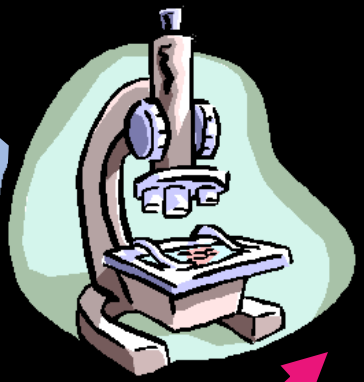


predictions

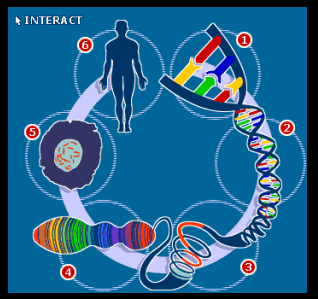
Other disciplines are similar: whole genomes, satellite maps, sensor networks, etc.

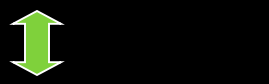


instruments

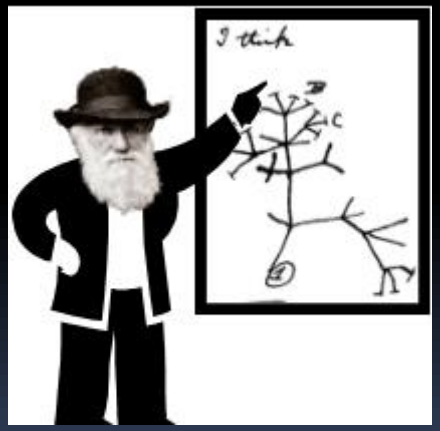


reality





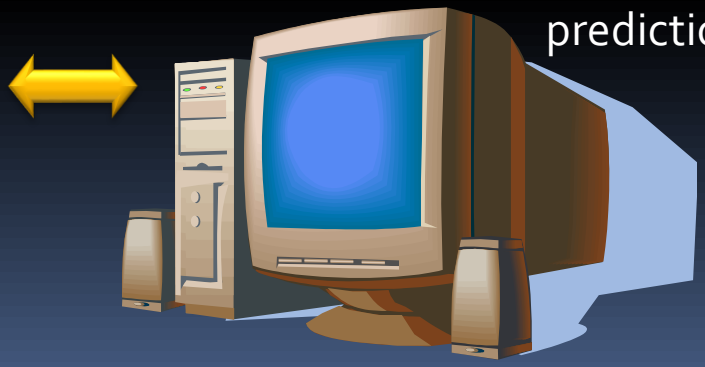
models



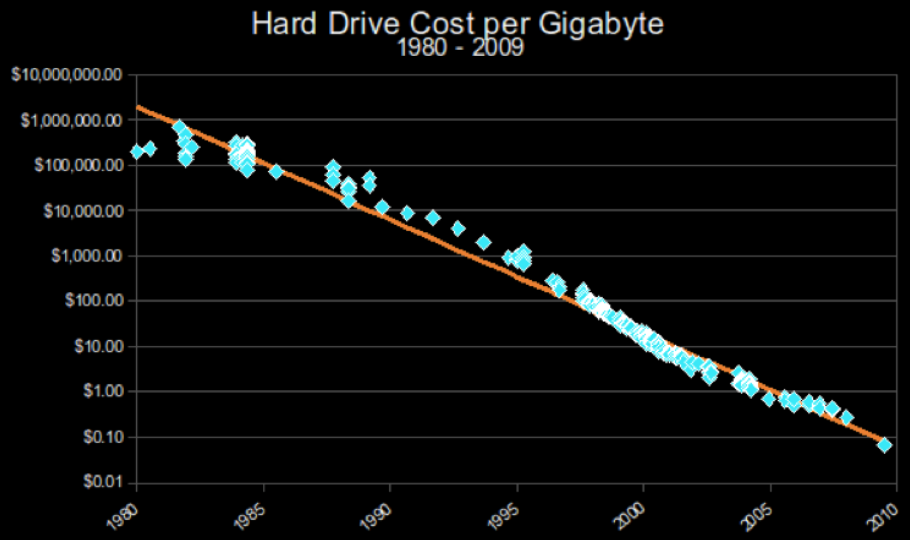
test

virtual reality

predictions

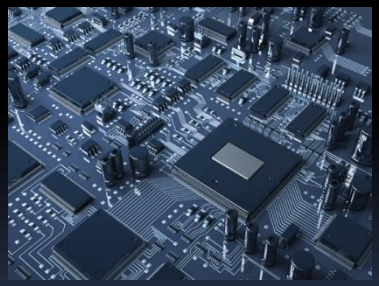


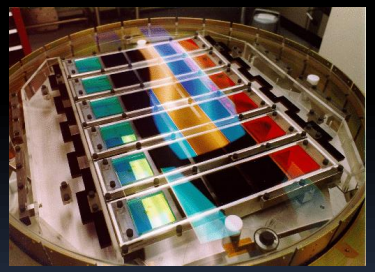
Exponentially cheaper devices



Moore's law

Exponential growth





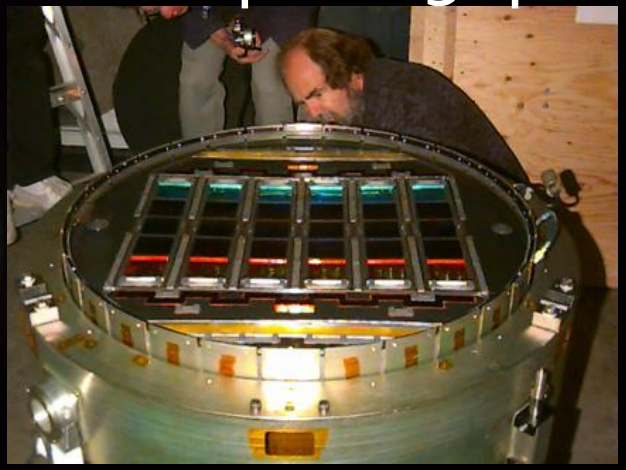
Elektronics → Sensors → Data

Prototype of data-intensive science project:

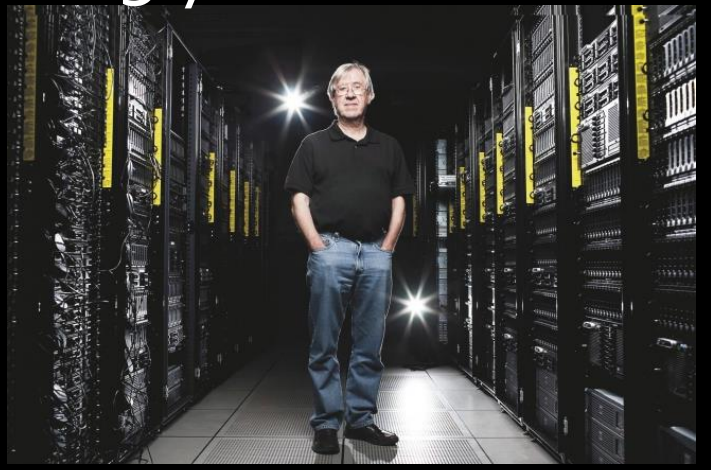
SLOAN DIGITAL SKY SURVEY (SDSS): THE 3D MAP OF THE UNIVERSE 1995-2005…

2.5m

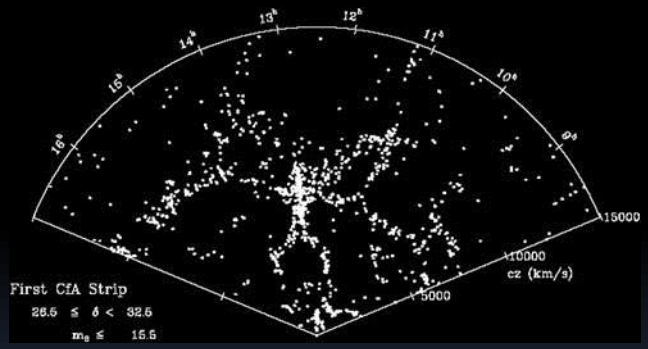
120Mp – 2.5Tp



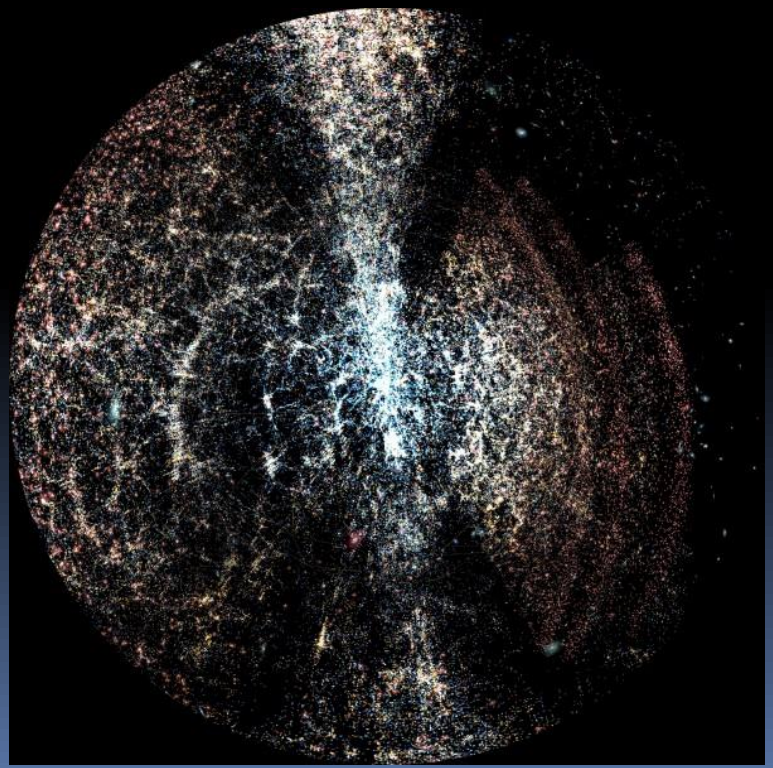
5 years:10TB



CfA 1989: 1100 galaxies



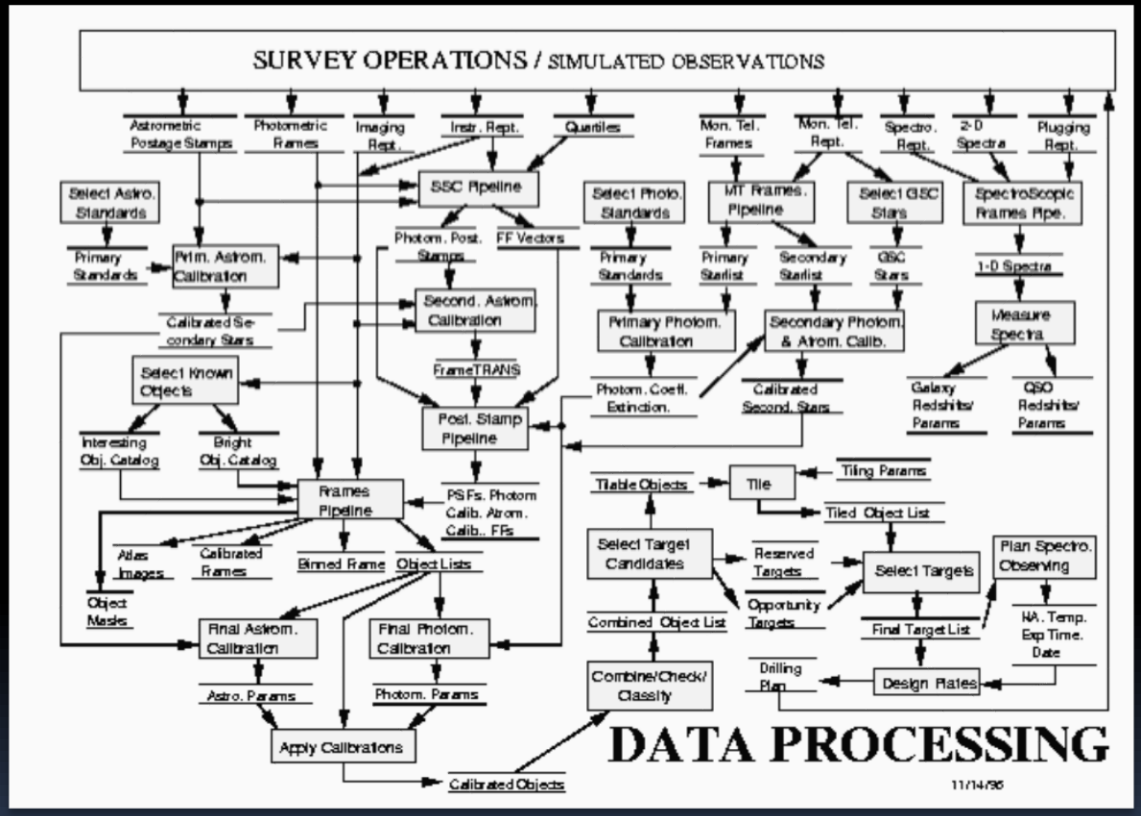
SDSS 2005: 1M galaxies



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")*



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)

New questions: astronomical data sets

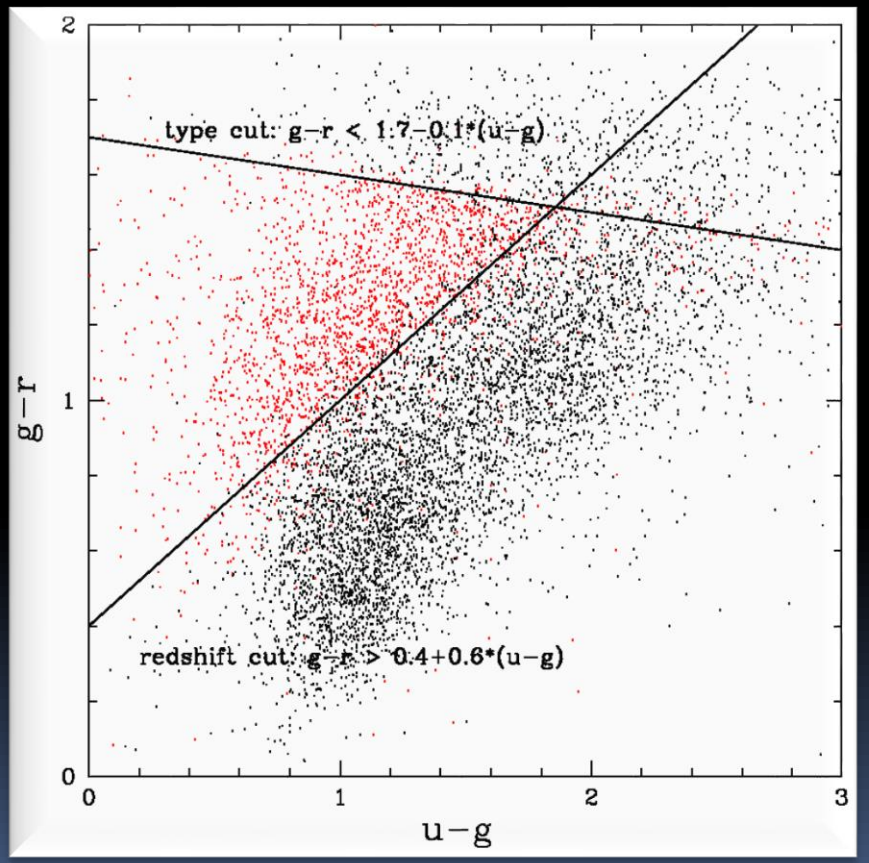
Star/galaxy separation
Quasar target selection

→ "cuts"

→ **Multi-dimensional polyhedra**

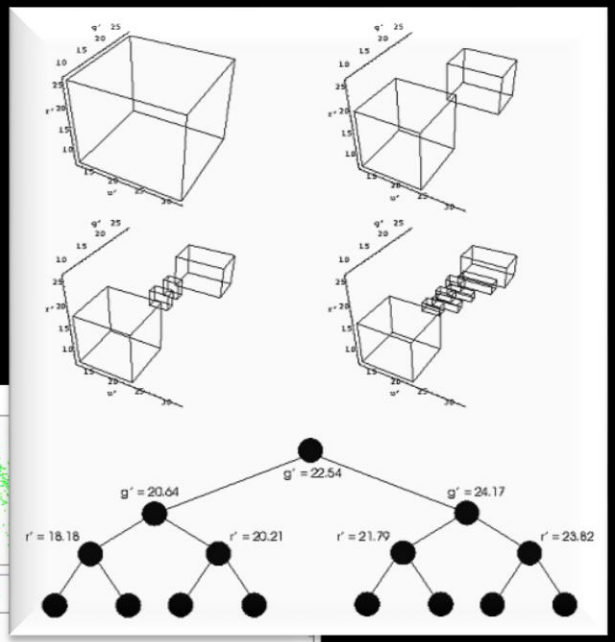
Skyserver: several million queries per year

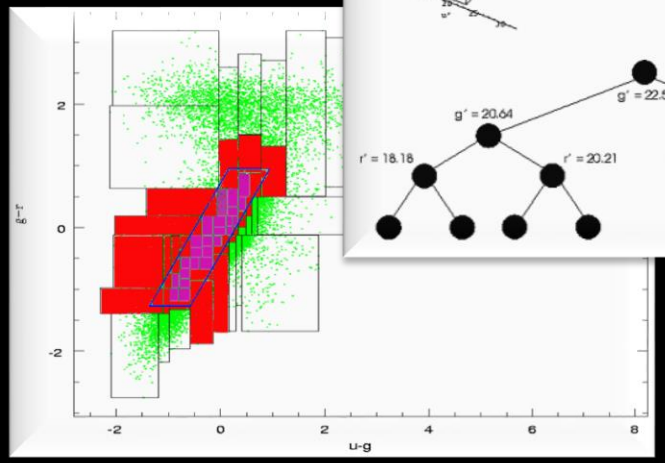
```
petroMag_i > 17.5 and (petroMag_r > 15.5 or petroR50_r > 2)
and (petroMag_r > 0 and g > 0 and r > 0 and i > 0) and (
(petroMag_r-extinction_r) < 19.2 and (petroMag_r -
extinction_r < (13.1 + (7/3) * (dered_g - dered_r) + 4 * (dered_r
- dered_i) - 4 * 0.18) ) and ( (dered_r - dered_i - (dered_g -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

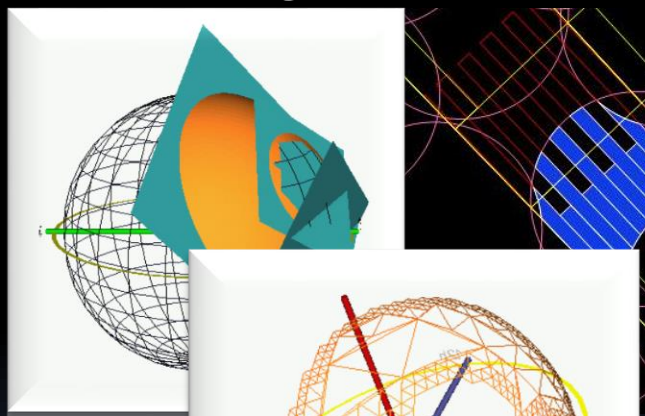


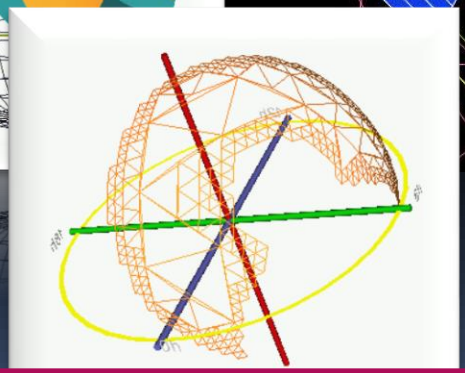
New skills: Indexing, databases

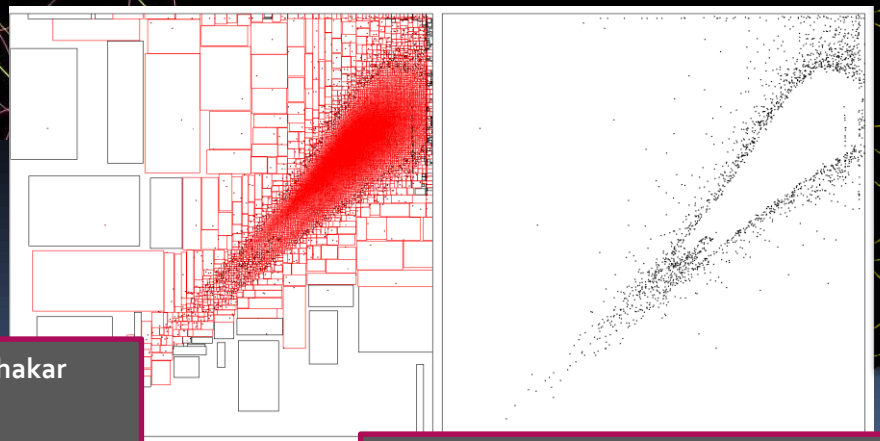
- SDSS data "read through"~1 day
- Astronomers should learn: Database programming, computer geometry, search trees, …
- Multidimensional- and spherical indexing

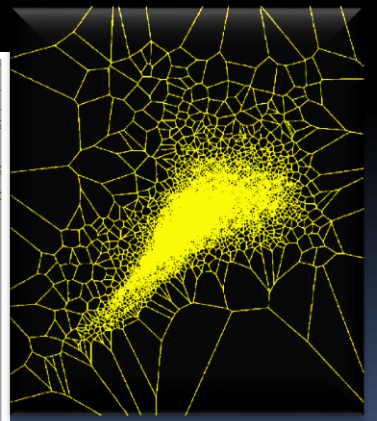












AS Szalay, J Gray, G Fekete,P Kunszt, P Kukol, A Thakar MSR -TR 123 (2005)

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)

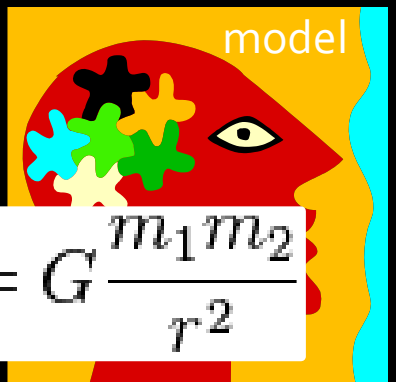
L Dobos, I Csabai, AS Szalay, T Budavári, N Li
Proceedings of the 25th International Conference on Scientific and Statistical Database Management, ACM, (2013)

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)

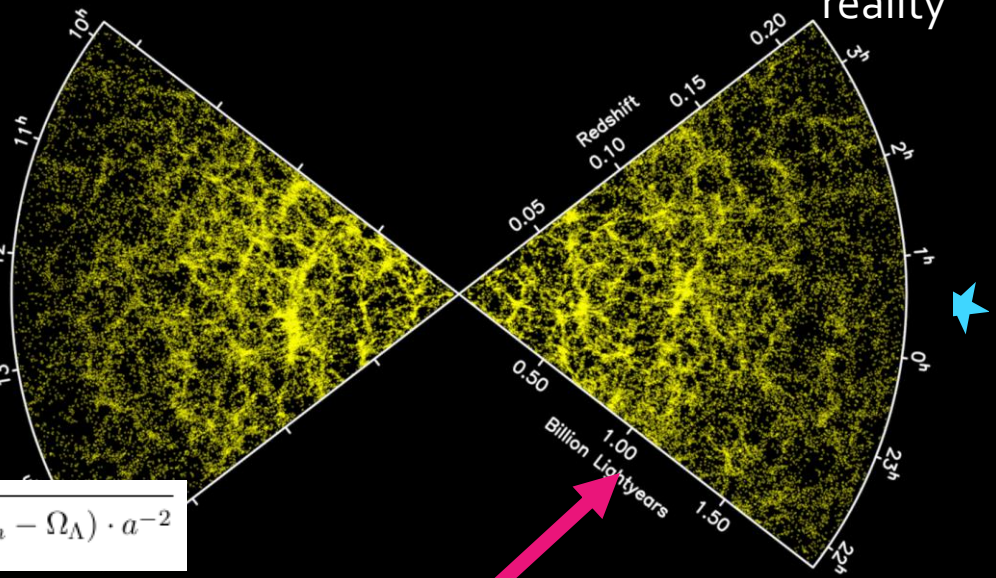
Models: N-body simulations



ob

expe

reality

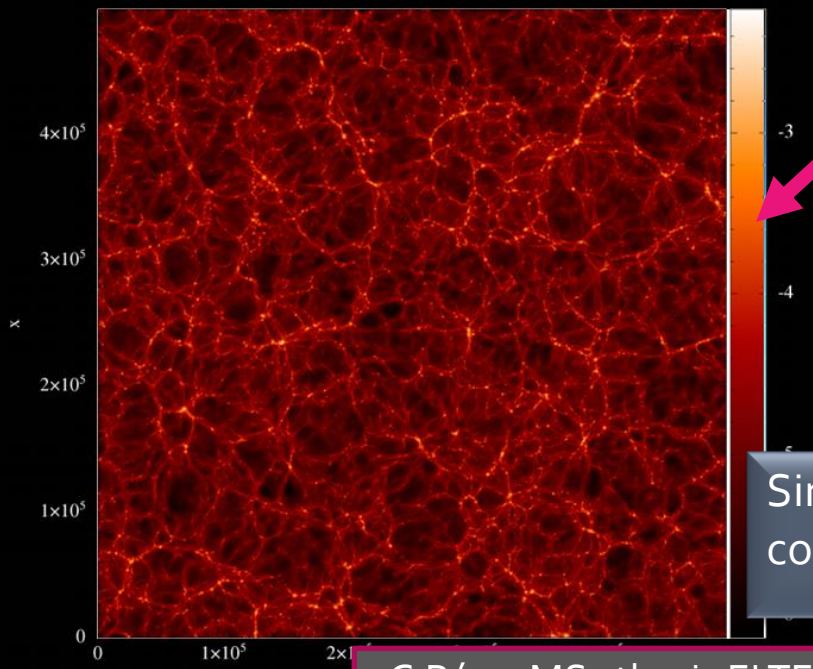


$$F = G\frac{m_1 m_2}{r^2}$$

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

model

test

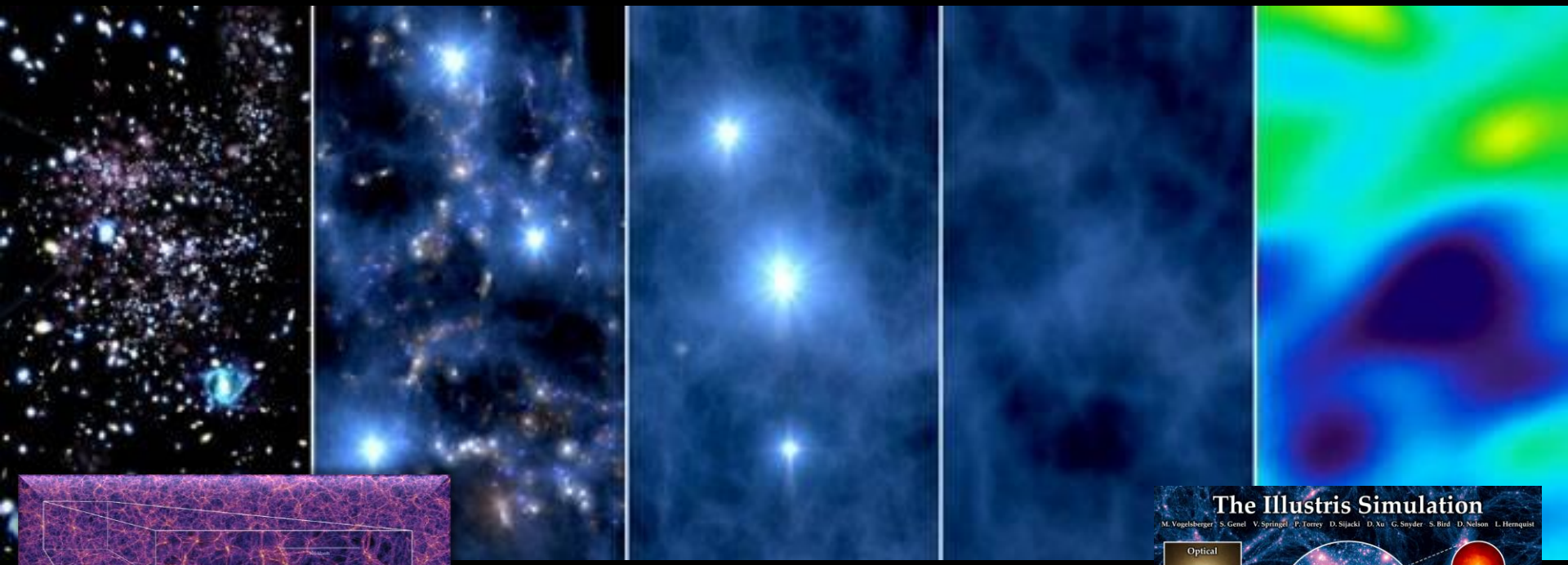


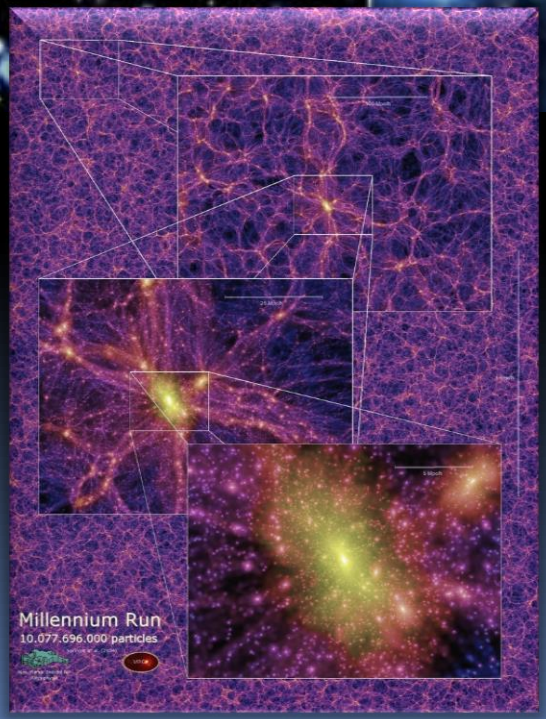
Virtual reality

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

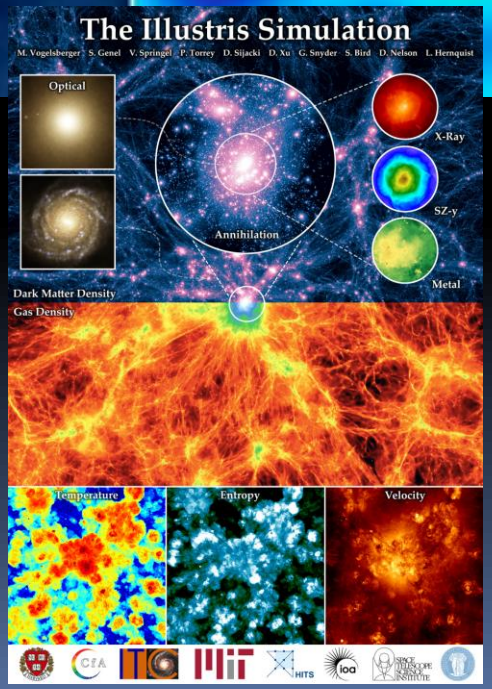
G Rácz, MSc thesis ELTE (2015)

Models: The Millennium Run



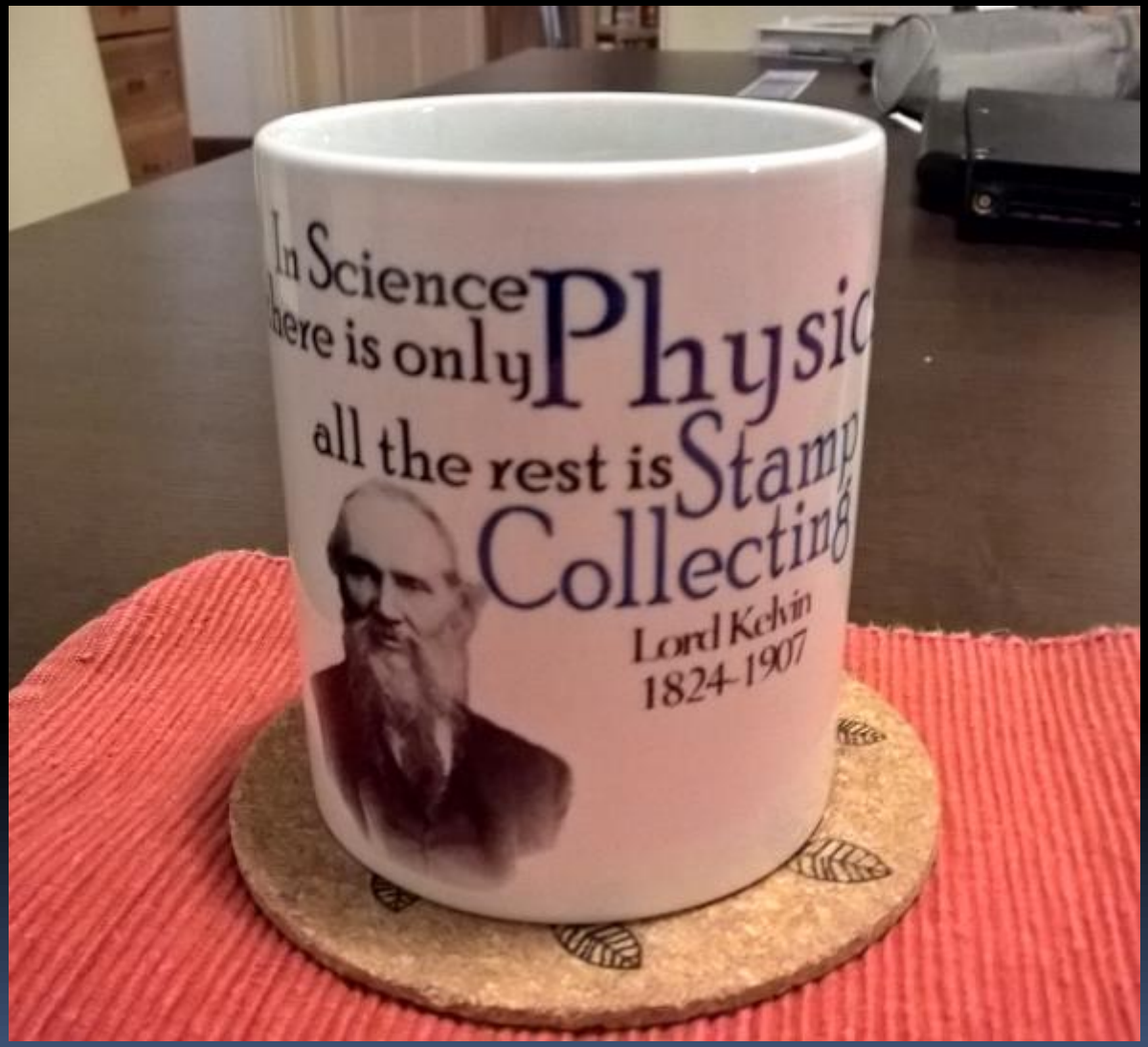


Simulation of the Universe
The Virgo Consortium
Dozen research groups, 10 billion "dark matter particles", 2 billion light years
1500 processors, 30 days, 25TB output



Real Universe – Virtual Universe

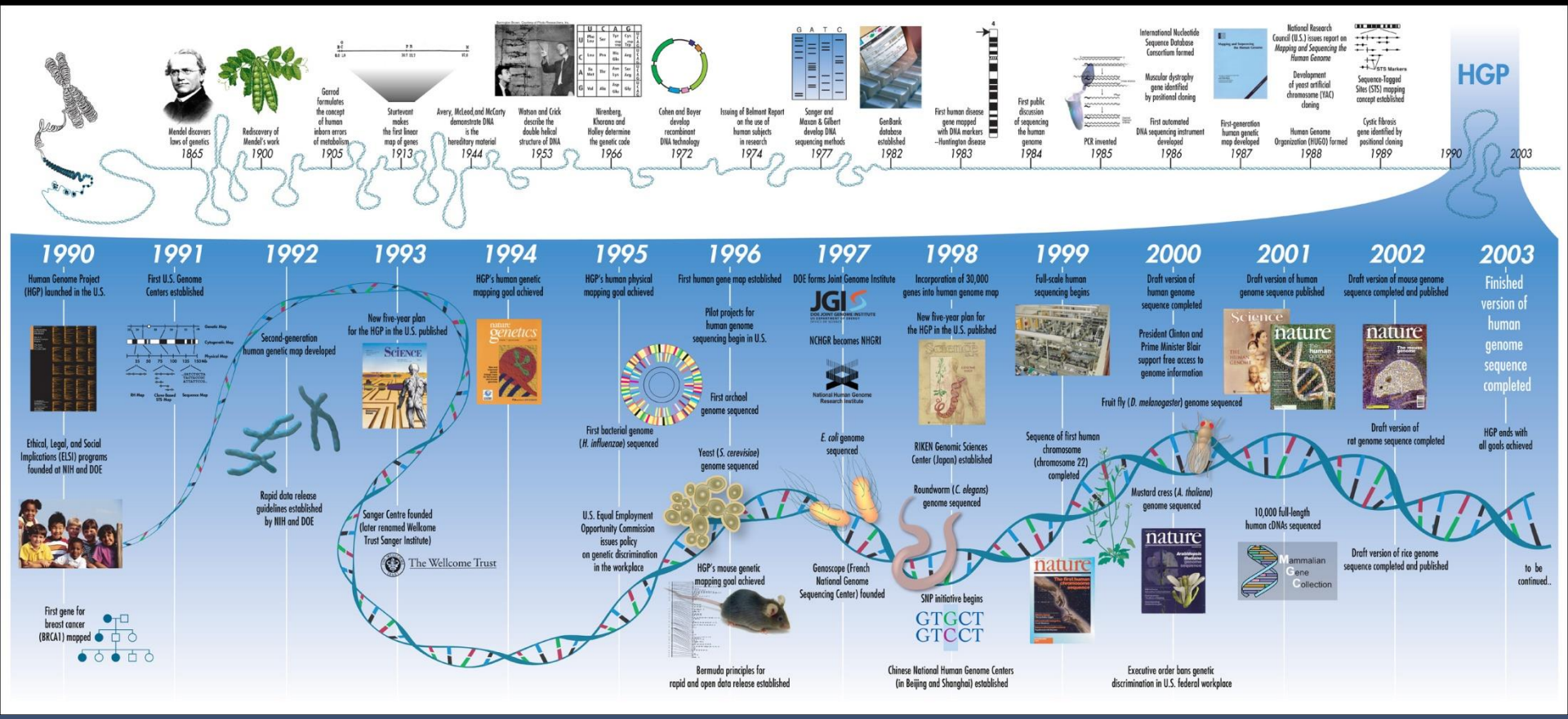
Is this real or simulated?



? Ernest Rutherford (1871—1937)

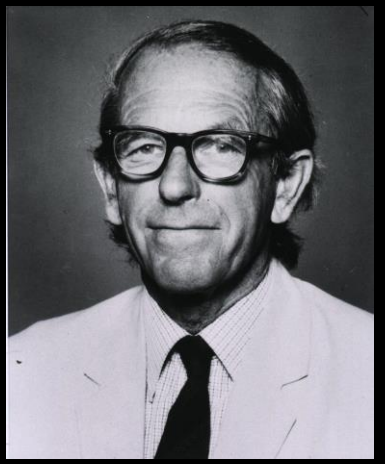
NEXT GENERATION SEQUENCING

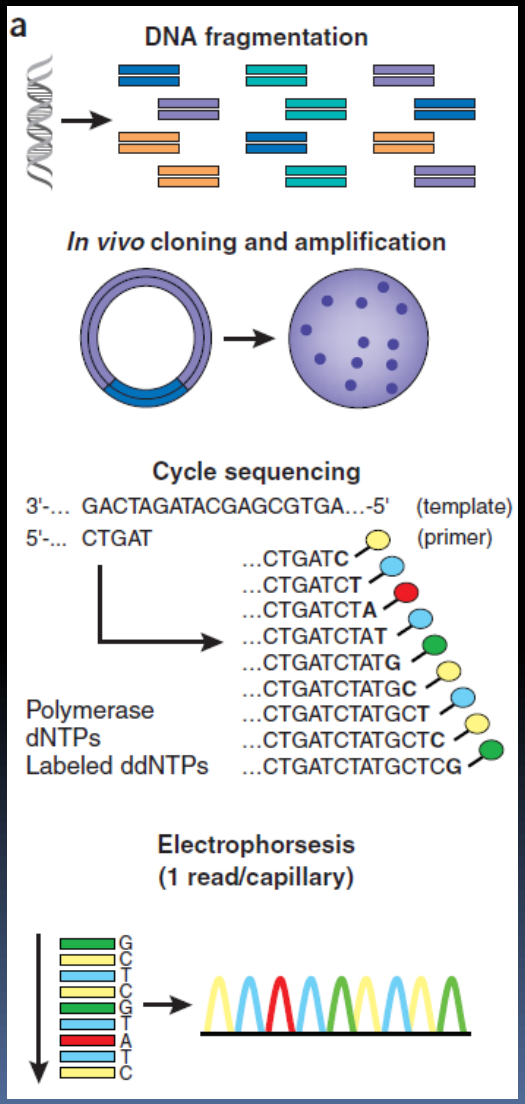
Map of the genome

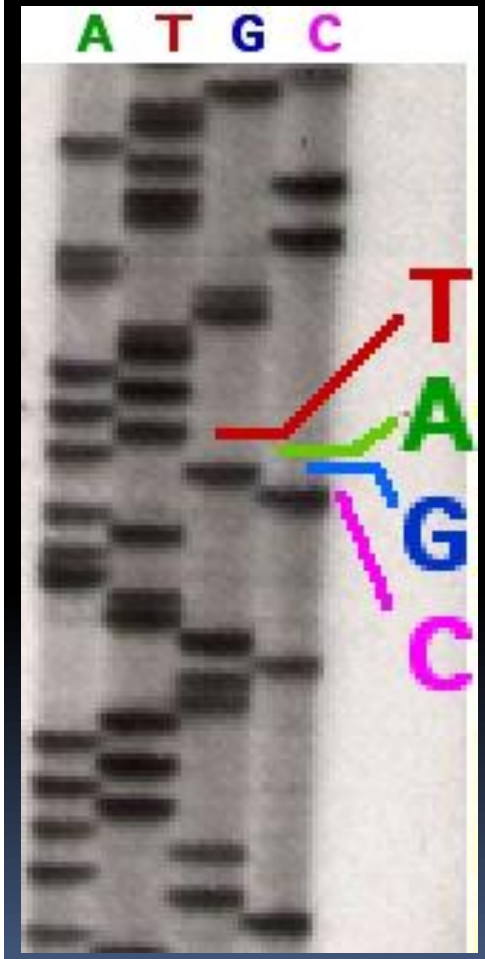


High througput sequencing history: Sanger

1977 Frederick_Sanger





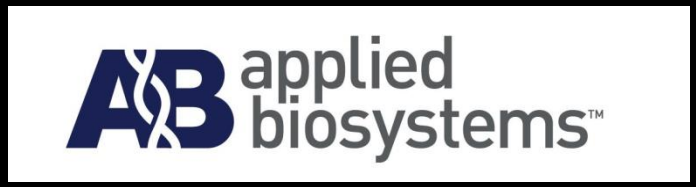


- DNA is fragmented
- Cloned to a plasmid vector
- Cyclic sequencing reaction
- Separation by electrophoresis
- Readout with fluorescent tags

Main technologies

„Past”:

http://www.youtube.com/watch?v=l99aKKHcxC4



Solid

http://www.youtube.com/watch?v=nlvyF8bFDwM

„Present”:

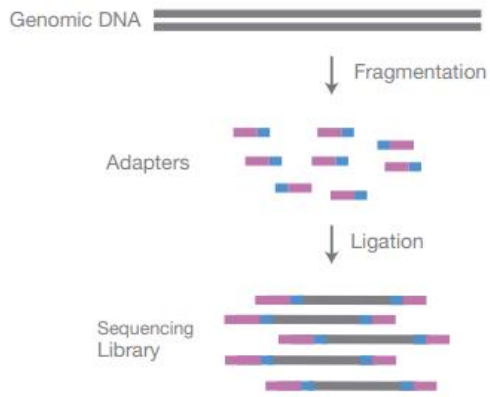
http://www.youtube.com/watch?v=yVf2295JqUg

„Future”:

http://www.nanoporetech.com/news/movies#movie-21-gridion-part-1

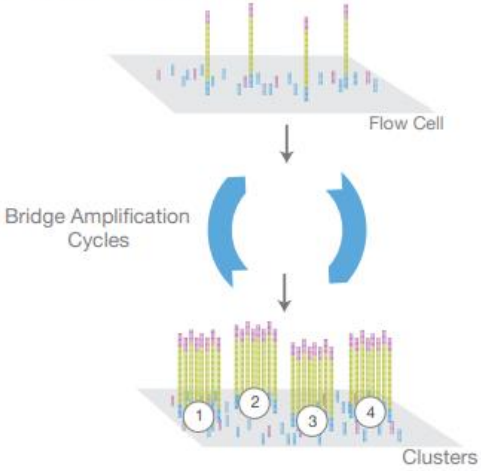
NGS overview (Illumina)

A. Library Preparation



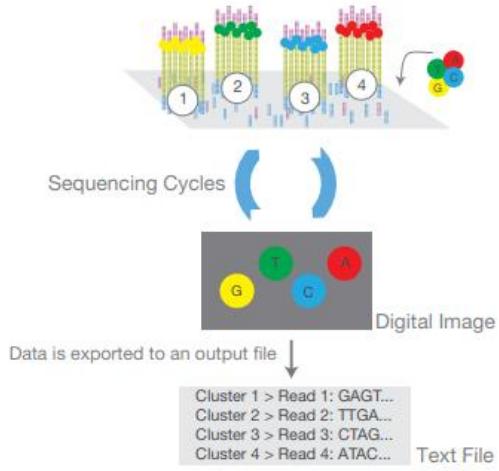
NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

B. Cluster Amplification



Library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

C. Sequencing



Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

D. Alignment & Data Anaylsis

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Barcoding/multiplexing (Illumina)

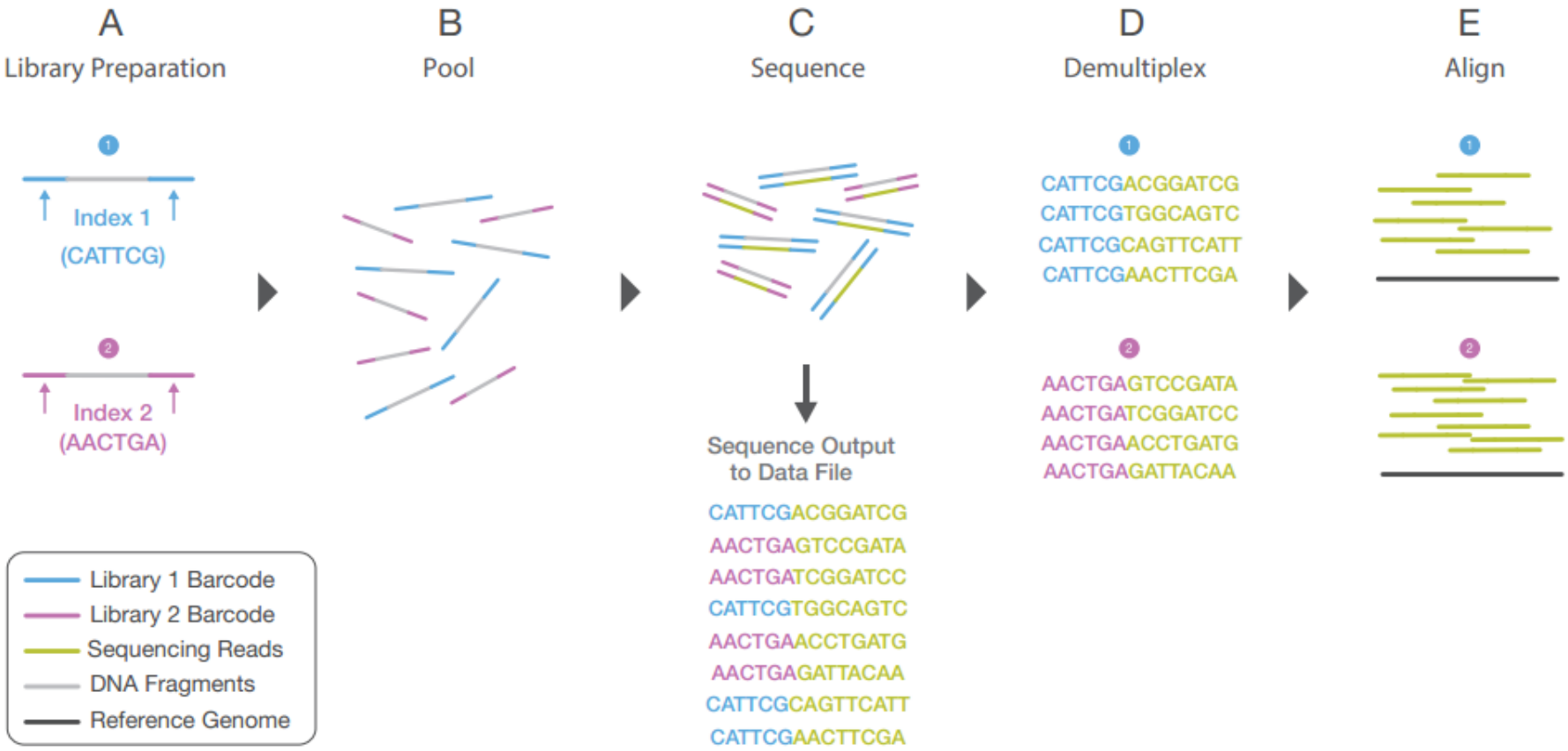


Figure 5: Library Multiplexing Overview.

a. Two distinct libraries are attached to unique index sequences. Index sequences are attached during library preparation.

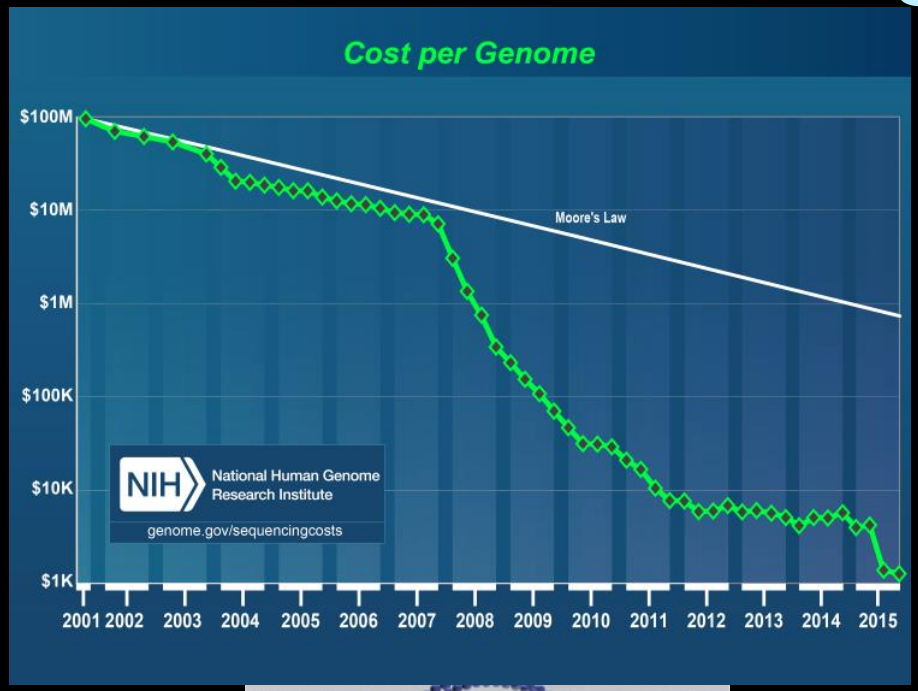
b. Libraries are pooled together and loaded into the same flow cell lane.

c. Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file.

d. A demultiplexing algorithm sorts the reads into different files according to their indexes.

e. Each set of reads is aligned to the appropriate reference sequence.

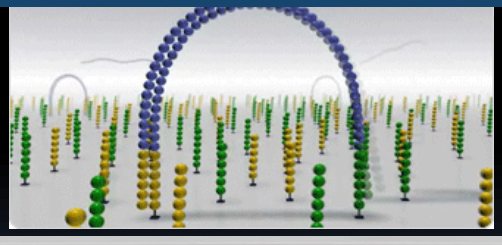
Moore's law in genetics

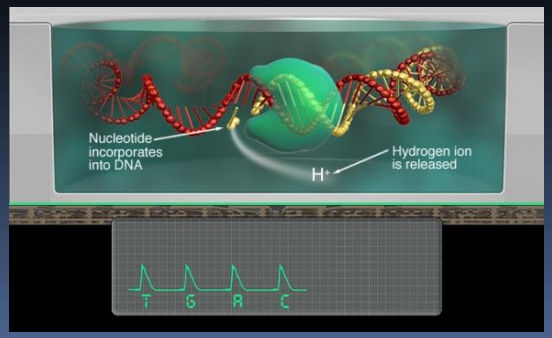


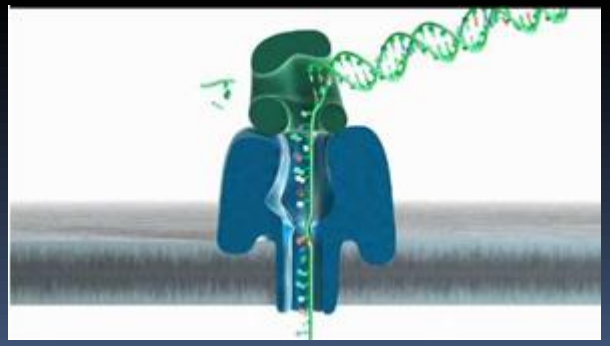
Human genome sequencing
1990-2003: 13yrs /2.7 Bn USD
2016: ~days/1000 USD
2020: ?????

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

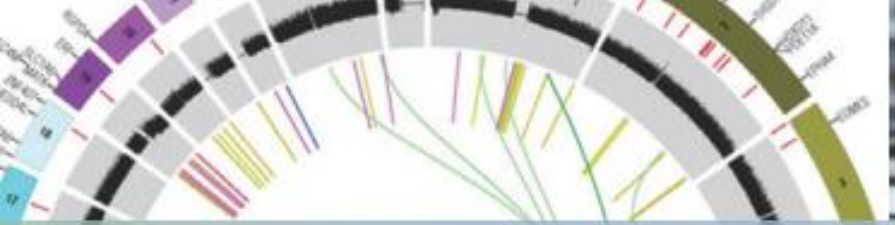
Oxford Nanopore
100Mb,$900







S.Spisak, K.Lawrenson, Y.Fu, II.Csabai, ... M. Freedmann. Nature Medicine doi:10.1038/nm.3975 (2015)



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.

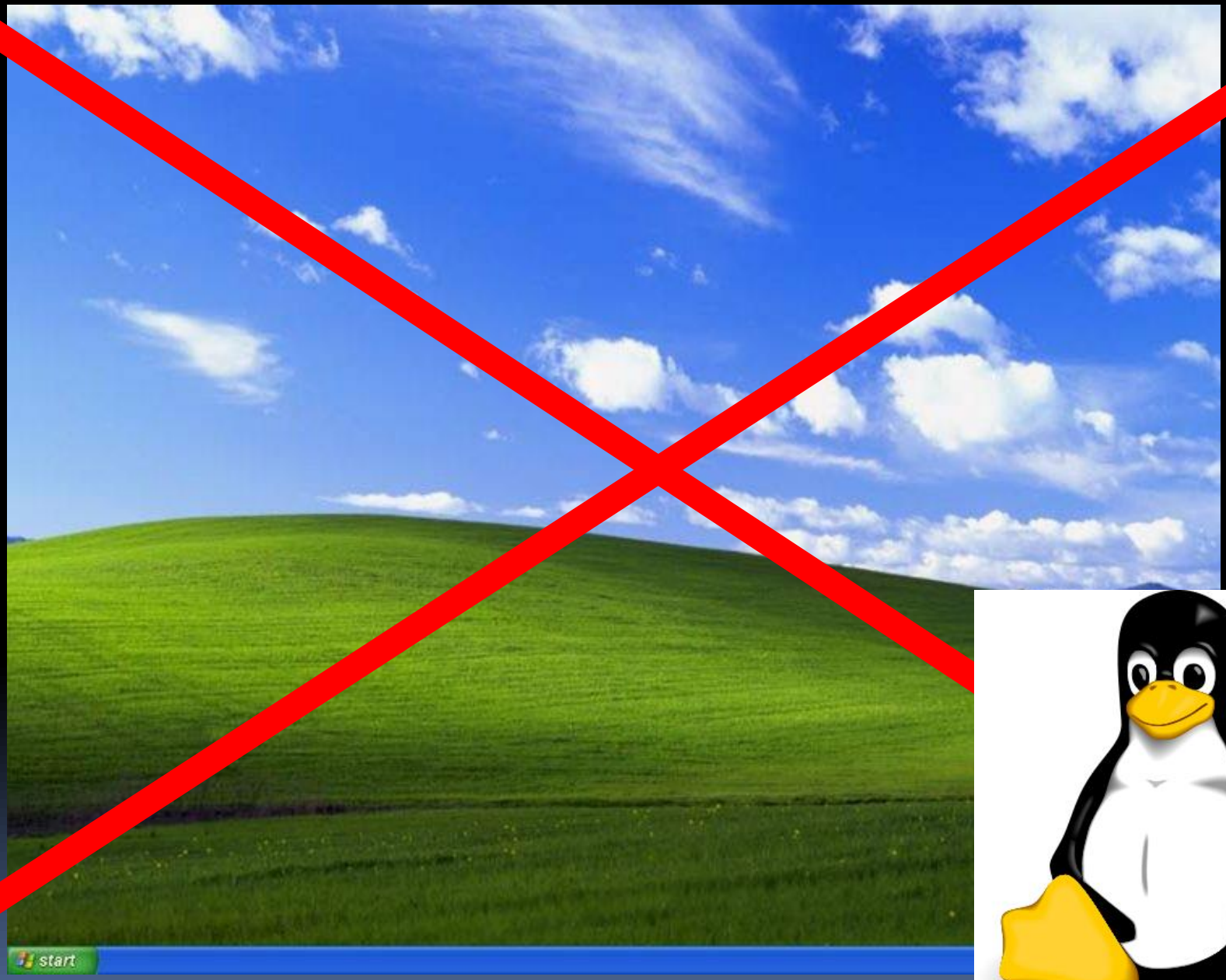
TCGA data are available in two data repositories: the TCGA Data Portal and the Cancer Genomics Hub. All data can be accessed directly from the TCGA Data Portal regardless of which repository houses the data file.

S.Spisak, K.Lawrenson,Y.Fu,II.Csabai, … M. Freedmann. Nature Medicine doi:10.1038/nm.3975 (2015)

NGS machnies

Massive amount of sequence data

http://www.flickr.com/photos/ohm17/162622755/


start

http://www.flickr.com/photos/esquimo_2ooo/5241744434/

Computing and storage challenge: clouds, virtualization, …

Crossbow

Genotyping from short reads using cloud computing

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Crossbow is a scalable software pipeline for whole genome resequencing analysis. It combines Bowtie, an ultrafast and memory efficient short read aligner, and SoapSNP, and an accurate genotyper. These tools are combined in an automatic, parallel pipeline that runs in the cloud (Elastic MapReduce in this case) on a local Hadoop cluster, or on a single computer, exploiting multiple computers and CPUs wherever possible. The pipeline can analyze over 35x coverage of a human genome in one day on a 10-node local cluster, or in 3 hours for about $85 using a 40-node, 320-core cluster rented from Amazon Web Services.

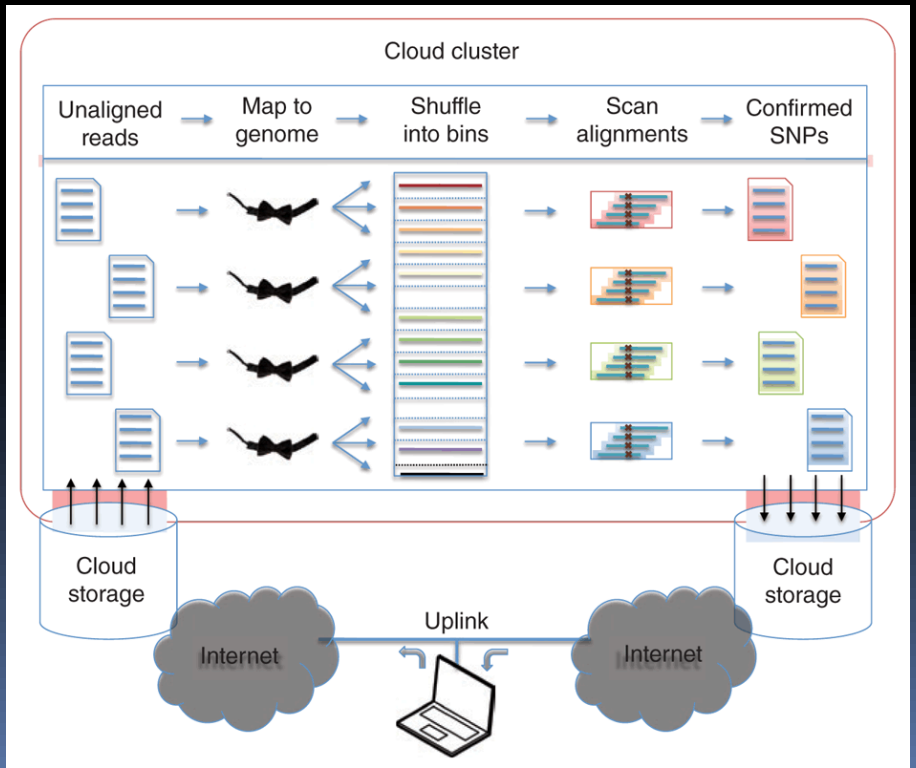
OSI certified

Version 1.1.0 - October 12, 2010

- Added `--discard-ref-bin` and `--discard-all` options, which can be helpful to reduce Crossbow running time when a run's chief purpose is to test whether it runs all the way through.
- Fixed a bug in soapsnp that caused a segmentation fault in the last partition of a chromosome when

Site Map

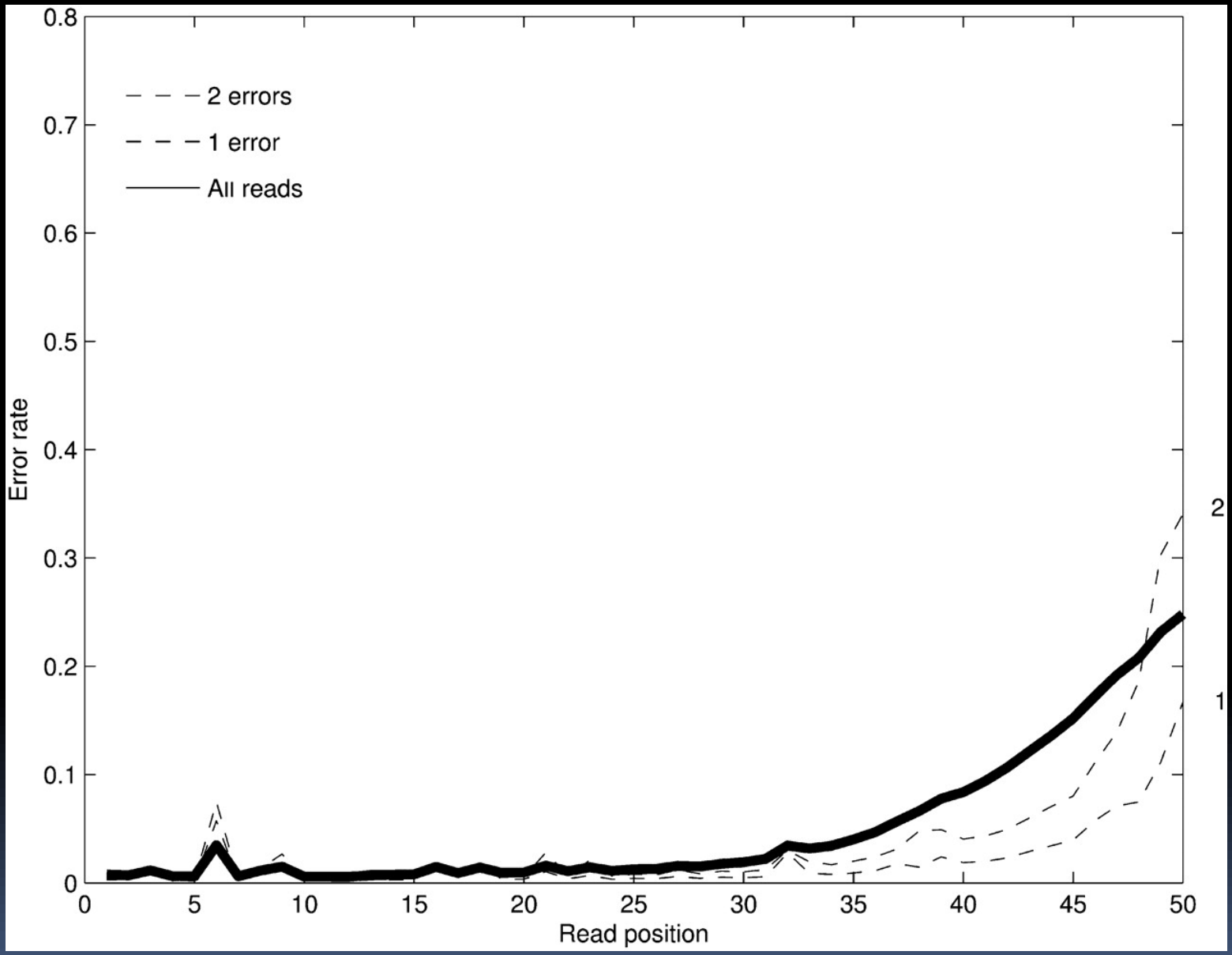
- Home
- Web interface
- News archive



Data formats: FASTQ

@IL31_4368:1:1:996:8507/2
TCCCTTACCCCCAAGCTCCATACCCTCCTAATGCCCACACCTCTTACCTTAGGA
+
FFCEFFFEEFFFFFFFEFFEFFFEFCFC<EEFEFFFCEFF<;EEFF=FEE?FCE
@IL31_4368:1:1:996:21421/2
CAAAAACTTTCACTTTACCTGCCGGGTTTCCCAGTTTACATTCCACTGTTTGAC
+
>DBDDB,B9BAA4AAB7BB?7BBB=91;+*@;5<87+*=/*@@?9=73=.7)7*
@IL31_4368:1:1:997:10572/2
GATCTTCTGTGACTGGAAGAAAATGTGTTACATATTACATTTCTGTCCCCATTG
+
E?=EECE<EEEE98EEEEAEEBD??BE@AEAB><EEABCEEDEC<<EBDA=DEE
@IL31_4368:1:1:997:15684/2
CAGCCTCAGATTCAGCATTCTCAAATTCAGCTGCGGCTGAAACAGCAGCAGGAC
+
EEEEDEEE9EAEEDEEEEEEEEEECEEAAEEDEE<CD=D=*BCAC?;CB,<D@,
@IL31_4368:1:1:997:15249/2
AATGTTCTGAAACCTCTGAGAAAGCAAATATTTATTTTAATGAAAAATCCTTAT
+
EDEEC;EEE;EEE?EECE;7AEEEEEE07EECEA;D6D>+EE4E7EEE4;E=EA
@IL31_4368:1:1:997:6273/2
ACATTTACCAAGACCAAAGGAAACTTACCTTGCAAGAATTAGACAGTTCATTTG
+
EEAAFFFEEFEFCFAFFAFCCFFEFEF>EFFFFB?ABA@ECEE=<F@DE@DDF;
@IL31_4368:1:1:997:1657/2
CCCACCTCTCTCAATGTTTTCCATATGGCAGGGACTCAGCACAGGTGGATTAAT
(...)

http://en.wikipedia.org/wiki/FASTQ_format


2 errors
1 error
All reads
Error rate
Read position
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
35
40
45
50
2
1

Mapping the short reads onto a reference genome

The typical task:

- 1 billion short reads, 30-200 nt long (R)
- Target genome: 1M-15G nt (G)
- Few to few thousand samples (S)

The naive solution scales with

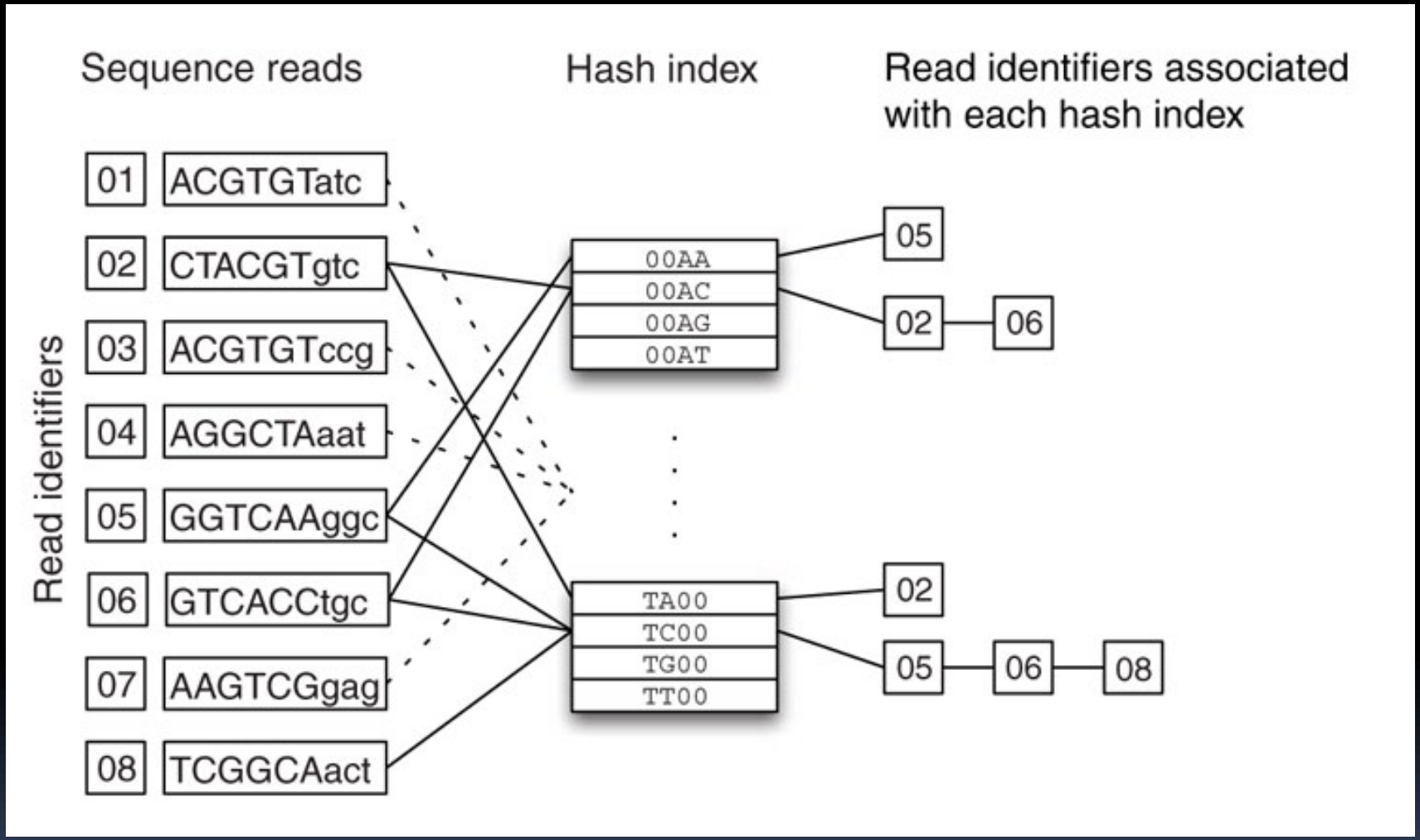
$S * R * G \sim 30*10^9*3*10^9*10 \sim 10^{21}$

This is too many computer operations even for the modern machines! And the gap is growing!

"Running these accurate alignment algorithms as a full search of all possible places where the sequence may map is computationally infeasible."

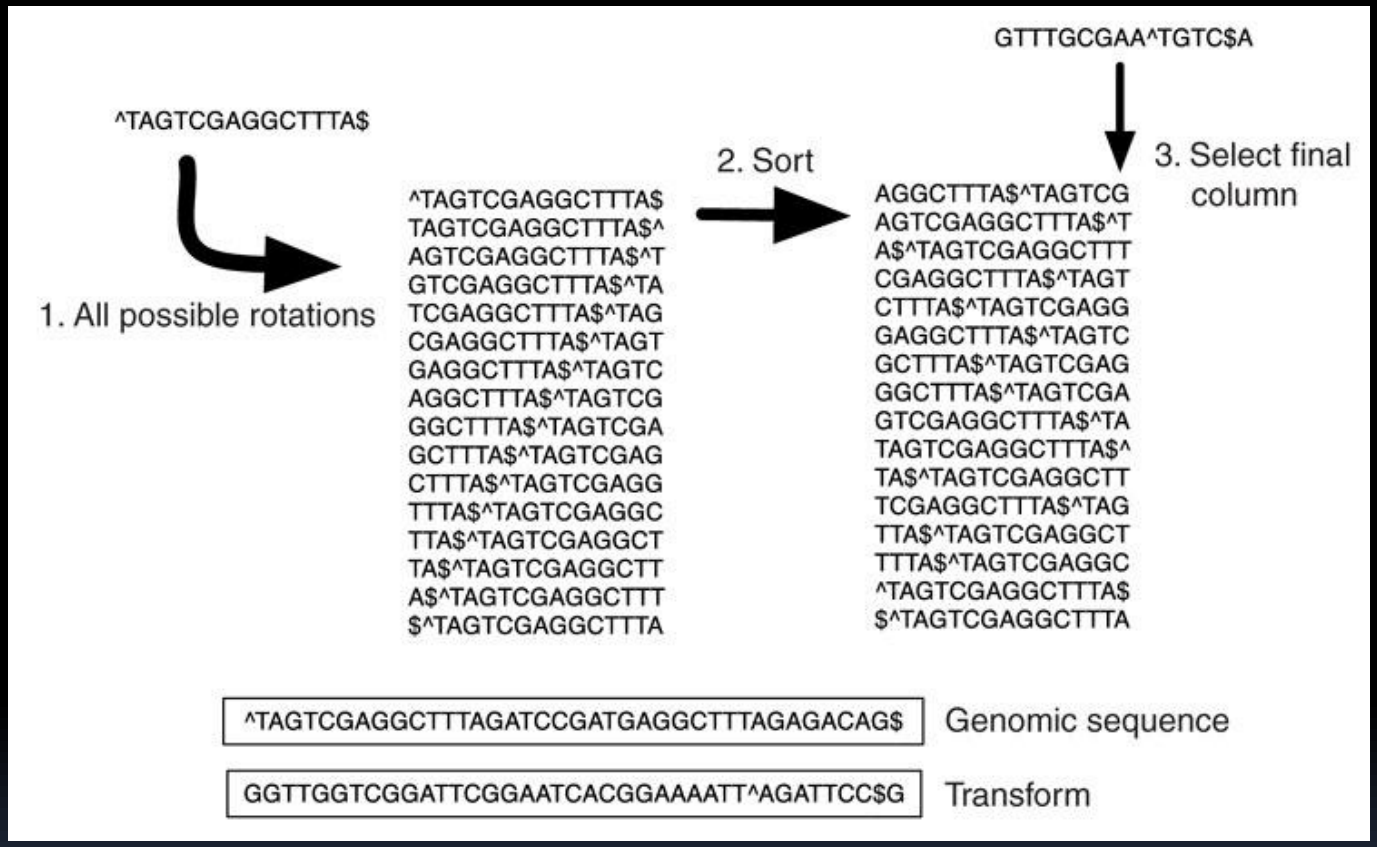
Sense from sequence reads: methods for alignment and assembly Paul Flicek & Ewan Birney Nature Methods 6, S6 - S12 (2009) Published online: 15 October 2009 Corrected online: 6 May 2010 doi:10.1038/nmeth.1376

Computer science tricks: HashTable



Sense from sequence reads: methods for alignment and assembly Paul Flicek & Ewan Birney Nature Methods 6, S6 - S12 (2009) doi:10.1038/nmeth.1376

Computer science tricks: Burrows-Wheeler



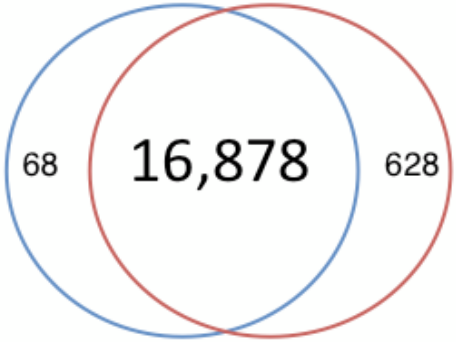
Sense from sequence reads: methods for alignment and assembly Paul Flicek & Ewan Birney Nature Methods 6, S6 - S12 (2009)
doi:10.1038/nmeth.1376

Genome alignment software

MAQ
SOAP2
Bowtie
BWA
Shrimp
…

http://en.wikipedia.org/wiki/List_of_sequence_alignment_software

Whole-exome NA12878 SNP overlaps: MAQ vs. BWA



MAQ	
Unique calls	68
No. in dbSNP	46 / 68 (68%)
Known Ti/Tv	2.54
No. filtered in other	58 / 68 (85%)
No. in dbSNP of filtered	91 / 249 (37%)

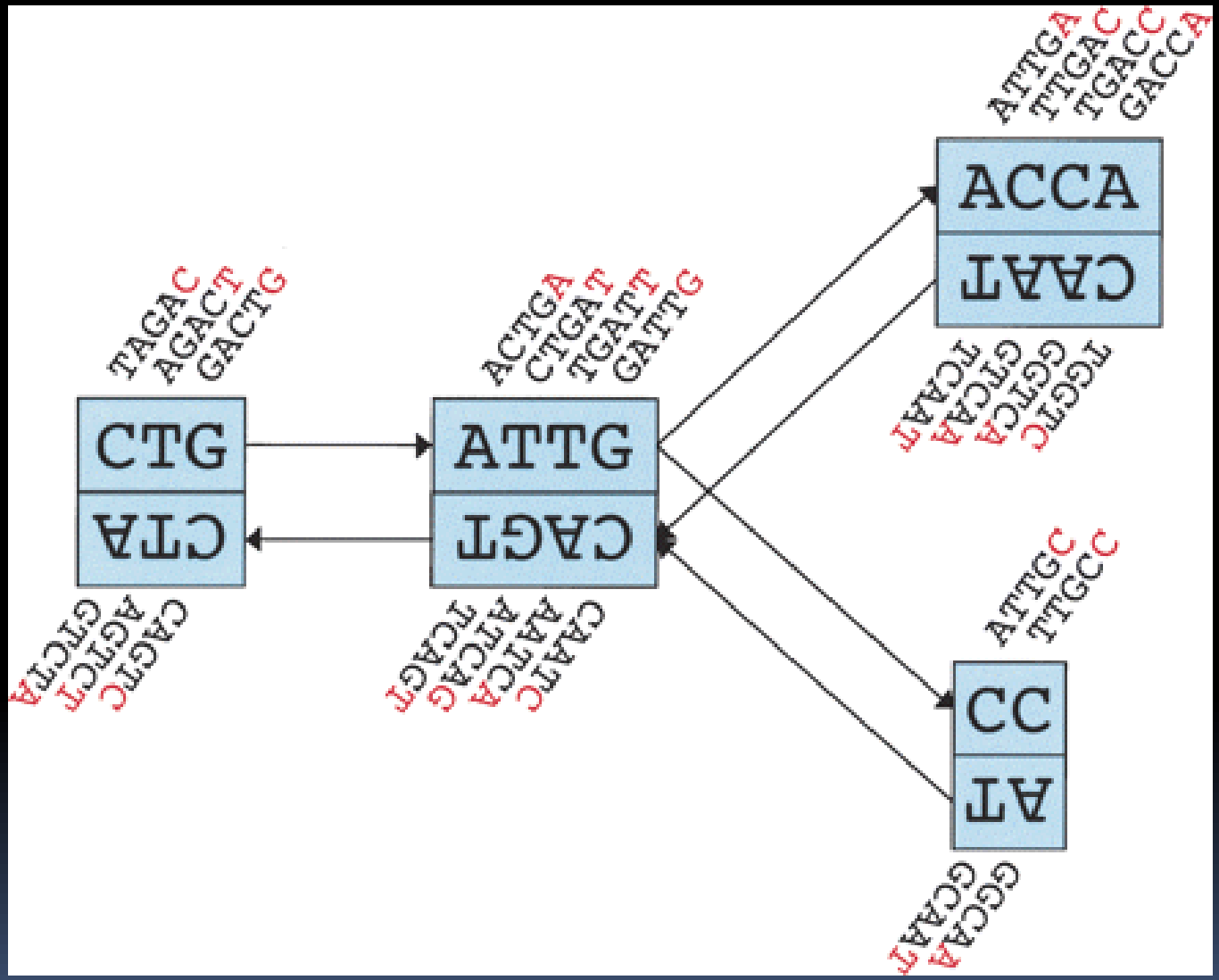
BWA	
Unique calls	628
No. in dbSNP	284 / 628 (45%)
Known Ti/Tv	2.26
No. filtered in other	411 / 628 (65%)
No. in dbSNP of filtered	229 / 1560 (15%)

Intersection	
Common calls	16878/ 17574 (96.7%)
No. in dbSNP	16617 / 16878 (98.5%)
Known Ti/Tv	3.29
Novel Ti/Tv	1.84

Slide design stolen from MAD

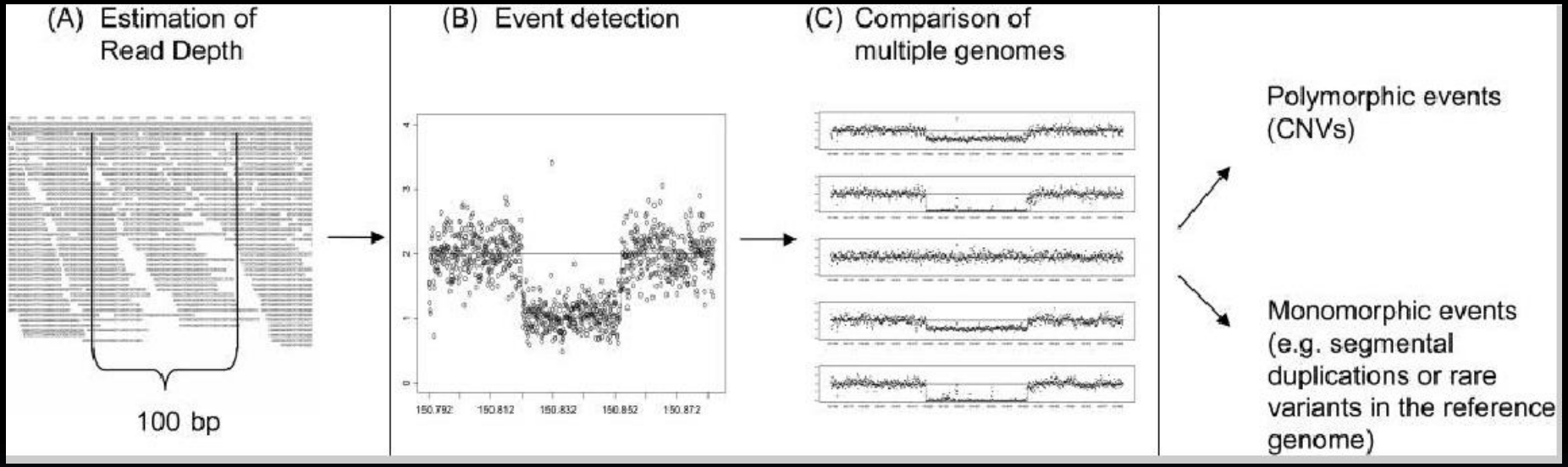
http://www.broadinstitute.org/gsa/wiki/index.php/File:ExampleDiagram.png

De novo sequencing: Bruijn graphs



Velvet: Algorithms for de novo short read assembly using de Bruijn graphs doi: 10.1101/gr.074492.107 Genome Res. 2008. 18: 821-829

Copy Number Variation detection

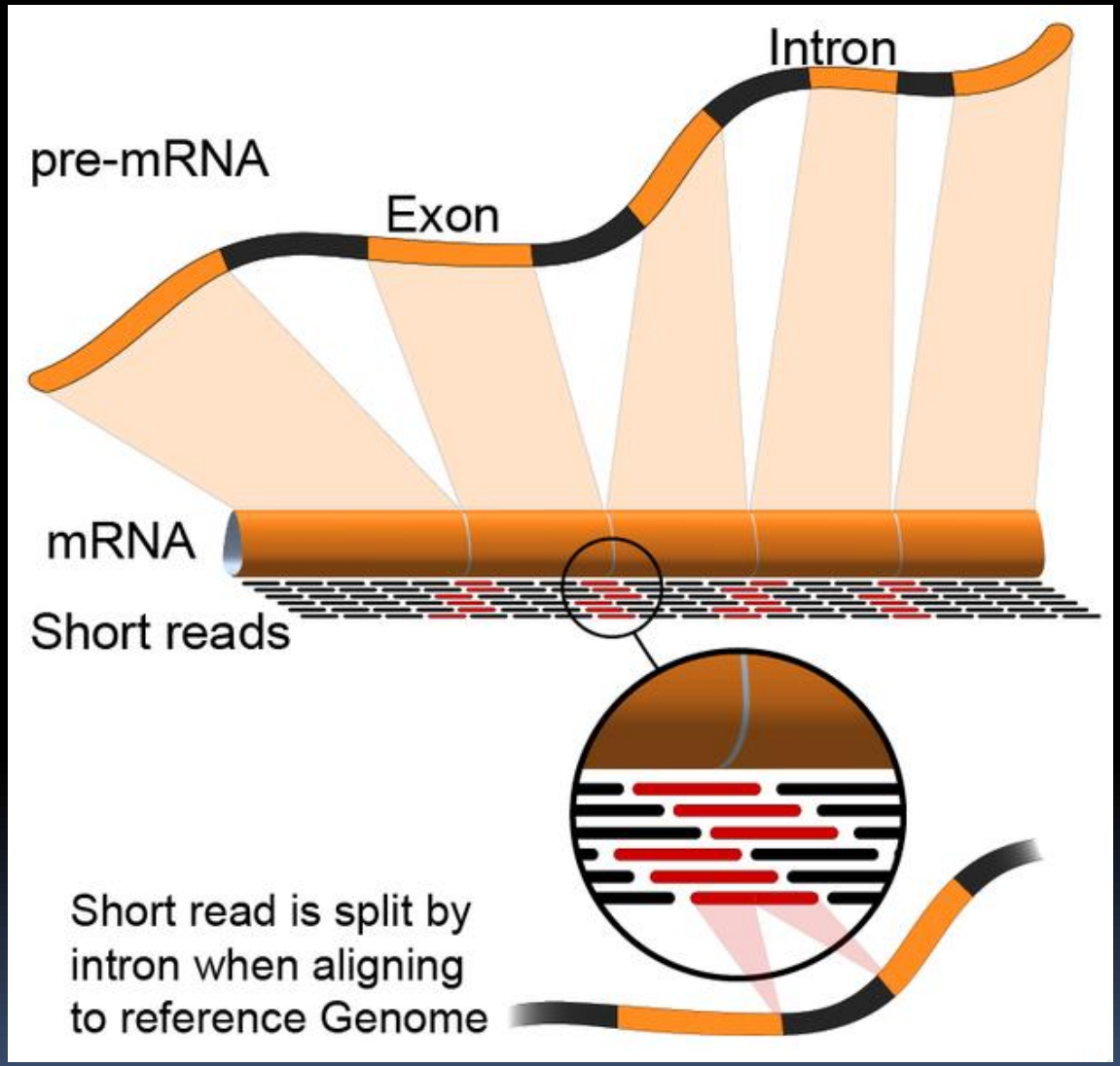


Genome Res. 2009 Sep;19(9):1586-92. Epub 2009 Aug 5.
Sensitive and accurate detection of copy number variants using read depth of coverage.

RNA-SEQ

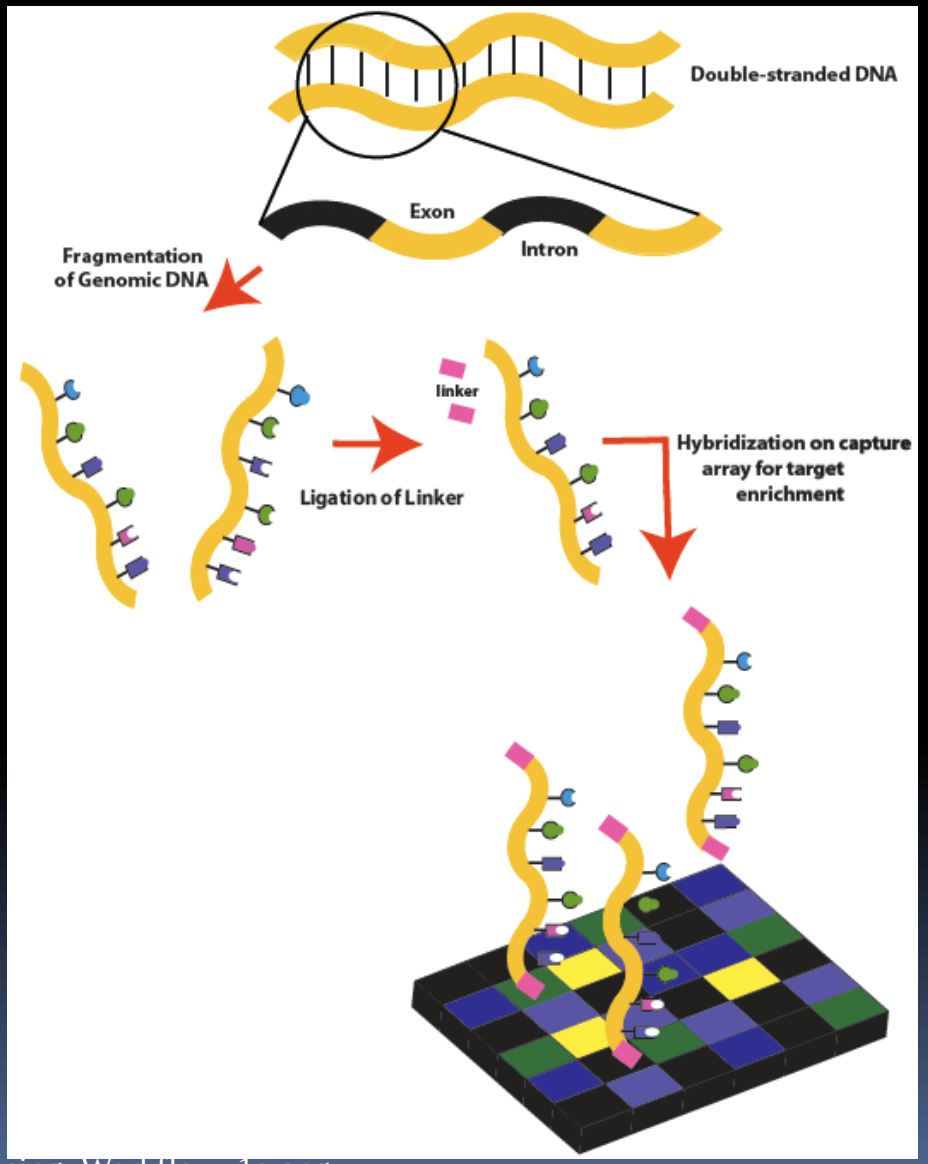
gene regulation

protein information



http://en.wikipedia.org/wiki/File:RNA-Seq-alignment.png

Exome Sequencing



http://en.wikipedia.org/wiki/File:Exome_Sequencing_Workflow_1a.png

Mate-pair / paired end
Epigenomic sequencing

And many other ….

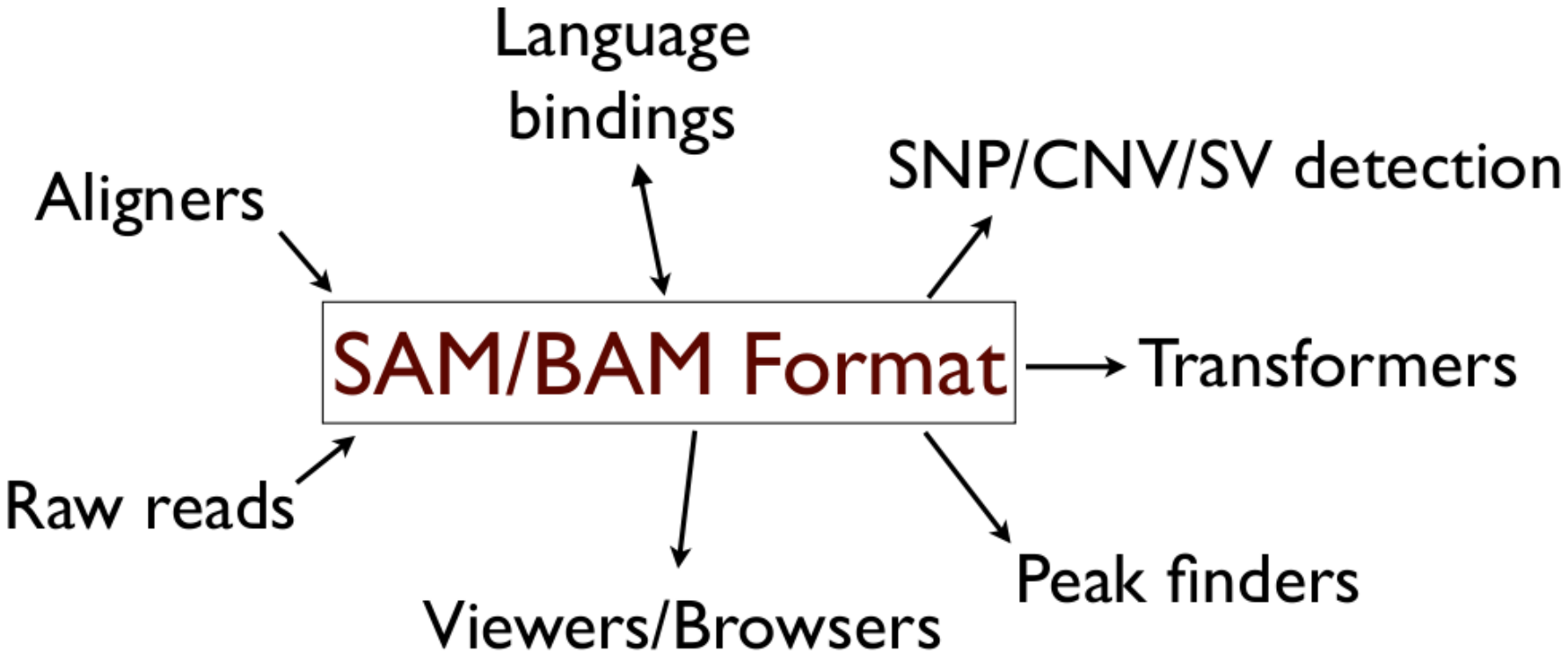
Output formats: SAM, BAM

```
DRR001355.27418 16    gi|323481965|ref|NC_015139.1|  135739  255   80M   *   0   0   CACGCGAGTGGGTAGGGTCTTCGCGTTGCNACGAATTAATCCGAATGCTCCGCCGCTTGTGCGGGCCCCCGTCAATTCCT   %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%B==B<@;:6@=BBA=927@AAAAB9@?AA   XA:i:0 MD:Z:0T0C0A0A0G0C0C0C3T0A3T7T4A0T8A2A0C7A28   NM:i:18
DRR001355.27418 0     gi|323481965|ref|NC_015139.1|  100946  255   80M   *   0   0   AGGAATTGACGGGGGCCCGCACAAGCGGCGGAGCATTCGGATTAATTCGTNGCAACGCGAAGACCCTACCCACTCGCGTG   AA?@9BAAAA@729=ABB=@6:;@<B==B%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%   XA:i:0 MD:Z:28T7G0T2T8A0T4A7A3T0A3G0G0G0C0T0T0G0A0   NM:i:18
DRR001355.43133 16    gi|323481965|ref|NC_015139.1|  132131  255   80M   *   0   0   CCTAGATGTCCAGTCATCTGCTGCGCCTCAACGCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACC   %%%%%%%%%%%%%%%%%%%%%%%?:4>?B?4B88543?:?B=ABBA9?BBCA;>CB>>4CBCCCCCBCBBBBACCBC;CA   XA:i:0 MD:Z:16A63   NM:i:1
DRR001355.43133 0     gi|323481965|ref|NC_015139.1|  104554  255   80M   *   0   0   GGTGAAATGCCACTCGAACCCAGAGCTAGCTGGTTCTCCCCGAAATGCGTTGAGGCGCAGCAGATGACTGGACATCTAGG   AC;CBCCABBBBCBCCCCCBC4>>BC>;ACBB?9ABBA=B?:?34588B4?B?>4:?%%%%%%%%%%%%%%%%%%%%%%%   XA:i:0 MD:Z:63T16   NM:i:1
DRR001355.52148 16    gi|323481965|ref|NC_015139.1|  130361  255   80M   *   0   0   CGGGATAGGGAGCGCACTGTCGCGCGACGTGCTGACCCCAGCTCGCGTGCCGCTTTAATGGGCGAACAGCCCAACCCTTG   %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%<9A6A54;A?BABBBB:BBB9BB@A9@A@<BB9A   XA:i:0 MD:Z:1C5T3C2A6T1A6T4A8A3A31   NM:i:10
DRR001355.52148 0     gi|323481965|ref|NC_015139.1|  106324  255   80M   *   0   0   CAAGGGTTGGGCTGTTCGCCCATTAAAGCGGCACGCGAGCTGGGGTCAGCACGTCGCGCGACAGTGCGCTCCCTATCCCG   A9BB<@A@9A@BB9BBB:BBBBAB?A;45A6A9<%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%   XA:i:0 MD:Z:31T3T8T4A6T1A6T2G3A5G1   NM:i:10
DRR001355.73033 16    gi|323481965|ref|NC_015139.1|  100328  255   80M   *   0   0   GGGGGCGACGGGGAGCCGGCCTGAGAGGGTGACCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAG   %%%%%%%%%%%%%%%%%%%%%%%%%%%=..44:44<05143>6=>A95;AAA5?=A:A99>>B?9==BB>>AABAABABA   XA:i:0 MD:Z:2C1A0T2T0C0A1T3T2T1C6A3T1A45   NM:i:13
DRR001355.73033 0     gi|323481965|ref|NC_015139.1|  136357  255   80M   *   0   0   CTGCCTCCCGTAGGAGTCTGGGCCGTGTCTCAGTCCCAGTGTGGCCGGTCACCCTCTCAGGCCGGCTCCCCGTCGCCCCC   ABABAABAA>>BB==9?B>>99A:A=?5AAA;59A>=6>34150<44:44..=%%%%%%%%%%%%%%%%%%%%%%%%%%%   XA:i:0 MD:Z:45T1A3T6G1A2A3A1T0G0A2A0T1G2   NM:i:13
```

DRR001355.27418 16
gi|323481965|ref|NC_015139.1|
135739 255 80M * 0 0
CACGCGAGTGGGTAGGGTCTTCGCGTTGCNACGAATTAATCCGAATGCTCCGCCGCTTGTGCGGGCCCCCGTCAATTCCT
%%%B==B<@;:6@=BBA=927@AAAAB
9@?AA
XA:i:0 MD:Z:0T0C0A0A0G0C0C0C3T0A3T7T4A0T8A2A0C7A28 NM:i:18

Field 1: Query name
Field 2: Flag
Field 3: Reference sequence name
Field 4: 1-based leftmost coordinate of the clipped sequence
Field 5: Mapping quality
Field 6: CIGAR strings
Field 7: Mate reference sequence name
Field 8: 1-based leftmost coordinate of the clipped sequence
Field 9: Insert size (5’ to 5’)
Field 10: Query sequence
Field 11: Sequence qualities

http://samtools.sourceforge.net/samtools.shtml



http://gorgonzola.cshl.edu/pfb/2010/LectureNotes/ngs2/ngs2.pdf

Pileup

```
seq1 272 T 24  ,.$....,,.,.,...,,,.,..^+. <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,.,.,...,,,.,..A <<<;<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,,.,.,...,,,.,...    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,,.,.,...,,,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,.,....  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,,.,.,.C.,,,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  ....,,.,.,...,,,.,....^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,.,..... ;75&<<<<<<<<<=<<<9<<:<<
```

Chrom

Position

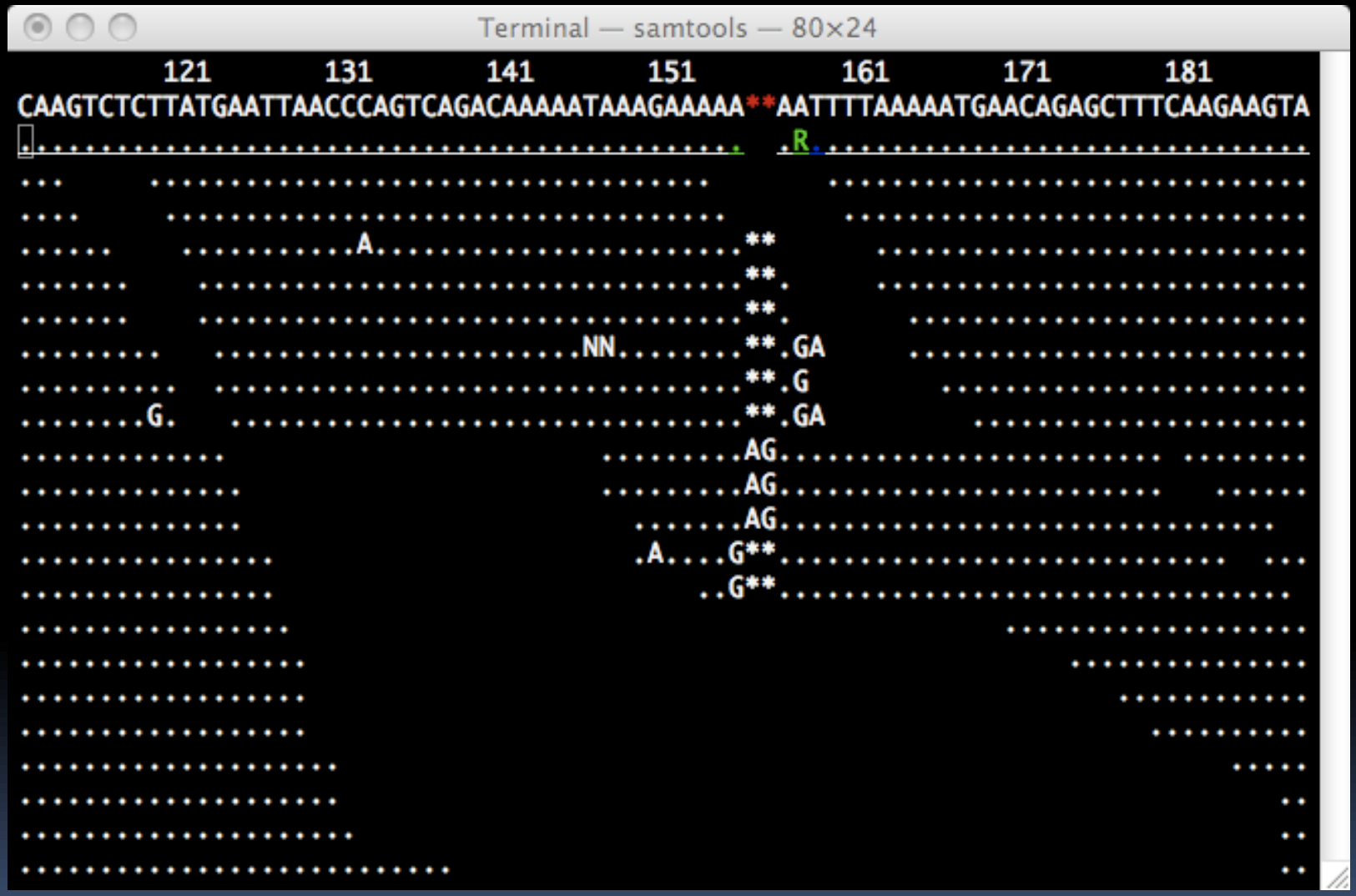
Ref

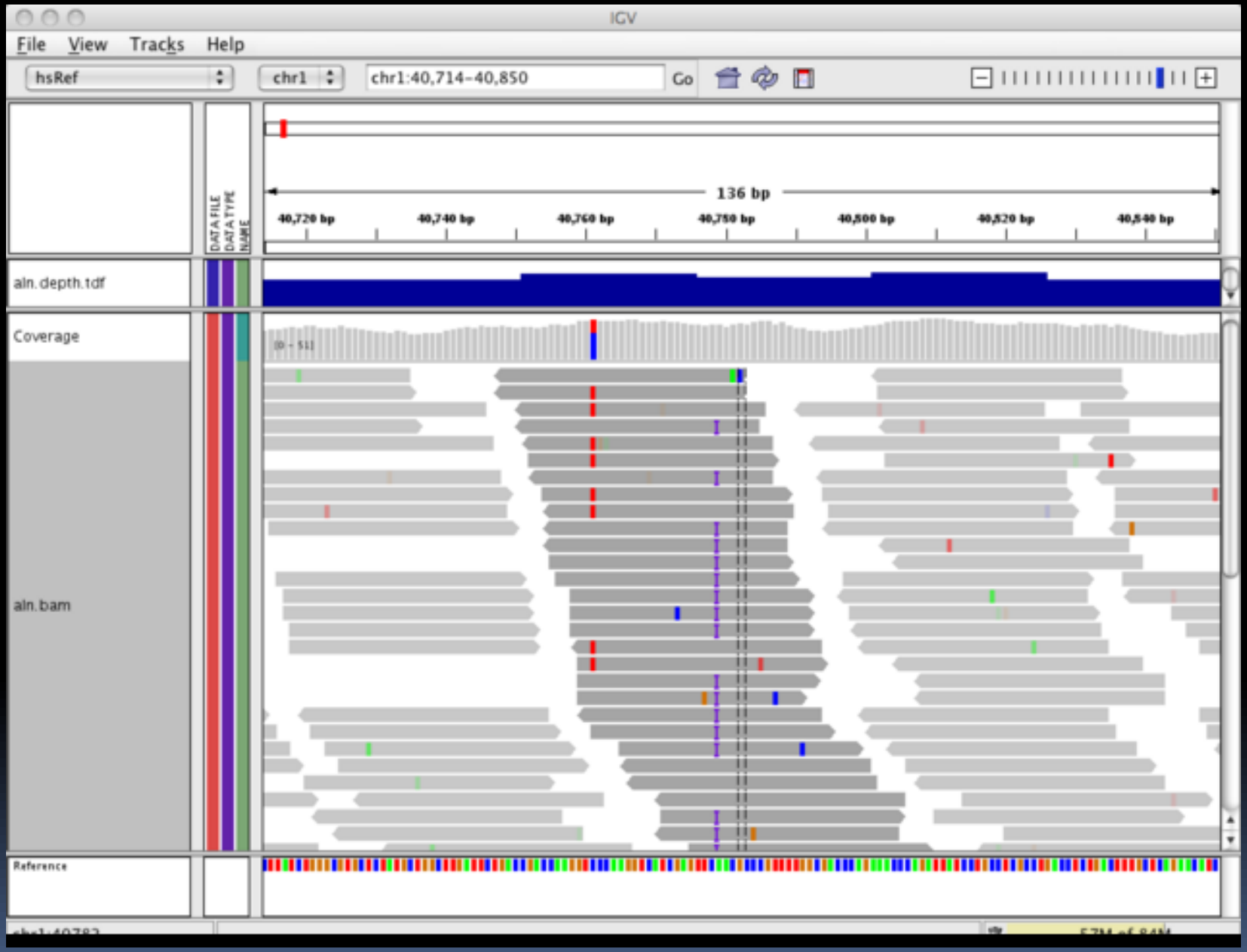
Coverage

Read bases

Qualities

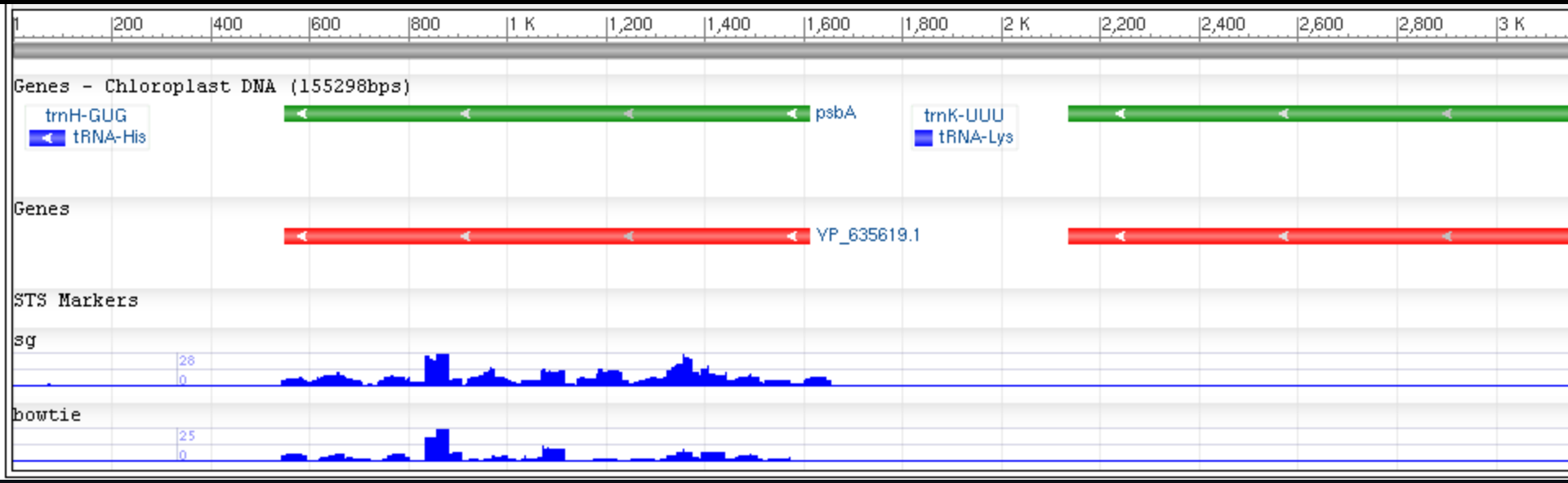
Samtools: TVIEW

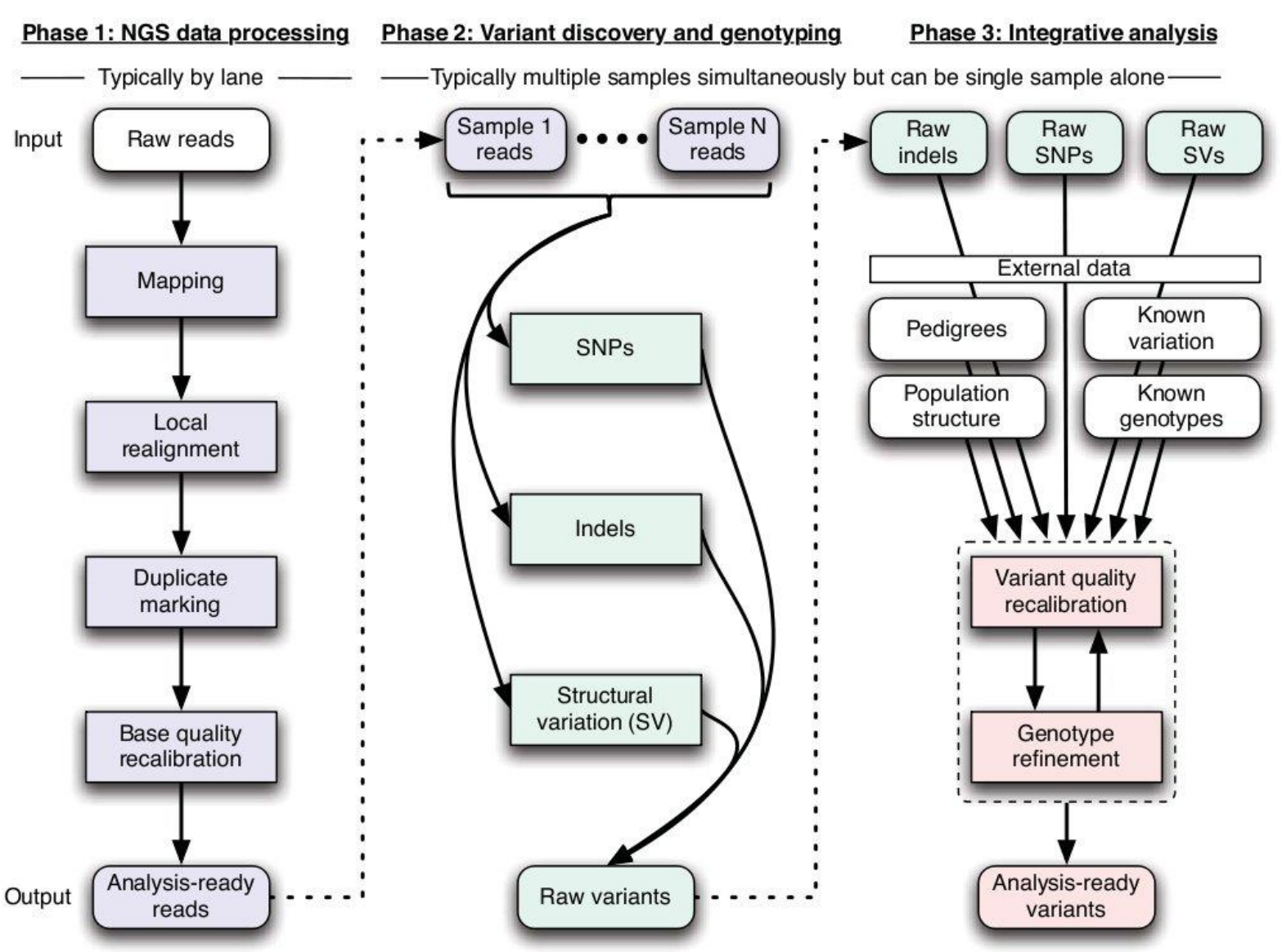




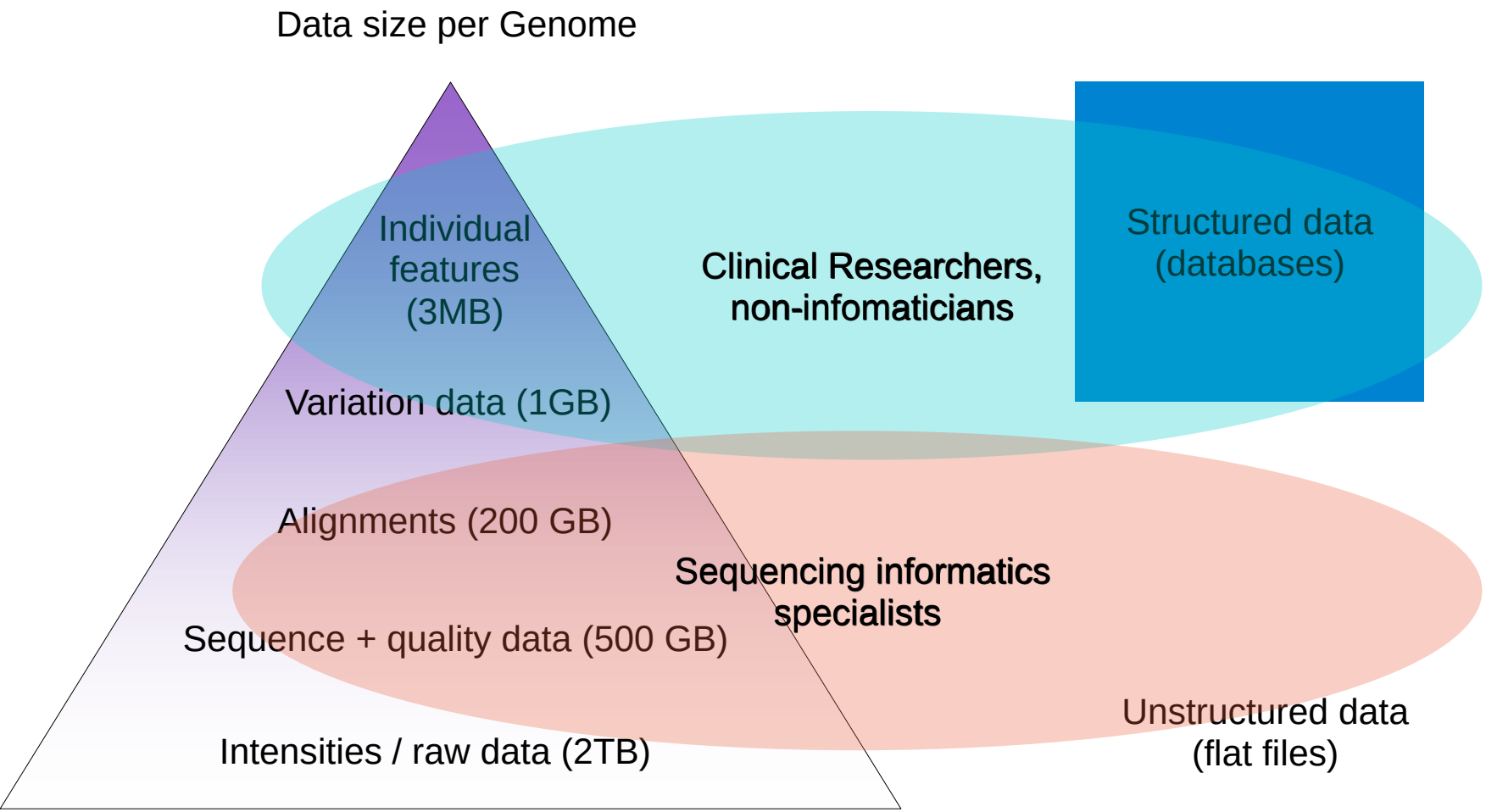
http://www.broadinstitute.org/software/igv/

Viewing the results: NCBI/UCSC genome browsers



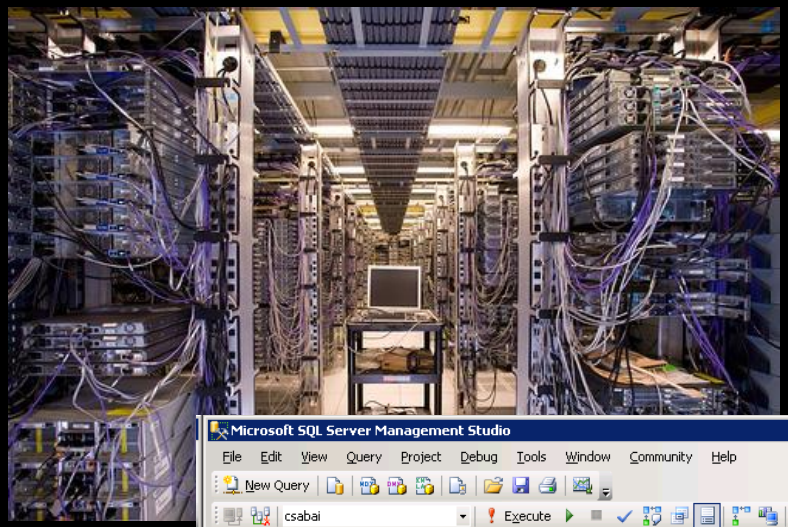

Phase 1: NGS data processing
Phase 2: Variant discovery and genotyping
Phase 3: Integrative analysis
Typically by lane
Typically multiple samples simultaneously but can be single sample alone
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Sample 1 reads
Sample N reads
SNPs
Indels
Structural variation (SV)
Raw variants
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants

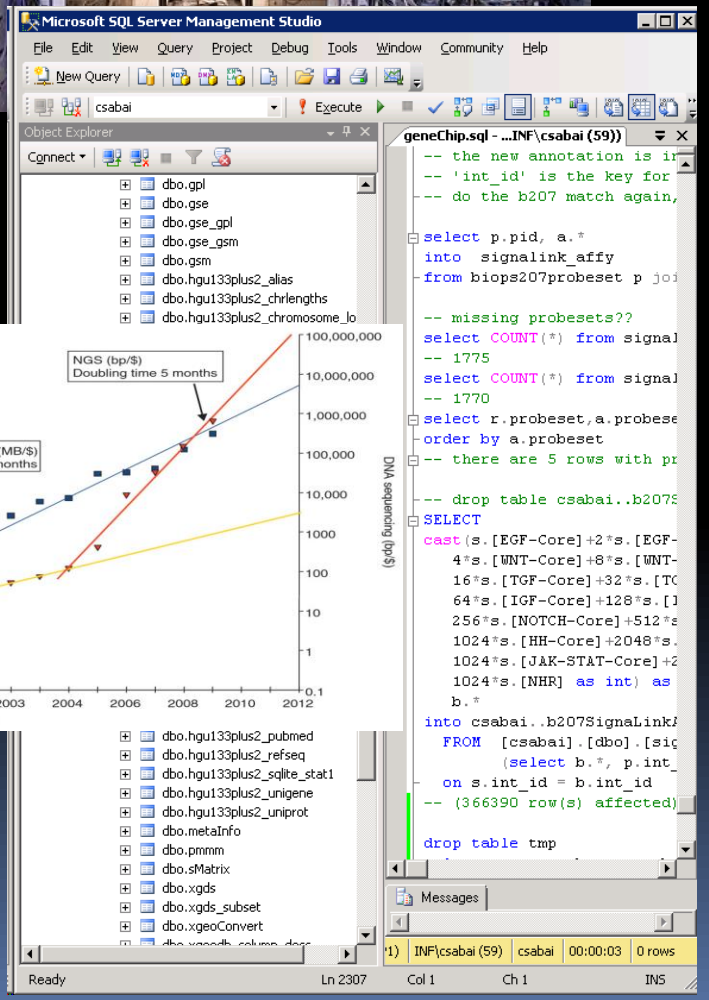
Genomics Data – Big Data Challenge

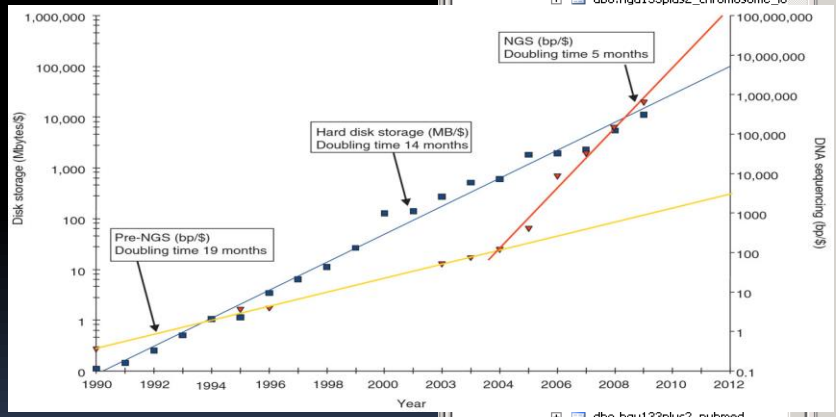


Source: Guy Coates, Wellcome Trust Sanger Institute

- Doubling rate:
  - Transistors/compute/storage: 18mo
  - Sequencing: 12mo
- Storage
  - Database technology
  - Cloud storage
  - Sharing: archives
- Alignment: 3Gb x 1Greads
  - Clever algorithms
  - Parallelization, cloud, GPU
- Bioinformatics
  - Who will write the code?
  - Who will know how to use?
  - Beware of "black boxes"



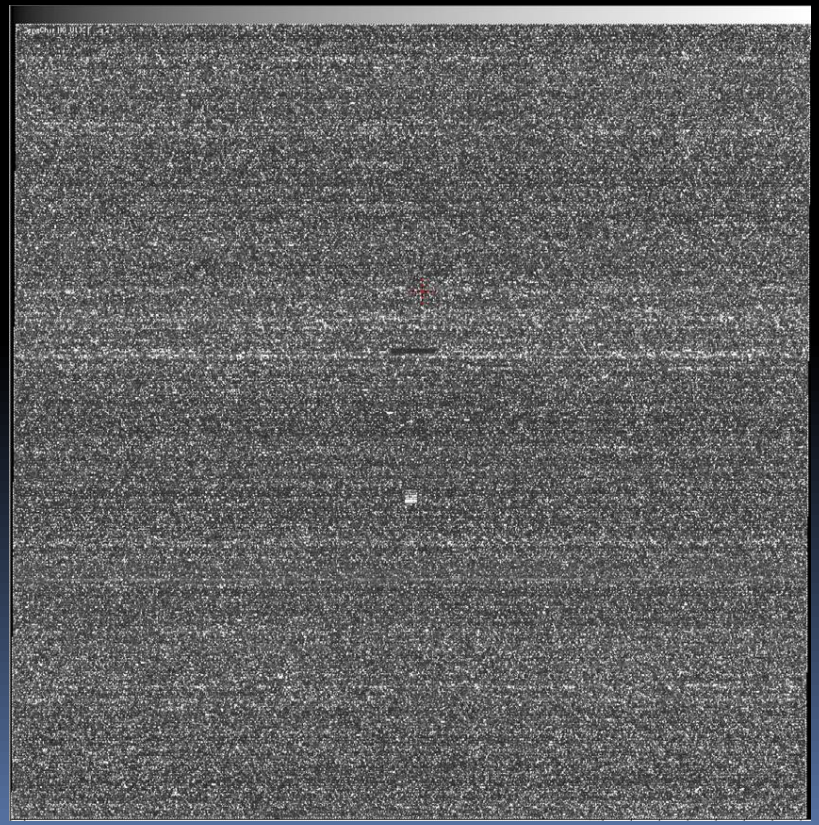


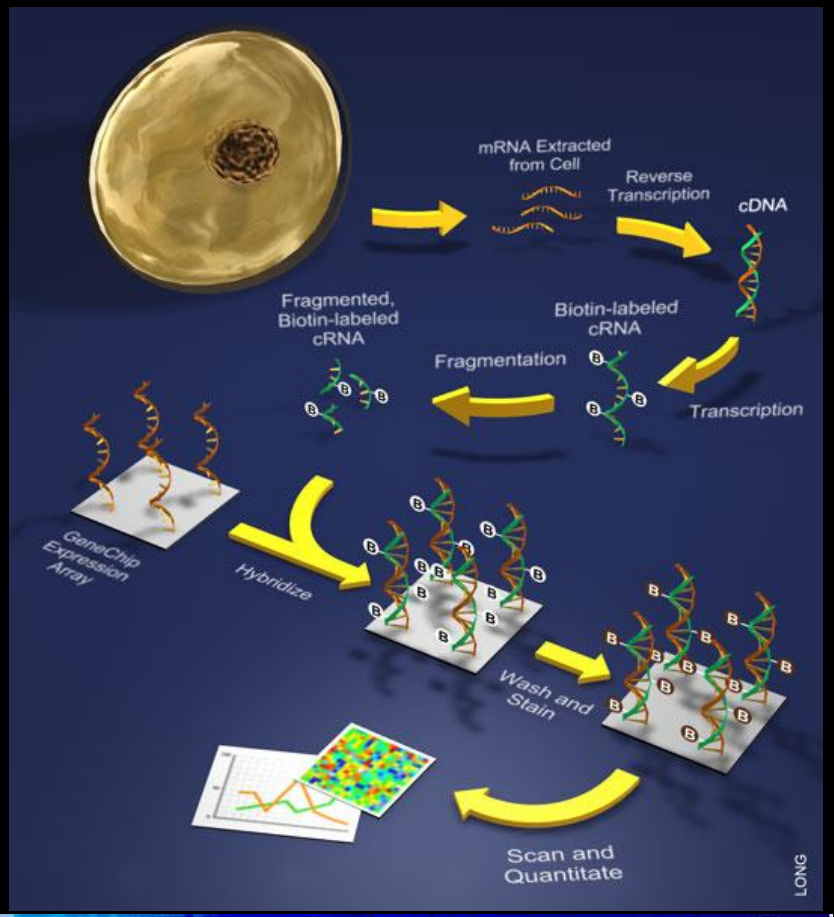


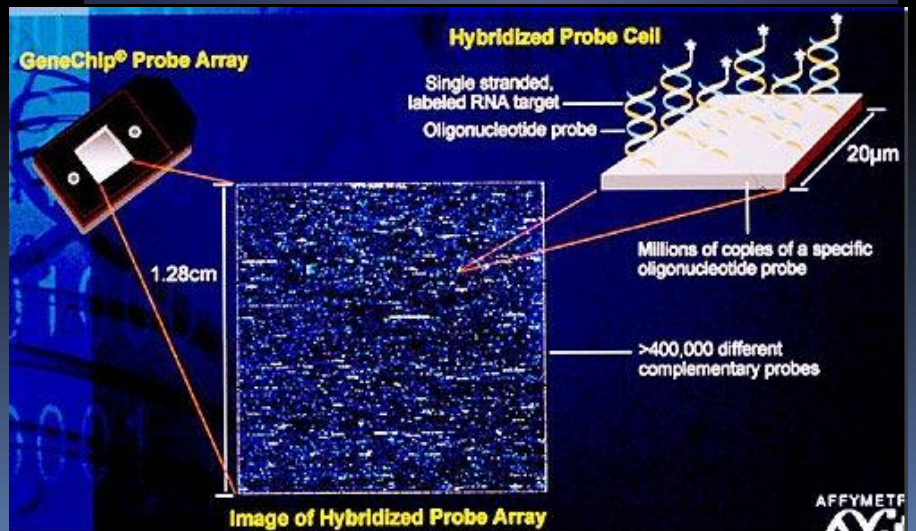
Dimension reduction - microarrays

Microarrays

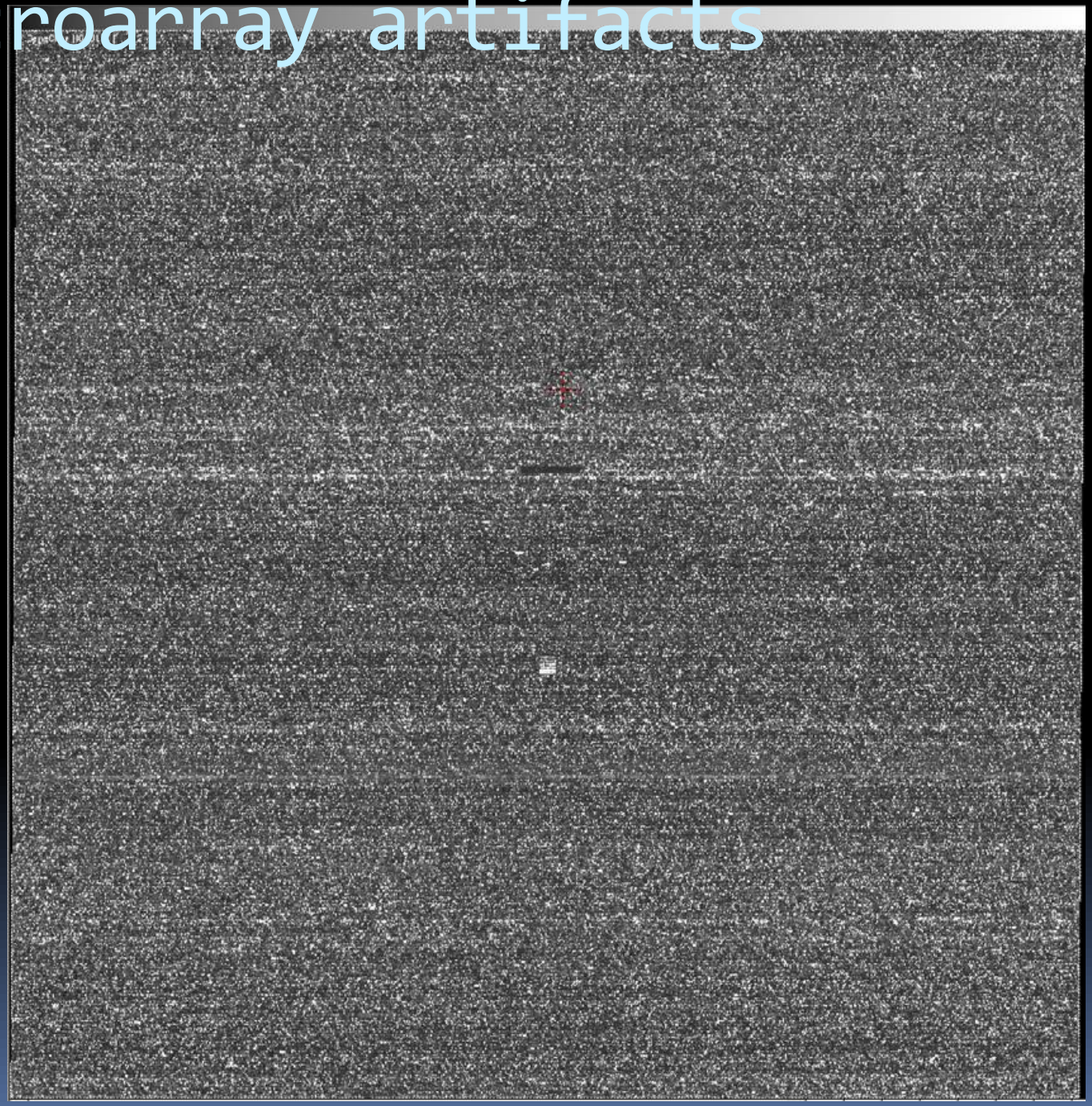
- Affymetrix HG U133 Plus2
  - Raw image 67Mpix (photometry!)
  - 604258 probes
  - 54675 probe sets





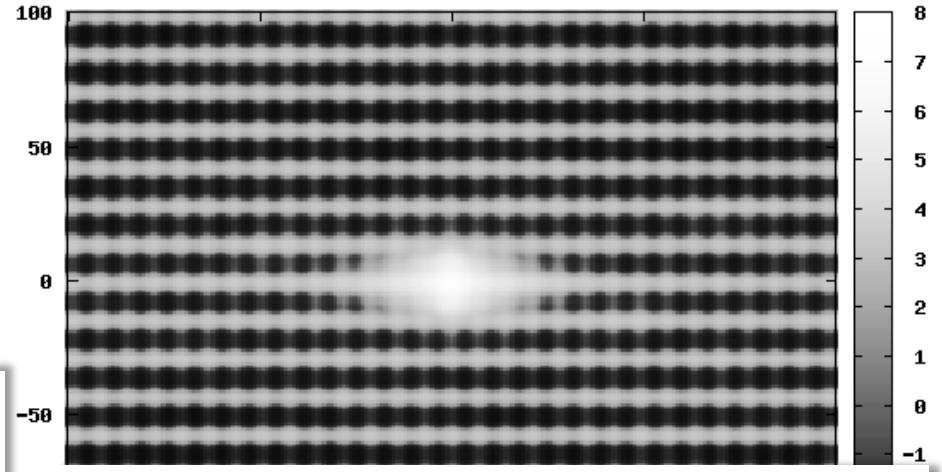


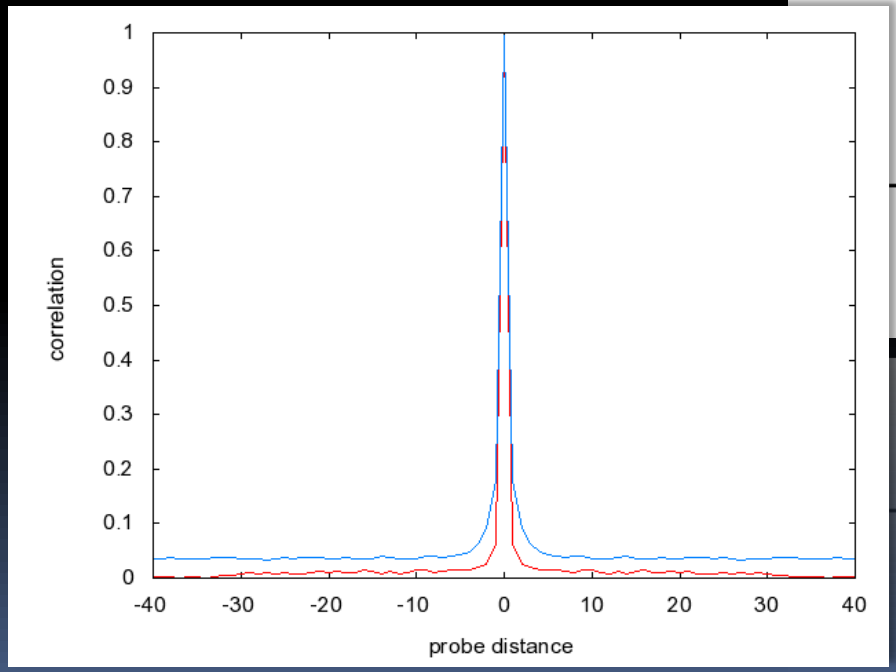
Microarray artifacts

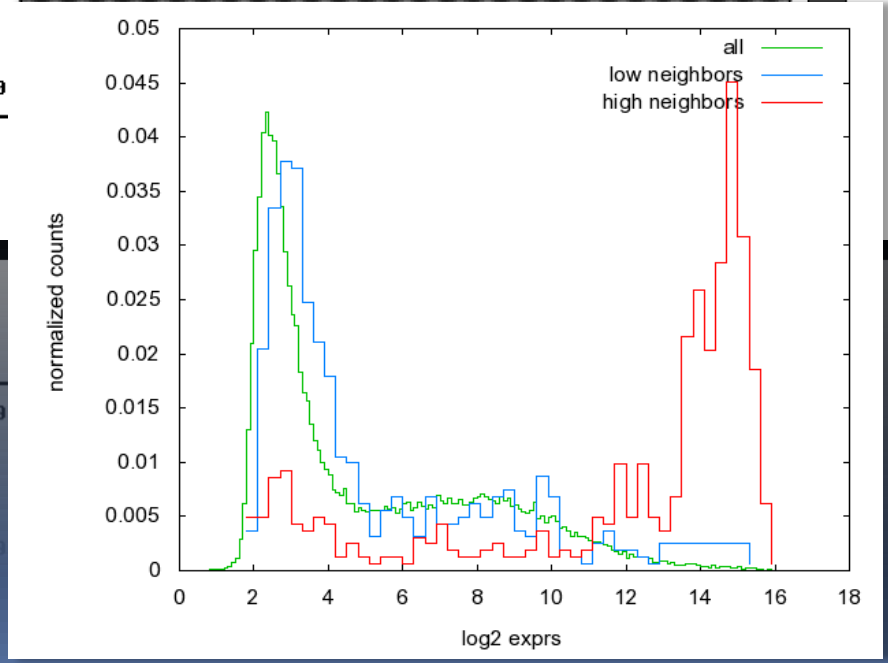


Microarray artifacts

- Raw image cross-correlation: bleeding of bright cells
- Can be seen in CEL/exprs data, too
- Leave out / deconvolution

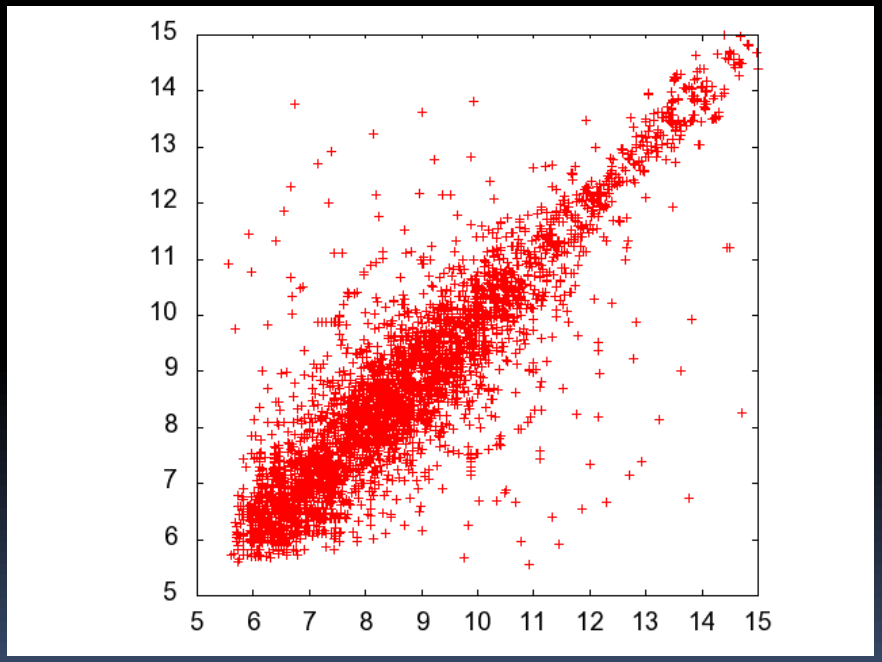




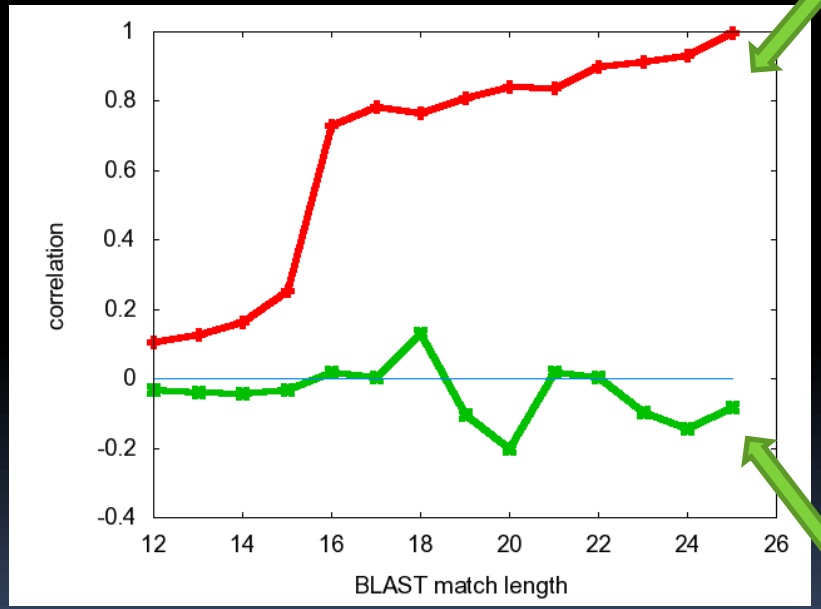


Cross-hybridization

- HGU133Plus2: 604,258 „perfect match" 25-mer sequence
- All pairs BLAST: 18M have longer than 12 overlap, 58138 has longer than 15 overlap
- Example: overlap=22, Corr.coeff: 0.92

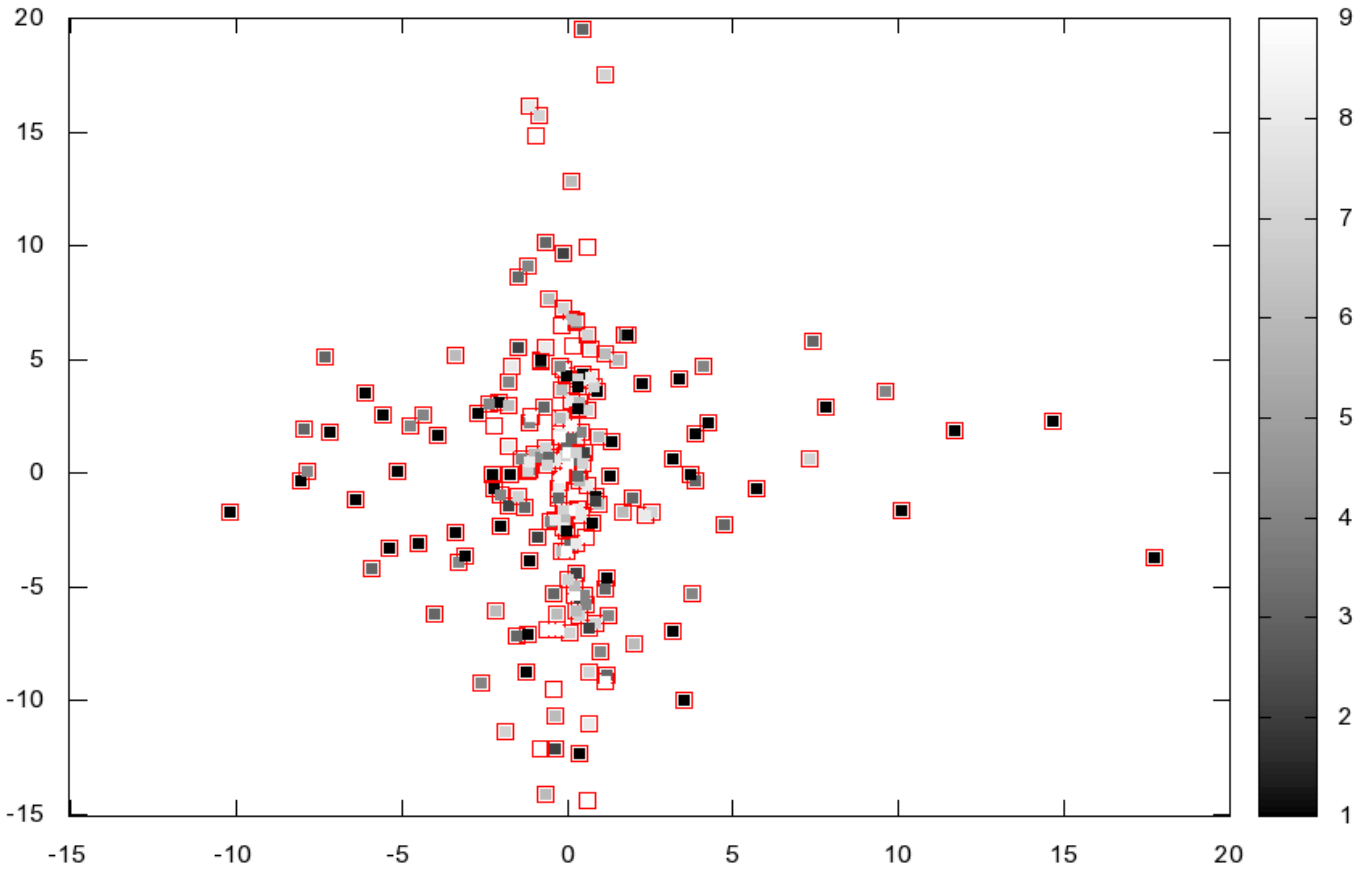


Normal BLAST: strong crosshybr for overlaps above 15



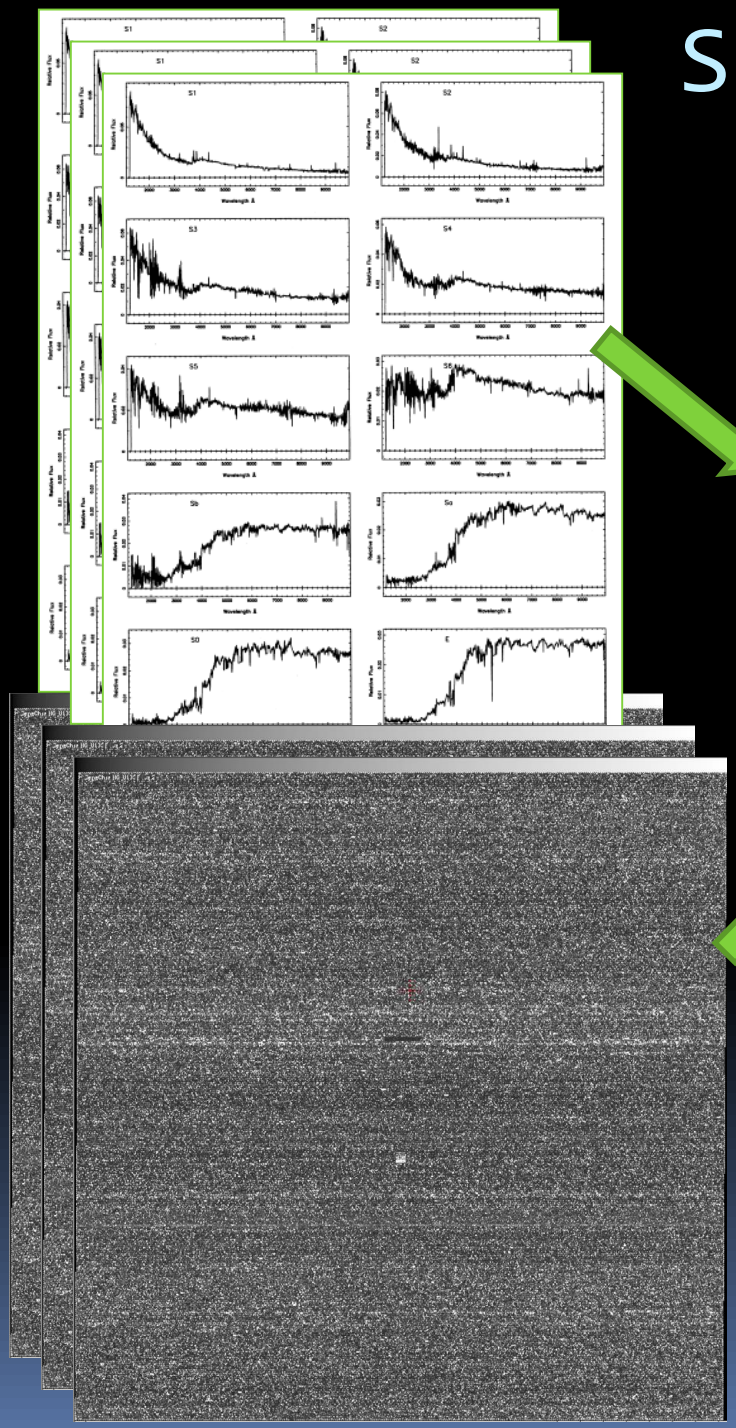
Reverse-complement BLAST: bulk hibridization?

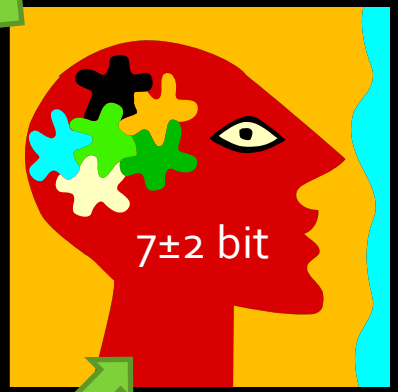
- After cleaning, preprocessing… still too much data
  - 54265 dimensional vectors for each sample

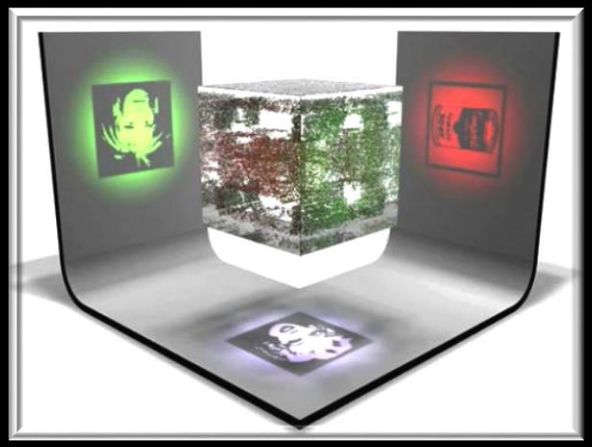


Similar challenges

- SDSS spectra: 1 million times 3000 dimensional vectors
- Microarray study: 207 times 54675 dimensional vectors







Compression : dimension reduction, matrix factorization, machine learning

Hope: There are "physical laws" at the background, so data points do not fill the space. They are constrained to subspaces/hyper-surfaces. This is our only chance to understand the world!

Principal components

Rotations in 54675 dimensions – projection to 2

2D rotation matrix:

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Advantages:

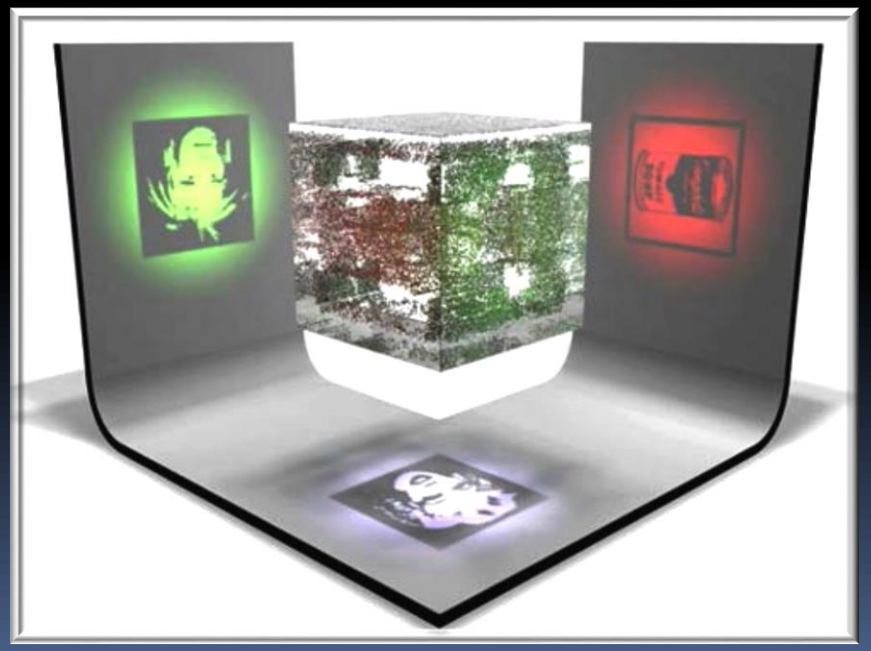
- "close enough" to raw data
- unsupervised

3D rotation matrices:

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$$

$$R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}$$

$$R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$



4D … 54675D rotation matrices:

SIGGRAPHASIA2009

Shadow Art

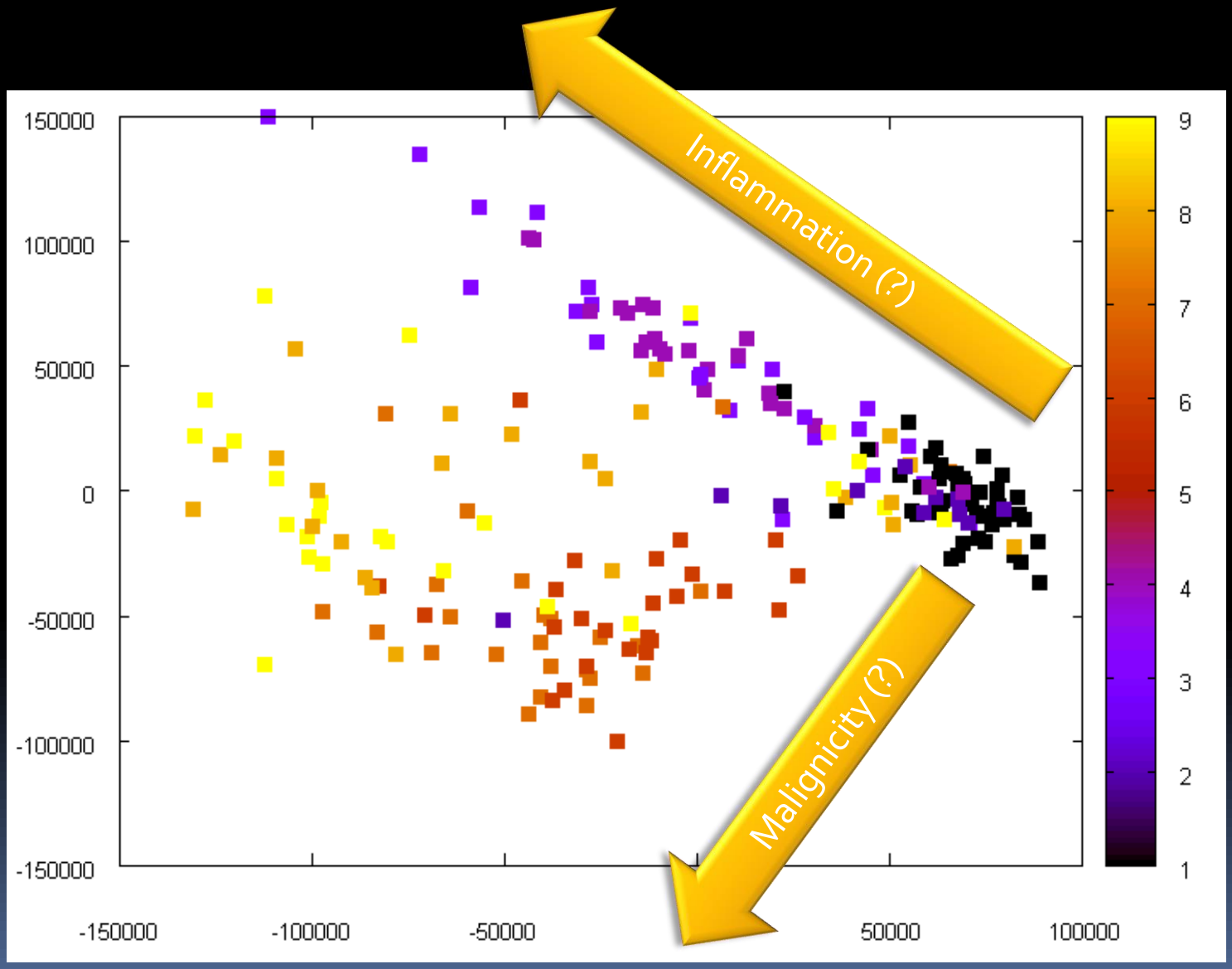
Niloy J. Mitra
IIT Delhi / KAUST

Mark Pauly
ETH Zurich

Principal components

Rotations in 54000 dimensions – projection to 2

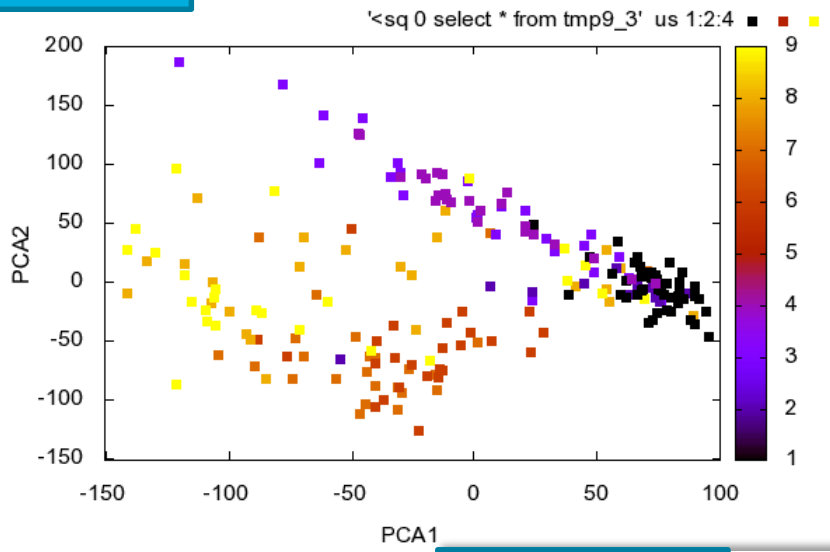
PCA1, PCA2


Inflammation (?)
Malignicity (?)
150000
100000
50000
0
-50000
-100000
-150000
-150000
-100000
-50000
50000
100000
9
8
7
6
5
4
3
2
1
CRC 2
CRC 1
AD2
AD1
IBD2
IBD1
NEG

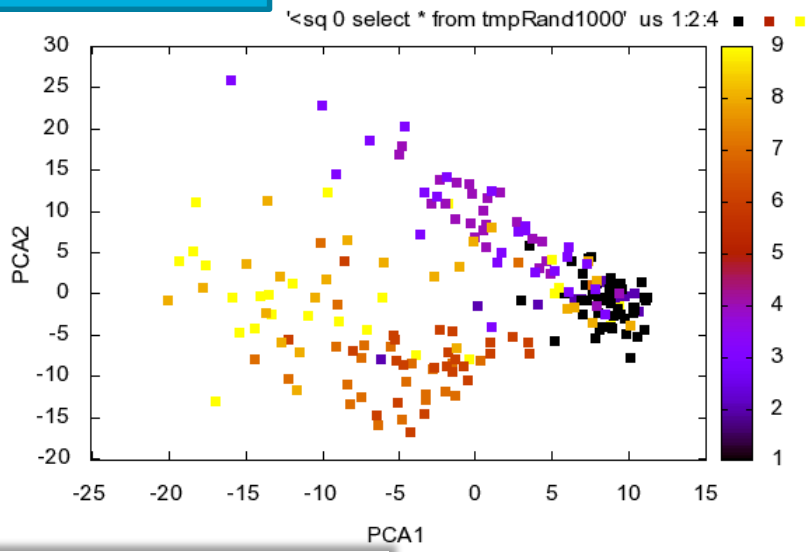
“Realize that everything connects to everything else.”

/Leonardo da Vinci/

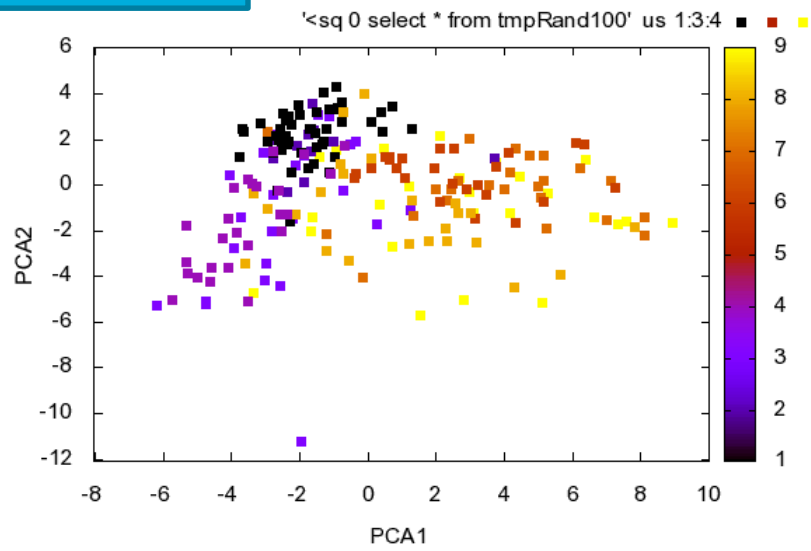
All 54265



Random 1000

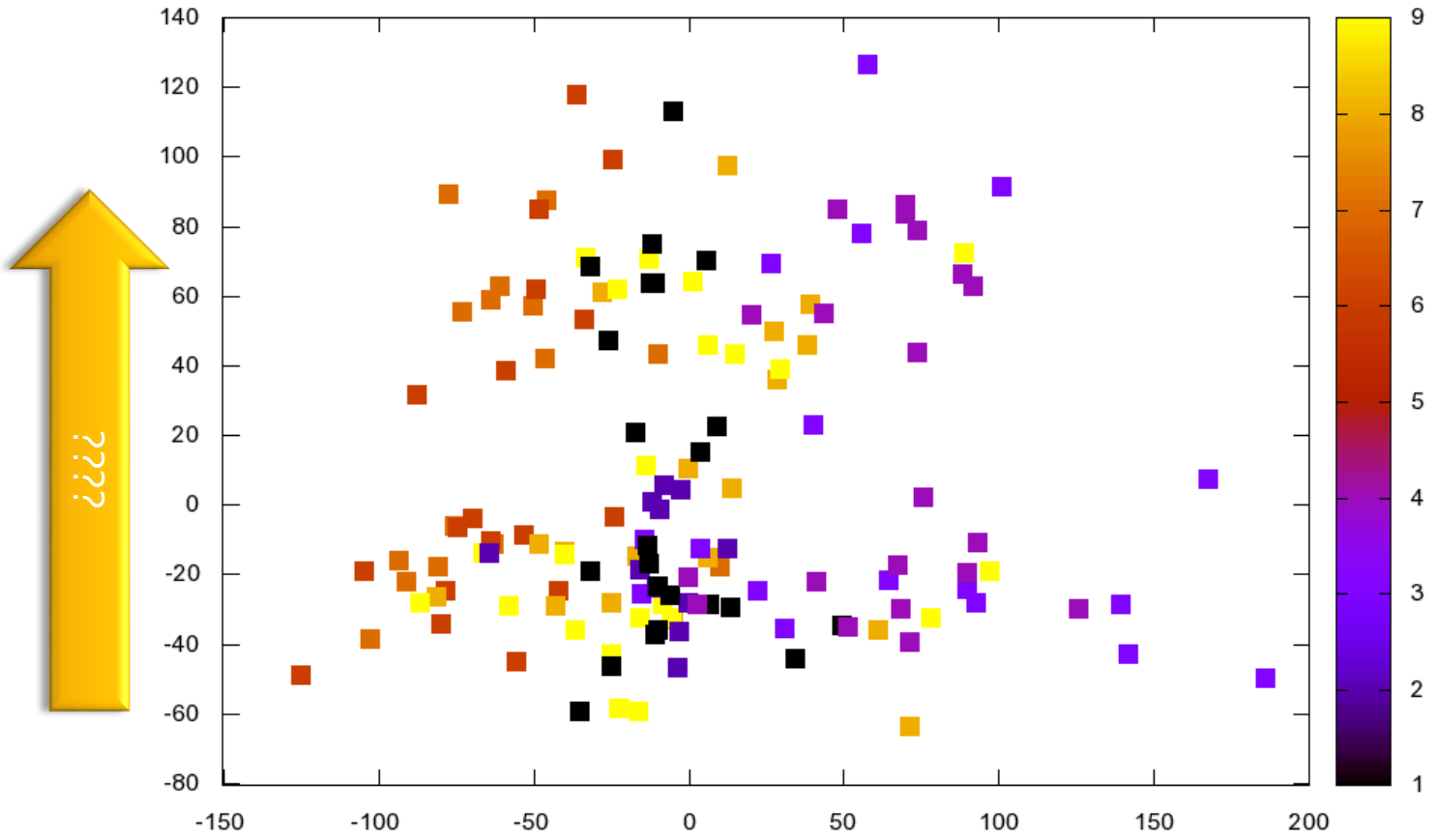


Random 100

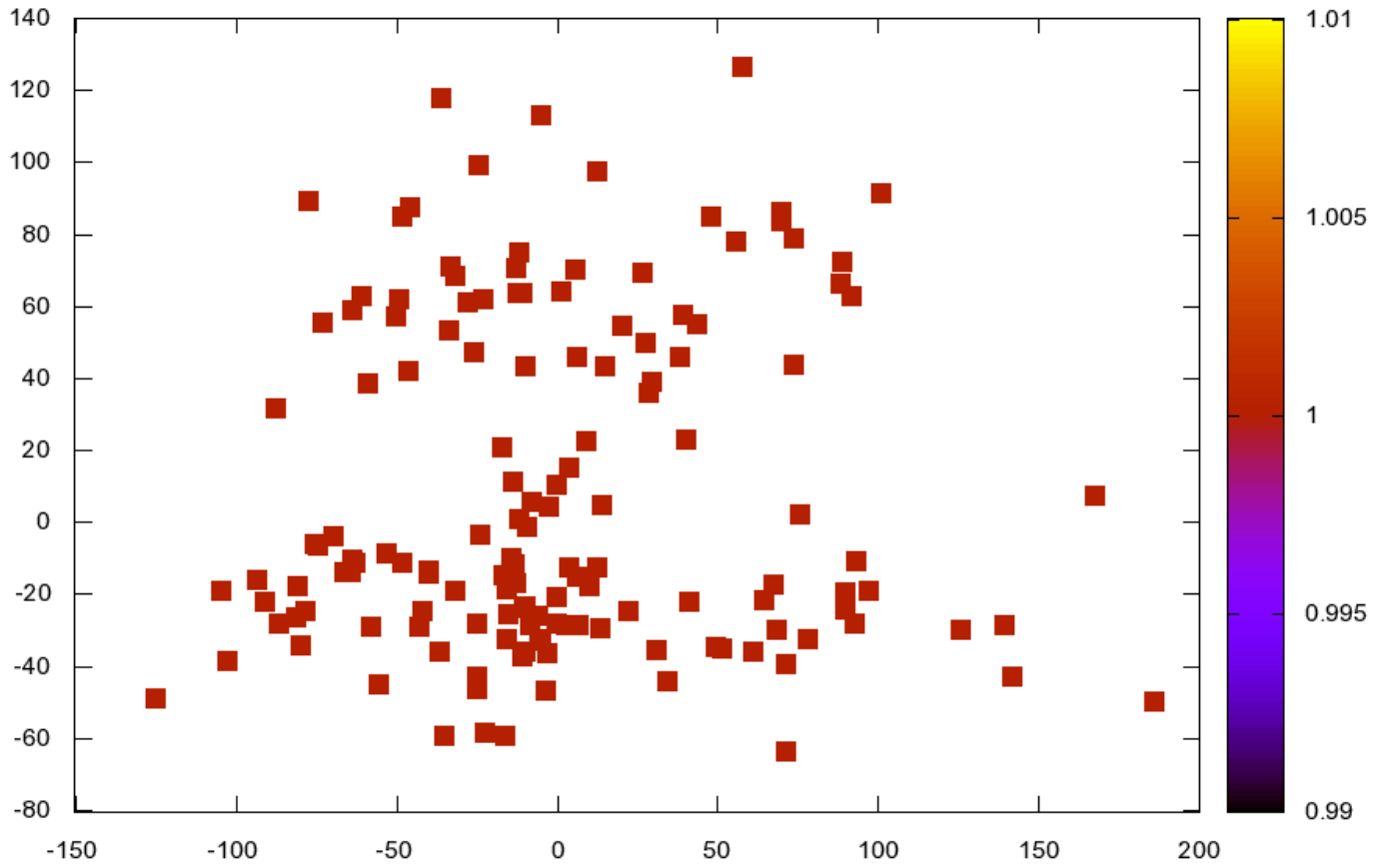


Marker genes?

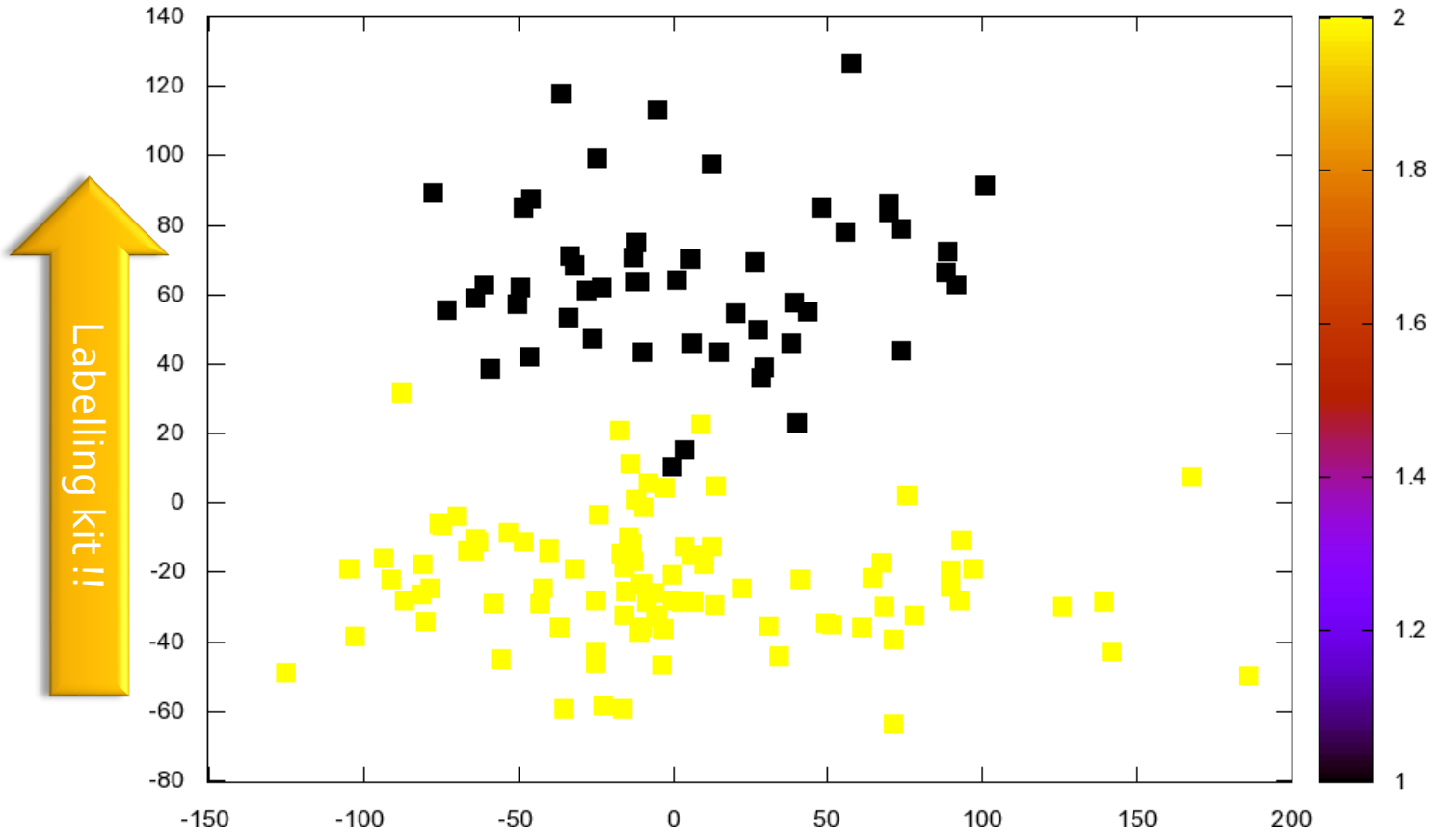
PCA2, PCA3


¿¿¿¿
140
120
100
80
60
40
20
0
-20
-40
-60
-80
-150
-100
-50
0
50
100
150
200
9
8
7
6
5
4
3
2
1

PCA2, PCA3

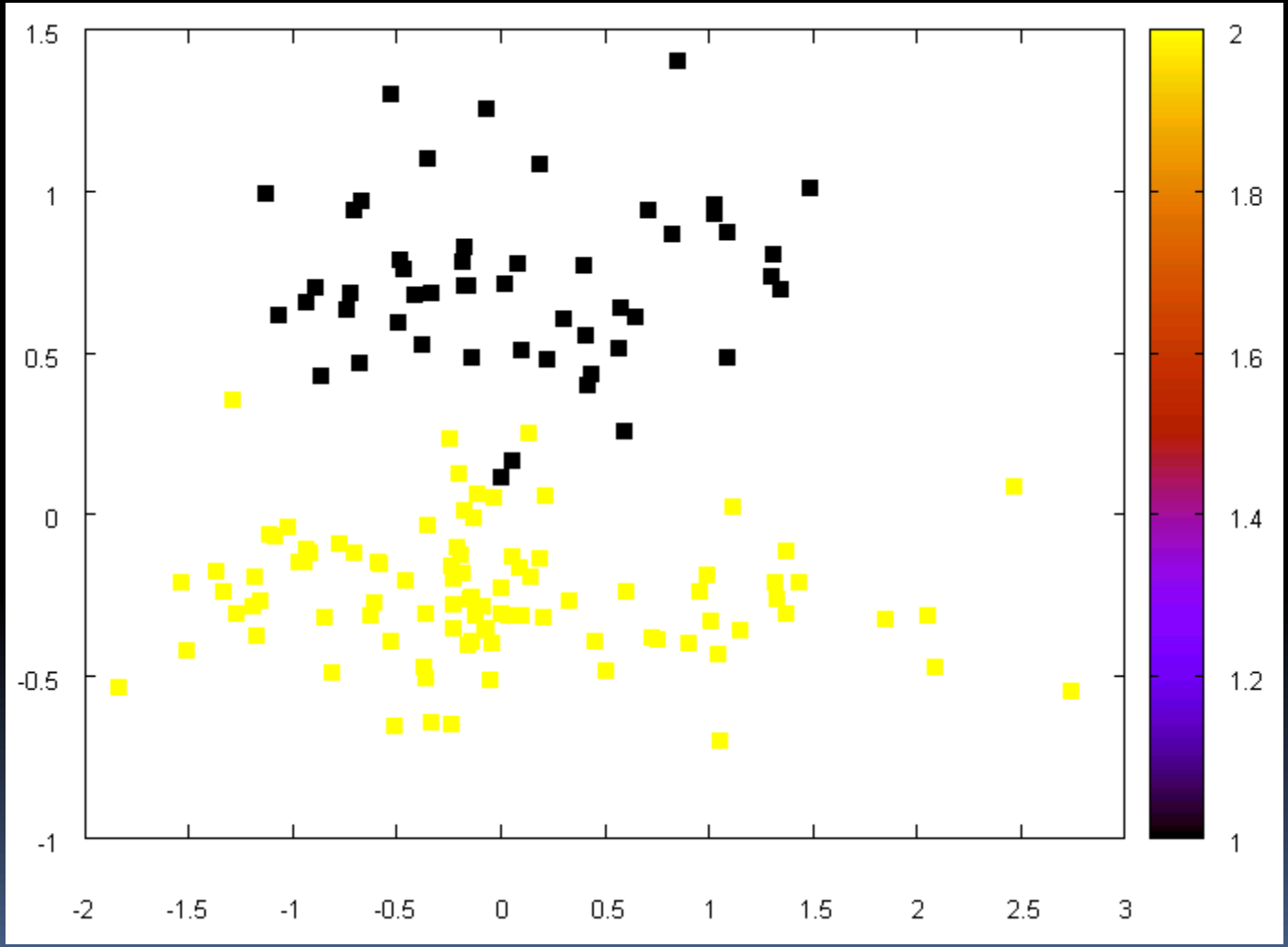

140
120
100
80
60
40
20
0
-20
-40
-60
-80
-150
-100
-50
0
50
100
150
200
1.01
1.005
1
0.995
0.99

PCA2, PCA3

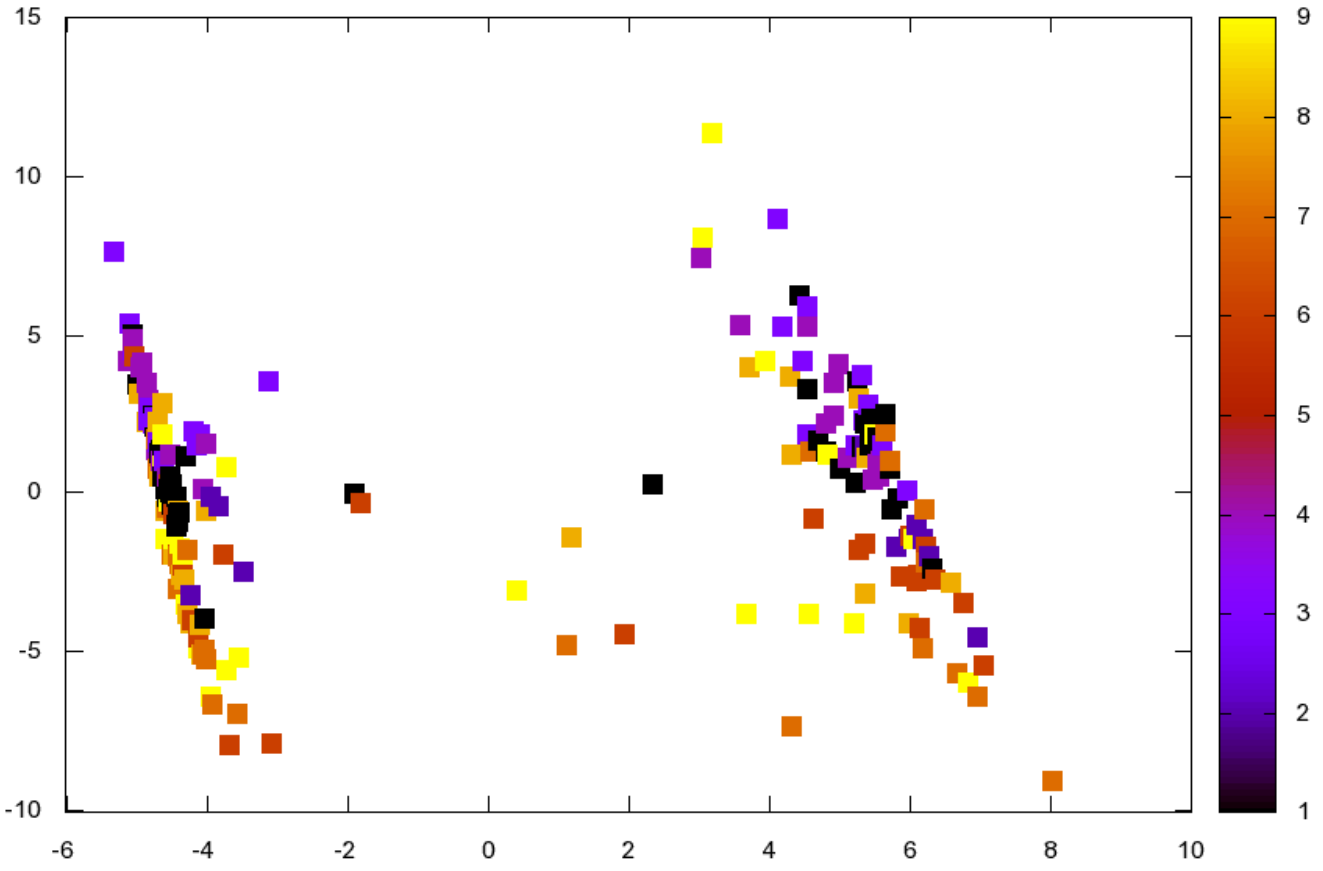

Labelling kit !!
140
120
100
80
60
40
20
0
-20
-40
-60
-80
-150
-100
-50
0
50
100
150
200
2
1.8
1.6
1.4
1.2
1

Főkomponens elemzés

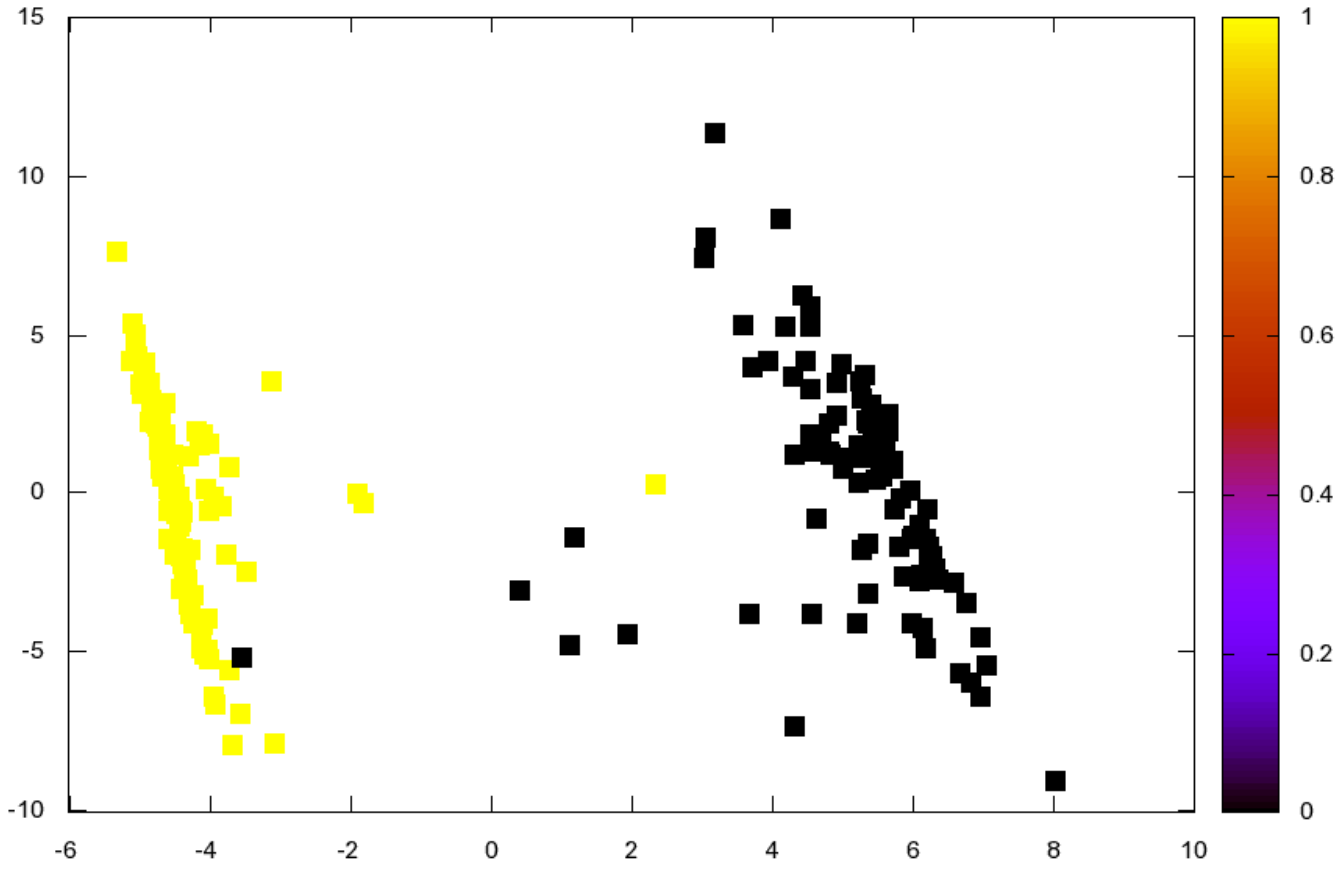
- Labelling kit

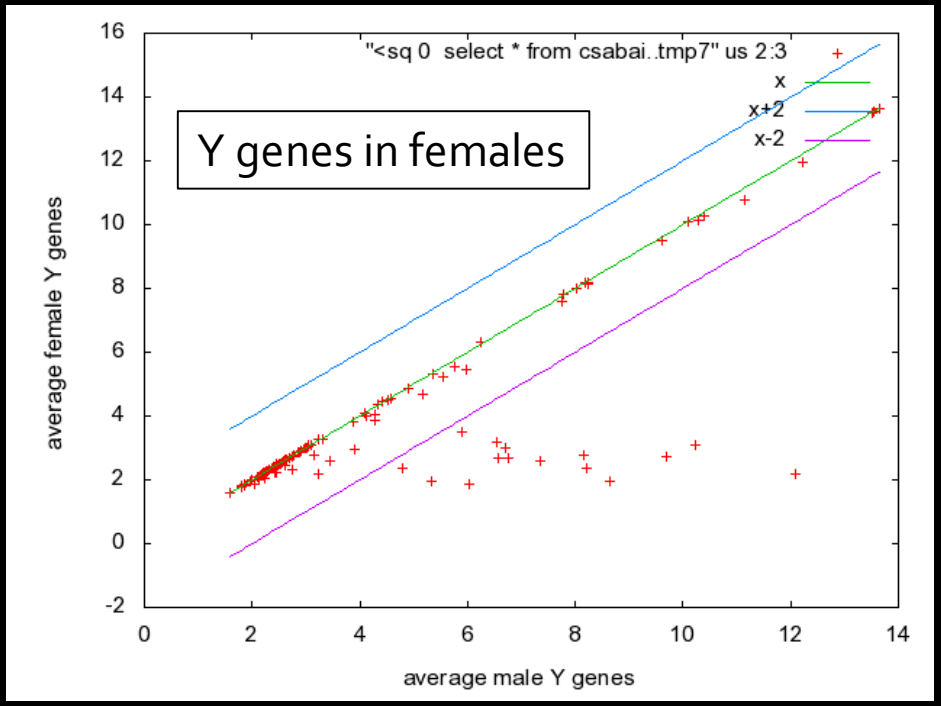


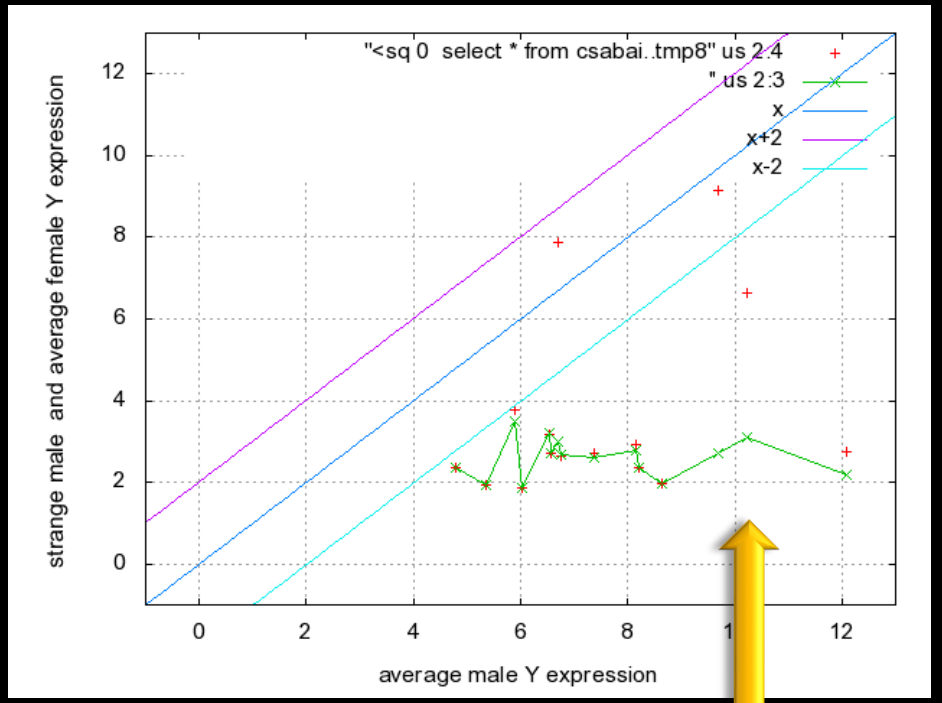
PCA – KEGG pathways (ribosome)

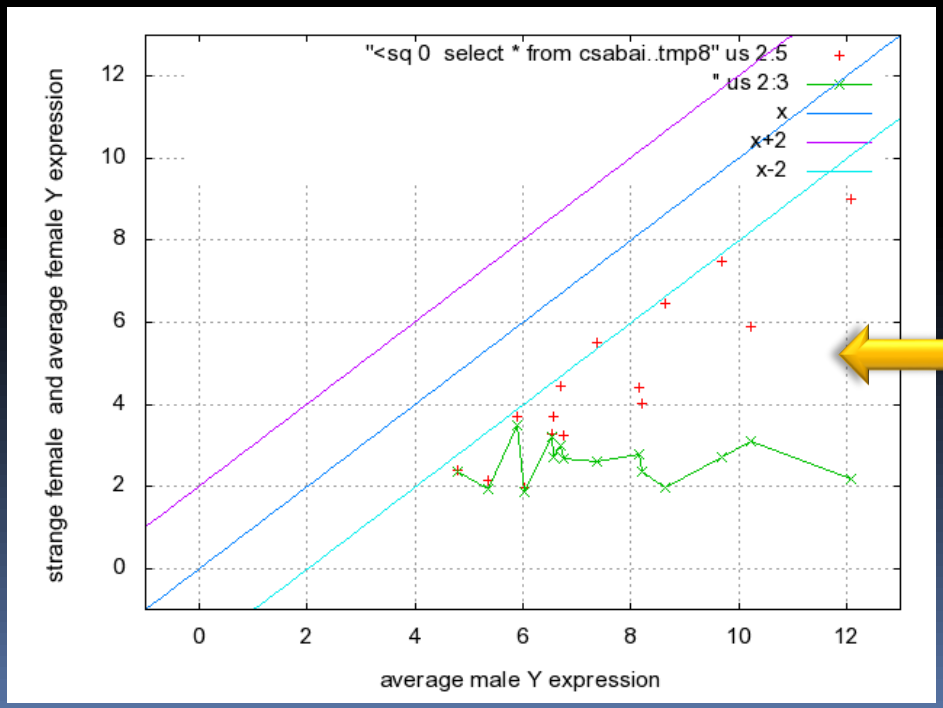


PCA – KEGG pathways (ribosome)





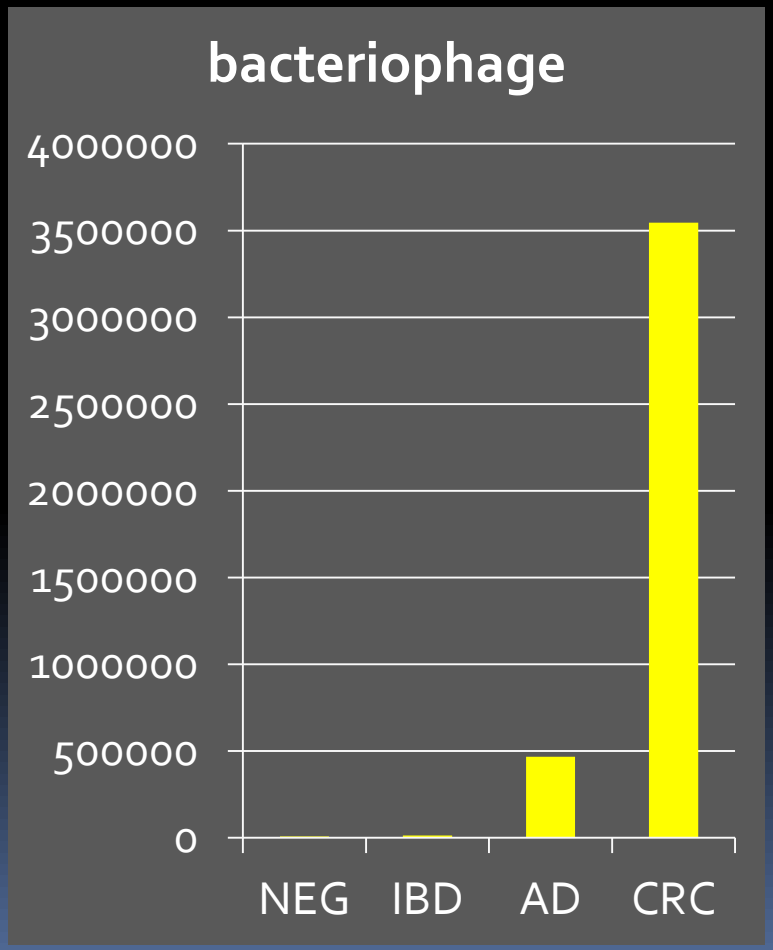


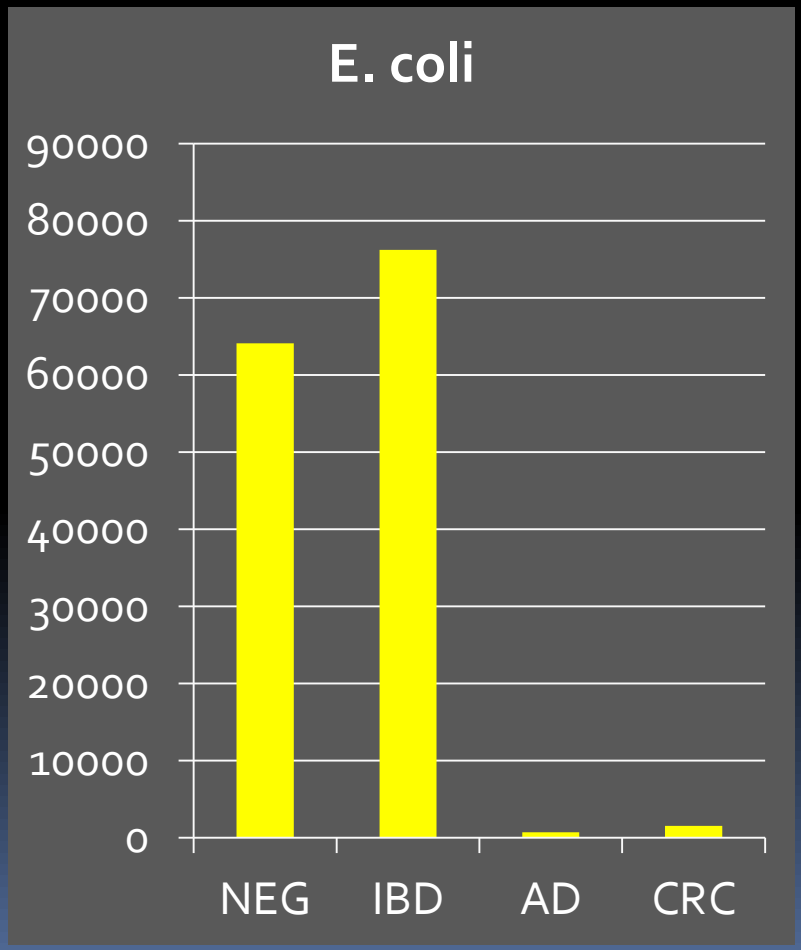


"strange male": CRC D, 78 yrs
"strange female": NEG, 43 yrs

Surprise in NGS study

- instead of marker genes – presence/composition of bacteria/phages





NEW DIRECTIONS

Source: Han Liang, Rice univ

Metagenomics

- Characterizing the biodiversity found on Earth, human metagenome
- The growing number of sequenced genomes enables us to interpret partial sequences obtained by direct sampling of specif environmental niches.
- Examples: ocean, acid mine site, soil, coral reefs, human microbiome which may vary according to the health status of the individual, sewage

THE METAGENOMICS PROCESS

Extract all DNA from microbial community in sampled environment

DETERMINE WHAT THE GENES ARE
(Sequence-based metagenomics)

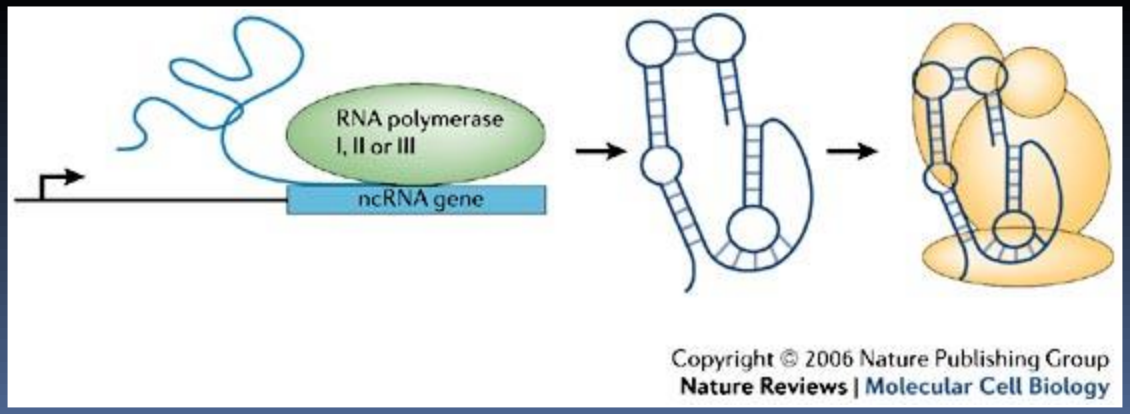
- Identify genes and metabolic pathways
- Compare to other communities
- and more…

DETERMINE WHAT THE GENES DO
(Function-based metagenomics)

- Screen to identify functions of interest, such as vitamin or antibiotic production
- Find the genes that code for functions of interest
- and more…

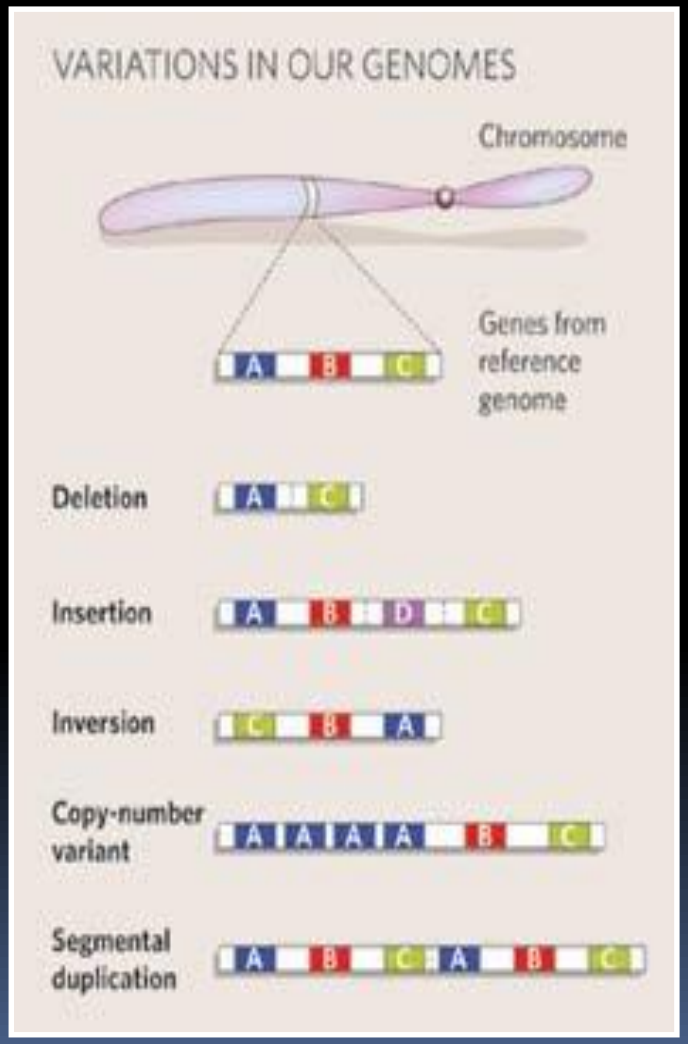
Discovering noncoding RNAs

- ncRNA presence in genome difficult to predict by computational methods with high certainty because the evolutionary diversity
- Detecting expression level changes that correlate with changes in environmental factors, with disease onset and progression, complex disease set or severity
- Enhance the annotation of sequenced genomes (impact of mutations more interpretable)



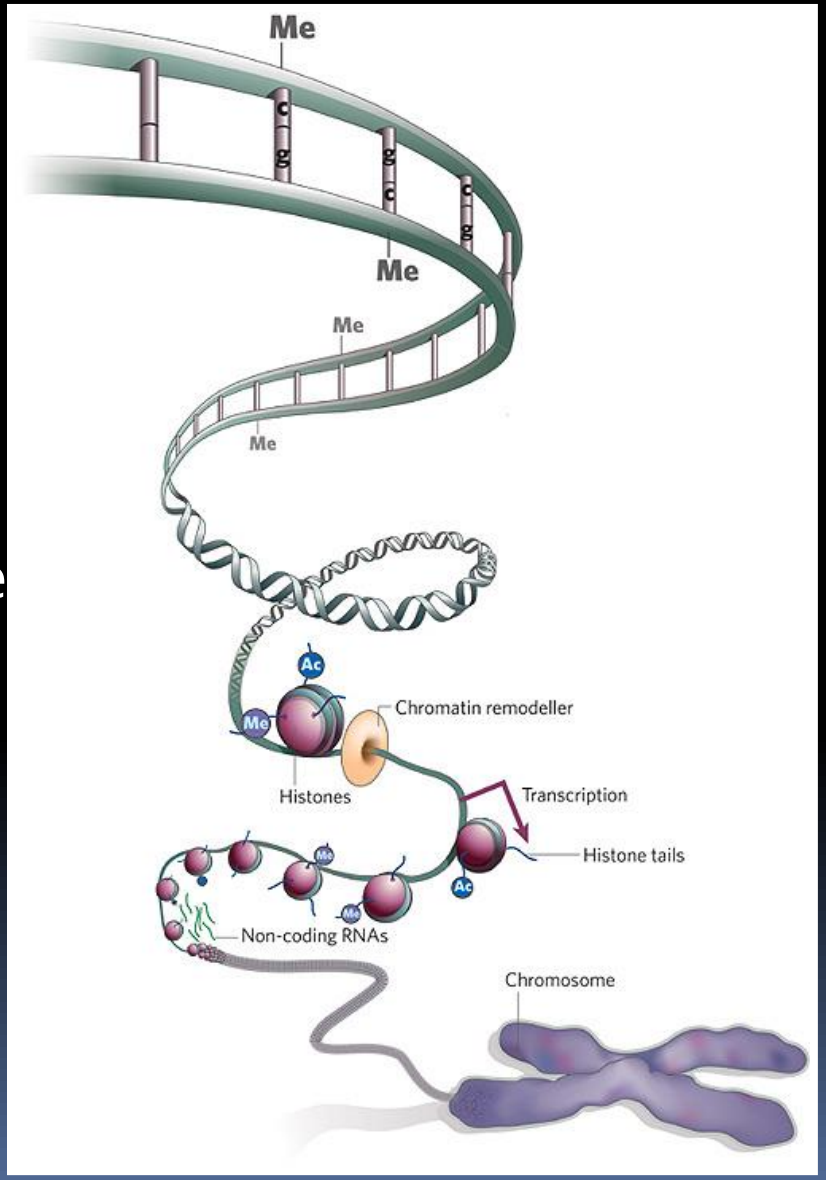
Defining variability in many human genomes

- Common variants have not yet completly explained complex disease genetics →rare alleles also contribute
- Also structural variants, large and small insertions and deletions
- Accelerating biomedical research
- 1000 Genome Project / COSMIC database

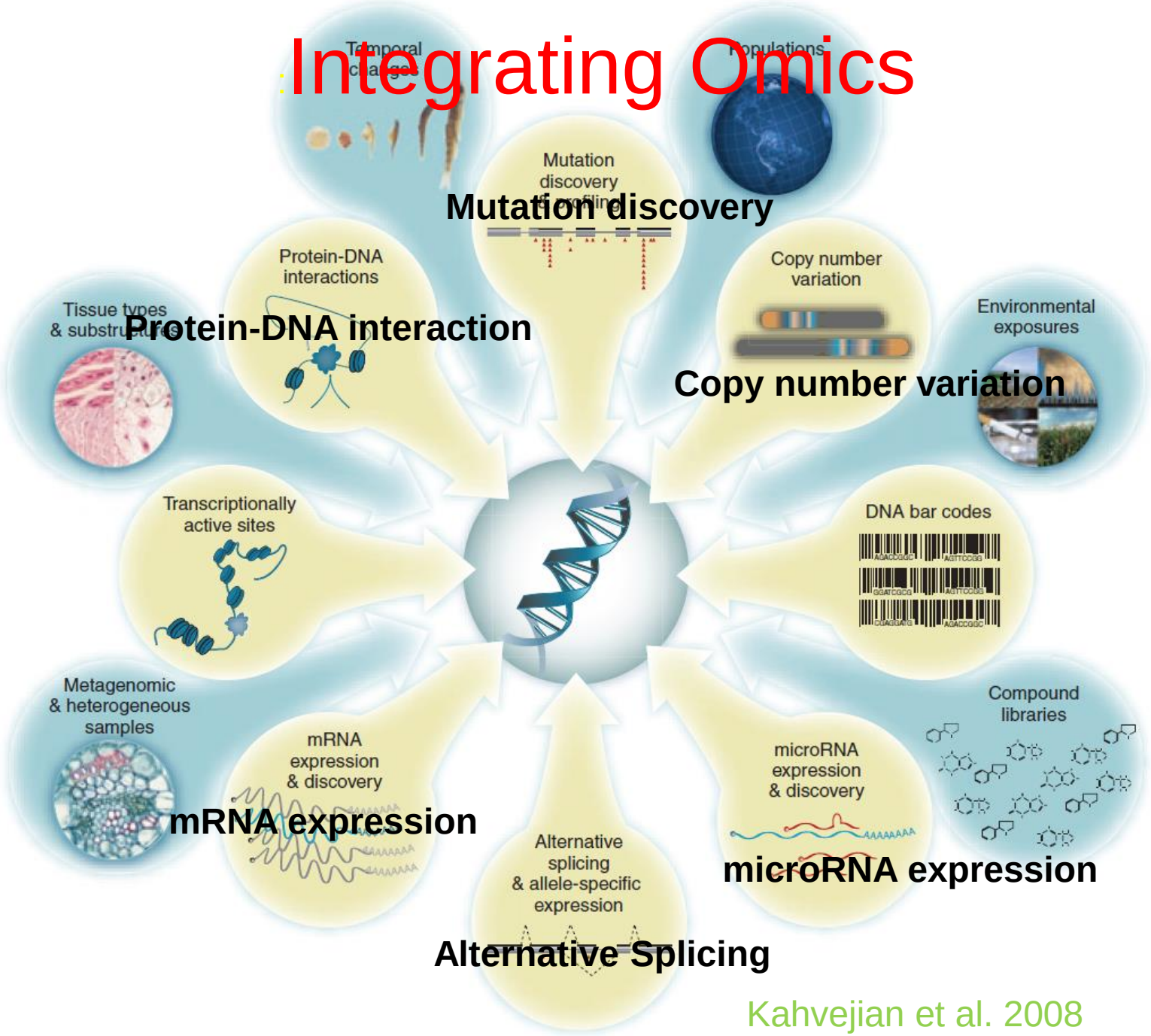


Epigenomic variation

- Enable of genome-wide patterns of methylation and how this patterns change through the course of an organism's development.
- Enhanced potential to combine the results of different experiments, correlative analyses of genome-wide methylation, histone binding patterns and gene expression, for example.
- ENCODE project!



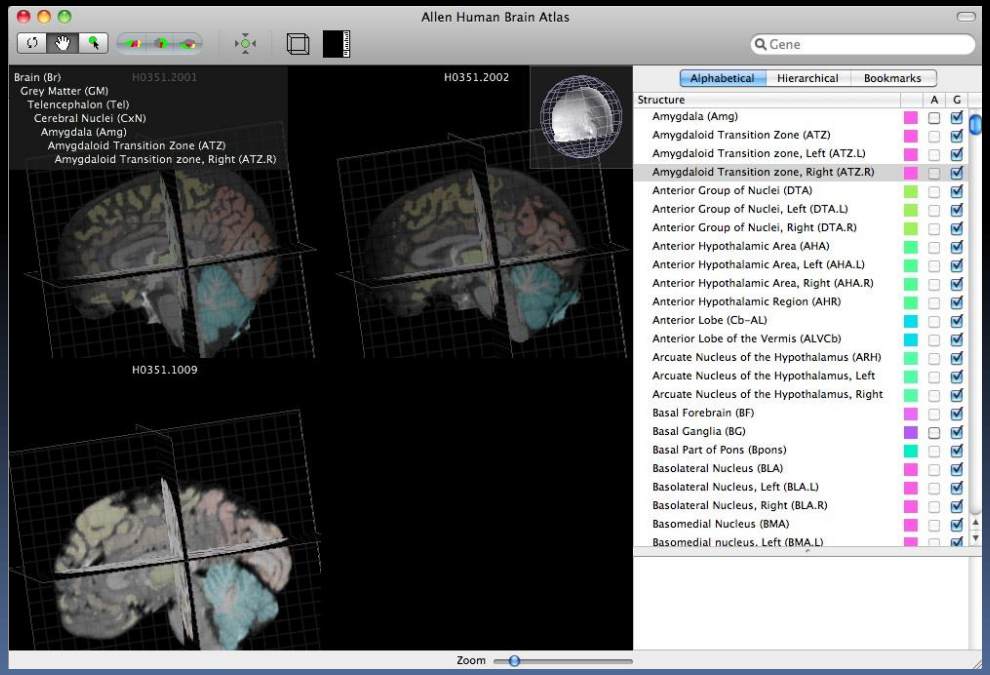
Integrating Omics



Kahvejian et al. 2008

NEUROINFORMATICS

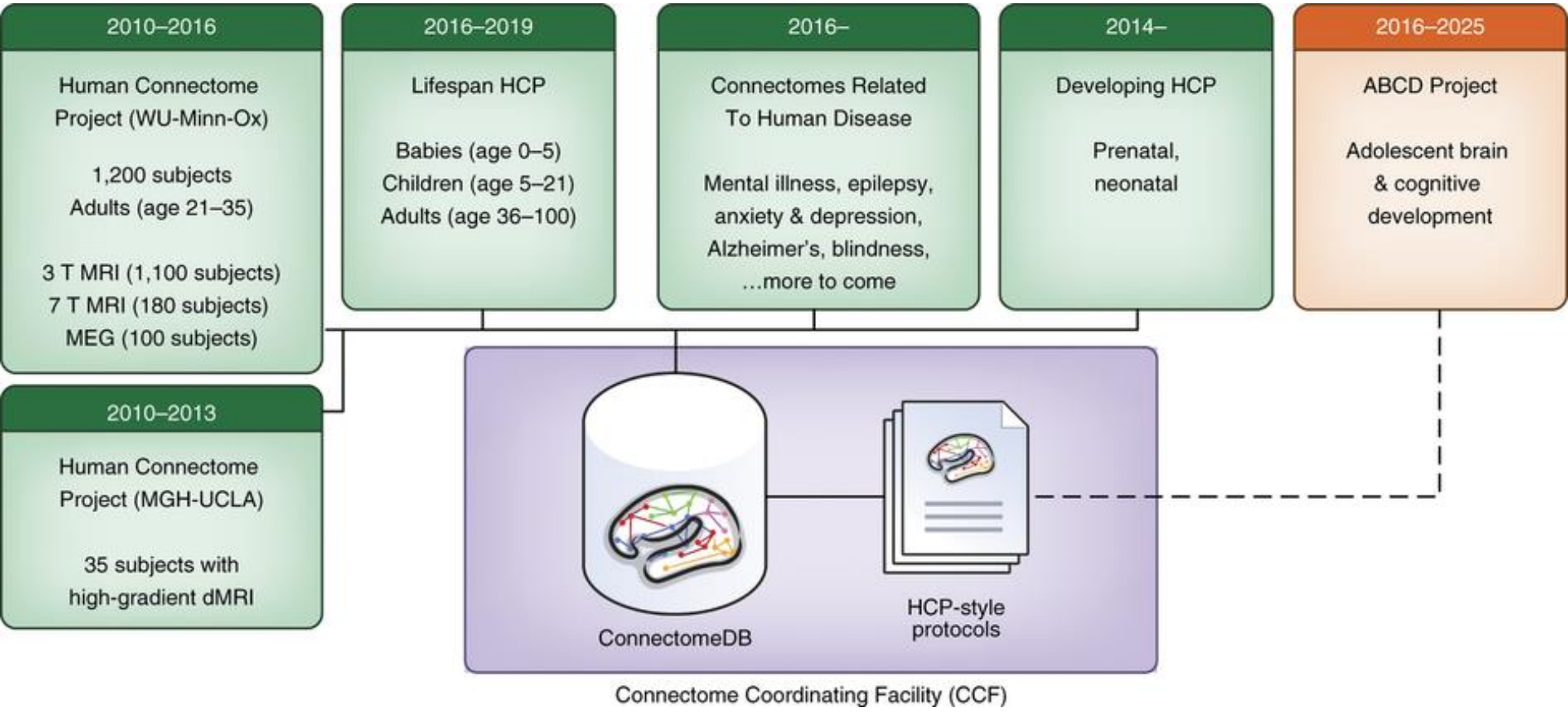
- EU: Horizon2020 – Human Brain Project (10 yrs, ~1BEur, Henry Markram, EPFL, Blue Brain)
- US: Human Brain Initiative (NIH)
  - Allen Brain Atlas – 2003-2013, mouse atlas ready



The Human Connectome Project's neuroimaging approach

•Matthew F Glasser, Stephen M Smith, Daniel S Marcus, Jesper L R Andersson, Edward J Auerbach, Timothy E J Behrens, Timothy S Coalson, Michael P Harms, Mark Jenkinson, Steen Moeller, Emma C Robinson, Stamatios N Sotiropoulos, Junqian Xu, Essa Yacoub, Kamil Ugurbil & David C Van Essen

Nature Neuroscience **19**, 1175–1187 (2016) doi:10.1038/nn.4361
Published online 26 August 2016



With these realities in mind, the HCP is distributing image data in three ways:

1. Via download through ConnectomeDB. HCP has made convenient data packages in 1 subject, 10, or 100 unrelated subject groups (by modality type), or all MEG Subjects to allow the user to "try out" the data without incurring a large download or data storage burden. The subjects in these set groups are unrelated to facilitate its use to investigators who want to analyze data without being concerned about family structure issues. The ConnectomeDB interface also allows the user to create their own subject groups of interest and download only the modalities necessary for analysis of that group.

2. "Connectome in a Box" (CinaB). Users can choose a complete 900 Subjects Release dataset, a starter dataset (~100 unrelated subjects), or MEG subjects dataset preloaded onto 8TB hard drives, that can be ordered and shipped to you for the cost of the hard drives + shipping. S900 Connectome in a Box can be ordered now and should be available for shipment by early 2016.

3. Amazon S3. Complete 900 Release data will be accessible on the cloud in the same organization as Connectome in a Box by end of 2015. See Accessing HCP Data on the Cloud for details.

For group-average functional connectivity data, because of their large size, we

HCP Data Sizes (per Subject)

Session	Format	.zip File Size
Structural	Unprocessed	71 MB
	Preprocessed	1 GB
Resting State fMRI (each of 2 runs)	Unprocessed	2 GB
	Preprocessed	5.8 GB
	FIX (compact)	3.9 GB
	FIX_extended	4.2 GB
Task fMRI (avg per Task)	Unprocessed	490 MB
	Preprocessed	1.9 GB
	Analyzed	400 MB
All 7 Tasks	Unprocessed	3.4 GB
	Preprocessed	13.1 GB
	Analyzed	2.8 GB
Diffusion	Unprocessed	2.6 GB
	Preprocessed	1.2 GB
Group-Average on U100 and R440	Additionally Processed	200 MB
Group-Average "dense" connectomes (each of 2)	Additionally Processed	33 GB
MEG	Unprocessed	9 GB
	Preprocessed	14 GB
	Source-level	35 GB
Total MEG (95 Subjects)	Unprocessed	780 GB
	Preprocessed	1.1 TB
	Both + Source	4.7 TB
Total (per Subject, MR only)	Unprocessed	10 GB
	Preprocessed	40 GB
	Both	50 GB
	Both+Analyzed	53 GB
Total (100 Subjects, MR only)	Unprocessed	1.2 TB
	Preprocessed	4.2 TB
	Both+Analyzed	5.7 TB
Total (All HCP scan datasets from 897 Subjects)	**Unprocessed**	**10.9 TB**
	Preprocessed	**36.7 TB**
	Both+Analyzed	**52.3 TB**

Blue Brain

The Blue Brain Project was launched by the Brain Mind Institute, EPFL, Switzerland and IBM, USA in May 2005, now over 4.8M WWW pages.

The EPFL Blue Gene is the 8th fastest supercomputer in the world.

Can simulate about 100M minimal compartment neurons or 10-50000 multi-compartmental neurons, with 10^3-10^4 x more synapses. Next generation BG will simulate $>10^9$ neurons with significant complexity.

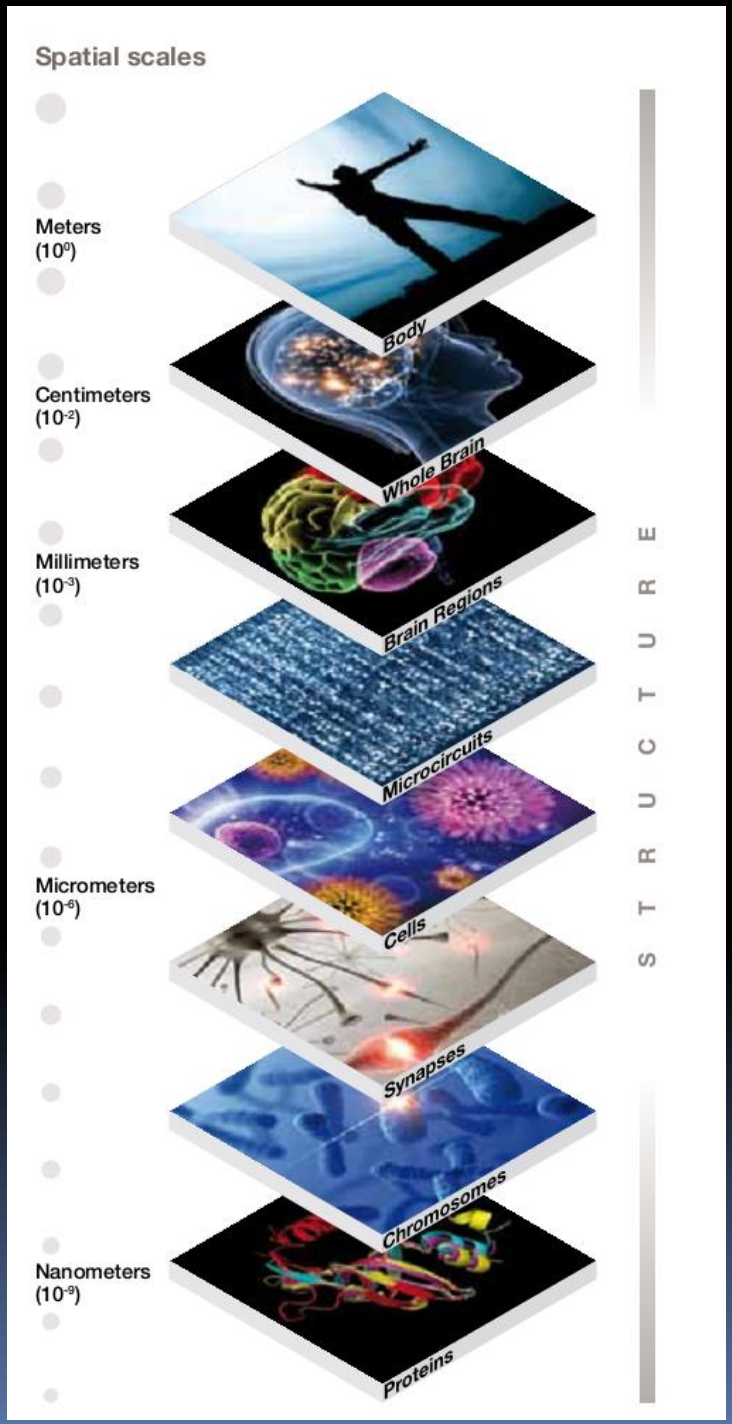
1. The Blue Synapse: A molecular level model of a single synapse.

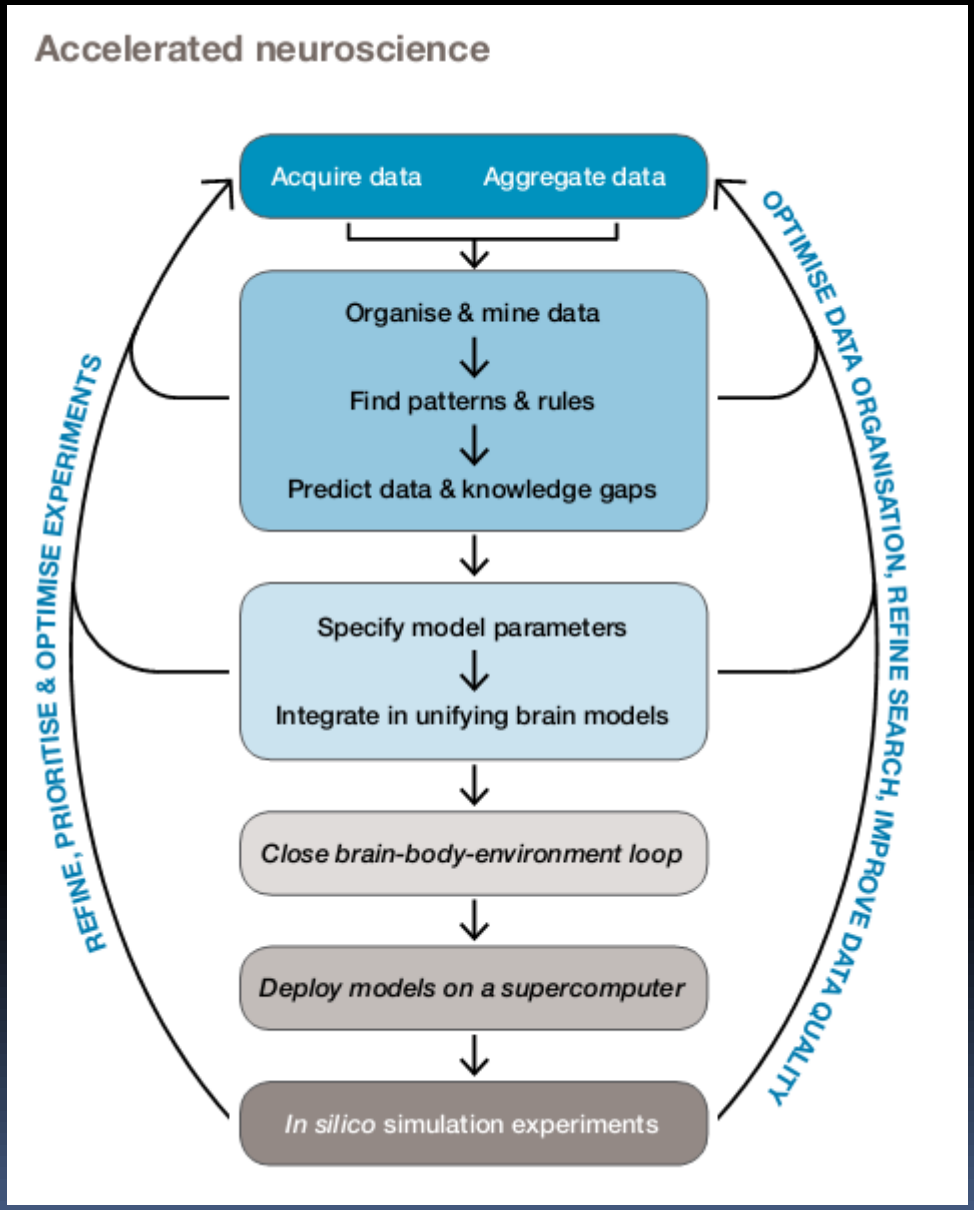
2. The Blue Neuron: A molecular level model of a single neuron.

3. The Blue Column: A cellular level model of the Neocortical column with 10K neurons, later 50K, 100M connections.

4. The Blue Neocortex: A simplified Blue Column will be duplicated to produce Neocortical regions and eventually and entire Neocortex.

5. The Blue Brain Project will also build models of other Cortical and Subcortical models of the brain, and sensory + motor organs.


Spatial scales
Meters
(10^0)
Body
Centimeters
(10^-2)
Whole Brain
Millimeters
(10^-3)
Brain Regions
Microcircuits
Micrometers
(10^-6)
Cells
Synapses
Chromosomes
Nanometers
(10^-9)
Proteins
STRUCTURE
Time scales
Years
(10^7)
Development & Aging
Days
(10^5)
Hours
(10^3)
Behaviour
Minutes
(10^2)
Learning
Synaptic Plasticity
Seconds
(10^0)
Milliseconds
(10^-3)
Metabolism
Microseconds
(10^-6)
Action Potential
Nanoseconds
(10^-9)
Vesicle release
Picoseconds
(10^-12)
Molecular dynamics
FUNCTION

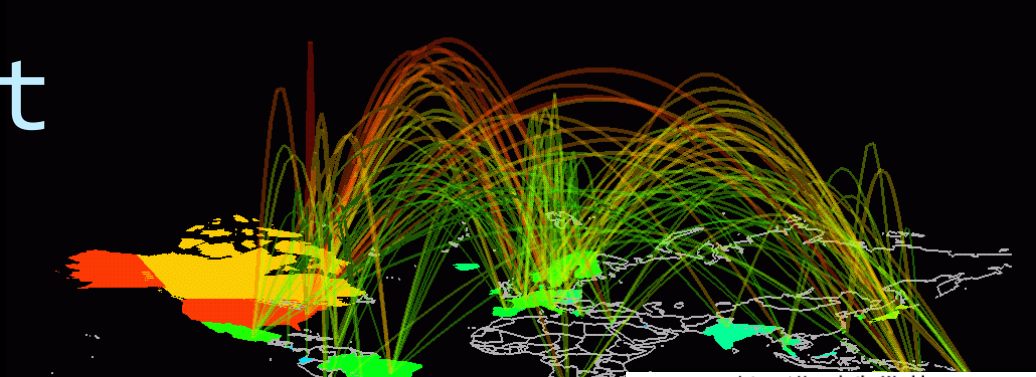

Accelerated neuroscience
Acquire data
Aggregate data
Organise & mine data
Find patterns & rules
Predict data & knowledge gaps
Specify model parameters
Integrate in unifying brain models
Close brain-body-environment loop
Deploy models on a supercomputer
In silico simulation experiments
REFINE, PRIORITISE & OPTIMISE EXPERIMENTS
OPTIMISE DATA ORGANISATION, REFINE SEARCH, IMPROVE DATA QUALITY

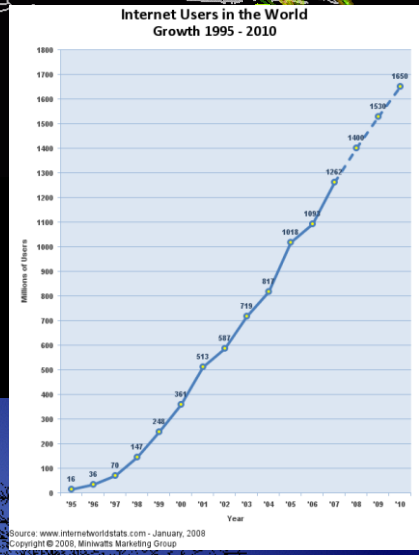
Manmade complex systems

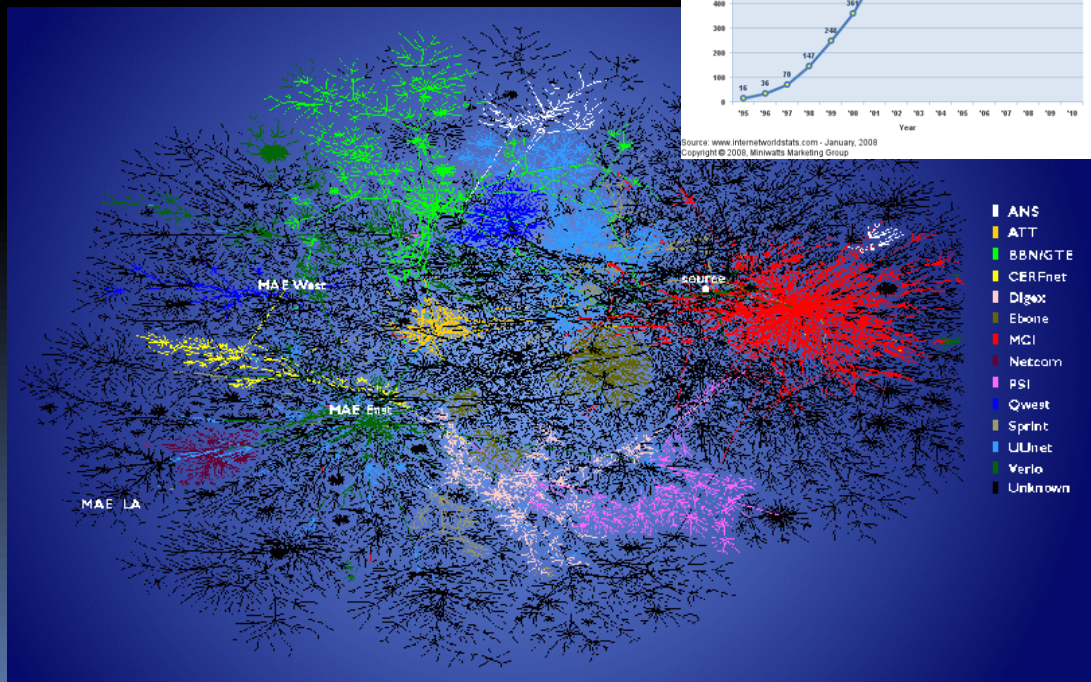
COMMUNICATION- SOCIAL- AND FINANCIAL NETWORKS

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 …

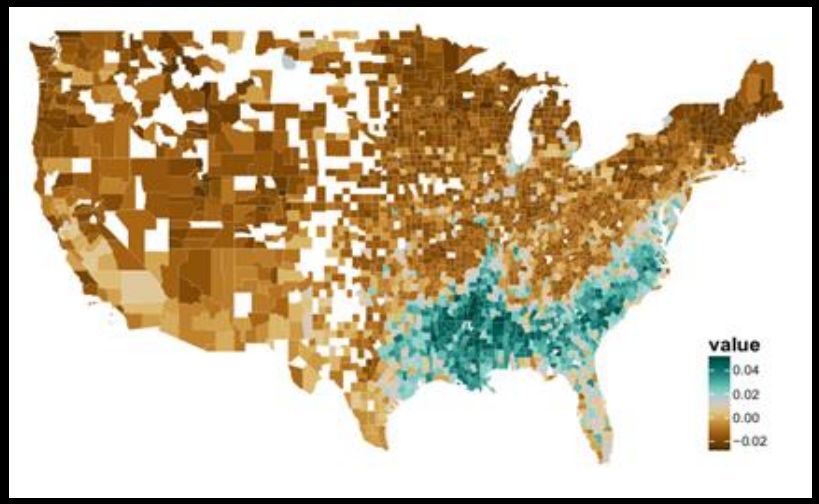


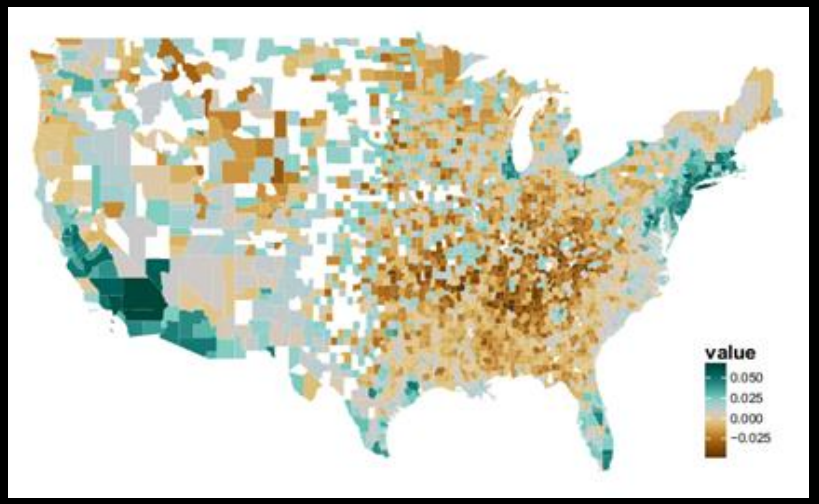


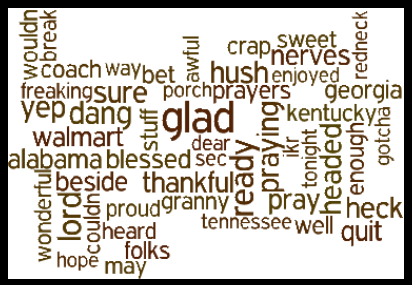


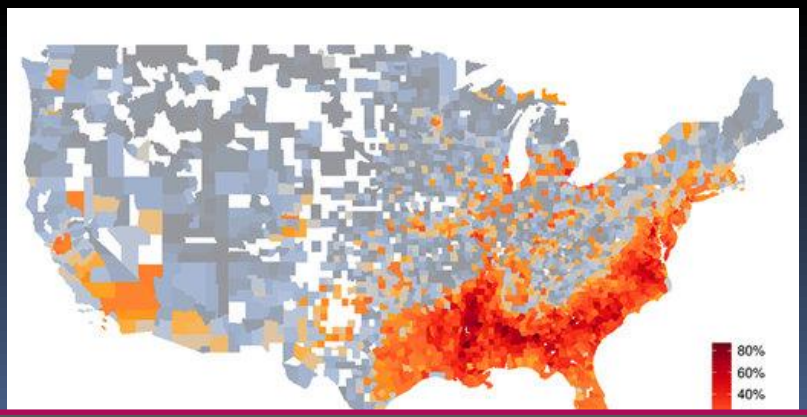
Map of society: TwitterDB

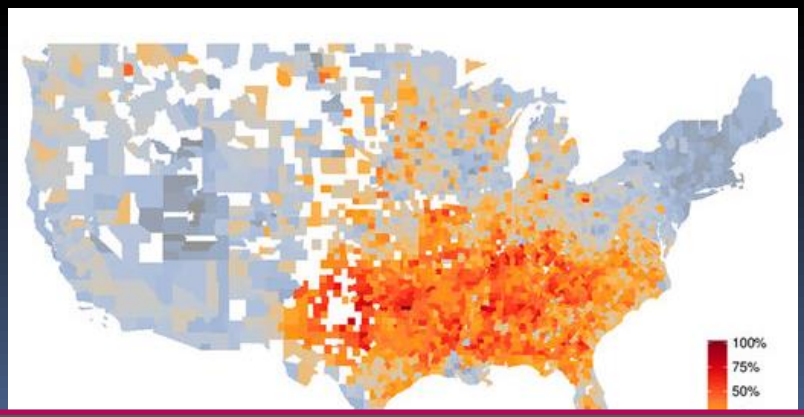
Principal dimensions: race, religion, urbanization







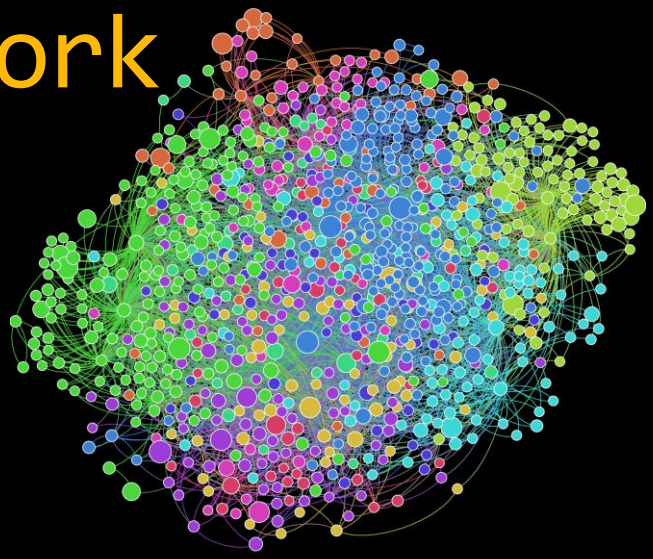




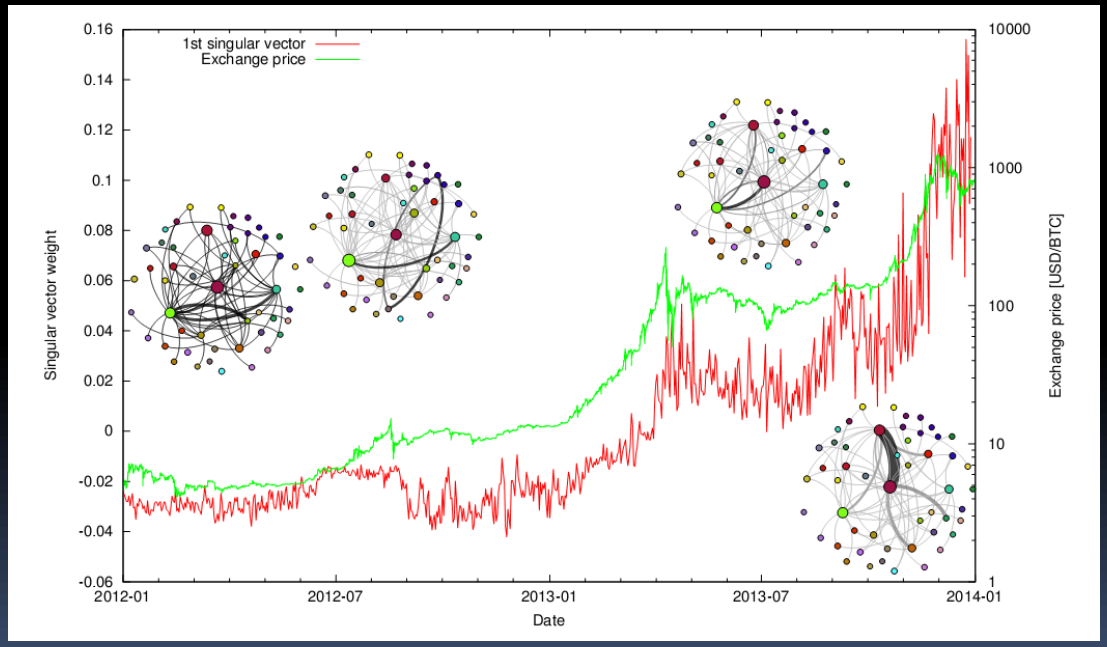
E Bokányi , D Kondor, L Dobos, T Sebők, J Stéger I Csabai, G Vattay Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States , Palgrave Communications 2, 16010 (2016)

Bitcoin financial network

Map of economy



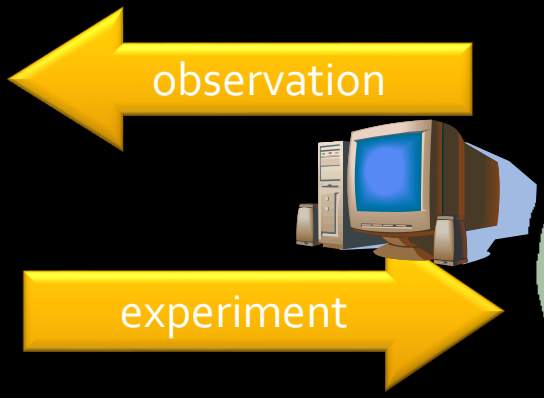
- All (50M) transactions are logged, public
- Dynamic evolving network
- Database
- Dimension reduction (graph non-negative factorization)



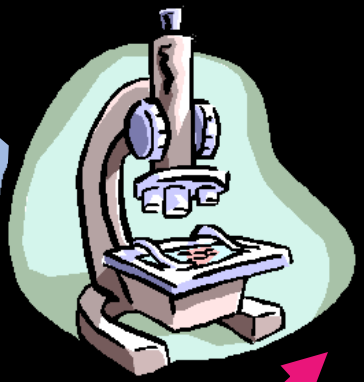
Strong random correlations in networks of heterogeneous agents; I Kondor, I Csabai, G Papp, E Mones, G Czimbalmos, MC Sándor Journal of Economic Interaction and Coordination 9 (2), 203-232 (2014)

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)

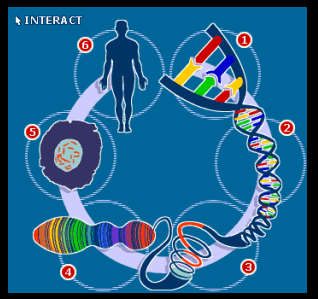
Big Data–Big Challenges–Big opportunities

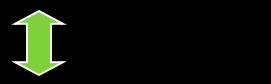


instruments

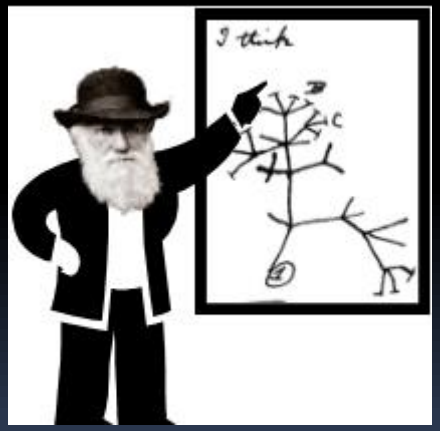


reality





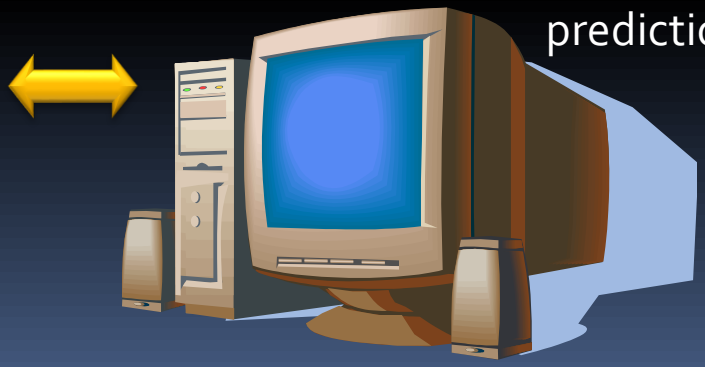
models

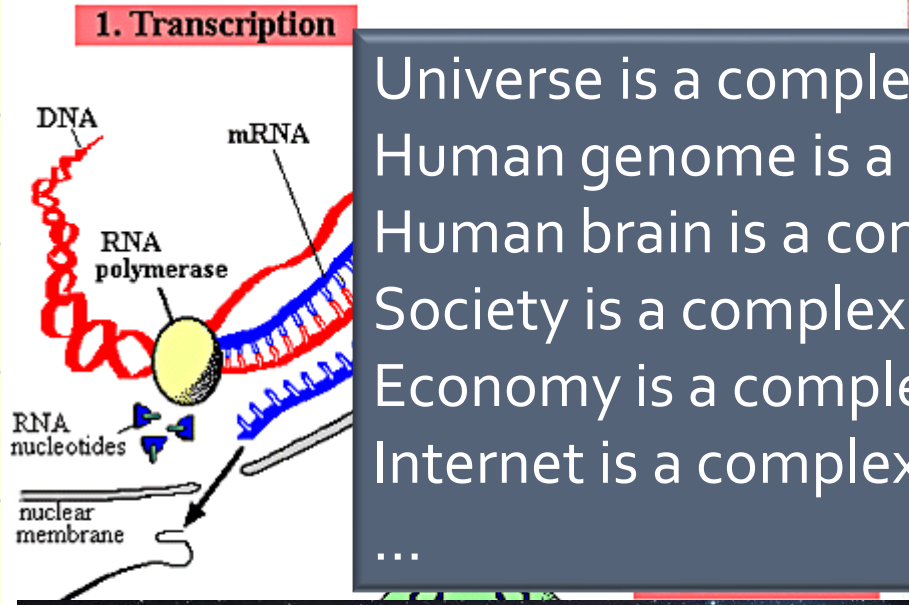


test

virtual reality

predictions





Universe is a complex system
Human genome is a complex system
Human brain 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 "

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

- **A:** NASA scientists
- **B:** Harvard doctors
- **C:** Sorcerers
- **D:** YOU!

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/