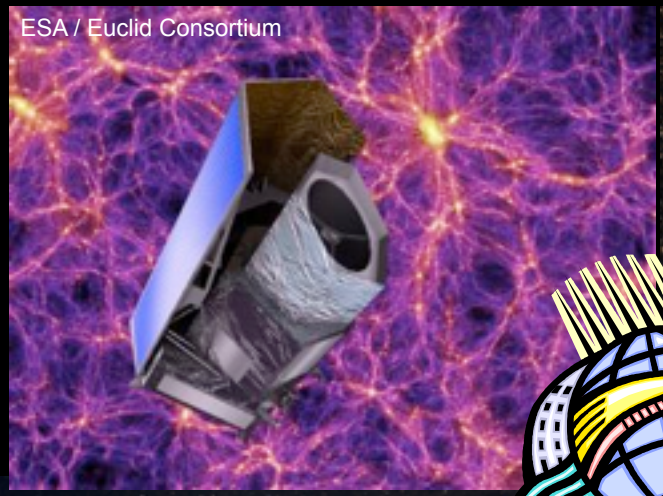
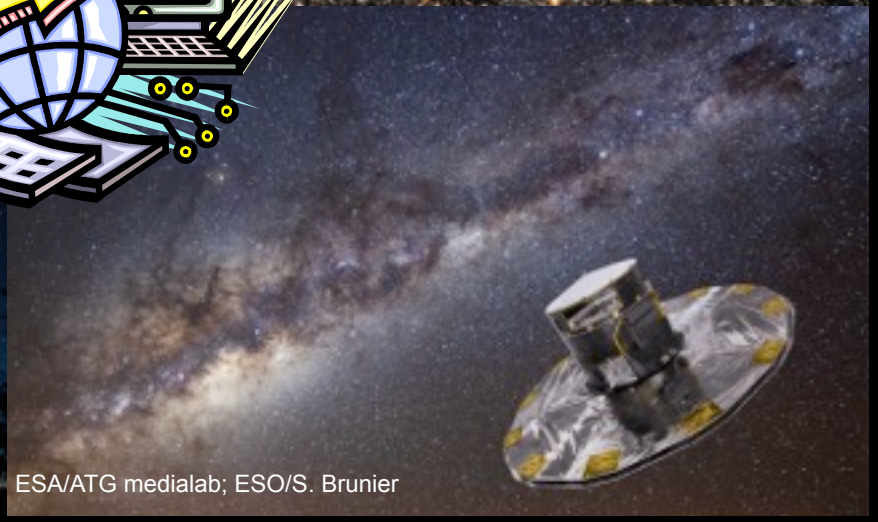
Big-Data as a Challenge for Astrophysics





Volker Beckmann
Institut National de Physique Nucléaire et de Physique des Particules
CNRS; beckmann@in2p3.fr

Outline

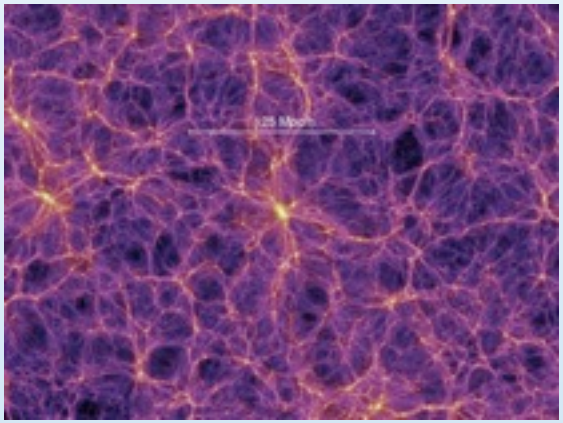
- BigData: Volume, Velocity, Variety, Veracity and Value
- Why does astrophysics need Big-Data?
- Science drivers
- Specific challenges in astrophysics
- Current and future experiments
- Outlook and conclusions

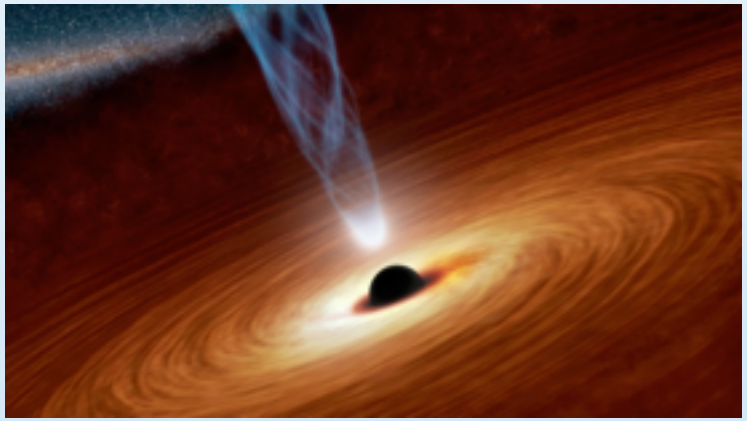
V. Beckmann

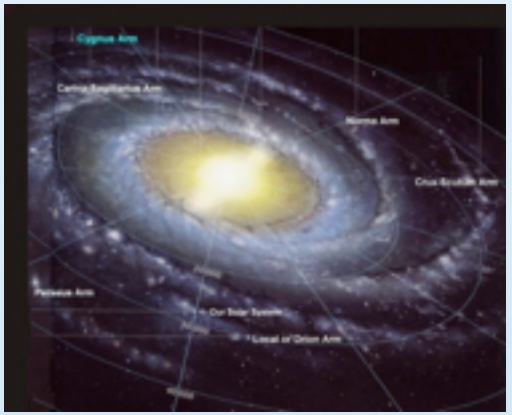
Key questions

Some key questions:

- Universe: origin, structure and evolution
- Objects in the universe: how's matter/energy converted, accretion and emission
- Milky Way: origin, structure and evolution
- Physics: dark matter, dark energy, accretion + emission processes, neutrinos, gravitational waves …

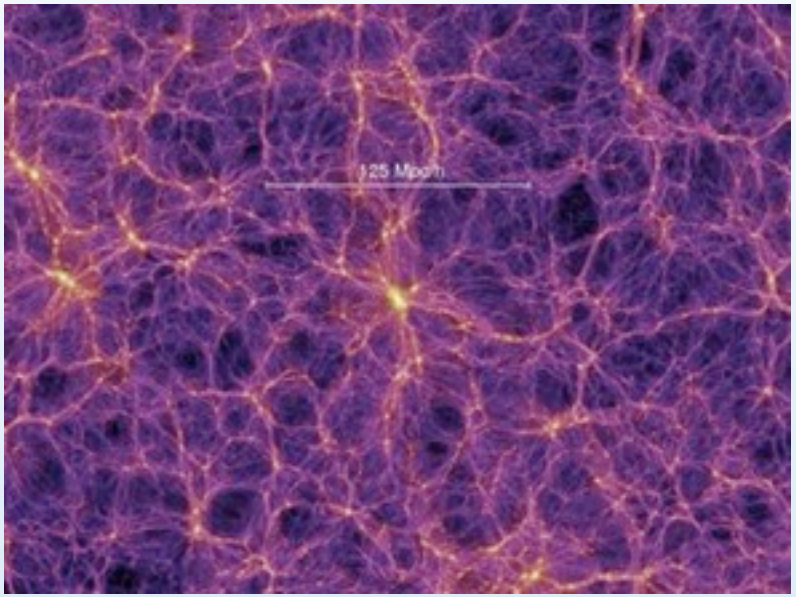






Universe Structure & Evolution

We can simulate the evolution and structure of the Universe to some extend



Springel et al. 2005
Millenium simulation
10 billion particles
20 million galaxies
Cube of 2 billion light years on a side
25 TByte output

- Using start-conditions, how does matter distribute?
- How often and how do galaxies merge?
- What are the resulting large scale structures?
- Where does star formation take place?

Universe Structure & Evolution

- How early did galaxies form?
- How did they merge?
- Super massive black holes
- Dark matter – where does it reside?
- Dark energy – does it evolve with time?

simulation → observation

- Deep images (deep = far into the past)
- Determine distances of galaxies (spectra)
- Combine information from different wavelengths
- > 100 million galaxies (better 1 billion: 1% of all galaxies in the observable Universe)

V. Beckmann

What's special?

- Old science, but small community
- 20,000 astronomers world wide (1.7 million researchers in EU)
- Broad science
- many colleagues work in many fields, across many techniques/wavelengths, everybody needs access to all data
- Strong networks
- Large experiments; e.g. Euclid >1000 colleagues, remote experiment sites
- Several data sets needed to identify objects → multi-messenger science
- Standards, open archives, open source software

V. Beckmann

One file format

- Flexible Image Transport System (FITS) introduced in 1980s
- Used for the transport, analysis, and archival storage of scientific data sets
- Multi-dimensional arrays: 1D spectra, 2D images, 3D+ data cubes
- Tables containing rows and columns of information
- Header (ASCII) keywords provide descriptive information about the data

FITS file format

fv: Summary of isgri_sky_ima.fits in /Users/volker/Documents/science/data/obs/test_cons/scw/08560...

File Edit Tools Help

Index	Extension	Type	Dimension	View
0	Primary	Image	0	Header | Image | Table
1	GROUPING	Binary	18 cols X 8 rows	Header | Hist | Plot | All | Select
2	ISGR-SKY.-IMA	Image	400 X 400	Header | Image | Table
3	ISGR-SKY.-IMA	Image	400 X 400	Header | Image | Table
4	ISGR-SKY.-IMA	Image	400 X 400	Header | Image | Table
5	ISGR-SKY.-IMA	Image	400 X 400	Header | Image | Table
6	ISGR-SKY.-IMA	Image	400 X 400	Header | Image | Table
7	ISGR-SKY.-IMA	Image	400 X 400	Header | Image | Table
8	ISGR-SKY.-IMA	Image	400 X 400	Header | Image | Table
9	ISGR-SKY.-IMA	Image	400 X 400	Header | Image | Table

FITS The Astronomical Image and Table Format
Flexible Image Transport System

V. Beckmann

FITS file format

```
XTENSION= 'IMAGE   '           / Image extension
BITPIX  =                   -32 / IEEE 32-bit floating point values
NAXIS   =                     2 / Number of data axes
NAXIS1  =                   400 / Length of data axis 1
NAXIS2  =                   400 / Length of data axis 2
PCOUNT  =                     0 / required keyword; must = 0
GCOUNT  =                     1 / required keyword; must = 1
EXTNAME = 'ISGR-SKY.-IMA'       / Extension name
EXTREL  = '6.5     '            / ISDC release number
BASETYPE= 'DAL_ARRAY'           / Data Access Layer base type
TELESCOP= 'INTEGRAL'            / Telescope or mission name
ORIGIN  = 'ISDC    '            / Origin of FITS file
INSTRUME= 'IBIS    '            / Instrument name
DETNAM  = 'ISGRI   '            / Name of the detector layer
ISDCLEVL= 'IMA     '            / ISDC level of data processing
CREATOR = 'ii_skyimage 5.4'     / Executable which created or modified this data
CONFIGUR= 'osa_8.0_2009-08-24T07:45:37' / Software system configuration
DATE    = '2009-12-15T17:48:07' / Creation or modification date
MJDREF  =                51544. / Modified Julian Date of time origin
TIMESYS = 'TT      '            / Time frame system
TIMEUNIT= 'd       '            / Time unit
TIMEREF = 'LOCAL   '            / Time reference frame
REVOL   =                   856 / Revolution number
SWID    = '085600190010'        / Science Window identifier
SW_TYPE = 'POINTING'            / Type of the Science Window
SWBOUND = 'OTF     '            / Reason for Science Window ending
OBTSTART= '0000023168874446848O' / OBT of the start of the Science Window
OBTEND  = '0000023169248998195 2' / OBT of the end of the Science Window
TSTART  =    3577.42810698031 / Start time of the Science Window
TSTOP   =    3577.46944960259 / End time of the Science Window
TFIRST  =    3577.42872901587 / Time of the first data element
TLAST   =     3577.4694495517 / Time of the last data element
TELAPSE =                3518. / [s] Total elapsed time of the data
ONTIME  =    3516.37097167969 / [s] Sum of good time intervals
DEADC   =   0.667853708897771 / Dead-time correction factor
EXPOSURE=    2092.22927527603 / [s] Effective exposure time
BSCALE  =                   1 / Real value=value*BSCALE + BZERO
BZERO   =                   0 / Offset applied to true pixel values
BUNIT   = 'counts/sec'          / Unit for pixel values
```

V. Beckmann

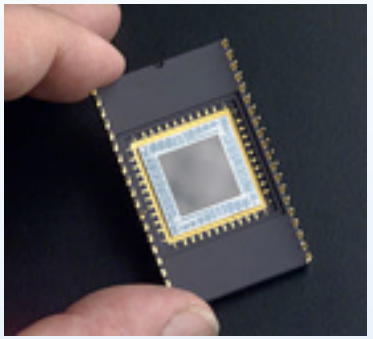
One file format

- Used for: images, spectra, light curves, tables, data cubes …
- Used in: space-based and ground-based astrophysics, across all disciplines
- Standards for key words, header, coordinate systems, …
- Fits i/o libraries
- tools to visualize and manipulate fits files
- http://fits.gsfc.nasa.gov/
- BigData: Volume, Velocity, ~~Variety~~, Veracity and Value

Big-Data before Big-Data

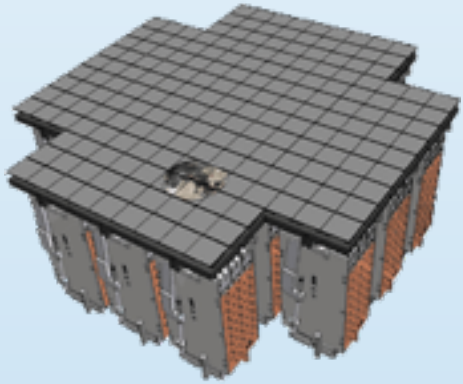
- *Bonner Durchmusterung* (1859-1903)
- Catalogue all stars in the sky down to apparent brightness V=10^{mag} on the northern sky (40 times fainter than naked eye)
- 7.7 cm telescope
- Position (± 0.002° = 7”)
- brightness (±10%)
- 325 000 stars

Data rate evolution



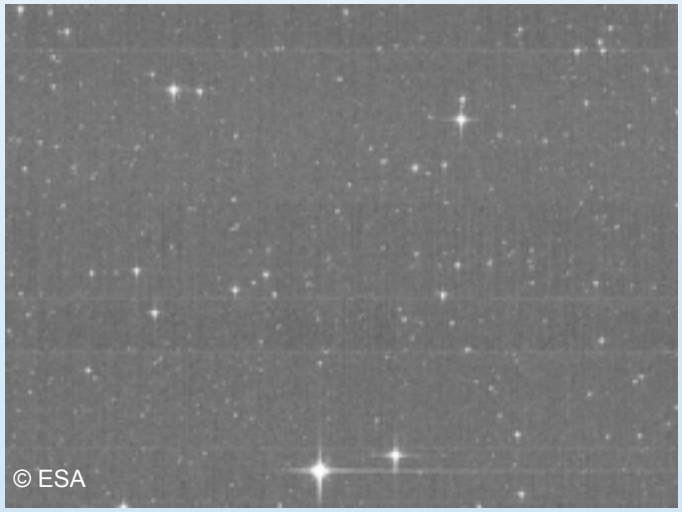
- Space based data limited by bandwidth
- 100 Mbps max (X-band), but see HTS
- INTEGRAL (2002): 1.2 Gbyte/day
- Hubble space telescope (1990): 15 Gbyte/day
- Gaia (2013), Euclid (2021): ~50 Gbyte/day

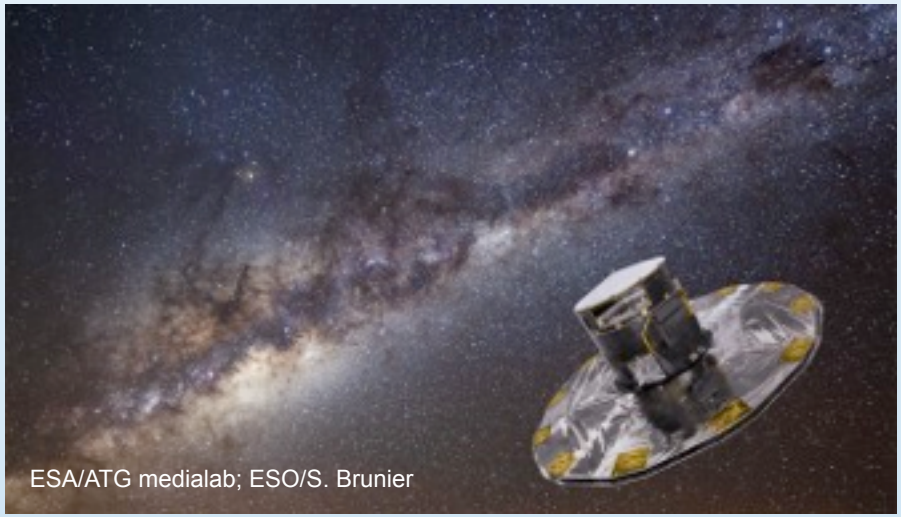
- Ground based: fast increase through fast read-out systems, multiple charge-coupled devices (CCDs)
- 1990s: 1 Mbyte / CCD frame
- LSST (2022): 3 Gbyte / exposure (15s)



Big-Data today: GAIA

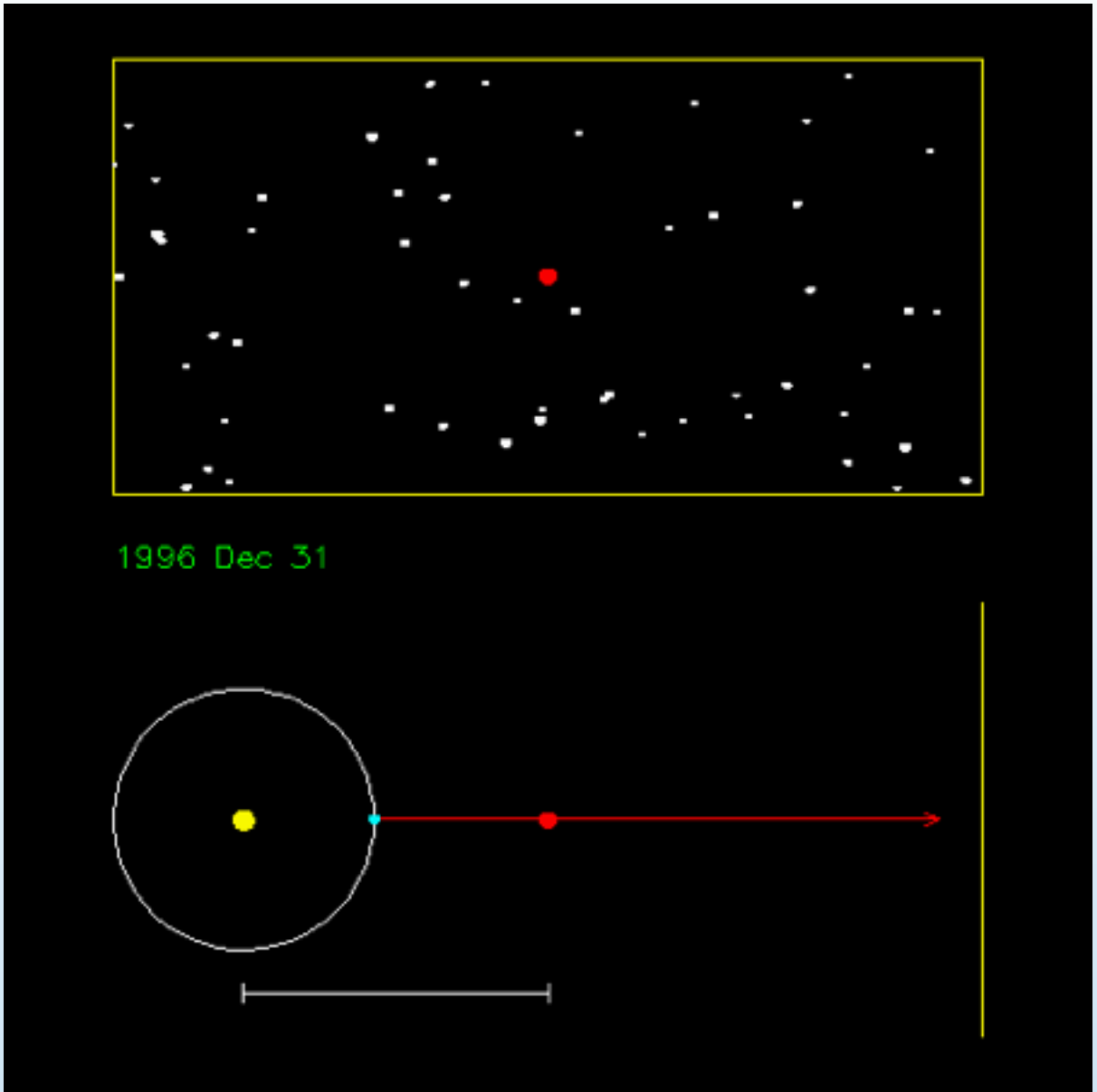
- GAIA satellite mission
- Map the milky way in 3D
- Stellar physics
- Dark matter
- Extrasolar planets
- 50 Gbyte/day; 1 Pbyte total data products





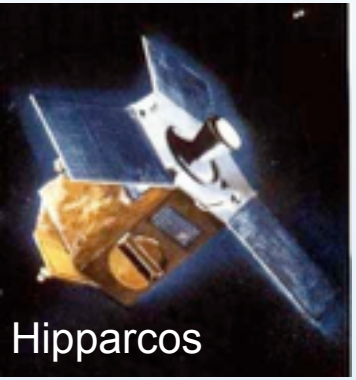
V. Beckmann

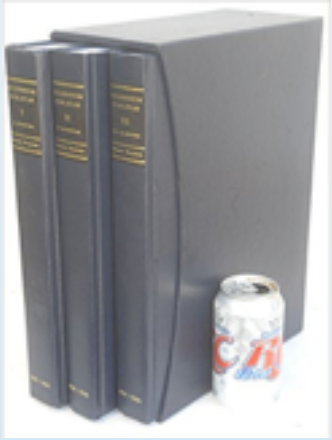
Big-Data today: GAIA



Distance determination of stars through parallax measurement

Big-Data today: GAIA


Hipparcos

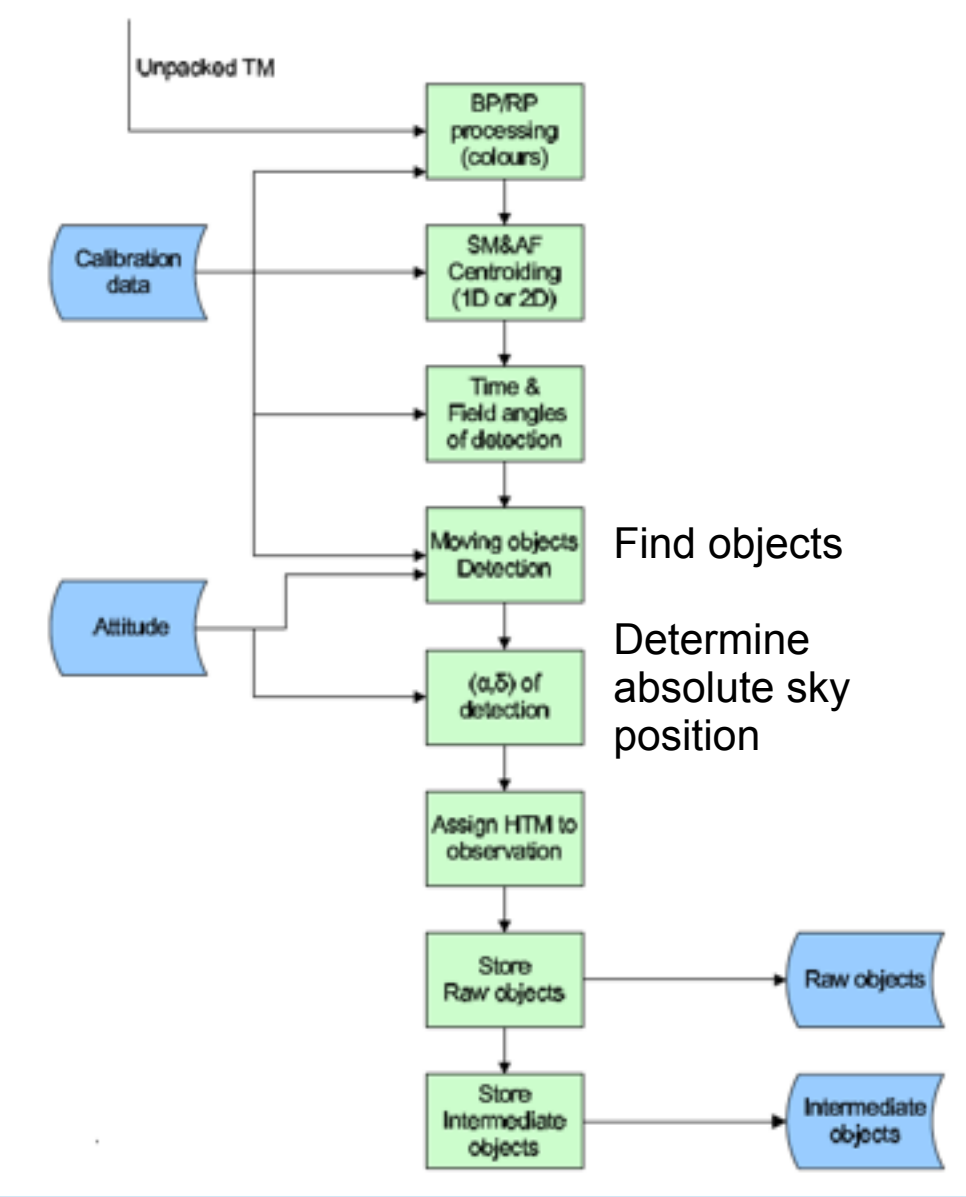


	HIPPARCOS (1989)	Gaia (2013)	
Magnitude limit	12	20 - 21	
Completeness limit	7.3 - 9	20	
Number of objects	120 000	35×10^6	$V < 15$
		350×10^6	$V < 18$
		1.3×10^9	$V < 20$
Astrometric accuracy	1 mas ($V < 9$)	7 μas	$V < 12$
	1-3 mas($V > 9$)	25 μas	$V = 15$
		300 μas	$V = 20$
$\sigma_\pi/\pi < 1\%$	150 stars	11×10^6 stars	
$\sigma_\pi/\pi < 5\%$	6,200 stars	77×10^6 stars	
$\sigma_\pi/\pi < 10\%$	21,000 stars	150×10^6 stars	
Radial velocity	–	$2 - 10$ km s^{-1}	$V < 17$
Spectro-photometry	–	$\simeq$ 25-colour	$V < 20$
Low resolution spectroscopy	–	R = 11,500	$V < 16 - 17$

V. Beckmann

Big-Data today: GAIA

100 Gbytes of image/spectra data per day



Find objects

Determine absolute sky position

V. Beckmann

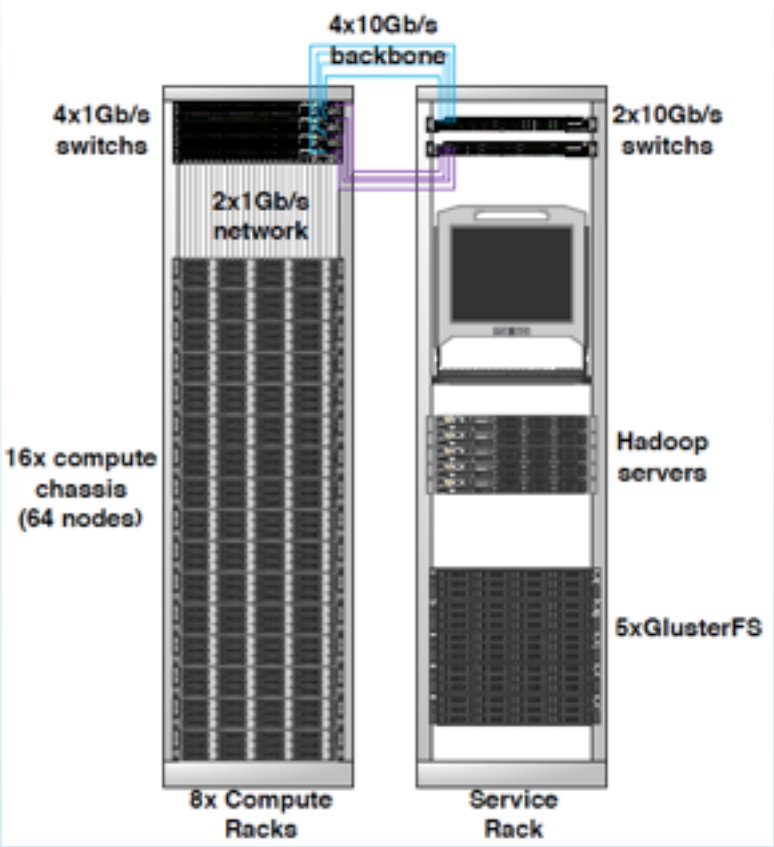
Big-Data today: GAIA

- Retrieve daily input data
- Launch daily processing workflow
- Send the results back to ESA
- Receive cycle processing data (every 6 months)
- Launch cycle processing
- Send the results back to ESA

Chosen solution: Hadoop

- Distributed file system for vast amount of data (HDFS)
- Software framework for distributed processing of large datasets (MapReduce)

Hardware:
1024 cores
10 Gbit/s network
1.5 Gbyte/core memory

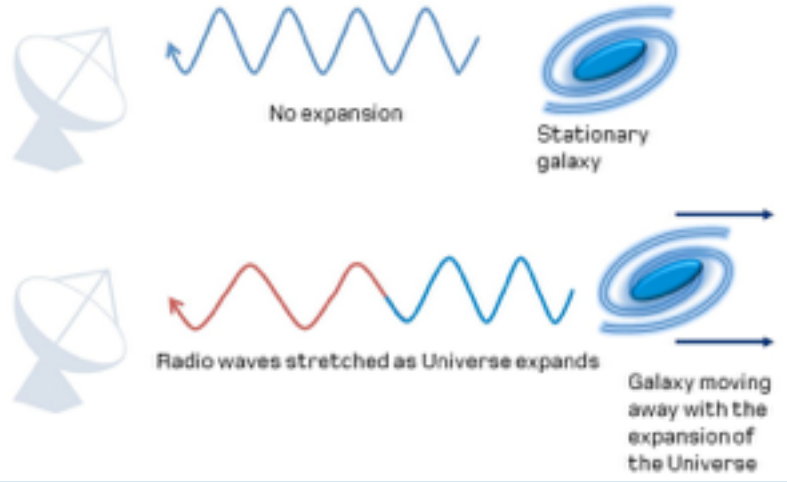


Brunet, Montmorry, Frezouls 2012

Big-Data today: LOFAR

LOFAR (Low Frequency Array) ground-based radio telescope (see van Haarlem et al. 2013)

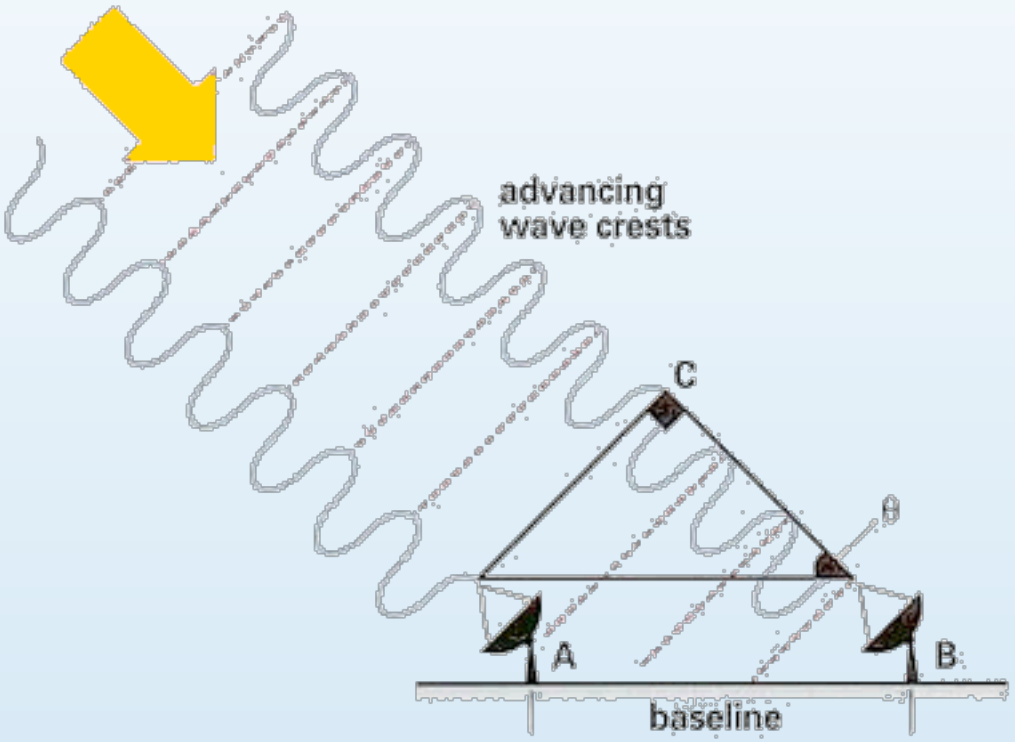
- Study the early universe → detect hydrogen HI from the epoch of re-ionization (redshift 6<z<20; 180 – 1000 million years after Big Bang)
- Radio transient phenomena
- 10-240 MHz
- Simple dipole antennas
- Signals from dipoles are combined digitally into phased array



V. Beckmann

Big-Data today: LOFAR

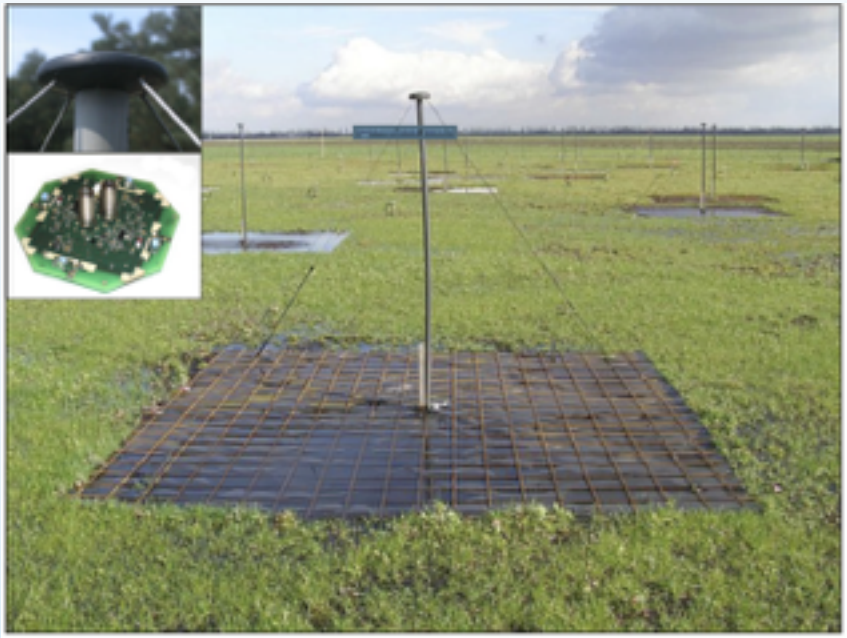
LOFAR ground-based radio telescope (see van Haarlem et al. 2013)



$$\theta = 1.220\frac{\lambda}{D}$$

Big-Data today: LOFAR

LOFAR ground-based radio telescope

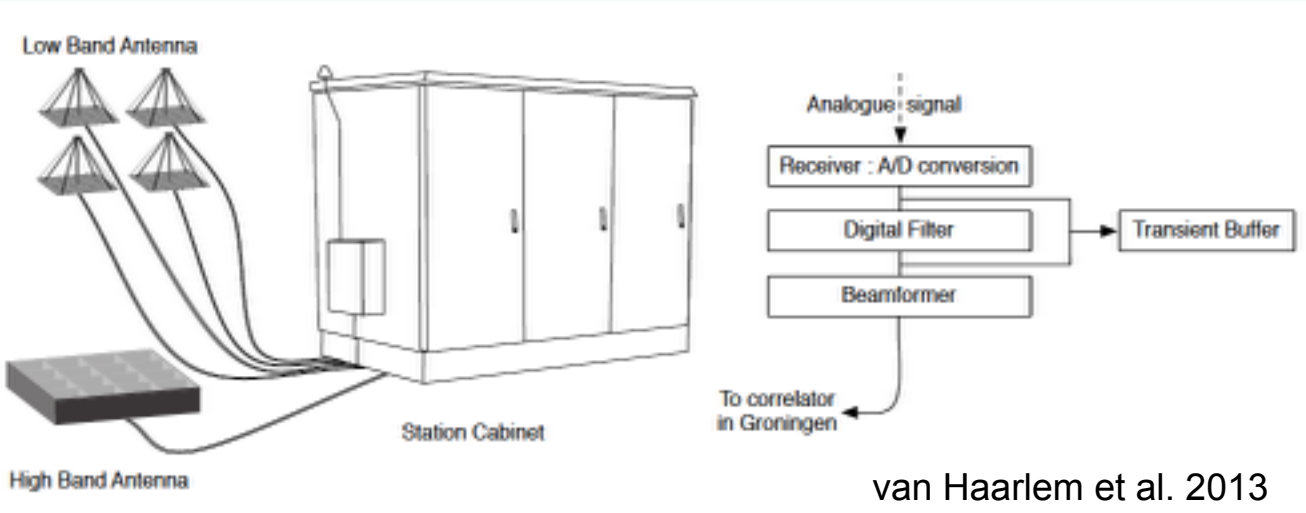


LOFAR LBA dipole
(van Haarlem et al. 2013)

LOFAR station Hamburg

Big-Data today: LOFAR

Big-Data today: LOFAR



van Haarlem et al. 2013

photo: LOFAR consortium

- Combination of on-site reduction and HPC
- 25 Pbyte/day raw data
- Local control unit: beam forming, digitization, filtering
- Correct for gain and phase differences → 250 Tbyte/day
- Central processing: three-rack IBM Blue Gene/ P
- 42 Tflop/s
- Time off-sets, FX correlator (Fourier transformation, time domain)

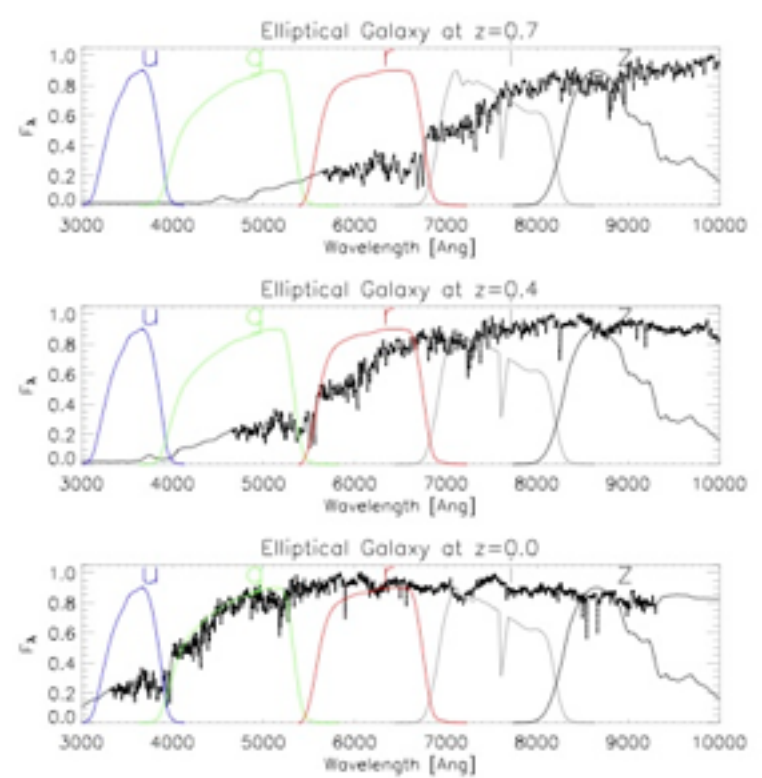
V. Beckmann

Big-Data tomorrow: Euclid

Study structure & evolution of Universe, dark matter, dark energy

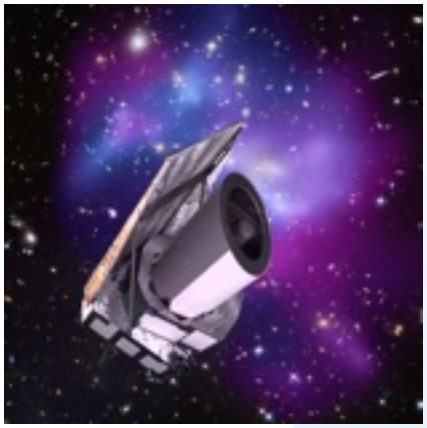
How to achieve science goals:

- Distribution of galaxies in space / distance (redshift)
- Significant fraction of the sky (1/3)
- Significant fraction of all galaxies (1% = 1 billion)
- Measure precise shapes
- Hubble Space Telescope image over 1/3 of the sky!
- Redshifts → photometric redshifts
- Add filters from ground-based observations
- Calibrate with some 10^5 spectra
- Add additional information



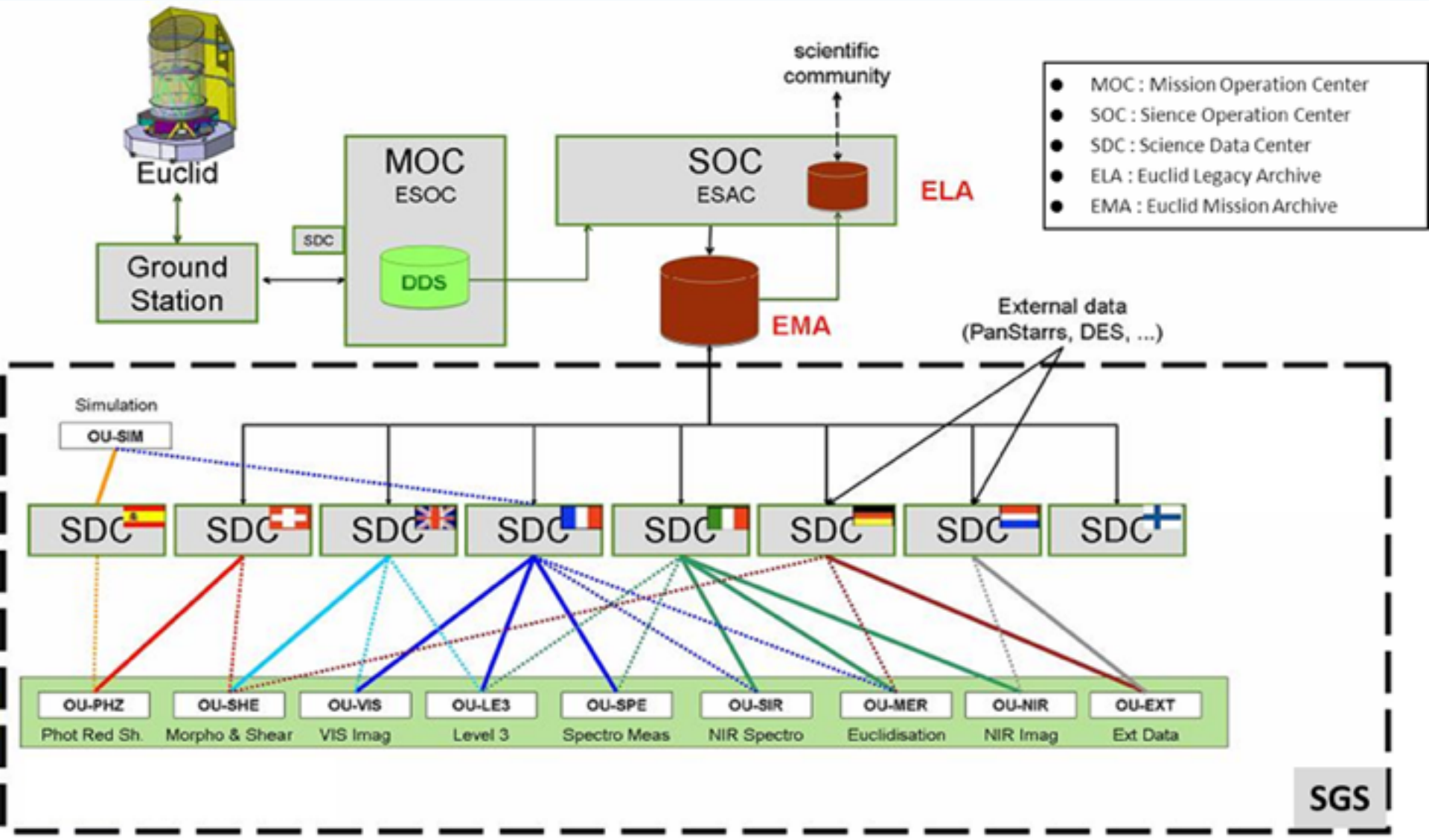
Padmanabhan (2007)

Big-Data tomorrow: Euclid



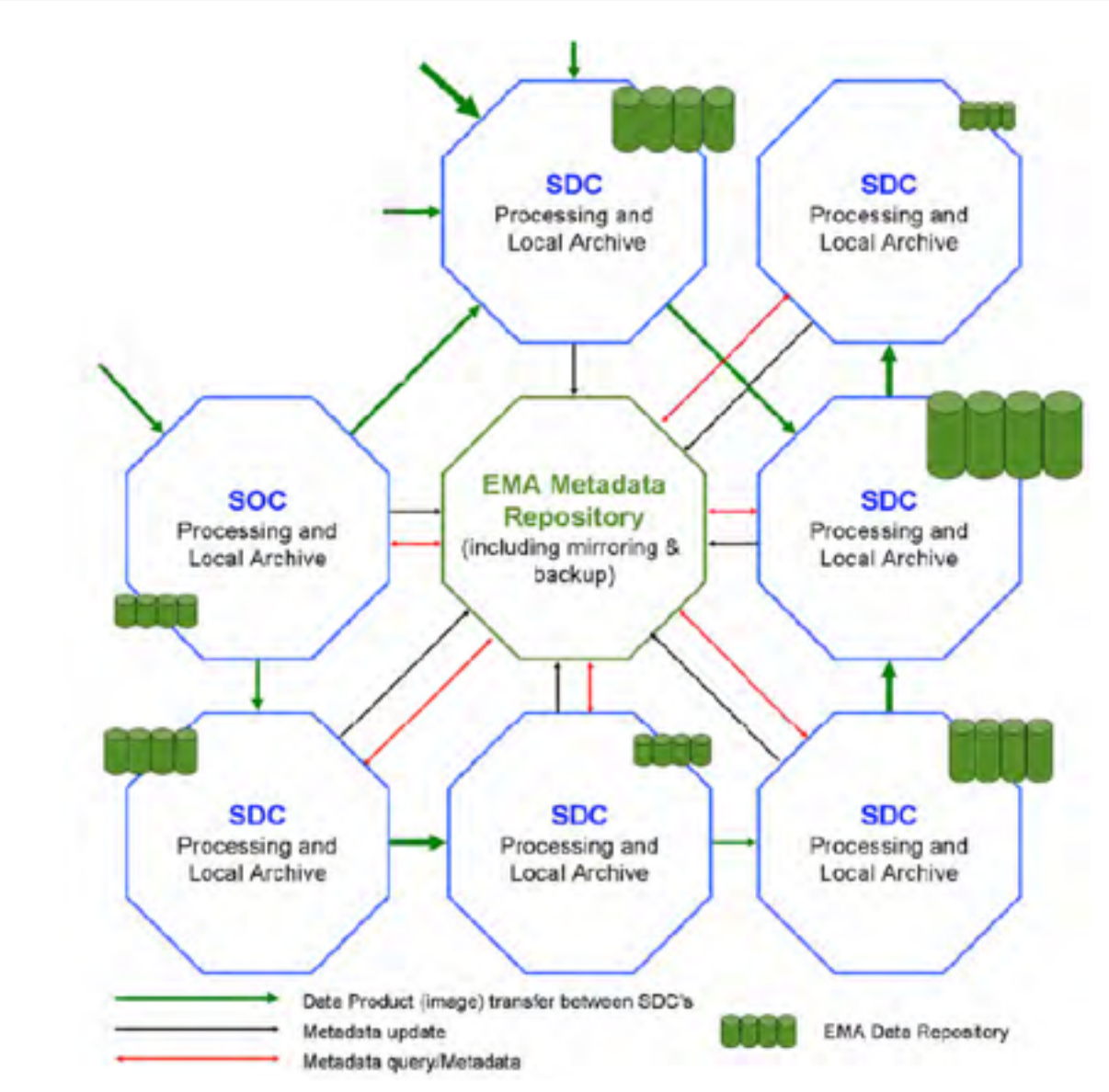
Euclid RedBook (2012)

Big-Data tomorrow: Euclid



Euclid RedBook (2012)

Big-Data tomorrow: Euclid

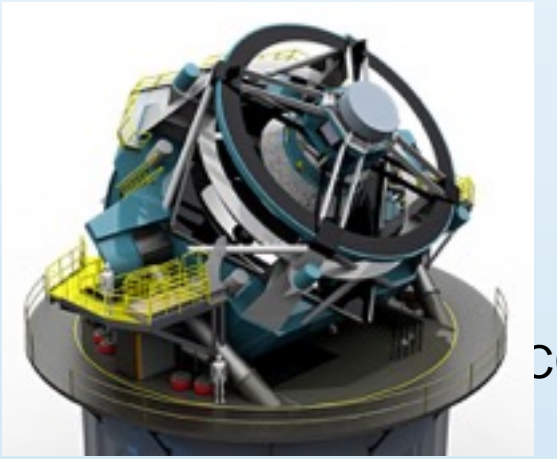


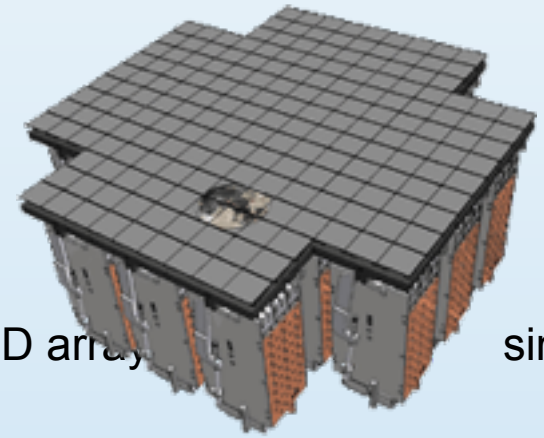
Euclid RedBook (2012)

Big-Data tomorrow: LSST

Ground based optical imaging telescope

Product	Factor	size
CCD image 4096x4096 pixel		64 MByte
Exposure	=189 CCD images	12.1 GByte
Visit	= 2 exposures	24.2 GByte
Nightly data	~1000 visits	25 TByte
One year	~300 nights	8300 TByte



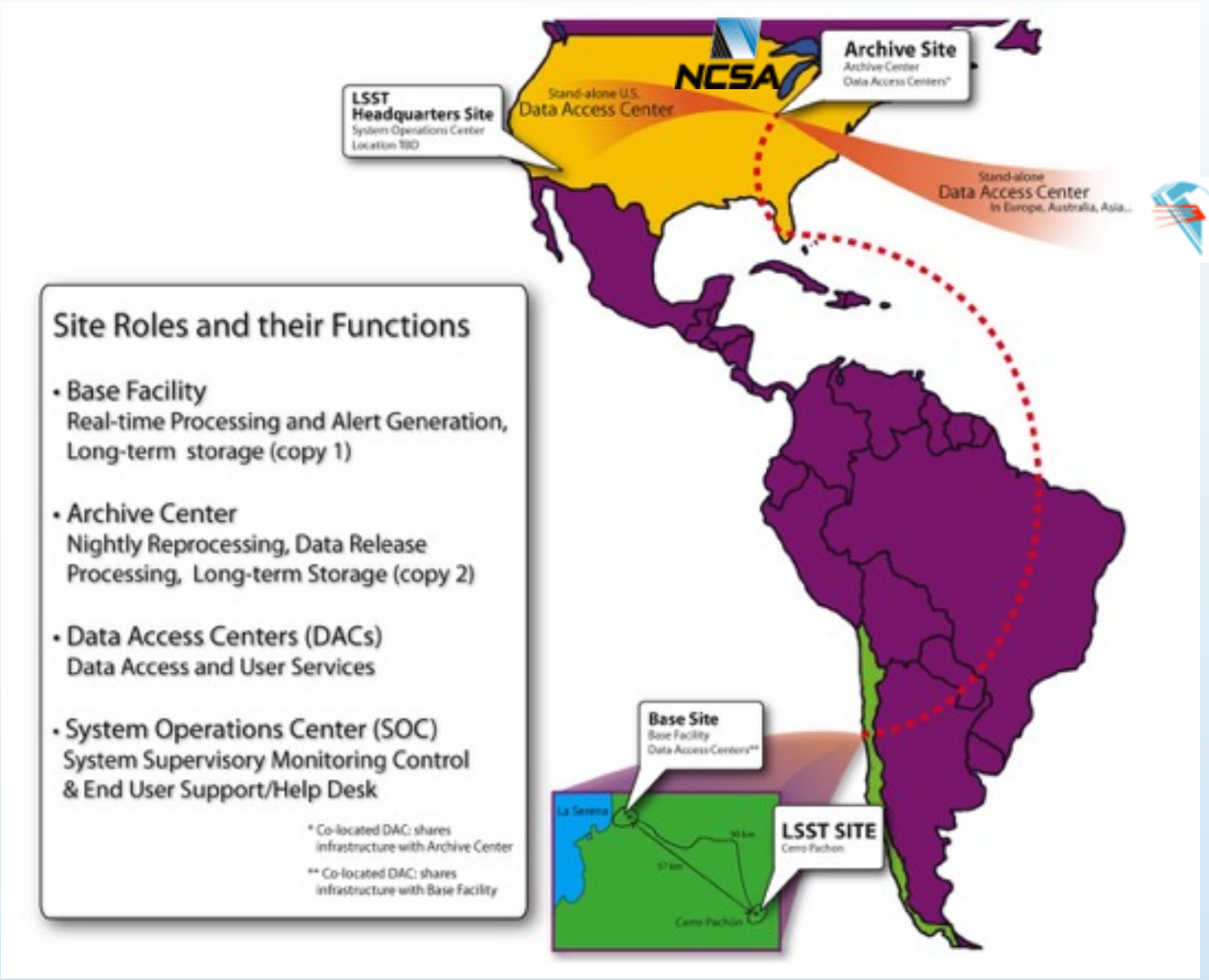


CCD array

simulat

V. Beckmann

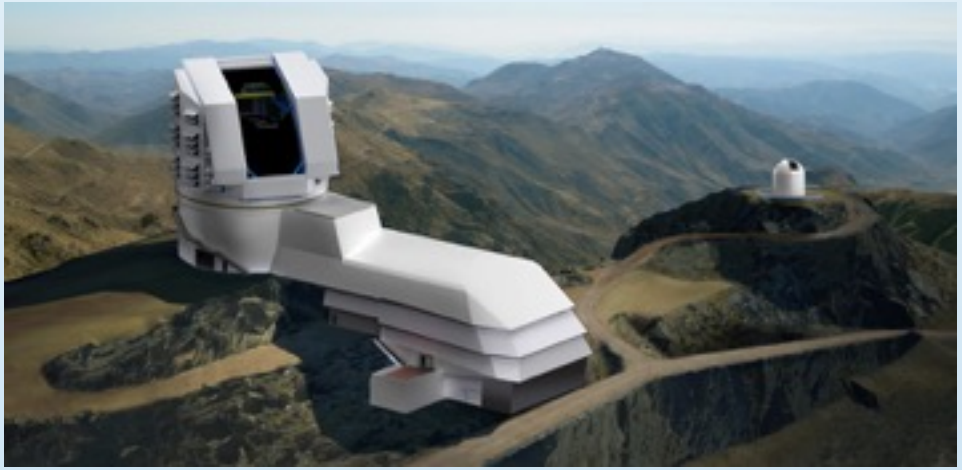
Big-Data tomorrow: LSST



Big-Data tomorrow: LSST

Computing Challenges:

- 8 Pbyte per year raw data → >100 Pbyte final products
- Detect transients and send out alerts within 60 seconds
- 20,000 deg^2 (half of the sky), 2000 times imaged in ten years
- Relational database tables (per sky field)
- 20 billion rows in the object table (unique sources)
- 3 trillion rows in the source table (all detections)
- Catalogue of 15 Pbyte
- Middleware should run on single laptop and on 100,000+ core cluster
- 150 Tflop/s



Big-Data tomorrow: LSST

approach:

- Middleware layer, isolating h/w, orchestrating s/w
- Massively parallel, distributed, fault tolerant, relational database
- Robust technologies: MySQL, xrootd
- Open source s/w project
- C++, python
- Modularity: everything is a python module
- 8 years prototyping, 40+ developers
- Two large HTC centers at NCSA (National Center for Supercomputing Applications at Univ. Illinois) and CC-IN2P3 (Lyon)

From presentation of Mario Juric (LSST data management project scientist)

Challenges

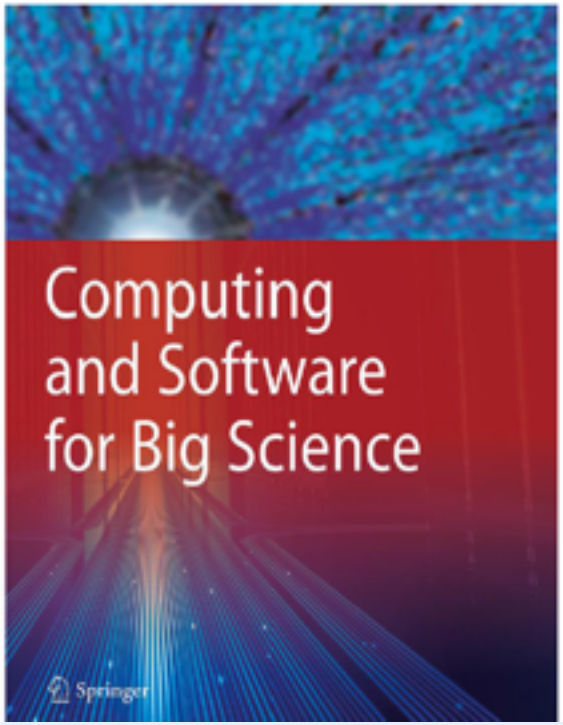
- Astrophyscis goes BigData: Gaia, LOFAR, Euclid, LSST, CTA, SKA, …
- Pbyte scale data with need of Tflop processing (HTC & HPC)
- Solutions depend on the science requirements
- Space or ground, remote or central, real-time processing or not…
- Advantage: community used to work together, file format standards, coding standards (C++, python)
- Development platforms
- Disadvantage: artificial splitting of data centers, astronomers

Approach

- Try to reduce data!
- On-site (ground-based; LOFAR, CTA, SKA, …), on-board satellite processing (Gaia) where possible
- Then: centralise if possible (Gaia, LOFAR, SKA)
- use as few as possible (LSST, Euclid)
- Make use of existing distributed computing infrastructure (CTA)
- GPUs: fast, but not good on i/o → Fourier transformation (e.g. LOFAR), template fitting
- GRID: infrastructure, middleware, relatively heavy to use → CTA
- HPC / HTC with Hadoop, Hive
- Cloud: virtualisation, flexibility, lower performance → project development,

Initiatives

- International Astronomical Union (IAU); coordinate science and approaches throughout the community
- Virtual Observatory: combine data and tools from many experiments, many wavelengths. Use on-line or off-line
- Research Data Alliance (RDA); research data sharing without barriers
- Astroparticle Physics European Consortium (ApPEC)
- GRID (EGI, NGIs, France Grilles, ...)



Computing and Software for Big Science

Editor-in-Chief: V. Beckmann; M. Elsing; G. Quast

- Explores emerging issues in big-science development
- Includes coverage of distributed data analysis and deep learning algorithms
- Presents articles on software benchmarking, performance assessment and on data-quality monitoring on or off-line
- Discusses aspects of evolving computing infrastructures

Addressing challenges ranging from data reduction via data sharing, to the need for data-driven modeling, this journal explores concepts for large-scale, collaborative computing and software development as well as new algorithms and techniques for data processing.

http://www.springer.com/physics/particle+and+nuclear+physics/journal/41781

The road ahead…

Astrophysics arrived in the world of Big-Data

User interfaces, usability by individual scientists

Maintain and build up expertise on computing science

Volker Beckmann