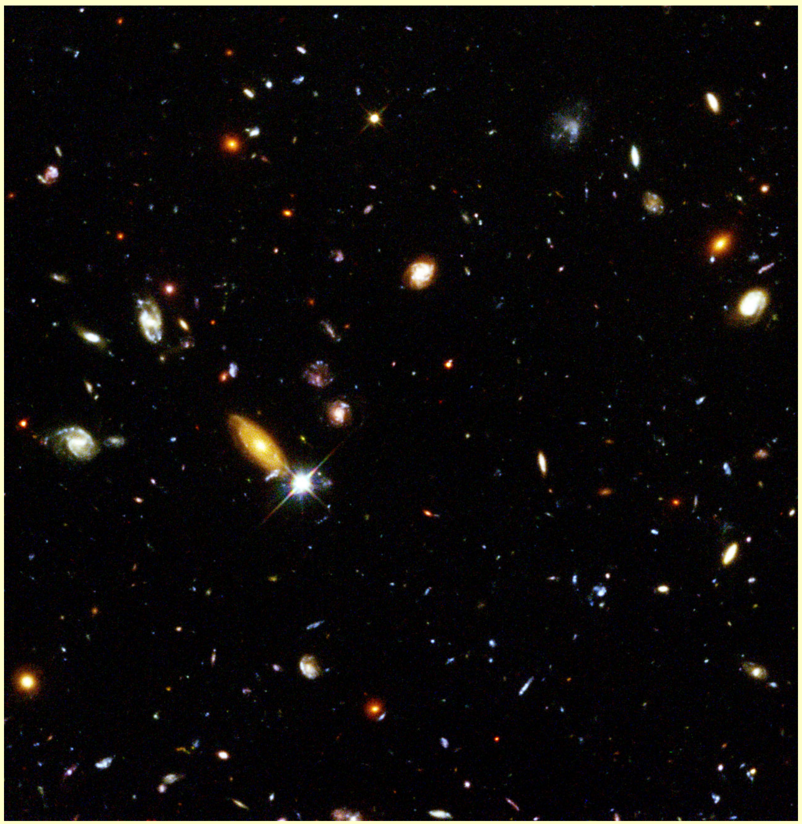
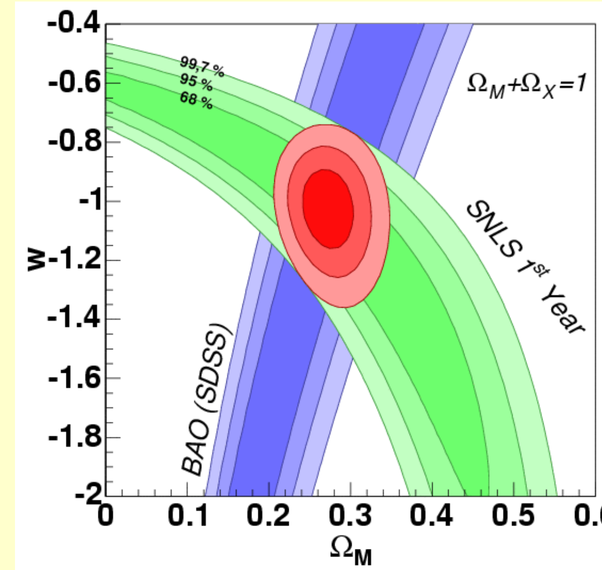
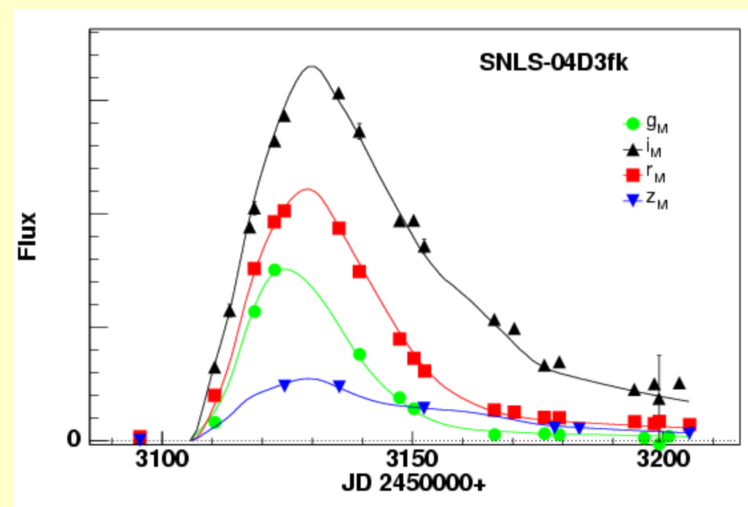
SNLS
Data Reduction



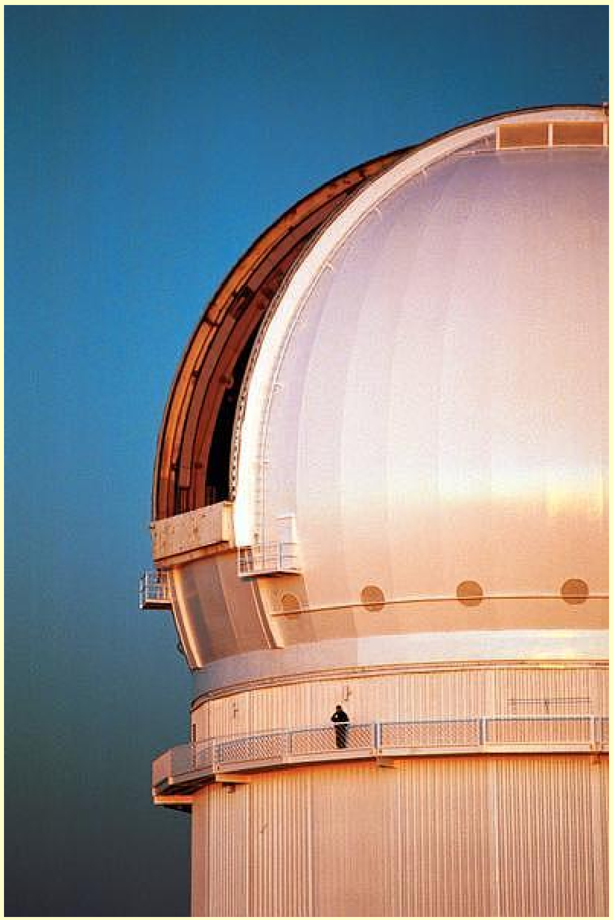
Pierre Astier
LPNHE/IN2P3/CNRS
Universités Paris 6&7



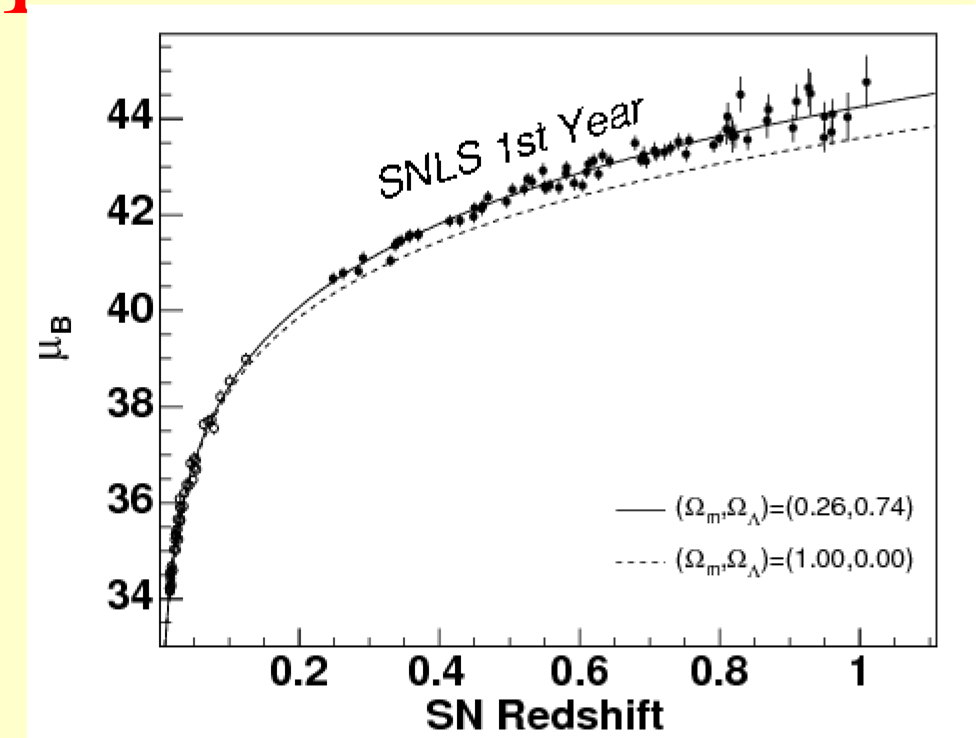


Aspera in Barcelona
30/05/2011

SuperNova Legacy Survey: SNe Ia @ 0.2<z<1



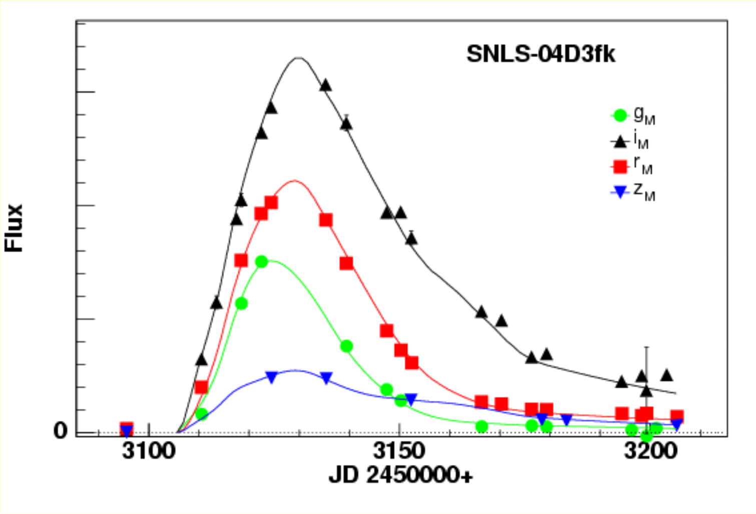
Distances photometry (CFHT, 3.6m)



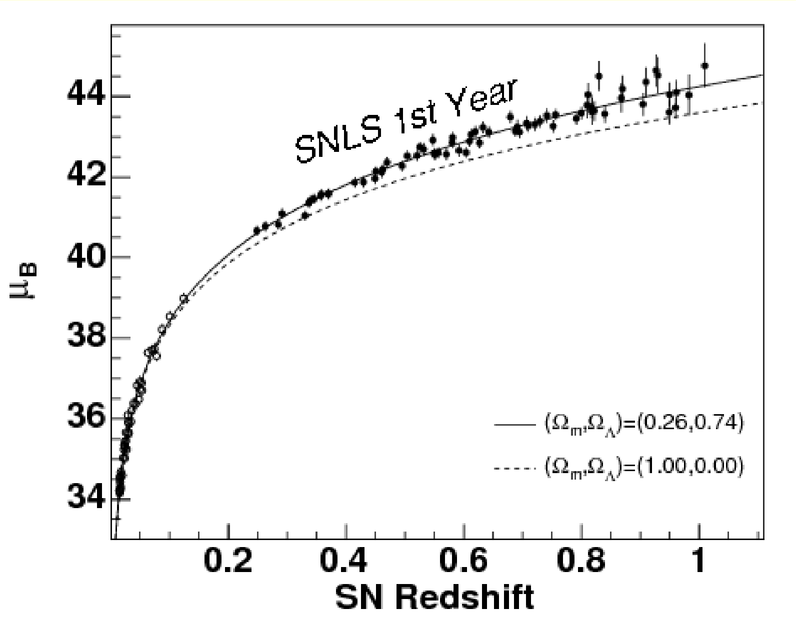
Redshifts : spectroscopy (VLT, Gemini, Keck)

Pierre Astier (30/05/2011 Aspera)

Hubble diagram : flux vs redshift



Peak flux



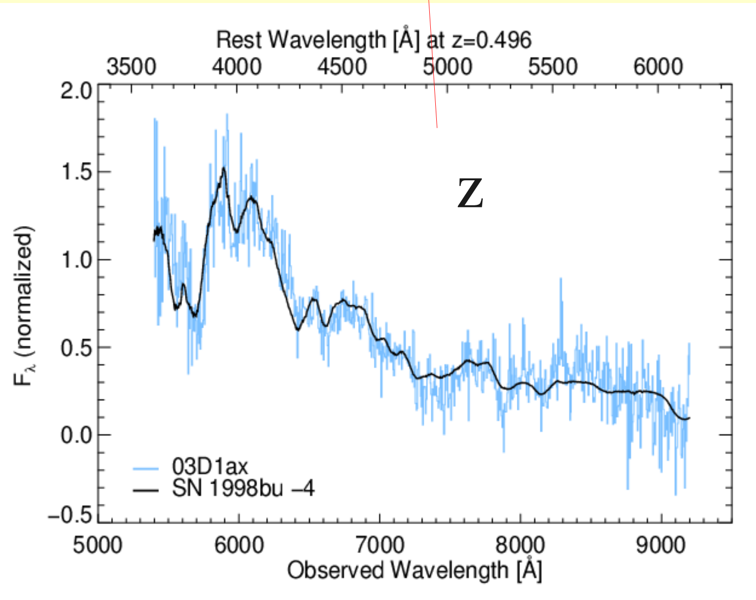
Multi-band photometry

=> distance

The Hubble diagram
probes the kinematics
of the expansion of
the universe …
… and hence its contents

spectroscopy:
- identification
- redshift

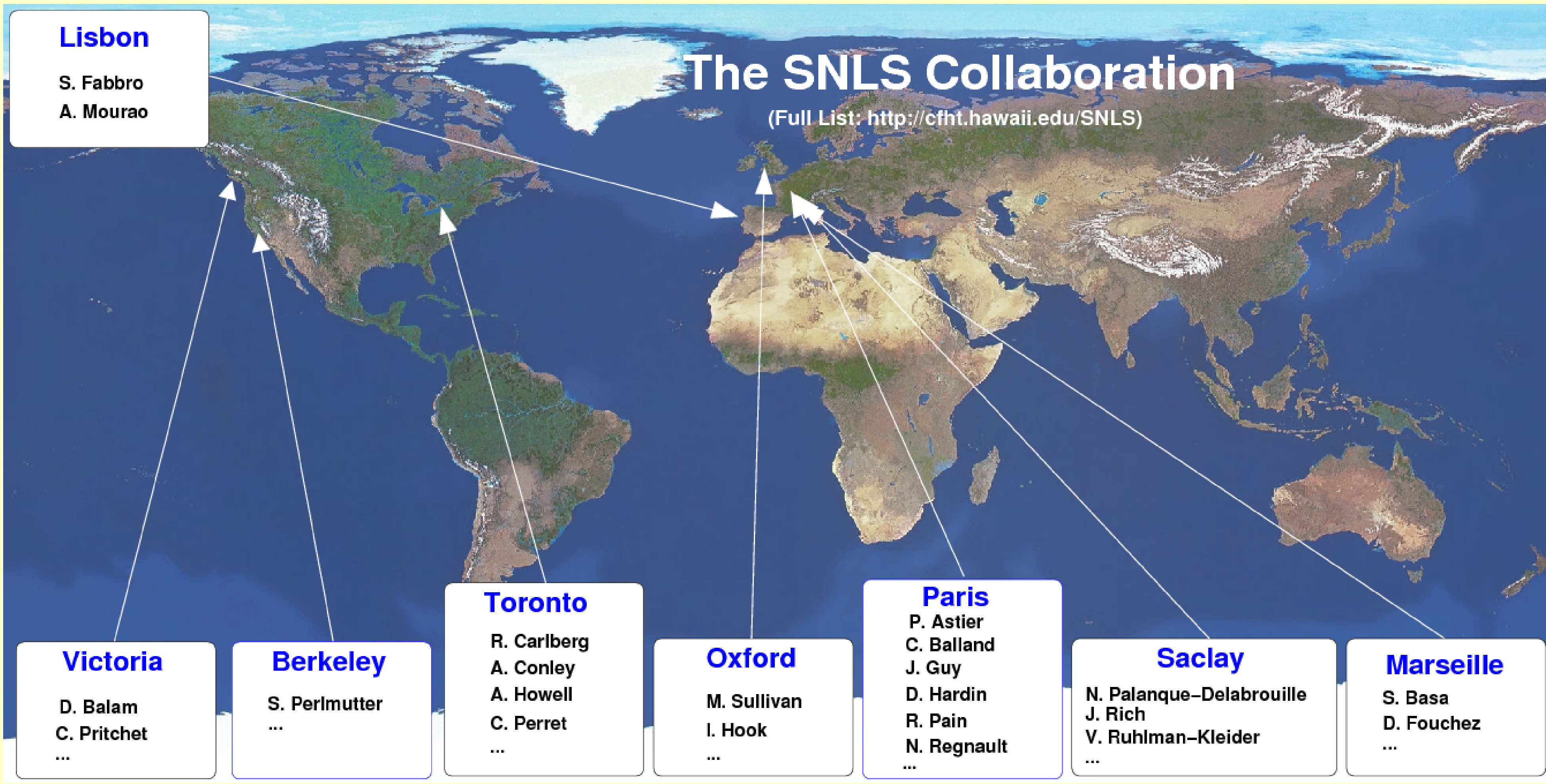
Z



Pierre Astier (30/05/2011 Aspera)

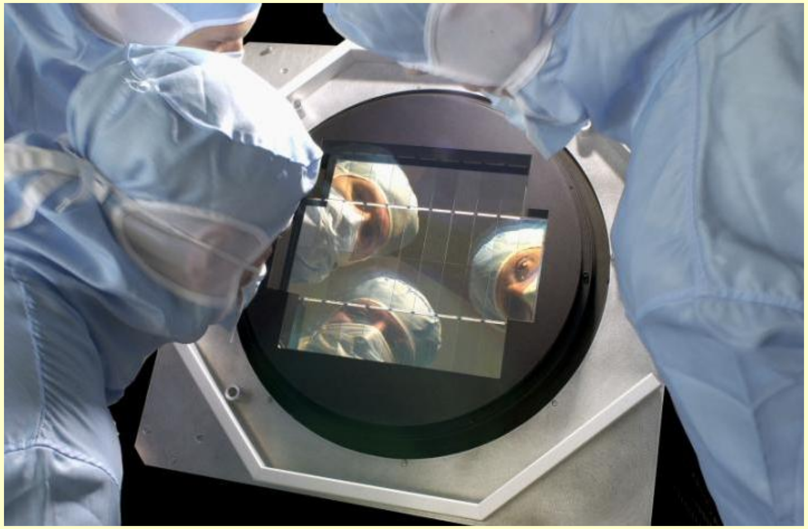
The SNLS Collaboration

(SuperNova Legacy Survey)

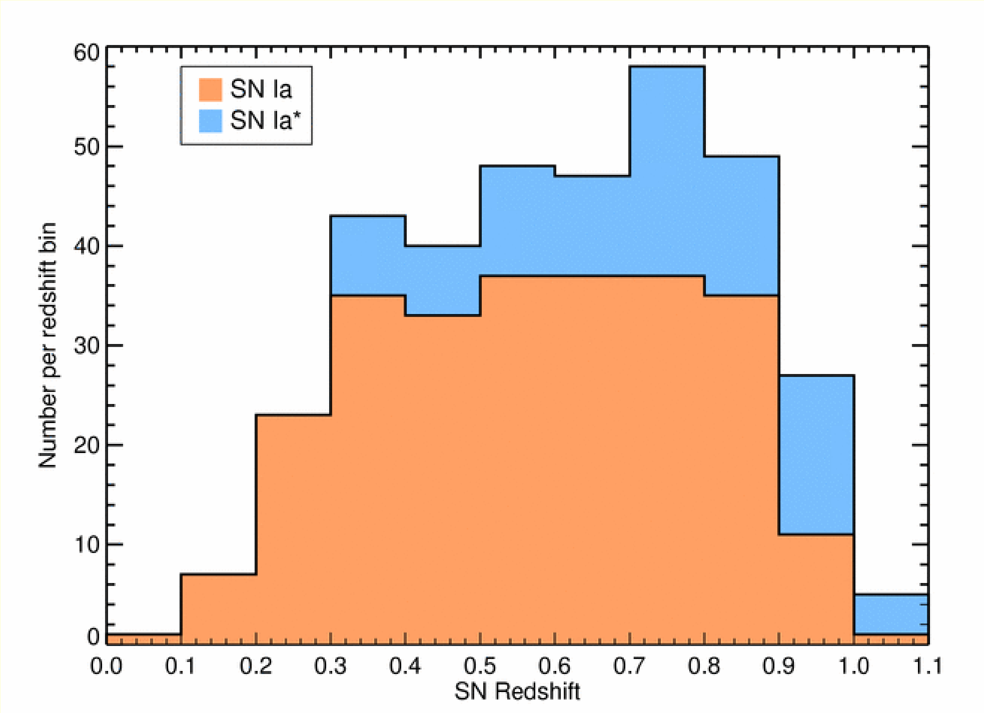


SNLS: statistics

- SN survey in "rolling search" mode
  (within CFHTLS, 5 years)
- Goal: ~500 SNe to constrain dark energy)
- Uses Megacam on CFHT (340 Mpixels)
  4 bands g,r,i,z / 4 nignts
- Spectroscopy on VLT,Gemini & Keck.



- Survey over (2003-2008)
- ~400 confirmed SNe Ia
  $0.1<z<1.05$
- ~1000 detected SNe



SNLS Data?

Spectroscopy (VLT, Gemini, Keck):

Small volumes (<200 Mb per event , ~500 events)

Reduction is semi-automatised.

Imaging (CFHT/Megacam) for photometry :

The telescope delivers "exposures" which contain 36 CCD

Science & calibration exposures

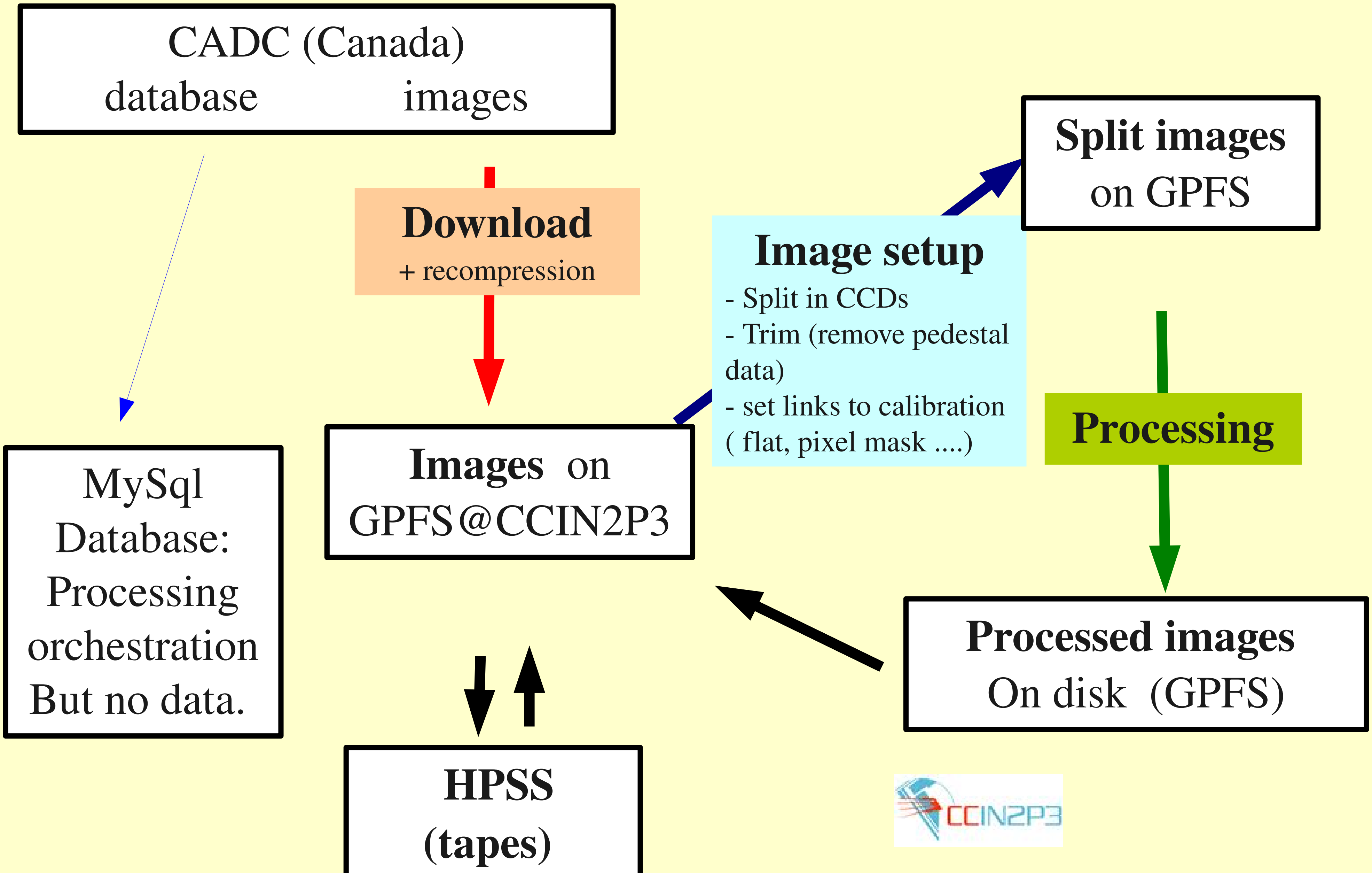
Total : ~20000 expositions of 340 Mpixels (680 Mb)

What do we do with images?

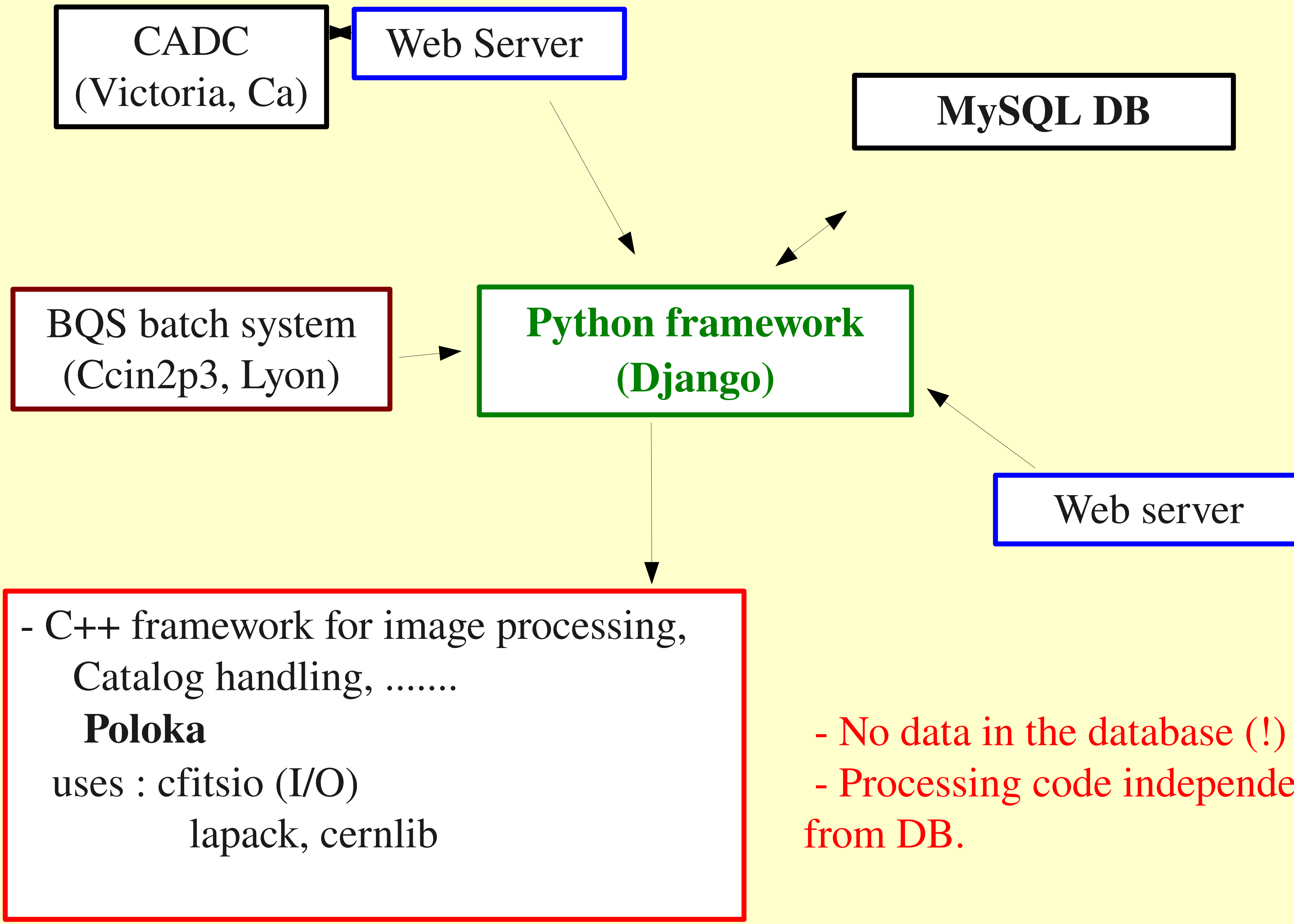
- Detecting supernovae (difference imaging run in Hawaii)
- Photometry of supernovae.
- Photometric calibration.
- Deep stacks.
- Add fake supernovae to images and run the detection.
- ....

SNLS : data transfers

Data flow


CADC (Canada)
database images
Download
+ recompression
MySql
Database:
Processing
orchestration
But no data.
Images on
GPFS@CCIN2P3
Image setup
- Split in CCDs
- Trim (remove pedestal
data)
- set links to calibration
(flat, pixel mask ....)
Split images
on GPFS
Processing
Processed images
On disk (GPFS)
HPSS
(tapes)
CCIN2P3

Software



- No data in the database (!)
- Processing code independent from DB.

The storage element : the “image”

An « image » = 1 CCD (9 Mpixels) from an exposure
One exposure = 36 “images” = 36 CCD.

An image = a directory
That contains :

- The image itself
- Weight image
- Flat, Bias, …. as links
- Several catalogs
- A set of masks
  (dead/satur/satellite/cosmics)
- Processing logs, ….

```
ccali46:tcsh[107] ls -lhrt /sps/snls13/PROD08/DATA/D2/2008-03-09/r/0/975277p21
total 34M
-rw-r--r--  1 snlsprod snovae   31 Apr 30 16:57 imcopy.log
-rw-r--r--  1 snlsprod snovae 8.5M Apr 30 16:57 elixir.fz
-rw-r--r--  1 snlsprod snovae  391 Apr 30 16:57 trim.log
lrwxrwxrwx  1 snlsprod snovae   55 May  6 21:11 flat.fits -> /sps/snls12/DETREND/
lrwxrwxrwx  1 snlsprod snovae   55 May  6 21:11 bias.fits -> /sps/snls12/DETREND/
lrwxrwxrwx  1 snlsprod snovae   54 May  6 21:11 dead.fits -> /sps/snls12/DETREND/
-rw-r--r--  1 snlsprod snovae  24K May  8 10:50 satur.fits.gz
-rw-r--r--  1 snlsprod snovae 138K May  8 10:50 segmentation.cv.fits.gz
-rw-r--r--  1 snlsprod snovae 129K May  8 10:50 segmentation.fits.gz
-rw-r--r--  1 snlsprod snovae 5.7K May  8 10:50 miniback.fits
-rw-r--r--  1 snlsprod snovae 9.8K May  8 10:50 low.fits.gz
-rw-r--r--  1 snlsprod snovae 868K May  8 10:50 se.list
-rw-r--r--  1 snlsprod snovae  25K May  8 10:50 cosmic.fits.gz
-rw-r--r--  1 snlsprod snovae 6.3K May  8 10:50 make_catalog_v1.0.log
-rw-r--r--  1 snlsprod snovae 2.5M May  8 20:33 aperse.list
-rw-r--r--  1 snlsprod snovae 121K May  8 20:33 standalone_stars.list
-rw-r--r--  1 snlsprod snovae  754 May  8 20:33 mkcat2_v1.0.log
-rw-r--r--  1 snlsprod snovae  20K May 15 02:18 satellite.fits.gz
-rw-r--r--  1 snlsprod snovae  13M May 15 02:18 weight.fz
-rw-r--r--  1 snlsprod snovae 1.1K May 15 02:18 findsatellite_v1.0.log
-rw-r--r--  1 snlsprod snovae  69K May 18 02:37 match_usno.dat
-rw-r--r--  1 snlsprod snovae 255K May 18 02:37 all_coords.dat
-rw-r--r--  1 snlsprod snovae 8.5M May 18 02:37 calibrated.fz
-rw-r--r--  1 snlsprod snovae 2.2K May 18 02:37 matchusno_v2.0.log
```

Volumes on disc

Input images : about 20 000 expositions x 330 Mpixels ~ 13 Tb

Images pre-reduced (size *= 2 …):
- science (SN fields) : 450 000 x 35 Mb = 16 Tb
- calibration : = 6 Tb

Resampled images (for light curves) = 12 Tb

2 productions (science, development) ~ 70 Tb

Other stuff : Fake supernovae, deep stacks, sandbox ..... ~ 10 Tb

==> less than 100 Tb. All on GPFS (at ccin2p3@lyon)
with backups on HPSS.

Image compression

Image files are encoded in the FITS format:

- Header (image metadata, ASCII) ~10 kb
- Pixels (9Mp x 16 bits) 18 Mb

→ Compressing the header is just useless.

A lossless compression reduces the file size by a factor of 2 :

+ disk space
+ network traffic
+ negligible impact on computing time.

We use the RICE compression (integrated to CFITSIO)

-> "block" compression → direct access to image lines.

-> acceptable increase of I/Os to access image cutouts.

Processing pipelines : examples

Pre-reduction : from raw image to “calibrated”

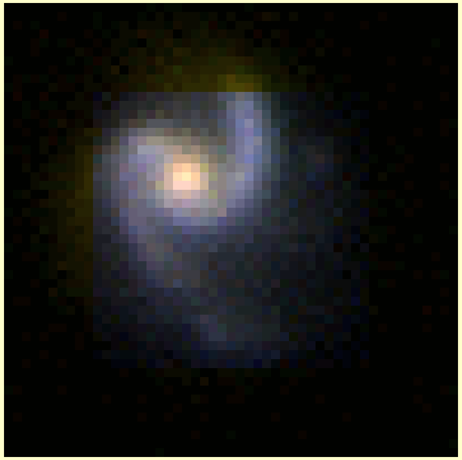
- make a catalogue, detect cosmics, planes, ...

.....

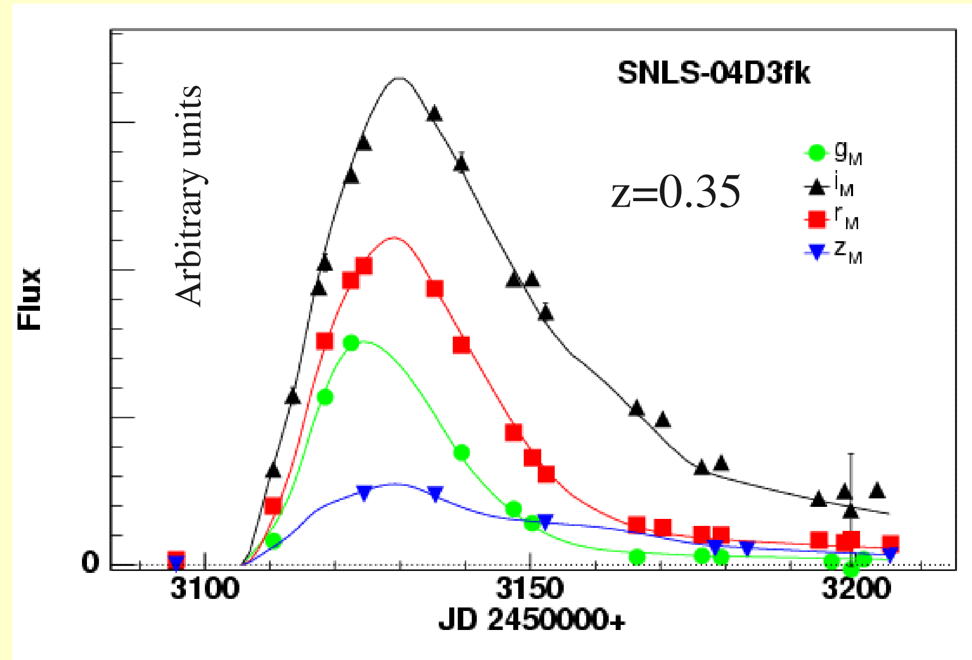
==> Traverse all pixels for short computations.
“Computing” time dominated by I/O's

Light curves : flux(time) for transients

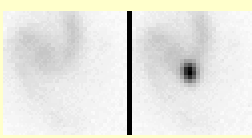
- 1) re-sample all images at the same pixel grid
  => 1 image read (20 Mb), 1 image written. CPU : ~ few s/image
- 2) evaluate convolution kernels
  => 1 image read, ~2kb written . CPU : ~ 30 s / image
- 3) simultaneous fit to all images (~700 images)
  => ~read 2000 stamps. CPU ~ 2 hours.



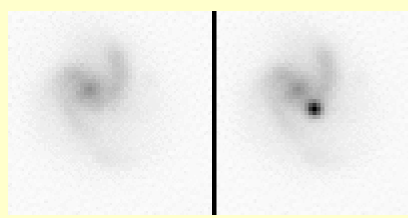
Light curves



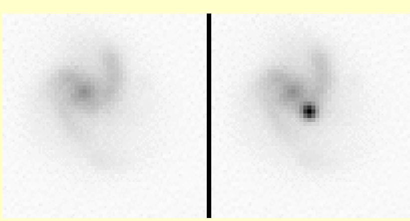
We process the 4 bands independently



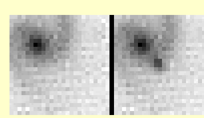
g



r



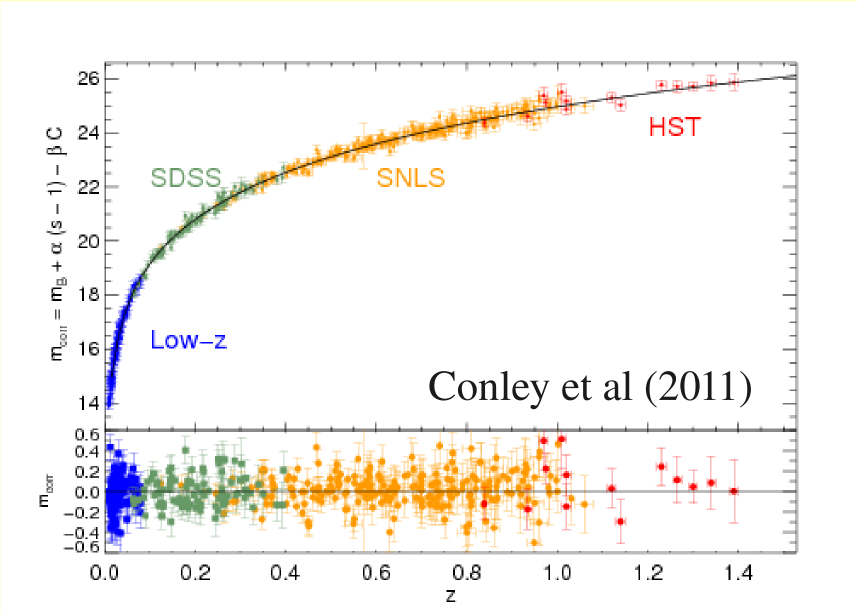
i

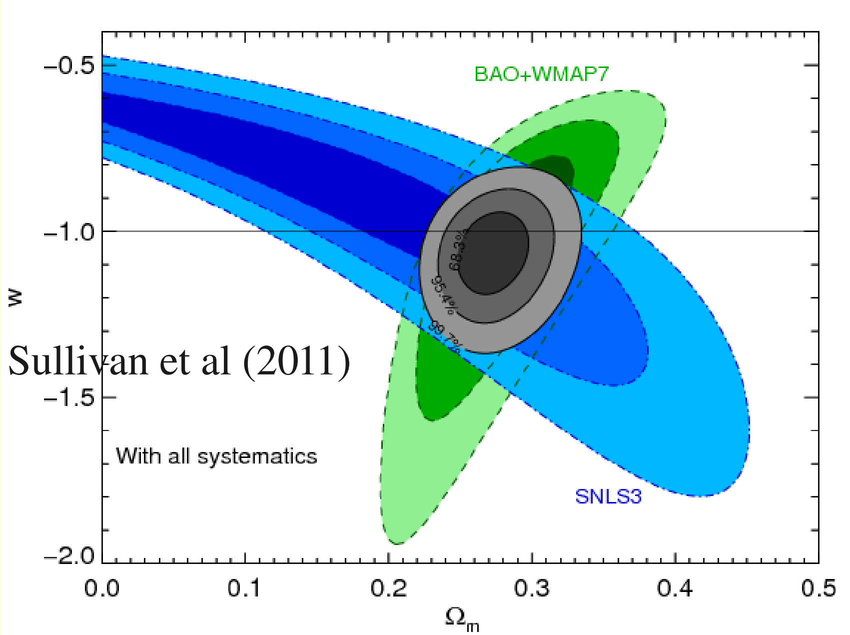


z

By-product:
“averaged” galaxy images

Cosmology fits





- For SNLS, cosmology fits may run on a desktop.

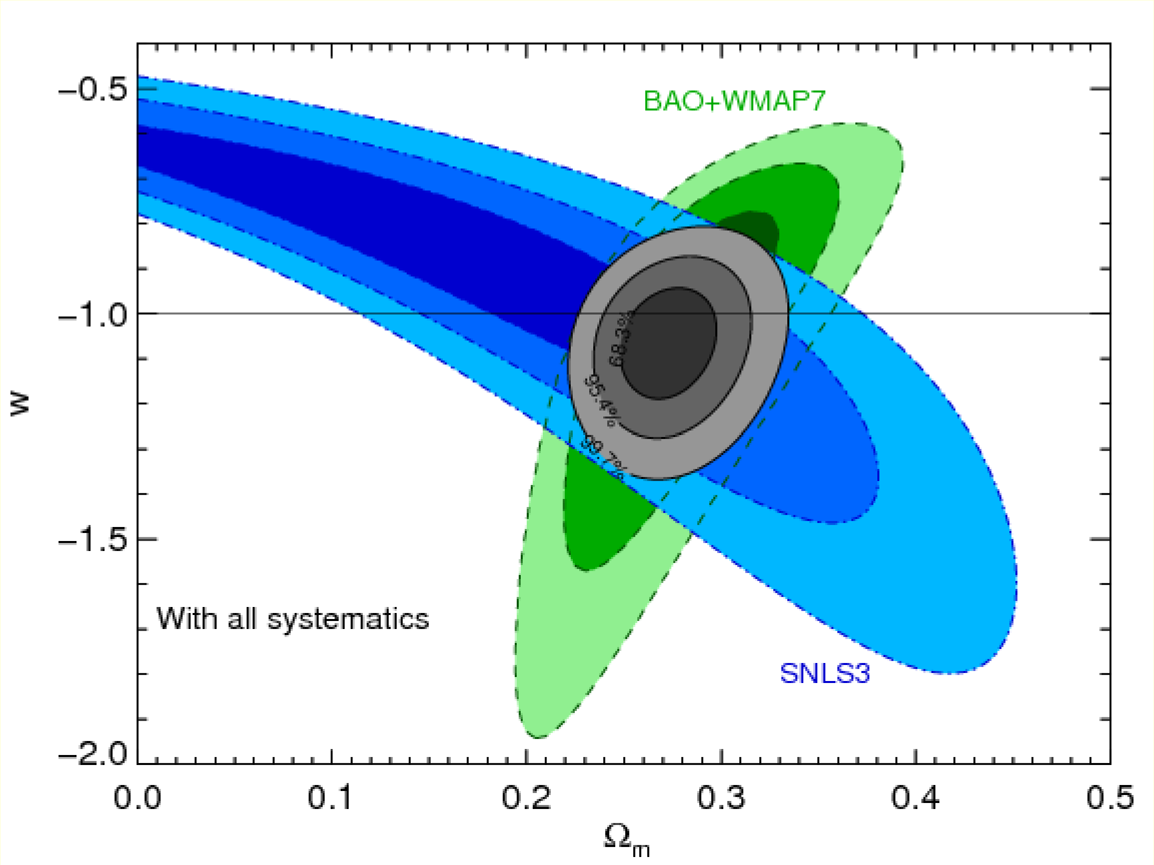
- Unlike essentially any cosmological analysis involving more than geometry.

- Computing the evolution of perturbations is CPU intensive, hence cosmology fits involving perturbations are CPU intensive.

→ joint analyses with CMB could be expensive.

Data reduction for SNLS : summary

- Modest amount of data by nowadays standards :
  few tens of Tb's
- Mixes I/O intensive and CPU intensive jobs.
- Algorithms require data to be on disk.
- Currently, our GPFS setup limits us to ~100 simultaneous jobs (i.e. 100 cores).
- The reduction typically takes a few weeks (+ problems)
- Scaling the data volume by 10 would require significant improvements in both hardware and software.



Regnault et al, 2009
Guy et al, 2010
Conley et al, 2011
Sullivan et al, 2011

By the way, the expansion of the universe is accelerating