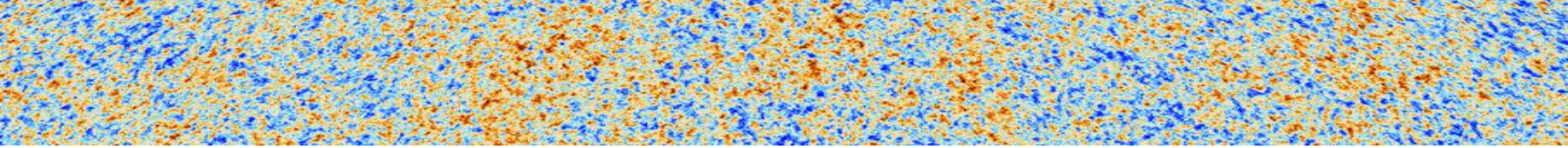


Analysis, simulations and combination of data

Paolo Natoli

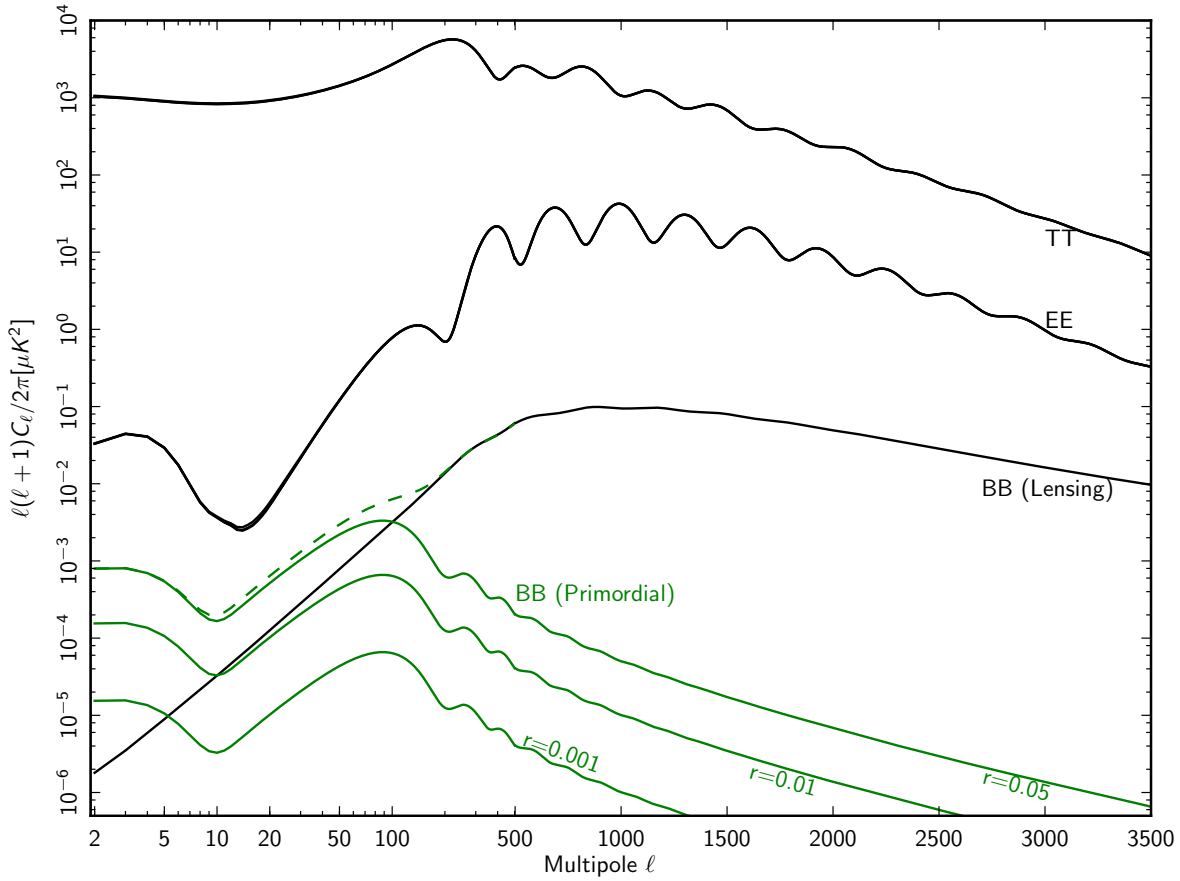
Università di Ferrara and INFN

Towards the European Coordination of the CMB programme

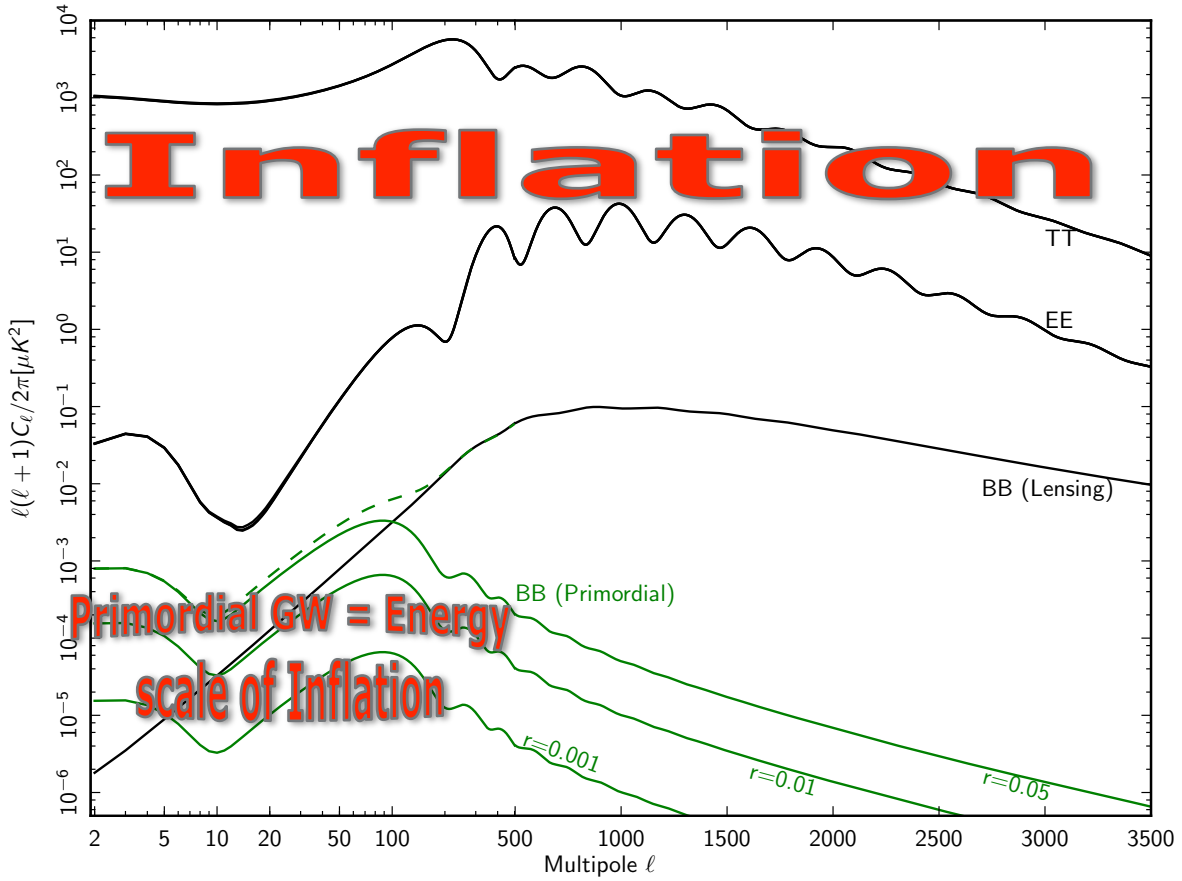
Villa Finaly, Firenze, 8 September 2016

Overview

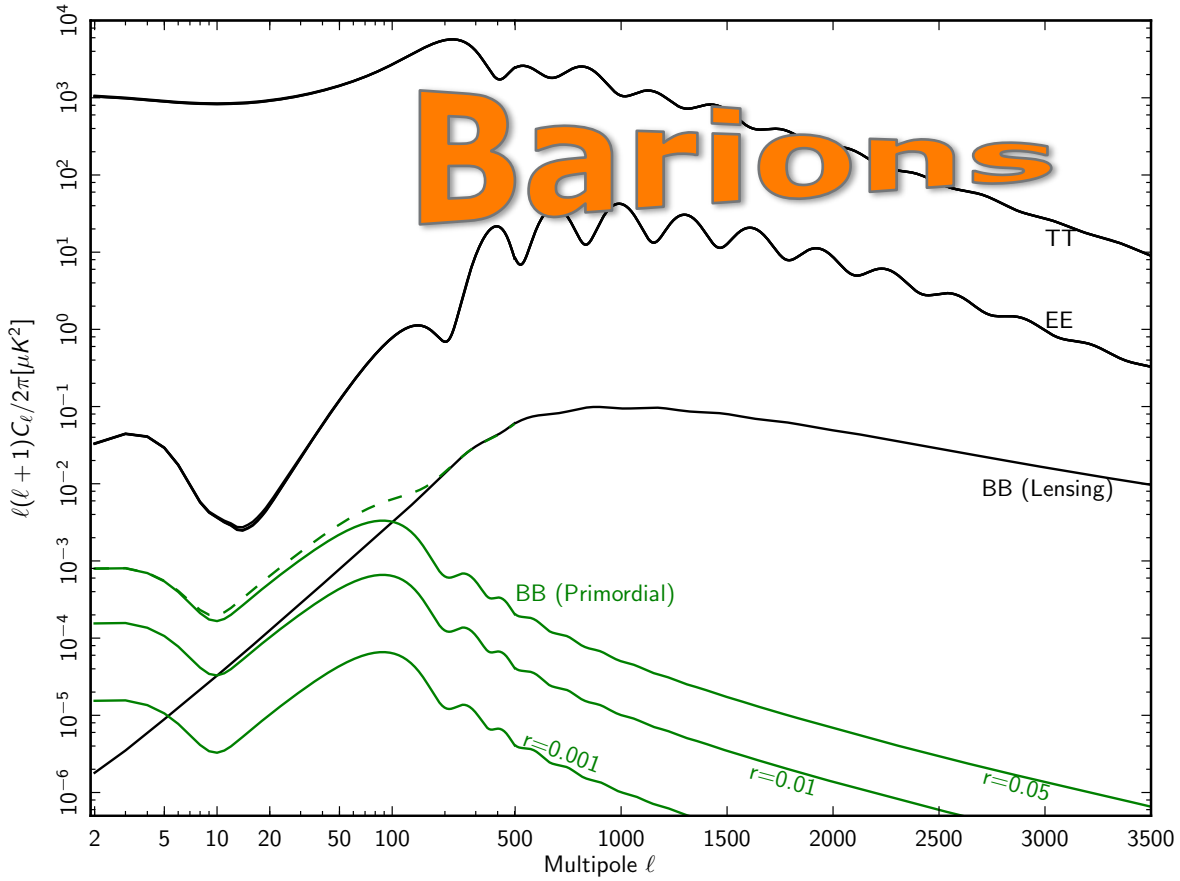
- Science targets and analysis challenges
- CMB pipelines and computational needs
- The role of simulations
- Future perspectives and conclusions


$\ell(\ell+1)C_\ell/2\pi[\mu K^2]$
Multipole ℓ
TT
EE
BB (Lensing)
BB (Primordial)
r=0.001
r=0.01
r=0.05

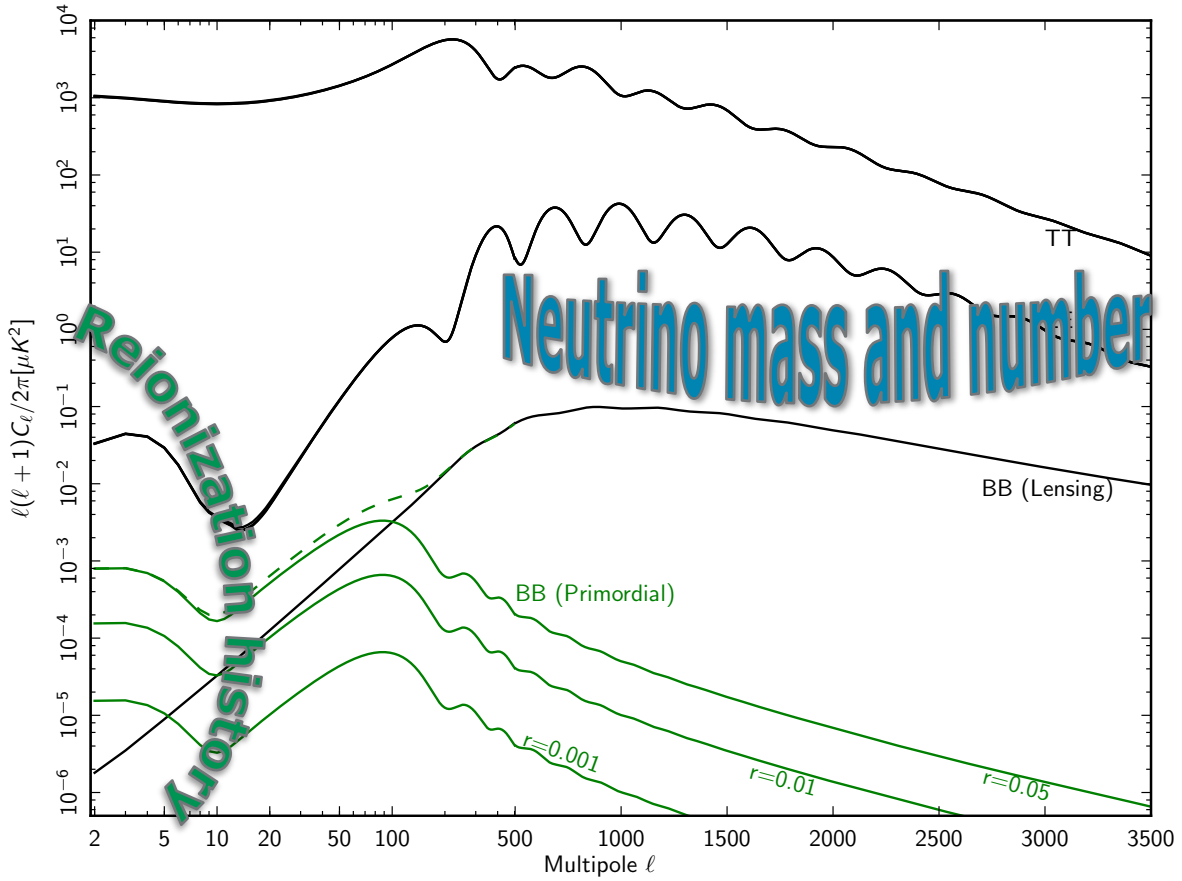
Science targets



Science targets

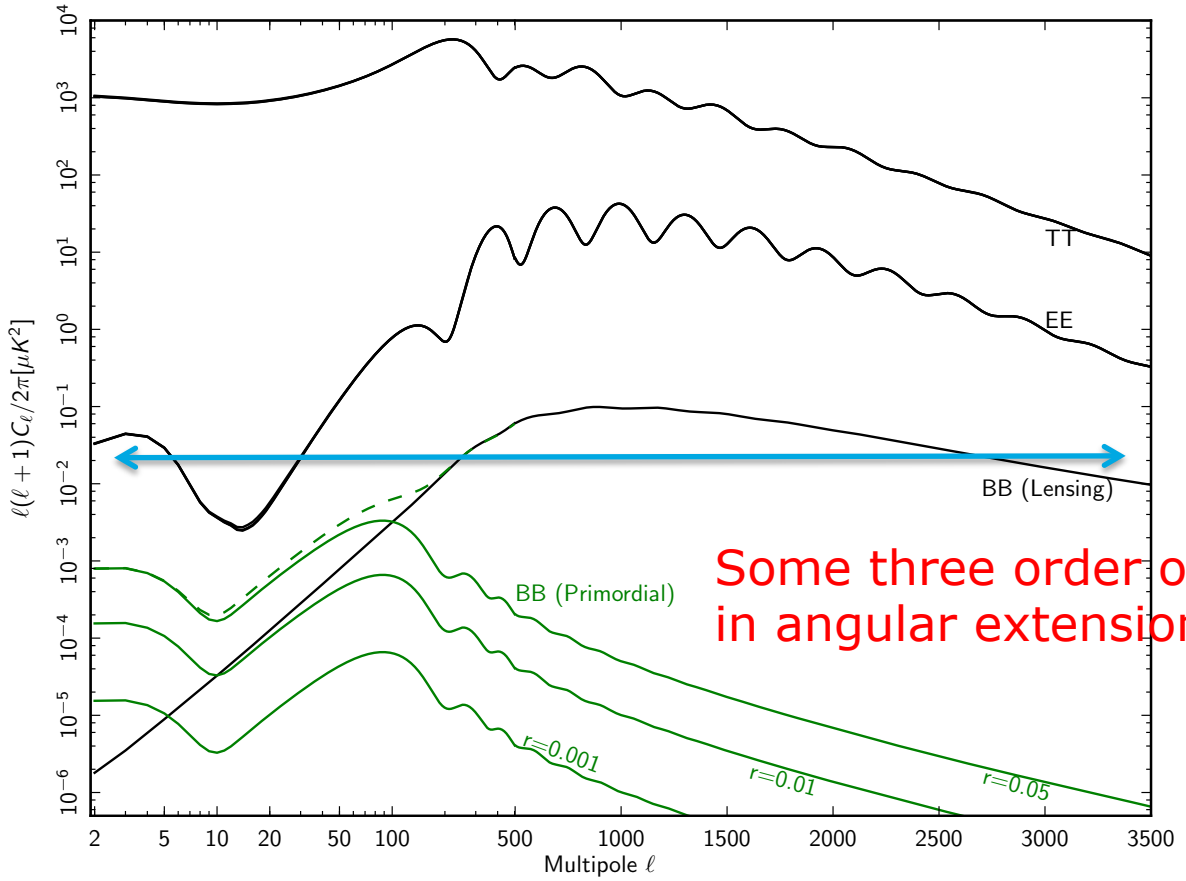


Science targets



CMB exploitation in a nutshell

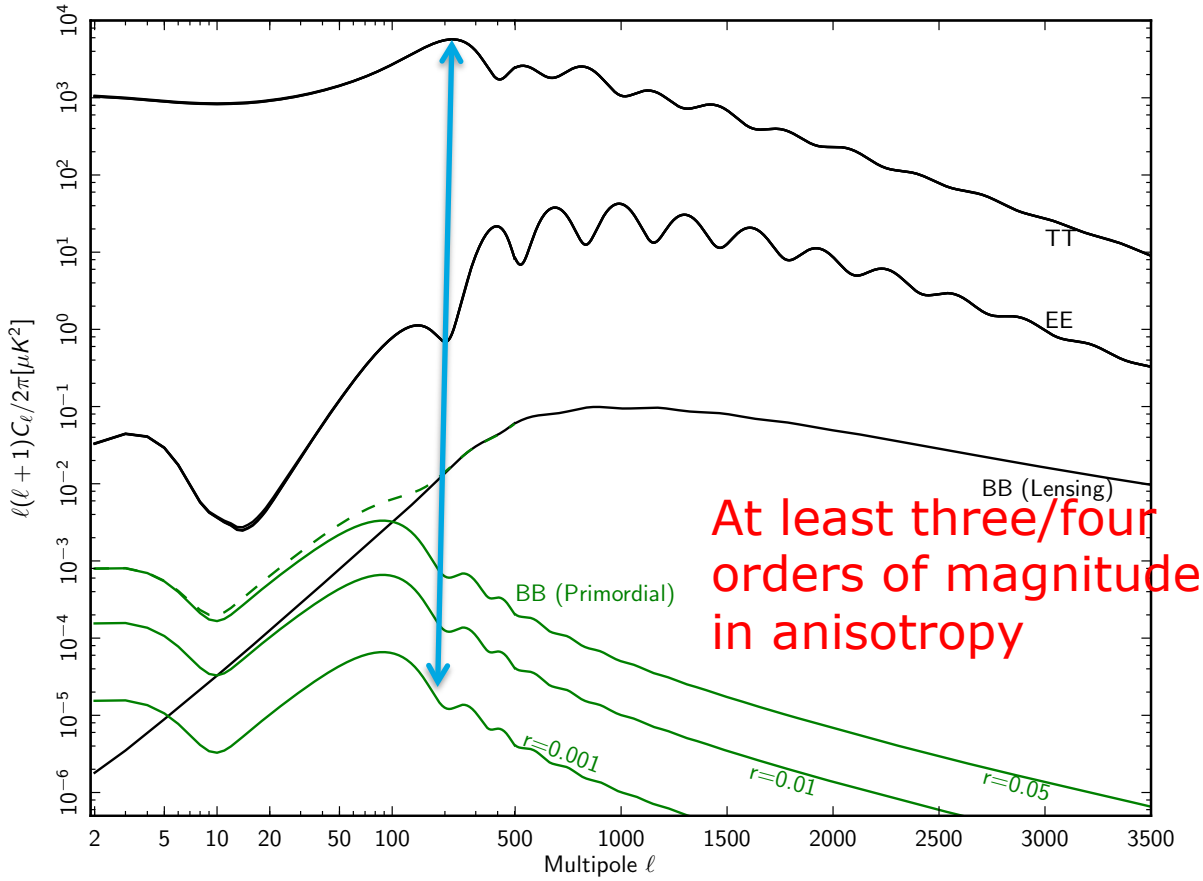
- Wide list of science targets in cosmology and fundamental physics
- … Arising from signals spread over several orders of magnitudes in amplitude and angular scale.



Some three order of magnitude in angular extension

Exploitation issues

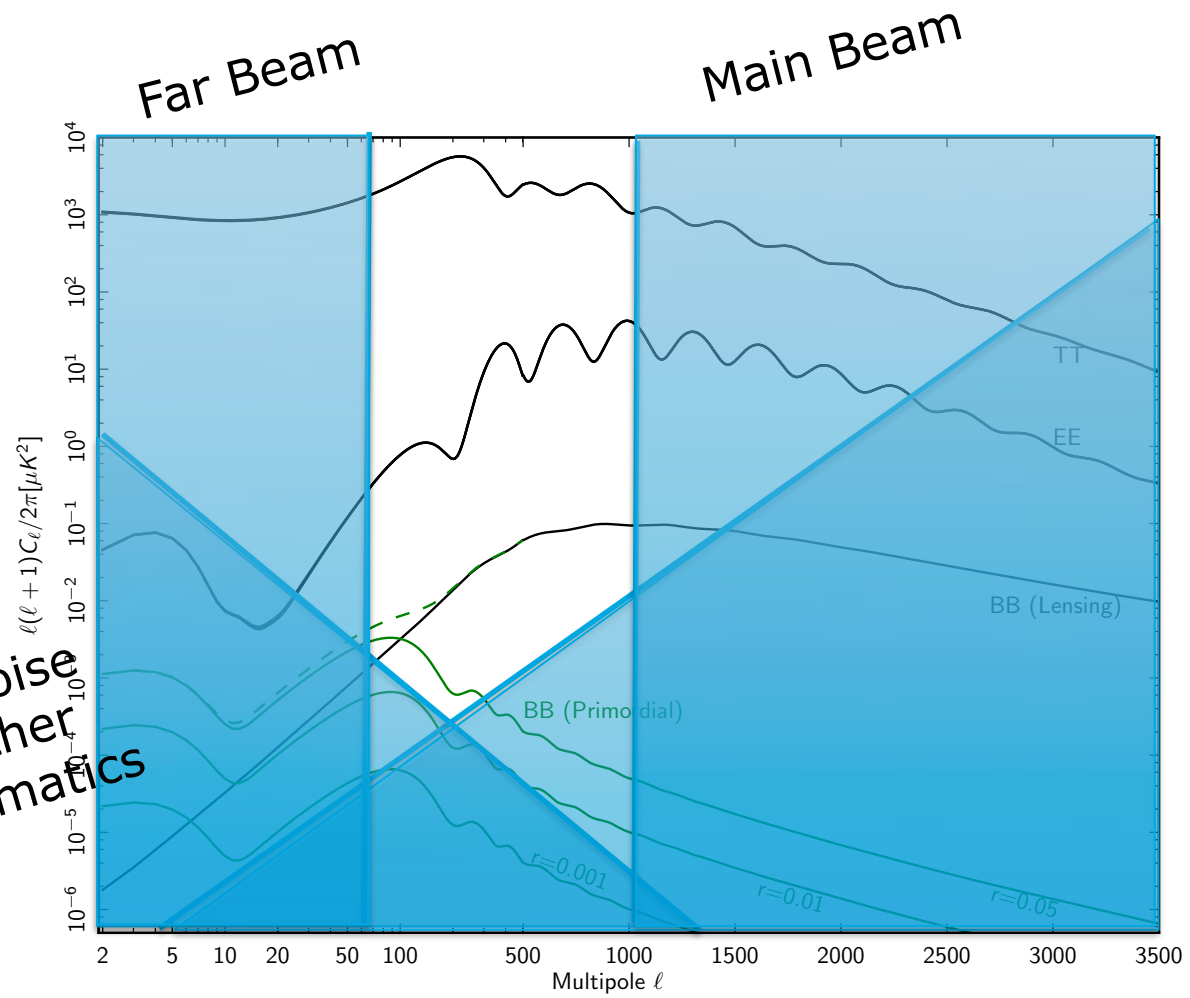
- Wide list of science targets in cosmology and fundamental physics
- Wide range of angular scales:
    - Large datasets: full sky maps (Mpix)


$\ell(\ell+1)C_\ell/2\pi[\mu K^2]$
Multipole ℓ
TT
EE
BB (Lensing)
BB (Primordial)
r=0.001
r=0.01
r=0.05
At least three/four orders of magnitude in anisotropy

Exploitation issues

- Wide list of science targets in cosmology and fundamental physics
- Wide range of angular scales:
  - Large datasets: full sky maps (Mpix)
- Signals ranking from faint to extremely faint
  - Large datasets: many detectors, long observations (Tb to Pb)

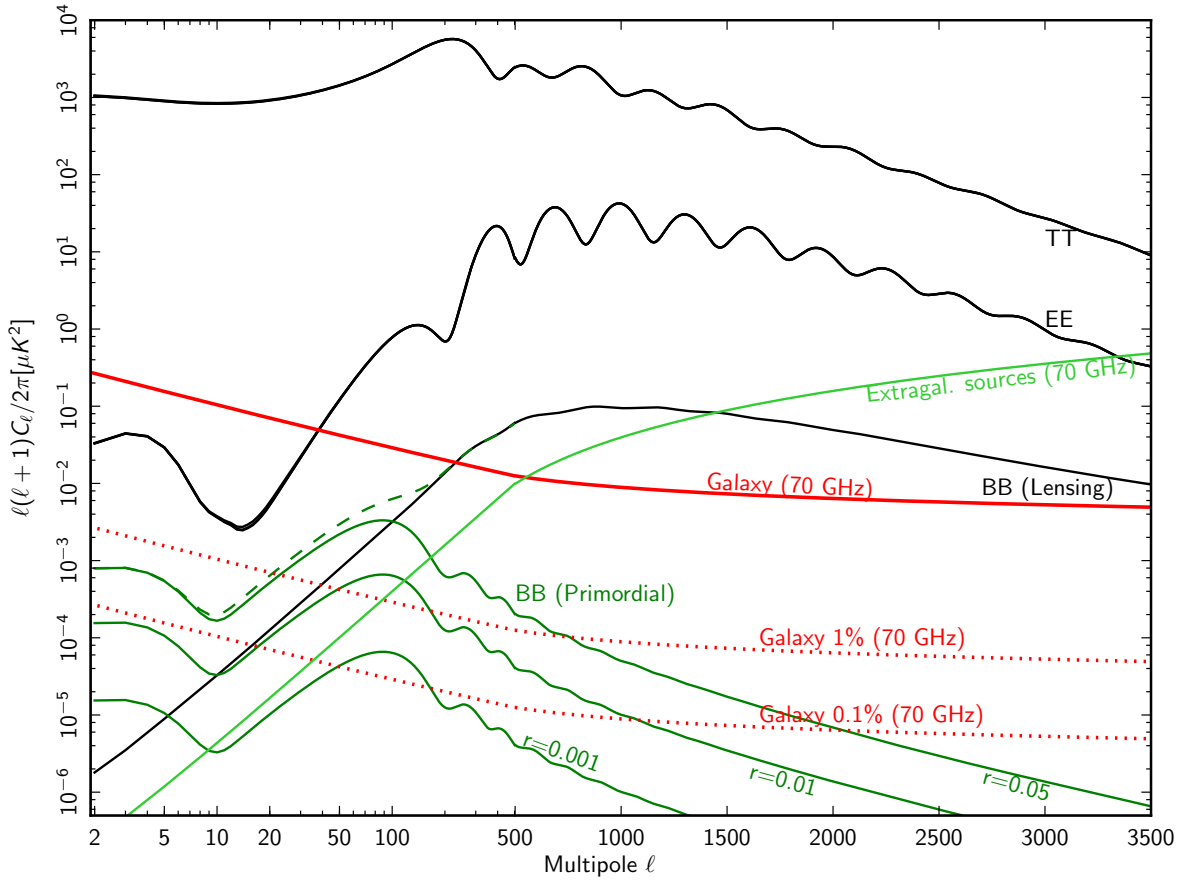
Systematics from the instrument



Correlated noise
Gain and other
nasty systematics

See talk by
J.-L Puget

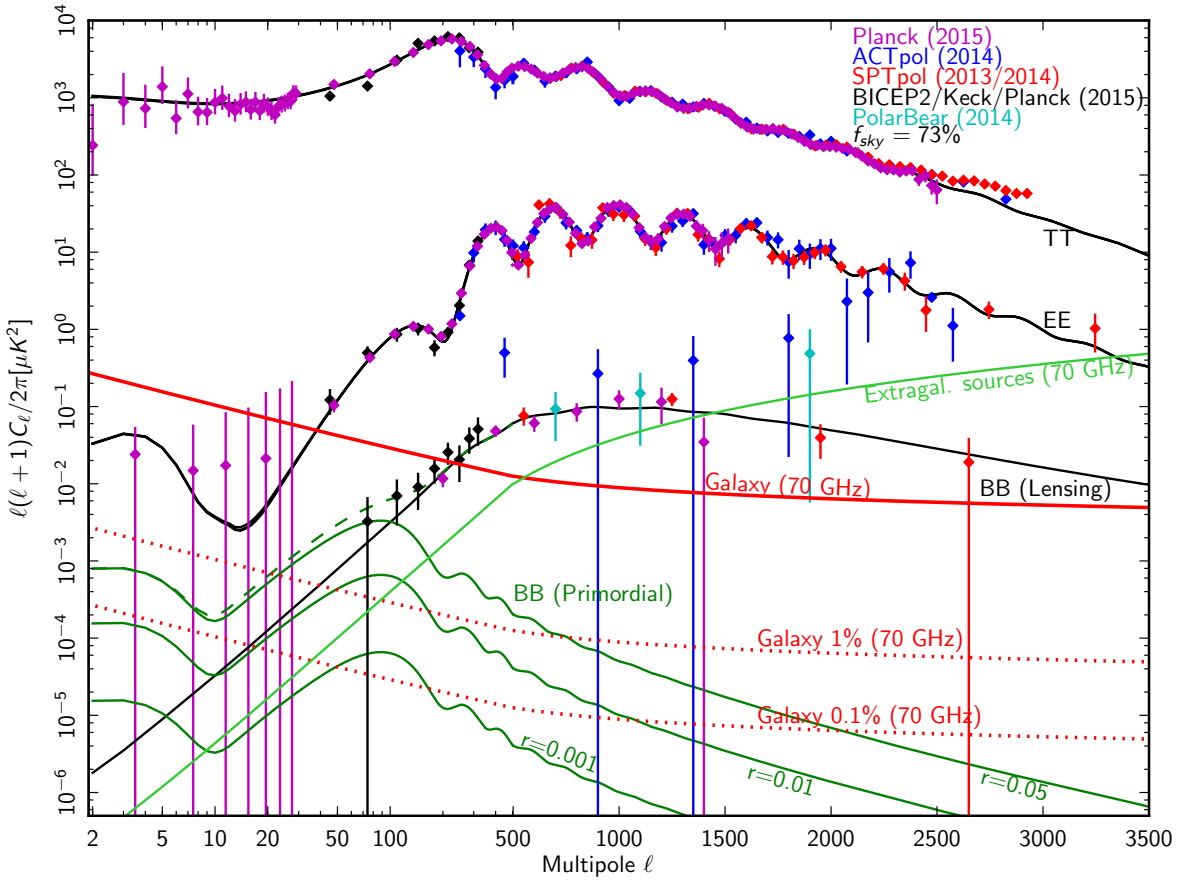
Systematics from the sky



Exploitation issues

- Wide list of science targets in cosmology and fundamental physics
- Wide range of angular scales:
  - Large datasets: full sky maps (Mega pixel)
- Signals ranking from faint to extremely faint
  - Large datasets: many detectors, long observations (Tb to Pb)
- Large, not huge. But analysis is *extremely* challenging:
  - Statistically optimal techniques needed: dense problem
  - Error budget dominated by systematics of instrumental and sky origin

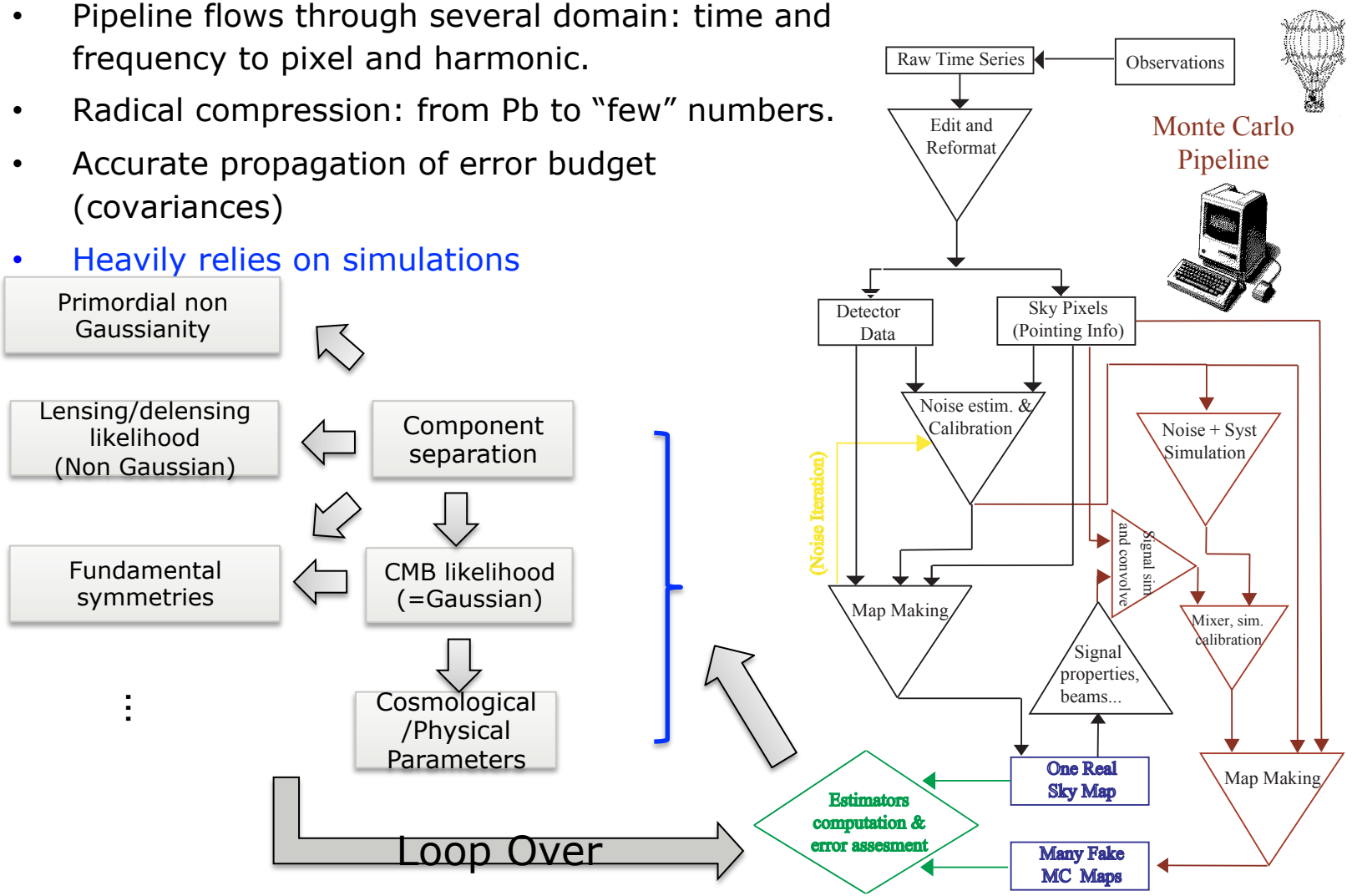
Systematics from the sky



D. Molinari+

CMB data analysis pipeline

- Pipeline flows through several domain: time and frequency to pixel and harmonic.
- Radical compression: from Pb to "few" numbers.
- Accurate propagation of error budget (covariances)
- Heavily relies on simulations



Approach vs computational cost and requirements

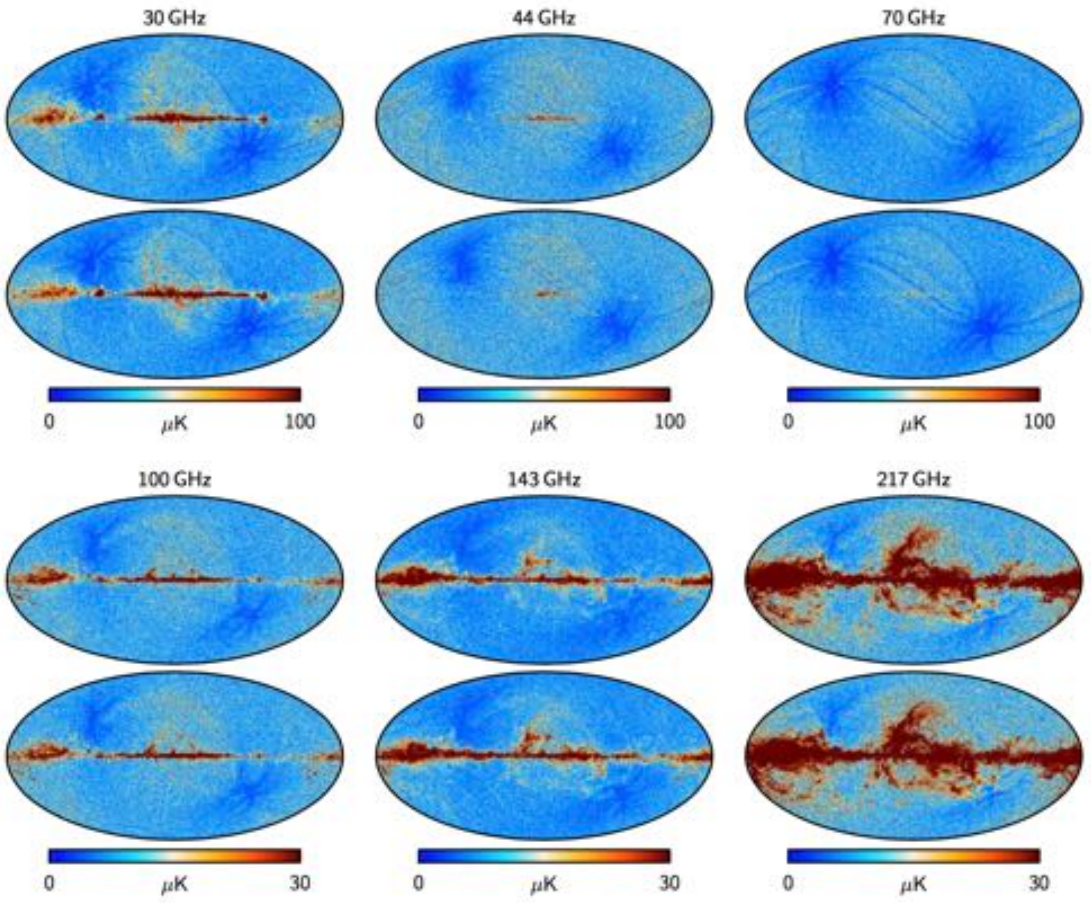
- "Exact" treatment totally unfeasible
  - ✓ Too costly (N_{pix}^3 or worse, for megapixel maps)
  - ✓ Error budget dominated by systematics anyway, already for present day experiments
- Have to rely on simulations methods
  - ✓ Computational cost dominated by simulation/map making level.
  - ✓ Scales as timeline length times number of detectors times number of simulations
  - ✓ Propagating systematics through MC is very costly and not always straightforward (c.f. Planck)
- Heavy dependence on supercomputers, precisely High Performance Computing:
  - ✓ Low latency, high bandwidth communication
  - ✓ Significant storage, fast I/O
  - ✓ No grid or share-at-home!

Use of simulations

- Pre launch mission definition:
  - ✓ To assess the impact of systematic effects (If I do not see it why worry?)
  - ✓ To validate algorithms and tools
- Actual data analysis:
  - ✓ To perform debiasing of estimators from spurious contributions (systematics and noise which cannot be modeled otherwise)
  - ✓ To assess uncertainties, including covariances that are notoriously tricky.

Planck Full Focal Plane simulations

1. End to end effort for all Planck Channels [arXiv:1509.06348]
2. Major computational burden was set of 10^4 Monte Carlo maps: 1 million CPU-days on world class super computer (NERSC and CSC)
3. Supported Planck cosmological analysis

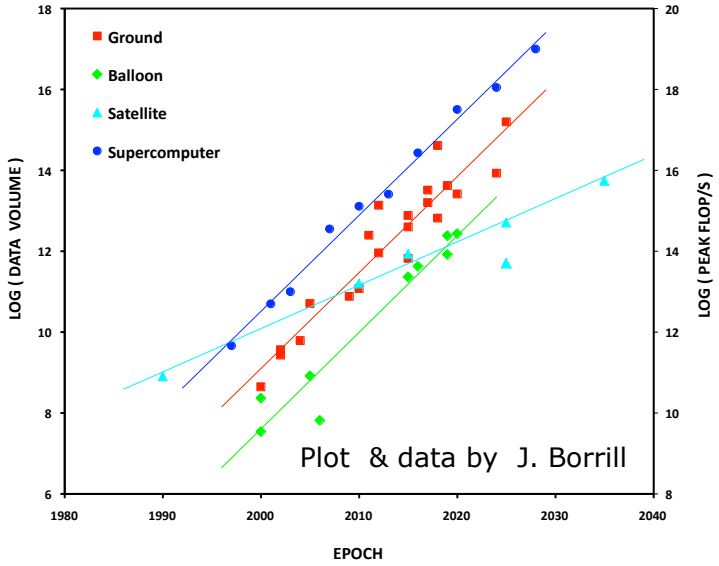


Scaling this up: Planck/FFP8 comparison to CORE

- Noise (timeline to map) Monte Carlo
  - ✓ 850 CPU-hours for single map set realization of the 52 detectors of the Planck HFI dataset (5 surveys).
  - ✓ For the minimal Core+ configuration, this would translate into about 40000 CPU-hours per map set (assuming 33% increase in sampling rate and same timeline length)**, on existing hardware.**
  - ✓ These are affordable numbers for forecasted HPC allocations even without advocating for aggressive versions of Moore's law
- CMB Monte Carlo
  - ✓ Dominated by beam convolutions
  - ✓ High number of hit per pixel for Core+ makes pixel space convolution via effective beam (a la FeBECOP) over appealing compared to harmonic methods
  - ✓ Monte Carlo generation for these methods scales with number of pixels in the beam and the map, not with timeline length (once precomputations are done).
  - ✓ More detectors means more beams, resolution (perhaps) smaller
  - ✓ The Planck FFP8 computational requirement is then in the ballpark of what we need: about 200 K Cpu-h for each 10000 MC set.

More detecotrs, higher computational needs

- Projected scaling up of computing power (based on some version of Moore's law) allows in principle to scale up to cover forthcoming ground based experiments…



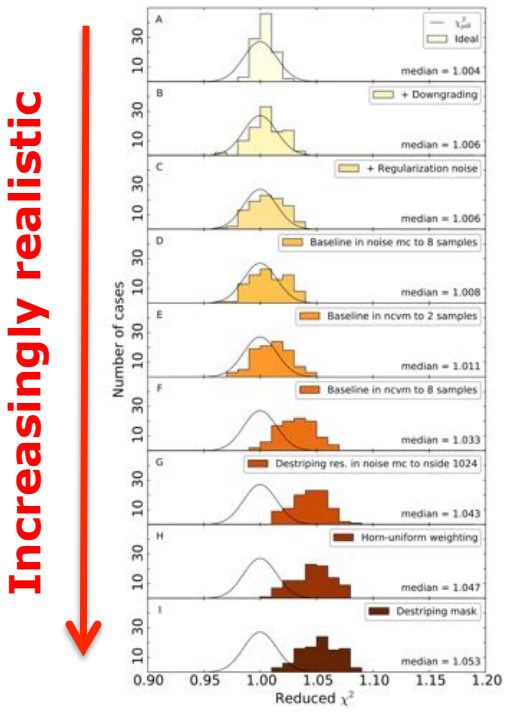
- All feasible? Beware of several things…

Issues and worries: computing power

- Difficult to forecast the evolution of supercomputing resources in the next decade, at a time when large system are already becoming severely energy constrained. Many experts speak openly of Moore's law coming to an end. We go into uncharted territory.
- Even ignoring the above, exploitation of available resources (when available) are limited by user concurrency and cost of flop unit.
  - ✓ *We are not the only community in need of significant computing power. We are already competing for resources*
  - ✓ *Must find a balance between cheap flops offered on clogged computers and costly dedicated service. Can European coordination play a role here?*
- Sheer size of data limits human direct intervention. Automatization is a must and complicates business.

Issues and worries: accuracy

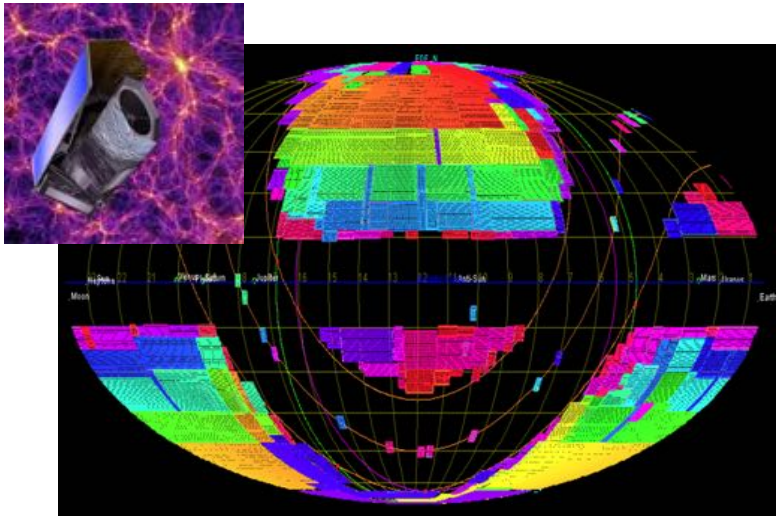
- Accuracy needs are already scaling up simulation volume. Analytic shortcuts are not always easy or found.



- Example: Planck LFI map making (arXiv:1502.01585)
- χ^2 for pixel-pixel noise covariance matrix (a fairly large guy!)
- No instrumental nor sky driven systematic
- Pure data analysis effect: the algorithm to build this matrix is just an approximation of the real stuff.
- Would be worse for bolometers/colored noise
- Exact treatment cannot be achieved analytically. Simulation driven correction is an option, but requires large number of realizations

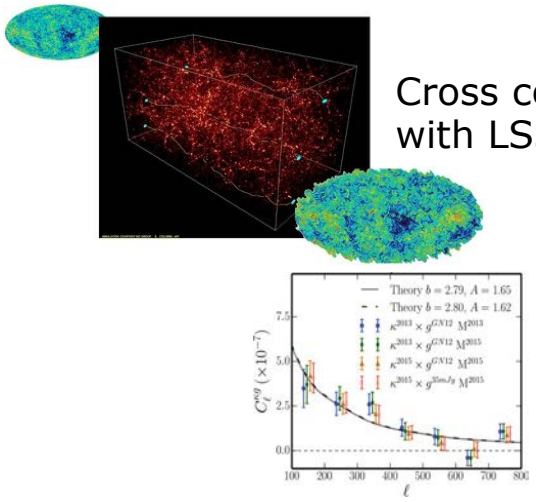
Issues and worries: data combination

- Data combination will inevitably scale up the needs and complicate analysis, especially when properly done
  - ✓ Combining final likelihood products is easy, correctly exploiting physically correlated datasets much much harder



Courtesy C. Baccigalupi

Covered CMB Fields



Cross correlation with LSS

from Bianchini et al., 2016

An historical touch

The Challenge Of Data Analysis For Future CMB Observations

J. Borrill, circa 1999

FUTURE PROSPECTS

We have seen that existing algorithms are capable of dealing with CMB datasets with at most 10^5 pixels. Over the next 10 years a range of observations are expected to produce datasets of 5×10^5 (BOOMERanG LDB), 10^6

TABLE 3. The computational requirements for one iteration of the Newton-Raphson algorithm to extract a 20-bin power spectrum for MAXIMA and BOOMERanG

Flight	N_p	Disc	RAM	Operations	Serial Time	Cray T3E Time
BOOMERanG NA	26,000	110 Gb	11 Gb	7.1×10^{14}	14 days	5 hours (64 PE)
MAXIMA 1	32,000	170 Gb	17 Gb	1.3×10^{15}	25 days	9 hours (64 PE)
MAXIMA 2	80,000	1 Tb	100 Gb	2.1×10^{16}	13 months	18 hours (512 PE)
BOOMERanG LDB	450,000	30 Tb	3 Tb	3.7×10^{18}	196 years	140 days (512 PE)

(MAP) and 10^7 (PLANCK) pixels that will necessarily require new techniques. This is an ongoing area of research in which some progress has been made in particular special cases.

A way out

- It wasn't computers that saved us but rather the development of fast and accurate data analysis tools:
  - ✓ *Fast transforms on the sphere (Muciaccia, Natoli & Vittorio 1996)*
  - ✓ *Fast pixelization scheme (HealPIX, Gorski et al 1998+)*
  - ✓ *Fast and accurate component separation (Gispert and Bouchet, 1997, …)*
  - ✓ *Fast and accurate map making (Natoli et al., 2001, Dore' et al 2001, …)*
  - ✓ *Fast power spectrum estimation (Hivon et al 2002, …)*
  - ✓ *….*
- These efforts have been severely driven by the *push* to analyze real data. We must be able to keep the trend alive. Live data is only way, cannot afford a 10 year gap.

Example of ongoing research: fast tackling of beam systematics

QuickPol: Fast effective beam matrices calculation for CMB polarization

Eric Hivon[1], Sylvain Mottet[1] & Nicolas Ponthieu[2,3]

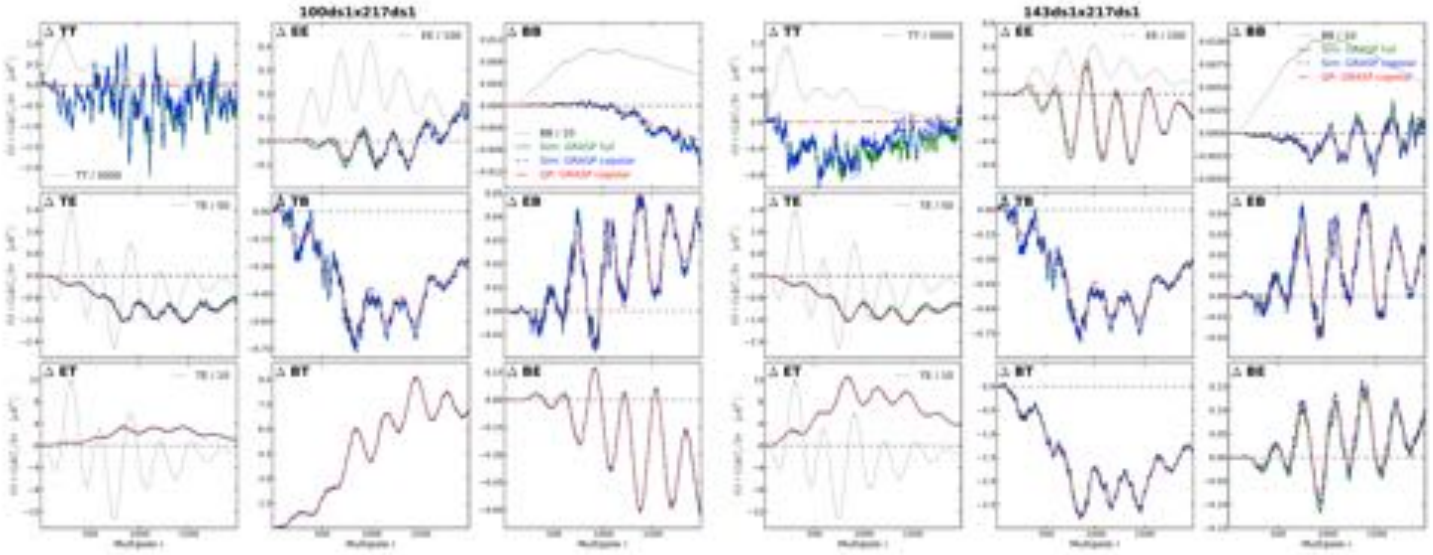


Fig. 5. Comparison to simulations for 100ds1x217ds1 (*lhs* panels) and 143ds1x217ds1 (*rhs* panels) cross power spectra, for computer simulated beams. In each panel is shown the discrepancy between the actual $\ell(\ell+1)C_\ell/2\pi$ and the one in input, smoothed on $\Delta\ell = 31$. Results obtained on simulations with either the full beam model (green curves) or the copolar beam model (blue dashes) are to be compared to `QuickPol` analytical results (red long dashes). In panels where it does not vanish, a small fraction of the input power spectrum is also shown as black dots for comparison.

arXiv:1608.08833v1

Conclusions

- Tenuous CMB signal targets imply larger, more complex datasets. Analysis requirements will scale accordingly.
  - ✓ The need to accurately exploit dataset combination will surely boost requirements up. Forecasting how much is hard.
- Efficient access to supercomputing power should be a key tier in coordination plans.
- Cannot underestimate the power of simulations but don't want to overestimate at the same time. With the systematic error budget dominating analysis, desire for advancement in methodology is growing
- Some people think theory can happily leave without data (at least for some while). Data analysts in general do not share the view: breakthroughs have historically been pushed by needs.