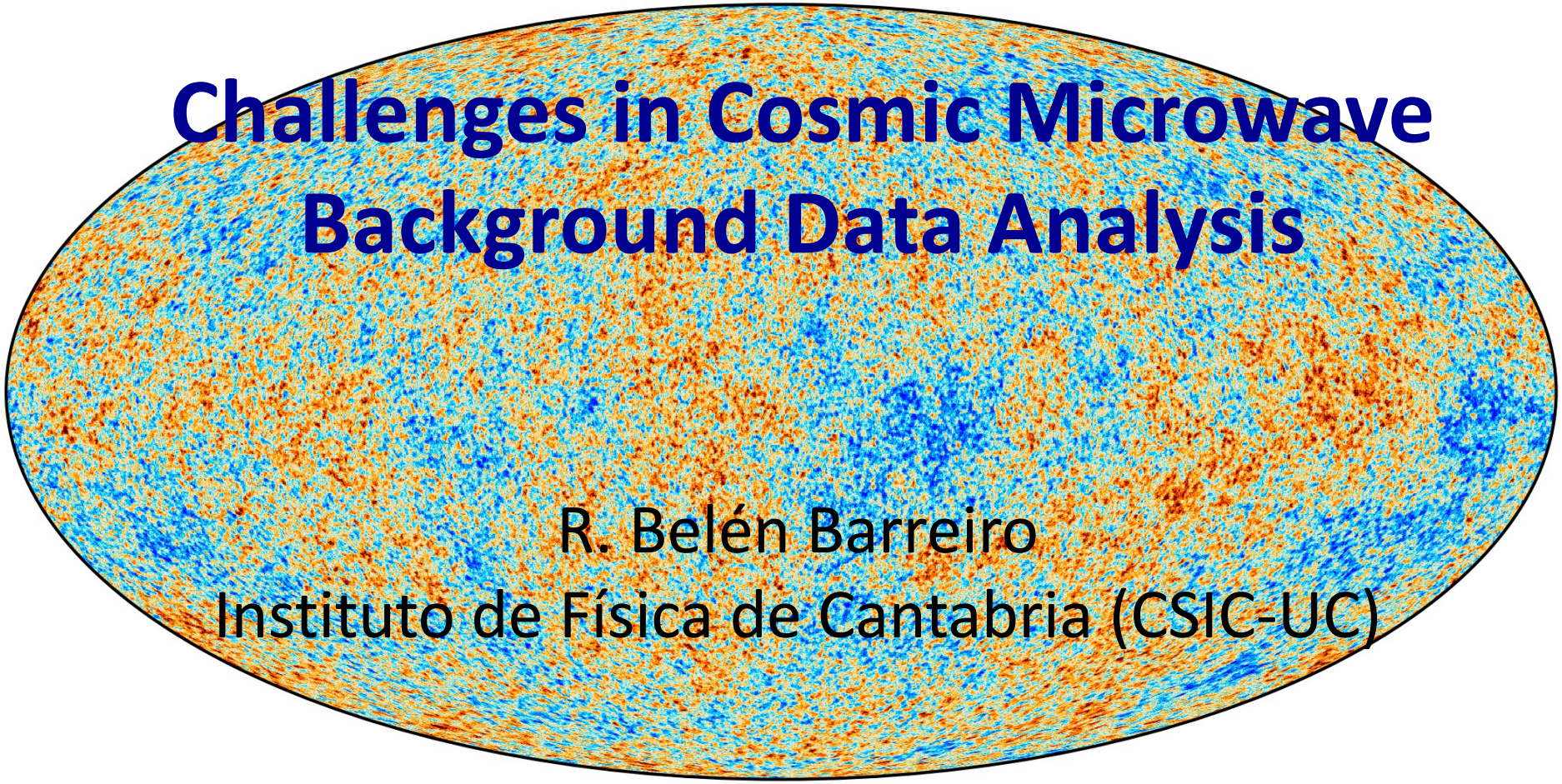


Challenges in Cosmic Microwave Background Data Analysis

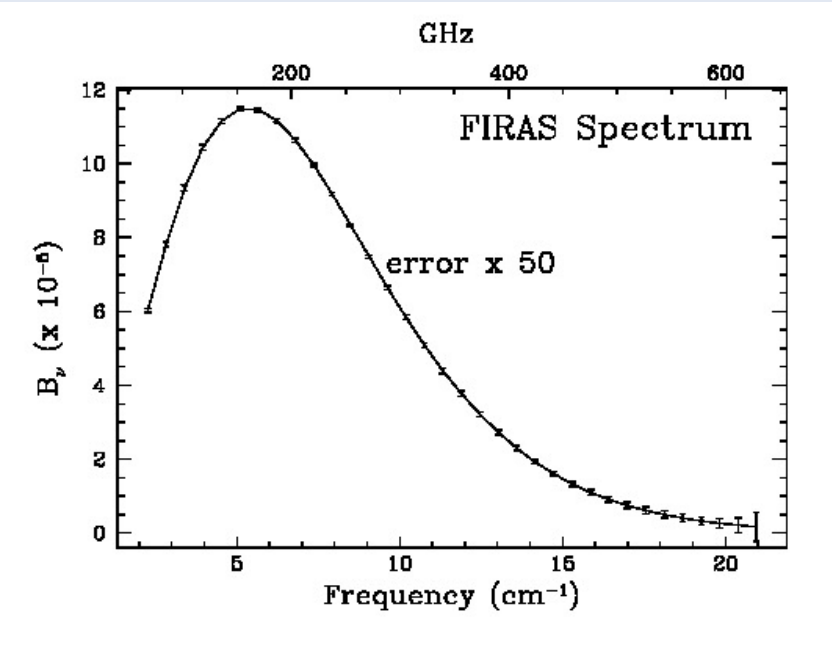
R. Belén Barreiro
Instituto de Física de Cantabria (CSIC-UC)

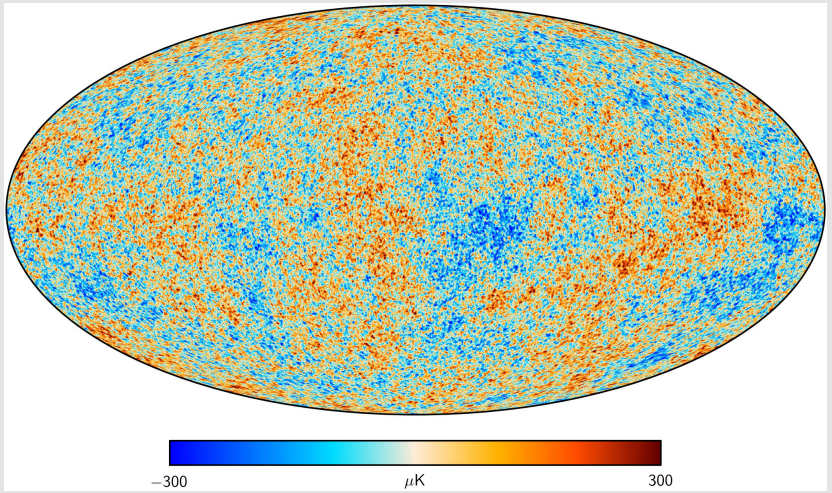
Outline

- Introduction
- Component separation problem
- Standard methodology
- Machine learning

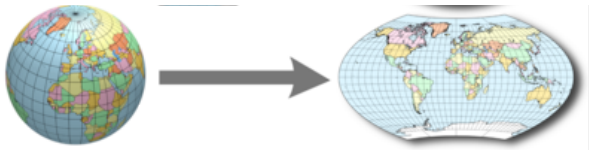
The Cosmic Microwave Background (CMB)

The **CMB** is a **homogenous and isotropic radiation** that has travelled to us shortly after the Big-Bang (when the universe **was 380,000 years-old),** with very little changes or interactions (besides its cooling or redshift due to the expansion of the universe). It has a blackbody spectrum with T_0=2.725K





The CMB presents small anisotropies at the level of ~10^{-5}, which encode a wealth of information about the early Universe, its content and evolution.
Moreover, the polarization of the CMB provides a unique probe of inflation.



Projection of the sky

Next frontier: B-mode polarization

- CMB is partially and linearly polarised
- Linear **polarisation is defined locally**, in terms of the so-called **Stokes parameters Q and U**
- Full-sky polarization maps can be decomposed into two components, the E-modes and the B-modes, (invariant under rotation) and are related to the Q and U Stokes parameters by a non-local transformation
- Primordial B-mode of polarization is sourced by gravitational waves (predicted by inflation) → **if we detect primordial B polarization, we have (indirect) proof of primordial gravitational waves !!**
- But B-mode is a extremely weak signal and appears mixed with different contaminants → very difficult to detect
- B-mode not detected yet, best constraint $r < 0.032$ (r → tensor-to-scalar ratio amplitude) [Tristram et al. 2022]

The microwave sky

$$b\,S \quad + b\,F \quad + \quad \boldsymbol{n} = \boldsymbol{d}$$

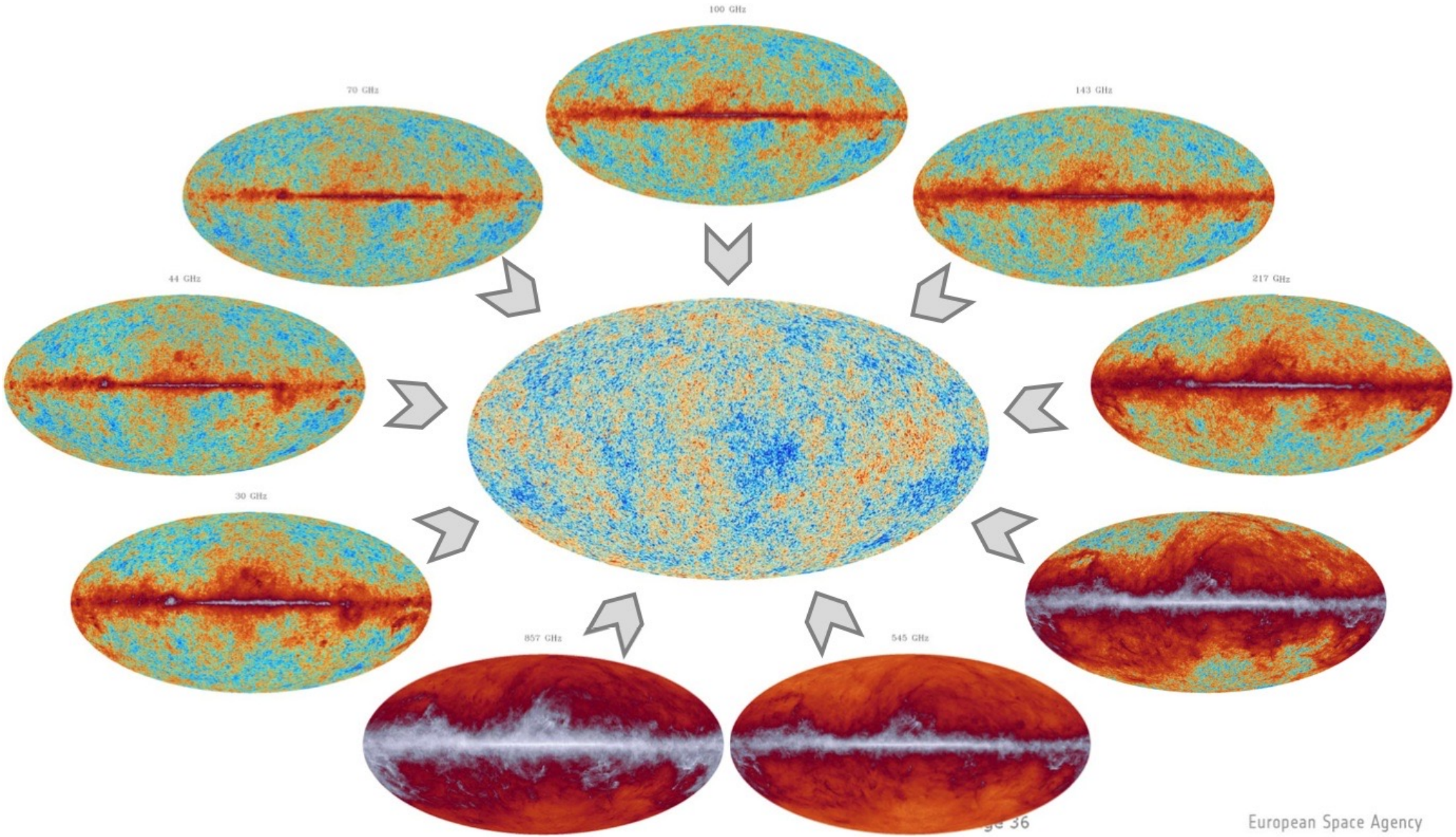
- The observed microwave sky is a combination of the CMB plus other astrophysical signals (foregrounds) along the line of sight
- The CMB and the foregrounds have a different frequency dependence
- Observe at different frequencies in order to separate the different components
- If signal not well separated → misinterpretation of the results (e.g. BICEP2, 2014)

The microwave sky

Planck: ESA satellite that provided the best full-sky CMB observations covering 9 frequencies from 30 to 857 GHz [50 million pixels per map]

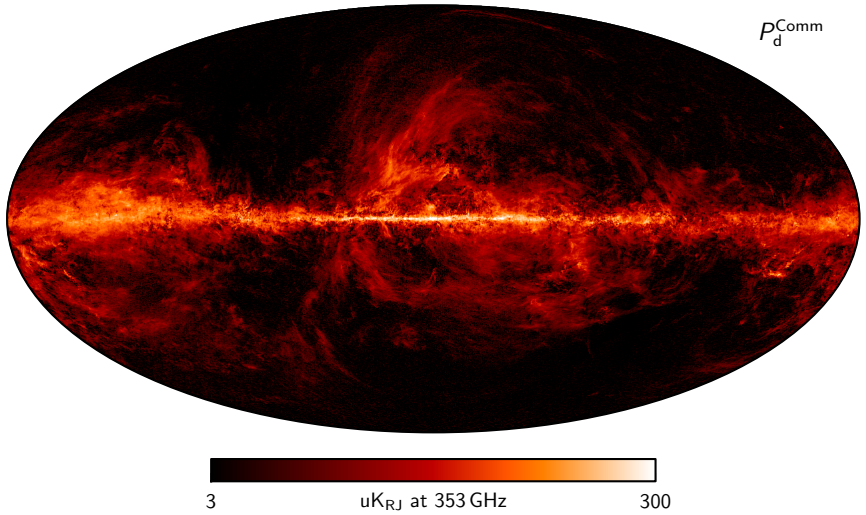
Video: https://sci.esa.int/s/A1YlJXA

Planck observations (Intensity)

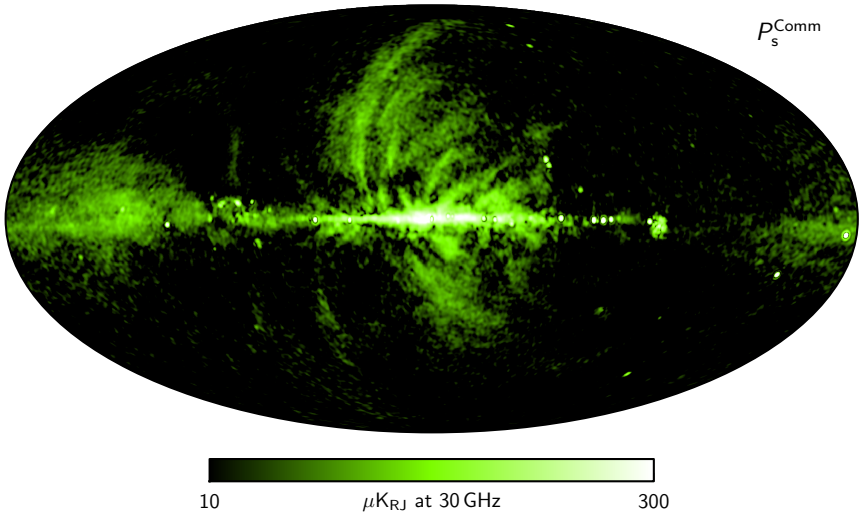


Main astrophysical contaminants

Polarized dust

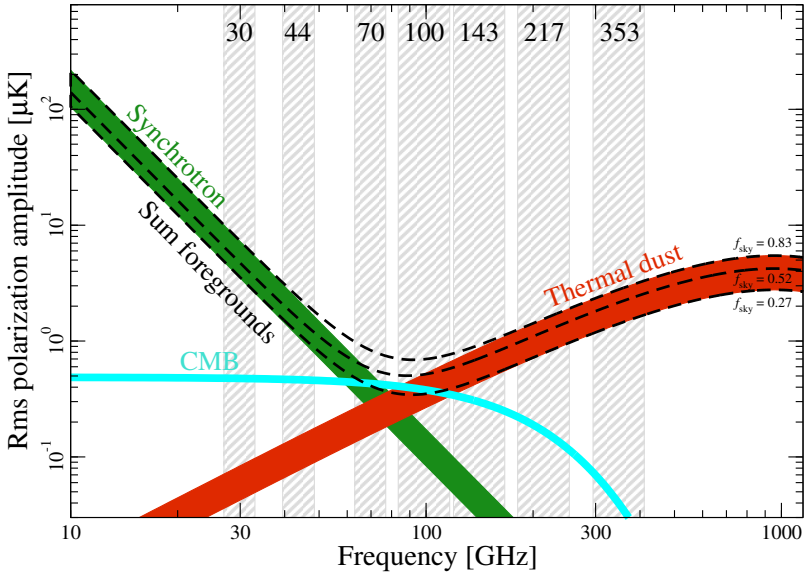


Polarized synchrotron



Diffuse emission from our galaxy

- More important at large scales
- More intense at Galactic plane
- For polarization, main emissions are synchrotron (low frequencies) and thermal dust (high frequencies)

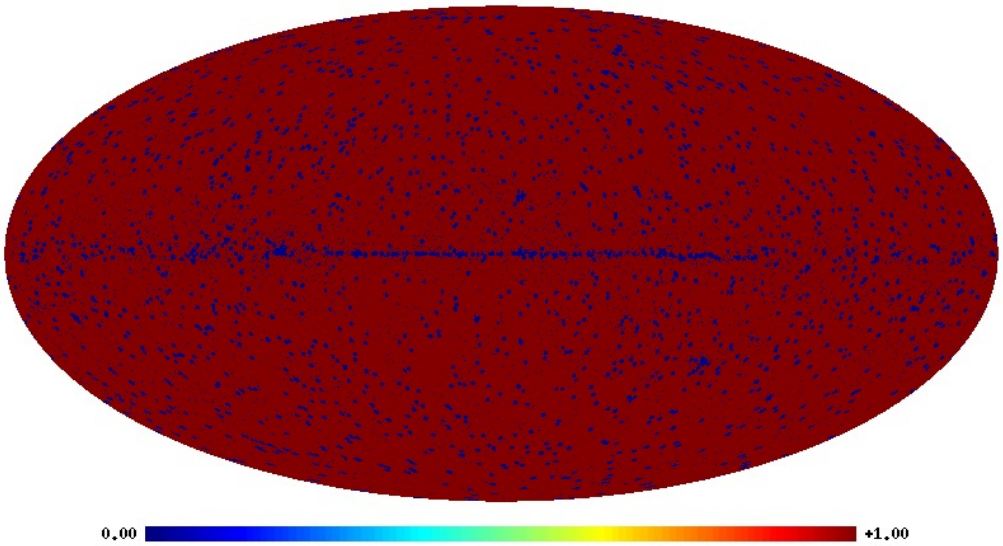


Planck 2018 results IV. Diffuse component separation

Main astrophysical contaminants

Compact (localised) sources

- More important at small scales
- Mainly emission from extragalactic point sources
  - Distributed over the whole sky
  - Point-like objects ⇨ beam profile
  - Different frequency dependence for each source ⇨ not well suited for global separation techniques
  - Polarised but in principle less relevant for B-modes (main signal of B-modes is at large scales)



Position of point sources in intensity for Planck frequencies [from M. López-Caniego]

How to deal with component separation

- Many different approaches:
  - To focus on extracting one component (e.g. CMB, point sources)
  - To recover all component at the same time
  - Blind methods (make minimal assumptions about the components)
  - Parametric methods (require a model of the components)
  - May work in real, harmonic or wavelet space
  - ….

Methodology

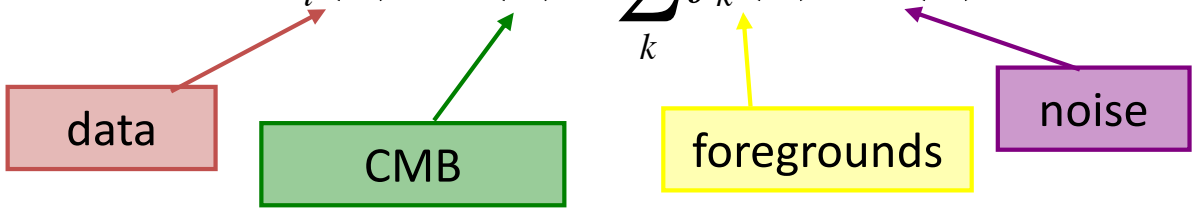
- Linear combination of data
  - Internal lineal combination (e.g. ILC, Bennett et al. 2003, NILC, Basak and Delabrouille 2013, GNILC Remazeilles et al. 2011)
  - Template fitting (e.g. Sevem, Fernandez-Cobos et al. 2012)
- Blind/semi-blind methods
  - SMICA [Cardoso et al. 2008]
- Non-blind methods
  - Parametric fitting (e.g. Commander, Eriksen et al. 2008; B-SeCRET, de la Hoz et al. 2020)
- Optimal filtering (focused on compact sources)
  - Matched filter [Tegmark & de Oliveira-Costa 1998]
  - Mexican Hat Wavelet (MHW_2) [López-Caniego et al. 2006]
  - Filtered fusion [Argüeso et al. 2009]
- Useful to have different methods to test robustness and quality of the reconstruction

See also: Leach et al. (2008) for a comparison of methods
Herranz & Vielva (2010) for a tutorial on compact object detection

Linear Combination

- Let us assume that we have a set of observations d_i at different frequencies

$$d_i(x) = s(x) + \sum_k f_k(x) + n(x)$$



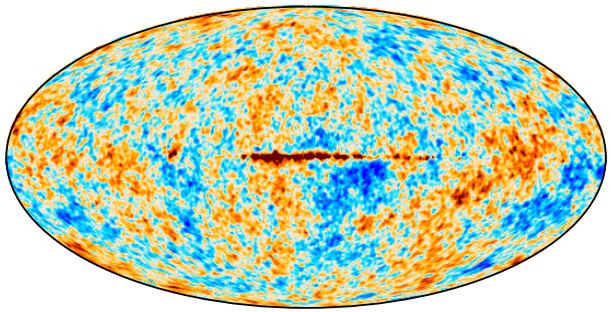
- Since the CMB is constant over frequency, if we are interested on recovering only this component, a natural solution is to make some kind of average of the different channels, which preserves the CMB signal while minimising the error

$$\hat{s}(x) = \sum_{i=1}^{N_\nu} w_i d_i(x)$$

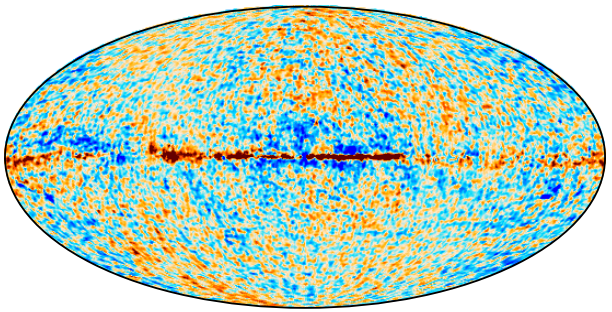
- A common approach is to obtain the w_i coefficients by minimising the variance of the reconstructed map subject to the constraint $\sum_{i=1}^{N_\nu} w_i = 1$

Linear combination: Planck results

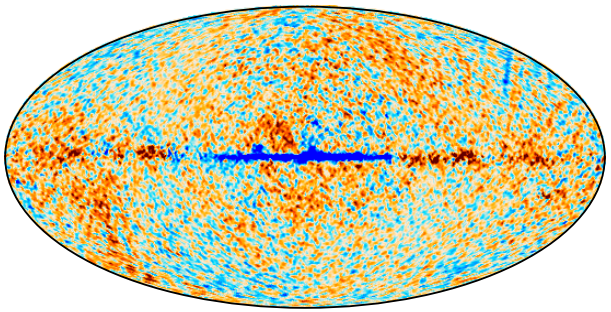
SEVEM I



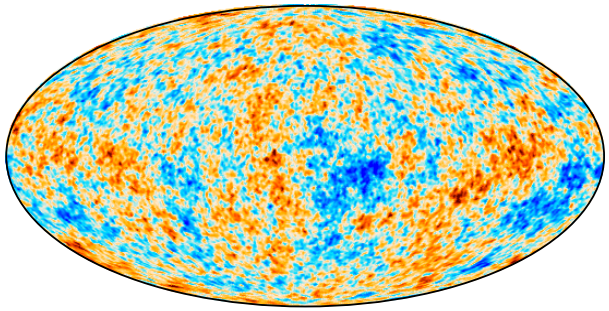
SEVEM Q



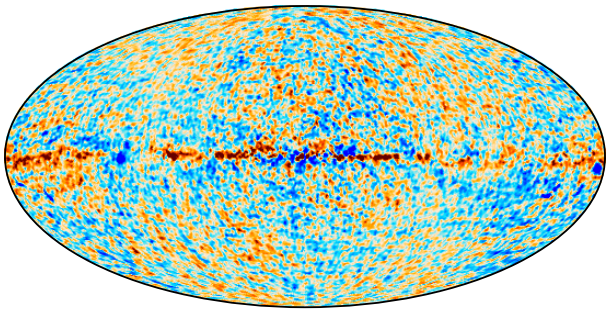
SEVEM U



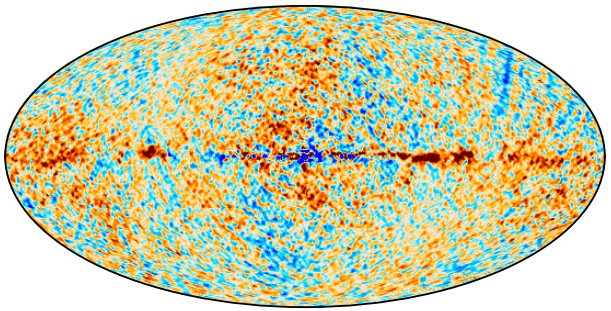
NILC I



NILC Q



NILC U



- No assumptions required → robust method
- Very fast → run through simulations to quantify errors
- Is this sufficient for future high-sensitive CMB data ?
- Only useful for components with well-known frequency dependence

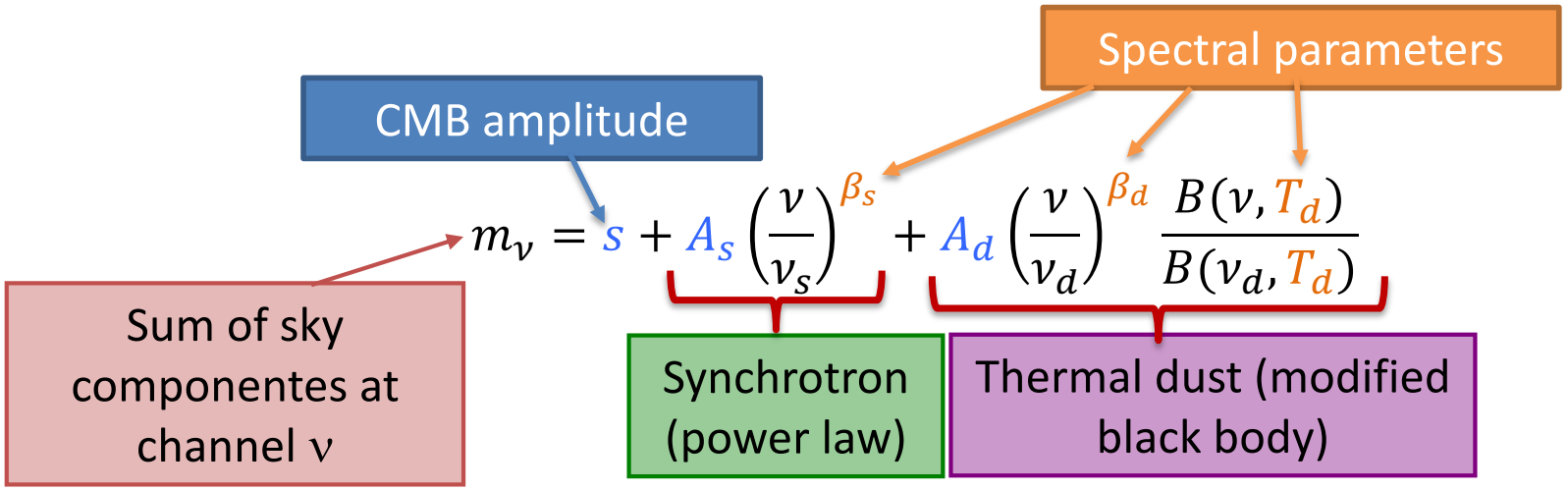
Bayesian parametric methods

They fit an explicit parametric model m(θ) to a set of observations d by evaluating the posterior distribution

$$P(\theta|d) \propto L(d|\theta)P(\theta)$$

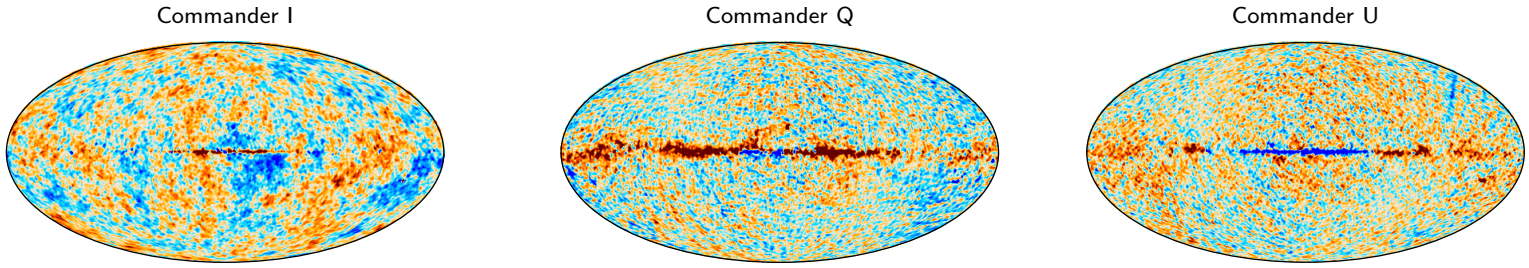
where L(d|θ) is the likelihood of data, P(θ) a set of priors and θ the free parameters of the model

For polarization, the model typically includes CMB, synchrotron and thermal dust emission, for each frequency channel, we have

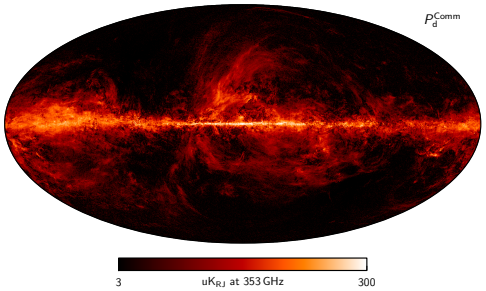


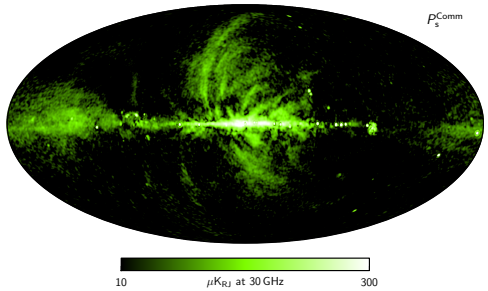
Note that both the amplitude and spectral indices can vary spatially, what allows to estimate the spectral dependence of the components pixel by pixel

Bayesian parametric methods: Planck results



Polarized dust





Polarized synchrotron

- Parametric methods are very powerful since take into account information from the different components
- They can recover all the foreground components
- Need modelisation of the different components → less robust
- Errors due to mismodelling difficult to quantify
- They can be very demanding regarding computational resources

Point source extraction

- Point sources are localised, individual objects, distributed over the whole sky, that present different frequency dependences
- The diffuse separation methods are generally not well suited to deal with emission from point sources
- The most common approach is to perform a pre-processing step of the data, where these objects are detected
- The objects can then be masked, inpainted or subtracted before carrying out the diffuse foreground separation

Point source extraction: linear filtering

- Linear filtering is commonly used to enhance the signal versus the background
- The filtered image w(x) can be obtained as the convolution of the data d(x) with the filter $\psi(x)$:

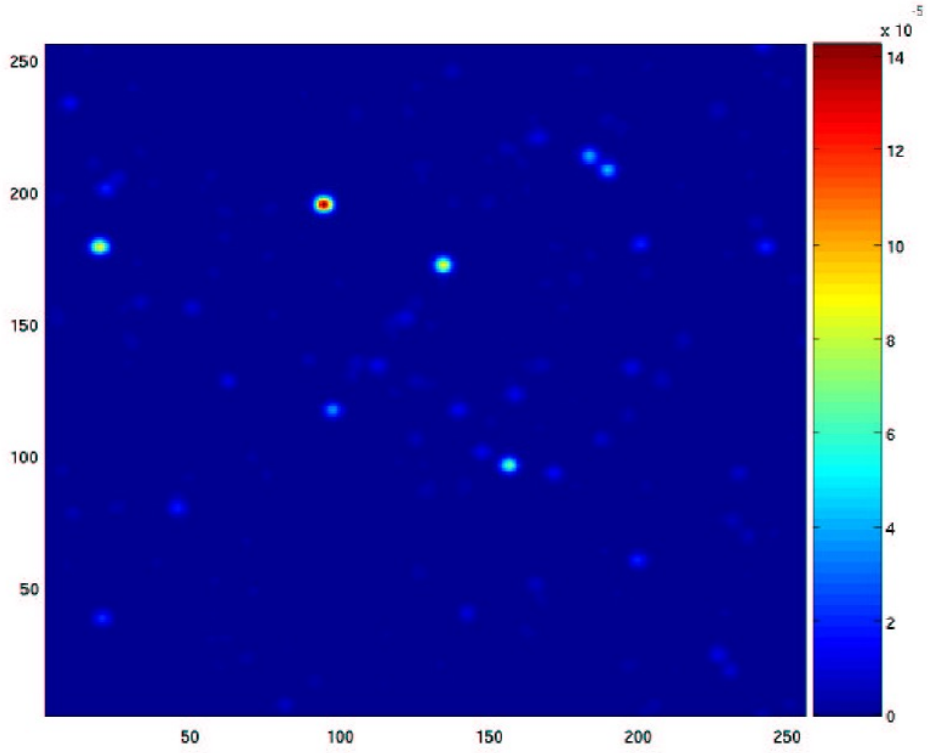
$$w(x) = \int d(u)\psi(x-u)\,du$$

- Those parts of data that follow the shape of the filter are enhanced
  ⇨ ψ should resemble the sought signal

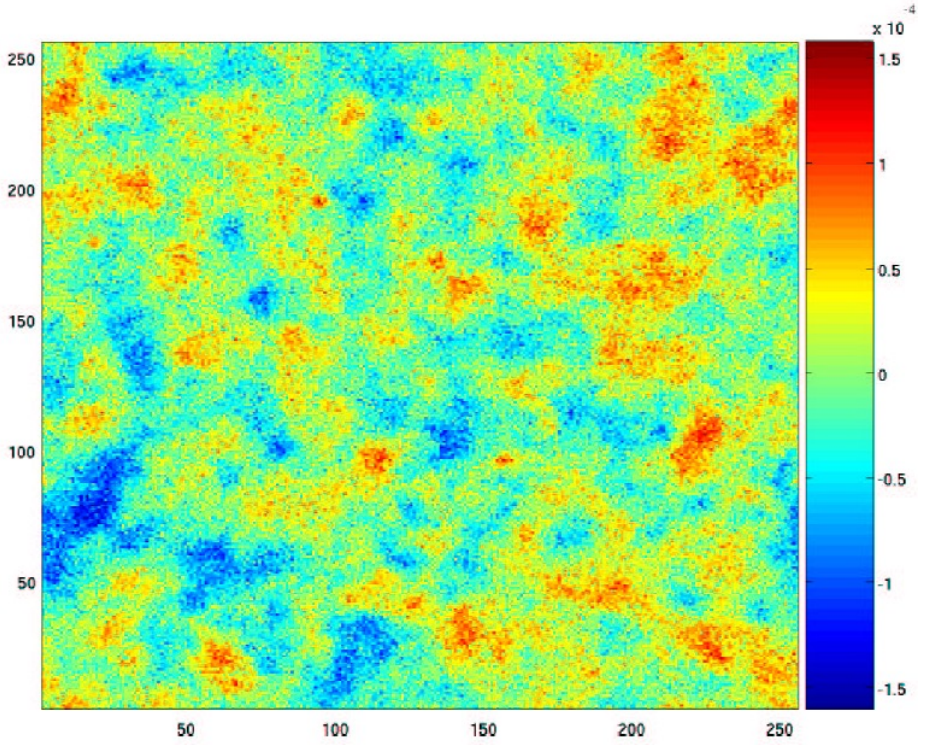
- Direct convolution is very CPU-time consuming ⇨ filtering is usually performed in Fourier (or spherical harmonic) space
- Some common filters are the Matched filter, the Mexican Hat Wavelet 2 or, specifically for polarization, the Filtered Fusion

Example: filtering with the MHW_2

Simulated Planck 70 GHz channel filtered with the MHW_2 at the optimal scale (intensity)



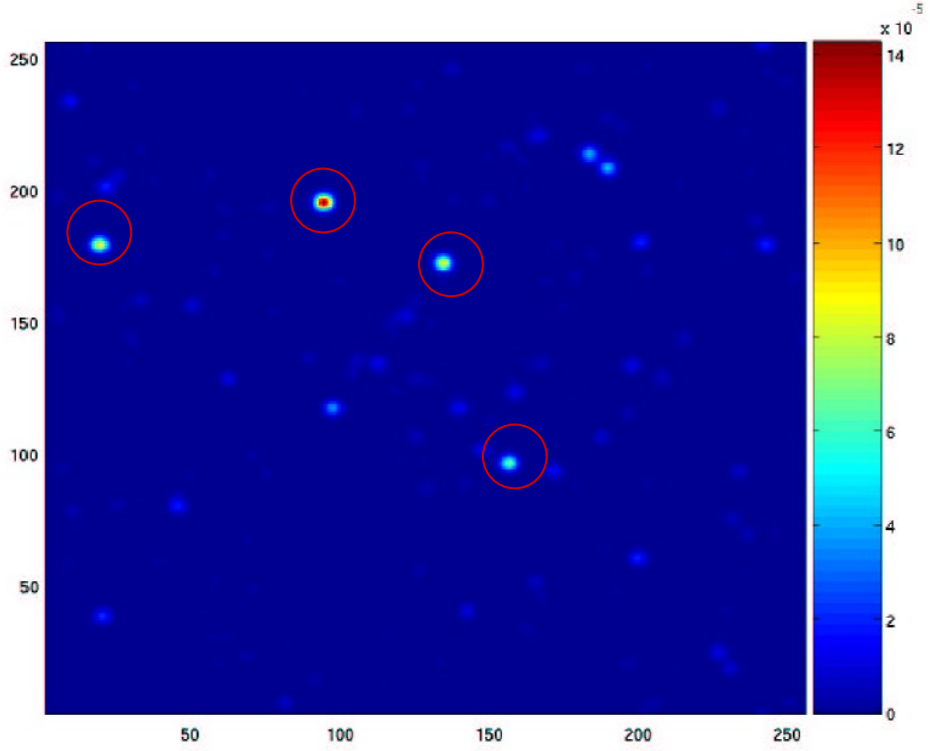
Input point sources



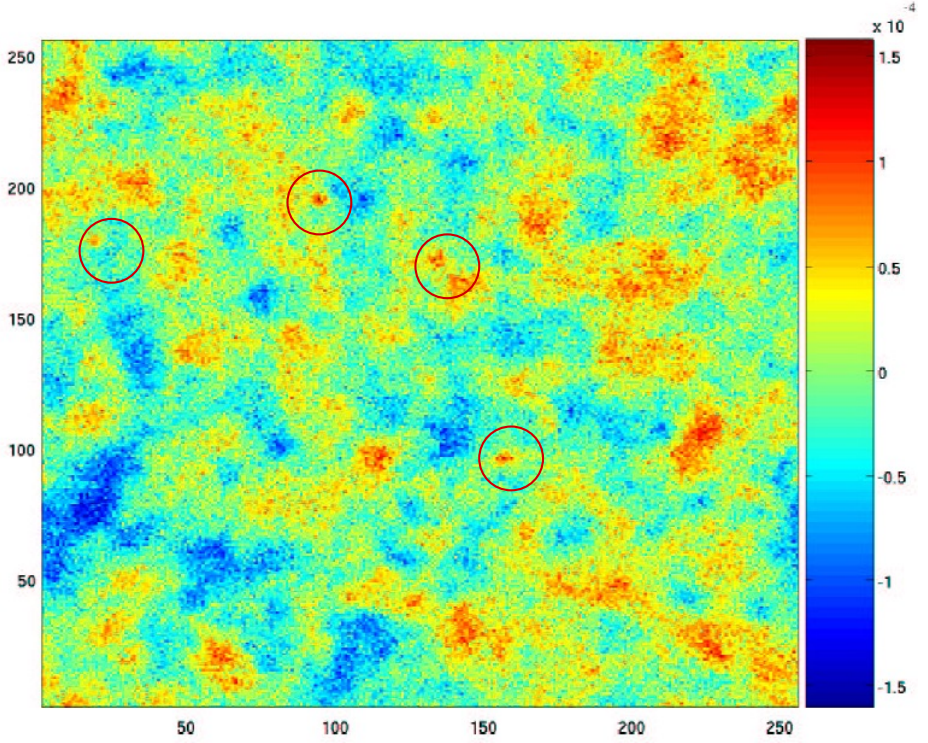
Simulated data at 70 GHz

Example: filtering with the MHW_2

Simulated Planck 70 GHz channel filtered with the MHW_2 at the optimal scale (intensity)



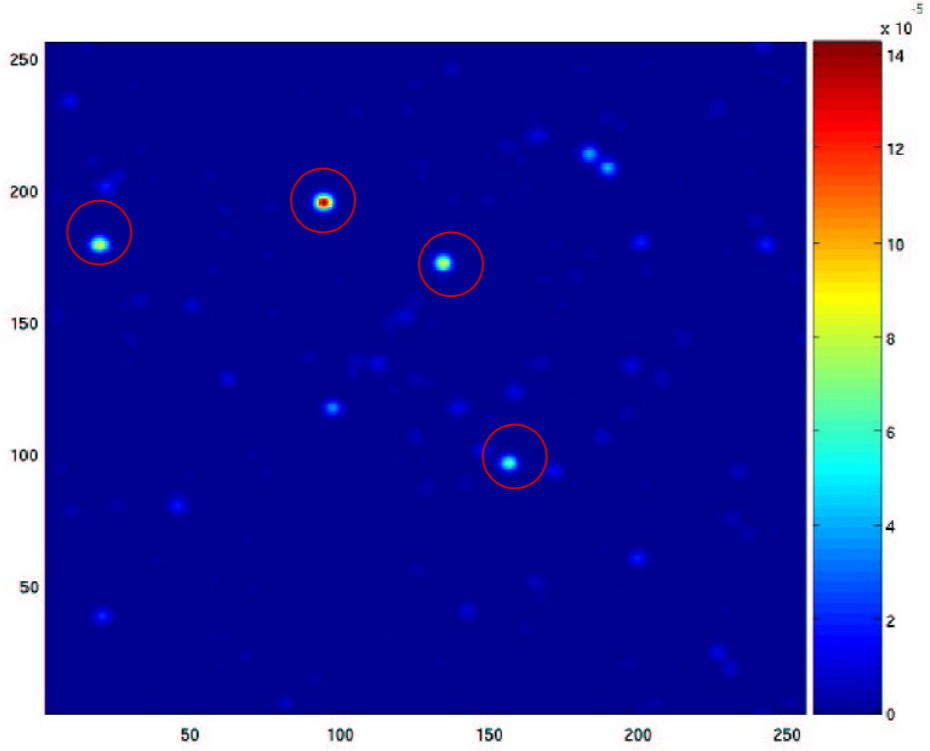
Input point sources



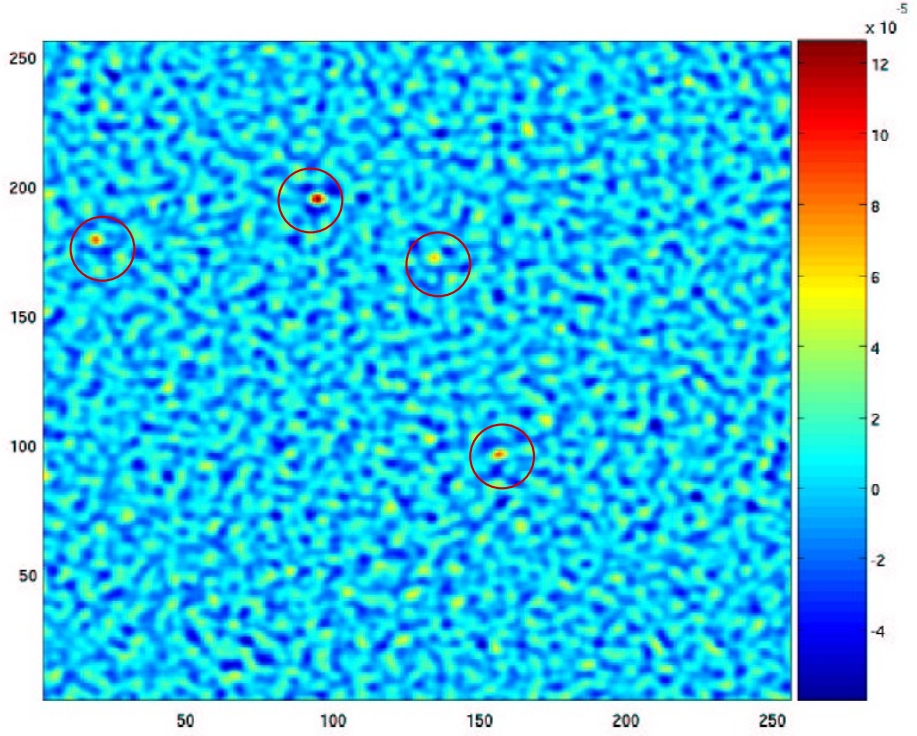
Simulated data at 70 GHz

Example: filtering with the MHW2

Simulated Planck 70 GHz channel filtered with the MHW2 at the optimal scale (intensity)



Input point sources



Simulated data at 70 GHz

López-Caniego et al. 2006

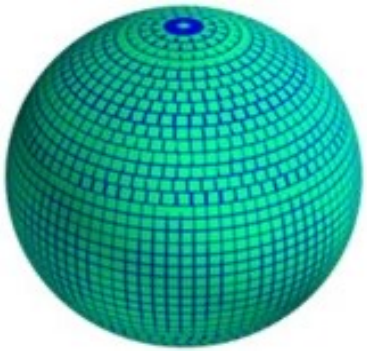
ML-based methods for CMB applications

- Mainly in the recent years, some ML-based techniques have been explored for different CMB applications including diffuse component separation, point source extraction or foreground characterization [e.g. Petroff et al. 2020, Casas J.M. et al. 2022, Farsian et al. 2020]

- On our side:
  - Working on a point souce detection technique [Diego-Palazuelos et al.]
  - Preliminary study for diffuse component separation
    - asuming a parametric model but using NN pixel by pixel to estimate the amplitude and spectral parameters → promising results, much faster than sampling [Casaponsa et al.]

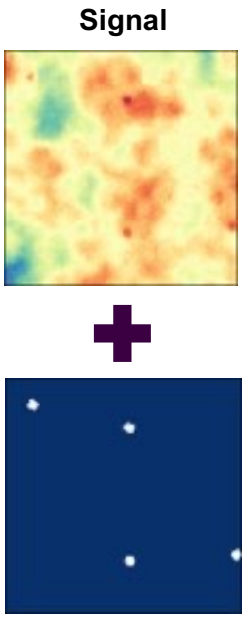
- Still a long way to go

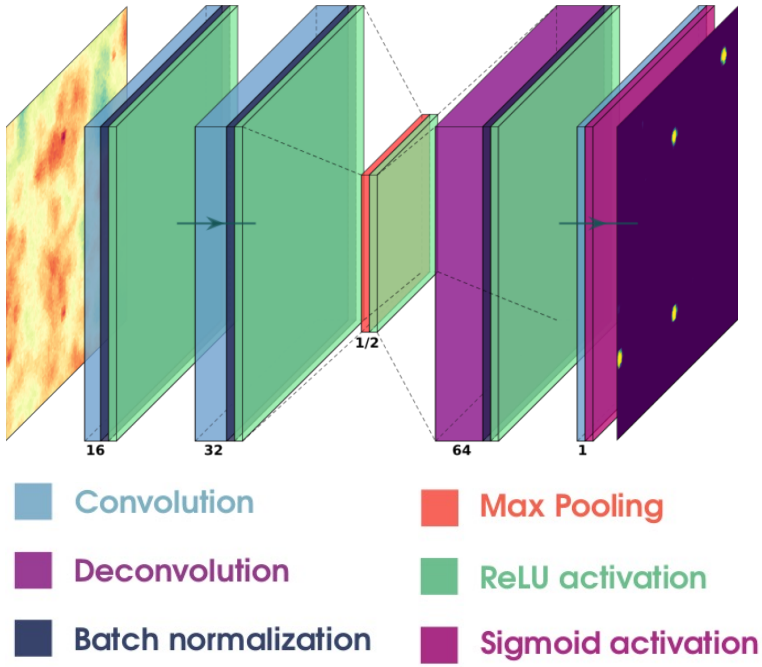
Point Source detection with ML



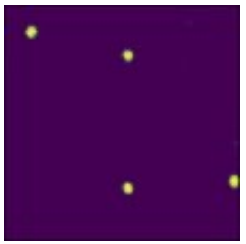
3.63° × 3.63°
2D patches

Treat the blind detection of extragalactic point sources (PS) on CMB temperature maps as an image segmentation problem





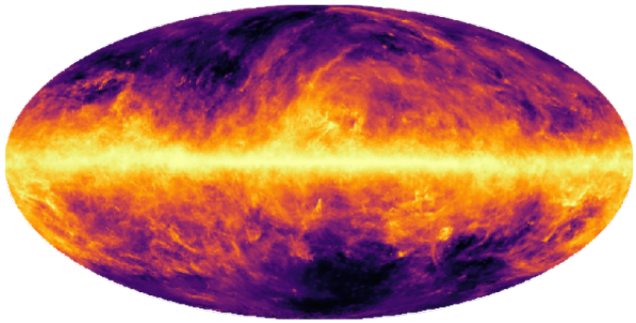
CNN prediction



Diego-Palazuelos et al. 2022

Specialized CNNs for progressively higher foreground emission

Galactic foregrounds @ 143 GHz

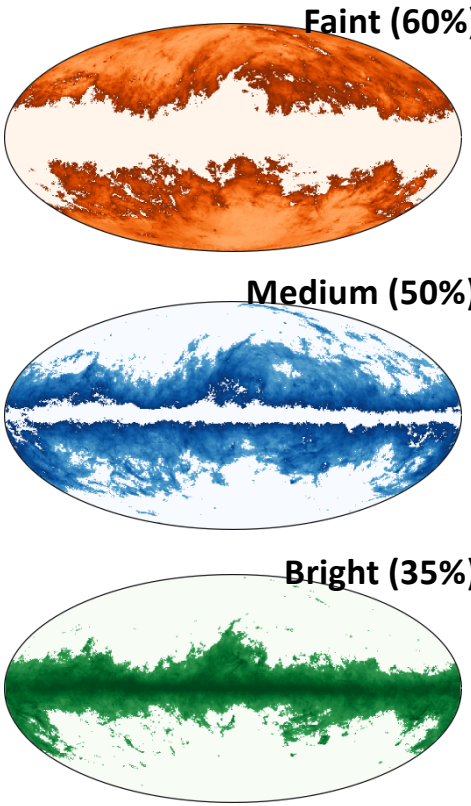


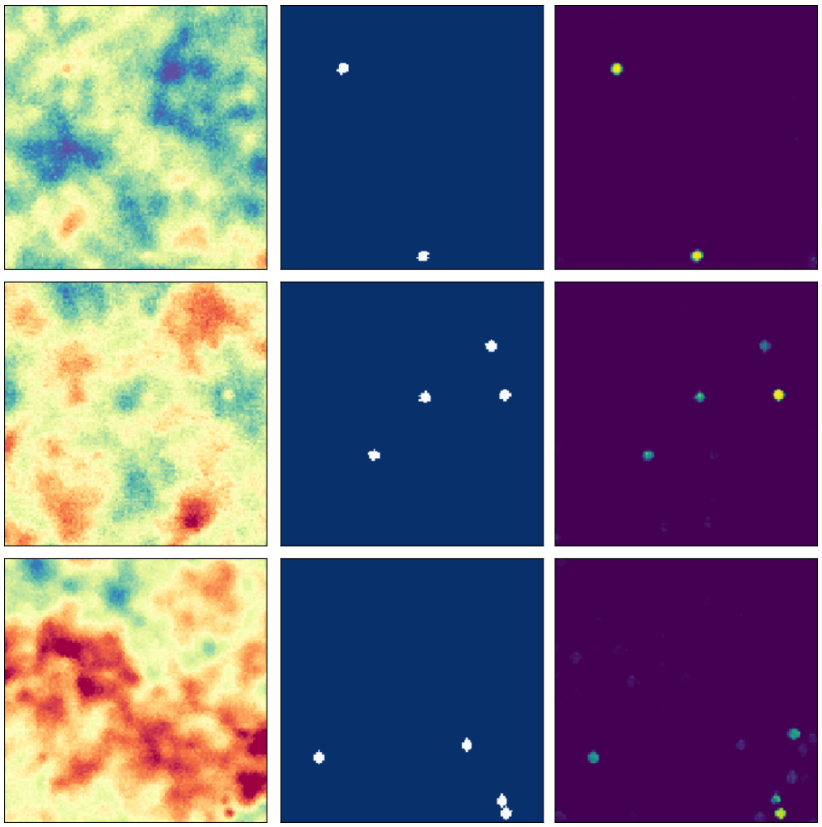
- **free-free**
- **synchrotron**
- **Spinning & thermal dust**

+

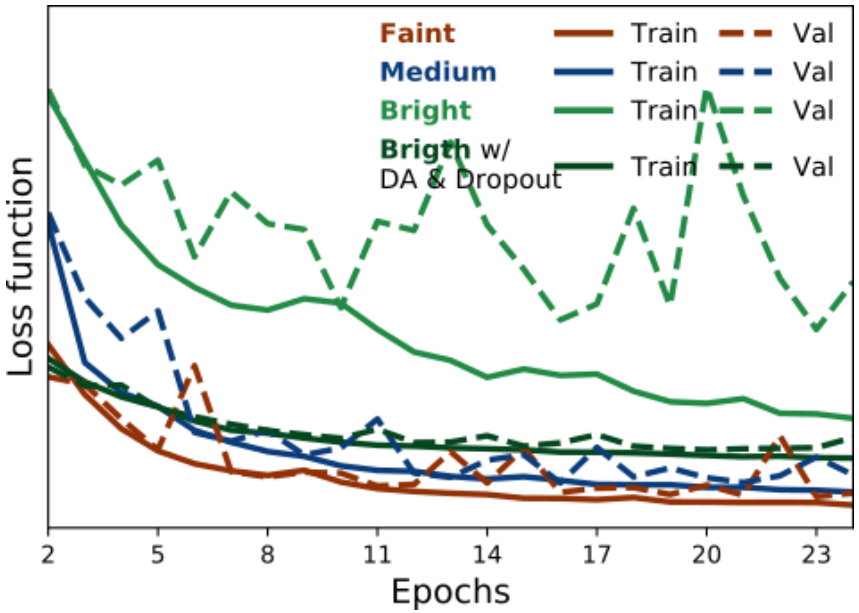
Extragalactic foregrounds

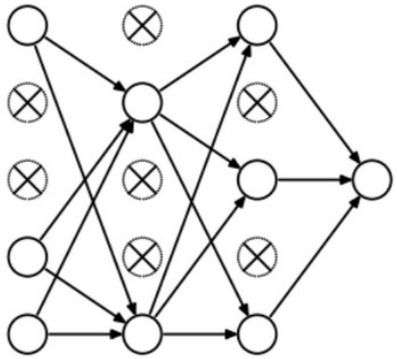
- **thermal & kinetic SZ**
- **undetectable PS (<177 mJy)**





Mitigating over-fitting with...

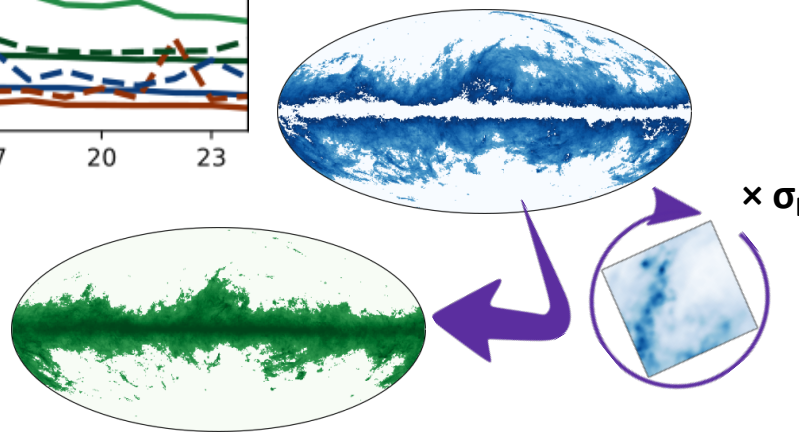




Dropout

+

Data augmentation



$\times \sigma_{Brigth}/\sigma_{Medium}$

Increase the size of the Bright training set by adding rotated and re-scaled patches from the fraction of sky in the Medium region that doesn't overlap with the Bright

Comparison with traditional methods

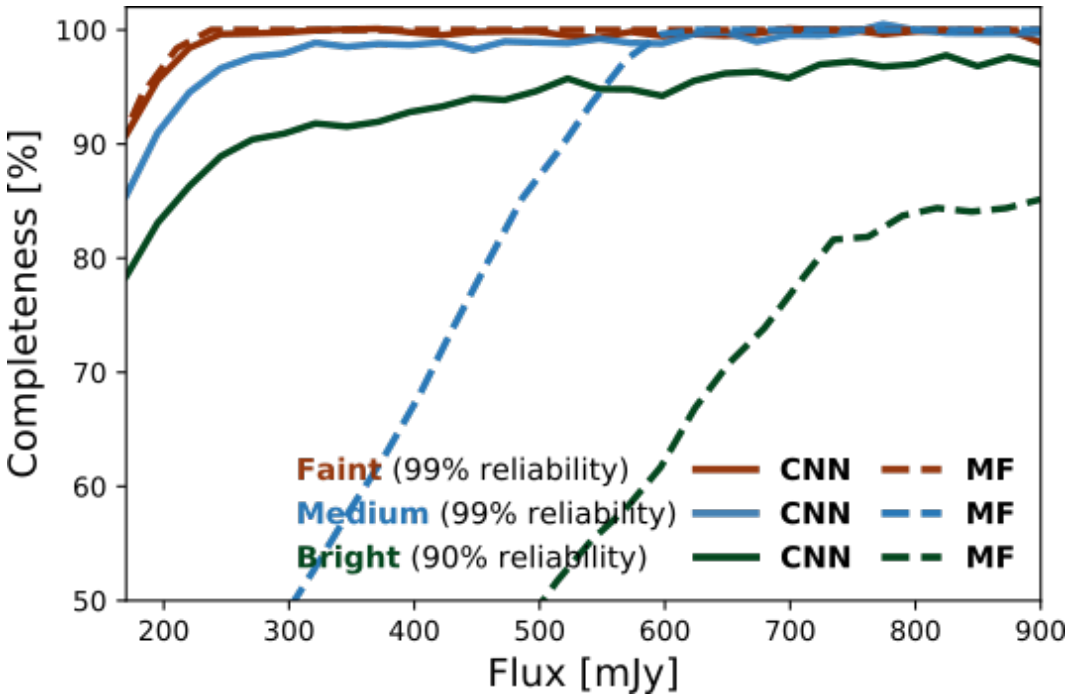
Conventional approach

Source detection through the combination of a matched filter (MF) and a detection criterion

MF ≡ optimal linear filter for maximizing S/N ratio against a stochastic background. It provides an unbiased estimate of the source's flux

ML approach

Source detection as binary segmentation. The sources' position is our only concern. We don't have sensitivity to the sources' flux



Where foregrounds dominate, the MF cannot correctly characterize the statistics of the background, but the CNN is still able to provide high levels of completeness even at low fluxes

Open questions for ML

- How to obtain reliable foregrounds simulations
  - Foregrounds are deterministic but there is a lack of knowledge regarding its statistical properties and frequency dependence
  - How do we avoid overfitting?
- How to obtain reliable errors including mismodelling of the foregrounds
- It is posible to use sime kind of unsupervised learning that do not need foreground modelling?

Final remarks

- Microwave observations contain a mixture of CMB, contaminants and noise that need to be separated in order to interpret correctly the data
- A large number of methods have been developed and showed to work well for intensity (mainly focused on the recovery of the CMB)
- However, high-sensitivity future experiments for the detection of the B-mode polarization of CMB require an extremely good characterisation of foregrounds
- Indeed, the component separation problem for the detection of the primordial B mode of polarization is extremely challenging and unavoidable
- Given the weakness of the signal, even small errors in the modellisation of the foregrounds can lead to significant biases on the measurement of the B-mode
- Can machine-learning based methods provide a solution and be competitive?