# Progress in statistical data analysis methods for 21cm surveys

**Phil Bull**

Jodrell Bank Centre for Astrophysics
and
Centre for Radio Cosmology, U. Western Cape

# Overview

Statistical challenges in 21cm analysis

Foreground removal and signal loss

Flagging and ringing

Bayesian anomaly detection

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 948764)

European Research Council
Established by the European Commission

# Statistical challenges in 21cm analysis

- Gigantic dynamic range between foregrounds and signal
  - Need to weight data carefully to improve SNR
  - Need to subtract foregrounds very carefully to avoid destroying signal
- Complexity of spectral/temporal response of instruments
  - Difficult to come up with accurate simulations (c.f. galaxy surveys, who can simulate their covariance matrices!)
  - Unknown systematics, calibration errors

# Statistical challenges in 21cm analysis

- Gigantic dynamic range between foregrounds and signal
  - Need to weight data carefully to improve SNR
  - Need to subtract foregrounds very carefully to avoid destroying signal
- Complexity of spectral/temporal response of instruments
  - Difficult to come up with accurate simulations (c.f. galaxy surveys, who can simulate their covariance matrices!)
  - Unknown systematics, calibration errors

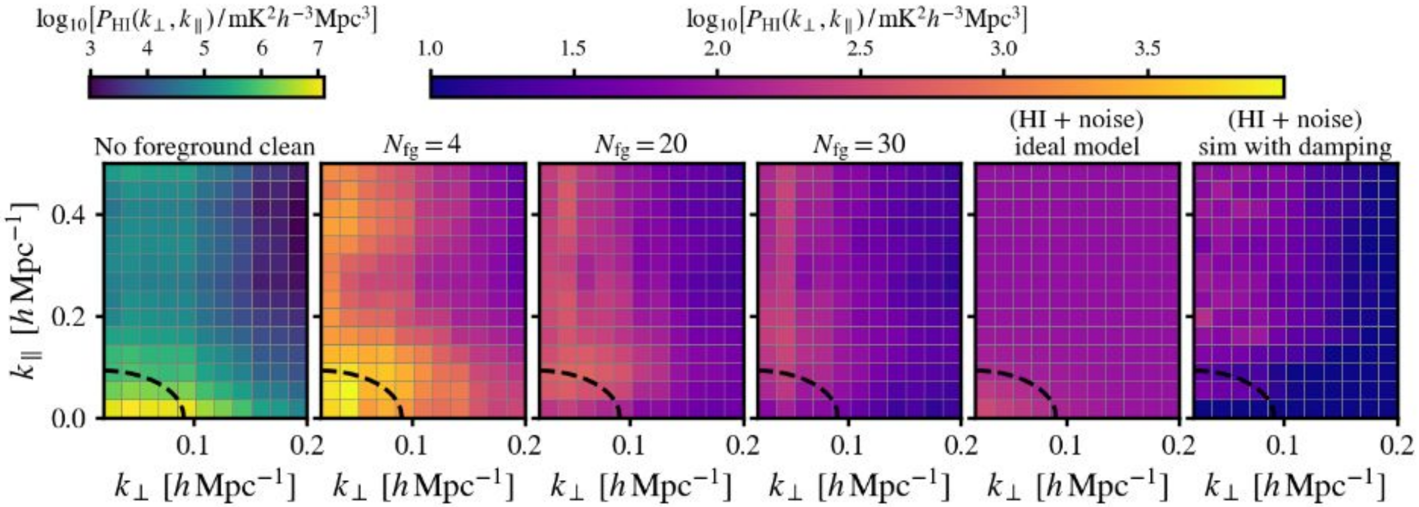
**Challenge:** How do we extract the extremely delicate 21cm signal:

(a) without wrecking everything; or

(b) at least knowing if/when we've wrecked everything!

# Foreground removal and signal loss

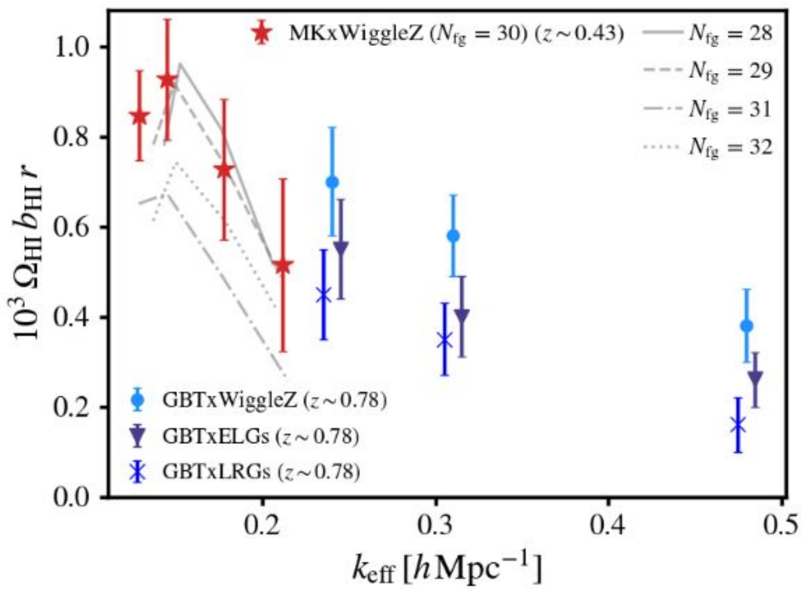
- Foregrounds: much brighter than 21cm signal + corrupted by instrument response
- "Blind" methods construct **filters** from data and subtract limited number of modes
- Modes are not orthogonal to 21cm signal, so some **signal loss** unavoidable




Cunnington, Li et al. [2206.01579]

# Foreground removal and signal loss

- Signal loss is a major problem for science interpretation
- **Do we trust methods that "undo" the loss?** (i.e. transfer function method)
    - *TF method: Inject mock data into real data, apply filter, cross-correlate with unfiltered mock to infer scale-dep.. signal loss transfer function T(k)*




Cunnington, Li et al. [2206.01579]

# Foreground removal and signal loss

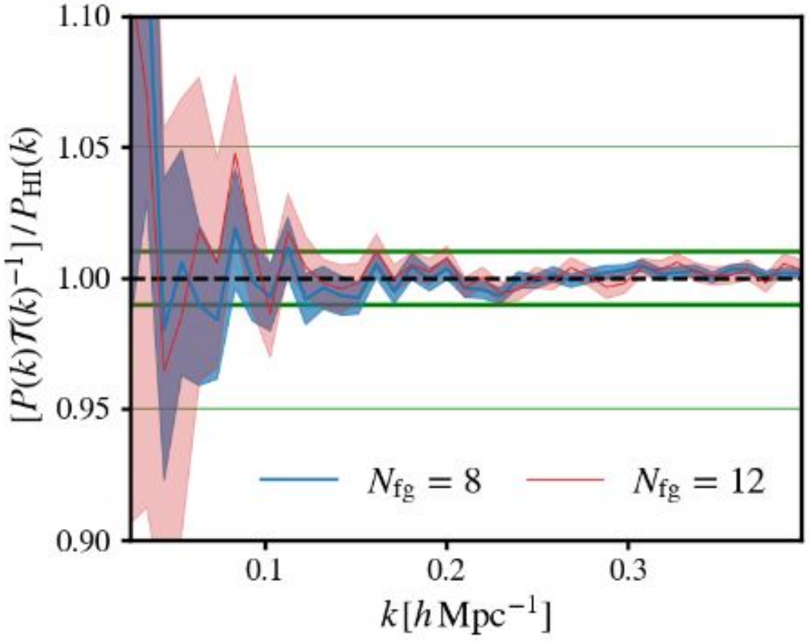
**Cunnington et al. [2302.07034]**

- TF method is robust if implemented properly!
- Need to account for variance of TF
- Beware over-application of TF!

```
P(auto) = P(clean) / T(k)
```

*not*

```
P(auto) = P(clean) / T²(k)
```



# Foreground removal and signal loss

## Cunnington et al. [2302.07034]

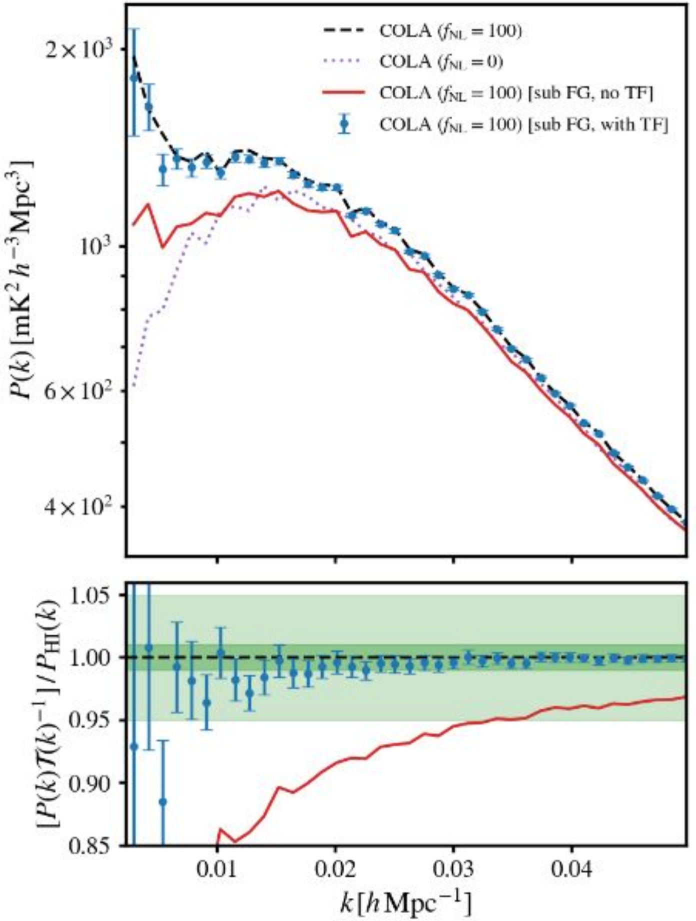
- TF method is robust if implemented properly!
- Need to account for variance of TF
- Beware over-application of TF!

```
P(auto) = P(clean) / T(k)
```

*not*

$$P(auto) = P(clean) / T^2(k)$$

- **Results are robust to simulated model**

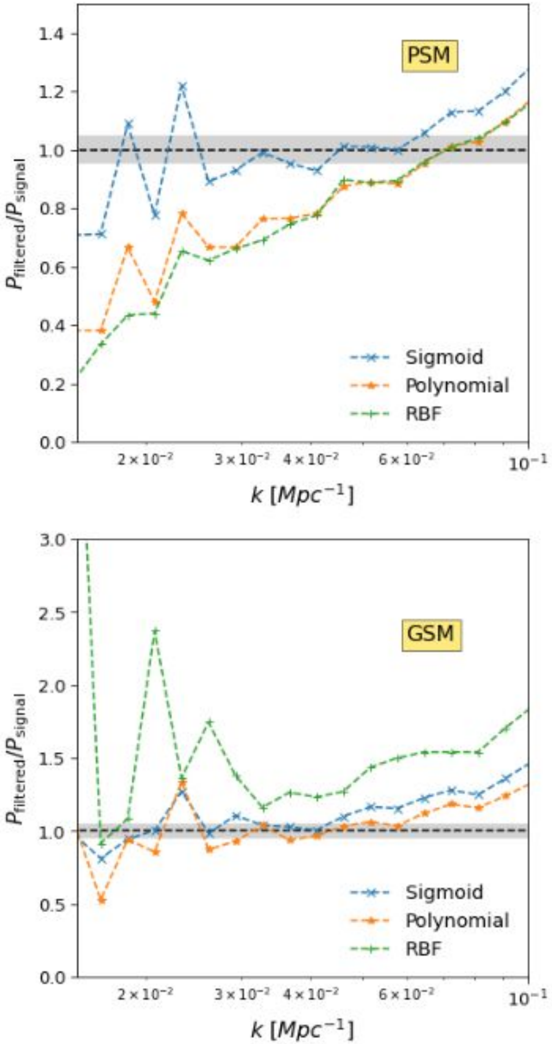


# Kernel PCA

- Many blind filtering methods are related to Principal Component Analysis (PCA)
    - Construct freq.-freq. covariance from observed data
    - Do eigendecomposition
    - Use highest-SNR modes as FG templates
- PCA is lossy, and can quickly eat the 21cm signal

# Kernel PCA

- Many blind filtering methods are related to Principal Component Analysis (PCA)
  - Construct freq.-freq. covariance from observed data
  - Do eigendecomposition
  - Use highest-SNR modes as FG templates
- PCA is lossy, and can quickly eat the 21cm signal
- **Kernel PCA** is a related method that permits non-linear combinations of the data to be used in constructing FG modes
- If tuned carefully, acts like “fractional” PCA




Irfan & Bull [2107.02267]

# Flagging and ringing

- Flagging of RFI-affected channels is unavoidable
- This is a major headache for harmonic analysis (e.g. power spectra)!
  - Missing data causes ringing (very bad for 21cm due to dynamic range)

# Flagging and ringing

- Flagging of RFI-affected channels is unavoidable
- This is a major headache for harmonic analysis (e.g. power spectra)!
    - Missing data causes ringing (very bad for 21cm due to dynamic range)
- What to do?
    - Lomb-Scargle / Least-Squares Spectral Analysis
    - In-painting (fill-in missing data)
    - Deconvolve the mask

All involve implicit models of missing data

# Flagging and ringing

- Flagging of RFI-affected channels is unavoidable
- This is a major headache for harmonic analysis (e.g. power spectra)!
  - Missing data causes ringing (very bad for 21cm due to dynamic range)
- What to do?
  - Lomb-Scargle / Least-Squares Spectral Analysis
  - In-painting (fill-in missing data)
  - Deconvolve the mask
  - **Infer the masked data** → *Gaussian constrained realisations*

All involve implicit models of missing data

# GCR and Gibbs sampling

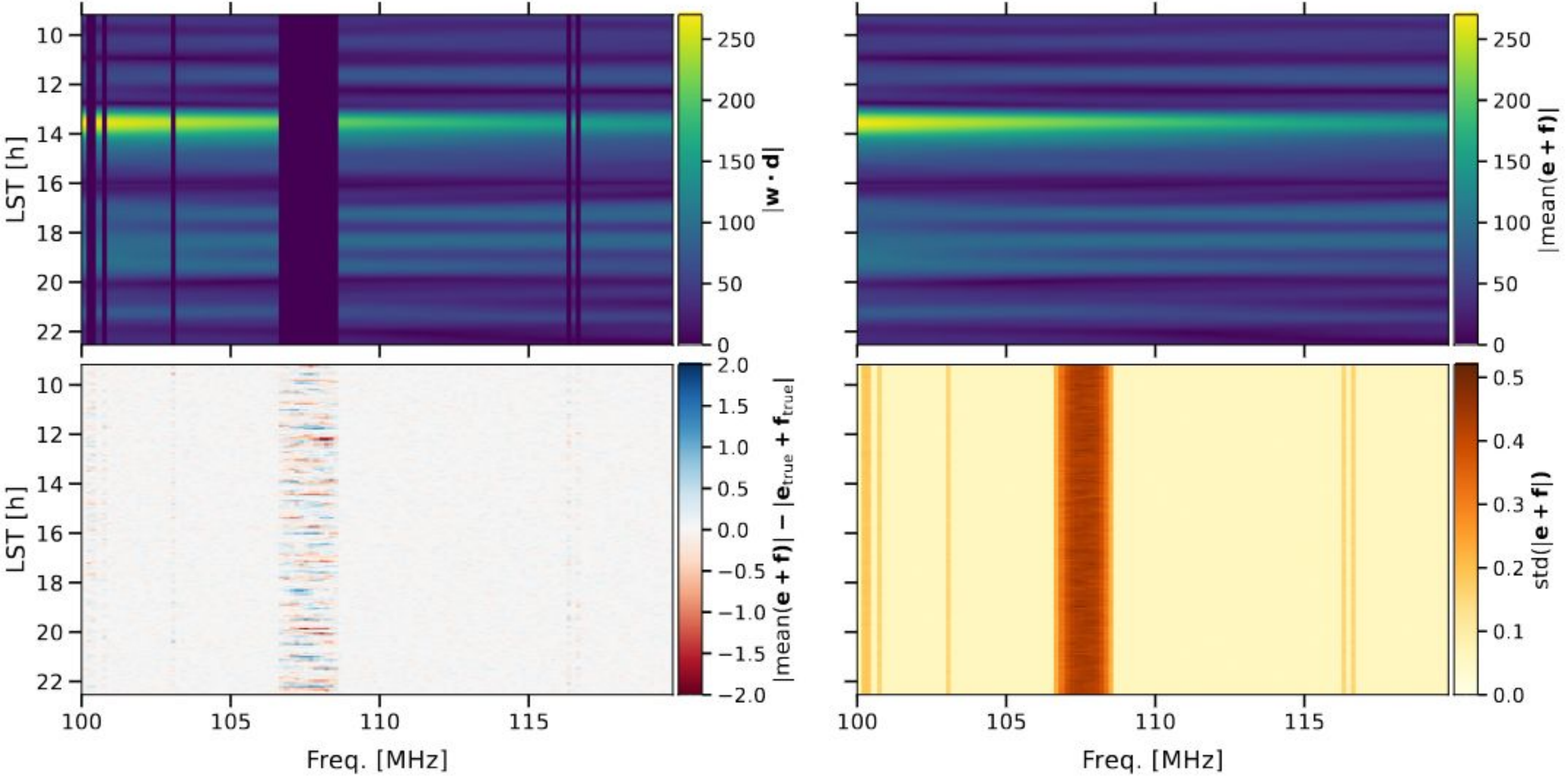
Kennedy et al. [2211.05088]

- **GCR:** Draw samples of the 21cm signal + foregrounds given **observed data**, **foreground basis functions**, **noise** and **21cm signal covariance** estimates

$$p(\mathbf{e}, \mathbf{a}_{\rm fg} | \mathbf{E}, \mathbf{g}_j, \mathbf{N}, \mathbf{d}) \propto p(\mathbf{d} | \mathbf{e}, \mathbf{a}_{\rm fg}, \mathbf{g}_j, \mathbf{N}) p(\mathbf{e} | \mathbf{E})$$

- Each sample has **no gaps**, so Fourier analysis can be applied exactly (no ringing). Repeat many times to build up statistical distribution.
- What if the 21cm signal covariance is poorly known? → **Gibbs sampling method**
  - Iteratively sample 21cm signal (+ foregrounds), then 21cm covariance

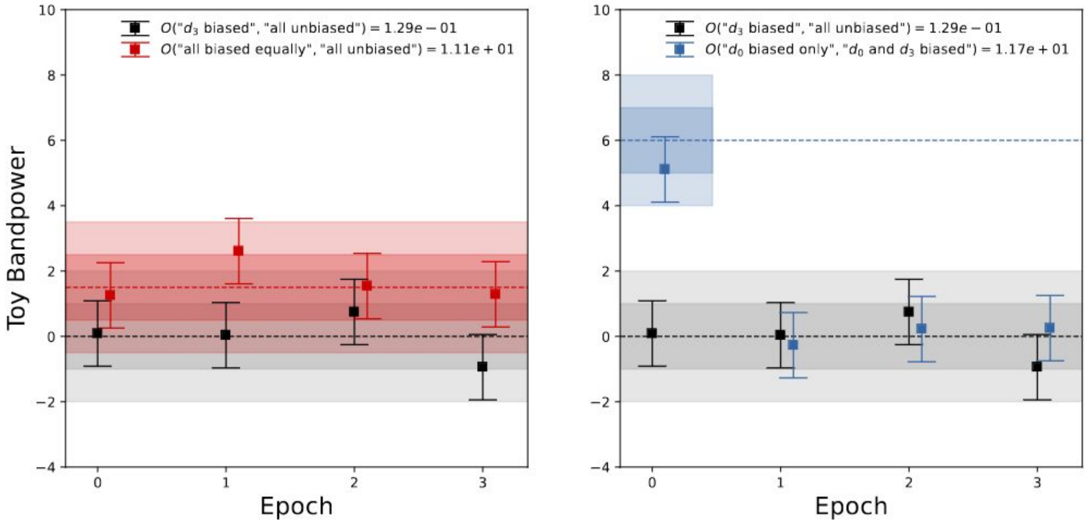
$$\mathbf{s}_{i+1} \leftarrow p(\mathbf{s}_i | \mathbf{S}_i, \mathbf{N}, \mathbf{d})$$
$$\mathbf{S}_{i+1} \leftarrow p(\mathbf{S}_i | \mathbf{s}_{i+1}).$$




Kennedy et al. [2211.05088]

# Bayesian anomaly detection

- So many ways of splitting the HERA data, not enough humans to inspect it all
- **Chiborg:** an automated, statistically-principled way of doing **null/jackknife** tests
  - Can handle a few subsets with different weights (i.e. not just equal halves)
  - Big hierarchy of hypotheses + simple parametrisation for “biased” data




Wilensky+ [2210.17351]

GitHub: [mwilensky768/chiborg](https://github.com/mwilensky768/chiborg)

# Bayesian anomaly detection

- Basic idea: enumerate every possible combination of systematic-affected vs not-affected subsets of the data, then calculate odds ratios
- Systematic level does not have to be known/assumed (i.e. can be drawn from a distribution, and hyperparameters of distribution can be marginalised)

$$P(\mathbf{d}|\mathbf{C}_0, \mathcal{H}_0) = \int_{\mathbb{R}} d\mu_0 P(\mu_0) P(\mathbf{d}|\mathbf{C}_0, \mu_0, \mathcal{H}_0)$$

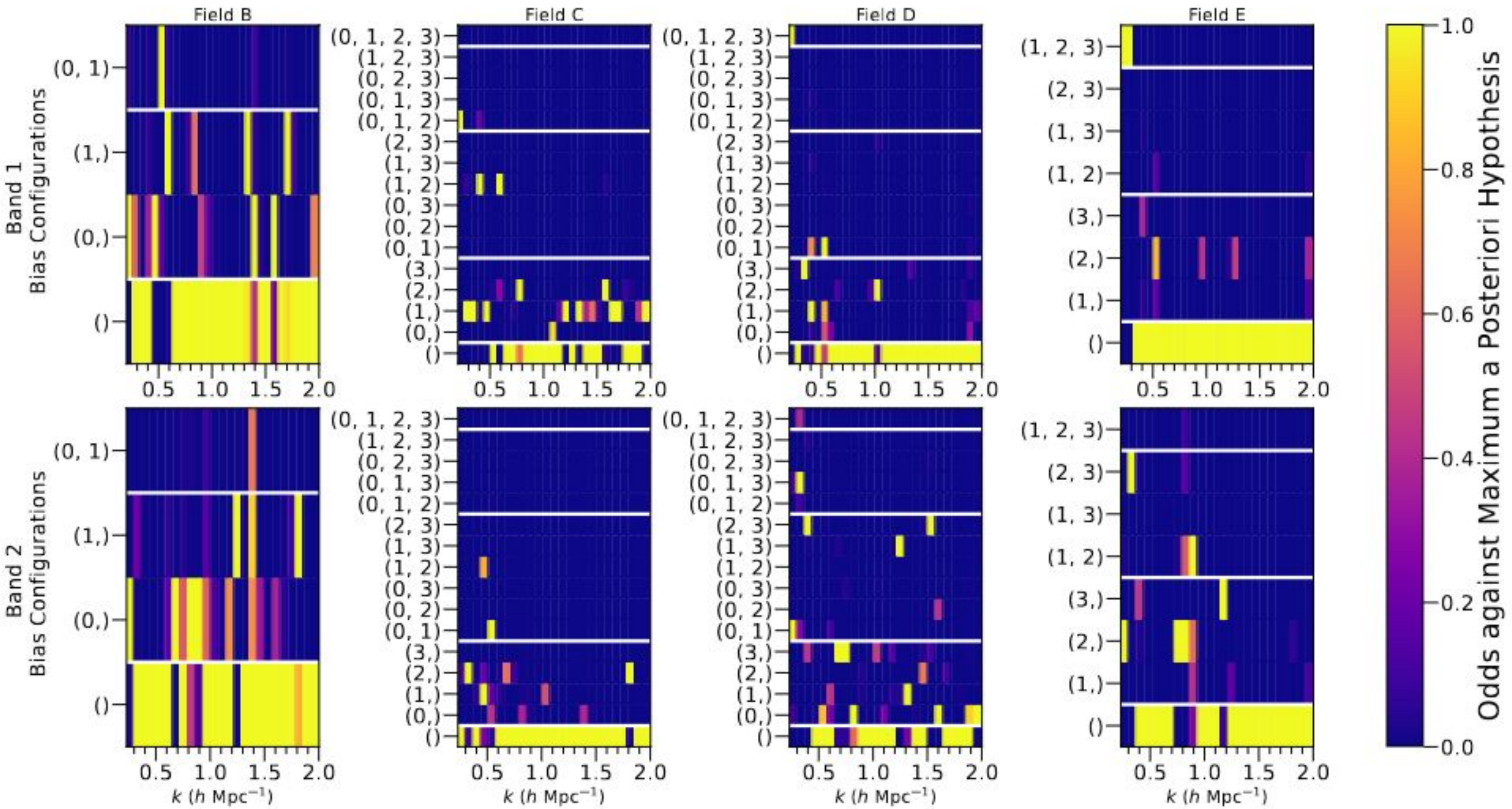
Null hypothesis
(no systematic)

$$P(\mathbf{d}|\mathbf{C}_0, \mathcal{H}_i) = \int_{\mathbb{R}} d\mu_0 P(\mu_0) \int_{\mathbb{R}^N} d\boldsymbol{\varepsilon} P(\boldsymbol{\varepsilon}|\mathcal{H}_i) P(\mathbf{d}|\mathbf{C}_0, \mu_0, \boldsymbol{\varepsilon}, \mathcal{H}_i)$$

Systematic-affected
hypothesis

Wilensky et al. [2210.17351]

# Which epochs of the HERA data have anomalous power? *(by field/band, vs scale cut)*




Wilensky et al. [2210.17351]

# Summary

- Building a statistical model of the data allows us to treat sensitive filtering and power spectrum estimation steps in a principled manner
- Principled approach improves robustness!
- Also allows sense-checking of data in automated fashion (null tests etc.)
- **Please use our software!**

**To get in touch: phil.bull@manchester.ac.uk**

# EXTRA SLIDES