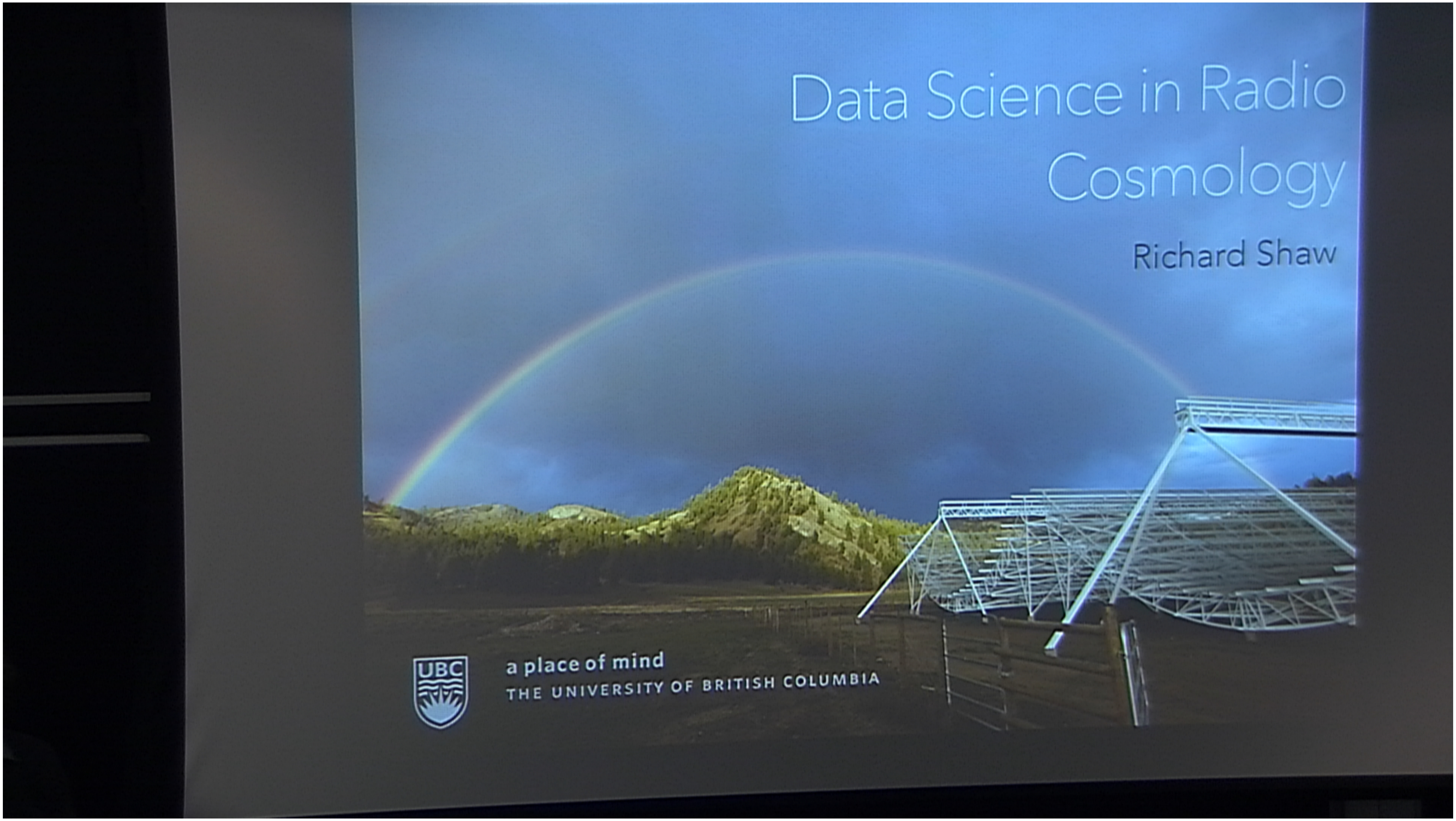
Title: Data Science in Radio Cosmology

Date: Mar 24, 2016 11:00 AM

URL: http://pirsa.org/16030130

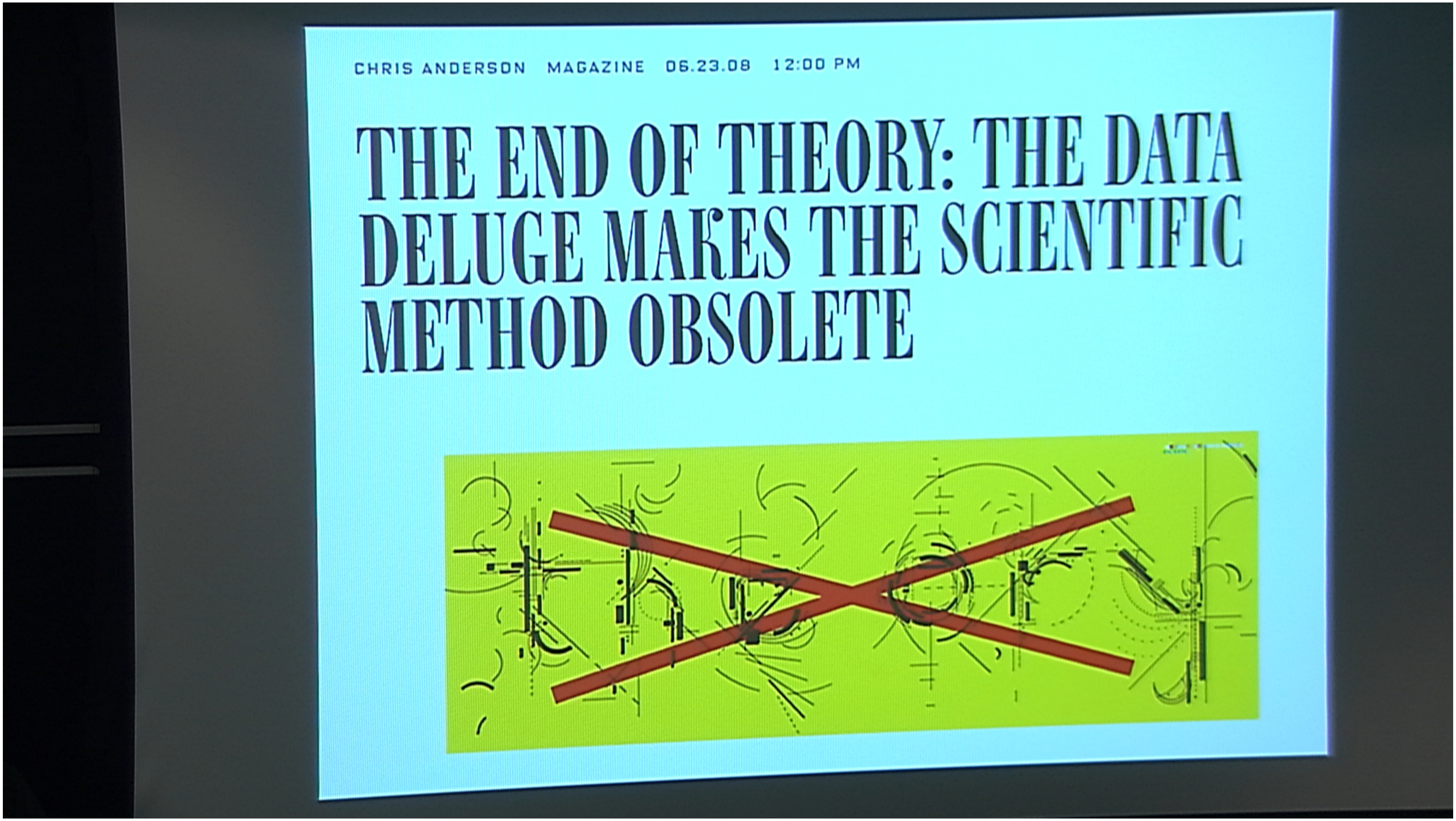
Abstract: <p>In recent decades probing for the subtle indications of new physics in
experimental data has become increasingly difficult. The datasets have gotten
much bigger, the experiments more complex, and the signals ever smaller. Success
stories, like LIGO and Kepler, require a sophisticated combination of statistics
and computation, coupled with an appreciation of both the experimental realities
and the theoretical framework governing the data.

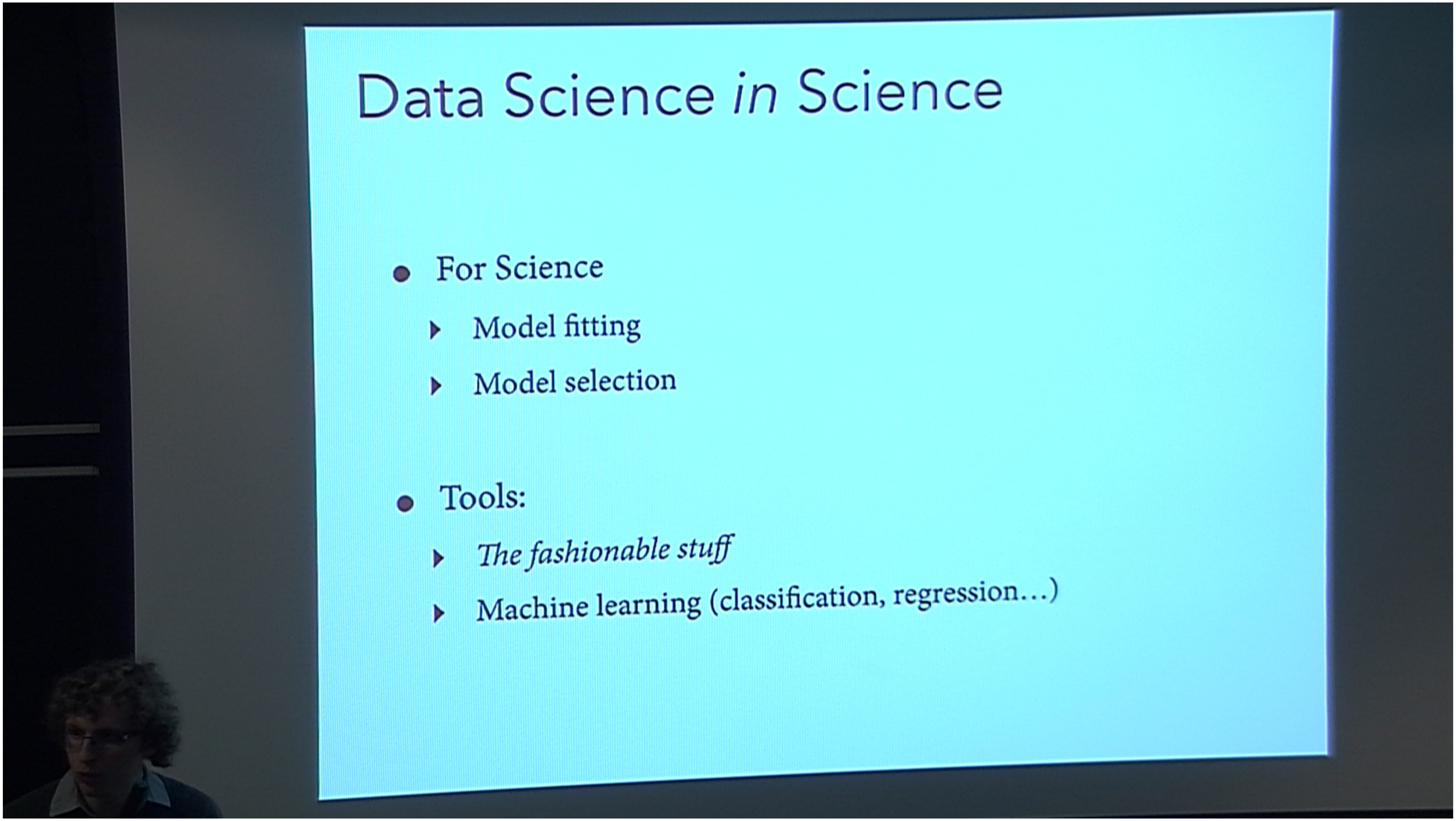
In this talk I will look broadly at data science in physics, and how and why it
has taken an increasingly central role. I'll highlight specifically my current
area of research, radio cosmology: discussing why it is one of the most
challenging areas for data science, and describing my work developing optimal
and efficient statistical methods for turning terabytes of timestreams into
cosmology.</p>

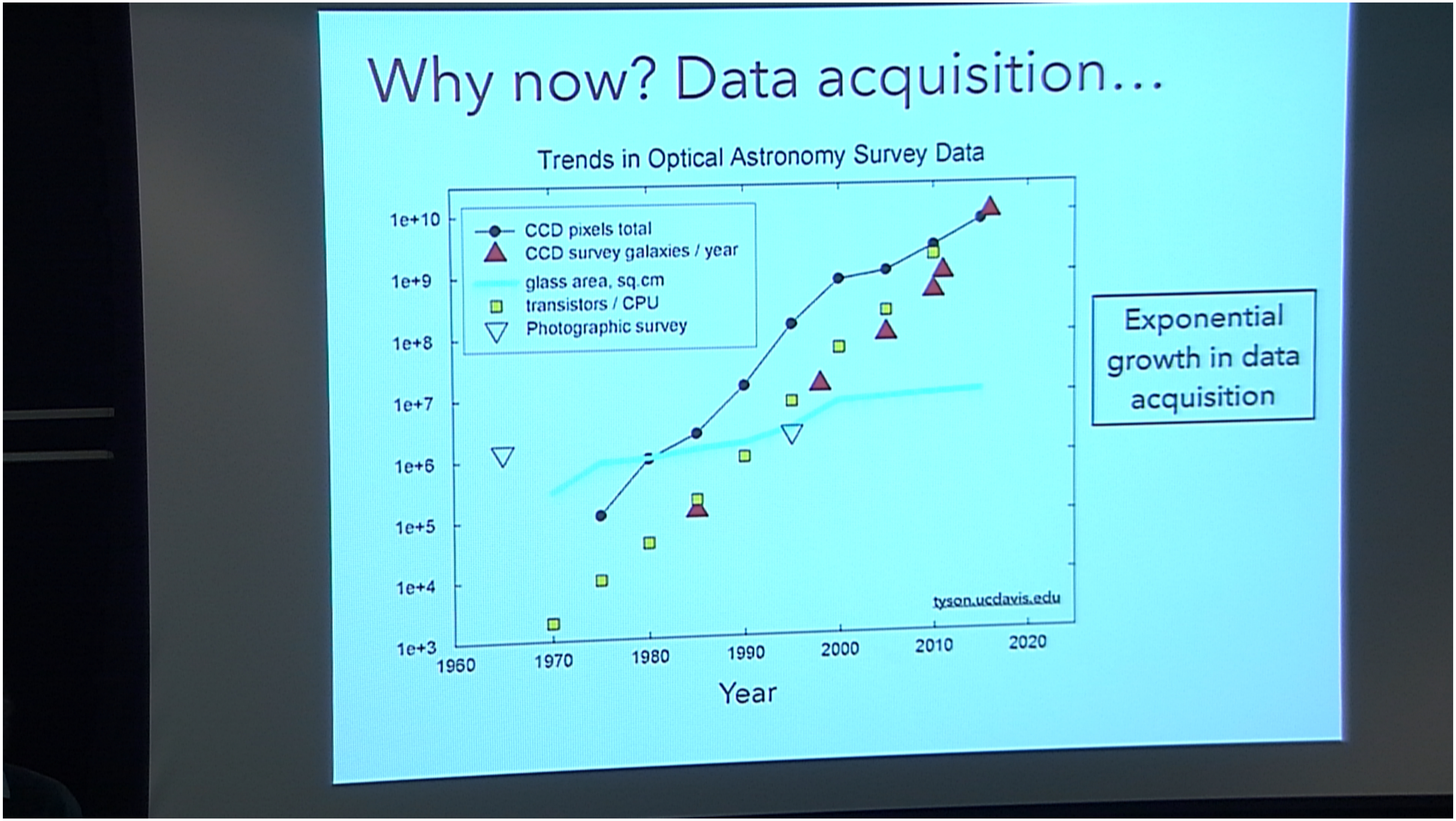

Data Science in Radio Cosmology
Richard Shaw
UBC
a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

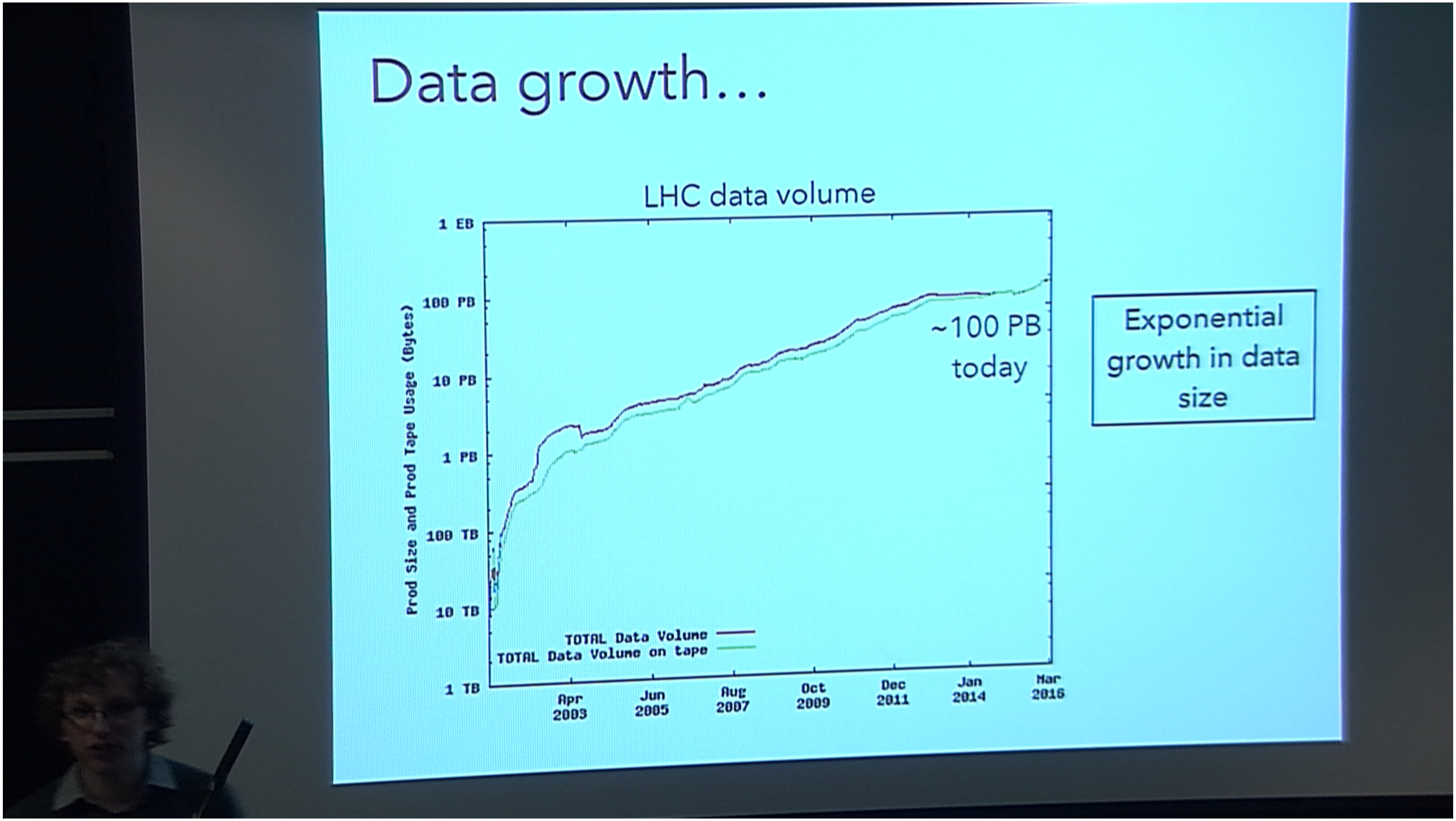
Data Science

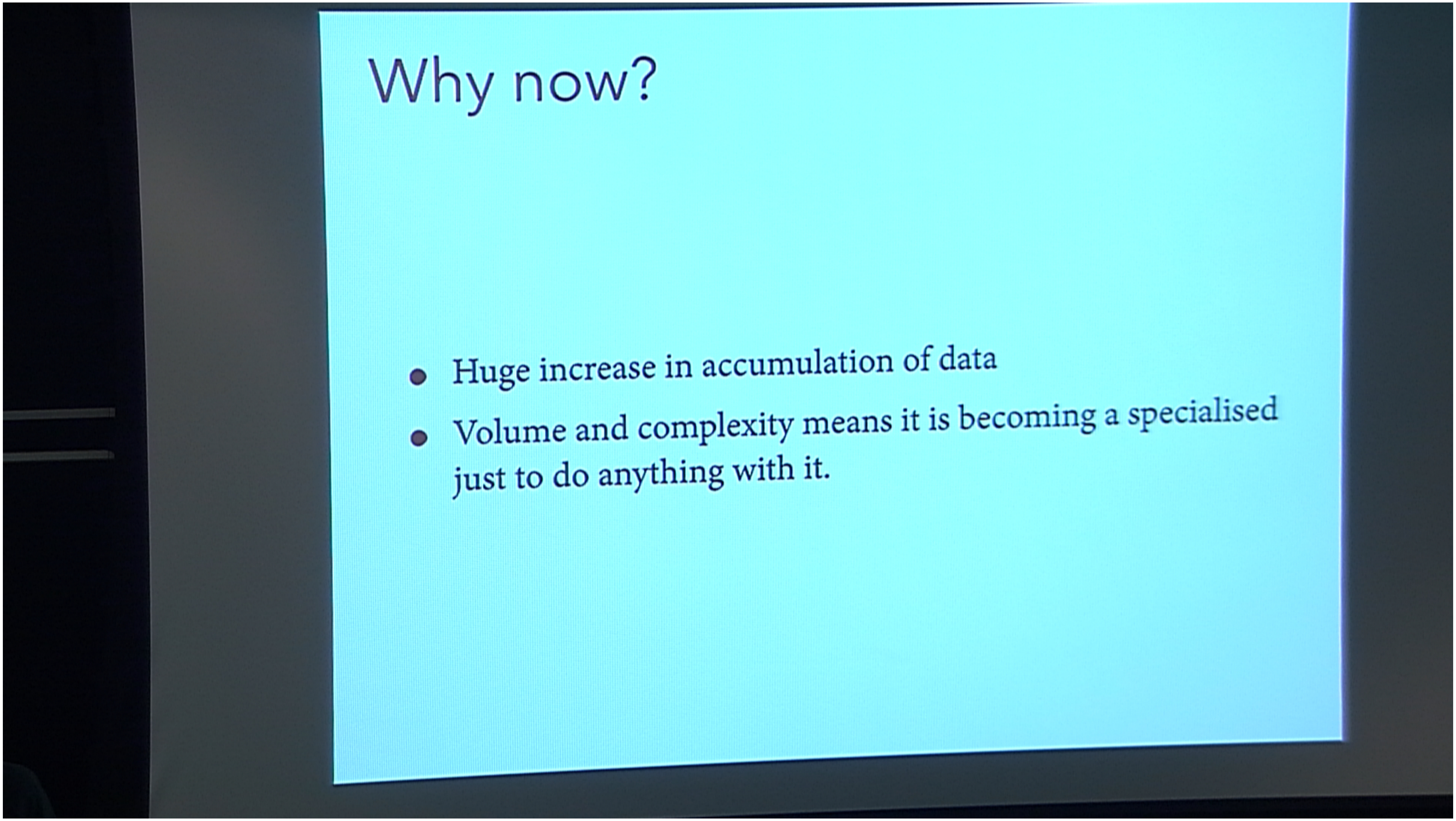
- Combination of:
  - Statistics
  - Signal processing
  - Machine learning
  - High performance computing
  - **Physics** (theory and experiment)


CHRIS ANDERSON MAGAZINE 06.23.08 12:00 PM
THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE


Data Science *in* Science
For Science
Model fitting
Model selection
Tools:
The fashionable stuff
Machine learning (classification, regression…)


Why now? Data acquisition…
Trends in Optical Astronomy Survey Data
CCD pixels total
CCD survey galaxies / year
glass area, sq.cm
transistors / CPU
Photographic survey
1e+10
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1960
1970
1980
1990
2000
2010
2020
Year
tyson.ucdavis.edu
Exponential growth in data acquisition


Data growth...
LHC data volume
1 EB
100 PB
10 PB
1 PB
100 TB
10 TB
1 TB
Prod Size and Prod Tape Usage (Bytes)
~100 PB today
Exponential growth in data size
TOTAL Data Volume
TOTAL Data Volume on tape
Apr 2003
Jun 2005
Aug 2007
Oct 2009
Dec 2011
Jan 2014
Mar 2016


Why now?
Huge increase in accumulation of data
Volume and complexity means it is becoming a specialised just to do anything with it.

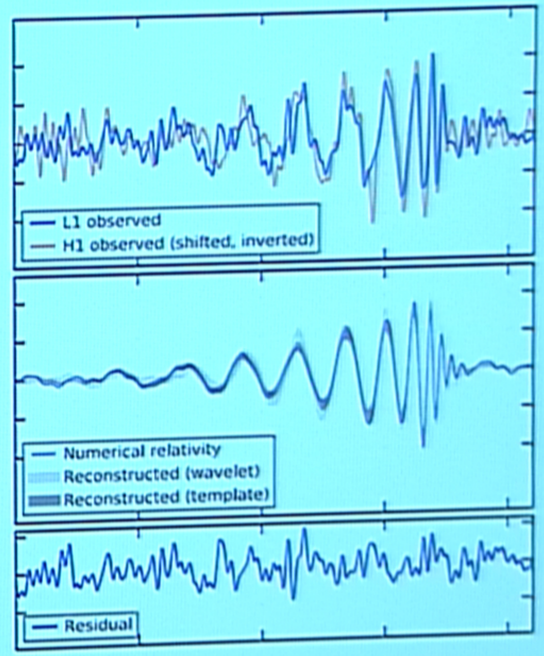
LHC

- Need to identify which collisions are interesting
- Each collision is described by many tens of parameters
- Machine classification problem (Boosted Decision Trees)
  - Trained on simulated data
  - Selects between events of interest and background events
- Used in real-time in software triggers (LHCb)
- Used for event selection for Higgs detection (CMS)

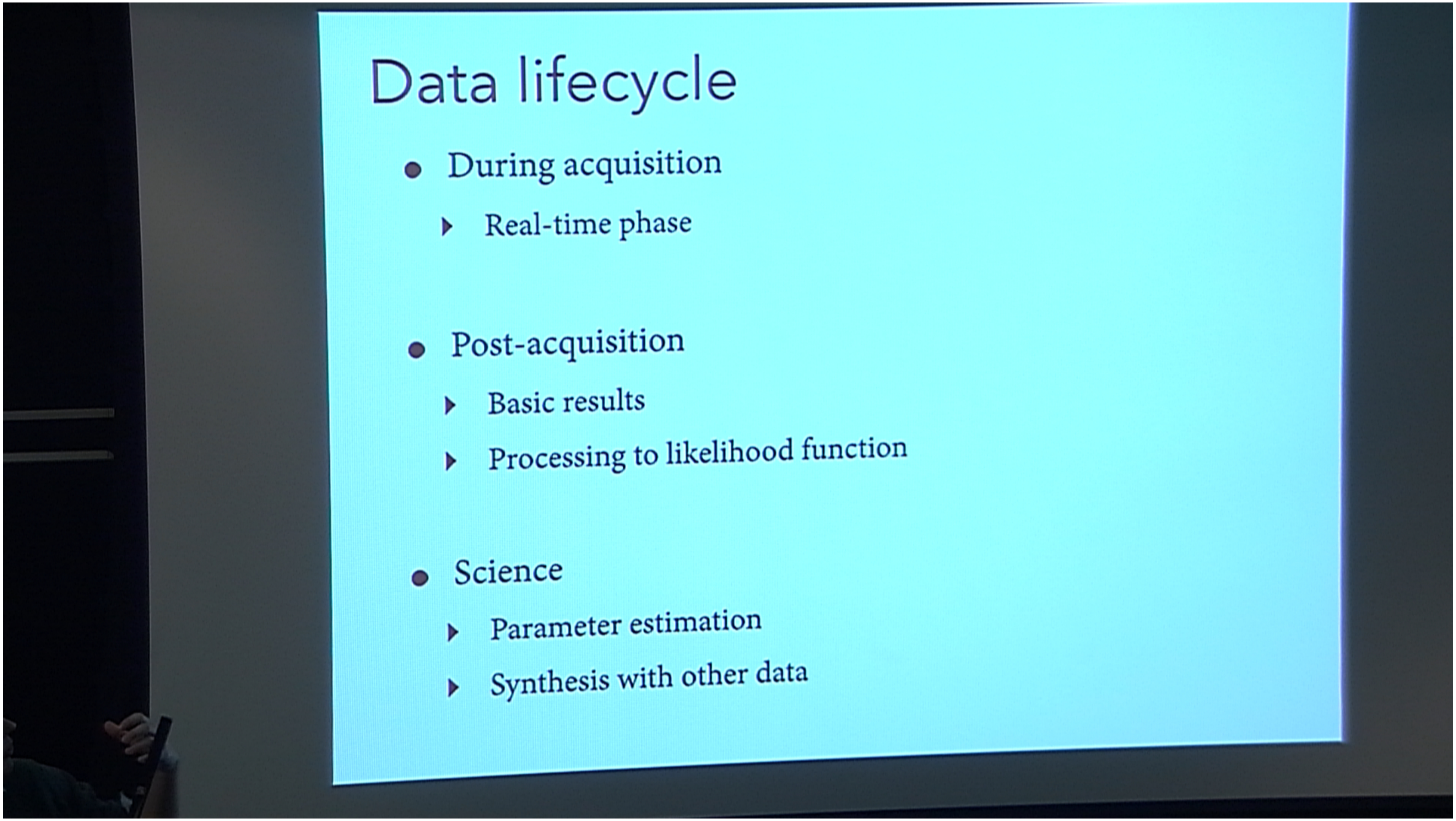
Gligorov 2014, CMS Collaboration 2012

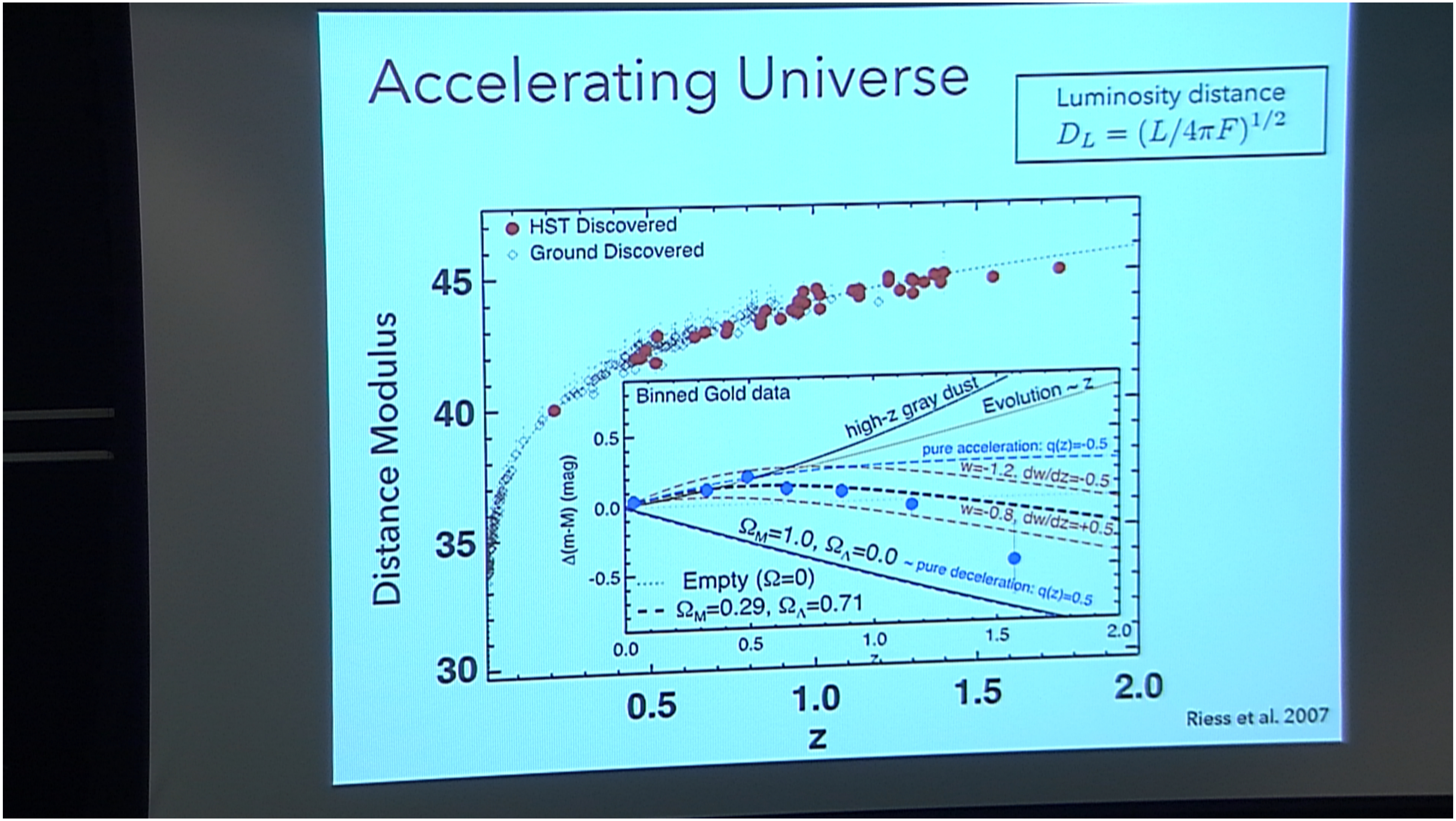
LIGO event detection

- Real-time matched filter
  - Effectively search against 250k template waveforms
  - Look for peaks in likelihood ratio, keep those that exceed some false positive rate
  - Keep only events coincident within 15ms in both detectors (~light travel time)



Abbott et al. 2016


Data lifecycle
During acquisition
Real-time phase
Post-acquisition
Basic results
Processing to likelihood function
Science
Parameter estimation
Synthesis with other data


Accelerating Universe
Luminosity distance
$D_L = (L/4\pi F)^{1/2}$
HST Discovered
Ground Discovered
Distance Modulus
45
40
35
30
0.5
1.0
1.5
2.0
z
Binned Gold data
high-z gray dust
Evolution ~ z
pure acceleration: q(z)=-0.5
w=-1.2, dw/dz=-0.5
w=-0.8, dw/dz=+0.5
Ω_M=1.0, Ω_Λ=0.0 ~ pure deceleration: q(z)=0.5
Empty (Ω=0)
Ω_M=0.29, Ω_Λ=0.71
Δ(m-M) (mag)
Riess et al. 2007

Probing Dark Energy

- Acceleration 'explained' by dark energy (cosmological constant, quintessence ...)

- Expansion is governed by Friedmann equation

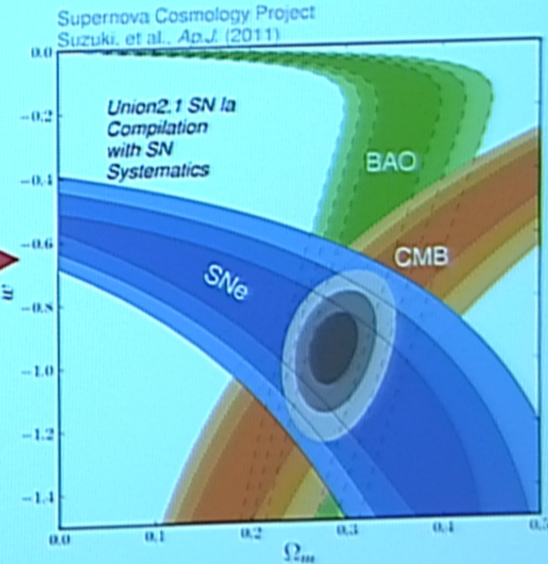
$$H(z)^2 \approx \Omega_m(1+z)^3 + \Omega_{DE} \exp\left[\int_0^z (1+w(z))\frac{dz}{1+z}\right]$$

- Fundamental physics is contained within the equation of state $w(\rho) = p/\rho < -1/3$

Probing Dark Energy

- Construct likelihood function for dataset i.e. Pr(supernovae|expansion)
- Use Markov-Chain Monte-Carlo to infer distribution of relevant parameters

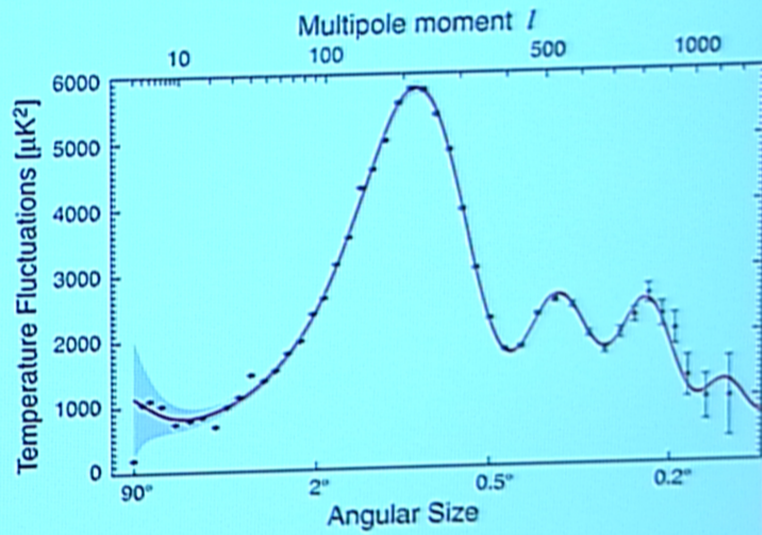
> - Propose parameters (w_0, w_a, Ω_m, …)
> - Generate expansion history
> - Compare to data (likelihood), accept parameters if likely
> - Repeat until enough samples



Baryon Acoustic Oscillations

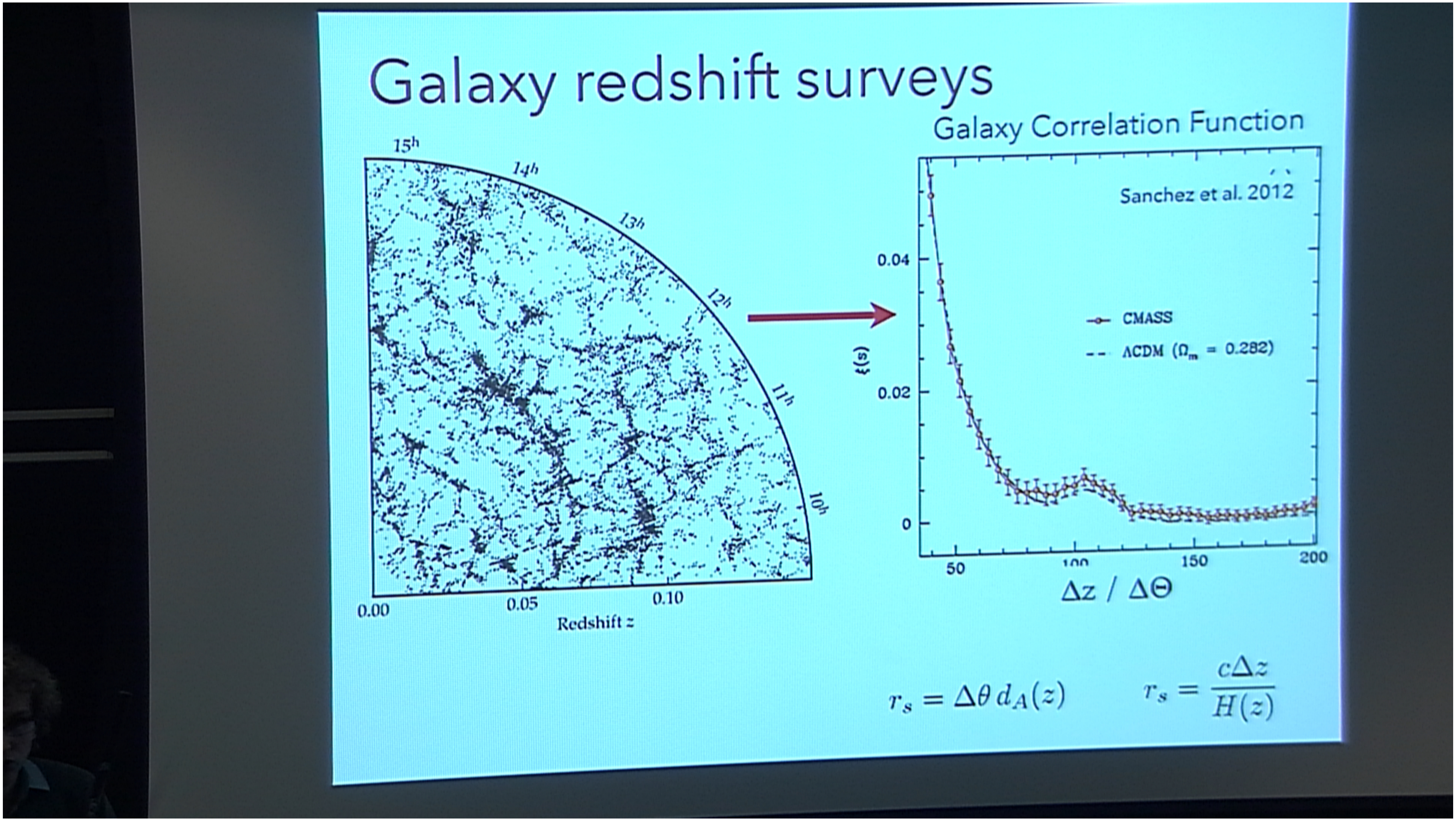
- Sounds waves propagating in the early Universe. Leave acoustic peaks in the CMB
- Weaker imprint left in the matter distribution
- Gives a standard (statistical) ruler

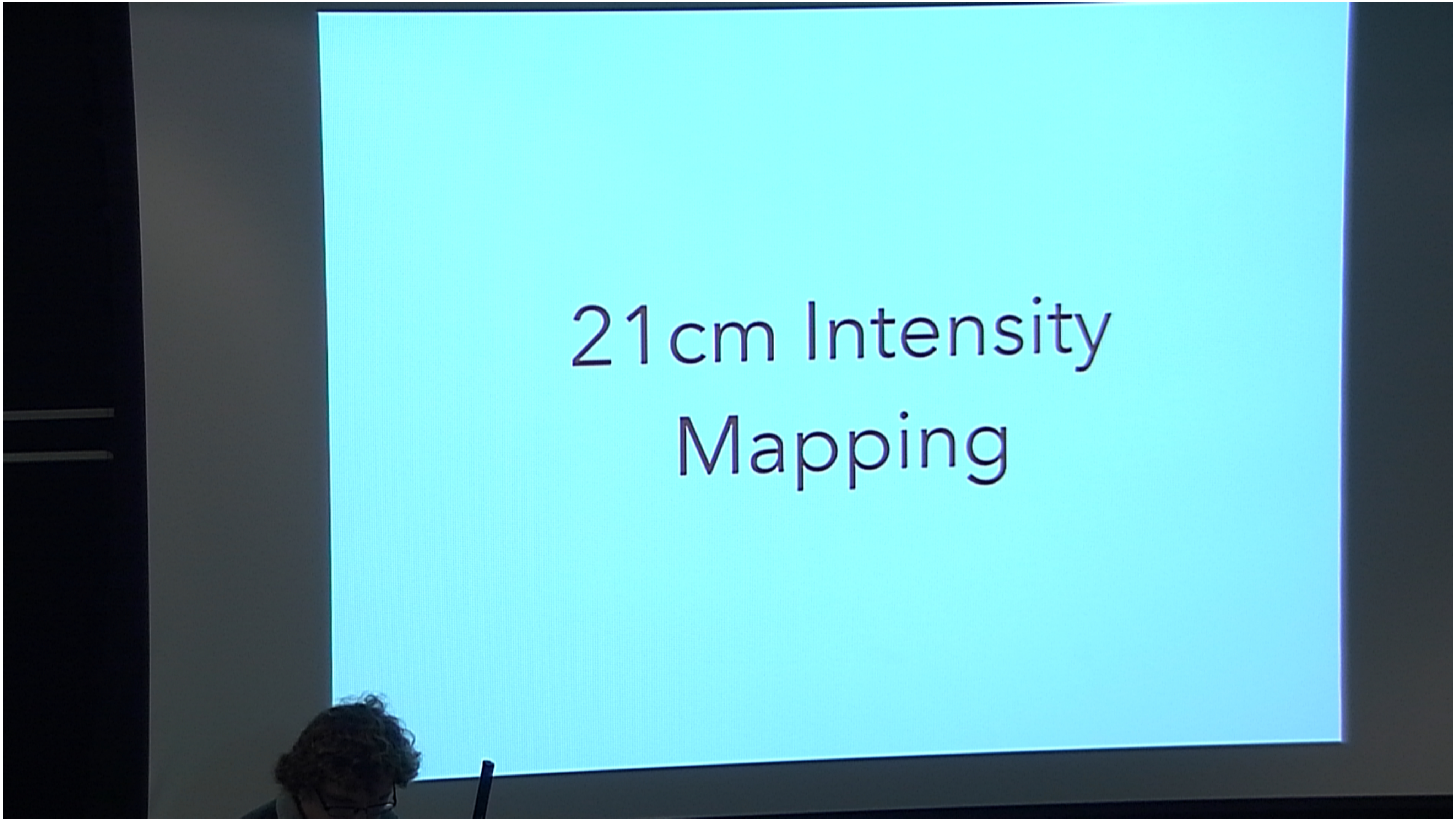
CMB angular power spectrum

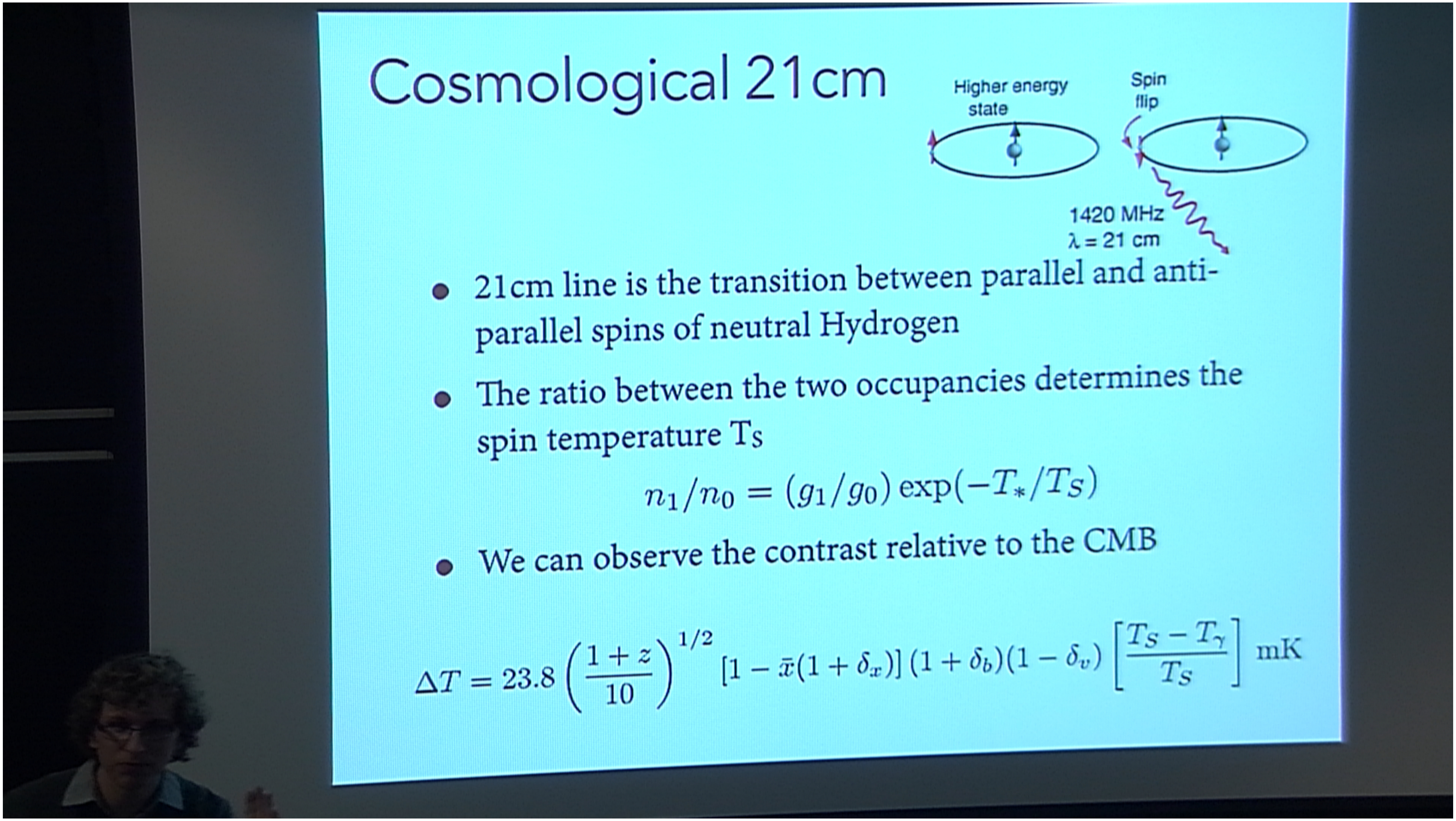


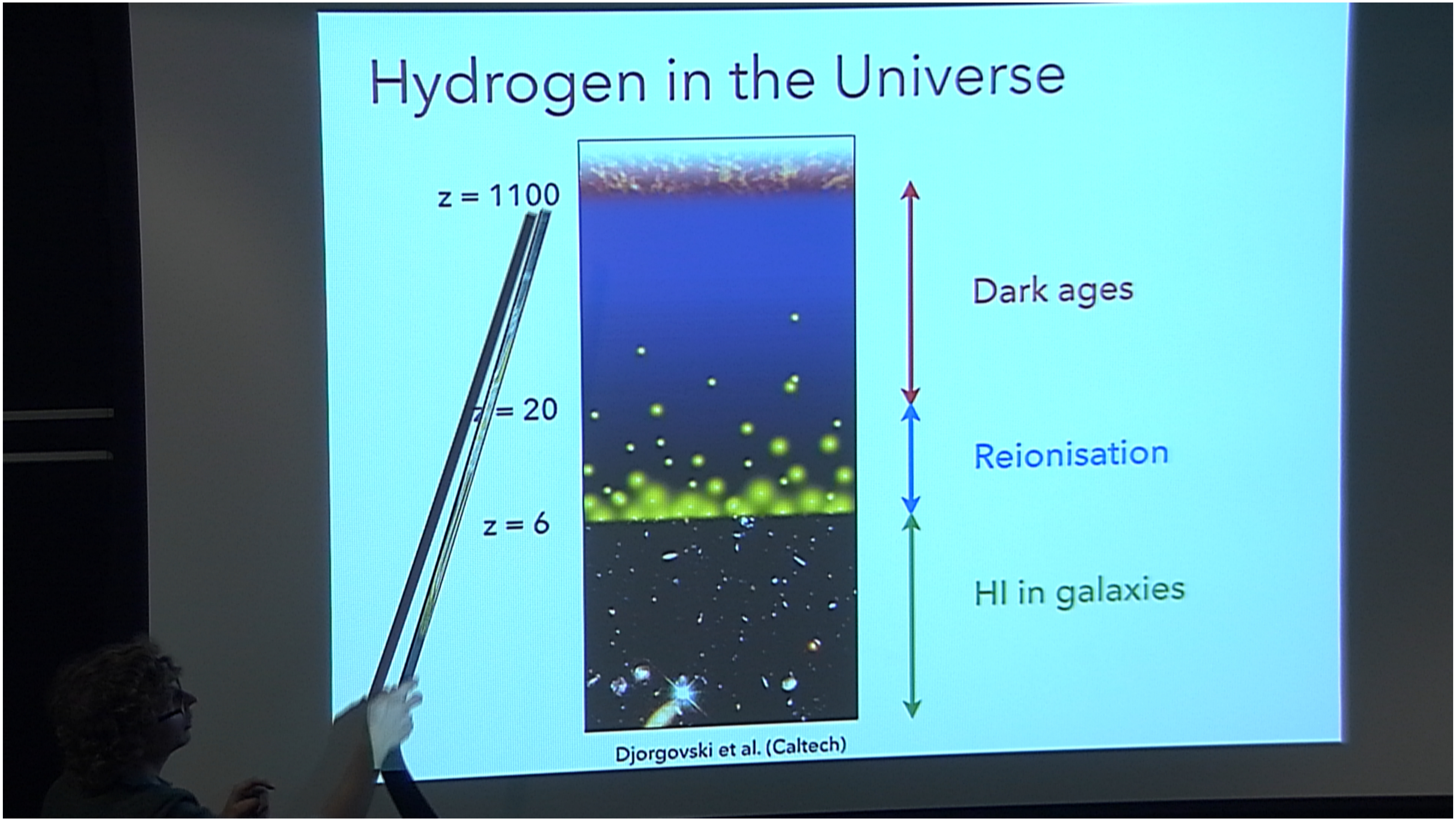
$$r_s = \int_0^{\tau_*} c_s d\tau \sim 100\, h^{-1}\,\mathrm{Mpc}$$

Known from CMB


Galaxy redshift surveys
Galaxy Correlation Function
15h
14h
13h
12h
11h
10h
0.00
0.05
0.10
Redshift z
Sanchez et al. 2012
CMASS
ΛCDM (Ω_m = 0.282)
ξ(s)
0.04
0.02
0
50
150
200
Δz / ΔΘ
$r_s = \Delta\theta\, d_A(z)$
$r_s = \frac{c\Delta z}{H(z)}$


21cm Intensity
Mapping

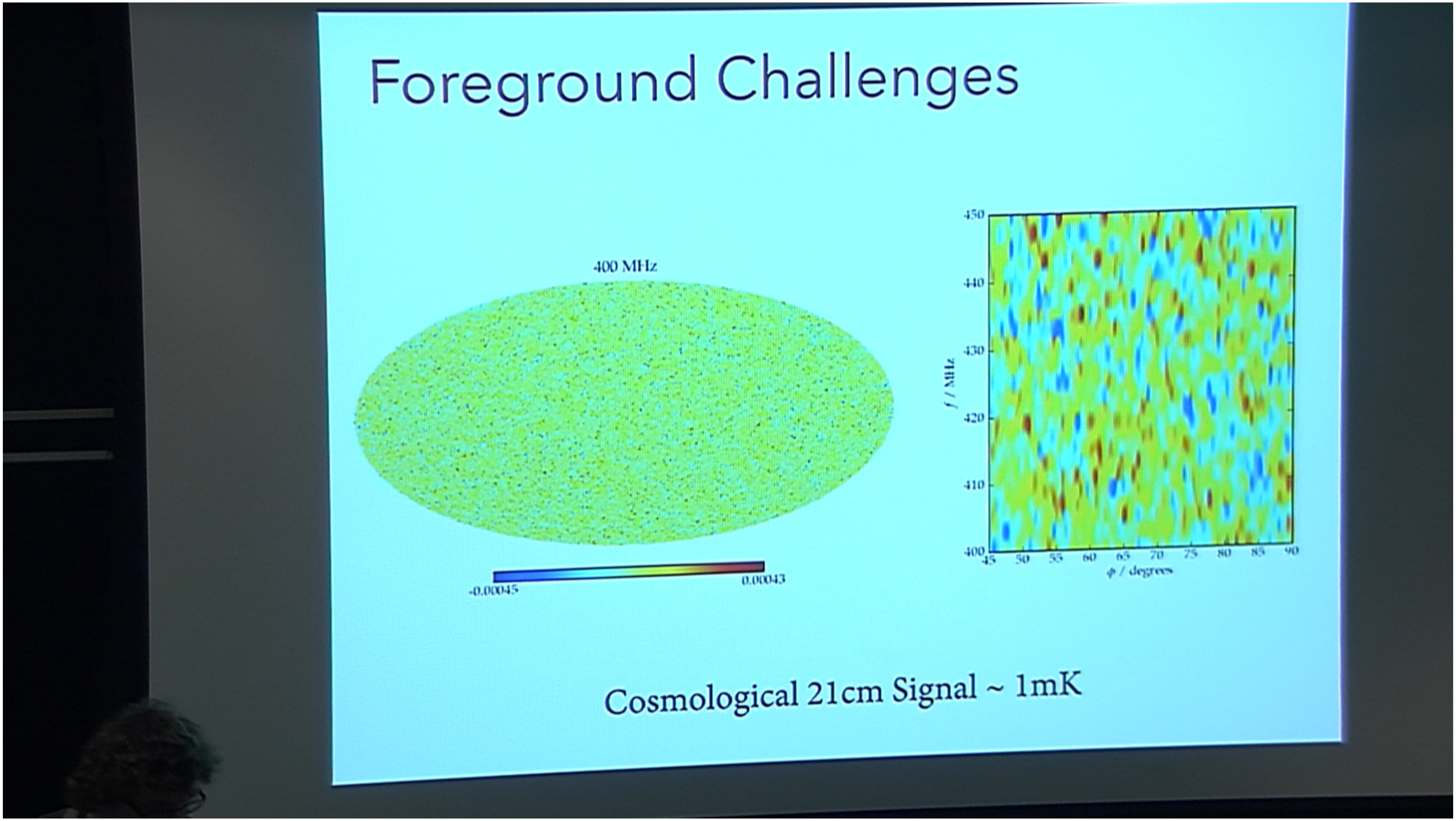

Cosmological 21cm
Higher energy state
Spin flip
1420 MHz
λ = 21 cm
21cm line is the transition between parallel and anti-parallel spins of neutral Hydrogen
The ratio between the two occupancies determines the spin temperature T_S
$n_1/n_0 = (g_1/g_0)\exp(-T_*/T_S)$
We can observe the contrast relative to the CMB
$\Delta T = 23.8\left(\frac{1+z}{10}\right)^{1/2}[1-\bar{x}(1+\delta_x)](1+\delta_b)(1-\delta_v)\left[\frac{T_S-T_\gamma}{T_S}\right]\,\mathrm{mK}$

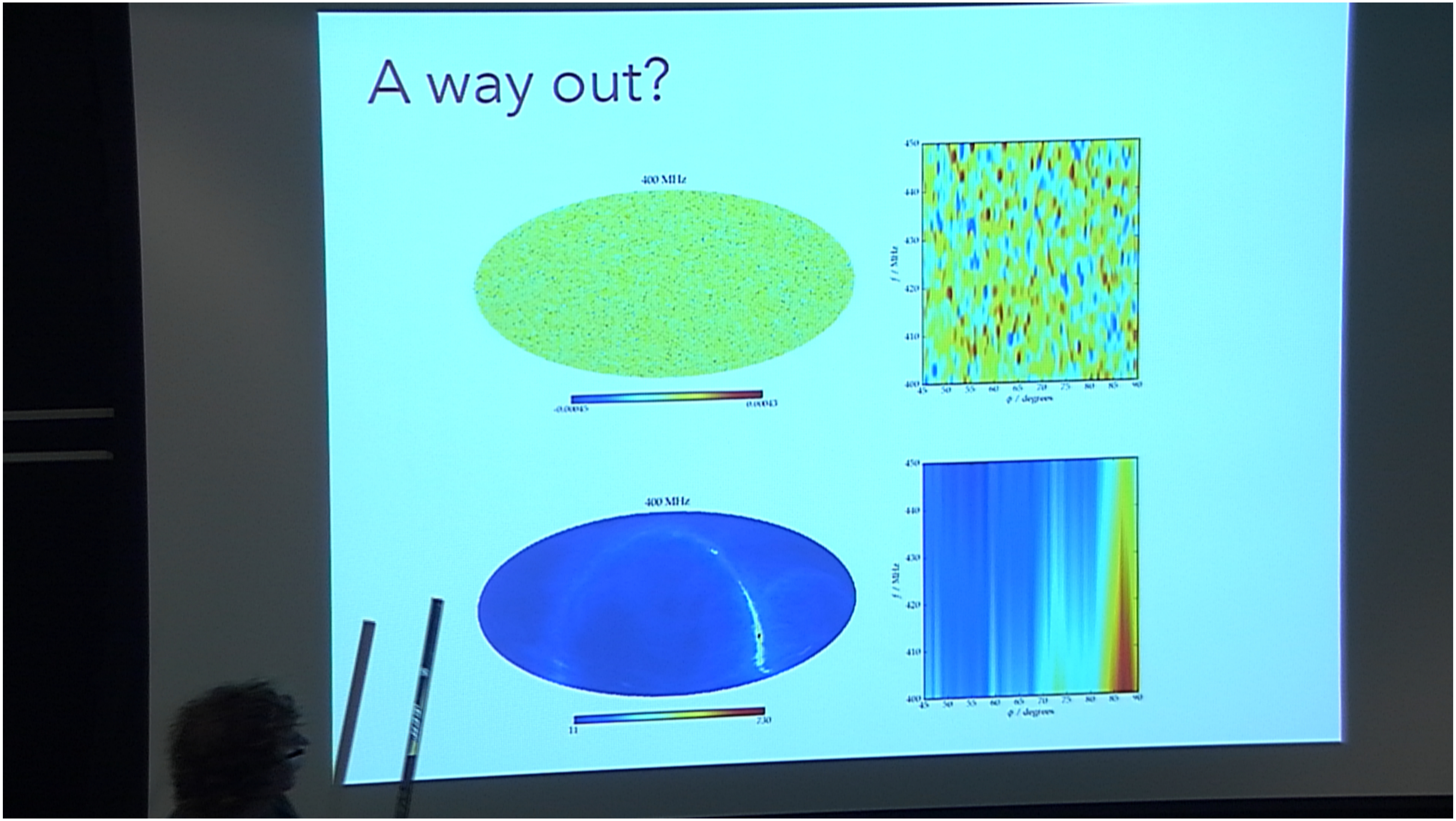

Hydrogen in the Universe
z = 1100
z = 20
z = 6
Dark ages
Reionisation
HI in galaxies
Djorgovski et al. (Caltech)

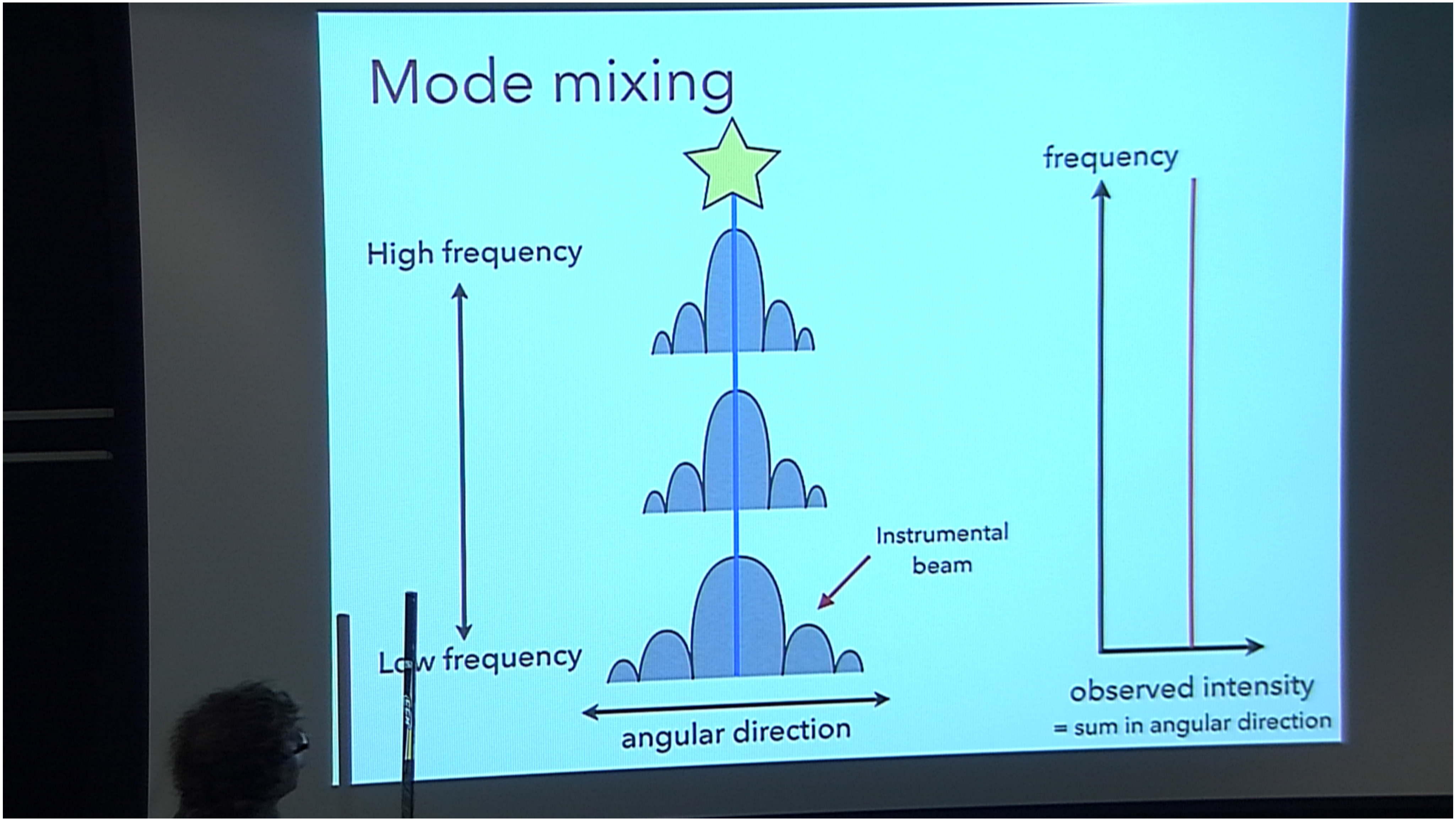
21cm Intensity Mapping

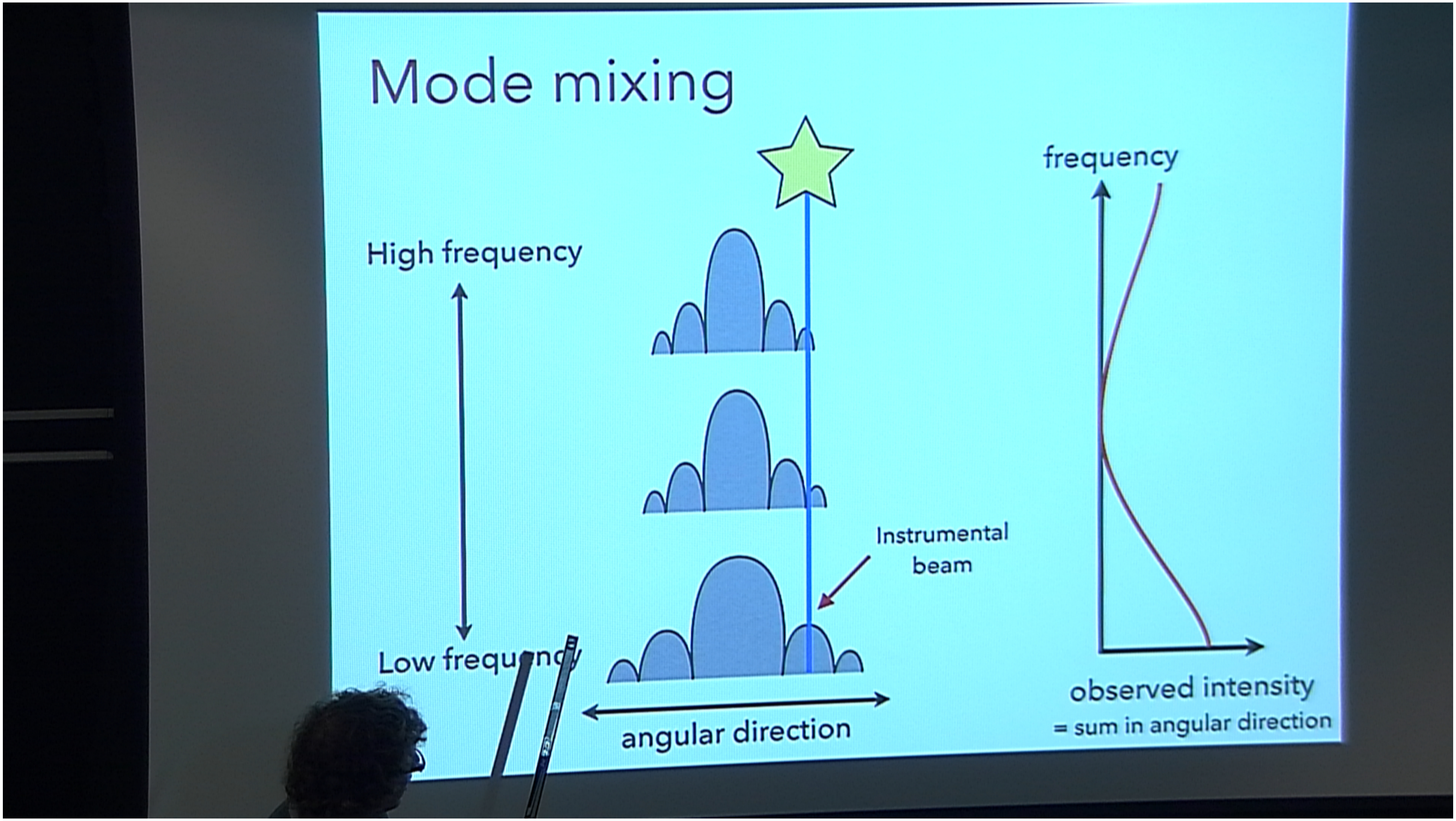
- In 21cm the frequency gives the redshift.
- Observe the diffuse emission from many unresolved galaxies
- Changes the game in telescope design:
  - Previously: large field of view, large collecting area, large angular resolution (SKA?)
  - Now: large field of view, large collecting area, modest angular resolution (compact arrays, single dishes).

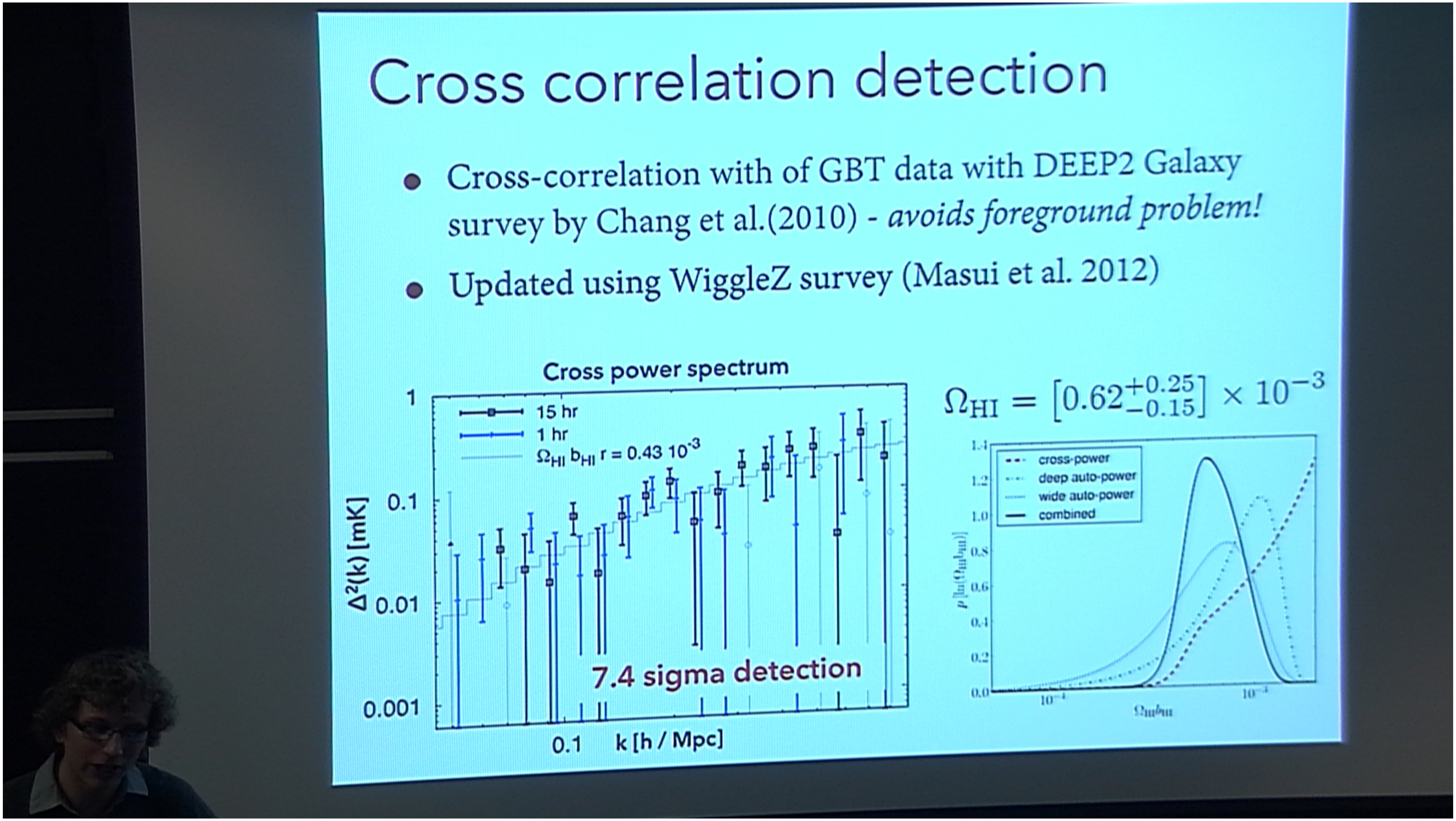
Chang, Pen, Peterson and McDonald , 2008, http://arxiv.org/pdf/0709.3672

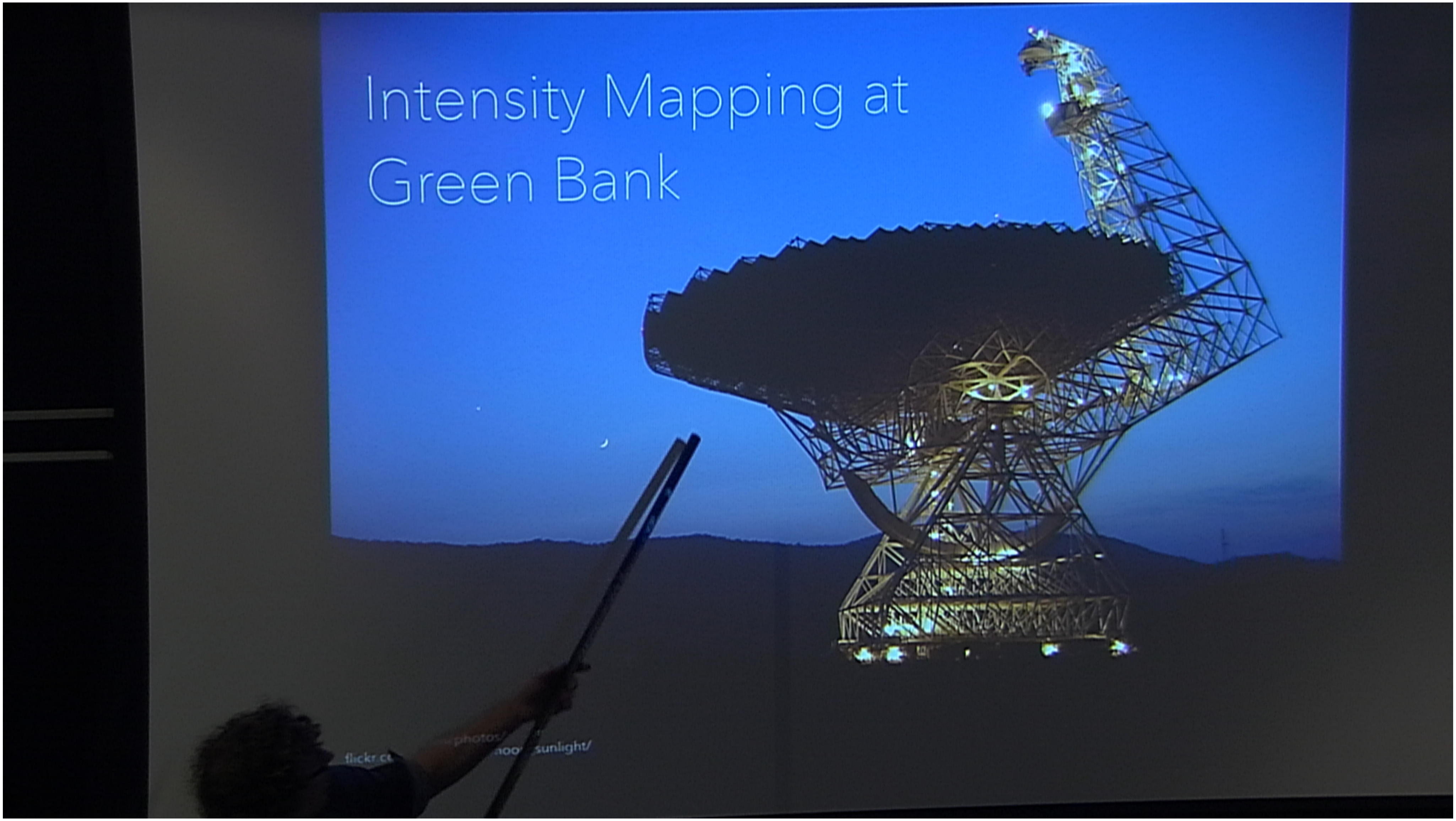

Foreground Challenges
400 MHz
-0.00045
0.00043
450
440
430
420
410
400
f / MHz
45 50 55 60 65 70 75 80 85 90
φ / degrees
Cosmological 21cm Signal ~ 1mK

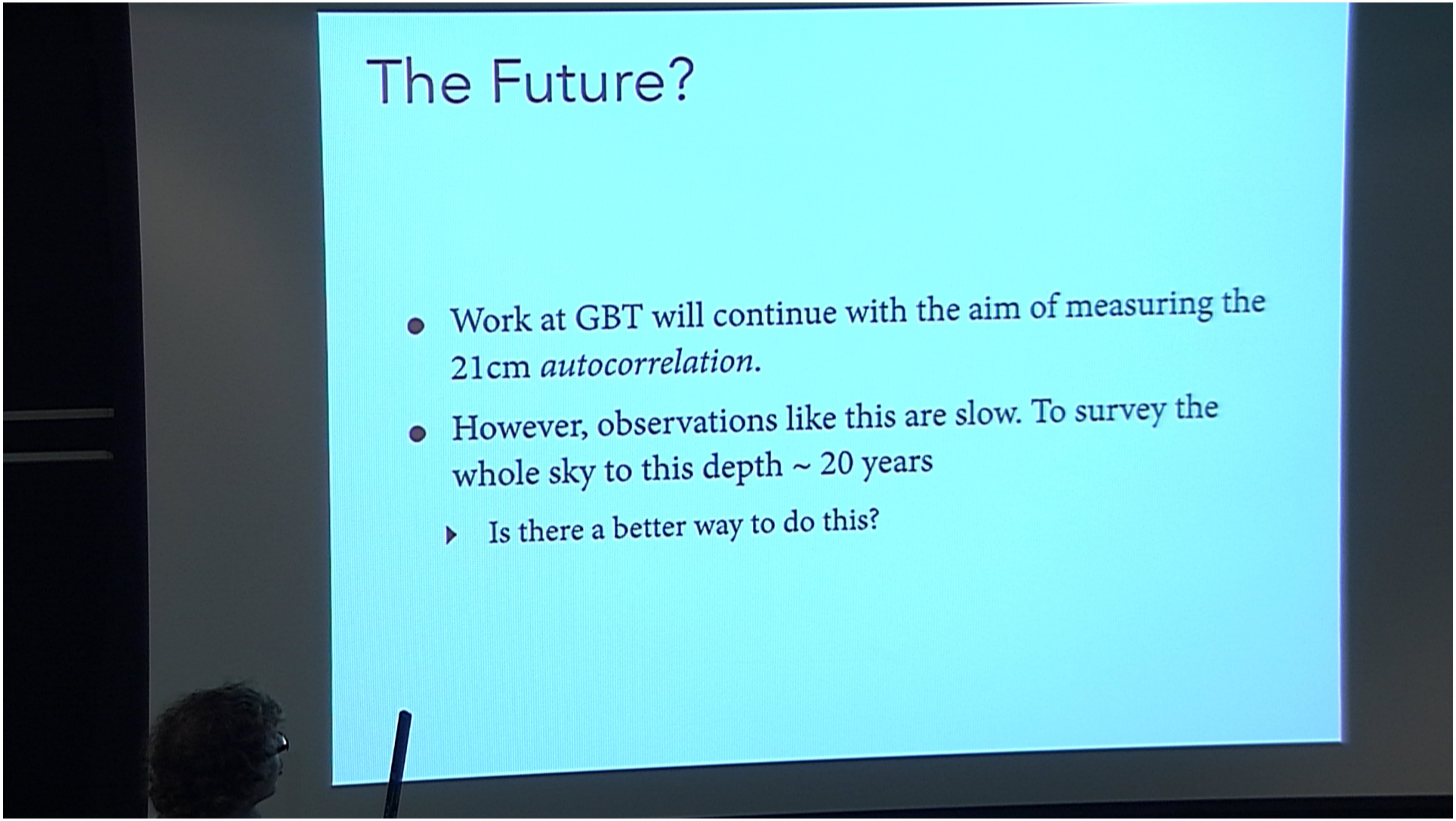

A way out?
400 MHz
400 MHz
φ / degrees
φ / degrees
f / MHz
f / MHz


Mode mixing
High frequency
Low frequency
Instrumental beam
angular direction
frequency
observed intensity
= sum in angular direction

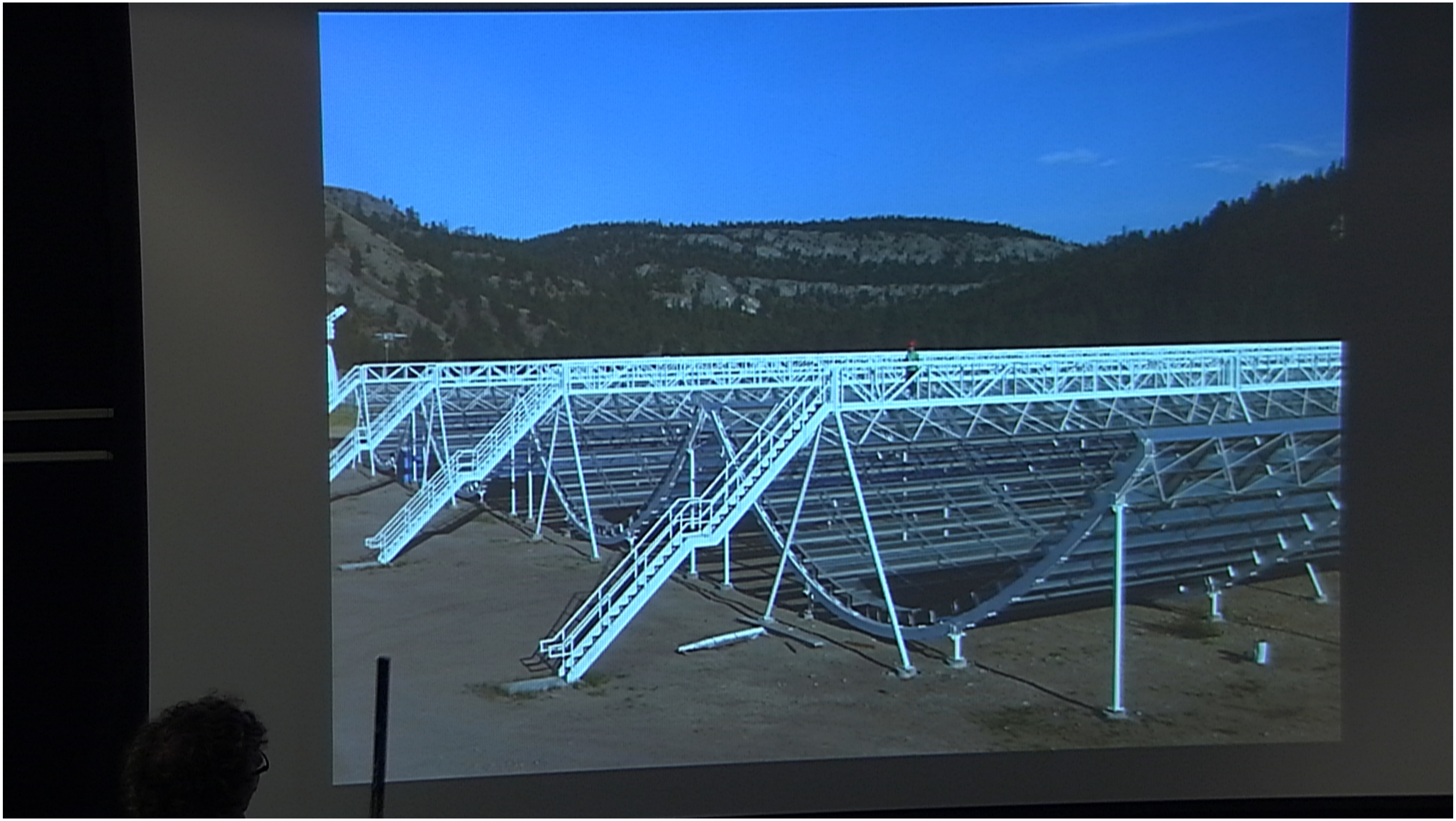

Mode mixing
High frequency
Low frequency
frequency
Instrumental
beam
observed intensity
= sum in angular direction
angular direction


Cross correlation detection
Cross-correlation with of GBT data with DEEP2 Galaxy survey by Chang et al.(2010) - avoids foreground problem!
Updated using WiggleZ survey (Masui et al. 2012)
Cross power spectrum
15 hr
1 hr
$\Omega_{HI}\, b_{HI}\, r = 0.43\ 10^{-3}$
$\Delta^2(k)$ [mK]
1
0.1
0.01
0.001
0.1
k [h / Mpc]
7.4 sigma detection
$\Omega_{\rm HI} = [0.62^{+0.25}_{-0.15}] \times 10^{-3}$
cross-power
deep auto-power
wide auto-power
combined

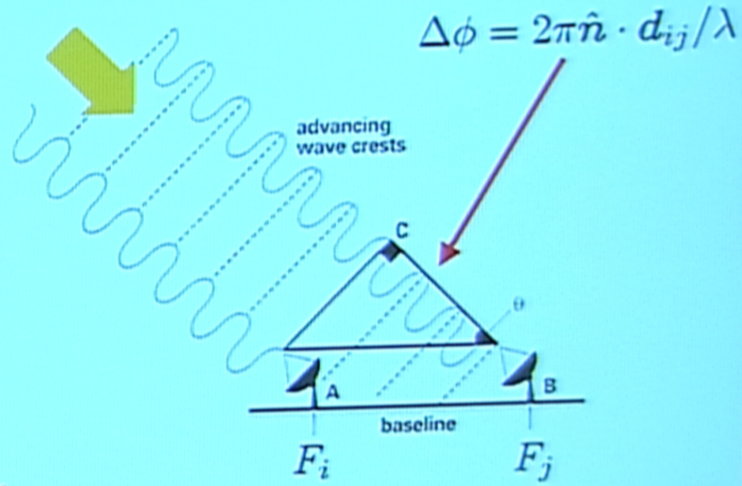

Intensity Mapping at
Green Bank


The Future?
Work at GBT will continue with the aim of measuring the 21cm autocorrelation.
However, observations like this are slow. To survey the whole sky to this depth ~ 20 years
Is there a better way to do this?


McGill
UBC
DRAO
NRC · CNRC
UNIVERSITY OF
TORONTO
chime



Interferometers

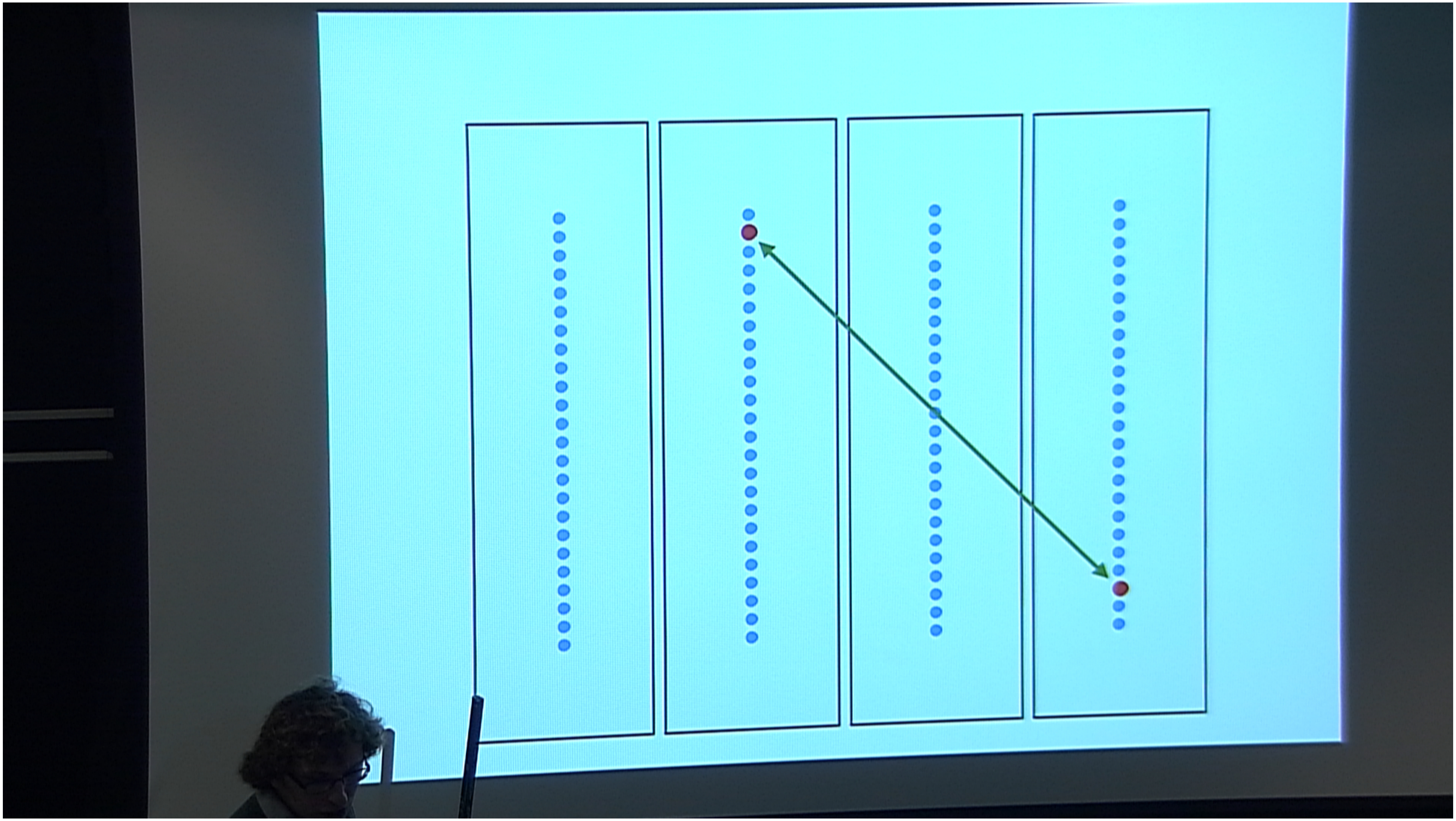


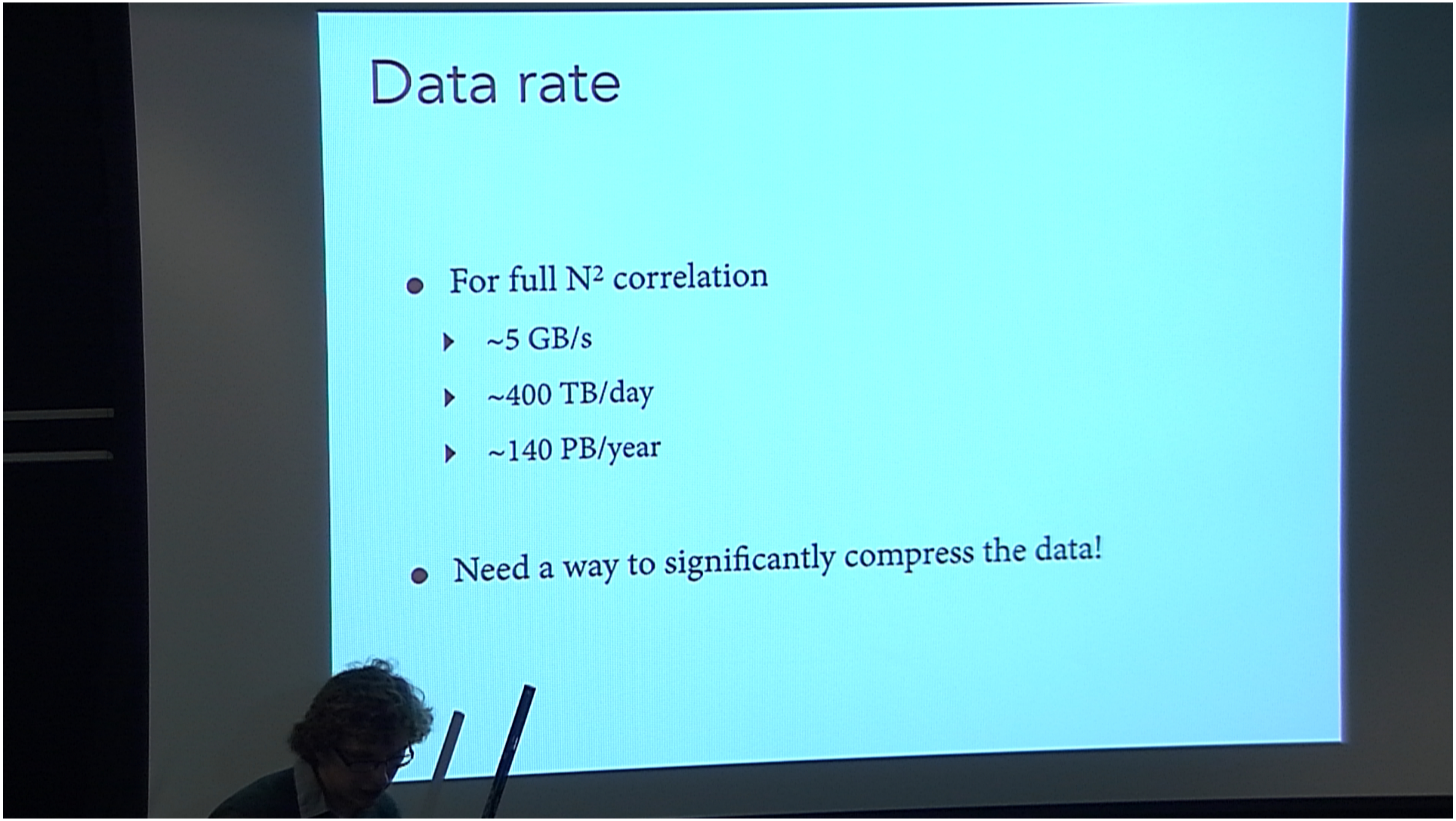
- Visibility is instantaneous correlation of 2 antennas

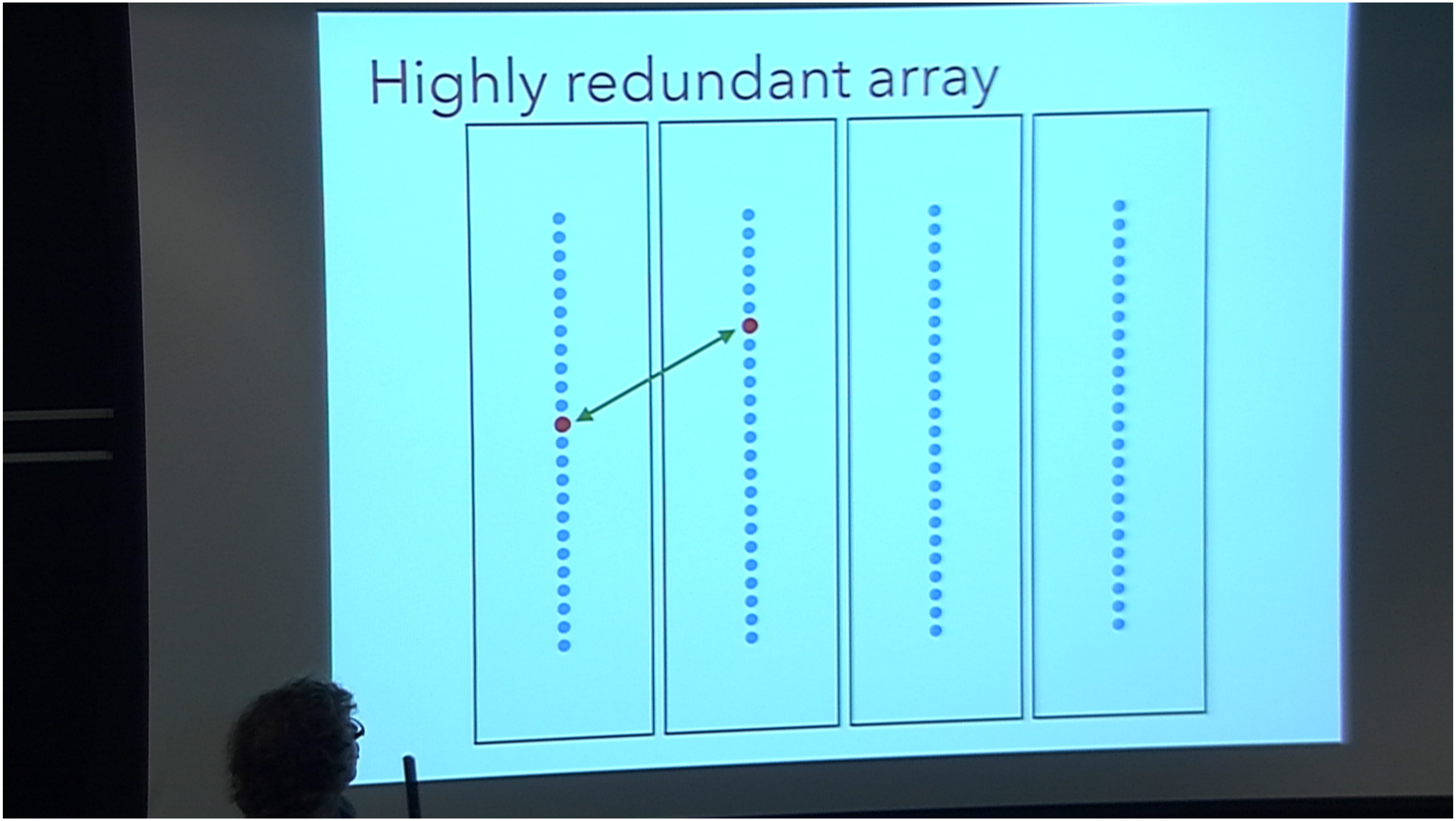
$$V_{ij} = \langle F_i F_j^* \rangle$$

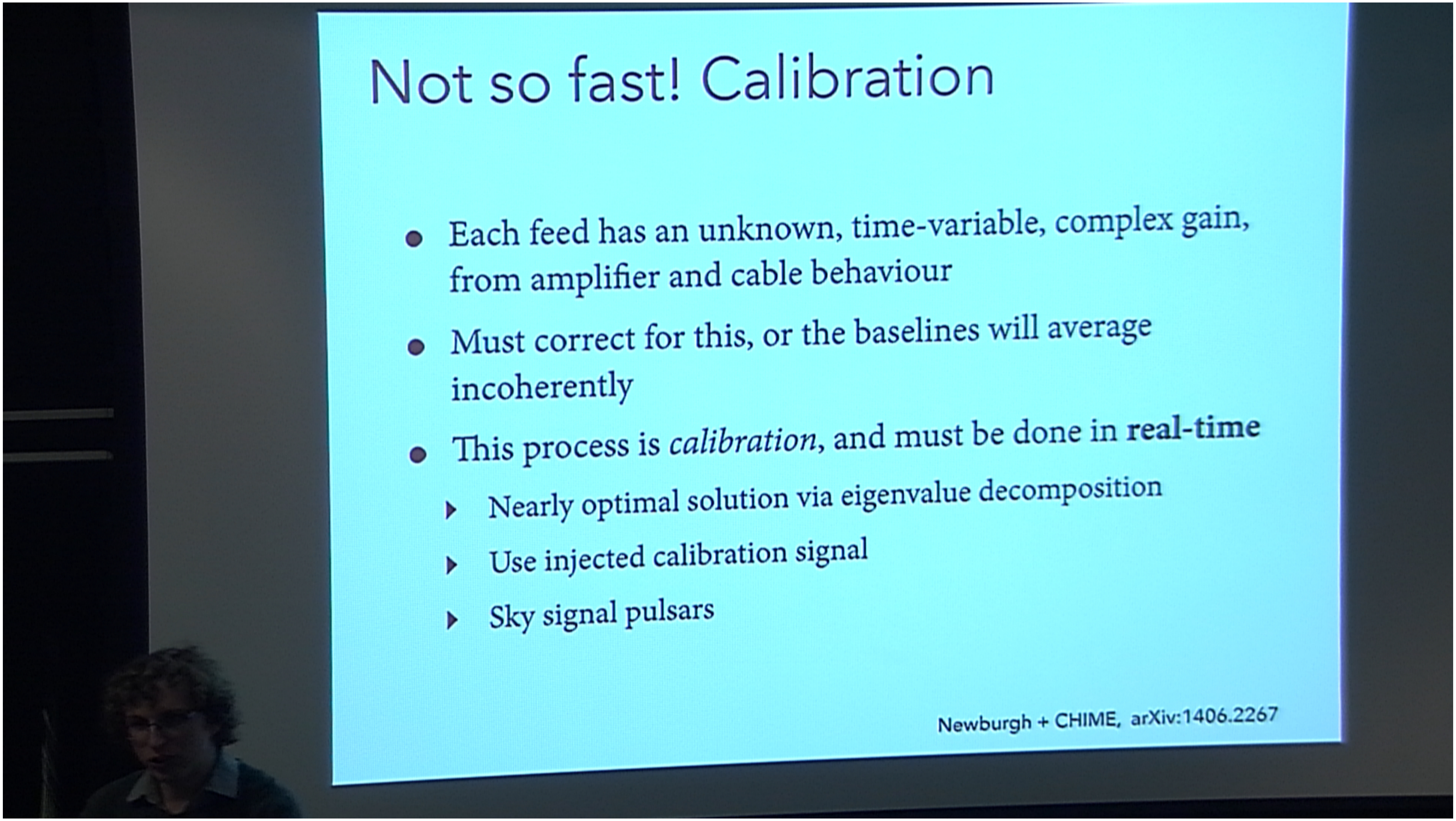
- Each pair measures a Fourier mode of the sky
- Written explicitly:

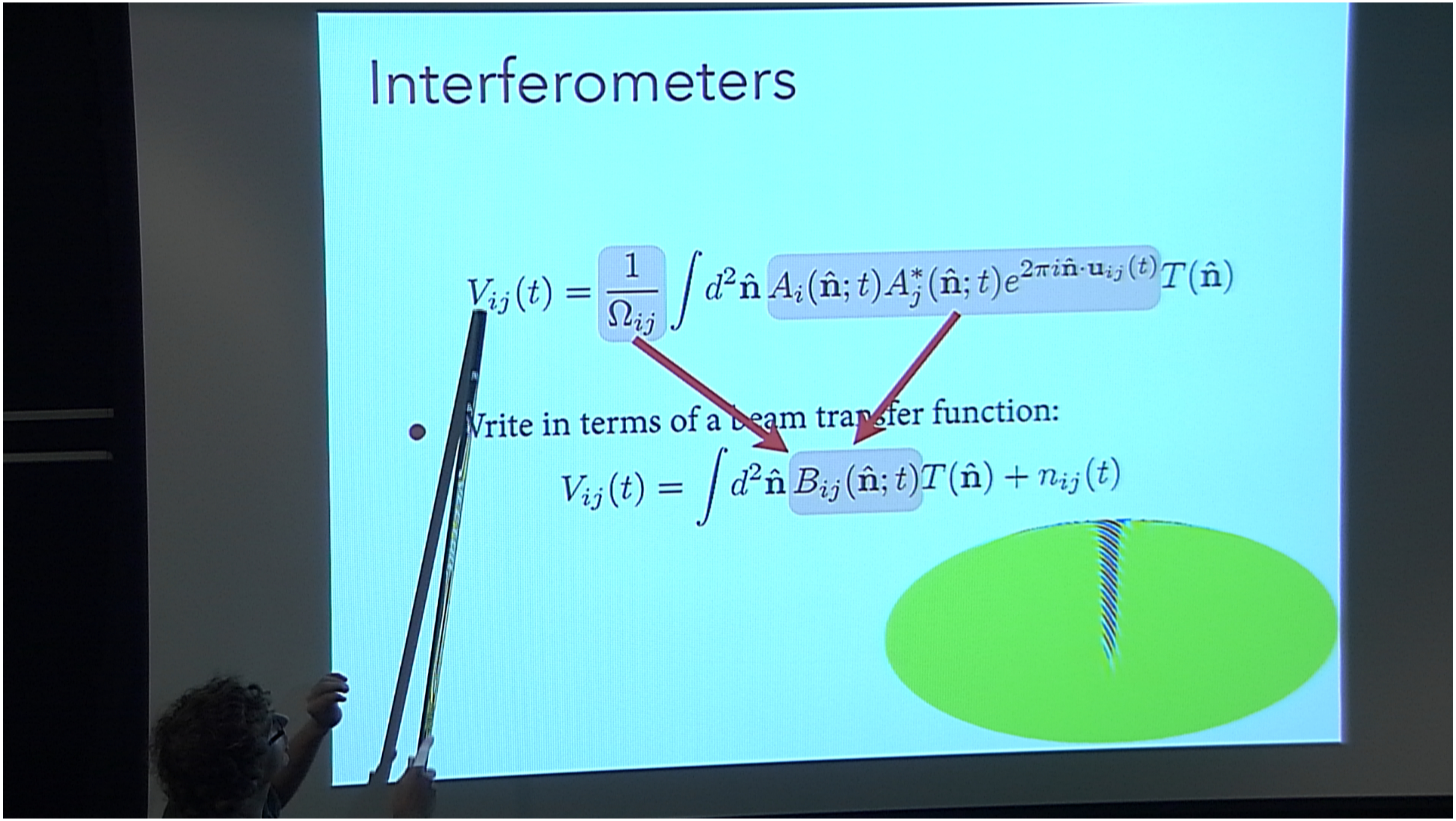
$$V_{ij}(t) = \frac{1}{\Omega_{ij}} \int d^2\hat{\mathbf{n}}\, A_i(\hat{\mathbf{n}};t) A_j^*(\hat{\mathbf{n}};t) e^{2\pi i \hat{\mathbf{n}}\cdot \mathbf{u}_{ij}(t)} T(\hat{\mathbf{n}})$$

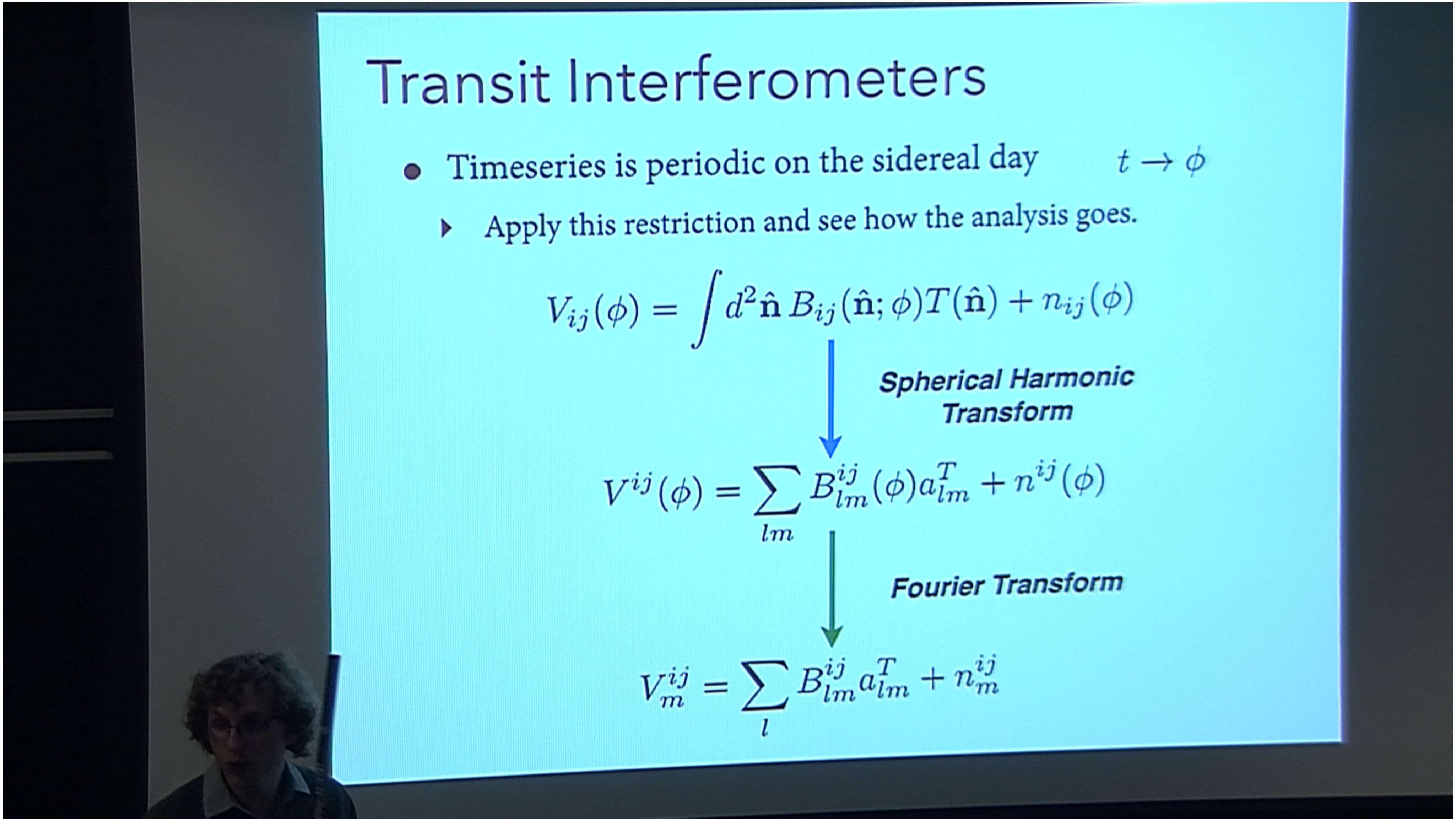


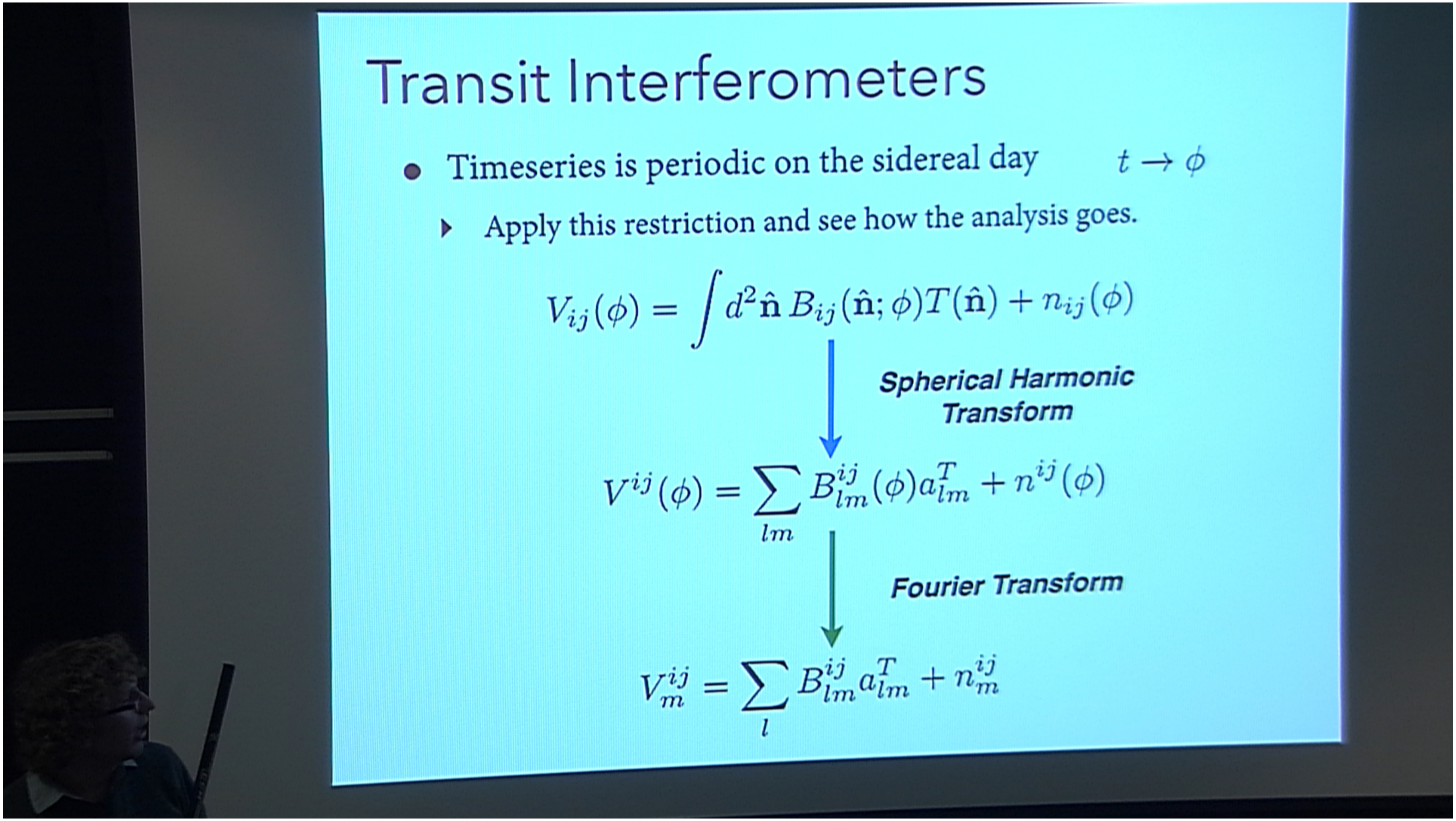

Data rate
For full N2 correlation
~5 GB/s
~400 TB/day
~140 PB/year
Need a way to significantly compress the data!


Highly redundant array


Not so fast! Calibration
Each feed has an unknown, time-variable, complex gain, from amplifier and cable behaviour
Must correct for this, or the baselines will average incoherently
This process is calibration, and must be done in real-time
Nearly optimal solution via eigenvalue decomposition
Use injected calibration signal
Sky signal pulsars
Newburgh + CHIME, arXiv:1406.2267


Interferometers
$V_{ij}(t) = \frac{1}{\Omega_{ij}} \int d^2\hat{\mathbf{n}}\, A_i(\hat{\mathbf{n}};t) A_j^*(\hat{\mathbf{n}};t) e^{2\pi i \hat{\mathbf{n}}\cdot\mathbf{u}_{ij}(t)} T(\hat{\mathbf{n}})$
Write in terms of a beam transfer function:
$V_{ij}(t) = \int d^2\hat{\mathbf{n}}\, B_{ij}(\hat{\mathbf{n}};t) T(\hat{\mathbf{n}}) + n_{ij}(t)$


Transit Interferometers
Timeseries is periodic on the sidereal day $t \to \phi$
Apply this restriction and see how the analysis goes.
$V_{ij}(\phi) = \int d^2\hat{\mathbf{n}}\, B_{ij}(\hat{\mathbf{n}};\phi) T(\hat{\mathbf{n}}) + n_{ij}(\phi)$
Spherical Harmonic Transform
$V^{ij}(\phi) = \sum_{lm} B^{ij}_{lm}(\phi) a^T_{lm} + n^{ij}(\phi)$
Fourier Transform
$V^{ij}_m = \sum_l B^{ij}_{lm} a^T_{lm} + n^{ij}_m$


Transit Interferometers
Timeseries is periodic on the sidereal day $t \to \phi$
Apply this restriction and see how the analysis goes.
$V_{ij}(\phi) = \int d^2\hat{\mathbf{n}}\, B_{ij}(\hat{\mathbf{n}};\phi)T(\hat{\mathbf{n}}) + n_{ij}(\phi)$
Spherical Harmonic Transform
$V^{ij}(\phi) = \sum_{lm} B^{ij}_{lm}(\phi)a^T_{lm} + n^{ij}(\phi)$
Fourier Transform
$V^{ij}_m = \sum_l B^{ij}_{lm}a^T_{lm} + n^{ij}_m$

m-mode transform

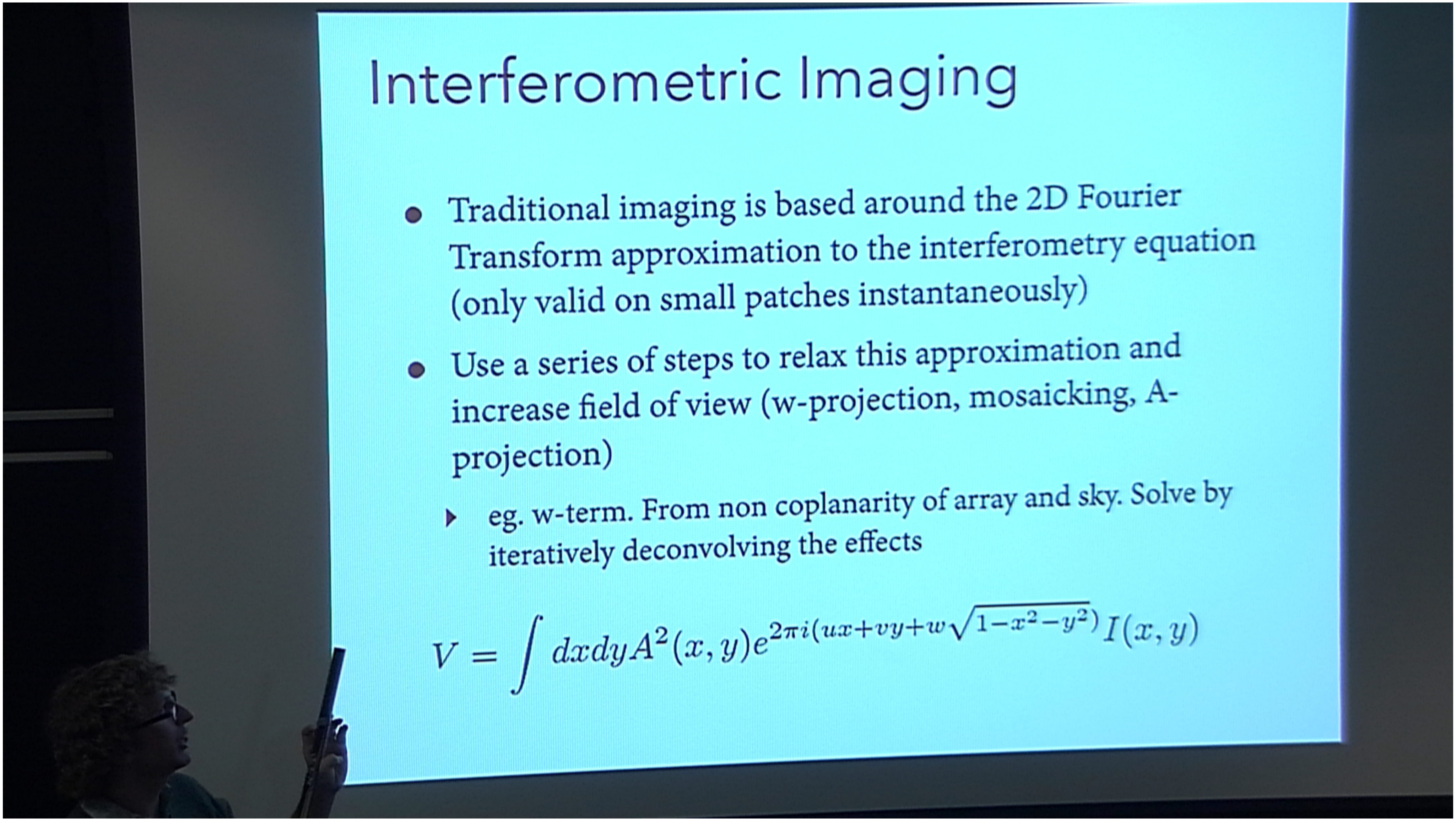
- Mapping does not mix m's (each is independent)

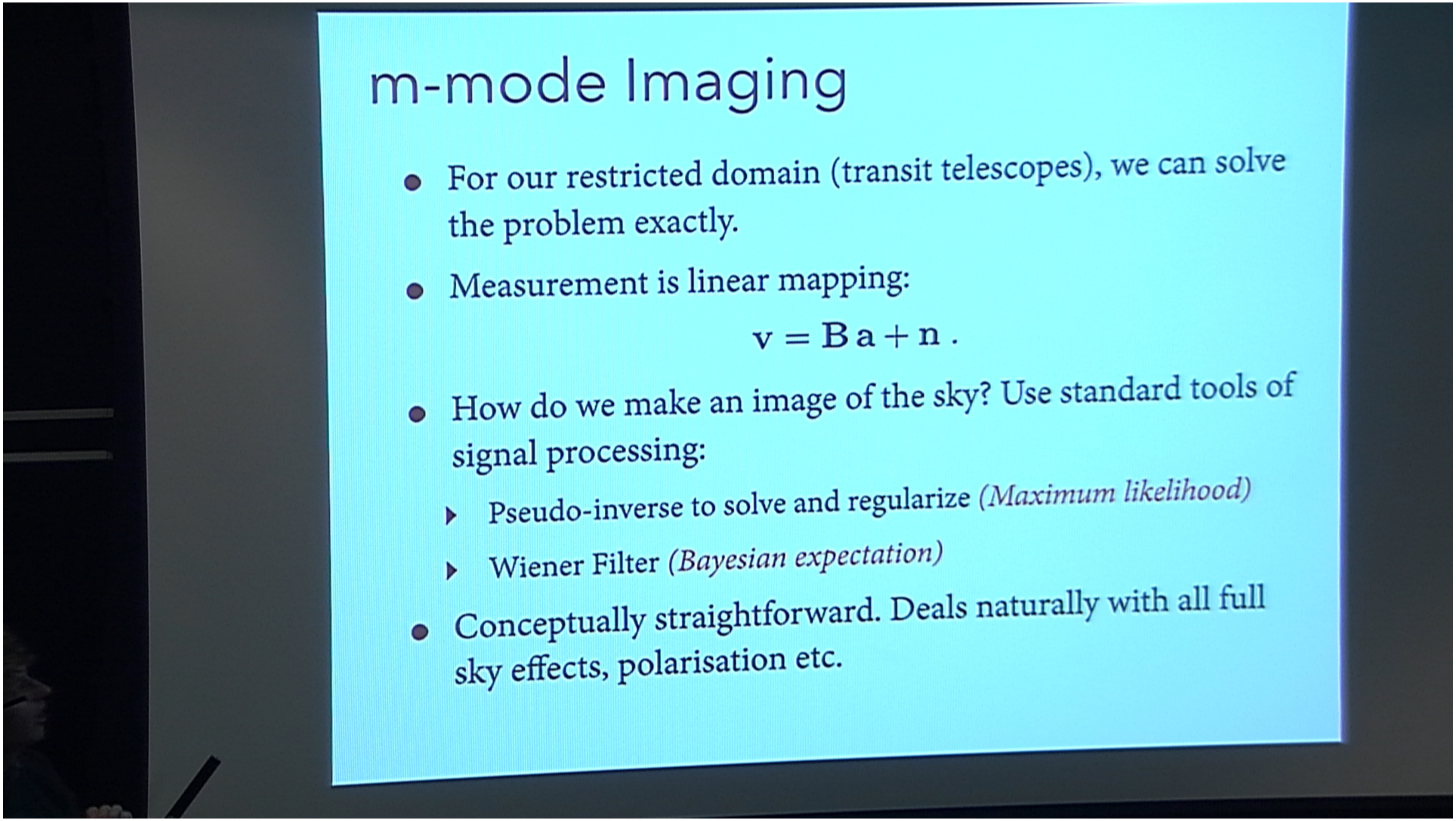
$$V_m^\alpha = \sum_l B_{lm}^\alpha a_{lm}^T + n_m^\alpha$$

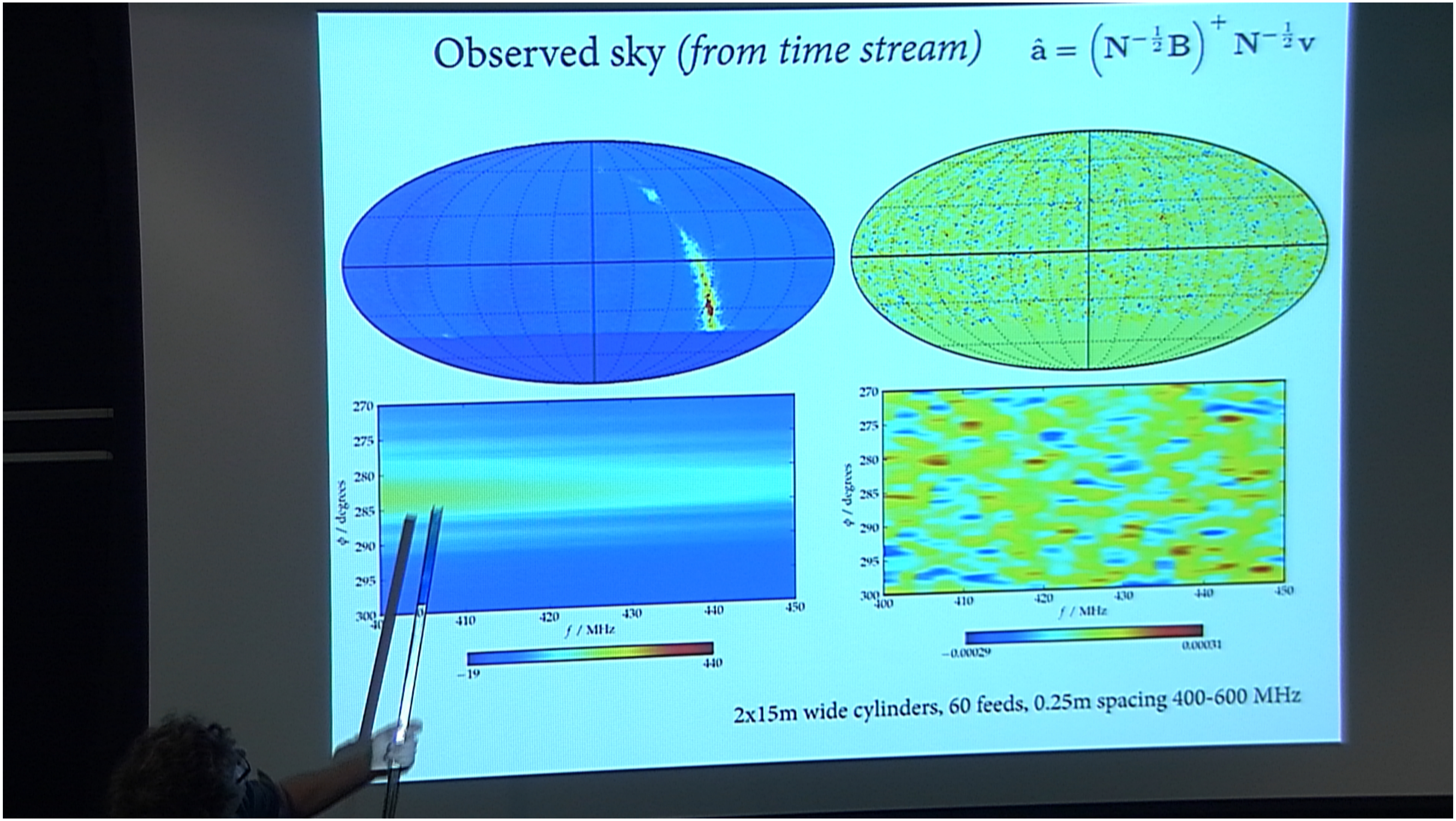
- Write in vector form

$$\mathbf{v} = \mathbf{B}\,\mathbf{a} + \mathbf{n}\,.$$

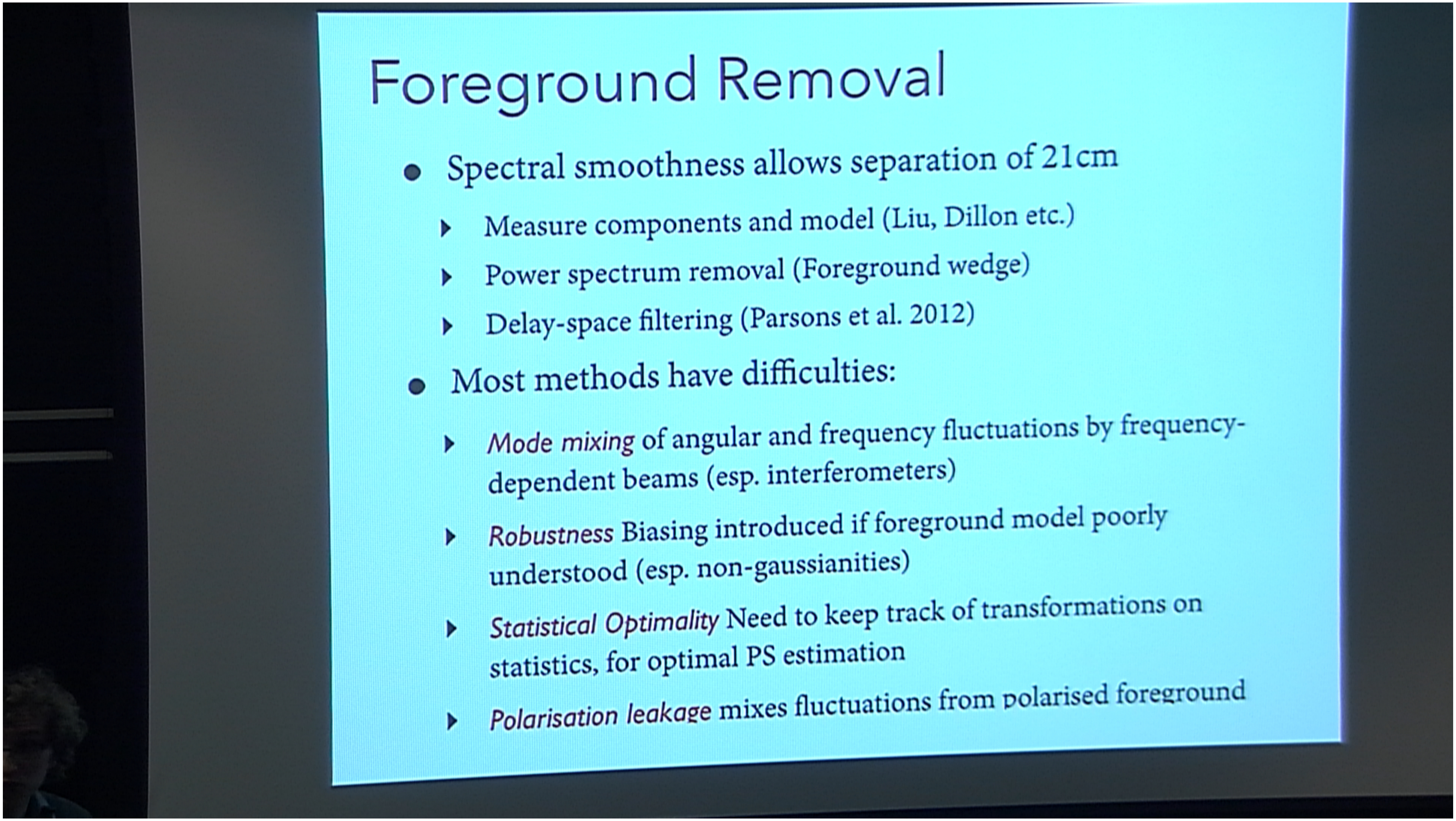
- Simple, linear mapping from the information on the sky, to the measured degrees of freedom
- Discrete relation, with finite number of degrees, can apply all the standard statistical, signal processing techniques.
- Computationally efficient: For 1000 m's an $O(N^3)$ matrix operation becomes 10^6 times faster

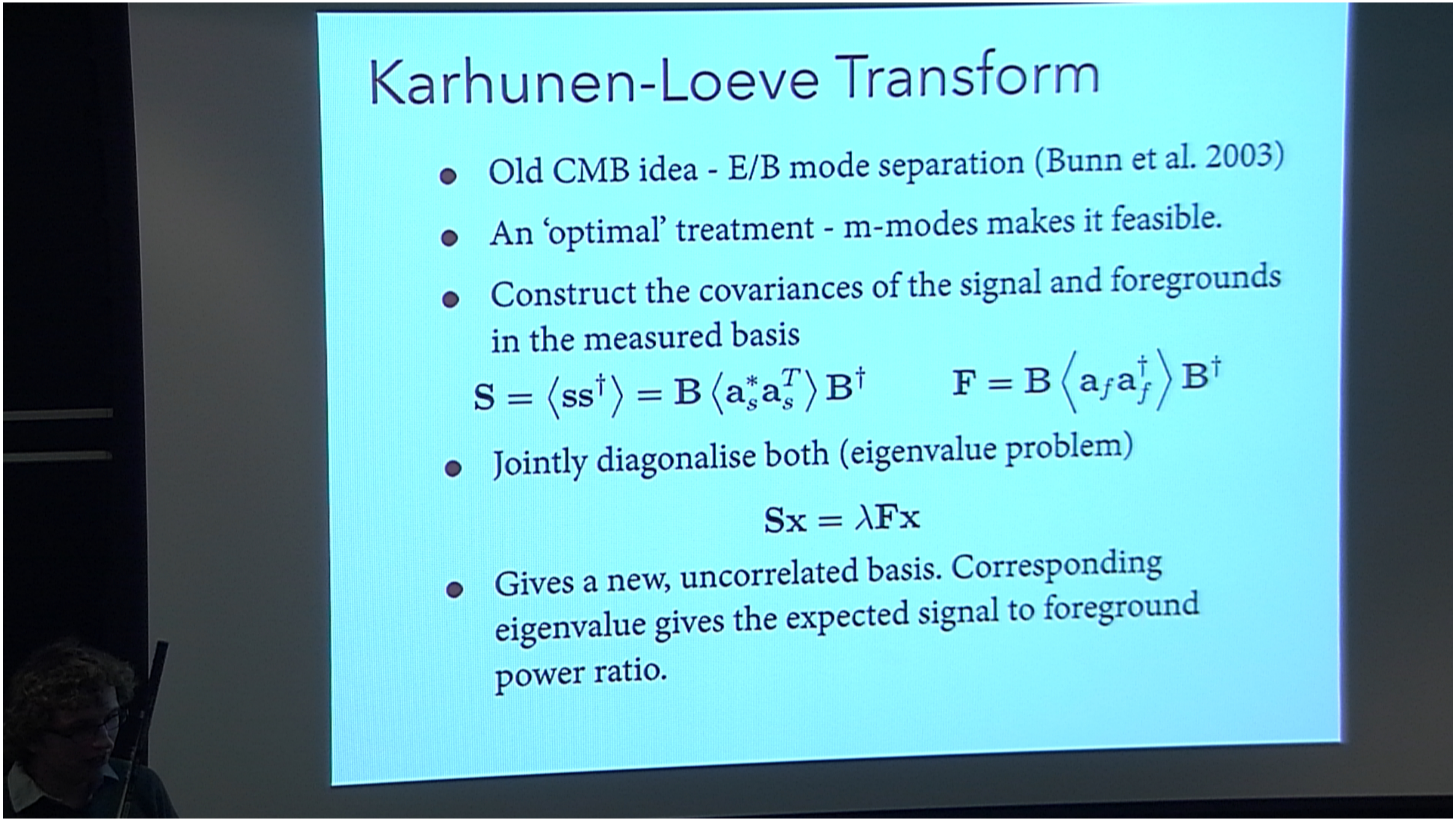

Interferometric Imaging
Traditional imaging is based around the 2D Fourier Transform approximation to the interferometry equation (only valid on small patches instantaneously)
Use a series of steps to relax this approximation and increase field of view (w-projection, mosaicking, A-projection)
eg. w-term. From non coplanarity of array and sky. Solve by iteratively deconvolving the effects
$V = \int dx dy A^2(x,y) e^{2\pi i (ux + vy + w\sqrt{1-x^2-y^2})} I(x,y)$

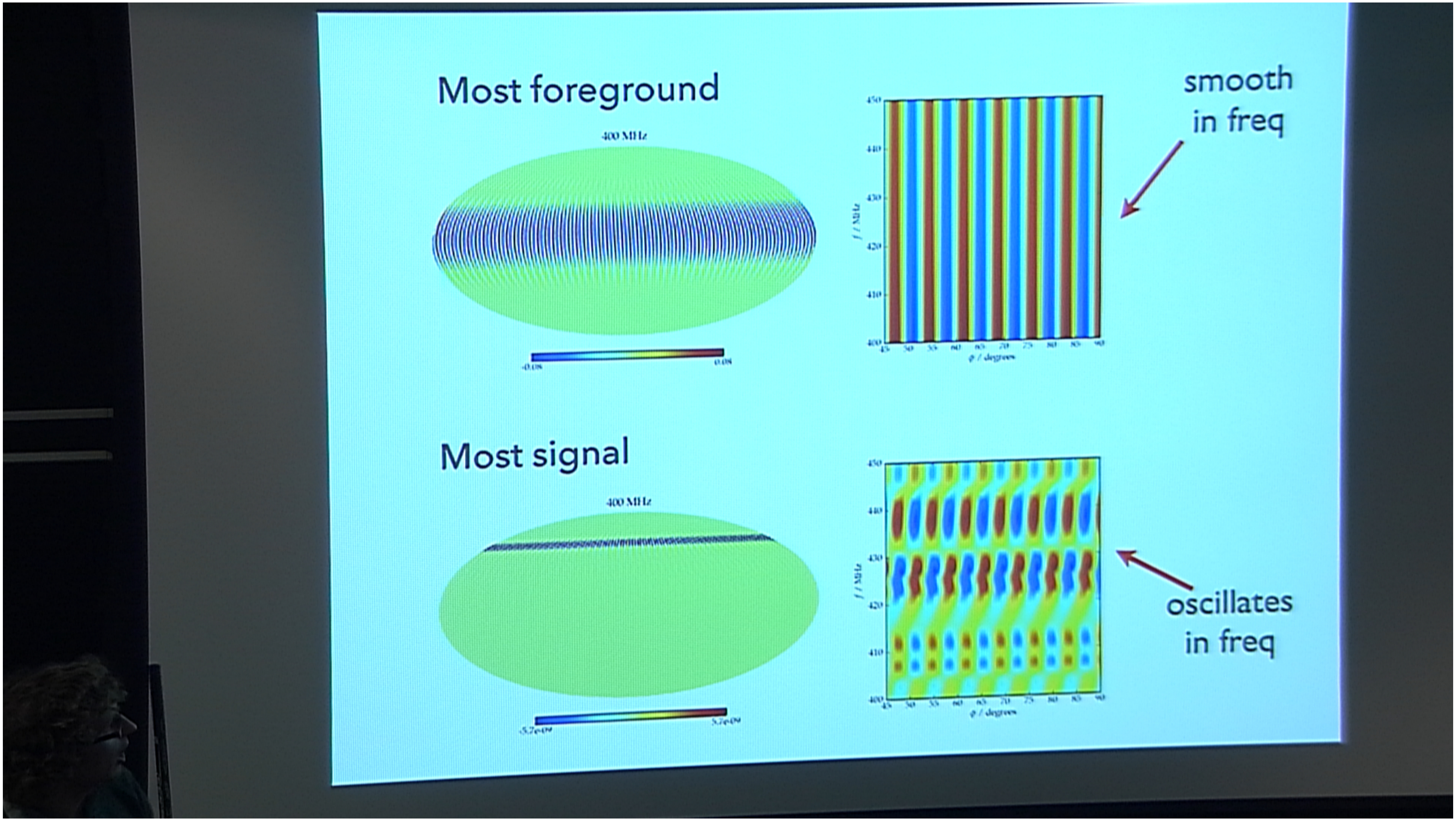

m-mode Imaging
For our restricted domain (transit telescopes), we can solve the problem exactly.
Measurement is linear mapping:
v = B a + n .
How do we make an image of the sky? Use standard tools of signal processing:
Pseudo-inverse to solve and regularize (Maximum likelihood)
Wiener Filter (Bayesian expectation)
Conceptually straightforward. Deals naturally with all full sky effects, polarisation etc.



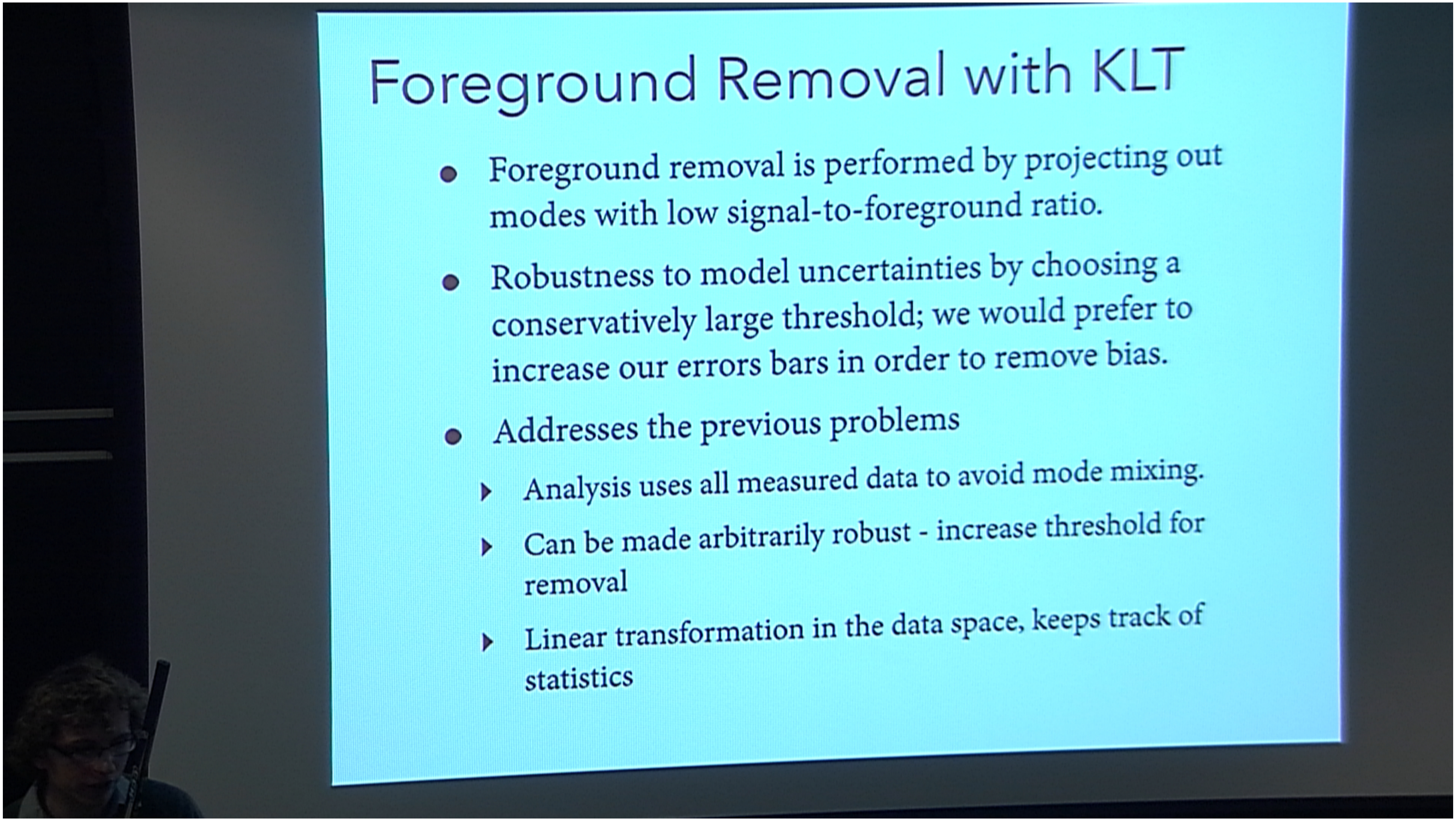
Observed sky (*from time stream*)
$\hat{\mathbf{a}} = \left(\mathbf{N}^{-\frac{1}{2}}\mathbf{B}\right)^{+}\mathbf{N}^{-\frac{1}{2}}\mathbf{v}$
ϕ / degrees
f / MHz
−19
440
−0.00029
0.00031
2x15m wide cylinders, 60 feeds, 0.25m spacing 400-600 MHz

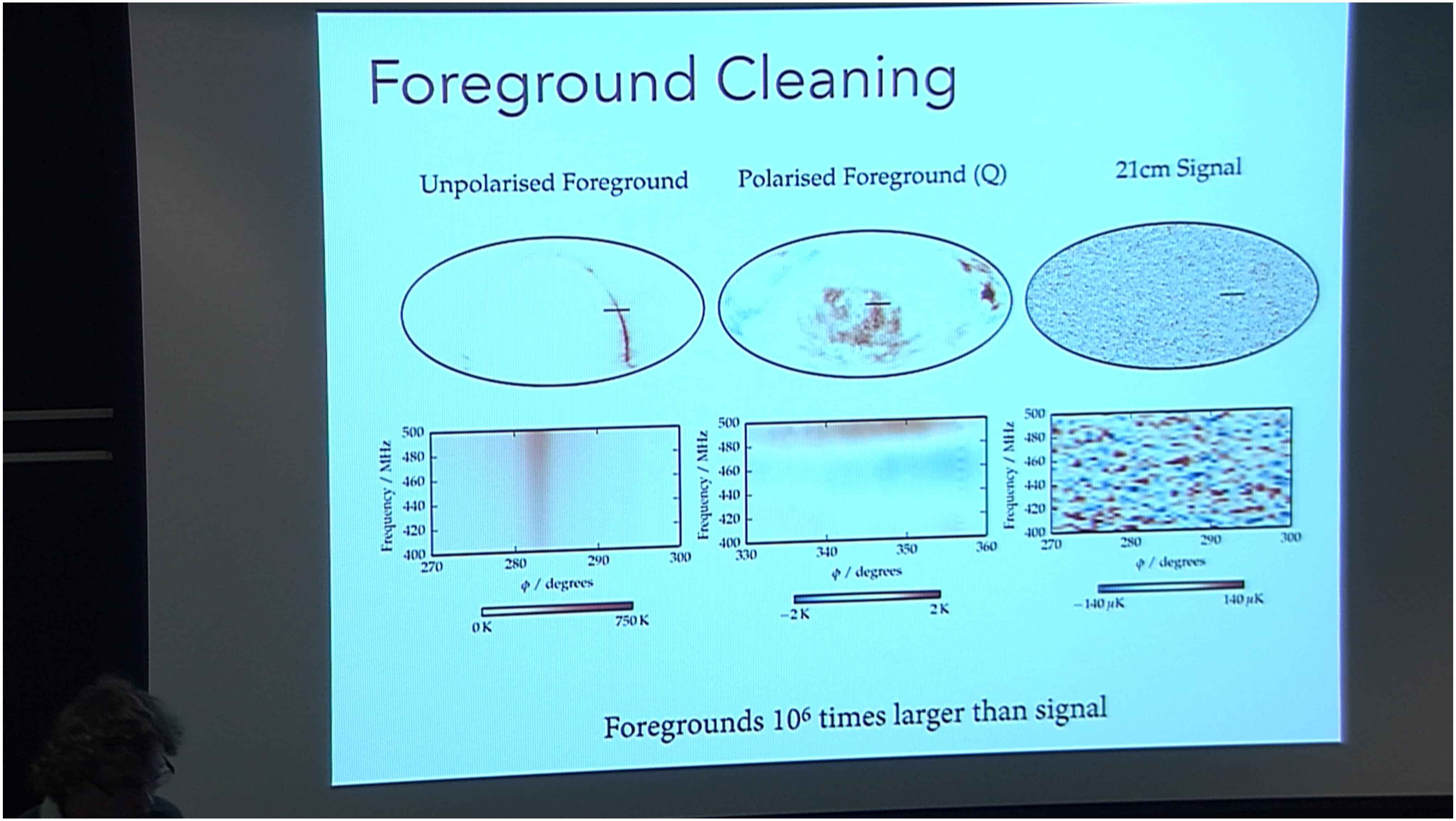

Foreground Removal
Spectral smoothness allows separation of 21cm
Measure components and model (Liu, Dillon etc.)
Power spectrum removal (Foreground wedge)
Delay-space filtering (Parsons et al. 2012)
Most methods have difficulties:
Mode mixing of angular and frequency fluctuations by frequency-dependent beams (esp. interferometers)
Robustness Biasing introduced if foreground model poorly understood (esp. non-gaussianities)
Statistical Optimality Need to keep track of transformations on statistics, for optimal PS estimation
Polarisation leakage mixes fluctuations from polarised foreground


Karhunen-Loeve Transform
Old CMB idea - E/B mode separation (Bunn et al. 2003)
An 'optimal' treatment - m-modes makes it feasible.
Construct the covariances of the signal and foregrounds in the measured basis
$\mathbf{S} = \langle \mathbf{s}\mathbf{s}^\dagger \rangle = \mathbf{B} \langle \mathbf{a}_s^* \mathbf{a}_s^T \rangle \mathbf{B}^\dagger$
$\mathbf{F} = \mathbf{B} \langle \mathbf{a}_f \mathbf{a}_f^\dagger \rangle \mathbf{B}^\dagger$
Jointly diagonalise both (eigenvalue problem)
$\mathbf{S}\mathbf{x} = \lambda \mathbf{F}\mathbf{x}$
Gives a new, uncorrelated basis. Corresponding eigenvalue gives the expected signal to foreground power ratio.

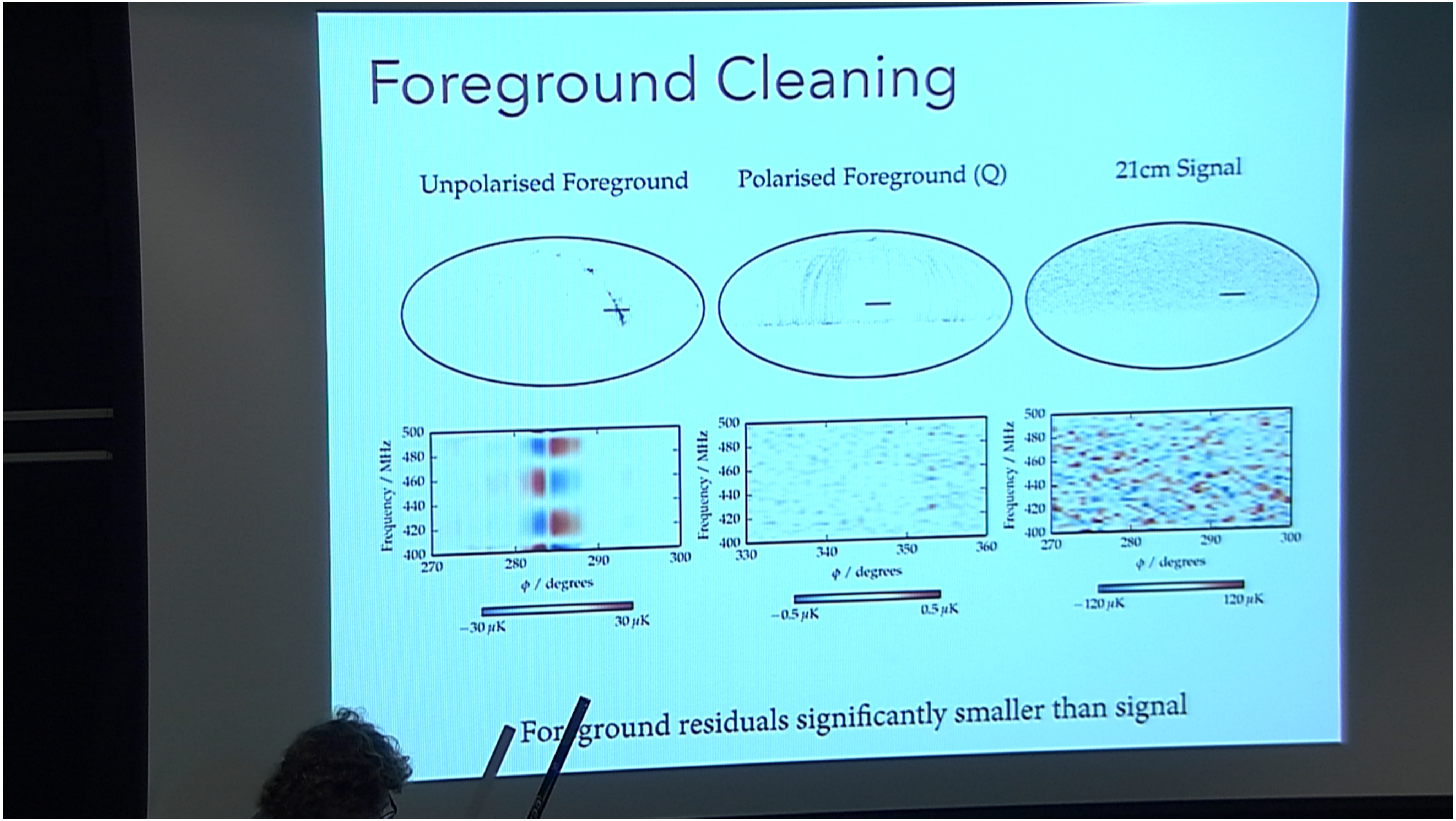


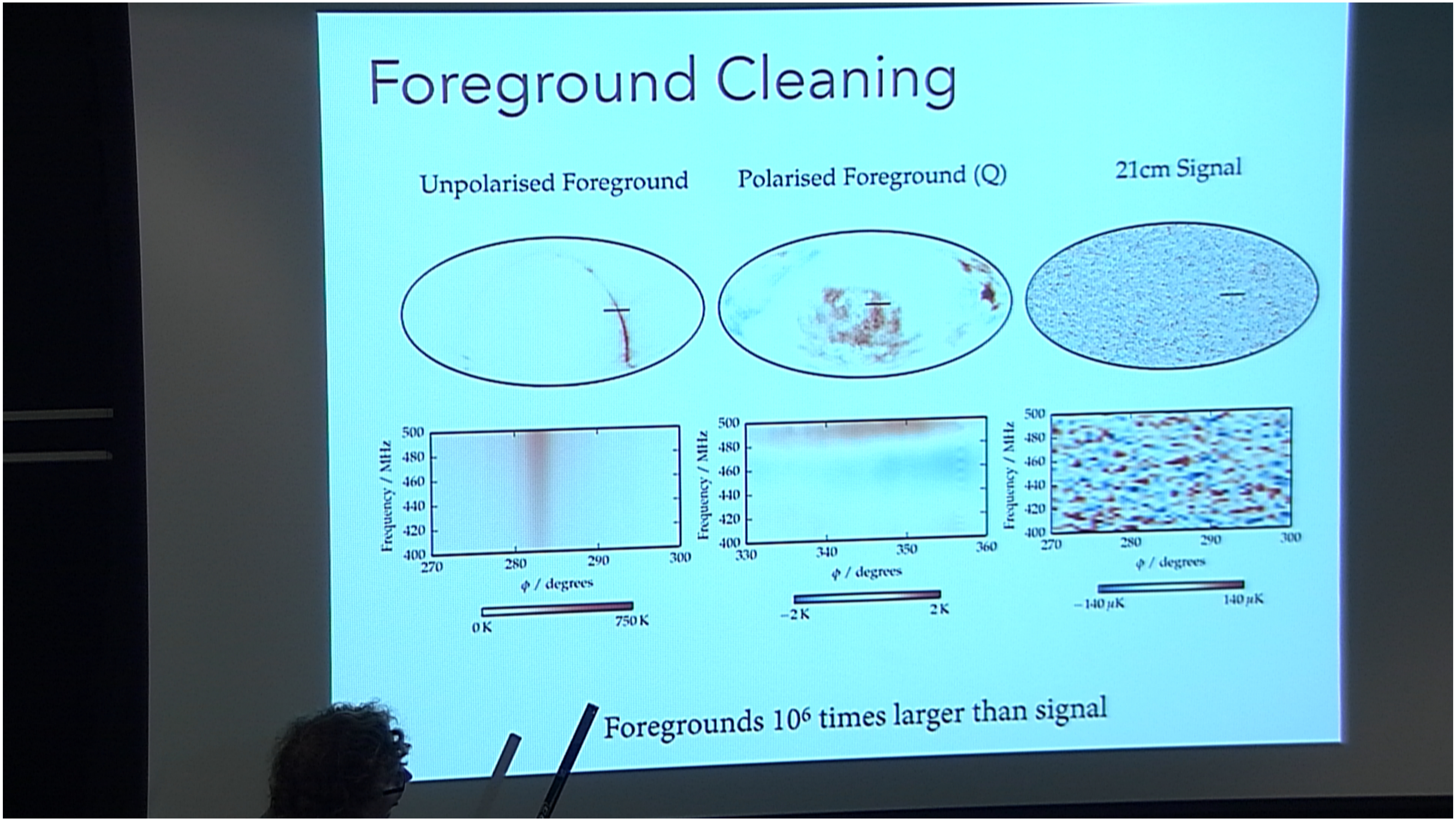
Most foreground
400 MHz
smooth
in freq
Most signal
400 MHz
oscillates
in freq

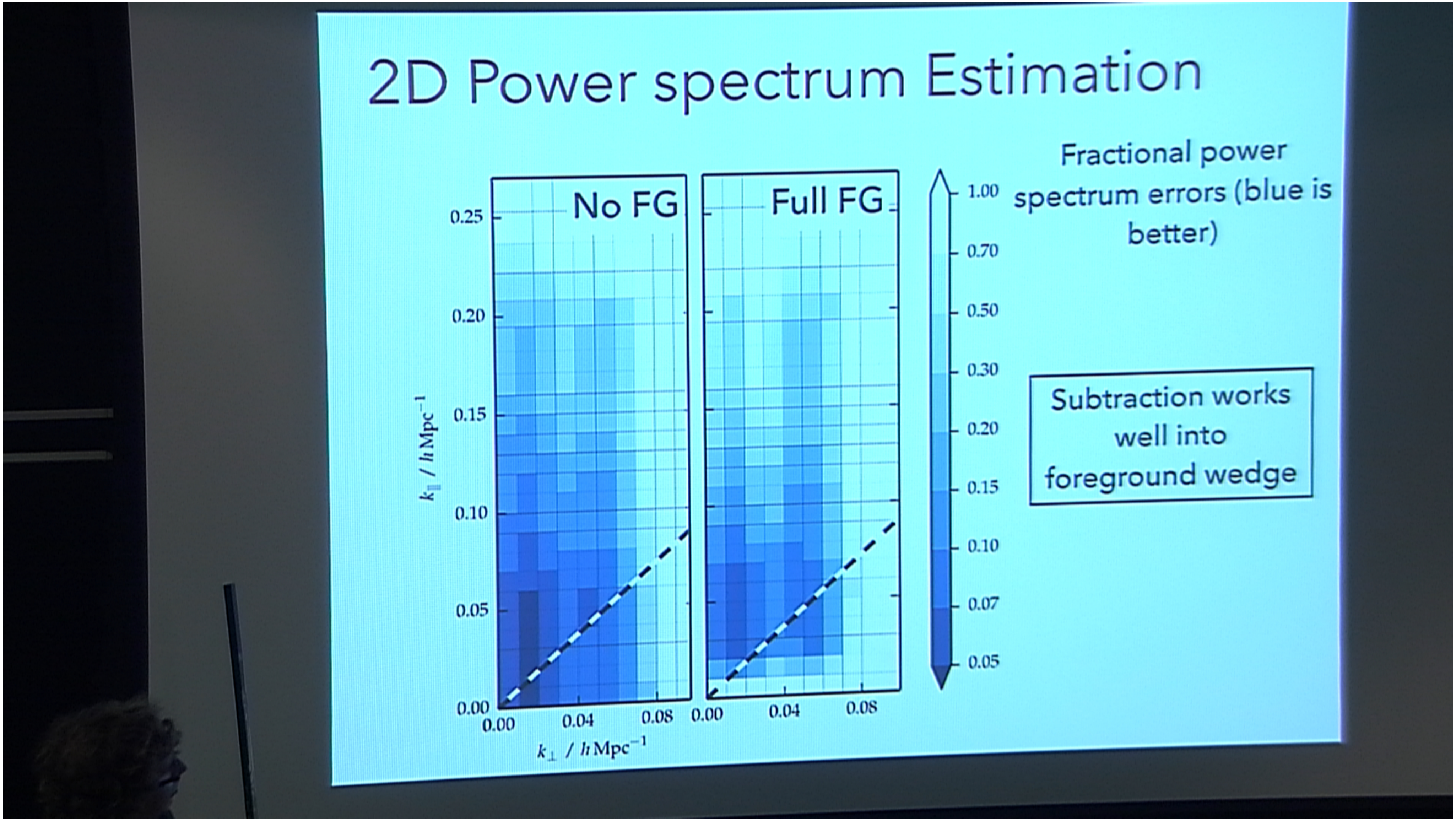

Foreground Removal with KLT
Foreground removal is performed by projecting out modes with low signal-to-foreground ratio.
Robustness to model uncertainties by choosing a conservatively large threshold; we would prefer to increase our errors bars in order to remove bias.
Addresses the previous problems
Analysis uses all measured data to avoid mode mixing.
Can be made arbitrarily robust - increase threshold for removal
Linear transformation in the data space, keeps track of statistics



Foreground Cleaning
Unpolarised Foreground
Polarised Foreground (Q)
21cm Signal
Frequency / MHz
500
480
460
440
420
400
270
280
290
300
330
340
350
360
φ / degrees
0 K
750 K
−2 K
2 K
−140 μK
140 μK
Foregrounds 10⁶ times larger than signal


Foreground Cleaning
Unpolarised Foreground
Polarised Foreground (Q)
21cm Signal
Frequency / MHz
ϕ / degrees
−30 μK
30 μK
−0.5 μK
0.5 μK
−120 μK
120 μK
For ground residuals significantly smaller than signal


Foreground Cleaning
Unpolarised Foreground
Polarised Foreground (Q)
21cm Signal
Frequency / MHz
φ / degrees
0 K
750 K
−2 K
2 K
−140 μK
140 μK
Foregrounds 10⁶ times larger than signal


2D Power spectrum Estimation
No FG
Full FG
Fractional power spectrum errors (blue is better)
Subtraction works well into foreground wedge
$k_\parallel$ / h Mpc^{-1}
$k_\perp$ / h Mpc^{-1}
0.00
0.05
0.10
0.15
0.20
0.25
0.04
0.08
1.00
0.70
0.50
0.30
0.20
0.15
0.10
0.07
0.05

Summary

- Data Science is becoming an increasingly large and distinct part of physics
- 21cm Intensity Mapping is a promising technique for mapping the Universe and measuring BAOs.
- Data volume and foregrounds are challenging
- New techniques, like the m-mode formalism show promise for surmounting them