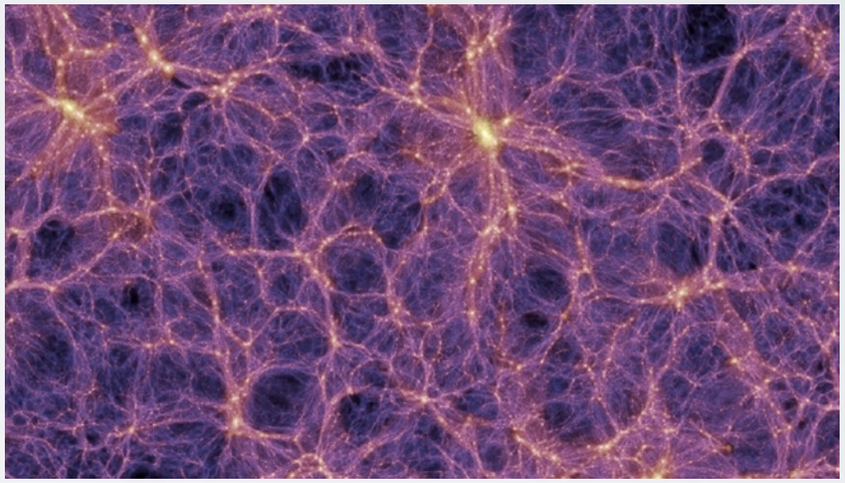
Donough Regan (Sussex University)

Dr StrangeCosmos
or:
How I learned to stop worrying and love Data Analytics



Sussex Data Analysis Forum Meeting
Tuesday 28 February 2017

Two Problems - One Ethos

(a) Calculation of redshift space distortions using the EFTofLSS

Standard perturbation theory breaks down on scales far larger than we can probe.

Can we extend the range using EFTofLSS?

Work with: L. Fonseca de la Bella , D. Seery, S. Hotchkiss

Ethos: Make the nonlinear calculations more **efficient**

(b) 3 point function to improve cosmological constraints

Optimal measurement the bispectrum can become intractable due to the covariance matrix.

Is there a sufficient statistic we can use without much information loss?

Work with: J.Byun, A. Eggemeier, D. Seery, R. Smith

Ethos: Make the measurements more **efficient**

Old approach

Define the problem
+ Do the calculations!

Prepare the report

Data Analysis Phases

- Get the data
- Clean the data
- Exploratory analysis
- Statistical analysis/Inferences
- Interpret + Challenge
- Report + Reproducible Research

Is that not what I've been doing?

~~New approach~~

(Usual) Data Analytics Pipeline

Data Mining
/Predictive Modelling

Problem Definition
+ ETL / Analytics

Usually dominant part

Data Output
(formatted)

Exploratory
Analysis

Stats,
Interpretations,
Challenge
Results

Reports
+ Paper writing
+ Reproducibility

Origin of perturbations

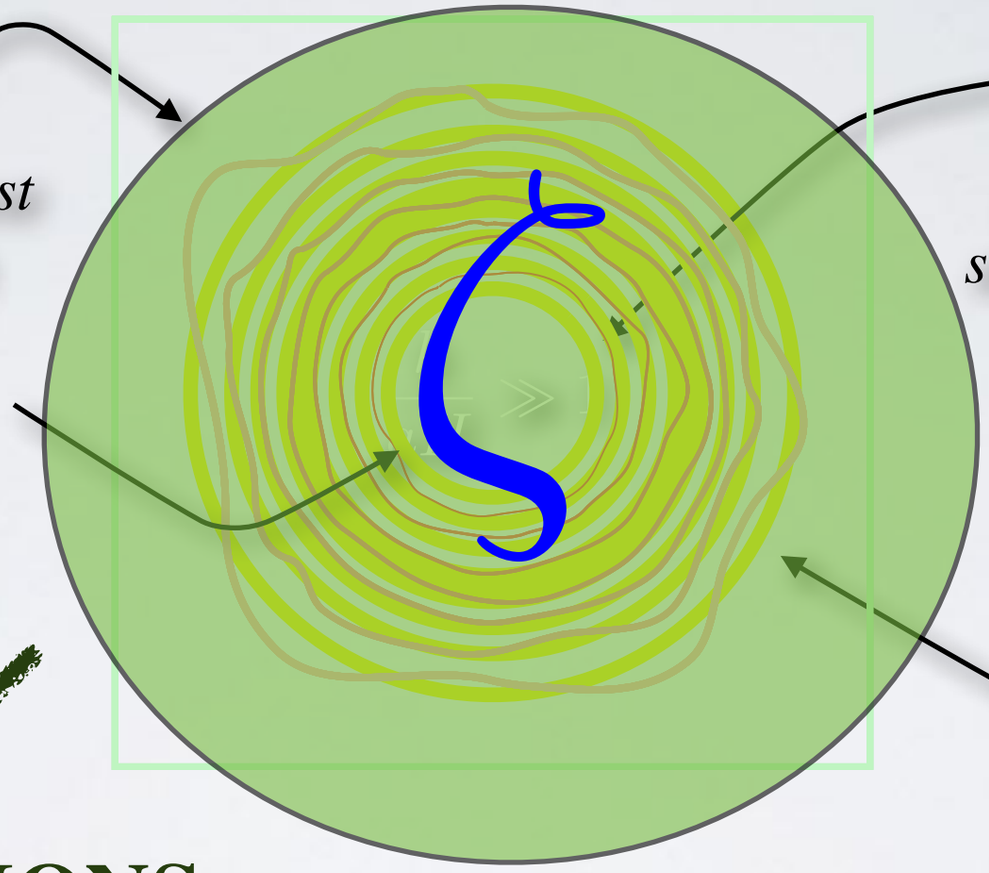
QUANTUM FLUCTUATIONS

The initial, almost flat, almost smooth inflationary patch

$$\frac{k}{aH} \ll 1$$

Horizon contracts

Modes outside the horizon stop oscillating and "freeze"

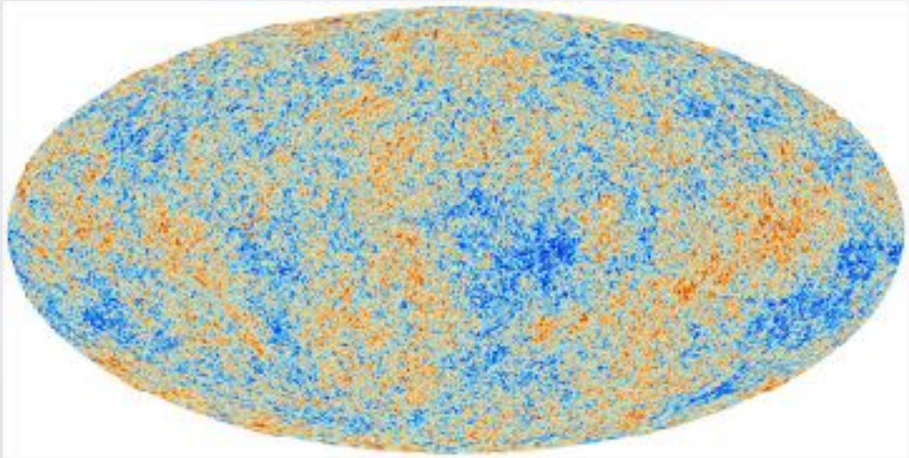


The universe reheats and the horizon expands again. Fluctuations inside the horizon set the initial conditions for radiation and matter perturbations

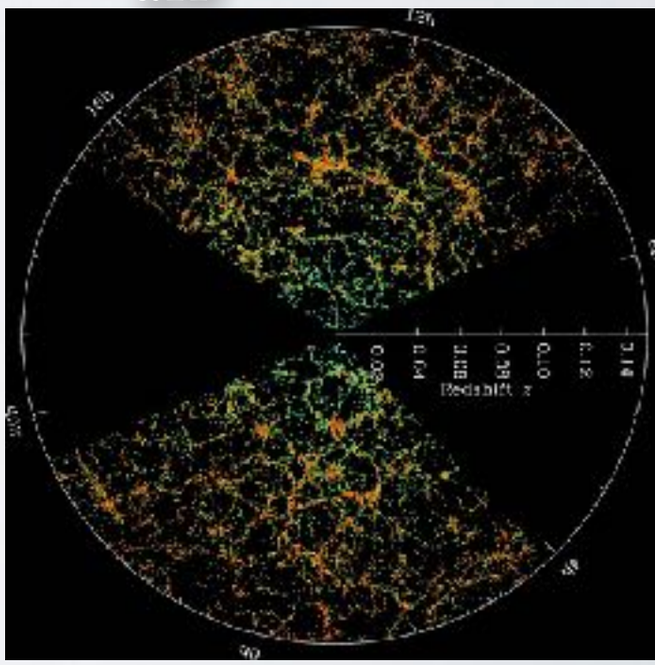
Initial horizon size

$$\frac{k}{aH} \simeq 1$$

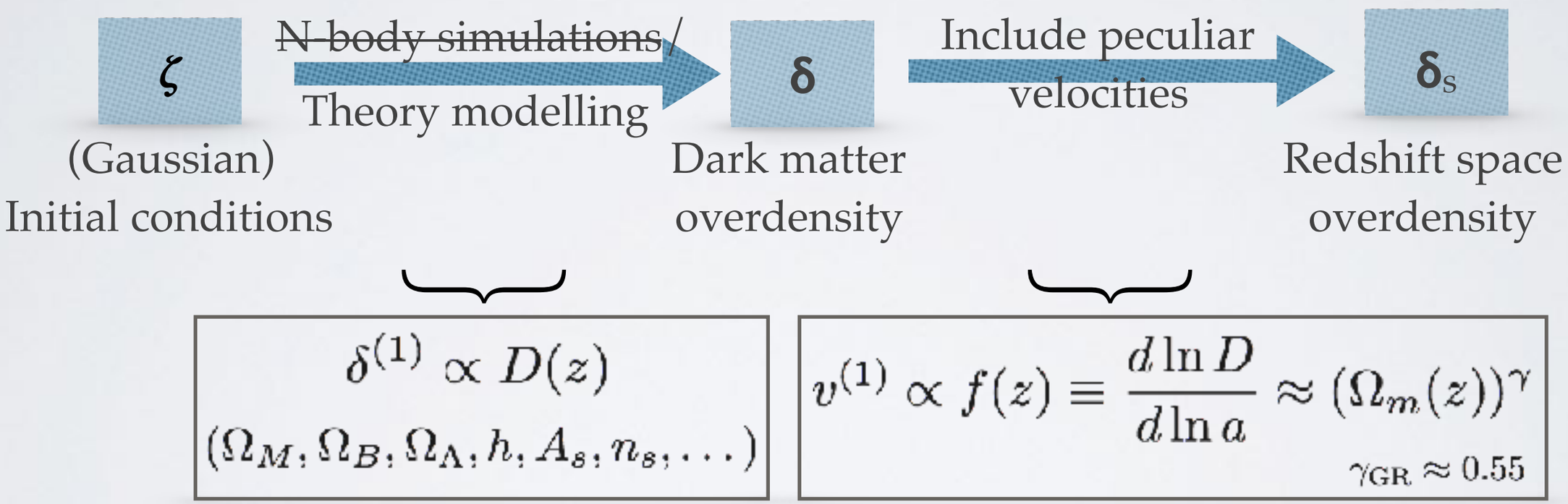
CMB FLUCTUATIONS



STARS & GALAXIES - LSS



(a) Redshift space distortions in the EFTofLSS



Standard Perturbation Theory

Effective Field Theory

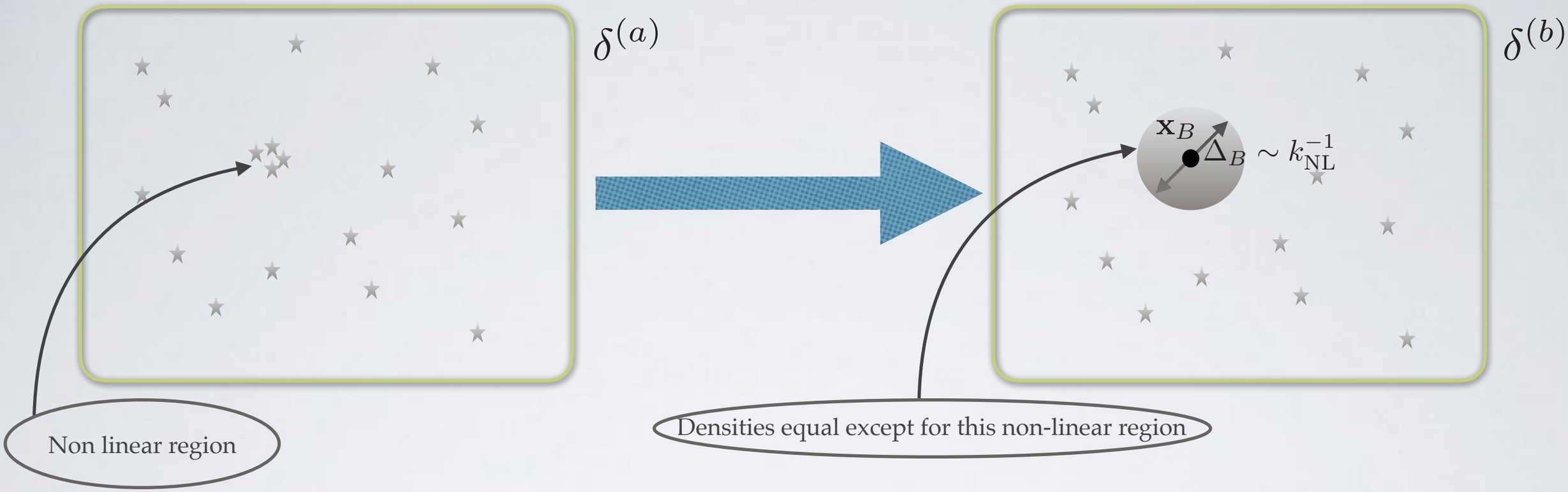
Super-horizon	Large scales SPT	Mildly non-linear EFT	UV regime

$k_{\max}$

- SPT breaks down in mild non-linear regime
- EFT tells us that by including all terms consistent with symmetry we can *parametrise our ignorance* of UV regime
- To leading order these terms go as $\partial^2\delta$

$$\mathrm{FT}[\delta^2](\mathbf{k}) = \int d^3\mathbf{q}\,\delta(\mathbf{k}-\mathbf{q})\delta(\mathbf{q}) + Z_{c.t.}k^2\delta(\mathbf{k})$$

Why contributions start at k²? A more physical picture… arXiv:1307.3220 (Mercolli & Pajer)



Conservation of mass and momentum:

$$\int d^3\mathbf{x}\delta^{(a)}(\mathbf{x}) = \int d^3\mathbf{x}\delta^{(b)}(\mathbf{x}) \quad \& \quad \int d^3\mathbf{x}\,\mathbf{x}\delta^{(a)}(\mathbf{x}) = \int d^3\mathbf{x}\,\mathbf{x}\delta^{(b)}(\mathbf{x})$$

$$\Longrightarrow \delta^{(a)}(\mathbf{k}) - \delta^{(b)}(\mathbf{k}) = \int d^3\mathbf{x} e^{-i\mathbf{k}.\mathbf{x}}[\delta^{(a)}(\mathbf{x}) - \delta^{(b)}(\mathbf{x})] = \int d^3\mathbf{y} e^{-i\mathbf{k}.(\mathbf{y}-\mathbf{x}_B)}[\delta^{(a)}(\mathbf{y}-\mathbf{x}_B) - \delta^{(b)}(\mathbf{y}-\mathbf{x}_B)]$$

$$\approx e^{i\mathbf{k}.\mathbf{x}_B}\int_0^{\Delta_B} d^3\mathbf{y}[\cancel{1} - \cancel{i\mathbf{k}.\mathbf{y}} + |\mathbf{k}.\mathbf{y}|^2 + \dots][\delta^{(a)}(\mathbf{y}-\mathbf{x}_B) - \delta^{(b)}(\mathbf{y}-\mathbf{x}_B)] \sim e^{i\mathbf{k}.\mathbf{x}_B}\mathcal{O}\left(\frac{k^2}{k_{\rm NL}^2}\right)$$

mass conservation momentum conservation

$$\delta^{Ren} = \delta - c_s^2\frac{k^2}{k_{\rm NL}^2}\delta$$

Calculating the Dark Matter Power Spectrum

Solve equations for perturbative solution (to 3rd order for 1 loop calc) + add counter term

$$\delta = \delta^{(1)} + \delta^{(2)} + \delta^{(3)} + \delta^{c.t.} \qquad (\delta^{(1)} = D(z)\delta_\star\,,\, \delta^{c.t.} = -2\pi c_s^2(z)k^2\delta_\star)$$

Philosophy: Correct the SPT calculation for back-reactions from UV regime

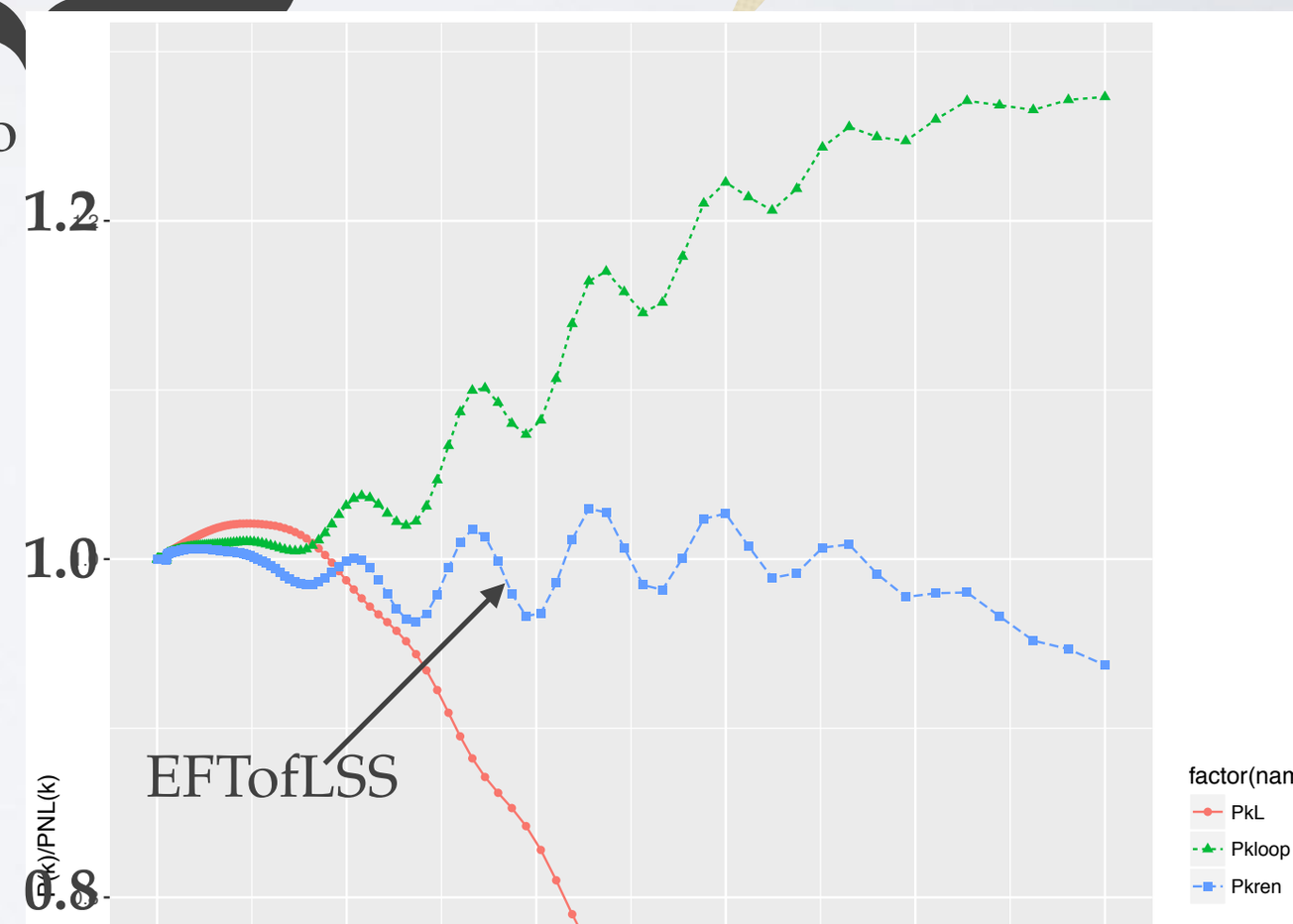
Gaussian ICs

$$\langle\delta\delta\rangle = \langle\delta^{(1)}\delta^{(1)}\rangle + 2\langle\delta^{(1)}\delta^{(3)}\rangle + \langle\delta^{(2)}\delta^{(2)}\rangle + 2\langle\delta^{(1)}\delta^{c.t.}\rangle$$

$$P^{\rm DM}(k) = P_{11}(k) + P_{13}(k) + P_{22}(k) - 2(2\pi)D(z)c_s^2(z)k^2P_\star(k)$$

SPT (to

Usual perspective: Fit c_s^2 for each redshift from simulations



What's causing the oscillations?

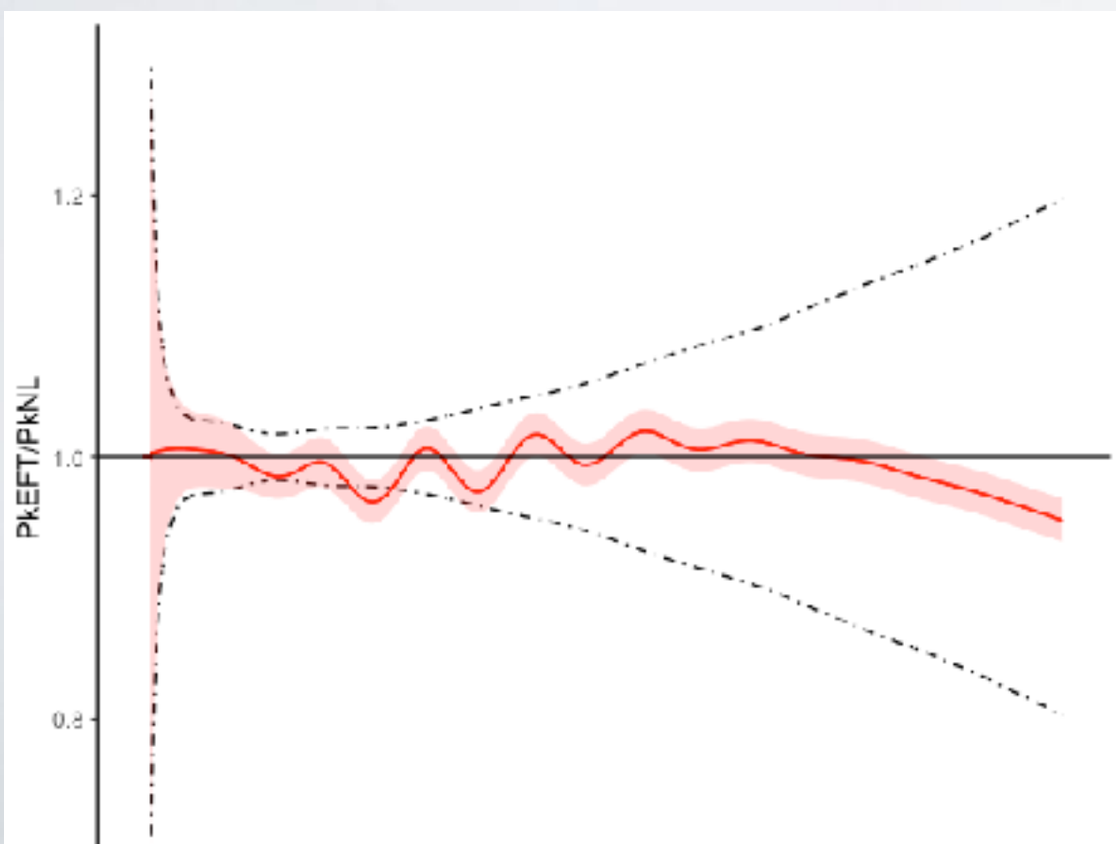
Small scale wiggles e.g. due to Baryon Acoustic Oscillations (BAOs) are amplified by growing displacement fields.

SPT treats densities and displacements equally. *This is a problem (displacements non-linear)!*

SAVING GRACE!

Lagrangian perturbation theory (*where we comove with the fluid elements*) treats displacement fields exactly => no such problem!

Expand LPT in terms of SPT/EFT



LPT tells us how to dampen the wiggle part (no difference to the no-wiggle part)

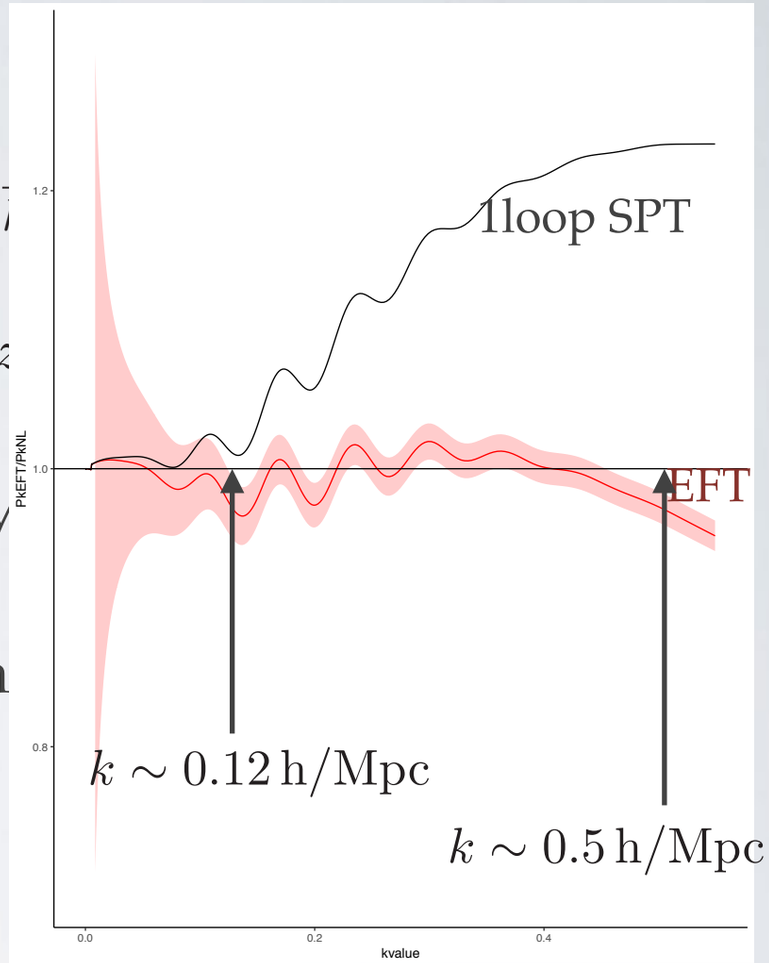
Devil in the details!

$$P^{\mathrm{IR}}(k) = P_{\mathrm{nw}}(k) + e^{-k^2\Sigma^2}\Big(\Delta P_{\mathrm{1loop,w}}(k) + k^2\Sigma^2\Delta P_{L,\mathrm{w}}($$

$$\Sigma^2 \approx \int \frac{dq}{6\pi^2} P_L(q)$$

Summary of DM Calculation

- EFT adjusts broadband power due to UV physics
- Symmetries tell us how to renormalise $\delta^{\mathrm{Ren}} = \delta + c_s^2(z)\frac{k^2}{k_{NL}^2}\delta + c_2(z)$
- Find time dependence using loop integrals e.g. P_{13} tells us $c_s^2(z$
- Fit for c_s^2 using the non-linear power spectrum (e.g. from CAM
- Residual oscillations largely eliminated using IR-resummation



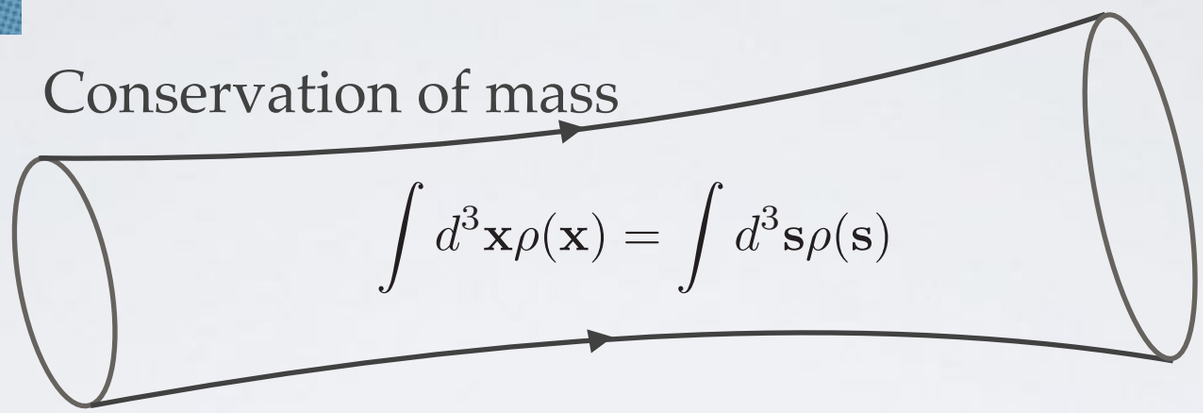
Large increase in scales

Increase in available modes by factor $\mathcal{O}(0.5/0.12)^3 = \mathcal{O}(50)$

Redshift Space Distortions

Real Space $\{x^i\}$ $\xrightarrow{\quad s^i = x^i + \hat{\mathbf{n}}.\mathbf{v}\hat{n}^i \quad}$ $\{s^i\}$ Redshift Space

Conservation of mass



$$\int d^3\mathbf{x}\rho(\mathbf{x}) = \int d^3\mathbf{s}\rho(\mathbf{s})$$

$$\delta(\mathbf{k}) \longrightarrow \delta_s(\mathbf{k}) \approx \delta(\mathbf{k}) - ik_z v_z(\mathbf{k}) + \frac{i^2}{2}k_z^2[v_z^2]_{\mathbf{k}} - \frac{i^3}{6}k_z^3[v_z^3]_{\mathbf{k}} - ik_z[v_z\delta]_{\mathbf{k}} + \frac{i^2}{2}k_z^2[v_z^2\delta]_{\mathbf{k}}$$

Convolution => UV contributions

$$\mathbf{k}.\hat{\mathbf{n}} \equiv k_z \equiv k\mu$$

Equations of motion $\Rightarrow$ $v_z(\mathbf{k})^{(1)} = -\frac{i\mathbf{k}.\hat{\mathbf{n}}}{k^2} f\delta^{(1)}$

growth rate

Steps same as before ...
(won't show them in detail here!!)

Follow same steps as for the DM power spectrum:

Calculate the power spectrum (**SPT**)

$$P = P_{s(11)} + P_{s(13)} + P_{s(22)}$$

Lot's of tricks to calculate these (see paper!!)

$$P_{s(11)}(k) = P_{11}(k)(1 + f\mu^2)^2$$

$$P_{s(22)}(k) = \sum_{\ell=0}^{4} P^{(\ell)}_{s(22)}(k)\mu^{2\ell}$$

$$P_{s(13)}(k) = \sum_{\ell=0}^{3} P^{(\ell)}_{s(13)}(k)\mu^{2\ell}$$

Only up to μ^6

Form of the counterterms (**EPT**)

$$\delta_s(\mathbf{k}) \to \delta_s^{\mathrm{Ren}}(\mathbf{k}) = \delta_s(\mathbf{k}) - \frac{k^2}{k_{\mathrm{NL}}^2}\delta(\mathbf{k})(c_0(z) + \mu^2 c_1(z) + \mu^4 c_2(z) + \mu^6 c_3(z))$$

4 counterterms

What about those oscillations?

Non-linear velocities (Fingers-of-God) will wash away small scale structures

CONCLUSIONS (a)

- EFTofLSS is *best* viewed as a way to renormalise the theory
- Can predict the time dependence of the counterterms (if virial scales decouple)
- IR resummation needed to correct for (too) large displacement fields
- Allows for much larger range of scales accessible by theory

- **Redshift Space Distortions**: Calculation follows the same principles
- Just 4 counterterms to fit (doing so right now!!)
- More controlled method to account for non-linear velocities (FoG)
- Lot's of nifty tricks to make our terms analytic in μ

- **Future**: Incorporate Modified Gravity into the formalism

Aside: Efficient method for 3D integrals

$$\int \frac{d^3\mathbf{q}}{(2\pi)^3} \frac{\mathcal{P}_\ell(\hat{\mathbf{k}}.\hat{\mathbf{q}})}{|\mathbf{k}-\mathbf{q}|^2} P(|\mathbf{q}|)$$

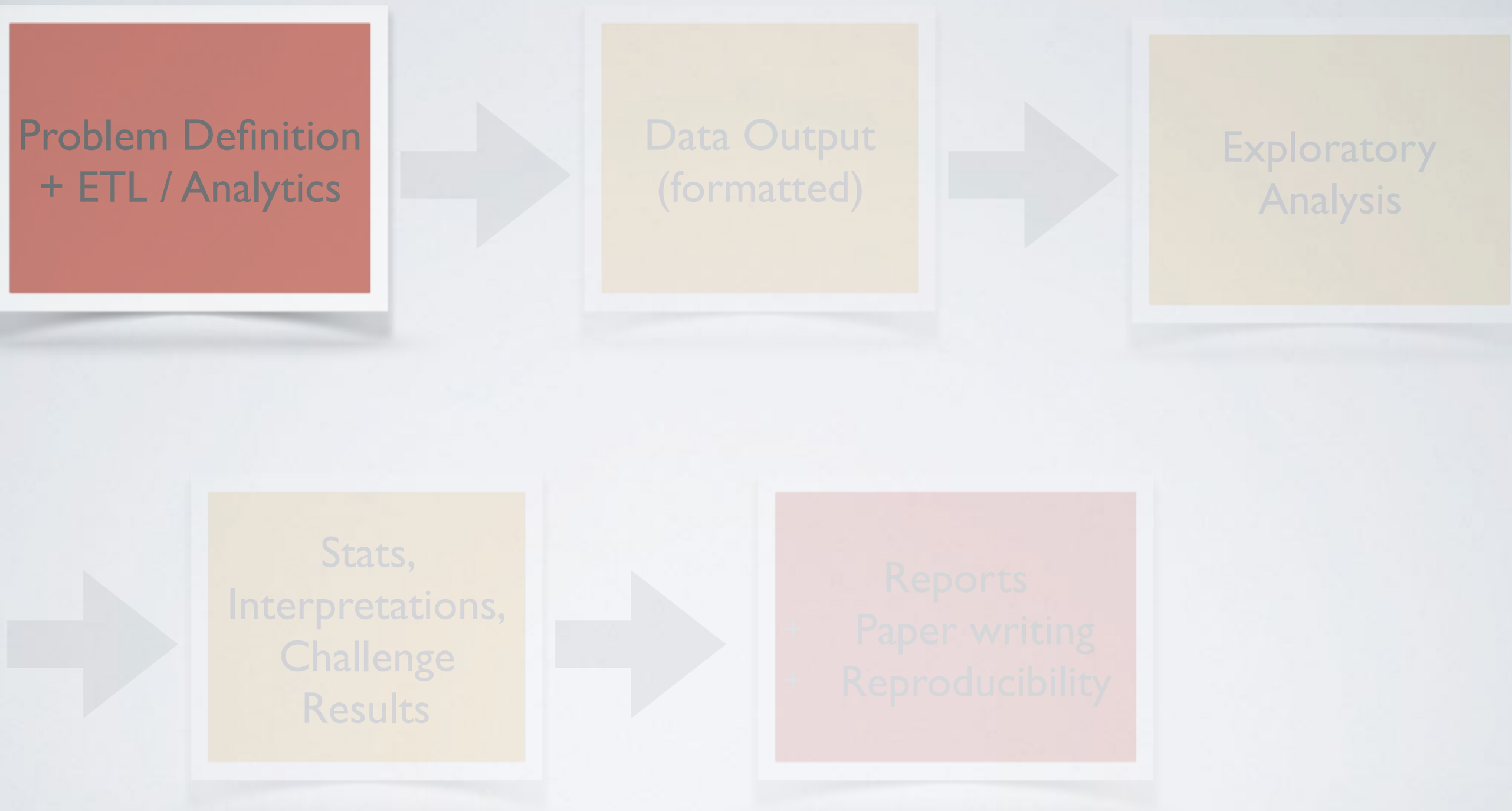
$$\int \frac{d^3\mathbf{s}}{(2\pi)^3} \int \frac{d^3\mathbf{q}}{(2\pi)^3} (2\pi)^3 \delta_D(\mathbf{k}-\mathbf{q}-\mathbf{s}) \frac{\mathcal{P}_\ell(\hat{\mathbf{k}}.\hat{\mathbf{q}})}{s^2} P(|\mathbf{q}|)$$

$$\int d^3\mathbf{x} e^{i\mathbf{k}.\mathbf{x}} \int \frac{d^3\mathbf{s}}{(2\pi)^3} e^{i\mathbf{s}.\mathbf{x}} \int \frac{d^3\mathbf{q}}{(2\pi)^3} e^{i\mathbf{q}.\mathbf{x}} \frac{\mathcal{P}_\ell(\hat{\mathbf{k}}.\hat{\mathbf{q}})}{s^2} P(|\mathbf{q}|)$$

$$\Rightarrow \int \frac{d^3\mathbf{q}}{(2\pi)^3} \frac{\mathcal{P}_\ell(\hat{\mathbf{k}}.\hat{\mathbf{q}})}{|\mathbf{k}-\mathbf{q}|^2} P(|\mathbf{q}|) \propto \int \frac{dq}{2\pi^2} q^2 P(q) \int dx x j_\ell(kx) j_\ell(qx)$$

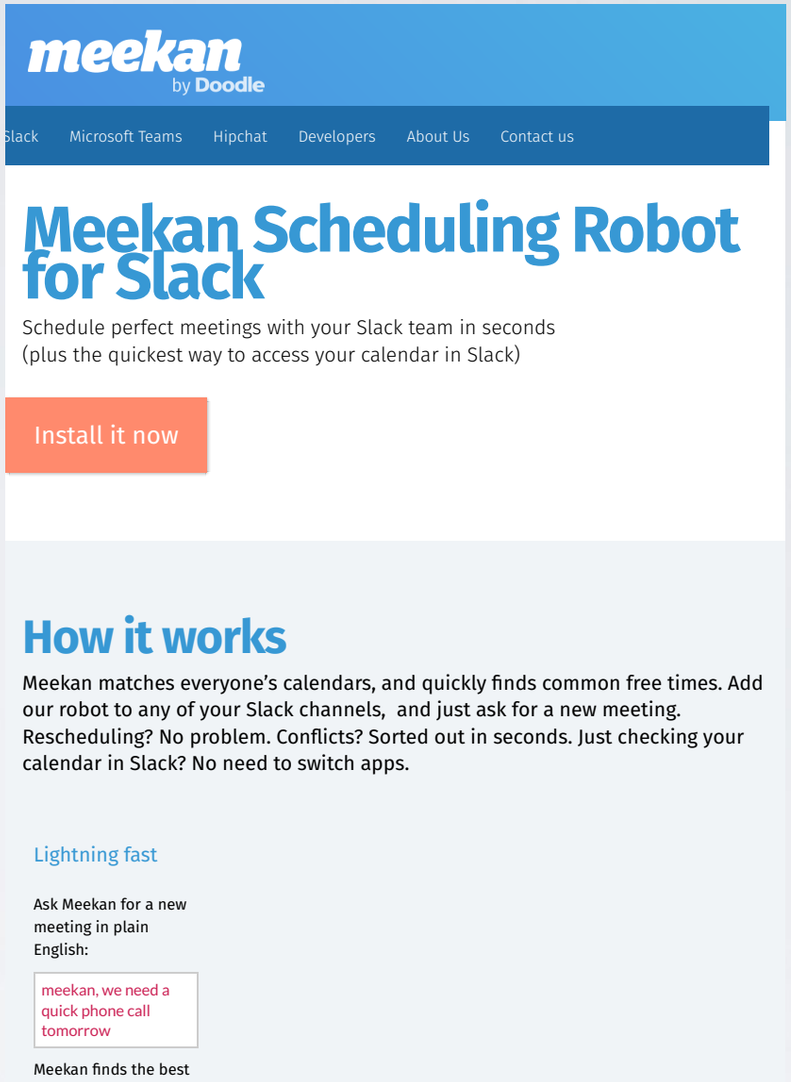
$$(\text{e.g. for } \ell = 0) \int \frac{d^3\mathbf{q}}{(2\pi)^3} \frac{1}{|\mathbf{k}-\mathbf{q}|^2} P(|\mathbf{q}|) \propto \int \frac{dq}{2\pi^2} q^2 P(q) \frac{\ln\left(\frac{k+q}{|k-q|}\right)}{2kq}$$

How did we get here?

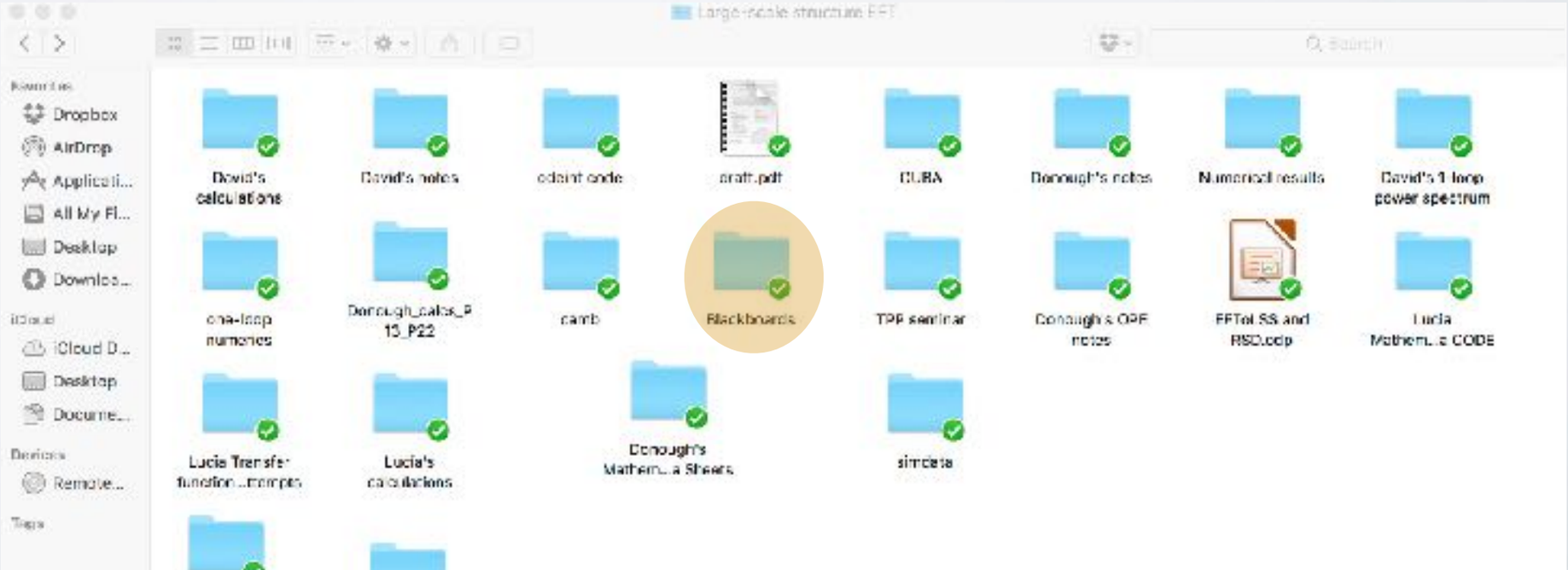

Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

Organising the team

Slack & meekan

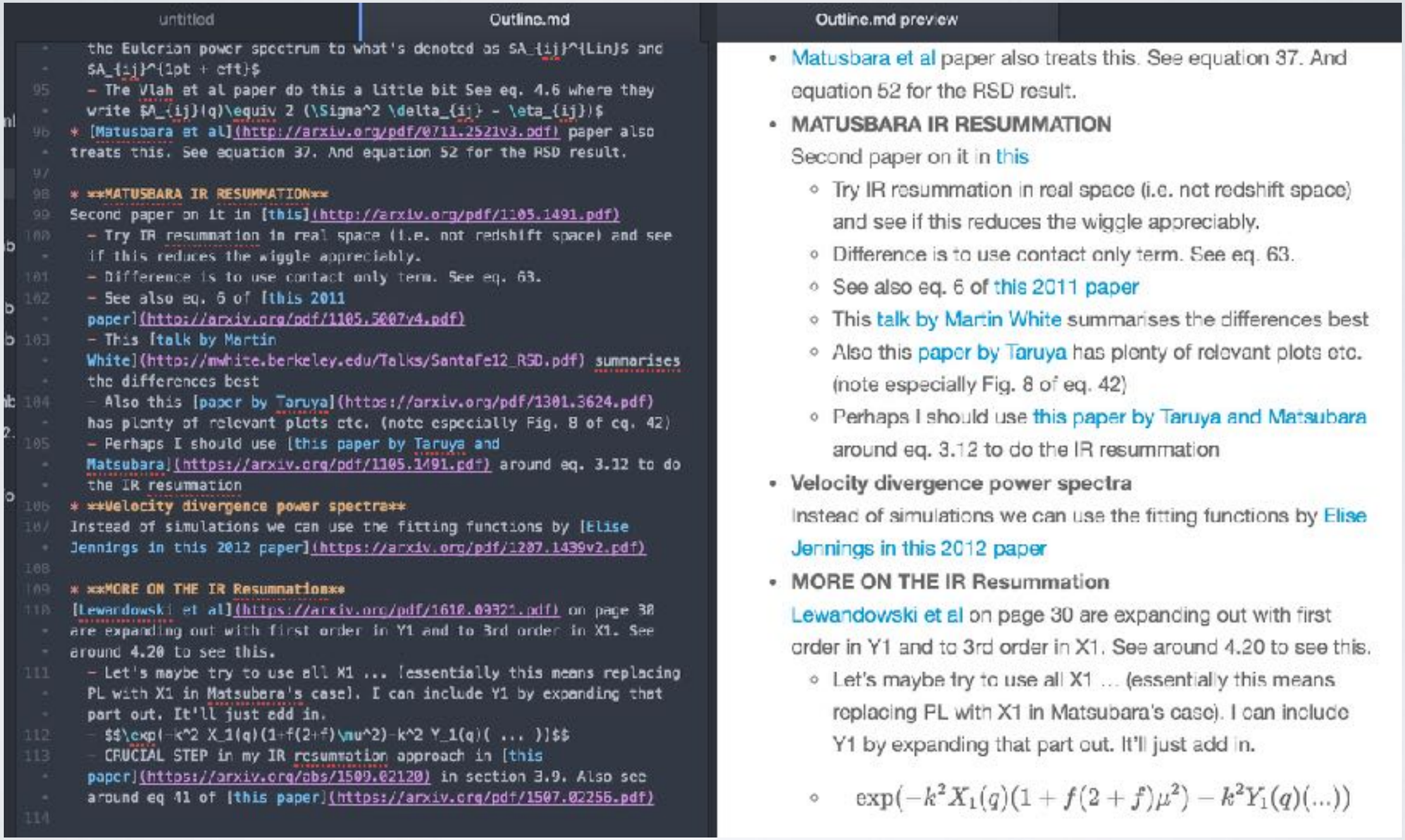


Sharing our findings



Summarising our meetings + quick notes

Markdown + latex (via Atom)



Preliminary Analytics

Pen & Paper + Mathematica

RSD_checkex3

RSD Checks Preliminaries (N.B. this sheet uses results from FabrikantImplemented.nb)

P_{22} terms (all these are operators to be pre-multiplied by $\frac{1}{\pi^3}\int dr\, P_L(r) \int dq\, P_L(q)$)

Kaiser × Kaiser

```
coeffs[n_] := Integrate[(A1 * (1 + muqr/2 * (q/r + r/q)) + B1 * (1 + muqr * (q/r + r/q) + muqr^2))^2 * LegendreP[n, muqr] * (2 * n + 1) / 2, {muqr, -1, 1}]
```

```
{coeffs[0], coeffs[1], coeffs[2], coeffs[3], coeffs[4]}
```

$$\left\{\frac{7A1^2}{6}+\frac{10\,A1\,B1}{3}+\frac{38\,B1^2}{15}+\frac{A1^2q^2}{12r^2}+\frac{A1\,B1\,q^2}{3r^2}+\frac{B1^2q^2}{3r^2}+\frac{A1^2r^2}{12q^2}+\frac{A1\,B1\,r^2}{3q^2}+\frac{B1^2r^2}{3q^2},\ \frac{(A1+2\,B1)\,(5\,A1+8\,B1)\,(q^2+r^2)}{5\,q\,r},\right.$$
$$\left.\frac{A1^2}{3}+\frac{8\,A1\,B1}{3}+\frac{68\,B1^2}{21}+\frac{A1^2q^2}{6r^2}+\frac{2\,A1\,B1\,q^2}{3r^2}+\frac{2\,B1^2q^2}{3r^2}+\frac{A1^2r^2}{6q^2}+\frac{2\,A1\,B1\,r^2}{3q^2}+\frac{2\,B1^2r^2}{3q^2},\ \frac{2\,B1\,(A1+2\,B1)\,(q^2+r^2)}{5\,q\,r},\ \frac{8\,B1^2}{35}\right\}$$

```
P22KK :=
  2 * r^2 * q^2 * ((7/6 * Amu^2 + 10/3 * Amu * Bmu + 38/15 * Bmu^2) * I000 + (1/12 * Amu^2 + 1/3 * Amu * Bmu + 1/3 * Bmu^2) * (q^2/r^2 + r^2/q^2) * I000 -
    1/5 * (Amu + 2 * Bmu) * (5 * Amu + 8 * Bmu) * (q/r + r/q) * I011 + (1/3 * Amu^2 + 8/3 * Amu * Bmu + 68/21 * Bmu^2) * I022 +
    (1/6 * Amu^2 + 2/3 * Amu * Bmu + 2/3 * Bmu^2) * (q^2/r^2 + r^2/q^2) * I022 - 2 * Bmu/5 * (Amu + 2 * Bmu) * (q/r + r/q) * I033 + 8 * Bmu^2/35 * I044)
```

[vδ] × [vδ]

[v^2] × [v^2]

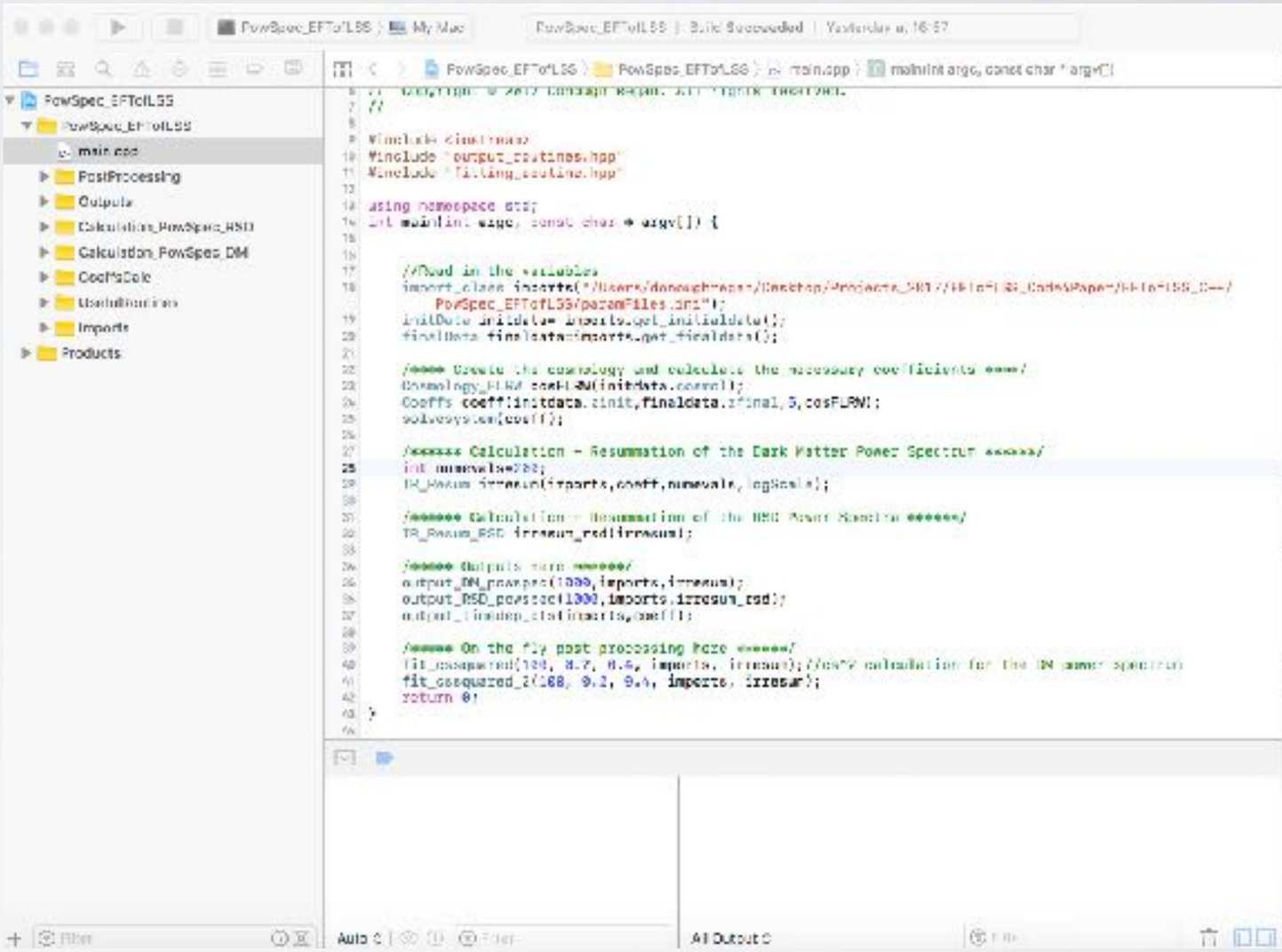
Kaiser × [vδ] times 2

Kaiser × [v^2] times 2

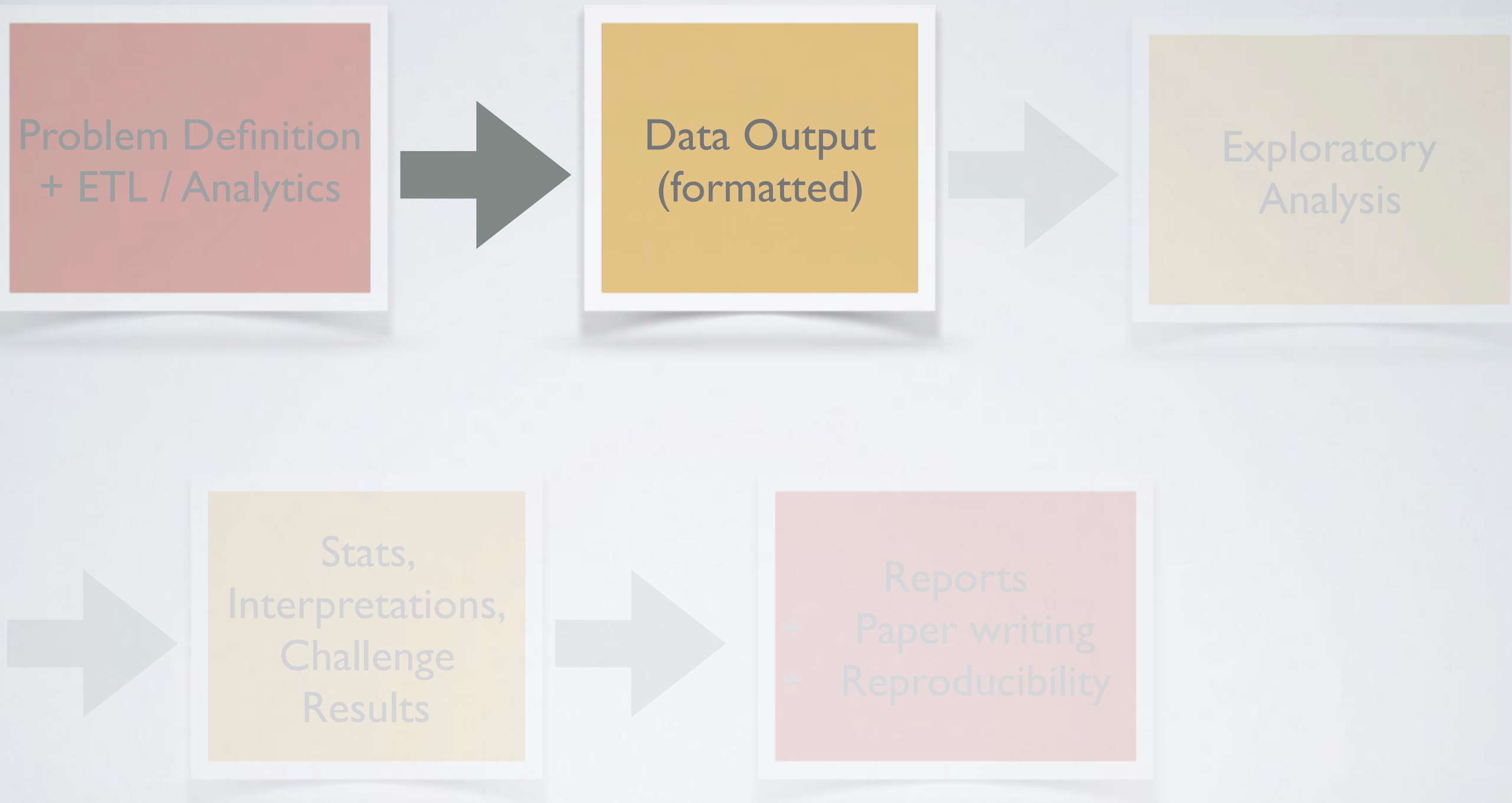
Producing the data (i.e. the power spectra)

3 pipeline codes: 2 in C++, 1 Mathematica

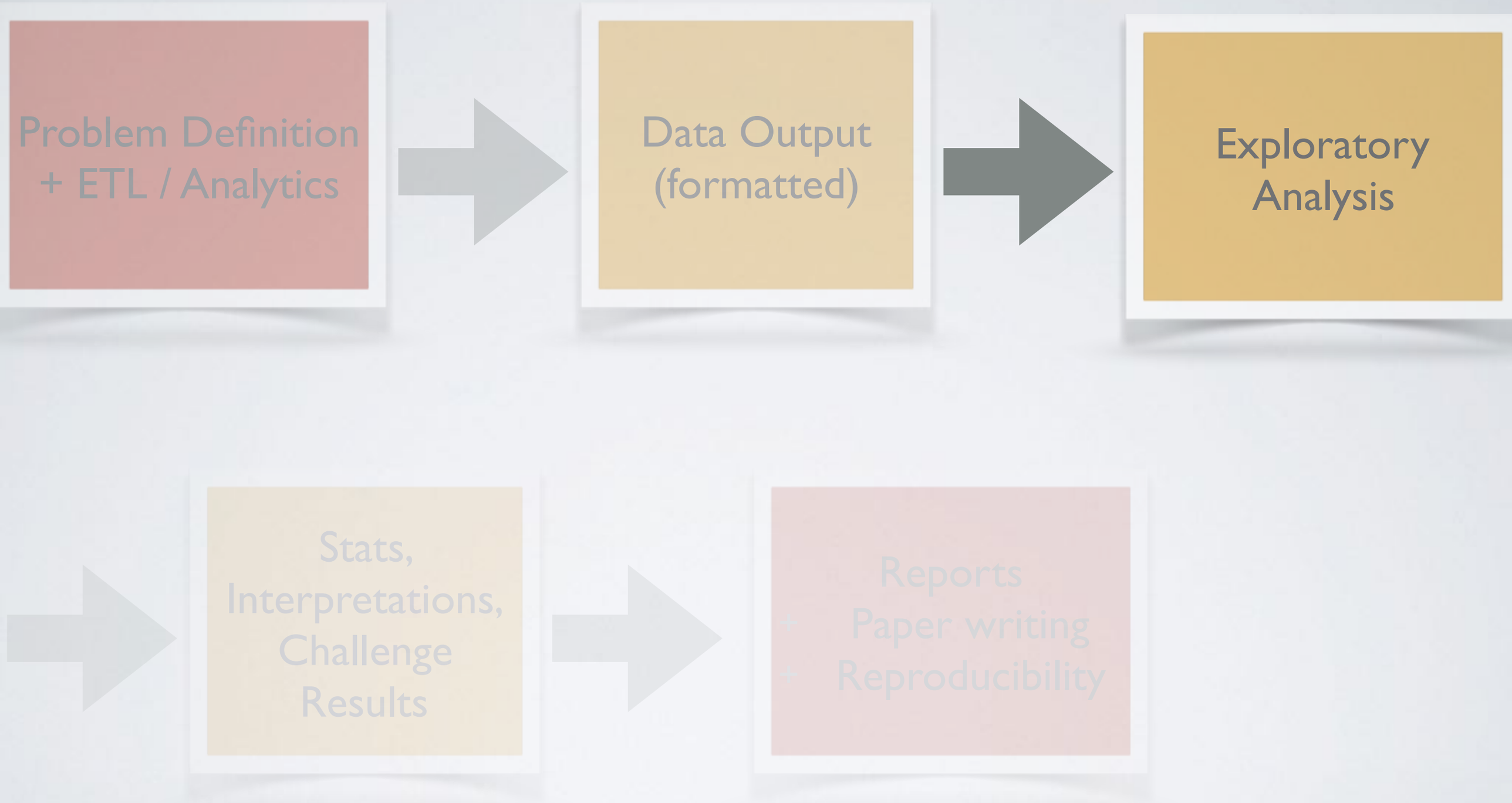
- Use IDEs (e.g. XCode,CLion) to code!



- Use version control e.g. GitHub


Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

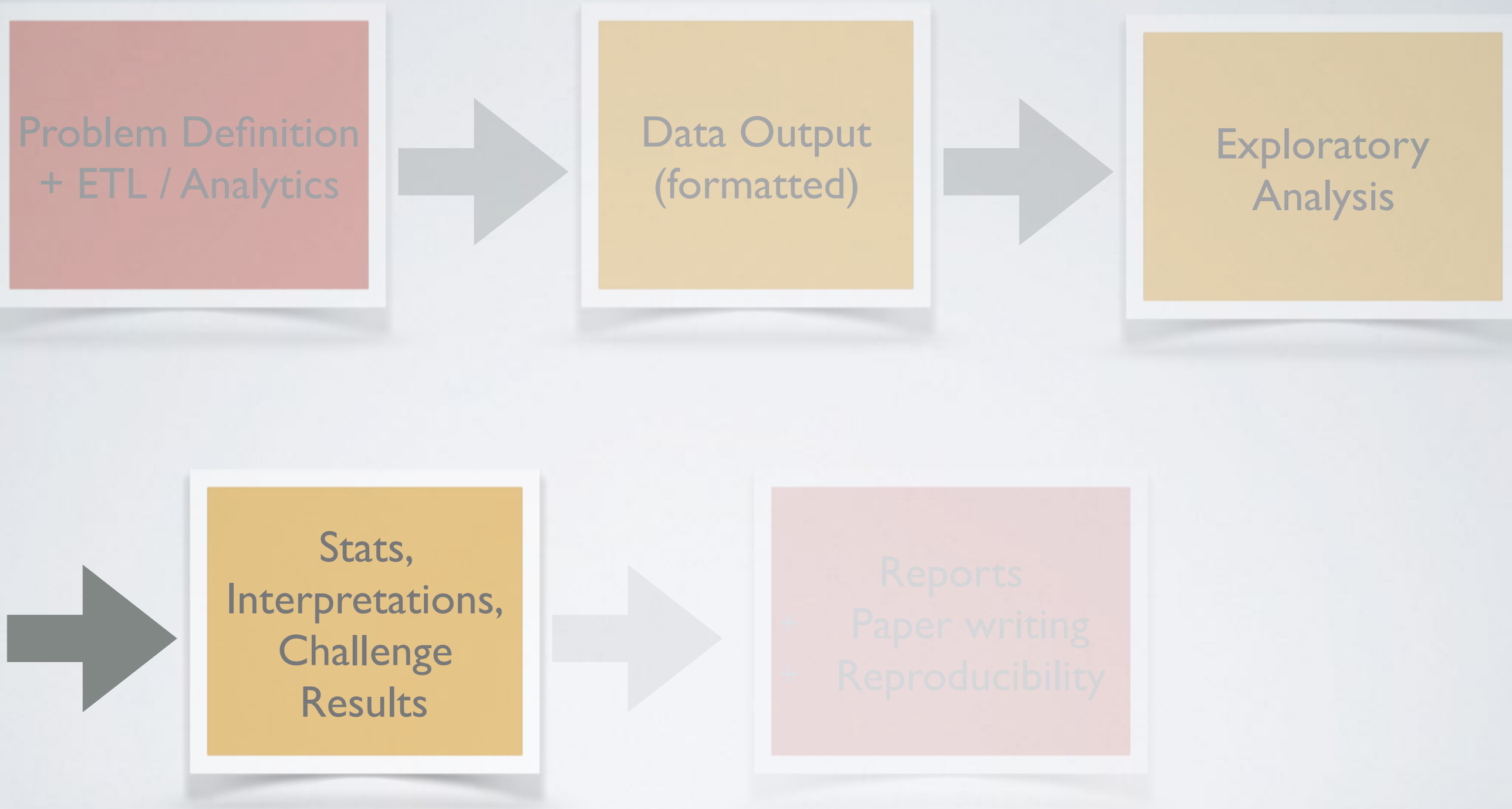
- Label files clearly
- Output to .csv files
- Instructive column naming (easier for analysis pipelines)


Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

$$P^{\mathrm{DM}}(k) = P_{1\,\mathrm{loop}}(k) - c_s^2 k^2 P_L(k)$$

Plot $$-\frac{P^{\mathrm{DM}}(k) - (P_{1\,\mathrm{loop}}(k))}{k^2 P_L(k)}$$

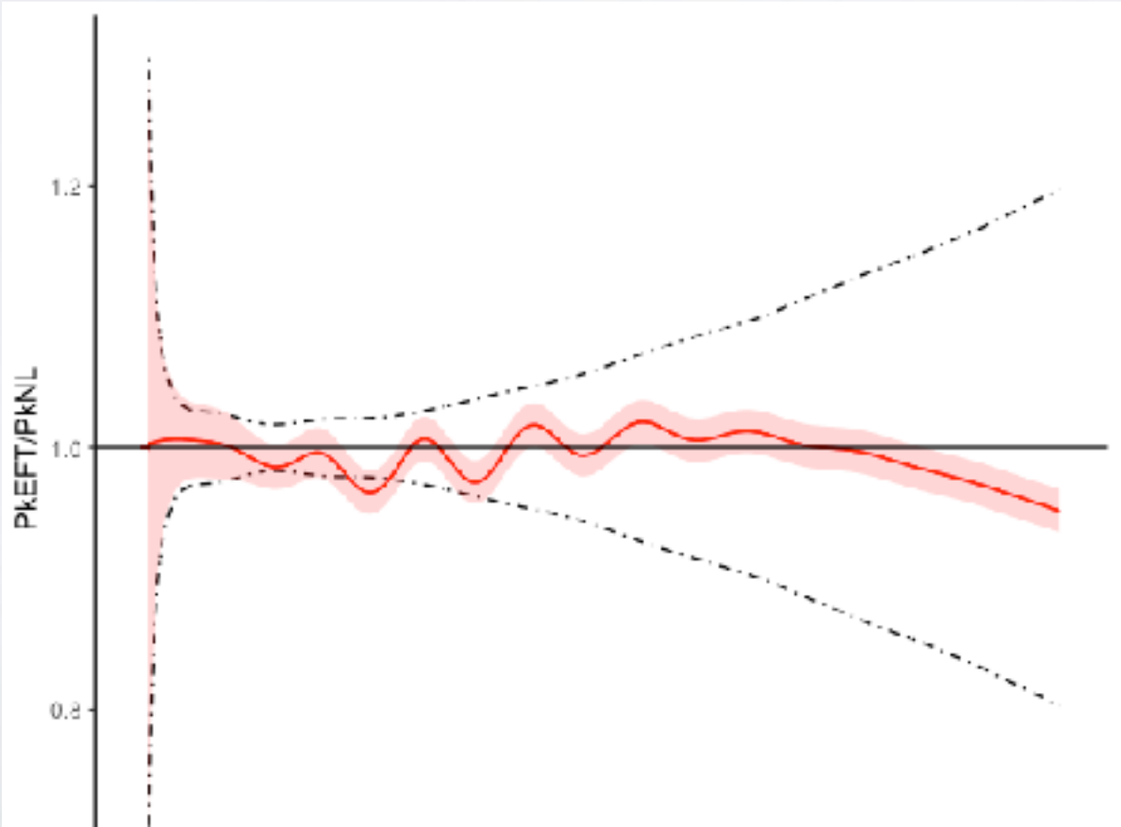
$c_s^2 = 3.4$

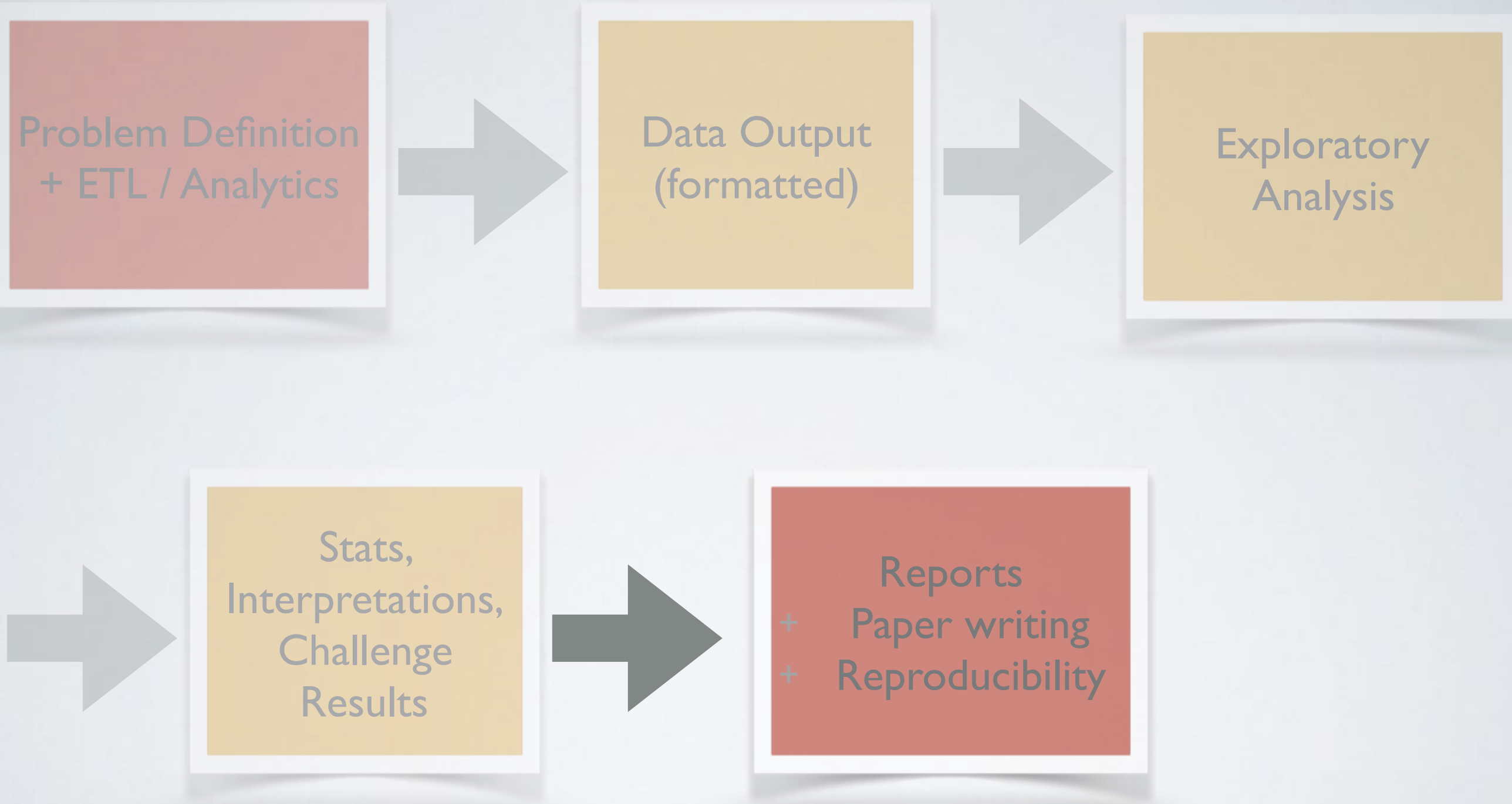

Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

Chi-squared fitting

```
> library(stats4)#for mle
> library(bbmle)#for mle2
> PkDMnew$PkNL<-pkstitted_real_cons:$Fkvals
> PkDMnew$PkDiff<-PkDMnew$PkNL-PkDMnew$Pk1.ocp
> LL<-function(csq,sigma1){
+     R=dnorm(PkDMnew$PkDiff,csq*(PkDMnew$kvalue^2)*PkDMnew$PkL,sigma1*sqrt(PkDMnew$PkDiff^2))
+     -sum(log(R))
+ }
> fit<-suppressWarnings(mle2(LL,start=list(csq=0,sigma1=1),fixed=list(sigma1=1)))#fit the sigma value... and so mu and it's error are the mle estimates for cs^2
> summary(fit)
Maximum likelihood estimation

Call:
mle2(minuslogl = LL, start = list(csq = 0, sigma1 = 1), fixed = list(sigma1 = 1))

Coefficients:
    Estimate Std. Error z value     Pr(z)
csq -3.45685    0.40105 -8.6194 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-2 log L: 979.959
```

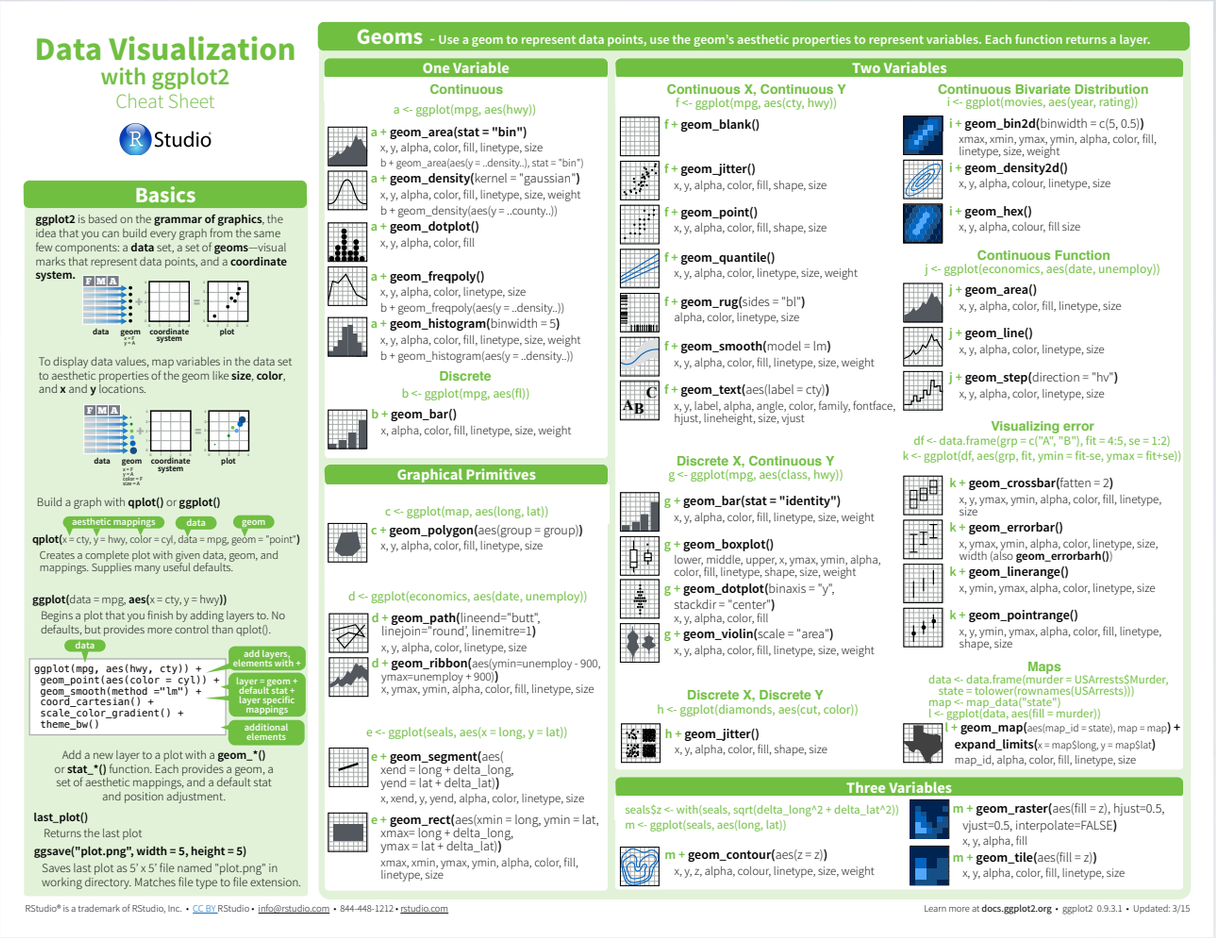
Theory errors



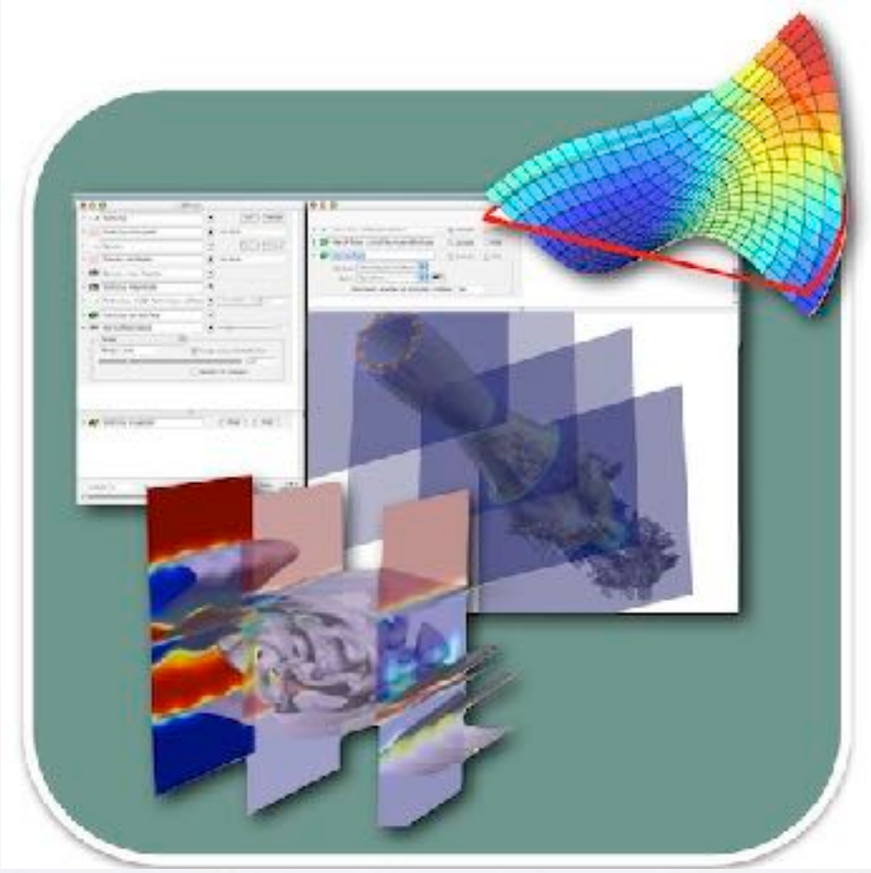

Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

Plotting

R & ggplot2



DataTank



Paper Writing

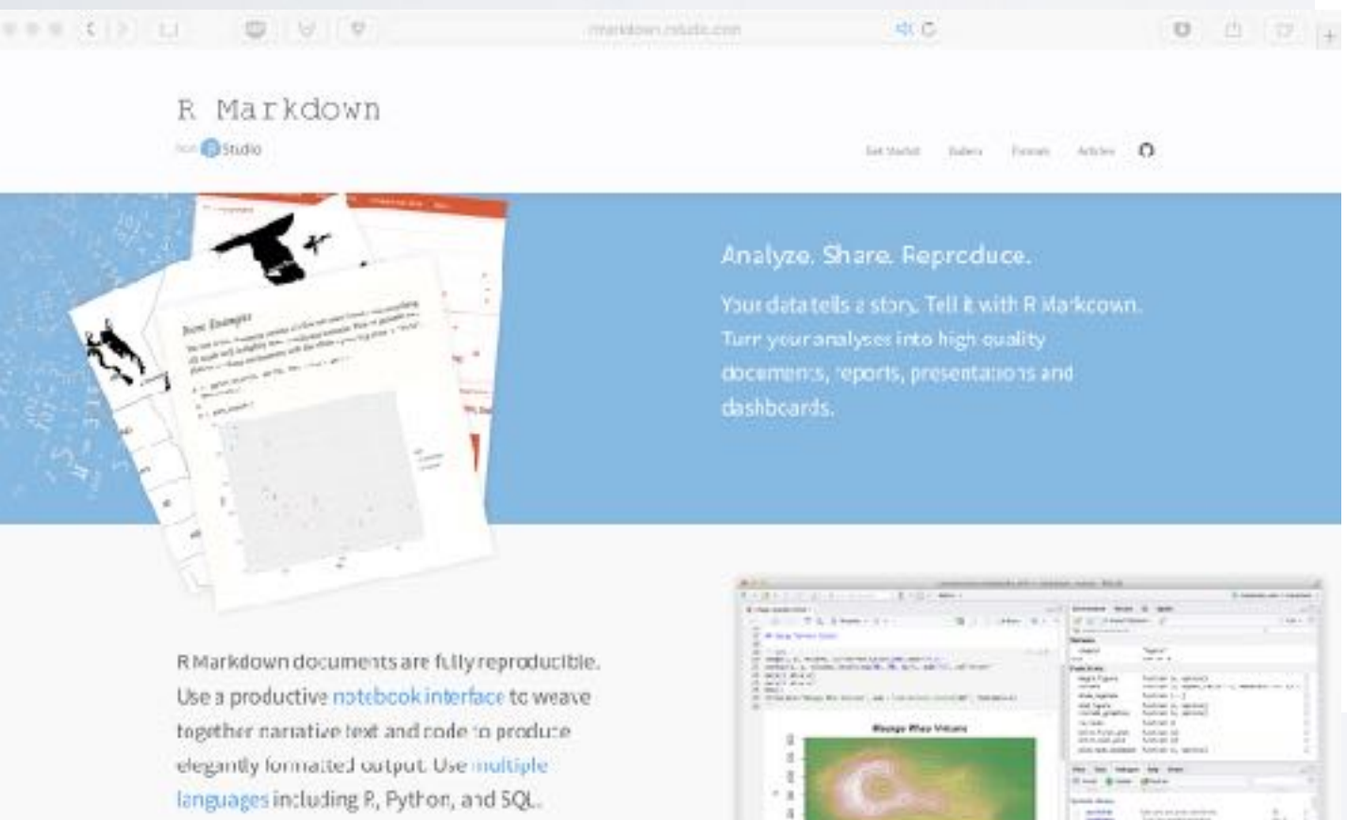
LaTeX + Dropbox

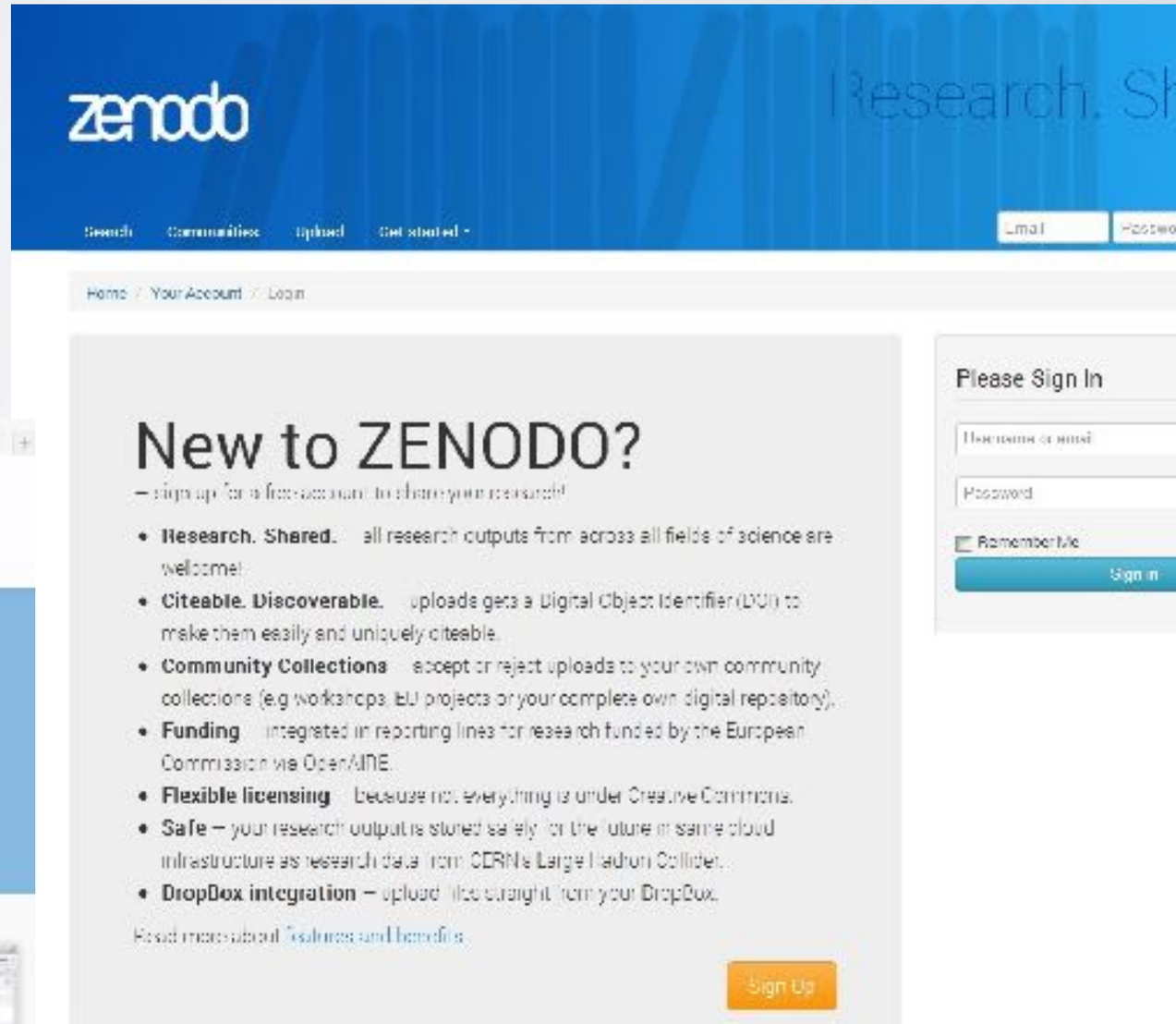
Reproducible Research

Code will be released on GitHub

+ via zenodo.org

+ via RMarkdown





Advantages vs Disadvantages

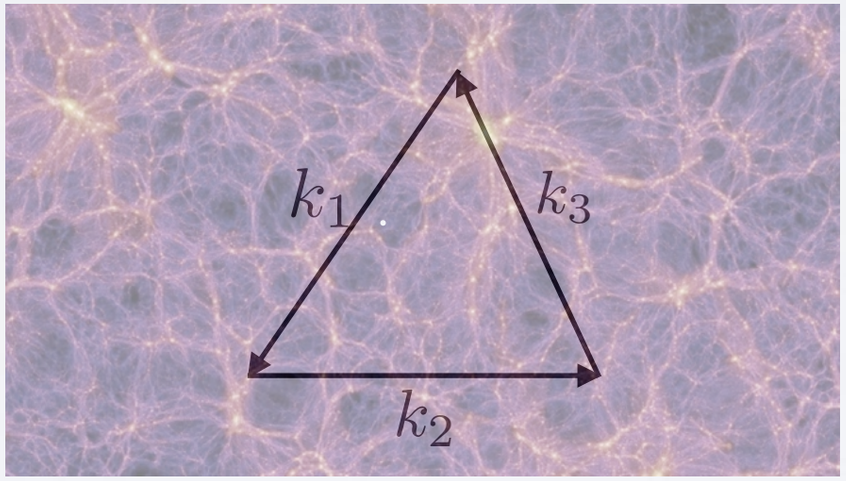
- More professional approach
- Better delineation of tasks
- Less chance of "reinventing the wheel"
- Reproducibility better adheres to the scientific process

BUT…

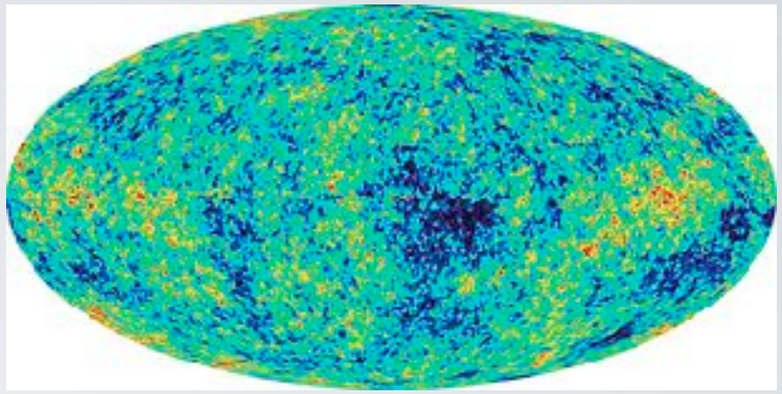
- Takes (slightly) more time (possibly!)

(b) 3 point function to improve cosmological constraints

$$\langle \delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\delta(\mathbf{k}_3)\rangle = (2\pi)^3 \delta_D(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) B_\delta(k_1, k_2, k_3)$$



Aside: On the CMB



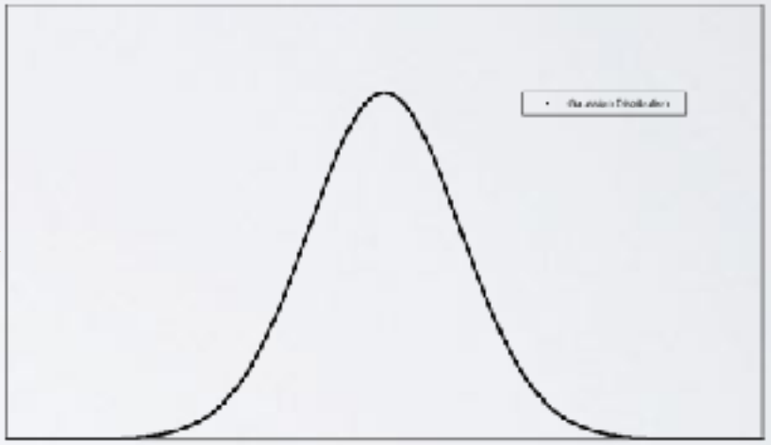
$\zeta(\mathbf{k})$ ⟷ $\Delta_l(k)$ ⟷ $\frac{\Delta T}{T}(\hat{\mathbf{n}})$

Initial curvature perturbation

Temperature fluctuations

Primordial to CMB via transfer function

Linear relation →

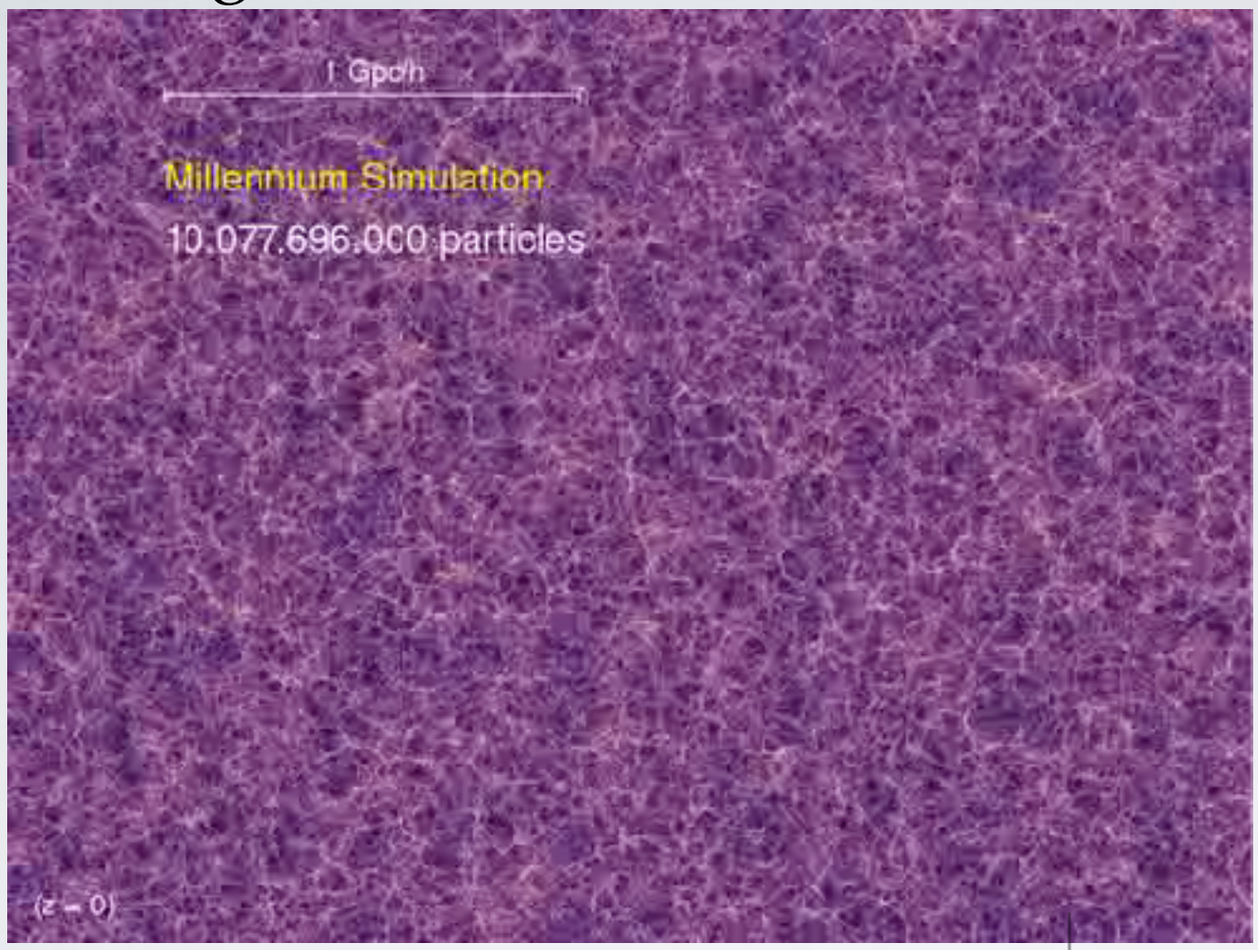


3pf of CMB ≈ 0 if
3pf of primordial ≈ 0

→

- Superb probe for primordial non-Gaussianity
- 2pf (Power Spectrum) of CMB contains all cosmological information

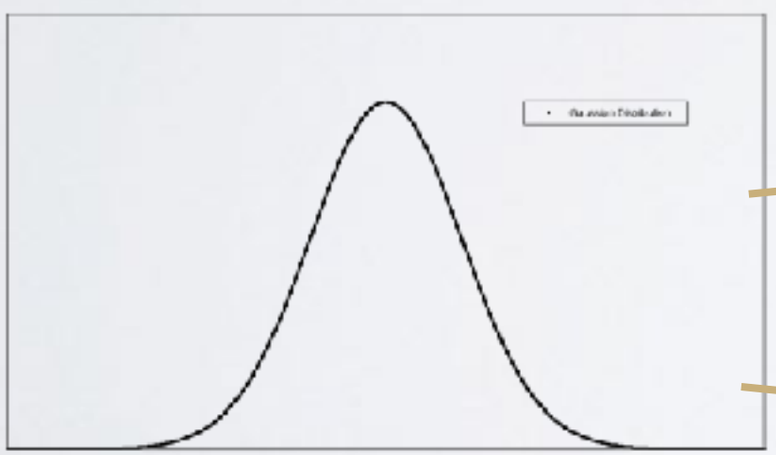
Large Scale Structure Evolution



Video: http://www.mpa-garching.mpg.de/galform/virgo/millennium

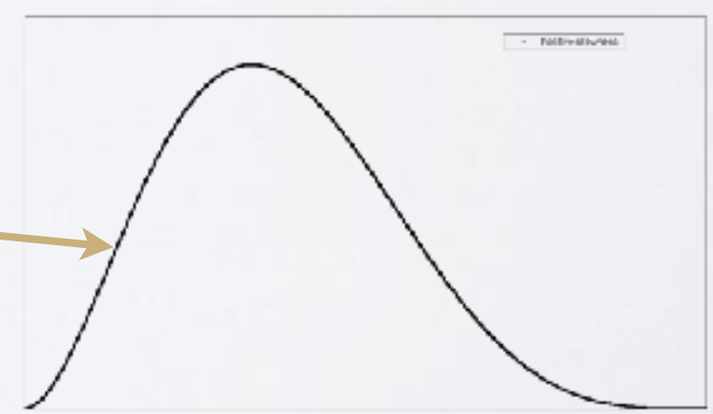
$$\dot{\delta} + \nabla.((1+\delta)\mathbf{v}) = 0$$

Non-Gaussianity



Gaussian distribution

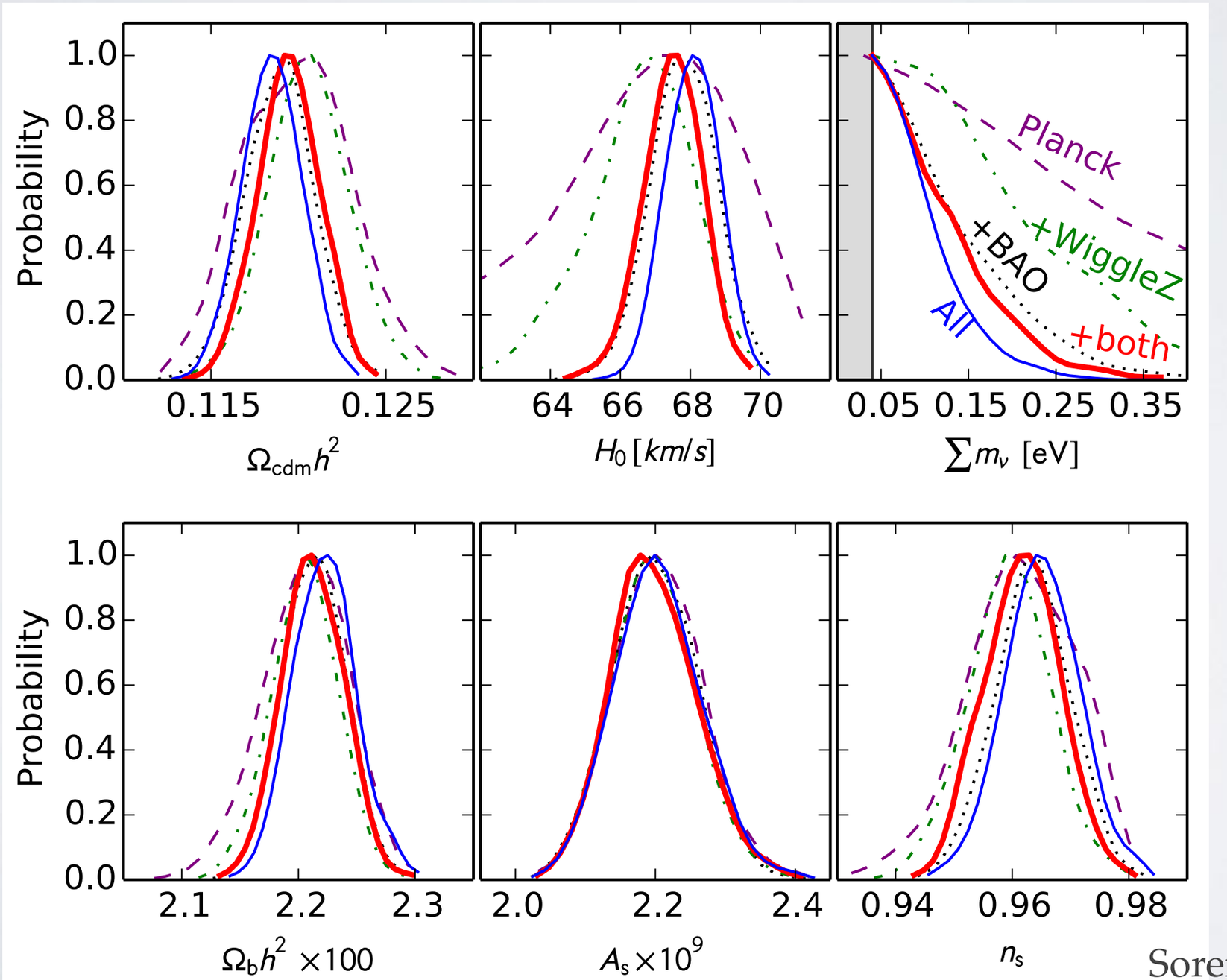
Negatively skewed



Positively skewed

Improving cosmological constraints with LSS

- CMB is best probe currently… but difficult to improve beyond Planck
- LSS is 3D data set (z dependence)… but nonlinear structure
- LSS breaks some degeneracies e.g. between matter density and expansion rate

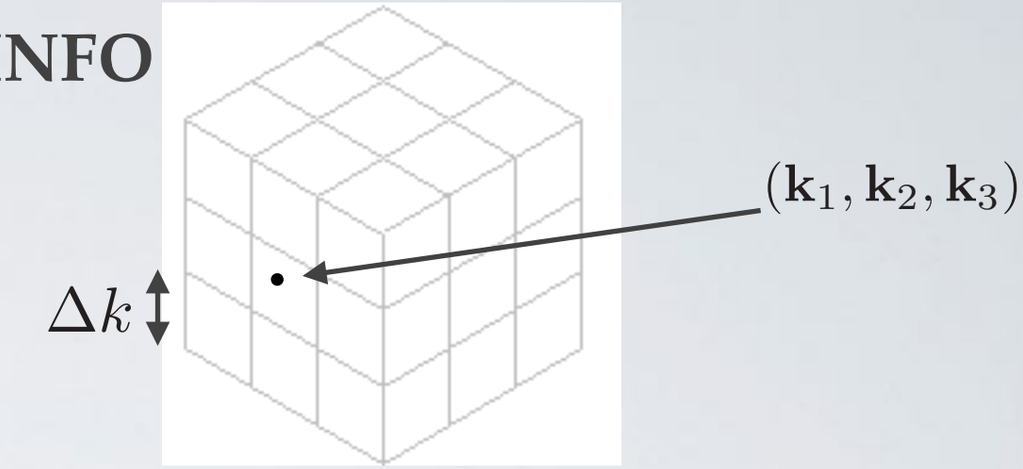


Using LSS power spectrum

arXiv:1306.4153
Sorensen, Parkinson, Davis

ISSUES WITH ADDING BISPECTRUM INFO

$$\hat{B}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = \frac{\delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\delta(\mathbf{k}_3)}{V}$$



…measured on a grid

$$\hat{B}_{\rm SE}(k_1, k_2, k_3) = \frac{(2\pi)^3}{V_B(k_1, k_2, k_3)} \int_{k_1} d^3\mathbf{q}_1 \int_{k_2} d^3\mathbf{q}_2 \int_{k_3} d^3\mathbf{q}_3 \delta_D(\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3)\hat{B}(\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3)$$

$$|\mathbf{q}_i| \in [k_i - \Delta k/2, k_i + \Delta k/2]$$

- Naively the complexity of this is $\sim N^6$ ISSUE 1: Complexity of the estimator
- Can get around this but … covariance matrix is prohibitive

Lots of mock catalogs required

ISSUE 2: Getting an accurate covariance matrix

+ ISSUE 3: Need accurate theory modelling to extract parameter constraints

$$\mathcal{L} \propto \exp\left[-(B_{\rm data} - B_{\rm model})^{\rm T} C^{-1} (B_{\rm data} - B_{\rm model})\right]$$

Inverse covariance matrix

vector of values across all measured configurations

State of the art: *SDSS III - Baryon Oscillation Spectroscopic Spectrum* arXiv:1606.00439

- 10,000 square degrees
- 2 patches in North & South Hemispheres
- BOSS Galaxies: LRGs
- LOWZ: $0.15 \leq z \leq 0.43$
- CMASS: $0.43 \leq z \leq 0.70$

$\sim 1.2 \times 10^6$ galaxies

ISSUE 1: Complexity of the estimator

$$\hat{B}_{\rm SE}(k_1,k_2,k_3) = \frac{1}{V_B(k_1,k_2,k_3)} \int_{k_1} d^3\mathbf{q}_1 \int_{k_2} d^3\mathbf{q}_2 \int_{k_3} d^3\mathbf{q}_3 \delta_D(\mathbf{q}_1+\mathbf{q}_2+\mathbf{q}_3)\hat{B}(\mathbf{q}_1,\mathbf{q}_2,\mathbf{q}_3)$$

Trick:

$$(2\pi)^3\delta_D(\mathbf{q}_1+\mathbf{q}_2+\mathbf{q}_3) = \int d^3\mathbf{x} e^{i\mathbf{x}.(\mathbf{q}_1+\mathbf{q}_2+\mathbf{q}_3)}$$

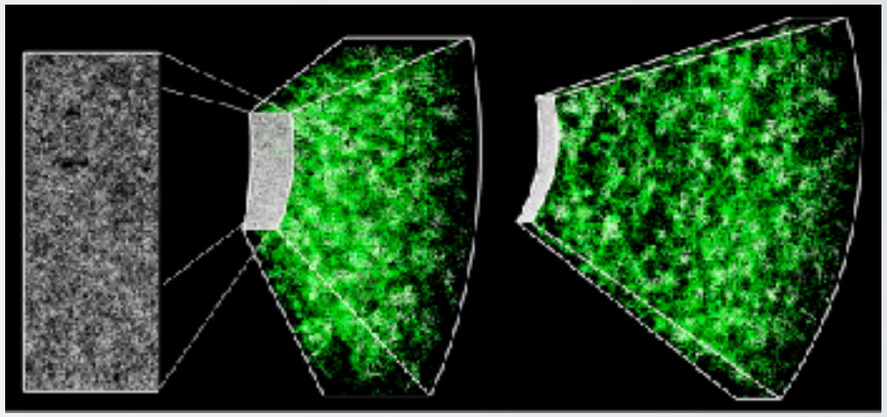


Image: Tinker et al, SDSS

$$\hat{B}_{\rm SE}(k_1,k_2,k_3) = \frac{k_f^3}{(2\pi)^6 V_B(k_1,k_2,k_3)} \int d^3\mathbf{x}\mathcal{D}_{k_1}(\mathbf{x})\mathcal{D}_{k_2}(\mathbf{x})\mathcal{D}_{k_3}(\mathbf{x}), \quad \text{where} \quad \mathcal{D}_{k_i}(\mathbf{x}) = \int_{k_i} d^3\mathbf{q}_i e^{i\mathbf{x}.\mathbf{q}_i}\delta(\mathbf{q}_i)$$

$\mathcal{O}(N^3)$ complexity

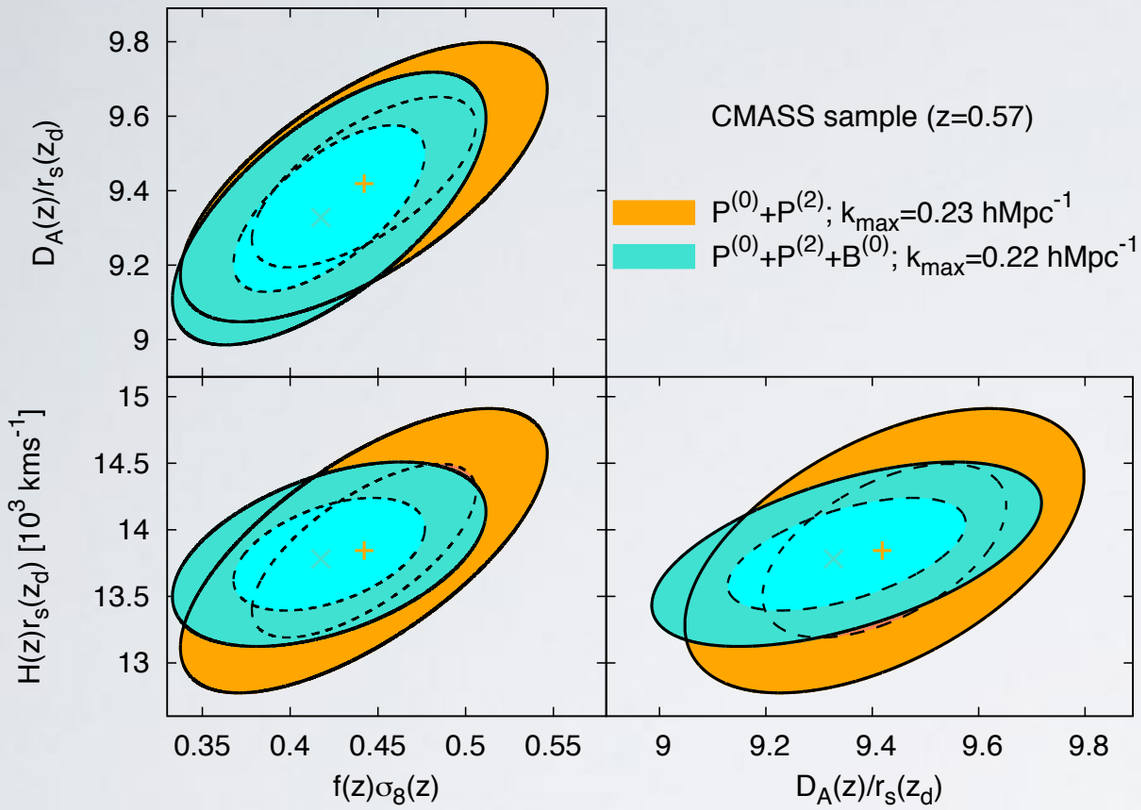
ISSUE 2: Getting an accurate covariance matrix

$\Delta k = 6k_f$, $k \in [0.03, 0.18]$h/Mpc

- 825 configurations.
- Estimated with 2048 PTHalo mocks

ISSUE 3: Systematic errors from theory modelling

Estimated from the average over the mock catalogs → Sets the truncation scale k_{max}



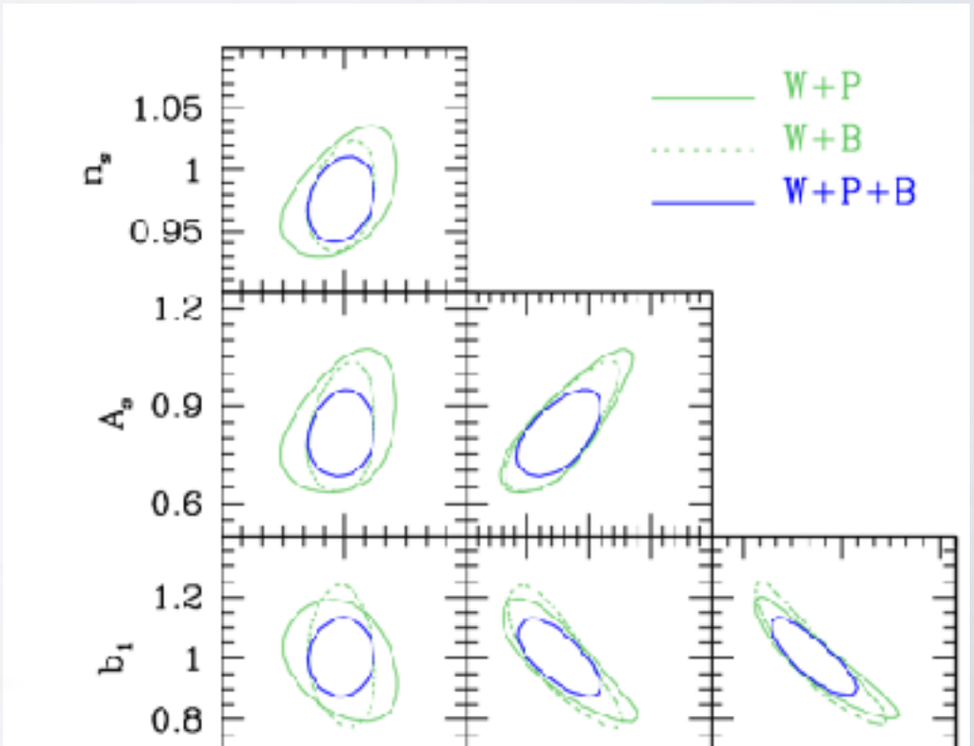
$f(z)$	$0.58 \pm 0.12\,(0.11)$
$\sigma_8(z)$	$0.668 \pm 0.078\,(0.056)$

Breaks degeneracy b / w fluctuation amplitude (σ_8) and growth rate (f)

<u>Can we improve on this?</u>

Sefusatti et al suggested there is lots more information

…but we must include more scales arXiv-ph/0604505



To progress we need more compressed (but hopefully optimal) statistics

SIMULATIONS (R.Smith)

Cosmological parameters

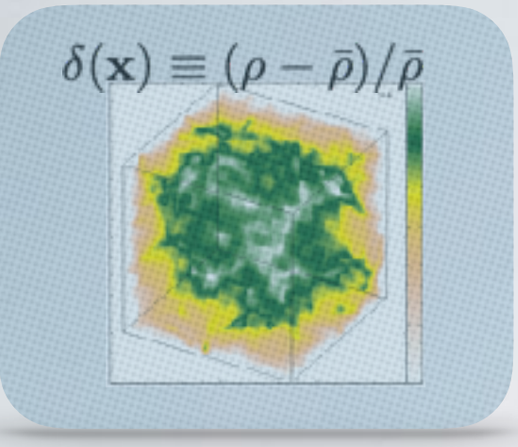
Parameter	Ω_m	Ω_b	w_0	w_a	σ_8	n_s	h
Fiducial	0.25	0.040	-1.0	0.0	0.8	1.00	0.70
Δ	± 0.05	± 0.005	± 0.2	± 0.1	± 0.1	± 0.05	± 0.05

- 200 realisations of fiducial cosmology (from ZBOX2)
- Cubes of side L=1500 Mpc/h with N=750^3 particles

✚

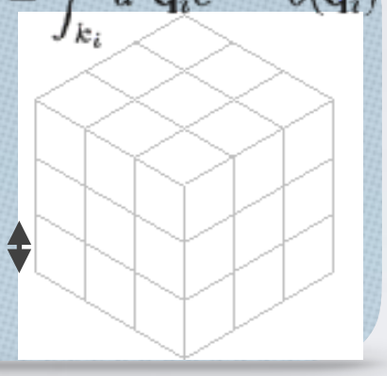
- 4 realisations for each varying parameter

Standard Estimator



Weighted FFT of side $\boldsymbol{\Delta k}$ for each (k_1,k_2,k_3)

$$\mathcal{D}_{k_i}(\mathbf{x}) = \int_{k_i} d^3\mathbf{q}_i e^{i\mathbf{x}.\mathbf{q}_i}\delta(\mathbf{q}_i)$$



$$|\boldsymbol{q}_i| \in [k_i - \Delta k/2, k_i + \Delta k/2]$$

Integrate over all points

$$\hat{B}(k_1,k_2,k_3) \propto \int d^3\mathbf{x}\mathcal{D}_{k_1}(\mathbf{x})\mathcal{D}_{k_2}(\mathbf{x})\mathcal{D}_{k_3}(\mathbf{x})$$

Repeat for many configurations

Total number is around 30000 for $\Delta k = k_f$

Covariance Matrix

$$\mathcal{L} \propto \exp\left[-(B_{\rm data} - B_{\rm model})^{\rm T} C^{-1} (B_{\rm data} - B_{\rm model})\right]$$

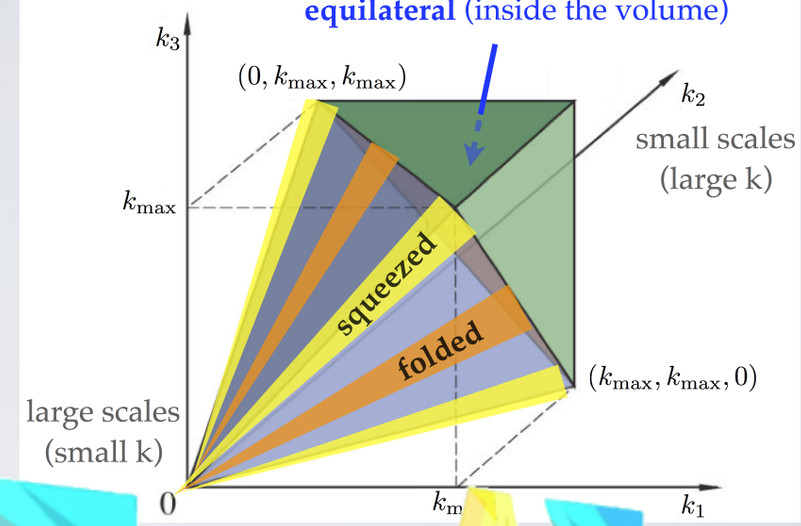
Use 200 simulations to get this ⇒ use ($\Delta k = 8k_f$) 95 configurations

Are we losing information?

Modal Estimator

arXiv: 1008.1730, 1108.3813,1207.5678

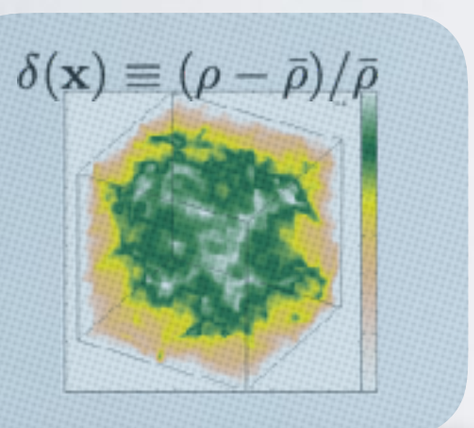
Define a basis + inner product in configuration space



Task: Find coefficients of basis expansion

What is this basis? Triples of 1D polynomials labelled $n \equiv \{n_1, n_2, n_3\}$

$$Q_n(k_1, k_2, k_3) = q_{n_1}(k_1) q_{n_2}(k_2) q_{n_3}(k_3)$$



Weighted FFT on full grid

$$\mathcal{M}_{n_i}(\mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} e^{i\mathbf{k}.\mathbf{x}} \frac{q_{n_i}(k)}{\sqrt{k\hat{P}(k)}} \delta(\mathbf{k})$$

Integrate over all points

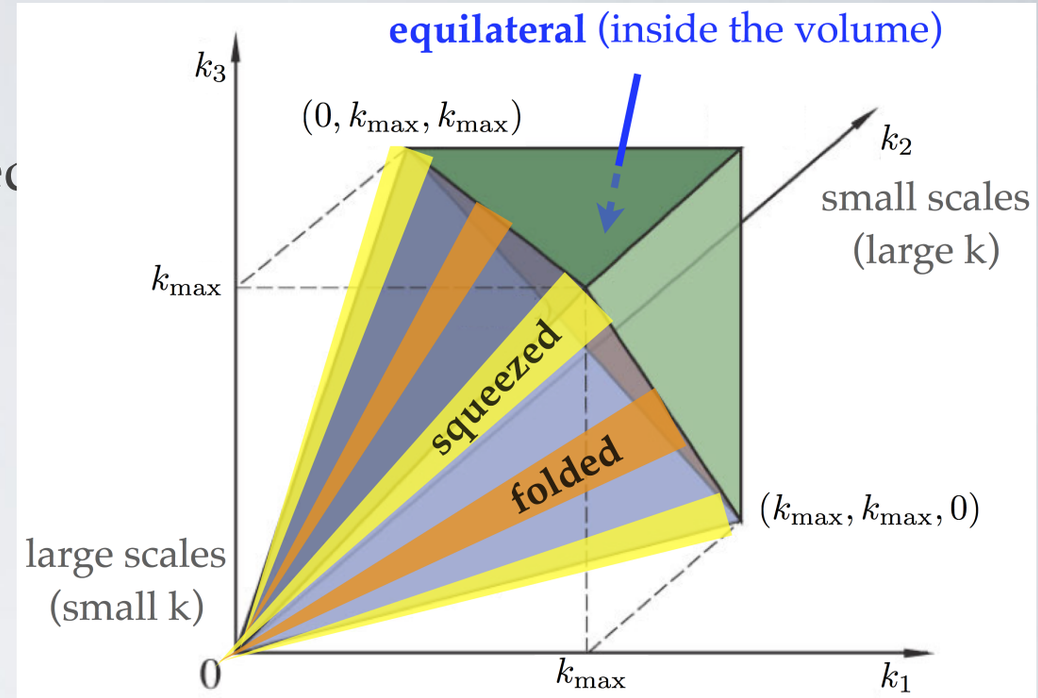
$$\beta_n \supset \int d^3\mathbf{x} \mathcal{M}_{n_1}(\mathbf{x}) \mathcal{M}_{n_2}(\mathbf{x}) \mathcal{M}_{n_3}(\mathbf{x})$$

Repeat for each mode n (just 50 used)

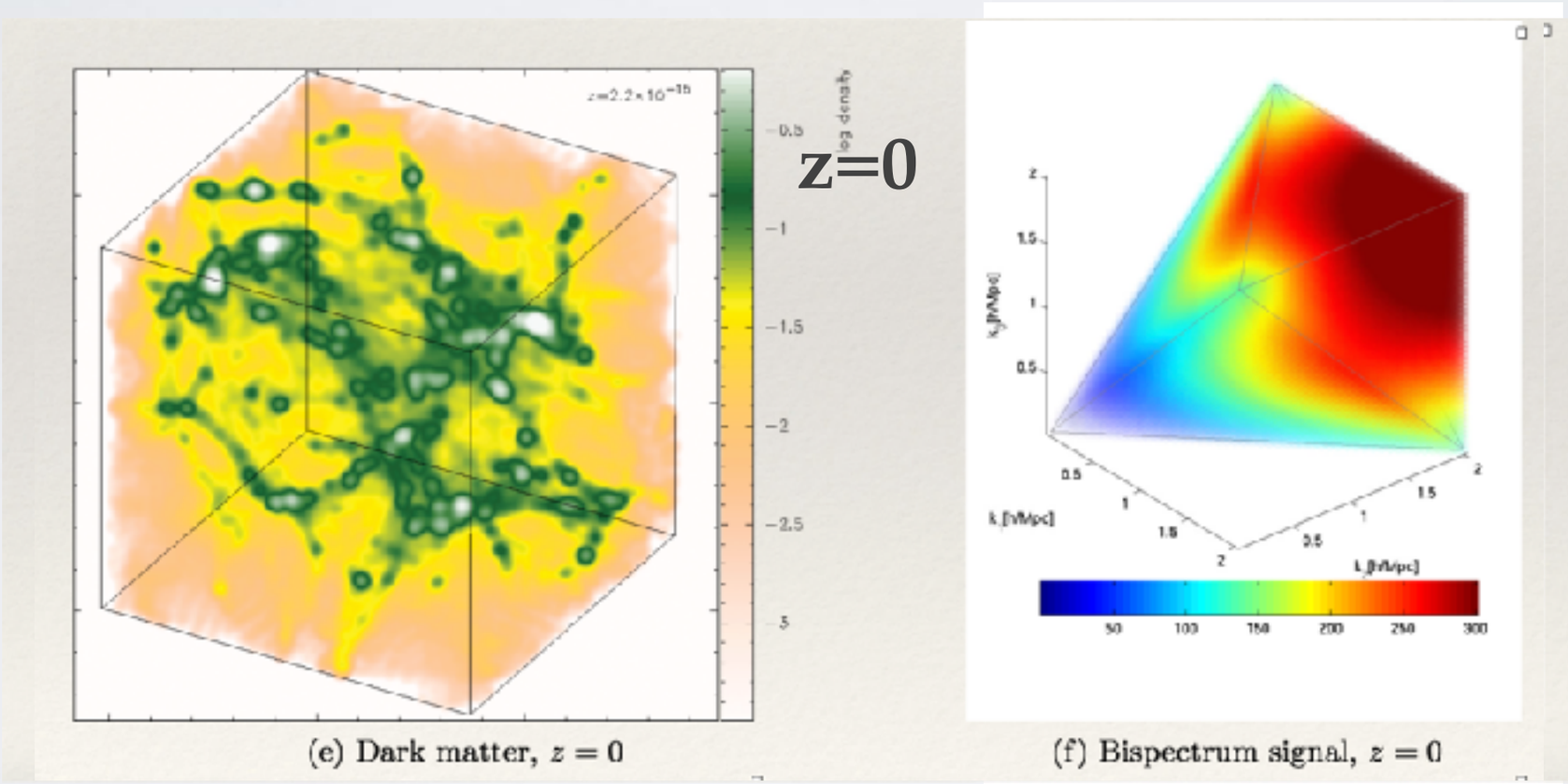
Reconstruction of bispectrum:

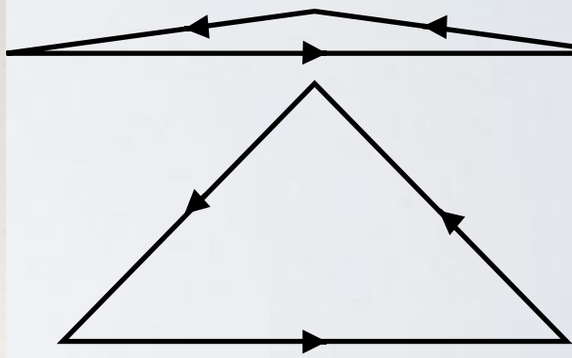
$$\hat{B}(k_1, k_2, k_3) = \sqrt{\frac{\hat{P}(k_1)\hat{P}(k_2)\hat{P}(k_3)}{k_1 k_2 k_3}} \sum_n \beta_n Q_n(k_1, k_2, k_3)$$

Appears to have little loss of signal …
but covariance matrix (50 x 50 matrix) not yet calculated
=> not yet used for parameter constraints



Useful visual interpretation of evolving bispectrum





Filaments & clusters appearing

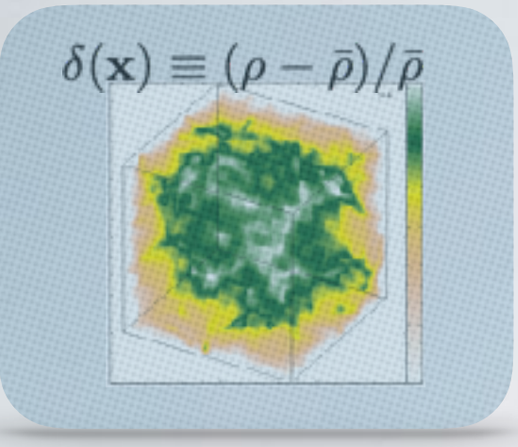
More clustering structure

Flattened & equilateral

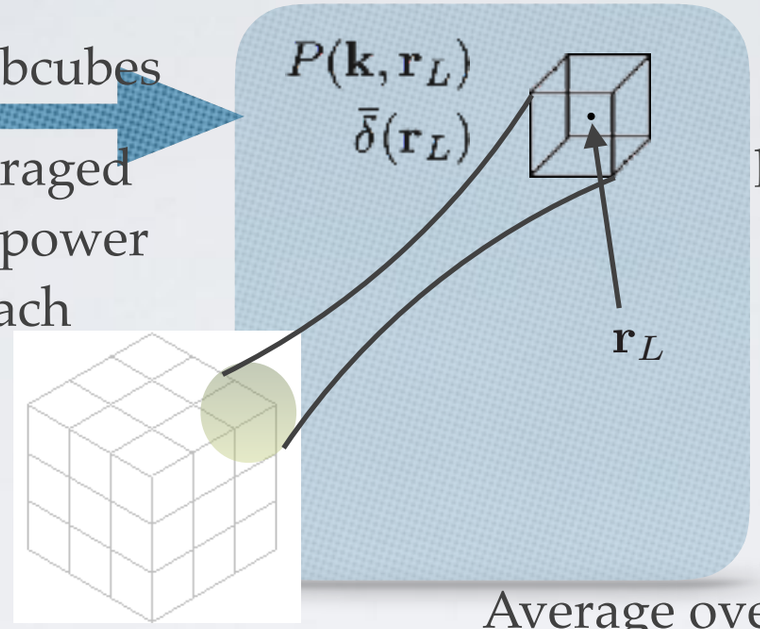
Flattened & equilateral

Integrated Bispectrum

arXiv:1403.3411 Chiang, Wagner, Schmidt, Komatsu



Split into subcubes and get averaged density and power spec in each



Average the product over all subcubes

$$iB_L(\mathbf{k}) = \langle P(\mathbf{k}, \mathbf{r}_L)\bar{\delta}(\mathbf{r}_L)\rangle$$

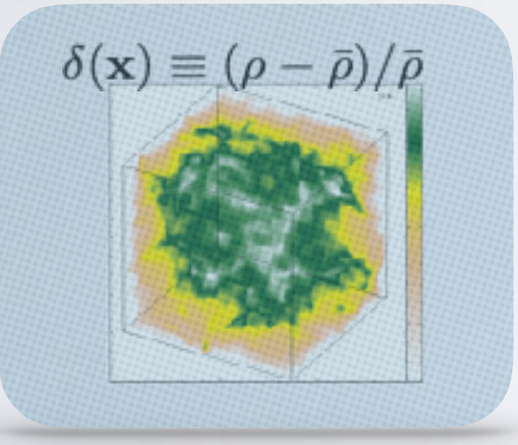
Average over the angular directions $\hat{\mathbf{k}}$

$$iB_L(k) = \int \frac{d^2\hat{\mathbf{k}}}{4\pi} iB_L(\mathbf{k})$$

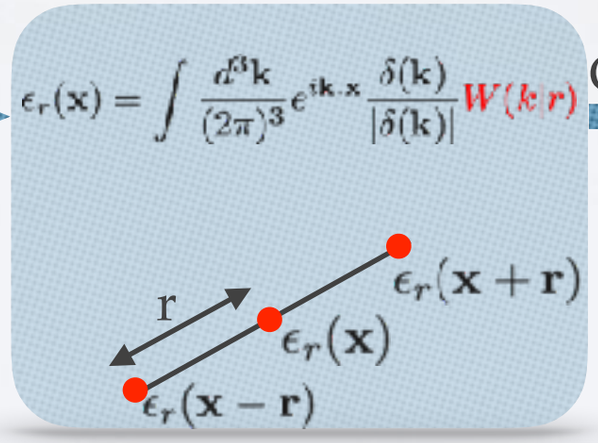
Line Correlation Function

arXiv:1409.3007 Obreschkow et al

Application to cosmological parameter estimation in arXiv:1611.01160 (Eggemeier & Smith)



Get the phase and smooth on scale r about **x**

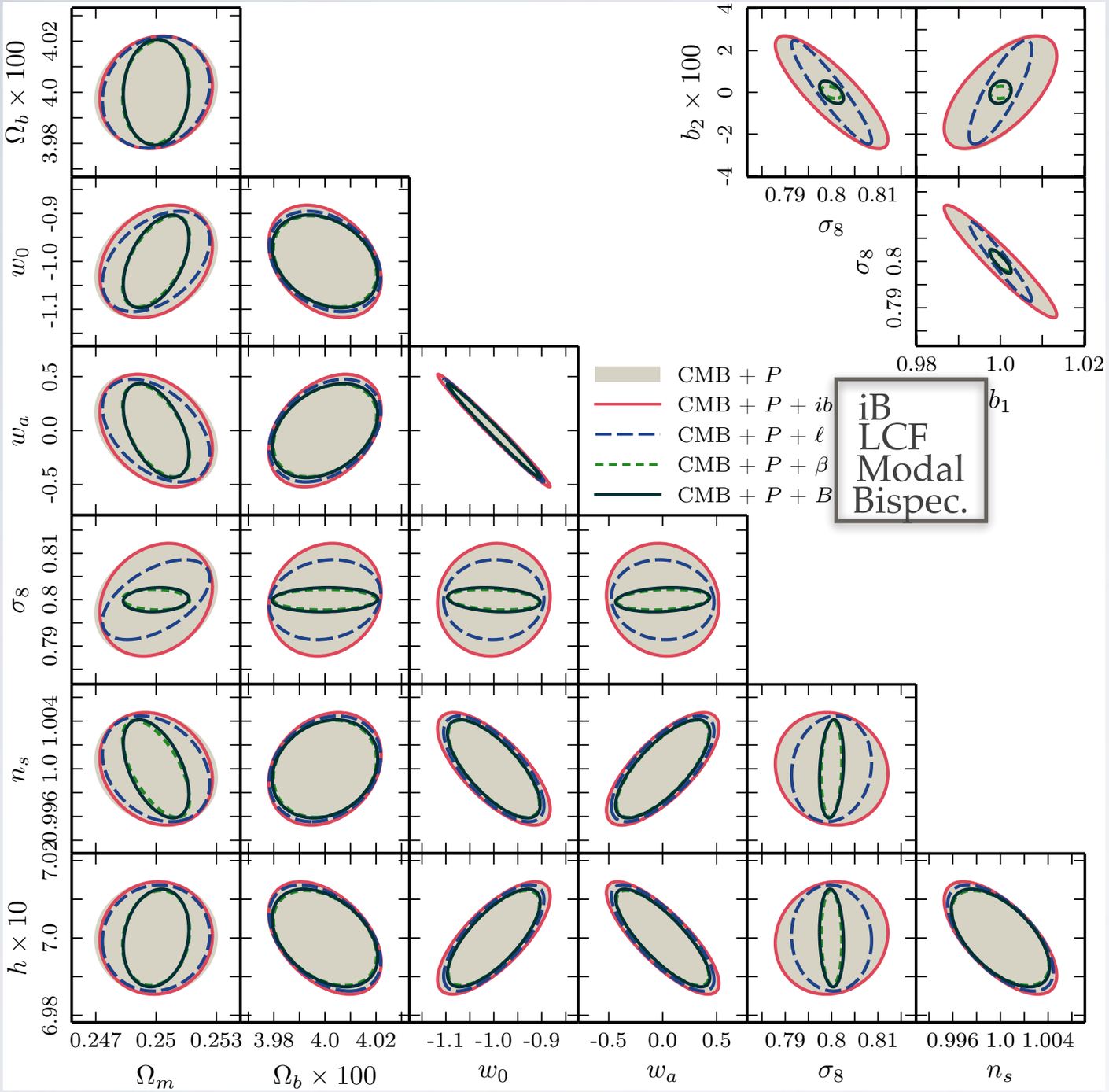


Get sum of points and average over orientations

$$\hat{\ell}(r) \propto \int d^2\hat{\mathbf{r}} \int d^3\mathbf{x}\, \epsilon_r(\mathbf{x}-\mathbf{r})\epsilon_r(\mathbf{x})\epsilon_r(\mathbf{x}+\mathbf{r})$$

Repeat for many scales r

PARAMETER ESTIMATES



- iB gives little info
- LCF improvement mainly due to degeneracy breaking
- Modal estimator optimal

just 10 modes is enough!!

Up to factor of 2 improvement possible!

CONCLUSIONS (b)

- Bispectrum offers a valuable resource… ***no reason to* ignore** it!!!

- Important to have several ***complementary*** estimators — differing noise sensitivities

- ***S/N*** can be misleading … Tips: (1) use B_φ (2) include cross-correlations

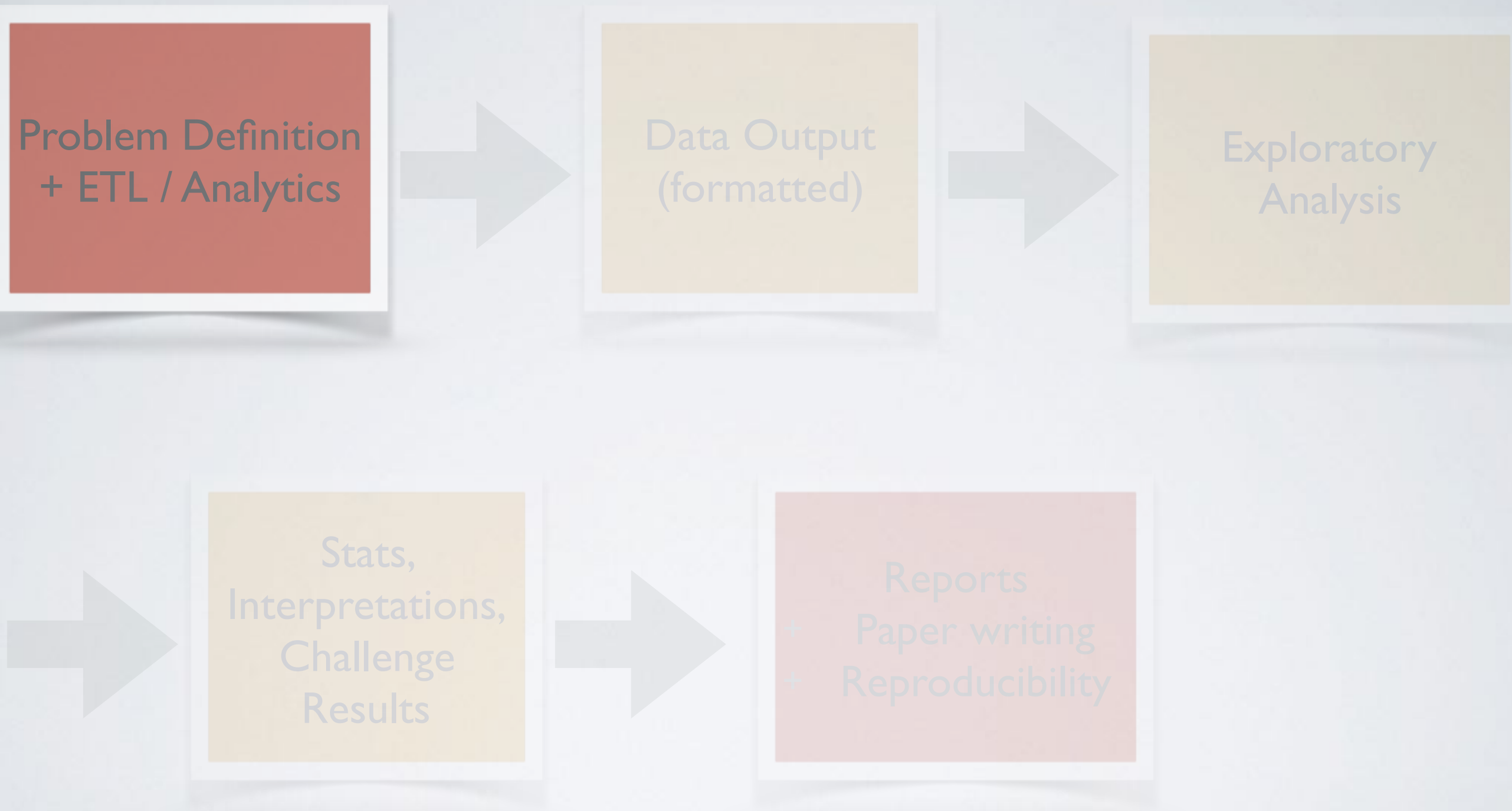
- Potential **improvement in error bars** by factor of 2 - only *partially* due to degeneracy breaking

- **Modal** estimator is ***optimal*** ⇒ large data *compression* ⇒ covariance matrix calc simpler (*maybe (future!) use for derivative fitting from simulations in place of theory*)

- *Theory modelling* errors (slightly) degrade these improvements

- Work to do (RSD, survey geometry, …) before confronting data … stay tuned!

How did we get here?


Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

Organising Team & Sharing findings

Slack/meekan/Dropbox/GitHub/Markdown!!

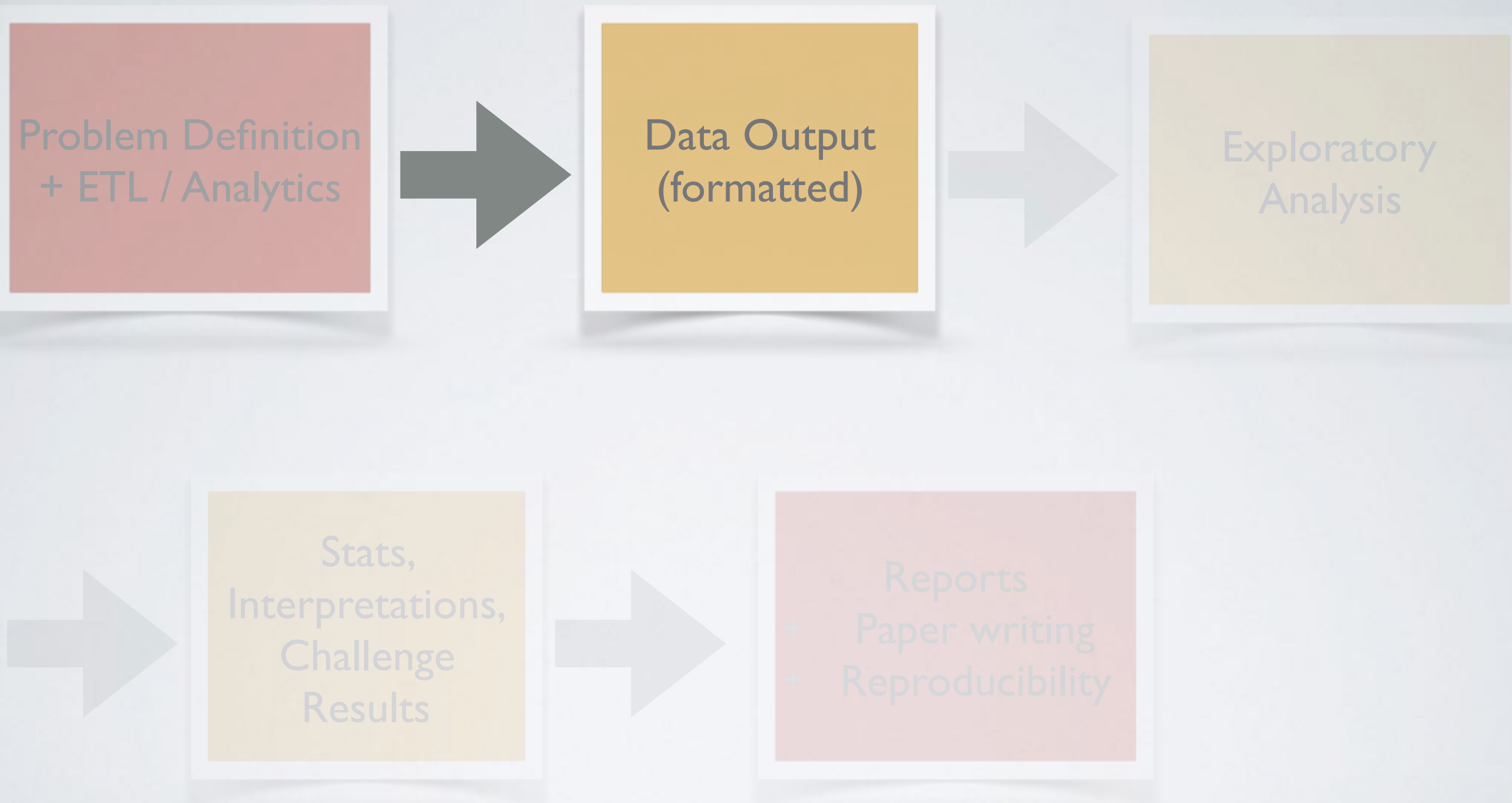
Estimators Code

Written in C++
Parallelise with OpenMP/MPI
Run on Apollo Supercomputer
(https://info.hpc.sussex.ac.uk/hpc-guide/access/index.html)

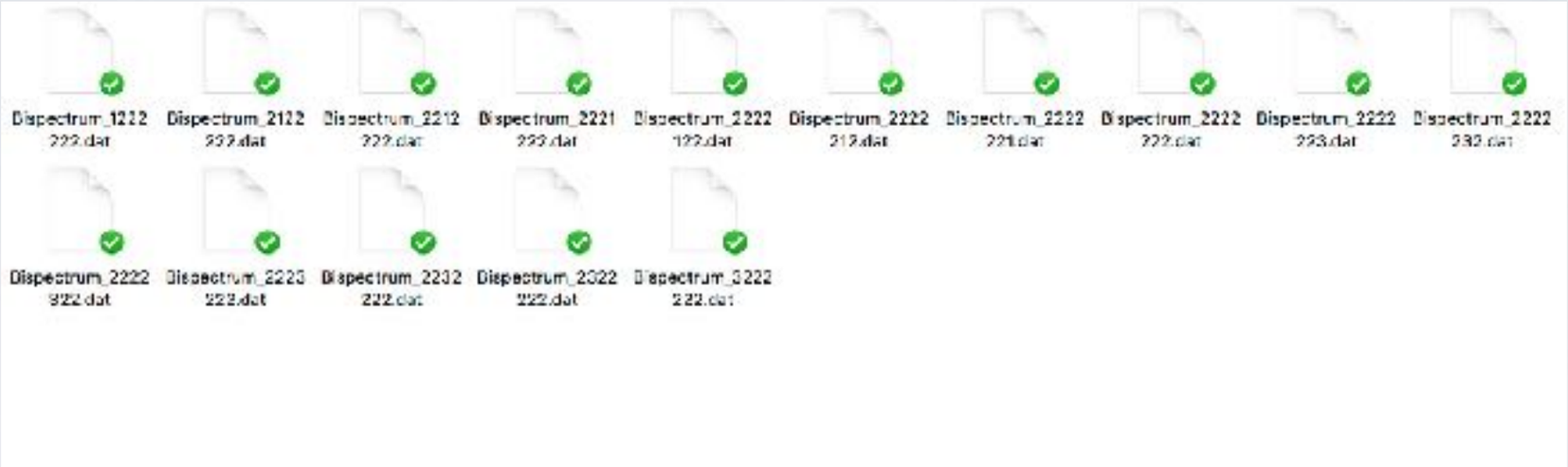
job.sh

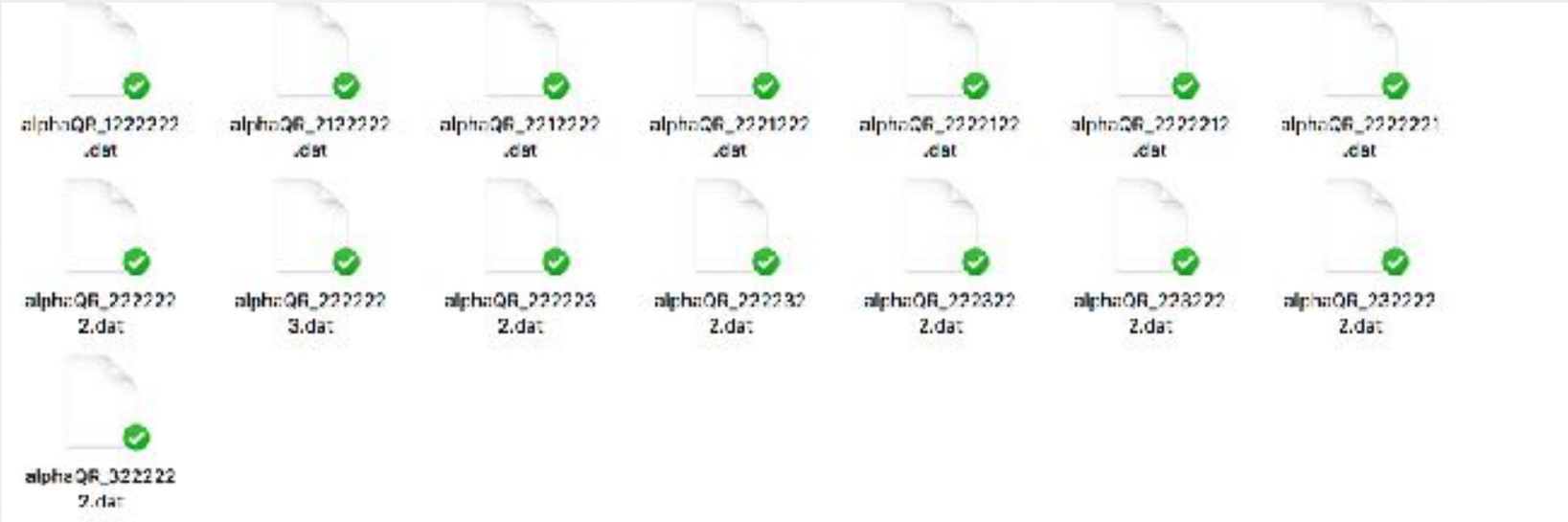
job.sh 〉 No Selection

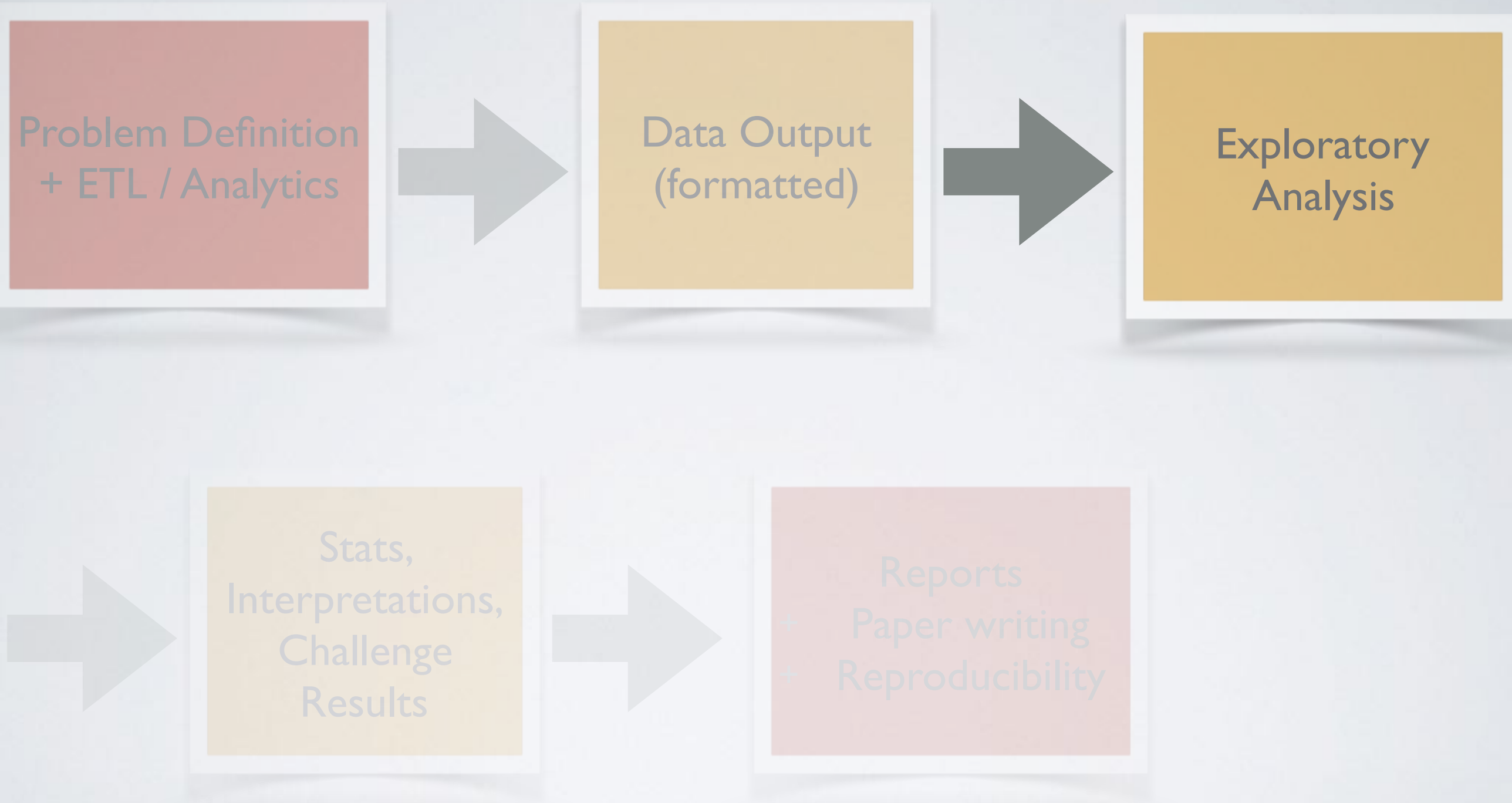
```
#!/bin/sh
#$ -N NAME_OF_YOUR_JOB
#$ -M youremail@sussex.ac.uk
#$ -m bea
#$ -cwd
#$ -pe openmpi_mixed_8 16  # this requests 8 cores per node and 16 cores in total, you can get list of other
    supported parallel environments with 'qconf -spl'
#$ -o output.dat


# queue:
#$ -q mps.q

# modules:
module add sge
module load gcc/4.8.1
module load openmpi/gcc/64/1.7.3
module load fftw3/openmpi/gcc/64/3.3.4

export OMP_NUM_THREADS=1

mpirun --npernode 8 -np ${NSLOTS} ./executable  # specify again how many cores per node you request
```


Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

- Label files clearly and consistently (makes pipeline analysis easier)





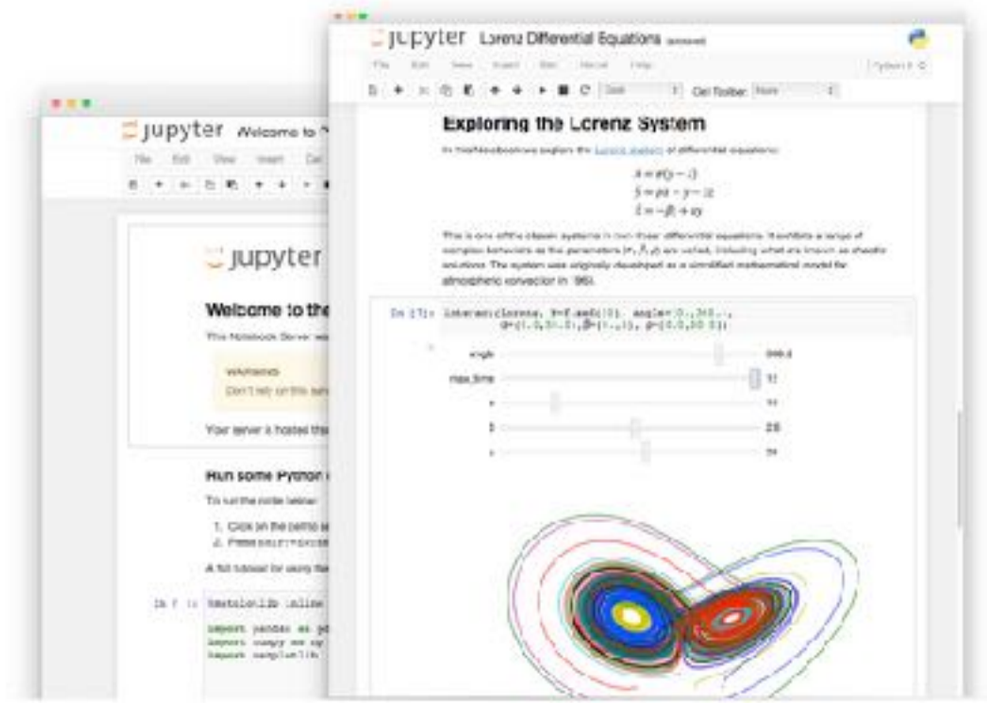

Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

jupyter.org

Project Jupyter | Home

jupyter

Install About Resources Documentation NBViewer Widgets Blog Donate



The Jupyter Notebook

The Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

Language of choice

The Notebook has support

Share notebooks

Notebooks can be shared

Interactive widgets

Code can produce rich

Big data integration

Leverage big data tools, such

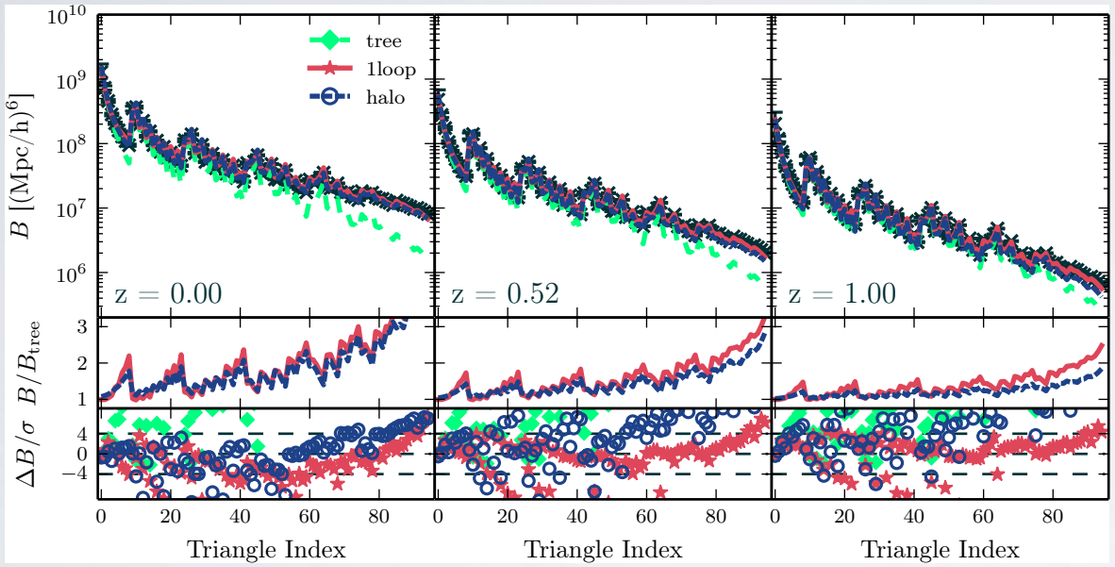
Fisher matrix analysis

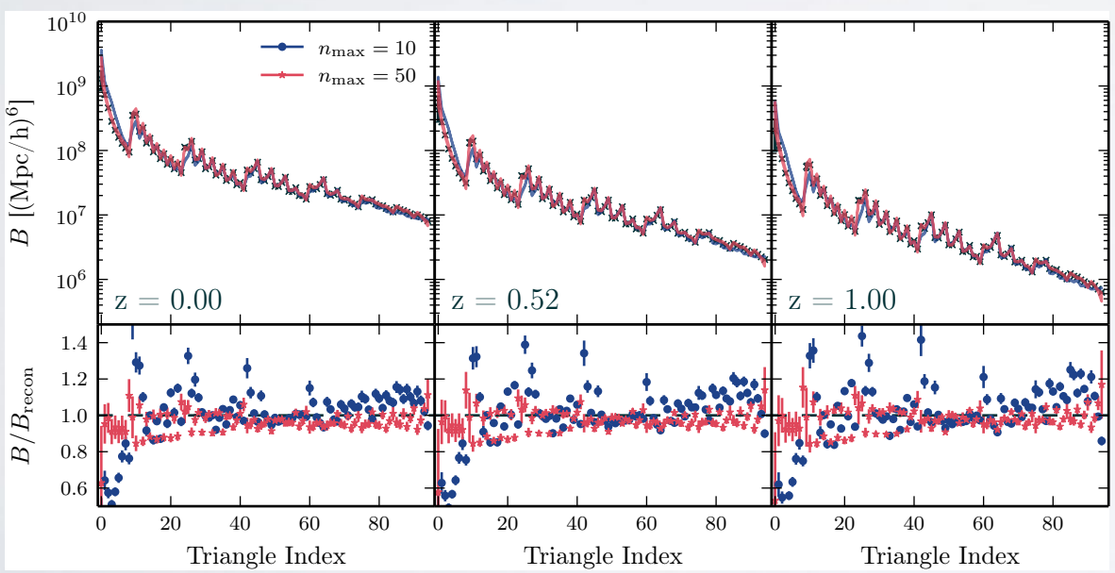
$$F_{\Omega_1\Omega_2} = \sum_{ij} \frac{\partial \mu_i}{\partial \Omega_1} C_{ij}^{-1} \frac{\partial \mu_j}{\partial \Omega_2}$$

$$\mu = (\underbrace{P_1, \ldots, P_N}_{\text{various P(k)}}, \underbrace{B_1, \ldots, B_M}_{\text{various configs of } B(k_1,k_2,k_3)})$$

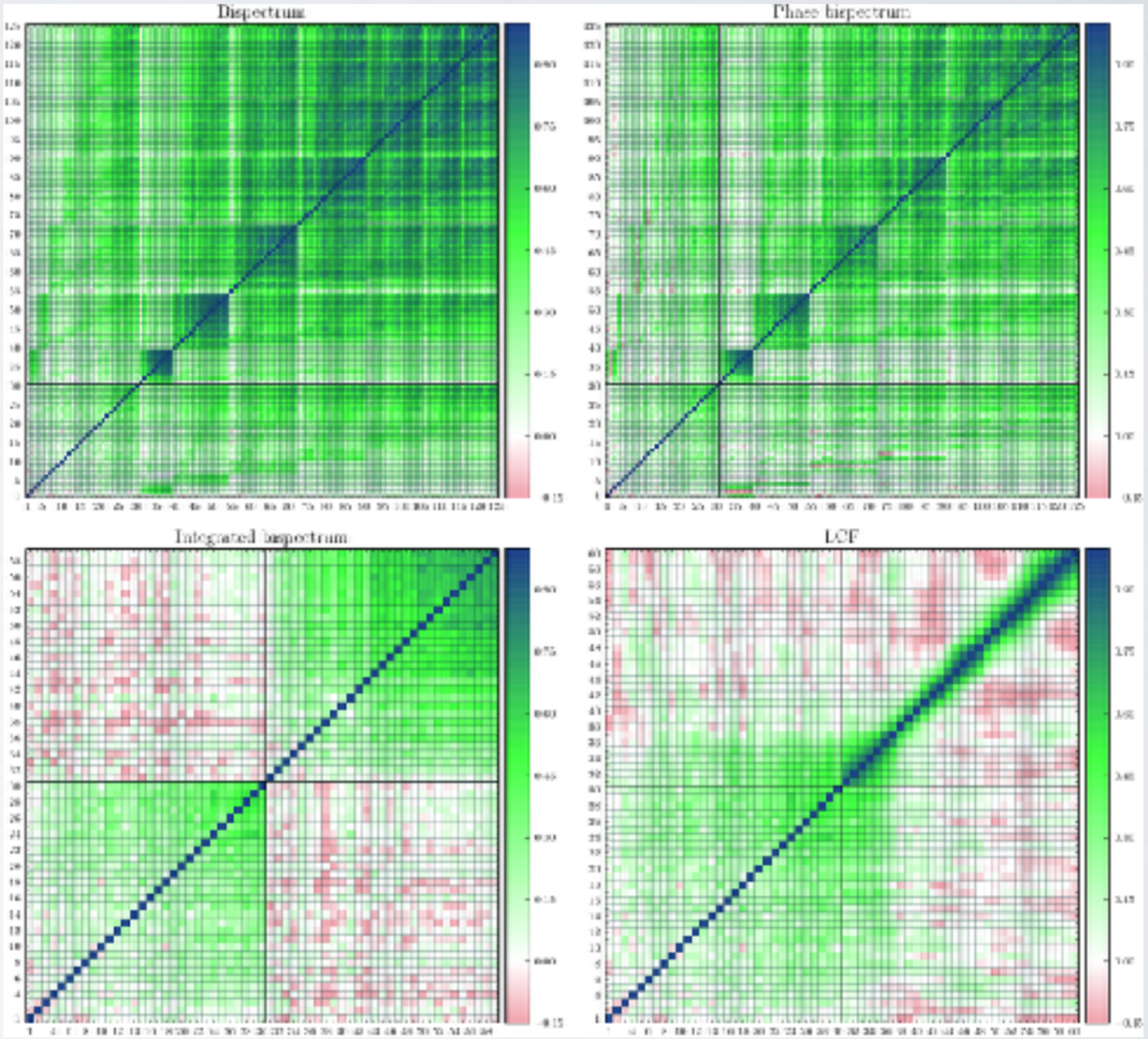
Need *derivatives* & *covariance* matrix

Means (data vs theory)





Covariance matrices



Branch: master ▾ notebooks / Signal-to-Noise & Fisher Matrices.ipynb Find file Copy path

aegge Add comparison between beta50 and beta10, add bar charts for model co... cf65fd9 on 9 Jan

1 contributor

8.62 MB Download History

In [1]:
```
#load modules
%run /home/a/ae/ae237/notebooks/modules.py
from stan import Stan
```

In [2]:
```
P = Stan('PSpec')
B = Stan('BSpec')
Bphase = Stan('BSpecPhase')
iB = Stan('iBSpec')
L = Stan('LC')
beta = Stan('BModes')

PB = Stan(['PSpec','BSpec'])
PBphase = Stan(['PSpec','BSpecPhase'])
PiB = Stan(['PSpec','iBSpec'])
PL = Stan(['PSpec','LC'])
Pbeta = Stan(['PSpec','BModes'])
```

Signal-to-Noise

Bispectrum

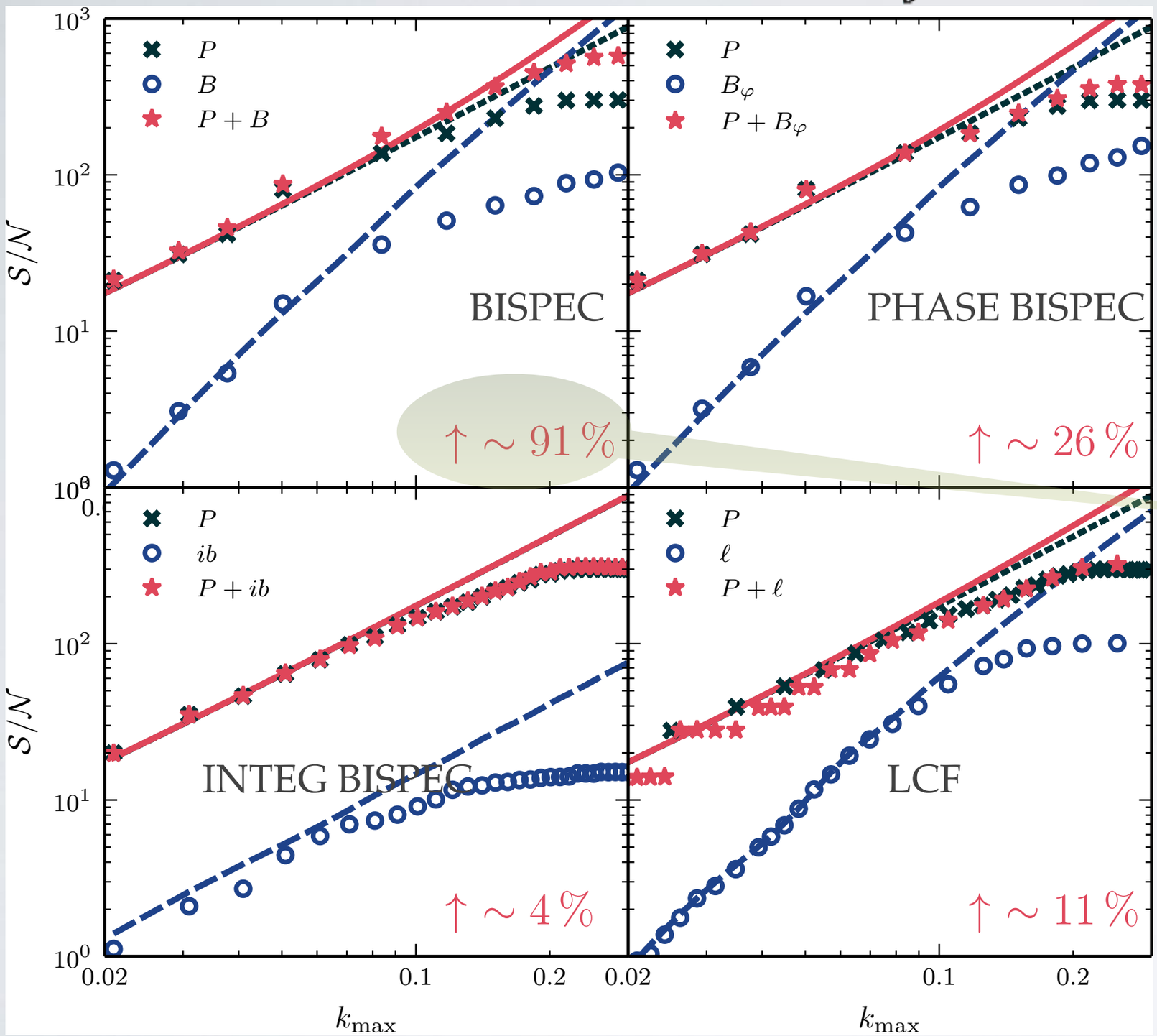
In [3]:
```
P.cumulative_signal2noise()
B.cumulative_signal2noise()
PB.cumulative_signal2noise()
```

In [25]:
```
# Gaussian signal-to-noise
kf = 2*pi/1500
k = kf/2*(1+2*arange(100))
S2NG_P = zeros(k.size)
for i in range(1,k.size+1):
    S2NG_P[i-1] = sqrt(2*pi*((k[:i]/kf)**2).sum())

S2NG_P_spline = interp1d(k,S2NG_P,kind='cubic')
```

In [16]:
```
def F2(k1, k2, k3):
    mu = (k3**2-k1**2-k2**2)/(2*k1*k2)
    return 5./7 + mu/2*(k1/k2+k2/k1) + 2./7*mu**2
```

SIGNAL-TO-NOISE (possibly misleading!!)

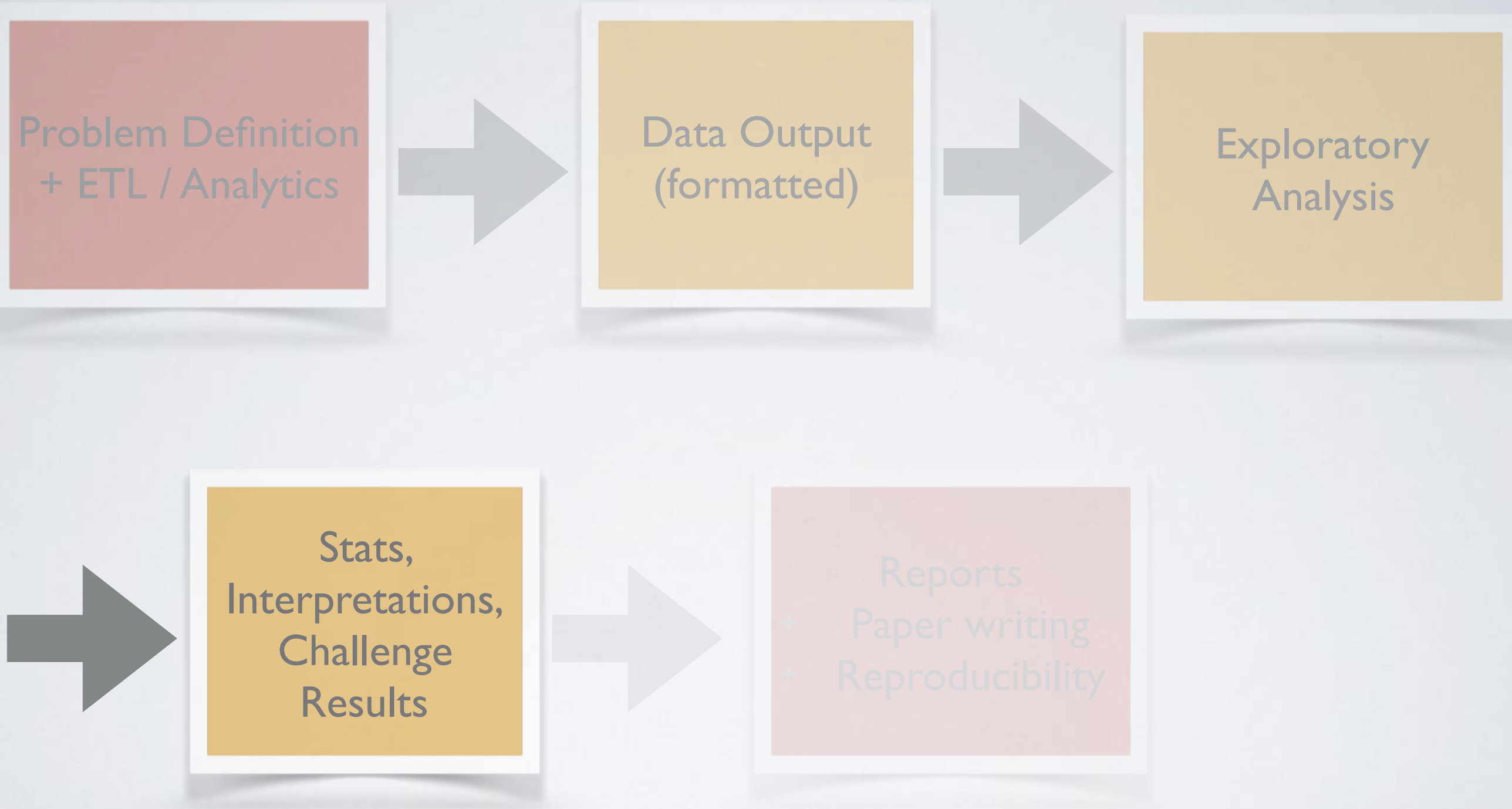
$$\mathcal{S}/\mathcal{N} = \sum_{ij} B_i C_{ij}^{-1} B_j$$



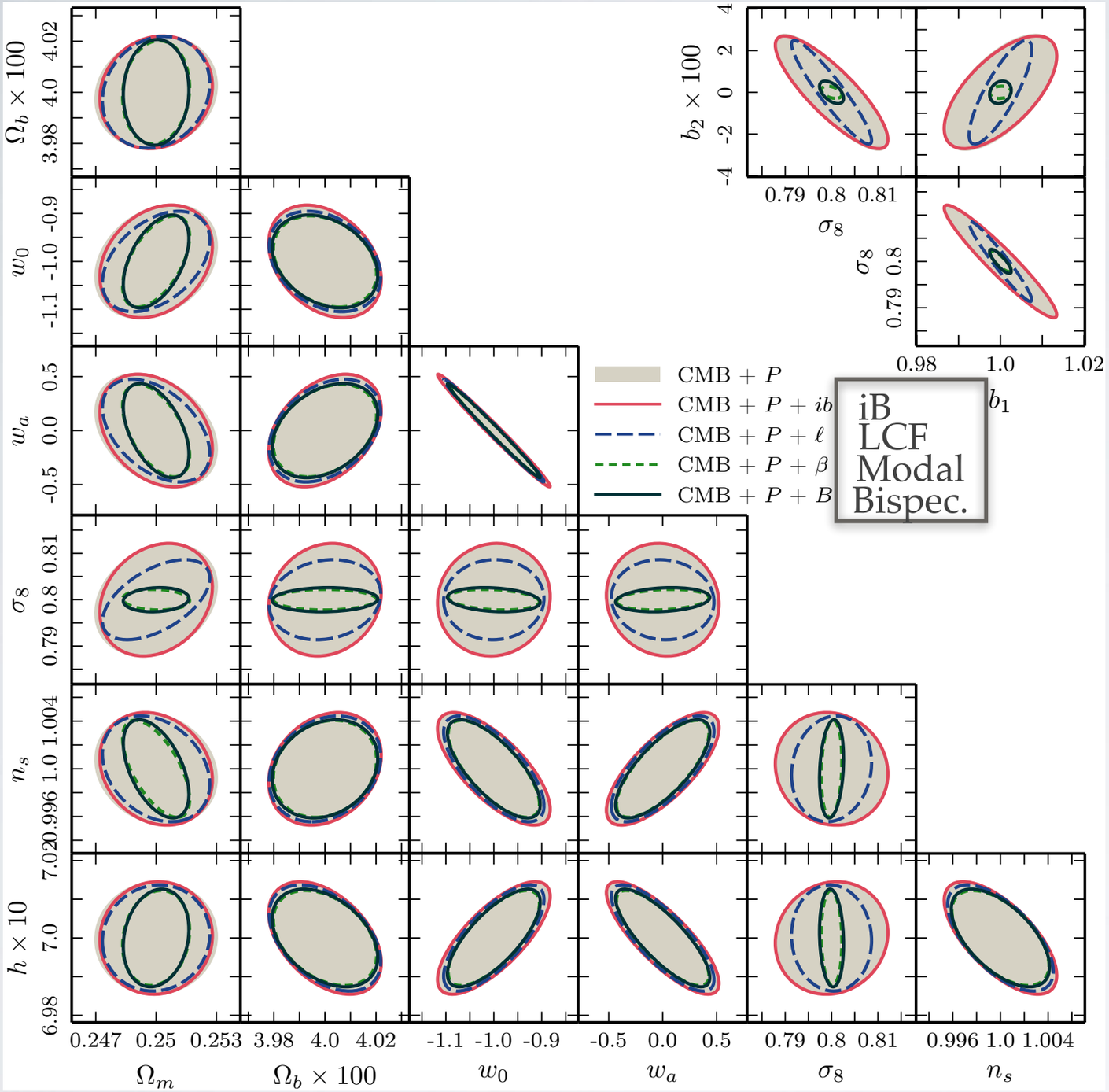
Underestimates B only
Overestimates P+B

4 lessons

1) Cross-correlations are vital
2) S/N of phase bispectrum is better proxy for information content
3) iB inferior
4) Modal estimator appears "optimal"


Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

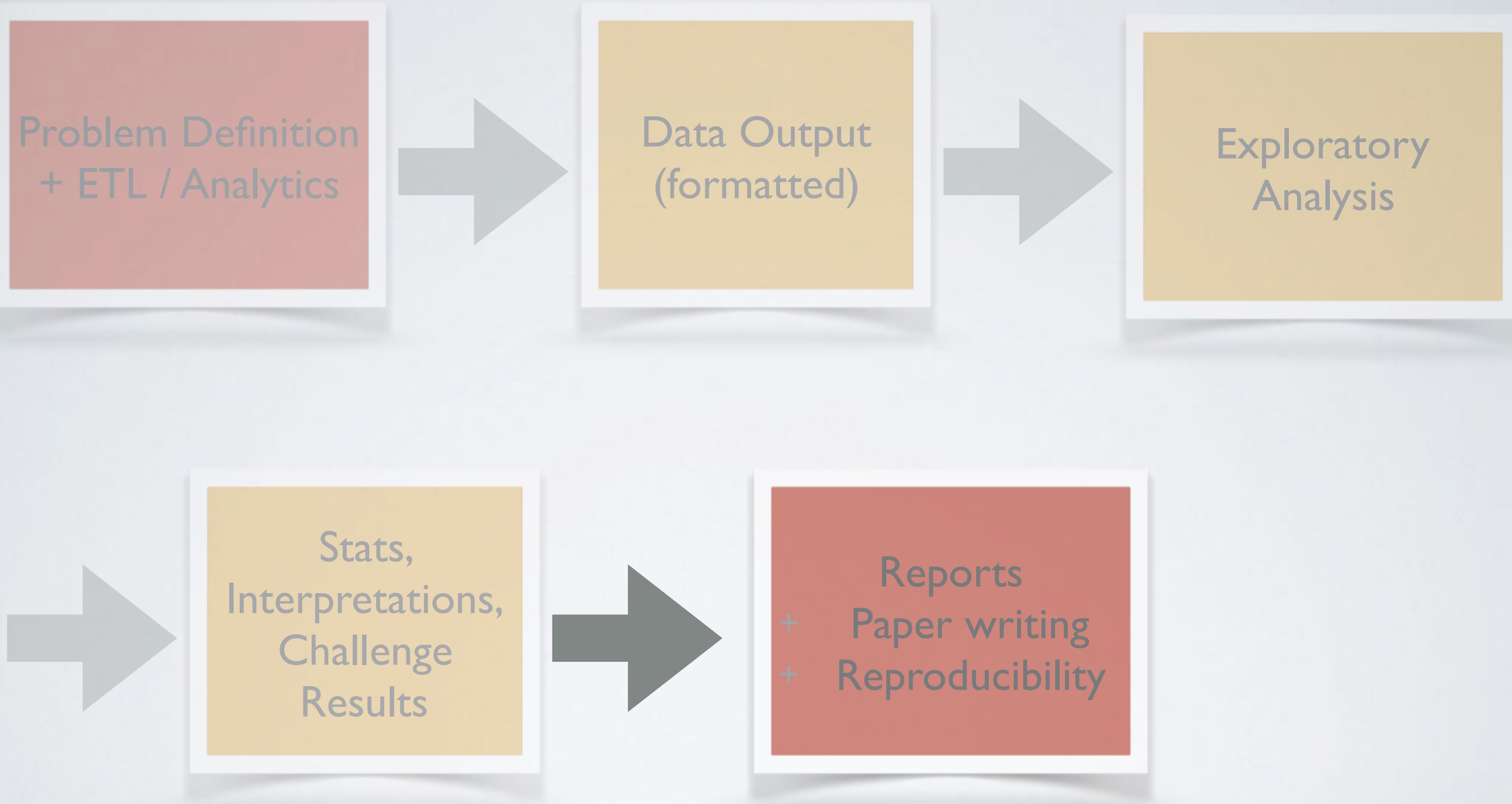
PARAMETER ESTIMATES



- iB gives little info
- LCF improvement mainly due to degeneracy breaking
- Modal estimator optimal

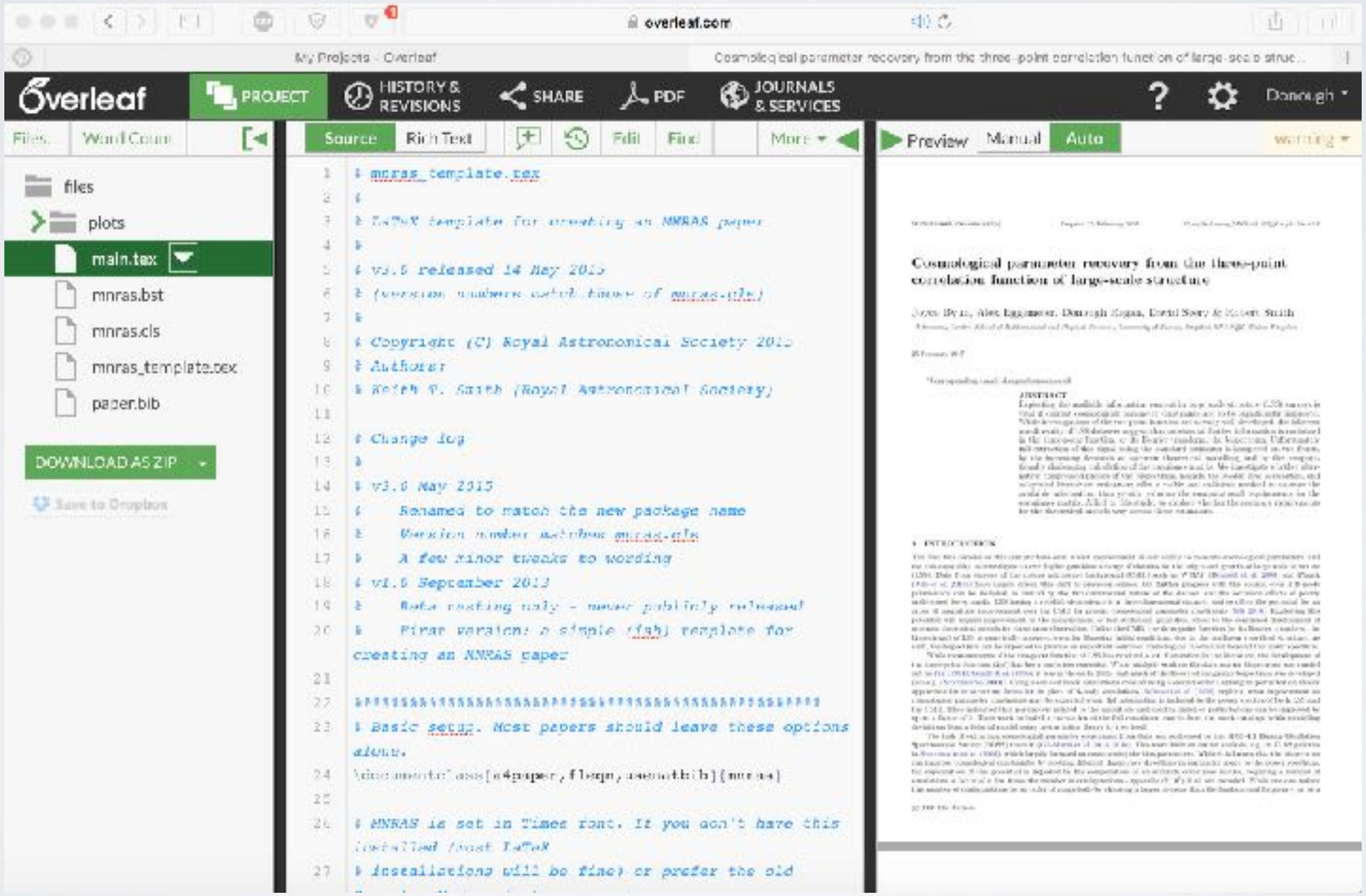
just 10 modes is enough!!

Up to factor of 2 improvement possible!


Problem Definition
+ ETL / Analytics
Data Output
(formatted)
Exploratory
Analysis
Stats,
Interpretations,
Challenge
Results
Reports
+ Paper writing
+ Reproducibility

Reproducible Research

Code will be released on GitHub/BitBucket

\+ via zenodo.org

Paper Writing