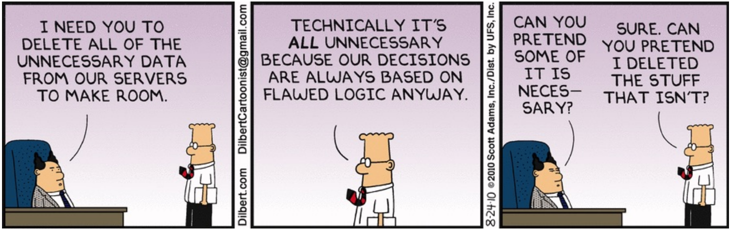
Astrostatistics and Astroinformatics

Big-Data in Astronomy and Astrophysics

Jason McEwen
www.jasonmcewen.org
@jasonmcewen

Mullard Space Science Laboratory (MSSL)
University College London (UCL)



UK Dark Energy Strategy 2020
Royal Astronomical Society, London, January 2016

Outline

What is big-data?

A. Gandomi, M. Haider / International Journal of Information Management 35 (2015) 137–144

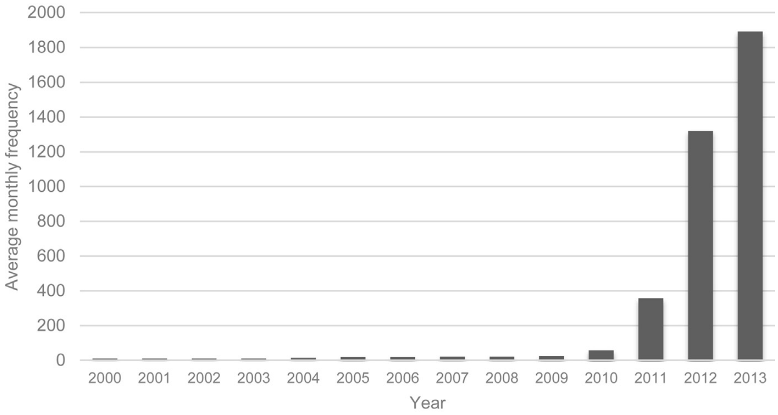


Fig. 1. Frequency distribution of documents containing the term "big data" in ProQuest Research Library.

What is big-data?

The nVs (originally 3Vs, then 6Vs, then 10Vs, . . .):

1. Volume: many bytes (e.g. typically peta, exabytes)
2. Variety: structural heterogeneity (e.g. sub-populations, variety of sources)
3. Velocity: rate of generation and analysis
4. Veracity: unreliability in sources
5. Variability: variation in data flow rate
6. Value: low value density
7. . . .

Typically (but not exclusively) characterised by:

- High-dimensional datum (wide)
- Massive number of datum (deep)

What is big-data?

The nVs (originally 3Vs, then 6Vs, then 10Vs, . . .):

1. Volume: many bytes (e.g. typically peta, exabytes)
2. Variety: structural heterogeneity (e.g. sub-populations, variety of sources)
3. Velocity: rate of generation and analysis
4. Veracity: unreliability in sources
5. Variability: variation in data flow rate
6. Value: low value density
7. . . .

Typically (but not exclusively) characterised by:

- High-dimensional datum (wide)
- Massive number of datum (deep)

What is big-data?

The nVs (originally 3Vs, then 6Vs, then 10Vs, . . .):

1. Volume: many bytes (e.g. typically peta, exabytes)
2. Variety: structural heterogeneity (e.g. sub-populations, variety of sources)
3. Velocity: rate of generation and analysis
4. Veracity: unreliability in sources
5. Variability: variation in data flow rate
6. Value: low value density
7. . . .

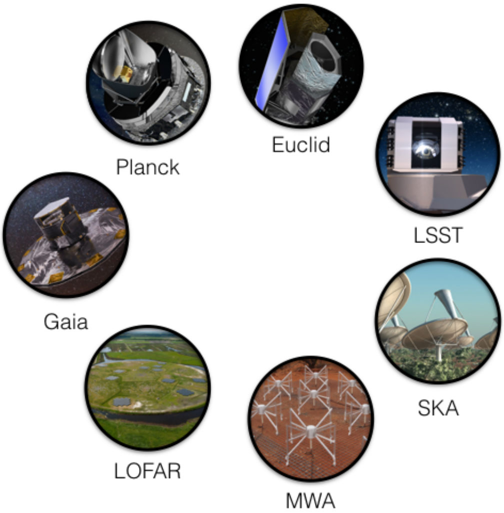
Typically (but not exclusively) characterised by:

- High-dimensional datum (wide)
- Massive number of datum (deep)

What is big-data in astronomy and astrophysics?

- Big machines (*e.g.* physical hardware, experiments)
- Big theory
- Big simulations
- Big parameter space
- Big algorithms
- Big collaborations
- Big engagement (*e.g.* outreach, industry)

What is big-data in astronomy and astrophysics?



Wide and deep observations (in addition to wide and deep data)

Challenges of big-data

A. Gandomi, M. Haider / International Journal of Information Management 35 (2015) 137–144

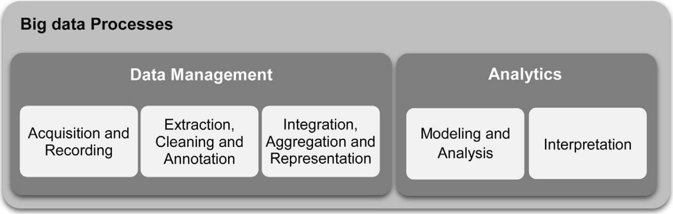


Fig. 3. Processes for extracting insights from big data.

Computational challenges:

- Data too big (to hold in memory)
- Access and analysis too slow (unfeasible)
- Too much power/energy required

Challenges of big-data

A. Gandomi, M. Haider / International Journal of Information Management 35 (2015) 137–144

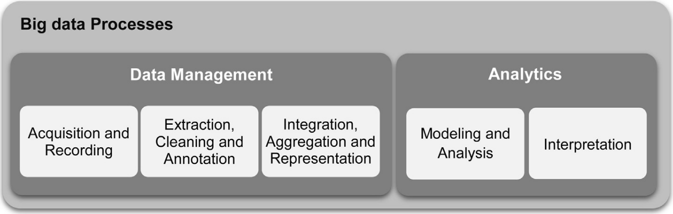


Fig. 3. Processes for extracting insights from big data.

Computational challenges:

- Data too big (to hold in memory)
- Access and analysis too slow (unfeasible)
- Too much power/energy required

Challenges of big-data

Analysis challenges (Fan et al. 2014):

1. Heterogeneity, e.g. sub-populations, different data sources, tension between data
2. Error accumulation, e.g. high-dimensional parameter spaces, bias
3. Spurious correlations, e.g. correlation vs causation, data dredging
4. Incident endogeneity, e.g. chance correlation between signal of interest and error

Challenges of big-data

Analysis challenges (Fan et al. 2014):

1. Heterogeneity, e.g. sub-populations, different data sources, tension between data
2. Error accumulation, e.g. high-dimensional parameter spaces, bias
3. Spurious correlations, e.g. correlation vs causation, data dredging
4. Incident endogeneity, e.g. chance correlation between signal of interest and error

Challenges of big-data

Analysis challenges (Fan et al. 2014):

1. Heterogeneity, e.g. sub-populations, different data sources, tension between data
2. Error accumulation, e.g. high-dimensional parameter spaces, bias
3. Spurious correlations, e.g. correlation vs causation, data dredging
4. Incident endogeneity, e.g. chance correlation between signal of interest and error

Challenges of big-data

Analysis challenges (Fan et al. 2014):

1. Heterogeneity, e.g. sub-populations, different data sources, tension between data
2. Error accumulation, e.g. high-dimensional parameter spaces, bias
3. Spurious correlations, e.g. correlation vs causation, data dredging
4. Incident endogeneity, e.g. chance correlation between signal of interest and error

Analysing big-data

Generic approaches to analysing big-data (Wang et al. 2015):

- Subsample
- Divide-and-conquer
- Stream processing

Additional approaches in astronomy and astrophysics:

- Exploit structure (geometry, symmetry, physics)
- Modelling:
  - Model-based consolidatory science
  - Model-agnostic exploratory science
- Approximation
- ...

Analysing big-data

Generic approaches to analysing big-data (Wang et al. 2015):

- Subsample
- Divide-and-conquer
- Stream processing

Additional approaches in astronomy and astrophysics:

- Exploit structure (geometry, symmetry, physics)
- Modelling:
  - Model-based consolidatory science
  - Model-agnostic exploratory science
- Approximation
- ...

Analysing big-data

Examples of specific methods:

- Bayesian analysis
- MCMC sampling
- Hierarchical probabilistic (Bayesian) models
- Variable selection
- Experimental design
- Machine learning
- Optimisation
- Wavelets
- Sparsity
- Compressed sensing
- ...

$\Rightarrow$ Astrostatistics and Astroinformatics

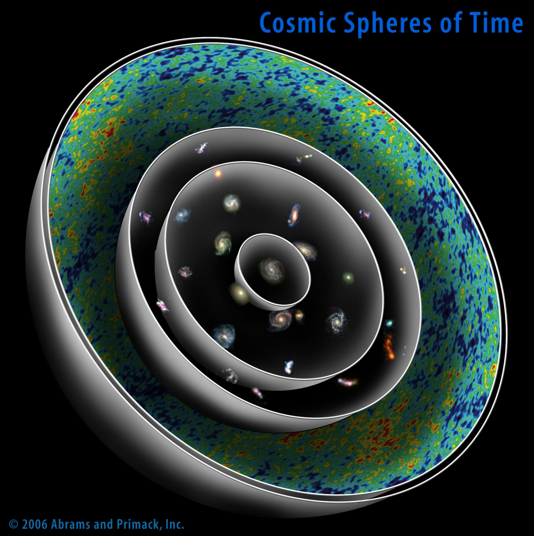
Outline

1. Big-data in astronomy and astrophysics

2. Illustrative analyses
   - Planck
   - Euclid
   - LSST
   - SKA

3. Concluding remarks

Observations made on the celestial sphere



Scale-discretised wavelets on the sphere

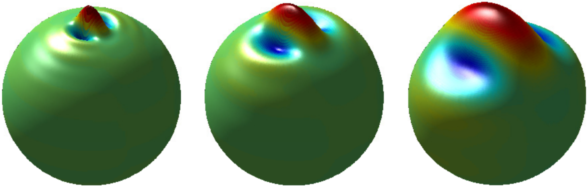
Transforms

- Spin scale-discretised wavelet transform is given by the projection onto each wavelet (Wiaux, McEwen *et al.* 2008, McEwen *et al.* 2013, McEwen *et al.* 2015):

$$\underbrace{\boxed{W^{{}_s\Psi^j}(\rho) = \langle {}_sf,\, \mathcal{R}_\rho\, {}_s\Psi^j \rangle}}_{\text{projection}} = \int_{\mathbb{S}^2} \mathrm{d}\Omega(\omega)\, {}_sf(\omega)\, (\mathcal{R}_\rho\, {}_s\Psi^j)^*(\omega)\ .$$

- Original function may be recovered exactly in practice from wavelet coefficients:

$${}_sf(\omega) = \underbrace{\boxed{\sum_{j=0}^{J}}}_{\text{finite sum}} \underbrace{\boxed{\int_{\mathrm{SO}(3)} \mathrm{d}\varrho(\rho)\, W^{{}_s\Psi^j}(\rho)\, (\mathcal{R}_\rho\, {}_s\Psi^j)(\omega)\ .}}_{\text{wavelet contribution}}$$



(a) $j = 4$ (b) $j = 3$ (c) $j = 2$

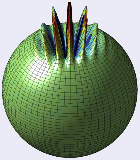
Figure: Scale-discretised wavelets on the sphere.

Scale-discretised wavelets on the sphere

Fast algorithms and codes

- Fast algorithms essential (McEwen, Leistedt *et al.* 2015, Leistedt, McEwen *et al.* 2013, McEwen *et al.* 2013, Leistedt McEwen *et al.* 2007, Wiaux, McEwen & Vielva 2007, Wiaux *et al.* 2005, Wandelt & Gorski 2001, Risbo 1996)

FastCSWT code `http://www.fastcswt.org`

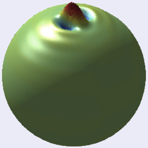


Fast directional continuous spherical wavelet transform algorithms

McEwen *et al.* (2007)

- Fortran
- Supports directional and steerable wavelets

S2DW code `http://www.s2dw.org`



Exact reconstruction with directional wavelets on the sphere

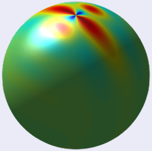
Wiaux, McEwen, Vandergheynst, Blanc (2008)

- Fortran
- Parallelised
- Supports directional and steerable wavelets
- Supports inversion

Scale-discretised wavelets on the sphere

Fast algorithms and codes

S2LET code http://www.s2let.org

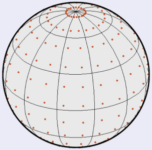


S2LET: Fast wavelet analysis on the sphere

McEwen, Leistedt, Büttner, Peiris & Wiaux (2015), Leistedt, McEwen, *et al.* (2012)

- C, Matlab, IDL, Python
- Supports directional and steerable wavelets, ridgelets and curvelets
- Supports inversion
- Supports spin
- Faster algorithms

SO3 code http://www.sothree.org



SO3: Fast Wigner transforms on the rotation group

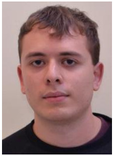
McEwen, Büttner, Leistedt, Peiris & Wiaux (2015)

- C, Matlab, Python
- Fast and exact Fourier transforms on the rotation group $\mathrm{SO}(3)$

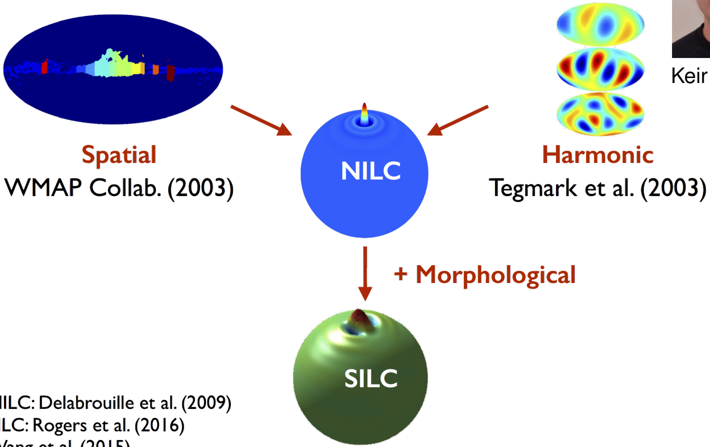
Planck component separation

SILC

- SILC: Blind Planck component separation via Scale-discretised, directional wavelet Internal Linear Combination (Rogers, Peiris, Leistedt, McEwen & Pontzen 2016)



Keir Rogers

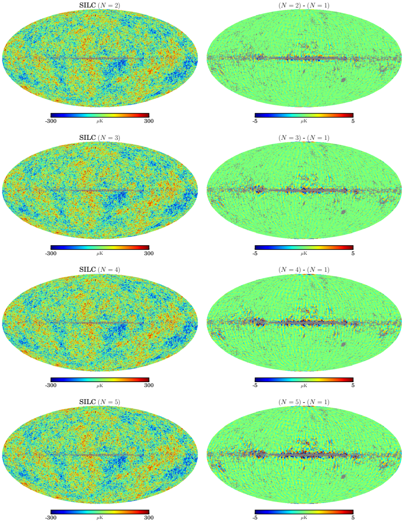


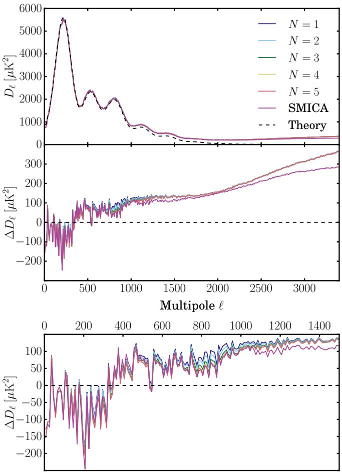
NILC: Delabrouille et al. (2009)
SILC: Rogers et al. (2016)
Wang et al. (2015)

Planck component separation

SILC

- SILC (R1) maps available for download: http://www.silc-cmb.org





E/B separation
Exploiting scale-discretised wavelets

- E/B separation with spin directional wavelets for CMB polarisation and cosmic shear (Leistedt, McEwen, Büttner & Peiris, in prep.)

Boris Leistedt

Algorithm to recover E/B signals using scale-discretised wavelets

1. Compute spin wavelet transform of $Q + \mathrm{i}U$:

$$(Q+\mathrm{i}U)(\omega) \xrightarrow[\text{S2LET}]{\text{Spin wavelet transform}} W^{{}_2\Psi^j}_{Q+\mathrm{i}U}(\rho)$$

2. Account for mask in harmonic and spatial domains simultaneously:

$$W^{{}_2\Psi^j}_{Q+\mathrm{i}U}(\rho) \xrightarrow{\text{Mitigate mask}} \widehat{W}^{{}_2\Psi^j}_{Q+\mathrm{i}U}(\rho)$$

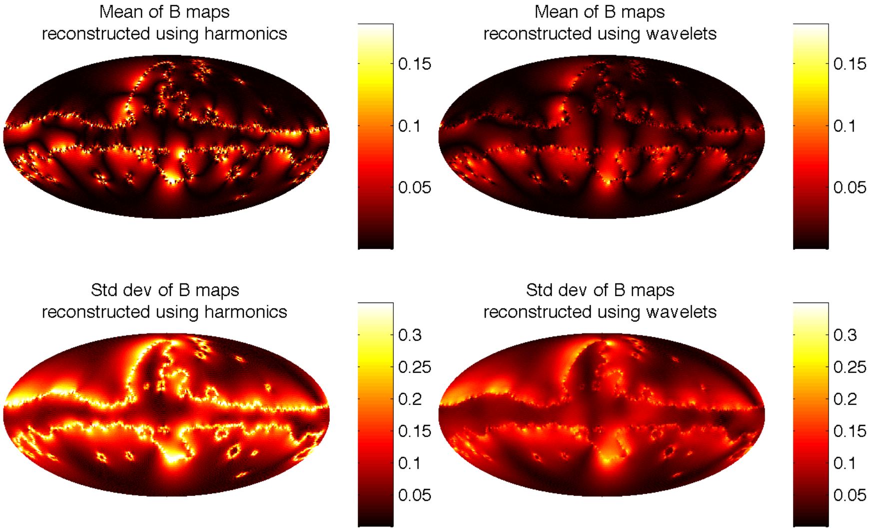
3. Construct E/B maps:

(a) $$W^{{}_0\tilde{\Psi}^j}_{\tilde{E}}(\rho) = -\mathrm{Re}\left[\widehat{W}^{{}_2\Psi^j}_{Q+\mathrm{i}U}(\rho)\right] \xrightarrow[\text{S2LET}]{\text{Inverse scalar wavelet transform}} \tilde{E}(\omega)$$

(b) $$W^{{}_0\tilde{\Psi}^j}_{\tilde{B}}(\rho) = -\mathrm{Im}\left[\widehat{W}^{{}_2\Psi^j}_{Q+\mathrm{i}U}(\rho)\right] \xrightarrow[\text{S2LET}]{\text{Inverse scalar wavelet transform}} \tilde{B}(\omega)$$

E/B separation

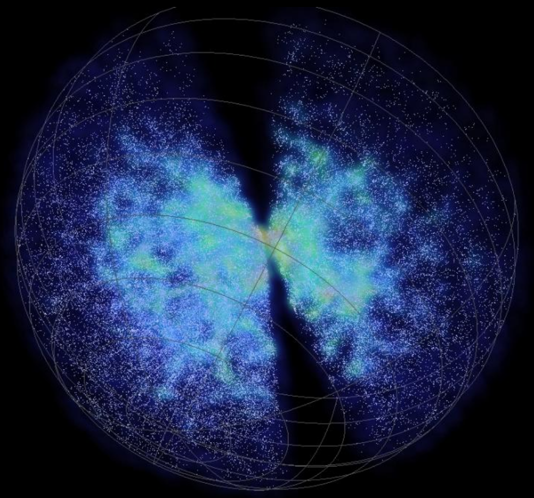
Preliminary results



Outline

1. Big-data in astronomy and astrophysics
2. Illustrative analyses
   - Planck
   - Euclid
   - LSST
   - SKA
3. Concluding remarks

LSS on the 3D ball



Credit: SDSS

Fourier-LAGuerre wavelets (flaglets) on the ball

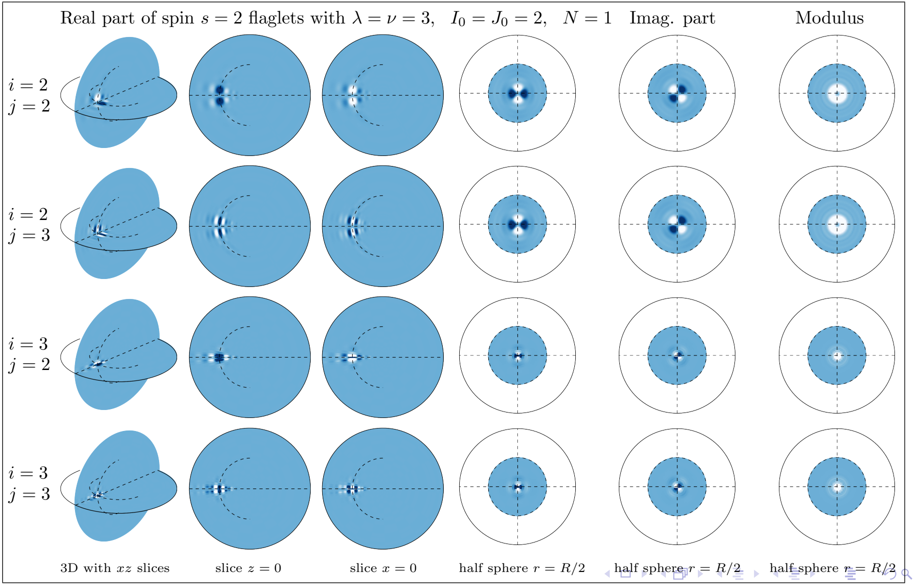
- Fourier-Laguerre wavelet (flaglet) transform is given by the projection onto each wavelet (Leistedt & McEwen 2012):

$$\underbrace{\boxed{W^{{}_s\Psi^{jj'}}(r,\rho) = \langle {}_sf,\ \mathcal{T}_{(r,\rho)}\ {}_s\Psi^{jj'}\rangle}}_{\text{projection}} = \int_{\mathbb{B}^3} \mathrm{d}^3\boldsymbol{r}\ {}_sf(\boldsymbol{r})(\mathcal{T}_{(r,\rho)}\ {}_s\Psi^{jj'})^*(\boldsymbol{r})\ .$$

- Original function may be recovered exactly in practice from wavelet coefficients:

$${}_sf(\boldsymbol{r}) = \underbrace{\boxed{\sum_{jj'}}}_{\text{finite sum}} \underbrace{\boxed{\int_{\mathrm{SO}(3)} \mathrm{d}\varrho(\rho) \int_{\mathbb{R}+} \mathrm{d}r\ W^{{}_s\Psi^{jj'}}(r,\rho)(\mathcal{T}_{(r,\rho)}\ {}_s\Psi^{jj'})(\boldsymbol{r})\ .}}_{\text{wavelet contribution}}$$

Fourier-LAGuerre wavelets (flaglets) on the ball



3D weak lensing

- 3D weak lensing with spin wavelets on the ball (Leistedt, McEwen, Kitching, Peiris 2015).
- Wavelet transform of 3D cosmic shear:

$$\boxed{W^{{}_2\Psi^{ij}}_{{}_2\gamma}(\boldsymbol{n}, r) = ({}_2\gamma \odot {}_2\Psi^{ij})(\boldsymbol{n}, r)}$$

- Wavelet covariance:

$$\boxed{C^{ij,i'j'}(\boldsymbol{n}, \boldsymbol{n}', r, r') = \langle\, W^{{}_2\Psi^{ij}}_{{}_2\gamma}(\boldsymbol{n}, r)\, W^{{}_2\Psi^{i'j'}\,*}_{{}_2\gamma}(\boldsymbol{n}', r')\,\rangle}$$

compute from data

- Theory wavelet covariance:

$$\boxed{C^{ij,i'j'}(\boldsymbol{n}\cdot\boldsymbol{n}', r, r') = \frac{2}{\pi}\sum_\ell \frac{(N_{\ell,2})^2}{4}\int_{\mathbb{R}+}\mathrm{d}k k^2\int_{\mathbb{R}+}\mathrm{d}k' k'^2\, C^{\phi\phi}_\ell(k,k')\, P_\ell(\boldsymbol{n}\cdot\boldsymbol{n}')\, {}_2\mathcal{H}^{ij}_\ell(k,r)\, {}_2\mathcal{H}^{i'j'\,*}_\ell(k',r')}$$

compute from theory

- Simultaneous spatial and scale representation (can handle complicated sky coverage and filter unreliable harmonic modes).

Outline

1. Big-data in astronomy and astrophysics

2. Illustrative analyses
   - Planck
   - Euclid
   - LSST
   - SKA

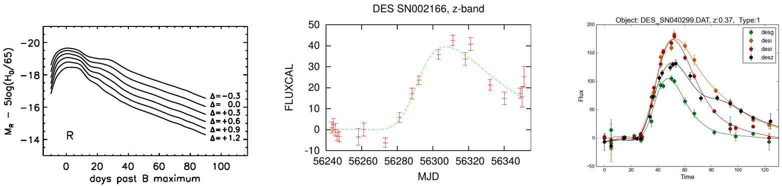
3. Concluding remarks

Photometric supernova classification

Machine learning

- Photometric supernova classification by machine learning (Lochner, McEwen, Peiris & Lahav, in prep.)
- Go beyond single techniques to study classes.
- Understand physical requirements (*e.g.* representative training, redshift).

Michelle Lochner

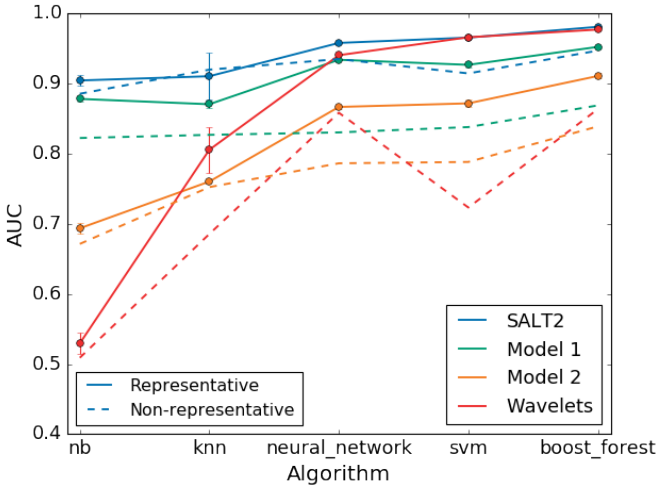


(a) Templates (b) Generic parameterisations (c) Wavelets (non-parametric)

Figure: Feature selection classes (in order of increasing model independence)

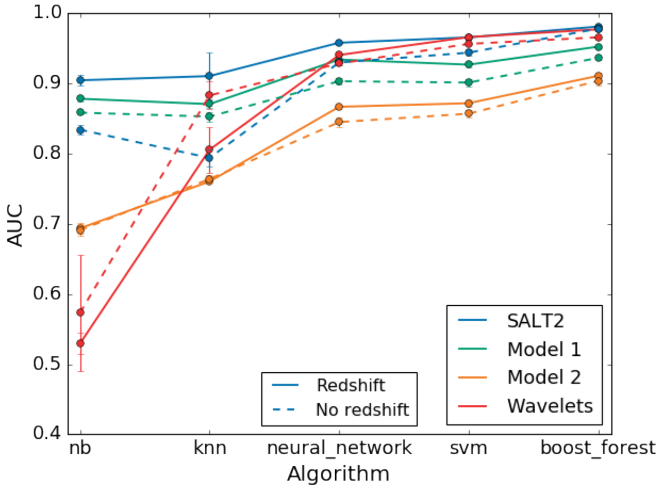
Photometric supernova classification

Importance of representative training data



Photometric supernova classification

Importance of redshift



Outline

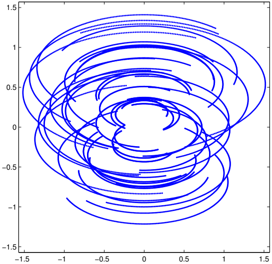
1. Big-data in astronomy and astrophysics

2. Illustrative analyses
   - Planck
   - Euclid
   - LSST
   - SKA

3. Concluding remarks

Radio interferometric telescopes acquire “Fourier” measurements

“Fourier” Measurements



Compressive sensing

- Developed by Candes *et al.* 2006 and Donoho 2006 (and others).
- Although many underlying ideas around for a long time.
- Exploits the sparsity of natural signals.
- Active area of research with many new developments.

SARA for radio interferometric imaging

Algorithm

- Sparsity averaging reweighted analysis (SARA) for RI imaging (Carrillo, McEwen & Wiaux 2012)

- Consider a dictionary composed of a concatenation of orthonormal bases, i.e.

$$\boxed{\Psi = \frac{1}{\sqrt{q}}[\Psi_1, \Psi_2, \dots, \Psi_q],}$$

  thus $\Psi \in \mathbb{R}^{N \times D}$ with $D = qN$.

- We consider the following bases: Dirac (*i.e.* pixel basis); Haar wavelets (promotes gradient sparsity); Daubechies wavelet bases two to eight.
  $\Rightarrow$ concatenation of 9 bases

- Promote average sparsity by solving the reweighted ℓ_1 analysis problem:

$$\boxed{\min_{\bar{\boldsymbol{x}} \in \mathbb{R}^N} \|W\Psi^T \bar{\boldsymbol{x}}\|_1 \quad \text{subject to} \quad \|\boldsymbol{y} - \Phi\bar{\boldsymbol{x}}\|_2 \leq \epsilon \quad \text{and} \quad \bar{\boldsymbol{x}} \geq 0\,,}$$

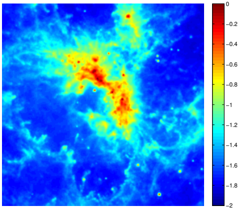
SARA

  where $W \in \mathbb{R}^{D \times D}$ is a diagonal matrix with positive weights.

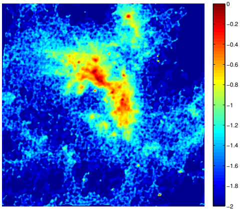
- Solve a sequence of reweighted ℓ_1 problems using the solution of the previous problem as the inverse weights $\rightarrow$ approximate the ℓ_0 problem.

SARA for radio interferometric imaging

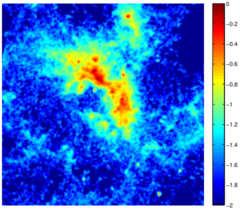
Results on simulations



(a) Original



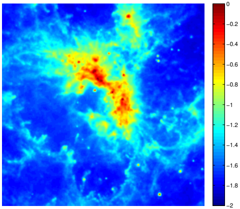
(b) “CLEAN” (SNR=16.67 dB)



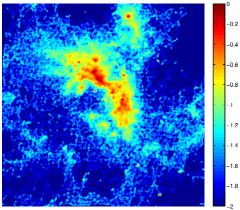
(c) “MS-CLEAN” (SNR=17.87 dB)

SARA for radio interferometric imaging

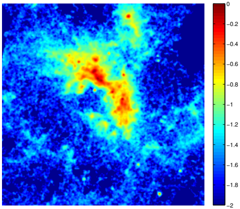
Results on simulations



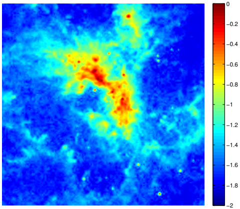
(a) Original



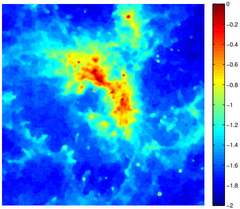
(b) “CLEAN” (SNR=16.67 dB)



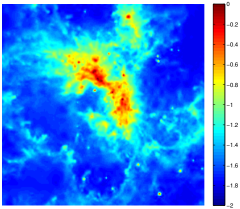
(c) “MS-CLEAN” (SNR=17.87 dB)



(d) BPDb8 (SNR=24.53 dB)



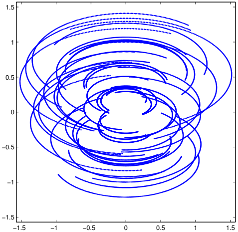
(e) TV (SNR=26.47 dB)



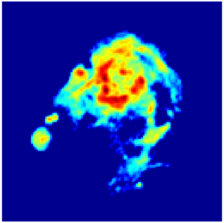
(f) SARA (SNR=29.08 dB)

Supporting continuous visibilities

Results on simulations



(a) Coverage

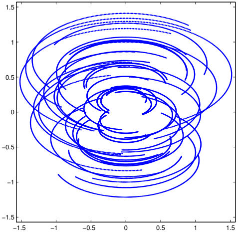


(b) M31 (ground truth)

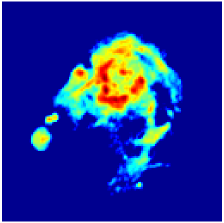
Figure: Reconstructed images from continuous visibilities.

Supporting continuous visibilities

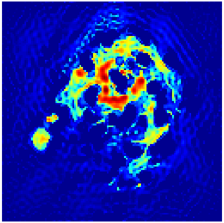
Results on simulations



(a) Coverage



(b) M31 (ground truth)

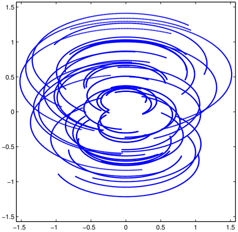


(c) "CLEAN" (SNR= 8.2dB)

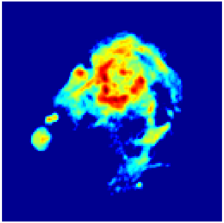
Figure: Reconstructed images from continuous visibilities.

Supporting continuous visibilities

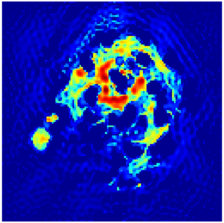
Results on simulations



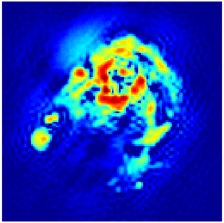
(a) Coverage



(b) M31 (ground truth)



(c) "CLEAN" (SNR= 8.2dB)

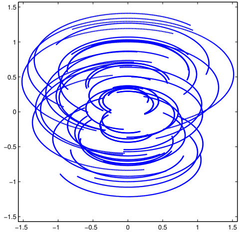


(d) "MS-CLEAN" (SNR= 11.1dB)

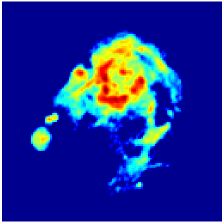
Figure: Reconstructed images from continuous visibilities.

Supporting continuous visibilities

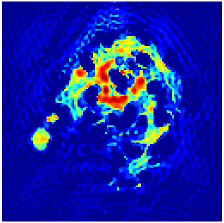
Results on simulations



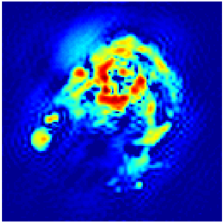
(a) Coverage



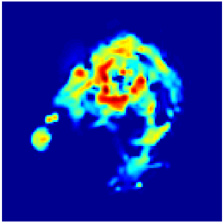
(b) M31 (ground truth)



(c) “CLEAN” (SNR= 8.2dB)



(d) “MS-CLEAN” (SNR= 11.1dB)



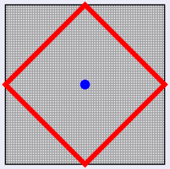
(e) SARA (SNR= 13.4dB)

Figure: Reconstructed images from continuous visibilities.

Distributed algorithms and codes

- Distributed storage and computation (Onose *et al.* 2016) by divide-and-conquer and sub-sampling techniques

SOPT code http://basp-group.github.io/sopt/



Sparse OPTimisation
Carrillo, McEwen, Wiaux

SOPT is an open-source code that provides functionality to perform sparse optimisation using state-of-the-art convex optimisation algorithms.

PURIFY code http://basp-group.github.io/purify/

Next-generation radio interferometric imaging
Carrillo, McEwen, Wiaux

PURIFY is an open-source code that provides functionality to perform radio interferometric imaging, leveraging recent developments in the field of compressive sensing and convex optimisation.

Concluding remarks

- Increasingly inter-disciplinary, drawing on statistics, applied mathematics, computer science, information engineernig, . . .
- Increasingly intra-disciplinary (*e.g.* Planck, Euclid, LSST, SKA, . . .)
- Many methodological synergies

Concluding remarks

How can we exploit synergies?

1. Open (unencumbered) data and open code
2. Develop best practices (*e.g.* code development, general codes, reproducible/replicable research, blinded analysis)
3. Explore HPC synergies (*e.g.* Dirac, Archer, Hartree, Google, Amazon, . . .)
4. Develop appropriate career progression routes
5. Go beyond individual techniques to understand properties of classes of approach
6. Develop common language
7. Promote inter- and intra-disciplinary collaboration and communication, *e.g.* Alan Turing Institute (ATI), workshops (*e.g.* BASP conference), Hackathons, . . .
8. . . .

Concluding remarks

How can we exploit synergies?

1. Open (unencumbered) data and open code
2. Develop best practices (*e.g.* code development, general codes, reproducible/replicable research, blinded analysis)
3. Explore HPC synergies (*e.g.* Dirac, Archer, Hartree, Google, Amazon, . . .)
4. Develop appropriate career progression routes
5. Go beyond individual techniques to understand properties of classes of approach
6. Develop common language
7. Promote inter- and intra-disciplinary collaboration and communication, *e.g.* Alan Turing Institute (ATI), workshops (*e.g.* BASP conference), Hackathons, . . .
8. . . .

Concluding remarks

How can we exploit synergies?

1. Open (unencumbered) data and open code
2. Develop best practices (*e.g.* code development, general codes, reproducible/replicable research, blinded analysis)
3. Explore HPC synergies (*e.g.* Dirac, Archer, Hartree, Google, Amazon, . . .)
4. Develop appropriate career progression routes
5. Go beyond individual techniques to understand properties of classes of approach
6. Develop common language
7. Promote inter- and intra-disciplinary collaboration and communication, *e.g.* Alan Turing Institute (ATI), workshops (*e.g.* BASP conference), Hackathons, . . .
8. . . .

Concluding remarks

How can we exploit synergies?

1. Open (unencumbered) data and open code
2. Develop best practices (*e.g.* code development, general codes, reproducible/replicable research, blinded analysis)
3. Explore HPC synergies (*e.g.* Dirac, Archer, Hartree, Google, Amazon, . . .)
4. Develop appropriate career progression routes
5. Go beyond individual techniques to understand properties of classes of approach
6. Develop common language
7. Promote inter- and intra-disciplinary collaboration and communication, *e.g.* Alan Turing Institute (ATI), workshops (*e.g.* BASP conference), Hackathons, . . .
8. . . .

Concluding remarks

How can we exploit synergies?

1. Open (unencumbered) data and open code
2. Develop best practices (*e.g.* code development, general codes, reproducible/replicable research, blinded analysis)
3. Explore HPC synergies (*e.g.* Dirac, Archer, Hartree, Google, Amazon, . . .)
4. Develop appropriate career progression routes
5. Go beyond individual techniques to understand properties of classes of approach
6. Develop common language
7. Promote inter- and intra-disciplinary collaboration and communication, *e.g.* Alan Turing Institute (ATI), workshops (*e.g.* BASP conference), Hackathons, . . .
8. . . .

Concluding remarks

How can we exploit synergies?

1. Open (unencumbered) data and open code
2. Develop best practices (*e.g.* code development, general codes, reproducible/replicable research, blinded analysis)
3. Explore HPC synergies (*e.g.* Dirac, Archer, Hartree, Google, Amazon, . . .)
4. Develop appropriate career progression routes
5. Go beyond individual techniques to understand properties of classes of approach
6. Develop common language
7. Promote inter- and intra-disciplinary collaboration and communication, *e.g.* Alan Turing Institute (ATI), workshops (*e.g.* BASP conference), Hackathons, . . .
8. . . .

Concluding remarks

How can we exploit synergies?

1. Open (unencumbered) data and open code
2. Develop best practices (*e.g.* code development, general codes, reproducible/replicable research, blinded analysis)
3. Explore HPC synergies (*e.g.* Dirac, Archer, Hartree, Google, Amazon, . . .)
4. Develop appropriate career progression routes
5. Go beyond individual techniques to understand properties of classes of approach
6. Develop common language
7. Promote inter- and intra-disciplinary collaboration and communication, *e.g.* Alan Turing Institute (ATI), workshops (*e.g.* BASP conference), Hackathons, . . .
8. . . .