Big Data and Complex Modeling:
Challenges in Astronomy and Solar Physics

David A. van Dyk

Statistics Section
Imperial College London

Los Alamos National Labs, March 2015

Introduction

Massive new data streams are opening up a world of opportunities for data scientists!

1. Astrostatistics: Data quality and quantity lead to more interesting statistical models
2. Data-driven versus Science-driven methods
3. Predictive models versus Descriptive models
4. Tradeoff: computational speed and statistical principles
5. These issues are not unique to astronomy!

Joint work with:

- CHASC International Center for Astrostatistics
  (Includes researchers at Harvard, Univ of California, NASA, Imperial, Crete, etc.)
- Imperial Centre for Inference in Cosmology

Imperial College London

Outline

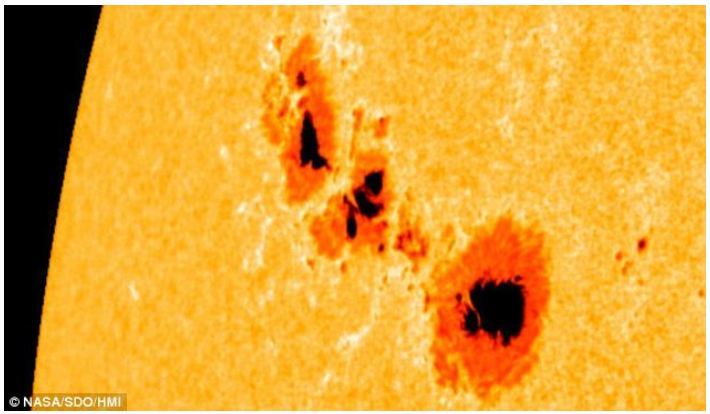
Massive Data Sets and Data Streams

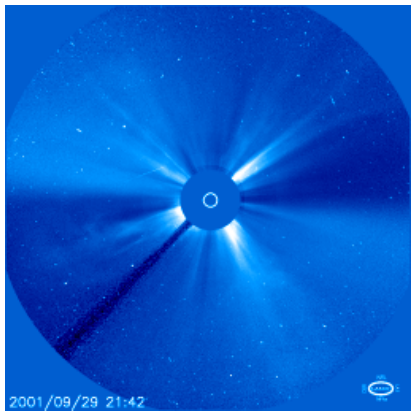
Dramatic increase in the quality and quantity of data:

- massive new surveys: catalogs containing T/PBs of data,
- high resolution spectrography and imaging across the electromagnetic spectrum,
- incredibly detailed movies of dynamic and explosive processes in the solar atmosphere,
- massive number of items and/or features,
- space-based telescopes tailored to specific scientific goals,
- data are *not just massive:* they are rich, deep, & complex.

Massive Challenges for Data Scientists!!

Example I: Thermal Structure in the Solar Corona[1]



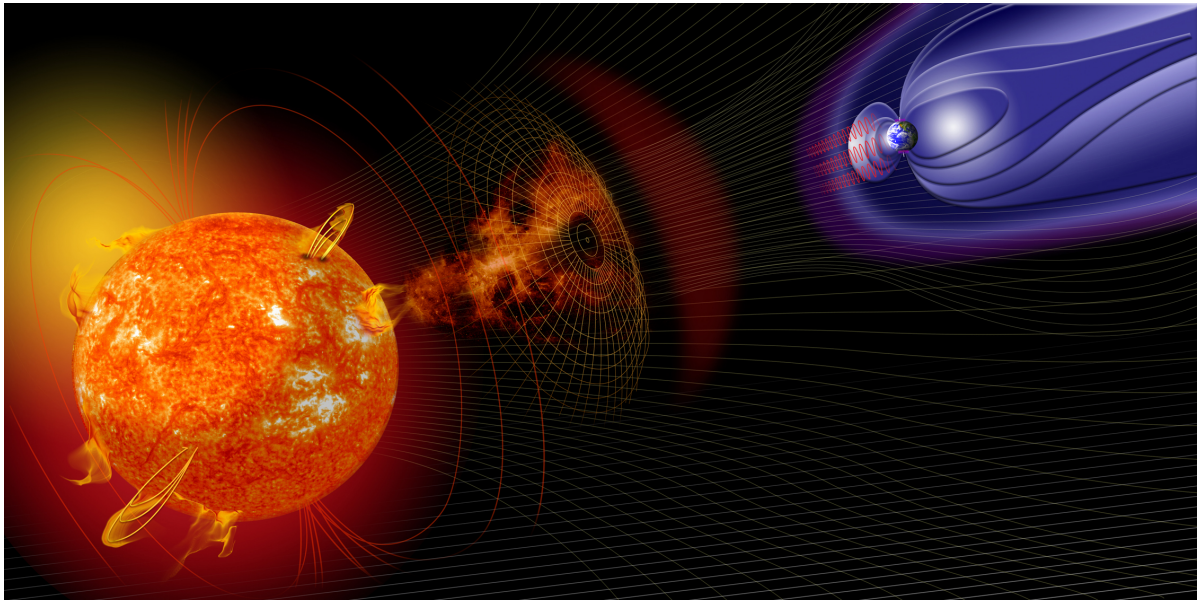


- *Solar Corona*: Highly energetic and violent, characterized by sunspots, solar flares, and coronal mass ejections.
- Solar storms can affect space weather, earth satellites, communication systems, and electric grids.
- *Goal:* Track solar activity with the aim of predicting storms and their effects on Earth.

Imperial College London

[1] N Stein, D Stenning, T Lee, XL Meng, V Kashyap, and CHASC

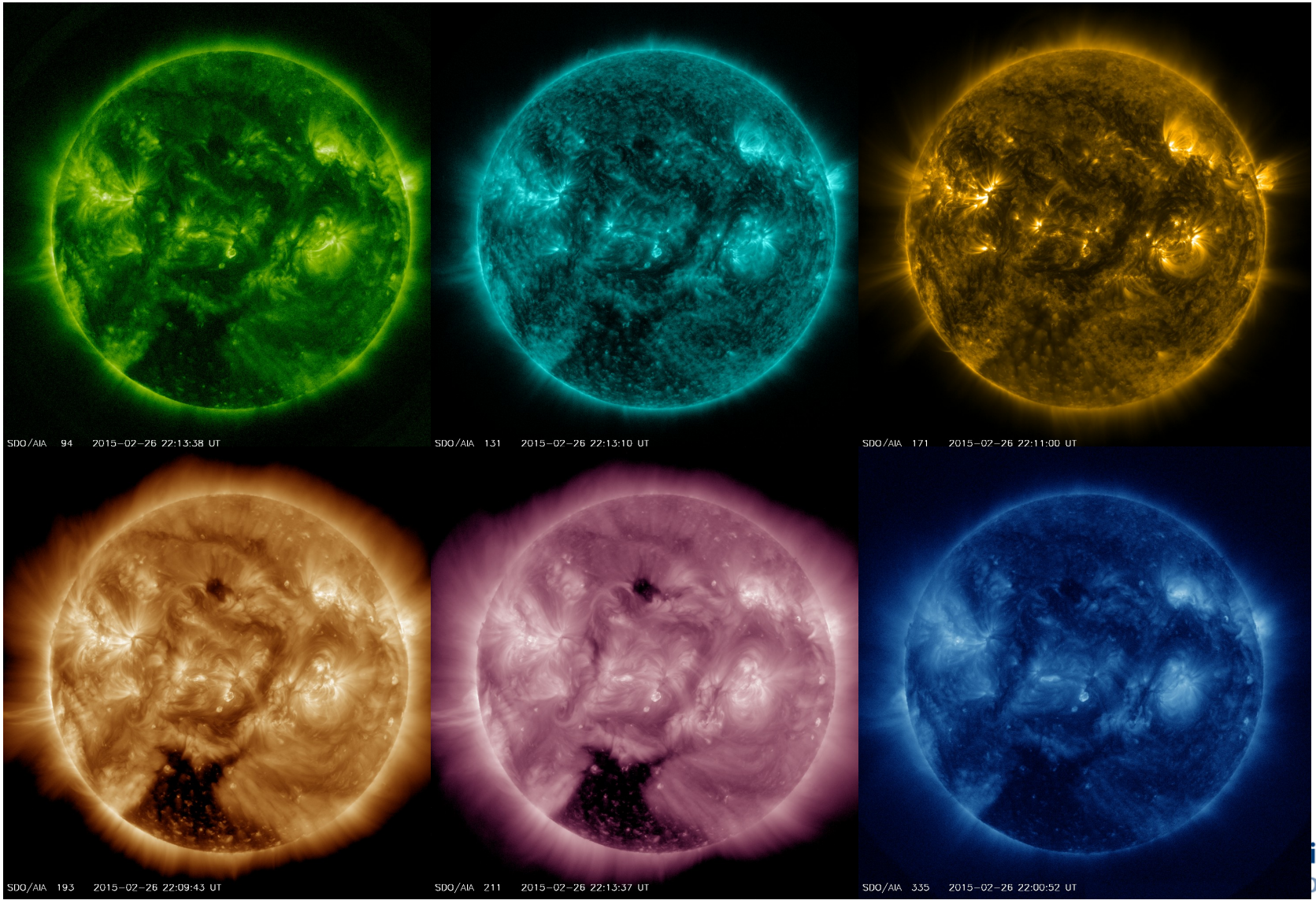
Space Weather Effects



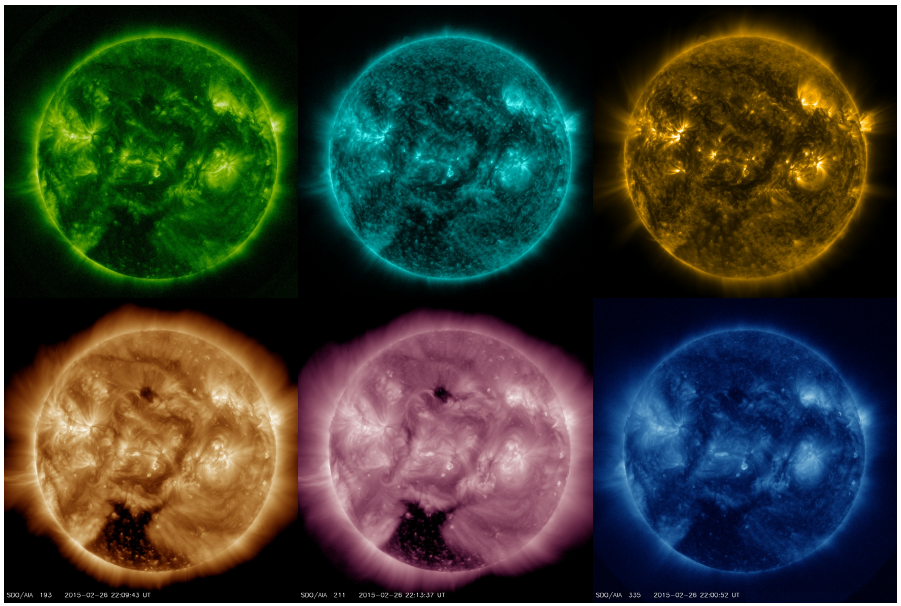
Artist illustration of events on the sun changing the conditions in Near-Earth space.
Image Credit: NASA

- The highly energetic particles released by solar flares and CMEs can impact the Earth's magnetosphere.
- These impacts cause radio interference and can damage satellites and electric power transmission.
- 1859 geomagnetic storm: Worldwide Aurorae, gold miners thought it was morning, telegraph machines failed, shocked operators, threw sparks, even if unplugged.

The Data: Pixel-by-Pixel Spectra



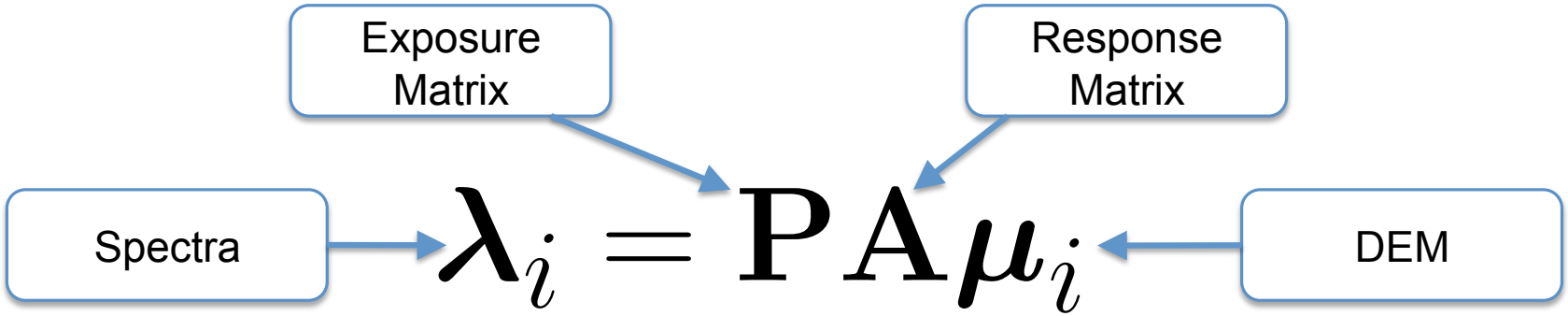
The Data: Pixel-by-Pixel Spectra



The Solar Dynamics Observatory

- NASA satellite launched February 2010
- Massive Data Stream: 1.4 TB/day of compressed data
- High Spatial and Temporal resolution
- Low Spectral resolution
- White light and magnetigram images

The Differential Emission Measure



DEM: *expected emission due to plasma of a given temperature.*
A: *expected spectra of plasmas at each temperature.*
Challenge: *Inversion is ill-posed.*

Normalized Spectra: $\pi_i = \dfrac{\mathbf{PA}\boldsymbol{\mu}_i}{\mathbf{1}^\top \mathbf{PA}\boldsymbol{\mu}_i}$, MLE is trivial.

Goal: *Cluster pixels with similar spectra.*

How Should we Cluster Probability Vectors?

K-means Algorithm:

Assignment: Assign units to clusters by minimizing the Euclidean distance to the centroid.

Update: Compute new centroids by minimizing the total Euclidean distance within each cluster.

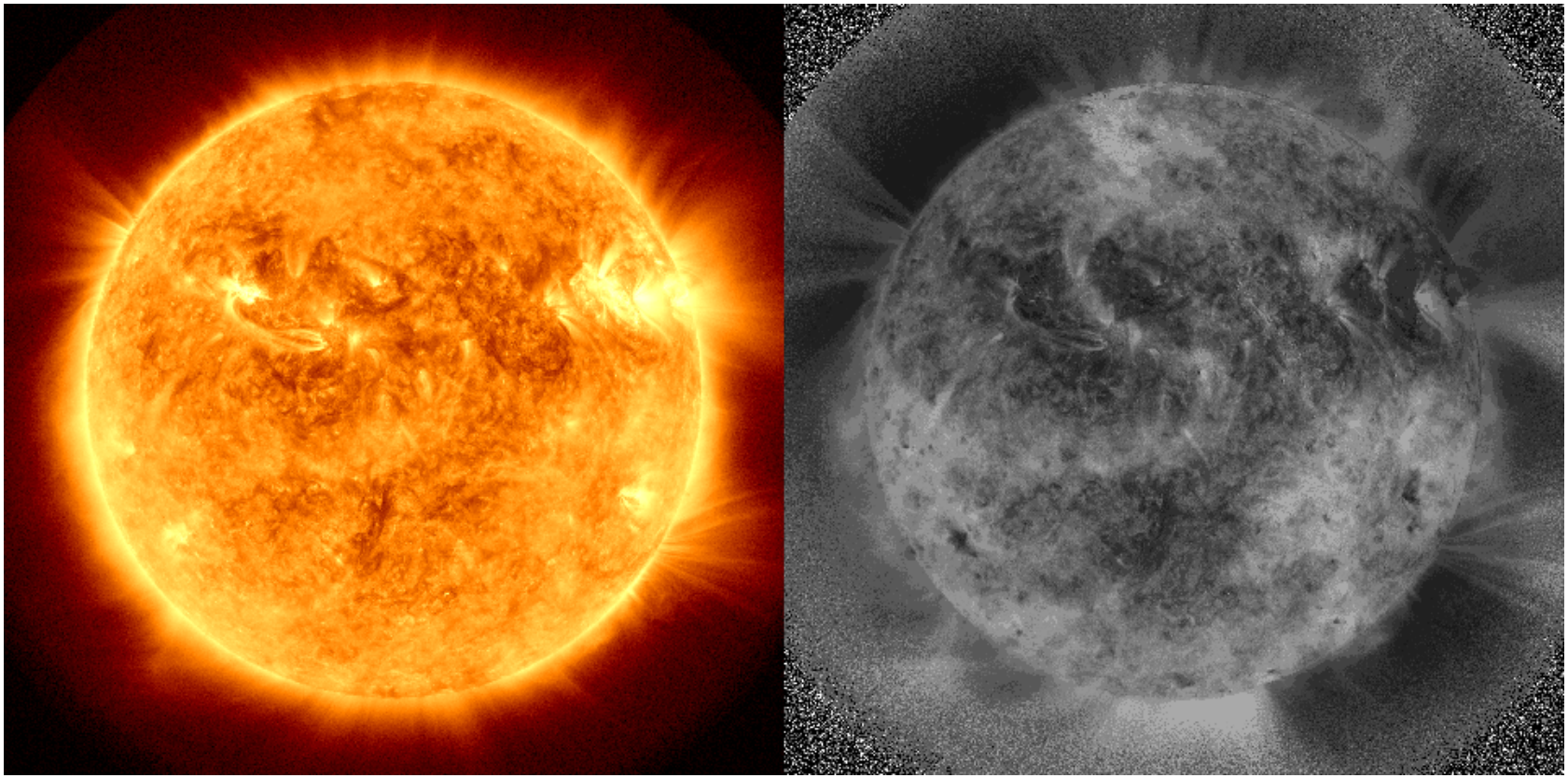
The H-means Algorithm:

- Replace Euclidean distance with **Hellinger Distance**: appropriate for probability vectors.
- Hellinger distance between $\hat{\pi}_i$ and $\boldsymbol{c}_j$:

$$d_H^2(\hat{\pi}_i, \boldsymbol{c}_j) = \frac{1}{2}\sum_k \left(\sqrt{\hat{\pi}_{ik}} - \sqrt{c_{jk}}\right)^2 = 1 - \sum_k \sqrt{\hat{\pi}_{ik} c_{jk}}.$$

- Both steps remain in closed form.

Never Before Seen Structure[2]

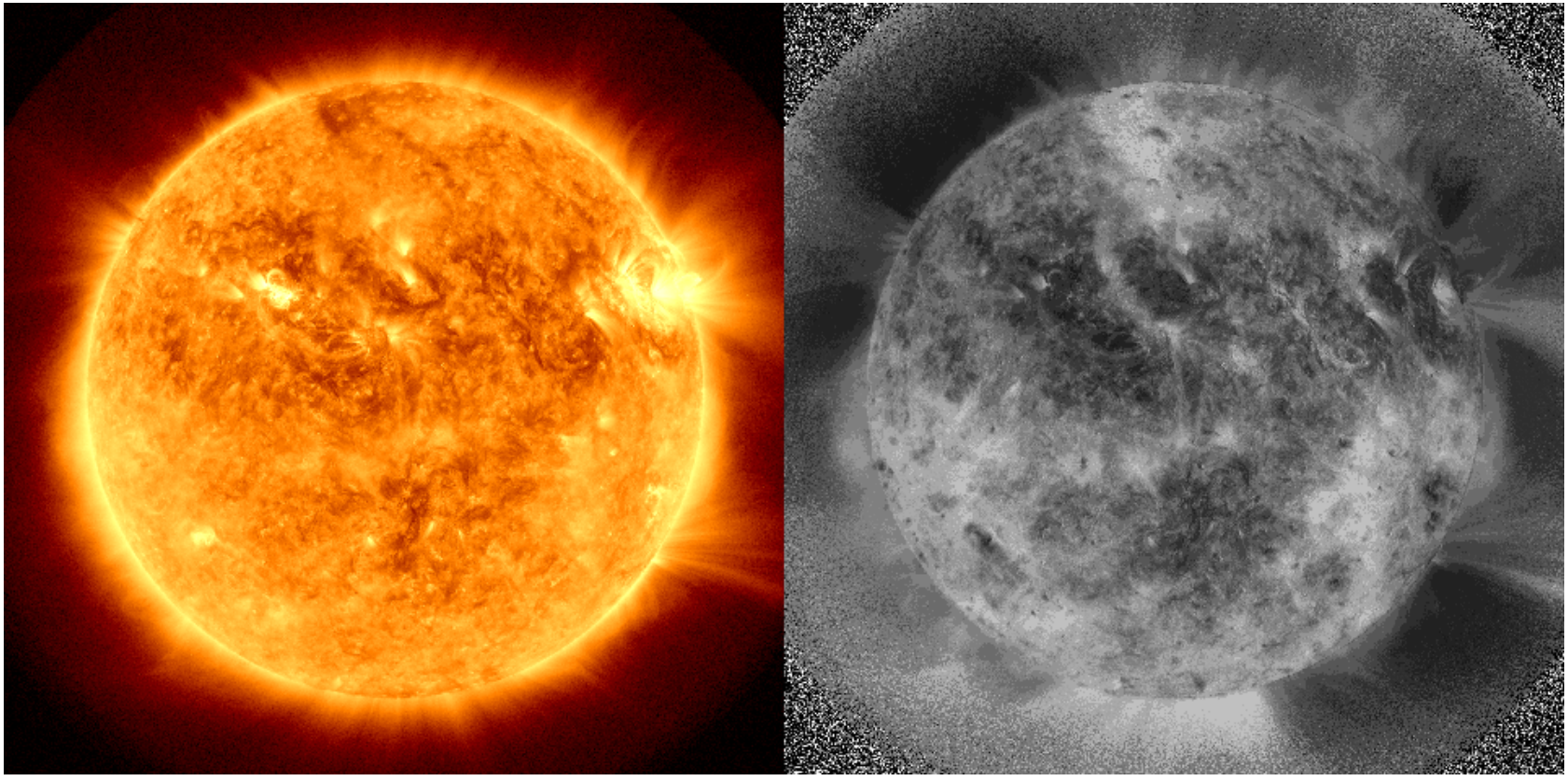


Grey Scale Images of Clusters: 2 Oct 2010 at 05.57

[2] Stein, N., Kashyap, V., Meng, X. L., and van Dyk, D. A. (2012). H-Means Image Segmentation to Identify Solar Thermal Features. *Proceedings of the 19th IEEE International Conference on Image Processing, ICIP 2012*.

Stein, N., Kashyap, V., and van Dyk, D. A. (2015). Clustering Latent Features: A Case Study in Solar Image Segmentation, in preperation

Never Before Seen Structure

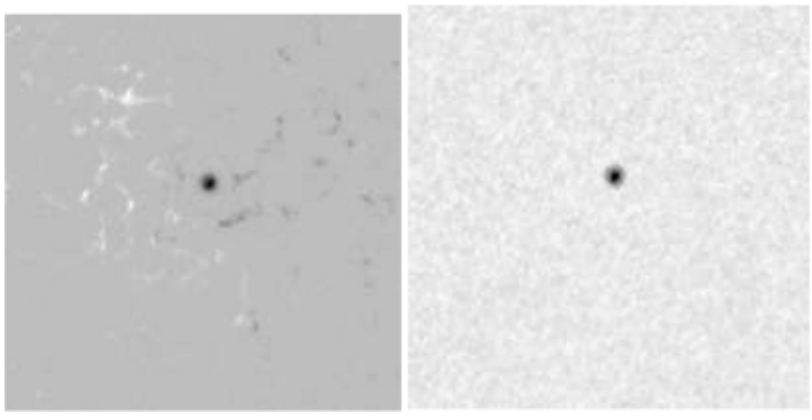


Grey Scale Images of Clusters: 2 Oct 2010 at 18.43

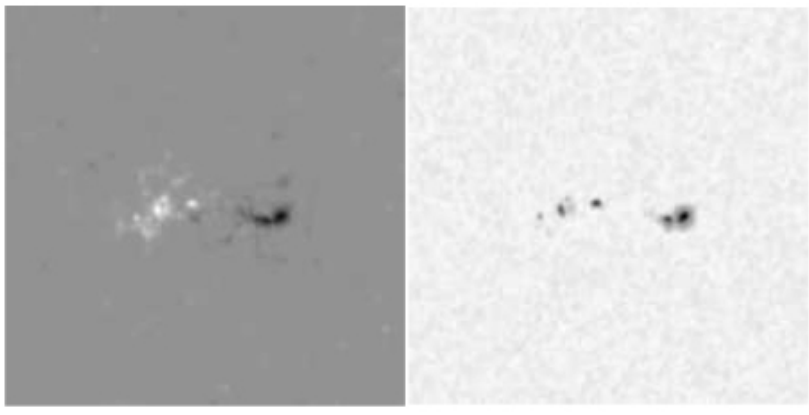
Goal: Model, Track, & Forecast structures.

Mt Wilson Classification of Sun Spots

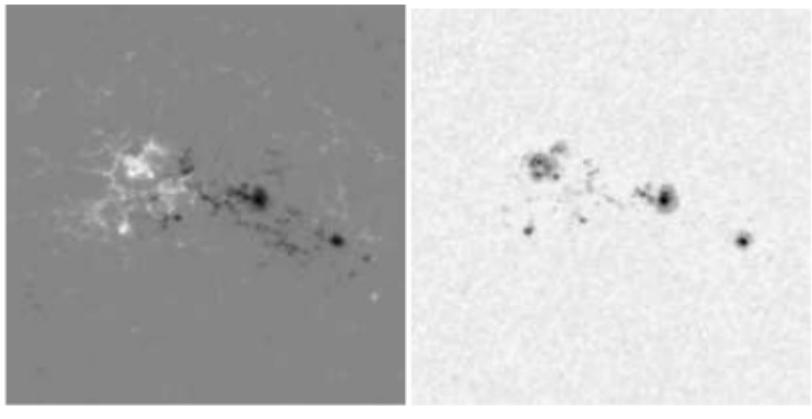
α class



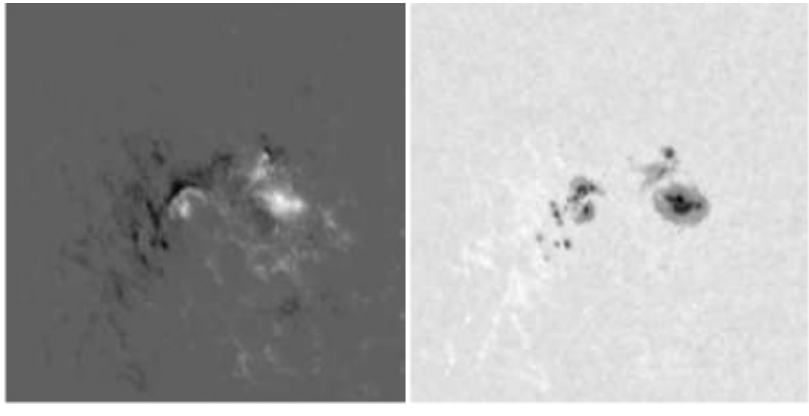
β class



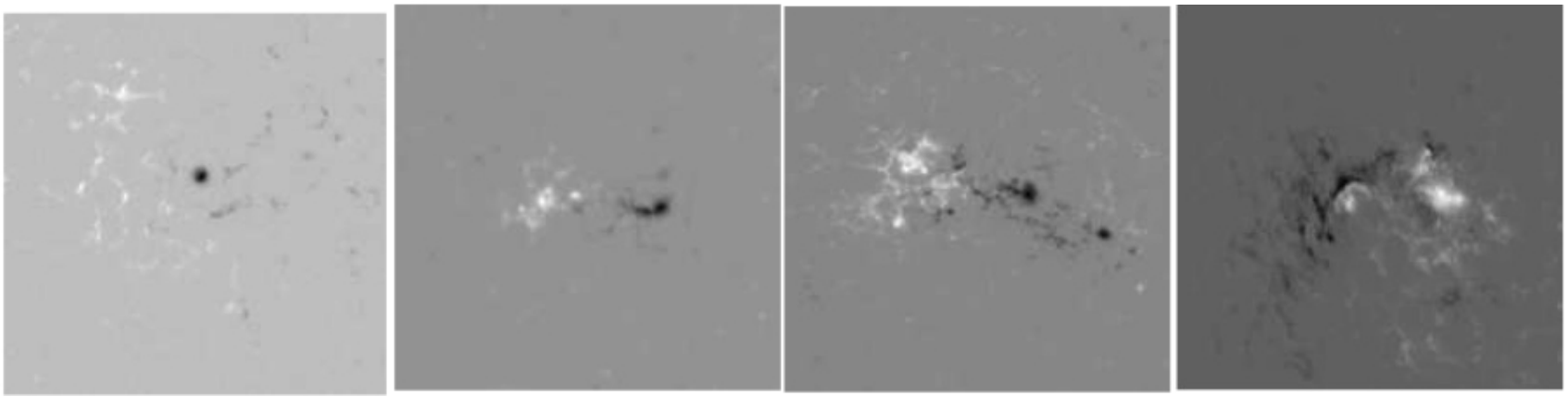
$\beta\gamma$ class



$\beta\gamma\delta$ class



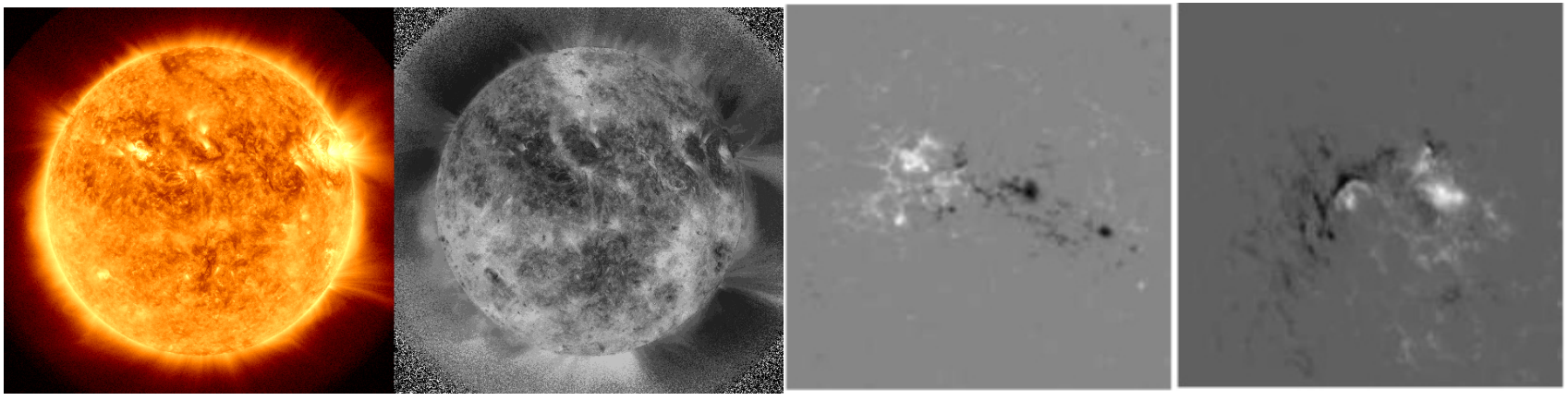
Automatic Classification of Sun Spots[3]



- Classification is predictive for activity in solar corona
- Sunspots are typically classified manually
- Automate classification: **science-driven feature selection**
- Features from mathematical morphology (e.g., area of overlap of polarities, roughness of separating line, etc.)
- Features have physical meaning beyond classification
- **Goal:** Use to model, track, and forecast evolution (movie)

[3] Stenning, D. C., Lee, T. C. M, van Dyk, D. A., Kashyap, V. L., Sandell, J., and Young, C. A. (2013). *Statistical Analysis and Data Mining*, **6**, 329–345.

Big Data Methods in Solar Physics



- Use science-driven models to inform data-driven methods.
  - E.g., cluster spectra for DEM; optimal choice of distance.
  - E.g., tuning feature extraction to Mt. Wilson classification.
- Reduce the complexity of data to understandable features.
  - E.g., summarize images w/ feature for secondary analysis.
- More efficient to split initial and secondary analyses.

Example II: Stellar Evolution[4]

Complex Data and Sophisticated Models

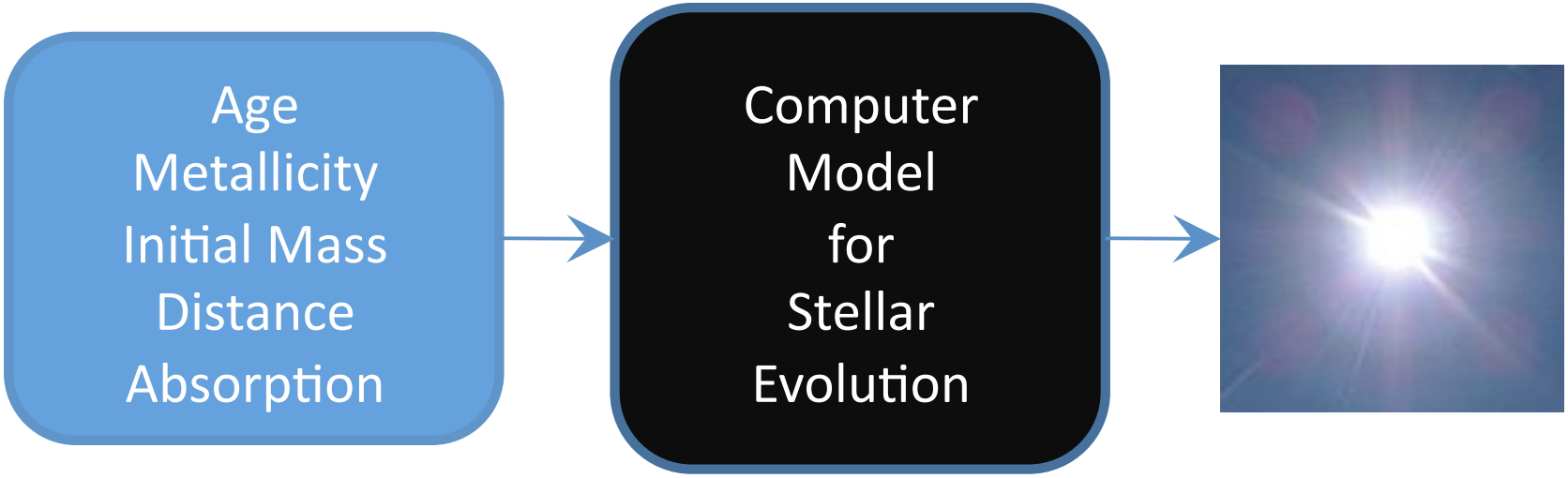
1. Complex computer models and simulations are taking the place of the analytic likelihood function.
2. Sophisticated data allows us to fit such models, but an entirely new set of statistical methods is required.
3. This sort of modeling, computing, and inference is coming to many more areas of Astronomy.
4. I will discuss one example in detail: stellar evolution.

Challenge is acute when complex models are combined with massive data streams.

Imperial College London

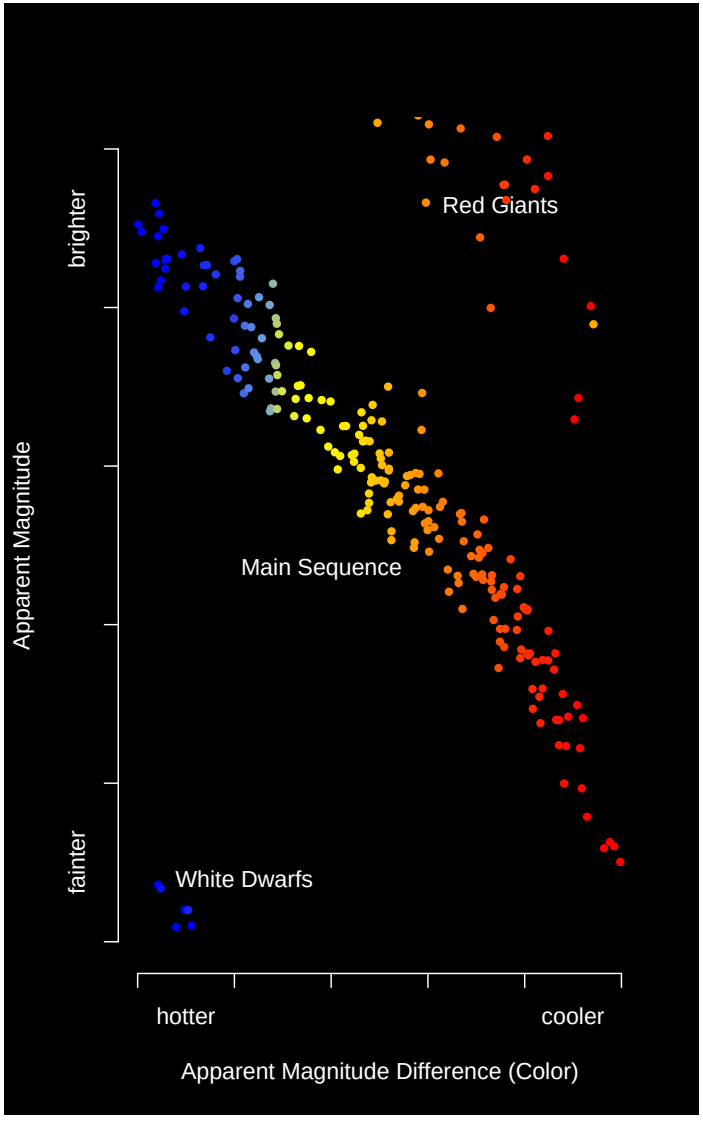
[4] N Stein, S DeGennaro, E Jeffery, W Jefferys, S Si, D Stenning, and T von Hippel

Compter Model for Sun-Like Stellar Evolution



- Computer model predicts how a the spectrum of a sun-like star evolves as a function of input parameters.
- We aim to embed these models into a sophisticated multi-level model for statistical inference.

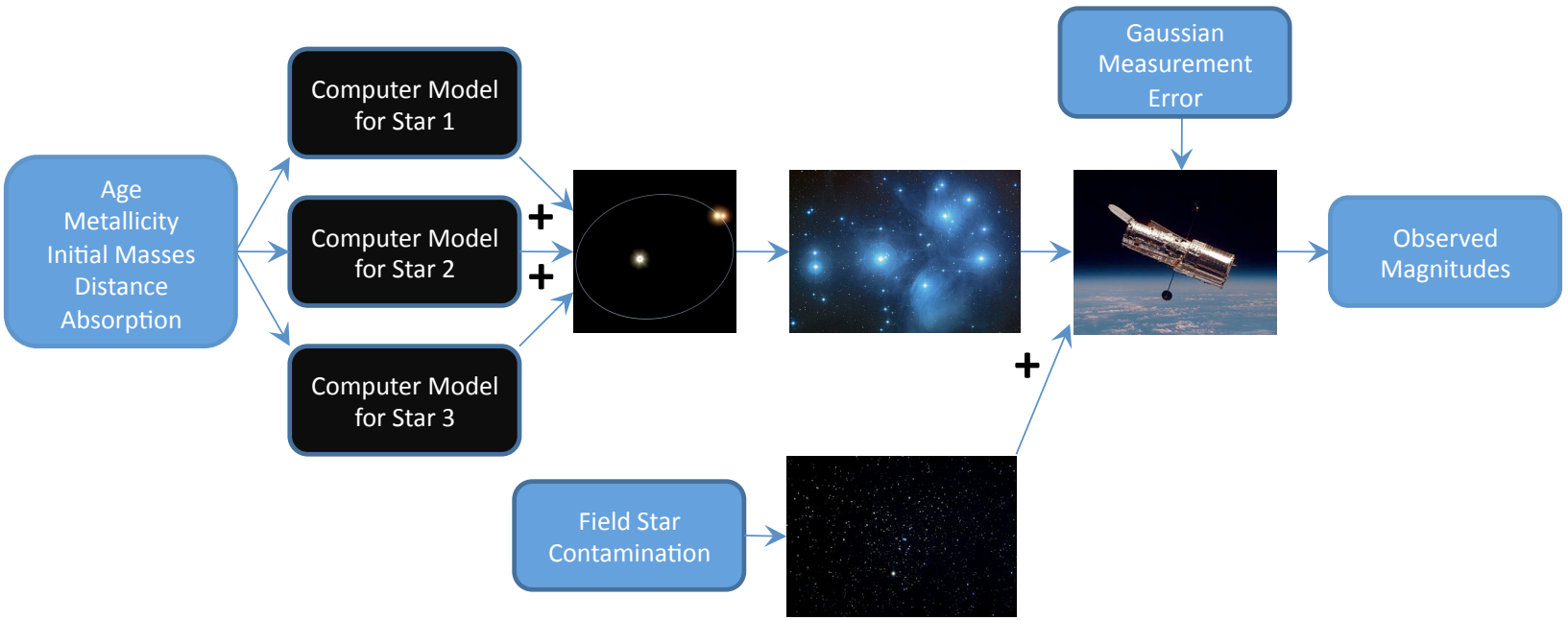
The Data: Color Magnitude Diagrams



Color-Magnitude Diagram

- Plot Magnitude Difference vs. Magnitude.
- Identifies stars at different stages of their lives.
- Evolution of a CMD.
- Facilitates physical intuition as to likely values of parameters.
- “Chi-by-eye” fitting.
- Can we avoid ad hoc methods?

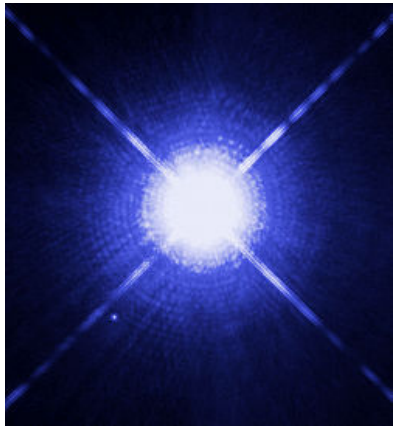
Embedding Computer Model into Statistical Model[5]



- Between 1/3 and 1/2 of "stars" are unresolved binaries.
- Star clusters: same age, metallicity, distance, & absorption.
- Cluster data is contaminated with field stars.
- Data observed with Gaussian measurement errors.

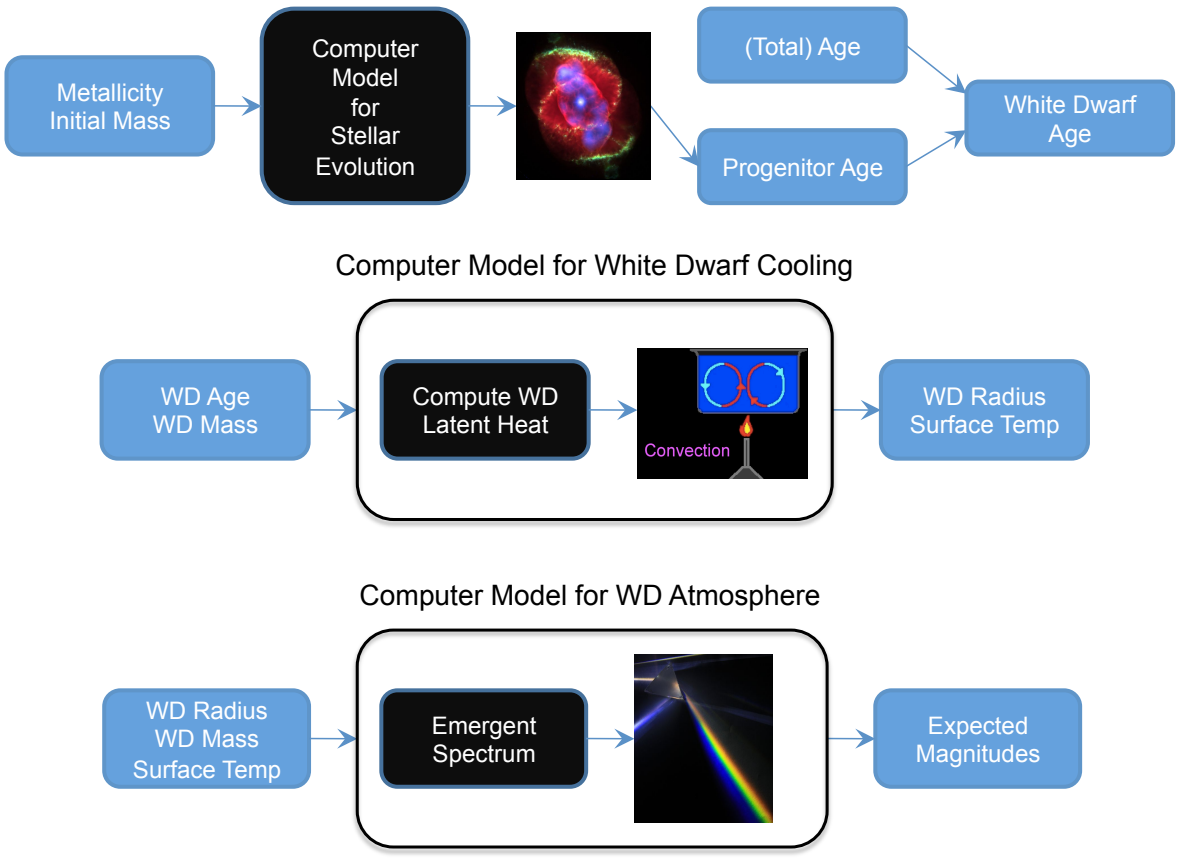
[5] van Dyk, D. A., DeGennaro, S., Stein, N., Jefferys, W. H., and von Hippel, T. (2009). Statistical Analysis of Stellar Evolution. *The Annals of Applied Statistics*, **3**, 117–143.

White Dwarfs Physics



- Sun-like stars are powered by thermal-nuclear reactions.
- White dwarfs are the cooling embers after reactions cease.
- Different physical processes require different models.

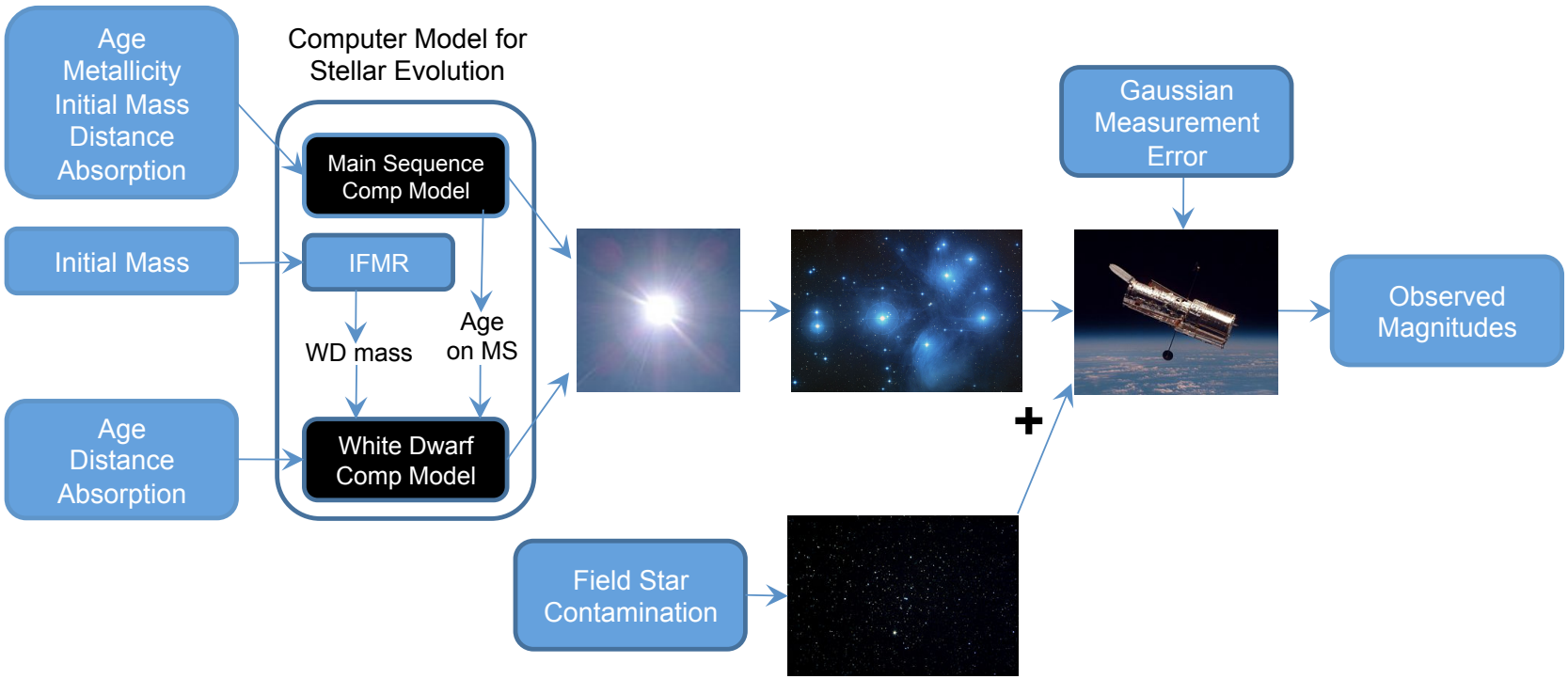
The Missing Link: White Dwarf Mass



- We must model: white dwarf mass $= f$(initial mass).
- Parametric Bridge between Computer Models.

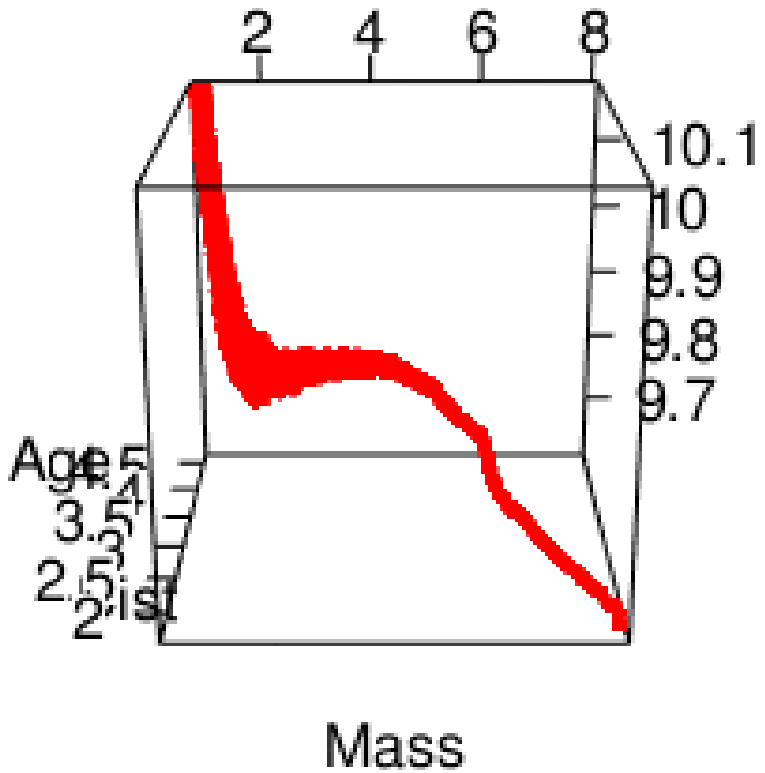
Imperial College London

Opening Up the Black Box: The Final Model[6]



[6] Stein, N. M., van Dyk, D. A., von Hippel, T. , DeGennaro, S., Jeffrey, E. J., and Jefferys, W. H. (2013). Combining Computer Models to Account for Mass Loss in Stellar Evolution. *Statistical Analysis and Data Mining*, **6**, 34–52.

Model Fitting: Complex Posterior Distributions[7]

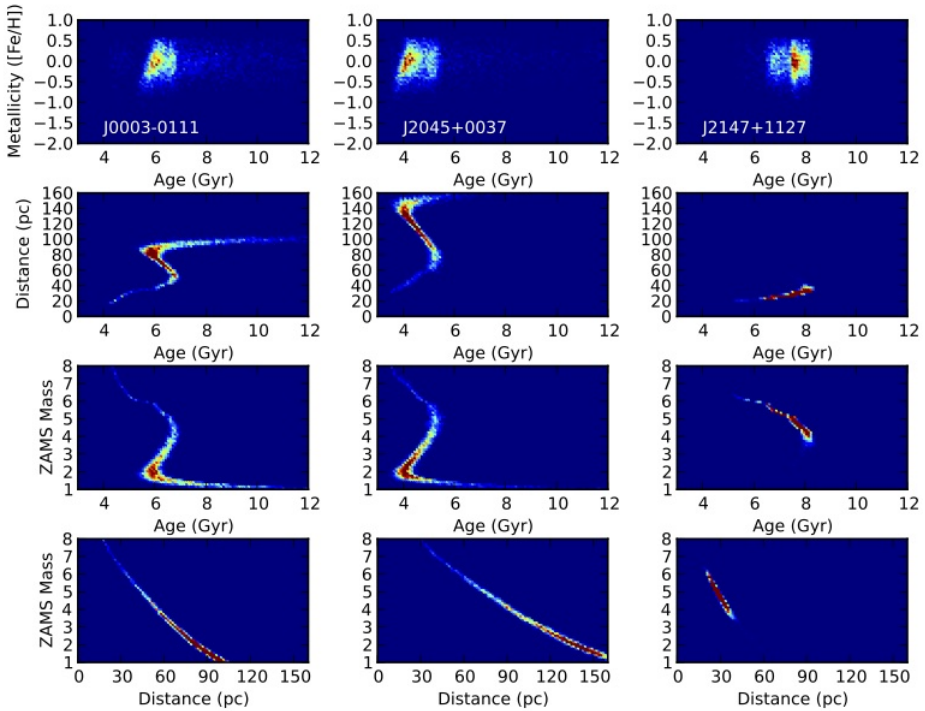


Highly non-linear relationships among stellar parameters.

Imperial College London

[7] O.Malley, E. M., von Hippel, T., and van Dyk, D. A. (2013). *The Astrophysical Journal*, **775**, 1–11.

Model Building and Learning



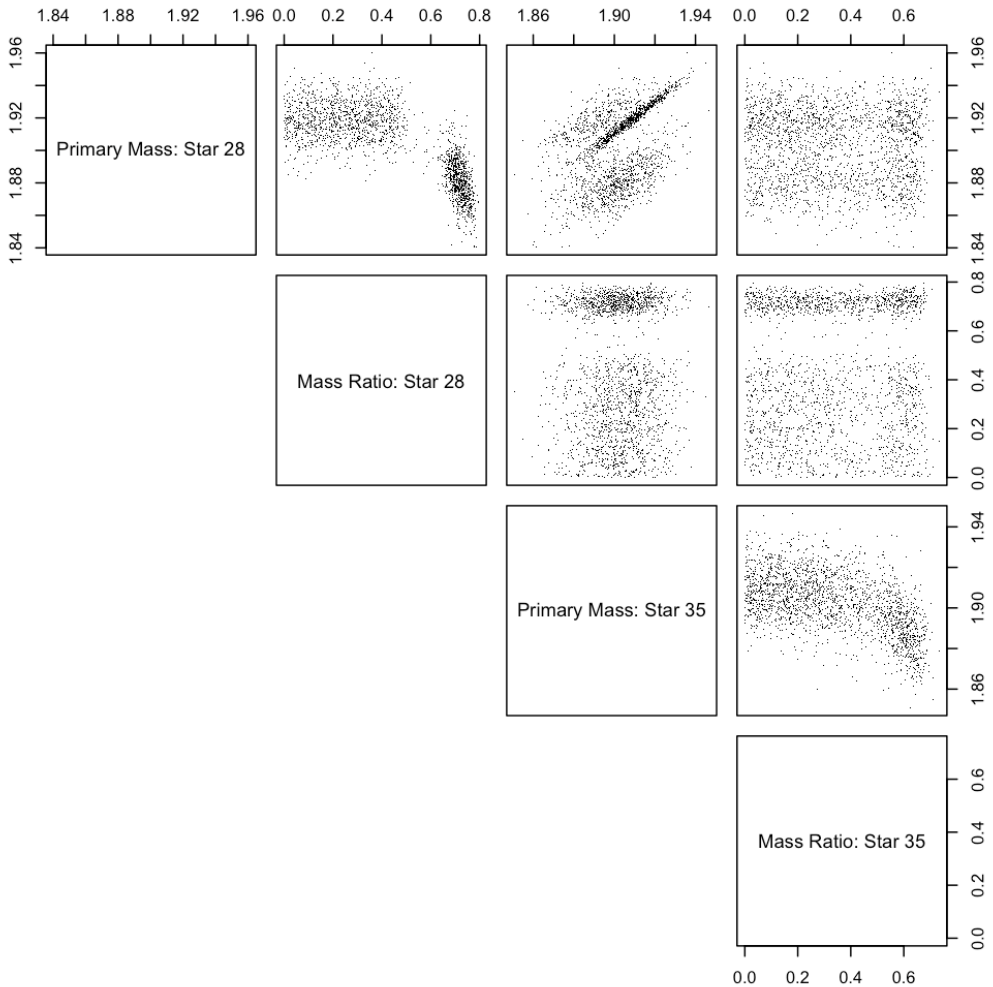
External distance information is very informative!

- Different computer models imply different parameter constraints: combining external information for many stars may be a powerful tool for model selection.
- Learn the age of galactic structures by combining information from multiple stars in a hierarchical model.[8]

Imperial College London

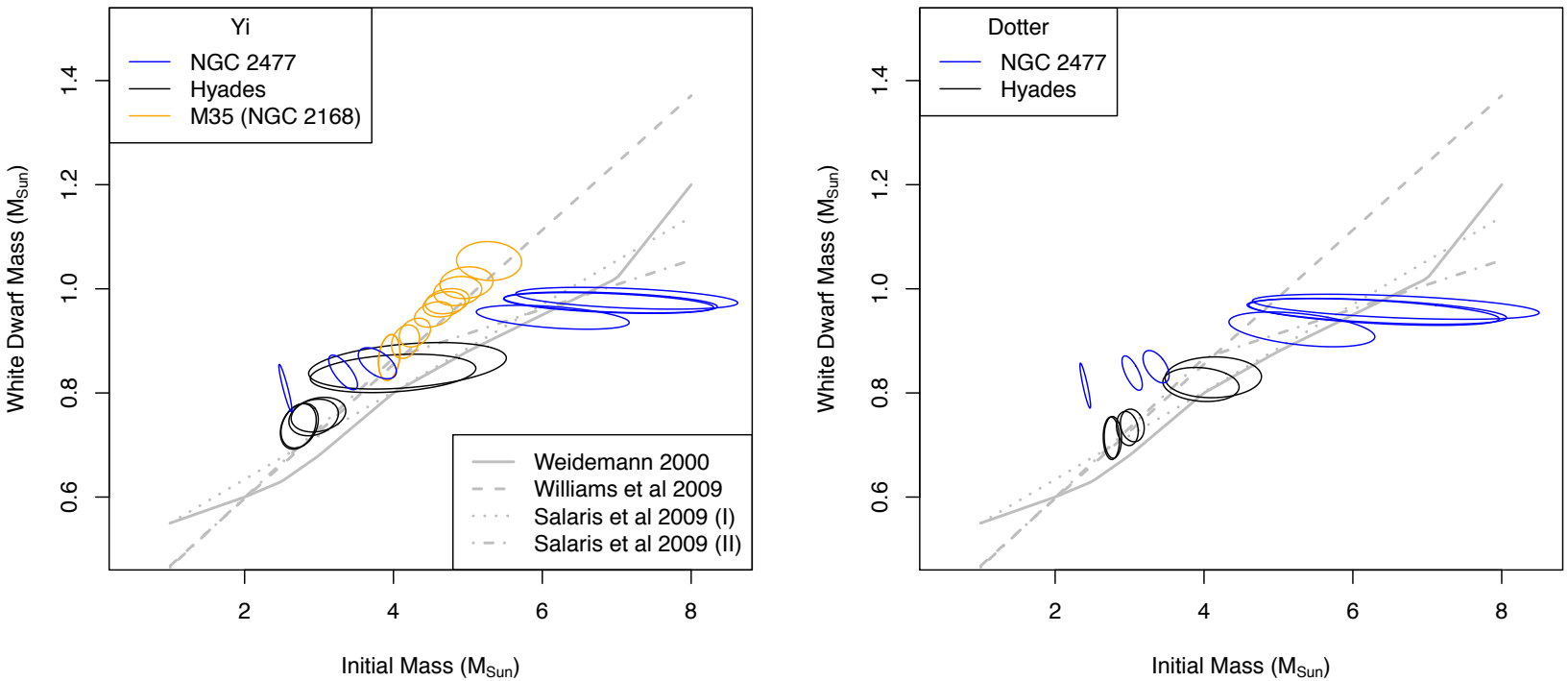
[8] Si, S., van Dyk, D., et al. (2015). Using White Dwarf Populations to Estimate Age of the Galactic Halo, in prep.

Model Fitting: Complex Posterior Distributions



The classification of certain stars as field or cluster stars can cause multiple modes in the distributions of other parameters.

Fitting the Initial-Final Mass Relationship[9]

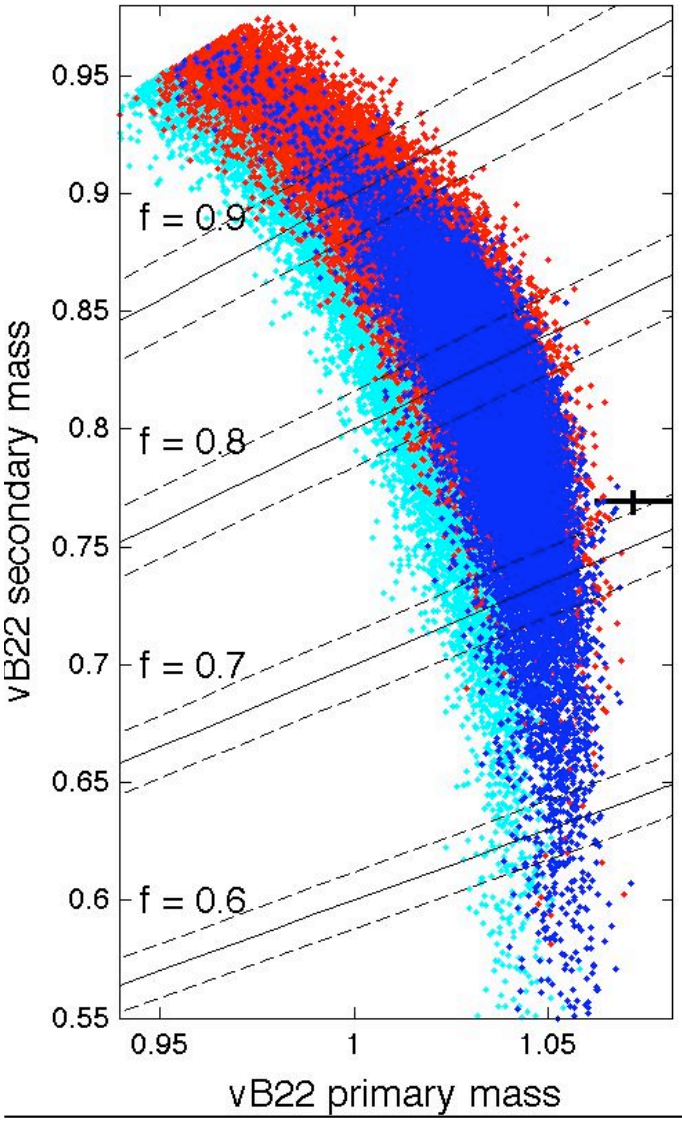


- How best to combine results from three clusters?
- Is there one relationship? Depend on other variables?

[9] Stein, N. M., van Dyk, D. A., von Hippel, T. , DeGennaro, S., Jeffrey, E. J., and Jefferys, W. H. (2013). *Statistical Analysis and Data Mining*, **6**, 34–52.

Imperial College London

Diagnosing Complex Models[10]

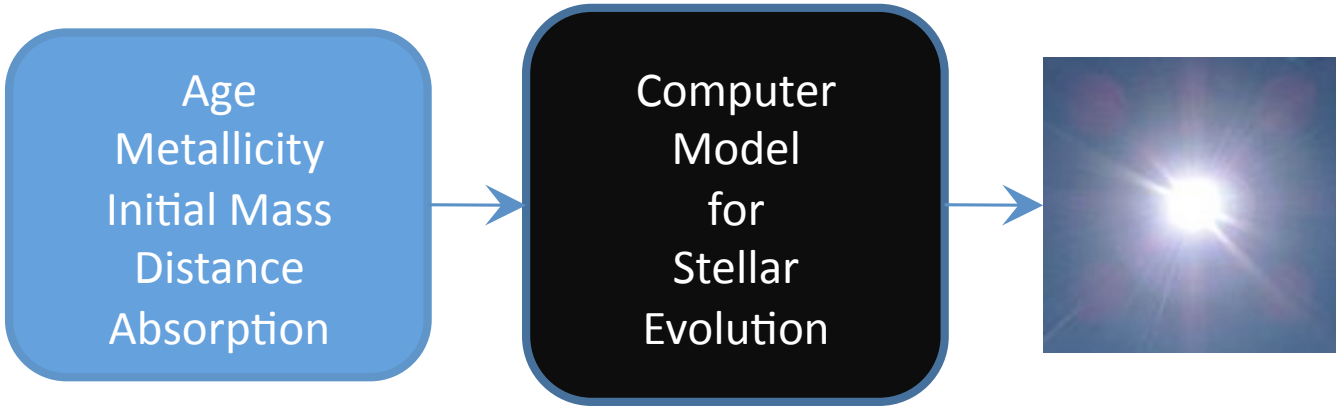


- Double-Line Eclipsing Binaries: direct measures of component masses.
- Double line Spectroscopic: direct measure of mass ratio.
- Direct check of a quantity that resides deep in our statistical model and is highly model dependent.
- Use discrepancies to diagnose and tune computer models, and/or build a joint model.

[10] DeGennaro, S., von Hippel, T., Jefferys, W. H., Stein, N., van Dyk, D. A., and Jeffery, E. (2009). *The Astrophysical Journal*, **696**, 12–23.

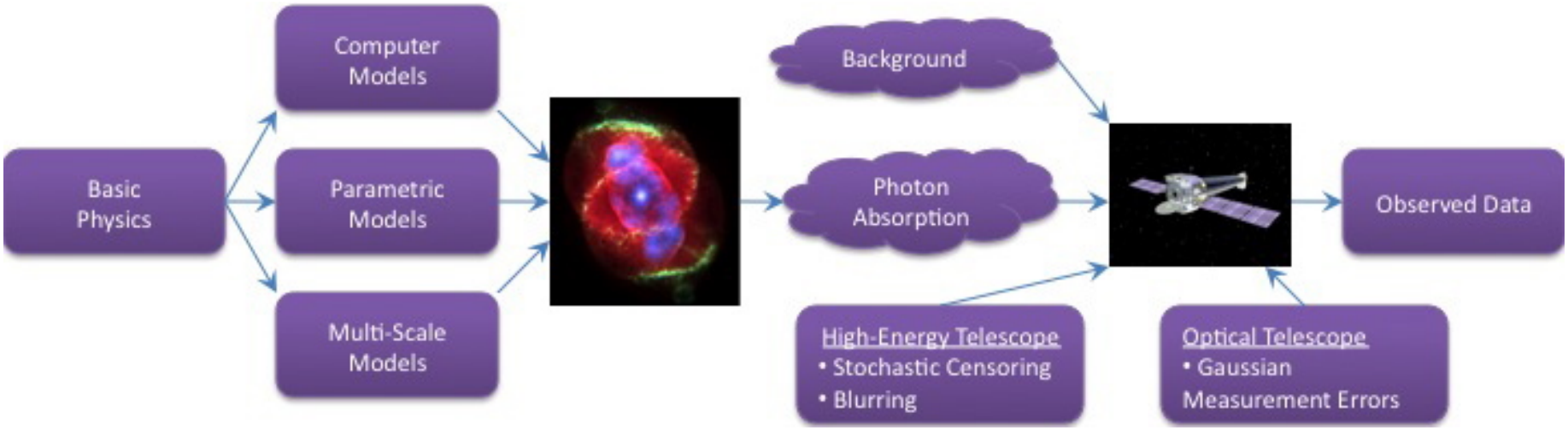
Imperial College London

Big Data Methods in Stellar Evolution



- Massive data sets allow us to cherry pick data that is most informative for a particular scientific question
  - E.g., estimating the age of galactic structures.
  - E.g., checking models using external data.
- Splitting initial & secondary analyses is more efficient.
  - E.g., learning about IFMR from cluster-specific analyses.
- Science-driven models are absolutely essential.

Example III: Calibration of X-ray Detectors[11]



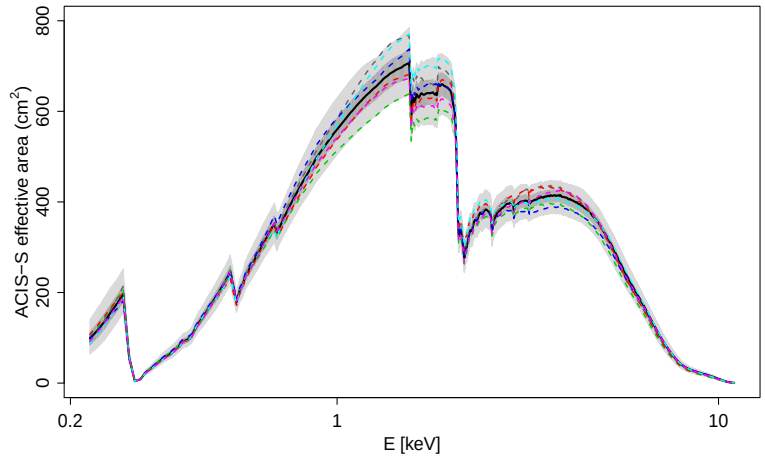
- Embed physics models into multi-level statistical models.
- Must account for complexities of data generation.
- State of the art data and computational techniques enable us to fit the resulting complex model.

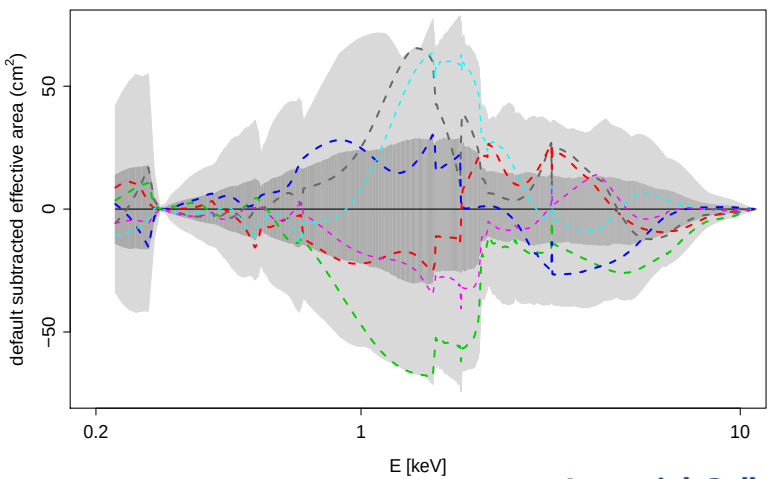
Imperial College London

[11] V Kashyap, D van Dyk, J Xu, A Connors, A Siegminowska, and CHASC

Derivation of Calibration Products

- Effective area records instrument sensitivity as a function of energy
- Complex computer models of subassembly components.
- Calibration scientists provide a sample representing uncertainty





Imperial College London

Simple Emulation of Computer Model[12]

We use *Principal Component Analysis* to represent uncertainly:

$$A \sim A_0 + \bar{\delta} + \sum_{j=1}^{m} e_j r_j \boldsymbol{v}_j,$$

A_0: default effective area,

$\bar{\delta}$: mean deviation from A_0,

r_j and $\boldsymbol{v}_j$: first m principle component eigenvalues & vectors,

e_j: independent standard normal deviations.

Capture 95% of variability with $m = 6 - 9$.

[12] Lee, H., Kashyap, V., van Dyk, D., Connors, A., Drake, J., et al. (2011). Accounting for Calibration Uncertainties in X-ray Analysis: Effective Areas in Spectral Fitting. *The Astrophysical Journal*, **731**, 126–144.

Two Possible Target Distributions[13]

We consider inference under:

A PRAGMATIC BAYESIAN TARGET: $\pi_0(A, \theta) = p(A)p(\theta|A, Y)$.

THE FULLY BAYESIAN POSTERIOR: $\pi(A, \theta) = p(A|Y)p(\theta|A, Y)$.

Concerns:

- **Statistical** Fully Bayesian target is "correct".
- **Cultural** Astronomers have concerns about letting the current data influence calibration products.
- **Computational** Both targets pose challenges, but pragmatic Bayesian target is easier to sample.
- **Practical** How different are $p(A)$ and $p(A|Y)$?

With MCMC we sample a different effective area curve at each iteration according to its conditional distribution.

[13] Xu, J., van Dyk, D., Kashyap, V., Siemiginowska, A., et al. (2014). A Fully Bayesian for Jointly Fitting Instrumental Calibration and X-ray Spectral Models. *The Astrophysical Journal*, **794**, 97.

Implementing the Fully Bayesian Analysis

Direct MH sampling is difficult. (Case-by case tuning of jumping rules.)

Pragmatic Bayesian posterior

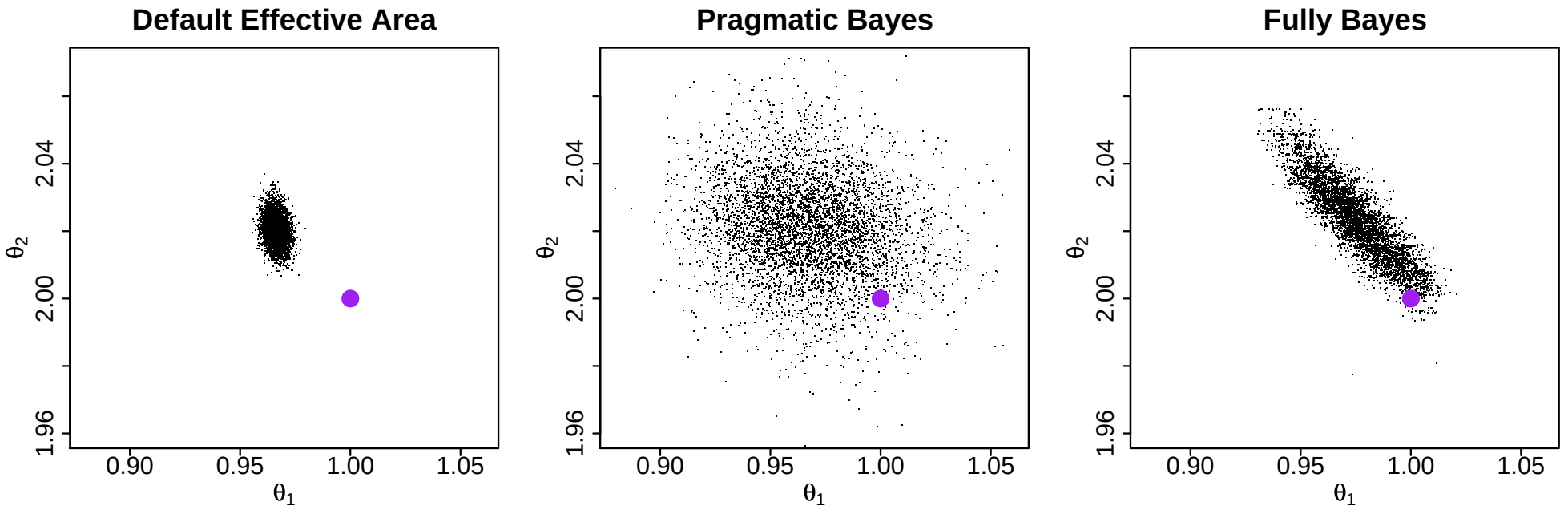
- We can easily sample from $\pi_0(A, \theta)$.
- Well suited proposal dist'n: over-dispersed relative to $\pi(A, \theta)$.
- But $\pi_0(A, \theta)$ cannot be evaluated

$$\pi_0(A, \theta) = p(\theta | Y, A)p(A) = \frac{p(Y|\theta, A)p(\theta)}{p(Y|A)}p(A)$$

This is a doubly intractable distribution.

- We construct a normal approximation ($\sim$ 20 dimensional).
- Use as jumping rule in an independence MH sampler.

Sampling From the Full Posterior

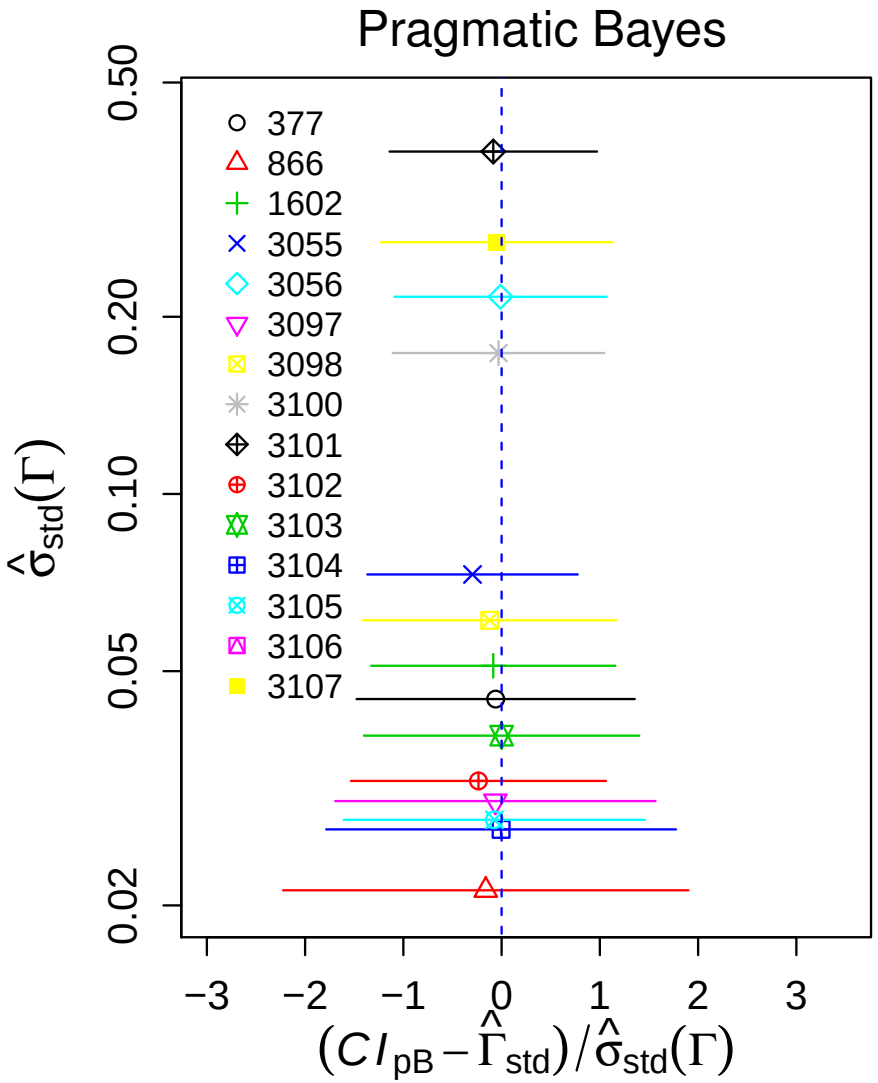


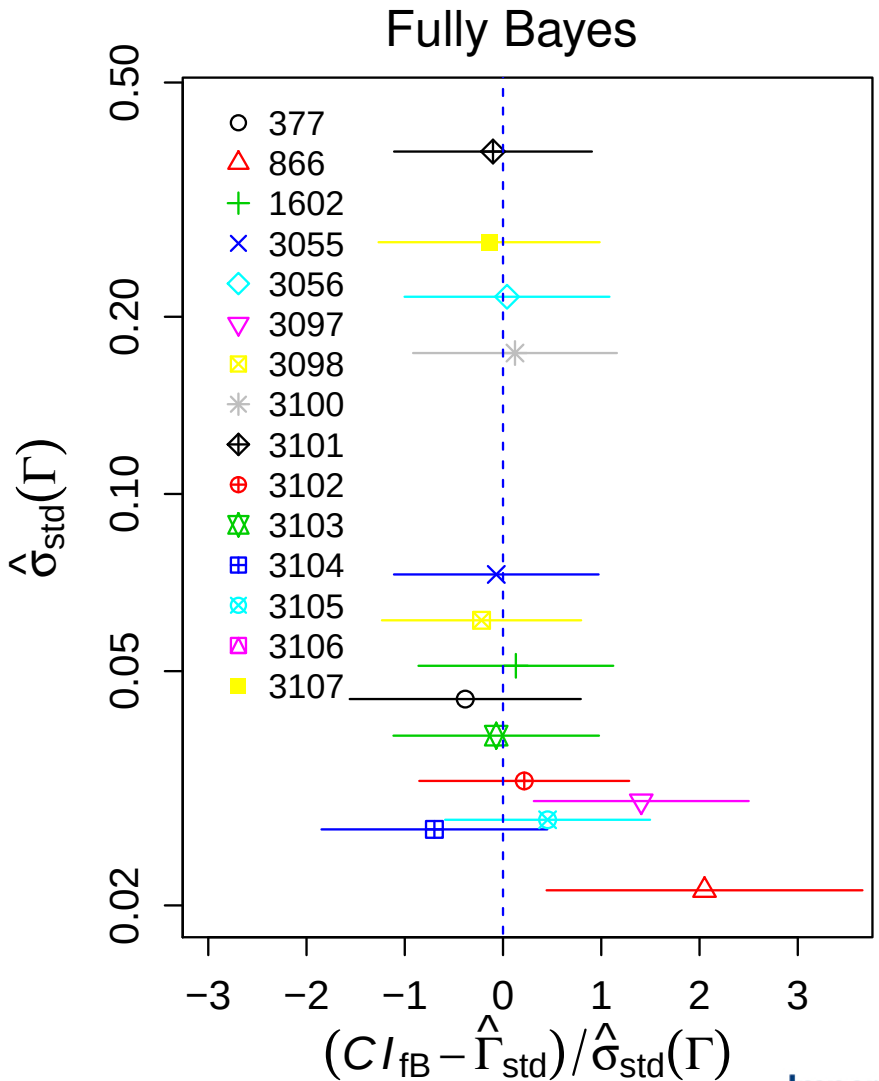
Spectral Model (purple bullet = truth):

$$f(E_j) = \theta_1 E_j^{-\theta_2}$$

Pragmatic Bayes is clearly better than standard method, but a Fully Bayesian Method is the ultimate goal.

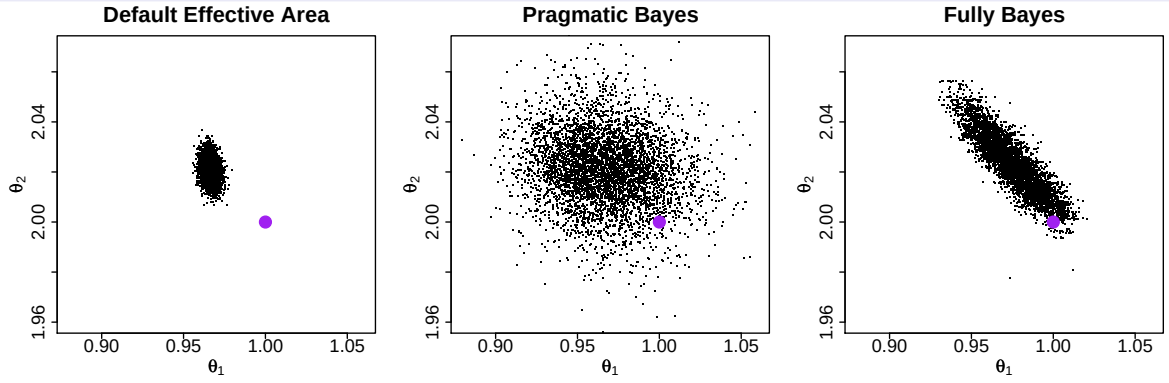
How it Works on a Sample of Radio-Loud Quasars





Imperial College London

Big Data Methods in X-ray Analysis



- Represent high-dimensional complexity with low-dimensional summaries
  - E.g., use PCA to emulate uncertainty in effective area; use same variance structure for all effective area curves.
- Use simplifying assumptions when possible
  - E.g., pragmatic Bayesian assumption
  - E.g., Gaussian approximation to the pragmatic posterior
- Mix science-driven models with data-driven methods.
  - E.g., we must understand the spectra, but only need to predict the instrumental characteristics.

Thanks...

Solar Features:

- Nathan Stein
- David Stenning
- Vinay Kashyap
- Thomas Lee
- Xiao-Li Meng

Instrument Calibration:

- Jin Xu
- Alanna Connors
- Vinay Kashyap
- Hyunsook Lee
- Aneta Siegminowska

Stellar Evolution:

- Nathan Stein
- David Stenning
- Shijing Si
- Elizabeth Jeffery
- William H. Jefferys
- Erin M. O'Malley
- Ted von Hippel

And

The CHASC International Astro-Statistics Collaboration

Imperial College London

For Further Reading I


Xu, J., van Dyk, D., Kashyap, V., Siemiginowska, A., Connors, A., Drake, J., et al.
A Fully Bayesian Method for Jointly Fitting Calibration and X-ray Spectral Models
The Astrophysical Journal, **794**, 97, 2014.


O.Malley, E. M., von Hippel, T., and van Dyk, D. A.
A Bayesian Approach to Deriving Ages of Individual Field White Dwarfs.
The Astrophysical Journal, **775**, 1–11, 2013.


Stenning, D., Lee, T., van Dyk, D., Kashyap, V., Sandell, J., and Young, C.
Morphological Feature Extraction for Statistical Learning [in Solar Images]
Statistical Analysis and Data Mining, **6**, 329–345, 2013.


Stein, N, van Dyk, D., von Hippel, T., DeGennaro, S., Jeffery, E., Jeffreys, W. H.
Combining Computer Models to Account for Mass Loss in Stellar Evolution.
Statistical Analysis and Data Mining, **6**, 34–52, 2013.


Stein, N., Kashyap, V., Meng, X. L., and van Dyk, D. A.
H-Means Image Segmentation to Identify Solar Thermal Features.
Proceedings of the 19th IEEE Intern'l Conf. on Image Processing, ICIP 2012


Lee, H., Kashyap, V., van Dyk, D., Connors, A., Drake, J., Izem, R., Min, S., et al.
Accounting for Calibration Uncertainties in X-ray [Spectral] Analysis
The Astrophysical Journal, **731**, 126–144, 2011.

New ASA Interest Group!

New! Astrostatistics Interest Group New!

At the JSM:

- Sunday at 4 PM: Bayesian Astrostatistics
- Wednesday at 8:30 AM: Big Data in Astrostatistics
- Wednesday at 10:30 AM: Informal Meeting outside the "Big Data in Astrostatistics" session room
- Wednesday at 2:00 PM: Analysis of Kepler Data at SAMSI
- Thursday at 8:30 AM: IOL: Astrostatistics

For more information:

`http://community.amstat.org/astrostats/home`

Get Involved!

Association of Astrostatisticians

- New ASA Interestgroup:
  `http://community.amstat.org/astrostats/home`
- International Astrostatistics Association (New! Working Groups!)
- Astrostatistics and Astroinformatics Portal:
  `http://asaip.psu.edu`

Data Challenges Competitions

- Banff Challenge 1 & 2: Davison and Sartori (Stat Sci, 2008)
  `http://www.birs.ca/events/2010/5-day-workshops/10w5068`
- GREAT08 & GREAT10: `arXiv:0908.0945v1` `arXiv:1202.5254v2`
- Strong Lens Time Delay: `http://timedelaychallenge.org`

Vast Public Data Resources

- The Virtual Observatory: `http://www.usvao.org`
- The Sun Today: `http://www.thesuntoday.org`

Mathematical Morphology for Solar Features

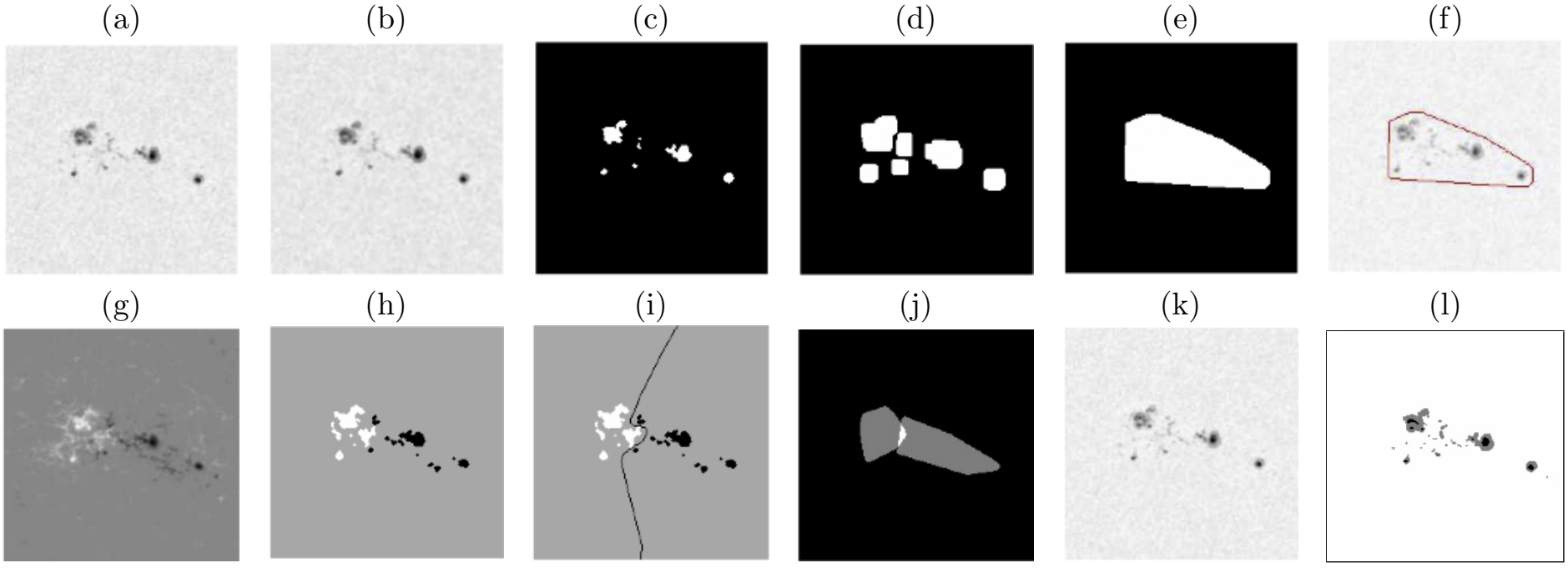


Figure 3 The original $\beta\gamma$ white light image (a) is cleaned (b) and thresholded to produce a binary representation of the sunspot group (c). This image is then dilated (d) and has a convex hull placed around the result (e) and the area inside the hull becomes the sunspot area (f) in the magnetogram. Then, in the $\beta\gamma$ magnetogram (g), morphological opening followed by thresholding on both the image and inverse image yields the trinary primal sketch of the active region in (h). Region growing gives the separating boundary in (i). Convex hulls are utilized to measure polarity mixture in (j). We smooth the white light image in (k) and apply thresholding iteratively in (l) to produce a representation of the umbrae and penumbrae that can be used to detect delta spots.