Statistical Learning Challenges in Astronomy and Solar Physics

David A. van Dyk

Statistics Section, Imperial College London

BIRS Workshop on Statistical Learning, December 2011

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 volume is growing.... astronomically!!

Massive statistical learning challenges!!

Scaling up to Massive Data Streams

Basic Techniques

- Dimension Reduction: Transform high dimensional data to simpler forms more amenable to standard analyses.
- Speed up "inner loop" computation of many methods for application to massive surveys (Alex Gray, Georgia Tech).
- Algorithm-based methods that can handle massive data.

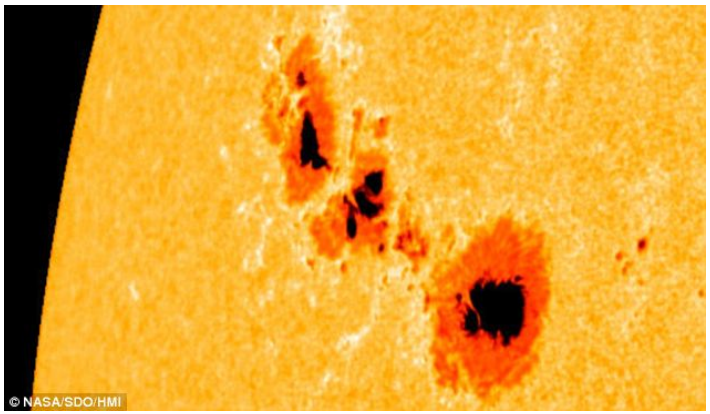
Examples

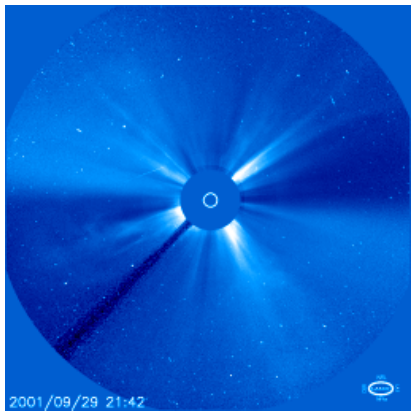
- **CMU and Berkeley Astrostat:** Reduce high dimensional data (e.g., light curves) to automatically identify and classify objects (e.g., variable stars).
- **Imperial Astrostat:** Search for anomalies in multiple massive semi-incompatible surveys–found most distant known quasar.

Complex Data and Sophisticated Models

- A great leap forward: *Large Synoptic Survey Telescope* (1.28 petabytes/year).
- But data are *not just massive:* they are rich, deep, & complex.
- LSST: Trigonometric Parallax, Proper Motion, and Photometric data in 5 bands.
- Require specialized models, methods, and computation.
- Idiosyncratic statistical challenges.

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



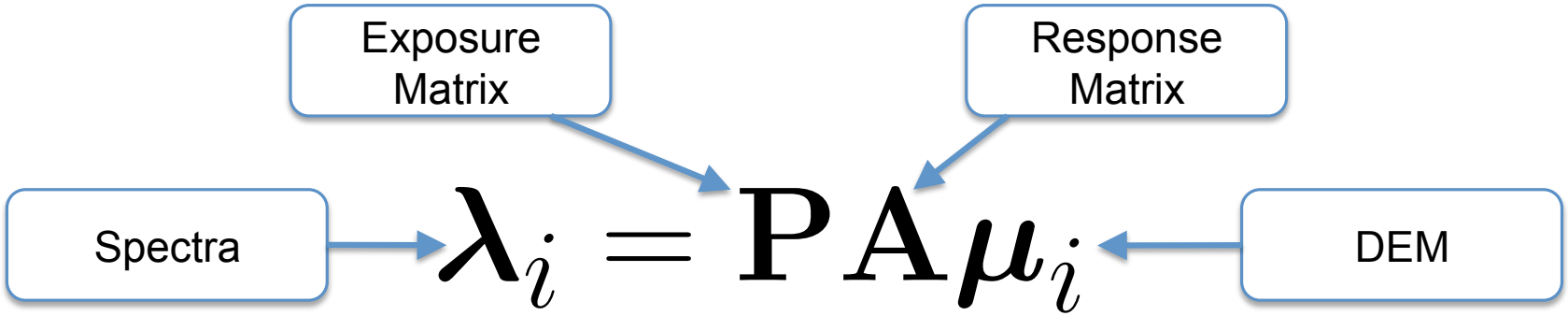


- Highly energetic and violent *solar corona* is characterized by sunspots, solar flares, and coronal mass ejections.
- Solar storms can affect space weather, earth satellites, communication systems, and electric grids.
- *Goal:* Track solar activity with the aim of predicting storms and their effects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

Imperial College London

[1] N Stein, XL Meng, V Kashyap, and iCHASC

The Diffuse Emission Measure

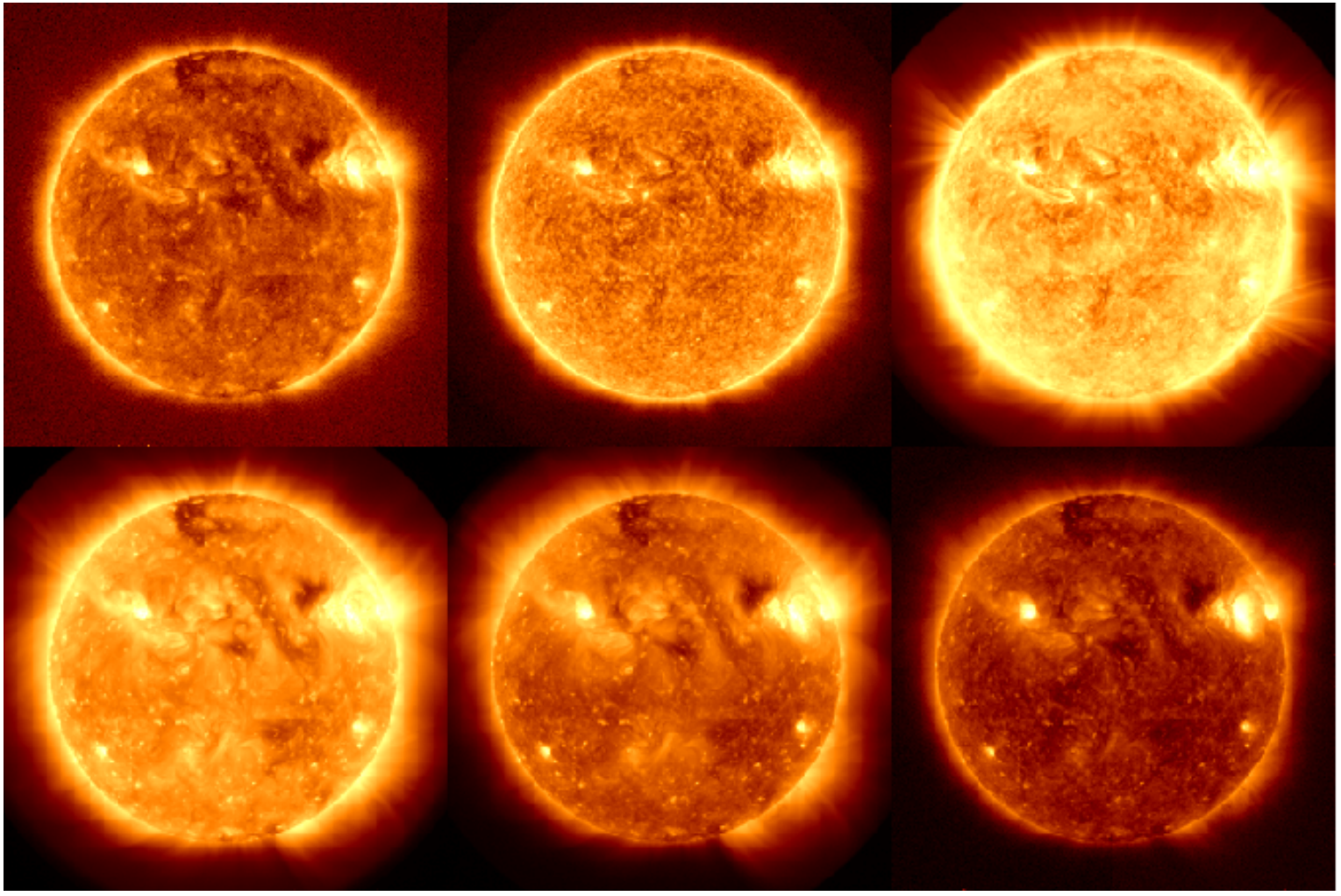


DEM: *expected emission due to plasma of a given temperature.*
A: *expected spectra of plasmas at each temperature.*

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.*

The Data: Pixel-by-Pixel Spectra



High-Resolution Images with Low-Resolution 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.

A Generalized K-means Algorithm:

- Replace Euclidean distance with appropriate alternative.
- Ideally both steps remain in closed form!!
- What distance should we use for probability vectors?

H-means

H-means Algorithm:

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

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

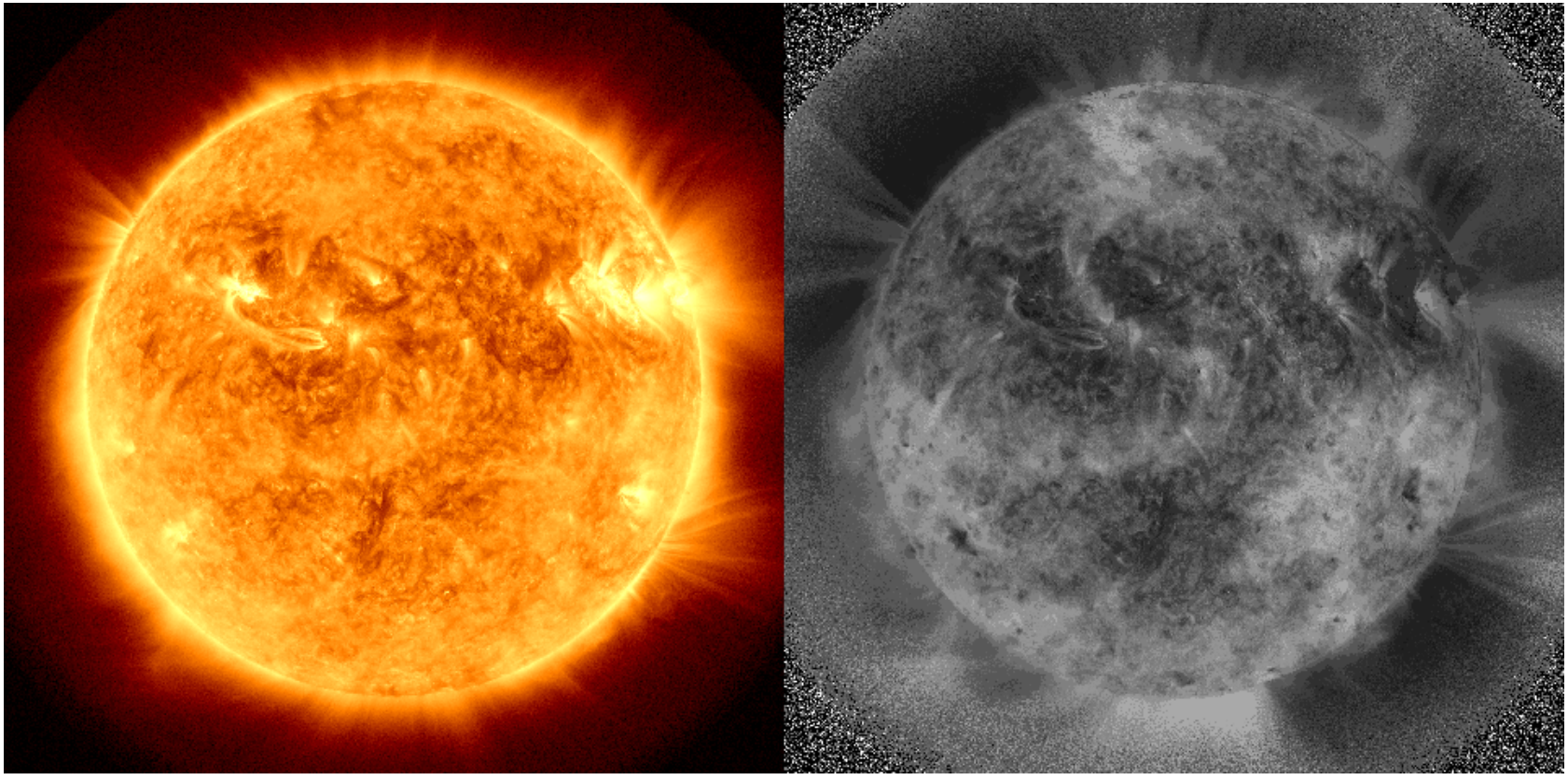
- **Hellinger** distance between $\hat{\pi}_i$ and $\boldsymbol{c}_j$:

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

- Update centroid for cluster j, $(c_{j1}, \ldots, c_{jK})$:

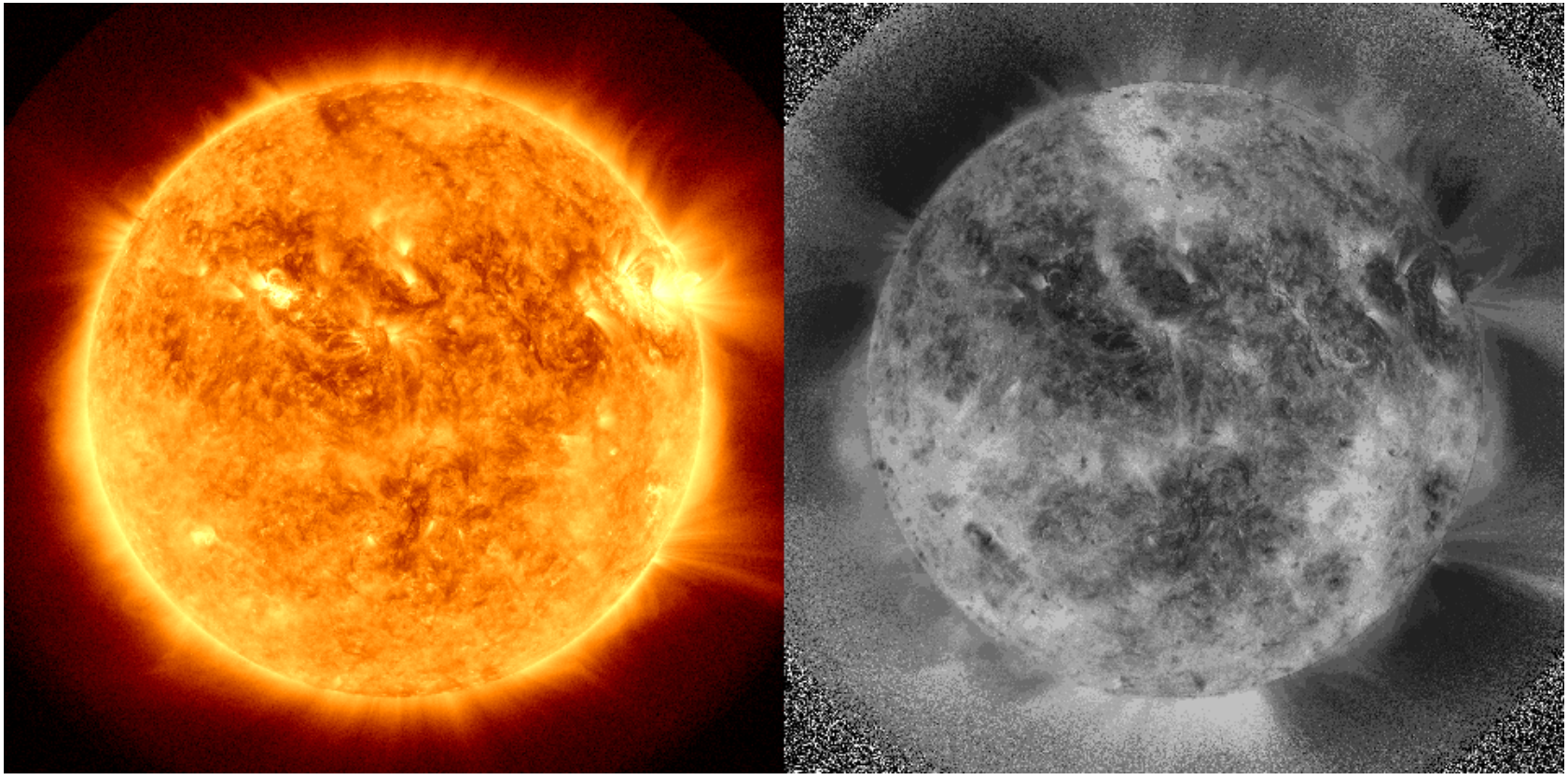
$$c_{jk} = \frac{\left(\sum_{i \in \text{ cluster } j} \sqrt{\hat{\pi}_{ik}}\right)^2}{\sum_{k'} \left(\sum_{i \in \text{ cluster } j} \sqrt{\hat{\pi}_{ik'}}\right)^2}.$$

Never Before Seen Structure



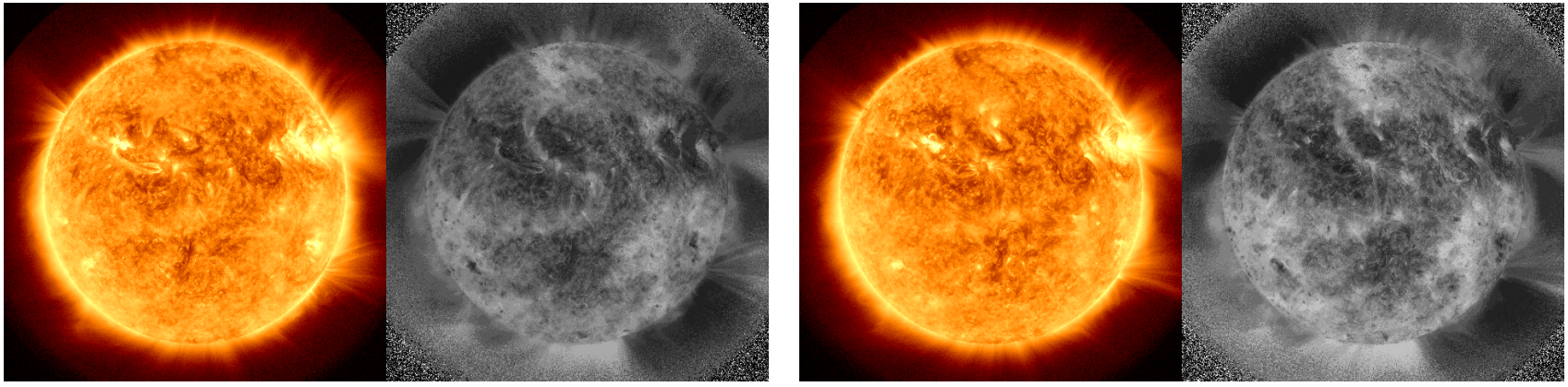
Grey Scale Images of Clusters: 2 Oct 2010 at 05.57

Never Before Seen Structure



Grey Scale Images of Clusters: 2 Oct 2010 at 18.43

Current Work



- Like K-means, H-means assumes cluster shapes and sizes are all the same. What is the effect?
- Ultimately we want to include spatial structure, track clusters, and predict events.
- Over time clusters evolve spatially and spectrally.
- Reconstruct the underlying DEM in individual clusters.

Example II: Stellar Evolution[2]

Complex Data and Sophisticated Models

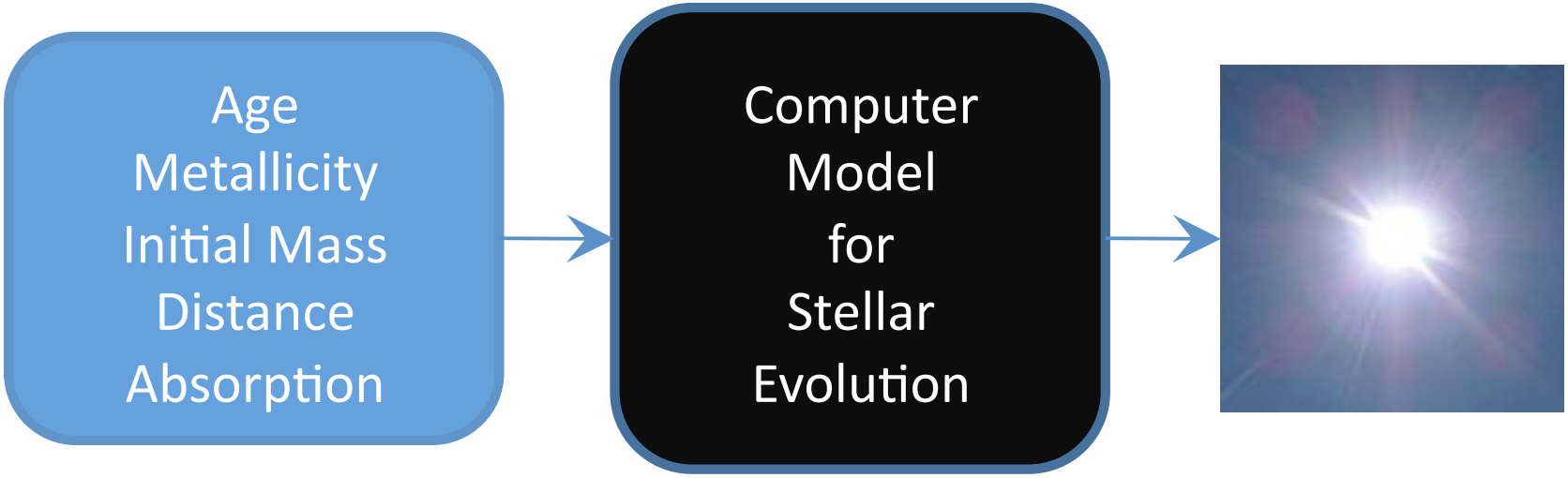
1. Complex computer models and simulations are taking the place of the analytic likelihood function.
2. Sophisticated data allows us to fit such models, but an entirely new set of methods is required.
3. This sort of modeling, computing, and inference is coming to many more areas of Astronomy.
4. I will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

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

Imperial College London

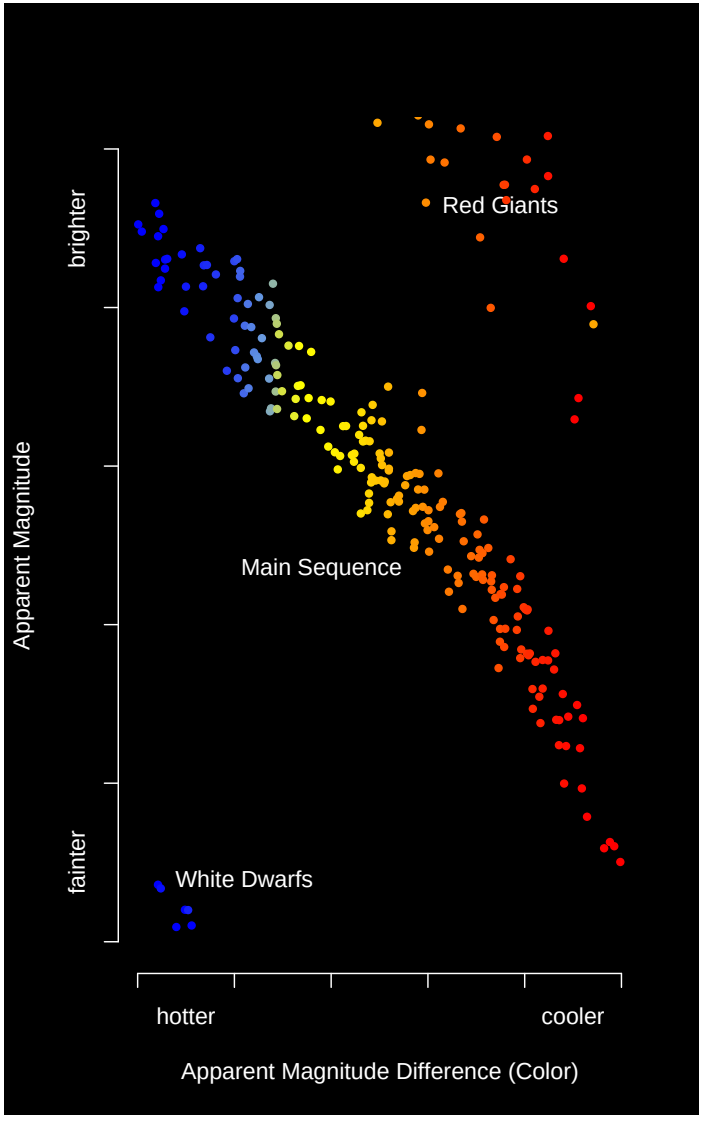
[2] N Stein, D van Dyk, S DeGennaro, E Jeffery, W Jefferys, and T von Hippel

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.

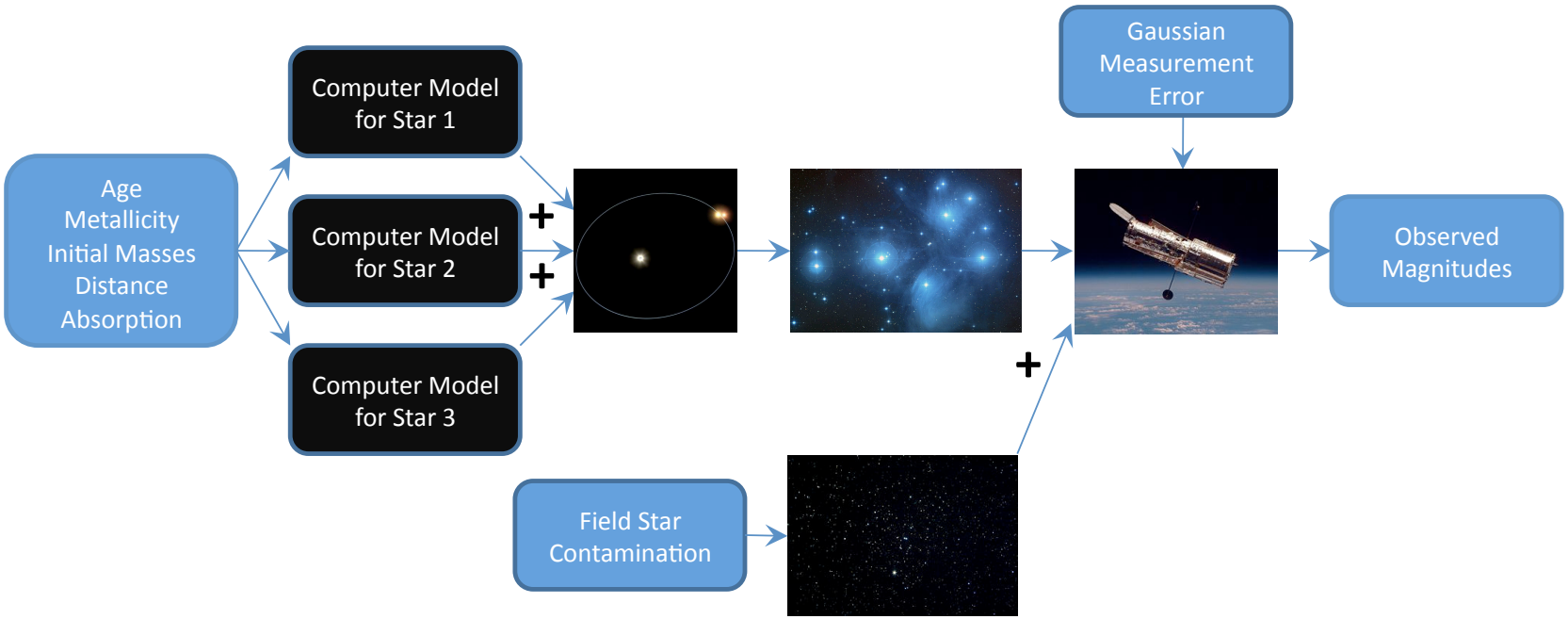
The Data: Color Magnitude Diagrams



Color-Magnitude Diagram

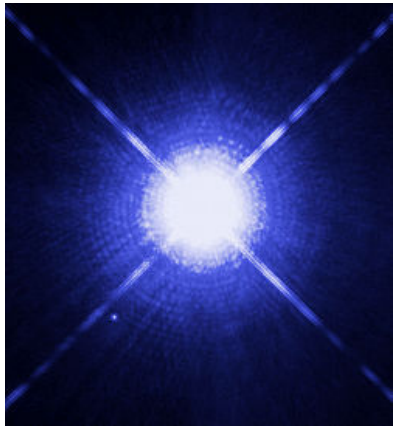
- Plot Magnitude Difference vs. Magnitude.
- Identifies stars at different stages of their lives.
- Evolution of a CMD.
- Facilitates physical intuition as to likely values of parameters.
- “Chi-by-eye” fitting.

Embedding Computer Model into Statistical Model



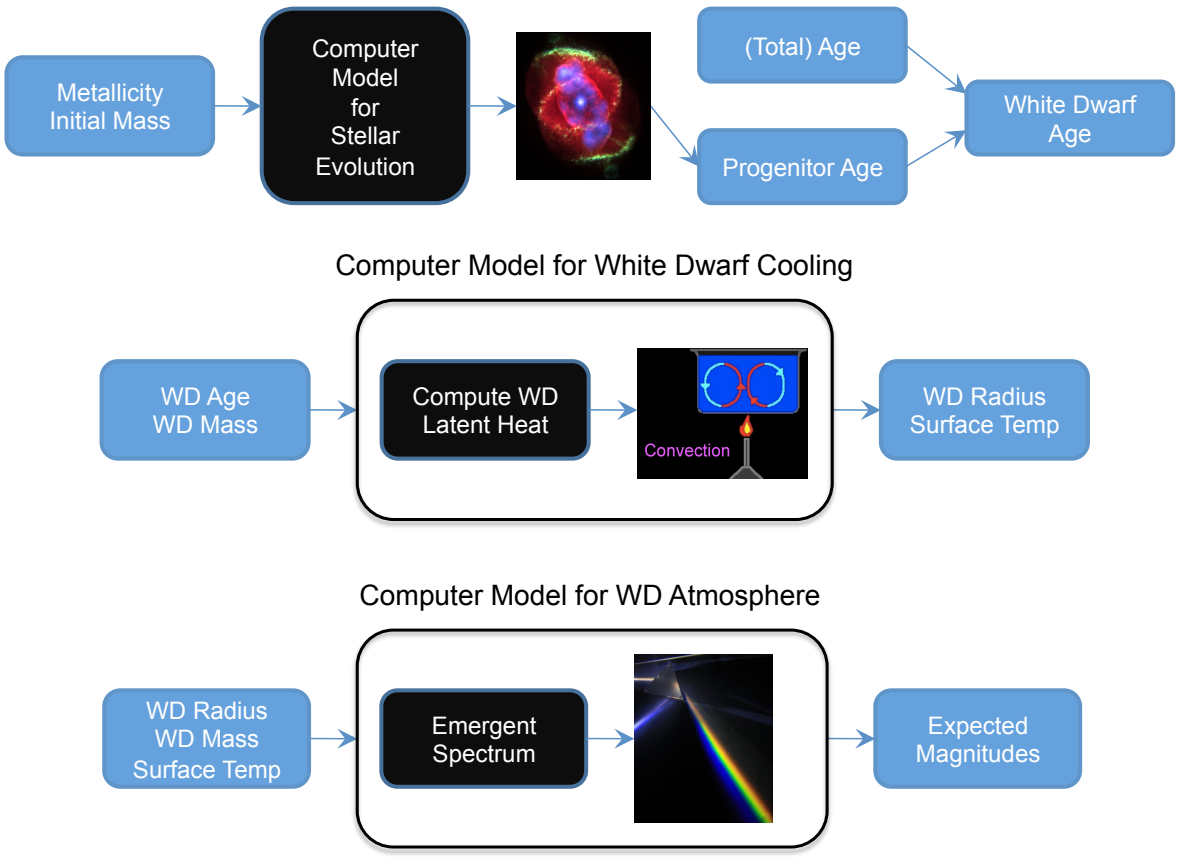
- Between 1/3 and 1/2 of “stars” are unresolved binaries.
- Star clusters: same age, metallicity, distance, & absorption.
- Cluster data is contaminated with field stars.
- Data observed with Gaussian measurement errors.

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.

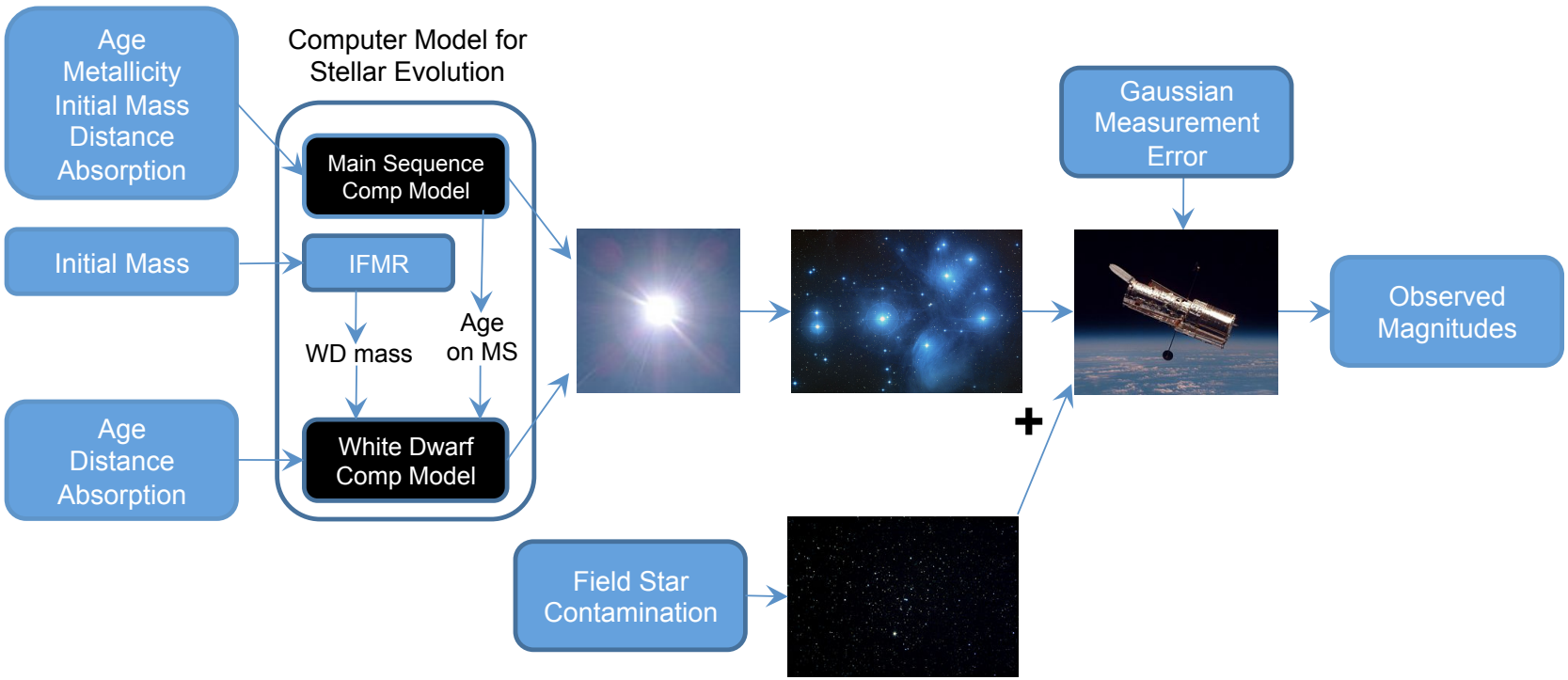
The Missing Link: White Dwarf Mass



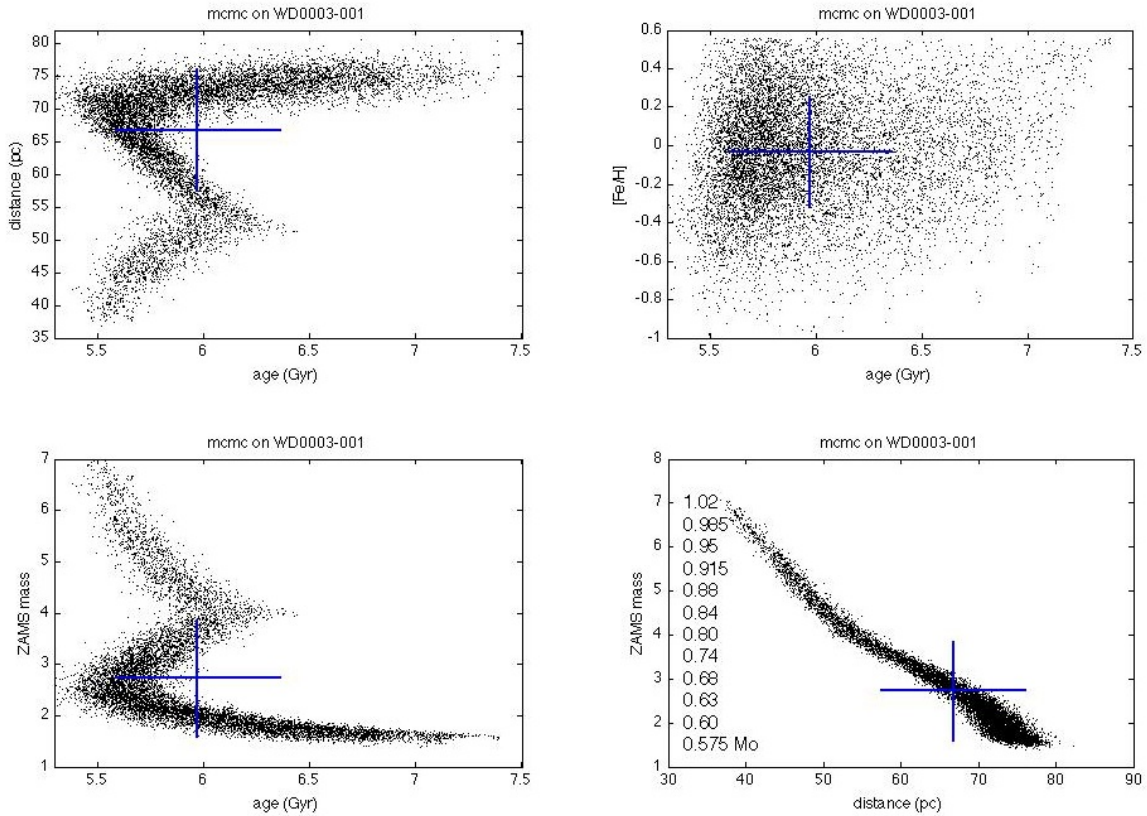
- We must model: white dwarf mass $= f$(initial mass).
- Parametric Bridge between Computer Models.

Imperial College London

Opening Up the Black Box: The Final Model

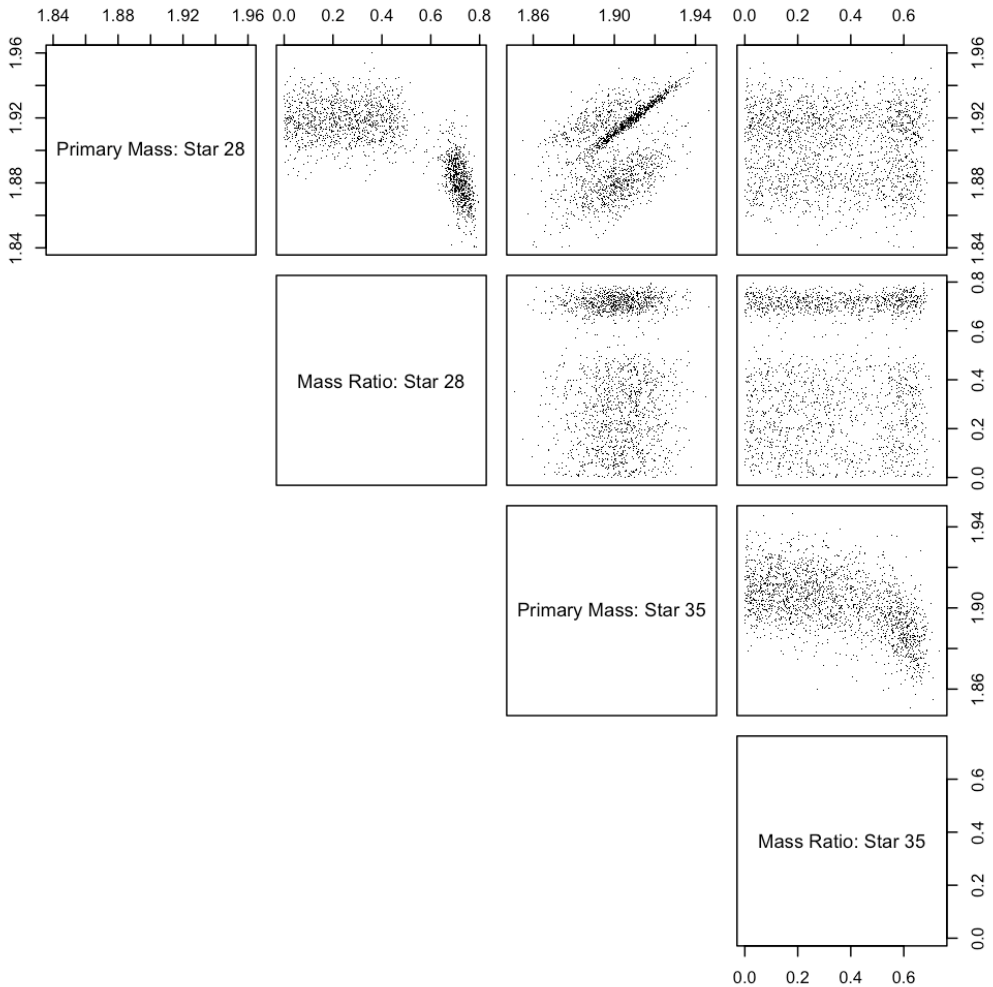


Model Fitting: Complex Posterior Distributions



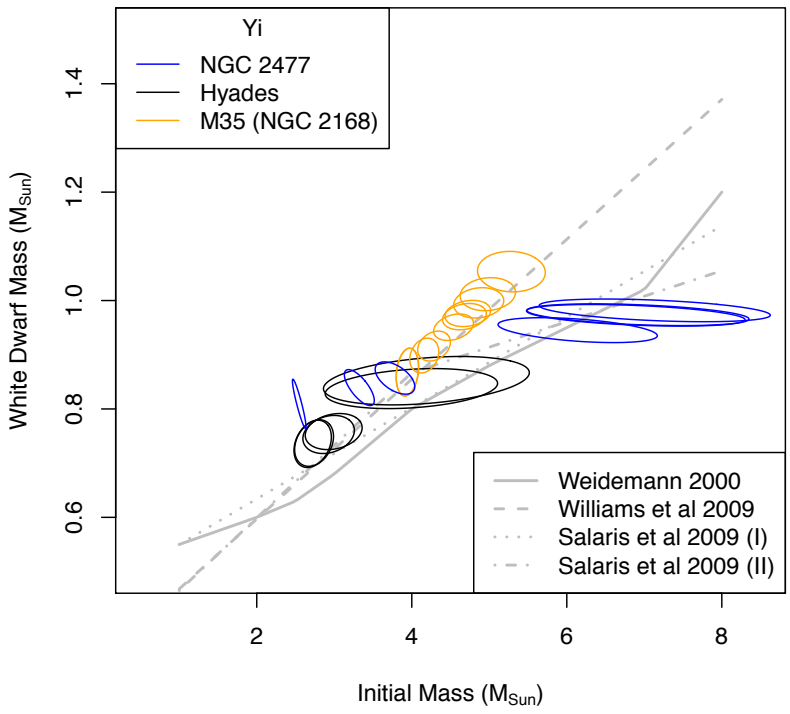
Highly non-linear relationship among stellar parameters.

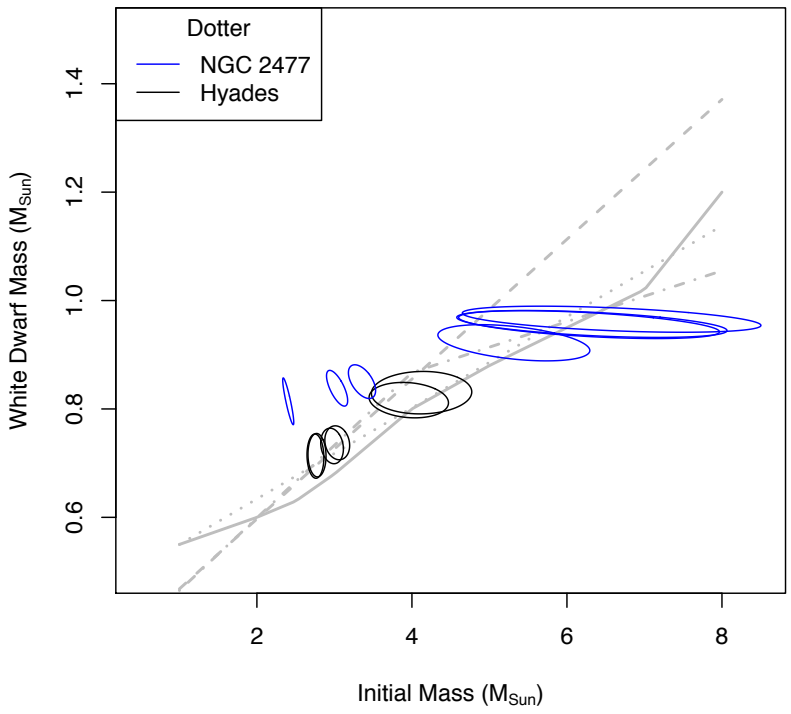
Model Fitting: Complex Posterior Distributions



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

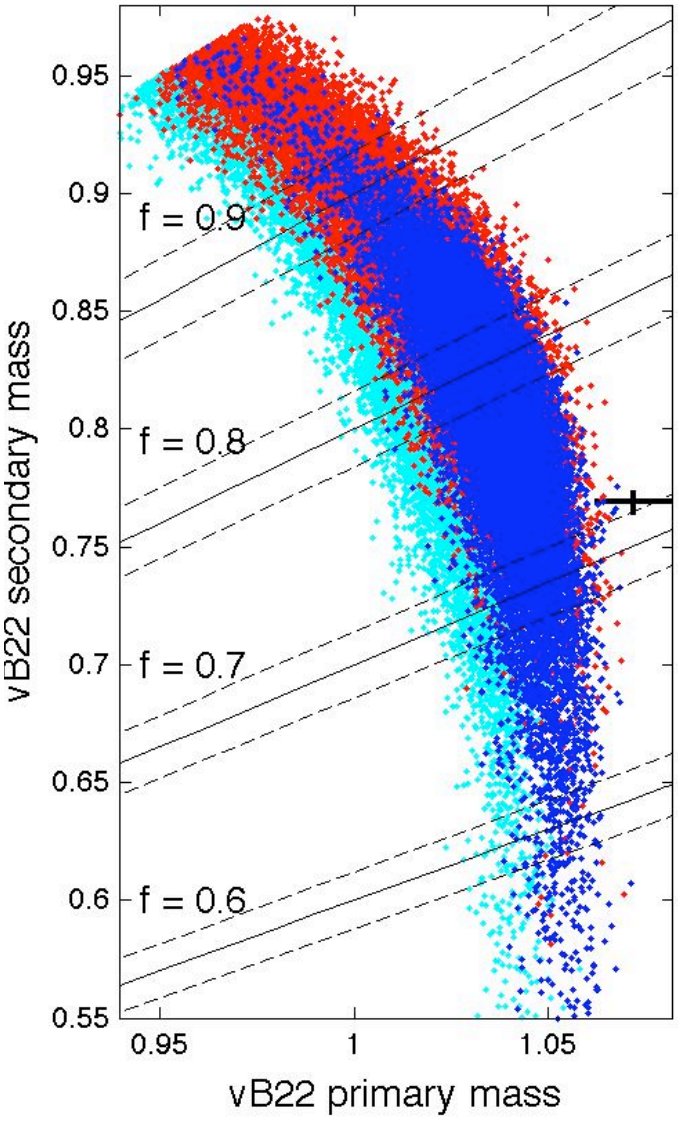




- How best to combine results from three clusters?
- Is there one relationship? Depend on other variables?

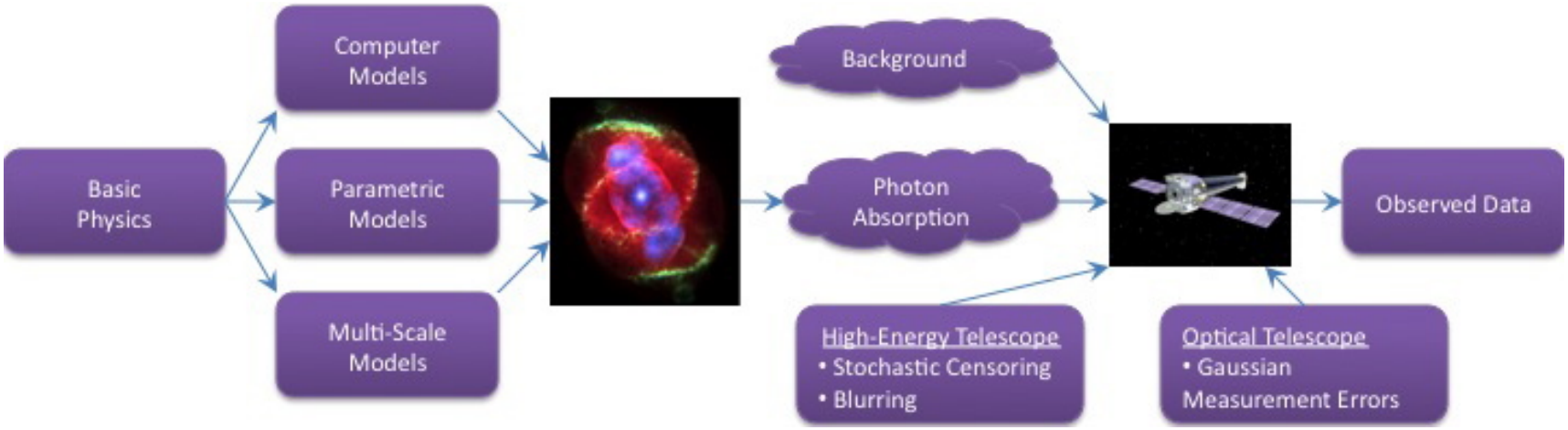
Imperial College London

Diagnosing Complex Models



- 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.

Example III: Calibration of X-ray Detectors[3]



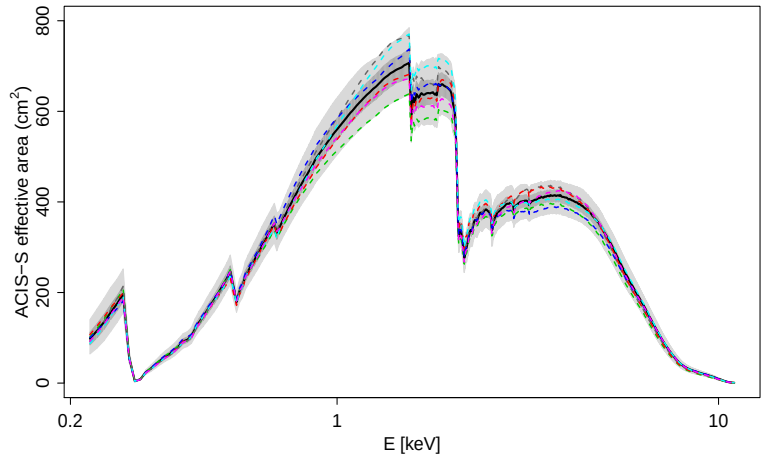
- Embed physics models into multi-level statistical models.
- X-ray and γ-ray detectors count a typically *small number of photons* in each of a *large number of pixels*.
- Must account for complexities of data generation.
- State of the art data and computational techniques enable us to fit the resulting complex model.

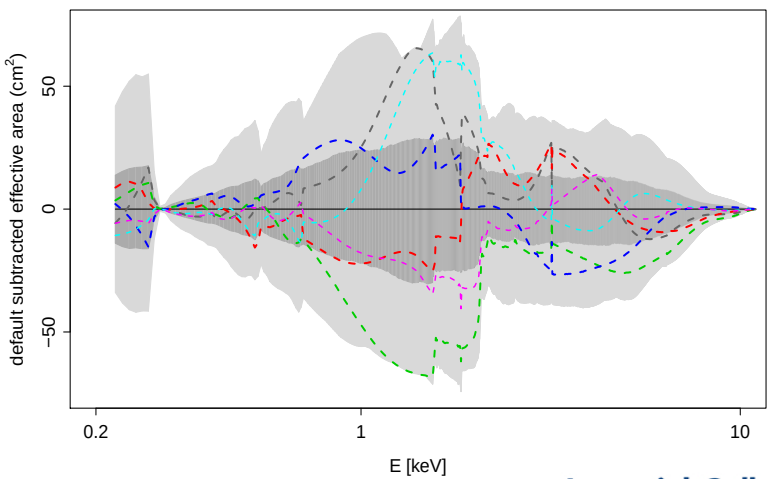
Imperial College London

[3] V Kashyap, D van Dyk, J Xu, A Connors, A Siegminowska, and iCHASC

Derivation of Calibration Products

- Effective area records instrument sensitivity as a function of energy
- Aim to capture deterioration of detectors over time.
- Complex computer models of subassembly components.
- Calibration scientists provide a sample representing uncertainty
- *Calibration Sample* is typically of size $M \approx 1000$.





Imperial College London

Simple Emulation of Computer Model

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 $m = 6 - 9$.

Two Possible Target Distributions

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.

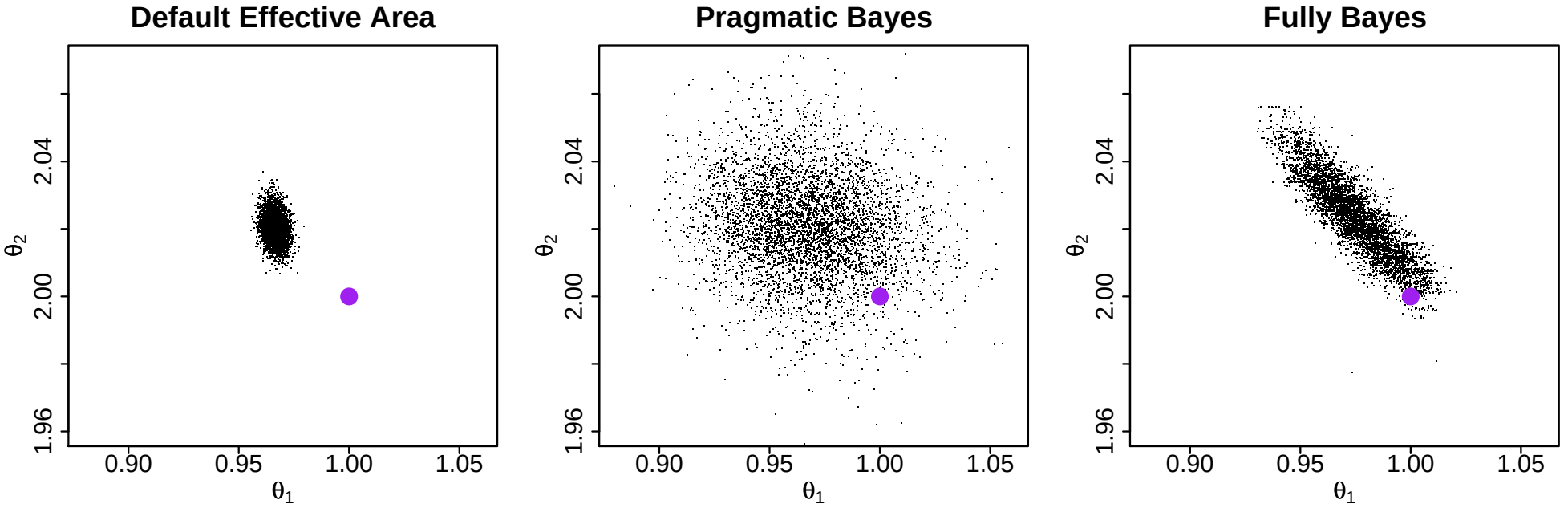
Sampling the Full Posterior Distribution

- Sampling $\pi(A, \theta) = p(A, \theta | Y)$ is complicated because we only have a computer-model generated sample of $p(A)$ rather than an analytic form.
- But PCA gives a *degenerate normal* approximation:

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

  where e_j are independent standard normals.
- PCA represents A as deterministic function of $e = (e_1, \ldots, e_m)$.
- We can construct an MCMC sampler of $p(e, \theta | Y)$.

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 current practice, but a Fully Bayesian Method is the ultimate goal.

Thanks...

Solar Thermal Structure:

- Nathan Stein
- Xiao-Li Meng
- Vinay Kashyap

Stellar Evolution:

- Nathan Stein
- Steven DeGennaro
- Elizabeth Jeffery
- William H. Jefferys
- Ted von Hippel

Instrument Calibration:

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

And

iCHASC:
Imperial-California-Harvard
AstroStatistics Collaboration

For Further Reading I


Stein, N, van Dyk, D., von Hippel, T., DeGennaro, S., Jeffery, E., Jeffreys, W. H.
Combining Computer Models in a Principled Bayesian Analysis: From Normal Stars to White Dwarf Cinders. Submitted.


Jeffery, E., von Hippel, T., DeGennaro, S., van Dyk, D., Stein, N., and Jefferys, W.
The White Dwarf Age of NGC 2477. *Astrophysical Journal*, **730**, 35–43, 2011.


Lee, H., Kashyap, V., van Dyk, D., Connors, A., Drake, J., Izem, R., Min, S., Park, T., Ratzlaff, P., Siemiginowska, A., and Zezas, A.
Accounting for Calibration Uncertainties in X-ray Analysis: Effective Area in Spectral Fitting. *The Astrophysical Journal*, **731**, 126–144, 2011.


van Dyk, D. A., DeGennaro, S., Stein, N., Jefferys, W. H., von Hippel, T.
Statistical Analysis of Stellar Evolution *The Annals of Applied Statistics* **3**, 117-143, 2009.


DeGennaro, S., von Hippel, T., Jefferys, W., Stein, N., van Dyk, D., and Jeffery, E.
Inverting Color-Magnitude Diagrams to Access Precise Cluster Parameters: A New White Dwarf Age for the Hyades. *Astrophysical Journal*, **696**, 12–23, 2009.