Future Directions of Astrostatistics

David A. van Dyk

Department of Statistics, University of California, Irvine
Statistics Section, Imperial College London

SCMA V, June 2011

Massive Data Sets and Data Streams

Are we drowning in our own data??

1. **D. Marinucci:** Exponential growth, esp. in cosmology.
2. **K. Borne:** More data is not just more-qualitatively different.
3. New Statistical/CS methods for massive data...
   - ML/AI methods: scalable but ad hoc.
   - Statistical methods: principled but *SLOW*.
   - We need to aim for the best of both worlds!
4. We are not alone: Massive data sets are ubiquitous.
5. Progress is being made!
   - **A. Lee:** Transform high-dim data to simpler form more amenable to standard analyses.
   - **A. Gray:** Speed up computation of many methods for application to massive surveys.
   - **J. Richards:** Automatic light curve classif'n.

Complex Data and Sophisticated Models

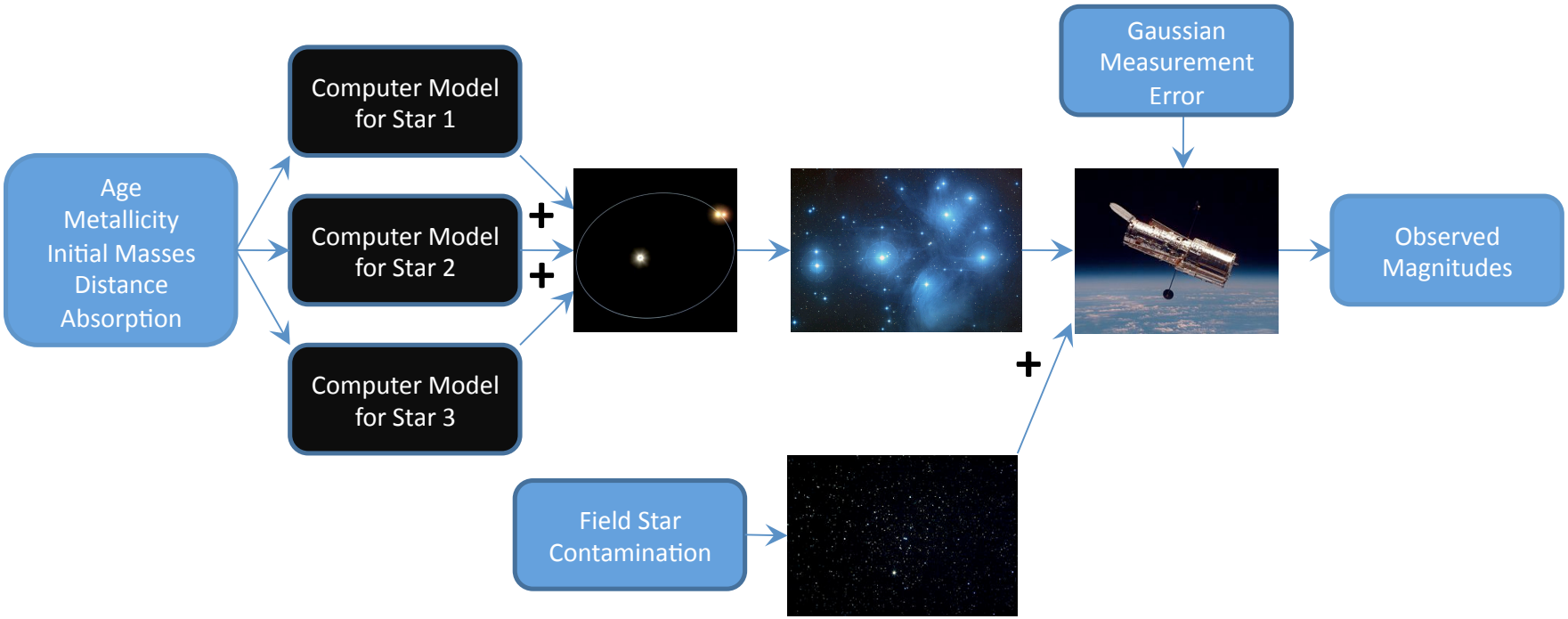
Just One Example...

- A great leap forward:
  Large Synoptic Survey Telescope
  (1.28 petabytes per year).
- *Not just massive:* rich & deep.
- LSST: Trigonometric Parallax, Proper Motion, and Photometric data in 5 bands.
- Rich data enables us to fit complex computer/simulation models.
- Echo Kirk Borne: “More data is not just more data: it is qualitatively different”

Complex Data and Sophisticated Models

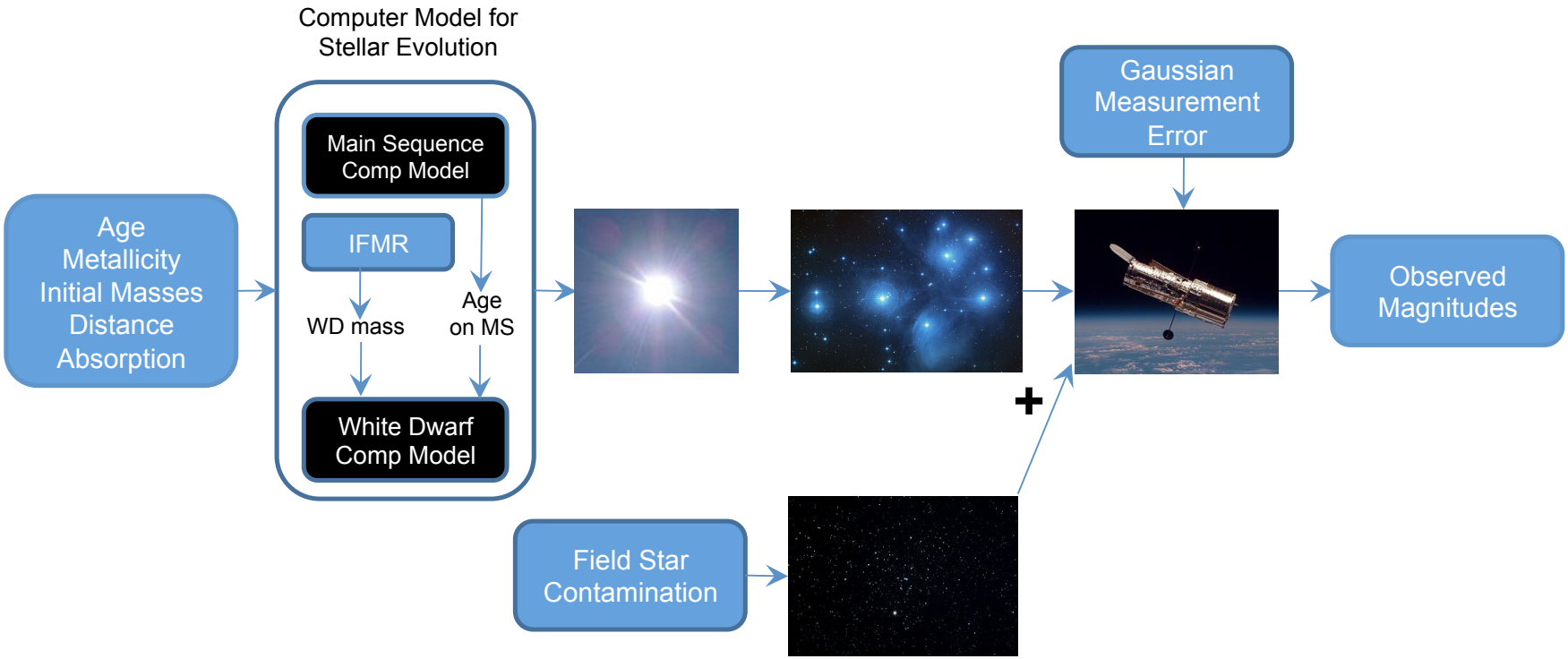
1. Complex computer models and simulations are taking the place of the analytical likelihood function.
2. Sophisticated data allows us to fit such models, but an entirely new set of methods are required.
   - **D. Higdon:** Gaussian process emulator of computer model.
   - **C. Schafer:** "Likelihood Free" / ABC: distribution of parameters that result in simulation close to observed data.
   - **V. Kashyap:** Uses PCA to analytically summarize calibration products generated with computer models.
   - **D. van Dyk/N. Stein:** Embedding computer models into multilevel model in a fully Bayesian setting.
3. **C. Graziani:** "This kind of computing is coming to many more areas of Astronomy"; "Challenge is acute when complex models are mixed with massive data."
4. Model fitting, model comparison, design, etc.
5. Future isn't with off-the-shelf methods or standard models.

Computer Models in a Principled Statistical Analysis



We embed computer models into a statistical likelihood function for a coherent analysis.

Improving Computer Models



1. Opening up the "black box" to improve the fit, handle errors properly, and improve understanding.
2. Treat computer models as any other component in a highly-structured multi-level model.

Integration of Statisticians into Astronomy

NEED statisticians, computer scientists, and methodologists to be a regular part of the process:

1. in design of data collection
2. in methodological development
3. in data analysis
4. funded by grants, hired as postdocs, brought in as faculty.

Models:

1. Econometricians, Psychometricians, and Biostatisticians reside in academic departments.
2. Data/Methods experts in Business, Engineering, and Biology departments.
3. And even in Astronomy: Imperial College London