Big Data, Big Challenges, and Big Ideas

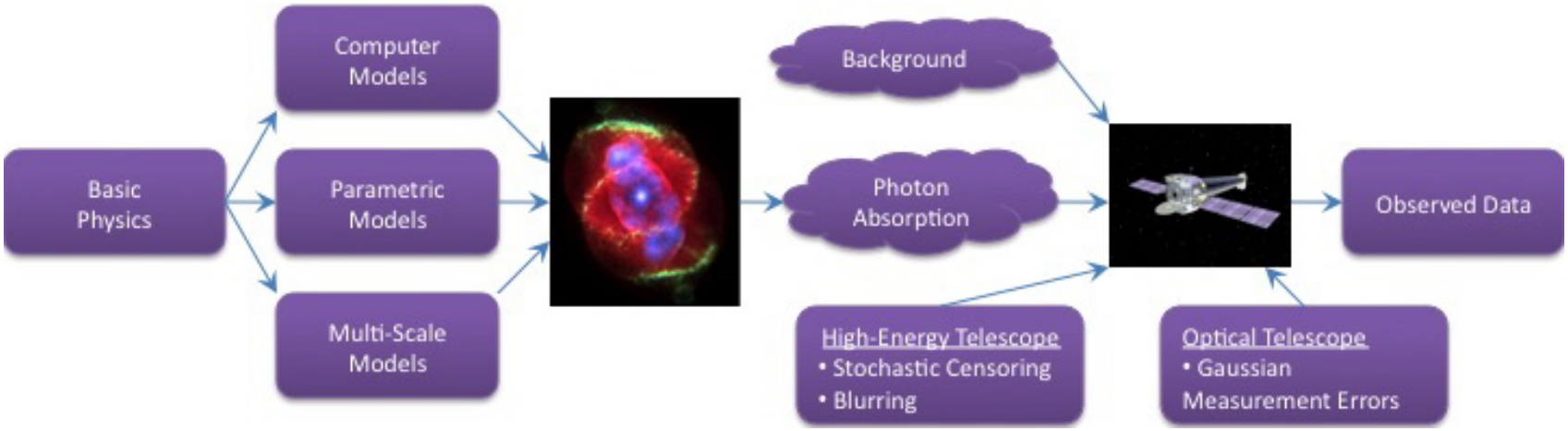
in 21st Century Astrostatistics

David A. van Dyk

Statistics Section and Data Science Institute
Imperial College London

Imperial College London

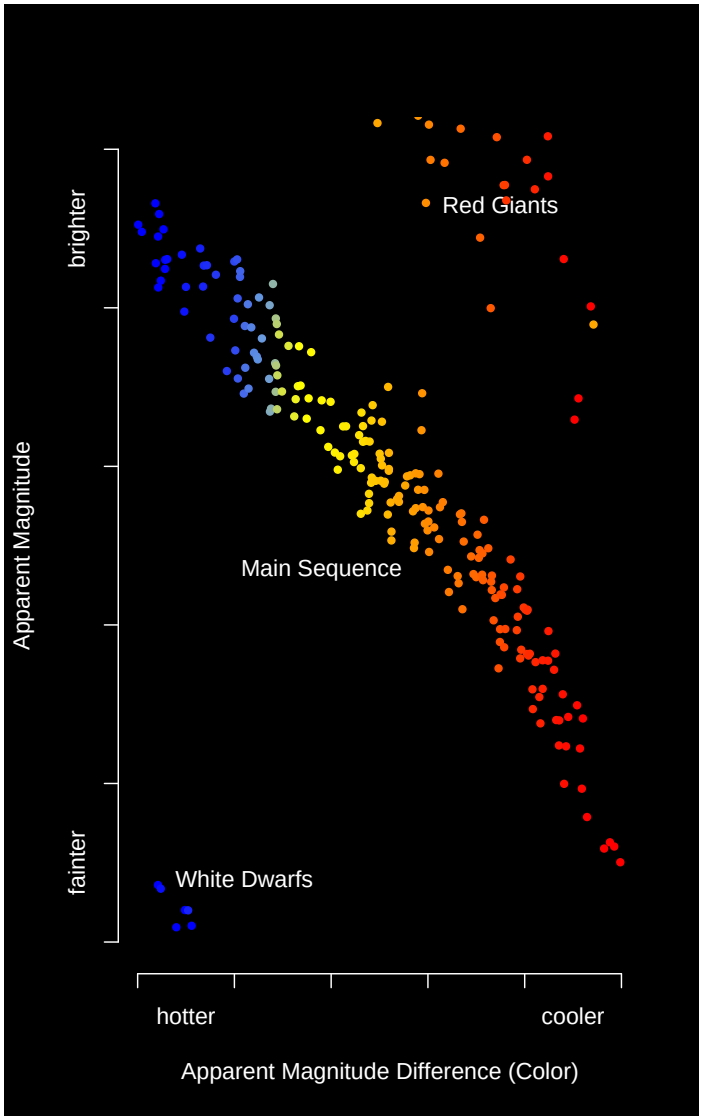
Complexity



- Complex physics, complex models, complex instruments, and complex questions.
- New reliance on state-of-the-art statistical methods.
- *Descriptive science-driven* versus *predictive* models.

Imperial College London

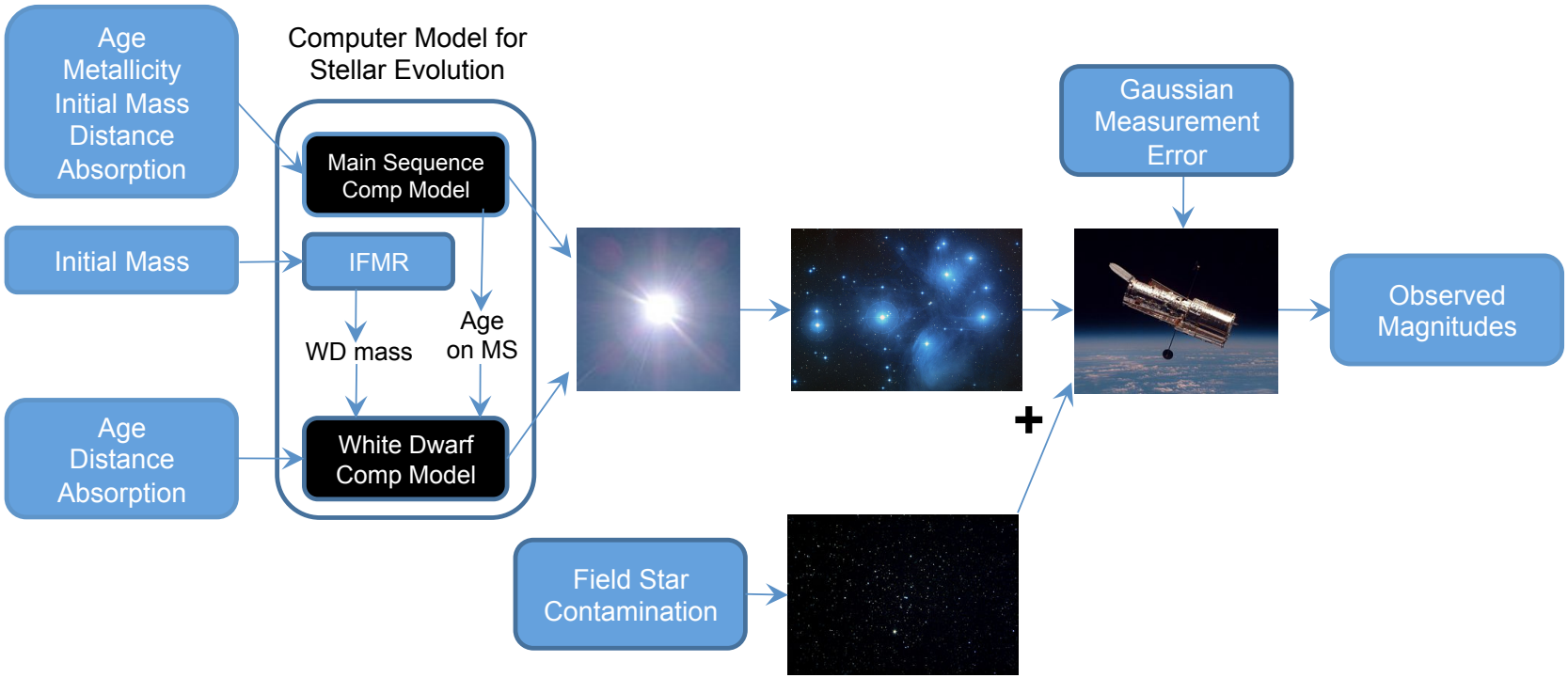
Computer Models



- Complex computer models and simulations are taking the place of the analytic models.
- E.g., to model
  - stellar evolution,
  - planetary, stellar atmospheres
  - reactions in interstellar clouds,
  - the emergence of galactic clusters and superclusters,
  - elements formed in Big Bang.
- Chi-by-eye fitting.
- Principled methods, fitted values, and error bars?

Imperial College London

Computer Models



- Embedding ensembles of computer models into a principled multi-level Bayesian model.
- Challenge is acute when complex models are combined with massive data streams.

Imperial College London

Massive ("BIG") Data

- A great leap forward:
  Large Synoptic Survey Telescope (1.28 petabytes/year).
- Data are *not just massive:*
  they are rich, deep, & complex.
- Require specialized models, methods, and computation.
- Big computational challenges.
- Automated data collection and model fitting (photo Z).
- Science-driven data reduction.

Imperial College London

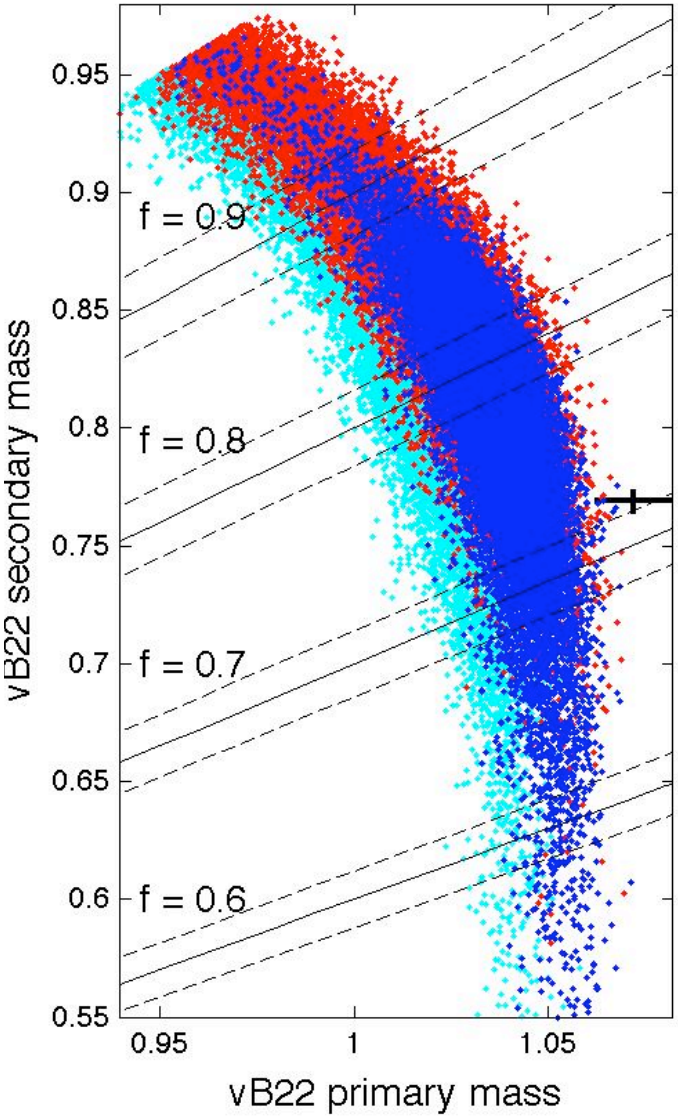
Massive ("BIG") Data

New instruments: *More data than we can analyze!*

Should more resources be devoted to computational facilities and methodological development?

Imperial College London

Systematic Errors

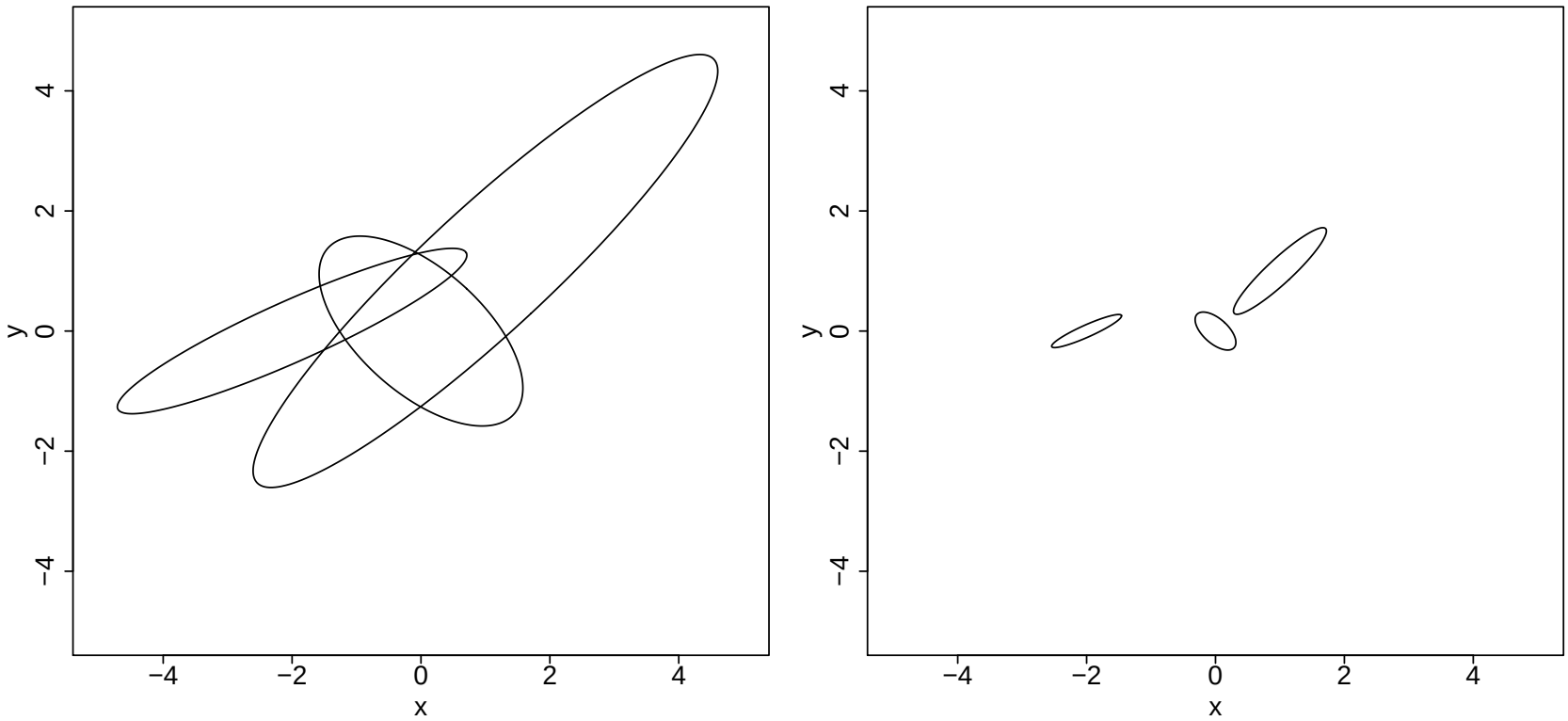


- More and more data sources can be used to compare results.
- With ever larger datasets, systematic errors may dominate statistical errors.
- Inconsistencies may appear between instruments and/or data sources.

Errors are easier to see, but still difficult to correct!

Imperial College London

Systematic Errors



As datasets grow, systematic errors swamp statistical errors and new disparities appear.

Imperial College London