Statistical challenges in modern astronomy

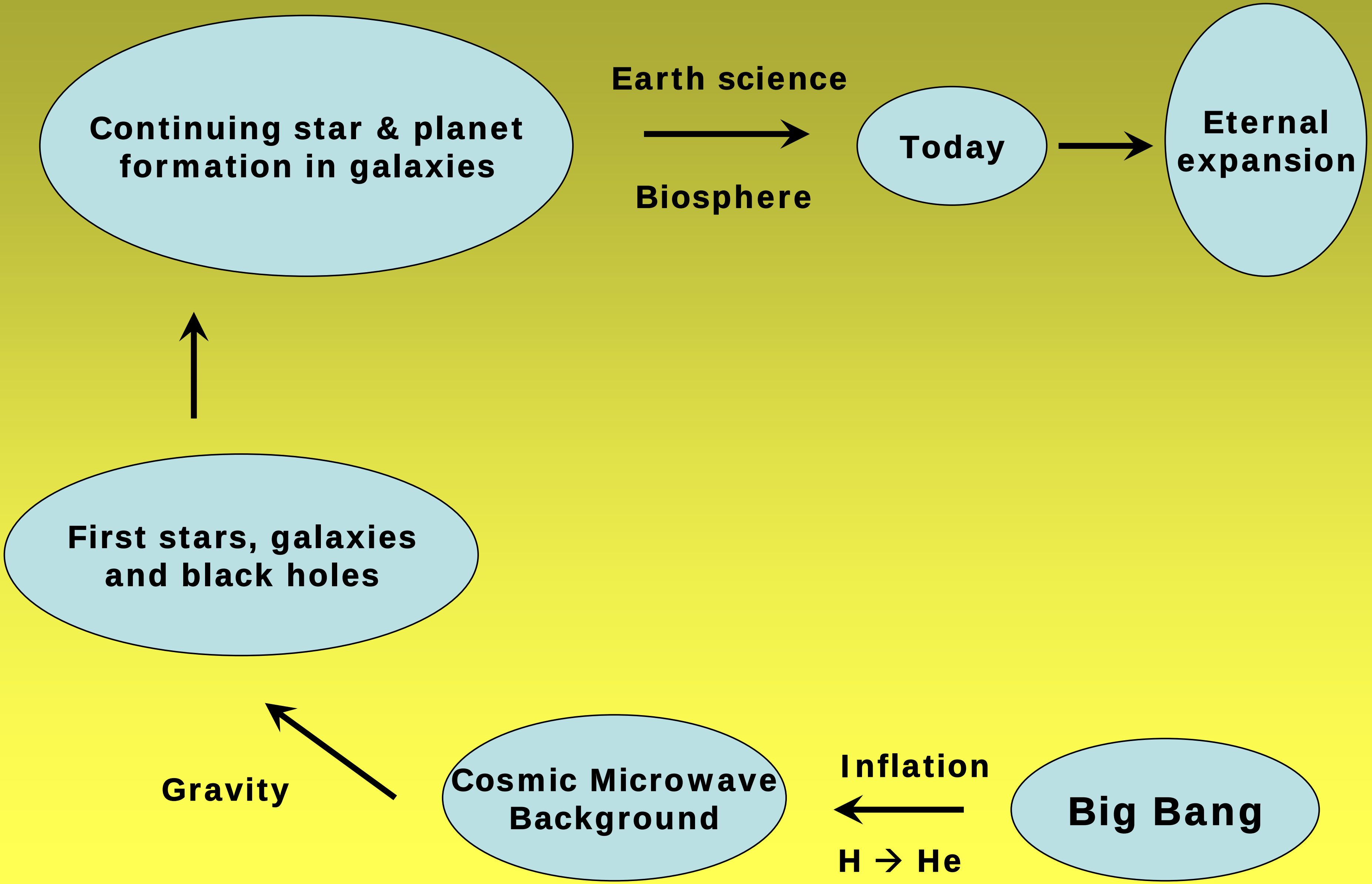
Eric Feigelson (Astro & Astrophys)

&

Jogesh Babu (Stat)

Penn State University

Overview of modern astronomy & astrophysics



Continuing star & planet formation in galaxies

Earth science

Biosphere

Today

Eternal expansion

First stars, galaxies and black holes

Gravity

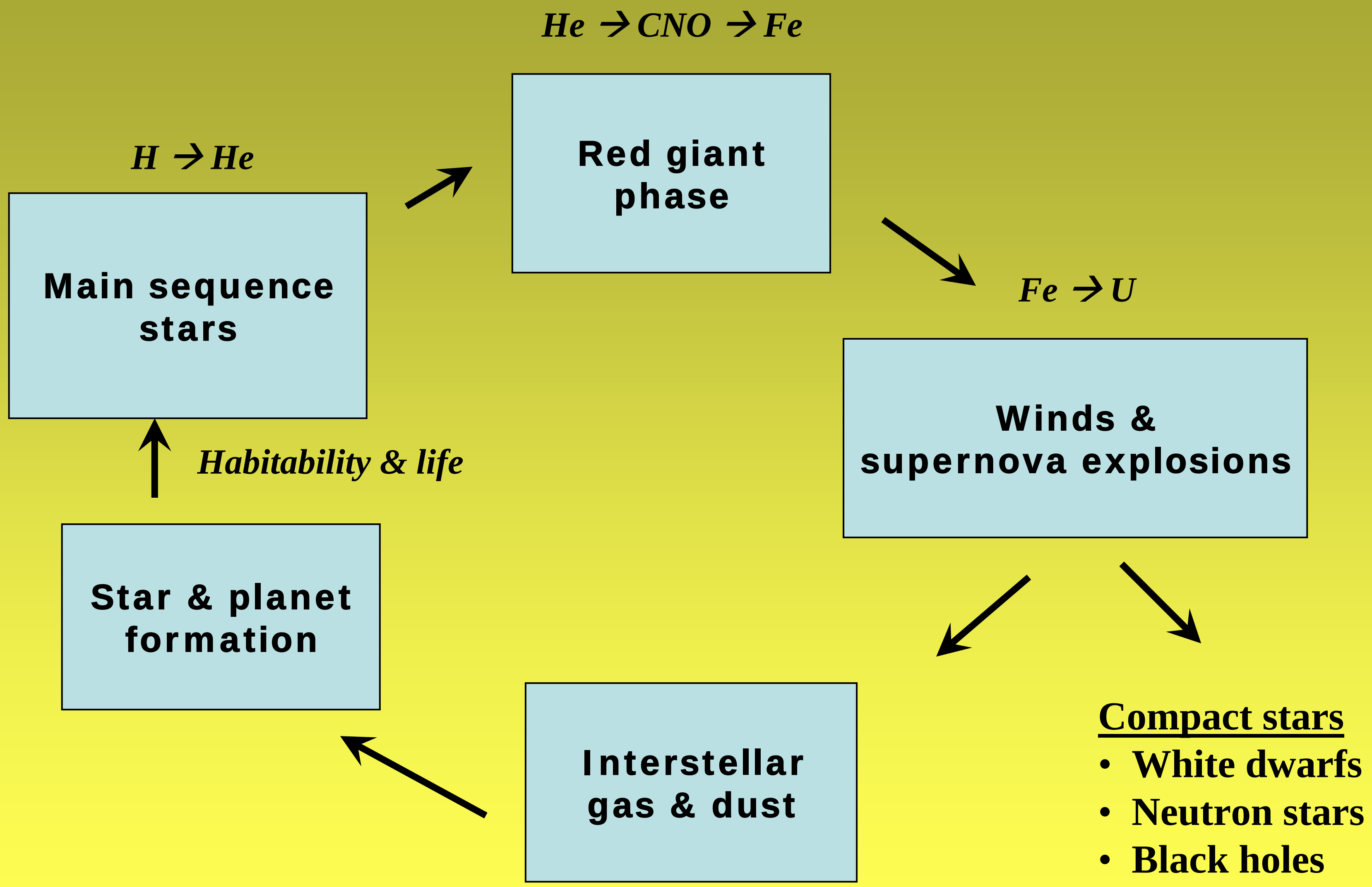
Cosmic Microwave Background

Inflation

H → He

Big Bang

Lifecycle of the stars



H → He

Main sequence stars

He → CNO → Fe

Red giant phase

Fe → U

Winds & supernova explosions

Compact stars

- White dwarfs
- Neutron stars
- Black holes

Interstellar gas & dust

Star & planet formation

Habitability & life

Astronomy & statistics: A glorious history

Hipparchus (4th c. BC): **Average via midrange of observations**

Galileo (1572): **Average via mean of observations**

Halley (1693): **Foundations of actuarial science**

Legendre (1805): **Cometary orbits via least squares regression**

Gauss (1809): **Normal distribution of errors in planetary orbits**

Quetelet (1835): **Statistics applied to human affairs**

But the fields diverged in the late 19-20th centuries,
astronomy → astrophysics (EM, QM)
statistics → social sciences & industries

Do we need statistics in astronomy today?

- **Are these stars/galaxies/sources an unbiased sample of the vast underlying population?**
- **When should these objects be divided into 2/3/… classes?**
- **What is the intrinsic relationship between two properties of a class (especially with confounding variables)?**
- **Can we answer such questions in the presence of observations with measurement errors & flux limits?**

Do we need statistics in astronomy today?

- **Are these stars/galaxies/sources an unbiased sample of the vast underlying population? Sampling**
- **When should these objects be divided into 2/3/… classes? Multivariate classification**
- **What is the intrinsic relationship between two properties of a class (especially with confounding variables)? Multivariate regression**
- **Can we answer such questions in the presence of observations with measurement errors & flux limits? Censoring, truncation & measurement errors**

- **When is a blip in a spectrum, image or datastream a real signal? Statistical inference**

- **How do we model the vast range of variable objects (extrasolar planets, BH accretion, GRBs, …)?**
  Time series analysis

- **How do we model the 2-6-dimensional points representing galaxies in the Universe or photons in a detector?**
  Spatial point processes & image processing

- **How do we model continuous structures (CMB fluctuations, interstellar/intergalactic media)?**
  Density estimation, regression

How often do astronomers need statistics?

(a bibliometric measure)

Of ~15,000 refereed papers annually:

1% have *`statistics'* in title or keywords

5% have *`statistics'* in abstract

10% treat variable objects

5-10% (est) analyze data tables

5-10% (est) fit parametric models

The state of astrostatistics today

The typical astronomical study uses:

- Fourier transform for temporal analysis (Fourier 1807)
- Least squares regression (Legendre 1805, Pearson 1901)
- Kolmogorov-Smirnov goodness-of-fit test (Kolmogorov, 1933)
- Principal components analysis for tables (Hotelling 1936)

Even traditional methods are often misused:

- Six unweighted bivariate least squares fits are used interchangeably in H_o studies with wrong confidence intervals
  Feigelson & Babu ApJ 1992
- Likelihood ratio test (F test) usage typically inconsistent with asymptotic statistical theory
  Protassov et al. ApJ 2002

But astrostatistics is an emerging discipline

- We organize cross-disciplinary conferences at Penn State *Statistical Challenges in Modern Astronomy (1991, 1996, 2001)*

- Fionn Murtagh & Jean-Luc Starck run methodological meetings & write monographs

- Alanna Connors runs statistics sessions as AAS meetings & we run astronomy sessions at JSM/ISI meetings

- Powerful astro-stat collaborations appearing in the 1990s:
  - Harvard/Smithsonian (David van Dyk, Chandra scientists, students)
  - CMU/Pitt = PICA (Larry Wasserman, Chris Genovese, Bob Nichol, …)
  - NASA-ARC/Stanford (Jeffrey Scargle, David Donoho)
  - Efron/Petrosian, Berger/Jeffreys/Loredo/Connors, Stark/GONG, …

A new imperative: Virtual Observatory

Huge, uniform, multivariate databases are emerging from specialized survey projects & telescopes:

- **10^9-object catalogs from USNO, 2MASS & SDSS opt/IR surveys**
- **10^6- galaxy redshift catalogs from 2dF & SDSS**
- **10^5-source radio/infrared/X-ray catalogs**
- **10^{3-4}-samples of well-characterized stars & galaxies with dozens of measured properties**
- **Many on-line collections of 10^2-10^6 images & spectra**
- **Planned Large-aperture Synoptic Survey Telescope will generate ~10 Pby**

The Virtual Observatory is an international effort underway to federate these distributed on-line astronomical databases.

Powerful statistical tools are needed to derive scientific insights from extracted VO datasets
(NSF FRG involving PSU/CMU/Caltech)

Some methodological challenges for astrostatistics in the 2000s

- Simultaneous treatment of measurement errors and censoring (esp. multivariate)
- Statistical inference and visualization with very-large-N datasets too large for computer memories
- A user-friendly cookbook for construction of likelihoods & Bayesian computation of astronomical problems
- Links between astrophysical theory and wavelet coefficients (spatial & temporal)
- Rich families of time series models to treat accretion and explosive phenomena

Structural challenges for astrostatistics

Cross-training of astronomers & statisticians

New curriculum, summer workshops

Effective statistical consulting

Enthusiasm for astro-stat collaborative research

Recognition within communities & agencies

More funding (astrostat gets <0.1% of astro+stat)

Implementation software

StatCodes Web metasite (www.astro.psu.edu/statcodes)

Standardized in R, MatLab or VOStat? (www.r-project.org)

Inreach & outreach

A Center for Astrostatistics to help attain these goals