Challenges of Modern Empirical Astrophysics

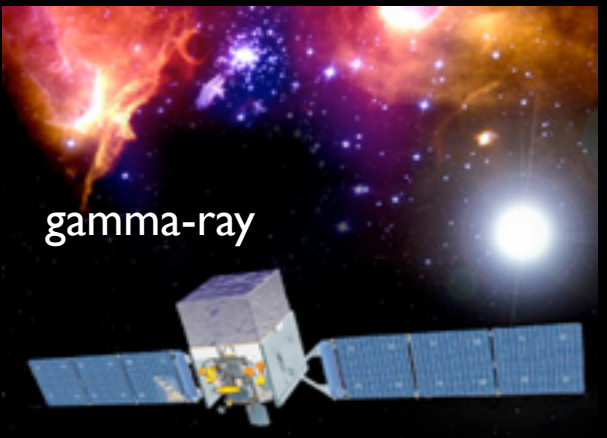
Prajval Shastri

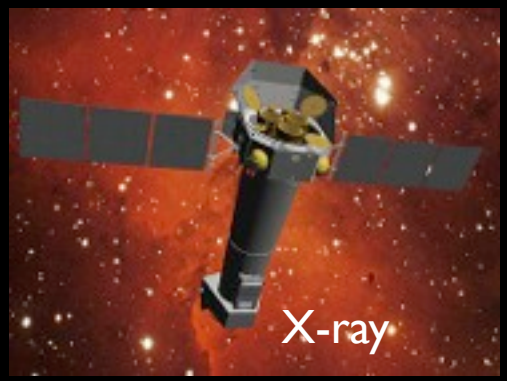
Indian Institute of Astrophysics

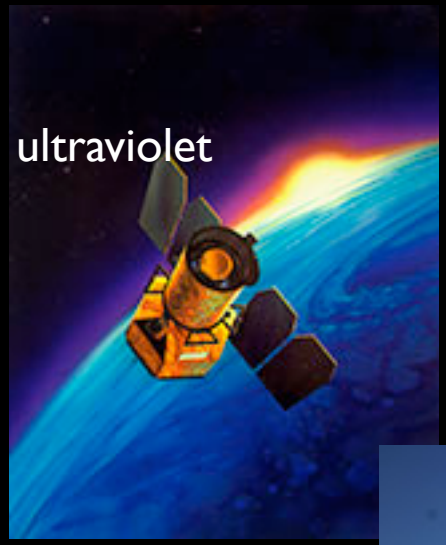
Astrophysical Data:

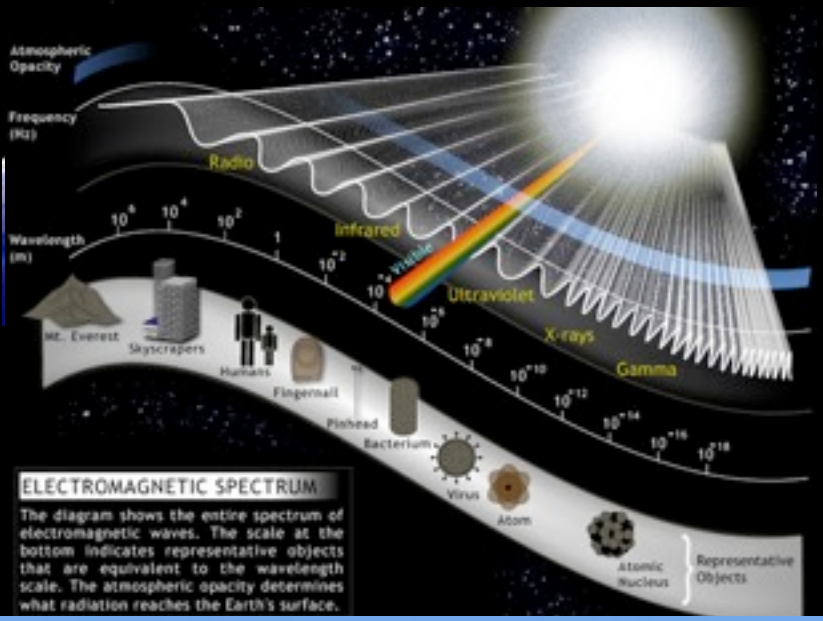
- *Position*
- *Intensity*
- *Polarisation*

- *Multi-epoch data -> measuring change -> Time Series*
- *Know the light wavelength -> Spectroscopy*
- *Multi-frequency: "broad brush spectroscopy"*


gamma-ray


X-ray


ultraviolet


Atmospheric Opacity
Frequency (Hz)
Wavelength (m)
Radio
Infrared
Visible
Ultraviolet
X-rays
Gamma
Mt. Everest
Skyscrapers
Humans
Fingernail
Pinhead
Bacterium
Virus
Atom
Atomic Nucleus
Representative Objects
ELECTROMAGNETIC SPECTRUM
The diagram shows the entire spectrum of electromagnetic waves. The scale at the bottom indicates representative objects that are equivalent to the wavelength scale. The atmospheric opacity determines what radiation reaches the Earth's surface.

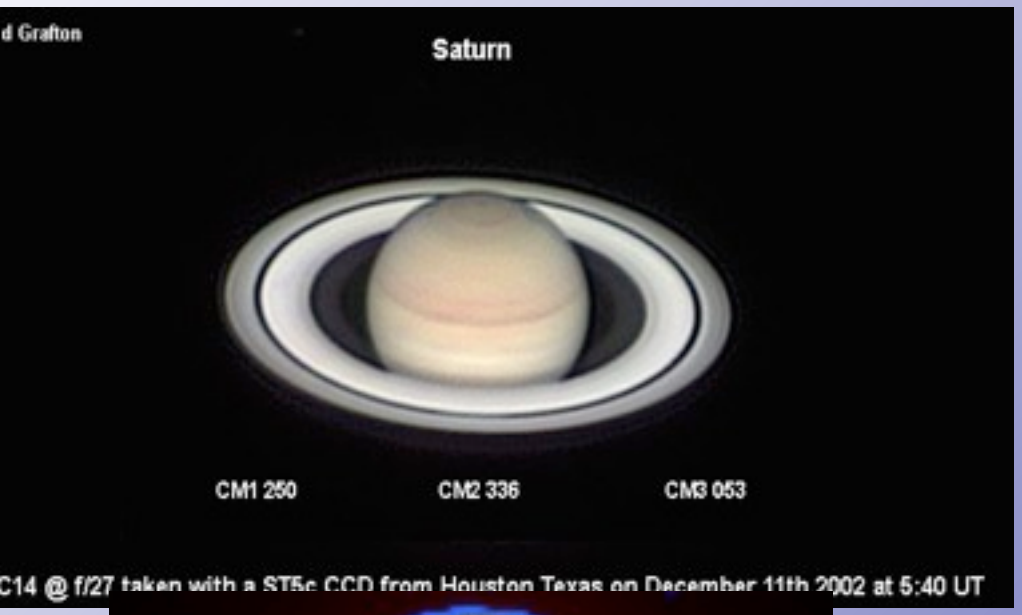

Infrared
Nasa/JPL-Caltech

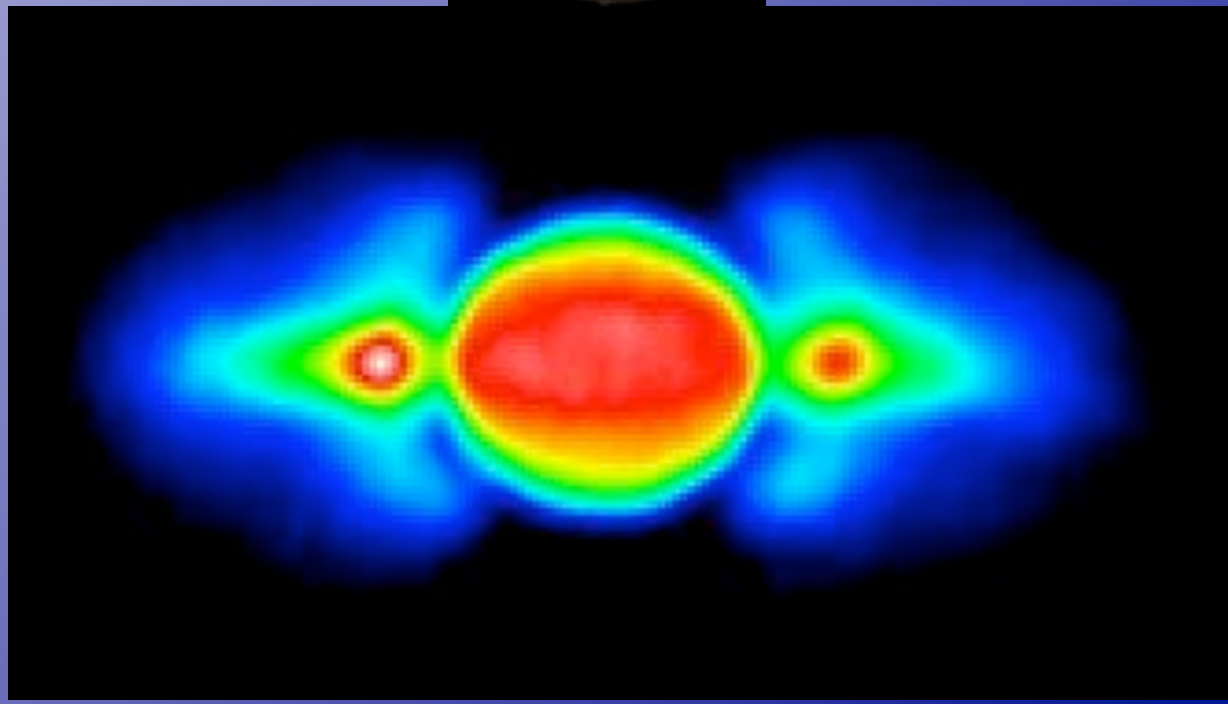

Radio

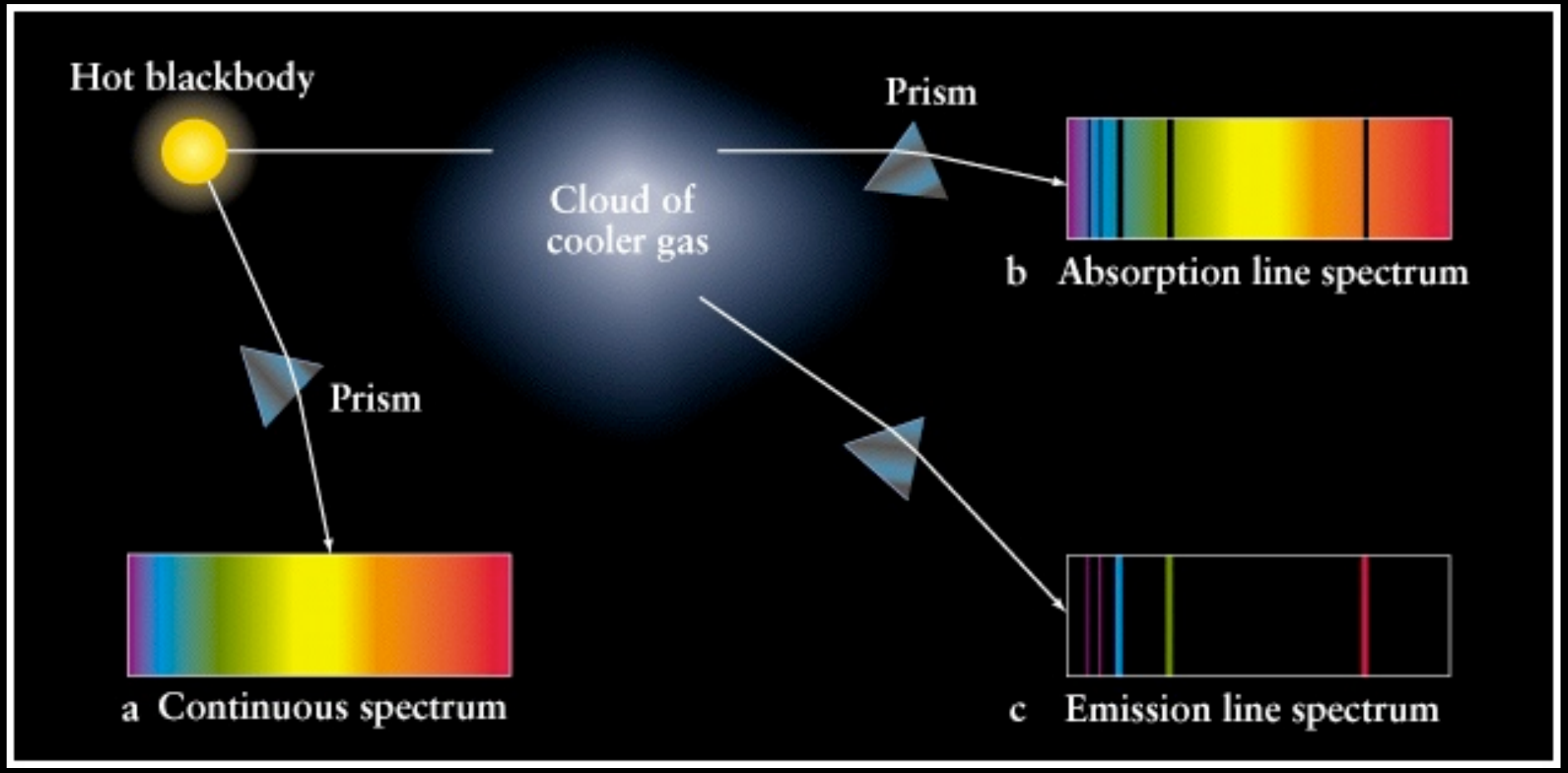
Astrophysical Data:

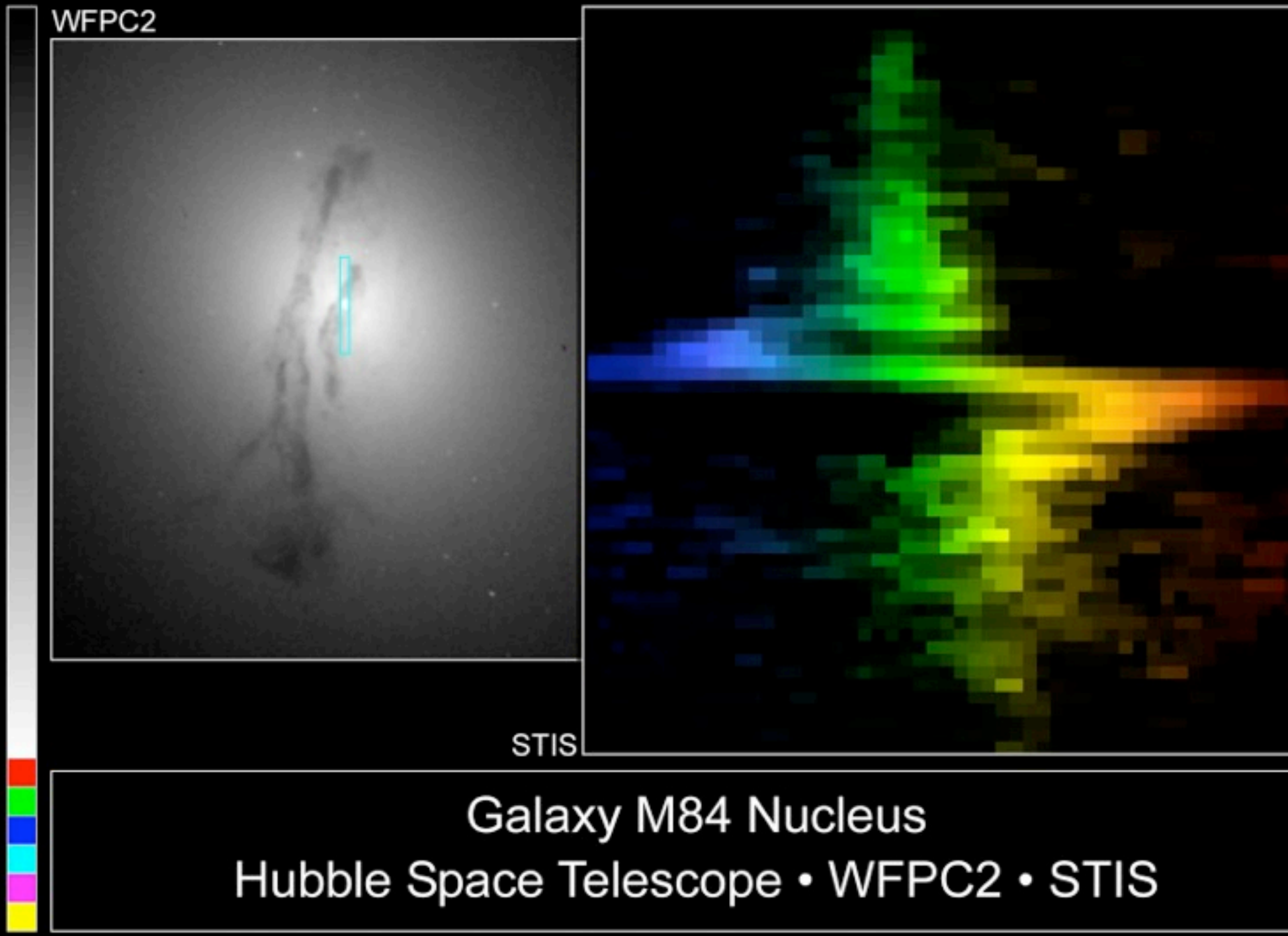
- *Position*
- *Intensity*
- *Polarisation*
  - *Multi-epoch data -> measuring change -> Time Series*
  - *Know the light wavelength -> Spectroscopy*
  - *Multi-frequency Astronomy: "broad brush spectroscopy"*

Electromagnetic wave astrophysics
+ Neutrino astrophysics
+ Gravitational wave astrophysics


d Grafton
Saturn
CM1 250
CM2 336
CM3 053
C14 @ f/27 taken with a ST5c CCD from Houston Texas on December 11th 2002 at 5:40 UT




Hot blackbody
Prism
Cloud of
cooler gas
b Absorption line spectrum
Prism
a Continuous spectrum
c Emission line spectrum



Galaxy M84 Nucleus
Hubble Space Telescope • WFPC2 • STIS

PRC97-12 • ST ScI OPO • May 12, 1997 • B. Woodgate (GSFC), G. Bower (NOAO) and NASA

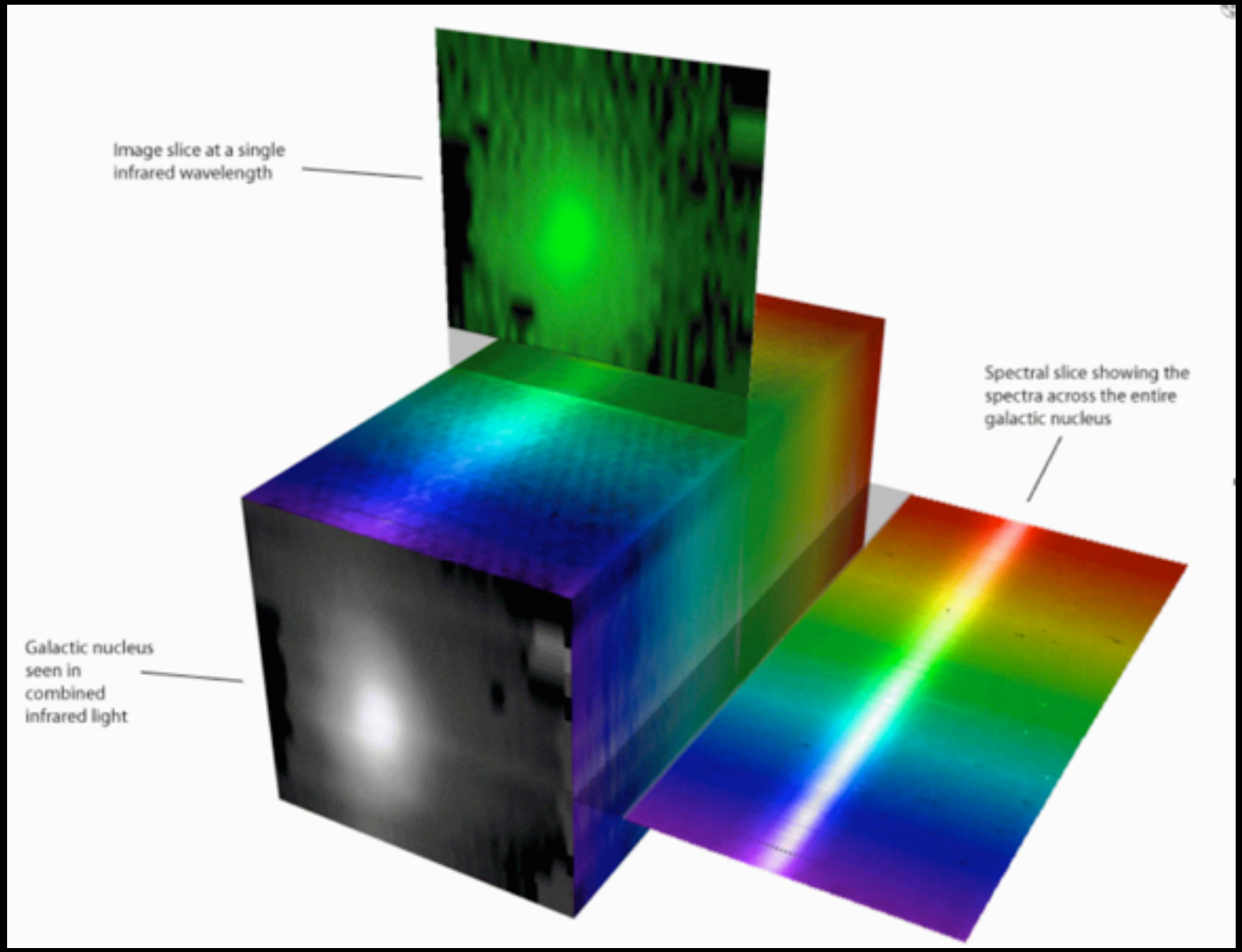


Image slice at a single infrared wavelength
Spectral slice showing the spectra across the entire galactic nucleus
Galactic nucleus seen in combined infrared light

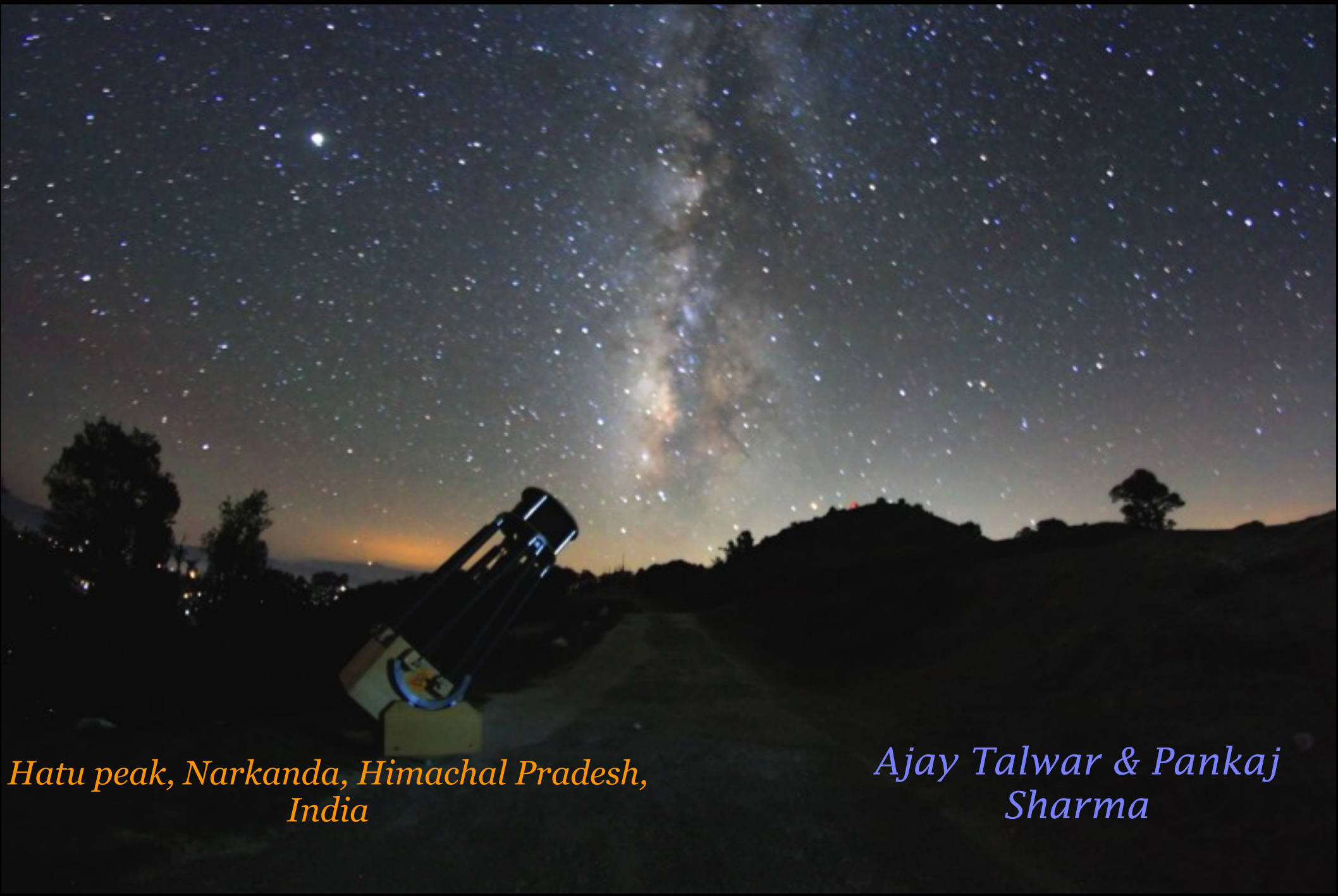
Astrophysicists *seek to understand the the universe...i.e., everything....*

and therefore also ourselves, because we are evolutes of the physical processes that occur on cosmological scales..

"the last bastion of the generalist"

- Ter Haar

ಚಂದ್ರ

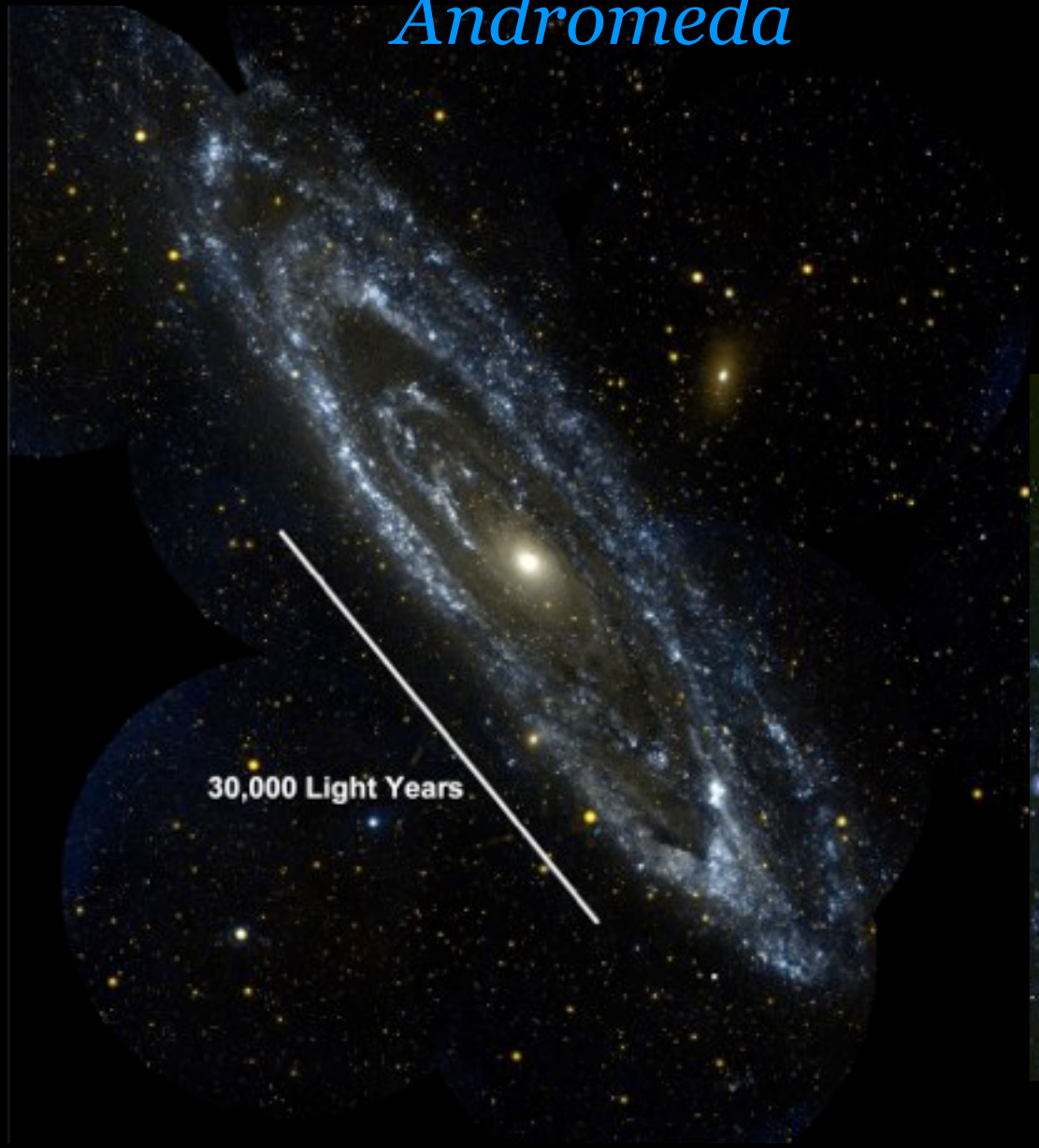


Hatu peak, Narkanda, Himachal Pradesh, India

Ajay Talwar & Pankaj Sharma

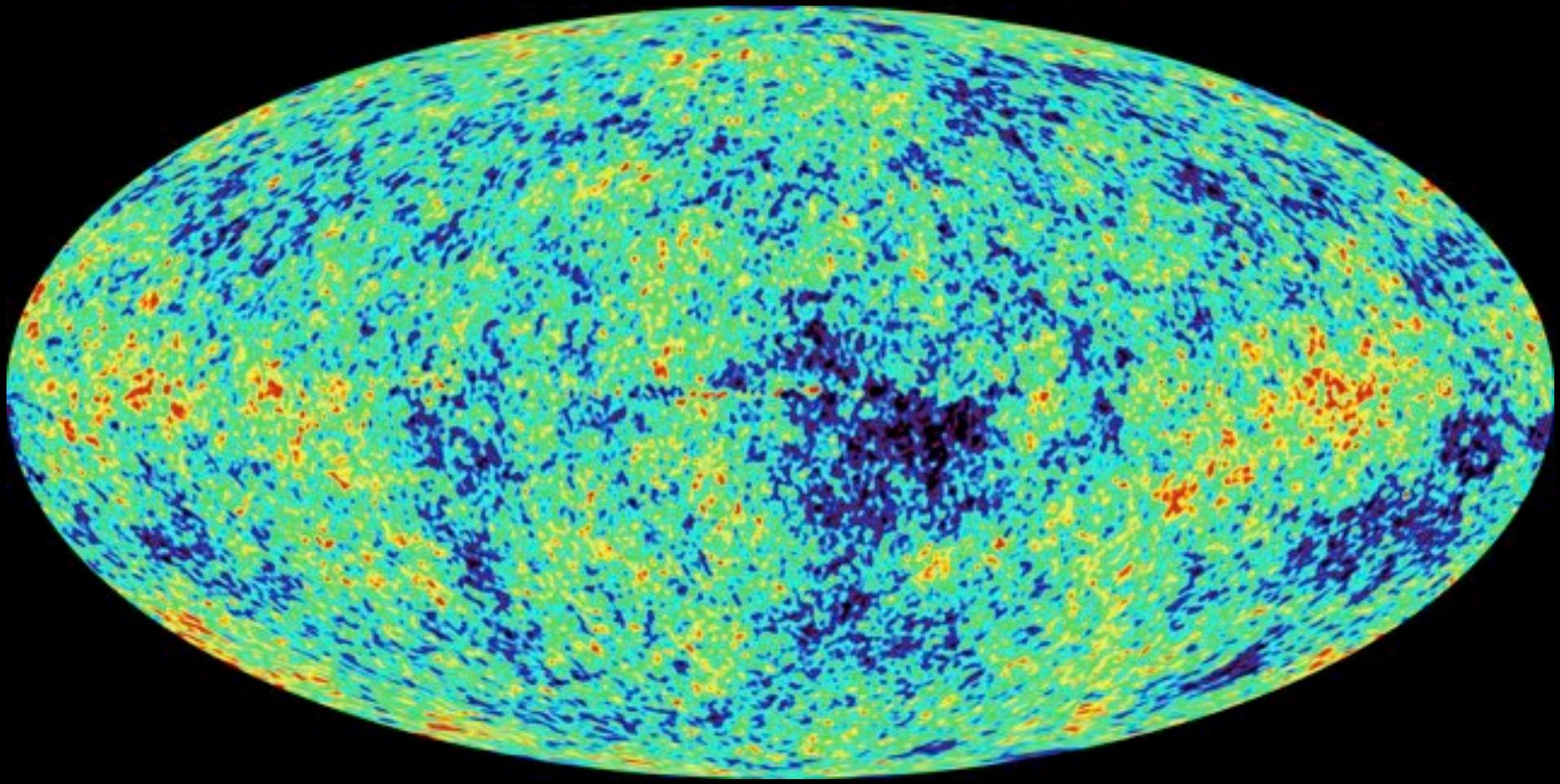
Spiral Galaxy NGC 4414

Andromeda



Messier 33

GALEX image



Cosmic Microwave Background

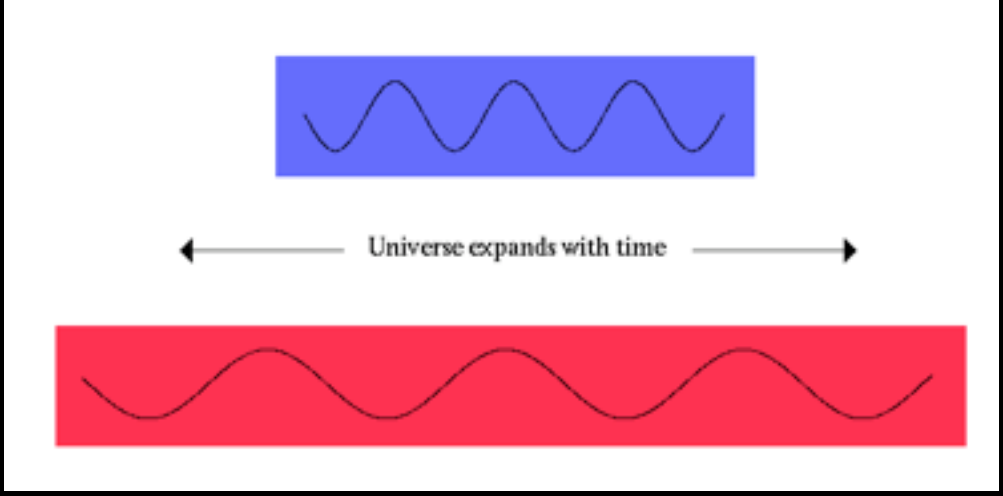
Nature's tools!

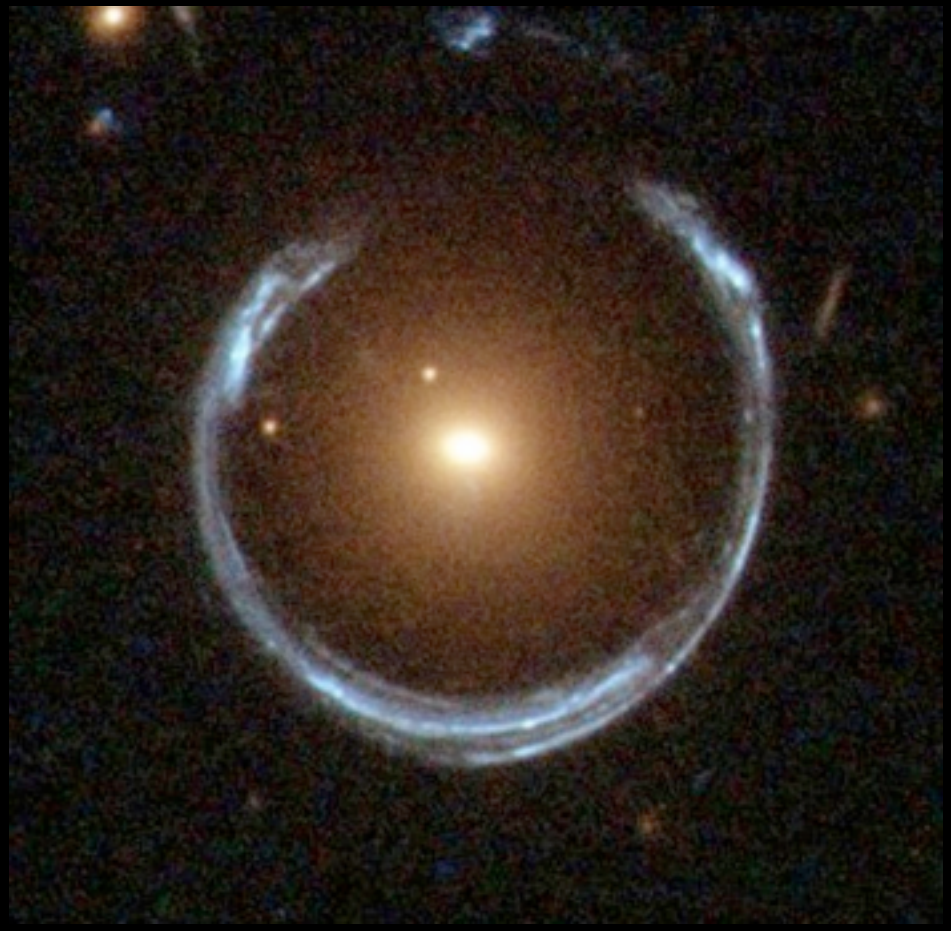
non-zero velocity of light :

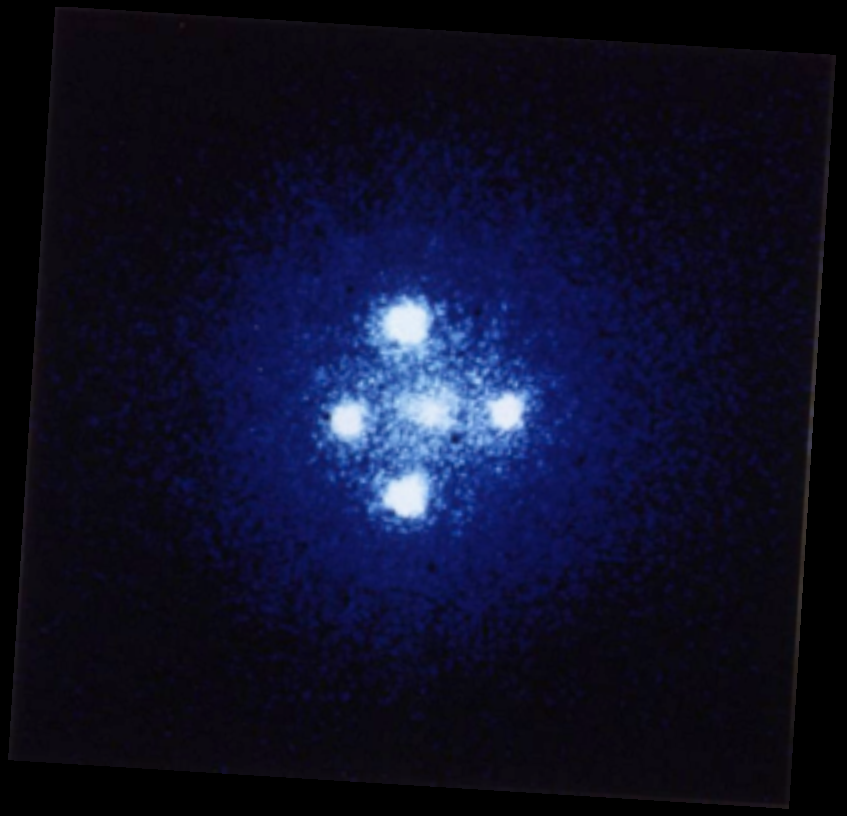
- *"There" is "here - then"*
- *Variability gives a limit on the size*
  (in the line of sight)

Expansion of the Universe:

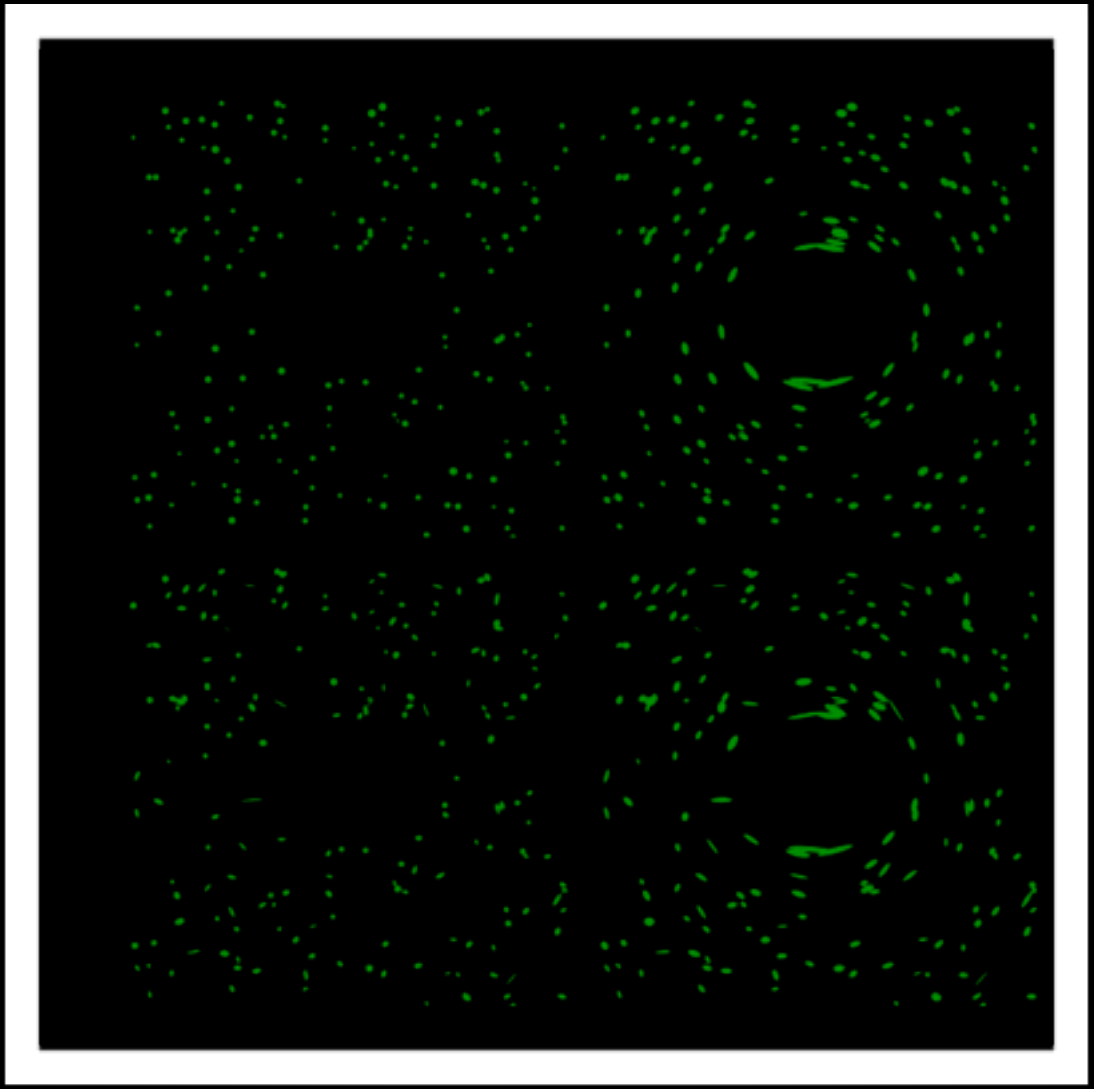
Cosmological Redshift:
Due to Stretching of Space

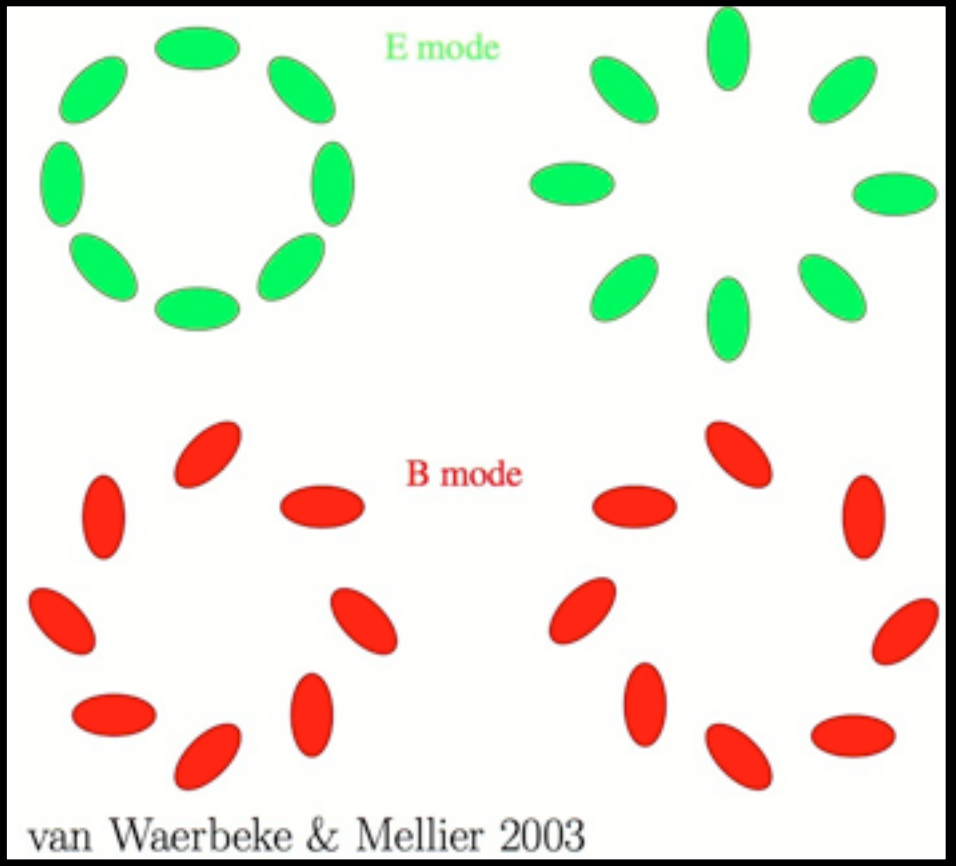


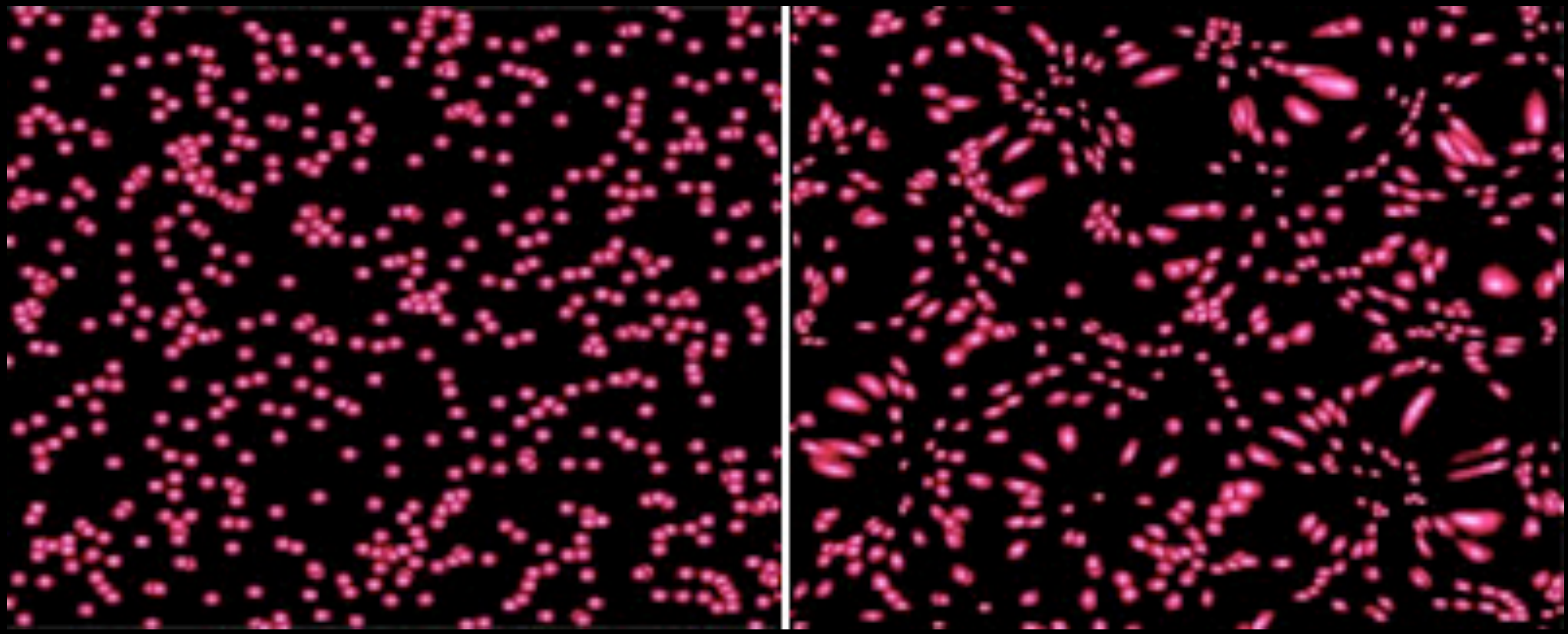




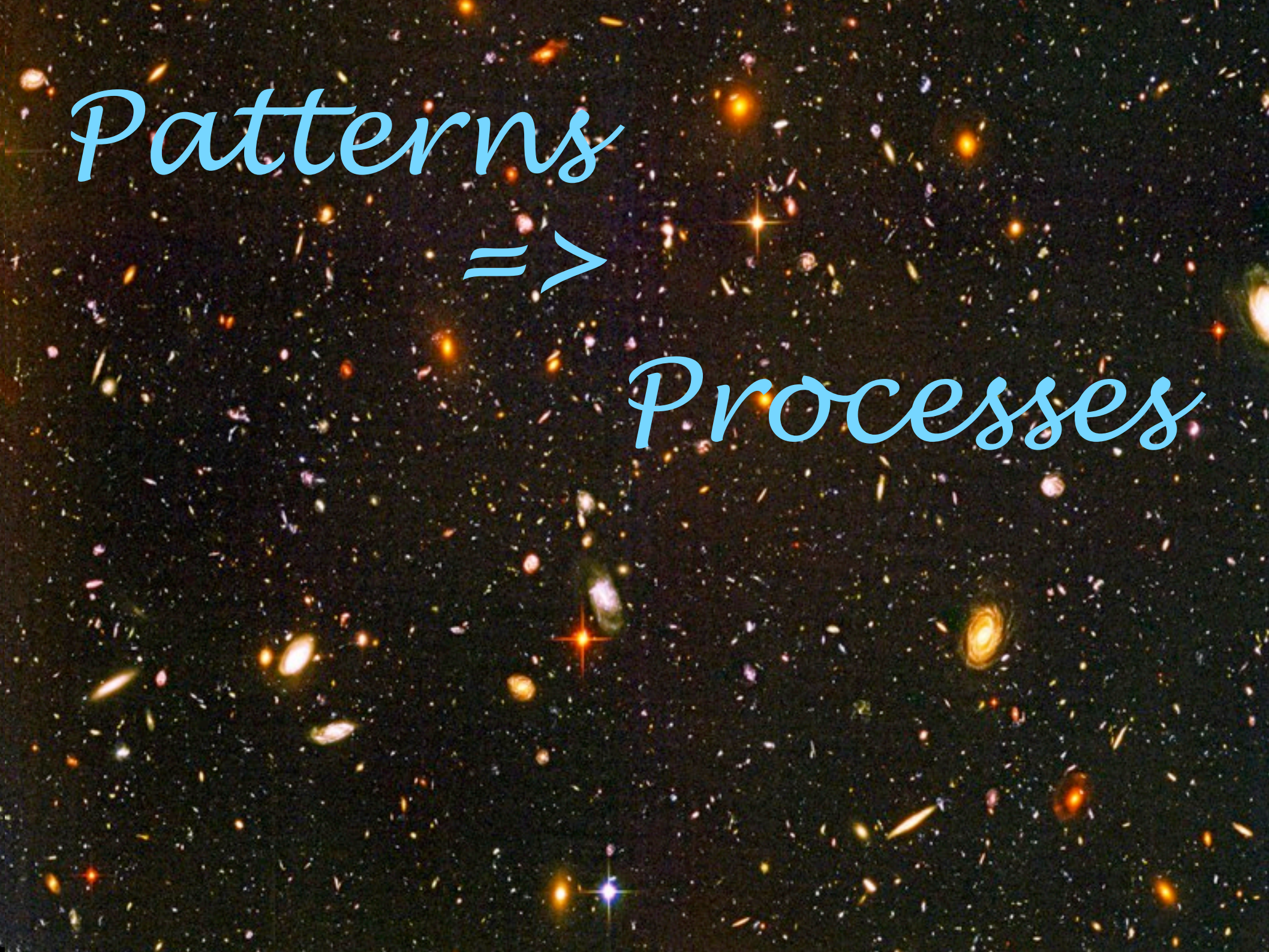
Horseshoe Einstein Ring: ESA/Hubble and NASA



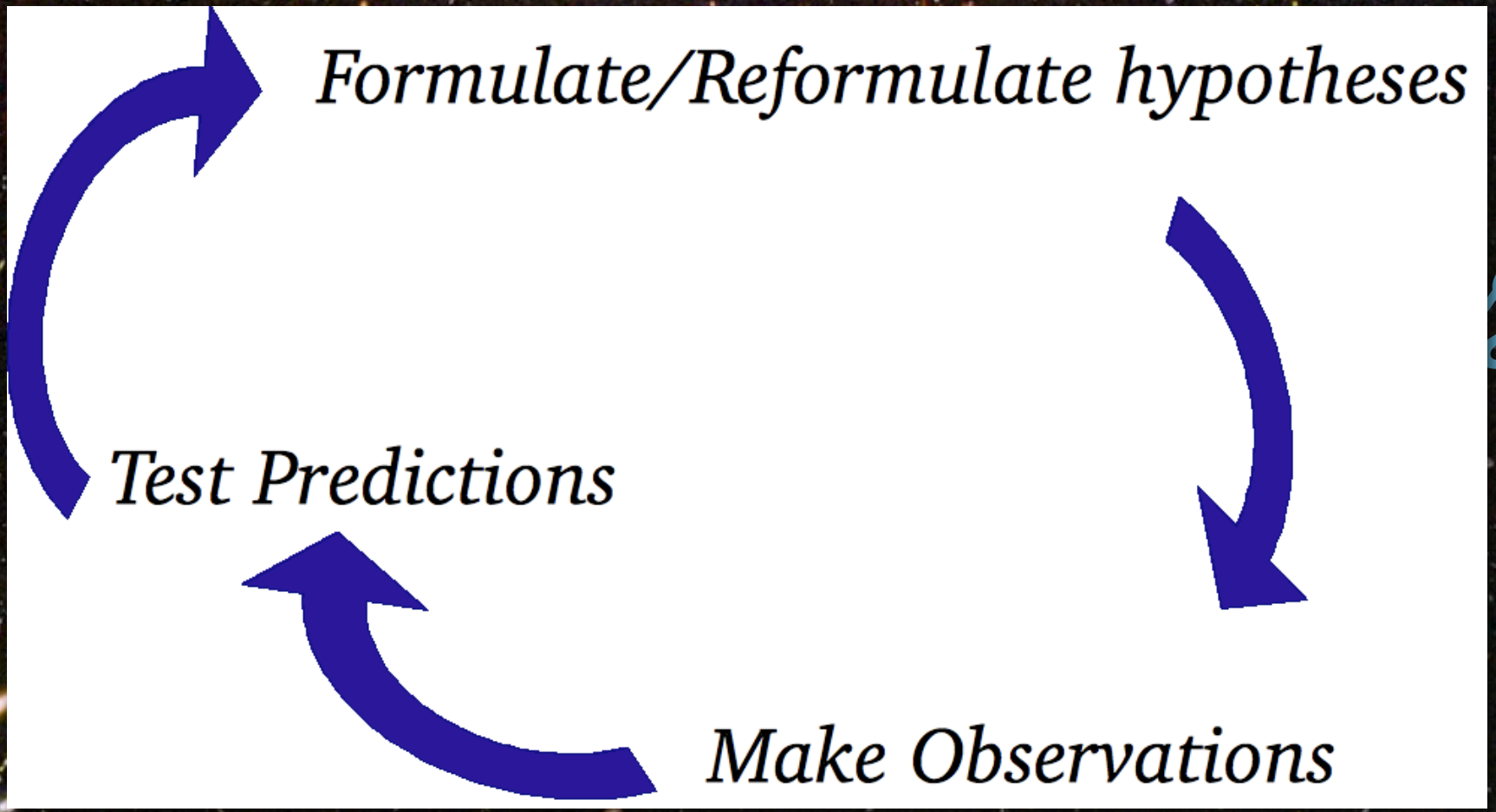

E mode
B mode
van Waerbeke & Mellier 2003



Credit: :LSST


Patterns
=>
Processes

Patterns


Formulate/Reformulate hypotheses
Make Observations
Test Predictions

- *Rigour in data analysis*
- *Multiwavelength measurements (increase in the dimensions)*
- *Increase in amount of data makes automation inevitable*

Rigour in data reductions & analysis

- *When is a blip in a spectrum, image or data stream a real signal?*
- *Are these stars/galaxies/sources an unbiased sample of the vast underlying population?*
- *When should these stars/galaxies/sources 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 when our data have measurement errors & flux limits?*
- *How is the (very common) variability in stars/galactic nuclei etc. to be modelled?*

Rigour in data reductions & analysis

- *When is a blip in a spectrum, image or data stream a real signal?* ***Statistical Inference***
- *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, Principal Component Analysis***
- *How is the variability in stars or galactic nuclei to be modelled?* ***Time Series Analysis***
- *Can we answer such questions when our data have measurement errors & flux limits?* ***Censoring, Truncation & Measurement Errors***

- **Maximum Entropy Method in imaging**
  Gull & Skilling 1984

seeks to extract as much information from a measurement as is justified by the data's signal-to-noise ratio

- **Two-point correlation function for galaxies**
  Bhavsar 1990
  - The data points are pairs of galaxies (ie galaxy co-ordiates in the sky), and to take into account the fact that the /N error bars are not independent, the bootstrap methodology is applied

- **Concept similar to Mahalanobis Distance in object detection**
  Babu, Mahabal, Djorgovski, Williams 2008

gives a very robust object detection technique that is capable of detecting faint sources especially in multi-epoch frames, i.e., even those objects that are not visible at all epochs (which would normally be smoothed out by traditional methods)

- **Oscillation Analysis of Solar Corona**
  Gissot & Hochedez 2008

ability of a motion estimation algorithm to explore and analyse the oscillating motions of coronal loops present in extreme Ultraviolet image sequences, using Morlet wave analysis.

- **Nonparametric Inference for the Cosmic Microwave Background**

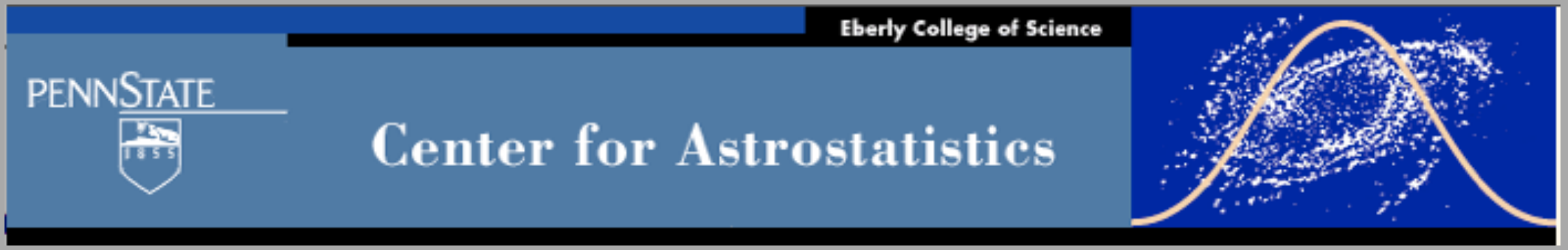
  Genovese, Miller, Nichol, Arjunwadkar & Wasserman 2004

  - construction of non-parametric confidence set for the unknown Cosmic Microwave Background Spectrum, to give an estimated spectrum based on minimal assumptions, leading to a wide range of additional inferences in addition to those similar to the cosmologists' model-based estimates.

- **Image reconstruction with error estimates**

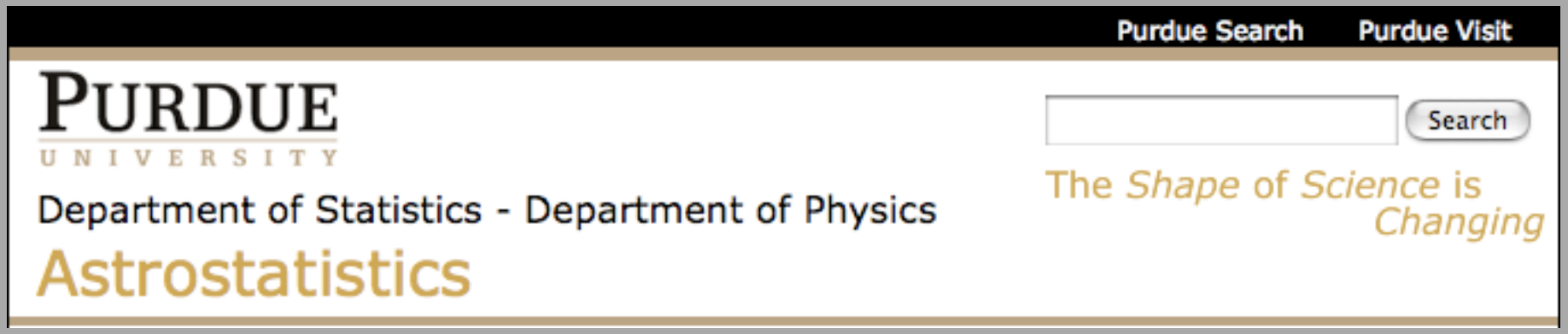
  van Dyk, Connors, Esch, et al 2007

explicitly model the complexities of both astronomical sources and the data generation mechanisms inherent in new high-tech instruments, i.e., non-uniform stochastic censoring, heteroscedastic errors in measurement, and background contamination.

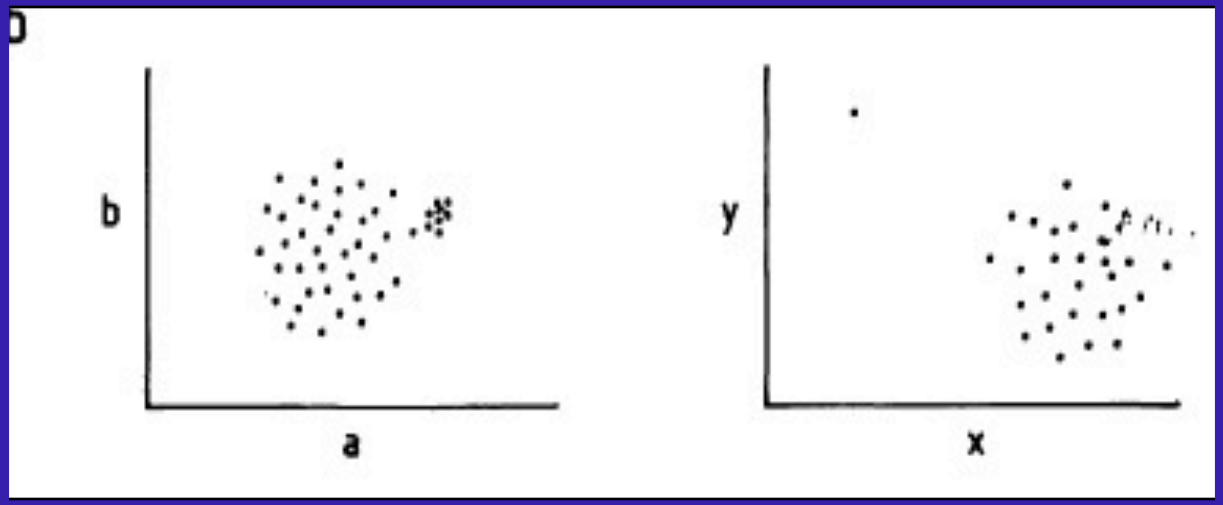


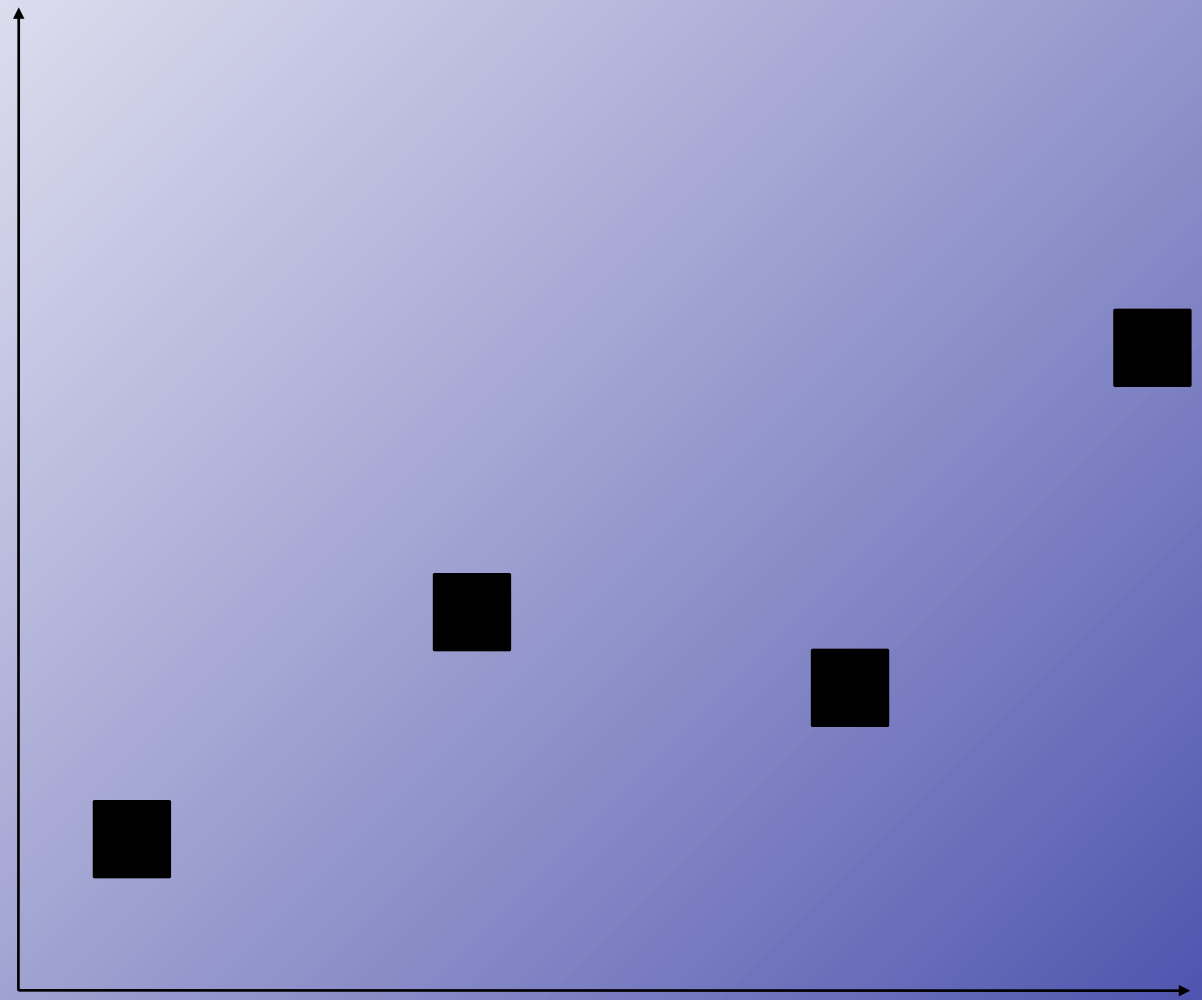
CHASC Astro-Statistics Collaboration

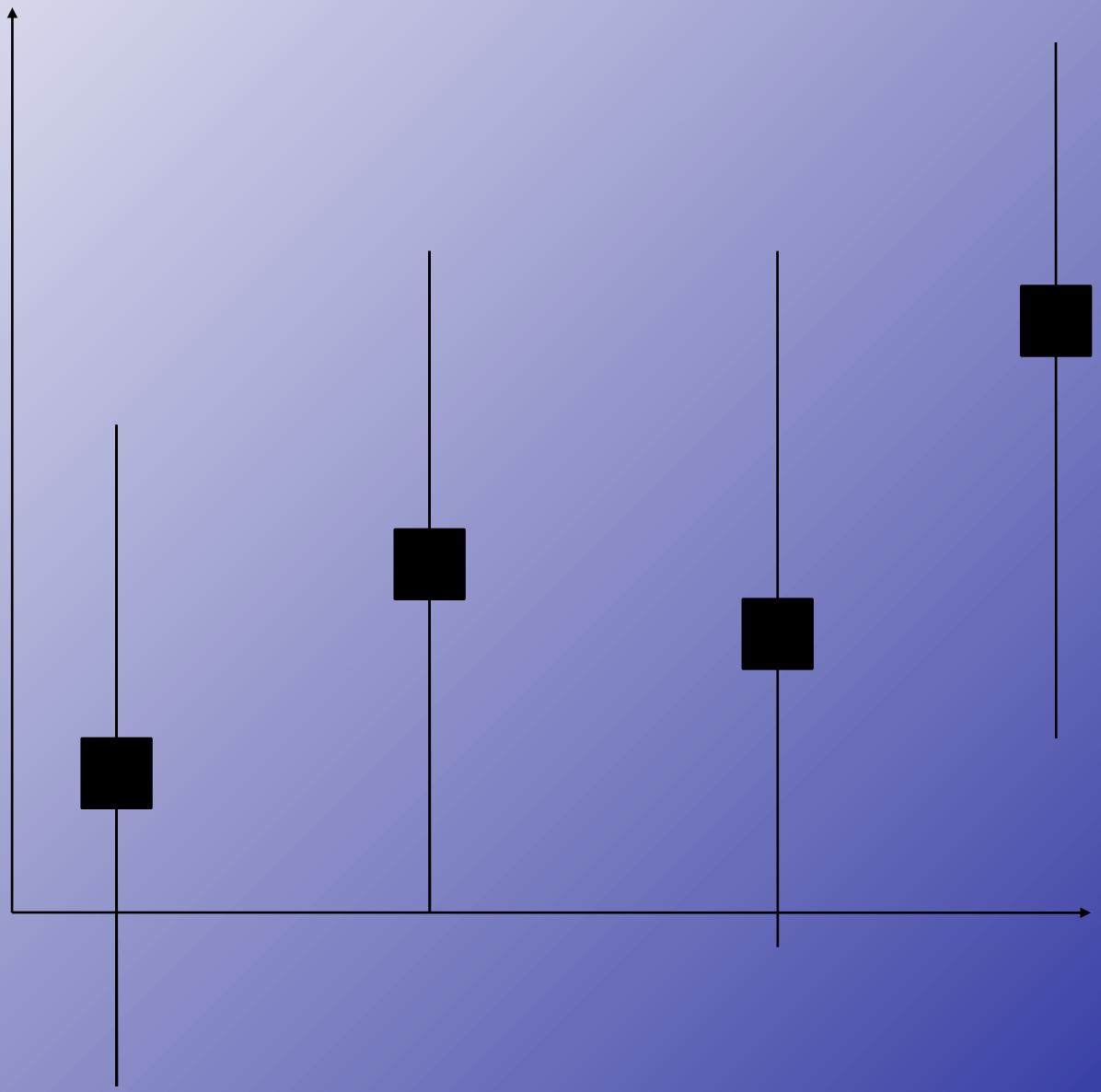
(California/Harvard/ASC AstroStatistics Collaboration)

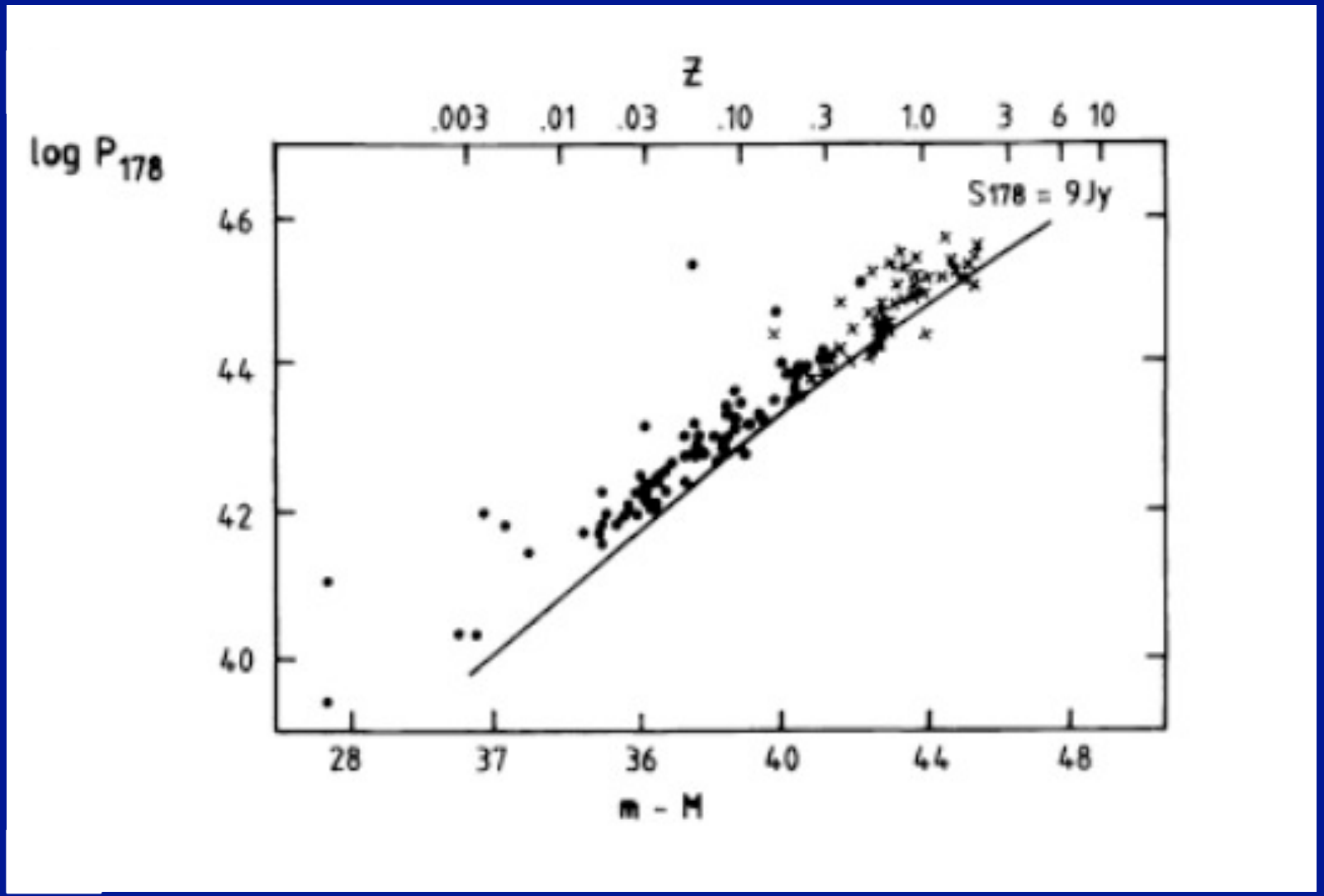


Carnegie-Mellon and University of Pittsburgh

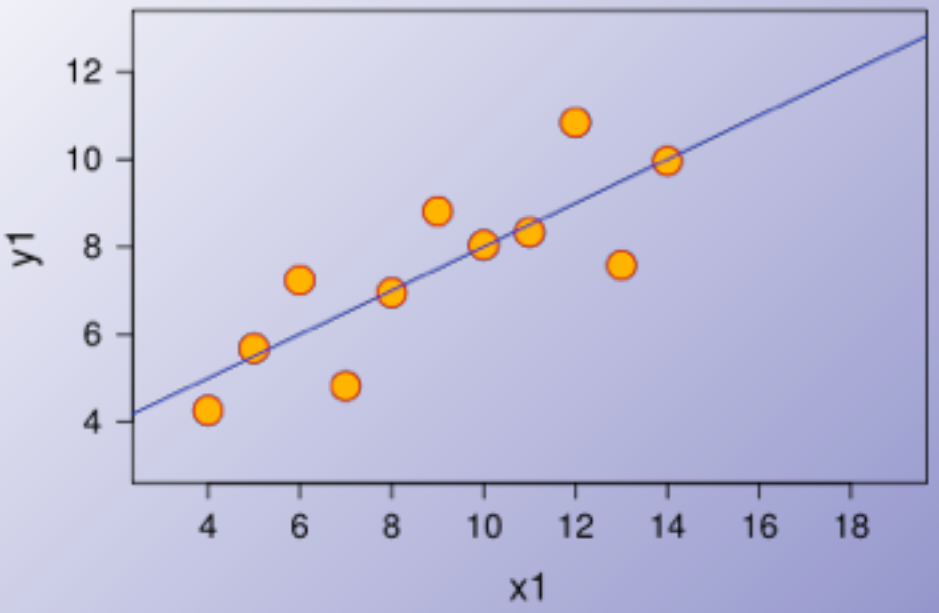

b
a
y
x

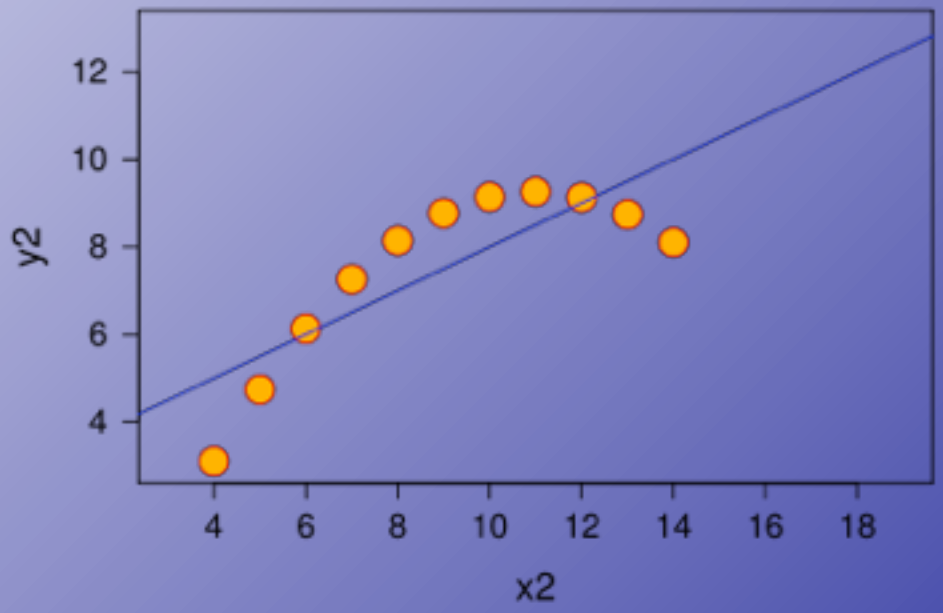


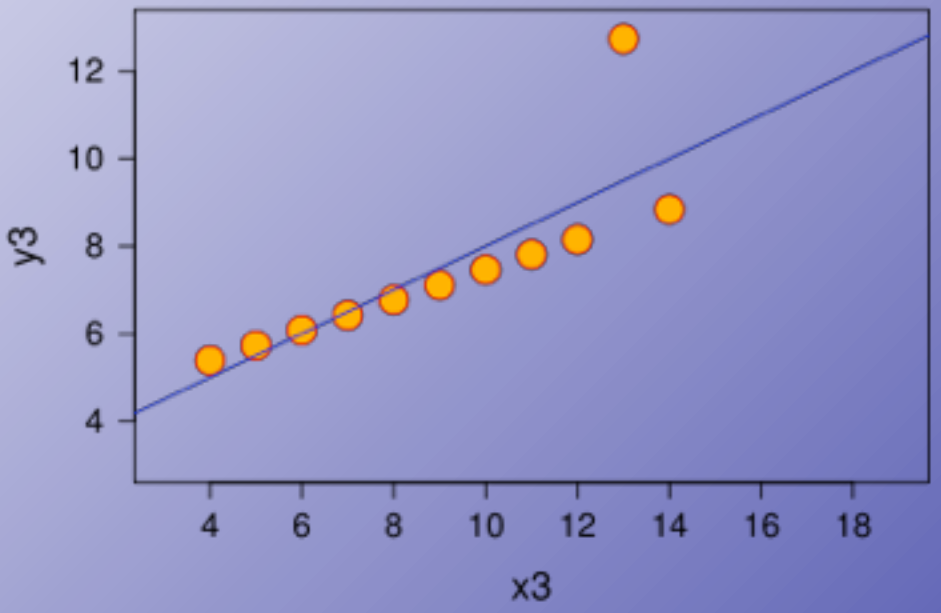


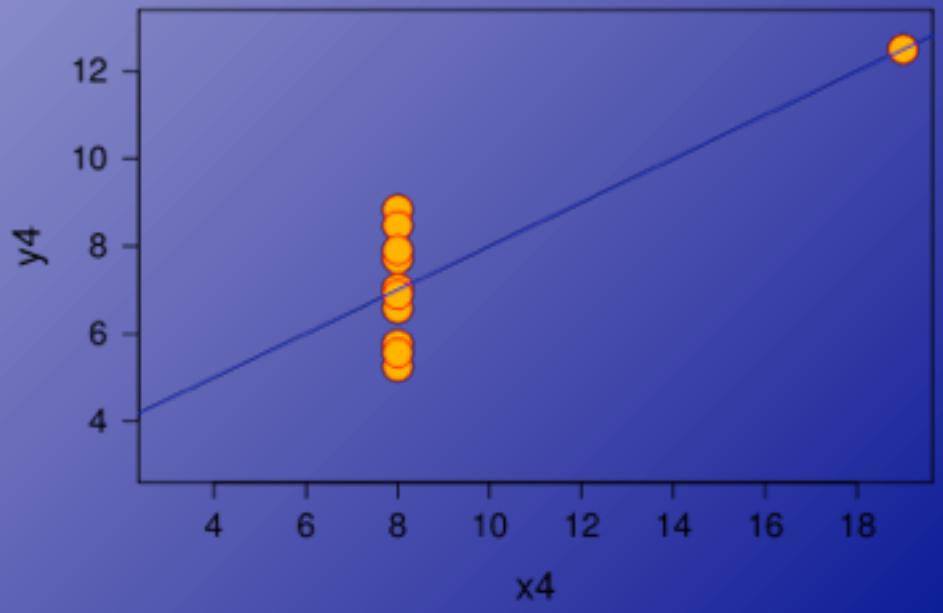

z
.003
.01
.03
.10
.3
1.0
3
6
10
log P178
S178 = 9 Jy
46
44
42
40
28
37
36
40
44
48
m - M

Anscombe's Quartet









Were taught that:

"we must use non-parametric tests"

But we tend to work in "recipe book style" !!

MOREOVER,

The Prevailing doctrine:

"Does the eye see much correlation? If not, calculation of a formal correlation statistic is probably a waste of time."

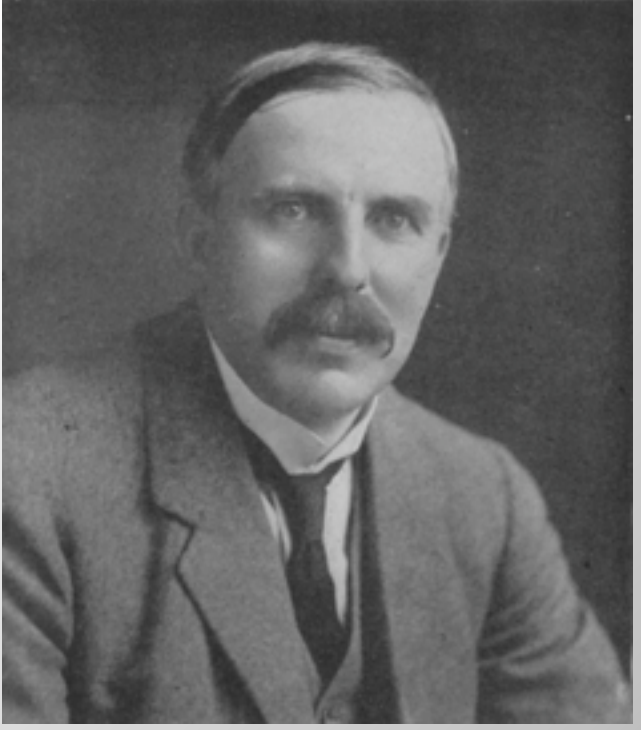
Thus, in general, the awareness and exposure is poor, and the ignorance is profound...and worse....quite unabashed!

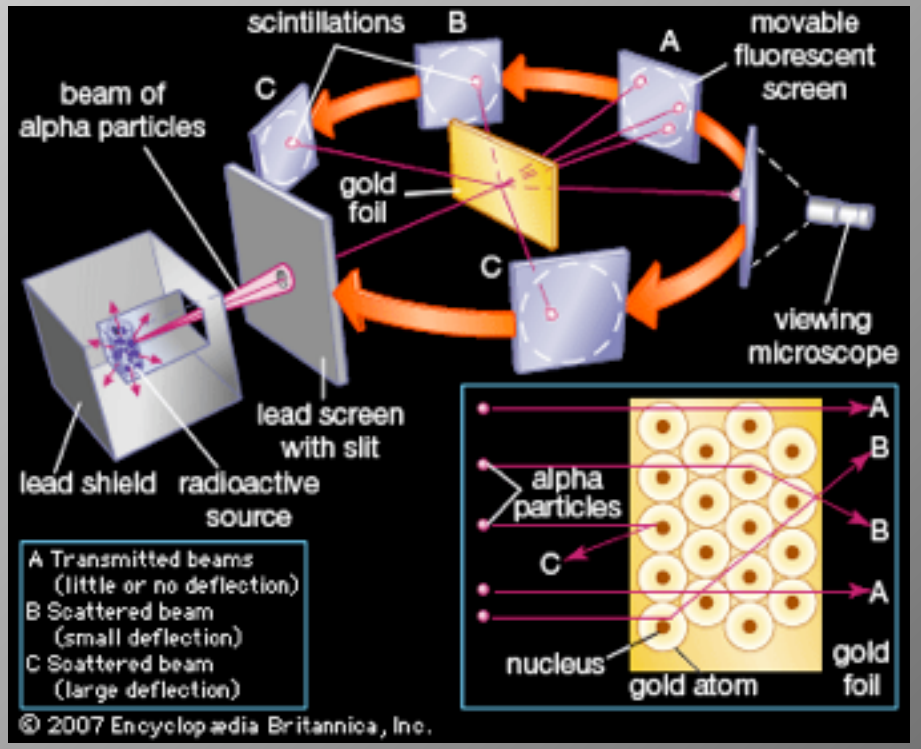
unabashed astrophysicist: Is the difference between these magnitudes significant?

eminent statistician: Don't ask me, go look at the data!

Institutional barriers

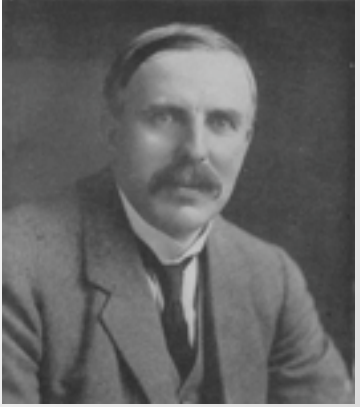
- Statistics is not part of an astrophysicist's formal training
- Astrophysicists tend to be housed in research institutes rather than in universities
- Astrophysicists come with their “physicist” baggage:





"If your experiment needs statistics, you ought to have done a better experiment."

– Ernest Rutherford



"If your experiment needs statistics, you ought to have done a better experiment."

- *Ernest Rutherford*

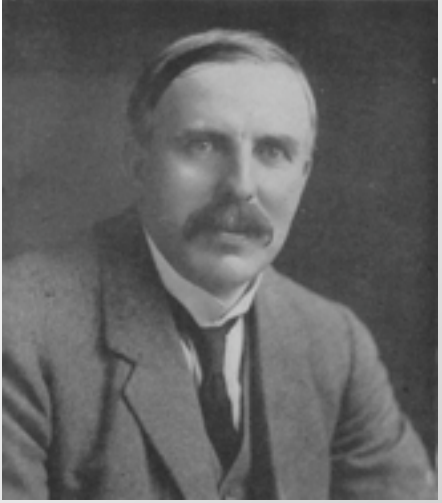
Empirical astrophysics differs from empirical physics:

- *it is an observational science, i.e., The variation in the information acquired is not in the control of the experimenter*

- *data are constantly gathered at the limit of the instrument capability*

In other observational disciplines, experimental design and inference, and hypothesis testing develop together, but this has only very rarely happened in astrophysics.

There is a GAP in the pedagogy!



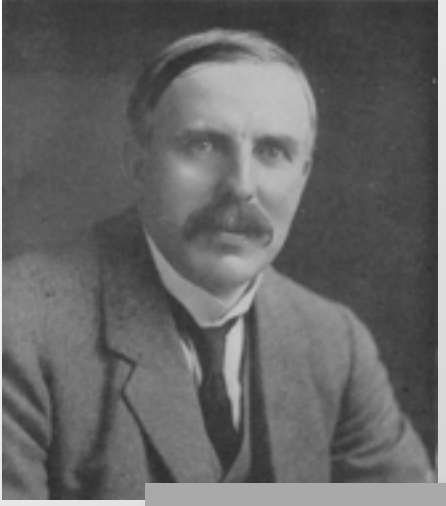
"If your experiment needs statistics, you ought to have done a better experiment."

– Ernest Rutherford

Acts are not only of **omission**: not using state-of-the-art statistical methodology, but

also of **commission**: methods are wrongly used or misused because of the recipe-book style with no understanding of the fundamentals

Often an astrophysicist does not know about the difference between the **sample** and the **population**, or the distinction between **Hypothesis Testing** & **Model Fitting**



"If your experiment needs statistics, you ought to have done a better experiment."

– Ernest Rutherford

Ac
of

als

Oft

Even traditional methods are often misused:

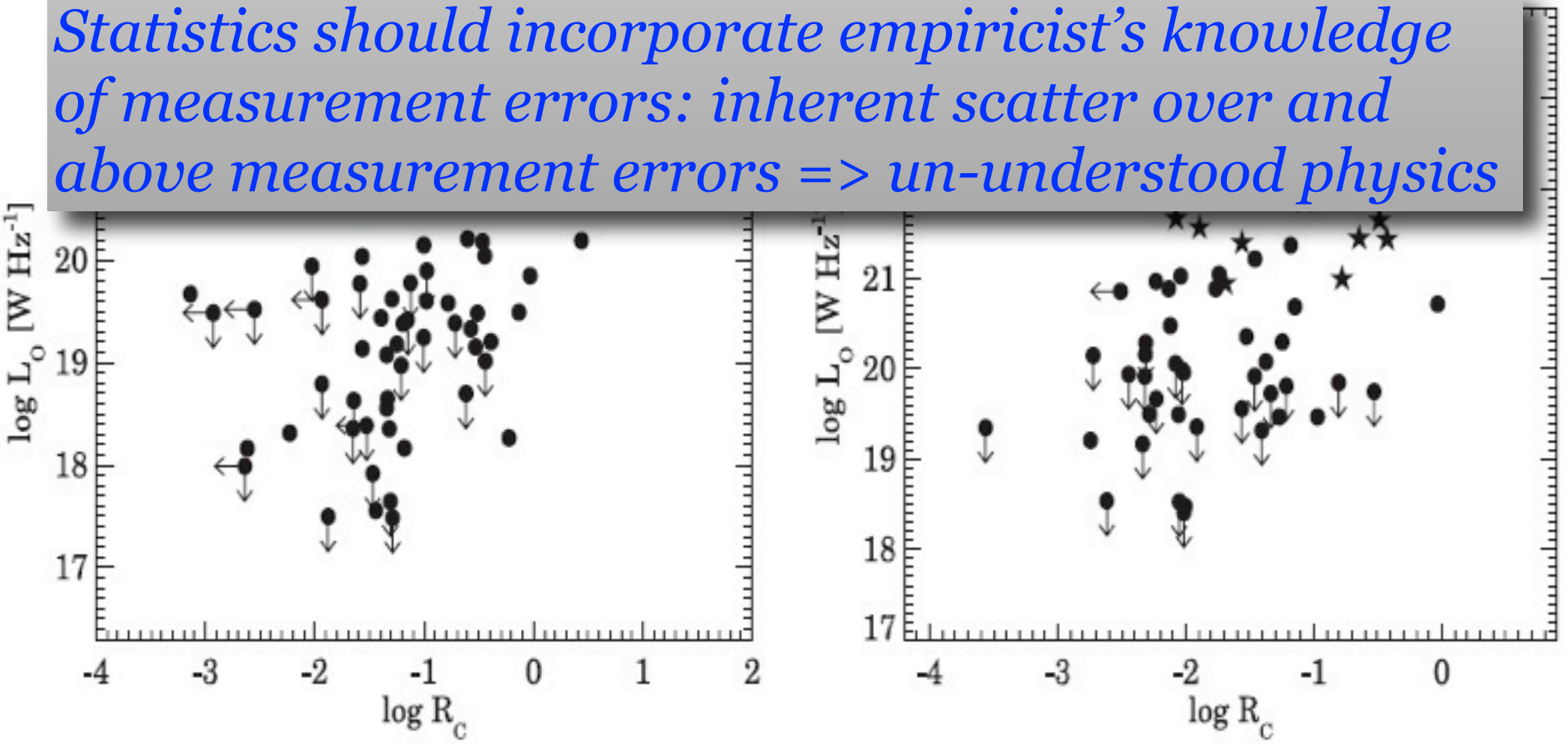
- **Unweighted bivariate least-squares fits are used interchangeably in Hubble constant studies with wrong confidence intervals**
  Feigelson & Babu ApJ 1992
- **Likelihood-ratio test (F test) usage typically inconsistent with asymptotic statistical theory**
  Protassov et al. ApJ 2002
- **Kolmogorov-Smirnov goodness-of-fit probabilities are inapplicable when the model is derived from the data**
  Babu & Feigelson ADASS 2006

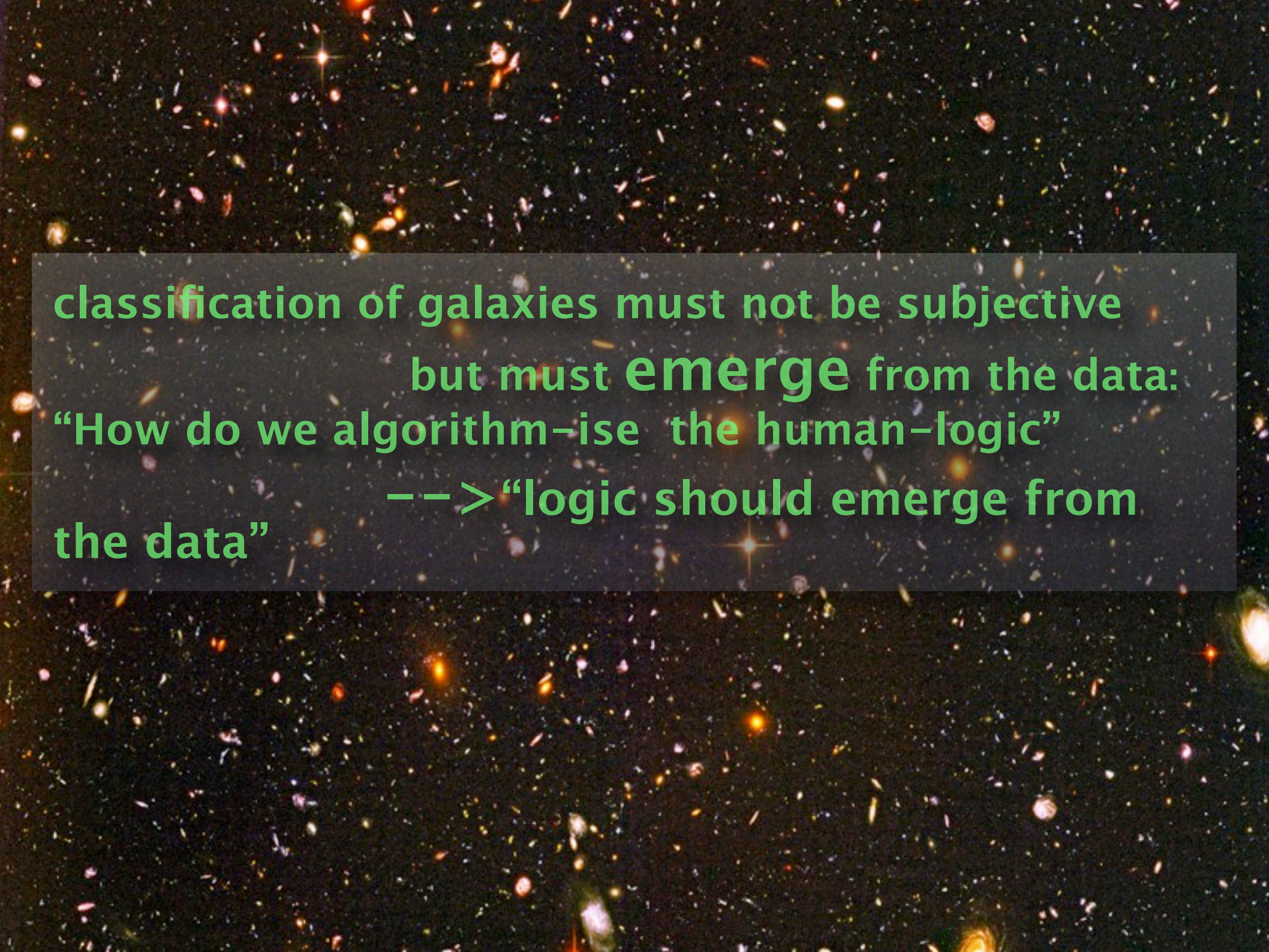
difference between the **sample** and the **population**, or the distinction between **Hypothesis Testing** & **Model Fitting**

Kharb & Shastri 2004

P. Kharb and P. Shastri: Optical nuclei and the F-R Divide

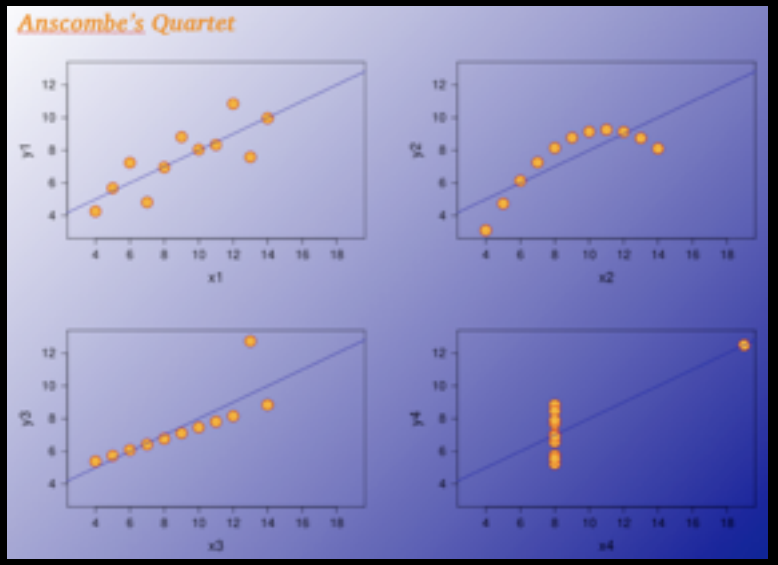
Statistics should incorporate empiricist's knowledge of measurement errors: inherent scatter over and above measurement errors => un-understood physics




classification of galaxies must not be subjective
but must emerge from the data:
“How do we algorithm-ise the human-logic”
-->“logic should emerge from the data”

Probability:

- Coin flips, conditional probabilities
- normal and chi-square distributions
- The Central Limit Theorem



Exploratory Data Analysis:

- *uncover the underlying structure*
- *detect outliers and anomalies*
- *extract important variables*
- *formulate hypotheses for testing*

Uses the R software environment

Statistical Inference:

- *Going beyond the immediate data*
- *Is the observed difference between groups dependable or could it have happened by chance?*

Bayesian Inference:

- *Taking prior knowledge into account*

Likelihood Estimation

Difference between likelihood and probability!

Fitting mathematical models to the data

Tuning the free parameters to obtain a good fit

Non-parametric statistics:

- which make no assumptions about the probability distributions of a population
- therefore applicability is wider

Model Selection:
Goodness of Fit: Bootstrap:
Cluster Analysis: Grouping, data mining

Multivariate Analysis: of data with two or more dependent variables

Monte-Carlo Markov Chain techniques that use
pseudo-random (simulated) values to estimate mathematical solutions

Time Series Analysis

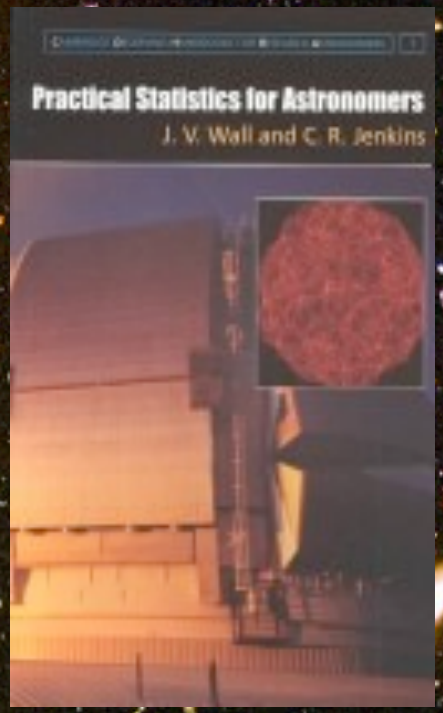
ASAIP

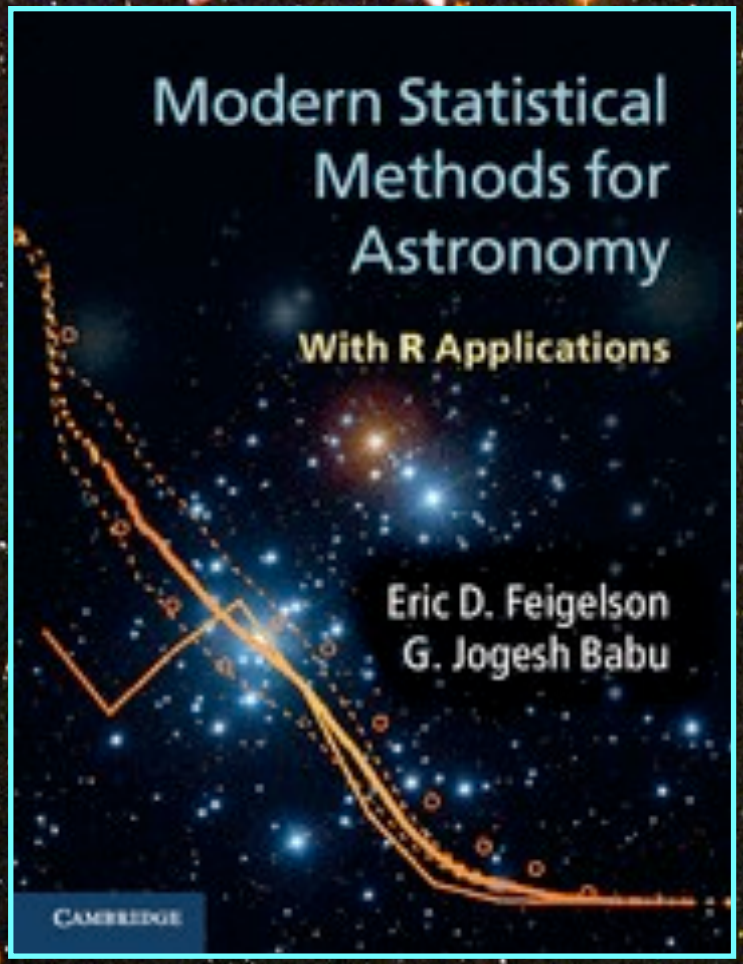
Astrostatistics & Astro-informatics Portal

(http://asaip.psu.edu)

astro.r Facebook group

(https://www.facebook.com/groups/astro.r)





Thankyou!

ASAIP :

Astrostatistics & Astro-informatics Portal

(http://asaip.psu.edu)

astro.r Facebook group

(https://www.facebook.com/groups/astro.r)