大數據天文學 — 時間序列分析

Michael Ting-Chang Yang 楊庭彰

首先我們要知道

在日常生活中，有許多事情跟時間有關

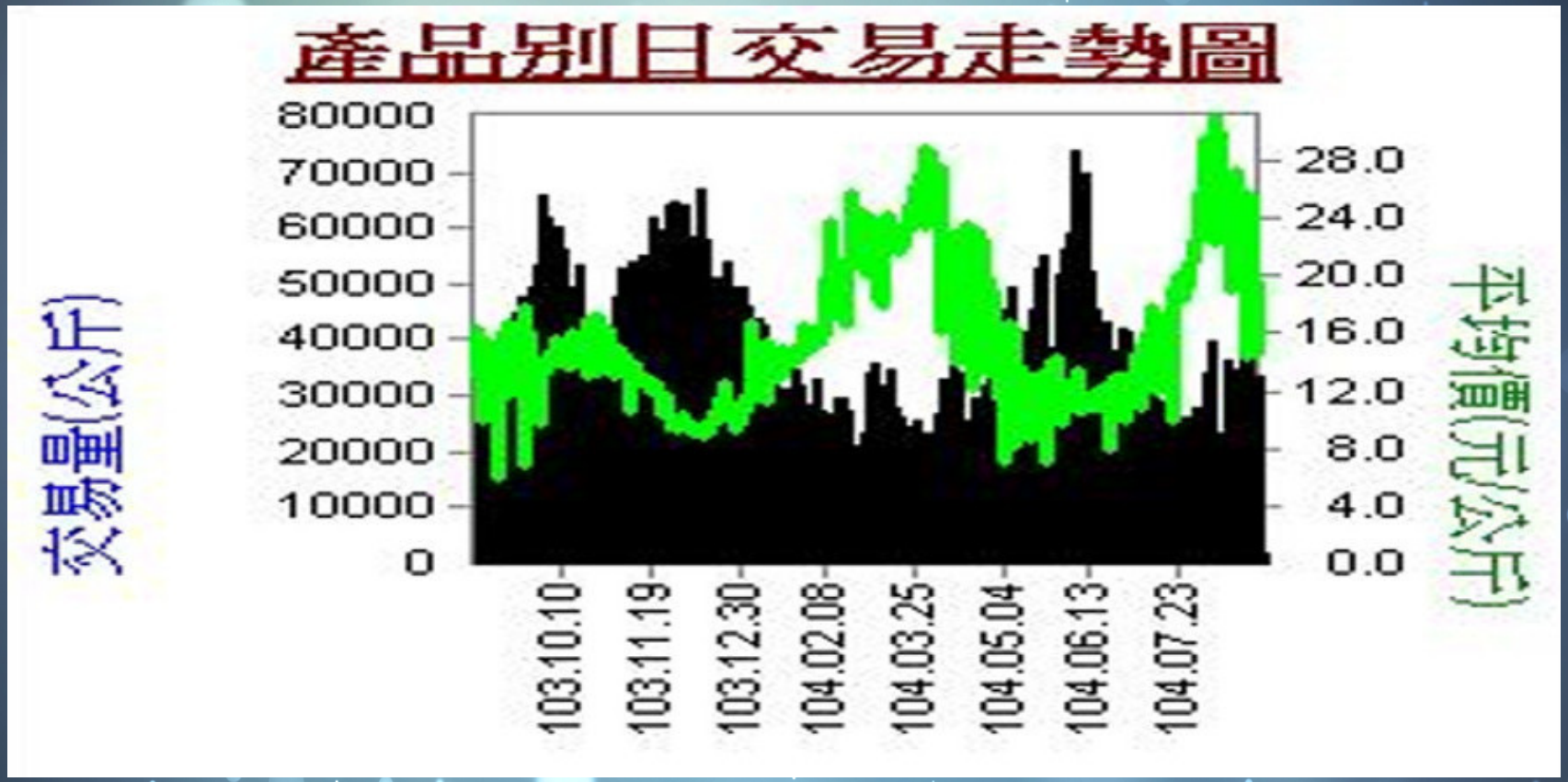
時間序列無所不在

- 個人
  - 體重變化
  - 身高變化
  - 年齡變化
- 社會、經濟
  - 股市
  - 匯率
  - 交通流量
  - 民調支持度
- 自然
  - 動物群數量
  - 全球氣溫
  - 海洋高度

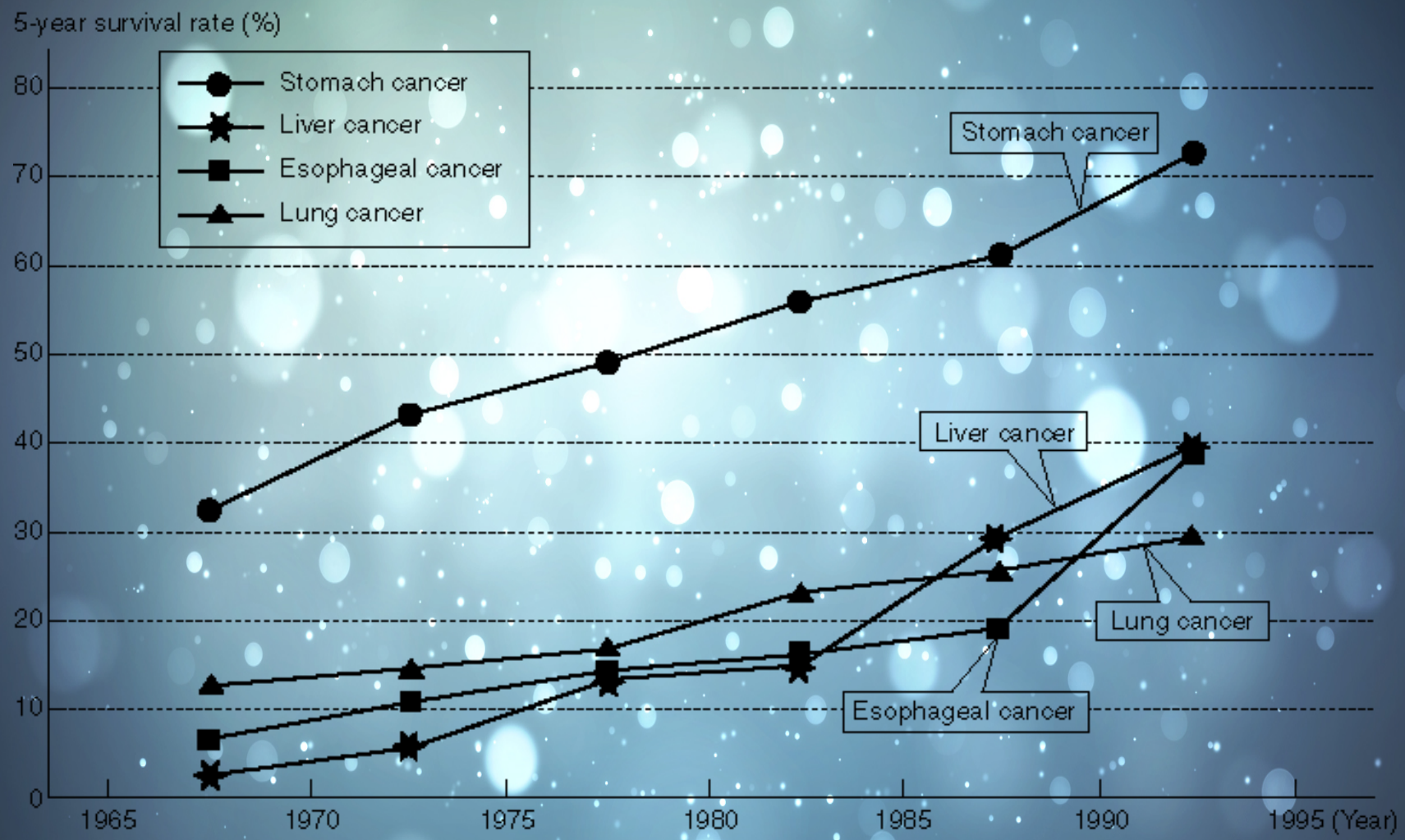
股市加權指數


Zoom 3m 6m 12m
加權指數 (0000) 日K線
玩股網 www.wantgoo.com
©WantGoo玩股網
收盤價 8,542.27
漲跌 ▲ 31.41
漲跌幅 ▲ 0.37 %
0.75
0.5
0.25
0
10,000
9,500
9,000
8,500
8,000
7,500
7,000
季線扣抵 9340.13
月線扣抵 8510.86
成交量(張)
成交量 868
MV5 880.34
MV20 919.47
KD
K9 20.43
D9 18.49
0
RSI
RSI6 29.93
RSI12 28.83

蔬果價格

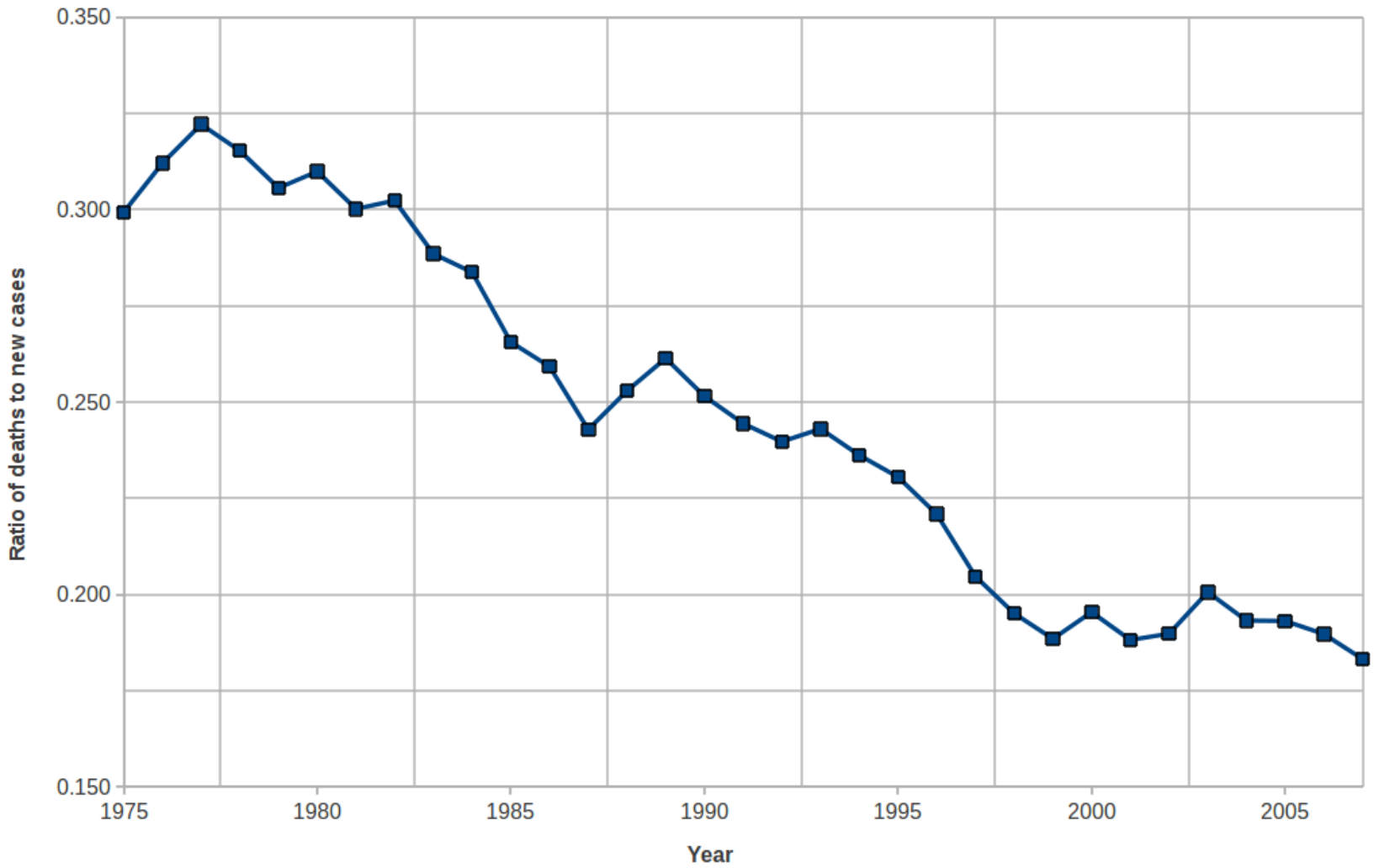

產品別日交易走勢圖
交易量(公斤)
平均價(元/公斤)
80000
70000
60000
50000
40000
30000
20000
10000
0
28.0
24.0
20.0
16.0
12.0
8.0
4.0
0.0
103.10.10
103.11.19
103.12.30
104.02.08
104.03.25
104.05.04
104.06.13
104.07.23

癌症統計

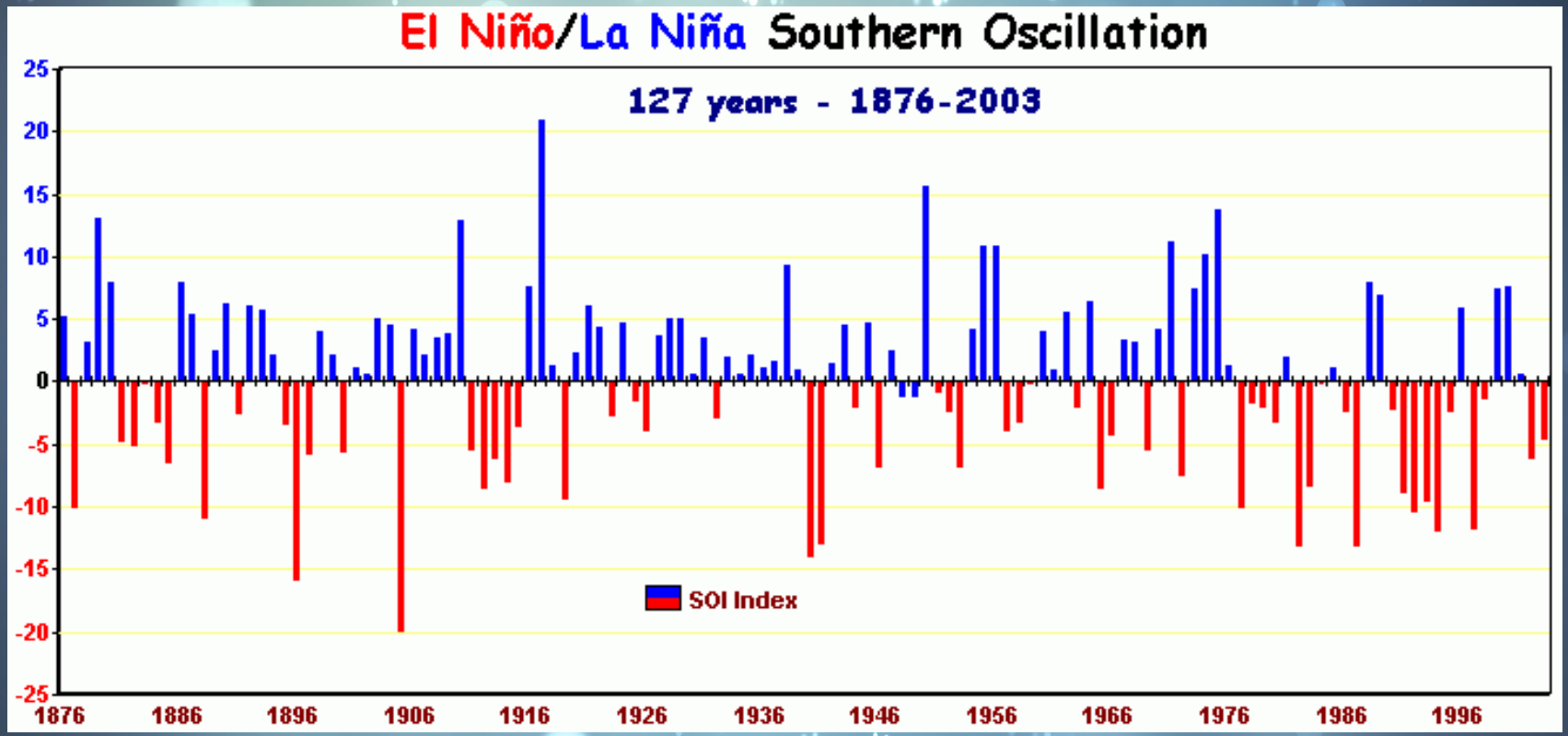


Source: The Central Hospital of National Cancer Center. The values are the survival rates after 5 years from the treatment during every 5 years since 1965 for the registered inpatients (males) in the hospital.

癌症統計 (cont.)


Ratio of Deaths to New Cases for Breast Cancer in United States
1975 to 2007
Ratio of deaths to new cases
0.350
0.300
0.250
0.200
0.150
1975
1980
1985
1990
1995
2000
2005
Year

South Oscillation – El Nino

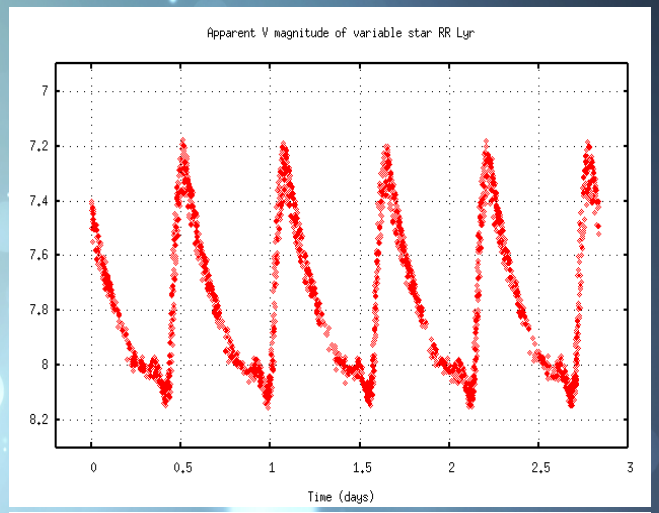


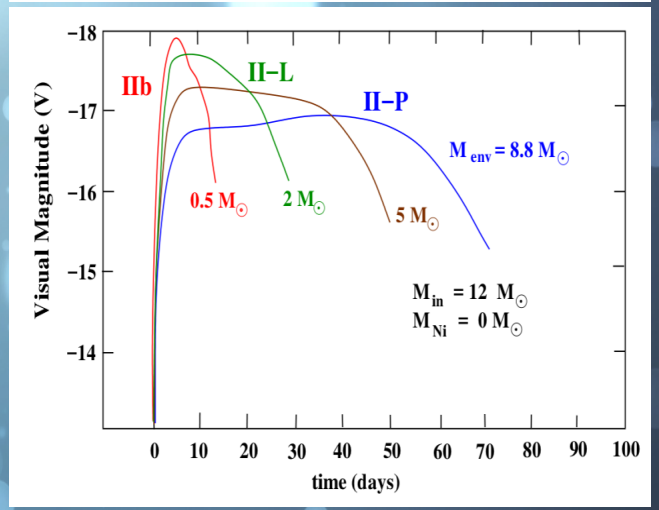
有了資料，然後呢？

- 大數據 → 將資料變成~~垃圾~~有價值的資訊
- 時間序列分析 → 將時間序列資料中有價值的部分找出來
- 既然時間序列的資料無所不在，所以時間序列分析的技巧也是相當重要

時間序列的特性

- 時間: time
- （物理）量: quantity
- 周期: period
- 振幅: amplitude
- 時間尺度: scale
- 形狀: shape
- 資料是有序的: ordered data
- 今日主題：天文時序資料

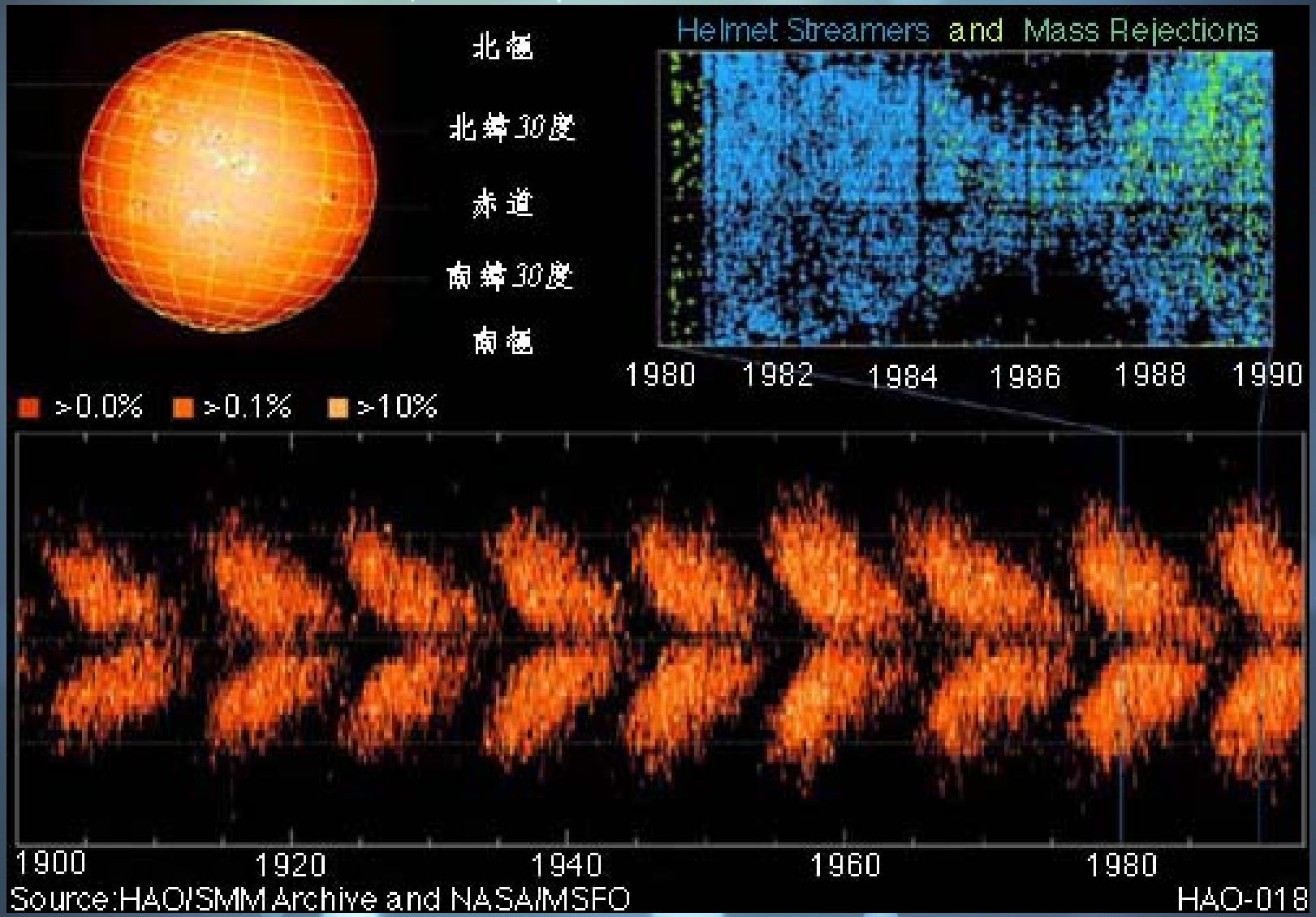




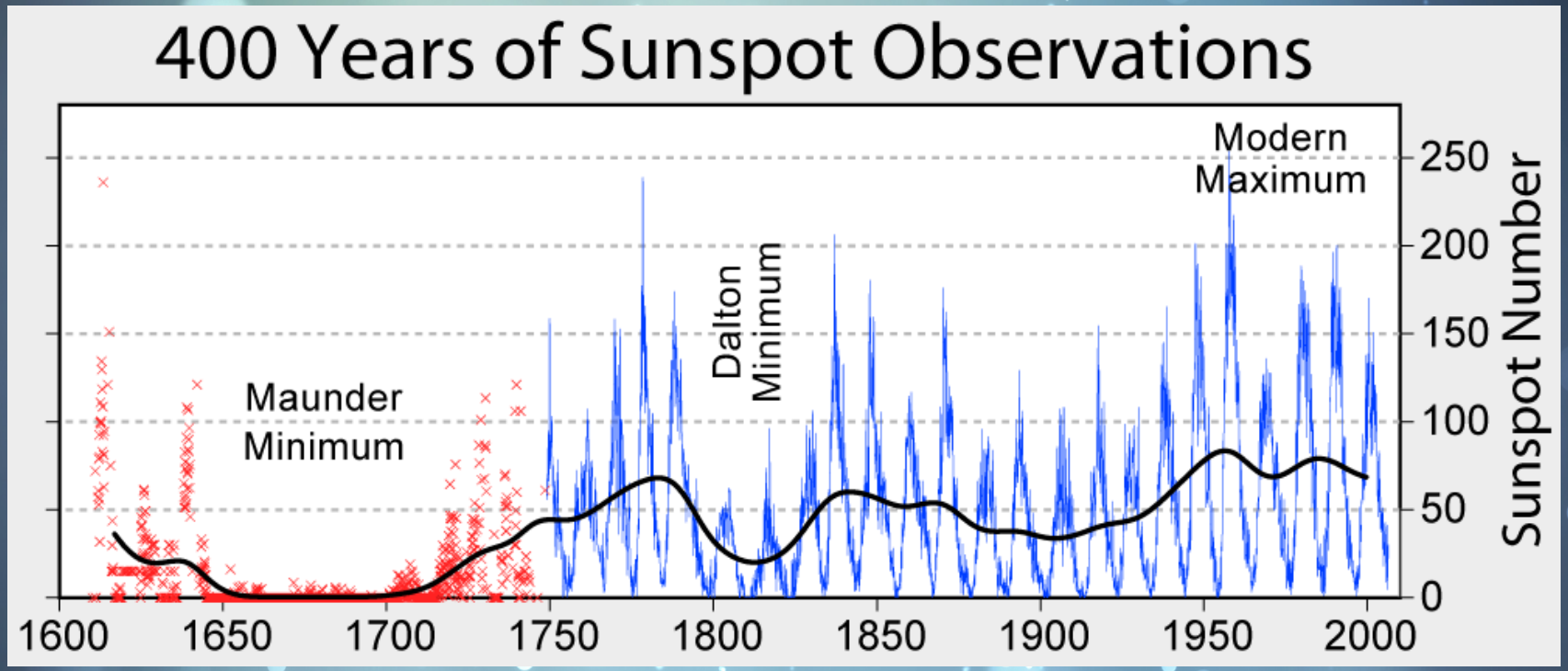
常見的天文時間序列

- 太陽黑子周期
- 光變曲線
- 色變曲線
- 半徑變化曲線
- 關鍵字：variation/variability

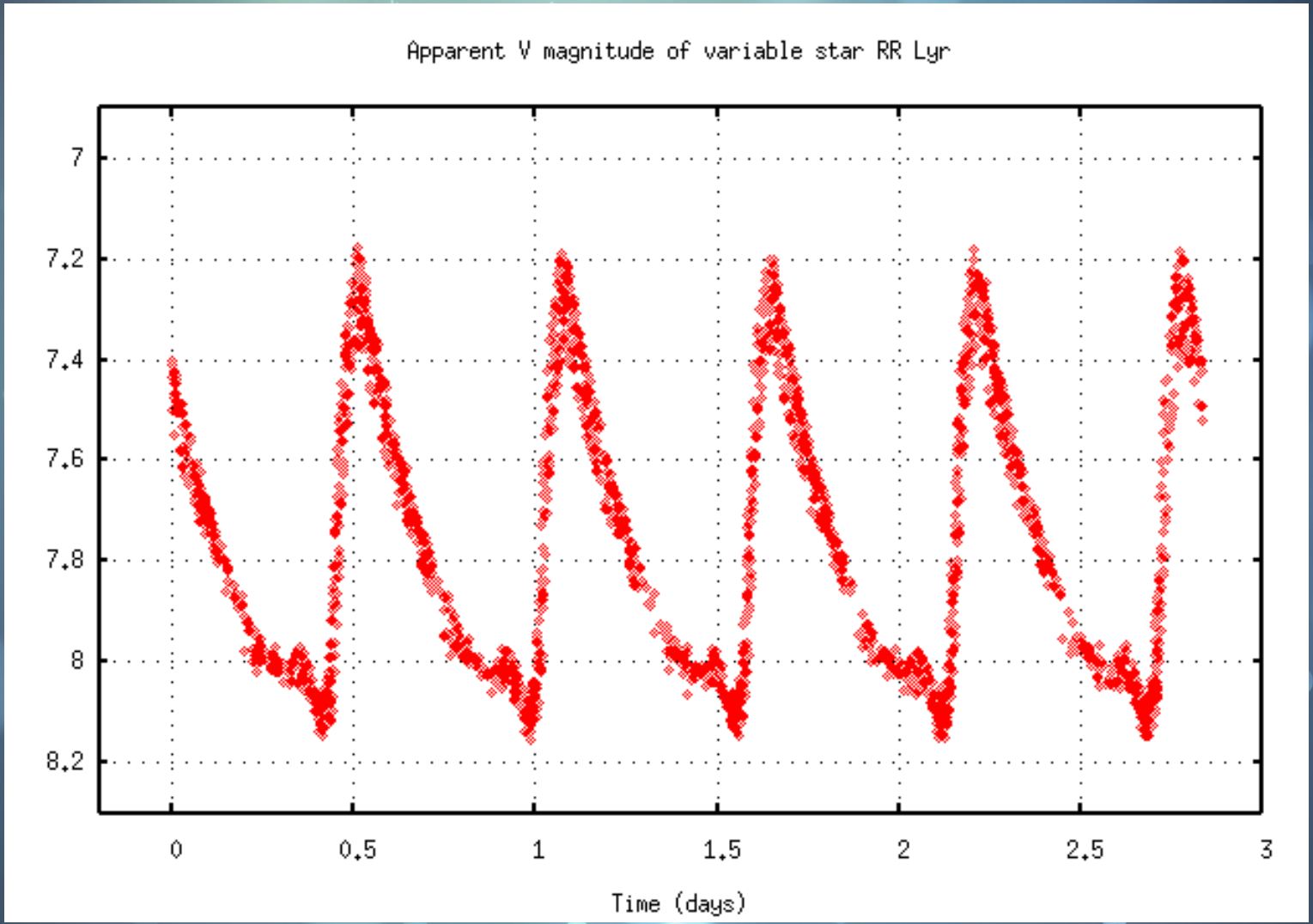
太陽黒子


北極
北緯30度
赤道
南緯30度
南極
Helmet Streamers and Mass Rejections
1980 1982 1984 1986 1988 1990
>0.0% >0.1% >10%
1900 1920 1940 1960 1980
Source:HAO/SMM Archive and NASA/MSFO
HAO-018

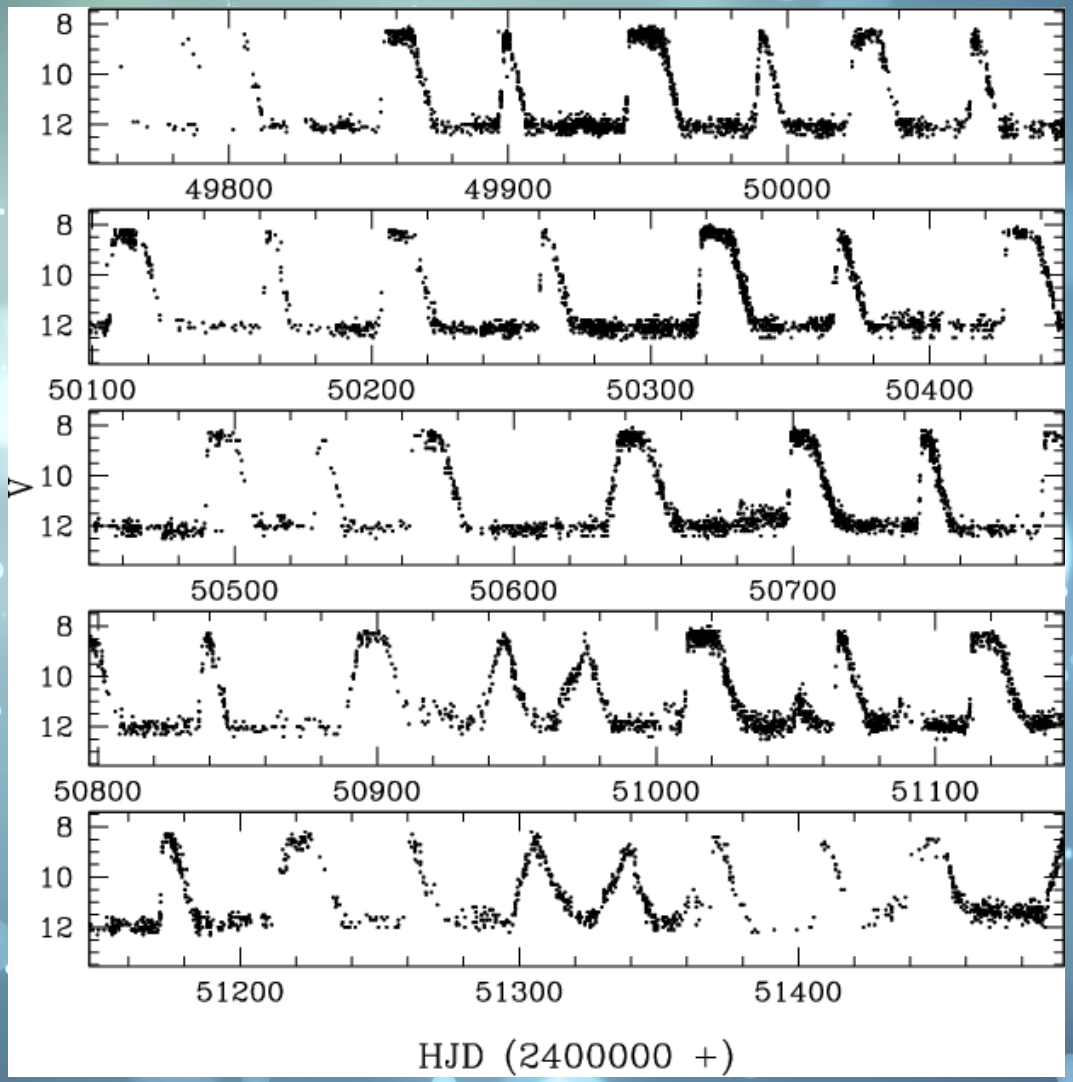
太陽黒子 (cont.)


400 Years of Sunspot Observations
Modern
Maximum
Dalton
Minimum
Maunder
Minimum
Sunspot Number
250
200
150
100
50
0
1600
1650
1700
1750
1800
1850
1900
1950
2000

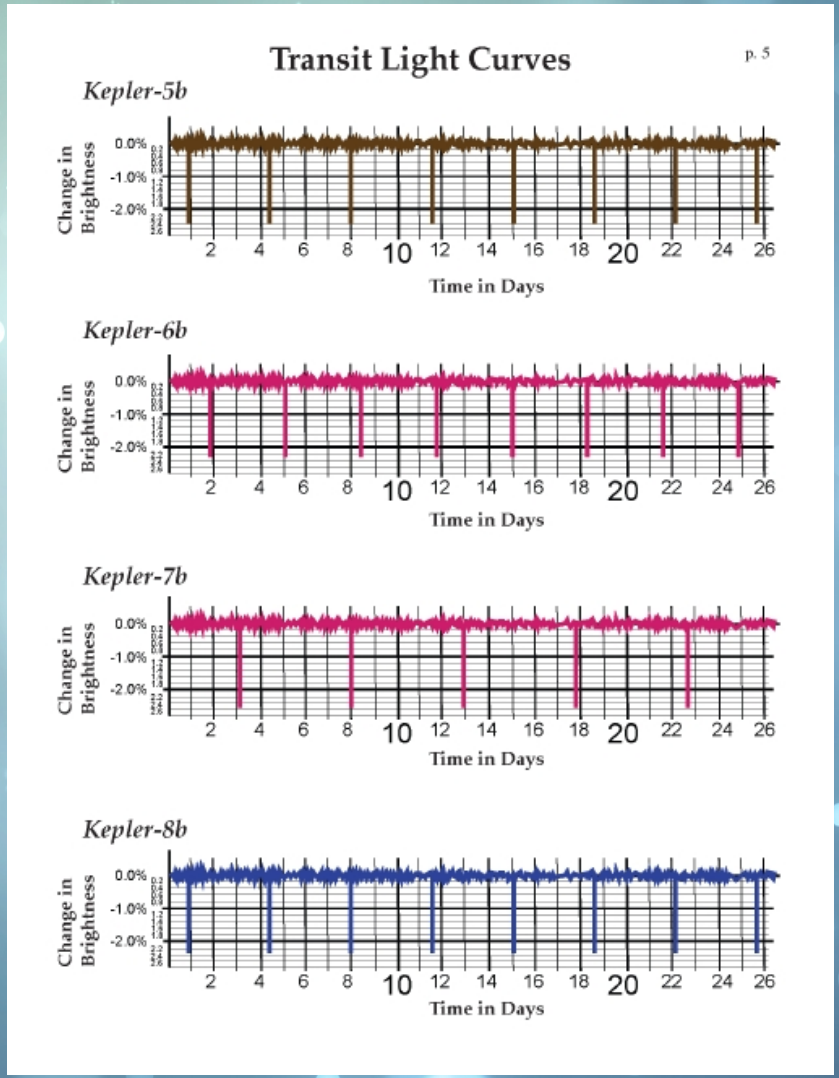
RR Lyrae 變星光變曲線



SS Cyg 變星光變曲線



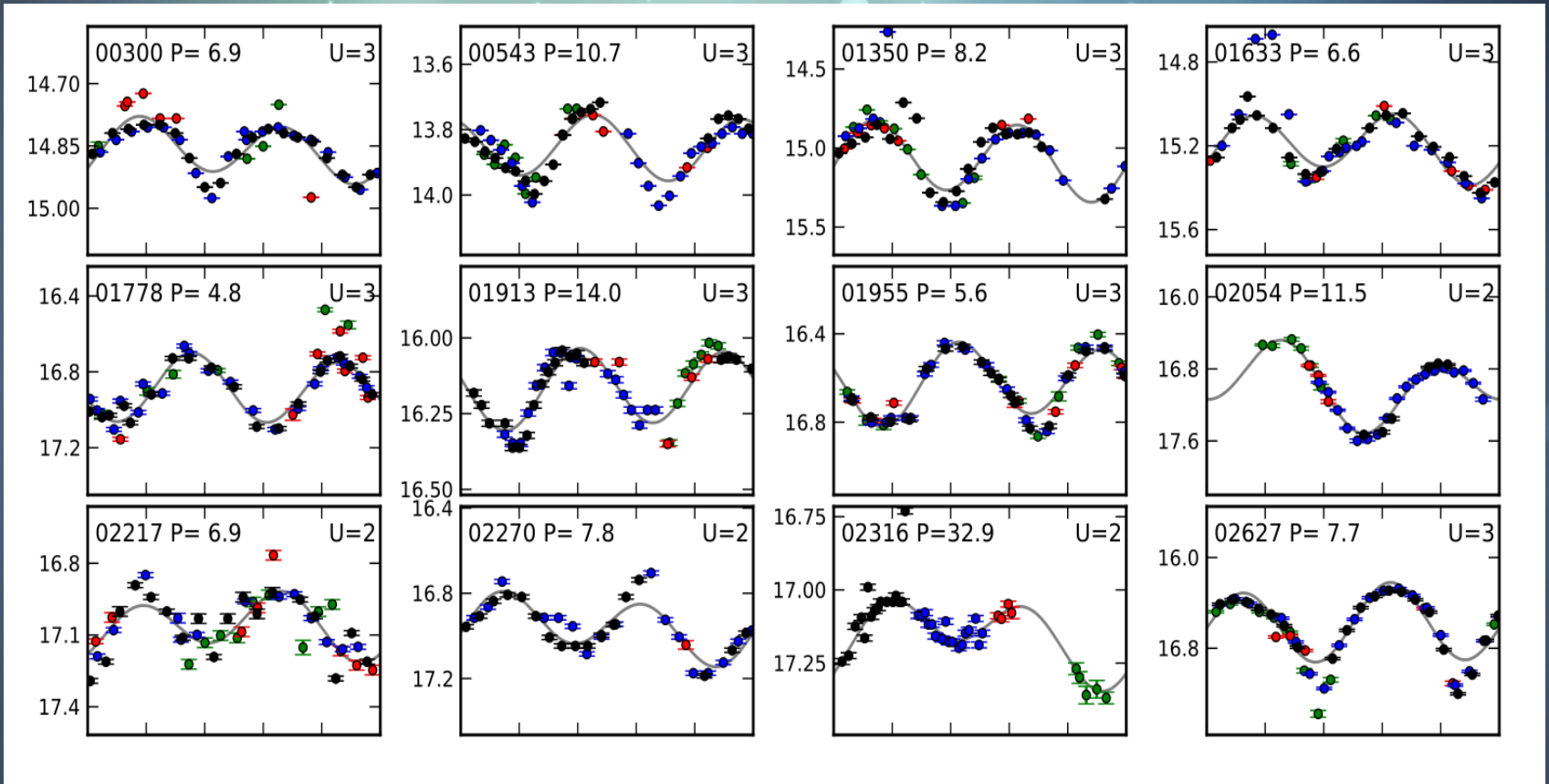
Kepler transit 光變曲線



光變曲線的應用

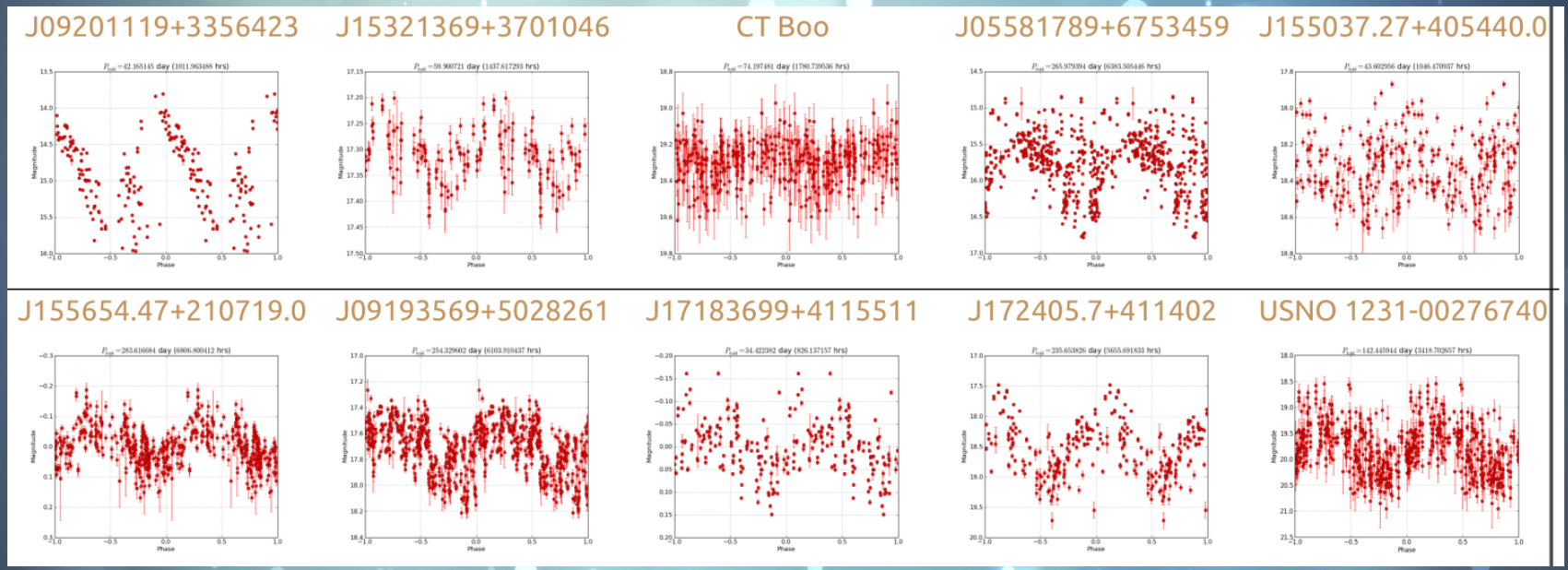
- 不同種類變星的研究
- 活躍星系核的活動
- 小行星旋轉曲線
- 系外行星
- 恒星閃焰
- 超新星爆發
- ...

小行星旋轉曲線



Chan-Kao Chang et al. (2014)

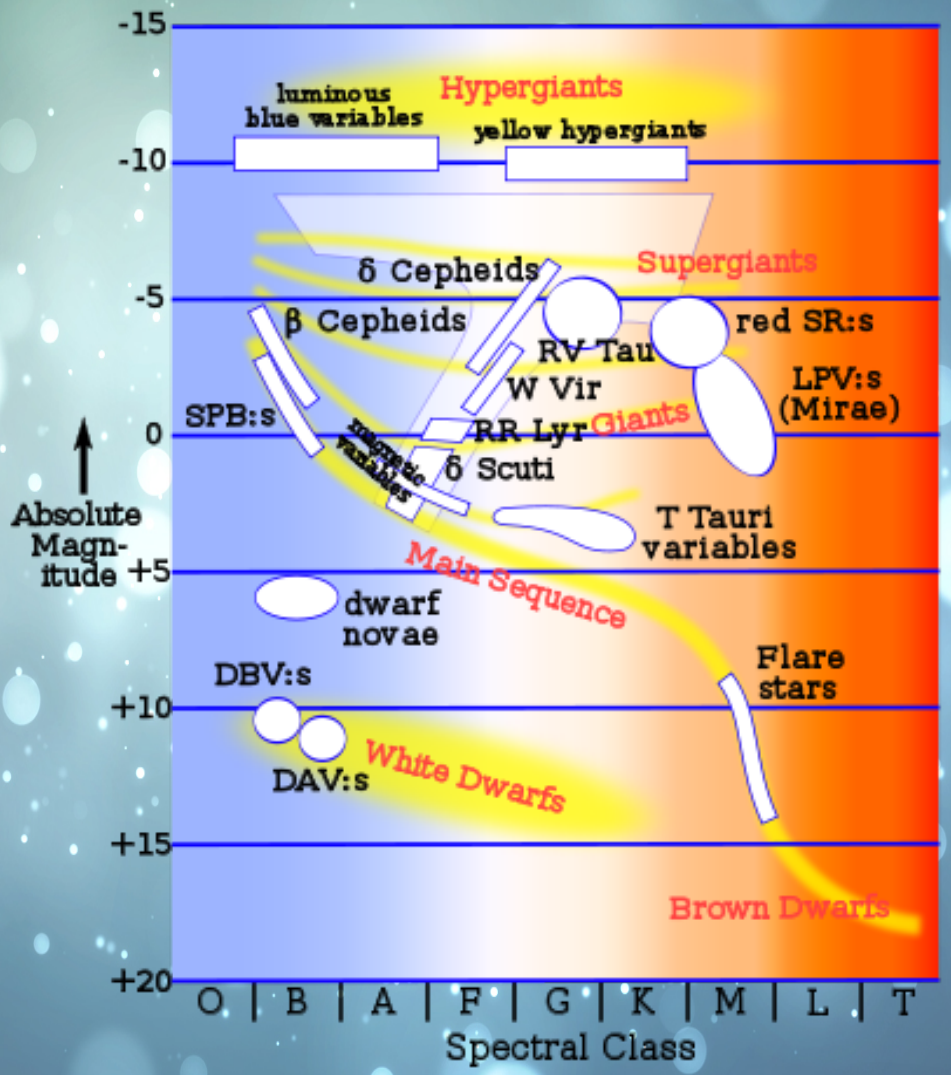
激變變星的長周期變化


J09201119+3356423
J15321369+3701046
CT Boo
J05581789+6753459
J155037.27+405440.0
J155654.47+210719.0
J09193569+5028261
J17183699+4115511
J172405.7+411402
USNO 1231-00276740

變星的分類


Variable Star Classification
Group
Class
Type
Variable Stars
Intrinsic
Pulsating Stars
Cepheids
RR Lyrae
RV Tauri
Long-Period Variables (LPVs)
Type I Classical
Type II W Virginis
Mira
Semiregular
Cataclysmic Stars
Supernovas
Novas
Recurrent Novas
Dwarf Novas
Symbiotic Stars
R Coronae Borealis
Extrinsic
Eclipsing Binaries
Rotating Variables

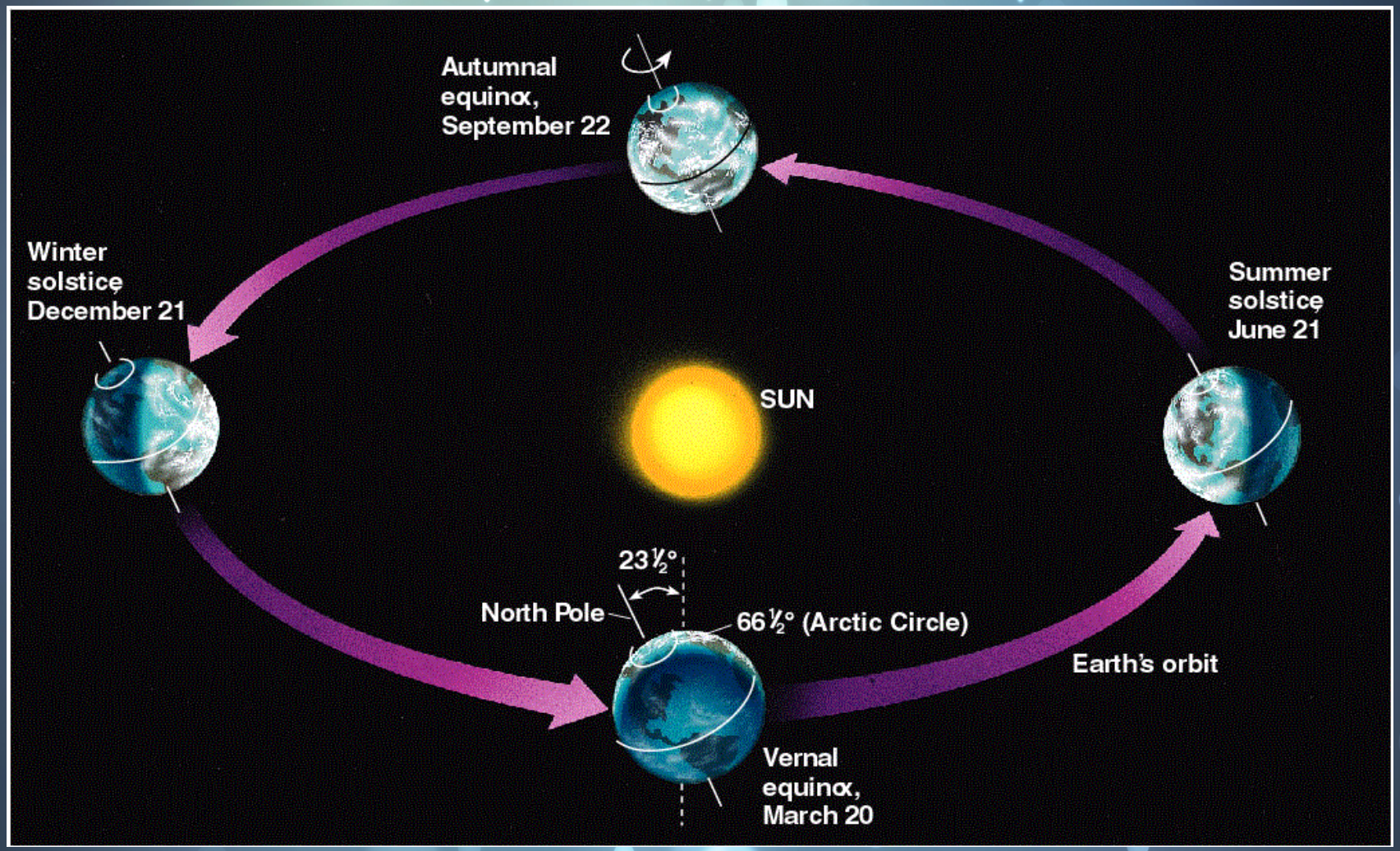
變星與赫羅圖


-15
-10
-5
0
+5
+10
+15
+20
Absolute Magn-itude
luminous blue variables
Hypergiants
yellow hypergiants
δ Cepheids
Supergiants
β Cepheids
red SR:s
RV Tau
W Vir
LPV:s (Mirae)
SPB:s
RR Lyr
Giants
δ Scuti
magnetic variables
T Tauri variables
Main Sequence
dwarf novae
Flare stars
DBV:s
DAV:s
White Dwarfs
Brown Dwarfs
O
B
A
F
G
K
M
L
T
Spectral Class

時間序列分析 – 周期分析

- 如果光變曲線有變化性，我們相信有其物理原因
- 所以首先要判別訊號中有沒有周期性
- 首先，要先取得及整理資料
  - 觀測
  - 測光
  - 前處理
    - data reduction
    - barycenter correction
    - etc
  - 周期分析

太陽系質心時間修正: Barycenter Correction



太陽系質心時間修正: Barycenter Correction (cont.)

Conversion	Convert from	Convert to	Light travel time	Time system conversion
Geocentric correction	Mission Elapsed Time	Geocentric time	±23 ms at maximum	From Mission Elapsed Time to Terrestrial Time (TT)
Barycentric correction	Geocentric time	Barycentric time	±500 s at maximum	From TT to Barycentric Dynamical Time (TDB)
Binary demodulation	Barycentric time	Binary-demodulated time	Depends on binary parameters	None

傅立葉分析 – Fourier Transformation

$$F(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx$$

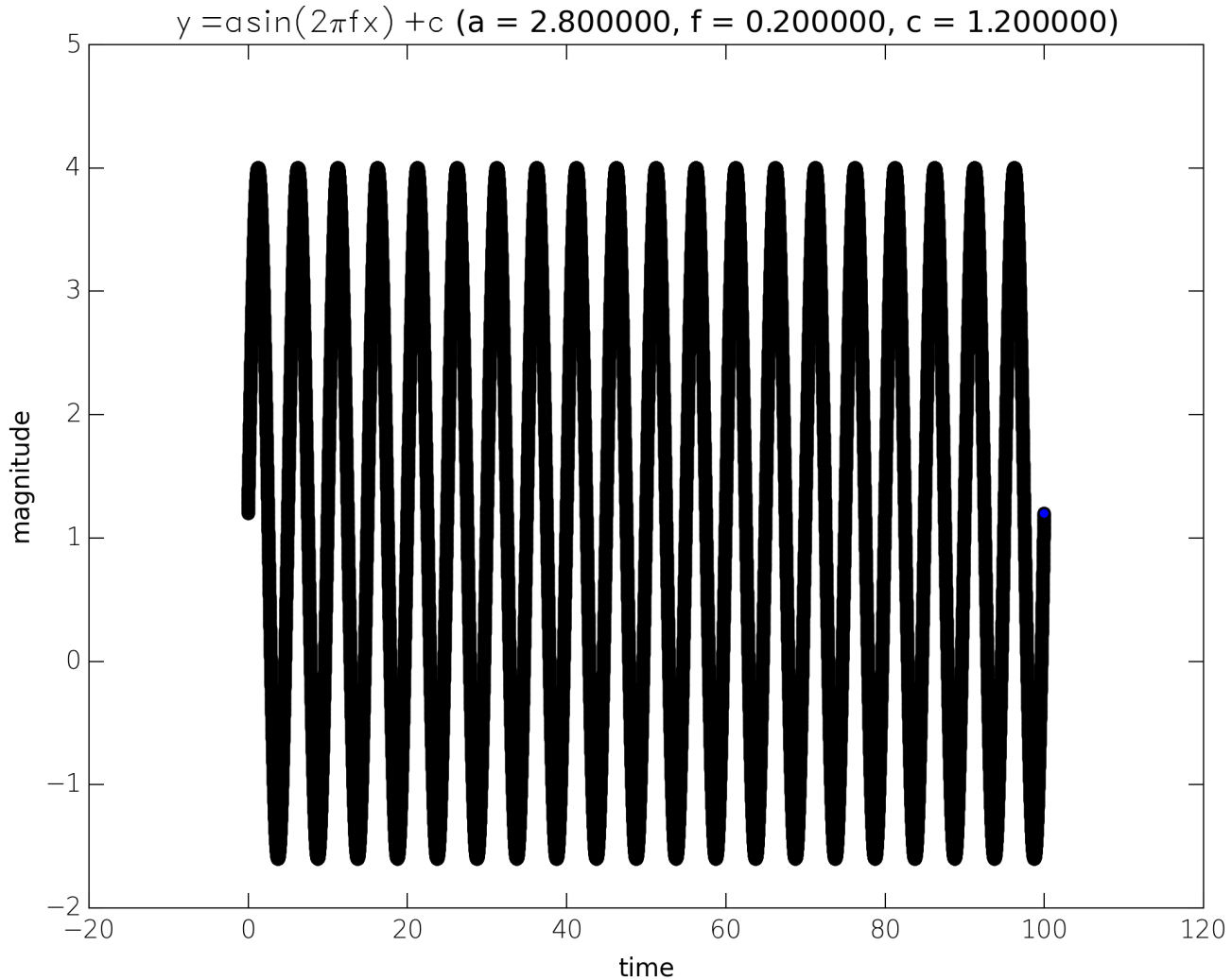
離散傅立葉分析 Discrete Fourier Transformation (DFT):

- 傅立葉分析一開始只用於連續周期函數
- 不過真實世界中，大部分的資料是離散的（因為觀測的緣故）
- 所以，離散傅立葉分析 (discrete Fourier transoformation) (DFT) 就派上用場

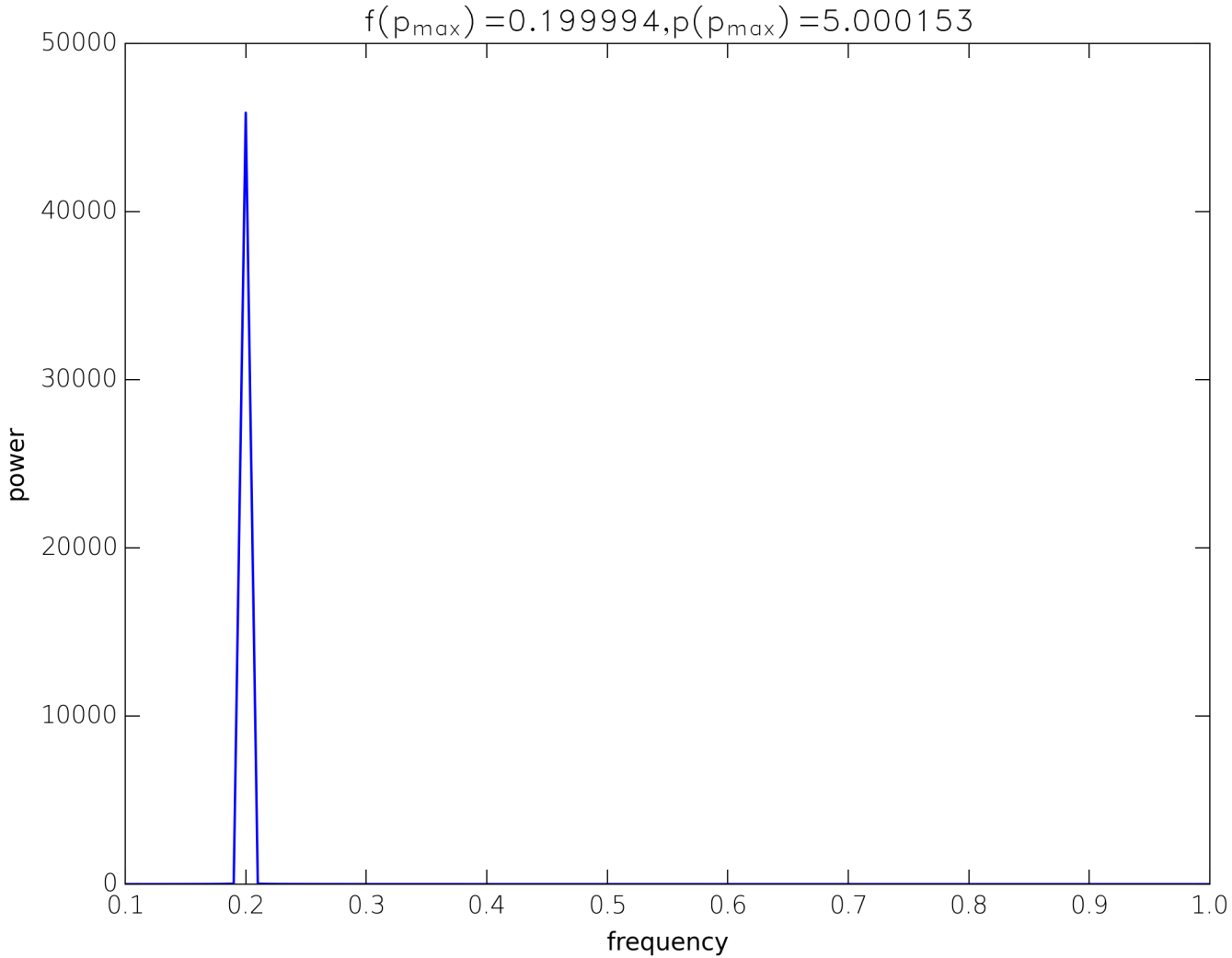
快速傅立葉分析 – Fast Fourier Transformation

- A fast Fourier transform (FFT) is an algorithm that computes the discrete Fourier transform (DFT) of a sequence, or its inverse.
- **最常見的演算法**: Cooley–Tukey FFT algorithm (Numerical Recipes, Ch. 12)
- **一些限制：**
  - **資料點數目**: 2^n
  - **資料點要均勻分布**
  - **如果資料點數目不為 2^n 的話，那一般在後面補零至 2^n 個資料點**

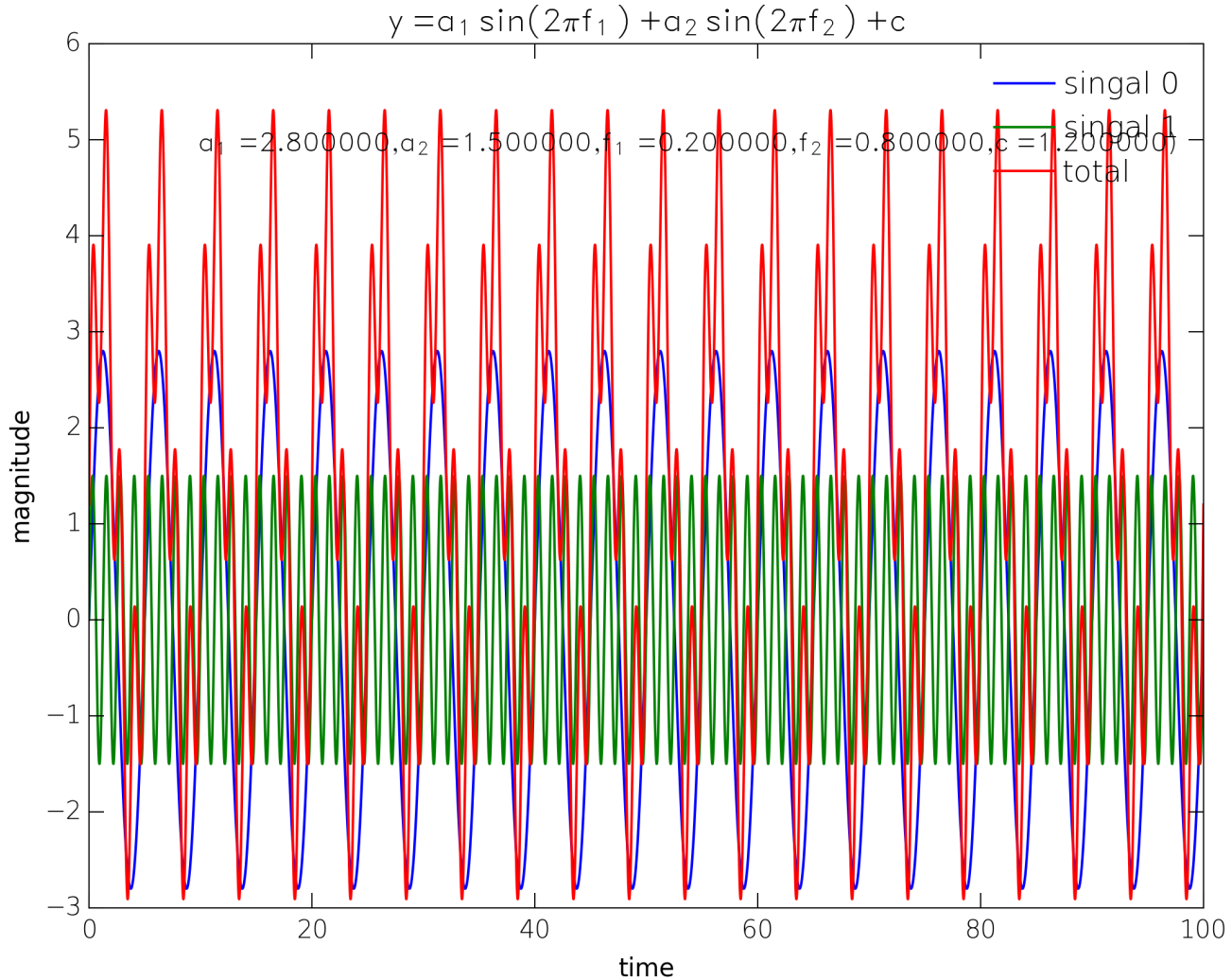
Light Curve: Single period



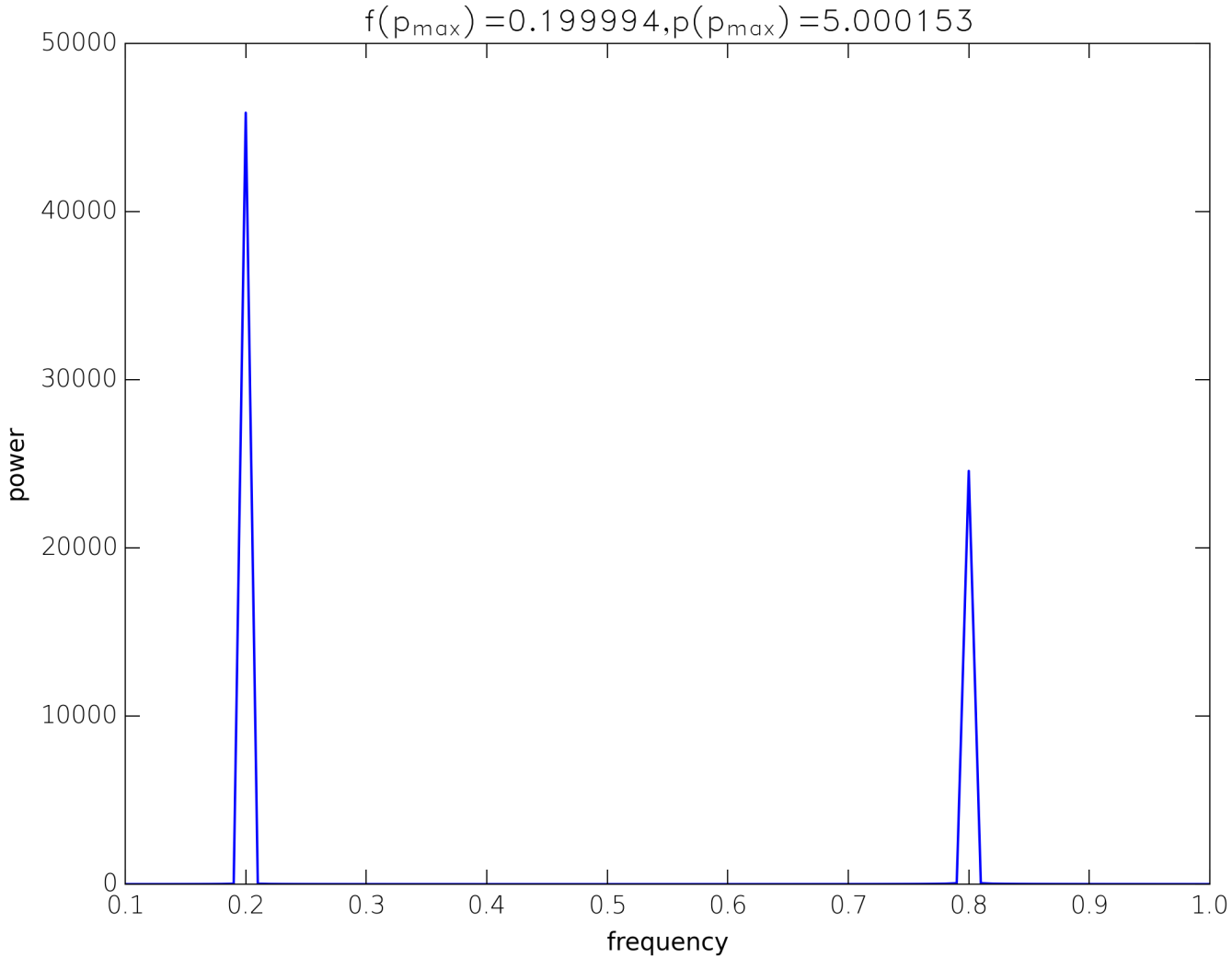
Fourier Power Spectrum



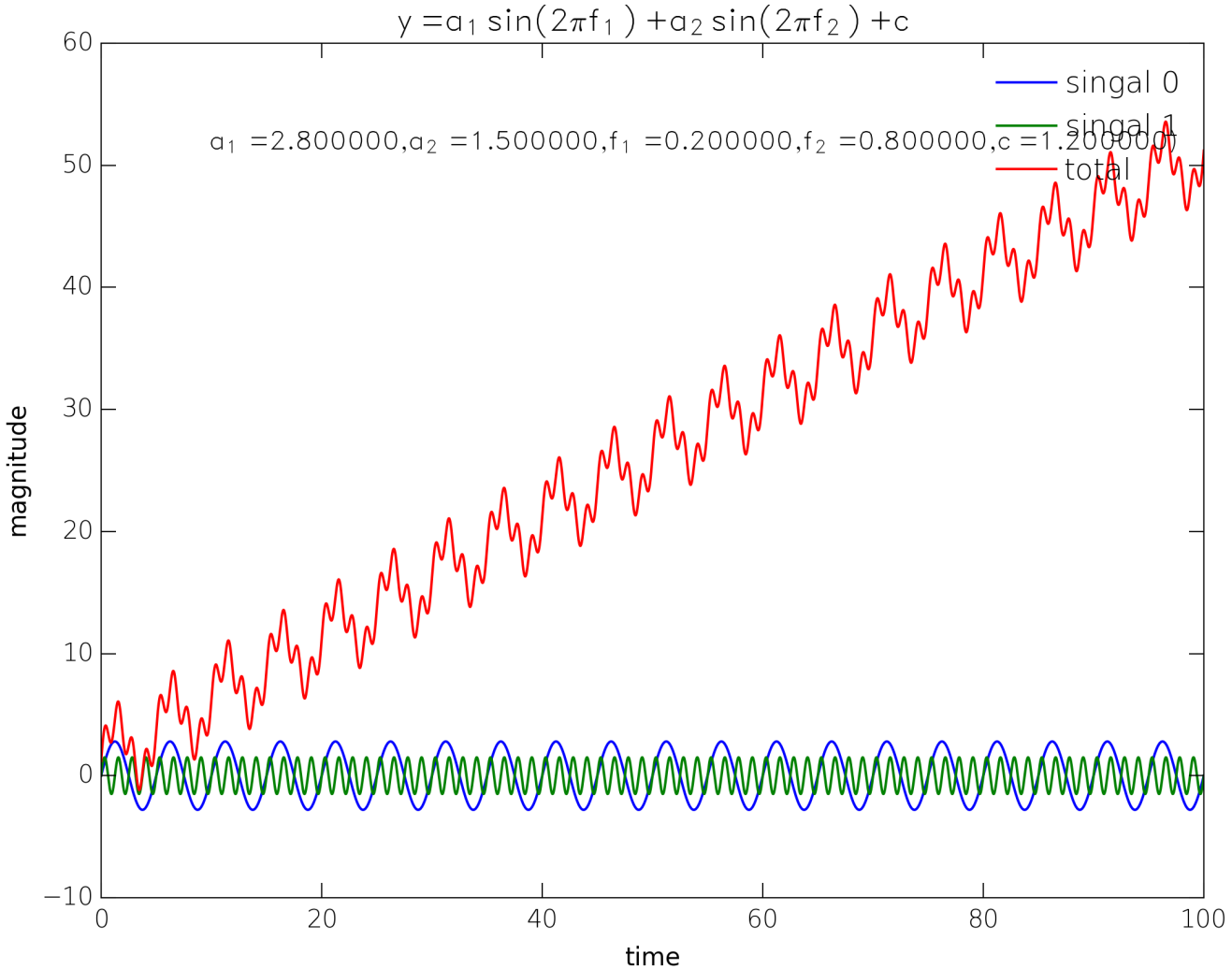
Light Curve: Two periods



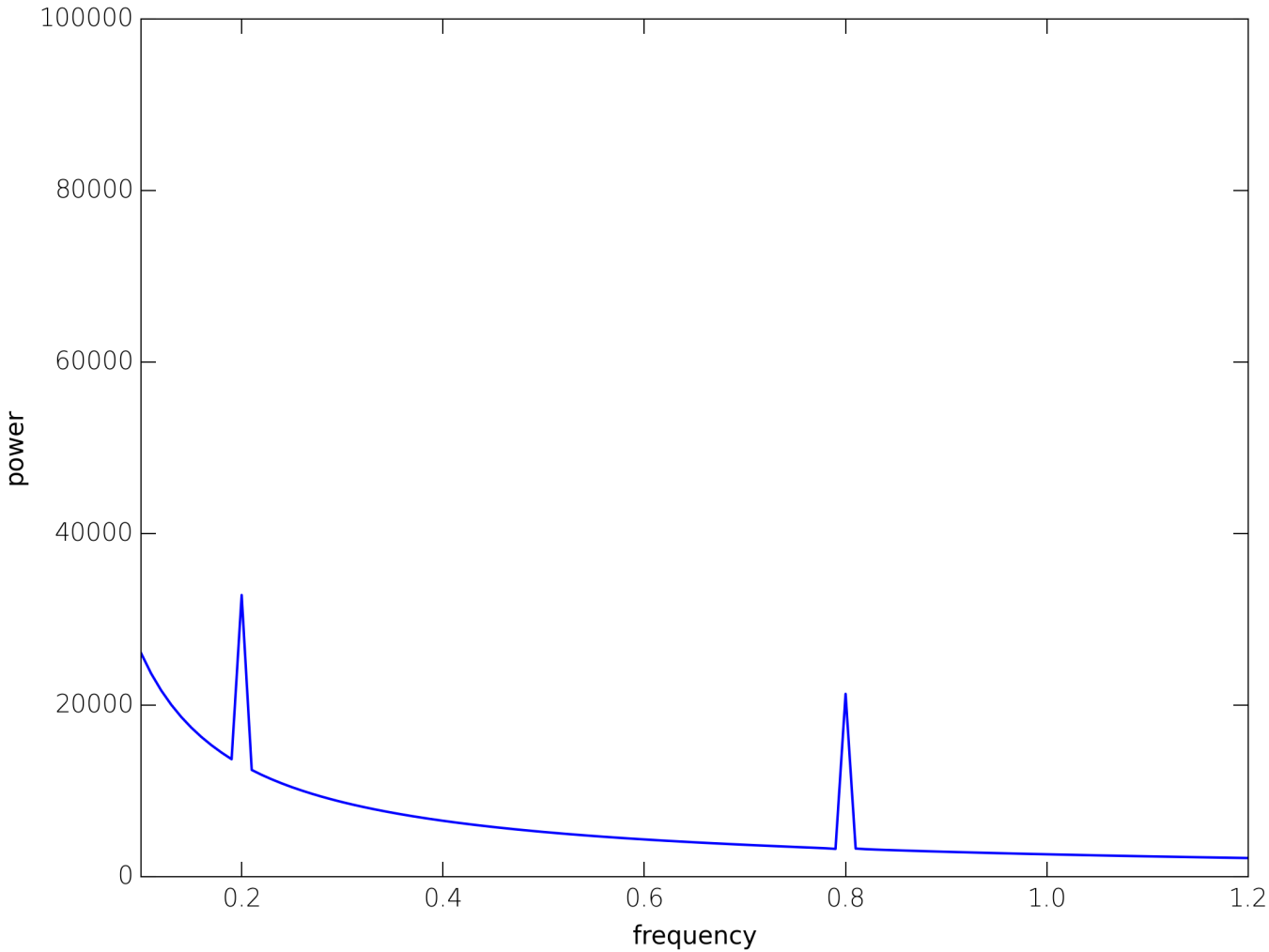
Fourier Power Spectrum



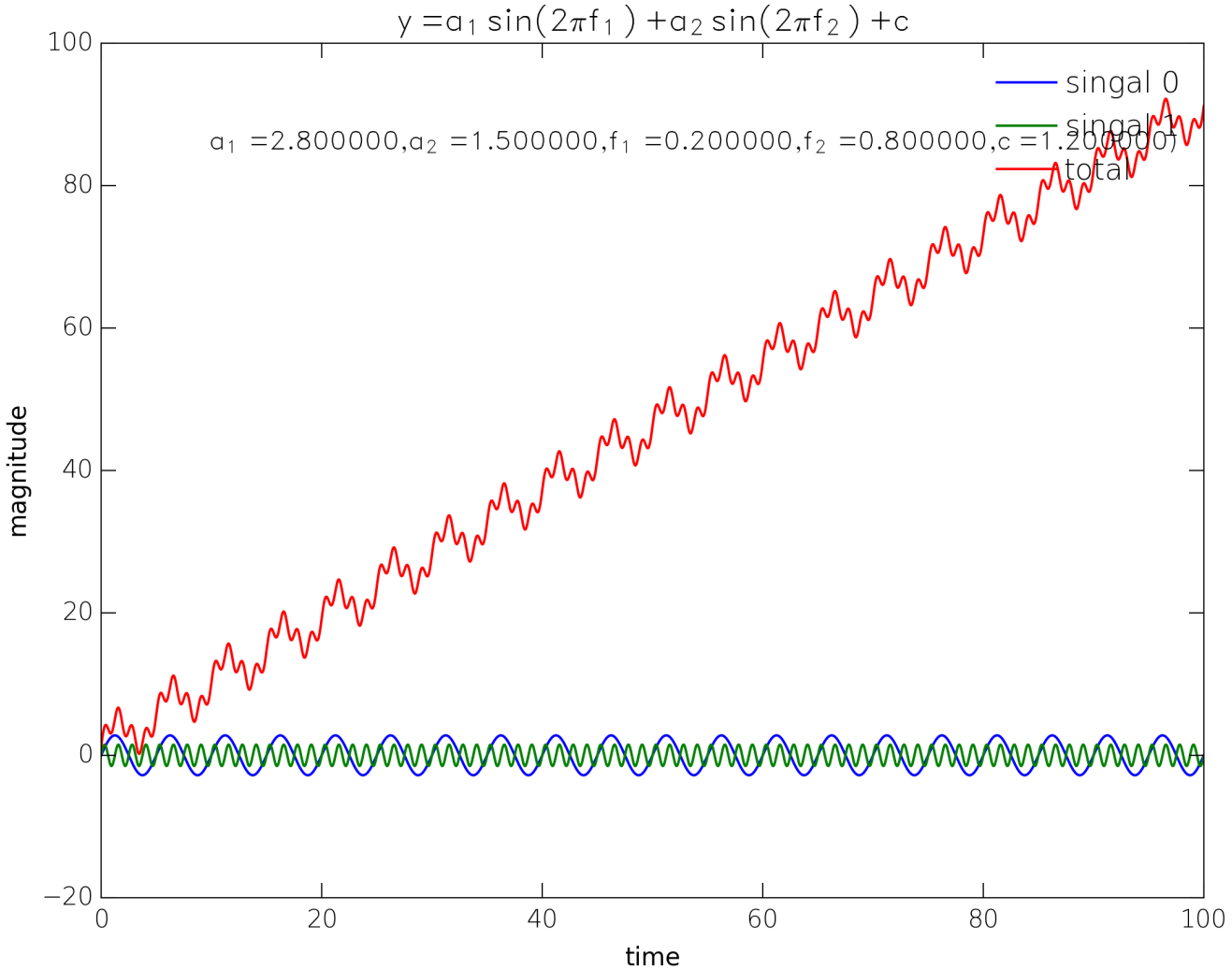
Light Curve: Two periods with small trend



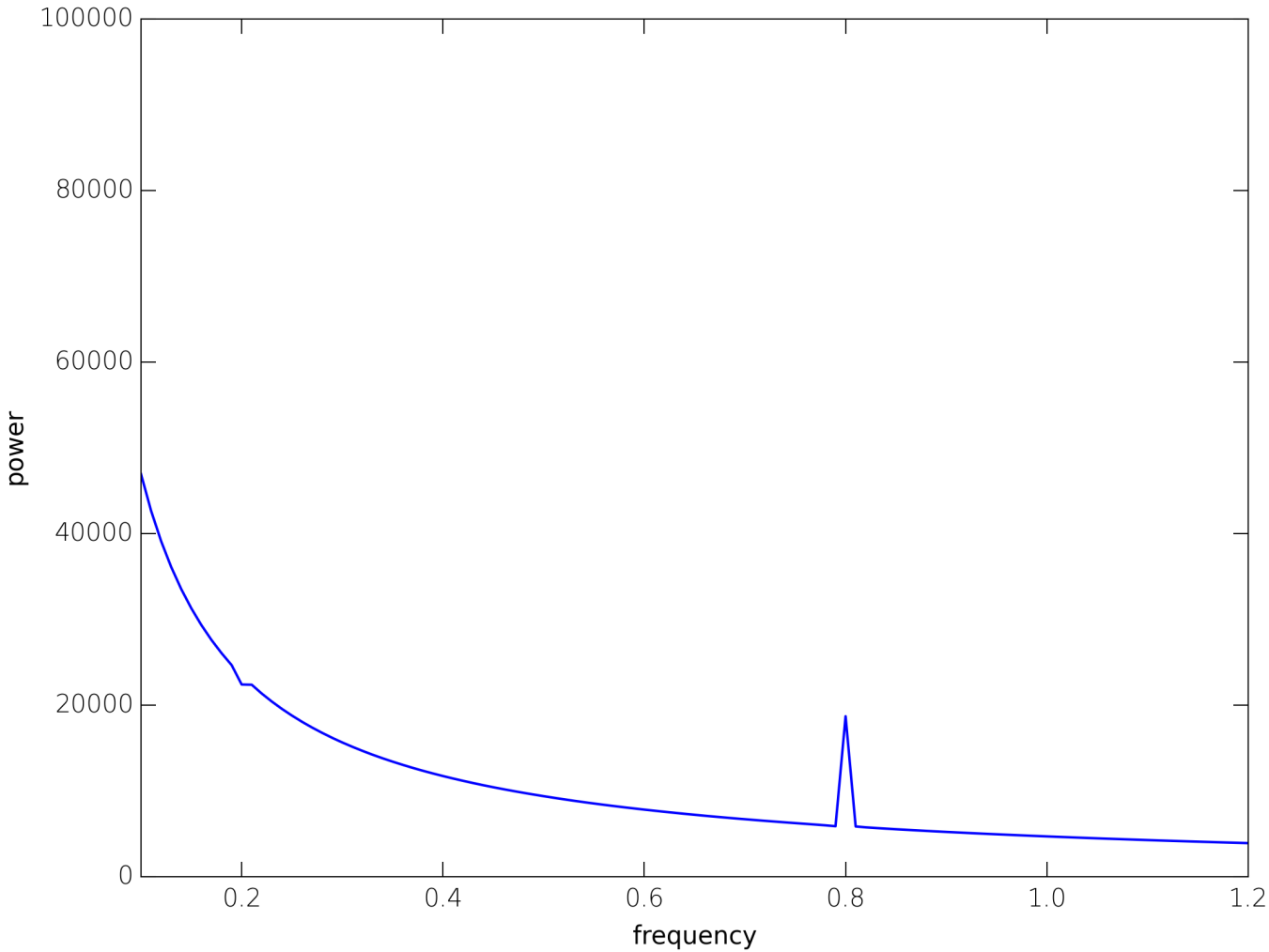
Fourier Power Spectrum



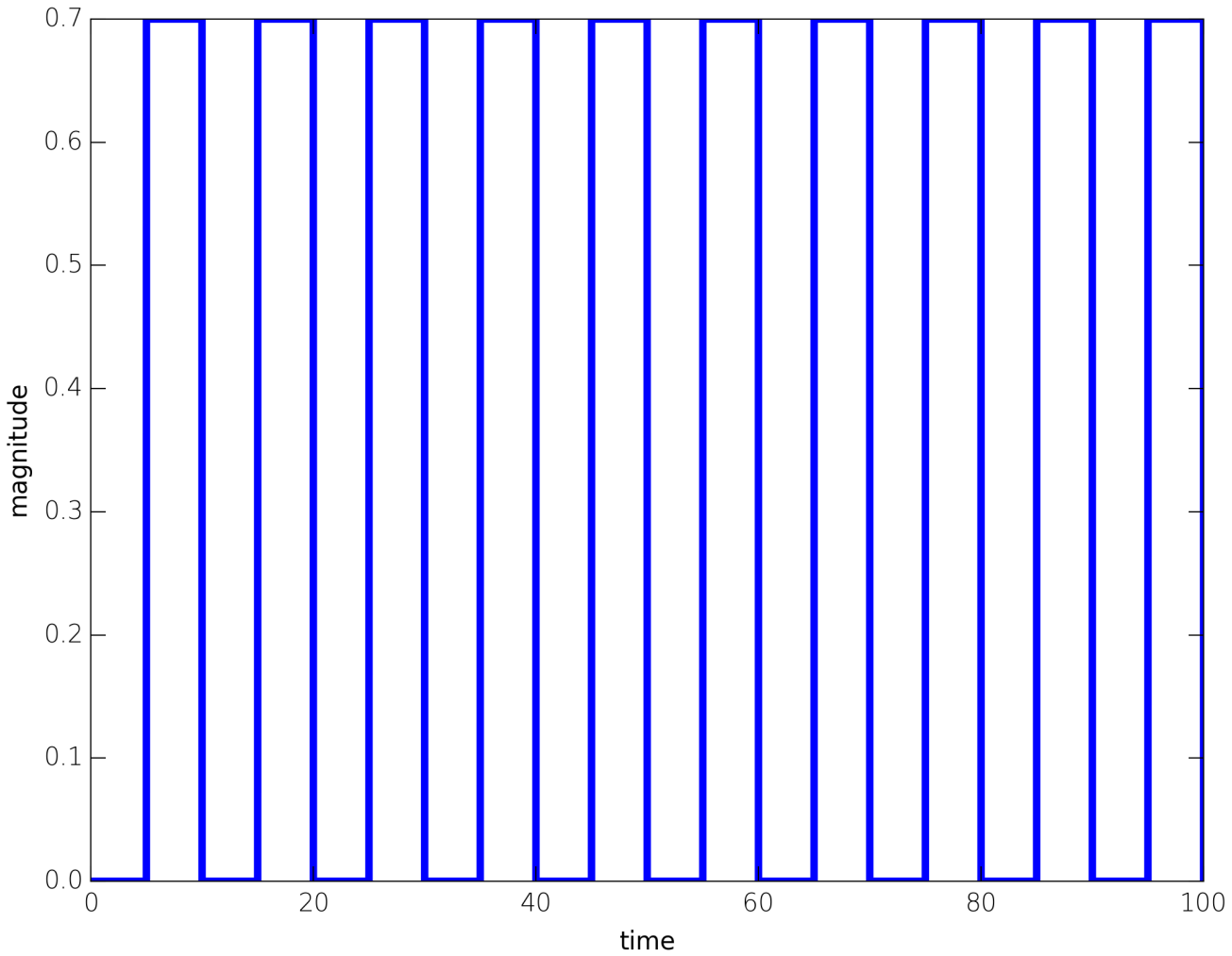
Light Curve: Two periods with big trend



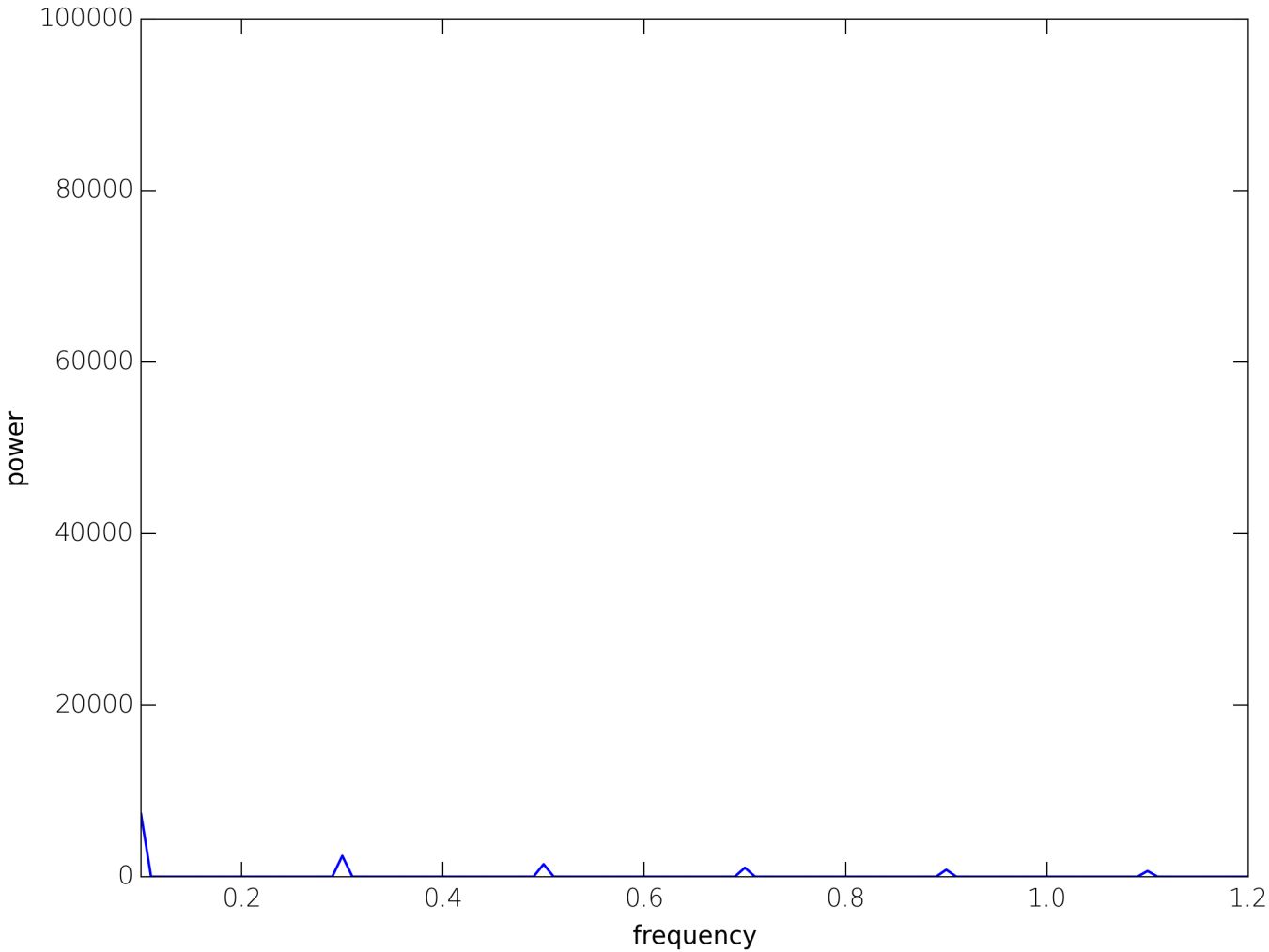
Fourier Power Spectrum



Light Curve: step function



Fourier Power Spectrum

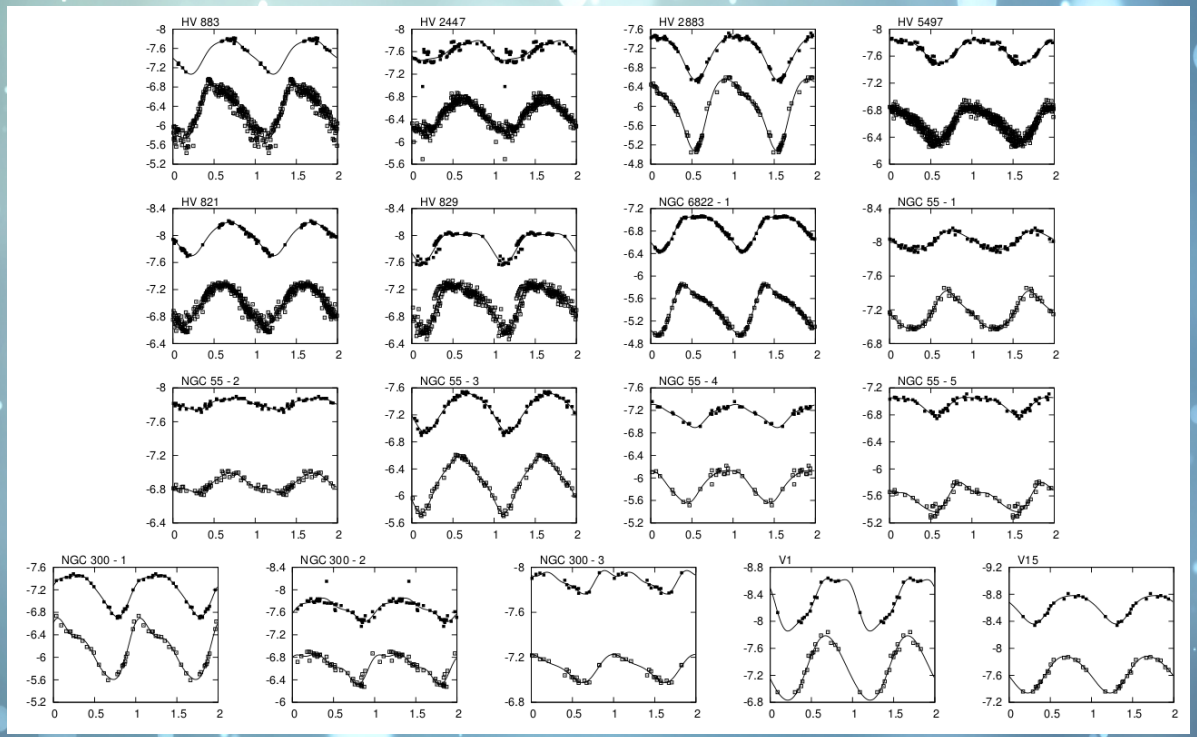


FFT 的優缺點

- 優點:
  - 快：Reduce the complexity from $O(n^2)$ to $O(n \log n)$
    - 如果原始資料點數為 10^8 則 FFT 比傳統 DFT 要快數百萬倍
  - 寫法簡單
  - 有許多現成函式庫可用
- 缺點:
  - 資料點要均勻分布
  - 資料點數要為 2^n
  - 原始資料中的 trend, gap 會造成在 power spectrum 中的一些偏差

Fourier Decomposition Technique

- Use fitting with different components of Fourier series to decompose the Fourier components



Ngeow et al. (2013)

Gibbs Effect in Fourier Transformation

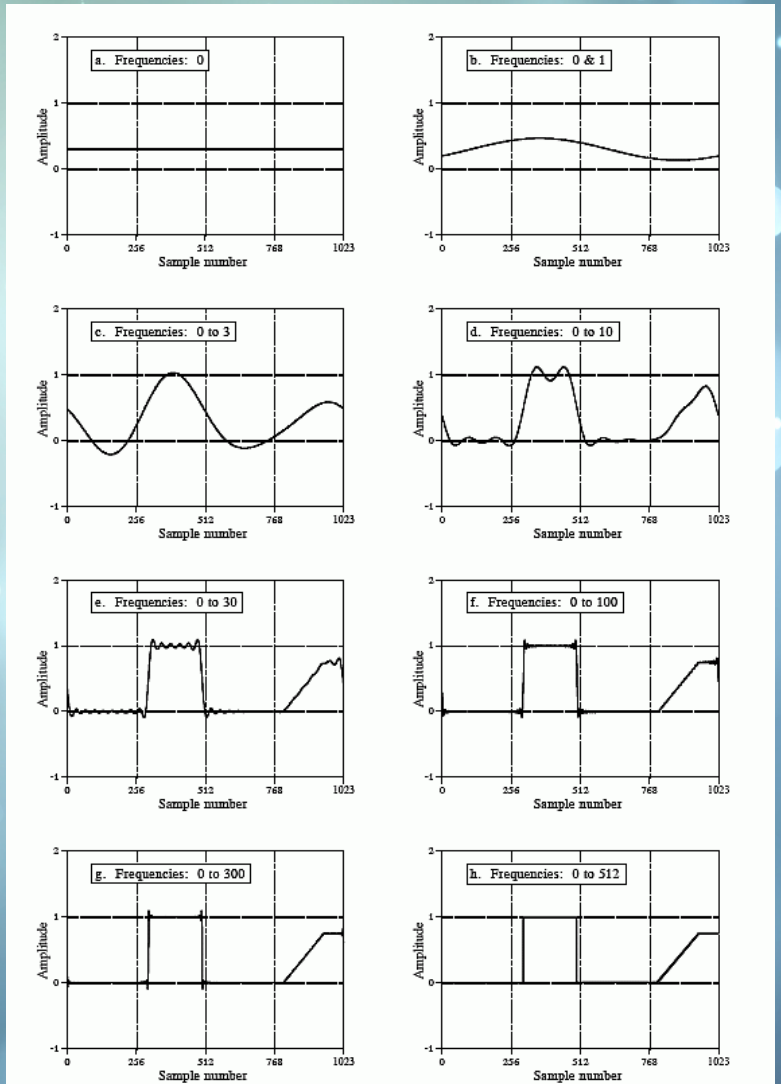


FIGURE 11-6.
The Gibbs effect.

有些事我們必須知道

- 觀測間隙 (observation gap) 在真實世界中一定存在
- 觀測資料點的時間分佈常常不是均勻的
- 資料趨勢不明 (uncertain trend of the raw data)

如何解決/減少這些問題？

- 切段
  - 可能衍生新問題：資料點不足、訊號不夠強
- Fitting
  - 可能的解法
- 很多科學/數學家嘗試解決這些問題，各有不同優缺點

Lomb's Paper

LEAST-SQUARES FREQUENCY ANALYSIS OF UNEQUALLY SPACED DATA

N. R. LOMB

School of Physics, University of Sydney, N.S.W., Australia

(Received 15 May, 1975)

Abstract. The statistical properties of least-squares frequency analysis of unequally spaced data are examined. It is shown that, in the least-squares spectrum of gaussian noise, the reduction in the sum of squares at a particular frequency is a χ_2^2 variable. The reductions at different frequencies are not independent, as there is a correlation between the height of the spectrum at any two frequencies, f_1 and f_2, which is equal to the mean height of the spectrum due to a sinusoidal signal of frequency f_1, at the frequency f_2. These correlations reduce the distortion in the spectrum of a signal affected by noise. Some numerical illustrations of the properties of least-squares frequency spectra are also given.

Scargle's Paper

STUDIES IN ASTRONOMICAL TIME SERIES ANALYSIS. II. STATISTICAL ASPECTS OF SPECTRAL ANALYSIS OF UNEVENLY SPACED DATA

JEFFREY D. SCARGLE
Theoretical and Planetary Studies Branch, Space Science Division, Ames Research Center, NASA
Received 1982 January 11; accepted 1982 April 26

ABSTRACT

Detection of a periodic signal hidden in noise is frequently a goal in astronomical data analysis. This paper does not introduce a new detection technique, but instead studies the reliability and efficiency of detection with the most commonly used technique, *the periodogram*, in the case where the observation times are unevenly spaced. This choice was made because, of the methods in current use, it appears to have the simplest statistical behavior. A modification of the classical definition of the periodogram is necessary in order to retain the simple statistical behavior of the evenly spaced case. With this modification, periodogram analysis and least-squares fitting of sine waves to the data are exactly equivalent. Certain difficulties with the use of the periodogram are less important than commonly believed in the case of detection of strictly periodic signals. In addition, the standard method for mitigating these difficulties (tapering) can be used just as well if the sampling is uneven. An analysis of the statistical significance of signal detections is presented, with examples.

Subject heading: numerical methods

Lomb-Scargle Periodogram

- Define a time delay τ

$$\tan 2\omega\tau = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j}$$

- The basic function will be orthogonal
- The power at frequency ω will be:

$$P_x(\omega) = \frac{1}{2}\left(\frac{[\sum_j X_j \cos\omega(t_j-\tau)]^2}{\sum_j \cos^2\omega(t_j-\tau)} + \frac{[\sum_j X_j \sin\omega(t_j-\tau)]^2}{\sum_j \sin^2\omega(t_j-\tau)}\right)$$

- Then, generate a periodogram for investigation

The advantages / disadvantages of LS periodogram

- Advantages:
  - Orthogonal set in each testing frequencies: avoid power leakage
  - Good for the unevenly spaced data
- Disadvantages:
  - Unweightted fitting
  - Trend
  - Originally designed for mono-frequency: power leakage if multiple frequencies are close enough

Fourier Analysis

- 基本上對於正弦波型式的變化會較有用
- 那非正弦波呢？

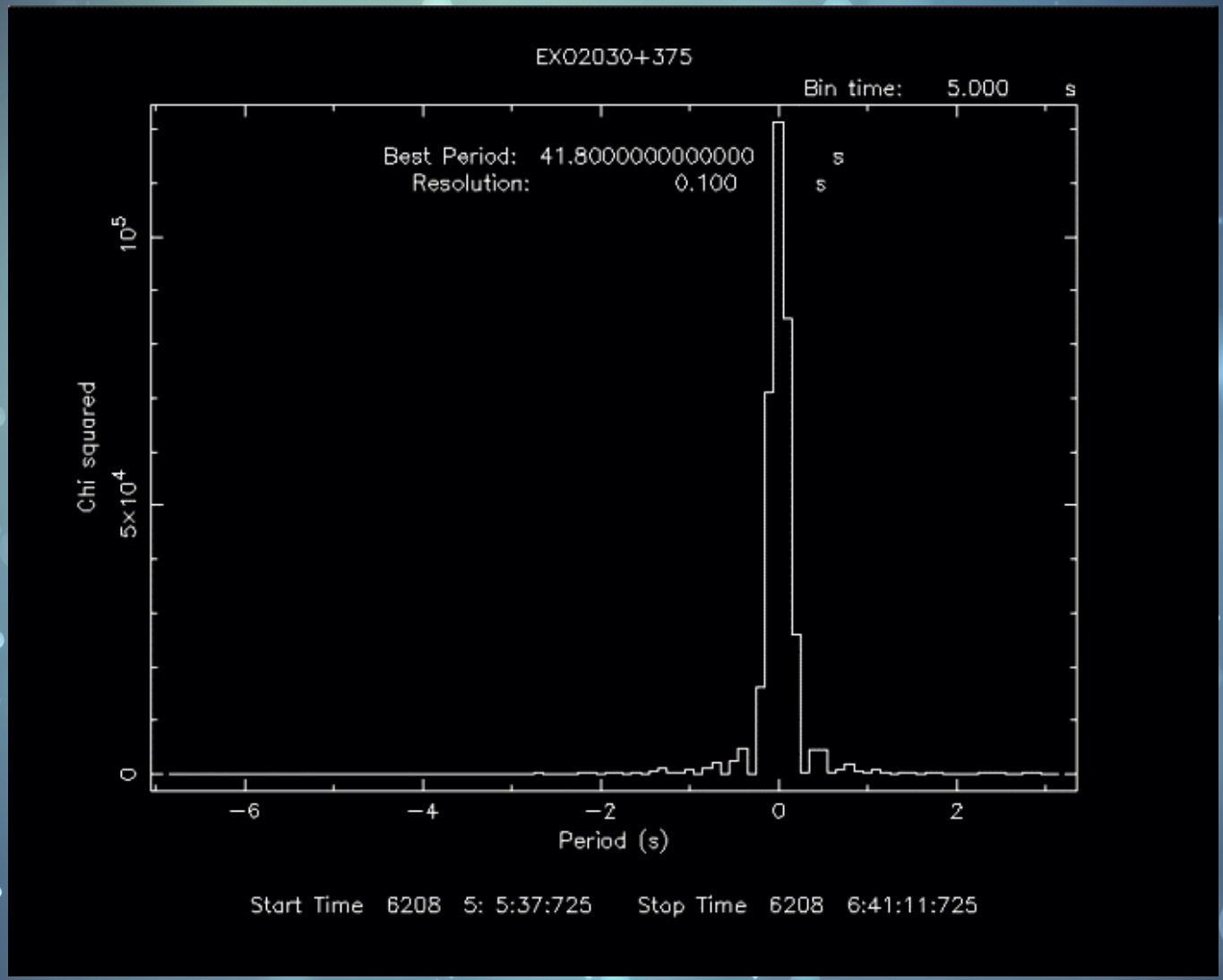
與相位有關的方法 Phase Based Methods

- Epoch folding: χ^2 maximazing
- Phase Dispersion Minimization (PDM)
- Analysis of Variance (AoV/ANOVA)

摺合法 Epoch Folding Search

- 我們要「猜」周期，所以先決定要猜的周期的範圍
- 將光變曲線對 trial period 作折合，決定每一點的相位
- 計算折合光變曲線 (folded light curve) 的 χ^2
- 假如 trial period 是「正確」的，那理論上折合光變曲線的 χ^2 會有最大值

摺合法 Epoch Folding Search (cont.)



相位差最小化 Phase Dispersion Minimization

PERIOD DETERMINATION USING PHASE DISPERSION MINIMIZATION

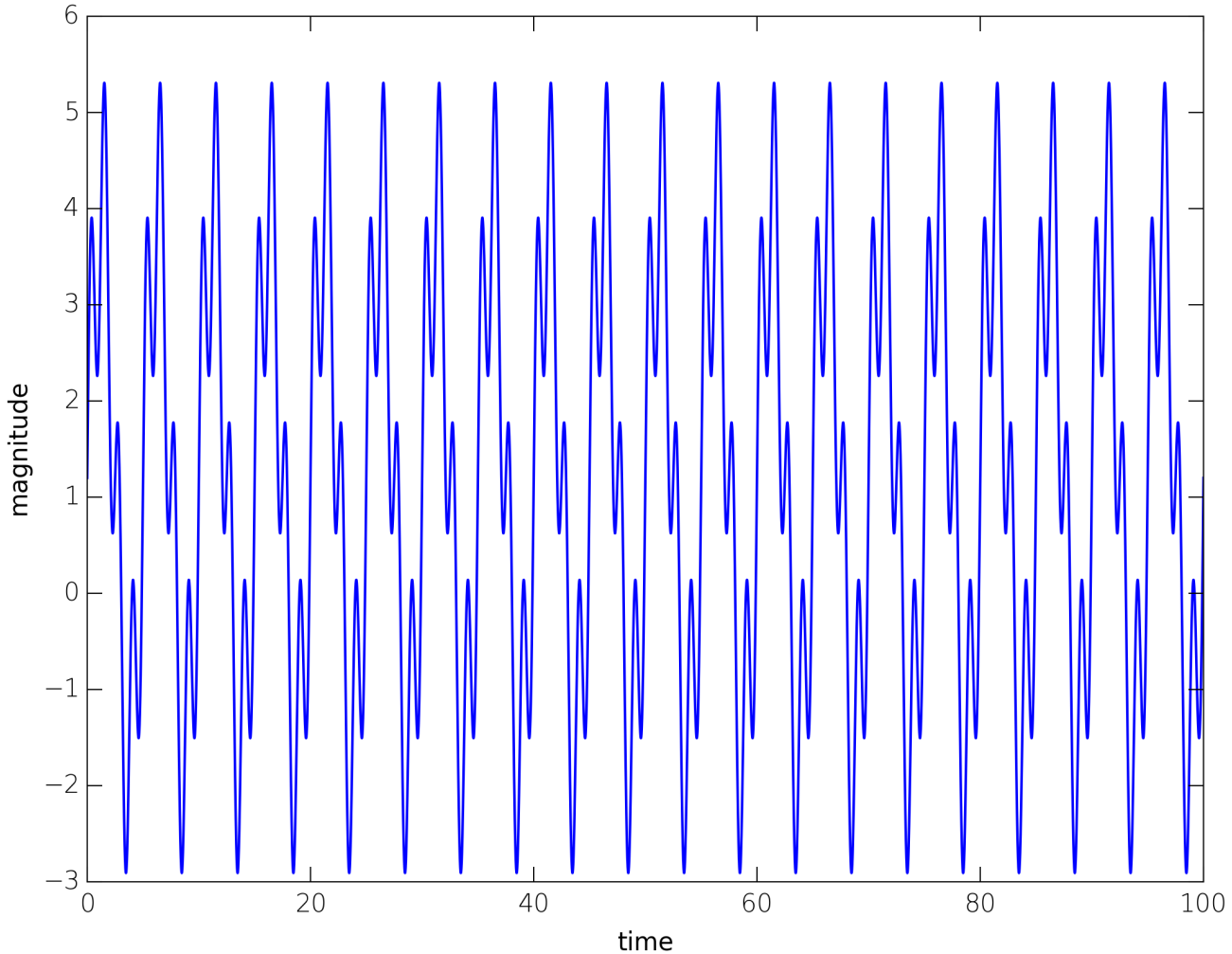
R. F. STELLINGWERF
Department of Physics and Astronomy, Rutgers University
Received 1978 February 6; accepted 1978 March 22

ABSTRACT

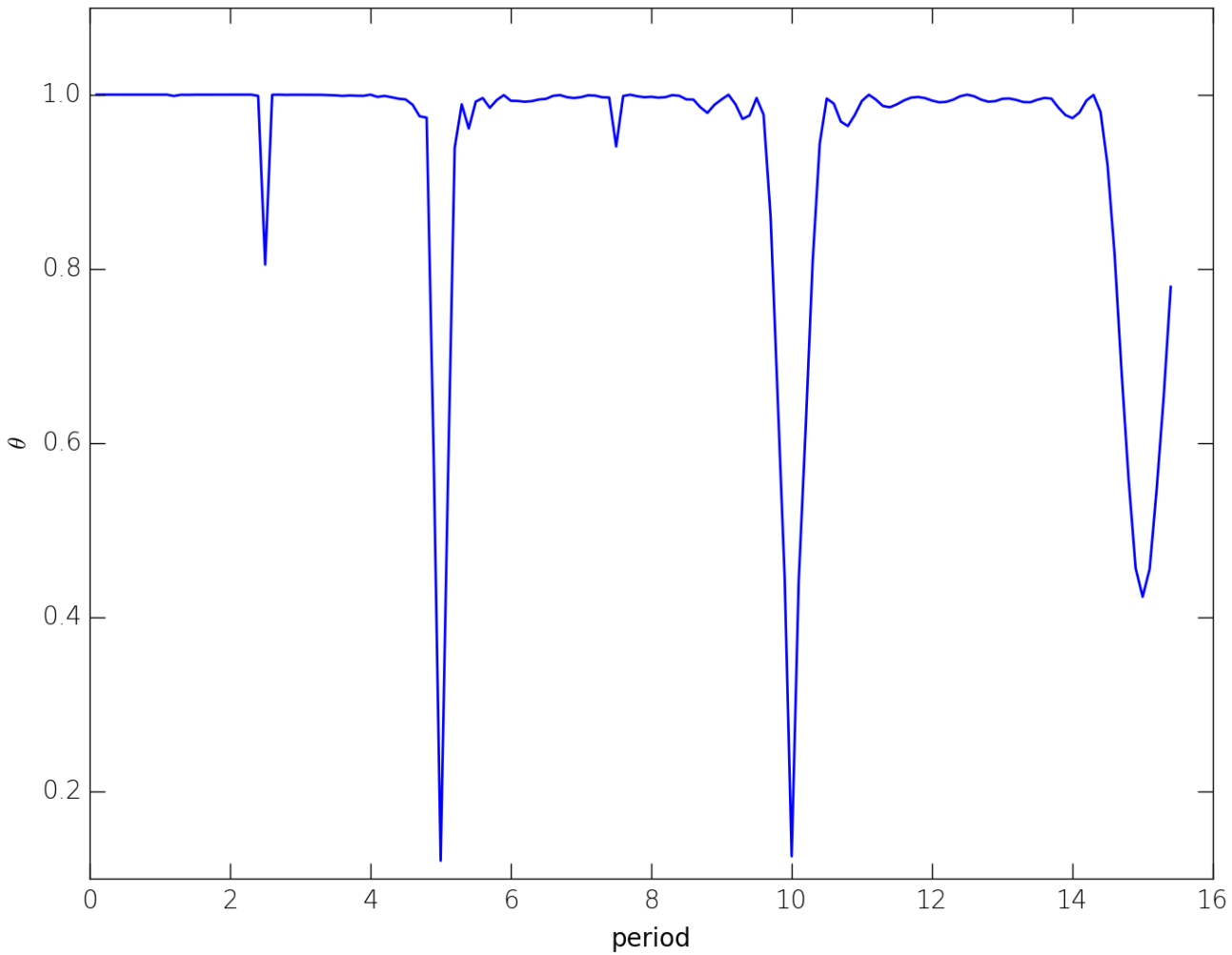
We derive a period determination technique that is well suited to the case of nonsinusoidal time variation covered by only a few irregularly spaced observations. A detailed statistical analysis allows comparison with other techniques and indicates the optimum choice of parameters for a given problem. Application to the double-mode Cepheid BK Cen demonstrates the applicability of these methods to difficult cases. Using 49 photoelectric points, we obtain the two primary oscillatory components as well as the principal mode-interaction term; the derived periods are in agreement with previous estimates.

Subject headings: stars: Cepheids — stars: individual — stars: pulsation

Light Curve: two sinsoidal function



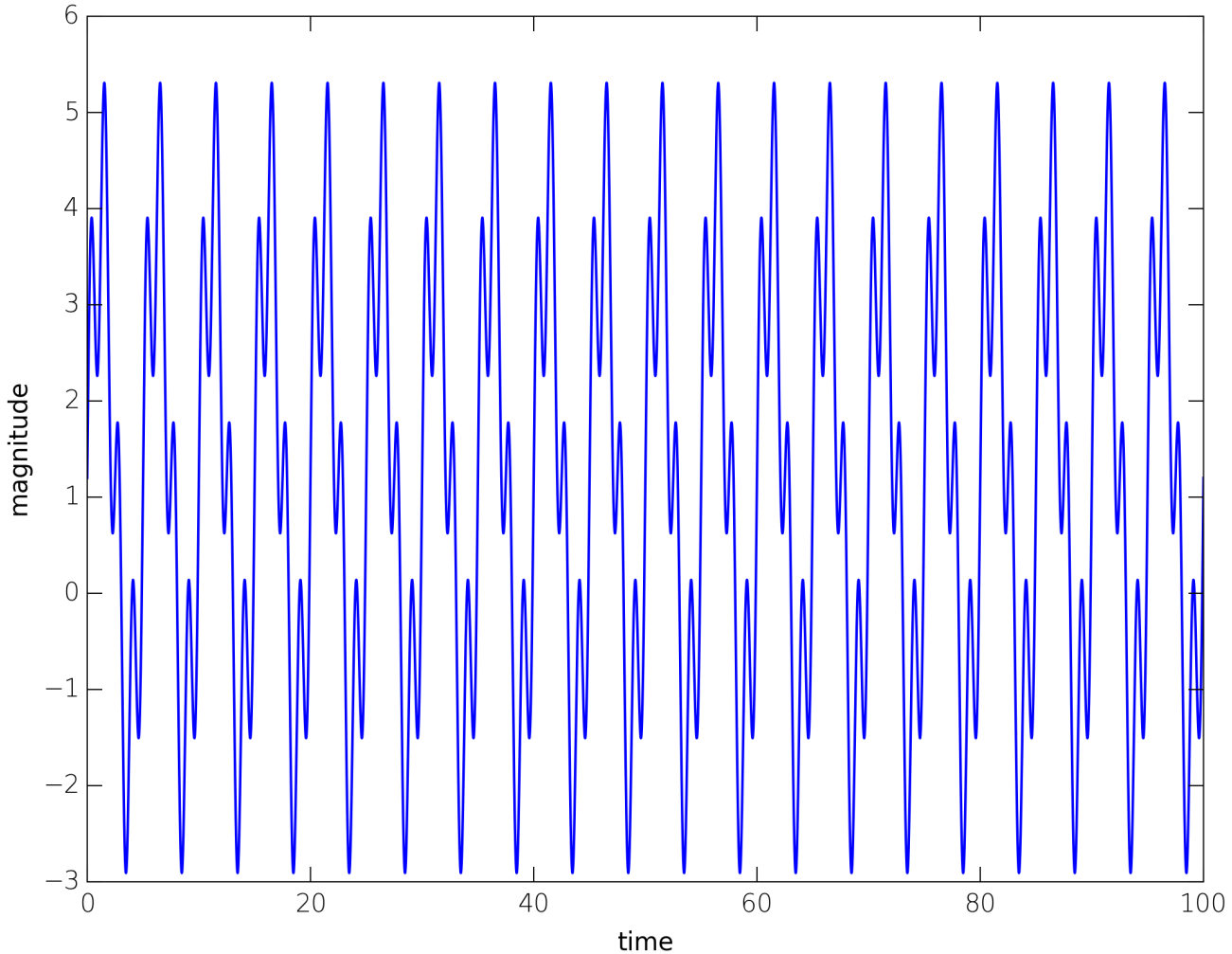
Periodogram



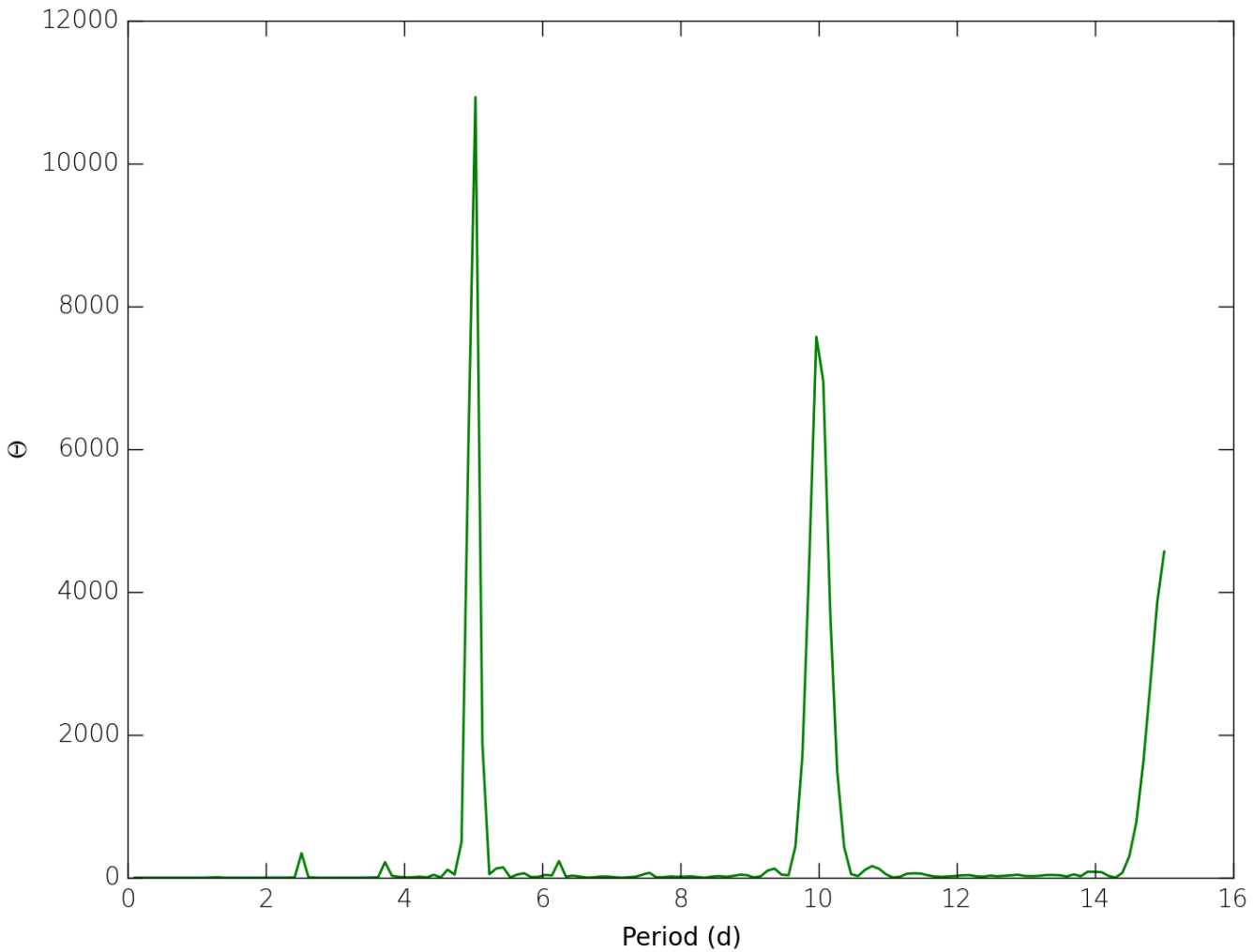
變異數分析 Analysis of Variance (AoV, ANOVA)

- Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated procedures, developed by statistician and evolutionary biologist Ronald Fisher.

Light Curve: two sinsoidal function



Periodogram



Periodogram / Spectrogram

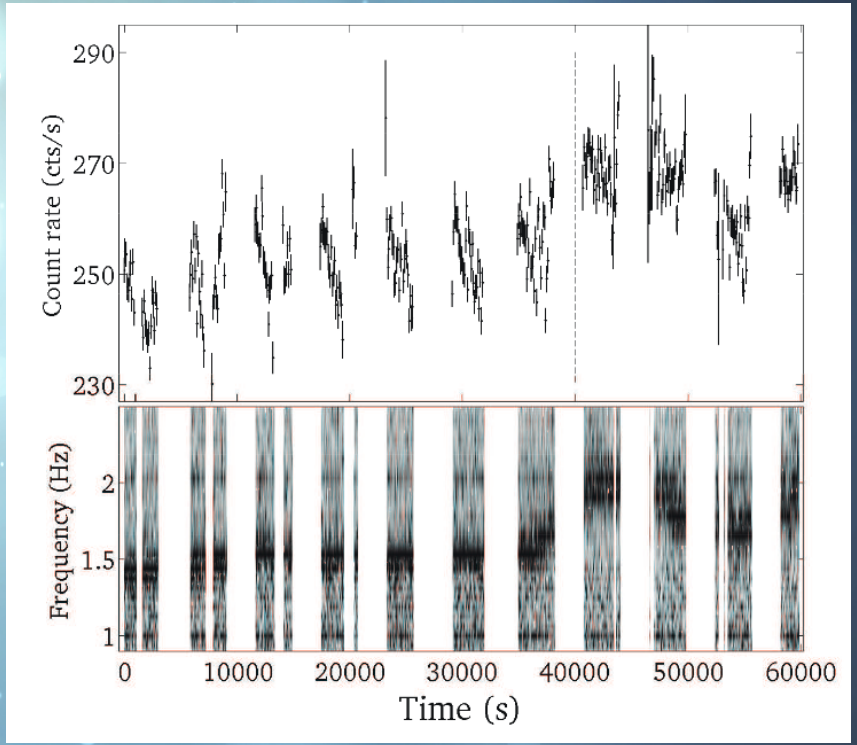
- Periodogram: Time – frequency
- Spectrogram: Time – frequency – power
- **所涵蓋的資訊範圍不同，應用層也不相同**

非穩定時序 – 周期非定值

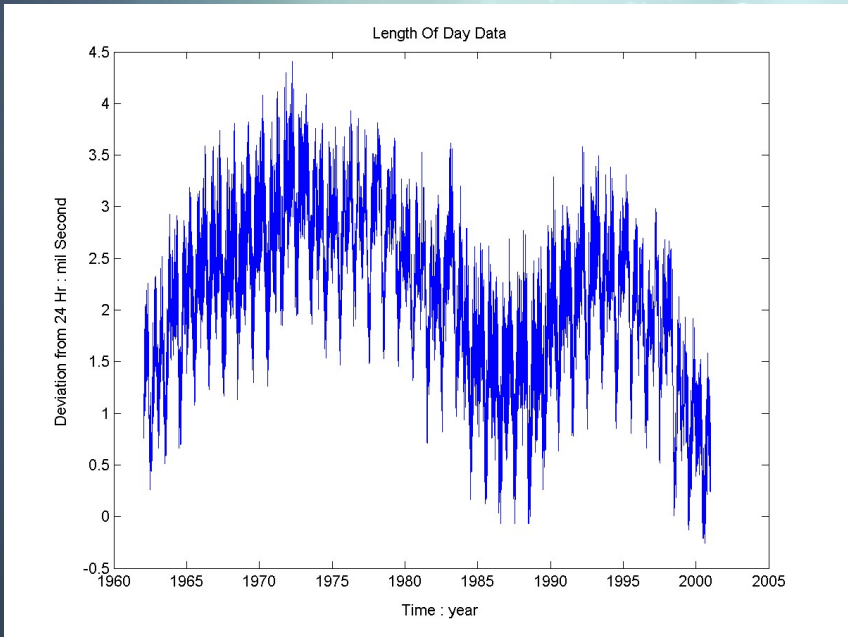
- **動態功率譜**: Dynamical Power Spectrum
- **希爾伯特-黃變換**: HHT
- **小波分析**: Wavelet

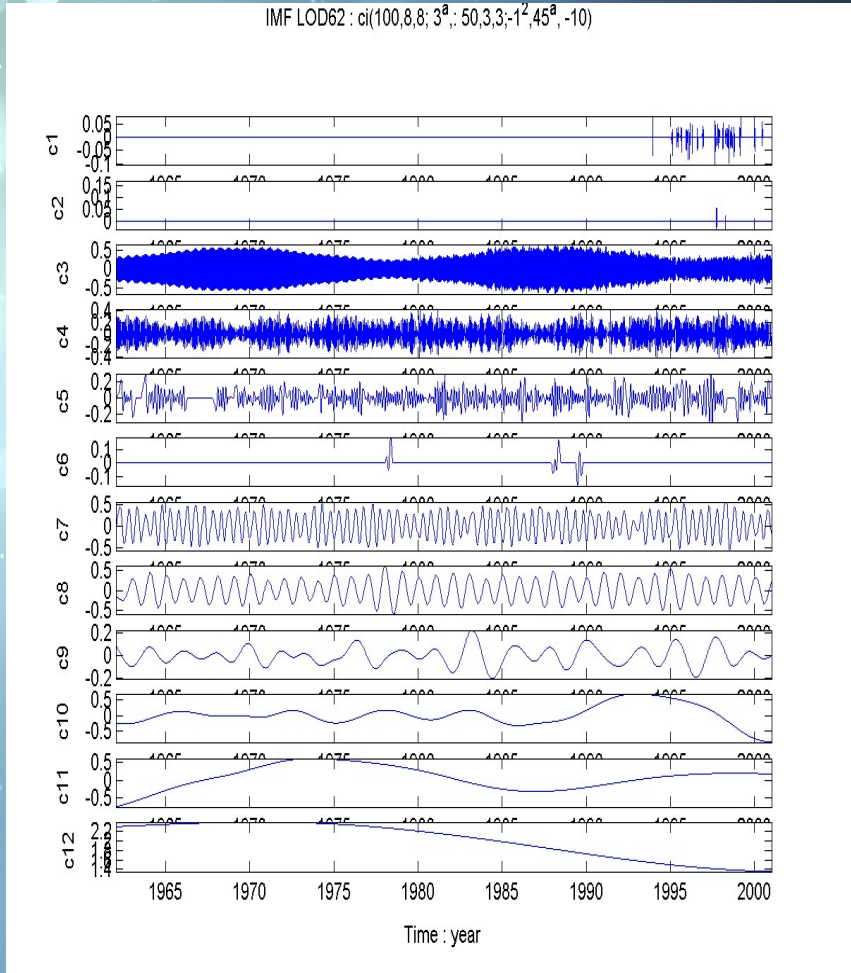
Dynamical Power Spectrum

- **將光變曲線依時間「分段」作周期分析**
- **將結果整合，即成** dynamical power spectrum

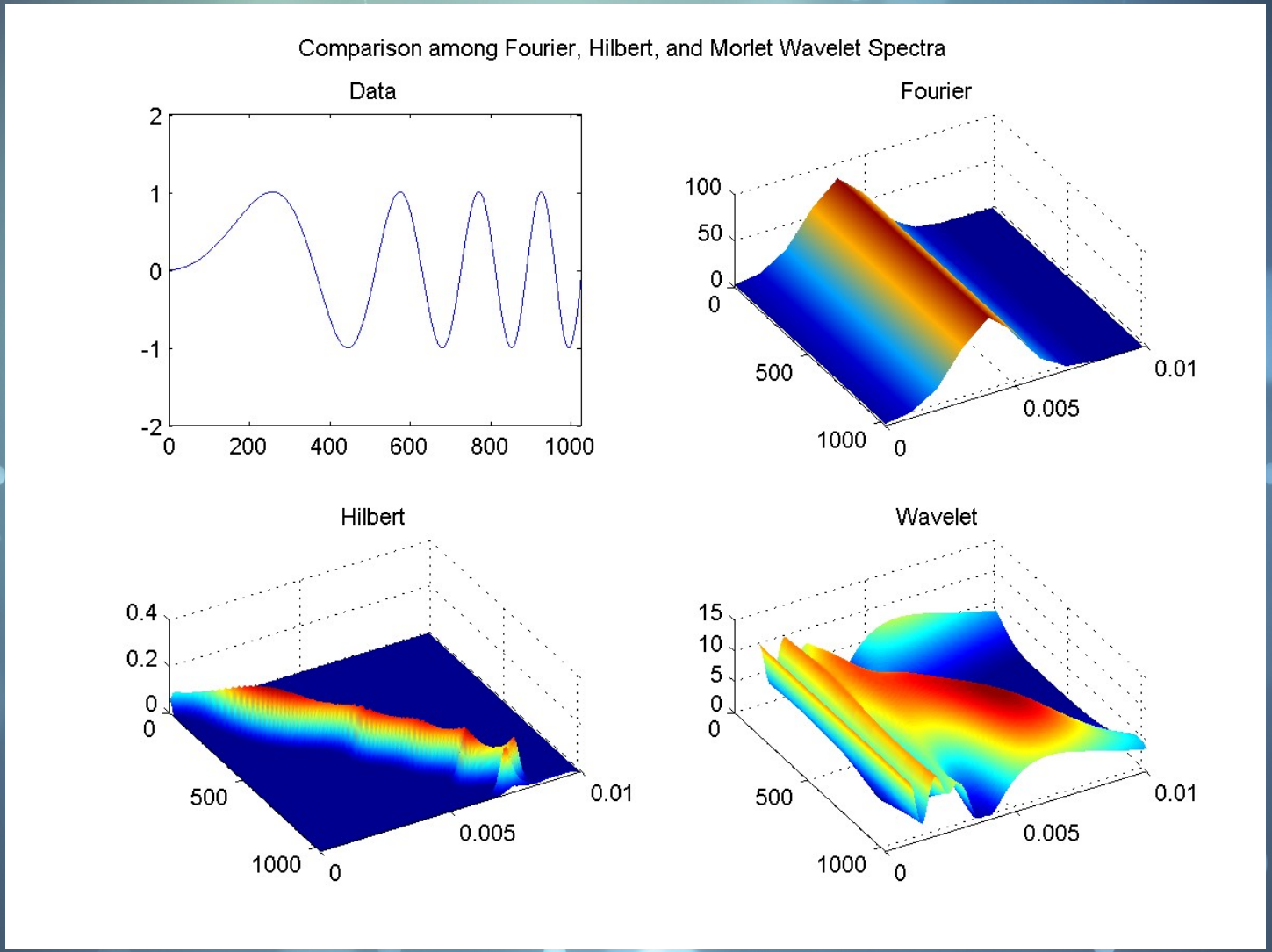


Hilbert-Huang Transformation





Comparison of Spectrograms



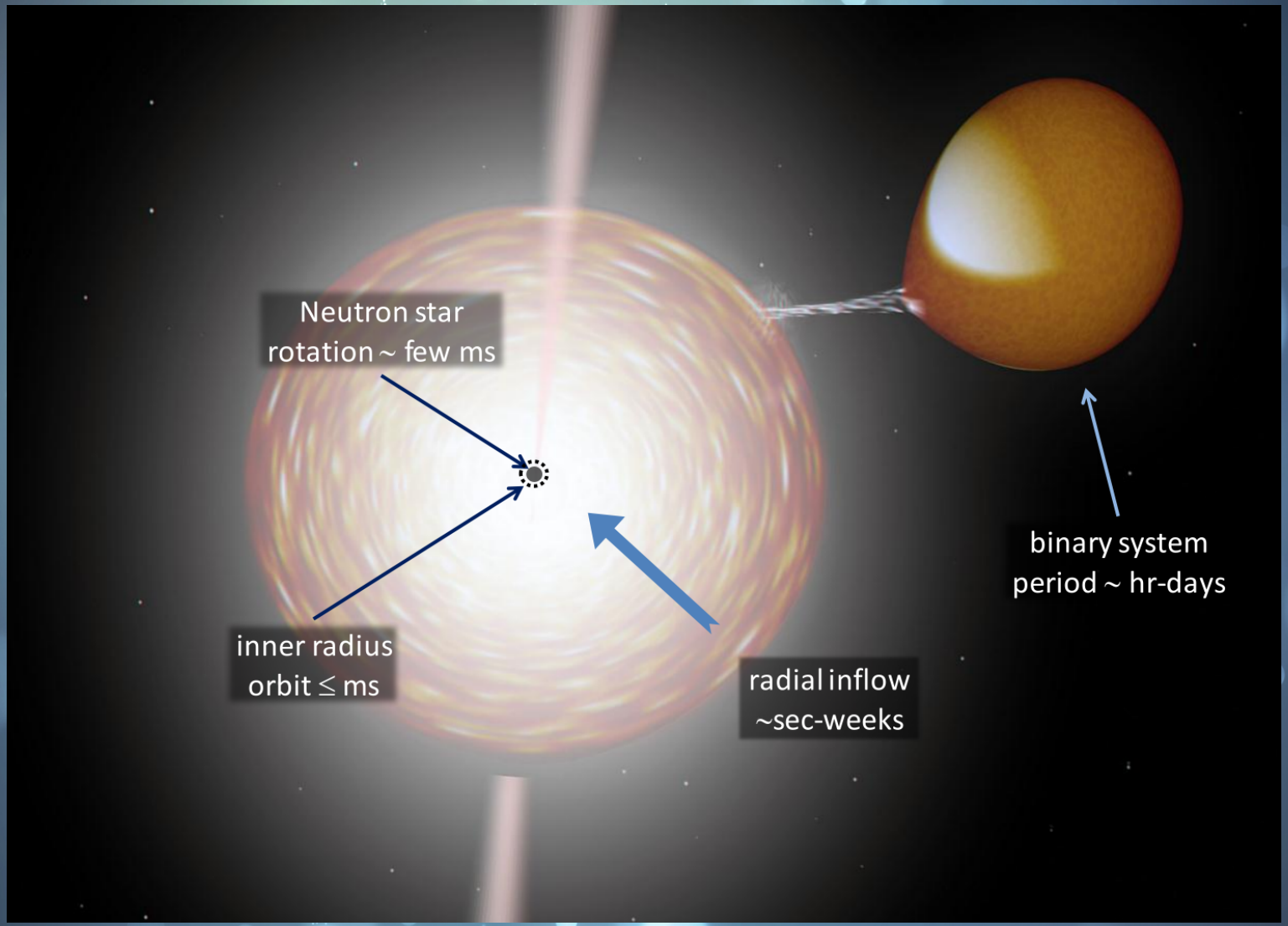
影響時序資料品質的因素

- Data Quality
  - Error of the measurement
  - Time span
  - sampling frequency
- Observation gap
- Alias

Better data:

- Data with smaller errors
- Continuous data: less gaps
- Longer time span of the data

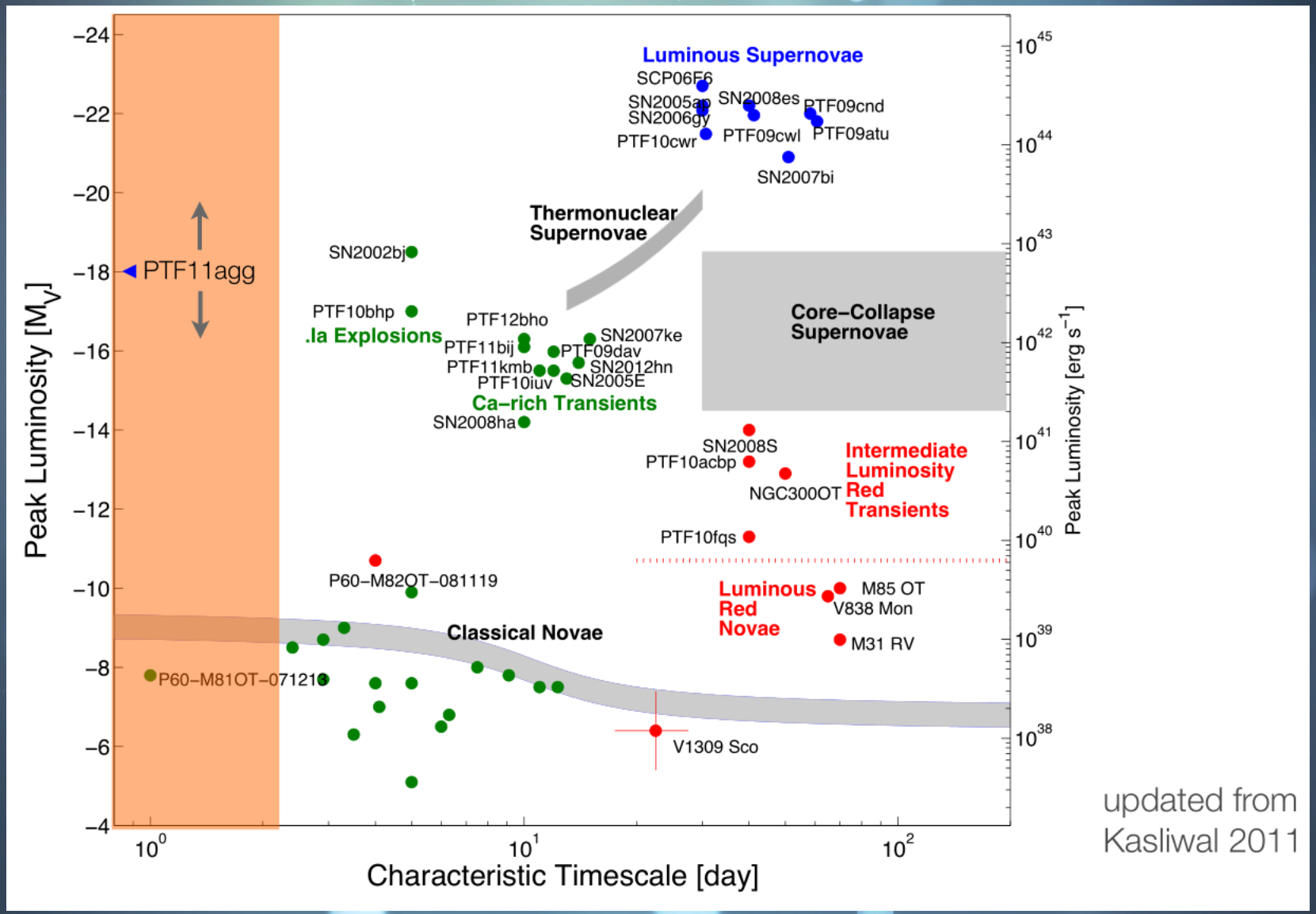
一個物理系統不同時間尺度 – 緻密雙星系統



時間序列分析流程

- 目標是什麼？可能是什麼？
- 前人研究的成果有那些？
- 作出光變曲線
- 肉眼觀察：有無變化，可見的準周期？形狀
- 測試可能的周期性
  - 正弦波形: Fourier Analysis
  - 極不規則形: Phase based analysis
  - 或兩個都試試？
  - 其他新的方法？
- 分析規納結果
- 再分析
- 再再分析
- ...

Phase Diagram of Astronomical Objects



Some Coding Resources

- Python:
  - `numpy.fft`
  - scipy.fftpack
  - scipy.signal.lombscargle
  - pynfftls
  - astroML.time_series.lomb_scargle
- C: FFTW
- R: lomb
- GUI: period04
  `https://www.univie.ac.at/tops/Period04/`
- Matlab: plomb, fft

相關天文計畫

- Kepler http://kepler.nasa.gov/
- PTF http://www.ptf.caltech.edu/
- CSS/CRTS http://crts.caltech.edu/
- ASAS http://www.astrouw.edu.pl/asas/
- Pan-STARRS http://pan-starrs.ifa.hawaii.edu/public/
- LSST http://www.lsst.org/
- GAIA http://sci.esa.int/gaia/
- **當然，還有很多很多...**

It's the Golden Era of Astronomy

- 目前是天文的黃金時期
- 天文學家可以拿到許多不同波段，不同目的的望遠鏡
- 目前主要欠缺人力來作適當的分析，及資訊獲取
- 歡迎你的加入