

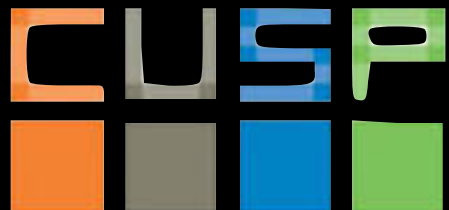
# Urban Informatics

Fall 2015

dr. federica bianco [fb55@nyu.edu](mailto:fb55@nyu.edu)



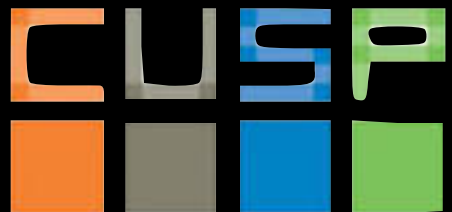
@fedhere



*don't forget to vote!*

## Recap:

- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- Basic statistics: distributions and their moments
- Hypothesis testing:  $p$ -value, statistical significance
- Statistical and Systematic errors
- Goodness of fit tests
- OLS, residual minimization
- Likelihood, chisq



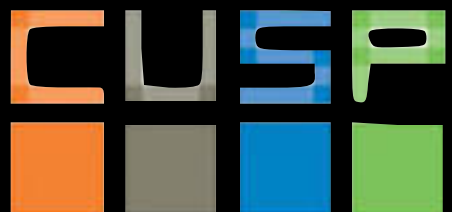
*don't forget to vote!*

## Recap:

- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- Basic statistics: distributions and their moments
- Hypothesis testing:  $p$ -value, statistical significance
- Statistical and Systematic errors
- Goodness of fit tests
- OLS, residual minimization
- Likelihood, chisq

## Today

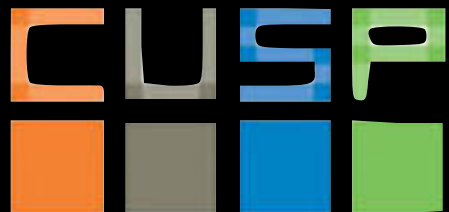
- Topics in (time) series analysis



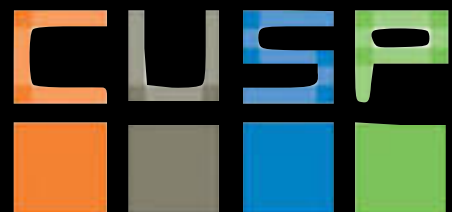
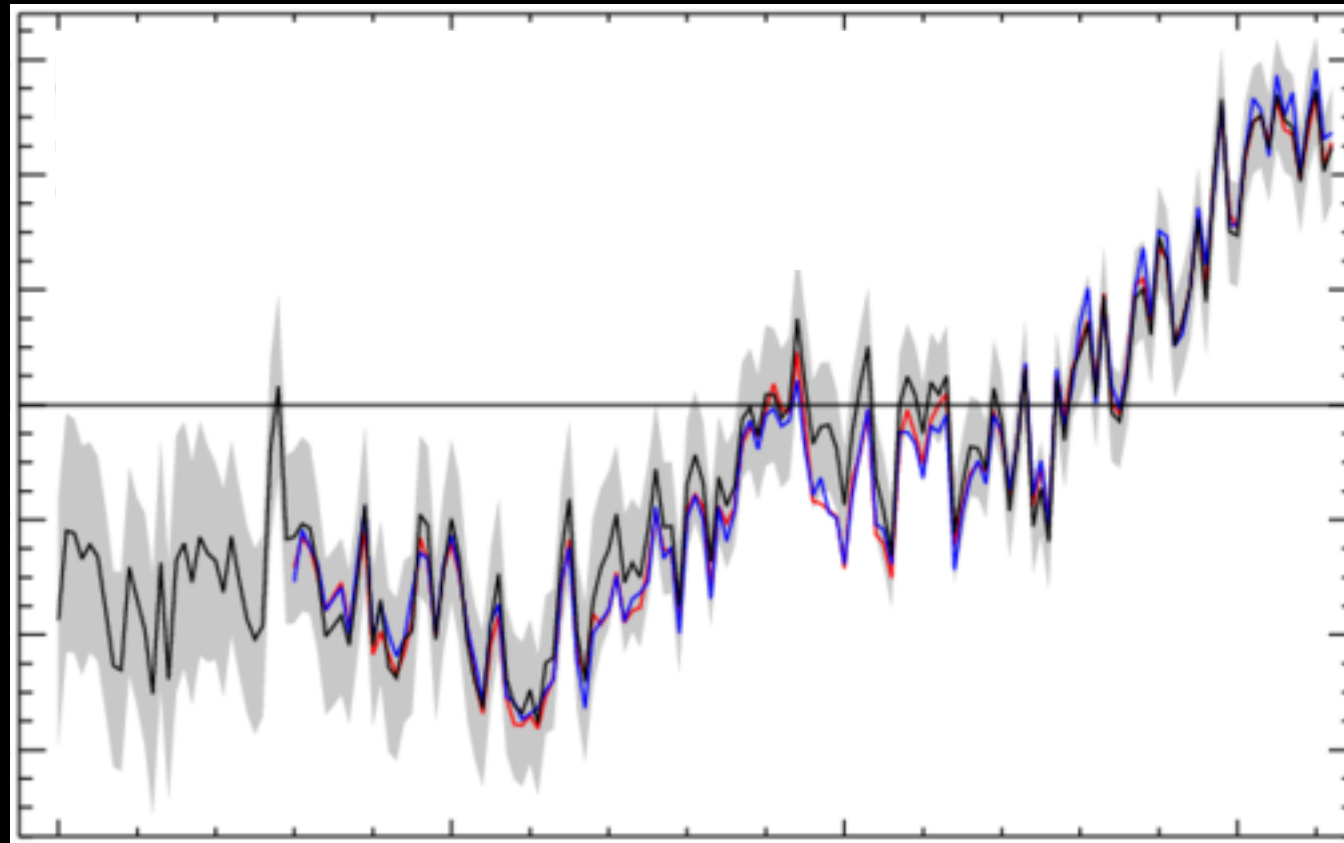
*don't forget to vote!*

# Topics in (time) series analysis

- smoothing
- de-trending
- event detection
- period finding (Fourier analysis)
- clustering

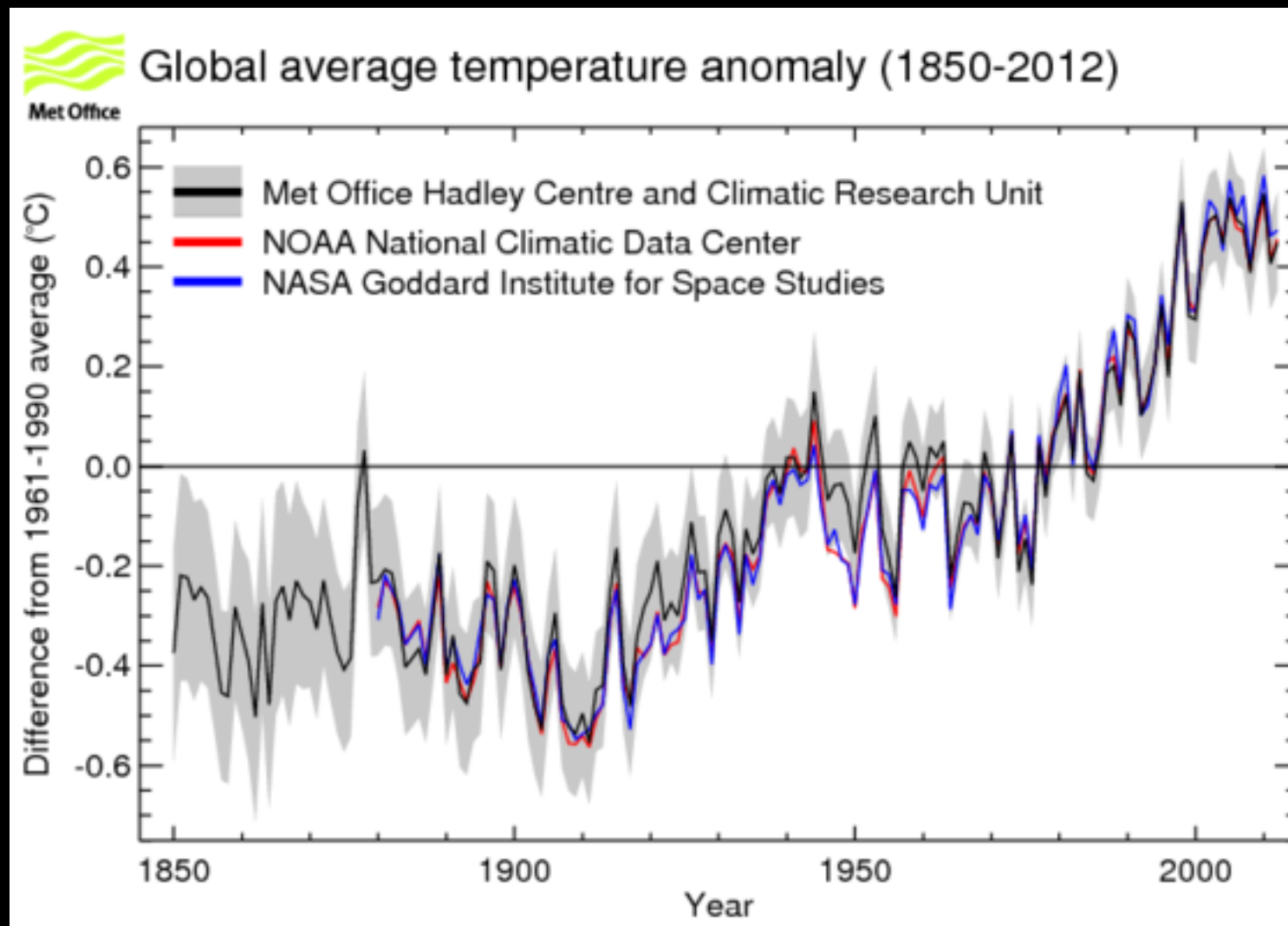


*don't forget to vote!*

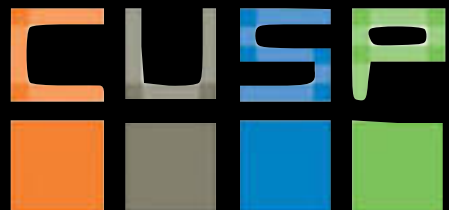


*don't forget to vote!*

VIII: Topics in Time series

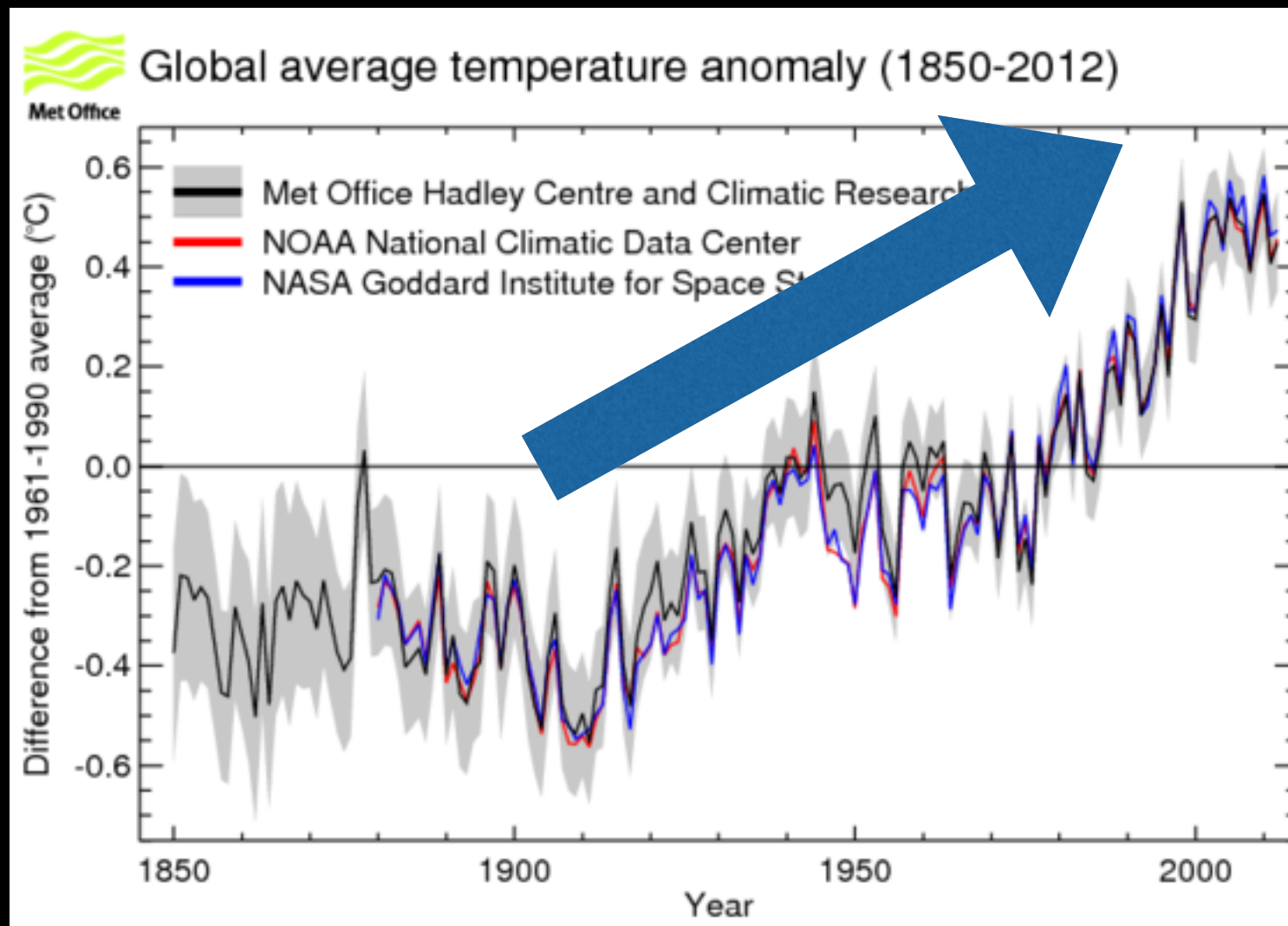


Trend

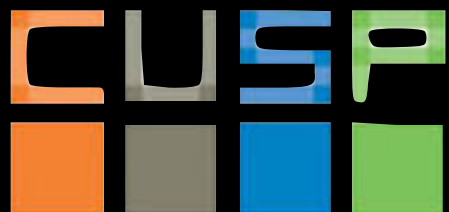


*don't forget to vote!*

VIII: Topics in Time series

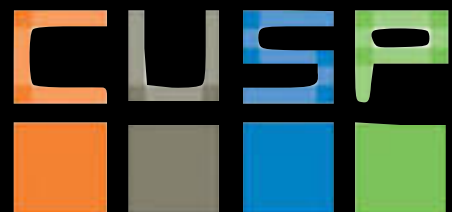
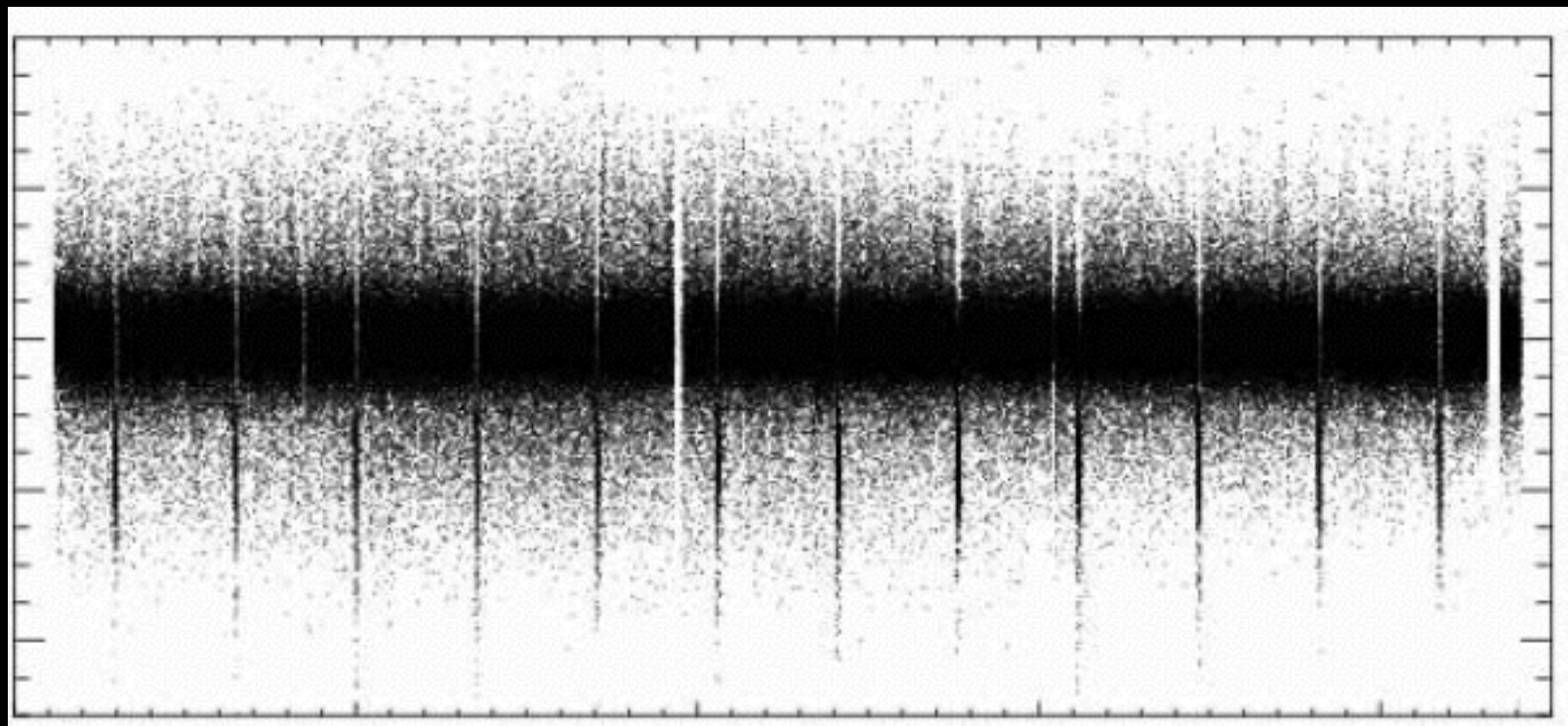


## Trends



*don't forget to vote!*

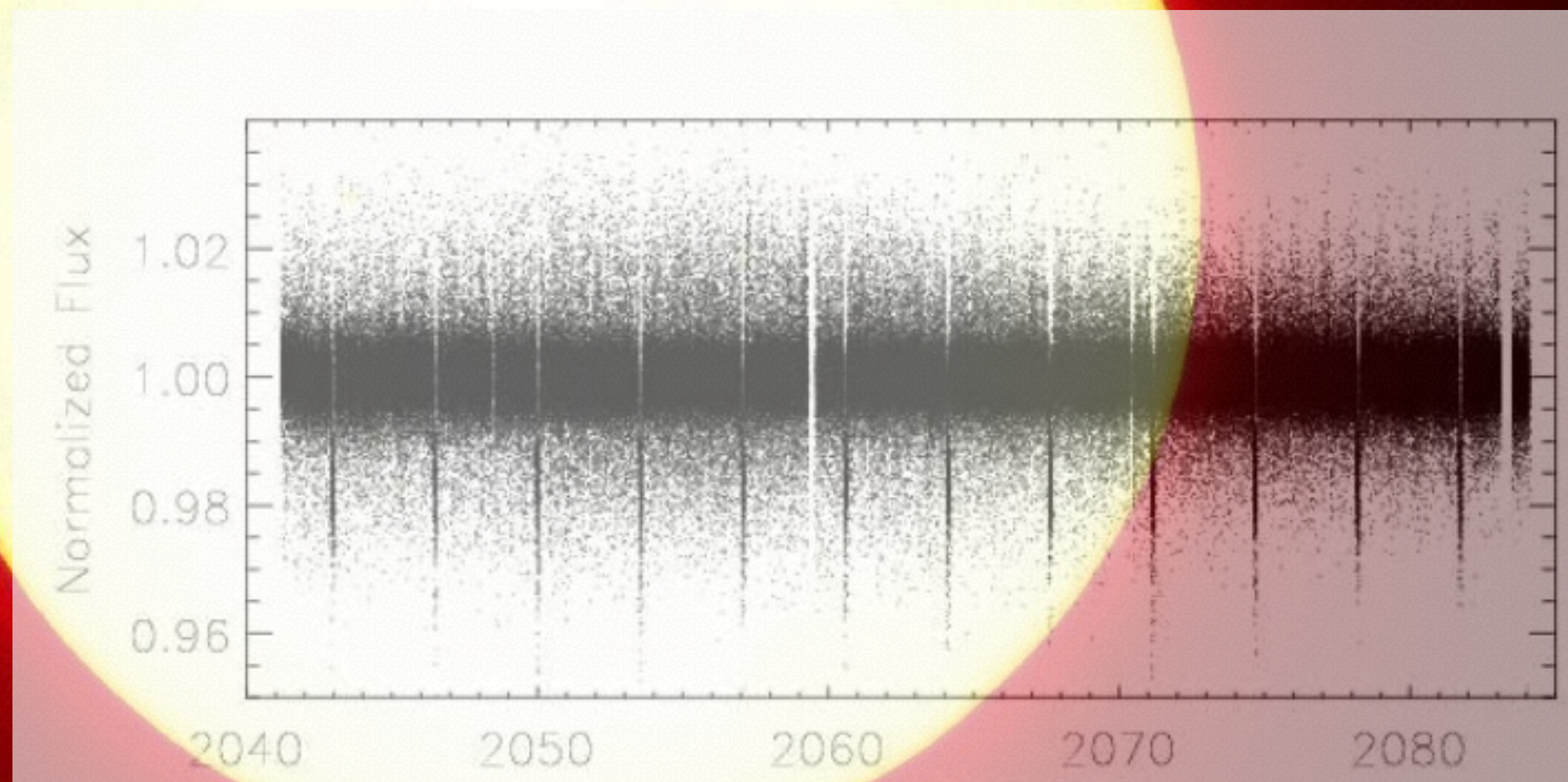
VIII: Topics in Time series



*don't forget to vote!*

VIII: Topics in Time series

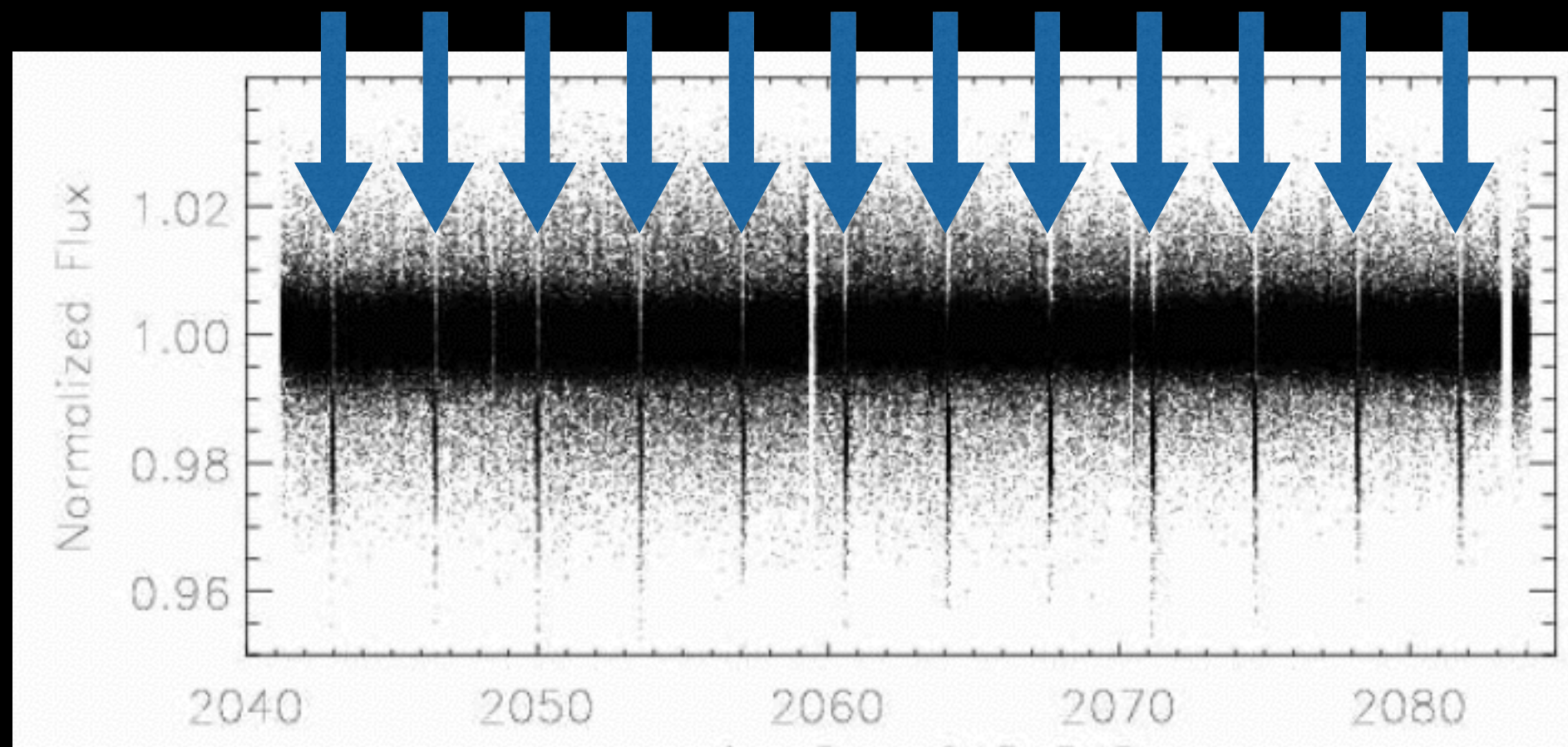




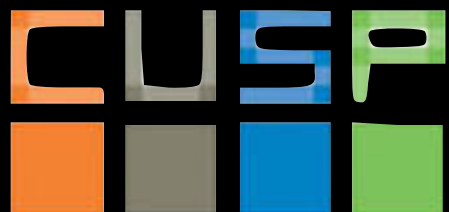
*don't forget to vote!*



HD 209458, the first transiting planet to be discovered.



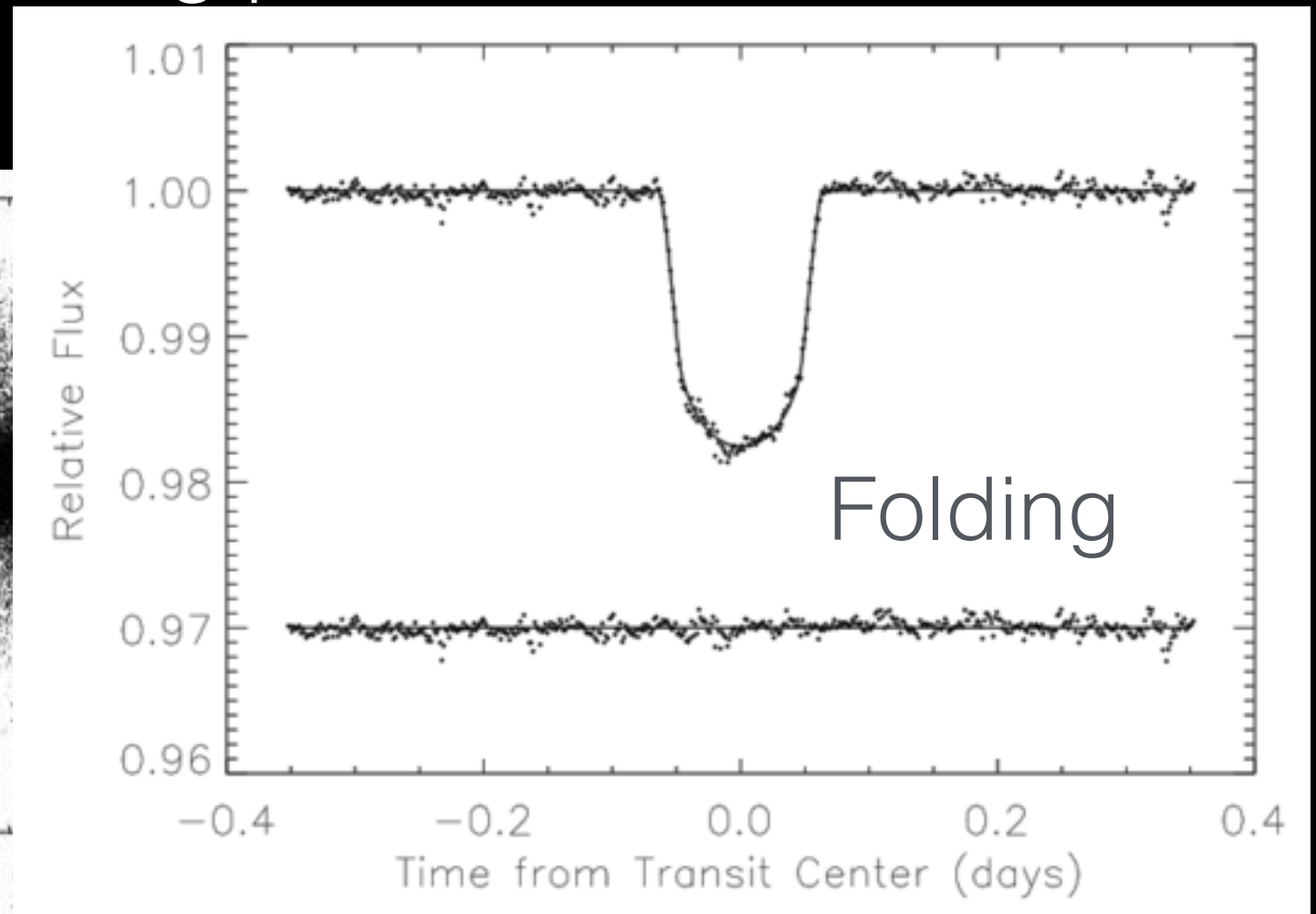
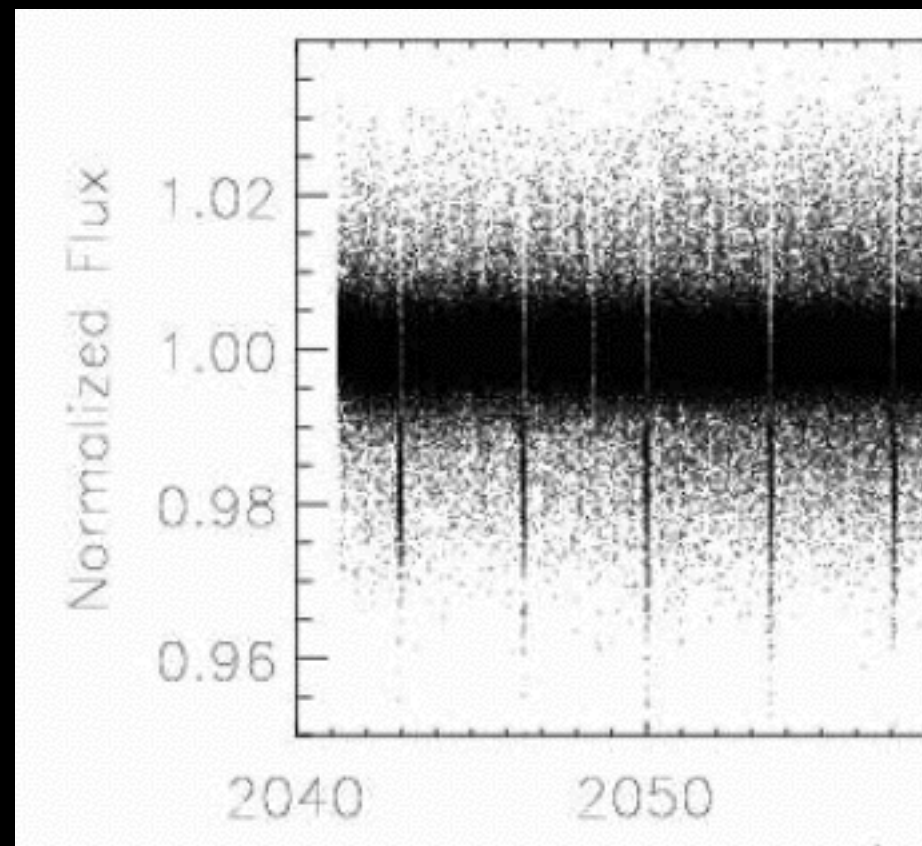
**Periodicity**



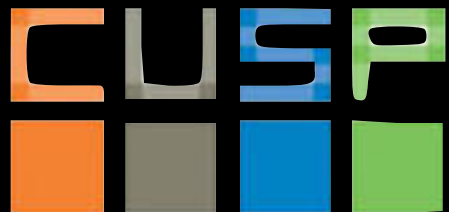
*don't forget to vote!*

VIII: Topics in Time series

HD 209458, the first transiting planet to be discovered.

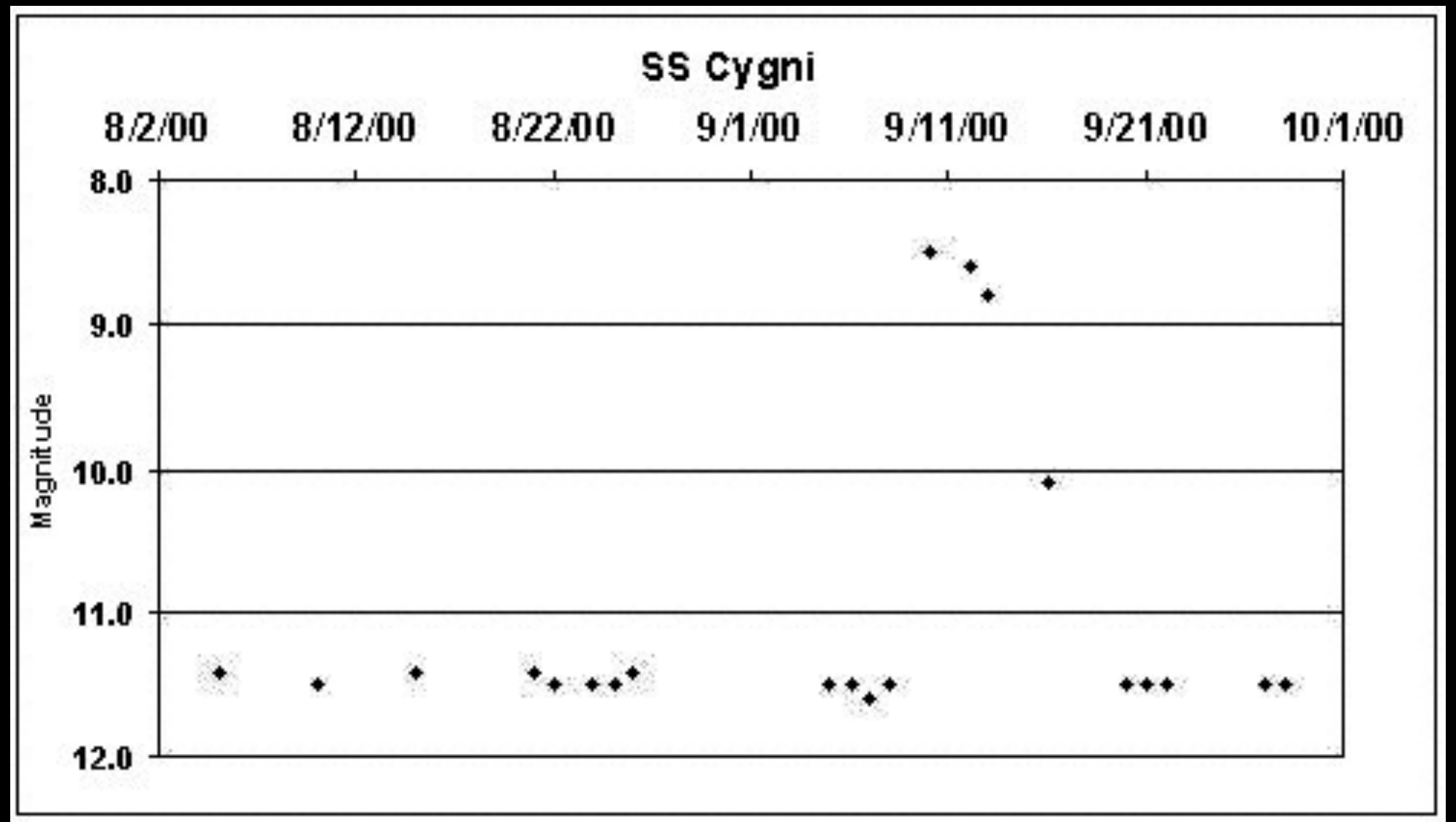


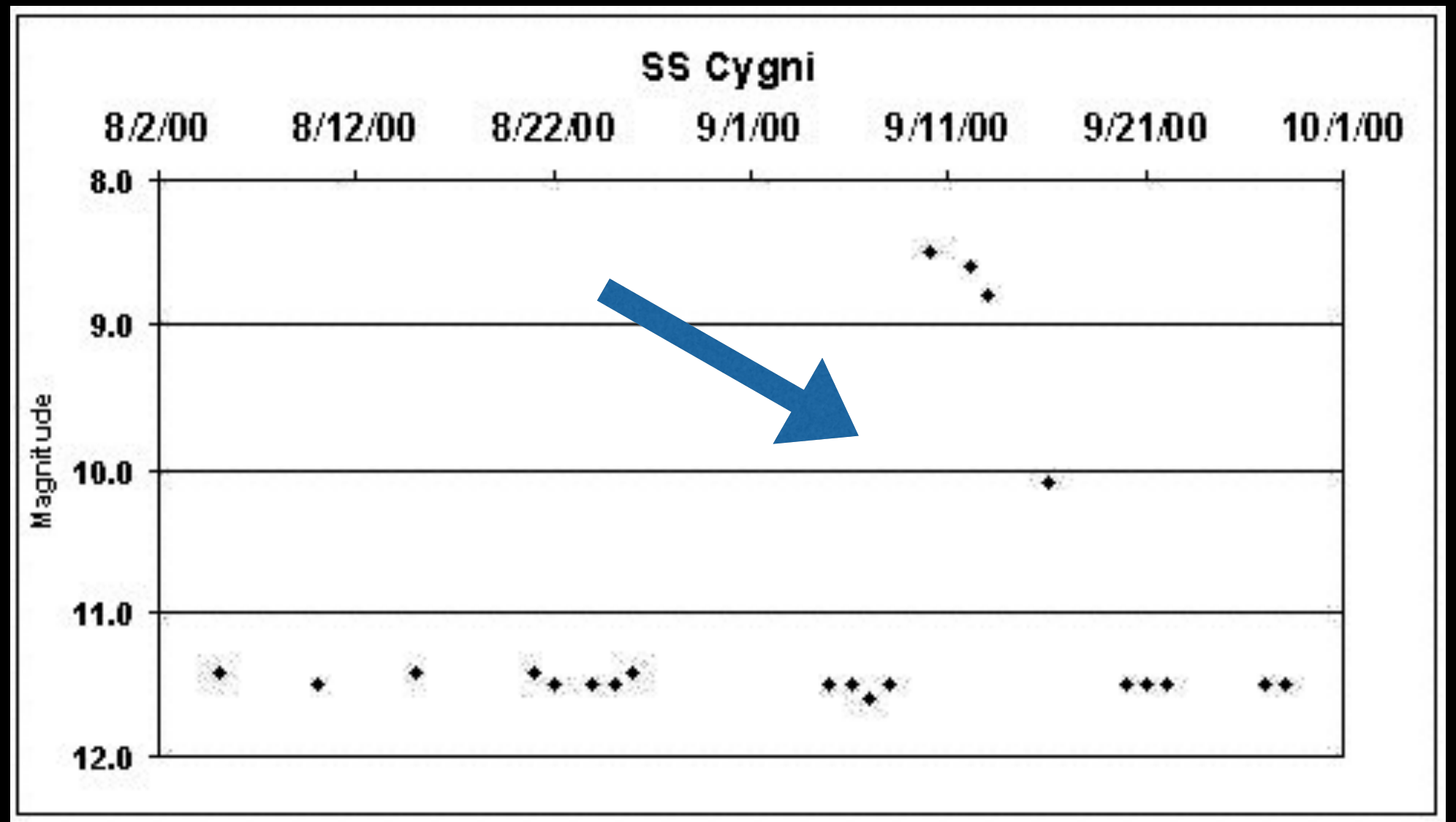
**Periodicity**



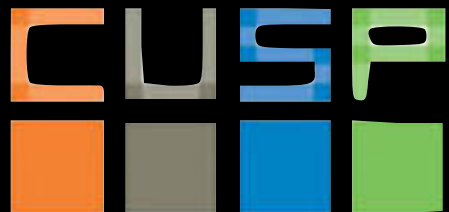
*don't forget to vote!*

VIII: Topics in Time series



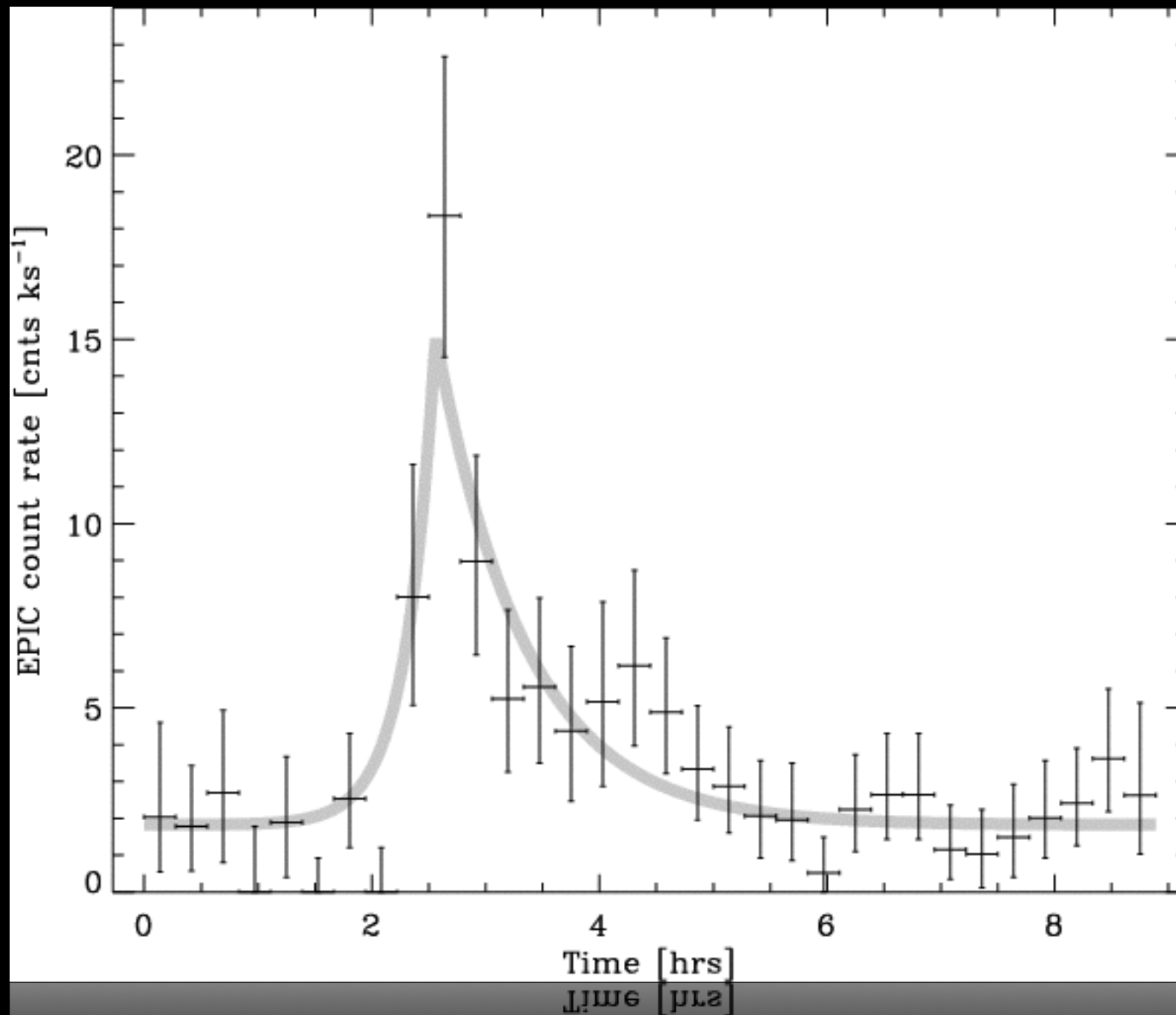


event detection

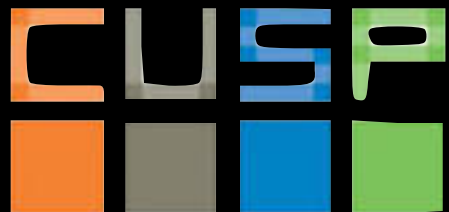


*don't forget to vote!*

VIII: Topics in Time series

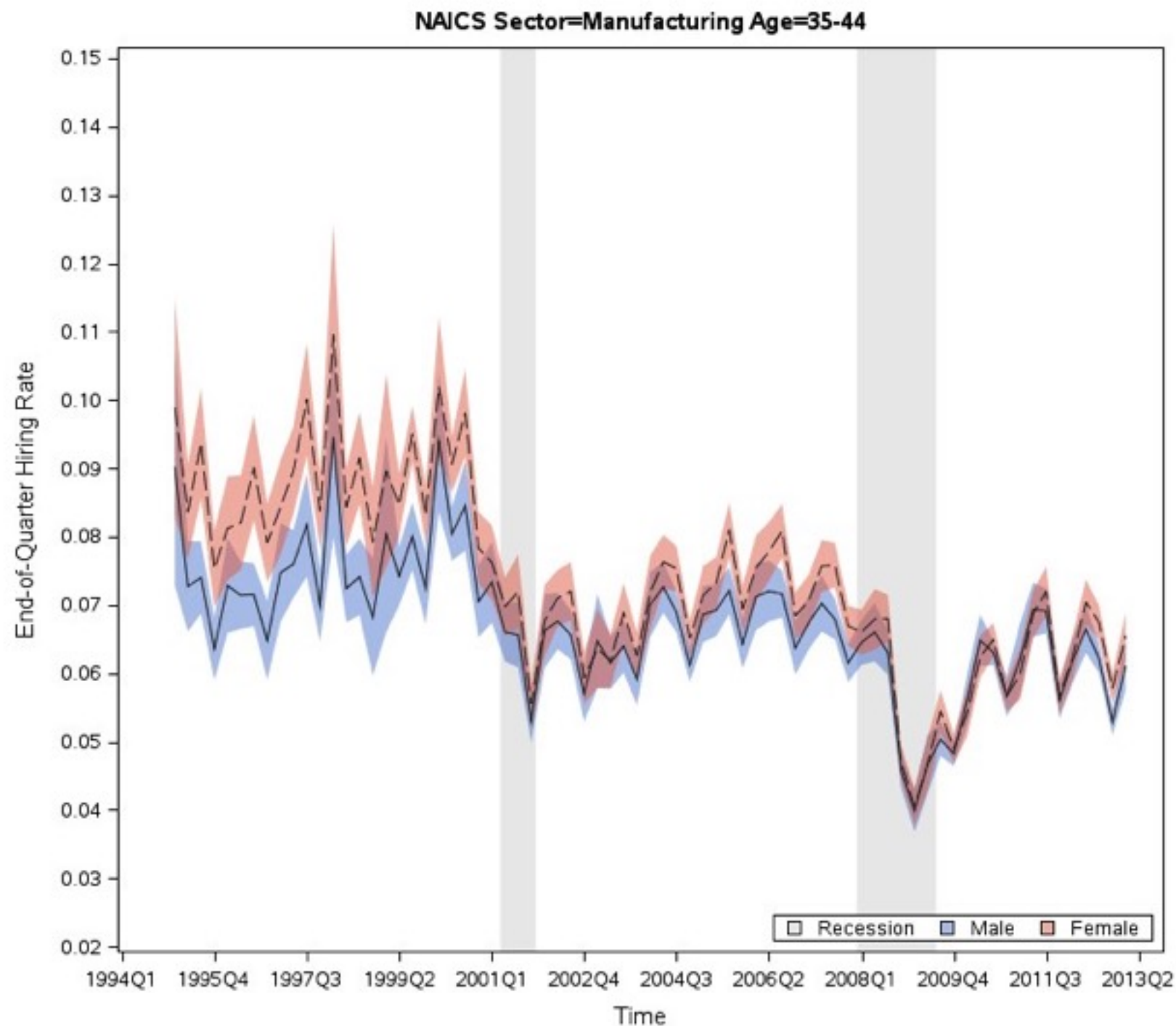


event detection



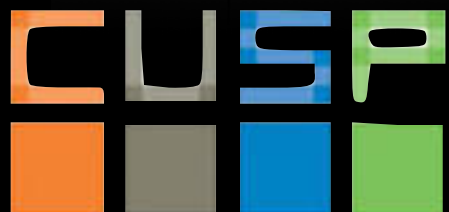
*don't forget to vote!*

VIII: Topics in Time series



event detection

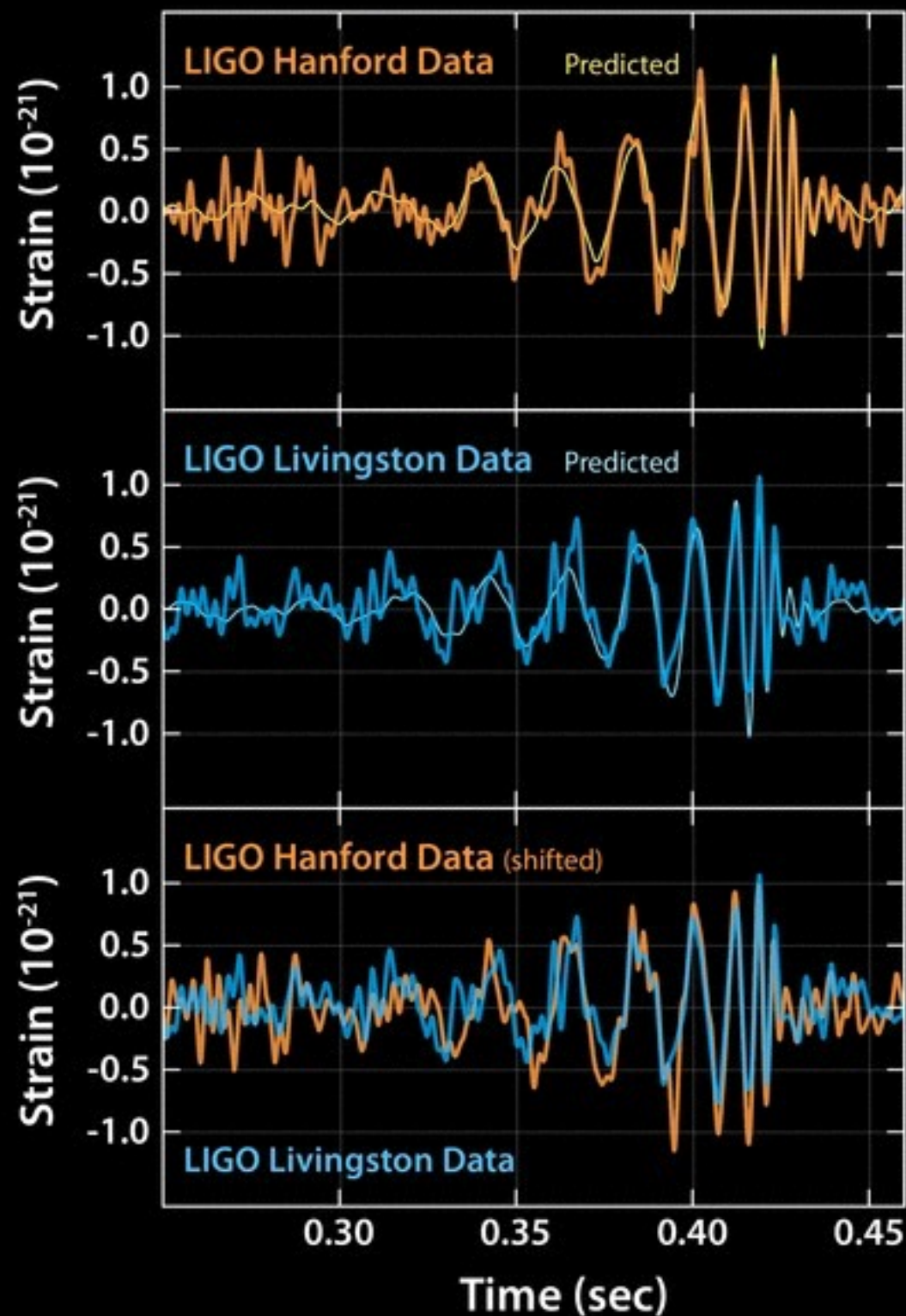
LEHD data (Prof. Julia Lane)



*don't forget to vote!*

VIII: Topics in Time series

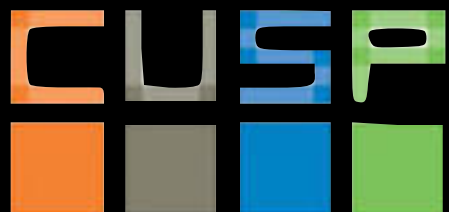




**event detection**

LIGO gravitational wave detection

Abbott et al. Physical Review Letters 116, 061102 (2016)

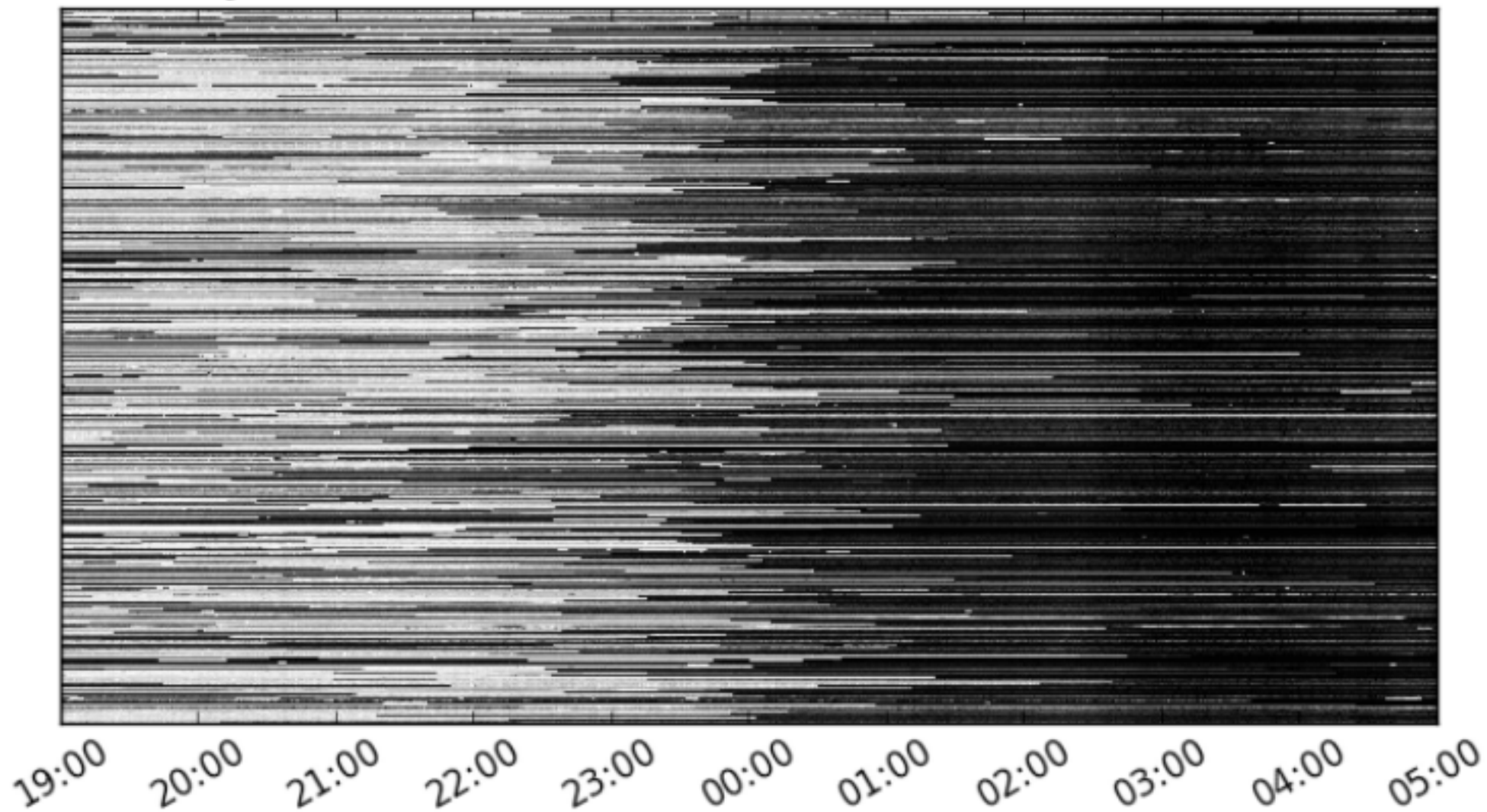


**don't forget to vote!**

VIII: Topics in Time series

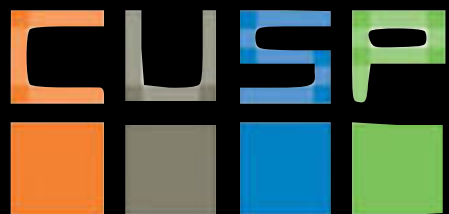


Monday



<http://www.sciencedirect.com/science/article/pii/S0306437915001167>

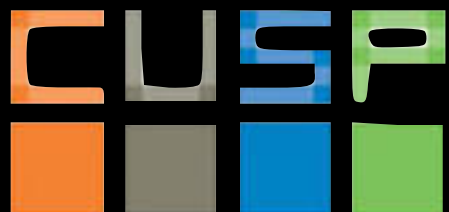
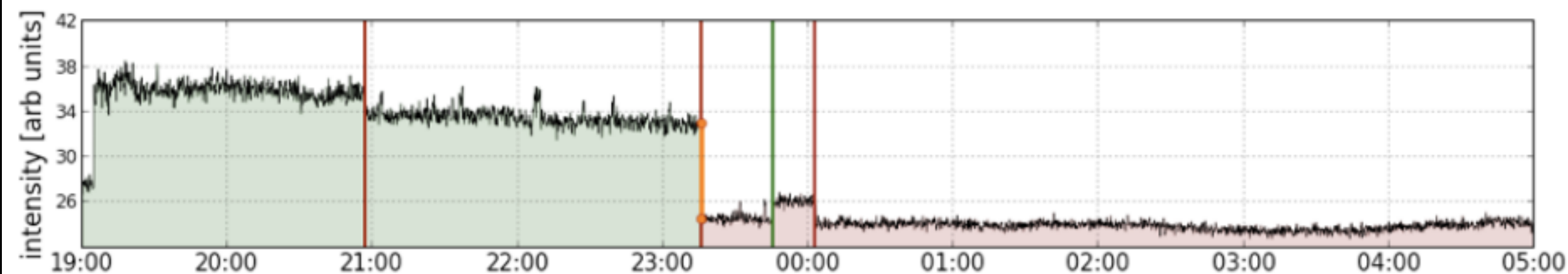
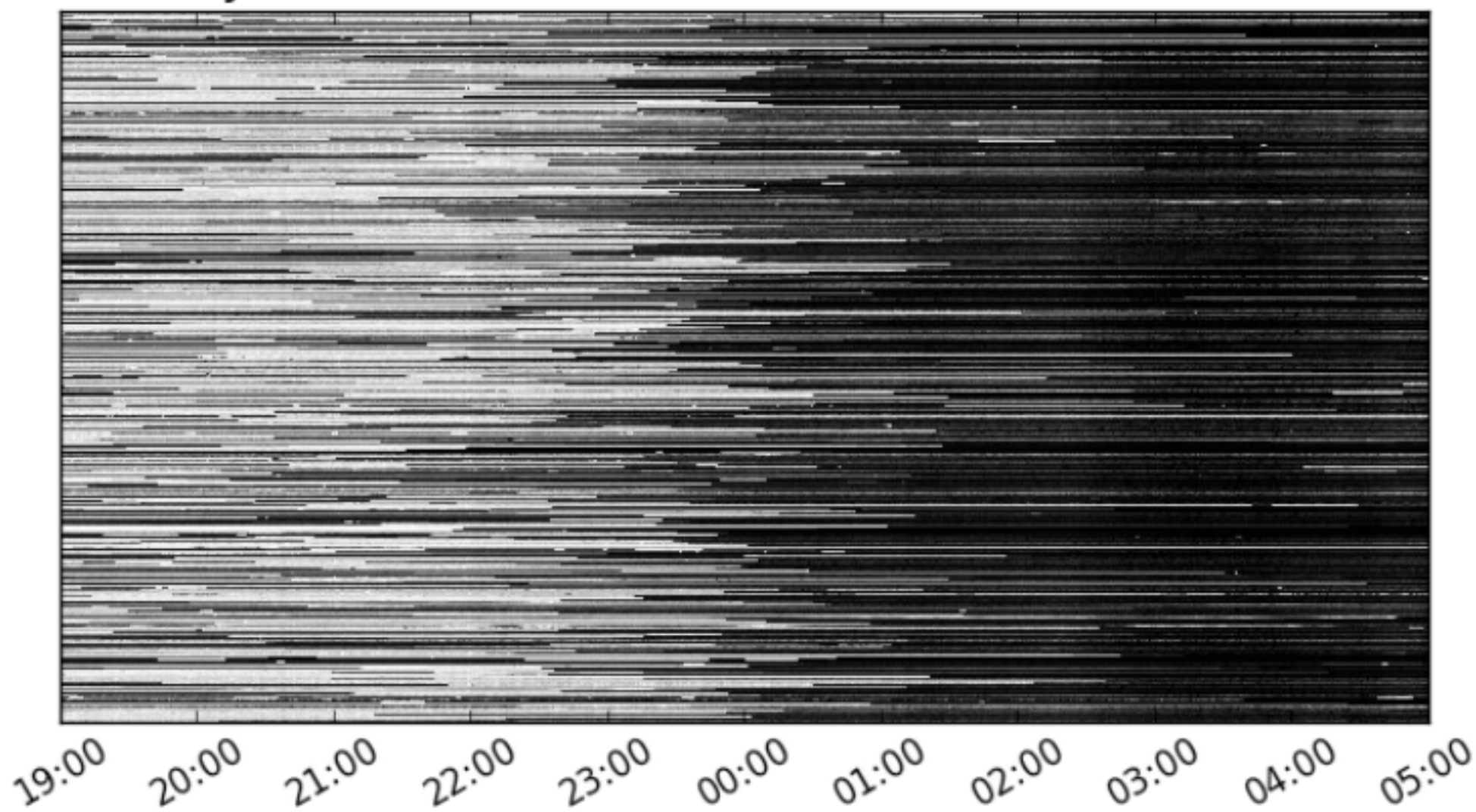
CUSP-UO



*don't forget to vote!*

VIII: Topics in Time series

Monday

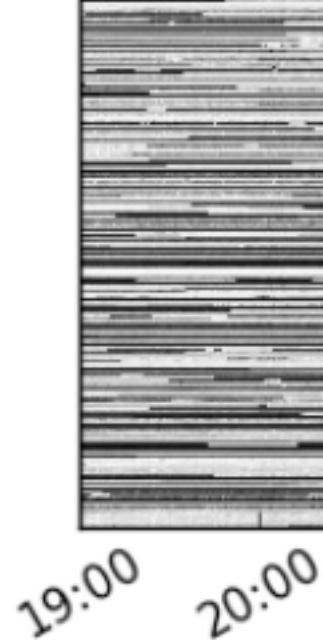


*don't forget to vote!*

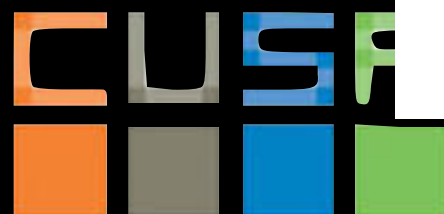
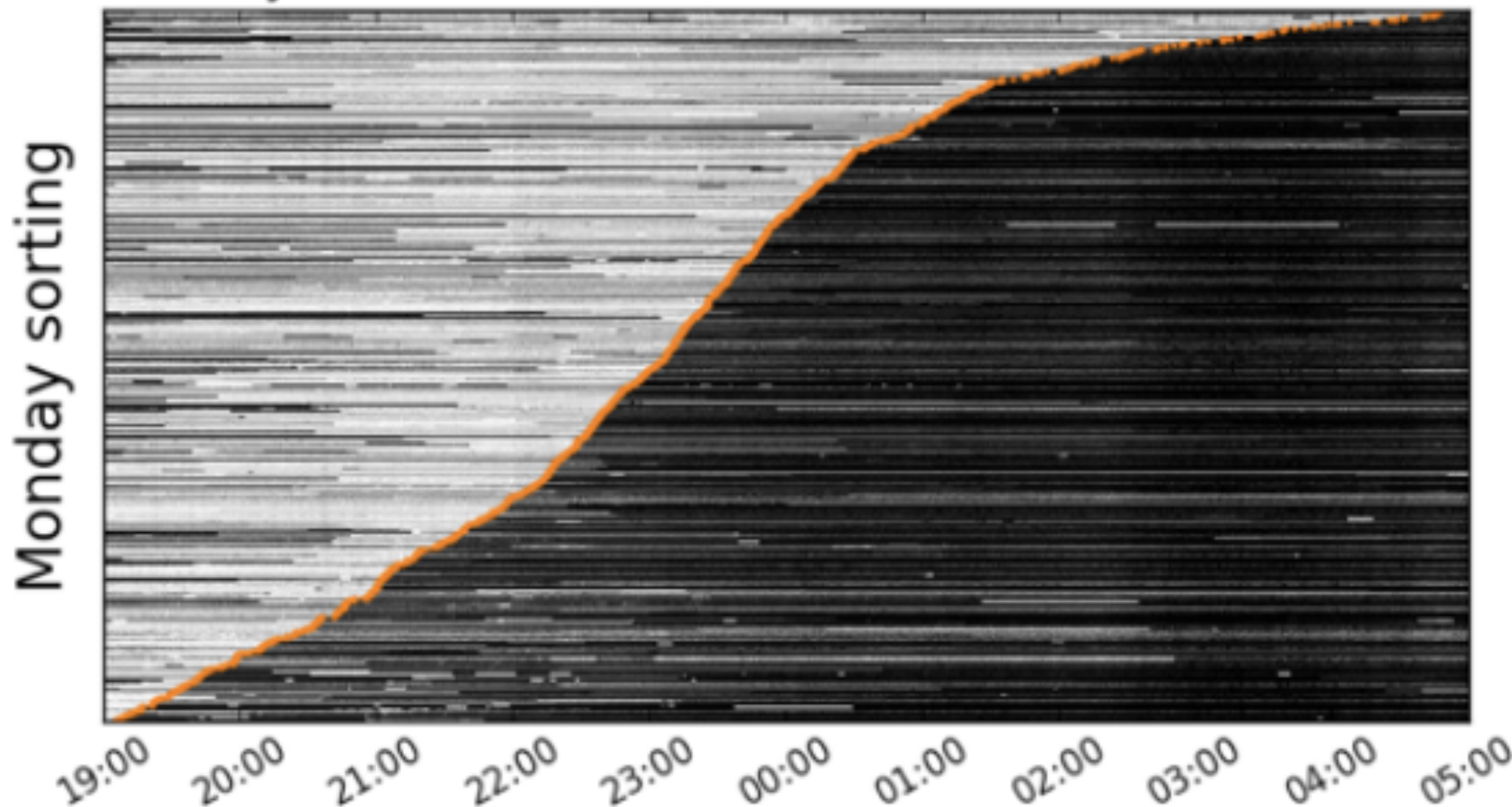
VIII: Topics in Time series



Monday

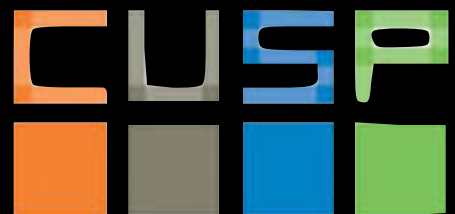
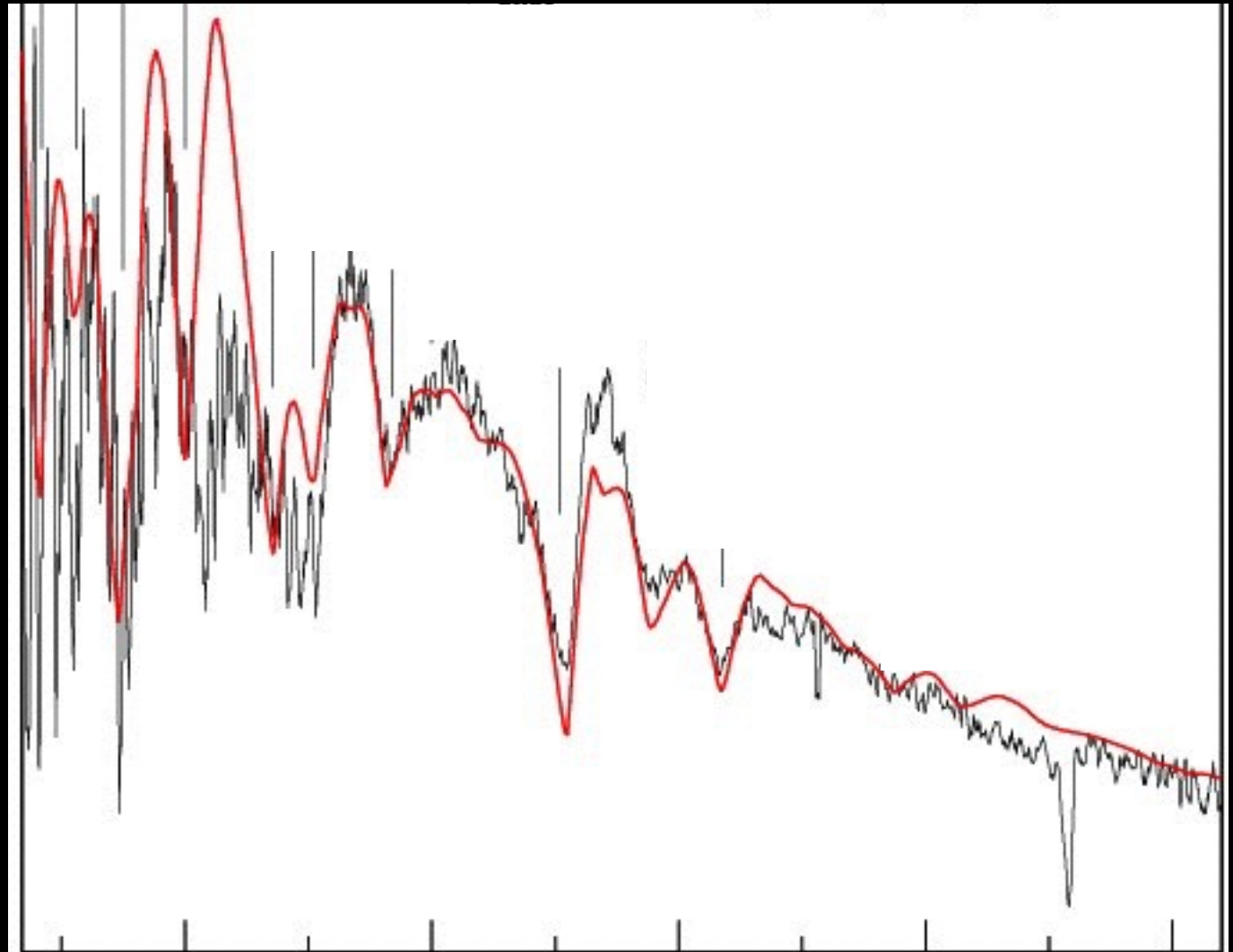


Monday



*don't forget to vote!*

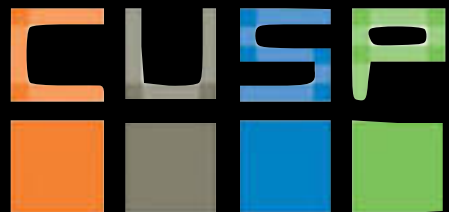
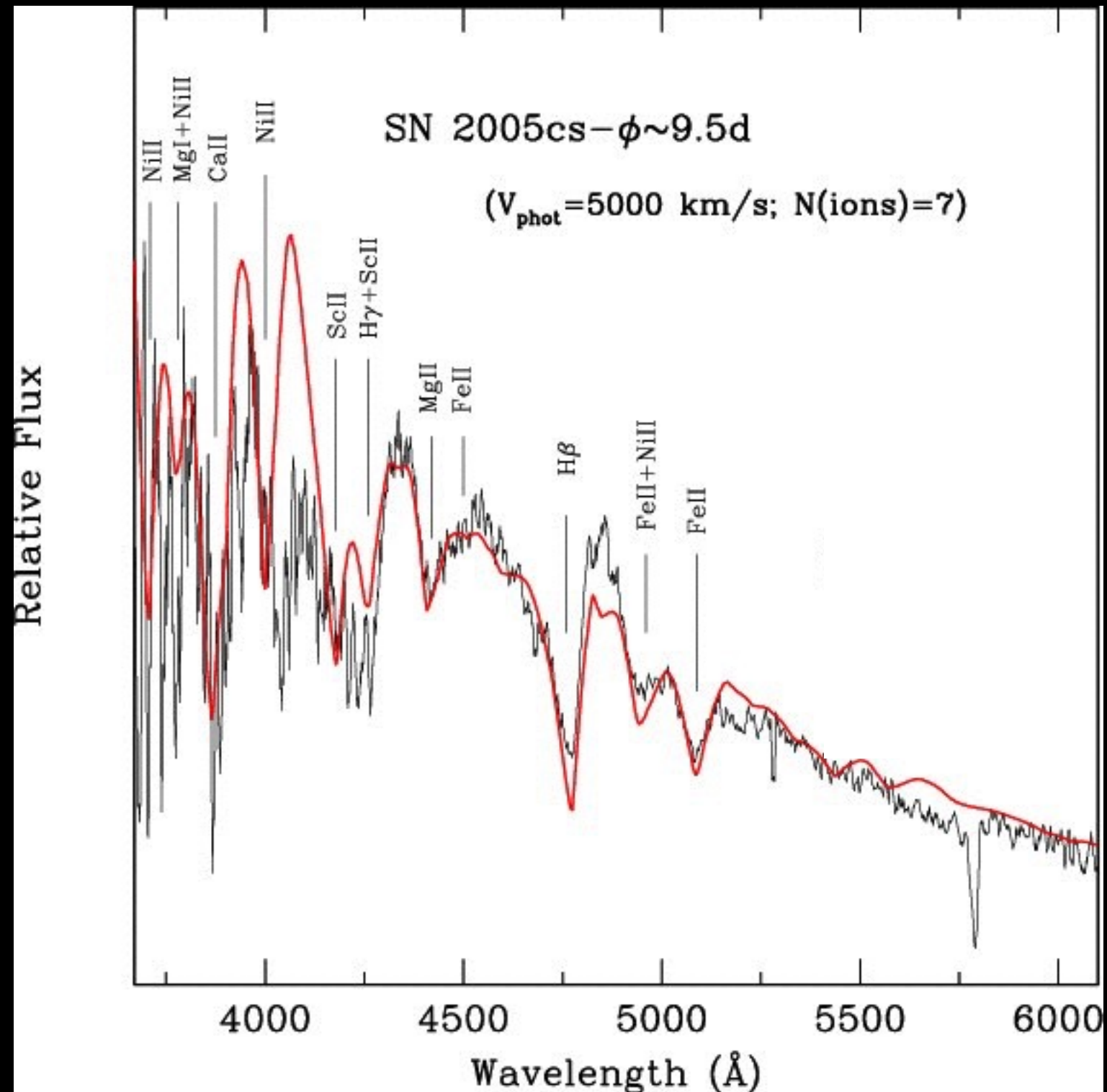
VIII: Topics in Time series



*don't forget to vote!*

VIII: Topics in Time series

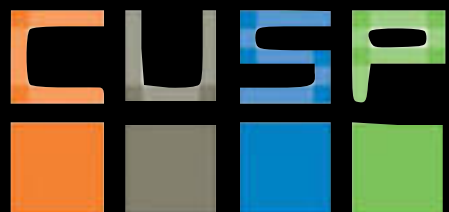
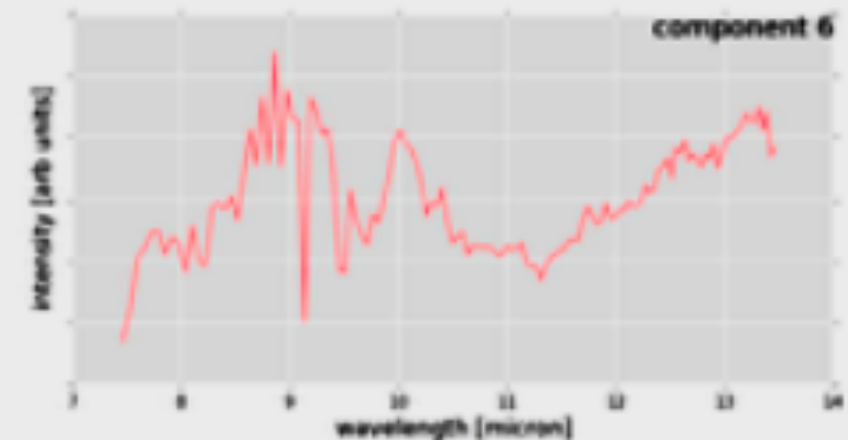
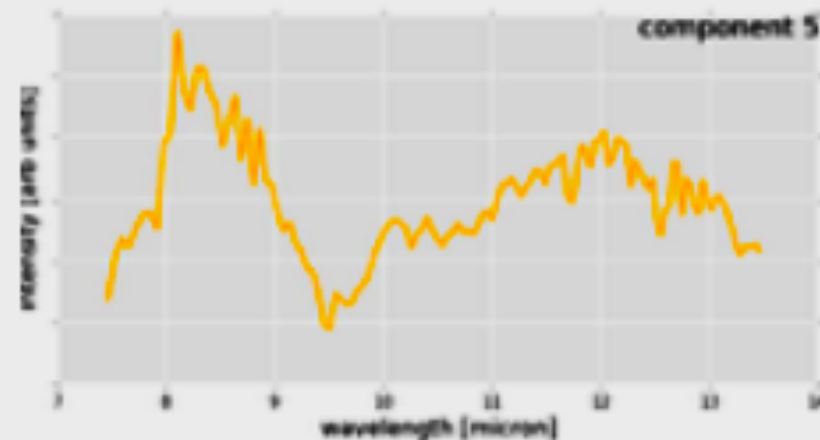
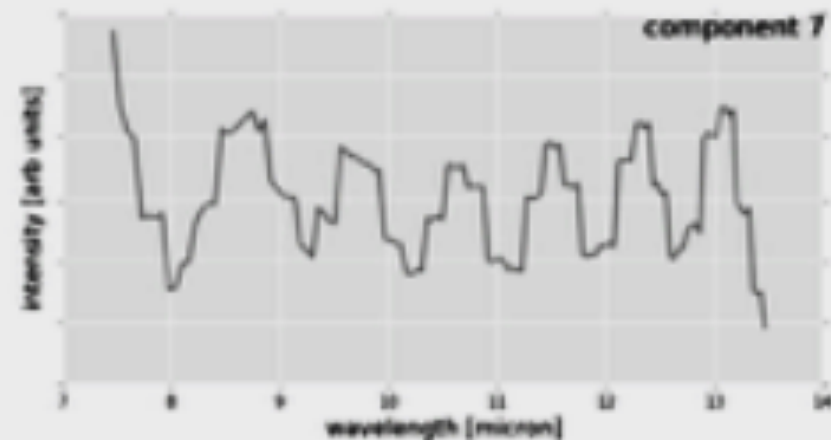
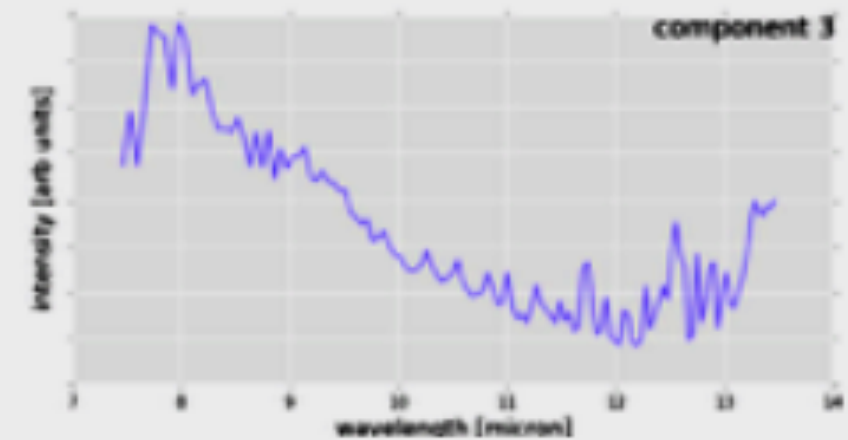
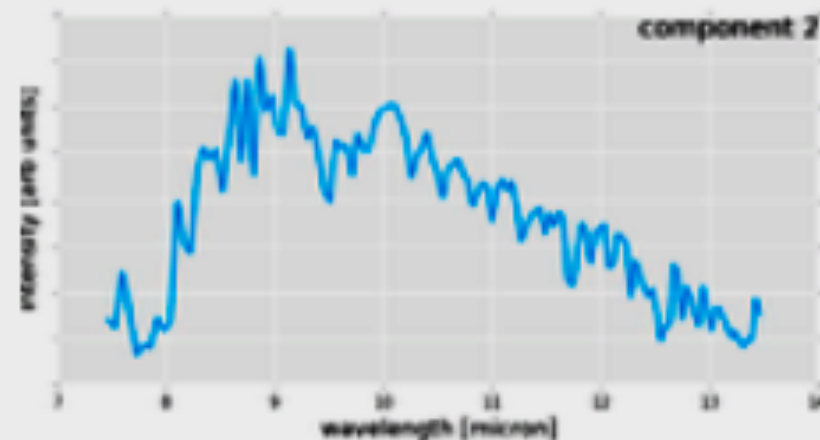
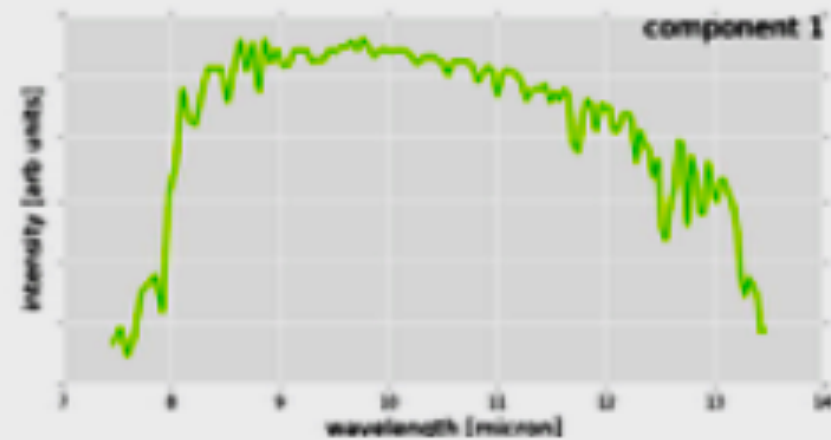
they do not have to be *TIME* series!



*don't forget to vote!*

VIII: Topics in Time series

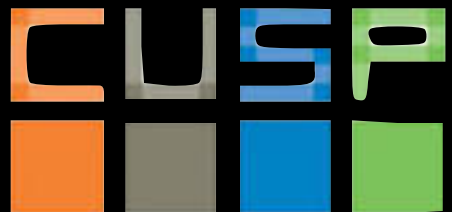
# CUSP-UO spectra of urban lights for light technology assessment



*don't forget to vote!*

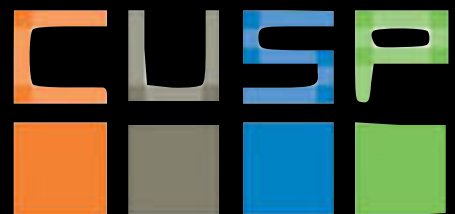
VIII: Topics in Time series

- event detection



*don't forget to vote!*

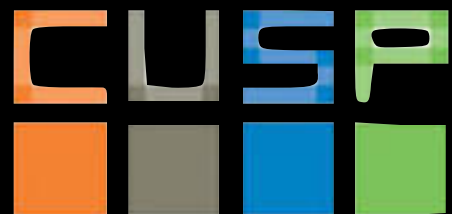
- event detection
- identification of trends



*don't forget to vote!*

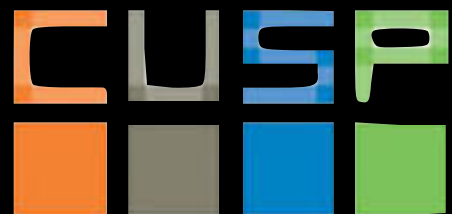


- event detection
- identification of trends
- periodicity detection



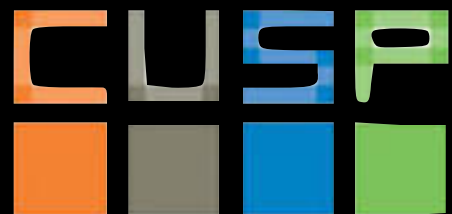
*don't forget to vote!*

- event detection
- identification of trends
- periodicity detection
- prediction



*don't forget to vote!*

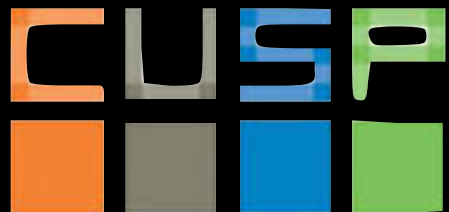
- event detection
- identification of trends
- periodicity detection
- prediction
- classification (clustering)



*don't forget to vote!*

- event detection

Thresholding



*don't forget to vote!*

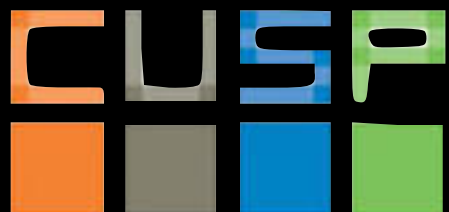
- event detection

## Thresholding



- take the mean (possibly a local mean)
- take the standard deviation (possibly a local stdev)
- find points that deviate from the mean by more than  $N$  standard deviation

[https://github.com/fedhere/Ulnotebooks/blob/master/  
FDNYdeaths.ipynb](https://github.com/fedhere/Ulnotebooks/blob/master/FDNYdeaths.ipynb)



*don't forget to vote!*

- event detection
- identification of trends

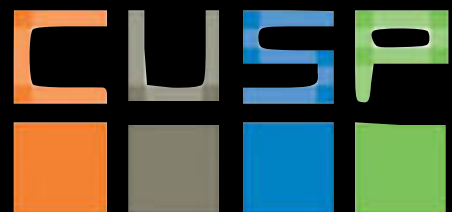
**Stationary data**

**Smoothing (Rolling mean)**

**ADFuller test for unit root (for non-stationarity)**



<https://github.com/fedhere/Ulnotebooks/blob/master/timeseries/stationarity.ipynb>



*don't forget to vote!*

VIII: Topics in Time series

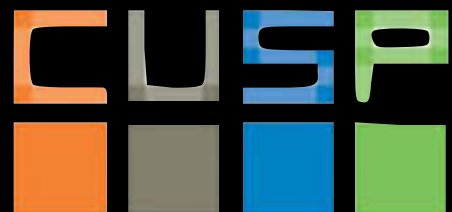
- event detection
- identification of trends
- periodicity detection

ARMA/ARIMA



<http://www.statsref.com/HTML/index.html?arima.html>

<http://www.econ.ohio-state.edu/dejong/note2.pdf>



*don't forget to vote!*

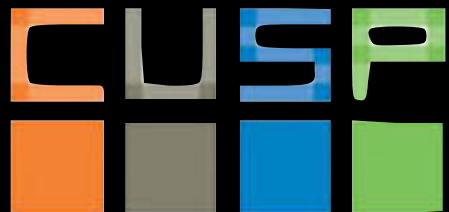
VIII: Topics in Time series

ARIMA

Moving Average Model

$$x(t) = \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t + \mu$$

 jupyter



*don't forget to vote!*

VIII: Topics in Time series



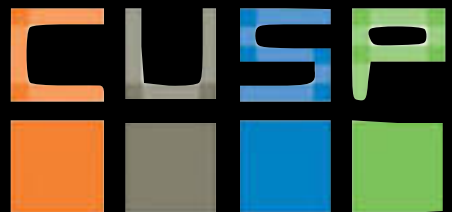
## ARIMA

Autoregression

$$x(t) = a_1 x(t-1) + \epsilon_t$$

Moving Average Model

$$x(t) = \sum_{i=1}^q \theta_i \epsilon_{t-i} + \epsilon_t + \mu$$



*don't forget to vote!*

VIII: Topics in Time series

## ARIMA

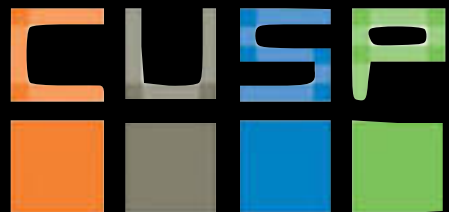
Autoregression

$$x(t) = a_1 x(t-1) + \epsilon_t$$

$$x(t) = a_1 x(t-1) + a_2 x(t-2) + \dots + a_n x(t-n) + \epsilon_t$$

Moving Average Model

$$x(t) = \sum_{i=1}^q \theta_i \epsilon_{t-i} + \epsilon_t + \mu$$



*don't forget to vote!*

VIII: Topics in Time series

# Integration

$$x'(t) = x(t) - x(t-i)$$

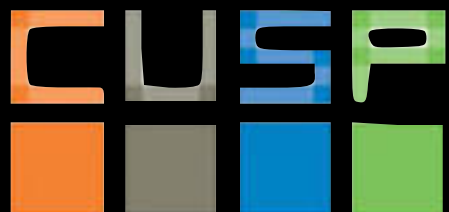
## ARIMA

### Autoregression

$$x(t) = \sum_{i=1}^p a_i x_{t-i} + \varepsilon_t$$

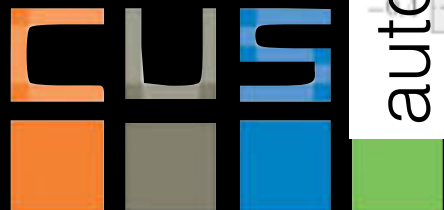
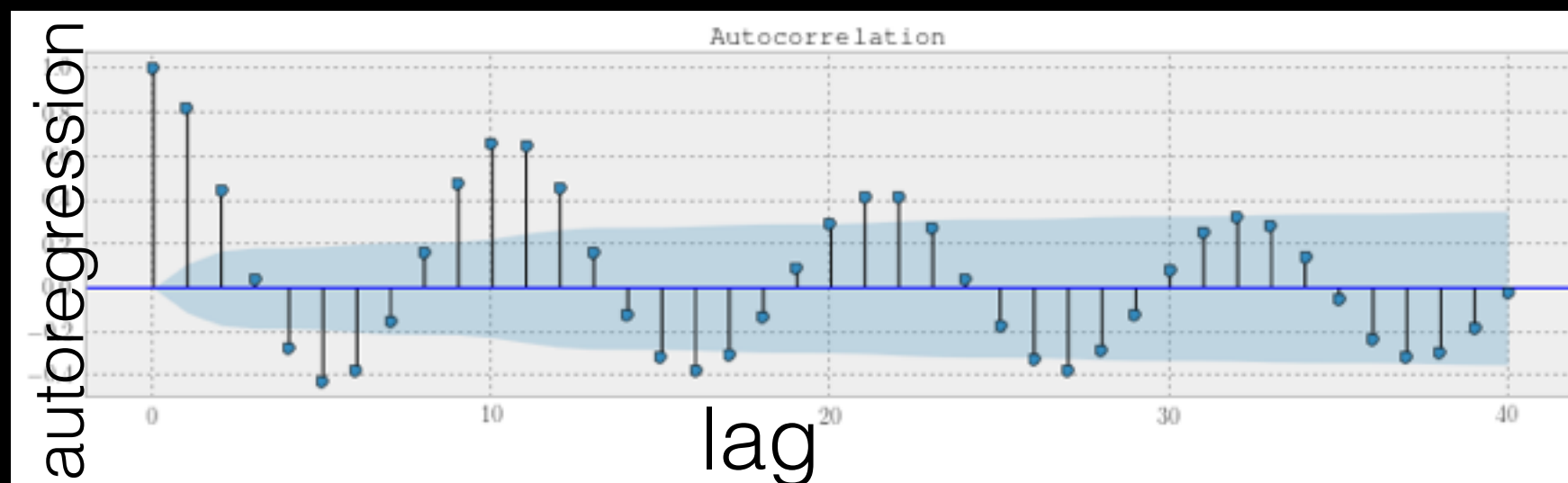
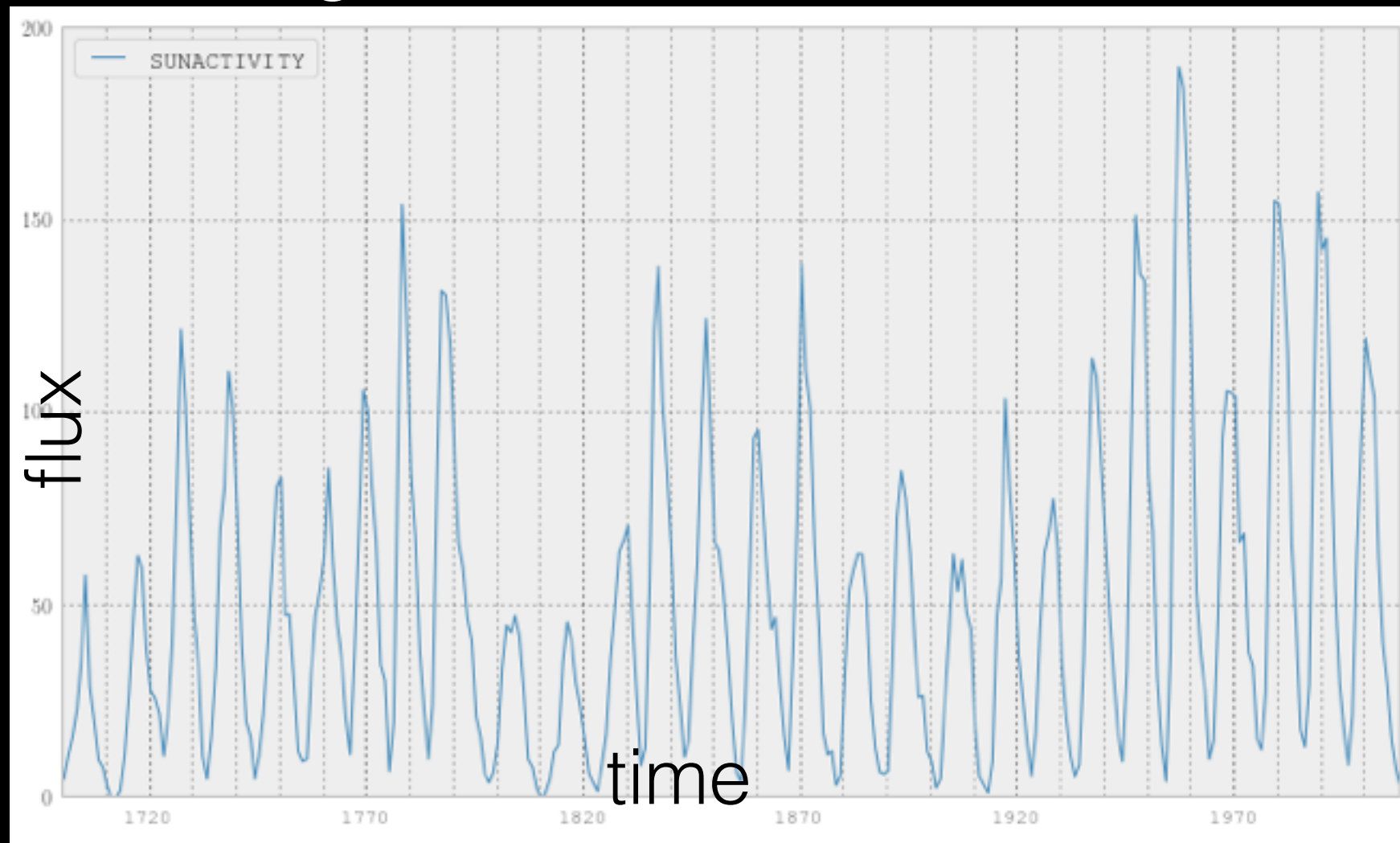
### Moving Average Model

$$x(t) = \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t + \mu$$



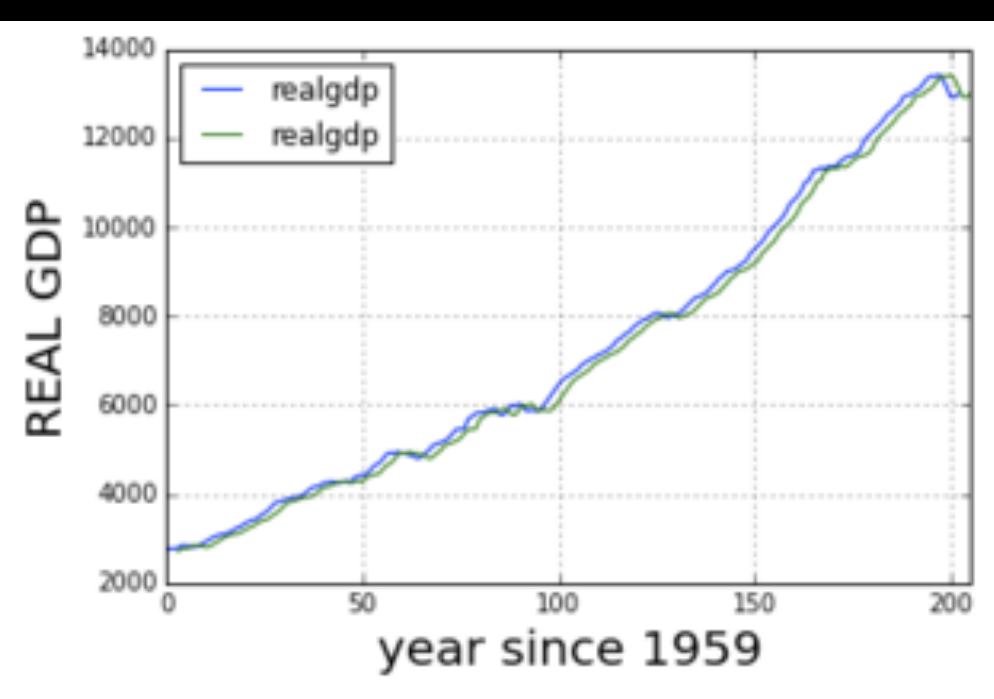
*don't forget to vote!*

[http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/tsa\\_arma\\_0.html](http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/tsa_arma_0.html)



***don't forget to vote!***

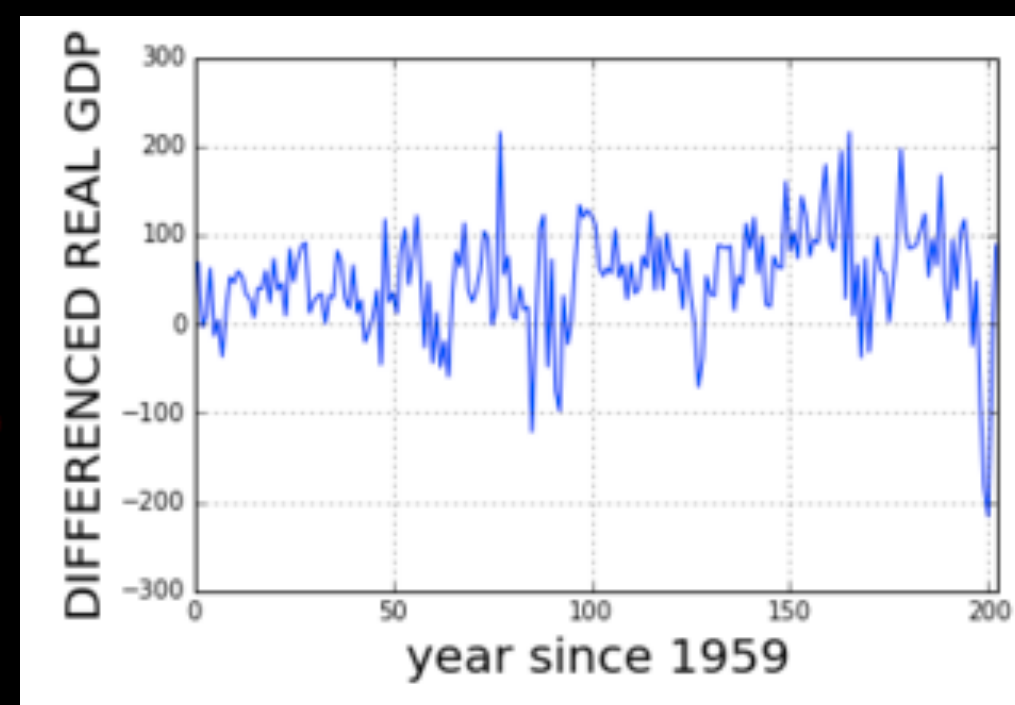
VIII: Topics in Time series



Integration

$$x'(t) = x(t) - x(t-i)$$

ARIMA



Autoregression

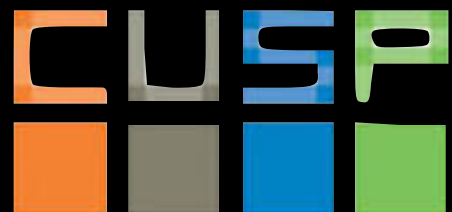
$$x(t) = \sum_{i=1}^p a_i x_{t-i} + \varepsilon_t$$

Moving Average Model

$$x(t) = \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t + \mu$$

 jupyter

[https://github.com/fedhere/Ulnotebooks/blob/master/ARMA\\_microdata.ipynb](https://github.com/fedhere/Ulnotebooks/blob/master/ARMA_microdata.ipynb)

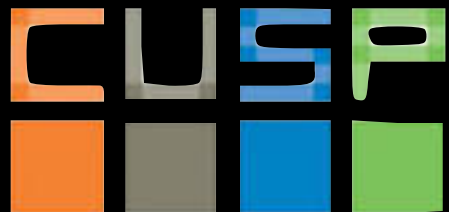


*don't forget to vote!*

VIII: Topics in Time series

## Fourier


$$F(\omega) = \frac{1}{2\pi} \int f(t) e^{-i\omega t} dt$$



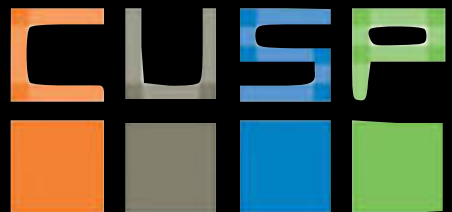
*don't forget to vote!*

VIII: Topics in Time series

## Fourier

$$F(\omega) = \frac{1}{2\pi} \int f(t) e^{-i\omega t} dt$$


takes a function in time domain



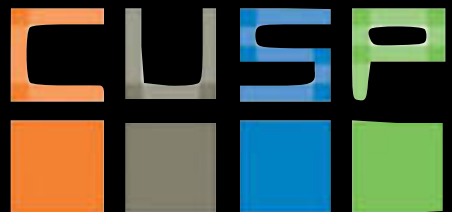

*don't forget to vote!*

VIII: Topics in Time series

## Fourier

$$F(\omega) = \frac{1}{2\pi} \int f(t) e^{-i\omega t} dt$$

takes a function in time domain  
to a function in frequency domain



*don't forget to vote!*

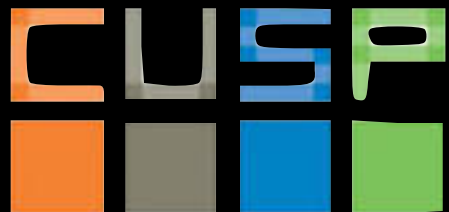
VIII: Topics in Time series



## Fourier

$$F(\omega) = \frac{1}{2\pi} \int f(t) e^{-i\omega t} dt$$

takes a function in space domain  
to a function in spatial frequency domain



*don't forget to vote!*

VIII: Topics in Time series

## Fourier

$$F(\omega) = \frac{1}{2\pi} \int f(t) e^{-i\omega t} dt$$

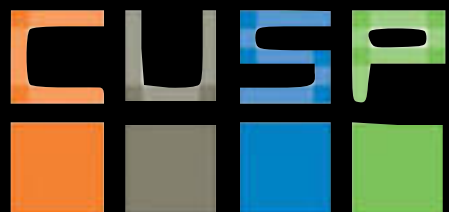
takes a function in space domain

$f(t)$  is measured in seconds

to a function in spatial frequency domain

$f(t)$  is measured in 1/seconds

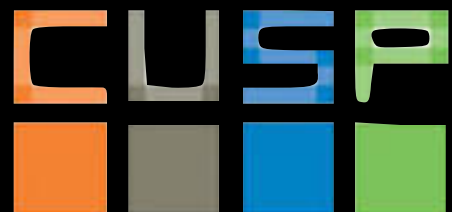
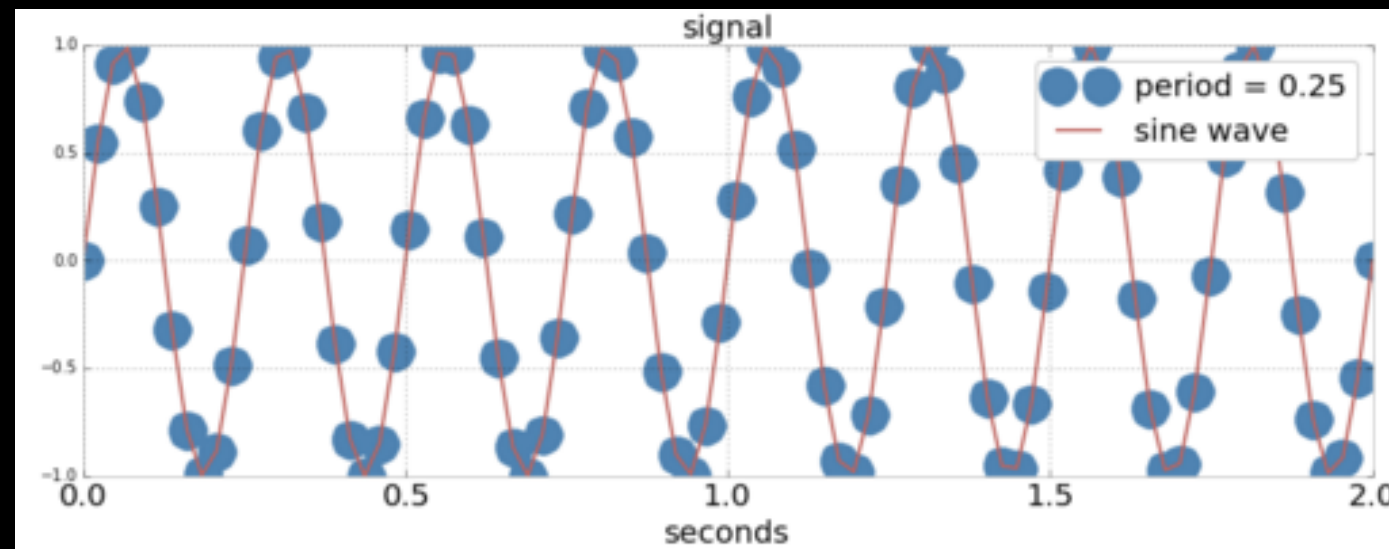
or Hz



*don't forget to vote!*

VIII: Topics in Time series

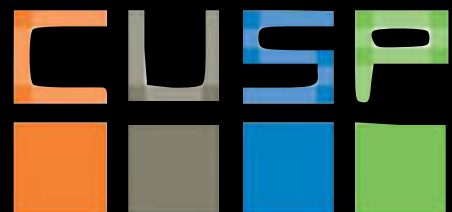
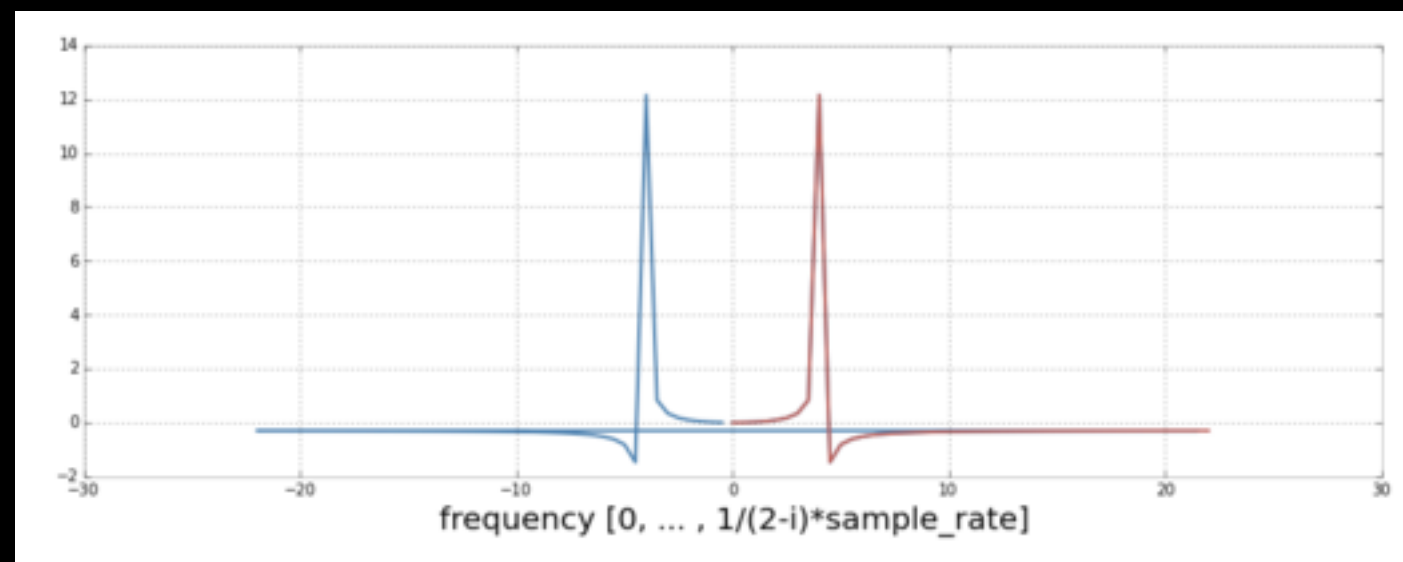
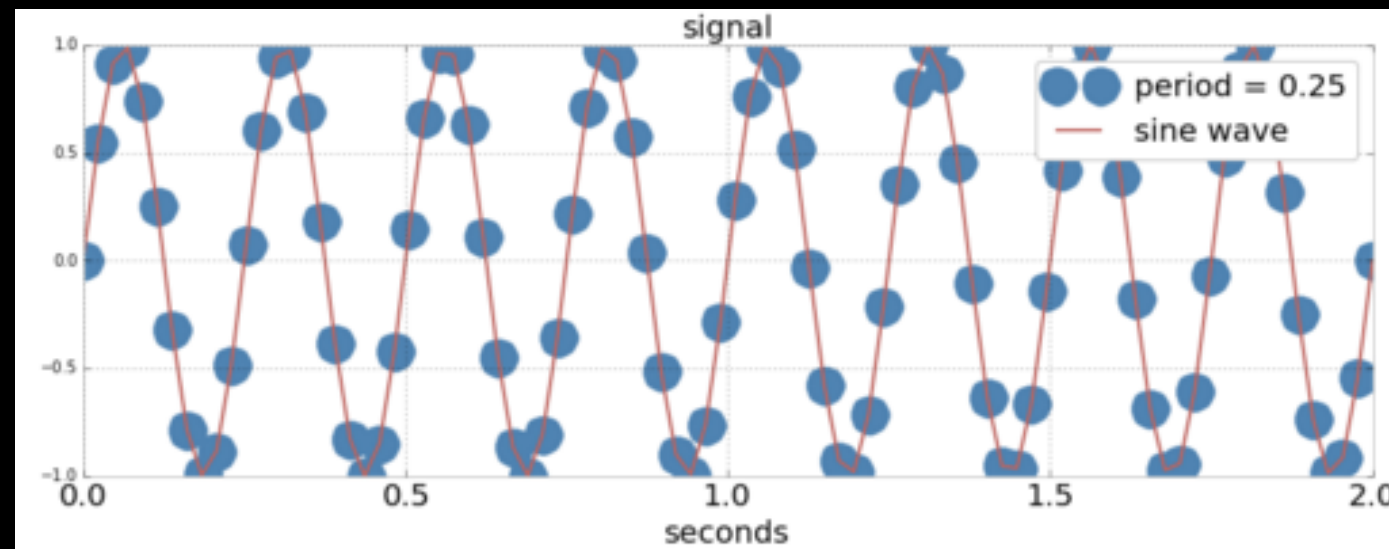
# Fourier



*don't forget to vote!*

VIII: Topics in Time series

# Fourier

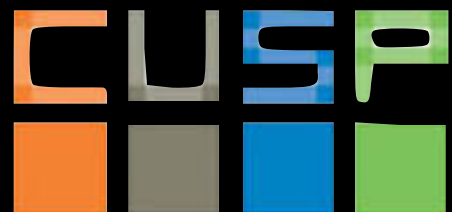
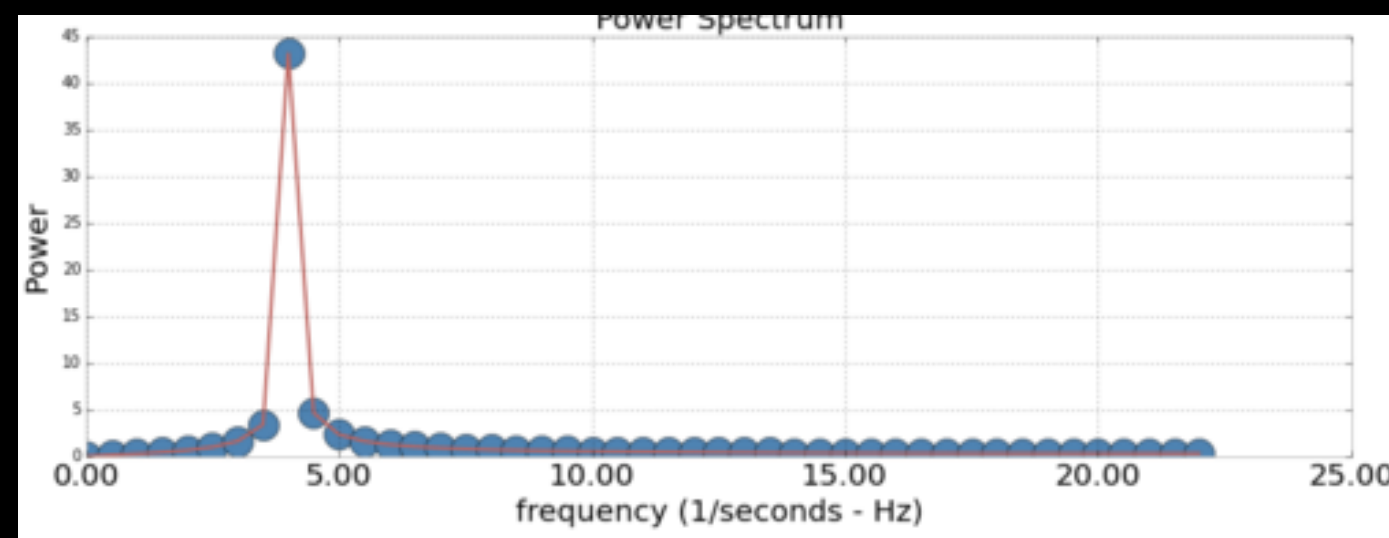
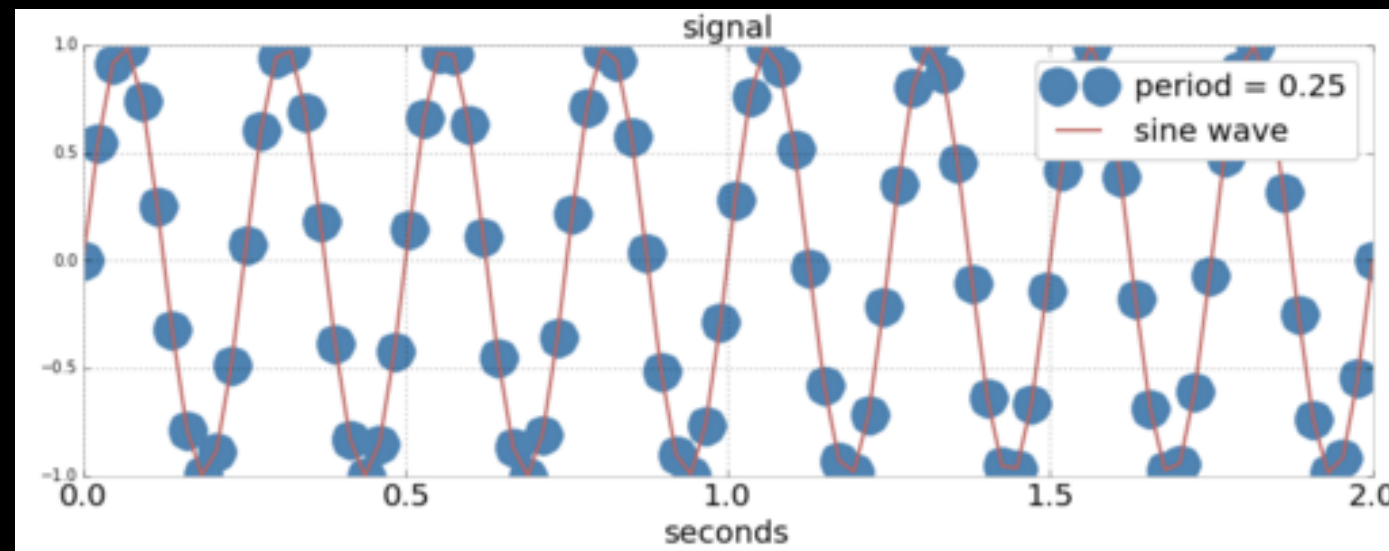


*don't forget to vote!*

VIII: Topics in Time series

The absolute value of the square of the Fourier transform  
this is called a Power spectrum.  
High value of the power spectrum indicate periodicity at the  
corresponding frequency

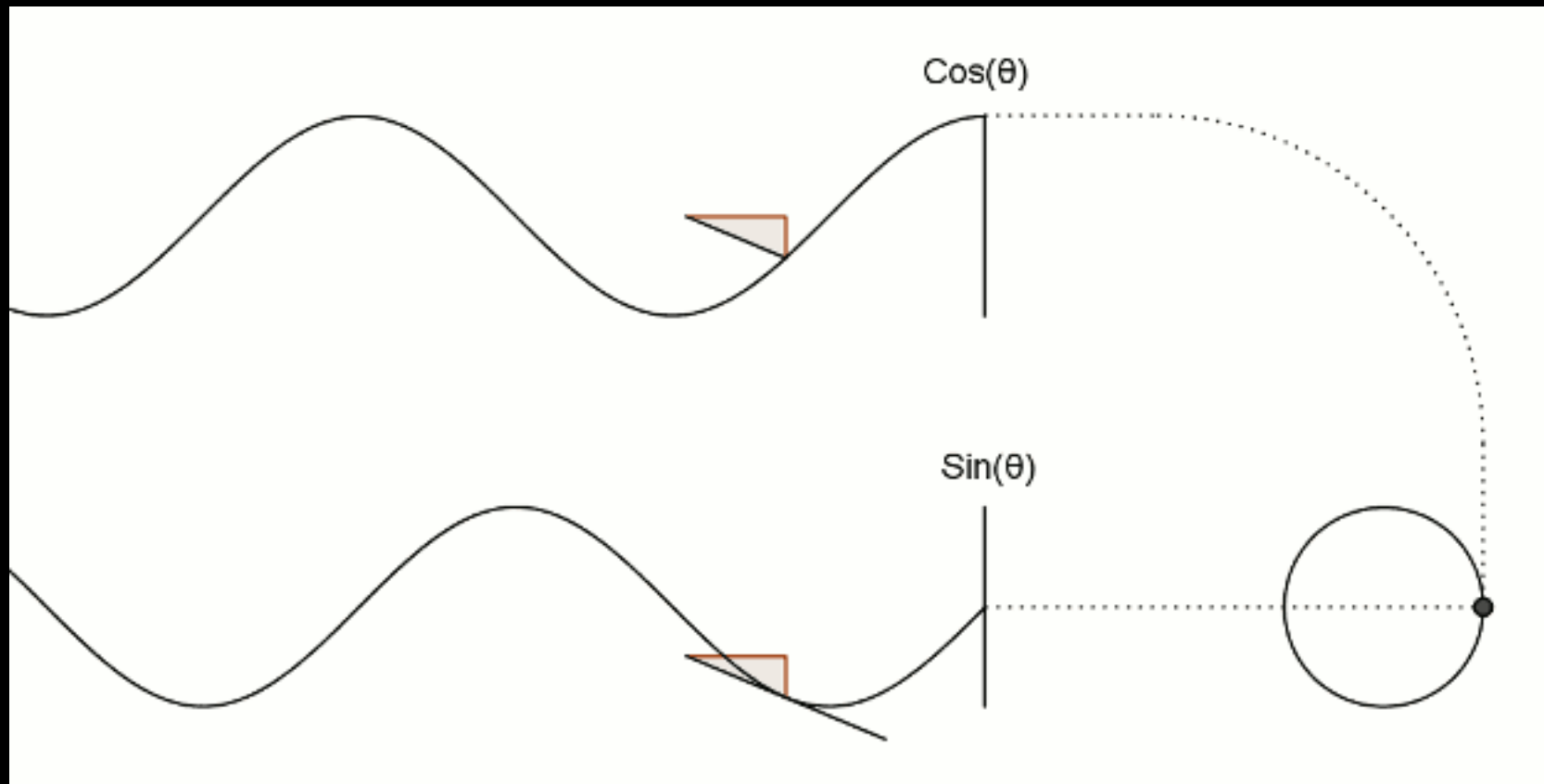
## Fourier



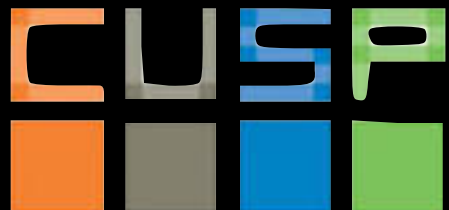
*don't forget to vote!*

VIII: Topics in Time series

# Cosine and Sine... just in case



<http://www.businessinsider.com/7-gifs-trigonometry-sine-cosine-2013-5>



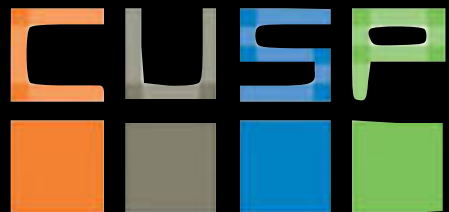
*don't forget to vote!*

VIII: Topics in Time series

Fourier



[https://github.com/fedhere/Ulnotebooks/blob/master/  
fourier.ipynb](https://github.com/fedhere/Ulnotebooks/blob/master/fourier.ipynb)



*don't forget to vote!*

VIII: Topics in Time series

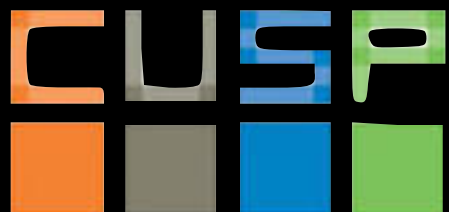
## Homework:

Technical reading on SM time analysis tools. Get through ARMA

[http://conference.scipy.org/proceedings/  
scipy2011/pdfs/statsmodels.pdf](http://conference.scipy.org/proceedings/scipy2011/pdfs/statsmodels.pdf)

Reading: an excellent analysis of time series  
by Jake Vander Plas  
(UW e-science center)

[https://jakevdp.github.io/blog/2014/06/10/is-  
seattle-really-seeing-an-uptick-in-cycling/](https://jakevdp.github.io/blog/2014/06/10/is-seattle-really-seeing-an-uptick-in-cycling/)



*don't forget to vote!*

VIII: Topics in Time series



# Homework:

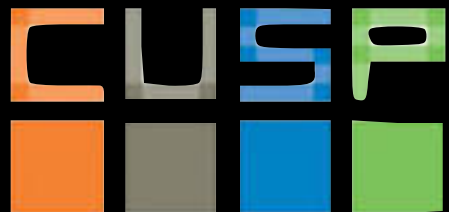
## Data:

MTA subway fares. It is a complete dataset of rides logged by card swipes for 600 Manhattan stations.

It contains 23 different subway card types (e.g. monthly pass, daily pass, Act for Disability pass... i will give you this as a list)

Each time series (per station, per ticket type) contains the number of swipes per week for 194 weeks from 05/21/2010 to 02/21/2014.

it is given to you as a python data cube.  
you can load it as `np.load("MTA_Fare.npy")` and you will end up with a python numpy array of shape (600,23,194)



*don't forget to vote!*

VIII: Topics in Time series

# Homework:

## Goal 1:

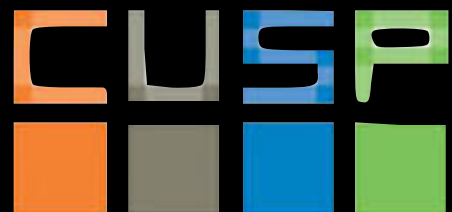
Event detection: Identify the most prominent event. There is a very significant drop ( $>3$ -sigma) in *all* time series. Identify it and figure out what it is due to.

## Goal 2:

Some of the time series are stationary, some show a downward trend: Identify the time series with the most prominent downward trend.

## Goal 3:

Several stations show a prominent annual periodicity. Identify the 5 stations that show the most prominent periodic trend on an annual period. Figure out what the increase in rides is due to.



*don't forget to vote!*

VIII: Topics in Time series

## Homework Hints:

### Goal 1:

Some of the time series are stationary, some show a downward trend: Identify the time series with the most prominent downward trend.

work with all time series individually. you can use the rolling mean to find trends: compare rolling mean near beginning and end of time series.

### Goal 2:

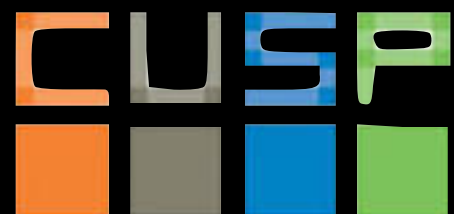
Event detection: Identify the most prominent event. There is a very significant drop ( $>3$ -sigma) in *all* time series.

Identify it and figure out what it is due to.

Since I am telling you the event is in all time series you can work with averages: for example average over all rise types per station. Since i am telling you it is a highly significant event you can find it by thresholding

### Goal 3:

Several stations show a prominent annual periodicity. Identify the 5 stations that show the most prominent periodic trend on an annual period. Figure out what the increase in rides is due to.



Work in Fourier space: find the series that have the most

prominent peak at  $\sim 1$  year frequency  
*don't forget to vote!*

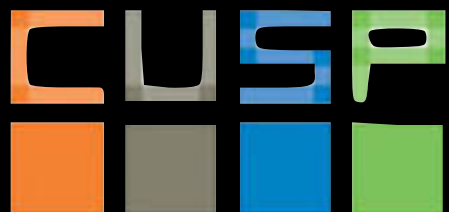
## Homework ExtraCredit:

### Cluster:

Cluster the time series: you can use KMeans for example to identify common trends. or PCA. Since this is extra credit I will leave it entirely to you to figure out the details.

for KMeans for e.g.:

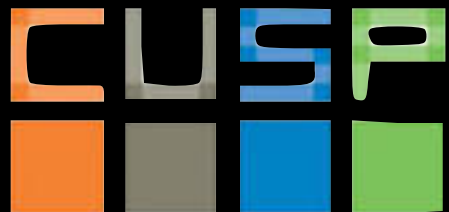
```
#i am flattening the first 2 dimensions of the cube to cluster all
light curves for all stations and all types
tots = data.transpose(2,0,1).reshape(data.shape[2],
data.shape[1]*data.shape[0]).T
#removing empty light curves
tots = tots[tots.std(1)>0]
#ith Kmeans you have to choose the number of clusters ahead km
= KMeans(n_clusters=10)
#and standardize the lightcurves before clustering
vals = ((tots.T - tots.mean(1))/tots.std(1)).T
km.fit(vals)
```



*don't forget to vote!*

## Key points:

- Time series analysis may be done for a number of purposes: classification, prediction, event detection, period finding
- smoothing, binning, detrending (difference, regression)
- prediction tools: autoregression, ARMA, ARIMA
- period finding (Fourier analysis)



*don't forget to vote!*

## References:

Statistical Analysis Handbook

<http://www.statsref.com/HTML/index.html>

Stationary and non stationary time series

<http://www.cas.usf.edu/~cconnor/geolsoc/html/chapter11.pdf>

ARMA & ARIMA

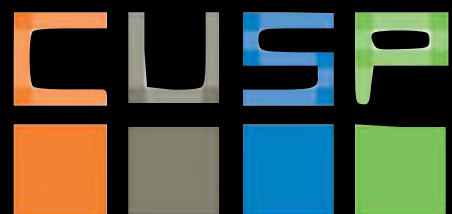
<http://www.econ.ohio-state.edu/dejong/note2.pdf>

A basic but quite intuitive Fourier Transform tutorial

<http://www.thefouriertransform.com/>

Fourier Transform for Imaging: it is actually a very common image analysis technique and urban science relies a lot on imaging and computer-vision techniques <http://homepages.inf.ed.ac.uk/rbf/HIPR2/fourier.htm>

Time series classification in python, not covered but you should read about it!  
<http://alexminnaar.com/time-series-classification-and-clustering-with-python.html>



***don't forget to vote!***

VIII: Topics in Time series



## References on clustering

Clustering: Science or Art??

Ulrike von Luxburg, Robert C. Williamson, Isabelle Guyon, 2009

<http://users.cecs.anu.edu.au/~williams/papers/P184.pdf>

Determining the number of groups from  
measures of cluster stability

G. Bel Mufti, P. Bertrand and L. El Moubarki, 2005

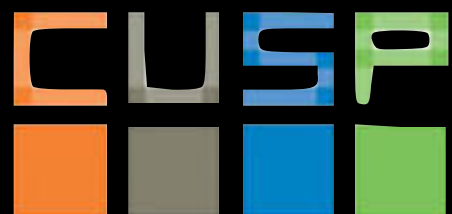
[http://citeseerx.ist.psu.edu/viewdoc/download?  
doi=10.1.1.98.4941&rep=rep1&type=pdf](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.4941&rep=rep1&type=pdf)

Clustering technique-based least square support vector machine for  
EEG signal classification

Siulya, Yan Lia, Peng (Paul) Wenb, 2010

(This is in the field of neuroscience, but it discusses clustering of time  
series. You should have access to it from an NYU internet connection)

<http://www.sciencedirect.com/science/article/pii/S0169260710002907>



***don't forget to vote!***

VIII: Topics in Time series