Kernels for Periodic Time Series Arising in Astronomy

Gabriel Wachman — Tufts University

Roni Khardon — Tufts University

Pavlos Protopapas — Harvard CfA and IIC

Charles A. Alcock — Harvard CfA

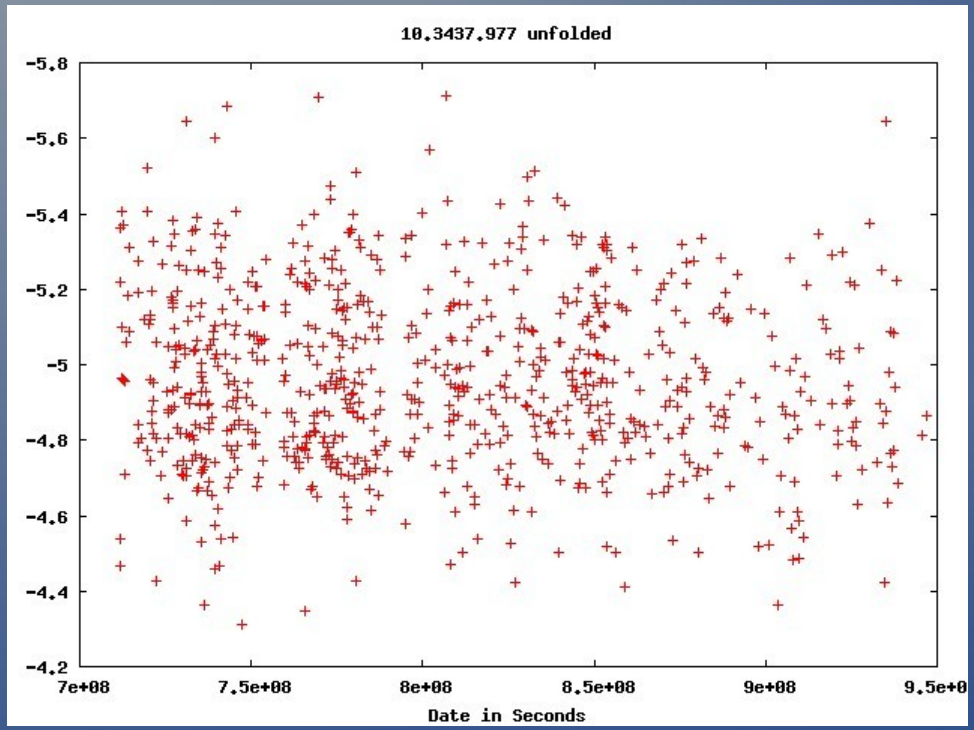
Acknowledgments

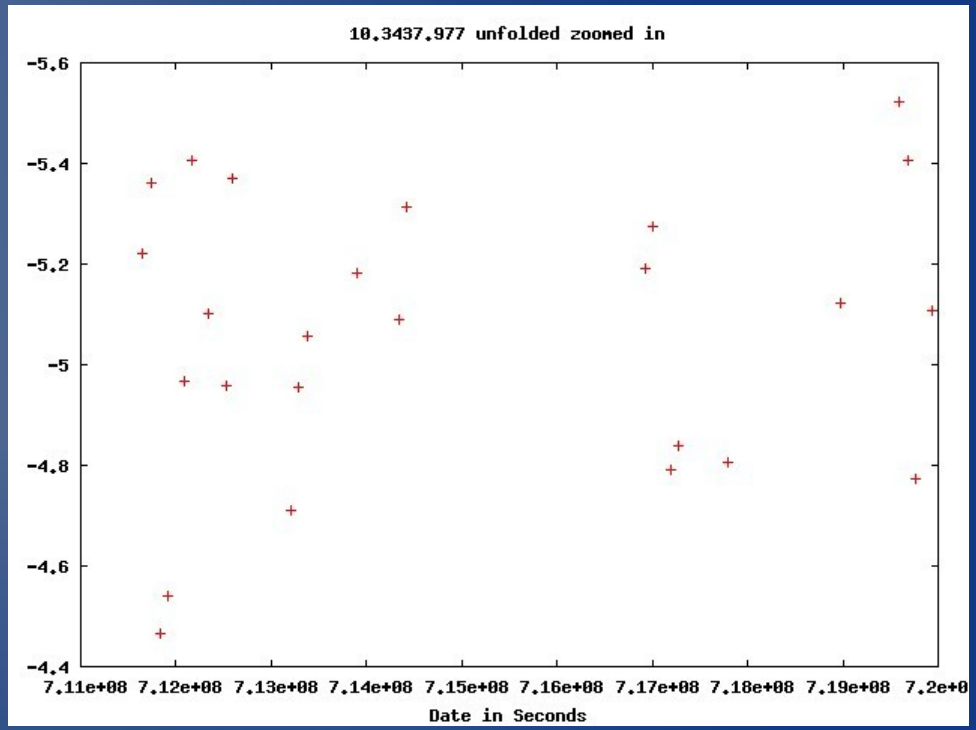
- Initiative for Innovative Computing
- NSF grant IIS-080340
- Odyssey cluster supported by the FAS Research Computing Group at Harvard
- Tufts High-performance computing research cluster

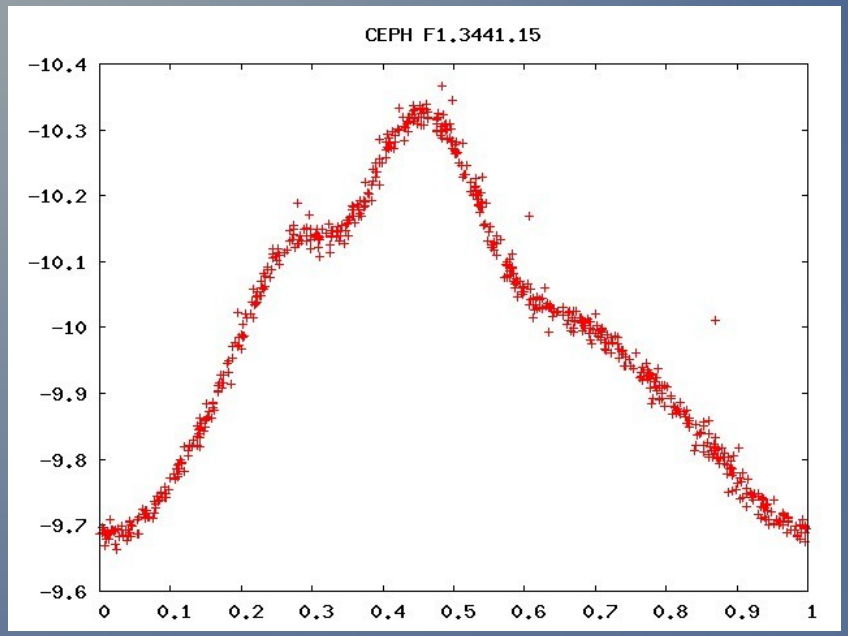
Motivation

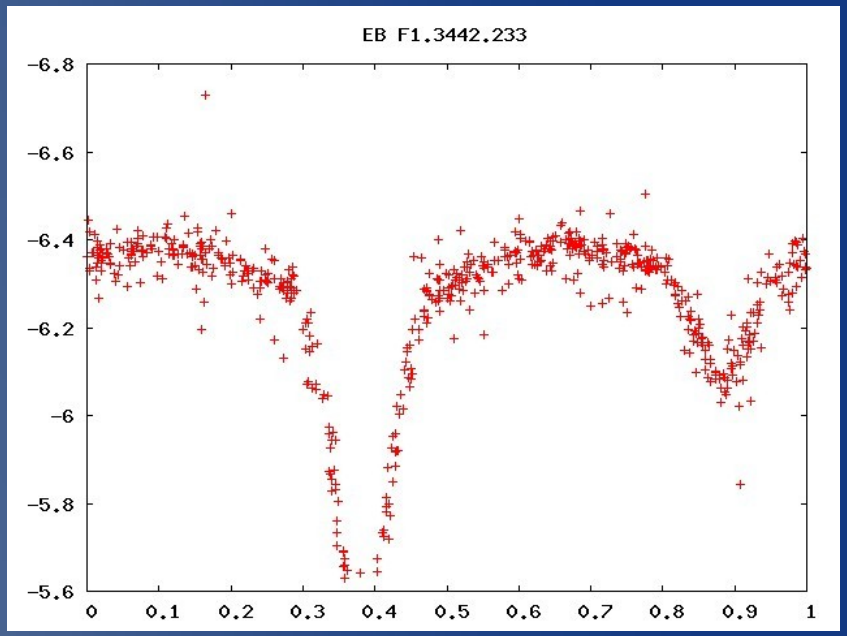
- Astronomy surveys contain millions of objects
- Significant time requirement to do manual classification
- Many stars remain unclassified
- Stars are represented as a *time series*

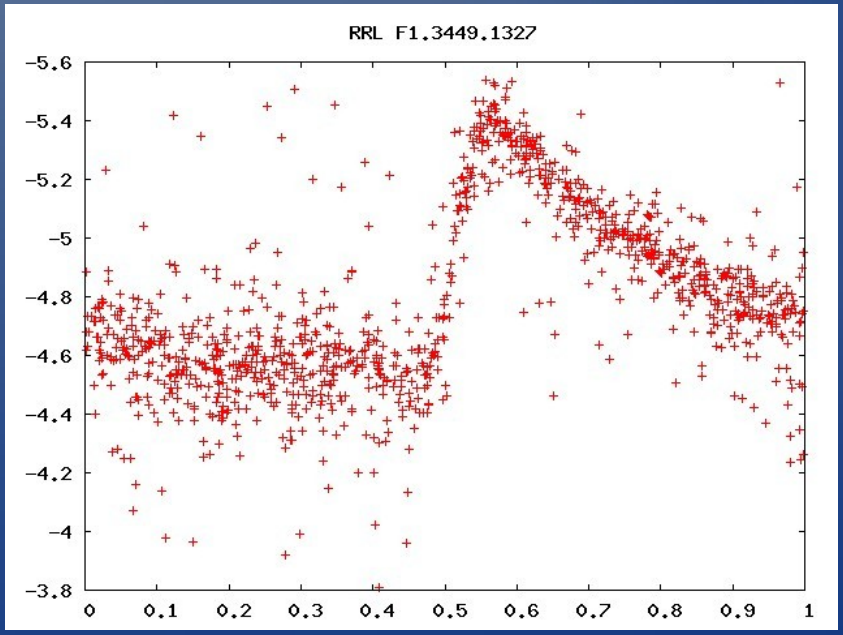
Periodic Variable Stars





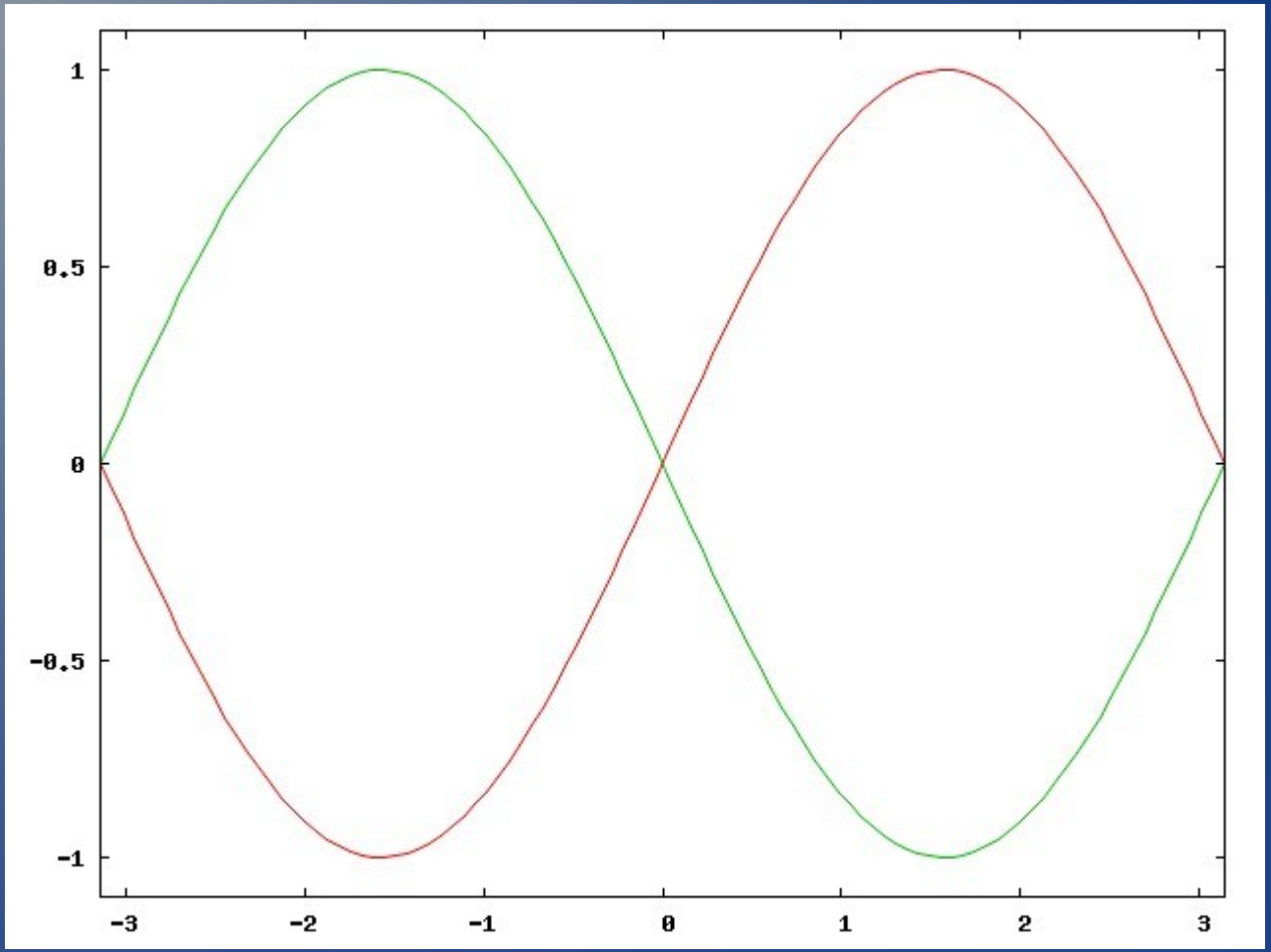

CEPH F1.3441.15
-10.4
-10.3
-10.2
-10.1
-10
-9.9
-9.8
-9.7
-9.6
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

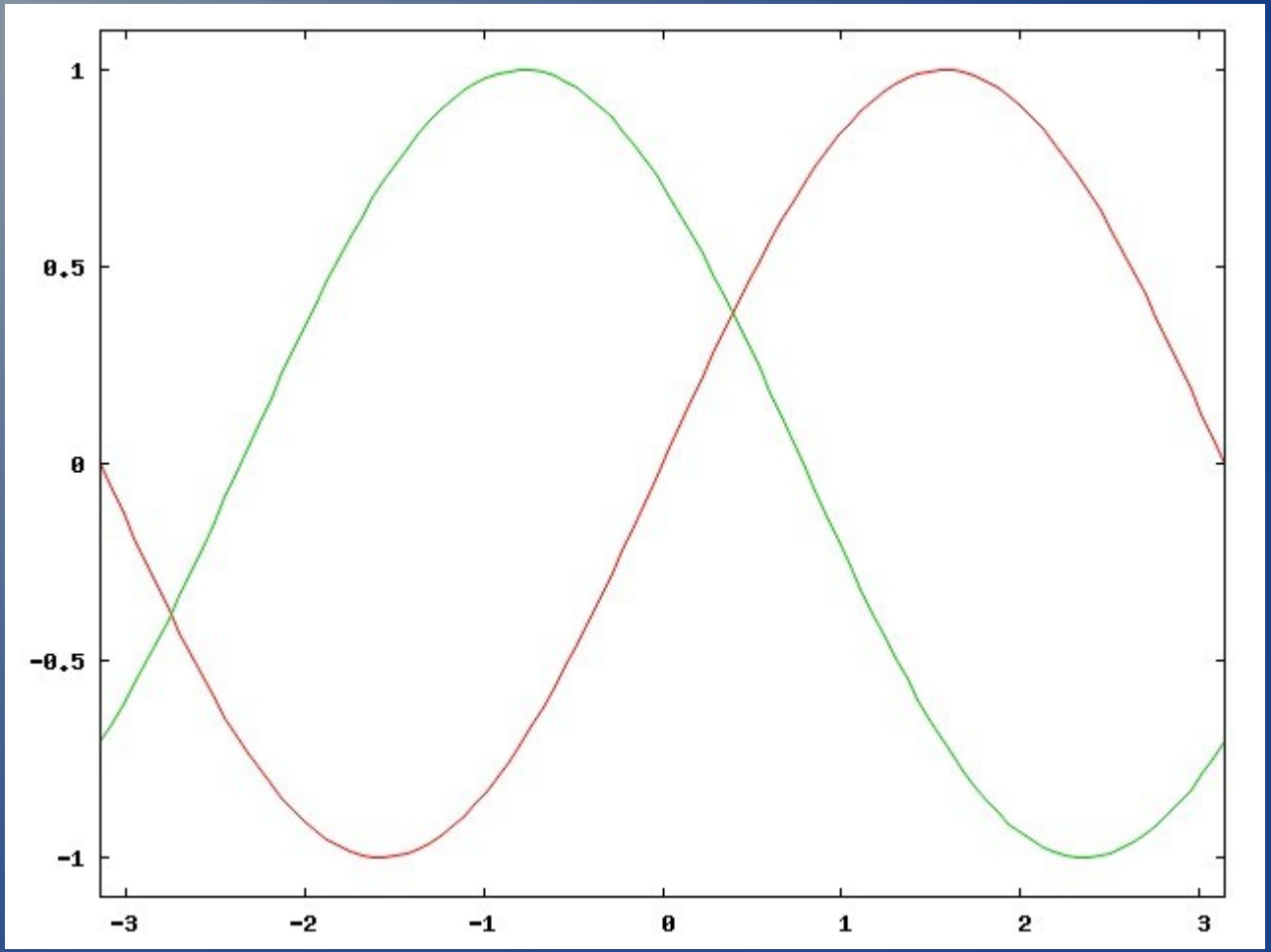

EB F1.3442.233
-6.8
-6.6
-6.4
-6.2
-6
-5.8
-5.6
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

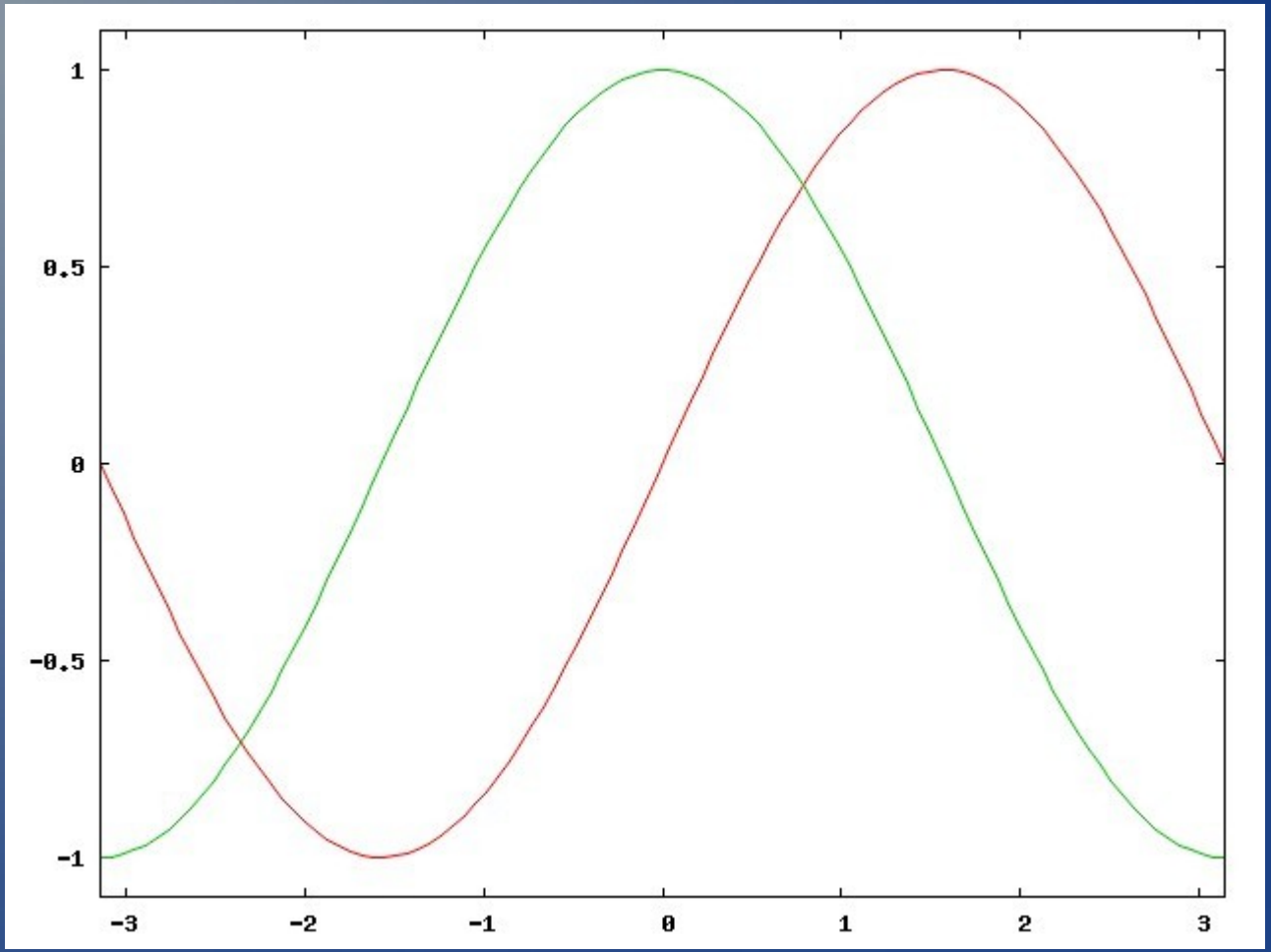

RRL F1.3449.1327
-5.6
-5.4
-5.2
-5
-4.8
-4.6
-4.4
-4.2
-4
-3.8
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

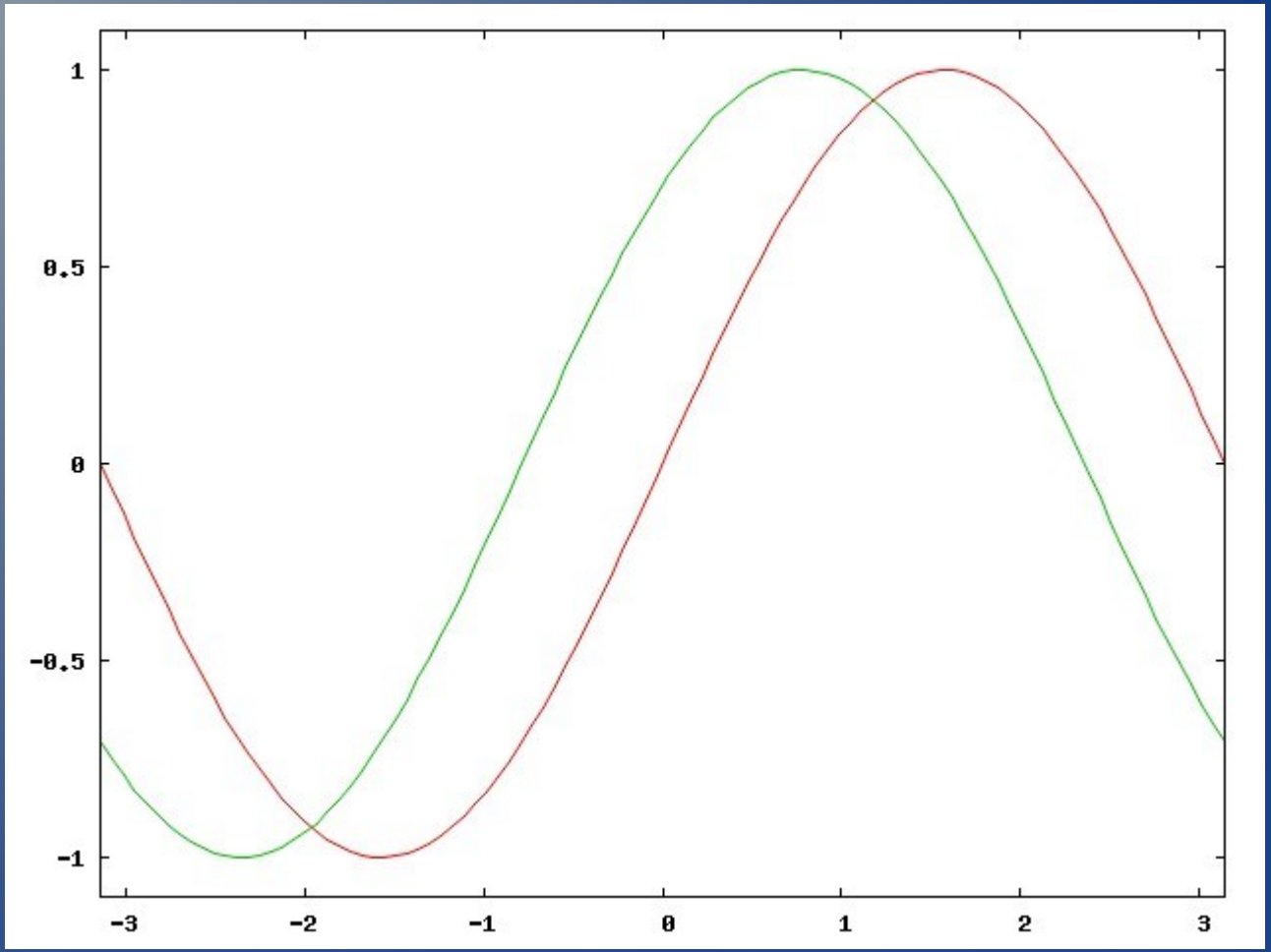
Agenda

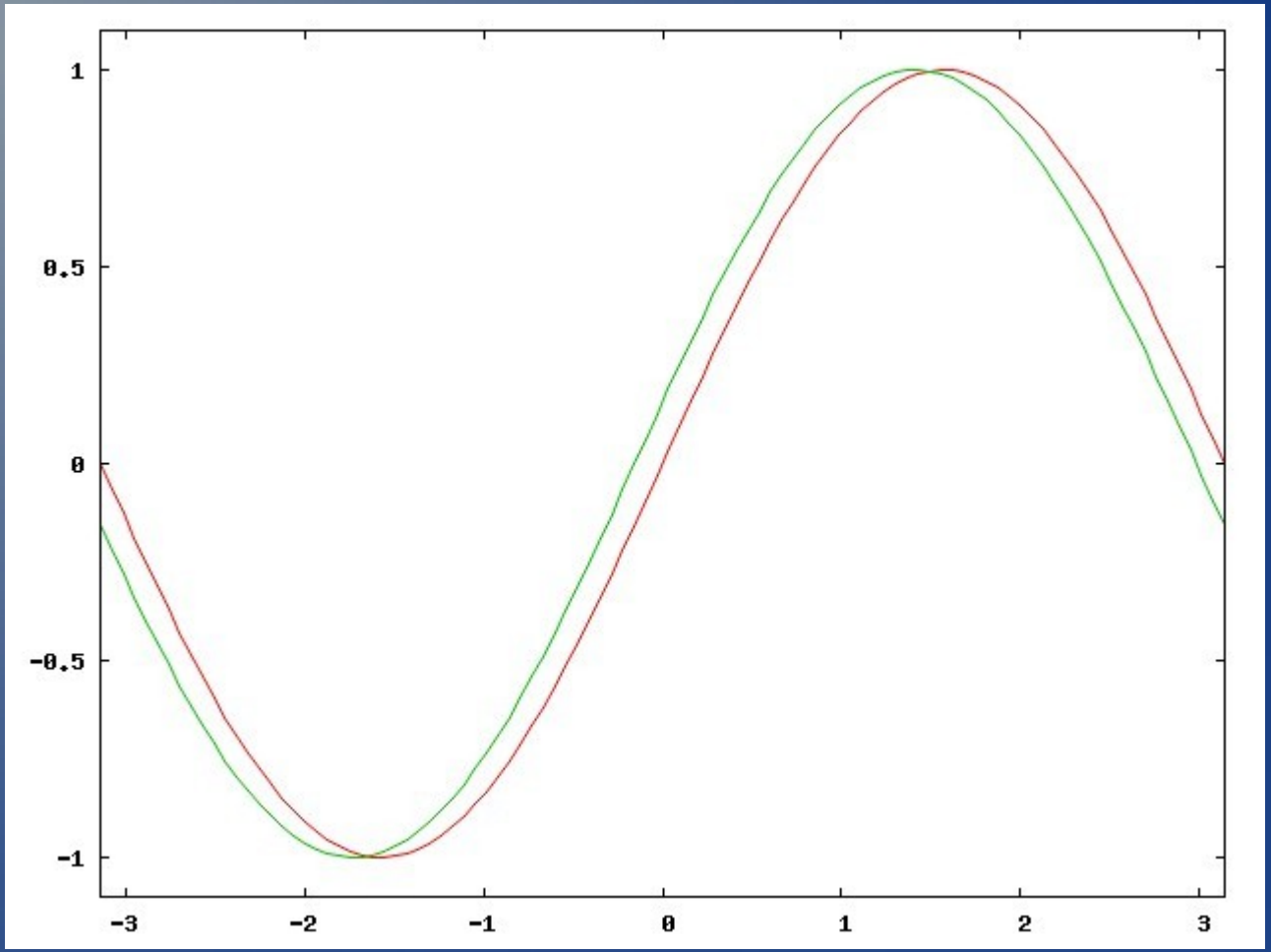
- Introduction and Motivation
- Similarity Measure for Time Series
- Kernel for Periodic Time Series
- Experiments and Results
- Big Picture
- Summary


1
0.5
0
-0.5
-1
-3
-2
-1
0
1
2
3


1
0.5
0
-0.5
-1
-3
-2
-1
0
1
2
3


1
0.5
0
-0.5
-1
-3
-2
-1
0
1
2
3


1
0.5
0
-0.5
-1
-3
-2
-1
0
1
2
3


1
0.5
0
-0.5
-1
-3
-2
-1
0
1
2
3

Kernel for Periodic Time Series

- We want to capture what it means for two time series to have the same "shape"
- S1 [Protopapas06]:

$$max_s \langle x , y_{+s} \rangle$$

- Does exactly what we want
- Can compute using FFT in O(nlogn)
- Is it a kernel?

Properties of S1

Theorem 1: S1 satisfies the Cauchy Schwartz inequality.

Theorem 2: Can construct a distance measure using S1 that satisfies triangle inequality.

Theorem 3: Any 3x3 Gram matrix of S1 is positive semidefinite.

Theorem 4: S1 is not positive semidefinite.

Kernel for Periodic Time Series

- K1:

$$\sum_{s=1}^{n} e^{\gamma \langle x, y_{+s} \rangle}$$

- Intuitively approximates maximum alignment
- Works as well in practice

Kernel for Time Series

- *Theorem 5:* K1 is positive semidefinite
- *Theorem 6:* K1 is computable in O(nlogn)

Shape Data

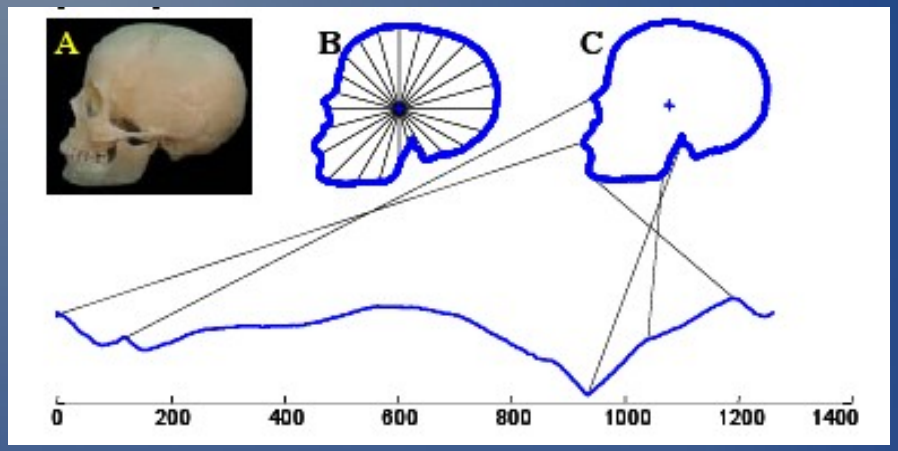


Image from **E. Keogh, L. Wei, X. Xi, S. Lee, M. Vlachos. LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures. VLDB 2006.**

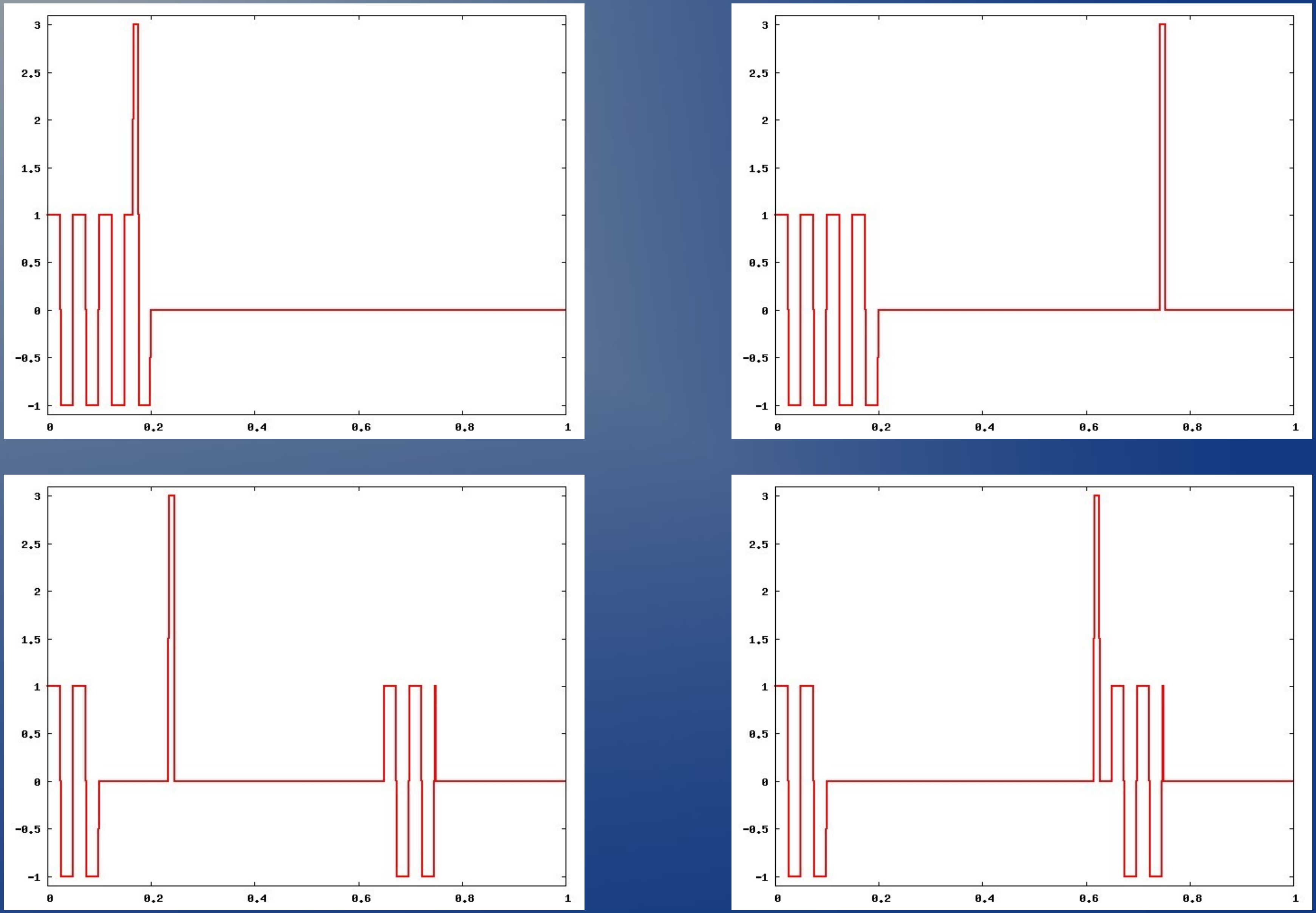
Experiments on Shape Data

	1-NN S1	1-NN K1	1-NN DTW	1-NN UP	SVM ED	SVM UP	SVM K1
A	0.54 ± 0.06	0.54 ± 0.08	0.33 ± 0.06	0.49 ± 0.05	0.20 ± 0.05	0.41 ± 0.05	0.63 ± 0.04
B	0.73 ± 0.04	0.73 ± 0.04	0.59 ± 0.08	0.70 ± 0.07	0.40 ± 0.10	0.65 ± 0.08	0.76 ± 0.08
I	0.98 ± 0.01	0.98 ± 0.01	0.84 ± 0.03	0.97 ± 0.02	0.47 ± 0.03	0.80 ± 0.02	0.91 ± 0.02

Experiments on Shape Data

	1-NN S1	1-NN K1	1-NN DTW	1-NN UP	SVM ED	SVM UP	SVM K1
A	0.54 ± 0.06	0.54 ± 0.08	0.33 ± 0.06	0.49 ± 0.05	0.20 ± 0.05	0.41 ± 0.05	0.63 ± 0.04
B	0.73 ± 0.04	0.73 ± 0.04	0.59 ± 0.08	0.70 ± 0.07	0.40 ± 0.10	0.65 ± 0.08	0.76 ± 0.08
I	0.98 ± 0.01	0.98 ± 0.01	0.84 ± 0.03	0.97 ± 0.02	0.47 ± 0.03	0.80 ± 0.02	0.91 ± 0.02

Breaking Universal Phasing



Experiments on Artificial Data

	1-NN S1	1-NN K1	1-NN UP	SVM UP	SVM K1
Artificial	0.99 ± 0.02	1.00 ± 0.00	0.65 ± 0.04	0.50 ± 0.07	0.997 ± 0.00
Artificial w/Noise	0.84 ± 0.14	0.84 ± 0.12	0.61 ± 0.09	0.53 ± 0.12	0.90 ± 0.05

Astronomy Data OGLE[Ud97,So03]

- 14087 periodic variables of type Cepheid, EB, RRL
- Periods given
- *Features* are average brightness, average color, and period

Experiments on Astronomy Data

	1-NN	SVM		1-NN	SVM
S1	0.844 ± 0.011	0.680 ± 0.011	features + S1	0.991 ± 0.002	0.998 ± 0.001
K1	0.901 ± 0.008	0.947 ± 0.005	features + K1	0.992 ± 0.002	0.998 ± 0.001
UP	0.827 ± 0.010	0.851 ± 0.006	features + UP	0.991 ± 0.002	0.997 ± 0.001
			features	0.938 ± 0.006	0.974 ± 0.004

Producing a Clean Catalog

	Ceph	EB	RRL
Ceph	3416	1	13
EB	0	3389	0
RRL	9	0	7259

	Ceph	EB	RRL
Ceph	3382	1	3
EB	0	3364	0
RRL	1	0	7195

	Ceph	EB	RRL
Ceph	3352	1	0
EB	0	3312	0
RRL	0	0	7138

Test Survey: MACHO

- ~25 million stars
- ~50,000 are periodic variables
- Two primary issues:
  - Eliminating *just* the other 24,950,000 stars
  - Finding the periods

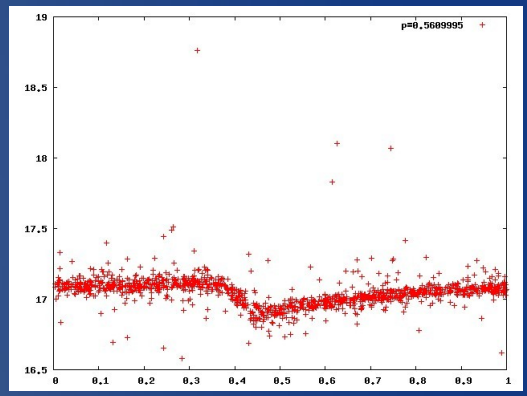
Astronomy Classification

- Eliminate non-variables, non-periodic, and any stars not of type Cepheid, RRL, or EB
- Find periods automatically
- Scalable approach

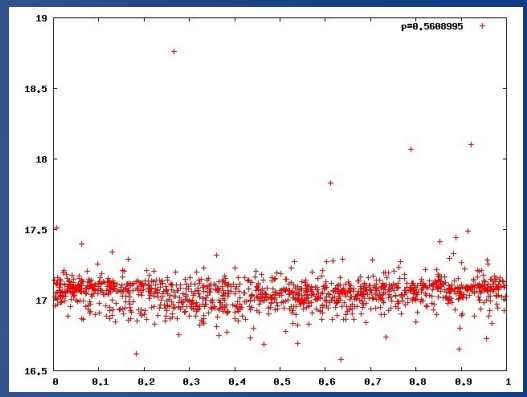
Period Estimation

- Challenges
  - Sensitivity of period
  - Range of potential periods
    - 0.25d – 20d
  - Computational cost

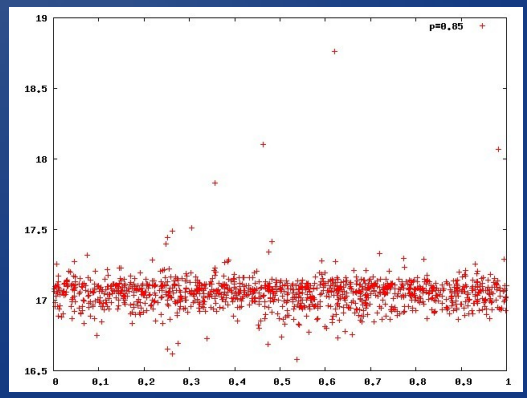
0.5609995



0.5608998



0.85



Summary

- Insight into cross-correlation
- Kernel for periodic time series
  - Similar to cross-correlation
  - Efficiently computable
  - Works well in multiple domains
- Integral piece of larger astronomy project
- Application of maximum alignment approximation to other domains

Questions

gabriel.wachman@tufts.edu