



Robust Automatic Classification of variable stars:

*Dealing with incomplete time series and
discovering deep patterns for light-curve
representation*

*Karim Pichara
Computer Science Department PUC*

Nicolás Castro, DCC UC - Andrés Riveros, DCC UC - Pavlos Protopapas, IACS

People developing this project



Nicolás Castro
Master Student



Andrés Riveros
Master Student



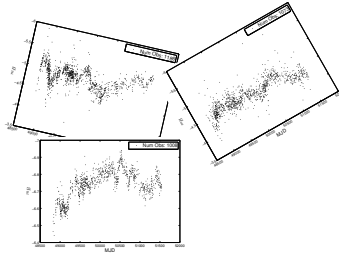
Pavlos Protopapas
IACS, Harvard



Karim Pichara
DCC, PUC

Lightcurve Representation

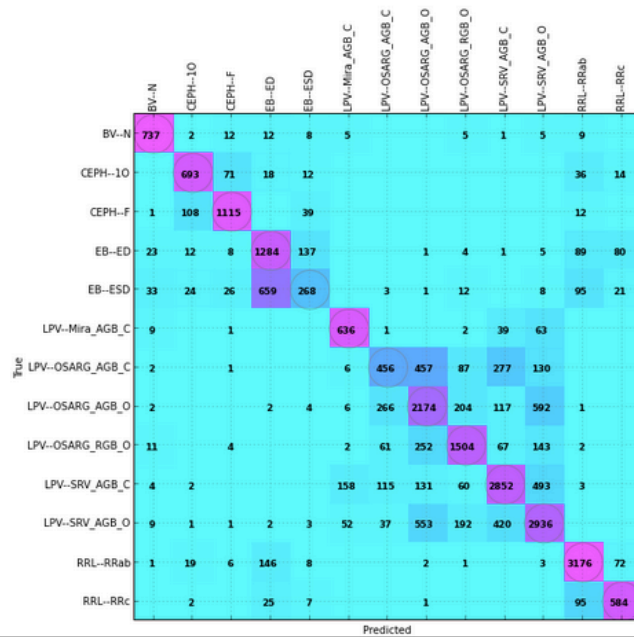
lightcurves



Dataset of Features

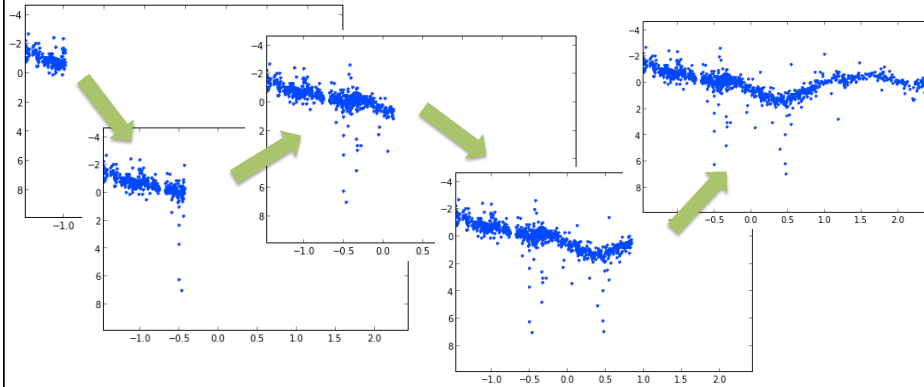
698	0.139	0.1339	0.0011	0.2864	0.0020	4.7916	0.0702	32.0	2350.0	17.5
1551	0.106	0.1439	0.0004	0.2653	0.0033	5.1840	0.0670	34.2	2251.1	4.9
3550	0.662	0.1492	0.0007	0.1961	0.0023	4.9771	0.0559	50.1	2072.2	8.0
247	0.096	0.1501	0.0001	0.3701	0.0039	7.7016	0.0960	31.1	2071.1	6.0
296	0.192	0.1511	0.0004	0.3700	0.0039	7.9344	0.0960	15.8	2033.0	4.6
319	0.010	0.1593	0.0004	0.3062	0.0040	6.8440	0.2020	37.0	2463.3	20.2
241	0.052	0.1623	0.0000	0.3603	0.0041	8.3225	0.1004	39.4	2401.1	8.0
330	0.010	0.1609	0.0001	0.4004	0.0043	9.1166	0.1056	12.5	2467.2	5.2
322	0.230	0.1603	0.0001	0.2272	0.0010	7.7039	0.0924	20.2	2411.1	6.6
312	0.110	0.1601	0.0007	0.3910	0.0043	9.2711	0.1063	15.3	2467.2	6.5
244	0.022	0.1701	0.0010	0.4302	0.0048	10.2034	0.1269	9.8	2392.2	9.5
147	0.106	0.1703	0.0004	0.4305	0.0045	9.9464	0.1011	11.0	2392.0	4.0
84	0.028	0.1729	0.0004	0.4365	0.0047	10.5683	0.1136	30.1	2353.0	3.8
696	0.070	0.1729	0.0022	0.2322	0.0049	5.8973	0.1342	52.2	2953.2	20.7
127	0.070	0.1741	0.0011	0.4701	0.0054	10.3697	0.1270	6.6	2397.3	17.1
94	0.022	0.1740	0.0004	0.4387	0.0049	10.3735	0.1269	10.0	2041.1	4.0
220	0.011	0.1749	0.0004	0.4865	0.0050	11.2214	0.1224	5.3	2053.2	3.7
230	0.001	0.1749	0.0004	0.7306	0.0056	6.4756	0.0807	39.2	2053.2	5.6
521	0.042	0.1776	0.0006	0.2779	0.0030	6.0037	0.0769	39.9	2050.9	5.2
97	0.030	0.1785	0.0011	0.4487	0.0053	11.4649	0.2097	6.4	2053.0	10.8
101	0.021	0.1786	0.0012	0.4287	0.0048	10.9211	0.1394	12.7	2461.1	11.0
431	0.000	0.1816	0.0007	0.3973	0.0035	7.6923	0.0913	35.2	2072.2	6.4
59	0.241	0.1817	0.0000	0.4466	0.0032	11.2271	0.1405	30.3	2063.3	7.4
44	0.030	0.1821	0.0007	0.4204	0.0050	11.3033	0.1331	9.8	2072.0	6.5
79	0.004	0.1824	0.0000	0.4889	0.0051	11.7949	0.1345	7.3	2072.2	3.1
307	0.030	0.1825	0.0001	0.4524	0.0053	12.0271	0.1447	5.2	2072.6	7.6
190	0.134	0.1820	0.0012	0.4709	0.0053	11.5811	0.1539	9.3	2072.2	11.1
67	0.040	0.1897	0.0012	0.4362	0.0050	11.2226	0.1594	9.8	2064.4	13.2
112	0.011	0.1869	0.0006	0.4900	0.0054	12.6219	0.1442	5.3	2143.3	5.0
1554	0.130	0.1975	0.0017	0.2531	0.0051	6.5142	0.1449	40.7	2701.1	15.1
59	0.034	0.1979	0.0013	0.3529	0.0053	8.8250	0.1503	32.0	2253.2	11.7

Automatic classification: Results so far

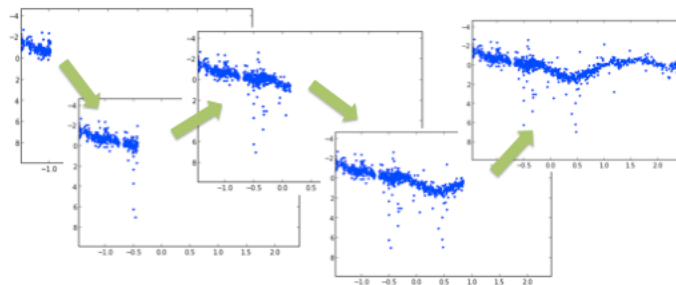


Motivation: Observational projects take several years to finish the lightcurves

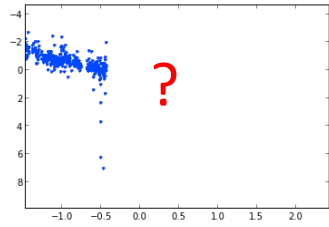
Few years ($5 < \text{Few} < 12$)



May be we can do something in the meantime?



Directly calculating the features from incomplete lightcurves seems not right

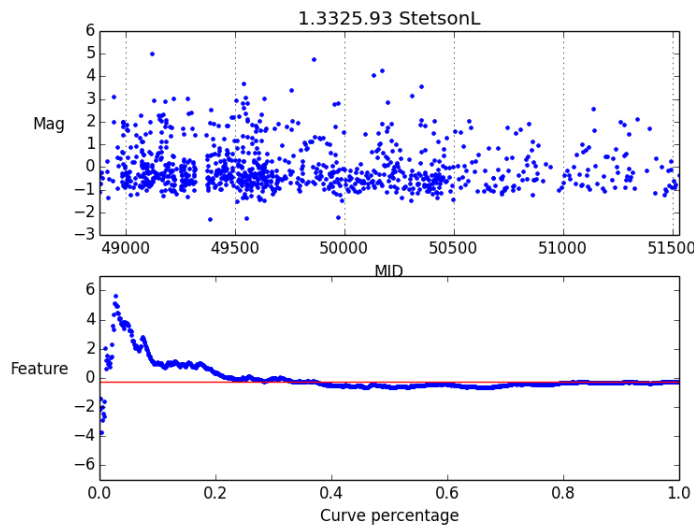


Dataset of Features

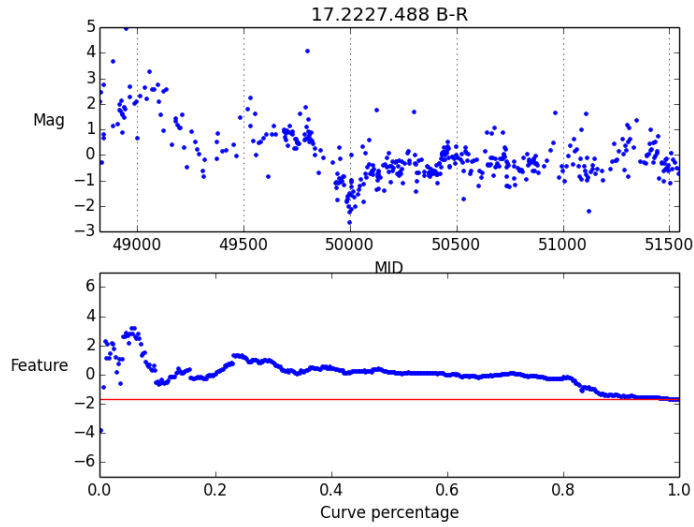
609	0.330	0.1309	0.00113	0.2366	0.0020	4.7916	0.0702	32.0	2109.8	37.5
1351	0.106	0.1409	0.0009	0.2013	0.0033	3.440	0.0670	34.2	2225.1	4.9
13508	0.02	0.1401	0.0007	0.1981	0.0023	0.771	0.0520	20.1	2307.2	8.0
247	0.006	0.1301	0.0005	0.3721	0.0009	7.7916	0.0060	13.1	2347.1	6.0
206	0.302	0.151	0.0004	0.3710	0.0039	7.9344	0.005	15.4	2403.8	4.6
139	0.015	0.0006	0.3062	0.0000	0.0460	0.0000	0.0000	17.8	2403.8	20.3
241	0.025	0.1623	0.0000	0.308	0.0041	8.1325	0.000	19.4	2403.1	8.8
330	0.01	0.1609	0.0003	0.400	0.0043	9.240	0.0035	12.5	2460.7	5.2
3262	0.20	0.1601	0.0007	0.3272	0.0030	0.0000	0.0004	20.2	2411.1	6.6
332	0.10	0.161	0.0007	0.3070	0.0043	9.2751	0.1063	15.3	2480.7	6.5
244	0.022	0.161	0.0007	0.4511	0.0040	10.2034	0.1070	9.0	2500.2	9.5
147	0.106	0.170	0.0005	0.430	0.0045	9.6464	0.101	11.9	2500.8	4.0
84	0.020	0.170	0.0004	0.434	0.0047	10.306	0.1130	10.1	2505.9	3.8
606	0.070	0.1770	0.0022	0.2322	0.0040	1.0000	0.1242	22.2	2505.3	20.7
107	0.036	0.191	0.0018	0.4270	0.0054	10.9657	0.1710	6.4	2672.1	17.1
96	0.020	0.190	0.0004	0.4307	0.0040	10.3753	0.1200	10.0	2684.1	4.0
230	0.011	0.170	0.0004	0.4665	0.0040	11.2010	0.084	5.3	2805.3	3.7
230	0.003	0.170	0.0006	0.2368	0.0046	0.0000	0.0000	39.5	2805.3	5.6
525	0.042	0.170	0.0006	0.2770	0.0030	0.0000	0.0000	39.9	2600.9	5.2
97	0.010	0.170	0.0013	0.607	0.0001	11.4044	0.0000	6.4	2808.8	40.6
110	0.021	0.170	0.0012	0.4207	0.0040	10.3021	0.1304	12.7	2800.1	11.0
437	0.000	0.1604	0.0007	0.3073	0.0013	7.6623	0.0913	2.2	2607.2	6.8
39	0.20	0.161	0.000	0.4666	0.0013	11.3271	0.1010	5.1	2808.3	7.8
43	0.004	0.161	0.0007	0.4204	0.0030	11.3033	0.1331	9.8	2672.0	6.5
79	0.004	0.161	0.0006	0.400	0.0011	11.3940	0.1345	7.3	2675.2	5.1
107	0.020	0.1625	0.0000	0.401	0.0033	11.378	0.1447	2.2	2675.6	7.6
106	0.134	0.160	0.0007	0.4374	0.0033	11.2091	0.1330	3.3	2670.2	11.1
47	0.040	0.1607	0.0007	0.4652	0.0030	11.2020	0.1504	4.8	2668.4	13.2
112	0.011	0.1606	0.0006	0.4000	0.0034	11.4019	0.1443	5.3	2714.3	5.0
104	0.130	0.1607	0.0017	0.2021	0.0031	0.2502	0.1449	46.7	---	---
50	0.014	0.1670	0.0013	0.3320	0.0013	8.6230	0.1500	32.0	---	---



Feature convergence

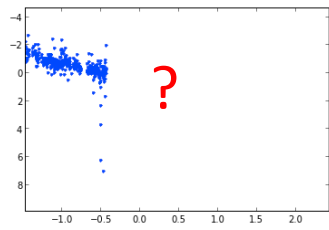


Feature convergence

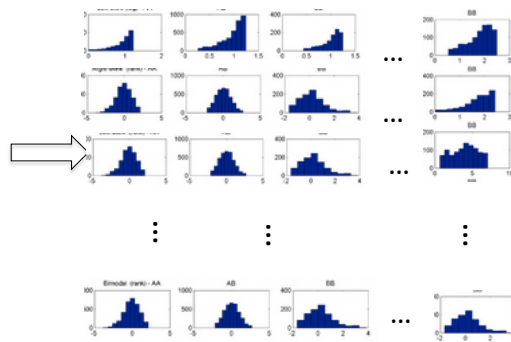


13

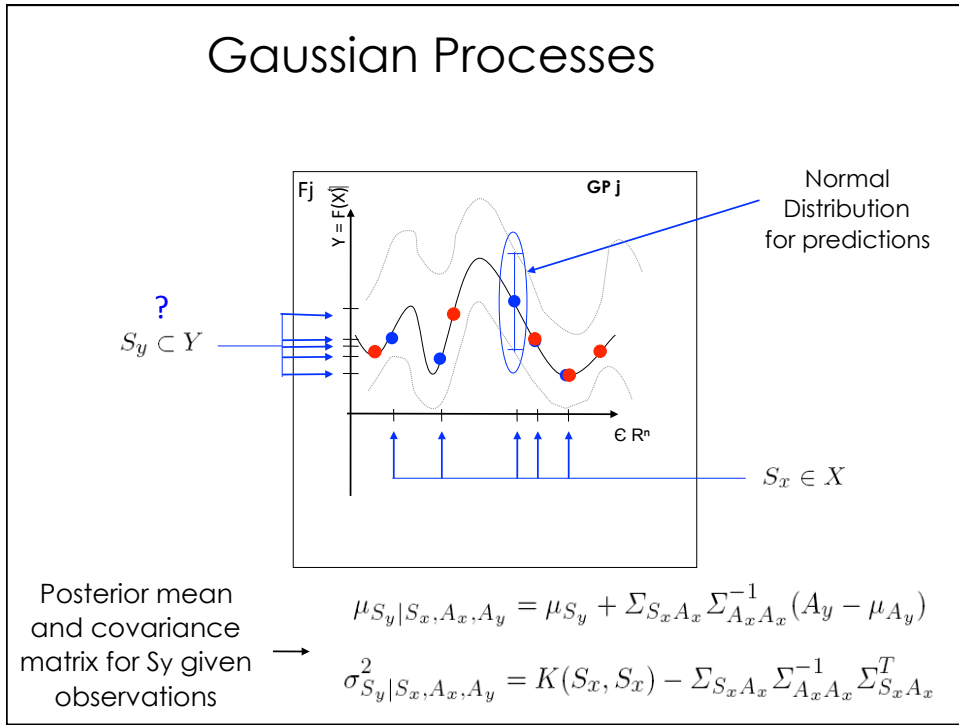
Idea: lets calculate distributions of features in order to model uncertainty



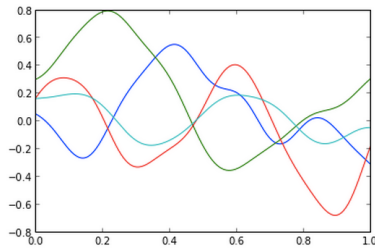
Dataset of distributions



Gaussian Processes

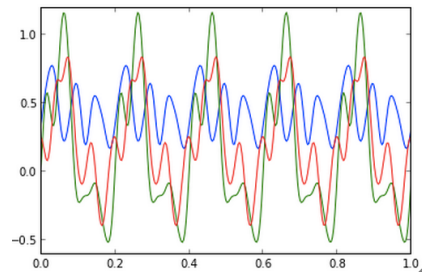


Gaussian Processes



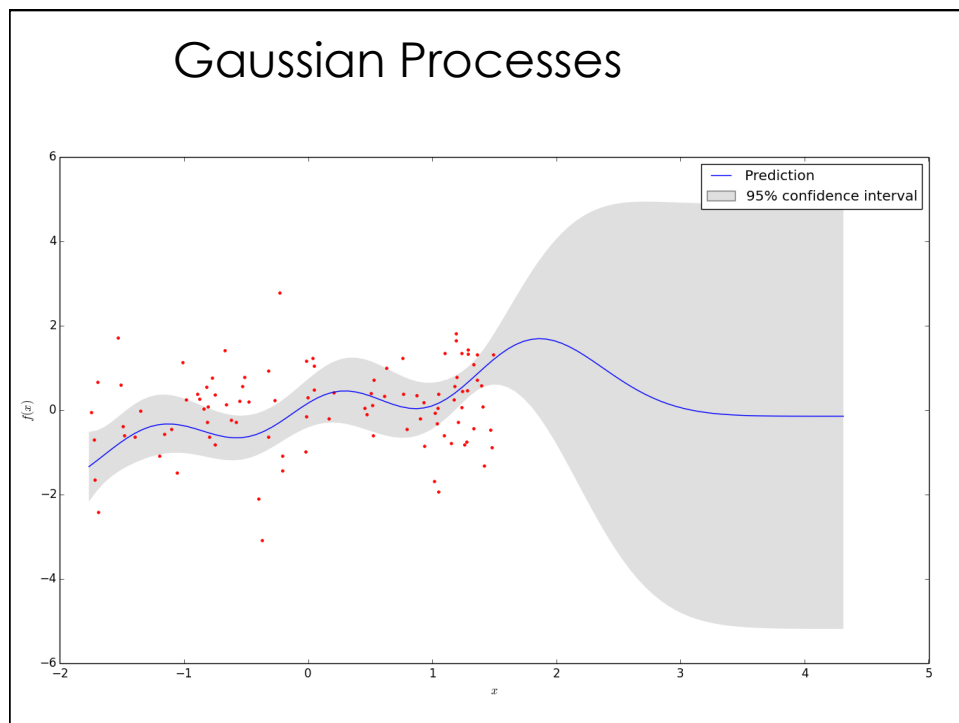
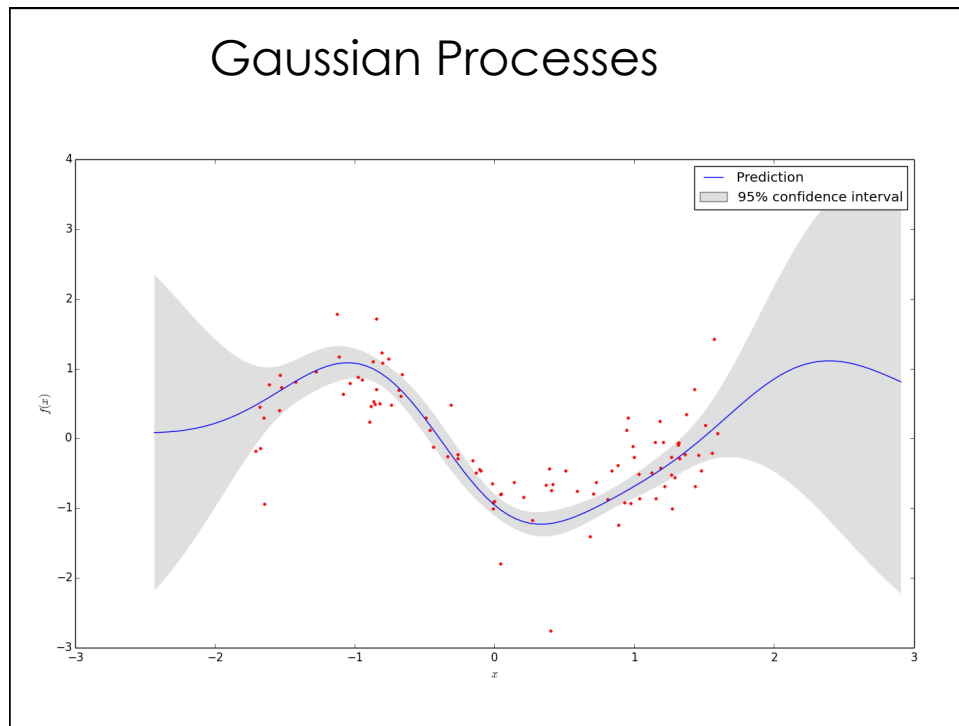
$$k_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$$

$$k_{Per}(x, x') = \sigma^2 \exp\left(-\frac{2 \sin^2(\pi(x-x')/p)}{\ell^2}\right)$$

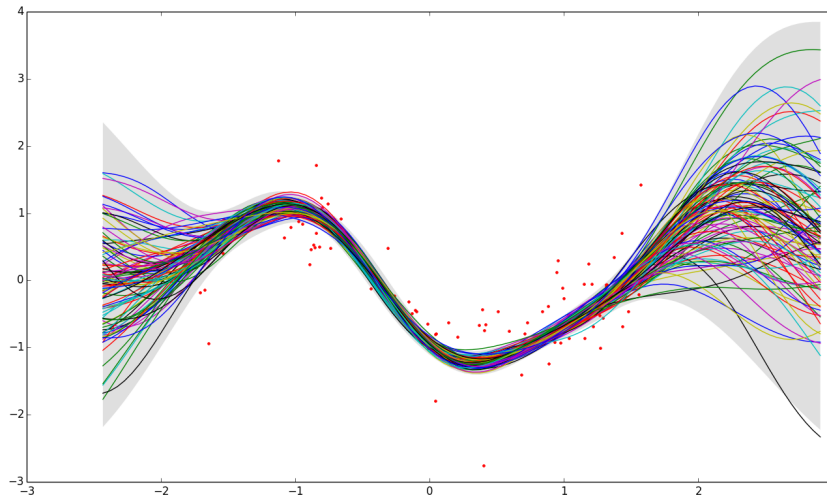


$$\mu_{S_y|S_x, A_x, A_y} = \mu_{S_y} + \Sigma_{S_x A_x} \Sigma_{A_x A_x}^{-1} (A_y - \mu_{A_y})$$

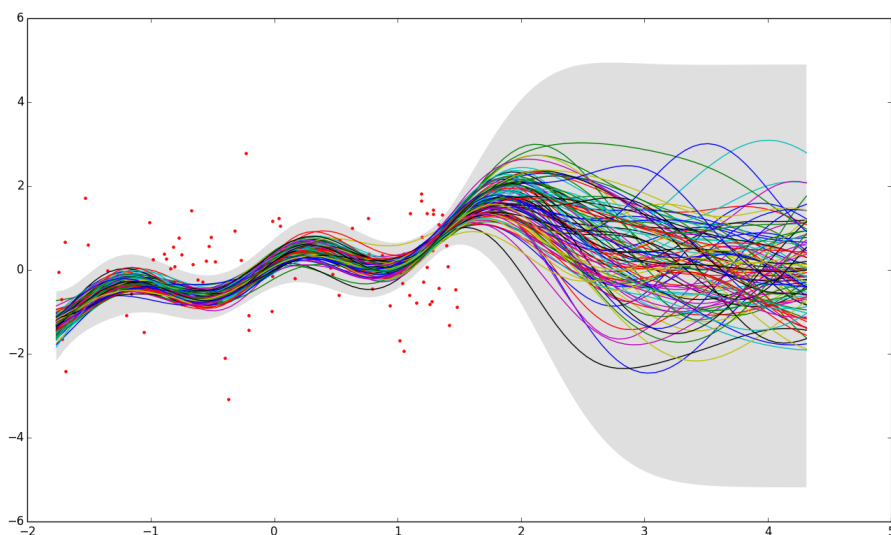
$$\sigma_{S_y|S_x, A_x, A_y}^2 = K(S_x, S_x) - \Sigma_{S_x A_x} \Sigma_{A_x A_x}^{-1} \Sigma_{S_x A_x}^T$$



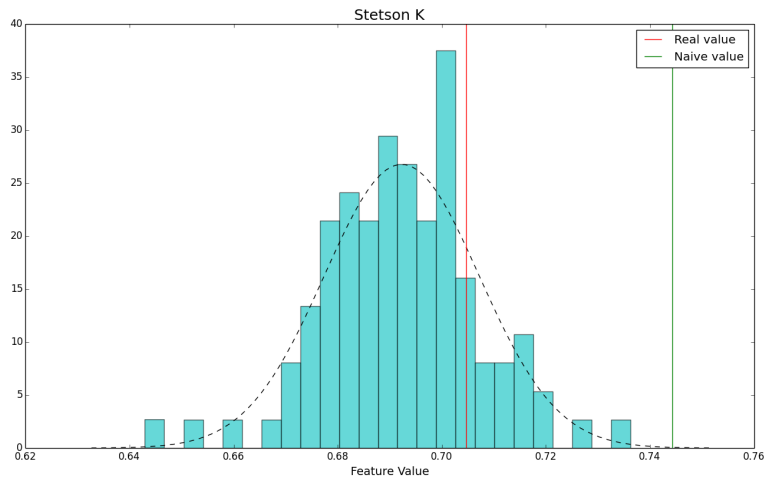
With Gaussian Processes we can take samples from lightcurves



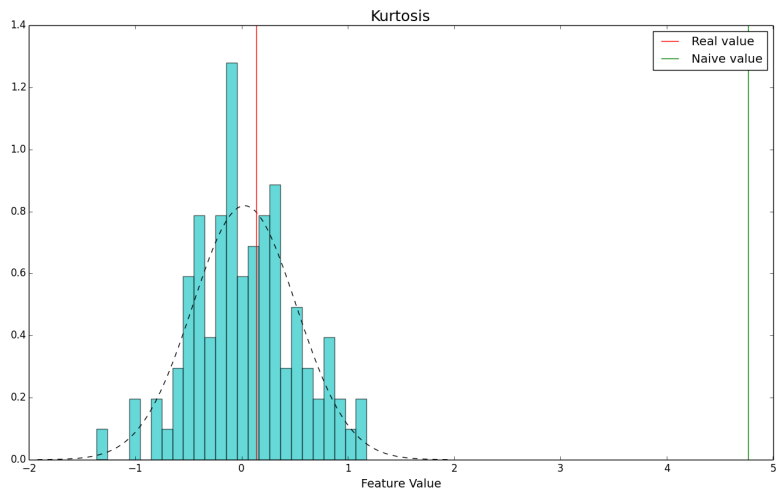
With Gaussian Processes we can take samples from lightcurves



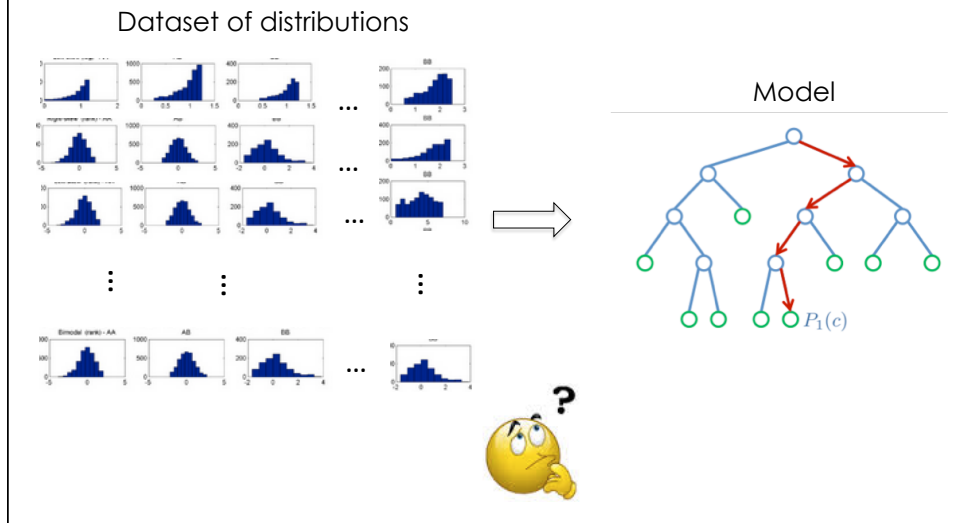
Having samples from lightcurves we can obtain samples from features



Having samples from lightcurves we can obtain samples from features

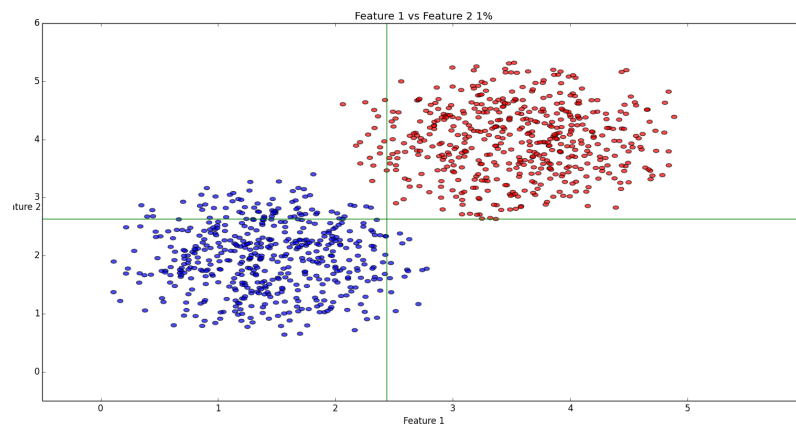


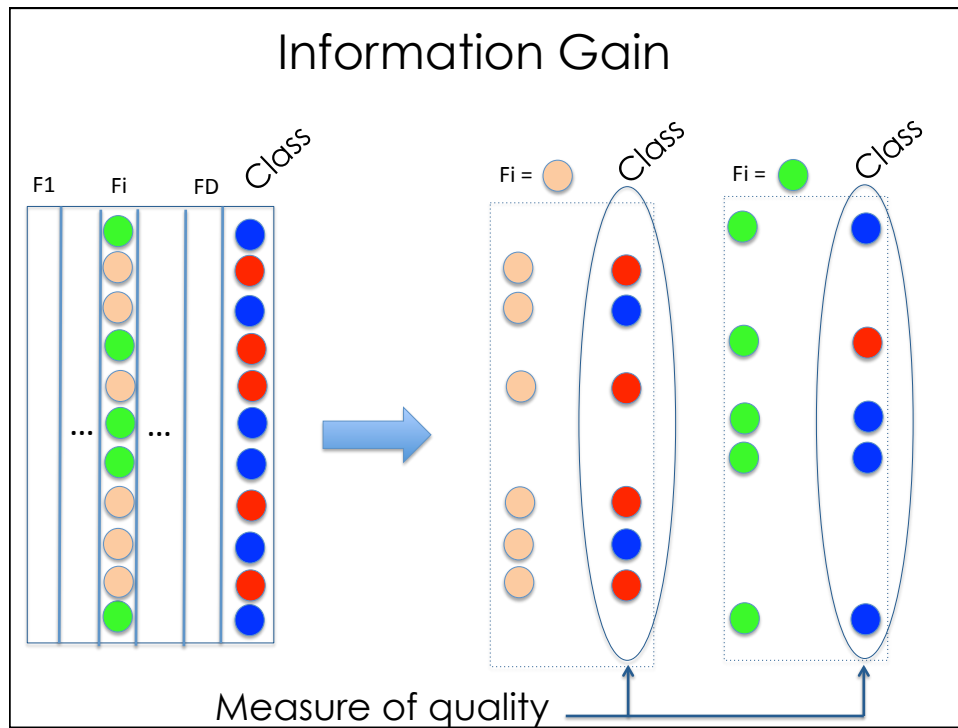
Once we have the features as distributions, how we can train a classifier?



Plain decision trees:

Search for splits in the variables that separate the data in different classes





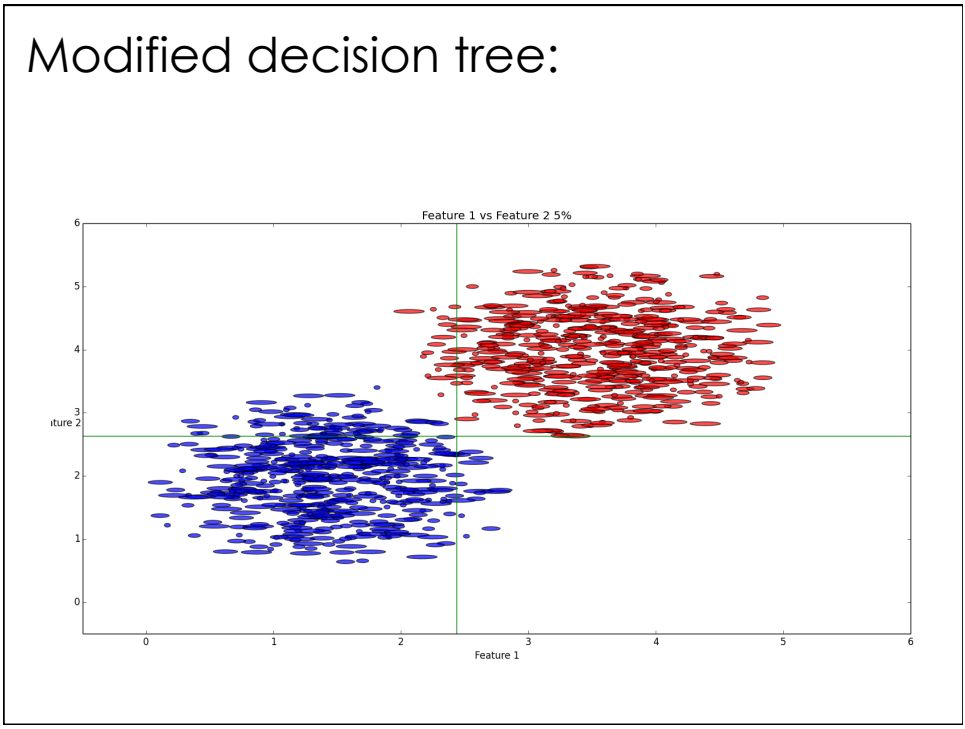
Measuring the quality of a split

Shannon's Entropy $H(D) = - \sum_{i=1}^m p_i \log_2(p_i)$

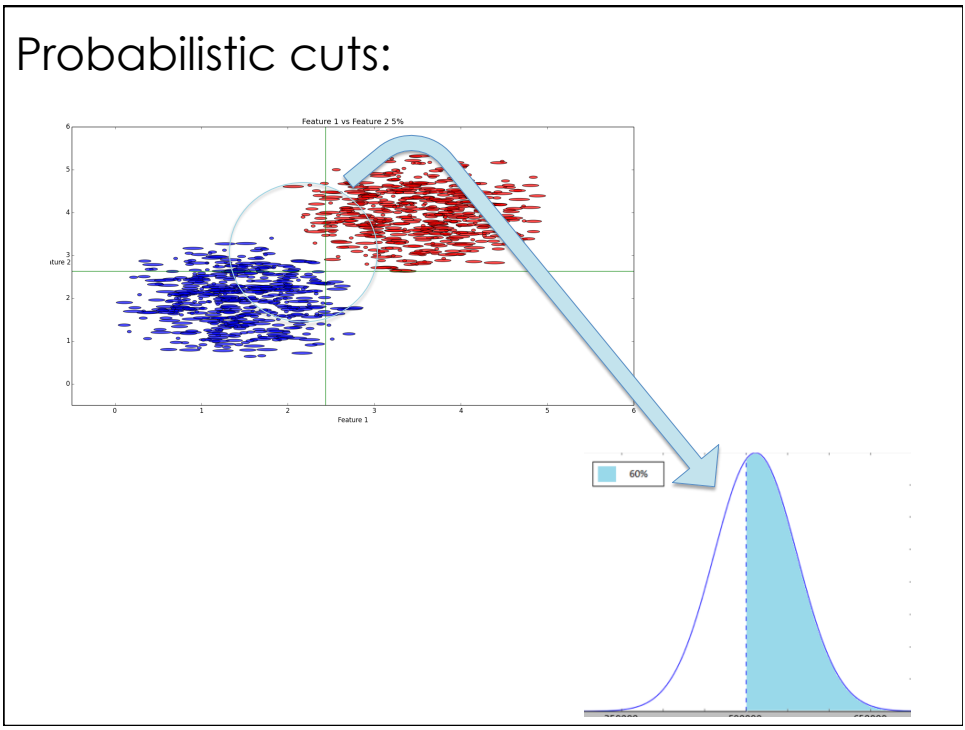
where p_i is the probability that an arbitrary tuple in D belongs to class i . It is estimated as:

$$p_i = \frac{|C_{i,D}|}{|D|} \quad , \quad |C_{i,D}| = \begin{array}{l} \text{Number of} \\ \text{samples of} \\ \text{class } i \text{ in } D \end{array}$$

Modified decision tree:



Probabilistic cuts:



Expected entropy

$$S(D) = - \sum_{i=1}^m p_i \log_2 p_i$$

The way p_i is estimated changes

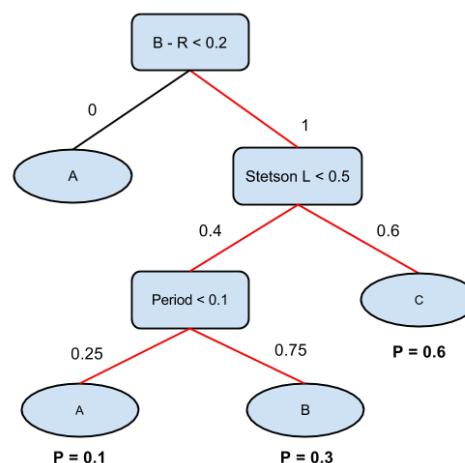
$$p_i = \frac{|C_{i,d}|}{|D|}, \quad |C_{i,d}| = \sum_{j \in i} \mathbf{w} \cdot \int_A(\mathbf{x}_j)$$

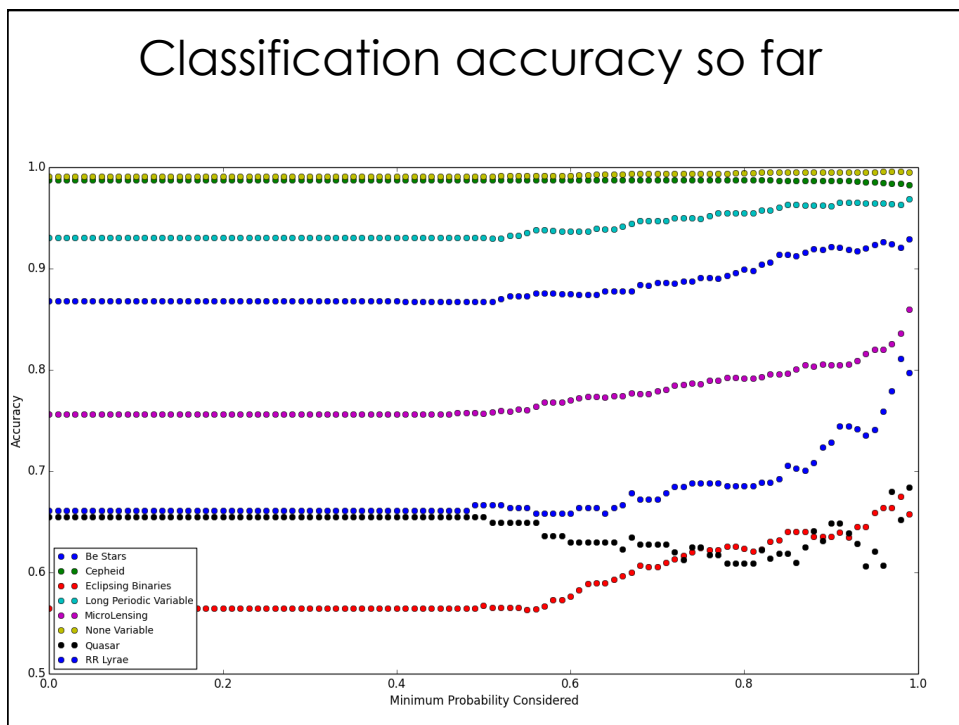
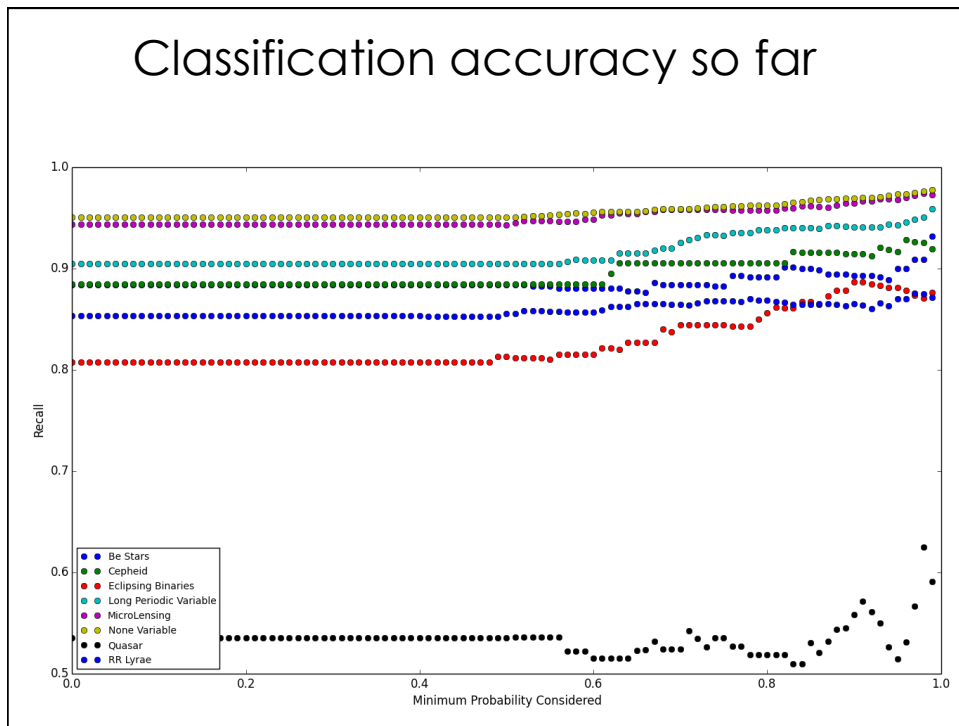
$$|D| = \sum_{\forall j} \int_A(\mathbf{x}_k)$$

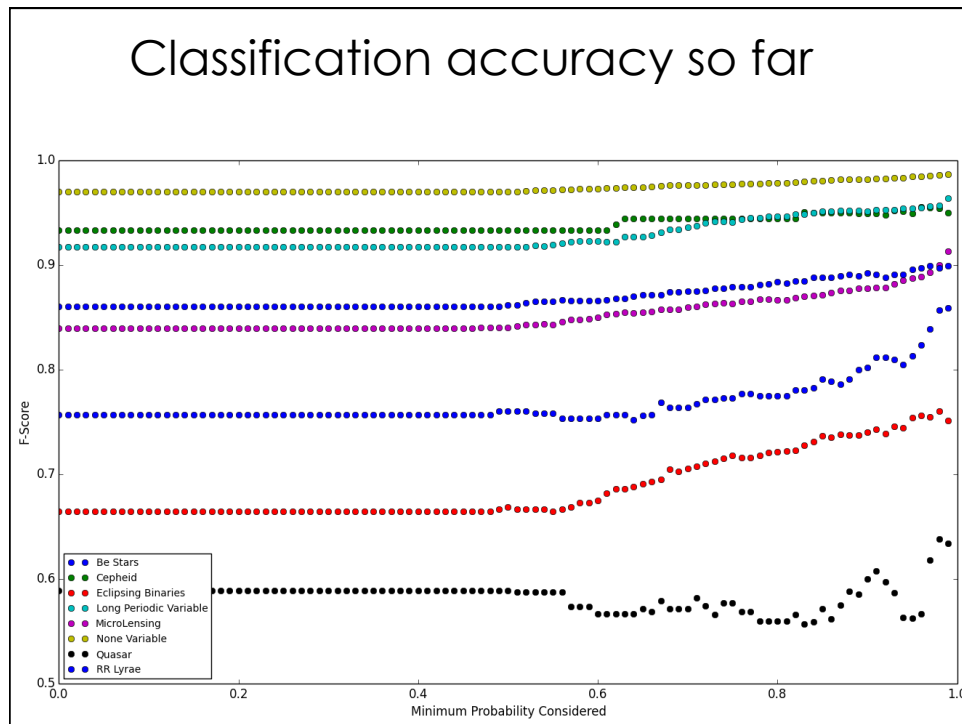
We weight the importance of each tuple by multiplying it by the probability that it will get to that given node.

After learning: Probabilistic classification

- A sample can go through many paths when being classified.
- The model chooses the class with highest probability.







Conclusions and Future Work

- We can perform automatic classification with “incomplete” lightcurves handling uncertain features
- For some classes, we can have good accuracies before completing the observational time
- We are working to increase our list of features with uncertainty

Conclusions and Future Work

- We expect to improve the fit of the Gaussian Processes in order to get better accuracies
- We will extend the uncertain decision tree to a uncertain Random Forest
- We will model the uncertainty of features with more suitable distributions

Conclusions and Future Work

- Still a lot to do with our models





Thank you!!!!



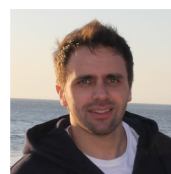
Nicolás Castro
Master Student



Andrés Riveros
Master Student



Pavlos Protopapas



Karim Pichara