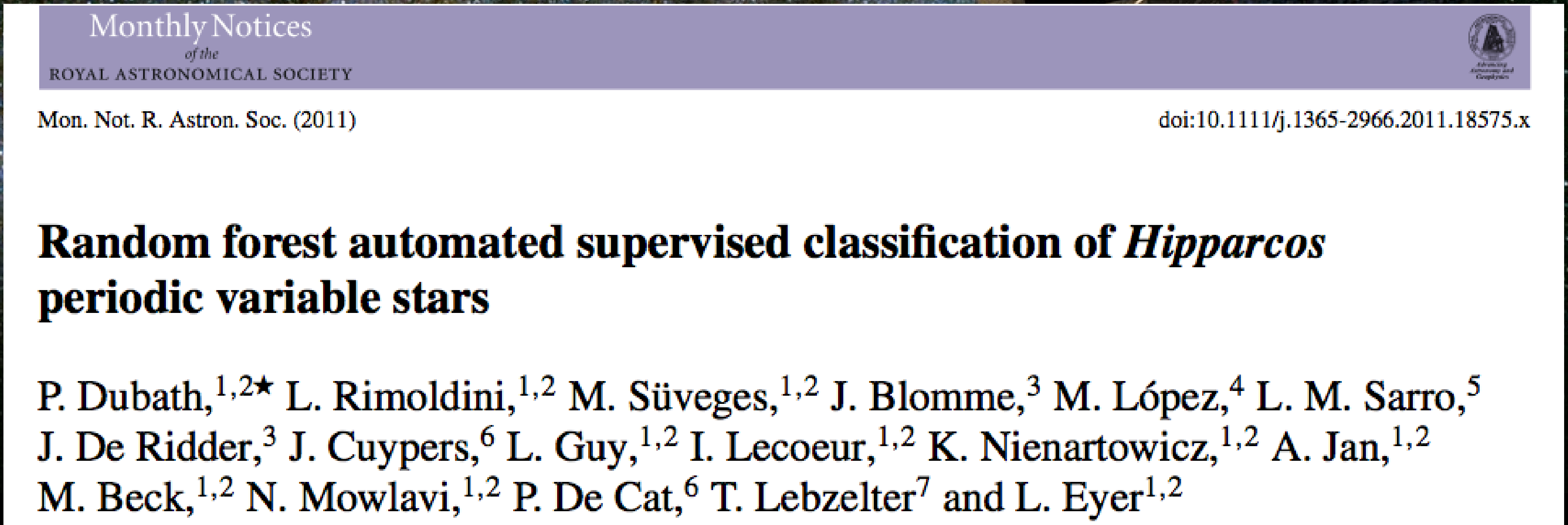
Variable Star Classification

Pierre Dubath

Observatory of the University of Geneva
Gaia CU7 team
Pierre.Dubath@unige.ch
June 1st, 2011



Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. (2011)   doi:10.1111/j.1365-2966.2011.18575.x

Random forest automated supervised classification of *Hipparcos* periodic variable stars

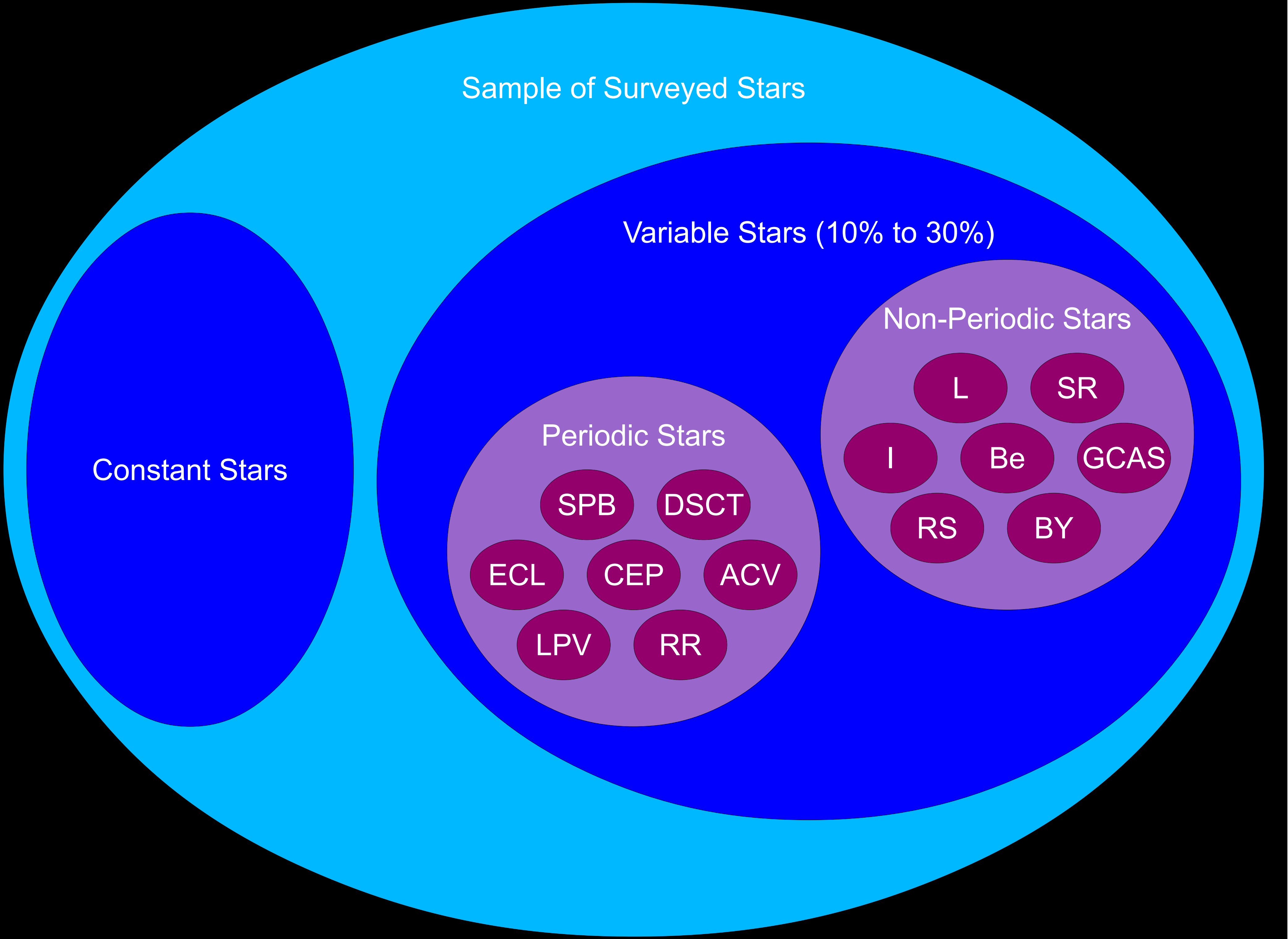
P. Dubath,[1,2★] L. Rimoldini,[1,2] M. Süveges,[1,2] J. Blomme,[3] M. López,[4] L. M. Sarro,[5] J. De Ridder,[3] J. Cuypers,[6] L. Guy,[1,2] I. Lecoeur,[1,2] K. Nienartowicz,[1,2] A. Jan,[1,2] M. Beck,[1,2] N. Mowlavi,[1,2] P. De Cat,[6] T. Lebzelter[7] and L. Eyer[1,2]

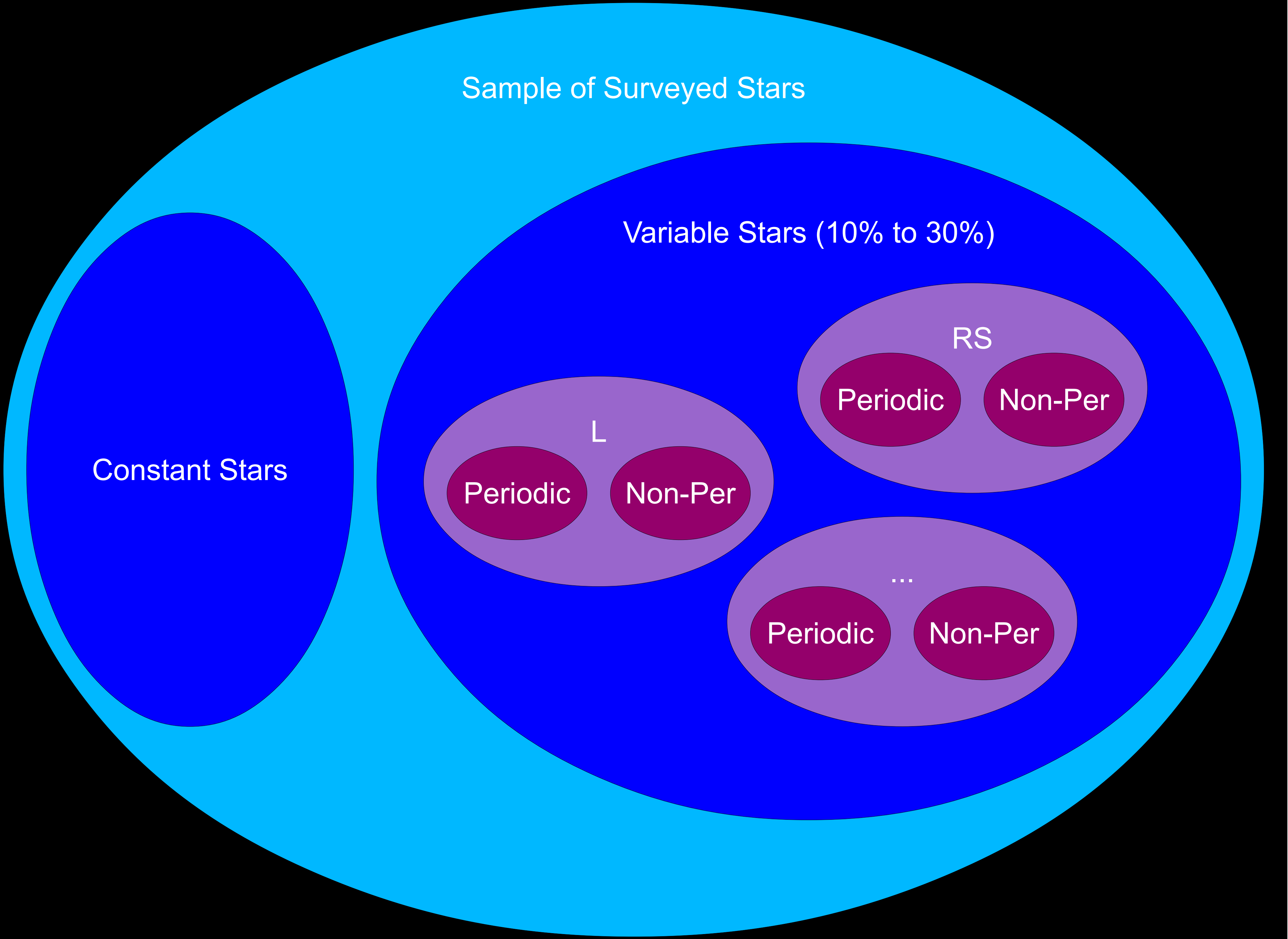
Context

- Work in progress of the CU7 team to get ready for the Gaia data analysis
- Using Hipparcos data as a control sample
  - ➔ Establish the best classification strategy
- Current choice: classification in three steps

Variable Star Classification

- Work in progress of the CU7 team to get ready for the Gaia data analysis (major contributions from Isabelle Lecoeur et Lorenzo Rimoldini)
- Random forest classification of Hipparcos periodic variables, Dubath et al. 2011, MNRAS
- Hipparcos unsolved variable classification, Rimoldini et al. in preparation
- Overall performance of a complete Hipparcos variable star classification


Sample of Surveyed Stars
Variable Stars (10% to 30%)
Constant Stars
Non-Periodic Stars
L
SR
I
Be
GCAS
RS
BY
Periodic Stars
SPB
DSCT
ECL
CEP
ACV
LPV
RR



Sample of Surveyed Stars

Constant Stars

Variable Stars (10% to 30%)

RS

Periodic

Non-Per

L

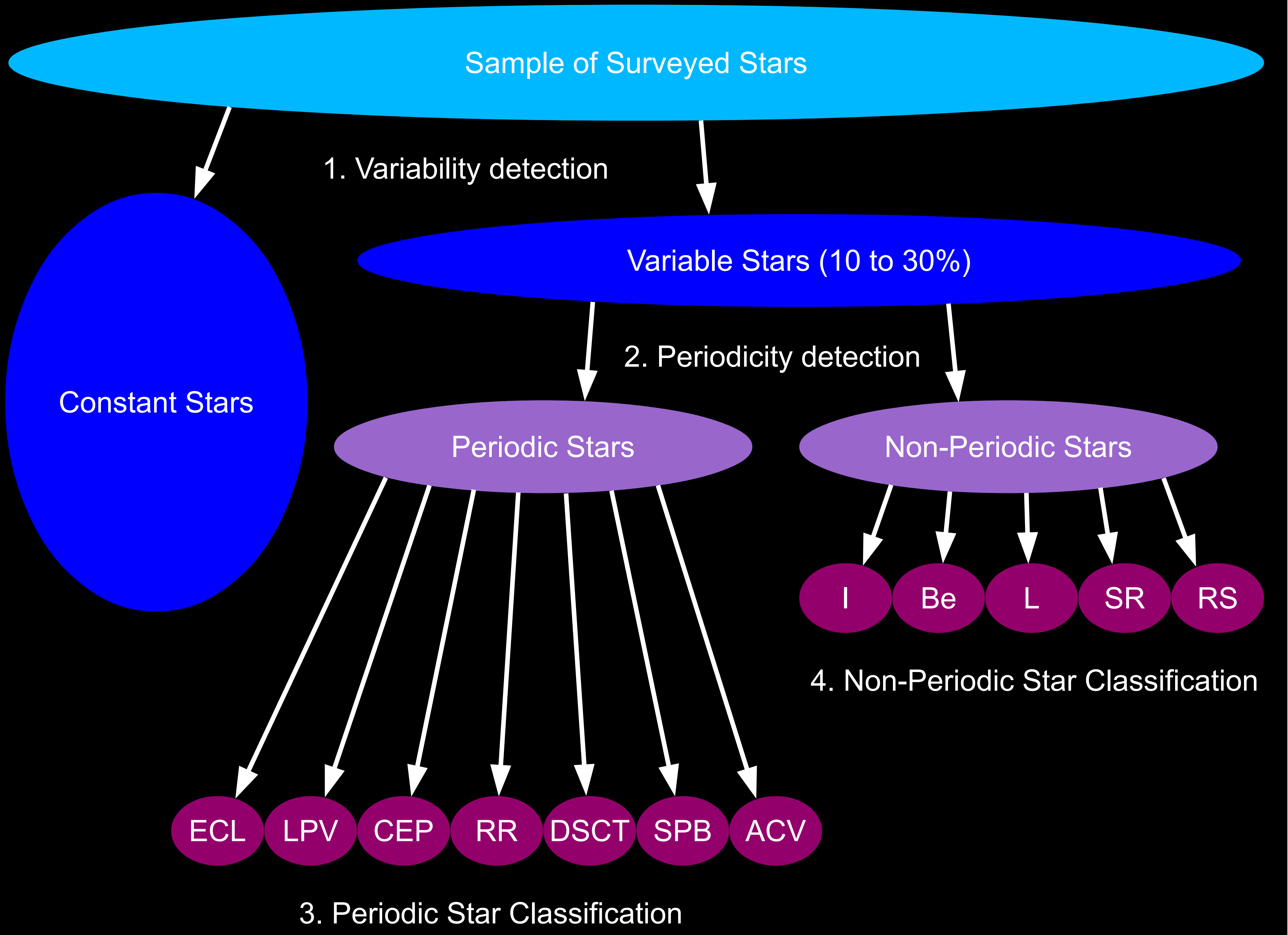
Periodic

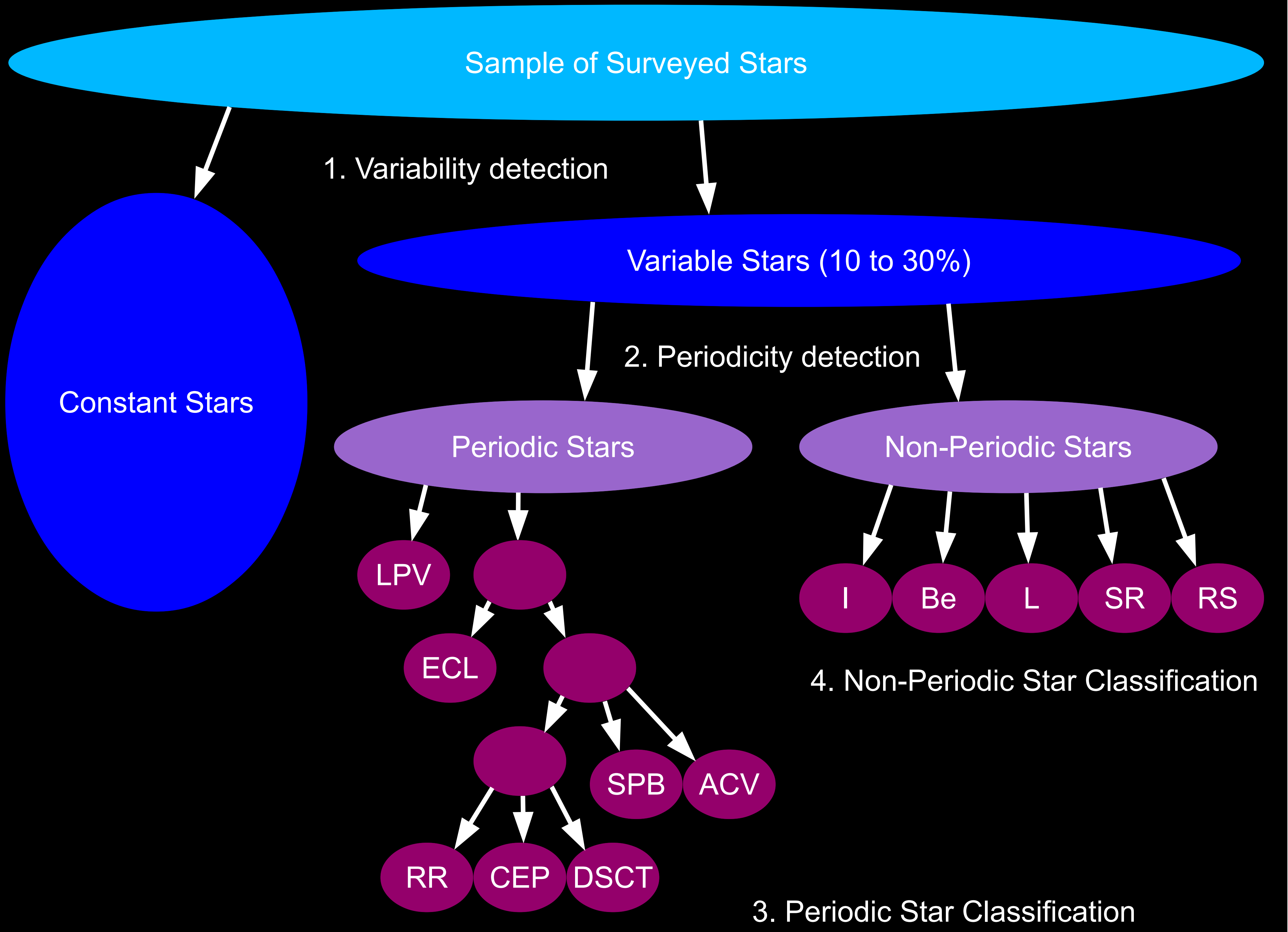
Non-Per

...

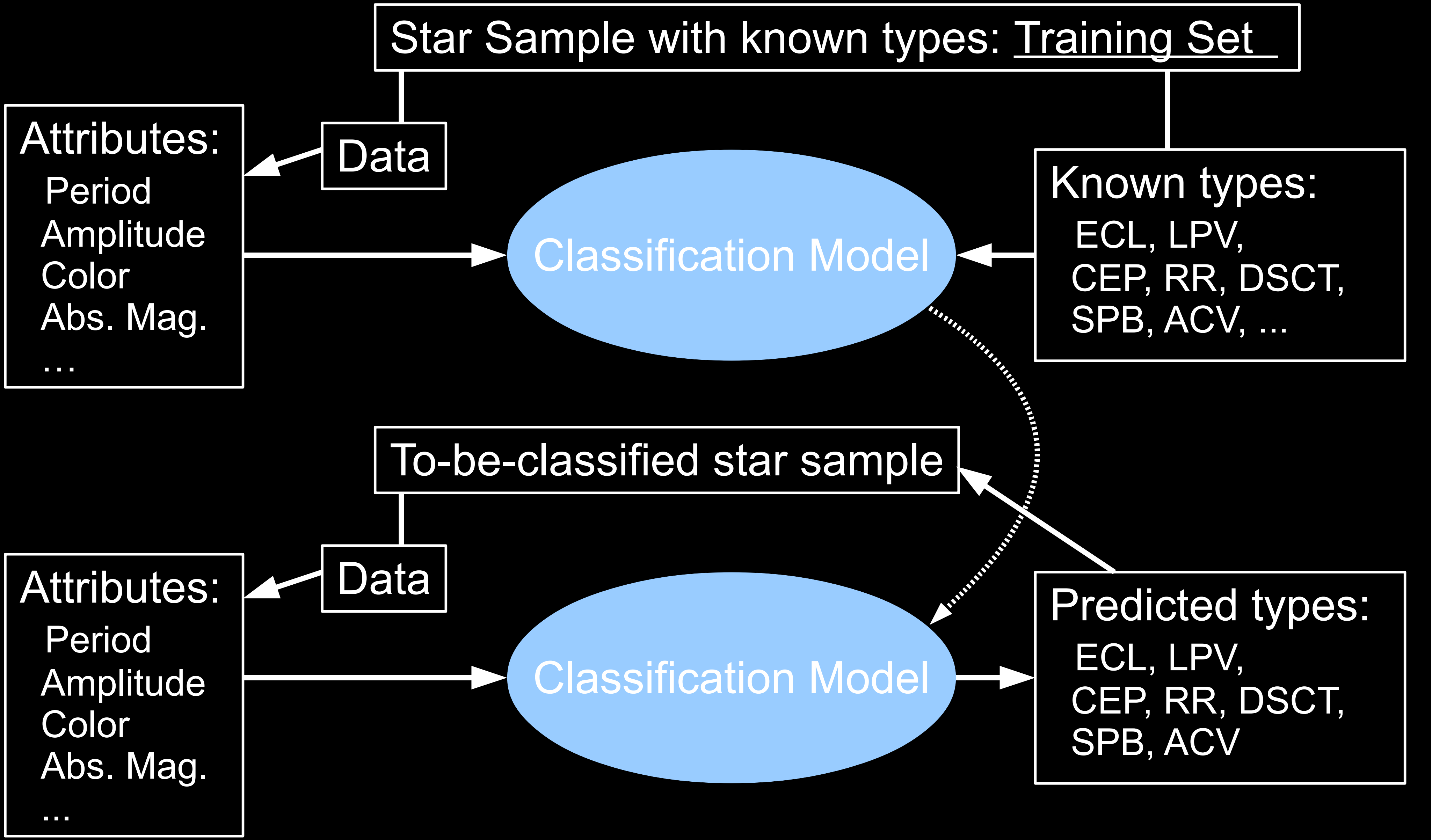
Periodic

Non-Per


Sample of Surveyed Stars
1. Variability detection
Constant Stars
Variable Stars (10 to 30%)
2. Periodicity detection
Periodic Stars
Non-Periodic Stars
I
Be
L
SR
RS
4. Non-Periodic Star Classification
ECL
LPV
CEP
RR
DSCT
SPB
ACV
3. Periodic Star Classification


Sample of Surveyed Stars
1. Variability detection
Constant Stars
Variable Stars (10 to 30%)
2. Periodicity detection
Periodic Stars
Non-Periodic Stars
LPV
ECL
SPB
ACV
RR
CEP
DSCT
I
Be
L
SR
RS
4. Non-Periodic Star Classification
3. Periodic Star Classification

Supervised Classification



Star Sample with known types: Training Set

Data

Attributes:
Period
Amplitude
Color
Abs. Mag.
…

Classification Model

Known types:
ECL, LPV,
CEP, RR, DSCT,
SPB, ACV, ...

To-be-classified star sample

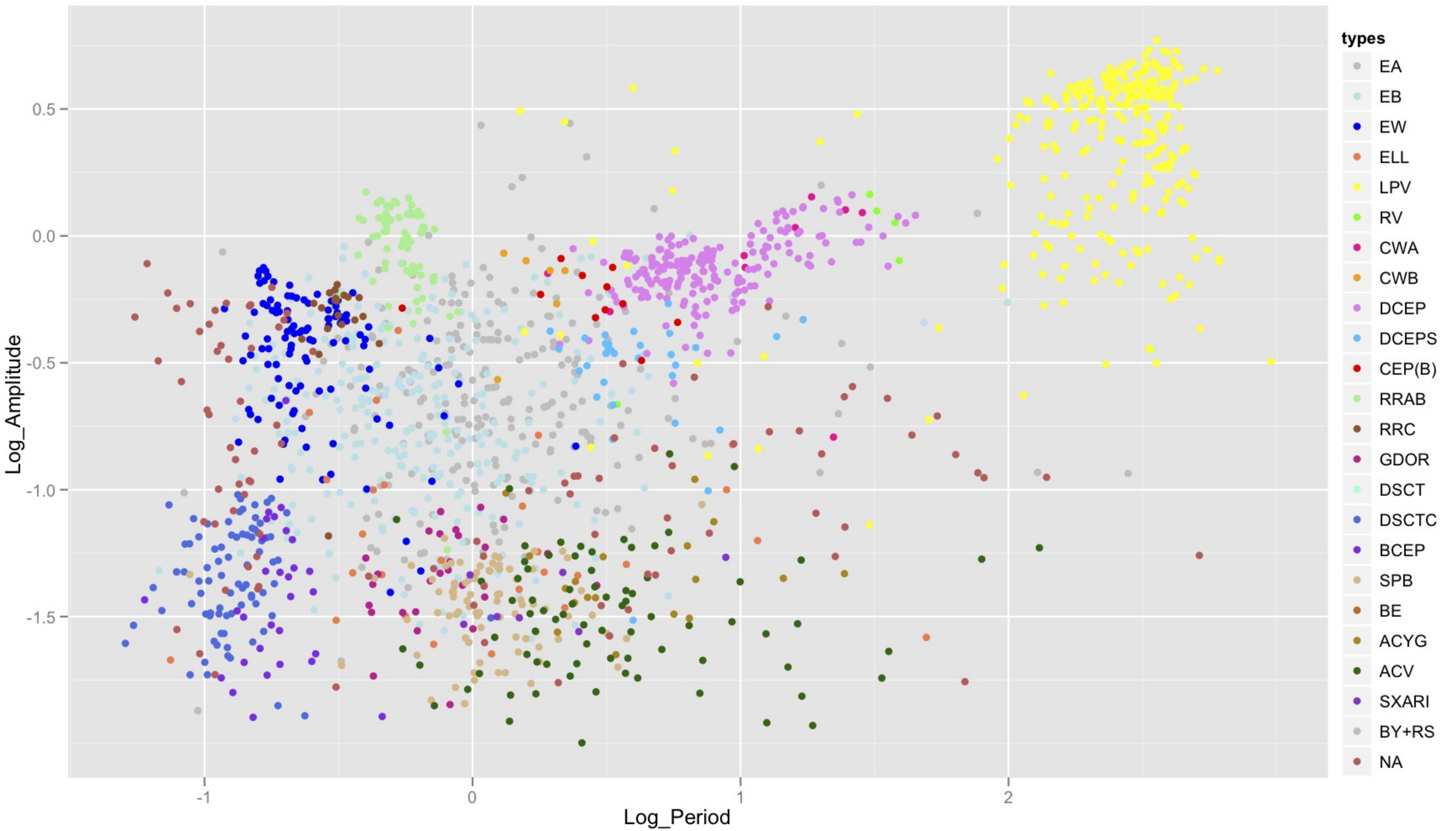
Data

Attributes:
Period
Amplitude
Color
Abs. Mag.
...

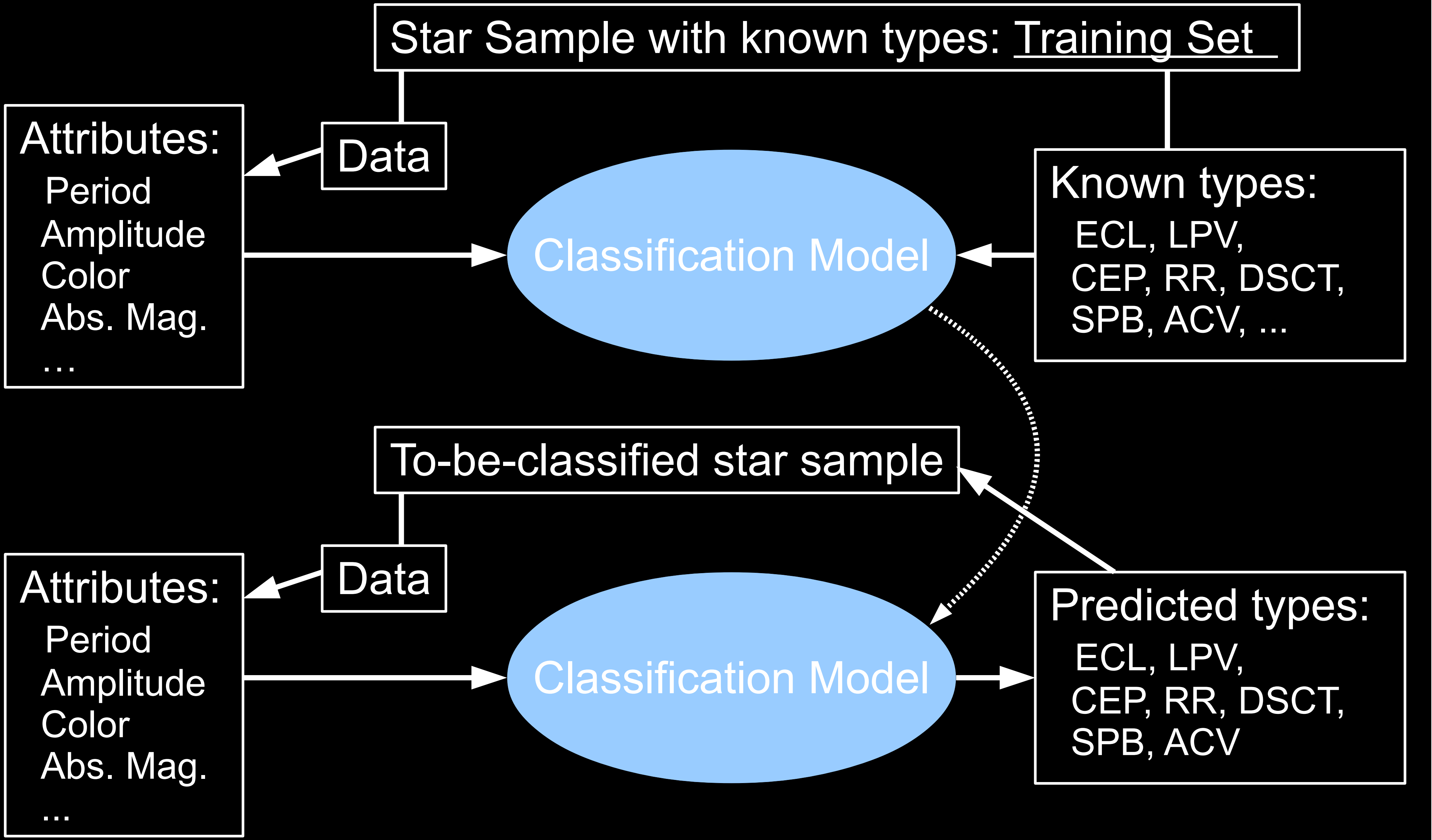
Classification Model

Predicted types:
ECL, LPV,
CEP, RR, DSCT,
SPB, ACV

Period-Amplitude diagram



Supervised Classification



Star Sample with known types: Training Set

Data

Attributes:
Period
Amplitude
Color
Abs. Mag.
…

Classification Model

Known types:
ECL, LPV,
CEP, RR, DSCT,
SPB, ACV, ...

To-be-classified star sample

Data

Attributes:
Period
Amplitude
Color
Abs. Mag.
...

Classification Model

Predicted types:
ECL, LPV,
CEP, RR, DSCT,
SPB, ACV

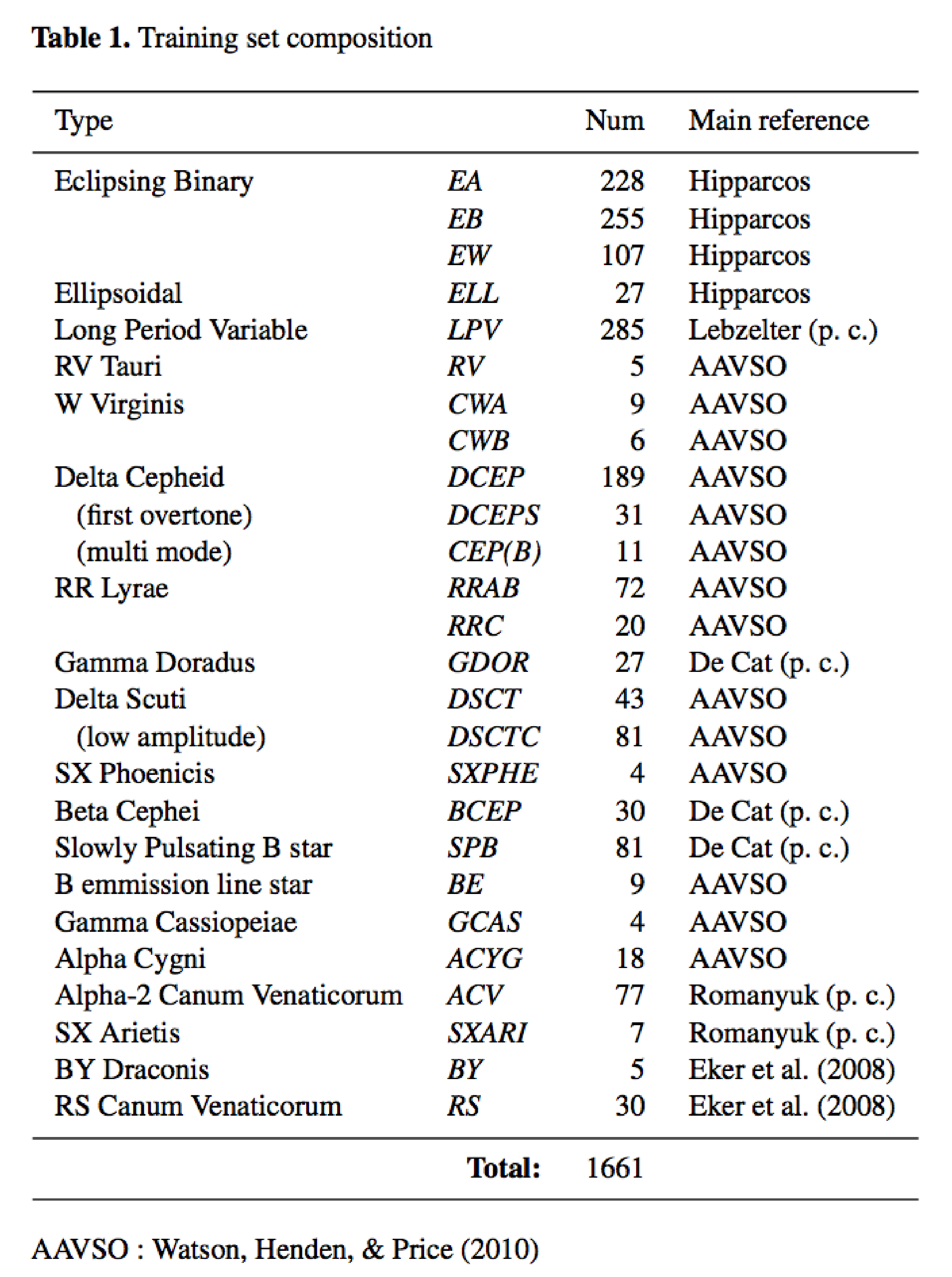
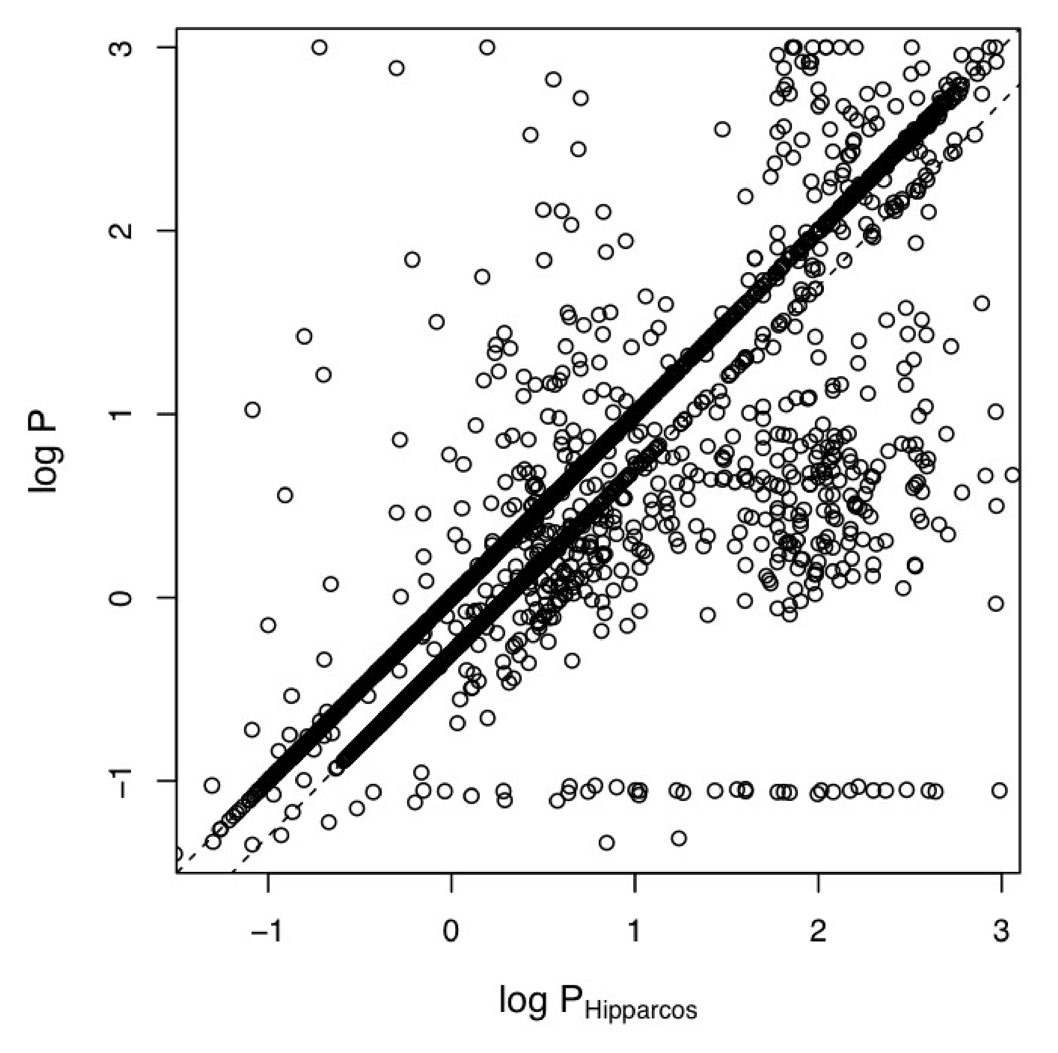


Table 1. Training set composition

Type		Num	Main reference
Eclipsing Binary	*EA*	228	Hipparcos
	EB	255	Hipparcos
	EW	107	Hipparcos
Ellipsoidal	*ELL*	27	Hipparcos
Long Period Variable	*LPV*	285	Lebzelter (p. c.)
RV Tauri	*RV*	5	AAVSO
W Virginis	*CWA*	9	AAVSO
	CWB	6	AAVSO
Delta Cepheid	*DCEP*	189	AAVSO
(first overtone)	*DCEPS*	31	AAVSO
(multi mode)	*CEP(B)*	11	AAVSO
RR Lyrae	*RRAB*	72	AAVSO
	RRC	20	AAVSO
Gamma Doradus	*GDOR*	27	De Cat (p. c.)
Delta Scuti	*DSCT*	43	AAVSO
(low amplitude)	*DSCTC*	81	AAVSO
SX Phoenicis	*SXPHE*	4	AAVSO
Beta Cephei	*BCEP*	30	De Cat (p. c.)
Slowly Pulsating B star	*SPB*	81	De Cat (p. c.)
B emmission line star	*BE*	9	AAVSO
Gamma Cassiopeiae	*GCAS*	4	AAVSO
Alpha Cygni	*ACYG*	18	AAVSO
Alpha-2 Canum Venaticorum	*ACV*	77	Romanyuk (p. c.)
SX Arietis	*SXARI*	7	Romanyuk (p. c.)
BY Draconis	*BY*	5	Eker et al. (2008)
RS Canum Venaticorum	*RS*	30	Eker et al. (2008)
	Total:	1661	

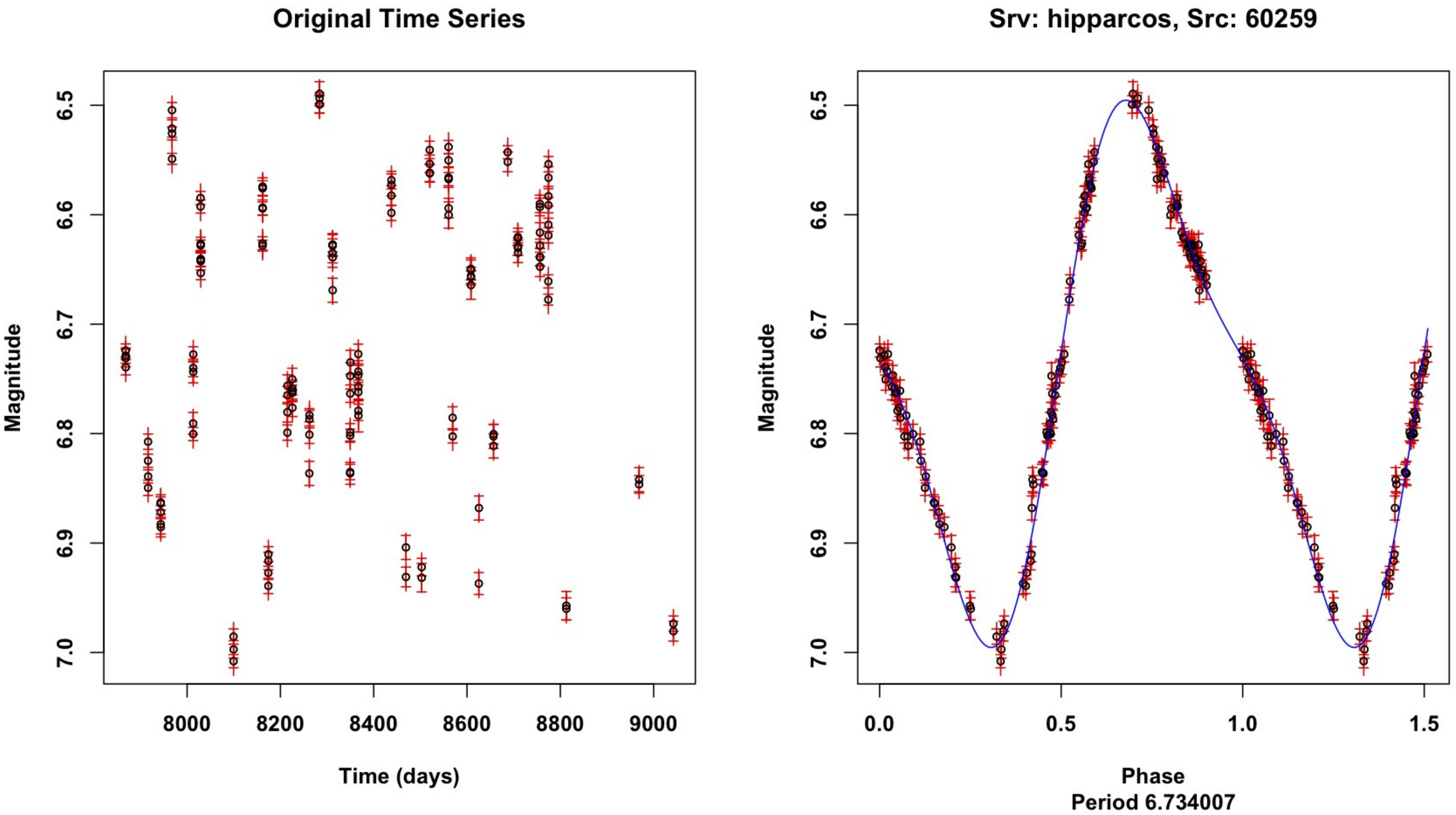
AAVSO : Watson, Henden, & Price (2010)

Lomb-Scargle period search

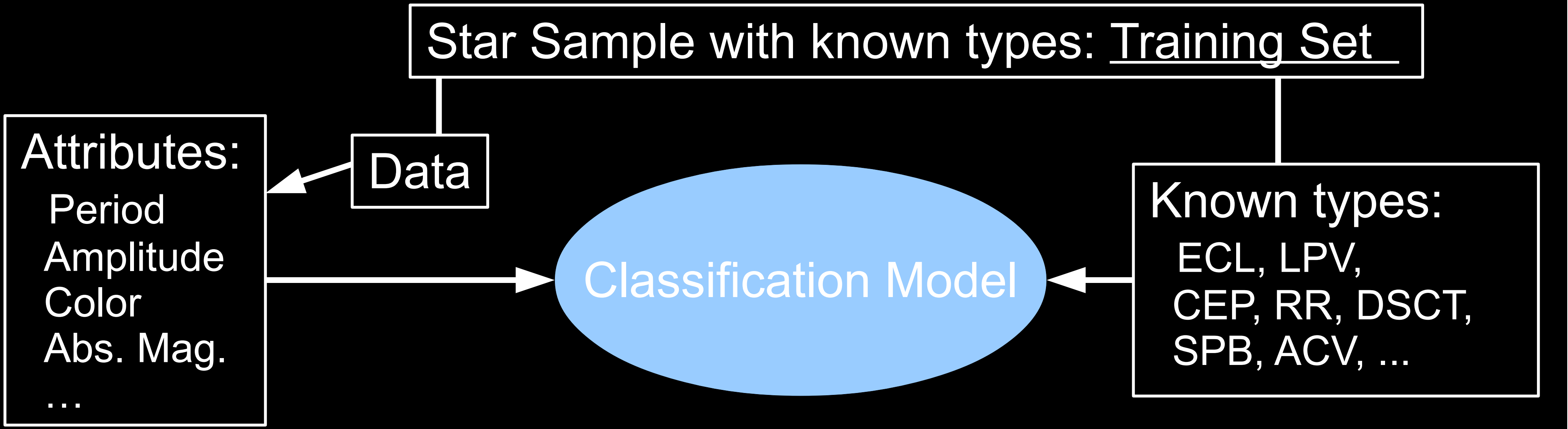


- (Zechmeister & Kűrster 2009)

Fourier series modeling



Supervised Classification



Star Sample with known types: Training Set

Data

Attributes:

Period
Amplitude
Color
Abs. Mag.
…

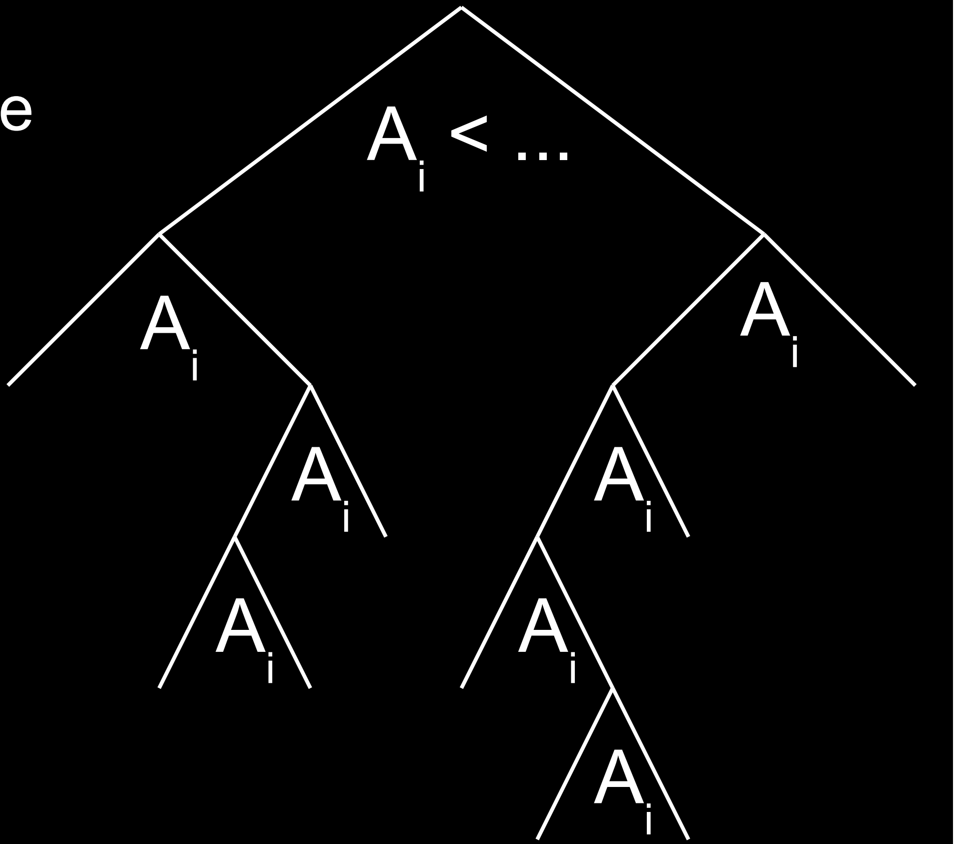
Classification Model

Known types:

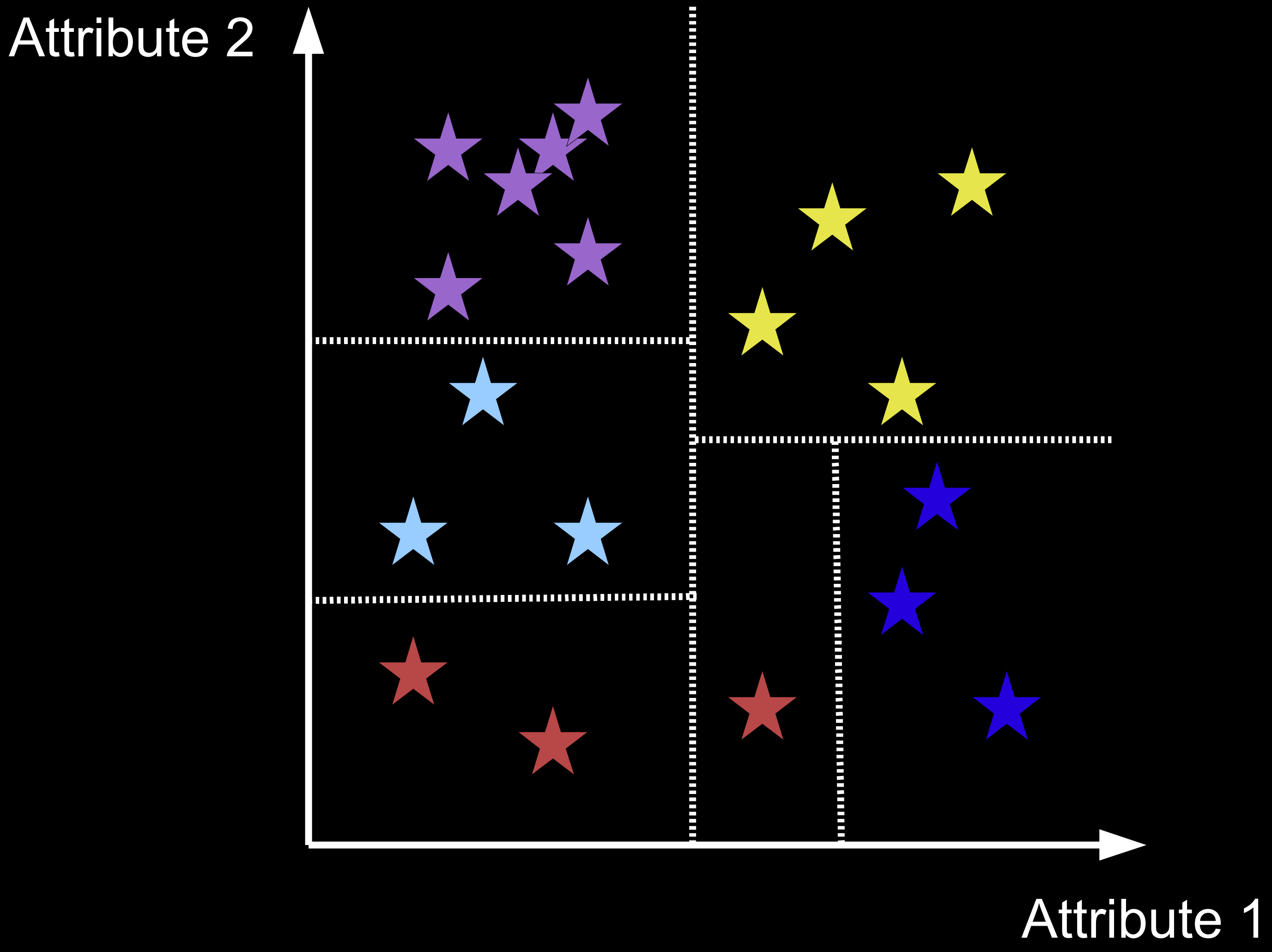
ECL, LPV,
CEP, RR, DSCT,
SPB, ACV, ...

Random Forest (1/3)

- Classification trees
- Binary partitions using one attribute
- Each split minimize impurity

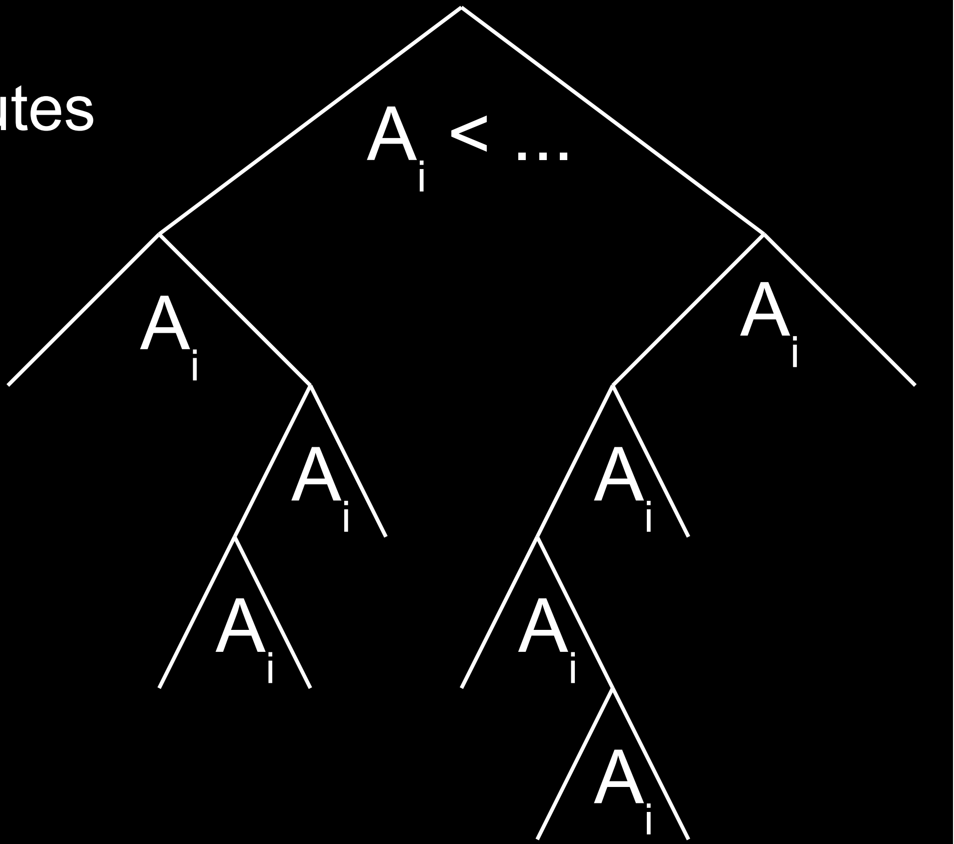


Random Forest (2/3)

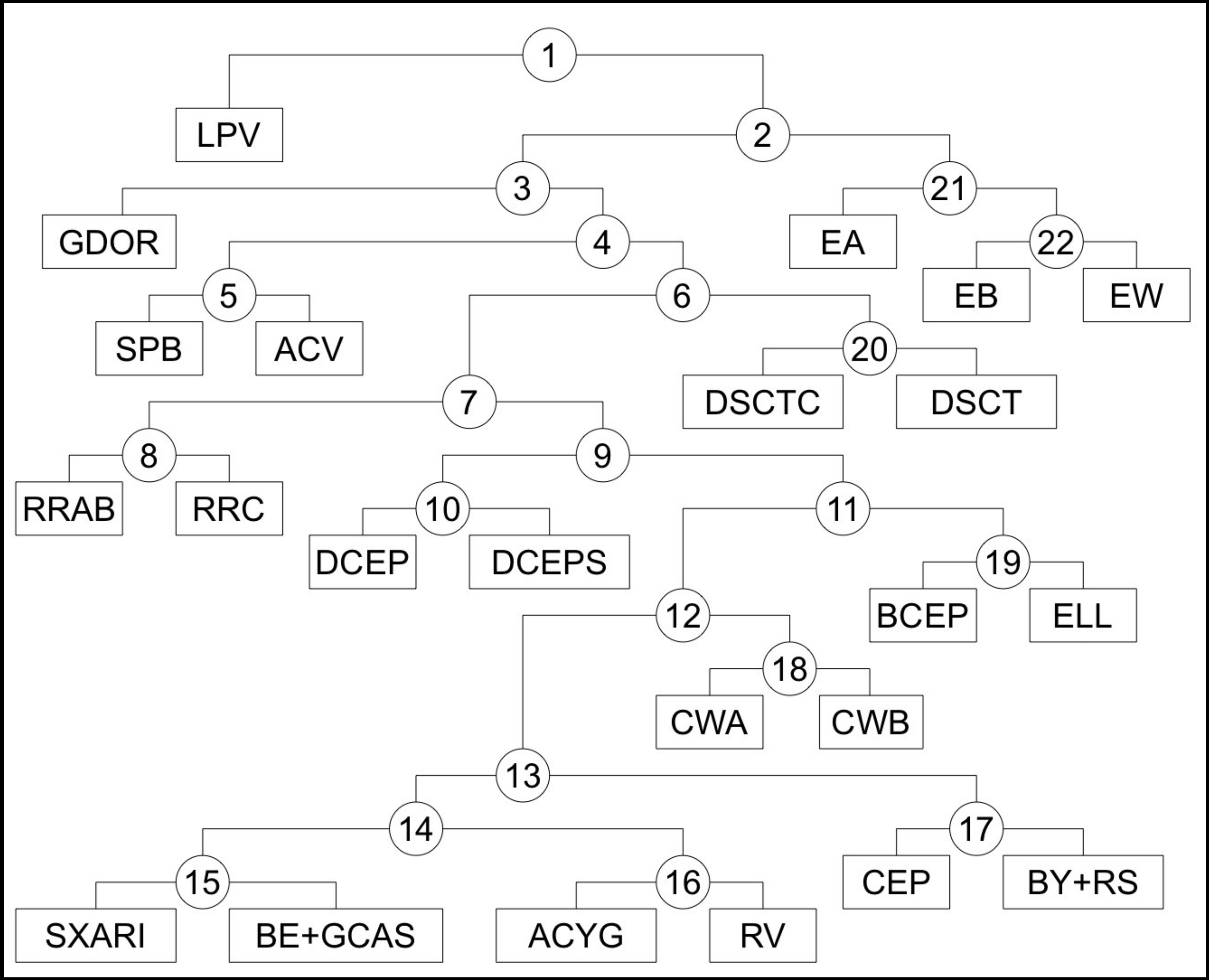


Random Forest (3/3)

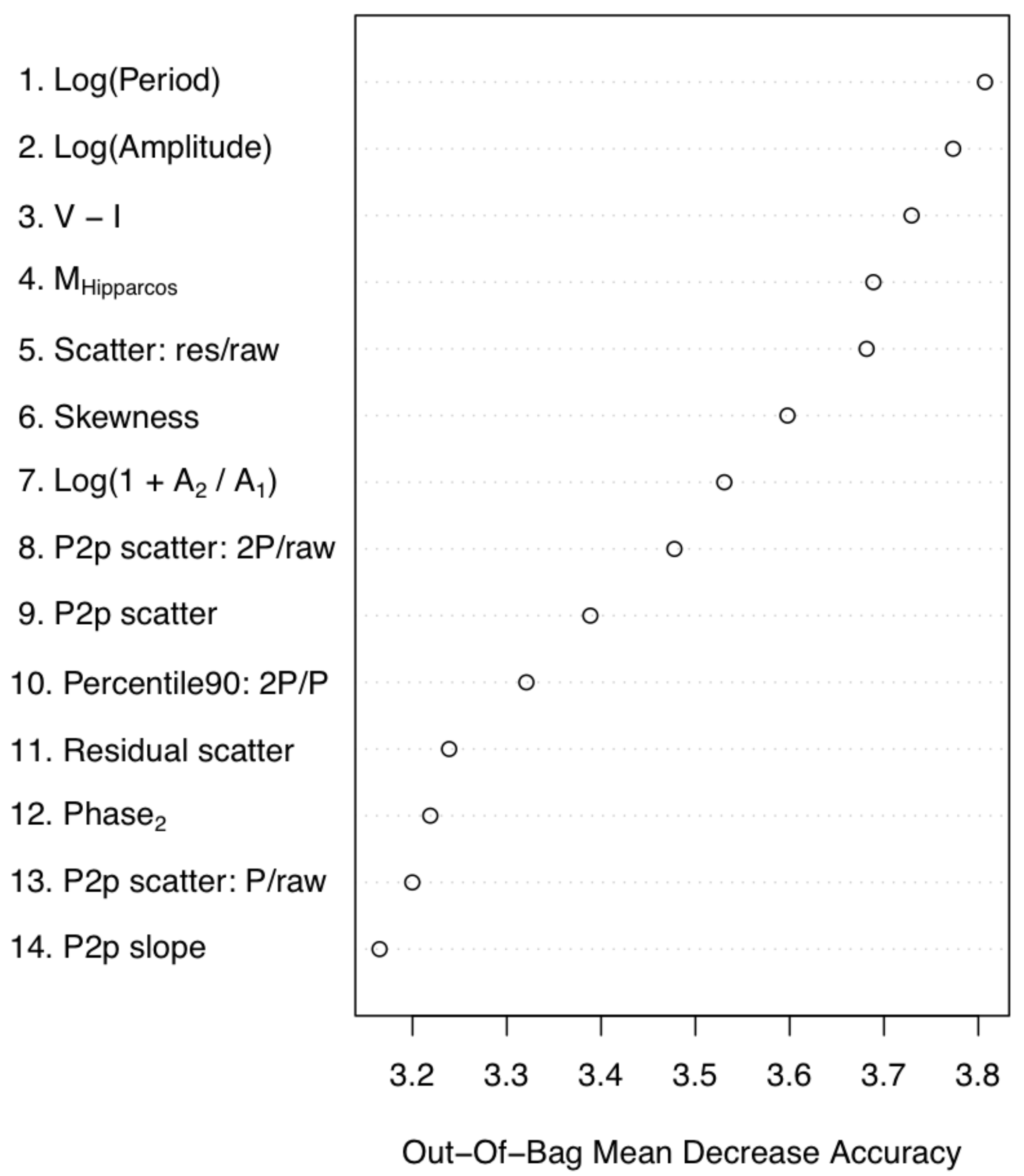
- Bootstrapping
- A_i from a random sub-set of attributes
- Average many trees



Multistage classification

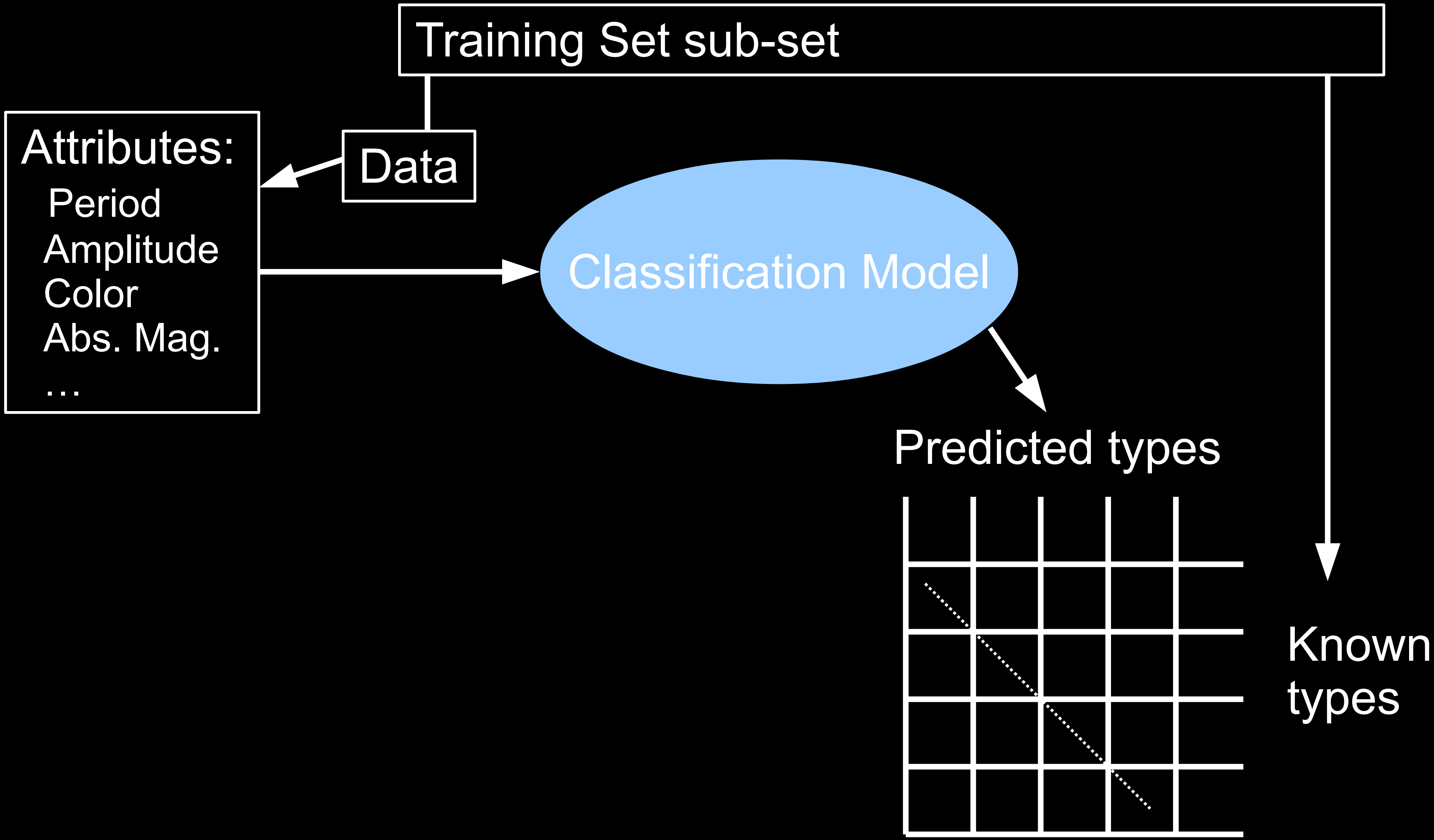


Random Forest Attribute Importance



- Attributes with Spearman correlation larger than 80% are trimmed

Cross-Validation



Training Set sub-set

Data

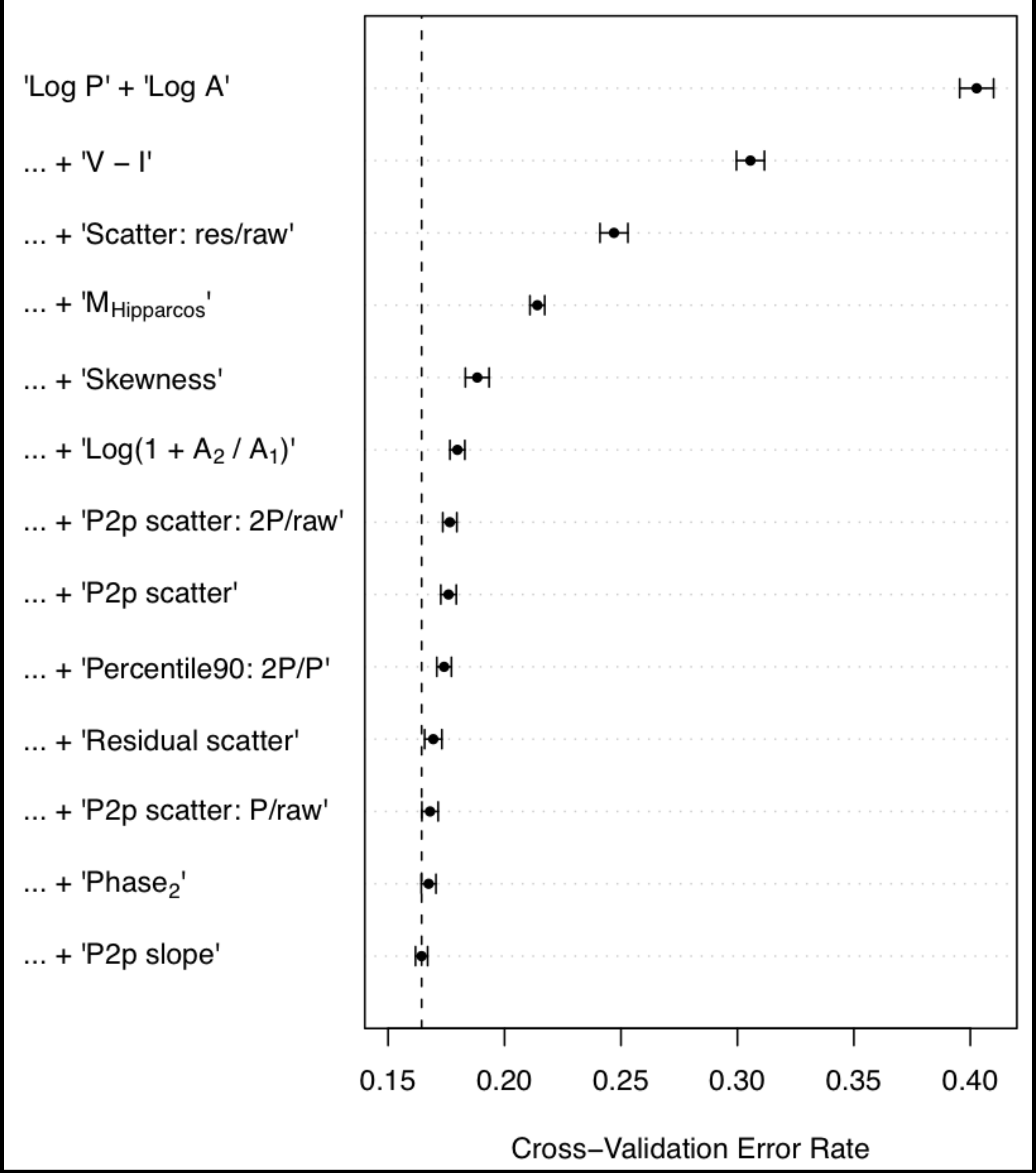
Attributes:
Period
Amplitude
Color
Abs. Mag.
…

Classification Model

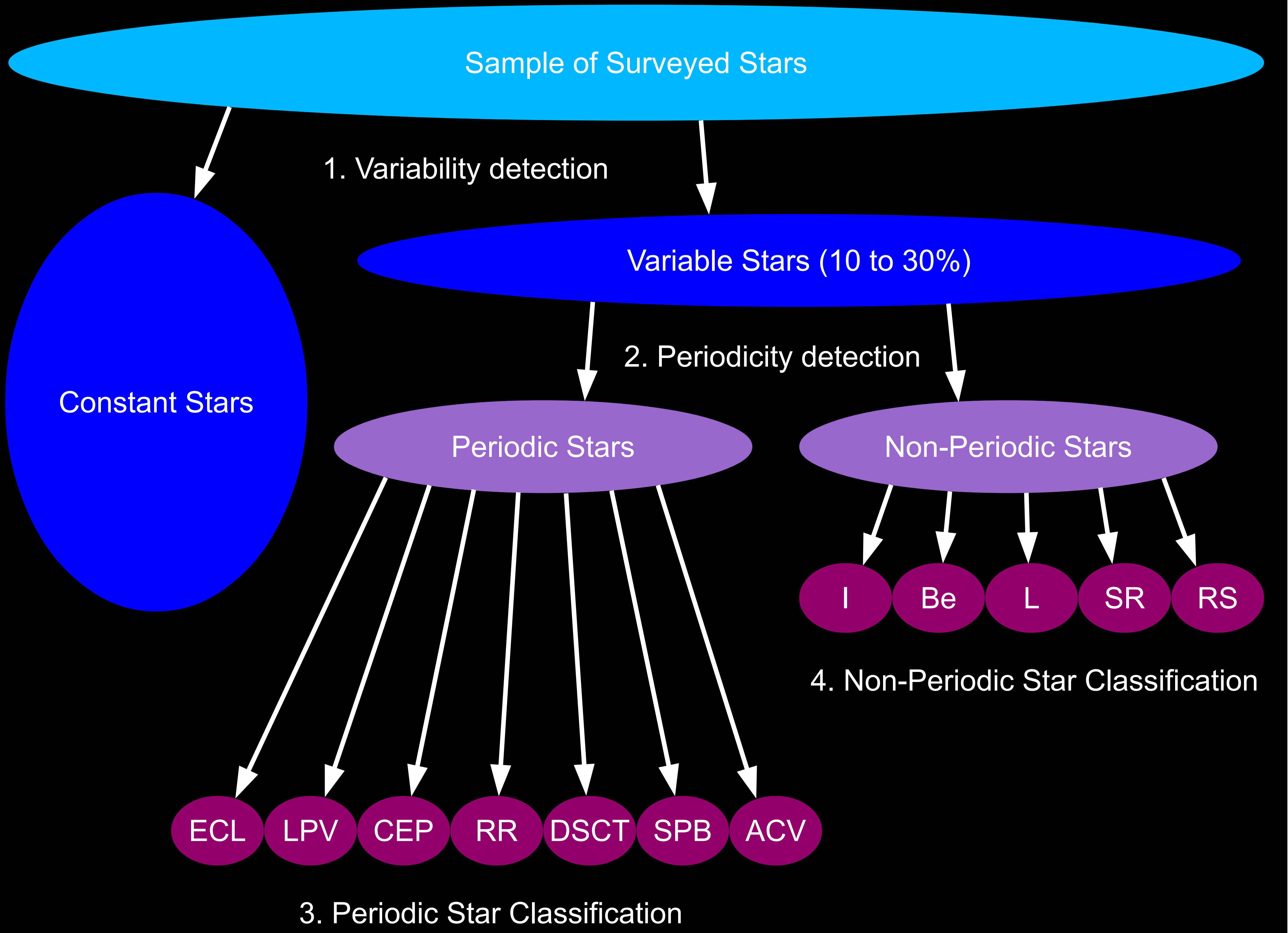
Predicted types

Known
types

Random Forest CV error rates



EA	EB	EW	ELL	LPV	RV	CWA	CWB	DCEP	DCEPS	CEP(B)	RRAB	RRC	GDOR	DSCT	DSCTC	BCEP	SPB	BE+GCAS	ACYG	ACV	SXARI	BY+RS	
214	13									1													EA
19	191	28	2	1				2					1		4		3		2	2			EB
	30	76							1														EW
	14			1									1		1		3			5		2	ELL
				285																			LPV
	1			1				2	1														RV
	2				1			5														1	CWA
	1						2	2	1														CWB
								183	5	1													DCEP
	1							11	17													2	DCEPS
	1							4		6													CEP(B)
	1										69	1					1						RRAB
	2	4									1	12		1									RRC
													27										GDOR
	1	1									1			32	12								DSCT
	1													1	77					2			DSCTC
	1	1													1	26	1						BCEP
			1													1	74		1	4			SPB
1									1								5		2	4			BE+GCAS
	1																	1	13	2		1	ACYG
	3								1				1				6			66			ACV
	2																2			3			SXARI
	1							1														33	BY+RS



Sample of Surveyed Stars

1. Variability detection

Constant Stars

Variable Stars (10 to 30%)

2. Periodicity detection

Periodic Stars

Non-Periodic Stars

I

Be

L

SR

RS

4. Non-Periodic Star Classification

ECL

LPV

CEP

RR

DSCT

SPB

ACV

3. Periodic Star Classification

Variability detection

- Variability criteria

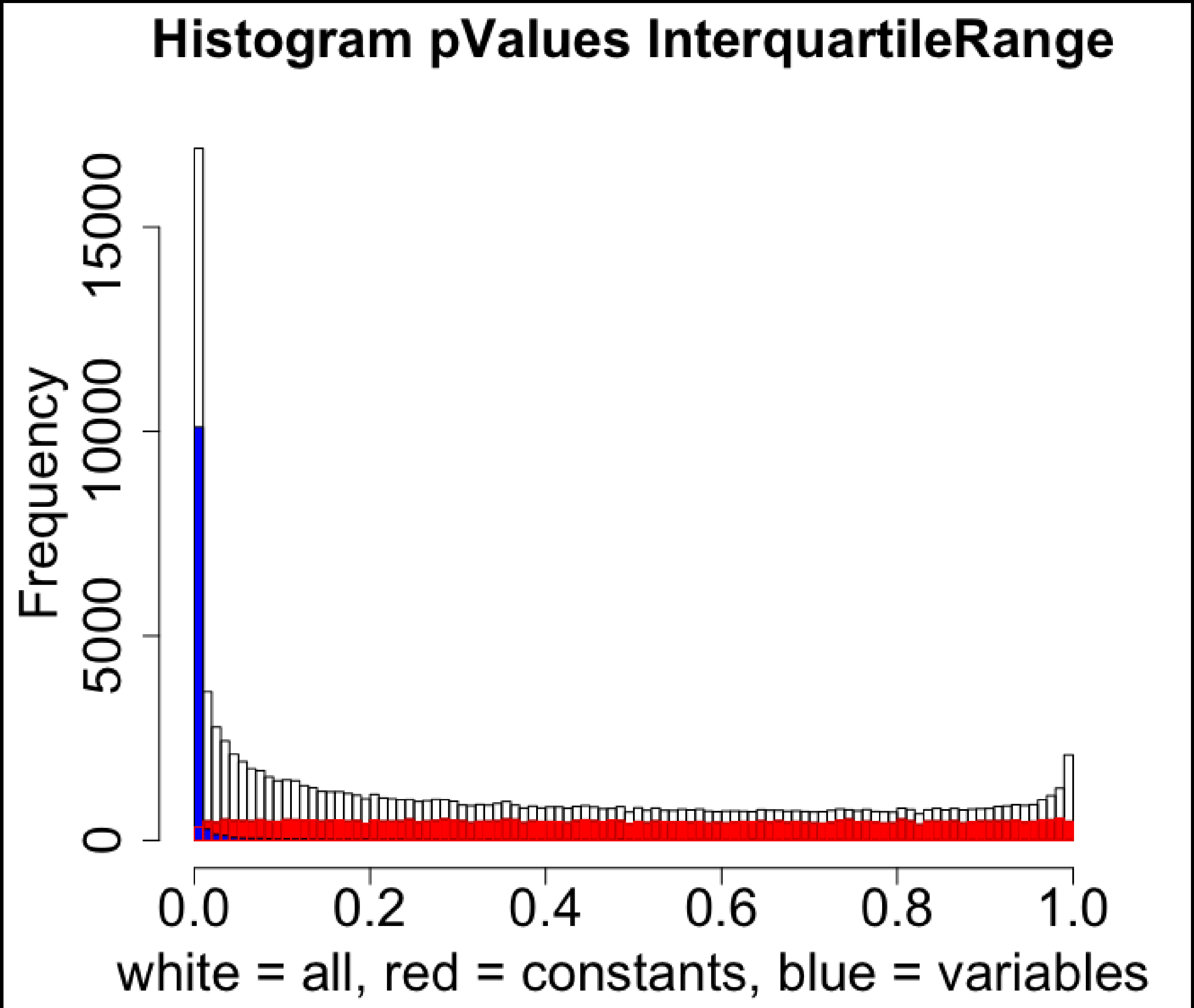
$$\chi^2 = \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma^2} \right)^2$$

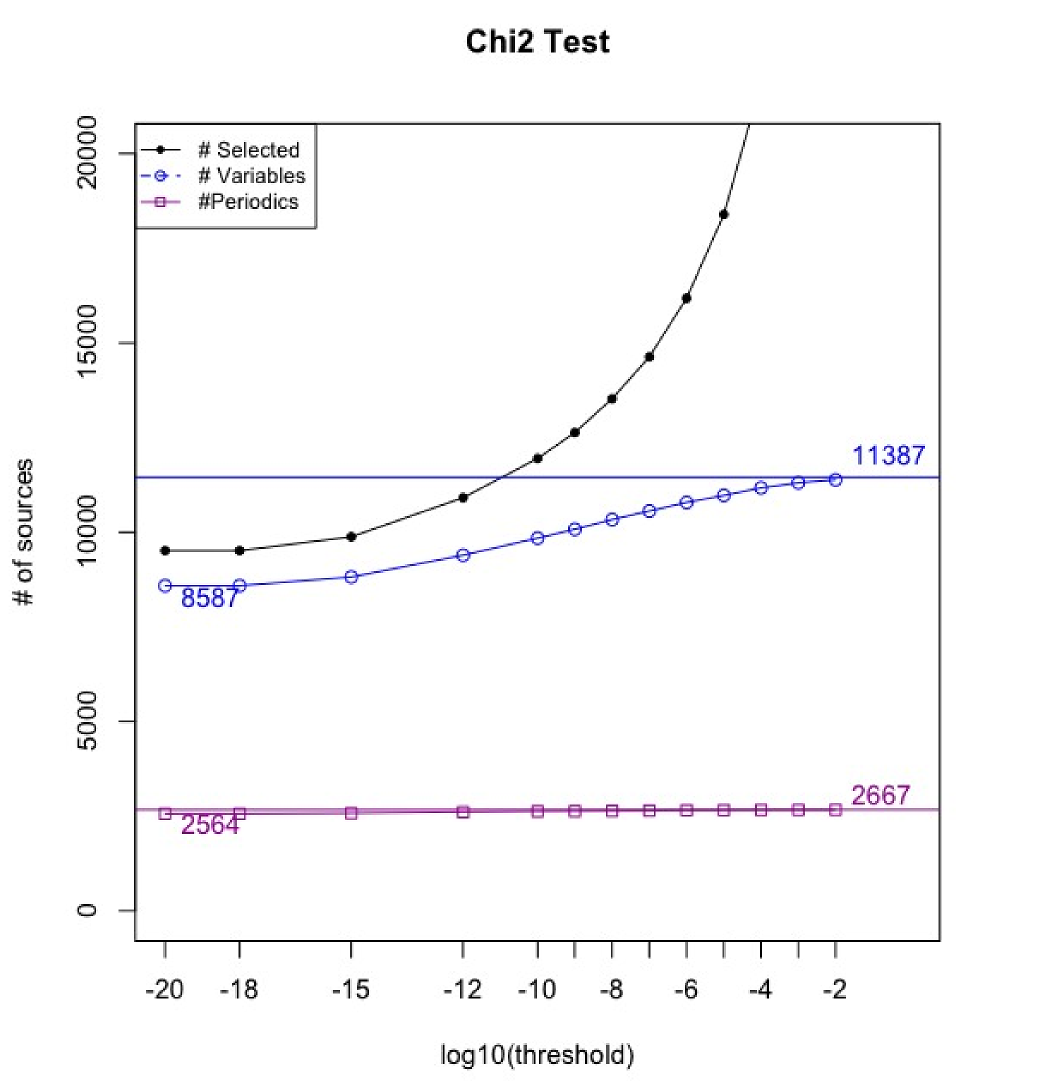
Using errors	Not using errors
Chi square	Abbe
Skewness	Skewness
Kurtosis	Kurtosis
Stetson	Inter-quartile range
Outlier median	B/R Correlation

$$Abbe = \frac{1}{2} \frac{\sum_{i=1}^{n-1} \left(x_{i+1} - x_i \right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}$$

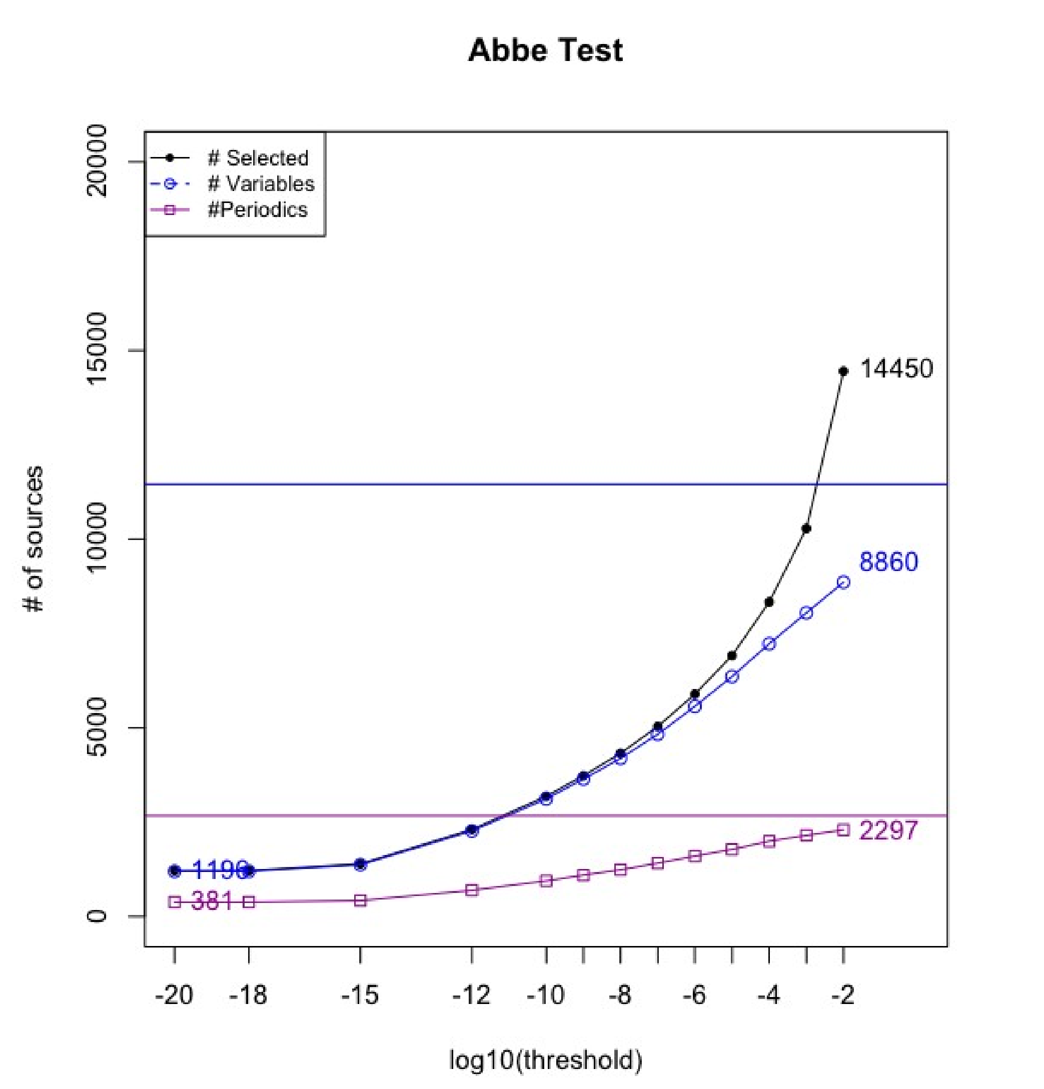
- Compute pValues: probability of the null hypothesis H_0 = constant star
  - If pValue < 1e-4 → variable star

Inter-quartile range Hypparcos pValues

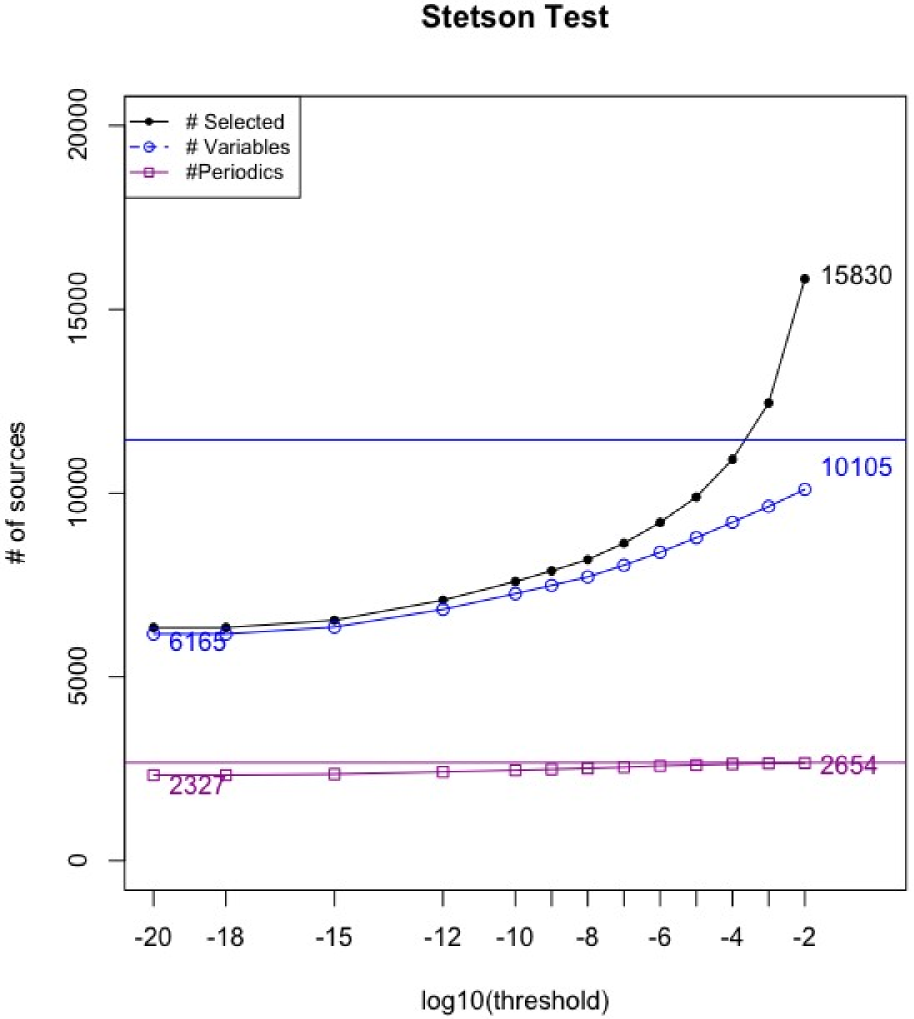


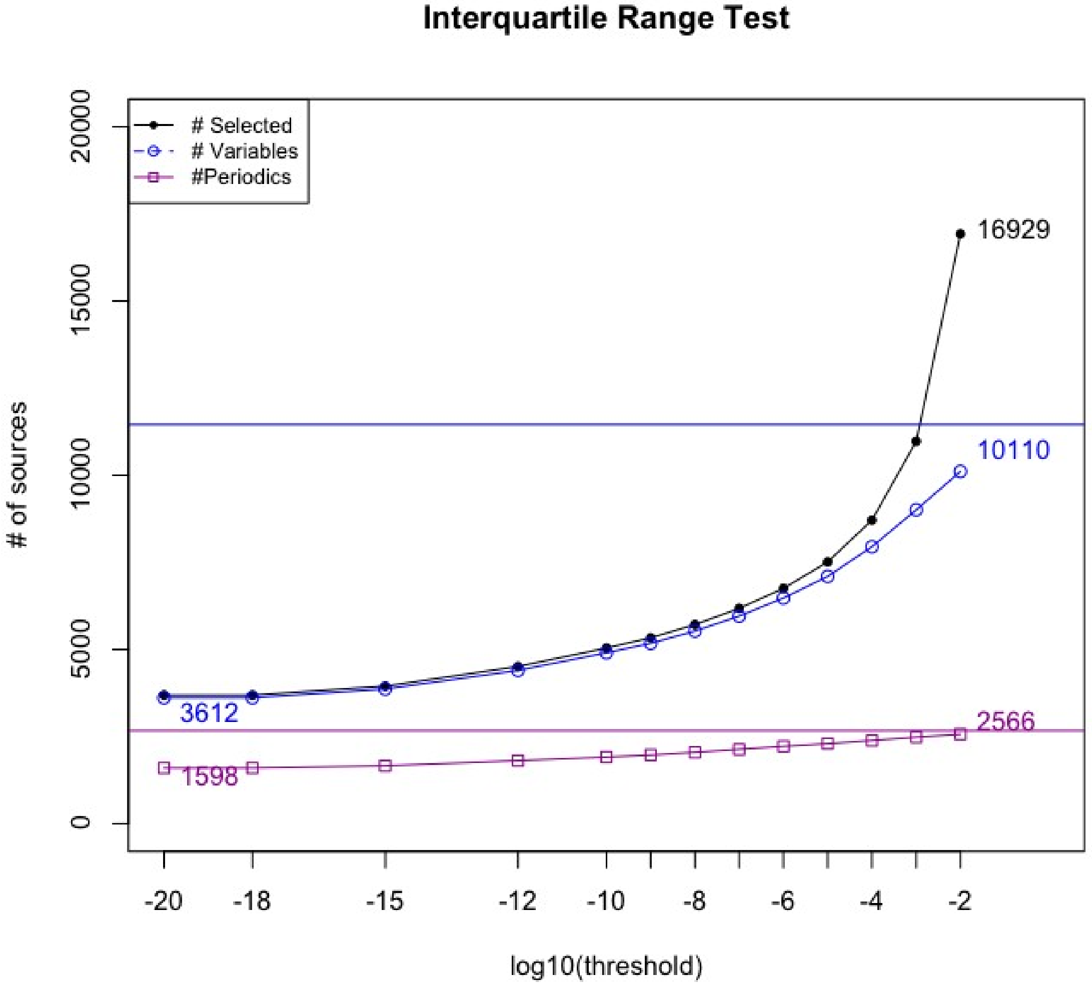


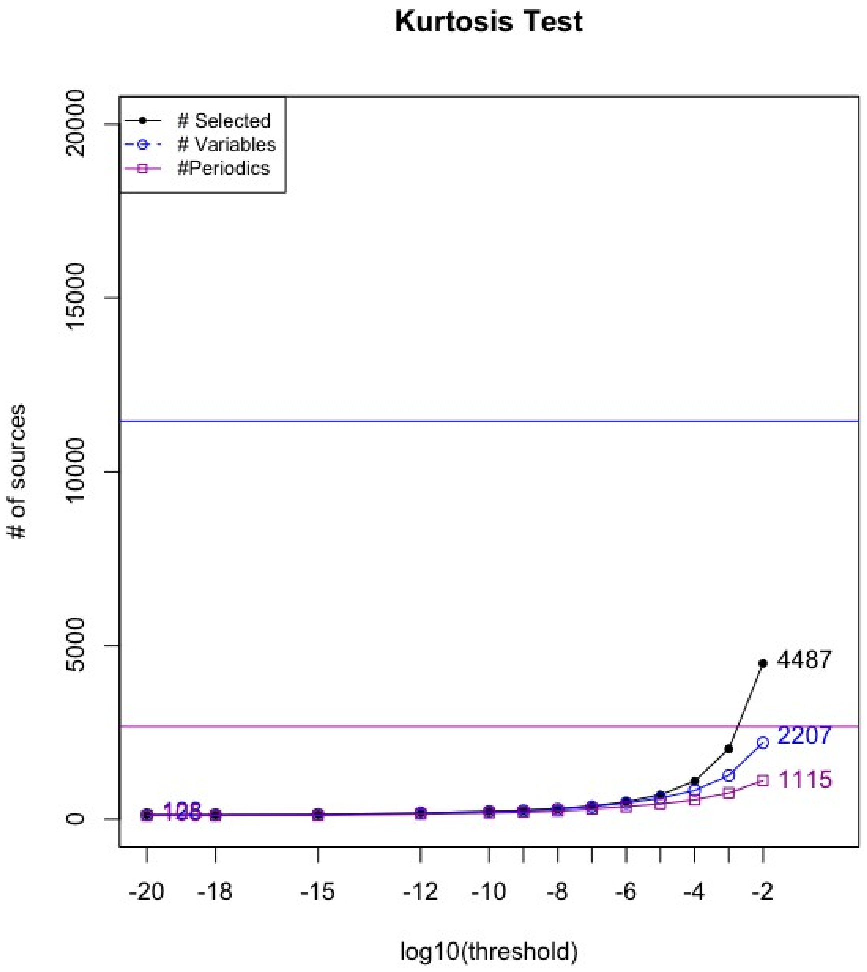
Chi2 Test

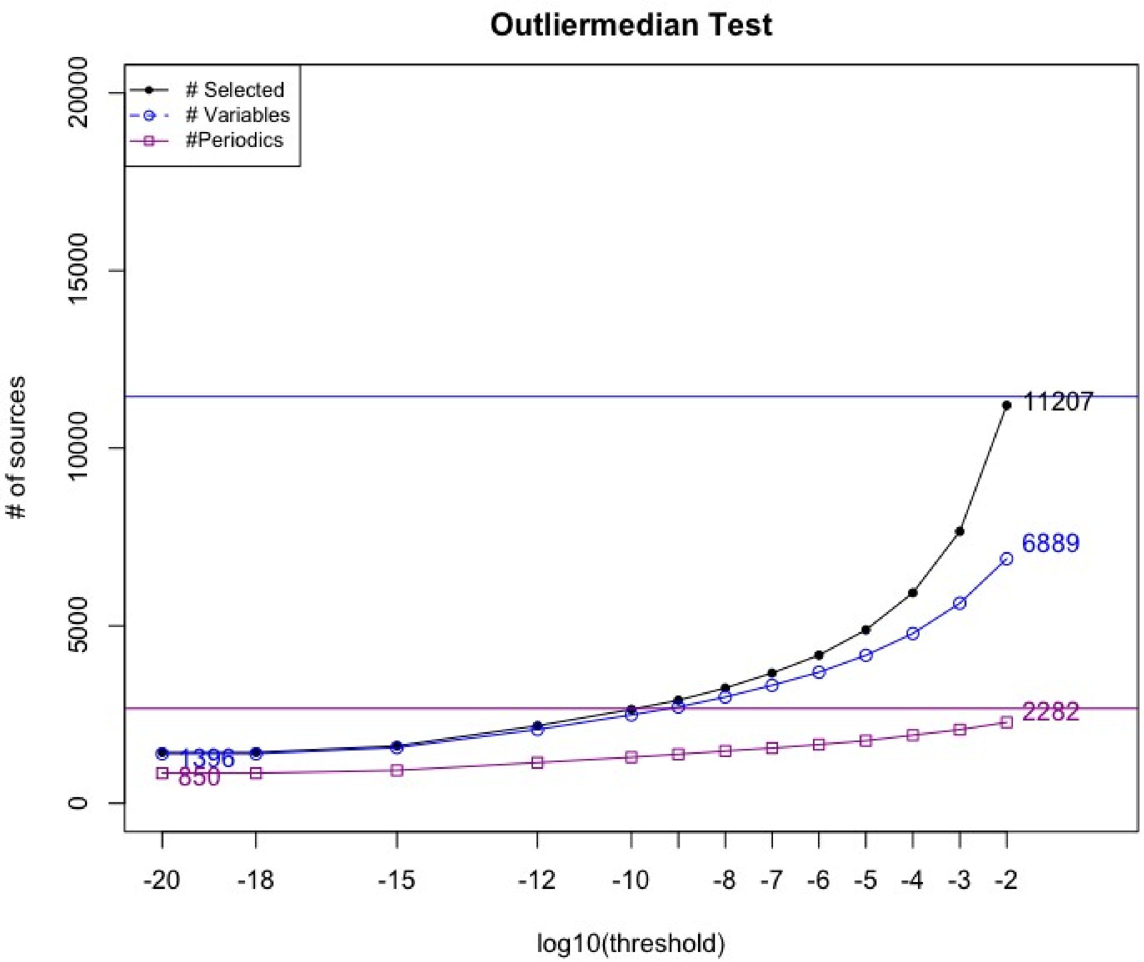


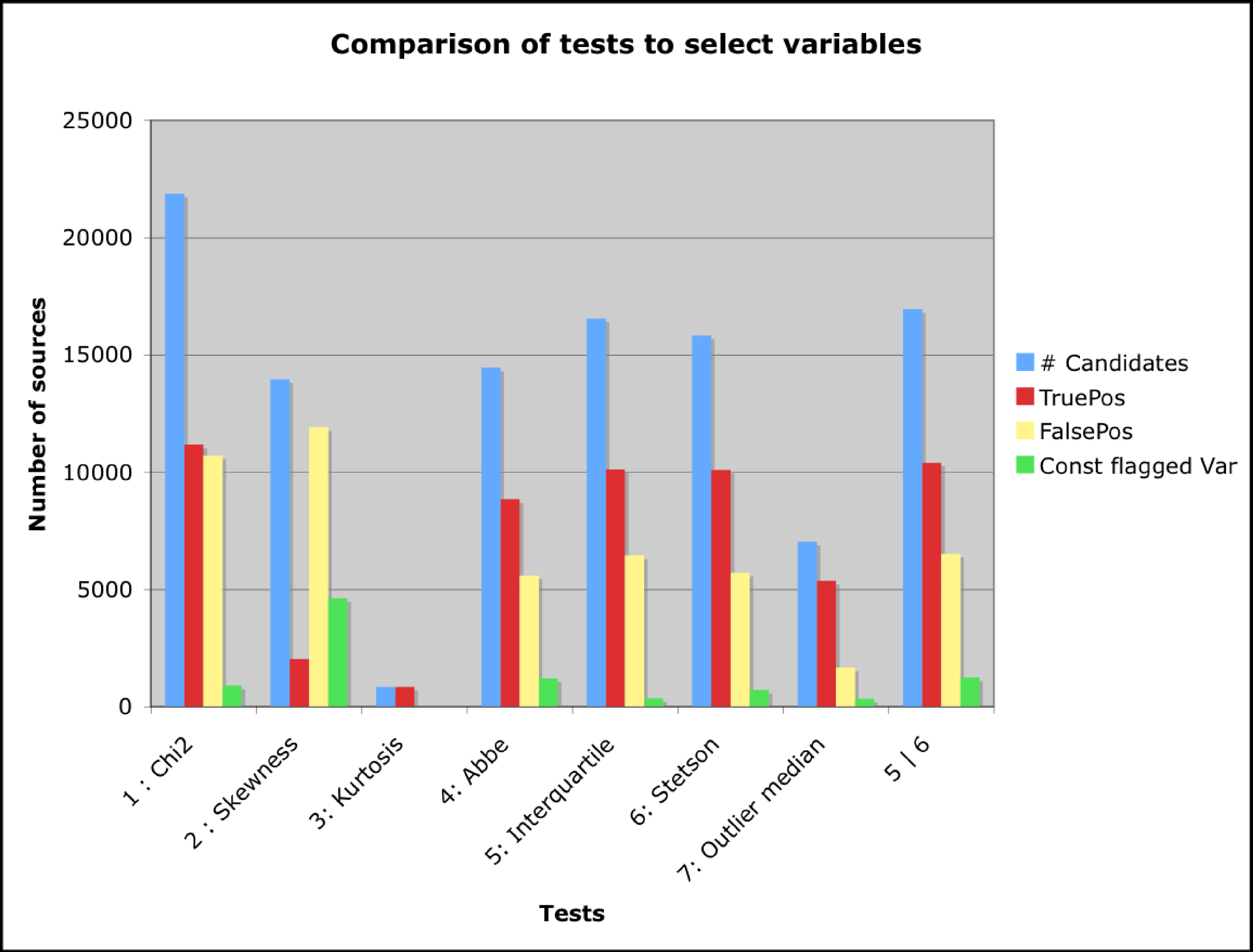
Abbe Test


Stetson Test
# Selected
# Variables
#Periodics
# of sources
0
5000
10000
15000
20000
15830
10105
6165
2327
2654
-20
-18
-15
-12
-10
-8
-6
-4
-2
log10(threshold)

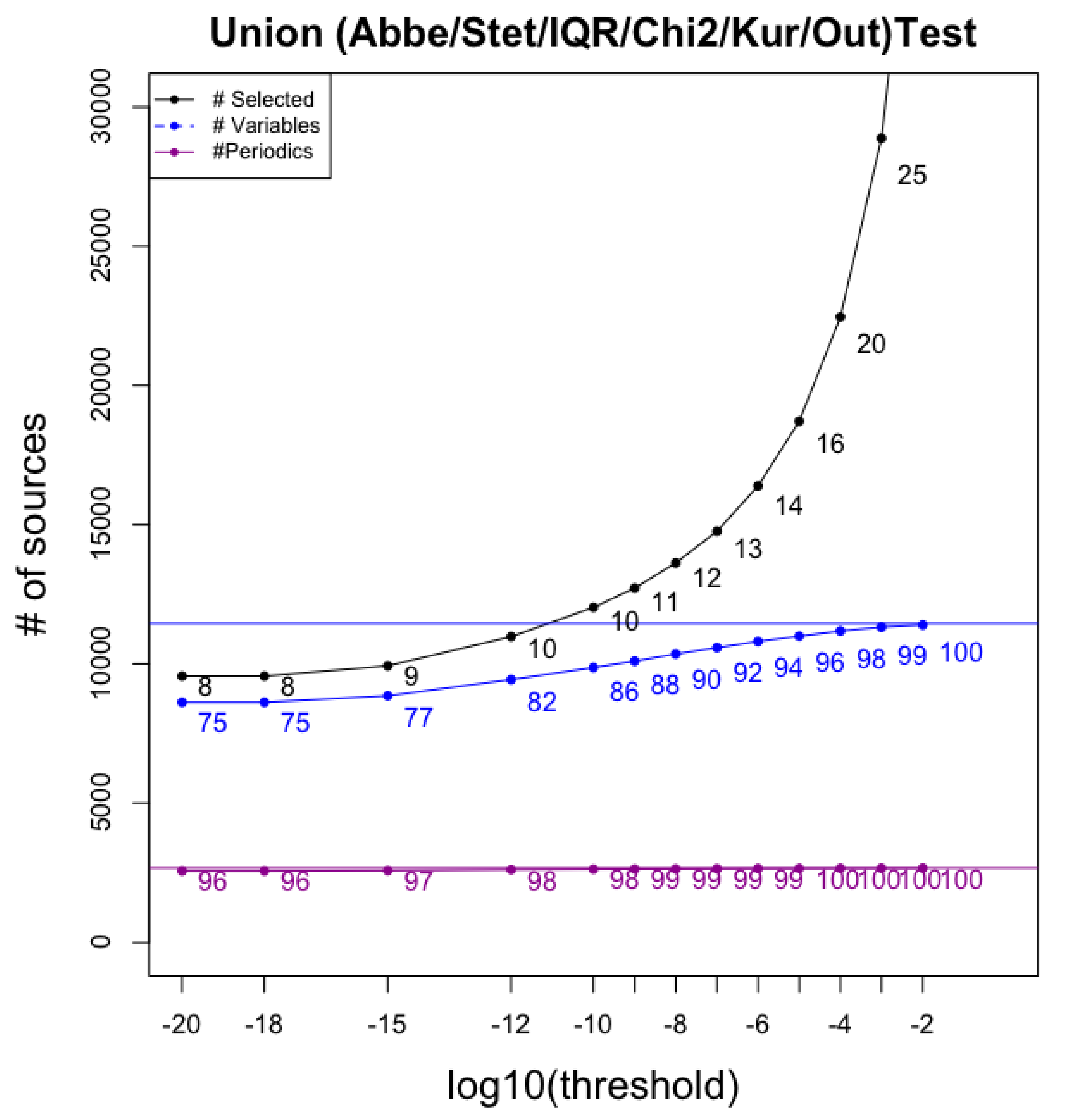

Interquartile Range Test
# Selected
# Variables
#Periodics
# of sources
0
5000
10000
15000
20000
16929
10110
3612
1598
2566
-20
-18
-15
-12
-10
-8
-6
-4
-2
log10(threshold)


Kurtosis Test
# Selected
# Variables
#Periodics
# of sources
0
5000
10000
15000
20000
4487
2207
1115
108
-20
-18
-15
-12
-10
-8
-6
-4
-2
log10(threshold)


Outliermedian Test
# Selected
# Variables
#Periodics
# of sources
0
5000
10000
15000
20000
11207
6889
2282
1396
850
-20
-18
-15
-12
-10
-8
-7
-6
-5
-4
-3
-2
log10(threshold)

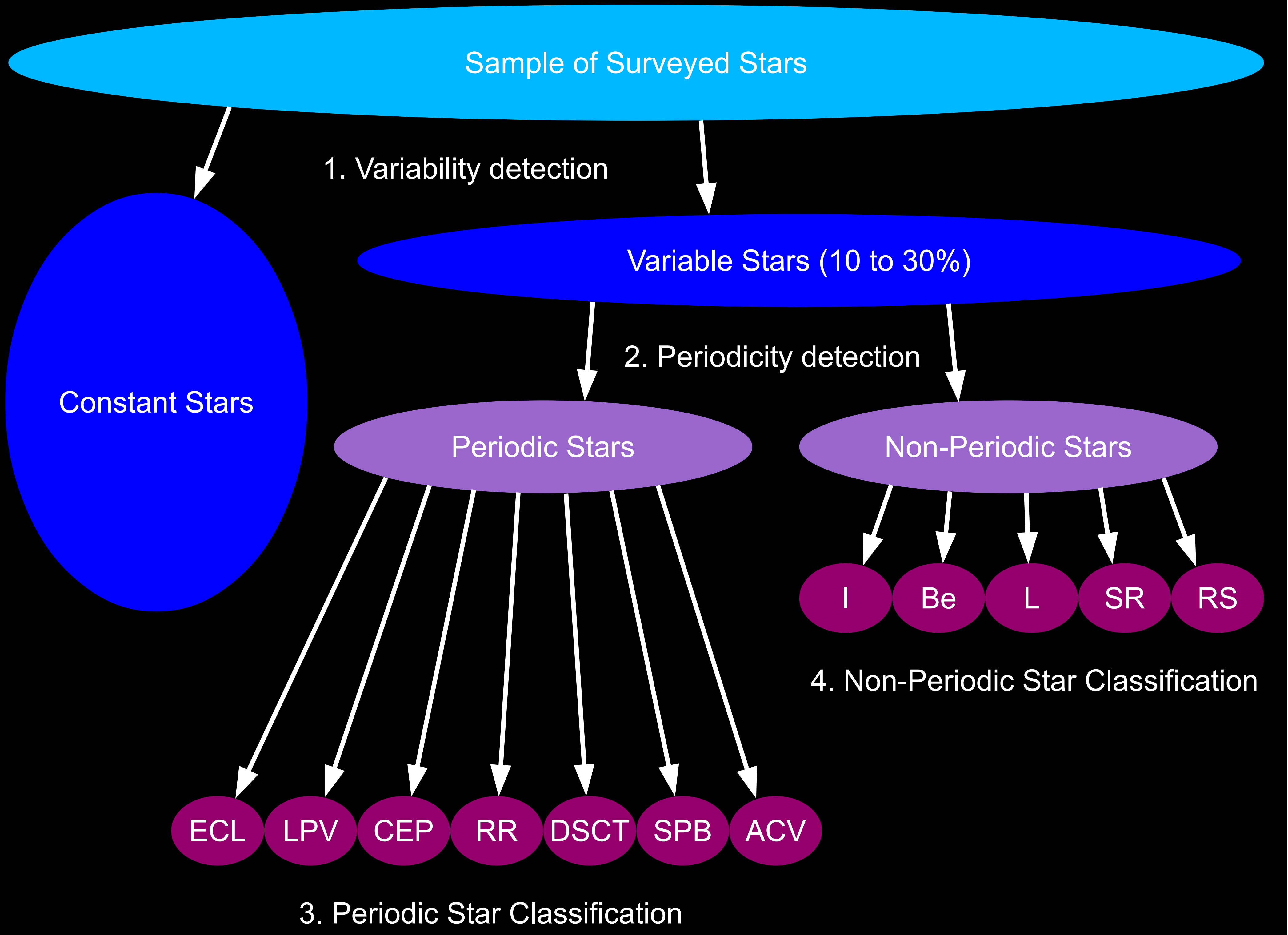


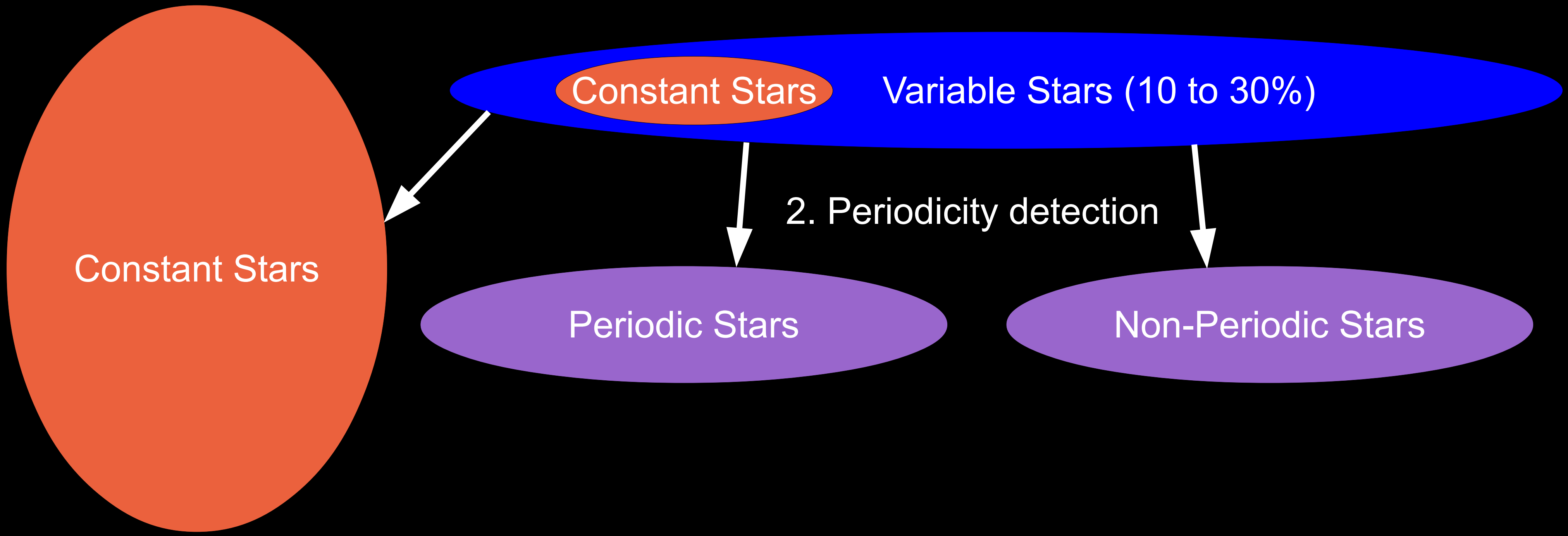
Comparison of tests to select variables


Union (Abbe/Stet/IQR/Chi2/Kur/Out)Test
# Selected
# Variables
#Periodics
# of sources
0
5000
10000
15000
20000
25000
30000
log10(threshold)
-20
-18
-15
-12
-10
-8
-6
-4
-2
8
8
9
10
10
11
12
13
14
16
20
25
75
75
77
82
86
88
90
92
94
96
98
99
100
96
96
97
98
98
99
99
99
99
100
100
100
100

Choice for variability detection

- Union of Stetson with pValues < 1e-2 and Inter-quartile range with pValues < 1e-3
  - 17'006 candidates (14.8 % of total)


Sample of Surveyed Stars
1. Variability detection
Constant Stars
Variable Stars (10 to 30%)
2. Periodicity detection
Periodic Stars
Non-Periodic Stars
I
Be
L
SR
RS
4. Non-Periodic Star Classification
ECL
LPV
CEP
RR
DSCT
SPB
ACV
3. Periodic Star Classification



Constant Stars
Variable Stars (10 to 30%)
Constant Stars
2. Periodicity detection
Periodic Stars
Non-Periodic Stars

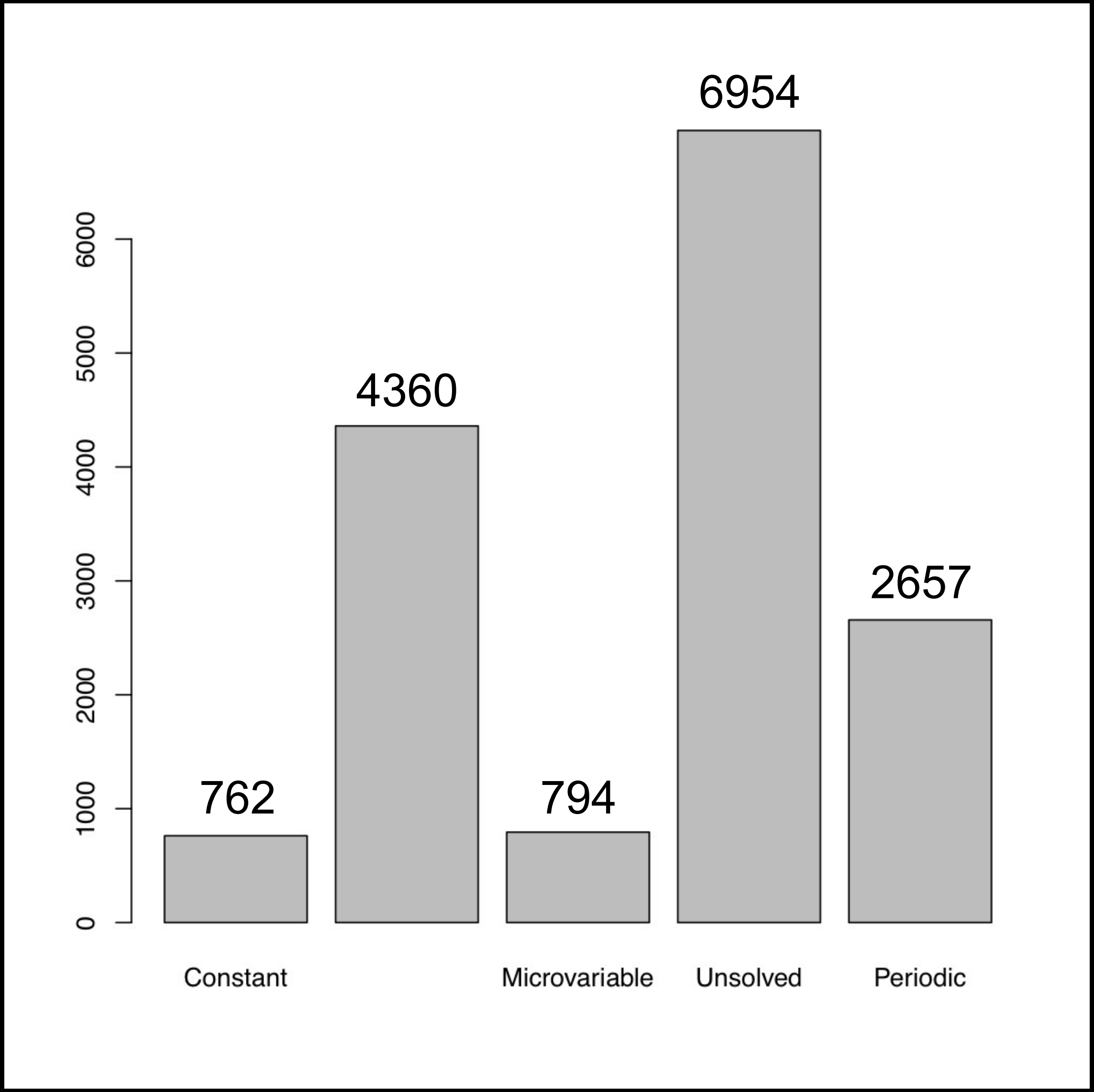
Periodicity detection through a supervised classification

- Compute a number of “attributes” characterizing objects and their light-curve
- Use attributes (features) as predicators (variables) in a supervised classifiers
- Train the classifier with a set of stars of known types
- Use a 10-fold Cross-Validation (CV) to evaluate the performance of this approach

Period search

- Generalized Lomb–Scargle method (Zechmeister & Kűrster 2009)
- Our Sample of 15'527 stars includes 3022 stars with known periods (3022 = 2657 P + 365 U)
- Recovery rate of 77 % (2323 out of 3022)
  - ➔ 1644 with correct period
  - ➔ 679 with twice the period values
- Recovery rate for the 2657 periodic = 86 % (i.e., 2300)

Hipparcos variability types

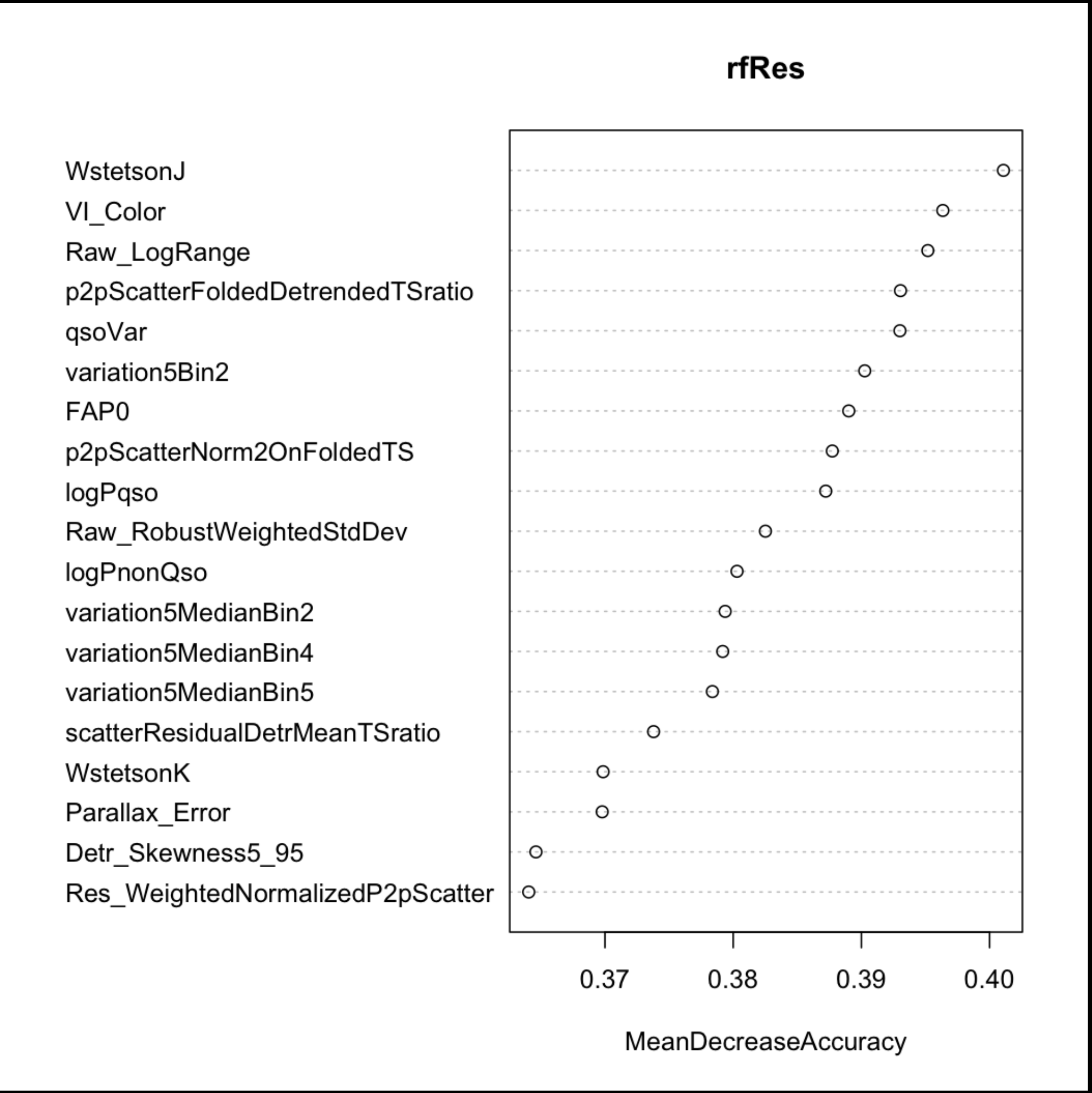


Total : 15527

P : 2657 != 2712

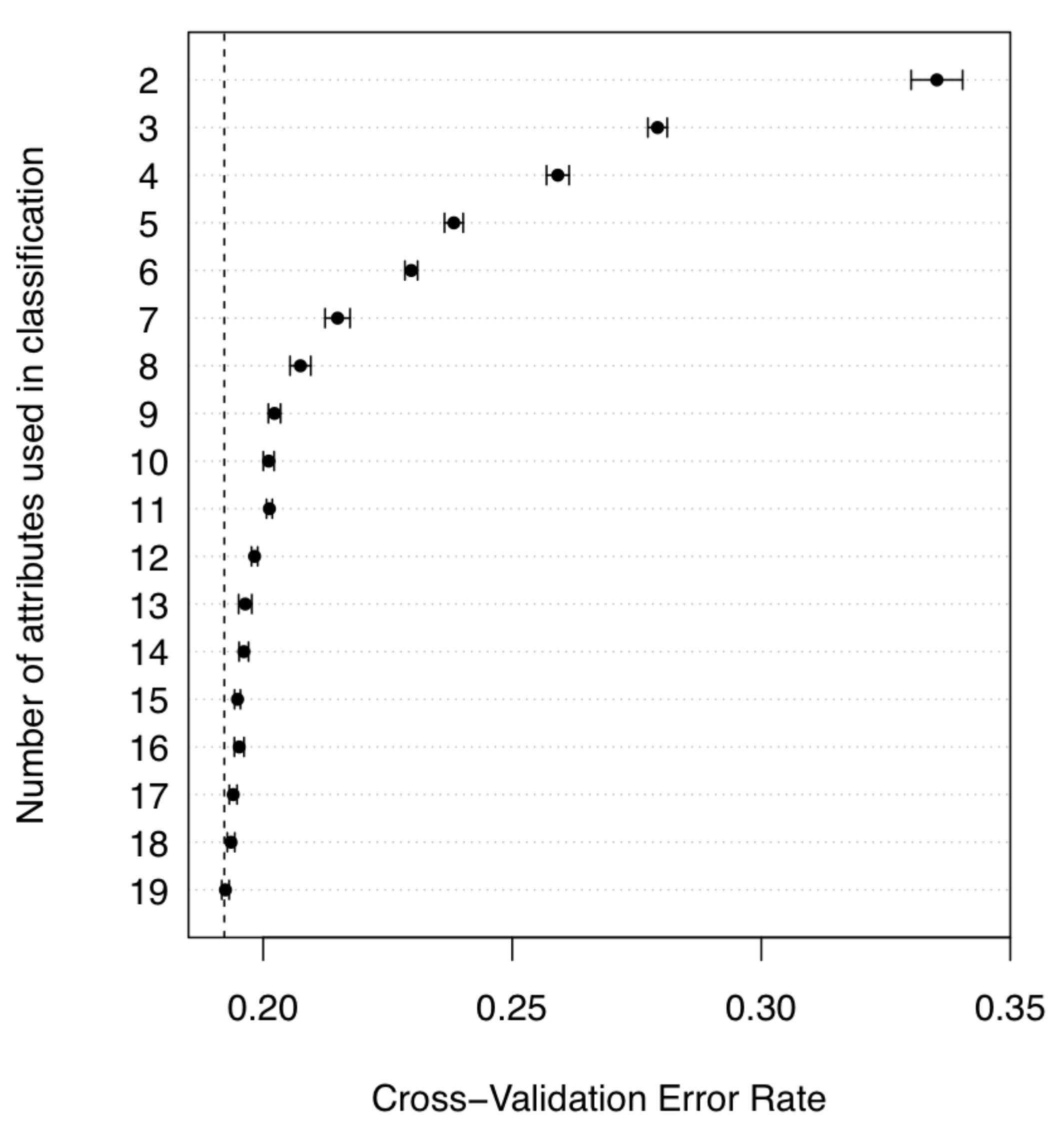
Hip Types R (662)
D (816)
...removed from TS!

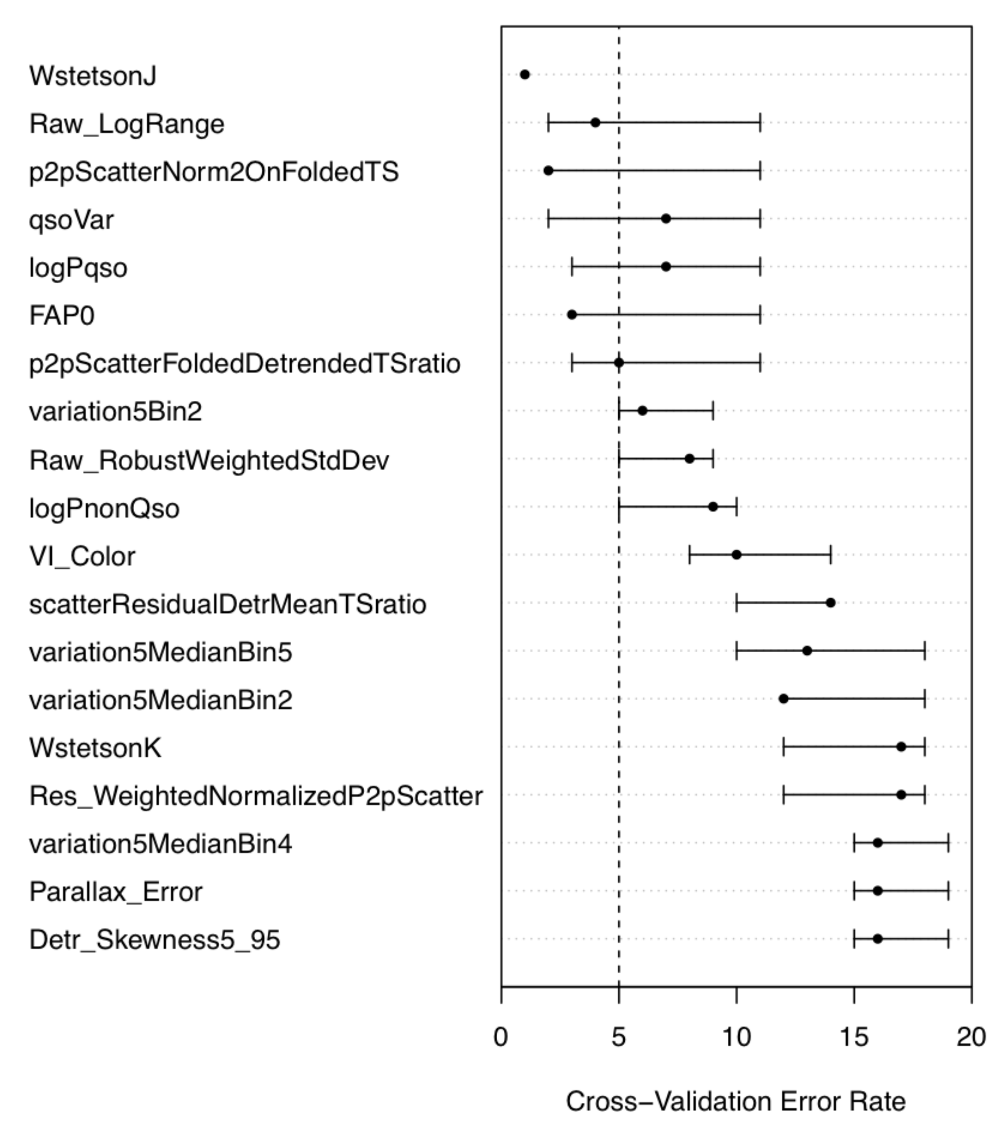
Random forest attribute importance



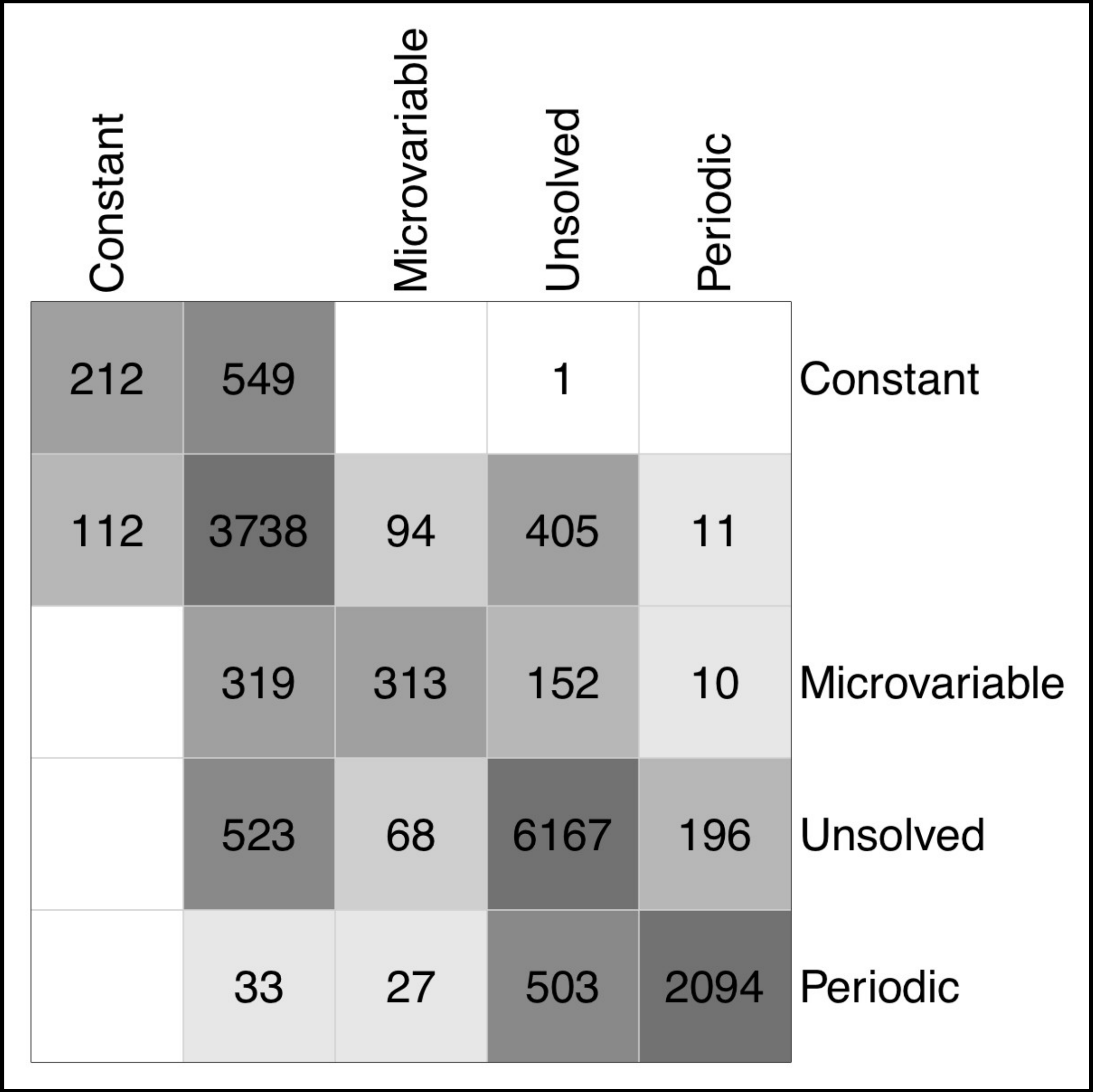
- Attributes with Spearman correlation larger than 80% are trimmed

Random forest CV error rates



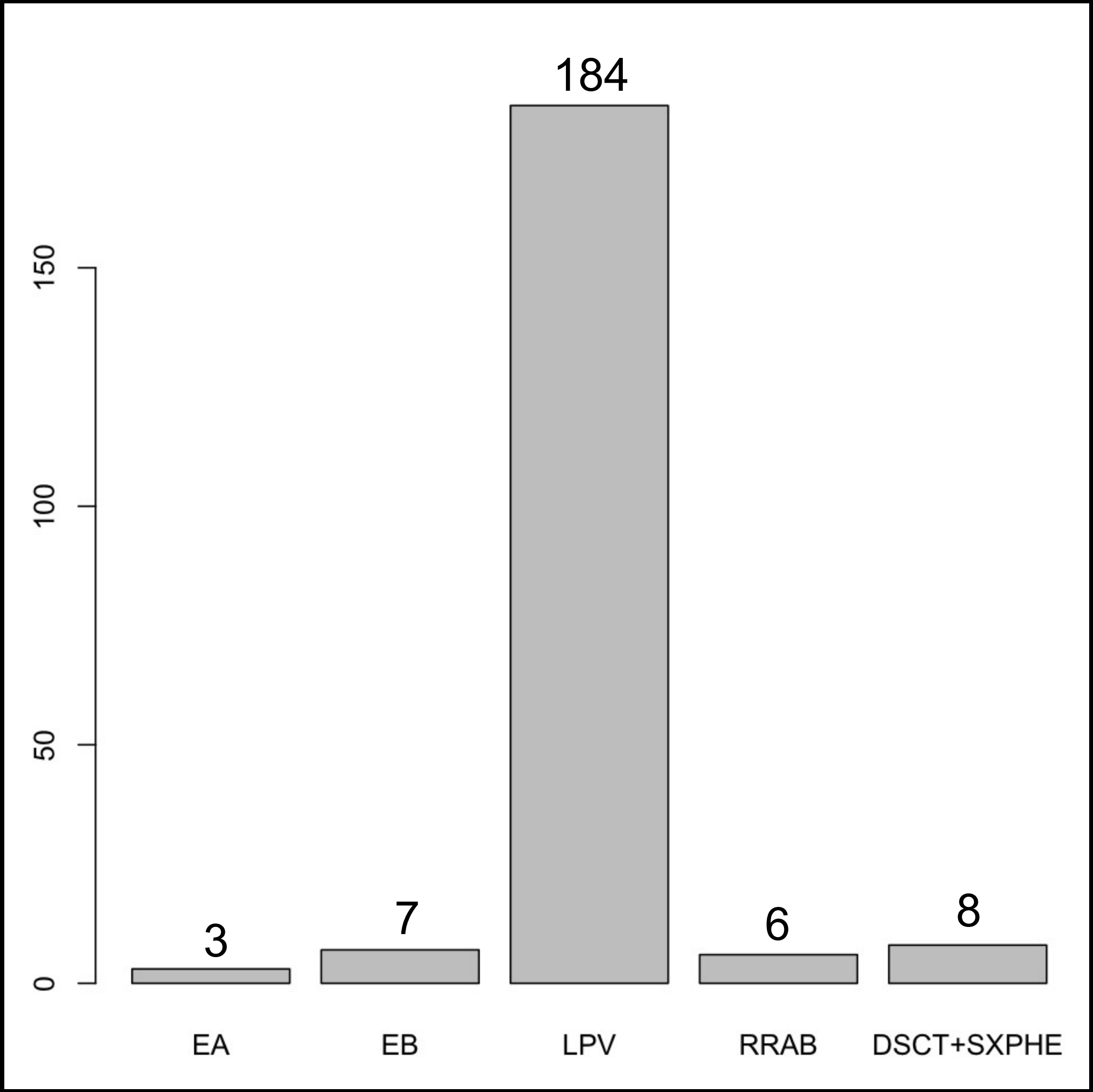


Random forest confusion matrix



Constant		Microvariable	Unsolved	Periodic	
212	549		1		Constant
112	3738	94	405	11	
	319	313	152	10	Microvariable
	523	68	6167	196	Unsolved
	33	27	503	2094	Periodic

Predicted types for non-periodic identified as periodic



Conclusion

- We established a complete scheme for variable star classification
- Optimized for Hipparcos data
  - ➔ Training set must be representative of the test set
- Hipparcos classification relatively easy: clean sample and well known stars
- Can be completed with additional information
  - ➔ Color light curves
  - ➔ Radial velocity time series
- Next step: apply our scheme to other surveys