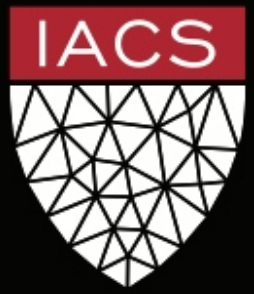


Harvard School of Engineering and Applied Sciences

IACS Institute for Applied Computational Science

Clustering-based feature learning

Pavlos Protopapas

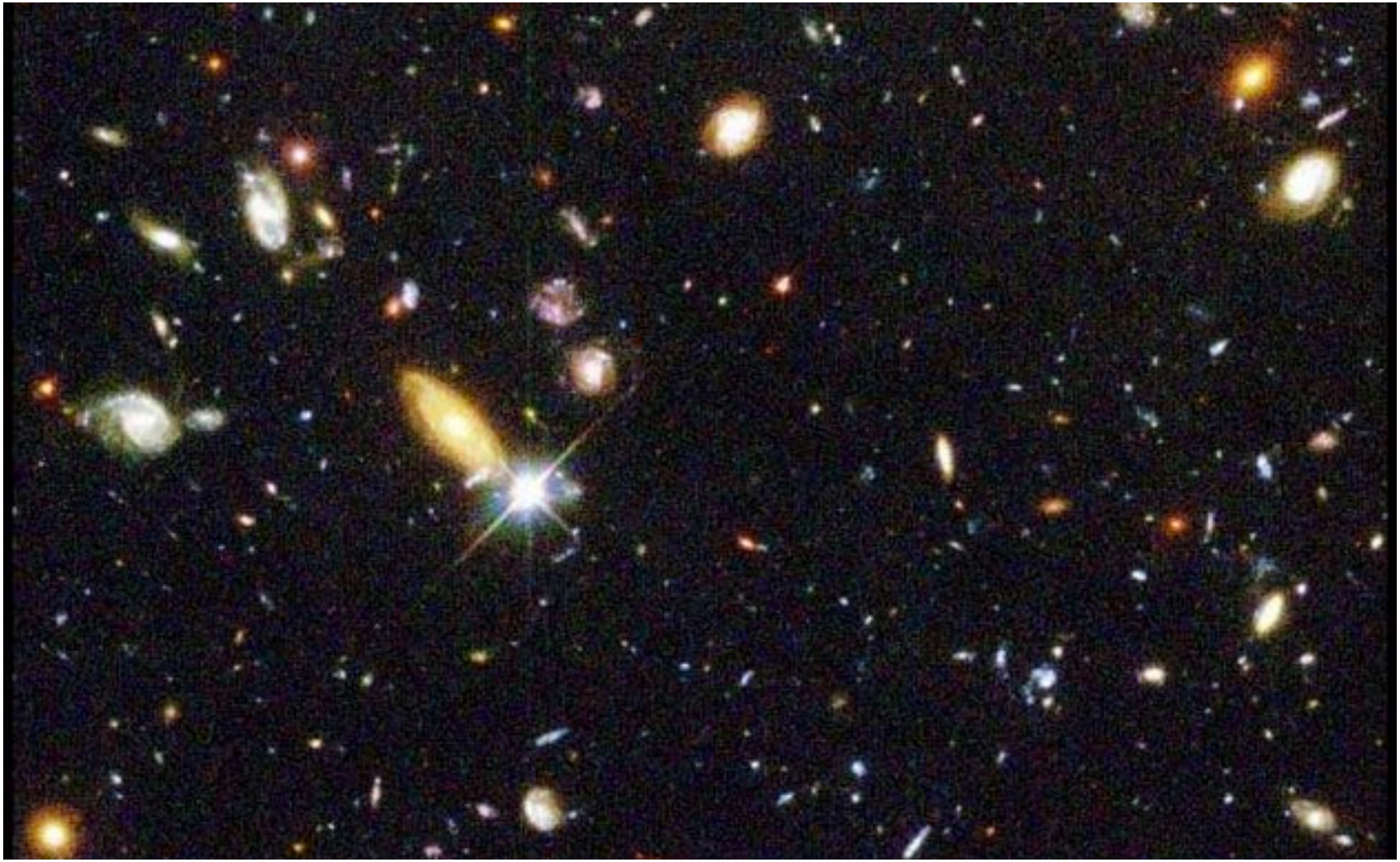
Institute for Applied Computational Science

Collaborators: Cristóbal Mackenzie

K.Pichara, D-W Kim, Isadora Nun, Xufei Wang, Nicolas Castro, Justin Yang,

June 7, 2016

Vision



Observable Universe

Breakout of the visible universe

Superclusters in the observable universe = 10 million

Galaxy groups in the observable universe = 25 billion

Large galaxies in the observable universe = 350 billion

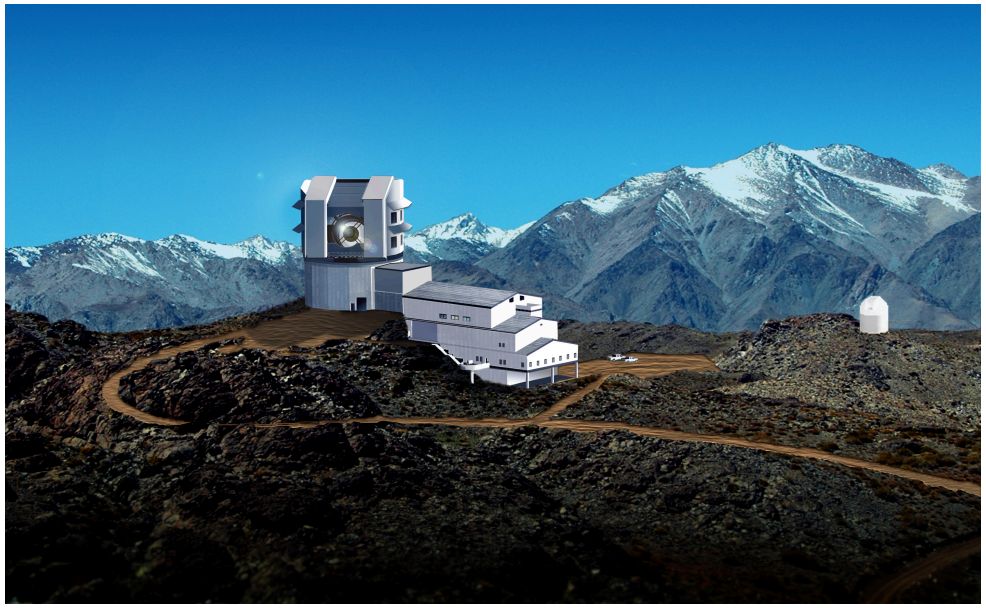
Dwarf galaxies in the observable universe = 7 trillion

Stars in the observable universe = **30 billion trillion**

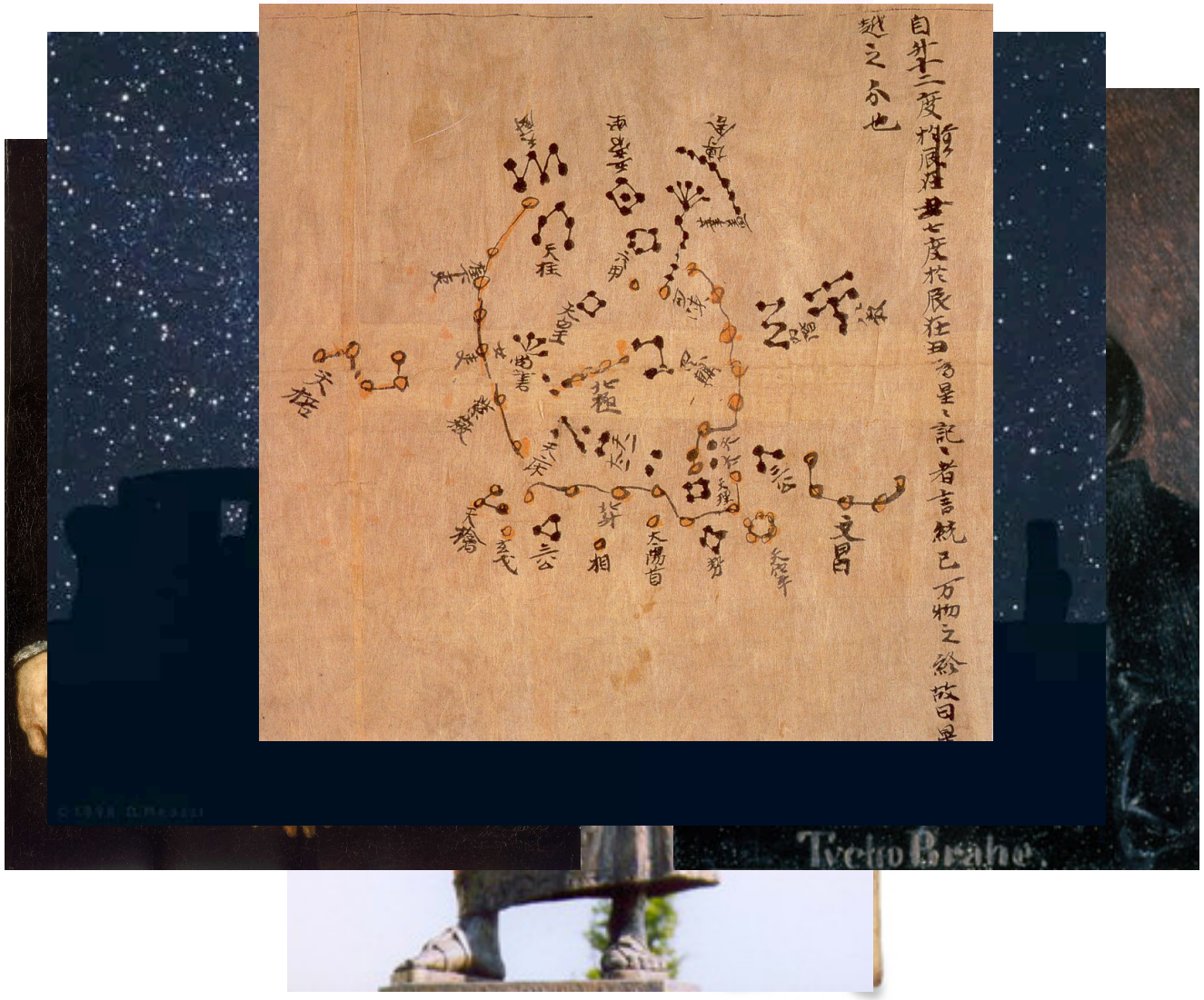
Real visible universe

The Large Synoptic Survey Telescope (LSST) will approximately see 40 billion objects

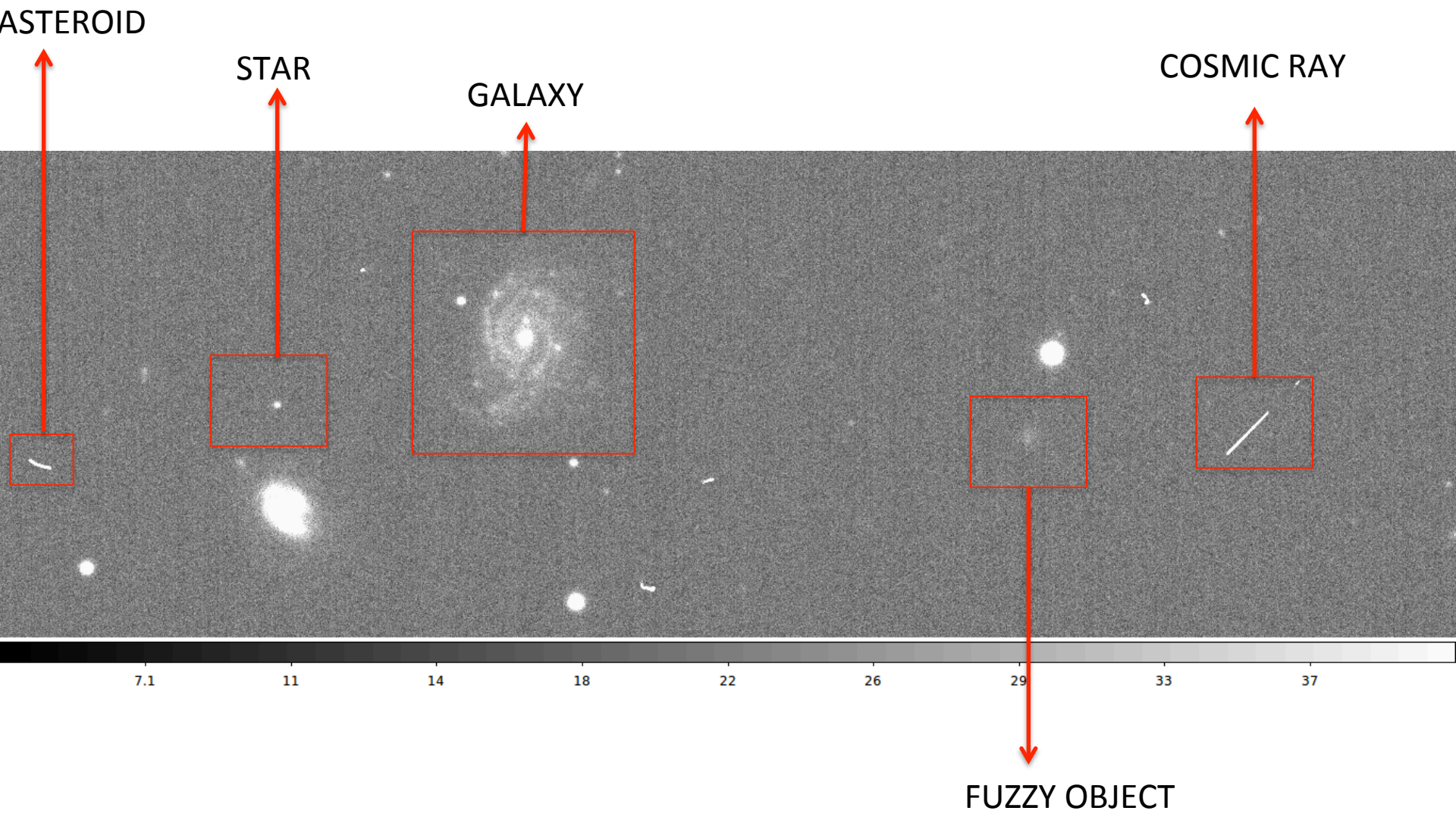
All telescopes together in the next 10 years: **1 Trillion objects**



Data collection in astronomy

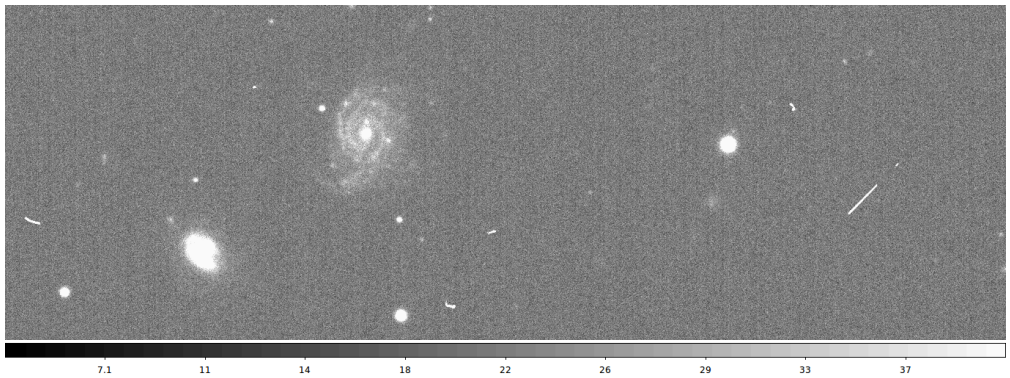


- Universe is huge but the number of objects to be observed is manageable
- Astronomy is not new in collecting data
- New trend is data science


ASTEROID
STAR
GALAXY
COSMIC RAY
7.1
11
14
18
22
26
29
33
37
FUZZY OBJECT

Data Collection and Processing

IMAGE

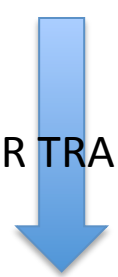


Photometry and astrometry

DETECTIONS

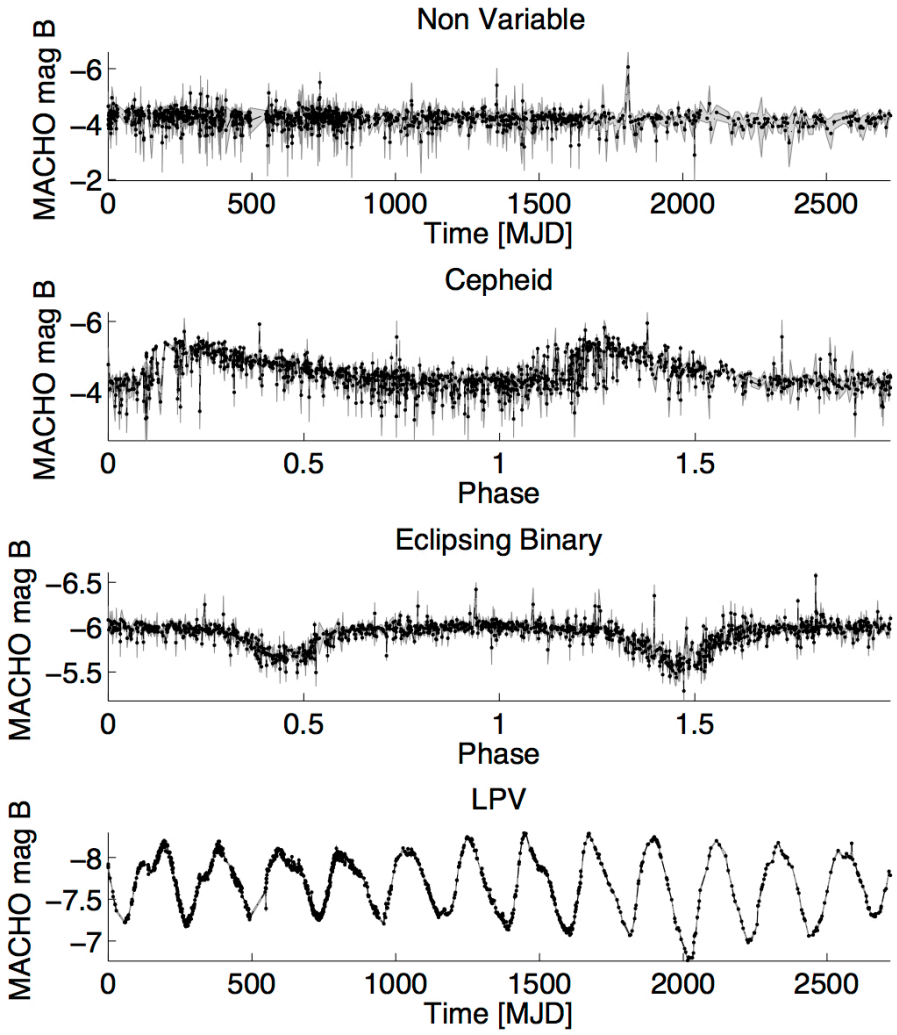
DETECTION ID	Time	POS –X (RA)	POS-Y (DEC)	BRIGHTNESS	…
1.323242.1	321211.1	02:30:21	-73:23:12	19.5	…
2.653991.3	321212.2	12:10:12	-63:23:12	21.2	…
4.992242.8	321211.8	02:30:20	-73:23:11	21.2	…
…	….	…	…	…	…

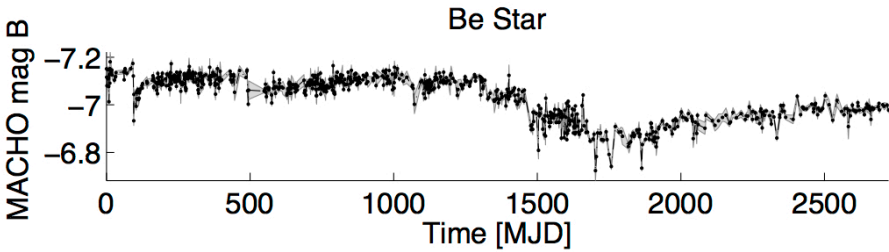
DETECTION ID	Time	POS –X (RA)	POS-Y (DEC)	BRIGHTNESS	…
1.323242.1	321211.1	02:30:21	-73:23:12	19.5	…
2.653991.3	321212.2	12:10:12	-63:23:12	21.2	…
4.992242.8	321211.8	02:30:20	-73:23:11	21.1	…
…	….	…	…	…	…



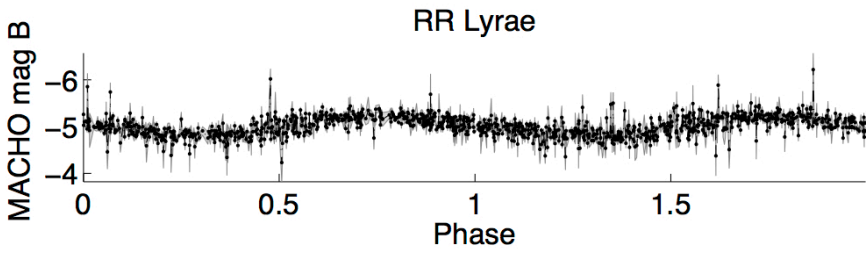
MOVING OBJECTS (ASTEROIDS OR TRANSIENTS) LIGHT CURVES (TIME SERIES)

DETECTION ID	Time	POS –X (RA)	POS-Y (DEC)	BRIGHTNESS	…
2.653991.3	321212.2	12:10:12	-63:23:12	21.2	…
…	….	…	…	…	…

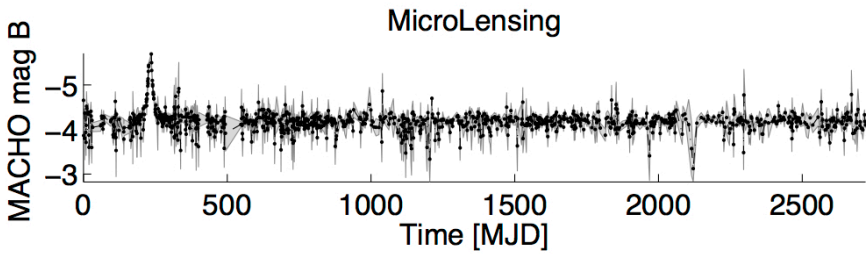

Non Variable
MACHO mag B
−6
−4
−2
0
500
1000
1500
2000
2500
Time [MJD]


Be Star
MACHO mag B
−7.2
−7
−6.8
0
500
1000
1500
2000
2500
Time [MJD]

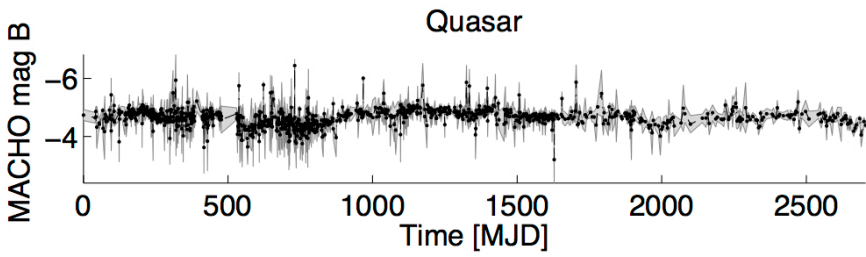
Cepheid
MACHO mag B
−6
−4
0
0.5
1
1.5
Phase


RR Lyrae
MACHO mag B
−6
−5
−4
0
0.5
1
1.5
Phase

Eclipsing Binary
MACHO mag B
−6.5
−6
−5.5
0
0.5
1
1.5
Phase


MicroLensing
MACHO mag B
−5
−4
−3
0
500
1000
1500
2000
2500
Time [MJD]

LPV
MACHO mag B
−8
−7.5
−7
0
500
1000
1500
2000
2500
Time [MJD]


Quasar
MACHO mag B
−6
−4
0
500
1000
1500
2000
2500
Time [MJD]

Reality

Today's astronomical surveys (Pan-STARRS, LSST etc) are and will be producing literally billions of time series.

Real questions

Can we classify them using automatic methods?

Can we do this in real time?

Science Questions

Having everything classified what can we do?

A. Create larger collection of rare objects so they can be studied in more details

B. Discover new novel physical phenomena

C. Resolve observational biases and therefore answer questions on abundance, relative populations

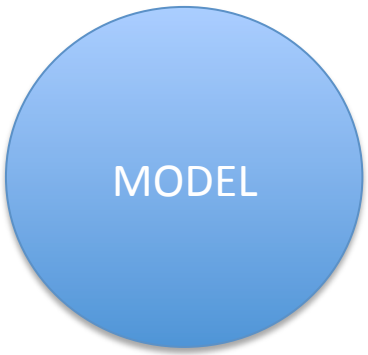
D. Reduce error bars on known cosmological quantities

Classification

TRAINING SET

OBJECT ID	TimeS	POS –X (RA)	POS-Y (DEC)	FEATURES	LABEL
LC22	[0,0.2…]	02:30:2	-73:23:12	[19.5,21.1, ..]	**CEPHEID**
LC24	[0, 0.2…]	12:10:1	-63:23:12	[21.2, 20.5, …]	**QUASAR**
…	….	…	…	…	…

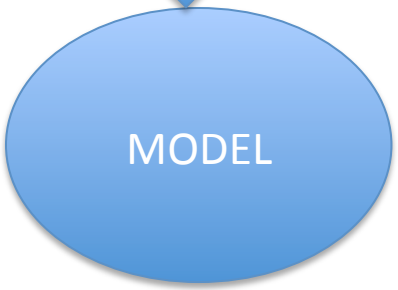
LEARN



DATABASE

OBJECT ID	TimeS	POS –X (RA)	POS-Y (DEC)	FEATURES	…
LC22	[0,0.2…]	02:30:21	-73:23:12	[19.5,21.1, ..]	…
LC24	[0, 0.2…]	12:10:12	-63:23:12	[21.2, 20.5, …]	…
…	….	…	…	…	…

APPLY



OBJECT ID	TimeS	POS –X (RA)	POS-Y (DEC)	FEATURES	PREDICTED LABELS
LC22	[0,0.2…]	02:30:21	-73:23:12	[19.5,21.1, ..]	CEPH
LC24	[0, 0.2…]	12:10:12	-63:23:12	[21.2, 20.5, …]	RRL
…	….	…	…	…	…

- CHALLENGES WITH HETEROGENOUS DATA
- DIFFERENT SENSORS
- DIFFERENT SCIENCE

What are the challenges?

Feature extraction, feature selection etc

Classification

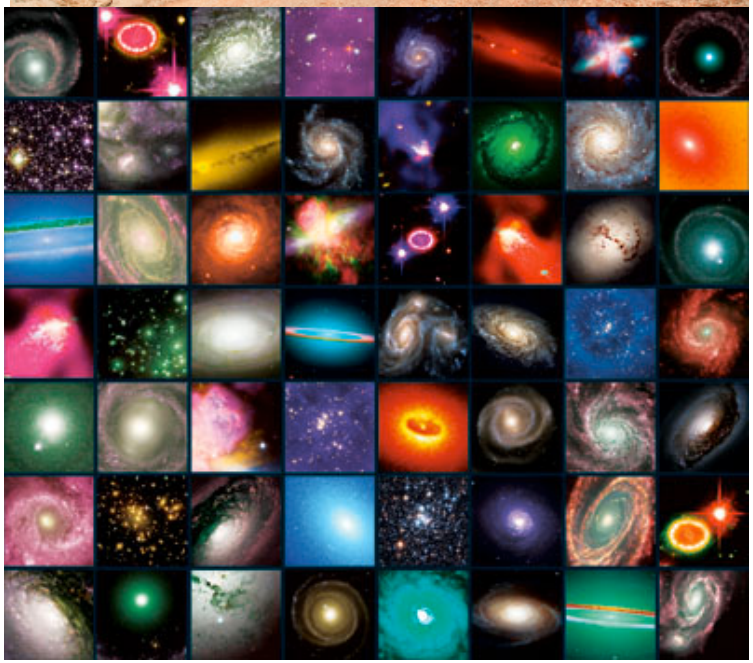
- Deal with uncertain features
- Deal with uncertain labels
- Streaming classification

Outlier detection

CLASSIFY THEM ALL (combining all data)

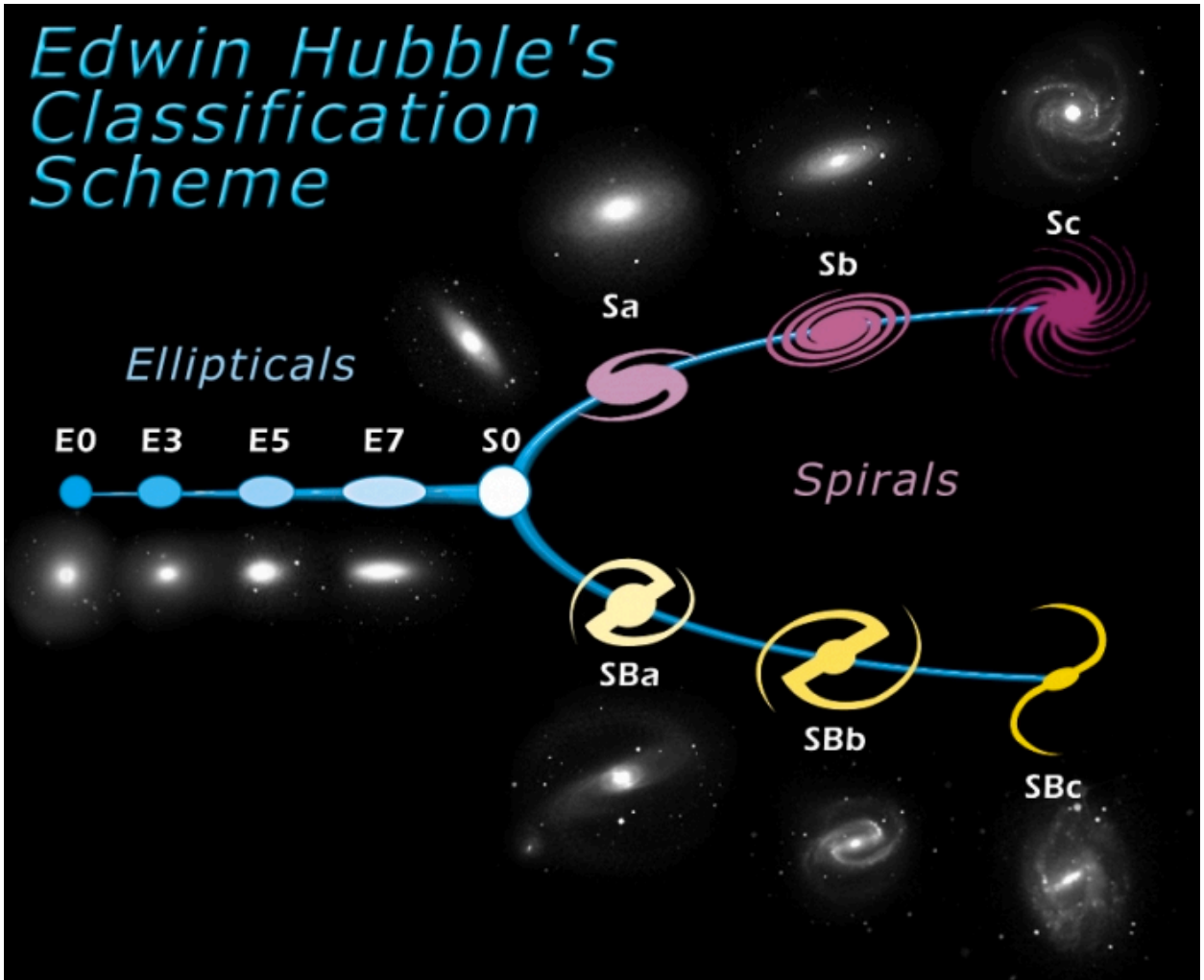
- Meta-classifier
- Missing data
- **Training set**

Experimental Design and Follow Ups

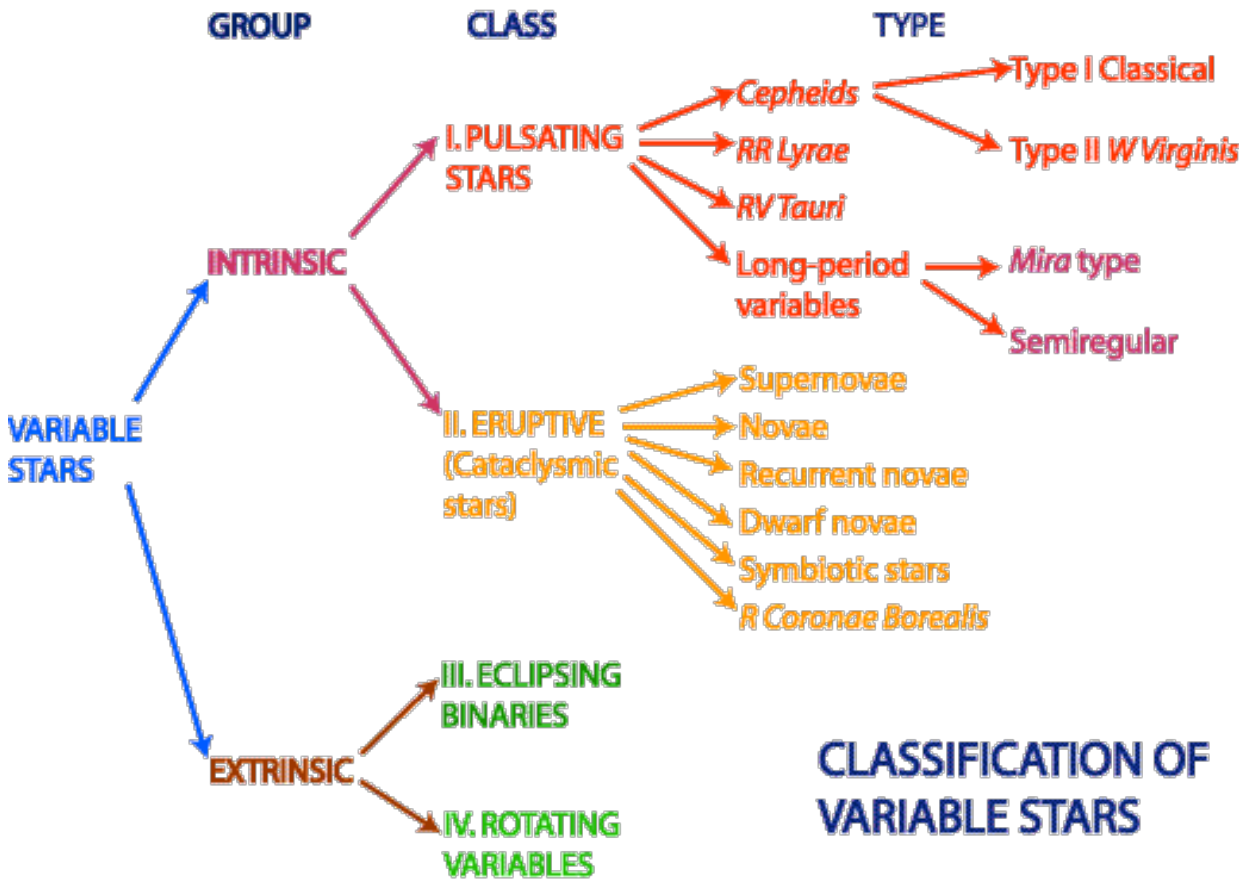


Training set- Labels. The zoo

- Labels (animals in kingdom)



Variable objects from astronomers perspective



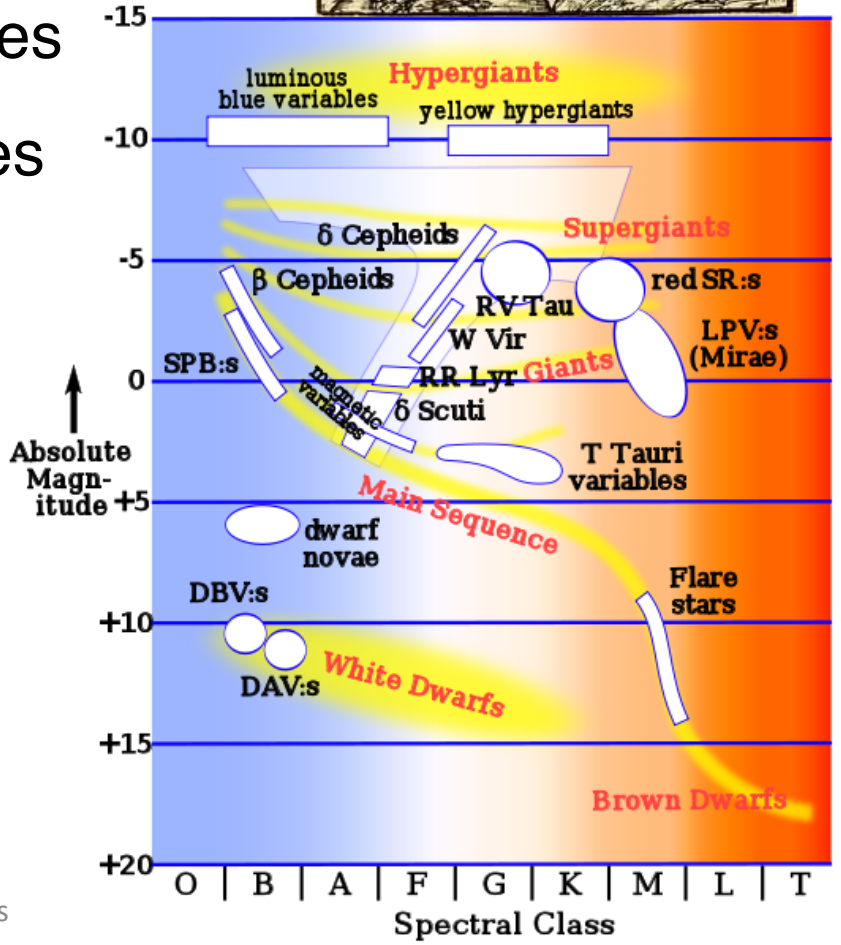
The art of building training sets

How do we create training sets?

Combination of many thing

- Looking at the color- magnitudes
- Manual inspection of lightcurves
- Sometimes spectra.
- Synthesize them from models

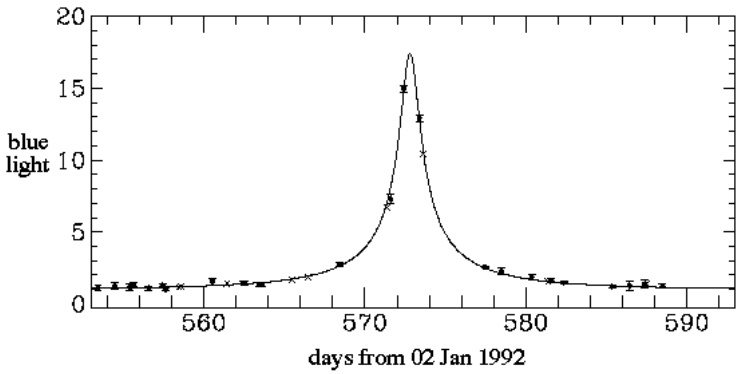
This is an experimental design question

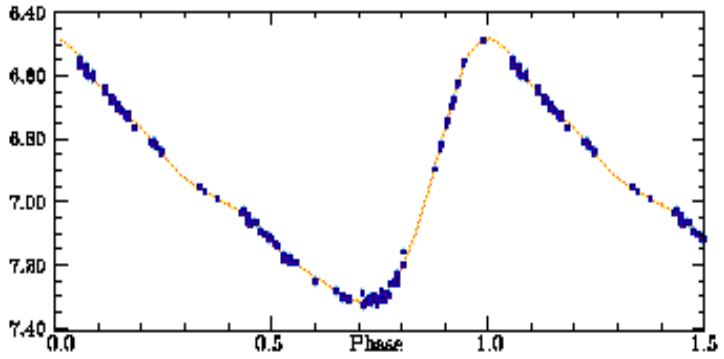


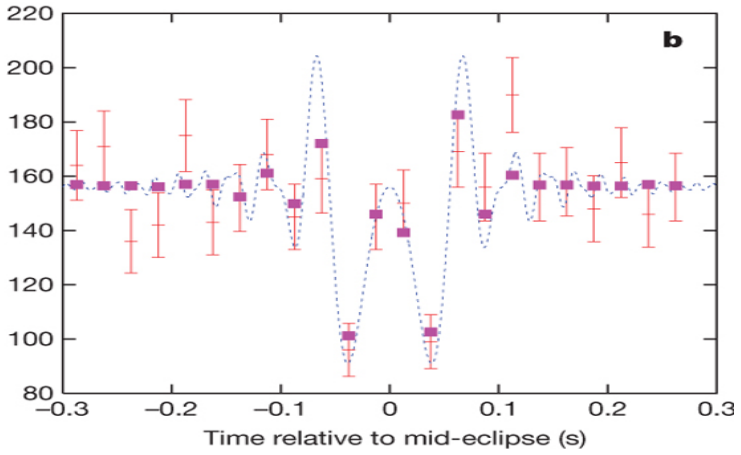
The secret

ASK YOUR MOM

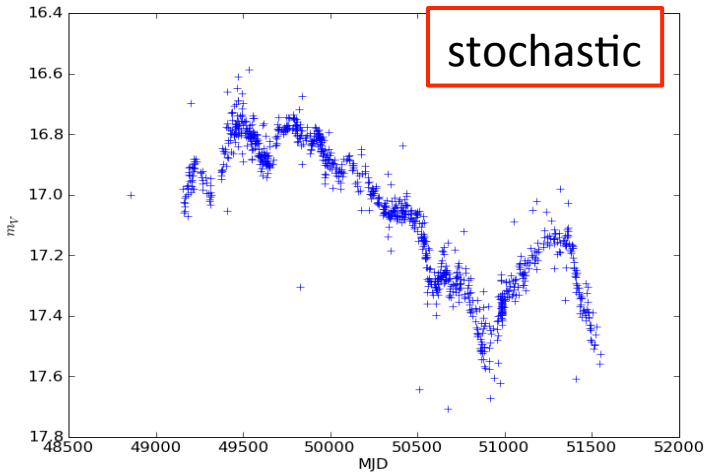
Time variability

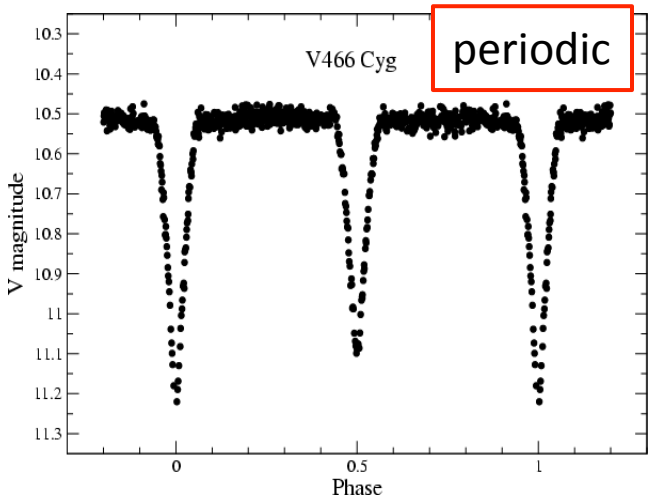


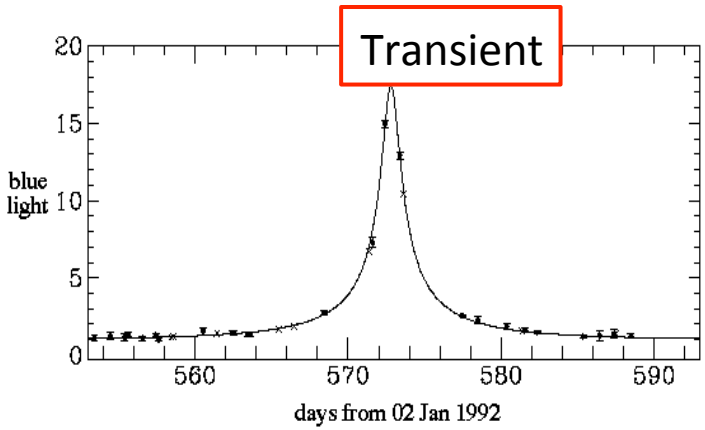


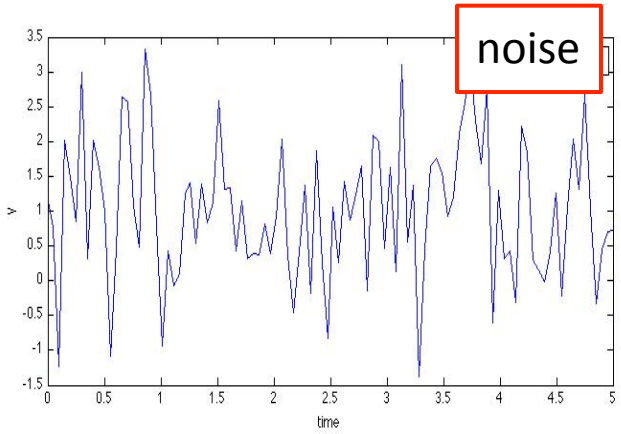


Variable objects from data scientist perspective



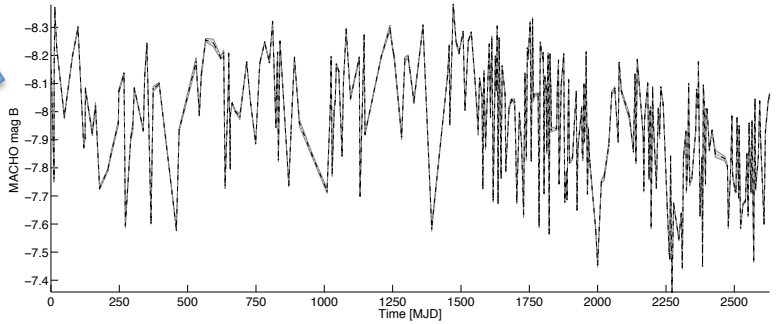






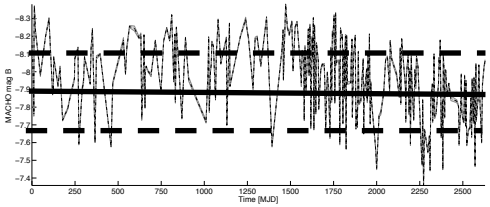
Features

Time-series features

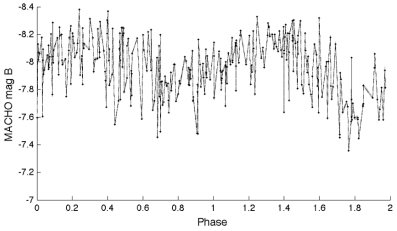


STD

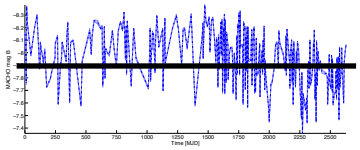
Standard Deviation



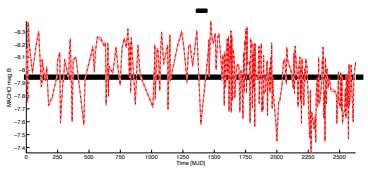
Period



Color

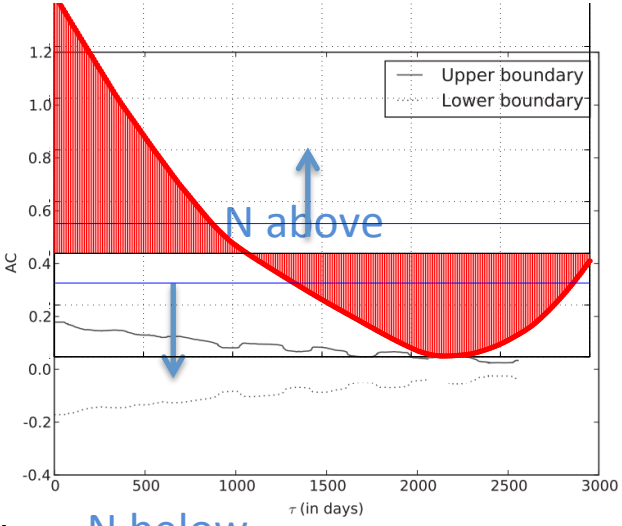


Mean B



Mean R

N



*Kim et al., 2011; Pichara et al., 2012, Nun et al 2015

FATS

Feature Analysis for Time Series (FATS)

http://isadoranun.github.io/tsfeat/FeaturesDocumentation.html

Institute for
Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

Feature Analysis for Time Series

Authors: Isadora Nun isadoranun@seas.harvard.edu , Pavlos Protopapas pavlos@seas.harvard.edu

Contributors: Daniel Acuña, Nicolás Castro, Rahul Dave, Cristobal Mackenzie, Adam Miller, Karim Pichara, Andrés Riveros, Brandon Sim and Ming Zhu

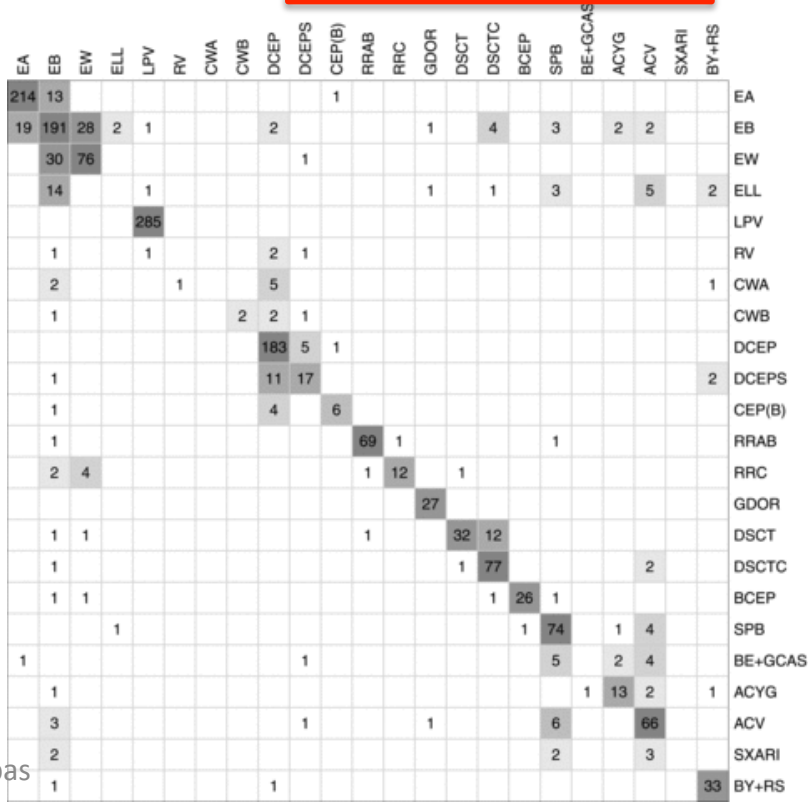
Periodic Variable (RF and SVM)

- Wachman et. al. 2009 ML Journal. Used OGLEII+MACHO periodic variables

	Ceph	EB	RRL	Ceph	EB	RRL	Ceph	EB	RRL	Ceph	EB	RRL
Cepheid	3416	1	13	3382	1	3	3363	1	3	3352	1	0
EB	0	3389	0	0	3364	0	0	3342	0	0	3312	0
RRL	9	0	7259	1	0	7195	0	0	7166	0	0	7138

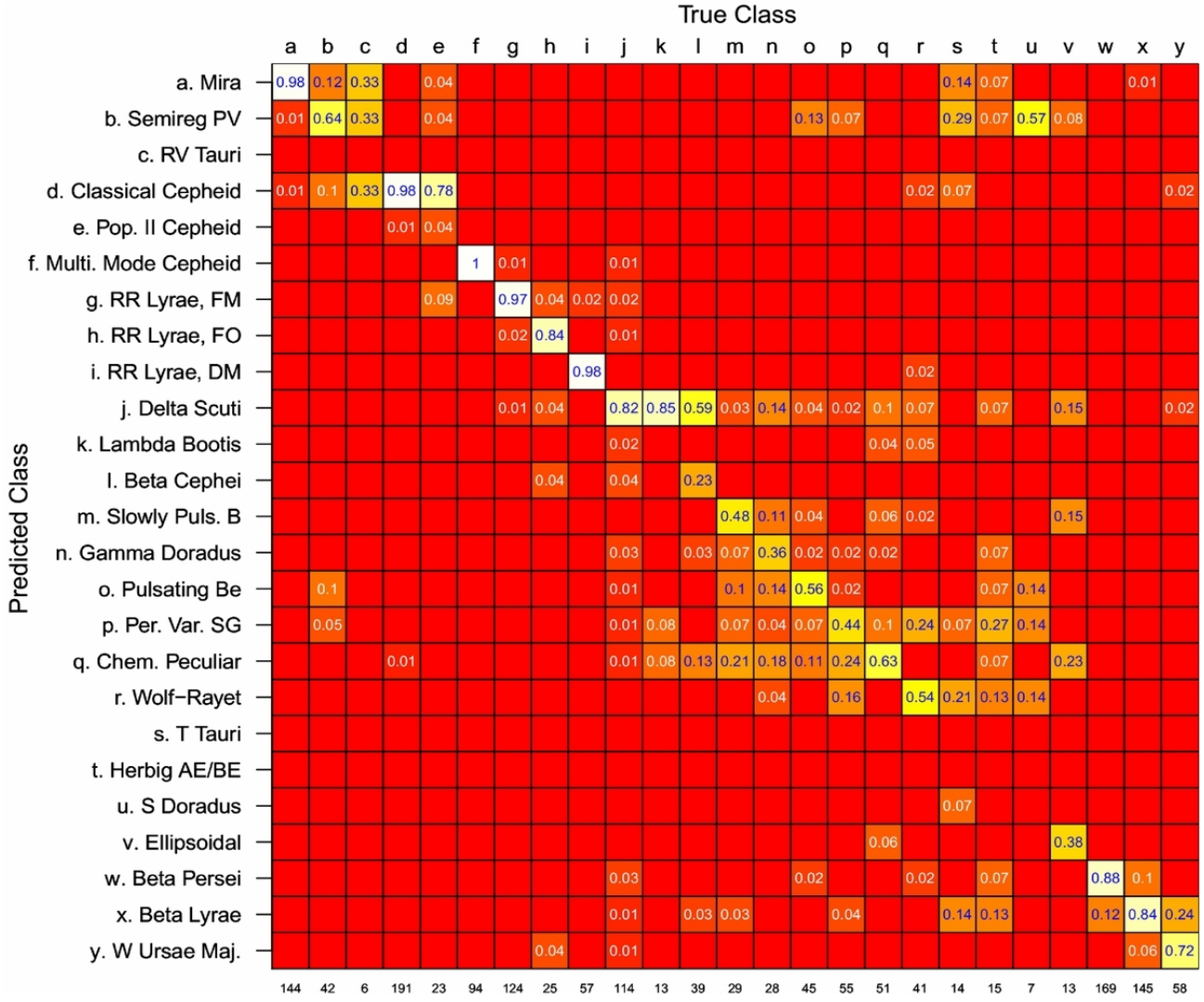
- P. Dubath 2011, MNRAS.

Used *Hipparcos* periodic variable stars



Monthly Notices of the Royal Astronomical Society
Volume 414, Issue 3, pages 2602-2617, 9 MAY 2011 DOI: 10.1111/j.1365-2966.2011.18575.x
http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2966.2011.18575.x/full#f8

Figure 5 from On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data
Joseph W. Richards et al. 2011 ApJ 733 10 doi:10.1088/0004-637X/733/1/10



Non periodic variables/Stochastic

- Pichara 2011, 2012 and Kim et al used SVM and RF to discriminate AGNs and Be stars from the rest in MACHO and EROS and recently to Pan-STARRS

  88% recall and 78% precision.

  Combining with other contextual information (mid-IR, xray) resulted in very high confidence candidates (missing data issue)

  .

- Mashala 2011, classification of SN and cataclysmic variables. Extremely high recall/precision

MANY MORE EXAMPLES

Unsupervised Feature Extraction

Feature extraction is expensive

Involves experts in choosing and designing them

Classification performance depends on these features

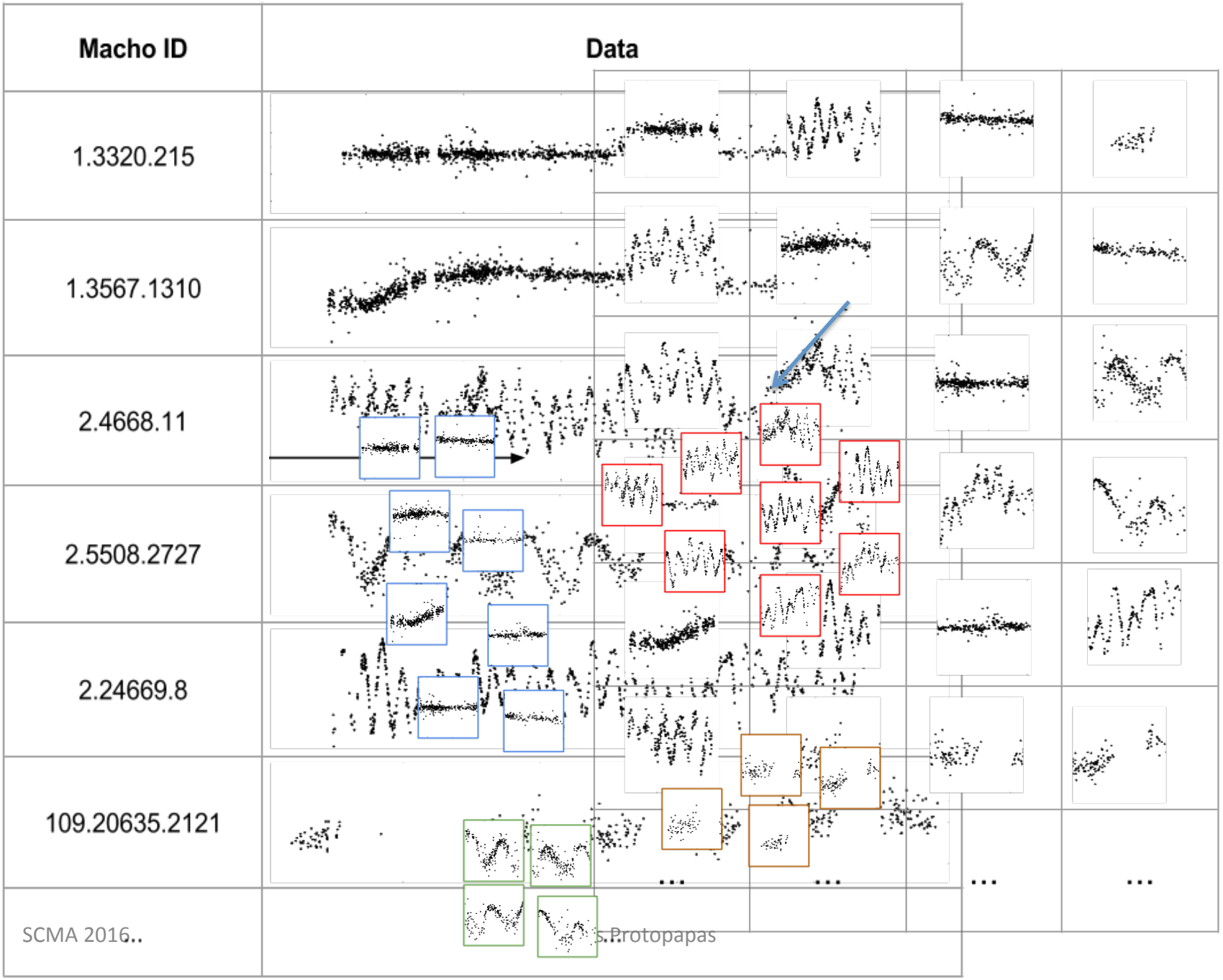
If features uncertainties are not uniform then the classification model must be adjusted (or the training set)

Unsupervised feature learning algorithm designed for variable stars

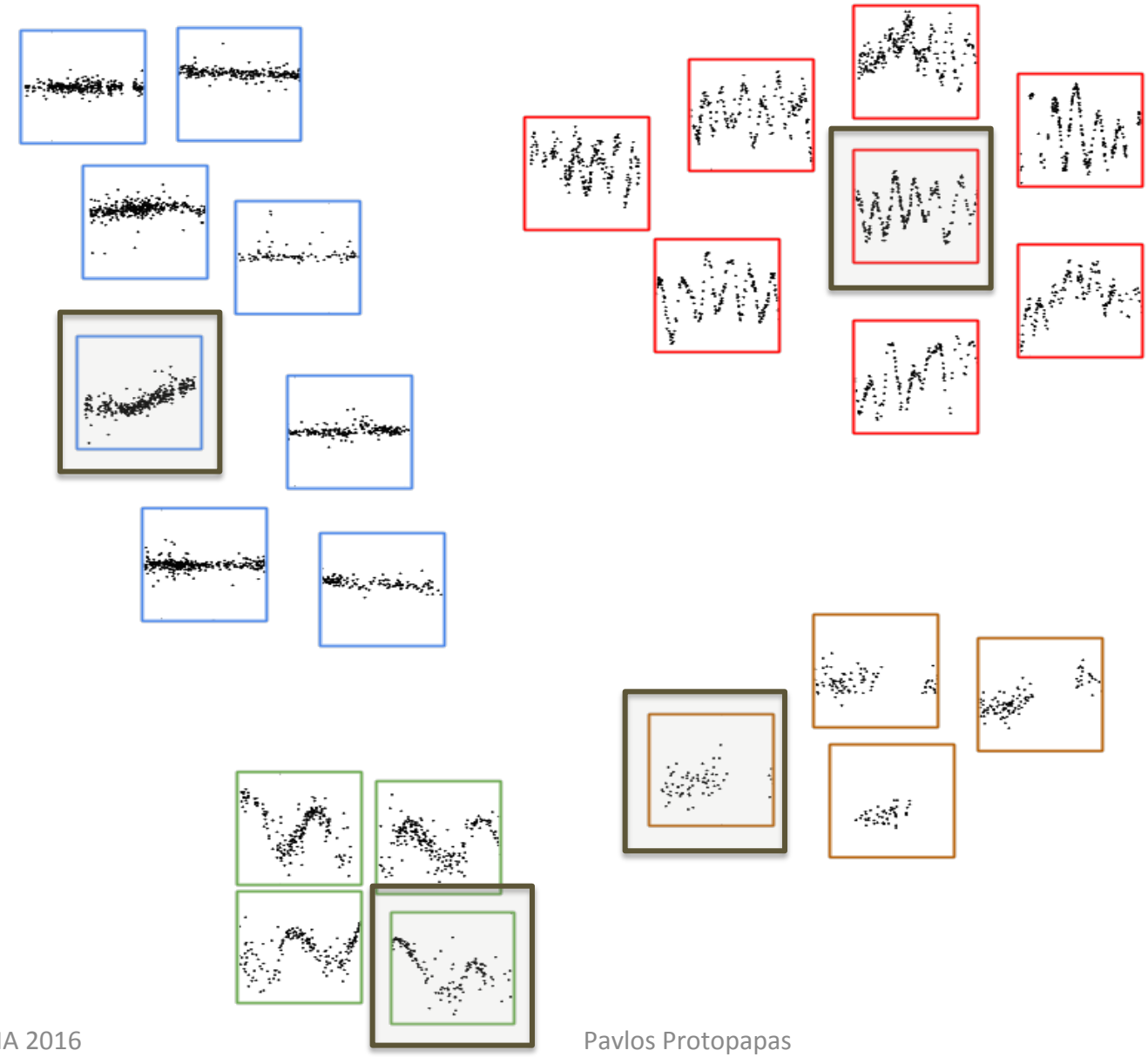
Develop methods that prescind from expert-designed features for features that are automatically learned from data

These methosd aims to use unlabeled data to learn a model that can be then used to transform data of the same kind to a new representation suitable for classification tasks.

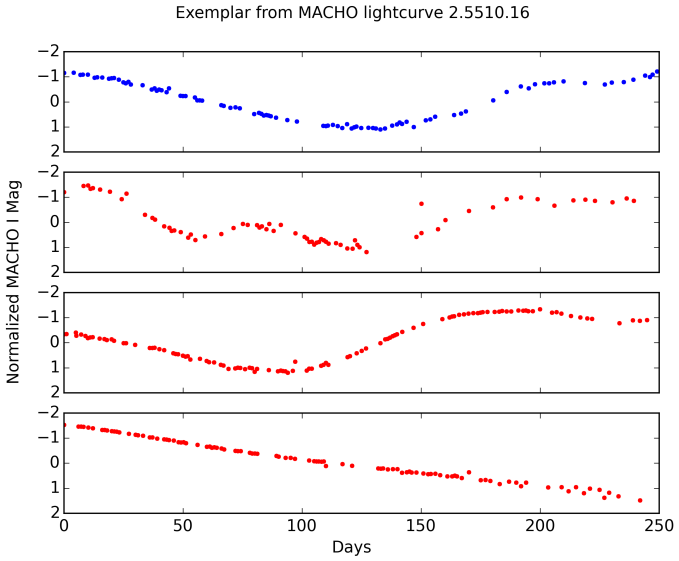
- Sparse Coding (Olshausen et al. 1996; Lee et al. 2006)
- Restricted Boltzmann Machine (Hinton & Salakhutdinov 2006; Hinton et al. 2006; Larochelle & Bengio 2008)
- Recurrent Neural Network (Hüsken&Stagge2003)
- Autoencoder (Poultney et al. 2006; Hinton & Salakhutdinov 2006; Bengio 2009)
- Clustering approaches (Coates & Ng 2012)

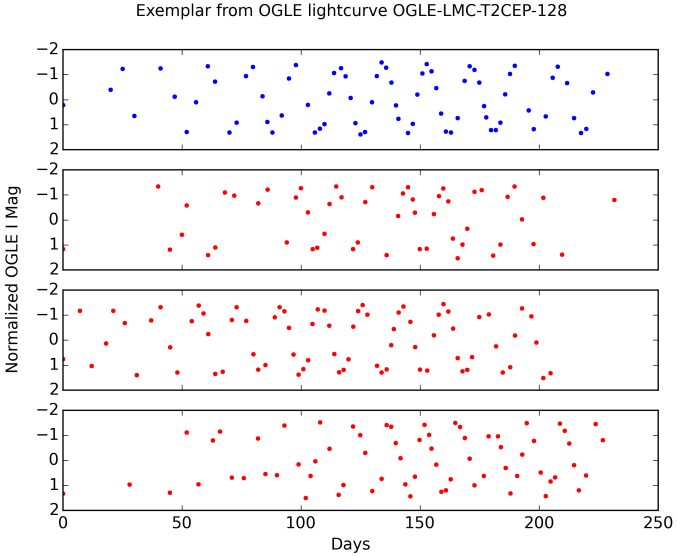

Macho ID
Data
1.3320.215
1.3567.1310
2.4668.11
2.5508.2727
2.24669.8
109.20635.2121

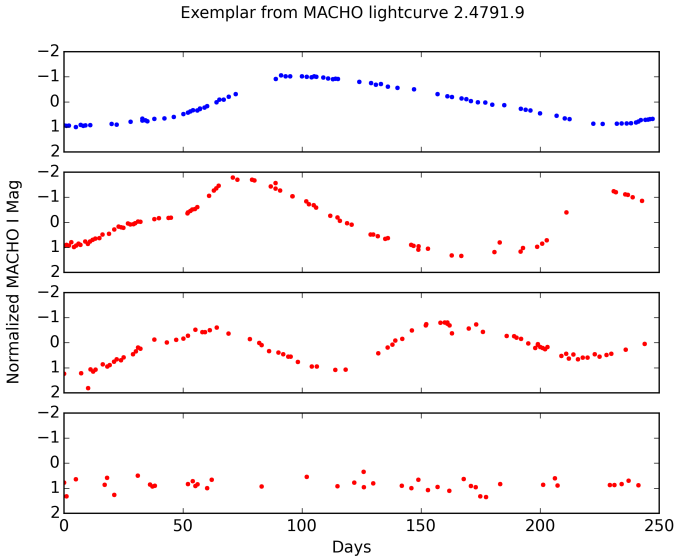
The exemplars

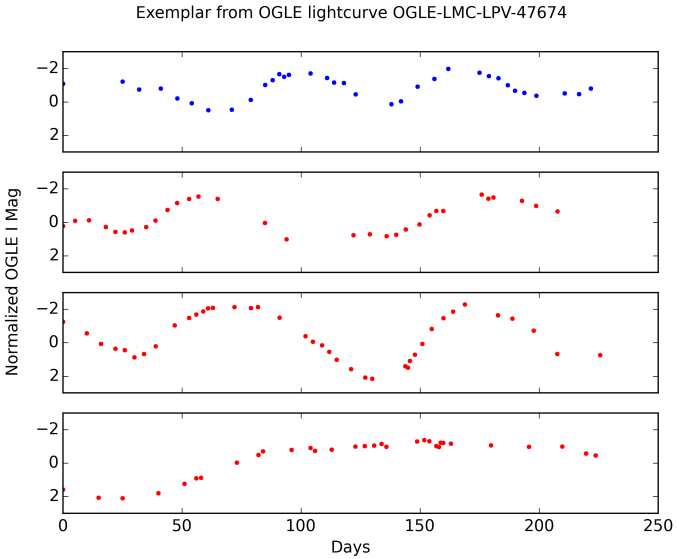


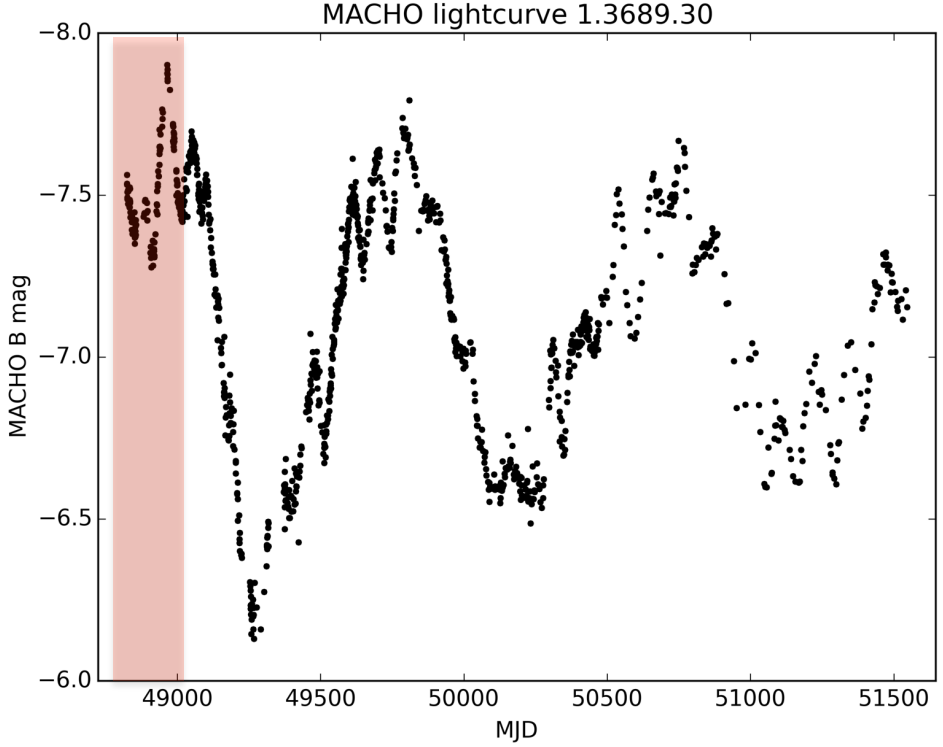
The exemplars

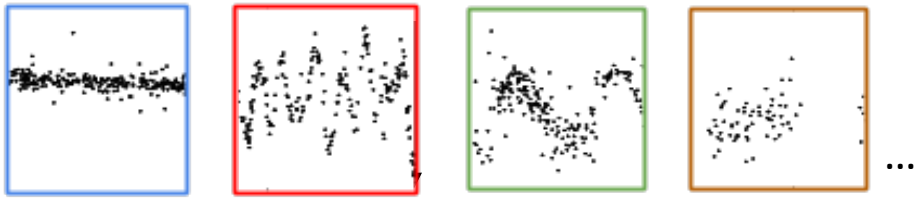








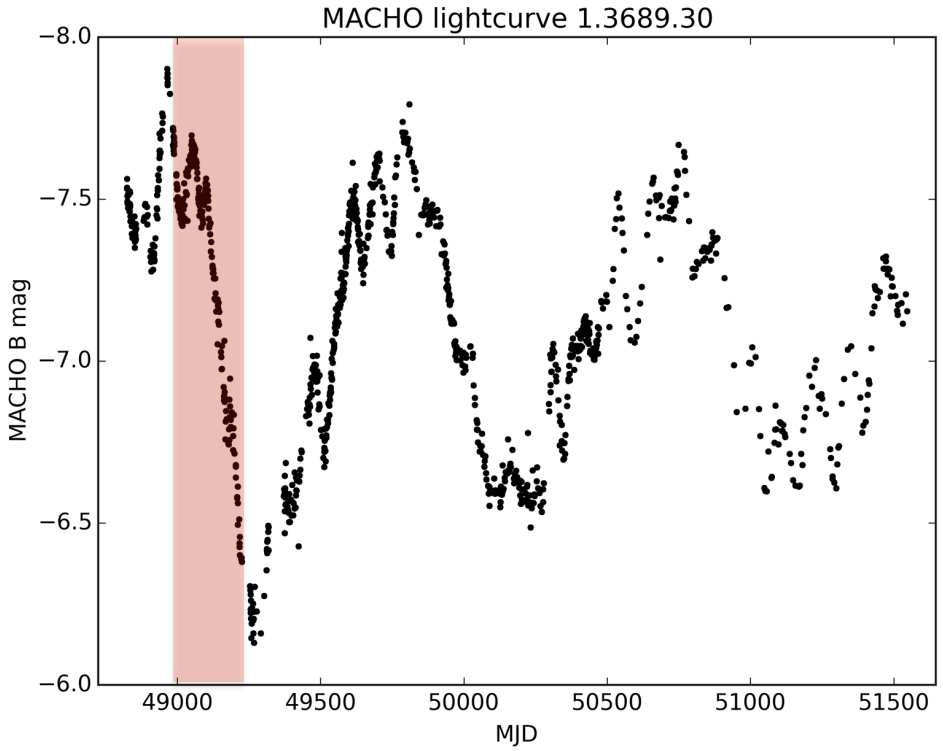


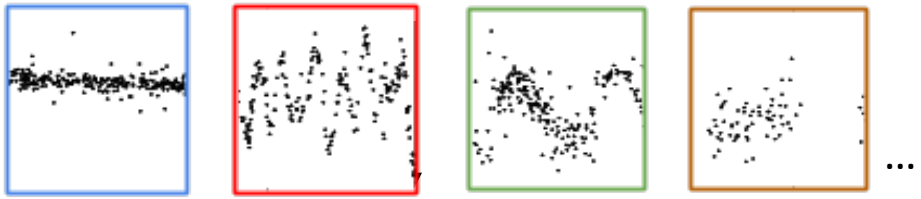


$$1 \quad f(x) \longrightarrow$$

$$\mathbb{R}^K$$

f_1	f_2	f_3	f_4	$\cdots$	f_K

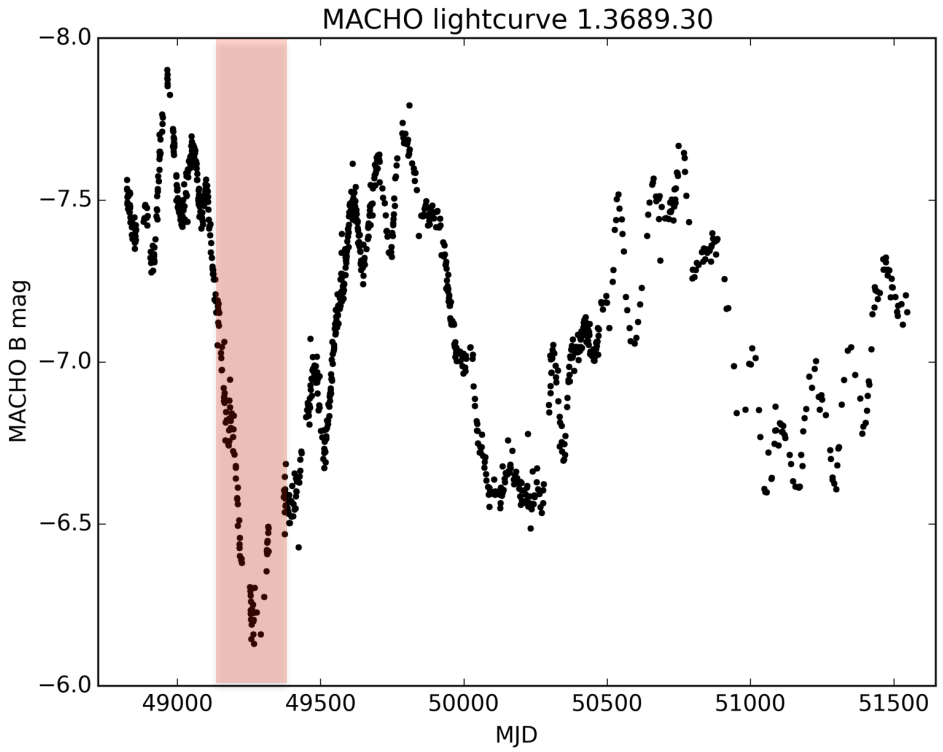


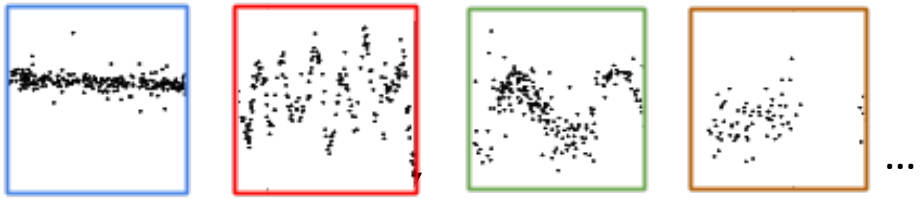


$$2 \quad f(x) \longrightarrow$$

$$\mathbb{R}^K$$

f_1	f_2	f_3	f_4	$\cdots$	f_K
f_1	f_2	f_3	f_4	$\cdots$	f_K

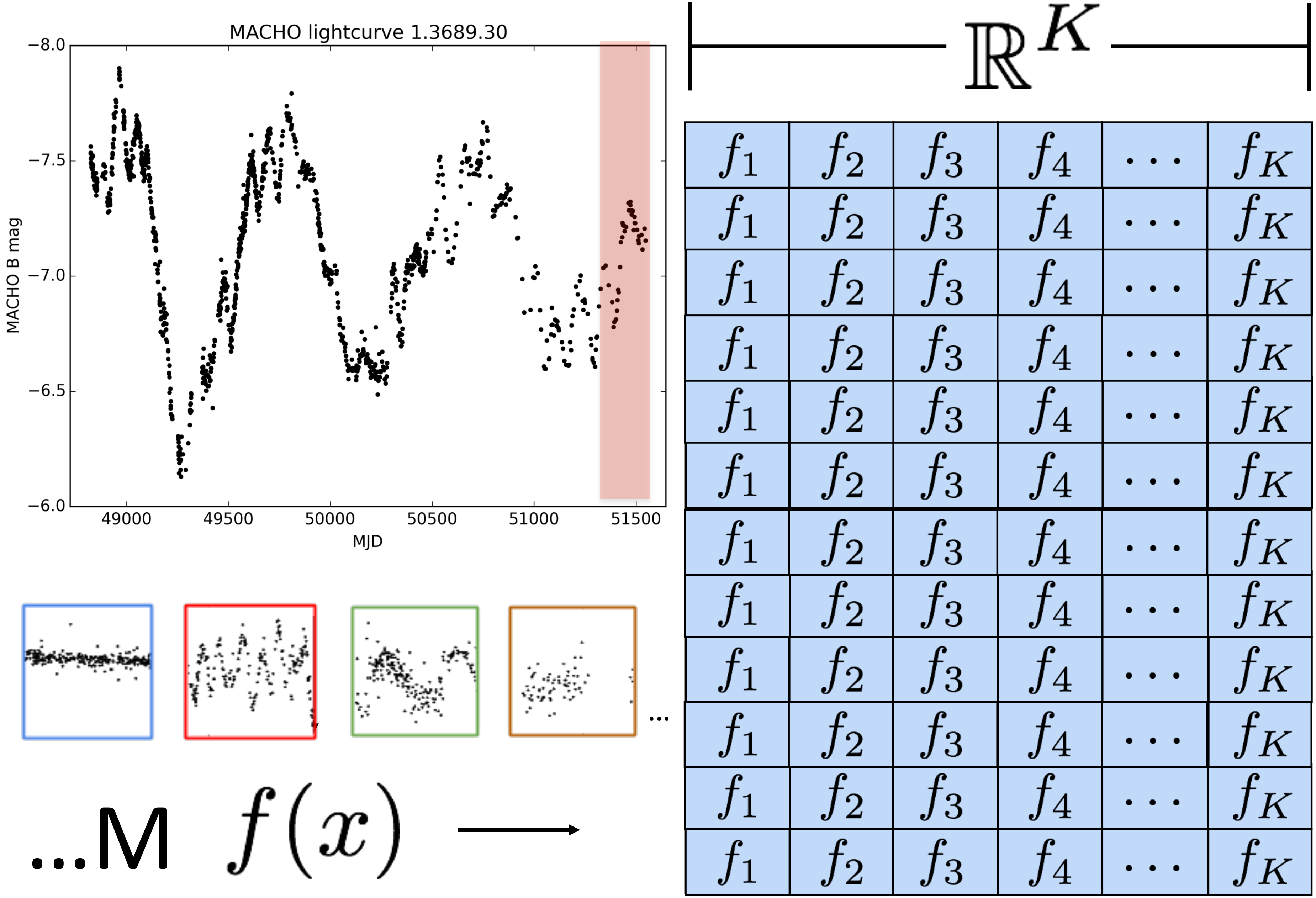




$$3 \quad f(x) \longrightarrow$$

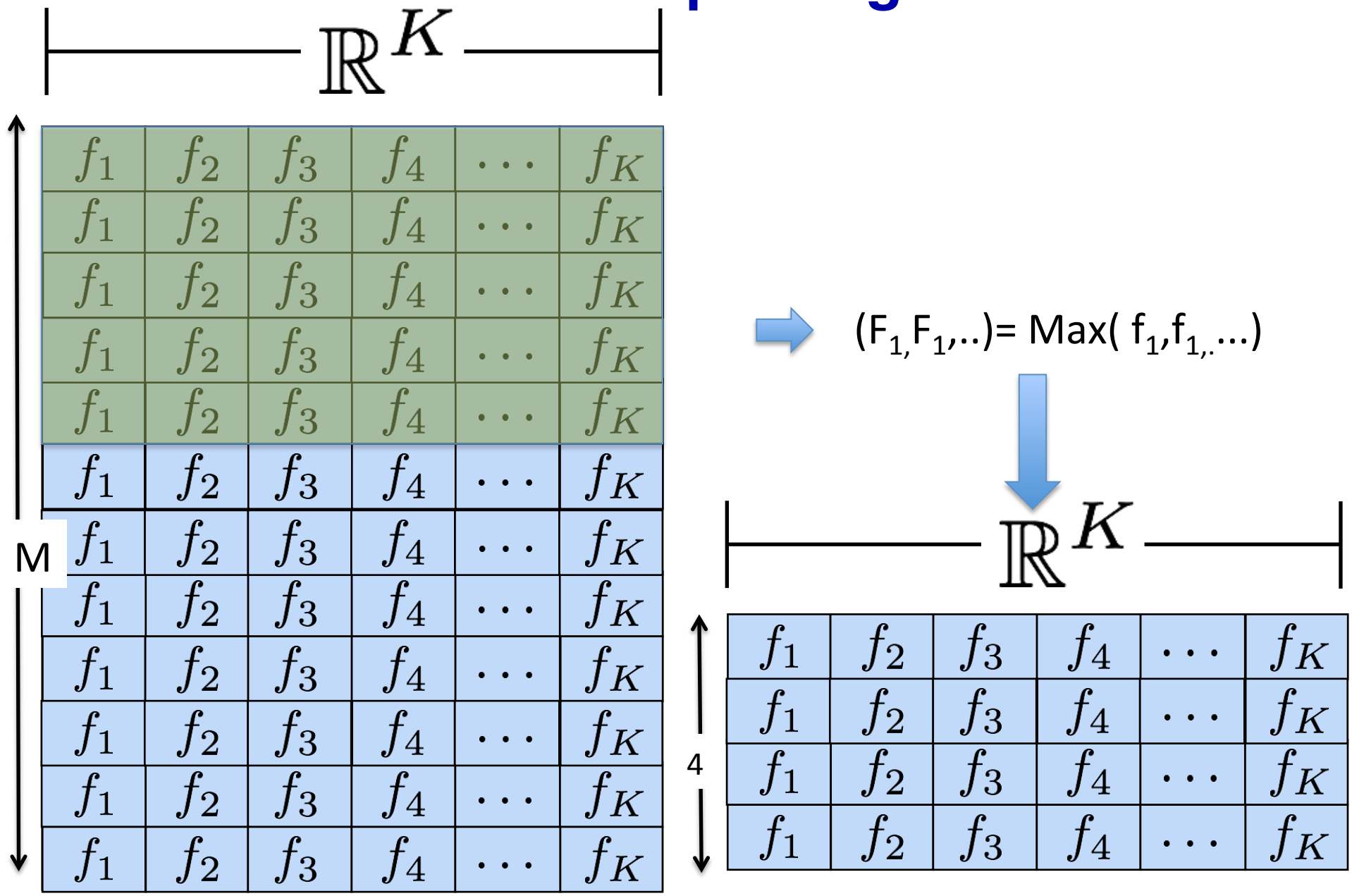
$$\mathbb{R}^K$$

f_1	f_2	f_3	f_4	$\cdots$	f_K
f_1	f_2	f_3	f_4	$\cdots$	f_K
f_1	f_2	f_3	f_4	$\cdots$	f_K



MACHO lightcurve 1.3689.30
MACHO B mag
−8.0
−7.5
−7.0
−6.5
−6.0
49000
49500
50000
50500
51000
51500
MJD
...M
$f(x)$
$\mathbb{R}^K$
f_1 f_2 f_3 f_4 $\cdots$ f_K

Feature pooling

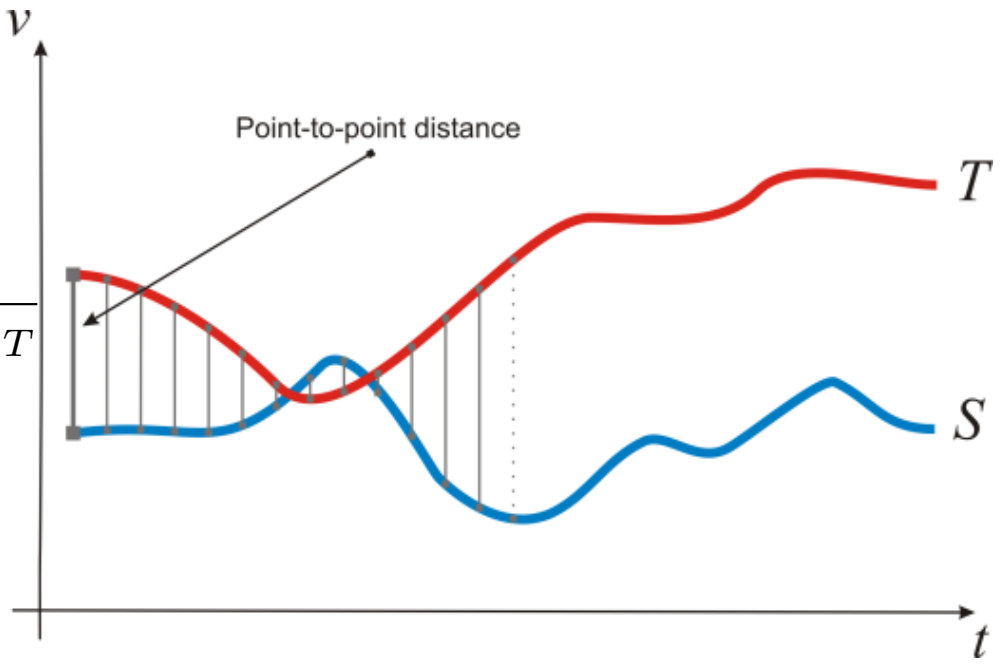


Distance

Clustering small snapshots into similar groups requires measuring their similarity or distance

Euclidean distance

$$d(T,S) = \sqrt{(T-S)(T-S)^T}$$



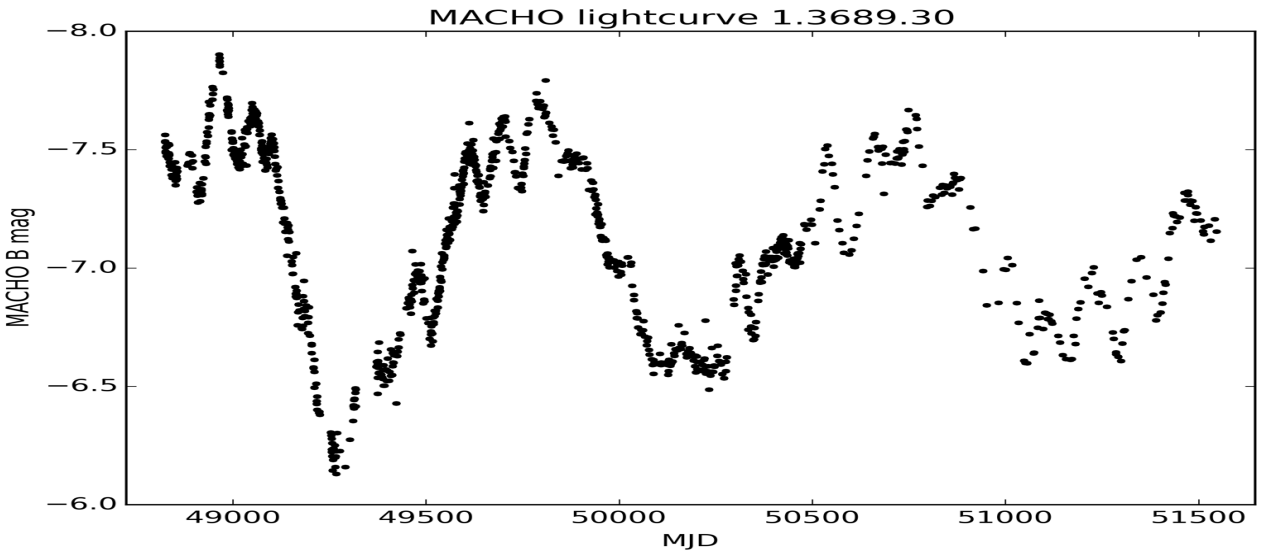
Elastic measure

Astronomical lightcurves are unevenly sampled

=> Euclidean distance is not even well defined for the comparison of these kind of data.

“elastic measures” tolerates uneven sampling and time series of different length

Time Warp Edit Distance (Marteau 2009) to be one of the most powerful and flexible for the case of unevenly sampled time series.



Edit Distance

We use Time Warp Edit Distance as the similarity measure for lightcurves in our experiments.

The Time Warp Edit Distance is based on the Levenshtein Distance (Levenshtein 1966), commonly known as **Edit Distance**

The basic idea behind **Edit Distance**. Compare pairs of values between two time series

For each *delete*, *insert* and *match* you pay a penalty Γ (this is user defined) and we minimize the cost in a greedy way.

$$\delta(X_1^p, Y_1^q) = min \begin{cases} \delta(X_1^{p-1}, Y_1^q) + \Gamma(x_p \to \Lambda) & delete \\ \delta(X_1^{p-1}, Y_1^{q-1}) + \Gamma(x_p \to y_q) & match \\ \delta(X_1^p, Y_1^{q-1}) + \Gamma(\Lambda \to y_q) & insert \end{cases}$$

Time Warp Edit Distance

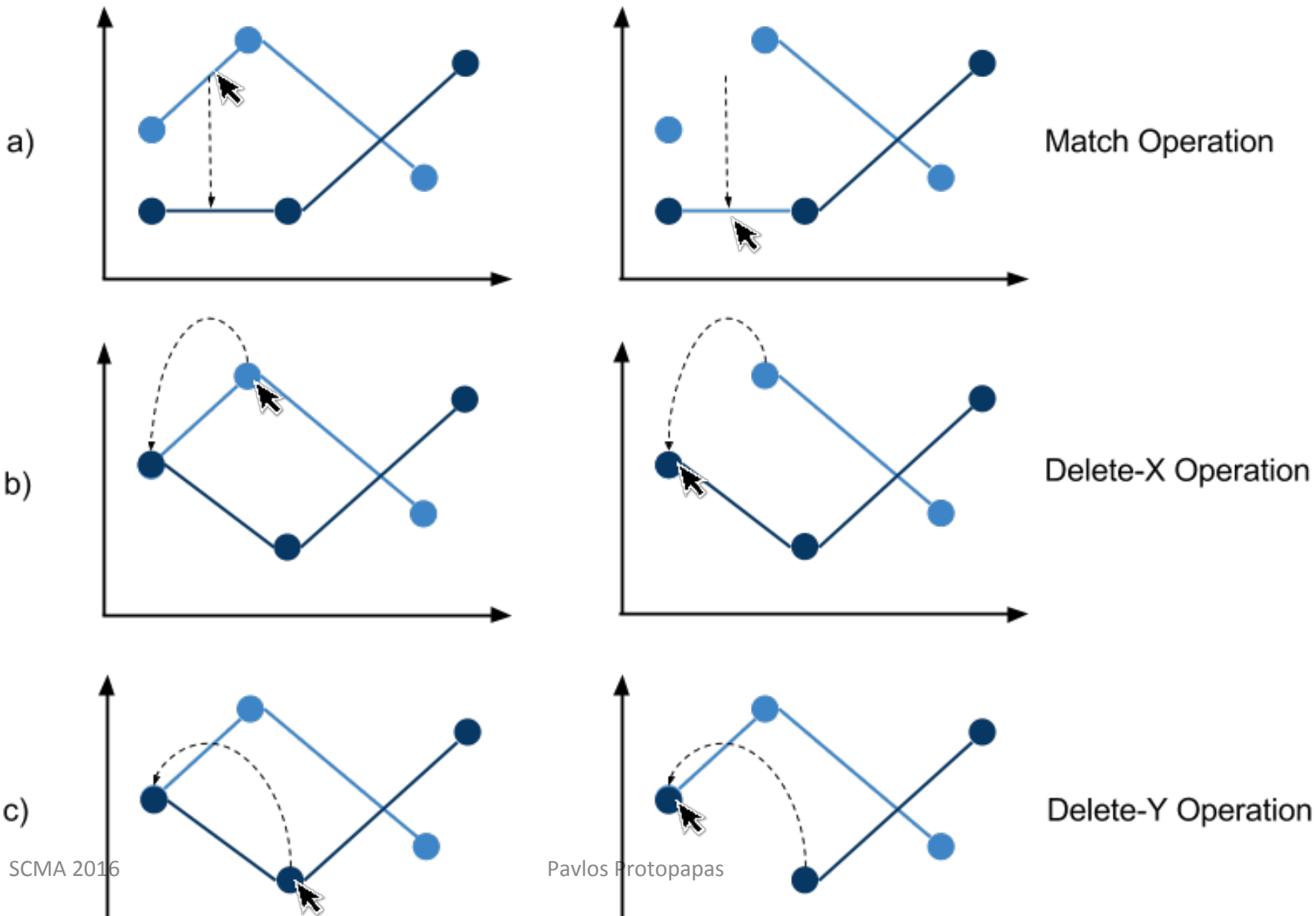
$$\delta_{\lambda,\gamma}(X_1^p, Y_1^q) = min \begin{cases} \delta_{\lambda,\gamma}(X_1^{p-1}, Y_1^q) + \Gamma_{\mathbf{x}} & del - X \\ \delta_{\lambda,\gamma}(X_1^{p-1}, Y_1^{q-1}) + \Gamma_{\mathbf{xy}} & match \\ \delta_{\lambda,\gamma}(X_1^p, Y_1^{q-1}) + \Gamma_{\mathbf{y}} & del - Y \end{cases}$$

$$\Gamma_{\mathbf{x}} = |m_{x_p} - m_{x_{p-1}}| + \gamma |t_{x_p} - t_{x_{p-1}}| + \lambda$$

Where m's are the brightness of the stars, t's are the time of operations and X,Y the pairs (m,t)

λ and γ are user specified parameters

Time Warp Edit Distance



Affinity Propagation (Frey & Dueck 2007)

Each data point is represented as a node in a network

Recursively transmits real-valued messages along the edges of the network until a satisfactory set of exemplar points emerges.

The magnitude of the transmitted messages reflects the “affinity” that one data point has for choosing another point as its exemplar and this is related to their similarity

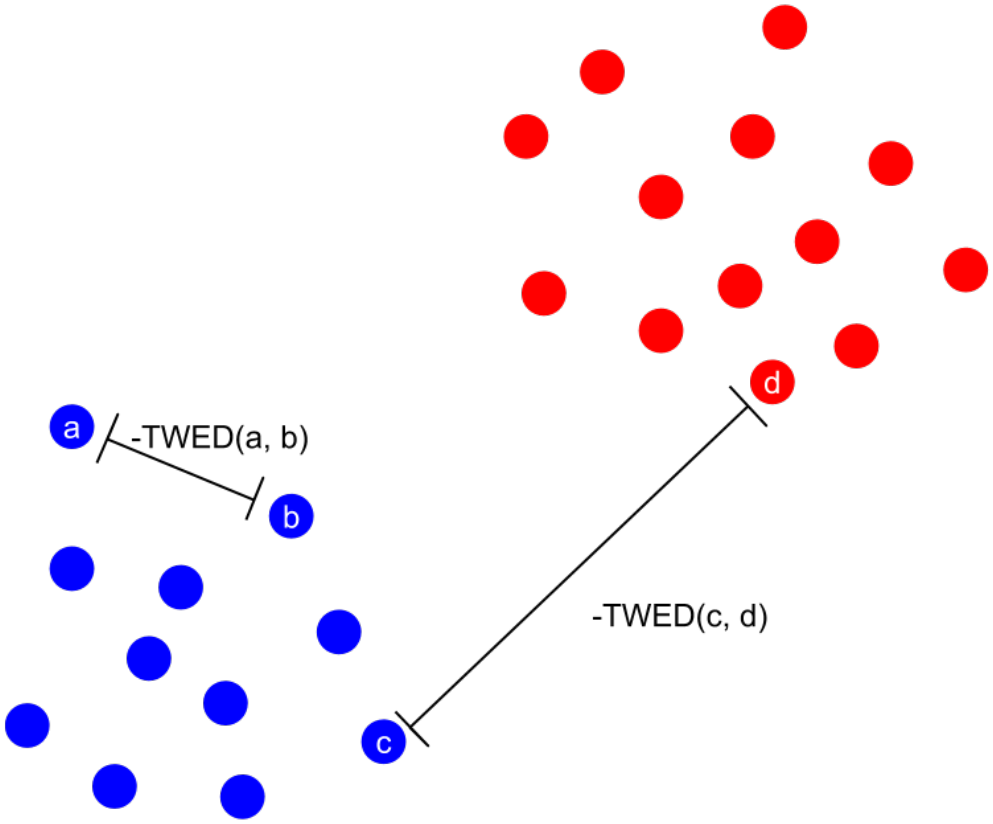
Data points exchange two different kinds of messages during clustering:

Responsibility r(i, k): accumulated evidence for how good point k is to serve as an exemplar to point i

Availability a(i,k): reflects how appropriate it would be for point i to choose point k as its exemplar.

Affinity Propagation

\



	Class	Number of Objects
1	Non Variable	3613
2	Quasars	17
3	Be Stars	55
4	Cepheid	103
5	RR Lyrae	551
6	Eclipsing Binaries	42
7	MicroLensing	173
8	Long Period Variable	281

TABLE 1: MACHO Training Set Composition

	Class	Number of Objects
1	Cepheid	992
2	Type 2 Cepheid	476
3	RR Lyrae	971
4	Eclipsing Binaries	982
5	Delta Scuti	980
6	Long Period Variable	957

TABLE 2: OGLE-III Training Set Composition.

	Class	TSF F-Score	LF F-Score
1	Non Variable	0.875	0.991
2	Quasars	0.217	0.296
3	Be Stars	0.625	0.717
4	Cepheid	0.936	0.871
5	RR Lyrae	0.797	0.953
6	Eclipsing Binaries	0.725	0.780
7	MicroLensing	0.468	0.980
8	Long Period Variable	0.802	0.975
	Weighted Average	0.807	0.975

TABLE 4: Classification performance on the MACHO training set using SVM.

	Class	TSF F-Score	LF F-Score
1	Cepheid	0.555	0.835
2	Type 2 Cepheid	0.467	0.651
3	RR Lyrae	0.649	0.749
4	Eclipsing Binaries	0.458	0.862
5	Delta Scuti	0.656	0.817
6	Long Period Variable	0.407	0.821
	Weighted Average	0.696	0.821

TABLE 5: Classification performance on the OGLE-III

Summary

It is feasible to classify everything

Unsupervised feature extraction that works as well for classification as traditional features

We have

- Classification models
- Outlier detection
- Meta-classification models

We need

- Training sets.
  - Transfer learning
  - Active learning
  - Experimental Design