Mario Pasquato
Marie Skłodowska-Curie Fellow
Padua Observatory
(Italy)

Michela Mapelli
Abbas Askar
Ammar Askar
Mario Spera
Federico Abbate
Piero Trevisan

Seeing in the dark

spotting dark remnant subsystems and intermediate mass black holes in star clusters with machine learning

MWSTREAMS Conference 20-11-2018 Heidelberg

Three projects

Predict what	Problem type	Learn on	Simulations	Paper
IMBH host? Yes/No you remember it if you were in Santorini	Classification	Density profiles	~2000 MOCCA Survey Database I	Pasquato et al. 2018 MNRAS (submitted, 1st ref. rep)
BH subsystem host? Yes/No you saw it on arXiv few days ago	Classification	Structural parameters	~2000 MOCCA survey Database I	Askar et al. 2018 MNRAS (submitted)
IMBH mass	Regression	Pulsar accelerarions, jerks	~200 N-body with hiGPUs	Pasquato & Spera 2018 (in prep.)

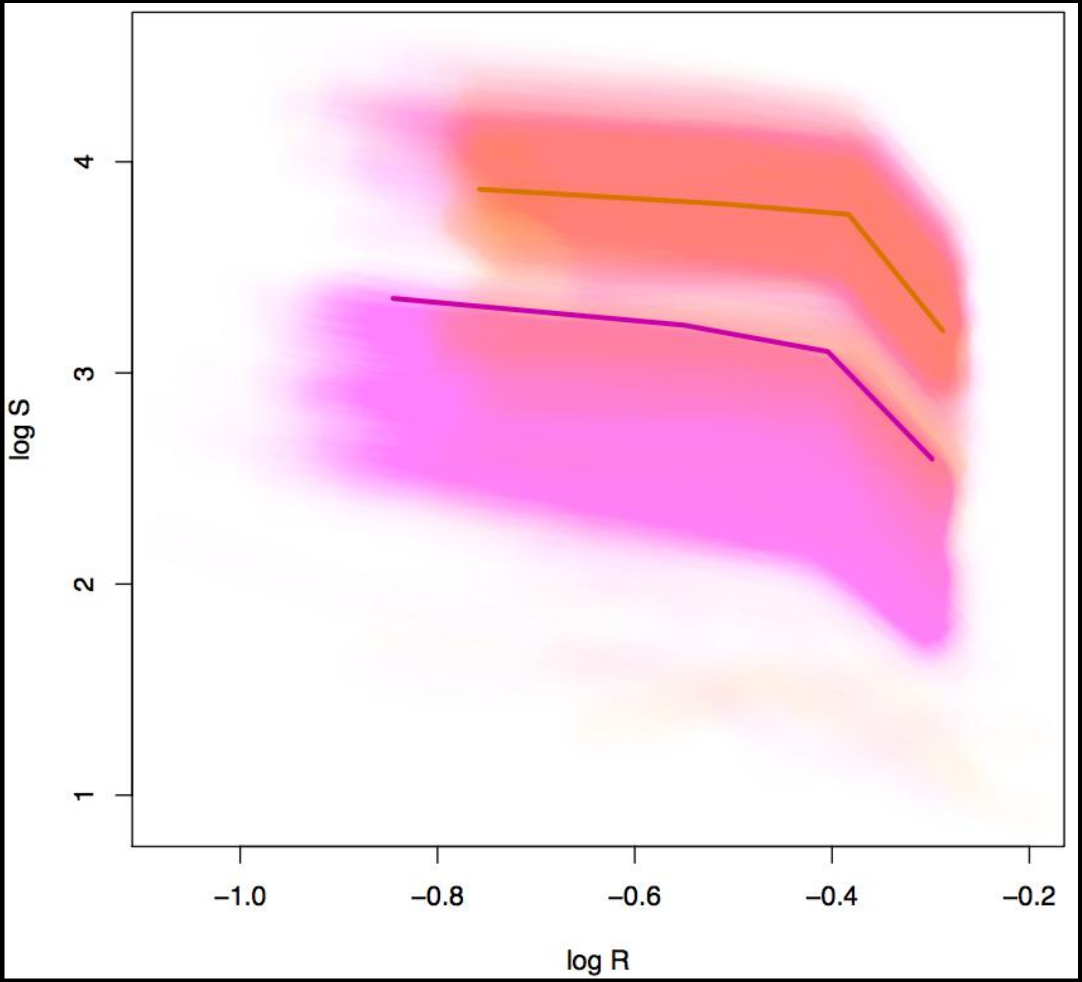
Three projects

Predict what	Problem type	Learn on	Simulations	Paper
IMBH host? Yes/No you remember it if you were in Santorini	Classification	Density profiles	~2000 MOCCA Survey Database I	Pasquato et al. 2018 MNRAS (submitted, 1st ref. rep)
BH subsystem host? Yes/No you saw it on arXiv few days ago	Classification	Structural parameters	~2000 MOCCA survey Database I	Askar et al. 2018 MNRAS (submitted)
IMBH mass	Regression	Pulsar accelerarions, jerks	~200 N-body with hiGPUs	Pasquato & Spera 2018 (in prep.)

Three projects

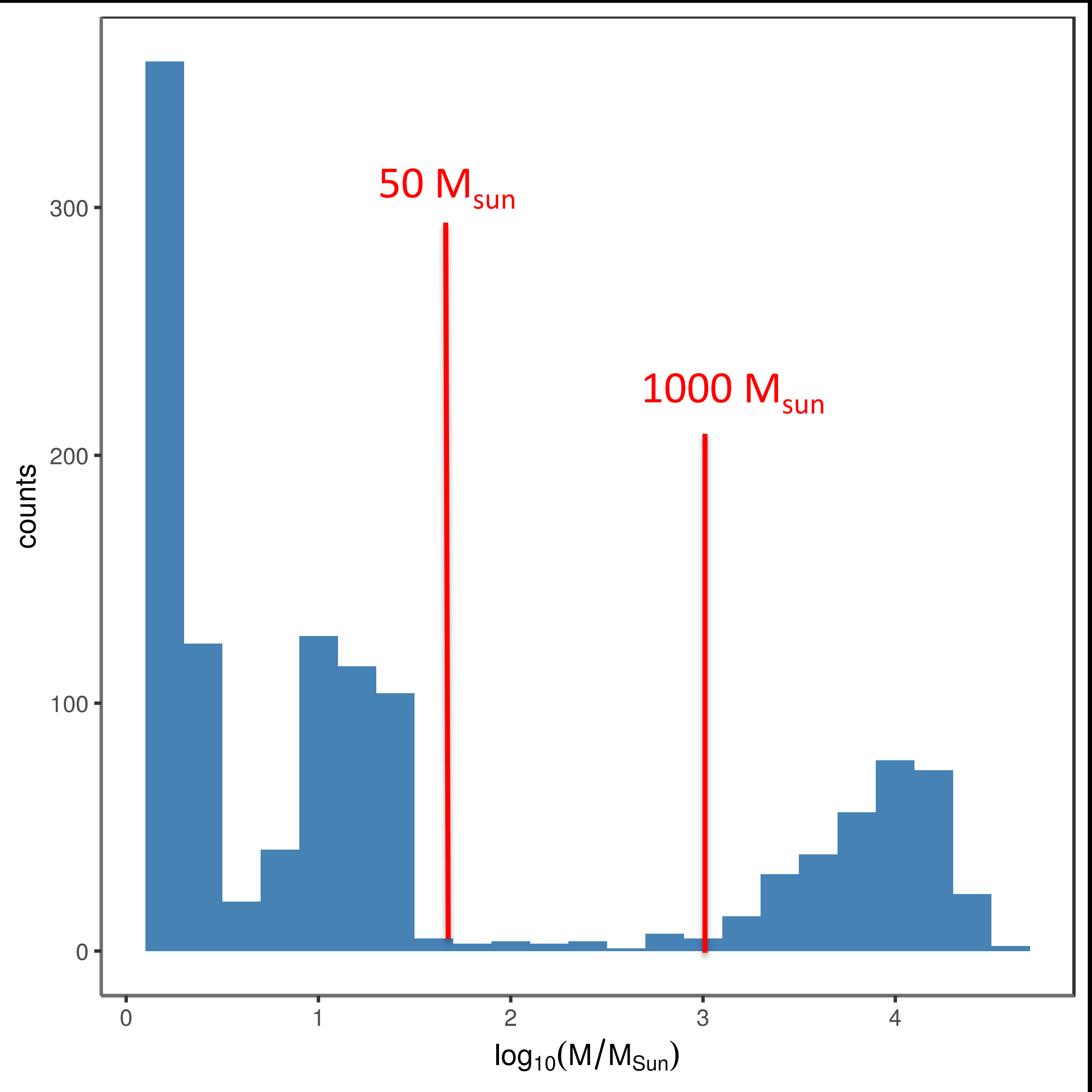
Predict what	Problem type	Learn on	Simulations	Paper
IMBH host? Yes/No you remember it if you were in Santorini	Classification	Density profiles	~2000 MOCCA Survey Database I	Pasquato et al. 2018 MNRAS (submitted, 1st ref. rep)
BH subsystem host? Yes/No you saw it on arXiv few days ago	Classification	Structural parameters	~2000 MOCCA survey Database I	Askar et al. 2018 MNRAS (submitted)
IMBH mass	Regression	Pulsar accelerarions, jerks	~200 N-body with hiGPUs	Pasquato & Spera 2018 (in prep.)

I. MOCCA simulations + feature based approach = IMBH Yes/No?



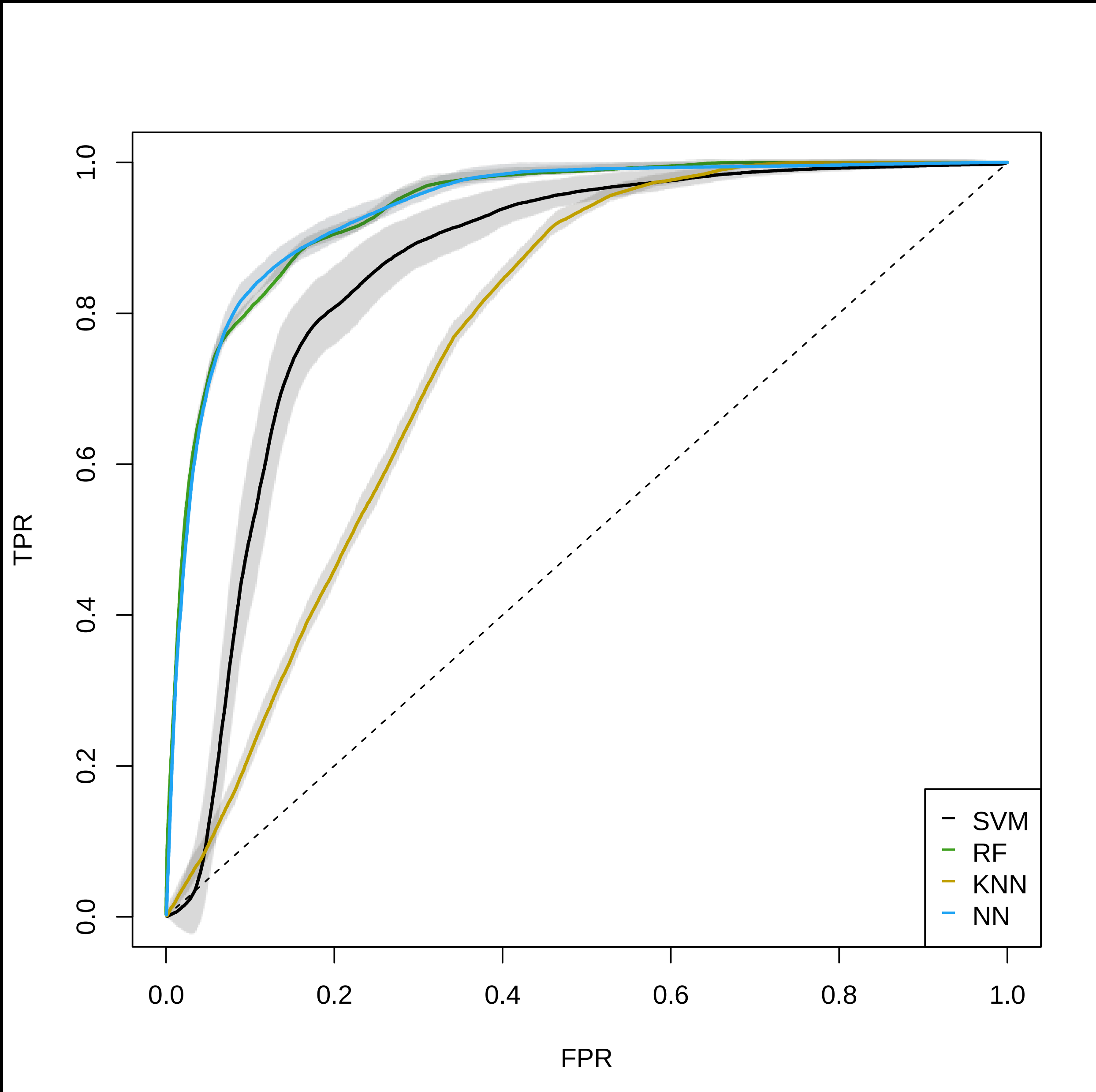
- 2000 MOCCA Survey Database I simulations
- Numeric features for learning: surface density profiles
- IMBH host / non host

Classification approach



- Simulation -> Density profile = 8 numbers
- Classify into IMBH host / non-host
- Test on unseen simulations

ROC curves for our classifieres



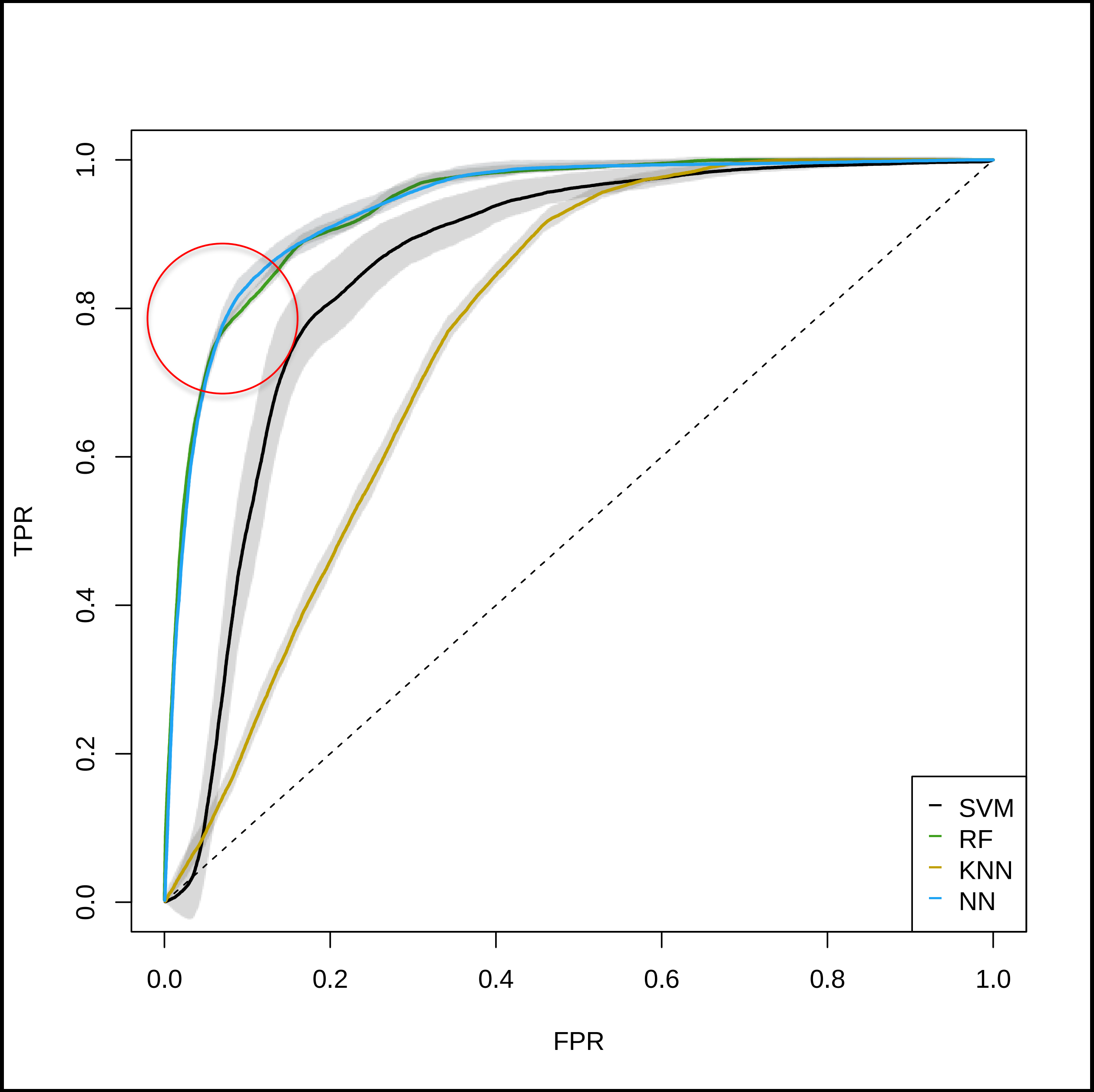
Model	AUC
Neural net	0.94
Random forest	0.94
Support vector machine	0.85
k-nearest neighbor	0.76

Perfomance of the two best classifiers (NN and RF) is very similar

No ad hoc tuning of classifier parameters

Performance measured on snapshots not seen in training

Interpretation



If we, say, accept 5% false positives
FPR = FP/(FP+TN)

With the random forest or the neural net

we get ~70% recall
i.e. 70% of the actual IMBH hosts are actually found
TPR = TP/(TP+FN)

Results

- Four classifiers: the best two (neural net, random forest) have very similar ROC curves
- At 5% FPR they yield 70% TPR
- Scenarios
  - IMBH prevalence 50%
  - IMBH prevalence 10%

Real hosts	Real non-hosts	False Positive Rate	True Positive Rate (Recall)	Claimed hosts	Correct claims	Correct claims / total (Precision)
100	100	5%	70%	75	70	93%
20	180	5%	70%	23	14	61%

To summarize

Based only on surface density profiles,
within the MOCCA Survey database,
our classifiers
without fine tuning,
on snapshots not seen in training,
catch ~70% of IMBH hosts
with a 5% false positive rate

II. MOCCA Sim. + Structural parameters as features = BH subsystem Yes/No?

Askar, Askar, Pasquato, Giersz 2018 MNRAS submitted; on arXiv, do read it, it's interesting

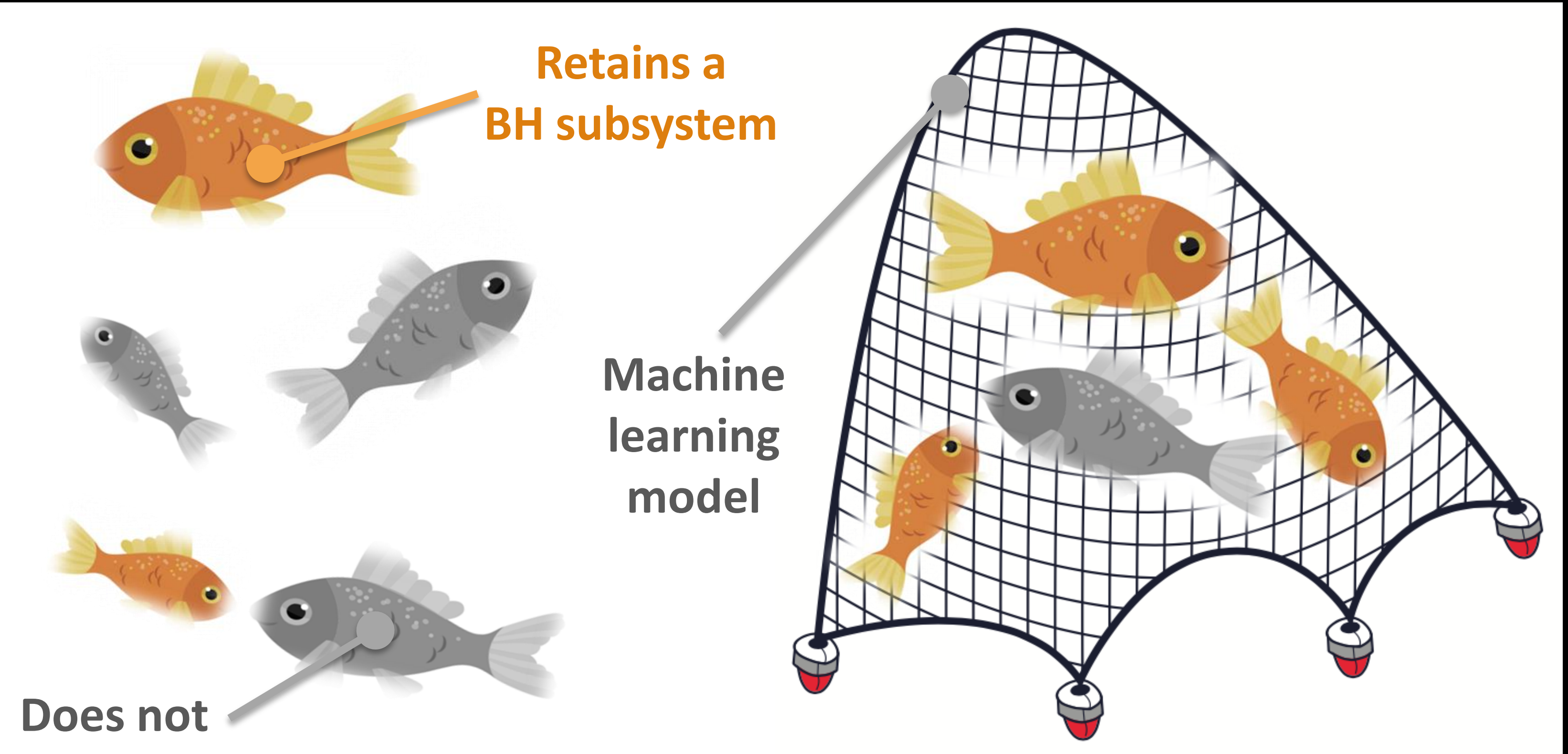
- Half-Light Radius
- Central Surface Brightness
- Central Velocity Dispersion
- Total Luminosity
- Relaxation Time
- Core Radius

Who retains a black hole subsystem?

BH subsystem -> GW, dynamical heating...

Catching BH subsystem hosts

- Which initial conditions + evolutionary history (as reflected by the structural parameters) lead to retaining a BH subsystem?



Desired model properties

- Catches (almost) only real BH subsystem hosts
- Interpretable

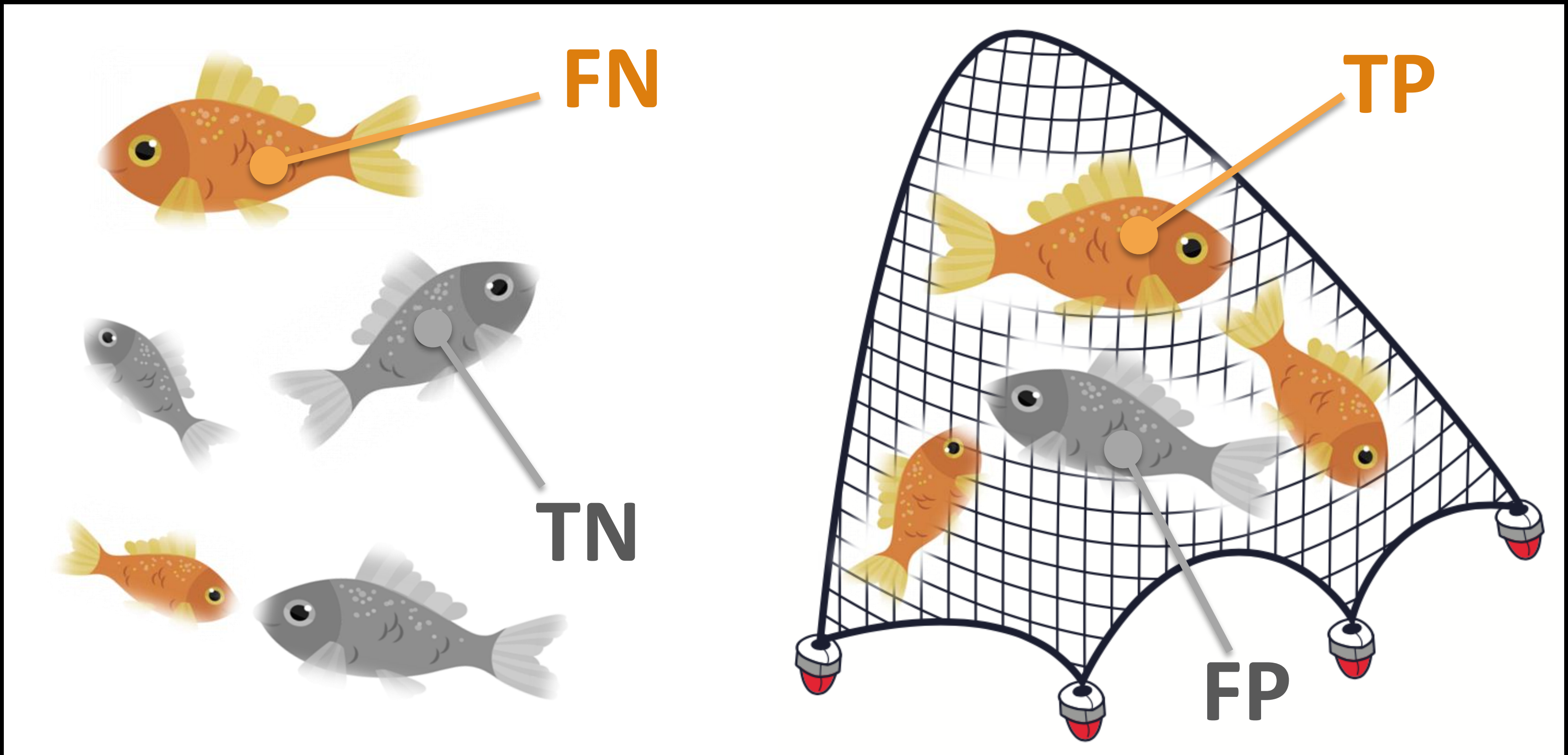
Desired model properties

- Catches (almost) only real BH subsystem hosts
- Interpretable

Performance metric
should weigh
precision

Performance metrics

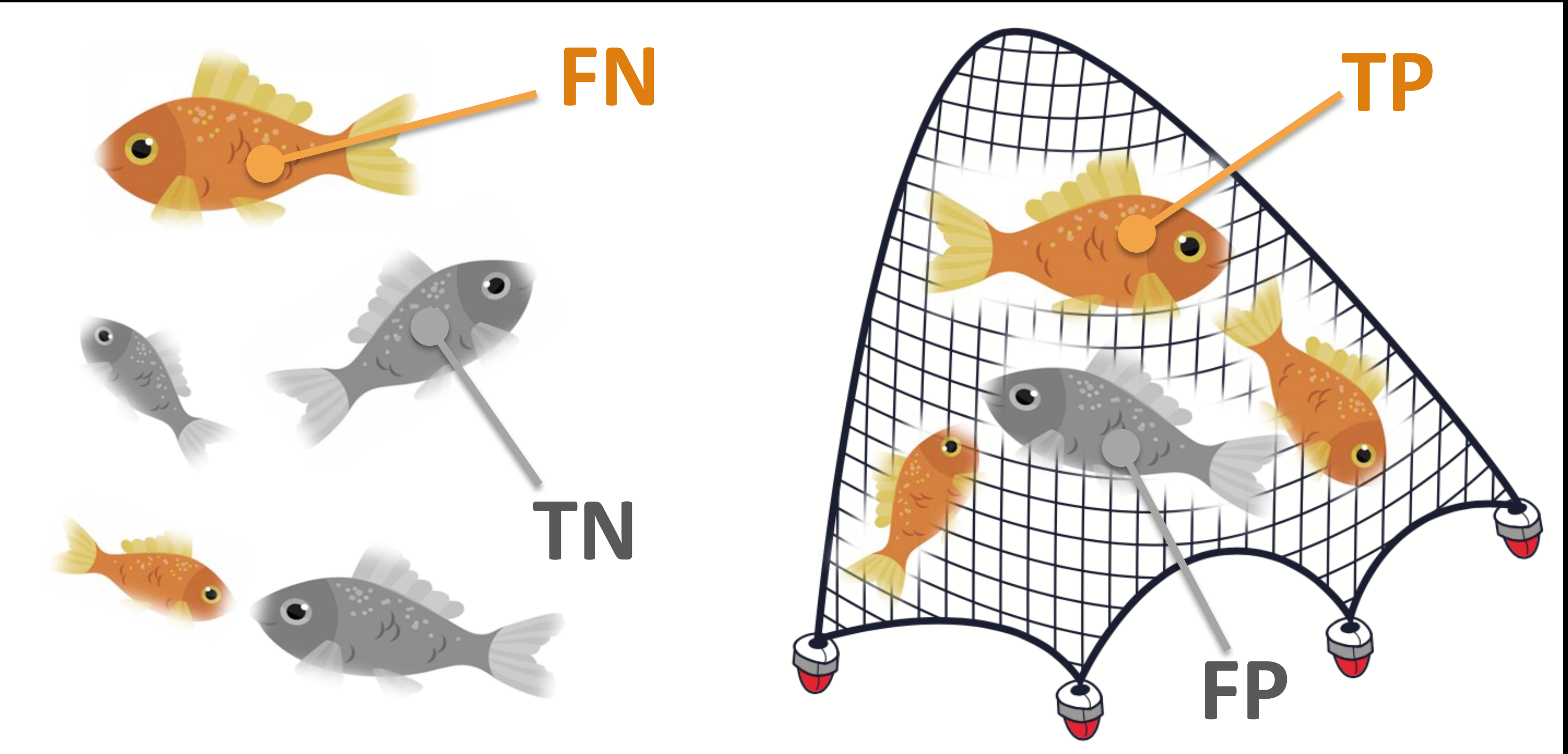
- Precision TP/(TP+FP) how clean is the catch?
- True Positive Rate or Recall TP/(TP+FN) how big is the catch?
- F-score = 1/(1/Precision + 1/Recall)



Performance metrics

- Precision 3/(3+1) = 3/4
- Recall 3/(3+2) = 3/5
- F-score = 1/(4/3 + 5/3) = 1/3

Example with numbers from this pic

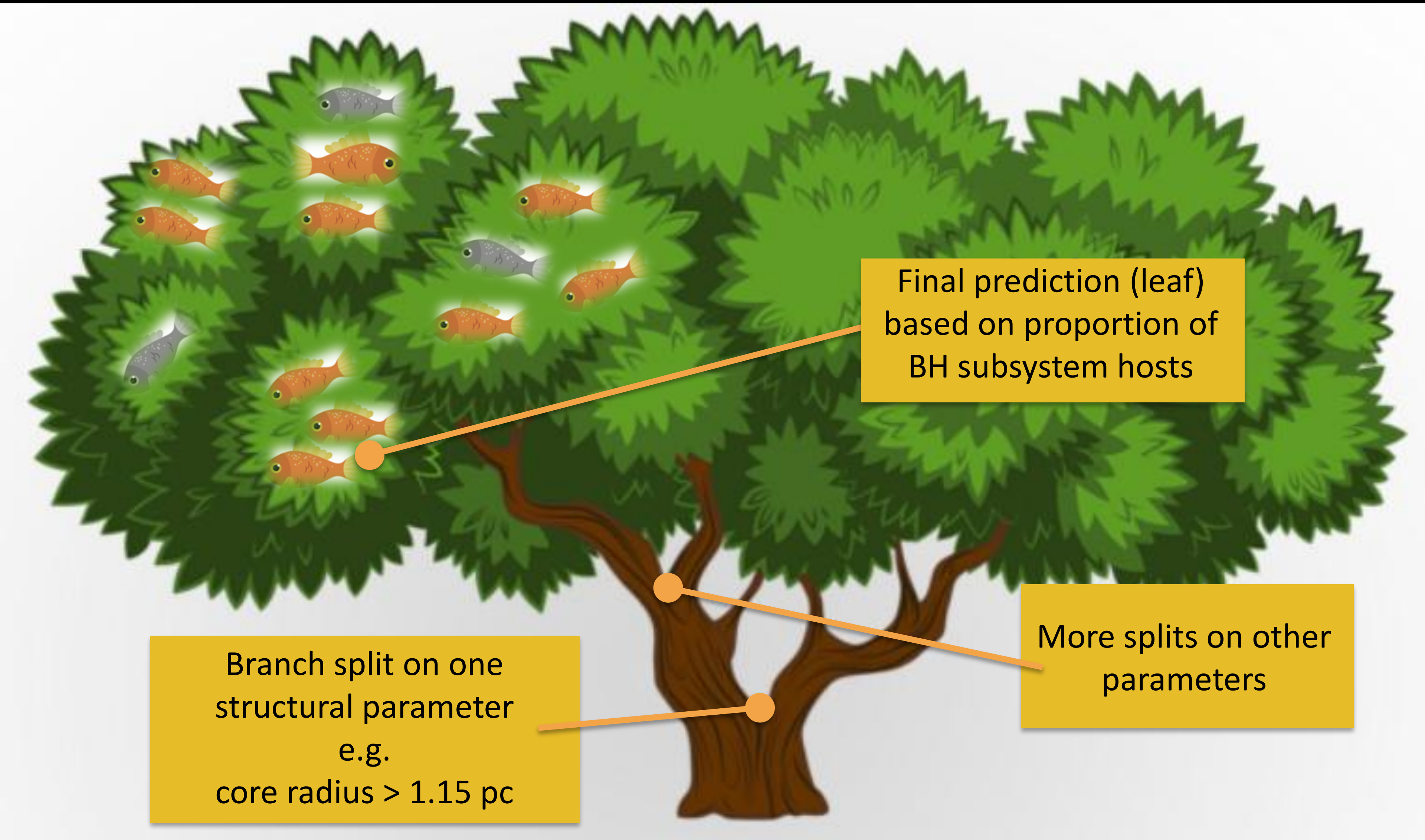


Desired model properties

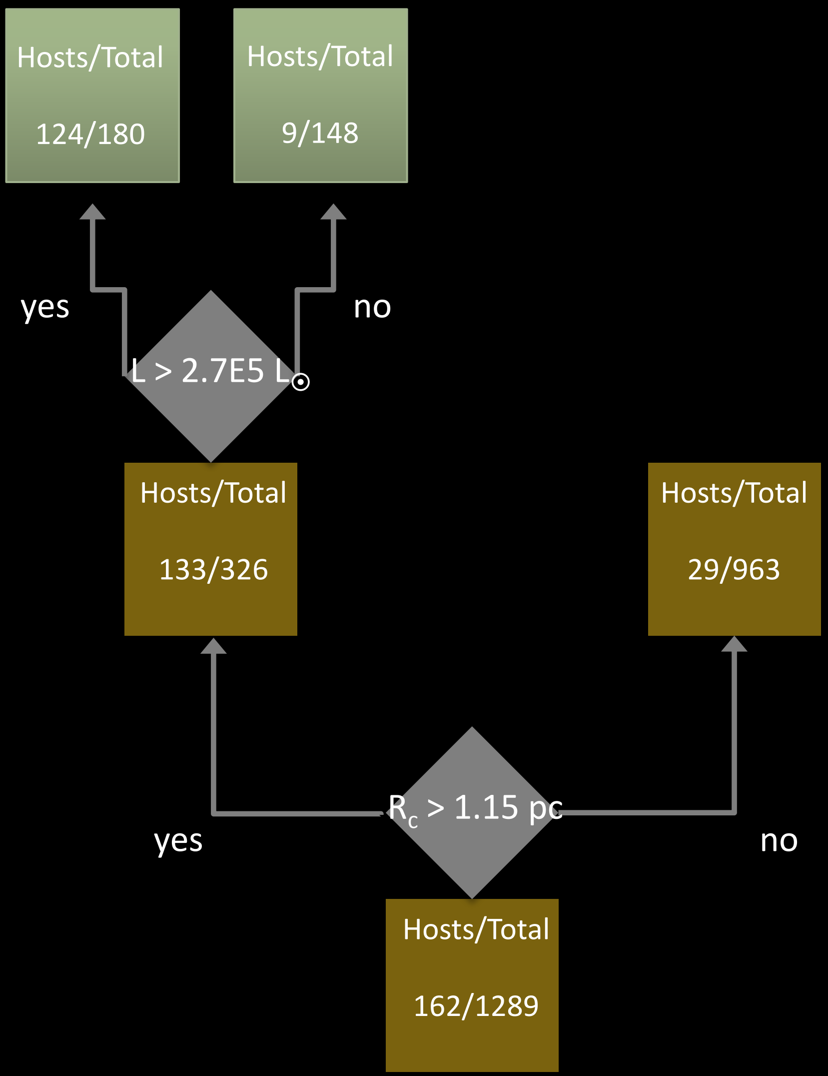
- Catches (almost) only real BH subsystem hosts
- Interpretable

What makes a star cluster
a BH subsystem host?

Tree-based model



Physical interpretation



First few branches of the learned tree

First split is on core radius: black hole subsystem hosts have large cores due to dynamical heating

Second split on total luminosity: big clusters produce more black holes, have higher retention due to higher escape velocity

Tree models also have good performance

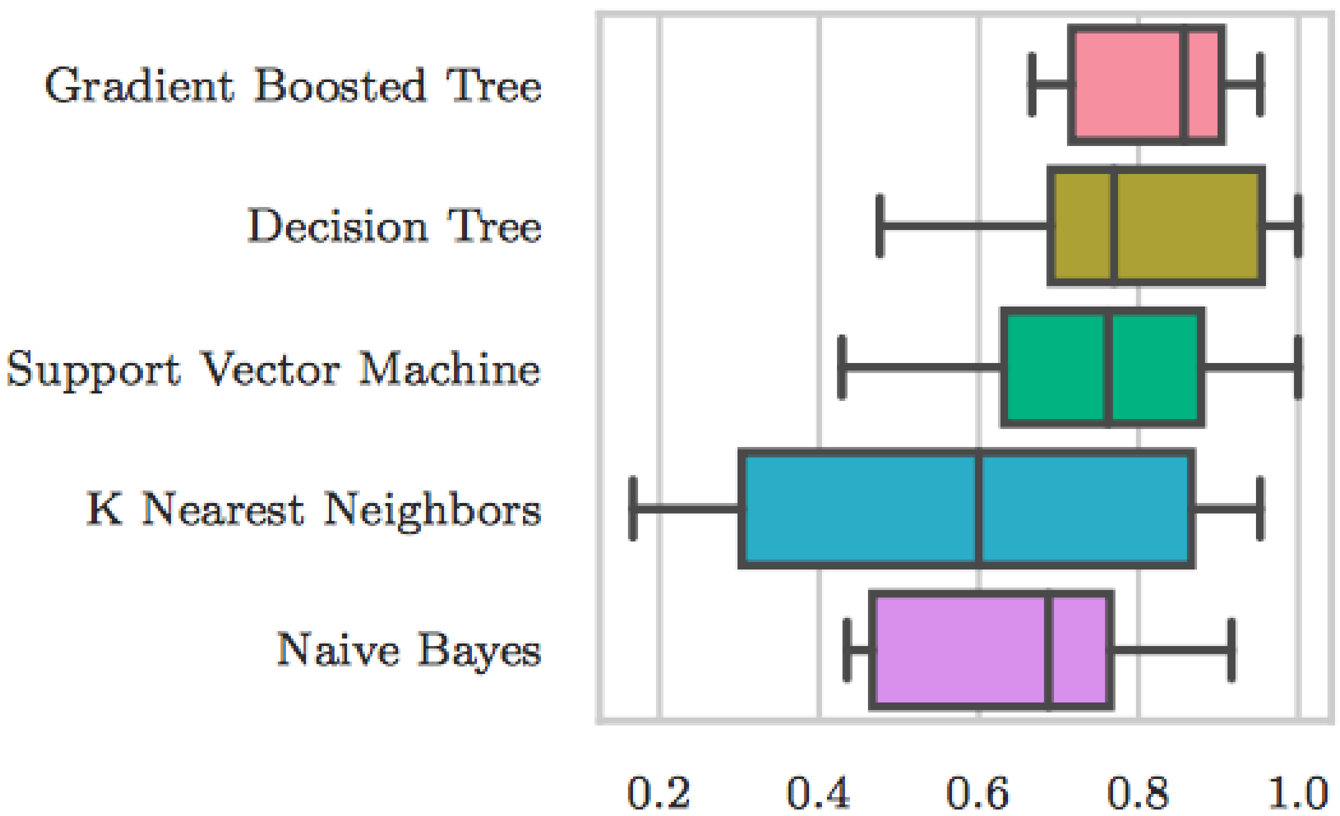


Figure 4. Comparison of each classifiers' f-score with 15-fold testing. An f-score of 1.0 is the best possible, 0.0 is the worst. The score is affected by how well the classifier can find all the BH subsystems and whether the identifications are false-positives.

Comparison with real GCs

Green row = predicted BH subsystem host by all models e.g. NGC 288, M10

Results compare well with other methods (Askar et al. marked with *)

Table 4. Predictions from the Harris (1996, *updated 2010*) and Baumgardt & Hilker (2018) datasets using the gradient boosted decision tree classifier. Entires where BHS presence was classified positively are shown. The *BHS* column represents the classifier trained on all simulation data whereas *Fallback* represents training on models where mass fallback was enabled and BH natal kicks were lower.

Cluster Name	BHS (Harris)	Fallback (Harris)	BHS (B&H)	Fallback (B&H)
IC 4499 *	✓	✓	✓	✓
NGC 288 *	✓	✓	✓	✓
NGC 3201 *	✓	✓	✗	✓
NGC 4372 *†	✓	✓	✓	✓
NGC 4590 (M68)	✗	✓	✗	✗
NGC 4833 *†	✗	✓	✓	✓
NGC 5139 (ω Cen)	✓	✓	✗	✓
NGC 5272 (M3) *	✗	✓	✓	✓
NGC 5286	✗	✓	✗	✓
NGC 5466 *	✗	✓	✗	✗
NGC 5897 *†	✓	✓	✗	✓
NGC 5904 (M5)	✗	✓	✓	✓
NGC 5927	✗	✗	✓	✓
NGC 5986 *†	✓	✓	✗	✓
NGC 6101 *†	✓	✓	✗	✗
NGC 6139 †	✓	✓	✗	✗
NGC 6144 *†	✗	✓	✓	✓
NGC 6205 (M13) *	✗	✓	✓	✓
NGC 6218 (M12)	✓	✓	✗	✗
NGC 6254 (M10)	✓	✓	✓	✓
NGC 6266 (M62)	✗	✗	✓	✗
NGC 6273 (M19) †	✗	✓	✓	✓
NGC 6287 †	✓	✓	✗	✗
NGC 6304 †	✗	✓	✓	✓
NGC 6316 †	✓	✓	✗	✗
NGC 6333 (M9) †	✓	✓	✗	✗
NGC 6356 †	✗	✓	✗	✓
NGC 6362 *	✗	✓	✗	✓
NGC 6380 †	✓	✓	✗	✗
NGC 6388	✗	✗	✓	✗
NGC 6401 *	✓	✓	✗	✗
NGC 6402 (M14) †	✓	✓	✓	✓
NGC 6426 *†	✗	✓	✗	✗
NGC 6440 †	✓	✓	✗	✗
NGC 6496 *†	✗	✓	✗	✗
NGC 6517 †	✗	✓	✗	✗
NGC 6539 (GCL 85)	✓	✓	✗	✗
NGC 6553	✓	✓	✗	✗
NGC 6569 *†	✓	✓	✓	✓
NGC 6584 *†	✓	✓	✗	✗
NGC 6656 (M22) *	✓	✓	✓	✓
NGC 6712 *	✗	✓	✓	✓
NGC 6723 *†	✓	✓	✓	✓
NGC 6760 †	✓	✓	✗	✗
NGC 6779 (M56) *	✗	✓	✓	✓
NGC 6809 (M55) *	✗	✓	✗	✓
NGC 6934 *	✓	✓	✗	✗
NGC 6981 (M72) *	✗	✓	✗	✗
NGC 7078 (M15)	✗	✗	✓	✓
NGC7089 (M2)	✗	✓	✗	✓
Pal11 *†	✓	✓	✓	✓
Terzan5 †	✗	✓	✗	✗

III. Pulsar acceleration and jerk + regression = IMBH mass?

- 196 star cluster simulations with direct N-body code hiGPUs
- Random mass cutoff of the Kroupa IMF, random number of particles, random IMBH mass (from 0 up to 25% of cluster mass)
- 4068 snapshots extracted at random times
- Accelerations (a_z), jerks (a'_z) recorded
- **Question: can we predict IMBH mass with machine learning?**

Variables to use

- I extract k particles at random, call them "pulsars" and assume we know
  - their x, y positions (sky projected)
  - their v_z, a_z, j_z (line of sight)
- I also assume we know the
  - total mass of the cluster
  - projected half-mass radius of the cluster

Predict this

name	M_{GC} (M_{sun})	R_h (pc)	x	y	v_z	a_z	j_z	m_{IMBH} (M_{sun})
10121_.6022081_95.3305/10000003.dat	6780	1.06	24.67	-13.84	0.14	-0.0035	-1.05e-05	600

x k

Train on these variables (features)
on train data (80%)

Predict this
on unseen
test data
(20%)

name	M_{GC} (M_{sun})	R_h (pc)	x	y	v_z	a_z	j_z	m_{IMBH} (M_{sun})
10121_.6022081_95.3305/10000003.dat	6780	1.06	24.67	-13.84	0.14	-0.0035	-1.05e-05	600

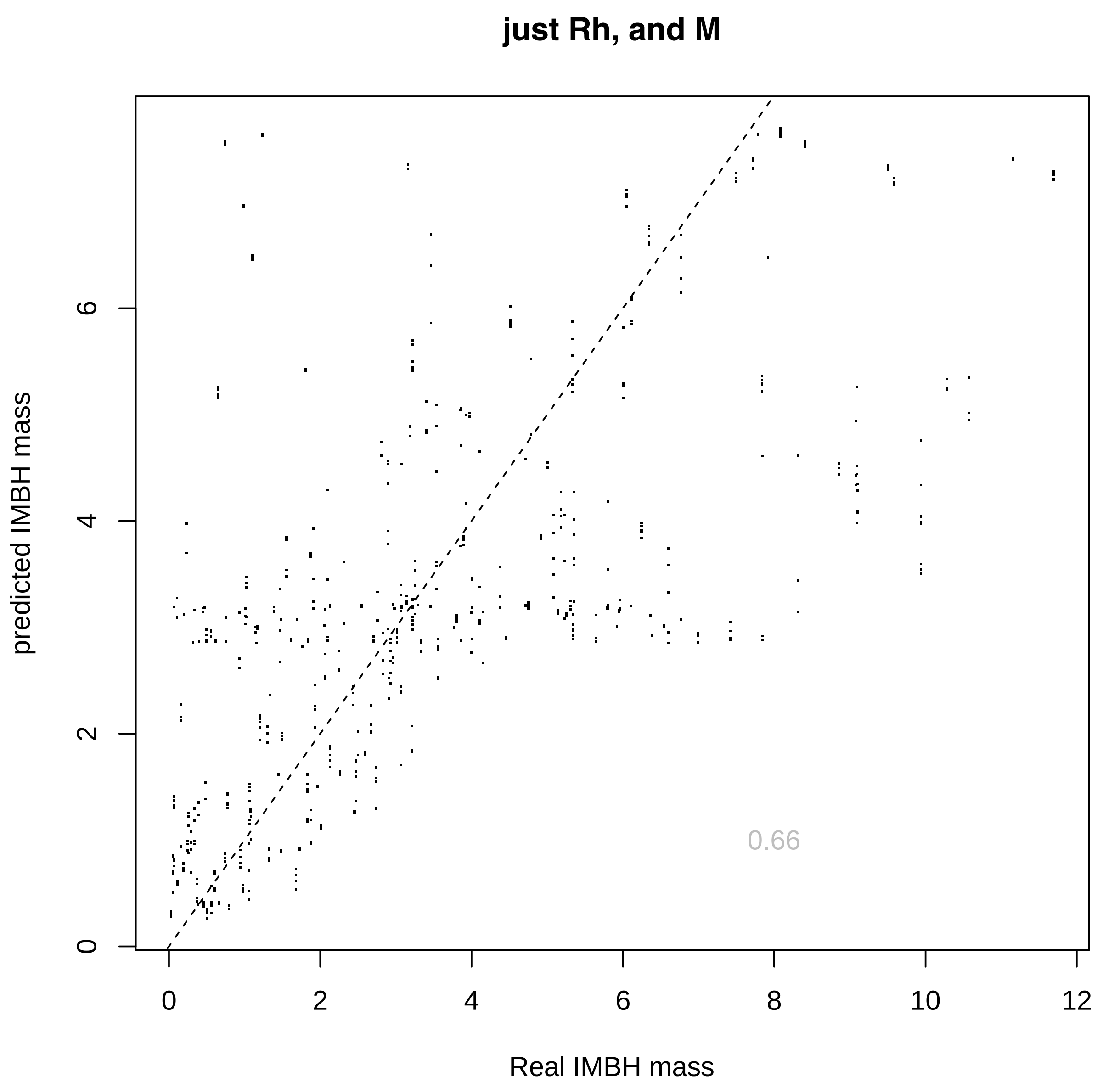
x k

Pulsars appear sorted by modulus of acceleration

Using support vector regression algorithm implemented in R library e1071
Chang, Chih-Chung and Lin, Chih-Jen *LIBSVM: a library for Support Vector Machines*
http://www.csie.ntu.edu.tw/~cjlin/libsvm

Hastie Tibshirani Friedman *The elements of statistical learning*, Chapter 12.3

With only R_h and M



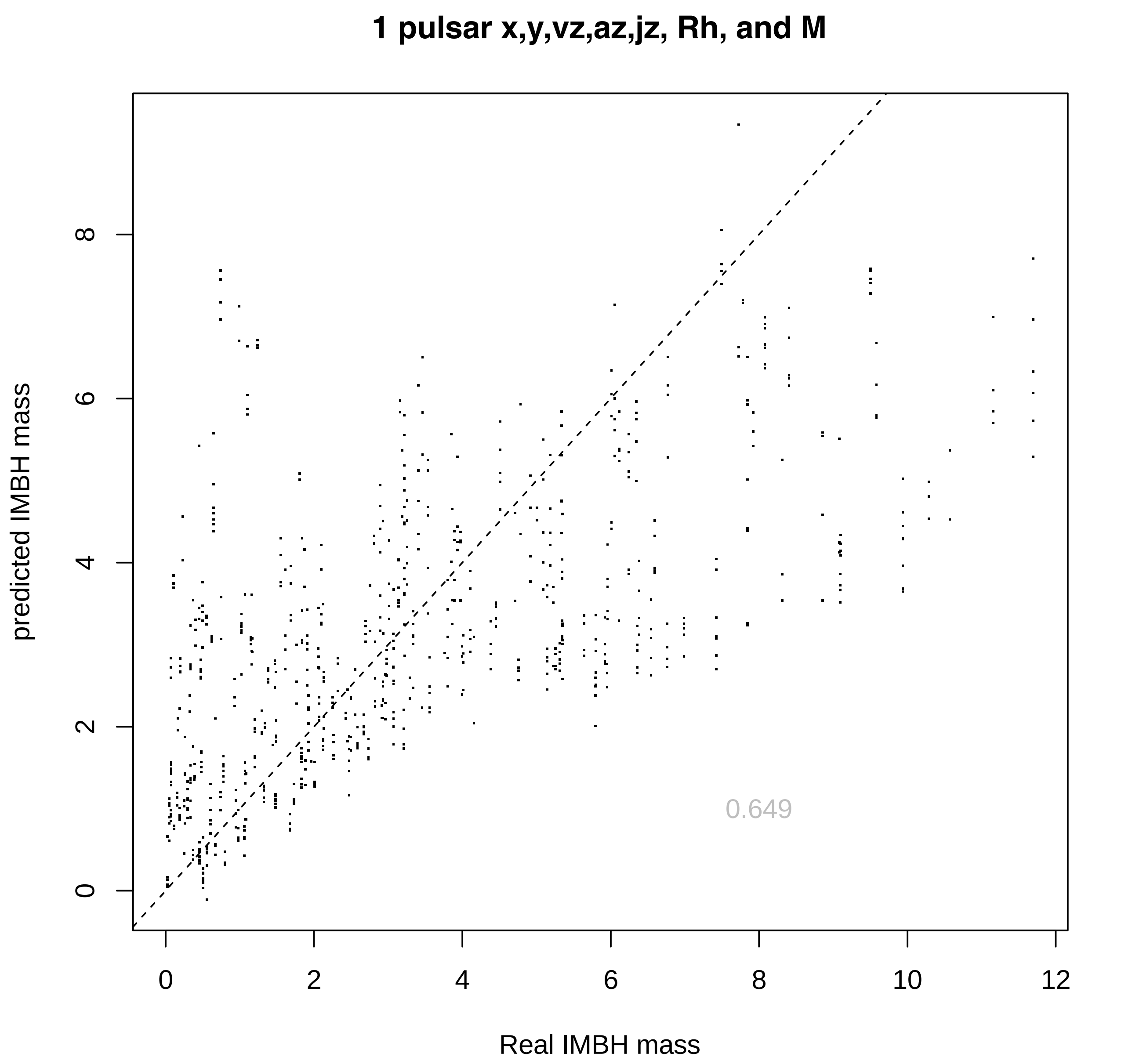
0 pulsars

This is our baseline

Masses are in 10^3 M_{sun}

Notice y-axis scale

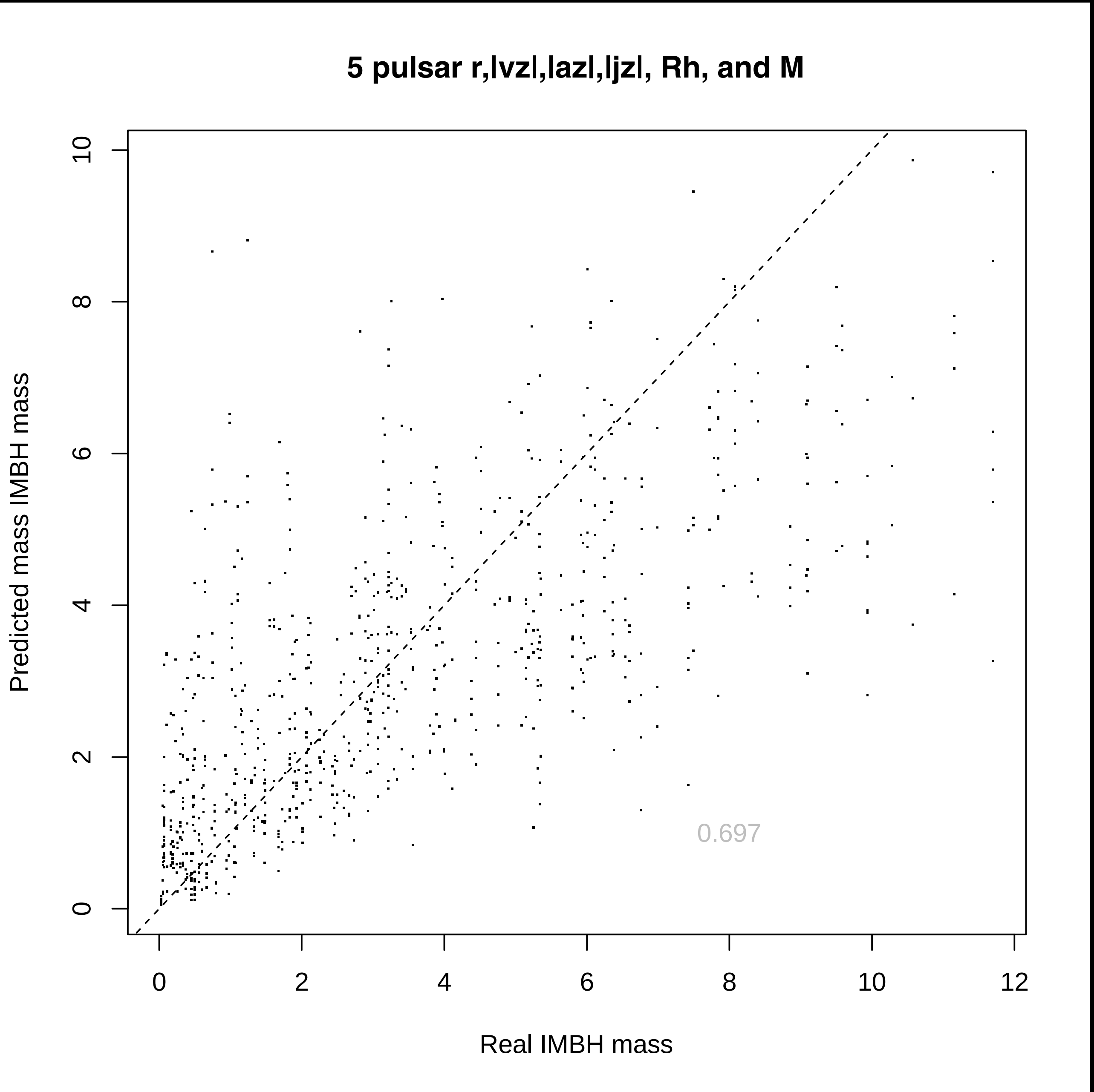
With R_h and M and 1 pulsar



1 pulsar

Slight improvement?

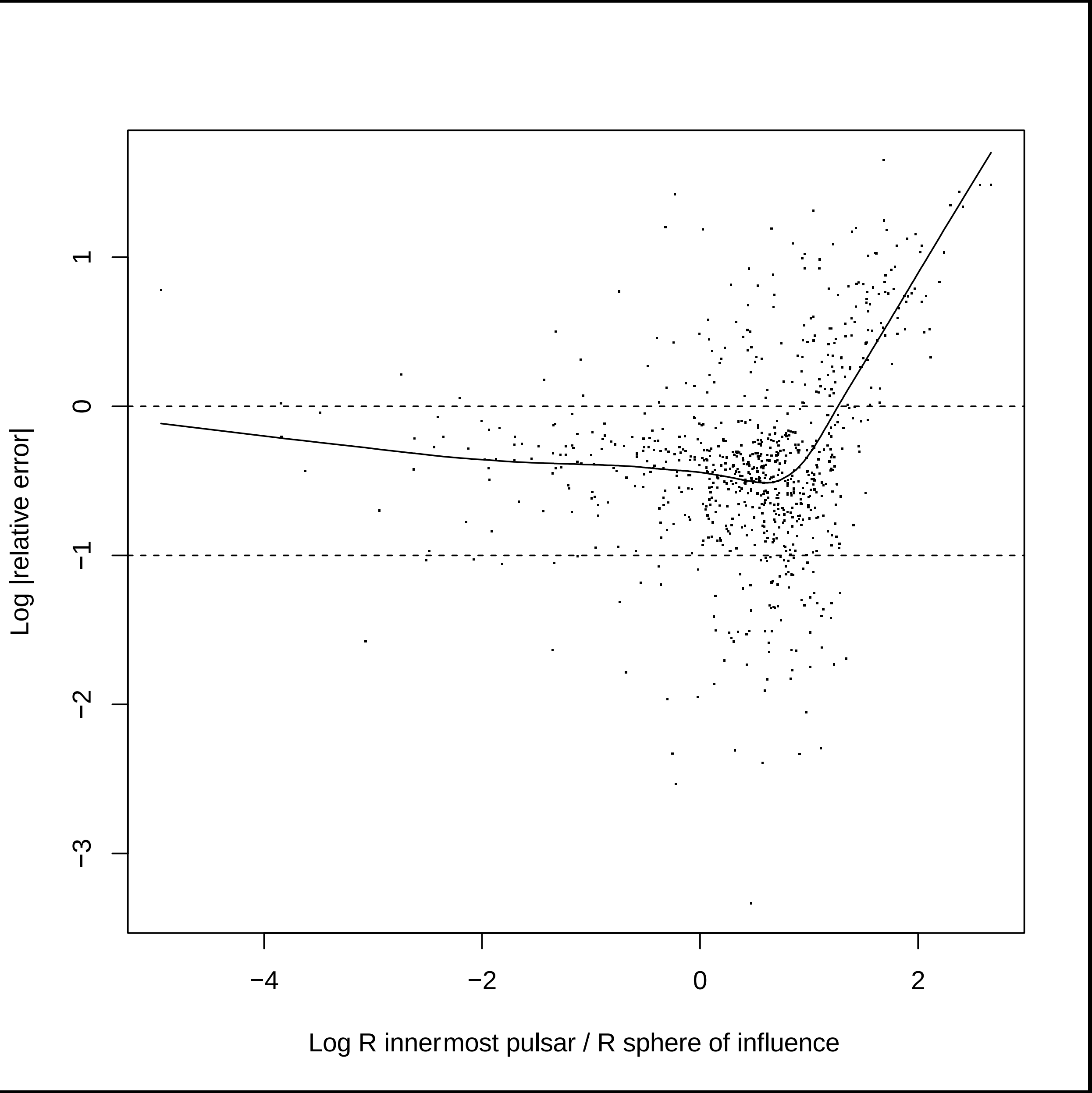
5 pulsars



5 pulsars

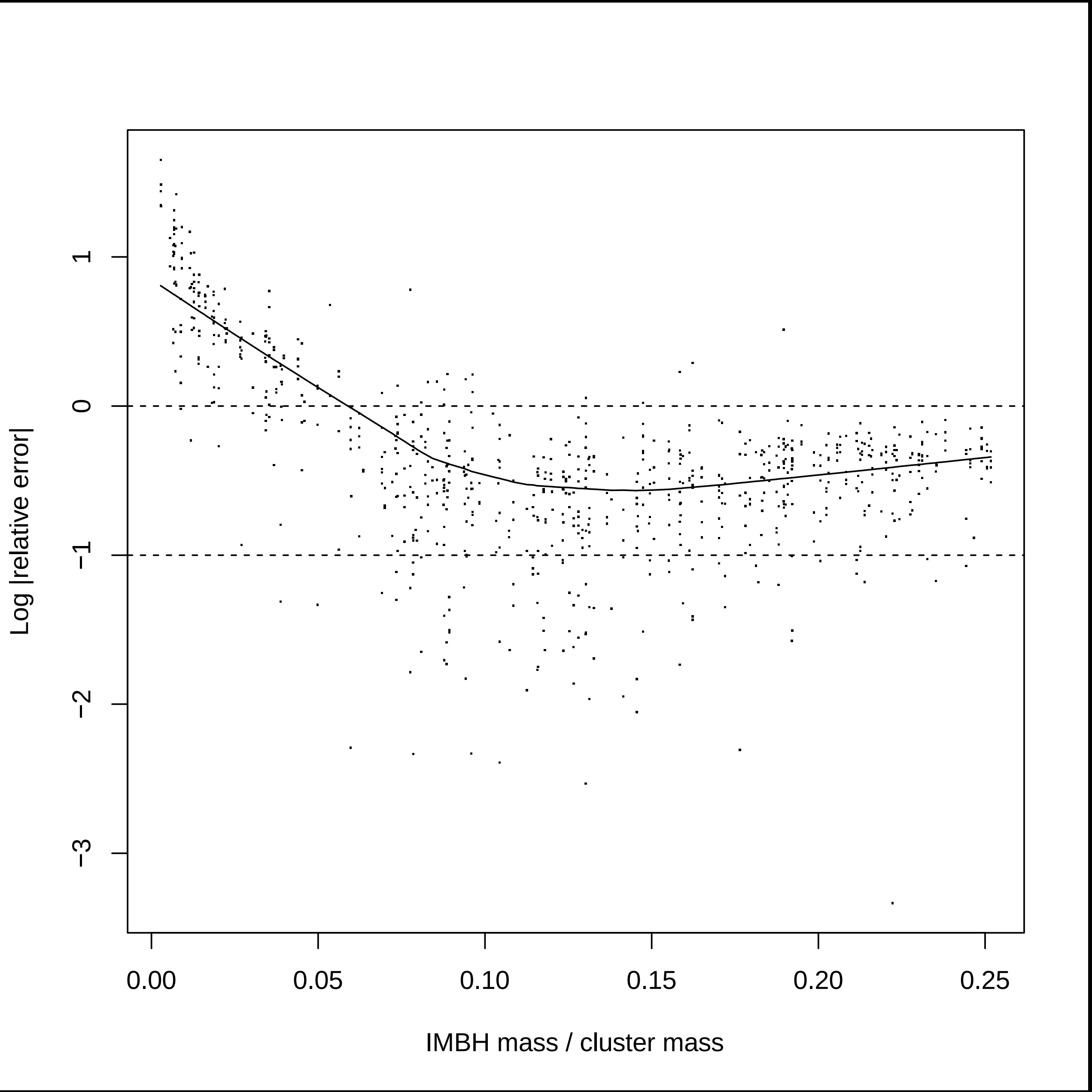
More improvement

If pulsars get near the IMBH enough...



Relative error mostly between 10% and 100% if the innermost pulsar is at least within 10 times the sphere of influence radius of the IMBH

If there actually is an IMBH...



Relative error mostly between 10% and 100% if the IMBH is at least 5% of the total cluster mass

...then we can predict IMBH mass

- IMBH mass predicted within a factor 2 (if...)
- Improvement:
  - more realistic simulations (pulsars are heavier and mass-segregate, stellar evolution, ...)
  - adding observationally realistic noise to accelerations, jerks
- Final goals: how many pulsars do observers need to constrain IMBH mass? NGC104, Terzan 5, ... IMBH mass?

Take home points

- We are on the brink of doing lots of new science with ML... join the ML revolution!
- Standard ML algorithms predict IMBH host-non-host on MOCCA simulations with good accuracy from surface density profiles only
- Interpretable tree model achieves high precision and recall on structural parameters, predicts BH subsystem host/non-host based on MOCCA simulations on N-body simulations and real clusters
- Pulsar acceleration/jerk data + structural parameters predict IMBH mass in N-body simulations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 664931.