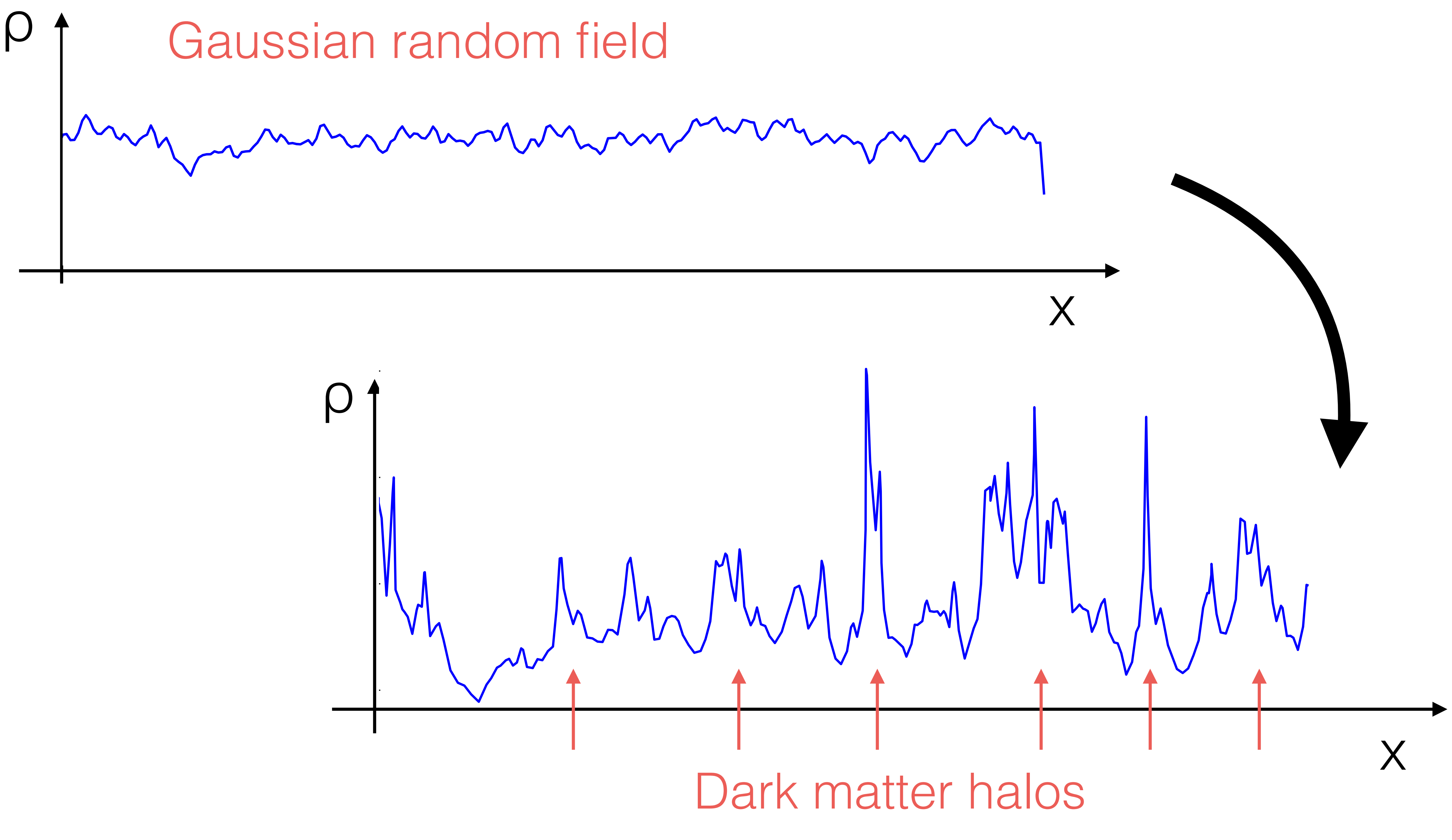
Machine Learning
Dark Matter Halo Formation

Luisa Lucie-Smith
University College London

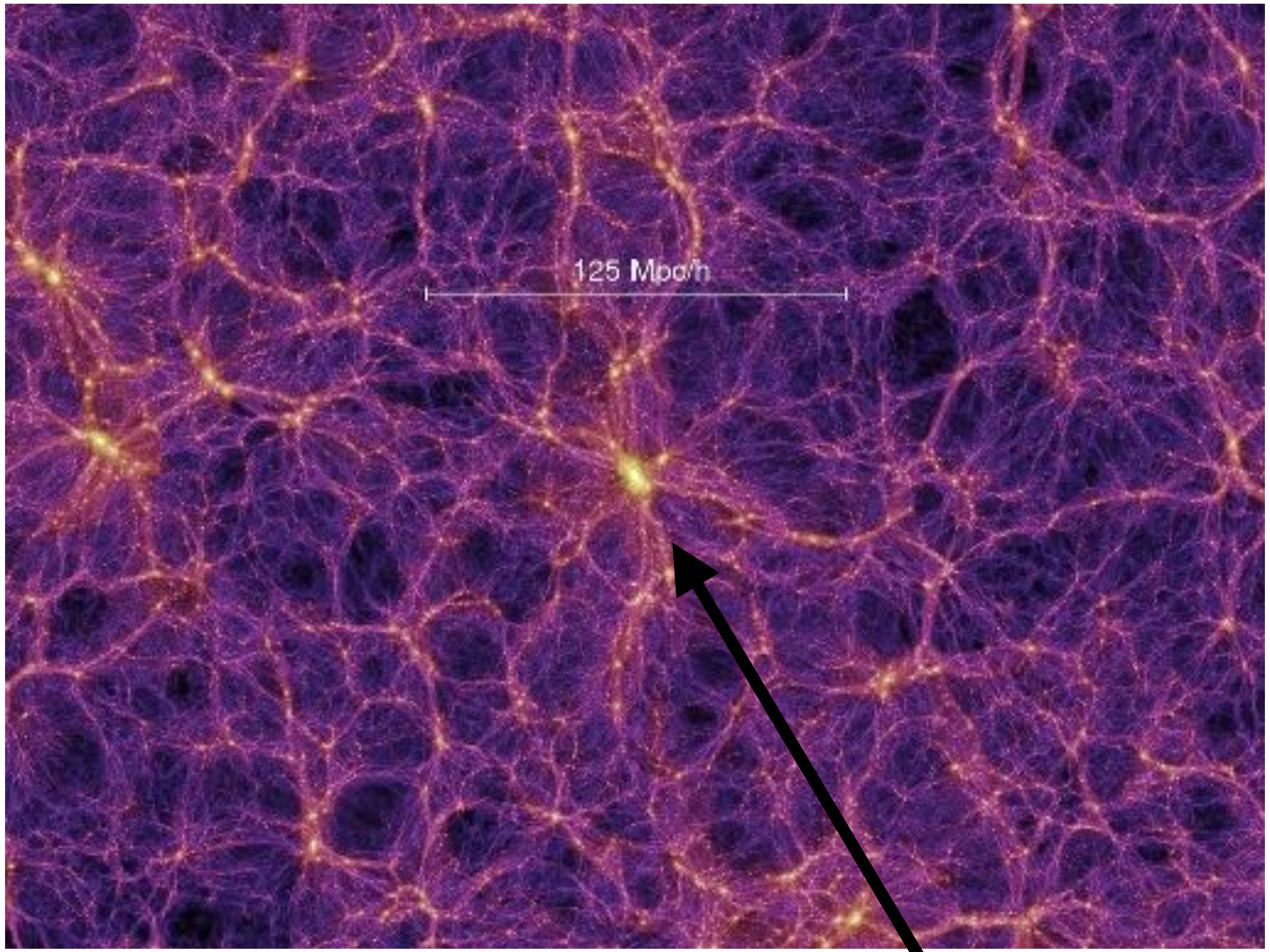
with H.V. Peiris, A. Pontzen, M. Lochner

arXiv:1802.04271

The Physics



N-body simulation



Evolve dark matter (DM) through cosmic time

Difficult ***physical*** interpretation

Dark matter halo

Figure: Springel et al. (2005)

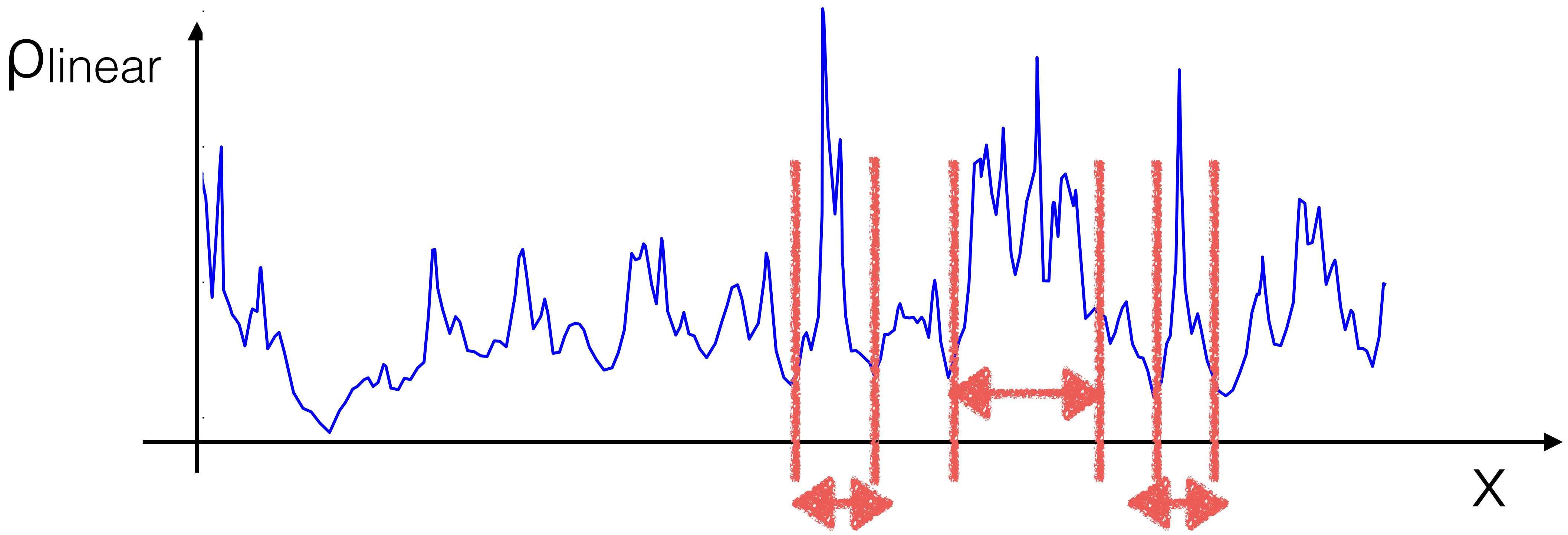
Outline

1. Train a machine learning algorithm to learn cosmological structure formation from N-body simulations

2. Investigate what aspects of early-Universe density field contain relevant information on dark matter halo formation

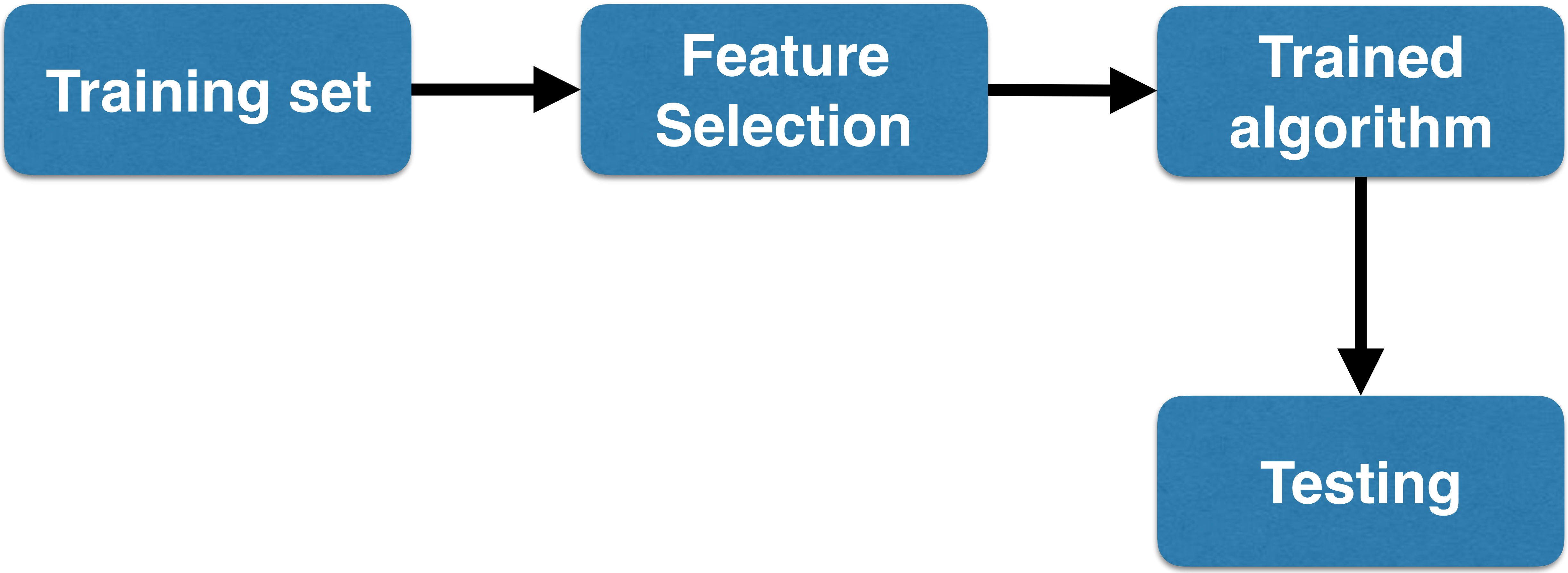
3. How we can go beyond existing analytic approximations of halo collapse

A machine learning approach



*Can a machine learning algorithm classify whether DM particles in the initial conditions will end up **IN** or **OUT** of halos of a given mass range at the end of a simulation?*

Supervised classification



Training set: N-body simulation

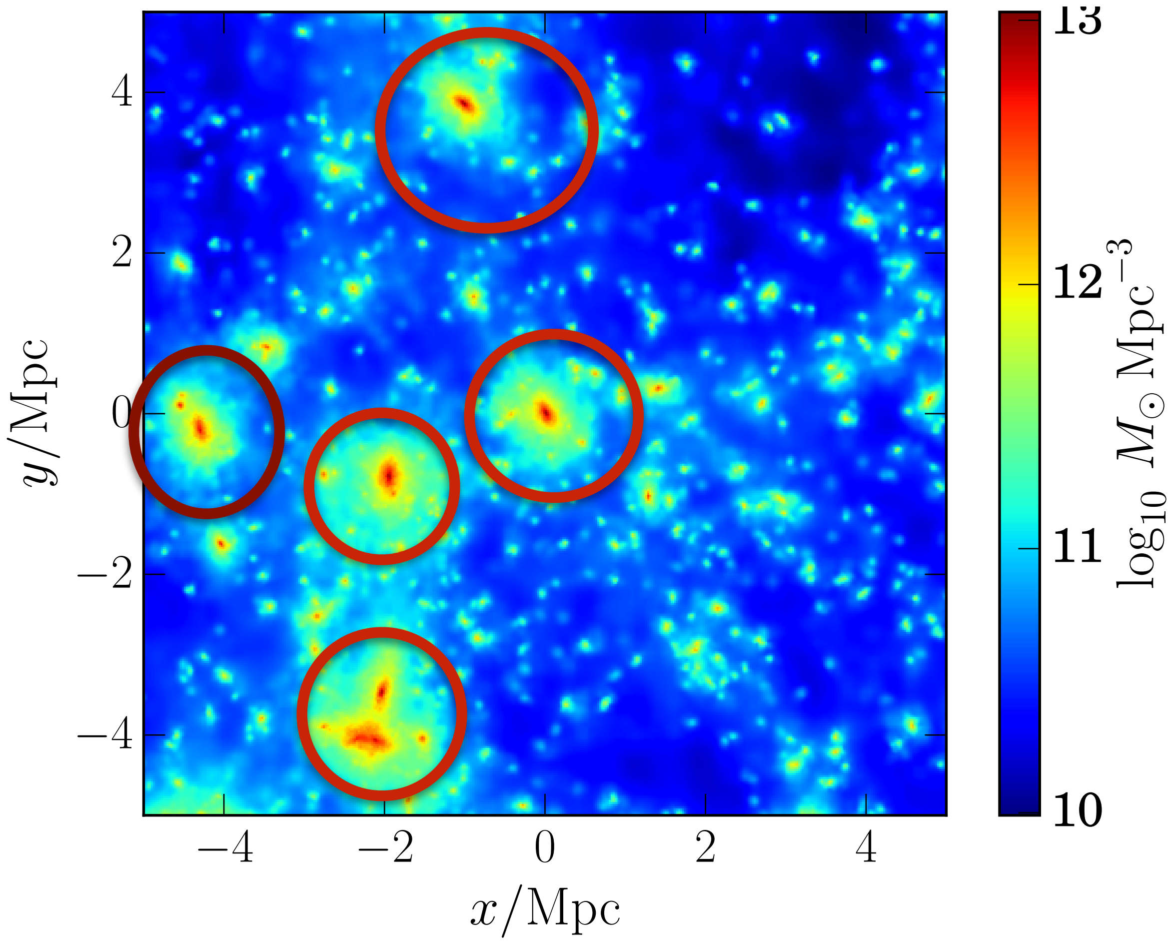
- *Samples*

  A subsample of the simulation's DM particles

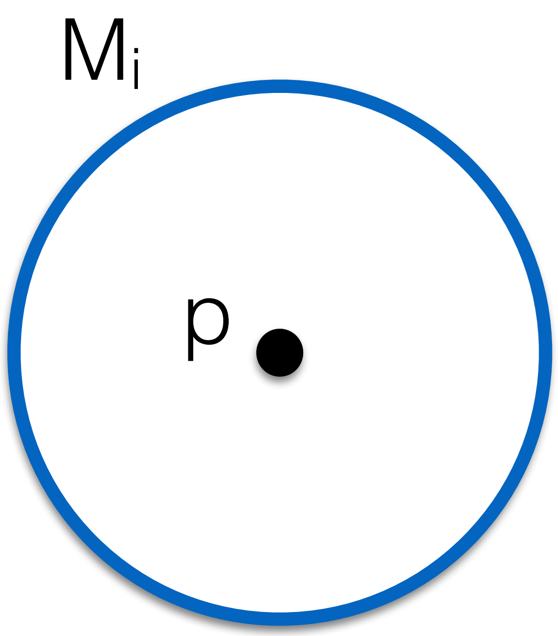
- *Class Labels*

  1. **IN** halos of mass M, s.t. $10^{12}\,M_\odot < M < 10^{14}\,M_\odot$

  2. **OUT**, otherwise.



Density features



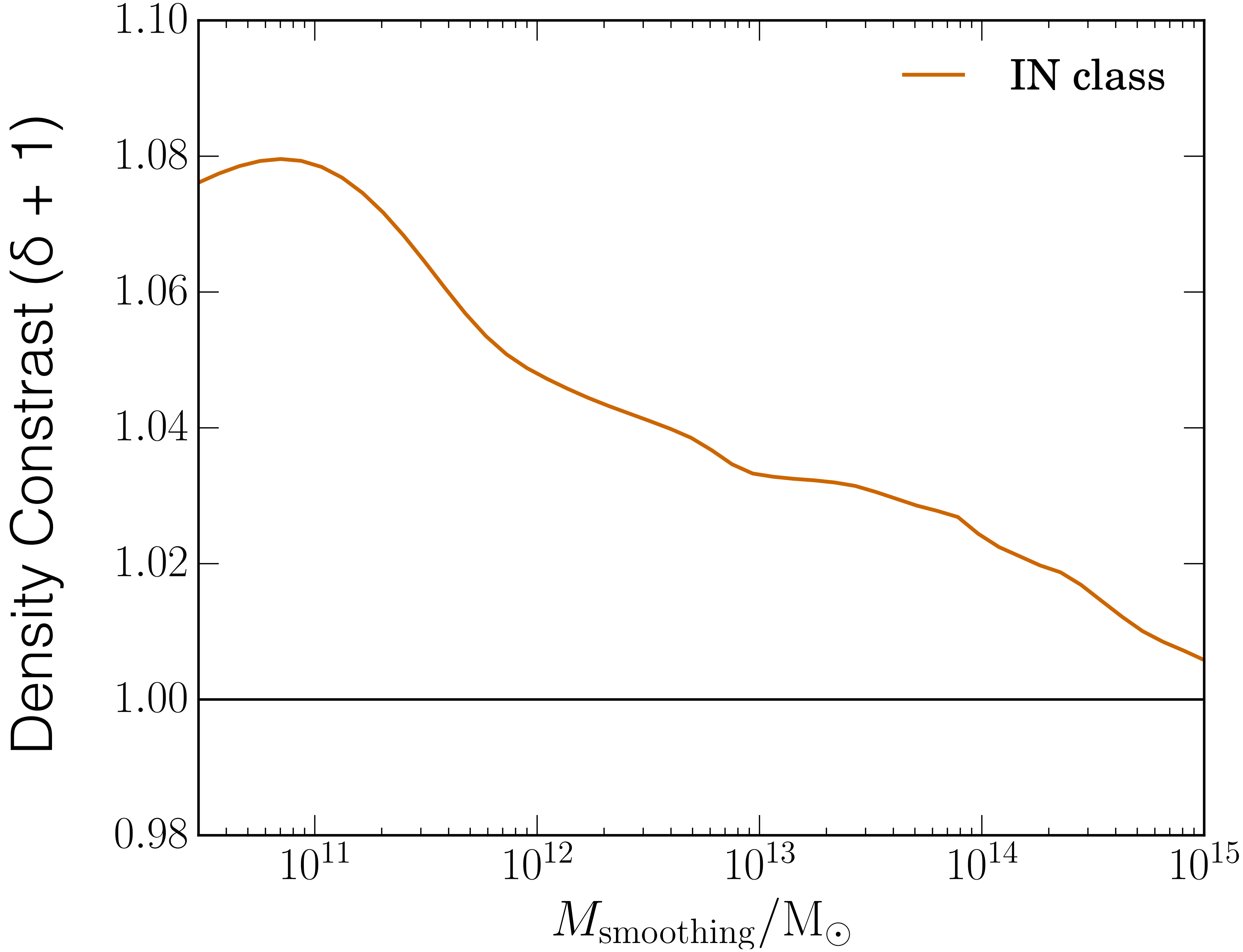
1. Smooth the density field ρ_i with a top-hat window function at mass scale M_i centred on particle p

2. Feature = density contrast,

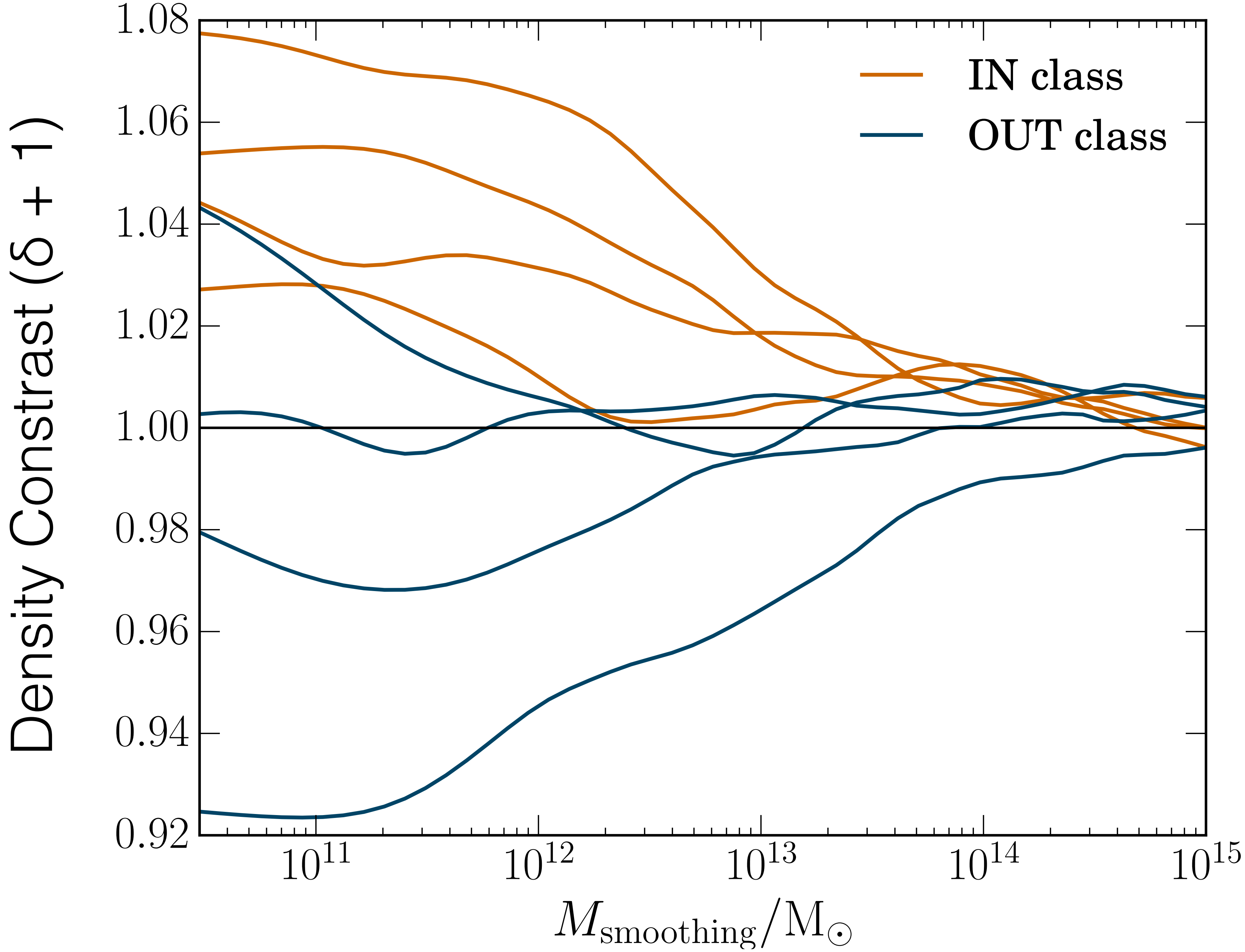
$$\delta_i = \frac{\rho_i - \bar{\rho}}{\bar{\rho}}$$

Do the same procedure for 50 mass scales

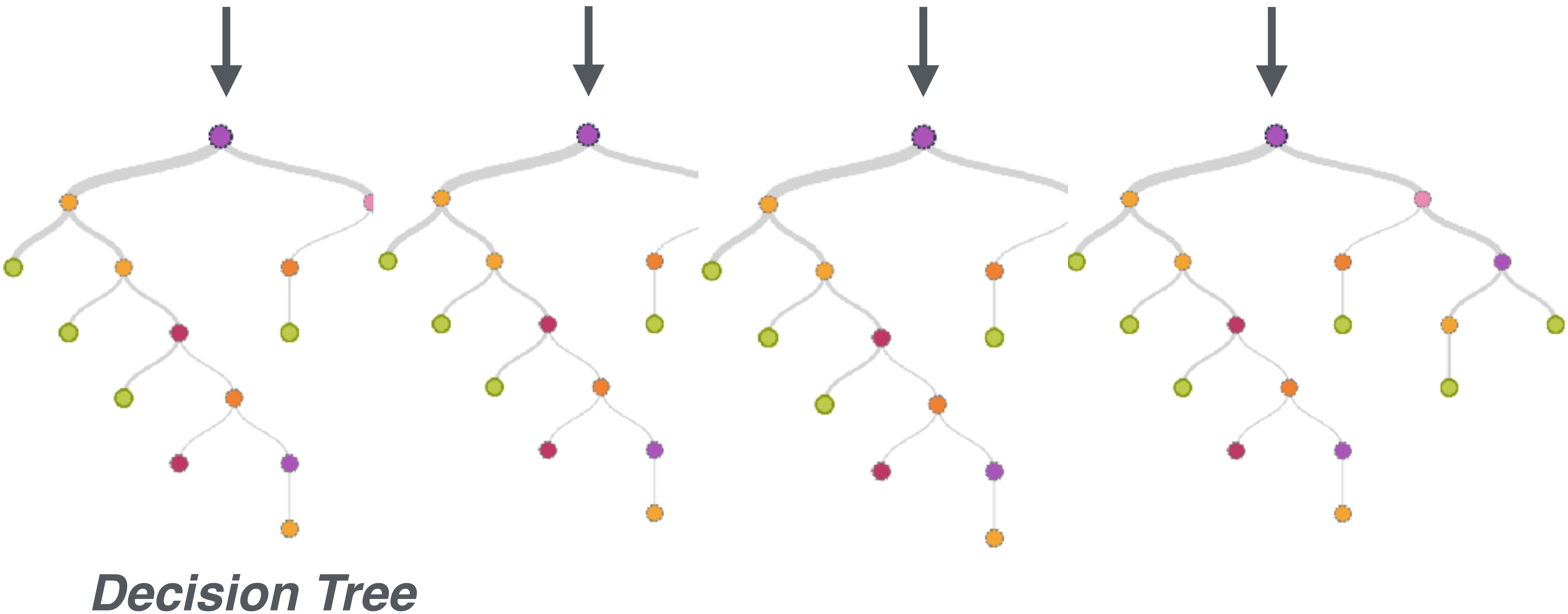
Density features



Density features

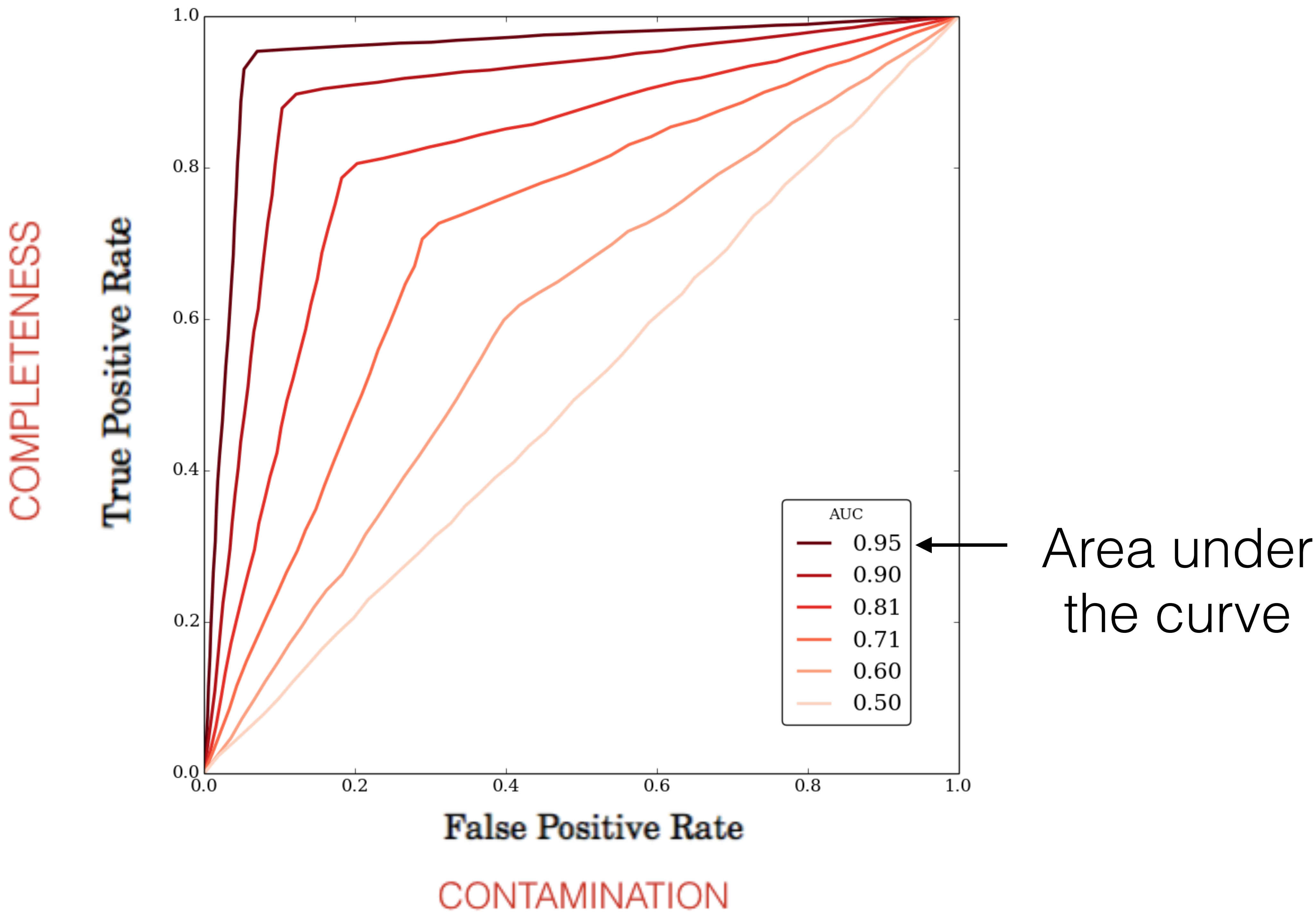


Random Forests



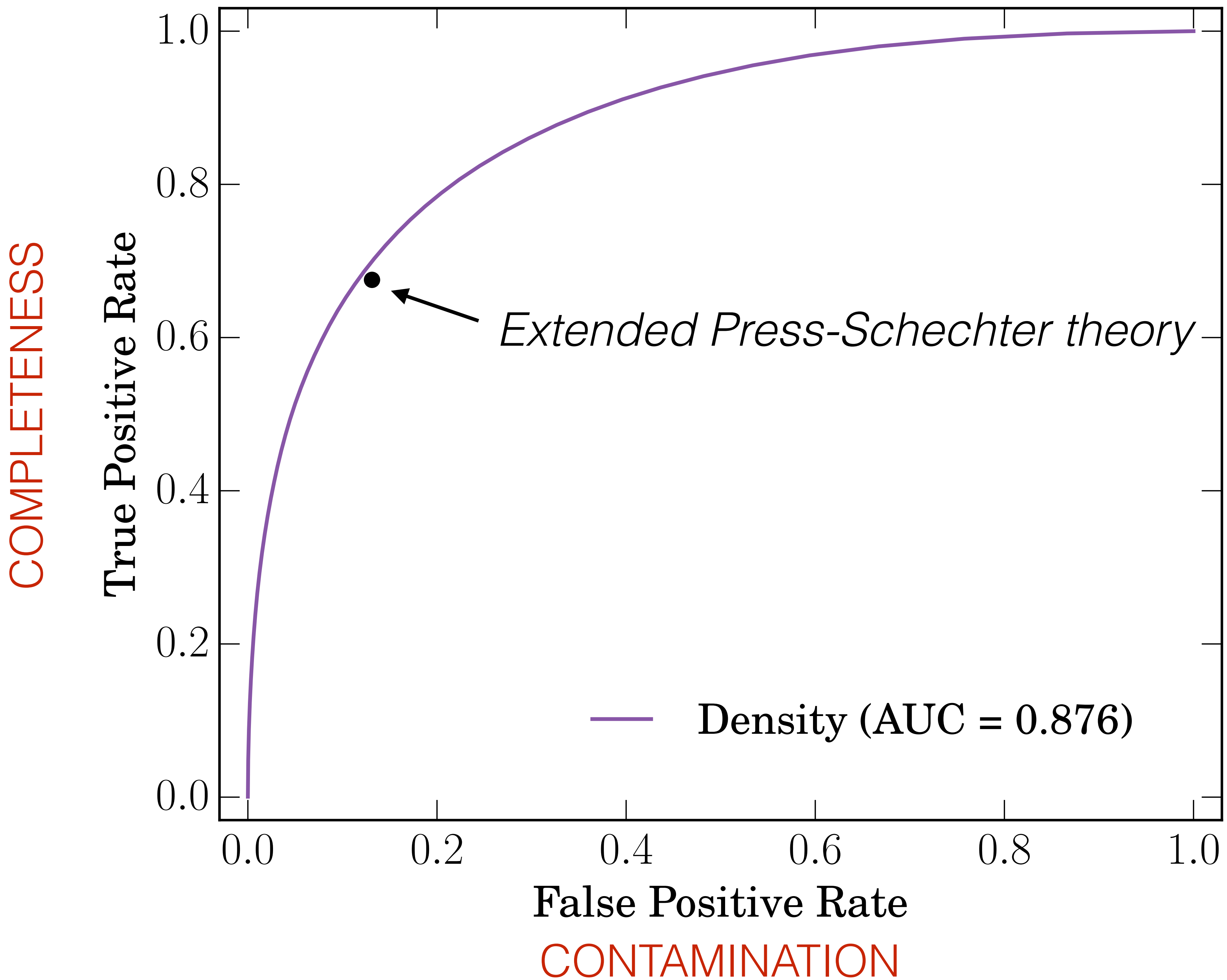
Final prediction =
average probabilistic predictions

Receiver Operating Characteristic (ROC) curves

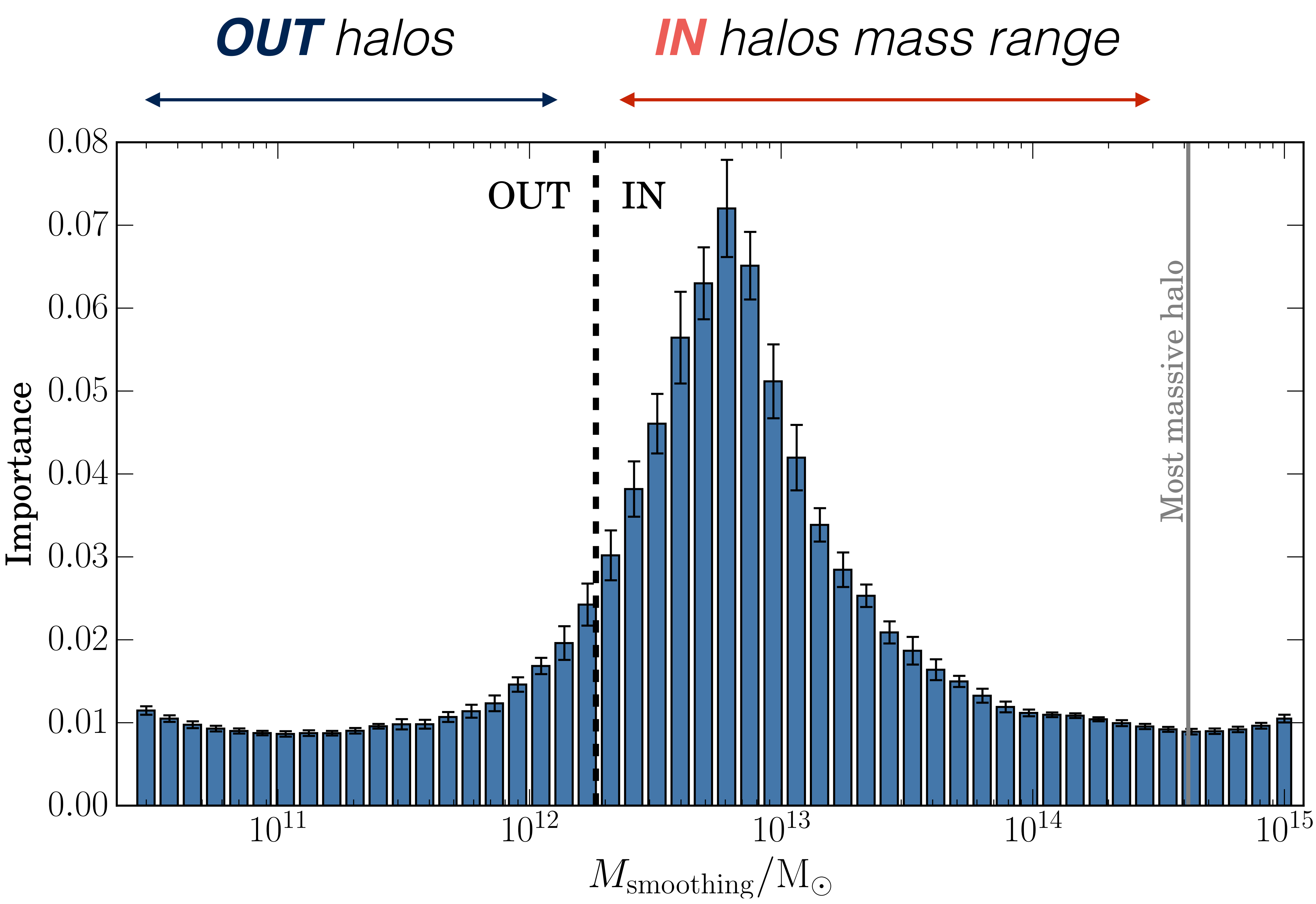


Credit: Michelle Lochner

Machine learning vs extended Press-Schechter



Density Importances



Additional physics

- **Tidal shear effects** affect the formation of dark matter halos. Motivated by *Sheth-Tormen theory* on ellipsoidal collapse

Difficult analytically ✗

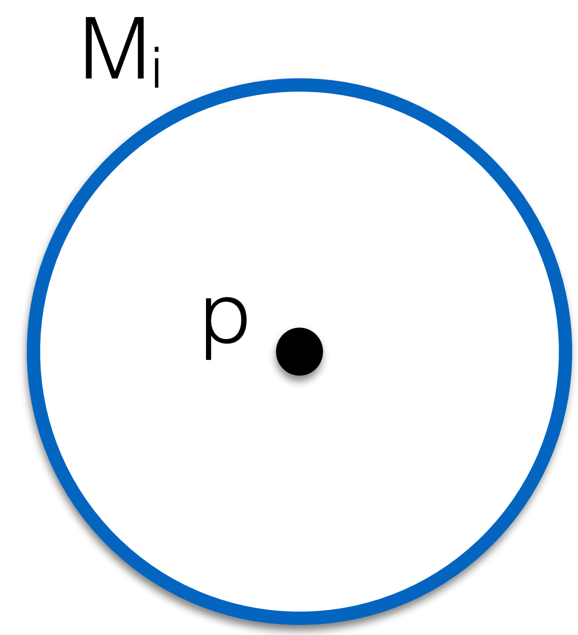
Straightforward with machine learning ✔

↓

Translate the shear field into new features!

The tidal shear

1. Smoothed density contrast δ_i at mass scale M_i centred on particle p



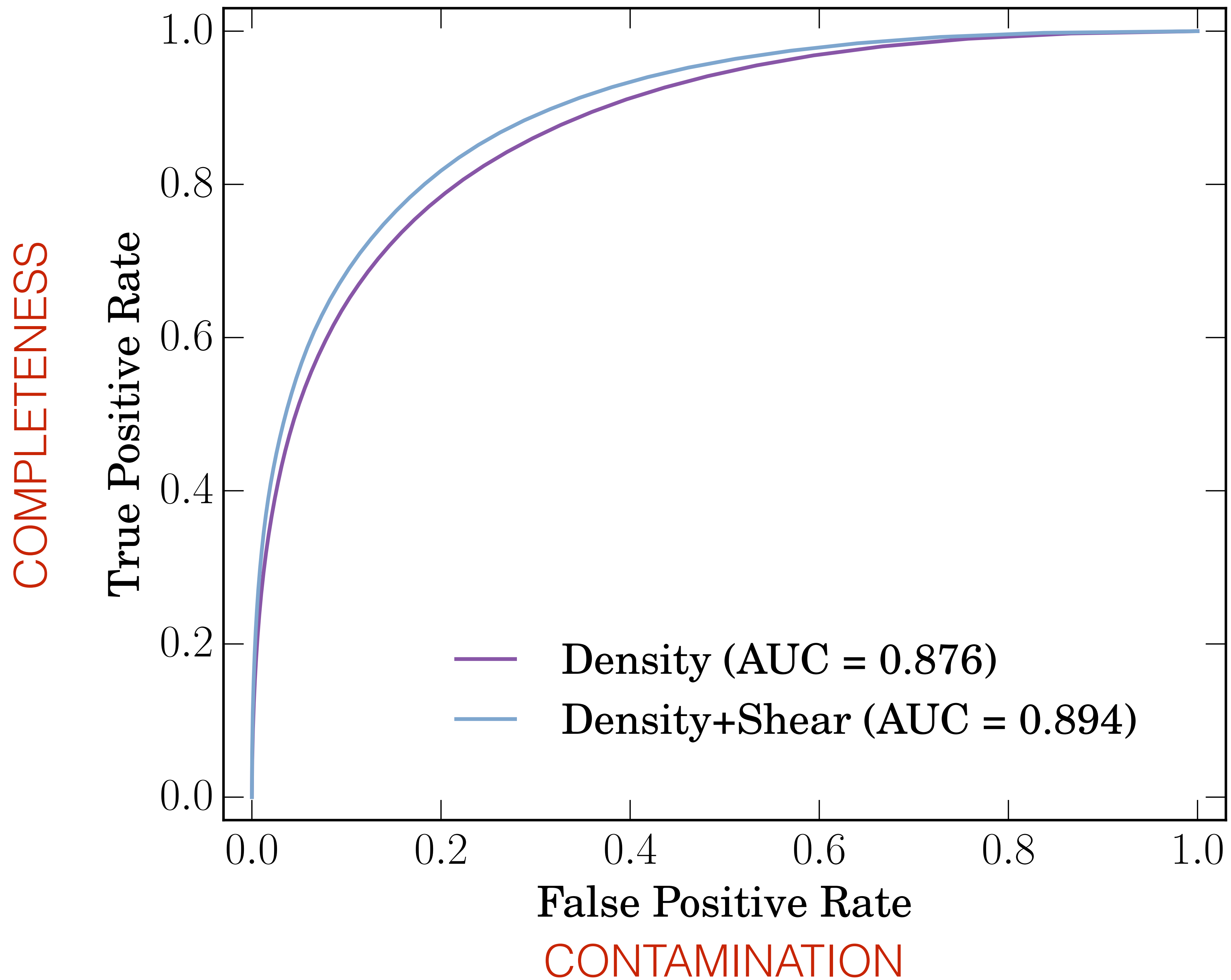
2. Solve Poisson's equation $\nabla^2 \Phi_i = \delta_i$

3. The tidal shear tensor

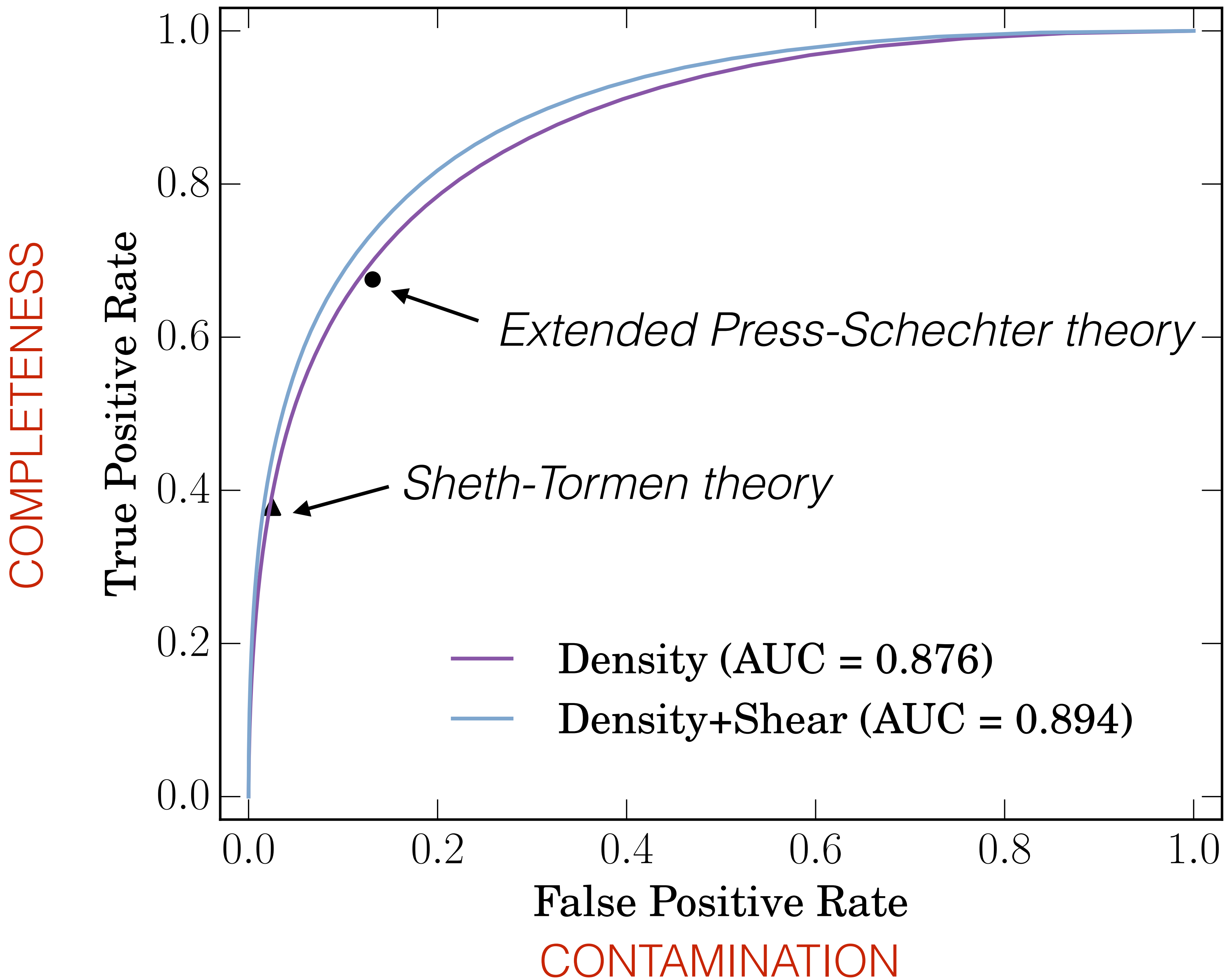
$$T_i^{\alpha\beta} = \frac{\partial^2 \Phi_i}{\partial x^\alpha \partial x^\beta}$$, with eigenvalues $\lambda_{i,1}$, $\lambda_{i,2}$, $\lambda_{i,3}$

4. Features = two independent linear combinations of the **eigenvalues** (*ellipticity* and *prolateness)*

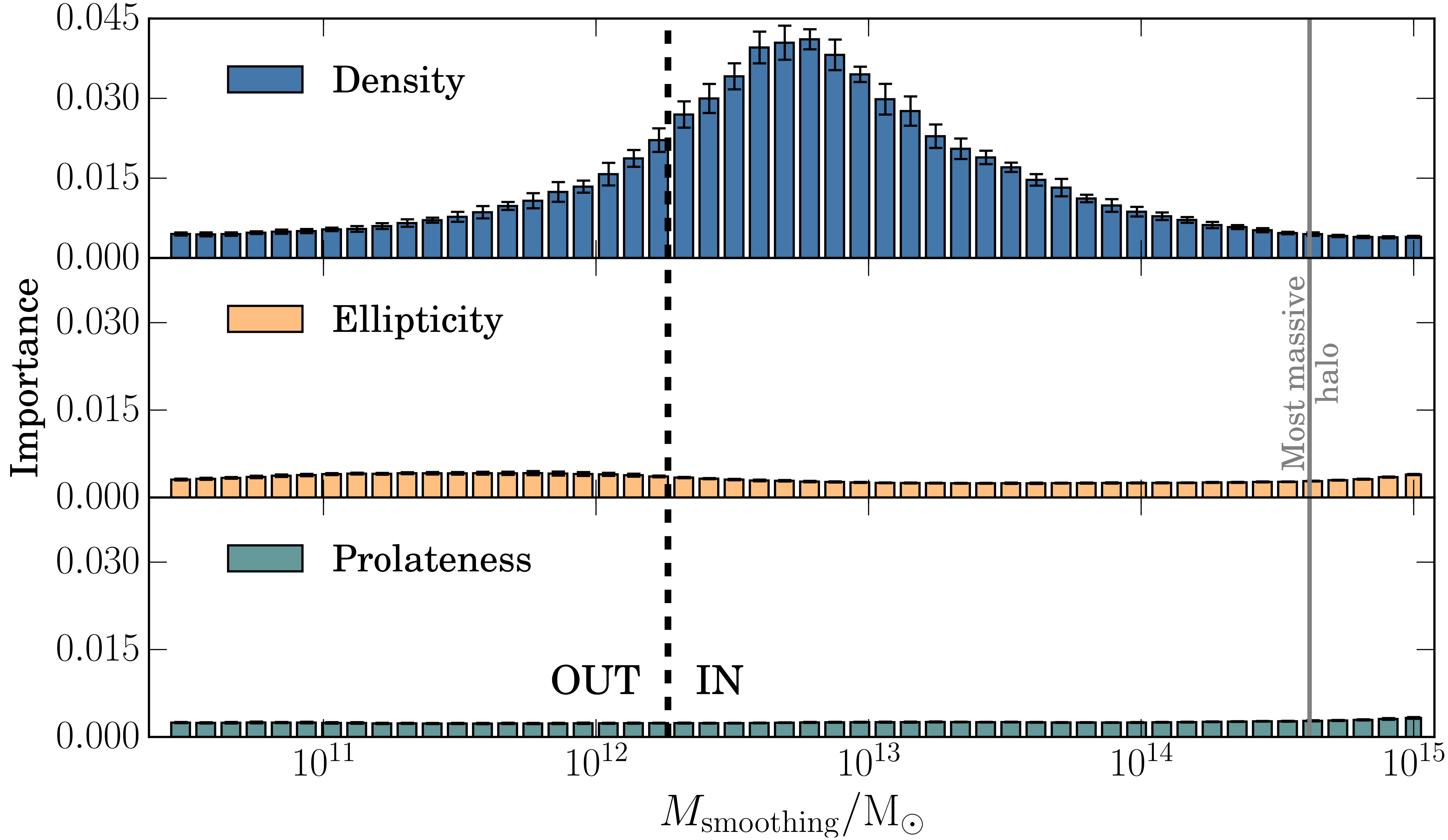
Adding the shear shows little improvement



What is the difference between ST and EPS?



Density + Shear importances



Conclusions

- Achieve comparable predictions to spherical and ellipsoidal approximations given only the linear density field
- Importance ranking shows which information improves predictions or not
- Ongoing work involves extending to regression and incorporating extra physical information which should allow better understanding of link between linear and non-linear universe

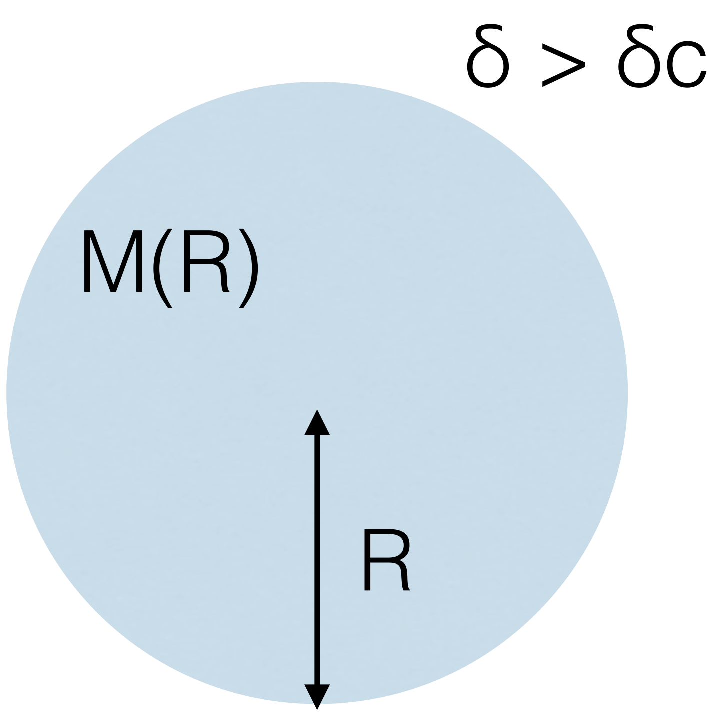
For more information see

arXiv:1802.04271

Extra Slides

The density field

Spherical collapse:



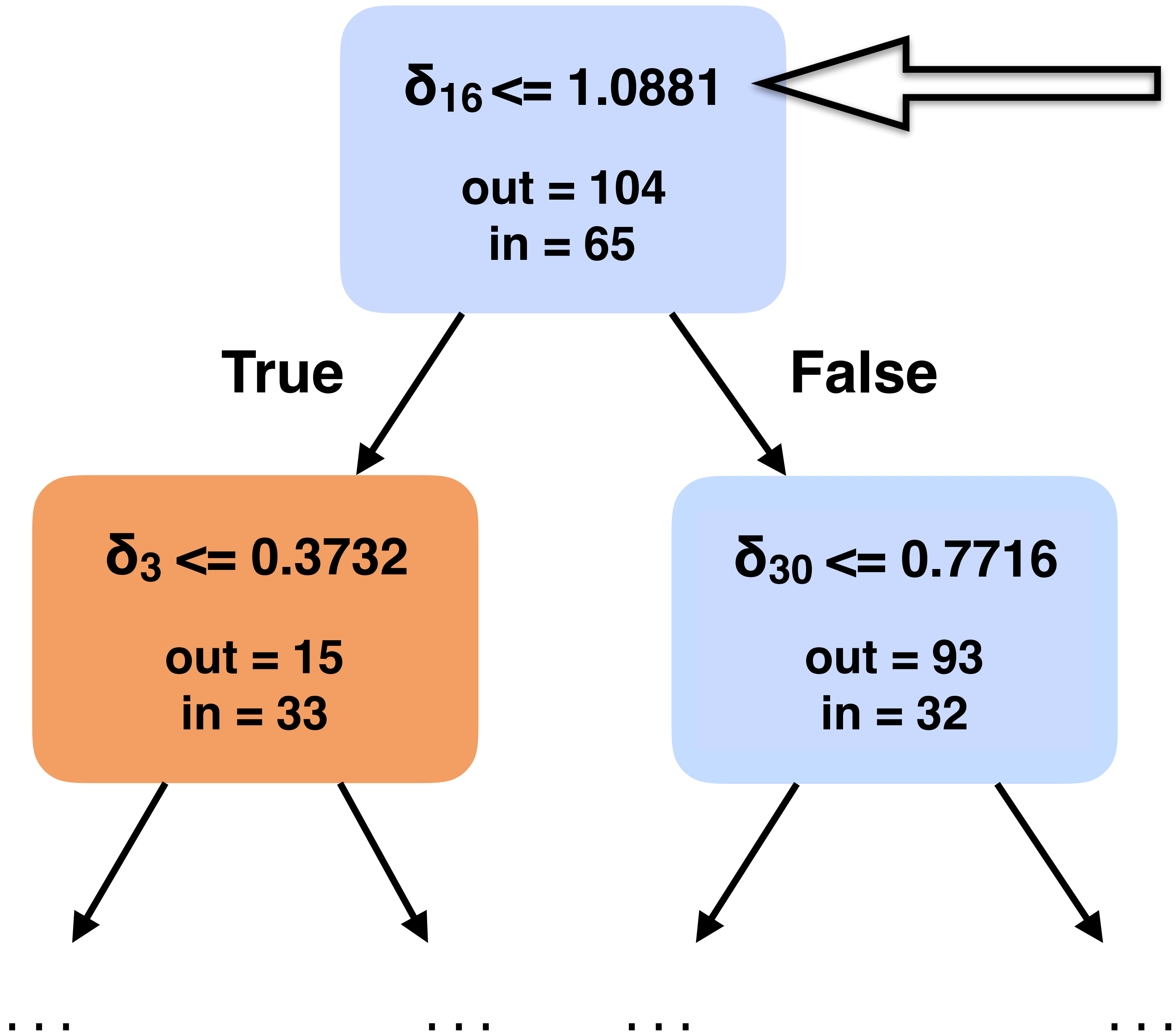
Regions where density contrast is above some threshold, δ_c

↕

Dark matter halo of mass M(R)

Extended Press-Schechter theory: *analytic* solution tested against simulations

Decision Trees



How to construct decision rules?

The decision rule at a node

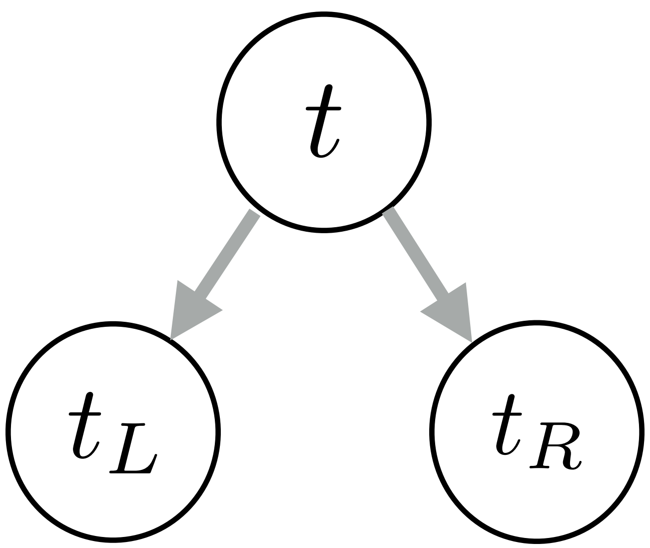
Feature's split	***Impurity Decrease Δi*** (Entropy or Gini impurity)
δ_1 <= -0.234	Δi = 0.321
δ_2 <= 0.7863	Δi = 0.87
δ_3 <= 0.0012	Δi = 0.56
…	…
…	…

Highest impurity decrease

↓

Choose feature 2!

The best split feature



Maximise impurity decrease

$$\Delta i = i(t) - p_L i(t_L) - p_R i(t_R)$$

Entropy

$$i_E(t) = -\sum_{j=1}^{c} p(j,t)\log_2 p(j,t)$$

Gini Impurity

$$i_G(t) = 1 - \sum_{j=1}^{c} p(j,t)^2$$

The tidal shear features

Define $t_{i,j} = \lambda_{i,j} - \delta_i/3$, where λ are the tidal shear eigenvalues.

Two new features per particle at mass scale M_i :

- ***Ellipticity***

$$e_i = 3(t_{i,1} - t_{i,3})$$

- ***Prolateness***

$$p_i = 3(t_{i,1} + t_{i,3})$$

Do the same procedure for 50 mass scales

Feature Importance

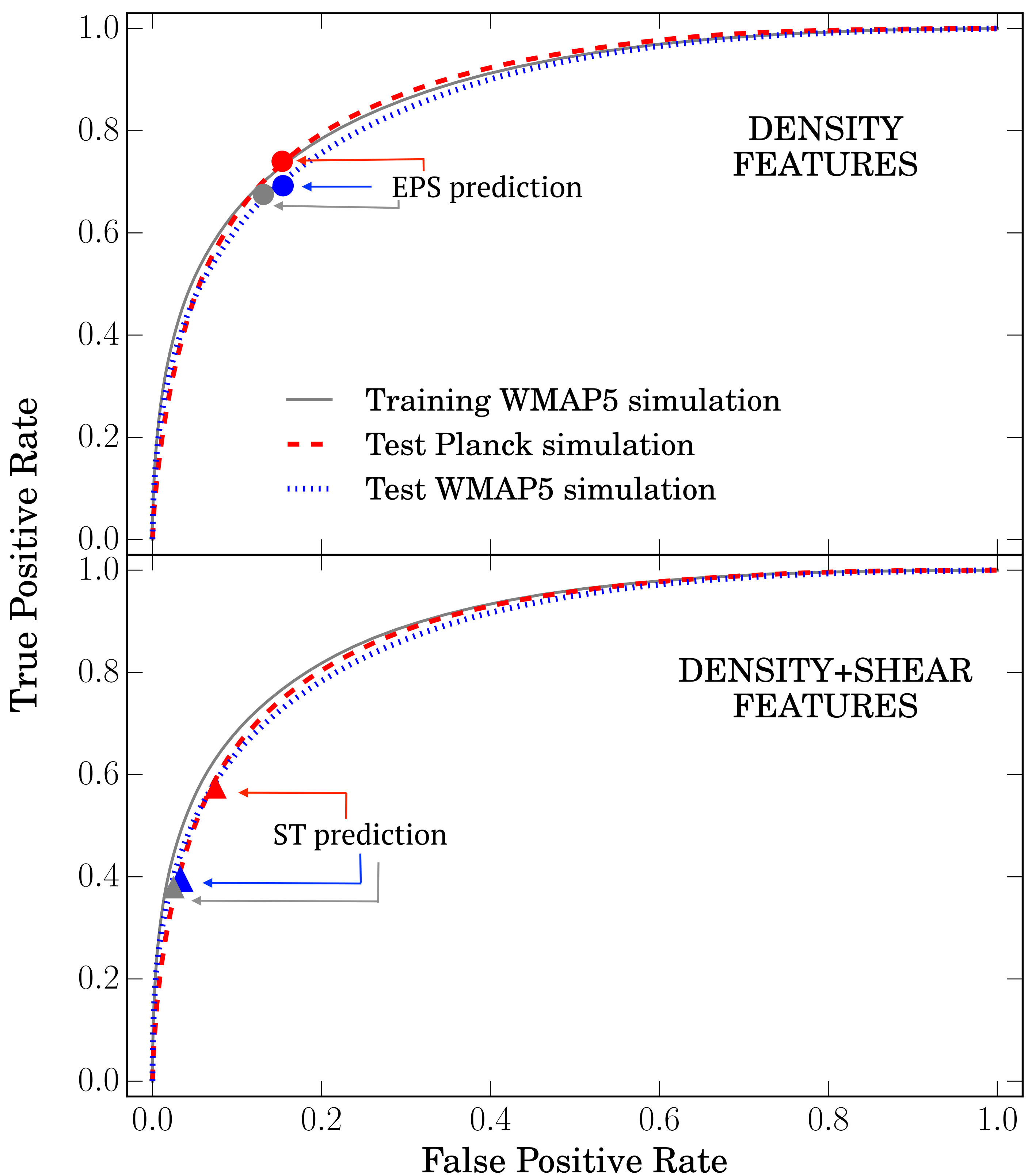
fraction of samples

$$\mathrm{Imp}(X) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : s_t = X} p(t) \Delta i(t)$$

Number of trees

Impurity decrease
(Entropy or Gini impurity)

Test on independent simulations



Supervised classification