Leveraging Supervised ML for Targeted Anomaly Search:

An Application to Stellar Infrared-Excess detection

Gabriella Contardo

Data Science Group @ SISSA, Trieste (Italy)

Work done in collaboration with D. W. Hogg (NYU/CCA)

SISSA DATASCIENCE
Machine Learning for the Natural Sciences

SISSA

On finding odd things

- Anomaly Detection: Two (at least?) flavors:
  - “Distribution divergence” compared to some model
  - **Outlier detections**
- Finding *“odd objects” :* Points with a low probability, or in a low density region of the data space
- From a Machine Learning perspective: **“unsupervised” learning problem**, using **density estimation** (or other forms of DE / using latent space)
- Caveats:
  - Obtaining reliable density estimates is non-trivial, especially in high-dimension.
  - Will retrieve all (potentially) rare objects, but not necessarily the “interesting” ones.
- Advantages: Relevant for **unknown unknowns.**

On finding odd things

- In some instances, we actually look for **“known”** (or targeted) **unknowns**:
  - Finding more examples of that “weird object” that we stumbled on
  - **Objects that deviate from “expectation” in a specific way / region** (conditional or contextual anomalies in some nomenclatures).
- “Knowing” our unknowns does not necessarily mean that we can turn this problem into a supervised (binary) classification one (or a simple selection cut):
  - We might not have good examples of such anomalies, or very few, or hard / expensive to model them
- However, we might leverage this to help our search by framing it back into a supervised problem, without supervised anomalies.

Infrared-Excess in Stars

- Infrared Excess (i.e. departure from “expected” infrared emission) in stars can be caused by protoplanetary disks, circumstellar dust, debris disks, ...
- **Extreme excess** have been observed in some stars (and some “not so young”): candidates for “**Extreme Debris Disks**” (EDD, potentially coming from planetary collisions?)
- Quite rare occurrence: previous search had ~0.01% occurrence rate. <20 candidates currently.

Finding Infrared-Excess in Stars

- Most searches for IR-excess rely on :
  - IR-observation : quality / SNR cut in those dataset, often removes *a lot* of data to search in. But: EDD excess can actually show in mid-IR.
  - Some modeling to estimate an excess. Proper stellar model fitting is prohibitively expensive, so template approximations.
- Our pipeline:
  - Focus on “non-young” (main sequence) FGK (Sun-like) stars.
  - Use mid-IR for determining the excess / anomalousness.
  - Define MIR-excess in a data-driven way.

Data Driven Search for MIR-excess

- Model Mid-IR emission *from the data*:
  - Fit a regression model (here Random Forest) predicting the MIR from optical and near-IR photometric (and other) features
- Look for stars that have **confidently incorrect** predictions:

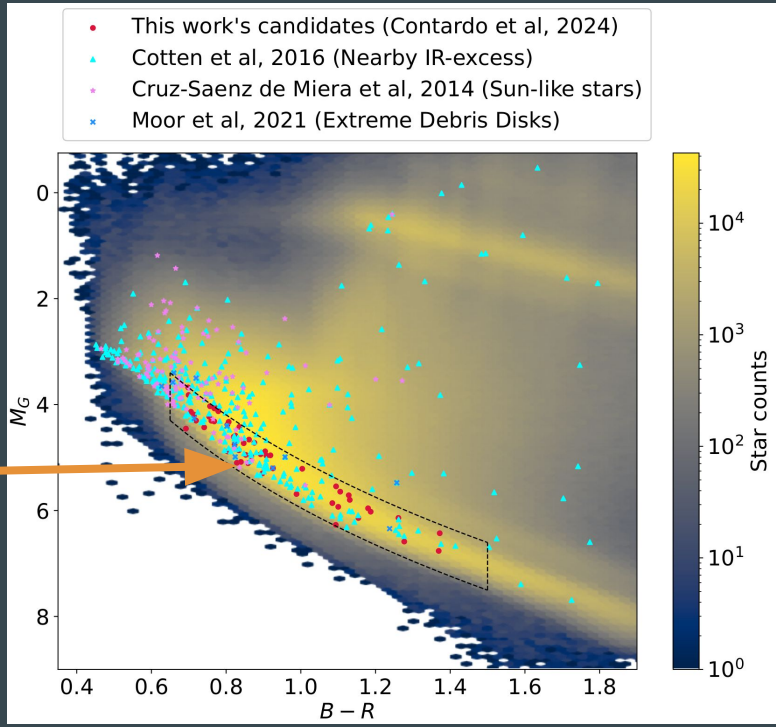
→ Anomalies *according to the data*, in a specific area of the entire feature space.

- Advantages: computationally efficient, bypass the need for stellar model(ing), sensitive to extreme outliers in the IR “leaking” in the MIR.
- Limitations: can only find outliers according to the data and the input features

Data

- Combine data from Gaia DR3, the Two Micron All Sky Survey (2MASS), unWISE and allWISE catalogs.
- G < 16 mag and 4000 K < Teff < 70000 K cut (FGK stars)
- Parallax error, ruwe, reddening, cross-match (2MASS, AllWISE) cuts.
- Main-Sequence cut
- Dust cut
- unWISE crossmatch and quality cut for MIR

4.9M stars



Method

- Input: photometric observations (magnitudes) and colors from Gaia DR3 and 2MASS, absolute magnitude M_G , *ruwe, parallax,* and reddening value from DR3
- Predict colors K-W1 and K-W2 with Random Forest regressors with default setup
    - Other methods did not show significant improvement in prediction quality on held-out sample.
- 8-fold split: different RFs are trained on each fold (~600,000 stars)

→ 7 “test predictions” for each stars

- Predicted magnitude per RF:

$$\widetilde{W}_{i,j} = -(RF_j(x_i) - K_i)$$

- Predicted magnitude combining RFS:

$$\widehat{W}_i = \mathrm{Median}(\{\widetilde{W}_{i,j}\}_{j \in \mathrm{F}_i})$$

Anomaly Criterion Cuts

- High prediction errors: anomalous (either excess or deficit), but we want to focus on **highly confident** incorrect predictions.

Additional criteria:

1. Have high prediction precision (low variance in prediction across folds)

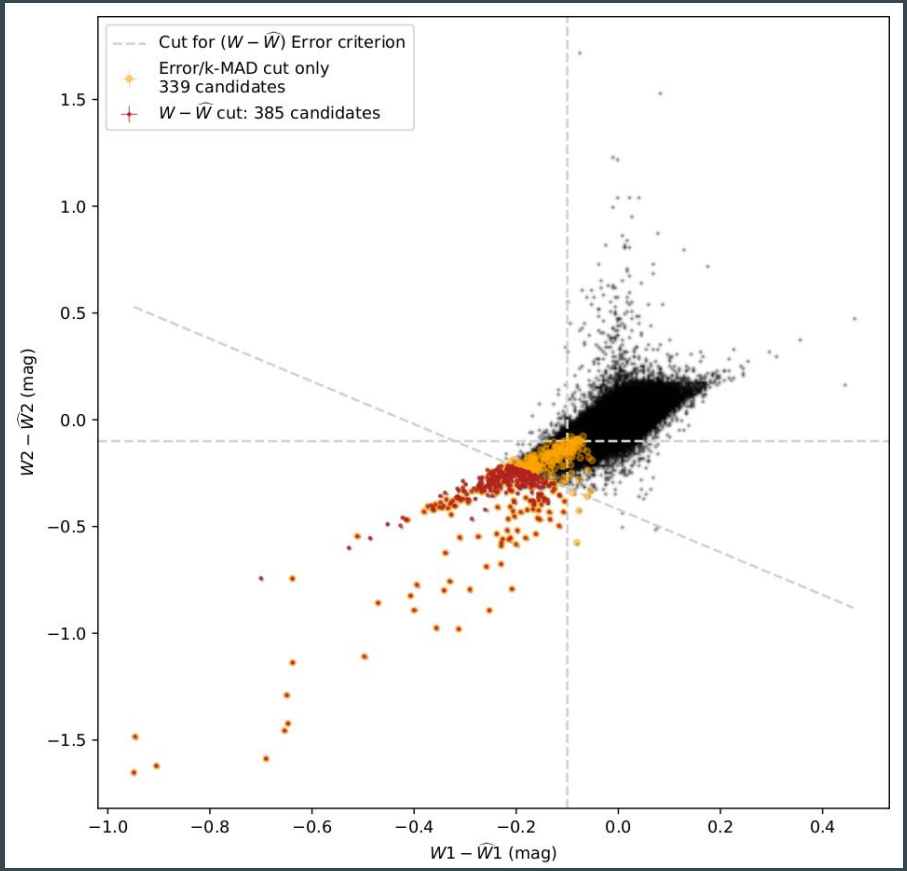
$$\text{fold-MAD}_i = \text{Median}(\{|\widehat{W}_i - \widetilde{W}_{i,j}|\}_{j \in \text{F}_i})$$

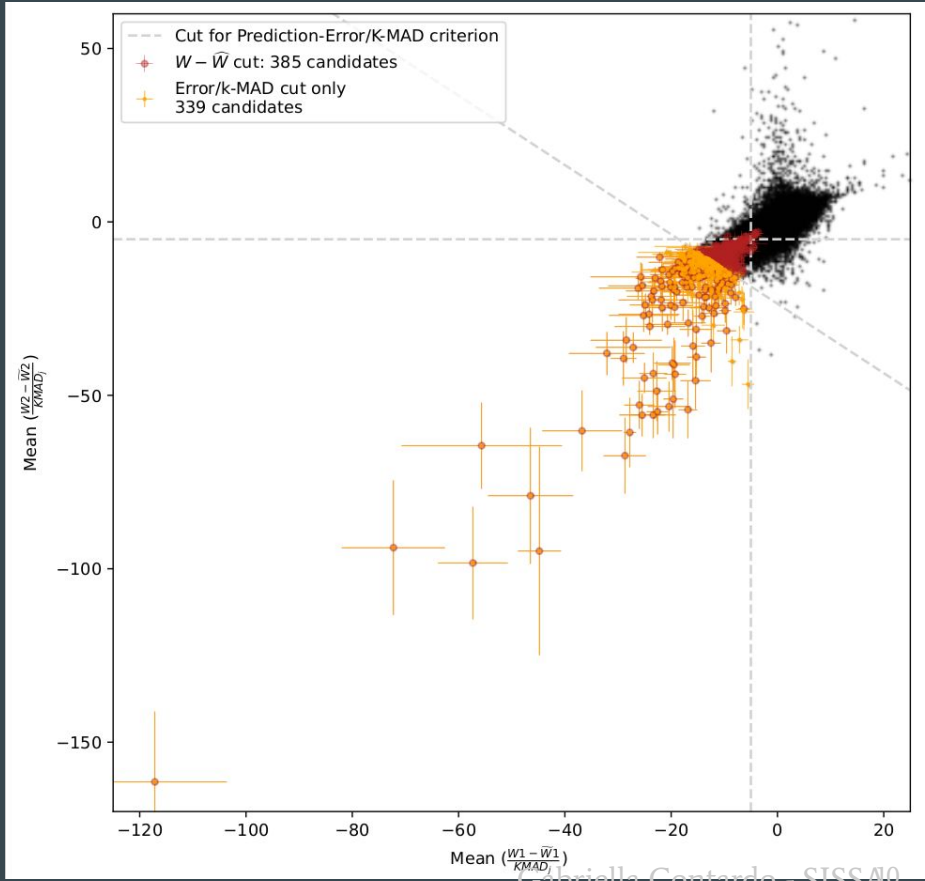
2. Are in *well-predicted* regions of the feature space: similar examples (kNN) have high accuracy

$$\text{K-MAD}_{i,j} = \text{Median}(\{|W_k - \widetilde{W}_{k,j}|\}_{k \in \text{NN-colour}(i,j)})$$

3. Are in a well-populated region of the dataset (i.e. they are not outliers in the feature space): mean Distance kNN < .1

Anomaly Criterion Cuts



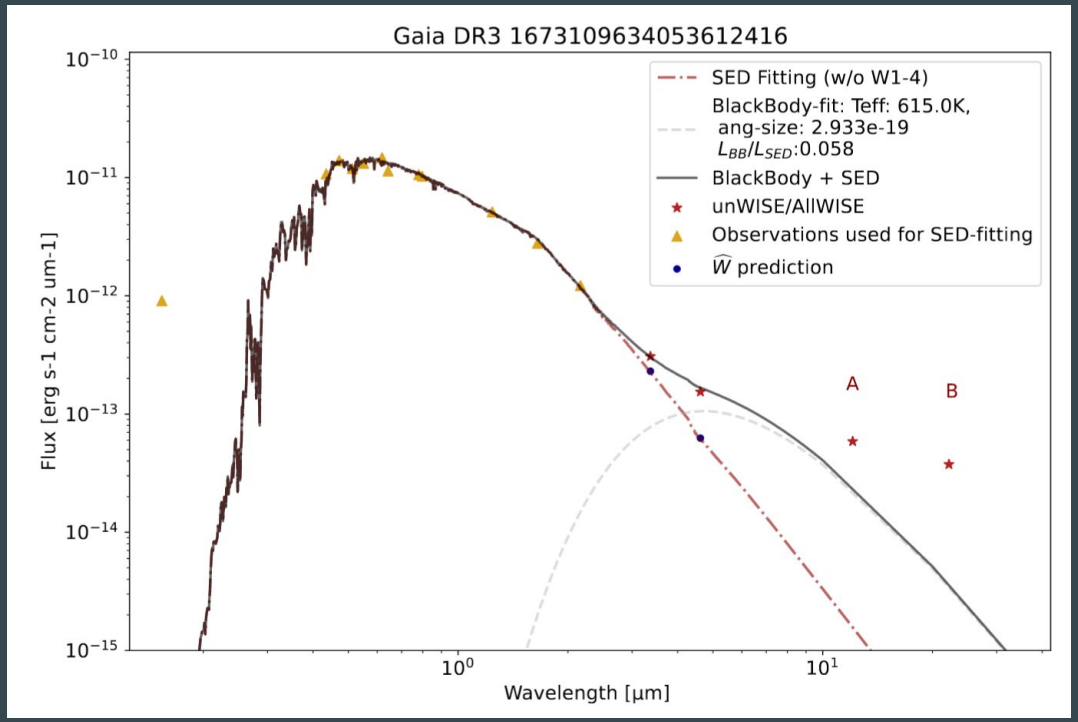


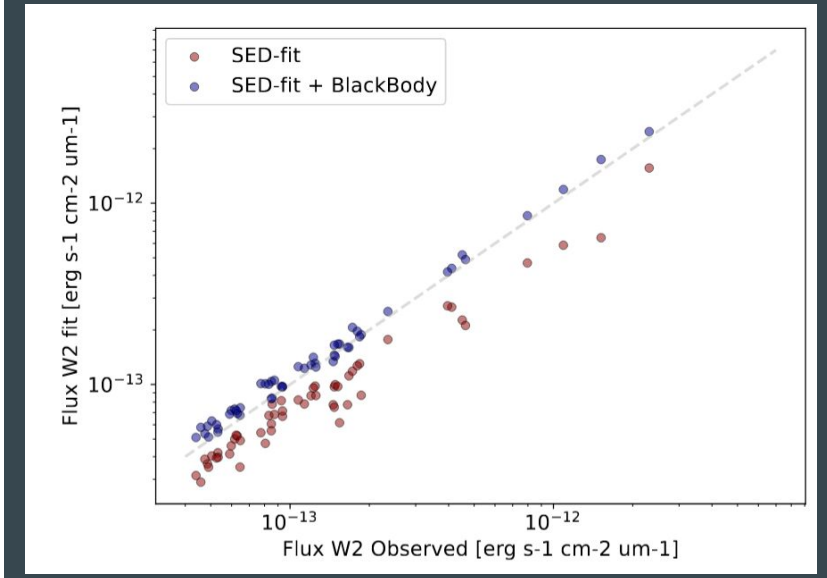
Anomaly Criterion Cuts – Additional cuts

- In addition of the error, precision, and well-predicted region cuts, we implement a serie of check to prevent potential false detection: **53 candidates** (out of 4.9M)

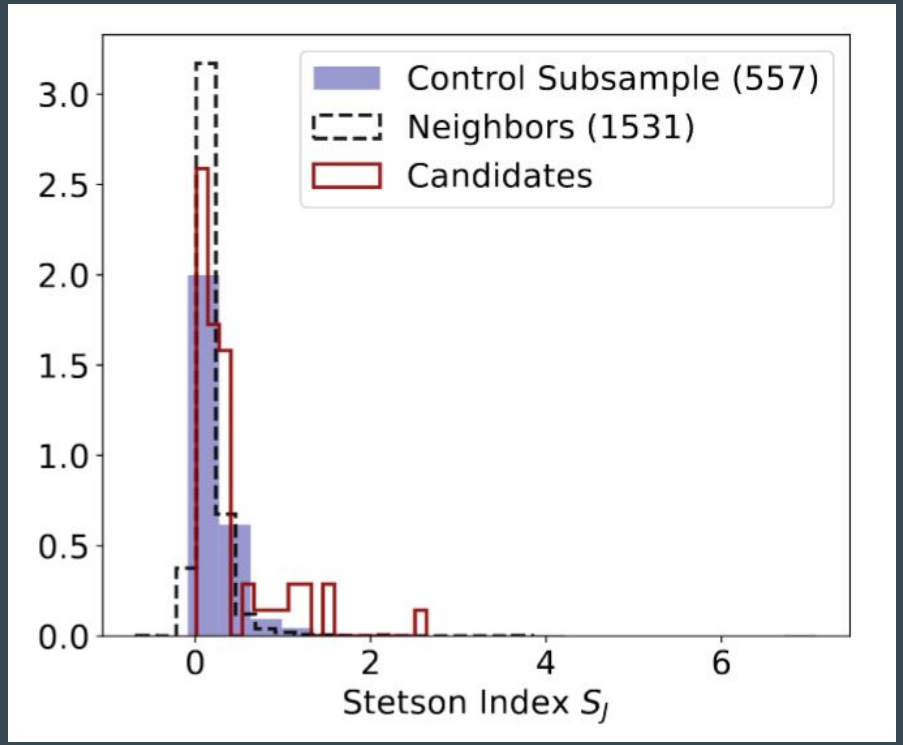
Criterion cut	Number of remaining candidates
Prediction Error cut (Eq. 3)	385
Mean (Prediction Error / k-MAD) cut (Eq. 4)	339
Error cut AND error/k-MAD cut	170
fold-MAD cut (Eq. 5)	127
Crowding cut at 5 arcsecond	87
FoM > 4 cut	87
Proper-Motion disagreement cut	78
Disagreement *AllWISE*/*unWISE* cut	76
Mean Distance k-NN $< .1$	66
$abs(b) > 10$	59
Removing binaries and binaries candidates (*Gaia*, Simbad)	55
Removing duplicated sources (*Gaia* DR3 flag)	53

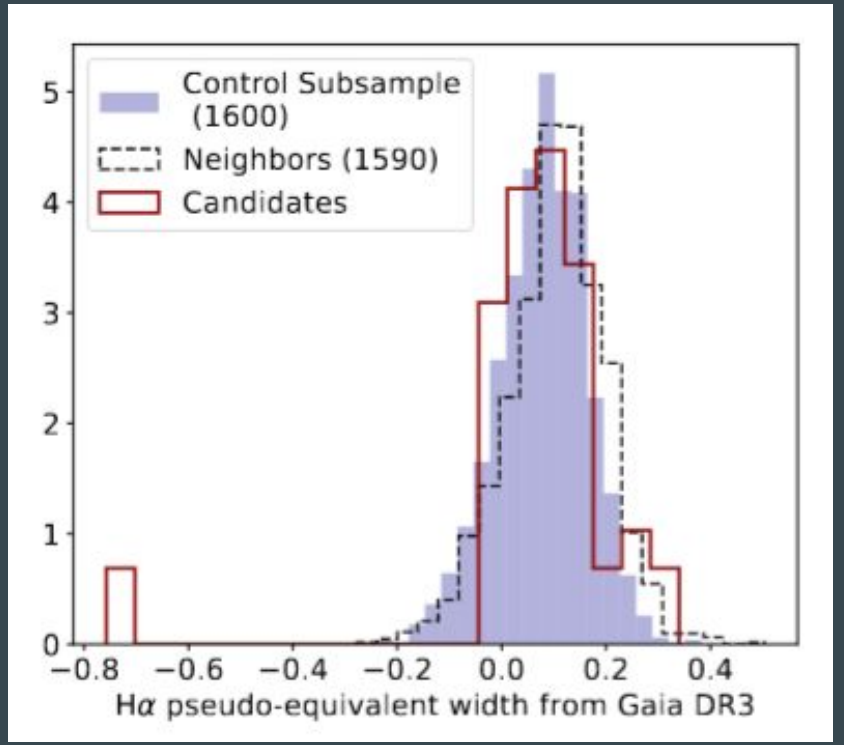
SED Fitting + Black Body fitting on (M)IR Residuals





Time-Variability and Hα emission of the Candidates



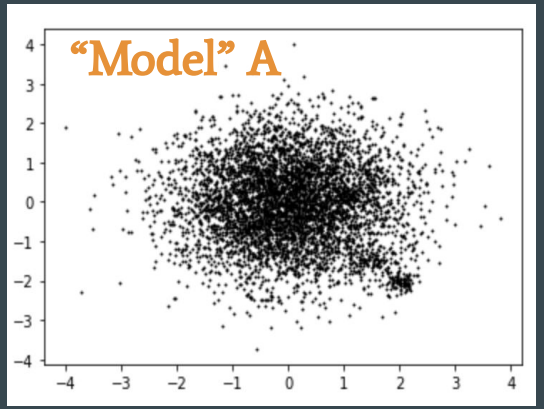


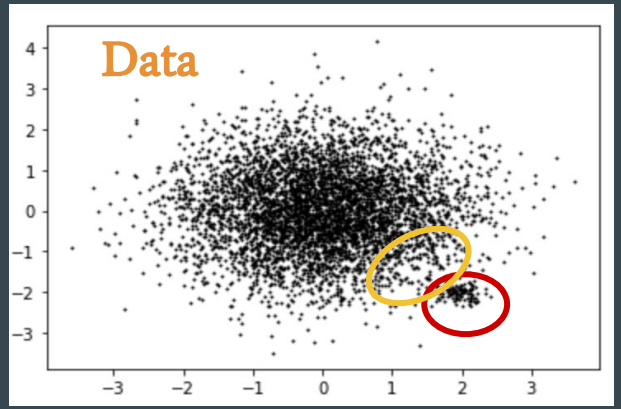
Discussion

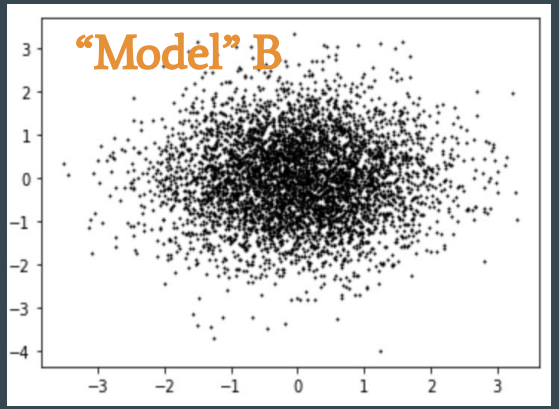
- A “methodological trick” to detect “contextual/conditional anomalies”
- Coherent trend when compared with doing conditional probability estimation: could be combined to find anomalies when the two tests disagree?
- Crosscheck with existing candidates show that we detect known-EDDs in our sample and all but 1 removed by our secondary cuts. Small overlap with previous searches.
- Large search compared to before. We could relax some cuts to yield potentially more candidates...
- **Next:** ideally get follow-up observations , deeper investigation for stellar age estimation (<- hard!).

Digression on the concept of Anomalies

- Some scientific communities focuses on “point-wise” anomalies (low density / rare)
- Other communities interested in finding divergences between distribution of observations vs “model” (e.g. bumps) // related to collective anomaly
- (Blind) search looking only for “outliers”: at risk of missing interesting things!?
- But what do we do when we don’t have (good) models? (Non-trivial even if we do have them!? Esp. in high dimension)
- Topographical features, class discovery, dimensionality reduction and what they preserve, ...



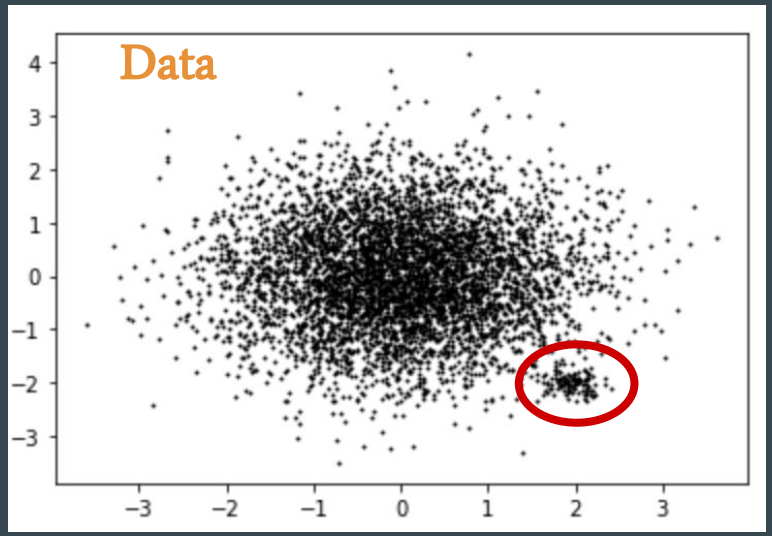


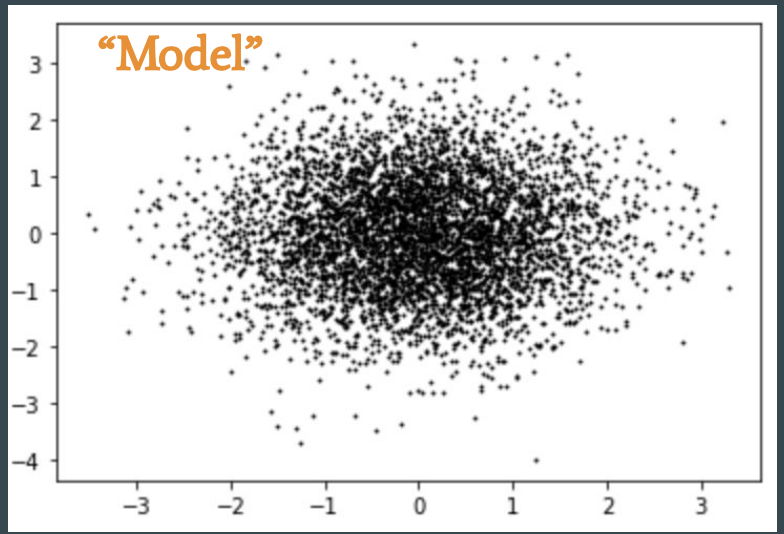


Thank you! Questions?

Digression on the concept of Anomalies

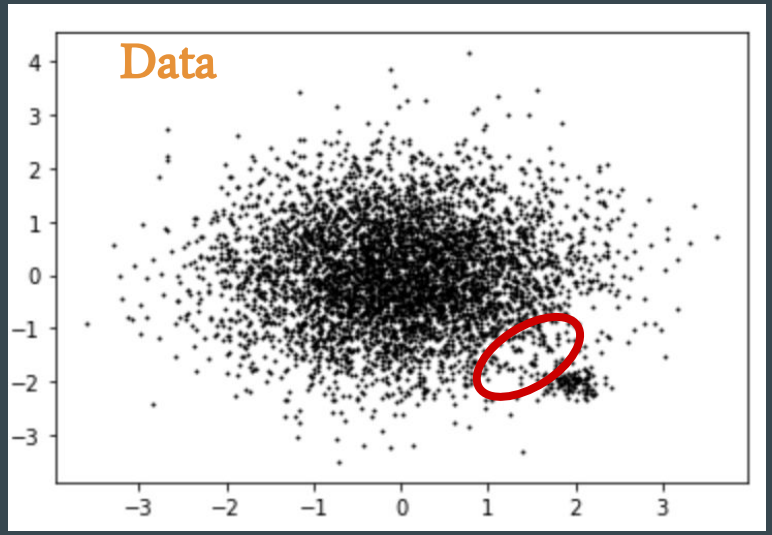
- Are anomalies always single-data-point with a low probability / density estimate wrt the data?

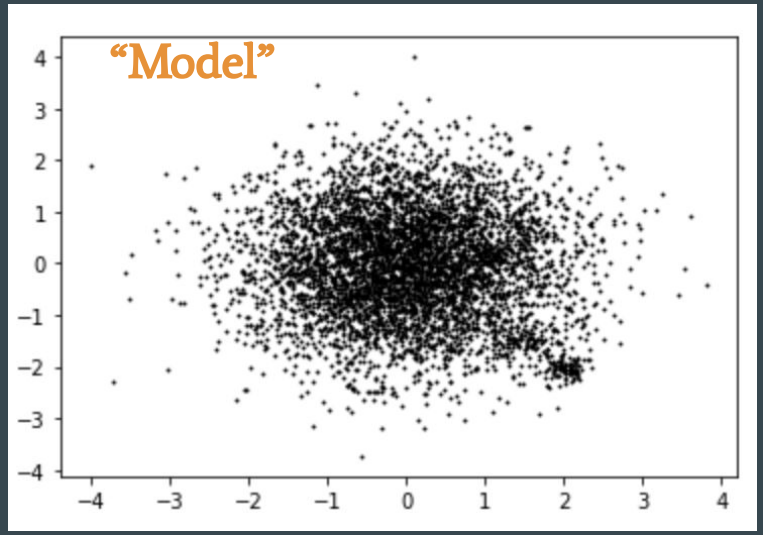




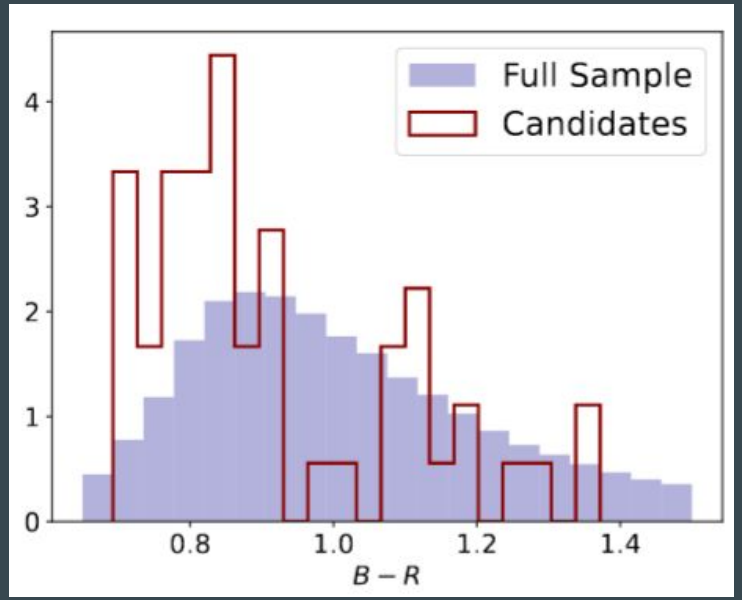
Digression on the concept of Anomalies

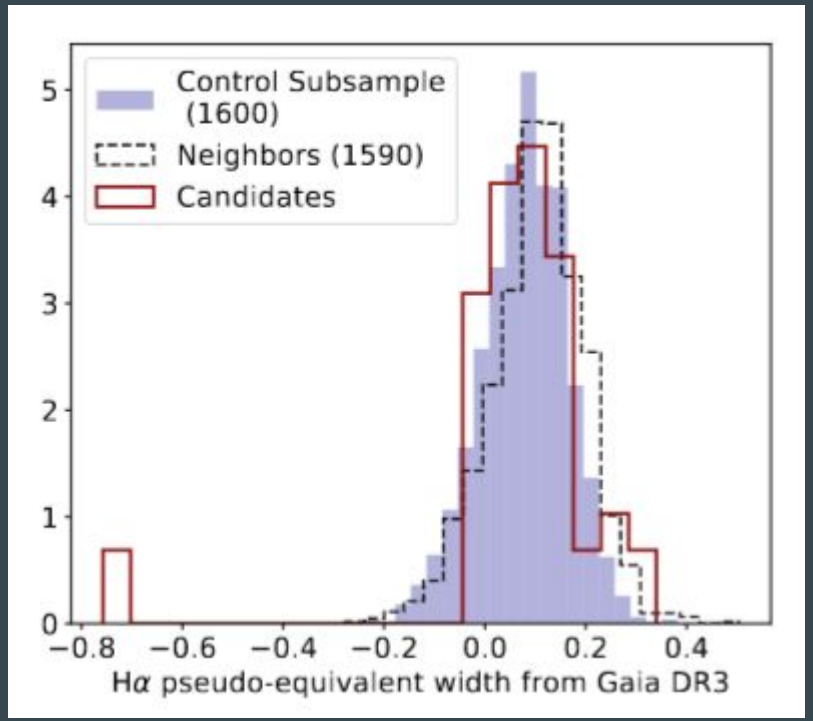
- Are anomalies always single-data-point with a low probability / density estimate wrt the data?
- Class Discovery,





Investigating the candidates population





Recovery Rate of Black Bodies