Data Mining in Astronomy

From Radio to Gamma Rays

Sabrina Einecke

THE UNIVERSITY of ADELAIDE

CTA-Linkage Meeting Adelaide
Nov 29th, 2019

Typical Process

Problem Definition
What would I like to find out?

Translation to Machine Learning
Determine category of machine learning problem.

Data Preparation
Create dataset.
Transformation of data.
Feature generation.
Feature pre-selection.

Exploratory Data Analysis
Statistical and visual analysis.
Outlier handling.
Error handling.

Modeling
Choose algorithm.
Tune settings.
Feature selection.

Validation
Determine performance.
Determine robustness.

Typical Process

Problem Definition
What would I like to find out?

Translation to Machine Learning
Determine category of machine learning problem.

Data Preparation
Create dataset.
Transformation of data.
Feature generation.
Feature pre-selection.

Exploratory Data Analysis
Statistical and visual analysis.
Outlier handling.
Error handling.

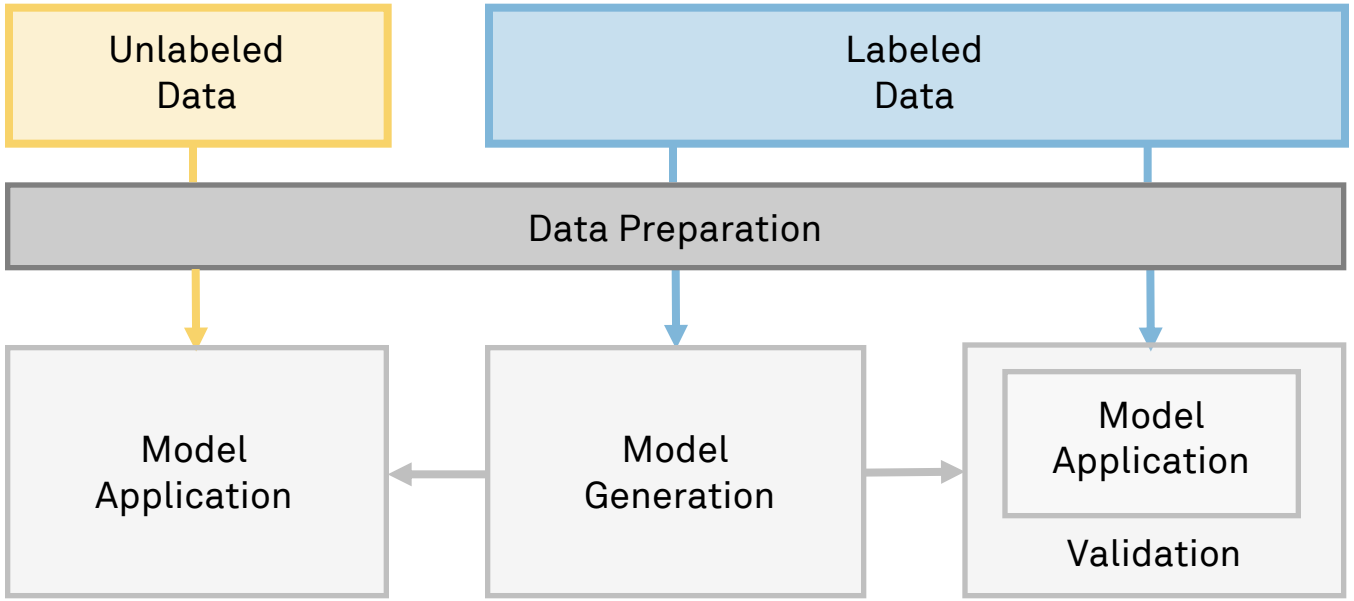
80% of work

Modeling
Choose algorithm.
Tune settings.
Feature selection.

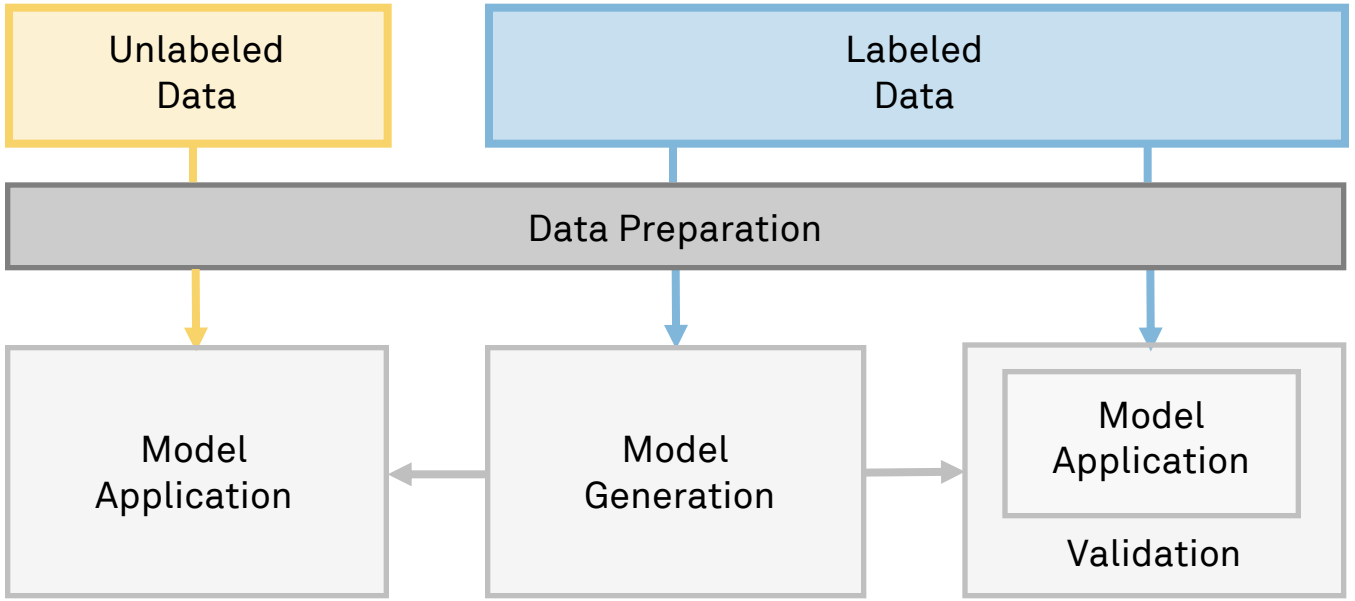
Validation
Determine performance.
Determine robustness.

20% of work

Supervised Learning

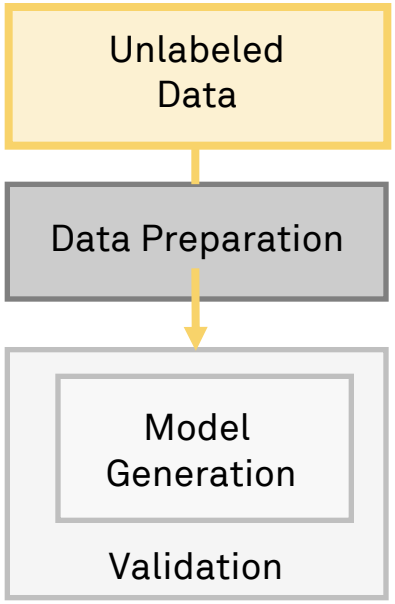


Supervised Learning



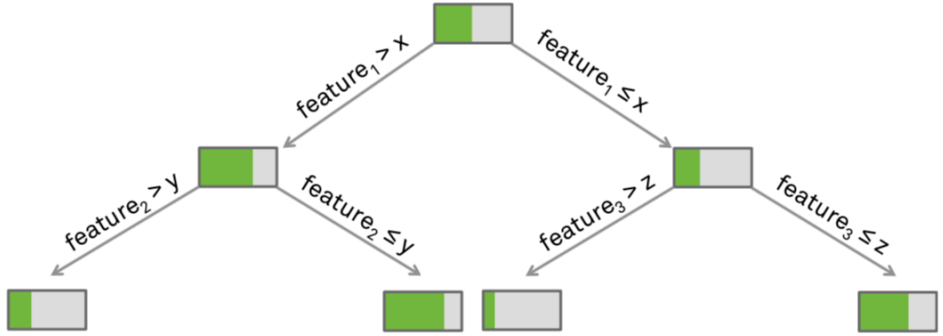
- Labels from simulations
- Labels from conventional methods etc.
- Labels from humans (citizen science!)

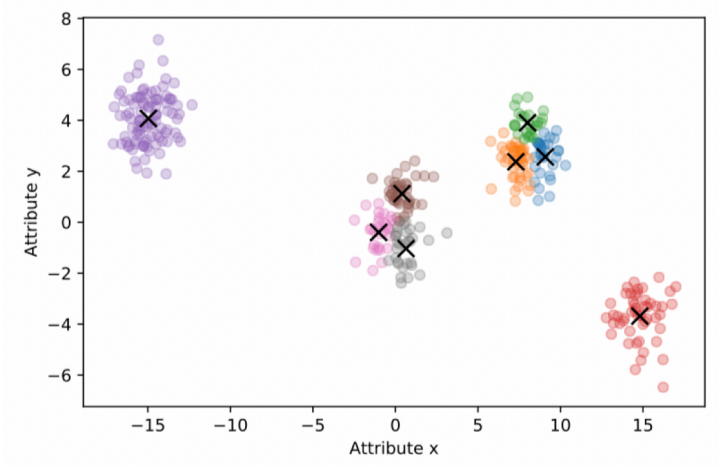
Unsupervised Learning

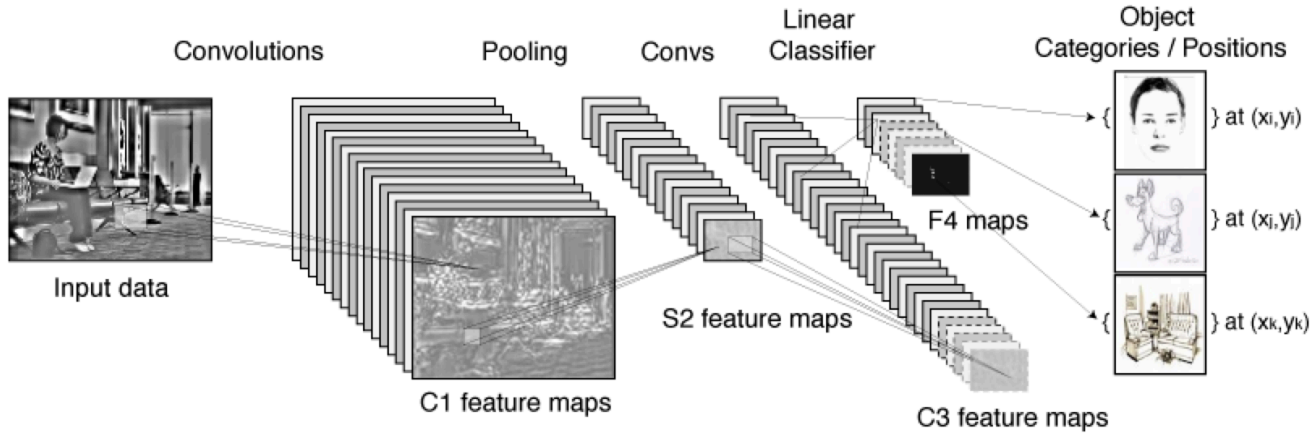


Machine Learning Categories

- Classification
- Regression
- Clustering
- Data Reduction
- Pattern Recognition
- Co-Occurrence Grouping
- ...





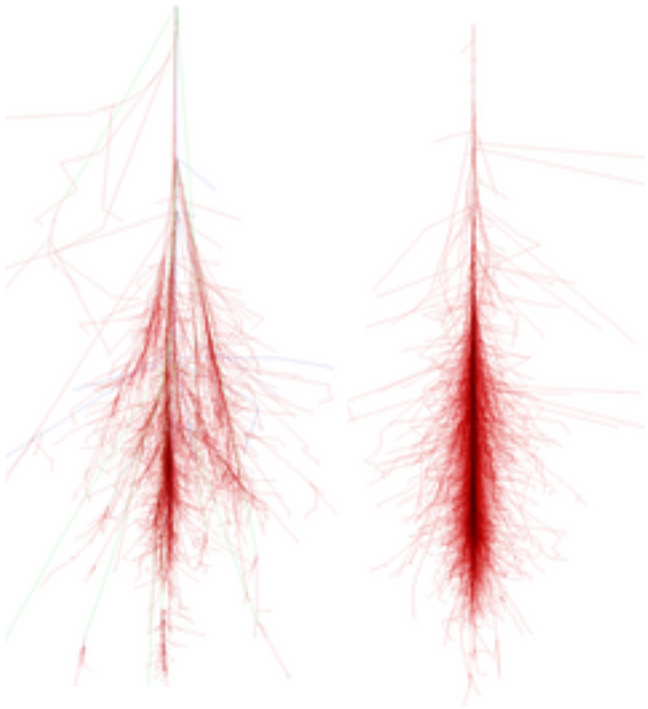


Classification / Regression

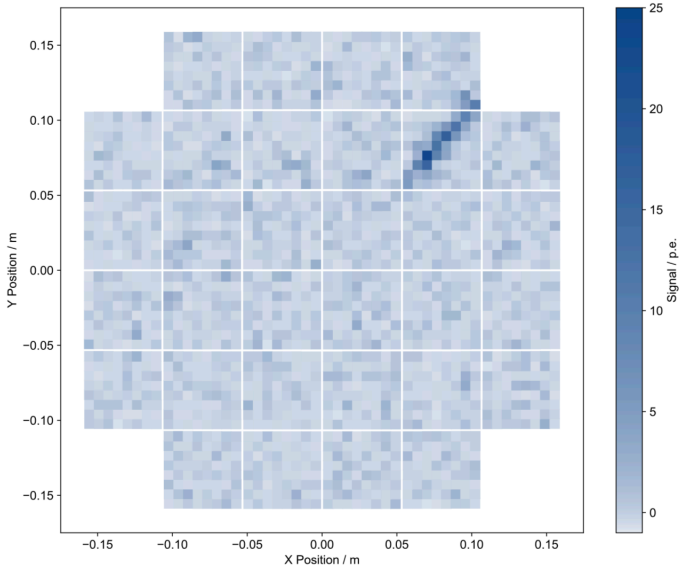
- **Aim:** Prediction of multiple classes / continuous parameter
- **Example:** Reconstruction in gamma-ray and neutrino astronomy

Random Forest
Neural Networks
k-Nearest Neighbours

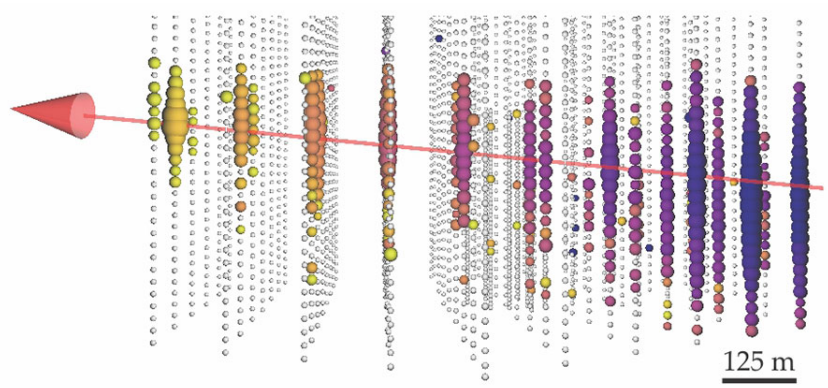
Proton vs. Gamma



Energy



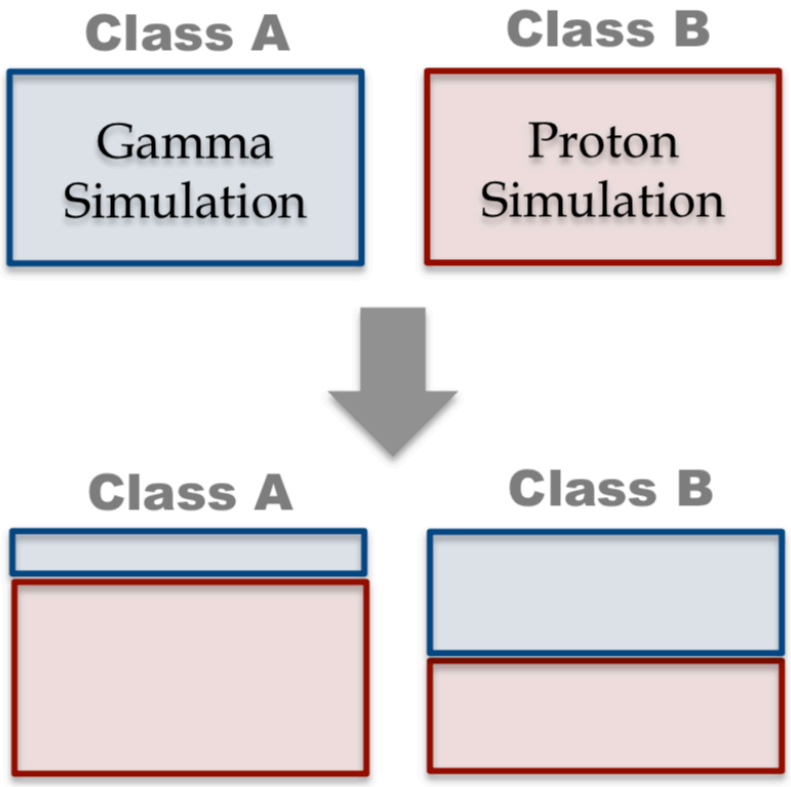
Direction



Classification / Regression

- **Aim:** Prediction of multiple classes / continuous parameter
- **Example:** Gamma / Hadron separation without simulations

Random Forest
Neural Networks
k-Nearest Neighbours



Default (supervised) training:
Gamma vs. Proton

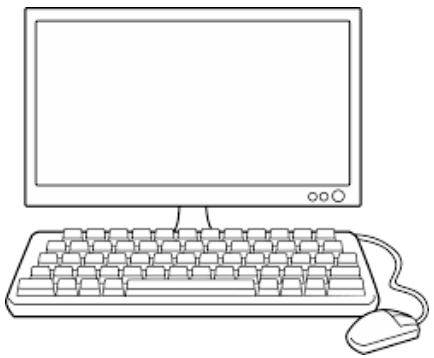
Weakly-supervised training:
On vs. Off region

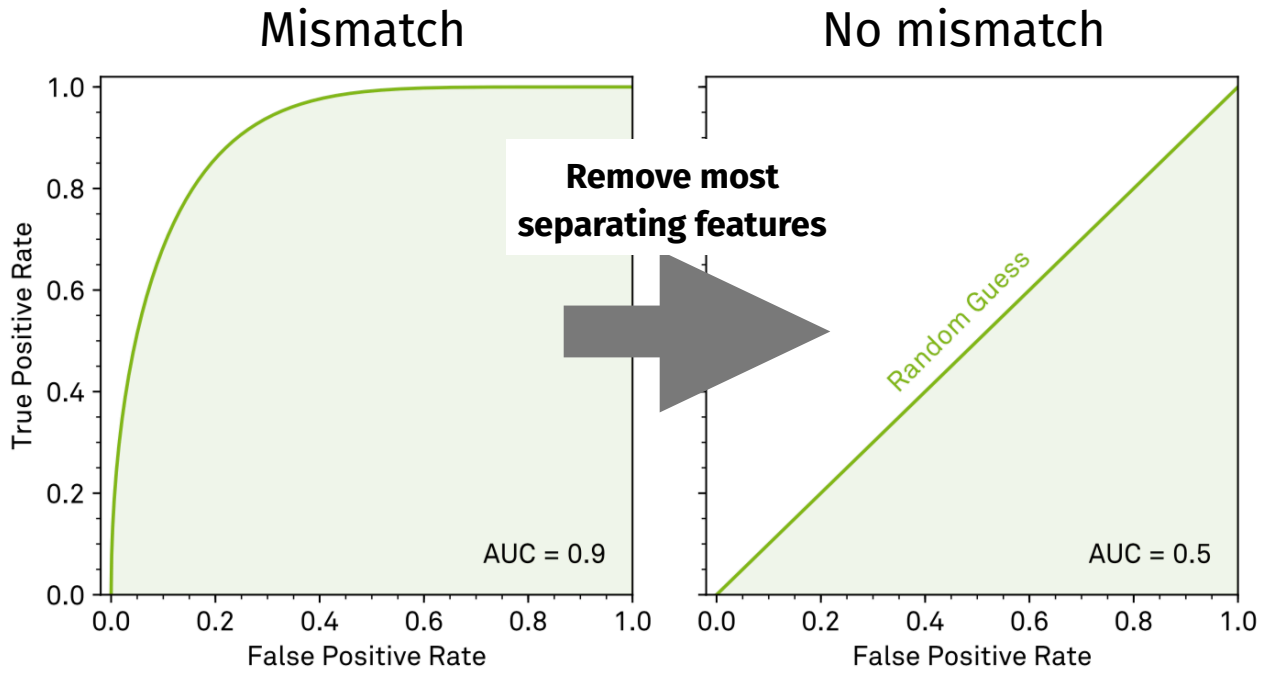
Classification / Regression

- **Aim:** Prediction of multiple classes / continuous parameter
- **Example:** Determination of (multi-variate) mismatches between data and simulation

Random Forest
Neural Networks
k-Nearest Neighbours

Training:
Simulation vs. Data



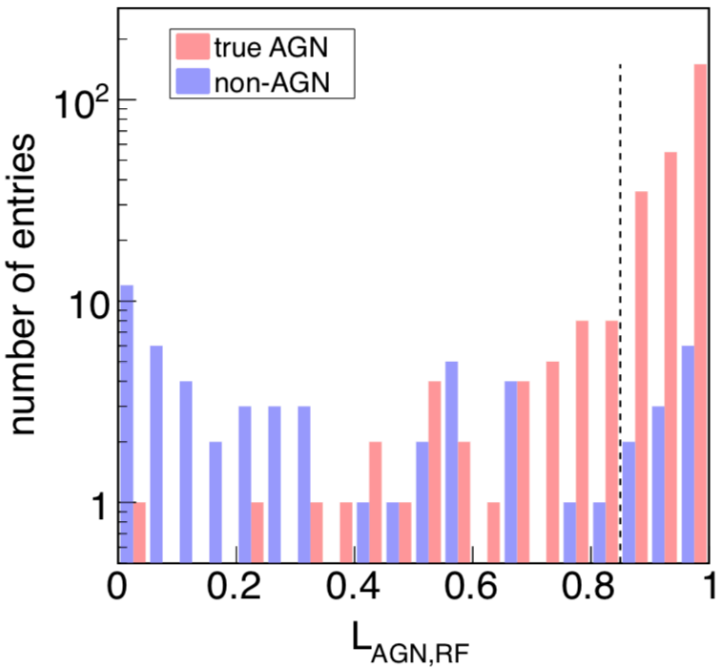


Classification / Regression

- **Aim:** Prediction of multiple classes / continuous parameter
- **Example:** Determination of source type based on catalog features

Random Forest
Neural Networks
k-Nearest Neighbours

AGN vs Non-AGN



https://arxiv.org/abs/1312.5726

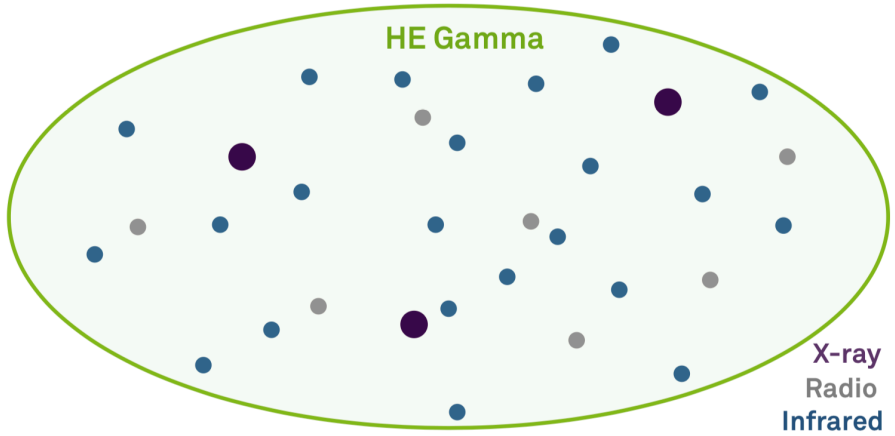
***Fermi*-LAT Source Catalog**

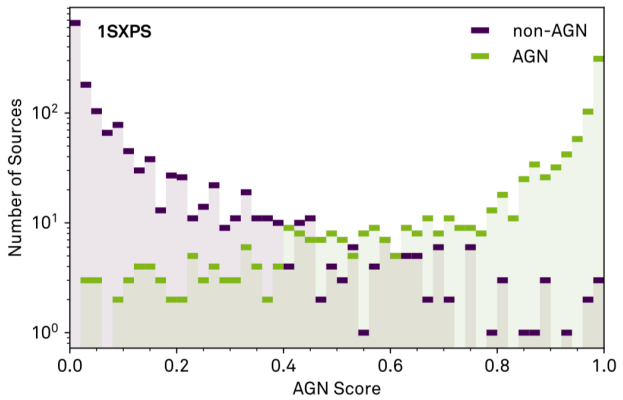
name	ra	dec	flux 1 100 gev [photon/cm^2/s]	flux 1 100 gev error [photon/cm^2/s]	spectral index	spectral index error	detection significance
3FGL J0542.2-8737	05 42 14.5	-87 37 07	4.17030e-10	8.23503e-11	2.03895	0.14339	6.4319139
3FGL J2108.6-8619	21 08 39.1	-86 19 03	1.97481e-10	8.12151e-11	1.74000	0.26993	4.7285366
3FGL J1026.4-8542	10 26 25.2	-85 42 55	9.07417e-10	1.01188e-10	2.01336	0.08438	13.4339647
3FGL J2337.2-8425	23 37 15.2	-84 2				0.23347	4.7041693
3FGL J0046.7-8419	00 46 45.0	-84 1				0.13341	6.2397938
3FGL J2202.4-8339	22 02 26.4	-83 39 22	1.98576e-09	1.43511e-10	2.42962	0.06724	23.0740681
3FGL J2237.5-8326	22 37 33.9	-83 26 14	5.32938e-10	9.96599e-11	2.43369	0.14117	6.1225739
3FGL J0533.6-8323	05 33 38.5	-83 23 05	8.11060e-10	9.91749e-11	2.29598	0.07827	11.6992741
3FGL J1036.0-8317	10 36 05.3	-83 17 17	4.65727e-10	8.58584e-11	2.33022	0.11465	6.6513314
3FGL J1224.6-8312	12 24 39.7	-83 12 33	4.48758e-10	8.75808e-11	2.69452	0.10959	7.4709144

Classification / Regression

- **Aim:** Prediction of multiple classes / continuous parameter
- **Example:** Determination of counterpart, based on multiple catalogs

Random Forest
Neural Networks
k-Nearest Neighbours





name	ra	dec	flux 1 100 gev [photon/cm^2/s]	flux 1 100 gev error [photon/cm^2/s]	spectral index	spectral index error	detection significance
3FGL J0542.2-8737	05 42 14.5	-87 37 07	4.17030e-10	8.23503e-11	2.03895	0.14339	6.4319139
3FGL J2108.6-8619	21 08 39.1	-86 19 03	1.97481e-10	8.12151e-11	1.74000	0.26993	4.7285366
3FGL J1026.4-8542	10 26 25.2	-85 42 55	9.07417e-10	1.01188e-10	2.01336	0.08438	13.4339647
3FGL J2337.2-8425	23 37 15.2	-84 2				0.23347	4.7041693
3FGL J0046.7-8419	00 46 45.0	-84 1				0.13341	6.2397938
3FGL J2202.4-8339	22 02 26.4	-83 39 22	1.98576e-09	1.43511e-10	2.42962	0.06724	23.0740681
3FGL J2237.5-8326	22 37 33.9	-83 26 14	5.32938e-10	9.96599e-11	2.43369	0.14117	6.1225739
3FGL J0533.6-8323	05 33 38.5	-83 23 05	8.11060e-10	9.91749e-11	2.29598	0.07827	11.6992741

***Fermi*-LAT Source Catalog**

RA	Decl	Err90	AstromType	l	b	OffAxis	DetFlag	Fieldflag	Rate_band0	HR1
1h 04m 27.27s	+38° 12′ 32.3″	3.5	0	179.83147	65.0313	1.1	0	0	45.1	-0.226
1h 04m 43.81s	+38° 14′ 48.7″	3.6	0	179.70554	65.07064	5	0	0	0.00807	-0.973
1h 04m 20.74s	+38° 04′ 46.3″	7.4					2	0	0.000576	-0.32
1h 04m 18.13s	+38° 20′ 46.6″	4.6						0	0.00173	-0.829
1h 04m 01.27s	+37° 43′ 28.7″	4.2						2	0.00483	-0.218
1h 03m 55.56s	+37° 43′ 35.0″	10.2	1	180.99643	65.07998	11.9	10	2	0.00609	-0.204
1h 04m 14.74s	+37° 42′ 39.6″	5.1	1	180.98661	65.14491	12.5	8	2	0.00209	0.536

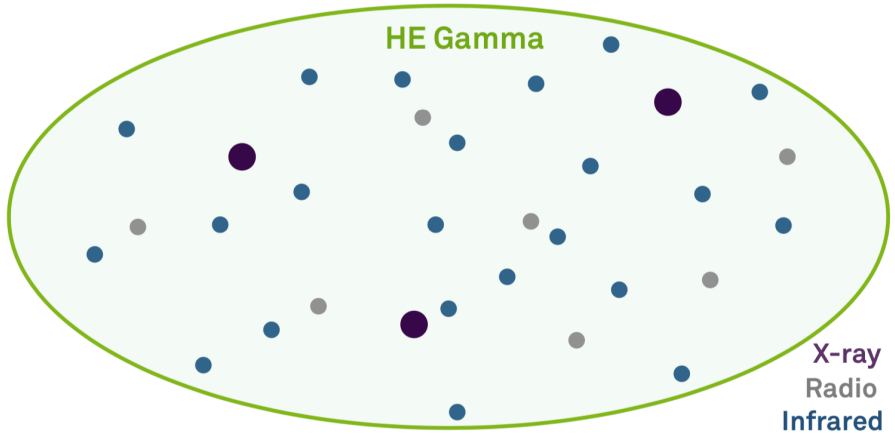
***Swift*-XRT Source Catalog**

Sabrina Einecke

Classification / Regression

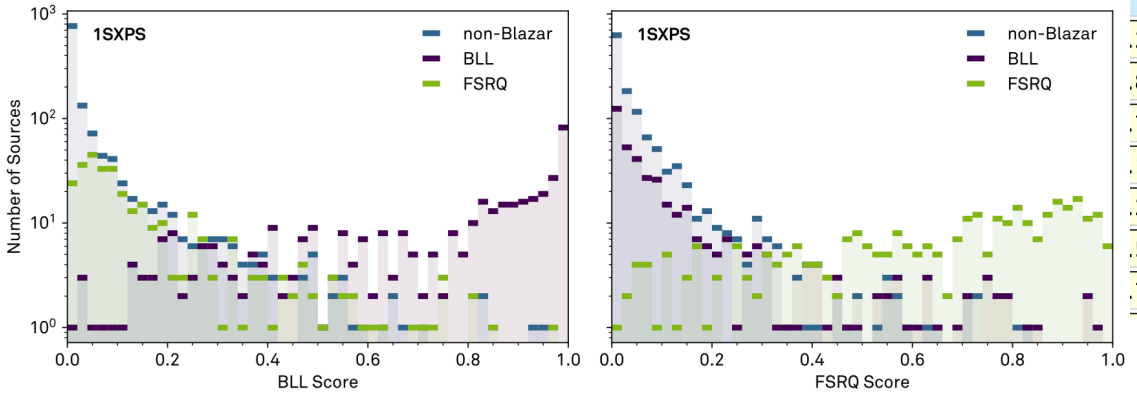
- **Aim:** Prediction of multiple classes / continuous parameter
- **Example:** Determination of counterpart, based on multiple catalogs

Random Forest
Neural Networks
k-Nearest Neighbours



***Fermi*-LAT Source Catalog**

name	ra	dec	flux 1 100 gev [photon/cm^2/s]	flux 1 100 gev error [photon/cm^2/s]	spectral index	spectral index error	detection significance
3FGL J0542.2-8737	05 42 14.5	-87 37 07	4.17030e-10	8.23503e-11	2.03895	0.14339	6.4319139
3FGL J2108.6-8619	21 08 39.1	-86 19 03	1.97481e-10	8.12151e-11	1.74000	0.26993	4.7285366
3FGL J1026.4-8542	10 26 25.2	-85 42 55	9.07417e-10	1.01188e-10	2.01336	0.08438	13.4339647
3FGL J2337.2-8425	23 37 15.2	-84 2				0.23347	4.7041693
3FGL J0046.7-8419	00 46 45.0	-84 1				0.13341	6.2397938
3FGL J2202.4-8339	22 02 26.4	-83 39 22	1.98576e-09	1.43511e-10	2.42962	0.06724	23.0740681
3FGL J2237.5-8326	22 37 33.9	-83 26 14	5.32938e-10	9.96599e-11	2.43369	0.14117	6.1225739
3FGL J0533.6-8323	05 33 38.5	-83 23 05	8.11060e-10	9.91749e-11	2.29598	0.07827	11.6992741



***Swift*-XRT Source Catalog**

	Decl	Err90	AstromType	l	b	OffAxis	DetFlag	Fieldflag	Rate_band0	HR1
.27s	+38° 12′ 32.3″	3.5	0	179.83147	65.0313	1.1	0	0	45.1	-0.226
.81s	+38° 14′ 48.7″	3.6	0	179.70554	65.07064	5	0	0	0.00807	-0.973
.74s	+38° 04′ 46.3″	7.4					2	0	0.000576	-0.32
.13s	+38° 20′ 46.6″	4.6					0	0	0.00173	-0.829
.27s	+37° 43′ 28.7″	4.2					9	2	0.00483	-0.218
.56s	+37° 43′ 35.0″	10.2	1	180.99643	65.07998	11.9	10	2	0.00609	-0.204
.74s	+37° 42′ 39.6″	5.1	1	180.98661	65.14491	12.5	8	2	0.00209	0.536

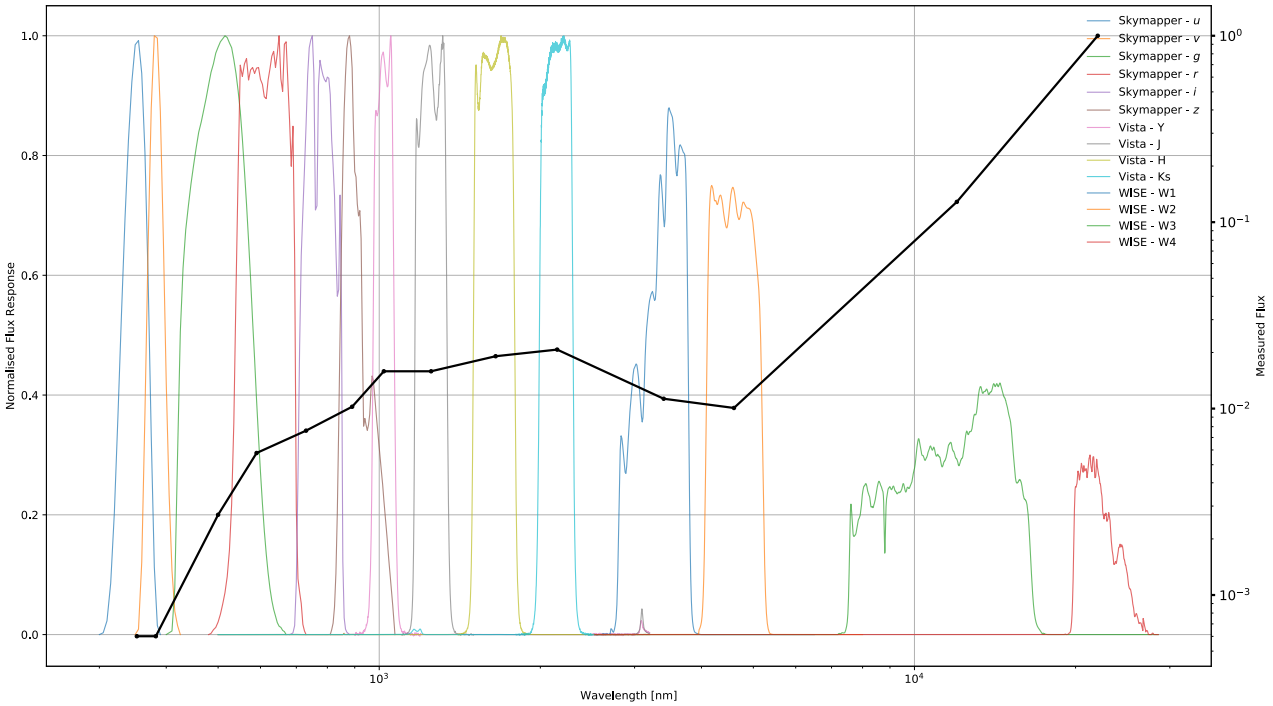
Sabrina Einecke

Classification / Regression

- **Aim:** Prediction of multiple classes / continuous parameter
- **Example:** Estimation of redshift from photometry

Random Forest
Neural Networks
k-Nearest Neighbours

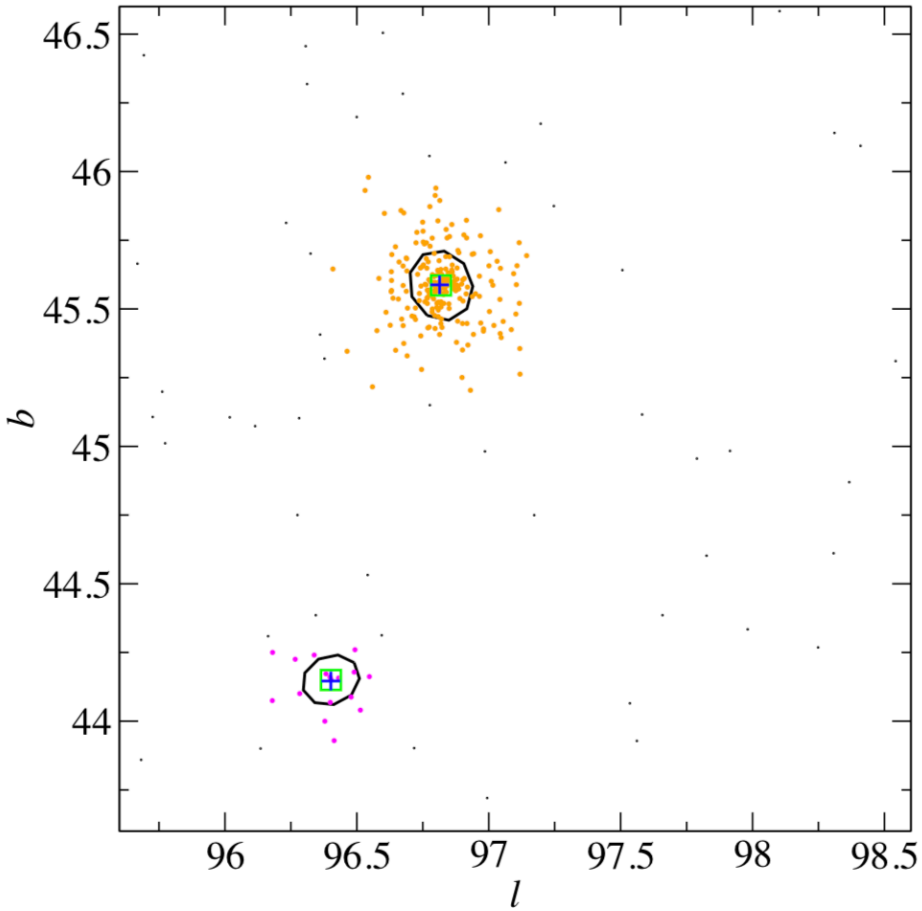
- kNN Regression - ∼6%
- Random Forest Regression - ∼10%
- kNN Classification - ∼5%
- Random Forest Classification - ∼8%



Kieran Luken, Ray Norris, Miroslav Filipovic

Clustering

- **Aim:** Group data by their similarity
- **Example:** Detection of sources, based on a set of measured photons



k-means
Gaussian Mixture Model
DBSCAN

https://arxiv.org/abs/1210.0522

Data Reduction

Principal Component Analysis
Autoencoder

- **Aim:** Reduction of features / dimensions (and keeping most of the information)
- **Example:** Gamma-ray camera image -> Camera parameters

Camera Image

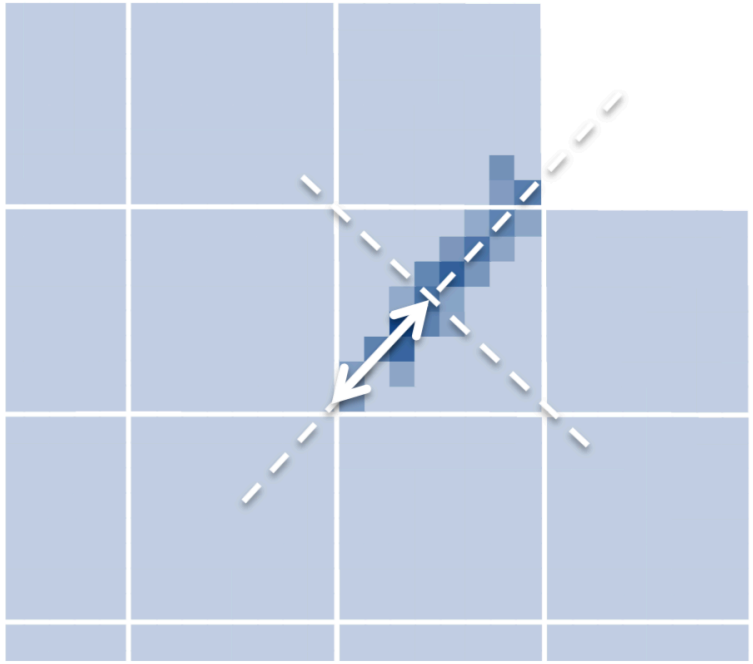
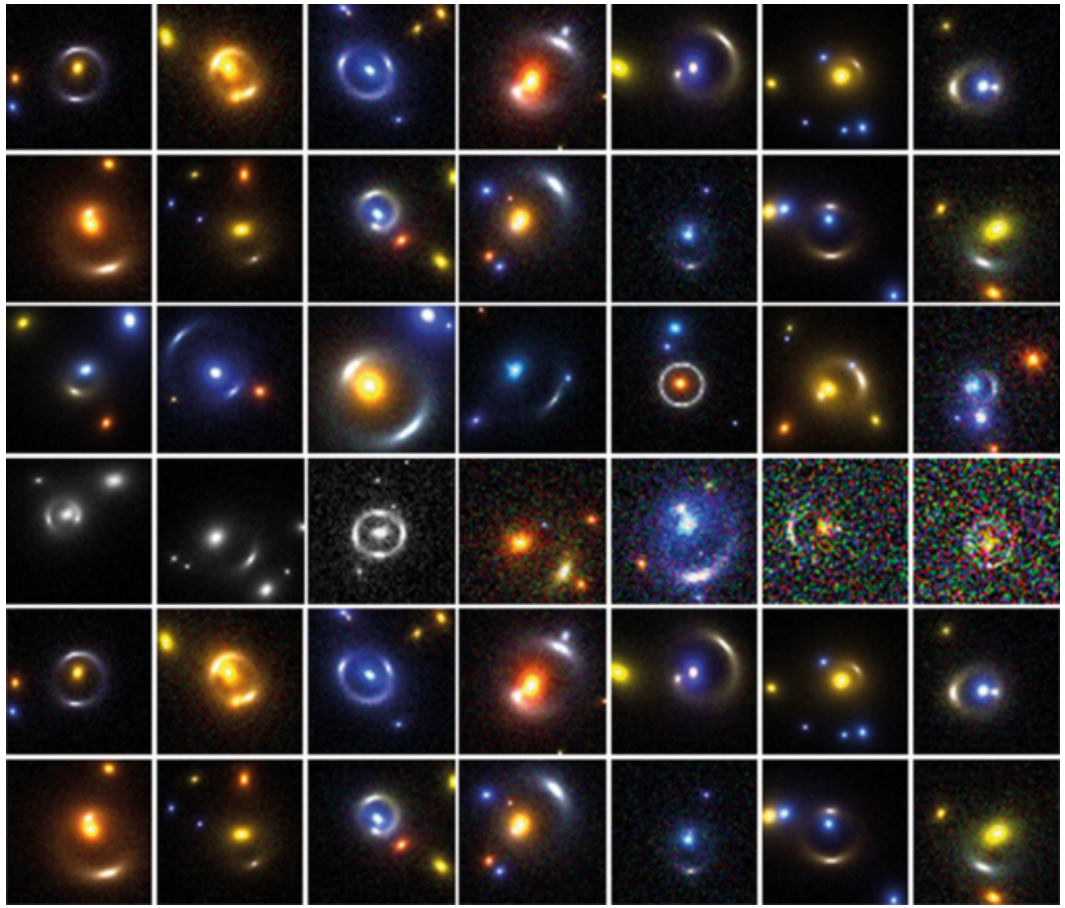


Image Parameters

telescope_events
array_event_id
run_id
intensity
x
y
r
phi
length
width
psi
skewness
kurtosis

Pattern Recognition

- **Aim:** Classify structures in data / images
- **Example:** Identification of gravitational lenses

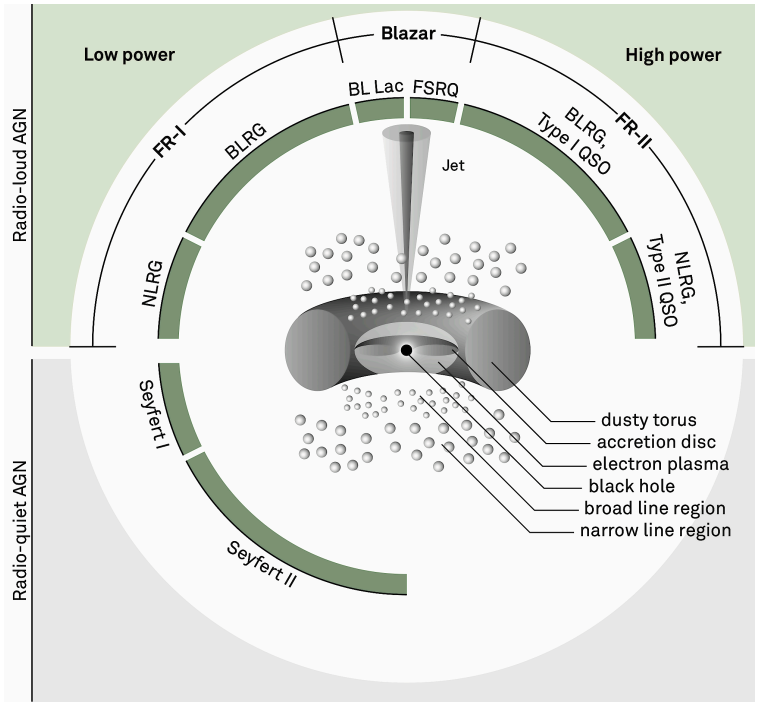


https://doi.org/10.1051/emsci/2017010

Co-Occurrence Grouping

- **Aim:** Discover relations between variables
- **Example:** ?

Population Studies



Identification of Potential Targets

If	Then
Luminosity (X-ray) > X A < Peak frequency < B Spectral Index (HE) > S Redshift < R	VHE counterpart

If	Then
Magnetic Field > Y C < Peak frequency < D Spectral Index (HE) > I Redshift < Z	PeVatron candidate

Conclusion

- Data Science / Machine Learning is becoming an integral part in exploring astronomical and astrophysical data!
- We have to define the problems!
- We have to support the methods with our expert knowledge!