ML Workflow

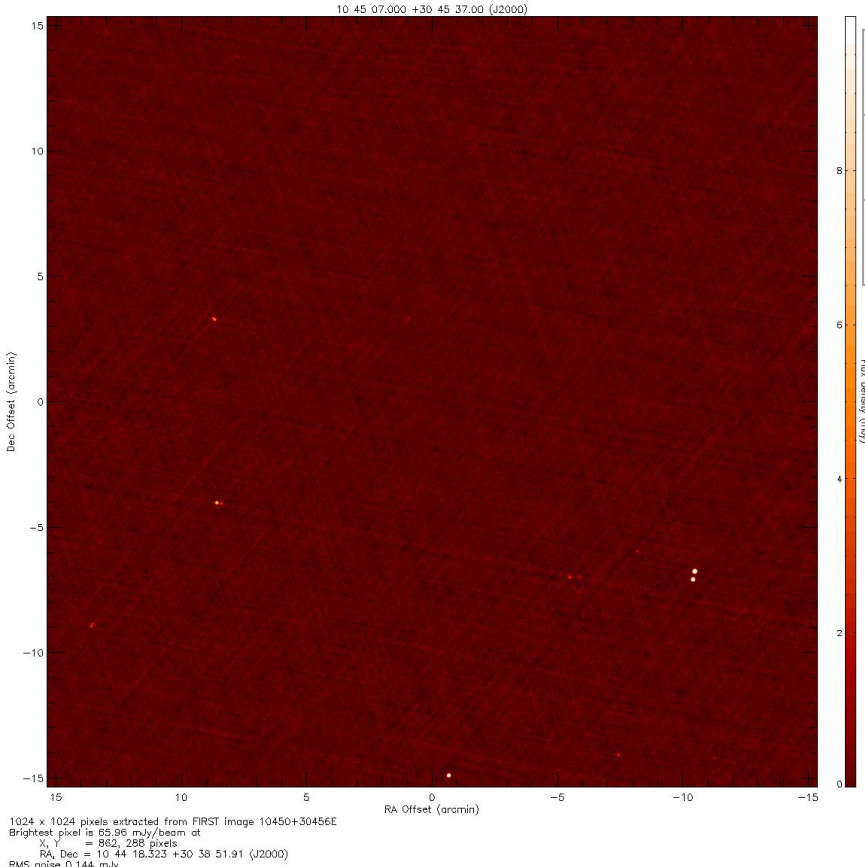
An **example** and **checklist**
to guide you in your own projects.

Checklist

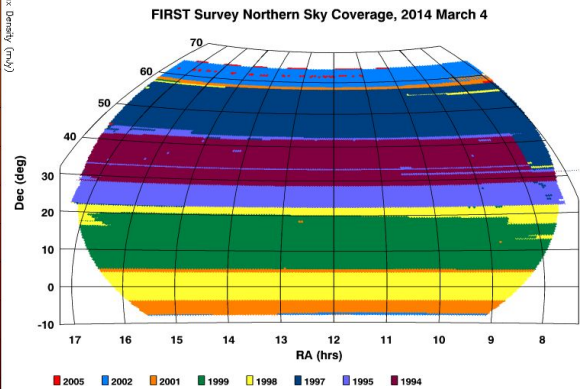
1. **Frame** the problem and look at the big picture
2. Get the **data**
3. **Explore** the data to get insights
4. **Prepare** the data to better expose the underlying data patterns to machine learning algorithms
5. **Explore** many different models and short-list the best ones.
6. **Fine-tune** your models and combine them to a great solution.
7. **Present** your solution
8. **Launch**, monitor and maintain your system.

This talk follows Appendix B of: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

1. Frame the Problem and Look at the Big Picture



VLA Data Rate	SKA Data Rate
<100 MB/s (360 GB/hr)	0.5 – 1 TB/s (>1.3 PB/hr)
https://science.nrao.edu/facilities/vla/docs/manuals/oss/performance/tim-res	https://doi.org/10.1098/rsta.2019.0060

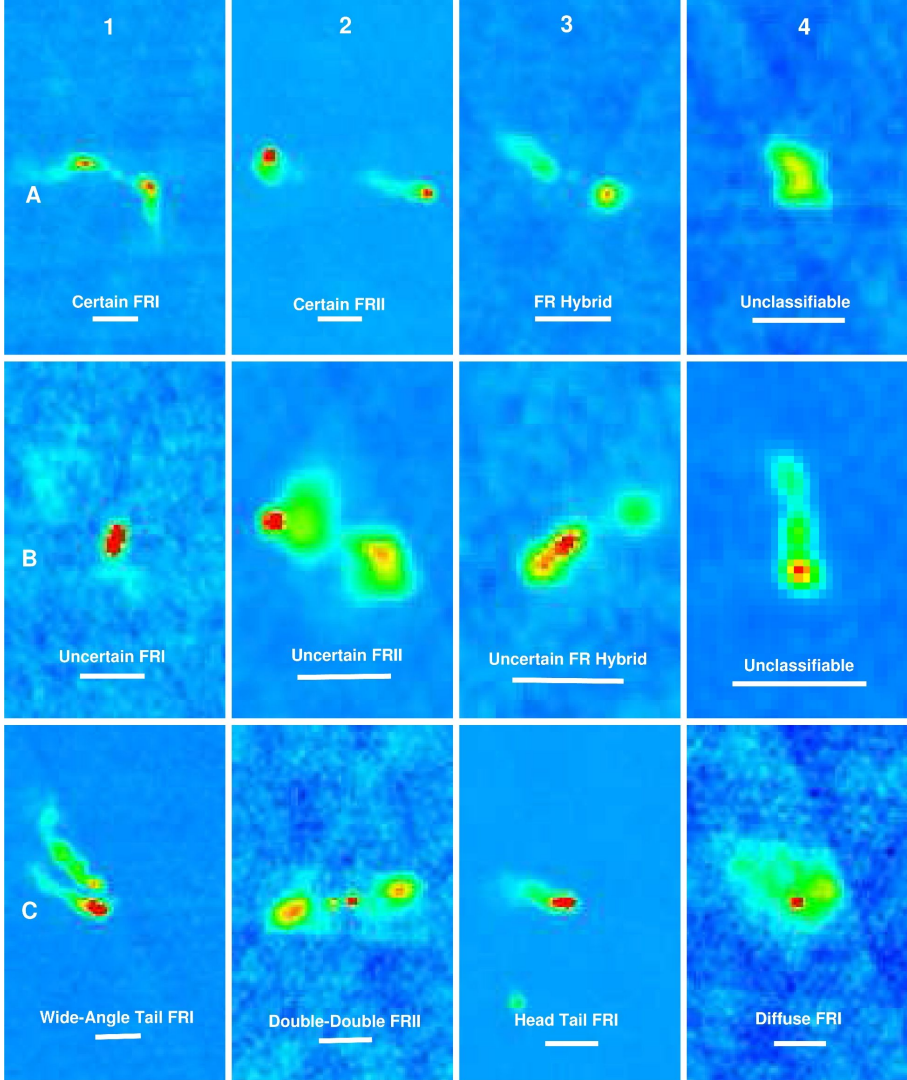


http://sundog.stsci.edu/index.html

2. Get Data

Source extractor, catalogues, private and public data sets.

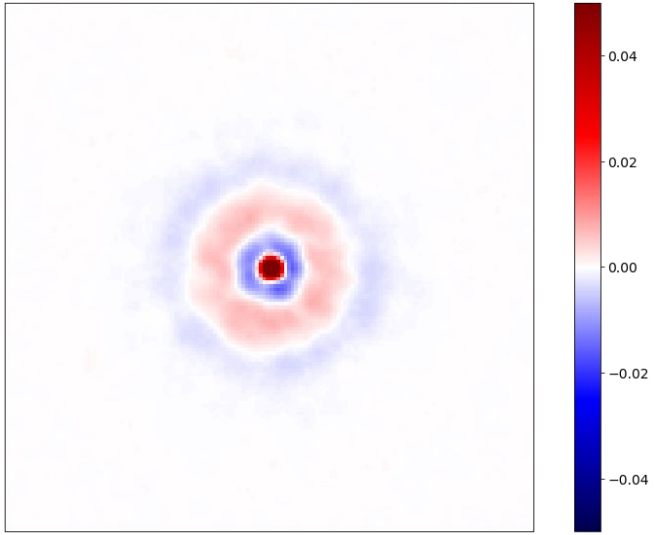
https://github.com/fmporter/MiraBest-full



3. Explore the Data to get Insights

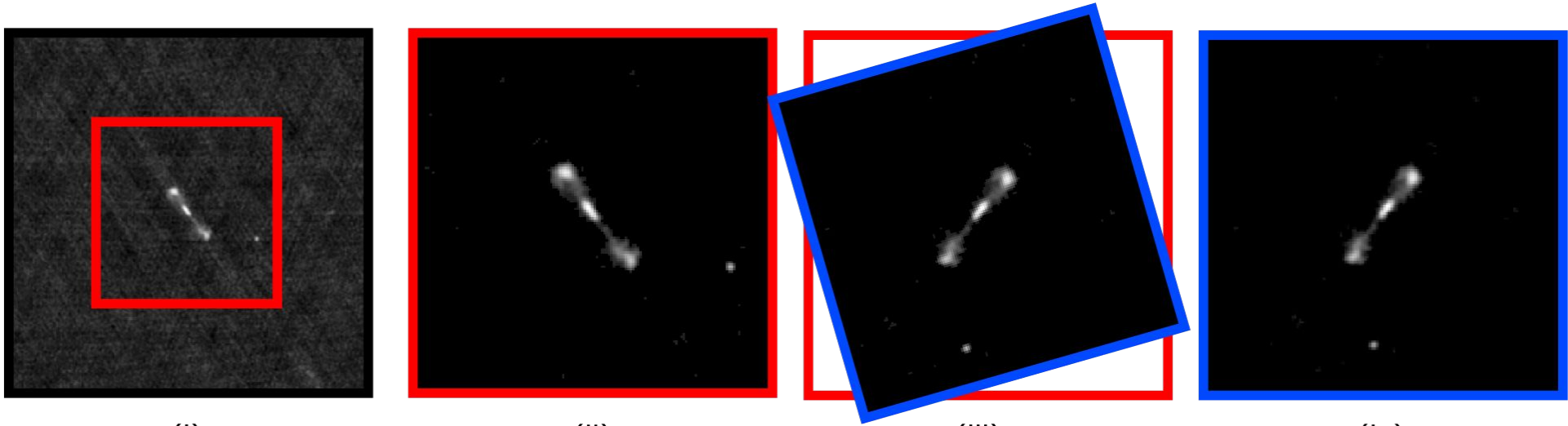
Look at the data.
Check biases.
Check distributions.

Label	No.	Class	Confidence	Morphology	No.	MiraBest Label
0	591	FRI	Certain	Standard	339	0
				Wide-Angle Tailed	49	1
				Head-Tail	9	2
			Uncertain	Standard	191	3
				Wide-Angle Tailed	3	4
1	631	FRII	Certain	Standard	432	5
				Double-Double	4	6
			Uncertain	Standard	195	7
NA	34	Hybrid	Certain	NA	19	8
			Uncertain	NA	15	9



4. Prepare the data [...]

“Feature Engineering” / “Cleaning”



5. Explore many different models [...]

Use your validation set.

Try ‘out of the box models’!

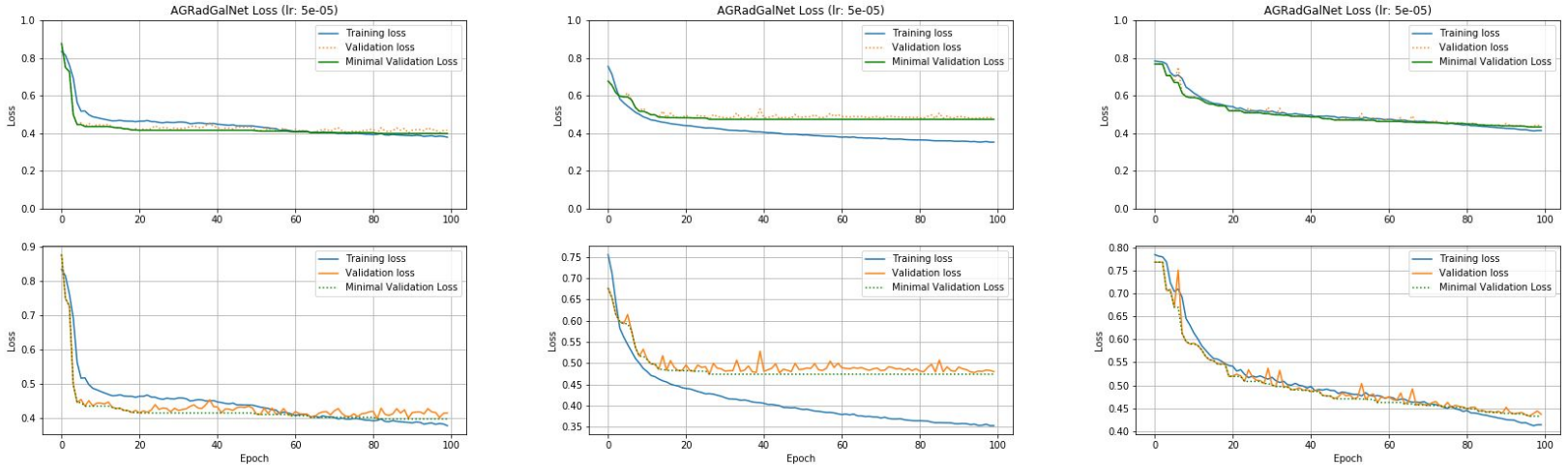
Norm. Class	Range Norm.		Standardisation		Sigmoid		Softmax	
F1 Score	0.81	0.83	0.69	0.71	0.83	0.84	0.85	0.87
Precision	0.84	0.80	0.72	0.70	0.85	0.83	0.89	0.84
Recall	0.77	0.87	0.67	0.73	0.81	0.86	0.81	0.91
Accuracy	82 %		70 %		84 %		86 %	
AUC	0.87		0.71		0.85		0.92	

Agg. Class	Mean		Concatenation		Deep Supervised		Fine Tuned	
F1 Score	0.78	0.82	0.82	0.84	0.79	0.78	0.79	0.82
Precision	0.85	0.77	0.86	0.81	0.77	0.80	0.83	0.79
Recall	0.72	0.88	0.78	0.88	0.81	0.75	0.76	0.85
Accuracy	80 %		83 %		78 %		81 %	
AUC	0.83		0.86		0.82		0.85	

6. Fine-tune your models and combine them [...]

Grid search using your validation sets.

Build an ‘Ensemble’ (e.g. Multiple good models vote for final class)



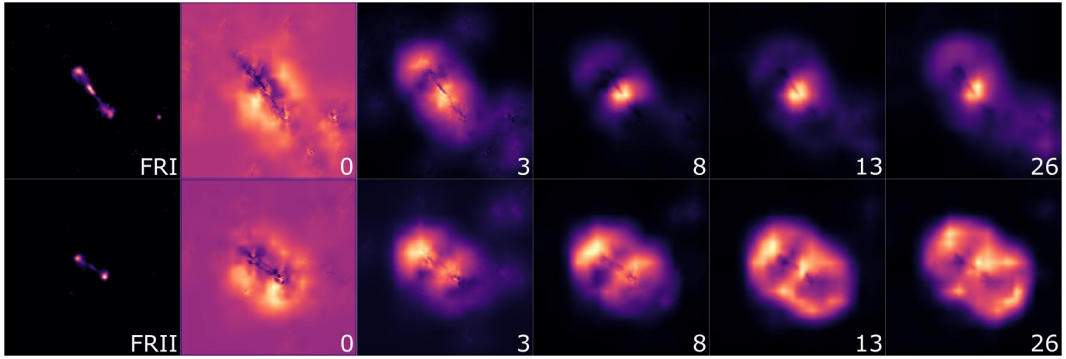
7. Present your solution

Present a (boring) table (but also …)

Network	Classic CNN		AG-CNN		AG-CNN		AG-CNN	
Data Set	FR-DEEP-F		FR-DEEP-F		MiraBest*		MiraBest	
Class	FRI	FRII	FRI	FRII	FRI	FRII	FRI	FRII
F1 Score	0.90 ± 0.03	0.88 ± 0.06	0.87	0.90	0.91	0.92	0.82	0.86
Precision	0.95 ± 0.02	0.83 ± 0.04	0.87	0.90	0.89	0.89	0.91	0.80
Recall	0.85 ± 0.02	0.94 ± 0.04	0.87	0.90	0.95	0.94	0.75	0.93
Accuracy	89 ± 1 %		88 %		92 %		84 %	
AUC	0.94		0.89		0.96		0.92	

Present **value**

(speed / reliability / adaptability / cost saving / scientific benefit)



8. Launch, monitor and maintain your system.

Use it!

Make it (publicly) available!

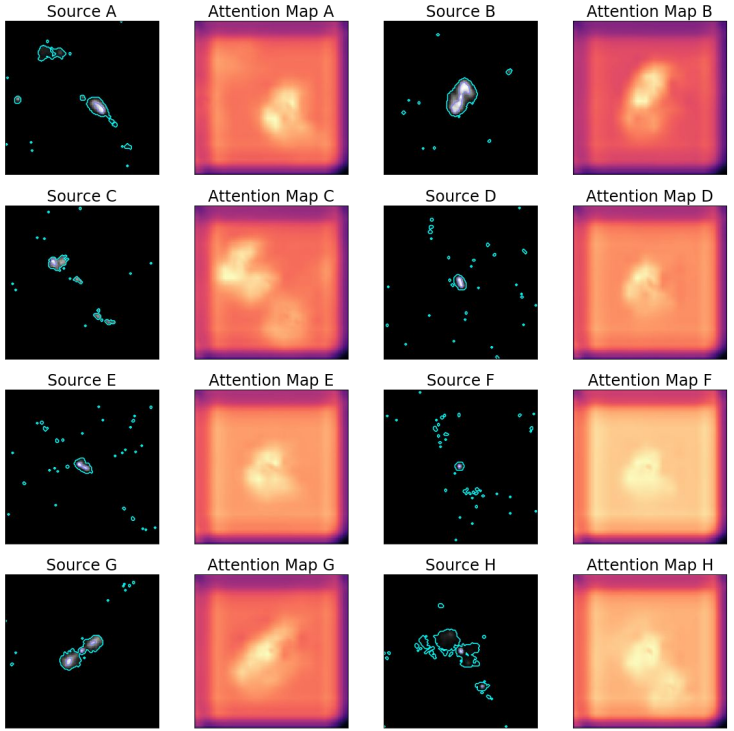
Make it as user friendly as possible.

Remove as many barriers to use as possible.

Continue to test it against the ‘norm’.

Continue training with incoming data if appropriate (or fully retrain if enough new data becomes available).

https://github.com/mb010/AstroAttention



Conclusion

Aim to solve the problem, not use a fancy technology.

(have fun along the way).

Discussion

Checklist

1. Frame the problem and look at the big picture
2. Get the data
3. Explore the data to get insights
4. Prepare the data to better expose the underlying data patterns to machine learning algorithms
5. Explore many different models and short-list the best ones.
6. Fine-tune your models and combine them to a great solution.
7. Present your solution
8. Launch, monitor and maintain your system.

https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/