

IQRM: smart and real-time channel masking for radio transient searches

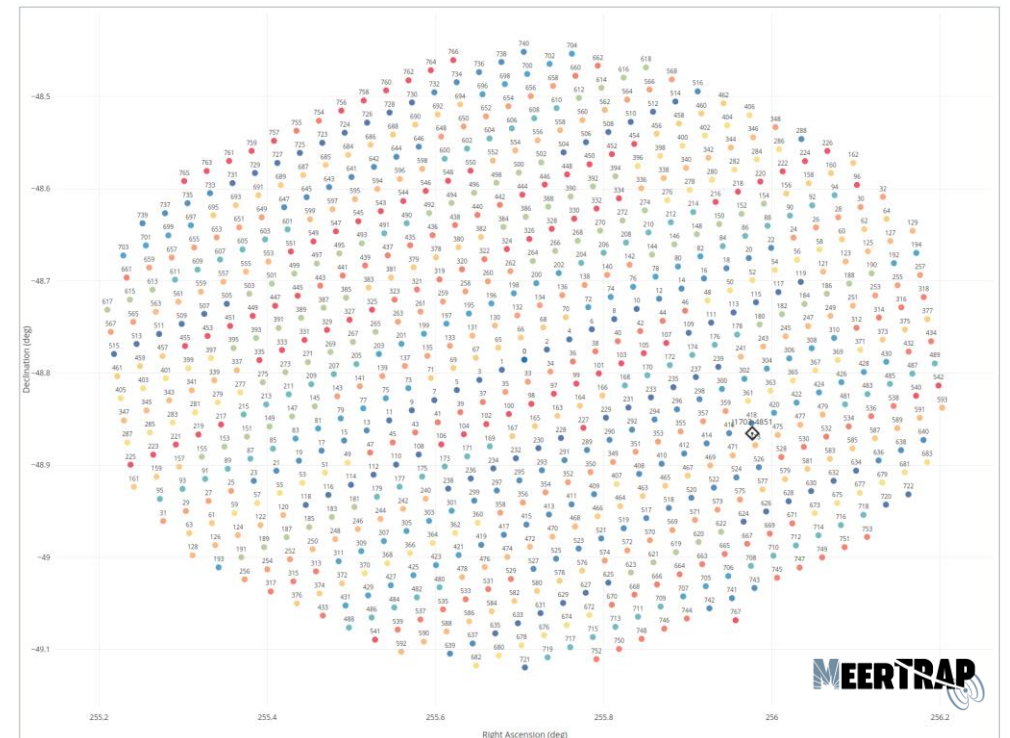
Vincent Morello, with Kaustubh Rajwade and Ben Stappers
arXiv: 2108.12434



European Research Council
Established by the European Commission

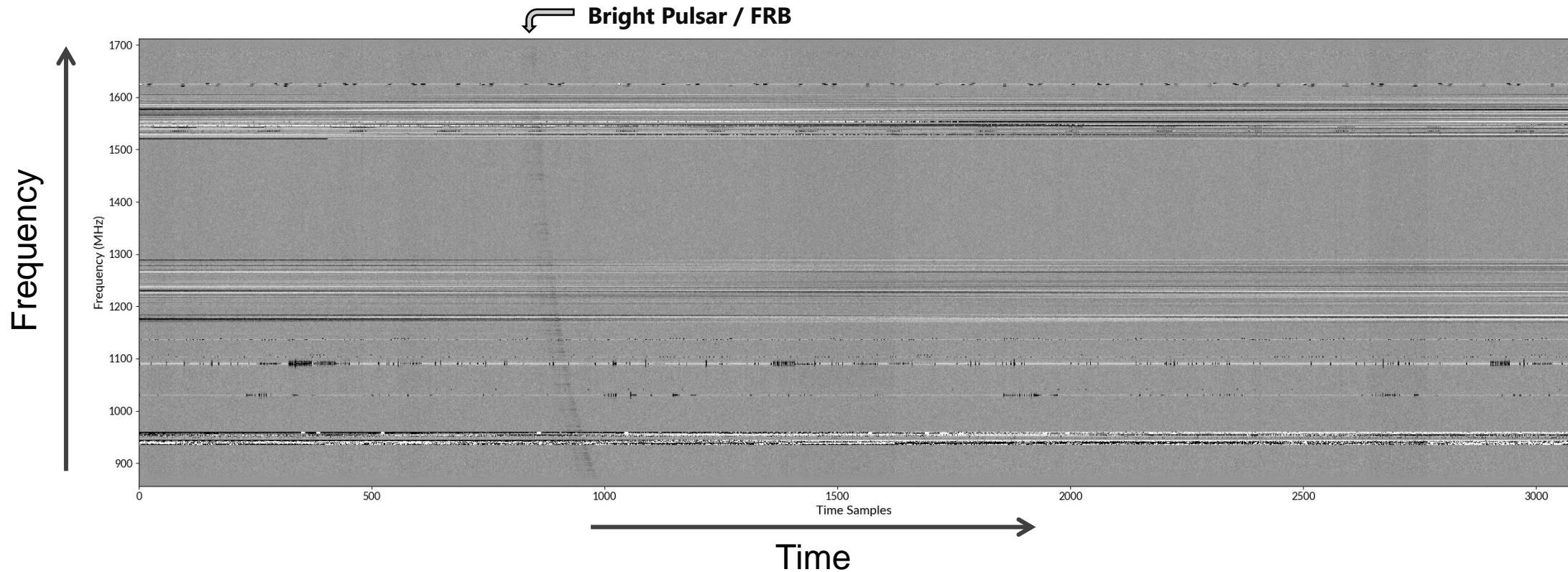
Context: Real-Time Radio Transient Searches

- Massively multi-beam systems: Data volumes so large that **real-time processing is the only option.**
- Example: MeerTRAP, commensal search backend of MeerKAT telescope. **Continuously searching 24/7:**
 - 768 Total beams tiled on sky
 - 1024 Channels, 306 μ s sampling time
 - **2.6 GB/s ingest rate**
- Dynamic RFI environment: Bad freq. channels vary as a function of time and pointing direction (cf. talk by Isaac Sihlangu)
- High dynamic range. A few unmasked bad channels throw off the search pipeline.



**Can't use a static list of bad channels to mask.
Must detect them in real-time with high accuracy**

Typical Search Mode Data (MeerKAT L-Band 856 - 1712 MHz)



High time resolution radio spectra = **Intensity vs. Frequency and Time**. Sampling interval $30 \mu\text{s} - 1,000 \mu\text{s}$.
Looking for millisecond, broadband pulses with significant dispersion: **lower freqs. arrive later**
Getting rid **automatically** of the visible RFI-affected channels is the main objective.

The IQRM Algorithm

IQRM: general idea

Adaptive, time-variable channel mask

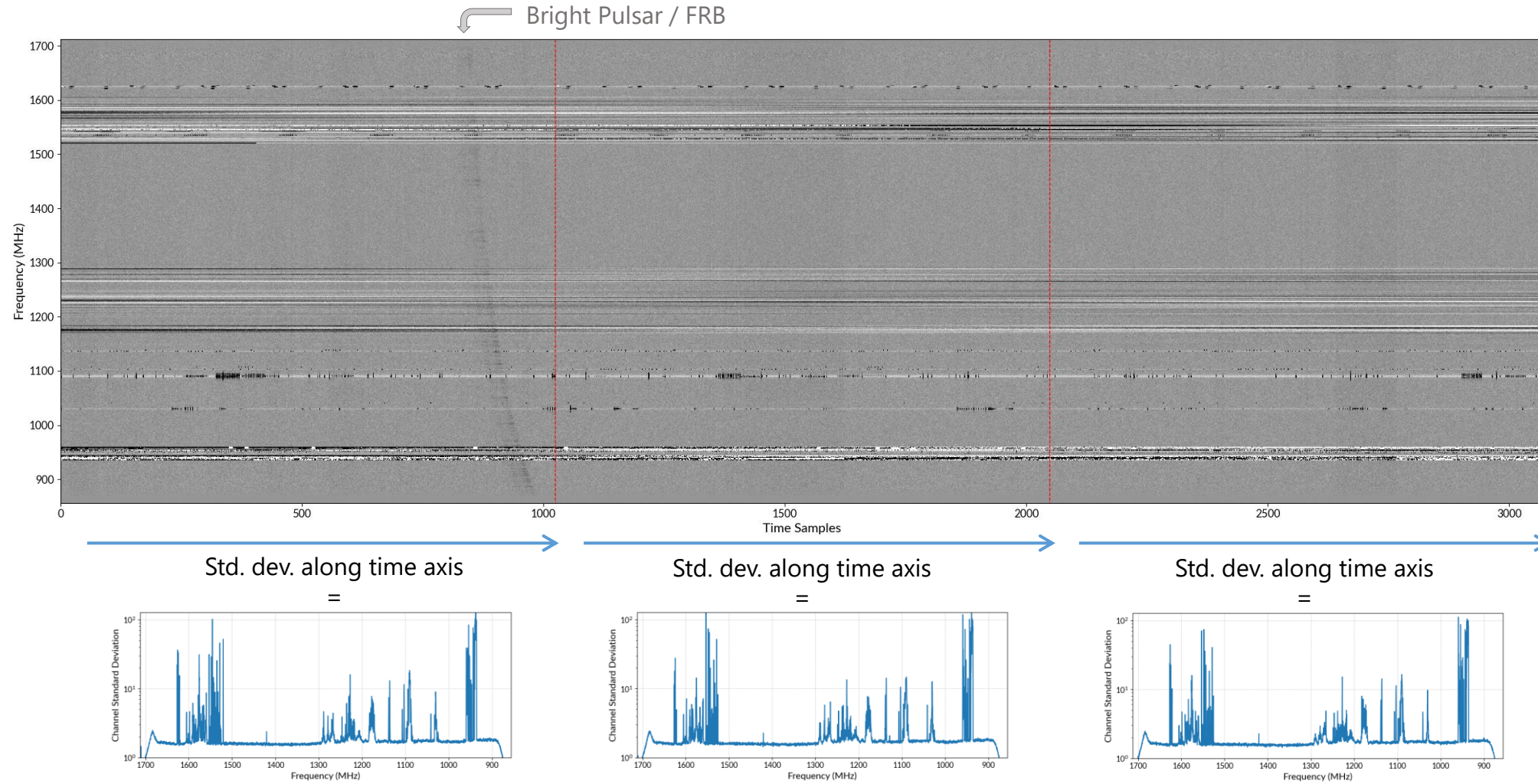
Calculated on short (~1 seconds) consecutive blocks of data

Ingest data block → detect & mask bad channels → send for processing → repeat

Step 1: Calculate a measure of RFI contamination in each channel

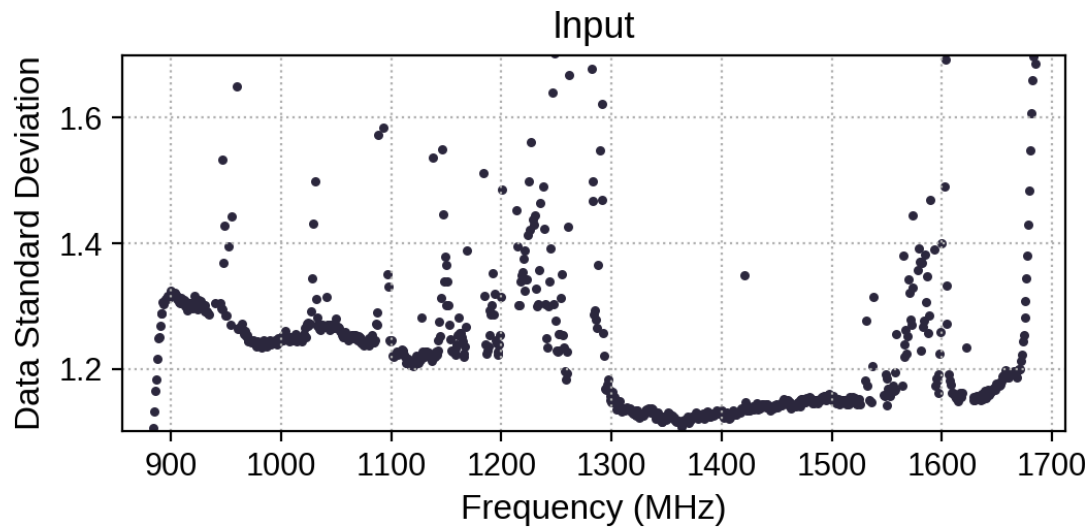
Step 2: Mask channels where said measure is "too high"

Example: Short-term Spectral Standard Deviation



The Statistics Problem to Solve

Task: finding high outliers in sequential data with a background trend

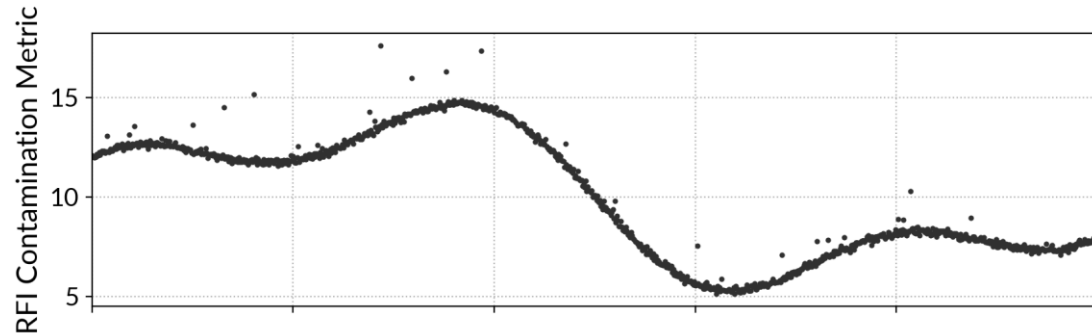


Trend caused by e.g. diffuse Galactic radio emission, Broadband RFI, instrumental effects, etc.

Algorithmic Catch-22 situation:

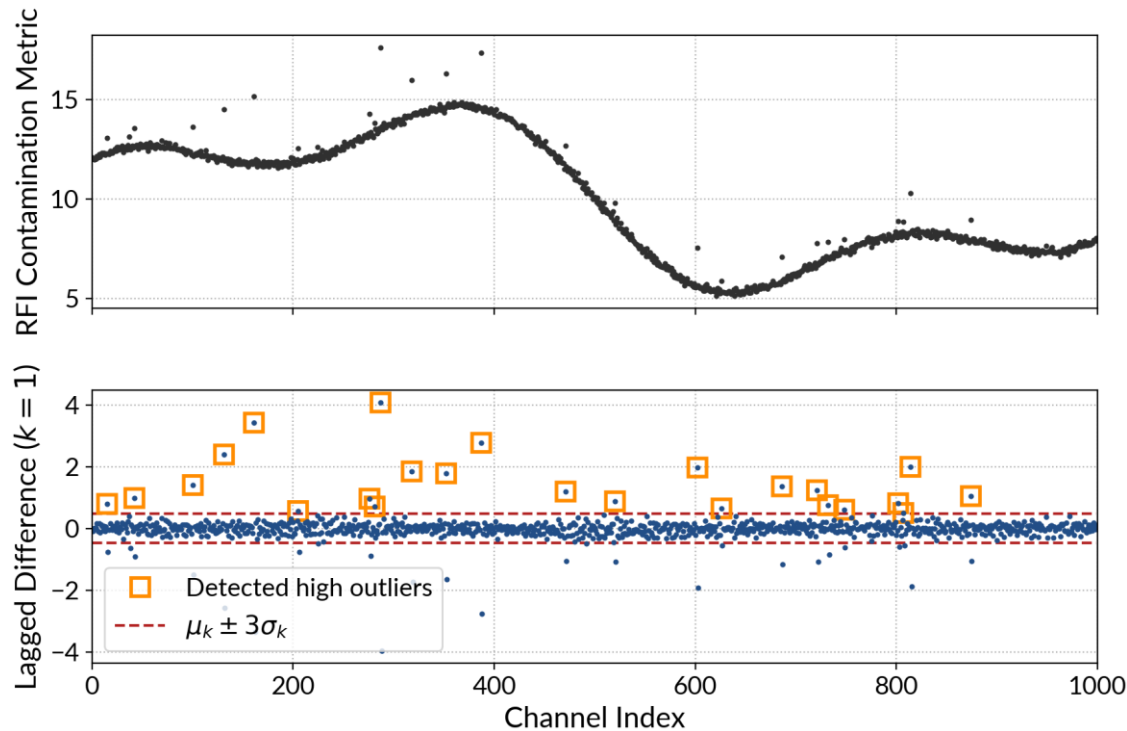
- Must fit and subtract trend to identify outliers
- Must remove outliers to fit trend

IQRM: Lagged Difference Step



Main Idea #1:
Eliminate the trend by taking lagged differences.

IQRM: Lagged Difference Step

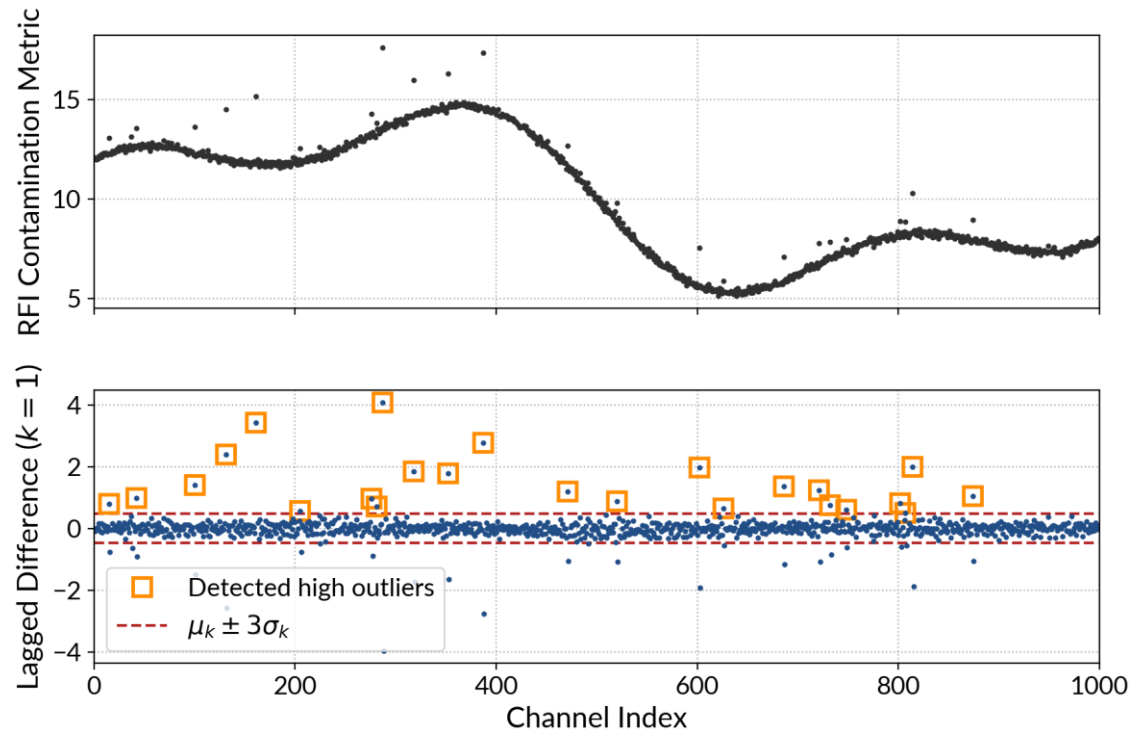


Main Idea #1:
Eliminate the trend by taking lagged differences.

Data minus copy of itself shifted by \mathbf{k} elements.
 \mathbf{k} is the lag.

$$\Delta_i^k = x_i - x_{i-k}$$

IQRM: Preliminary Flagging Step



Identifying outliers now much easier. **Tukey's rule:**

$$\Delta_i^k - m > 3\sigma \Rightarrow \text{high outlier}$$

m : median
 σ : std. dev.

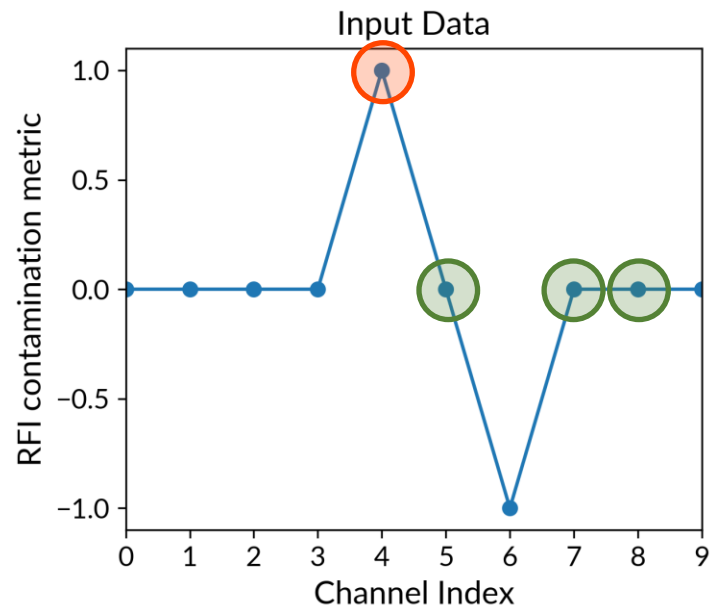
Standard deviation **estimated robustly** from Inter-Quartile Range (IQR): avoid influence of outliers.

Repeat process for lags k in $[-r, +r]$
 r = radius parameter of the algorithm.

IQRM: “Voting” Step

“Equivalence problem”: after taking lagged difference, can’t distinguish between:
Legitimate High outlier and **Neighbour of Low outlier**

Example below: if we **only** looked at lagged differences up to radius $r = 2$,
points #5, #7 and #8 would look like high outliers: **need to perform extra checks**

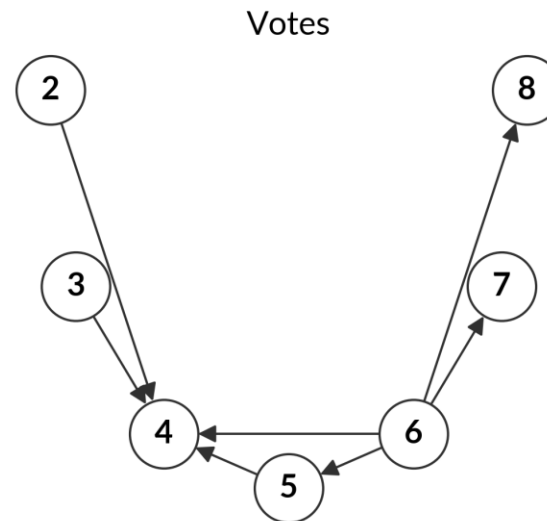
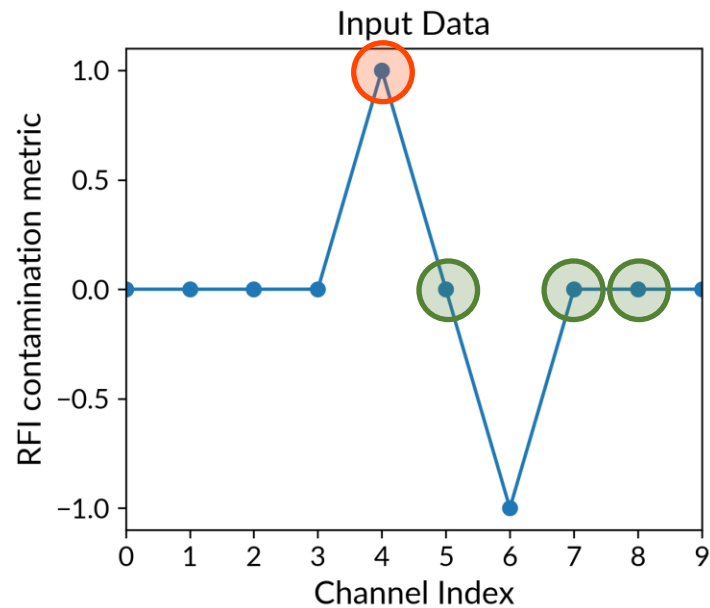


IQRM: “Voting” Step

Main Idea #2: Voting System to break equivalence problem

Vote $i \rightarrow j$ means: “From the point of view of \mathbf{x}_i , \mathbf{x}_j is a high outlier”

Votes can be represented as a directed graph.

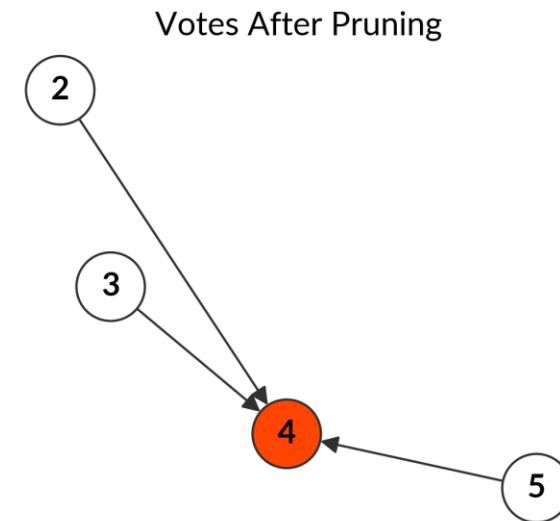
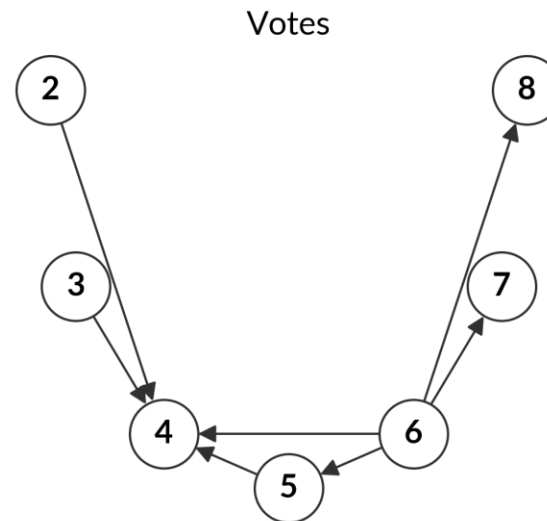
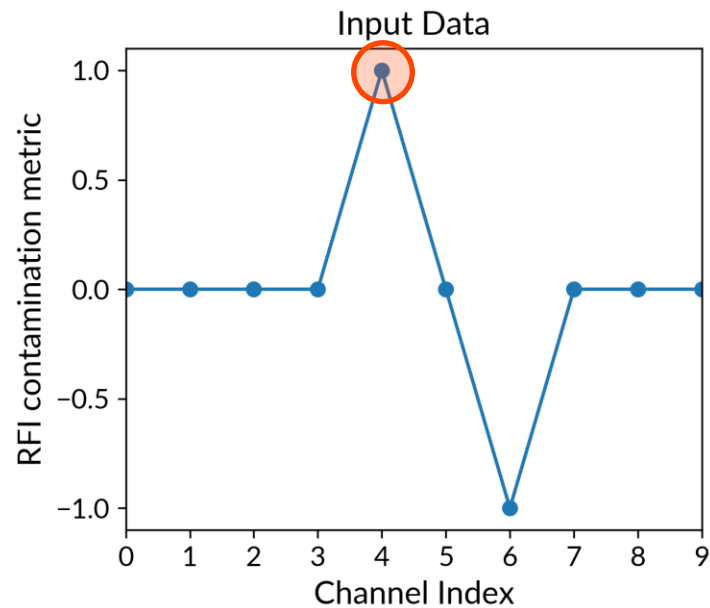


IQRM: “Voting” Step

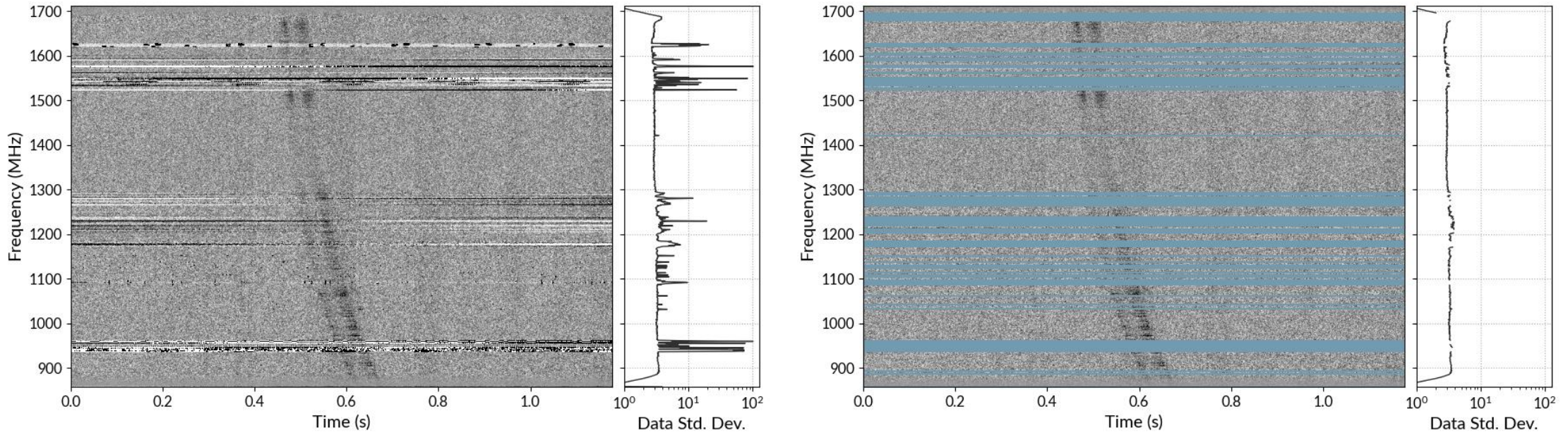
Vote $i \rightarrow j$ valid only if: i has cast strictly less votes than j has received

(Here: All votes from #6 are invalid and discarded)

Any point with ≥ 1 vote against it is flagged as a high outlier (Here: only #4)



Example: Single pulse recorded at MeerKAT L-Band



Note: Same color scale left and right, clipped due to high dynamic range. Pulsar: PSR J1226–3223.

Using spectral standard deviation as the RFI contamination metric,
all visually identifiable “bad” channels identified without any prior information.

Differences & Similarities with Spectral Kurtosis

Acceptable range of values for spectral kurtosis can be found from first principles.
(Nita et al. 2007, 2010)

... But what if we want to use other spectral statistics ?

MeerKAT: spectral standard deviation works better

Jodrell Bank: spectral autocorrelation works better

Could try other, more sophisticated functions of the data !

However:

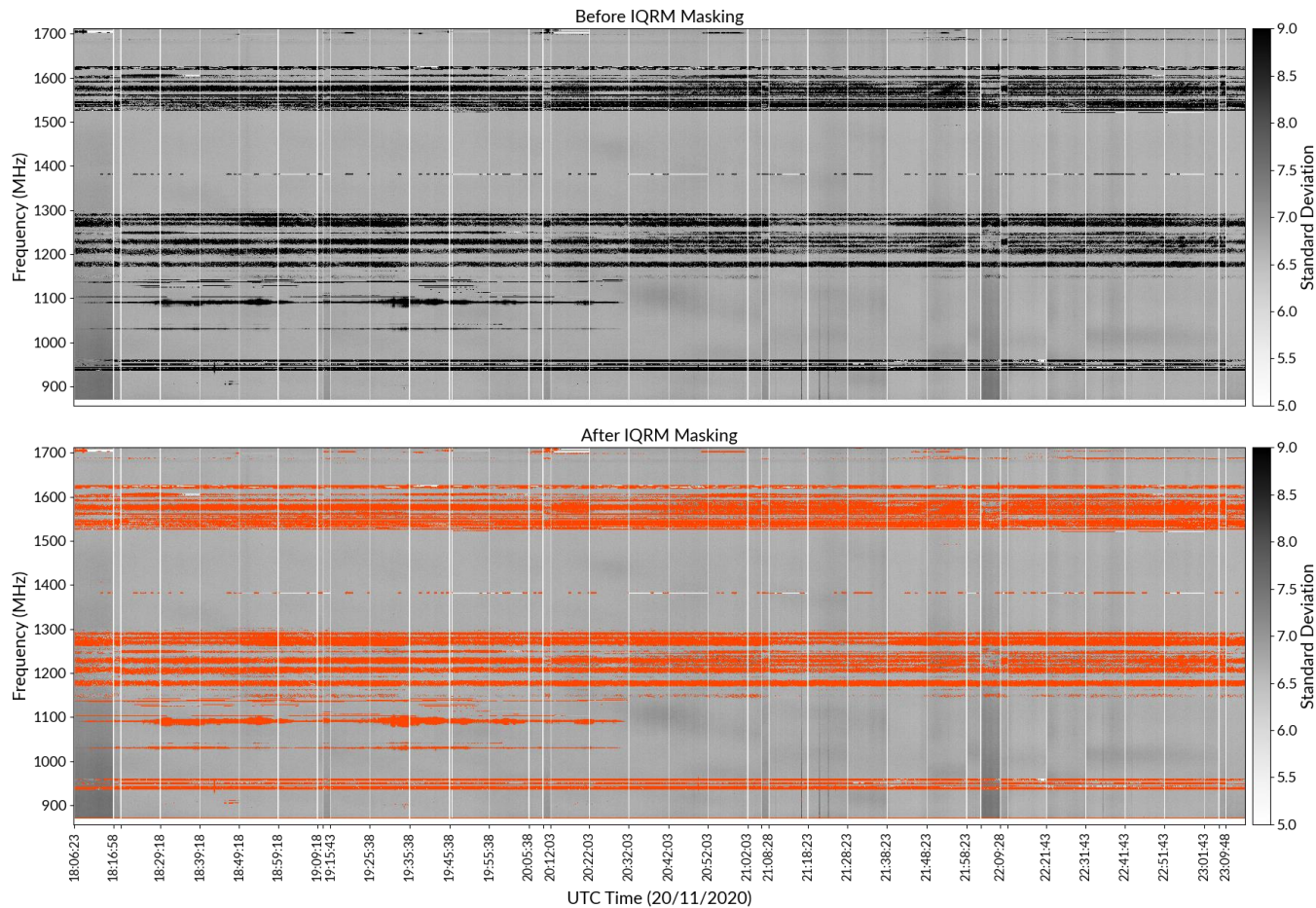
No more theoretical guarantees.

Above what threshold is the contamination level "too high" ?

Must be inferred from the data: **Outlier detection problem. That's where IQRM comes in.**

Tests on Real Data

MeerTRAP L-Band (856 - 1712 MHz)



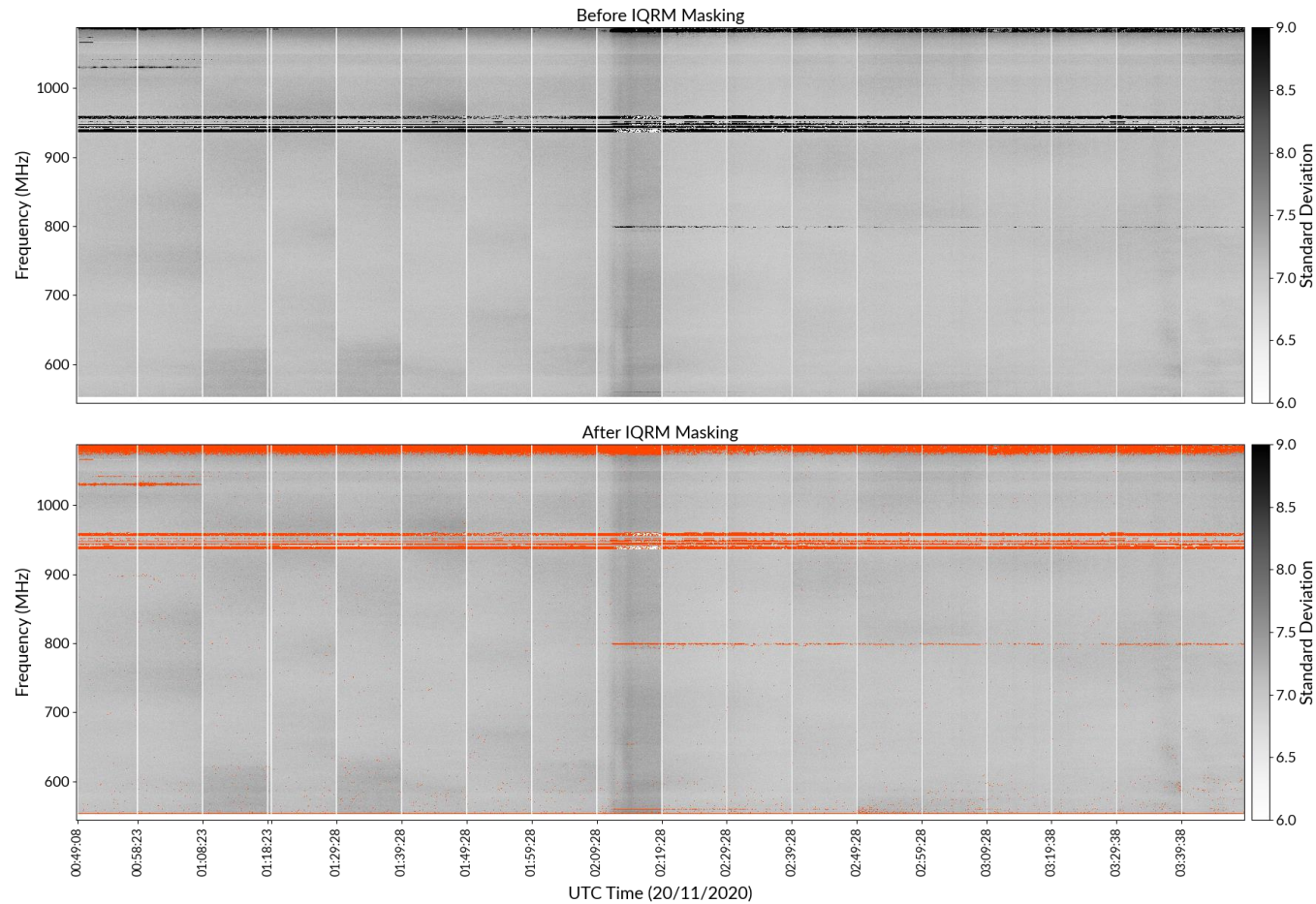
Spectral standard deviation over consecutive 5-second blocks for the central coherent beam. Note: color scale clipped due to high dynamic range

8-bit data, 306 μ s sampling time, 1024 channels.

6 hours worth of data with many target changes (correspond to vertical white lines).

Orange: channels flagged by IQRM

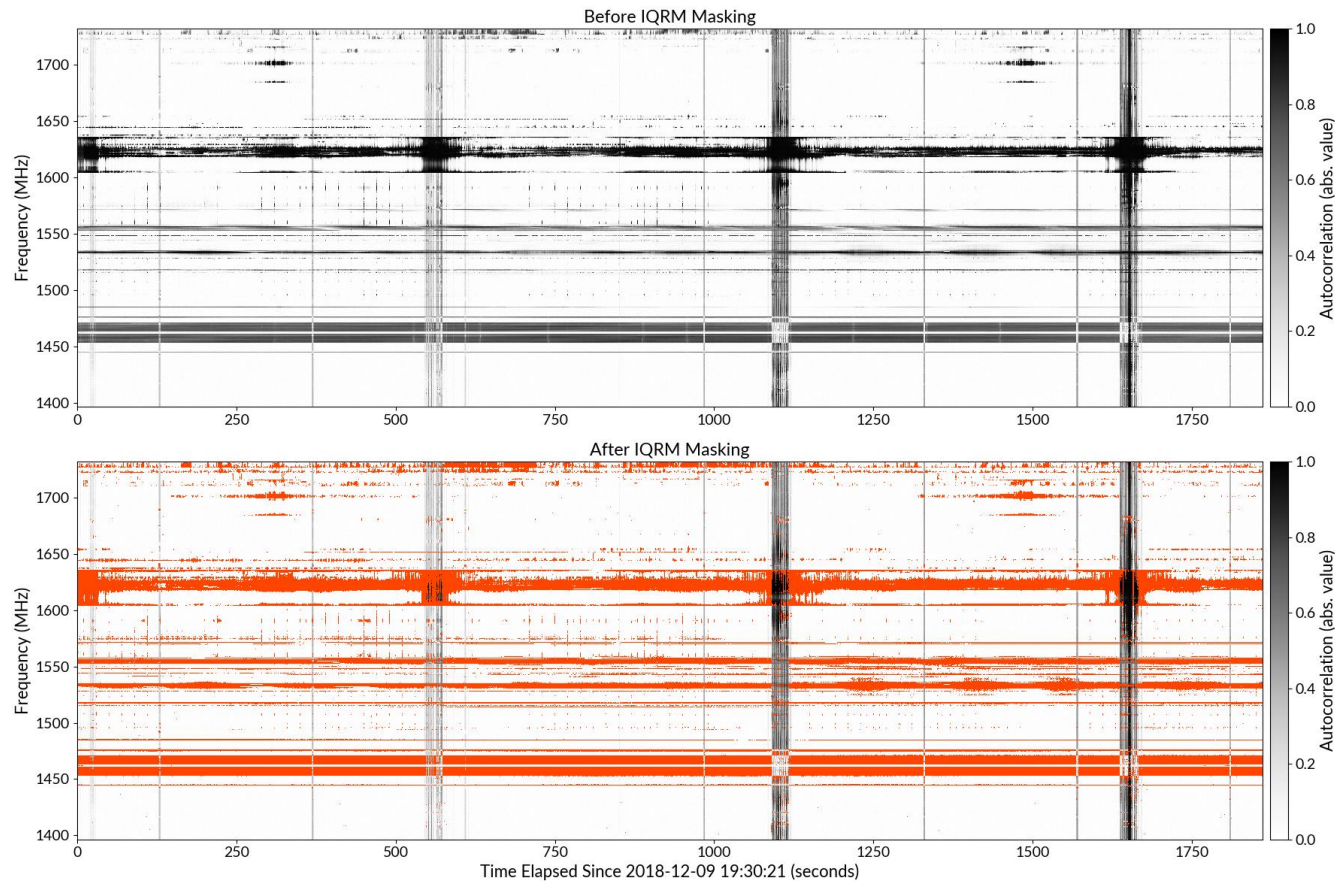
MeerTRAP UHF (544 - 1088 MHz)



Same thing at UHF band
4 hours worth of data on multiple sources
Note: color scale clipped due to high dynamic range

Orange: channels flagged by IQRM

Jodrell Bank L-Band (1396 - 1732 MHz)



Spectral autocorrelation function with 1-sample lag (ACF1)

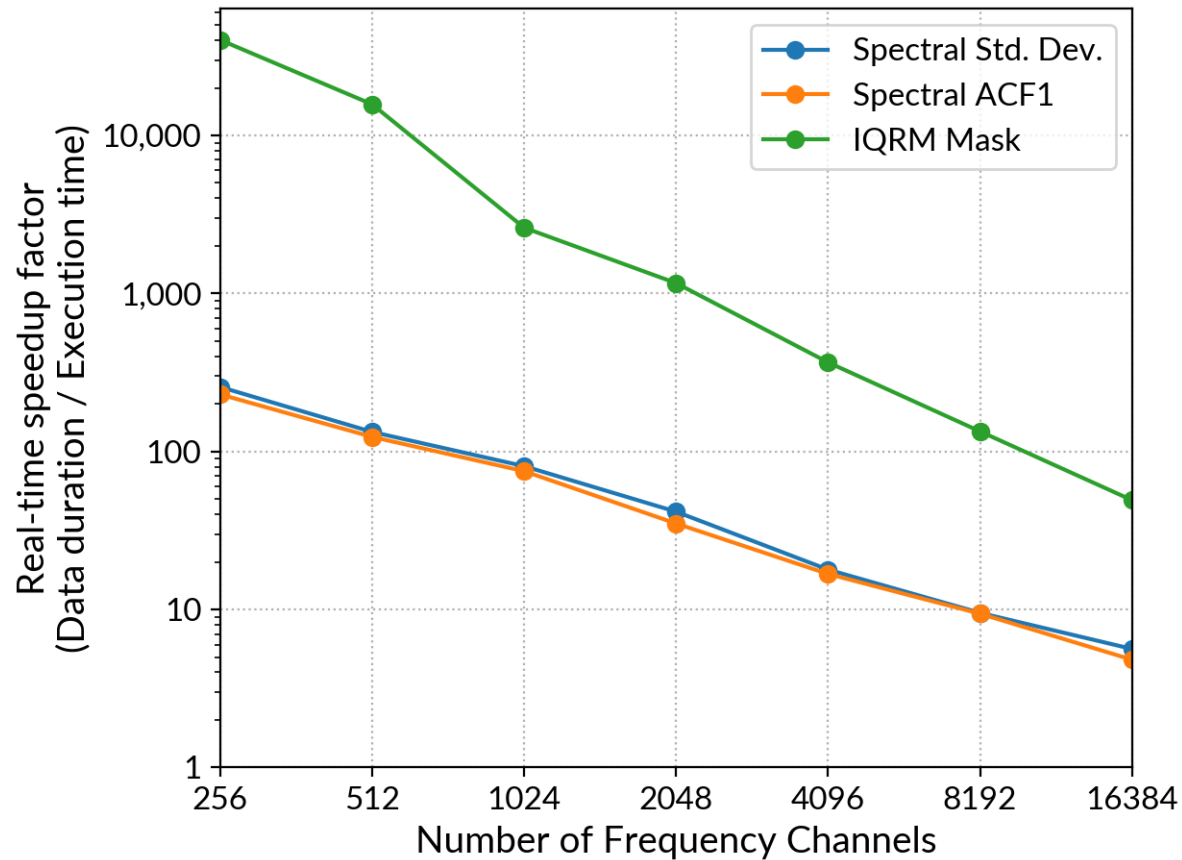
on consecutive 0.8-second blocks.

30-minute observation of PSR B0611+22

8-bit data, 256 μ s sampling time, 672 channels.

Orange: channels flagged by IQRM
Predictable edge case: fails when the majority of channels are contaminated by RFI.

Speed Benchmarks



Benchmarks on fake data:

- Gaussian noise
- 256 μ s sampling time
- **Using just one CPU core**

Implementation:

https://gitlab.com/kmrajwade/iqrm_apollo

The spectral statistic calculation is the the most expensive task.

The calculation of the channel mask by IQRM has negligible cost (and scales only with N_{chan}).

Overall much faster than real time on 1 CPU core even for 4K+ channels.

Need more in-depth testing

The “RFI ground truth” is almost never available:

The algorithm looks effective from the above plots, but it only sees a **proxy** for RFI contamination.

Rigorous evaluation of RFI mitigation can only be done w.r.t. science goal

Here: Finding astrophysical radio transients

We want: More genuine transients **and** Less spurious candidates.

Tests on Single Pulse Searches

Experimental Setup - 1

Data:

- Four known pulsar observations from Jodrell Bank 76-m telescope
- Pulsars chosen so that we detect a mix of faint and bright single pulses
- 30-minute integration time, 672 freq. channels, 256 μ s sampling time.

Pipeline:

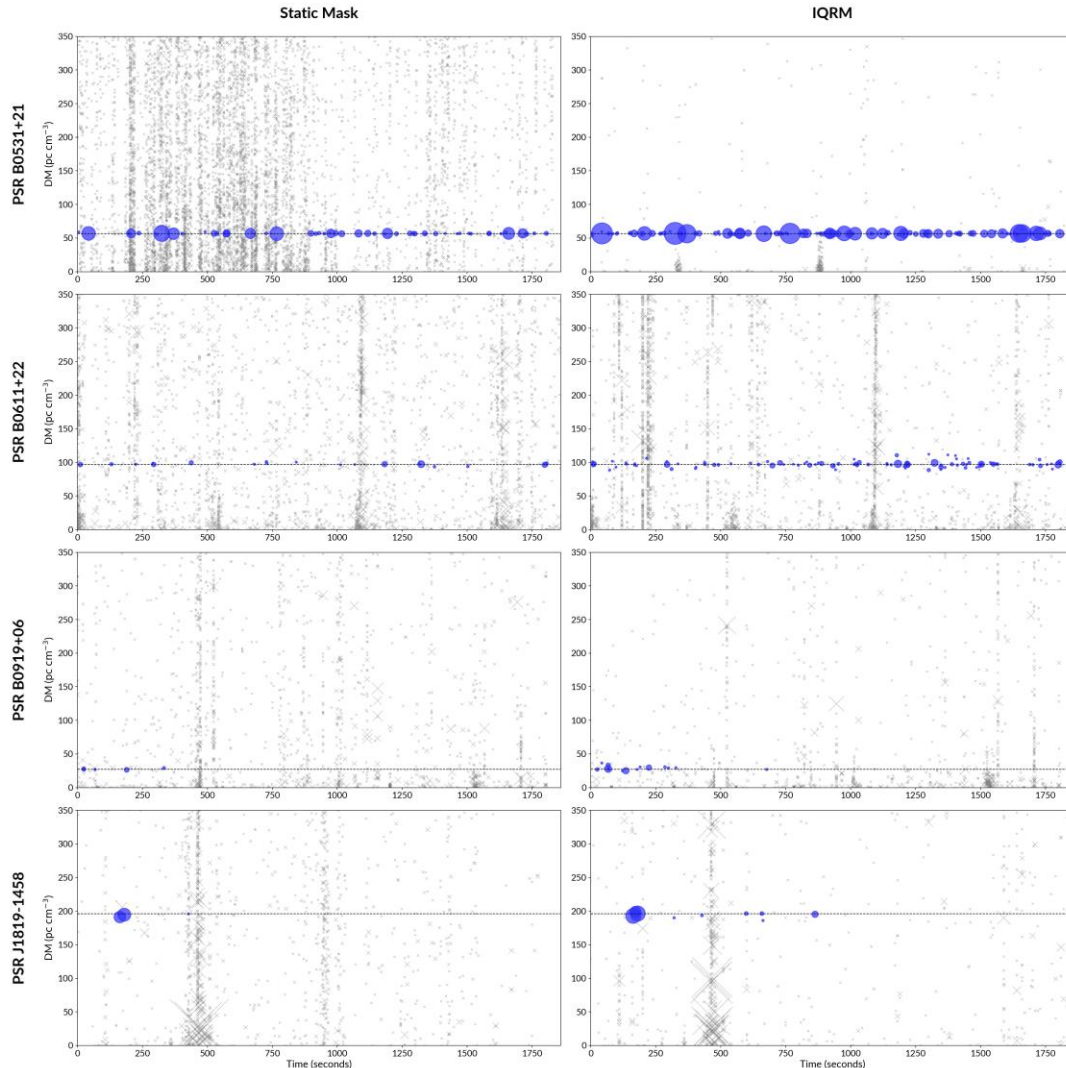
- Search code: **Heimdall** with default RFI mitigation params
<https://sourceforge.net/projects/heimdall-astro/>
- S/N threshold: 6
- Candidate classifier: **FETCH** (Deep Neural Network, see Aggarwal et al. 2020)
- Candidates reported as positive by FETCH are visually inspected for confirmation. Rest is marked as spurious.

Experimental Setup - 2

We ran the **exact same pipeline** on:

1. Original observation files, with a static list of known bad channels masked.
Static Mask is the “established standard” used in pulsar timing observations at Jodrell Bank.
2. Copies of files prealably cleaned by IQRM, based on spectral ACF1

Single Pulse Search Output



Blue Circles: Confirmed pulse detections

Grey Crosses: RFI / spurious candidates

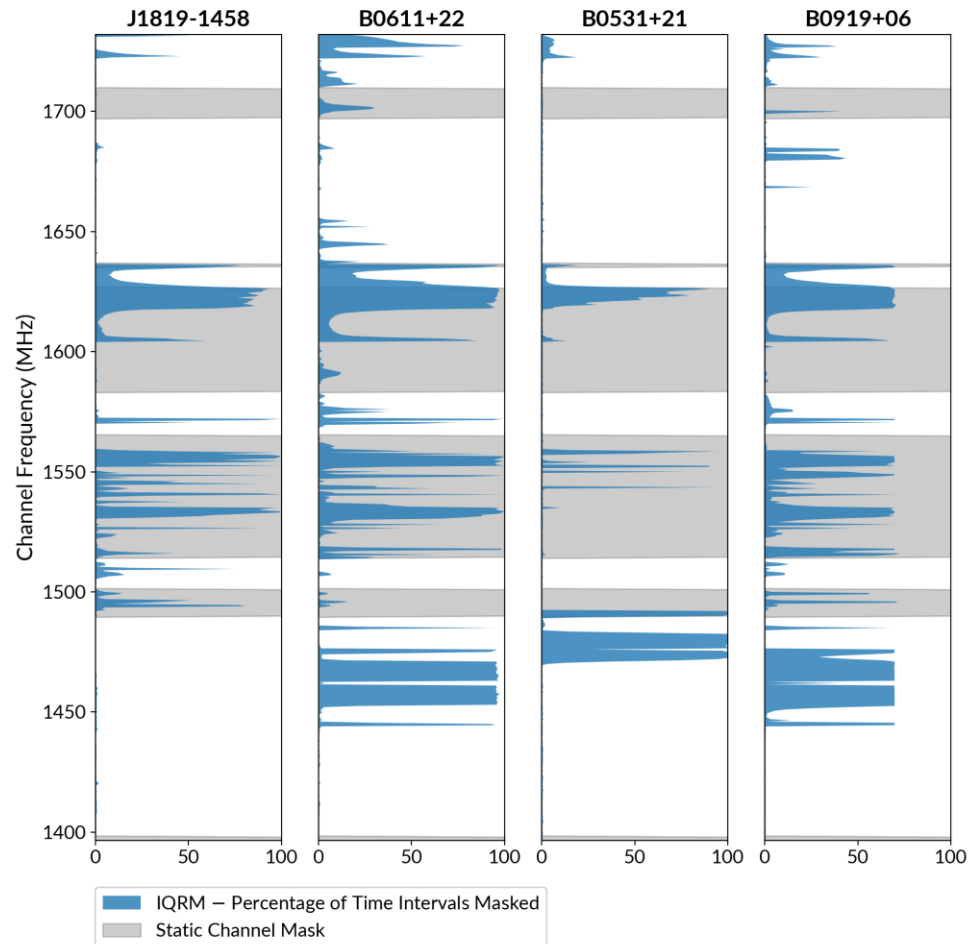
Size proportional to signal-to-noise (S/N) ratio of detected pulse.

Overall:

~3x More genuine pulses detected

~3x Less spurious candidates

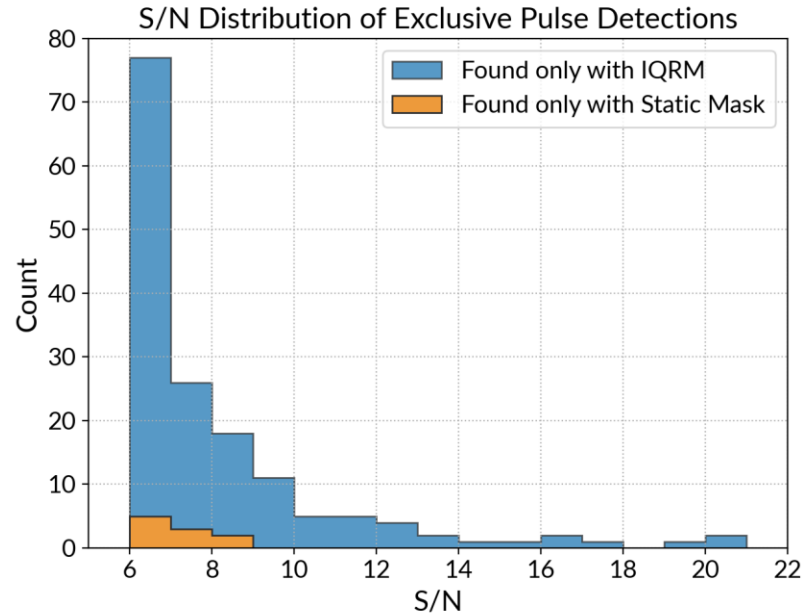
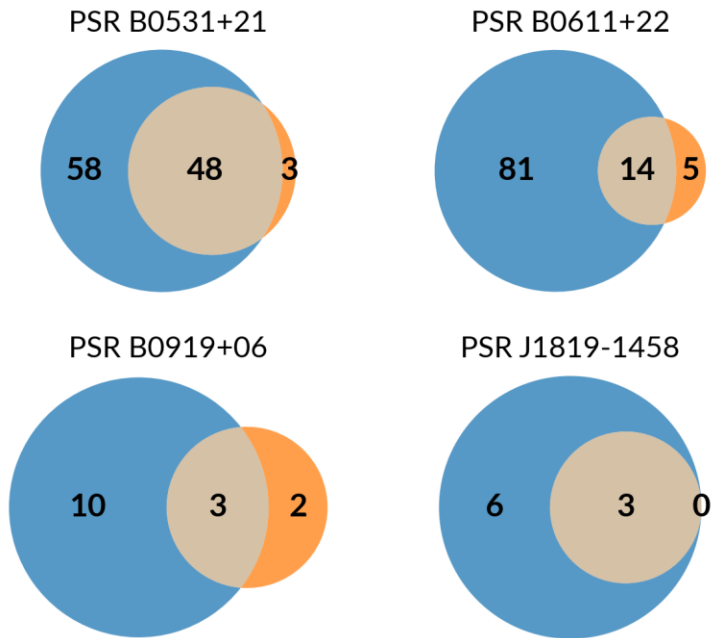
Difference between static channel mask and IQRM



RFI environment looks even more dynamic here than on the previously shown "RFI heatmap".

Contaminated channels change as a function of time and pointing direction.

Comparing Detections of Real Pulses



- Very few pulses are missed with IQRM, and they are all faint (Orange)
- In contrast: lots of pulses missed with Static Mask, including “obvious” bright ones (Blue)

Almost perfect result. Note: a few of the pulses found only with Static Mask were “RFI boosted”.

Conclusion and Future Work

Summary

- IQRM is a fast and highly effective channel flagging algorithm
 - Successfully used in Real-Time processing (2+ years continuously running on MeerTRAP)
 - Non-parametric / should generalize well to other observatories
 - More astrophysical detections & Less spurious candidates compared to static mask
- Not a “silver bullet”: Only masks channels. Should be used in conjunction with other techniques.
- Other mitigation algorithms are better suited to removing **fainter** forms of RFI. They benefit greatly from running IQRM **first** (this is what we do in the MeerTRAP pipeline)

**Paper: “IQRM: real-time adaptive RFI masking for radio transient and pulsar searches”
(accepted)**

<https://arxiv.org/abs/2108.12434>

Ideas of Further Work

- Test with more spectral statistics / RFI contamination metrics
- Compare against more advanced detection methods
- Extend to perform multi-dimensional outlier detection
e.g. could use: spectral std.dev. + autocorrelation + Kurtosis
- Retain and use information from recently processed blocks:
detect cases where all channels are bad and the block must be discarded
- Could apply similar algorithm along the time axis