RawHash

Implementing U-Test for Segmentation

Markus Lacher

Can Firtina

SAFARI

P&S genome sequencing on mobile devices

ETH zürich

Outline

Background on RawHash

U-test implementation

Results

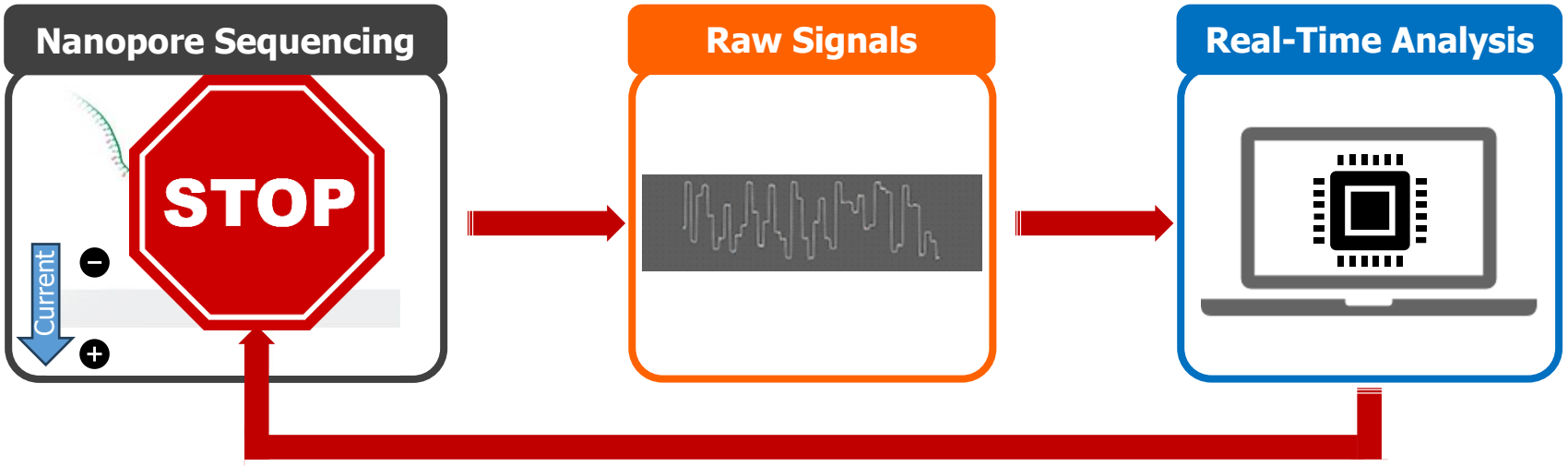
Future steps

Nanopore Sequencing

Nanopore Sequencing: a widely used sequencing technology

- Can sequence large fragments of nucleic acid molecules (up to >2Mbp)
- Offers high throughput
- Cost-effective
- Enables **real-time genome analysis**

Real-Time Analysis with Nanopore Sequencing



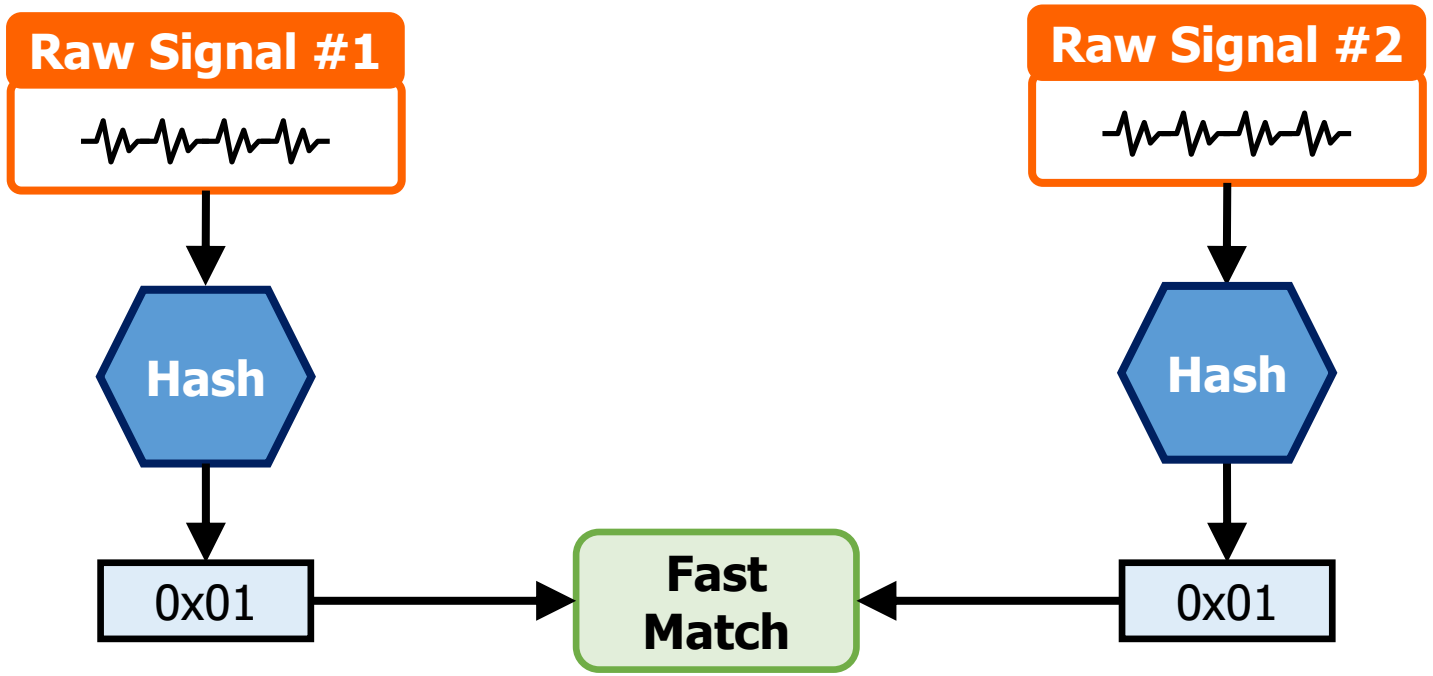
Raw Signals: Ionic current measurements generated at a certain **throughput**

Real-Time Analysis: Analyzing all raw signals by **matching the throughput**

Real-Time Decisions: Stopping sequencing **early** based on real-time analysis

RawHash – Key Idea

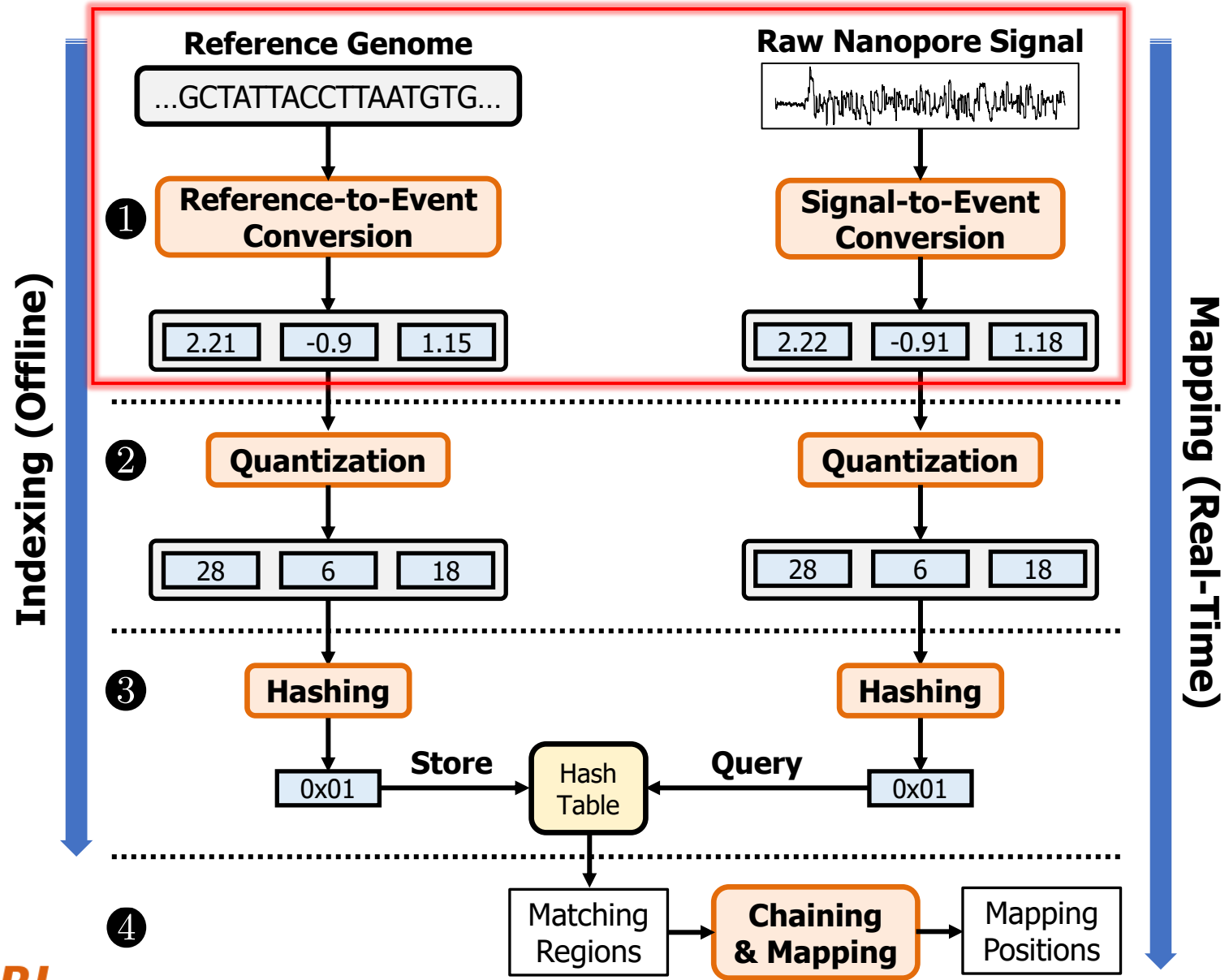
Key Observation: **Identical** nucleotides generate **similar** raw signals



Challenge #1: Generating the **same** hash value for **similar enough** signals

Challenge #2: **Accurately** finding similar regions **as few as possible**

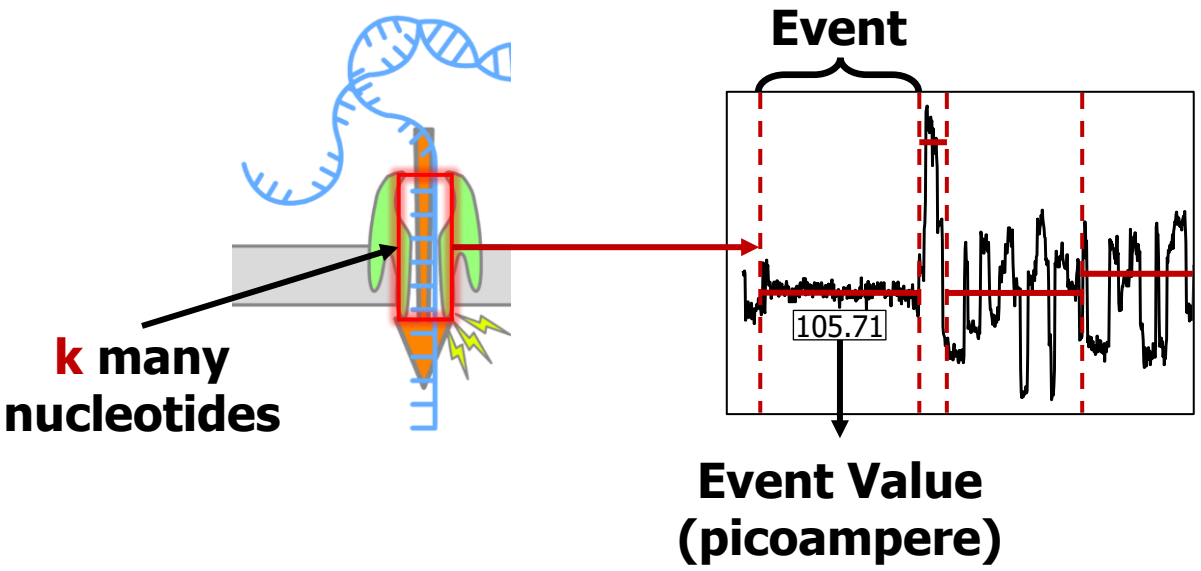
RawHash Overview



Events in Raw Nanopore Signals

- **Event:** A **segment** of the raw signal
  - Corresponds to a **particular k**-mer

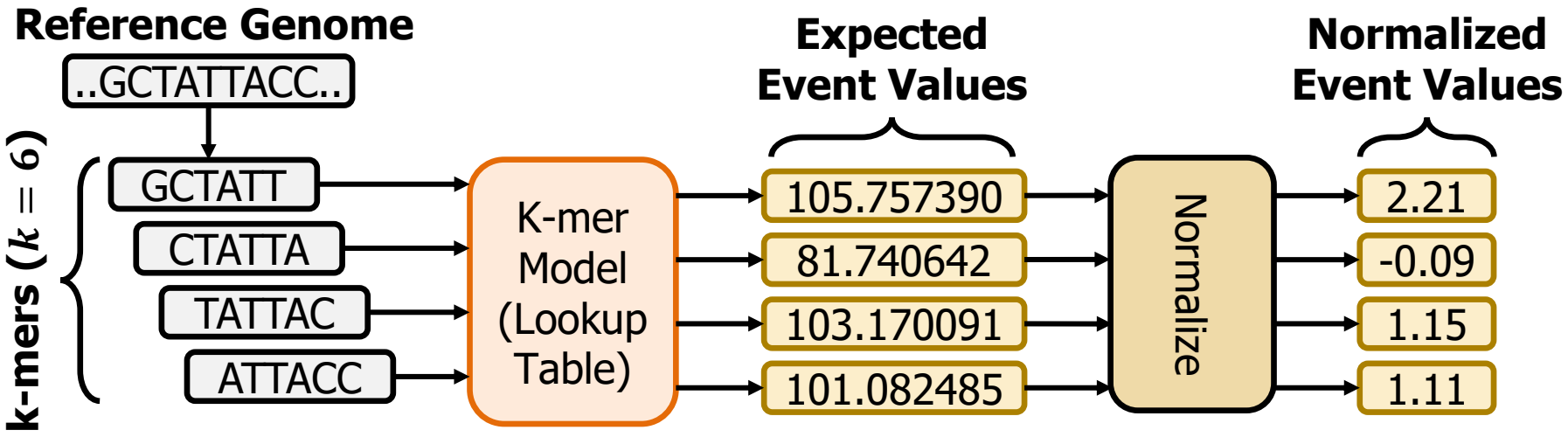
- **Event detection** finds these segments to identify **k**-mers
  - Start and end positions are marked by abrupt signal changes
  - Statistical methods identify these abrupt changes
  - **Event value:** average of signals **within an event**



Reference-to-Event Conversion

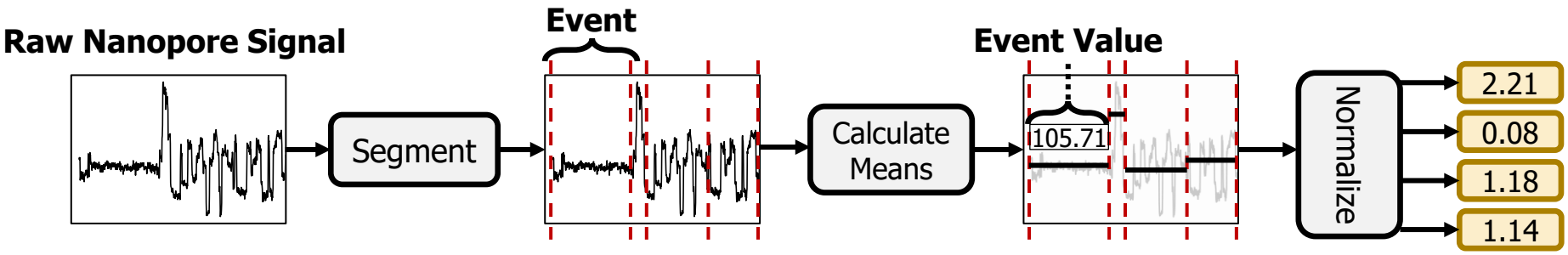
- **K-mer model:** Provides **expected** event values **for each k-mer**
  - Preconstructed based on nanopore sequencer characteristics

- Use the **k-mer model** to convert **all k-mers** of a reference genome to their **expected** event values



Signal-to-Event Conversion

- **Event detection:** Identifies signal regions corresponding to specific k-mers
  - Uses statistical test (**segmentation**) to spot abrupt signal changes



- Consecutive events ➔ consecutive k-mers

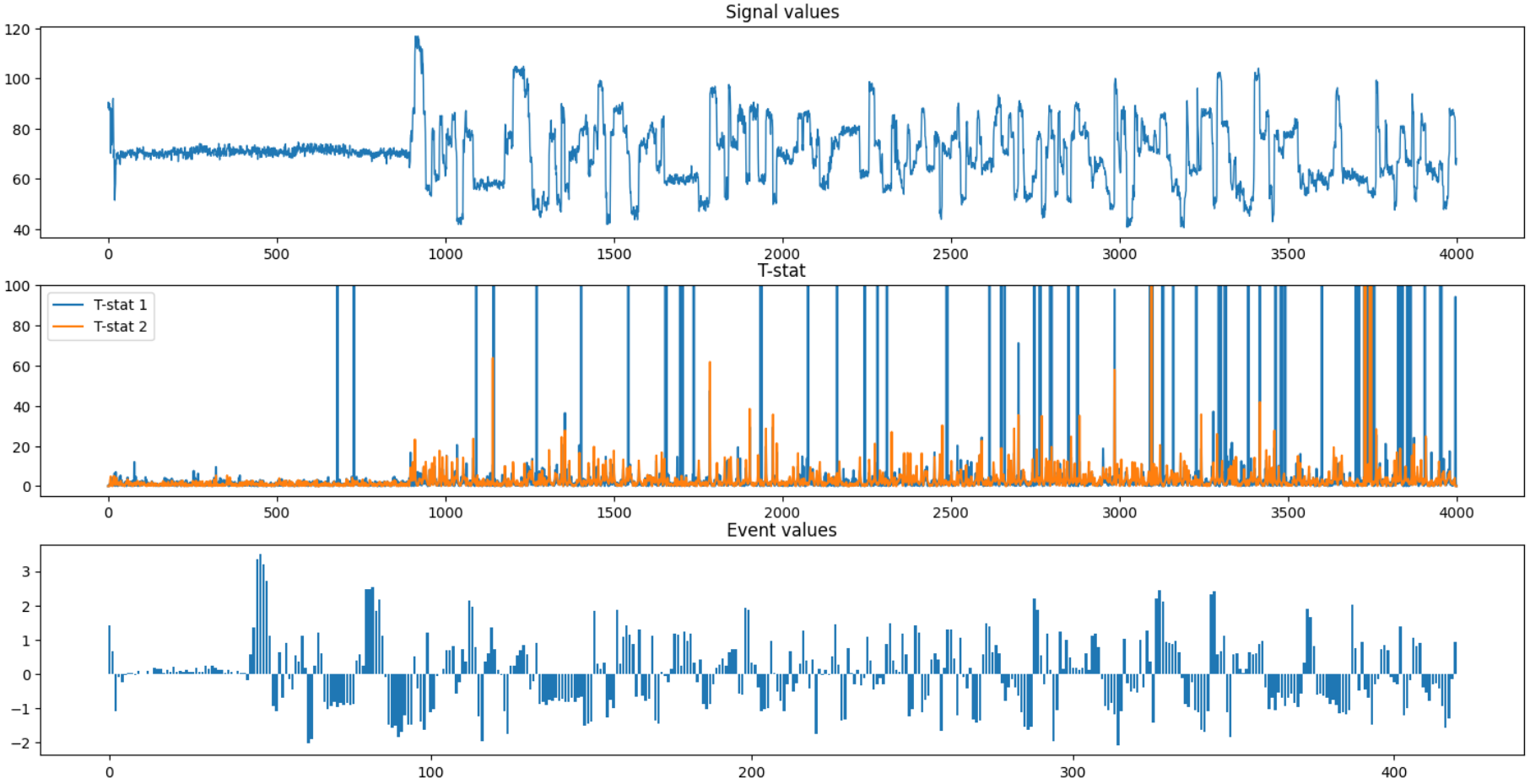
T-test

- A t-test is a statistical method used to determine if there is a **significant difference between the means** of two groups

$$T = \frac{\overline{x_1} + \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- T-test value is generated by **two adjacent sliding windows Peaks in the T-test** values are identified as beginning of an event
- Results in two configurable parameters, **threshold** and **window length**

Sample Data



First 4‘000 values of a nanopore signal. (From top to bottom: Signal values, T-test values, Event values)

Peak detection

- Peak detection function uses **two T-tests with different parameters.**
- Zoomed in from the previous plot. **Red dots represent peaks** found by rawhash.

Executive Summary

Problem: Is T-test the best statistical tool for detecting events?

Goal: Try **U-test** as a different statistical tool and evaluate performance.

Key observations:

1) T-Test detects differences in the mean of two distributions
2) T-Test uses two different window sizes and thresholds to detect events.
3) Events are clearly visible as peaks in the T-test.

Outline

Background on RawHash

U-test implementation

Results

Future steps

Mann-Whitney U-test

Goal: The goal of the Mann-Whitney U test is to assess whether there is a **significant difference** between **two independent groups**' distributions.

Difference to T-test: U-test **does not rely** on the assumptions that the samples are **normal distributed** and have approximately the **same variance.**

Mann-Whitney U-test

Assuming **two independent non-normal distributed** datasets → compare the ranksum of both distributions.

T_1: ranksum of datapoints n_1
T_2: ranksum of datapoints n_2

$$U_1 = n_1 * n_2 + \frac{n_1(n_1 + 1)}{2} - T_1$$

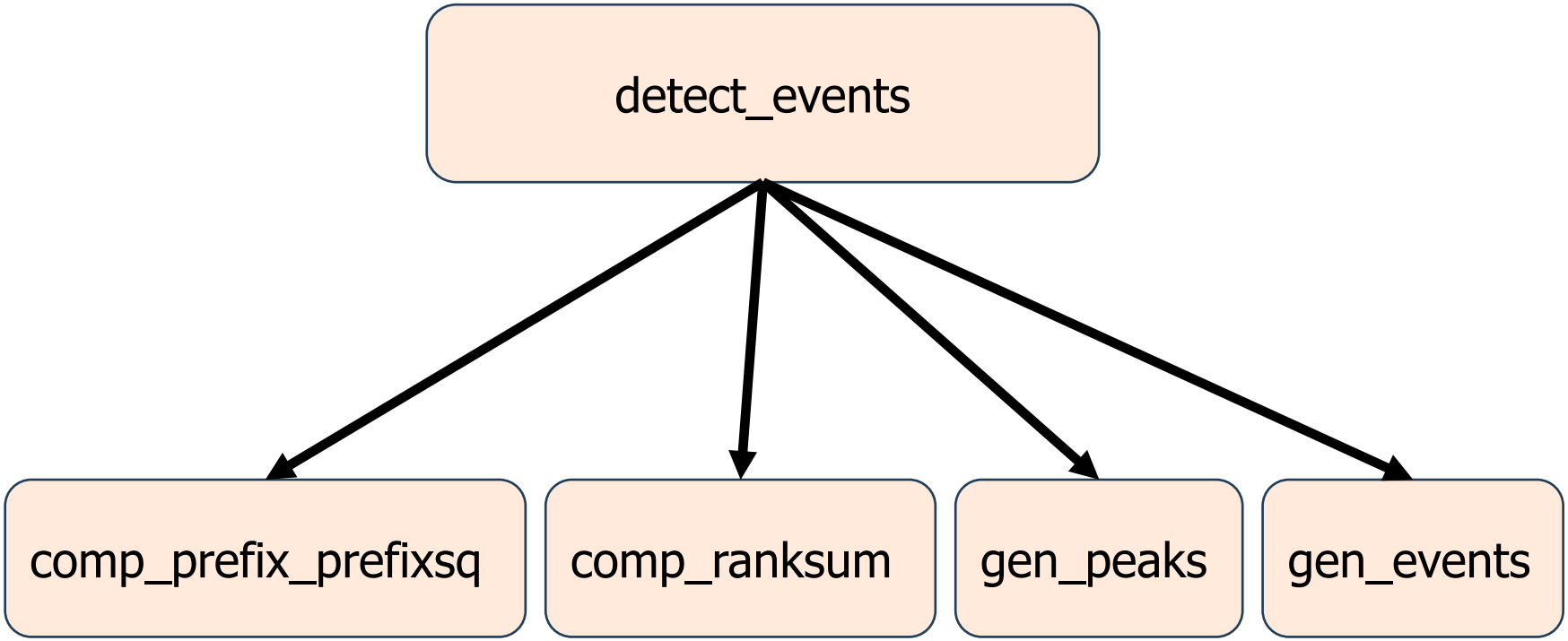
$$U_2 = n_1 * n_2 + \frac{n_2(n_2 + 1)}{2} - T_2$$

$$U = \min\{U_1, U_2\}$$

Low U-test value: **significant difference in the two distributions**

Functions in RawHash code

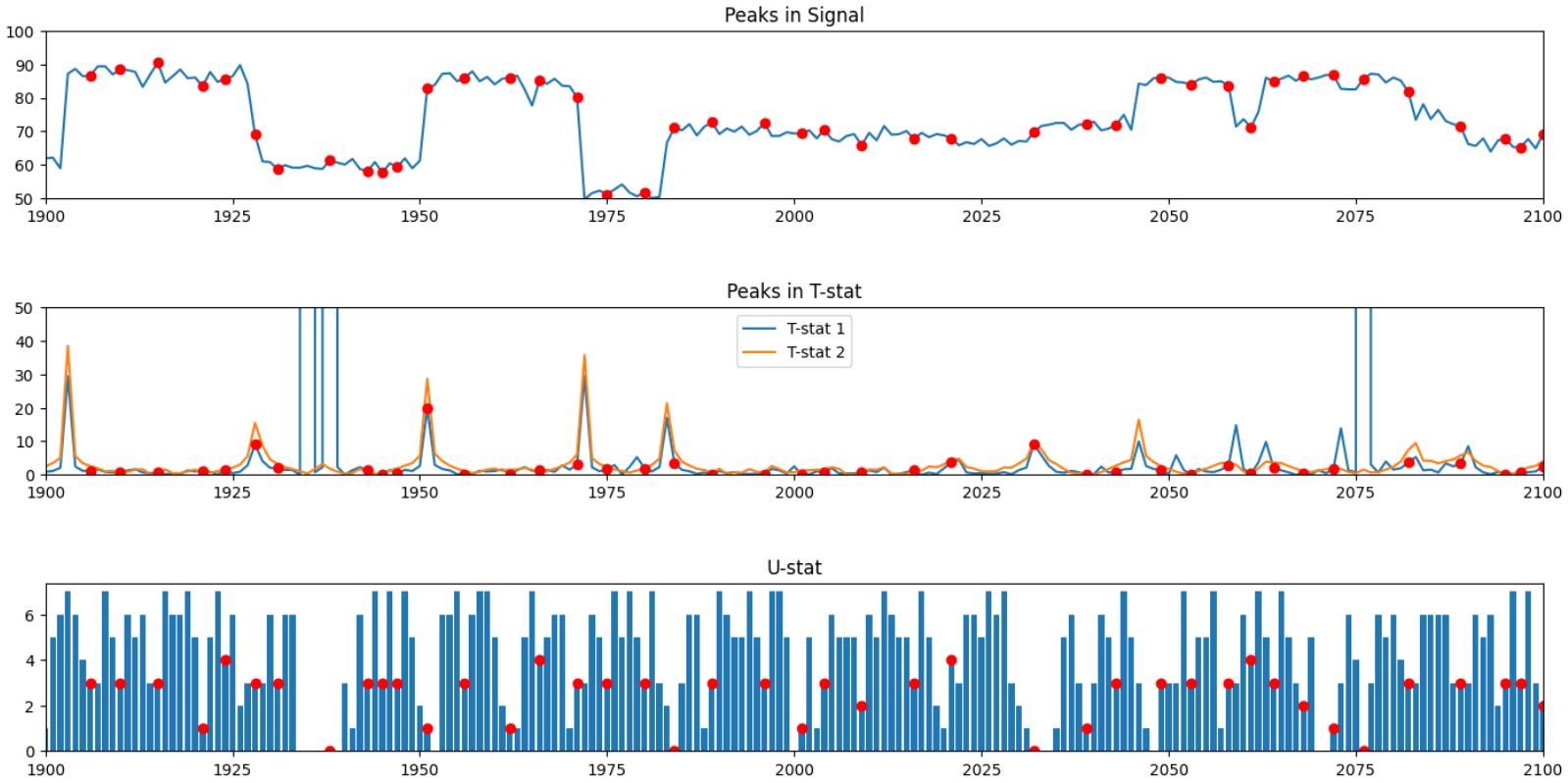
All functions are located in revent.c



U-test peak detection

- First version **only looks at U-values**
- If U-values fall below threshold a peak is detected
- If multiple values below threshold → **take center**

Red dots represent peaks found with U-test



Drawbacks

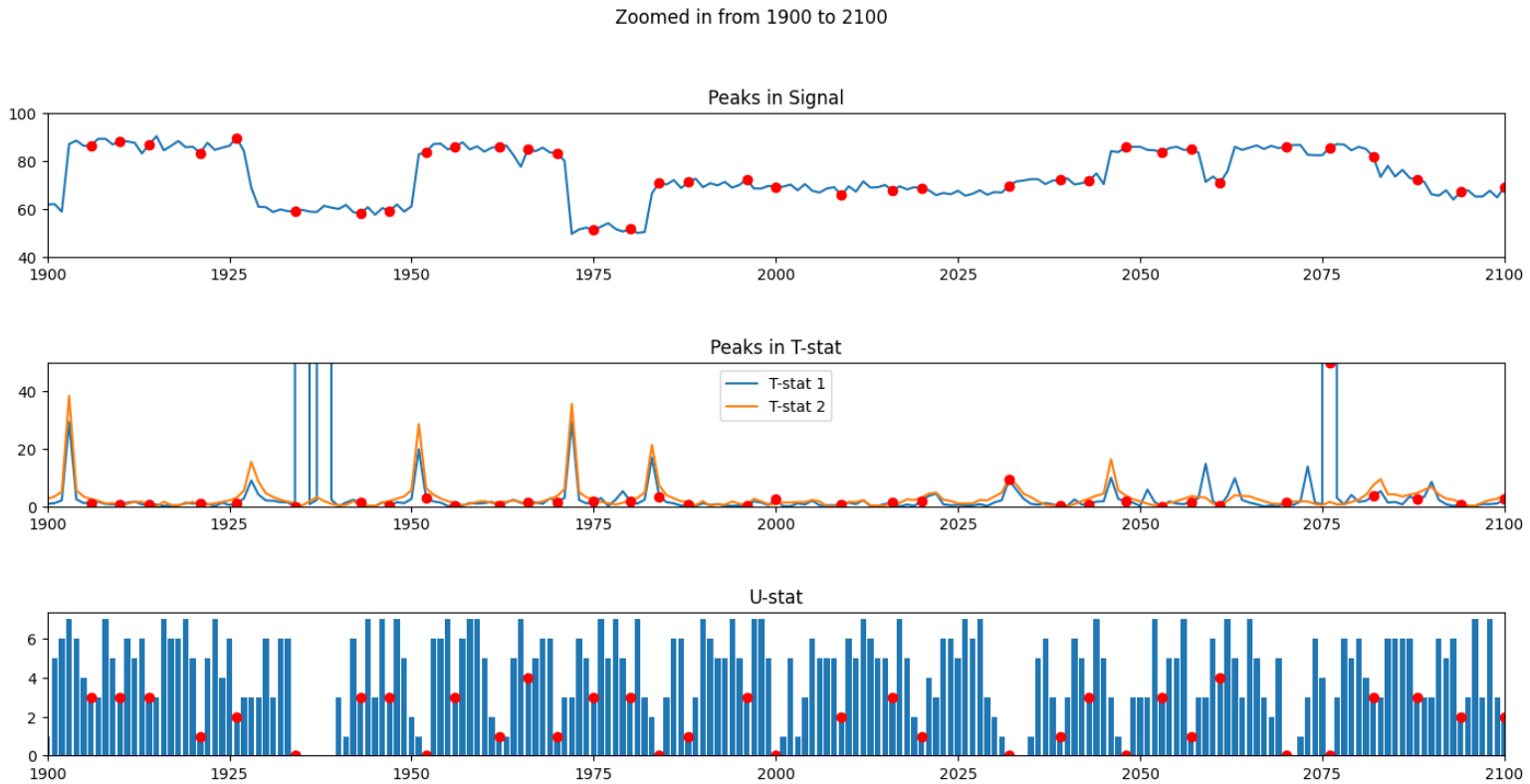
- Peaks are not well spaced
- taking the center point does not always take the minimum U-value.
- Resulting performance was bad.

→ Implement peak detection similar to T-test

U-test peak detection

- **Space peaks according to window length**. If new peak to close old peak: discard it.
- Search maximum U-value between two peaks. **Difference** between maximum and current U-value **must be higher than threshold.**

Red dots represent peaks found with new variant.



Outline

Background on RawHash

U-test implementation

Results

Future steps

Evaluation Methodology

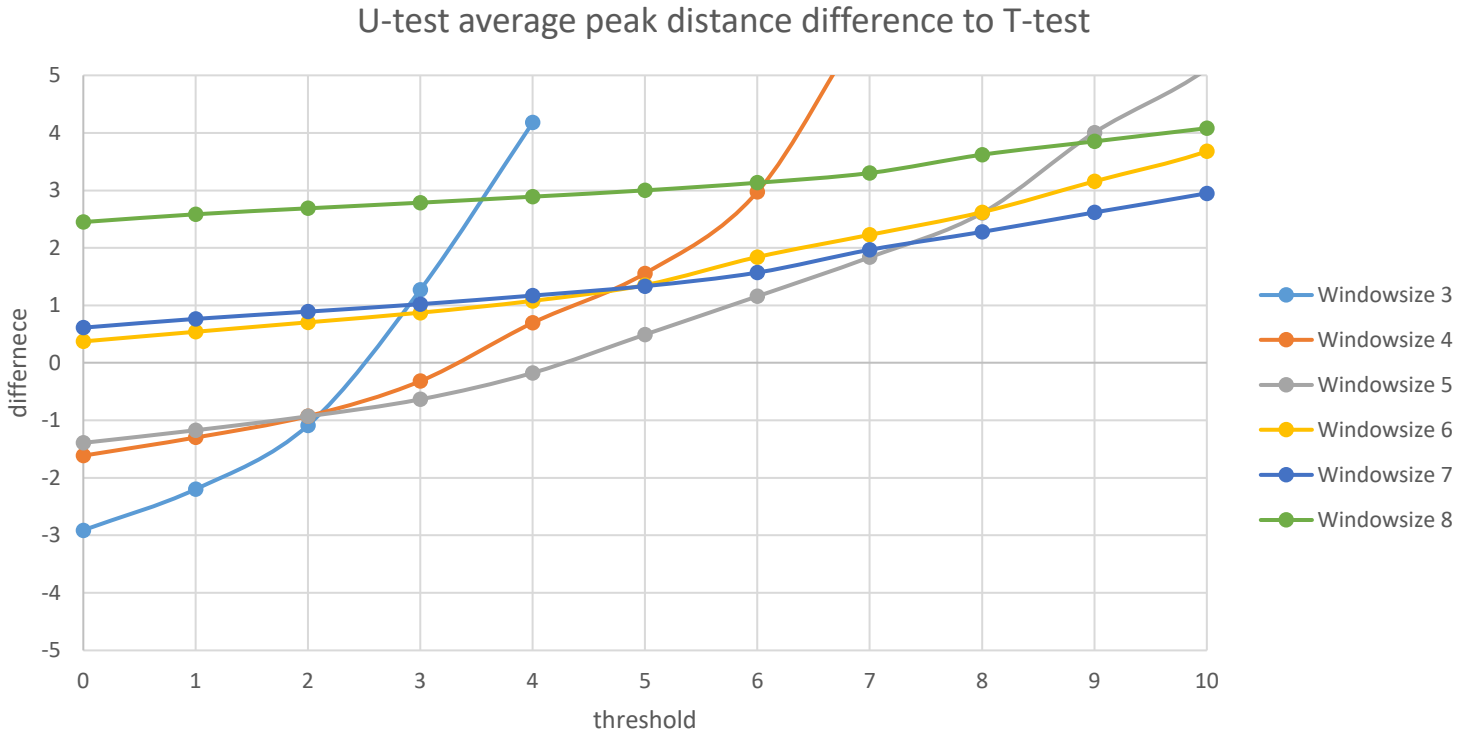
- Evaluation metrics:
  - **Accuracy**
    - **Baseline:** Mapping basecalled reads using minimap2
    - Precision, recall, and F1 scores
- **Datasets:**
  - Only D2 E. coli

	Organism	Reads (#)	Bases (#)	Genome Size
	Read Mapping			
D1	*SARS-CoV-2*	1,382,016	594M	29,903
D2	*E. coli*	353,317	2,365M	5M
D3	*Yeast*	49,989	380M	12M
D4	*Green Algae*	29,933	609M	111M
D5	*Human HG001*	269,507	1,584M	3,117M
	Relative Abundance Estimation			
	D1-D5	2,084,762	5,531M	3,246M
	Contamination Analysis			
	D1 and D5	1,651,523	2,178M	29,903

- Results were generated by sweeping the **window length from 2-20** and threshold accordingly to a **significance of 20%**.

Average peak distance

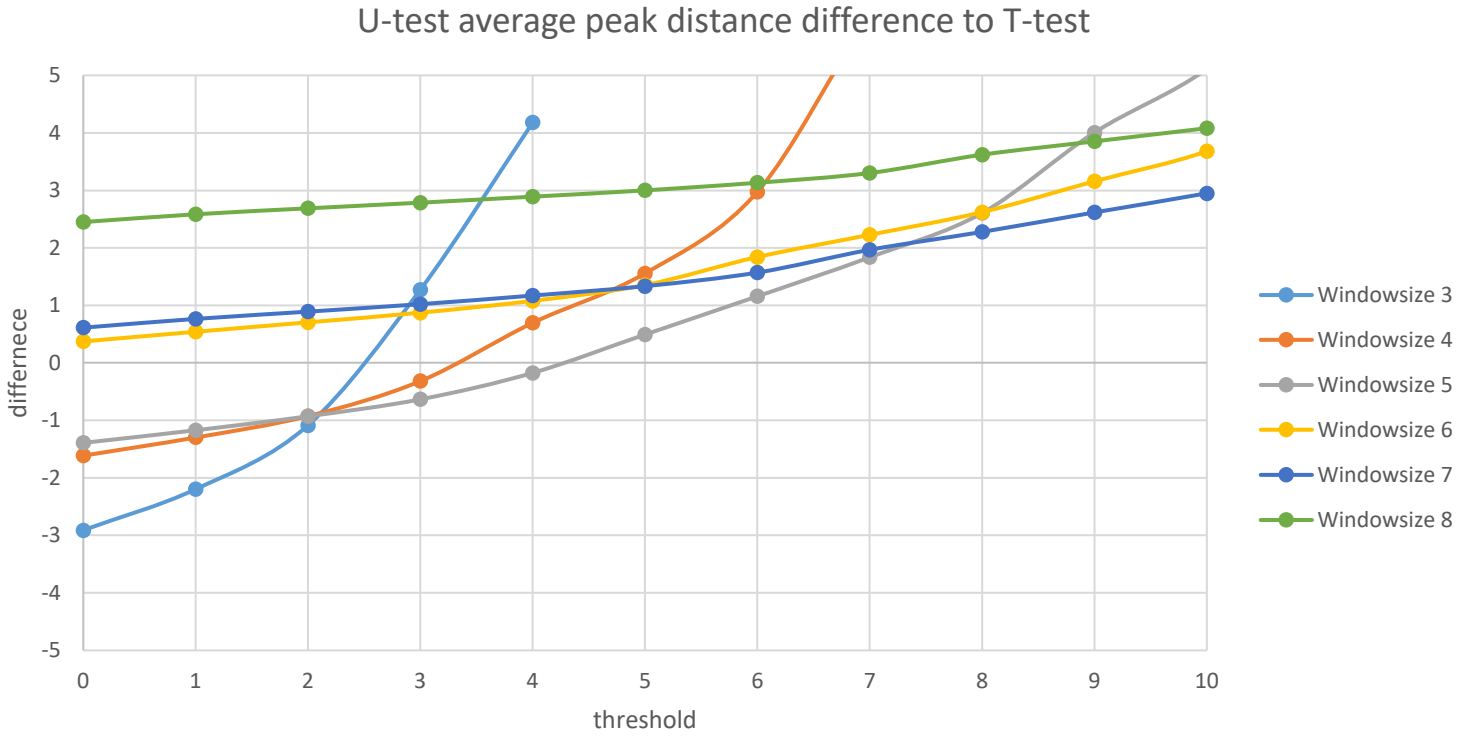
- Average peak distance for T-test on d2 is **8.912**



Observation: single step in threshold has big impact on average distance, which results in a big change in the number of peaks found.

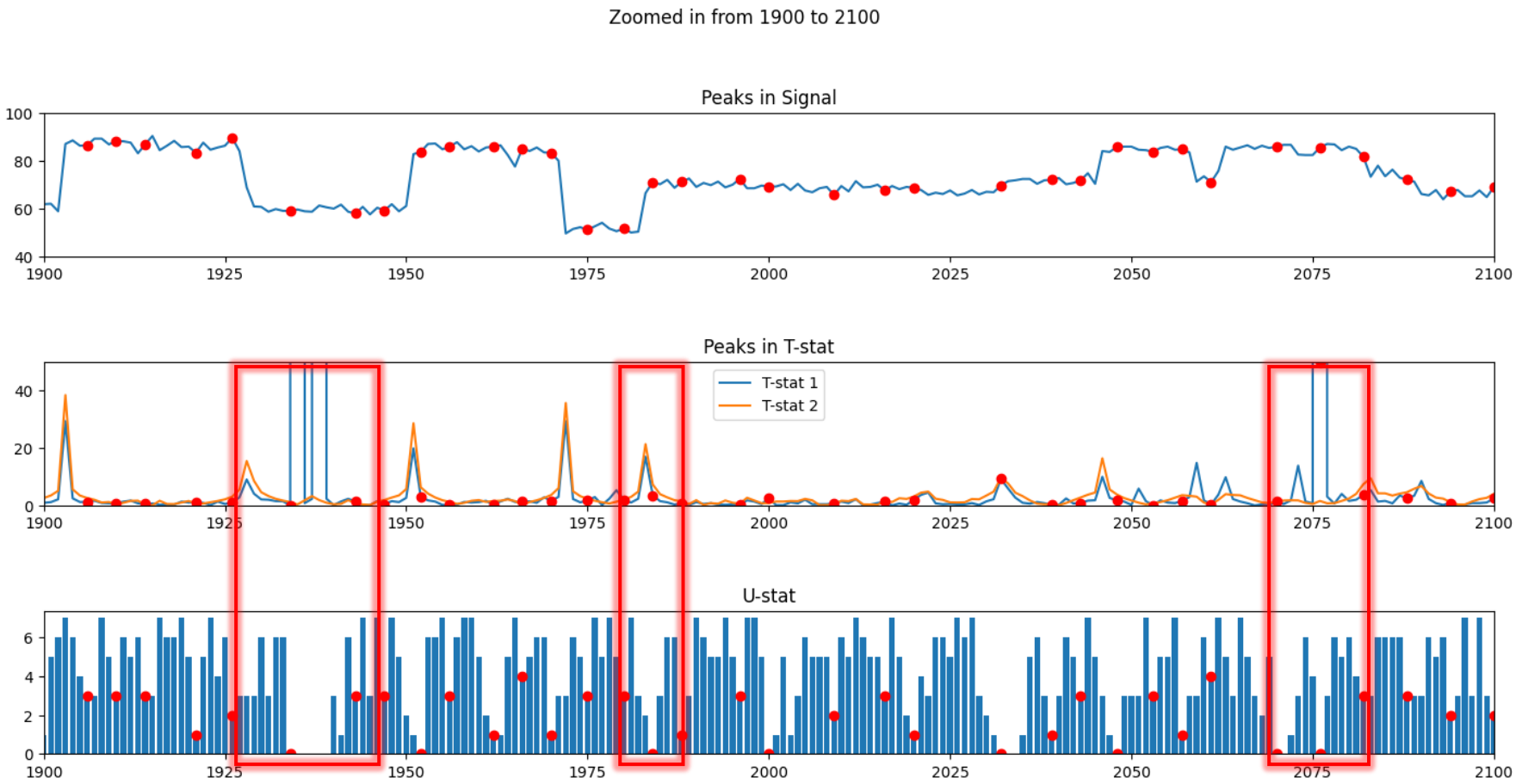
Average peak distance

- Average peak distance for T-test on d2 is **8.912**

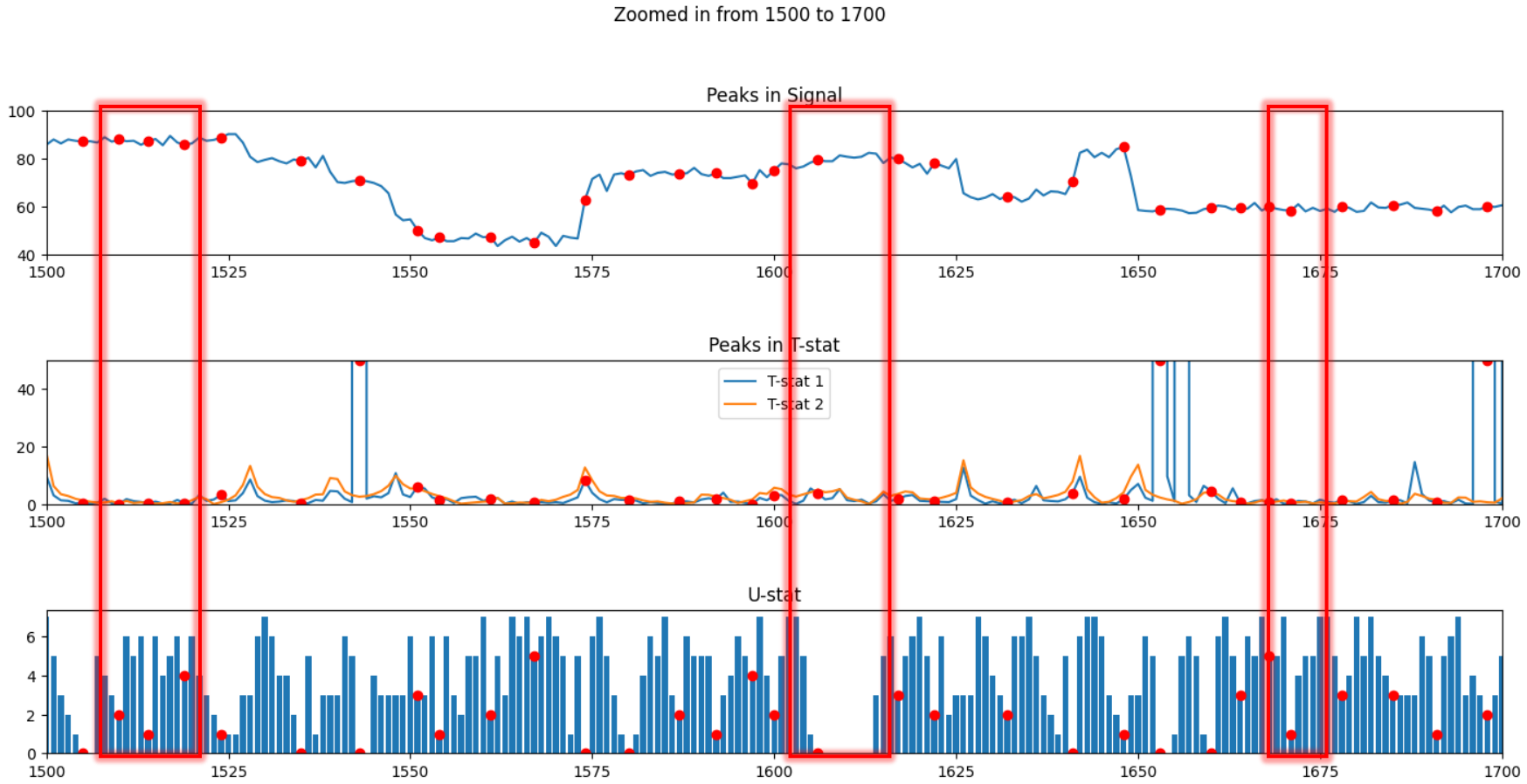


Observation: This problem occurs because U-test uses discrete values.

Problem: Is U-test to sensitive to parameter changes?



Observation: There is some correspondence between T-test and U-test



Observation: regions were signal is continuously increasing / decreasing result in low U-test value

→ Result in bad correlation to T-test

Accuracy

Top 10 parameter combinations sorted by F-1 score > precision > recall

Windowsize	Threshold	Precision	Recall	F-1 score
3	1	0.013123	0.001389	0.002512
10	22	0.013699	0.001085	0.00201
6	3	0.012739	0.001091	0.00201
8	7	0.012158	0.001095	0.00201
6	4	0.011976	0.001097	0.00201
6	15	0.012739	0.001091	0.002009
10	21	0.010417	0.000813	0.001508
6	8	0.010239	0.000814	0.001508
14	32	0.009868	0.000816	0.001508
13	3	0.009615	0.000818	0.001508

Observation: Accuracy is low

Executive Summary

Goal: Try **U-test** as a different statistical tool and evaluate performance.

Key observations:

1) Changes in threshold have big impact on average peak distance → T-test to sensitive to parameter changes
2) There exits correspondence between T-test und U-test signals
3) Overall accuracy is low

Outline

Background on Rawhash

U-test implementation

Results

Future steps

Future steps

- **Improve peak detection algorithm**
  - Use multiple window sizes
  - Try to match average peak distance to T-test results

- **Try to make end-to-end itegration more accurate**
  - double check U-test implementation

- **Try different statistical tests**
  - Non-discrete Tests, which can be parameterized more sensitively
  - E.g. ….

Title