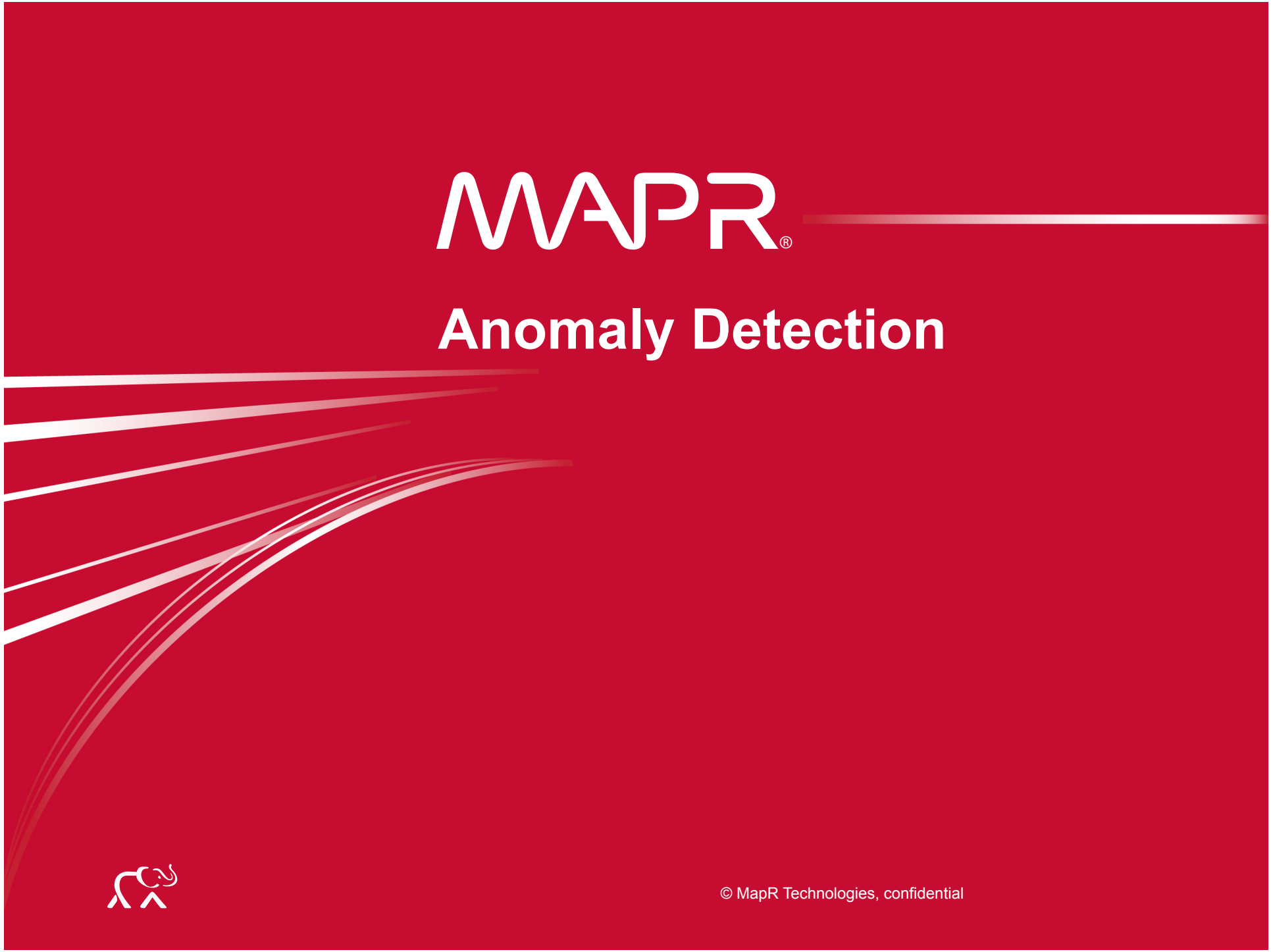


Agenda

- What is anomaly detection?
- Some examples
- Some generalization
- More interesting examples
- Sample implementation methods

Who I am

- Ted Dunning, Chief Application Architect, MapR
  tdunning@mapr.com
  tdunning@apache.org
  @ted_dunning

- Committer, mentor, champion, PMC member on several Apache projects
- Mahout, Drill, Zookeeper others

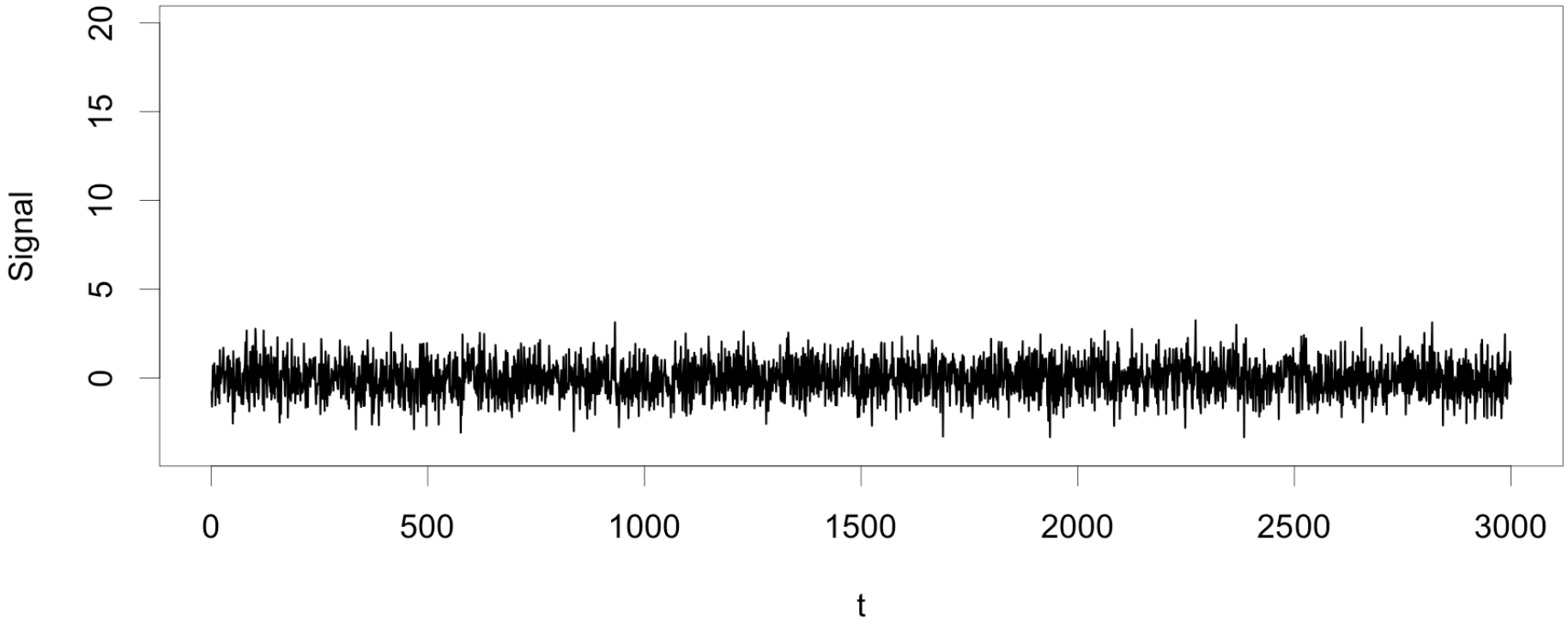
Who we are

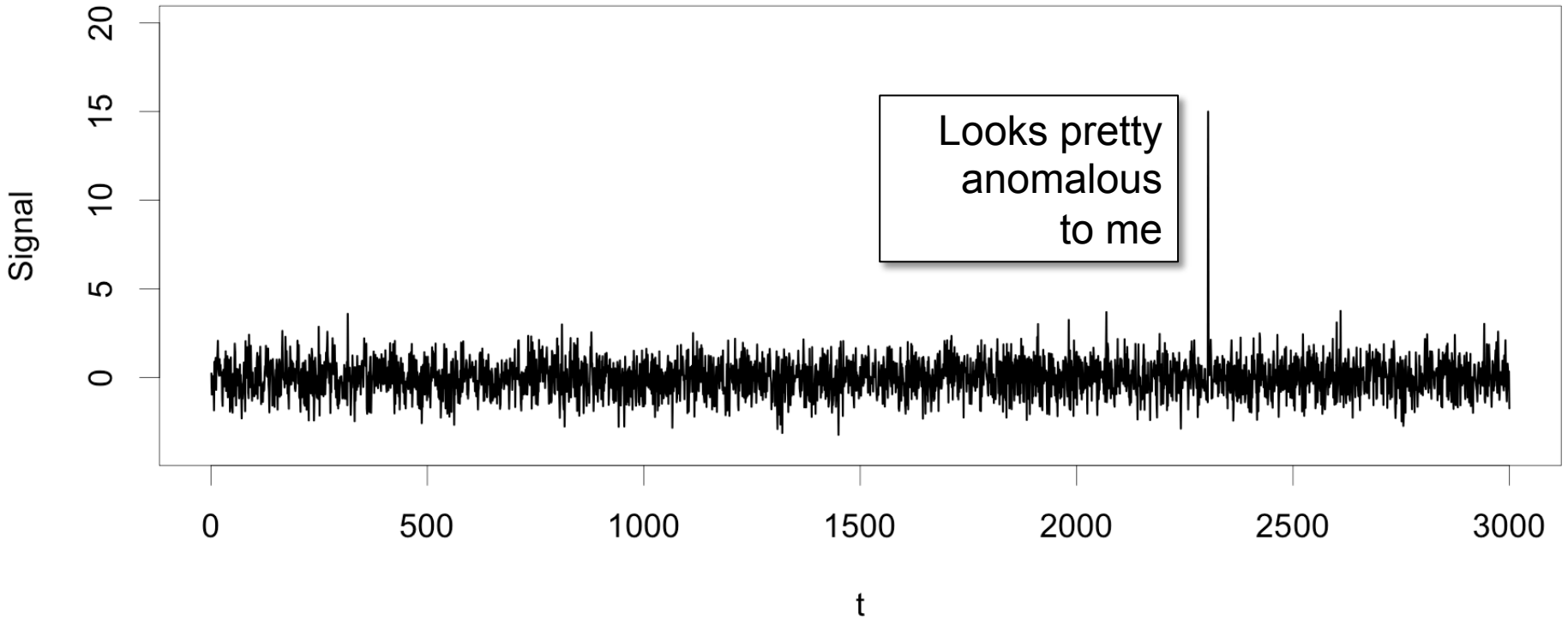
- MapR makes the technology leading distribution including Hadoop
- MapR integrates real-time data semantics directly into a system that also runs Hadoop programs seamlessly
- The biggest and best choose MapR
  - Google, Amazon
  - Largest credit card, retailer, health insurance, telco
  - Ping me for info

What is Anomaly Detection?

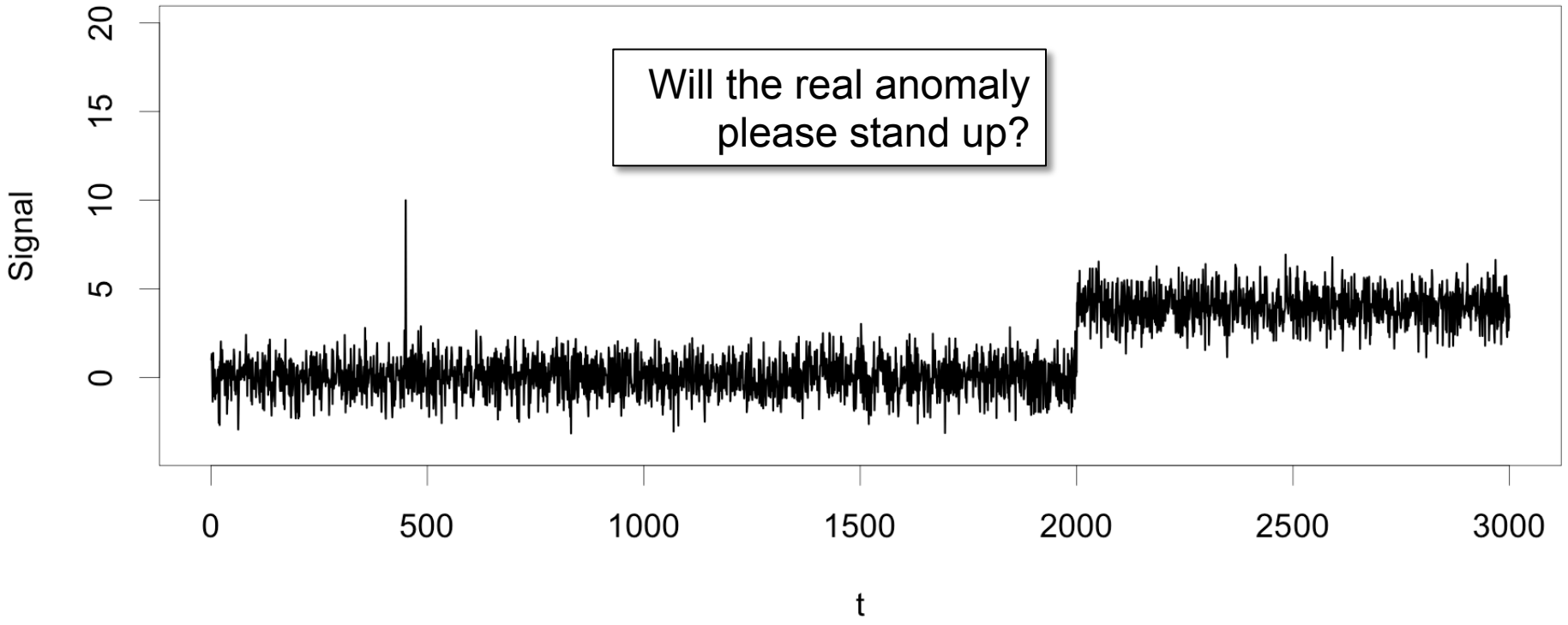
- What just happened that shouldn't?
  - but I don't know what failure looks like (yet)
- Find the problem before other people see it
  - especially customers and CEO's
- But don't wake me up if it isn't really broken

MAPR





MAPR



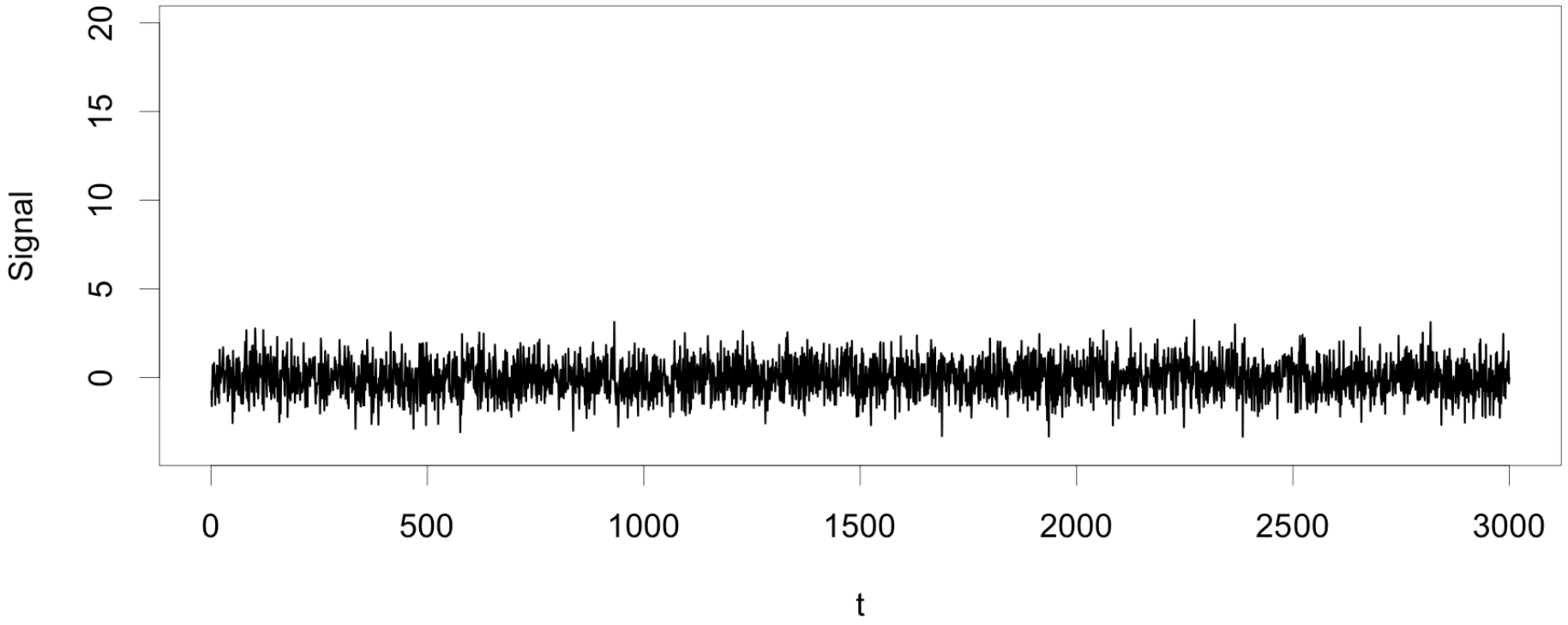
MAPR

What Are We Really Doing

- We want action when something breaks
  (dies/falls over/otherwise gets in trouble)
- But action is expensive
- So we don't want false alarms
- And we don't want false negatives

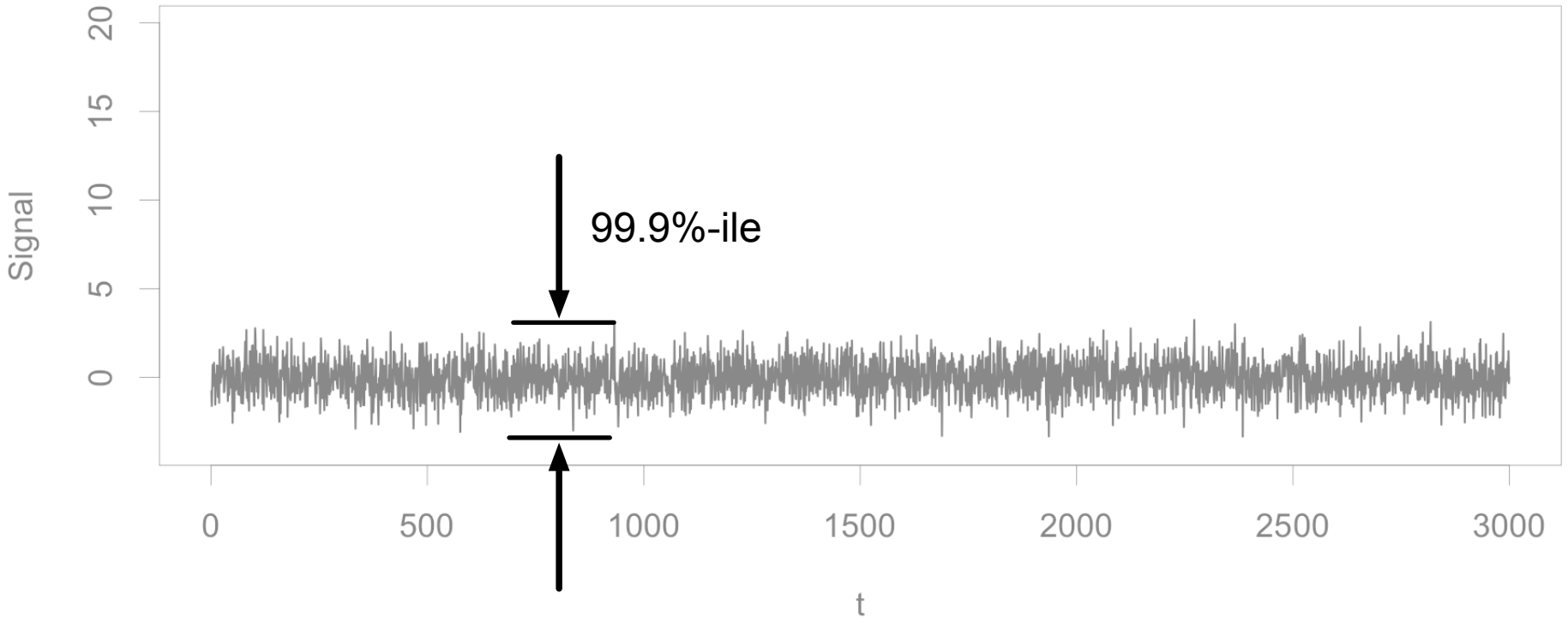
- We need to trade off costs

A Second Look



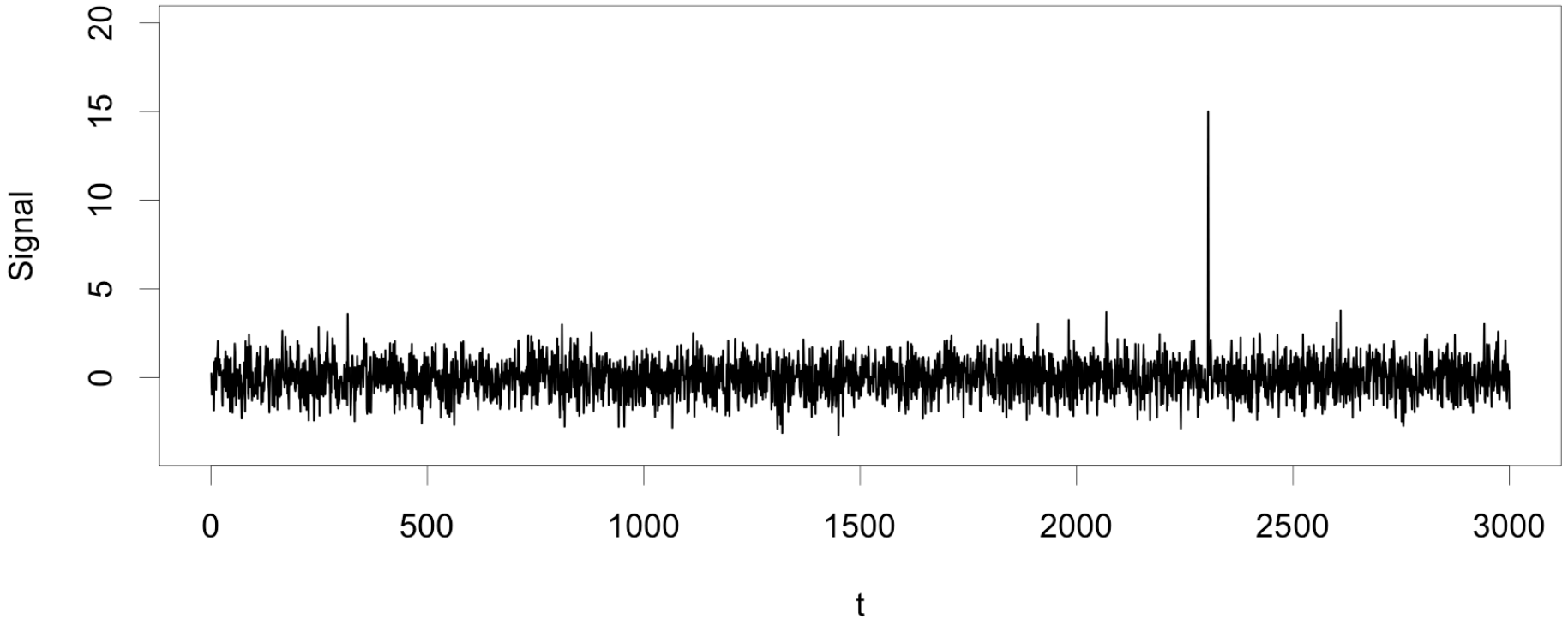
MAPR

A Second Look



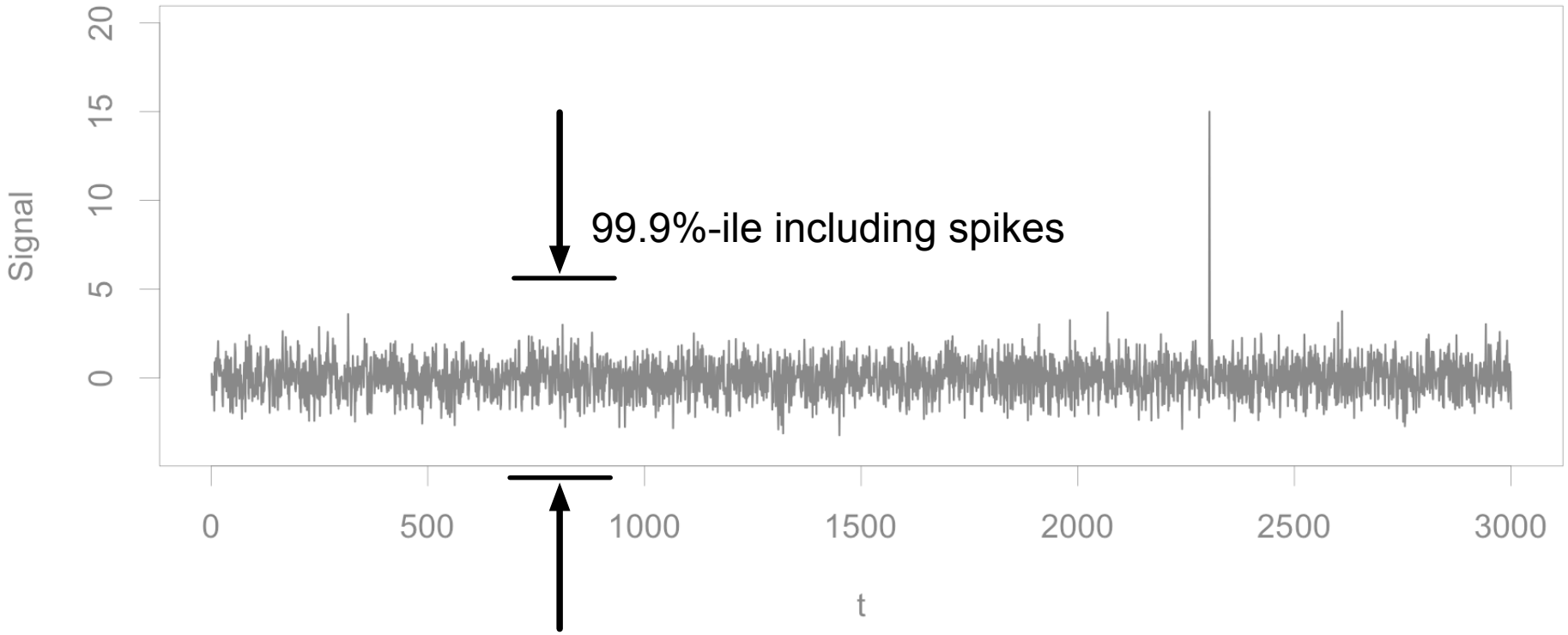
MAPR

With Spikes



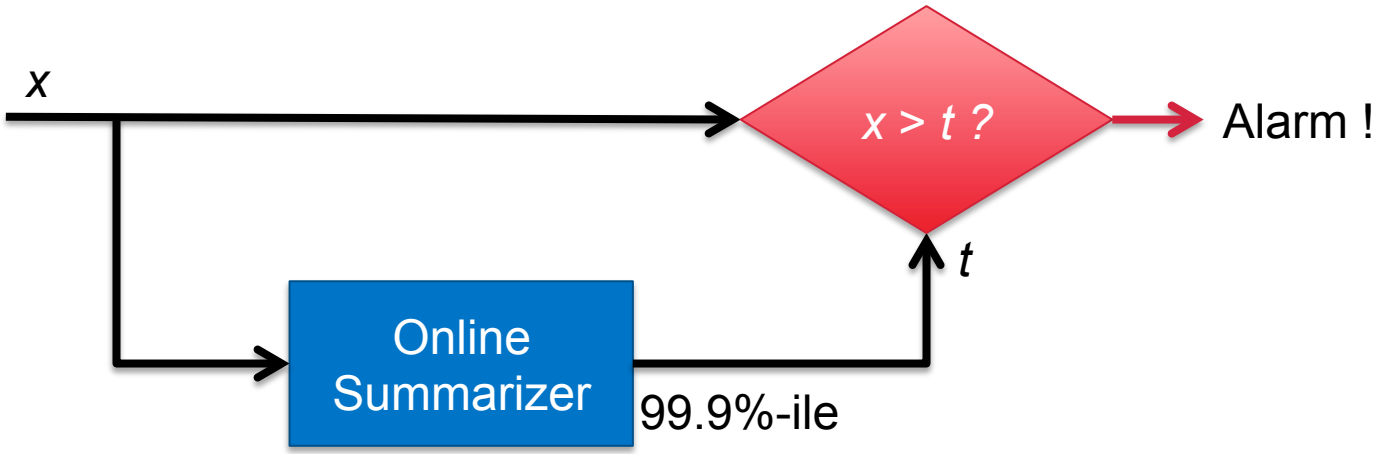
MAPR

With Spikes



MAPR

How Hard Can it Be?

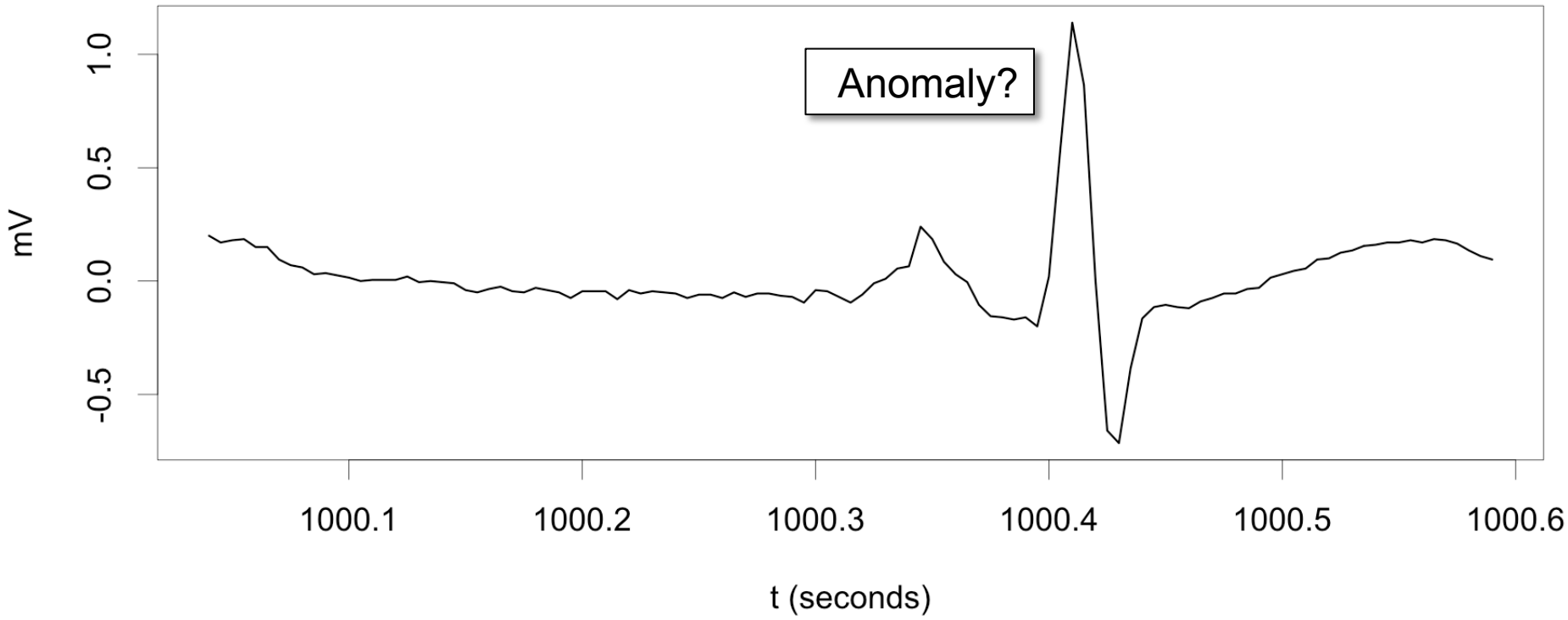


MAPR

On-line Percentile Estimates

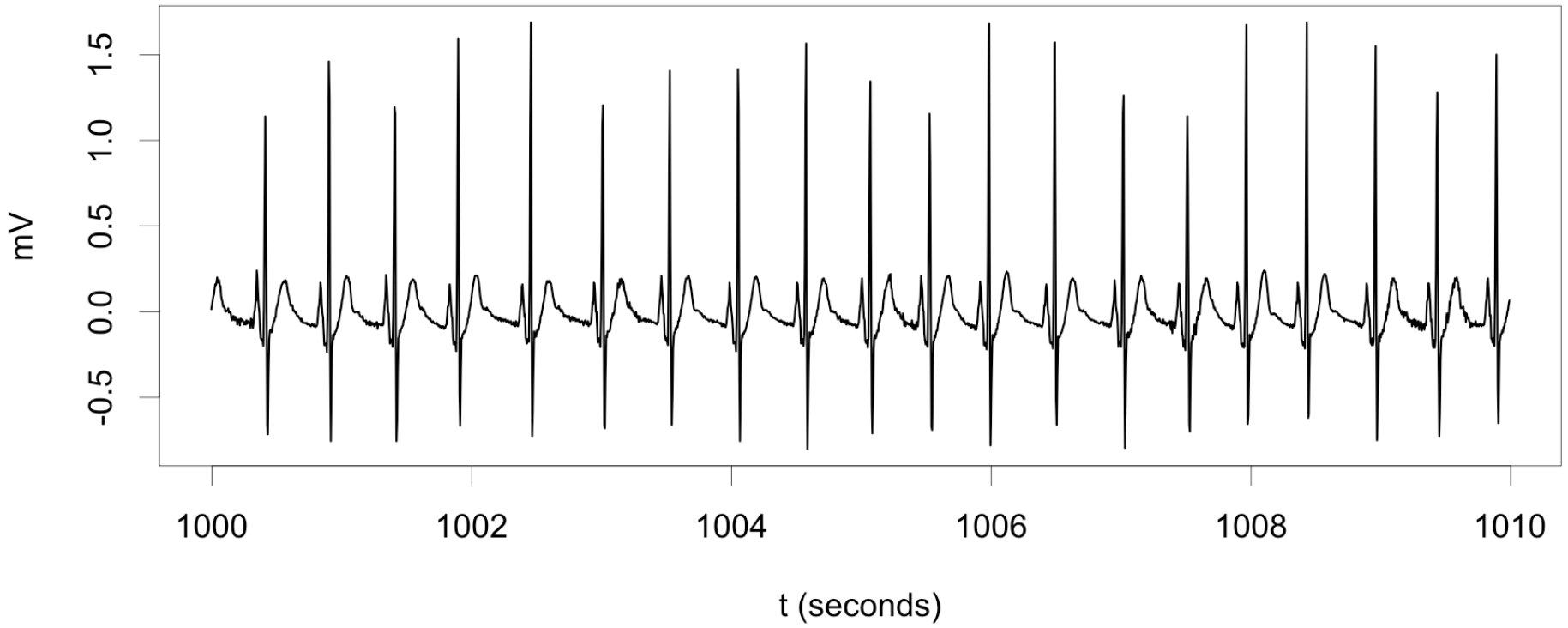
- Apache Mahout has on-line percentile estimator
  - very high accuracy for extreme tails
  - new in version 0.9 !!
- What's the big deal with anomaly detection?
- This looks like a solved problem

Spot the Anomaly



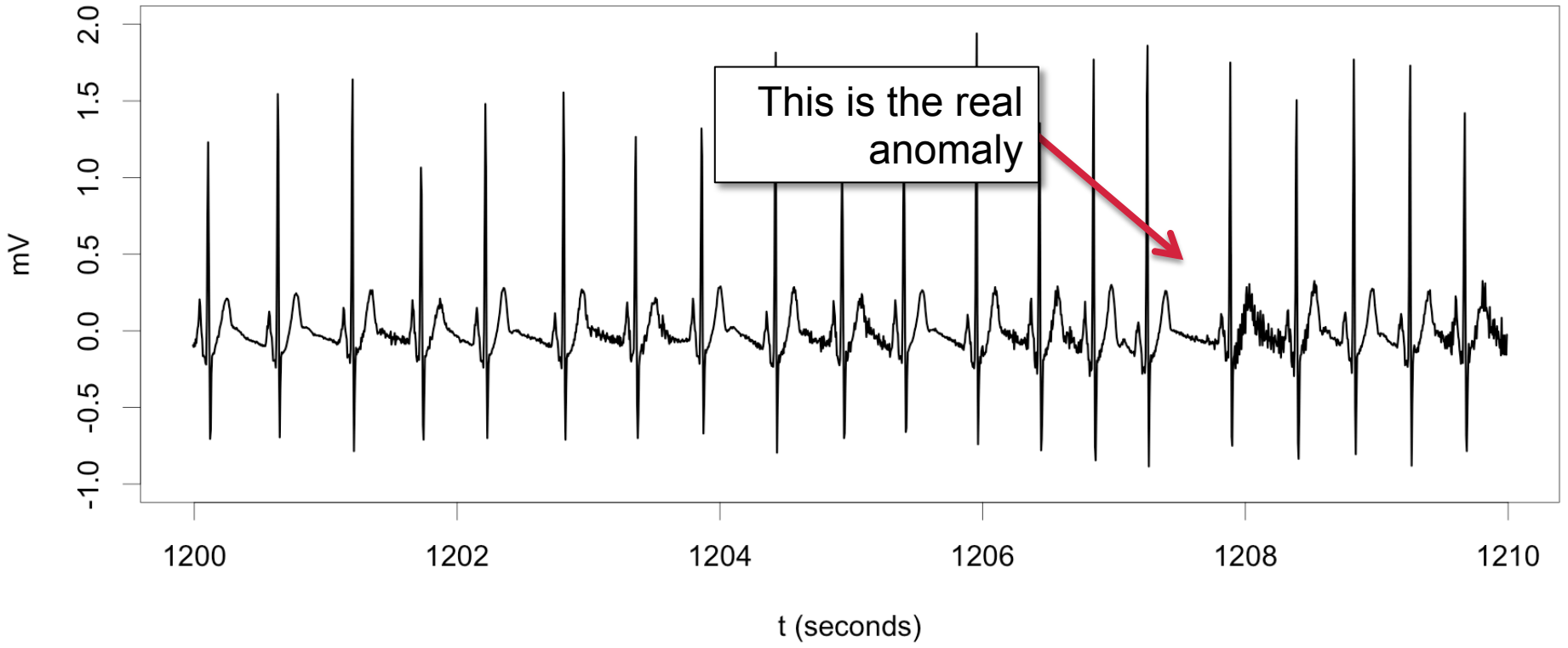
MAPR

Maybe not!



MAPR

Where's Waldo?



MAPR

Normal Isn't Just Normal

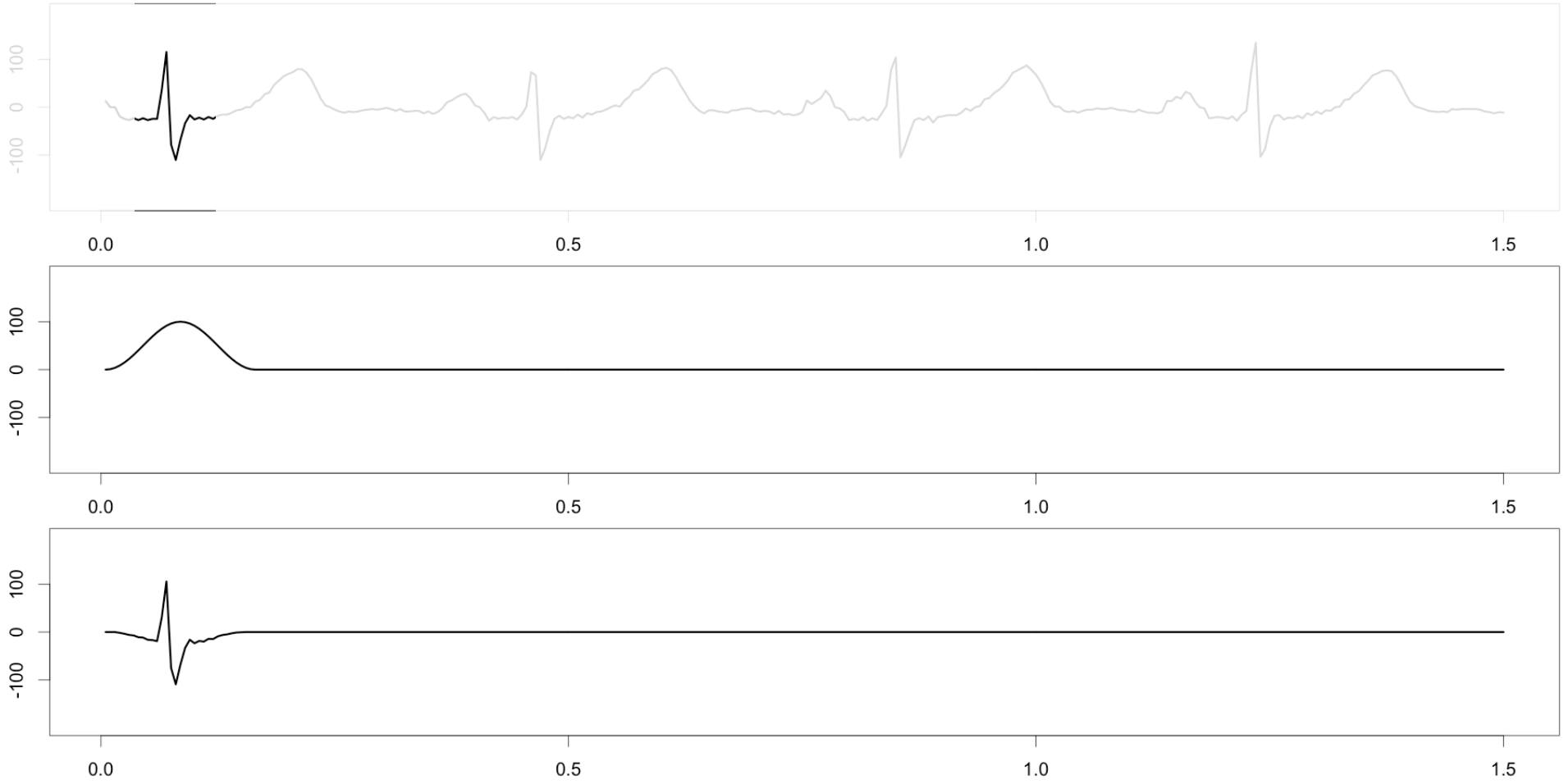
- What we want is a *model* of what is normal
- What doesn't fit the model is the *anomaly*
- For simple signals, the model can be simple …

$$x \sim N(0, \varepsilon)$$

- The real world is rarely so accommodating

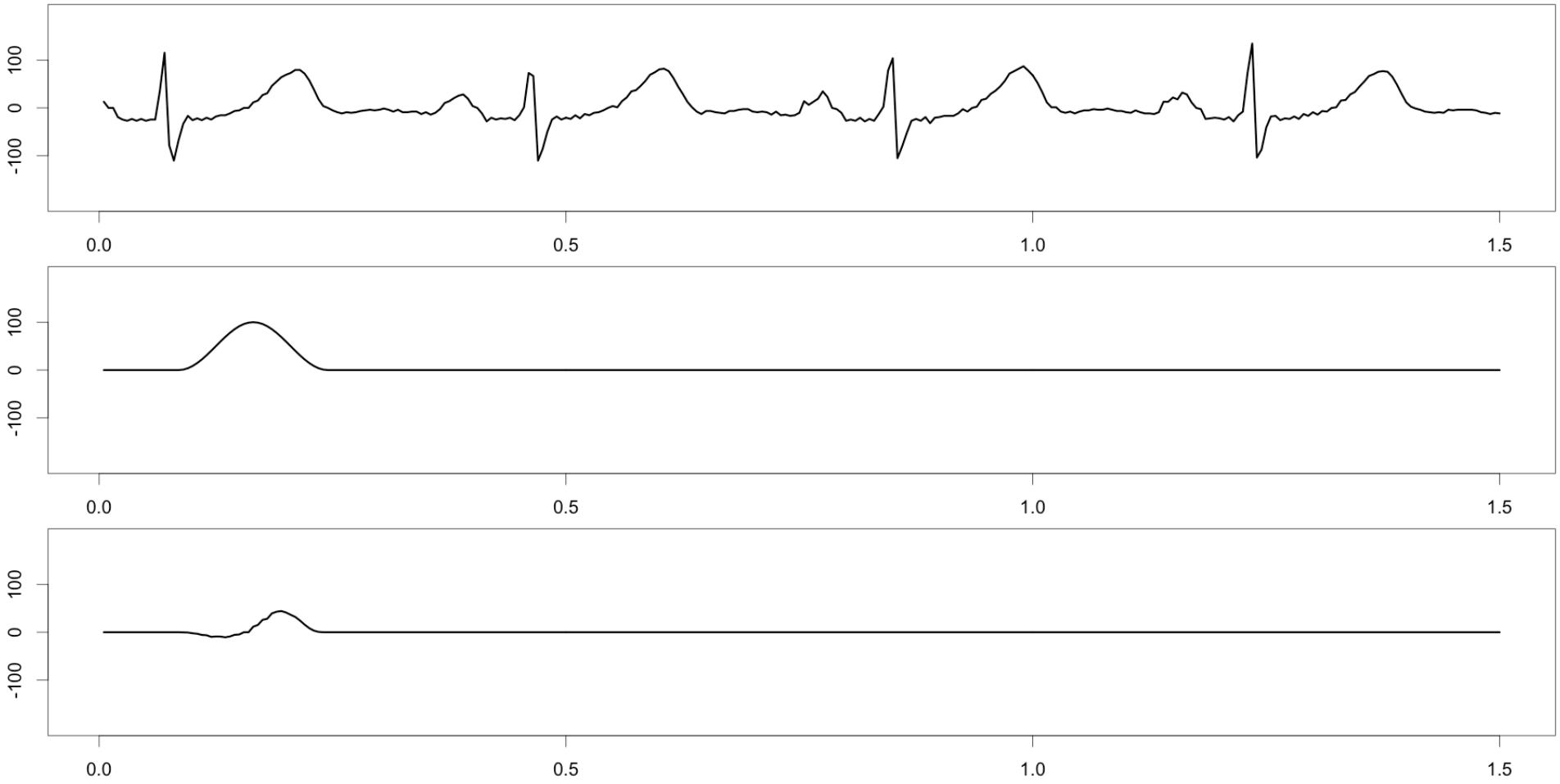
MAPR

We Do Windows



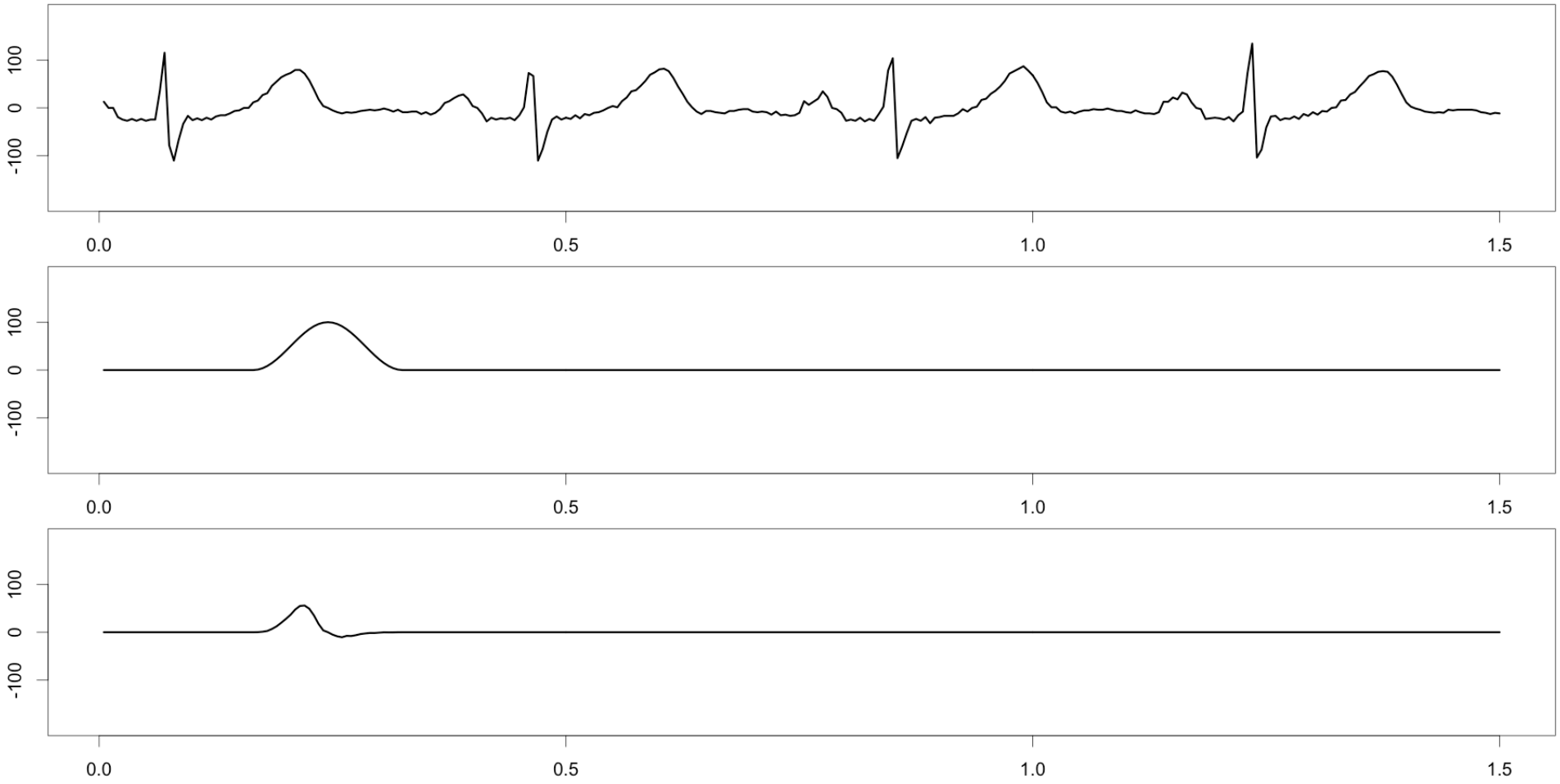
MAPR

We Do Windows



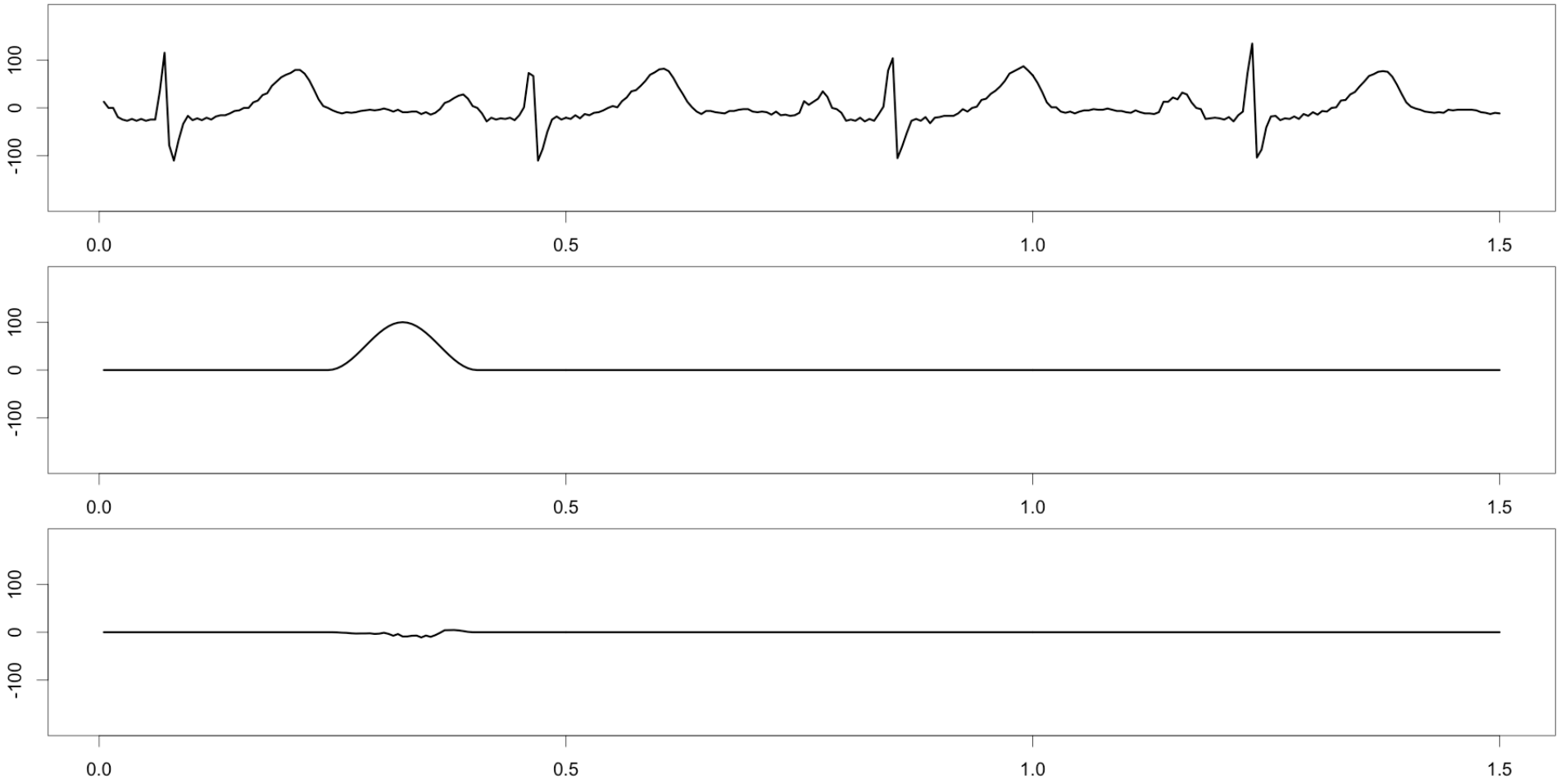
MAPR

We Do Windows



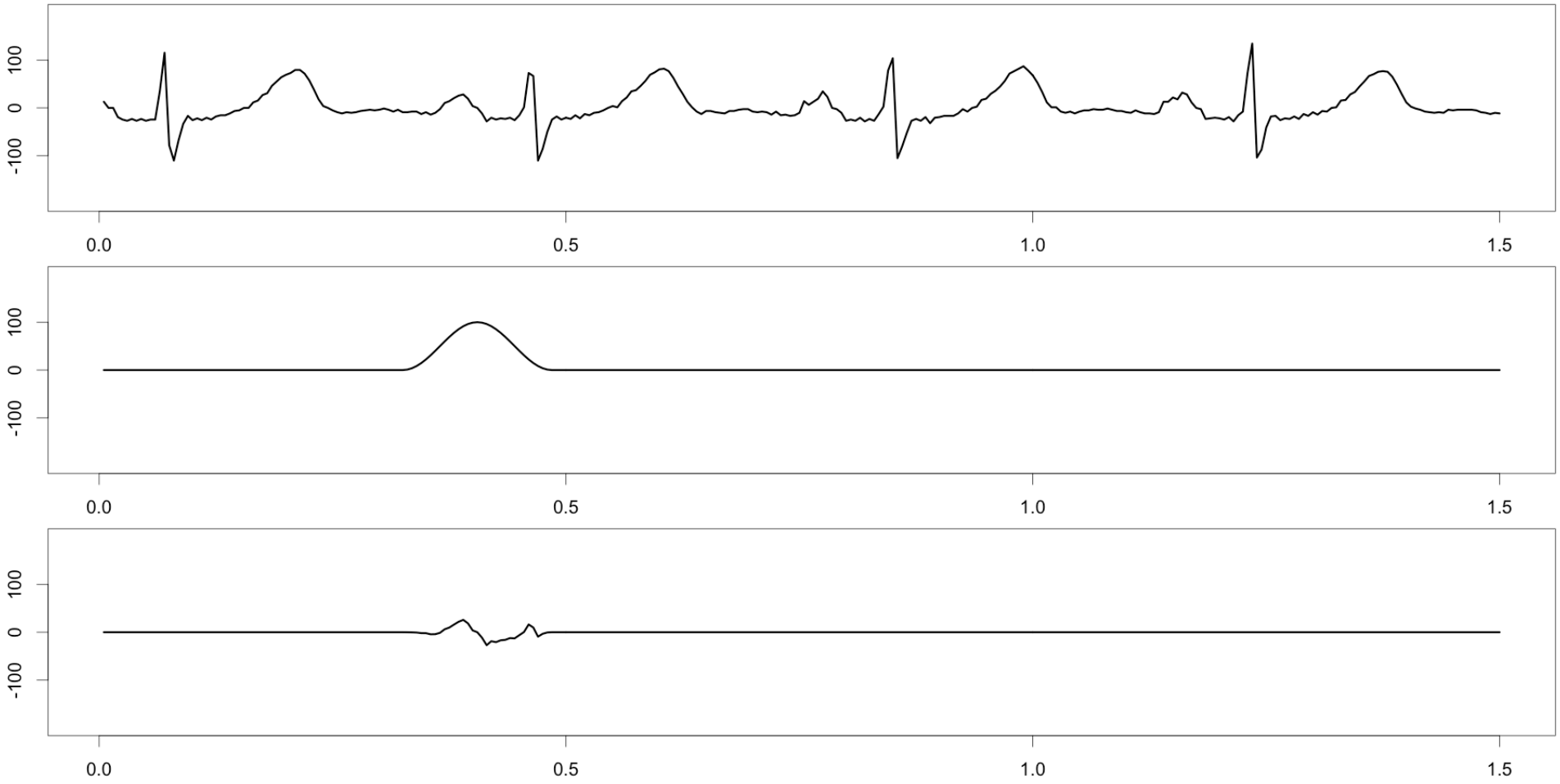
MAPR

We Do Windows



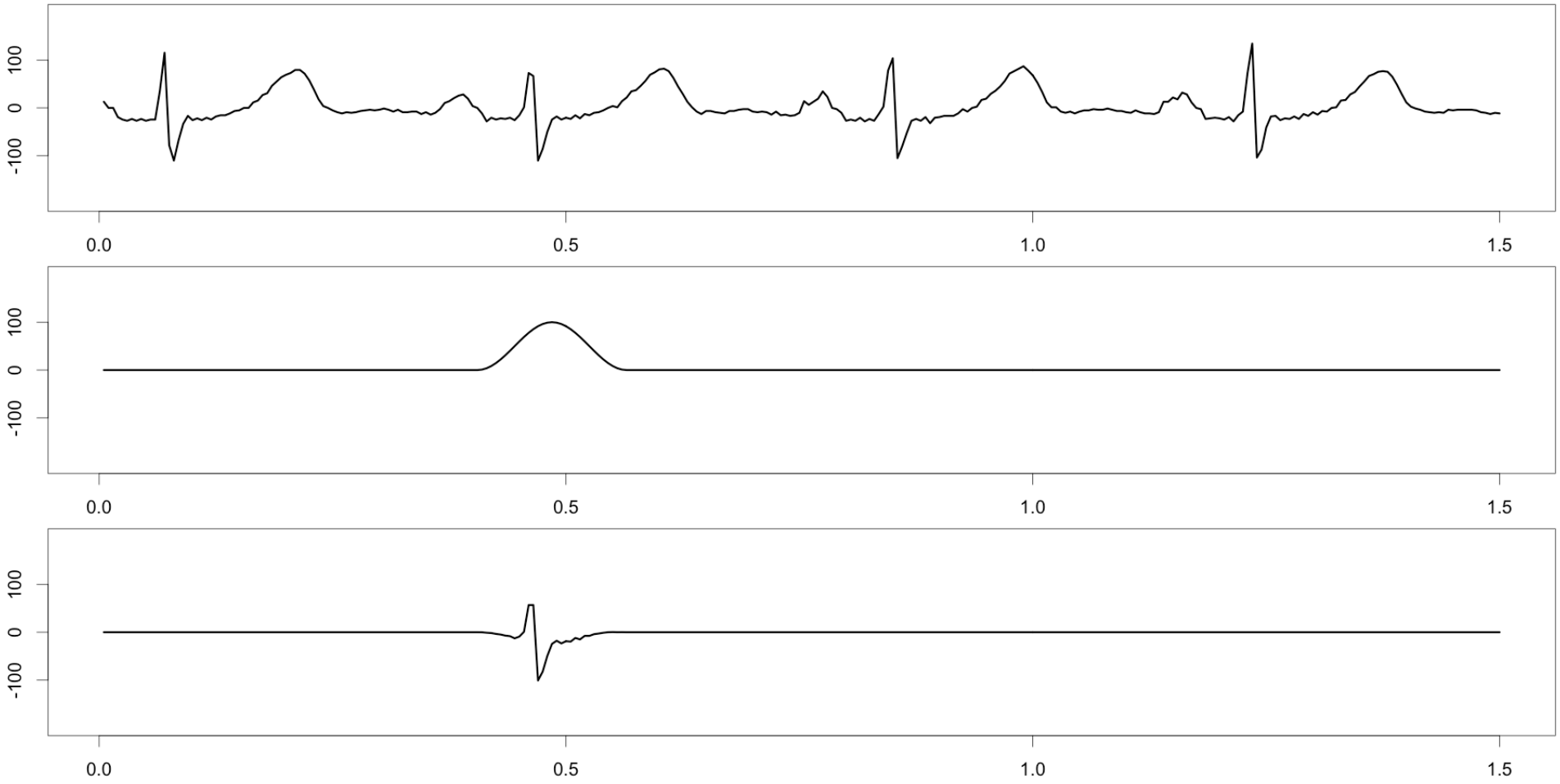
MAPR

We Do Windows



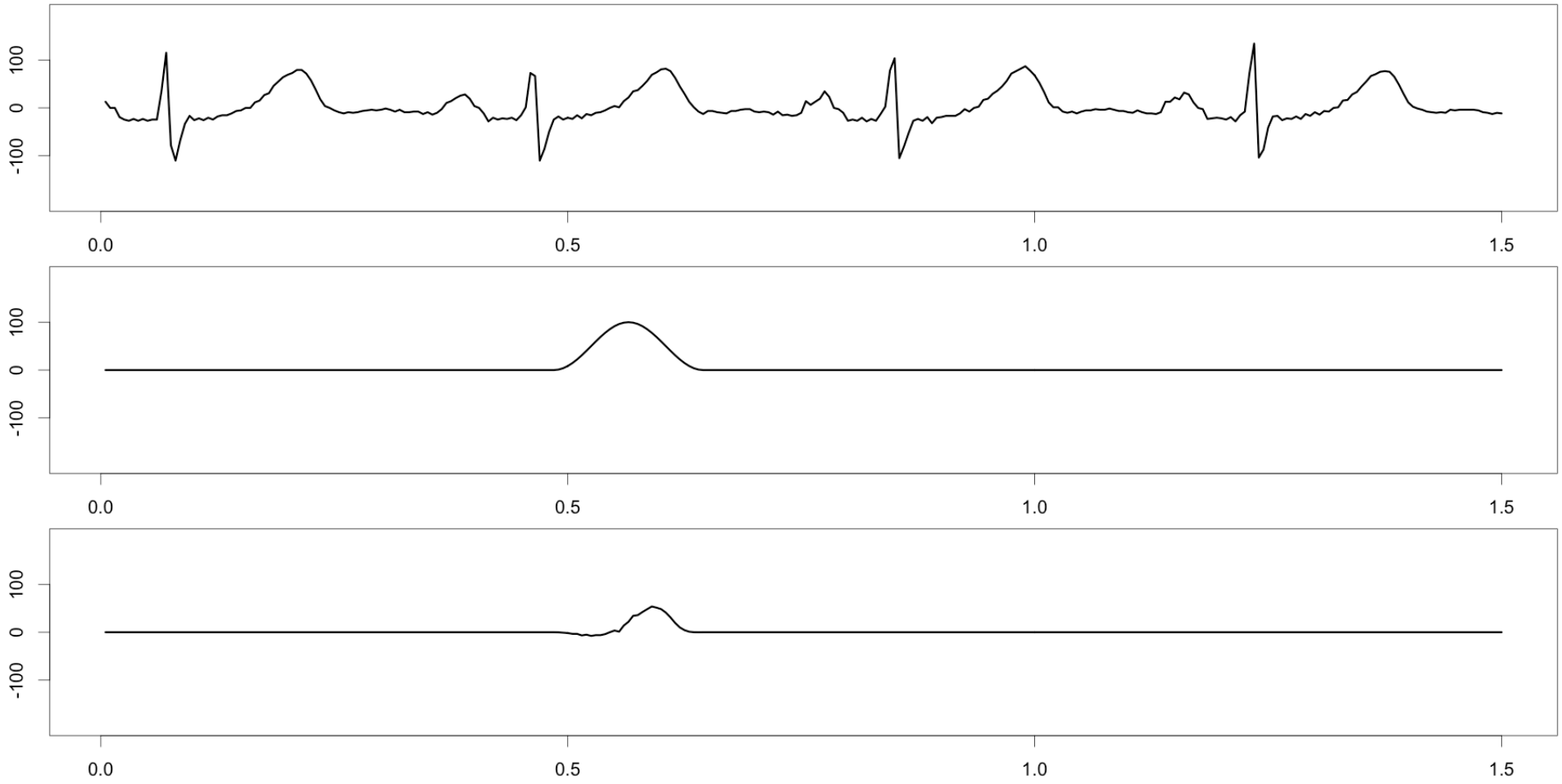
MAPR

We Do Windows



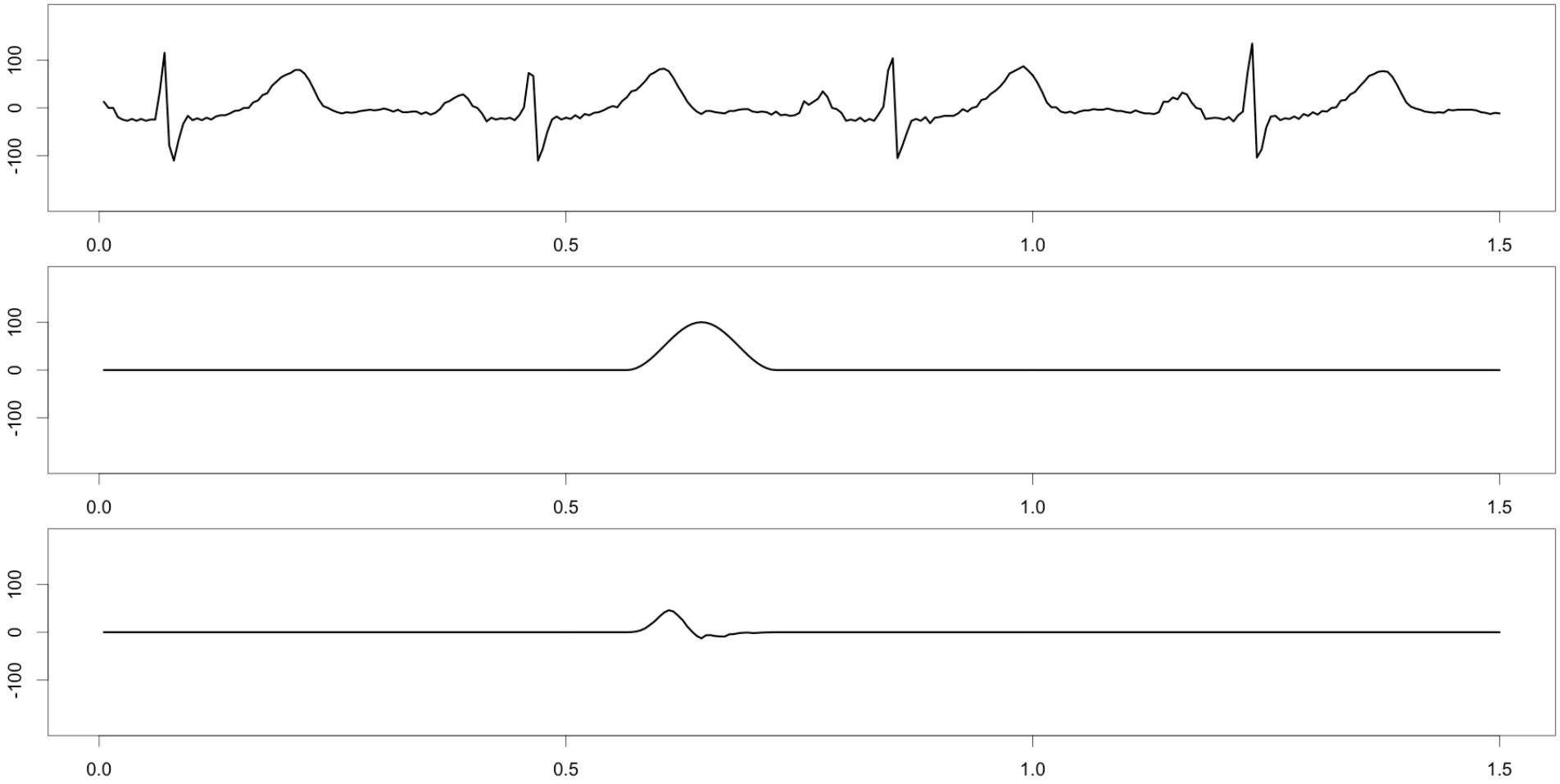
MAPR

We Do Windows



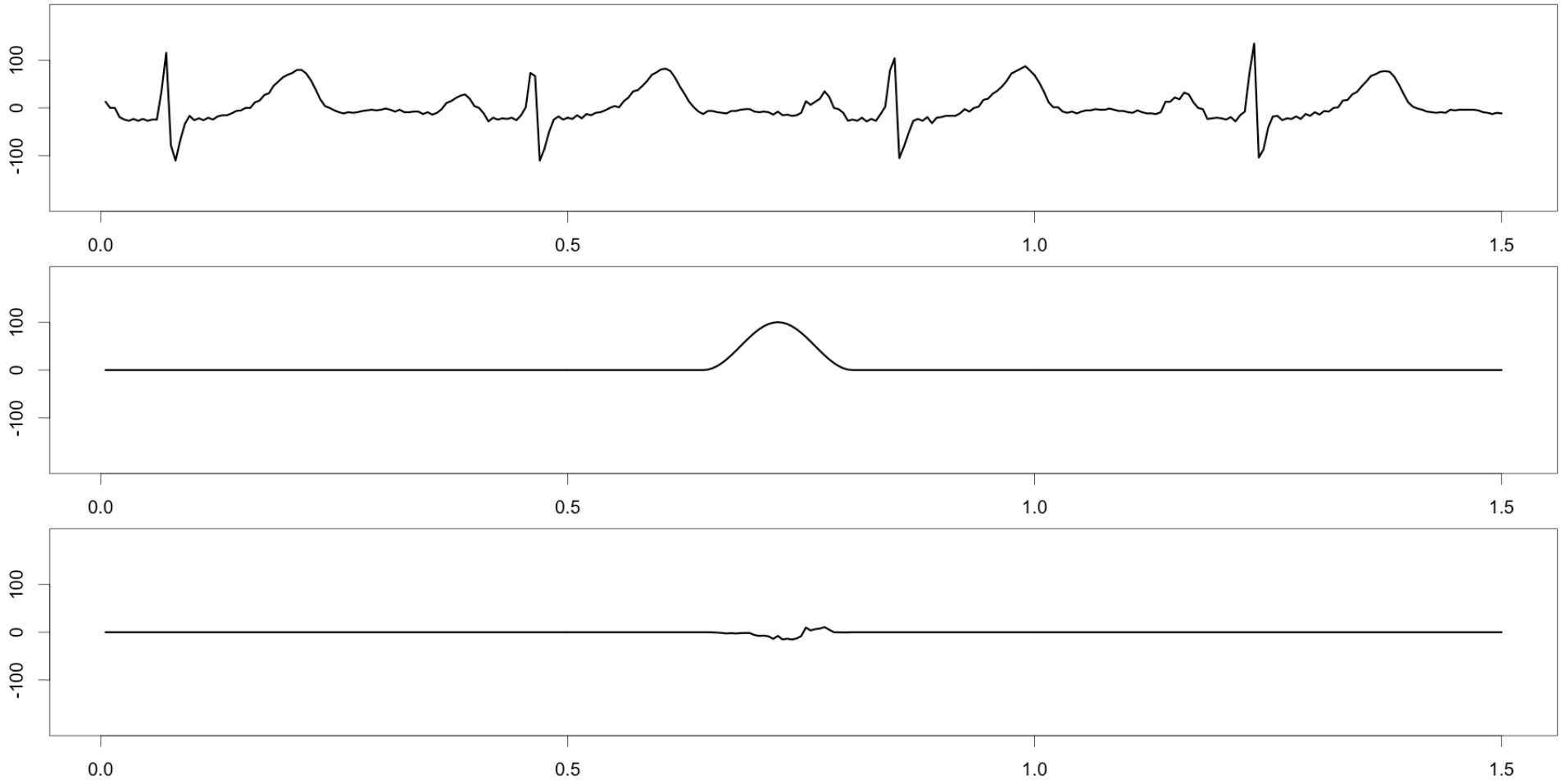
MAPR

We Do Windows



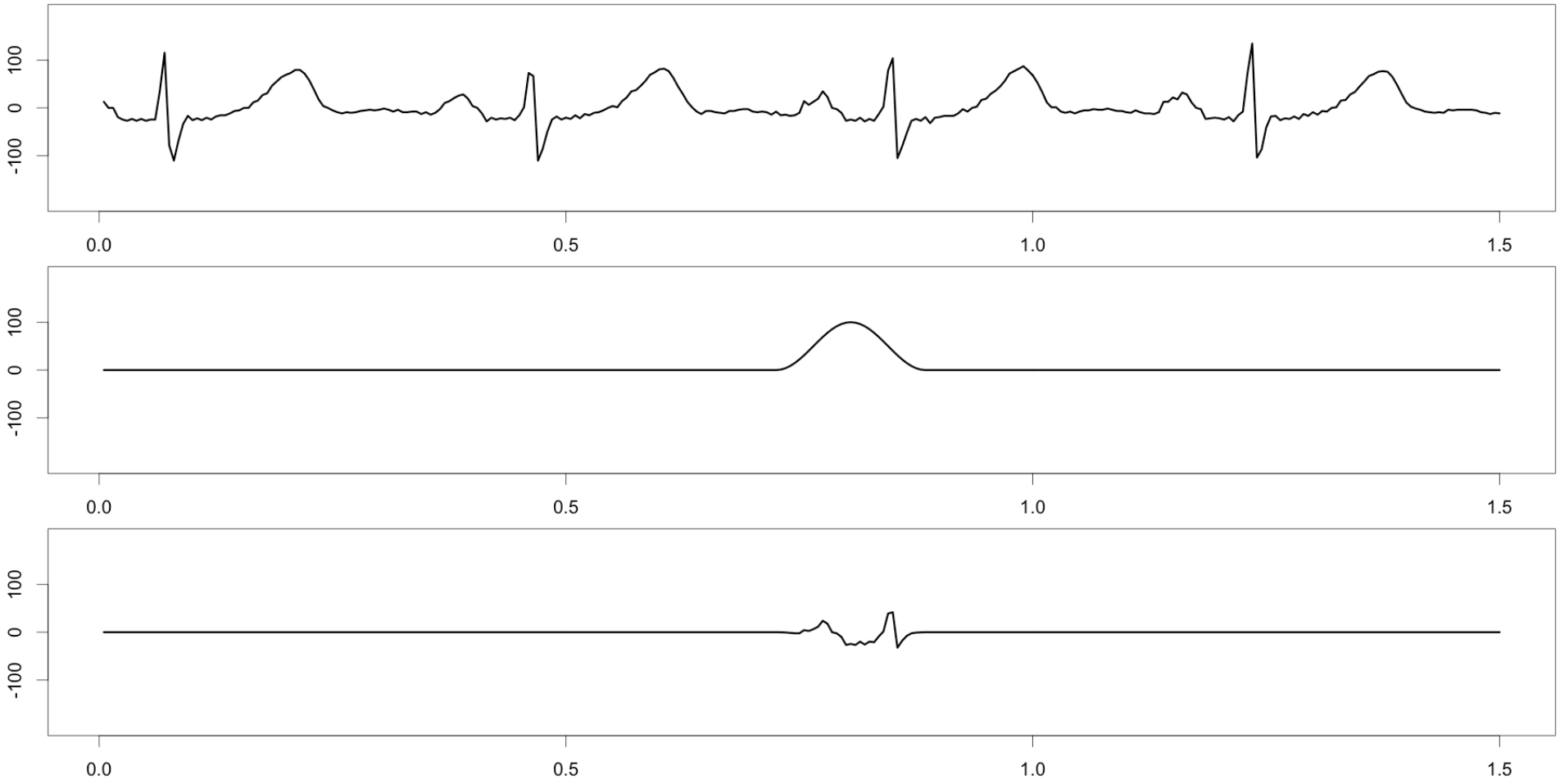
MAPR

We Do Windows



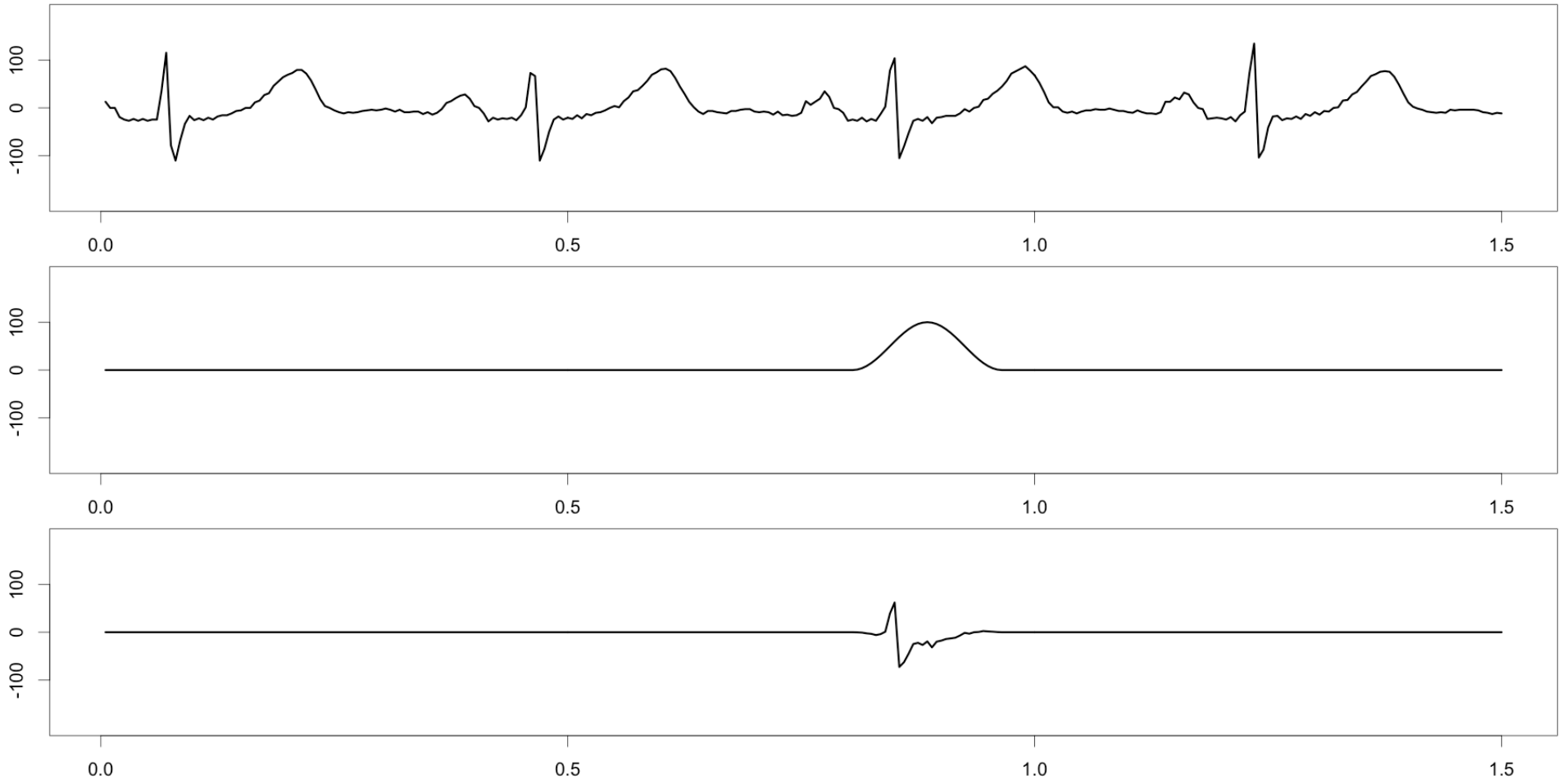
MAPR

We Do Windows



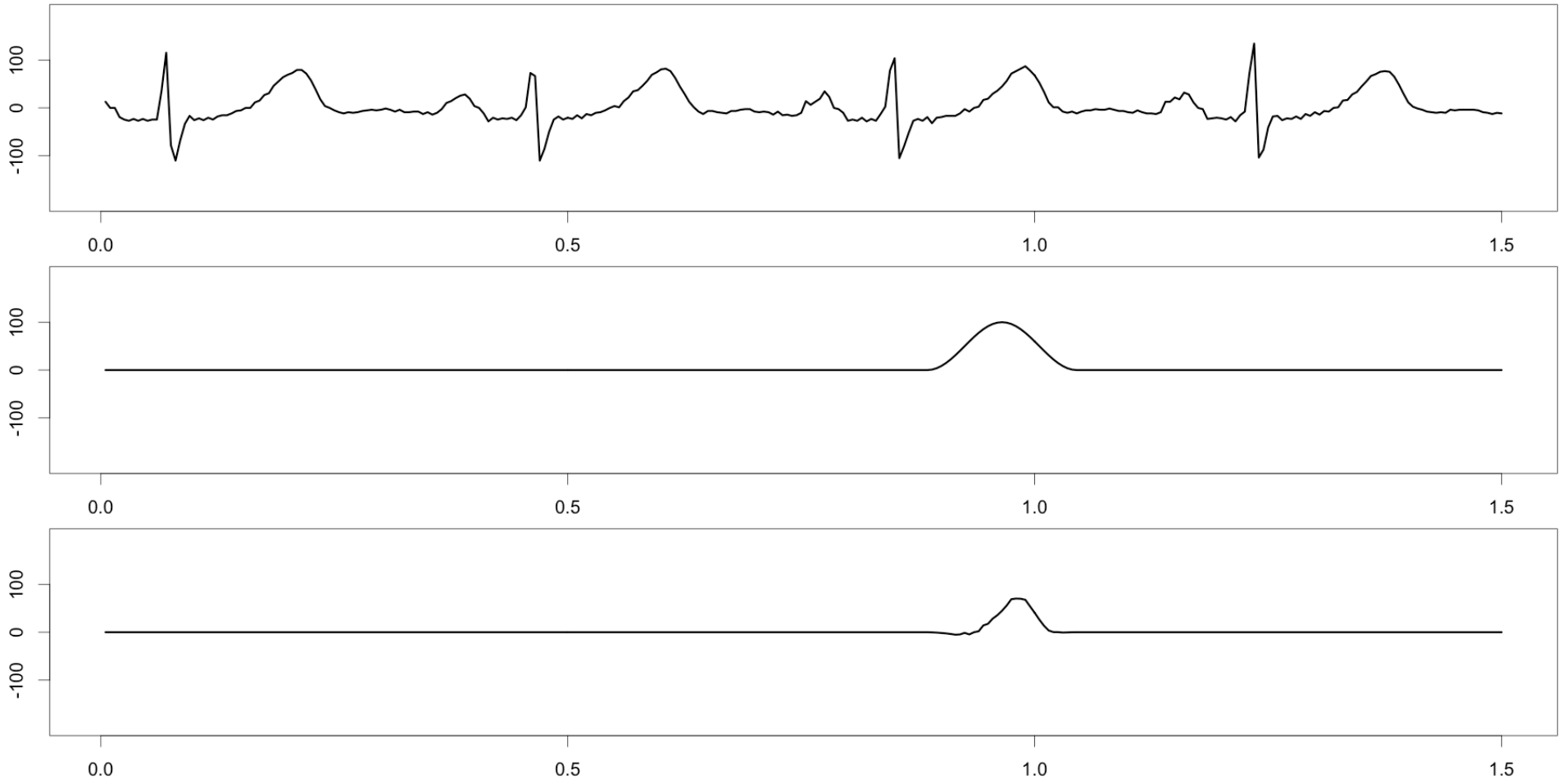
MAPR

We Do Windows



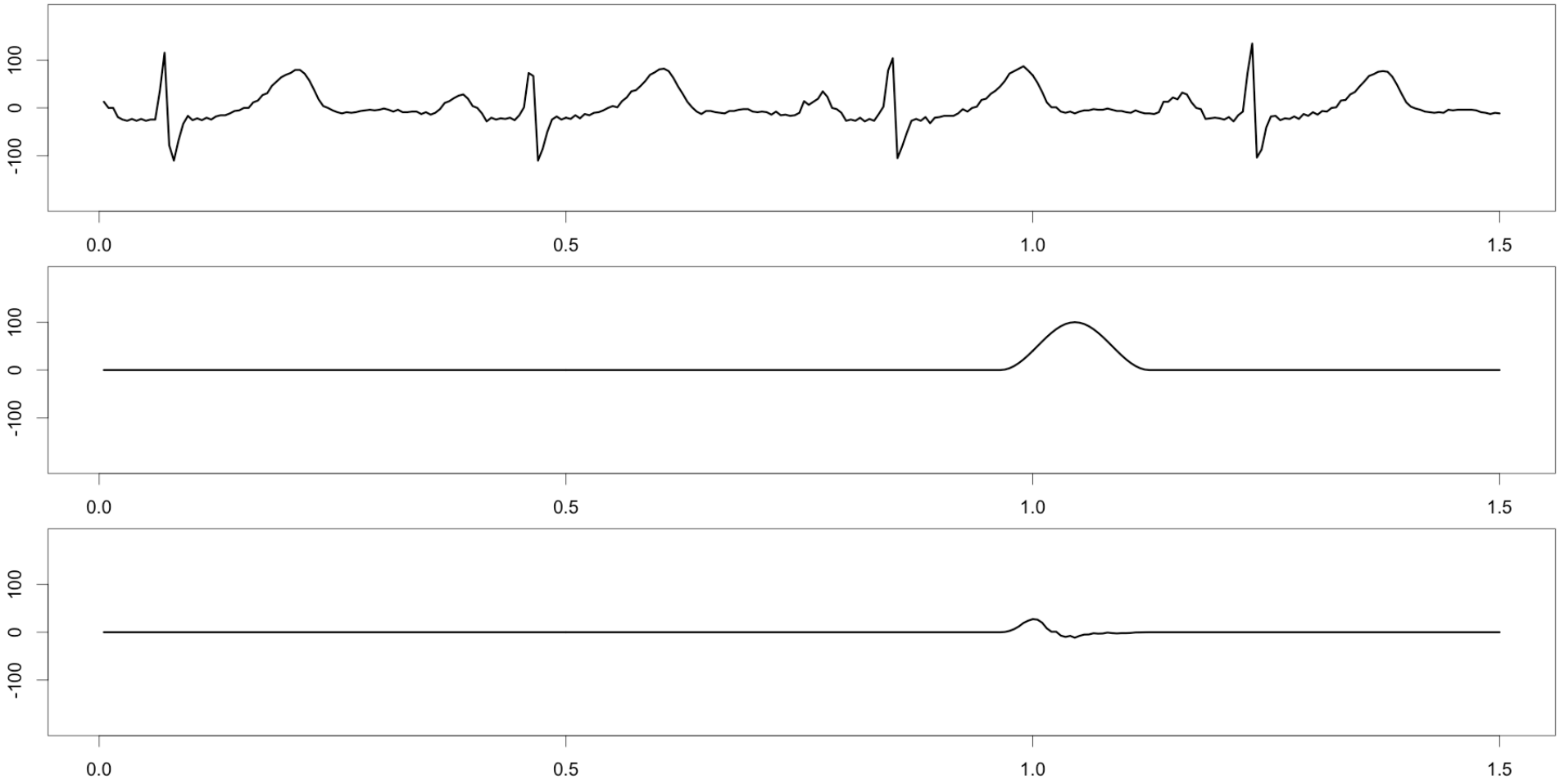
MAPR

We Do Windows



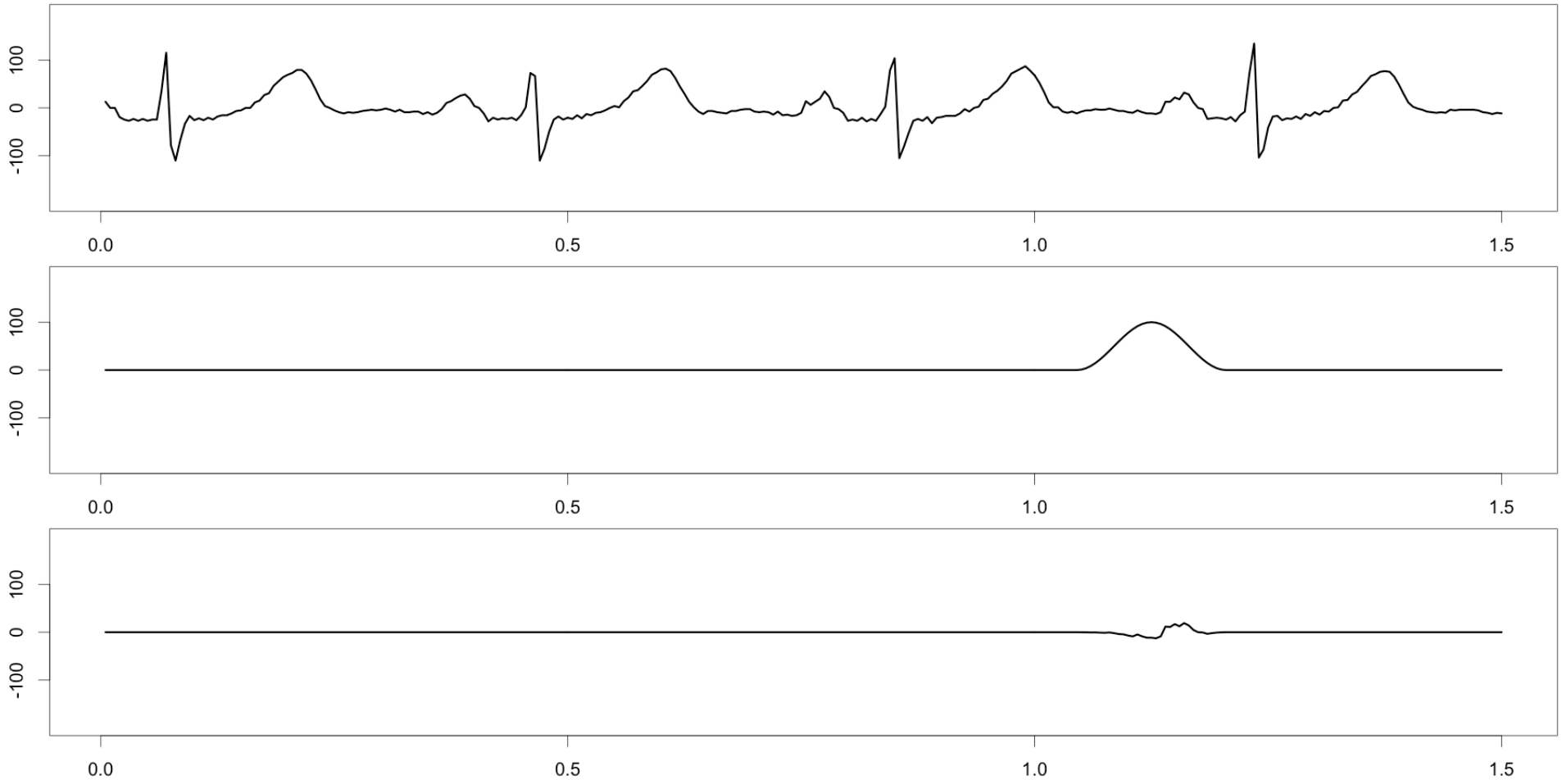
MAPR

We Do Windows



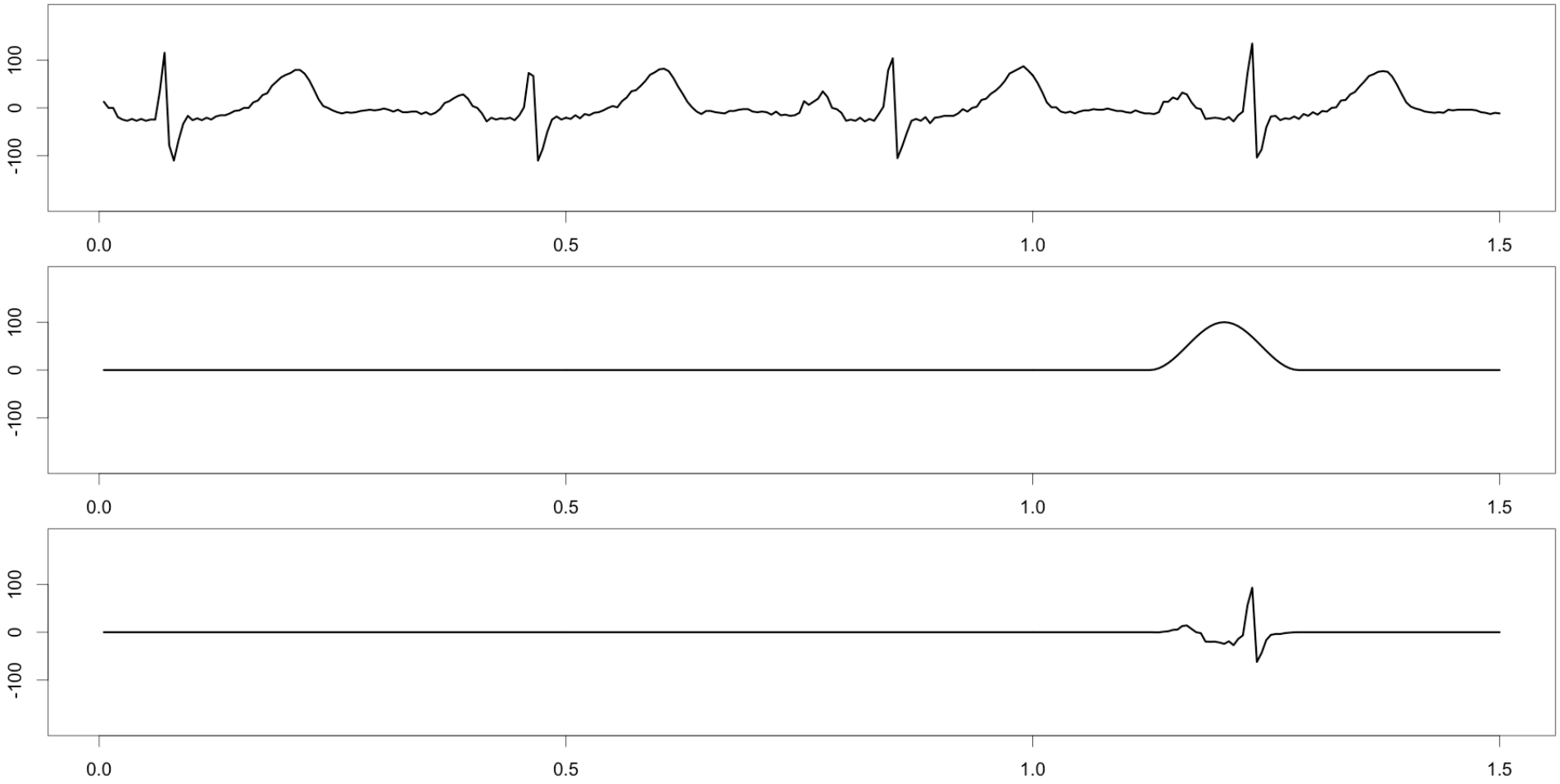
MAPR

We Do Windows



MAPR

We Do Windows



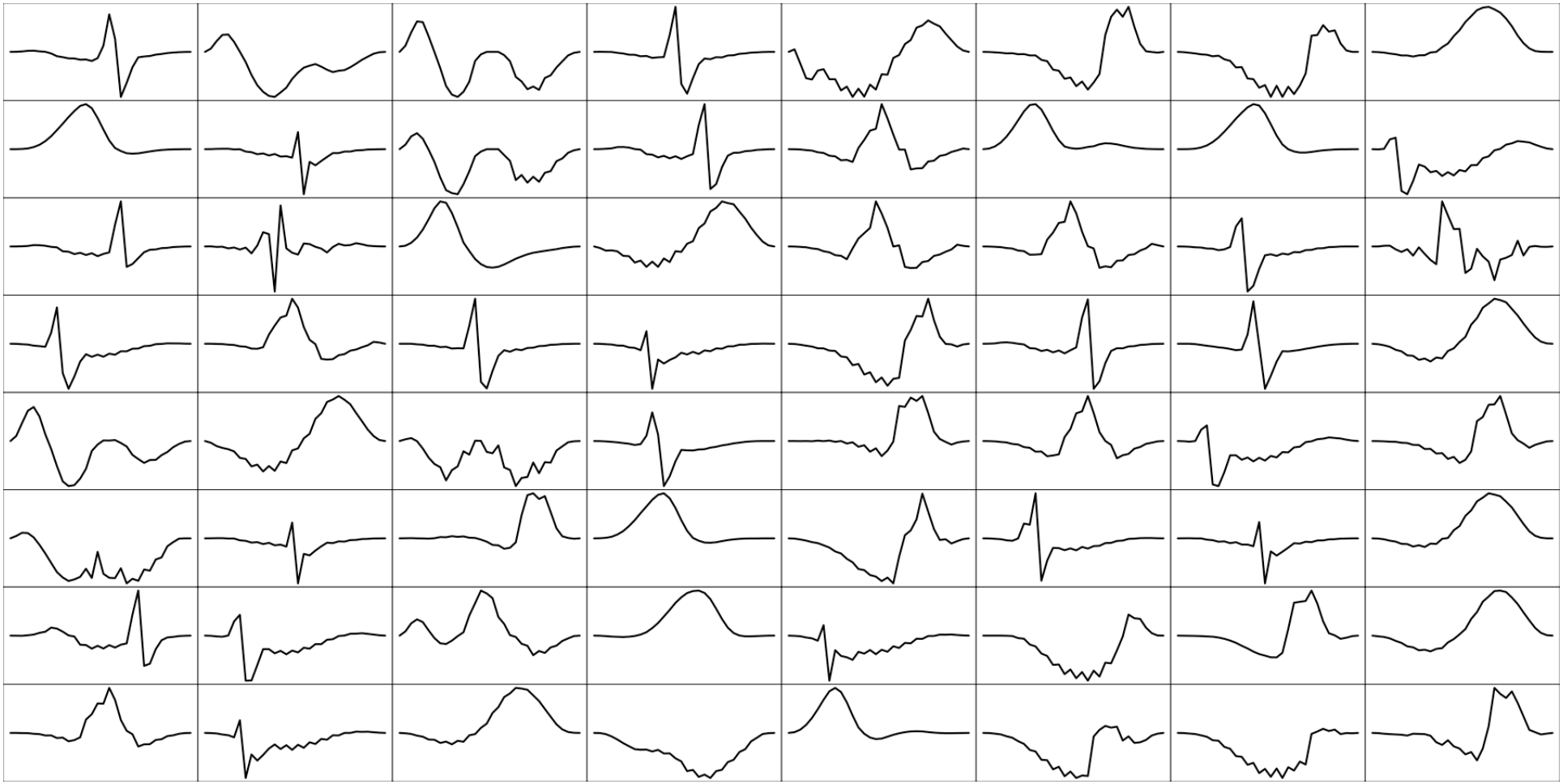
MAPR

Windows on the World

- The set of windowed signals is a nice model of our original signal
- Clustering can find the prototypes
  - Fancier techniques available using sparse coding

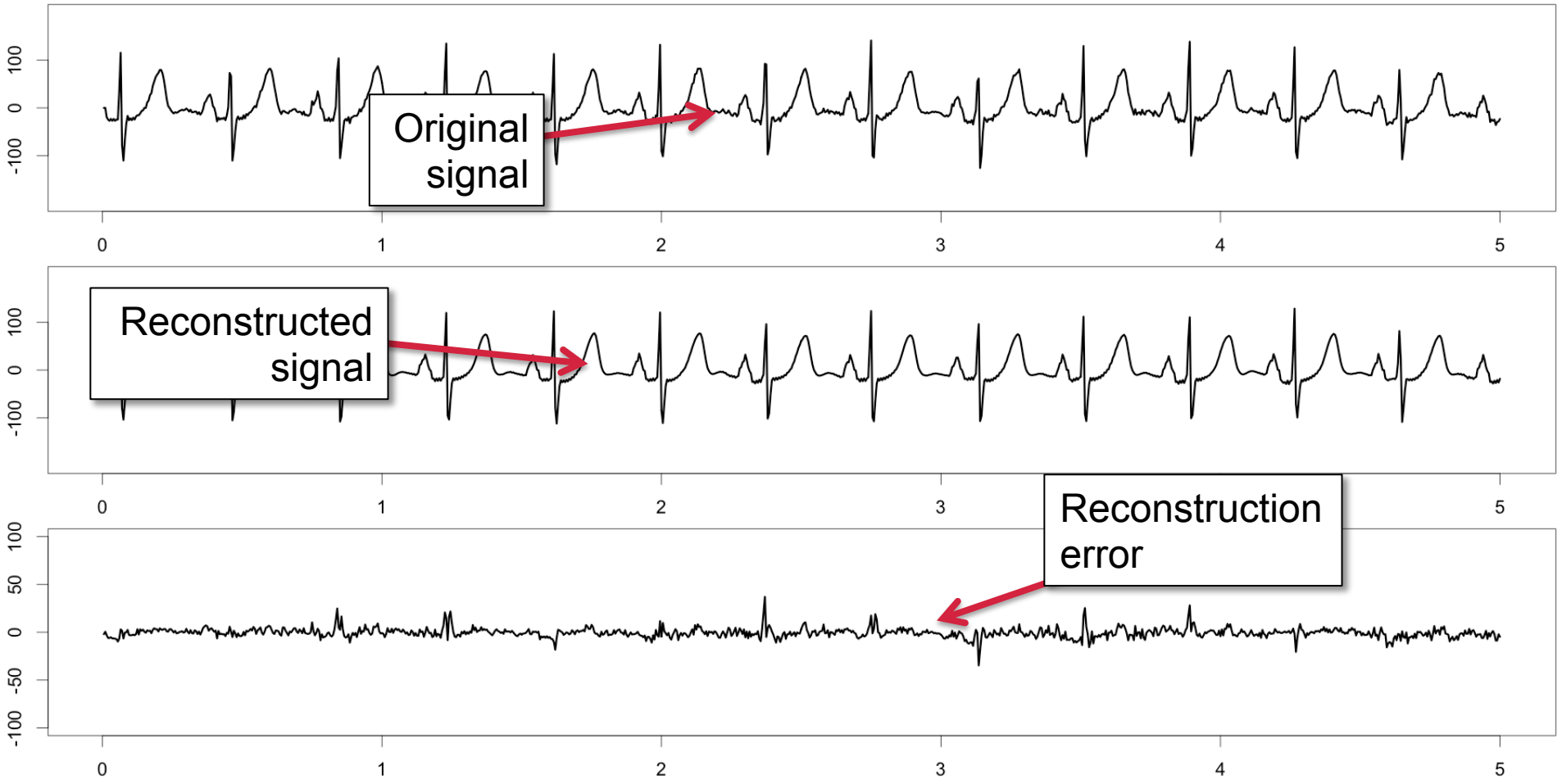
- The result is a dictionary of shapes
- New signals can be encoded by shifting, scaling and adding shapes from the dictionary

Most Common Shapes (for EKG)



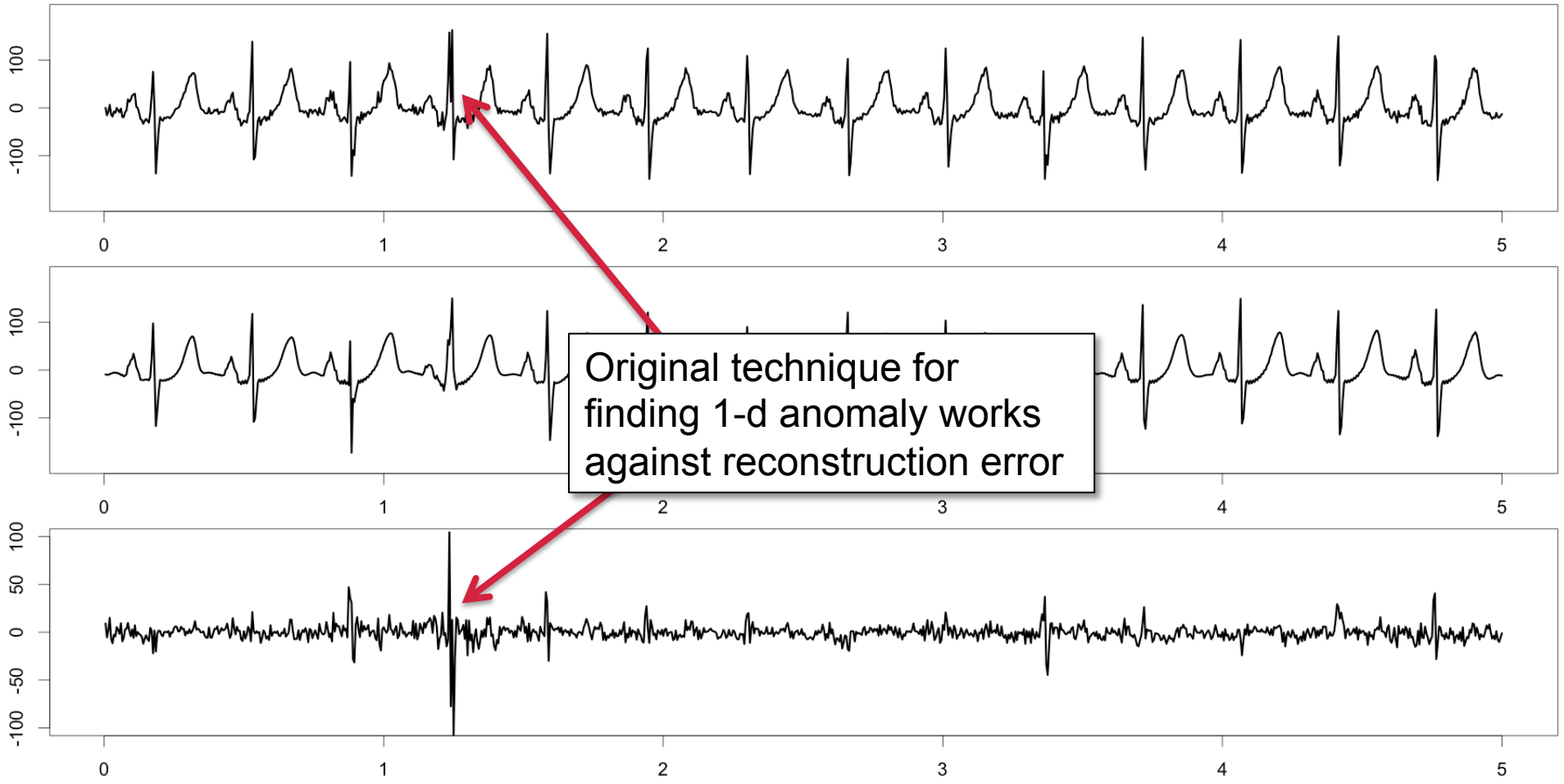
MAPR

Reconstructed signal



MAPR

An Anomaly

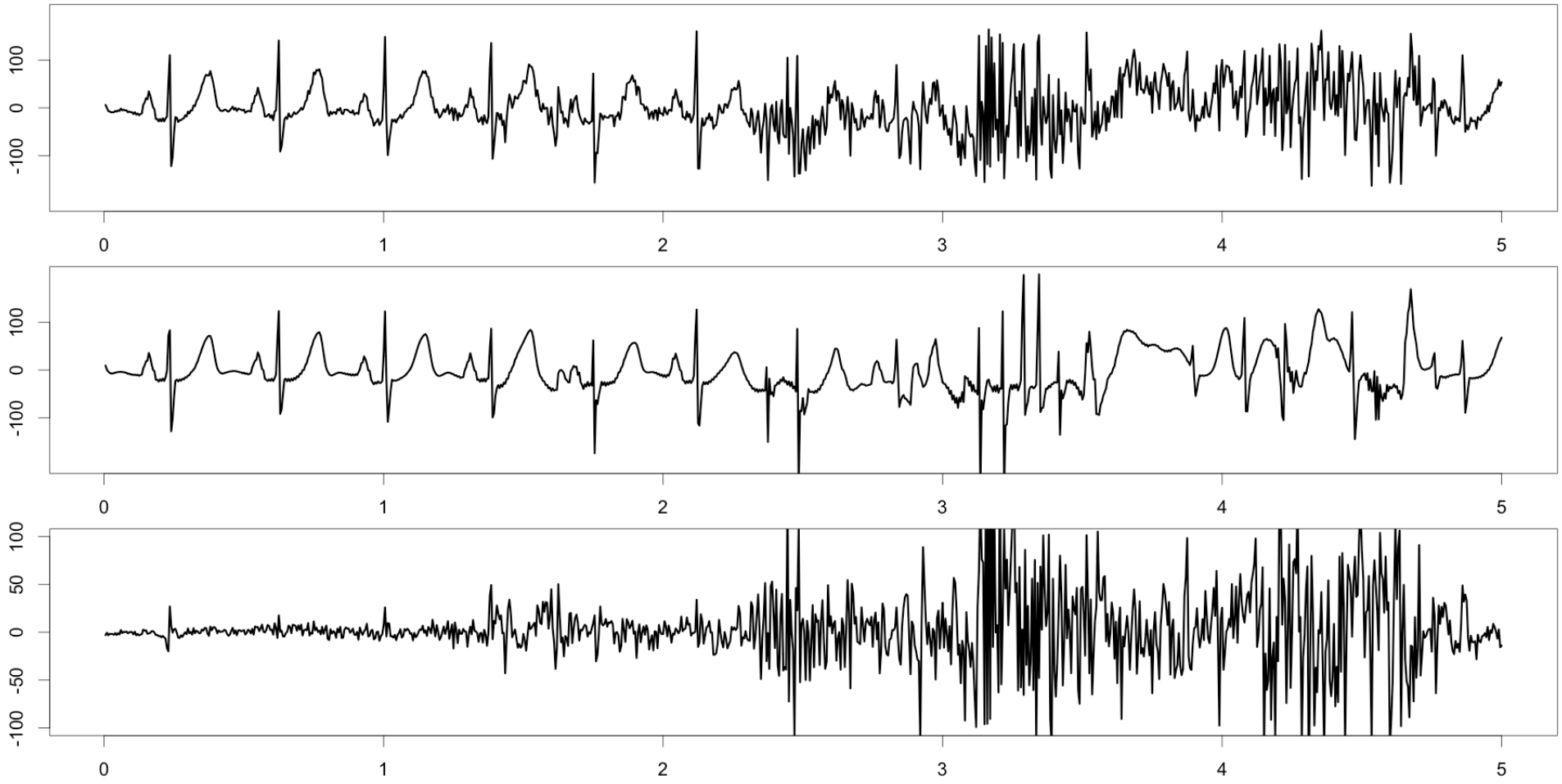


MAPR

Close-up of anomaly

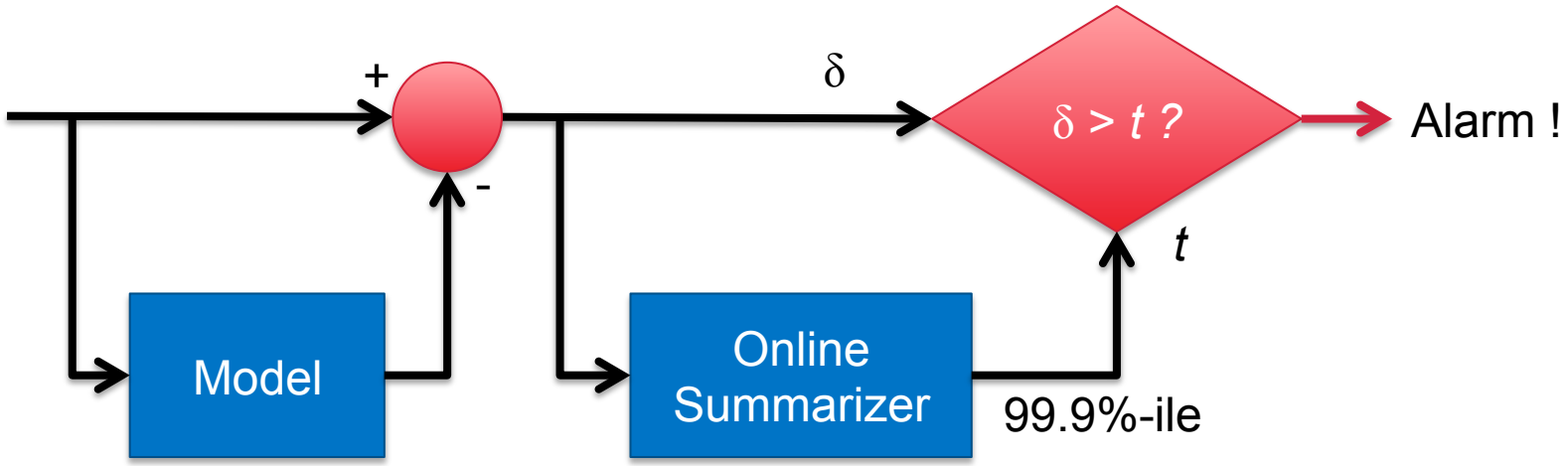
MAPR

A Different Kind of Anomaly



MAPR®

Model Delta Anomaly Detection



MAPR

The Real Inside Scoop

- The model-delta anomaly detector is really just a mixture distribution
  - the model we know about already
  - and a normally distributed error
- The output (delta) is (roughly) the log probability of the mixture distribution
- Thinking about probability distributions is good

Example: Event Stream (timing)

- Events of various types arrive at irregular intervals
  - we can assume Poisson distribution
- The key question is whether frequency has changed relative to expected values
- Want alert as soon as possible

Poisson Distribution

- Time between events is exponentially distributed

$$\Delta t \sim \lambda e^{-\lambda t}$$

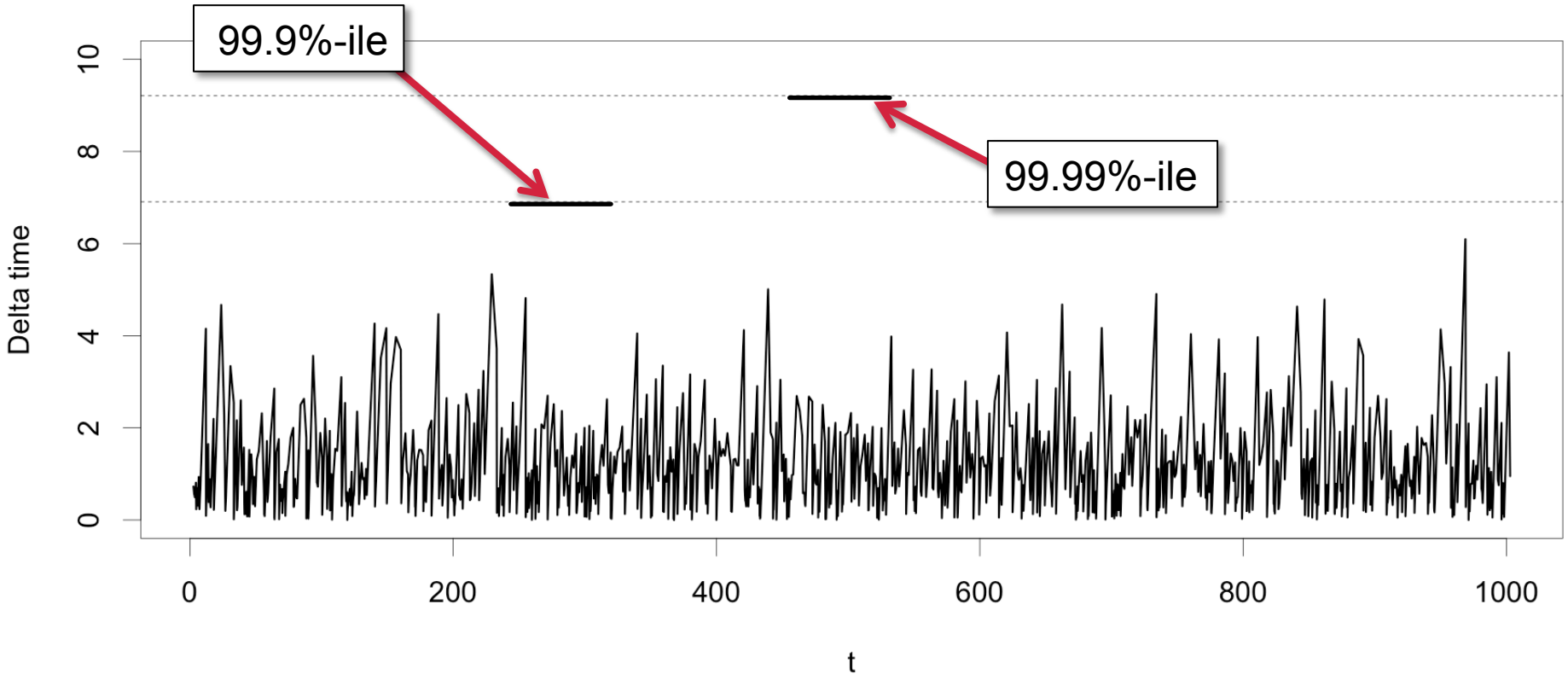
- This means that long delays are exponentially rare

$$P(\Delta t > T) = e^{-\lambda T}$$

$$-\log P(\Delta t > T) = \lambda T$$

- If we know λ we can select a good threshold
  - or we can pick a threshold empirically

Converting Event Times to Anomaly

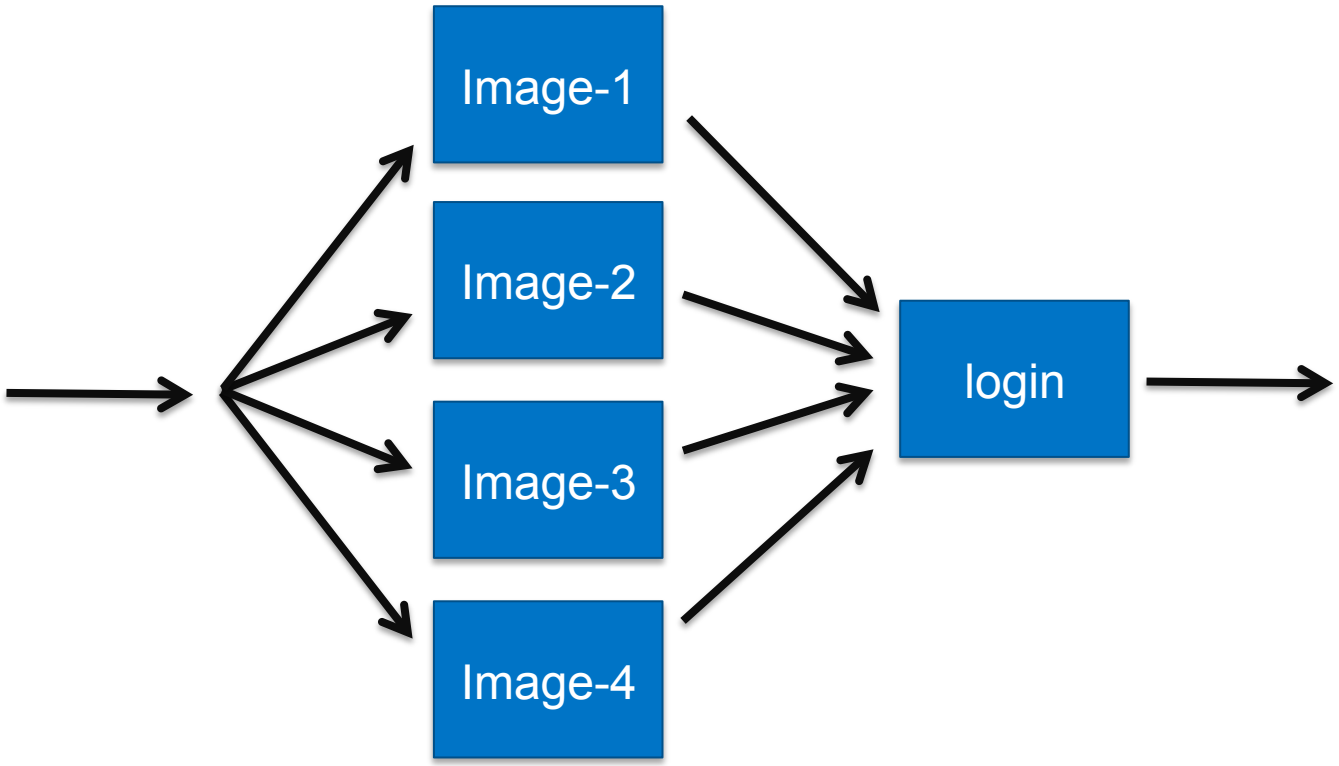


MAPR

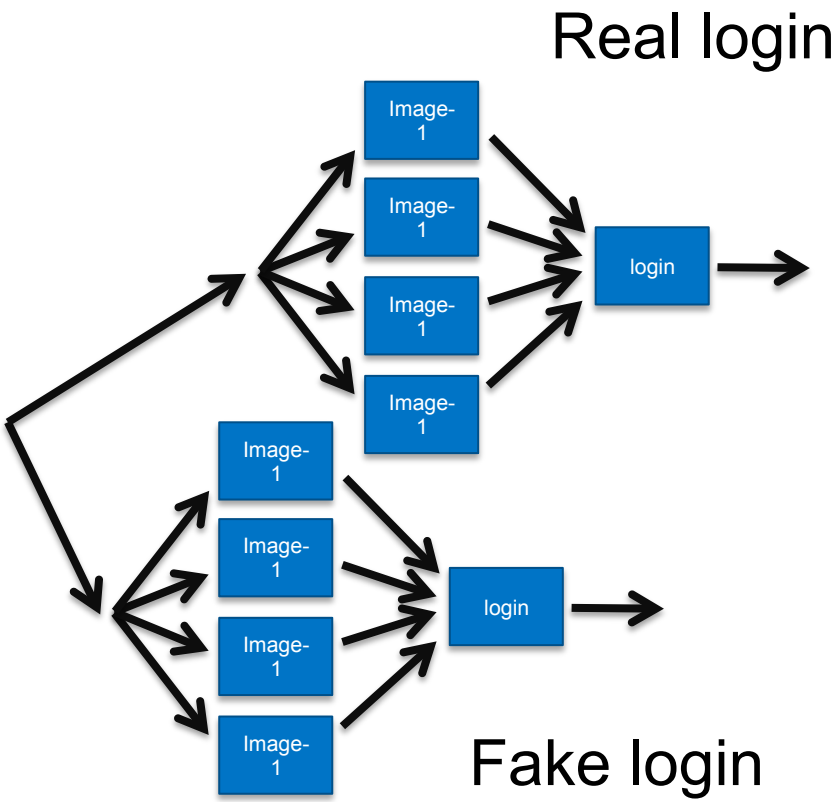
Example: Event Stream (sequence)

- The scenario:
  - Web visitors are being subjected to phishing attacks
  - The hook directs the visitors to a mocked login
  - When they enter their credentials, the attacker uses them to log in
  - We assume captcha's or similar are part of the authentication so the attacker has to show the images on the login page to the visitor
- The problem:
  - We don't really know how the attack works

Normal Event Flow



Phishing Flow



MAPR

Key Observations

- Regardless of exact details, there are patterns
- Event stream per user shows these patterns

- Phishing will have different patterns at much lower rate

- Measuring statistical surprise gives a good anomaly (fraud or malfunction) indicator

Recap (out of order)

- Anomaly detection is best done with a probability model
- Deep learning is a neat way to build this model
  - converting to symbolic dynamics simplifies life
- -log p is a good way to convert to anomaly measure
- Adaptive quantile estimation works for auto-setting thresholds

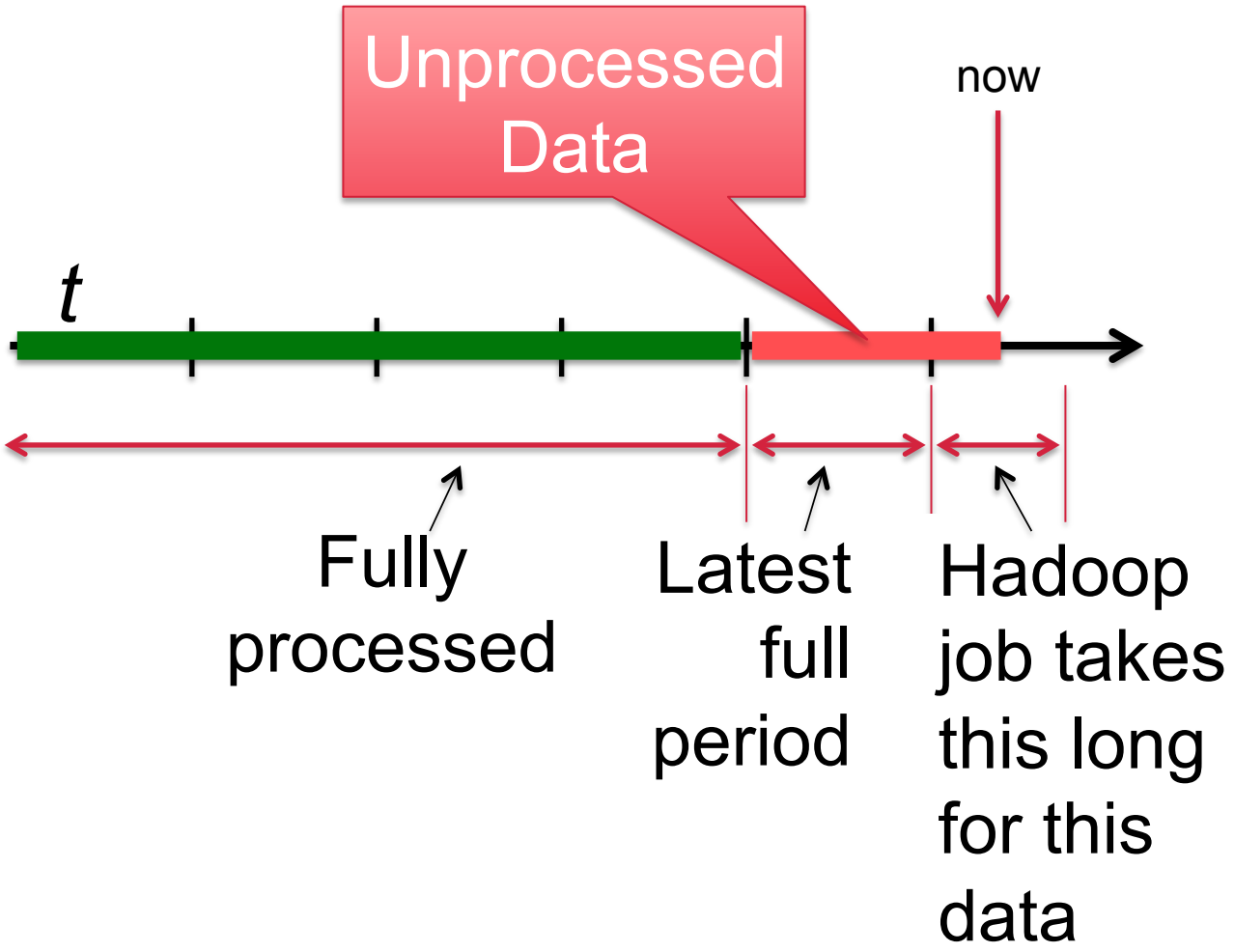
Recap

- Different systems require different models
- Continuous time-series
  - sparse coding or deep learning to build signal model
- Events in time
  - rate model base on variable rate Poisson
  - segregated rate model
- Events with labels
  - language modeling
  - hidden Markov models

How Do I Build Such a System

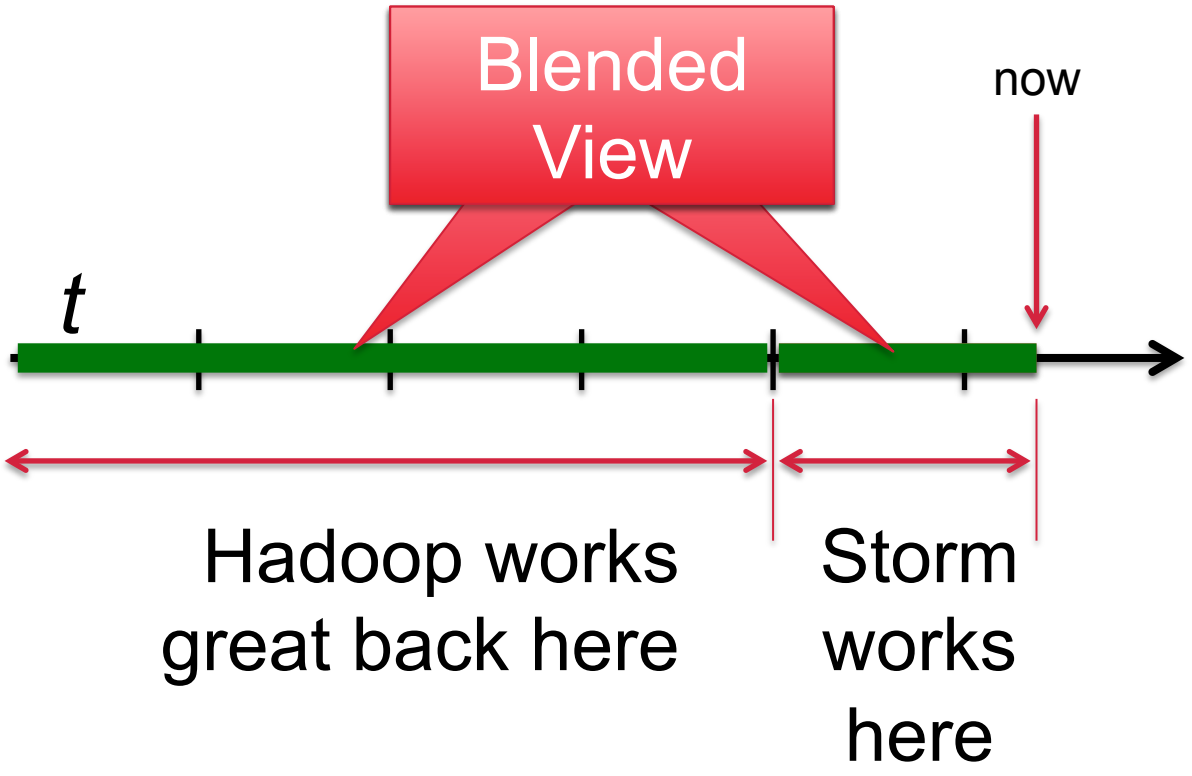
- The key is to combine real-time and long-time
  - real-time evaluates data stream against model
  - long-time is how we build the model
- Extended Lambda architecture is my favorite
- See my other talks on slideshare.net for info
- Ping me directly

Hadoop is Not Very Real-time



MAPR

Real-time and Long-time together



Who I am

- Ted Dunning, Chief Application Architect, MapR
  tdunning@mapr.com
  tdunning@apache.org
  @ted_dunning

- Committer, mentor, champion, PMC member on several Apache projects
- Mahout, Drill, Zookeeper others

MAPR®