High Performance Time Series Databases

© MapR Technologies, confidential

Agenda

- What is anomaly detection?
- Some examples
- Some generalization
- Compression == Truth
- Deep dive into deep learning
- Why this matters for time series databases

© MapR Technologies, confidential

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

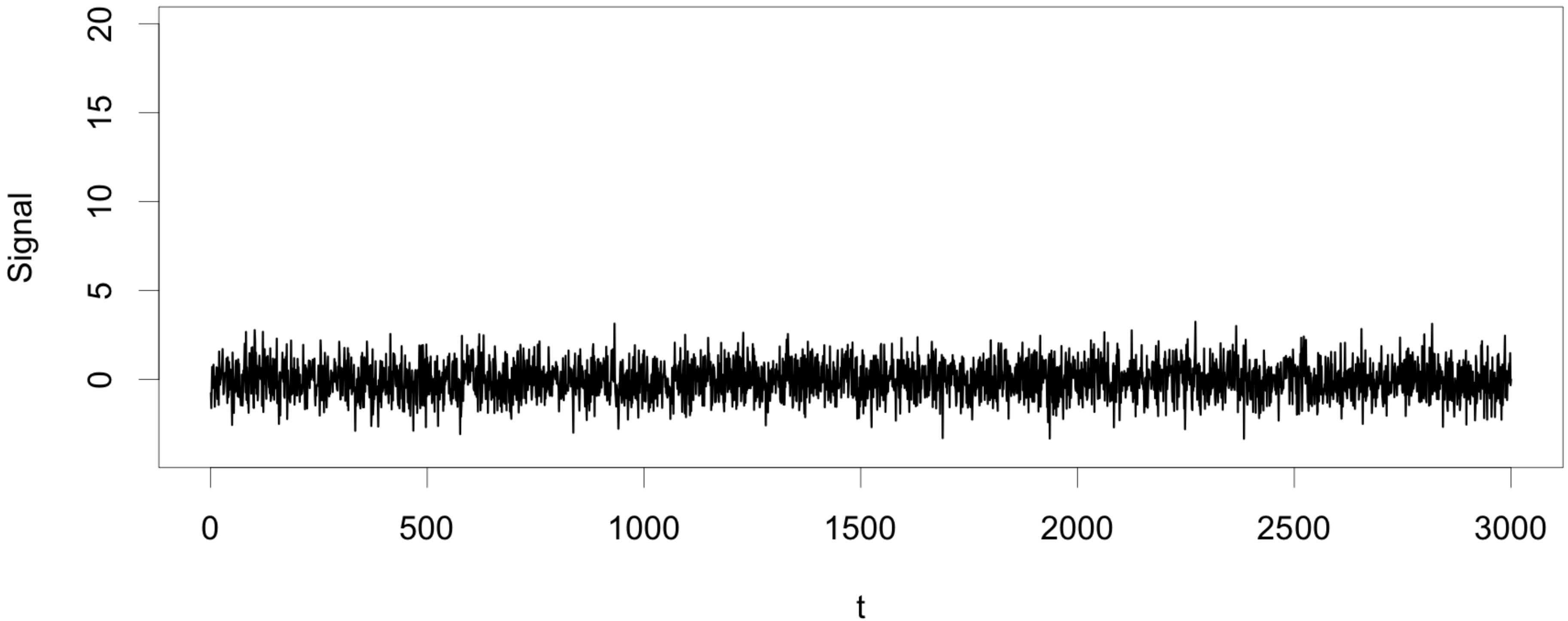
© MapR Technologies, confidential

Who we are

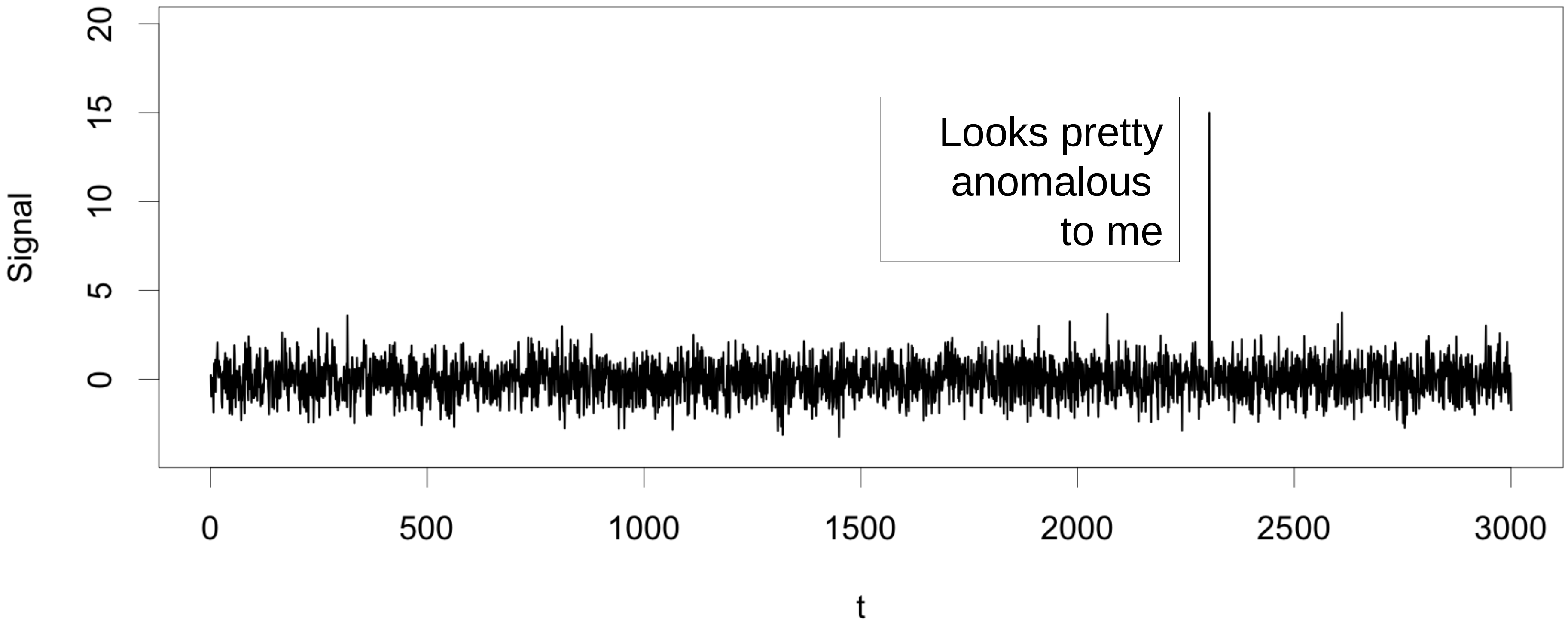
- MapR makes the technology leading distribution including Hadoop
- MapR integrates real-time data semantics directly into a system that also runs Hadoop programs seamlessly
- The biggest and best choose MapR
  - Google, Amazon
  - Largest credit card, retailer, health insurance, telco
  - Ping me for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

What is Anomaly Detection?

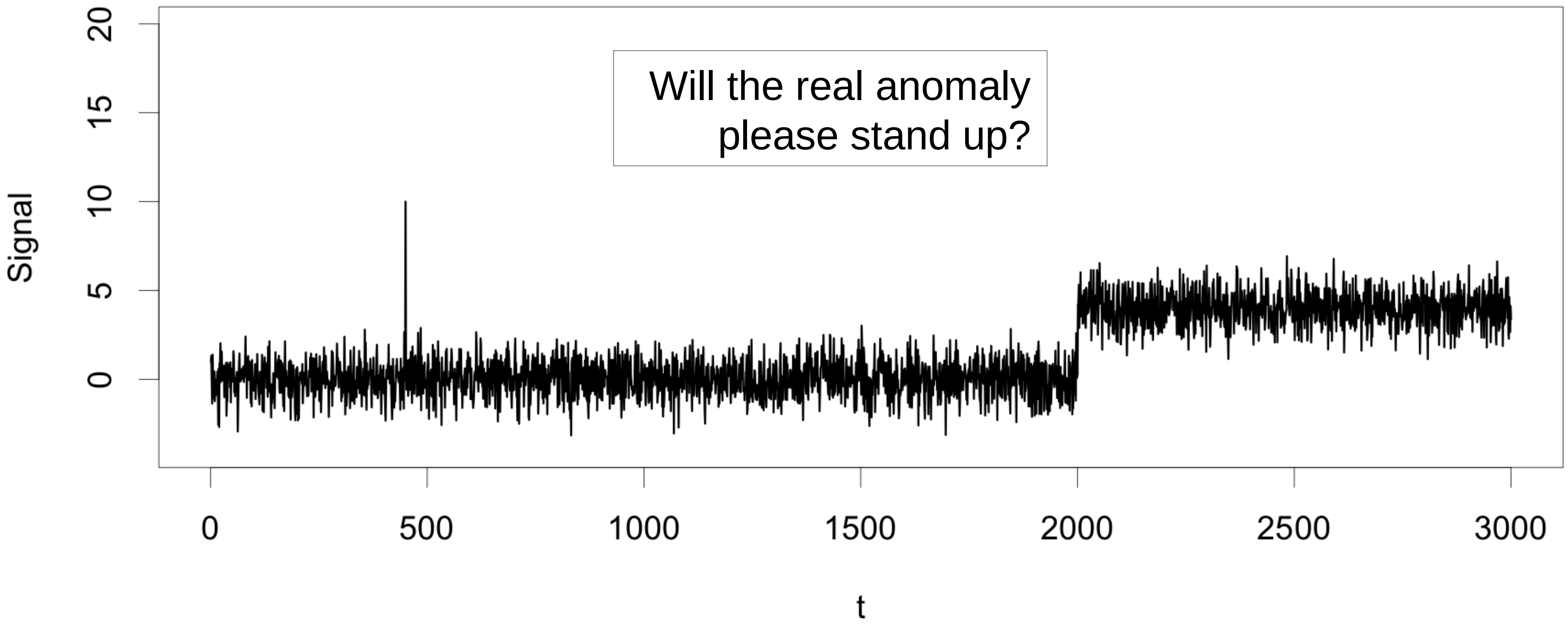
- What just happened that shouldn't?
  - but I don't know what failure looks like (yet)
- Find the problem before other people see it
  - especially customers and CEO's
- But don't wake me up if it isn't really broken



© MapR Technologies, confidential



© MapR Technologies, confidential



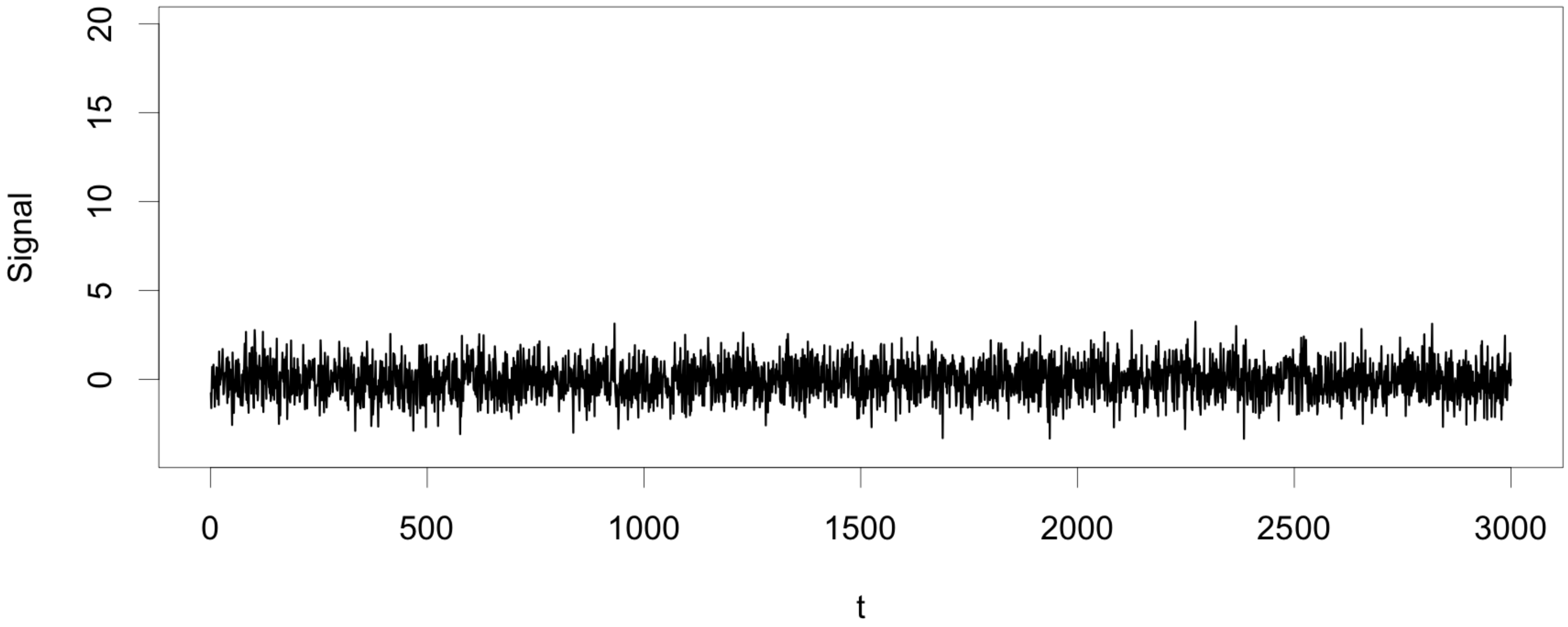
© MapR Technologies, confidential

What Are We Really Doing

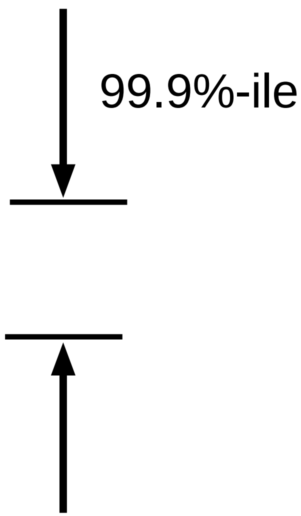
- We want action when something breaks
  (dies/falls over/otherwise gets in trouble)
- But action is expensive
- So we don't want false alarms
- And we don't want false negatives

- We need to trade off costs

A Second Look

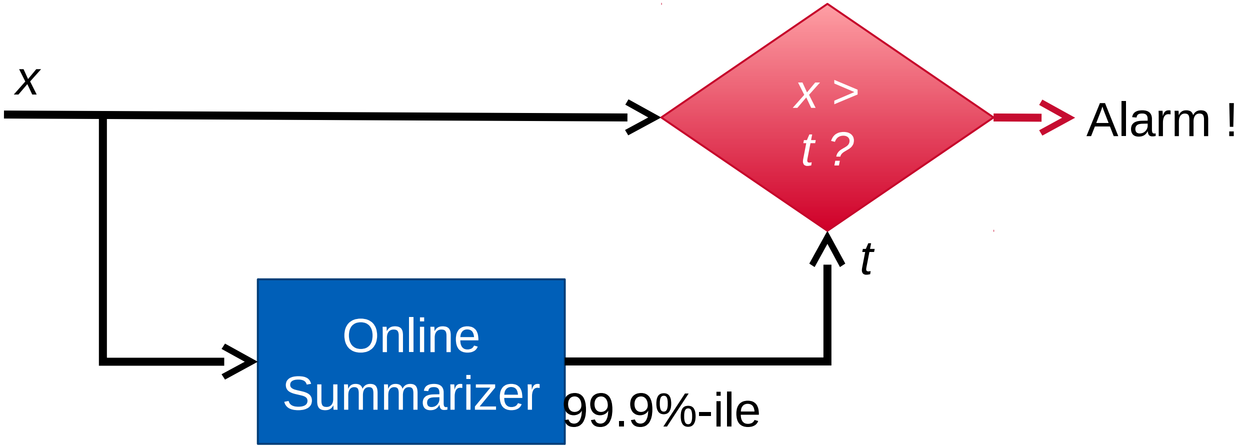


A Second Look



© MapR Technologies, confidential

How Hard Can it Be?



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?

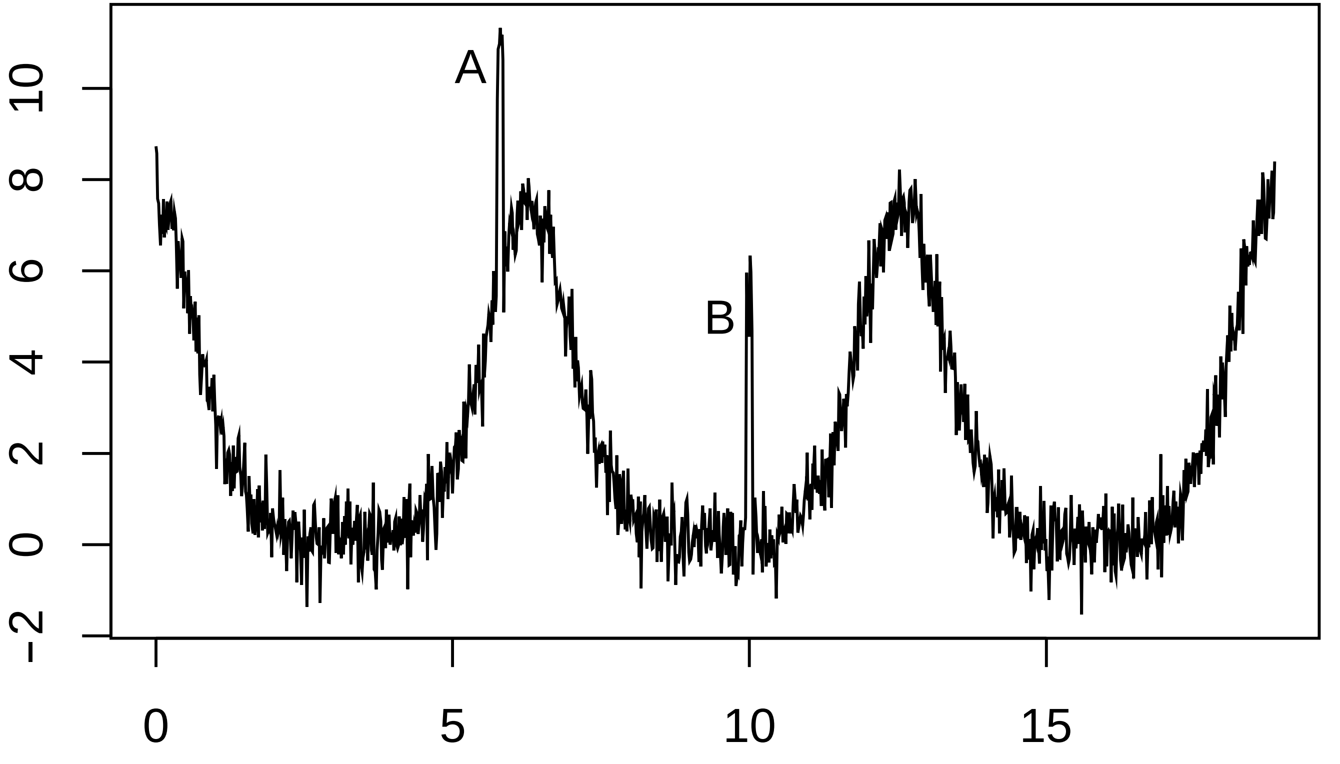
- This looks like a solved problem

© MapR Technologies, confidential

Already Done? Etsy Skyline?

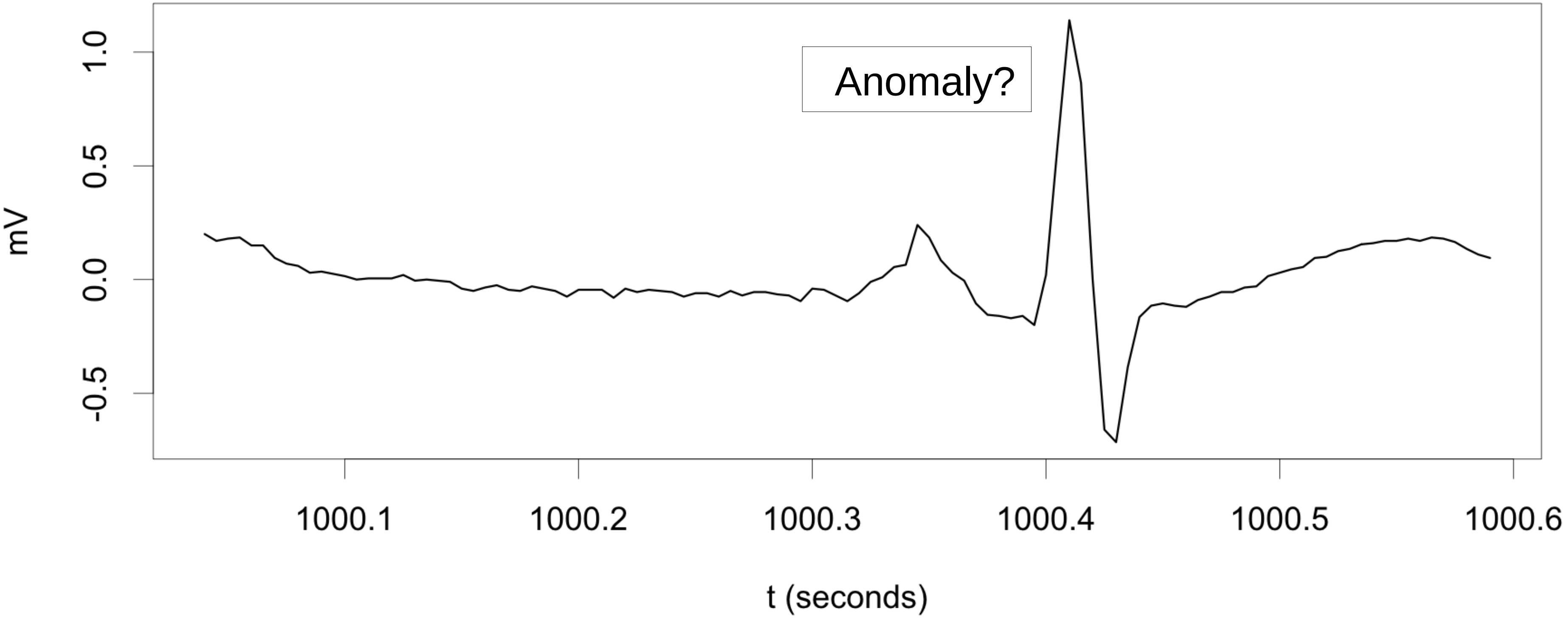
© MapR Technologies, confidential

What About This?



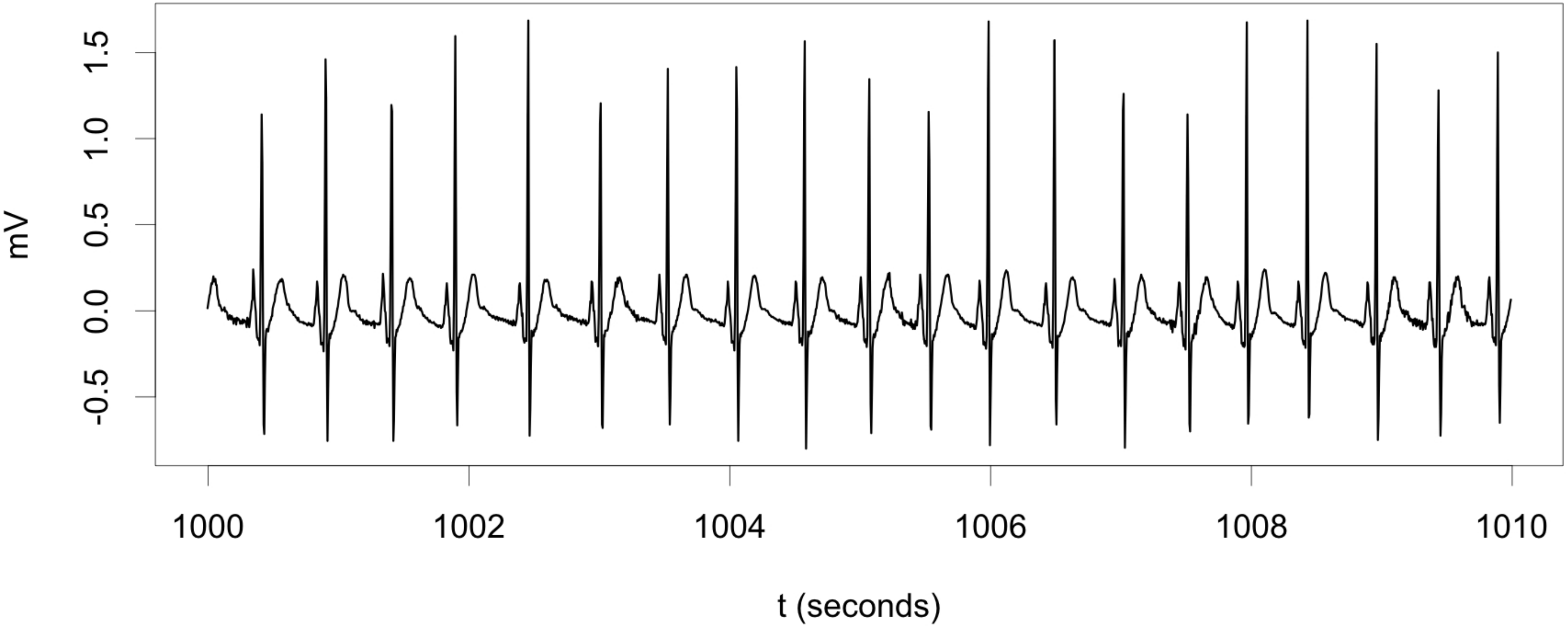
© MapR Technologies, confidential

Spot the Anomaly

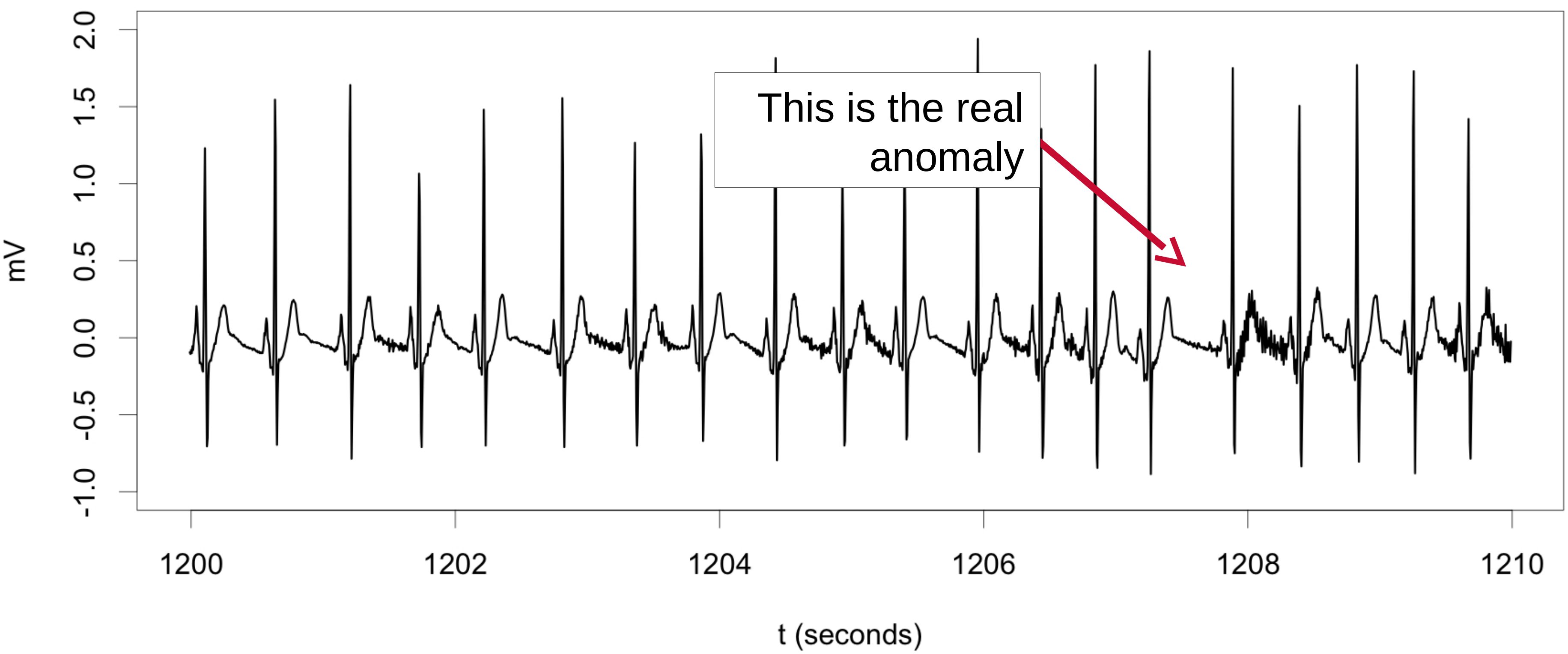


© MapR Technologies, confidential

Maybe not!



Where's Waldo?



Normal Isn't Just Normal

- What we want is a *model* of what is normal

- What doesn't fit the model is the *anomaly*

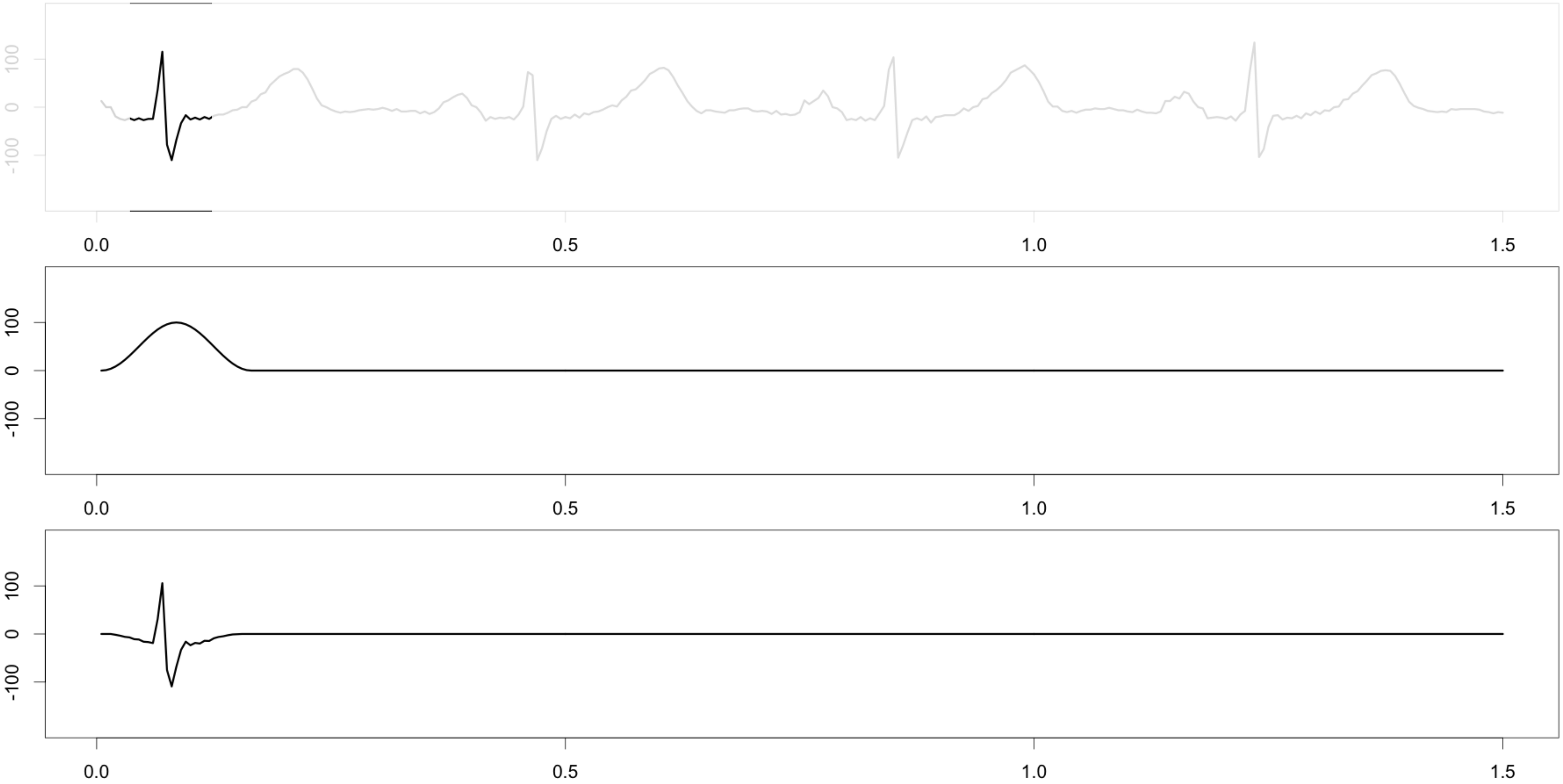
- For simple signals, the model can be simple …

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

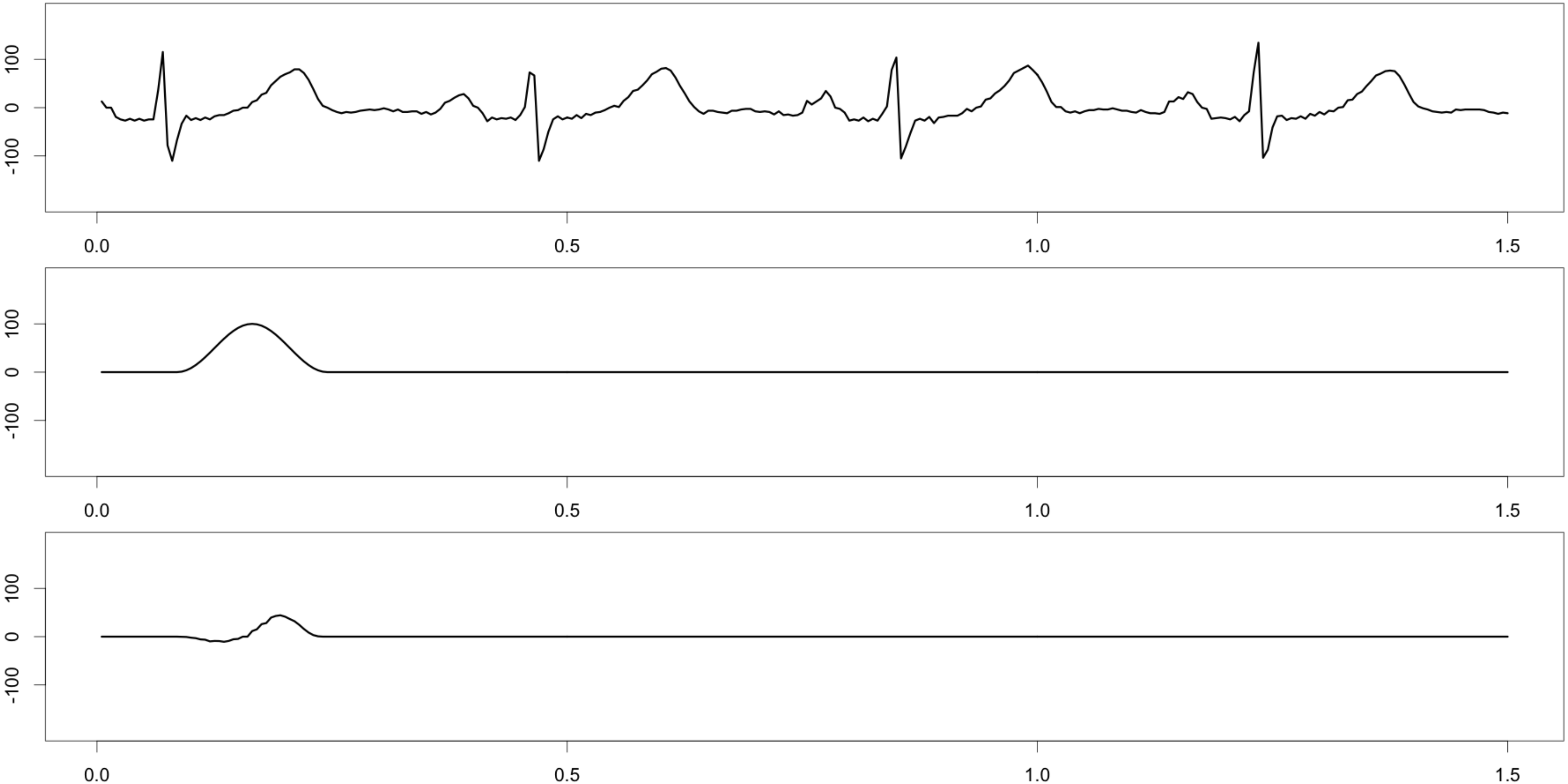
- The real world is rarely so accommodating

© MapR Technologies, confidential

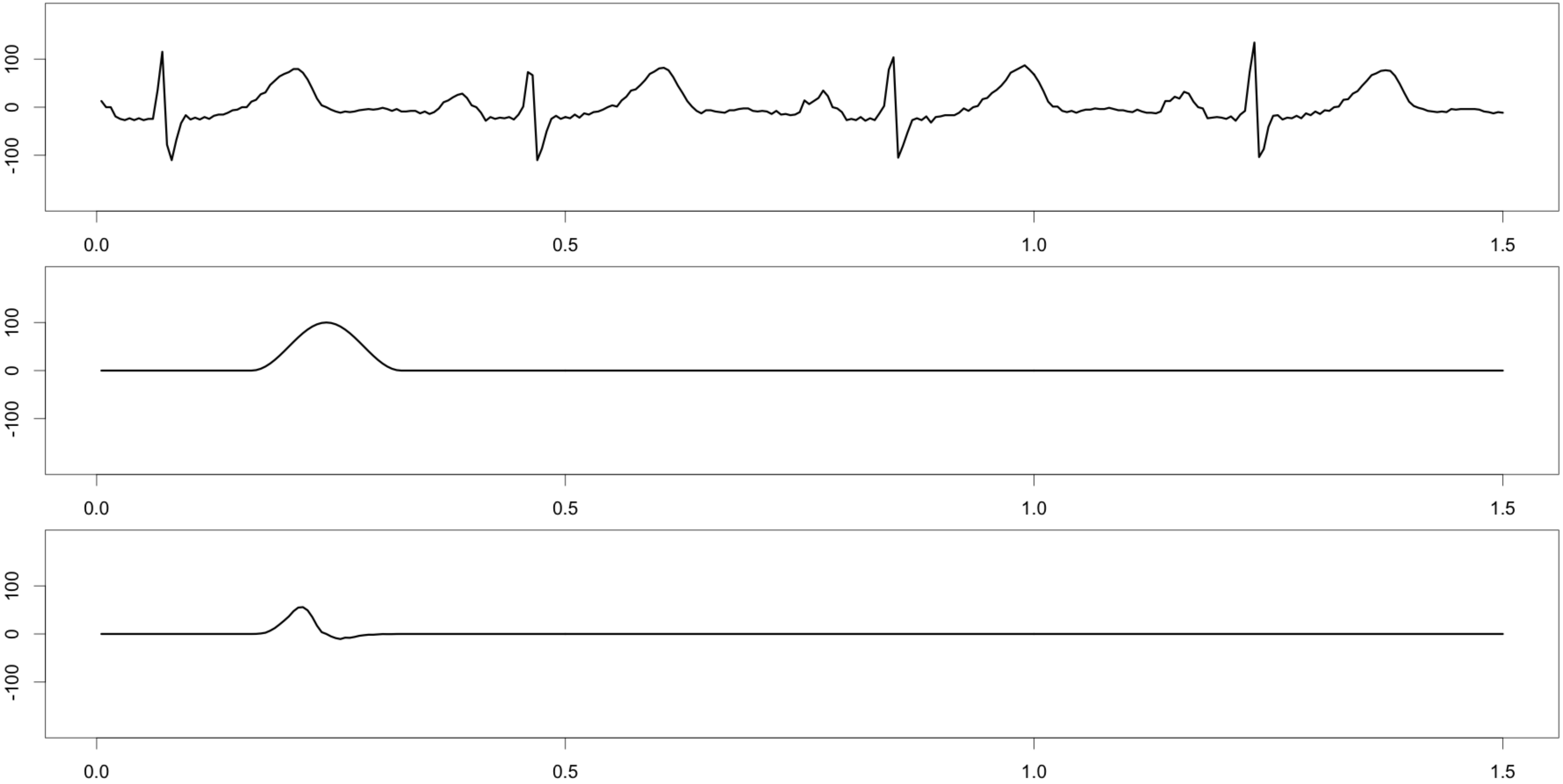
We Do Windows



We Do Windows

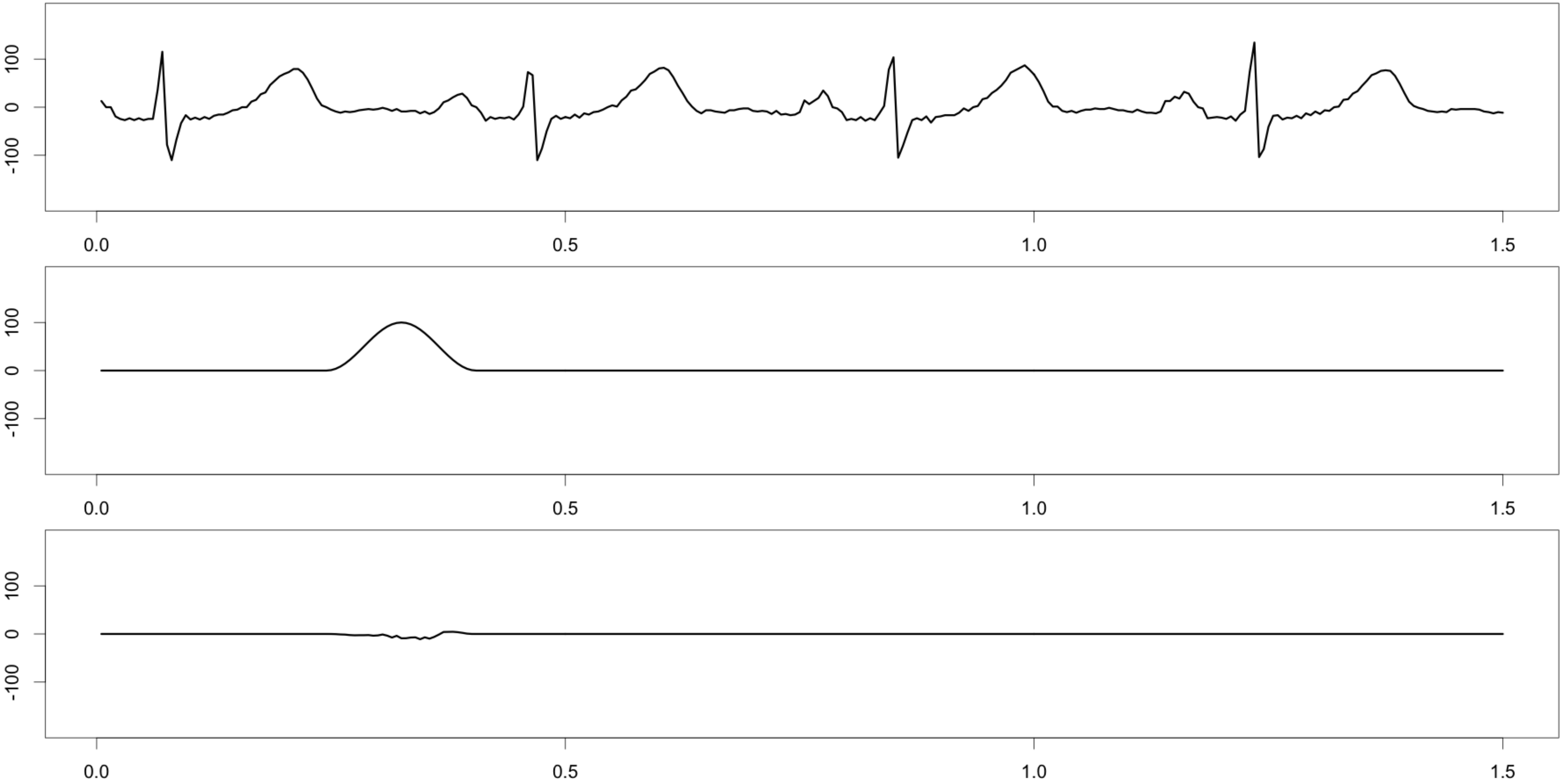


We Do Windows

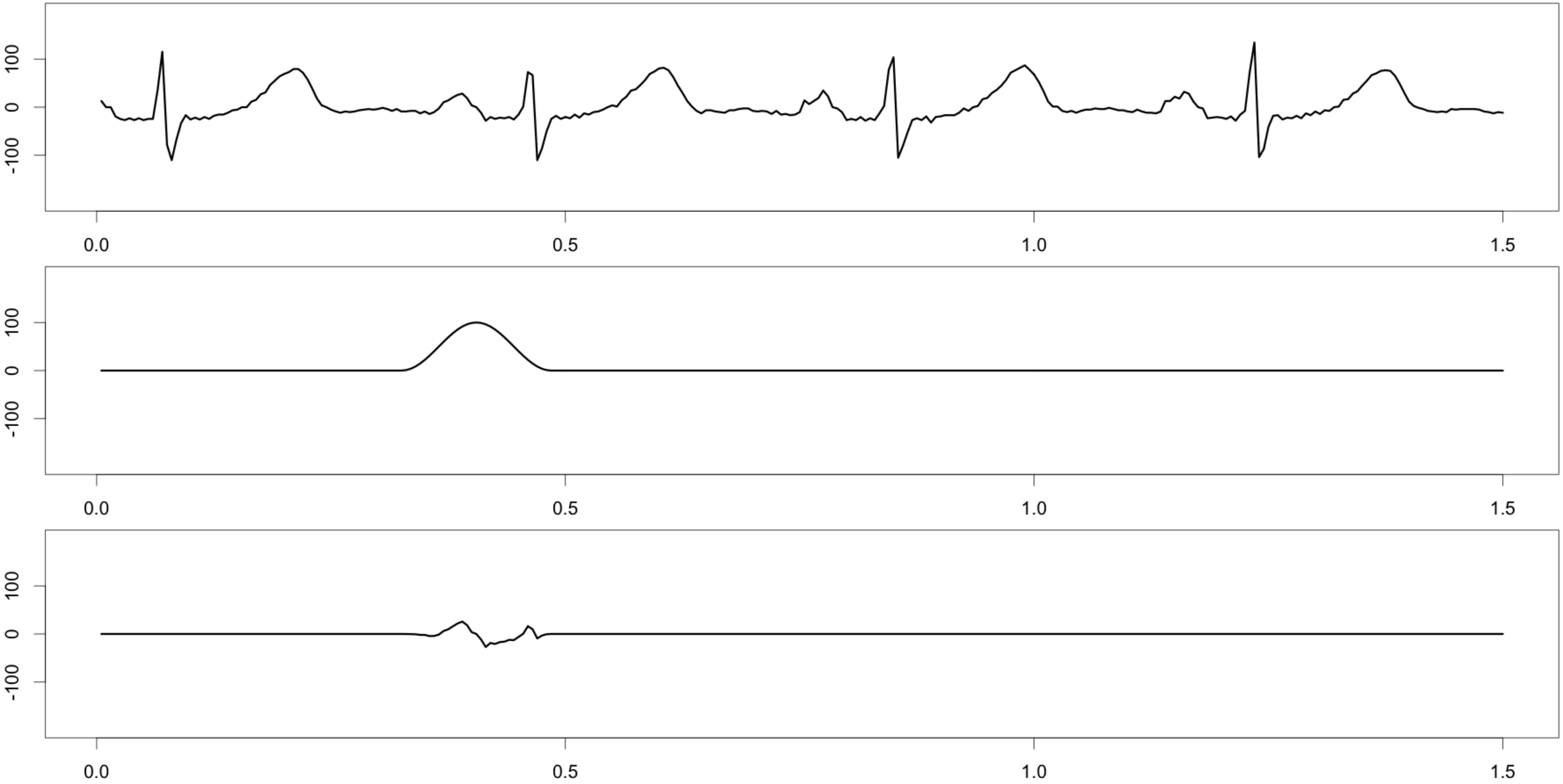


© MapR Technologies, confidential

We Do Windows

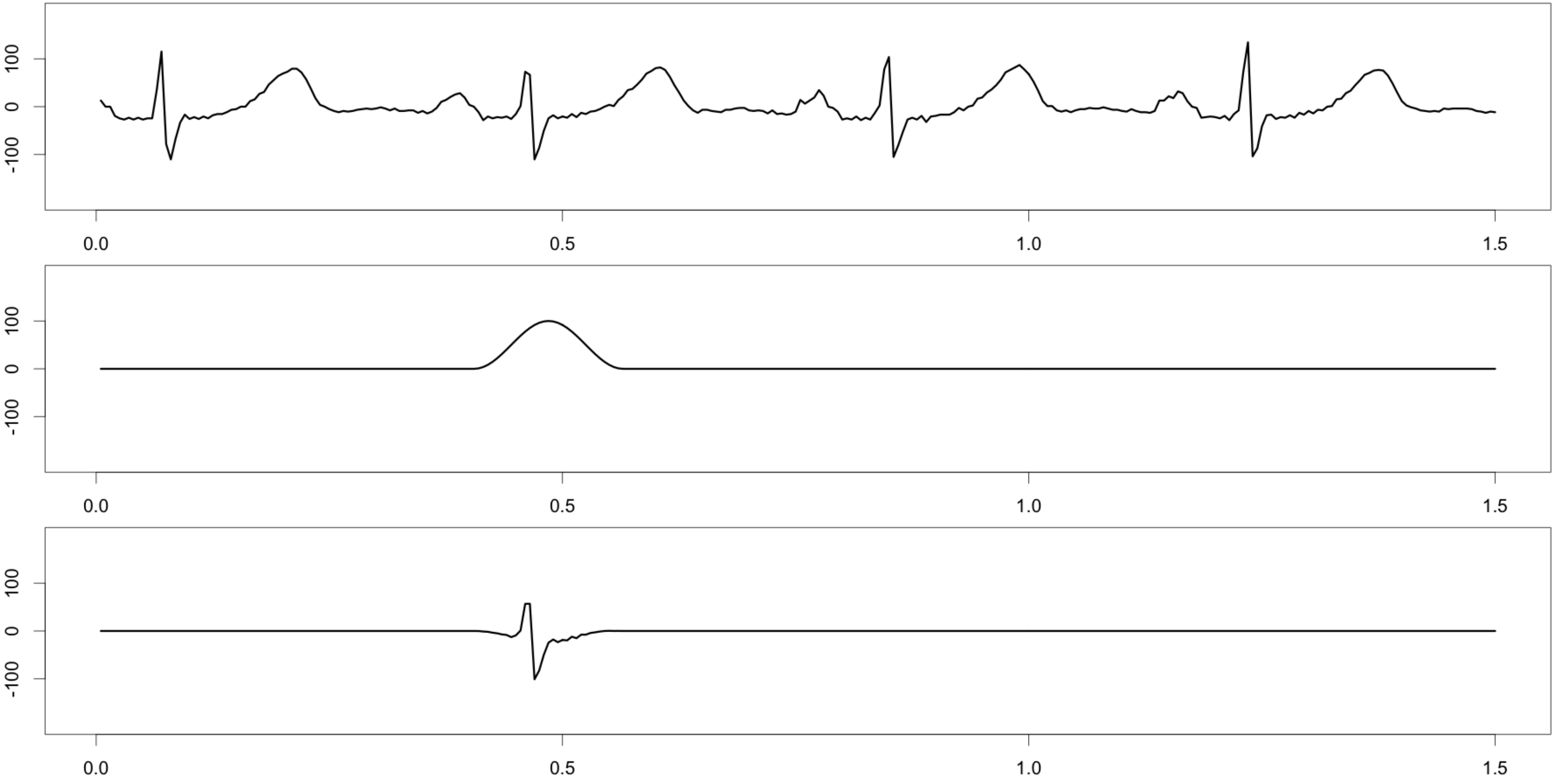


We Do Windows

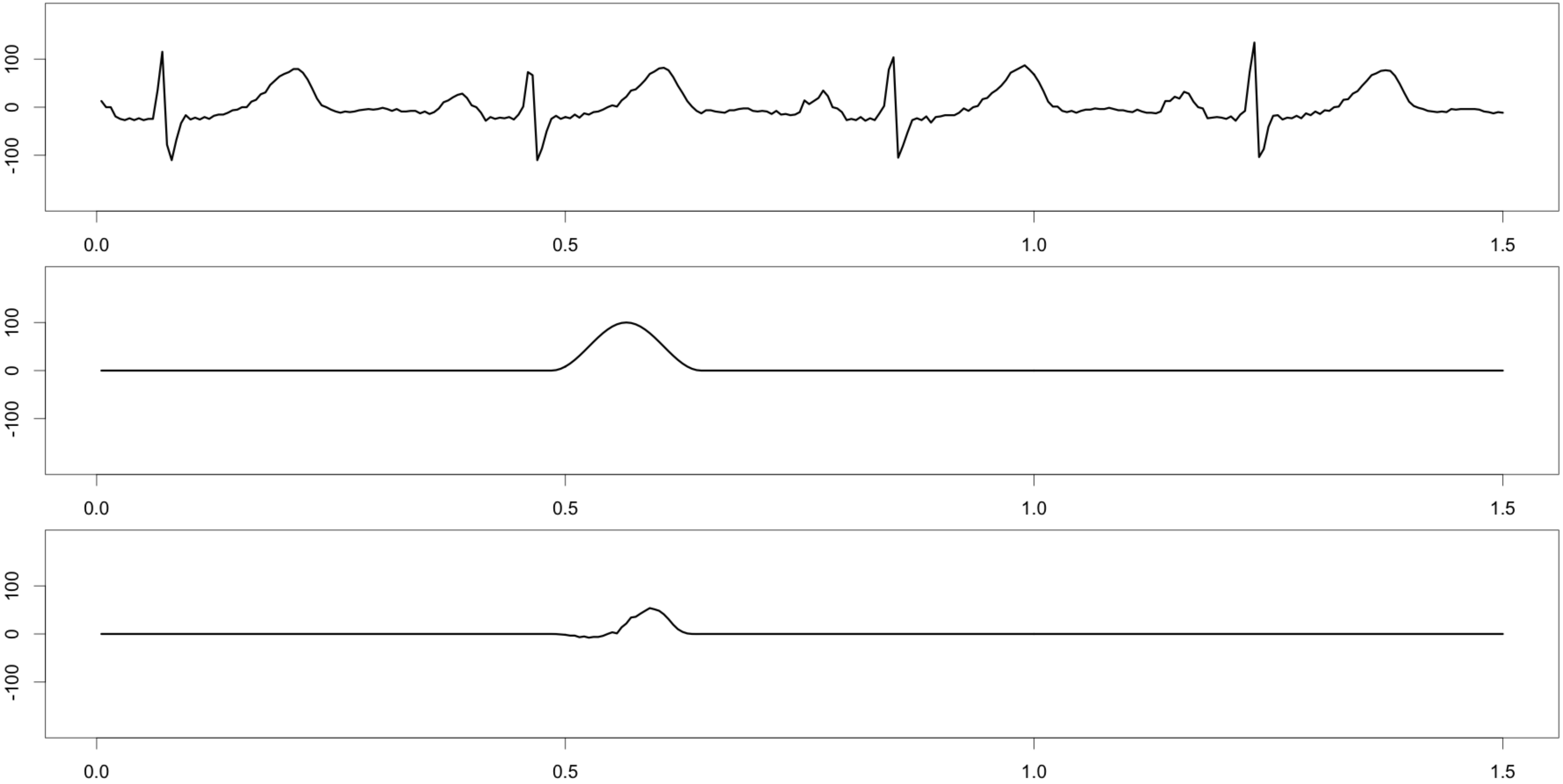


© MapR Technologies, confidential

We Do Windows

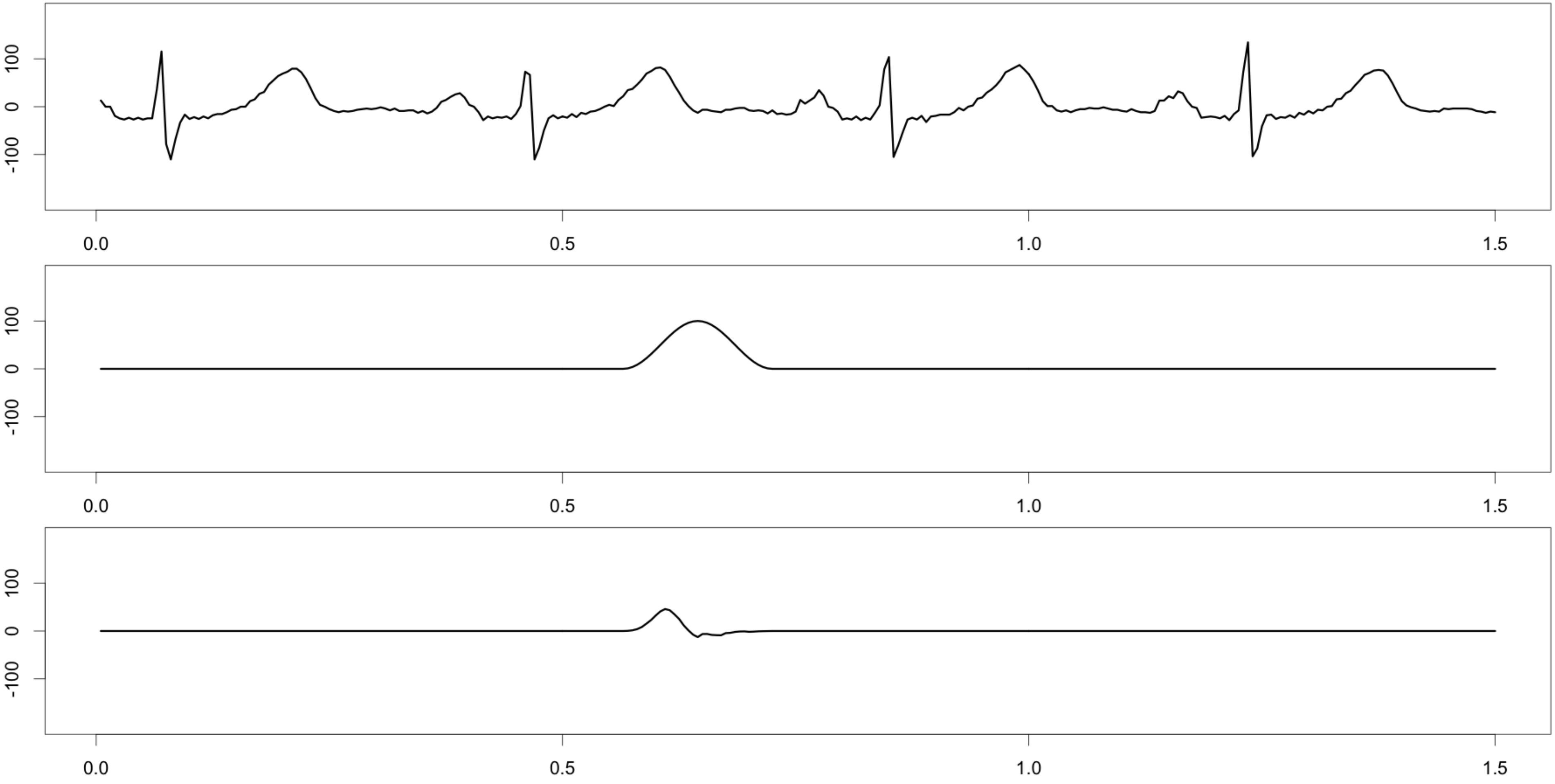


We Do Windows

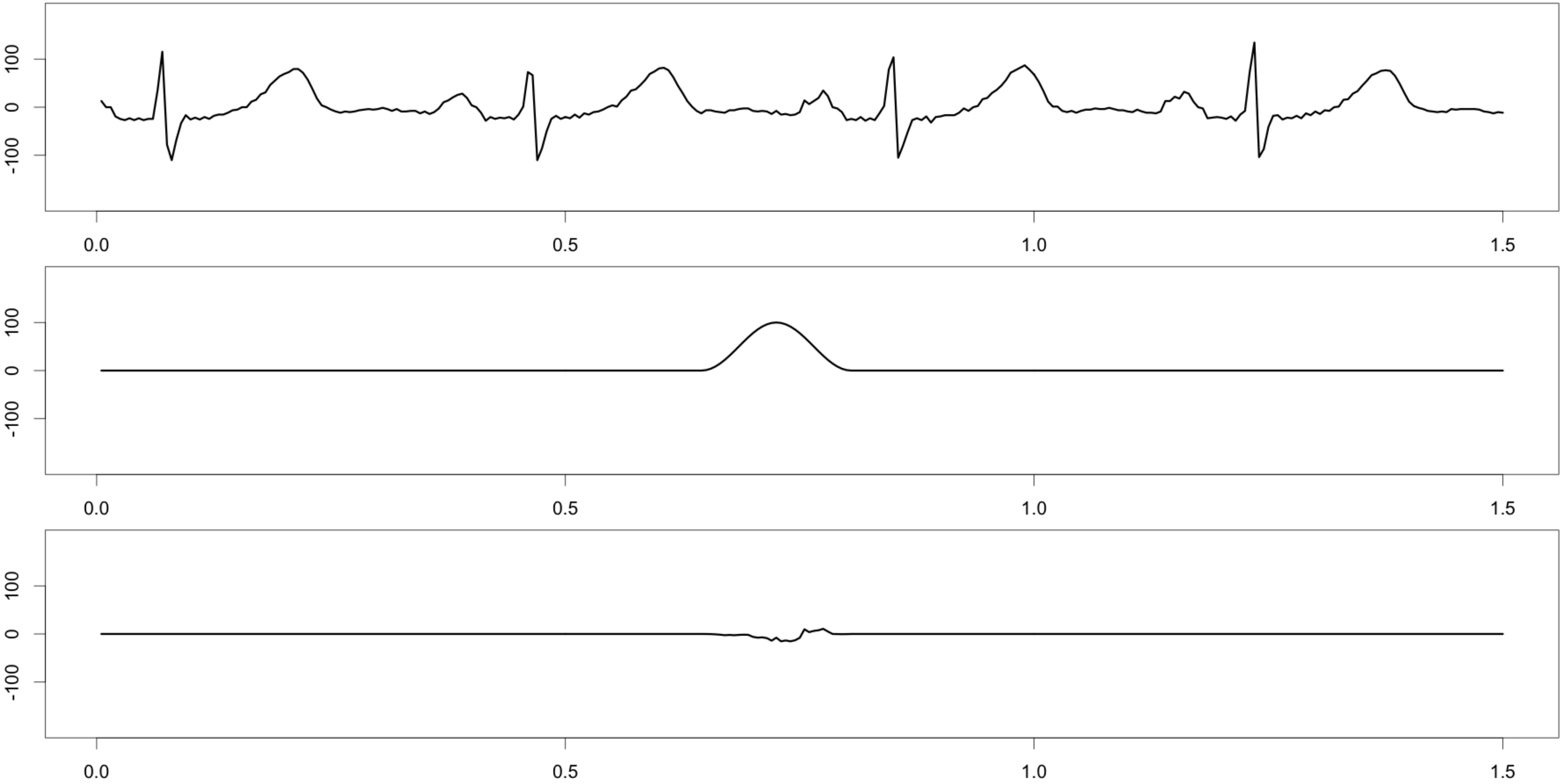


© MapR Technologies, confidential

We Do Windows

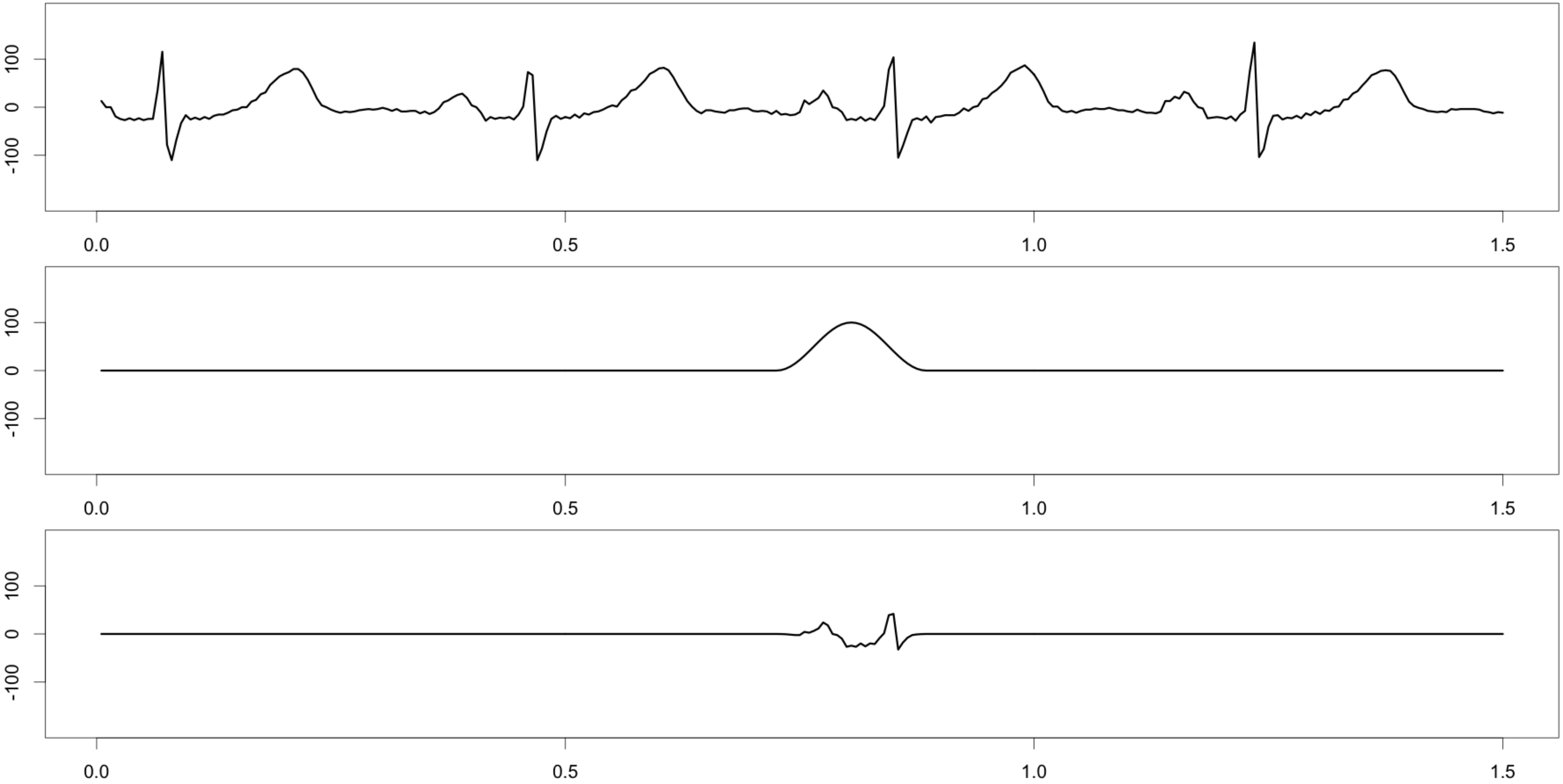


We Do Windows



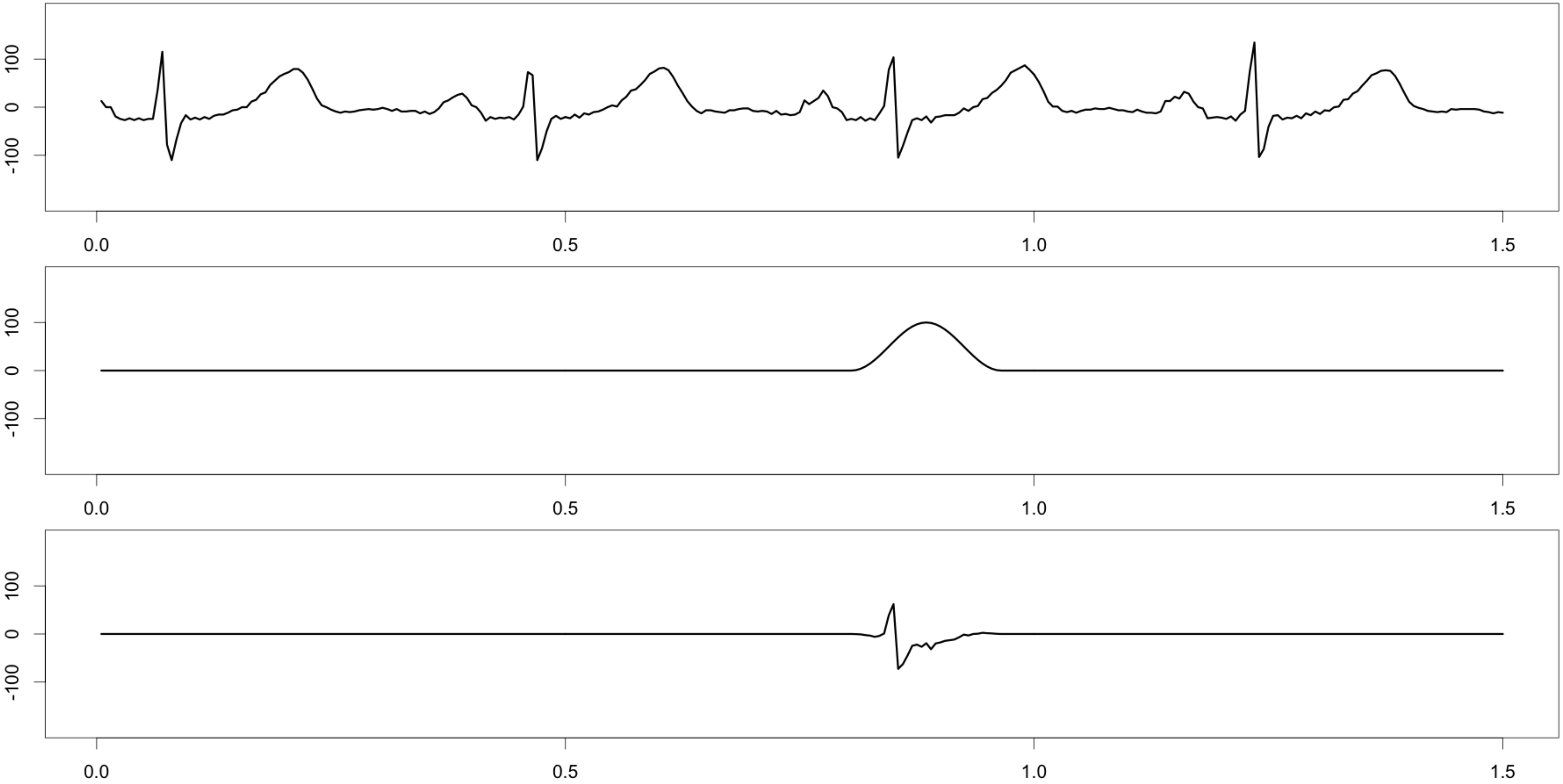
© MapR Technologies, confidential

We Do Windows



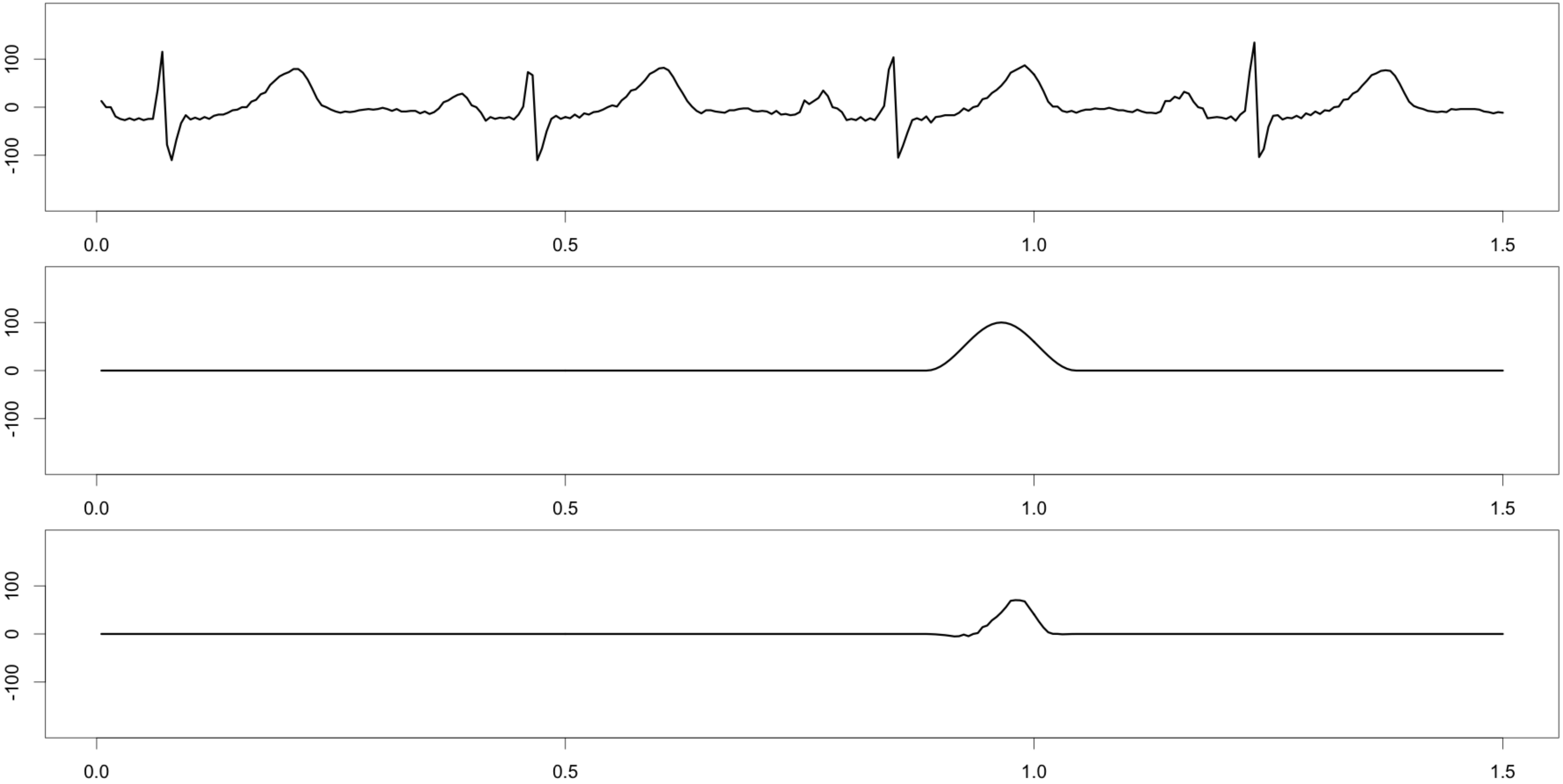
© MapR Technologies, confidential

We Do Windows

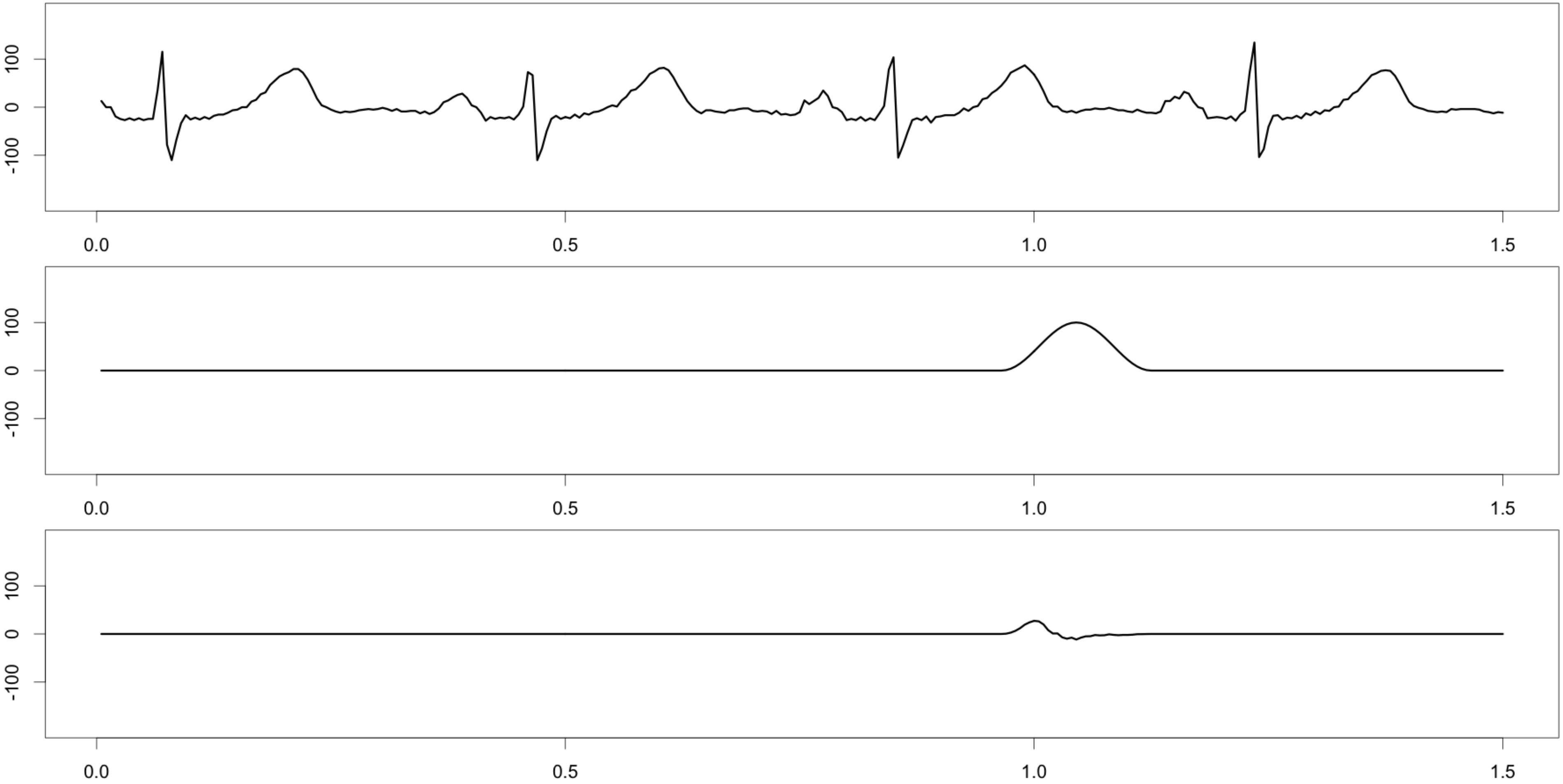


© MapR Technologies, confidential

We Do Windows

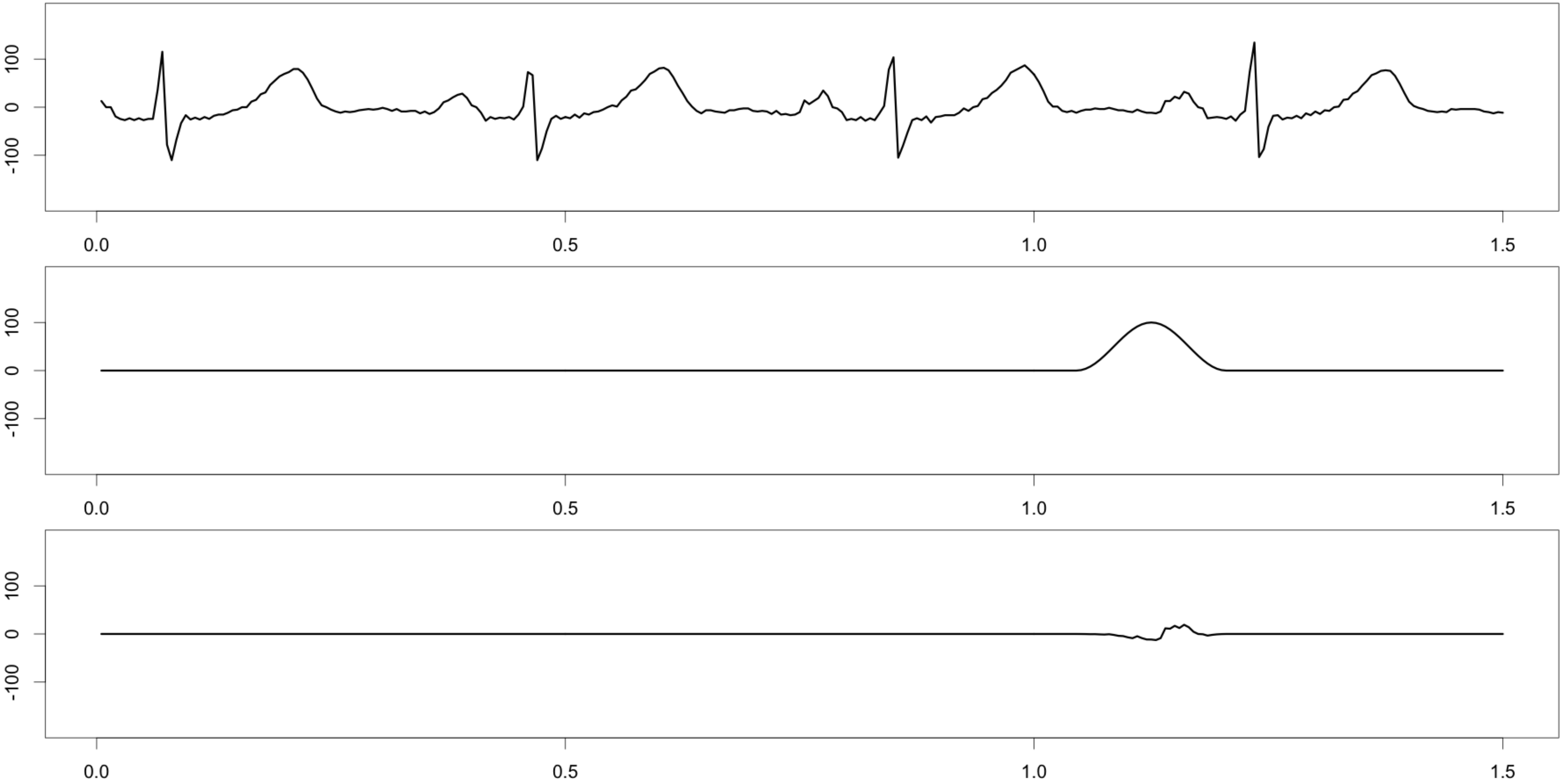


We Do Windows



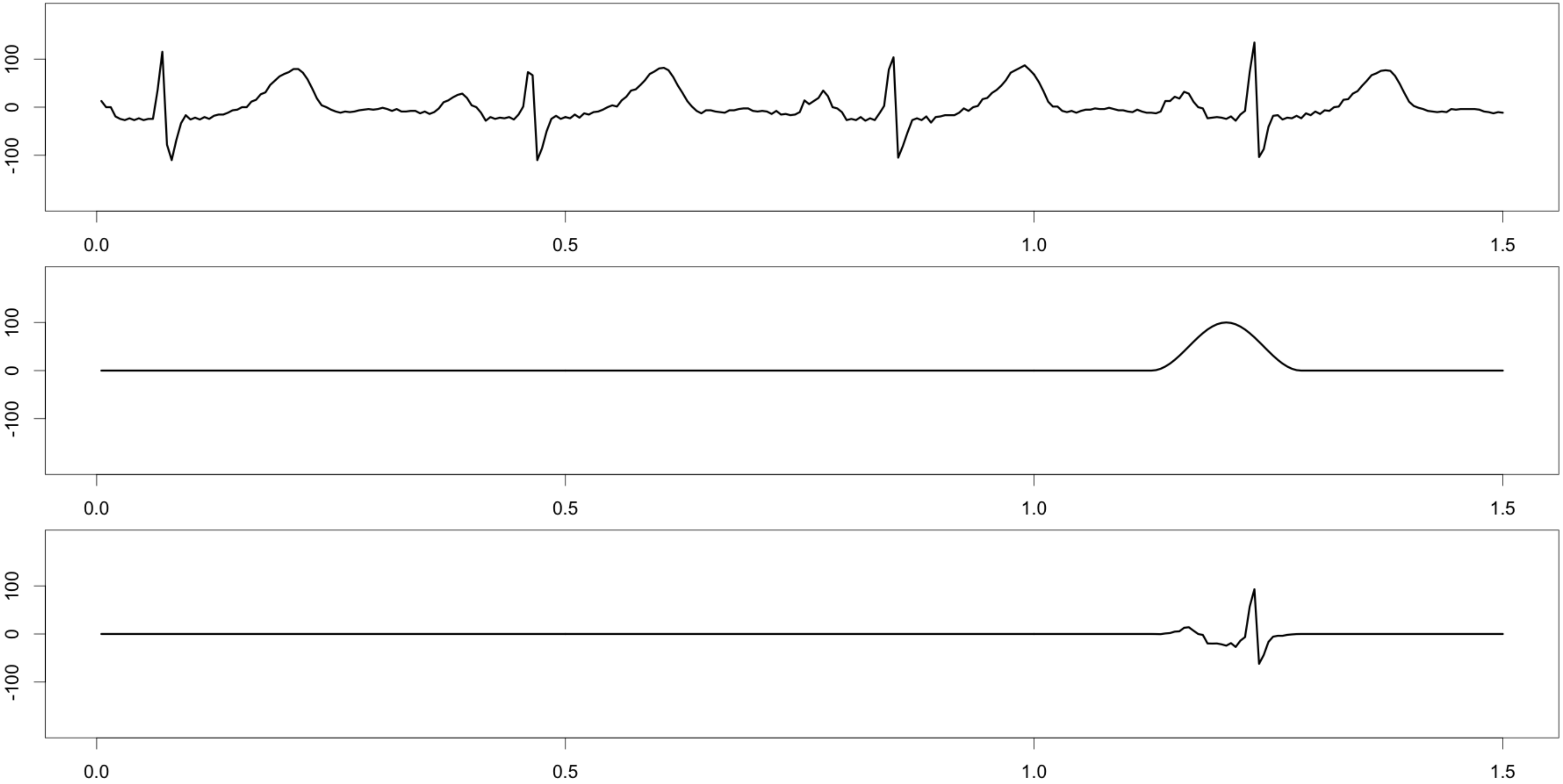
© MapR Technologies, confidential

We Do Windows



© MapR Technologies, confidential

We Do Windows



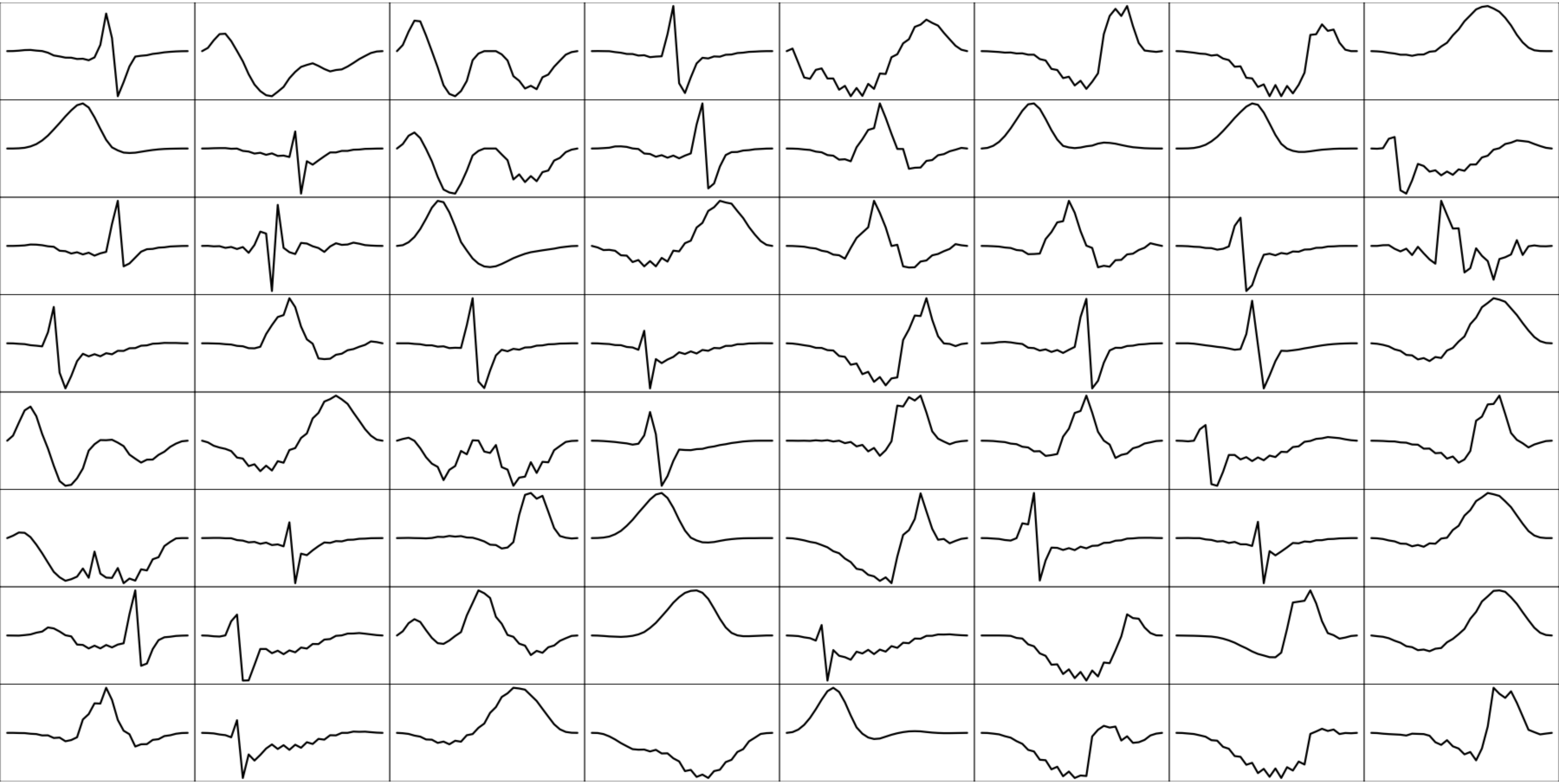
Windows on the World

- 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

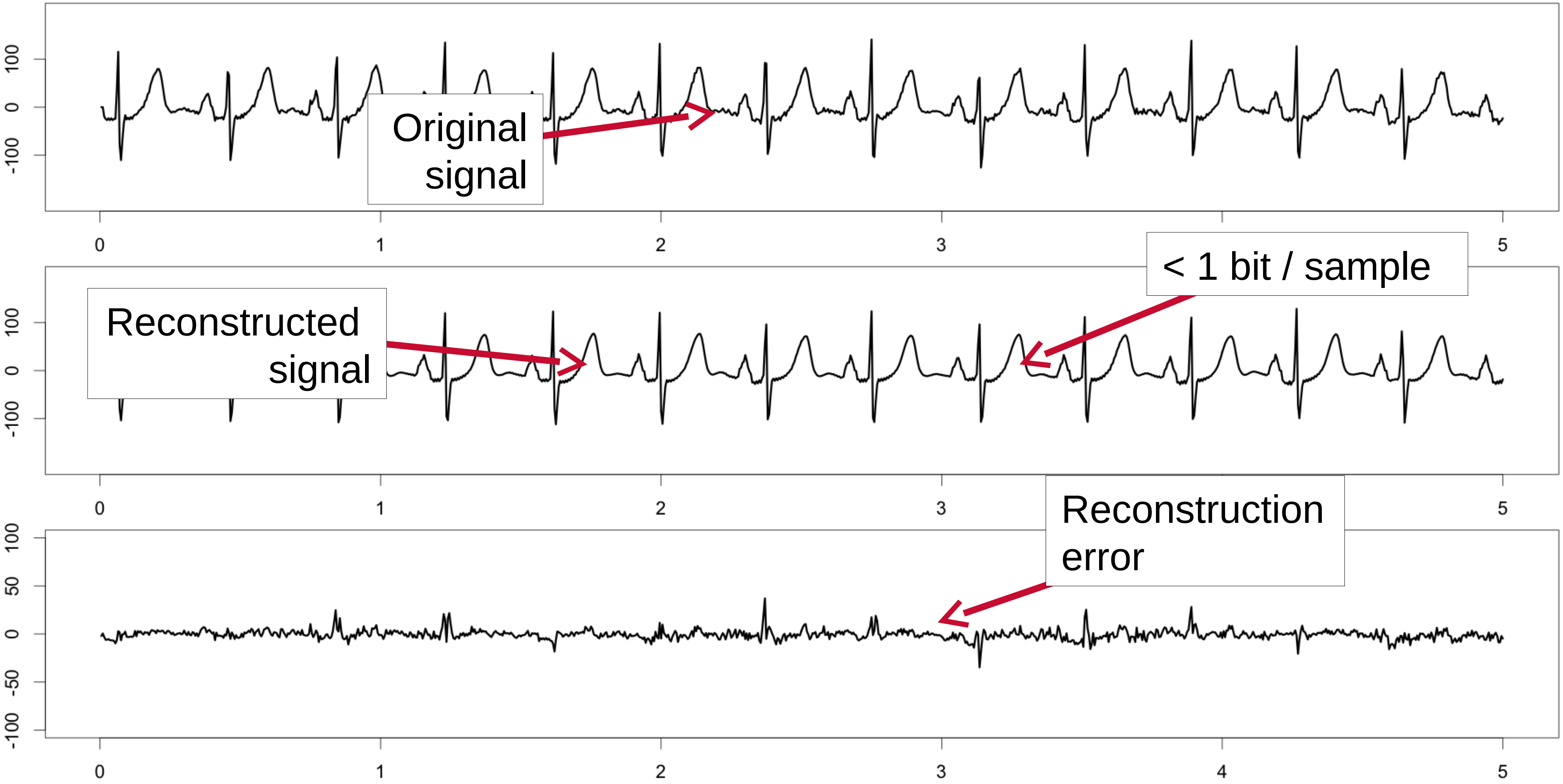
© MapR Technologies, confidential

Most Common Shapes (for EKG)



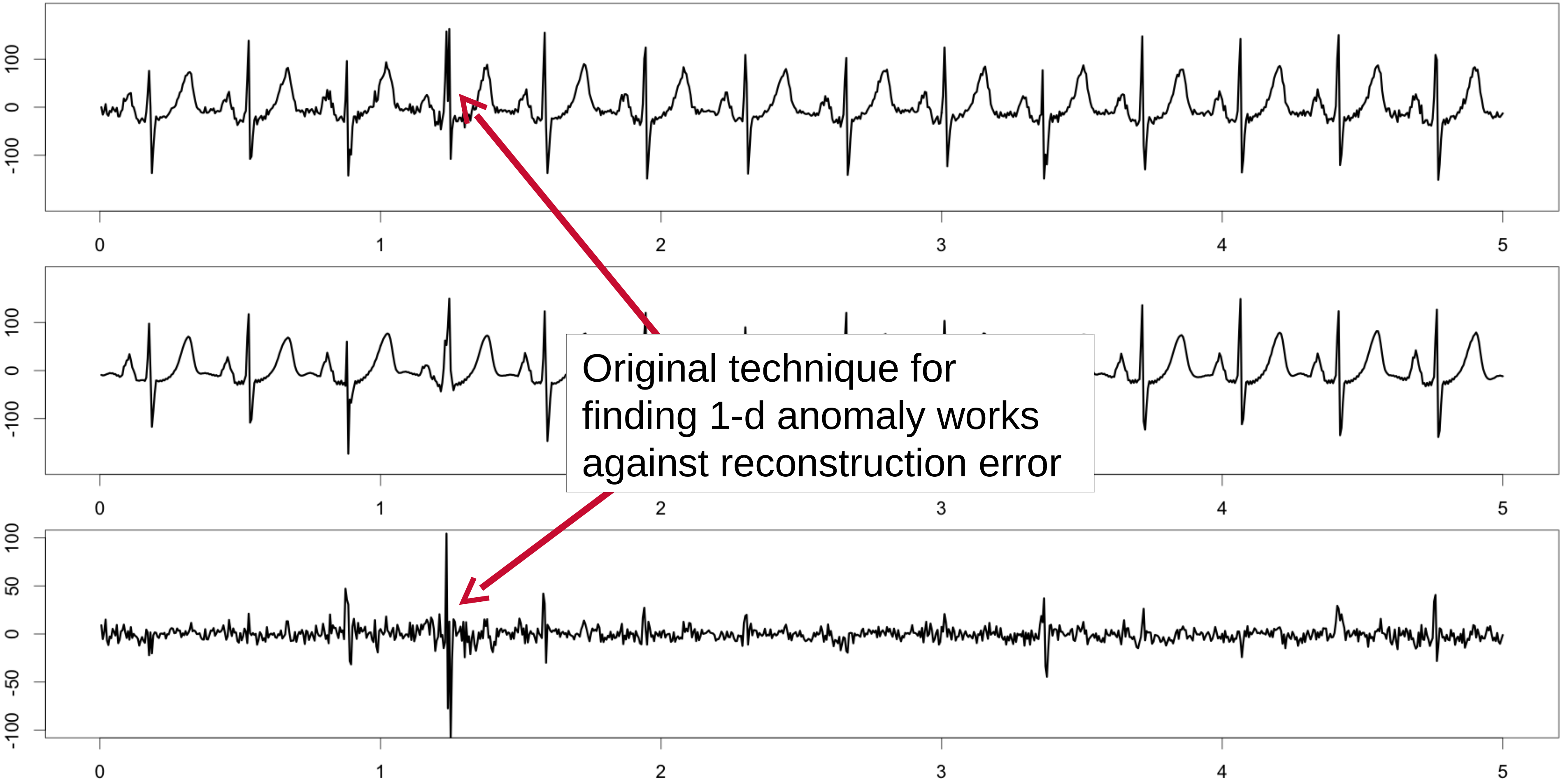
© MapR Technologies, confidential

Reconstructed signal



© MapR Technologies, confidential

An Anomaly


100
0
-100
0
1
2
3
4
5
Original technique for finding 1-d anomaly works against reconstruction error
100
50
0
-50
-100

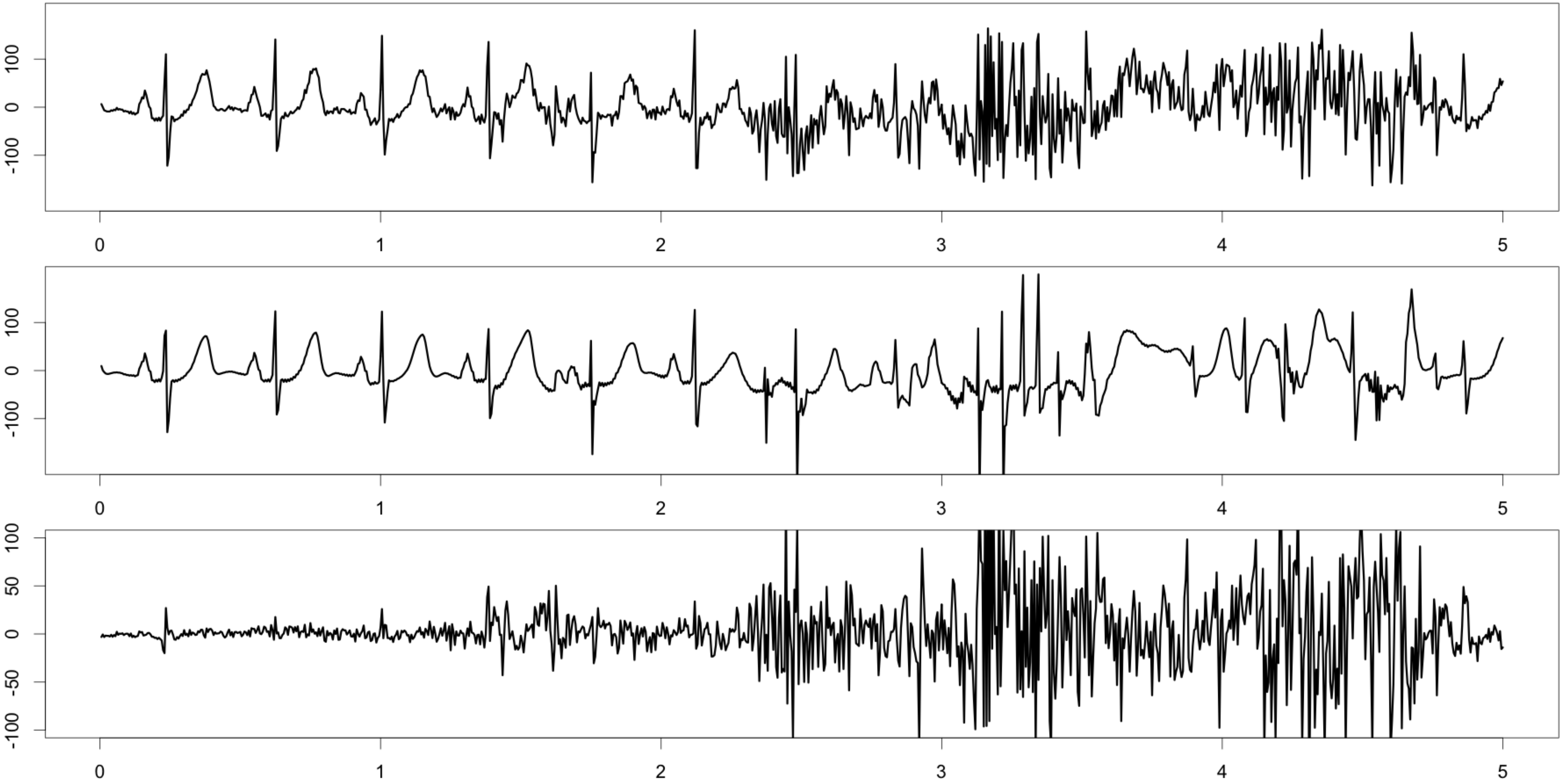
Close-up of anomaly

Not what you want your heart to do.

And not what the model expects it to do.

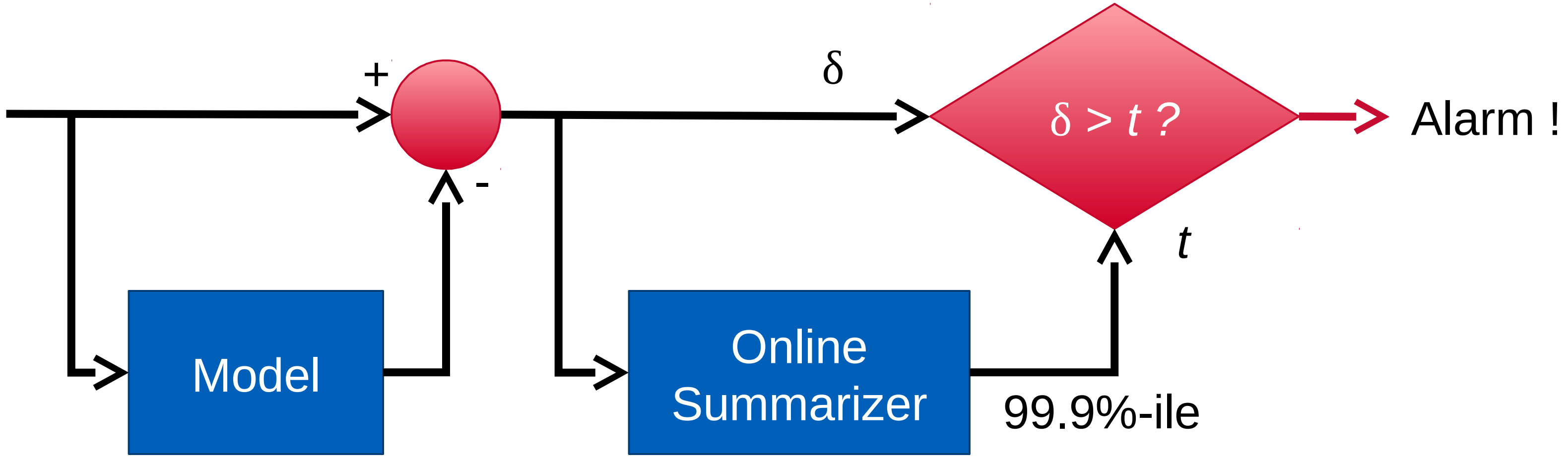
© MapR Technologies, confidential

A Different Kind of Anomaly



© MapR Technologies, confidential

Model Delta Anomaly Detection



© MapR Technologies, confidential

The Real Inside Scoop

- The model-delta anomaly detector is really just a sum of random variables
  - the model we know about already
  - and a normally distributed error

- The output (delta) is (roughly) the log probability of the sum distribution (really δ2)

- 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

© MapR Technologies, confidential

Recap (out of order)

- Anomaly detection is best done with a probability model
- -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 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

© MapR Technologies, confidential

But Wait! Compression is Truth

- Maximizing log πk is minimizing compressed size
  - (each symbol takes -log πk bits on average)

- Maximizing log πk happens where $\pi k = pk$
  - (maximum likelihood principle)

© MapR Technologies, confidential

But Auto-encoders Find Max Likelihood

- Minimal error => maximum likelihood
- Maximum likelihood => maximum compression
- So good anomaly detectors give good compression

© MapR Technologies, confidential

In Case You Want the Details

$$E[p_k \log \pi_k] = \sum_k p_k \log \pi_k$$

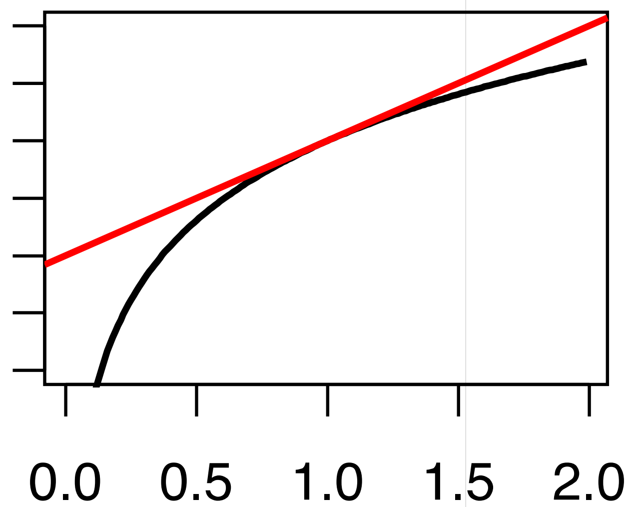
$$\log x \le x - 1$$

$$\sum_k p_k \log \frac{\pi_k}{p_k} \le \sum_k p_k \left(1 - \frac{\pi_k}{p_k}\right) = \sum_k p_k - \sum_k \pi_k = 0$$

$$\sum_k p_k \log \pi_k - \sum_k p_k \log p_k \le 0$$

$$\sum_k p_k \log \pi_k \le \sum_k p_k \log p_k$$

$$E[p_k \log \pi_k] \approx \frac{1}{n} \sum_i \log \pi_{x_i}$$

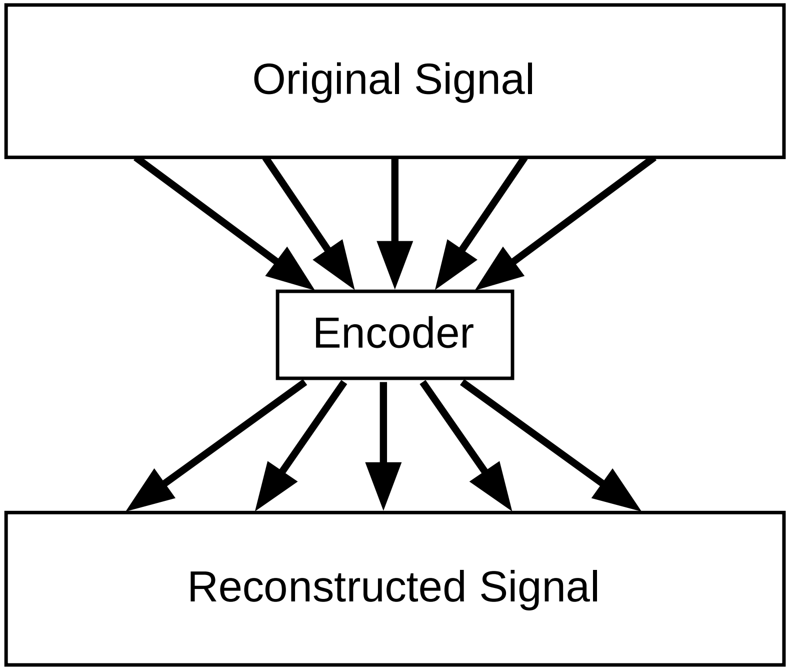


Pause To Reflect on Clustering

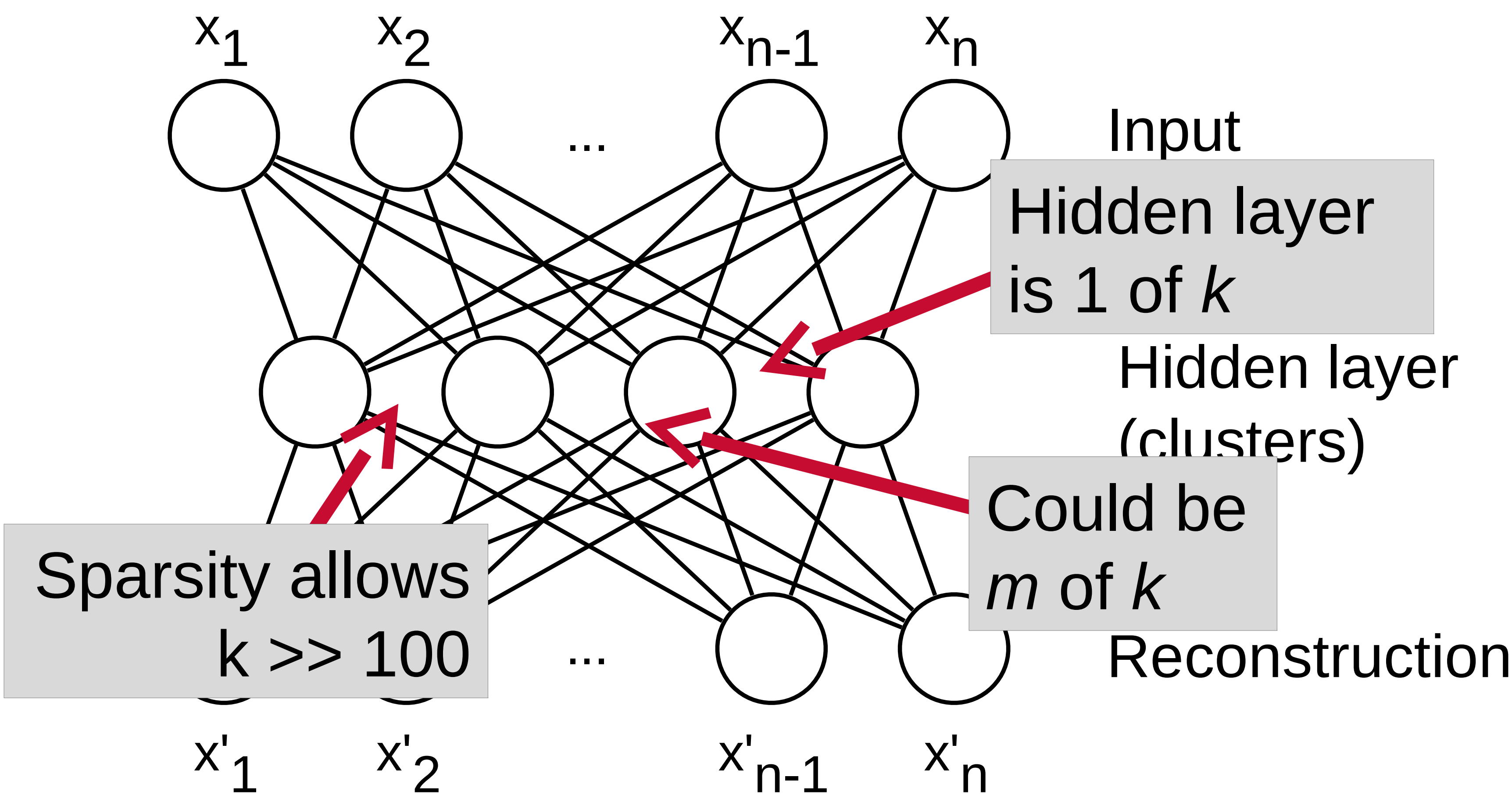
- Use windowing to apportion signal
  - Hamming windows add up to 1
- Find nearest cluster for each window
  - Can use dot product because all clusters normalized
- Scale cluster to right size
  - Dot product again
- Subtract from original signal

© MapR Technologies, confidential

Auto-encoding - Information Bottleneck

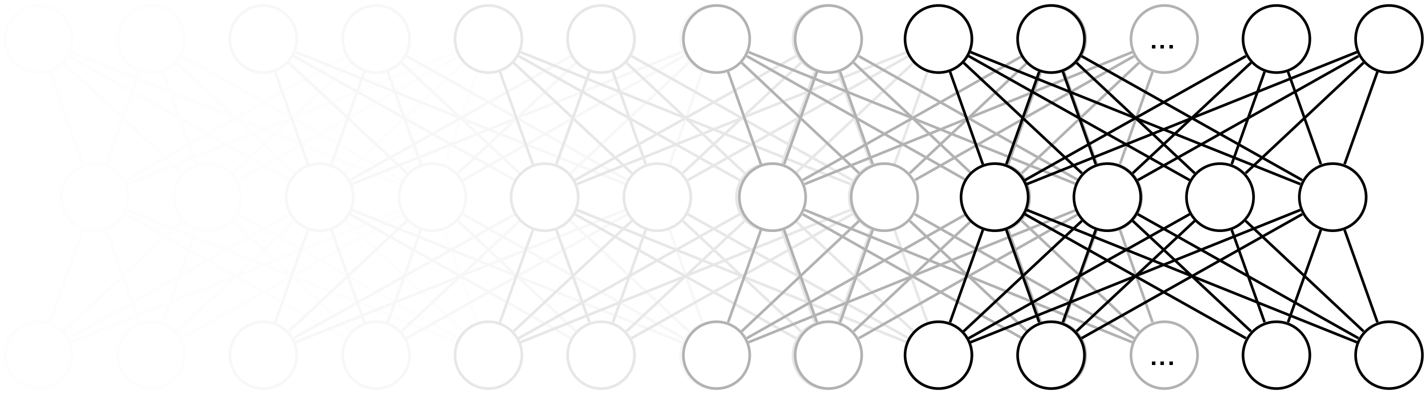


Clustering as Neural Network



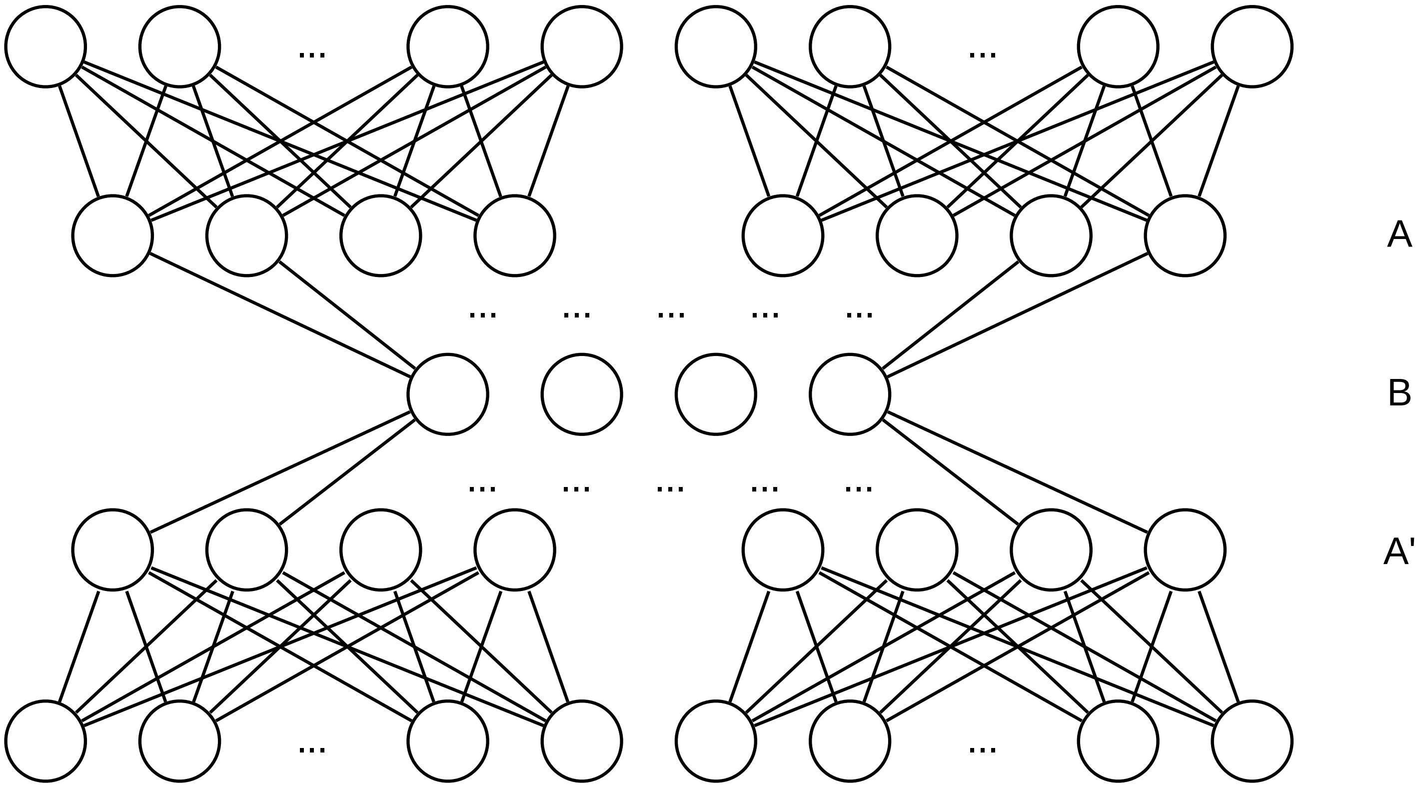
Overlapping Networks

Time series input



Reconstructed time series

Deep Learning



What About the Database?

- We don’t have to keep the reconstruction
- We can keep the first level nodes
  – And the reconstruction error
- To keep the first level nodes
  – We can keep the second level nodes
  – Plus the reconstruction error

© MapR Technologies, confidential

What Does it Matter?

- Even one level of auto-encoding compresses
  - 30-50x in EKG example with k-means
- Multiple levels compress more
  - Understanding => Truth => Compression
- Higher levels give semantic search

© MapR Technologies, confidential

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

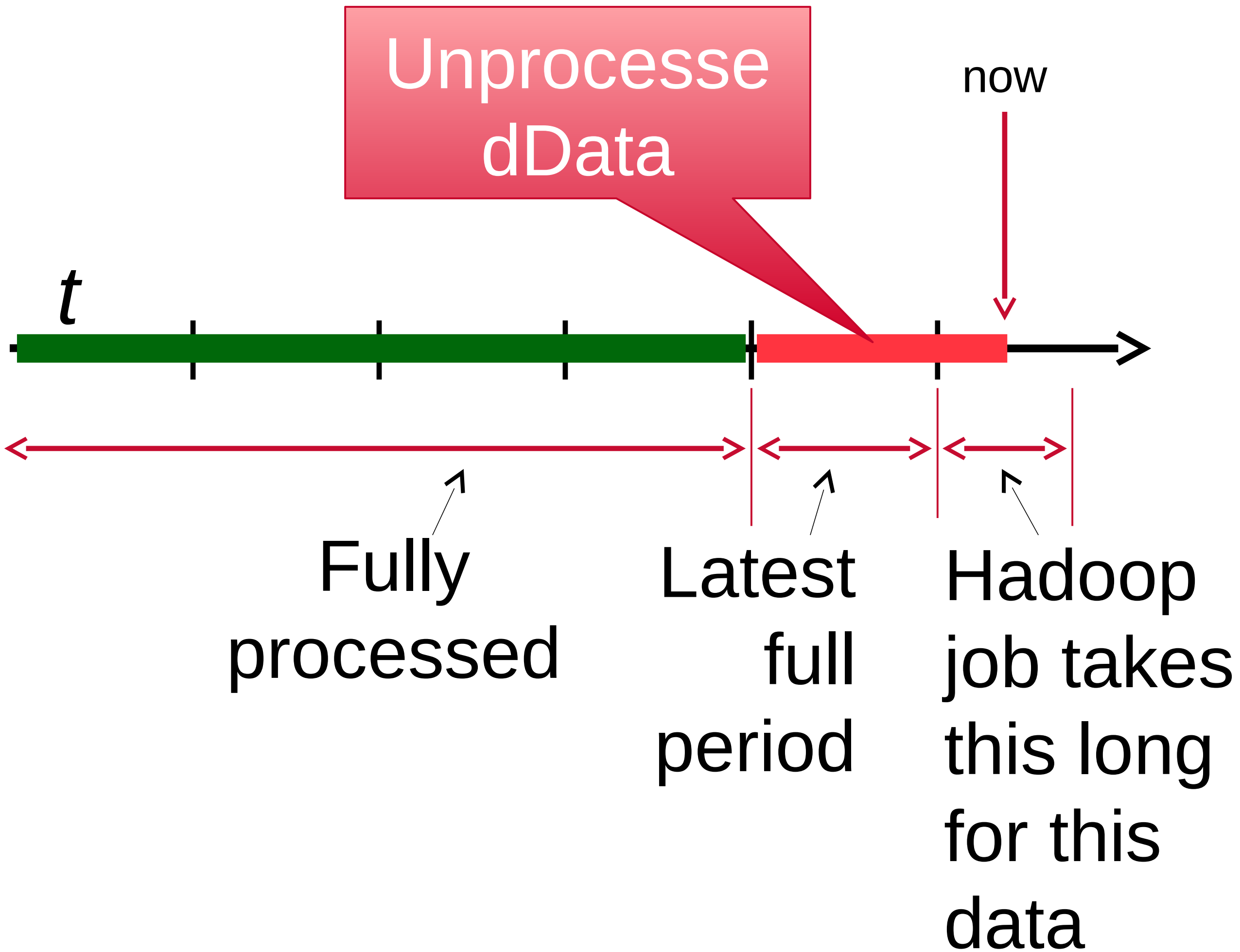
- Extended Lambda architecture is my favorite

- See my other talks on slideshare.net for info

- Ping me directly

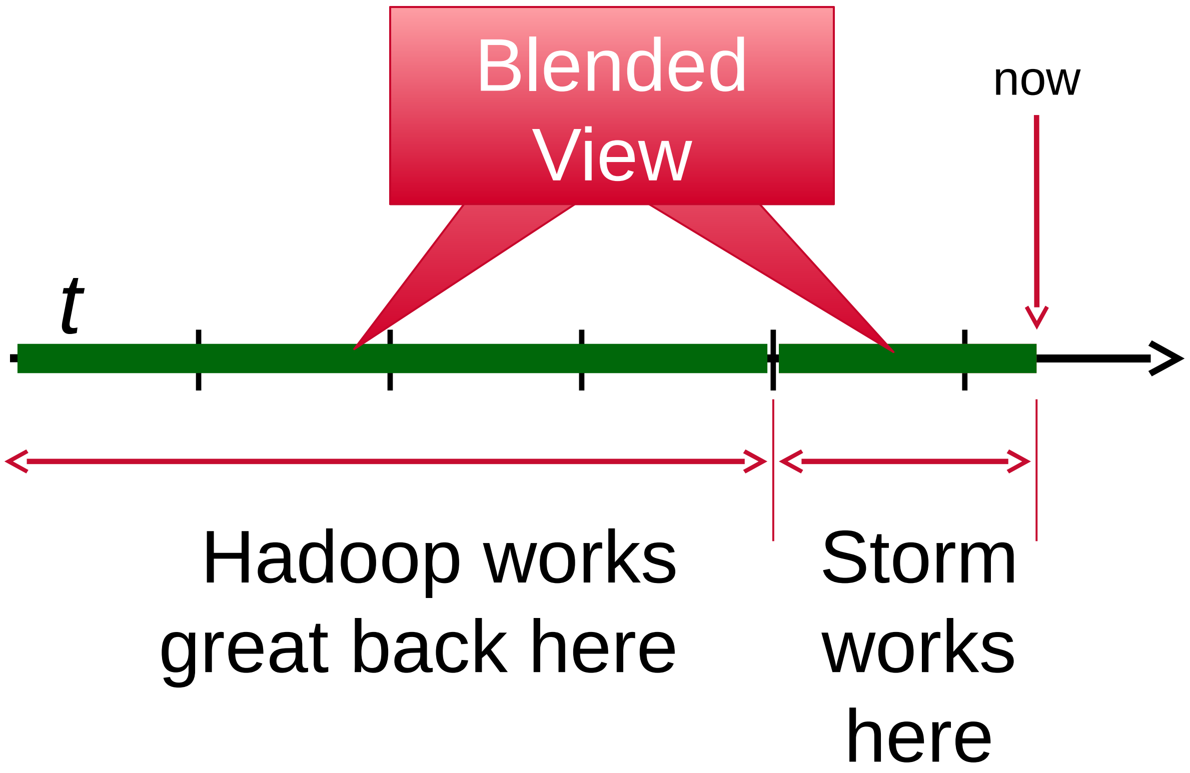
© MapR Technologies, confidential

Hadoop is Not Very Real-time



© MapR Technologies, confidential
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 Technologies, confidential