High Performance Time Series Databases
Agenda

• What is anomaly detection?
• Some examples
• Compression == Truth
• Deep dive into deep learning
• Time series databases and deep learning
Who I am

• Ted Dunning, Chief Application Architect, MapR
dunning@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
Looks pretty anomalous to me
Will the real anomaly please stand up?
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

![Graph showing a signal over time](image-url)
A Second Look

99.9%-ile
How Hard Can it Be?

Online Summarizer

99.9%-ile

x

x > t?

Alarm!

© MapR Technologies, confidential
Already Done? Etsy Skyline?

- **past hour**
  - name of anomalous metric
- **past 24 hours**
- **red line shows detected anomalous datapoint**
- **holy anomaly!**
- **timeseries datapoint that triggered detection**
- **list of anomalous metrics**
What About This?
Spot the Anomaly
Maybe not!
Where’s Waldo?

This is the real anomaly
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 …

\[ x \sim N(0, \sigma) \]

• The real world is rarely so accommodating
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
We Do Windows
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
Most Common Shapes (for EKG)
Reconstructed signal

Original signal

Reconstructed signal

Reconstruction error

< 1 bit / sample
An Anomaly

Original technique for finding 1-d anomaly works against reconstruction error.
Close-up of anomaly

Not what you want your heart to do.
And not what the model expects it to do.
Model Delta Anomaly Detection

Model

Online Summarizer

\[ \delta > t ? \]

99.9%-ile

Alarm!
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.
Agenda

• What is anomaly detection?
• Some examples
• Compression == Truth
• Deep dive into deep learning
• Time series databases and deep learning
Probability Models

- We make observations, $x_i \in E$
- Each observation has underlying probability $p(e)$
- We want to estimate these probabilities,
  \[ p(e) \]
- We call $p(e)$ “truth”
Compression

• Our probability estimate lets us compress our observations
• The compressed size of one event is
  \[ |e| \log_2 e \]

• The size of a bunch of events is
  \[ |c(e_1...e_n)| \sum_i \log_2 e_i \]
But Wait! Compression is Truth

• Maximizing $E[\log_2 e]$ minimizes compressed size

• Maximizing $E[\log_2 e]$ happens exactly when $e = p(e)$
But Auto-encoders Find Max Likelihood

• Minimal error ⇒ maximum likelihood

• Maximum likelihood ⇒ maximum compression

• So good anomaly detectors give optimal compression
In Case You Want the Details

\[ E[\log_e x] = p_e \log_e x \]

\[ \log x = x - 1 \]

\[ p_e \log_e \frac{e}{p_e} = p_e \frac{1}{p_e} = e = 0 \]

\[ p_e \log_e p_e^k = 0 \]

\[ p_e \log_e p_e = p_e \log p_e \]

\[ E[p_e \log_e x] = \frac{1}{n} \sum_{i=1}^{n} \log x_i \]
Conclusion

Compression = truth
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
Clustering as Neural Network

- Input
- Hidden layer is 1 of $k$
- Hidden layer (clusters)
- Could be $m$ of $k$
- Reconstruction

Sparsity allows $k >> 100$
Clustering as Neural Network

Input

Hidden layer (clusters)

Reconstruction is lookup

Dot product = distance

Reconstruction
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
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
System Architecture

\[ \delta_1 \ldots \delta_n (x_1 \ldots x_n) \]

DB

Model

© 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