Adaptive Time Series Representations

By Harrison Boyer and Benjamin Deckey

Motivation

- The motivation behind our work is to derive an efficient representation and compare it to various state of the art representation techniques
- Representations compared against in this project:
  - Piecewise Aggregate Approximation(PAA)
  - Adaptive Piecewise Constant Approximation(APCA)

Current State of the Art Techniques

- Data Adaptive:
  - APCA
  - SVD
  - Piecewise Polynomials (interpolation, regression)
  - Symbolic(SAX, iSAX)
- Non Data Adaptive:
  - PAA
  - Wavelets
  - Spectral Mappings (DFT)

Limits of these techniques

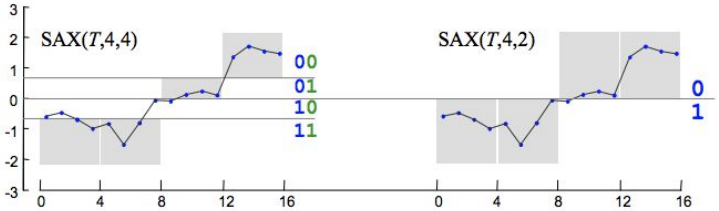
- Tradeoff between accuracy and compression
- The more compressed the representation becomes the more information is lost about the original series
- Industry wide representations mitigate this tradeoff in different ways
- APCA performs well with both dimensionality reduction and reconstruction error however fails to yield an efficient indexing scheme
- SAX is an extension of PAA and thus does not adapt to different scenarios easily
- Our goal is to develop a data adaptive representation that minimizes reconstruction error with regards to various similarity metrics.

Adaptive Piecewise Constant Approximation(APCA)

- Breaks TS down into segments that are each represented by their length and mean value
- Segments are adapted to volatility in data values
- Outlines various methods to derive segmentation:
  - Linear Interpolation (Stan's paper *Approximate queries and representations for large data sequences*)
  - Wavelet transform(this is how they decided to implement segmentation)

Symbolic Aggregate Approximation (SAX)

- Builds symbols or words from TS as representation
- SAX uses a piecewise aggregate segmentation(PAA) and then assigns symbols to each segment :



- iSAX is an extension of SAX that improves on the indexing scheme of SAX

Our Representation

- Uses a k-means clustering algorithm to match each point in the TS to a cluster, the series is then represented by the clusters
- This algorithm is data adaptive as the only parameter it requires is the amount of clusters you desire to represent the series

Metrics

- Norms are a good way to measure ‘distance’ of our representation to the original series
- We use Lebesgue norms to measure our representations

$$\|x\|_p = \left(\sum_{i \in I} |x_i|^p \right)^{1/p}$$

- Where p=1(Manhattan Distance), p=2 (Euclidean Distance), and p=∞ (Chebyshev distance)
- $x = TS_{Approx} - TS_{orginal}$

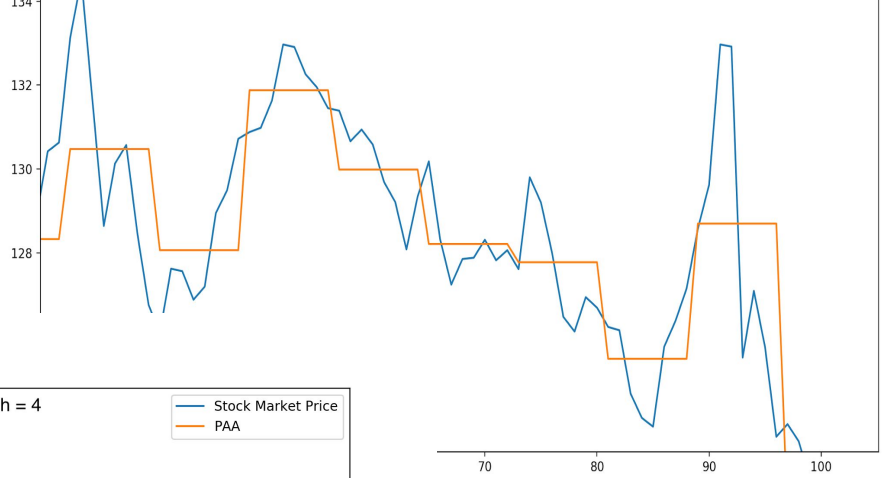
Dynamic Time Warping - Norms are not equipped at dealing with shifted series, thus DTW adds another metric to measure our accuracy

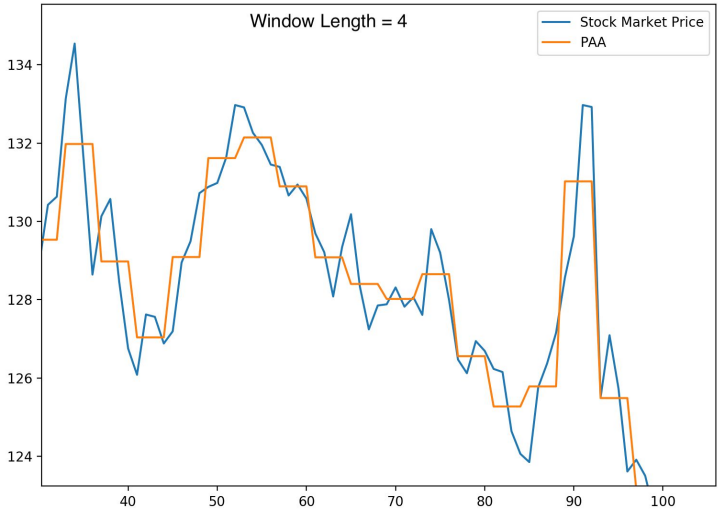
Comparison and Parameterization

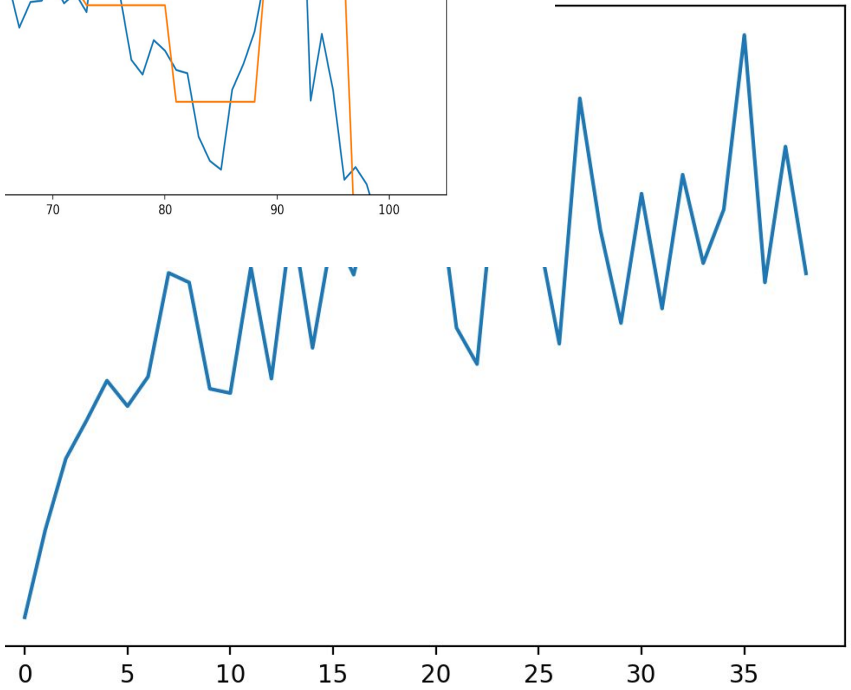
- As each of these representations have parameters that can be tuned to data for better results we wanted to test the error each representation has over a range of different parameter values
- This way we can have a better apples to apples comparison if we are using each representations tuned parameters

Comparis

PAA is a represe



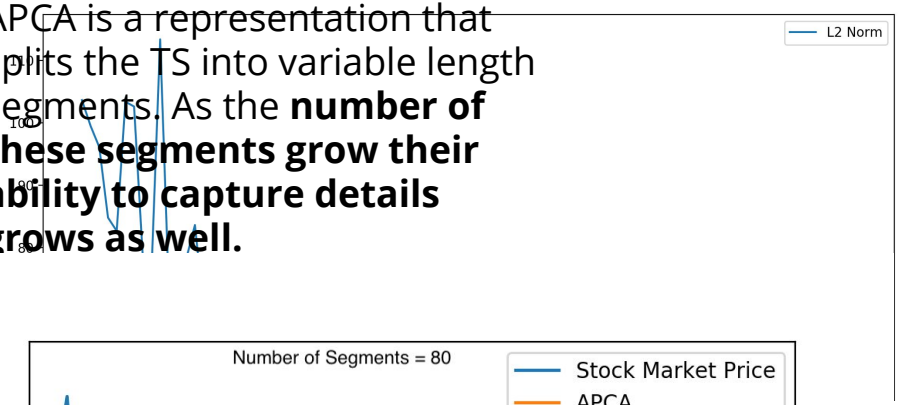


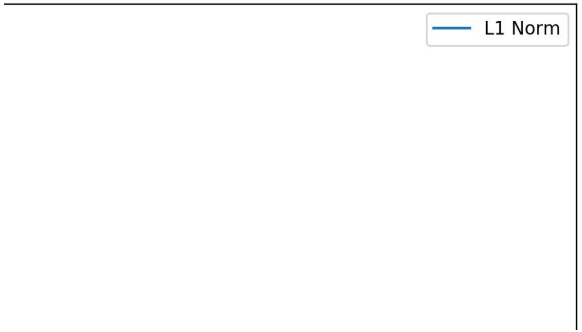


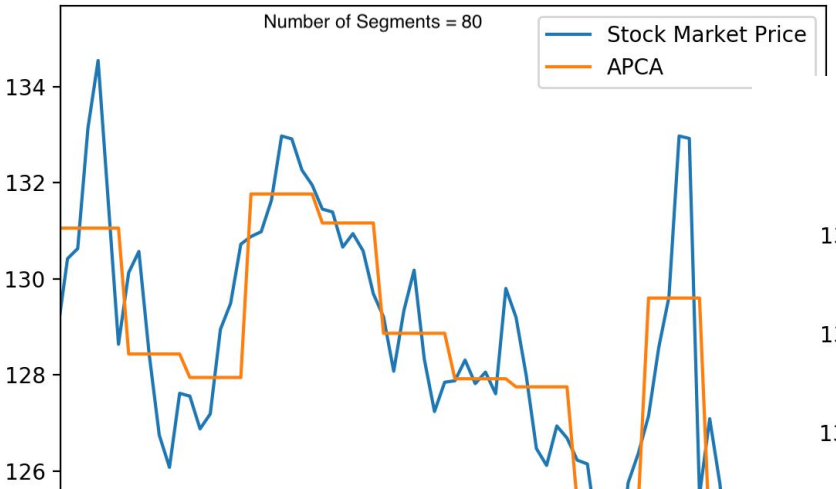
Comparison: APCA

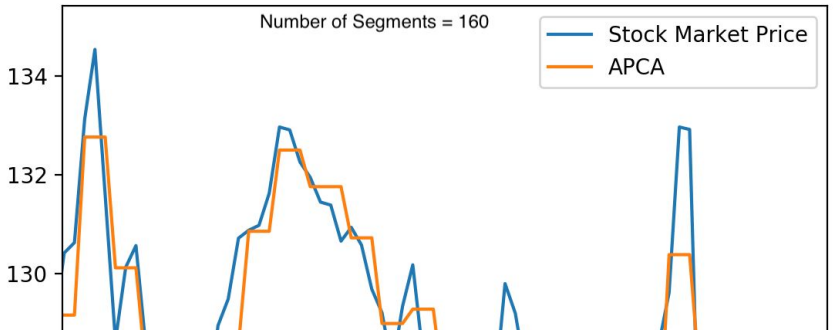
f Segments - 1 - 90

APCA is a representation that splits the TS into variable length segments. As the **number of these segments grow their ability to capture details grows as well.**



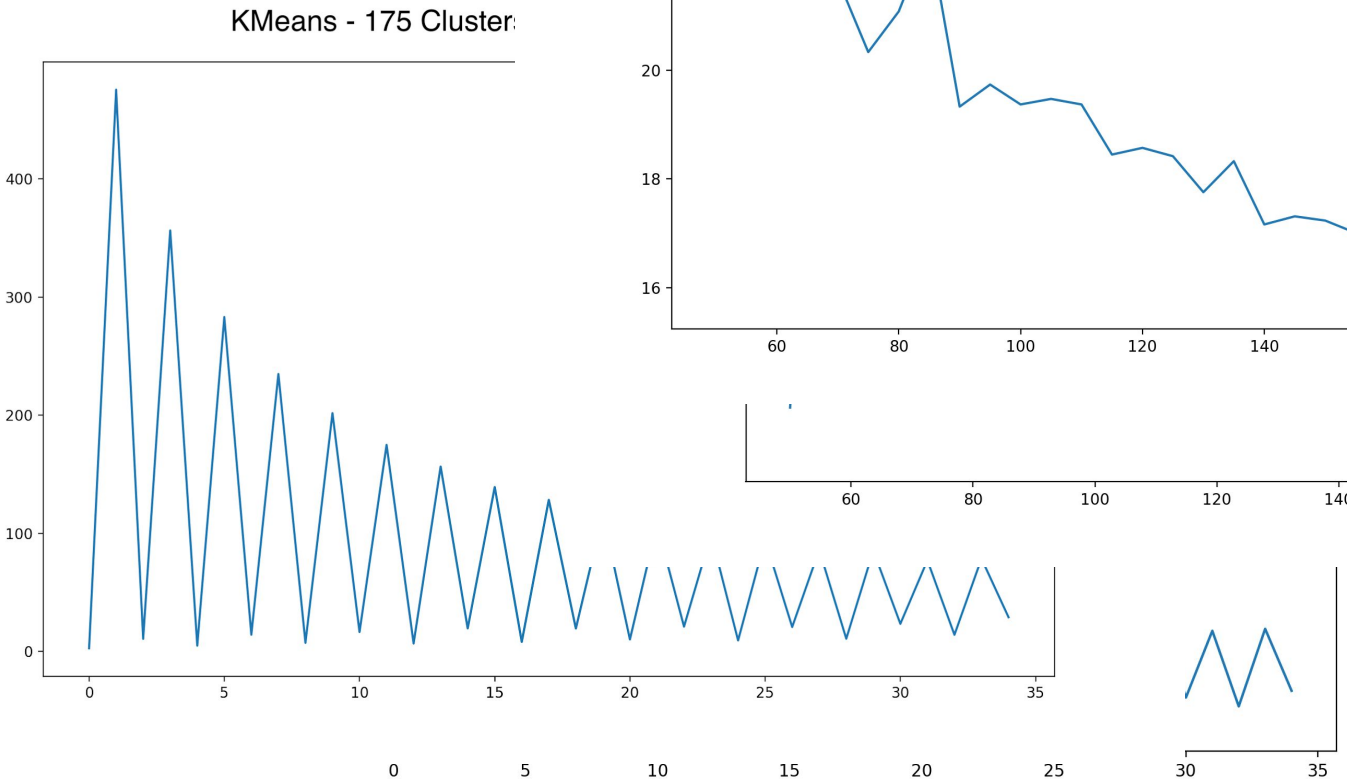




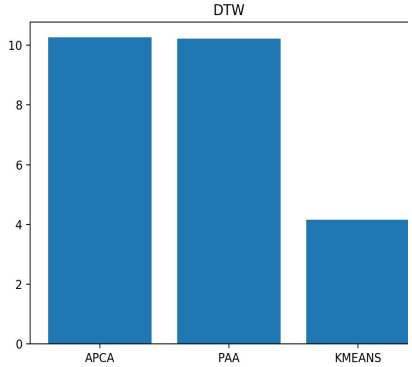


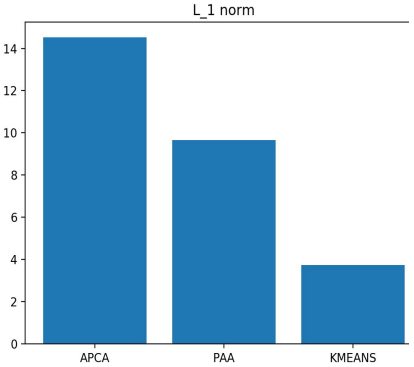
Comparison: K-Means

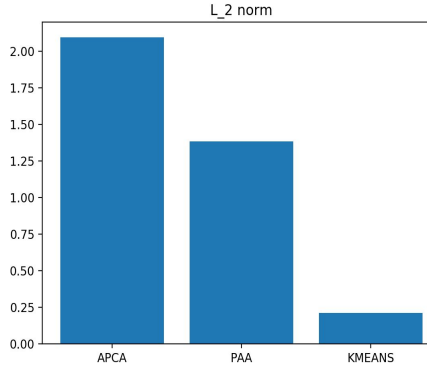
K-Means looks
series through
fixed length wir
These windows
clustered toget
on the similarit
features. **As th
size shrinks,** e
sees more simp
the TS and **may
able to recons
complex chan
data**. We also h
number of clus
will change hov
the matching o
algorithm is.

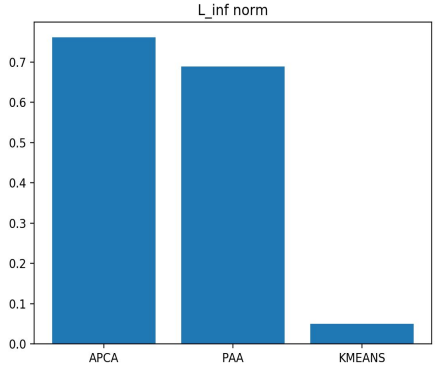


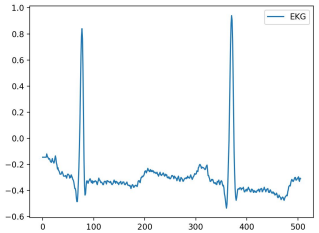
Results: bar graphs representing how well representations perform given different metrics for EKG data



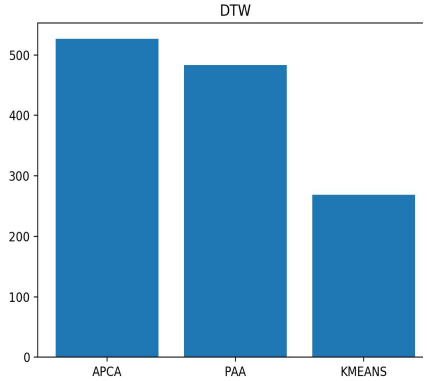


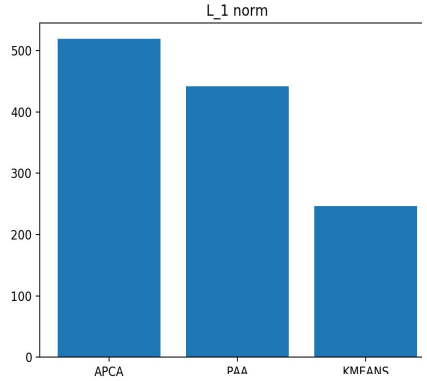


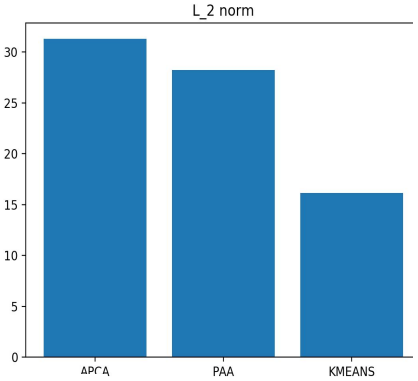


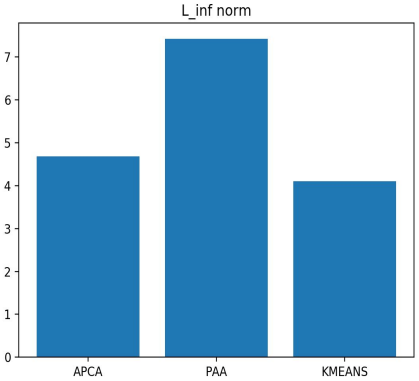


Results: bar graphs representing how well representations perform given different metrics for Stock data









Conclusion/Questions

- We developed a k-means clustering algorithm that successfully reduces the dimensions of the TS while maintaining sufficient information
- We compared our method to industry standard methods using different Lebesgue norms and DTW, k-means out performed APCA and PAA on all metrics
- Future work involves determining how best to tune the K-Means parameters to see the best results
- We have no real world application of our algorithm, so we would also like to see how well our algorithm performs on a much larger scale w.r.t. speed and efficiency

Thanks!