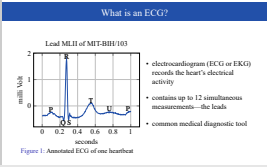
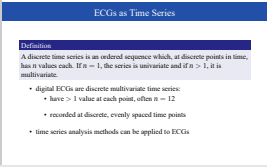
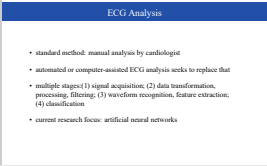


Figure 1: Annotated ECG of one heartbeat

- P wave: blood entering the heart
- QRS complex: heart contraction pumping blood
- T: return of ventricle to polarized state
- U: present in 25%
- muscle contractions caused by electric pulses
- electric pulse can be measured on the skin
- electrodes form leads (need 2 to measure anything)
- most types of heart disease can be detected
- diagnosis and analysis is performed by trained cardiologists
- **datasets available online; contain 2 or more leads (the most significant ones)**
- **I will be using online datasets for my analysis**
- heart diseases are some of the most deadly ones, thus ECG are really important



- modern ECGs have at least 2, most have 12
- digital ones have set sampling frequencies, even the machines have set frequencies
- multivariate: measure more than 1 lead per time point
- discrete: set sample frequency in the machines
- discrete: because measured at discrete moments in time
- time series: they are data measured at equal time intervals
- n measurements per point in time (i.e. leads)
- $n = 1$ is univariate, $n > 1$ is multivariate

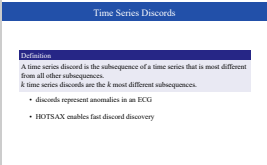


- is relatively slow; time is of the essence
- lots of training required
- error prone
- maybe not feasible for long ECGs
- can speed up process
- can pick up details humans miss
- digitizing paper ECGs or recording digital ones
- filtering to remove various types of noise
- reduce complexity of the data
- select important features and neglect irrelevant ones to ease analysis
- often added, figure out if there is some disease present or not
- balance between accuracy and complexity needed
- ann: hand all the steps discussed to a NN; use as good classifier too

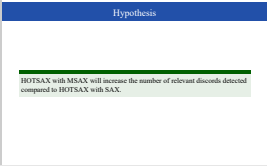
SAX, MSAX, and HOTSAX

- Lin *et al.* (2003):
  Symbolic Aggregate Approximation (SAX)—simplified, symbolic representation
- Anacleto *et al.* (2020):
  Multivariate SAX (MSAX)—expands SAX to multivariate time series
- Keogh *et al.* (2005):
  Heuristically Ordered Time series using SAX (HOTSAX)—discord discovery algorithm for SAX

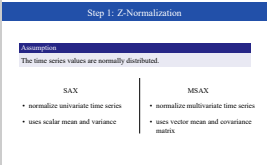
- ecg as letters that mean same thing as original
- guaranteed to behave like the original data
- works on univariate time series
- has been used on ECGs
- takes the correlation between ecg leads into account
- cov mat: covariance between each lead and variance on diag
- uses sax representation to make the finding of discords easier
- can use MSAX just as well



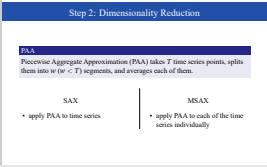
- these can be diseases, noise, etc
- the discord does not discern



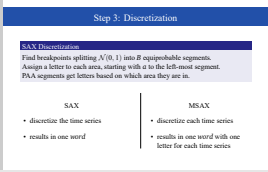
- **mention that MSAX to ECGs in particular is new**
- **mention that HOTSAX with MSAX is new**
- THIS METHOD WILL NOT BE SUPER ACCURATE; MANY ECG changes are relatively small and would get lost in the SAX process



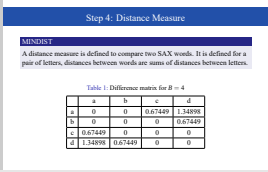
- say that the process is the same as MSAX based on SAX
- this is assumed and this worked for other people who applied SAX to ECGs
- to compare time series, normalization is the accepted step
- what is this
- takes into account the correlation between leads



- this reduces complexity
- PAA form of time series is shorter and simpler
- it still somewhat corresponds to the original



- result is called word
- N is the alphabet size
- big thing here is that this gives defined probability to each letter; makes no sense for real numbers (like PAA values)
- simplifies time series even more
- creates discrete categories, can be more useful


Step 4: Distance Measure

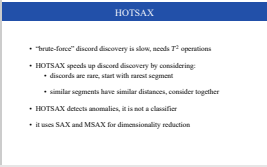
MINDIST

A distance measure is defined to compare two SAX words. It is defined for a pair of letters, distances between words are sums of distances between letters.

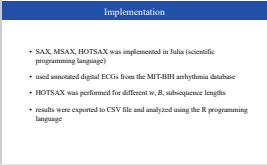
Table 1: Difference matrix for $B = 4$

	a	b	c	d
a	0	0	0.67449	1.34898
b	0	0	0	0.67449
c	0.67449	0	0	0
d	1.34898	0.67449	0	0

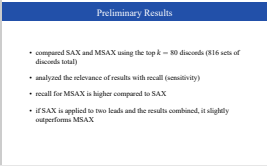
- distance is based on letter pairs
- SAX: sqrt of sum of squared distance
- MSAX: sqrt of sum of squared distance; also sum all leads
- this lower-bounds the euclidean distance, meaning that results in SAX should hold true for the real data too



- this is the basic idea that can speed up the process
- it is not guaranteed to do so, but it does not decrease efficiency
- this speeds up the process even more as we have fewer elements
- because of lower bounding, it still gives accurate results



- fast, type support, great libraries, JIT compilation
- ecgs have all heart beats annotated
- know which are normal, diseases, noise, etc
- 48 recordings of 30 minutes
- *w* - paa segments; *B* - alphabet size; subsequence length for HOTSAX



- how many relevant items are selected
- recall = true positive / (true positive + false negative)
- this is done because for medical things it is more useful to look at a couple too many segments than not enough

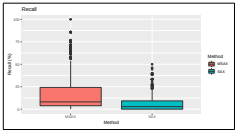


Figure 4: Boxplot comparing Recall for MSAX and single-lead SAX

- msax: average = 17.5%
- sax: average = 6.4%

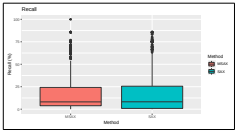


Figure 5: Boxplot comparing Recall for MSAX and dual-lead SAX

- msax: average = 17.5%
- sax: average = 18.5%

Remaining Tasks

- perform tests for statistical significance of the result
- analyze the outliers visible in the boxplots
- compute more sets of discords with different parameters
- explore the influence of the parameters on the result

- for example t-test, biserieal correlation