Time Series Classification:

Applying Deep Learning Techniques to Polysomnographic Sleep Data

Deep Learning in Medicine
Final Project May 8th, 2018
Mark Grivainis & Izzy Metzger

motivation behind the challenge

sleep deprivation

Early symptoms of sleep deprivation include difficulty reading and speaking, poor judgment, altered mood, and impaired memory.

Induced by a multitude of causes such as environmental factors, medications, illness, as well as sleep disorders.

Some sleep disorders include *obstructive sleep apnea, hypopnea, sleep related hypo-ventilation, circadian rhythm sleep-wake disorders, non-rapid eye movement sleep arousal disorders, nightmare disorder, rapid eye movement sleep, and restless legs syndrome*. Certain sleep disorders can be secondary to other health issues such obesity, post traumatic stress disorder, and generalized anxiety disorder.

challenge: identify non-apnea related arousals

The goal of the PhysioNet 2018 challenge is to identify non-apnea related arousals from data collected during polysomnographic sleep studies. Sleep apnea, the stopping of breathing for ten seconds or more. Sleep apnea has been well-studied in comparison to the other types of arousals.

There are many types of non-apnea related arousals such as *Bruxism (teeth grinding), Cheyne-Stokes breathing, Hypoventilation, Noise, Partial airway obstruction, periodic leg movement (PLM), Snoring, spontaneous arousals, and Respiratory effort (RERA).*

These regions (annotated as class 1) are part of a large unbalanced dataset, where class 0 represented non-arousal/apnea related arousal region dominates at around 92-95% of the data per reading.

A **polysomnography** is a multi-parameter assessment including electrooculography (EOG), electroencephalography (EEG), electromyography, nasal pressure and airflow and also includes REM onset.

EXCITING NEWS! Top AUPRC scores in the Unofficial Phase released yesterday: Matthew HP and Bahareh Pourbabaee with a score of 0.439, Yang Liu and Runnan He with a score of 0.244, and Márton Görög, Bálint Varga, and Péter Hajas with a score of 0.228.

The classification task:

- **Identify non-apnea arousals:**

target arousals are defined as regions where **either** of the following conditions were met:

- From 2 seconds before a **RERA** arousal begins, up to 10 seconds after it ends **or,**
- From 2 seconds before a **non-RERA**, **non-apnea** arousal begins, up to 2 seconds after it ends (2 seconds equals 24,000 samples)

RERA represented the majority of the target arousals (43,822 regions out of 44,005 of the non-apnea sleep arousals 99.5% of class 1).

The class labeled 0 included non-target arousals and non-arousals. Apnea related-arousals are classified under hypopnea (which is the most common arousal in the data set at 56,936 regions), central apnea, mixed apnea and obstructive apnea.

Regions were pre-computed and annotations were provided in the training set.

- Class 1 = arousal regions
- Class 0 = non-arousal regions
- Class -1 = regions that will not be scored

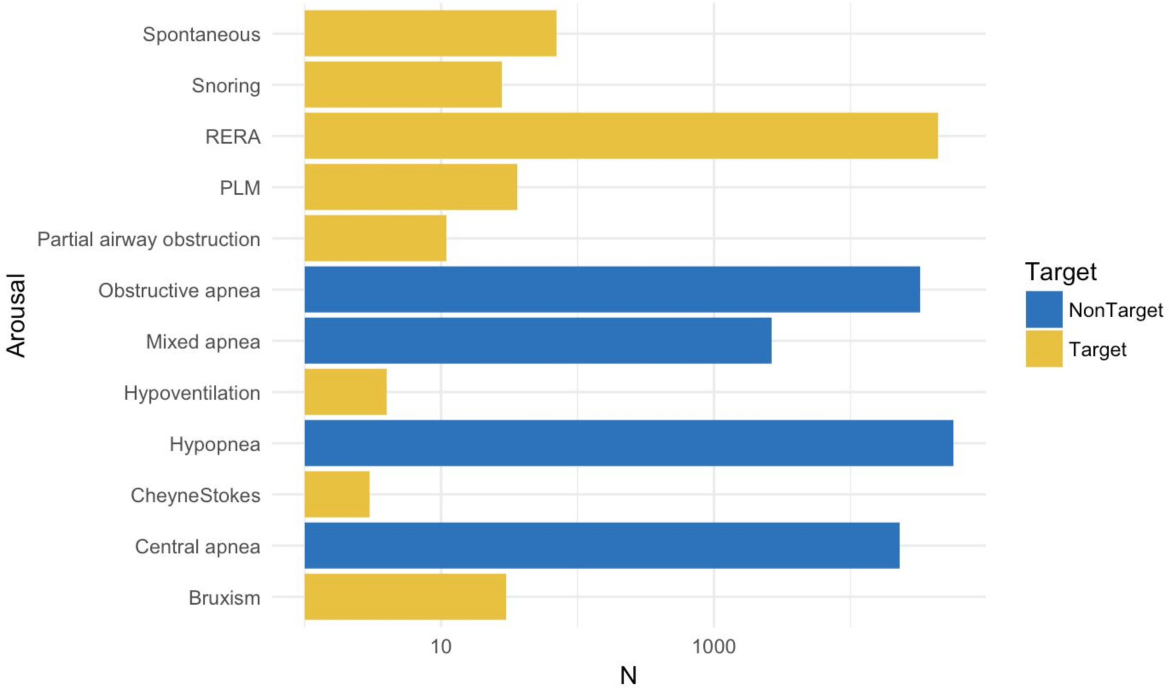
the data

- Annotated Polysomnography data and other types of clinical data (such as age and gender) from **1,985 patients.**

Annotations made by clinical staff at Mass General Hospital:

- **arousal:**
  1. **arousal**
  2. **non-arousal**
- **Sleep stages:**
  1. Wake
  2. Non-REM-I
  3. Non-REM-II
  4. Non-REM-III
  5. REM
  6. un-defined
- **arousal type**

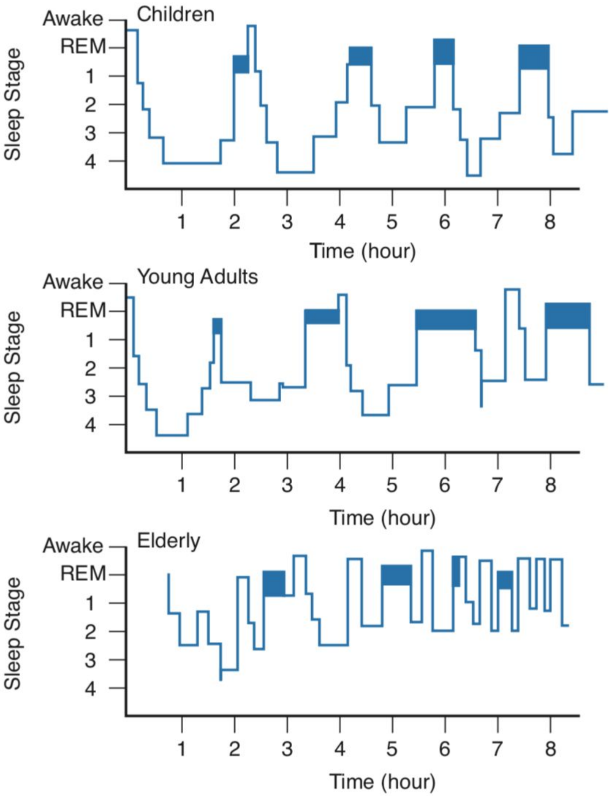
Counts Arousal Types in the Entire Physionet Challenge Dataset on a Log Scale



Patient demographic data such as age and gender were provided. Arousals are shown to increase in age. (ADD CITATION) We did not utilize any of the patient demographic data in designing out research project, but it may be interesting to investigate the relationship with age and the polysomnographic data.

Annotations made by clinical staff at MGH:

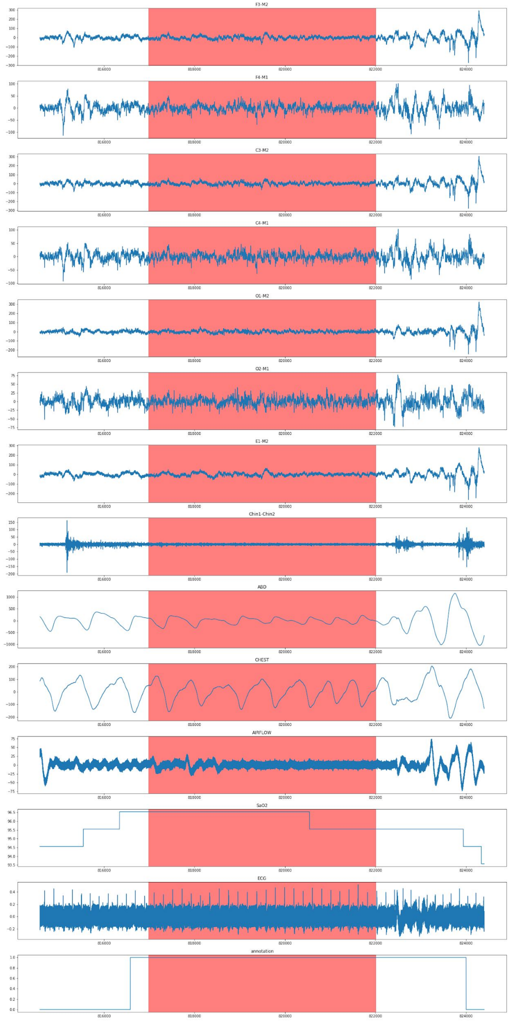
- arousal:
  1. arousal
  2. non-arousal
- **Sleep stages:**
  1. Wake
  2. Non-REM-I
  3. Non-REM-II
  4. Non-REM-III
  5. REM
  6. un-defined
- arousal



Normal Sleep Cycles

Koda-Kimble and Young's Applied therapeutics. 10th ed. 714-718.

Data Exploration of polysomnographic data



Rera type arousal which lasted approximately 25 seconds

The challenge required 2 seconds before the arousal to be labeled positive and 10 seconds after which can be seen in the last row.

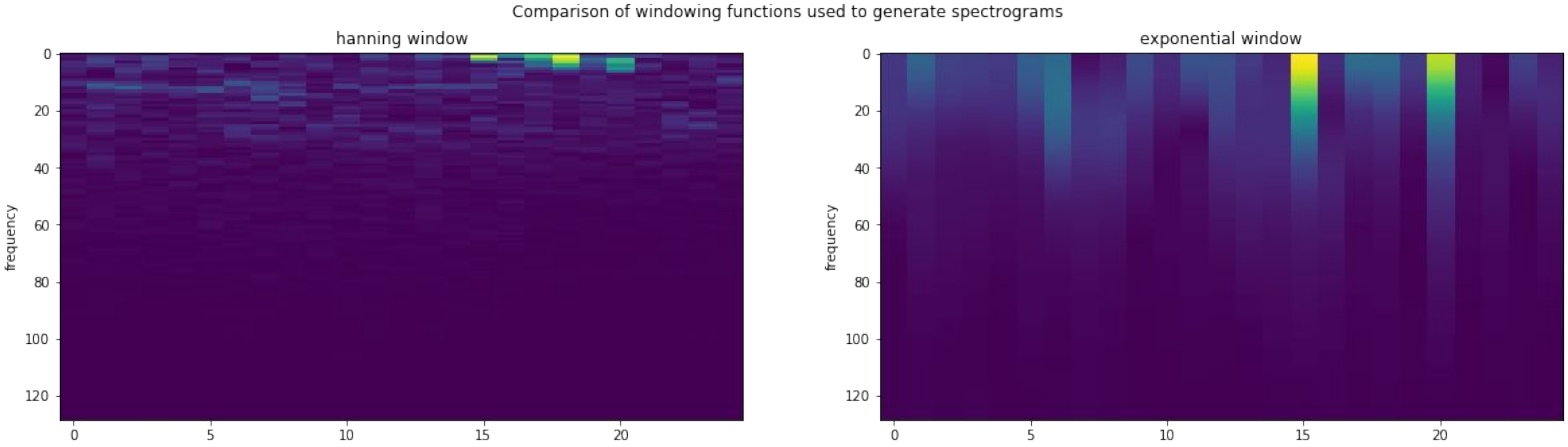
TABLE I

Description of Variables

Var	unit	Type of Test
SaO2	%	Oxygen saturation
ABD	V	Electromyography, a measurement of abdominal movement
CHEST	V	Electromyography, measure of chest movement
Chin1-Chin2	V	Electromyography, a measure of chin movement
AIRFLOW	V	A measure of respiratory airflow
ECG	mV	Electrocardiogram, a measure of cardiac activity
E1-M2	V	Electrooculography, a measure of left eye activity
O2-M1	V	Electroencephalography, a measure of posterior activity
C4-M1	V	Electroencephalography, a measure of central activity
C3-M2	V	Electroencephalography, a measure of central activity
F3-M2	V	Electroencephalography, a measure of frontal activity
F4-M1	V	Electroencephalography, a measure of frontal activity
O1-M2	V	Electroencephalography, a measure of posterior activity

Short Time Fourier-Transformation Convolutional Neural Network

Using Short Time Fourier Transforms to convert signals into images



There are multiple methods of applying a window function in fourier transformations. These spectrograms depict two different window functions. We found that the hanning window (left) provided better results than the exponential window (right). The hanning window was shown to provide more variation in lower frequencies, prompting us to explore this method more because the target-arousal signals might appear more.

Methods Overview for Final Model

Data Exploration and Pre-processing

Convert the signals into spectrograms and normalize each spectrogram as a whole. Spectrograms are then sliced into equal slices

Splitting of Data

180: 10: 10
training set: validation set: test set

Hyperparameters: snapshots during validation set to choose best model

Architecture and Parameters

2*13 CNNs in the first layer
2 CNNs in the second layer
1 dense layer
1 output layer

Window Selection

The mean label for each spectrogram is used to determine whether a window is chosen. Windows with a mean lower than a cutoff still have a random chance of being used for training and validation.

Software Used

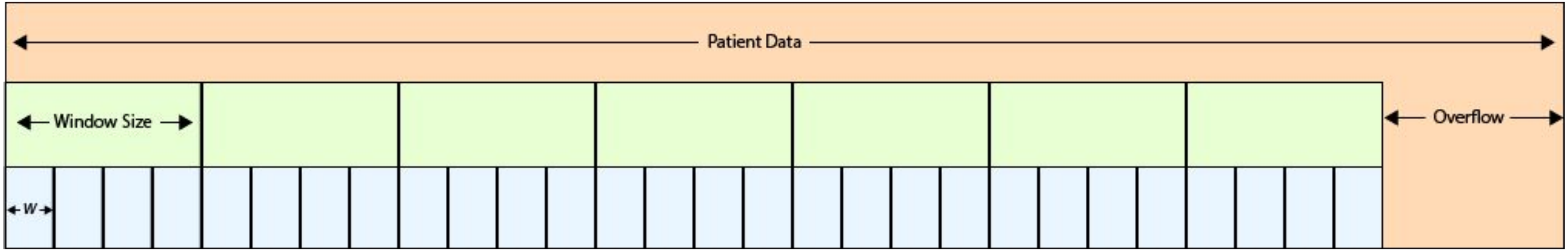
Python 3.5 programming language with PyTorch framework for Deep Learning and scikit-learn library.

R used for statistical analysis, specifically the packages MLmetrics and ROCR.

Evaluation Metrics

All regions classified as -1 are dropped (not cored). The gross AUPRC (area of precision recall curve), F-score, recall (sensitivity) and precision. AUCROC (area of receiver operating curve) and Accuracy (specificity) also determined. We also provide AUPRCs of individual participants.

Data Augmentation



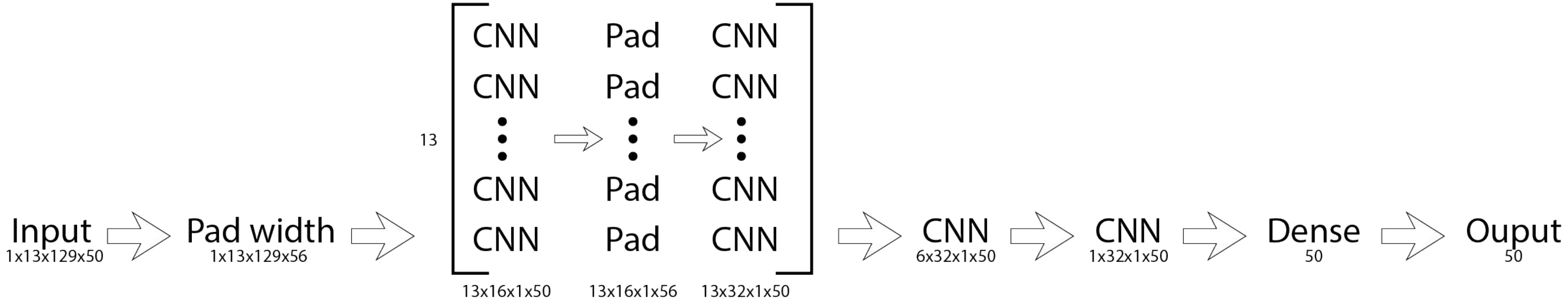
To prepare time contiguous data (you can think of time as moving from left to right), we split the data into windows and then using each window generate a spectrogram per signal.

As the spectrogram will further divide each window into bins we used the following formula to ensure that the window sizes match the bin dimensions

$$AW = WS + (WF - (WS + WF)_{mod} WF)$$

WS: Window Size;
AW: Adjusted Window
WF: Windowing Function Width

Hyperparameters and Window Selection for final architecture



Hanning Window (HW) = 256
Temporal bins = 50

Model Parameters: 317,731

Adam Optimizer:

- Learning rate: 1e-3
- Weight decay: 1e-3

Binary Cross Entropy Loss

scoring:

The PhysioNet Challenge emphasizes predicting target arousals, the class 1 regions, which is the unbalanced class in our case and places importance on the area under the precision-recall curve (PRAUC). We follow their standards in evaluating the model’s performance.

$$R_j = \frac{\text{number of arousal samples with predicted probability } (j/1000) \text{ or greater}}{\text{total number of arousal samples}}$$

$$P_j = \frac{\text{number of arousal samples with predicted probability } (j/1000) \text{ or greater}}{\text{total number of samples with predicted probability } (j/1000) \text{ or greater}}$$

$$AUPRC = \sum_i P_j(R_j - R_{j+1})$$

gross **AUPRC (i.e., for each possible value of *j*, the precision and recall are calculated for the *entire test database*)**

- not an average of the AUPRC for each record

Results

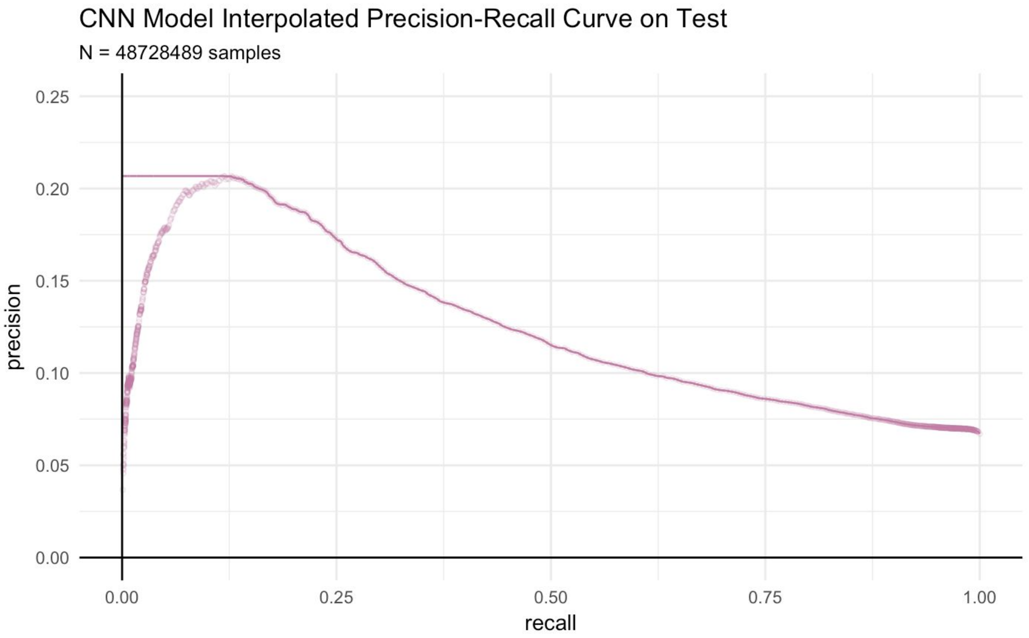
The prediction probabilities from each model and their ground truths were first processed by dropping the classes labeled -1 (regions that are not scored), thus leaving only 0 (N = 45459423) and 1 classes (N = 3269066). Class 1 (target arousal region) represented 6.7 percent of the test samples.

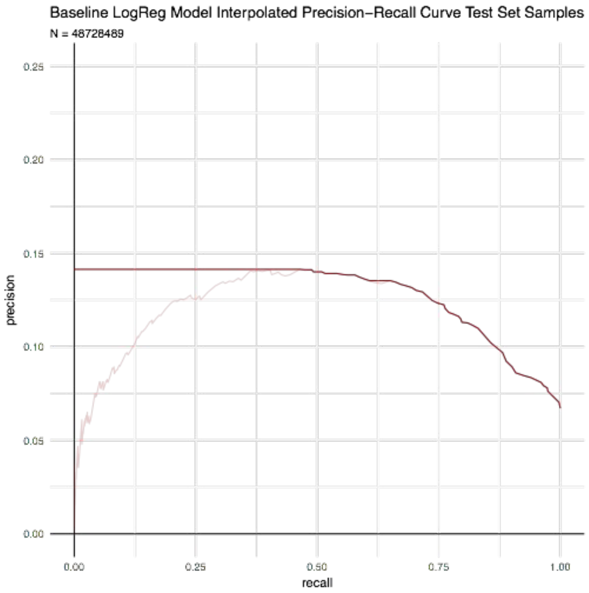
The Accuracy for the model results in a score of 0.9329 with a Area under the ROC curve as 0.651.

For this reason the authors and creators of the challenge are using the Area Under the Precision Recall Curve (AUPRC), to evaluate performance. In particular, the challenge calls for the gross AUPRC (the average of all precision-recall scores for all samples within a test set, not averages of scores samples in each patient).

Applying the classifier at the patient level and combining the time-series data with clinical features based on the patient's profile would be interesting to explore.

Gross Area Under the Precision-Recall Curve CNN Model on our Test Set





interpolated precision-recall curves for the CNN model and the baseline model

	F1 Score	Precision	Recall	AUPRC	AUROC	Accuracy
Baseline	**0.0027**	**1.000**	**0.0013**	**0.1164**	**#**	**#**
Our Model	**0.0365**	**0.9931**	**0.0186**	**0.1246**	**0.651**	**0.9329**

Although both scores are very low, our model does show a slight increase of + 0.0082 . Additionally, our Fourier Transformation-CNN model had an F1 score of 0.0365, depicting a + 0.0338 from the baseline F1 score of 0.0027.

The maximum F1 was 0.2078 for the CNN model.

thanks!
any questions?

References

- In: Koda-Kimble and Young’s Applied therapeutics. 10th ed. 714-718.
- Ahmed, Imran, and Michael Thorpy. "Clinical features, diagnosis and treatment of narcolepsy." Clinics in chest medicine 31.2 (2010): 371-381.
- PhysioBank, PhysioToolkit, and PhysioNet
- Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng and H. Eugene Stanley. Circulation. 2000;101:e215-e220, originally published June 13, 2000 https://doi.org/10.1161/01.CIR.101.23.e215.
-