ETH zürich

Machine Learning for Biomarker Discovery in Clinical Time Series

Karsten Borgwardt

ETH Zürich, D-BSSE

Berlin, May 20, 2019

Machine Learning and Personalized Medicine

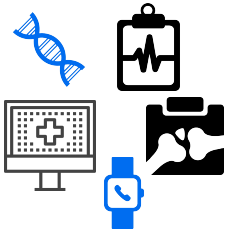
Goals

- Machine Learning tries to detect statistical dependencies in large datasets.

Machine Learning and Personalized Medicine

Goals

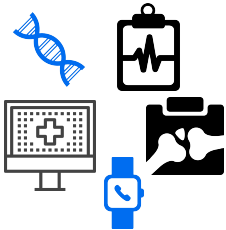
- Machine Learning tries to detect statistical dependencies in large datasets.



Machine Learning and Personalized Medicine

Goals

- Machine Learning tries to detect statistical dependencies in large datasets.



- Personalized Medicine tries to exploit wealth of health data for improved diagnosis, prognosis and therapy decisions, tailored to the properties of each patient.

Machine Learning in Medicine

Key Topics

Machine Learning in Medicine

Key Topics

- Automation of diagnoses

Original Investigation

December 12, 2017

Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer

Babak Ehteshami Bejnordi, MS[1]; Mitko Veta, PhD[2]; Paul Johannes van Diest, MD, PhD[3]; et al

» Author Affiliations | Article Information

JAMA. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

Machine Learning in Medicine

Key Topics

- Automation of diagnoses
- Biomarker discovery

Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth

Marcel Adam Just, Lisa Pan, Vladimir L. Cherkassky, Dana L. McMakin, Christine Cha, Matthew K. Nock & David Brent

Nature Human Behaviour **1**, 911–919 (2017)
doi:10.1038/s41562-017-0234-y
Download Citation

Received: 06 February 2017
Accepted: 04 October 2017
Published online: 30 October 2017

Machine Learning in Medicine

Key Topics

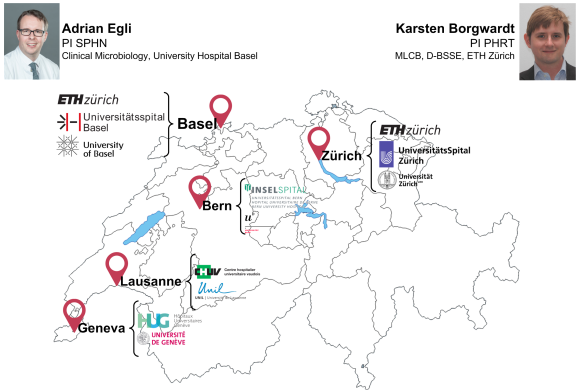
- Automation of diagnoses
- Biomarker discovery
- Biomedical data management

Roche to **buy Flatiron** Health for $1.9 billion to expand cancer care ...
Reuters - 15.02.2018
Roche to **buy Flatiron** Health for $1.9 billion to expand cancer care portfolio ... S) said on Thursday it would **buy** the rest of U.S. cancer data company **Flatiron** Health for $1.9 billion to speed development of cancer medicines and support its efforts to ... Privately held **Flatiron**, backed by Alphabet Inc (GOOGL.

Machine Learning in Medicine

Key Topics

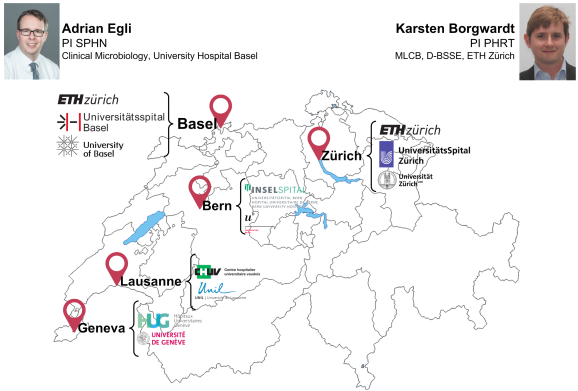
- Automation of diagnoses
- Biomarker discovery
  1. Personalized Swiss Sepsis Study
- Biomedical data management



Machine Learning in Medicine

Key Topics

- Automation of diagnoses
- Biomarker discovery
  1. Personalized Swiss Sepsis Study
- Biomedical data management
  2. Personalized Swiss Sepsis Study



The Need for Biomarkers for Sepsis

Predicting Sepsis

Background: What is sepsis and why is it relevant?

- Sepsis is a life-threatening organ dysfunction, caused by a dysregulated host response to infection (Singer et al., 2016).
- Identification of a bacterial species in blood still takes between 24h and 48h after blood sampling (Osthoff et al., 2017).
- From onset each hour of delayed effective antibiotic treatment increases mortality (Ferrer et al., 2014).

Predicting Sepsis

Background: What is sepsis and why is it relevant?

- Sepsis is a life-threatening organ dysfunction, caused by a dysregulated host response to infection (Singer et al., 2016).
- Identification of a bacterial species in blood still takes between 24h and 48h after blood sampling (Osthoff et al., 2017).
- From onset each hour of delayed effective antibiotic treatment increases mortality (Ferrer et al., 2014).

→ The first hours of sepsis are of critical importance.
→ Currently, when sepsis is detected, organ damage has already progressed.
→ **Detecting and treating sepsis earlier** and better identifying high-risk subgroups could be of highest clinical impact.

The Problem

Data Collection

- Temperature
- Heart rate
- Blood pressure
- Respiratory rate
- O_2 *saturation*

Sepsis

- High mortality
- High morbidity
- High health costs

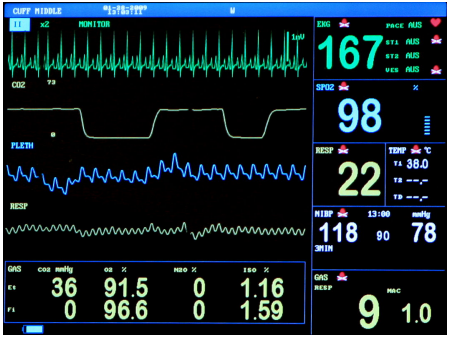


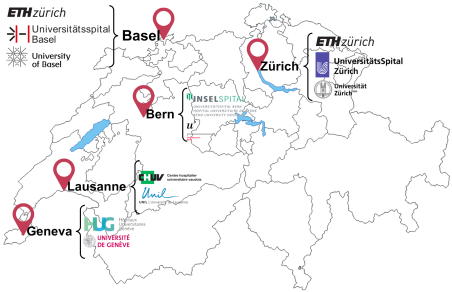
Figure: Monitoring of vital parameters

Personalized Swiss Sepsis Study

- Consortium of 22 research labs and 5 university hospitals in Switzerland
- Goal: Predict sepsis and sepsis-related mortality
- Approach: Integrate clinical data and molecular data for joint biomarker discovery

Adrian Egli
PI SPHN
Clinical Microbiology, University Hospital Basel

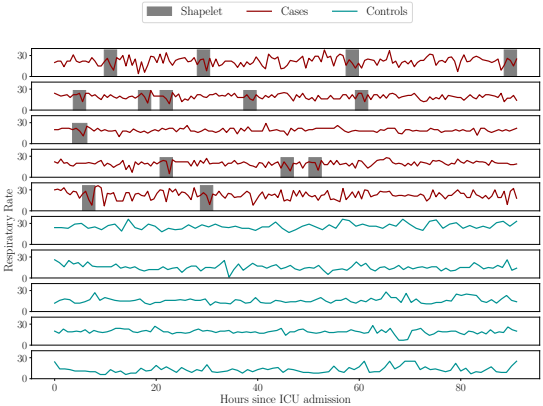
Karsten Borgwardt
PI PHRT
MLCB, D-BSSE, ETH Zürich



- Duration:
  3 years
  (2018-2021)
- Total funding:
  5.3 Million CHF

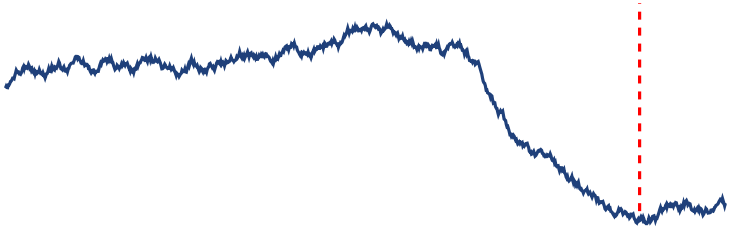
Biomarker Discovery in Longitudinal Data with Machine Learning

Predicting Sepsis

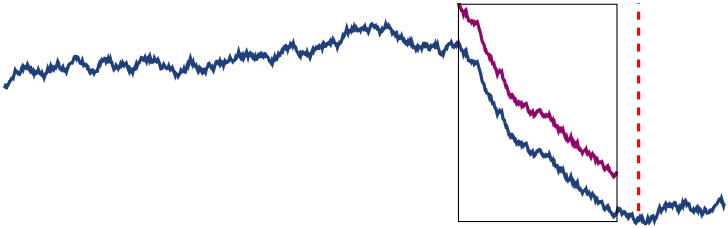


- Goal: To detect patterns in clinical time series that are statistically significantly associated with sepsis (Bock et al., Bioinformatics 2018).

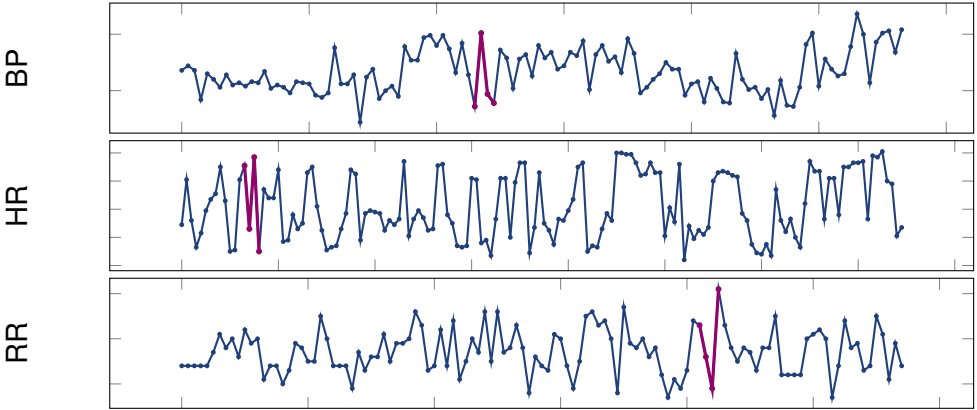
Expressive Subsequences — The Intuition



Expressive Subsequences — The Intuition

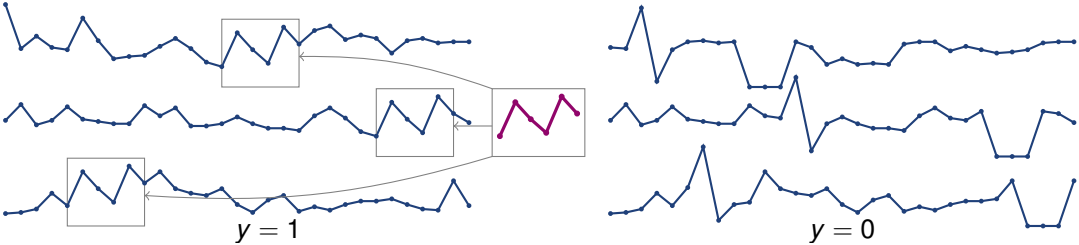


Expressive Subsequences — Reality

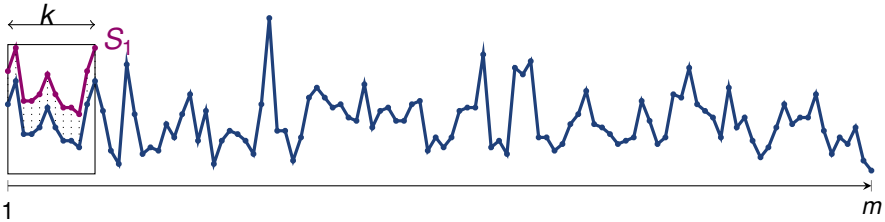


Time Series Shapelets

Shapelet: A subsequence that maximizes predictive power (Ye & Keogh, 2009)

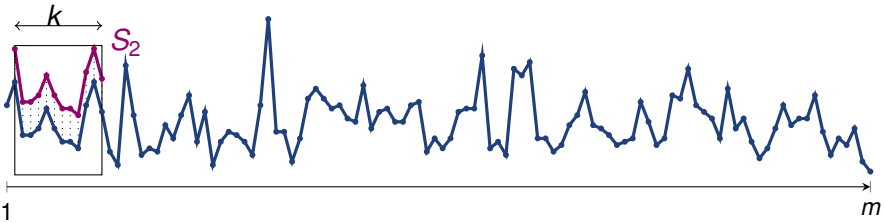


Shapelet Candidate Generation



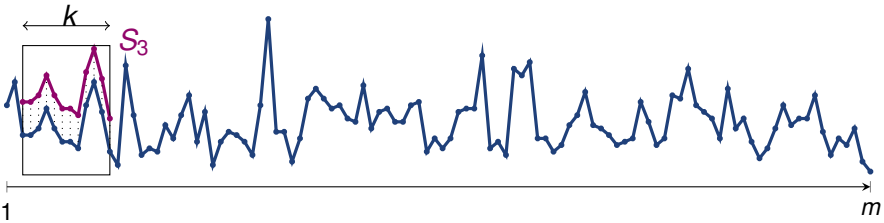
$$\text{Data set } D := \{\mathcal{T}, \mathcal{Y}\}$$
$$\text{Time series } \mathcal{T} := \{T_1, ..., T_n\}$$
$$\text{Labels } \mathcal{Y} := \{y_1, ..., y_n\}$$

Shapelet Candidate Generation



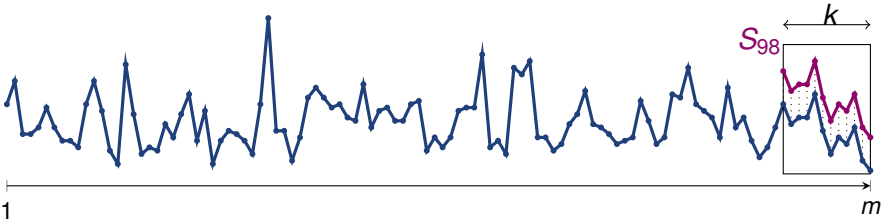
$$\text{Data set } D := \{\mathcal{T}, \mathcal{Y}\}$$
$$\text{Time series } \mathcal{T} := \{T_1, ..., T_n\}$$
$$\text{Labels } \mathcal{Y} := \{y_1, ..., y_n\}$$

Shapelet Candidate Generation



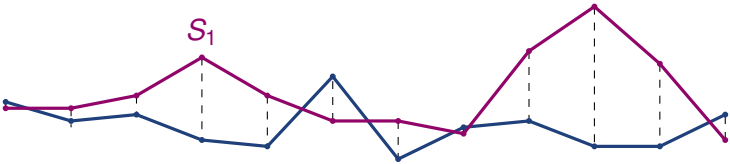
$$\text{Data set } D := \{\mathcal{T}, \mathcal{Y}\}$$
$$\text{Time series } \mathcal{T} := \{T_1, ..., T_n\}$$
$$\text{Labels } \mathcal{Y} := \{y_1, ..., y_n\}$$

Shapelet Candidate Generation



$$\text{Data set } D := \{\mathcal{T}, \mathcal{Y}\}$$
$$\text{Time series } \mathcal{T} := \{T_1, ..., T_n\}$$
$$\text{Labels } \mathcal{Y} := \{y_1, ..., y_n\}$$

Distance between Subsequence and Time Series

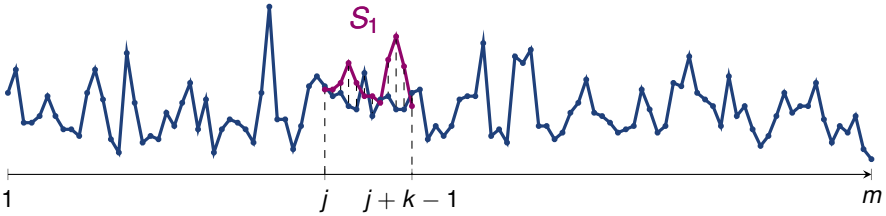


Candidate $S := \{s_1, ..., s_k\}$

Time series $T := \{t_1, ..., t_k\}$

$$dist(S, T) := \sum_{j=1}^{k} (S[j] - T[j])^2$$

Distance between Subsequence and Time Series

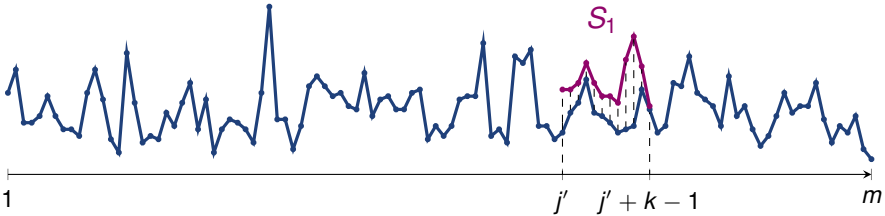


Candidate $S := \{s_1, ..., s_k\}$

Time series $T := \{t_1, ..., t_m\}$

$$dist(S, T) := \min_{j'} \sum_{j=j'}^{j'+k-1} (S[j] - T[j])^2$$

Distance between Subsequence and Time Series

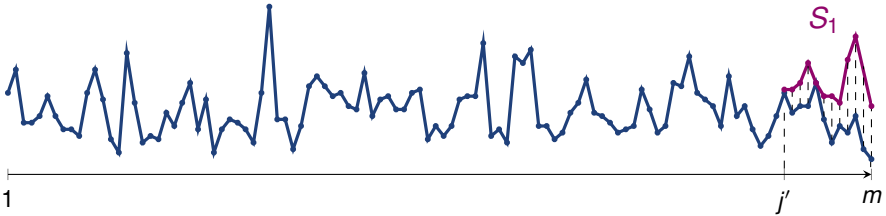


Candidate $S := \{s_1, ..., s_k\}$

Time series $T := \{t_1, ..., t_l\}$

$$dist(S, T) := \min_{j'} \sum_{j=j'}^{j'+k-1} (S[j] - T[j])^2$$

Distance between Subsequence and Time Series



Candidate $S := \{s_1, ..., s_k\}$

Time series $T := \{t_1, ..., t_m\}$

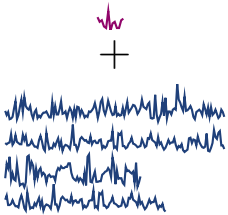
$$dist(S, T) := \min_{j'} \sum_{j=j'}^{j'+k-1} (S[j] - T[j])^2$$

Statistical Association Test

	$\mathcal{I} = 1$	$\mathcal{I} = 0$	
$y = 1$	a_S	b_S	n_1
$y = 0$	d_S	c_S	n_0
	r_S	q_S	n

1. Define a pattern indicator variable $\mathcal{I}$
2. Choose a *test statistic* (e.g. χ^2) and a significance threshold α
3. Evaluate the statistical test to obtain a p-value
4. If $p < \alpha$, the pattern and the outcome are *statistically significantly associated*

Our method



$$\xrightarrow{\theta_1 := dist(S, T_1)}$$
$$\xrightarrow{\theta_2 := dist(S, T_2)}$$
$$\vdots$$
$$\xrightarrow{\theta_n := dist(S, T_n)}$$

$$\begin{bmatrix} dist(S, T_1) \\ dist(S, T_2) \\ \vdots \\ dist(S, T_n) \end{bmatrix}$$

Contingency table for shapelet S and threshold θ_1

	$dist(S,T) \leq \theta_1$	$dist(S,T) > \theta_1$	
$y = 1$	a_S	b_S	n_1
$y = 0$	d_S	c_S	n_0
	r_S	q_S	n

$\vdots$

Multiple Hypothesis Testing

- 100 time series of length 100
  - $\approx 10^4$ patterns of size 5, with 100 possible thresholds
  - $\approx 10^6$ statistical tests
- with $\alpha = 0.05 \Rightarrow 50,000$ false positives

Multiple Hypothesis Testing

- Bonferroni correction:
  $\delta_{\text{bon}} = \frac{\alpha}{\text{number of all hypotheses}} = \frac{0.05}{10^6} = 5 \times 10^{-8}$
- Tarone correction: $\delta_{\text{tar}} = \frac{\alpha}{\text{number of } \textit{testable} \text{ hypotheses}}$

Contingency table with fixed margins (black)

	$S = 1$	$S = 0$	
$y = 1$	a_S	b_S	n_1
$y = 0$	d_S	c_S	n_0
	r_S	q_S	n

Minimum attainable *p*-value

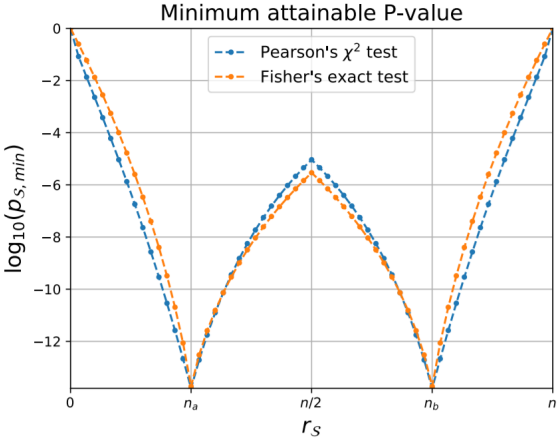


Figure: Minimum attainable *p*-value as a function of pattern frequency r_S (Terada et al., 2013)

S3M Algorithm

Key steps

- **Goal:** Determine all testable and all significant shapelets
- **Naive strategy:** Consider all shapelets and all possible binarization thresholds
- **More efficiently:** Prune all thresholds that lead to too frequent or too infrequent shapelets
- **Even more efficiently:** When computing the number of occurrences of a shapelet, stop once you have reached the frequency that makes it untestable.

Biomarker Discovery in ICU data with S3M

Data Selection

MIMIC-III Data Set (Johnson et al., 2016)

Exclusion Criteria

- Patient $<$ 15 y/o
- Missing chart values
- Patients logged with CareVue

Vital Signs (First 75 hours)

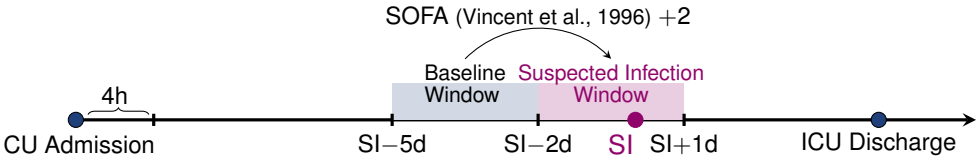
- Heart Rate
- Respiratory Rate
- Systolic Blood Pressure

Cases	355
Controls	355

Data Selection

Case Definition

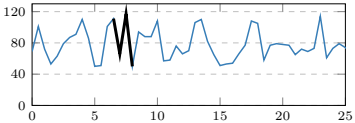
Sepsis-3 definition by Seymour et al. (2016)



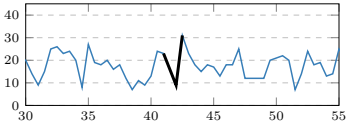
Control Definition

- Neither SI nor SOFA score increase
- Only SI or only SOFA score increase

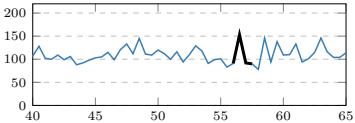
Results: Most significant shapelets



(a) Heart Rate
($p = 7.91 \times 10^{-18}$)



(b) Respiratory Rate
($p = 5.42 \times 10^{-20}$)



(c) Systolic Blood Pressure
($p = 8.27 \times 10^{-13}$)

- Heart Rate : Long term HRV might indicate sepsis
- Respiratory Rate: Sudden drop into abnormal regime (Kellett et al., 2017) with sharp increase
- Systolic Blood Pressure: Characteristic spike

Software development

S3M — Easy to install and use

Installation (www.tiny.cc/s3m)

macOS

```
$ brew install BorgwardtLab/mlcb/s3m
```

Debian and Ubuntu

```
$ sudo apt install s3m-latest.deb
```

Arch Linux

```
$ pacaur -S s3m
```

Docker

```
$ docker build -t s3m_container -f code/cpp/packages/DOCKER/Dockerfile .
```

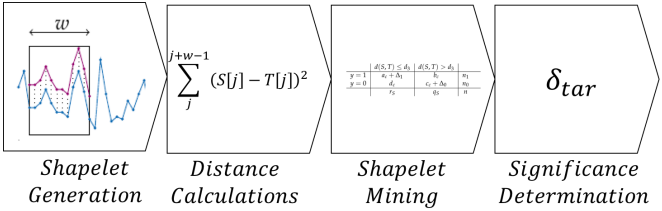
Usage

```
$ s3m -i data/example/synthetic.csv -m 15 -o results/example.json
```

S3M — Output

```
2018-Jun-14 11:12:54.472417: Loading input from data/example/synthetic.csv
2018-Jun-14 11:12:54.501429: Read 100 time series from training input file
2018-Jun-14 11:12:54.501512: Extracting shapelets with length [15:15]
2018-Jun-14 11:12:54.501532: n = 100, n1 = 50
2018-Jun-14 11:12:54.503180: Obtained 3600 candidate shapelets
2018-Jun-14 11:12:54.503207: Maximum length of input time series is 50
2018-Jun-14 11:12:54.503225: Window size correction factor is 1
2018-Jun-14 11:12:54.503264: Naive Bonferroni correction factor is 2.75028e-10
--------------------------------------------------
--------------------------------------------------
--------------------------------------------------
FWER               = 0.0056
Tarone             = 2.7e-08
Testable patterns  = 205200
--------------------------------------------------
0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************
2018-Jun-14 11:12:57.452131: Only keeping normal $p$-values
2018-Jun-14 11:12:57.502777: Detected 407 significant shapelets
```

Conclusion

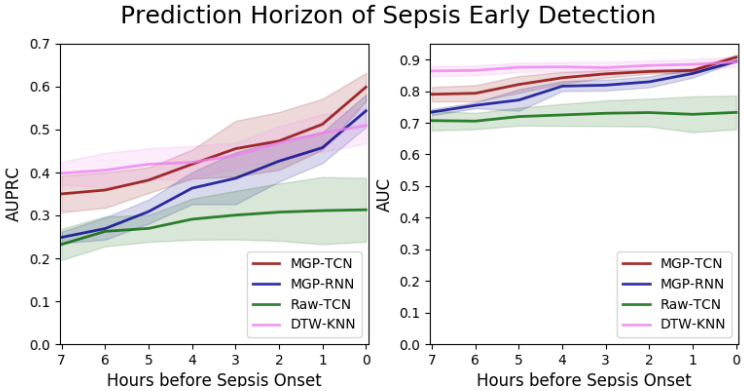


1. Novel association mapping algorithm for time series shapelet mining
2. An original contingency table based pruning criterion to improve efficiency
3. Statistically significant shapelets are interpretable and potentially clinically relevant
4. Accessible and usable at *tiny.cc/s3m*
5. Extendable to dependence between tests (Llinares-Lopez et al., 2015), categorical covariates (Papaxanthos et al., 2016), new approaches to GWAS (Llinares-Lopez et al., 2015, 2017), and available in software package CASMAP (Llinares-Lopez et al., 2019)

Outlook

Time Series Classification

- Classify the entire time series instead of finding motifs (Moor et al., arXiv 2019)



Data Mining in the Life Sciences

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.
- Data growth in three dimensions will pose extreme new challenges in Data Mining in Genetics and Medicine:
  - Population-scale datasets of individuals
  - Life-long recordings of health state
  - Highest-resolution information of the health state

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.
- Data growth in three dimensions will pose extreme new challenges in Data Mining in Genetics and Medicine:
  - Population-scale datasets of individuals
  - Life-long recordings of health state
  - Highest-resolution information of the health state
- How to mine (handle and use) this data?

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.
- Data growth in three dimensions will pose extreme new challenges in Data Mining in Genetics and Medicine:
  - Population-scale datasets of individuals
  - Life-long recordings of health state
  - Highest-resolution information of the health state
- How to mine (handle and use) this data?
- Many branches of the Life Sciences face very similar or analogous problems.

Data Mining in Genetics, Medicine and the Life Sciences

Outlook

- Automation, biomarker discovery, and biomedical data management will remain key research topics.
- Data growth in three dimensions will pose extreme new challenges in Data Mining in Genetics and Medicine:
  - Population-scale datasets of individuals
  - Life-long recordings of health state
  - Highest-resolution information of the health state
- How to mine (handle and use) this data?
- Many branches of the Life Sciences face very similar or analogous problems.

Plenty of opportunities for Data Mining in the Life Sciences

Thank you

- Sponsors: ERC-backup Scheme of Swiss National Science Foundation, Krupp-Stiftung, European Union (MSCA), SPHN/PHRT

References I

- L. Ye, E. Keogh (ACM, 2009), pp. 947–956.
- C. W. Seymour, *et al.*, *JAMA* **315**, 762 (2016).
- A. E. W. Johnson, *et al.*, *Scientific Data* **3** (2016).
- J. Kellett (Springer, 2017), pp. 63–85.
- C. Bock, *et al.*, *Bioinformatics (Oxford, England)* **34**, i438 (2018).
- C. E. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilita* (1936). Published: (Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze. 8) Firenze: Libr. Internaz. Seeber. 62 S. (1936).
- R. Ferrer, *et al.*, *Critical Care Medicine* **42**, 1749 (2014).
- T. F. Mackay, J. H. Moore, *Genome Medicine* **6**, 42 (2014).
- T. A. Manolio, *et al.*, *Nature* **461**, 747 (2009).
- S. Lee, *et al.*, *The American Journal of Human Genetics* **95**, 5 (2014).
- F. Llinares-López, *et al.*, *KDD* (2015), pp. 725–734.

References II

- F. Llinares-López, *et al.*, *Bioinformatics* **31**, i240 (2015).
- F. Llinares-López, *et al.*, *Bioinformatics (Oxford, England)* (2017).
- M. Osthoff, *et al.*, *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **23**, 78 (2017).
- L. Papaxanthos, *et al.*, *NIPS*, D. D. Lee, *et al.*, eds. (2016), pp. 2271–2279.
- C. W. Seymour, *et al.*, *JAMA* **315**, 762 (2016).
- M. Singer, *et al.*, *JAMA* **315**, 801 (2016).
- M. Sugiyama, *et al.*, *SIAM Data Mining*, S. Venkatasubramanian, J. Ye, eds. (SIAM, 2015), pp. 37–45.
- R. E. Tarone, *Biometrics* **46**, 515 (1990).
- A. Terada, *et al.*, *Proceedings of the National Academy of Sciences* **110**, 12996 (2013).
- P. H. Westfall, *et al.*, *Biometrics* **49**, 941 (1993).

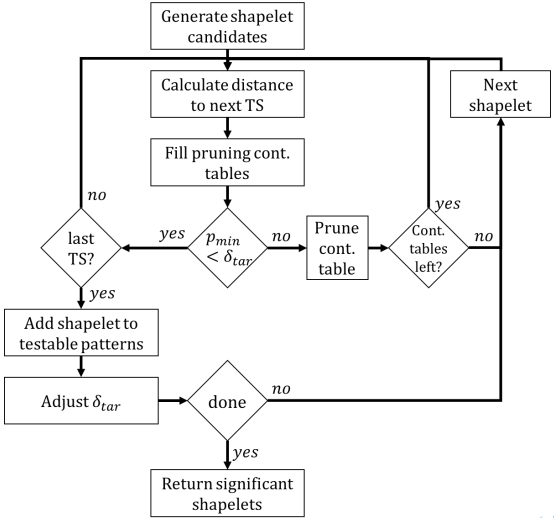
SOFA Score — Organ Systems

- Respiratory system
  - PaO2/FiO2
- Nervous system
  - Glasgow Coma Scale
- Cardiovascular system
  - Mean arterial pressure **or**
  - administration of vasopressors required
- Liver
  - Bilirubin
- Coagulation
  - Platelets
- Kidneys
  - Creatinine **or**
  - urine output

Predictive accuracy

Vital Sign	S3M	# shapelets	gRSF	# shapelets
Heart rate	0.70	1	**0.74**	3030
Respiratory rate	0.71	1	**0.76**	3406
Systolic blood pressure	**0.75**	1	0.74	971

Algorithm Detailed



Tarone vs. Bonferroni

Vital Sign	S3M	δ_{tar}	gRSF	δ_{bon}
Heart Rate	200	2.51×10^{-10}	0	1.28×10^{-15}
Respiratory Rate	514	4.47×10^{-10}	0	1.33×10^{-15}
Systolic Blood Pressure	58	2.55×10^{-9}	0	4.35×10^{-14}

Update Procedure

$\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathbf{S}\}$, $\widehat{\alpha} = \widehat{\delta}_{\text{tar}} \cdot |\mathcal{S}|$
if $\widehat{\alpha} > \alpha$ **then**
  repeat
    $\widehat{\delta}_{\text{tar}} \leftarrow$ next value from $\mathcal{P}$
    Remove untestable patterns from $\mathcal{S}$
    $\widehat{\alpha} = \widehat{\delta}_{\text{tar}} \cdot |\mathcal{S}|$
  until $\widehat{\alpha} \leq \alpha$
end if
return $\mathcal{S}$, $\widehat{\delta}_{\text{tar}}$

Algorithm 1: `Update Procedure`

Minimum attainable p-value for χ^2

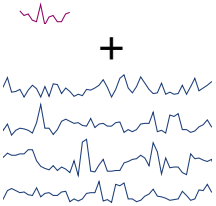
Letting $n_a := \min(n_1, n - n_1)$ and $n_b := \max(n_1, n - n_1)$, we have

$$p_{\min}(r_S) := \begin{cases} 1 - F_{\chi^2}\left((n-1)\frac{n_b}{n_a}\frac{r_S}{n-r_S}\right) & \text{if } 0 \leq r_S < n_a \\ 1 - F_{\chi^2}\left((n-1)\frac{n_a}{n_b}\frac{n-r_S}{r_S}\right) & \text{if } n_a \leq r_S < \frac{n}{2}, \\ 1 - F_{\chi^2}\left((n-1)\frac{n_a}{n_b}\frac{r_S}{n-r_S}\right) & \text{if } \frac{n}{2} \leq r_S < n_b, \\ 1 - F_{\chi^2}\left((n-1)\frac{n_b}{n_a}\frac{n-r_S}{r_S}\right) & \text{otherwise.} \end{cases} \tag{1}$$

where $F_{\chi^2}(\cdot)$ denotes the cumulative density function of a χ^2-distribution with one degree of freedom.

Pruning Shapelet Candidates

Partially filled Contingency Tables



$$\begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}$$

Scenario 1

All remaining T with $y = 1$ will fall in column 1.

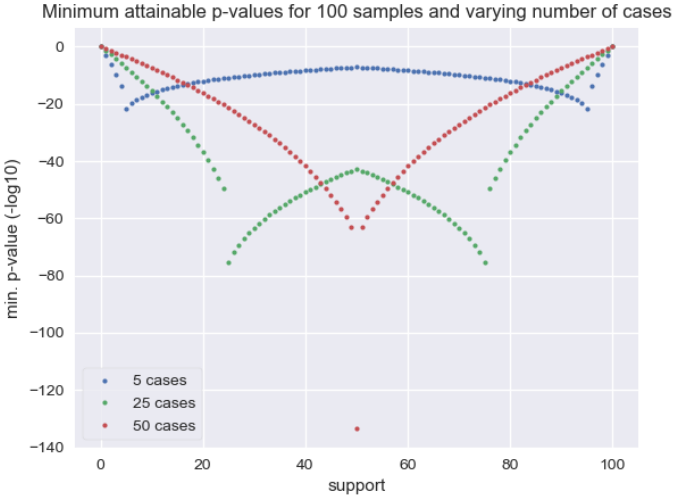
	$d(S,T) \leq d_3$	$d(S,T) > d_3$	
$y = 1$	$a_c + \Delta_1$	b_c	n_1
$y = 0$	d_c	$c_c + \Delta_0$	n_0
	r_S	q_S	n

Scenario 2

All remaining T with $y = 1$ will fall in column 2.

	$d(S,T) \leq d_3$	$d(S,T) > d_3$	
$y = 1$	a_c	$b_c + \Delta_1$	n_1
$y = 0$	$d_c + \Delta_0$	c_c	n_0
	r_S	q_S	n

Minimum Attainable *p*-value and Data Set Balance



S3M Algorithm - key step

Require: Data $\mathcal{D}$, minimal and maximal shapelet length k_{min} and k_{max}, Family-wise error rate α

Ensure: Significance threshold $\widehat{\delta}_{\text{tar}}$

candidates $\mathcal{C} \leftarrow$ GENERATE_CANDIDATES$(\mathcal{D}, k_{min}, k_{max})$, $\widehat{\delta}_{\text{tar}} \leftarrow 1$
for candidate c *in* $\mathcal{C}$ **do**
    pairs $\{(c, \theta)\} \leftarrow$ GET_DISTANCES$(c, \mathcal{D})$
    for (c, θ) that is testable under $\widehat{\delta}_{\text{tar}}$ **do**
        $\mathcal{T} \leftarrow$ Add (c, θ) to $\mathcal{T}$
        Update $\widehat{\delta}_{\text{tar}} := \frac{\alpha}{|\mathcal{T}|}$
        Remove no-longer testable patterns from $\mathcal{T}$
    end for
end for

Algorithm 2: S3M