Centre for
Data Analytics

Insight

Explainable Machine Learning Models for Structured Data

Dr Georgiana Ifrim

georgiana.ifrim@insight-centre.org

(joint work with Severin Gsponer, Thach Le Nguyen, Iulia Ilie)

30 July 2018

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN sfi Science Foundation Ireland

Overview

- **Structured Data**
  - Symbolic Sequences (e.g., DNA, malware)
  - Numeric Sequences (e.g., time series)

- **Explainable Learning Models**
  - Black-Box vs Linear Models with Rich Features

- **SEQL: Sequence Learning with All-Subsequences**
  - Framework for Sequence Classification & Regression

Structured Data: Sequences & Time Series

Many Applications:

- **DNA**

Value	Data points
290.507	AGGGCATCATGGAGCTGTCCAG
679.305	ATCACAATTTTGCCGAGAGCGA
1998.715	GTACACCCCGTTCGGCGGCCCA
447.803	CCTTTAGCCCATCGTTGGCCAA

- **Malware**

Byte sequence

Class	Data points
+1	C7 01 24 04 5F 0E EA DC 00 E9 D6 4A 00 0C 66 89
+1	74 13 BA EF 01 00 06 68 95 14 88 B7 00 0F 0E EA
-1	08 F9 C8 1A 80 C1 8B 48 40 00 89 51 10 B8 04 00
-1	B8 00 00 00 00 50 E8 D8 00 00 00 83 C4 04 53 FF

- **Sensors**

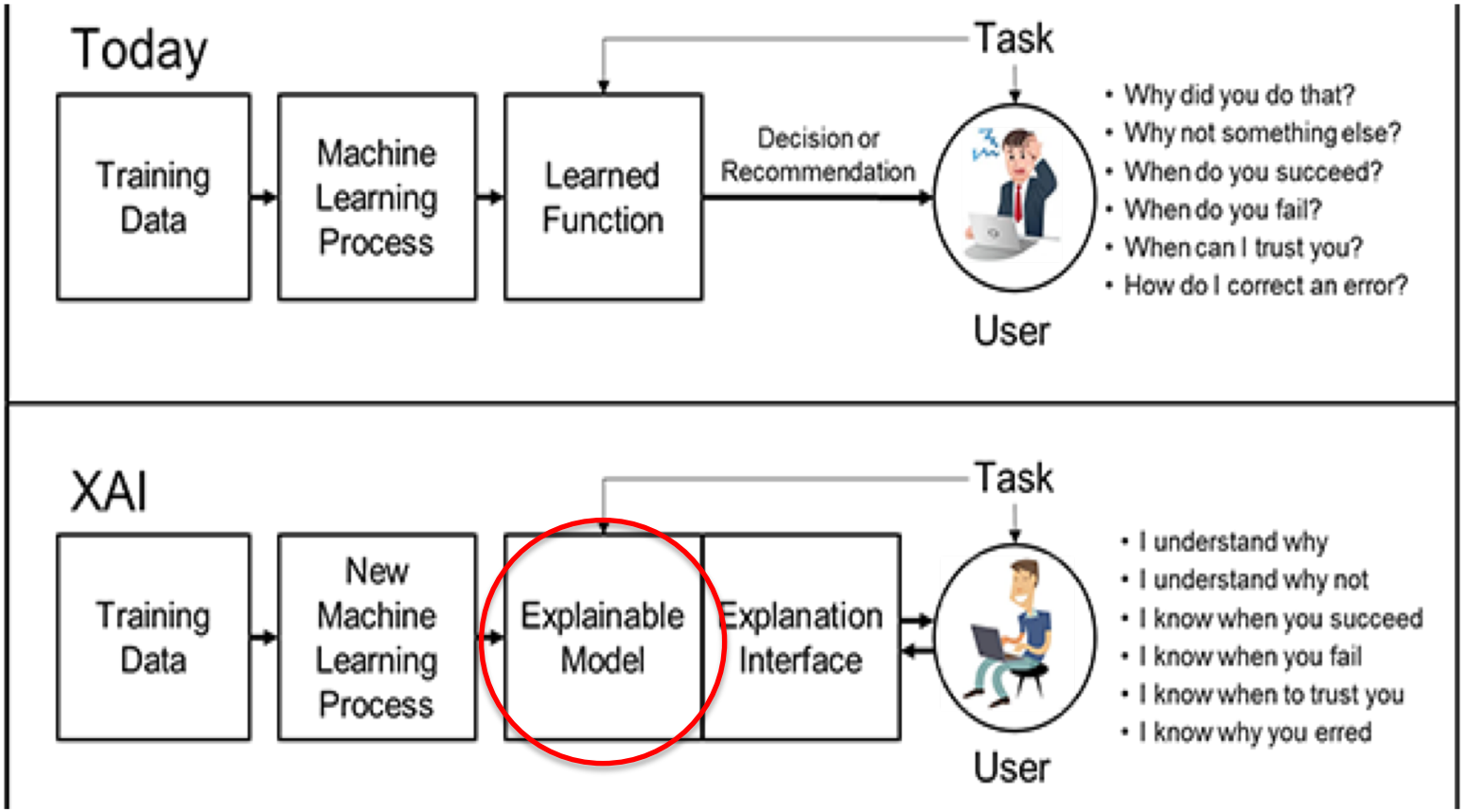
0	-0.26927	-0.26927	-0.26927	-0.26927	-0.26927
1	-0.46887	2.748	1.6263	-0.46887	-0.46887
0	2.2429	-0.39296	-0.39296	-0.39296	-0.39296
0	-0.45836	2.4229	-0.45836	2.5162	1.9876
0	-0.58609	-0.58609	-0.58609	-0.58609	-0.58609
0	1.8657	-0.44769	-0.44769	-0.44769	1.7914
0	1.3541	1.9638	-0.53962	-0.53962	-0.53962

Explainable Machine Learning Models

- **Accuracy & Efficiency:**
  - Many accurate algorithms: e.g., ensembles (Random Forest), Deep Neural Networks; but hard to interpret big, complex models
  - Large volumes of data, need efficient models

- **Interpretability:**
  - White box (linear models) vs black box (deep nets)
  - Interpretable AI is a big deal: Darpa Explainable AI (XAI; 2016), EU GDPR legislation (May 2018)

Darpa Explainable AI (XAI)



[**Source:** http://www.darpa.mil/program/explainable-artificial-intelligence]

SEQL: Sequence Learning with All-Subsequences

Key Idea: Linear Models with Rich Features are Accurate and Interpretable

- Linear models are **interpretable and well understood** (linear regression, logistic regression).
- Linear models with rich features are **accurate** (similar accuracy to ensembles, kernel-SVM, deep nets).
- **Efficiently optimize linear models:** We exploit the structure of a massive feature space (all-subsequences) to quickly select good features.

SEQL: Linear Models for Symbolic Sequences

Solution Approach

SEQL: all-subsequences are candidate features; focus on selecting good features quickly

Score	Sequence
290.5	AGTC**CACAA**GGCTAGGATAGCTA**TCCG**GATCGA
315.1	TATCCTGCAGTACAAG**TCCG**TAATT**CACAA**TCCA
805.6	AGTCCGC**TAGGCT**AGGATAGCTAGCCCGATCGA
799.7	AGCCAAGACCTGAAA**TAGGCT**CCTGAGATACAG
???	CGGGTCGTA**TCCG**CACTGAATATC**TAGGCT**TACG

Goal is to learn a mapping:

$$f : S \rightarrow \mathbb{R}$$

SEQL Model:

Weight	*k*-mer
796.6	TAGGCT
402,5	CACAA
-125.3	TCCG

Linear model (weighted sum of features):

$f(x) = \beta^t x$, with β the feature weights and x the feature vector

SEQL: Linear Models for Symbolic Sequences

Add features iteratively with greedy coordinate descent + branch-and-bound (bound the search for the best feature)

Algorithm 1 Coordinate Descent with Gauss Southwell Selection

1: Set $\beta^{(0)} = 0$
2: **while** termination condition not met **do**
3: Calculate objective function $L(\beta^{(t)})$
4: Find coordinate j_t with maximum gradient value
5: Find optimal step size η_{j_t}
6: Update $\beta^{(t)} = \beta^{(t-1)} - \eta_{j_t} \frac{\partial L}{\partial \beta_{j_t}}(\beta^{(t-1)}) e_{j_t}$
7: Add corresponding feature to feature set
8: **end while**

How do we find coordinate j_t efficiently?

Key Ideas

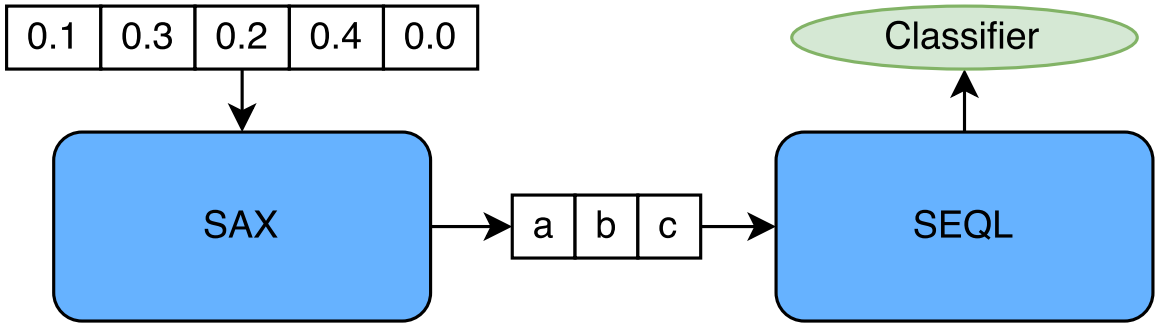
Bound gradient of k-mer using only information about its sub-k-mers.

Example

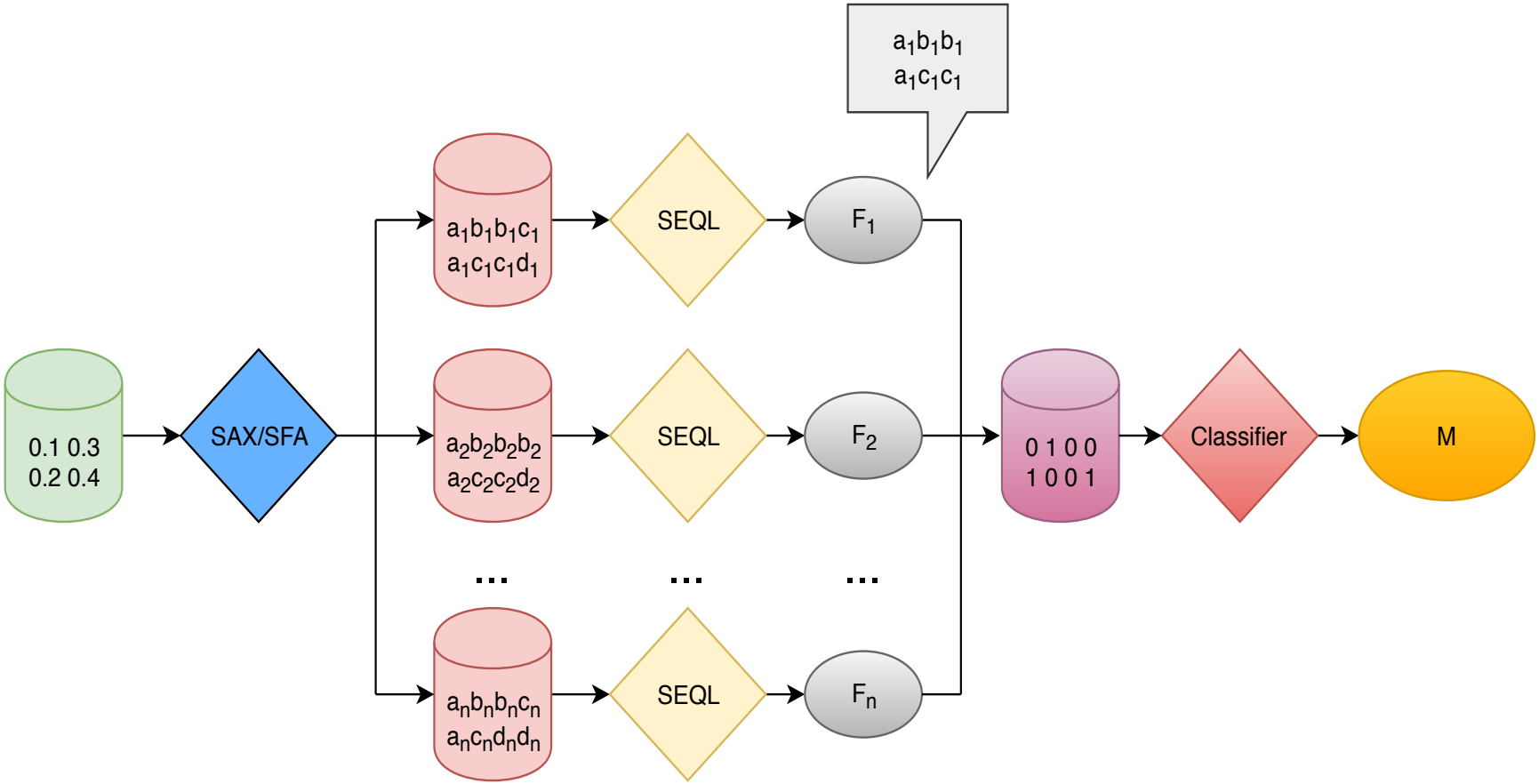
Given: $s_p = "ACT"$
Calculate bound: $\mu(s_p)$
$s_1 = "ACTC"$ -> $gradient(s_1) \leq \mu(s_p)$
$s_2 = "AACT"$ -> $gradient(s_2) \leq \mu(s_p)$
$s_3 = "TACTG"$ -> $gradient(s_3) \leq \mu(s_p)$

SEQL for Time Series Classification

Time Series → Discretisation (SAX, SFA) → Symbolic Sequence → Sequence Learner (SEQL)



SEQL for Time Series Classification



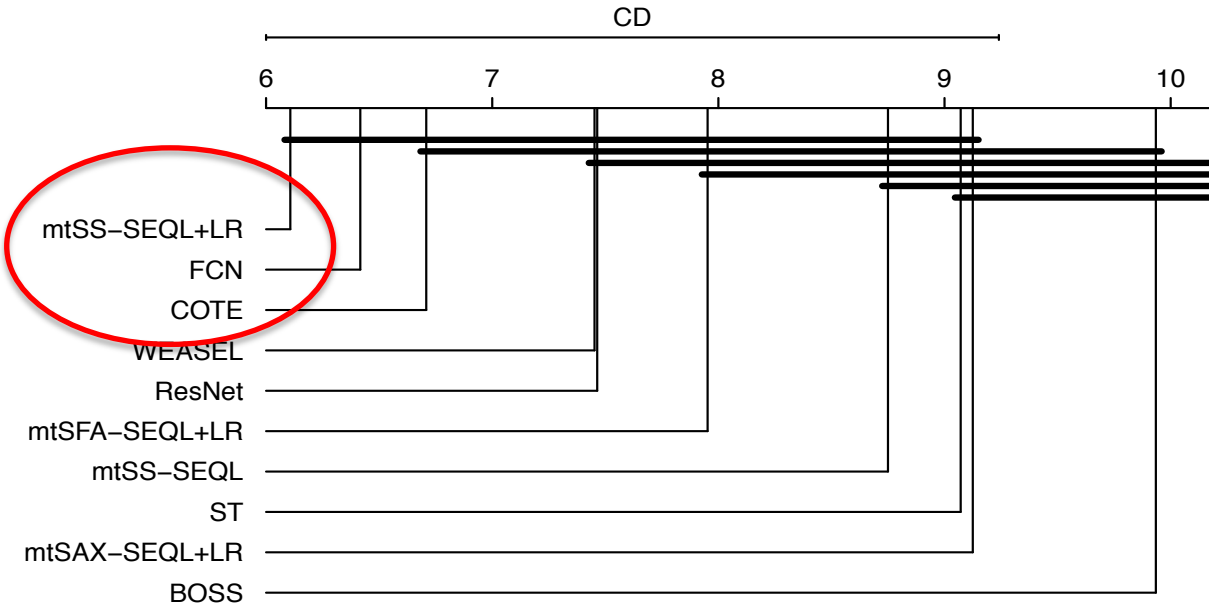
Evaluation on Time Series Classification

Ranking of learning algorithms by Accuracy

UCR Archive (85 TSC datasets: sensors, images, ECG)

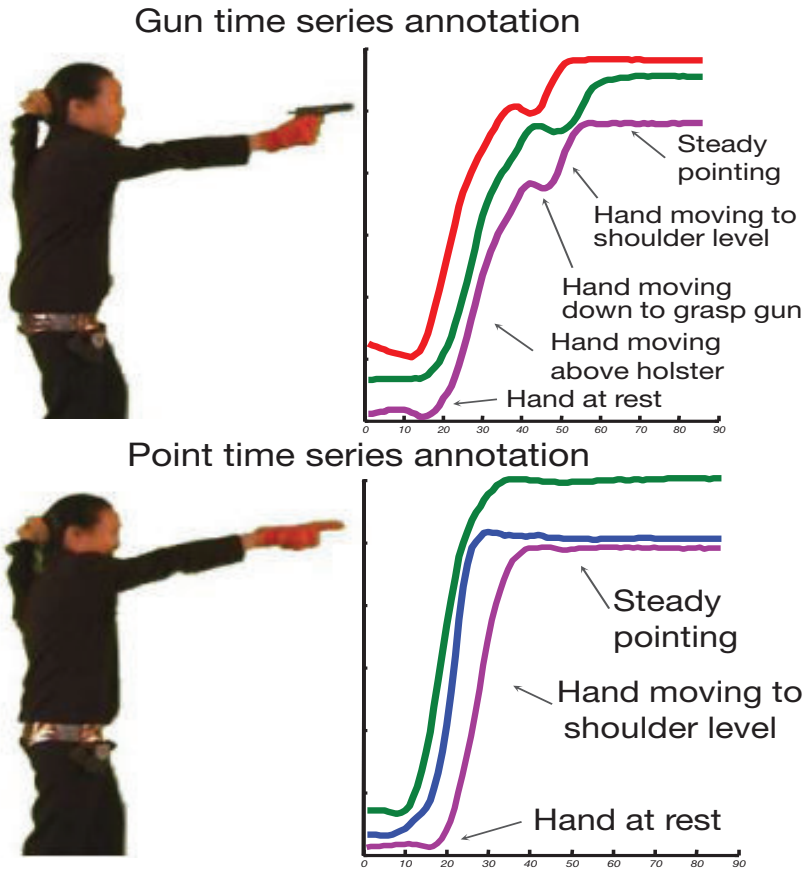
Top-3 models:

1. mtSS-SEQL+LR (our method, a linear model)
2. FCN (deep neural network)
3. COTE (ensemble of 35 classifiers)



Interpretability

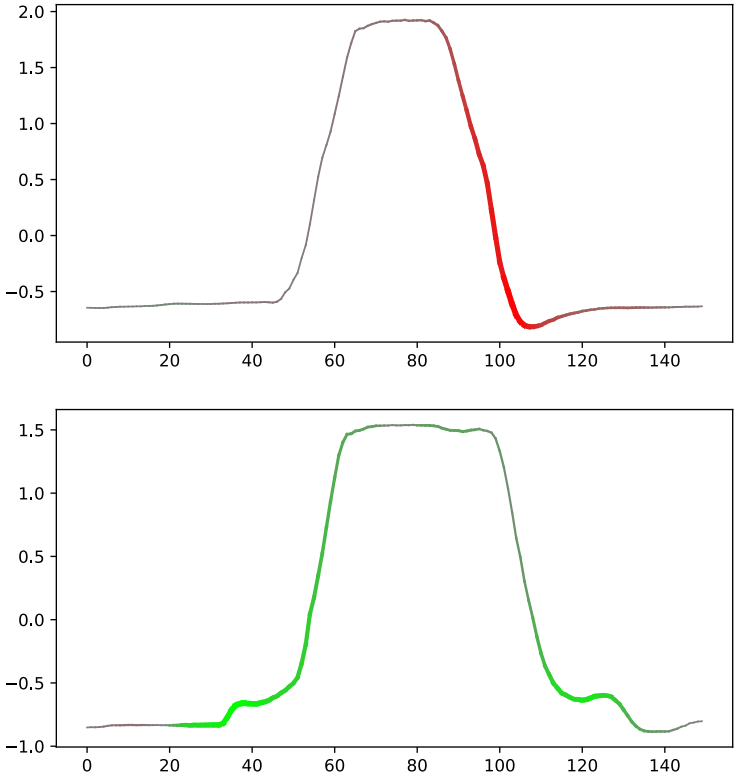
- GunPoint dataset tracking hand movement w/o Gun



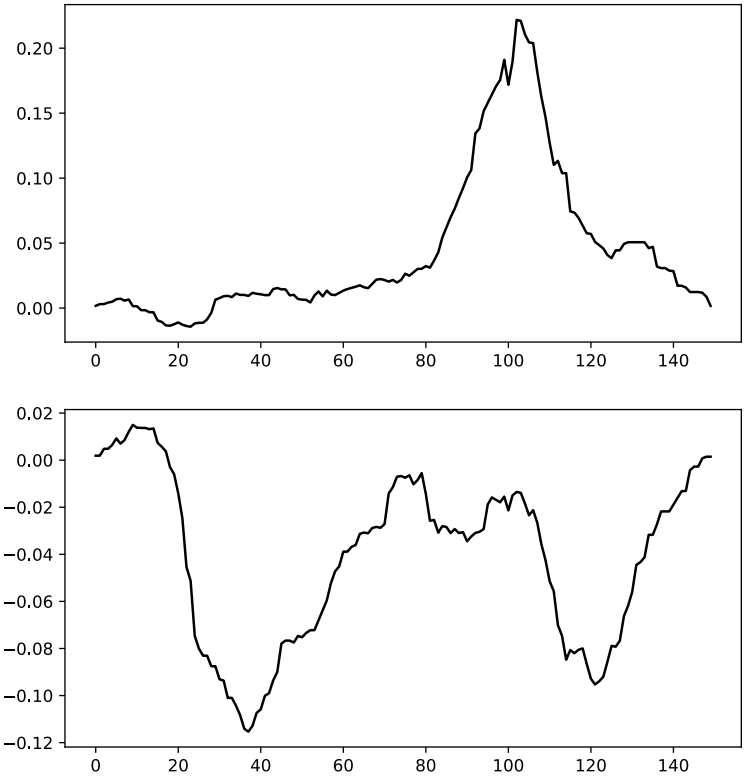
Interpretability

Coefficients	Subsequences
0.065 84	cbaab
0.062 47	db
0.062 23	ddddb
0.062 00	da
0.059 72	bbbbbbbbbbcdddd
−0.053 72	aaaaaabbbb
−0.054 39	bbbbaaaaaa

Point (top) and Gun (bottom)



Salient Region for Classification Decision



Github code for our work: https://github.com/heerme?tab=repositories

Recap SEQL

- Family of machine learning algorithms to train/predict (with) linear models for sequences
- Coordinate descent with Gauss-Southwell feature selection + Branch-and-bound for efficient feature search

- **Sequence Classificati**on (KDD08, KDD11): Logistic loss, l2-SVM loss
- **Sequence Regression** (ECMLPKDD17): Least-squares loss
- **Time Series Classification** (ICDE17): SEQL + SAX discretization

- Future Work:
  - Multi-dimensional Sequences

References

- **[DMKD18, Under review]** T Le Nguyen, S Gsponer, I Ilie, G Ifrim, **Interpretable Time Series Classification using All-Subsequence Learning and Symbolic Representations in Time and Frequency Domains,** DMKD18, 2018.
- **[In prep]** S Gsponer, B Smyth , G Ifrim**, Symbolic Sequence Classification with Gradient Boosted Linear Models,** 2018
- **[ECMLPKDD17]** S Gsponer,, B Smyth, G Ifrim. **Efficient Sequence Regression by Learning Linear Models in All-Subsequence Space**, ECML-PKDD, 2017.
- **[ICDE17]** T Le Nguyen, S Gsponer, G Ifrim, **Time Series Classification by Sequence Learning in All-Subsequence Space,** ICDE, 2017.
- **[PlosOne14]** BP Pedersen, G Ifrim, P Liboriussen, KB Axelsen, MG Palmgren, P Nissen, C. Wiuf, C. Pedersen, **Large scale identification and categorization of protein sequences using structured logistic regression**, PloS one 9 (1), 2014.
- **[KDD11]** G Ifrim, C Wiuf, **Bounded coordinate-descent for biological sequence classification in high dimensional predictor space**, KDD, 2011.
- **[KDD08]** G. Ifrim, G. Bakir, and G. Weikum, **Fast logistic regression for text categorization with variable-length n-grams**, KDD, 2008.