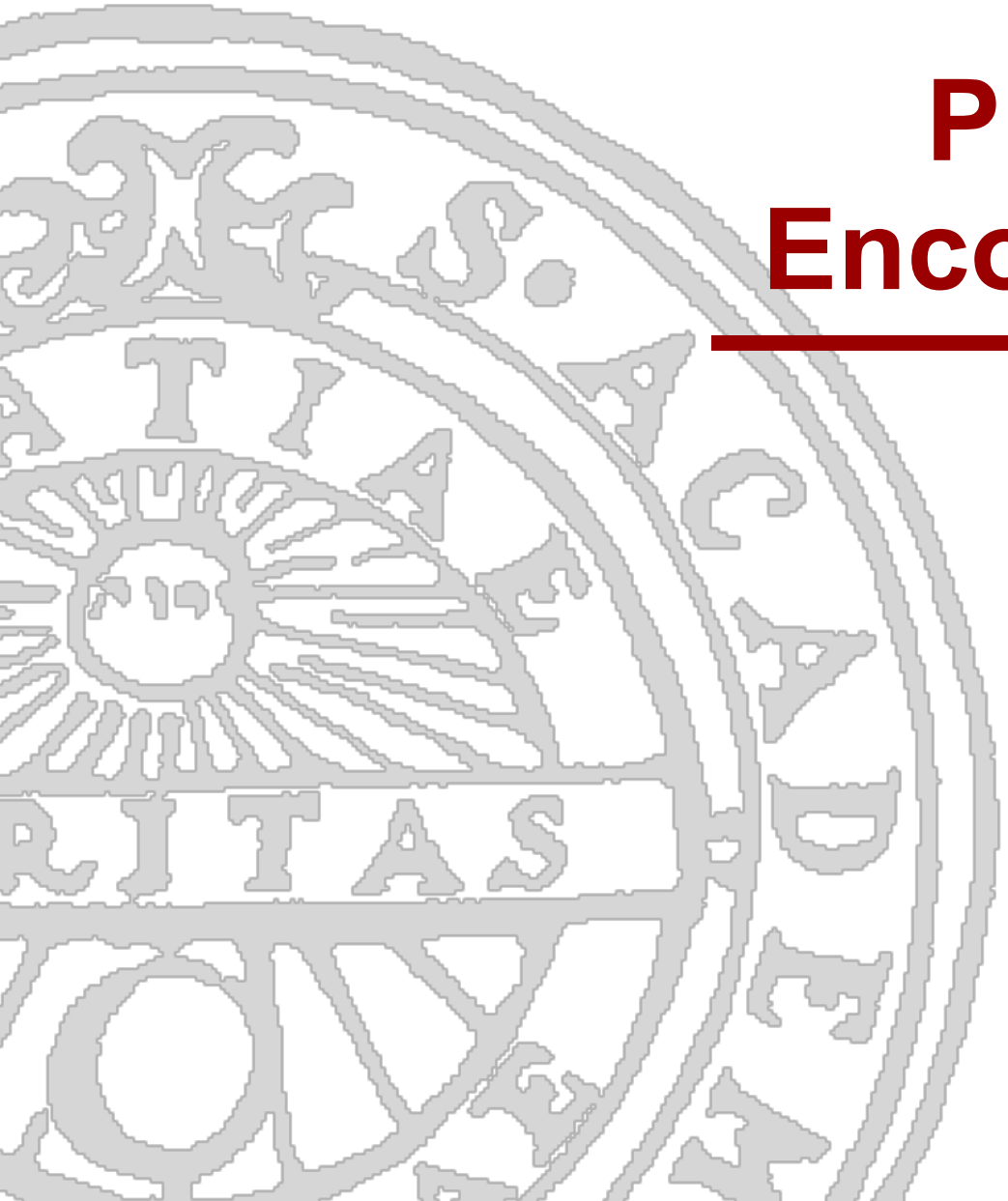


Project Report
Encore-SVMLight

Phuc Vo-Huu

UPPSALA UNIVERSITET

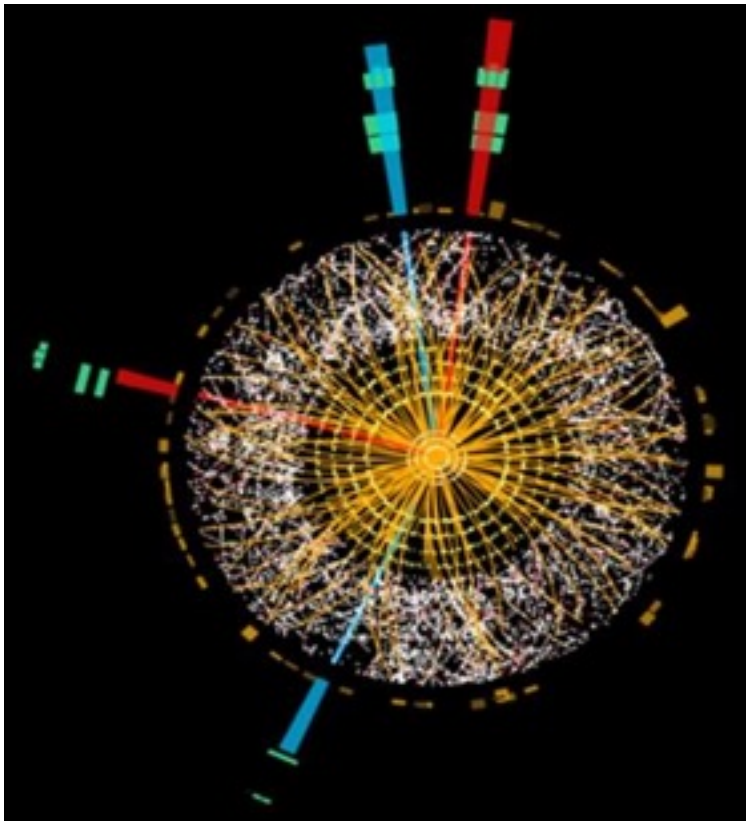
Informationsteknologi

Overview

- Background
- Approach
- Implementation
- Validation

ATLAS Higgs Boson Dataset

- Dataset includes
  - Simulated signals
  - Background events
- Challenge
  - to separate
    - signal
    - background
- Use/develop algorithm
  - to achieve best result



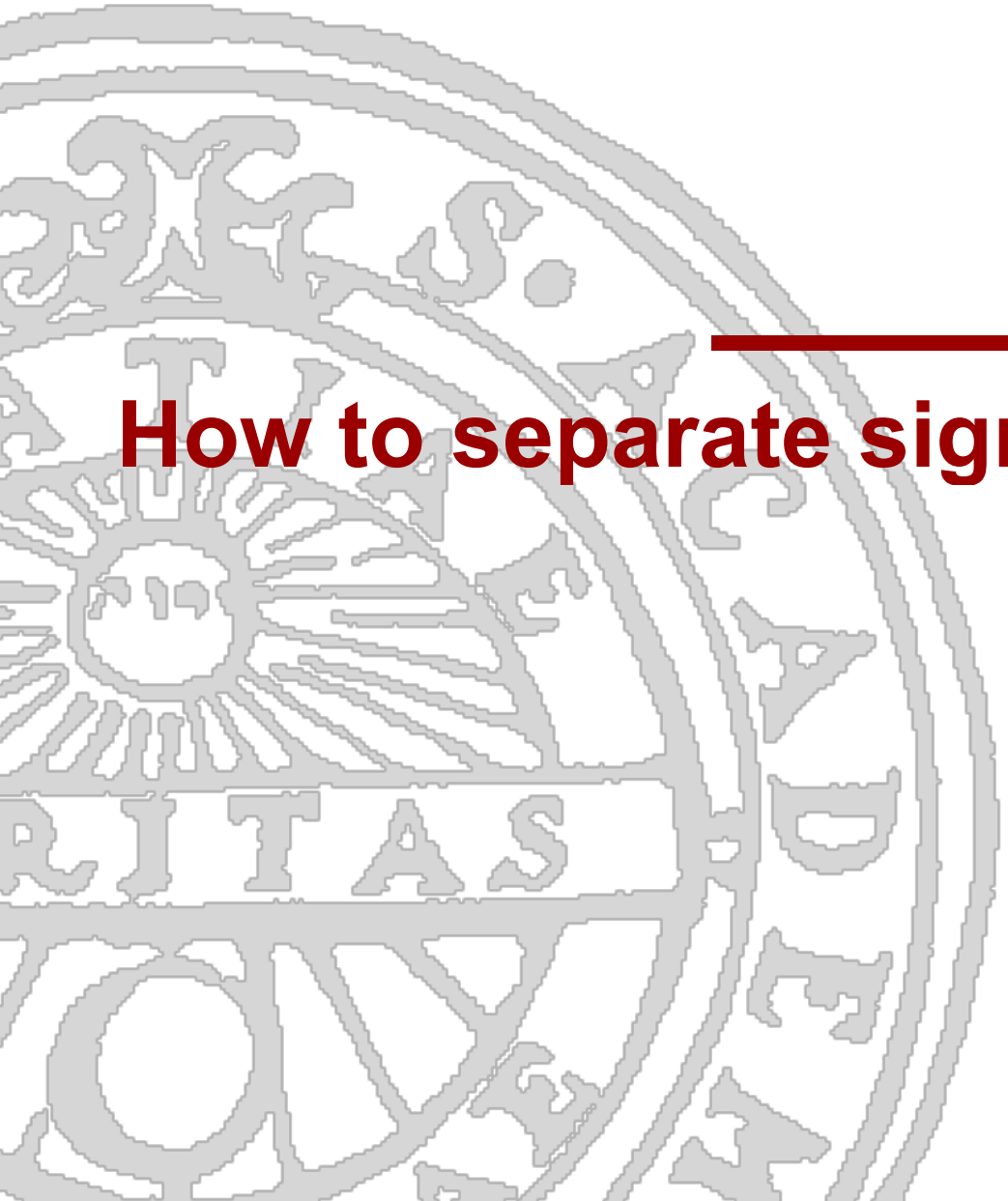
[CERN: http://goo.gl/DSznjP]

ATLAS Higgs Boson Dataset

- Training set
  - 250.000 events, 30 features
  - 1 weight, 1 label, 1 ID
- Testing set
  - 550.000 events
  - 30 features, 1 ID

higg_training_small.csv

```
1  EventId,DER_mass_MMC,DER_mass_transverse_met_lep,DER_mass_vis,DER_pt_h,DER_deltaeta_jet_jet,DER_mass_jet_jet,DER_prodeta_jet_jet,DER_deltar_tau_lep,DER_pt_tot,DER_sum_pt,DER_pt_ratio_lep_tau,DER_met_phi_centrality,DER_lep_eta_centrality,PRI_tau_pt,PRI_tau_eta,PRI_tau_phi,PRI_lep_pt,PRI_lep_eta,PRI_lep_phi,PRI_met,PRI_met_phi,PRI_met_sumet,PRI_jet_num,PRI_jet_leading_pt,PRI_jet_leading_eta,PRI_jet_leading_phi,PRI_jet_subleading_pt,PRI_jet_subleading_eta,PRI_jet_subleading_phi,PRI_jet_all_pt,Weight,Label
2  100000,138.47,51.655,97.827,27.98,0.91,124.711,2.666,3.064,41.928,197.76,1.582,1.396,0.2,32.638,1.017,0.381,51.626,2.273,-2.414,16.824,-0.277,258.733,2,67.435,2.15,0.444,46.062,1.24,-2.475,113.497,0.00265331133733,s
```



How to separate signal/background?

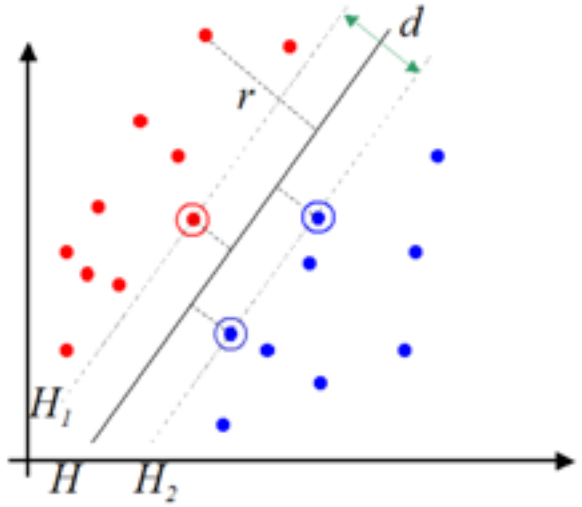
SVM

- Support Vector Machine (SVM)
  - Machine learning via hyperplane separation
  - Define a classifying function
    - maximizes margin d between classes
- Several SVM tools
  - LIBSVM
  - SVMLight
  - SVMTorch

[http://www.svms.org/software.html]

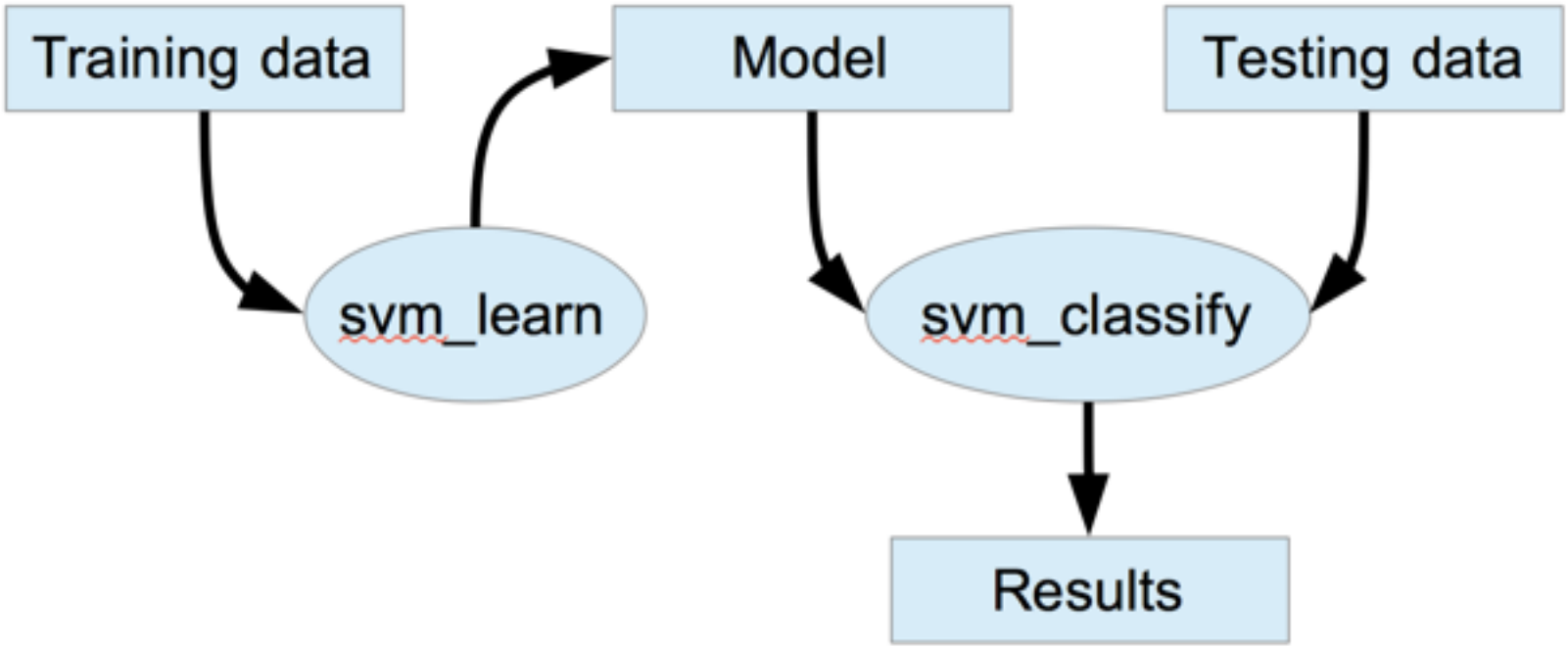
- *SVMLight is chosen*

[Tufts ATLAS Group (Whitehouse, Sliwa)]



SVMLight

- Model



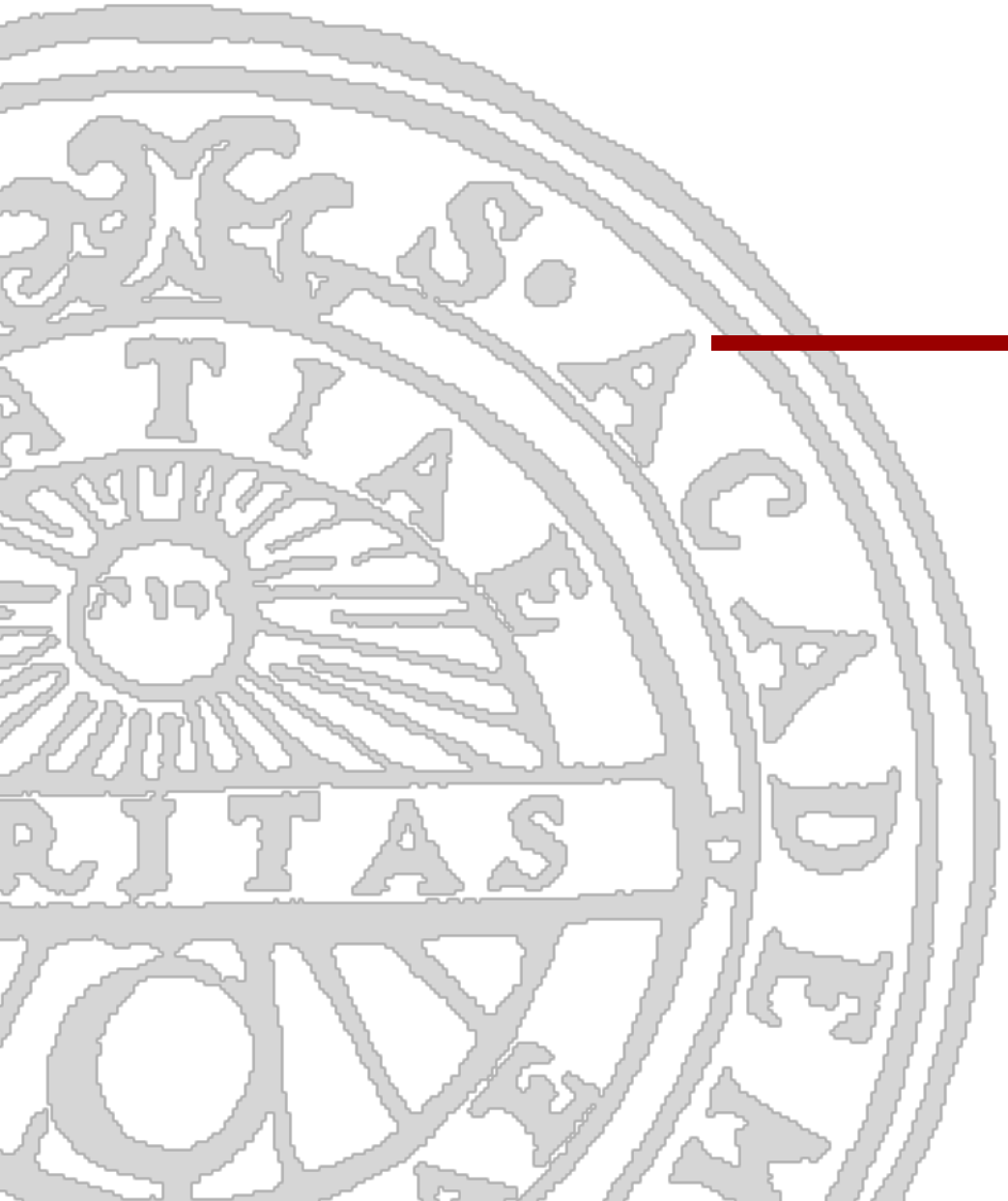
UPPSALA UNIVERSITET

Informationsteknologi

SVMLight

- Training & Testing dataset
  - SVMLight format
- Each features/values pairs
  - separated by a space character
  - ordered by increasing feature number

```
1  # 0 1:EventId 2:DER_mass_MMC 3:DER_mass_transverse_met_lep 4:DER_mass_vis 5:DER_pt_h
   6:DER_deltaeta_jet_jet 7:DER_mass_jet_jet 8:DER_prodeta_jet_jet 9:DER_deltar_tau_lep
   10:DER_pt_tot 11:DER_sum_pt 12:DER_pt_ratio_lep_tau 13:DER_met_phi_centrality
   14:DER_lep_eta_centrality 15:PRI_tau_pt 16:PRI_tau_eta 17:PRI_tau_phi 18:PRI_lep_pt
   19:PRI_lep_eta 20:PRI_lep_phi 21:PRI_met 22:PRI_met_phi 23:PRI_met_sumet 24:PRI_jet_num
   25:PRI_jet_leading_pt 26:PRI_jet_leading_eta 27:PRI_jet_leading_phi
   28:PRI_jet_subleading_pt 29:PRI_jet_subleading_eta 30:PRI_jet_subleading_phi
   31:PRI_jet_all_pt
2  0 1:350000 2:-999.0 3:79.589 4:23.916 5:3.036 6:-999.0 7:-999.0 8:-999.0 9:0.903 10:3.036
   11:56.018 12:1.536 13:-1.404 14:-999.0 15:22.088 16:-0.54 17:-0.609 18:33.93 19:-0.504 20:
   -1.511 21:48.509 22:2.022 23:98.556 24:0 25:-999.0 26:-999.0 27:-999.0 28:-999.0 29:-999.
   0 30:-999.0 31:-0.0
```



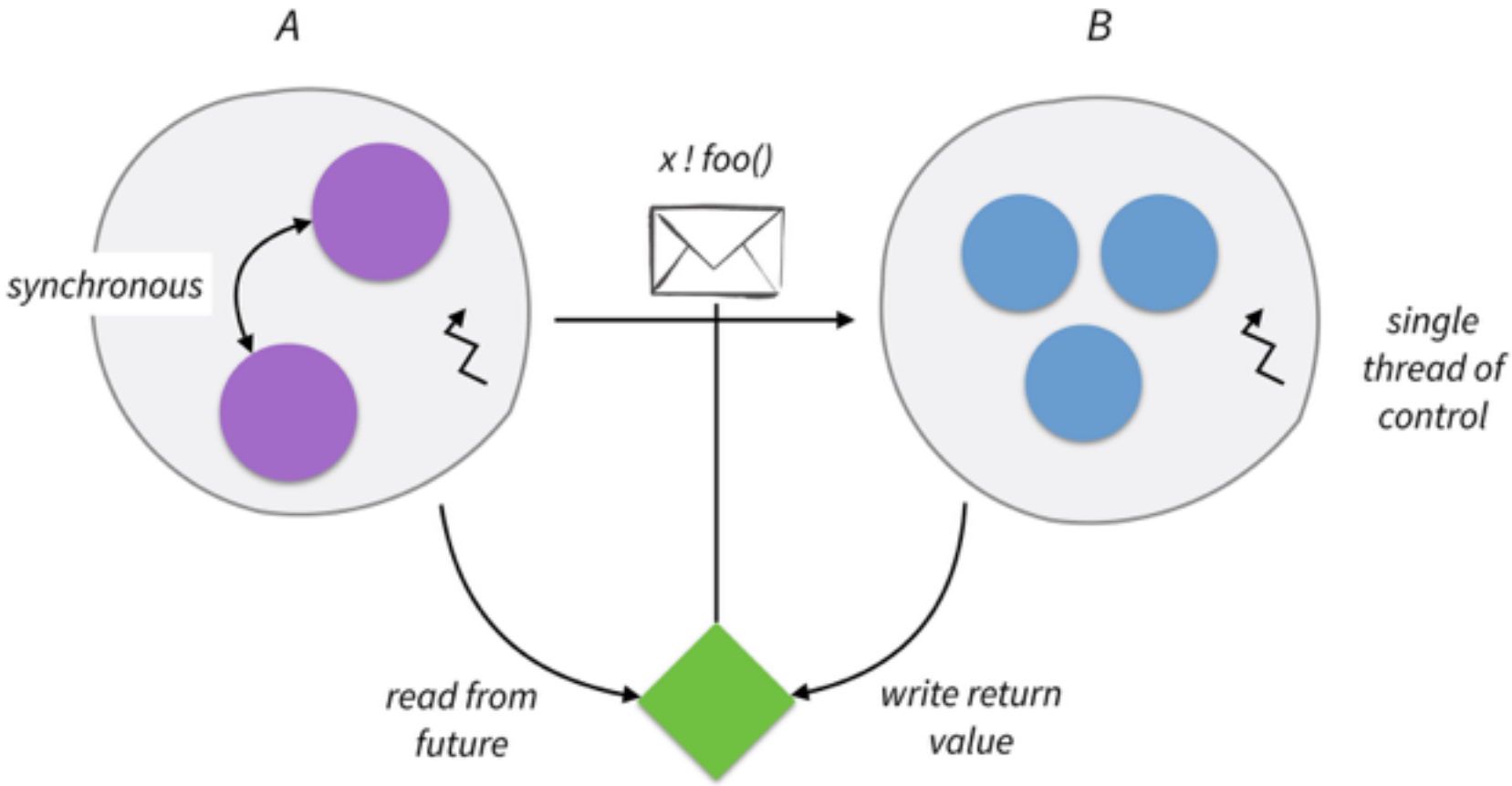
Approach

ENCORE

- Developed by
  - Programming Language Design Lab
- Object-oriented parallel programming language
  - Parallelism-by-default implicitly
    - Active objects
    - Passive objects
- Passive object
  - Does not have thread of control
  - Sync method calls

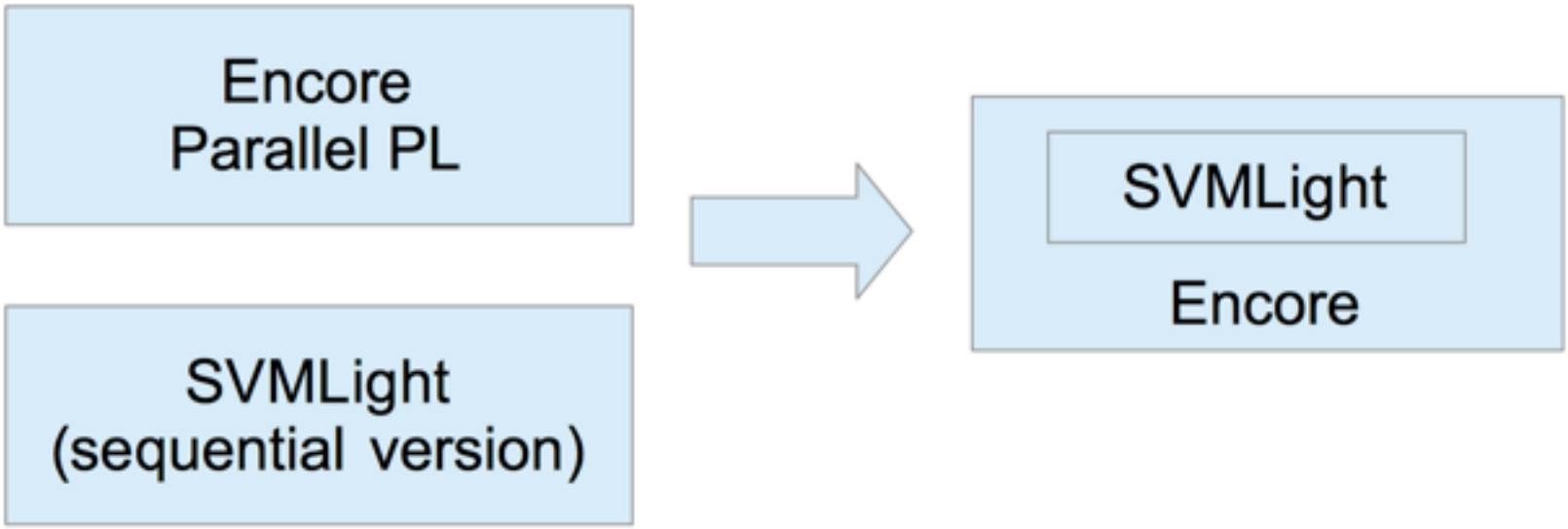
ENCORE (cont)

- Active-object based parallelism



Approach

- Using Encore to transform an SVM library
  - from sequential version
  - to parallel version
- **SVMLight**
  - among top 3 popular SVM libraries
  - LIBSVM, SVMLight, SVMTorch

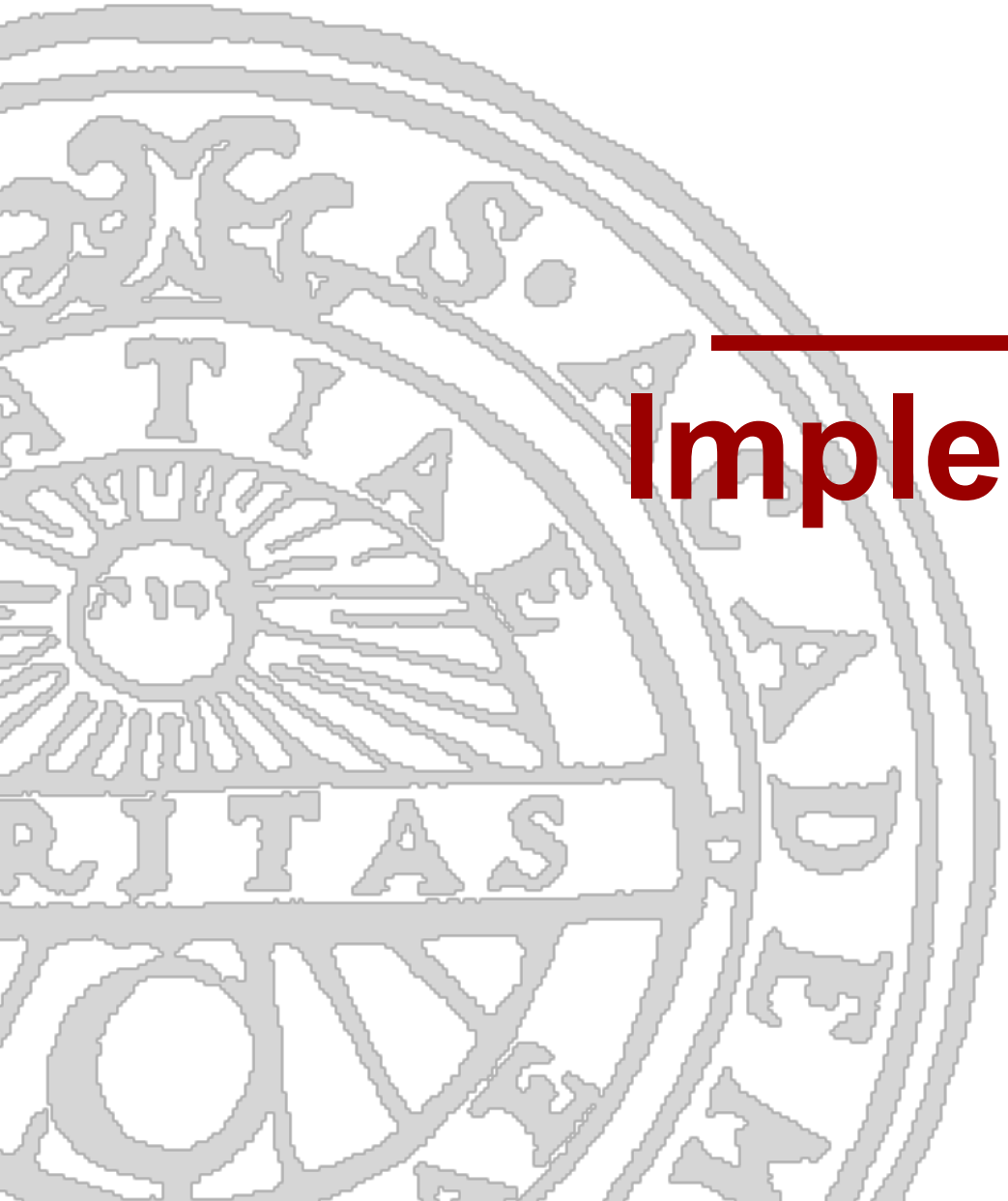


UPPSALA UNIVERSITET

Informationsteknologi

Approach (cont.)

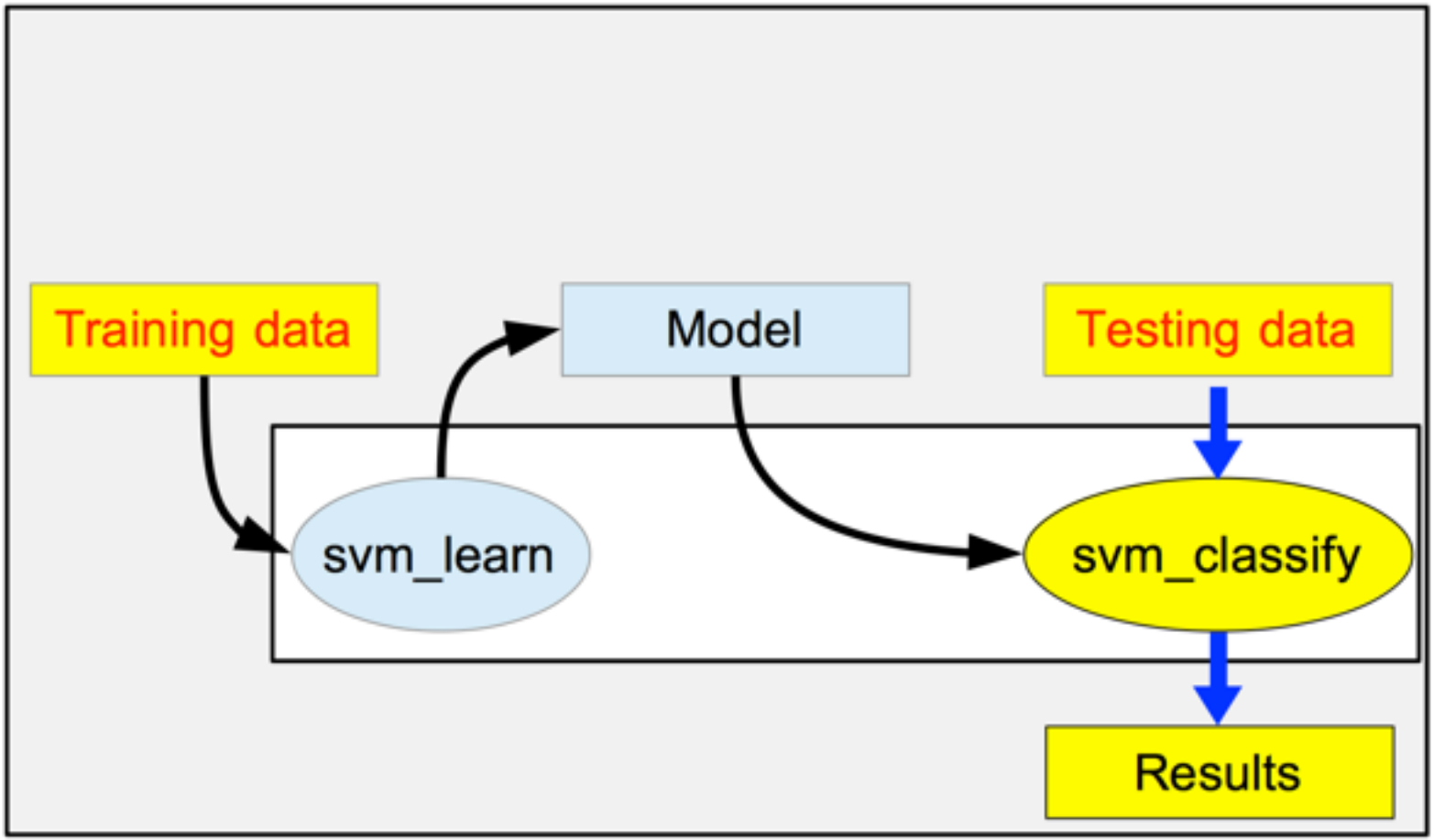
- Transform sequential SVMLight
  - sequential methods transformed to parallel methods
    - using active objects
  - num. of transformations depends on
    - dependencies between sequential methods
- Evaluate by
  - different Higgs Boson datasets
- Conclude with
  - correctness
  - data scalability
  - performance



Implementation

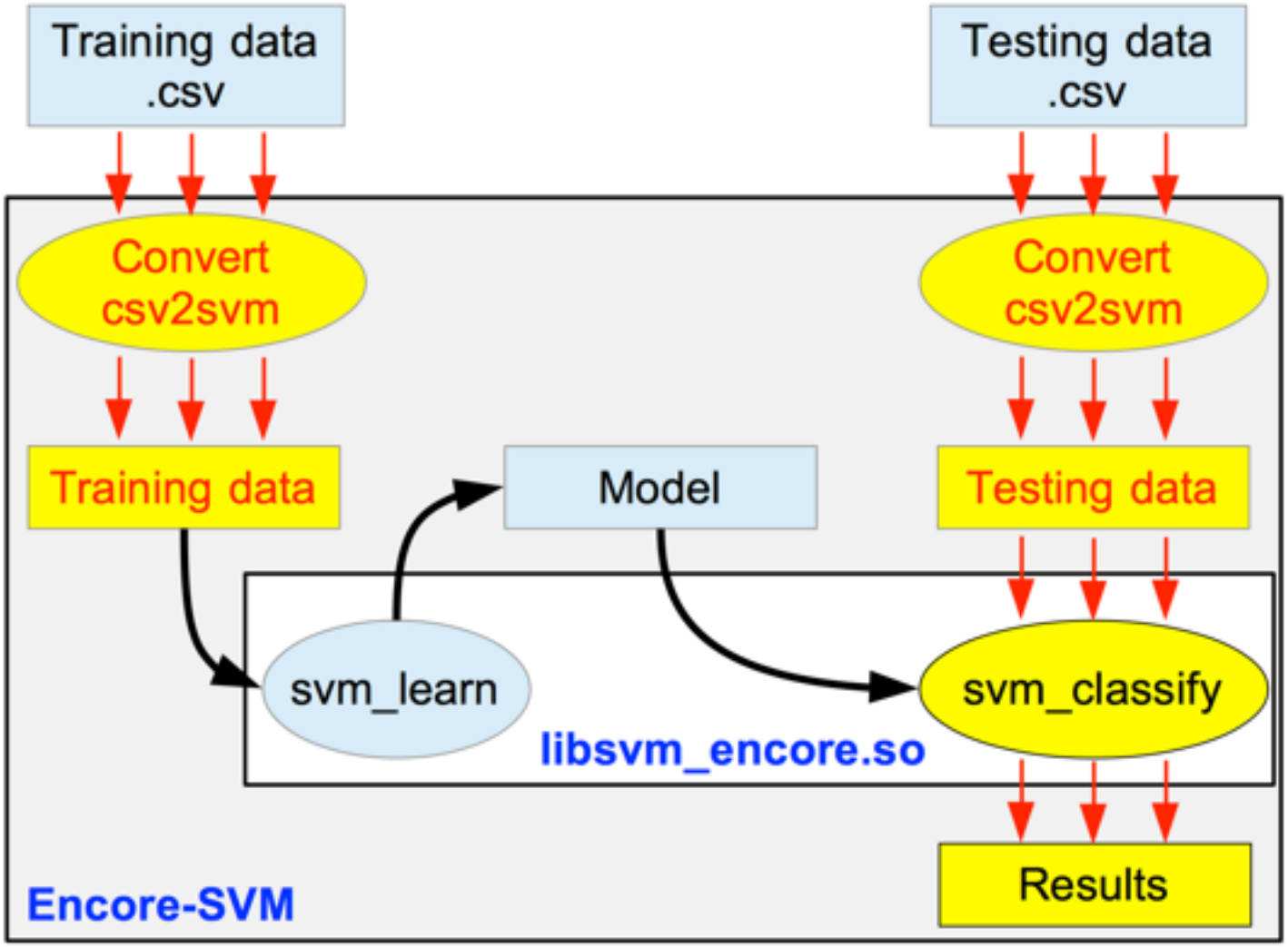
Implementation

- Current model



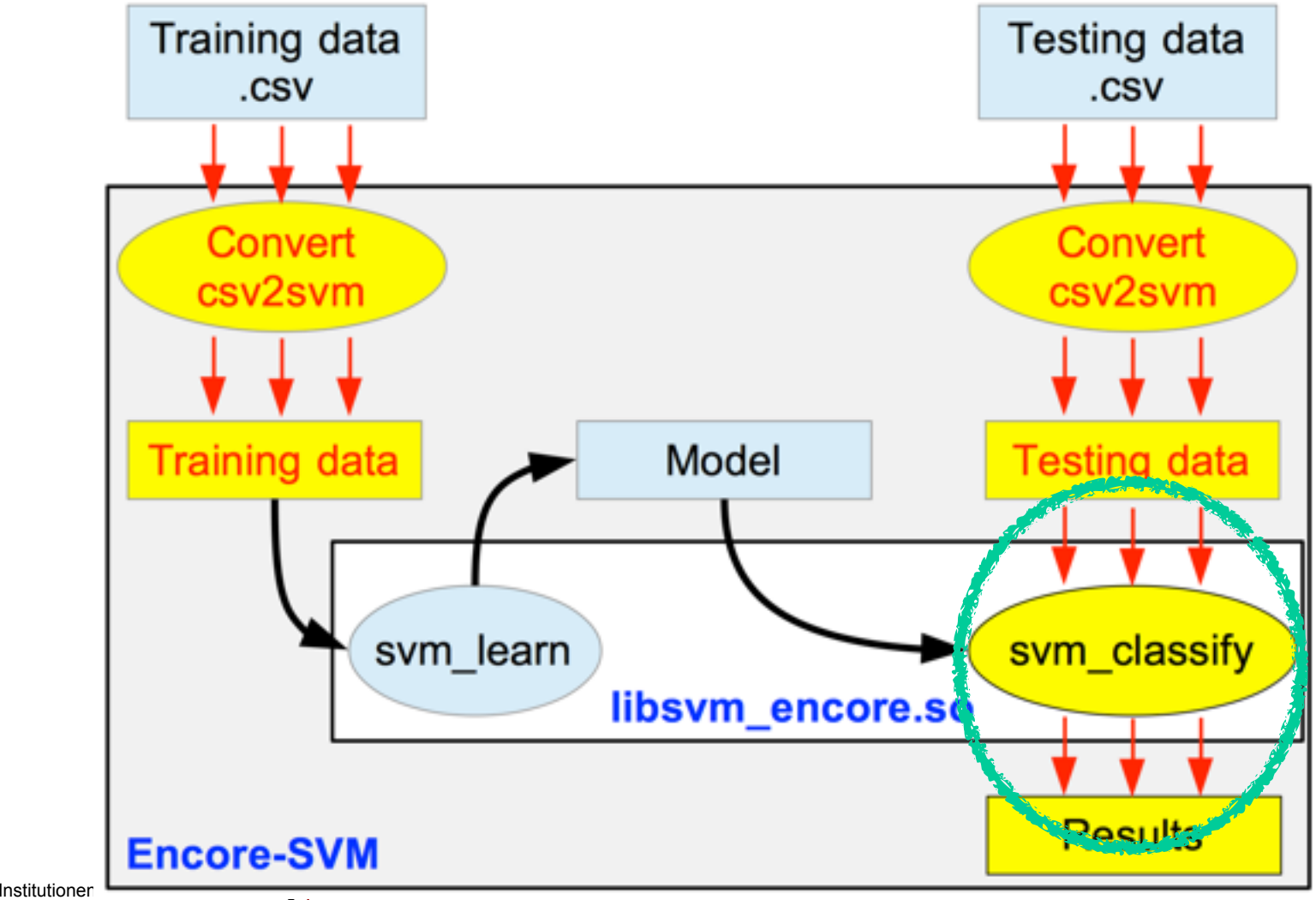
Implementation

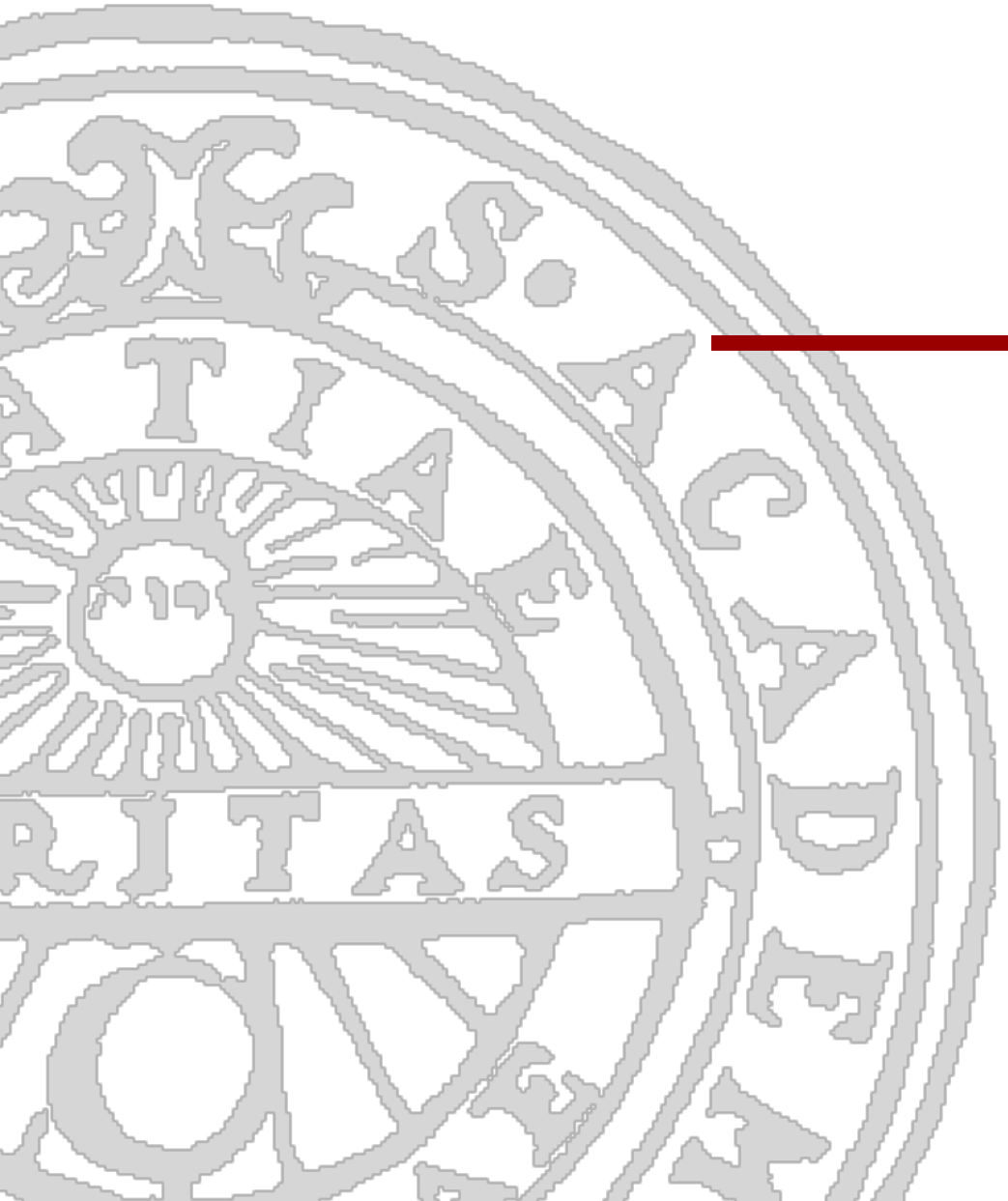
- https://github.com/PhucVH888/uu-projects/tree/master/SVM_Encore



Implementation

- Can be manually parallelize





Validation

Check correctness - # 1

- Learn which Reuters articles are about "corporate acquisitions”
  - Number of features: 9947
  - Training dataset:
    - 1000 positive events
    - 1000 negative events
  - Testing dataset
    - 600 test examples

```
1 6:0.01984032535866 71 15:0.0339873732306071 29:0.0360280968798065
31:0.0378103484117687 41:0.0456787263779904 63:0.021442413608662 74:0.
0813238108919922 75:0.0201048944012214 81:0.0603996615380116 142:0.
0102897706466067 172:0.0777948548082322 174:0.072717200608936 179:0.
```

Check correctness - #1

- Encore-SVM

```
phuc:SVM_Encore vo$ encorec -c svm_encore.enc; ./svm_encore svm_learn -ei data/ex-svms/
example1/train.dat -eo data/ex-svms/example1/model -vv
Importing module Converter from ./Converter.enc
Importing module Core from ./Core.enc
Importing module Params from ./Params.enc
Importing module String from /Users/vo/Dropbox/code/encore/bundles/standard/String.enc
=== Initializing ===
=== Parsing ===
=== Selected mode : svm_learn ===
=== Loading library ===
=== Setting up arguments ===
=== Training ===
       fin  = data/ex-svms/example1/train.dat
       fout = data/ex-svms/example1/model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..700..800..900..1000..1100.
.1200..1300..1400..1500..1600..1700..1800..1900..2000..OK. (2000 examples read)
Setting default regularization parameter C=1.0000
Optimizing.....................................................................
.....................................................................
.....................................................................
.....................................................................
.....................................................d
one. (425 iterations)
Optimization finished (5 misclassified, maxdiff=0.00085).
Runtime in cpu-seconds: 0.17
Number of SV: 878 (including 117 at upper bound)
L1 loss: loss=35.67674
Norm of weight vector: |w|=19.55576
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=383.42791
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=5.85% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>95.40% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>93.07% (rho=1.00,depth=0)
Number of kernel evaluations: 45954
Writing model file...done
=== Done ===
```

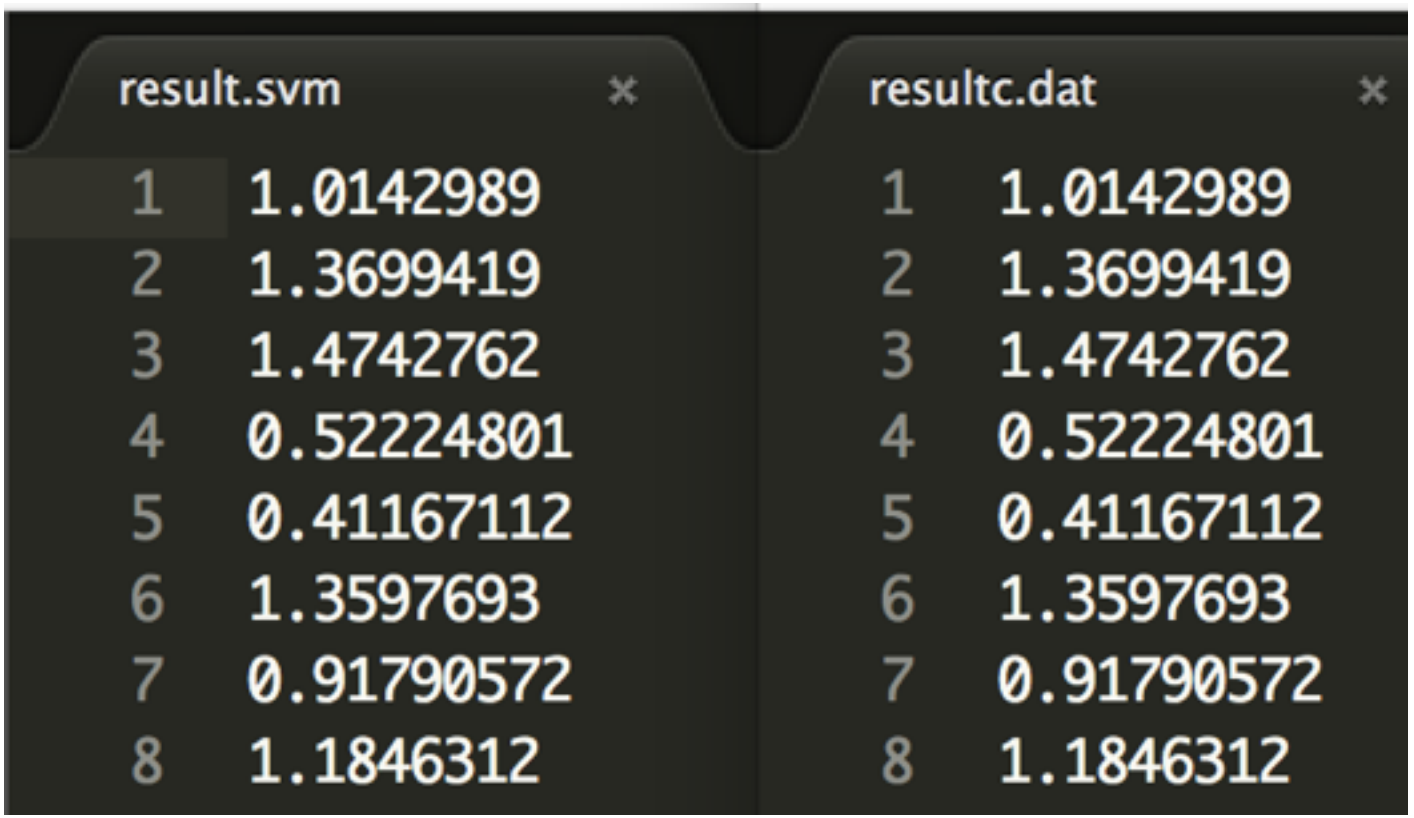
SVMLight

```
phuc:SVM_Encore vo$ ./svm_learn data/ex-svms/example1/train.dat data/ex-
svms/example1/model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..700..800..9
00..1000..1100..1200..1300..1400..1500..1600..1700..1800..1900..2000..OK
. (2000 examples read)
Setting default regularization parameter C=1.0000
Optimizing.....................................................................
.....................................................................
.....................................................................
.....................................................................
.....................................................................
.....................................................................
..done. (425 iterations)
Optimization finished (5 misclassified, maxdiff=0.00085).
Runtime in cpu-seconds: 0.08
Number of SV: 878 (including 117 at upper bound)
L1 loss: loss=35.67674
Norm of weight vector: |w|=19.55576
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=383.42791
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=5.85% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>95.40% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>93.07% (rho=1.00,depth=0)
Number of kernel evaluations: 45954
Writing model file...done
```

- More demos @ my github [https://goo.gl/cxsKfk]

Check correctness - #1

- Results
  - Encore-SVM
  - SVMLight



result.svm

1	1.0142989
2	1.3699419
3	1.4742762
4	0.52224801
5	0.41167112
6	1.3597693
7	0.91790572
8	1.1846312

resultc.dat

1	1.0142989
2	1.3699419
3	1.4742762
4	0.52224801
5	0.41167112
6	1.3597693
7	0.91790572
8	1.1846312

Higgs Boson Data

- Training set
  - 250.000 events, 30 features
  - 1 weight, 1 label, 1 ID
    - Training time: > 12 hours
- Testing set
  - 550.000 events
  - 30 features, 1 ID
    - Testing time: a few seconds

Higgs Boson Data

- Classifying time: Encore-SVM

```
phuc:SVM_Encore vo$ dist/./svm_encore svm_classify -ei data/large/test.svm -im data/large/model.dat -eo data/large/resu
lt-encore-18-apr.dat -vv
=== Initializing ===
=== Parsing ===
=== Selected mode : svm_classify ===
=== Loading library ===
=== Setting up arguments ===
=== Classifying ===
        fin1  = data/large/test.svm
        fin2  = data/large/model.dat
        fout = data/large/result-encore-18-apr.dat
Reading model...OK. (171402 support vectors read)
Classifying test examples..100..200..300..400..500..600..700..800..900..1000..1100..1200..1300..1400..1500..1600..1700.
.1800..1900..2000..2100..2200..2300..2400..2500..2600..2700..2800..2900..3000..3100..3200..3300..3400..3500..3600..3700
..3800..3900..4000..4100..4200..4300..4400..4500..4600..4700..4800..4900..5000..5100..5200..5300..5400..5500..5600..570
0..5800..5900..6000..6100..6200..6300..6400..6500..6600..6700..6800..6900..7000..7100..7200..7300..7400..7500..7600..77
00..7800..7900..8000..8100..8200..8300..8400..8500..8600..8700..8800..8900..9000..9100..9200..9300..9400..9500..9600..9
00..543200..543300..543400..543500..543600..543700..543800..543900..544000..544100..544200..544300..544400..544500..544
600..544700..544800..544900..545000..545100..545200..545300..545400..545500..545600..545700..545800..545900..546000..54
6100..546200..546300..546400..546500..546600..546700..546800..546900..547000..547100..547200..547300..547400..547500..5
47600..547700..547800..547900..548000..548100..548200..548300..548400..548500..548600..548700..548800..548900..549000..
549100..549200..549300..549400..549500..549600..549700..549800..549900..550000..done
Runtime (without IO) in cpu-seconds: 0.17
=== Done ===
```

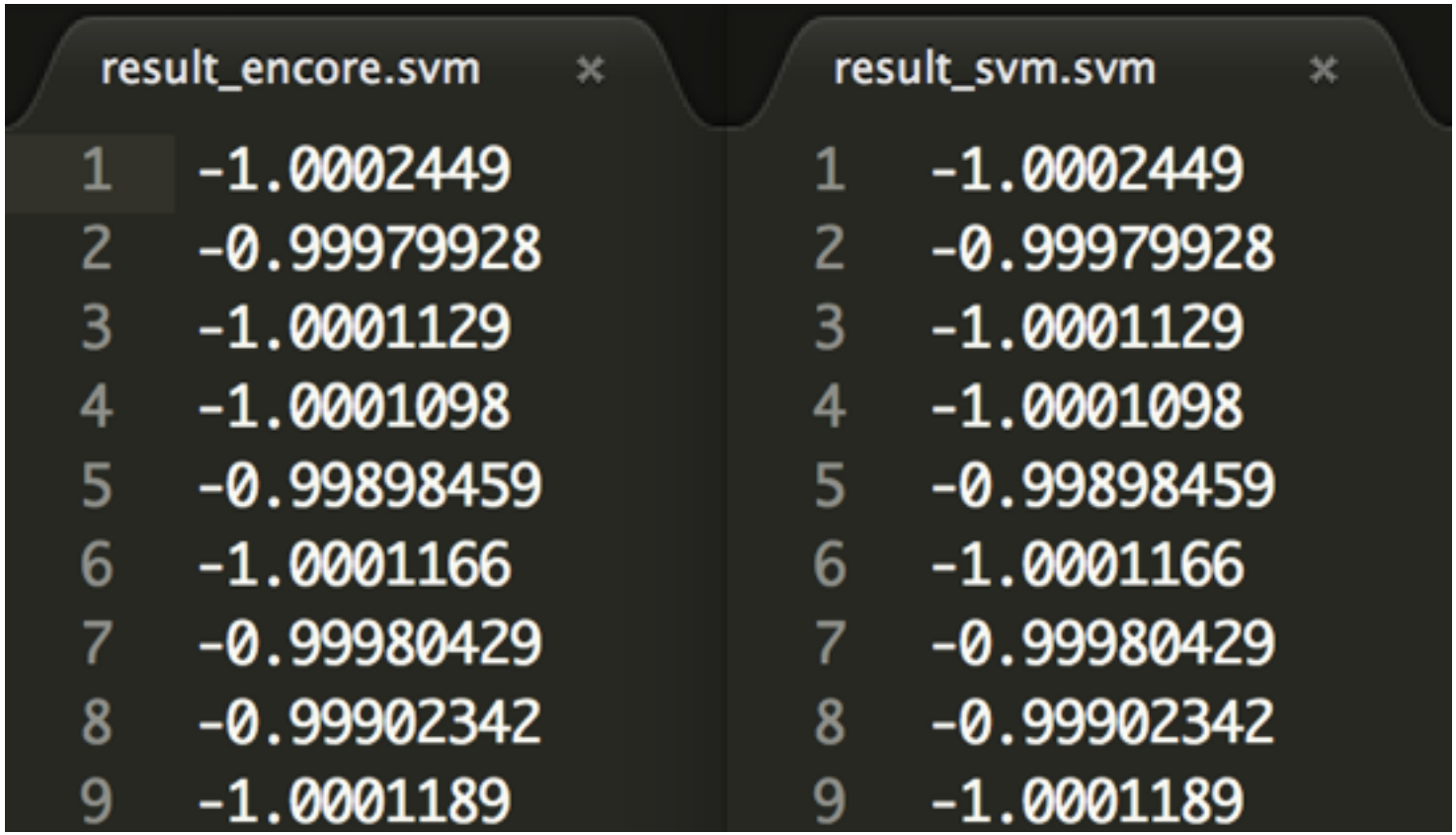
Higgs Boson Data

- Classifying time: SVMLight

```
phuc:SVM_Encore vo$ dist/./svm_classify data/large/test.svm data/large/model.dat data/large/result-svml-18-apr.dat
Reading model...OK. (171402 support vectors read)
Classifying test examples..100..200..300..400..500..600..700..800..900..1000..1100..1200..1300..1400..1500..1600..1700..
.1800..1900..2000..2100..2200..2300..2400..2500..2600..2700..2800..2900..3000..3100..3200..3300..3400..3500..3600..3700
..3800..3900..4000..4100..4200..4300..4400..4500..4600..4700..4800..4900..5000..5100..5200..5300..5400..5500..5600..570
0..5800..5900..6000..6100..6200..6300..6400..6500..6600..6700..6800..6900..7000..7100..7200..7300..7400..7500..7600..77
00..7800..7900..8000..8100..8200..8300..8400..8500..8600..8700..8800..8900..9000..9100..9200..9300..9400..9500..9600..9
0..541700..541800..541900..542000..542100..542200..542300..542400..542500..542600..542700..542800..542900..543000..5431
00..543200..543300..543400..543500..543600..543700..543800..543900..544000..544100..544200..544300..544400..544500..544
600..544700..544800..544900..545000..545100..545200..545300..545400..545500..545600..545700..545800..545900..546000..54
6100..546200..546300..546400..546500..546600..546700..546800..546900..547000..547100..547200..547300..547400..547500..5
47600..547700..547800..547900..548000..548100..548200..548300..548400..548500..548600..548700..548800..548900..549000..
549100..549200..549300..549400..549500..549600..549700..549800..549900..550000..done
Runtime (without IO) in cpu-seconds: 0.22
```

Higgs Boson Data

- Encore-SVM SVMLight



result_encore.svm

1	-1.0002449
2	-0.99979928
3	-1.0001129
4	-1.0001098
5	-0.99898459
6	-1.0001166
7	-0.99980429
8	-0.99902342
9	-1.0001189

result_svm.svm

1	-1.0002449
2	-0.99979928
3	-1.0001129
4	-1.0001098
5	-0.99898459
6	-1.0001166
7	-0.99980429
8	-0.99902342
9	-1.0001189

Conclusion

- Expected Outcomes
  - Better performance than sequential SVMLight version
    - on varied scalable datasets
  - Classifications of
    - background processes
    - signal processes
- Validation
  - Performance
    - execution time of original and Encore versions
  - Accuracy
    - check the results of both implementation

UPPSALA UNIVERSITET

Informationsteknologi

References

. [1] Atlas Higgs Challenge 2014. http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014.

. [2] Stephan Brandauer et al. “Formal Methods for Multicore Programming: 15th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2015, Bertinoro, Italy, June 15-19, 2015, Advanced Lectures”. In: ed. by Marco Bernardo and Broch Einar Johnsen. Cham: Springer International Publishing, 2015. Chap. Parallel Objects for Multicores: A Glimpse at the Parallel Language Encore, pp. 1–56.

. [3] CERN. https://root.cern.ch.

. [4] Encore-SVM. https://github.com/PhucVH888/uu-projects/tree/master/SVM_Encore.

. [5] Higgs Datasets 2014. https://archive.ics.uci.edu/ml/datasets.

. [6] Higgs Training database. https://archive.ics.uci.edu/ml/machine-learning-databases/ 00280/.

. [7] Statistical packages. https://en.wikipedia.org/wiki/List_of_statistical_packages.

. [8] SVMLight. http://svmlight.joachims.org.

. [9] SVMs tools. http://www.svms.org/software.html.

Informationsteknologi

Thank you!

Check correctness - # 2

- Learn which Reuters articles are about "corporate acquisitions”
  - Number of features: 9947
  - Training dataset:
    - 5 positive events
    - 5 negative events
  - Testing dataset
    - 600 test examples

Check correctness - # 2

- Encore-SVM

```
phuc:SVM_Encore vo$ encorec -c svm_encore.enc; ./svm_encore svm_learn -ei data/ex-svms/example2/train_tra
-eo data/ex-svms/example2/model -vv
Importing module Converter from ./Converter.enc
Importing module Core from ./Core.enc
Importing module Params from ./Params.enc
Importing module String from /Users/vo/Dropbox/code/encore/bundles/standard/String.enc
=== Initializing ===
=== Parsing ===
=== Selected mode : svm_learn ===
=== Loading library ===
=== Setting up arguments ===
=== Training ===
       fin  = data/ex-svms/example2/train_transduction.dat
       fout = data/ex-svms/example2/model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..OK. (610 examples read)
Setting default regularization parameter C=1.0066

Deactivating Shrinking due to an incompatibility with the transductive
learner in the current version.

Optimizing.done
Classifying unlabeled data as 300 POS / 300 NEG.
Retraining.....................................................................
..............................................................................
.....................................done
Increasing influence of unlabeled examples to 25.251168% ..........................done
Retraining.............................................done
Increasing influence of unlabeled examples to 37.876752% ..........................done
Retraining.....................................................done
Increasing influence of unlabeled examples to 56.815129% ............done
Retraining.....................................................done
Increasing influence of unlabeled examples to 85.222693% ......done
Retraining.....................................................done
Increasing influence of unlabeled examples to 100.000000% .....done
Retraining.....................................................................done
Writing prediction file...done
Number of switches: 60
done. (4374 iterations)
Optimization finished (2 misclassified, maxdiff=0.00086).
Runtime in cpu-seconds: 1.83
Number of SV: 369 (including 26 at upper bound)
L1 loss: loss=6.88538
Norm of weight vector: |w|=12.71364
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=162.63666
xacrit>=1: labeledpos=0.00000 labeledneg=0.00000 default=50.00000
xacrit>=1: unlabelpos=1.00000 unlabelneg=3.33333
xacrit>=1: labeled=0.00000 unlabled=4.33333 all=4.26230
xacritsum: labeled=39.66289 unlabled=28.58790 all=28.76946
r_delta_sq=1.00000 xisum=6.92909 asum=168.56461
Number of kernel evaluations: 244477
Writing model file...done
=== Done ===
```

SVMLight

```
phuc:SVM_Encore vo$  ./svm_learn data/ex-svms/example2/train_t
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..O
Setting default regularization parameter C=1.0066

Deactivating Shrinking due to an incompatibility with the tran
learner in the current version.

Optimizing.done
Classifying unlabeled data as 300 POS / 300 NEG.
Retraining.....................................................
..............................................................
.....................................done
Increasing influence of unlabeled examples to 0.001500% .....
319 positive -> Switching labels of 1 POS / 1 NEG unlabeled examples..done
Retraining.....done
Increasing influence of unlabeled examples to 25.251168% ...............
Retraining.........................................done
Increasing influence of unlabeled examples to 37.876752% ...............
Retraining.....................................................done
Increasing influence of unlabeled examples to 56.815129% ............done
Retraining.....................................................done
Increasing influence of unlabeled examples to 85.222693% ......done
Retraining.....................................................done
Increasing influence of unlabeled examples to 100.000000% .....done
Retraining.....................................................................
Writing prediction file...done
Number of switches: 60
done. (4374 iterations)
Optimization finished (2 misclassified, maxdiff=0.00086).
Runtime in cpu-seconds: 1.52
Number of SV: 369 (including 26 at upper bound)
L1 loss: loss=6.88538
Norm of weight vector: |w|=12.71364
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=162.63666
xacrit>=1: labeledpos=0.00000 labeledneg=0.00000 default=50.00000
xacrit>=1: unlabelpos=1.00000 unlabelneg=3.33333
xacrit>=1: labeled=0.00000 unlabled=4.33333 all=4.26230
xacritsum: labeled=39.66289 unlabled=28.58790 all=28.76946
r_delta_sq=1.00000 xisum=6.92909 asum=168.56461
Number of kernel evaluations: 244477
Writing model file...done
```

UPPSALA UNIVERSITET

Informationsteknologi

Demo

- Correctness
- Performance