DCAFPilot overview, analytics R&D

Valentin Kuznetsov, Cornell University

May 18th, CMS R&D meeting

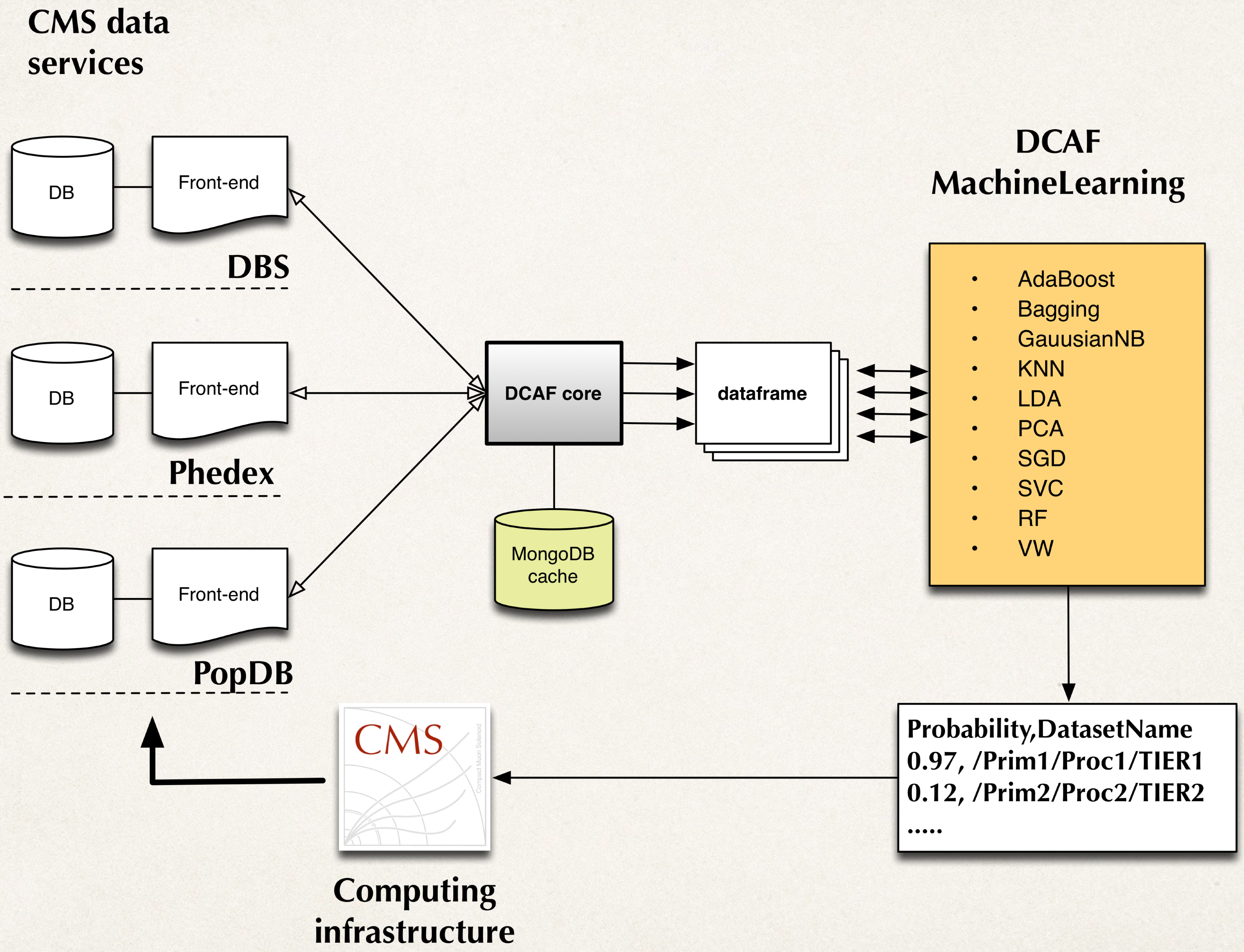
- Review of DCAFPilot internal
- Collected data
- Model selection
- Automation
- Challenges
- Plans

Problem statement

- We would like to predict dataset popularity once datasets created
- This information can be used by DDM to seed initial replicas, once historical information will be acquired replicas can be adjusted
- Investigate tools/techniques to perform these studies and evaluate necessary metrics
    - define popularity metrics
    - identify algorithm(s) to perform dataset popularity
    - setup machinery to perform this action on regular intervals

DCAFPilot internals

- DCAFPilot package has set of tools to collect data from CMS data-services, process/transform them, run ML algorithms, basic web service.
- Code it written in python, depends on numpy, scipy, pandas, MongoDB, Vowpal Wabbit, xgboost
  - R language can be used for data exploration and ML training
  - https://github.com/dmwm/DMWMAnalytics/tree/master/Popularity/DCAFPilot



CMS data services

DBS

Phedex

PopDB

DCAF MachineLearning

- AdaBoost
- Bagging
- GauusianNB
- KNN
- LDA
- PCA
- SGD
- SVC
- RF
- VW

```
Probability,DatasetName
0.97, /Prim1/Proc1/TIER1
0.12, /Prim2/Proc2/TIER2
.....
```

Computing infrastructure

DCAFPilot toolset

- Data collection: *dataframe, cron4dataframe.sh, popular_datasets*

- Data transformation: *merge_csv, transform_csv, csv2libsvm, csv2vm, vw_pred2csv, slice_data*

- Data modeling: *model, cron4models.sh, cron4verify.sh, run_models, check_prediction, verify_prediction*

Data collection

- Query CMS services: 800K queries, 200K datasets, 900 releases, 500 site entries and 5k user DNs from SiteDB (no-clean up)
  - Use DBS, Phedex, SiteDB, PopDB, Dashboard data-services
- Anonymized and factorization of all data, e.g. use DBS ORACLE ids, hashes
- 2013/2014/2015 data are available, one file per calendar week ***/afs/cern.ch/user/v/valya/workspace/analytics/data***
  - each file consists ~1K of popular dataset + 10K supplement datasets from DBS
  - one year of meta-data translates into 78x600000 data-frame
  - initially data were collected manually, new data are generated via cronjob

Modeling

- DCAFPilot provides tools to train ML models based on scikit-learn python ML library
  - 8 regressors and 18 classifiers are available
  - *model --learner=RandomForestClassifier --idcol=id --target=target —train-file=train_clf.csv.gz --scaler=StandardScaler --newdata=valid_clf.csv.gz --predict=pred.txt*
- External ML libraries are easy to use, e.g. used VW (Yahoo) and xgboost libraries (I learnt about them from kaggle competitions)
- *run_models* script runs 5 models: RandomForest, SGDClassifier, LinearSVC, VW, xgboost
- It is possible to run models on regular basis via cronjob but w/o fine-tuning

Automation

- Reached the point to achieve full automation in data collection and ML training
- DCAFPilot.spec is part of cmsdist, RPMs are available in my private repository and deployed to VM
- Code is deployed to VM (bc-stack.cern.ch) and 3 cronjobs are set-up on weekly basis:
    - cron4dataframes.sh collects new data for past calendar week
    - cron4models.sh runs 5 ML algorithms on fresh merged dataset
    - cron4verify.sh verify results from previous week wrt popDB data
- Pull data/predictions from VM to lxplus via acrontab into */afs/cern.ch/user/v/valya/workspace/analytics* area

Automation results train on 2014, predict 2015, naccess>10

	RF	SGD	LinearSVC	VW	xgboost
accuracy	0.97	0.96	0.97	0.92	0.97
precision	0.82	0.81	0.81	0.83	0.86
recall	0.95	0.86	0.94	0.46	0.93
F1-score	0.88	0.83	0.87	0.59	0.89

Tiers/classifiers breakdown

```
RandomForest classifier

TIER          TP(%)  TN(%)   FP(%)   FN(%)
---------------------------------------
AOD           37.21  55.77   0.67    6.35
AODSIM        18.13  74.32   0.63    6.92
MINIAOD        3.19  93.62   0.27    2.93
MINIAODSIM    38.11  45.00   0.29   16.60
USER           8.97  84.82   1.01    5.20

TOTAL          12.6  80.4    0.9     6.1
```

```
SDGClassifier

TIER          TP(%)  TN(%)   FP(%)   FN(%)
---------------------------------------
AOD           35.34  56.22   0.22    8.22
AODSIM        16.96  74.73   0.22    8.09
MINIAOD        3.46  93.62   0.27    2.66
MINIAODSIM    38.76  45.15   0.14   15.95
USER           8.77  85.46   0.37    5.41

TOTAL         12.3   81.0    0.3     6.4
```

```
LinearSVC classifier

TIER          TP(%)  TN(%)   FP(%)   FN(%)
---------------------------------------
AOD           36.50  56.22   0.22    7.06
AODSIM        18.15  74.70   0.25    6.90
MINIAOD        3.46  93.35   0.53    2.66
MINIAODSIM    41.24  45.14   0.15   13.47
USER           9.39  85.37   0.45    4.78

TOTAL         13.1   80.9    0.4     5.6
```

```
Xgboost classifier

TIER          TP(%)  TN(%)   FP(%)   FN(%)
---------------------------------------
AOD           36.76  56.37   0.07    6.80
AODSIM        18.39  74.90   0.05    6.65
MINIAOD        3.46  93.88   0.00    2.66
MINIAODSIM    36.70  45.27   0.02   18.01
USER           9.77  85.78   0.05    4.40

TOTAL         13.2   81.2    0.0     5.5
```

RandomForest classifier

```
TIER                             TP(%)  TN(%)  FP(%)  FN(%)
-----------------------------------------------------------
ALCAPROMPT                        0.00 100.00   0.00   0.00
ALCARECO                          0.21  99.32   0.38   0.09
AOD                              37.32  56.00   0.45   6.24
AODSIM                           18.45  74.64   0.31   6.59
DIGI-RECO                         0.00 100.00   0.00   0.00
DQM                               0.00  99.72   0.28   0.00
DQMIO                             0.00  99.77   0.23   0.00
DQMROOT                           0.00 100.00   0.00   0.00
FEVT                             24.29  74.29   0.00   1.43
FEVTHLTALL                        0.00 100.00   0.00   0.00
GEN                               0.14  99.40   0.23   0.24
GEN-RAW                           0.57  98.94   0.49   0.00
GEN-RAWDEBUG                      0.00 100.00   0.00   0.00
GEN-SIM                           1.87  96.75   0.40   0.98
GEN-SIM-DIGI                      0.00 100.00   0.00   0.00
GEN-SIM-DIGI-RAW                  4.68  94.50   0.19   0.63
GEN-SIM-DIGI-RAW-HLTDEBUG         1.36  97.28   0.23   1.13
GEN-SIM-DIGI-RAW-HLTDEBUG-RECO    0.00 100.00   0.00   0.00
GEN-SIM-DIGI-RECO                 0.56  98.31   0.56   0.56
GEN-SIM-RAW                      31.15  64.19   0.31   4.35
GEN-SIM-RAW-HLT                   0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG              0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG-RECO         0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG-RECODEBUG    0.00 100.00   0.00   0.00
GEN-SIM-RAW-RECO                 16.90  83.10   0.00   0.00
GEN-SIM-RAWDEBUG                  0.00 100.00   0.00   0.00
GEN-SIM-RECO                      8.90  87.51   0.36   3.23
GEN-SIM-RECODEBUG                 8.39  87.27   0.00   4.35
LHE                               0.00 100.00   0.00   0.00
MINIAOD                           3.46  93.62   0.27   2.66
MINIAODSIM                       37.76  45.23   0.06  16.95
PREMIX-RAW                        0.00 100.00   0.00   0.00
PREMIXRAW                         0.00 100.00   0.00   0.00
RAW                               5.08  91.94   0.45   2.54
RAW-HLT                           0.00 100.00   0.00   0.00
RAW-RECO                          3.39  94.92   0.28   1.41
RAWRECOSIMHLT                     0.00 100.00   0.00   0.00
RECO                              8.13  90.08   0.33   1.45
RECODEBUG                         0.00 100.00   0.00   0.00
USER                              9.51  85.44   0.39   4.66

ALL tiers                         9.49  86.35   0.34   3.81
```

SGD classifier

```
TIER                             TP(%)  TN(%)  FP(%)  FN(%)
-----------------------------------------------------------
ALCAPROMPT                        0.00 100.00   0.00   0.00
ALCARECO                          0.20  99.61   0.09   0.11
AOD                              34.03  56.41   0.04   9.53
AODSIM                           15.00  74.89   0.06  10.04
DIGI-RECO                         0.00 100.00   0.00   0.00
DQM                               0.00  99.97   0.03   0.00
DQMIO                             0.00  99.91   0.09   0.00
DQMROOT                           0.00 100.00   0.00   0.00
FEVT                             21.43  74.29   0.00   4.29
FEVTHLTALL                        0.00 100.00   0.00   0.00
GEN                               0.06  99.59   0.03   0.32
GEN-RAW                           0.49  99.35   0.08   0.08
GEN-RAWDEBUG                      0.00 100.00   0.00   0.00
GEN-SIM                           1.49  97.11   0.04   1.36
GEN-SIM-DIGI                      0.00 100.00   0.00   0.00
GEN-SIM-DIGI-RAW                  4.17  94.69   0.00   1.14
GEN-SIM-DIGI-RAW-HLTDEBUG         1.00  97.44   0.08   1.49
GEN-SIM-DIGI-RAW-HLTDEBUG-RECO    0.00 100.00   0.00   0.00
GEN-SIM-DIGI-RECO                 0.28  98.88   0.00   0.84
GEN-SIM-RAW                      27.65  64.46   0.04   7.85
GEN-SIM-RAW-HLT                   0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG              0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG-RECO         0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG-RECODEBUG    0.00 100.00   0.00   0.00
GEN-SIM-RAW-RECO                 16.90  83.10   0.00   0.00
GEN-SIM-RAWDEBUG                  0.00 100.00   0.00   0.00
GEN-SIM-RECO                      8.15  87.80   0.07   3.98
GEN-SIM-RECODEBUG                 7.76  87.27   0.00   4.97
LHE                               0.00 100.00   0.00   0.00
MINIAOD                           2.66  93.88   0.00   3.46
MINIAODSIM                       26.37  45.27   0.02  28.34
PREMIX-RAW                        0.00 100.00   0.00   0.00
PREMIXRAW                         0.00 100.00   0.00   0.00
RAW                               4.22  92.26   0.12   3.40
RAW-HLT                           0.00 100.00   0.00   0.00
RAW-RECO                          2.63  95.01   0.19   2.16
RAWRECOSIMHLT                     0.00 100.00   0.00   0.00
RECO                              6.12  90.35   0.07   3.47
RECODEBUG                         0.00 100.00   0.00   0.00
USER                              7.91  85.76   0.07   6.27

ALL tiers                         7.82  86.63   0.06   5.48
```

LinearSVC classifier

```
TIER                                TP(%)  TN(%)  FP(%)  FN(%)
-------------------------------------------------------------
ALCAPROMPT                           0.00 100.00   0.00   0.00
ALCARECO                             0.20  99.61   0.09   0.11
AOD                                 34.03  56.41   0.04   9.53
AODSIM                              15.00  74.89   0.06  10.04
DIGI-RECO                            0.00 100.00   0.00   0.00
DQM                                  0.00  99.97   0.03   0.00
DQMIO                                0.00  99.91   0.09   0.00
DQMROOT                              0.00 100.00   0.00   0.00
FEVT                                21.43  74.29   0.00   4.29
FEVTHLTALL                           0.00 100.00   0.00   0.00
GEN                                  0.06  99.59   0.03   0.32
GEN-RAW                              0.49  99.35   0.08   0.08
GEN-RAWDEBUG                         0.00 100.00   0.00   0.00
GEN-SIM                              1.49  97.11   0.04   1.36
GEN-SIM-DIGI                         0.00 100.00   0.00   0.00
GEN-SIM-DIGI-RAW                     4.17  94.69   0.00   1.14
GEN-SIM-DIGI-RAW-HLTDEBUG            1.00  97.44   0.08   1.49
GEN-SIM-DIGI-RAW-HLTDEBUG-RECO       0.00 100.00   0.00   0.00
GEN-SIM-DIGI-RECO                    0.28  98.88   0.00   0.84
GEN-SIM-RAW                         27.65  64.46   0.04   7.85
GEN-SIM-RAW-HLT                      0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG                 0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG-RECO            0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG-RECODEBUG       0.00 100.00   0.00   0.00
GEN-SIM-RAW-RECO                    16.90  83.10   0.00   0.00
GEN-SIM-RAWDEBUG                     0.00 100.00   0.00   0.00
GEN-SIM-RECO                         8.15  87.80   0.07   3.98
GEN-SIM-RECODEBUG                    7.76  87.27   0.00   4.97
LHE                                  0.00 100.00   0.00   0.00
MINIAOD                              2.66  93.88   0.00   3.46
MINIAODSIM                          26.37  45.27   0.02  28.34
PREMIX-RAW                           0.00 100.00   0.00   0.00
PREMIXRAW                            0.00 100.00   0.00   0.00
RAW                                  4.22  92.26   0.12   3.40
RAW-HLT                              0.00 100.00   0.00   0.00
RAW-RECO                             2.63  95.01   0.19   2.16
RAWRECOSIMHLT                        0.00 100.00   0.00   0.00
RECO                                 6.12  90.35   0.07   3.47
RECODEBUG                            0.00 100.00   0.00   0.00
USER                                 7.91  85.76   0.07   6.27

ALL tiers                            7.82  86.63   0.06   5.48
```

xgboost classifier

```
TIER                                TP(%)  TN(%)  FP(%)  FN(%)
-------------------------------------------------------------
ALCAPROMPT                           0.00 100.00   0.00   0.00
ALCARECO                             0.26  99.65   0.05   0.05
AOD                                 36.72  56.37   0.07   6.84
AODSIM                              18.39  74.90   0.05   6.66
DIGI-RECO                            0.00 100.00   0.00   0.00
DQM                                  0.00  99.97   0.03   0.00
DQMIO                                0.00 100.00   0.00   0.00
DQMROOT                              0.00 100.00   0.00   0.00
FEVT                                22.86  74.29   0.00   2.86
FEVTHLTALL                           0.00 100.00   0.00   0.00
GEN                                  0.17  99.61   0.01   0.20
GEN-RAW                              0.57  99.43   0.00   0.00
GEN-RAWDEBUG                         0.00 100.00   0.00   0.00
GEN-SIM                              1.88  97.10   0.05   0.97
GEN-SIM-DIGI                         0.00 100.00   0.00   0.00
GEN-SIM-DIGI-RAW                     4.68  94.69   0.00   0.63
GEN-SIM-DIGI-RAW-HLTDEBUG            1.46  97.41   0.10   1.02
GEN-SIM-DIGI-RAW-HLTDEBUG-RECO       0.00 100.00   0.00   0.00
GEN-SIM-DIGI-RECO                    0.42  98.74   0.14   0.70
GEN-SIM-RAW                         31.28  64.50   0.00   4.22
GEN-SIM-RAW-HLT                      0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG                 0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG-RECO            0.00 100.00   0.00   0.00
GEN-SIM-RAW-HLTDEBUG-RECODEBUG       0.00 100.00   0.00   0.00
GEN-SIM-RAW-RECO                    16.90  83.10   0.00   0.00
GEN-SIM-RAWDEBUG                     0.00 100.00   0.00   0.00
GEN-SIM-RECO                         9.04  87.85   0.02   3.09
GEN-SIM-RECODEBUG                    8.39  87.27   0.00   4.35
LHE                                  0.00 100.00   0.00   0.00
MINIAOD                              2.66  93.88   0.00   3.46
MINIAODSIM                          35.10  45.29   0.00  19.61
PREMIX-RAW                           0.00 100.00   0.00   0.00
PREMIXRAW                            0.00 100.00   0.00   0.00
RAW                                  5.32  92.39   0.00   2.29
RAW-HLT                              0.00 100.00   0.00   0.00
RAW-RECO                             3.39  95.20   0.00   1.41
RAWRECOSIMHLT                        0.00 100.00   0.00   0.00
RECO                                 7.90  90.41   0.00   1.69
RECODEBUG                            0.00 100.00   0.00   0.00
USER                                 9.38  85.78   0.04   4.80

ALL tiers                            9.36  86.66   0.04   3.95
```

Challenges

- The biggest challenge is model selection and proper tuning, e.g. automation procedure only collect data and runs pre-defined algorithms

- Model strategy is not clear, e.g. train for all tiers or subset

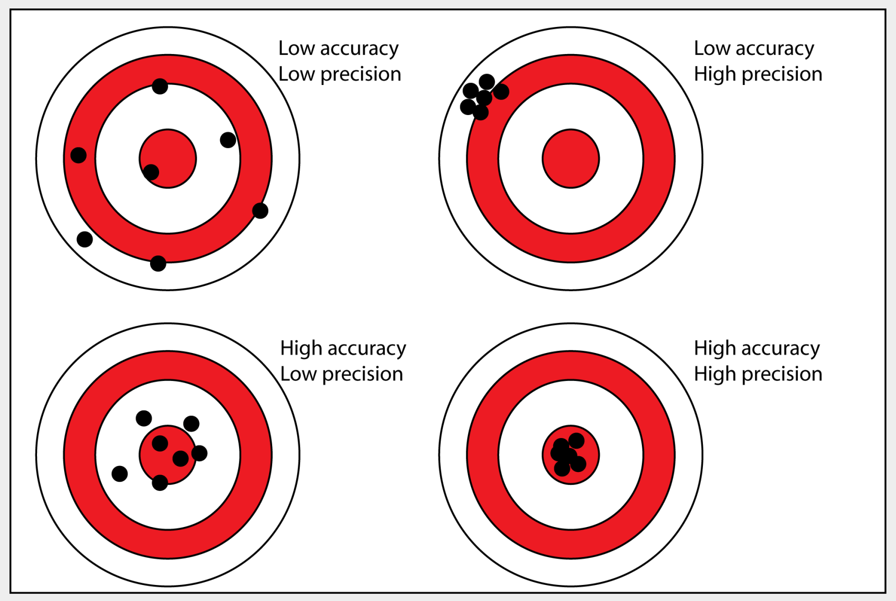
- It is unclear if collected meta-data is sufficient for decent outcome, e.g. plenty of information is available from McM service

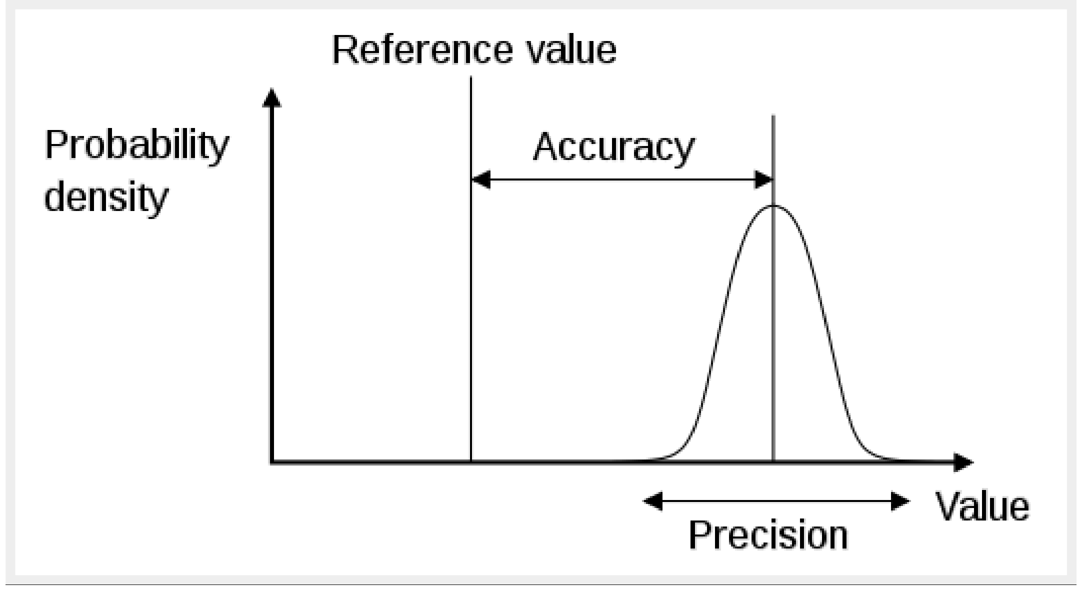
- Data seasonality effect, e.g. conference influence on dataset popularity

Plans

- I have one CS/MEng student at Cornell working on seasonality effect. The conference data were collected from CINCO and transformed into aggregated counters, e.g. how many conference we have in upcoming 1week, 2week, 1month
- We got two CERN summer students in collaboration with CERN openlab project
  - one student will work on data streaming from CMS data-services into Hadoop cluster
  - another student will work on algorithm selection problem via exploring Apache Spark MLlib framework

ACCURACY, PRECISION, RECALL AND F1





TP: true positive, TN: true negative, FP: false positive (false alarm), FN: false negative (miss)

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, Precision = \frac{TP}{TP+FP}$,

$Recall = \frac{TP}{TP+FN}$, a.k.a sensitivity, fraction of relevant instances that are retrieved

$F1 = \frac{2*Precision*Recall}{Precision+Recall}\frac{2TP}{2TP+FP+FN}$, a.k.a weighted average of the precision and recall