PREDICTING CMS DATASET POPULARITY

Valentin Kuznetsov (Cornell Univ.)

Daniele Bonacorsi (Bologna Univ.), Tony Wildish (Princeton Univ.),
Luca Giommi (Bologna Univ.), Ting Li (Cornell Univ.), Siddha Ganju (CERN openlab)

Presented by: D. Bonacorsi

ACAT 2016
18-22 January 2016
UTFSM, Valparaíso (Chile)
Chile/Continental timezone

ANALYTICS IN CMS

- Long-term goal (2-3 years)
  - build adaptive data-driven models of CMS Data and Workflow Management
  - make predictions: predict future behaviours from measurements of past performances
- Short-term goal (in Run-2)
  - support CMS Computing operations
- Why adaptive modelling?
  - models of the past aren't going to apply to the future for long..
  - only adaptive modelling will give us confidence and predictive power in the long term

POPULARITY: PROBLEM STATEMENT

- In Run I we collected 10B raw data events, and produced O(15B) MC
- Data Transfer throughput peeks at 5GB/s (on average, few PB/week), distributing more than 100PB of replicas to a complex topology of sites
- On average 250 users send up to 200K jobs/day, accessing distributed data
- We collect O(TB) of meta-data/year
- We can use historical information to predict dataset **popularity** in future
  - improve data placement; better site utilization and job scheduling

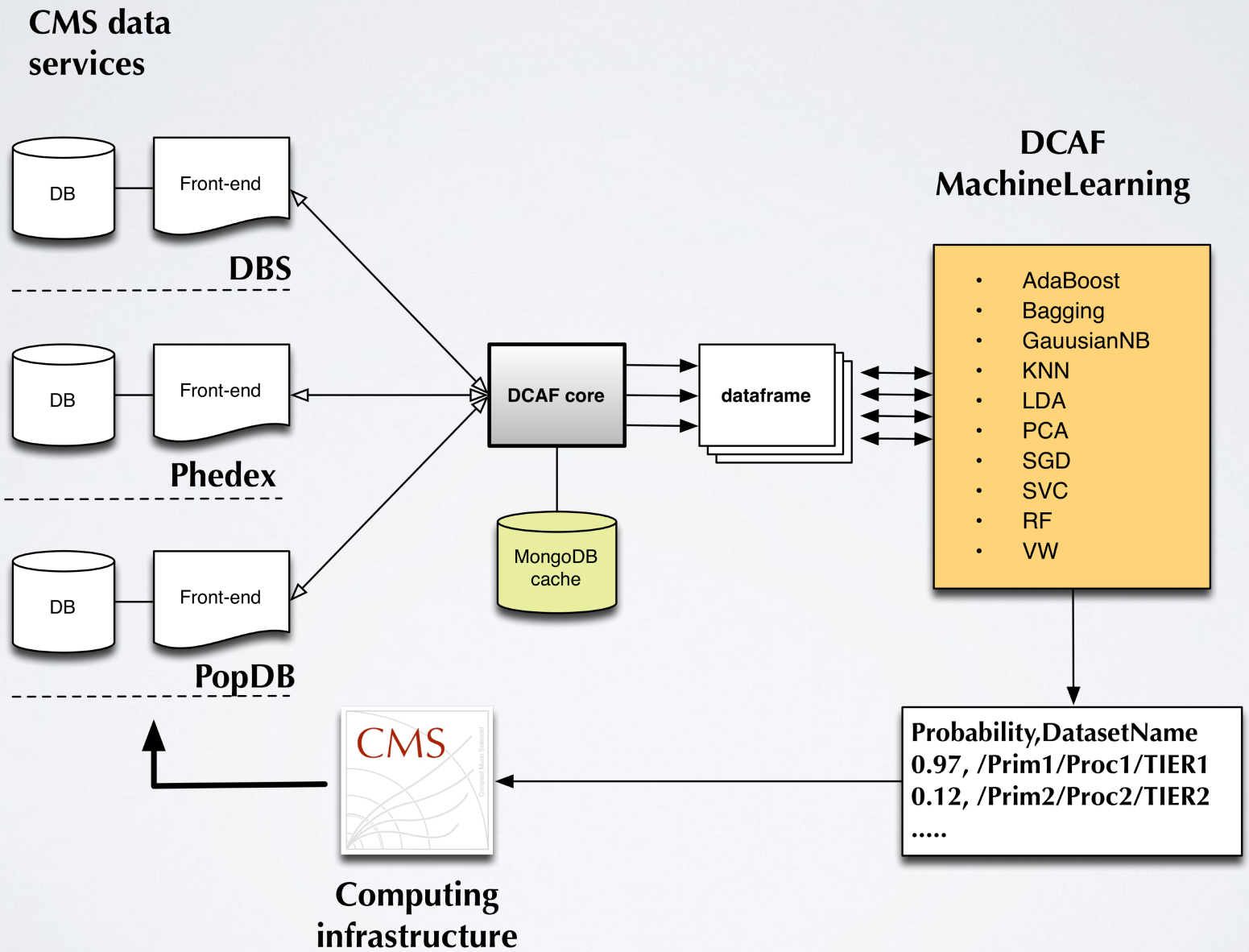
DATA SOURCES

- STRUCTURED sources:
  - **DBS**: data bookkeeping system knows about dataset meta-data
  - **PhEDEx**: data location and transfer system
  - **DAS**: data aggregation system, what users are looking for
  - **SiteDB**: people DNs and site info
  - **Dashboard**: job information system (and much more)
  - **PopularityDB**: historical data on access counts
- UNSTRUCTURED sources: **twikies**, **HN** lists, CERN **indico**, …
  - at this point we're not (yet?) working on any of them (would require data-mining on its own)

DATA COLLECTION

- Query CMS services:
  - 800K queries, 200K datasets, 900 releases, 500 site entries and 5000 user DNs
- Anonymize and factorize all data
- One year of data translates into 82x600000 data frame, in total worth of 1.5M measurements and cover 2013-2015 years, total size of dataset is ~80MB
- Complement data with CINCO conference counts
  - # conferences in N-th week from date of a given dataset access

DATA FLOW


CMS data services
DB
Front-end
DBS
DB
Front-end
Phedex
DB
Front-end
PopDB
DCAF core
MongoDB cache
dataframe
DCAF MachineLearning
AdaBoost
Bagging
GauusianNB
KNN
LDA
PCA
SGD
SVC
RF
VW
Probability,DatasetName
0.97, /Prim1/Proc1/TIER1
0.12, /Prim2/Proc2/TIER2
.....
CMS
Computing infrastructure

ROLLING FORECAST

- Collect chunk of data and train the forecasting model
  - use different ML algorithms (RF, SGD, SVC, VW, xgboost) and check their performance
  - Compare algorithms results, either choose best or use ensemble model
- Predict data for up-coming week
- Verify predictions once data become available in PopDB
- Merge predicted week into original chunk, collect new data and repeat

POPULARITY METRICS

- The definition of popularity in terms of metrics will influence model precision and cost function
- Use PopDB information and optimise its metrics against False Positive (FP) yield
  - FP rate can be translated into data transfer overhead, while FN can be treated as job latency overhead
  - Perform studies of different cuts based on #access, #users/day, totcpu metrics from PopDB

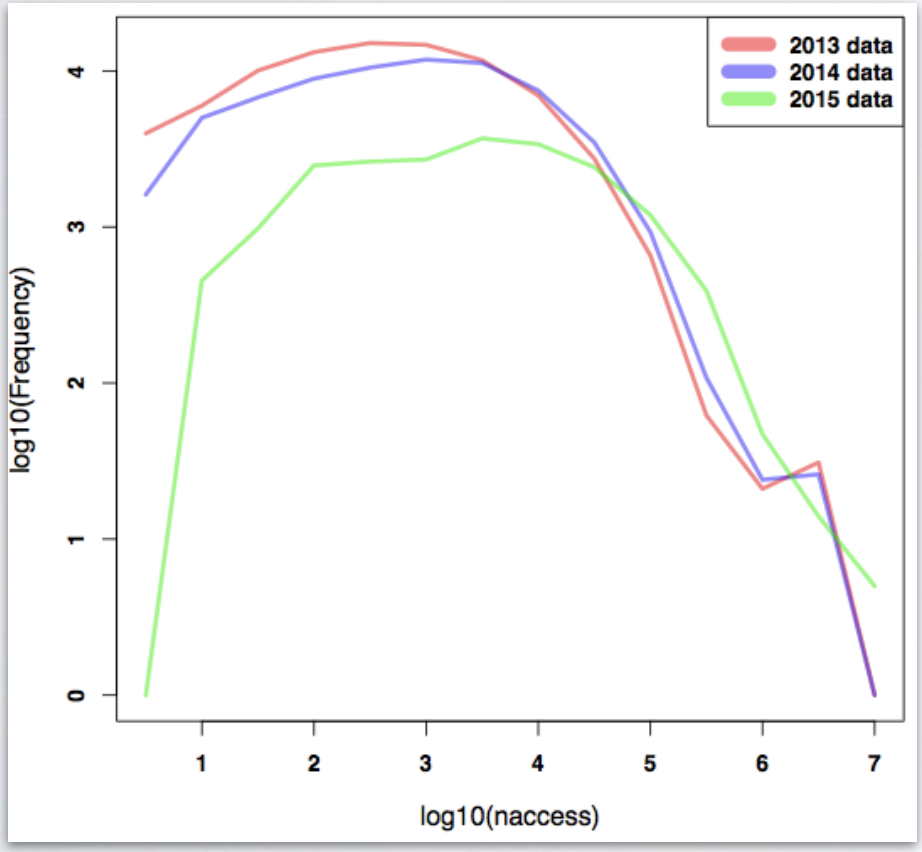
MORE ON FP, FN RATES

- FP, FN rates can be easily translated into data-transfer overhead or latency of CRAB jobs, respectively
- In 2014 we recorded ~565 ± 300 datasets every week in DBS. Using 1% FB rate and ~2TB average size of dataset this translates into ~10TB of additional data transfer/site usage.
- FN rate can be interpreted as a "normal" latency of CRAB (the CMS distributed analysis toolkit) jobs waiting for "hot" datasets if such datasets reside on a single site.

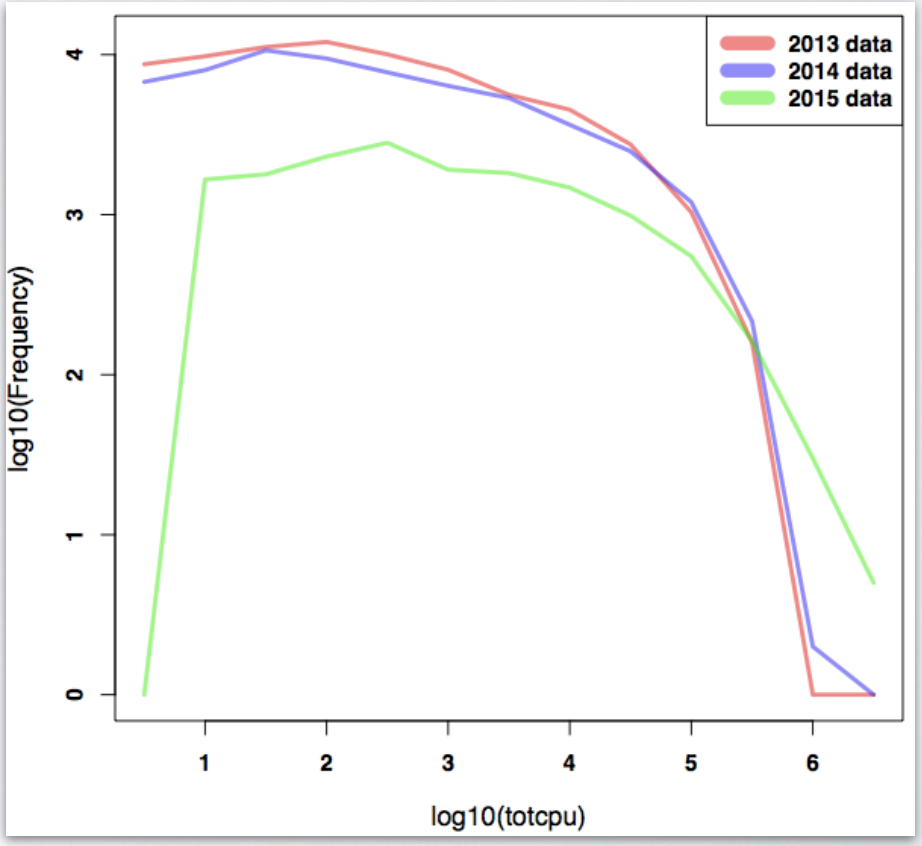
POPULARITY METRICS

NOTE: data collection is only the start, it needs validation, cleaning, transformation: so far, for 2015 we show only first half of the year

e.g. # accesses

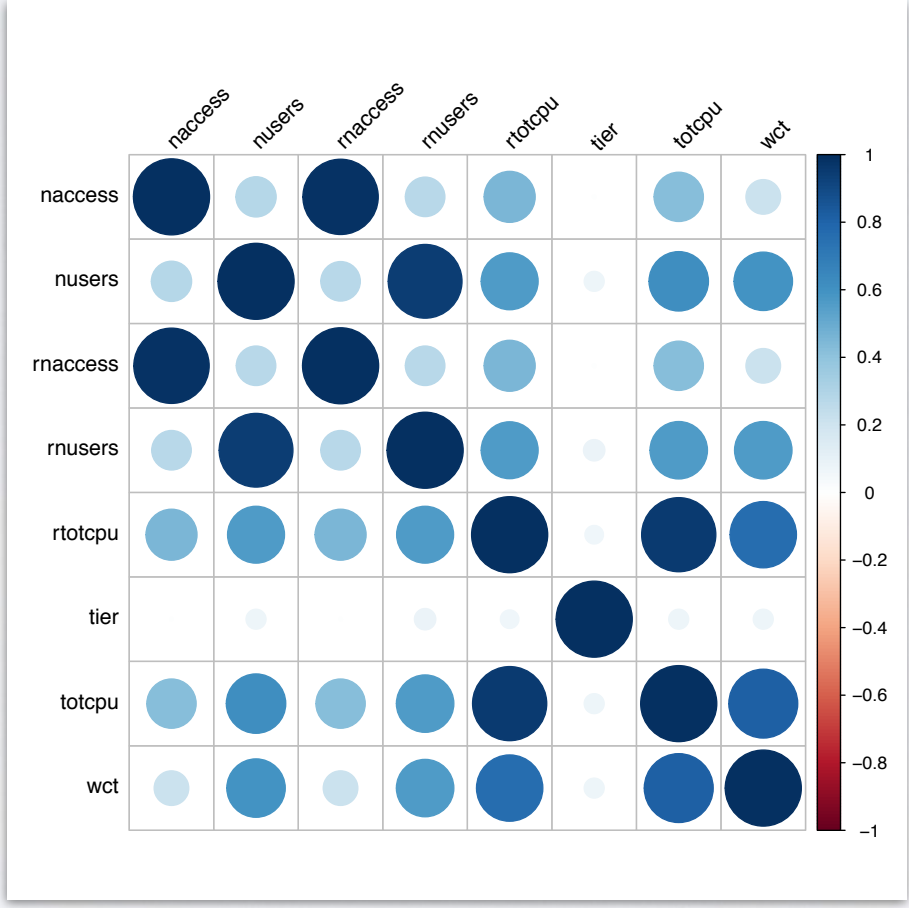


e.g. tot CPU hrs



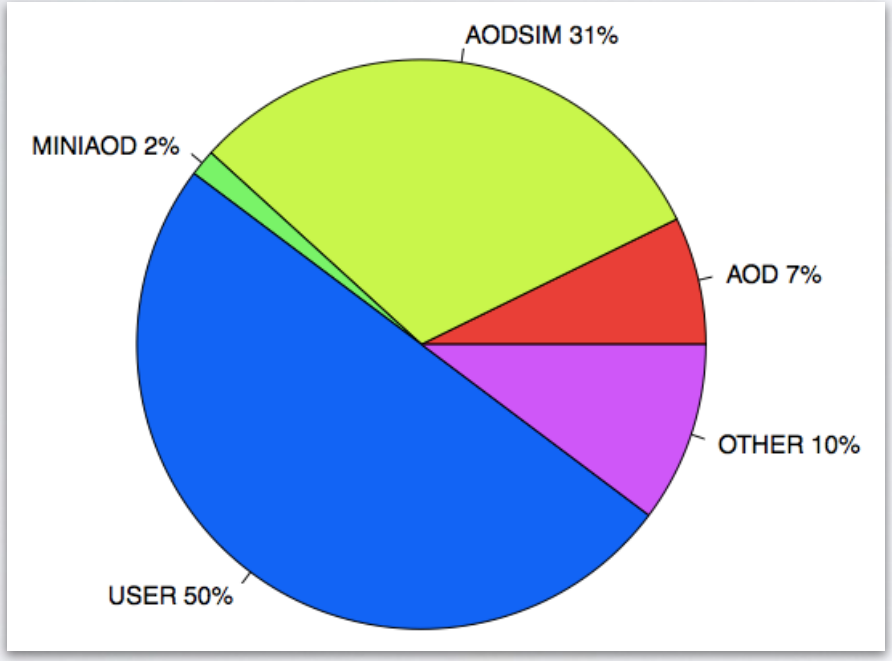
POPULARITY METRICS

The collected raw data do not always look good.
They also have plenty of correlations which we studied

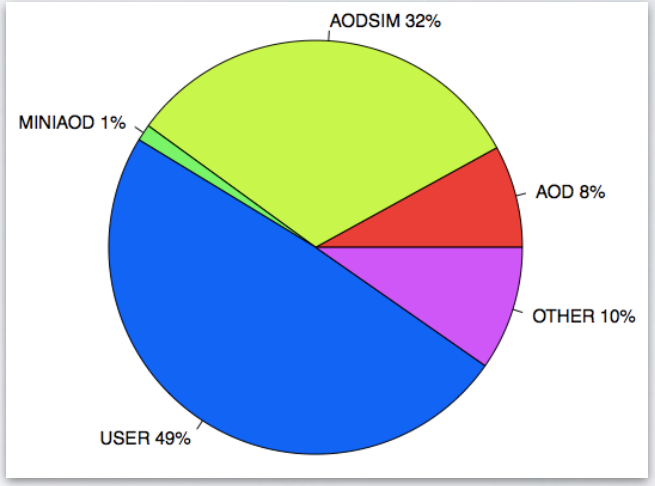


POPULARITY METRICS

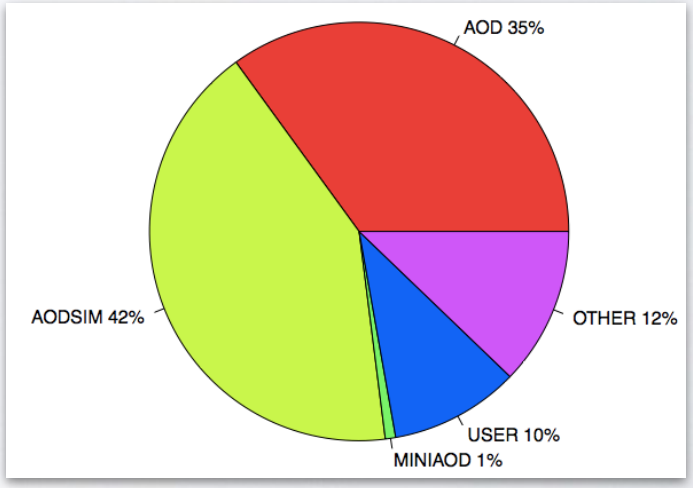
2014, prior to any cut



accesses > 10

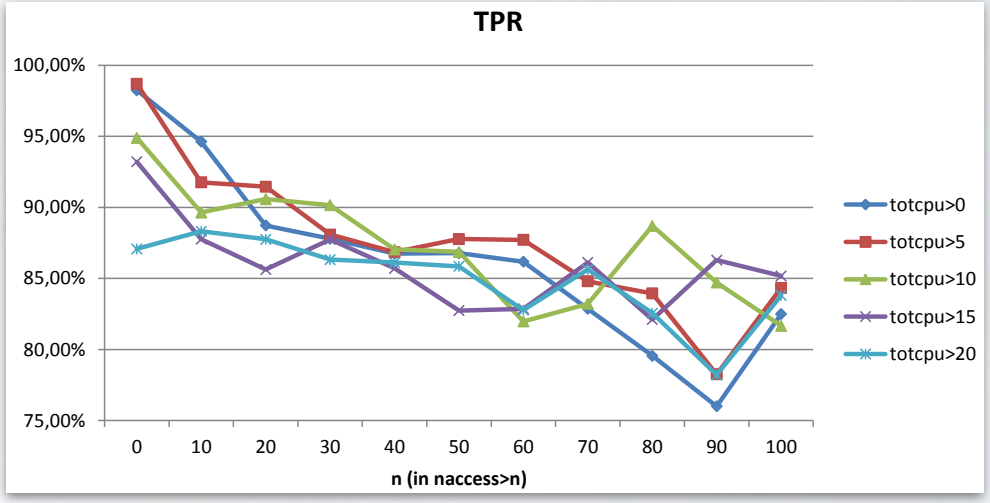


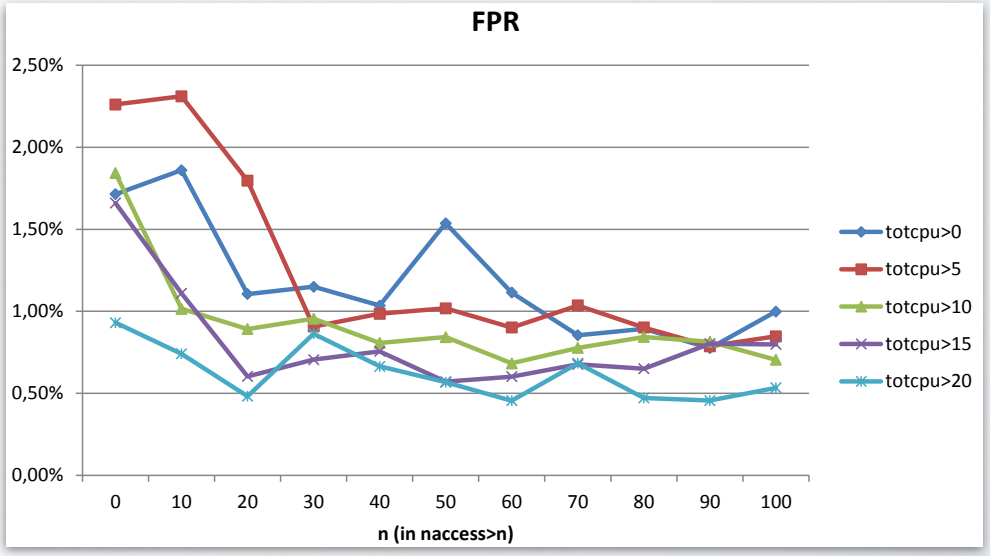
users > 2



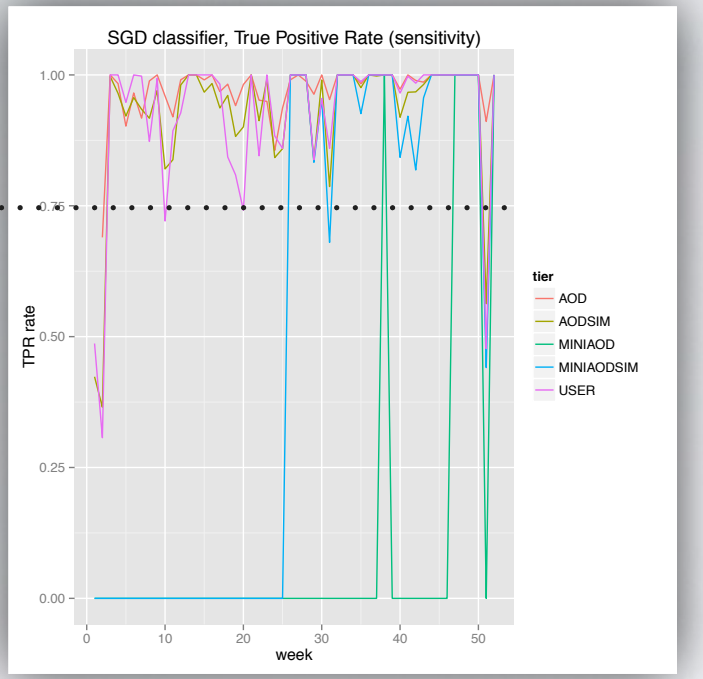
POPULARITY METRICS

- Define popular dataset as those which passed a given cut, perhaps even combined cut
- A good choice:
  - # accesses>10 & tot CPU hrs >10
- Train model (see next) and look at FP yield
- Study cut effect on yield of FP vs data tier

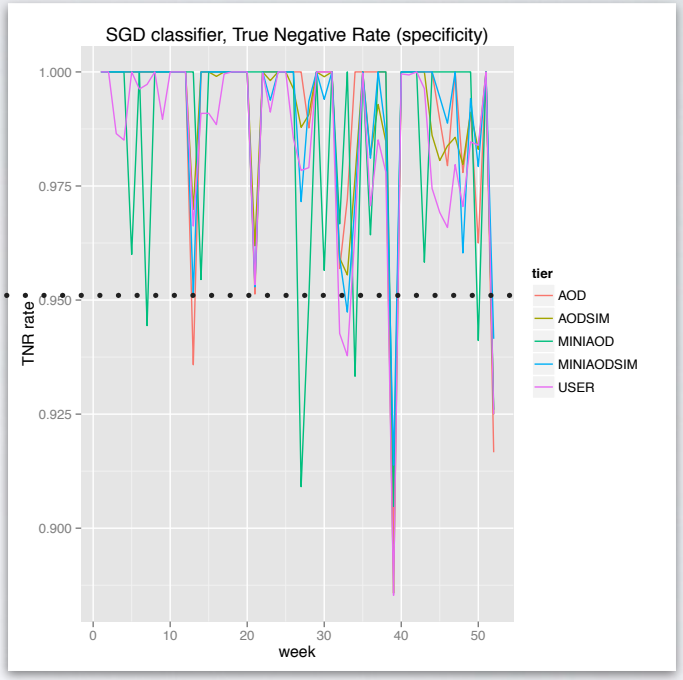




75%



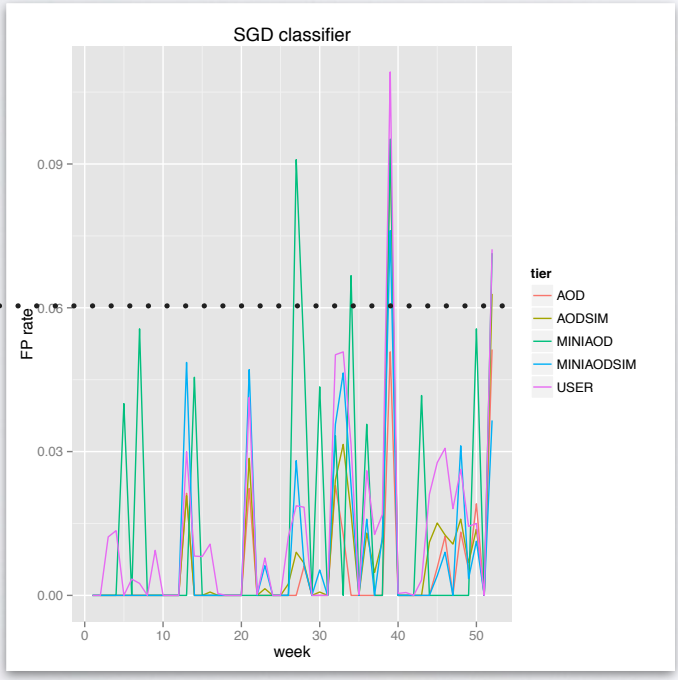
95%



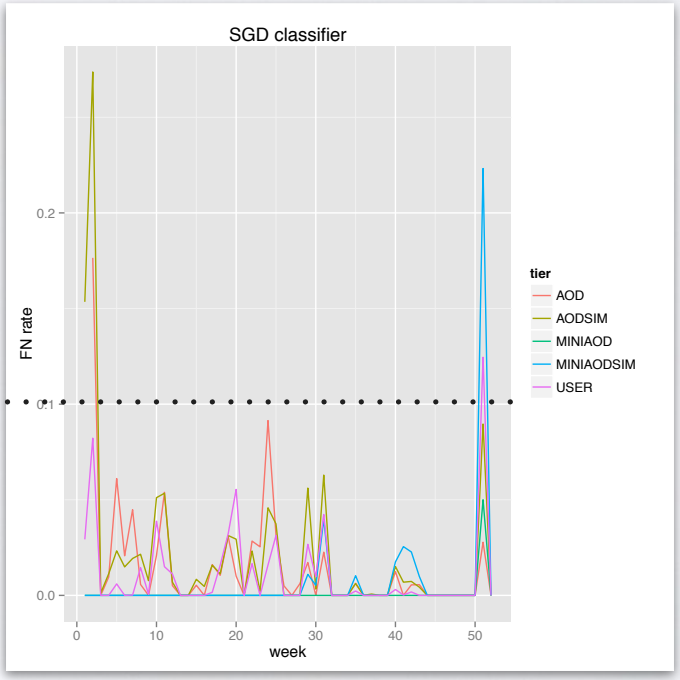
TPR=TP/(TP+FN)

TNR=TN/(TN+FP)

6%



10%



FP

FN

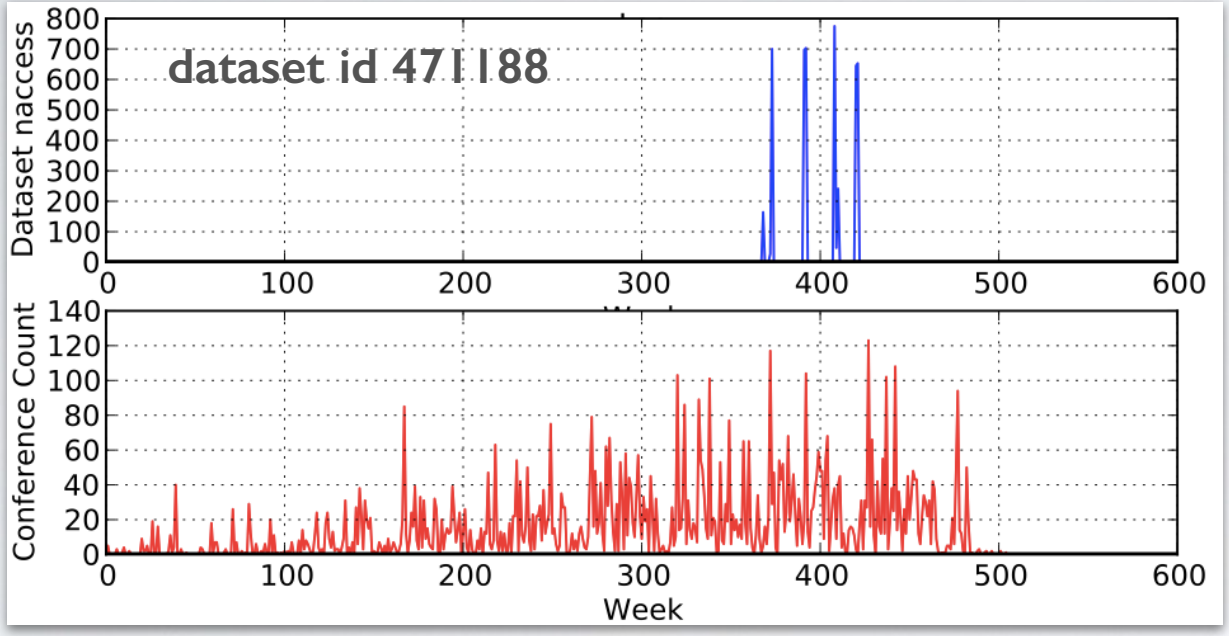
OBTAINED STATISTICS

Data Tier	TPR= TP/(TP+FN)	TNR= TN/(TN+FP)	FPR= FP/(FP+TN)	PPV= TP/(TP+FP)	NPV= TN/(TN+FN)	FP	FN
AOD	0.97+-0.05	0.99+-0.02	0.01+-0.02	0.99+-0.02	0.97+-0.06	0.005+-0.011	0.015+-0.029
AODSIM	0.93+-0.13	0.99+-0.02	0.01+-0.02	0.97+-0.06	0.97+-0.05	0.008+-0.016	0.021+-0.045
MINIAOD	0.11+-0.32	0.99+-0.02	0.01+-0.03	0.09+-0.28	0.99+-0.01	0.014+-0.026	0.001+-0.007
MINIAODSIM	0.49+-0.48	0.99+-0.02	0.01+-0.02	0.47+-0.47	0.99+-0.04	0.009+-0.016	0.007+-0.031
USER	0.93+-0.15	0.98+-0.02	0.02+-0.02	0.90+-0.15	0.99+-0.03	0.014+-0.021	0.011+-0.023

SEASONALITY EFFECT

- Extract conference counters from the CINCO conference database, append to data frame with datasets metadata and study effect of periodicity
  - for every dataset access in a given week, we added to the data frame also the # of conferences in {1,2,4,6,10+5*N, until 70} weeks in the future
- Extract datasets with more than 10 records of access in a future time
- Use Discrete Fourier Transform (DFT) on time series by FFT algorithm and search for significant spikes that presents the frequency of seasonality

SEASONALITY EFFECT



DFT of access count/dataset
covers 2013-2015 years
Periodicity 10-20 weeks

DFT of conference count
covers 2006-2015 years
Periodicity every 50.5 weeks

Cross-correlation between conference count series and access count series of a dataset were also studied. We found that for some datasets cross correlation is peaking at a positive lag which means that future conference schedules can affect the current dataset access. This is more evident than the observation that past conferences still have residual influence on the current dataset access, which is anyway observed for a subset of datasets.

SEASONALITY EFFECT

- Data on datasets access patterns shows some seasonality effect
- Conference counters can be used for prediction without other meta-data attributes, but they're less significant with respect to CMS meta-data attributes. Cross correlation need to be explored in more depth
- We see seasonality effect for some datasets but further analysis (e.g. for specific data-tier) should be pursued

SUMMARY

- We equipped CMS with a machinery that allows to extract precious information to feed into adaptive models of CMS Computing Operations in selected areas, and used it for the first use-case (data popularity)
- We understand and are able to predict CMS dataset popularity on regular basis
  - we are focussing on AOD+USER datasets
  - We clearly observed a MINIAOD/MINIAODSIM popularity ramp
- There is evidence of seasonality effect for some datasets, but more work is needed to exploit this in terms of predictive potentiality