ECML-PKDD 2015

BIKE SHARING MODEL REUSE FRAMEWORK

FOR TREE-BASED ENSEMBLES

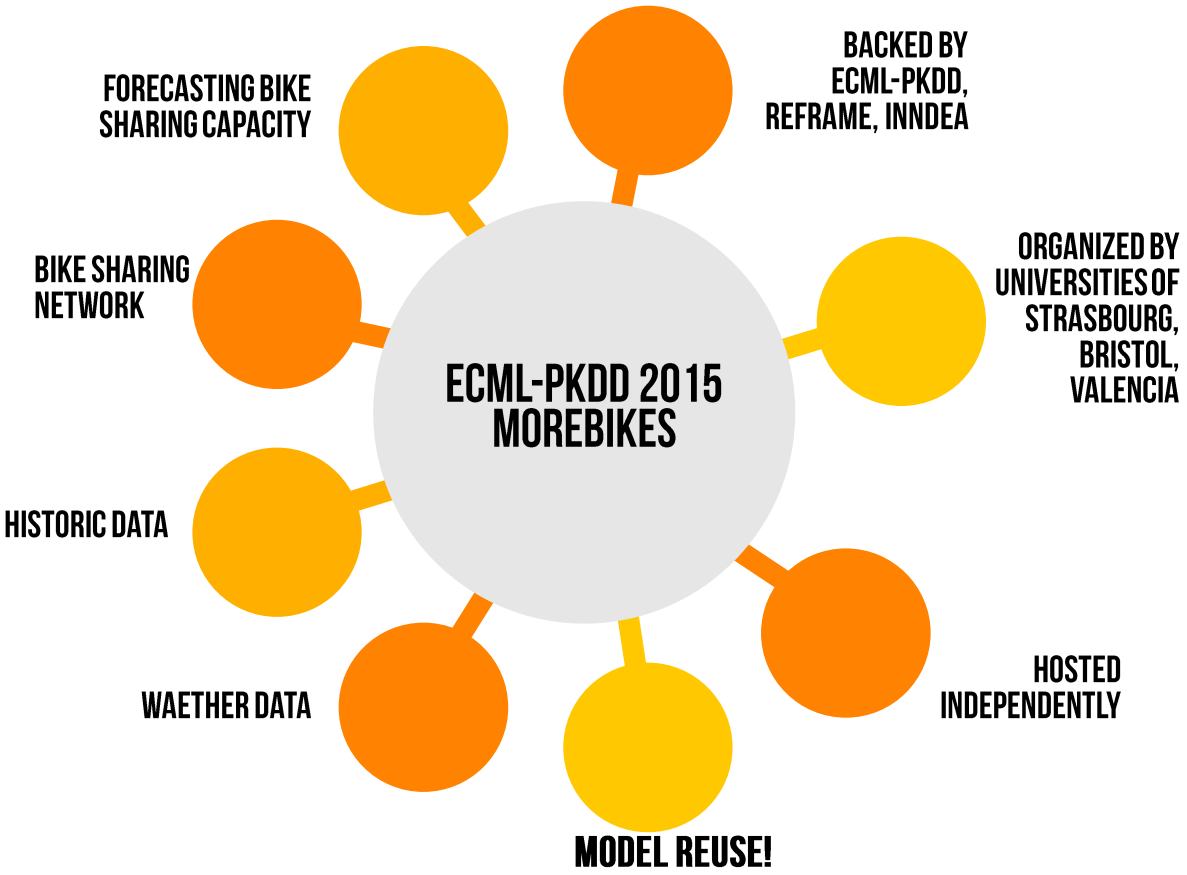
energyforecasting.io

dmlab

GERGO BARTA
BARTA@TMIT.BME.HU

MOREBIKES - MODEL REUSE WITH BIKE RENTAL STATION DATA

MOREBIKES - MODEL REUSE WITH BIKE RENTAL STATION DATA



3RD PLACE
WIND POWER FORECASTING
GEFCOM 2012

2ND PLACE
ELECTRIC LOAD FORECASTING
RWE NPOWER 2015

2ND PLACE
SOLAR & WIND POWER FORECASTING
GEFCOM 2014

MOTIVATION

ERASMUS, VLC, 2009

3RD PLACE
WIND POWER FORECASTING
GEFCOM 2012

2ND PLACE
ELECTRIC LOAD FORECASTING
RWE NPOWER 2015

2ND PLACE
SOLAR & WIND POWER FORECASTING
GEFCOM 2014

CHALLENGE APPROACH

THE DATA

1. 1-10  2.5 YRS
2. 11-200  1 MONTH
3. 201-275  1 MONTH  3 MONTHS TEST
4. BASELINE & LINEAR MODELS 
5. ERROR RATES 

THE DATA

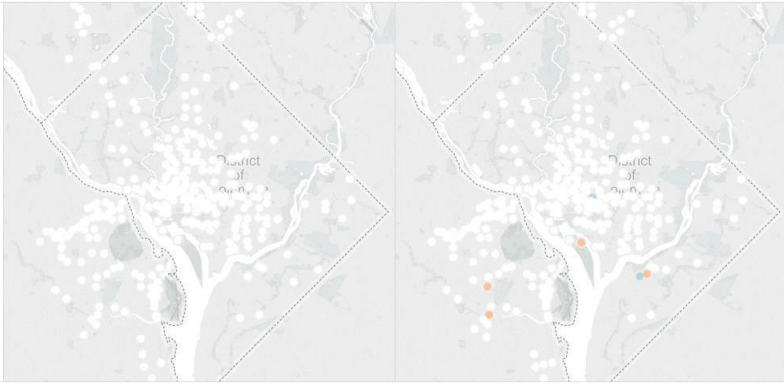
1. 1-10
2. 11-200 1 MONTH
3. 201-275 1 MONTH
4. BASELINE & LINEAR MODELS
5. ERROR RATES

3 MONTHS TEST

EXPLORING THE DATA

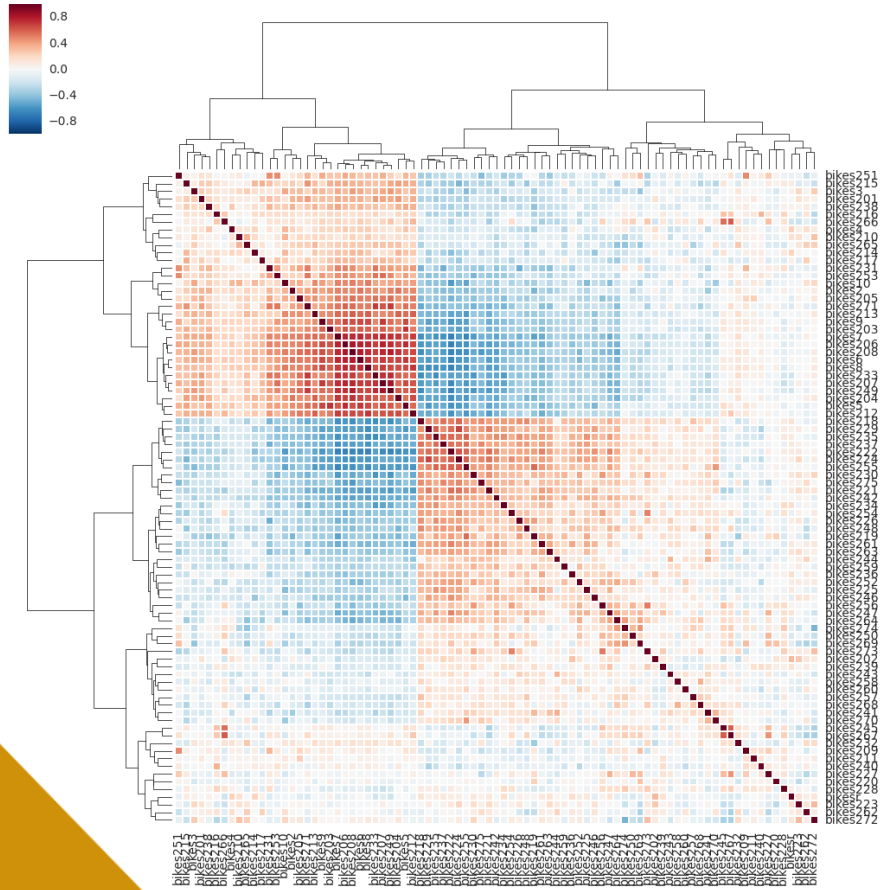
1. BIKE SHARING NETWORK
2. OBSERVING BEHAVIOUR
3. +/- CORRELATIONS (PEARSON,DTW)
4. STANDARDIZE DATA BY STATION SIZE

0



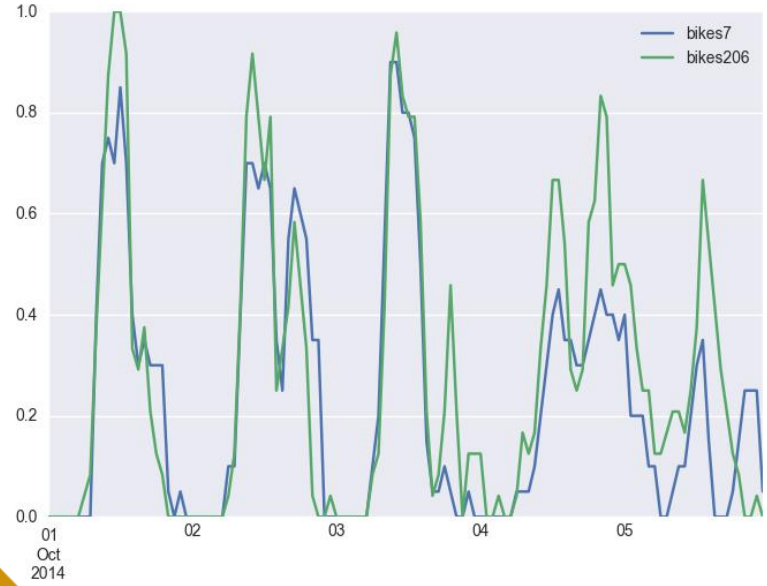
EXPLORING THE DATA

1. BIKE SHARING NETWORK
2. OBSERVING BEHAVIOUR
3. +/- CORRELATIONS (PEARSON,DTW)
4. STANDARDIZE DATA BY STATION SIZE



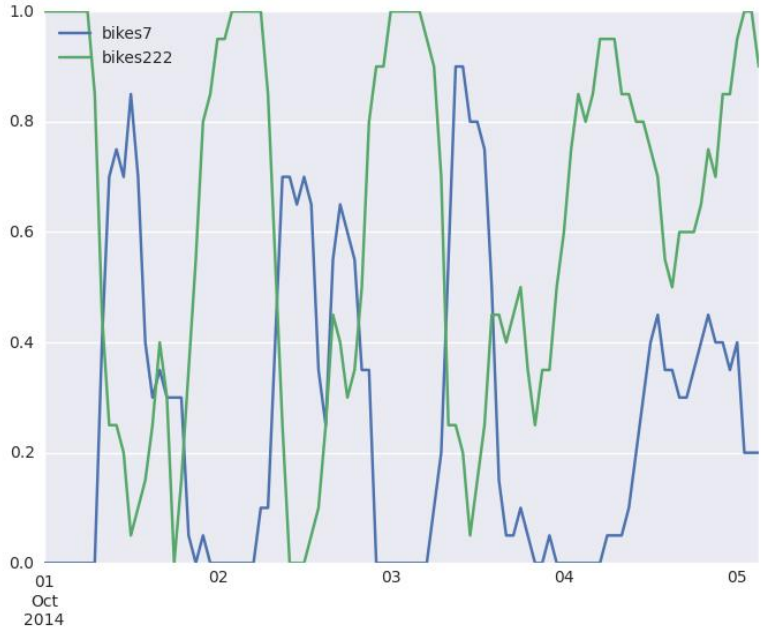
EXPLORING THE DATA

1. BIKE SHARING NETWORK
2. OBSERVING BEHAVIOUR
3. +/- CORRELATIONS (PEARSON,DTW)
4. STANDARDIZE DATA BY STATION SIZE

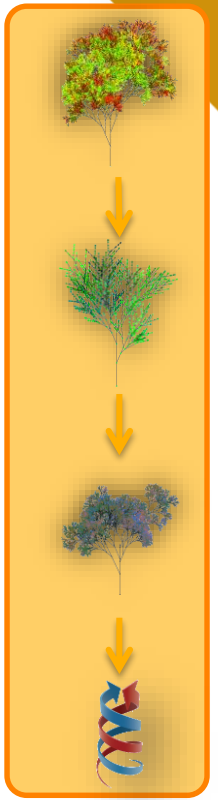


EXPLORING THE DATA

1. BIKE SHARING NETWORK
2. OBSERVING BEHAVIOUR
3. +/- CORRELATIONS (PEARSON,DTW)
4. STANDARDIZE DATA BY STATION SIZE



FRAMEWORK ARCHITECTURE



GRADIENT BOOSTING REGRESSOR

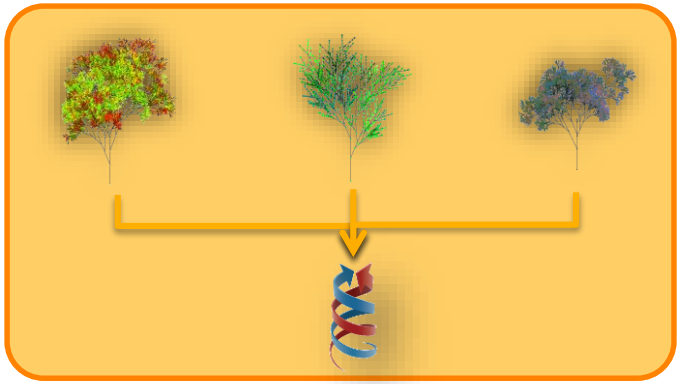
1. FRIEDMAN, 1999
2. ROBUST COMBINATION OF WEAK LEARNERS
3. RESISTANT TO OVERFITTING, FEW TUNING PARAMETERS, PROVIDES VARIABLE IMPORTANCE
4. QUANTILE REGRESSION CAPABLE!
5. EMPLOYS BOOSTING

STACKING

GBR + RF

7 BASELINE MODELS

ENSEMBLING



RANDOM FOREST REGRESSOR

1. BREIMAN, 2001
2. ROBUST COMBINATION OF WEAK LEARNERS
3. RESISTANT TO OVERFITTING, FEW TUNING PARAMETERS, PROVIDES VARIABLE IMPORTANCE
4. QUANTILE REGRESSION CAPABLE!
5. EMPLOYS BAGGING

CHALLENGES

1. DEVELOP A LOCAL VALIDATION SCHEME
2. INCLUDE RELEVANT INFORMATION
3. REDUCE NOISE
4. INCREASE ACCURACY

CHALLENGES

1. DEVELOP A LOCAL VALIDATION SCHEME
2. INCLUDE RELEVANT INFORMATION
3. REDUCE NOISE
4. INCREASE ACCURACY

SOLUTION

- 2 EVALUATION PERIODS: H2 & Q4 OCT
- PERFORMS SUPRISINGLY WELL

CHALLENGES

1. DEVELOP A LOCAL VALIDATION SCHEME
2. INCLUDE RELEVANT INFORMATION
3. REDUCE NOISE
4. INCREASE ACCURACY

SOLUTION

- FIND SUITABLE TRAINING SET DONORS (PEARSON)
- SWAP CAPACITY ATTRIBUTES, INCLUDE NEGATIVELY CORRELATED DONORS

Input variable	Transformation applied
bikes	1 - bikes
bikes_3h_ago	1 - bikes_3h_ago
full_profile_bikes	1 - full_profile_bikes
short_profile_bikes	1 - short_profile_bikes
short_profile_3h_diff_bikes	(-1) * short_profile_3h_diff_bikes
full_profile_3h_diff_bikes	(-1) * full_profile_3h_diff_bikes

CHALLENGES

1. DEVELOP A LOCAL VALIDATION SCHEME
2. INCLUDE RELEVANT INFORMATION
3. REDUCE NOISE
4. INCREASE ACCURACY

SOLUTION

- INDIVIDUAL MODEL(!) FOR EACH TEST CASE (500)
- W/ SAME PARAMETERS
- DISCARD IRRELEVANT DATA POINTS
- EFFICIENT NEAREST NEIGHBOUR APPROACH (KDTREE)

CHALLENGES

1. DEVELOP A LOCAL VALIDATION SCHEME
2. INCLUDE RELEVANT INFORMATION
3. REDUCE NOISE
4. INCREASE ACCURACY

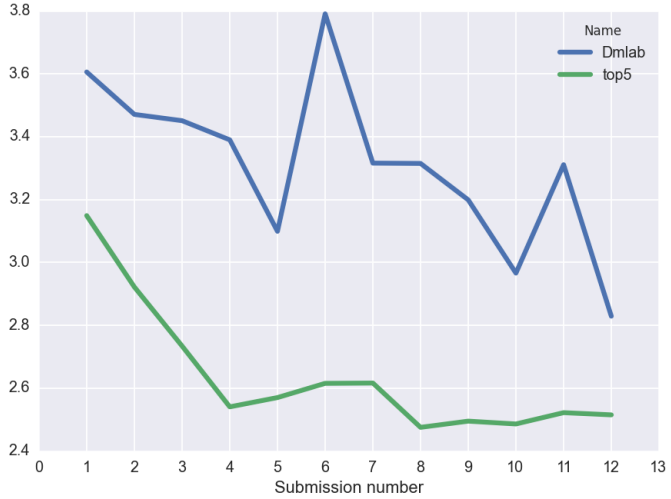
SOLUTION

- 3-FOLD OPTIMIZATION USING LOCAL CV
  - TRAINING SET DONOR SELECTION
  - MODEL PARAMETERS (2X)
  - NEAREST NEIGHBOURS

METHOD EVOLUTION & RESULTS

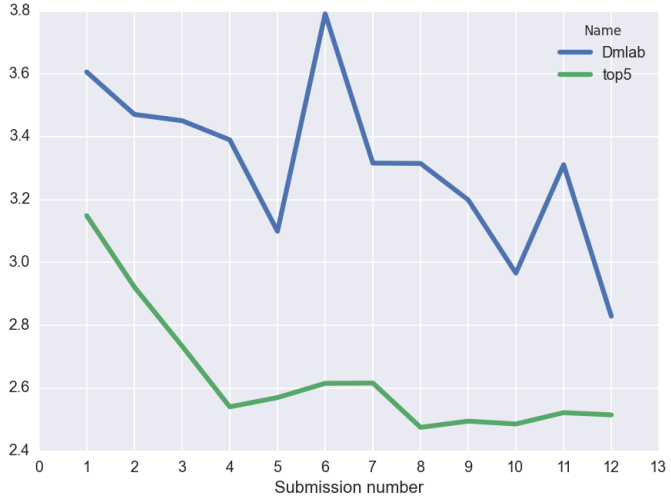
METHOD EVOLUTION & RESULTS

LEDERBOARD SCORE BY SUBMISSION



METHOD EVOLUTION & RESULTS

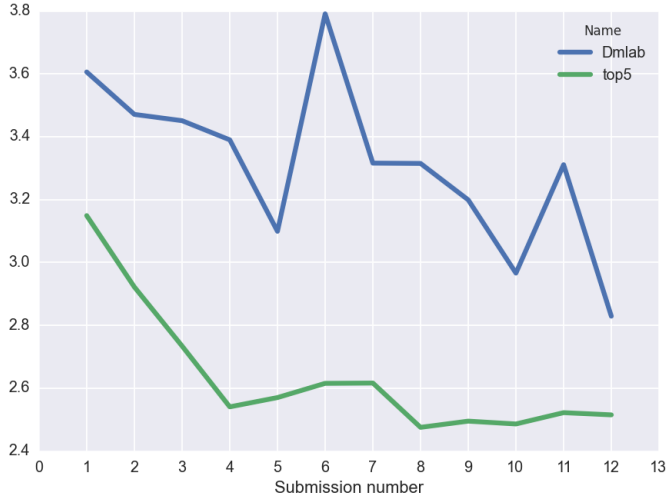
LEDERBOARD SCORE BY SUBMISSION



OCTOBER DATA ONLY (BASELINE)

METHOD EVOLUTION & RESULTS

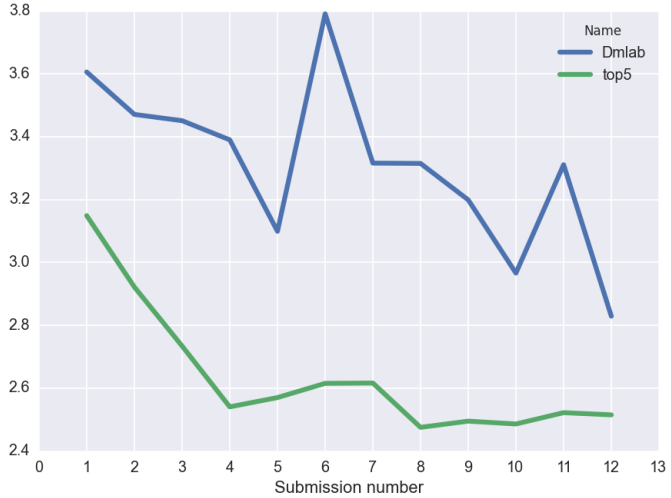
LEDERBOARD SCORE BY SUBMISSION



NEGATIVE CORRELATIONS

METHOD EVOLUTION & RESULTS

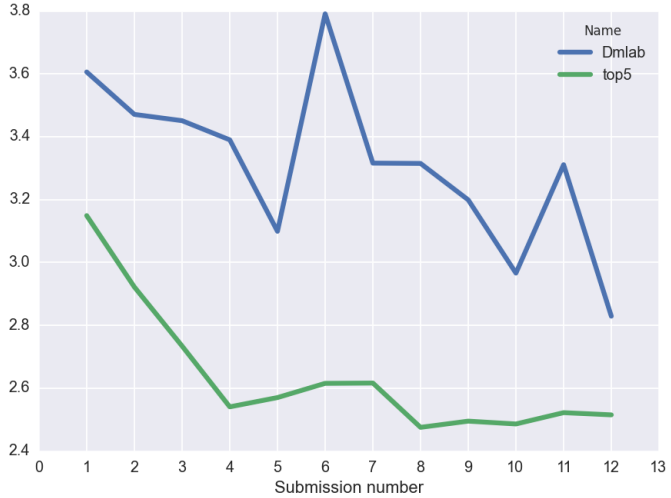
LEDERBOARD SCORE BY SUBMISSION



EXPERT MODELS + NEIGHBOUR FILTERING

METHOD EVOLUTION & RESULTS

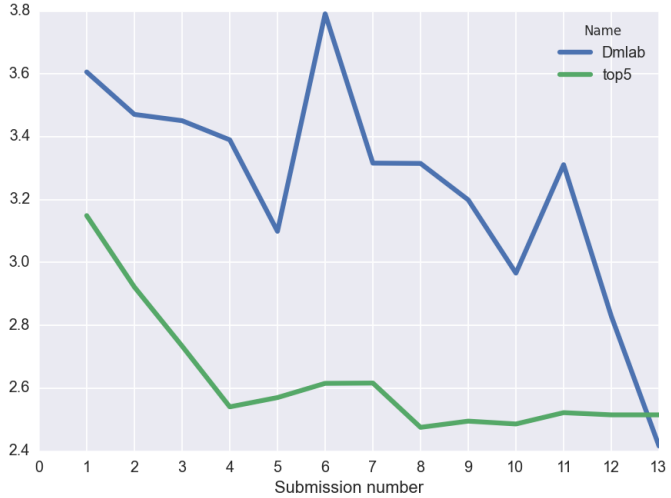
LEDERBOARD SCORE BY SUBMISSION



OPTIMIZING NEIGHBOUR ATTRIBUTES

METHOD EVOLUTION & RESULTS

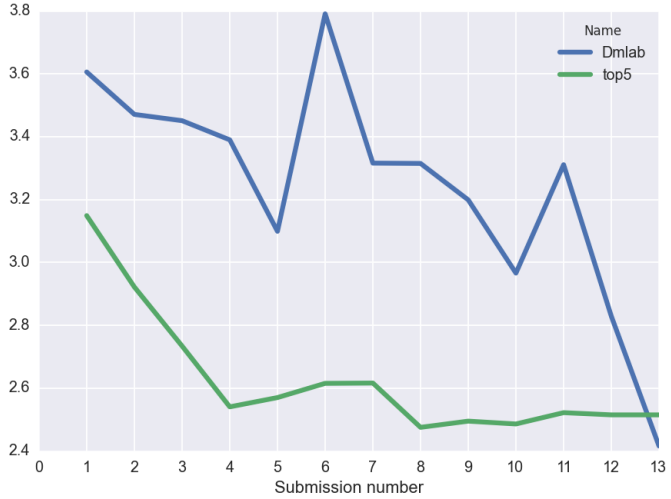
LEDERBOARD SCORE BY SUBMISSION



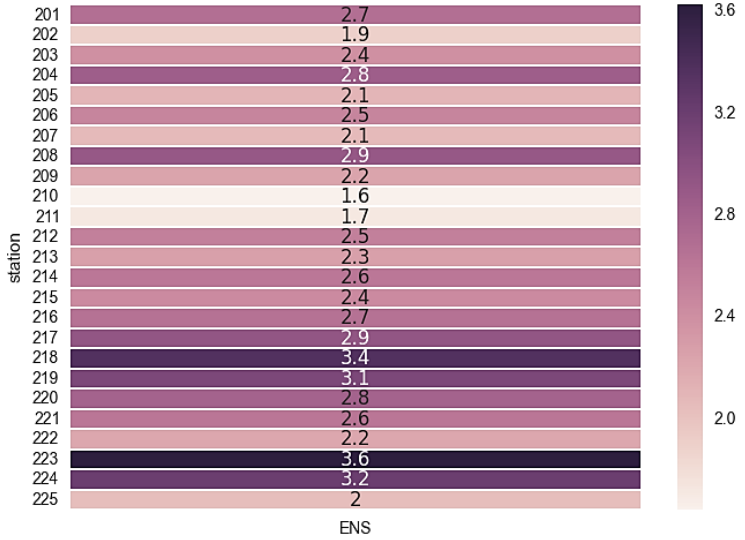
OPTIMIZING NEIGHBOUR ATTRIBUTES

METHOD EVOLUTION & RESULTS

LEDERBOARD SCORE BY SUBMISSION

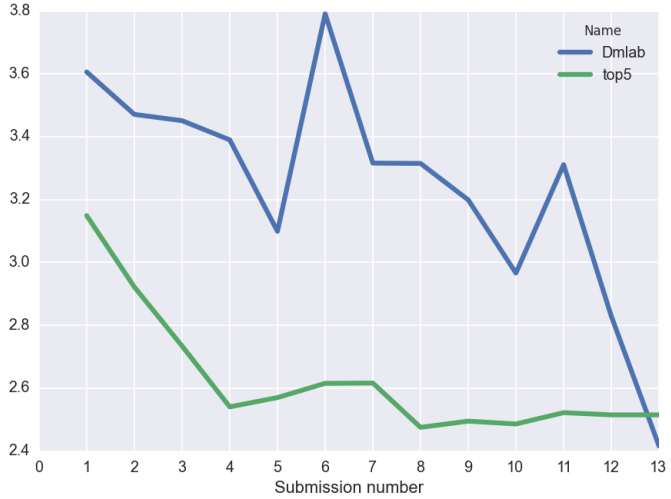


LOCAL CV ERROR BY STATION

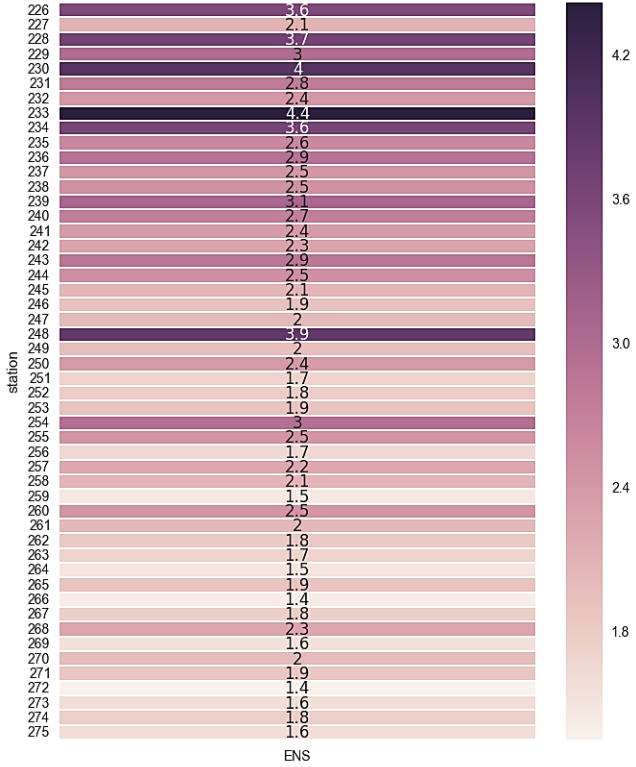


METHOD EVOLUTION & RESULTS

LEDERBOARD SCORE BY SUBMISSION



LOCAL CV ERROR BY STATION

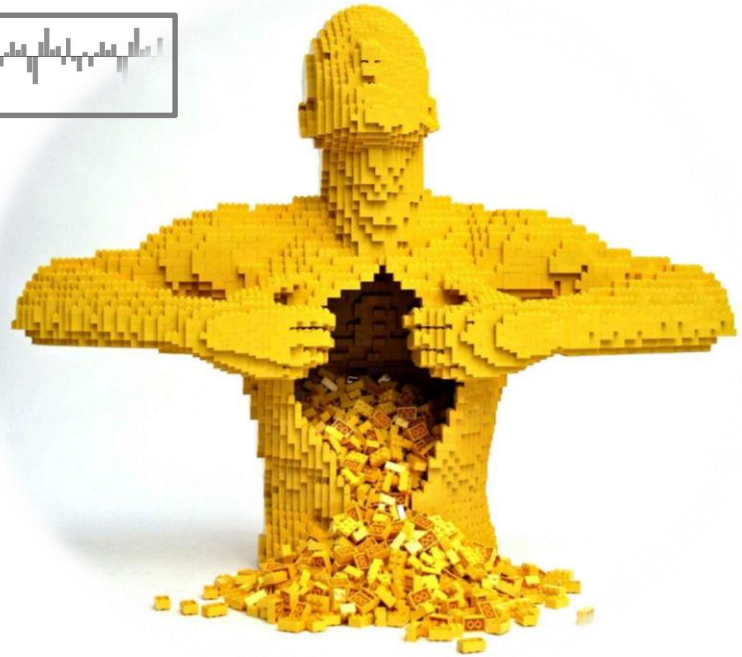


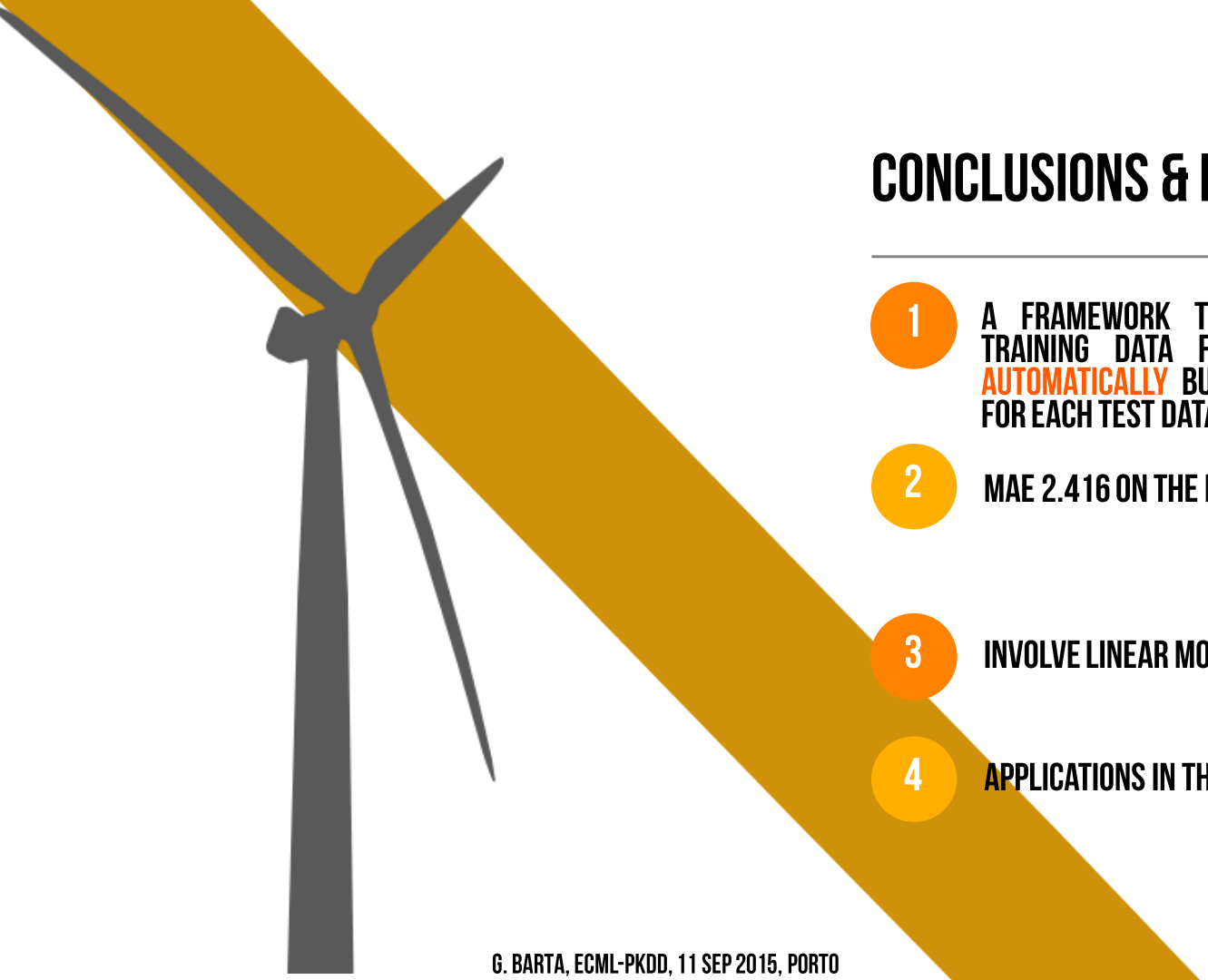
THE DATA

1. 1-10
2. 11-200 — 1 MONTH
3. 201-275 — 1 MONTH — 2-3 MONTHS TEST
4. BASELINE & LINEAR MODELS
5. ERROR RATES

THE DATA

1. 1-10
2. 11-200 1 MONTH
3. 201-275 1 MONTH 2-3 MONTHS TEST
4. BASELINE & LINEAR MODELS
5. ERROR RATES





CONCLUSIONS & FUTURE WORK

1. A FRAMEWORK THAT EFFICIENTLY PICKS SUITABLE TRAINING DATA FROM A LARGE POPULATION AND AUTOMATICALLY BUILDS ACCURATE ENSEMBLE MODELS FOR EACH TEST DATA POINT
2. MAE 2.416 ON THE FINAL TEST SET
3. INVOLVE LINEAR MODELS (STACKING, FALLBACK)
4. APPLICATIONS IN THE ENERGY FIELD (WIND)

WWW.ENERGYFORECASTING.IO

THANK YOU FOR YOUR KIND ATTENTION

GERGO BARTA – BARTA@TMIT.BME.HU