Moisture Magic

💦 🧙

UFZ “Hohes Holz” Soil Moisture Time-Series Analysis

HIDA Datathon, 2020

Maximilian Graf, Alexander Merdian-Tarko, Julius Polz, Christian Werner
@Max_Grave_, @jpolz3, @cwerner76

Source: https://github.com/HIDA-Datathon/moisturemagic.git

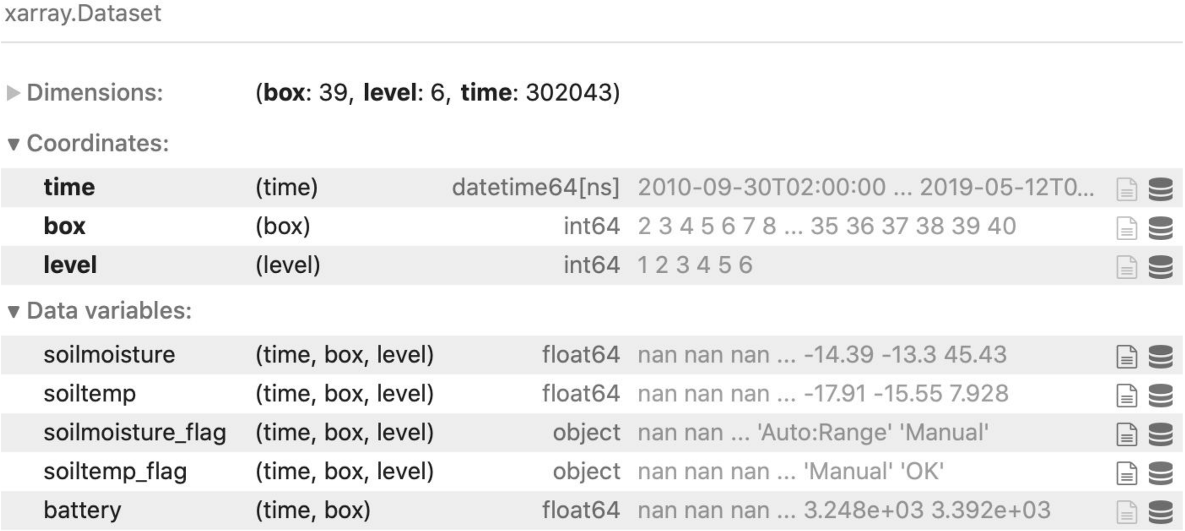
Steps

- Transform raw data into coherent netcdf format (xarray)
- Exploratory data analysis
- Semi-Unsupervised Time-Series Classification (UMAP)

Resources:
https://umap-learn.readthedocs.io/en/latest/
http://xarray.pydata.org/en/stable/

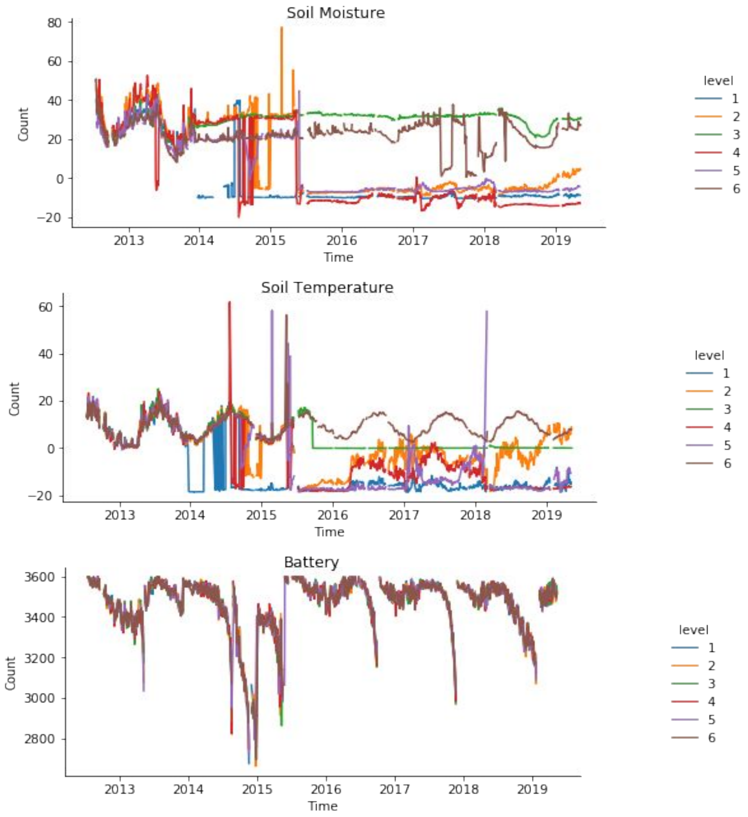
Preprocessing

- Regularize and convert raw data
- Resample time-interval to fixed 15min steps
- Coordinates: time, box (one profile), level (vertical sensor position)
- Export to netCDF file



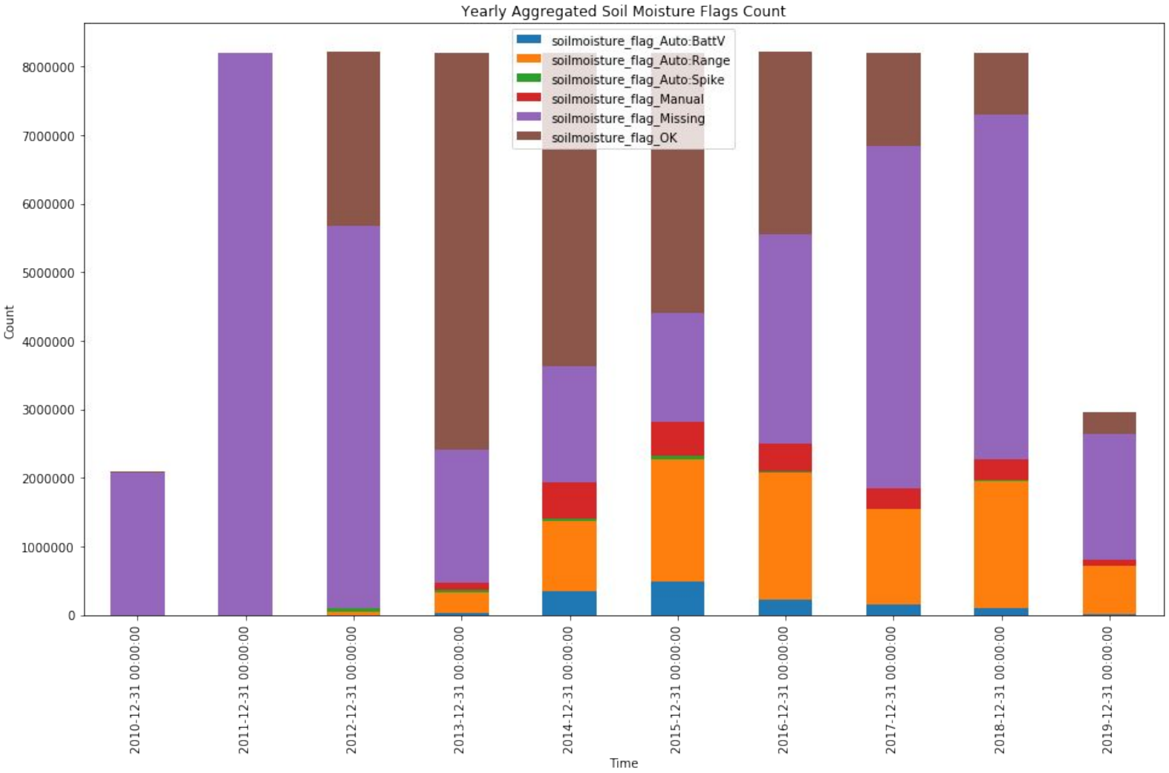
```
xarray.Dataset

▶ Dimensions:          (box: 39, level: 6, time: 302043)
▼ Coordinates:
    time               (time)              datetime64[ns]  2010-09-30T02:00:00 ... 2019-05-12T0...
    box                (box)                        int64  2 3 4 5 6 7 8 ... 35 36 37 38 39 40
    level              (level)                      int64  1 2 3 4 5 6
▼ Data variables:
    soilmoisture       (time, box, level)         float64  nan nan nan ... -14.39 -13.3 45.43
    soiltemp           (time, box, level)         float64  nan nan nan ... -17.91 -15.55 7.928
    soilmoisture_flag  (time, box, level)          object  nan nan ... 'Auto:Range' 'Manual'
    soiltemp_flag      (time, box, level)          object  nan nan nan ... 'Manual' 'OK'
    battery            (time, box)                float64  nan nan nan ... 3.248e+03 3.392e+03
```

How our data looks..



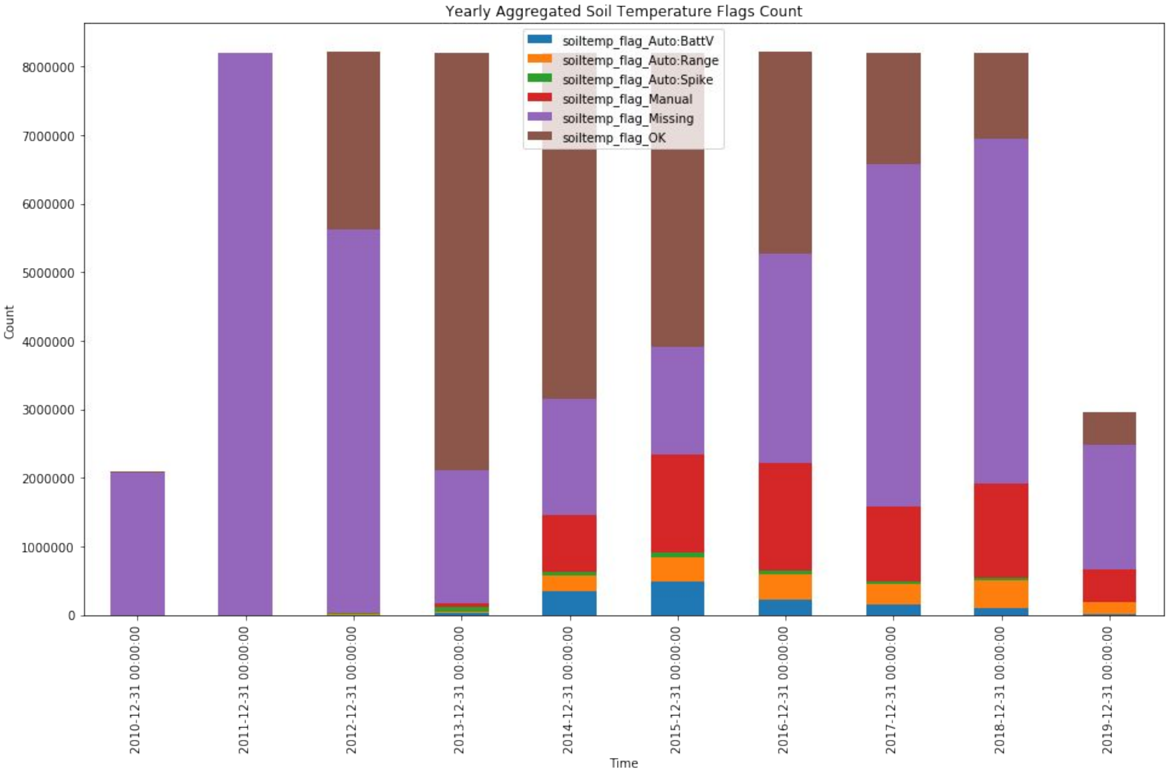
.. kinda wild 😱

Data Exploration - Soil Moisture Flags



- Yearly occurrence of soil moisture flags for almost all boxes and sensors over the entire period (2010 - 2019)
- Period 2014 - 2015 has the best data in terms of number of missing values and availability of manual flags

Data Exploration - Soil Temperature Flags



- Yearly occurrence of soil temperature flags for almost all boxes and sensors over the entire period (2010-2019)

- Period 2014-2015 has the best data in terms of number of missing values and availability of manual flags

Experimental Setup

Input

Windows of 40 time steps (10h)

One Sensor only, no neighbour data

Soil moisture+Temp+Battery

→ Input.shape = (n_samples, 40, 3)

Dataset

Train: All sensors 2014

→ n_samples = 160.000

Test: All sensors 2015

→ n_samples = 158.000

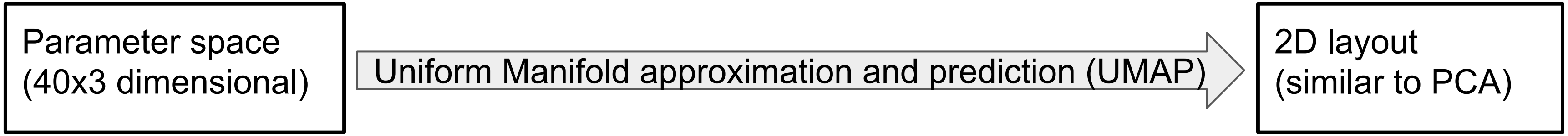
Reference

contains a temp **or** moisture flag

Goal

Unsupervised detection of flags

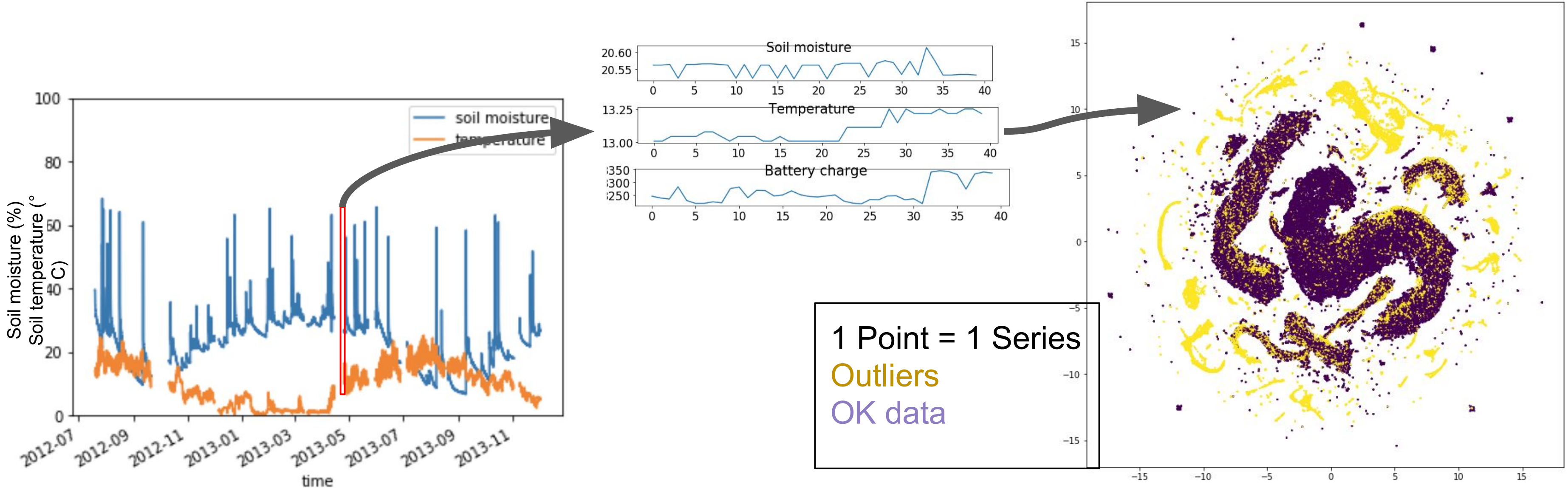
Semi-Unsupervised TS Classification using UMAP



Parameter space (40x3 dimensional)

Uniform Manifold approximation and prediction (UMAP)

2D layout (similar to PCA)



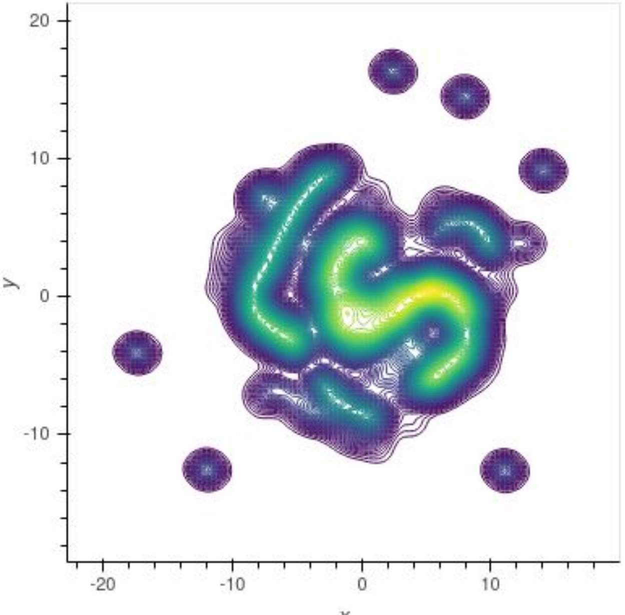
1 Point = 1 Series

Outliers

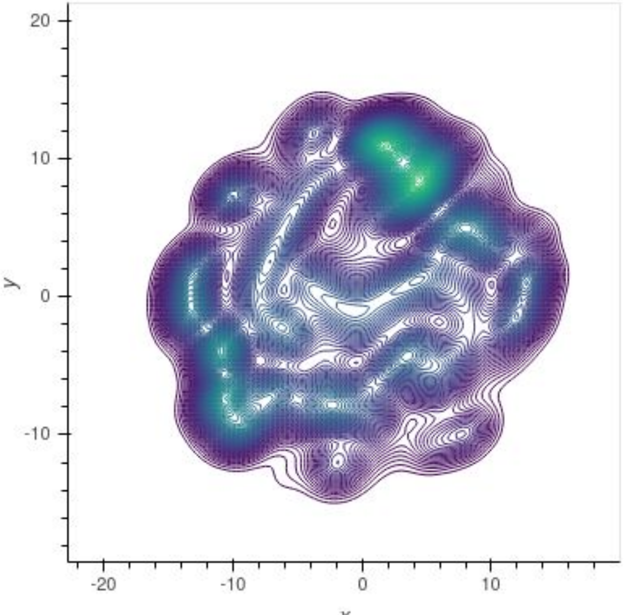
OK data

Density of points in 2D layout shows differences

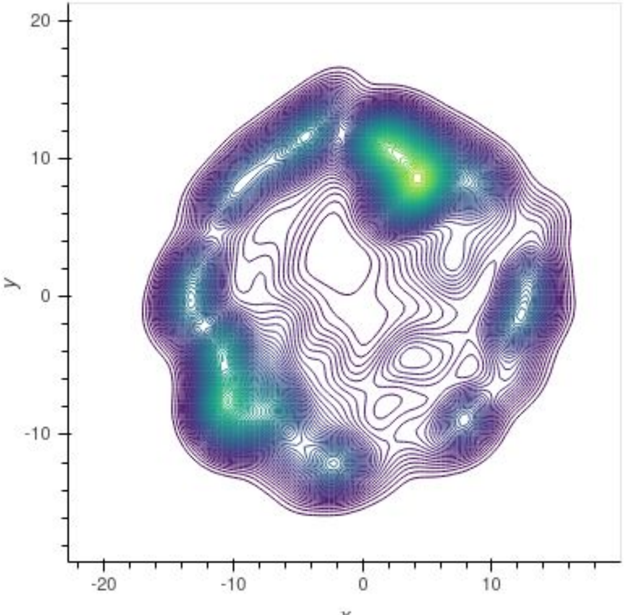
Flag “OK”



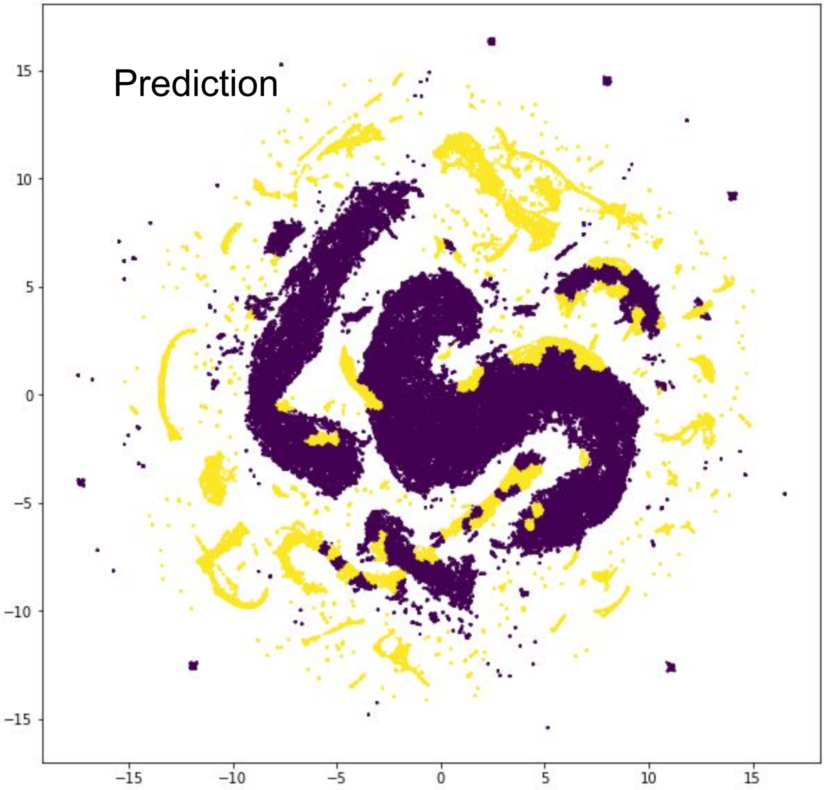
Flag “Manual”

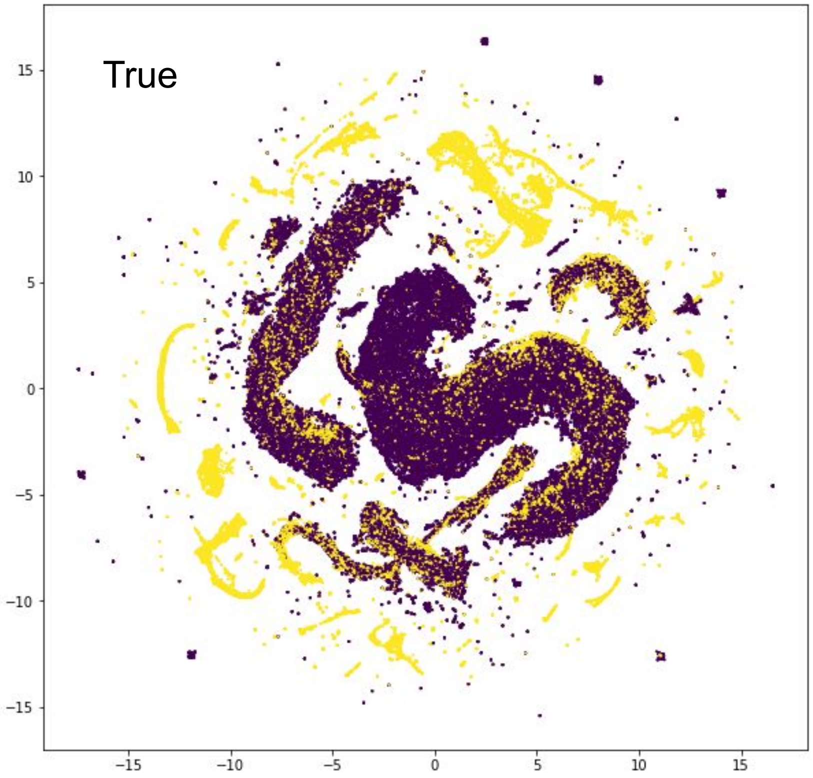


Flag “Auto:XXXX”

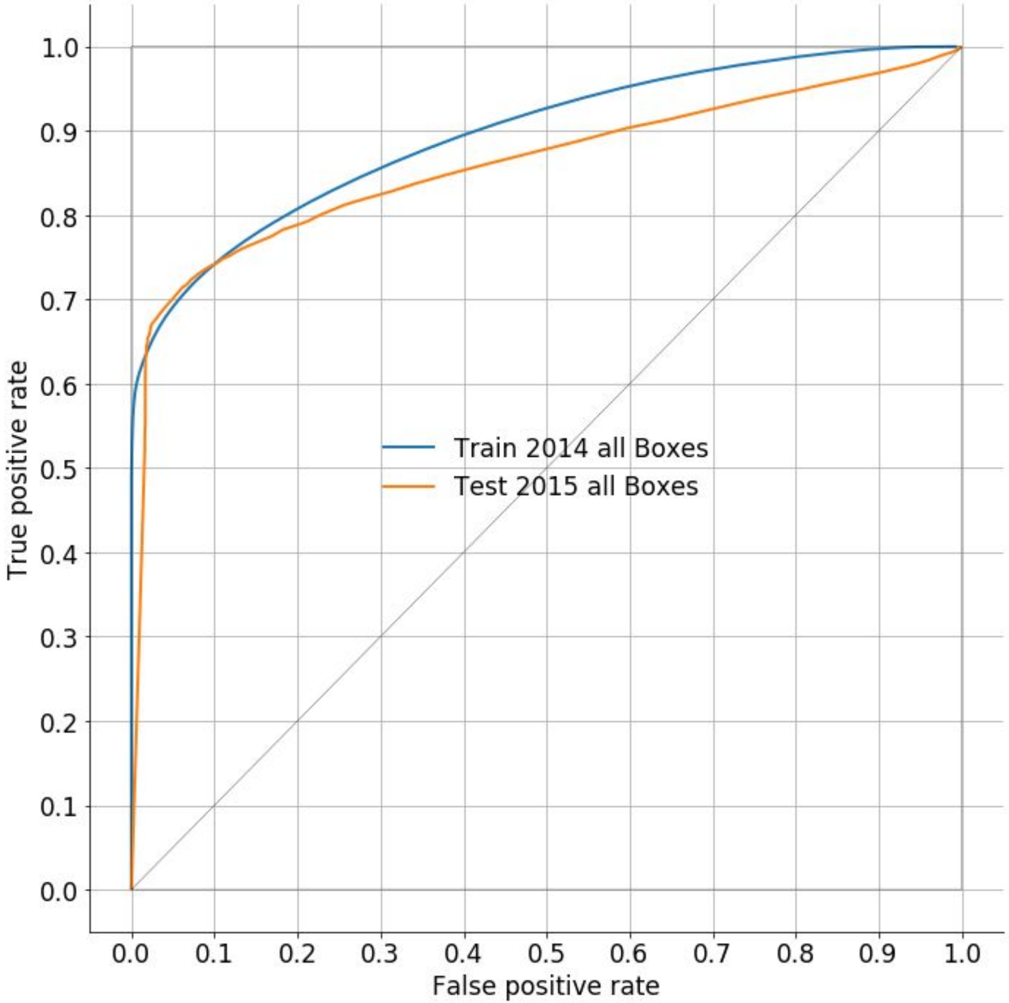


Clustering by k-means (supervised part)



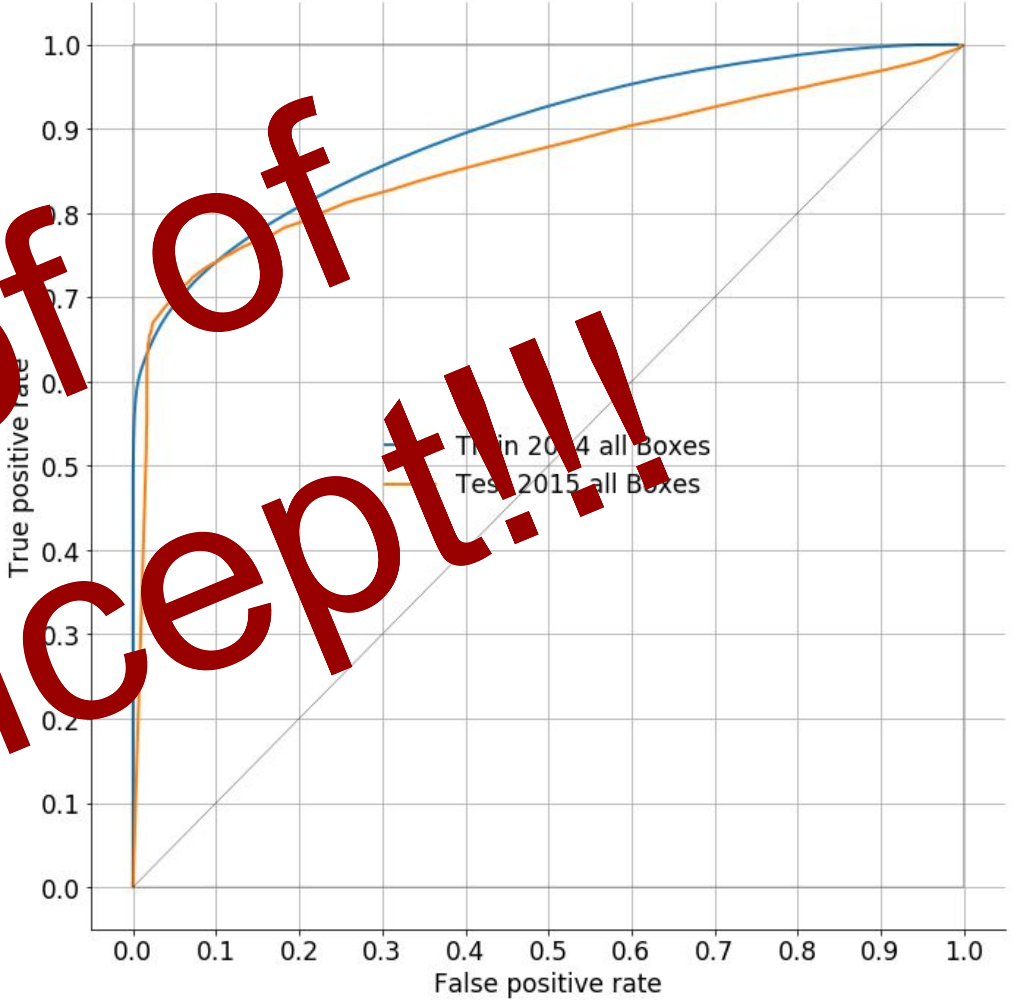


Receiver **O**perating **C**haracteristic



Positive = Outlier

Receiver **O**perating **C**haracteristic



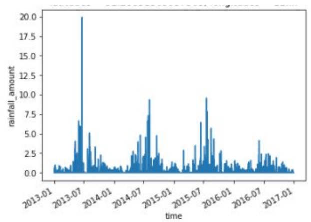
Positive = Outlier

Proof of concept!!!

What could be next?

- Conceptual: use all data vs. use trustworthy data
    - Select trustworthy periods
    - Select trustworthy boxes/sensors

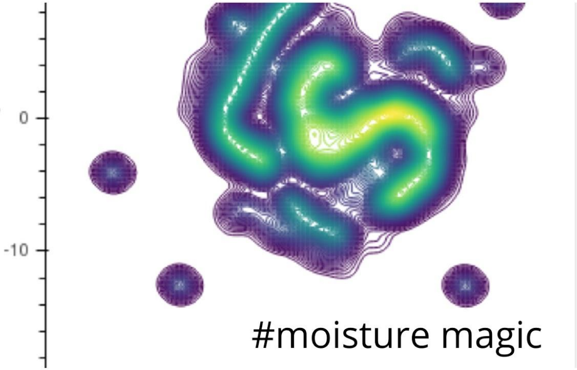
- Use additional information in TS classification
  e.g. rainfall data



- UMAP: Many opportunities to optimize. E.g. neighbouring sensors can be used easily → Should improve performance

Moisture Magic

💦 🧙



#moisture magic