EnWeather Data Science: Using Machine Learning models for Weather Forecast

Envision Digital
DS Team

Content

Part 1

ML quick intro.

Part 2

XGBoost intro.

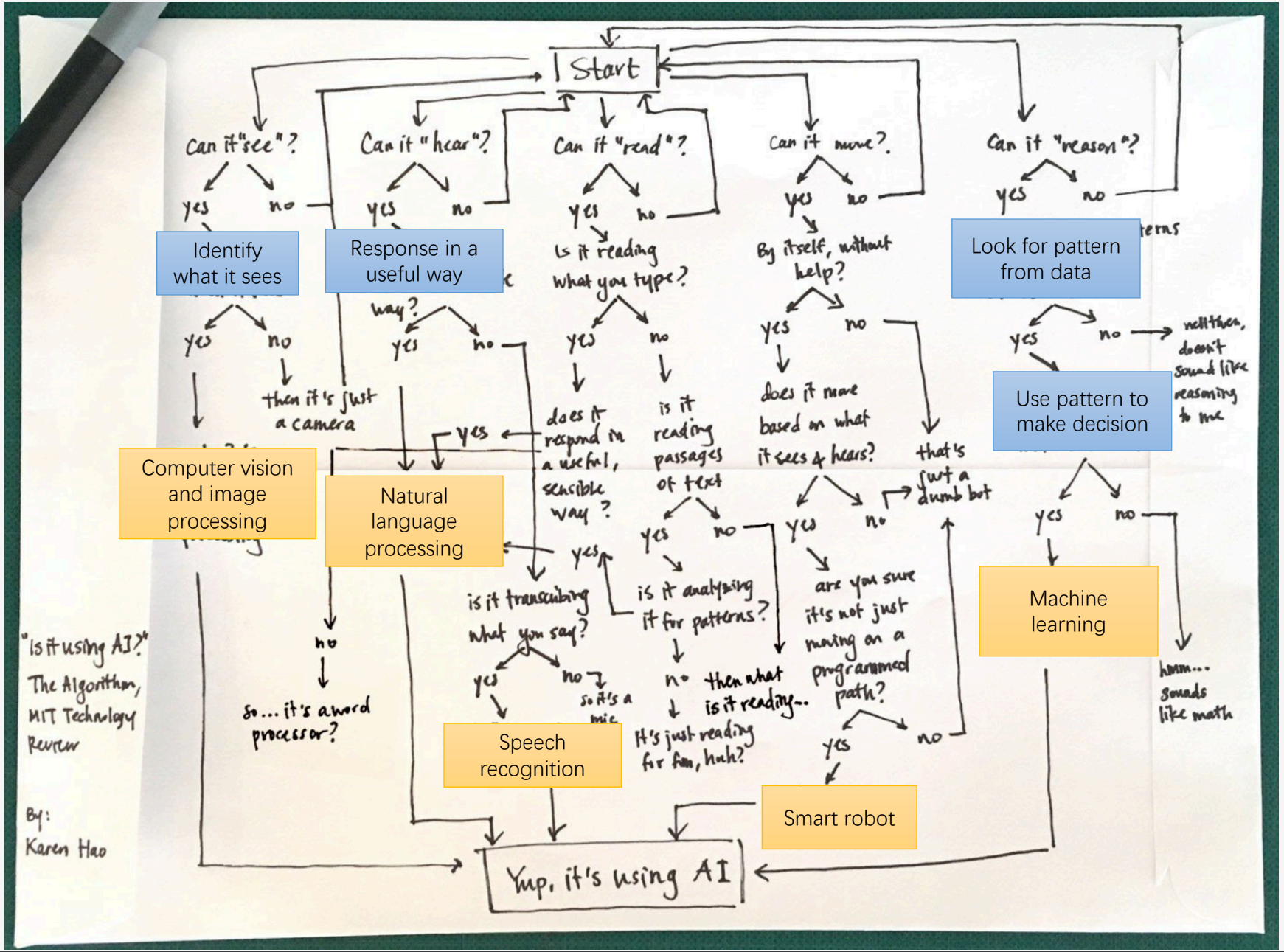
Part 3

XGBoost for wind forecast

Part 4

XGBoost for rainfall forecast

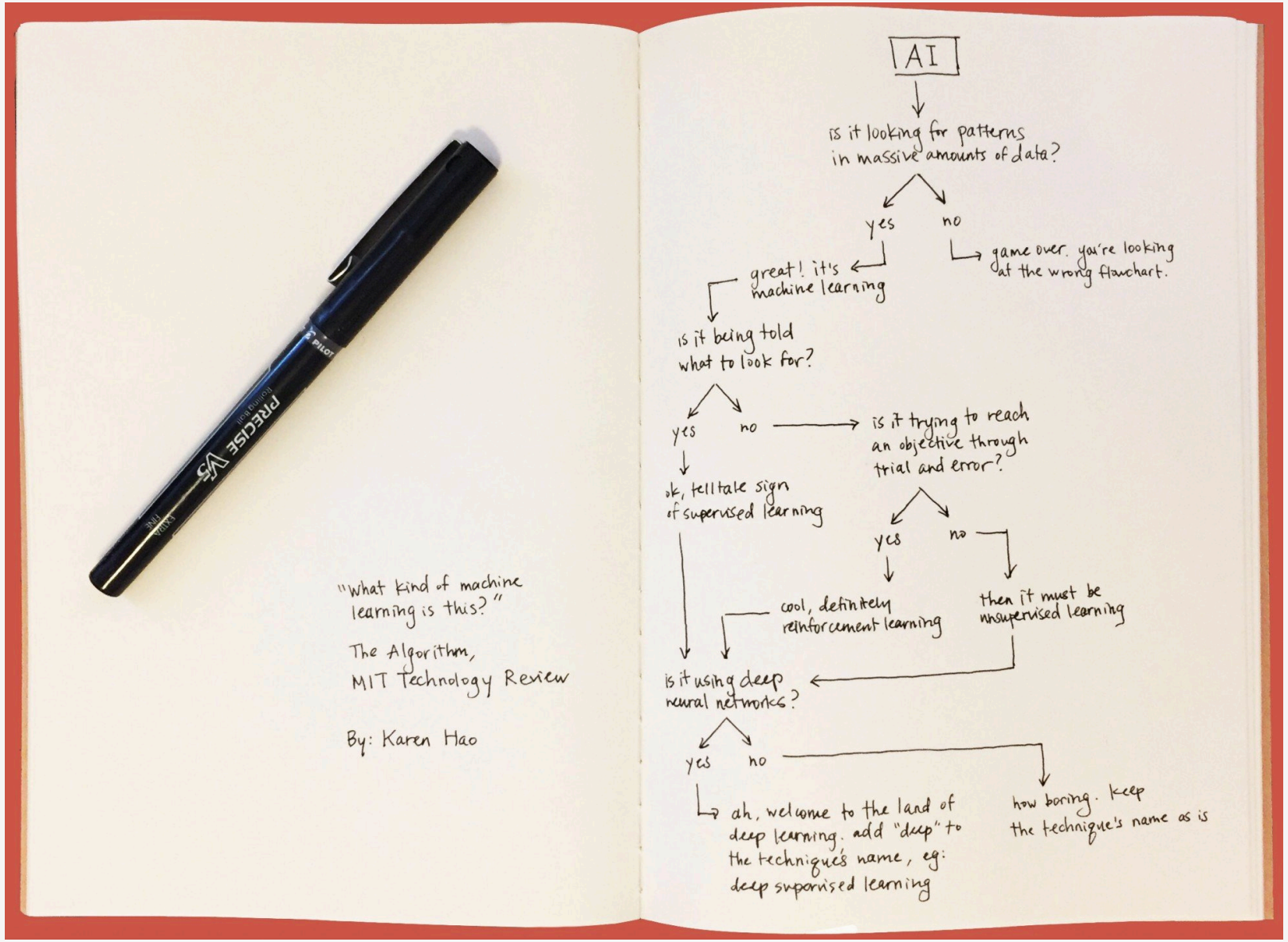
AI flowchart



https://www.technologyreview.com/s/612404/is-this-ai-we-drew-you-a-flowchart-to-work-it-out/

Machine Learning Flowchart

From massive amount of data to find useful patterns!



https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/

Machine Learning Algorithm Classification

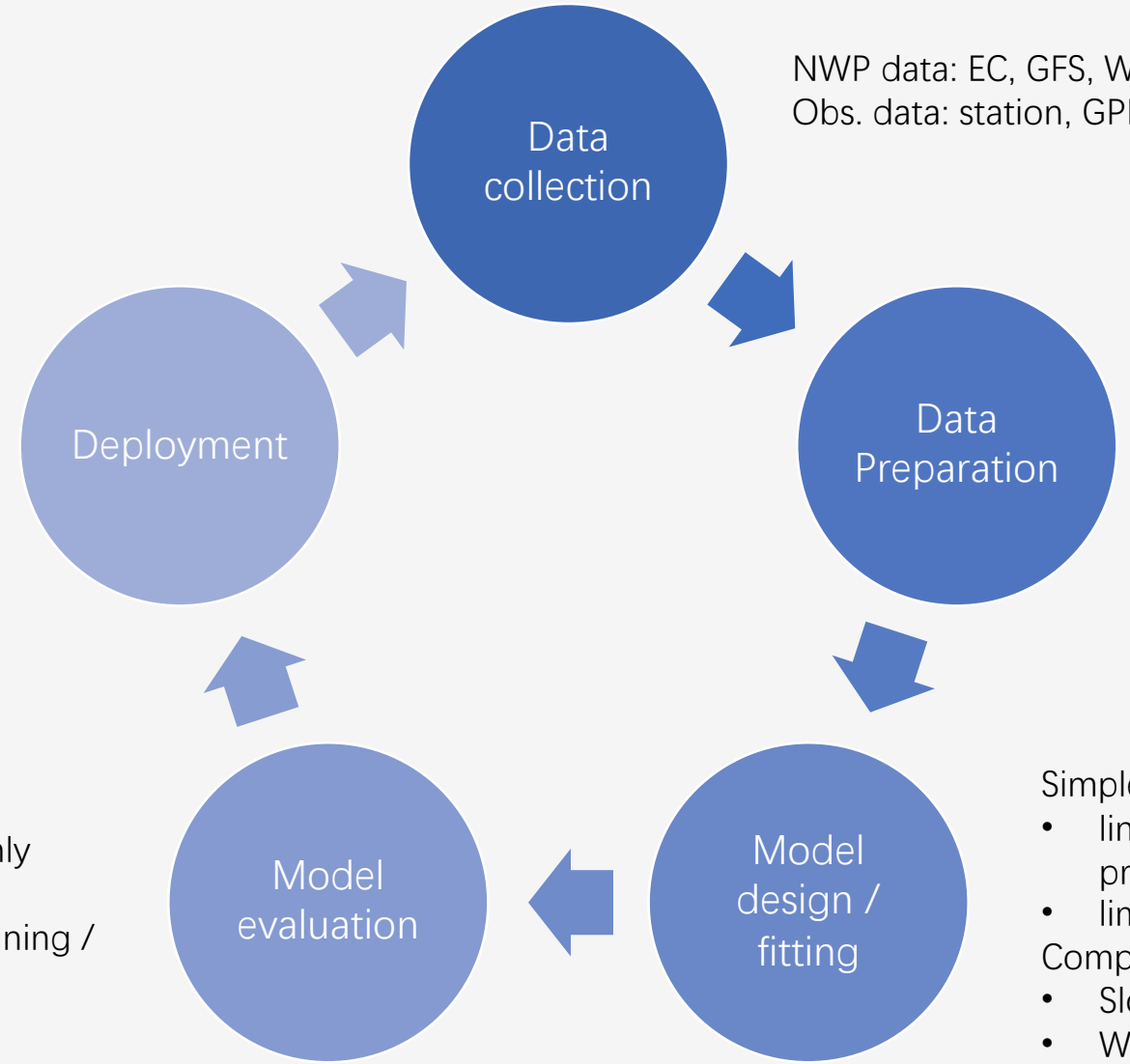
- Unsupervised learning
  - No labelled data is used
  - The machine just looks for whatever patterns it can find
  - Application: cybersecurity https://www.readitquik.com/articles/security-2/why-unsupervised-machine-learning-is-the-future-of-cyber-security/

- Supervised learning
  - Use labeled data or ground truth
  - tell the machine exactly what patterns it should look for
  - Many popular applications
    - Netflix: find similar show for you

- Semi-supervised learning
  - Mix small amount of labelled data with large unlabelled data

- Reinforcement learning
  - learns by trial and error to achieve a clear objective.
  - It tries out lots of different things and is rewarded or penalized depending on whether its behaviors help or hinder it from reaching its objective.
  - AlphaGo

Build Machine Learning Model Key Processes



NWP data: EC, GFS, WRF
Obs. data: station, GPM, radar, etc.

Data cleaning and preprocessing
- Regrid GPM data, handle nan/missing values, data transformation, etc.

Simple model:
- linear model: fast build, fast deployment, fast in production
- limited capability

Complicated model
- Slow build, complicated deployment,
- With high potentials

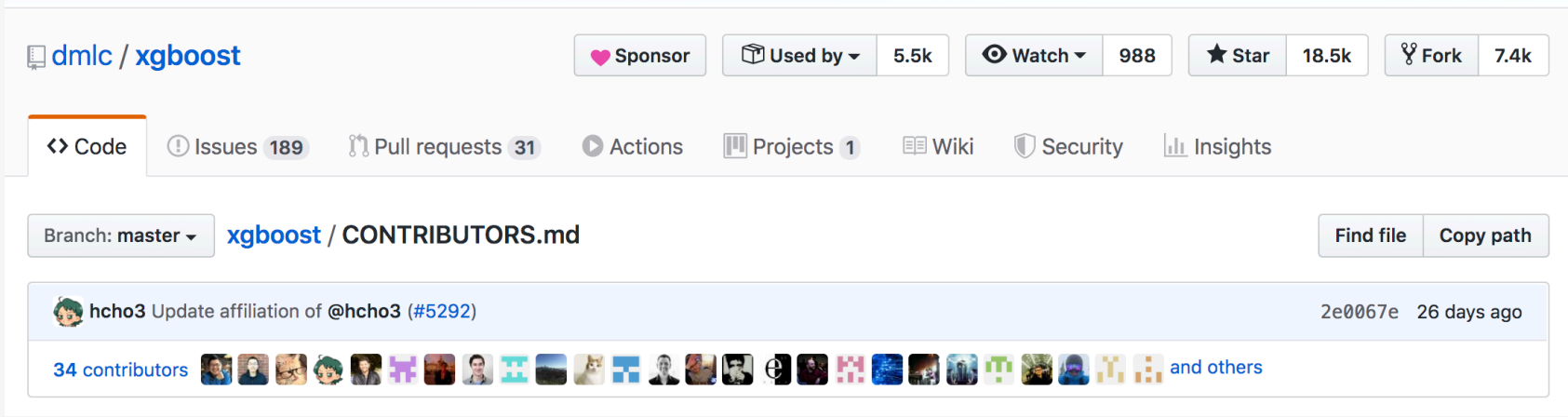
- Daily, weekly, monthly evaluation
- Help to do model tuning / retaining

- AWS, HPC,
- Crontab: with daily forecast,

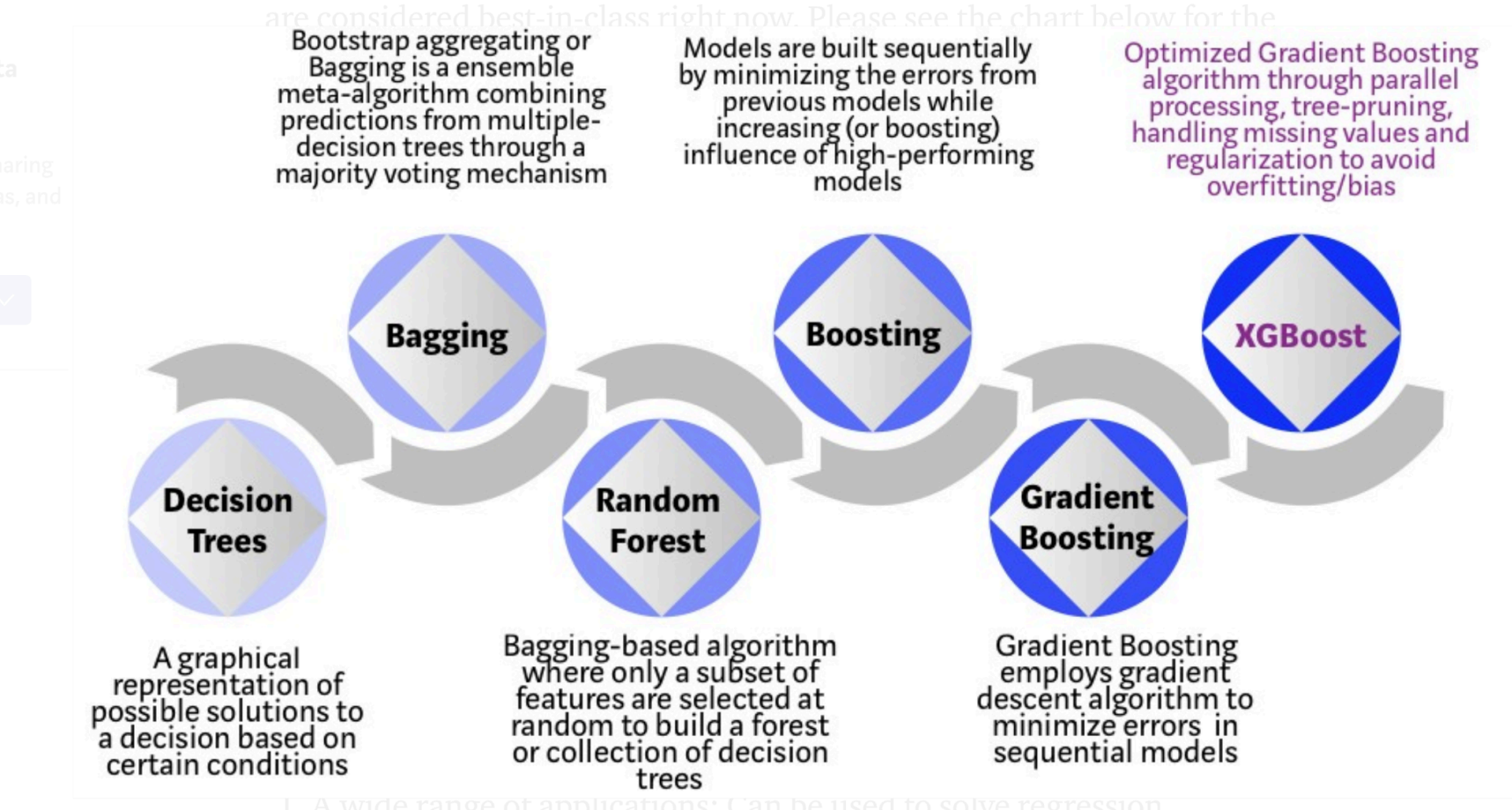
XtremeGradientBoost (XGBoost) Introduction

- XGBoost is a ***decision-tree*** based ***ensemble*** Machine Learning algorithm that uses a ***gradient boosting*** framework
- Developed by a research project at the University of Washington
- First presented by Tianqi Chen and Carlos Guestrin at SIGKDD conference in 2016
- Extremely popular in Kaggle competitions and many industry applications
- XGBoost open source project
  - https://github.com/dmlc/xgboost

"When in doubt, use XGBoost" — Owen Zhang, Winner of Avito Context Ad Click Prediction competition on Kaggle



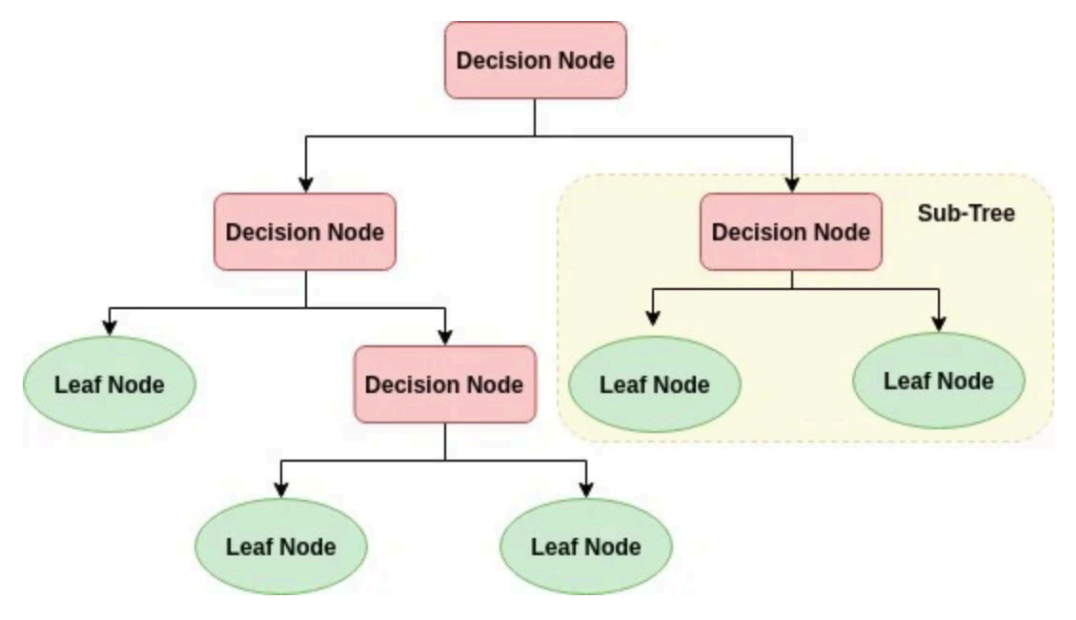
Evolution of XGBoost algorithm



Source: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Evolution of XGBoost algorithm (1)

- Decision Tree
  - A supervised learning algorithm
  - Has tree representation.
  - Each **internal node** of the tree corresponds to an attribute, and each **leaf node** corresponds to a class label.
- Bagging
  - Boostrap aggregation
  - constructs n classification trees using bootstrap sampling of the training data
    - Resample data with replacement
  - combines their predictions to produce a final meta-prediction
- Random forecast
  - A supervised learning algorithm
  - Based on DT and Bagging
  - For each tree, a subset of features are used.



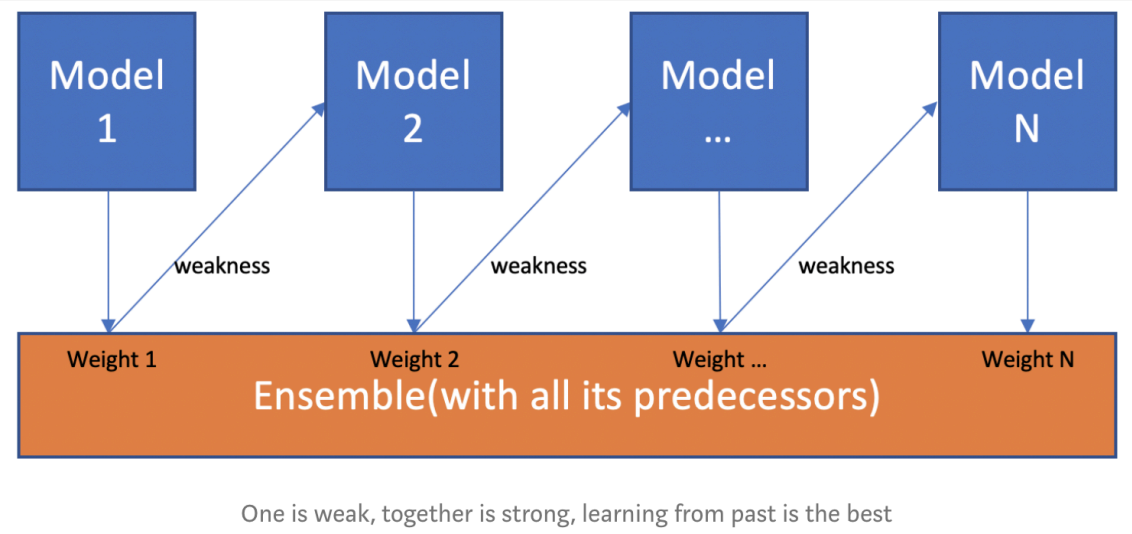
Evolution of XGBoost algorithm

- Boosting
  - A generic algorithm rather than a specific one
  - From a series of weak learners to a strong learner
  - Model are built sequentially by minimizing the error from previous model

- Types of Boosting algorithm

1. AdaBoost
   - For classification problem
   - Identify miss-classified points and increase their weights in the next round
2. Gradient Boosting
   - For both classification and regression problem
   - Define weak by the different between pred. and actual
   - Need differential loss function to calculate error.

- Xgboost
  - DT + Gradient Boosting + a series of optimization



One is weak, together is strong, learning from past is the best

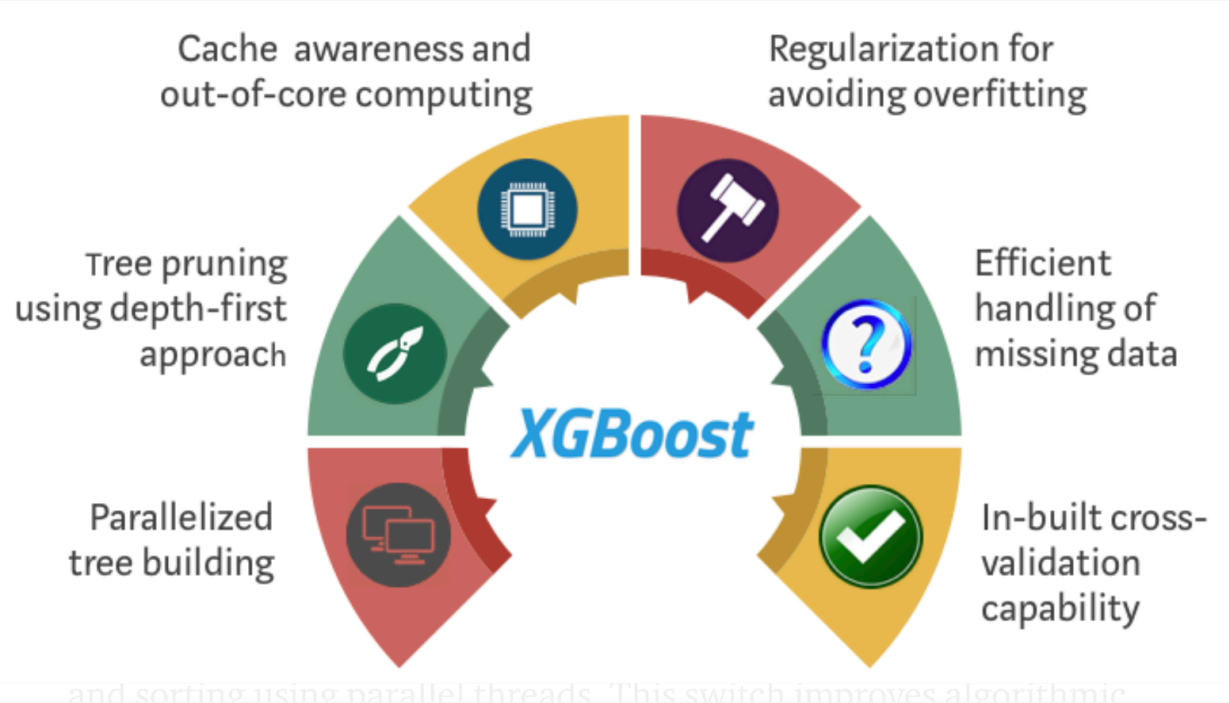
Why XGBoost algorithm is so good

Parallelization
- Utilize parallel processing, use multiple CPUs to execute the model
- A series of tree cannot build in parallization but build each tree can be parallelized

Regularization
- L1: lasso regularization, select features (force weights toward 0)
- L2: ridge regularization, smooth features (force small weights over param.)
- To avoid too complicated tree

Efficient tree pruning
- max depth to control tree grow
- prune the tree backwards and remove splits beyond which there is no positive gain.

……



How to use XGBoost in Weather Forecast

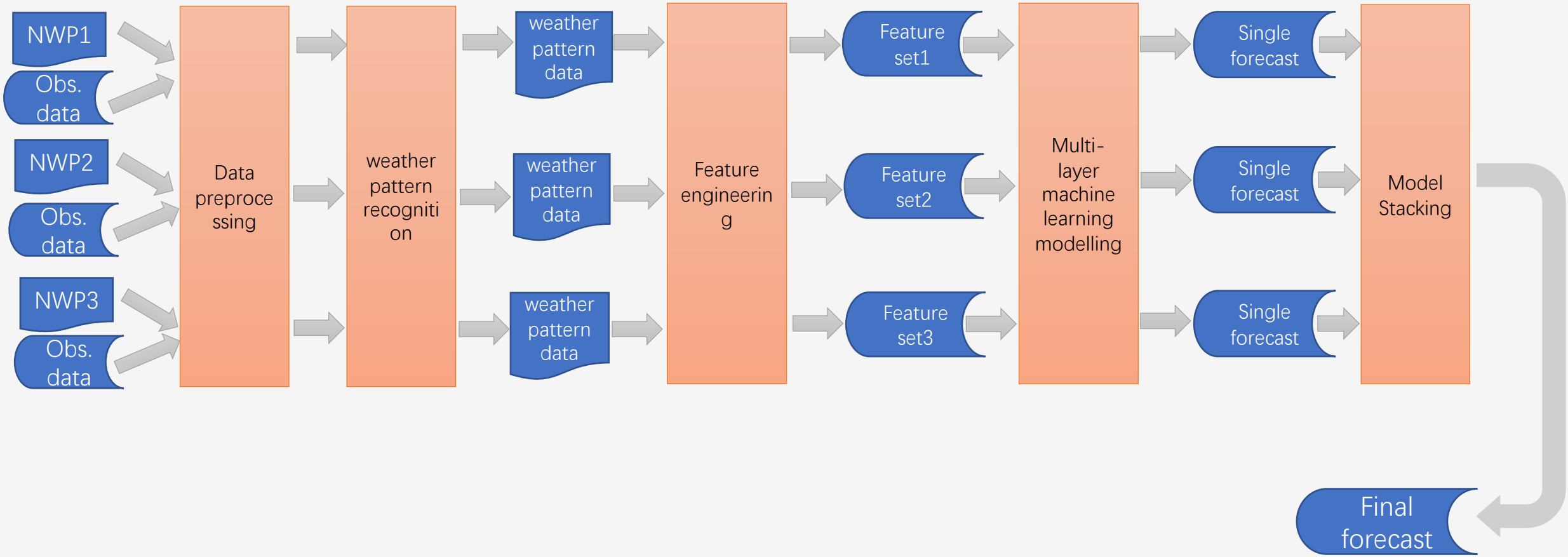
Use XGBoost for wind speed forecast

- Use cases
  - Optimize wind speed forecast for each wind turbine -> help to estimate wind power
- Stacking of single layer machine learning models
  - use multiple nwps
  - reduce the bias for each nwp

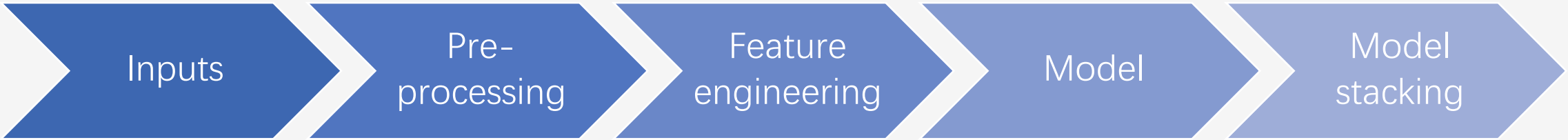
Use XGBoost for rainfall forecast

- Multi-layer single machine learning models
  - First layer rainfall probability
  - Second layer rainfall volume
- Each nwp have very different rainfall pattern
  - EC tends to have more rain
  - IBM tends to have less rain
  - Merge them together may get inconsistent results

Weather Forecast Pipeline using Machine Learning



Use Xgboost for wind speed forecast: entry-level



NWP:
- Nearest grid forecasts: ws, wd, tmp, pres, rho
- Nearby forecast

Obs
- Wind speed from turbines

- Match nwp + obs by time
- Handle invalid ws
- Smooth ws to reduce errors

Basic features
- Nearest/nearby grid attrs.
- Temporal shifted nearby grid attrs.

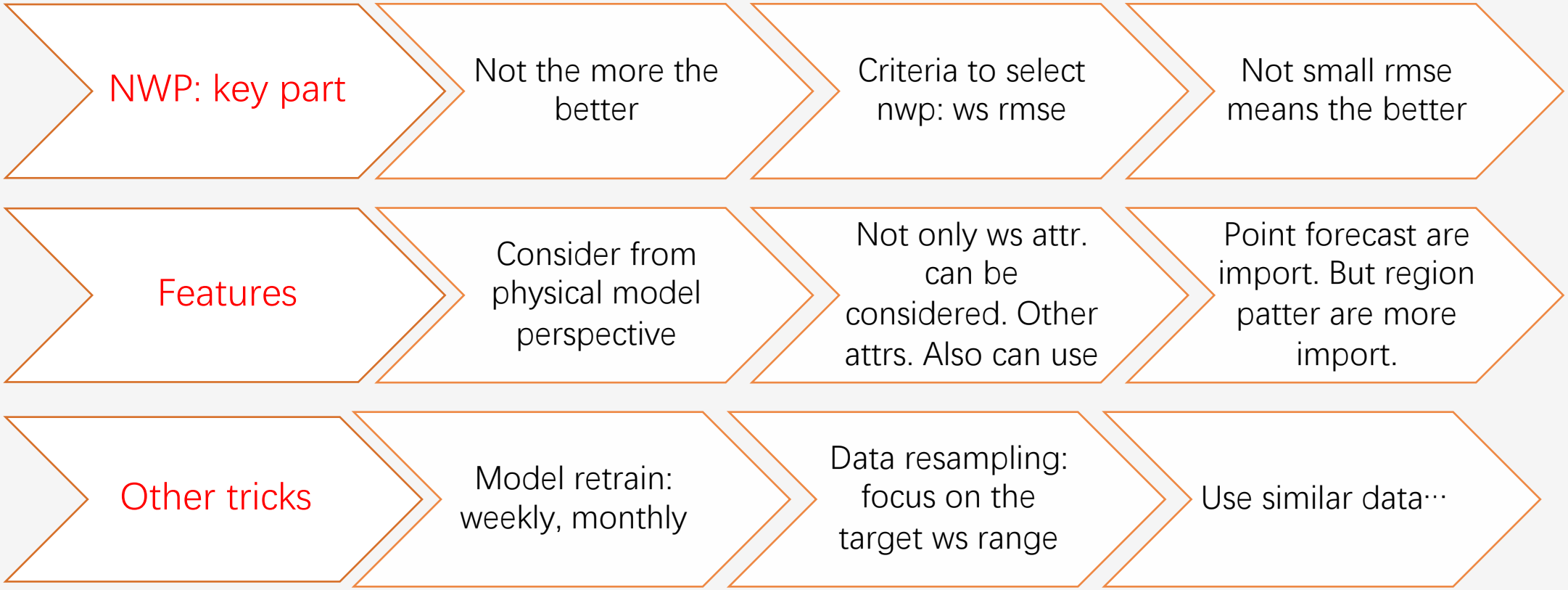
Delta features:
- changes within spatial-temporal aspect

- Define model type: reg. vs clas.
- Define para. Space
- Search for the best param.
- Generate the model

- Use different nwp + obs comb.
- Generate multiple models
- Use a simple model to combine the results: LR

What makes a good wind speed model

Building a wind speed xgboost mode is easy but…



Use XGBoost for rainfall forecast: advanced level

Rainfall forecast challenge
- Location + intensity

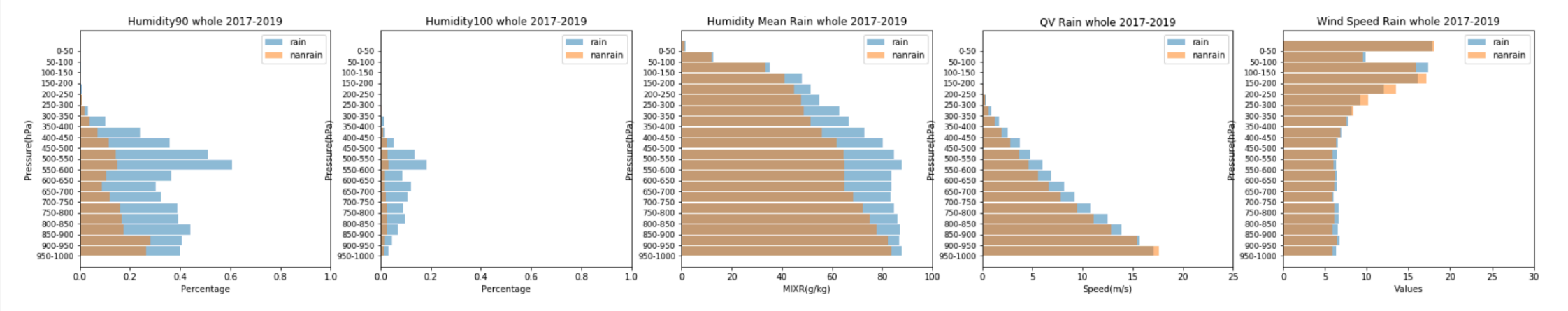
Difference with wind speed forecast
- More weather attributes are used
- Complicated data preprocessing to handle noise and errors
- Advanced feature engineering
  - point level features + region level features
- Two layers of modelling
  - Use XGBoost for both classification problem and regression problem

Use XGBoost for rainfall forecast: inputs

Surface level attributes + pressure level attributes

- Features from different levels are more important
- Possible attributes can be
  - Humidity, QV, QC, wind speed, etc.

Humidity shows a significant difference for rain and non-rain case!



Use XGBoost for rainfall forecast: preprocessing

Data cleaning/noise filtering

- Invalid rainfall: single and very small region rainfall
- Unsupervised learning to filter noise
  - DBSCAN (density based clustering algorithm)

Reduce spatial-temporal errors

- For both NWP data and GPM data
- Time domain smoothing: fix window smooth, (+-1hour)
- Spatial domain smoothing: redistribute rainfall in nearby area. Gaussian filter

Aggregation

- Use a few hour aggregation
- Spatial aggregation: max pooling, etc.

Use XGBoost for rainfall forecast: features

Observation based features

- Past rainfall patterns
- Frequent rainfall periods
- Shifted rainfall volumes
- Etc.

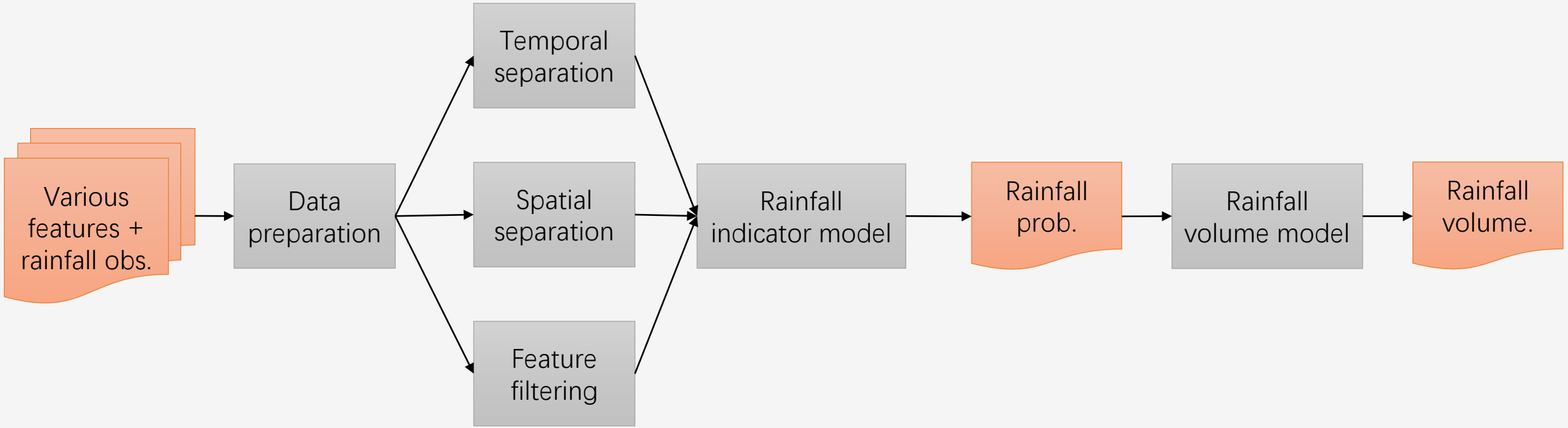
Point based features

- From point to point calc. features
- Various fcst. Attr.
- Attr. Statistics
- Attr. Shift
- Etc.

Region based features

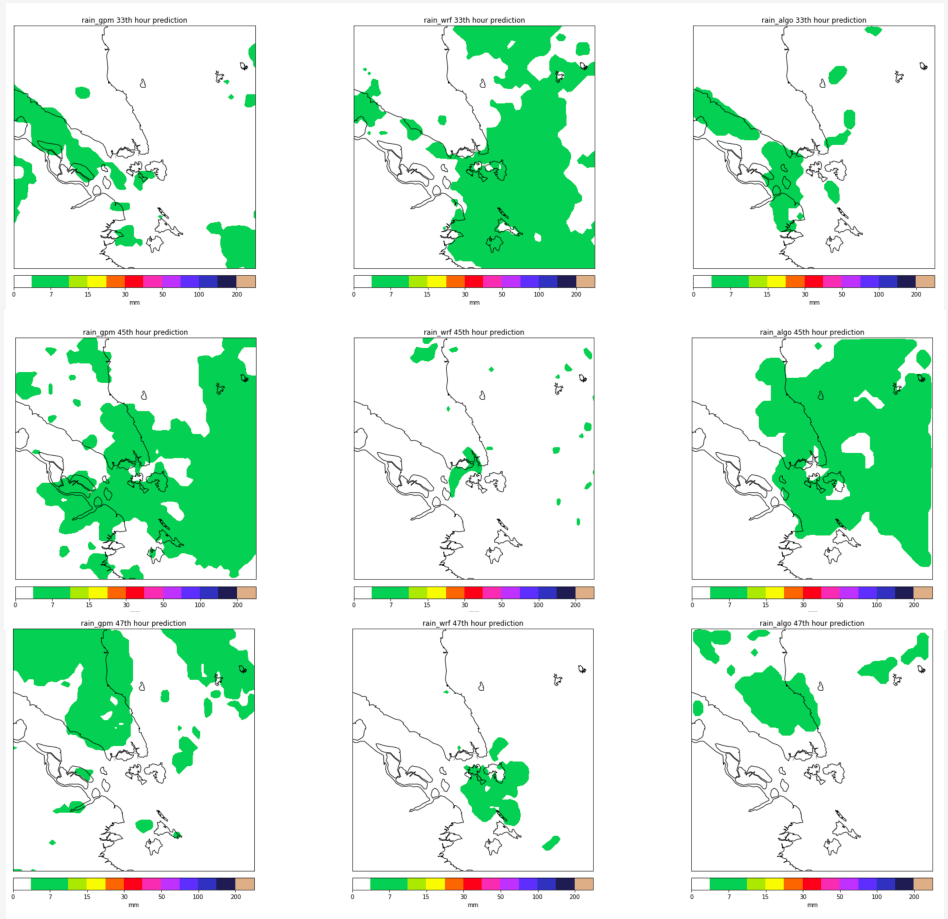
- Rainfall is a regional attr.
- Extract rainfall in terms of regions
- Calc. region stats
  - Min/max/size
- Rainfall region vs attr region
- Etc.

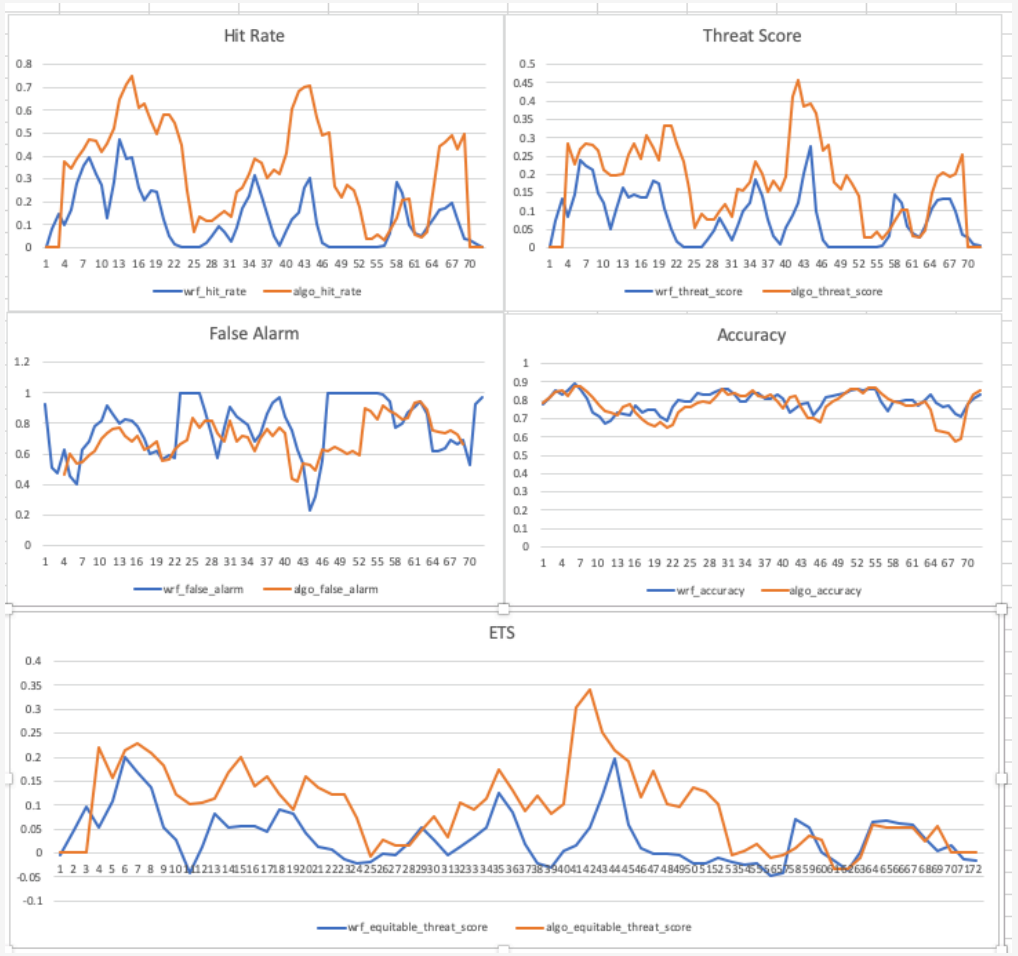
Use XGBoost for rainfall forecast: modelling



Use XGBoost for rainfall forecast: performance

GPM | WRF | ML





Thank You