ACM SIGKDD CUP 2018

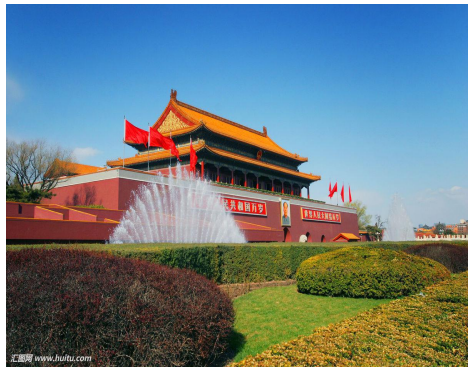
北京交通大学 信息科学研究所

参赛学生： 李有儒、李福臻、孔德强、孟莹莹、李超

指导老师： 朱振峰

背景

- 今年的KDD关注点事空气质量问题。
- 在过去几年中，空气质量问题已经影响了很多发展中国家的大城市。
- 在众多空气污染物中，悬浮颗粒（particulate matters，简称PM）是最致命的一种之一。直径小于或等于2.5μm的悬浮颗粒可以进入肺部深处，进入血管，导致DNA突变和癌症，中枢神经系统损伤，和过早死亡（premature death）。
- 所以，**精确地监测和预测PM2.5和其他颗粒及污染物**变得非常重要。如果可以准确预测污染事件，市民和政府可以随之作出适当的决策，例如关闭学校或减少室外运动，从而减少污染带来的损害。



任务描述

- 2018/4/1 - 2018/4/30： 第一阶段：练习赛。选手可以通过API获取数据，提交答案。本阶段为练习赛，结果不计入最后成绩。此阶段的主要作用为探索模型框架，开发和稳定数据获取、模型训练、预测和用程序提交结果。
- 2018/5/1 0:00 a.m. (UTC) - 2018/5/31 23:59 p.m. (UTC)： 第二阶段：决赛。在决赛的31天中，选手每天都需提交对未来空气质量的预测，本阶段的成绩将决定KDD Cup 2018最终排名。

评测指标

- 每天，提交的结果将会和真实空气质量数据（也就是空气监测站测量的污染物浓度）比较，并根据Symmetric mean absolute percentage error评分：

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(A_t + F_t)/2}$$

总体解决方案


城市污染
BEIJING
LONDON
多源数据
精准预测
160
140
120
100
80
60
40
20
0
10
20
30
40
Good
Moderate
Unhealthy Sensitive Groups
Unhealthy
Very Unhealthy
Hazardous

数据

历史空气质量数据

	PM2.5	PM10	NO2	CO	O3	SO2
count	44019.000000	36849.000000	44562.000000	44645.000000	44372.000000	44544.000000
mean	60.588996	130.234715	42.182196	0.714488	83.344744	6.015715
std	52.075653	99.546856	28.690830	0.499248	57.737356	7.137139
min	3.000000	6.000000	2.000000	0.100000	2.000000	2.000000
25%	24.000000	65.000000	20.000000	0.400000	40.000000	2.000000
50%	45.000000	112.000000	35.000000	0.600000	77.000000	3.000000
75%	84.000000	170.000000	58.000000	0.900000	113.000000	8.000000
max	705.000000	2030.000000	249.000000	4.800000	342.000000	300.000000

BEIJING

	PM2.5 (ug/m3)	PM10 (ug/m3)	NO2 (ug/m3)
count	14541.000000	15913.000000	14530.000000
mean	16.681968	24.564042	40.258238
std	12.391101	14.250281	26.660233
min	-9.400000	-1.800000	-5.000000
25%	7.800000	14.100000	20.100000
50%	13.000000	21.600000	34.600000
75%	22.000000	31.800000	54.900000
max	100.700000	252.000000	223.900000

LONDON

历史观测气象与网格气象数据

stationName	longitude	latitude	utc_time	temperature	pressure	humidity	wind_direction	wind_speed/kph
beijing_grid_646	118.0	40.6	2018-05-30 23:00:00	15.0	953.0137	53.0	335.8	5.81
beijing_grid_647	118.0	40.7	2018-05-30 23:00:00	15.0	957.6644	53.0	328.97	5.43
beijing_grid_648	118.0	40.8	2018-05-30 23:00:00	15.0	959.0803	53.0	326.91	5.54
beijing_grid_649	118.0	40.9	2018-05-30 23:00:00	16.0	957.2614	52.0	329.89	6.07
beijing_grid_650	118.0	41.0	2018-05-30 23:00:00	16.0	955.4424	52.0	332.37	6.62

BEIJING

	stationName	longitude	latitude	utc_time	temperature	pressure	humidity	wind_direction	wind_speed/kph
1210460	london_grid_856	2.0	52.1	2018-05-30 19:00:00	16.59	1014.1452	86.0	117.51	4.53
1210461	london_grid_857	2.0	52.2	2018-05-30 19:00:00	16.46	1014.1493	86.0	127.88	5.22
1210462	london_grid_858	2.0	52.3	2018-05-30 19:00:00	16.37	1014.1598	86.0	134.17	6.05
1210463	london_grid_859	2.0	52.4	2018-05-30 19:00:00	16.31	1014.1767	86.0	137.62	6.93
1210464	london_grid_860	2.0	52.5	2018-05-30 19:00:00	16.24	1014.1936	86.0	140.28	7.83

LONDON

➢ 数据

- 气象预报网格数据

stationName	longitude	latitude	utc_time	temperature	pressure	humidity	wind_direction	wind_speed/kph
beijing_grid_646	118.0	40.6	2018-05-30 23:00:00	15.0	953.0137	53.0	335.8	5.81
beijing_grid_647	118.0	40.7	2018-05-30 23:00:00	15.0	957.6644	53.0	328.97	5.43
beijing_grid_648	118.0	40.8	2018-05-30 23:00:00	15.0	959.0803	53.0	326.91	5.54
beijing_grid_649	118.0	40.9	2018-05-30 23:00:00	16.0	957.2614	52.0	329.89	6.07
beijing_grid_650	118.0	41.0	2018-05-30 23:00:00	16.0	955.4424	52.0	332.37	6.62

BEIJING

	stationName	longitude	latitude	utc_time	temperature	pressure	humidity	wind_direction	wind_speed/kph
1210460	london_grid_856	2.0	52.1	2018-05-30 19:00:00	16.59	1014.1452	86.0	117.51	4.53
1210461	london_grid_857	2.0	52.2	2018-05-30 19:00:00	16.46	1014.1493	86.0	127.88	5.22
1210462	london_grid_858	2.0	52.3	2018-05-30 19:00:00	16.37	1014.1598	86.0	134.17	6.05
1210463	london_grid_859	2.0	52.4	2018-05-30 19:00:00	16.31	1014.1767	86.0	137.62	6.93
1210464	london_grid_860	2.0	52.5	2018-05-30 19:00:00	16.24	1014.1936	86.0	140.28	7.83

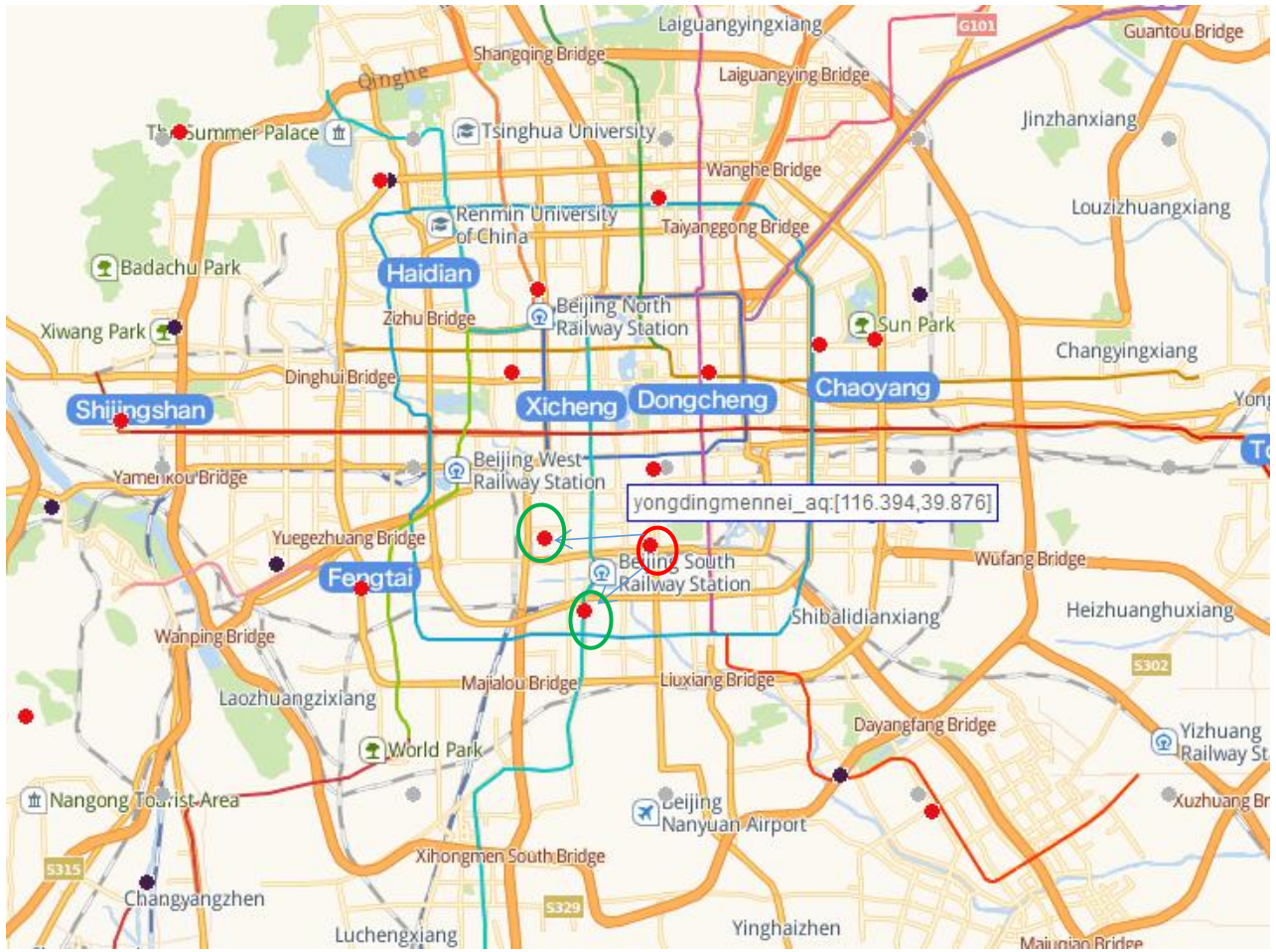
LONDON

➢ 挑战

- 数据大面积缺失，多任务、小样本
- 空气质量的人为影响，长期预测，模型性能要求

➢ 预处理

- 缺失值填补——利用时空依赖性进行迭代补全

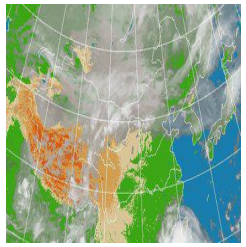


t

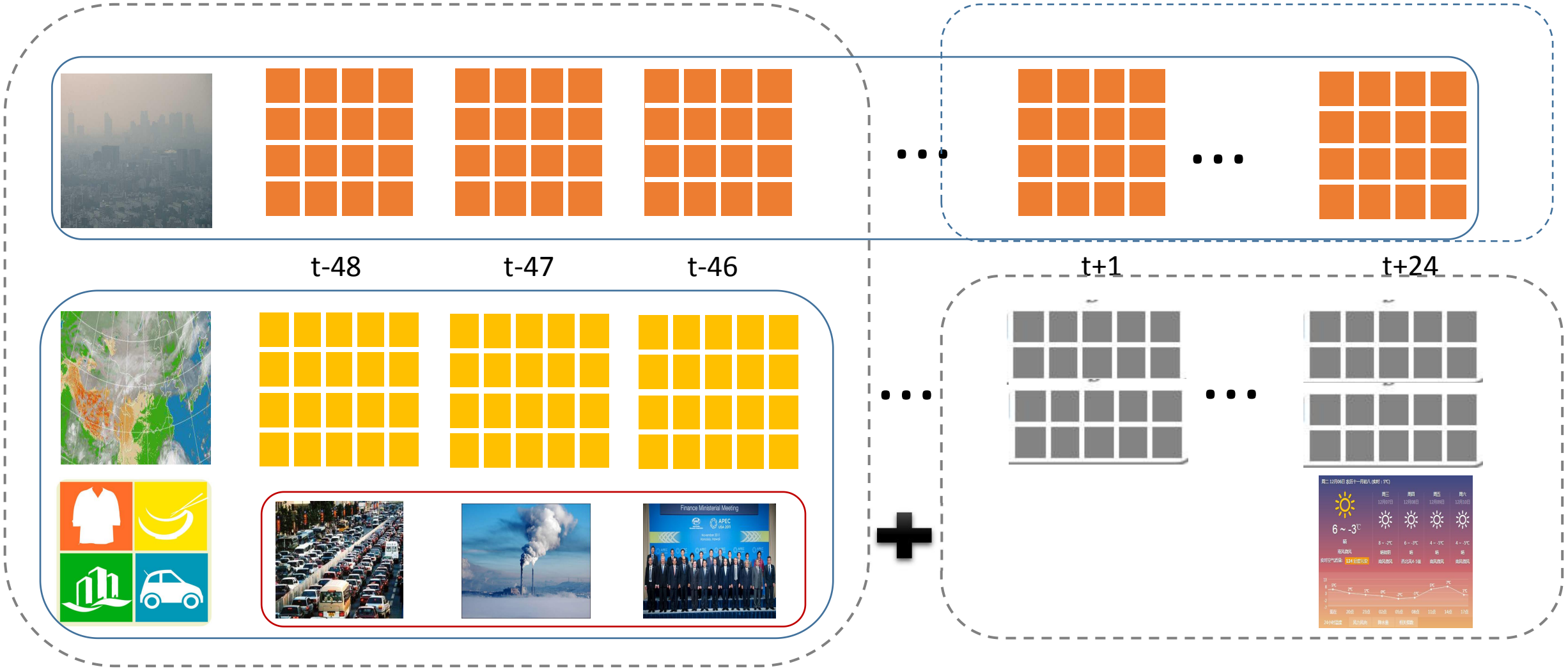
➢ 特征表示

- 训练样本构成要素

历史空气质量	历史气象信息	气象信息预报
☐ PM2.5,PM10 ☐ NO2、O3 ☐ CO，SO2	☐ 温度,气压 ☐ 风速、风向 ☐ 空气湿度	☐ 温度,气压 ☐ 风速、风向 ☐ 空气湿度



➢ 特征表示



➢ 特征表示

- 历史空气质量统计特征：
  - 最大值
  - 最小值
  - 极差
  - 平均值
  - 方差
  - 标准差
  - 中位值
  - AQI变换值
- 时间特征：
  - 是否交通时段
  - 是否生产高峰时段
  - 是否节假日
  - 是否重大活动
- 历史气象特征
- 气象预报特征
- 量化特征与变换特征

➢ 预测模型与集成

- 模型层次集成
  GBRT,XGBoost,RF
  不同参数模型间、不同模型间
- 样本层次集成
  2017/1/1 —— 2018/05
  2018/1/1 —— 2018/05
- 特征层次集成
- 平均、加权平均

➢ 最终成绩

- **Specialized Prize for the second-day prediction: TOP 20 (20/4188)**
- **Specialized Prize for the last 10 days: TOP 1% (41/4188)**
- **Ranking for 30 days prediction: TOP 1% (38/4188)**

Thank You

Q&A