The Simpler The Better: A Unified Approach to Predicting Original Taxi Demands based on Large-Scale Online Platforms

Yongxin Tong[1], Yuqiang Chen[2], Zimu Zhou[3], Lei Chen[4], Jie Wang[5], Qiang Yang[2,4], Jieping Ye[5], Weifeng Lv[1]

[1] SKLSDE Lab, Beihang University, [2] 4Paradigm Inc., [3] ETH Zurich,

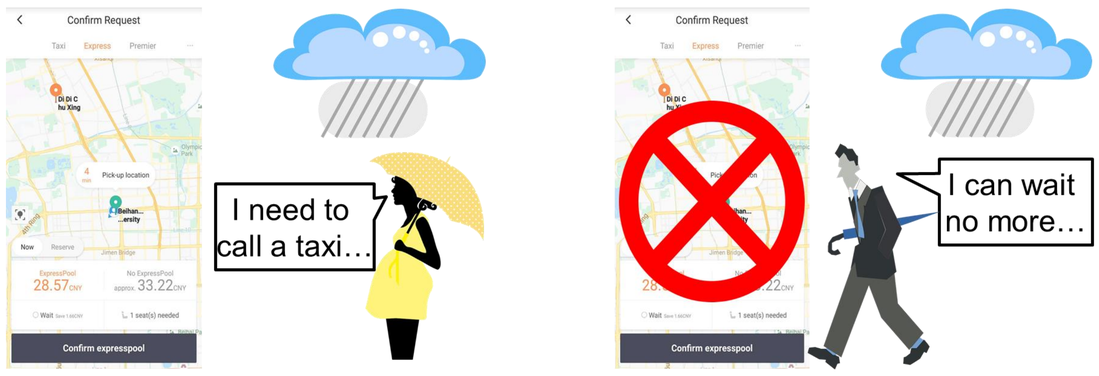
[4] Hong Kong University of Science and Technology, [5] Didi Research

[1] {yxtong, lwf}@buaa.edu.cn, [2] chenyuqiang@4paradigm.com, [3] zimu.zhou@tik.ee.ethz.ch

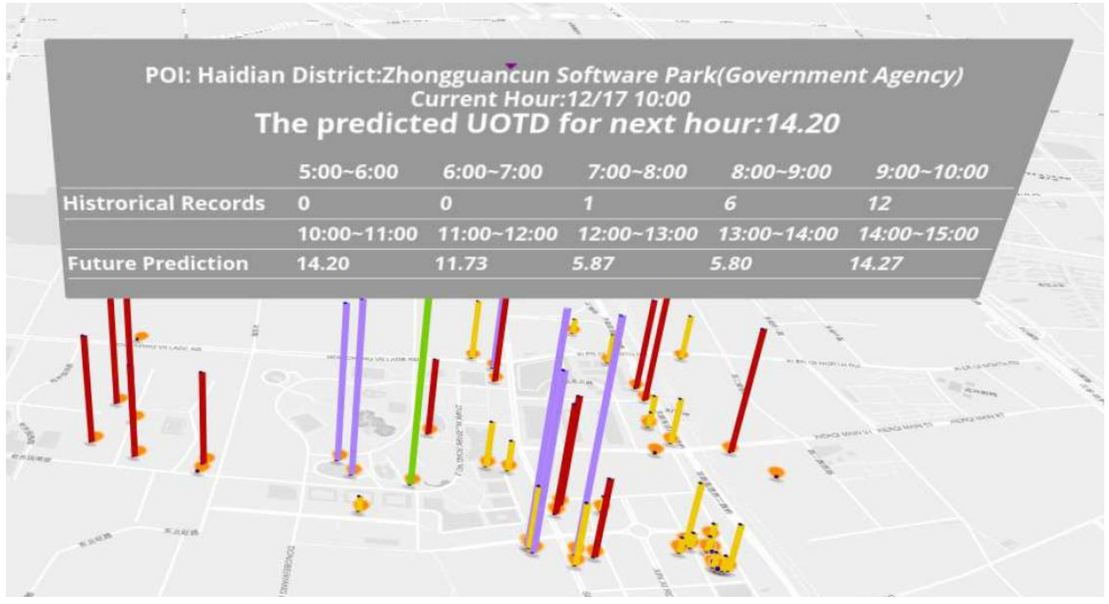
[4] {leichen,qyang}@cse.ust.hk, [5] {wangjiejacob,yejieping}@didichuxing.com

Target: UOTD Prediction

- OTD (Original Taxi Demand)
  - The number of taxi-calling orders *submitted* to the online taxicab platform
  - Including orders that are cancelled finally



- UOTD (Unit Original Taxi Demand)
  - The number of taxi-calling orders *submitted* to the online taxicab platform *per unit time and per unit region*

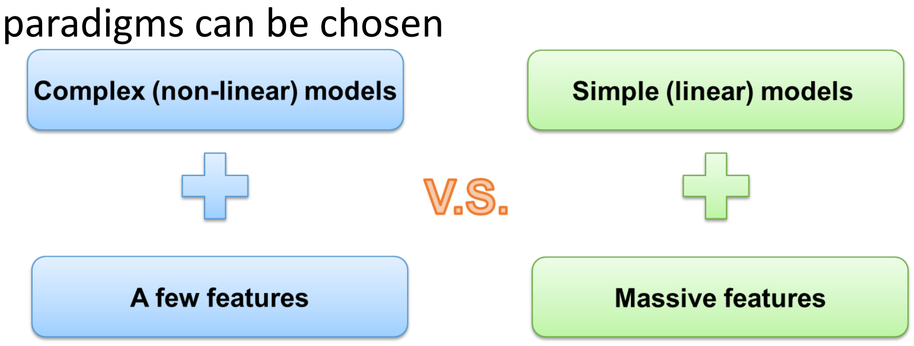


- Why we need to predict UOTD
  - Expand Potential Market
  - Assess Incentive Mechanisms
  - Guide Taxi Dispatching



Key Methodology

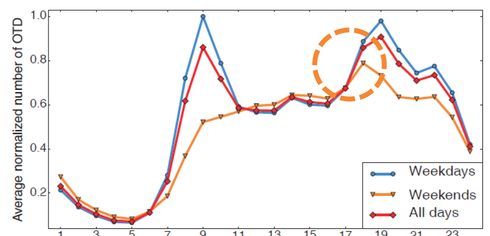
- Two paradigms can be chosen



- In industrial practice, the latter one (Simple models + Massive features) is preferred
  - As it can Transform *Model Redesign* to *Feature Redesign*

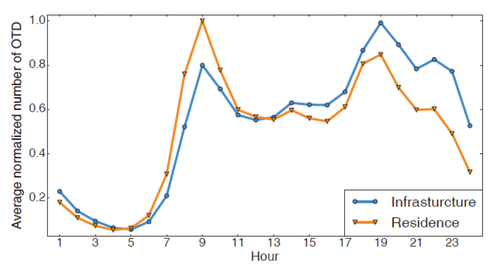
Feature Engineering

- Basic Features
  - Temporal Features
  - Spatial Features
  - Meteorological Features
  - Event Features
- Combinational Features
  - Combine basic features based on business logics and data analysis
  - Express the joint influences of different basic features in a simple model



Weekdays: Two peaks
Weekends: One peak

UOTD is influenced by *Day of week* and *Hour of day jointly*



Infrastructures: More at evening peak
Residences: More at morning peak

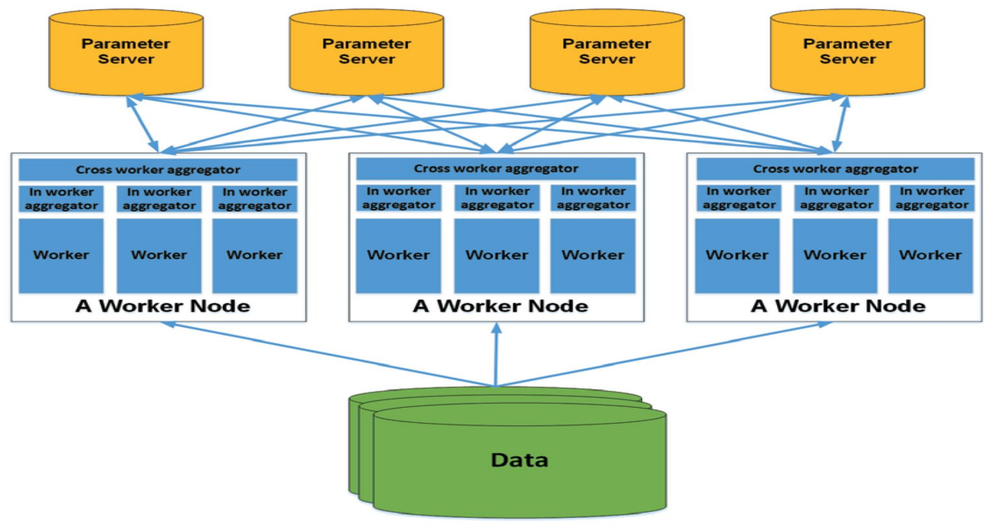
UOTD is influenced by *Typo of POIs* and *Hour of day jointly*

Model & Training

- Model
  - A linear regression model with high-dimensional features and a spatiotemporal regularizer

$$\text{obj}_{spatio\text{-}temporal}(\boldsymbol{w}) = \sum_{X \subseteq D} \phi(X) var(\{\boldsymbol{w}'\boldsymbol{x} | \boldsymbol{x} \in X\})$$

- Distributed Training



Experiments

Dataset	Method	ER	SMAPE	RMLSE
Beijing	HA	0.96957864	0.44033822	0.52884659
	ARIMA	0.89574376	0.42708392	0.50064628
	Markov	0.81039261	0.37087309	0.65547612
	GBRT	0.73525391	0.43042413	0.42926168
	NN	0.81226708	0.43515638	0.43978603
	HP-MSI	0.72515736	0.38083785	0.44228373
	LinUOTD	**0.6466814**	**0.35701066**	**0.40665828**
Hangzhou	HA	0.70616373	0.45098107	0.55787302
	ARIMA	3.16414193	0.46414572	0.59576175
	Markov	0.83794771	0.44441837	0.83023651
	GBRT	0.52536404	0.54445512	0.50110505
	NN	0.61526469	0.56586680	0.50200963
	HP-MSI	0.63366671	**0.43352982**	0.51046835
	LinUOTD	**0.54730029**	0.44870624	**0.49750043**