Anomaly Detection in Infrequently Occurred Patterns

Dong Wang

Principal Architect, wangdong13@baiduc.com

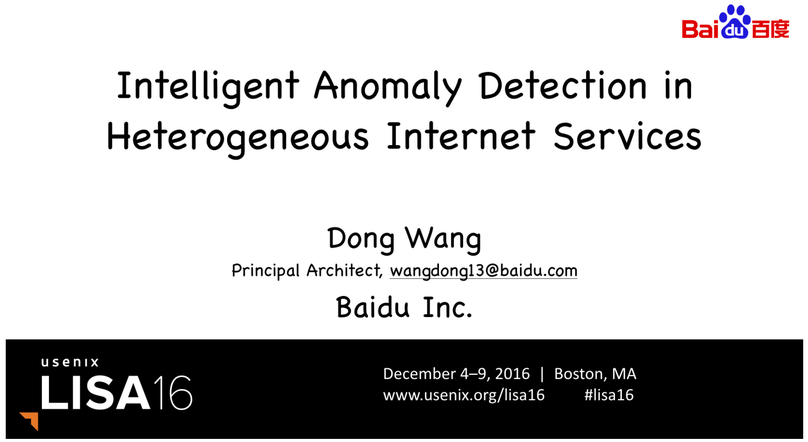
Baidu Inc.

SRECON AMERICAS

San Francisco
March 13-14, 2017

Background

- Detailed explanation to a case mentioned in the talk
  - ✓ Dong Wang, "Intelligent Anomaly Detection in Heterogeneous Internet Services", LISA, Boston, U.S.A. Dec.2016 (https://2459d6dc103cb5933875-c0245c5c937c5dedcca3f1764ecc9b2f.ssl.cf2.rackcdn.com/lisa16/wang.mp4)



Agenda

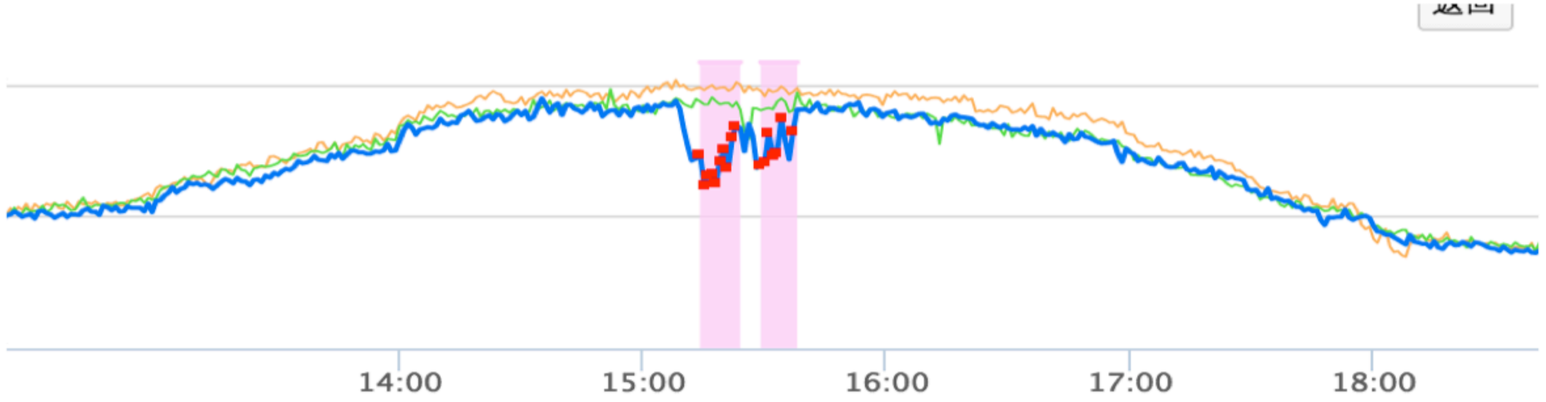
- Introduction to Baidu
- What the problem is
- The idea and solution
- Results

Introduction to Baidu

- One of the largest search engines in the world
  - ✓ Web/Image/Video/News/…
- Besides search, we also have
  - ✓ Location Based Service – Maps
  - ✓ Social/Knowledge – Tieba/Zhidao
  - ✓ Online to Offline – Nuomi/Waimai
  - ✓ Finance/Payment – Wallet
  - ✓ Cloud computing – Cloud
- Covers more than 1 billion users in total

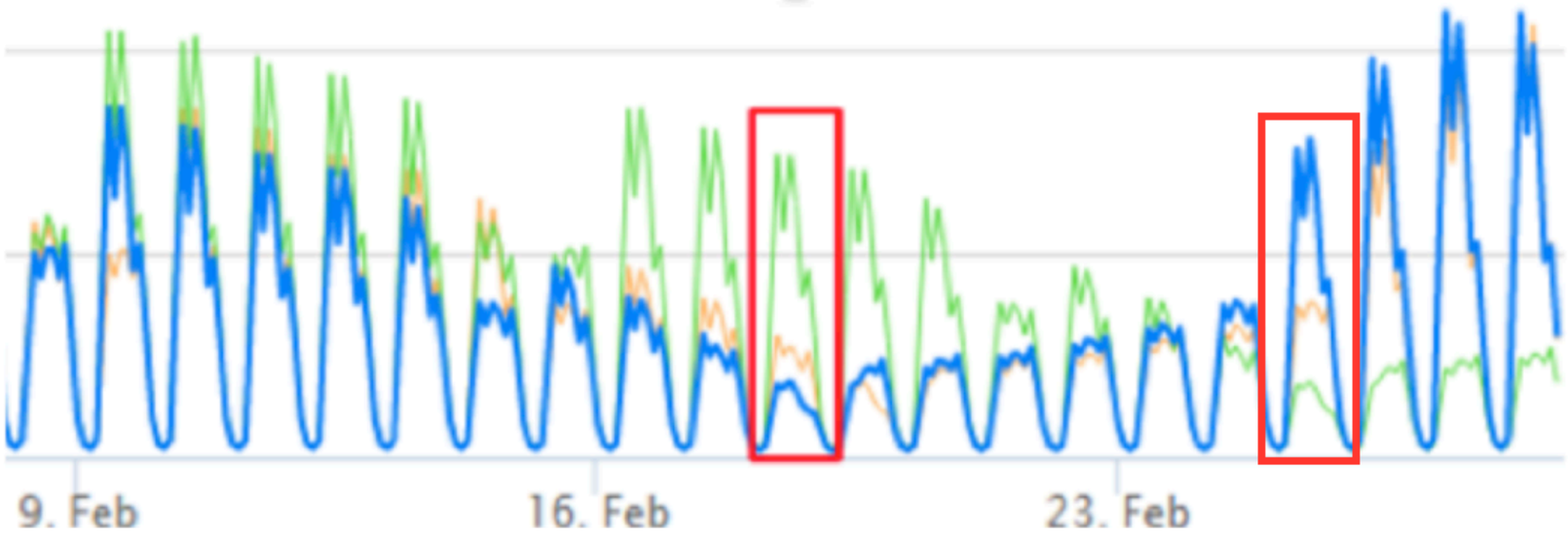
The problem - Anomaly Detection

- In theory it should not be a difficult problem ...



Example problem in Reality

- A metric's curve around Spring Festival

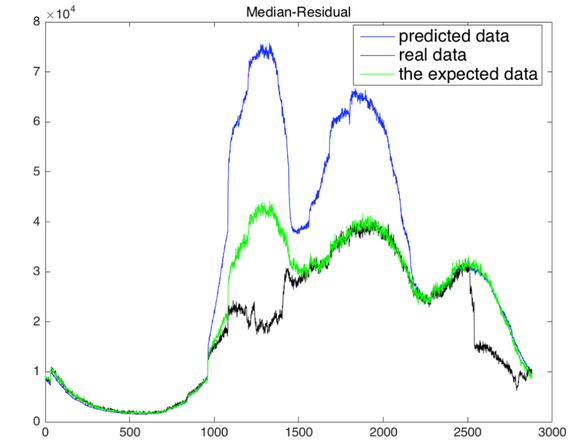


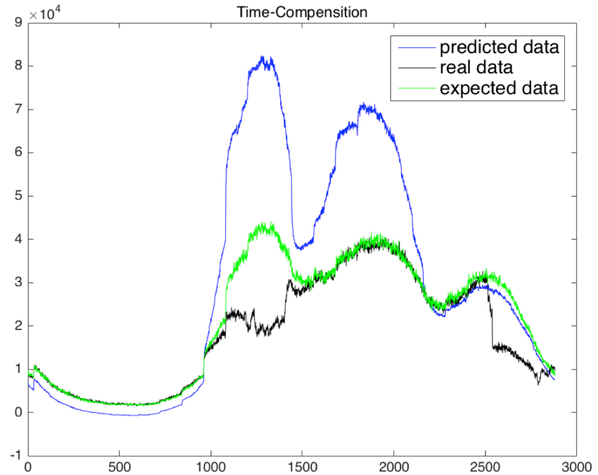
- The results were lots of missed or false alarms

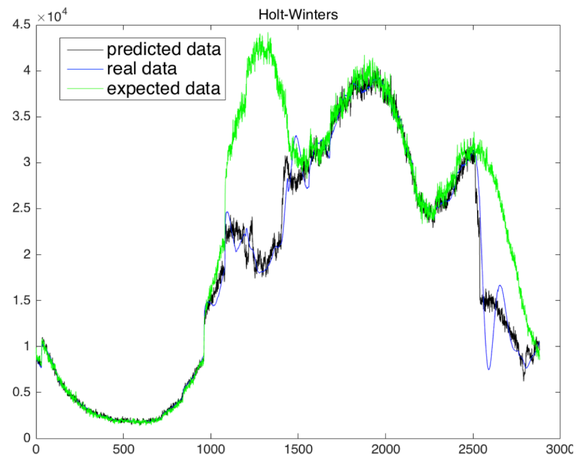
Some tried but failed Ideas

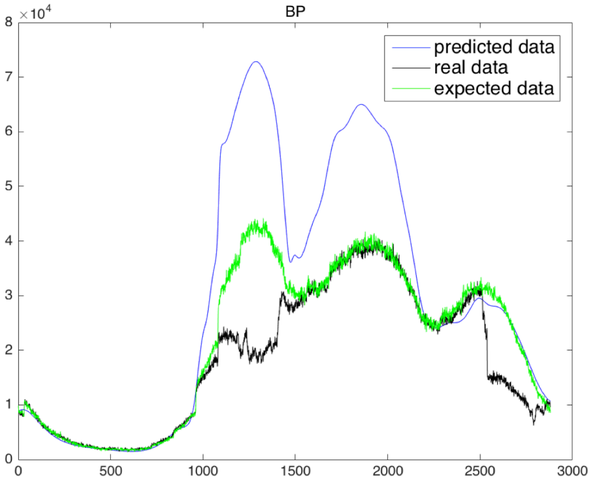
- Median compensation
- Time compensation
- Holt-Winters
- BP

- So we turn to some data mining methods.









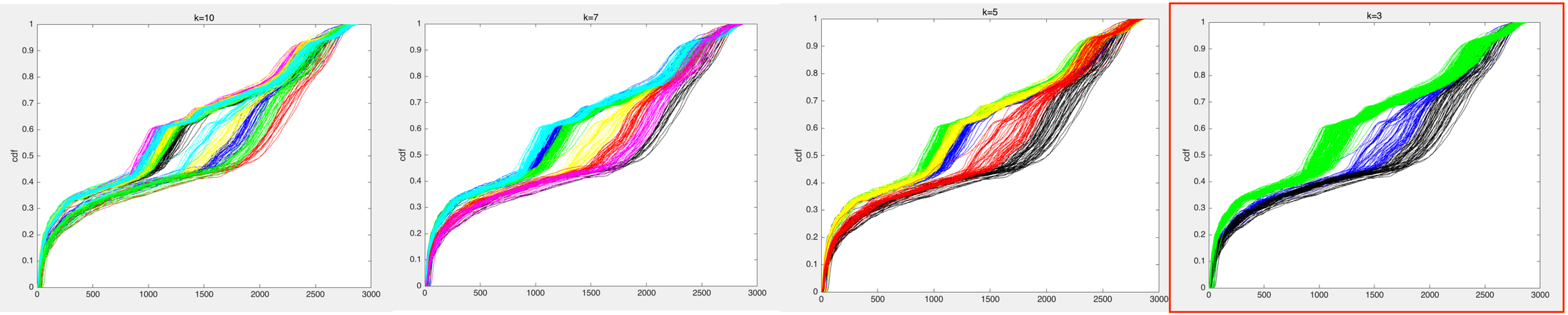
The difficulties

- Infrequency means NO enough training data
  - ✓ Holidays are infrequent
- We also cannot use the seasonality of time sequence
  - ✓ In China, holiday dates are not fixed

Year	Spring Festival	Dragon boat	Mid Autumn
2015	Feb. 19	Jun. 20	Sept. 27
2016	Feb. 8	Jun. 9	Sept. 15
2017	Jan. 28	May 30	Oct. 4

First Idea - Date Clustering

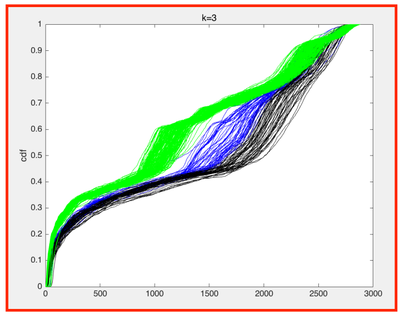
- Can we find as many as possible "similar" dates?
  - ✓ Clustering on CDF of everyday's data curve (k-means)



K=10 K=7 K=5 K=3

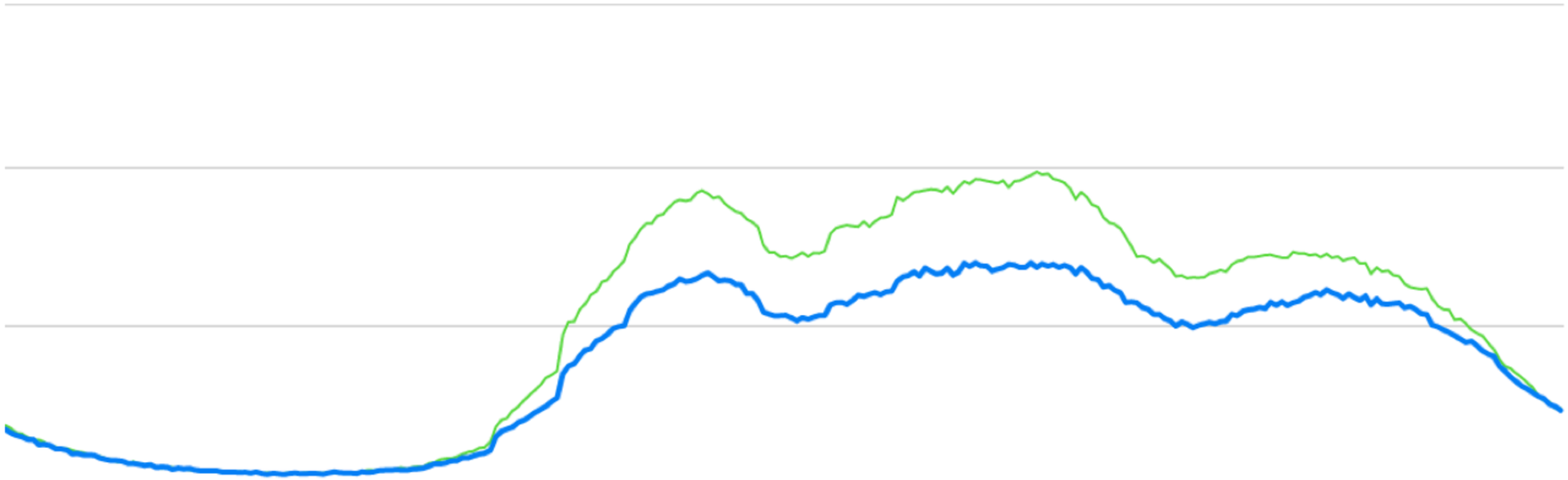
More details

Date	Green	Blue	Black
Working day	208	0	0
Sat.	0	30	7
Sun.	0	10	29
Festivals	0	1	21
Specials	0	1	3



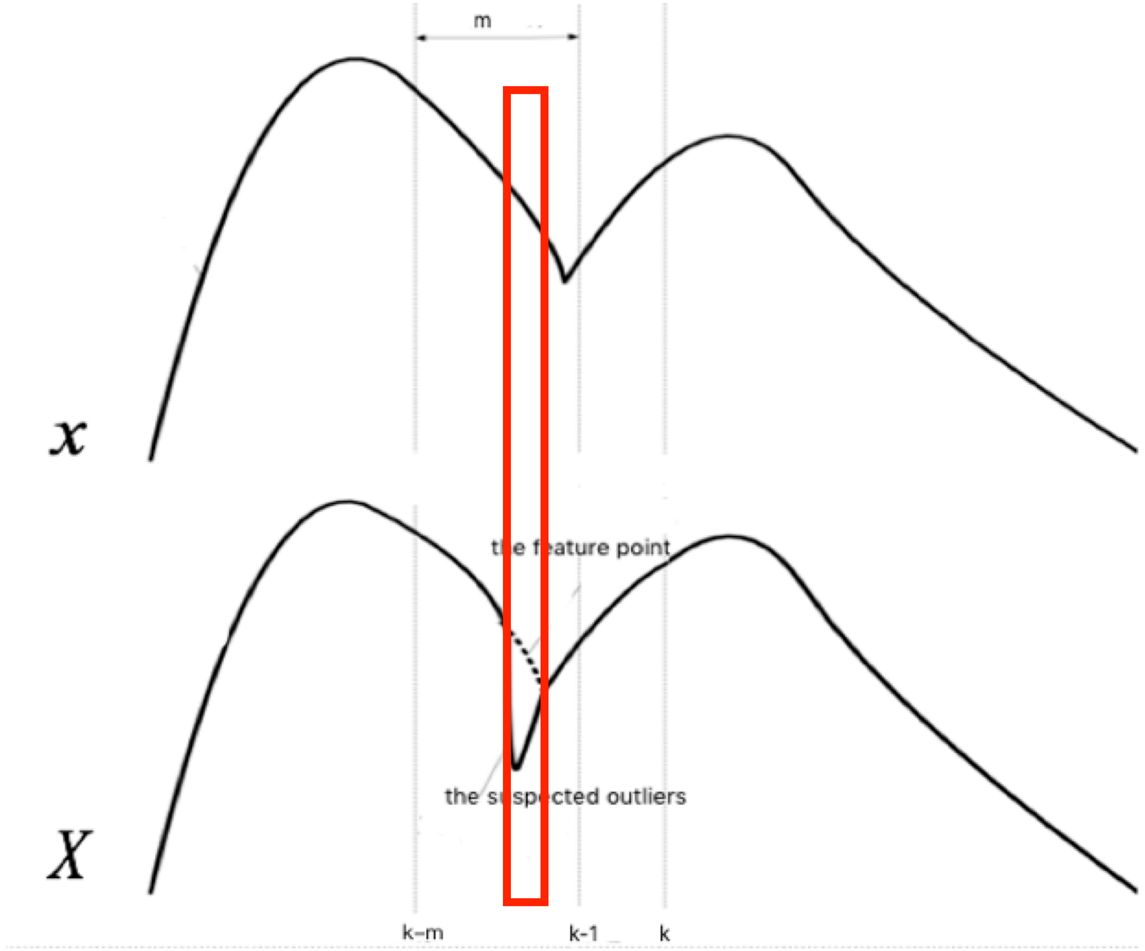
(3/1/2015 - 1/4/2016, Total: 310 days)

However, still some gap



- CDF reflects trend, not the exact values in points

Second Idea - Real time fixing

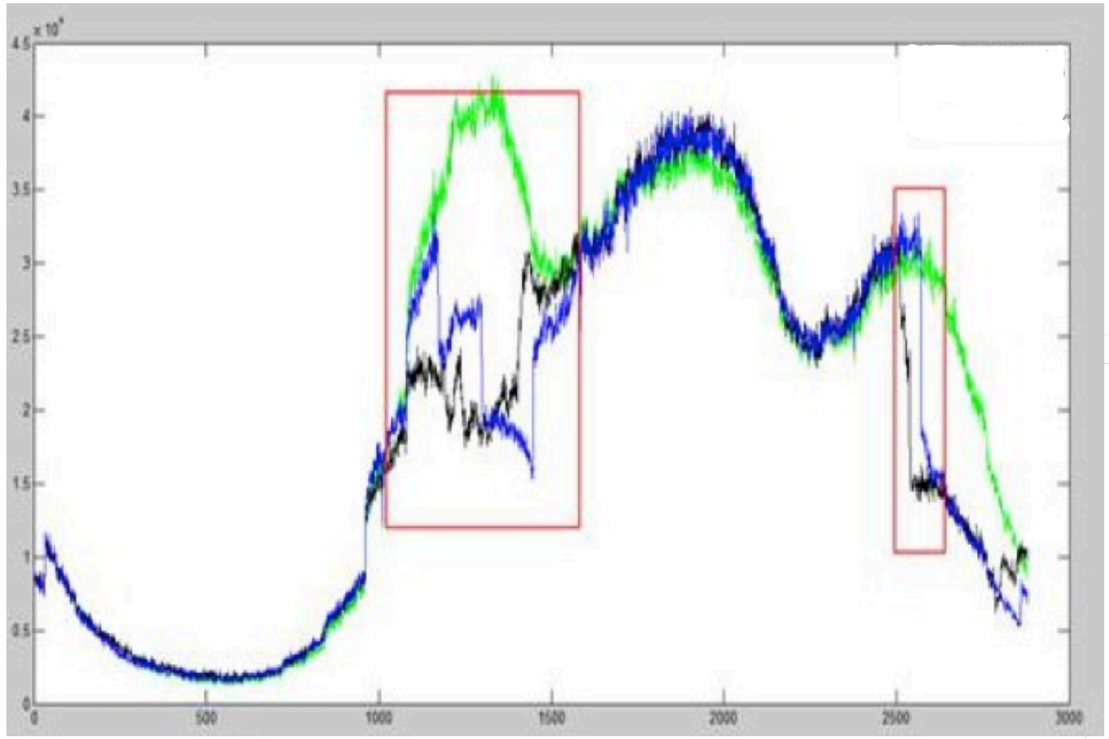


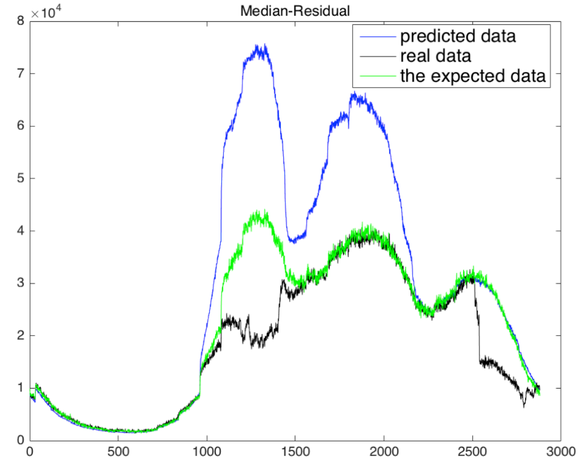
$$\frac{\hat{X}(k)}{X(k-1)} = \frac{x(k)}{x(k-1)}$$

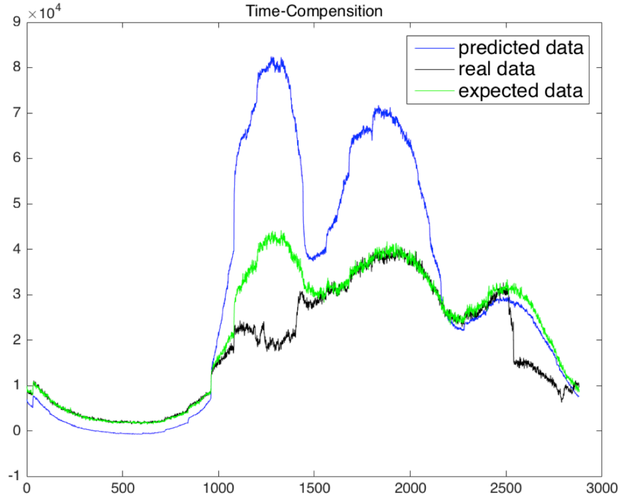
$$\frac{\hat{X}(k)}{\sum_{R(1)}^{R(len\_R)} X(l) + \sum_{W(1)}^{W(len\_W)} \hat{X}(L)} = \frac{x(k)}{\sum_{k-m+1}^{k} x(j)}$$

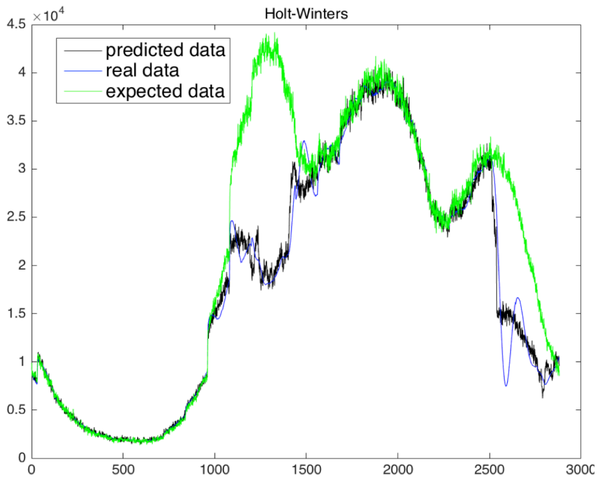
$$\frac{\hat{X}(k)}{\sum_{k-m+1}^{k} X(l)} = \frac{x(k)}{\sum_{k-m+1}^{k} x(j)}$$

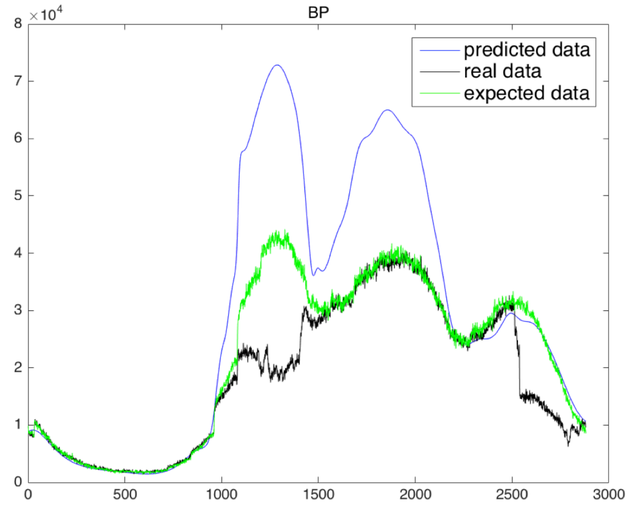
The experiment result





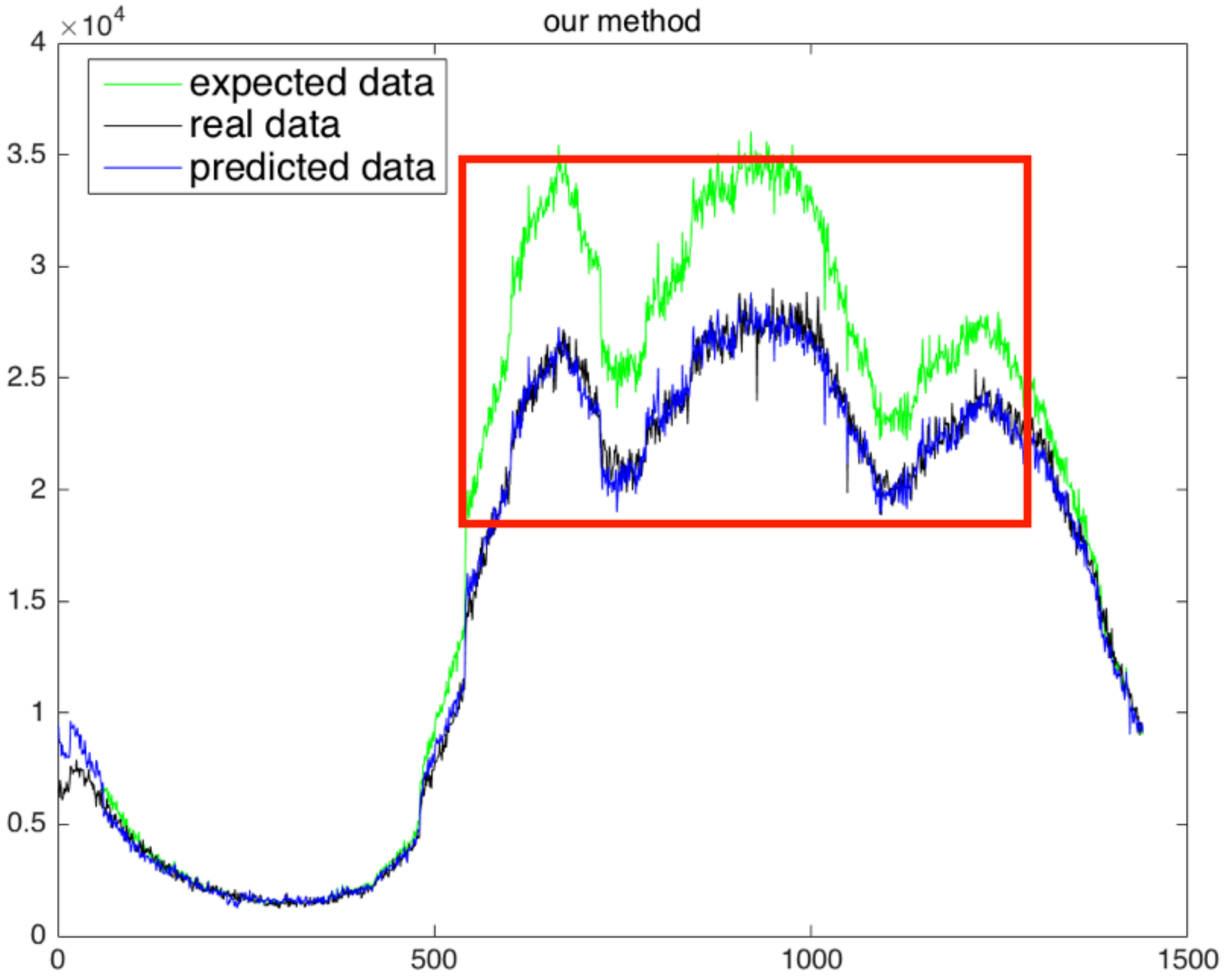






The real deployment

- Jan. 1st, 2017



Thanks for your Attention

and

any questions?

wangdong13@baidu.com