Air pollution monitoring, forecasting, and causality analysis with spatiotemporal (ST) urban big data

Julie Yixuan Zhu
Email: zhuyx08@gmail.com

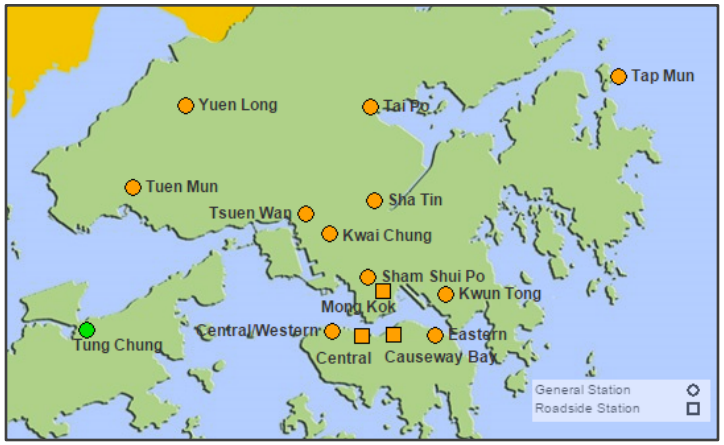
Air pollution has been a major problem in China

China World Trade Center Tower III, Beijing

China World Trade Center Tower III, a 1,000 plus foot skyscraper that's one of the tallest in Beijing. Jan 12th,2013. Photo courtesy of Bill Bishop/Sinocism China Newsletter

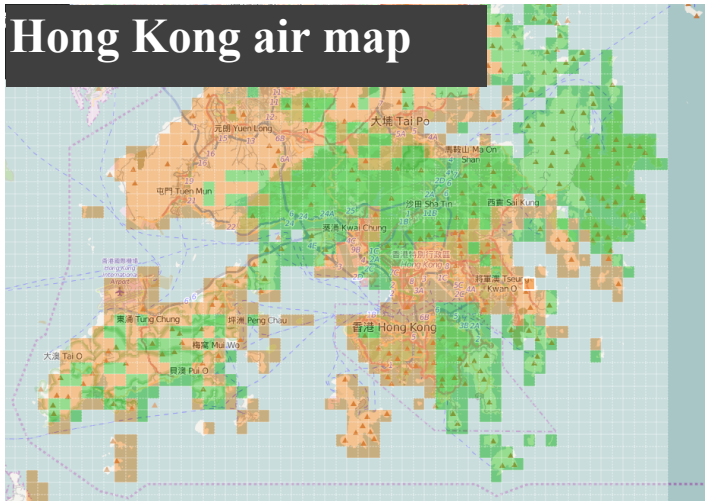
Data collected from urban monitoring systems

- However, few monitoring stations
  - 15 stations in Hong Kong
  - 35 stations in Beijing
- High cost for a city-wide monitoring system



Hong Kong

http://www.epd.gov.hk



What can urban big data do?

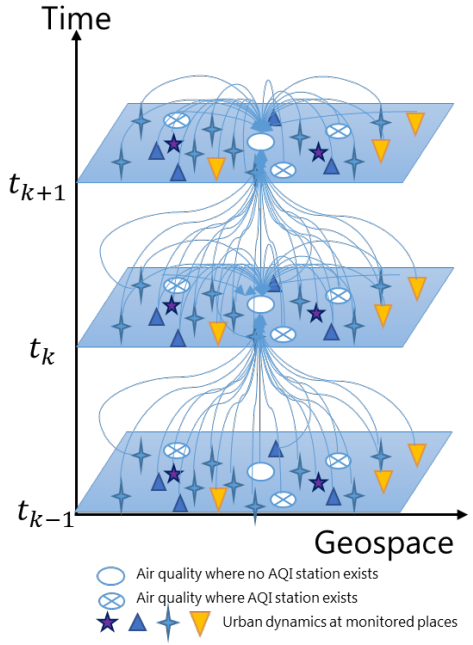
Three projects
- Causality-based air quality monitoring
- Causality-based air quality forecasting
- Identifying the causality for air pollutants

- 1. Millions of monitoring records in one city

Type	Category	Urban dynamics	Tuple format	Amounts	Source
Air pollution	1	AQI	(VALUE, S,T)	133315	SLEN & Cooperative dataset from SEMC
	2	PM2.5	(VALUE, S,T)	133315	
	3	PM10	(VALUE, S,T)	133315	
	4	NO2	(VALUE, S,T)	133315	
	5	CO	(VALUE, S,T)	133315	
	6	O3	(VALUE, S,T)	133315	
	7	SO2	(VALUE, S,T)	133315	
Meteorology	8	Pressure	(VALUE, S,T)	36796207	SZMB
	9	Humidity	(VALUE, S,T)	36796207	
	10	Temperature	(VALUE, S,T)	36796207	
	11	1 hour rain	(VALUE, S,T)	36796207	
	12	R24H	(VALUE, S,T)	36796207	
	13	Wind	(VALUE, S,T)	36796207	
Traffic	14	Traffic speed	(VALUE, S,T)	7892434	SUMAP
	15	Traffic index	(VALUE, S,T)	7892434	
Geography	16	POI	(VALUE, S)	153225	BAIDU Map
	17	Urban morphology	(VALUE, S)	500000	
	18	Roadmap	(VALUE, S)	500000	SECL

Basic information of urban dynamics datasets

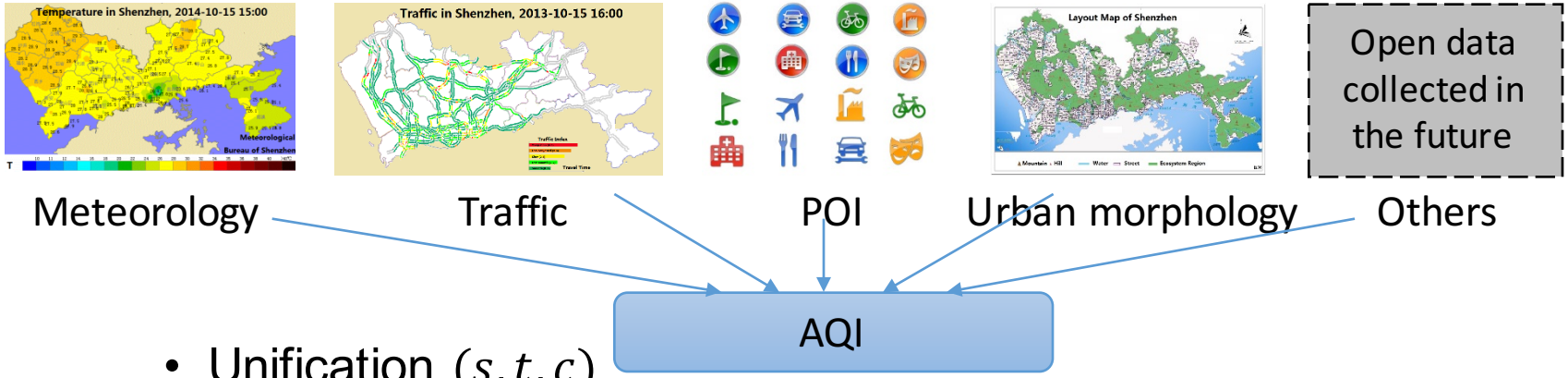
Note - Yellow: not completely open data, Green: open data



- Two philosophies of big data processing:
  - Process everything, with vast amount of computing resources.
  - Process some of the data, with less computing resources, but get approximate results.

1. *Causality* based air quality monitoring

1. Deal with data diversity



- Unification (s, t, c)

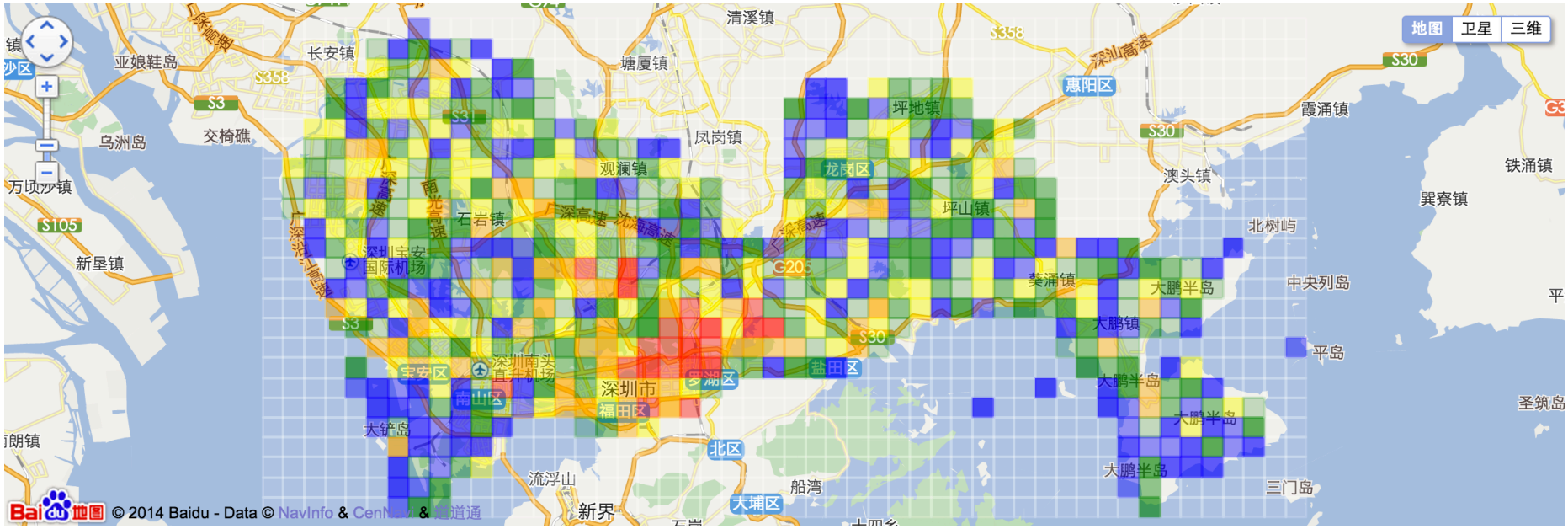
2. Transform “Big data” into “most influential data” based on causality, regarding to:

- Category c.
- Space s.
- Time t.

Map view of causality levels from all the other grids for a target grid

airinshenzhen.azurewebsites.net/professional/

Data in Shenzhen | HongHu | Liyuan | Huaqiaocheng | Nanyou | Yantian | Longgang | Xixiang | Nanao | Kuiyong | Meisha | Guanlan | Basic observation | contact



Description : This map illustrates the causality level of all the places in shenzhen with respect to HongHu in colors
There are 10 levels in total,each represented by a color. The index is shown by the image on the right.

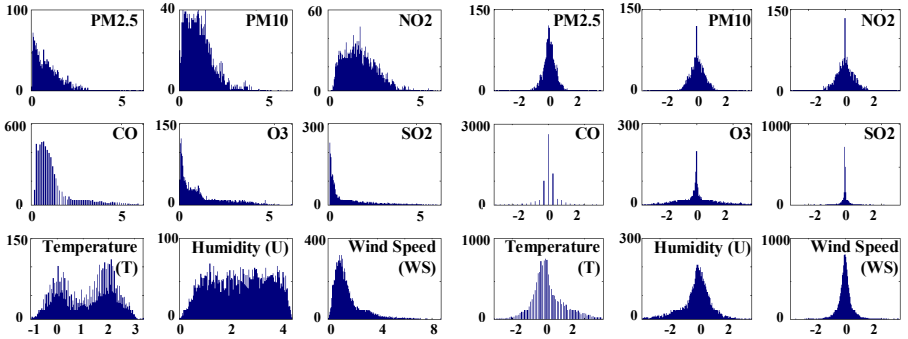
color index

1 2 3 4 5 6 7 8 9 10

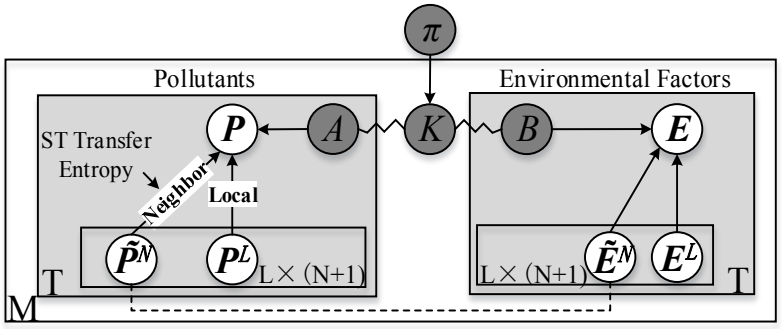
Back to top • updated time : 2014/10/24 17:00
• station : HongHu monitoring station

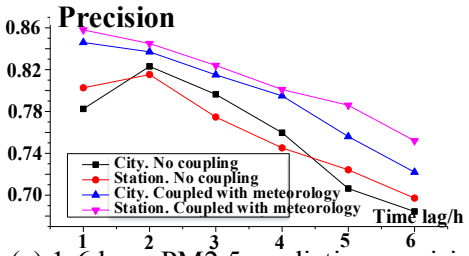
2. *Causality* based air quality forecasting

- Traditional time series models (ARMA, regression, SVM,ANN…)
- Physical models (Box model, dispersion model…)
- Causality model
  - 1) Model the Gaussian mixtures
  - 2) Use urban big data for better parameter learning.
  - A variable *K* that simultaneously affects the dynamics of pollutants and environmental factors.



(a) Original values normalized by standard deviation (b) Normalized 1-hour difference



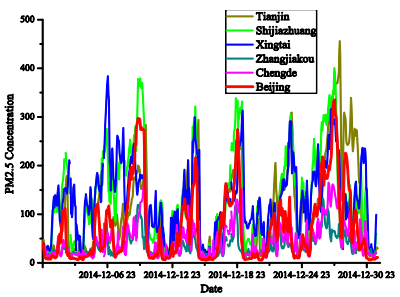


(a) 1-6 hour PM2.5 prediction precision coupled or not with meteorology

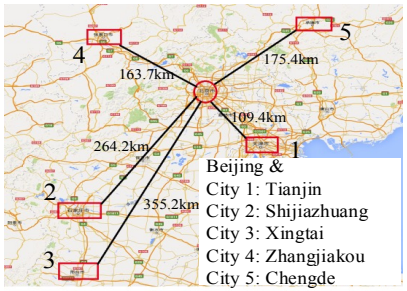
Method	1-hour prediction precision	
	City	Station
Our method	0.804	0.859
Our method without coupling	0.789	0.832
Hill climbing	0.528	--
MCMC	0.597	--
K2 + PS	0.684	0.753
CI test:	0.382	0.298

(b) 1-hour prediction precision compared with four baselines

3. Identify the *causality*

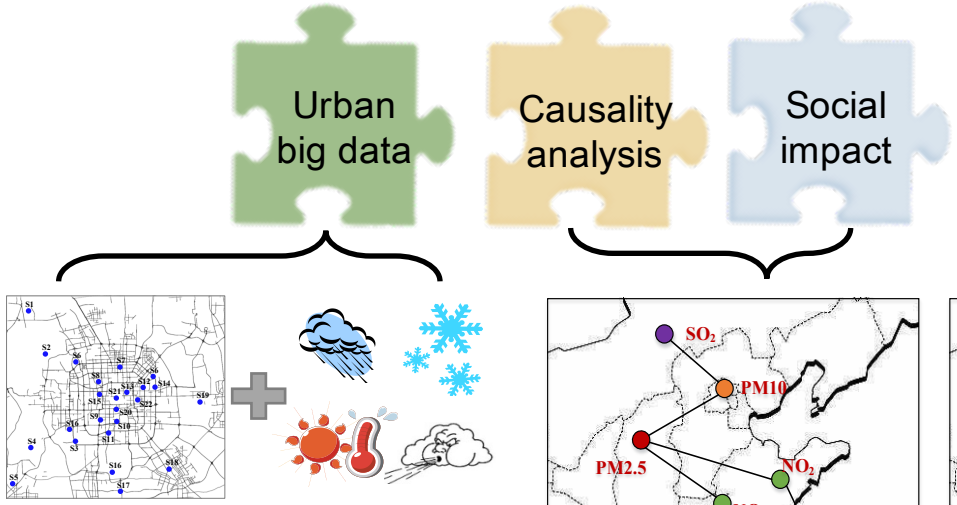


A) PM2.5 in Beijing and 5 neighbor cities

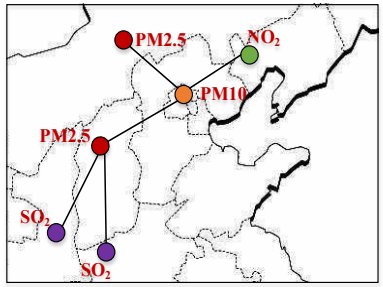


B) Relative locations and distances

The time series of pollutants show similar trends even 300km far away!



(a) Causal pathways (wind < 5m/s)



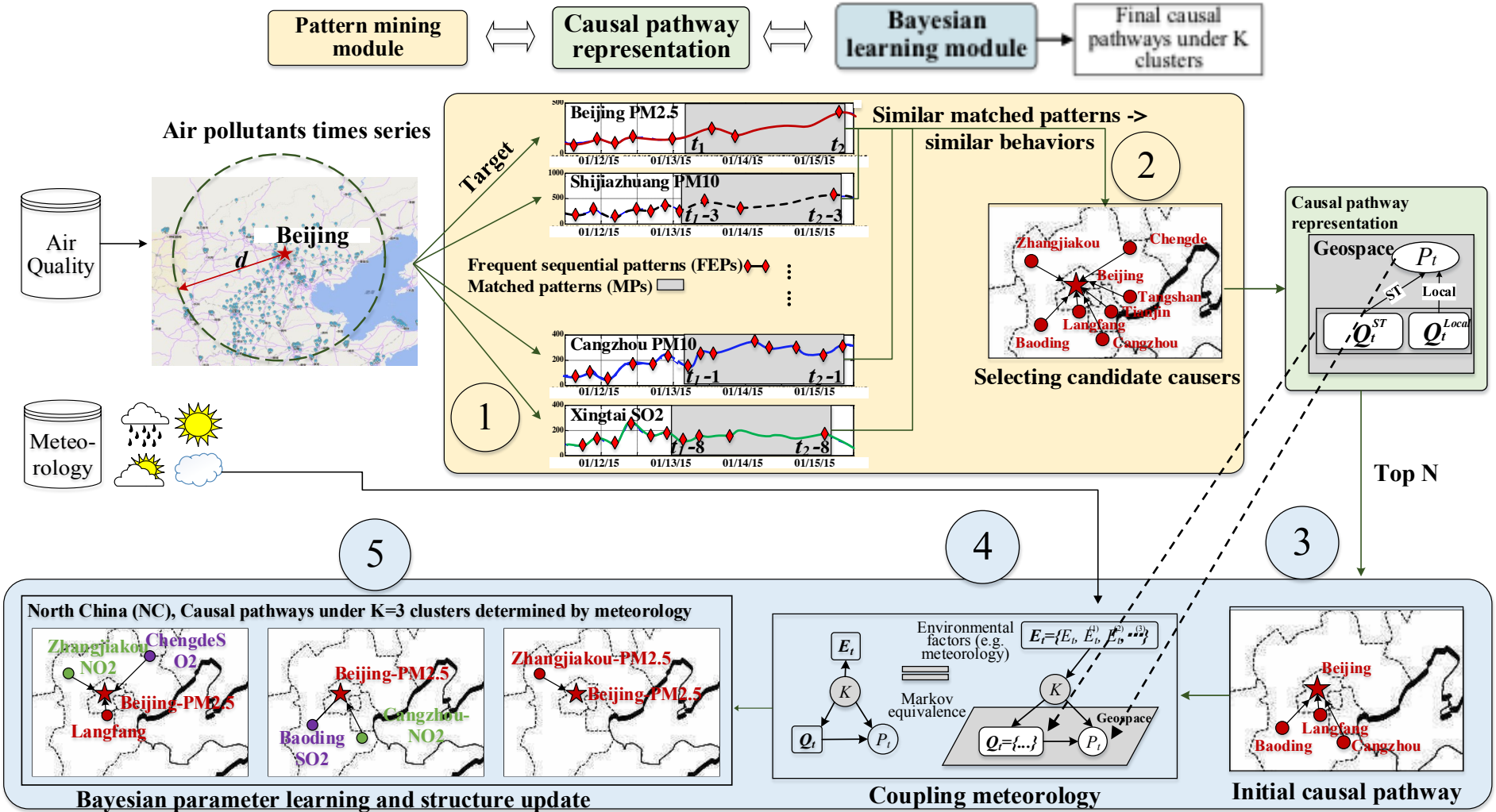
(b) Causal pathways (wind > 5m/s)

- Causality analysis with urban big data.
  - Inform the public policy making
  - Where do the pollutants come from?
  - How much do they cause each other?

Question is:
I know the air quality is bad, but I am living in Beijing. What can I do?

We propose p-Causality: a pattern-aided causality analysis approach.

- Combining the strengths of pattern mining and statistical modeling.



Conclusion

- Air quality monitoring and causality analysis based on urban big data
  - It is not needed process "all" the data.
  - Better precision and time efficiency can be achieved when transforming "big data" into "the most influential data".
- Open issues
  - More and more spatiotemporal data generated every day, e.g., complex system and the internet of things (IoT). Data can advance research and industry like never before.
  - Geographically sparse data
  - Can the methodology be transferred to other types of data? How do we model these data?