Towards Spatial Data Science

John W Z SHI
Dept. of Land Surveying and Geo-Informatics
The Hong Kong Polytechnic University
E-mail: lswzshi@polyu.edu.hk

Outline

- Science
- Data Science
- Spatial Data Science
- Basic scientific issues and challenges
- Uncertainty handling
- Summary

Science

- “Knowledge covering general truths of the operation of general laws, especially as obtained and tested through scientific method ” (Webster's New Collegiate Dictionary)

- Knowledge in the form of testable explanations and predictions about the universe.

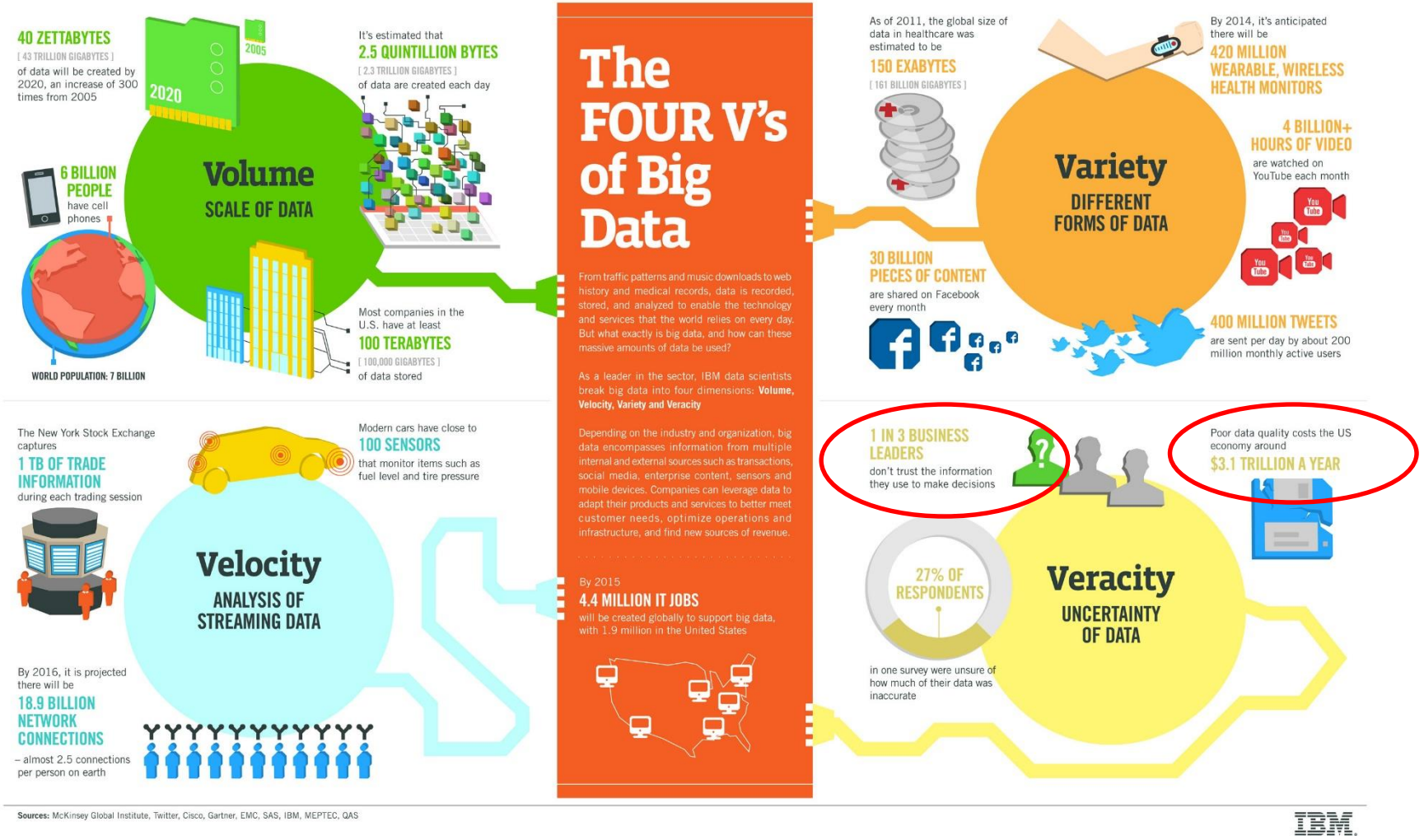
Science

- Knowledge itself that can be rationally explained and reliably applied
- Philosophy of nature, or natural science
- A disciplined way to study the natural world, such as physics
- A broad sense to denote reliable and teachable knowledge about a topic, such as computer science

Data Science

- Extraction of knowledge from data
- Initially as a substitute for computer science (Peter Naur 1960)
- "Statistics = Data Science?" (C.F. Jeff Wu 1997)
- Multidiscipline: mathematics, statistics, information theory and information technology

Big Data



From: https://www.google.com/search?hl=zh-CN&site=imghp&tbm=isch&source=hp&biw=881&bih=434&q=BIG+Data+IBM&oq=BIG+Data+IBM&gs_l=img.12...0.0.0.2179.0.0.0.0.0.0.0.0..0.0....0...1ac..64.img..0.0.0.Brd1z3N-yPw

Spatial Big Data

Spatial data subject to uncertainty

奇华饼家

地址：红磡黄埔花园11期聚宝坊地下G3D铺

交通：尚无公交信息　　交通/驾车路线

奇华

地址：红磡德康街6号黄埔新天地黄埔花园11期地下G3D号铺

交通：尚无公交信息　　交通/驾车路线

Go: 35min； Back: 20min

Spatial Data Science (SDS)

Can be defined as a science of discovering spatial knowledge and explaining spatial laws about the universe based on spatial (big) data.

Fundamental Scientific Issues of SDS

- **Spatial discovery and explanation**
  - Example: Discover and explain human motion behavior and regularities in urban spatially and temporally

- **Spatio-temporal prediction**
  - Example: Predict trend of human motion and social events

A multidisciplinary nature of SDS

- The basic supporting theories include
  - data science,
  - geographic information science
  - computing science
  - spatial statistics
  - machine learning
  - spatial data mining
  - …
- SDS is multidisciplinary

Challenges

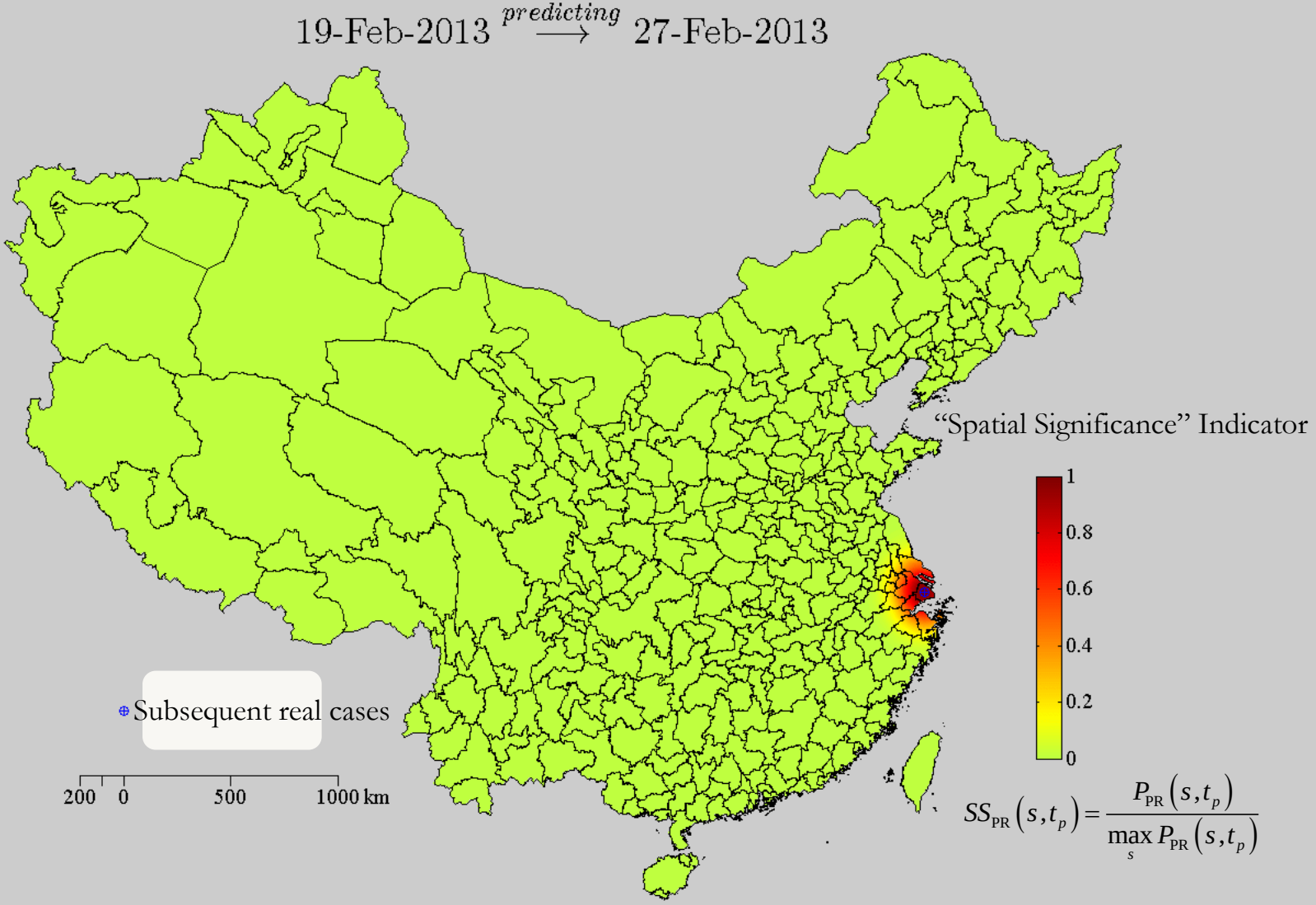
- Spatial Description and Representation
- Spatial Big Data Analytics and Discovery
- Heterogeneous Information Integration
- Uncertainty Handling in Spatial Big Data
- Machine Learning based on Spatial Big Data
- Towards Spatial Data Science

Research infinitives for developing SDS

- Spatial big data analytics and discovery
- Spatio-temporal prediction of natural and human phenomena
- Description and representation of spatial big data, especially for unstructured spatial data
- Spatial visual analytics
- Integration of heterogeneous spatial data and information
- Uncertainty modeling for spatial big data

Example: spatio-temporal prediction

An example of spatio-temporal prediction: H7N9



A **Spatiotemporal Proximity** Integrated Framework for Predicting the Infection Risk of Emerging Infectious Disease

(1) Retrospective Inference

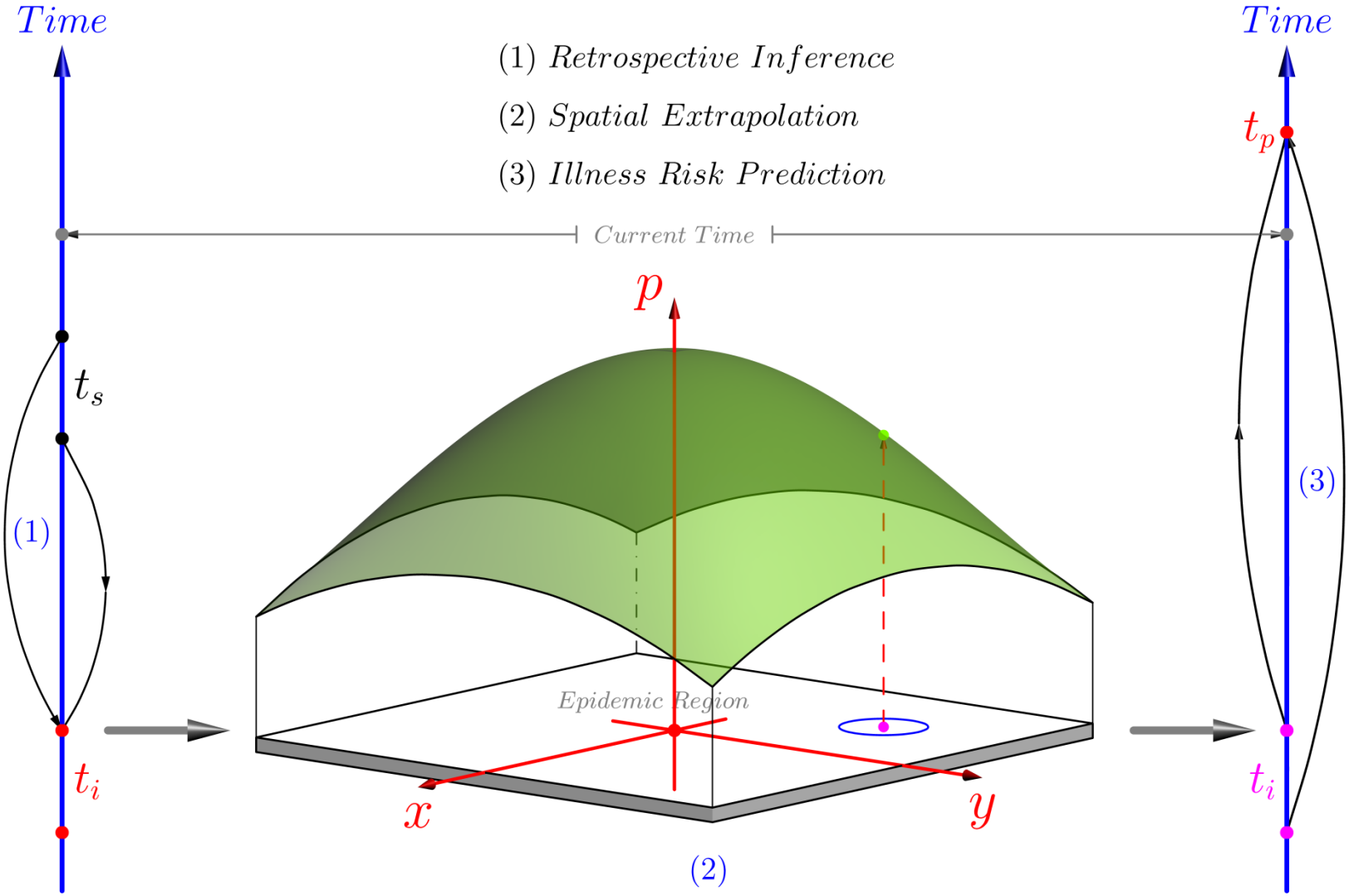
(2) Spatial Extrapolation (Kernel Density Estimation)

(3) Illness Risk Prediction

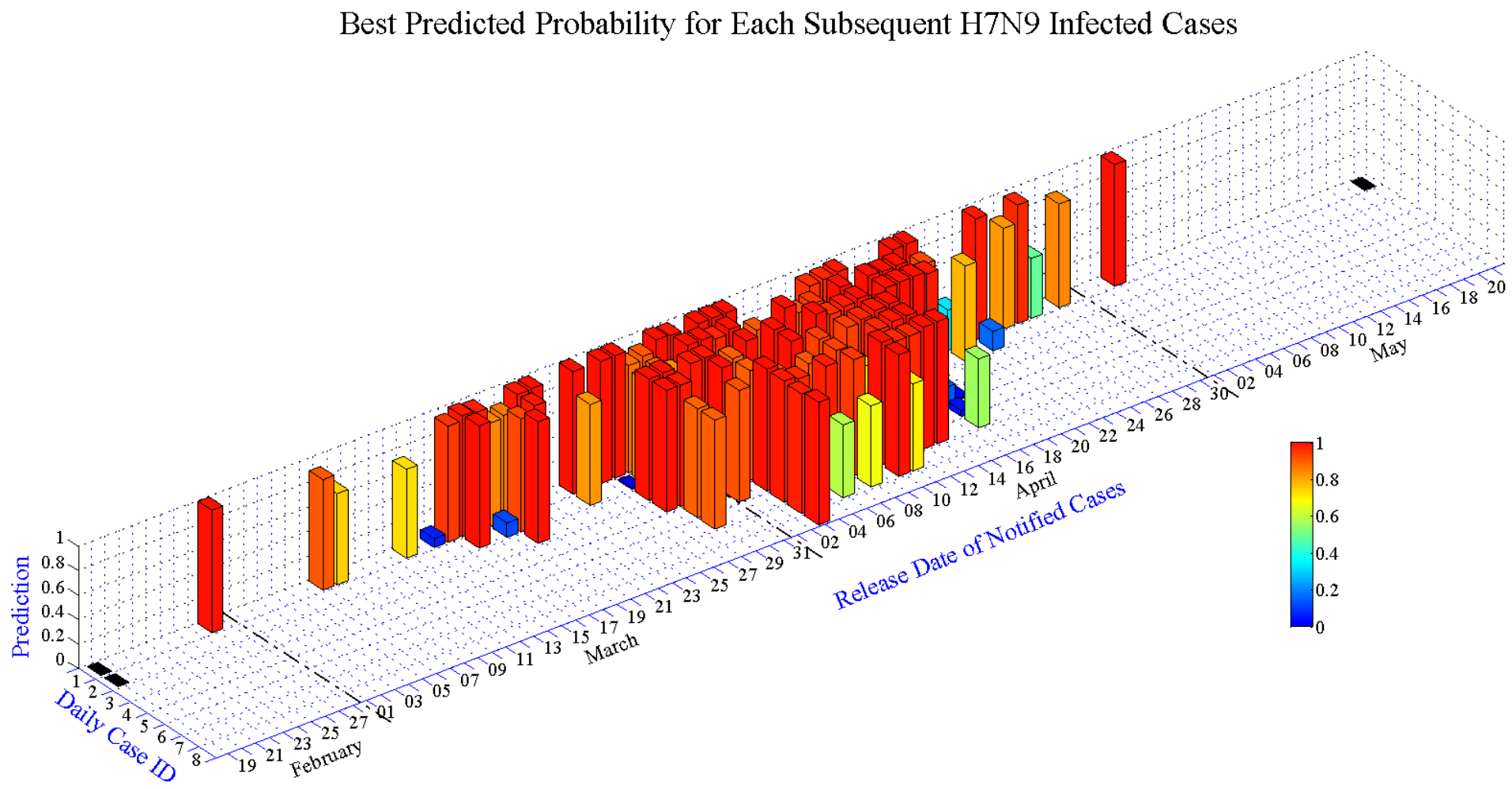
Experimental Data Set

Laboratory-confirmed cases infected with the avian influenza A H7N9, February to May 2013 in eastern China.

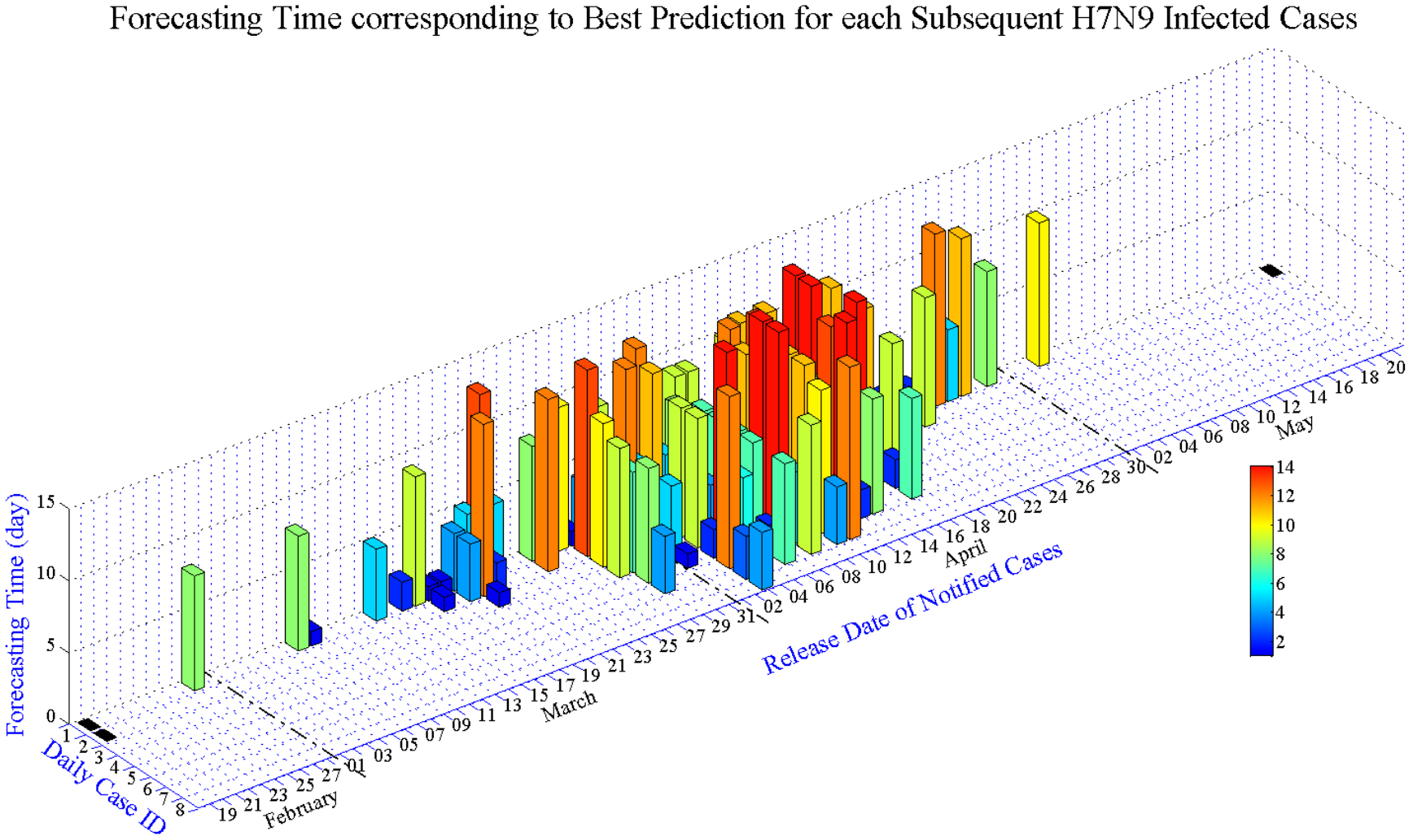
Operational Scheme



Experiments and Discussions



Experiments & Discussions

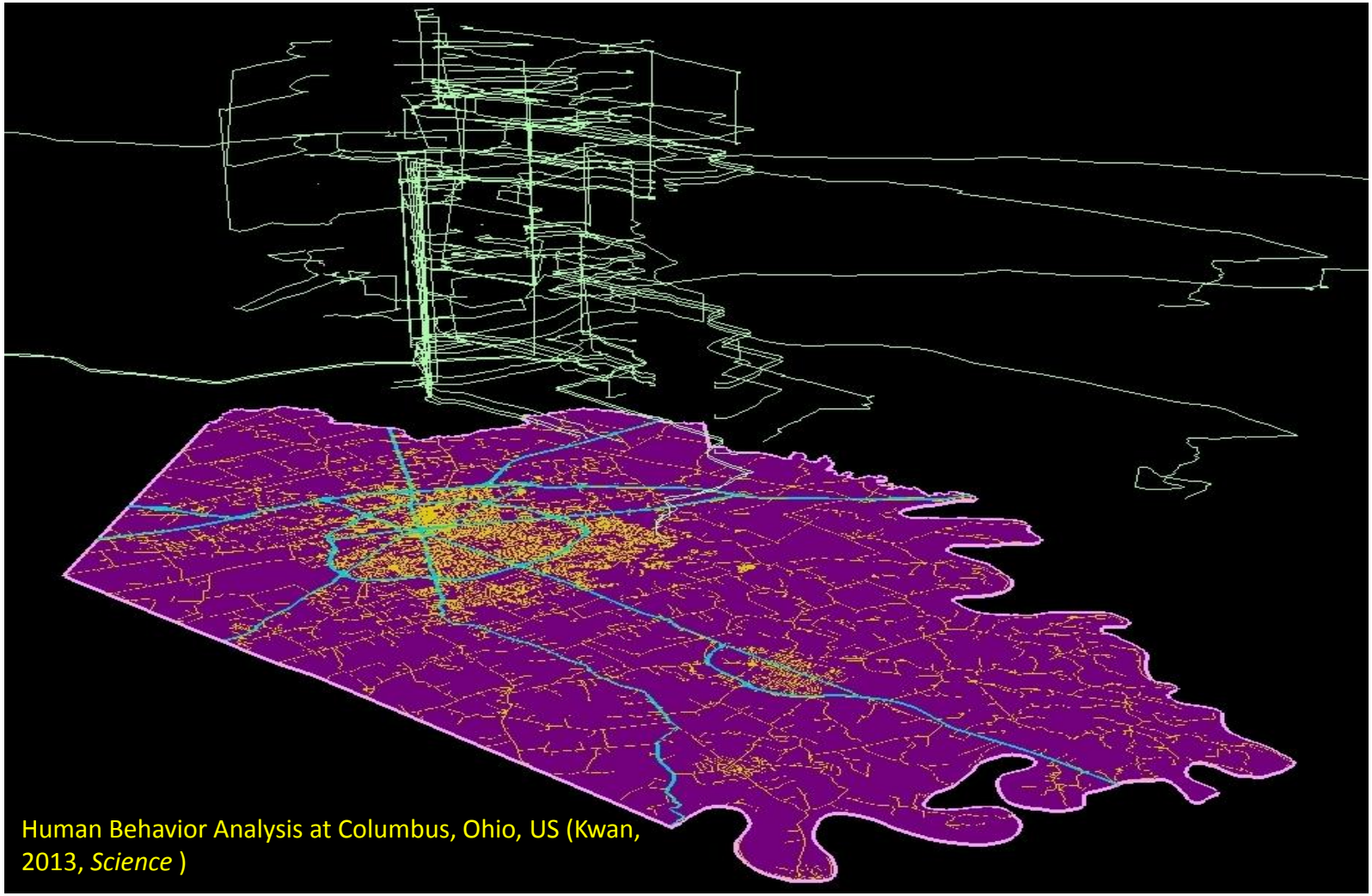


Note

- A model has been proposed for studying the impact of spatio-temporal proximity upon the infection risk of Emerging Infectious Disease.
- Experiments upon avian influenza A H7N9, February to May 2013 in Eastern China, demonstrates that the proposed model can provide 70% correct prediction for the coming 5 days.
- Findings can be used for exploring the spatio-temporal propagation pattern of EID, making short-term predictions.

Example: spatial discovery

An example of spatial discovery



Human Behavior Analysis at Columbus, Ohio, US (Kwan, 2013, *Science*)

Example: uncertainty modelling

Logic flow for uncertainty modellin

Presentation

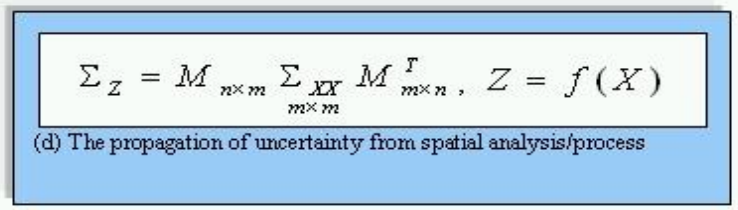
Data quality control

Propagation of uncertainty

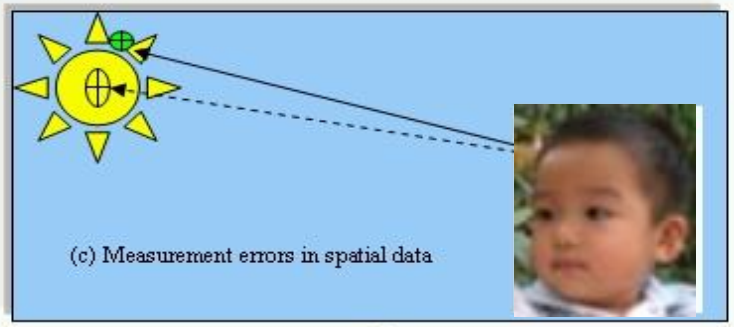
Spatial distribution

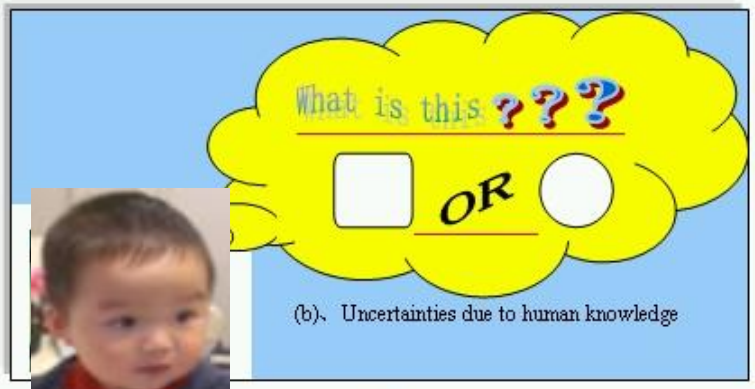
Uncertainty in cognition

Uncertainty in reality



(d) The propagation of uncertainty from spatial analysis/process


(c) Measurement errors in spatial data



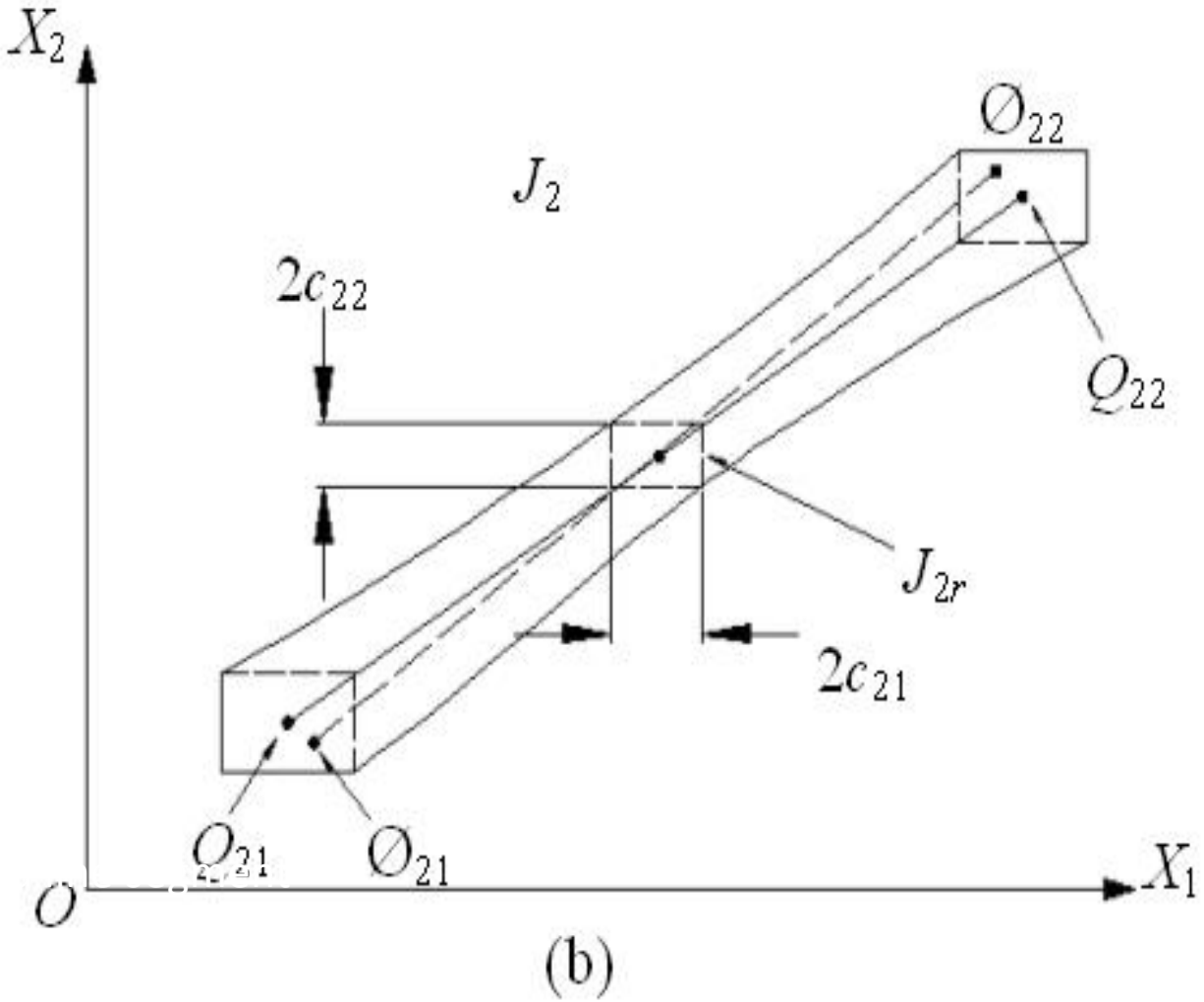
(b)、Uncertainties due to human knowledge


(a)、Uncertainties inherent in the real world

Uncertainty Vs Error

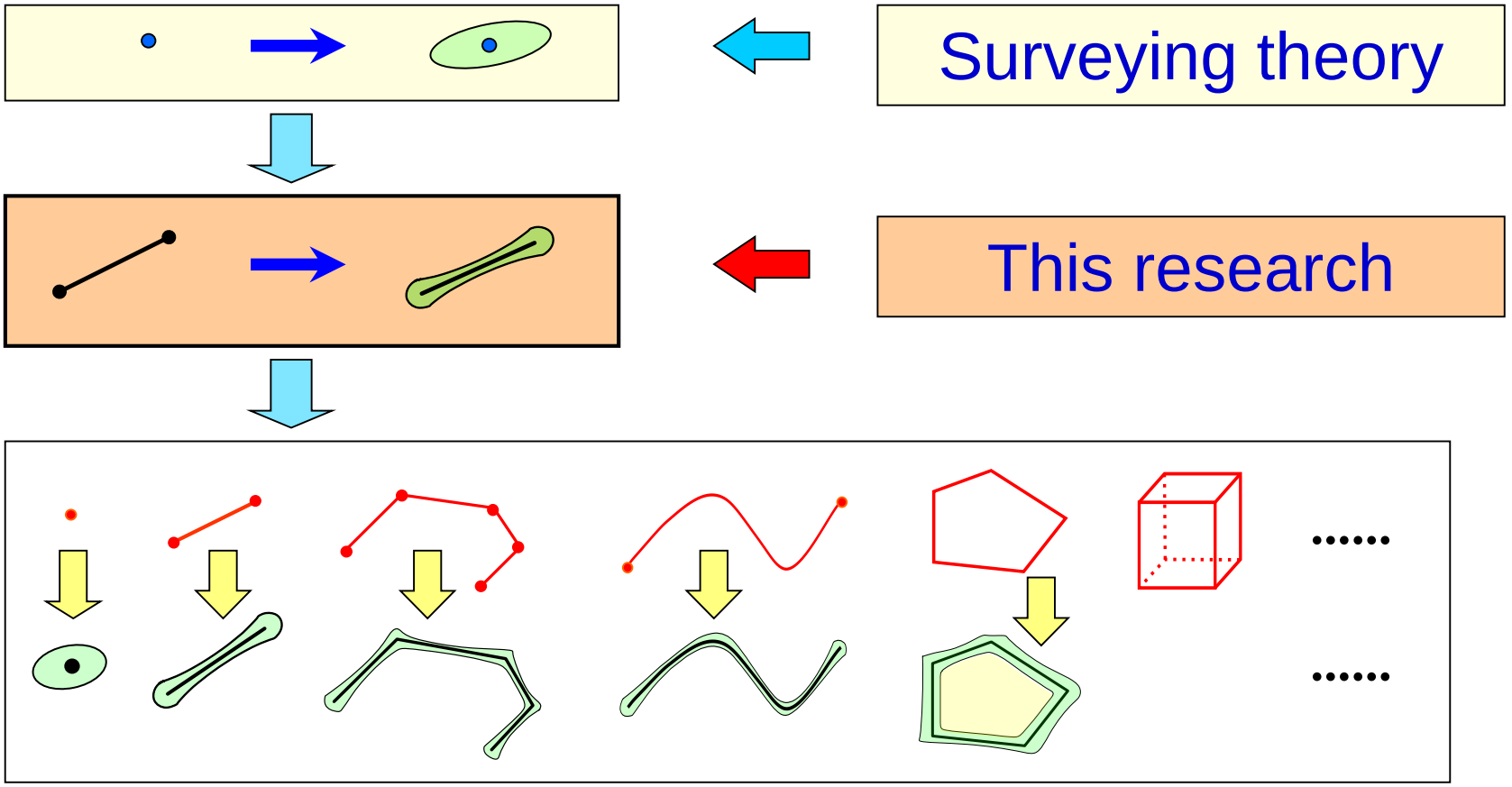
	Classification	Definition
Uncertainty	**Imprecision**	The level of variation of a set of measurements, or the lack of quality precision
	Ambiguity	Associating with either one or many relationships, or with a form of lack of clarity, implying one or more meanings.
	Vagueness	Lack of clarity in meaning and associating with the difficulty of making a sharp or precise distinction in relation to an object..
Error	**Random**	Irregular errors in terms of magnitude and sign
	Systematic	The magnitude and type of error following a regular pattern
	Gross	Mistakes

The confidence region model



(b)

Significant of the research development



Four breakthroughs

- From determine- to uncertainty-based representation
- From modelling uncertainty in static data to dynamic spatial analyses
- From modelling uncertainty for spatial data to spatial models
- From uncertainty description to spatial data quality control

Example: uncertainty-based spatial data mining

Mining spatial association rules from data with uncertainty

- (Spatial) association rule mining face risk of falsely 'discovering' numerous spurious rules
  - Typically 10%~90% resultant rules are fake
- Statistical method on (spatial) association rules
  - Over-conservative, rendering loss of true rules (much severer with imprecise data)

Mining spatial association rules from data with uncertainty

- Achievement recovering >50% true rules that are lost due to uncertainty in data
  - E.g. rules of land use changes vs. socio economic data, Massachusetts

(Rules discovered from error-free data = 100%)

Concluding remarks

- Spatial Data Science (SDS) is a trend for our discipline in the era of Big Data.
- GIS and then SDS?
- Fundamental scientific issues of SDS: a) spatial discovery and explanation and b) spatio-temporal prediction.
- Uncertainty is one of the key issues for SDS.
- One nature of SDS is multidisciplinary.
- SDS will have a significant impact on both natural and social sciences in the future.