Unstructured Data integration capabilities of GIS

Dr. Kingshuk Srivastava
Bhagwant Singh
Ankit Vishnoi

|March 2016|

What is unstructured data?

a generic label for describing any corporate information that is not in a database

information that either does not have a pre-defined data model or is not organized in a pre-defined manner

unstructured data : Data that does not reside in fixed locations

any data that has no identifiable structure

The value of unstructured data sources

- Provide a rich source of information about people, households and economies

- May enable the more accurate and timely measurement of a range of demographic, social, economic and environmental phenomena
  - Combined with traditional data sources
  - As a replacement for traditional data sources

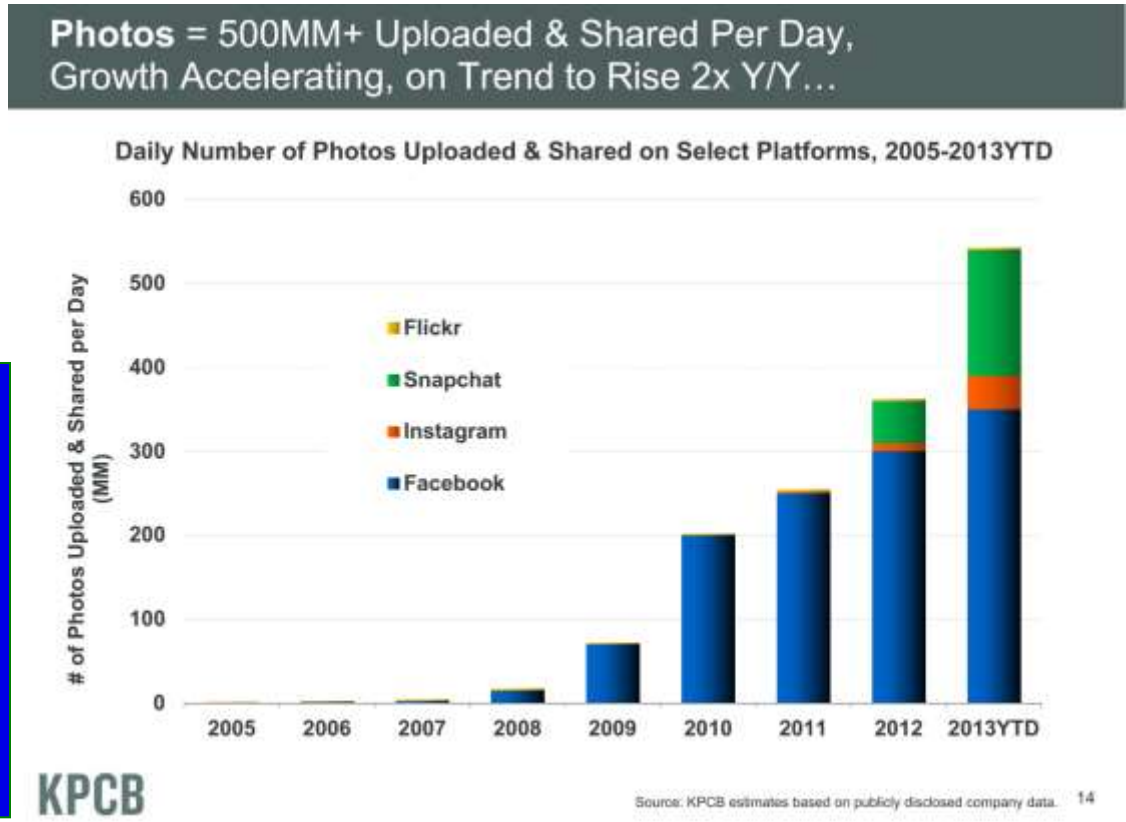
- So presents unprecedented opportunities for official statistics to
  - Improve delivery of current statistical outputs
  - Create new information products not possible with traditional data sources

Why is unstructured data important?

- Unstructured data doubles every three months
- 7 million web pages are added every day
- 80% of business is conducted on unstructured information
- 85% of all data stored is held in an unstructured format

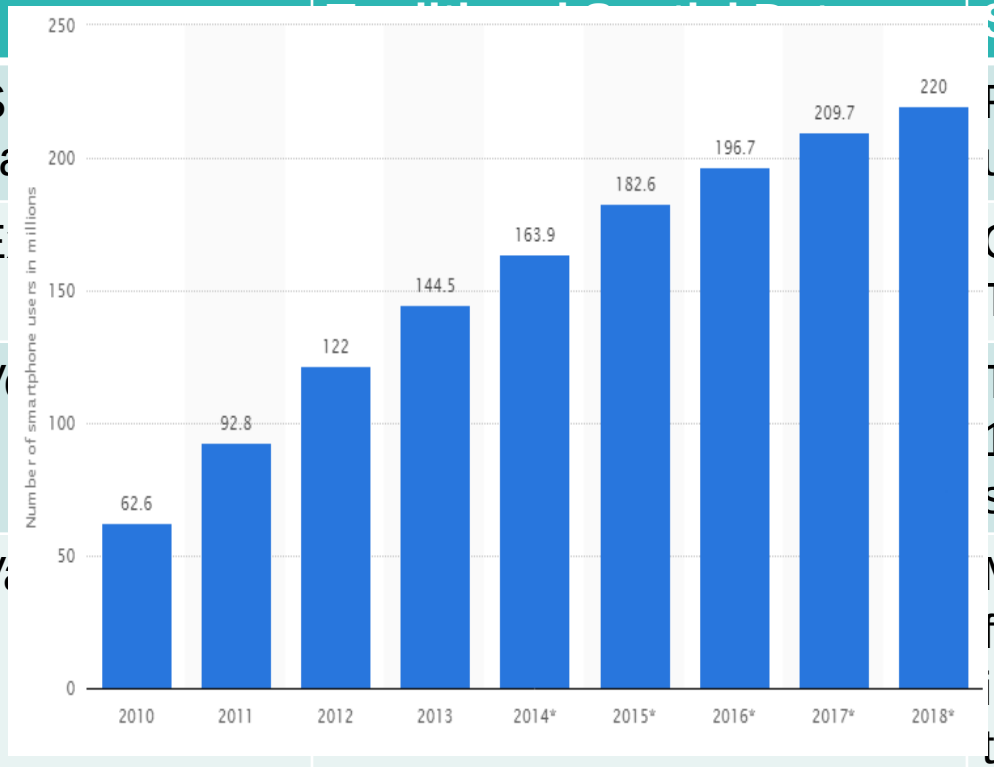
Some Trends

- Images are a major source of Unstructured Data
  - Radar
  - Light Synchrotrons
  - Smart Phones
  - Bio-imaging



- **Hadoop and HDFS dominant**
- **Business – main emphasis at NIST – interested in analytics and assume HDFS**
- **Academia also extremely interested in data management**
- **Clouds v. Grids**

Spatial Data vs. Spatial Big Data



		Spatial Big
S / ca		Real time r / user-genera ze
E		Check-ins / Temporally
V		Temporally detailed maps can reach 10^{13} items per year; GPS traces from smart phone
V		M / fr / ir / tv
Velocity	Limited velocity (census-decade)	H / m / to
Dimensionality	2D, 3D	T

Smart phone GPS-trace data size estimation

Time(date, clock), **Location**(x,y,z), **Metadata**

64 Bytes

10 min: 64 X (160 X 10^6) X (6 X 24) -> 1.5 TB per day

10 sec: 90 TB per day

100 sec: 900 TB per day, close to 1 PB

Big Data vs. Spatial Big Data

	Big Data	Spatial Big Data
Examples	Facebook/Twitter posts Google search terms	Geo-located tweets and posts Open Street Map
Data Types	Text keywords Web logs	GPS traces; geo-located social platform posts Temporally detailed roadmaps Frequently collected satellite/UAV imagery
Questions	Google brain: Does an image contain a cat?	Are there any hotspots of recent disaster-related tweets? Where?
Representative Computational Paradigms	Hadoop Hashing Sub-problem optimization (learning)	Spatial Hadoop, GIS in Hadoop Declustering Spatial partitioning

spatial queries: partitioning
data skew; boundary objects

Spatial Big Data (SBD)

- SBD are important to society
  - Ex. Eco-routing, Public Safety & Security, Understanding Climate Change
- SBD exceed capacity of current computing systems
- DBMS Challenges
  - Privacy vs. Utility Trade-offs
- Platform Challenges
  - Map-reduce – expensive reduce not suitable for iterative computations
  - Load balancing is harder for maps with polygons and line-strings
  - Spatial Hadoop ?

Relational to Spatial DBMS to SBD Management

- 1980s: Relational DBMS
- Spatial customer (e.g. NASA, USPS) faced challenges
  - Semantic Gap
- New ideas emerged in 1990s
  - Spatial data types and operations
- SBD may require new thinking for
  - Temporally detailed roadmaps
  - Eco-routing queries
  - Privacy vs. Utility Trade-off

Opportunities

- 1990s: Data Mining
  - Scale up traditional models (e.g., Regression) to large relational databases (460 T-bytes)
  - New pattern families: Associations : Which items are bought together? (Ex. Diaper, beer)

- Spatial customers
  - Walmart: Which items are bought just before/after events, e.g. hurricanes?
    - Where is a pattern (e.g., (diaper-beer) prevalent?
  - Global climate change

Challenges & Questions

Challenges

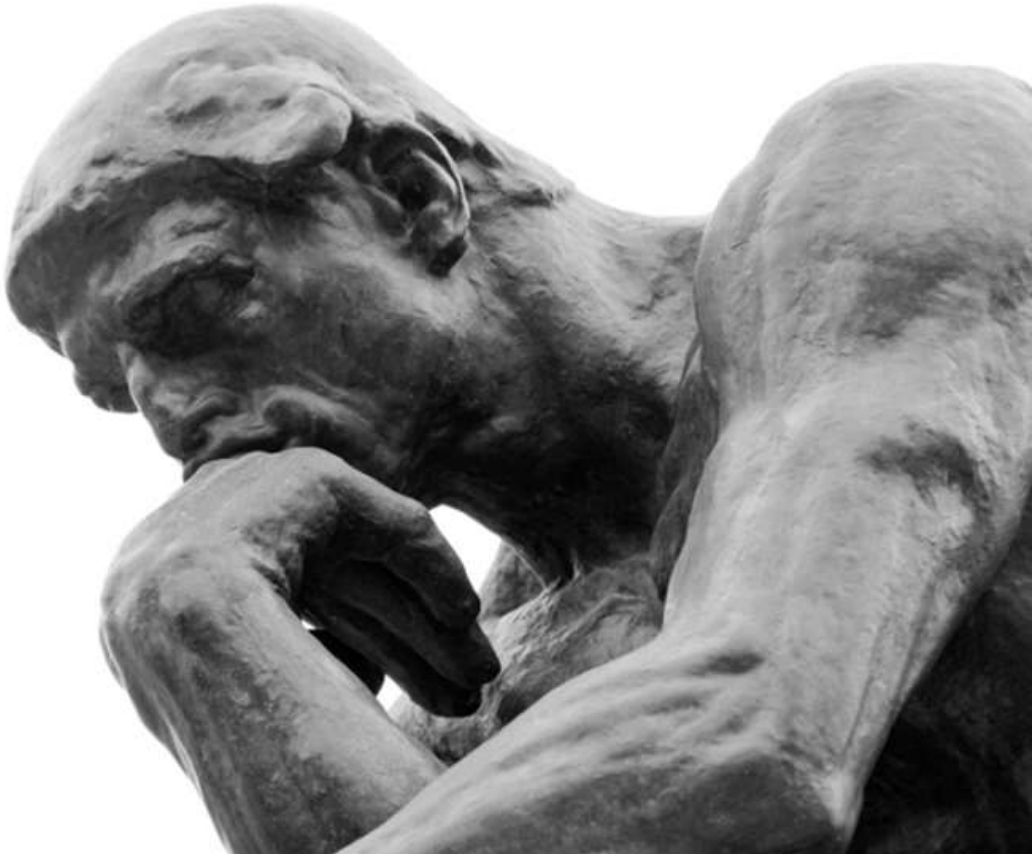
- Independent Identical items
- Distribution assumption not reasonable for spatial data
- Transactions, i.e. disjoint partitioning of data, not natural for continuous space
- This led to Spatial Data Mining (last decade)

Questions

- Does SBD facilitate better spatial models?
- (When) Does bigger spatial data lead to simpler models, e.g. database as a model ?
- On-line Spatio-temporal Data Analytics

What are we looking forward to?

- Case 1: Compute Spatial-Autocorrelation Simpler to Parallelize
  - Map-reduce is okay
  - Should it provide spatial de-clustering services?
  - Can query-compiler generate map-reduce parallel code?
- Case 2: Harder : Parallelize Range Query on Polygon Maps
  - Need dynamic load balancing beyond map-reduce
  - MPI or OpenMP is better!
- Case 3: Estimate Spatial Auto-Regression Parameters, Routing
  - Map-reduce is inefficient for iterative computations due to expensive "reduce"!
    - Ex. Golden section search, Determinant of large matrix
    - Ex. Eco-routing algorithms, Evacuation route planning
  - Option 1: Develop non-iterative formulations of spatial problems
  - Option 2: Alternative Platform: MPI, OpenMP, Pregel or Spatial Hadoop



Questions???