Spatial Data Science and Transportation

Shashi Shekhar
CTS Scholar & McKnight Distinguished University Professor
Dept. of Computer Sc. and Eng., University of Minnesota
shekhar@umn.edu

Acknowledgement: Slides prepared by Xun Tang, Yan Li. This material is based upon work supported by the National Science Foundation, the USDOD, the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, the NIH, and the UMN Center for Transportation Studies.
A Spatial Data Science Story

1854: What causes Cholera?

Collect & Curate Data

Discover Patterns, Generate Hypothesis

Test Hypothesis (Experiments)

Develop Theory

Impact:
sewage system, drinking water supply …

Q? What are Choleras of today?
Q? How may Spatial Data Mining Help?

What is new since Snow’s map? Spatial Big Data

- 1980s: USDOD opens GPS for civilian use
  - 1990s: use in Intelligent Transportation Systems
- Today: **2 billion** GPS receivers in use (7 billion by 2022).
  - Many share location every second
  - Generating a large volume of location traces

- GPS also provides **reference time** for many infrastructure
  - Airlines, Telecommunications, Banks
- GPS is the single point of failure for the entire modern economy.
- 50,000 incidents of deliberate (GPS) jamming last two years
  - Against Ubers, Waymo’s self-driving cars, delivery drones from Amazon

**Bloomberg Businessweek**
July 25, 2018, 4:00 AM CDT

The World Economy Runs on GPS. It Needs a Backup Plan

Large Constellations of Small Satellites

- Hi-frequency (e.g., daily or hourly) time-series of imagery of entire earth
  - Monitor illegal fishing, forest fires, crops (2017 DARPA Geospatial Cloud Analytics)
- Large Constellations
  - 2017: Planet Labs: 100 satellites: daily scan of Earth at 1m resolution in visible band

Cheap (or free) satellite data on cloud computers

- 2008: USGS gave away 35-year LandSat satellite imagery archive
  - Analog of public availability of GPS signal in late 1980s
- 2017: Many cloud-based Virtual collaboration environment
  - Explosion in machine learning on satelliite imagery to map crops, water, buildings, roads, …

<table>
<thead>
<tr>
<th>Elevation, Landsat, LOCA, MODIS, NAIP</th>
<th>Google Earth Engines</th>
<th>NEX</th>
<th>AWS Earth</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOAA</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>AVHRR, FIA, GIMMM, GlobCover, NARR, TRIMM, Sentinel-1</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>IARPA, GDELT, MOGREPS, OpenStreetMap, Sentinel-2, SpaceNet (building/road labels for ML)</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>CHIRPS, GeoScience Australia, GSMap, NASS, Oxford Map, PSDI, WHRC, WorldClim, WorldPop, WWF, BCCA, FLUXNET</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
Spatial Big Data has Big Value

New Ways to Exploit Raw Data May Bring Surge of Innovation, a Study Says  (May 13, 2011)

The study estimates that the use of personal location data could save consumers worldwide more than $600 billion annually by 2020. Computers determine users’ whereabouts by tracking their mobile devices, like cellphones. The study cites smartphone location services including Foursquare and Loopt, for locating friends, and ones for finding nearby stores and restaurants.

But the biggest single consumer benefit, the study says, is going to come from time and fuel savings from location-based services — tapping into real-time traffic and weather data — that help drivers avoid congestion and suggest alternative routes. The location tracking, McKinsey says, will work either from drivers’ mobile phones or GPS systems in cars.

U.P.S. Embraces High-Tech Delivery Methods (July 12, 2007)

By “The research at U.P.S. is paying off. …….— saving roughly three million gallons of fuel in good part by mapping routes that minimize left turns.”
Spatial Big Data is transforming our Society!
## A few Questions in Transportation Domain

<table>
<thead>
<tr>
<th>Role</th>
<th>Questions</th>
<th>Pattern Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traveler, Commuter</td>
<td>What will be the travel time on a route?</td>
<td>Prediction</td>
</tr>
<tr>
<td>Transportation manager</td>
<td>Which corridors are accident-prone?</td>
<td>Hotspot</td>
</tr>
<tr>
<td></td>
<td>Where and when are traffic flow anomalies?</td>
<td>Spatial Outlier</td>
</tr>
<tr>
<td>Traffic engineering</td>
<td>Which loop detector stations are very different from their neighbors?</td>
<td>Spatial Outlier</td>
</tr>
<tr>
<td></td>
<td>Where are the congestion (in time and space)?</td>
<td>Hotspot</td>
</tr>
<tr>
<td>Planner and researchers</td>
<td>What will be travel demand in future?</td>
<td>Prediction</td>
</tr>
<tr>
<td></td>
<td>How many trucks are there in a parking lot?</td>
<td>Prediction</td>
</tr>
<tr>
<td></td>
<td>What road types are co-located? Where are they?</td>
<td>Co-location</td>
</tr>
<tr>
<td>Vehicle engineers</td>
<td>Which locations have high NOx emission?</td>
<td>Hotspot, Co-location</td>
</tr>
</tbody>
</table>
Spatial Data Mining

• Challenge:
  – (Data Volume) >> (Number of Human Analysts)
  – Need automated methods
  – Need tools to amplify human capabilities

• Spatial Data are ubiquitous & important

• Current Data Science Tools are inadequate
  – Gerrymandering, Spatial Auto-correlation, …

• Practitioners in fields including:
  – Transportation, agriculture, weather, environment, …

Details: A UCGIS Call to Action: Bringing the Geospatial Perspective to Data Science Degrees and Curricula.
Defining Spatial Data Mining

• The process of discovering
  • interesting, useful, non-trivial patterns
    • patterns: non-specialist
    • exception to patterns: specialist
  • from large spatial datasets

• Spatial pattern families
  A. Hotspots, Spatial clusters
  B. Spatial outlier, discontinuities
  C. Co-locations, co-occurrences
  D. Spatial classification, prediction
  E. Object detection
  F. …


A. Hotspots, Spatial clusters

- **Question:** Which corridors are accident-prone?

- **Data:**
  - 43 Pedestrian fatalities in Orlando, FL (2000-9)
  - USDOT Fatality Analysis Reporting System
    https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

- **Patterns:**
  - Circular results from SaTScan
  - Linear hotspots

- **Interpretation:**
  - Unsafe pedestrian walkway

Report shows that pedestrian safety is a major concern on Minnesota's American Indian reservations

More residents get around on foot, often on well-traveled roads

By Kelly Smith | FEBRUARY 18, 2019 — 5:25PM


https://www.completecommunitiesde.org/planning/complete-streets/winter-maintenance-2/
A. Hotspots, Spatial clusters: Case Study on Hennepin County Crashes

**Question:** Which corridors are accident-prone?

**Data:**
- 1345 crashes on Hennepin County road intersections (2010 - 2015)
- Source: Hennepin County Public Works

Data Source: https://www.hennepin.us/business/work-with-henn-co/transportation-planning-design
A. Hotspots, Spatial clusters: Case Study on Hennepin County Crashes

- **Data:**
  - 1345 crashes on Hennepin County major intersections (2010-2015)
  - Source: Hennepin County PWD

- **Patterns:**
  - Linear hotspots (p-value = 0.05)
    - Minimum length: 500 meters
    - No turns over 45 degrees in the path (constrained on single street)

- **Interpretation:**
  - Intersections to corridors
  - Feasibility study

- **Next:**
  - Include other roads
  - Consider traffic volume

Data Source: https://www.hennepin.us/business/work-with-henn-co/transportation-planning-design
Dot sizes fool human eye but not algorithms

dot size = 0.25

dot size = 0.5

dot size = 1

dot size = 2
B. Spatial outlier, Discontinuities

- **Question:** Which loop detector stations are very different from their neighbors?

- **Data:**
  - 900 stations (with 1 to 4 loop detectors each).

- **Pattern:**
  - Spatial outlier at Station 9.

- **Interpretation:**
  - Hypothesis: faulty loop detector?
  - Action: Test station 8 detectors

Discovering Sub-time-series Co-occurrence Patterns of Non-compliance

**Given:**
- A set of multivariate event trajectories and a set of non-compliant windows
- A cross-k function threshold $\varepsilon$
- A time lag $\delta$
- A minimum support threshold $\text{minsupp}$

**Find:**
- Co-occurrence patterns whose cross-K function at distance $\delta$ exceeds $\varepsilon$ and whose support exceed $\text{minsupp}$

C. Hotspots, Co-locations, Co-occurrences

**Question:** Where are high transit-NOx emissions? What is co-located there?

**Data:**
- On Board Diagnostics Data from Metro-Transit Buses

Variables sampled every second:
- GPS location
- Speed
- Vehicle Load
- Engine and Heater Fuel Flow
- Exhaust Temp and Mass Flow
- Intake Temp And Mass Flow
- Engine Torque and RPM
- Engine Coolant Temp
- Odometer
- NOx emission
- …measurements on 200+ variables

C. Emission Hotspots, Co-locations

Red color: NO\textsubscript{x} emission exceeds EPA regulations

Legend: gNO\textsubscript{x}/m

0.016

0.000

C. Co-locations, Co-occurrences

Case Study: Test feasibility of road use charging system

- **Use Case:** Impact of EV on Gas Tax:
  - Test technology for road-type based road-usage based charging.

- **Q?** Can GPS distinguish road-types?
  - Which road types are closely co-located? Where?

- **Input:** Road map with road-types
- **Pattern:** Co-location of road-types
D. Spatial Classification, Prediction

- **Question:** Are there natural groups for UPS delivery trajectories?
- **Data:** A set of historical trajectories with on-board diagnostic data from UPS trucks.
- **Pattern:** Clusters of trajectories with similar spatial properties.
- **Interpretation:** Delivery zones are small, but the distance between each delivery zone and UPS depots is different.

E. Geospatial Object Detection

Q: ? How many trucks are there in a lot? City?

Ex.: Estimate truck supply in a city (CH Robinson)

Data:

- Aerial imagery (3 inch pixels)
  - Hennepin & Ramsey counties
- NAIP Imagery (1 meter pixels, 2017)

Pattern:

- Detected geospatial objects
  - Cars, trucks,
  - Houses, …


University of Minnesota
Driven to Discover®
Data science promises new insights, helping transform information into knowledge that can drive science and industry.

BY FRANCINE BERMAN, ROB RUTENBAR, BRENT HAILPERN, HENRIK CHRISTENSEN, SUSAN DAVIDSON, DEBORAH ESTRIN, MICHAEL FRANKLIN, MARGARET MARTONOSSI, PADMA RAGHAVAN, VICTORIA STODDEN, AND ALEXANDER S. SZALAY

Realizing the Potential of Data Science

Berman F. et al.,
Realizing the Potential of Data Science,
Communications of the ACM,
April 2018, Vol. 61 No. 4, pp. 67-72,
10.1145/3188721
Teaching Data Science: Many Flowers Blooming

- **University of California, Berkeley:**
  - Recently established division of data science (same level as college and school)
  - Opened Introductory, foundational, and advanced courses.
  - **Undergraduate** program in Data Science

- **University of Michigan, Ann Arbor:**
  - **Undergraduate** program in Data Science

- **Columbia University:**
  - **Master** of Data Science offered by Data Science Institute

- **University of Illinois, Urbana-Champaign:**
  - **Master** of Computer Science in Data Science offered as an online professional course

- **University of Chicago:**
  - **Master** of Science in Computational Analysis and Public Policy program

Berman F. et al.,
*Realizing the Potential of Data Science, Communications of the ACM*,
April 2018, Vol. 61 No. 4, pp. 67-72, 10.1145/3188721
Data Life Cycle

The data life cycle and surrounding data ecosystem from the *Realizing the Potential of Data Science Report.*

{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

**Acquire**
- Create, capture, gather from:
  - Lab
  - Fieldwork
  - Surveys
  - Devices
  - Simulations
  - More

**Clean**
- Organize
- Filter
- Annotate
- Clean

**Use/Reuse**
- Analyze
- Mine
- Model
- Derive much more additional data
- Visualize
- Decide
- Act
- Drive:
  - Devices
  - Instruments
  - Computers

**Publish**
- Share:
  - Data
  - Code
  - Workflows
  - Disseminate
  - Aggregate
  - Collect
  - Create portals, databases, and more
  - Couple with literature

**Preserve/Destroy**
- Store to:
  - Preserve
  - Replicate
  - Ignore
  - Subset, compress
  - Index
  - Curate
  - Destroy

Berman F. et al., *Realizing the Potential of Data Science, Communications of the ACM*, April 2018, Vol. 61 No. 4, pp. 67-72, 10.1145/3188721
Data Science Skills

The data life cycle
{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

Acquire
- Survey
- Sensor
- Citizen Science

Clean
- Filter
- Annotation

Judicious use/reuse
- Coding
- Querying
- Machine learning
- **Data mining**
- Statistics
- Optimization
- Visualization
- Spatial data analysis
- Interpretation
- Decision Making

Publish
- Portal
- Share

Preserve & Destroy
- Curation
- Indexing
## Data Science Tools

<table>
<thead>
<tr>
<th>Skills</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>• Python</td>
</tr>
<tr>
<td></td>
<td>• Matlab</td>
</tr>
<tr>
<td>Querying</td>
<td>• SQL</td>
</tr>
<tr>
<td></td>
<td>• Hive</td>
</tr>
<tr>
<td>Machine learning</td>
<td>• Scikit-learn</td>
</tr>
<tr>
<td></td>
<td>• Tensorflow</td>
</tr>
<tr>
<td></td>
<td>• MLlib for Spark</td>
</tr>
<tr>
<td>Data mining</td>
<td>• Rapid miner</td>
</tr>
<tr>
<td></td>
<td>• Oracle data mining</td>
</tr>
<tr>
<td></td>
<td>• Weka</td>
</tr>
<tr>
<td>Statistics</td>
<td>• R</td>
</tr>
<tr>
<td></td>
<td>• SAS</td>
</tr>
<tr>
<td>Optimization</td>
<td>• Cplex</td>
</tr>
<tr>
<td></td>
<td>• GAMS</td>
</tr>
<tr>
<td></td>
<td>• GUrobi</td>
</tr>
<tr>
<td>Spatial data analysis</td>
<td>• ArcGIS</td>
</tr>
<tr>
<td></td>
<td>• QGIS</td>
</tr>
<tr>
<td></td>
<td>• SaTScan</td>
</tr>
</tbody>
</table>
# Education in Data Science - UMN

<table>
<thead>
<tr>
<th>Name of Degrees</th>
<th>Focused skills</th>
<th>Name of Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bachelor</strong></td>
<td>Coming soon</td>
<td></td>
</tr>
<tr>
<td><strong>Certificate (12 credits)</strong></td>
<td>Post-Baccalaureate Certificate in Data Science</td>
<td>Coding, Querying, Machine learning, Data mining</td>
</tr>
<tr>
<td><strong>Master (31 credits)</strong></td>
<td>Master’s of Science in Data Science</td>
<td>Interpretation, Decision making</td>
</tr>
<tr>
<td></td>
<td>Master of Science in Business Analytics</td>
<td>Optimization, Decision making</td>
</tr>
<tr>
<td></td>
<td>M.S. in Industrial and Systems Engineering - Analytics Track</td>
<td></td>
</tr>
</tbody>
</table>
References

- Berman F. et al., Realizing the Potential of Data Science, Communications of the ACM, April 2018, Vol. 61 No. 4, pp. 67-72, 10.1145/3188721
Spatial Big Data driven Eco-Routing

Spatially oriented datasets exceeding capacity of current routing systems

- Due to Volume, Velocity (Update-rate) and, Variety

Waze.com