Data Science for Cities

Juliana Freire

Visualization and Data Analysis (ViDA) Lab
Computer Science & Engineering
Center for Urban Science & Progress (CUSP)
Center for Data Science
New York University
World Urbanization Prospects

United Nations, Department of Economic and Social Affairs, Population Division: World Urbanization Prospects, the 2011 Revision
Understanding Cities

City = Infrastructure + Environment + People

Infrastructure:
Condition, operations

Environment:
Meteorology, pollution, noise, flora, fauna

People:
Relationships, economic activities, health, nutrition, opinions, …

Open urban data exhaust

City code

NYU POLYTECHNIC SCHOOL OF ENGINEERING

San Francisco Data

City of Chicago Data Portal

OpenData

London.gov.uk

Startseite Portal der Stadt Zürich

Armazém Dados

RIO Prefeitura
Urban Data: Success Stories

- Developed at the University of Washington
- Real-time arrival predictions
- 94% reported increased or greatly increased satisfaction with public transit
- Significant decrease in actual wait time per user, and an even greater decrease in perceived wait time
- 78% of riders reported increased walking --- a significant public health benefit

http://onebusaway.org
New York City gets 25,000 illegal-conversion complaints a year, but it has only 200 inspectors to handle them.

Flowers’ group integrated information from 19 different agencies that provided indication of issues in buildings.

Result: hit rate for inspections went from 13% to 70%
Urban Data: Success Stories

- The NYU Furman Center
- Analysis of the impact and benefits of subsidized housing on the surrounding neighborhoods → influenced city spending decisions
- Assessment of crime data and property-level foreclosure data led to the finding that neighborhoods with concentrated foreclosures see an uptick in crime for each foreclosure notice issued → updates to policing strategies

http://furmancenter.org/
Our Vision: Data Science for Cities

• Democratize information technology in urban governance and management
  • Data-intensive research and practice must become widely adopted and applied by social science researchers, policy makers and urban residents
  • An enormous fraction of the population is interested, invested, and capable of participating in data-driven decision making

• Challenges
  • Substantial effort is required to gather, integrate, and analyze relevant data
  • Complex interactions among multiple components are hard to model
  • Centrally-controlled smart cities efforts are insufficient
  • Projects often constrained to well-defined questions and confirmatory analyses over few data slices
  • Need to significantly increase the level of automation, interactivity, scalability, and usability of the tools
Big Urban Data: Taxis as Sensors for NYC

- NYC taxis as sensors for can city life: economic activity, human behavior, mobility patterns, …
- Taxi data are “big”, complex and dirty
  - ~500k trips/day – 0.5 billion trips in 3 years
  - Multiple variables: spatial temporal + trip attributes
- Domain scientists and decision makers are unable to explore the whole data

Number of Trips for the years of 2011, and 2012
A Study of NYC Taxis: Data Preparation

<table>
<thead>
<tr>
<th></th>
<th>SQLite</th>
<th>Postgre SQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage Space in GB</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Building Indices in Minutes (One Year)</td>
<td>3,120</td>
<td>780</td>
</tr>
</tbody>
</table>

- Spatio-temporal index based on out-of-core kd-tree [Ferreira et al., TVCG 2013]
- Deployed at TLC and DoT!
- New index that leverages GPU – 2 orders of magnitude speedup [Vo and Doraiswami, in progress]

**Fig. 11.** Scalability of multiple GPUs for out-of-core queries. We obtain close to 60-fold speedup using three GPUs.

```javascript
db.trips.find({
    $or: [
        {pickup:
            { $gt: start1, $lt: end1 }},
        {pickup:
            { $gt: start2, $lt: end2 }},
        {pickup:
            { $gt: start3, $lt: end3 }},
        {pickup:
            { $gt: start4, $lt: end4 }}
    ],
    pickup:
        { $geoWithin: { $polygon: [ [mx1, my1], ..., [mxn, myn] ] } },
    $dropoff:
        { $geoWithin: { $polygon: [ [jx1, jy1], ..., [jxk, jyk], [lx1, ly1], ..., [lxp, lyp] ] } }
});
```

**Fig. 13.** MongoDB query to find all trips from lower Manhattan to JFK and LGA airports during all Sundays of May 2011. In this query, the time intervals corresponding to the four Sundays are specified using the $or identifier. The taxi data stores the time using the Unix time format. So it is represented as a number. The polygon corresponding to lower Manhattan is given by the vertices {(mx1, my1), ..., (mxn, myn)}. The polygons corresponding to JFK and LGA airports are specified by the vertices {(jx1, jy1), ..., (jxk, jyk)} and {(lx1, ly1), ..., (lxp, lyp)}, respectively.

**Fig. 14.** Scalability of multiple GPUs for queries on the 64-bit index. We obtain speedup of around 30 times using three GPUs.

<table>
<thead>
<tr>
<th>Query</th>
<th>MongoDB (1 GPU)</th>
<th>MongoDB (3 GPUs)</th>
<th>PostgreSQL (1 GPU)</th>
<th>PostgreSQL Speedup (3 GPUs)</th>
<th>ComDB Speedup (1 GPU)</th>
<th>ComDB Speedup (3 GPUs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time (sec)</td>
<td>Time (sec)</td>
<td>Speedup</td>
<td></td>
<td>Speedup</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.237</td>
<td>0.103</td>
<td>141.8</td>
<td>598</td>
<td>1376</td>
<td>1329</td>
</tr>
<tr>
<td>2</td>
<td>0.199</td>
<td>0.065</td>
<td>129.2</td>
<td>649</td>
<td>1987</td>
<td>119.6</td>
</tr>
<tr>
<td>3</td>
<td>0.202</td>
<td>0.093</td>
<td>97.1</td>
<td>480</td>
<td>1044</td>
<td>39.4</td>
</tr>
<tr>
<td>4</td>
<td>0.183</td>
<td>0.069</td>
<td>103.7</td>
<td>566</td>
<td>1502</td>
<td>25.6</td>
</tr>
<tr>
<td>5</td>
<td>0.361</td>
<td>0.159</td>
<td>106.3</td>
<td>294</td>
<td>668</td>
<td>23.8</td>
</tr>
<tr>
<td>6</td>
<td>0.325</td>
<td>0.174</td>
<td>102.6</td>
<td>315</td>
<td>589</td>
<td>28.9</td>
</tr>
</tbody>
</table>

**TABLE I.** Comparison of query execution times of spatio-temporal queries on different database systems. Using the KD-tree index on MongoDB, we obtain at least two orders of magnitude speedup with one GPU. Queries 5 and 6 do not have a constraint on the dropoff location. Note that with increasing constraints, the performance of the KD-tree index improves, and we obtain up to three orders of magnitude speedup.

- Spatio-temporal index based on out-of-core kd-tree [Ferreira et al., TVCG 2013]
- Deployed at TLC and DoT!
- New index that leverages GPU – 2 orders of magnitude speedup [Vo and Doraiswami, in progress]
A Study of NYC Taxis: Data Exploration

SELECT * 
FROM trips 
WHERE pickup_time in (5/1/11,5/7/11) 
  AND 
  dropoff_loc in "Times Square" 
  AND 
  pickup_loc in "Gramercy"

Interactively explore data through the map view and plot widgets

New, scalable, map rendering infrastructure

[Ferreira et al., IEEE TVCG 2013]
A Study of NYC Taxis: Event Detection

Identify and Explore Events (Month of May)

5 Borough Bike Tour 2011 (1 May 2011)

Work in progress
The Road Ahead

- Data-intensive science can have substantial impact on cities
- Need interdisciplinary collaborations
- For the Taxi project: visualization, data management, graphics, computational topology, traffic, economics
- Need usable tools and infrastructure to empower a broad range of data enthusiasts to explore the urban data exhaust

Data Science Methodologies → Transform Discovery

Scientific Areas

Career Paths
Education and Training
Software Tools
Reproducibility and Open Science
Working Spaces and Culture
Evaluation

Data Science
Thanks