Better Data for Better Cities

Brian Riordan - CS Lead

metro.strava.com
Agenda

Introductions: 10:15am
Overview of Strava & Metro: 10:15am – 10:45am
Overview of Product: 10:45am – 11:15am
Removing Noise: 11:15am – 11:30am
Comfort break: 11:30am – 12:00pm
Correlation (Crash Points, Counters, Speed): 12:00pm – 12:20pm
Finding Missing Geo: 12:30pm – 12:40pm
Delta Analysis: 12:40pm – 1:00pm
Discussion around potential urban research uses for Strava data: 1:00pm – 1:30pm
Lunch and networking
What and Who is Strava?
What is Strava?
The social network for cyclists and runners.
What is Strava Metro?

Data-Driven Bike and Pedestrian Planning

Aggregated, anonymized activity data from Strava’s millions of users

Analyze popular or avoided routes, peak commute times, intersection wait times, and origin/destination zones

Processed for compatibility with geographic information system (GIS) environments
Strava Metro’s Mission Statement

To produce state-of-the-art spatial data products and services to make cycling, running, and walking in cities better.
Why Build Strava Metro?

Data-Driven Bicycle and Pedestrian Planning

- Global need for consistent cycling data
- Continues to serve the Strava user
- Further bonds the cycling and pedestrian community
- It’s the right thing to do
How Do We Protect Users’ Privacy?

User privacy as the foremost concern

Focused on streets, not individuals

No way to bring a Metro record back to Strava

Opt out switches on Strava (less than 0.01%)
Just use the heatmap?

Not so fast!

- No temporal scale
- Point Saturation not Use Saturation
- Large cropping of ride start and ends
- Mix of ride types

What is it good for?

- Showing that people ride bikes
- Starting dialogues with city councils
- Keeping track of where you rode this year
- Editing your basemap / finding missing geometry
Data Restrictions & Use

- Contract is built to allow collaboration with the data throughout the county
- Data can be utilized by engineering firms, universities, and other contracted groups
- The data cannot be posted up on a FTP site for public download
- Aggregated results are fine for public display and download
Big Data
Strava By The Numbers

- Over 8 million activities uploaded per week
- Tens of Millions of Active Users
- 1 Trillion+ second-by-second GPS points globally
Metro Products
Strava Metro Data

- **Streets**: Minute-by-minute counts across your entire network
- **Origin / Destination**: Understand activity starting and ending points, by region
- **Intersections**: Activity counts and wait times at every intersection
Email Delivery
Contains key information about the data format, location, and coverage

- Time frames and roll-up
  - Below are the date and hourly ranges for the AM/PM data used when generating this product.
    - On-Season: April - October
      - Early AM hours: 12am – 5:59am --> Labeled as _0
      - AM Peak hours: 5am – 9:59am --> Labeled as _1
      - Late Evening Hours: 8pm – 11:59pm

- FTP file location
  - https://stravametro.exavault.com/files#/%2FNA%2FUS%2FCA%2FLA%2FLACounty

- Data Vis

- User Manual

- Date Range: July 1, 2015 - June 30, 2016
Metro File Structure & Naming Conventions

<table>
<thead>
<tr>
<th>TABLE NAME</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;identifier&gt;_edges</td>
<td>Line</td>
</tr>
<tr>
<td>&lt;identifier&gt;_nodes</td>
<td>Point</td>
</tr>
<tr>
<td>&lt;identifier&gt;_od_polygons</td>
<td>Polygon</td>
</tr>
<tr>
<td>&lt;identifier&gt;<em>&lt;ride/run&gt;<em>rollup_month</em>&lt;year&gt;</em>&lt;month&gt;_weekend</td>
<td>Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;<em>&lt;ride/run&gt;<em>rollup_month</em>&lt;year&gt;</em>&lt;month&gt;</td>
<td>Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;_&lt;ride/run&gt;<em>rollup_season</em>&lt;month&gt;_weekend</td>
<td>Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;_&lt;ride/run&gt;<em>rollup_season</em>&lt;month&gt;</td>
<td>Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;_&lt;ride/run&gt;_rollup_total_weekend</td>
<td>Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;_&lt;ride/run&gt;_rollup_total_weekday</td>
<td>Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;_&lt;ride/run&gt;_rollup_total</td>
<td>Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;<em>edges</em>&lt;ride/run&gt;_data</td>
<td>.csv/.sql/Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;<em>edges_nodes</em>&lt;ride/run&gt;_data</td>
<td>.csv/.sql/Table</td>
</tr>
<tr>
<td>&lt;identifier&gt;_od_polygons_data</td>
<td>.csv/.sql/Table</td>
</tr>
</tbody>
</table>
# Street Level Data

- Minute by minute tabular data of cycling behavior
- Preferred route direction
- Unique bike trips
- Unique user counts
- Trip purpose: commute and recreation
- Time/Speed (seconds)

<table>
<thead>
<tr>
<th>EDGE_ID</th>
<th>ATHCNT_0</th>
<th>RATHCNT_0</th>
<th>ACTCNT_0</th>
<th>RACTCNT_0</th>
<th>TATHCNT_0</th>
<th>TACTCNT_0</th>
<th>ACTTIME_0</th>
<th>RACTTIME_0</th>
<th>CMTCNT_0</th>
<th>RCMTCNT_0</th>
<th>ATHCNT_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>406645</td>
<td>139286</td>
<td>367</td>
<td>8</td>
<td>421</td>
<td>28</td>
<td>373</td>
<td>449</td>
<td>17.45127726...</td>
<td>256</td>
<td>5</td>
<td>219</td>
</tr>
<tr>
<td>406624</td>
<td>139267</td>
<td>363</td>
<td>10</td>
<td>418</td>
<td>25</td>
<td>368</td>
<td>443</td>
<td>3.5172256535</td>
<td>247</td>
<td>8</td>
<td>213</td>
</tr>
<tr>
<td>406647</td>
<td>139290</td>
<td>348</td>
<td>12</td>
<td>392</td>
<td>32</td>
<td>356</td>
<td>424</td>
<td>19.06984566...</td>
<td>247</td>
<td>12</td>
<td>185</td>
</tr>
<tr>
<td>406646</td>
<td>139289</td>
<td>346</td>
<td>13</td>
<td>397</td>
<td>27</td>
<td>353</td>
<td>424</td>
<td>11.1145850376</td>
<td>251</td>
<td>8</td>
<td>189</td>
</tr>
<tr>
<td>396263</td>
<td>127740</td>
<td>345</td>
<td>0</td>
<td>512</td>
<td>0</td>
<td>345</td>
<td>512</td>
<td>15.93356673...</td>
<td>NULL</td>
<td>294</td>
<td>0</td>
</tr>
<tr>
<td>406648</td>
<td>139291</td>
<td>343</td>
<td>36</td>
<td>394</td>
<td>53</td>
<td>348</td>
<td>411</td>
<td>0.9767940987</td>
<td>0.9307139932</td>
<td>241</td>
<td>22</td>
</tr>
<tr>
<td>406633</td>
<td>139276</td>
<td>340</td>
<td>14</td>
<td>394</td>
<td>41</td>
<td>363</td>
<td>435</td>
<td>7.0942751252</td>
<td>6.3904346334</td>
<td>249</td>
<td>27</td>
</tr>
<tr>
<td>403775</td>
<td>136242</td>
<td>334</td>
<td>0</td>
<td>487</td>
<td>0</td>
<td>334</td>
<td>487</td>
<td>12.38249701...</td>
<td>NULL</td>
<td>282</td>
<td>0</td>
</tr>
</tbody>
</table>
Origin/Destination Polygons

- Polygonal starting and ending points of trips
- Reported by the minute
- Trip purpose flag
- Array of intersected polygon IDs
OD Data

Use the starting and ending points to locate key zones

- Select all polygons that start in the morning
  - "HOUR" < 10 and "HOUR" > 4
- Summarize and export
- Join to polygon table
- Display
## Intersections

- Wait times at intersection
- Congregation of users at intersections
- Minute by minute with purpose flag
Demographics Data

- Rolled-up counts of users in the data files
- Breakdown of age and gender from users
Commutes are the #1 requested data feature in Strava Metro. 75% of all Strava users upload Commute data.

Activities in urban areas are commutes 60% - 40% of the time (High of 90% in London).

Commutes and recreation rides in urban areas have very high correlation in route choice.

Commutes > 6 miles have the same street choices as < 6 miles they are just longer trips.
Users tend to not select the commute flag on Strava.com

3 ways to detect commute activities:
Activities tagged as a commute on Strava.com

Activities with key words in the titles “To Work”, “Commute To Store”

Starting and Ending Points more than 1km apart within a distance and time threshold (This can be user defined but we find 30 miles and 90 minutes to be a good top threshold)
Macro & Micro Levels of Spatial and Temporal Details

LA Full Core Routes

Key Commute Routes and Corridors
Working with the Rollups & Data Structures

Refer to the User Guide for naming conventions and field definitions
Metro Matching Process

- Node structure w/ false node
- Overlay points
- Intersection of points and nodes
Working with Joins

Entire system is built on joining the data files back to the Node, Edges, and OD Spatial layers
Working with Edge Data
Working with Nodes Data
Node Data

Use the intersections to locate slow down times

- Join ID --> node_id
- Sort by total % .75
- Any Noise? (5914 seconds....)
  - Remove all values with aths < 5
Node Data

Use the intersections to locate slow down times
Strava Metro: Data Clean-up
Remove the noise!

1. Join the edges total to edges
   a. Symbolize by TATHCNT
   b. Lots to get lost in...

2. Select a subset in a region you're interested in
   a. Export & re-symbolize
Strava Metro: Data Clean-up
What are these crazy values....

1. Calculate Speed
   a. Recalculate length in miles
   b. Create 2 new fields: spd & rspd as double
   c. Open field Calc: length/((activity time in seconds/60)/60)
   d. Beware the divide by zero rules....
   e. Locate and remove erroneous data
Strava Metro: Data Clean-up
What are these crazy values....

1. Hints
   a. Very small segments (<10 meters) /w high counts
   b. Very high speeds with low counts < 5
Helpful Calculations for Field Calculator

Calculate Speed Per Edge from seconds:
\[
\text{length}/((\text{activity time in seconds}/60)/60)
\]

Filter Points:
\[
\text{start\_date\_local} < '2015-09-1' \text{ AND start\_date\_local} > '2015-08-01'
\]

Percent Change:
\[
((\text{TACTCNT\_12\_13}-\text{TACTCNT})/ \text{TACTCNT})*100
\]
Break!
Part 2

Correlation (Crash Points, Counters, Speed): 12:00pm – 12:20pm
Finding Missing Geo: 12:30pm – 12:40pm
Delta Analysis: 12:40pm – 1:00pm
As custom build product it’s designed to be merged with local datasets: traffic, crashes, proposed bike paths, etc.
Strava Metro Correlation to Counting Programs

- Strava Metro’s use and impact is multiplied when the data is used in conjunction with an established counting program
- Counting programs only show saturation at a single point and dilute from there
- Metro shows the rest of the network - it’s like a counter on every corner
Strava Metro Correlation to Counting Programs Cont’
Correlating Strava to Counting Programs Cont’

Using counting programs with the Metro data allows the data to become even more useful. Strava correlation with counting programs is statistically amazing, with r-squared values typically around 0.8.
How far can we push this? --- Total Miles Traveled in SDOT by Bike in 2014:

63,253,198
Bike Count Correlation: Merging data feeds

1. You need specific count data
   a. We will use the mock data to show how to do this
   b. 4 Locations /w 12 months of Counts
1. Locate Street objectIDs for Metro Counts
   a. Zoom to streets with bike count spots
   b. Grab the objectID (x4)
   c. Beware of dual count locations vs one-sided

Bike Count Correlation Cont’
**Bike Count Correlation Cont’**

Grab the month counts per street to Merge back with counter locations

```sql
select edge_id, EXTRACT(YEAR FROM datetime)|| ',' ||EXTRACT(MONTH FROM datetime) g_date,
sum(total_activity_count) acts,
sum(CASE when (EXTRACT(DOW FROM datetime) not in (0,6)) then total_activity_count end) as acts_weekday,
sum(CASE when (EXTRACT(DOW FROM datetime) in (0,6)) then total_activity_count end) as acts_weekend
from la_edges_ride_data
where edge_id in (315937,520926,146497,520706)
group by edge_id, g_date
order by edge_id, g_date
```

<table>
<thead>
<tr>
<th>edge_id</th>
<th>g_date</th>
<th>acts</th>
<th>acts_weekday</th>
<th>acts_weekend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>146497</td>
<td>328</td>
<td>63</td>
<td>65</td>
</tr>
<tr>
<td>2</td>
<td>146497</td>
<td>257</td>
<td>202</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>146497</td>
<td>161</td>
<td>135</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>146497</td>
<td>362</td>
<td>324</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>146497</td>
<td>387</td>
<td>302</td>
<td>85</td>
</tr>
<tr>
<td>6</td>
<td>146497</td>
<td>431</td>
<td>267</td>
<td>44</td>
</tr>
<tr>
<td>7</td>
<td>146497</td>
<td>187</td>
<td>139</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>146497</td>
<td>243</td>
<td>186</td>
<td>57</td>
</tr>
<tr>
<td>9</td>
<td>146497</td>
<td>296</td>
<td>244</td>
<td>52</td>
</tr>
</tbody>
</table>
Bike Count Correlation Cont’

Create a regression table to understand R2 and % Correlation

- This is typically done in a spreadsheet program like excel
- The key here is to end up with you site specific multiplier for extrapolation
Bike Count Correlation Cont’

Match this back to the street network via AOI Polygons
Bike Count Correlation Cont’

- Create 4 Extrapolation Zones
- Create new Extra Field
- Calculate Total Counts
- Display via jenks
- Create Summed Values
  - Total Miles
  - Total Time
  - Carbon offset?
Bike Count Correlation Cont’
Bike Count Correlation Cont’

Total Miles: 100,872,889
Basemaps

Important to think about multi-modal transportation

- Should include all:
  - Streets
  - Roads
  - Trails
  - Paths

- And should break at all intersections (decision points)

Strava Heatmap, Delaware County, Indiana
Basemaps

Important to think about multi-modal transportation

- Should include all:
  - Streets
  - Roads
  - Trails
  - Paths

- And should break at all intersections (decision points)

Missing trails file in the Delaware County basemap
Basemaps

Important to think about multi-modal transportation

- Should include all:
  - Streets
  - Roads
  - Trails
  - Paths

- And should break at all intersections (decision points)
Misaligned basemap geometry will cause count discrepancies in the Strava Metro data
Metro provides key insights into how the cycling population is adapting to new cycleways, protected lanes and surging car populations. The left image shows the GPS points before (red) and after (blue) a new section of cycleway was opened. The Metro data on the right shows the actual change in percent with blue losing trips and red gaining trips.
Further exploration of investment in dedicated cycleways.
Changes to cycling infrastructure impact routing throughout the network
Strava Metro: Crash Cluster Analysis

2012 Crash Dataset

Locate Streets touched

ID Trends and focus streets
Strava Metro: Crash Cluster Analysis

Subset of streets /w high Commute counts
Strava Metro: Bike Facility Hot Spots
Deep Dive & Out of the Box Thinking

What happens when we take just the network and look at it with use?

How do we do this spatial match?

Can you find the top points were connectivity challenges are causing issues?
Part 3

Discussion around potential urban research uses for Strava data: 1:00pm – 1:30pm
Afternoon Session - Hands On

1) Removing Noise
2) Correlation (Crash Points, Counters, Speed)
3) Finding Missing Geo
4) Delta Analysis
Relational databases

CREATE TABLE la_edges_ride_data
(  edge_id integer,  year integer,  day integer,  hour integer,  minute integer,  athlete_count integer,  rev_athlete_count integer,  activity_count integer,  rev_activity_count integer,  total_activity_count integer,  activity_time numeric,  rev_activity_time numeric,  commute_count integer)

WITH (OIDS=FALSE);

ALTER TABLE la_edges_ride_data
OWNER TO postgres;
Working with the Raw Data

Copy Data In:
```bash
copy la_edges_ride_data FROM 'C:\LA\la_c_201507_201606_ride_data.csv' DELIMITERS ',' CSV HEADER;
```

Update Table with Datetime Field and Populate:
```sql
ALTER TABLE la_edges_ride_data ADD COLUMN datetime timestamp without time zone;
```

Update Date Field:
```sql
UPDATE la_edges_ride_data set datetime = cast(timestamp '2014-12-31' + interval '1 day' * day + interval '1 hour' * hour + interval '1 minute' * minute as timestamp) where year = 2015;
UPDATE la_edges_ride_data set datetime = cast(timestamp '2015-12-31' + interval '1 day' * day + interval '1 hour' * hour + interval '1 minute' * minute as timestamp) where year = 2016;
```

Indexes:
```sql
CREATE INDEX la_c_ride_data_datetime_idx  ON la_edges_ride_data  USING btree  (datetime);
CREATE INDEX la_c_ride_data_edge_id_idx  ON la_edges_ride_data  USING btree  (edge_id);
CREATE INDEX la_c_ride_data_hour_idx  ON la_edges_ride_data USING btree (hour);
```
Counts by Day

SELECT day, count(*)
FROM la_c_201507_201606_ride_data
GROUP BY day
ORDER BY day
SELECT EXTRACT(dow from datetime) dow, count(*)
FROM la_c_201507_201606_ride_data
GROUP BY EXTRACT(dow from datetime)
ORDER BY dow
SELECT EXTRACT(HOUR from datetime),
SUM(total_activity_count)
FROM la_c_201507_201606_ride_data
WHERE EXTRACT(dow from datetime) in (1,2,3,4,5)
GROUP BY EXTRACT(HOUR from datetime)
ORDER BY EXTRACT(HOUR from datetime)
SELECT EXTRACT(YEAR FROM datetime) || ',' || EXTRACT(MONTH FROM datetime) || ',' || EXTRACT(day FROM datetime) || ',' || EXTRACT(hour FROM datetime) g_date,
    SUM(athlete_count) aths, SUM(activity_count) acts, SUM(rev_activity_count) rev_acts, median(activity_time) time1,
    median(rev_activity_time) time2
FROM la_c_201507_201606_ride_data
WHERE edge_id IN (521911,522002,521913,521914,521915,521918,521921,521961,521971)
    and datetime > '2015-09-07' and datetime < '2015-09-17'
GROUP BY g_date
ORDER BY g_date
Simple SQL with Metro

Use for a Group of Streets
Using Metro to Gain Temporal Movement Trends

**Weekday Commute and Weekend Activities**

- Weekday Commutes
- Weekend Activities

**Time of Day**

**Week to Week Trends**

Explore the two distinct user groups on Strava: weekday commuters and weekend recreation cyclists

**Commutes by Day**

- Commutes

**Yearly Commute Trends**

Explore seasonality patterns - bike to work month, summer lulls, holidays, etc.
Strava Metro: Corridor Analysis
Lower San Gabriel Trail
Strava Metro: Corridor Analysis
Lower San Gabriel Trail

SELECT EXTRACT(YEAR FROM datetime )|| ',' ||EXTRACT(MONTH FROM datetime )|| ',' ||EXTRACT(day FROM datetime ) || ',' ||EXTRACT(hour FROM datetime) || ',' ||EXTRACT(minute FROM datetime) g_date,
SUM(athlete_count) aths, SUM(activity_count) acts, SUM(rev_activity_count) rev_acts, median(activity_time) time1, median(rev_activity_time) time2
FROM la_c_201507_201606_ride_data
WHERE edge_id IN
(521911,522002,521913,521914,521915,521918,521921,521961,521971)
and datetime > '2016-05-08' and datetime < '2016-05-14'
GROUP BY g_date
ORDER BY g_date
Strava Metro: Corridor Analysis
Lower San Gabriel Trail

Minute by Minute Activity Counts May 8 - May 14, 2016

Activities
Strava Metro: Corridor Analysis
Lower San Gabriel Trail
Strava Metro: Corridor Analysis

Lower San Gabriel Trail

Average speed ranges from 12 - 18 mph