The Internet of Things

Aaron Hart
Dr. Rosaria Silipo
Phil Winters

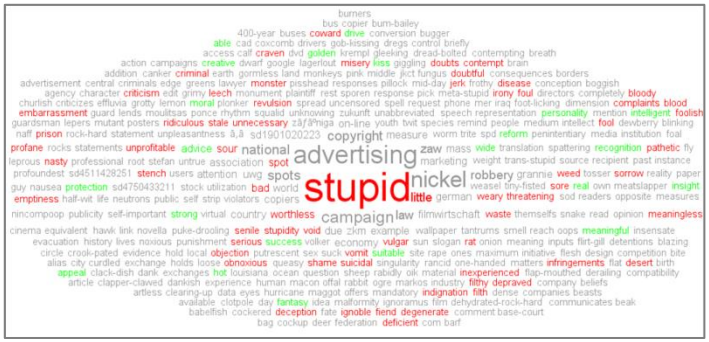
At the 2012 KNIME UGM……

Dr. Killian Thiel
Dr. Tobias Kötter

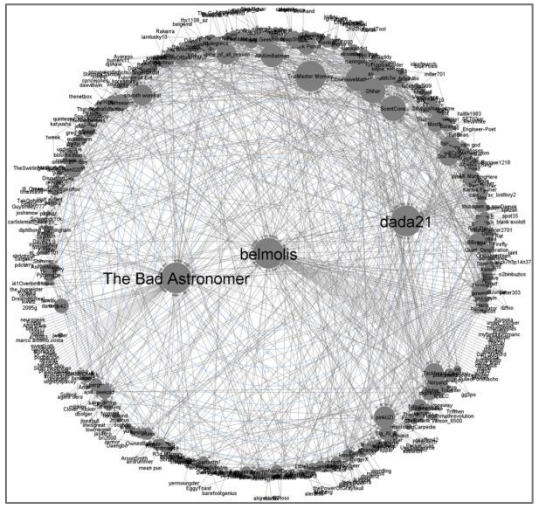
TEXT MINING MEETS NETWORK MINING

White Papers and complete workflows available on the KNIME Public Server!

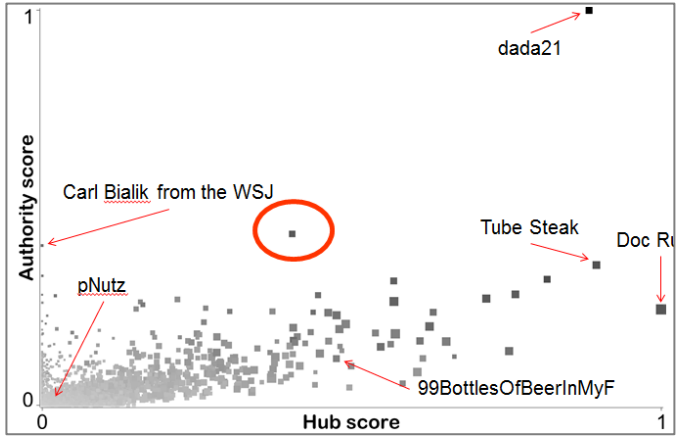
Text Mining for Sentiment



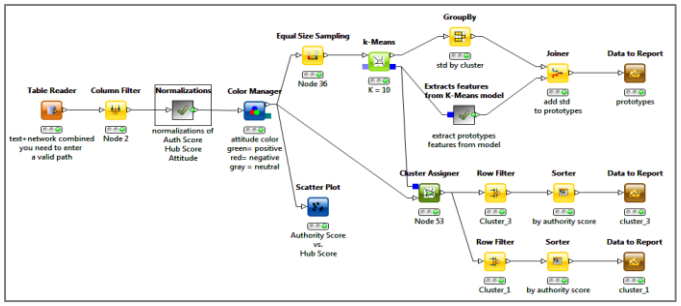
Network Mining for Relevance



Drill Down on special cases



Analytics for Prediction



Last year….

Time Series + Machine Learning + Big Data

Dr. Rosaria Silipo

Manufacturing
Chemical
Life Science
Transportation
Utilities
Automotive
Cyber Security

The Irish Energy Trials

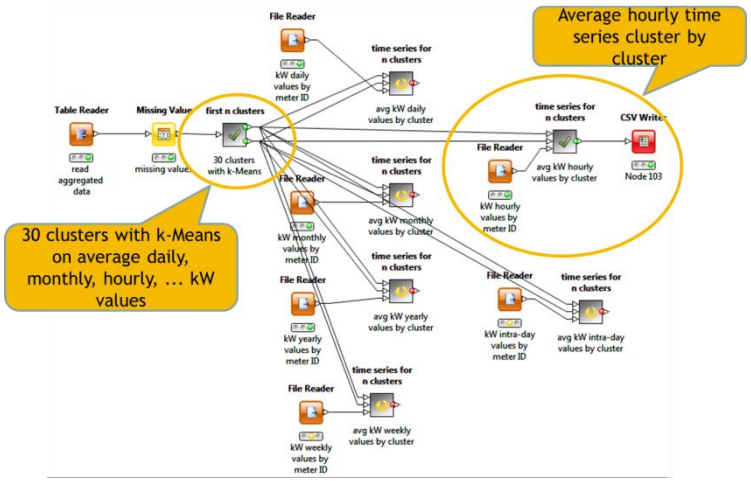
White Paper and Complete Workflows Available !

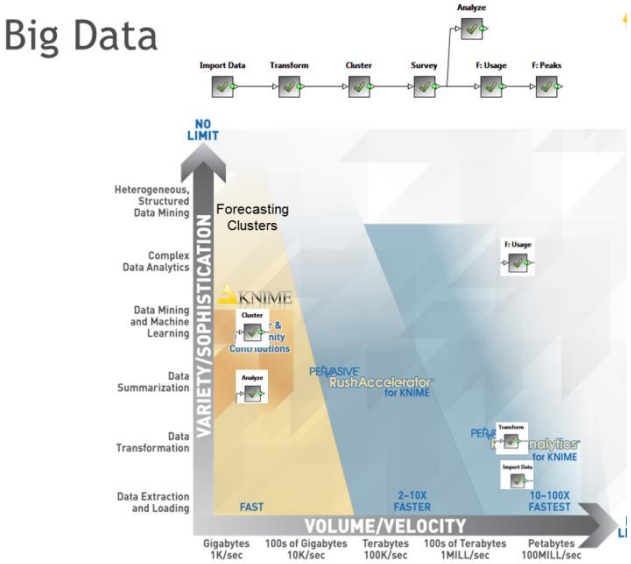
- Telemetry Data
- Time Series Analysis with clustering
- Measurable / Applied to the Business
- SENSIBLE usages of Big Data

Cluster 16: **Let's do Lunch…..**

	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Avg daily KW	Avg hourly kW	size
Cluster 23	16%	18%	18%	17%	16%	**8%**	**6%**	81	3.4	153
Cluster 29	12%	16%	16%	17%	17%	16%	**4%**	80	3.3	93
Cluster 16	18%	21%	21%	19%	17%	**5%**	**3%**	55	2.3	71
Cluster 11	15%	17%	17%	16%	16%	**11%**	**9%**	102	4.3	127
Cluster 25	**2%**	**2%**	**2%**	**25%**	**34%**	**65%**	**58%**	21	0.9	5

	Avg daily kW	Avg hourly kW	21-07	07-09	09-13	13-17	17-21	size
Cluster 21	26	1.1	19%	5%	**33%**	21%	21%	81
Cluster 23	81	3.4	19%	5%	**30%**	**30%**	14%	153
Cluster 29	80	3.3	8%	2%	**33%**	**39%**	17%	93
Cluster 16	55	2.3	**7%**	4%	**40%**	**38%**	11%	71





The Internet of Things

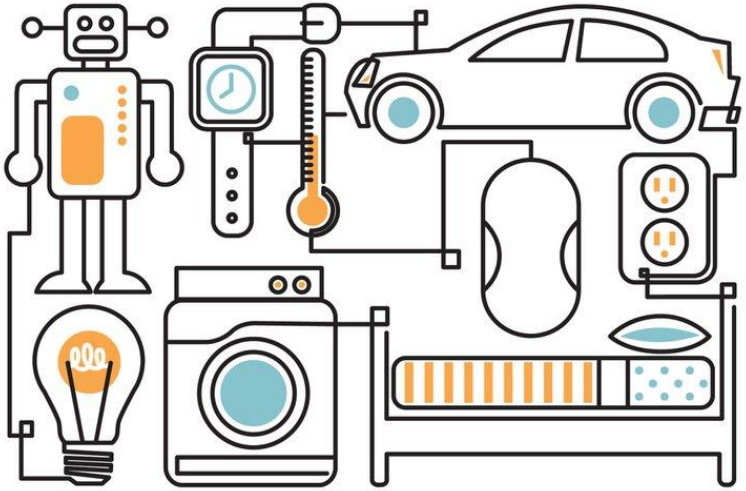
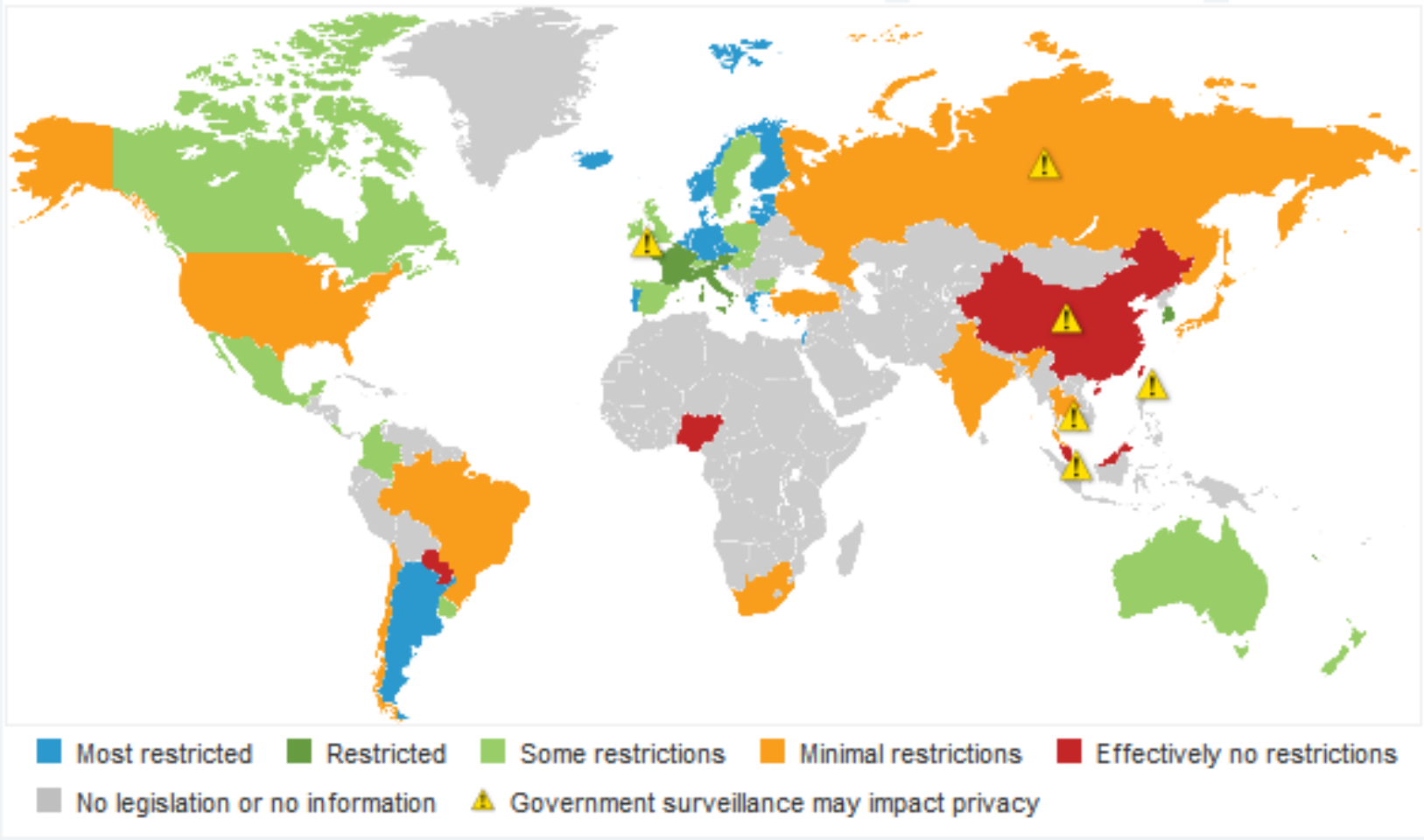


Illustration by CRISTINA BYVIK

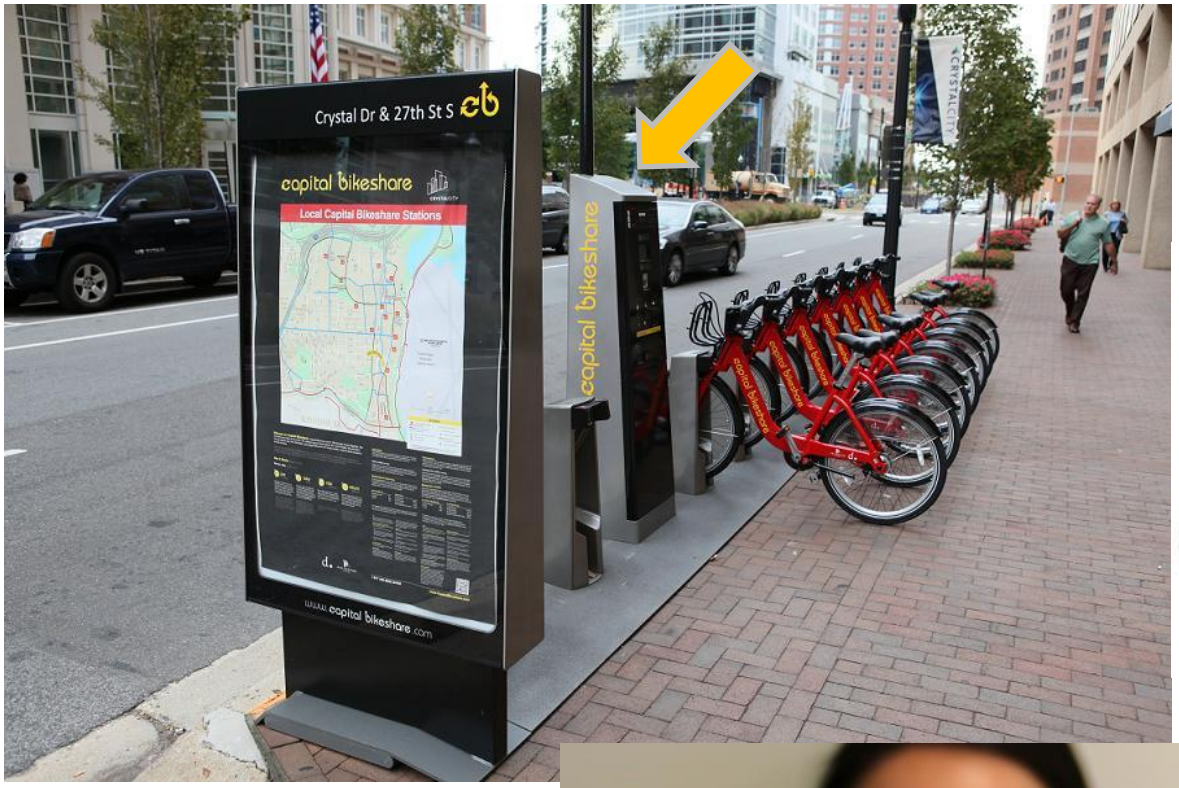
Use Public Data Please….



Original:

FORRESTER
PRIVACY AND DATA PROTECTION BY COUNTRY
For Security & Risk Professionals

Washington DC


Crystal Dr & 27th St S
capital bikeshare
Local Capital Bikeshare Stations
capital bikeshare
www.capitalbikeshare.com


capital bikeshare


Get keyed into
capital
bikeshare™

Sensors!

capital bikeshare

ABOUT FAQ CONTACT STORE MEMBER LOGIN RENEW JOIN

Home | How It Works | Station Map | Pricing | Explore By Bike | Safety | Partners | News | System Data

join for a day, 3 days, month, or year

take any available bike

ride where and when you want

return your bike at any station

What Is Capital Bikeshare?

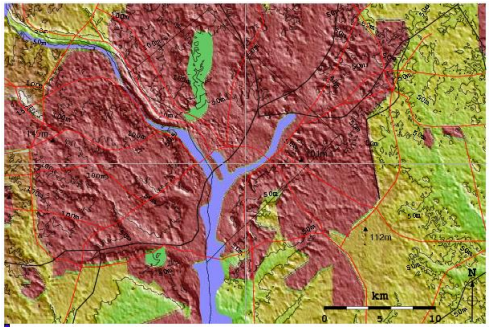
Capital Bikeshare puts over 2500 bicycles at your fingertips. You can choose any of the over 300 stations across Washington, D.C., Arlington and Alexandria, VA and Montgomery County, MD and return it to any station near your destination. Check out a bike for your trip to work, Metro, run errands, go shopping, or visit friends and family. Join Capital Bikeshare for a day, 3 days, a month, a year or try our new daily key option, and have access to our fleet of bikes 24 hours a day, 365 days a year. The first 30 minutes of each trip are free. Each additional 30 minutes incurs an additional fee.

join

for a day, 3 days, month, or year.

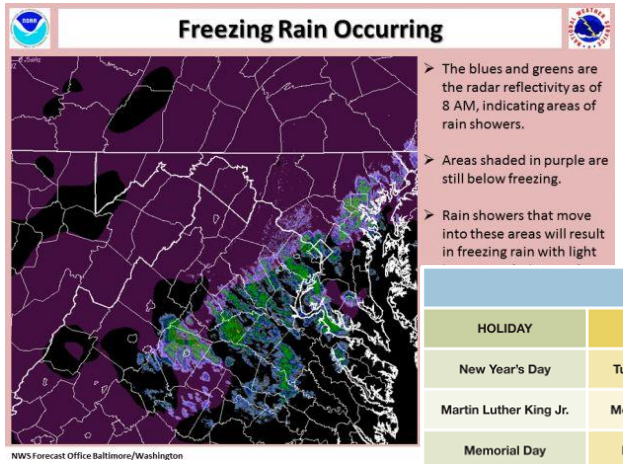
Street Maps

Topology / Elevations



Commuters

Weather



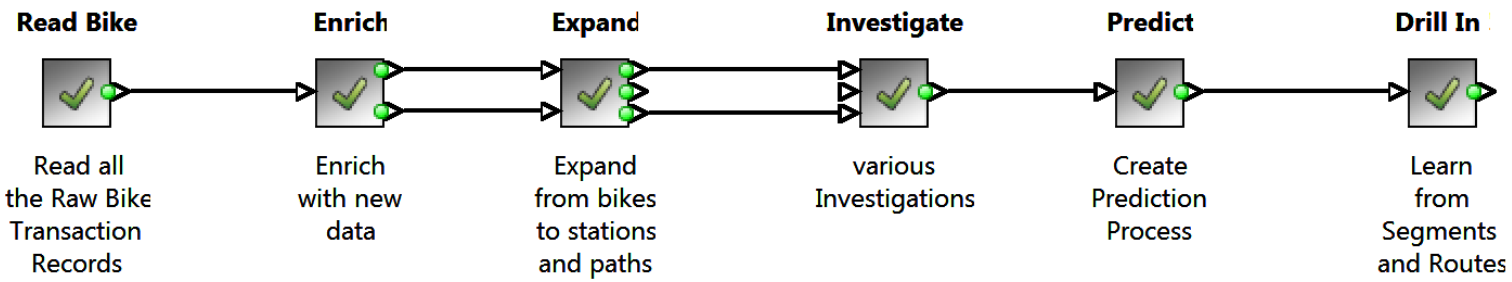
Holiday Schedules

HOLIDAY SCHEDULE

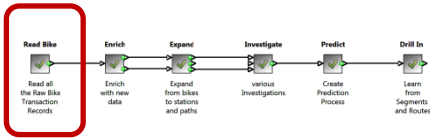
HOLIDAY	2013	2014	2015
New Year's Day	Tuesday, January 1	Wednesday, January 1	Thursday, January 1
Martin Luther King Jr.	Monday, January 21	Monday, January 20	Monday, January 19
Memorial Day	Monday, May 27	Monday, May 26	Monday, May 25
Independence Day	Thursday, July 4	Friday, July 4	Friday, July 3
Labor Day	Monday, September 2	Monday, September 1	Monday, September 7
Thanksgiving	Thursday, November 28 Friday, November 29	Thursday, November 27 Friday, November 28	Thursday, November 26 Friday, November 27
Christmas	Tuesday, December 24 Wednesday, December 25	Wednesday, December 24 Thursday, December 25	Thursday, December 24 Friday, December 25
New Year's Eve	Tuesday, December 31	Wednesday, December 31	Thursday, December 31

Tourists

KNIME and the Internet of Things



Reading the Sensor Data



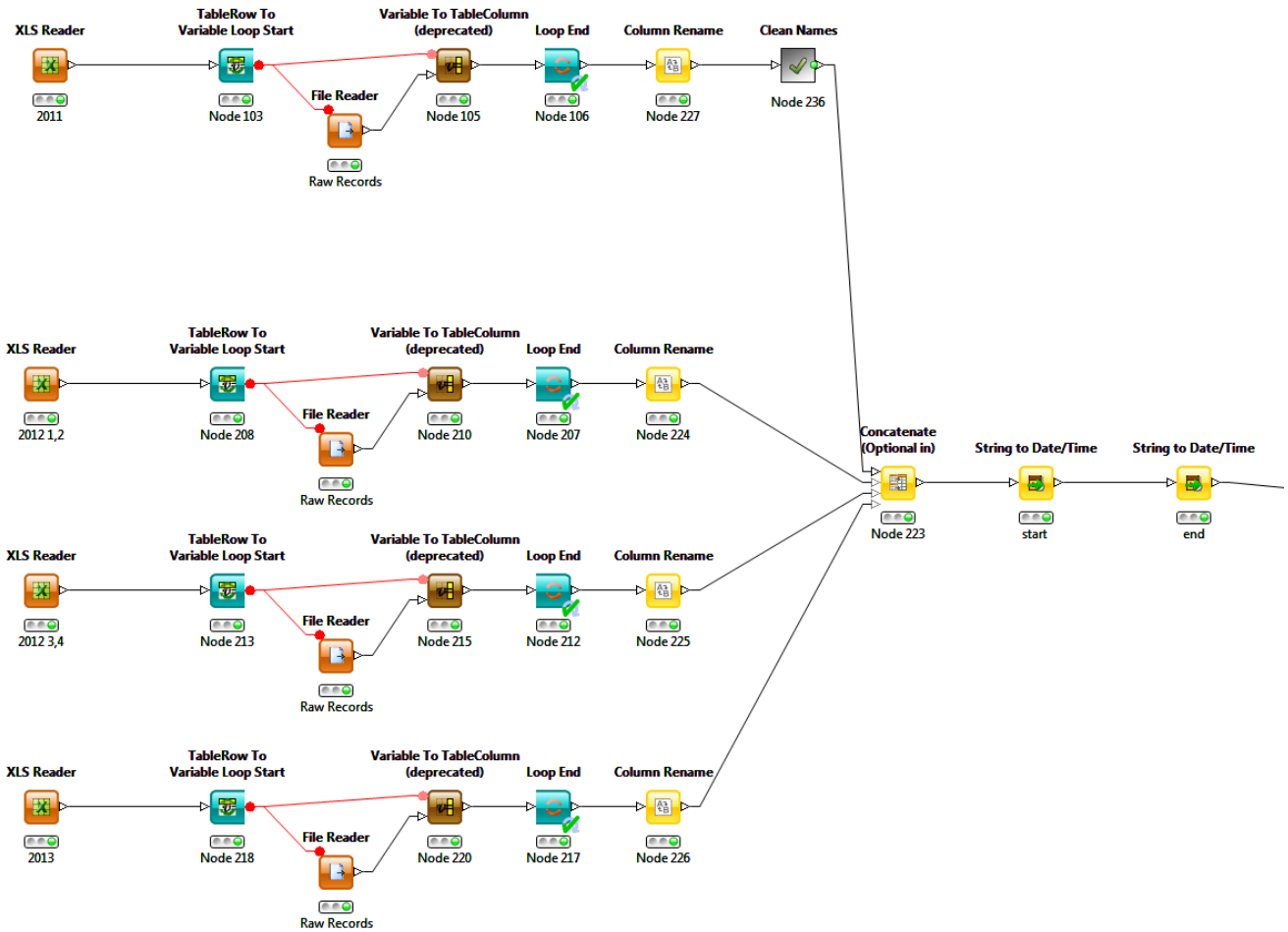
Output data - 3:224 - Column Resorter

File

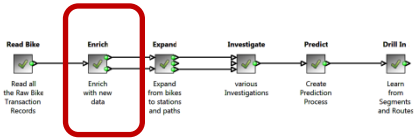
Table "default" - Rows: 5878390 | Spec - Columns: 8 | Properties | Flow Variables

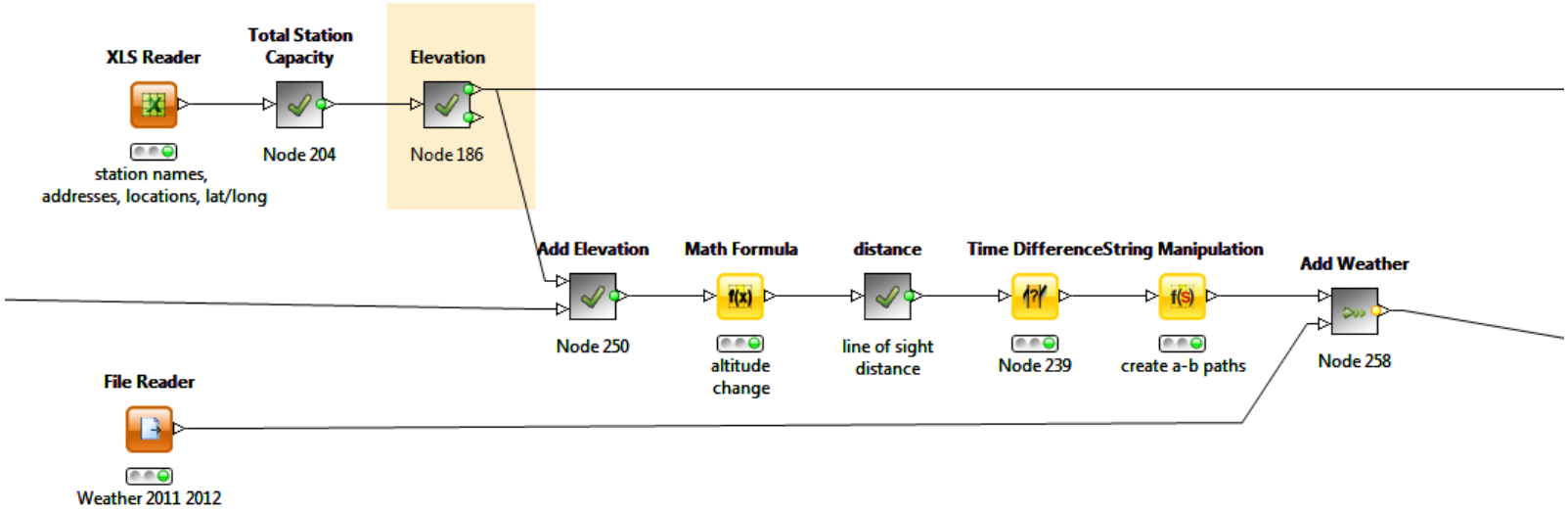
...	Bike#	Start date_time	Start station	Start te...	End station	End terminal	End date_time	Member type
..	W00749	31.Mar.2011 23:58:00	14th & Harvard St NW	31105	16th & Harvard St NW	31103	01.Apr.2011 00:00:00	Registered
..	W01048	31.Mar.2011 23:52:00	19th & L St NW	31224	7th & Water St SW / SW Waterfront	31609	01.Apr.2011 00:08:00	Casual
..	W00340	31.Mar.2011 23:47:00	Lincoln Park / 13th & East Capitol St NE	31619	13th & H St NE	31611	31.Mar.2011 23:50:00	Registered
..	W00981	31.Mar.2011 23:45:00	14th & R St NW	31202	Massachusetts Ave & Dupont Circle NW	31200	31.Mar.2011 23:50:00	Registered
..	W00449	31.Mar.2011 23:34:00	Park Rd & Holmead Pl NW	31602	15th & P St NW	31201	31.Mar.2011 23:43:00	Registered
..	w00583	31.Mar.2011 23:32:00	4th St & Massachusetts Ave NW	31604	Adams Mill & Columbia Rd NW	31104	31.Mar.2011 23:52:00	Registered
..	W00318	31.Mar.2011 23:31:00	37th & O St NW / Georgetown University	31236	21st & M St NW	31212	31.Mar.2011 23:39:00	Registered
..	W00577	31.Mar.2011 23:30:00	14th & Harvard St NW	31105	14th & V St NW	31101	31.Mar.2011 23:33:00	Registered
..	W00416	31.Mar.2011 23:29:00	Lamont & Mt Pleasant NW	31107	7th & T St NW	31109	31.Mar.2011 23:39:00	Registered
..	W00903	31.Mar.2011 23:28:00	13th St & New York Ave NW	31227	Georgia Ave and Fairmont St NW	31207	31.Mar.2011 23:45:00	Registered
..	W00662	31.Mar.2011 23:25:00	New York Ave & 15th St NW	31222	20th & E St NW	31204	31.Mar.2011 23:29:00	Registered
..	W00389	31.Mar.2011 23:21:00	Calvert St & Woodley Pl NW	31106	14th St & Spring Rd NW	31401	31.Mar.2011 23:44:00	Registered
..	W00191	31.Mar.2011 23:21:00	14th & V St NW	31101	14th & Rhode Island Ave NW	31203	31.Mar.2011 23:24:00	Registered
..	W00418	31.Mar.2011 23:18:00	17th & K St NW [formerly 17th & L St NW]	31213	15th & P St NW	31201	31.Mar.2011 23:26:00	Registered
..	W00627	31.Mar.2011 23:17:00	21st & I St NW	31205	Massachusetts Ave & Dupont Circle NW	31200	31.Mar.2011 23:23:00	Registered
..	W00492	31.Mar.2011 23:16:00	21st & M St NW	31212	17th & Corcoran St NW	31214	31.Mar.2011 23:20:00	Registered
..	W00343	31.Mar.2011 23:15:00	11th & Kenyon St NW	31102	14th St & Spring Rd NW	31401	31.Mar.2011 23:19:00	Registered
..	W00798	31.Mar.2011 23:12:00	14th & Rhode Island Ave NW	31203	Adams Mill & Columbia Rd NW	31104	31.Mar.2011 23:26:00	Casual

Reading Sensor data is always messy…… but KNIME is good at it!

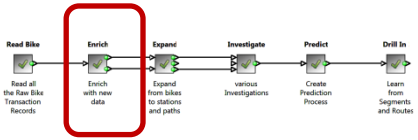


Enrich

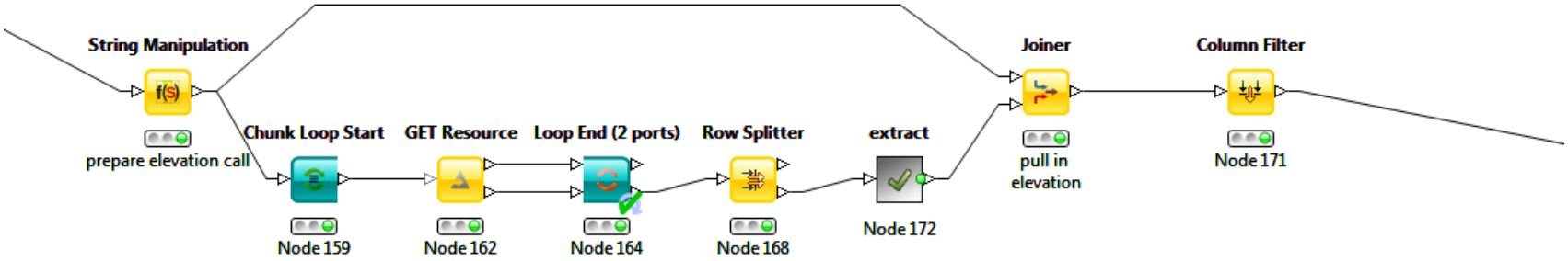

Read Bike
Enrich
Expand
Investigate
Predict
Drill In
Read all the Raw Bike Transaction Records
Enrich with new data
Expand from bikes to stations and paths
various Investigations
Create Prediction Process
Learn from Segments and Routes


XLS Reader
Total Station Capacity
Elevation
station names, addresses, locations, lat/long
Node 204
Node 186
Add Elevation
Math Formula
distance
Time Difference
String Manipulation
Add Weather
Node 250
altitude change
line of sight distance
Node 239
create a-b paths
Node 258
File Reader
Weather 2011 2012

Enrich



Use REST GET Resources to retrieve elevation



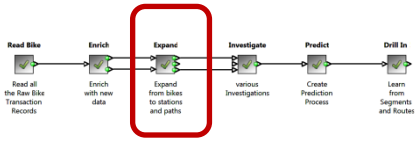
Appended table - 3:2:186:172:170 - String Manipulation(Temp for rejoining)

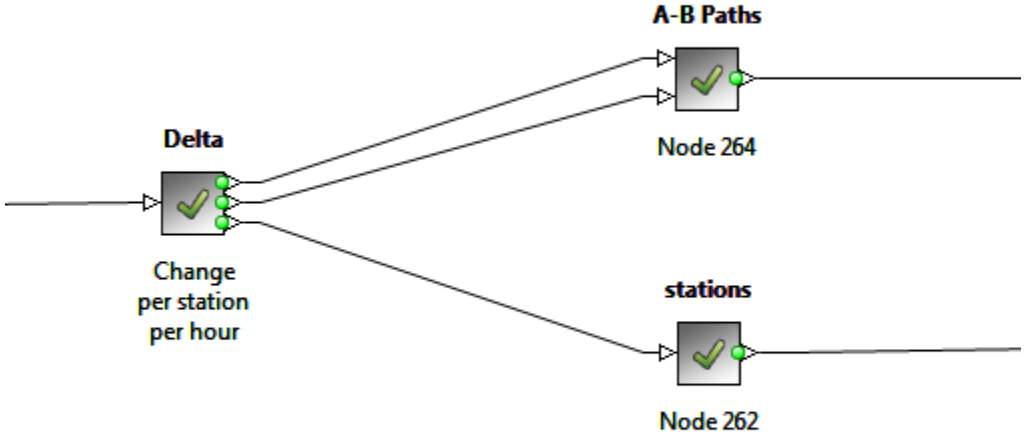
File

Table "default" - Rows: 214 | Spec - Columns: 5 | Properties | Flow Variables

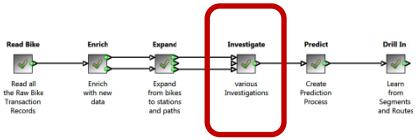
Row ID	S new column2	D long	D lat	D elevation	I Iteration
Row0#0	http://maps.googleapis.com/maps/api/elevation/json?locations=38.8561,-77.0512&sensor=false	-77.051	38.856	15.476	0

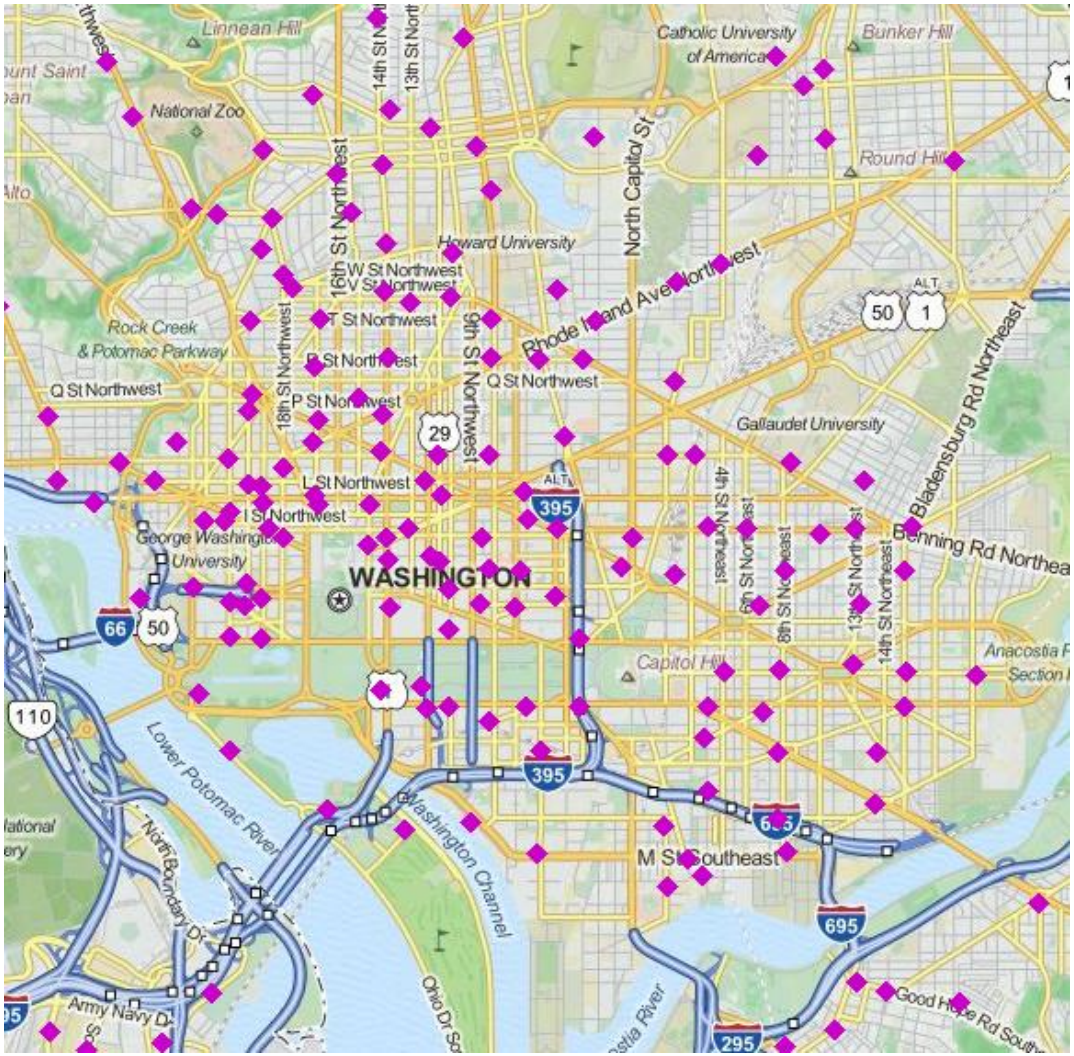
Expand


Read Bike
Read all the Raw Bike Transaction Records
Enrich
Enrich with new data
Expand
Expand from bikes to stations and paths
Investigate
various Investigations
Predict
Create Prediction Process
Drill In
Learn from Segments and Routes


A-B Paths
Node 264
Delta
Change per station per hour
stations
Node 262

Station and Bike Facts





<u>Over 3 years</u>

307 Stations
2963 Bikes
19.4% Casual Bikers

5.9m Bike Moves

The 30 minute rule…

What Is Capital Bikeshare?

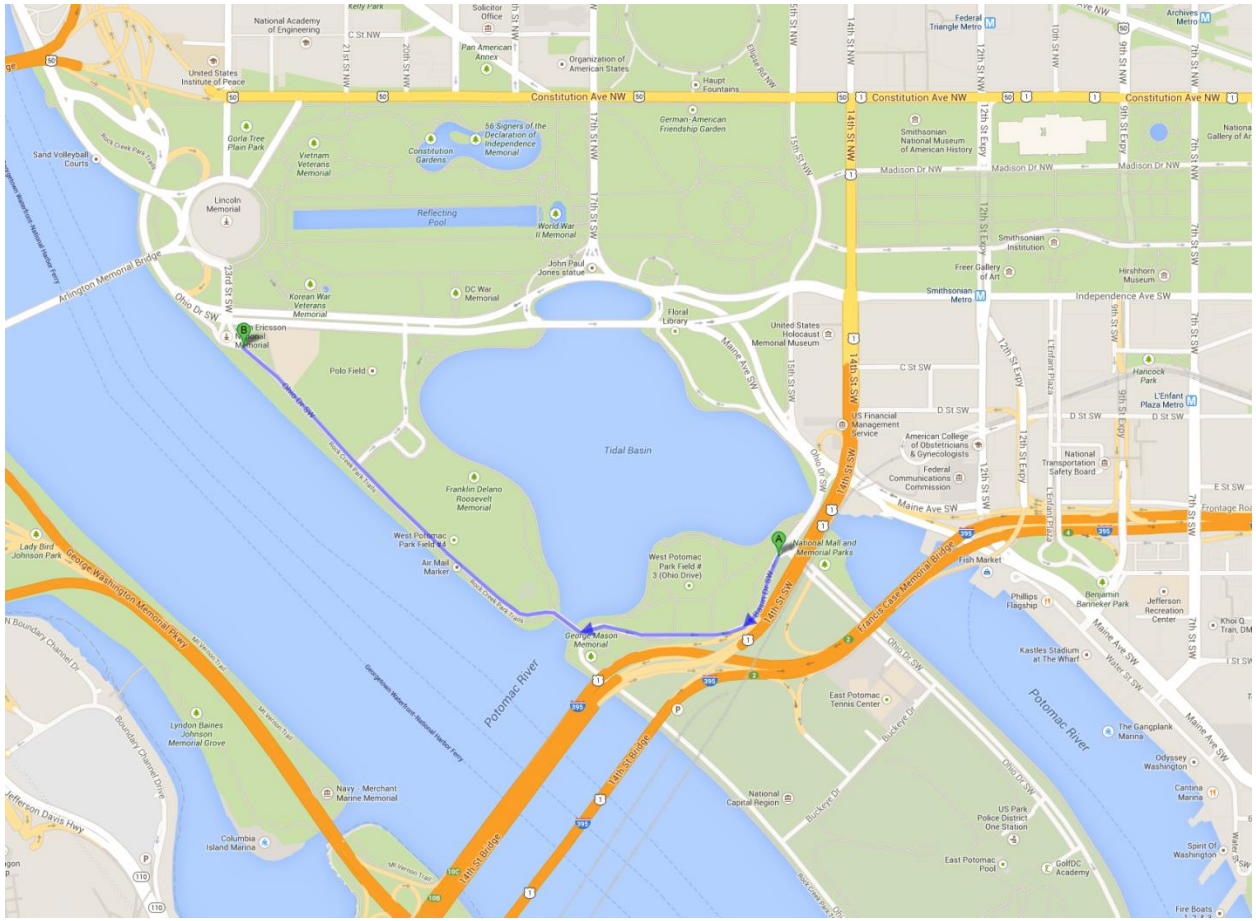
Capital Bikeshare puts over 2500 bicycles at your fingertips. You can choose any of the over 300 stations across Washington, D.C., Arlington and Alexandria, VA and Montgomery County, MD and return it to any station near your destination. Check out a bike for your trip to work, Metro, run errands, go shopping, or visit friends and family. Join Capital Bikeshare for a day, 3 days, a month, a year or try our new daily key option, and have access to our fleet of bikes 24 hours a day, 365 days a year. The first 30 minutes of each trip are free. Each additional 30 minutes incurs an additional fee.

	Under 30	30 and Over
Casual	63%	37%
Subscriber	98%	2%
Overall	91%	9%

Most Popular Segment:

Jefferson Memorial to the Lincoln Memorial

89% Casual Users!



The Business Challenge:

Even MORE of a Business Challenge

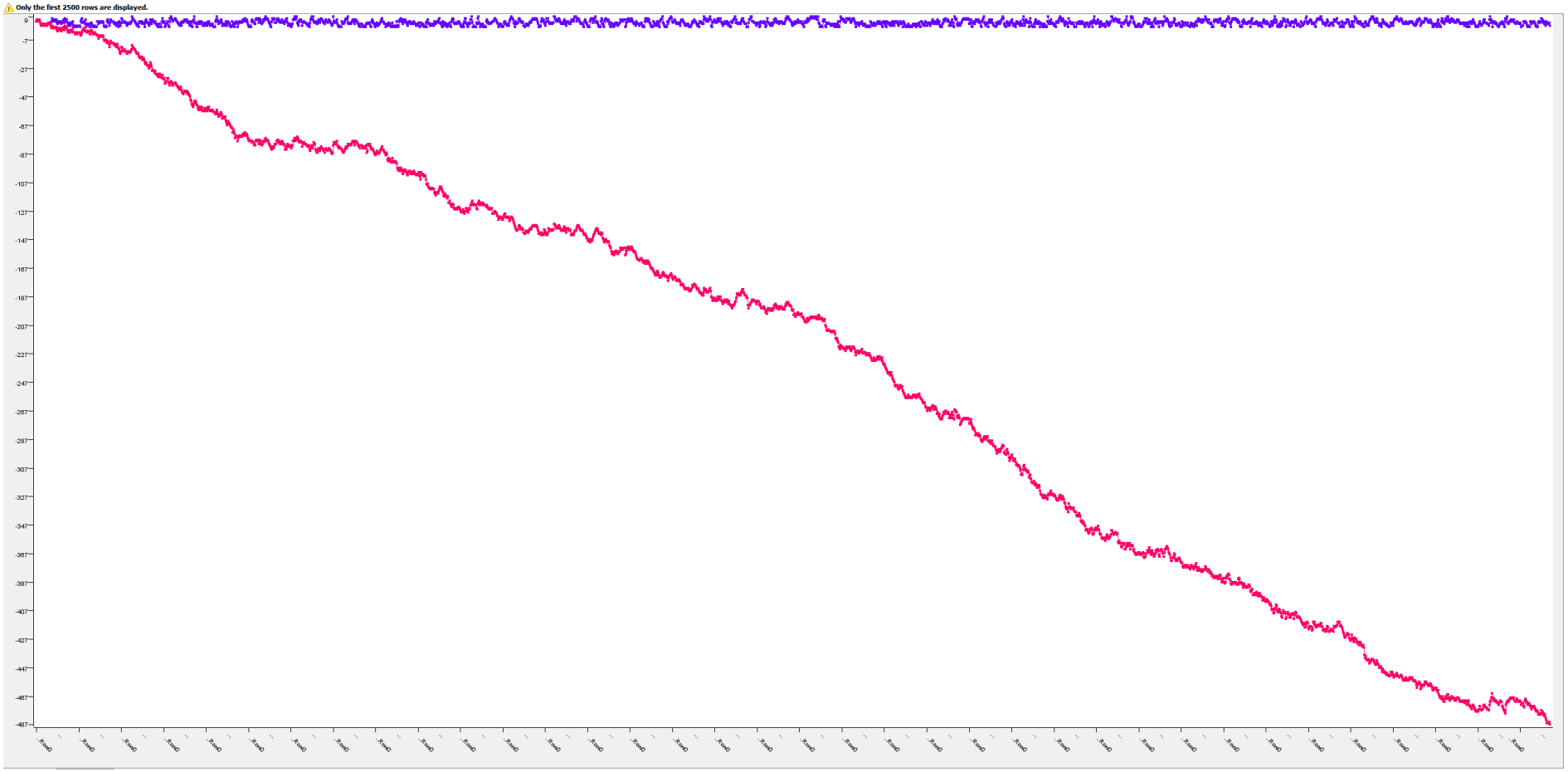
Any Station without bikes for 1 hour:

$XXXX Per Violation

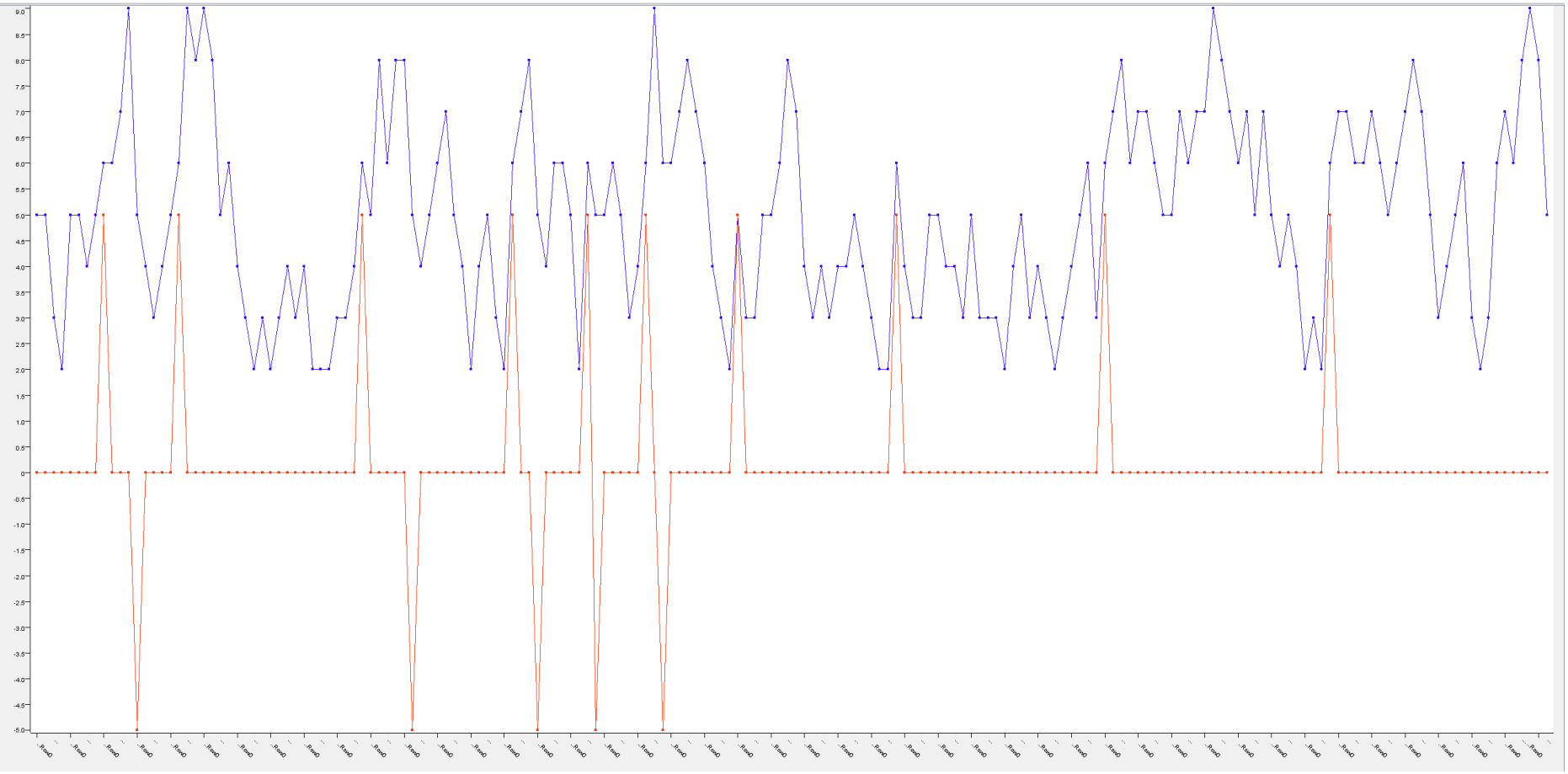
Any Station with no free slots for 1 hour:

$XXXX Per Violation

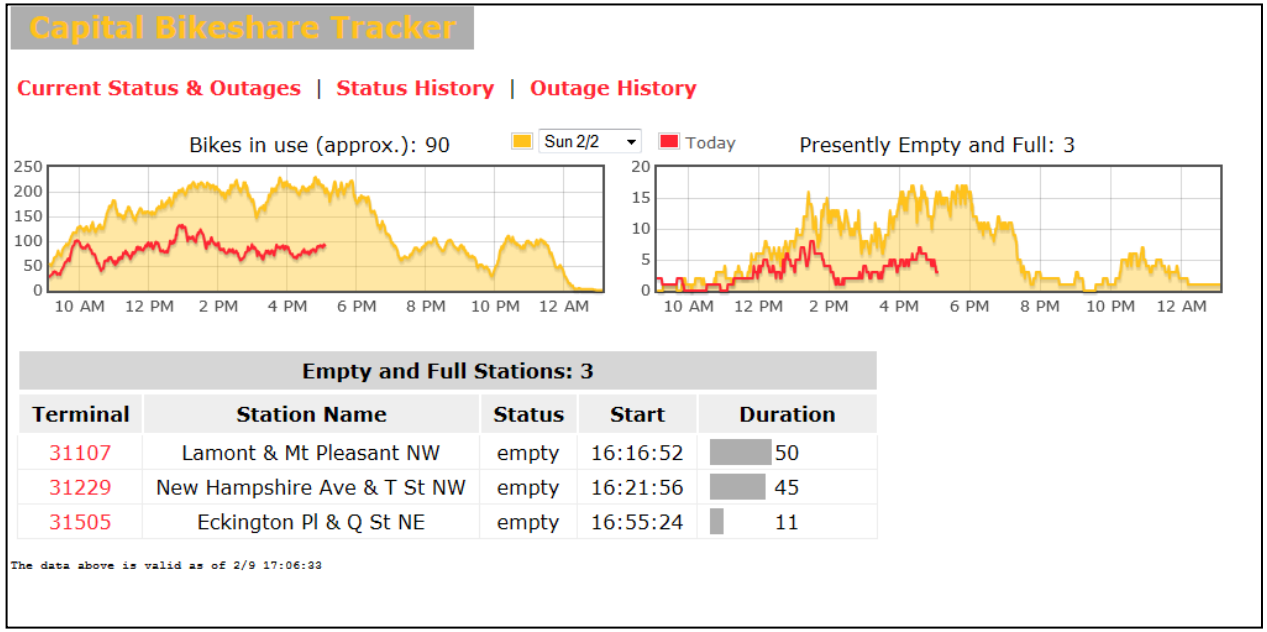
The Business Challenge: 12th and Bell St. -587 Bikes



The Business Challenge: 12th Street and Bell -587 Bikes



“Helpful” Bike Enthusiasts

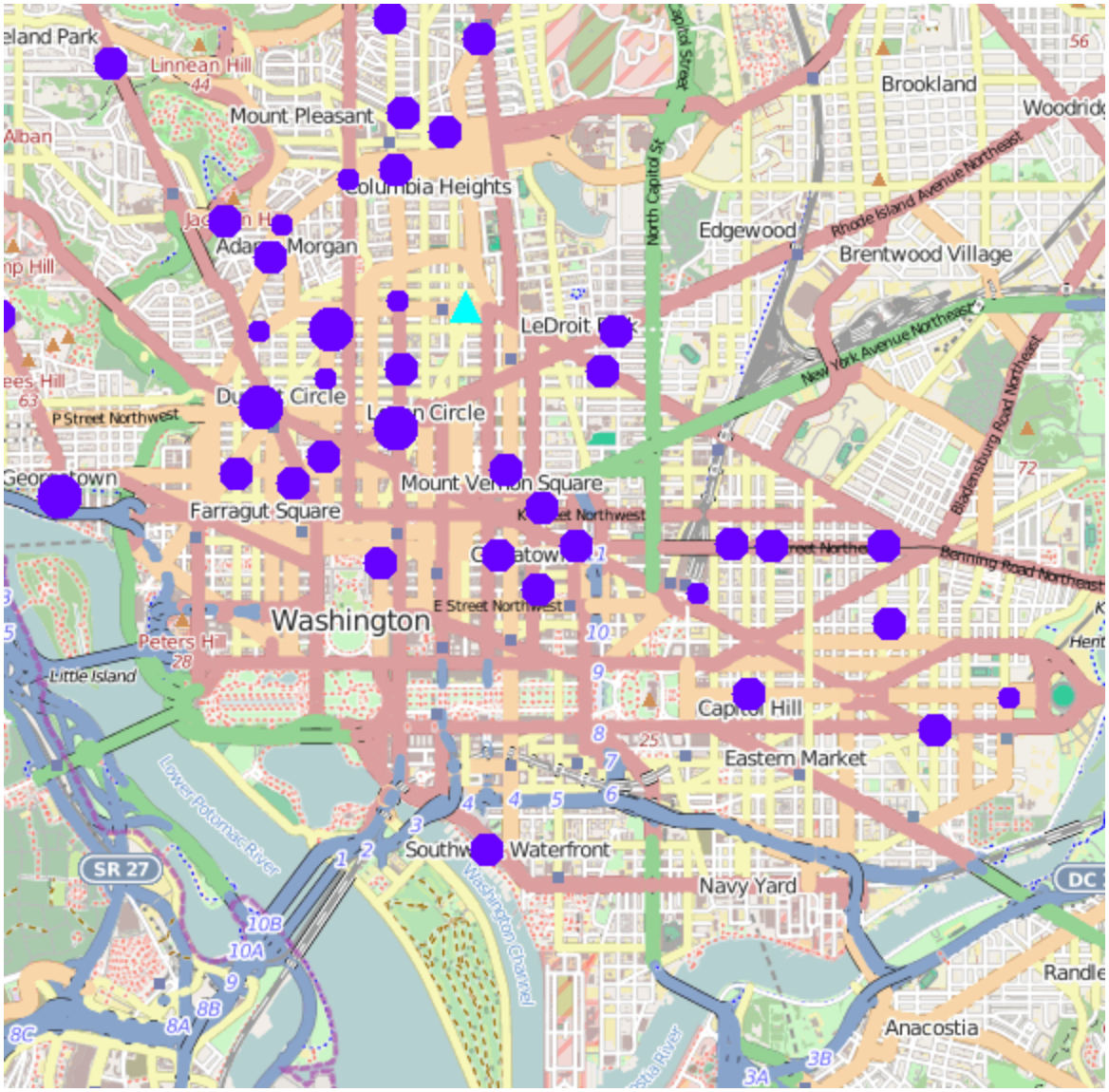


http://www.cabitracker.com/status.php

Capital Bikeshare Response

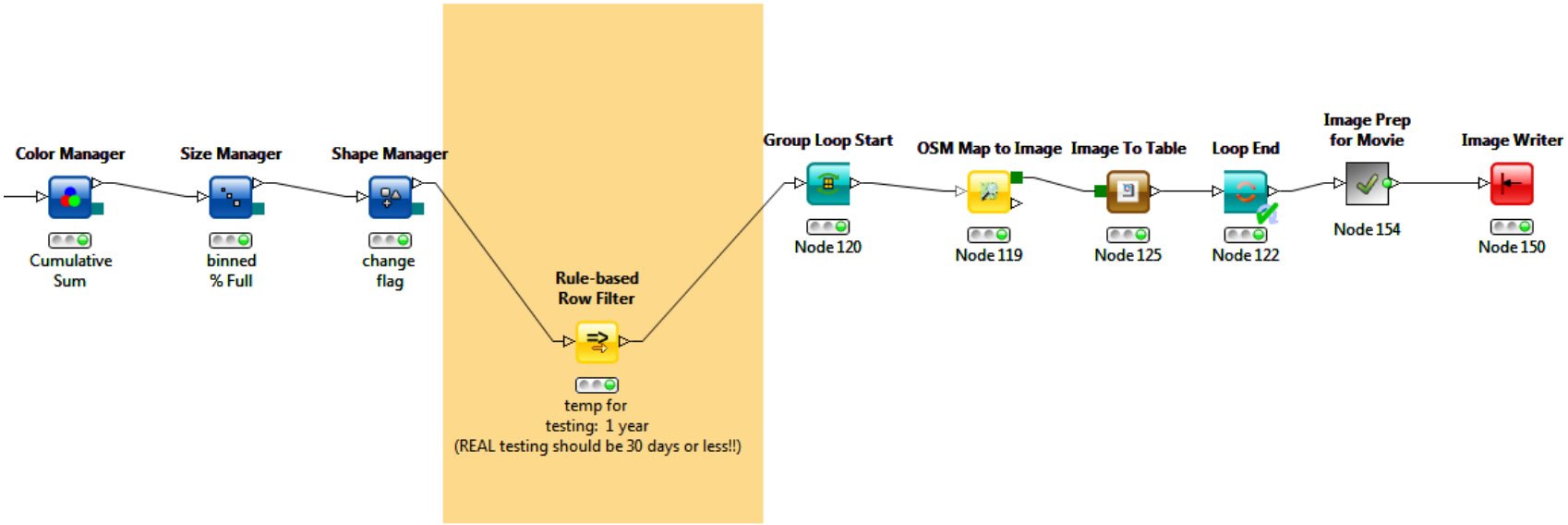
http://bikeportland.org/2013/03/10/behind-the-scenes-of-capital-bikeshare-84006

Station Totals per hour

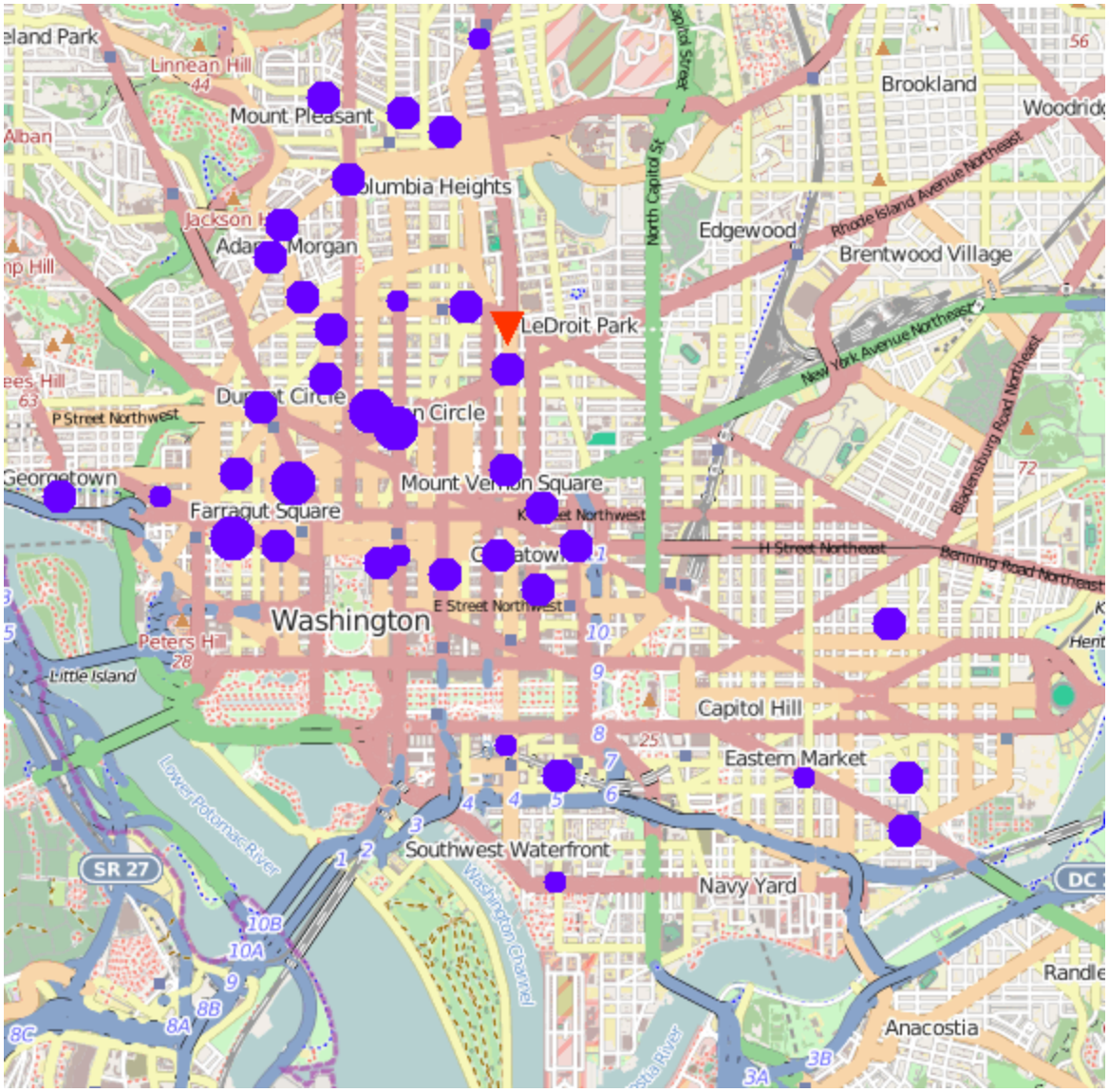


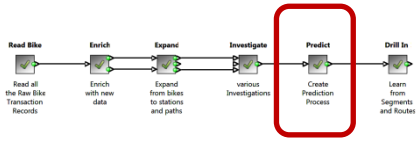
Creating the Video…..

OSM
Image Processing
Group Loop

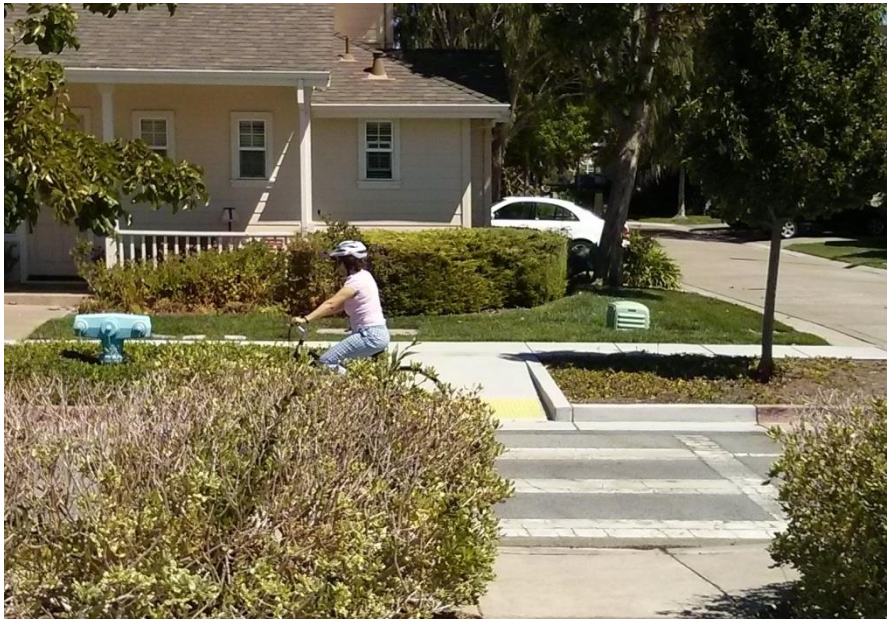


Station Totals per hour




Read Bike
Enrich
Expand
Investigate
Predict
Drill In
Read all the Raw Bike Transaction Records
Enrich with new data
Expand from bikes to stations and paths
various Investigations
Create Prediction Process
Learn from Segments and Routes

ROSARIA



Lean Restocking Alert System

Goal: Implement an alert signal for when restocking (both adding and removing bikes) might be needed in a station.

Target Variable. Flag and Shifted contain the current human operated restocking information.

We want 1 hour warning! Lag(flag-1)

Lean Restocking Alert System

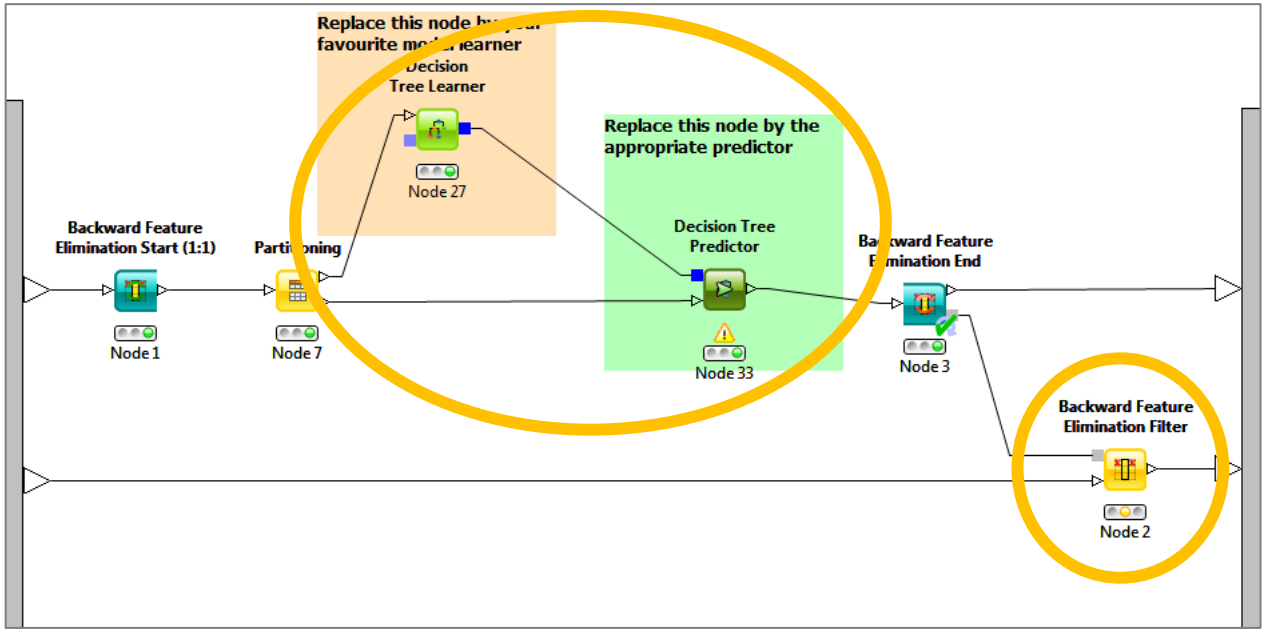
Input Features.

- Weather related features
- Number of registered and casual people showing up
- Station infos (name and max. number of docks)
- Calendar infos (working day, holiday, date)
- Past infos
- Count as the number of bikes added and removed at each hour
- Adjusted cumulative sum as the number of bikes available at the station at a given hour
- Bike ratio = adjusted cumulative sum/total docks available.
- Move to Percentage

Feature Elimination Loop

Two options:

1. Use all the input features (no thinking required, just a powerful machine)
2. Select the most useful input features via the “Feature Elimination” loop



At each step one input feature is removed- i.e. the input feature whose removal produces the smallest error increase.

“Backward Feature Elimination Filter” node

Dialog - 0:43:2 - Backward Feature Elimination Filter

File

Column Selection | Flow Variables | Memory Policy

- [ ] Include target column
- (•) Select features manually
- ( ) Select features automatically by error threshold

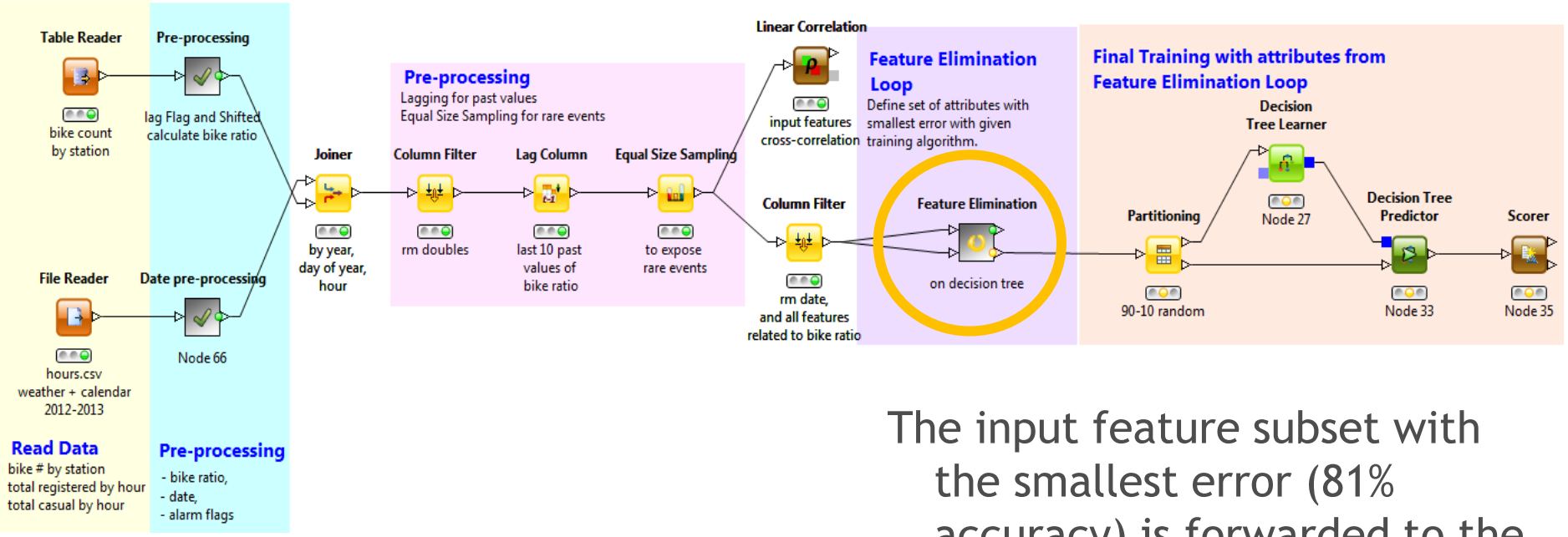
Prediction error threshold 0,5

Error	Nr. of features
0,183	4
0,19	3
0,192	7
0,198	8
0,199	9
0,201	10
0,205	11
0,207	12
0,208	13
0,212	14
0,216	15
0,219	16
0,22	17
0,22	19
0,22	18
0,221	20
0,222	21
0,222	22
0,223	23
0,223	24
0,224	25
0,224	26
0,225	27
0,225	2

I	season
I	mnth
I	hr
I	holiday
I	weekday
I	workingday
I	weathersit
D	temp
D	atemp
D	hum
D	windspeed
I	casual
I	registered
I	Year
I	Day of month
I	Day of year
I	Terminal
S	station
S	Flag(-1)
D	bike ratio
D	bike ratio(-1)
D	bike ratio(-2)
D	bike ratio(-3)
D	bike ratio(-4)
D	bike ratio(-5)
D	bike ratio(-6)
D	bike ratio(-7)
D	bike ratio(-8)

OK | Apply | Cancel | ?

“Lean Restocking Alert System” workflow



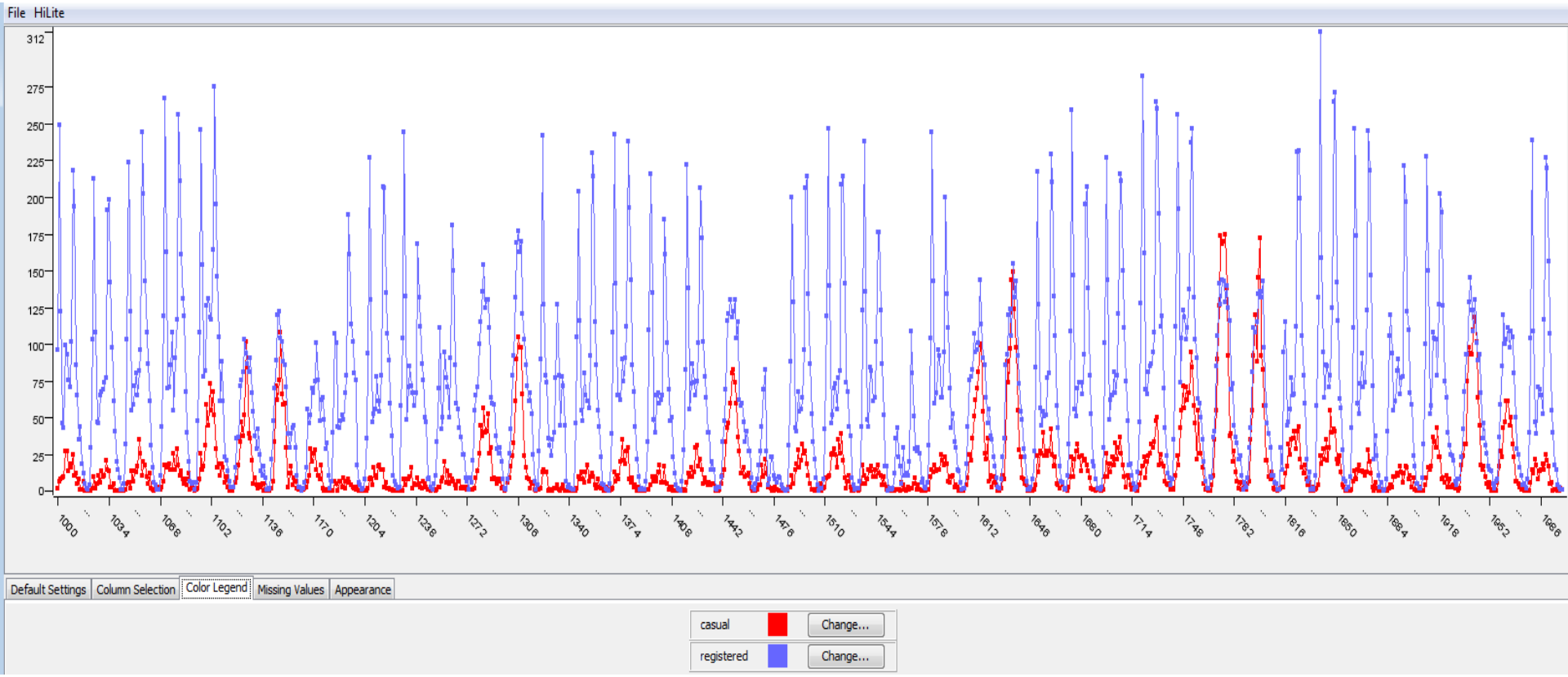
The input feature subset with the smallest error (81% accuracy) is forwarded to the final model training block:

- Hour of the day
- working day (Y/N)
- Current Bike Ratio
- Terminal (station code)

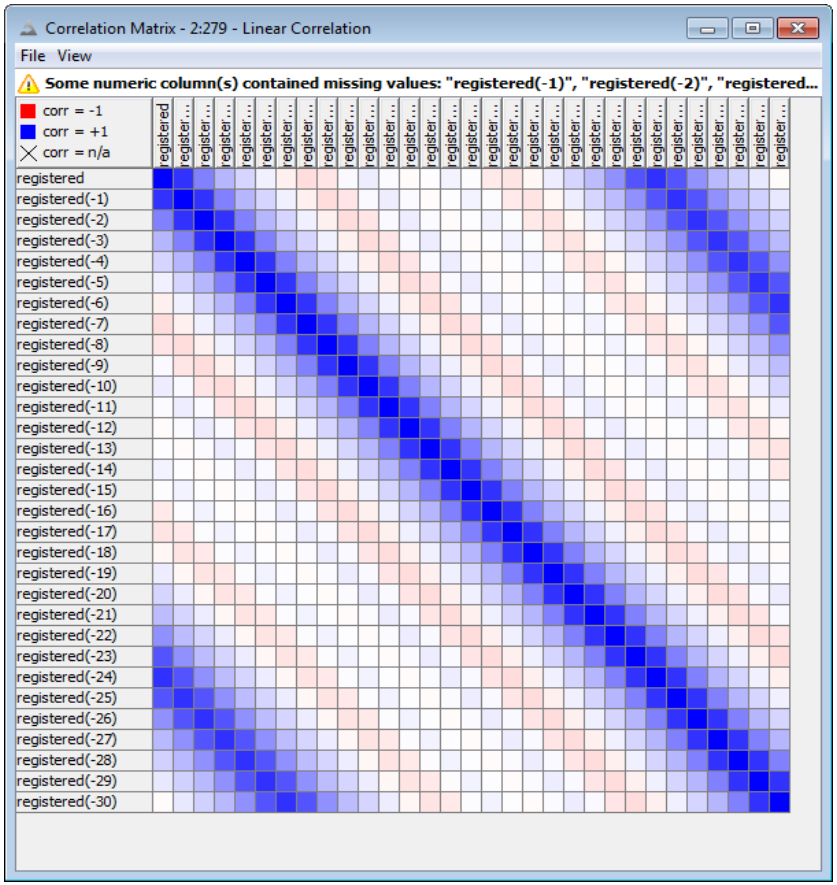
Bike Ratio past and weather infos do not seem to be relevant!

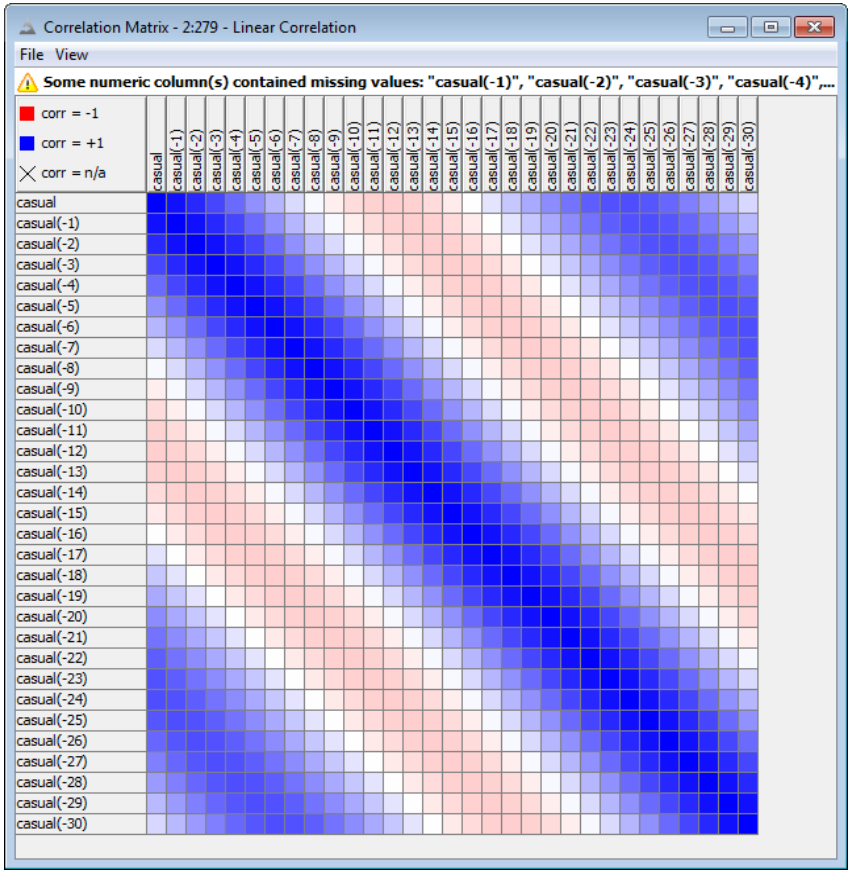
Total Bikers Number Prediction by the Hour

Registered (blue) vs. Casual (red) Bikers



Casual vs. Registered: Auto-Correlation Matrix (24h)



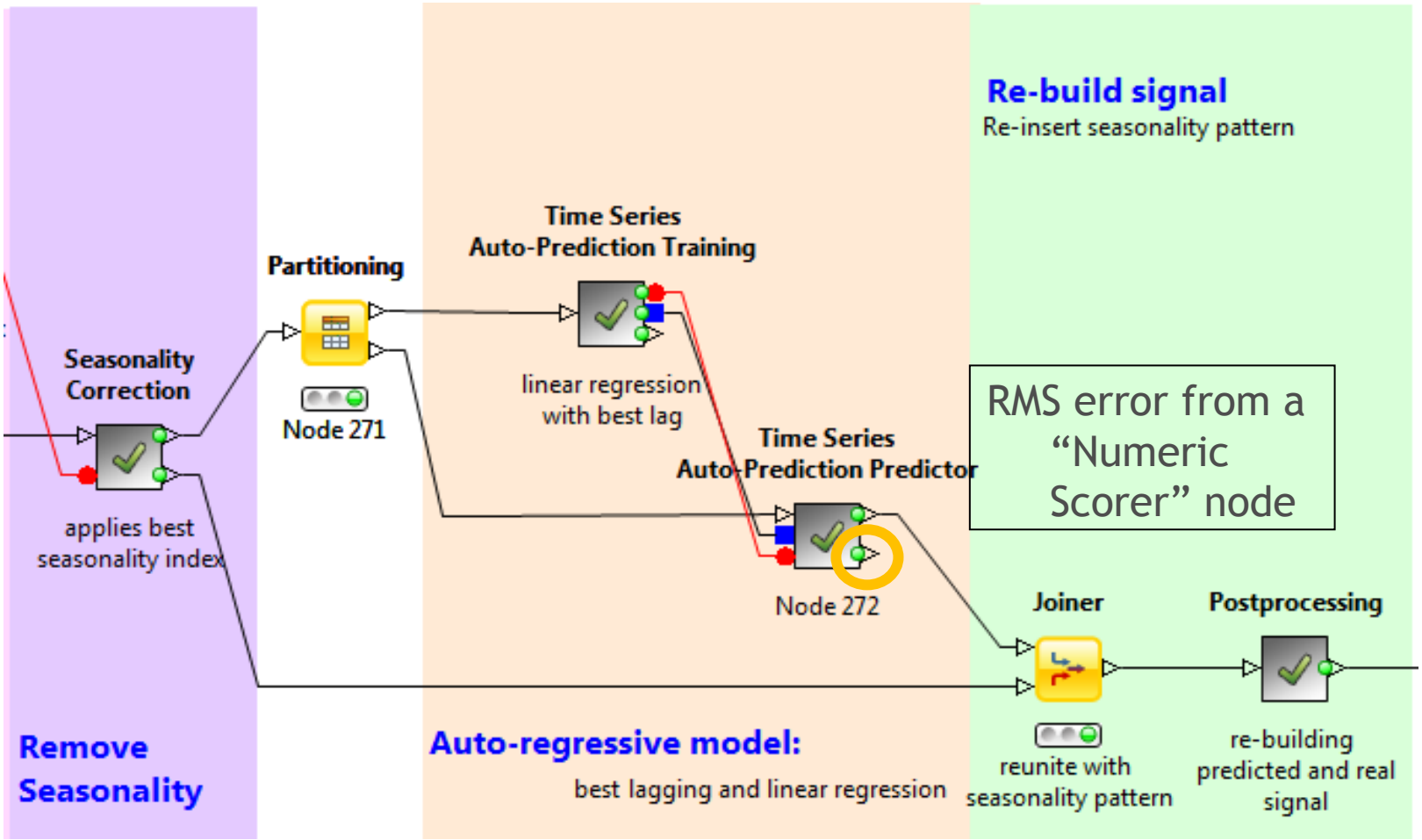


Registered Bikers

Casual Bikers

Time Series Prediction

(from last year's KNIME UGM):



Time Series Prediction

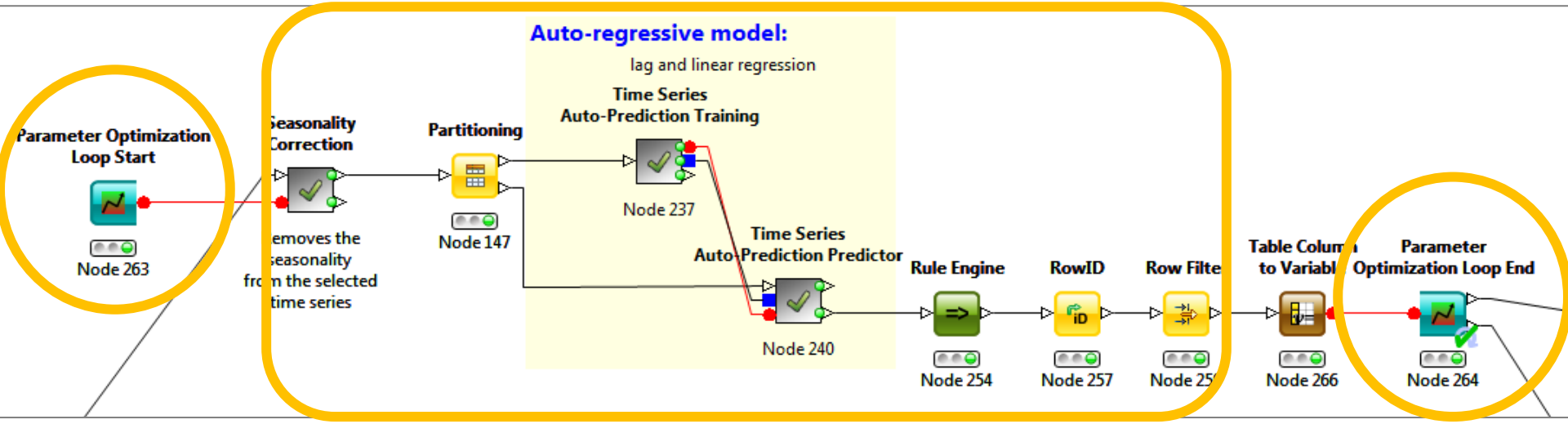
(from last year's KNIME UGM):

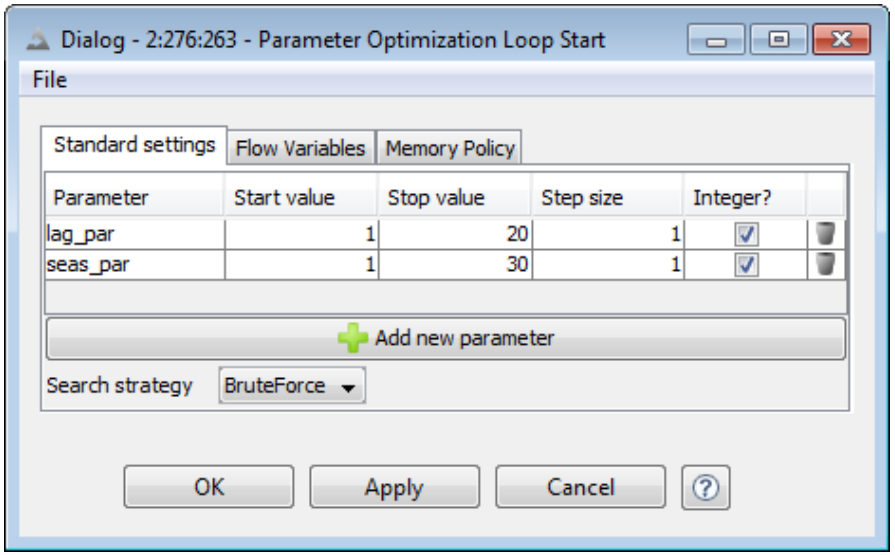
- Past lags and seasonality index had to be defined by hand.
- We tried a few values and we chose the best.
- Hypotheses for the bikers data: 24h seasonality index, 10-20 past hour values needed

This year's UGM:

We select the best values programmatically using the **Optimization Parameter Loop**!

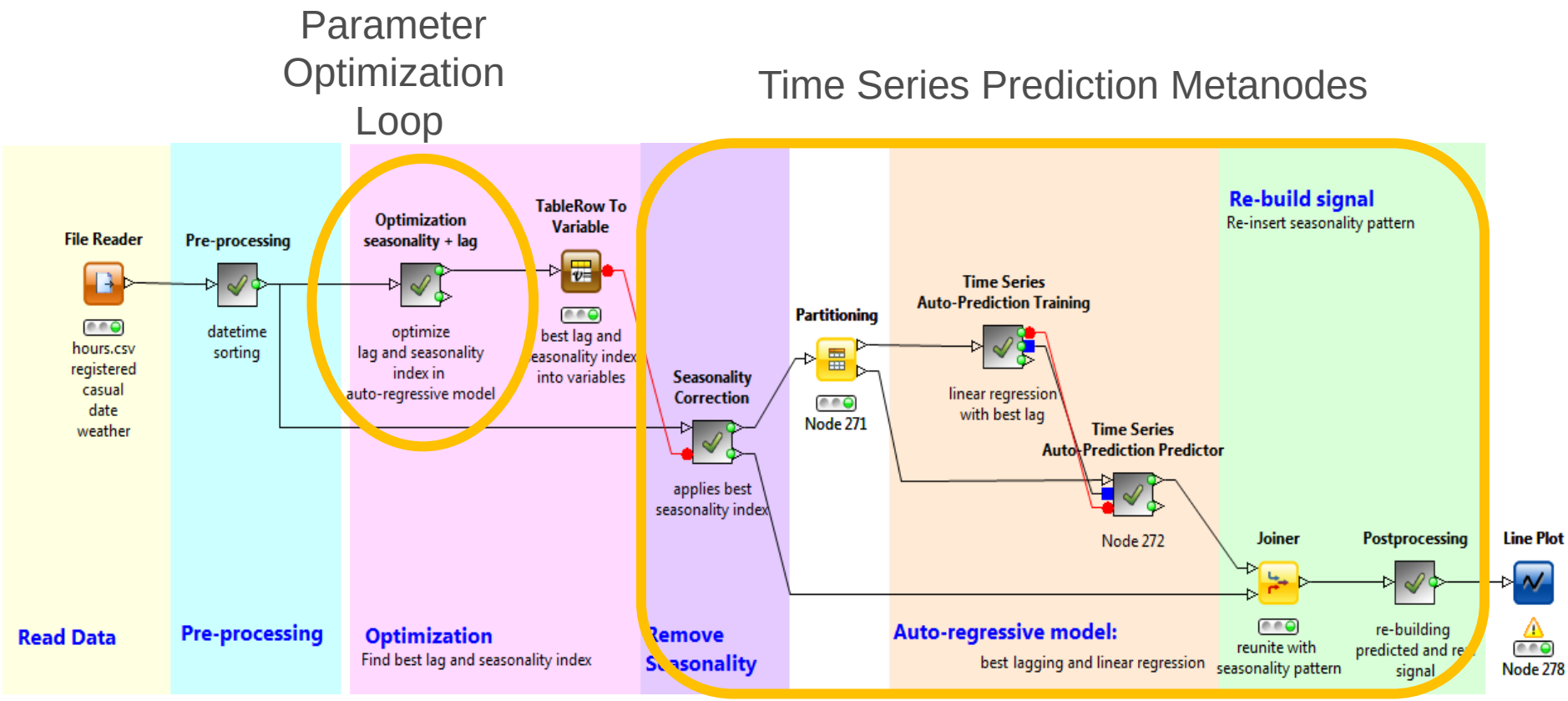
Parameter Optimization Loop





- Looping over different values for lagging and seasonality index (Brute Force or Hillclimbing).
- Collecting the RMS error at the end of the loop.
- Selecting the parameter set with minimum RMS error.

“Bikers Prediction” workflow



	Seasonality index	Past lags	RMS Error
casual	1hour	20	13.8
registered	24 hours	10	68.3

Conclusions….

- Best Input Feature Subset using the "Feature Elimination" Metanode
- Best Parameter Set for Time Series Prediction using the "Parameter Optimization" node
- Weather influence on bikers is overrated!
- Casual Bikers 24 hours Seasonality is also not so relevant!

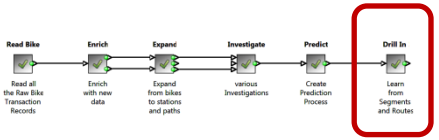
Top Net Bike Change Stations

Filtered table - 0:289 - Column Filter

File

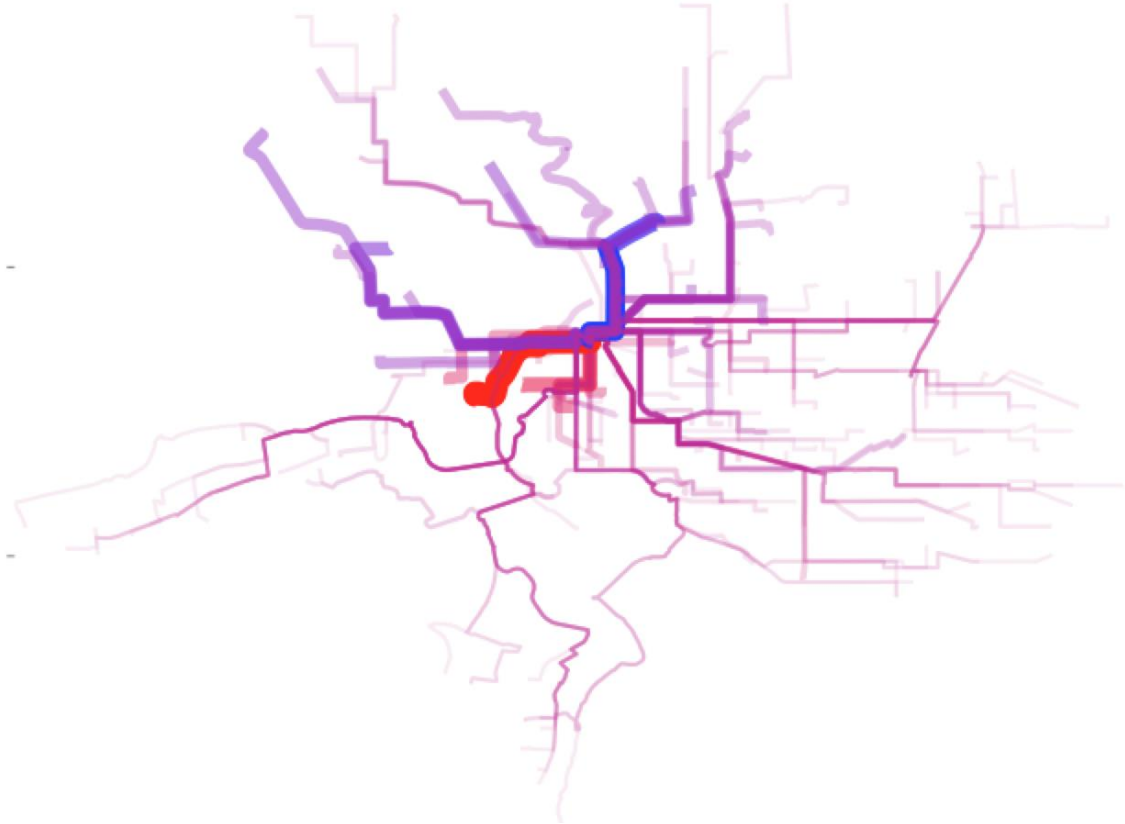
Table "default" - Rows: 10 | Spec - Columns: 3 | Properties | Flow Variables

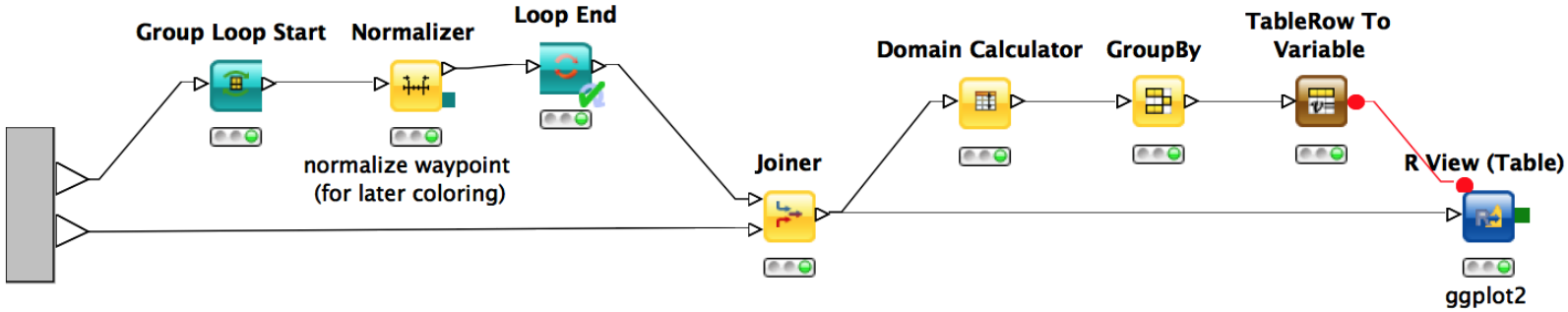
Row ID	S name	I Terminal	I Count(station)
Row98_Row49	Massachusetts Ave & Dupont Circle NW	31200	366953
Row230_Row...	Columbus Circle / Union Station	31623	278418
Row99_Row22	15th & P St NW	31201	255757
Row111_Row57	17th & Corcoran St NW	31214	206565
Row77_Row14	14th & V St NW	31101	187431
Row126_Row...	New Hampshire Ave & T St NW	31229	187031
Row220_Row76	Eastern Market Metro / Pennsylvania Ave & 7th St SE	31613	181194
Row80_Row17	Adams Mill & Columbia Rd NW	31104	169684
Row101_Row24	14th & Rhode Island Ave NW	31203	160891
Row125_Row90	8th & H St NW	31228	159584


Read Bike
Read all the Raw Bike Transaction Records
Enrich
Enrich with new data
Expand
Expand from bikes to stations and paths
Investigate
various Investigations
Predict
Create Prediction Process
Drill In
Learn from Segments and Routes

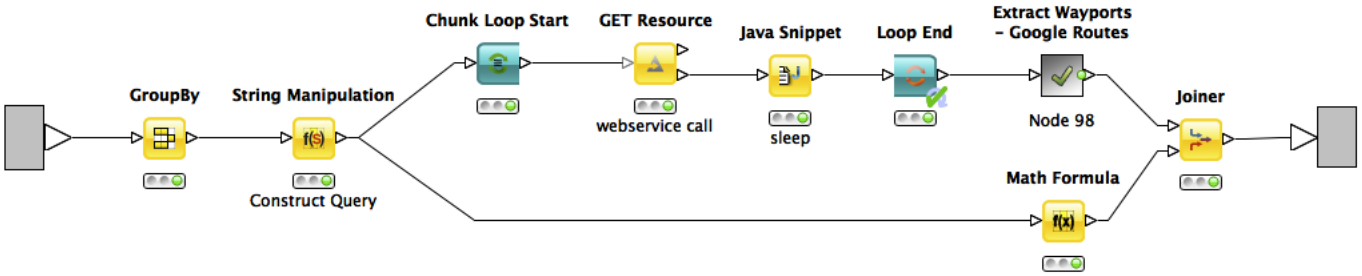
Aaron

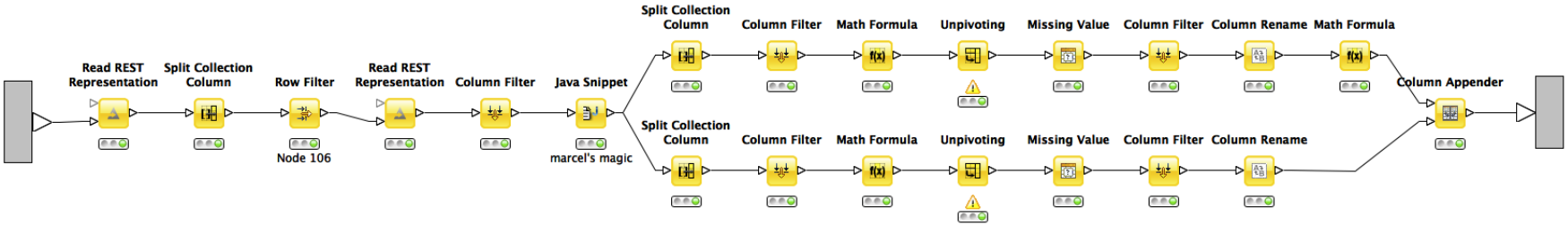
An aside on visualization





Finding route data (Just google it)





REST Webservices with KREST

http://maps.googleapis.com/maps/api/directions/json?
&origin=lat,long&destination=lat,long&mode=bicycling

Appended table - 5:169:98:97:91 - Colum...

File

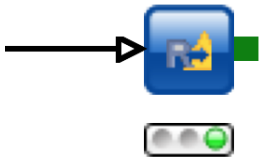
Table "default" - Rows: 6957

Row ID	D Lat	D Long	I waypo...
Row0	38.91	-77.044	0
Row1	38.91	-77.045	1
Row2	38.91	-77.045	2
Row3	38.91	-77.045	3
Row4	38.911	-77.045	4
Row5	38.911	-77.045	5
Row6	38.911	-77.044	6
Row7	38.911	-77.044	7
Row8	38.911	-77.04	8
Row9	38.911	-77.037	9
Row10	38.911	-77.035	10
Row11	38.91	-77.035	11
Row12	38.91	-77.035	12
Row167	38.91	-77.035	13
Row168	38.91	-77.035	14
Row169	38.91	-77.035	15
Row170	38.91	-77.036	16

https://developers.google.com/maps/documentation/directions/

Plotting with ggplot2

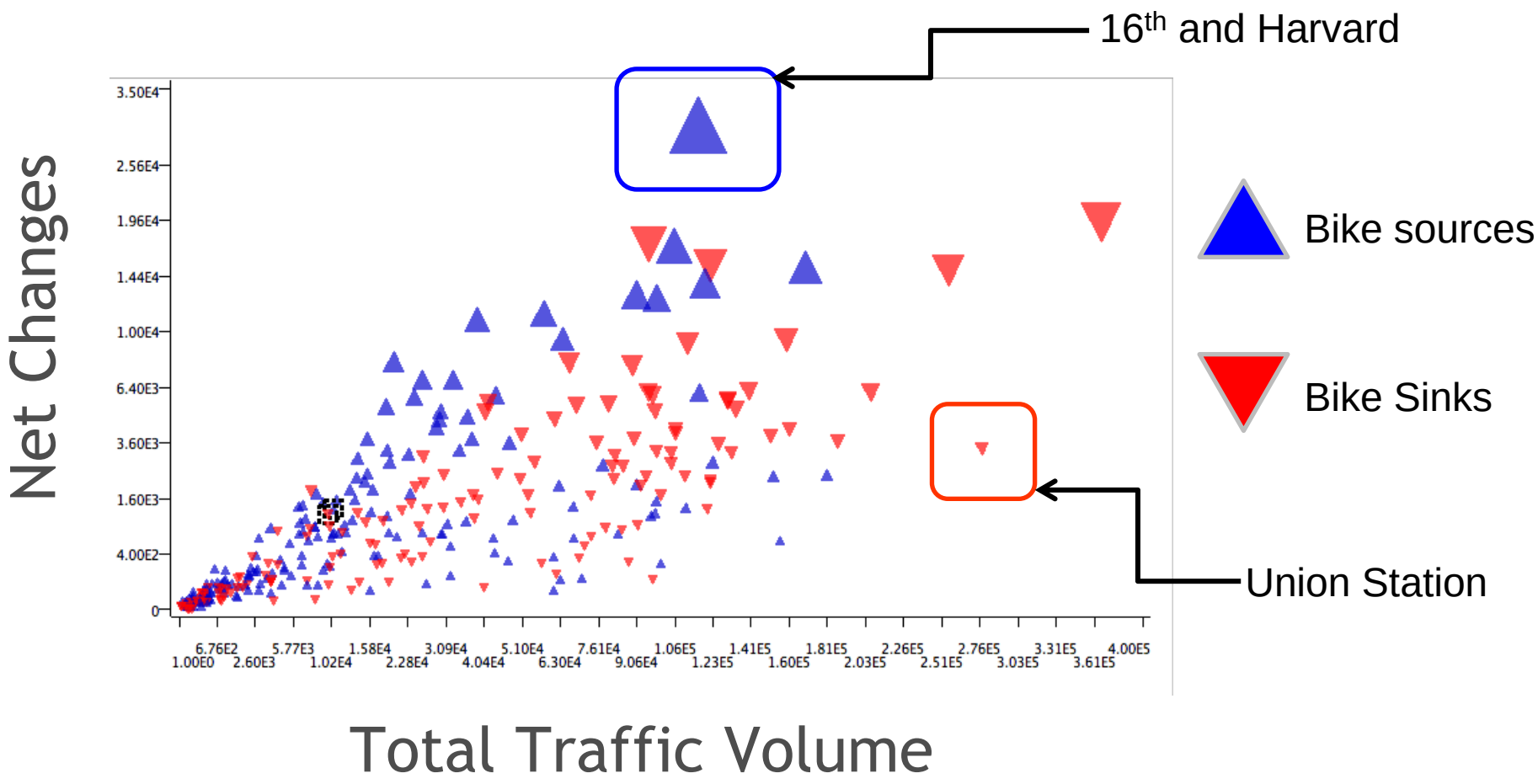
R View (Table)



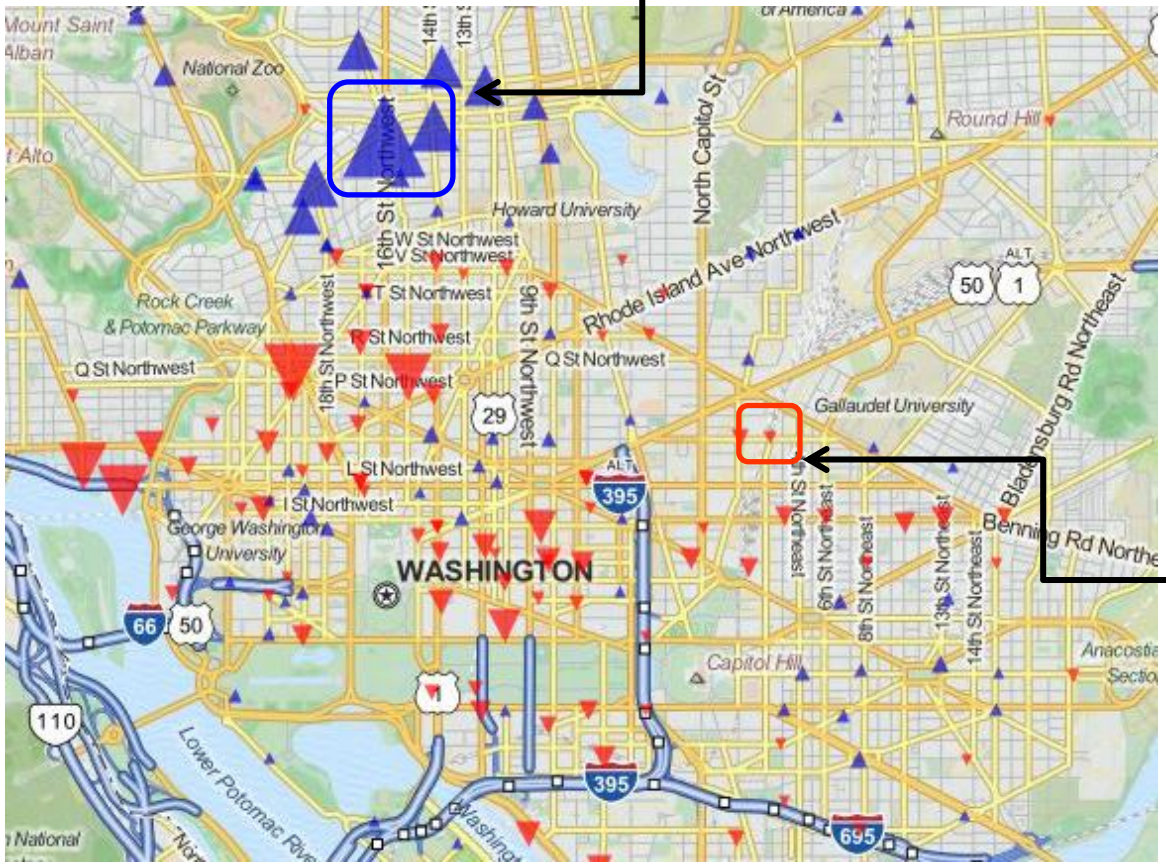
ggplot2

```
require(ggplot2)

clean_theme = theme(text = element_blank(),
                    panel.background = element_blank(),
                    panel.border = element_blank(),
                    axis.title.x = element_blank(),
                    axis.title = element_text(size=14, face="italic", colour = "black"),
                    axis.text = element_text(size=12, face="italic", colour = "black"),
                    legend.text = element_blank(),
                    legend.key = element_blank(),
                    legend.position = "none",
                    panel.grid = element_blank()
                    )

plot = ggplot() + geom_path(aes(knime.in$"Lat",
                                knime.in$"Long",
                                group = knime.in$"route",
                                alpha = knime.in$"Count*(a)",
                                color = factor(knime.in$"route")),
                            knime.in)

plot + clean_theme
```

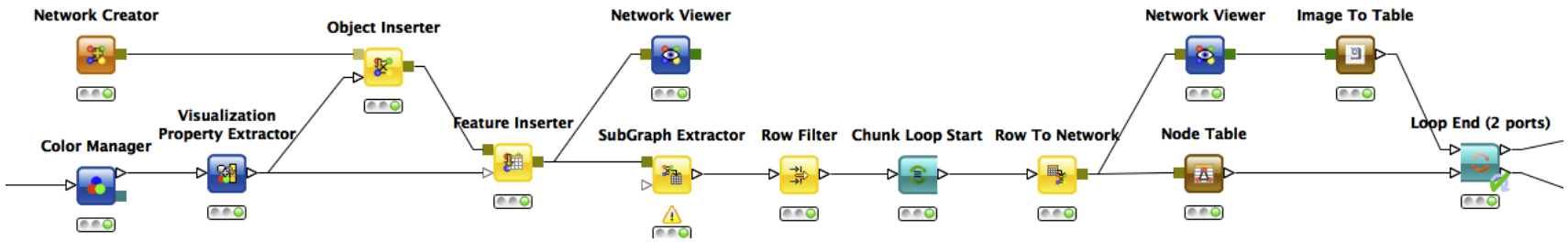
And now some analysis…



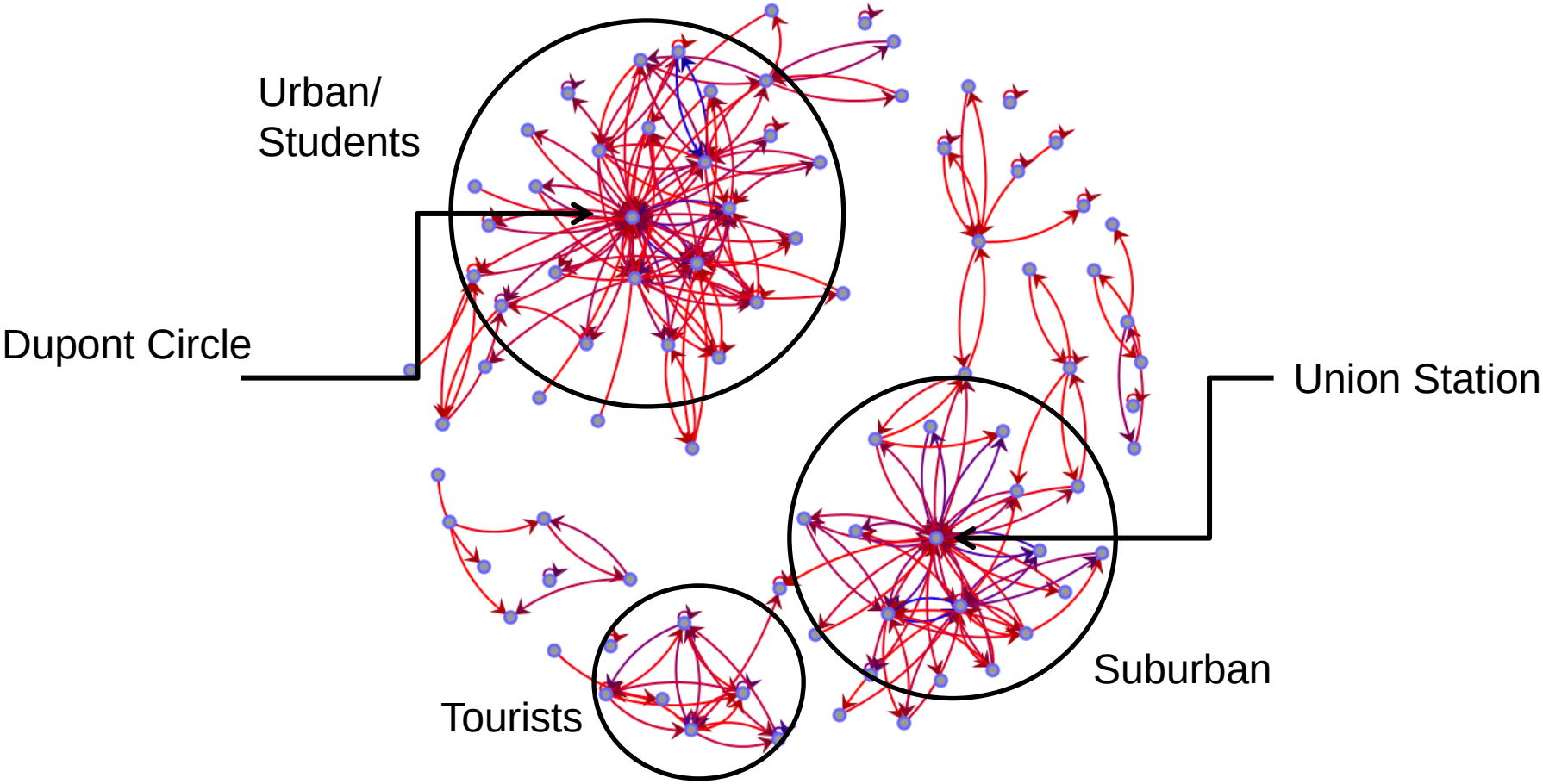
Stations with deficits and surpluses



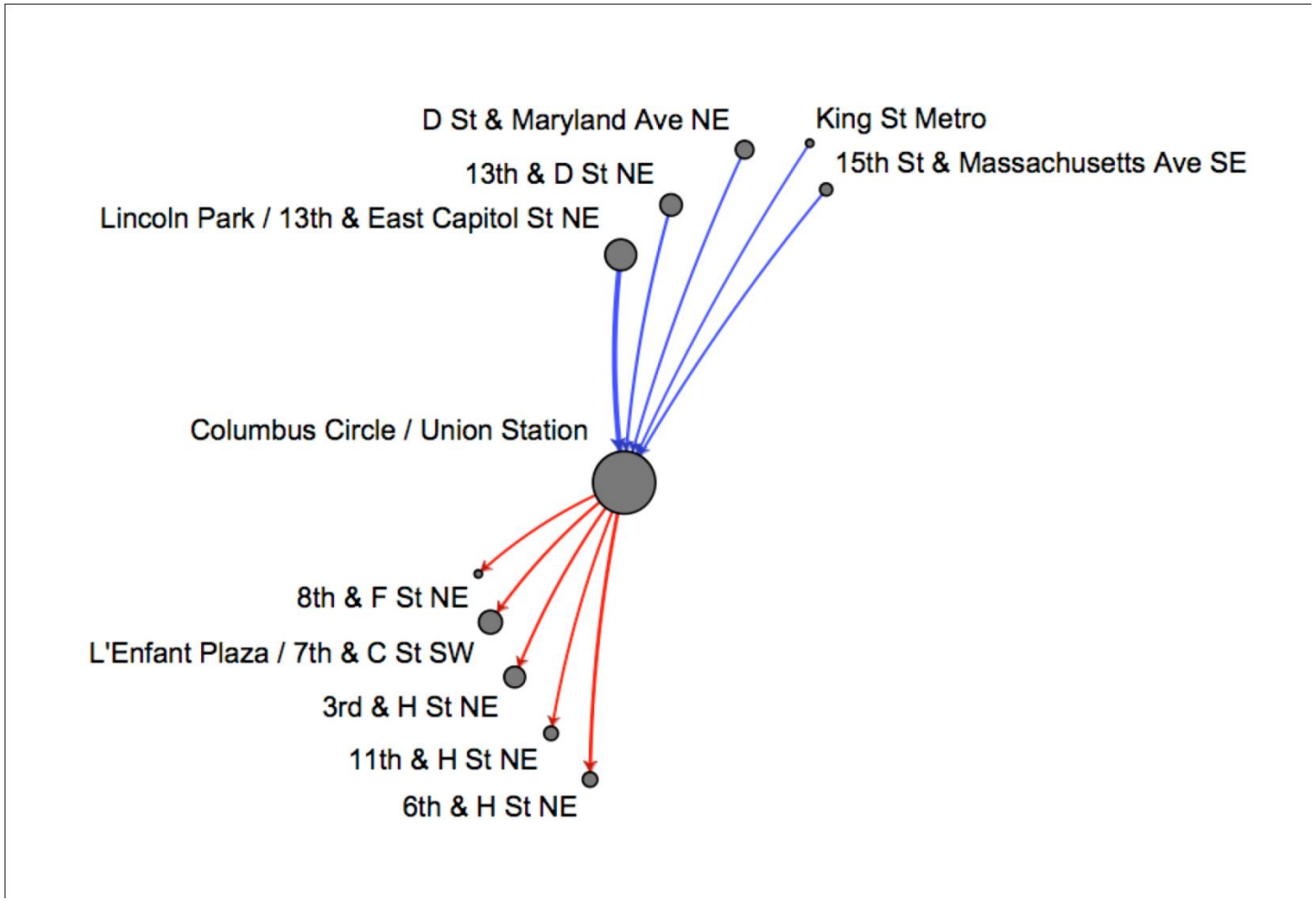
Network Analysis



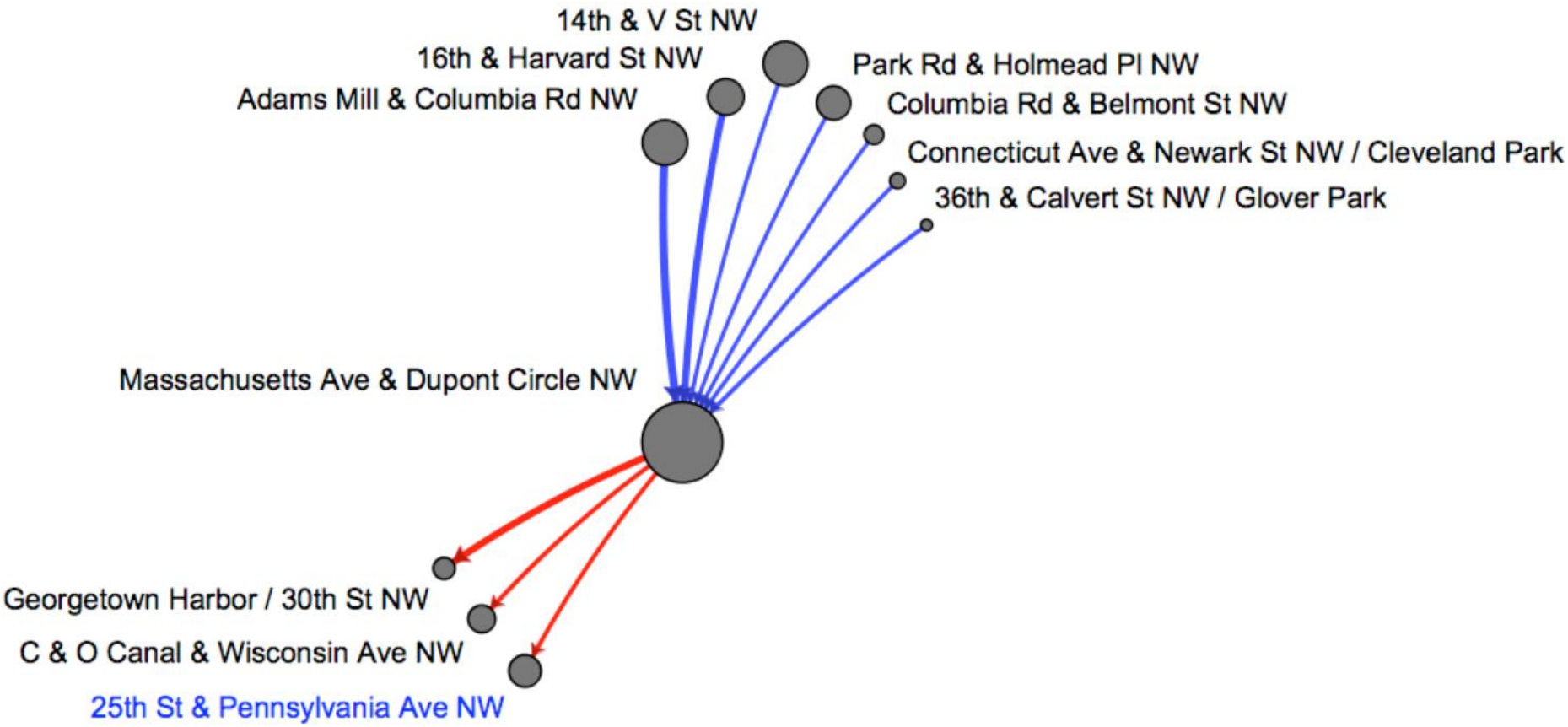
Top 250 Routes



Union Station Subnet



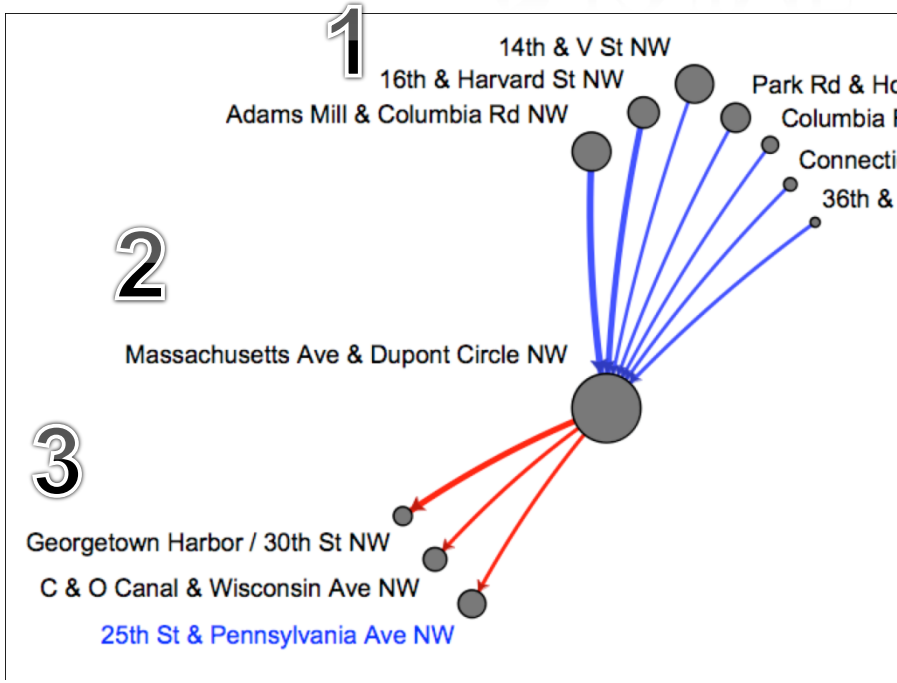
Dupont Circle Subnet



1) 16th and Harvard

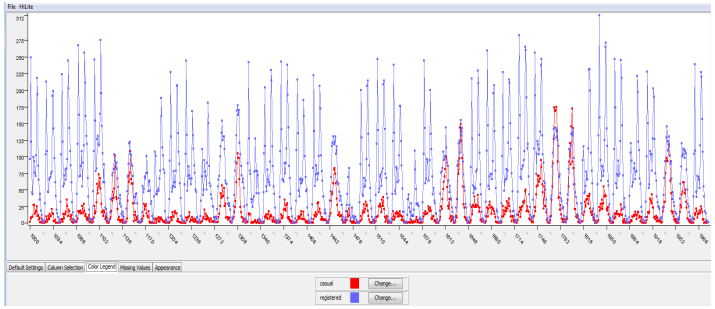
2) Dupont Circle

3) Georgetown Harbor

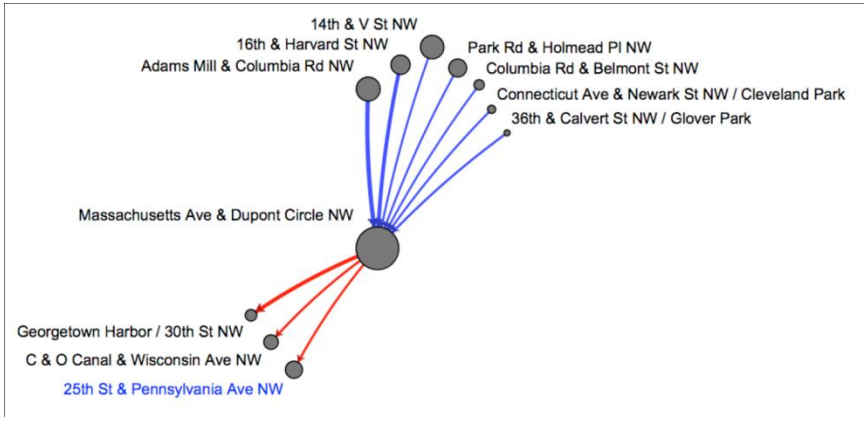


Possible Next Steps

Combine Prediction with Path !!!!

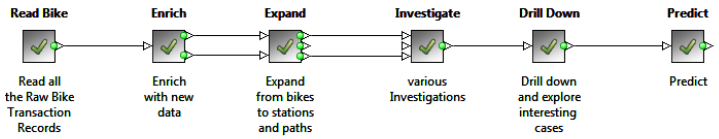


+



Enrich through analyzing images

Lessons Learned



Sensor Data is simple… but can still be messy

- get it in, get it clean!

Connect it!

- Common keys, Lat/long, time, etc.
- No common key? Network analysis + Imaging

Enrich / Expand

- KNIME Transformation
- External Sources; REST calls; Palladian, etc.

Explore

- Visualization (over time!) OSM, Imaging
- Network Analysis
- Correlation matrix!

Prediction

- Feature Elimination + “Classic Techniques”
- “Brute Force” Feature Elimination + “Classic Techiques

The Internet of Things

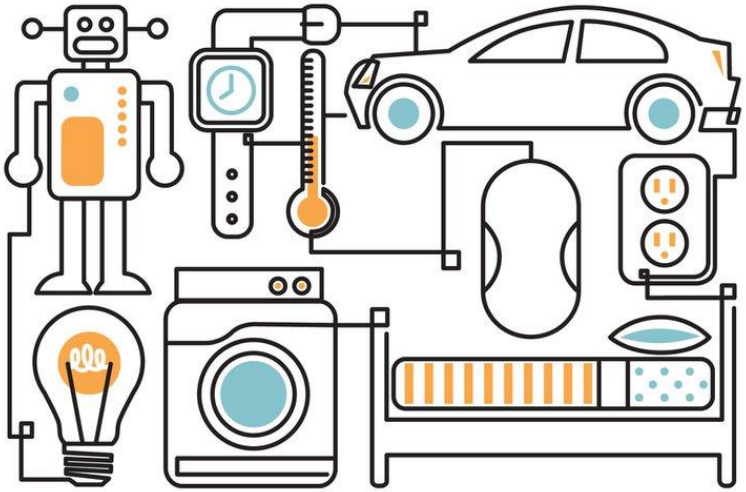


Illustration by CRISTINA BYVIK

Use Public Data Please….

The Internet of Things:
White paper, Workflows, Data will be available