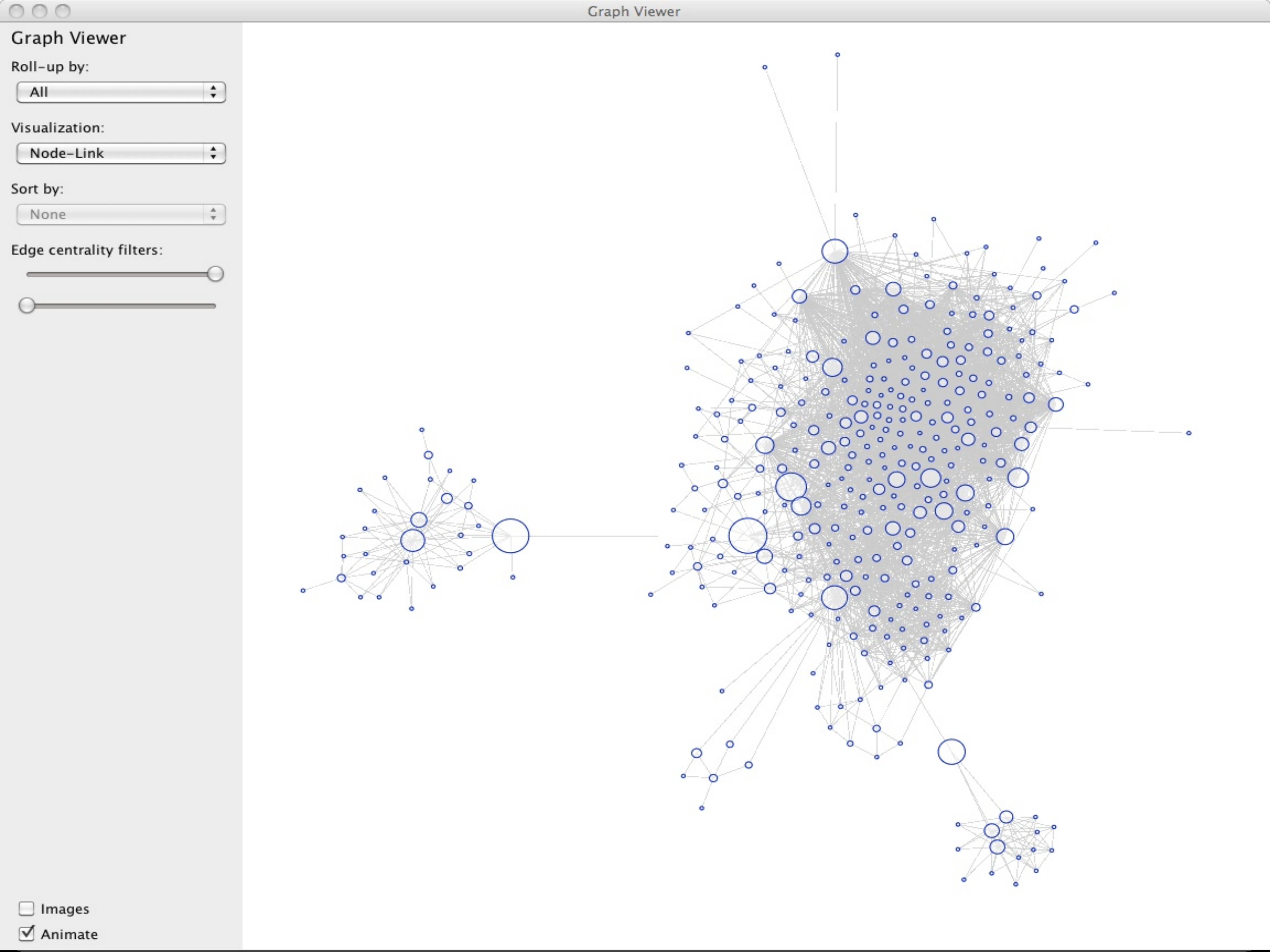
Interactive
^ Data Analysis

Jeffrey Heer
Stanford University


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate

Graph Viewer

Graph Viewer

Roll-up by:

All

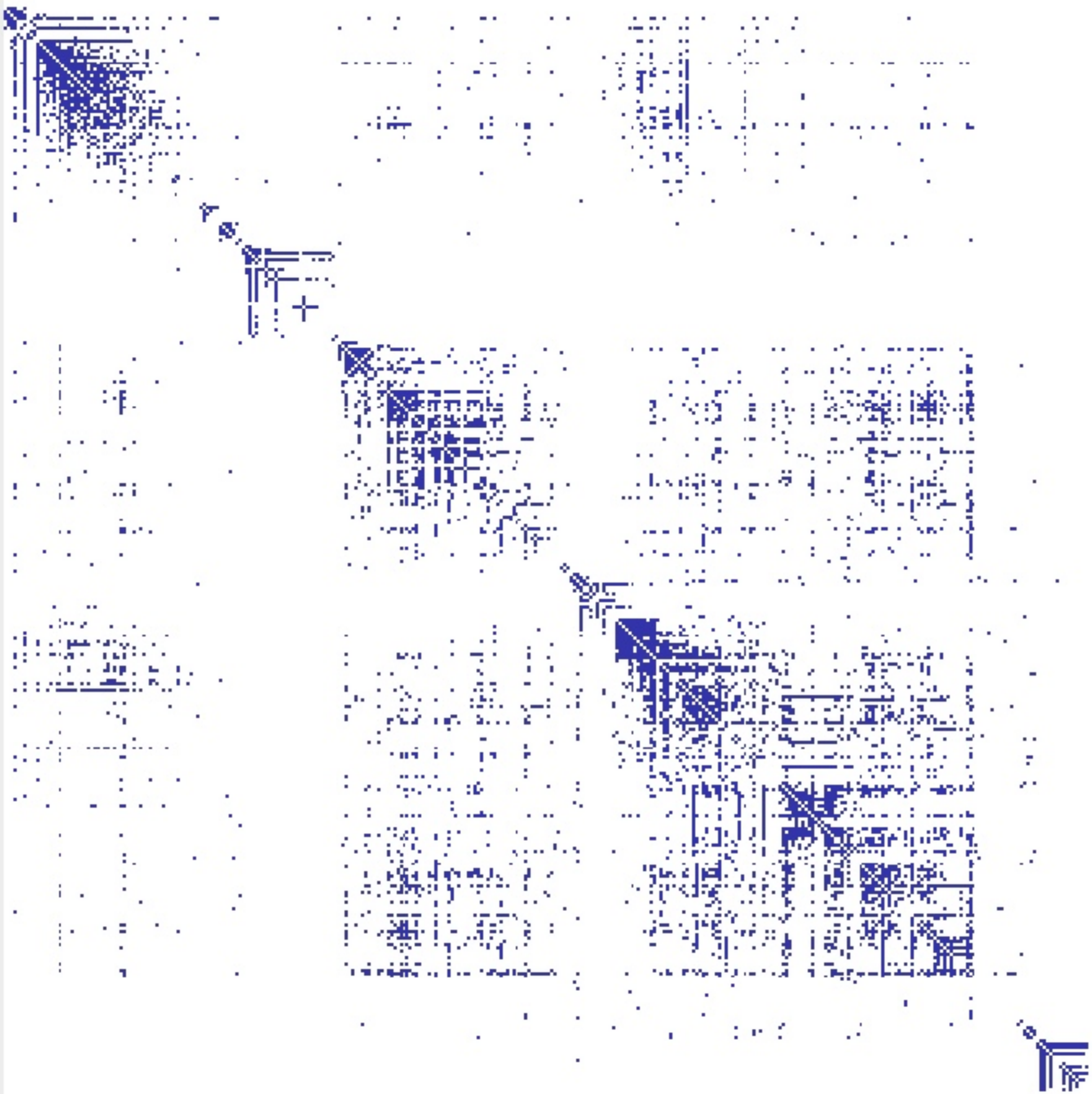
Visualization:

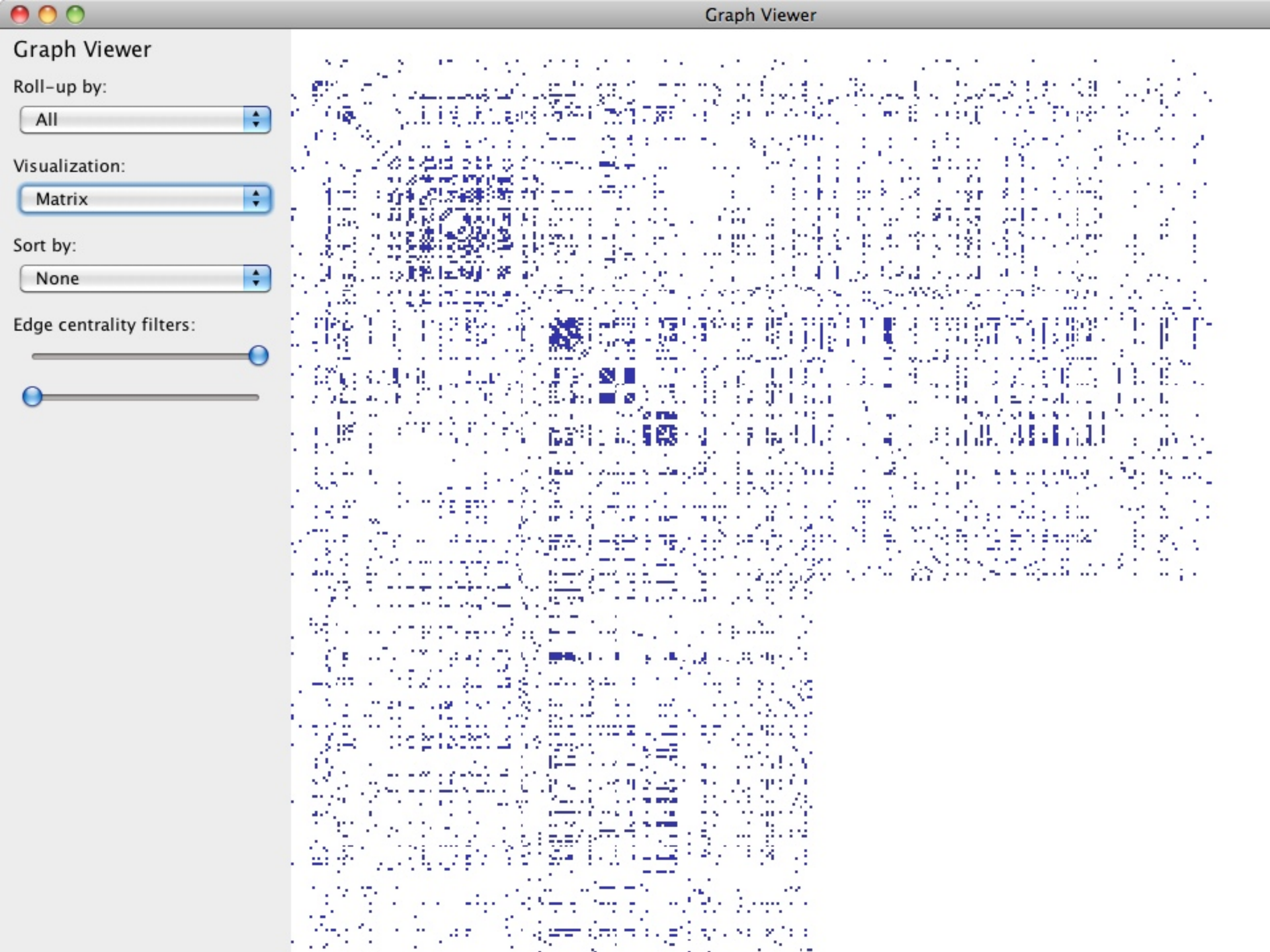
Matrix

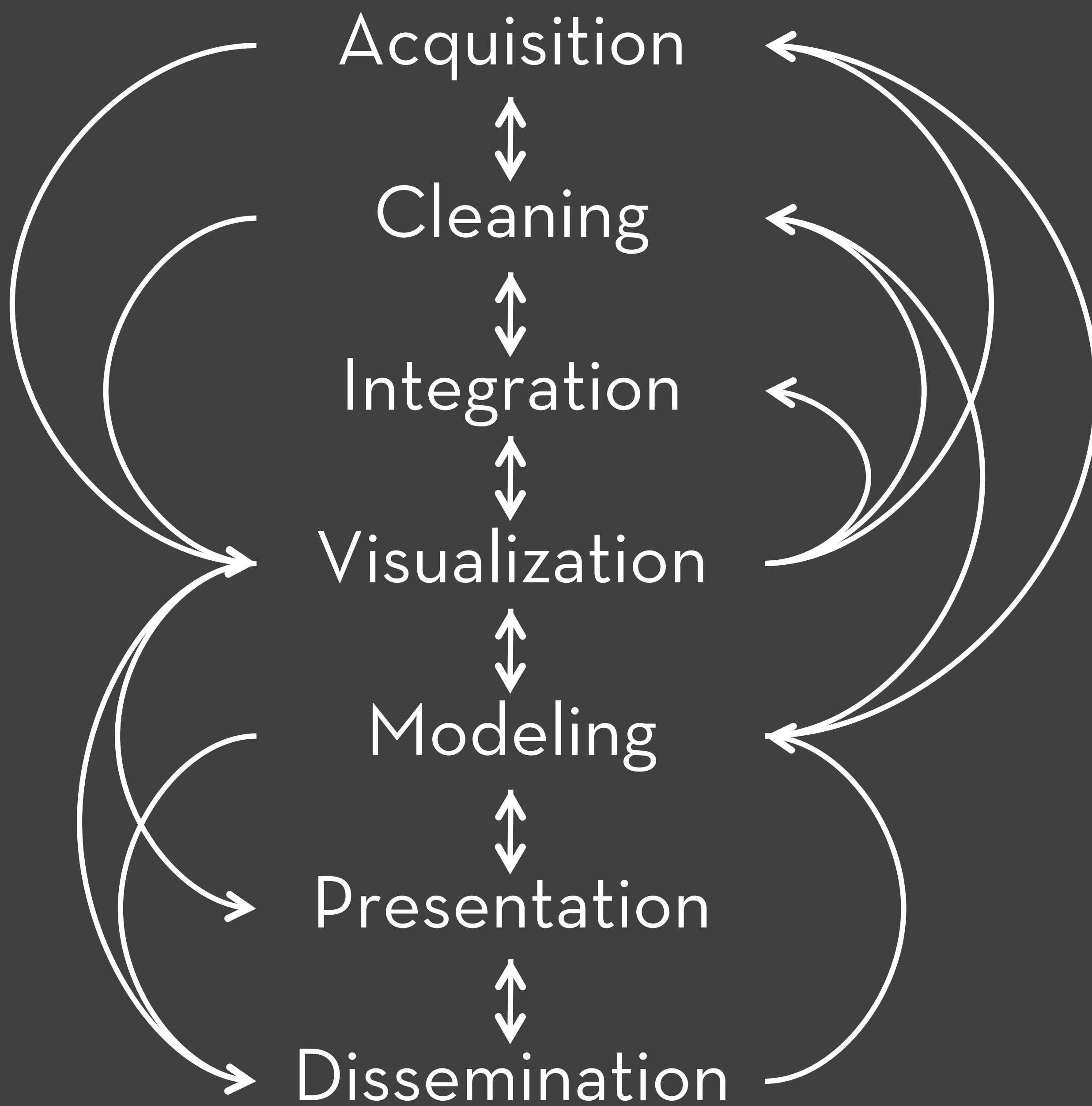
Sort by:

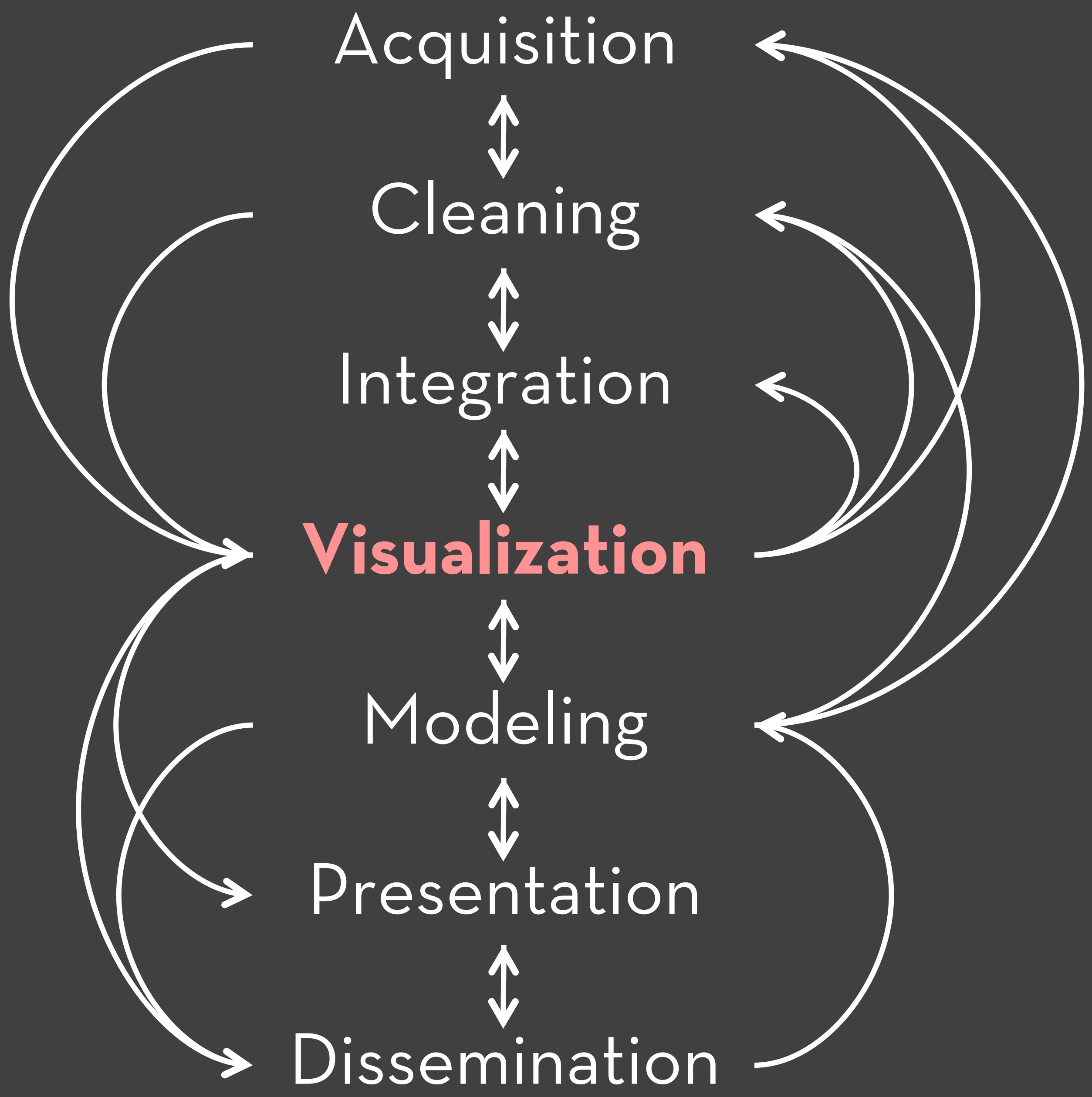
Linkage

Edge centrality filters:

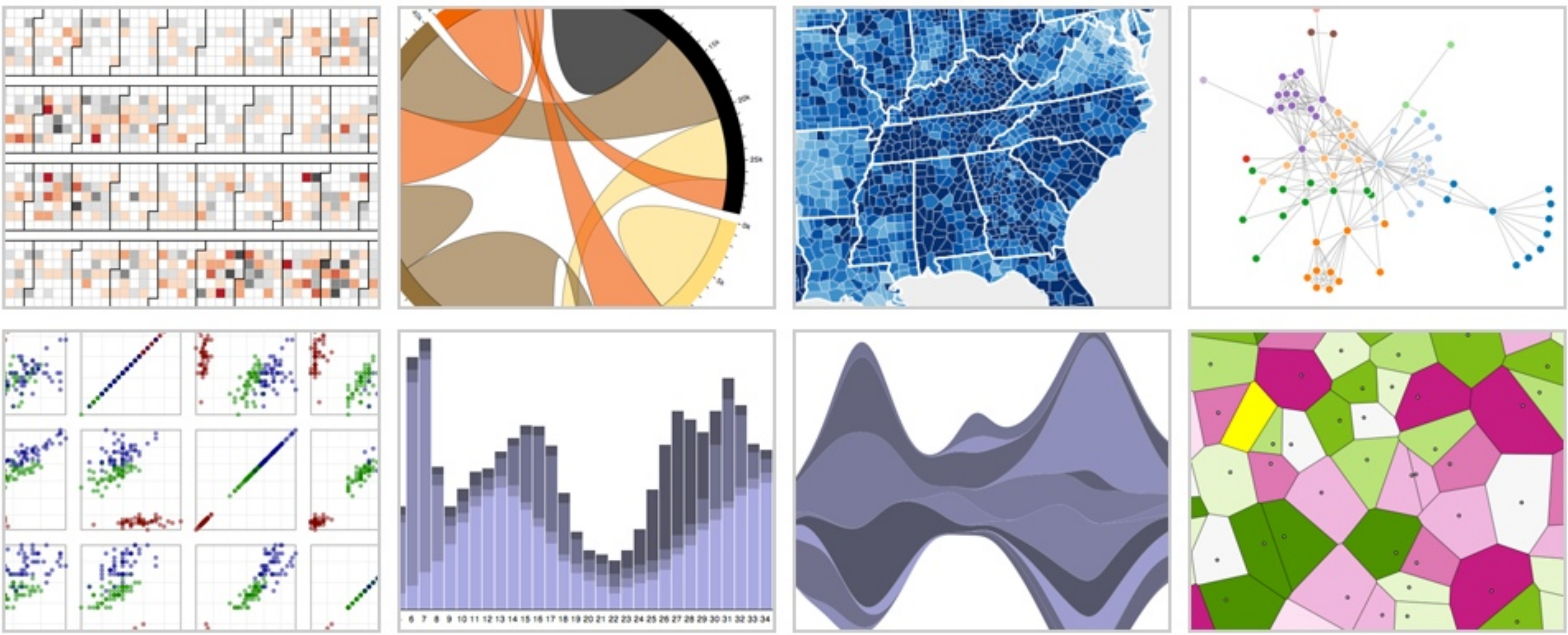



Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
None
Edge centrality filters:

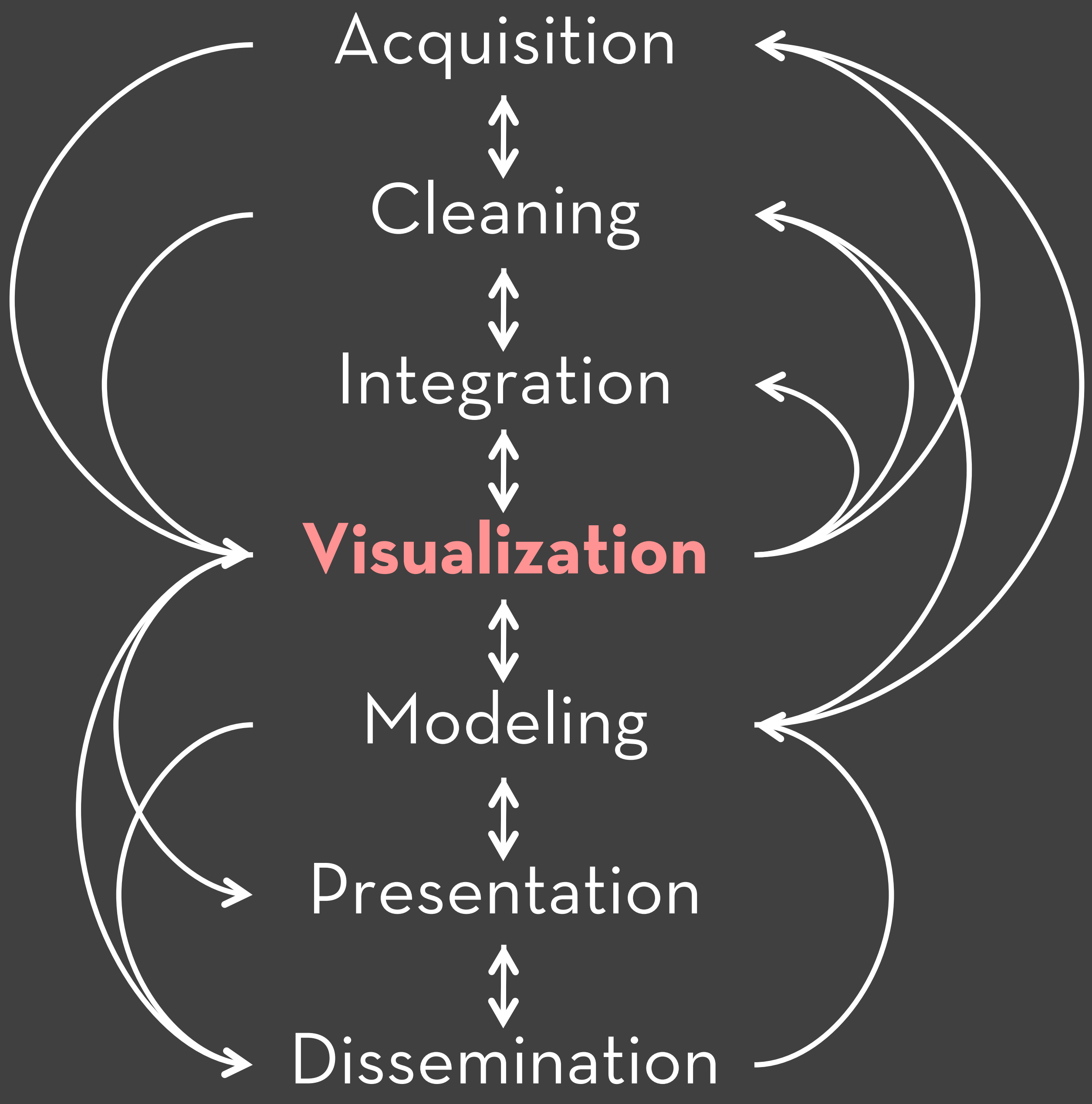

Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

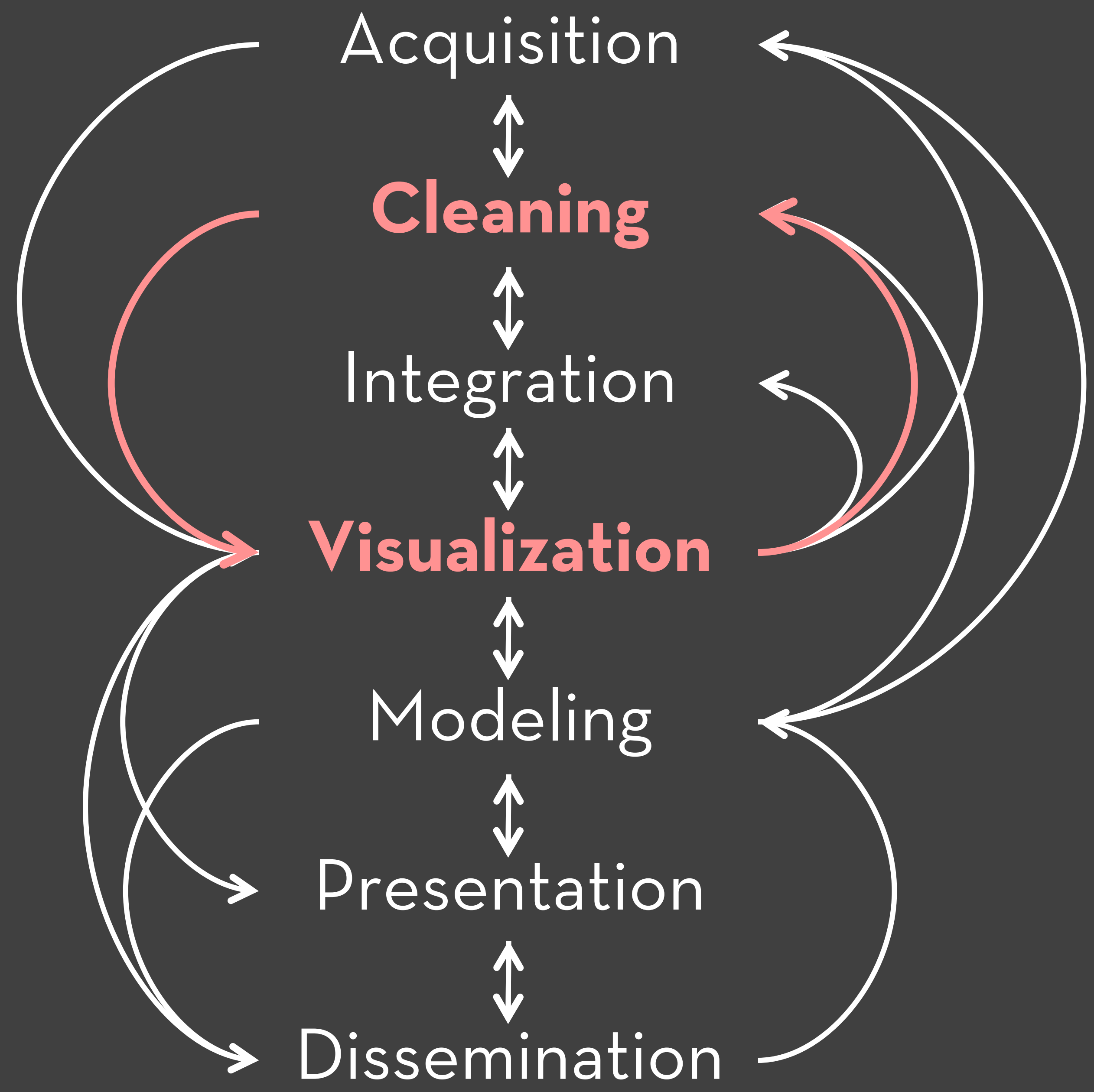

Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

d3.js Data-Driven Documents



with **Mike Bostock** & **Vadim Ogievetsky**


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900	955.8	2656	289
2006	4599030	3937	968.9	2645.1	322.9
2007	4627851	3974.9	980.2	2687	307.7
2008	4661900	4081.9	1080.7	2712.6	288.6

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755	3370.9	573.6	2456.7	340.6
2005	663253	3615	622.8	2601	391
2006	670053	3582	615.2	2588.5	378.3
2007	683478	3373.9	538.9	2480	355.1
2008	686293	2928.3	470.9	2219.9	237.5

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879	5073.3	991	3118.7	963.5
2005	5953007	4827	946.2	2958	922
2006	6166318	4741.6	953	2874.1	914.4
2007	6338755	4502.6	935.4	2780.5	786.7
2008	6500180	4087.3	894.2	2605.3	587.8

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

Reported crime in California

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	35842038	3423.9	686.1	2033.1	704.8
2005	36154147	3321	692.9	1915	712
2006	36457549	3175.2	676.9	1831.5	666.8
2007	36553215	3032.6	648.4	1784.1	600.2
2008	36756666	2940.3	646.8	1769.8	523.8

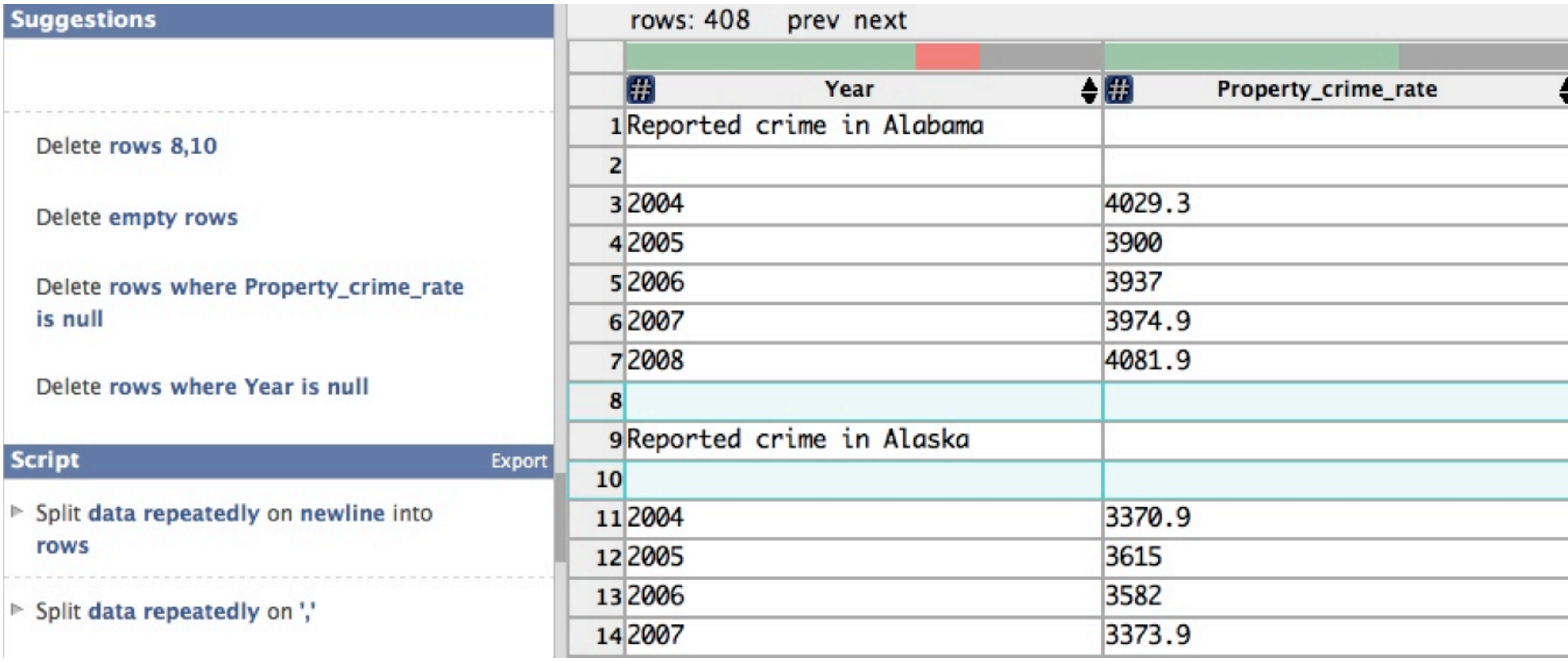
Reported crime in Colorado

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any "analysis" at all.

Anonymous Data Scientist
from our interview study, 2012

DataWrangler



with **Sean Kandel**, Philip Guo, Andreas Paepcke & Joe Hellerstein

Wrangler in 2 Parts…

1. Declarative data transformation language
   Tuple mapping – split, merge, extract, delete
   Reshaping – fold, unfold (cross-tabulation)
   Lookups & joins – e.g., FIPS code to US state
   Sorting, aggregation, *etc.*

Informed by prior work in databases:
Potter's Wheel, SchemaSQL, AJAX

Wrangler in 2 Parts…

1. Declarative data transformation language

+

2. Mixed-initiative interface for data transforms

   User: **Selects** data elements of interest

   System: **Suggests** applicable transforms via search over the space of viable transforms

   Enable rapid **preview and refinement**

Transform Suggestion

Interaction

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

Transform Suggestion

Interaction

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

- Text Selection
- Text Editing
- Row Selection
- Column Selection
- Transform Menu
- Click Quality Meter

Transform Suggestion

Interaction

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

Map user input to transform operands.

Example: text highlight maps to row, column, and text selections.

Inferred text selections include string indices and regular expressions.

Text Selection Inference

Series Id: LNU02000000

Text Selection Inference

```
Series Id: LNU02000000
```

-> ^ STR WS STR SYM WS STR NUM $

Text Selection Inference

```
Series Id: LNU02000000
```

-> ^ STR WS STR SYM WS STR NUM $

```
Series Id: LNU02000000
```

Text Selection Inference

`Series Id: LNU02000000`

-> ^ STR WS STR SYM WS STR NUM $

`Series Id: LNU02000000`

MATCH Indices 11-22

Text Selection Inference

`Series Id: LNU02000000`

-> ^ STR WS STR SYM WS STR NUM $

`Series Id: LNU02000000`

MATCH Indices 11-22

MATCH `LNU02000000`

Text Selection Inference

`Series Id: LNU02000000`

-> ^ STR WS STR SYM WS STR NUM $

`Series Id: LNU02000000`

MATCH Indices 11–22

MATCH `LNU02000000`

MATCH `LNU` NUM

MATCH STR NUM

Text Selection Inference

`Series Id: LNU02000000`

-> ^ STR WS STR SYM WS STR NUM $

`Series Id: LNU02000000`

MATCH Indices 11-22

MATCH LNU02000000

MATCH LNU NUM

MATCH STR NUM

AFTER : WS

Transform Suggestion

Interaction

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

Map user input to transform operands.

Example: text highlight maps to row, column, and text selections.

Inferred text selections include string indices and regular expressions.

Transform Suggestion

Interaction

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

Enumerate transforms that accept inferred operands as input.

Set unmatched params to default values.

Apply filter heuristics: No-ops, delete-all, and overly sparse outputs.

Transform Suggestion

Interaction

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

Sort transforms by:

- Toolbar selection
- Specification difficulty
- Frequency in corpus

Transform Suggestion

Interaction

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

Extract from unnamed_1 once between positions 17,25

Extract from unnamed_1 once on whitespace Alabama

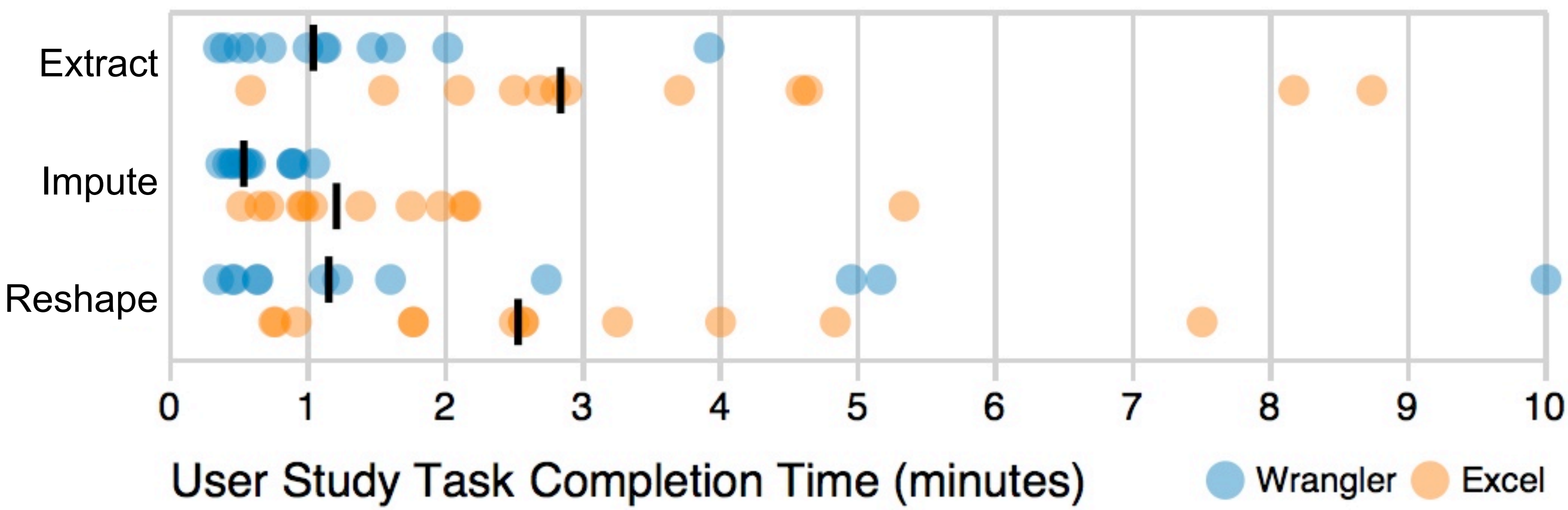
Cut from unnamed_1 once between positions 17,25

Cut from unnamed_1 once on whitespace Alabama

Split unnamed_1 once between positions 17,25 into columns

Split unnamed_1 once on whitespace Alabama into columns

Comparative Evaluation with Excel



Median completion time for Wrangler at least **twice as fast** in all tasks ($p < 0.001$).

Suggestions and visual previews used heavily.

Difficult Transforms: Table Reshaping

Fold

	Boys	Girls
Australia	1	2
Austria	3	4
Belgium	5	6
China	7	8

Australia	Boys	1
Australia	Girls	2
Austria	Boys	3
Austria	Girls	4
Belgium	Boys	5
Belgium	Girls	6
China	Boys	7
China	Girls	8

Pivot

Proactive Wrangling

Proactive transform suggestion [UIST'11]
Guide users to a proper relational table

Empty cells

Delimiters

$$S(T) = \left(1 - \frac{\sum_{c \in C} H_c(T)}{|C|}\right) + \frac{E(T) + D(T)}{|R|\,|C|}$$

Type homogeneity

$$H_c = \sum_{Type} \left(\frac{|i \in R : c_i \in Type|}{|R|}\right)^2$$

Proactive Wrangling

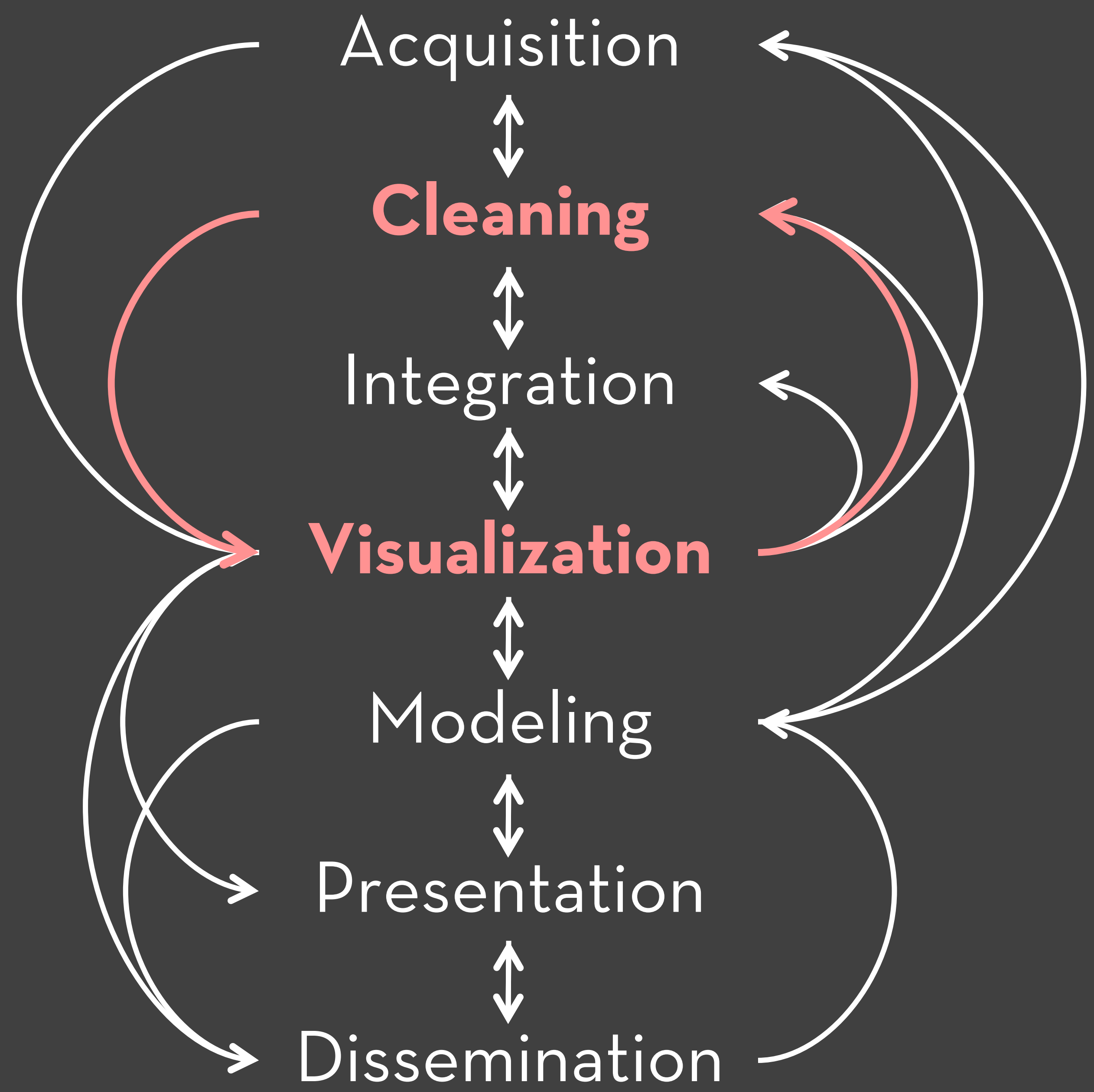
Proactive transform suggestion [UIST'11]
Guide users to a proper relational table

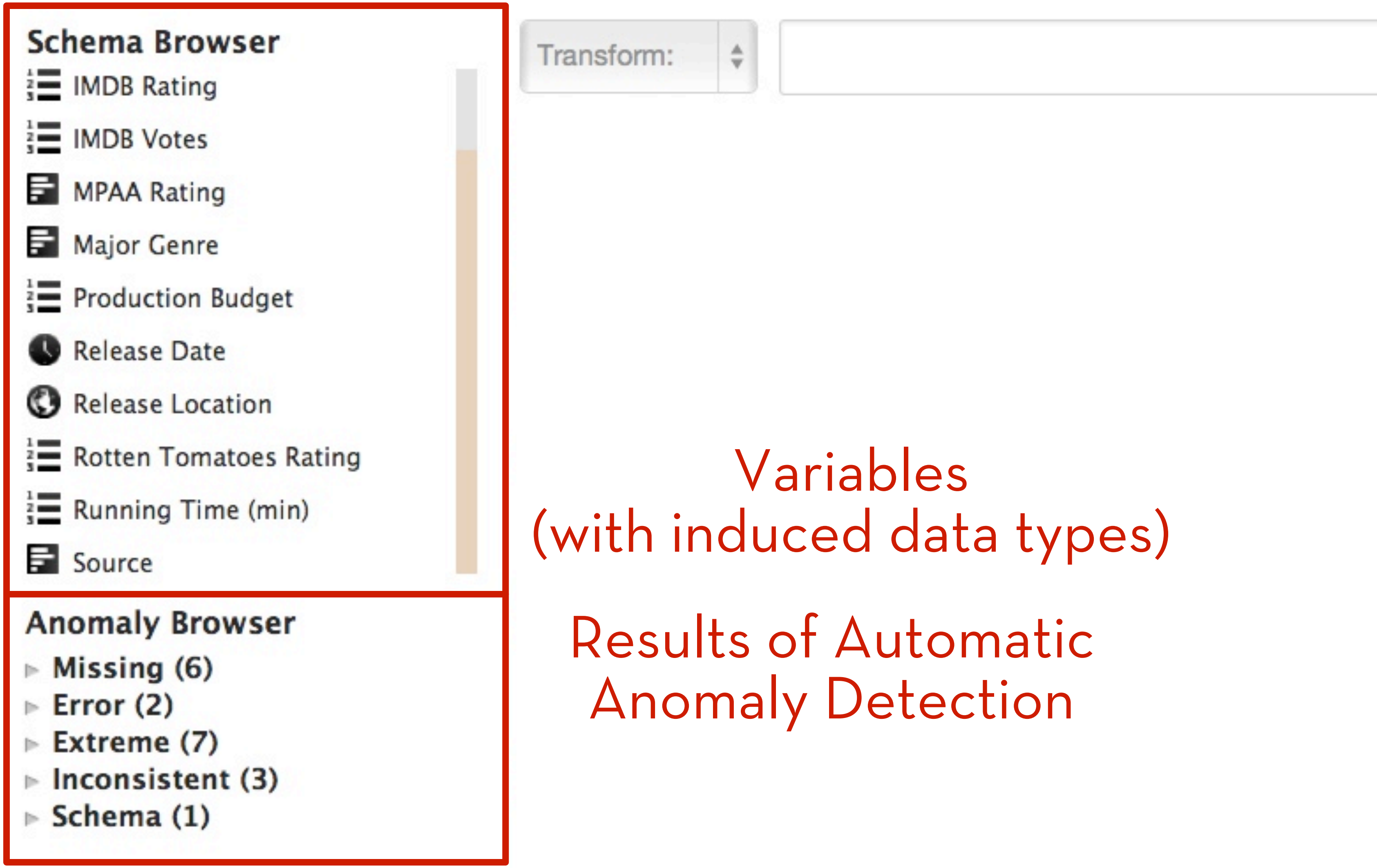
EVALUATION:

Compare automatic vs. manual transformation

53% of transforms automatically suggested

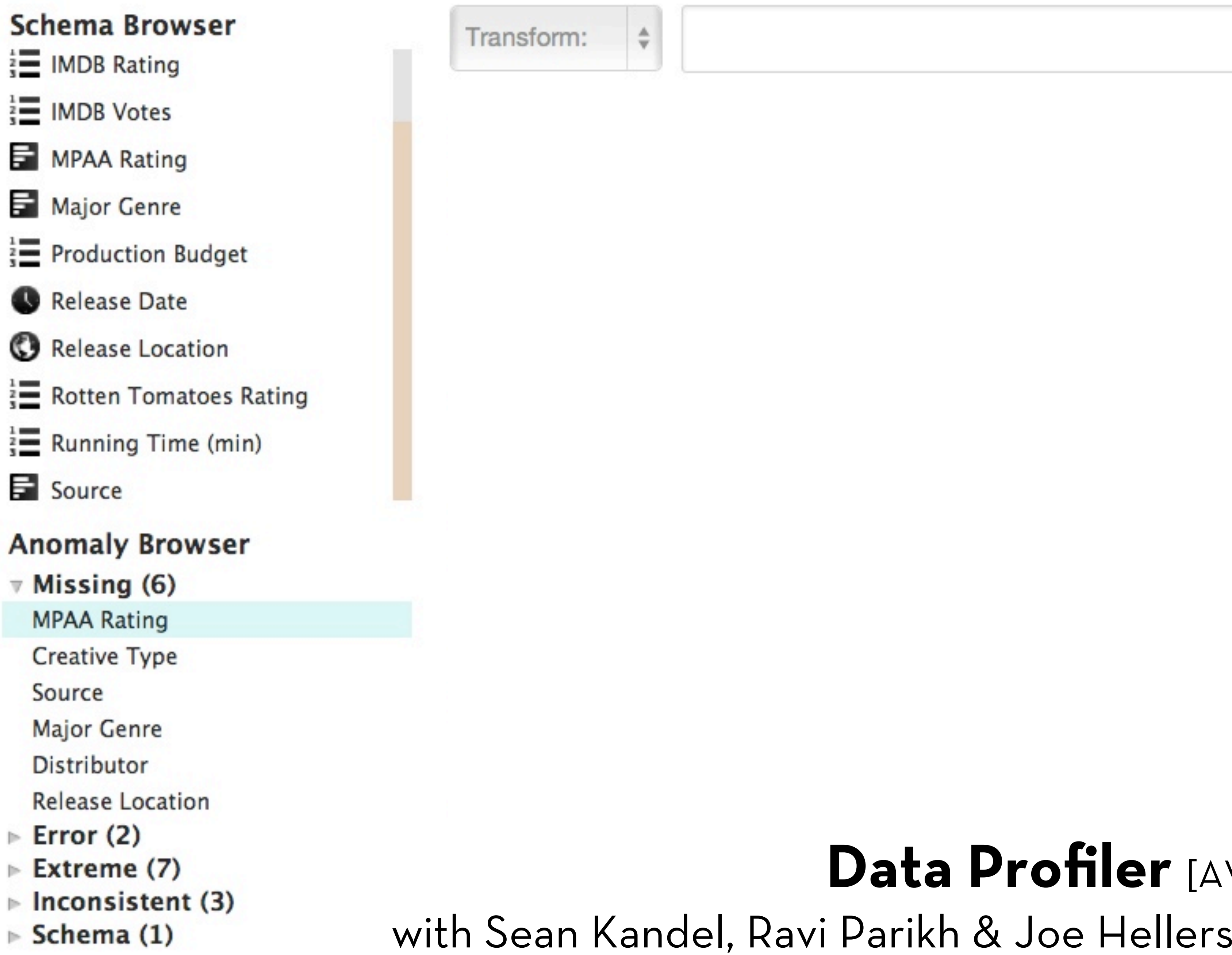
In those cases, the top-ranked suggestion is preferred **77%** of the time (**mean rank: 1.6**).


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination



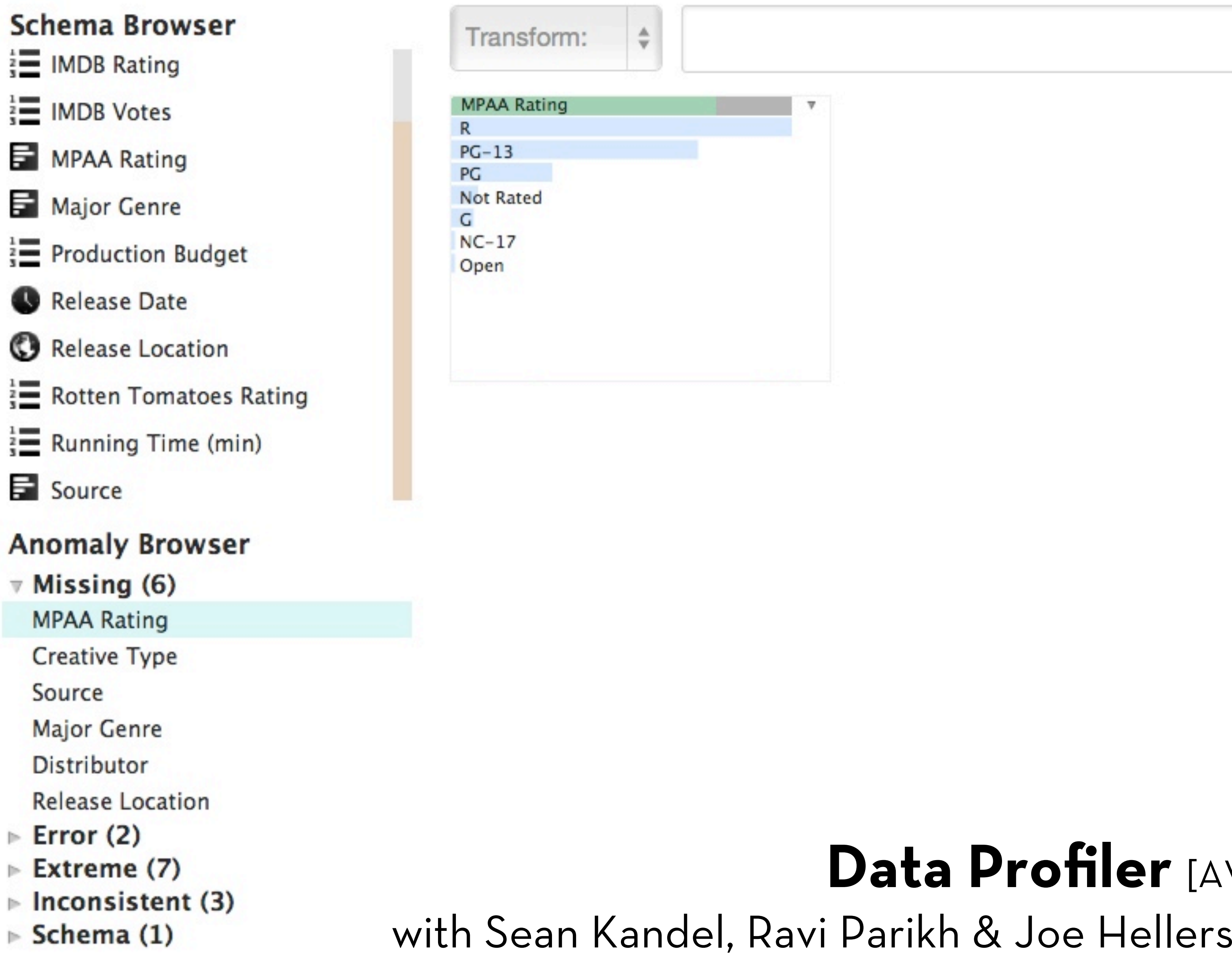
Data Profiler [AVI'12]

with Sean Kandel, Ravi Parikh & Joe Hellerstein



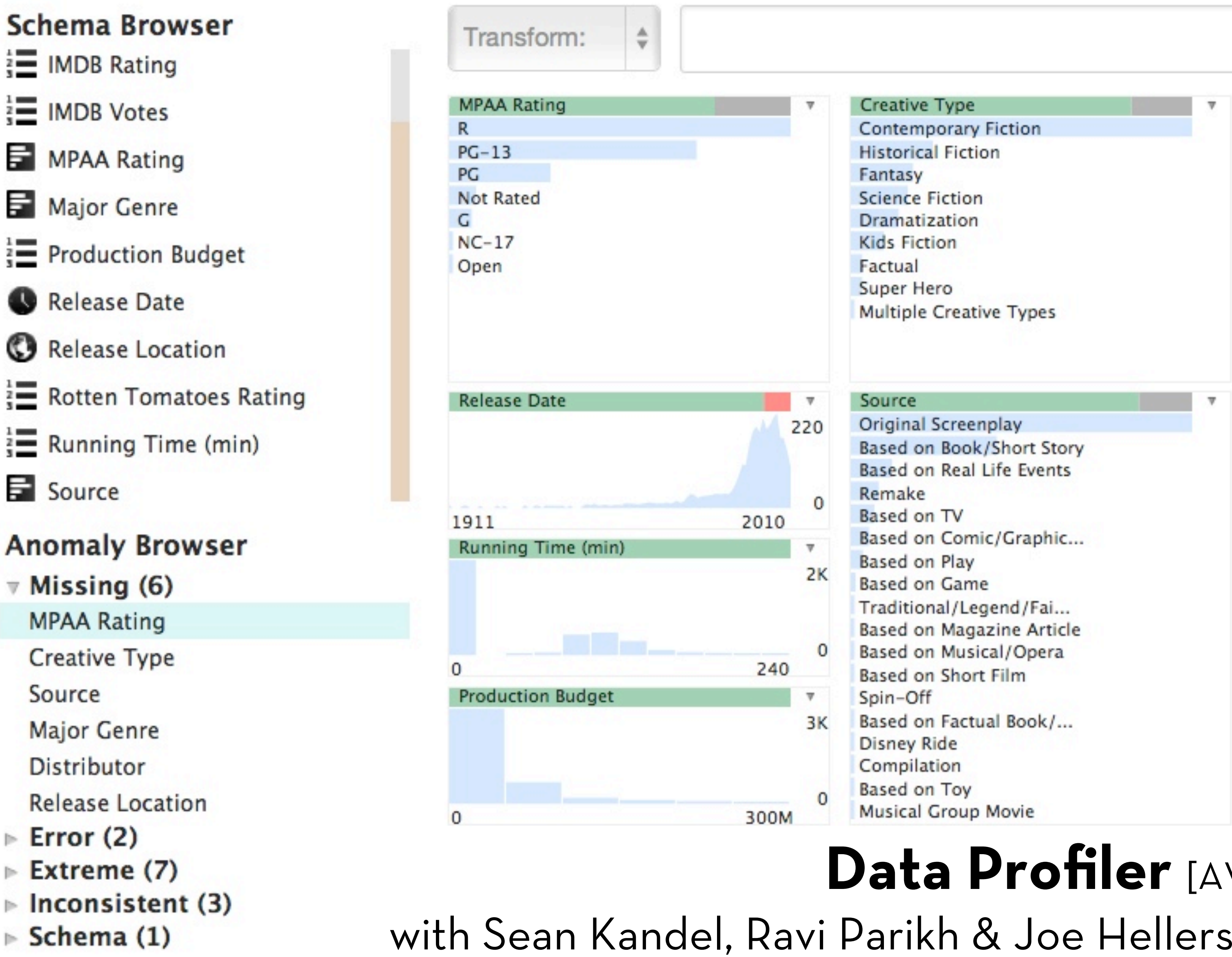
Data Profiler [AVI'12]

with Sean Kandel, Ravi Parikh & Joe Hellerstein



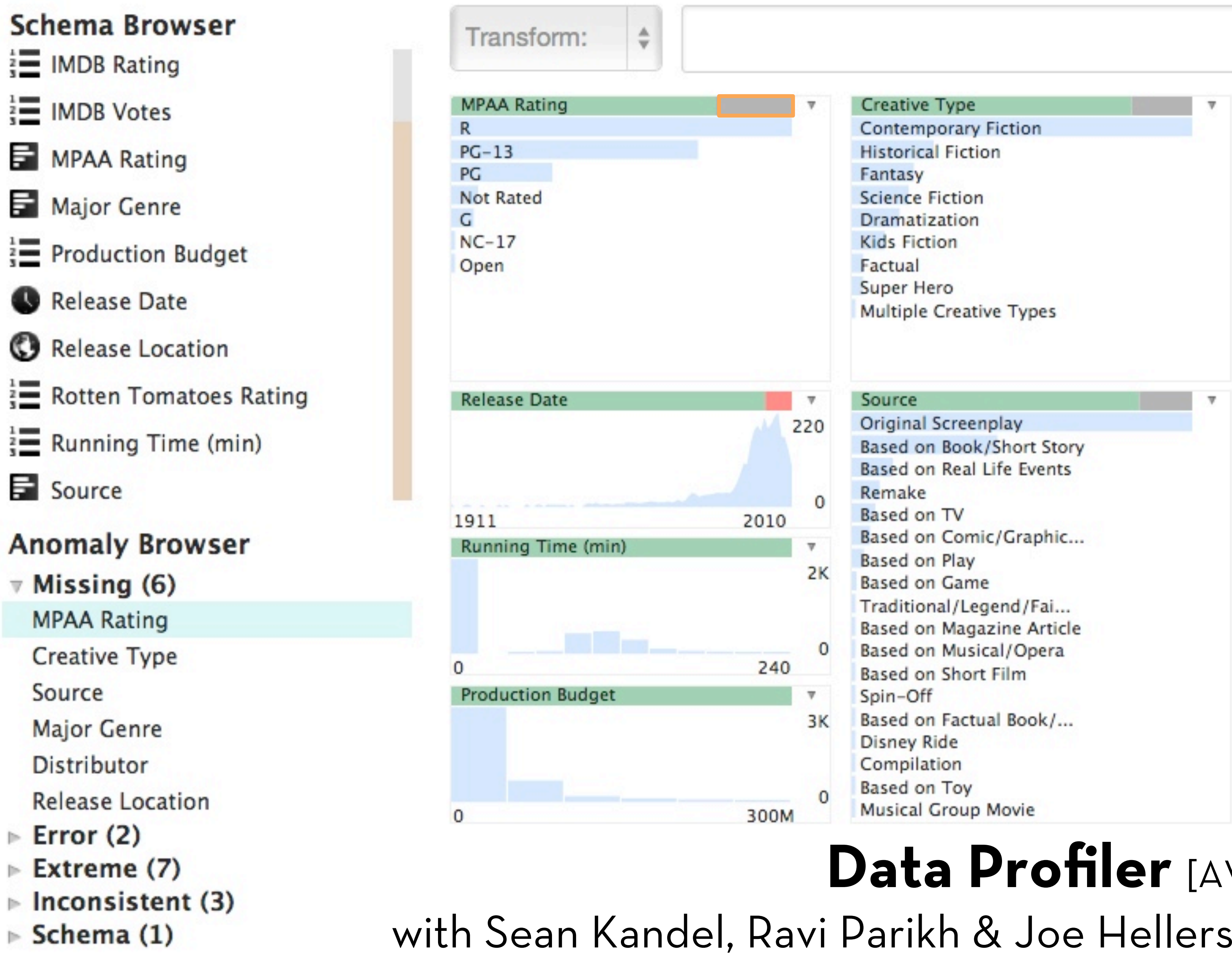
Data Profiler [AVI'12]

with Sean Kandel, Ravi Parikh & Joe Hellerstein



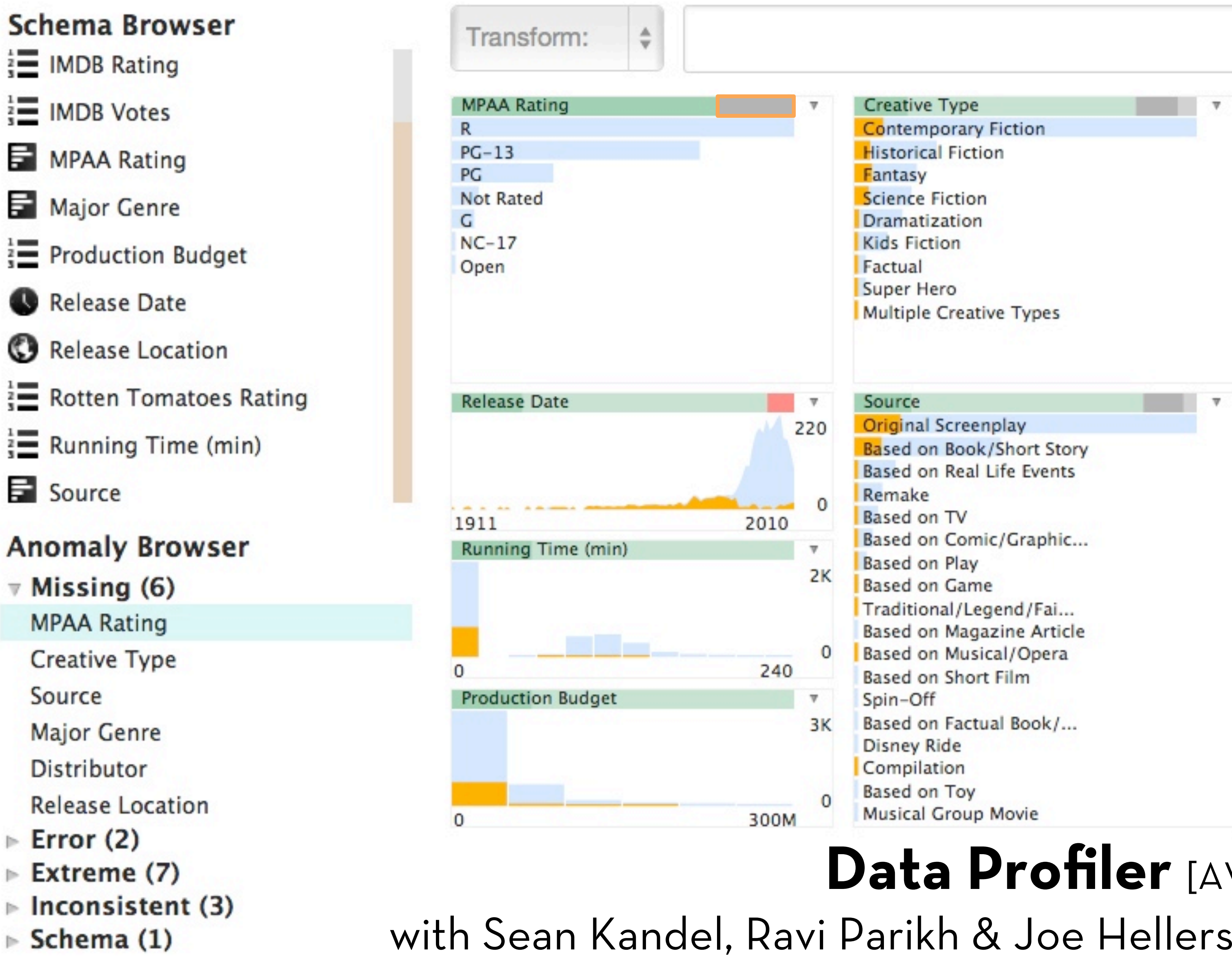
Data Profiler [AVI'12]

with Sean Kandel, Ravi Parikh & Joe Hellerstein



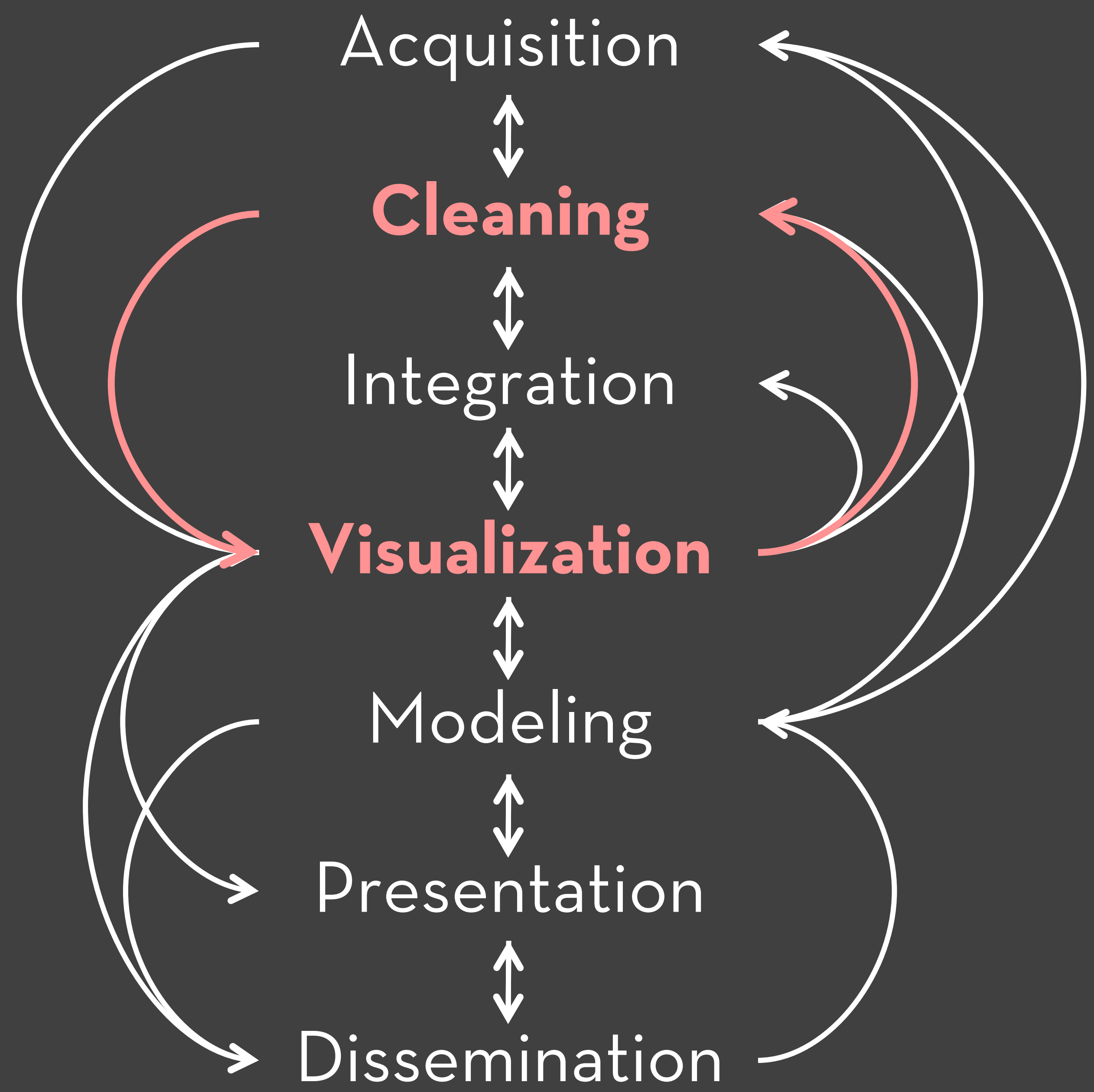
Data Profiler [AVI'12]

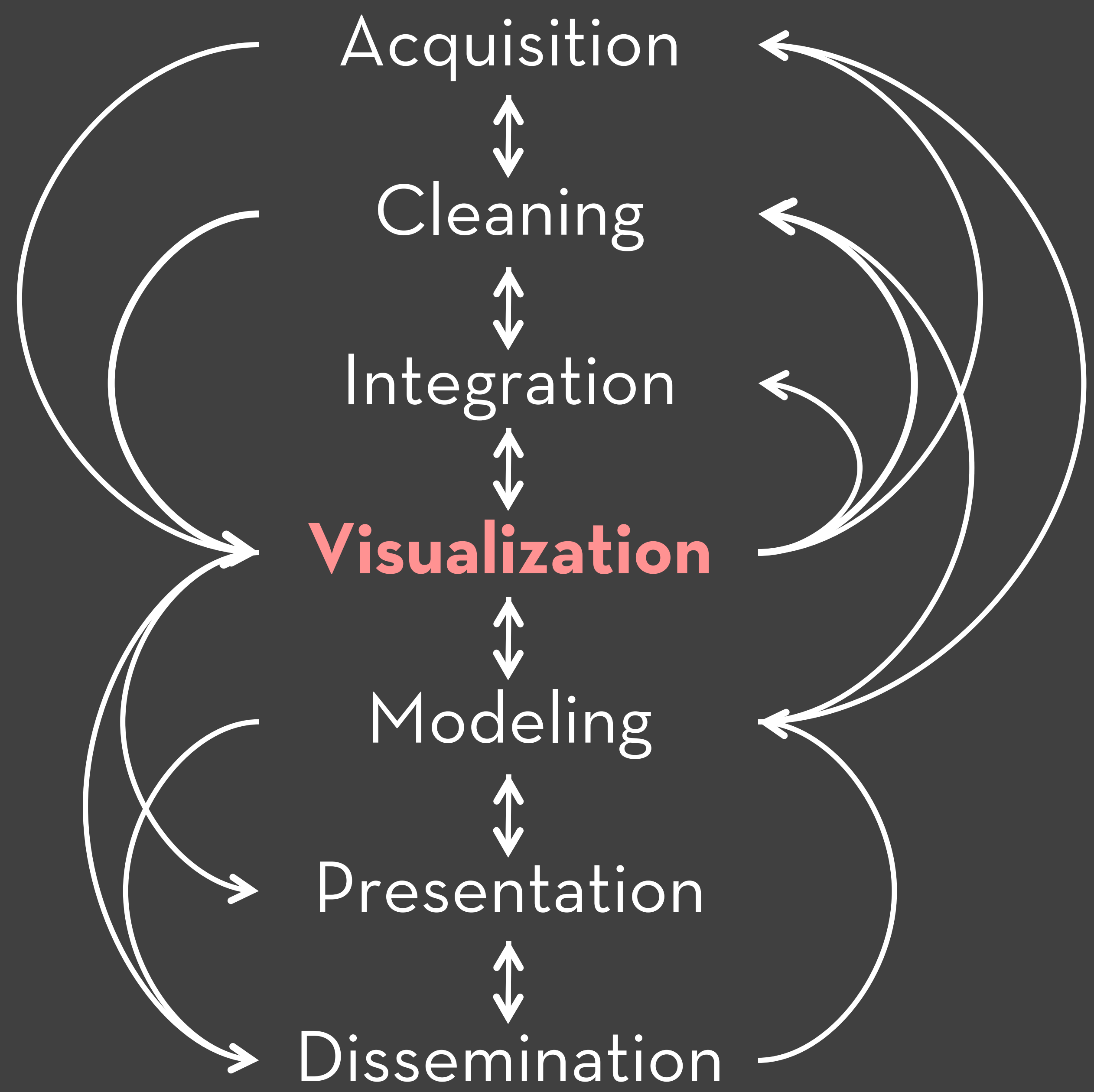
with Sean Kandel, Ravi Parikh & Joe Hellerstein

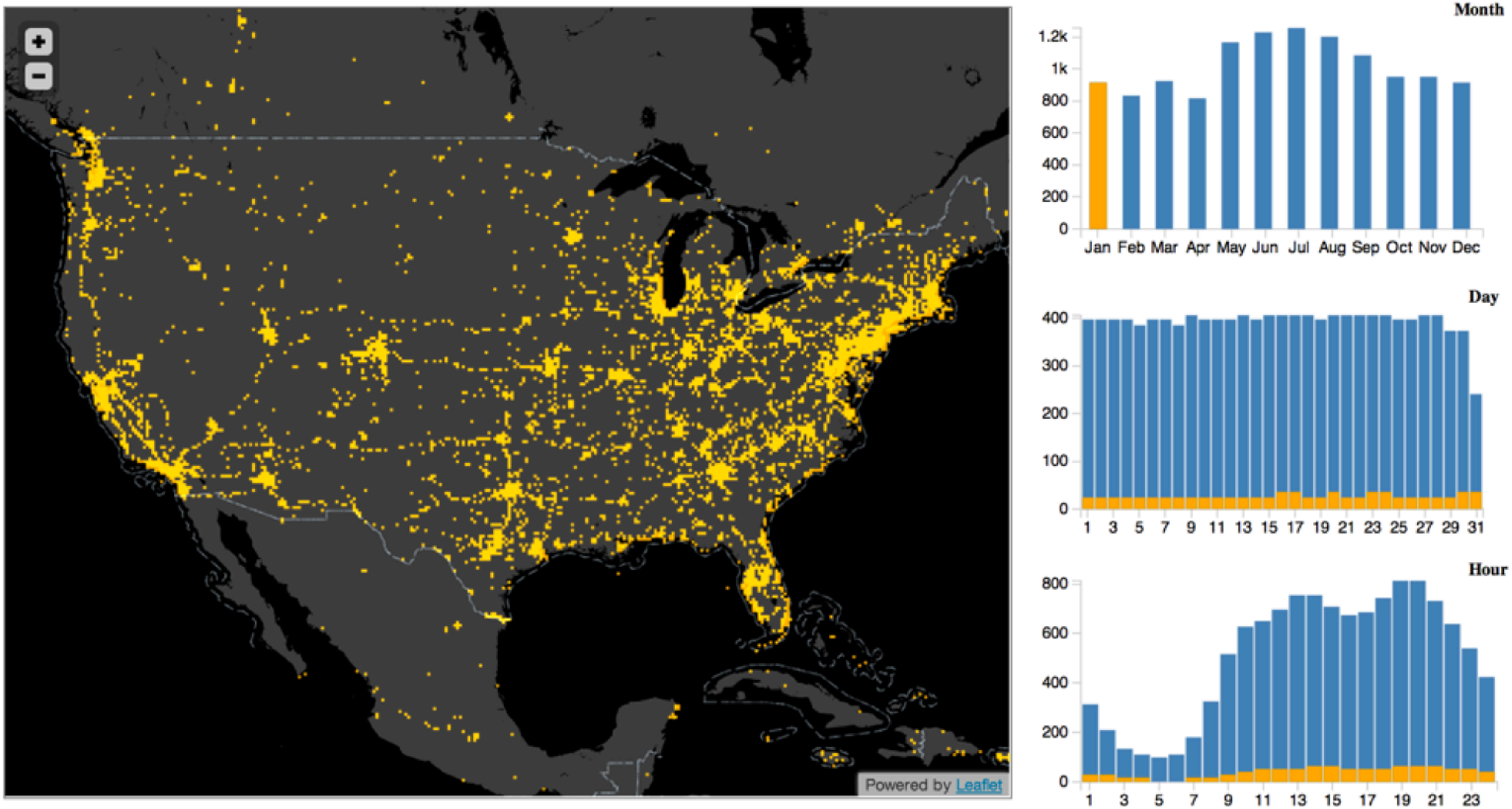


Data Profiler [AVI'12]

with Sean Kandel, Ravi Parikh & Joe Hellerstein


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

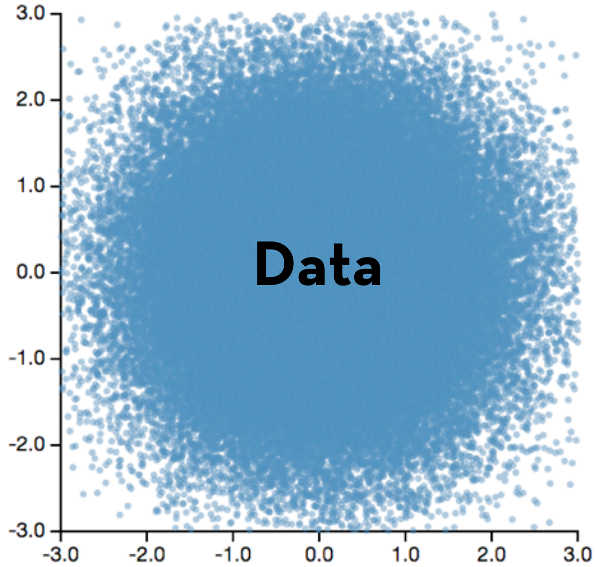

Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination



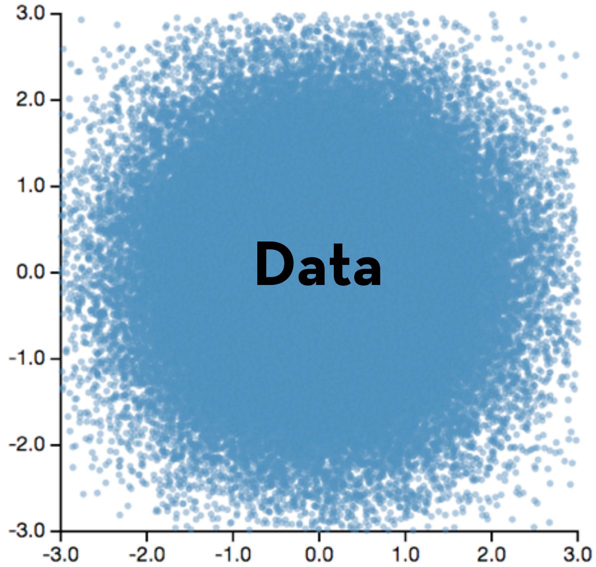
imMens: Real-Time Visual Querying of Big Data

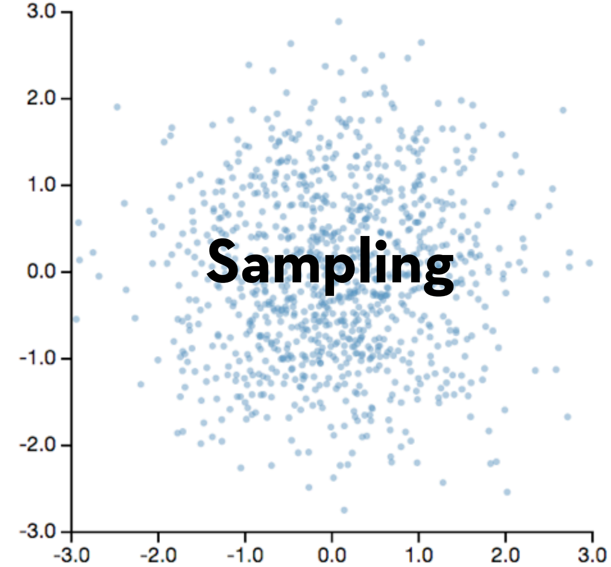
with Zhicheng (Leo) Liu & Biye Jiang

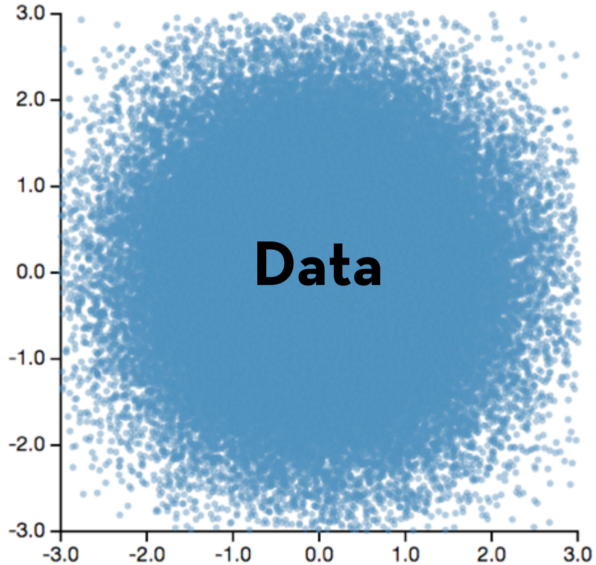
Perceptual and interactive scalability should be limited by the chosen resolution of the visualized data, not the number of records.

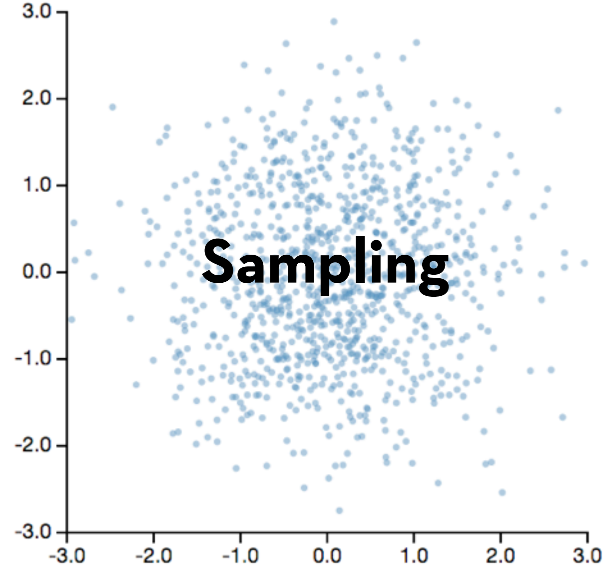


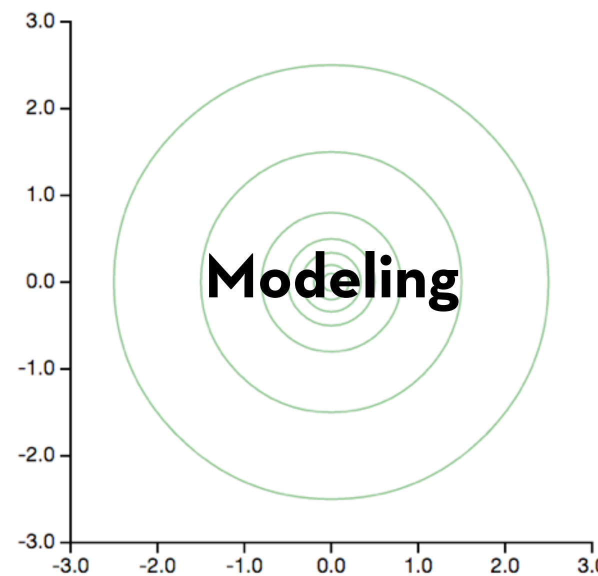
Data
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

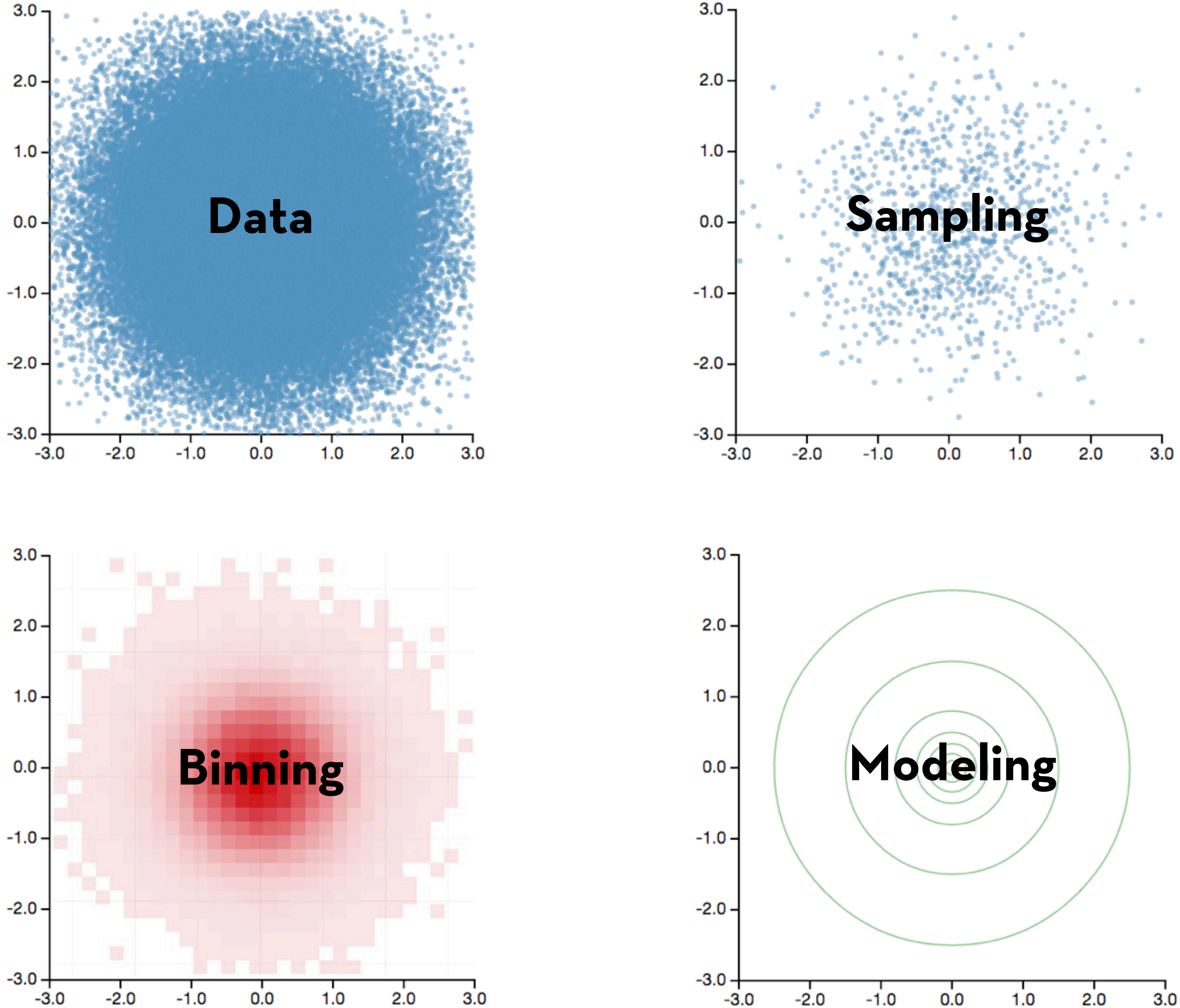

3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
Data
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

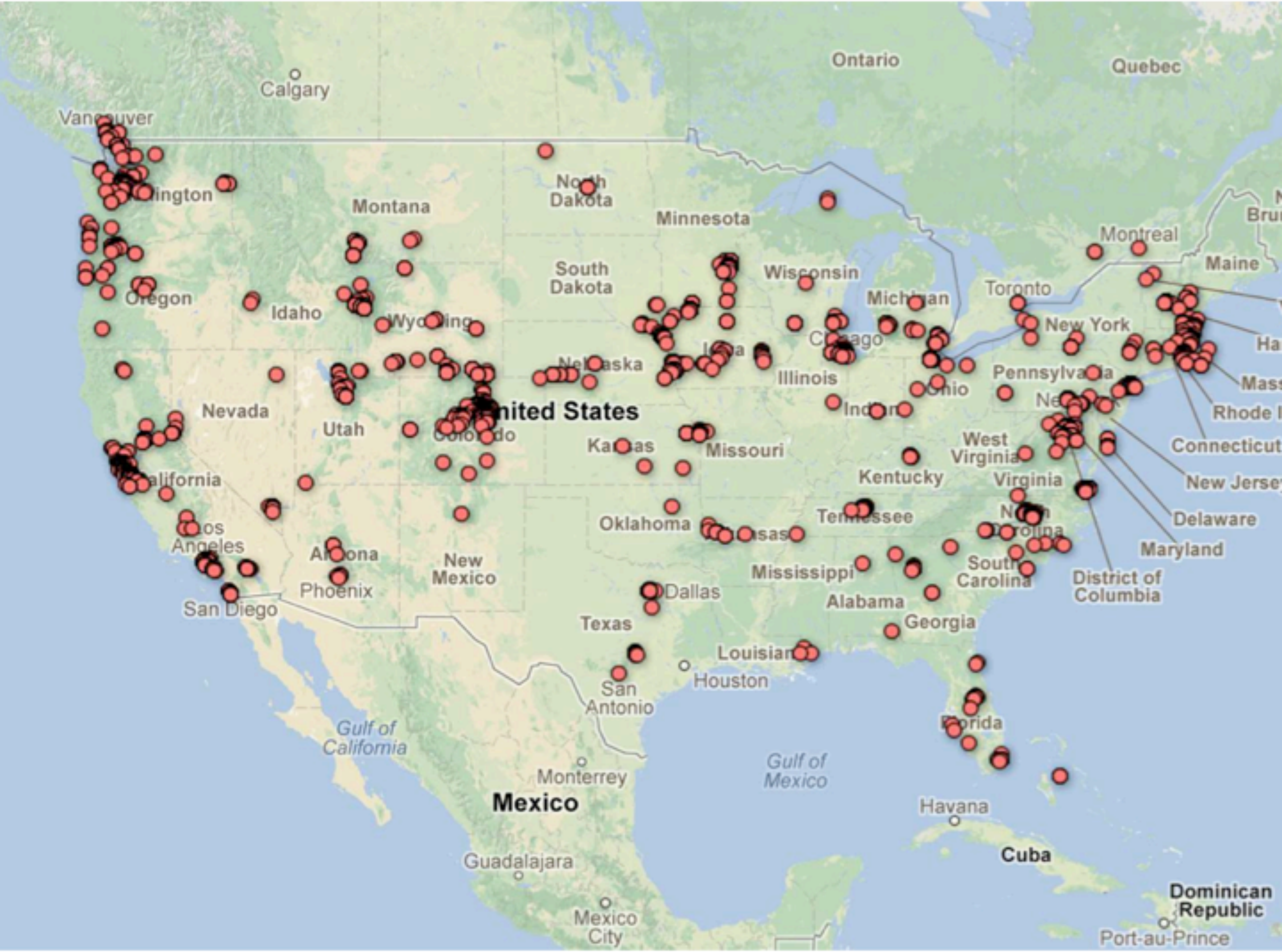

3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
Sampling
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0


Data
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

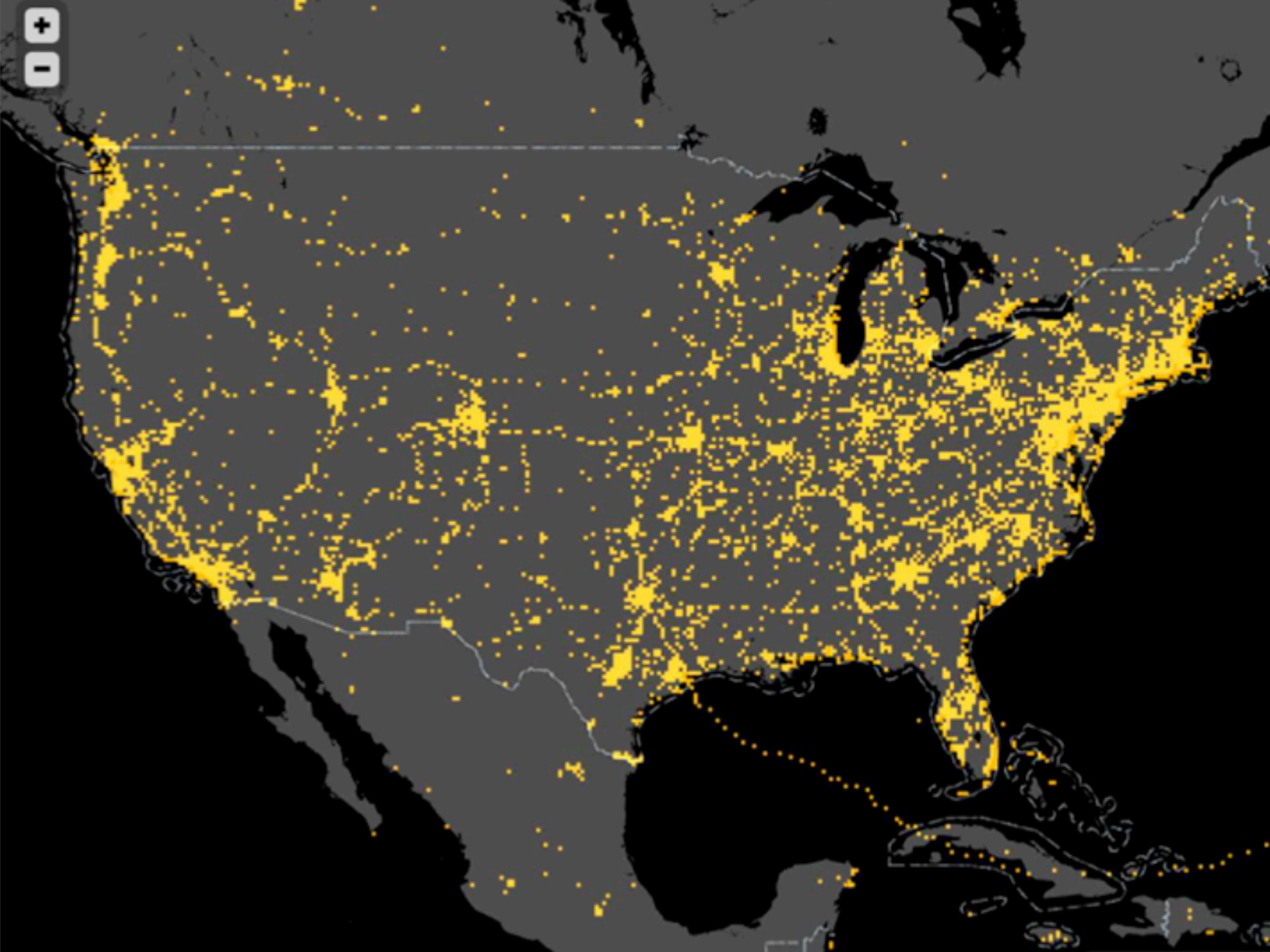

Sampling
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

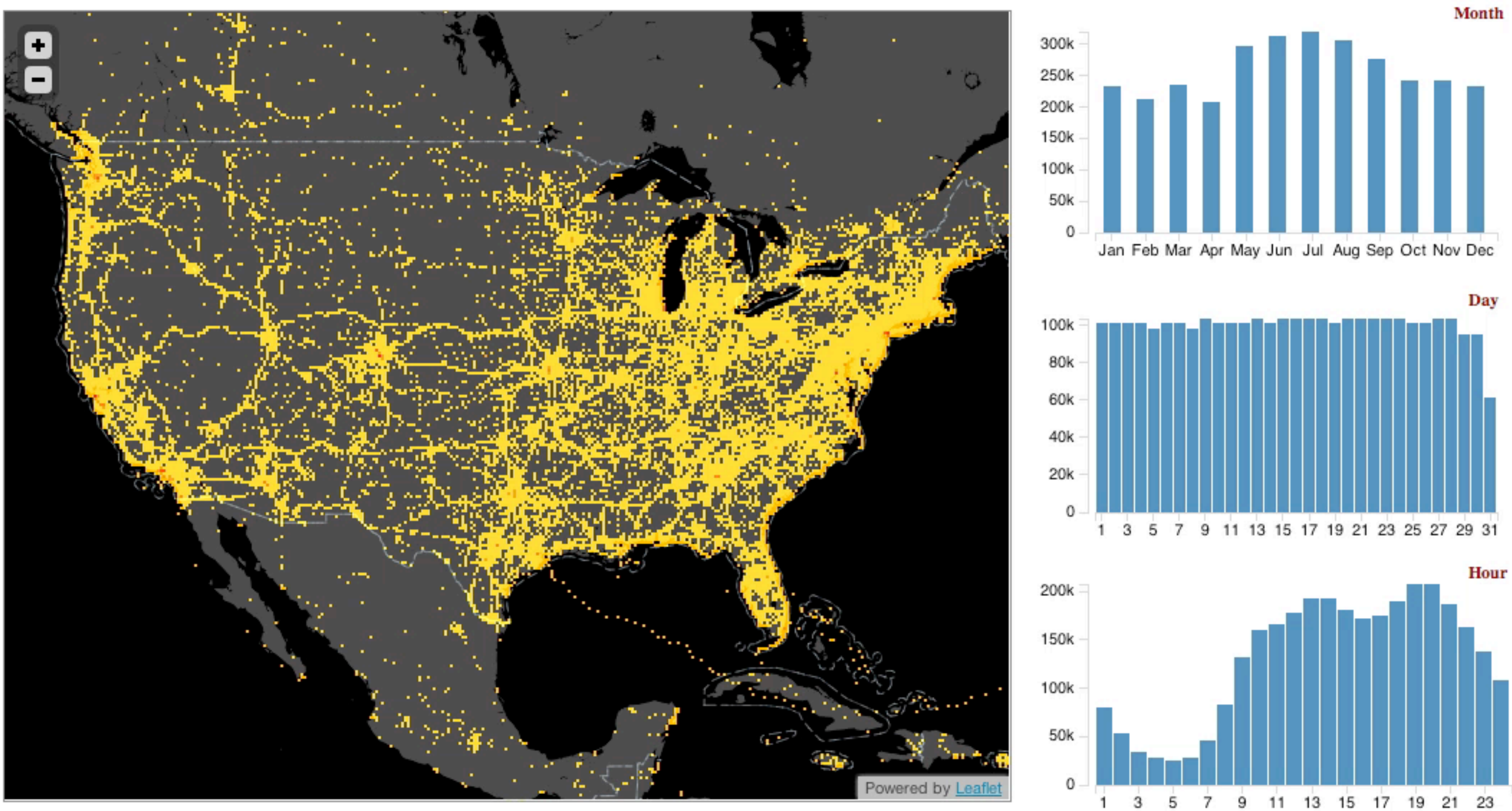

Modeling
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

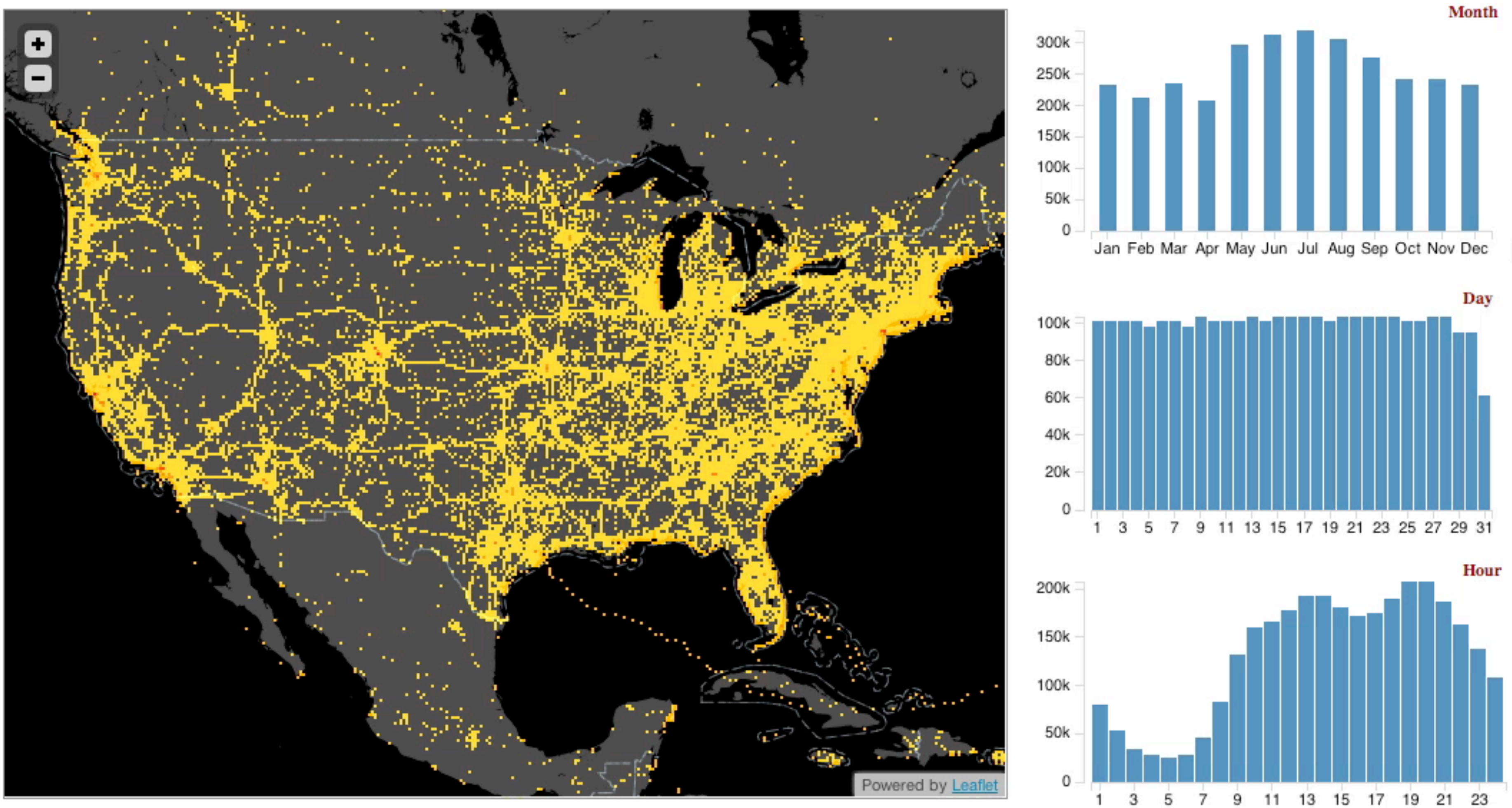

Data
Sampling
Binning
Modeling
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0 -2.0 -1.0 0.0 1.0 2.0 3.0



Ontario
Quebec
Calgary
Montana
North Dakota
Minnesota
South Dakota
Idaho
Toronto
New York
Montreal
Maine
Illinois
Missouri
Nevada
Utah
Kentucky
West Virginia
Virginia
Oklahoma
Mississippi
Alabama
Georgia
South Carolina
Texas
Dallas
Houston
San Antonio
Phoenix
San Diego
New Mexico
Connecticut
Delaware
Maryland
District of Columbia
Gulf of California
Gulf of Mexico
Monterrey
Mexico
Havana
Cuba
Guadalajara
Mexico City
Dominican Republic
Port-au-Prince



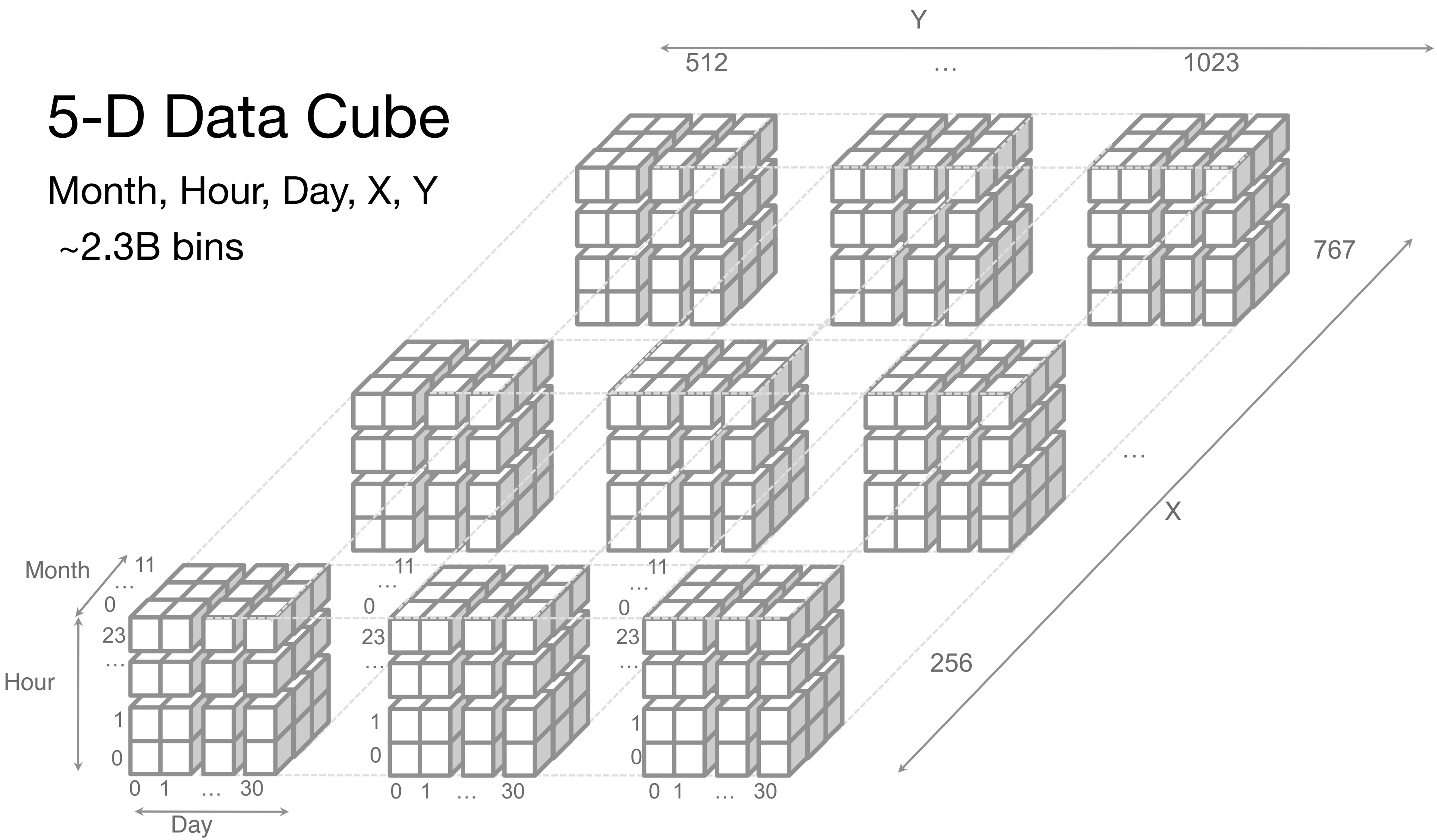

Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23
Powered by Leaflet


Powered by Leaflet
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

5-D Data Cube

Month, Hour, Day, X, Y

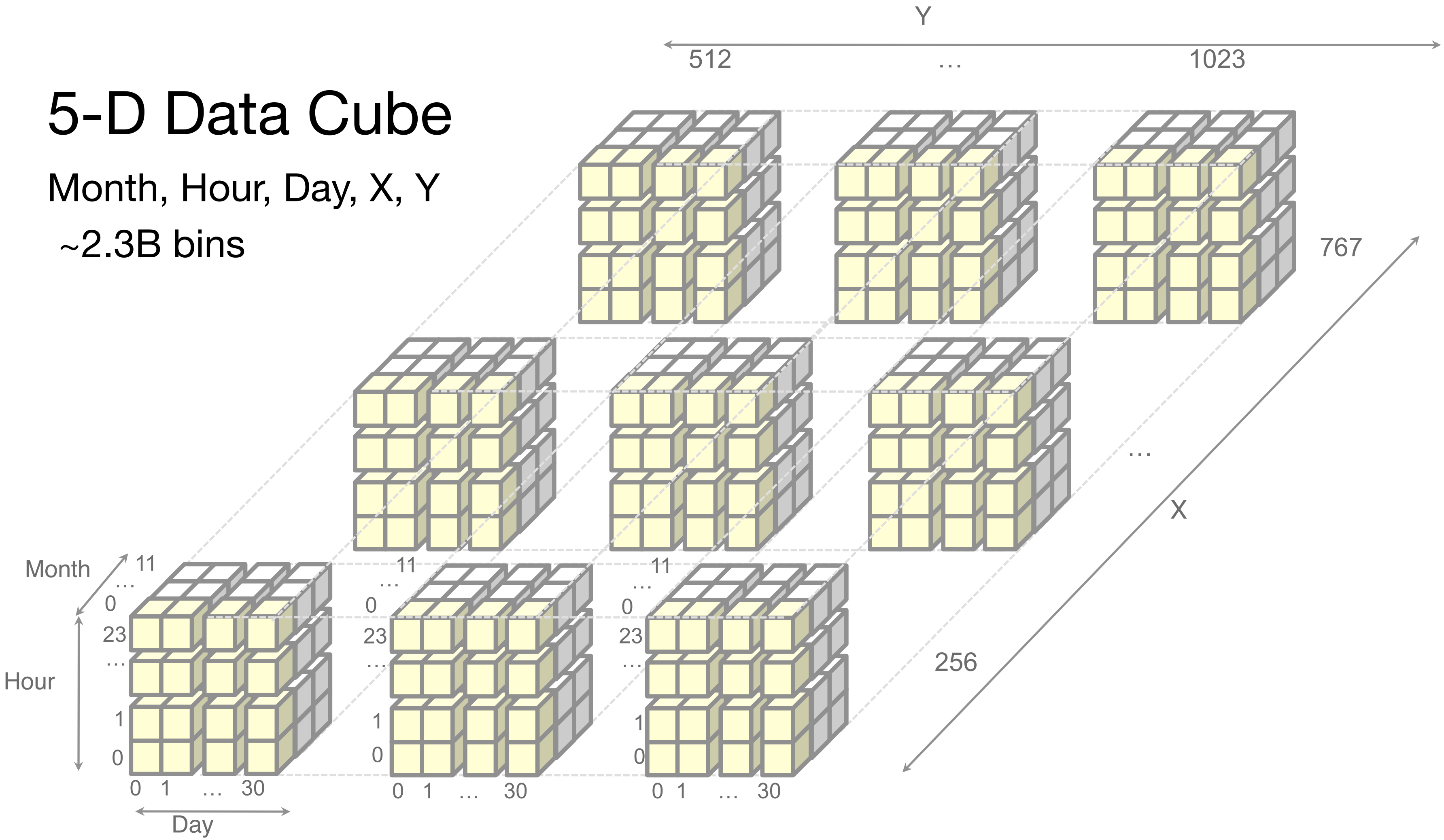
~2.3B bins

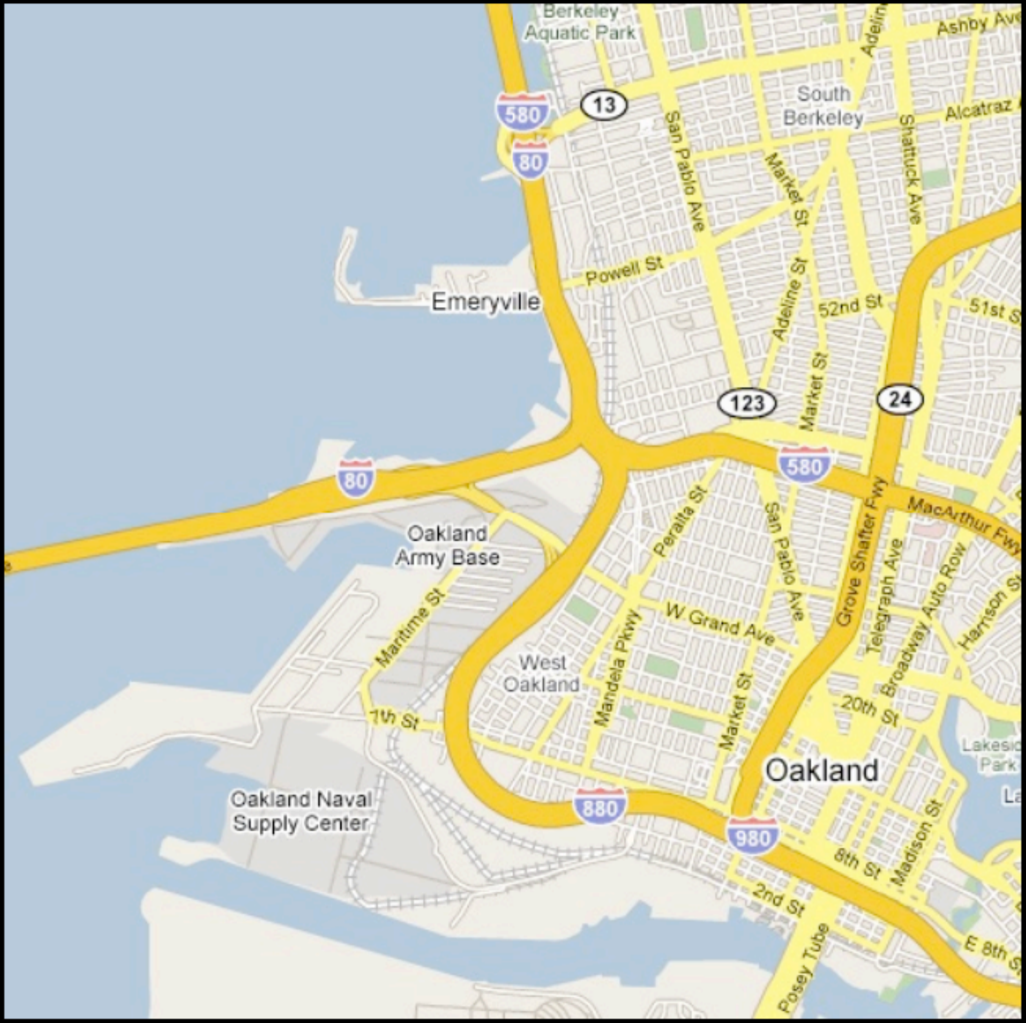


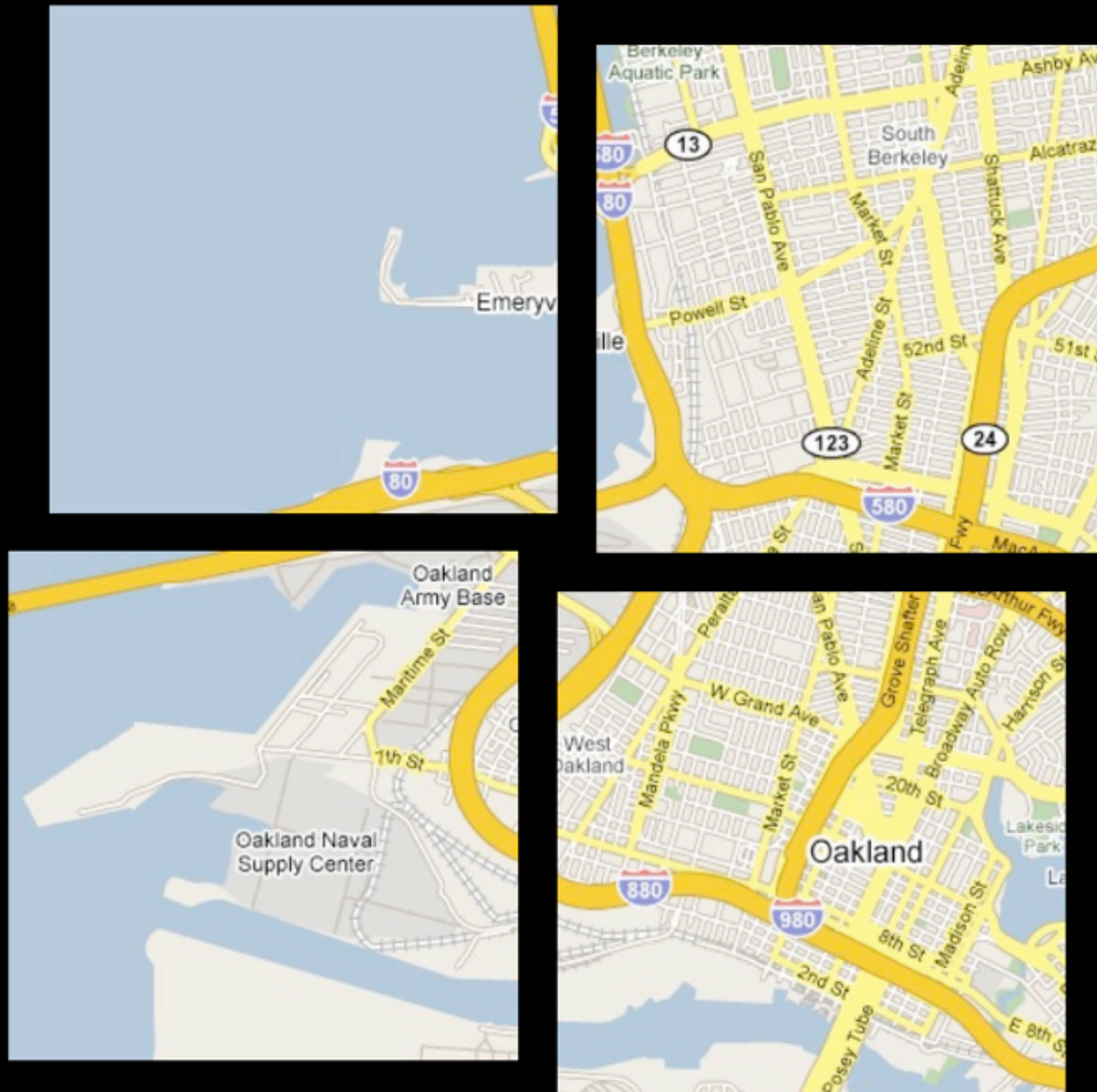
5-D Data Cube

Month, Hour, Day, X, Y

~2.3B bins

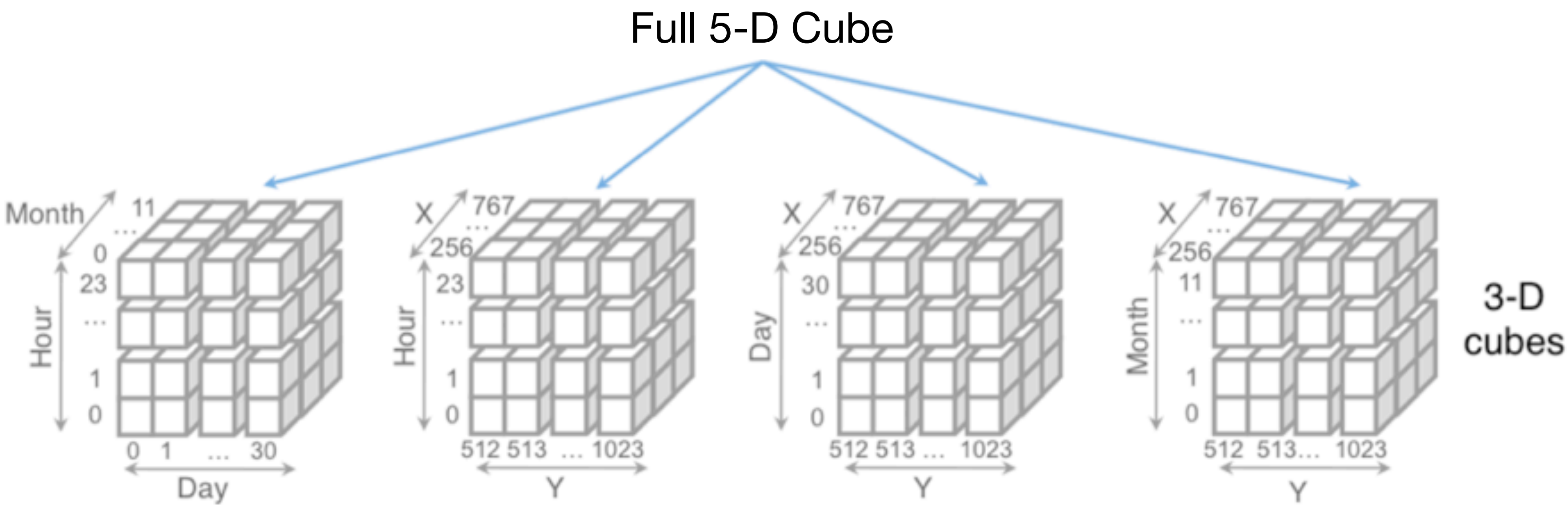



Berkeley
Aquatic Park
Ashby Ave
South
Berkeley
Alcatraz
San Pablo Ave
Market St
Shattuck Ave
Powell St
Emeryville
Adeline St
52nd St
51st St
Market St
123
24
13
580
80
Oakland
Army Base
Peralta St
San Pablo Ave
Grove Shafter Fwy
MacArthur Fwy
Telegraph Ave
Broadway Auto Row
Harrison St
Maritime St
W Grand Ave
West
Oakland
Mandela Pkwy
Market St
20th St
7th St
Oakland
Lakeside
Park
Oakland Naval
Supply Center
880
980
8th St
Madison St
2nd St
Posey Tube
E 8th St

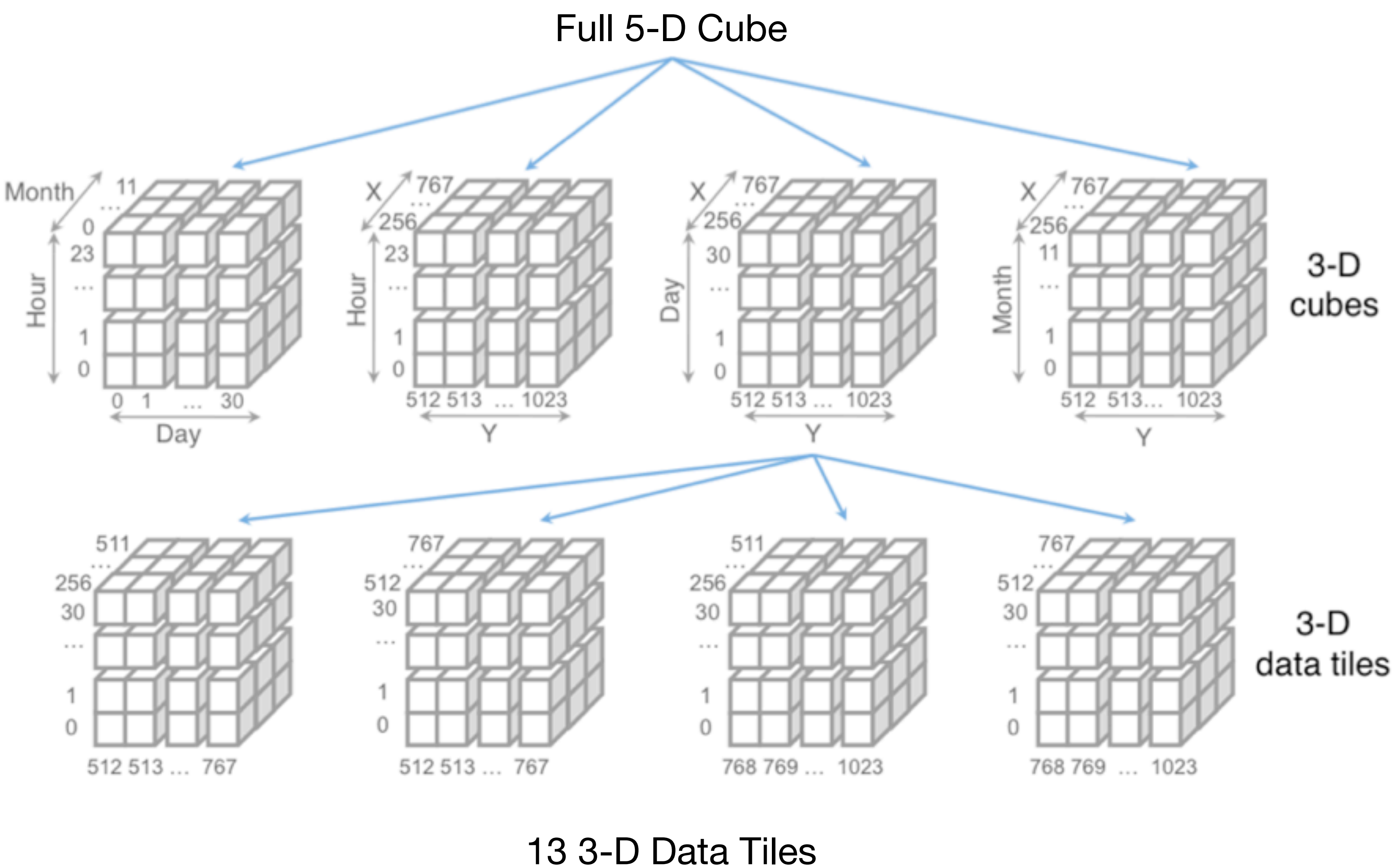


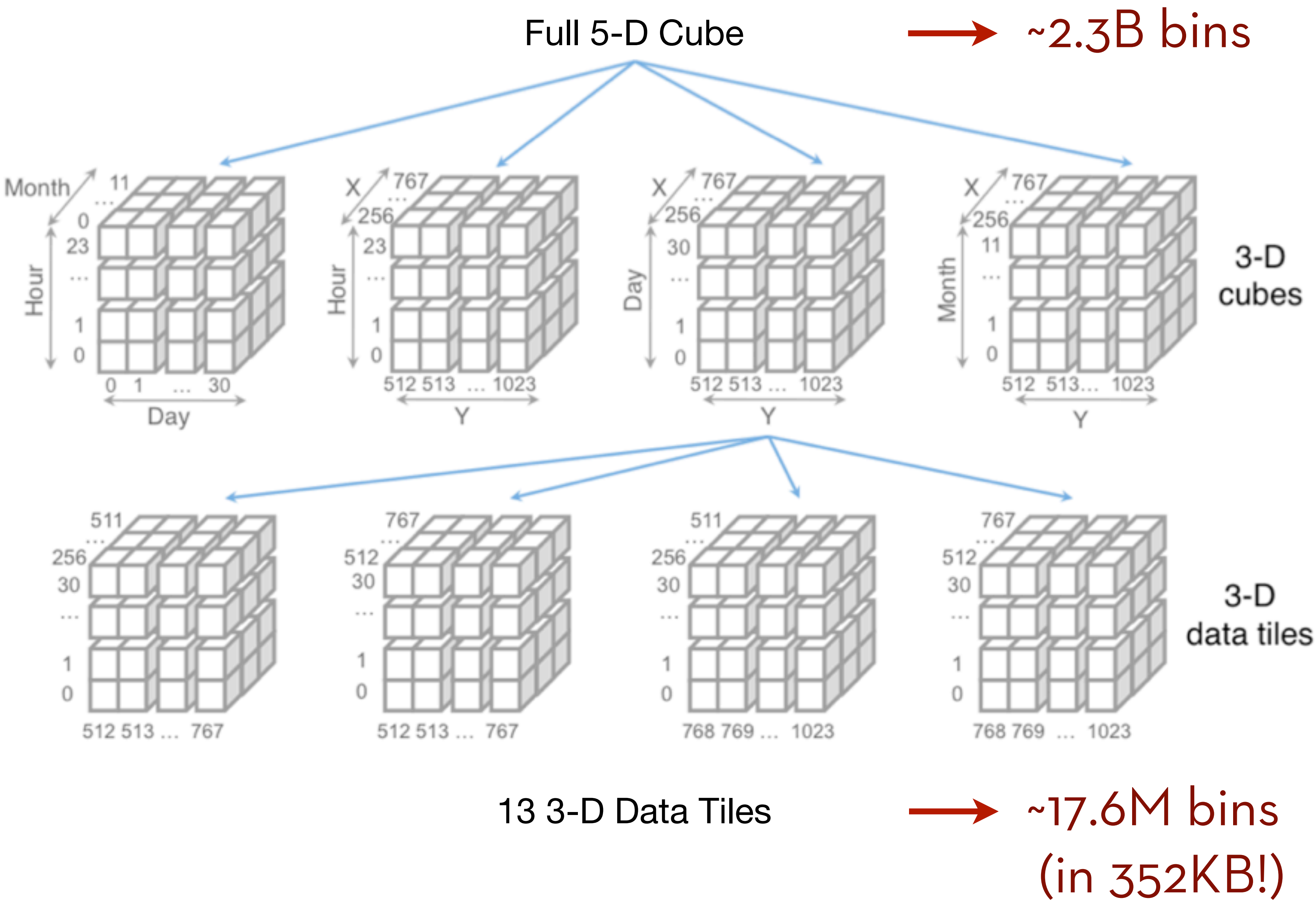
Emeryv
Berkeley
Aquatic Park
13
80
South
Berkeley
Ashby Ave
Alcatraz
San Pablo Ave
Market St
Shattuck Ave
Powell St
Adeline St
52nd St
51st S
123
24
580
Oakland
Army Base
Maritime St
Oakland Naval
Supply Center
W Grand Ave
West
Oakland
Mandela Pkwy
Market St
Grove Shafter
Telegraph Ave
Broadway Auto Row
Harrison St
20th St
Oakland
880
980
8th St
Madison St
2nd St
Posey Tube
E 8th S

Full 5-D Cube

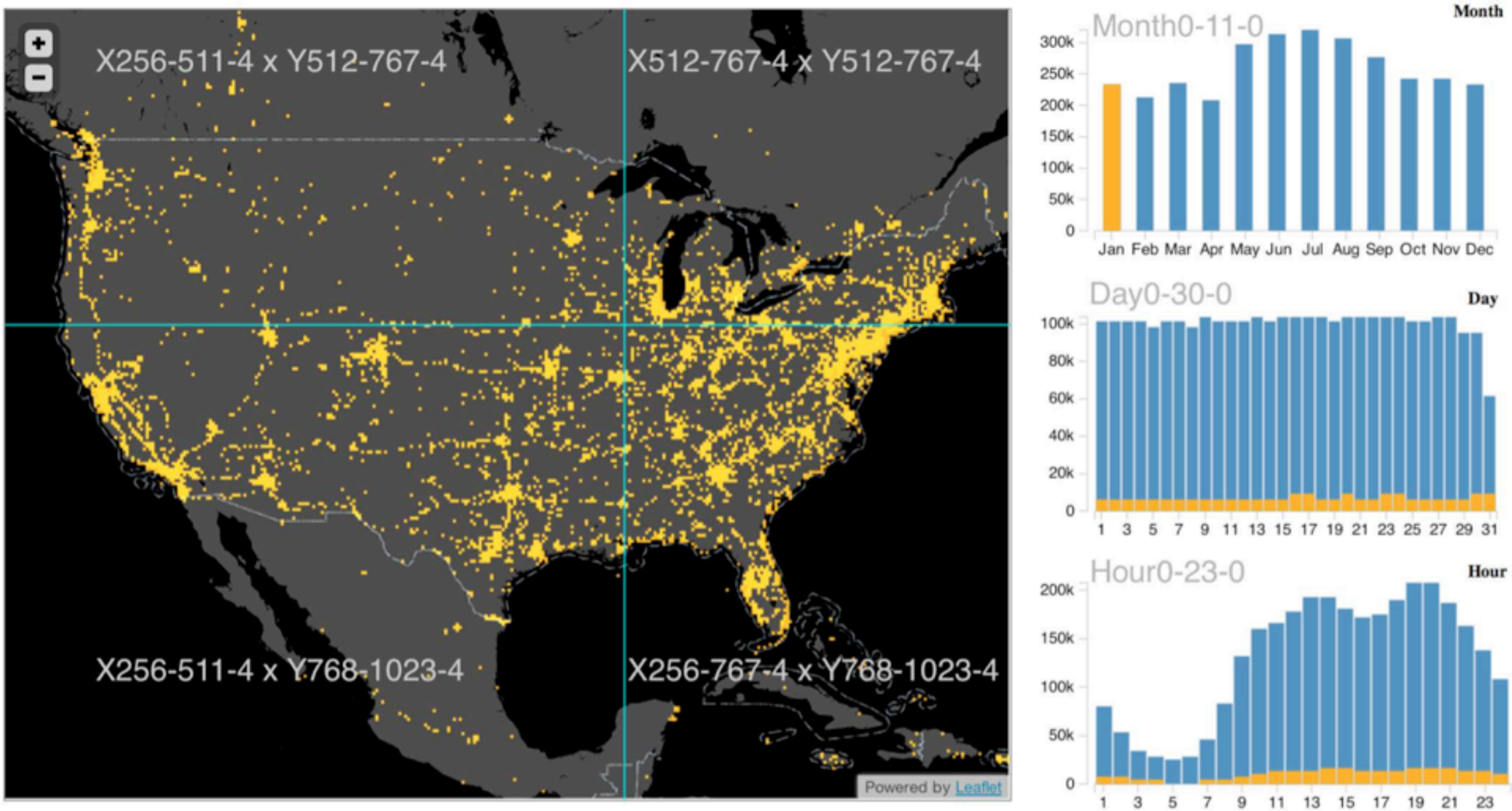


For any pair of 1D or 2D binned plots, the maximum number of dimensions needed to support brushing & linking is **four**.

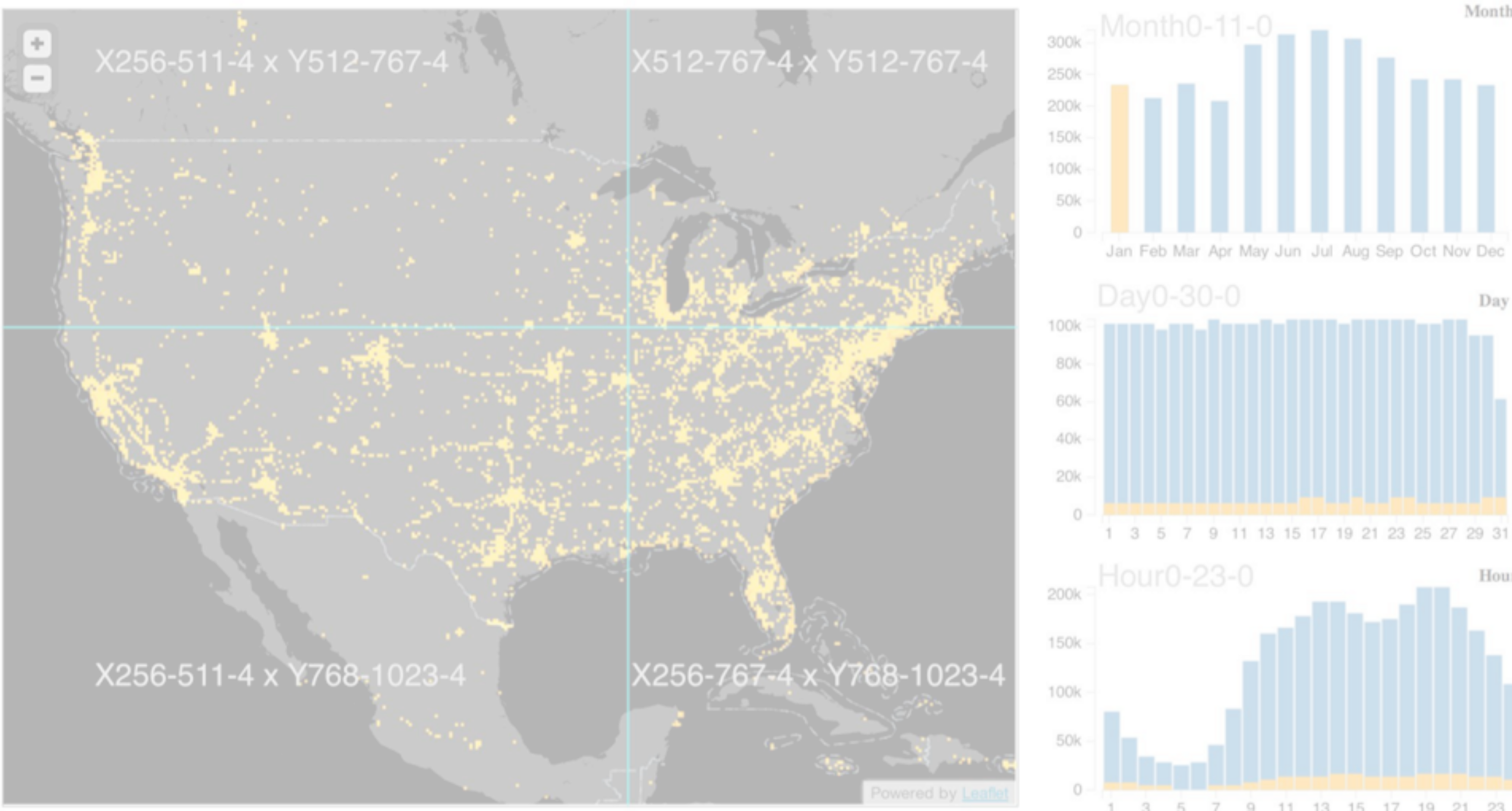

Full 5-D Cube
Month
11
...
0
23
...
1
0
Hour
0 1 ... 30
Day
X
767
...
256
23
...
1
0
Hour
512 513 ... 1023
Y
X
767
...
256
30
...
1
0
Day
512 513 ... 1023
Y
X
767
...
256
11
...
1
0
Month
512 513... 1023
Y
3-D
cubes
511
...
256
30
...
1
0
512 513 ... 767
767
...
512
30
...
1
0
512 513 ... 767
511
...
256
30
...
1
0
768 769 ... 1023
767
...
512
30
...
1
0
768 769 ... 1023
3-D
data tiles
13 3-D Data Tiles



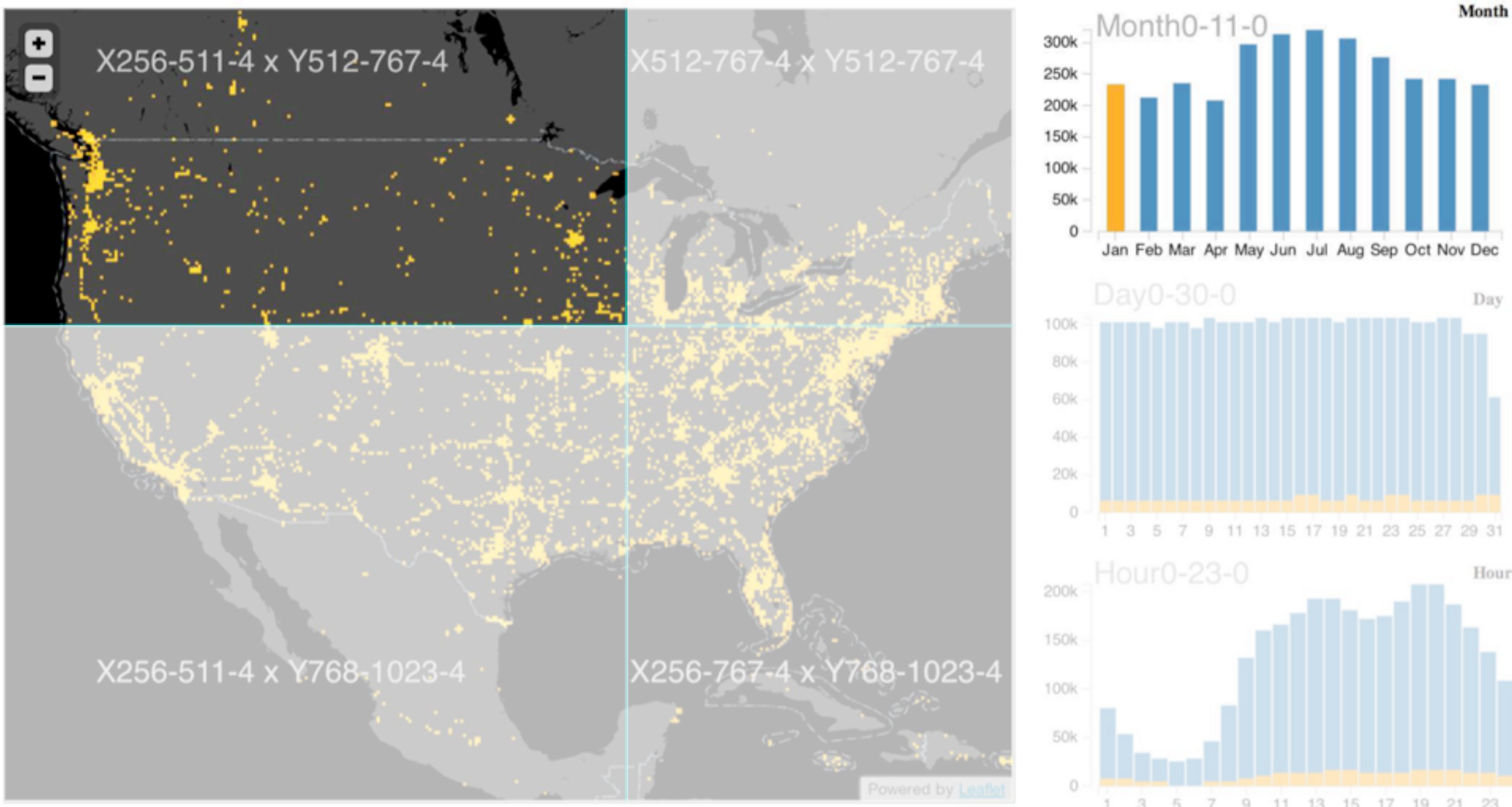
Full 5-D Cube
~2.3B bins
Month
Hour
Day
X
Y
3-D cubes
3-D data tiles
13 3-D Data Tiles
~17.6M bins
(in 352KB!)

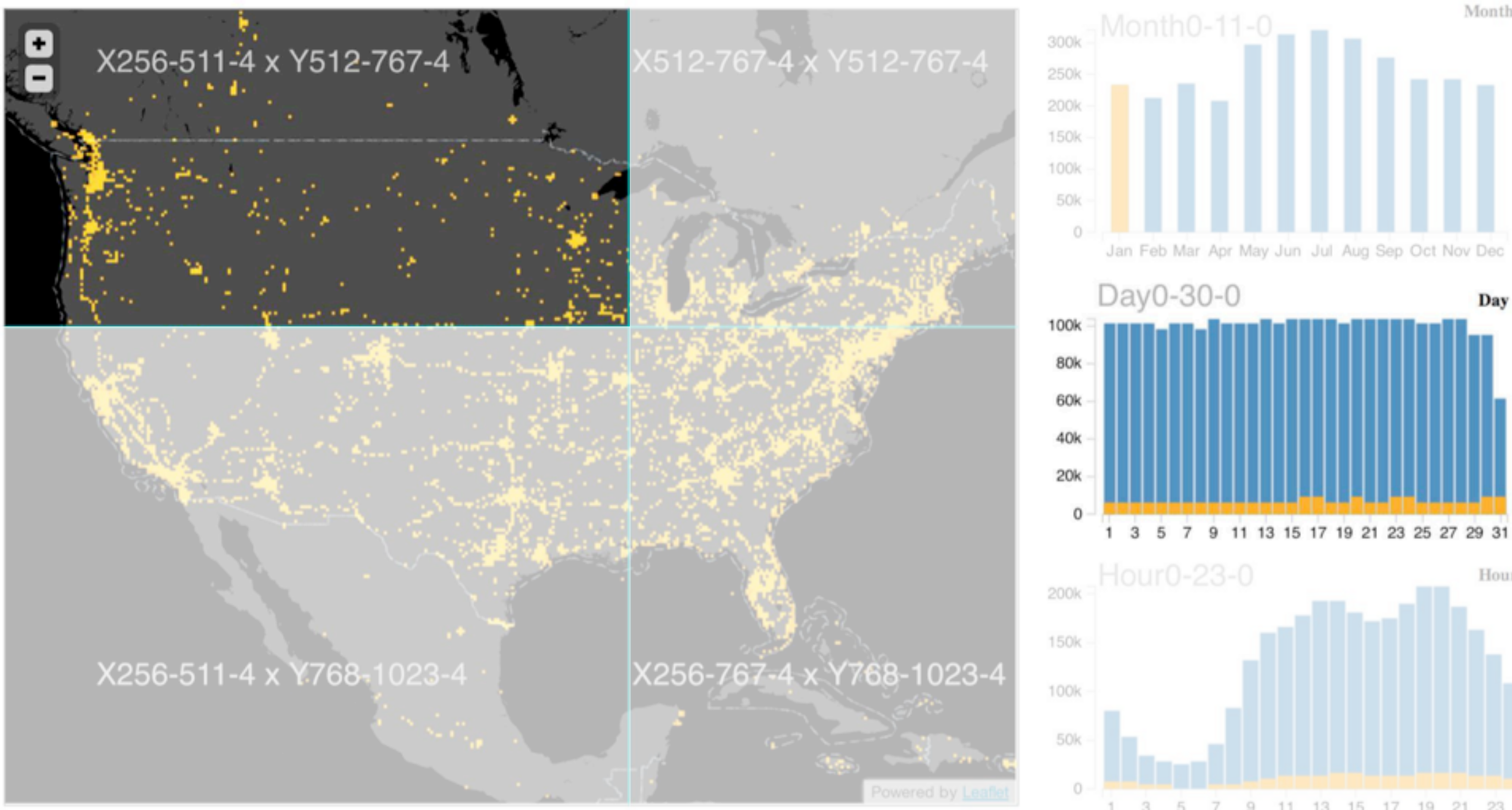


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
0 50k 100k 150k 200k 250k 300k
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
0 20k 40k 60k 80k 100k
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
0 50k 100k 150k 200k
1 3 5 7 9 11 13 15 17 19 21 23

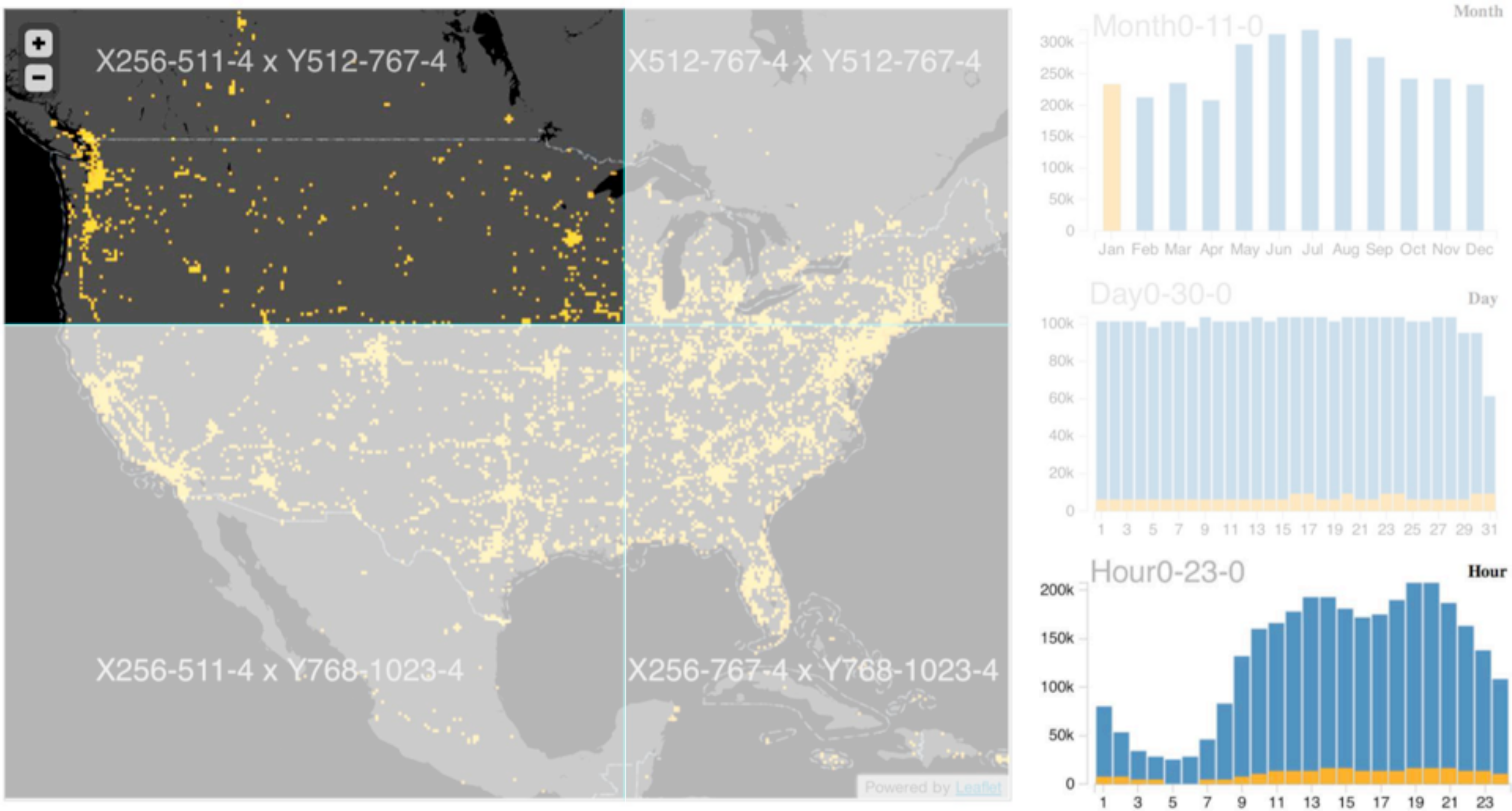


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k 250k 200k 150k 100k 50k 0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k 80k 60k 40k 20k 0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k 150k 100k 50k 0
1 3 5 7 9 11 13 15 17 19 21 23

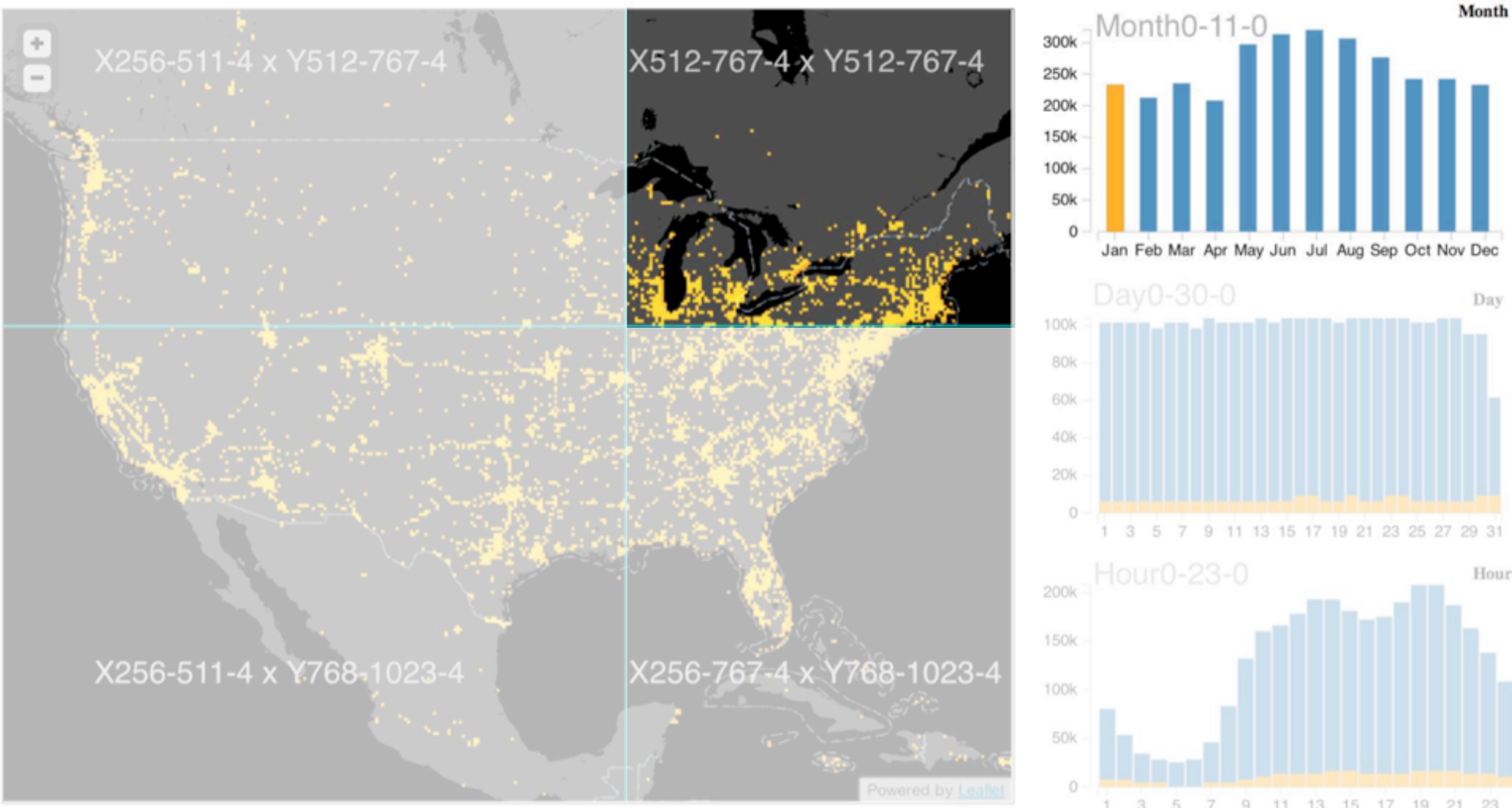

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23



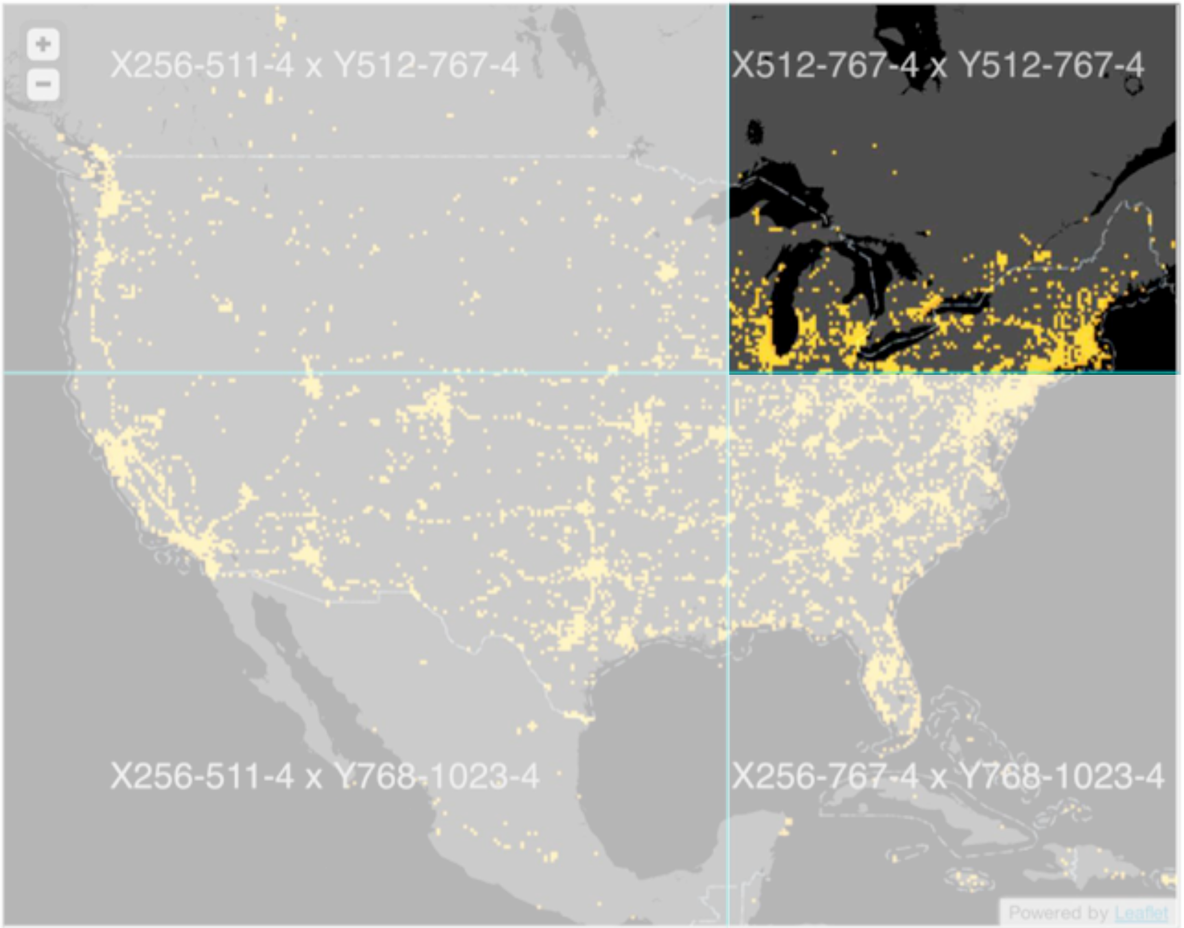
X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

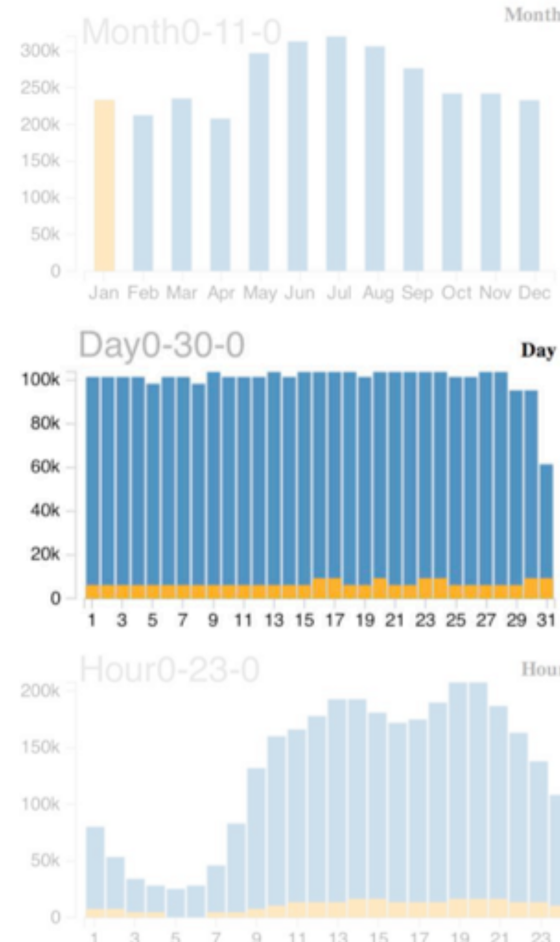


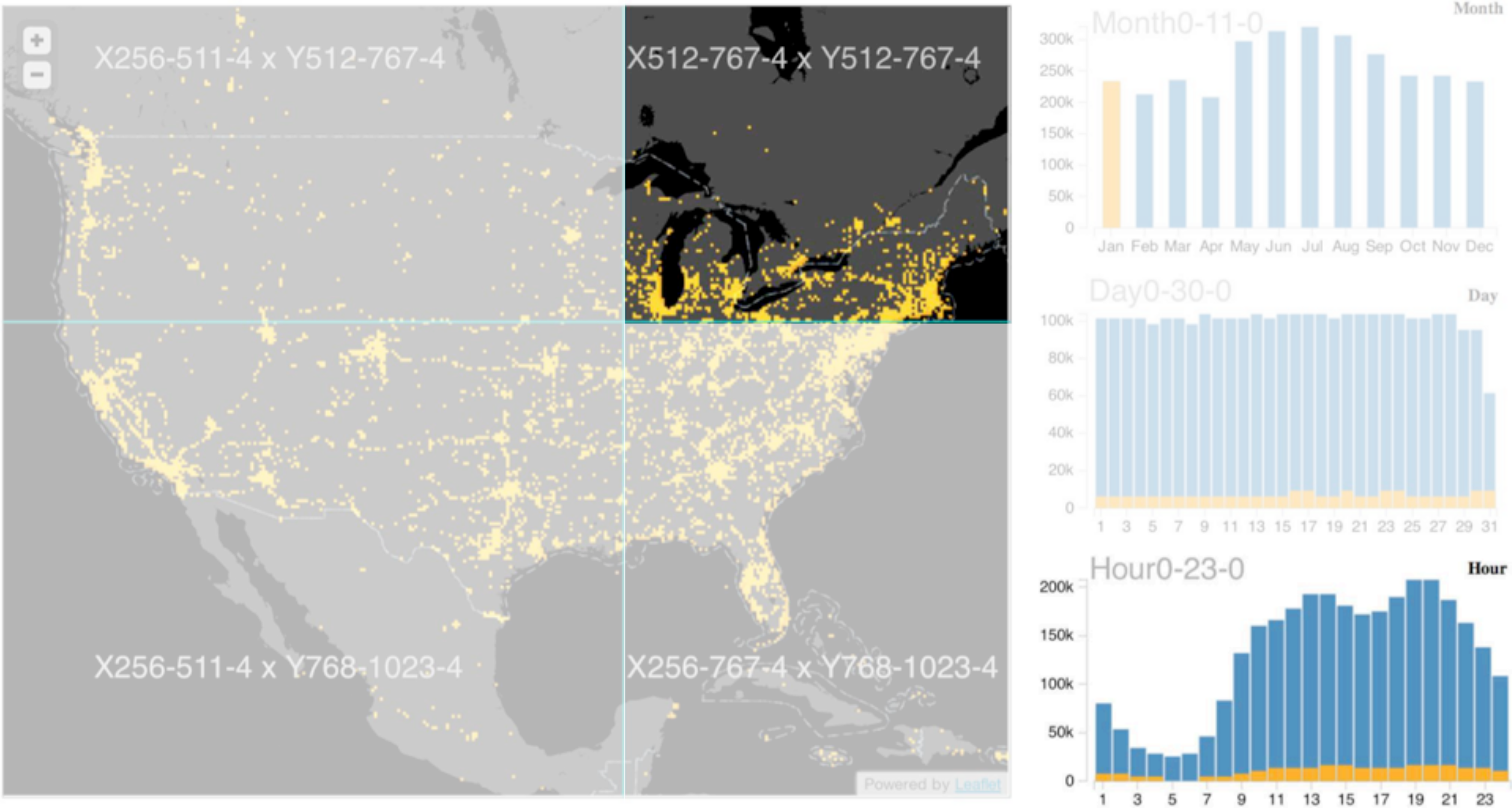
X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month
Month0-11-0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
Day0-30-0
Hour
Hour0-23-0

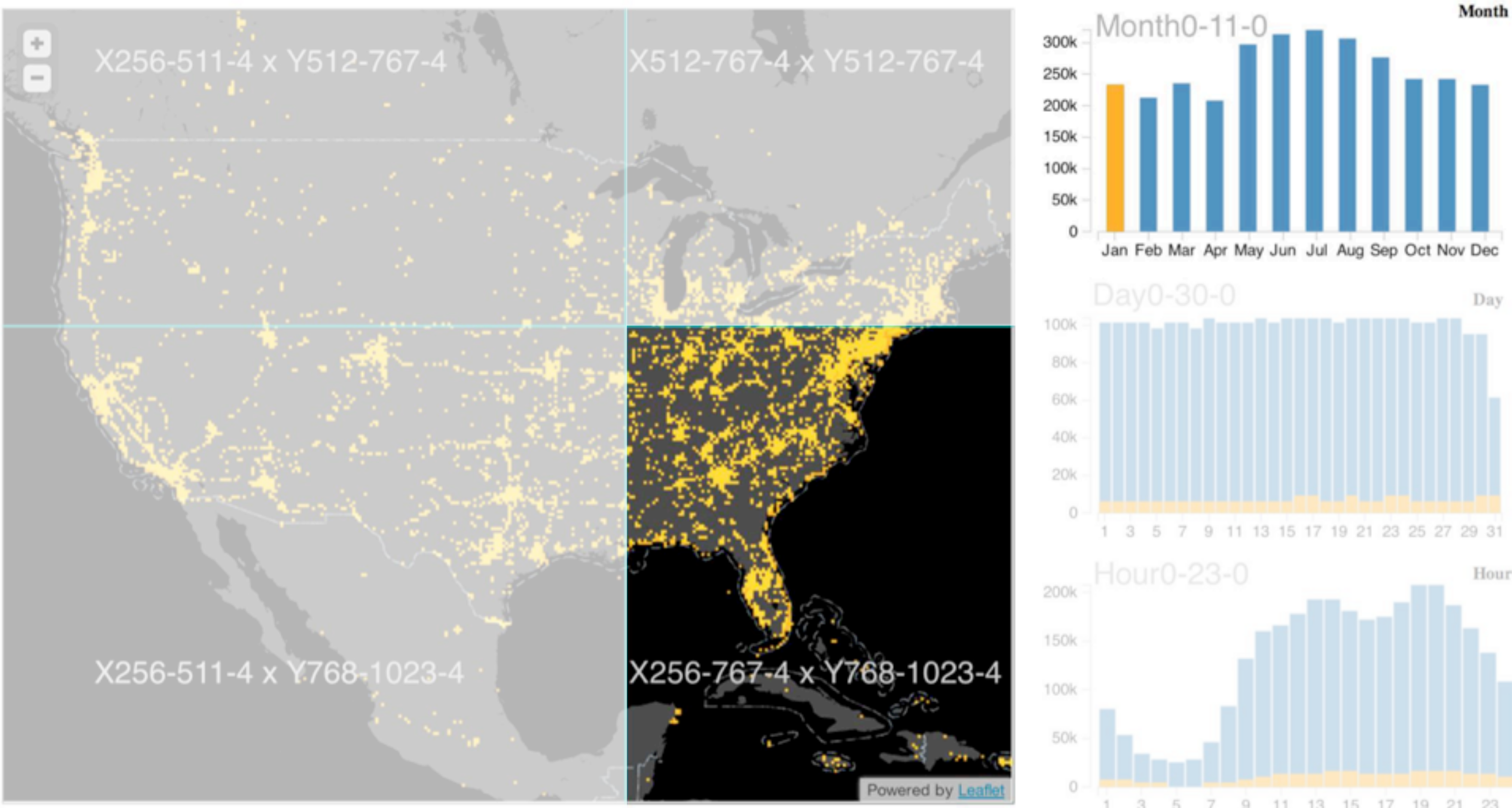


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

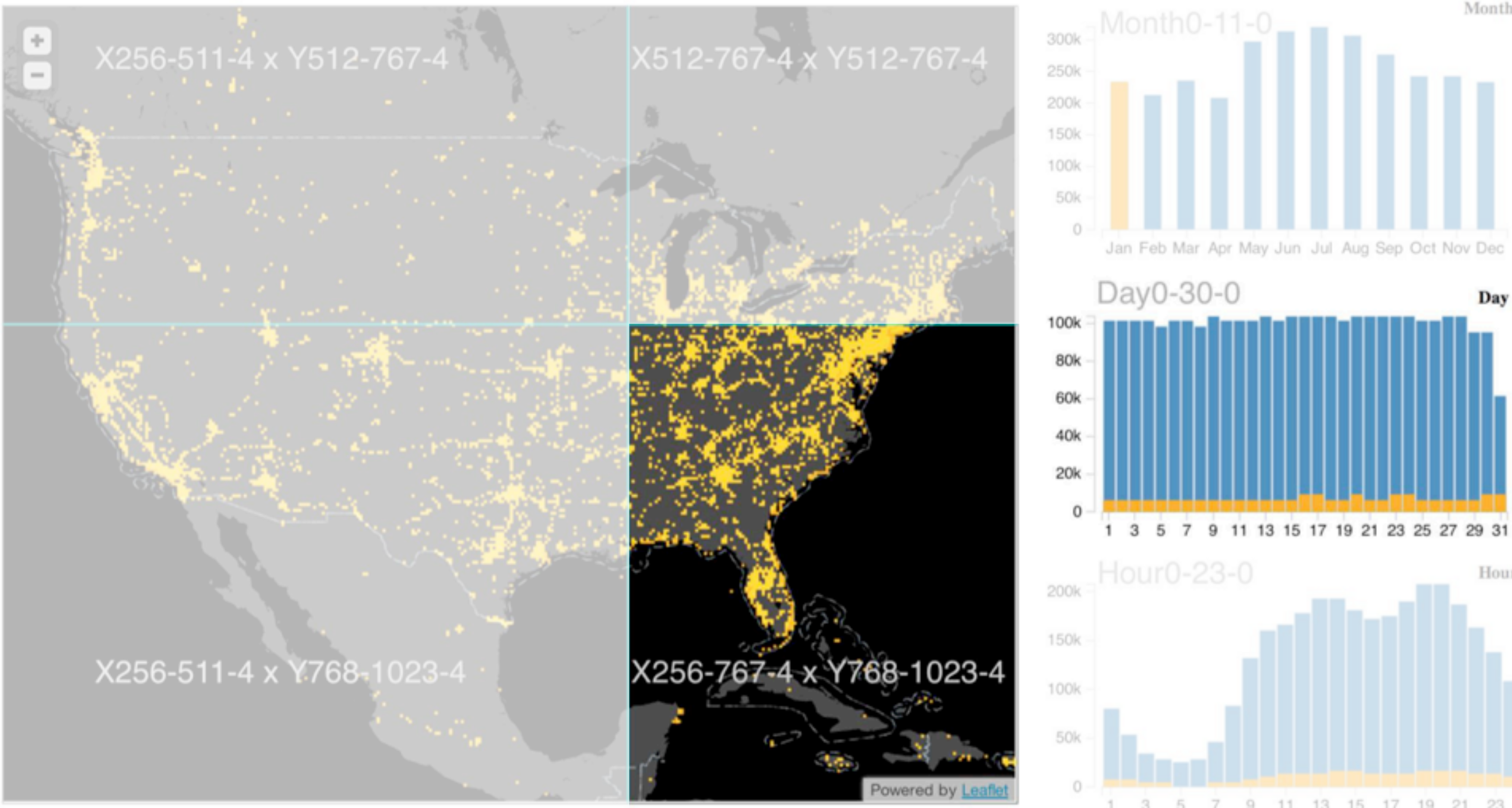

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet


Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

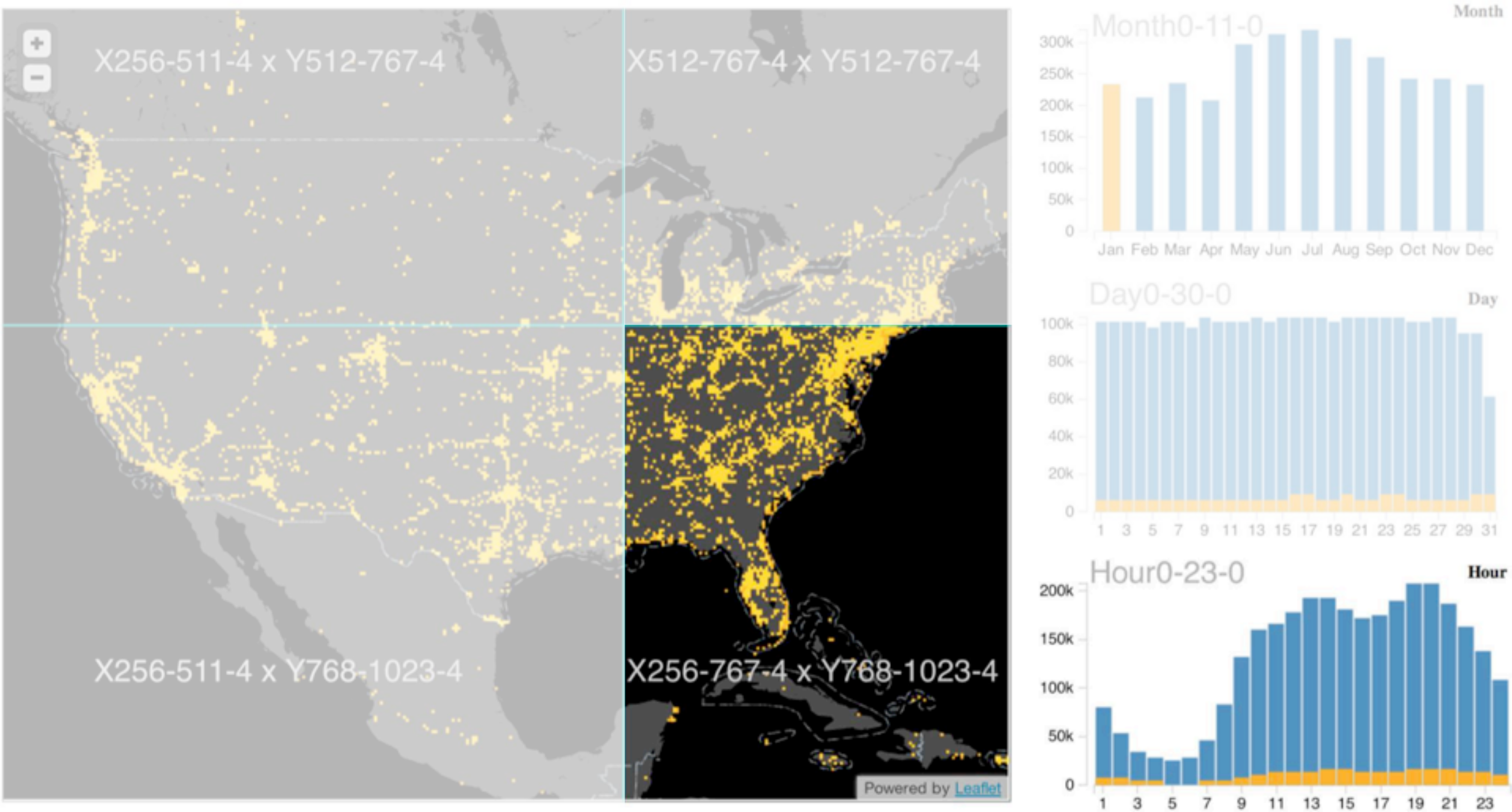

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

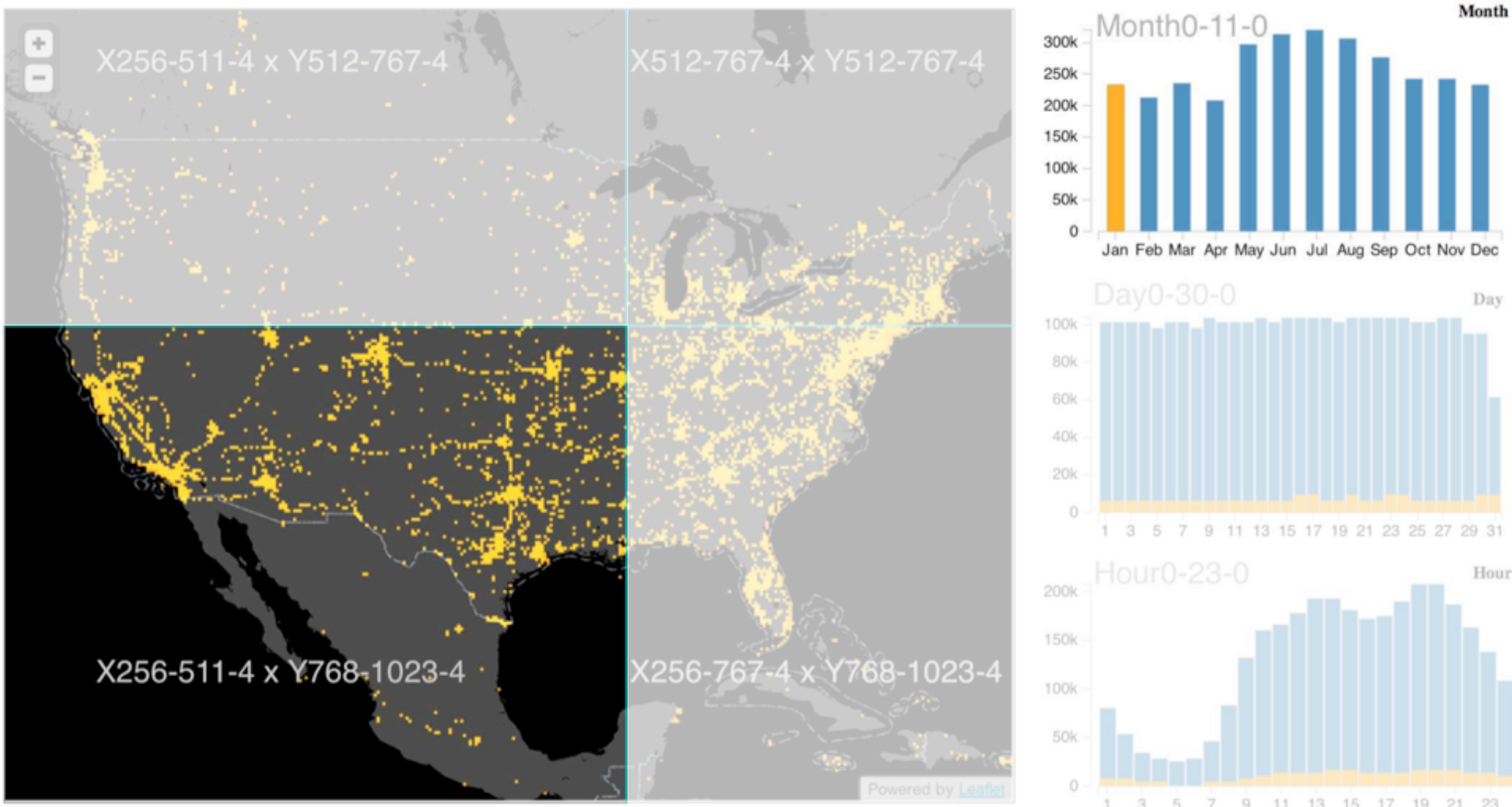


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

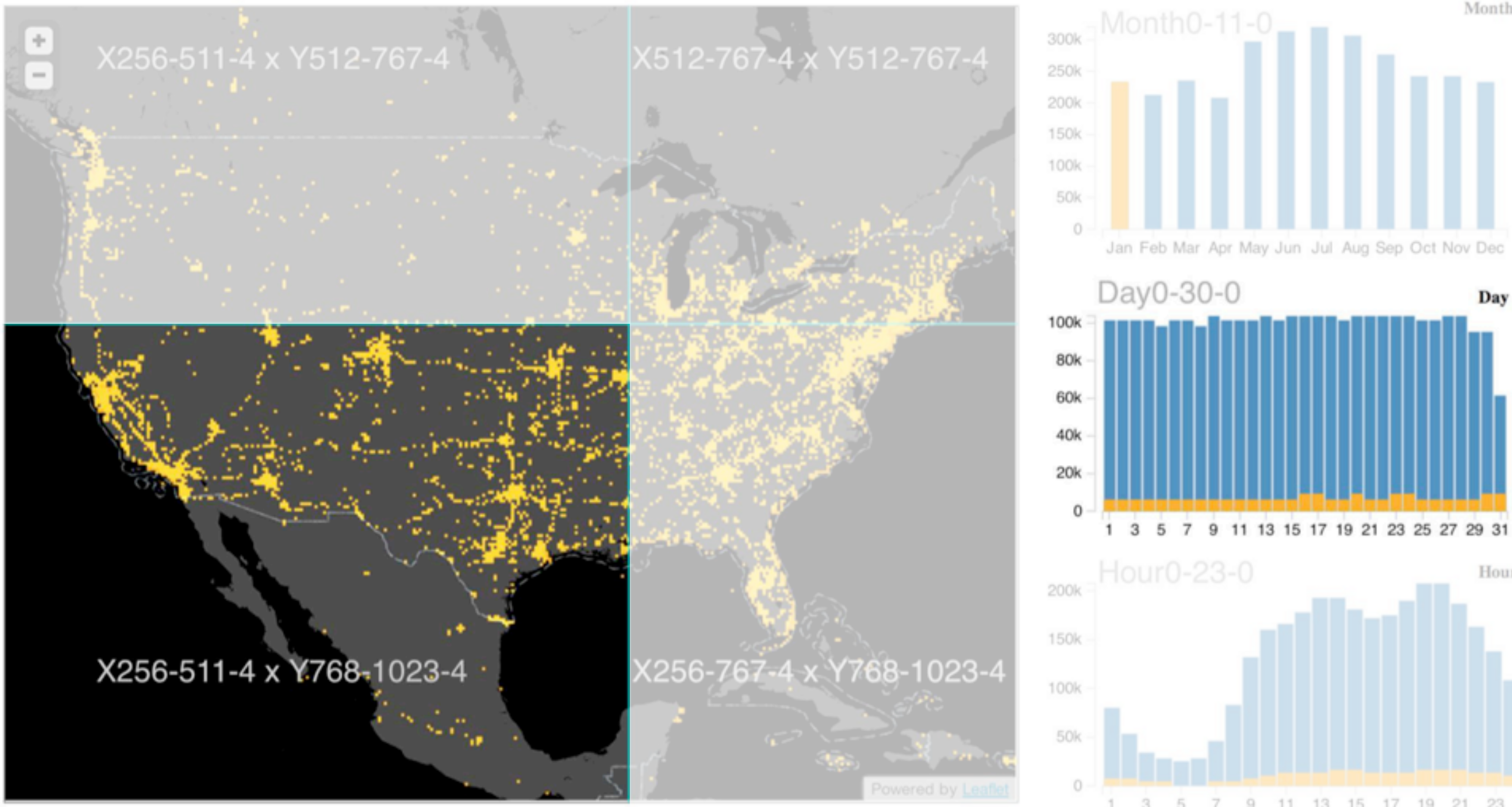


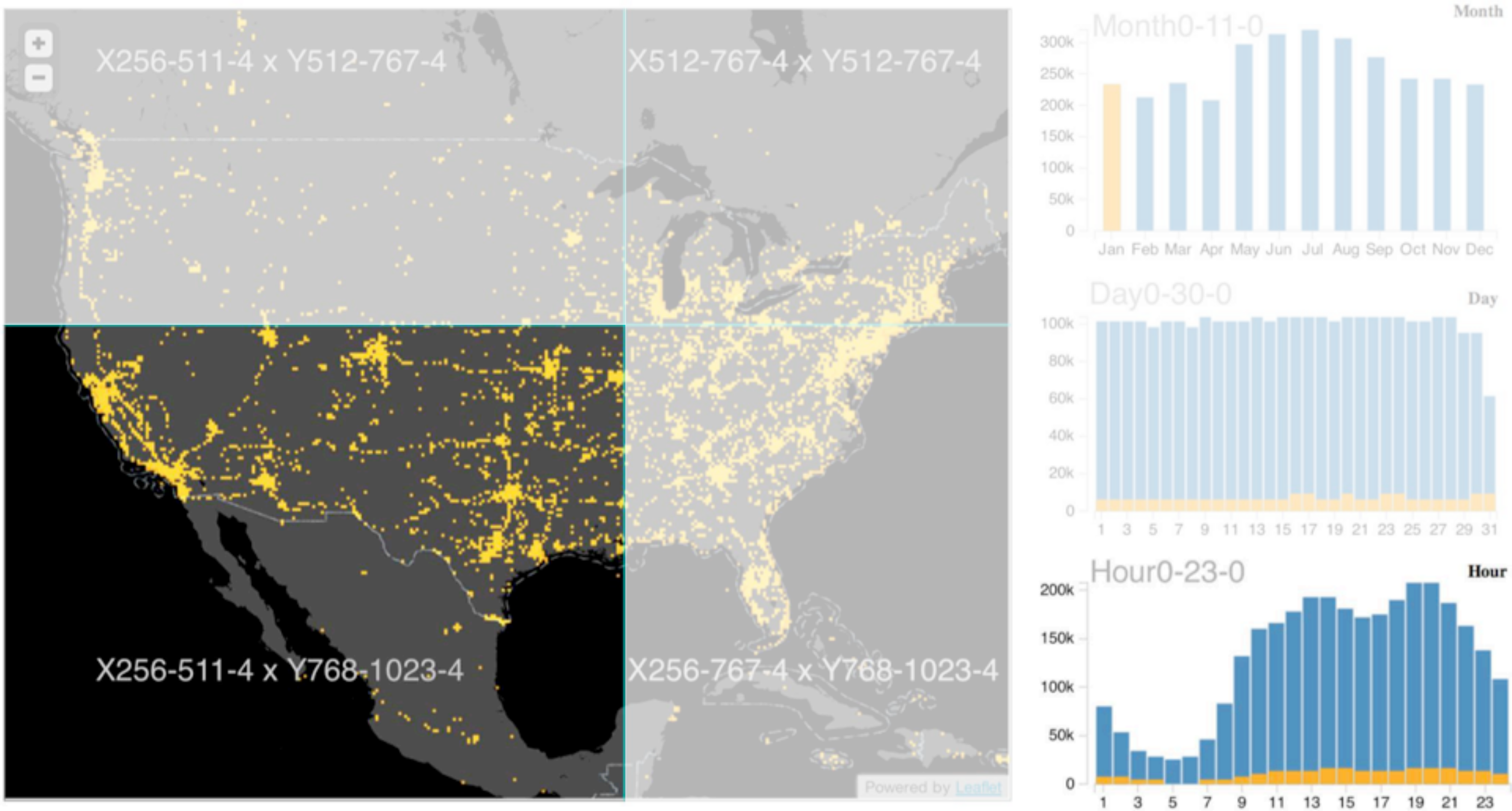
X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
Hour0-23-0
Hour

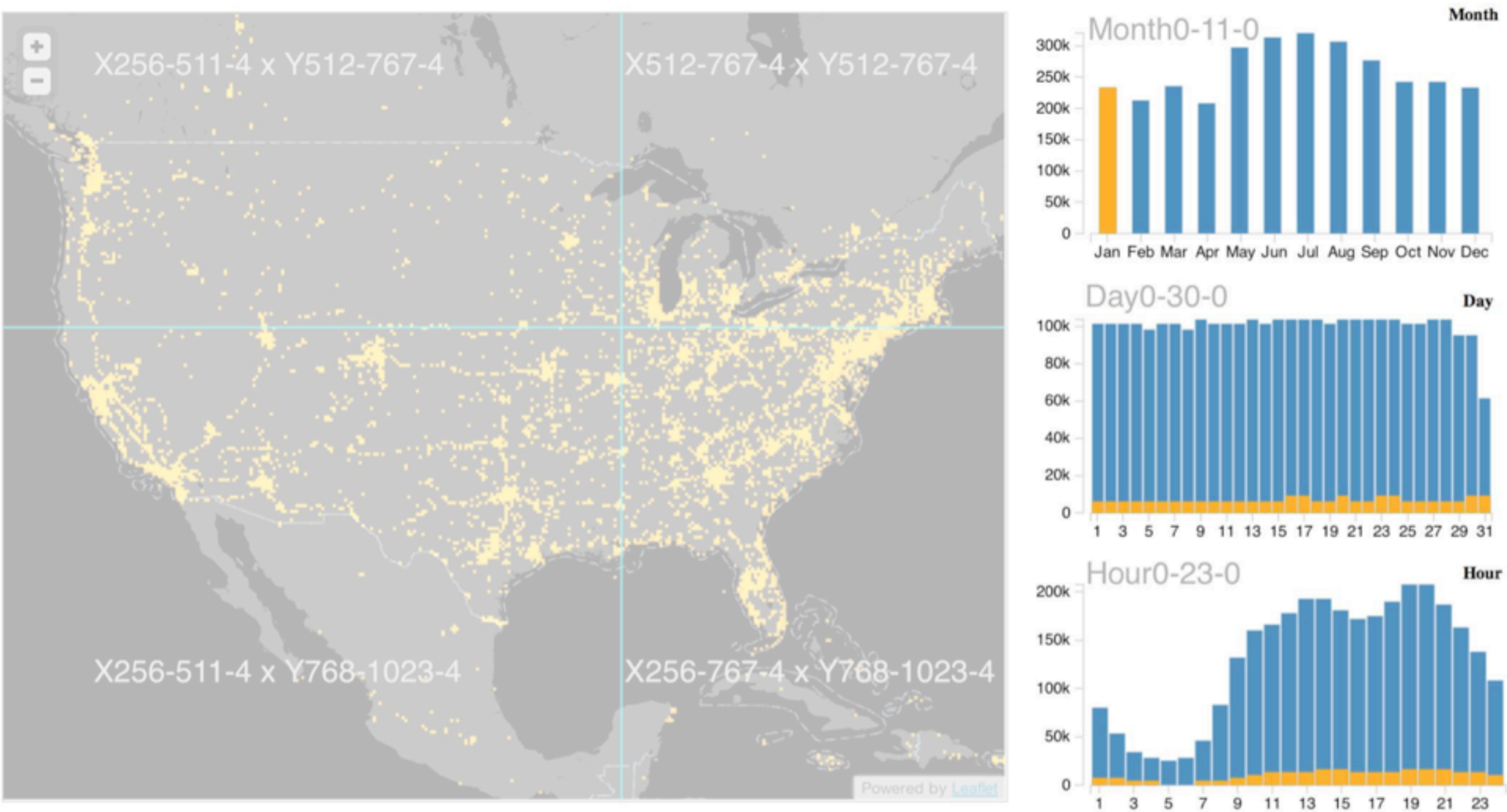

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23



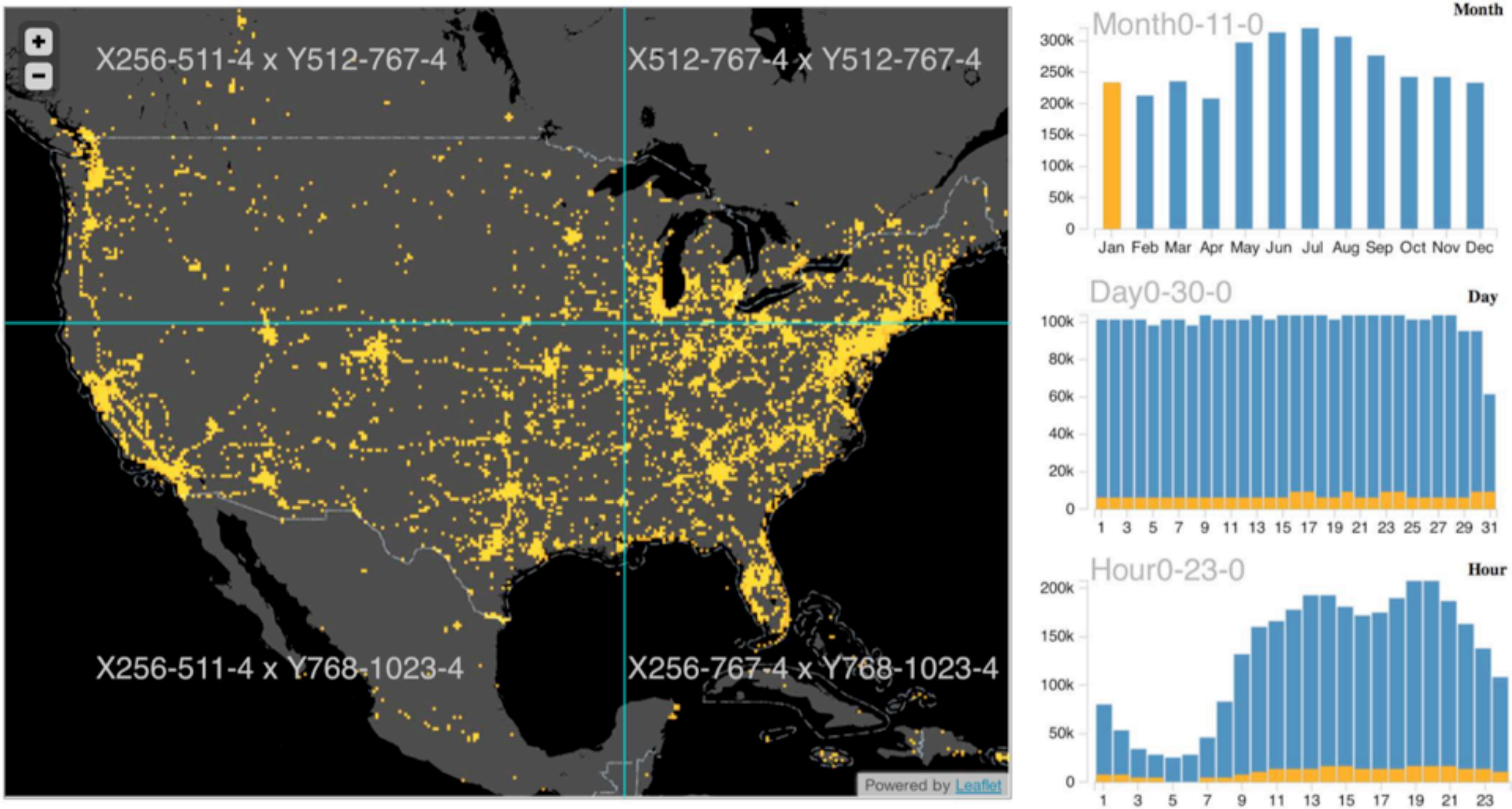
X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
Hour0-23-0
Hour

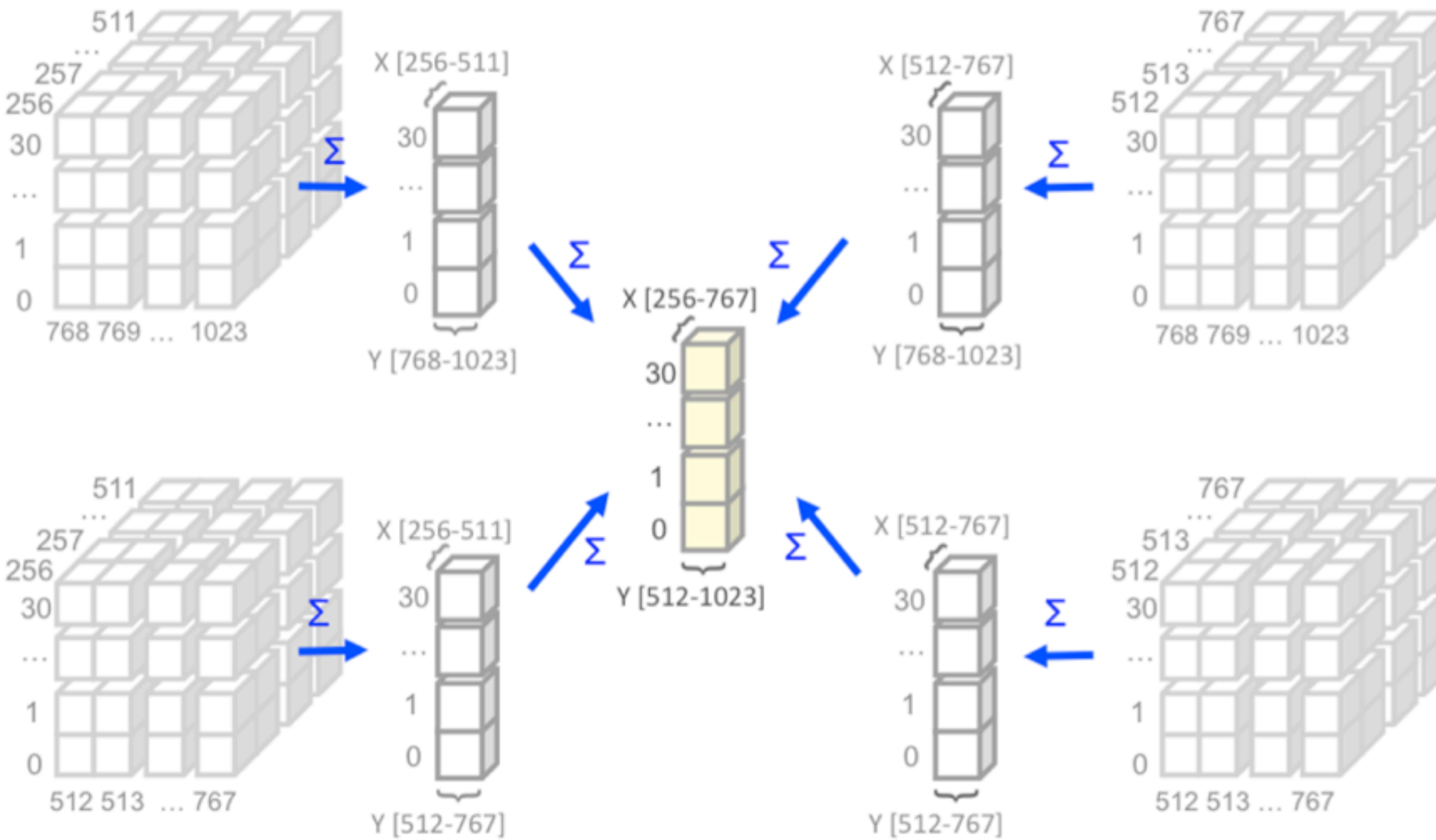

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
0 50k 100k 150k 200k 250k 300k
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
0 20k 40k 60k 80k 100k
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
0 50k 100k 150k 200k
1 3 5 7 9 11 13 15 17 19 21 23



X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
0
50k
100k
150k
200k
250k
300k
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
0
20k
40k
60k
80k
100k
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
0
50k
100k
150k
200k
1 3 5 7 9 11 13 15 17 19 21 23


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23


Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31


511
257
256
30
1
0
768 769 … 1023
X [256-511]
Y [768-1023]
Σ
767
513
512
X [512-767]
Y [768-1023]
X [256-767]
Y [512-1023]
512 513 … 767
Y [512-767]

index	X	Y	Day	Count
0	256	512	0	378
1	256	512	1	0
...	...	...	...	...
30	256	512	30	1209
31	256	513	0	76
...	...	...	...	...
7935	256	767	30	0
7936	257	512	0	0
...	...	...	...	...
2031615	511	767	30	466

index	X	Y	Day	Count
0	256	512	0	378
1	256	512	1	0
...	...	...	...	...
30	256	512	30	1209
31	256	513	0	76
...	...	...	...	...
7935	256	767	30	0
7936	257	512	0	0
...	...	...	...	...
2031615	511	767	30	466

sparse →

X	Y	Day	Count
256	512	0	378
...	...	...	...
256	512	30	1209
256	513	0	76
...	...	...	...
511	767	30	466

index	X	Y	Day	Count
0	256	512	0	378
1	256	512	1	0
...	...	...	...	...
30	256	512	30	1209
31	256	513	0	76
...	...	...	...	...
7935	256	767	30	0
7936	257	512	0	0
...	...	...	...	...
2031615	511	767	30	466

sparse →

X	Y	Day	Count
256	512	0	378
...	...	...	...
256	512	30	1209
256	513	0	76
...	...	...	...
511	767	30	466

dense →

378	0	...	1209	76	...	0	0	...	466

index	X	Y	Day	Count
0	256	512	0	378
1	256	512	1	0
...	...	...	...	...
30	256	512	30	1209
31	256	513	0	76
...	...	...	...	...
7935	256	767	30	0
7936	257	512	0	0
...	...	...	...	...
2031615	511	767	30	466

sparse

X	Y	Day	Count
256	512	0	378
...	...	...	...
256	512	30	1209
256	513	0	76
...	...	...	...
511	767	30	466

dense

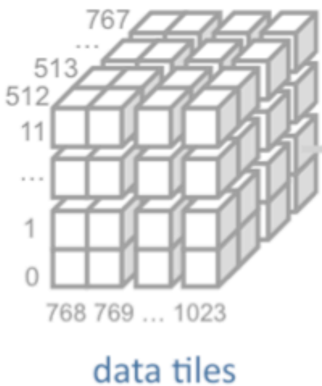
378	0	...	1209	76	...	0	0	...	466

Dense packing more efficient if:

density > 25% in 3D tiles

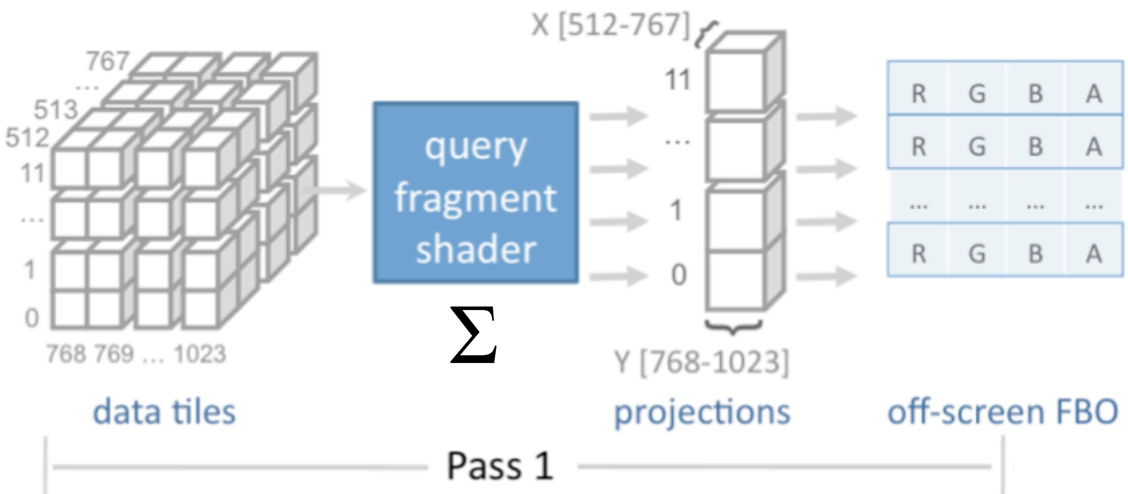
density > 20% in 4D tiles

Query & Render on GPU via WebGL



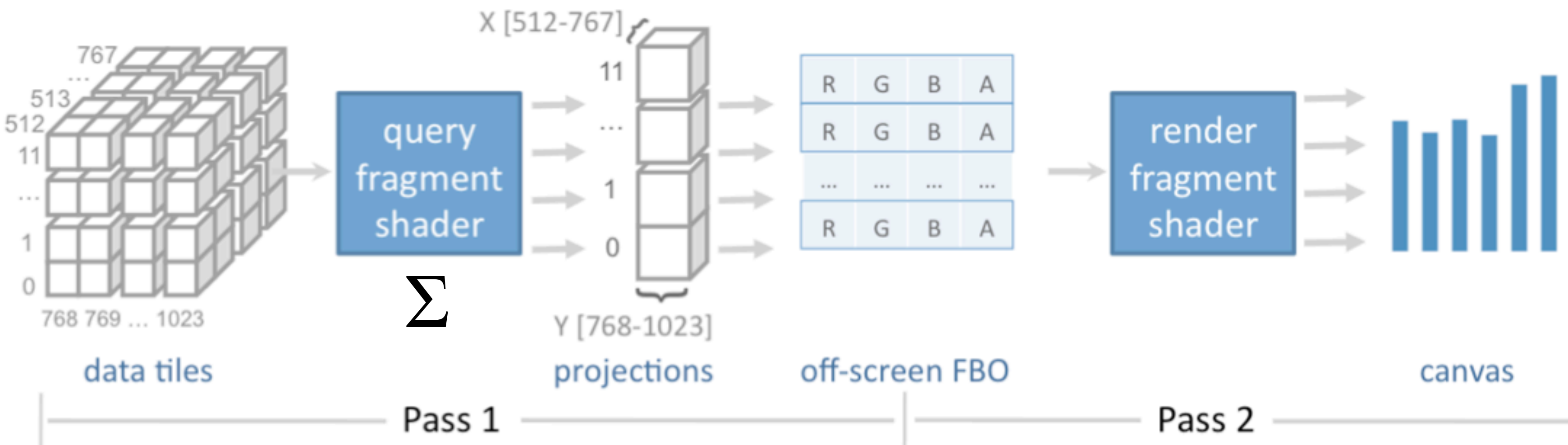
Pack data tiles as PNG image files,
bind to WebGL as image textures.

Query & Render on GPU via WebGL

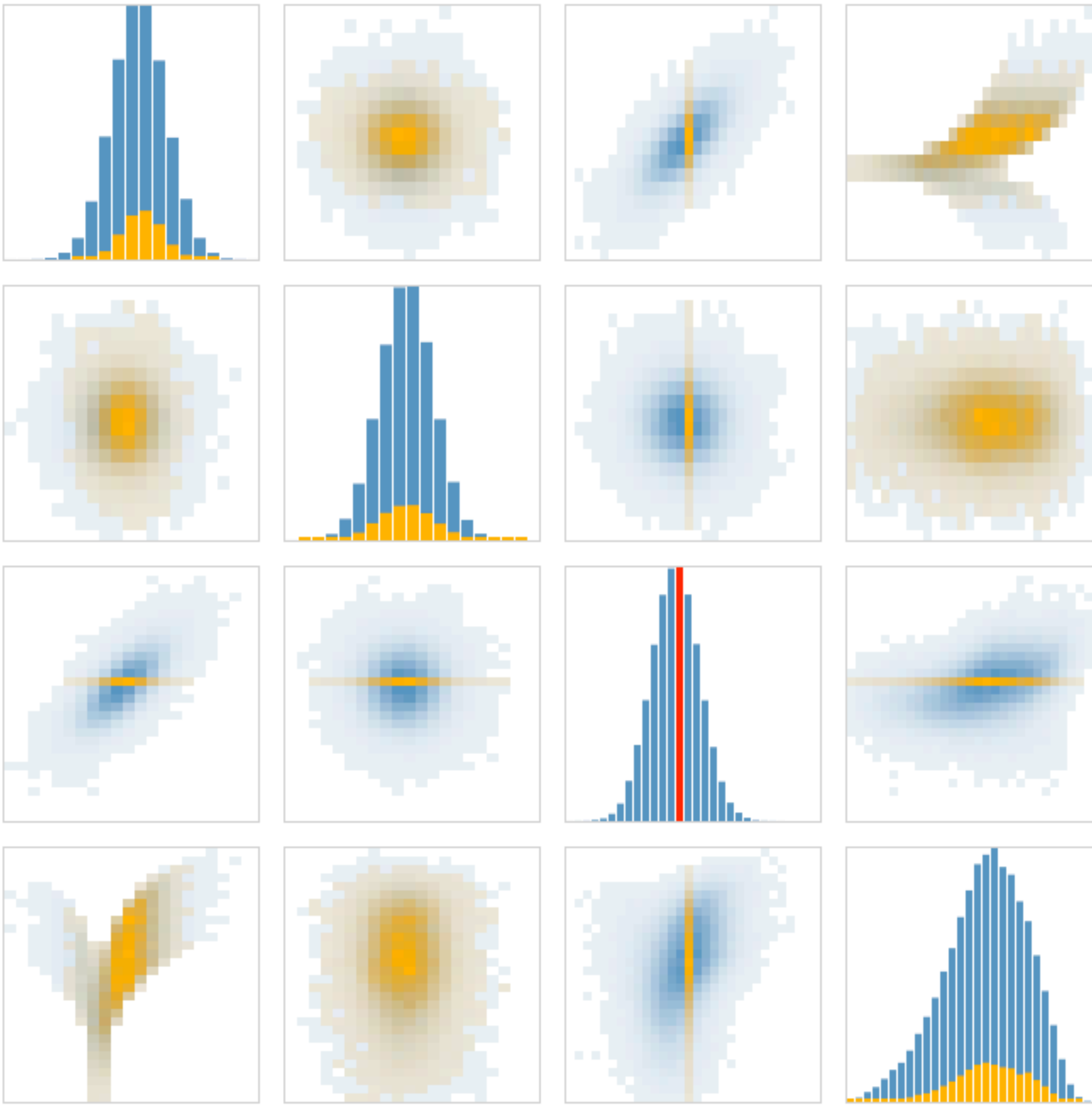


Invoke program for each output bin.
Executes in parallel on GPU.

Query & Render on GPU via WebGL



Performance Benchmarks



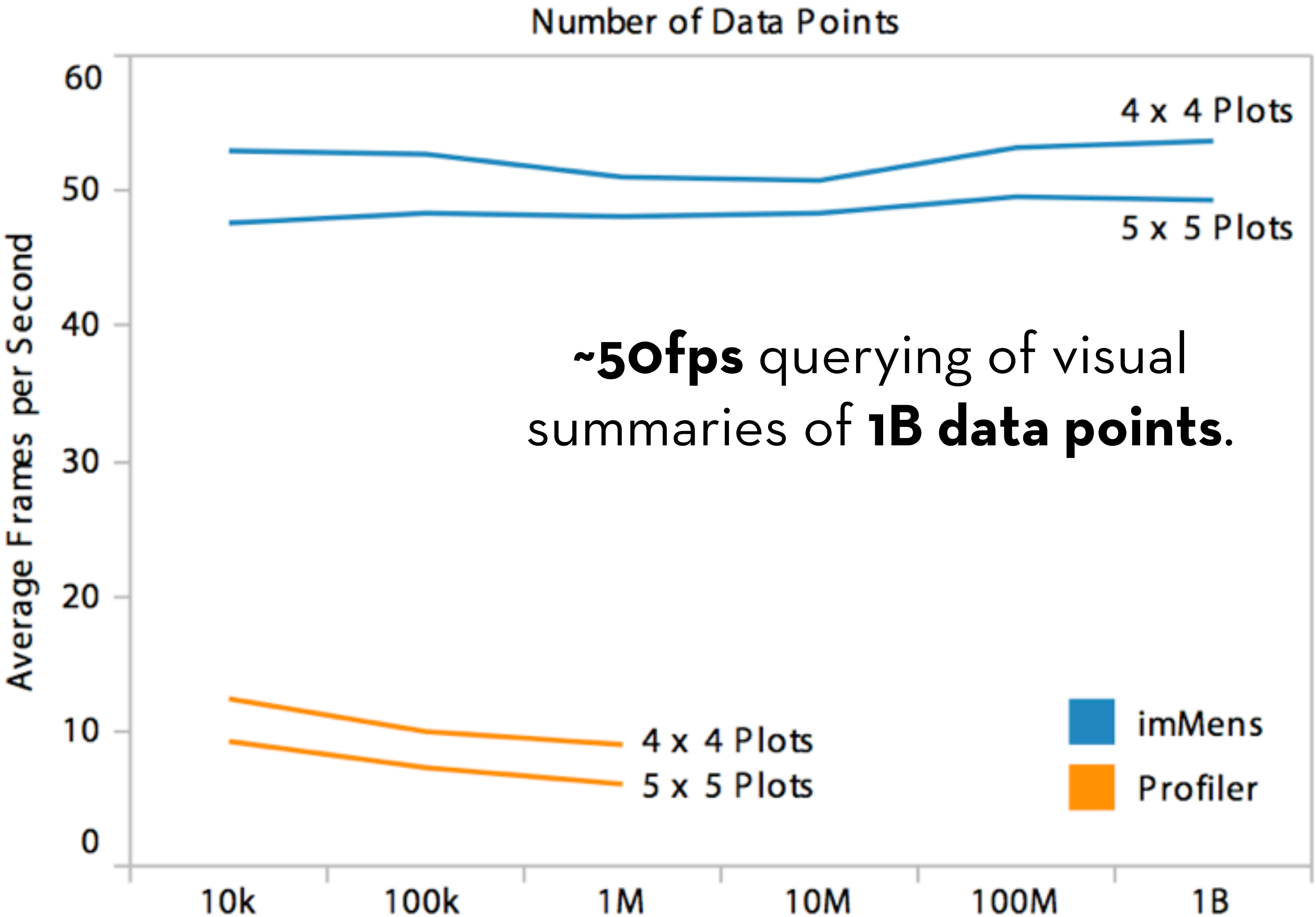
Simulate interaction: brushing & linking across binned plots.

- imMens vs. Profiler
- 4x4 and 5x5 plots
- 10 to 50 bins

Measure time from selection to render.

Test setup:

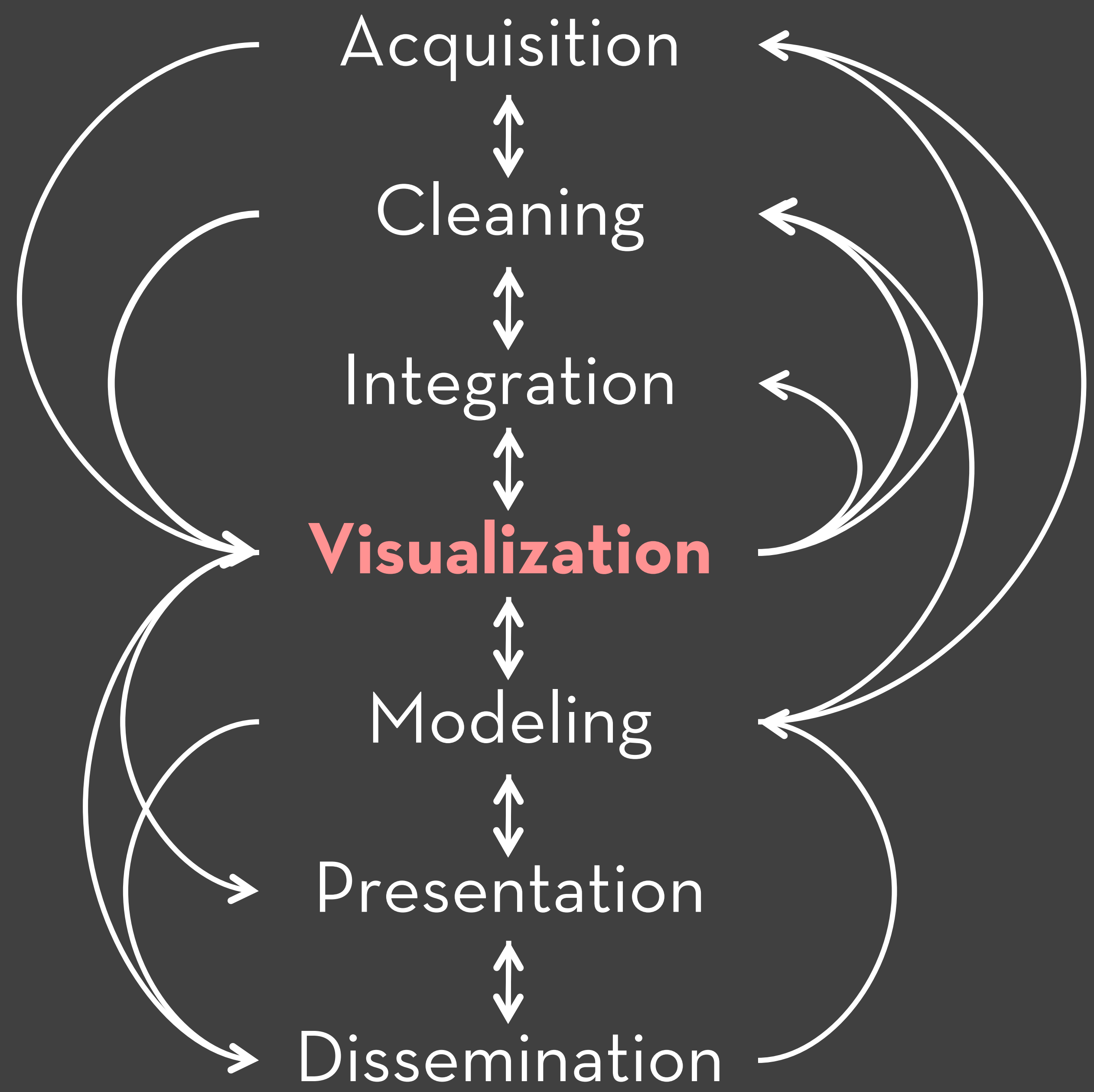
2.3 GHz MacBook Pro (4-core)
NVIDIA GeForce GT 650M
Google Chrome v.23.0

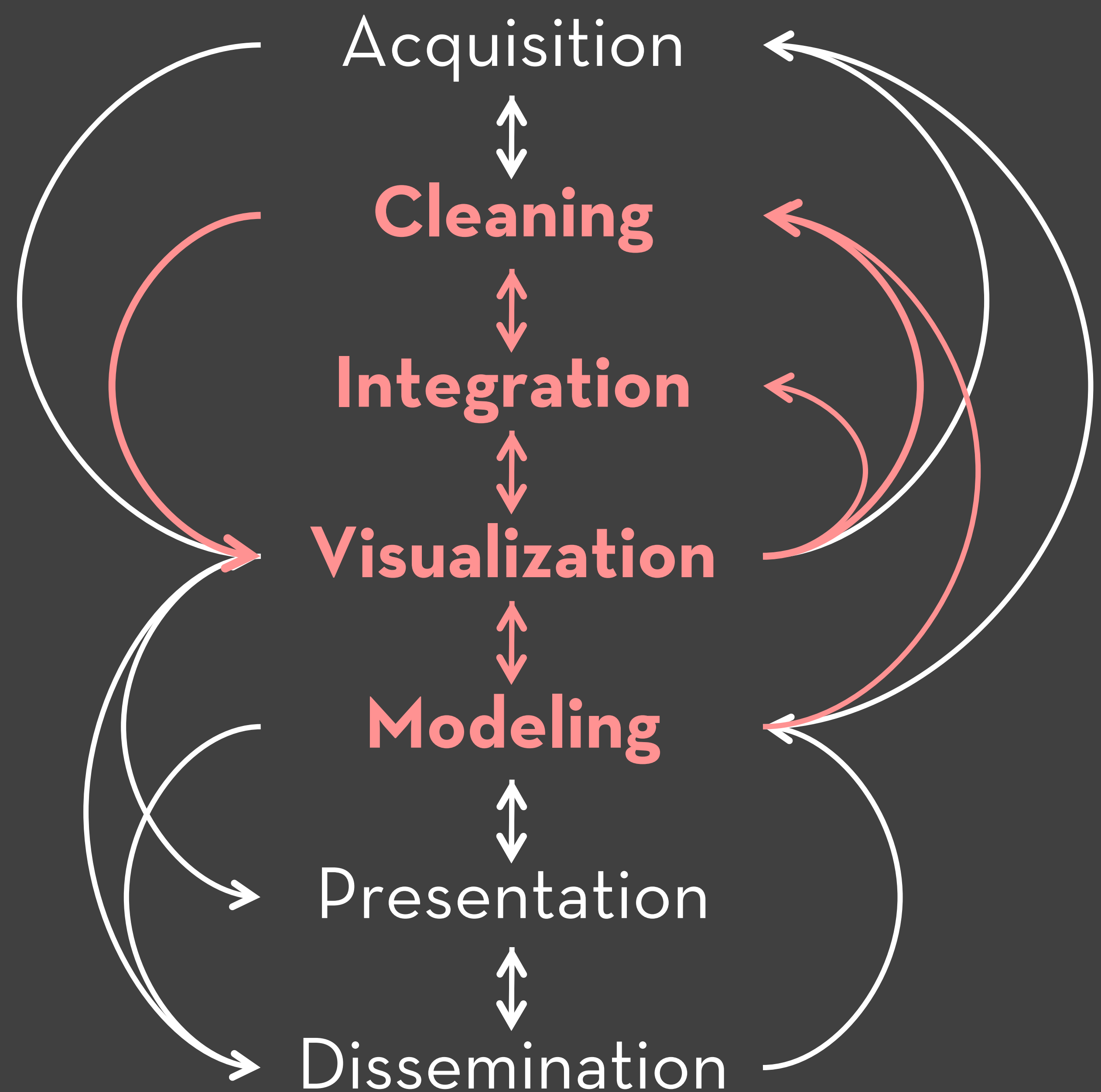


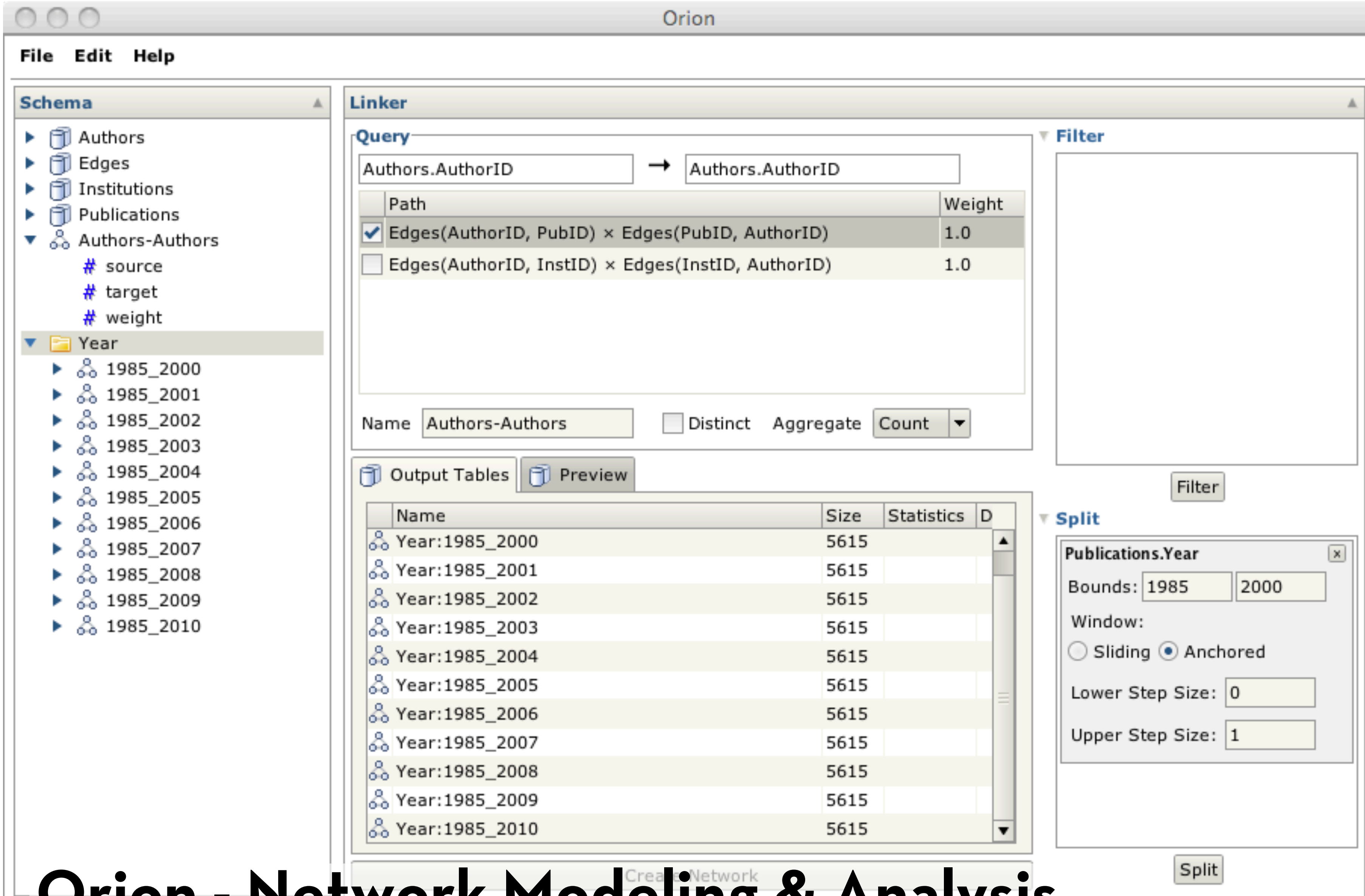
~50fps querying of visual summaries of **1B data points**.

Future Work

- Visualization specification interface
- Optimization considering resource constraints
- Integration with backend databases
- Server-side tile generation policies
- Activity modeling & prefetching schemes

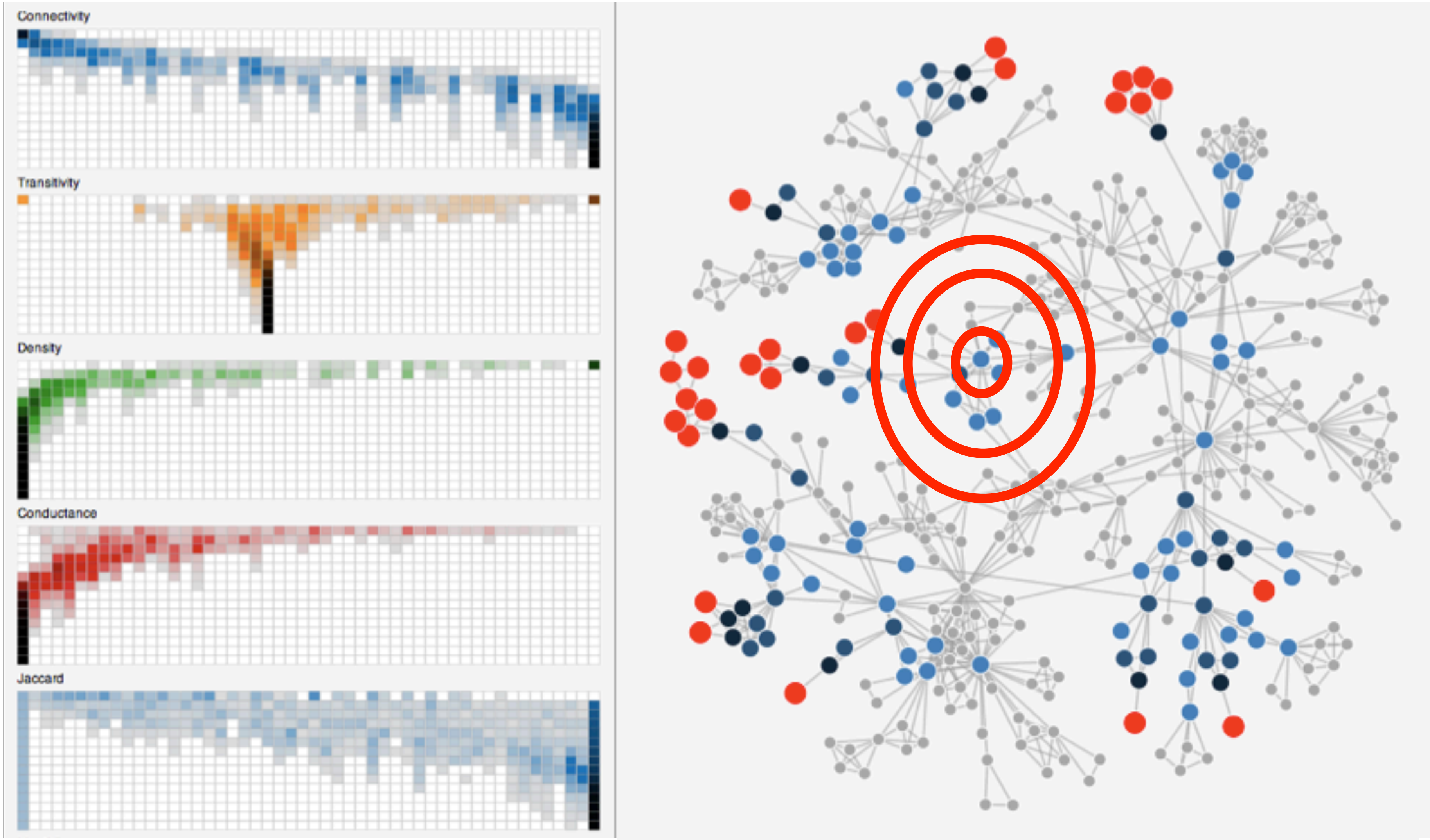

Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination



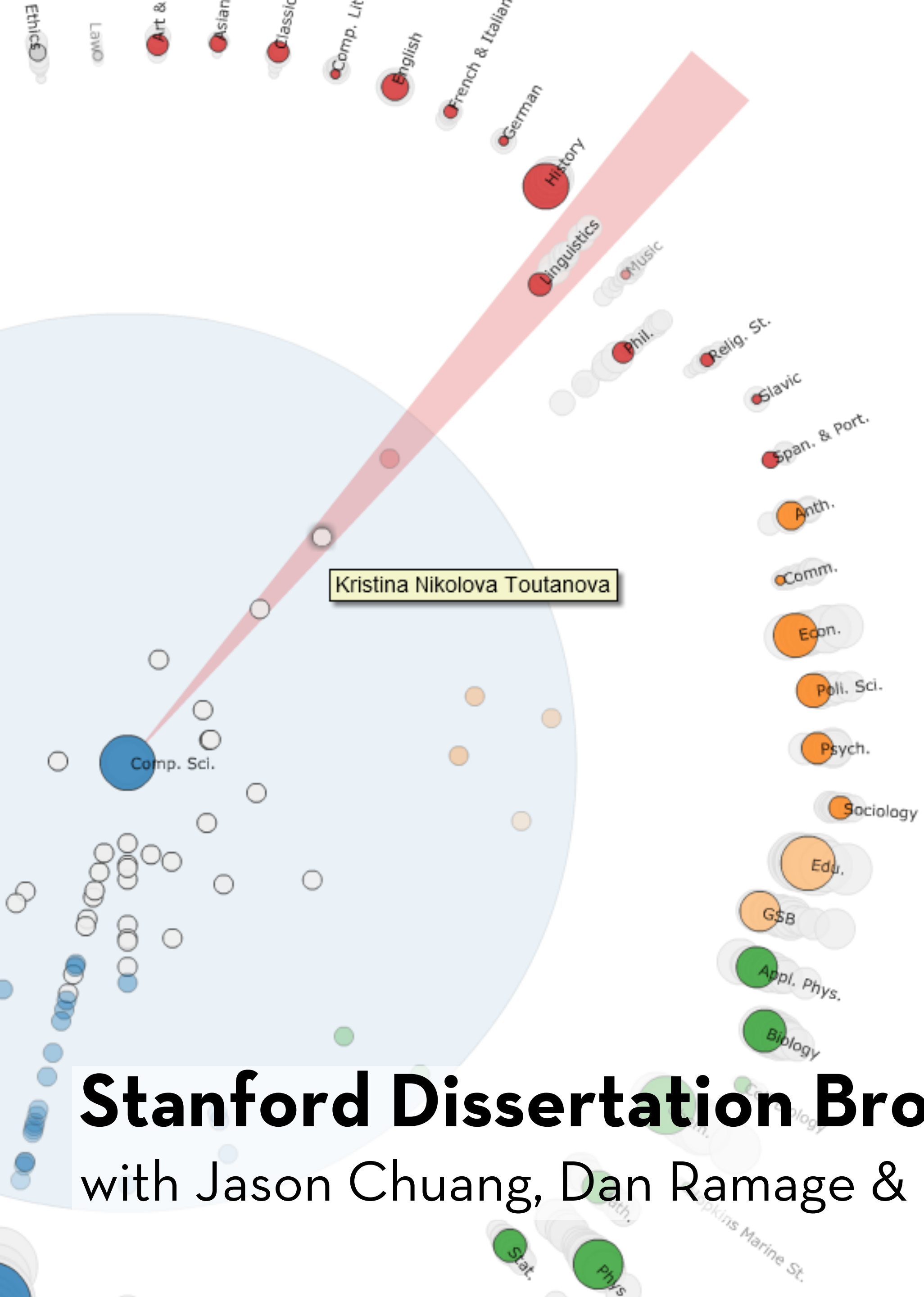
Orion - Network Modeling & Analysis

with Adam Perer [VAST'11]



GraphPrism

with Sanjay Kairam, Diana MacLean & Manolis Savva [AVI'12]



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

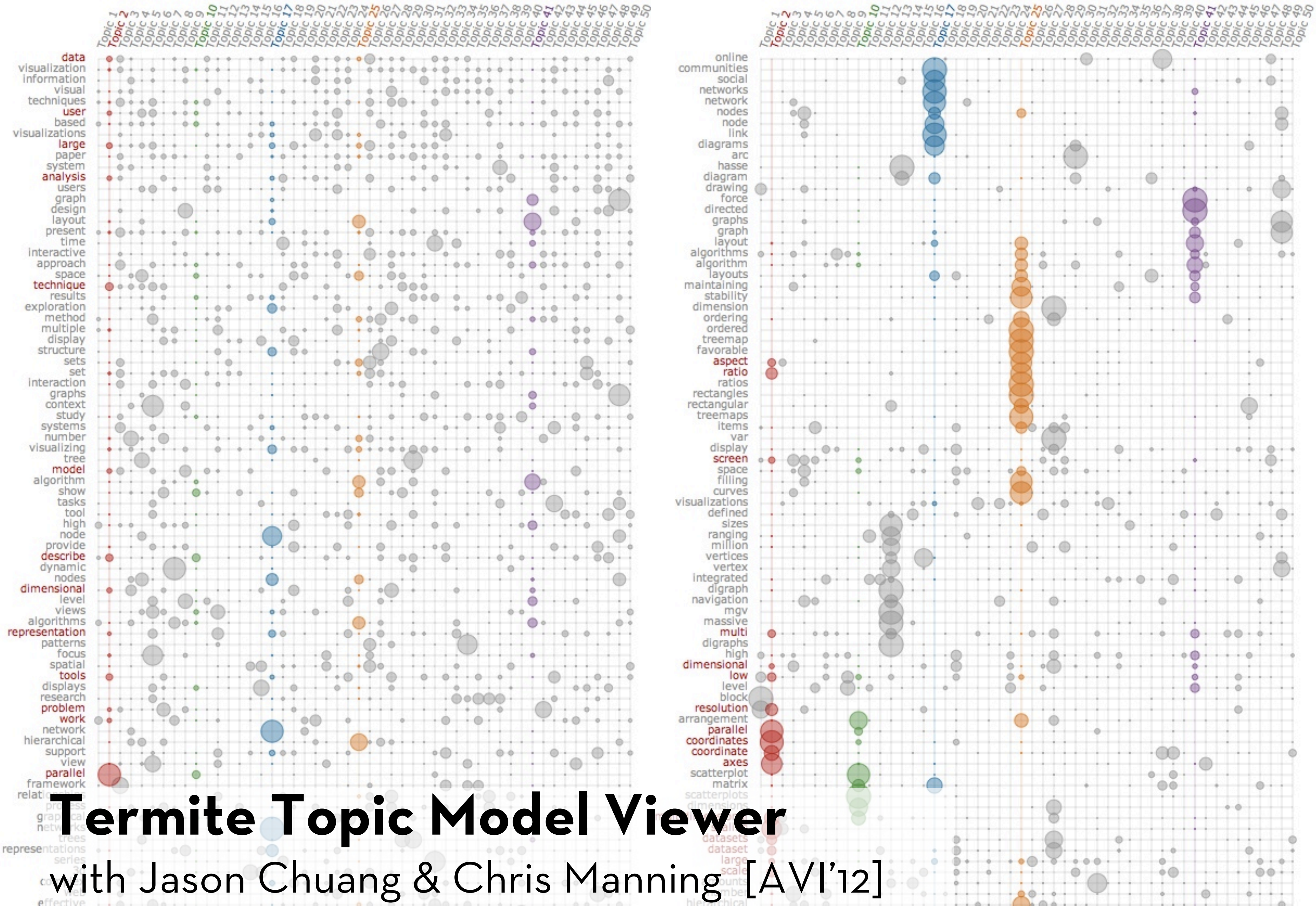
Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks--sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.

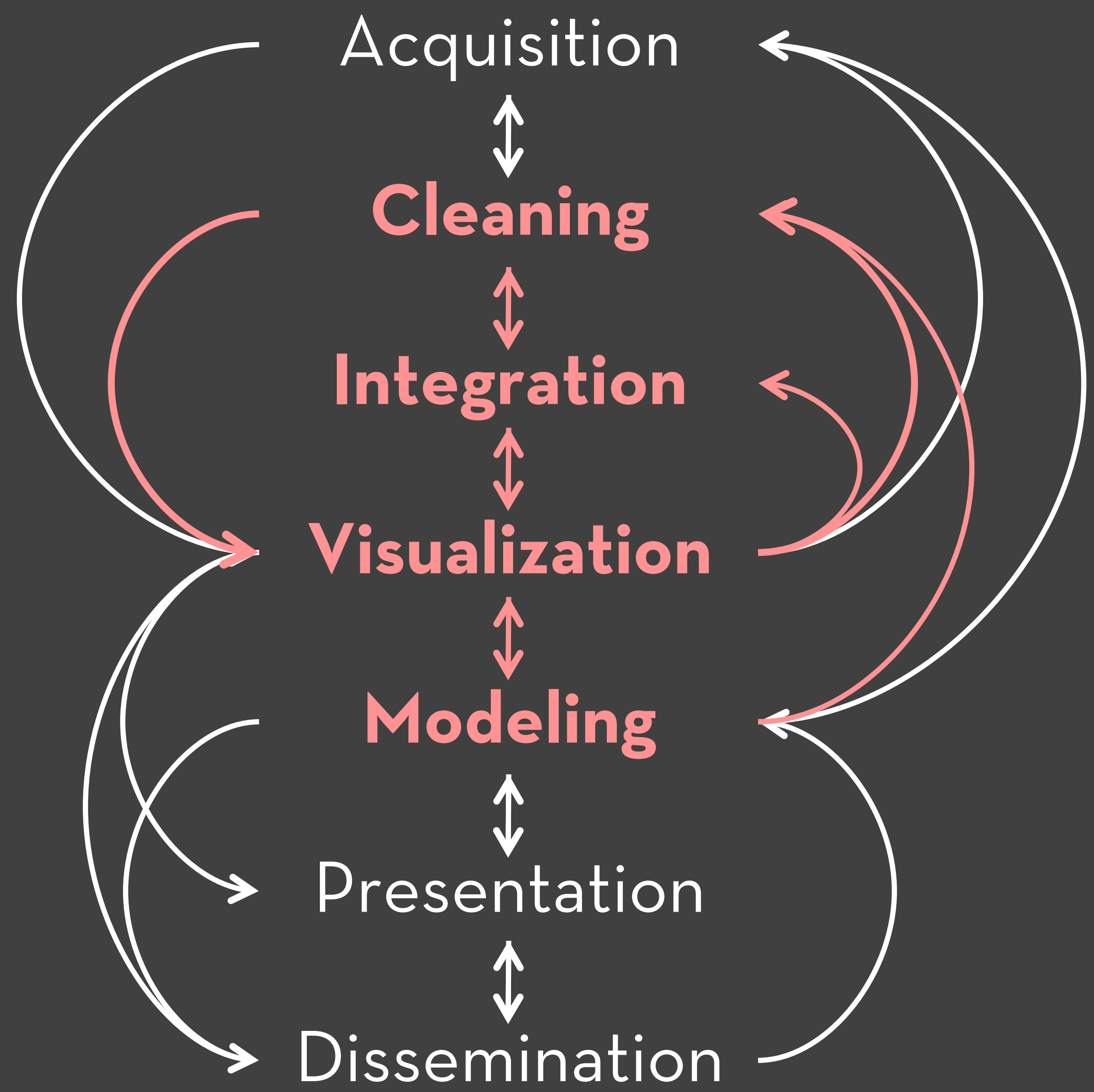
Stanford Dissertation Browser

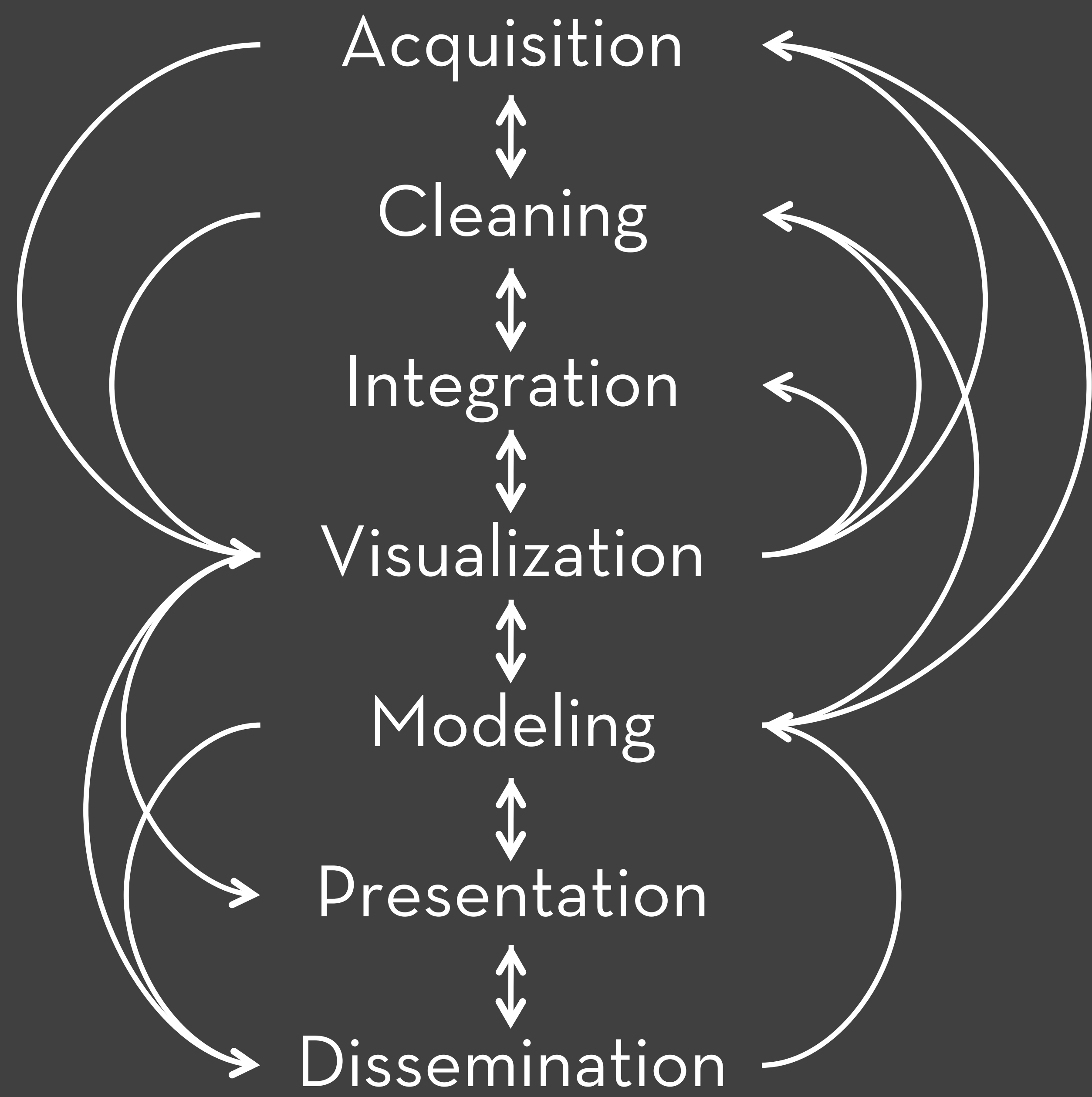
with Jason Chuang, Dan Ramage & Chris Manning [CHI'12]



Termite Topic Model Viewer

with Jason Chuang & Chris Manning [AVI'12]


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

Interactive

^ **Data Analysis**

http://vis.stanford.edu