Interactive ^ Data Analysis

Jeffrey Heer
@jeffrey_heer

Interactive Data Analysis

Jeffrey Heer
Stanford University

Interactive Data Analysis

Jeffrey Heer
University of Washington

How much data (bytes) did we produce in 2010?

2010: 1,200 exabytes

Gantz et al, 2008, 2010

2010: 1,200 exabytes

10x increase over 5 years

Gantz et al, 2008, 2010

The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, … because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

Graph Viewer

Graph Viewer

Roll-up by:

All

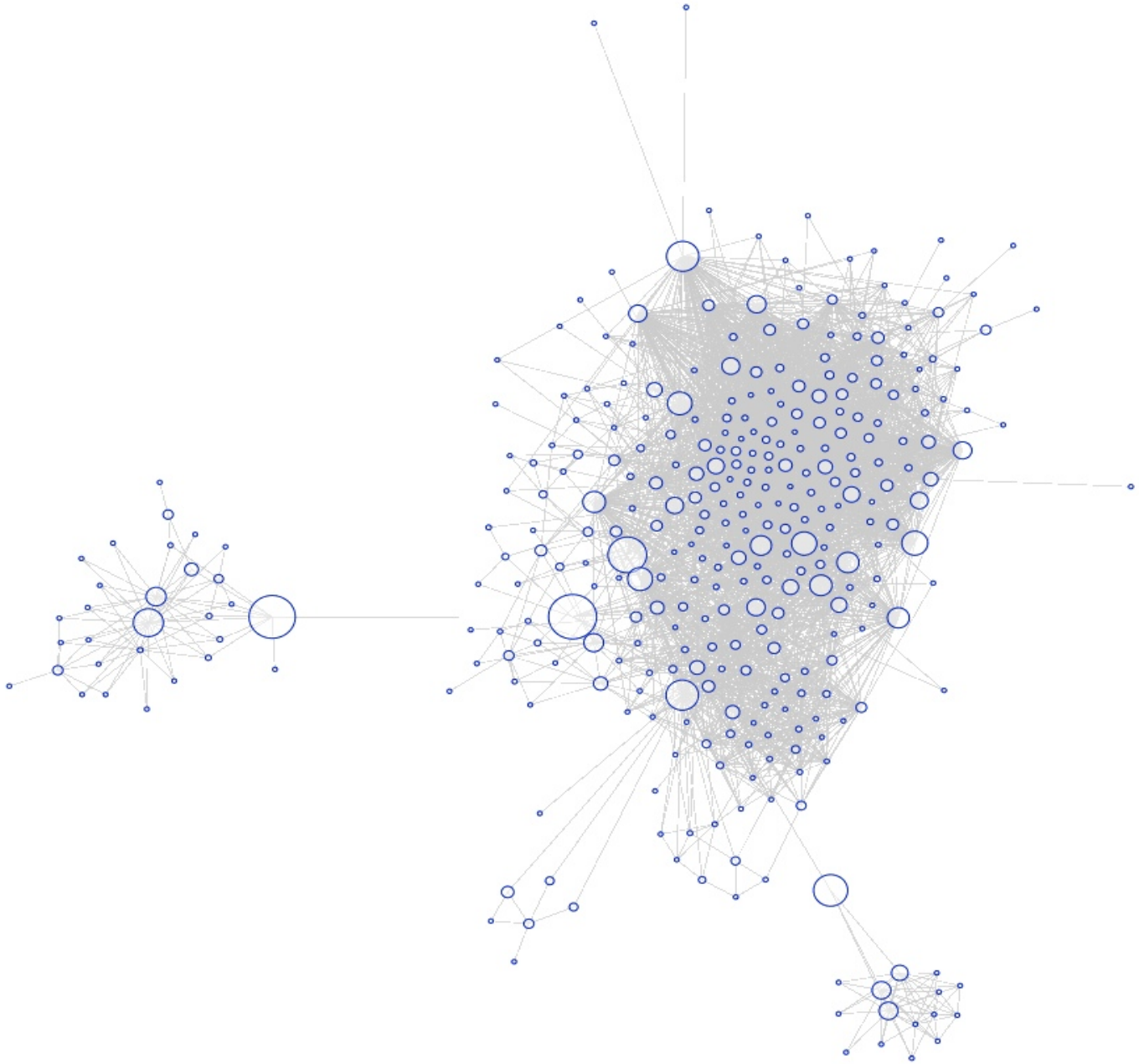
Visualization:

Node-Link

Sort by:

None

Edge centrality filters:



Images

Animate

Graph Viewer

Graph Viewer

Roll-up by:

All

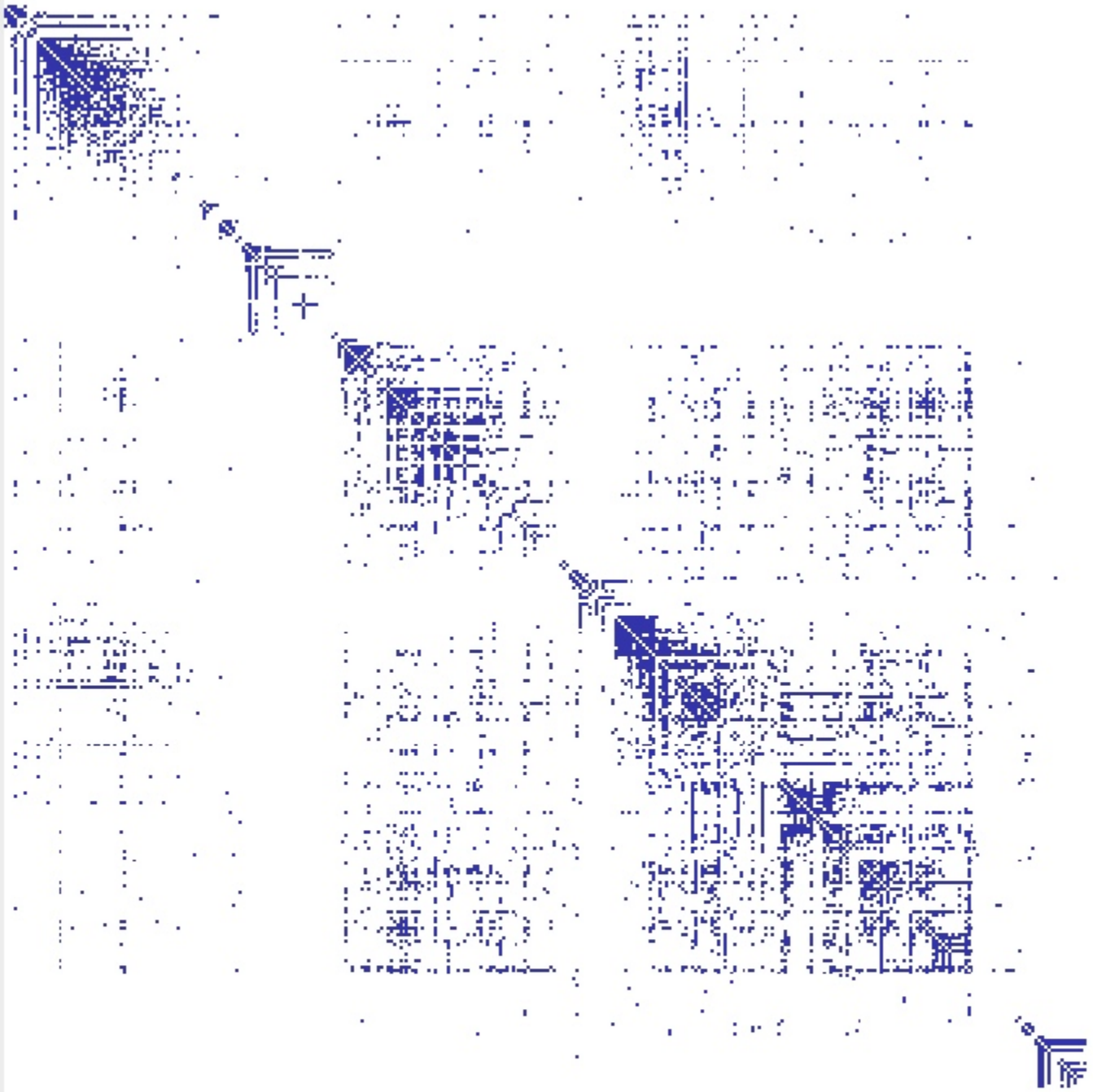
Visualization:

Matrix

Sort by:

Linkage

Edge centrality filters:



Graph Viewer

Graph Viewer

Roll-up by:

All

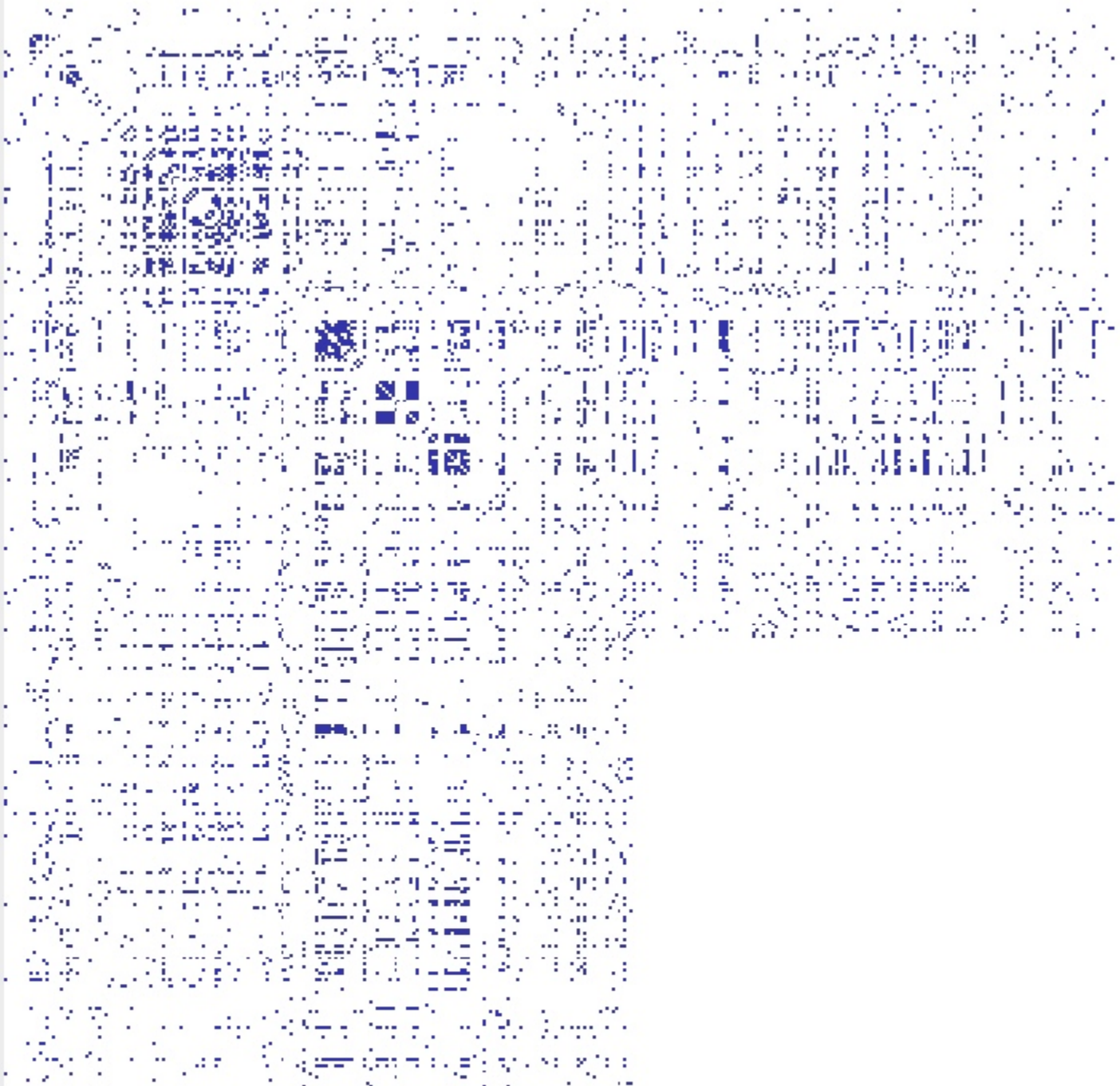
Visualization:

Matrix

Sort by:

None

Edge centrality filters:



Visualization

Acquisition

Cleaning

Integration

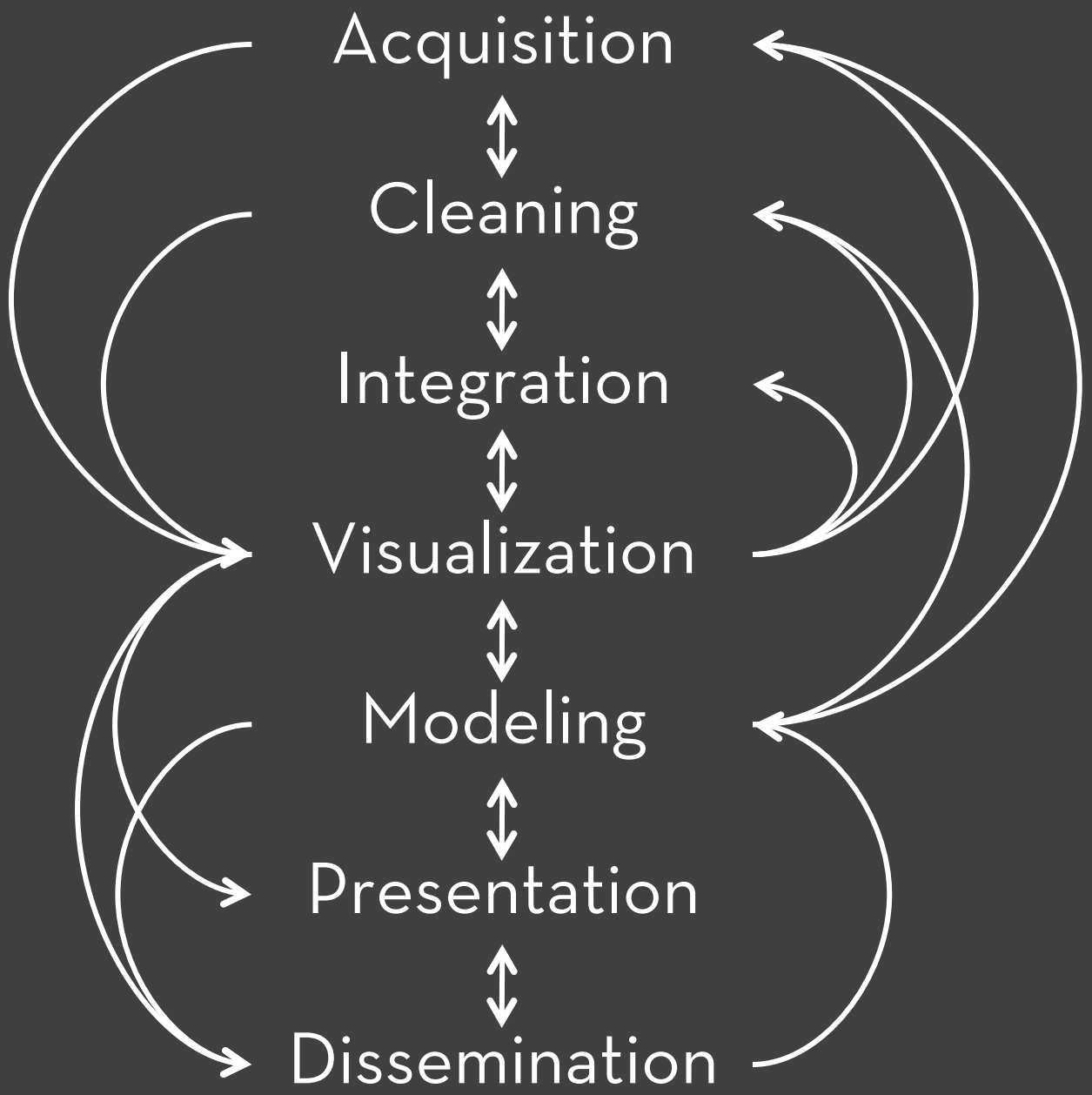
Visualization

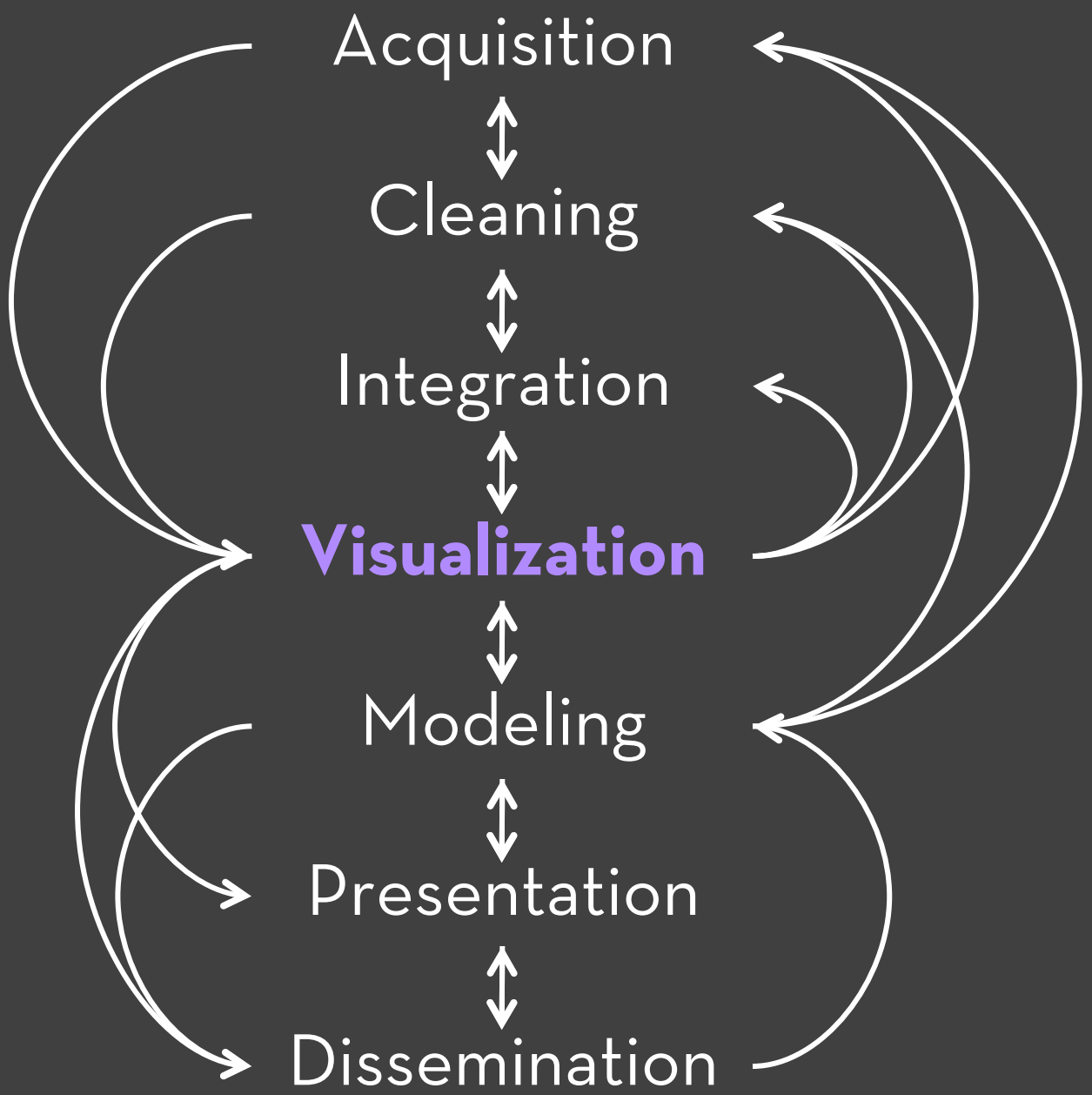
Modeling

Presentation

Dissemination

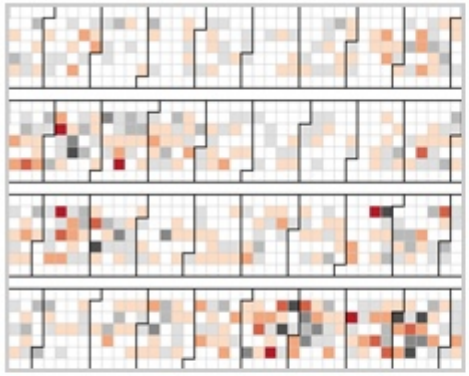
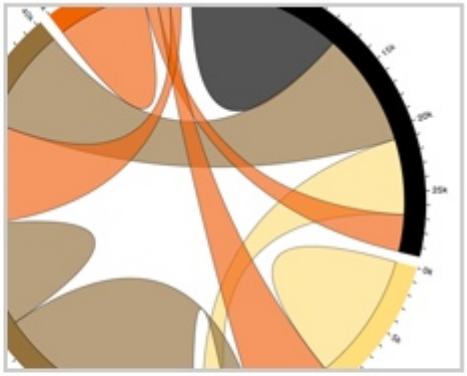
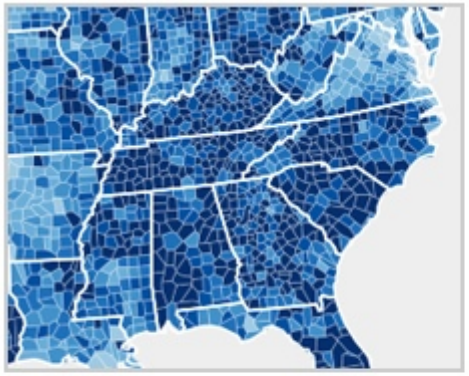
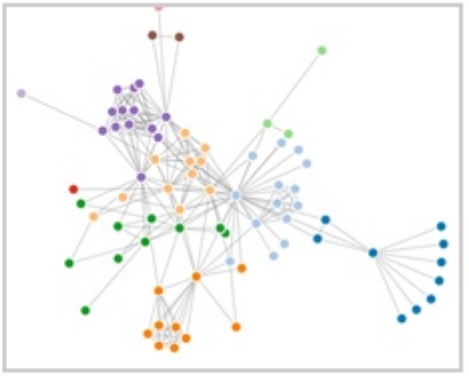
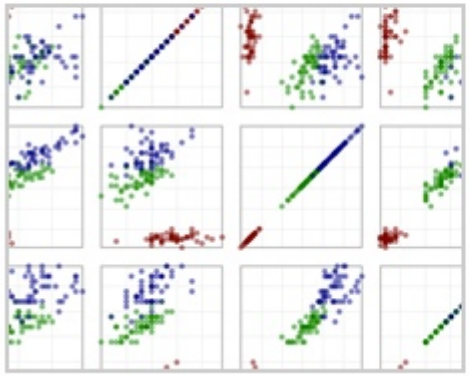
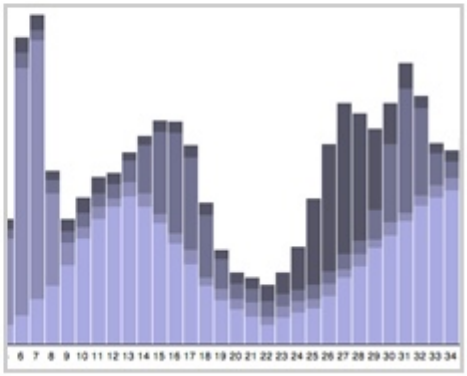
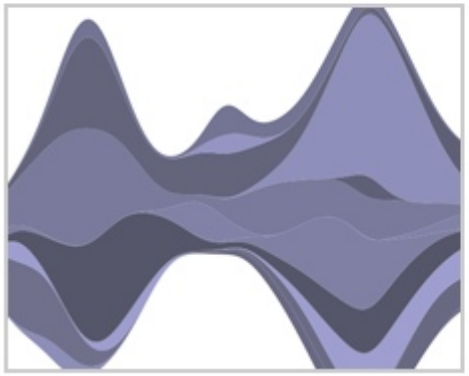
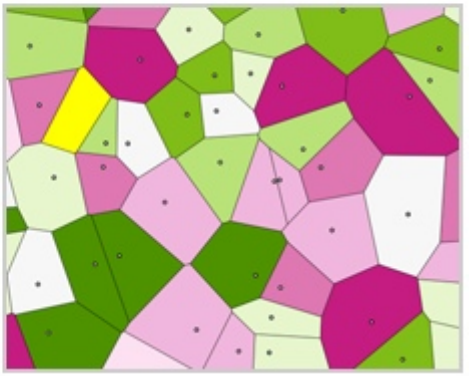

Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

How might we support **expressive** and **effective** visualization designs?

d3.js Data-Driven Documents

with Mike Bostock & Vadim Ogievetsky

Popular Watched Repositorie

GitHub, Inc. [US] https://github.com/popular/watched

coffee-script

jashkenas 4,733 watching

Unfancy JavaScript

socket.io

LearnBoost 4,454 watching

Realtime application framework for Node.JS, with HTML5 WebSockets and cross-browser fallbacks support.

gitignore

github 4,437 watching

A collection of useful .gitignore templates

underscore

documentcloud 4,434 watching

JavaScript's utility _ belt

d3

mbostock 4,286 watching

A JavaScript visualization library for HTML and SVG.

linux

torvalds 4,280 watching

Linux kernel source tree

symfony

symfony 4,209 watching

The Symfony2 PHP framework

paperclip

thoughtbot 4,031 watching

Easy file attachment management for ActiveRecord

redis

antirez 3,941 watching

Redis is an in-memory database that persists on disk. The data model is key-value, but many different kind of values are supported: Strings, Lists, Sets, Sorted Sets, Hashes

Popular Watched Repositorie

GitHub, Inc. [US] https://github.com/popular/watched

coffee-script jashkenas 4,733 watching

Unfancy JavaScript

socket.io LearnBoost 4,454 watching

Realtime application framework for Node.JS, with HTML5 WebSockets and cross-browser fallbacks support.

gitignore github 4,437 watching

A collection of useful .gitignore templates

underscore

JavaScript's utility _ belt

5th most starred project on GitHub

d3 mbostock 4,286 watching

A JavaScript visualization library for HTML and SVG.

linux torvalds 4,280 watching

Linux kernel source tree

symfony symfony 4,209 watching

The Symfony2 PHP framework

paperclip thoughtbot 4,031 watching

Easy file attachment management for ActiveRecord

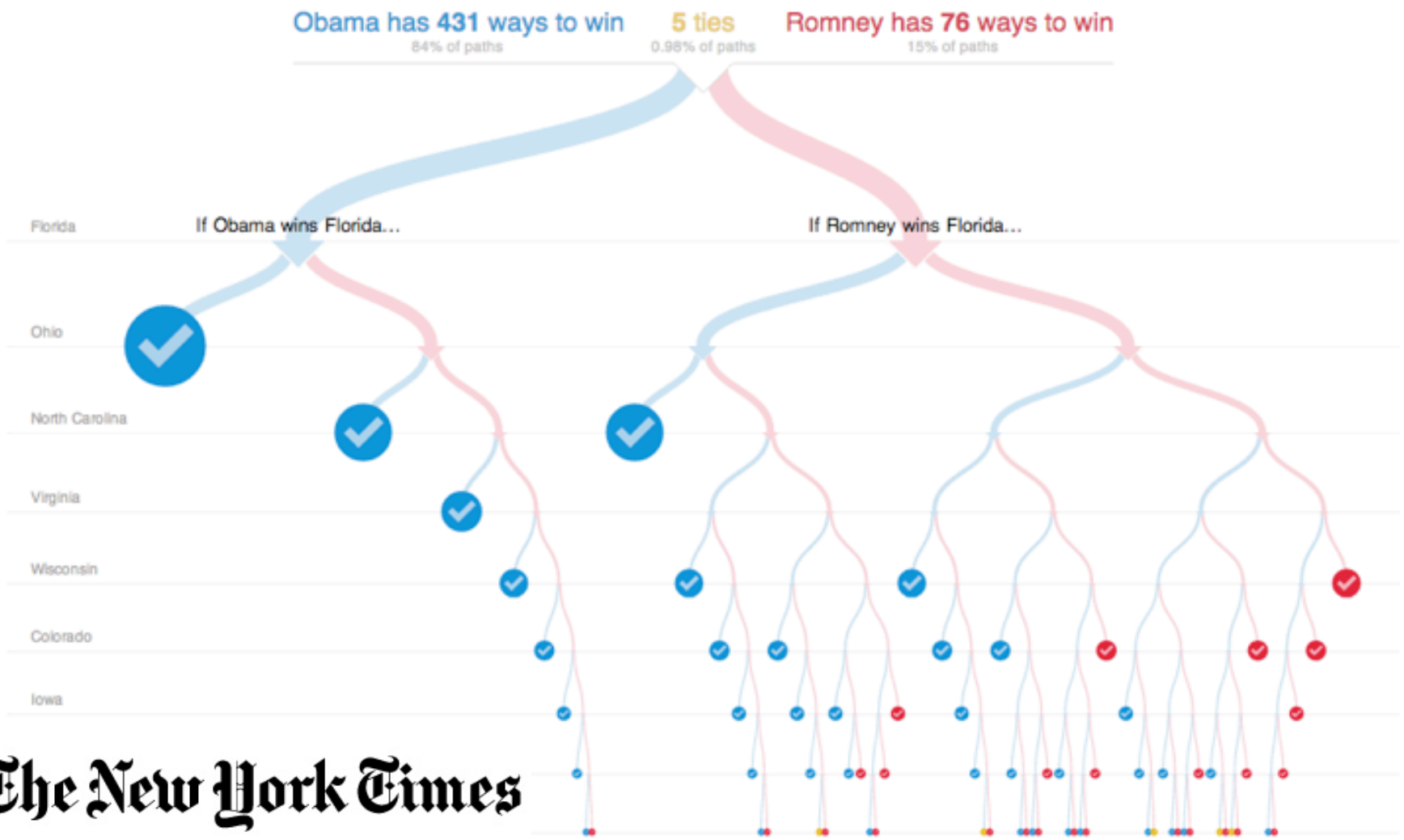
redis antirez 3,941 watching

Redis is an in-memory database that persists on disk. The data model is key-value, but many different kind of values are supported: Strings, Lists, Sets, Sorted Sets, Hashes

512 Paths to the White House

Select a winner in the most competitive states below to see all the paths to victory available for either candidate.

Fla.	Ohio	N.C.	Va.	Wis.	Colo.	Iowa	Nev.	N.H.
Dem Rep	Dem Rep	Dem Rep	Dem Rep	Dem Rep	Dem Rep	Dem Rep	Dem Rep	Dem Rep

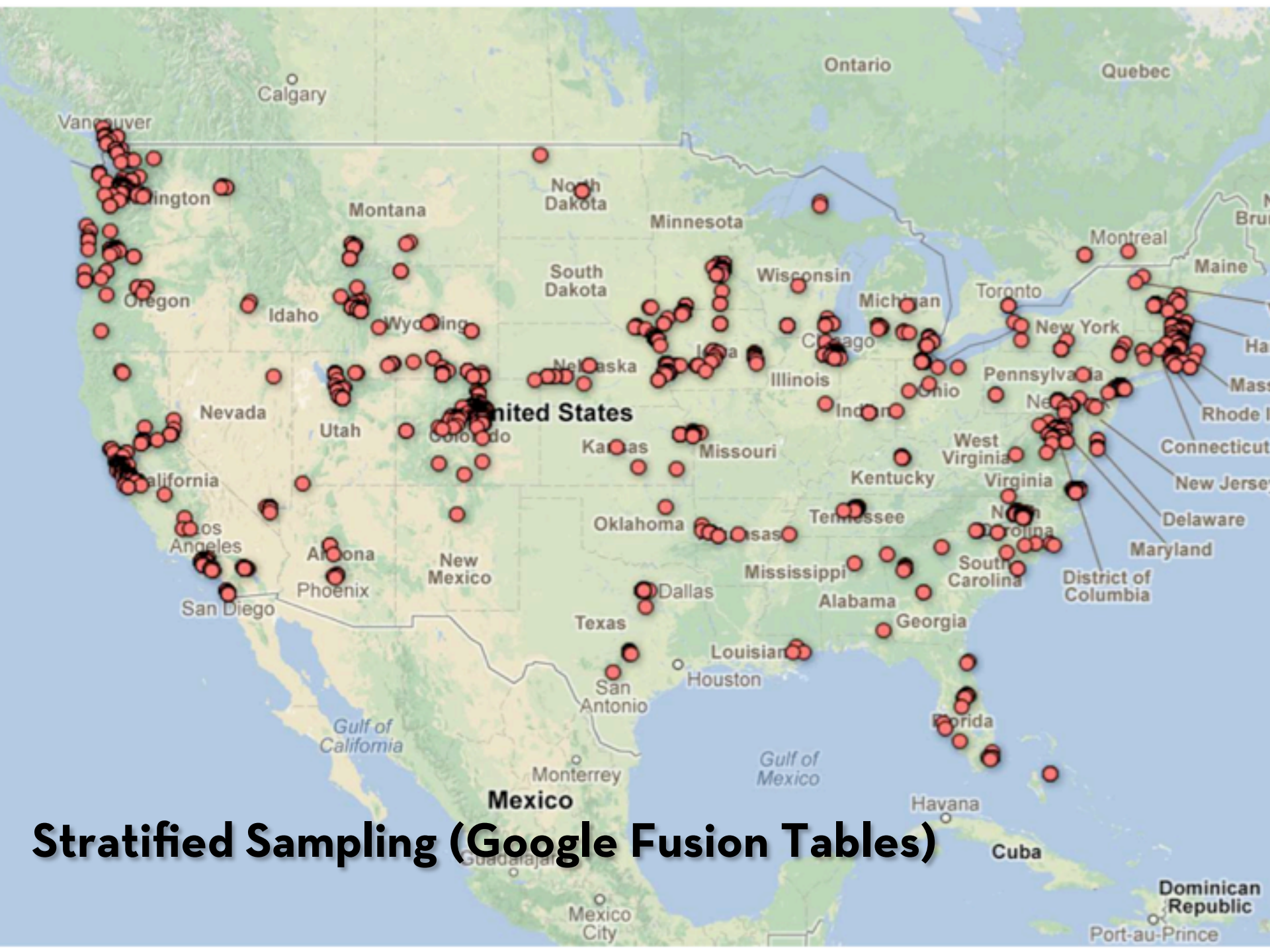


The New York Times

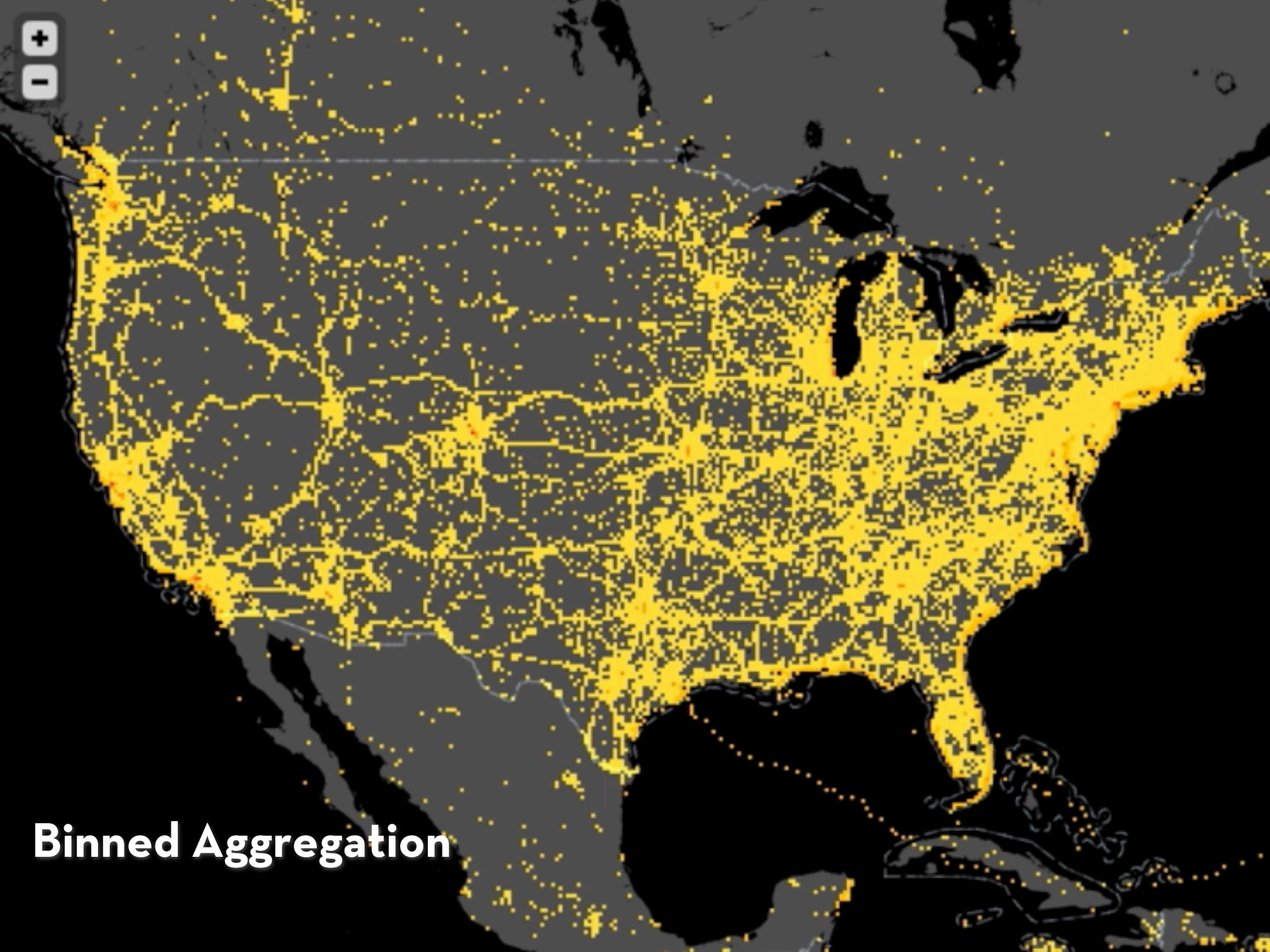

stamen design


stamen design
d3
d3

How can we visualize and interact with **billion+ record** databases in real-time?

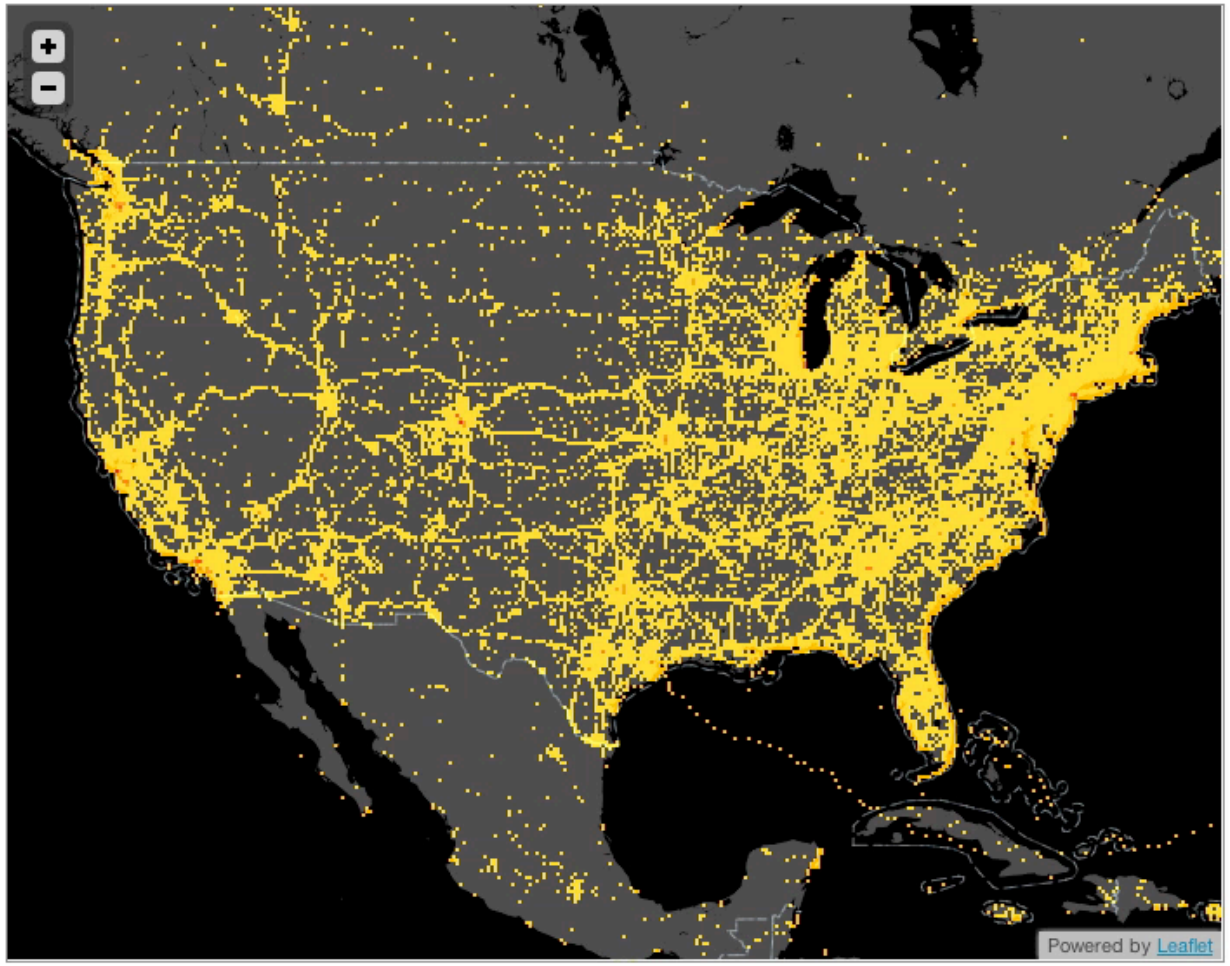


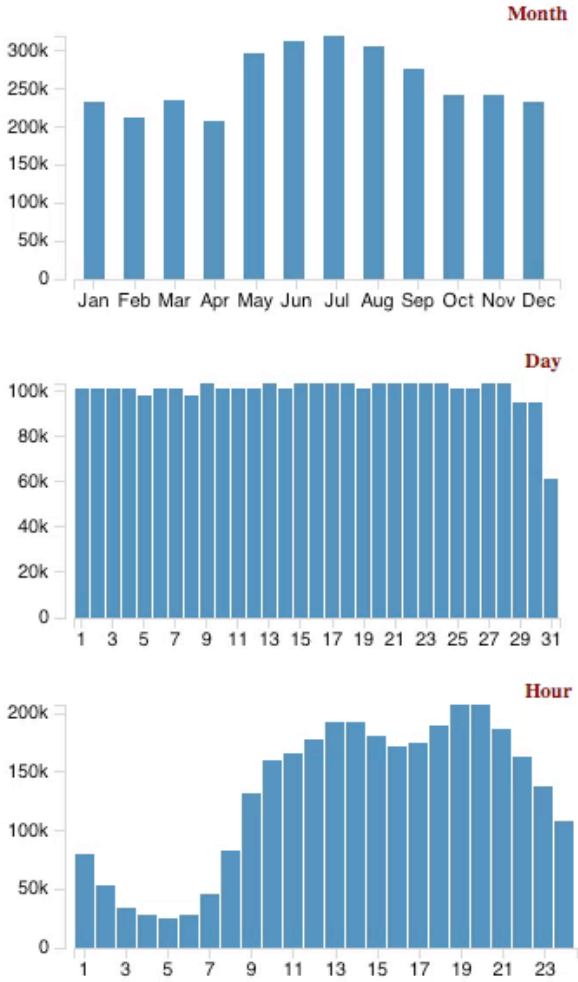
Stratified Sampling (Google Fusion Tables)


Binned Aggregation

imMens: Real-Time Visual Querying of Big Data

with Zhicheng (Leo) Liu & Biye Jiang





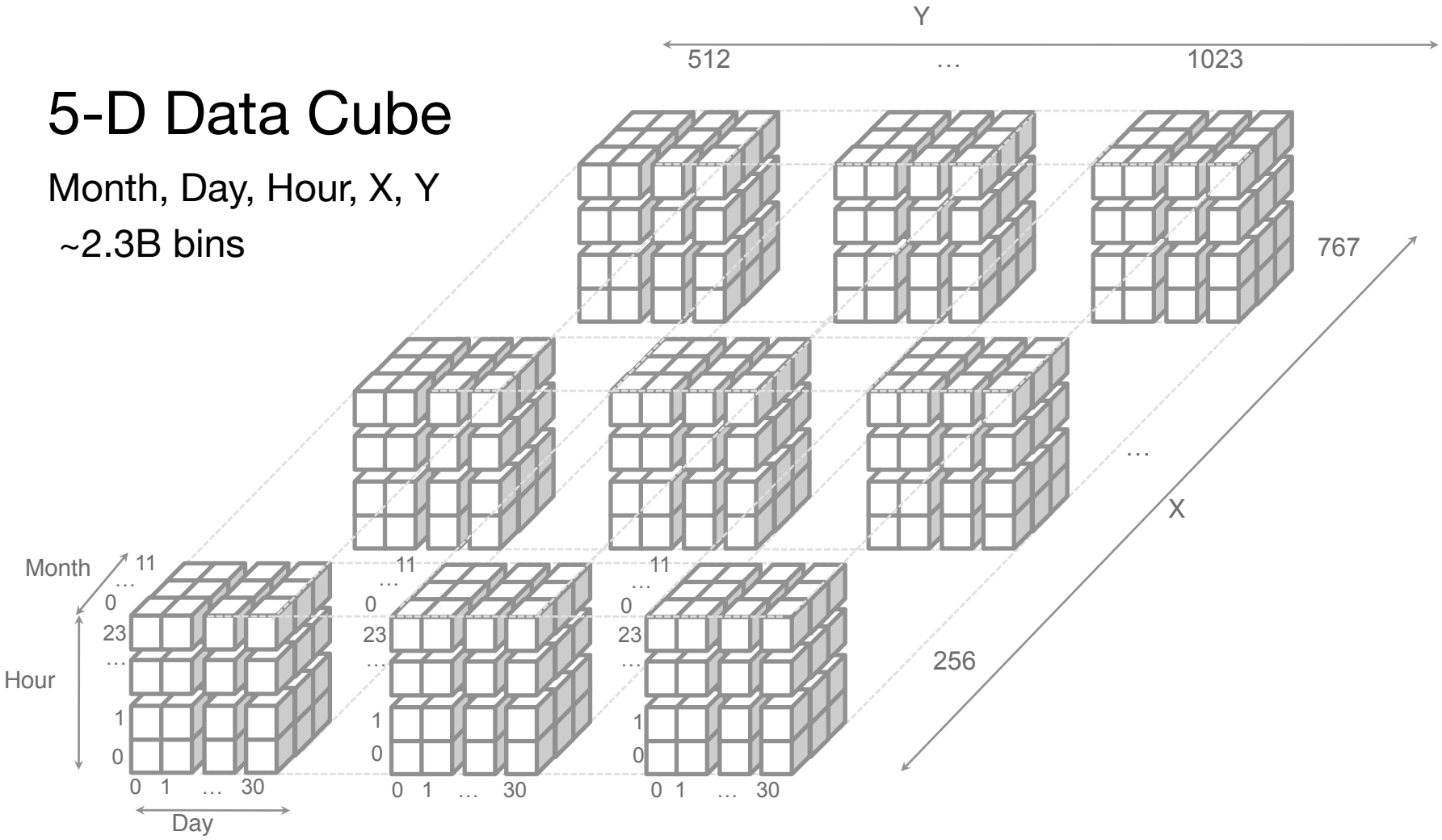
imMens: Real-Time Visual Querying of Big Data

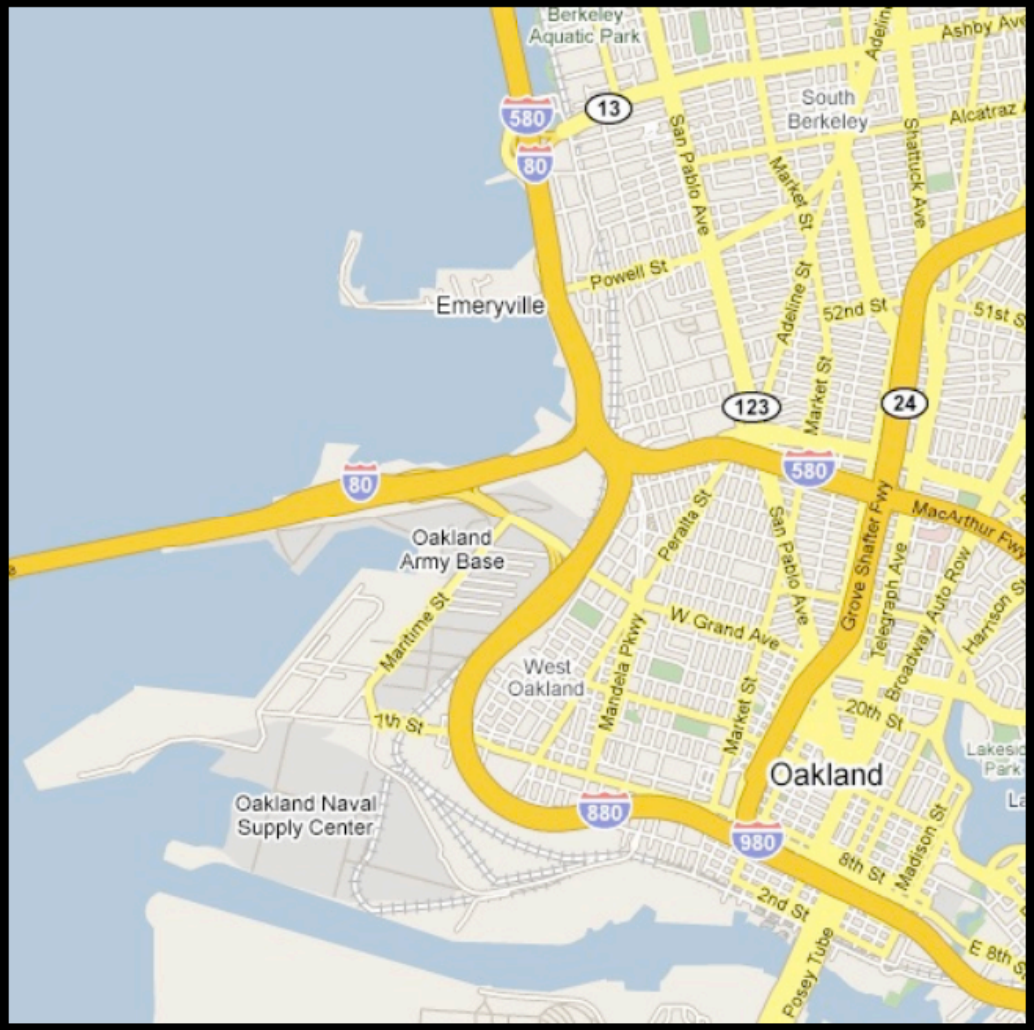
with Zhicheng (Leo) Liu & Biye Jiang

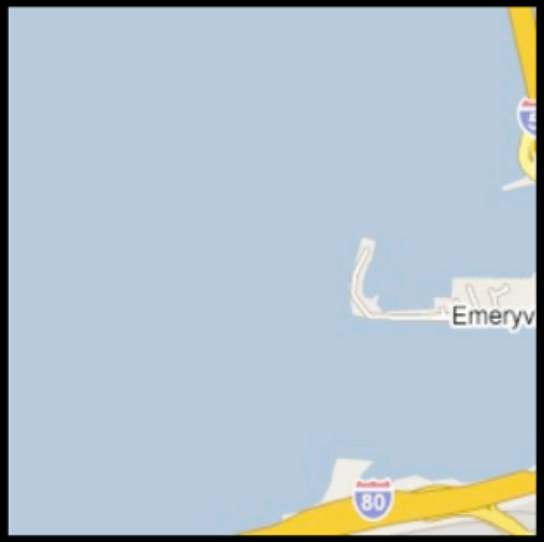
5-D Data Cube

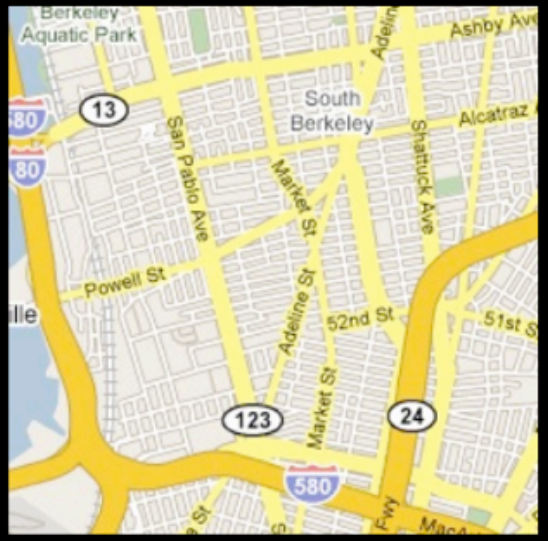
Month, Day, Hour, X, Y

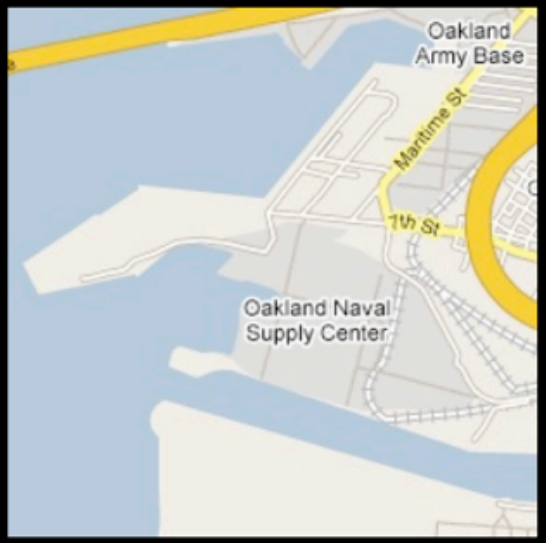
~2.3B bins

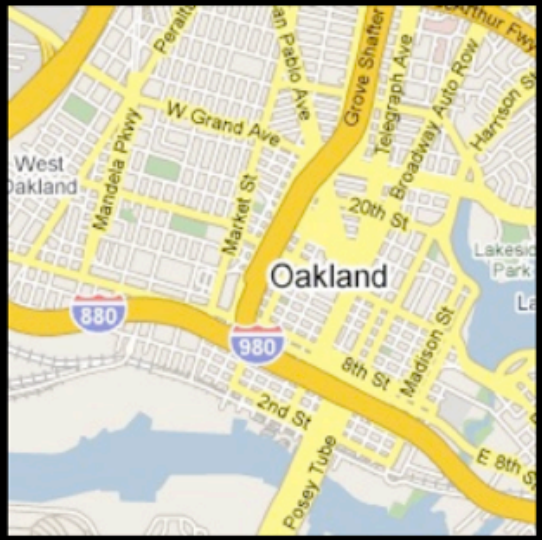



Berkeley
Aquatic Park
Ashby Ave
580
13
South
Berkeley
Alcatraz
80
San Pablo Ave
Market St
Shattuck Ave
Powell St
Emeryville
Adeline St
52nd St
51st St
123
Market St
24
80
580
MacArthur Fwy
Oakland
Army Base
Peralta St
San Pablo Ave
Grove Shafter Fwy
Telegraph Ave
Broadway Auto Row
Harrison St
Maritime 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
E 8th St
Posey Tube

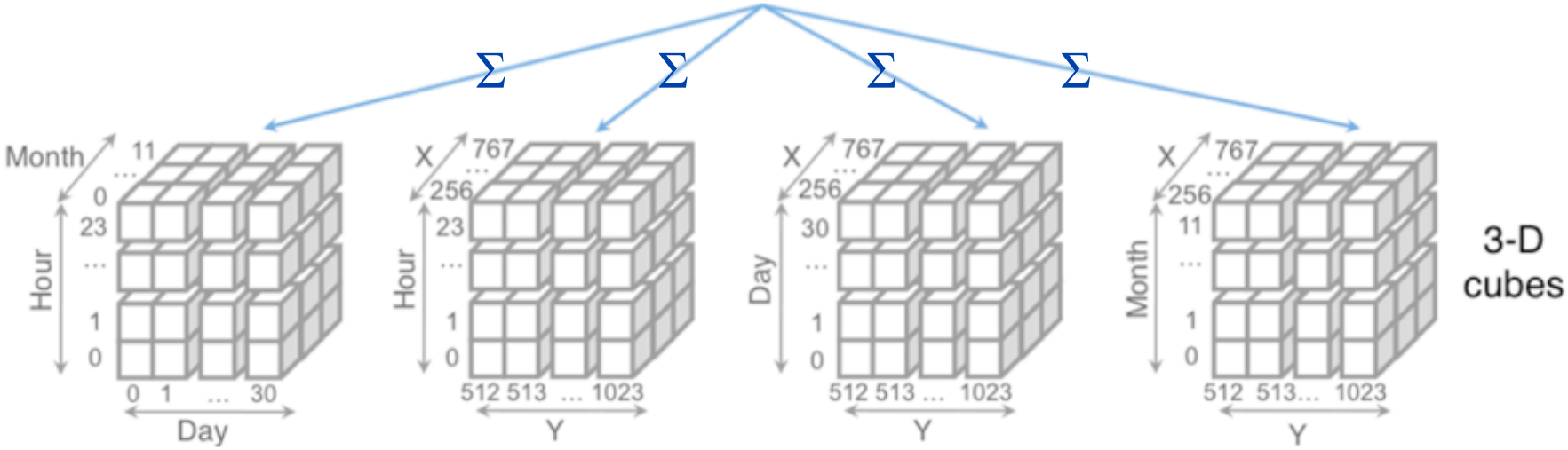

Emeryv
80

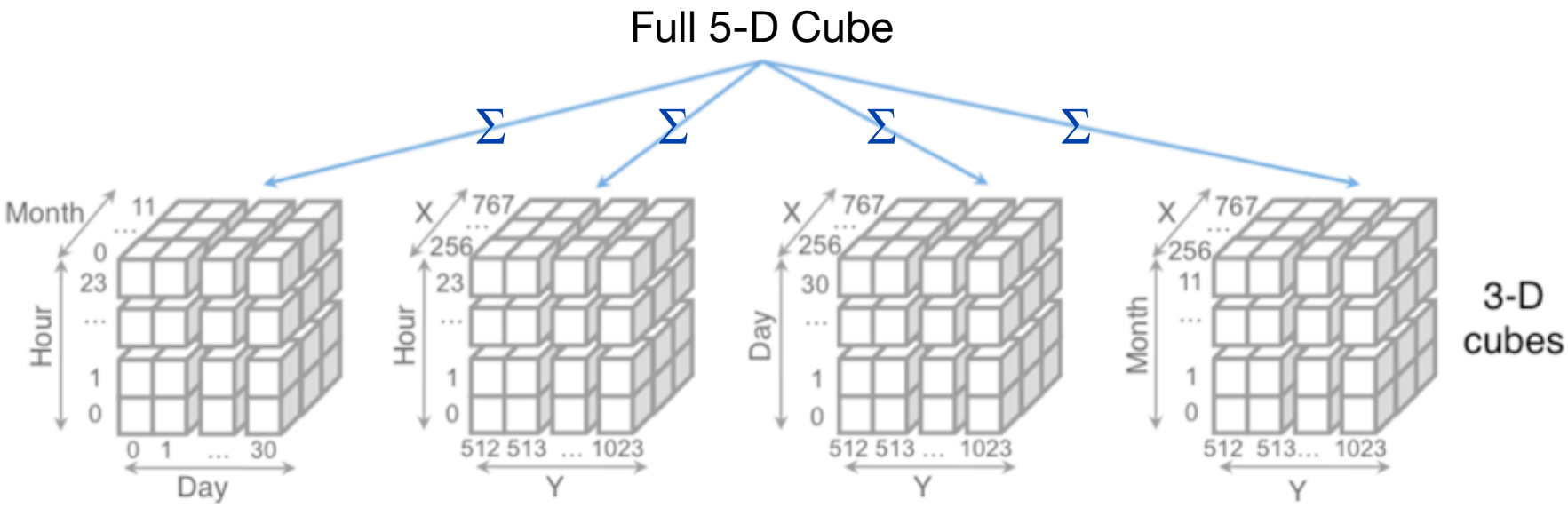

Berkeley
Aquatic Park
Ashby Ave
13
South
Berkeley
Alcatraz
San Pablo Ave
Market St
Shattuck Ave
Powell St
Adeline St
52nd St
51st S
123
24
580


Oakland
Army Base
Maritime St
7th St
Oakland Naval
Supply Center

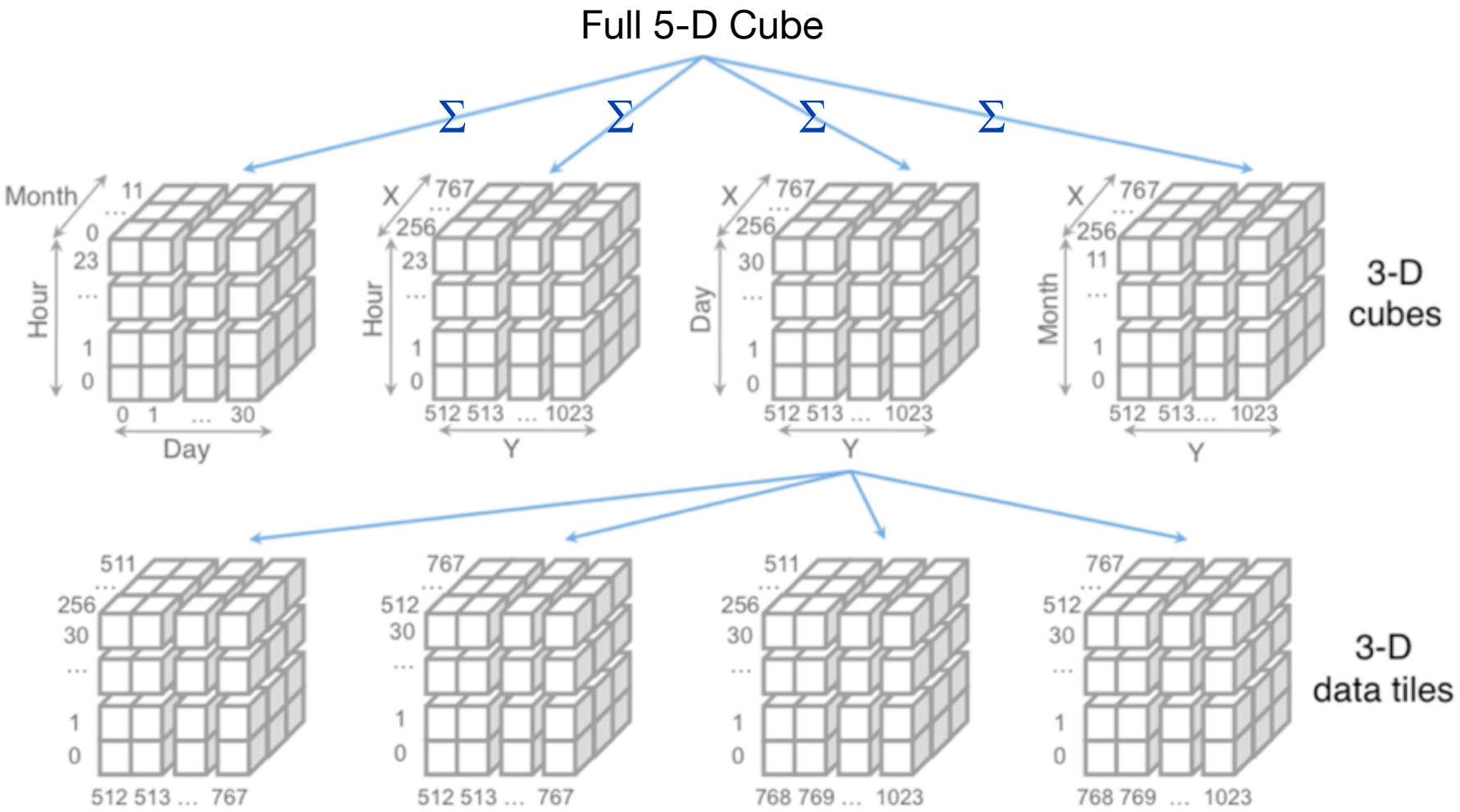

W Grand Ave
Grove Shafter
Telegraph Ave
Broadway Auto Row
Harrison St
Mandela Pkwy
West
Oakland
Market St
20th St
Oakland
880
980
8th St
Madison St
2nd St
E 8th St
Posey Tube

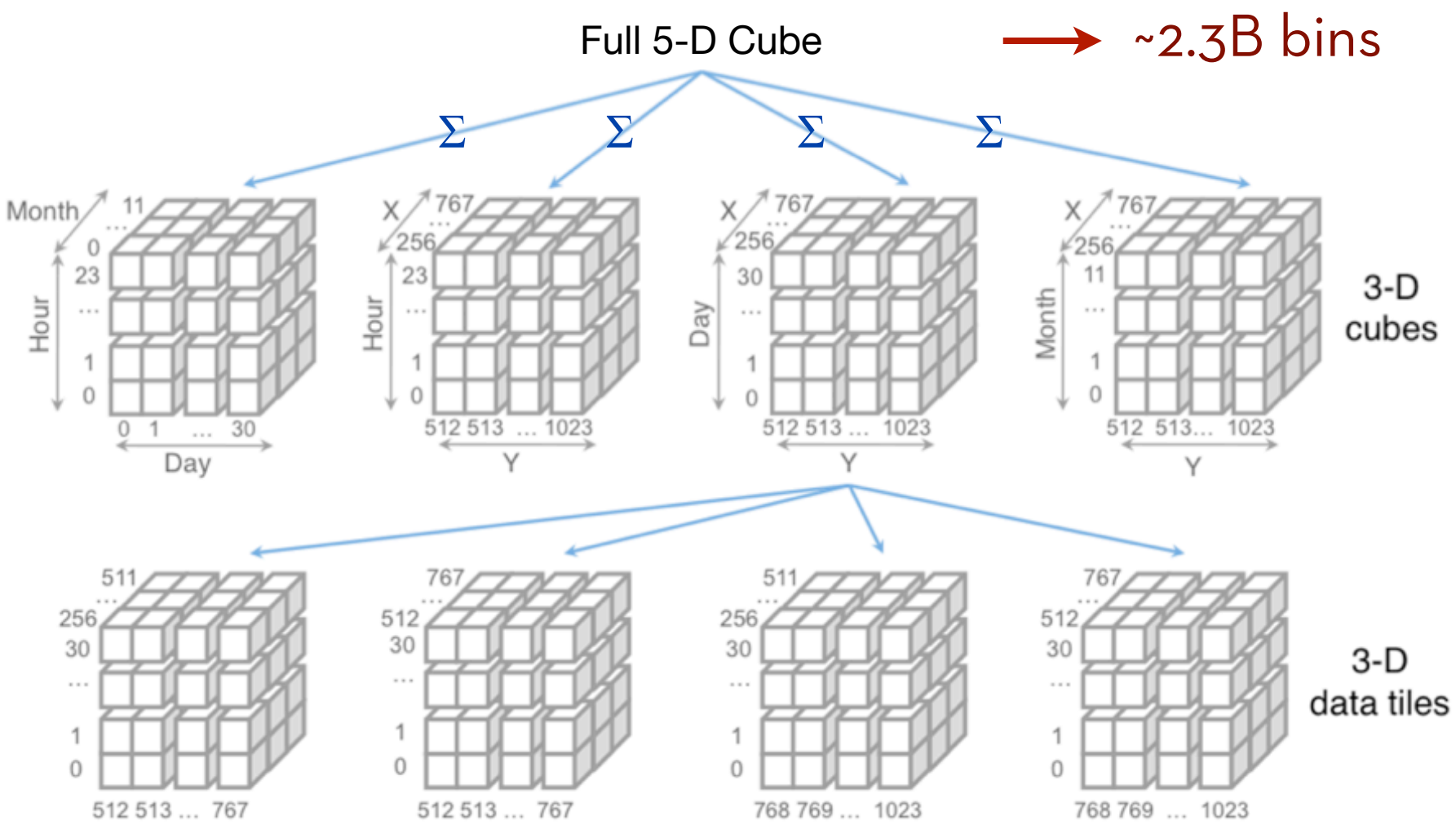
Full 5-D Cube

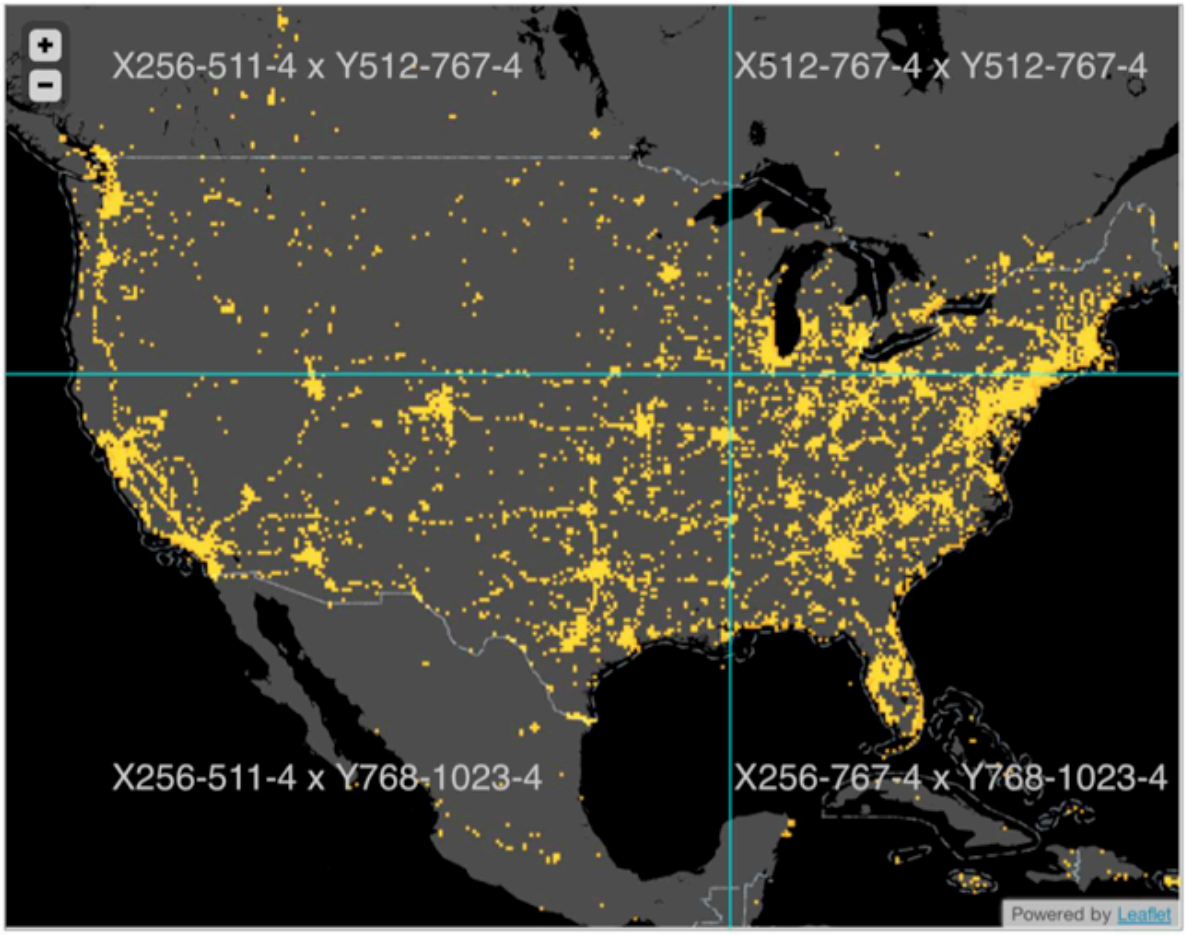

Full 5-D Cube
Σ
Σ
Σ
Σ
Month
11
0
Hour
23
1
0
0 1 … 30
Day
X
767
256
Hour
23
1
0
512 513 … 1023
Y
X
767
256
Day
30
1
0
512 513 … 1023
Y
X
767
256
Month
11
1
0
512 513… 1023
Y
3-D cubes

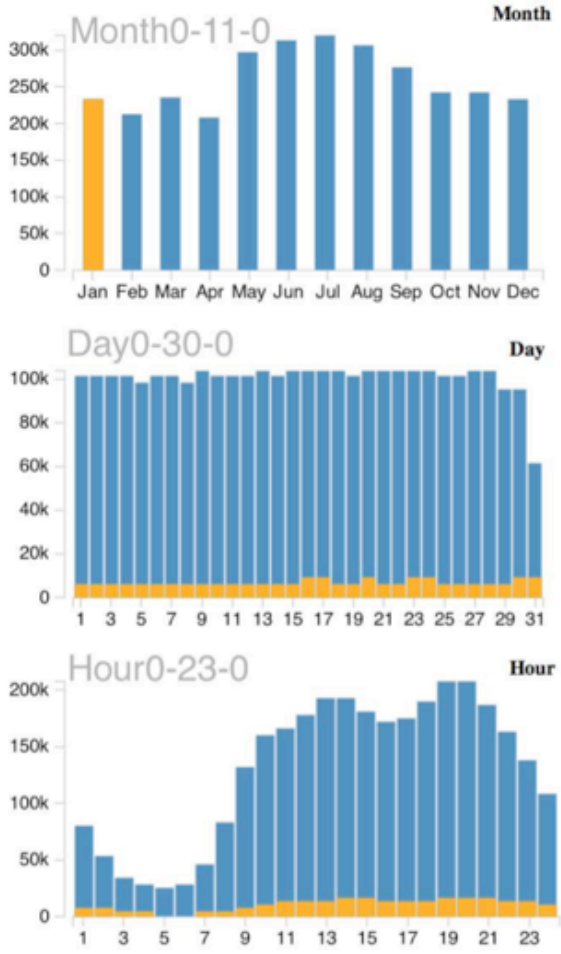


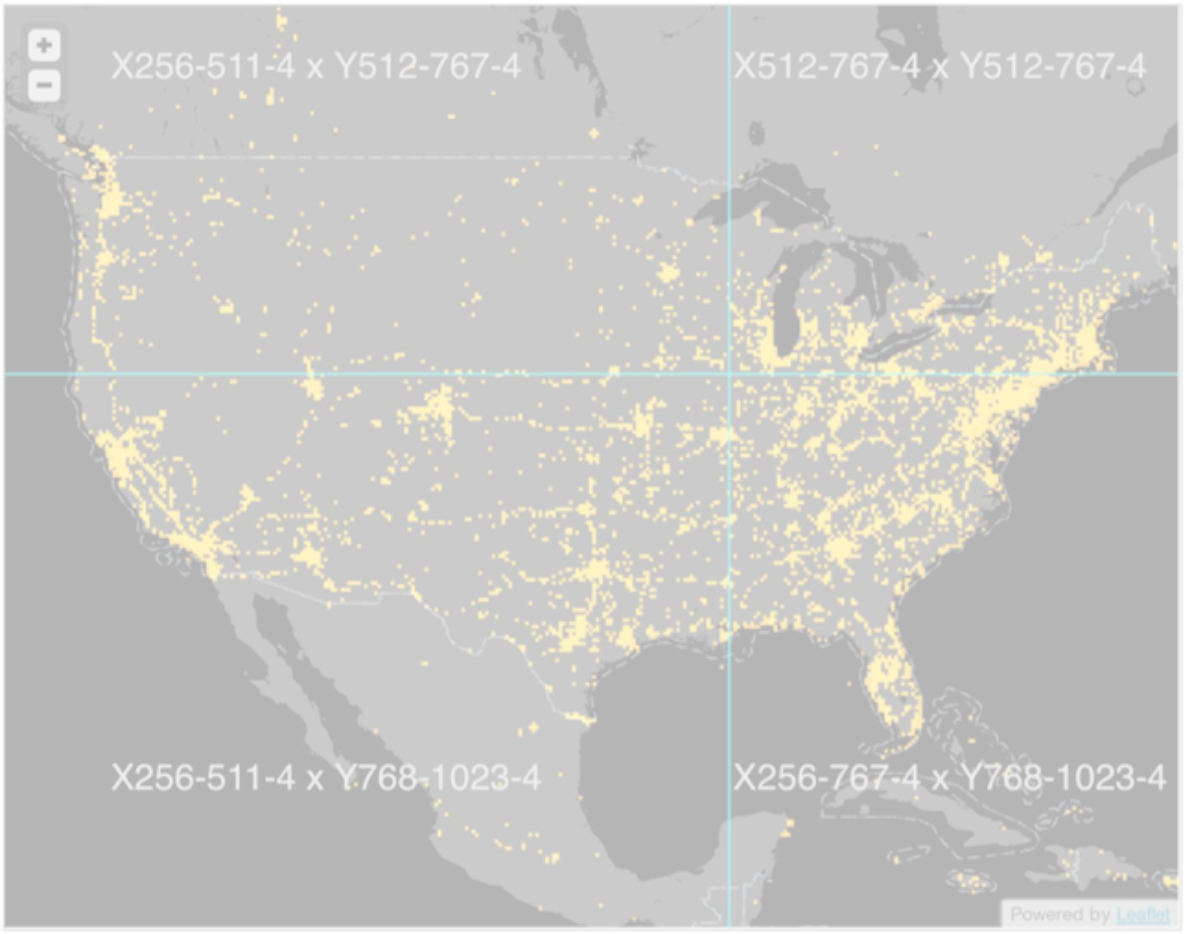
For any pair of 1D or 2D binned plots, the maximum number of dimensions needed to support brushing & linking is **four**.

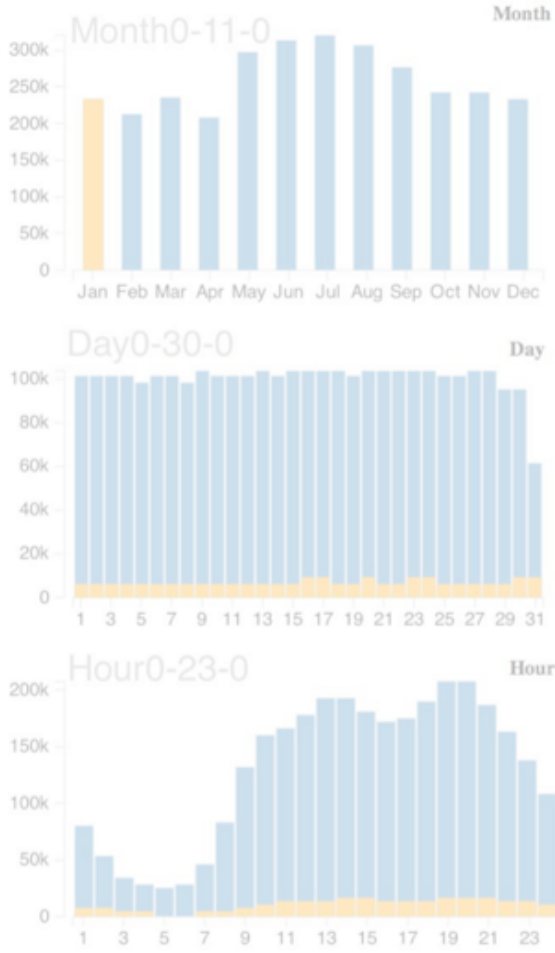

Full 5-D Cube
Σ
Σ
Σ
Σ
Month
Hour
Day
X
Hour
Y
X
Day
Y
X
Month
Y
3-D cubes
3-D data tiles
13 3-D Data Tiles

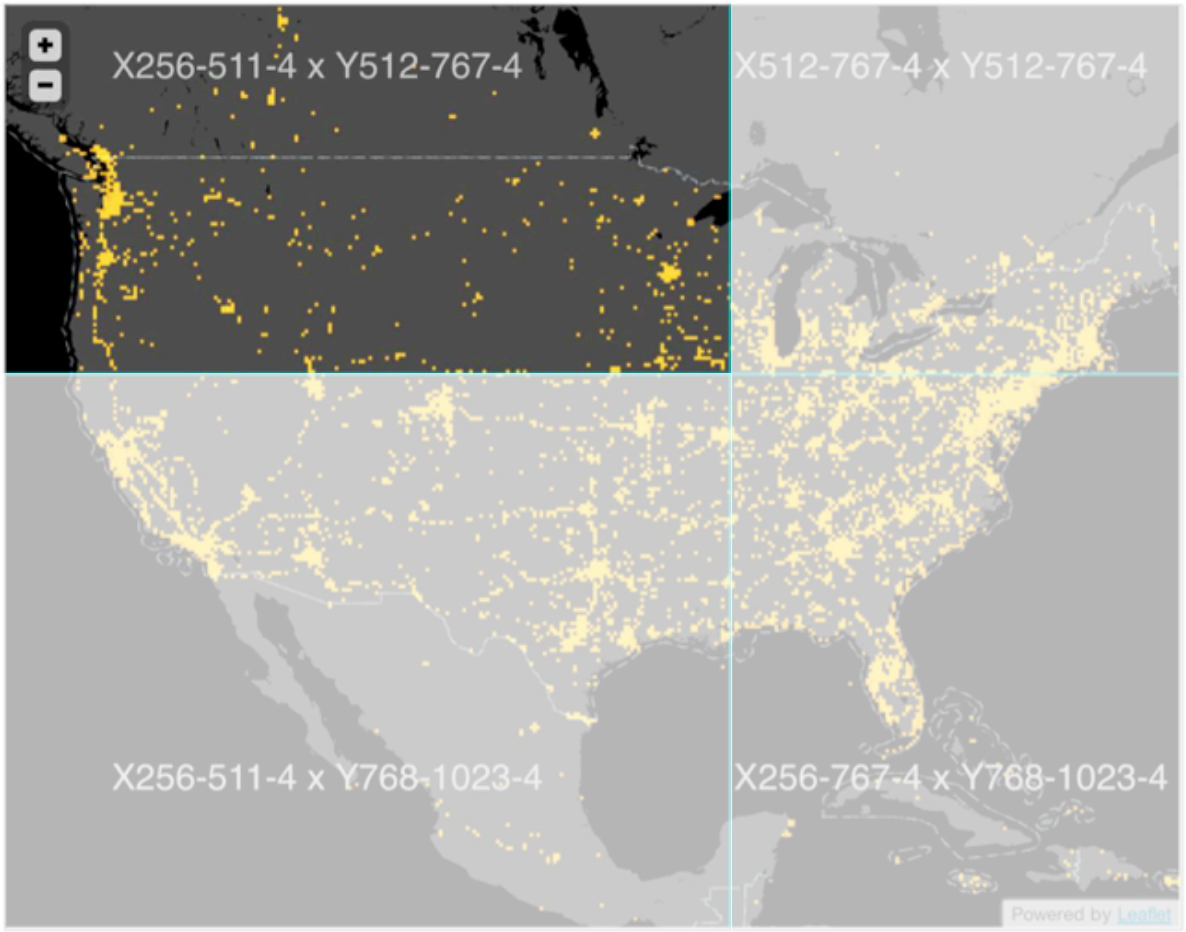

Full 5-D Cube
~2.3B bins
Σ
Σ
Σ
Σ
Month
Hour
Day
X
Hour
Y
X
Day
Y
X
Month
Y
3-D cubes
3-D data tiles
13 3-D Data Tiles
~17.6M bins
(in 352KB!)

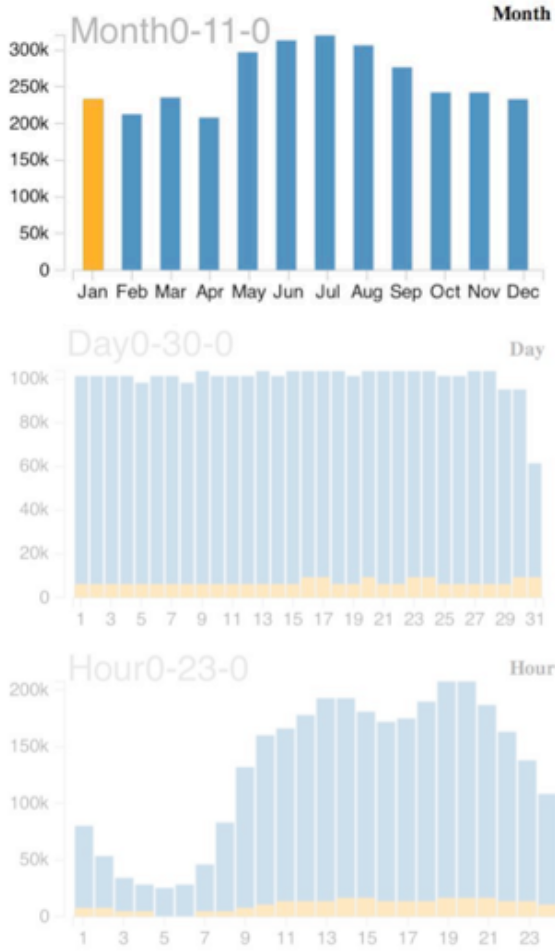

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

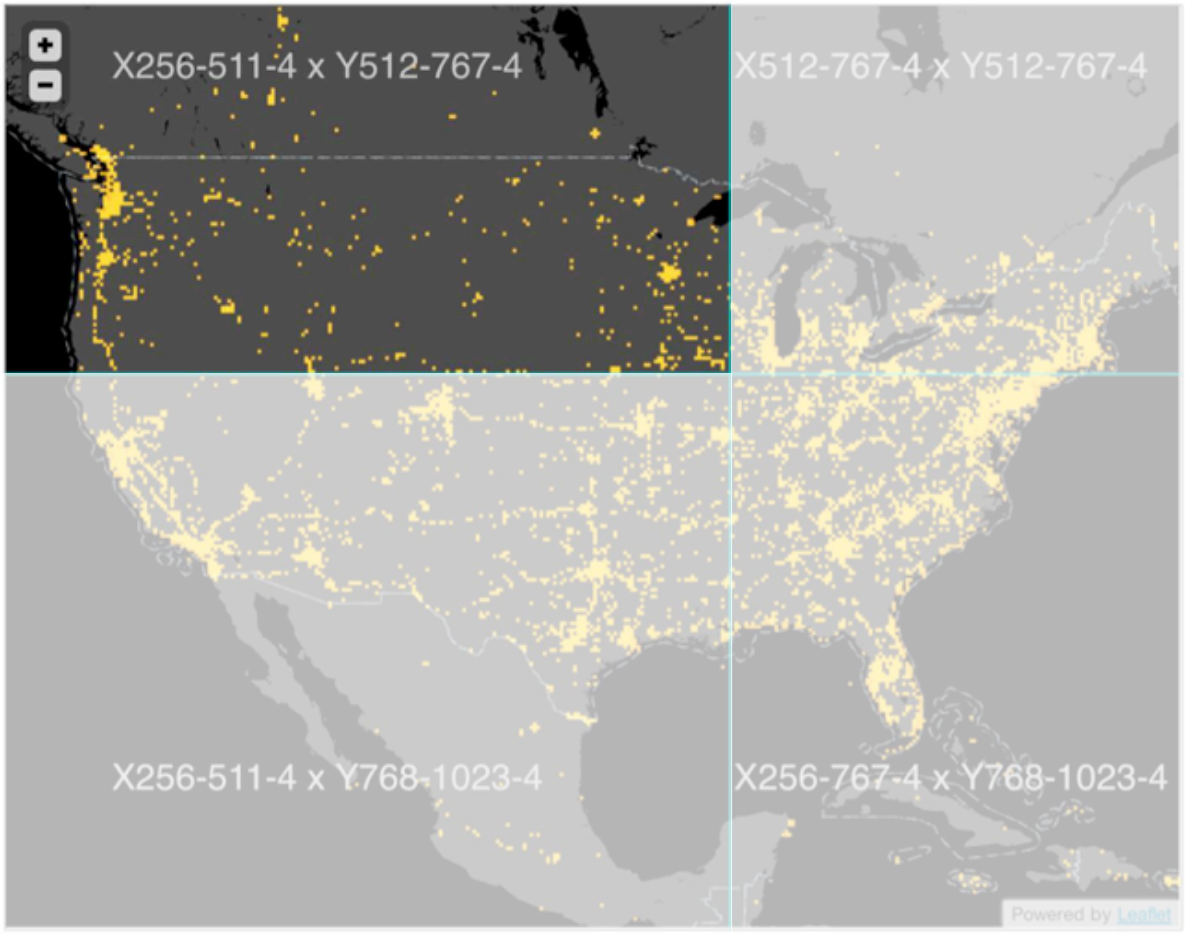

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
Day0-30-0
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour
Hour0-23-0
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

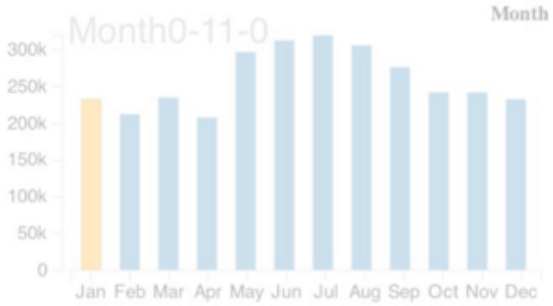

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

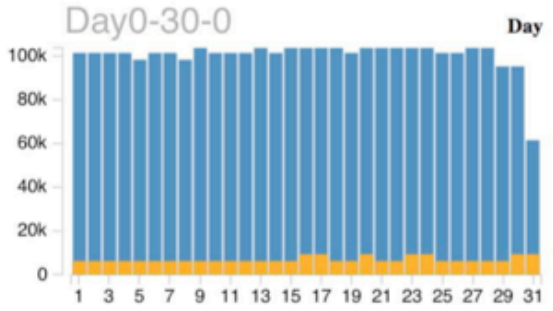

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
Day0-30-0
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour
Hour0-23-0
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

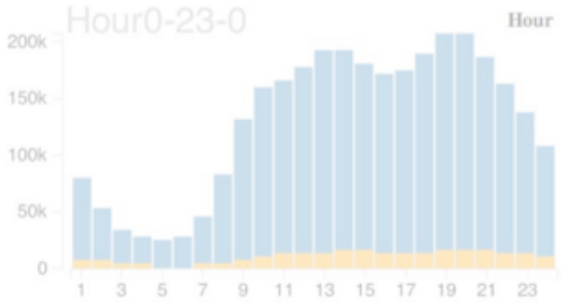

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

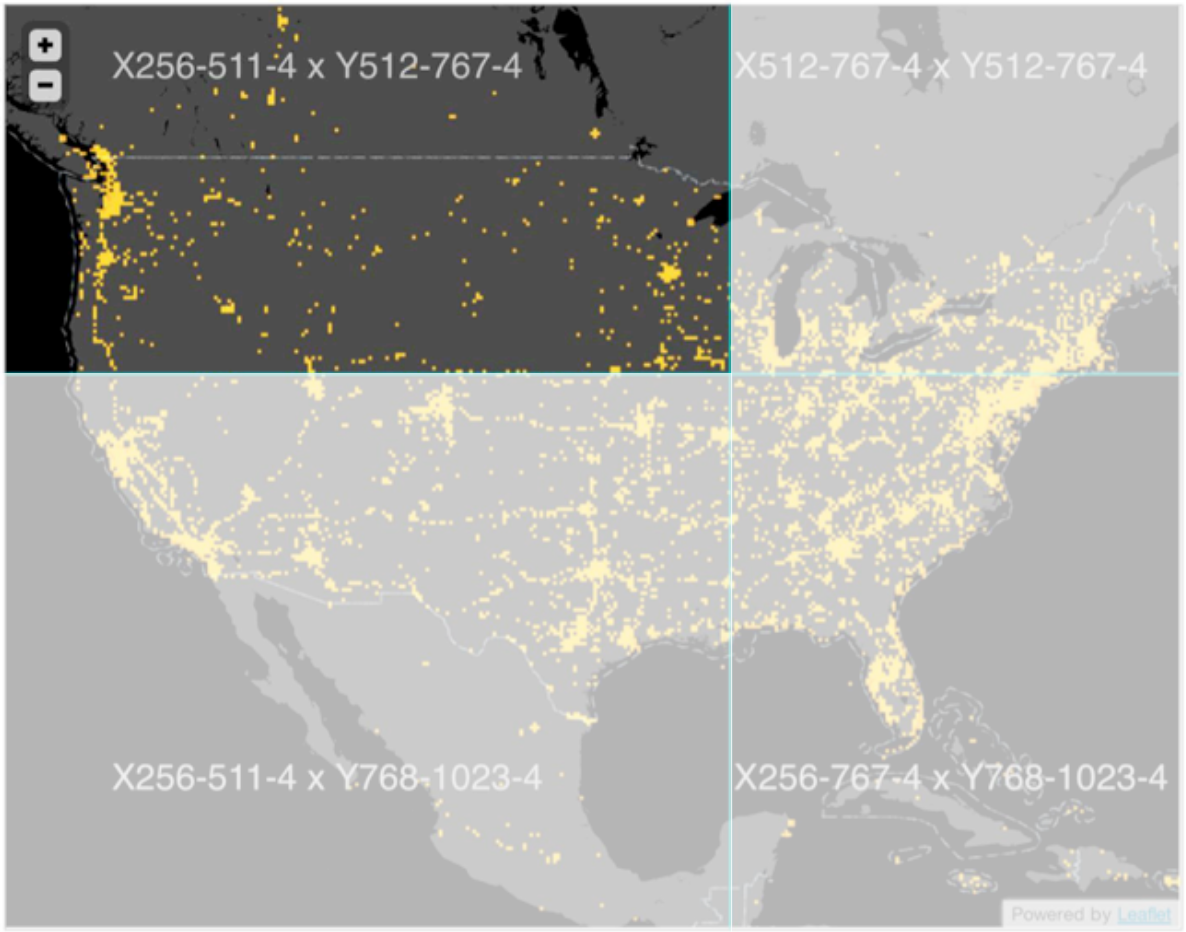

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23


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

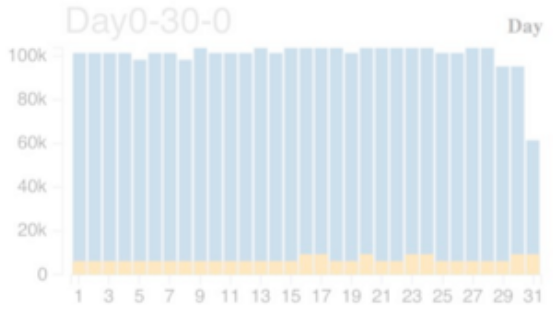

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

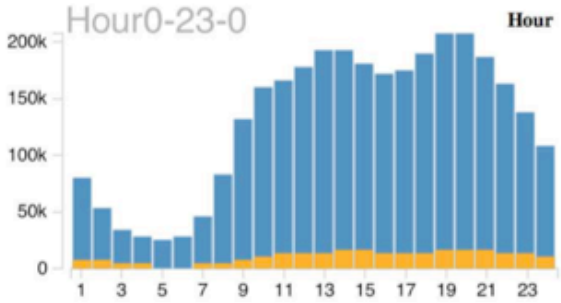

Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31

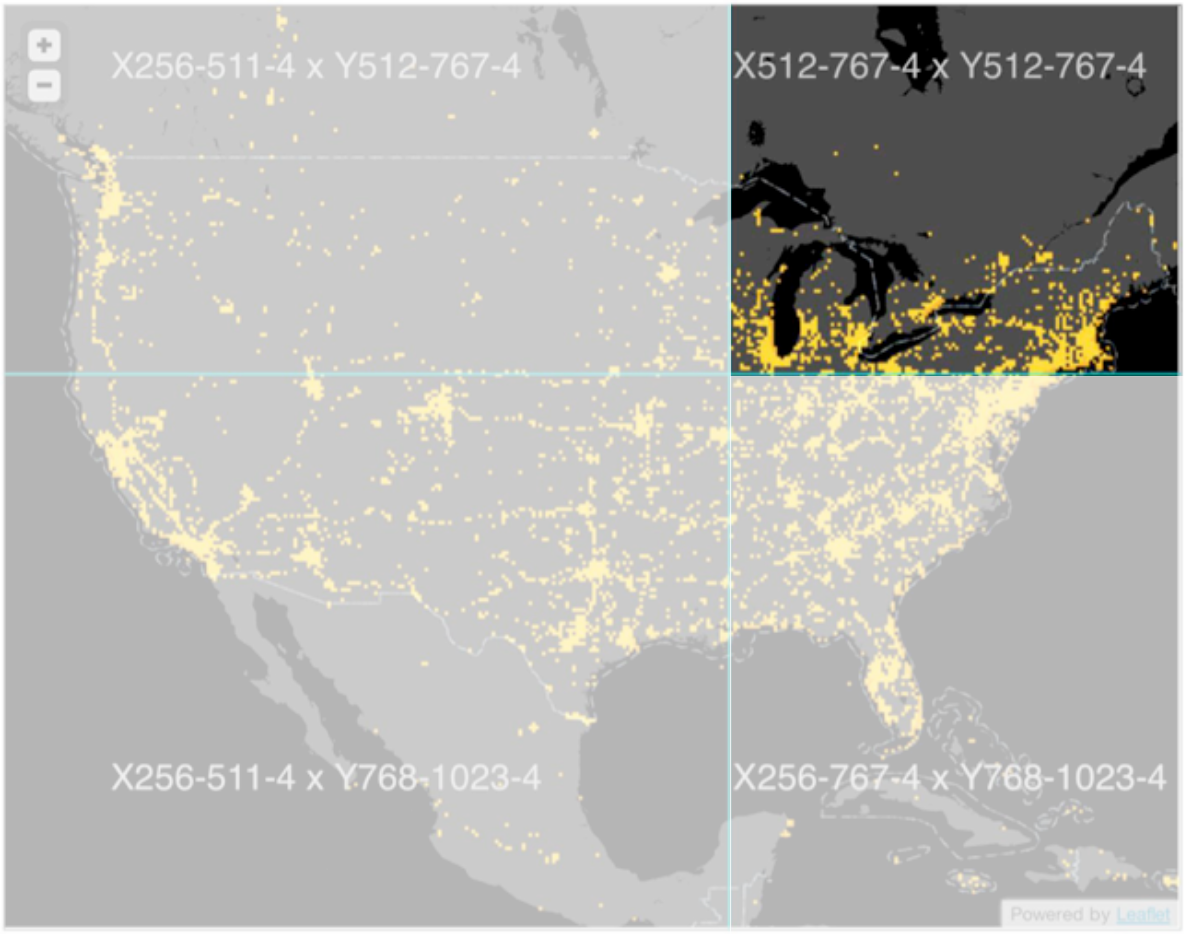

Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

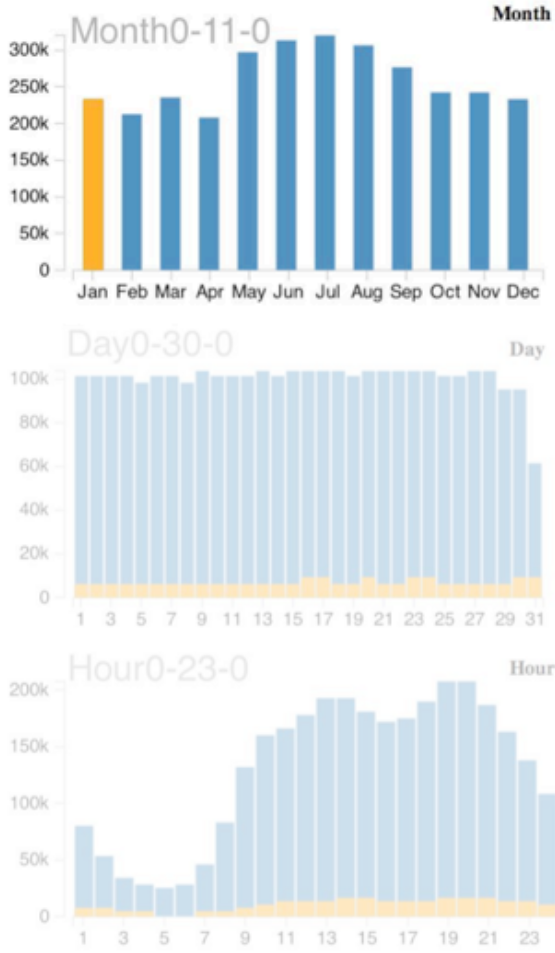

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

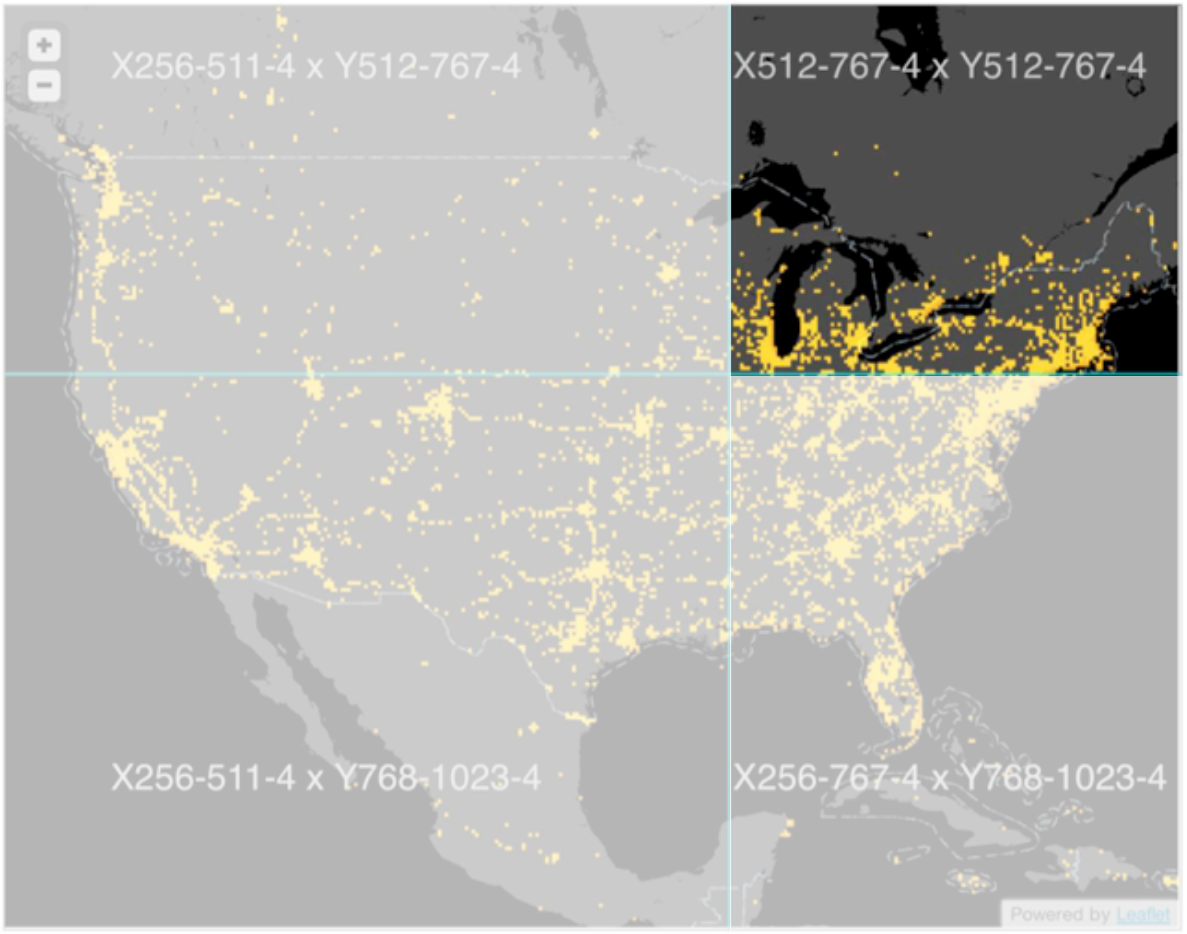

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

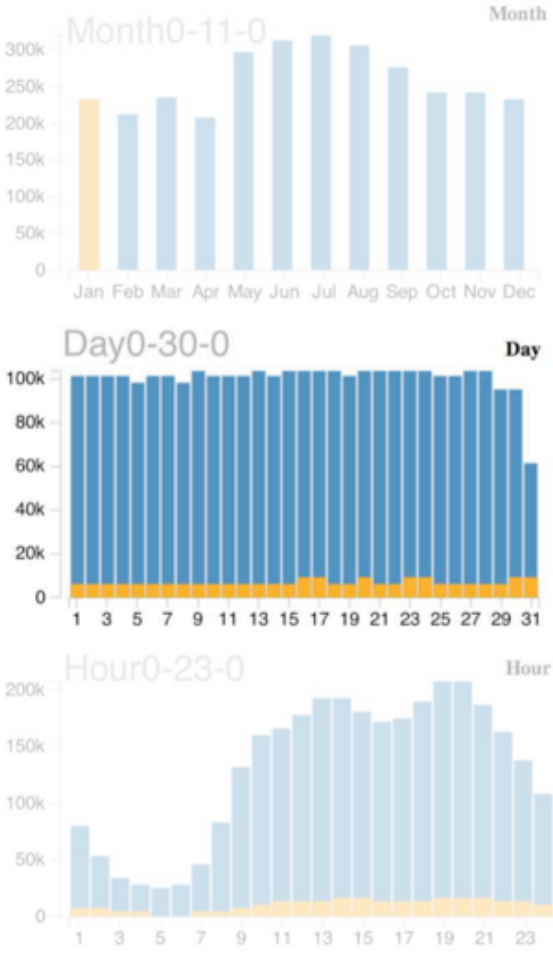

Day
Day0-30-0
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31

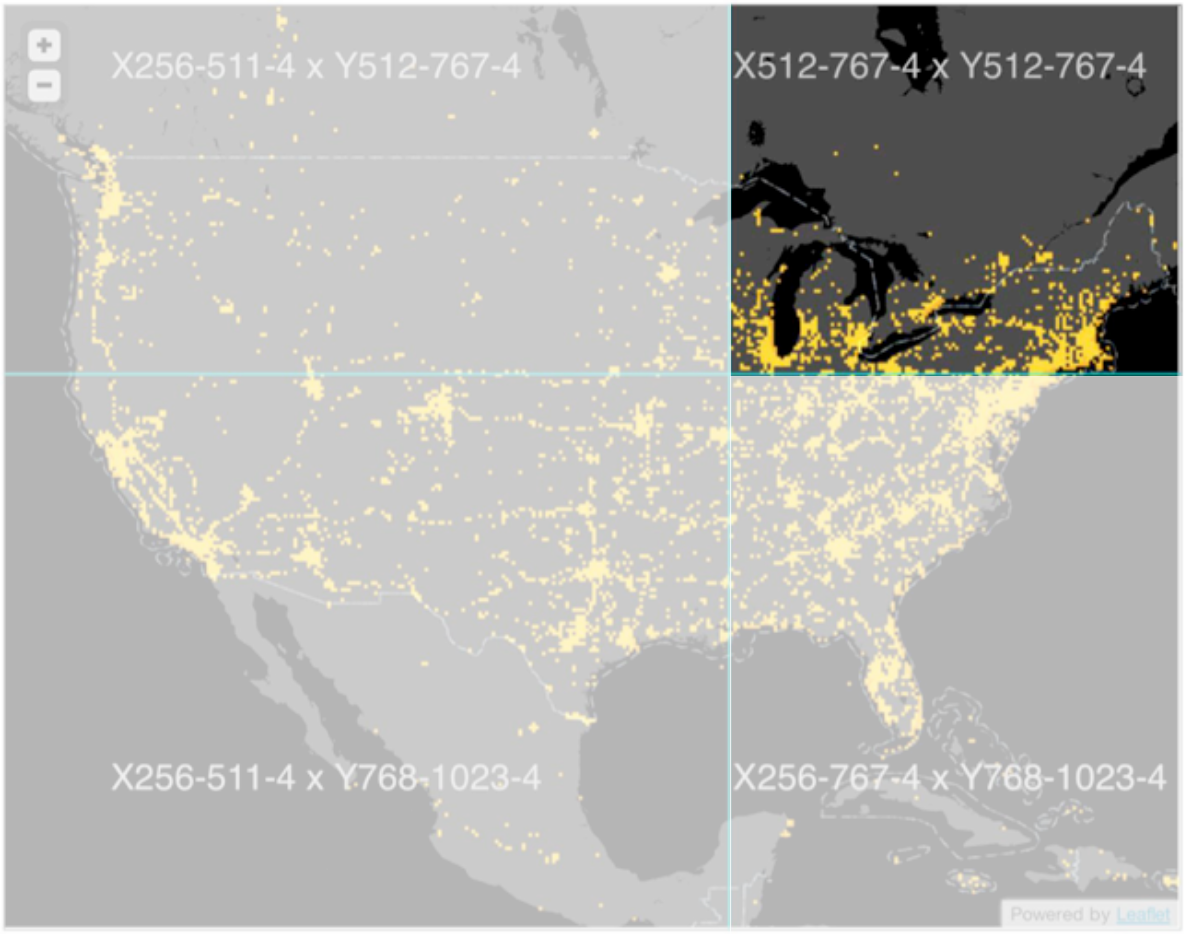

Hour
Hour0-23-0
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23


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

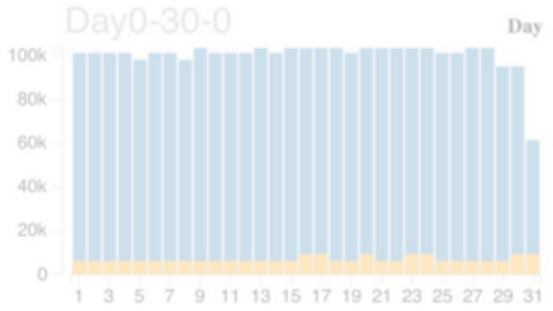

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
Day0-30-0
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour
Hour0-23-0
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

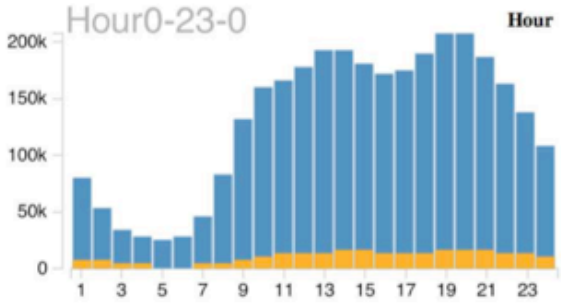

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

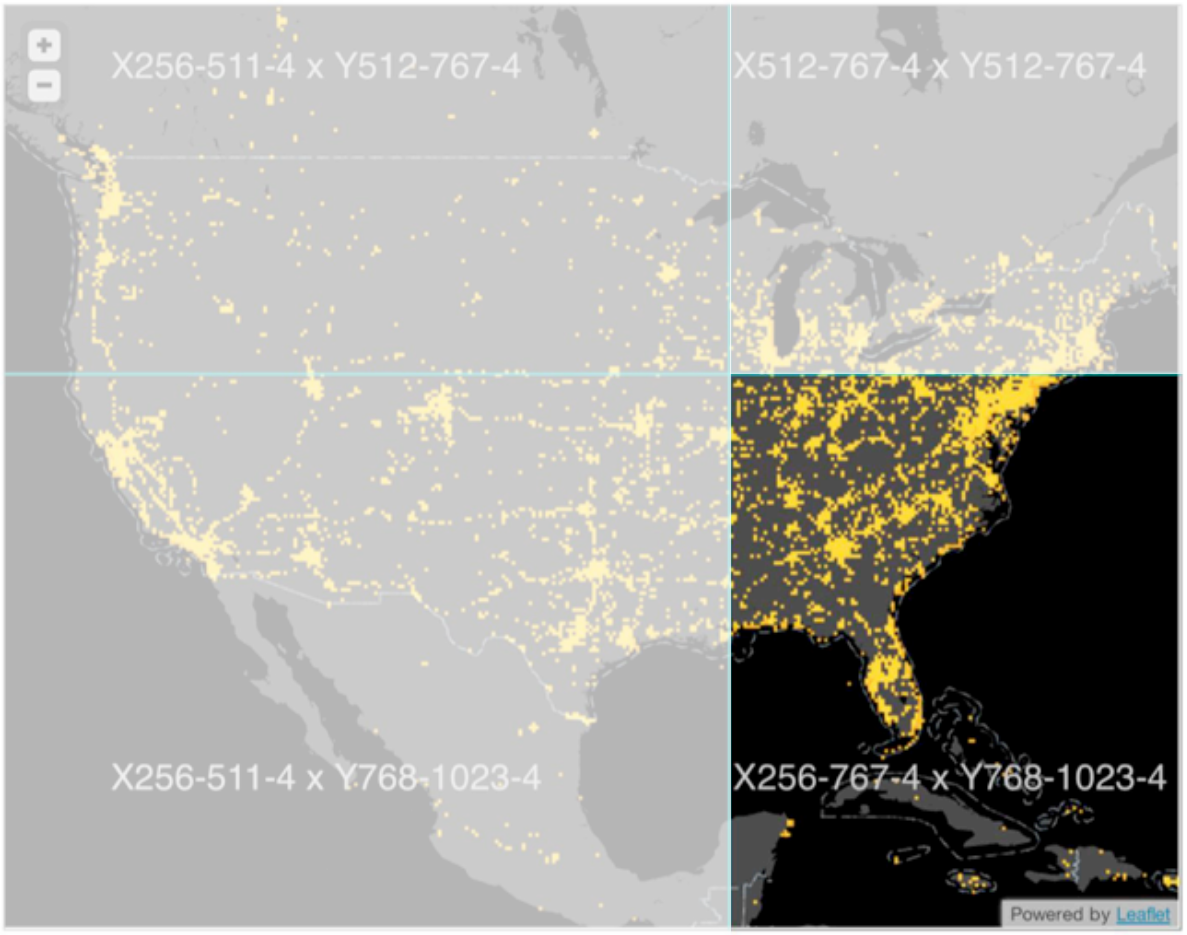

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

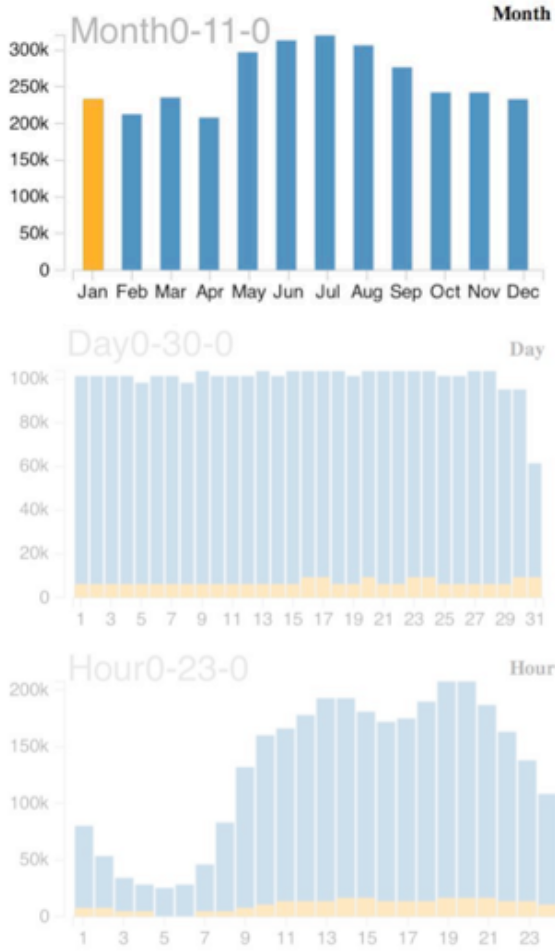

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

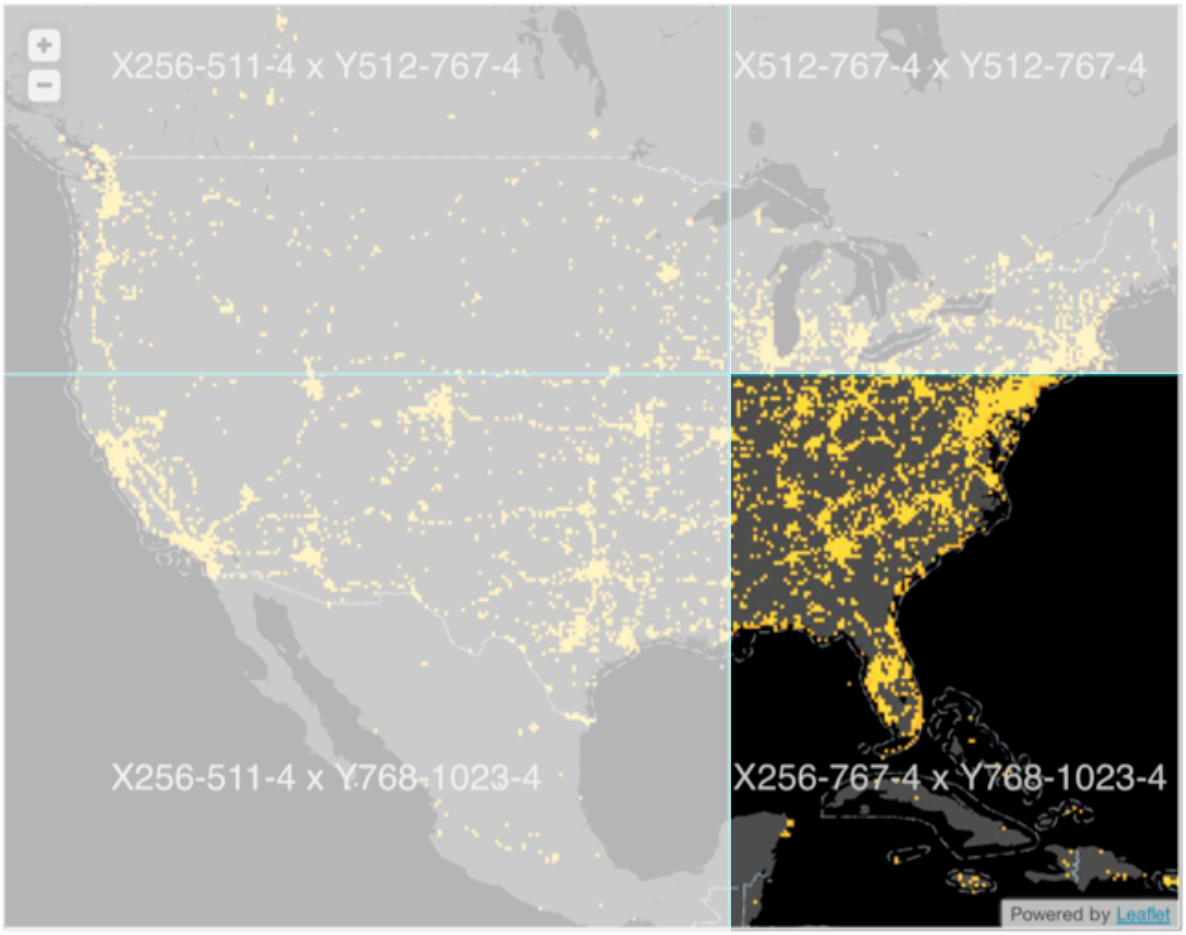

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

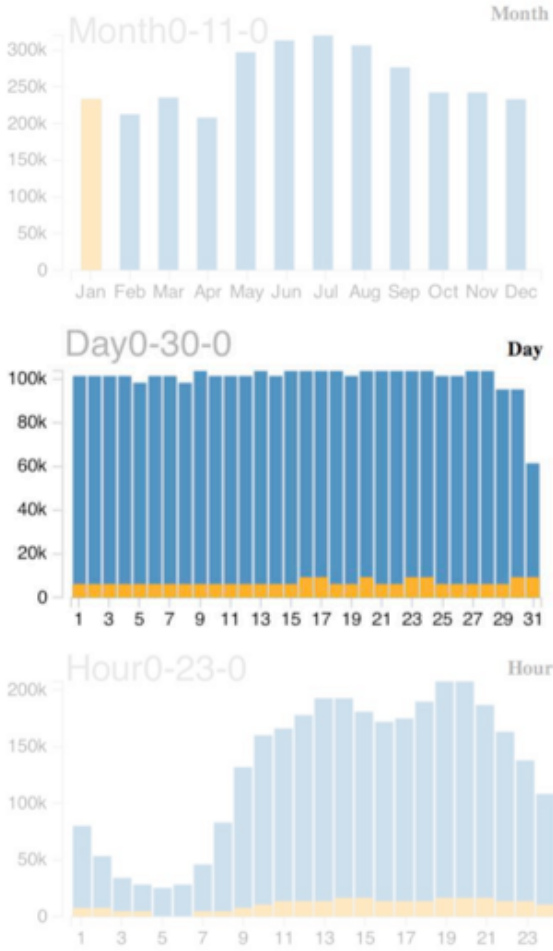

Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31

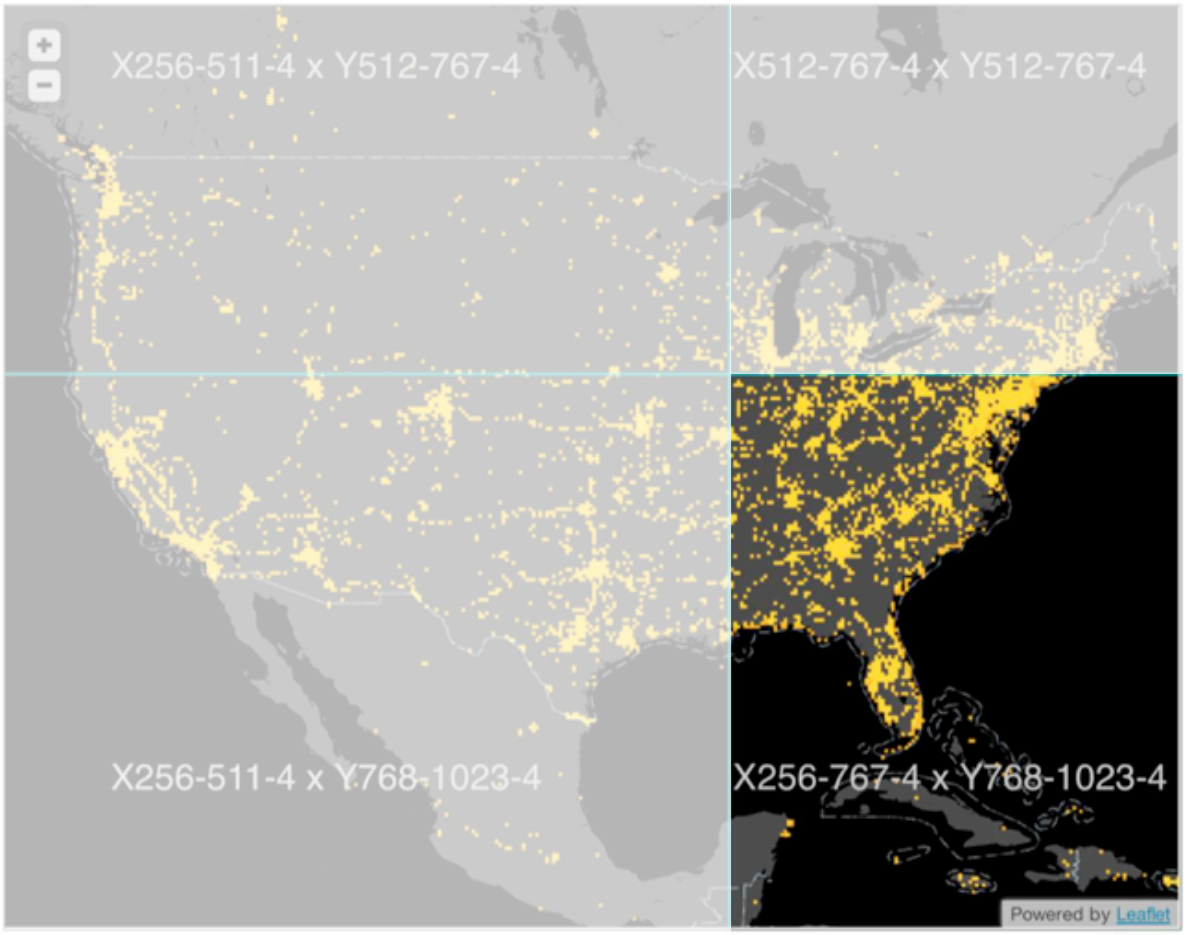

Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

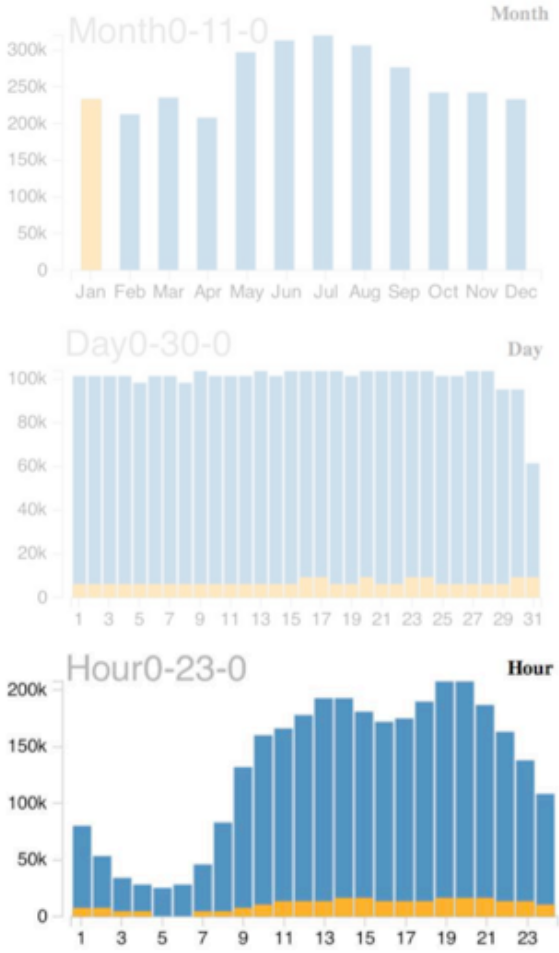

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

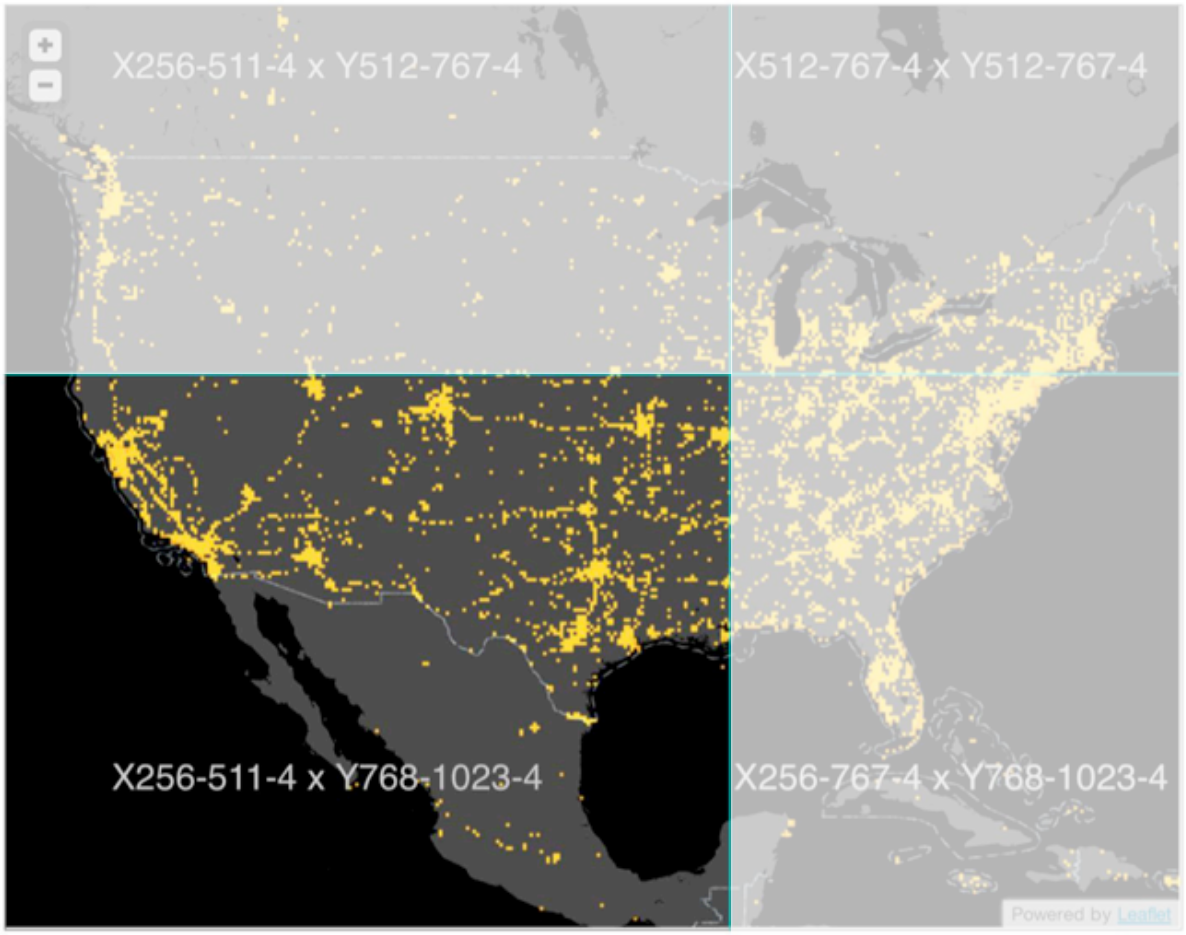

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

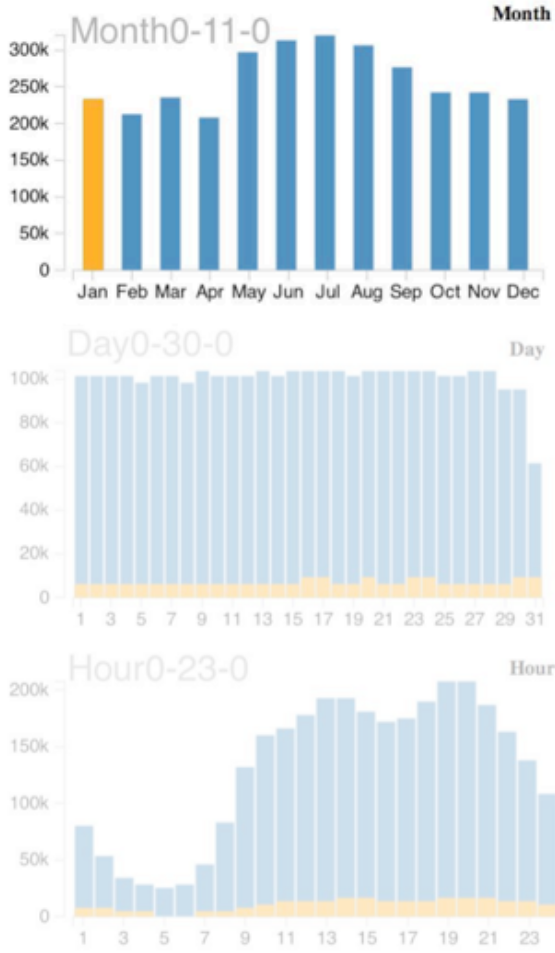

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

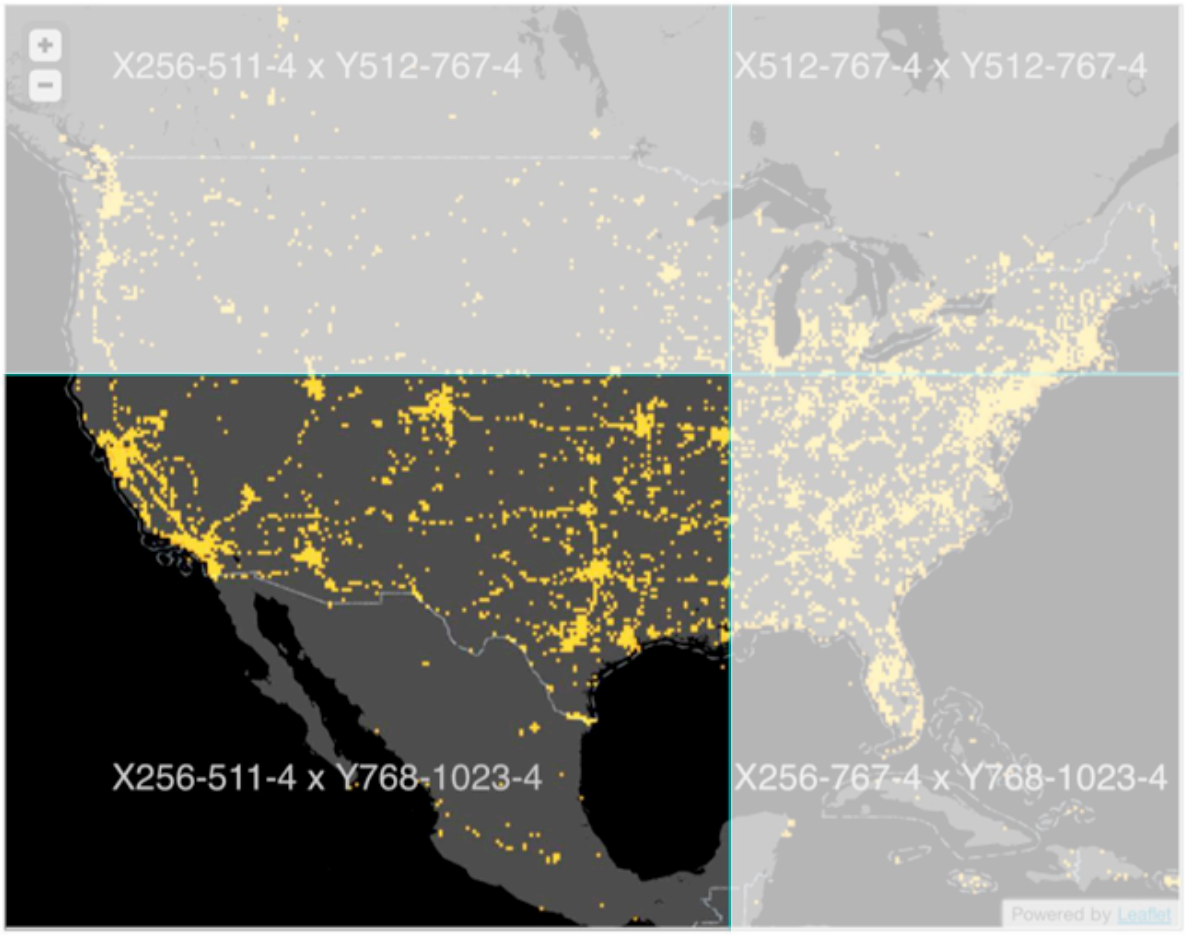

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

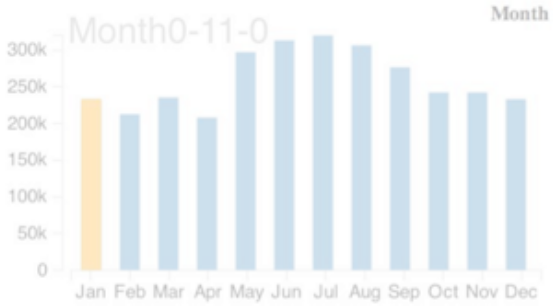

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

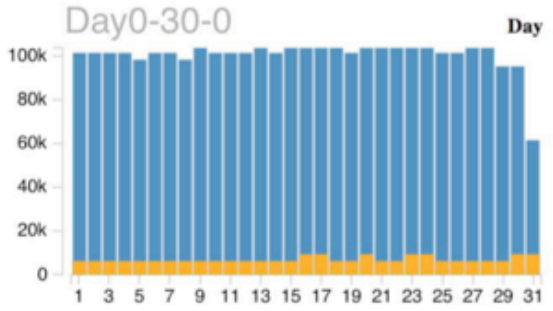

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
Day0-30-0
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour
Hour0-23-0
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

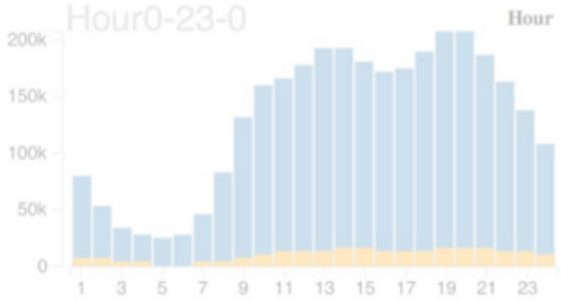

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

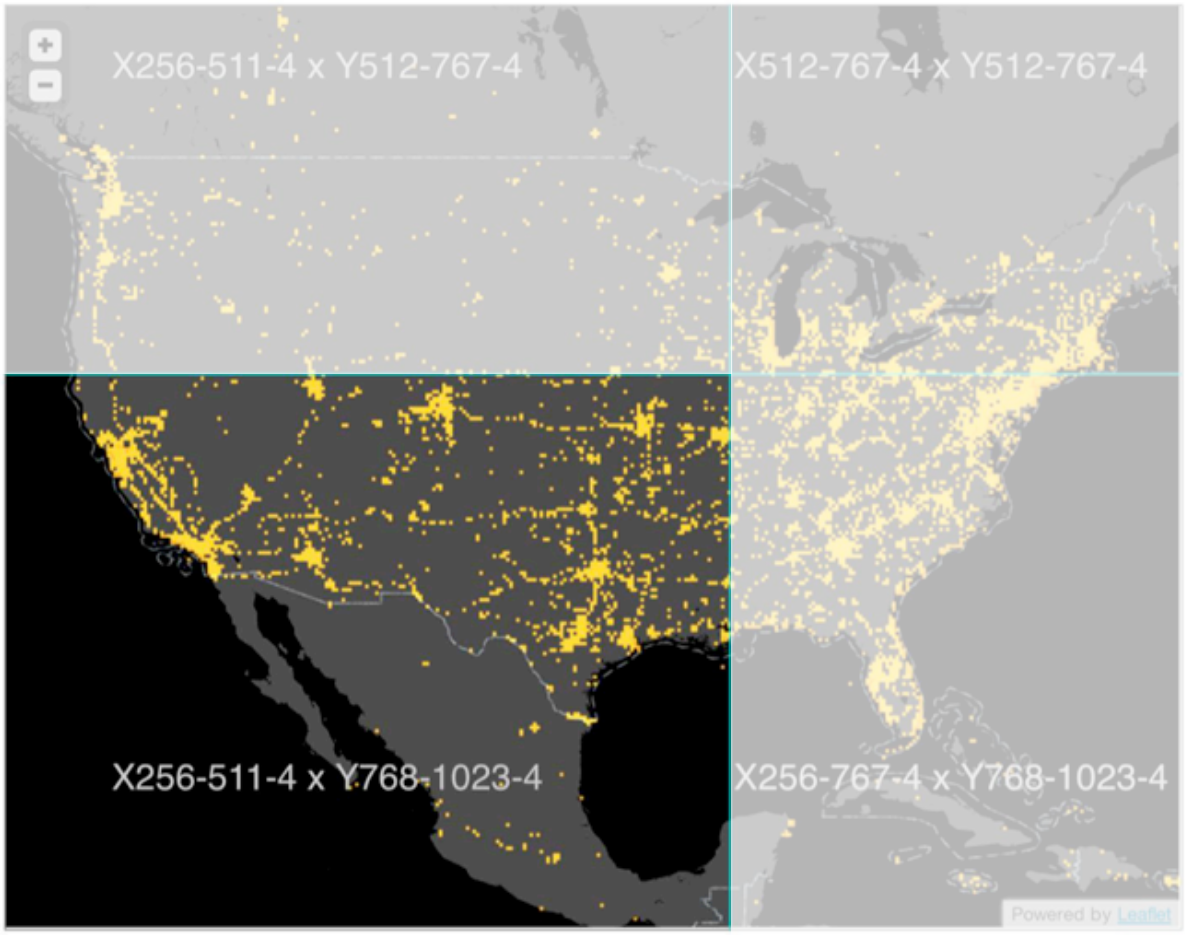

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
Day0-30-0
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour
Hour0-23-0
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23


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

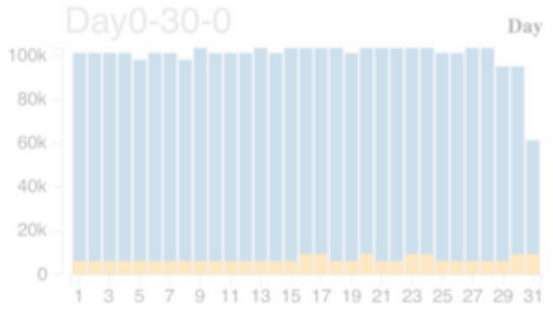

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

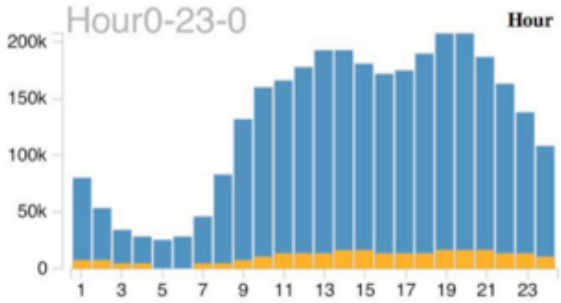

Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31

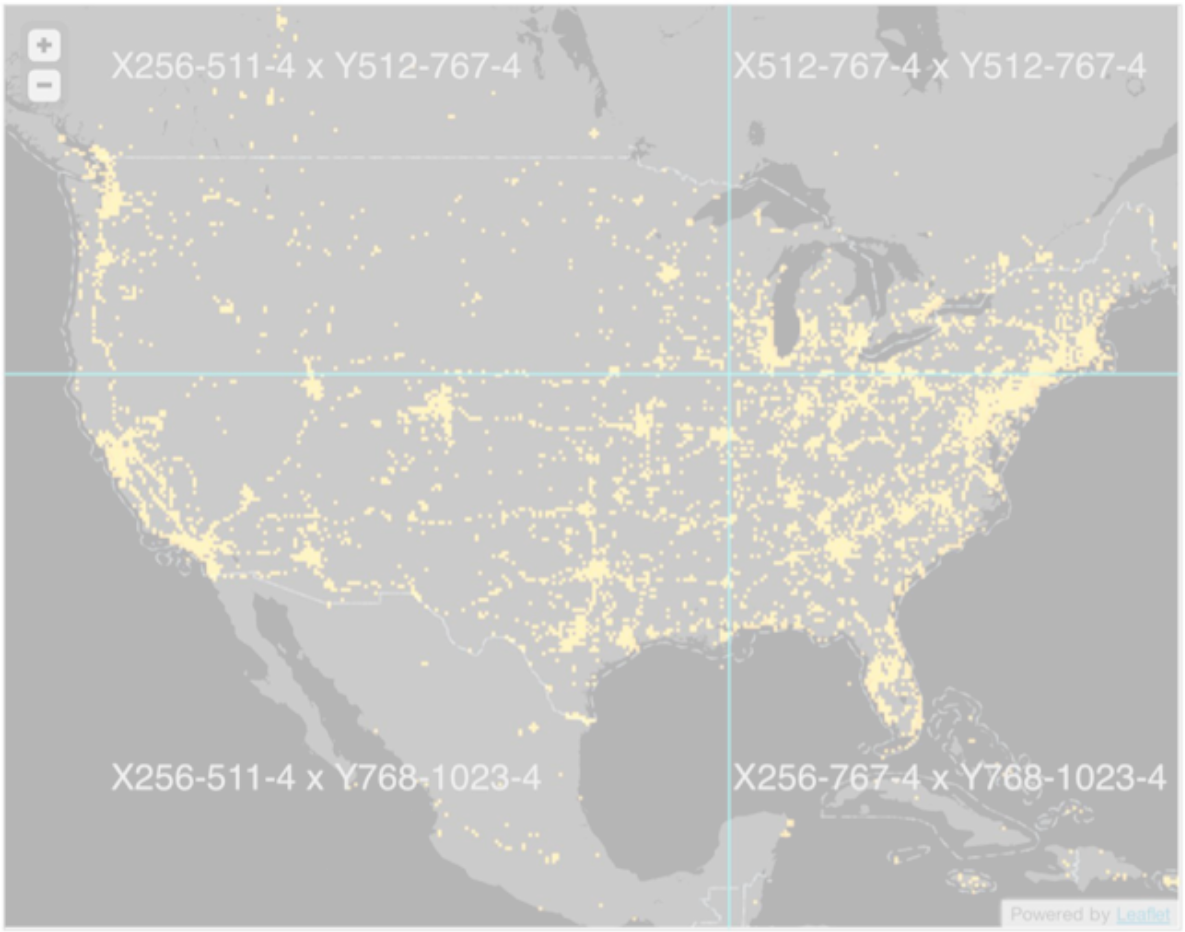

Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

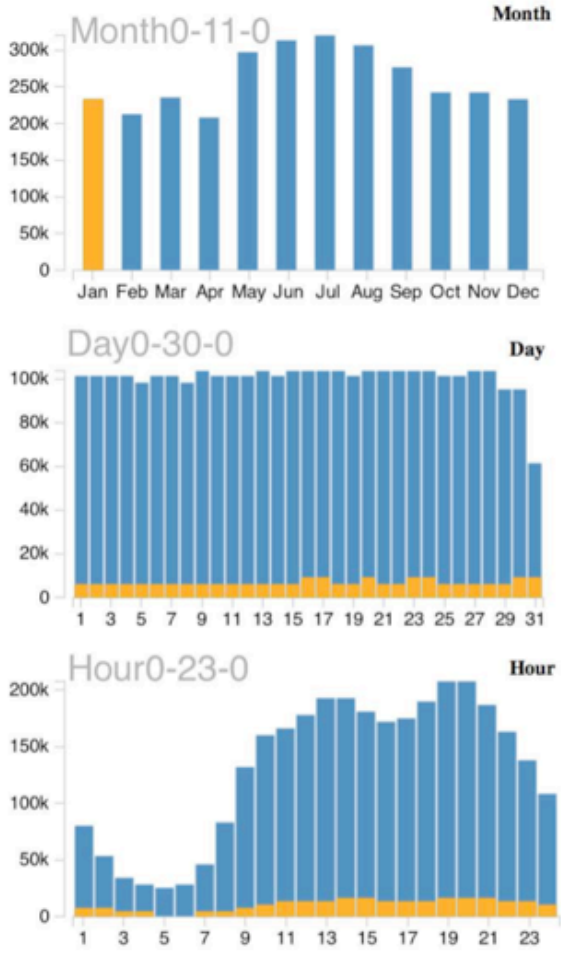

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet

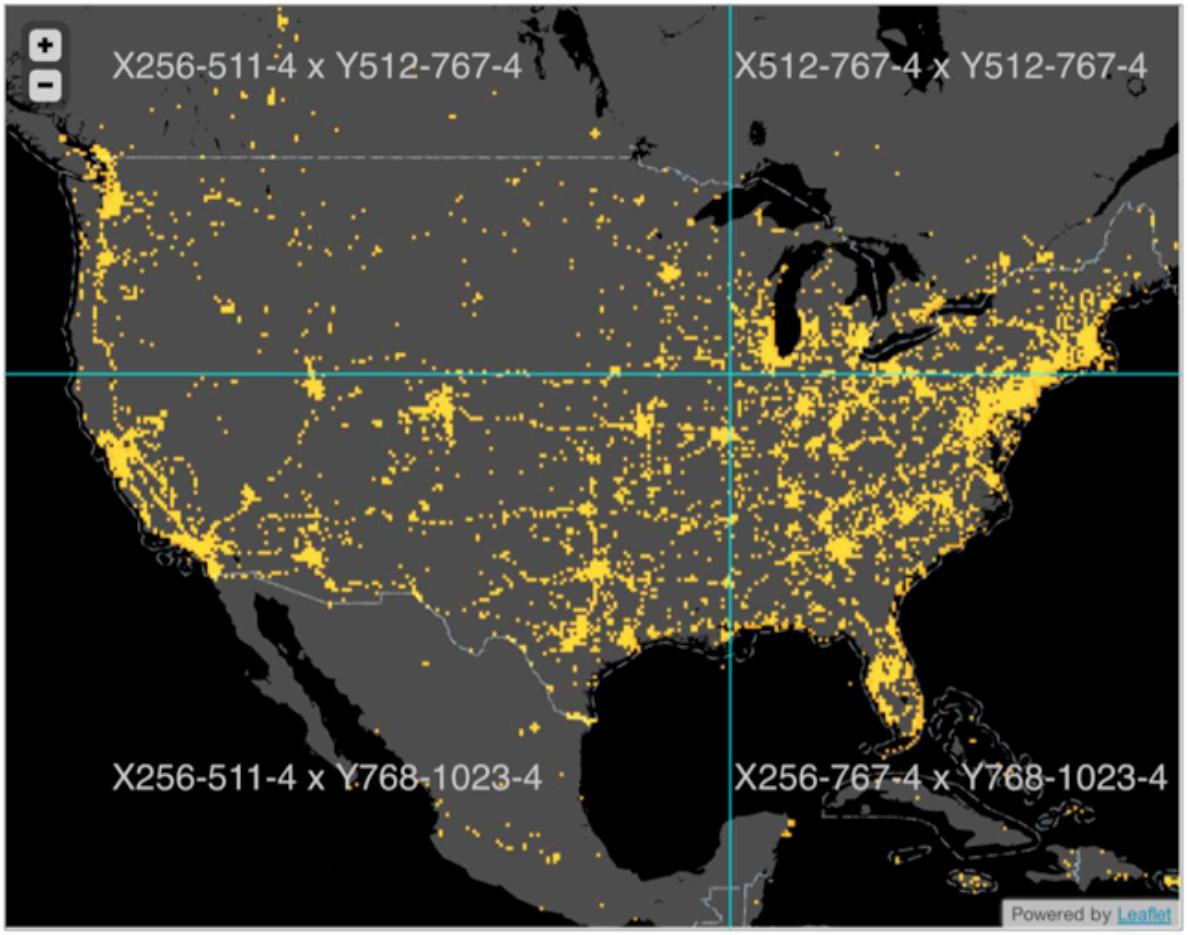

Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

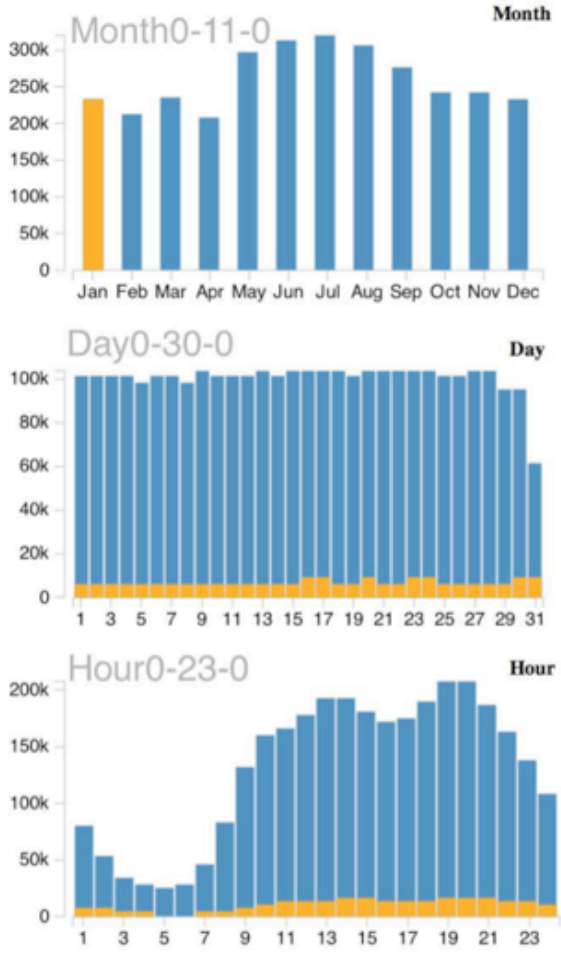

Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31


Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

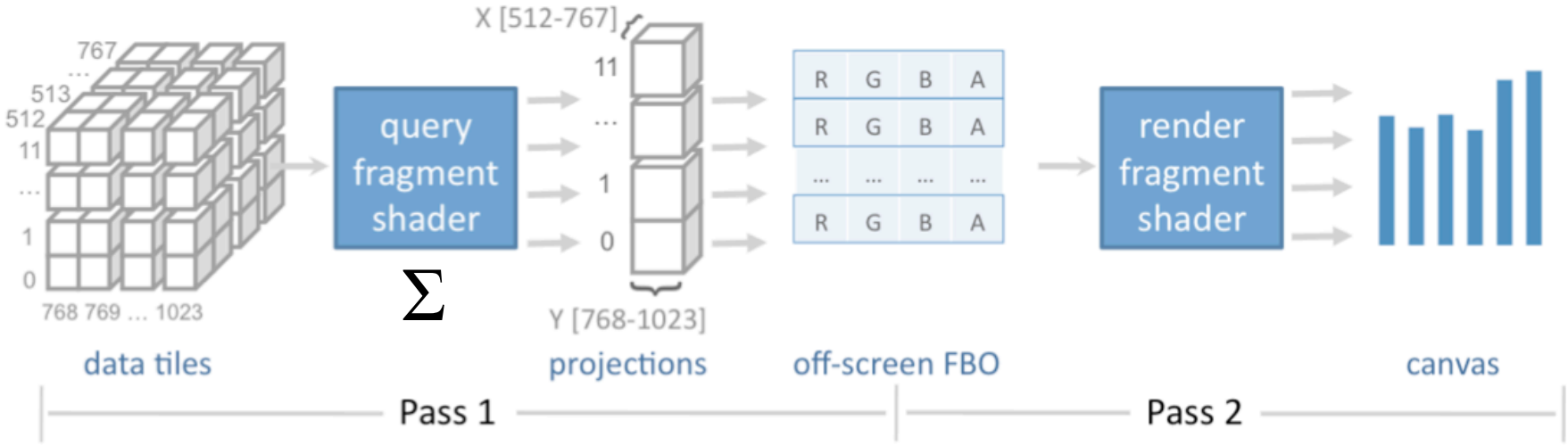

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet


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

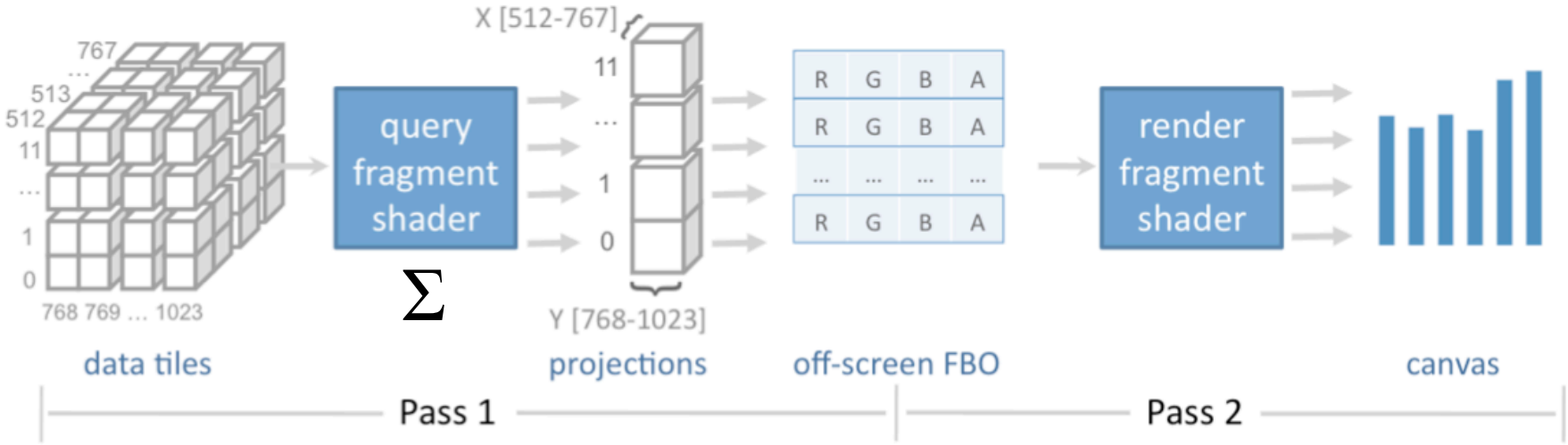

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet


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

Query & Render on GPU via WebGL

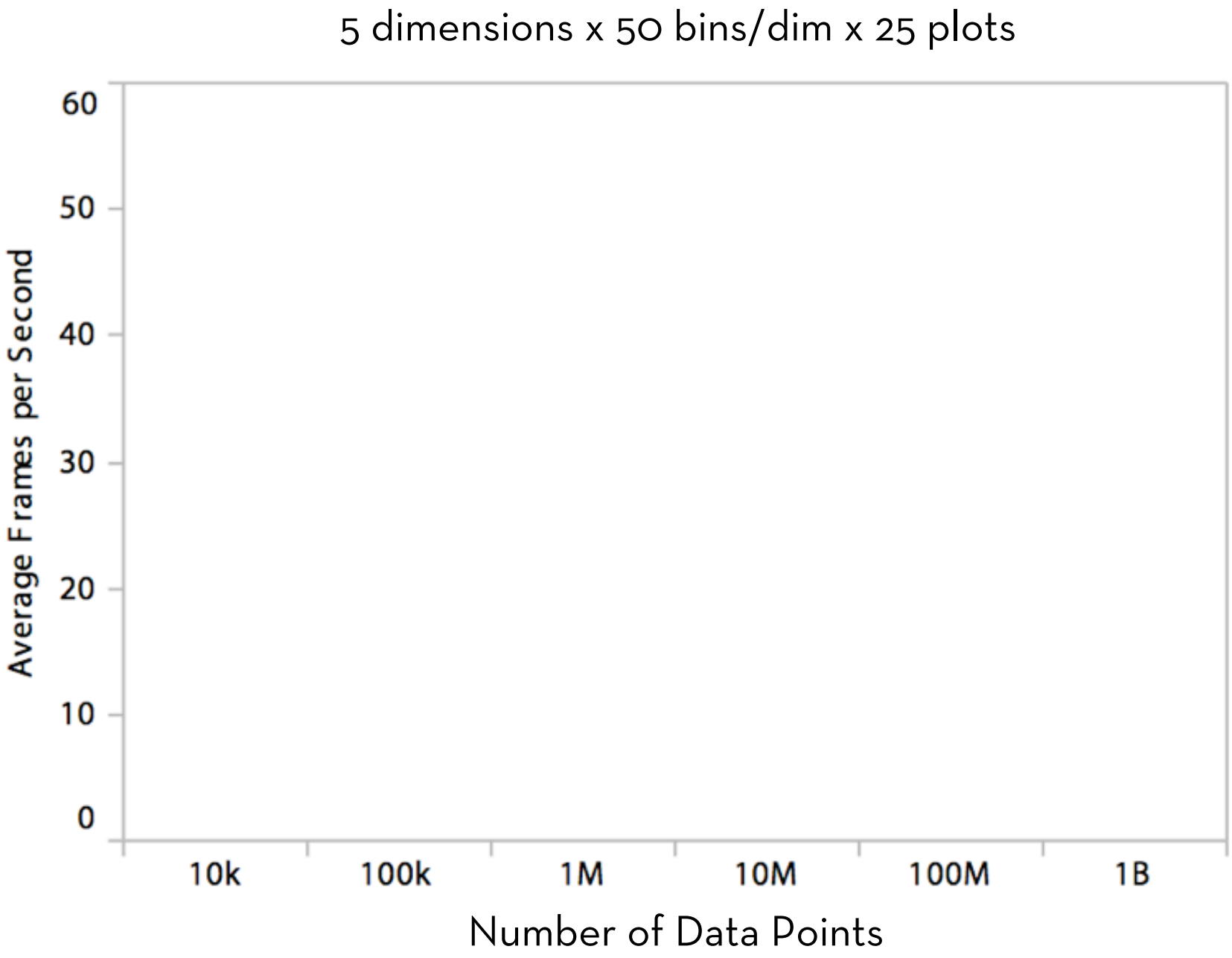


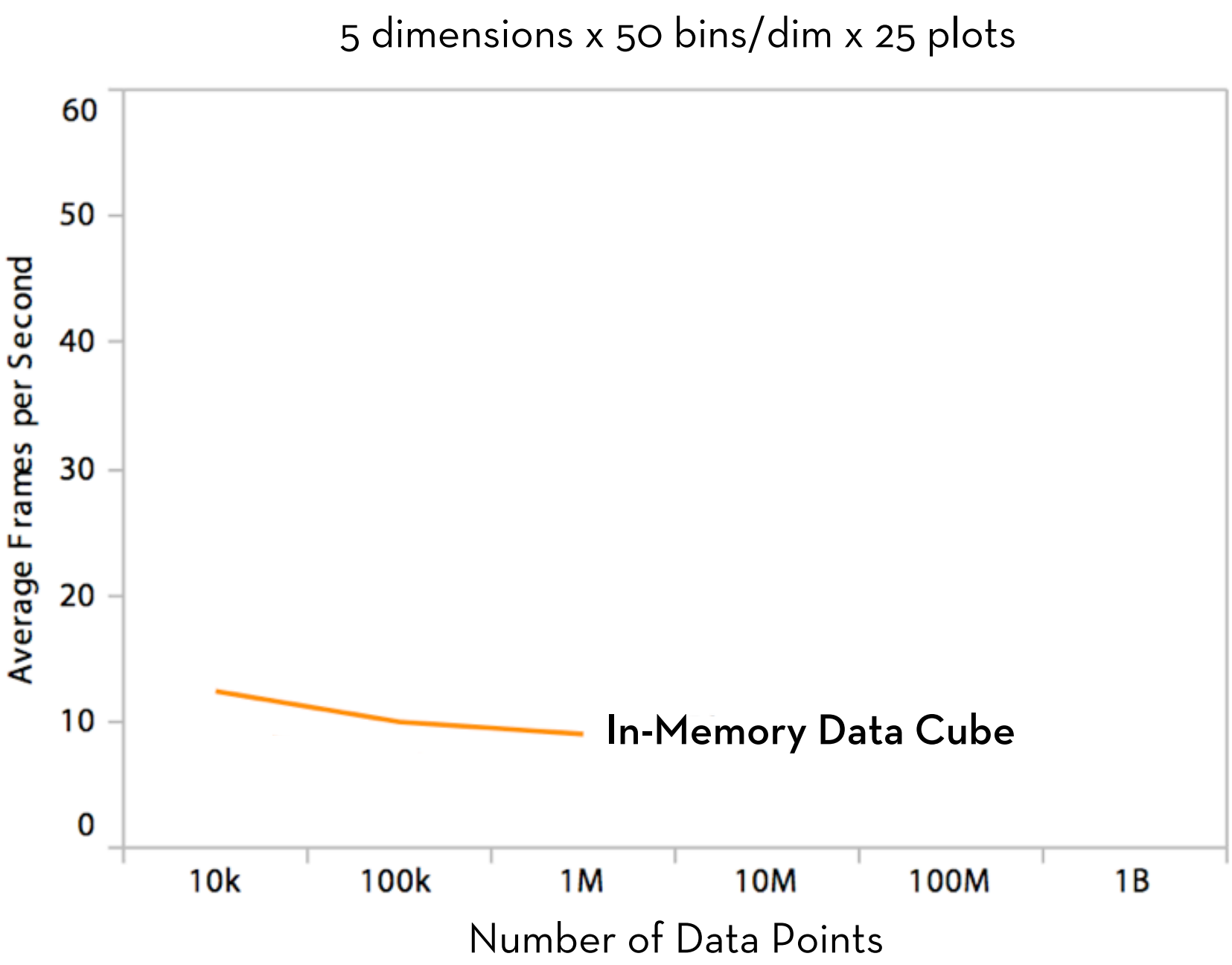
Query & Render on GPU via WebGL

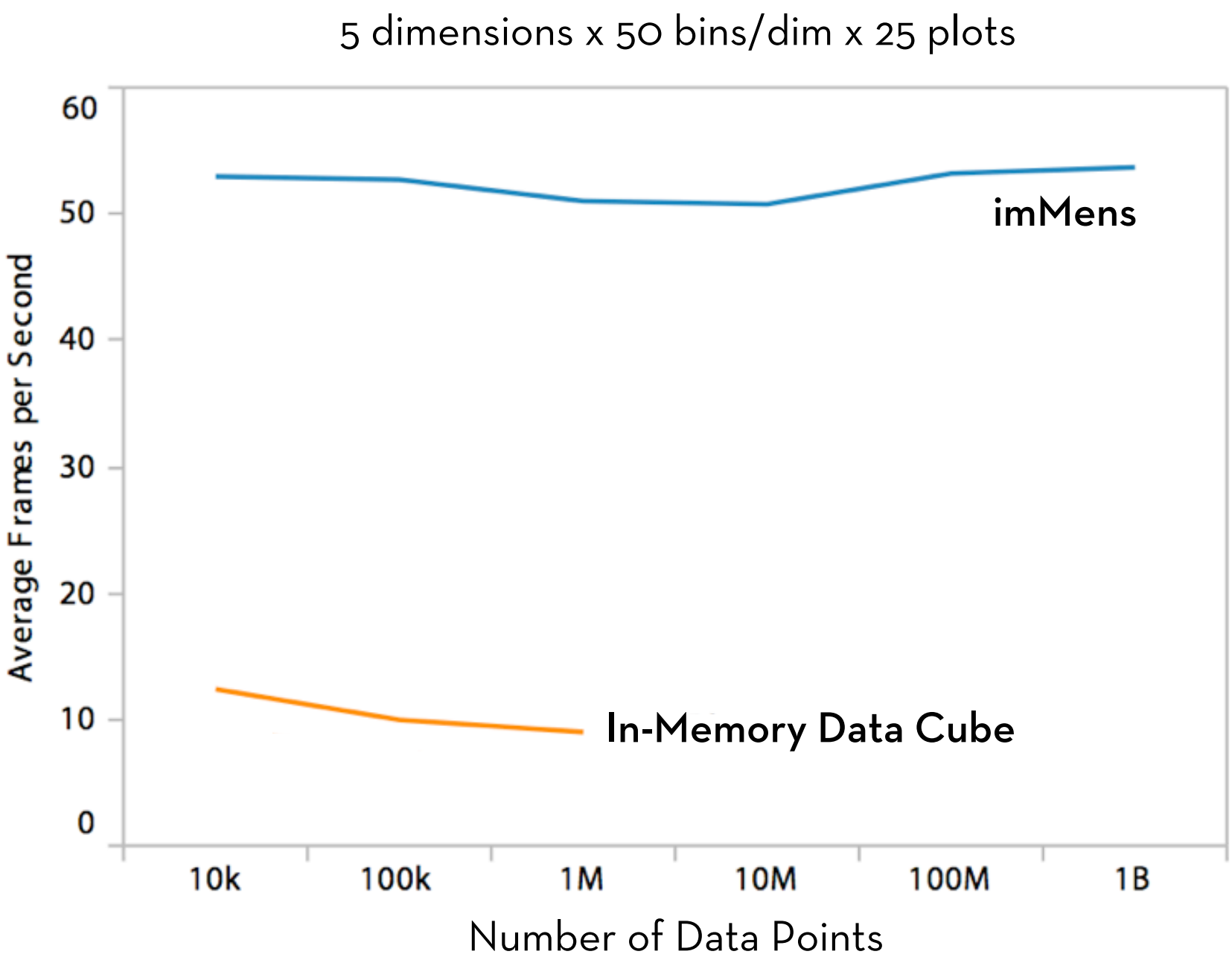


Bind data tiles as image textures.

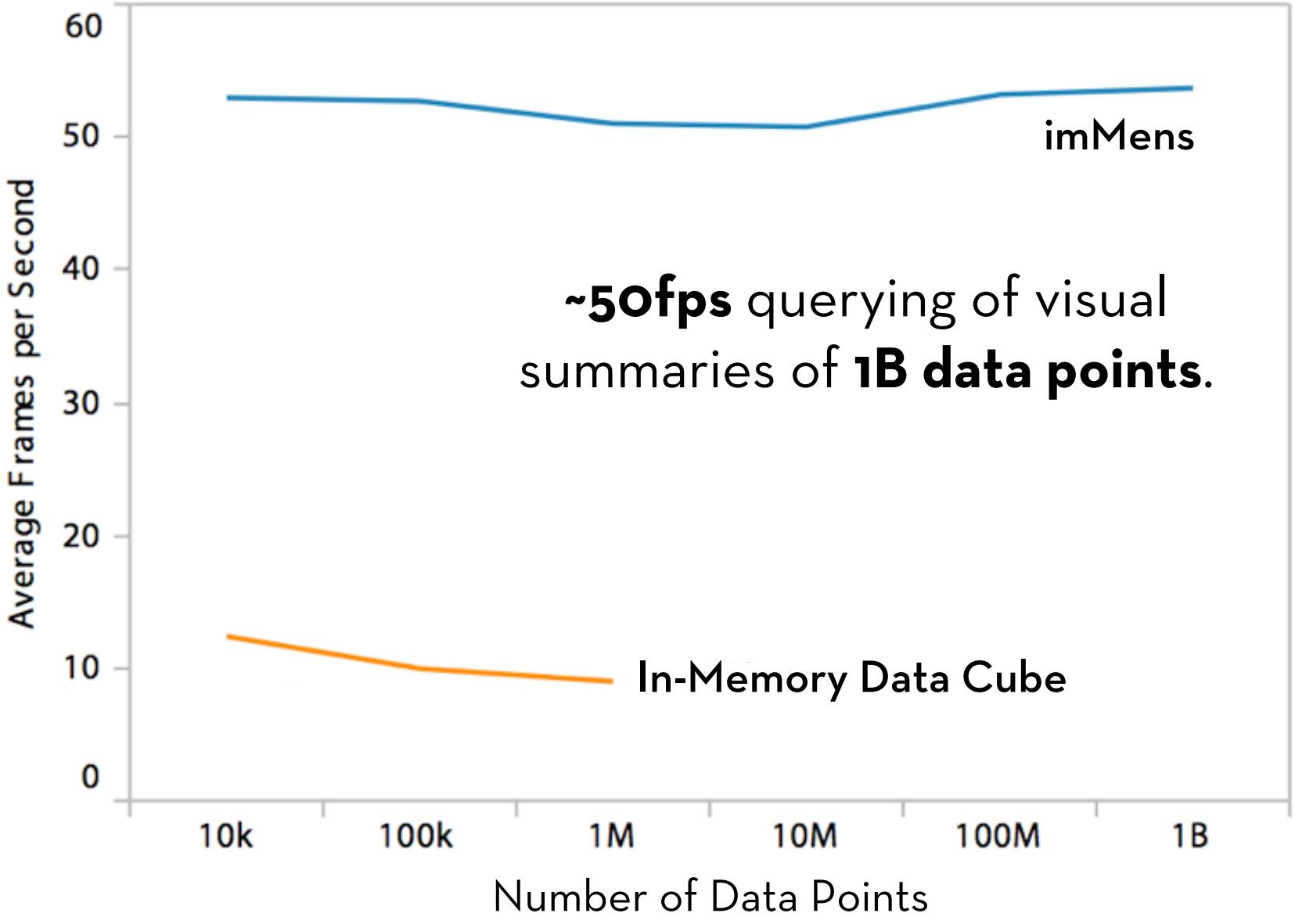
Compute aggregates in parallel on GPU.

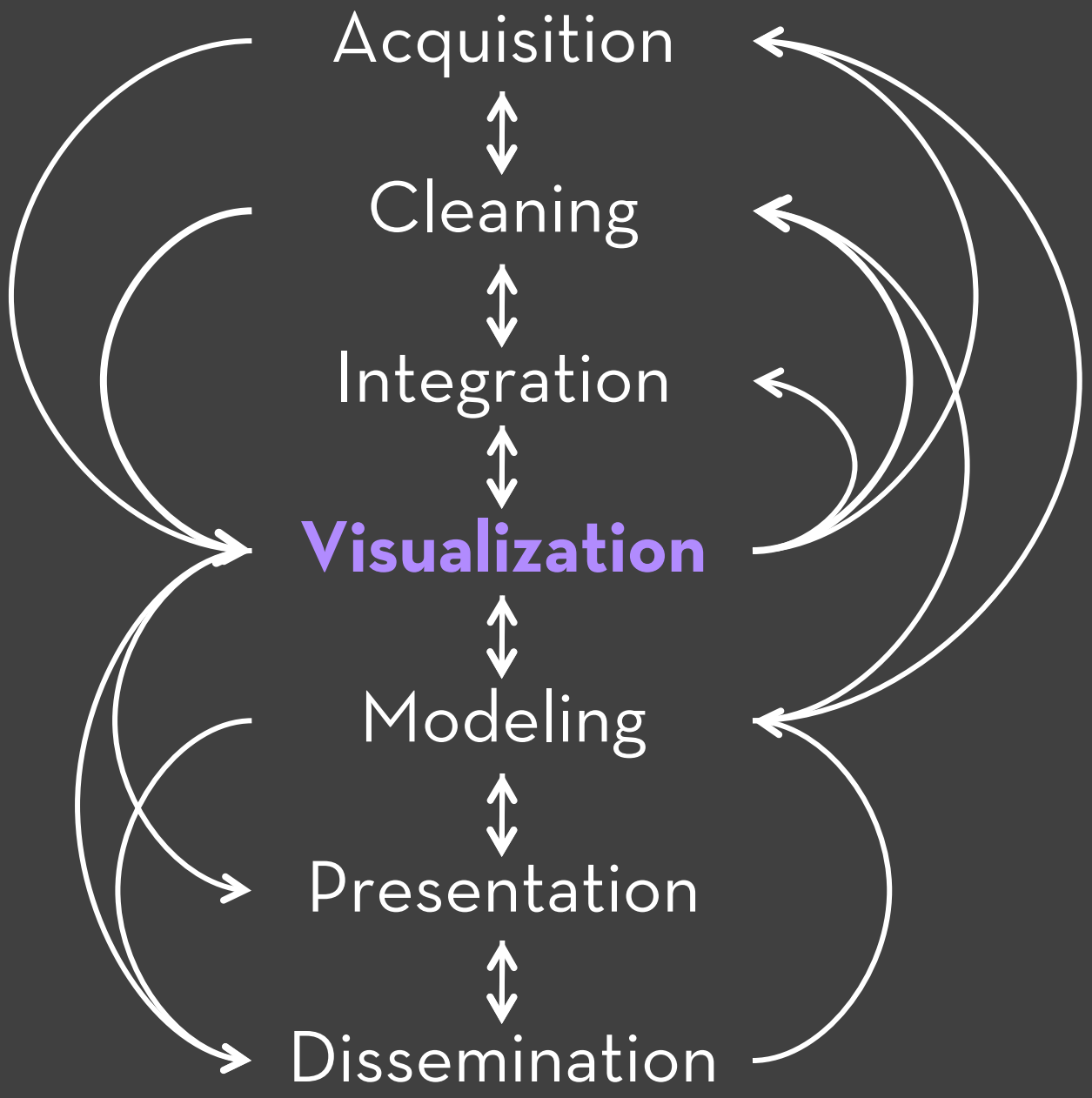

5 dimensions x 50 bins/dim x 25 plots
Average Frames per Second
60
50
40
30
20
10
0
10k
100k
1M
10M
100M
1B
Number of Data Points

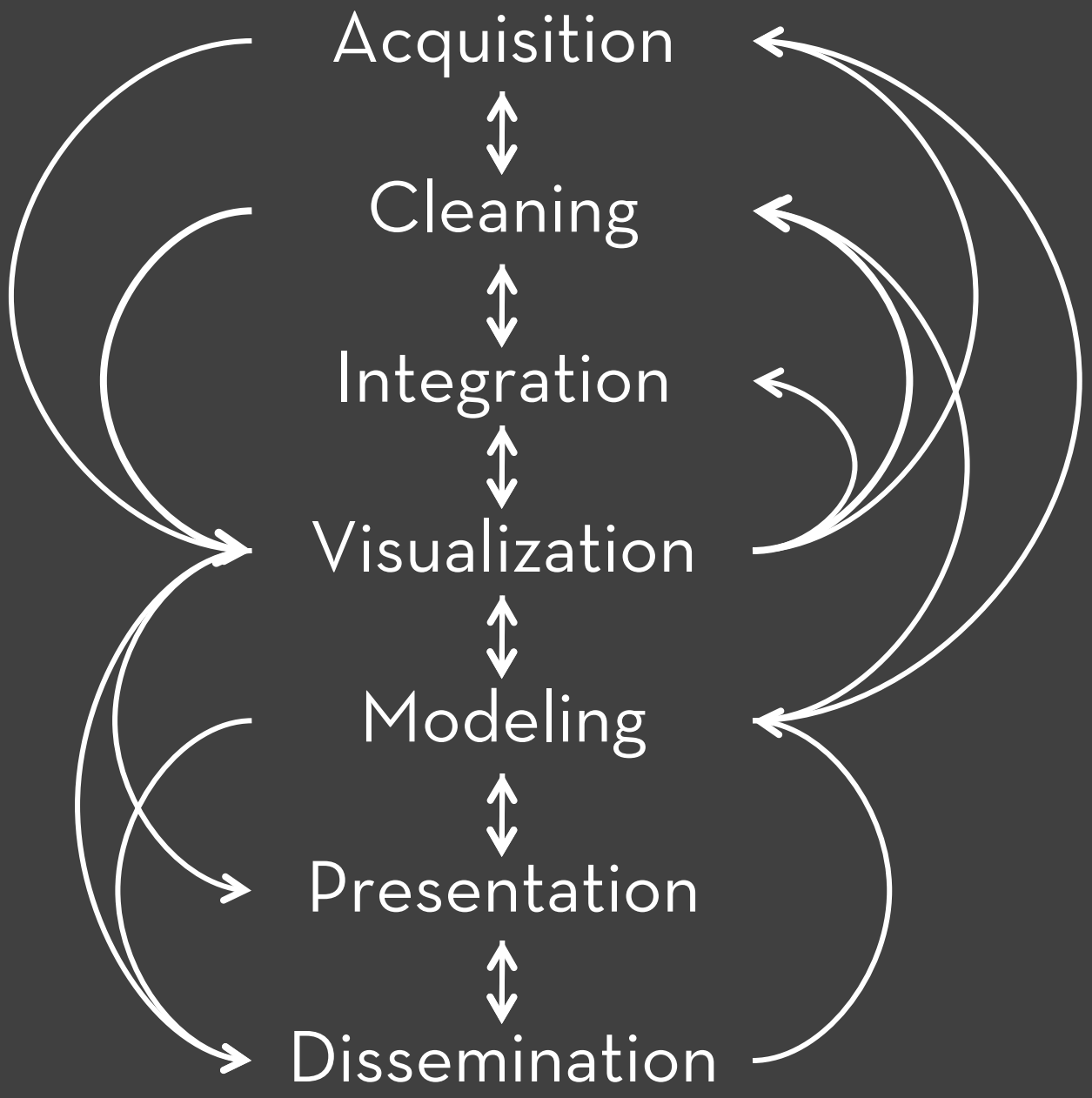

5 dimensions x 50 bins/dim x 25 plots
60
50
40
30
20
10
0
Average Frames per Second
In-Memory Data Cube
10k
100k
1M
10M
100M
1B
Number of Data Points

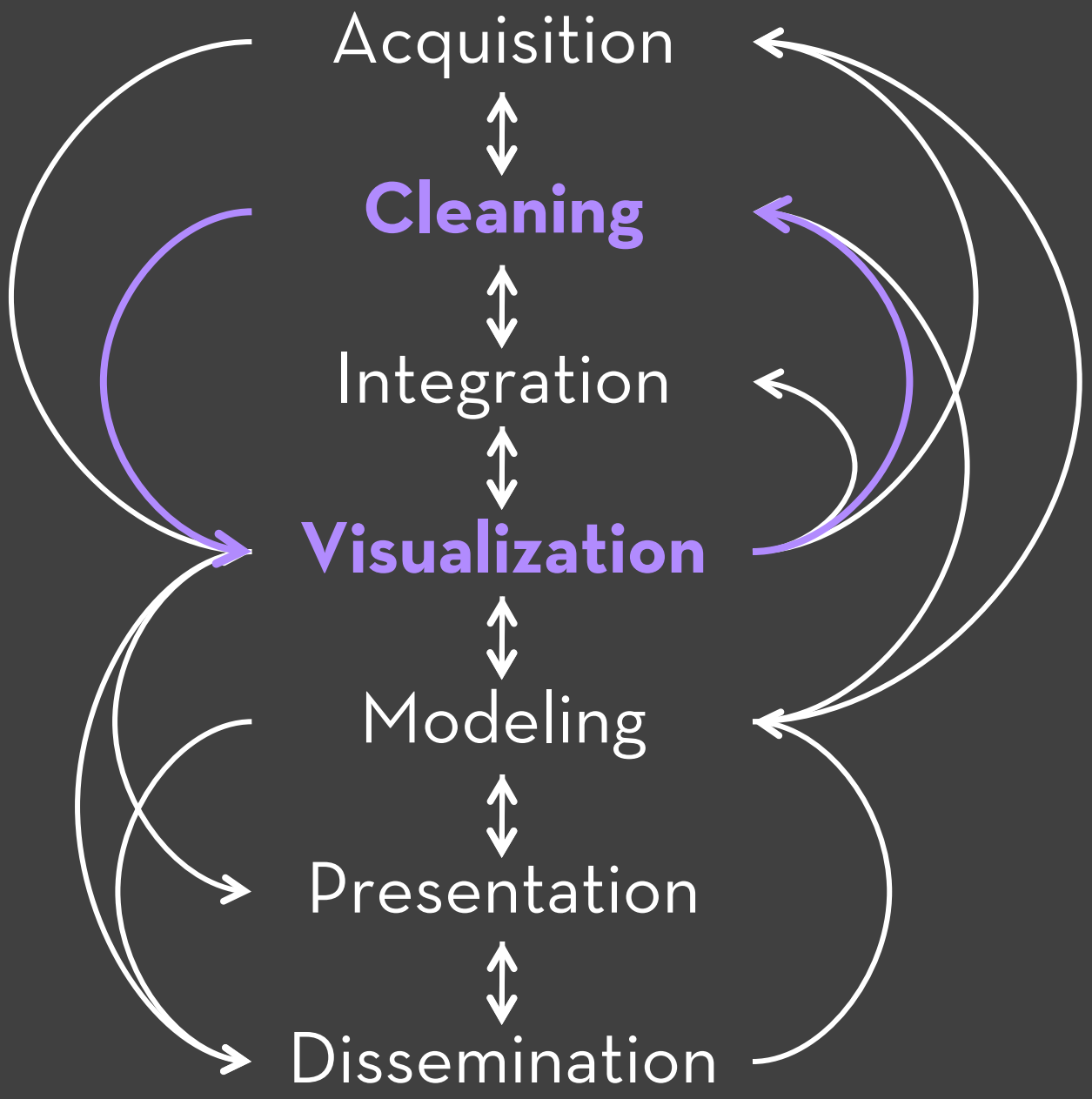

5 dimensions x 50 bins/dim x 25 plots
60
50
40
30
20
10
0
Average Frames per Second
imMens
In-Memory Data Cube
10k
100k
1M
10M
100M
1B
Number of Data Points

5 dimensions x 50 bins/dim x 25 plots




Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

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

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		

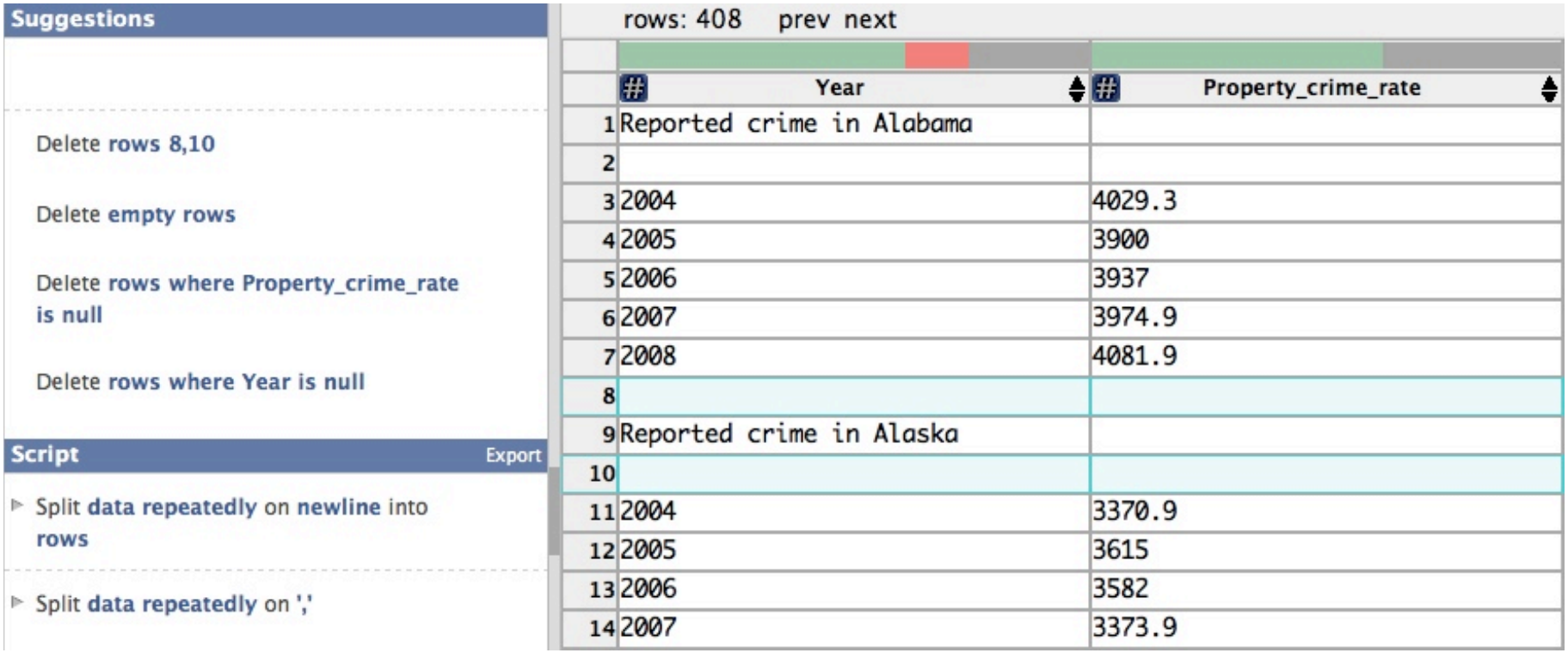
Reported crime in California

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		

DataWrangler



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

Wrangler in 2 Parts...

1. Data transformation language
   Map Operations – split, merge, extract, drop
   Reshaping – fold, pivot (cross-tabulate)
   Lookups & Joins – e.g., FIPS code to US state
   Sorting, Aggregation, *etc.*

Wrangler in 2 Parts...

1. 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 **visual preview and refinement**

Comparative Evaluation with Excel

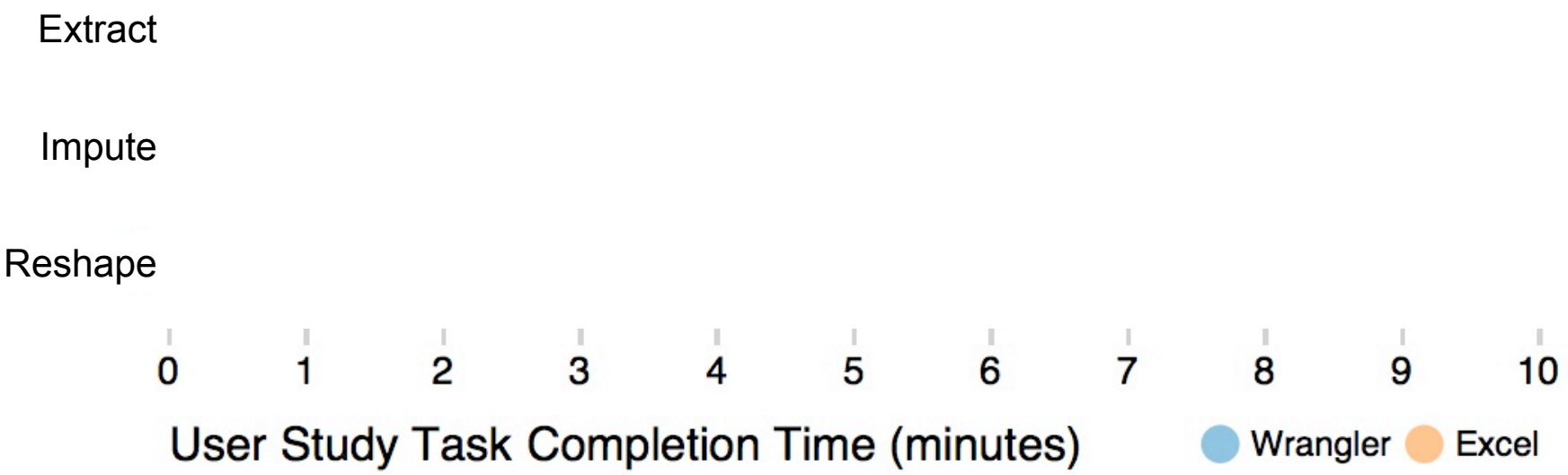
Comparative Evaluation with Excel

Extract

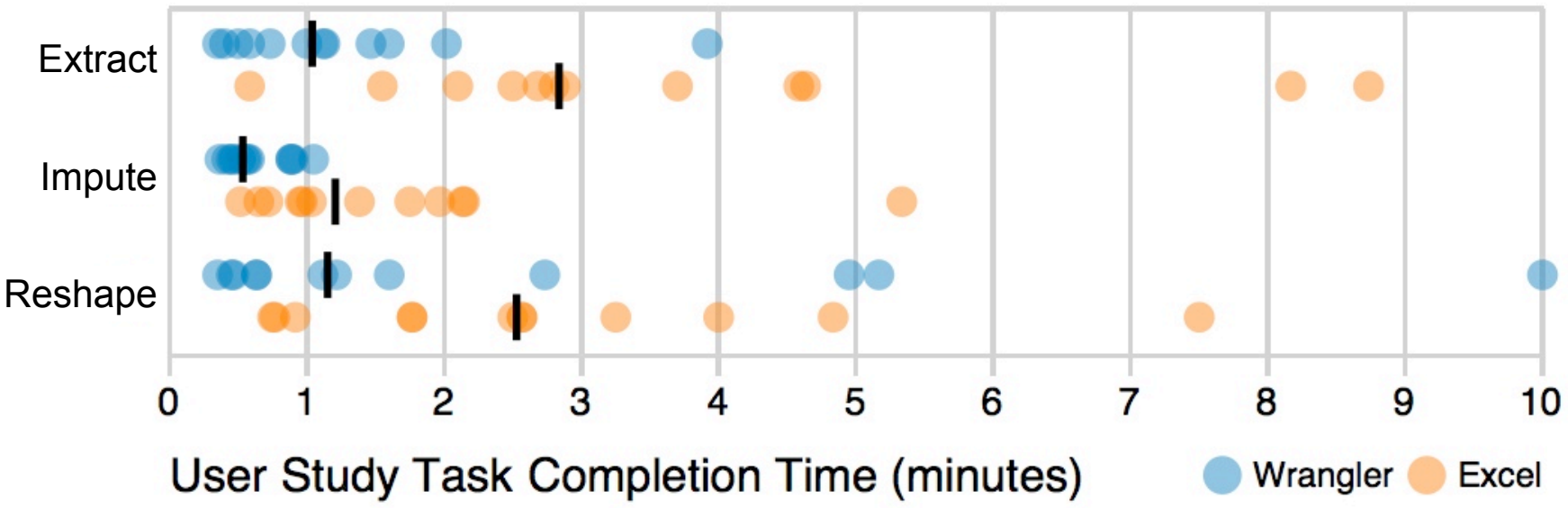
Impute

Reshape

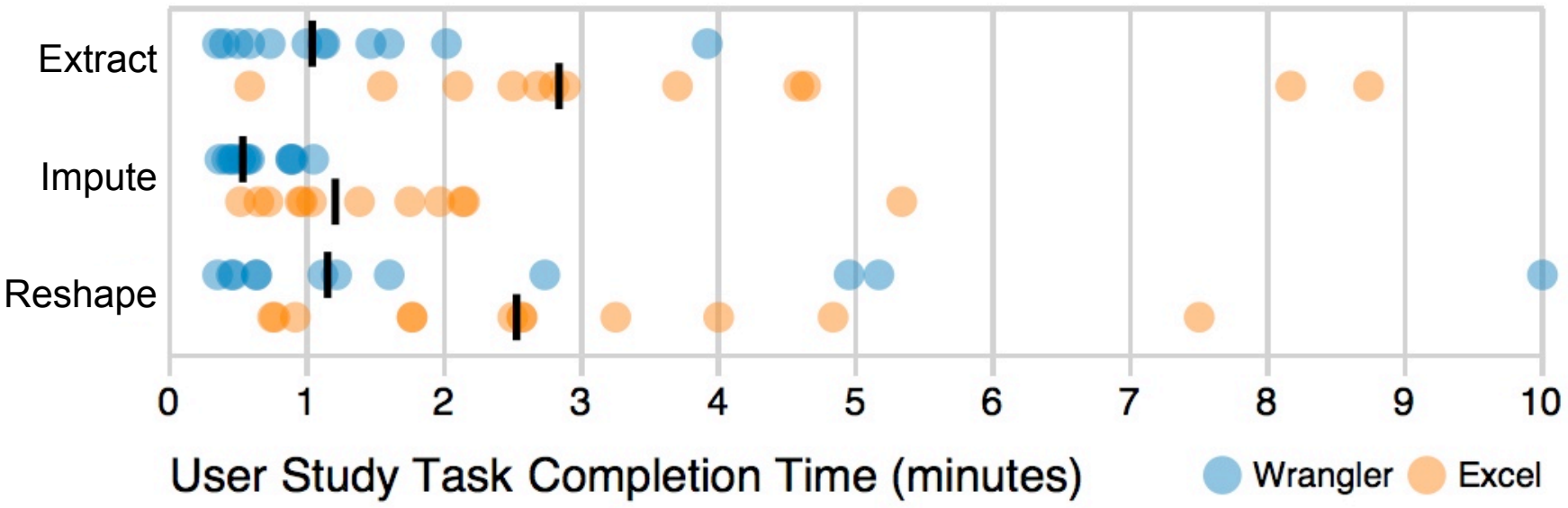
Comparative Evaluation with Excel



Comparative Evaluation with Excel



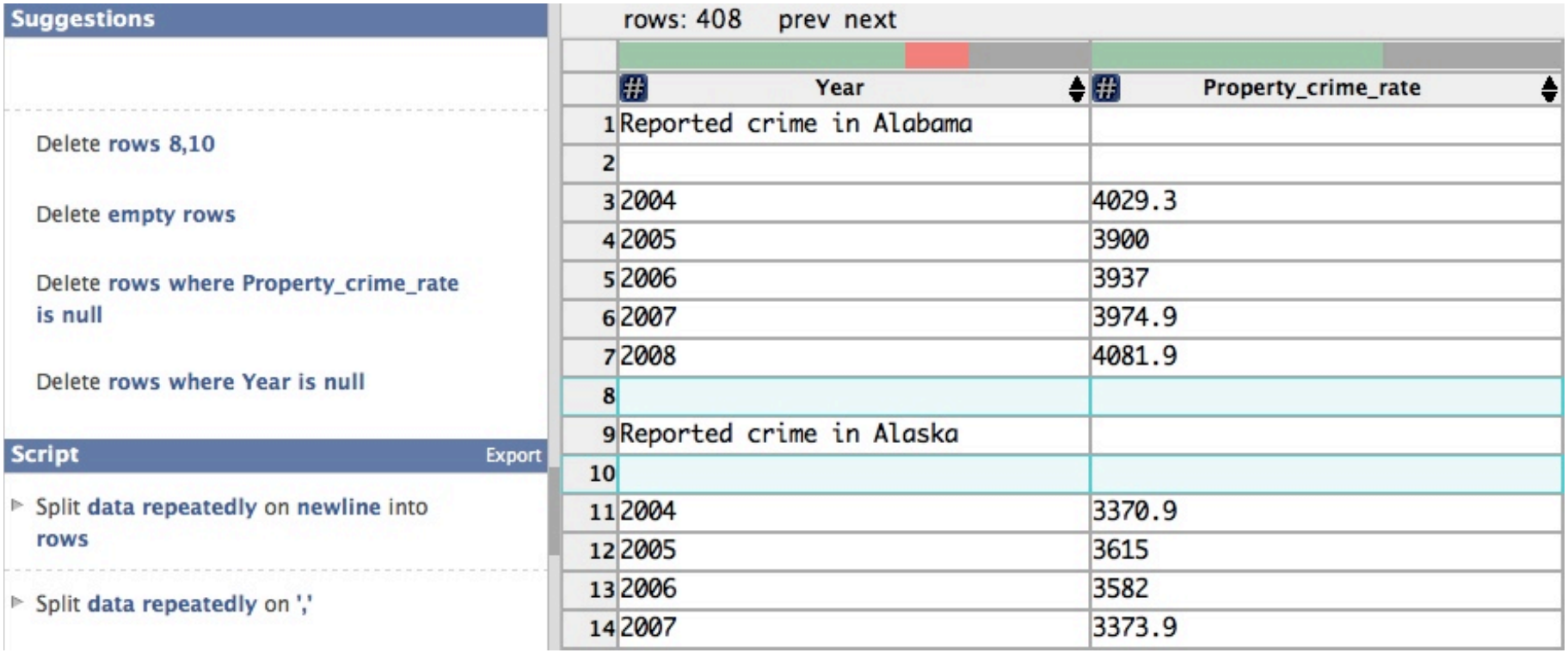
Comparative Evaluation with Excel



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

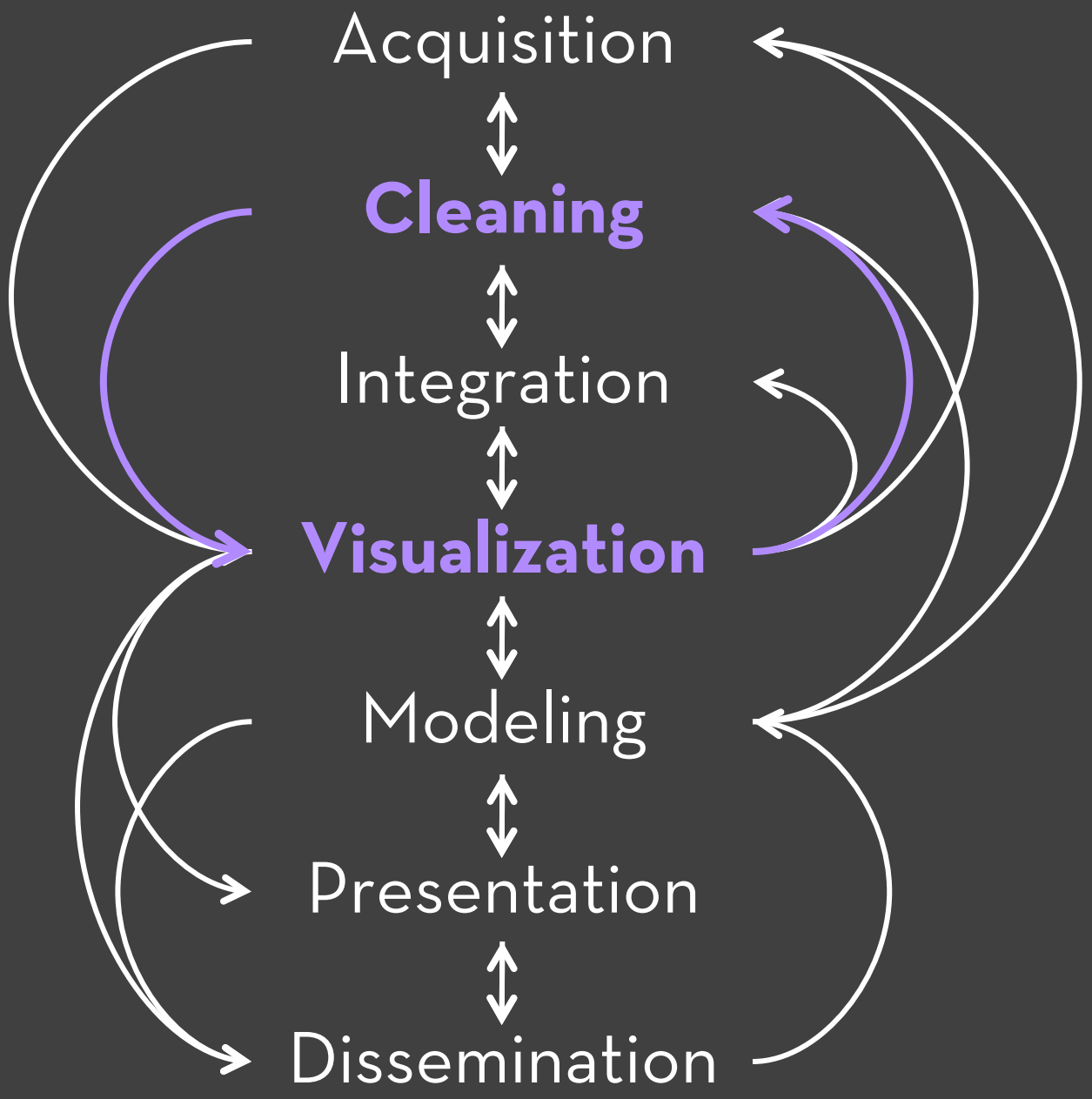
Suggestions and visual previews used heavily.

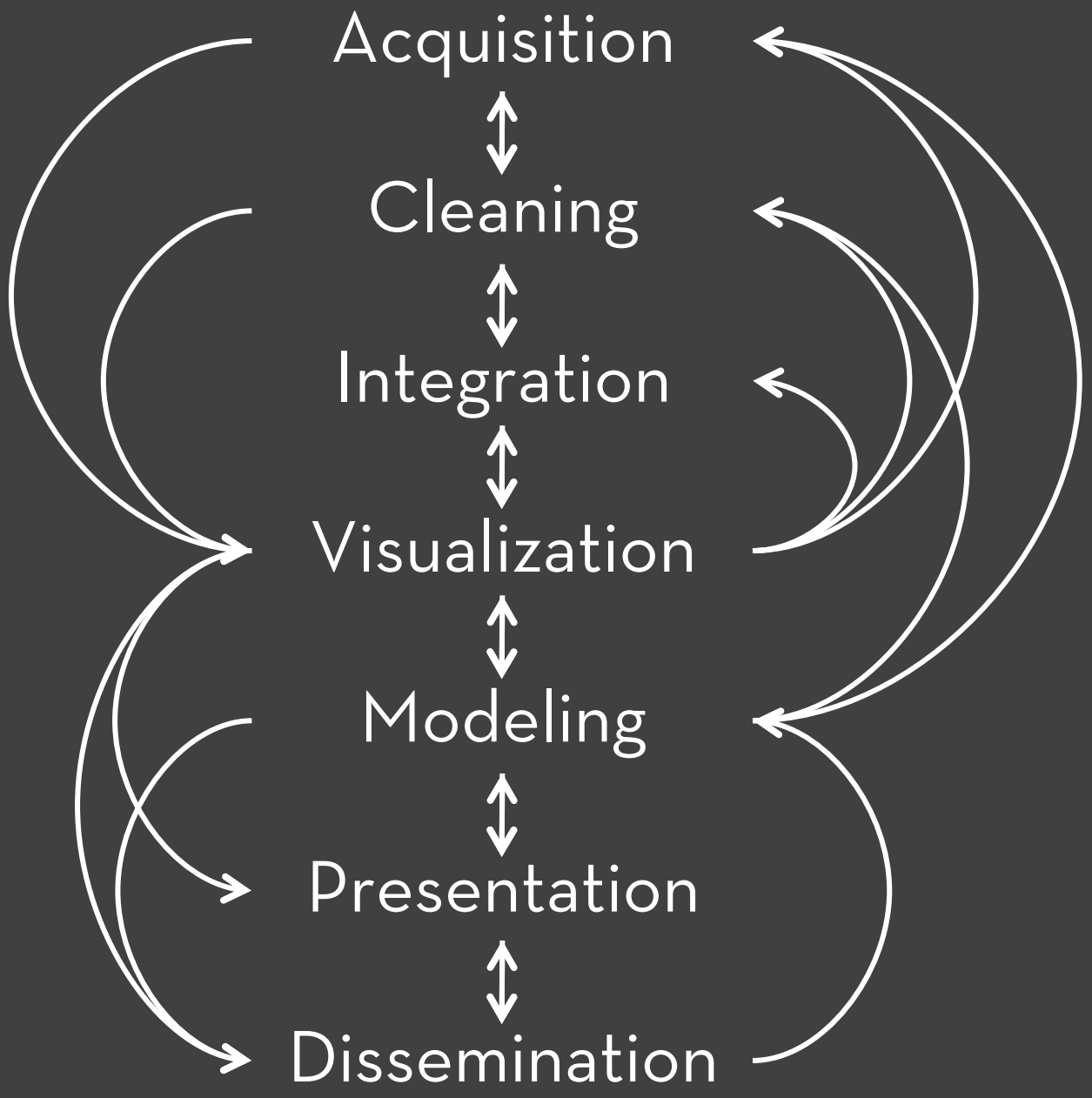
DataWrangler



http://vis.stanford.edu/wrangler

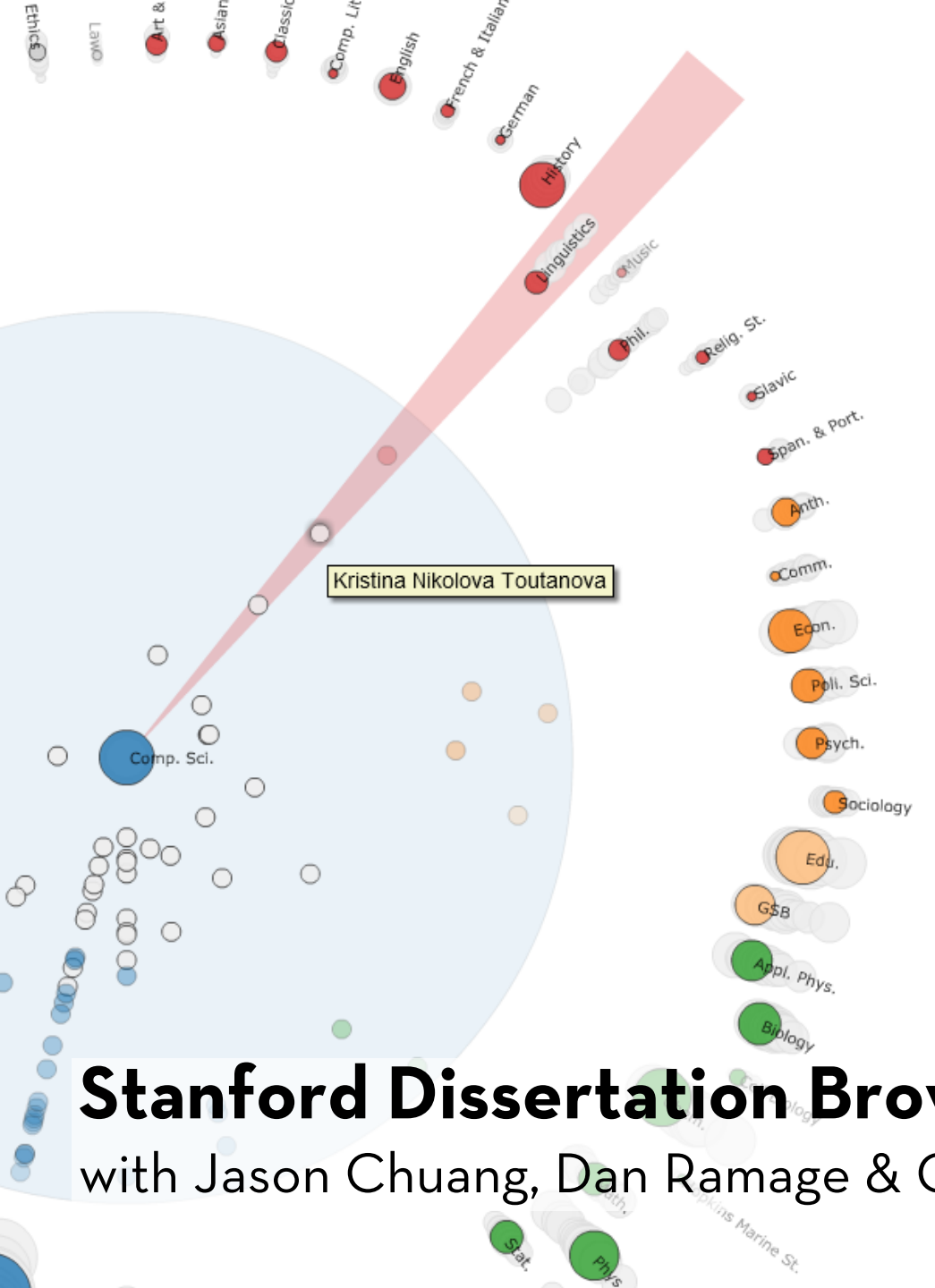

TRIFACTA


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

Interactive ∧ Data Analysis

Jeffrey Heer
http://idl.cs.washington.edu



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

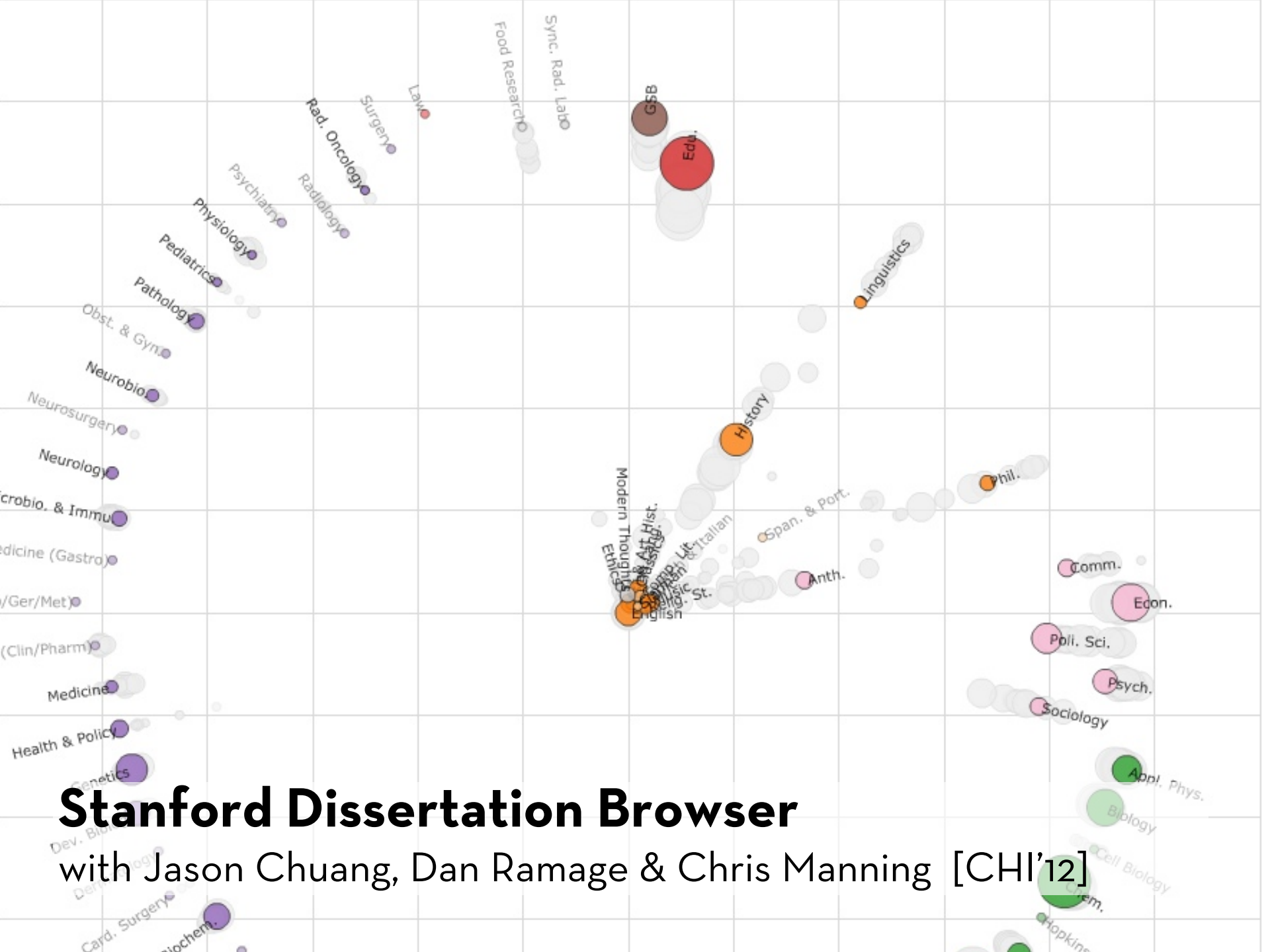
Keywords: Syntactic, Semantic, Tree kernels, Parsing

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]



Stanford Dissertation Browser

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