Big data visual analytics systems

CS594: Big Data Visualization & Analytics

Fabio Miranda

https://fmiranda.me

COMPUTER SCIENCE

Visual analytics

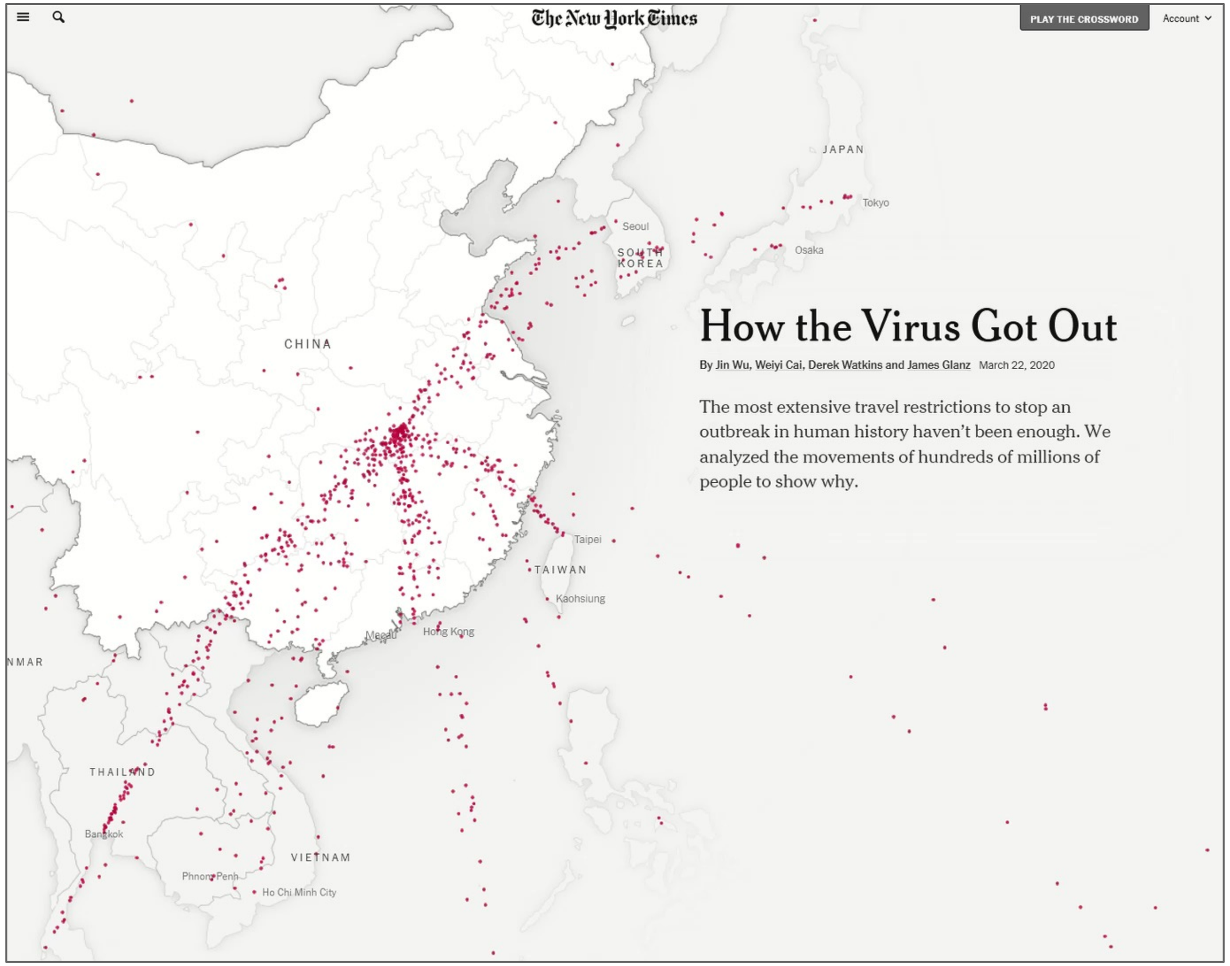
"*Visual analytics is the formation of abstract visual metaphors in combination with a human information interaction that enables <u>detection of the expected</u> and <u>discovery of the unexpected</u> within massive, dynamically changing information spaces.*"

[Wong and Thomas, 2004]

Visual analytics, infovis, scivis

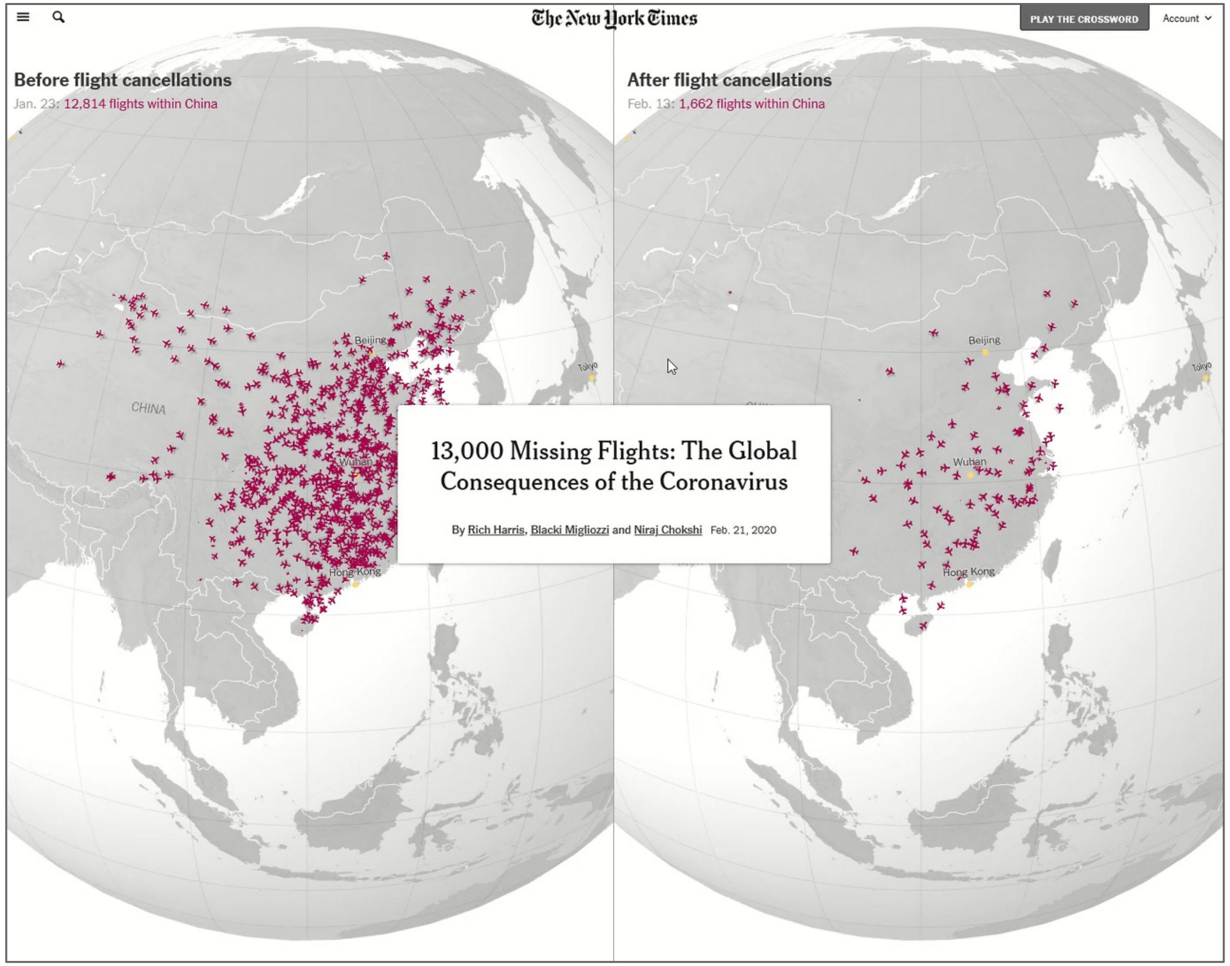
- Scientific visualization: data that has a natural geometric structure (wind flows, MRI).
- Information visualization: abstract representations (trees, graphs).
- Visual analytics: interactive visual representations <u>**and**</u> underlying analytical processes (data mining, stats, etc.)

Example: communicating findings



https://www.nytimes.com/interactive/2020/03/22/world/coronavirus-spread.html

Example: communicating findings

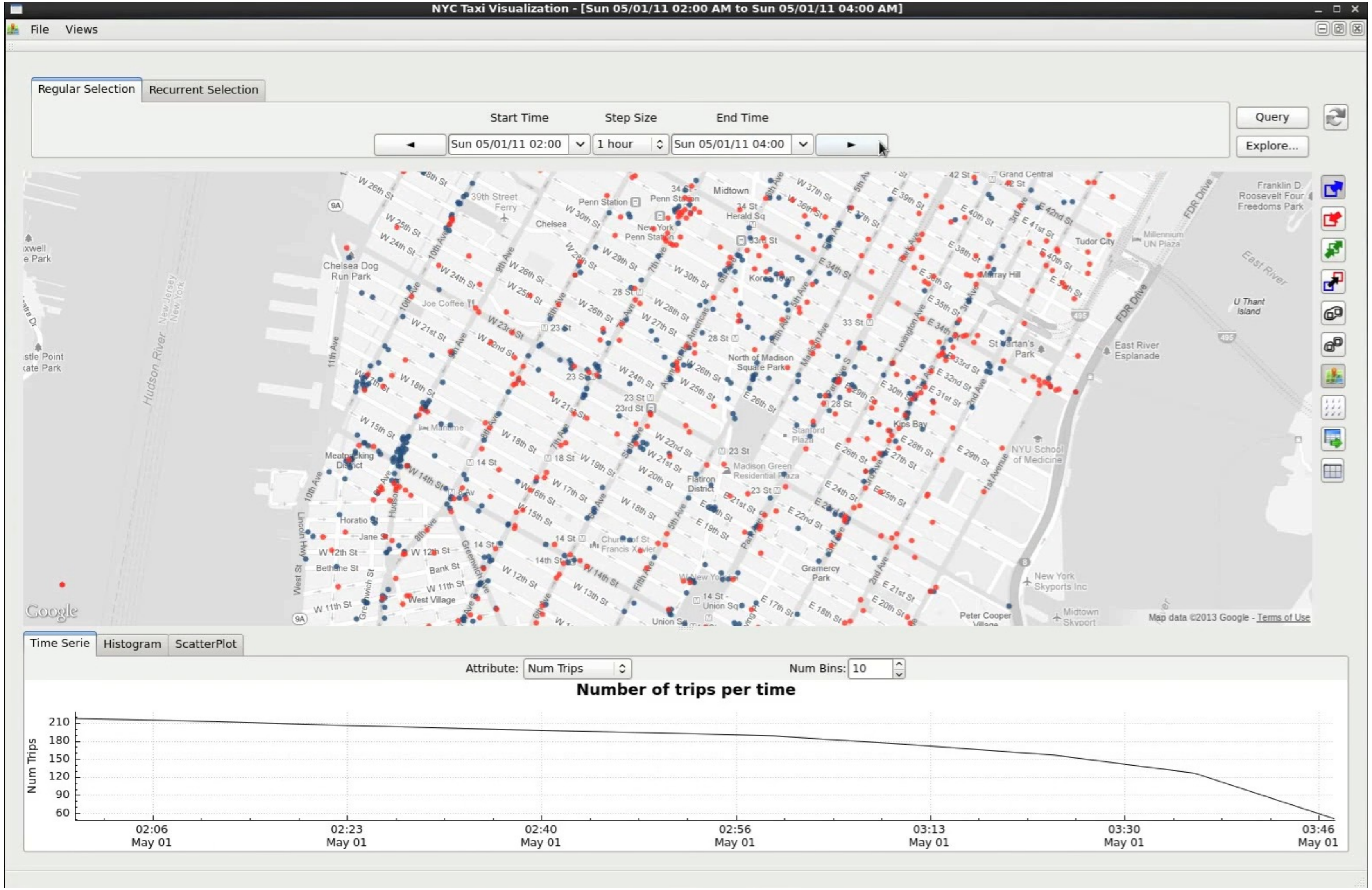


https://www.nytimes.com/interactive/2020/02/21/business/coronavirus-airline-travel.html

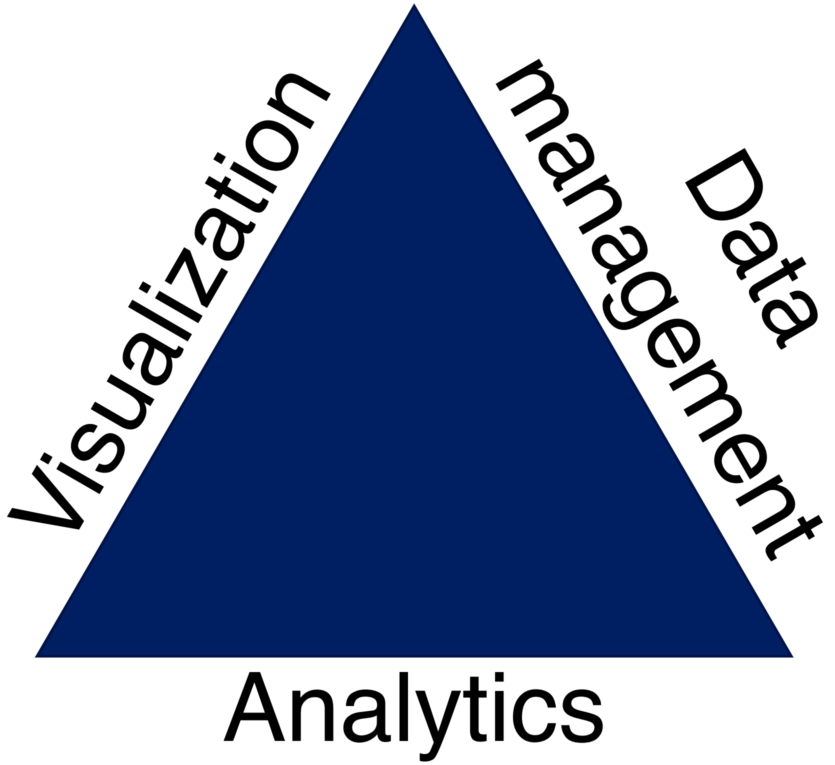
Example: exploration + communicating findings

https://www.nytimes.com/interactive/2016/12/21/upshot/Mapping-the-Shadows-of-New-York-City.html

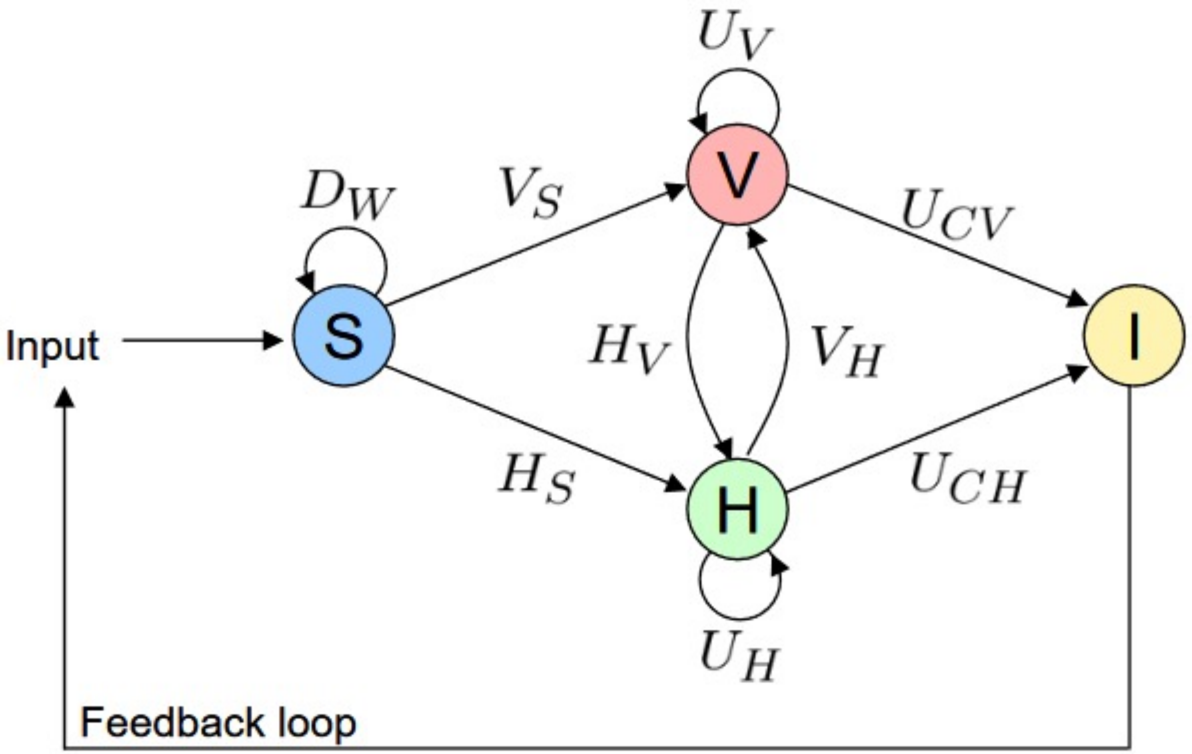
Example: visual analysis of data



Visual analytics



Visual analytics process



[“Visual Analytics: Scope and Challenges”, Kleim et al., 2008]

$F: S \rightarrow I$
$f \in \{D_w, V_x, H_y, U_z\}$
S: data, I: insight

D_w: data pre-processing functionalities (data transformation, data cleaning, data selection, data integration)

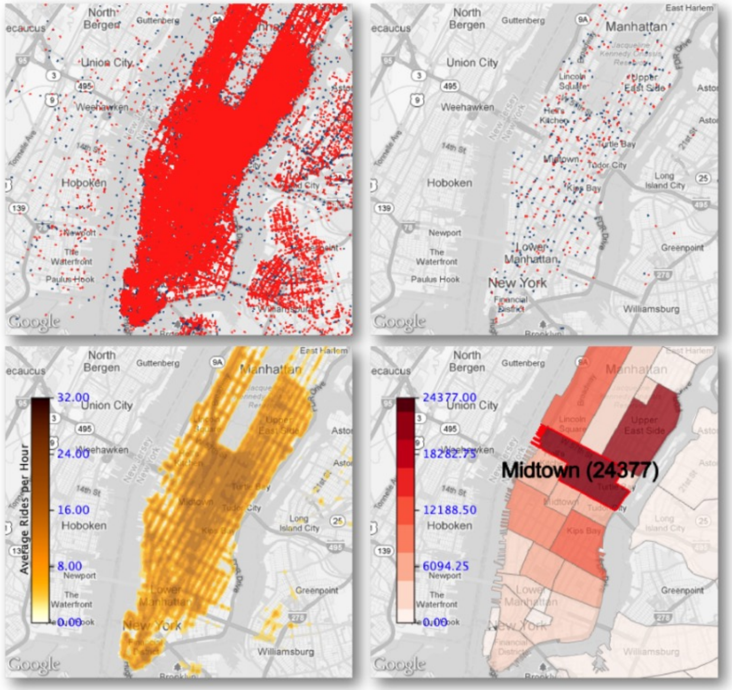
$V_x, x \in \{S, H\}$: visualization functions
- V_S: functions visualizing data ($S \rightarrow V$)
- V_H: functions visualizing hypotheses ($\mathrm{H} \rightarrow V$)

$H_y, y \in \{S, V\}$: hypotheses generation process
- H_S: hypotheses from automatic process ($S \rightarrow H$)
- H_v: hypotheses from visualization ($\mathrm{V} \rightarrow H$)

$U_z, z \in \{V, H, CV, CH\}$: user interactions
- U_V: visualization interaction ($V \rightarrow V$)
- U_H: hypotheses interaction ($\mathrm{H} \rightarrow H$)

U_{cv}: insight concluded from visualization
U_{ch}: insight concluded from hypothesis

Visual analysis process: taxi data



Distribution of NYC Taxi Pickups
and Dropoffs in Manhattan

Visual analysis process: taxi data

- Data: ~500,000 trips / day; ~1,000,000,000 trips in 5 years
  - Spatiotemporal: pickup + dropoff
  - Trip attributes: distance traveled, fare, tip, etc.
- Government, policy makers and scientists are usually unable to *interactive* explore the *whole* data.
  - Too many slices to examine.

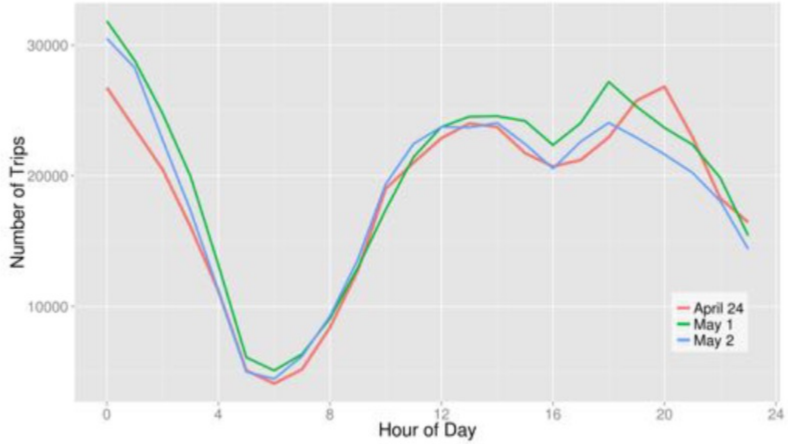
Visual analysis process: taxi data

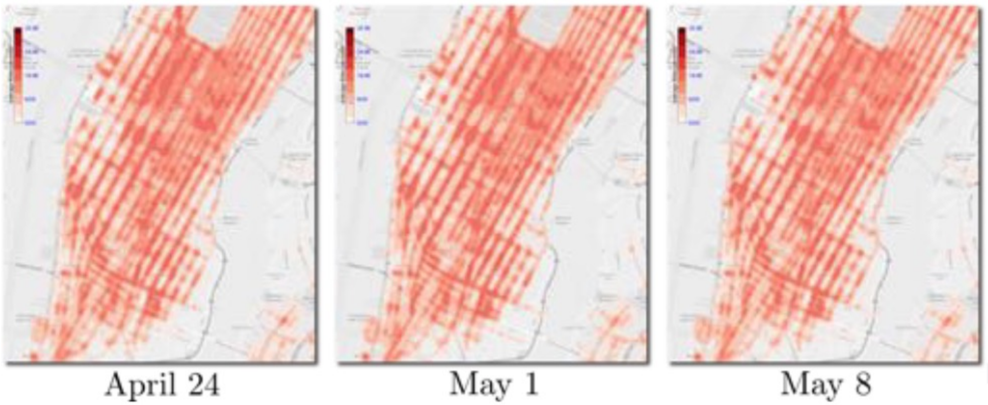
(a) (b) (c)

- Taxi pickups and drop-offs: data sets $S_1, S_2, \dots, S_n$
- Pre-processing
  - Data cleaning (e.g., remove points outside NYC): $d_c(S_1, \dots, S_n)$
  - Data integration (e.g., add weather column): $d_i(S_1, \dots, S_n)$
  - Data transformation (e.g., raw csv's to binary): $d_t(S_1, \dots, S_n)$
  - Data selection (e.g., only Manhattan trips): $d_s(S_1, \dots, S_n)$

Visual analysis process: taxi data

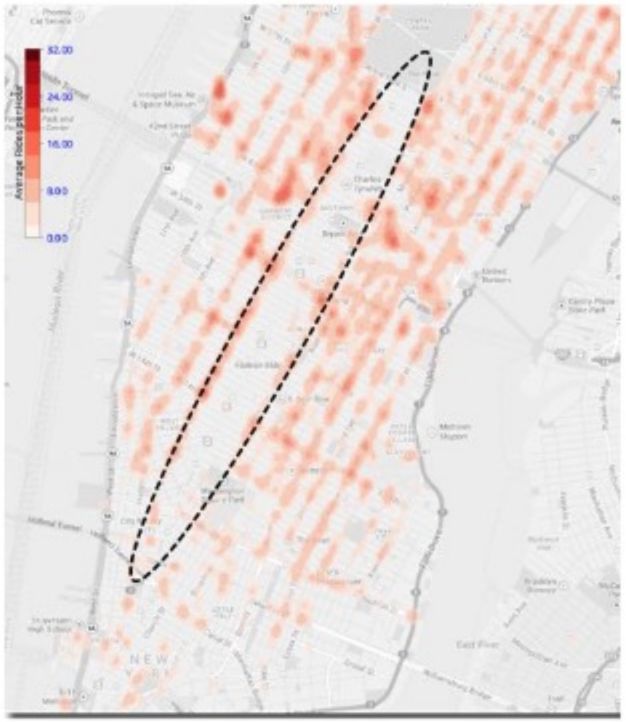
- Too many slices
- 365*24 1-hour slices in just one year
- Which slices are interesting?





Visual analysis process: taxi data

- Too many slices
- 365*24 1-hour slices in just one year
- Which slices are interesting?



May 1 (8-9am)

Visual analysis process: taxi data

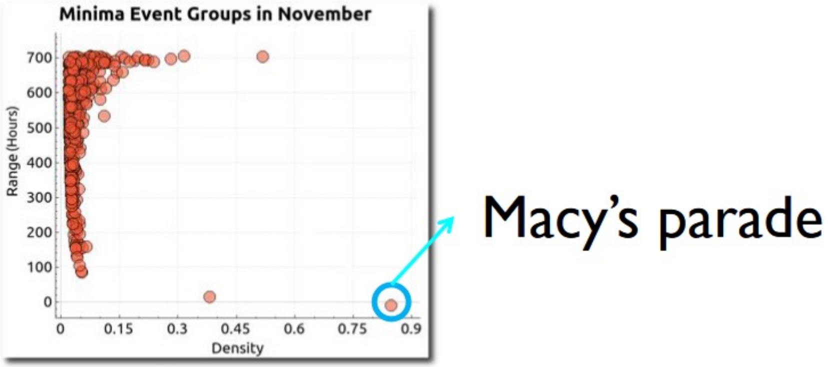
- Too many slices
- 365*24 1-hour slices in just one year
- Which slices are interesting?

- Guide users towards *interesting* data slices
- Data mining:
  - Identifying events with arbitrary spatial structure at multiple temporal scales:

$$h' = h_s(S')$$
$$with\ S' = d_s(d_t(d_i(d_c(S_1, \ldots, S_n))))$$

Visual analysis process: taxi data

- Visualization:
  - Visualizing the previous hypothesis: $v' = v_h(h')$
  - Adjusting parameters: $v'' = u_v(v')$
  - Hypotheses from visualization: $h'' = h_v(v'')$, leading to an insight (unusual taxi activity around Macy's parade).



Visual analysis process: taxi data

- Taxi pickups and drop-offs: data sets $S_1, S_2, \dots, S_n$
- Pre-processing
  - Data cleaning (e.g., removing): $d_c(S_1, \dots, S_n)$
  - Data integration (e.g., adding extra attributes): $d_i(S_1, \dots, S_n)$
  - Data transformation (e.g., format suitable for analysis): $d_t(S_1, \dots, S_n)$
  - Data selection (e.g., subset of sensors): $d_s(S_1, \dots, S_n)$
- Data mining:
  - Identifying events with arbitrary spatial structure at multiple temporal scales:

$$h' = h_s(S'), with\ S' = d_s\left(d_t\left(d_i\left(d_c(S_1, \dots, S_n)\right)\right)\right)$$

- Visualization:
  - Visualizing the previous hypothesis: $v' = v_h(h')$
  - Adjusting parameters: $v'' = u_v(v')$
  - Hypotheses from visualization: $h'' = h_v(v'')$

$$h_v(u_v(v_h(h_s(d_s(d_t(d_i(d_c(S_1, \dots, S_n))))))))$$

Visual analytics mantra

- Visualization mantra:

"Overview first, zoom / filter, details on demand"

[Shneiderman, 1996]

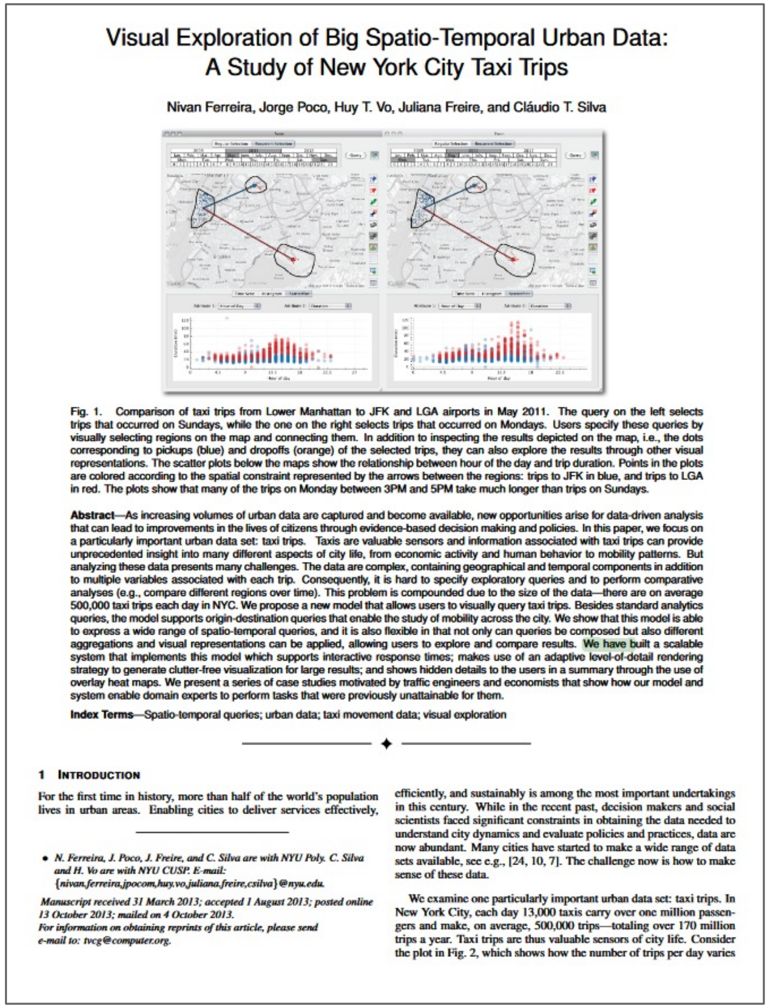
- Visual analytics mantra:

"Analyze first - show the important - zoom, filter and analyze further - details on demand"

[Kleim et al., 2008]

Visual analytics + big data

"We propose a new model that allows users to visually query taxi trips. Besides standard analytics queries, the model supports origin-destination queries that enable the study of mobility across the city. We show that this model is able to express a wide range of spatiotemporal queries, and it is also flexible in that not only can queries be composed but also different aggregations and visual representations can be applied, allowing users to explore and compare results."



Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips

Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Cláudio T. Silva

Fig. 1. Comparison of taxi trips from Lower Manhattan to JFK and LGA airports in May 2011. The query on the left selects trips that occurred on Sundays, while the one on the right selects trips that occurred on Mondays. Users specify these queries by visually selecting regions on the map and connecting them. In addition to inspecting the results depicted on the map, i.e., the dots corresponding to pickups (blue) and dropoffs (orange) of the selected trips, they can also explore the results through other visual representations. The scatter plots below the maps show the relationship between hour of the day and trip duration. Points in the plots are colored according to the spatial constraint represented by the arrows between the regions: trips to JFK in blue, and trips to LGA in red. The plots show that many of the trips on Monday between 3PM and 5PM take much longer than trips on Sundays.

Abstract—As increasing volumes of urban data are captured and become available, new opportunities arise for data-driven analysis that can lead to improvements in the lives of citizens through evidence-based decision making and policies. In this paper, we focus on a particularly important urban data set: taxi trips. Taxis are valuable sensors and information associated with taxi trips can provide unprecedented insight into many different aspects of city life, from economic activity and human behavior to mobility patterns. But analyzing these data presents many challenges. The data are complex, containing geographical and temporal components in addition to multiple variables associated with each trip. Consequently, it is hard to specify exploratory queries and to perform comparative analyses (e.g., compare different regions over time). This problem is compounded due to the size of the data—there are on average 500,000 taxi trips each day in NYC. We propose a new model that allows users to visually query taxi trips. Besides standard analytics queries, the model supports origin-destination queries that enable the study of mobility across the city. We show that this model is able to express a wide range of spatio-temporal queries, and it is also flexible in that not only can queries be composed but also different aggregations and visual representations can be applied, allowing users to explore and compare results. We have built a scalable system that implements this model which supports interactive response times; makes use of an adaptive level-of-detail rendering strategy to generate clutter-free visualization for large results; and shows hidden details to the users in a summary through the use of overlay heat maps. We present a series of case studies motivated by traffic engineers and economists that show how our model and system enable domain experts to perform tasks that were previously unattainable for them.

Index Terms—Spatio-temporal queries; urban data; taxi movement data; visual exploration

1 INTRODUCTION

For the first time in history, more than half of the world's population lives in urban areas. Enabling cities to deliver services effectively, efficiently, and sustainably is among the most important undertakings in this century. While in the recent past, decision makers and social scientists faced significant constraints in obtaining the data needed to understand city dynamics and evaluate policies and practices, data are now abundant. Many cities have started to make a wide range of data sets available, see e.g., [24, 10, 7]. The challenge now is how to make sense of these data.

We examine one particularly important urban data set: taxi trips. In New York City, each day 13,000 taxis carry over one million passengers and make, on average, 500,000 trips—totaling over 170 million trips a year. Taxi trips are thus valuable sensors of city life. Consider the plot in Fig. 2, which shows how the number of trips per day varies

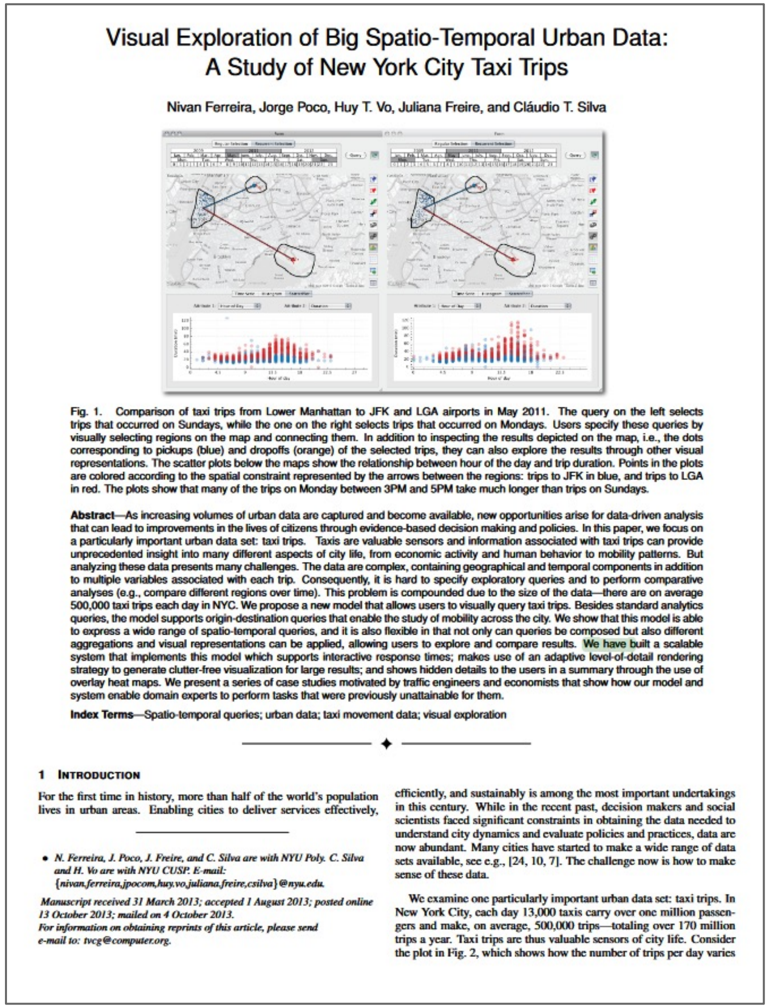
- *N. Ferreira, J. Poco, J. Freire, and C. Silva are with NYU Poly. C. Silva and H. Vo are with NYU CUSP. E-mail: {nivan.ferreira,jpocom,huy.vo,juliana.freire,csilva}@nyu.edu.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

[Ferreira et al., 2013]

TaxiVis

- Interaction capabilities that enable users to pose queries over all the dimensions of the data and flexibly explore the associated attributes.
- Compare spatiotemporal slices through multiple coordinated views.
- Interactively compose and refine queries, and generalize them by performing parameter sweeps.
- Efficient data stored, and adaptive level-of-detail rendering to provide clutter-free visualizations.



Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips

Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Cláudio T. Silva

Fig. 1. Comparison of taxi trips from Lower Manhattan to JFK and LGA airports in May 2011. The query on the left selects trips that occurred on Sundays, while the one on the right selects trips that occurred on Mondays. Users specify these queries by visually selecting regions on the map and connecting them. In addition to inspecting the results depicted on the map, i.e., the dots corresponding to pickups (blue) and dropoffs (orange) of the selected trips, they can also explore the results through other visual representations. The scatter plots below the maps show the relationship between hour of the day and trip duration. Points in the plots are colored according to the spatial constraint represented by the arrows between the regions: trips to JFK in blue, and trips to LGA in red. The plots show that many of the trips on Monday between 3PM and 5PM take much longer than trips on Sundays.

Abstract—As increasing volumes of urban data are captured and become available, new opportunities arise for data-driven analysis that can lead to improvements in the lives of citizens through evidence-based decision making and policies. In this paper, we focus on a particularly important urban data set: taxi trips. Taxis are valuable sensors and information associated with taxi trips can provide unprecedented insight into many different aspects of city life, from economic activity and human behavior to mobility patterns. But analyzing these data presents many challenges. The data are complex, containing geographical and temporal components in addition to multiple variables associated with each trip. Consequently, it is hard to specify exploratory queries and to perform comparative analyses (e.g., compare different regions over time). This problem is compounded due to the size of the data—there are on average 500,000 taxi trips each day in NYC. We propose a new model that allows users to visually query taxi trips. Besides standard analytics queries, the model supports origin-destination queries that enable the study of mobility across the city. We show that this model is able to express a wide range of spatio-temporal queries, and it is also flexible in that not only can queries be composed but also different aggregations and visual representations can be applied, allowing users to explore and compare results. We have built a scalable system that implements this model which supports interactive response times; makes use of an adaptive level-of-detail rendering strategy to generate clutter-free visualization for large results; and shows hidden details to the users in a summary through the use of overlay heat maps. We present a series of case studies motivated by traffic engineers and economists that show how our model and system enable domain experts to perform tasks that were previously unattainable for them.

Index Terms—Spatio-temporal queries; urban data; taxi movement data; visual exploration

1 INTRODUCTION

For the first time in history, more than half of the world's population lives in urban areas. Enabling cities to deliver services effectively, efficiently, and sustainably is among the most important undertakings in this century. While in the recent past, decision makers and social scientists faced significant constraints in obtaining the data needed to understand city dynamics and evaluate policies and practices, data are now abundant. Many cities have started to make a wide range of data sets available, see e.g., [24, 10, 7]. The challenge now is how to make sense of these data.

We examine one particularly important urban data set: taxi trips. In New York City, each day 13,000 taxis carry over one million passengers and make, on average, 500,000 trips—totaling over 170 million trips a year. Taxi trips are thus valuable sensors of city life. Consider the plot in Fig. 2, which shows how the number of trips per day varies

- *N. Ferreira, J. Poco, J. Freire, and C. Silva are with NYU Poly. C. Silva and H. Vo are with NYU CUSP. E-mail: {nivan.ferreira,jpocom,huy.vo,juliana.freire,csilva}@nyu.edu.*

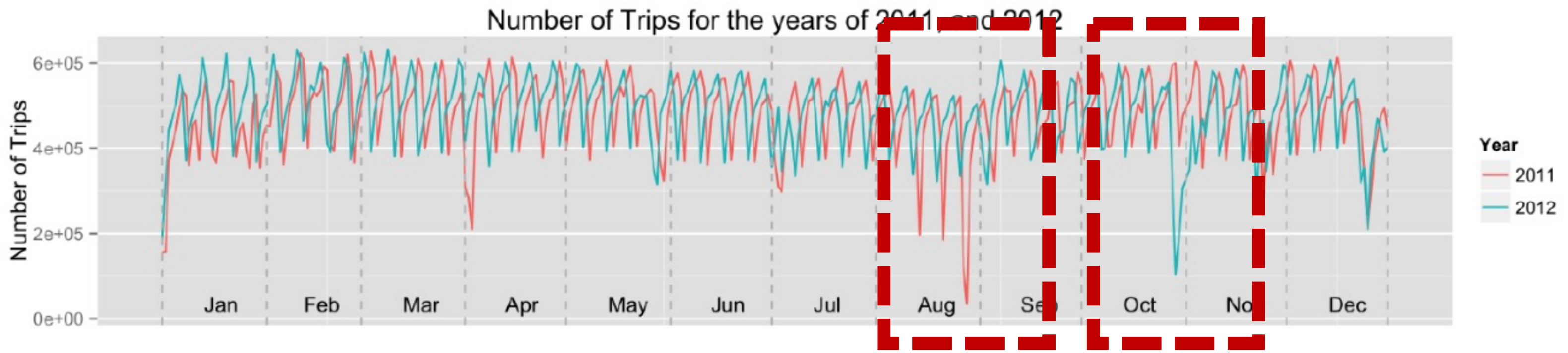
Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

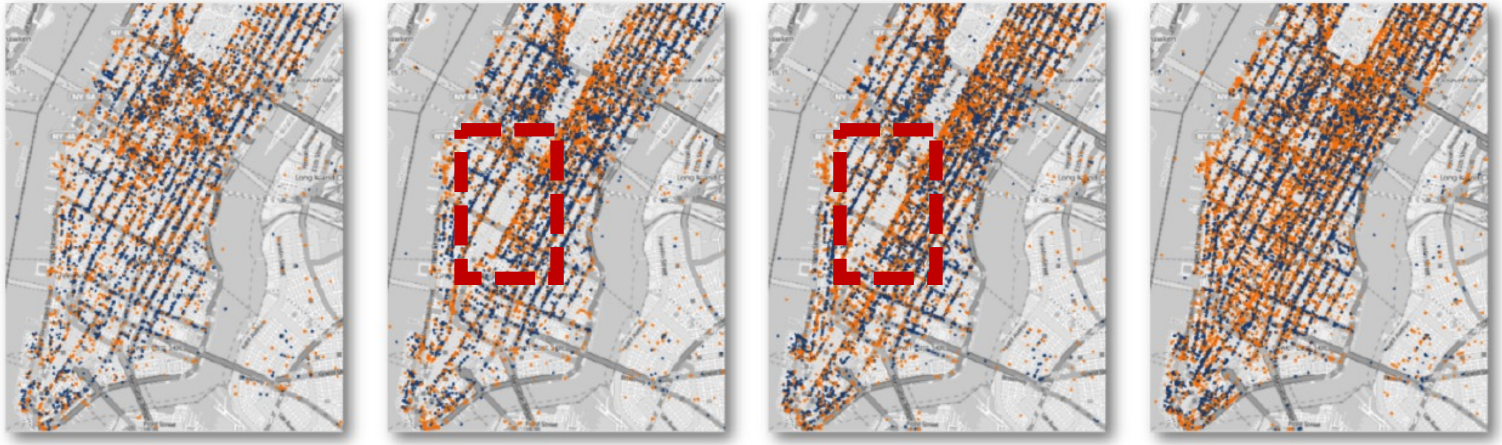
[Ferreira et al., 2013]

Taxi as sensors of city life

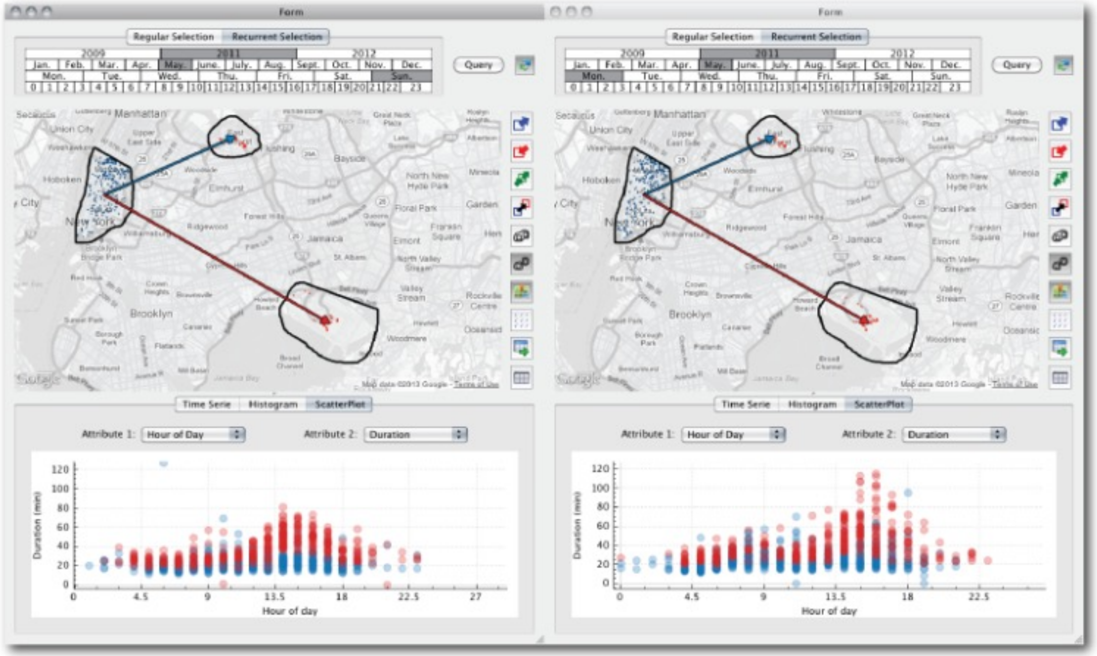
- Understanding the dynamics of the city, how different aspects of the data vary over space and time:
  - "*What is the average trip time from Midtown to the airports during weekdays?*"
  - "*How does the movement changes between Midtown and JFK throughout the day, over different days of the week?*"
  - "*How does the taxi fleet activity vary during weekdays?*"
- Ability to quickly test hypotheses: starting with one query about a specific place ("*movement patterns between Midtown and JFK*") and generalize to all neighborhoods.

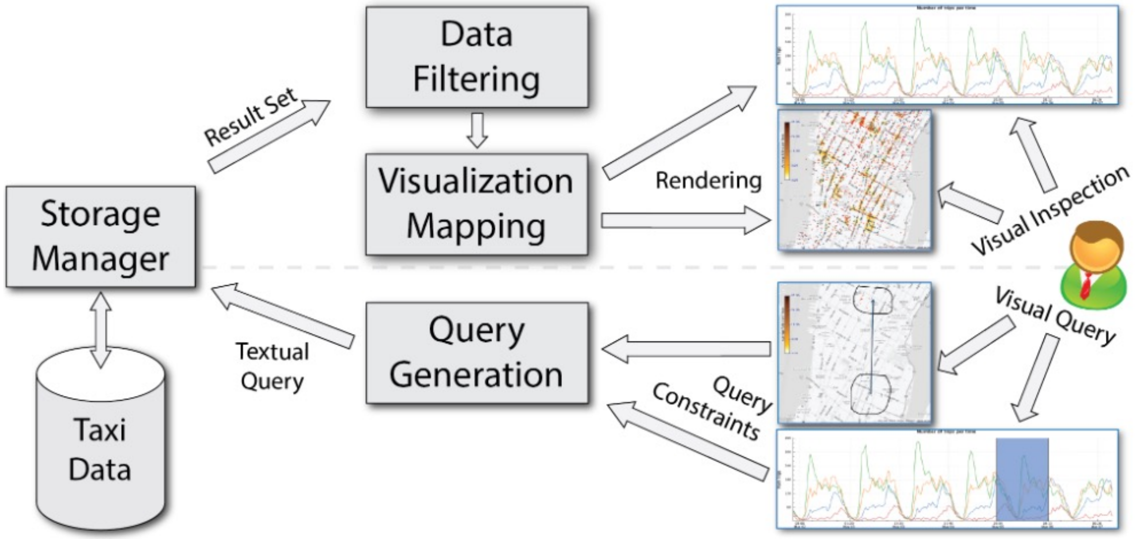
Taxis as sensors of city life



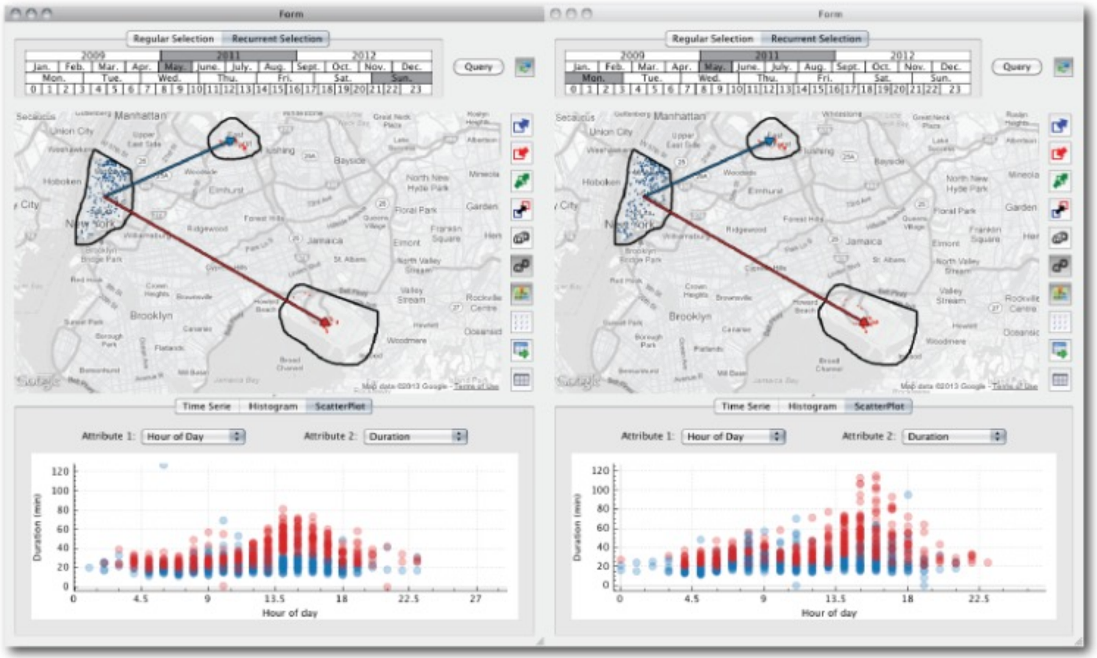


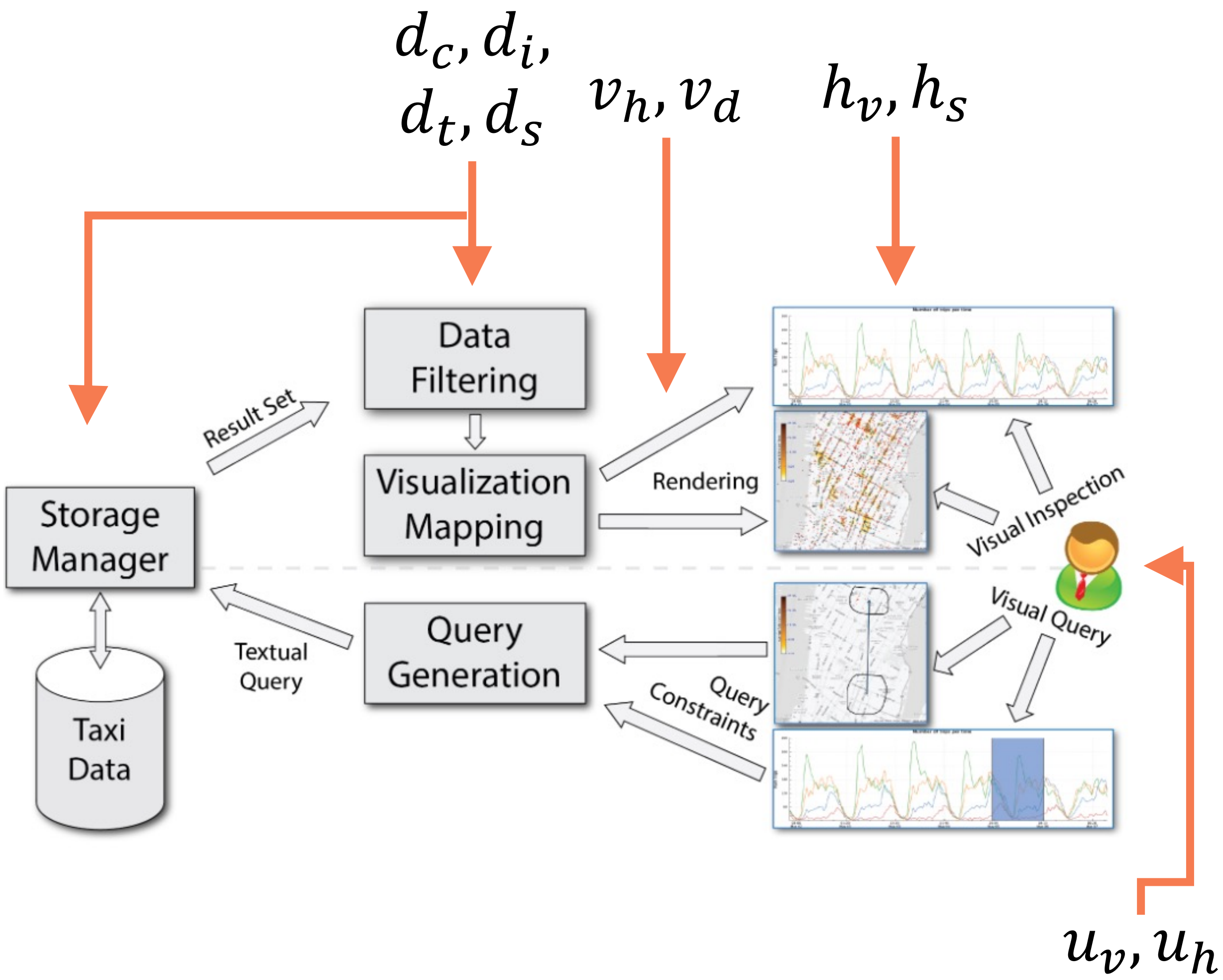
TaxiVis




Data Filtering
Visualization Mapping
Storage Manager
Taxi Data
Query Generation
Result Set
Textual Query
Rendering
Visual Inspection
Visual Query
Query Constraints

TaxiVis





$d_c, d_i,$
d_t, d_s
v_h, v_d
h_v, h_s
Data Filtering
Visualization Mapping
Query Generation
Storage Manager
Taxi Data
Result Set
Textual Query
Rendering
Query Constraints
Visual Inspection
Visual Query
u_v, u_h

Usability through visual operations

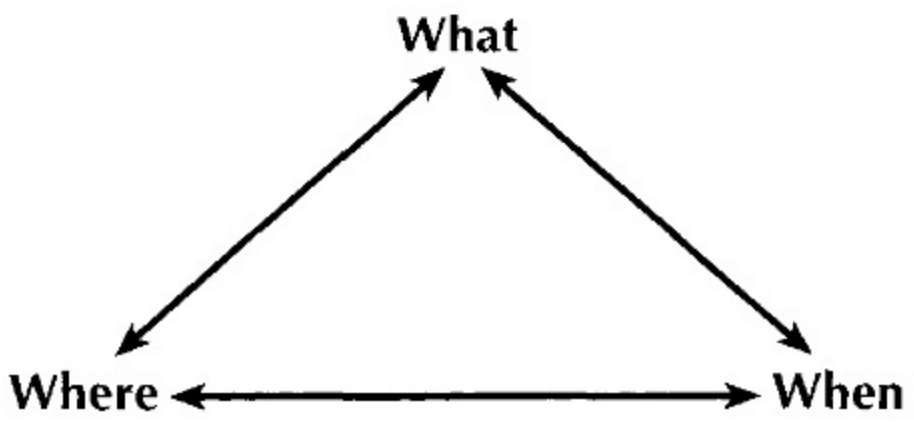
```
SELECT *
FROM trips
WHERE pickup_time in (5/1/11,5/7/11)
AND dropoff_loc in “Times Square”
AND pickup_loc in “Gramercy”
```

Data selection and result exploration are unified

Visual query model

- Expressive Triad framework:
  - *Where + when → what*: describe the objects (what) that are present at certain locations and times.
  - *When + what → where*: describe the locations occupied by objects at a given time.
  - *Where + what → when*: describe the times that objects occupied a given location.

THE BASIC VIEW COMPONENTS OF THE TRIAD FRAMEWORK



[Peuquet, 1994]

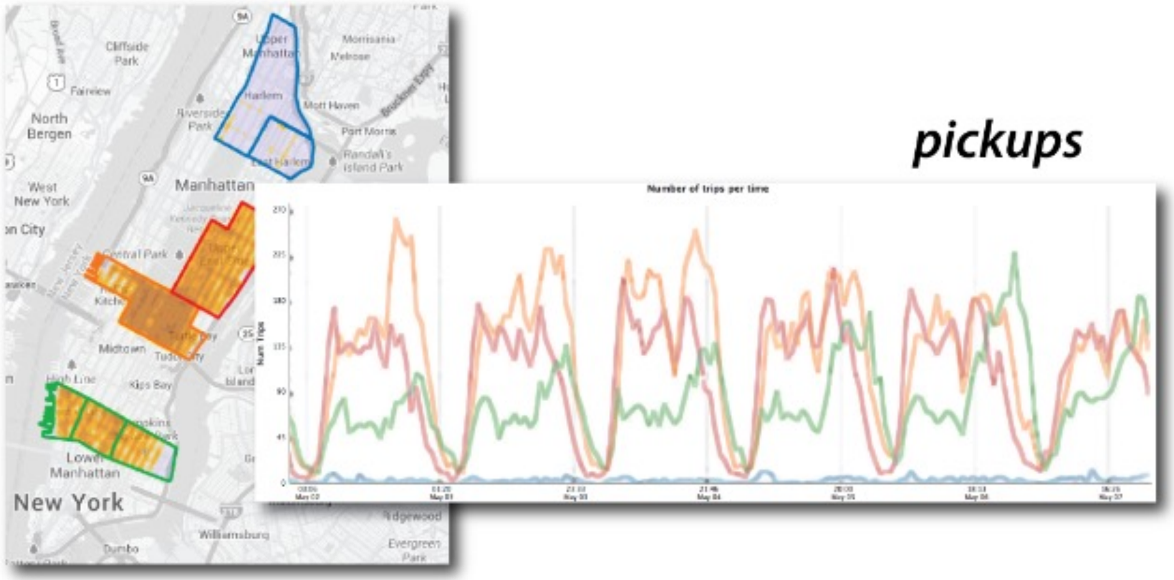
Visual query model

- Expressive Triad framework:
  - *Where + when → what*: “What is the average trip time from Midtown to the airports during weekdays?”
  - *When + what → where*: “Where are the hot spots in Manhattan in the weekends?”
  - *Where + what → when*: “When were activities restored in Lower Manhattan after the Sandy hurricane?”

Where? When?

What?

Comparison of taxi trips: characterizing neighborhoods



- Midtown and Upper East are the most active areas.
- Midtown is a popular destination in the morning and downtown in the night.
- Over the weekend, increase number of drop-offs in Downtown.
- Harlem is underserved by taxis.

Ability to quickly test hypotheses: starting with one query about a specific place and generalize to all neighborhoods.

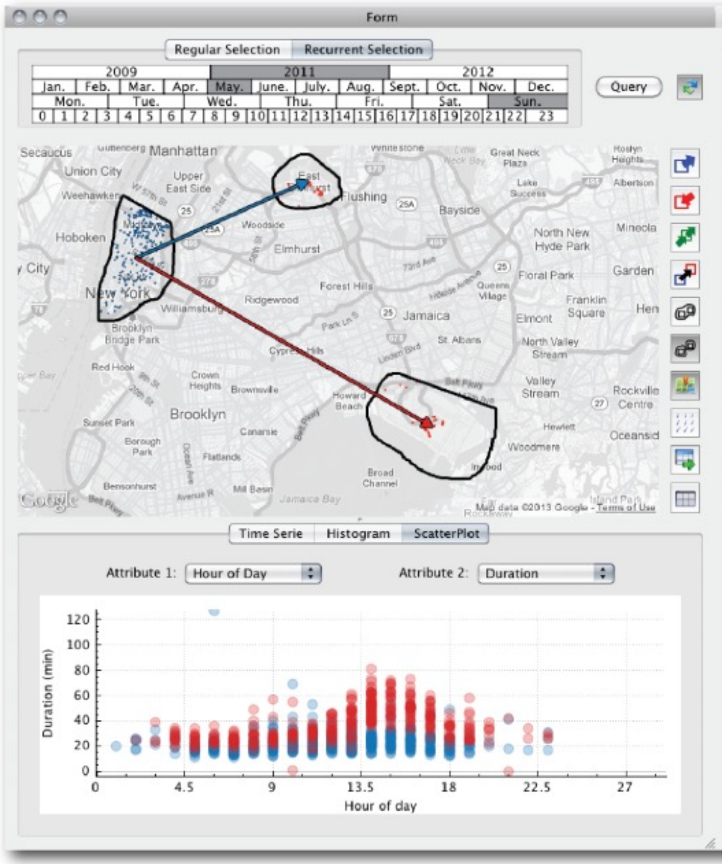
Challenge: interactive query evaluation

Typical query: find all trips that occurred between lower Manhattan and the two airports, JFK and LGA during all Sundays in May 2011.

Query time (sec)	PostgreSQL	ComDB
	503.9	20.6

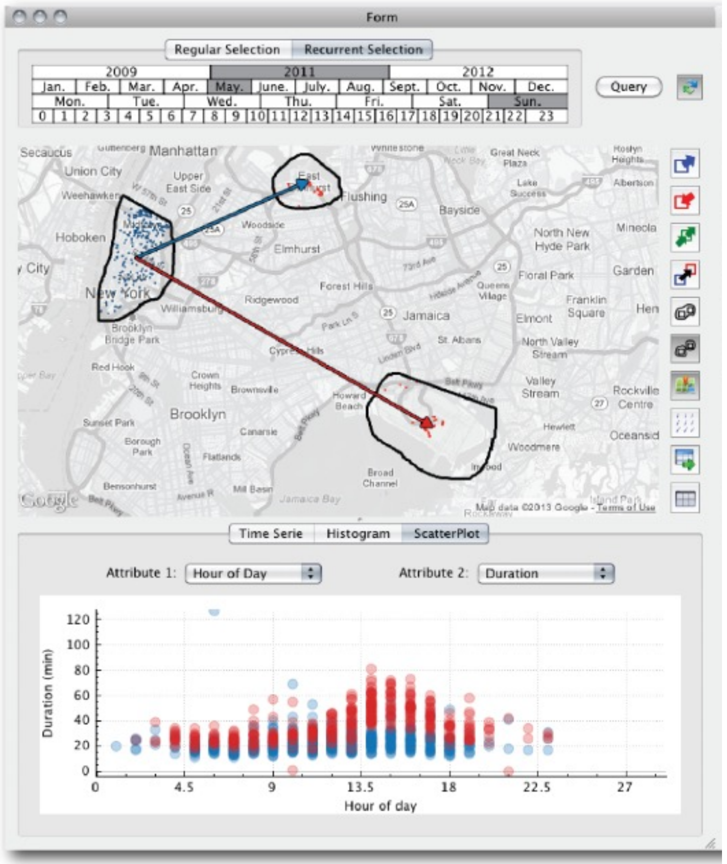
"Increased latency reduces the rate at which users make observations, draw generalizations and generate hypotheses"

[Liu and Heer, 2014]



Challenge: interactive query evaluation

Typical query: find all trips that occurred between lower Manhattan and the two airports, JFK and LGA during all Sundays in May 2011.



Query time (sec)	PostgreSQL	ComDB
	503.9	20.6

Goal: support interactive queries

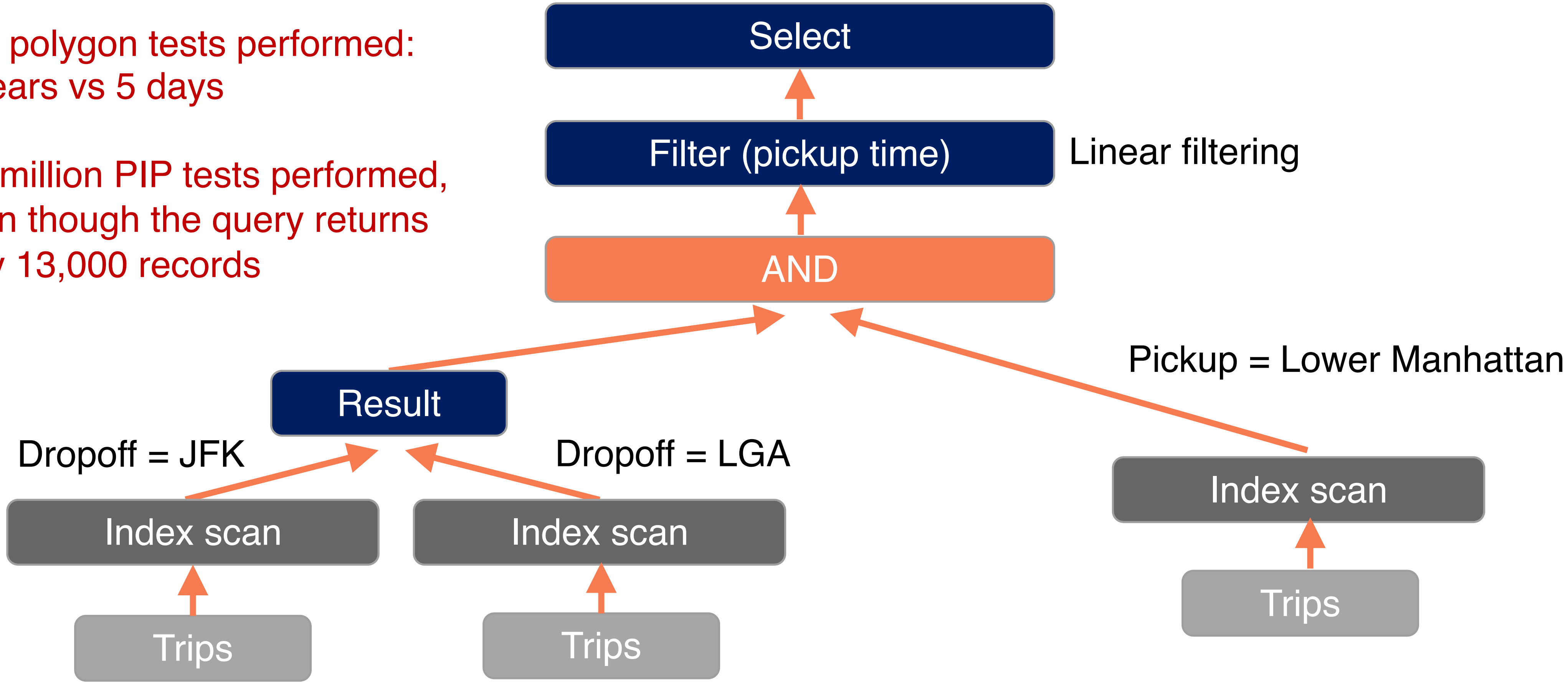
UIC COMPUTER SCIENCE

Challenge: interactive query evaluation

Typical query: find all trips that occurred between lower Manhattan and the two airports, JFK and LGA during all Sundays in May 2011.

# of polygon tests performed:
5 years vs 5 days

6.5 million PIP tests performed, even though the query returns only 13,000 records



Challenge: interactive query evaluation

SQL query

Company	Units	Value
10	2	1.34
13	5	2.00
18	9	1.90
21	1	0.37
18	5	1.46
5	3	7.30
18	2	5.30

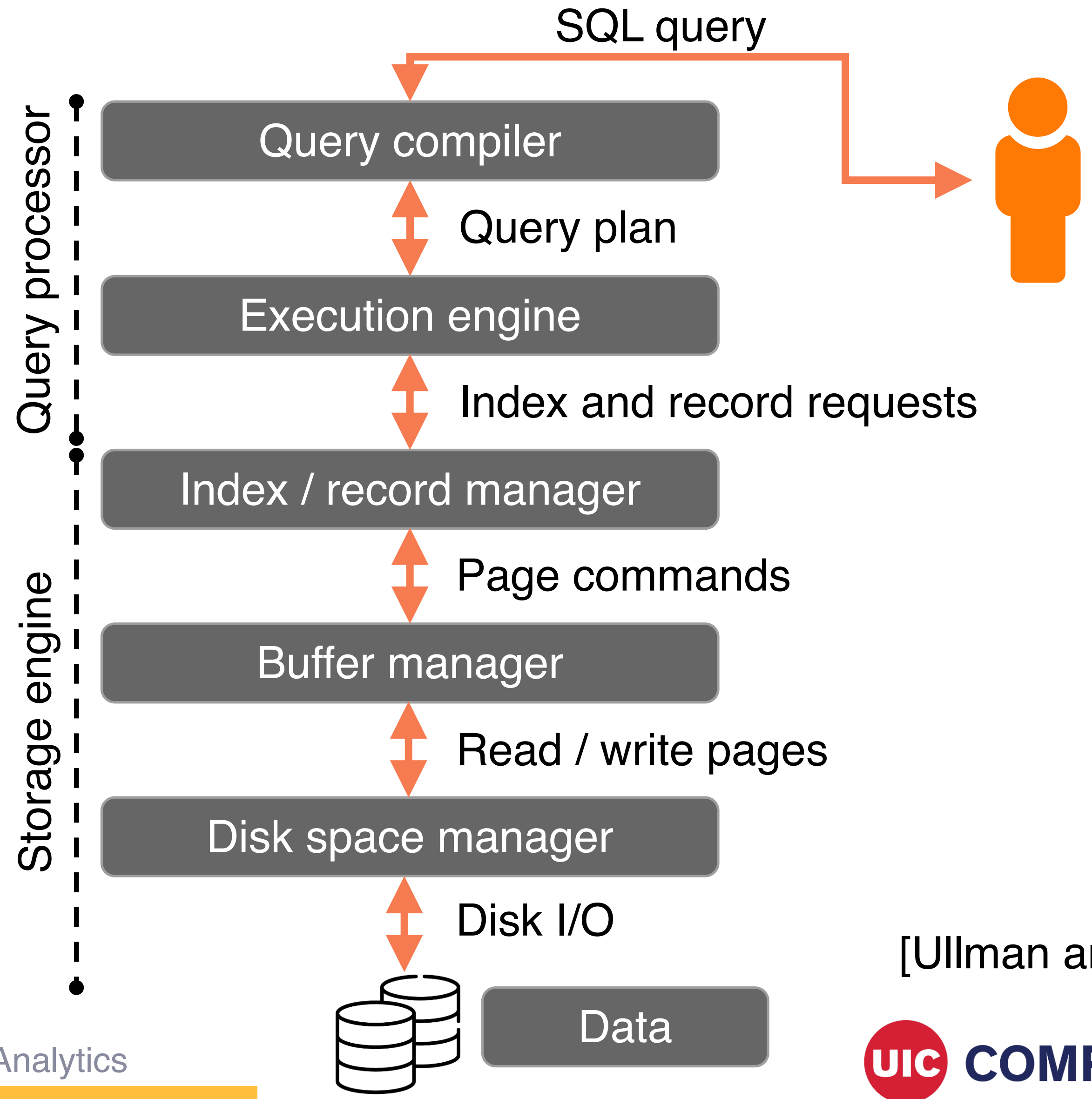
SQL query

Company	Pointer
5	5
10	0
13	1
18	2
18	4
18	6
21	3

Pointer	Company	Units	Value
0	10	2	1.34
1	13	5	2.00
2	18	9	1.90
3	21	1	0.37
4	18	5	1.46
5	5	3	7.30
6	18	2	5.30

Challenge: interactive query evaluation

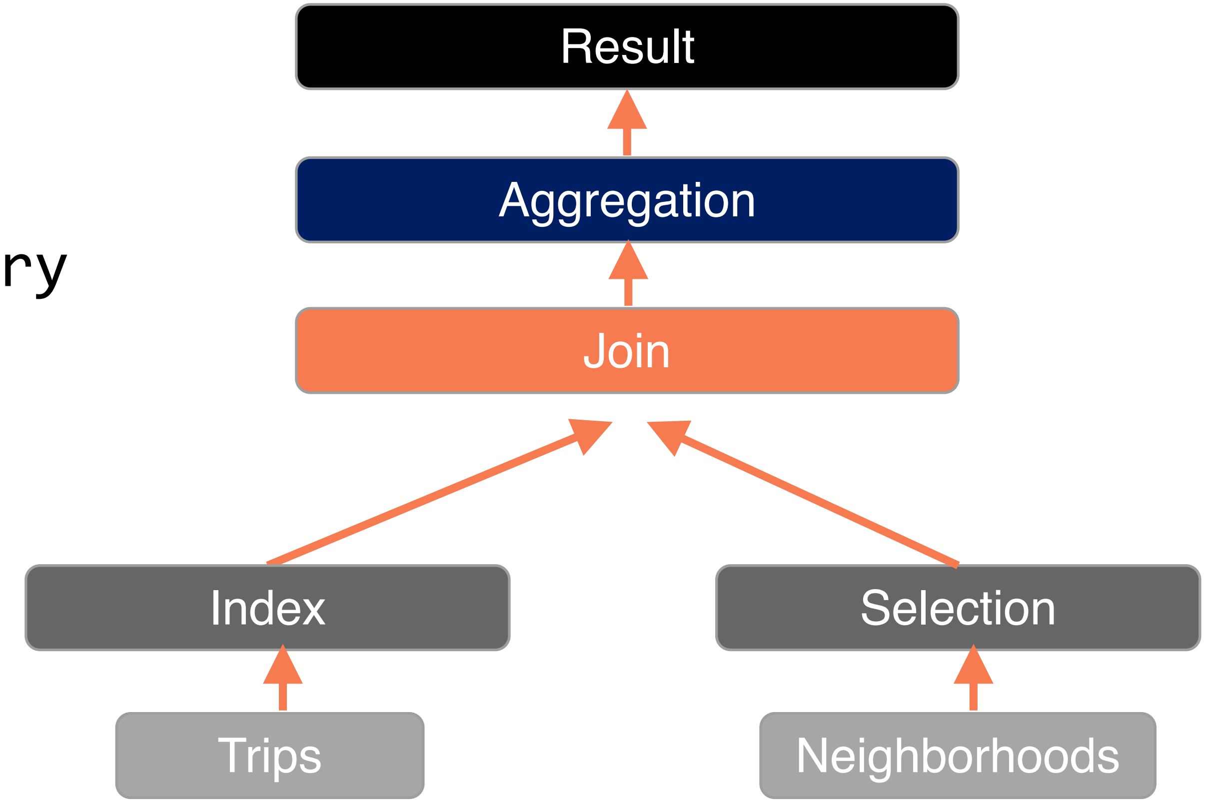
Query processor:



[Ullman and Widow]

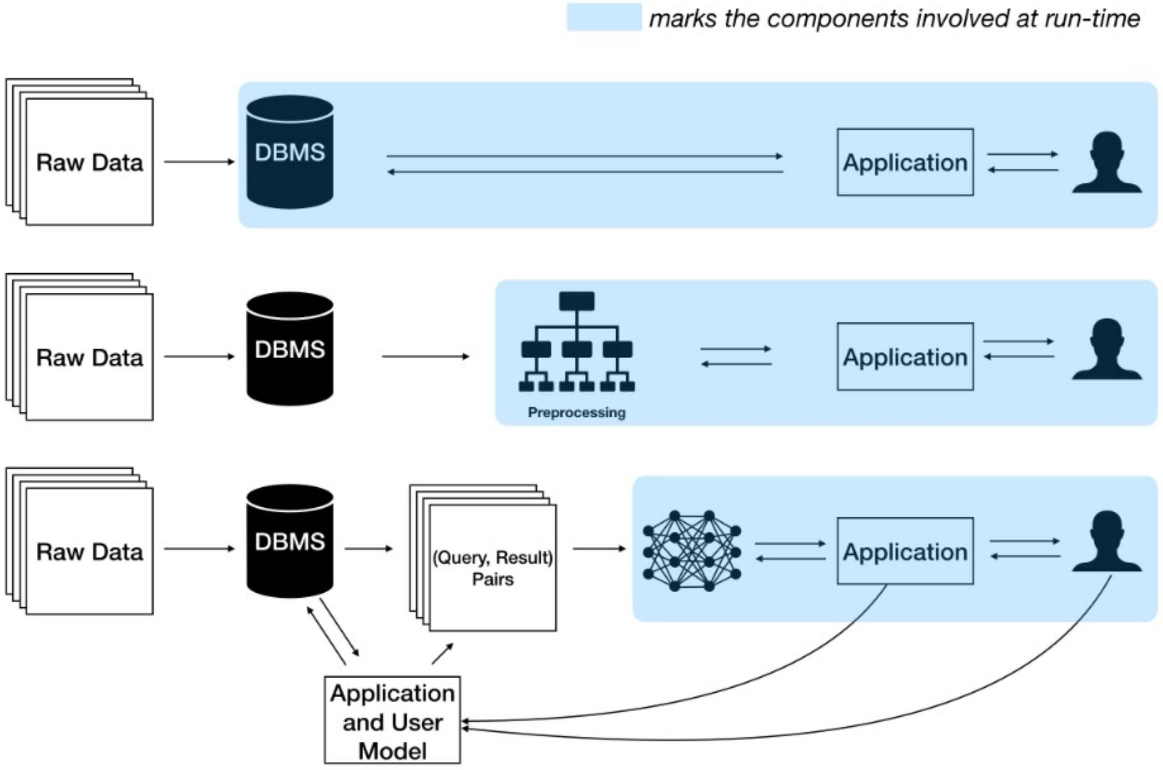
Challenge: interactive query evaluation

```
SELECT COUNT(*)
FROM taxi T, neighborhoods N
WHERE T.pickup INSIDE N.geometry
GROUP BY N.id
```



- Solution in existing spatial databases
- Slow: several seconds / minutes

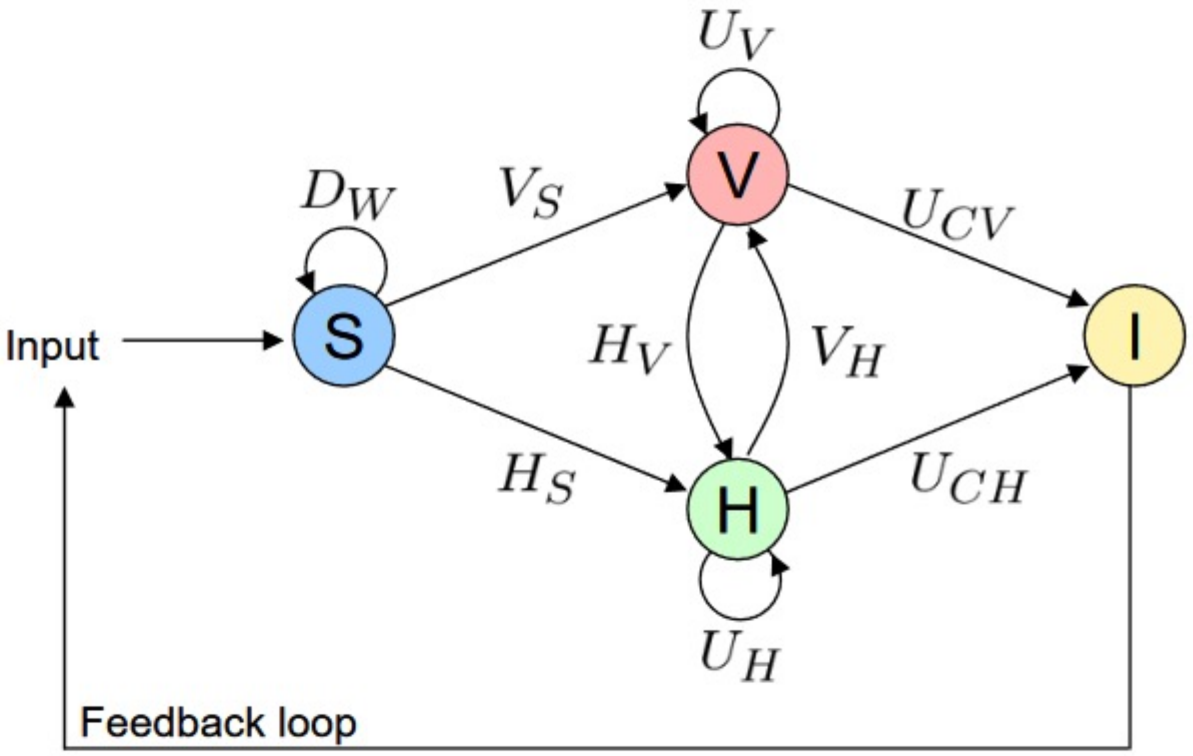
Challenge: interactive query evaluation



["NeuralCubes", Wang et al., 2019]

- Solutions:
  - Pre-compute aggregations
  - Use massively parallel architectures
  - Approximate queries
  - (all of the above)

Finding interesting features



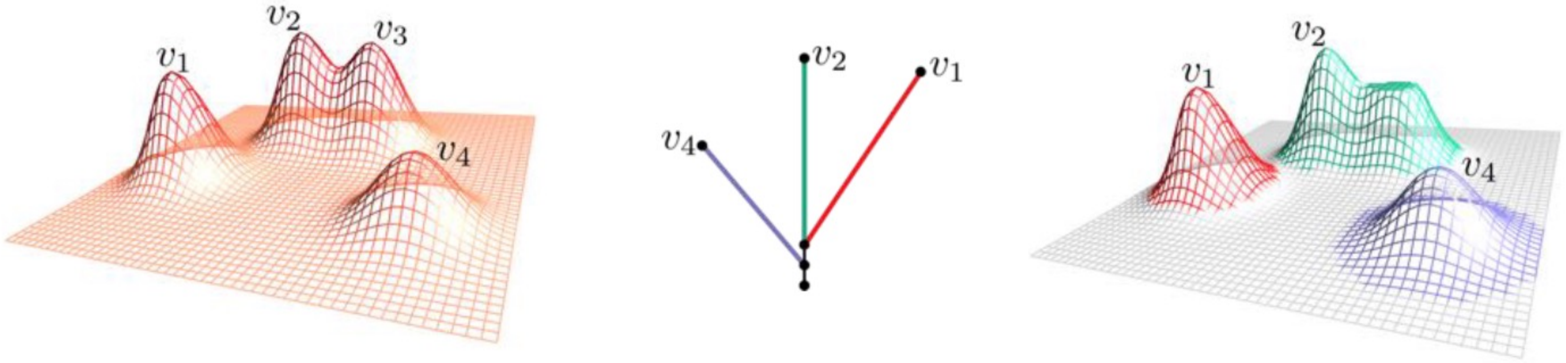
- How to tackle h_s (hypotheses from automatic methods)?
- Goal: guide users towards *interesting* data slices.

Identifying potential events

- Model data as a time-varying scalar function defined on a graph.
  - Taxi data: graph = road network; function = density of taxis
  - Subway data: graph = track network; function = delay of trains

Taxi data: potential events

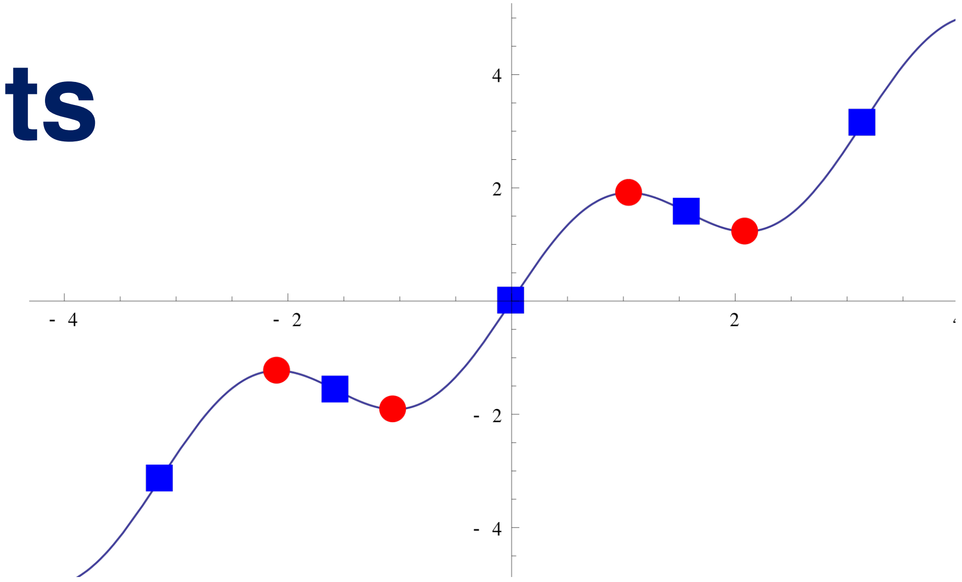
- Merge trees to efficiently identify events in each time step.
- Set of potential events: regions corresponding to maxima and minima.
    - A region is interesting if its behavior differs from its neighborhood.
    - Unimportant events can be simplified.

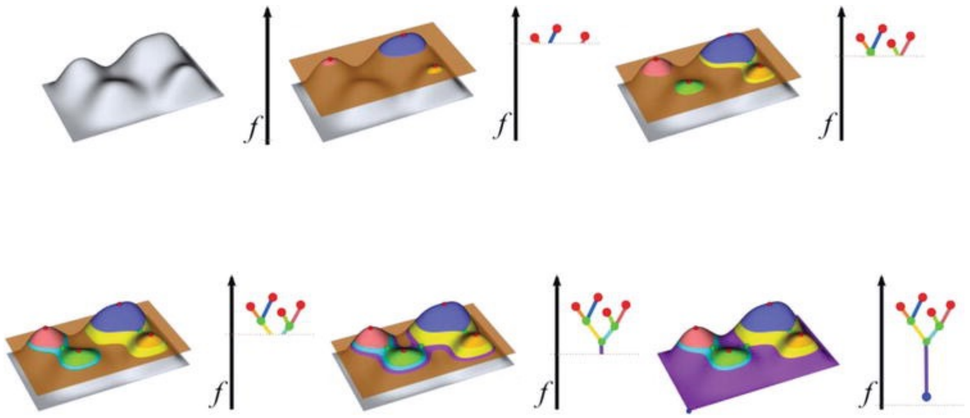


[Doraiswamy et al., 2014]

Taxi data: potential events

- Merge trees can be used to efficiently represent regions
- Topological changes occur at critical points (i.e., where the function is not differentiable or derivative equal to zero).
- Trees can be simplified to remove noise.
- Merge trees track level sets and capture local maxima as function height is reduced.

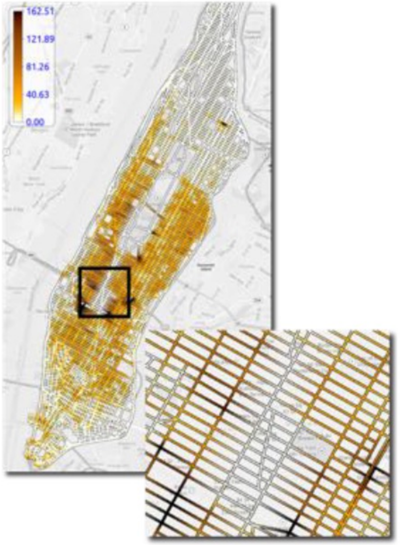




[Thomas et al., 2018]

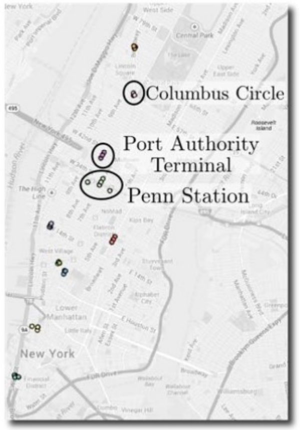
Taxi data: potential events

- Minima: lack of taxis
  - Density lower than local neighborhood.
  - Can denote road blocks.

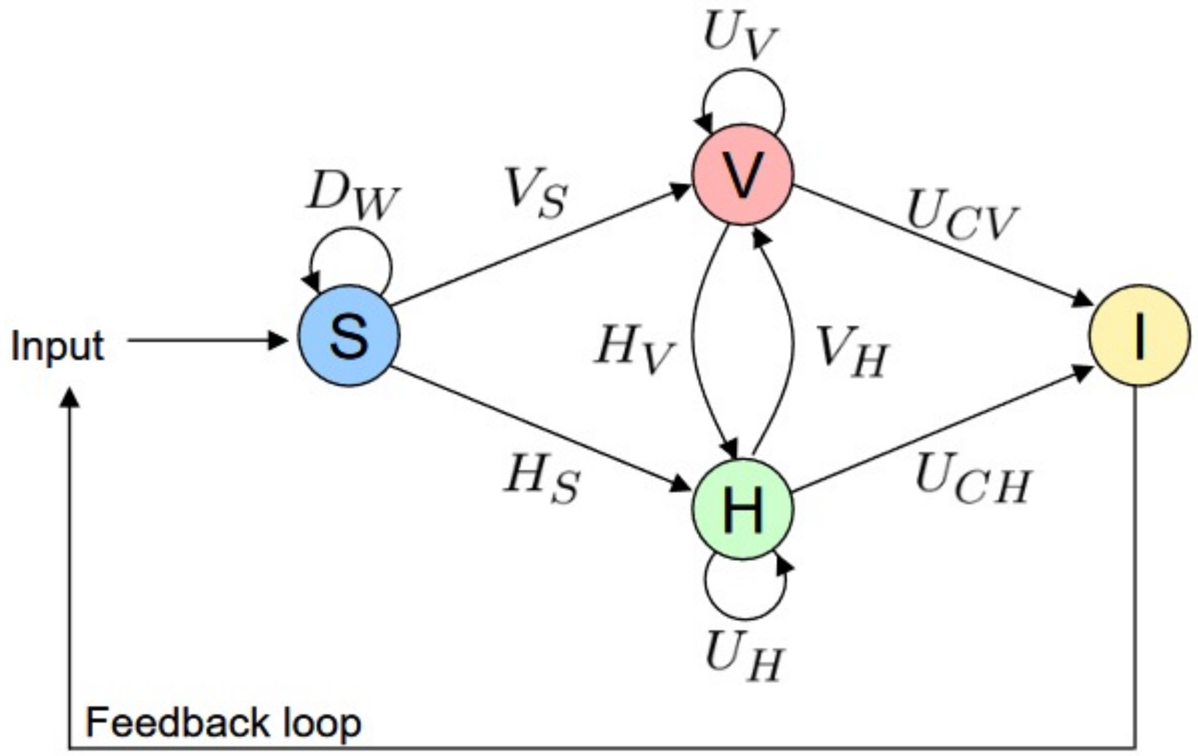


Maxima: popular taxi locations

- Density higher than local neighborhood.
- Can denote tourist locations, train stations.



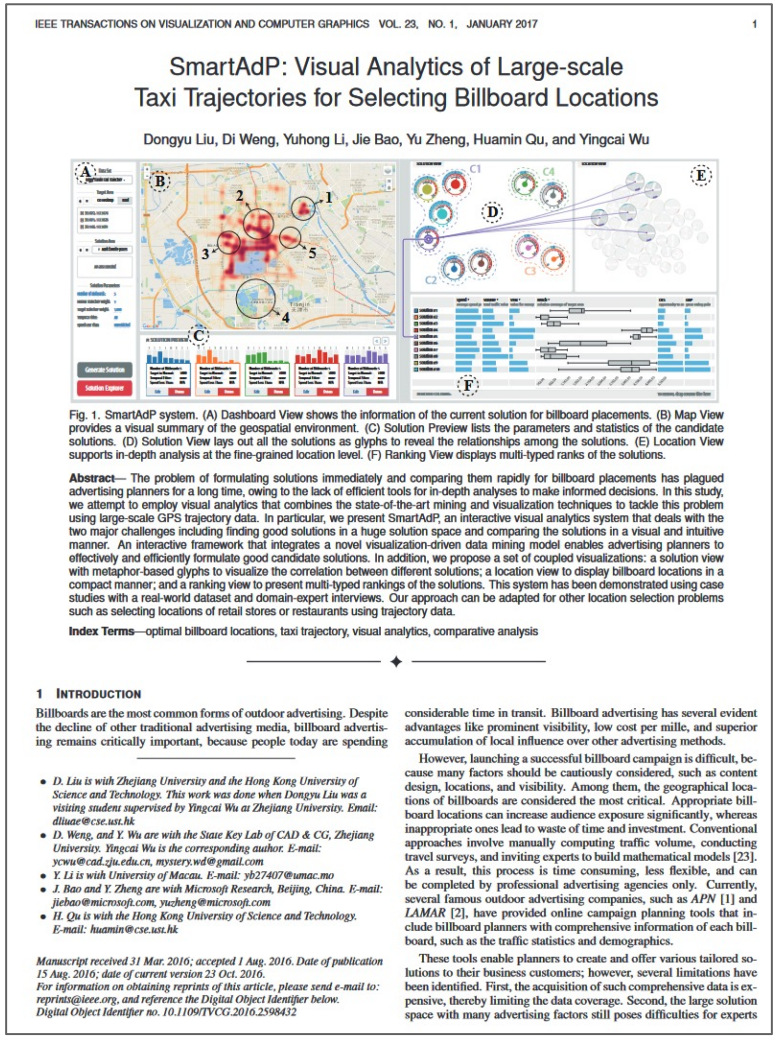
TaxiVis



- Visual analytics tool that includes:
  - h_s: hypotheses using computational topology
  - h_v: hypotheses from visualization
  - v_s: visualizing data
  - v_h: visualizing hypotheses
  - u: interaction

SmartAdP: taxi trajectories

“ … an interactive visual analytics system that deals with the two major challenges including finding good solutions in a huge solution space and comparing the solutions in a visual and intuitive manner. An interactive framework that integrates a novel visualization-driven data mining model enables advertising planners to effectively and efficiently formulate good candidate solutions. (…) This system has been demonstrated using case studies with a real-world dataset and domain-expert interviews.”



IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS VOL. 23, NO. 1, JANUARY 2017 1

SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations

Dongyu Liu, Di Weng, Yuhong Li, Jie Bao, Yu Zheng, Huamin Qu, and Yingcai Wu

Fig. 1. SmartAdP system. (A) Dashboard View shows the information of the current solution for billboard placements. (B) Map View provides a visual summary of the geospatial environment. (C) Solution Preview lists the parameters and statistics of the candidate solutions. (D) Solution View lays out all the solutions as glyphs to reveal the relationships among the solutions. (E) Location View supports in-depth analysis at the fine-grained location level. (F) Ranking View displays multi-typed ranks of the solutions.

Abstract— The problem of formulating solutions immediately and comparing them rapidly for billboard placements has plagued advertising planners for a long time, owing to the lack of efficient tools for in-depth analyses to make informed decisions. In this study, we attempt to employ visual analytics that combines the state-of-the-art mining and visualization techniques to tackle this problem using large-scale GPS trajectory data. In particular, we present SmartAdP, an interactive visual analytics system that deals with the two major challenges including finding good solutions in a huge solution space and comparing the solutions in a visual and intuitive manner. An interactive framework that integrates a novel visualization-driven data mining model enables advertising planners to effectively and efficiently formulate good candidate solutions. In addition, we propose a set of coupled visualizations: a solution view with metaphor-based glyphs to visualize the correlation between different solutions; a location view to display billboard locations in a compact manner; and a ranking view to present multi-typed rankings of the solutions. This system has been demonstrated using case studies with a real-world dataset and domain-expert interviews. Our approach can be adapted for other location selection problems such as selecting locations of retail stores or restaurants using trajectory data.

Index Terms—optimal billboard locations, taxi trajectory, visual analytics, comparative analysis

1 INTRODUCTION

Billboards are the most common forms of outdoor advertising. Despite the decline of other traditional advertising media, billboard advertising remains critically important, because people today are spending considerable time in transit. Billboard advertising has several evident advantages like prominent visibility, low cost per mille, and superior accumulation of local influence over other advertising methods.

However, launching a successful billboard campaign is difficult, because many factors should be cautiously considered, such as content design, locations, and visibility. Among them, the geographical locations of billboards are considered the most critical. Appropriate billboard locations can increase audience exposure significantly, whereas inappropriate ones lead to waste of time and investment. Conventional approaches involve manually computing traffic volume, conducting travel surveys, and inviting experts to build mathematical models [23]. As a result, this process is time consuming, less flexible, and can be completed by professional advertising agencies only. Currently, several famous outdoor advertising companies, such as *APN* [1] and *LAMAR* [2], have provided online campaign planning tools that include billboard planners with comprehensive information of each billboard, such as the traffic statistics and demographics.

These tools enable planners to create and offer various tailored solutions to their business customers; however, several limitations have been identified. First, the acquisition of such comprehensive data is expensive, thereby limiting the data coverage. Second, the large solution space with many advertising factors still poses difficulties for experts

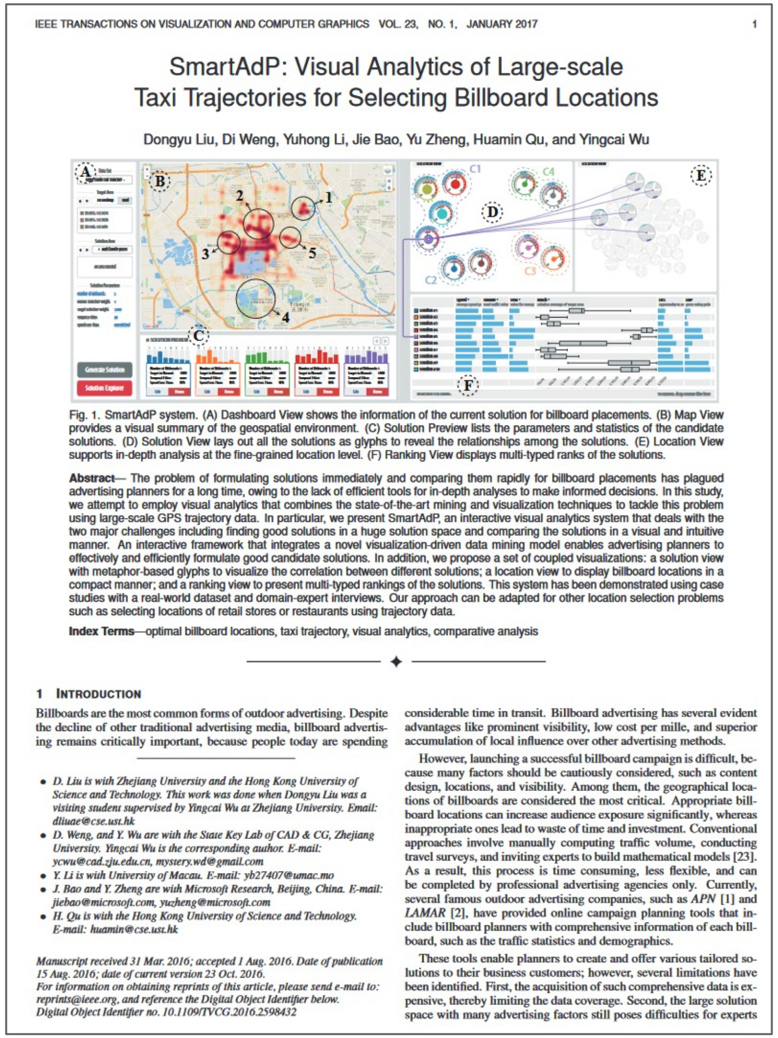
• D. Liu is with Zhejiang University and the Hong Kong University of Science and Technology. This work was done when Dongyu Liu was a visiting student supervised by Yingcai Wu at Zhejiang University. Email: dliuae@cse.ust.hk
• D. Weng, and Y. Wu are with the State Key Lab of CAD & CG, Zhejiang University. Yingcai Wu is the corresponding author. E-mail: ycwu@cad.zju.edu.cn, mystery.wd@gmail.com
• Y. Li is with University of Macau. E-mail: yb27407@umac.mo
• J. Bao and Y. Zheng are with Microsoft Research, Beijing, China. E-mail: jiebao@microsoft.com, yuzheng@microsoft.com
• H. Qu is with the Hong Kong University of Science and Technology. E-mail: huamin@cse.ust.hk

Manuscript received 31 Mar. 2016; accepted 1 Aug. 2016. Date of publication 15 Aug. 2016; date of current version 23 Oct. 2016. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TVCG.2016.2598432

[Liu et al., 2017]

SmartAdP: taxi trajectories

- Characterization of the problem of billboard location selection using taxi trajectory data.
- Interactive framework to generate billboard solutions with novel visualization-driven data mining model and application-specific data index mechanism.



IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS VOL. 23, NO. 1, JANUARY 2017 1

SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations

Dongyu Liu, Di Weng, Yuhong Li, Jie Bao, Yu Zheng, Huamin Qu, and Yingcai Wu

Fig. 1. SmartAdP system. (A) Dashboard View shows the information of the current solution for billboard placements. (B) Map View provides a visual summary of the geospatial environment. (C) Solution Preview lists the parameters and statistics of the candidate solutions. (D) Solution View lays out all the solutions as glyphs to reveal the relationships among the solutions. (E) Location View supports in-depth analysis at the fine-grained location level. (F) Ranking View displays multi-typed ranks of the solutions.

Abstract— The problem of formulating solutions immediately and comparing them rapidly for billboard placements has plagued advertising planners for a long time, owing to the lack of efficient tools for in-depth analyses to make informed decisions. In this study, we attempt to employ visual analytics that combines the state-of-the-art mining and visualization techniques to tackle this problem using large-scale GPS trajectory data. In particular, we present SmartAdP, an interactive visual analytics system that deals with the two major challenges including finding good solutions in a huge solution space and comparing the solutions in a visual and intuitive manner. An interactive framework that integrates a novel visualization-driven data mining model enables advertising planners to effectively and efficiently formulate good candidate solutions. In addition, we propose a set of coupled visualizations: a solution view with metaphor-based glyphs to visualize the correlation between different solutions; a location view to display billboard locations in a compact manner; and a ranking view to present multi-typed rankings of the solutions. This system has been demonstrated using case studies with a real-world dataset and domain-expert interviews. Our approach can be adapted for other location selection problems such as selecting locations of retail stores or restaurants using trajectory data.

Index Terms—optimal billboard locations, taxi trajectory, visual analytics, comparative analysis

1 INTRODUCTION

Billboards are the most common forms of outdoor advertising. Despite the decline of other traditional advertising media, billboard advertising remains critically important, because people today are spending considerable time in transit. Billboard advertising has several evident advantages like prominent visibility, low cost per mille, and superior accumulation of local influence over other advertising methods.

However, launching a successful billboard campaign is difficult, because many factors should be cautiously considered, such as content design, locations, and visibility. Among them, the geographical locations of billboards are considered the most critical. Appropriate billboard locations can increase audience exposure significantly, whereas inappropriate ones lead to waste of time and investment. Conventional approaches involve manually computing traffic volume, conducting travel surveys, and inviting experts to build mathematical models [23]. As a result, this process is time consuming, less flexible, and can be completed by professional advertising agencies only. Currently, several famous outdoor advertising companies, such as *APN* [1] and *LAMAR* [2], have provided online campaign planning tools that include billboard planners with comprehensive information of each billboard, such as the traffic statistics and demographics.

These tools enable planners to create and offer various tailored solutions to their business customers; however, several limitations have been identified. First, the acquisition of such comprehensive data is expensive, thereby limiting the data coverage. Second, the large solution space with many advertising factors still poses difficulties for experts

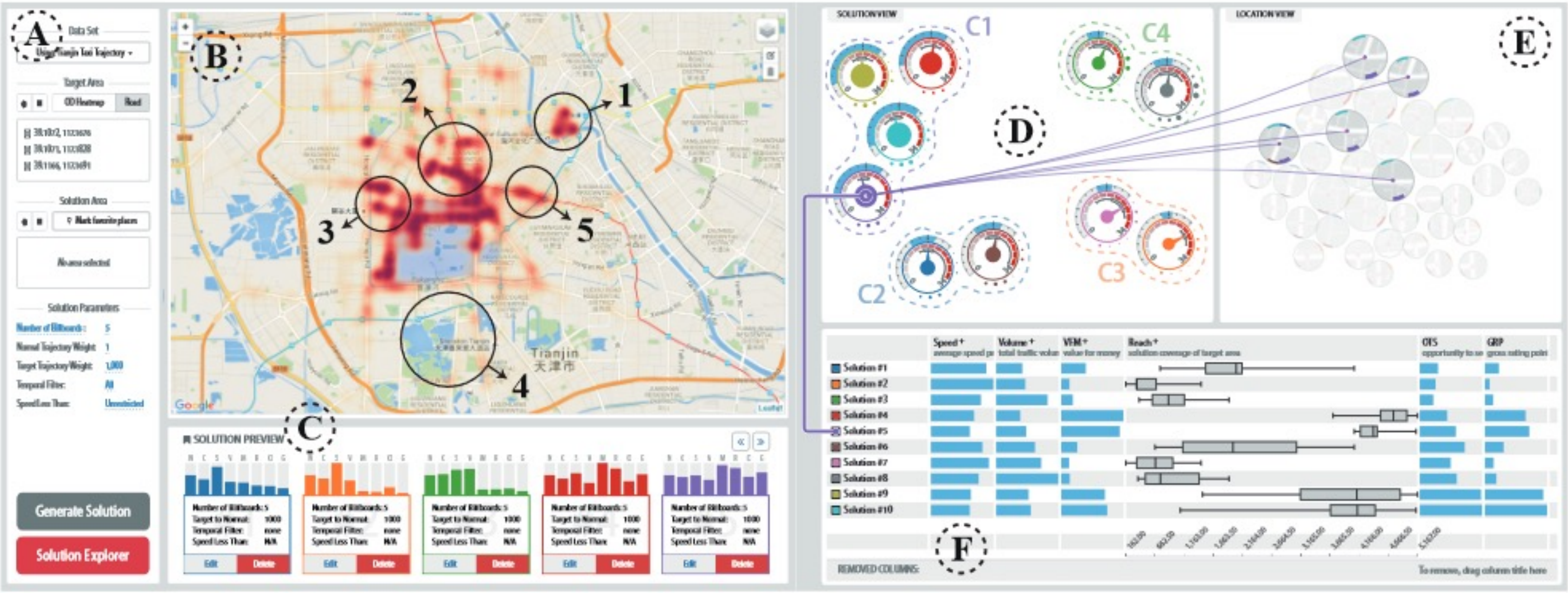
- *D. Liu is with Zhejiang University and the Hong Kong University of Science and Technology. This work was done when Dongyu Liu was a visiting student supervised by Yingcai Wu at Zhejiang University. Email: dliuae@cse.ust.hk*
- *D. Weng, and Y. Wu are with the State Key Lab of CAD & CG, Zhejiang University. Yingcai Wu is the corresponding author. E-mail: ycwu@cad.zju.edu.cn, mystery.wd@gmail.com*
- *Y. Li is with University of Macau. E-mail: yb27407@umac.mo*
- *J. Bao and Y. Zheng are with Microsoft Research, Beijing, China. E-mail: jiebao@microsoft.com, yuzheng@microsoft.com*
- *H. Qu is with the Hong Kong University of Science and Technology. E-mail: huamin@cse.ust.hk*

Manuscript received 31 Mar. 2016; accepted 1 Aug. 2016. Date of publication 15 Aug. 2016; date of current version 23 Oct. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2598432

[Liu et al., 2017]

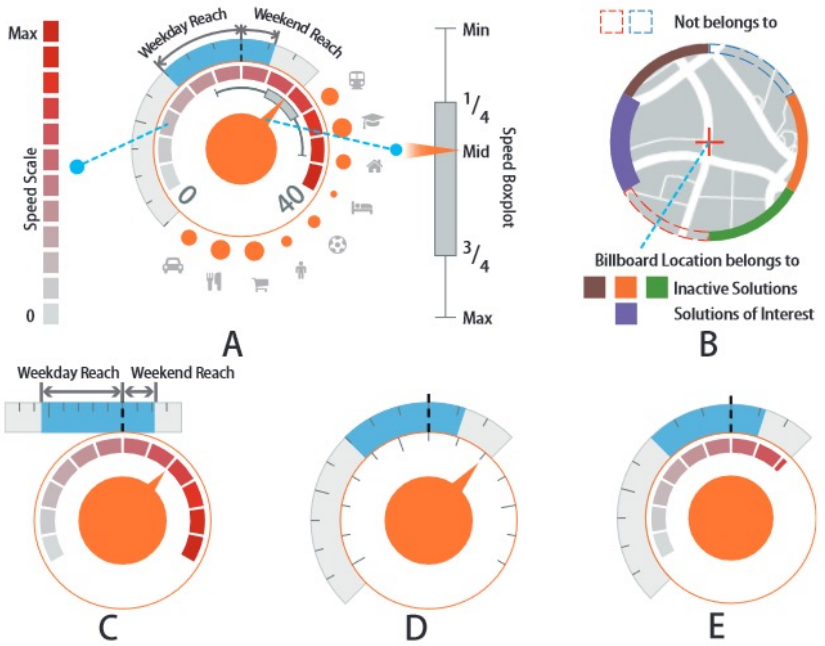
Billboard location

- Billboard location is a decisive factor for a billboard campaign. How to select the best locations (i.e., best intersections)?



Billboard location

- Data:
  - Traffic volume, traffic speed, traffic OD, environment, cost of billboard.
- Indicators:
  - *Coverage / reach*: percentage of covered trajectories among all trajectories.
  - *Opportunities to see*: average number of billboard contacts.
  - *Gross rating points*: average number of contacts that 100 target trajectories produce.
  - *Value for money*: value of the cost.



SmartAdP data structure

- K-location query:
  - Given a set of candidate vertices and target OD, select k locations.
  - Step 1: select trajectories with origin and destination within target OD.
    - All trajectories going from A to B.
  - Step 2: for each trajectory, update vertex coverage if it goes through vertex.
    - Does current trajectory go through candidate vertex? Update its coverage.

SmartAdP data structure

- How to support these operations?
  - We need to find trajectories within target OD (i.e., *vertex ⟶ trajectory).*
  - We need to find where a trajectory goes through (i.e., *trajectory ⟶ vertex*).
- MongoDB indices:
  - Trajectory-edge index: trajectory spatial points mapped to road network.
    Trajectory ⟶ road segment
  - Trajectory-vertex index: covered vertices of trajectories.
    Trajectory ⟶ vertices
  - Vertex-trajectory index: covered trajectories of vertices.
    Vertices ⟶ trajectories

Urban Mosaic: street-level images

" … a tool for exploring the urban fabric through a spatially and temporally dense data set of 7.7 million street-level images from New York City, captured over the period of a year. Working in collaboration with professional practitioners, we use Urban Mosaic to investigate questions of accessibility and mobility, and preservation and retrofitting."

Urban Mosaic: Visual Exploration of Streetscapes Using Large-Scale Image Data

Fabio Miranda[1], Maryam Hosseini[2], Marcos Lage[3], Harish Doraiswamy[1], Graham Dove[1], Cláudio T. Silva[1]

[1]New York University; [2]Rutgers University; [3]Universidade Federal Fluminense

[1]{fmiranda,harishd,grahamdove,csilva}@nyu.edu; [2]mary.hosseini@rutgers.edu; [3]mlage@ic.uff.br

ABSTRACT

Urban planning is increasingly data driven, yet the challenge of designing with data at a city scale and remaining sensitive to the impact at a human scale is as important today as it was for Jane Jacobs. We address this challenge with Urban Mosaic, a tool for exploring the urban fabric through a spatially and temporally dense data set of 7.7 million street-level images from New York City, captured over the period of a year. Working in collaboration with professional practitioners, we use Urban Mosaic to investigate questions of accessibility and mobility, and preservation and retrofitting. In doing so, we demonstrate how tools such as this might provide a bridge between the city and the street, by supporting activities such as visual comparison of geographically distant neighborhoods, and temporal analysis of unfolding urban development.

Author Keywords

Urban planning; Interactive visualization; Data analysis; Urban data

CCS Concepts

•**Human-centered computing → Human computer interaction (HCI);** *Visual analytics; Visualization toolkits;*

"A sense of place is built up, in the end, from many little things too, some so small people take them for granted, and yet the lack of them takes the flavor out of the city."

(Jane Jacobs, Downtown is for People)

INTRODUCTION

For those of us living in or visiting the world's major cities, their dynamism and complexity are immediately apparent. Yet urban planners and designers must work in a context where any single intervention, perhaps aimed at altering just one aspect, can have a wide ranging impact on a variety of interrelated components [12], affecting things at both a macro *city scale* and a micro *human scale*. Examples of changes at a city scale include, suburbanization, economic deconcentration, modification to transport infrastructure, rezoning and/or gentrification of neighborhoods, and major renewal projects. Examples of changes at a human scale, on the other hand, are reflected in the city's urban fabric, and include aspects that make a city livable, encourage walking, and contribute to the perception of safety; in other words affect the day to day lives of its inhabitants. This might be manifested in lighting, shadow, sky exposure, open-front shops, the details of building facades, etc. [32].

Responding to these challenges at the city scale, planners and designers have been aided by a rapid growth in data from urban environments, and so are able to turn to computational methods and large-scale data analysis, which increase understanding by quantifying different aspects of the city (e.g., [36, 13, 14, 29, 70, 67, 48]). However, while sensitivity to the impact of change at a human scale remains as important today as it was for Jane Jacobs and others in the 1950s and 1960s [47], analyses of suitable data, which emphasize qualitative, visual details, are often difficult and time-consuming to perform. Although there is an increasing availability of street-view images, which support a degree of virtual assessment and auditing of the built environment, their distribution is often temporally sparse and so analysis is limited.

This paper introduces Urban Mosaic, a tool for visually exploring the urban fabric. It responds to the challenges practitioners face by employing a newly available spatially and temporally dense data set of street-level images from New York City (NYC) Urban Mosaic is a visual exploration system designed to help practitioners in urban planning and design gain insight into the human scale impact of changes in the urban fabric. It utilizes state-of-the-art computer vision techniques for image similarity search and clustering, together with efficient spatio-temporal selection and aggregation over the image metadata to visually explore and map this image data set. It further allows the analysis to be augmented using spatio-temporal urban data from a variety of other sources (e.g., census, transport, crime, weather, housing market, zoning, noise complaints). The image data set used in this work contains 7.7 million images captured in the Manhattan and Brooklyn boroughs between April 2016 and April 2017 using car-mounted cameras. Urban Mosaic has been developed as a collaboration between researchers in urban planning, visual analytics, and HCI. We include a detailed

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.

http://dx.doi.org/10.1145/3313831.3376399

[Miranda et al., 2020]

Urban Mosaic: street-level images

- Demonstrates the potential for images from large-scale street-view data sets to help bridge data-driven urban planning and design at city scale and at the human level.
- Exploration of the urban fabric using large spatiotemporal collection of images.
- Visual comparison of geographically distant areas to analyze unfolding urban developments.

Urban Mosaic: Visual Exploration of Streetscapes Using Large-Scale Image Data

Fabio Miranda[1], Maryam Hosseini[2], Marcos Lage[3], Harish Doraiswamy[1], Graham Dove[1], Cláudio T. Silva[1]
[1]New York University; [2]Rutgers University; [3]Universidade Federal Fluminense
[1]{fmiranda,harishd,grahamdove,csilva}@nyu.edu; [2]mary.hosseini@rutgers.edu; [3]mlage@ic.uff.br

ABSTRACT
Urban planning is increasingly data driven, yet the challenge of designing with data at a city scale and remaining sensitive to the impact at a human scale is as important today as it was for Jane Jacobs. We address this challenge with Urban Mosaic, a tool for exploring the urban fabric through a spatially and temporally dense data set of 7.7 million street-level images from New York City, captured over the period of a year. Working in collaboration with professional practitioners, we use Urban Mosaic to investigate questions of accessibility and mobility, and preservation and retrofitting. In doing so, we demonstrate how tools such as this might provide a bridge between the city and the street, by supporting activities such as visual comparison of geographically distant neighborhoods, and temporal analysis of unfolding urban development.

Author Keywords
Urban planning; Interactive visualization; Data analysis; Urban data

CCS Concepts
•**Human-centered computing → Human computer interaction (HCI);** *Visual analytics; Visualization toolkits;*

"A sense of place is built up, in the end, from many little things too, some so small people take them for granted, and yet the lack of them takes the flavor out of the city."

(Jane Jacobs, Downtown is for People)

INTRODUCTION

For those of us living in or visiting the world's major cities, their dynamism and complexity are immediately apparent. Yet urban planners and designers must work in a context where any single intervention, perhaps aimed at altering just one aspect, can have a wide ranging impact on a variety of interrelated components [12], affecting things at both a macro *city scale* and a micro *human scale*. Examples of changes at a city scale include, suburbanization, economic deconcentration, modification to transport infrastructure, rezoning and/or gentrification of neighborhoods, and major renewal projects. Examples of changes at a human scale, on the other hand, are reflected in the city's urban fabric, and include aspects that make a city livable, encourage walking, and contribute to the perception of safety; in other words affect the day to day lives of its inhabitants. This might be manifested in lighting, shadow, sky exposure, open-front shops, the details of building facades, etc. [32].

Responding to these challenges at the city scale, planners and designers have been aided by a rapid growth in data from urban environments, and so are able to turn to computational methods and large-scale data analysis, which increase understanding by quantifying different aspects of the city (e.g., [36, 13, 14, 29, 70, 67, 48]). However, while sensitivity to the impact of change at a human scale remains as important today as it was for Jane Jacobs and others in the 1950s and 1960s [47], analyses of suitable data, which emphasize qualitative, visual details, are often difficult and time-consuming to perform. Although there is an increasing availability of street-view images, which support a degree of virtual assessment and auditing of the built environment, their distribution is often temporally sparse and so analysis is limited.

This paper introduces Urban Mosaic, a tool for visually exploring the urban fabric. It responds to the challenges practitioners face by employing a newly available spatially and temporally dense data set of street-level images from New York City (NYC) Urban Mosaic is a visual exploration system designed to help practitioners in urban planning and design gain insight into the human scale impact of changes in the urban fabric. It utilizes state-of-the-art computer vision techniques for image similarity search and clustering, together with efficient spatio-temporal selection and aggregation over the image metadata to visually explore and map this image data set. It further allows the analysis to be augmented using spatio-temporal urban data from a variety of other sources (e.g., census, transport, crime, weather, housing market, zoning, noise complaints). The image data set used in this work contains 7.7 million images captured in the Manhattan and Brooklyn boroughs between April 2016 and April 2017 using car-mounted cameras. Urban Mosaic has been developed as a collaboration between researchers in urban planning, visual analytics, and HCI. We include a detailed

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376399

[Miranda et al., 2020]

Temporally-dense street-level images

Objectives

Support the interactive analysis of the city at the micro scale, over geographically distant regions.

Comparison of the urban fabric in different regions

Assessment of features in the built environment

Assessment of walkability and accessibility

Image query composition

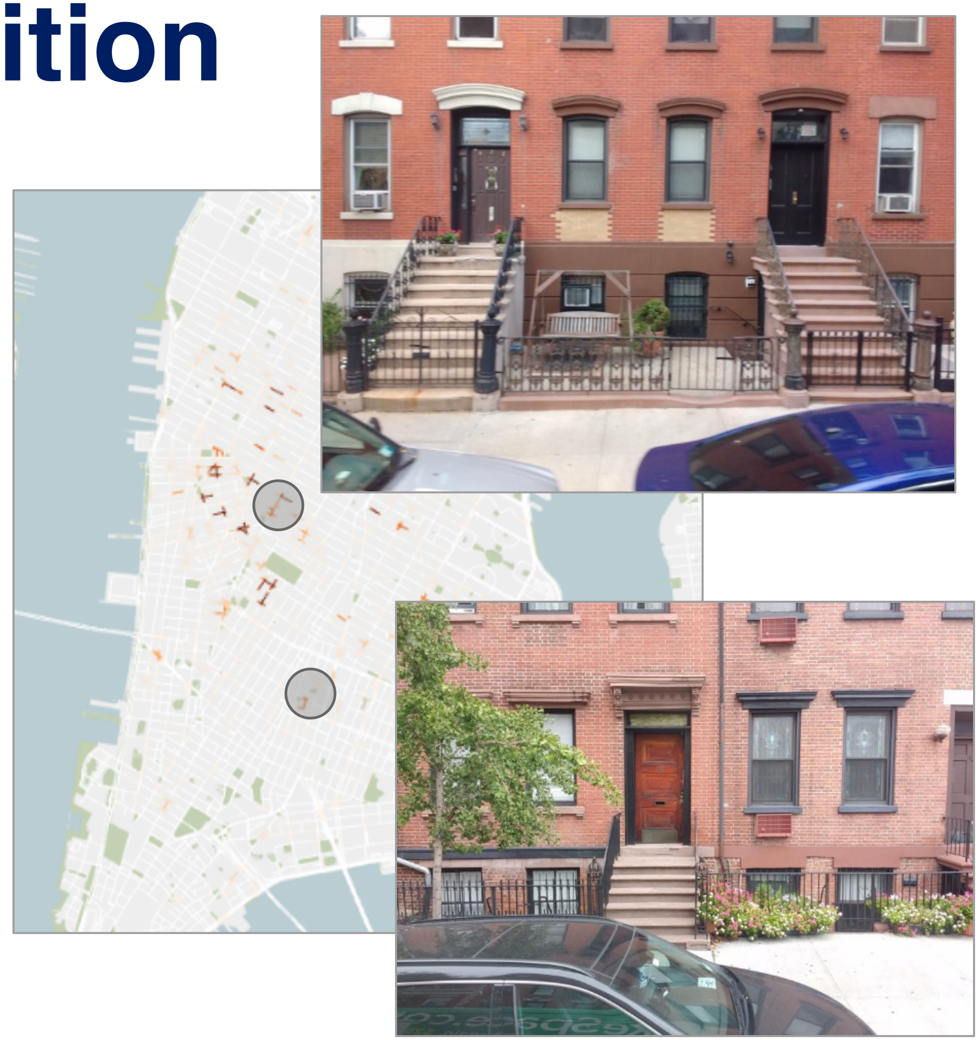
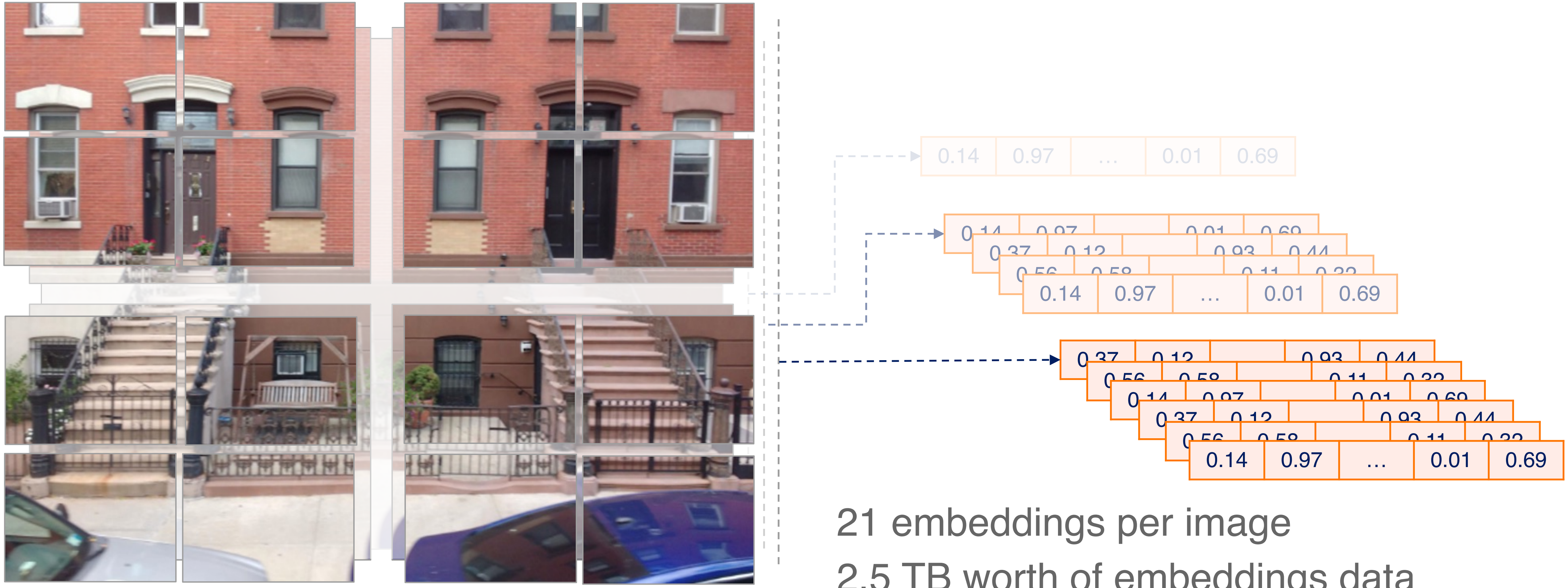


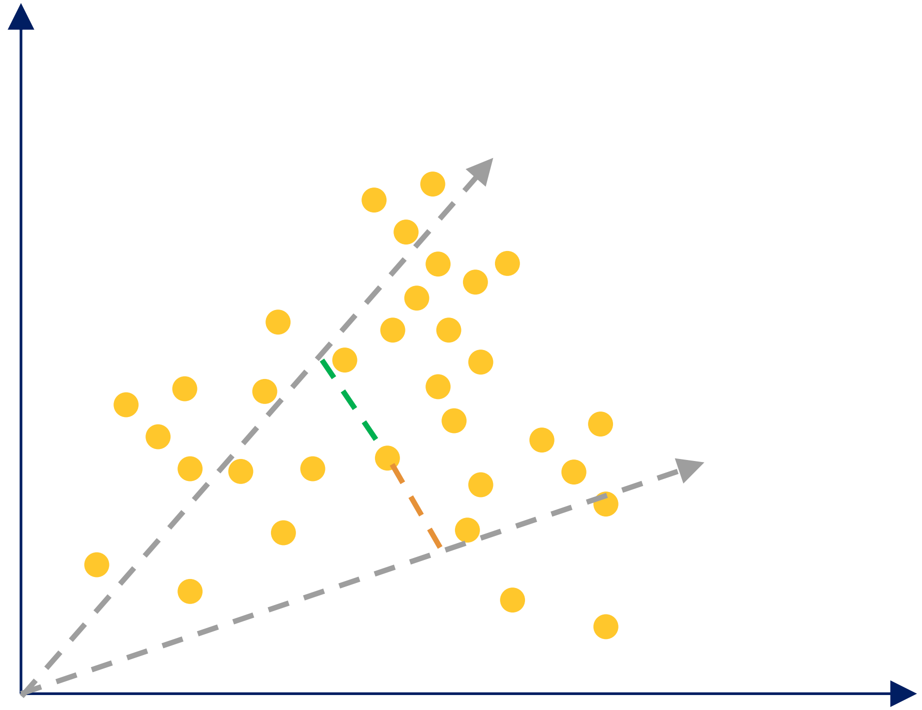
Image embeddings



21 embeddings per image

2.5 TB worth of embeddings data

Locality sensitive hashing



0.37	0.12	...	0.93	0.44

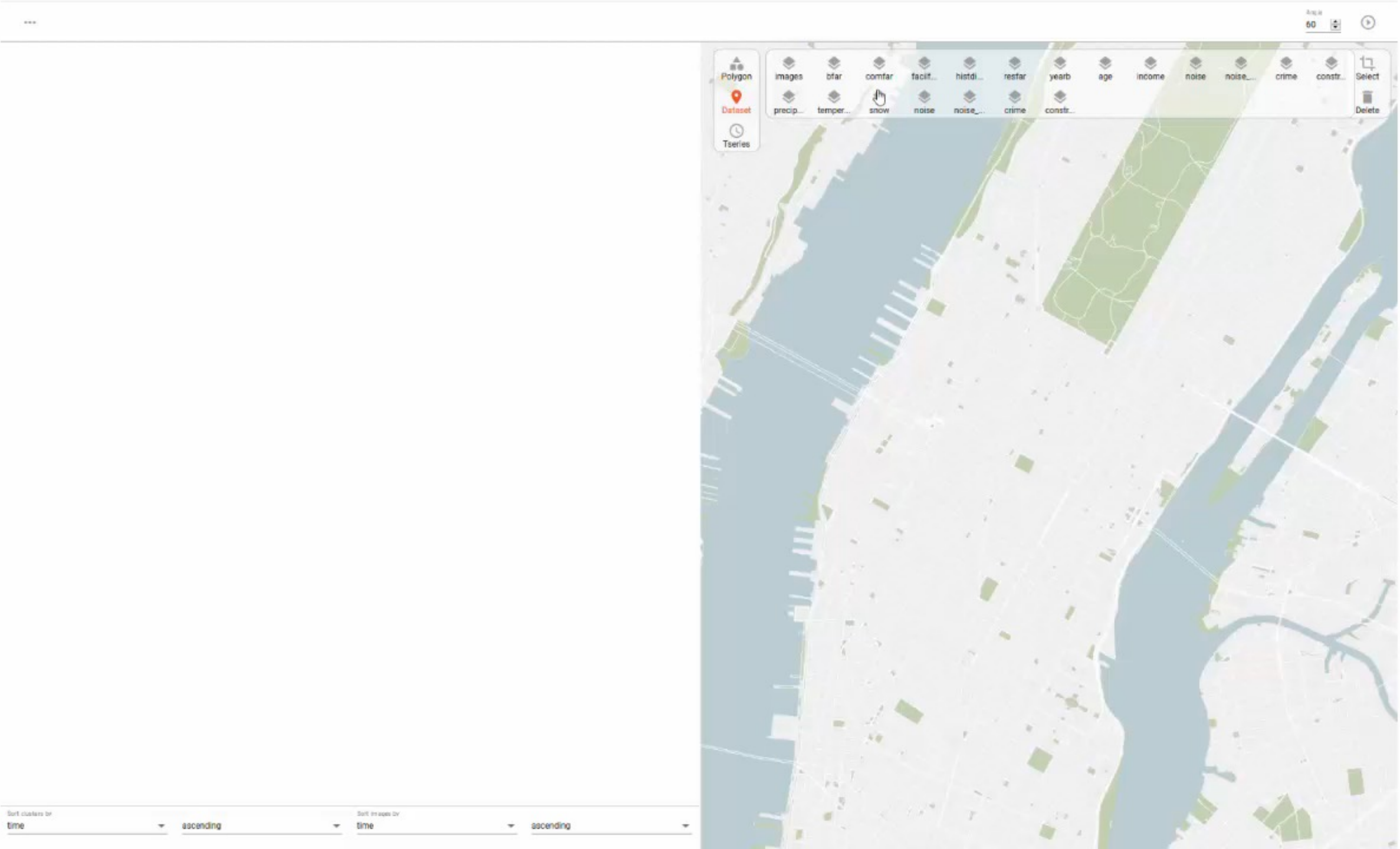
4 kbytes per embedding

0	0	...	1	0

64 bytes per embedding

$$\alpha_{1,2} = \cos^{-1}\left(\frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{|\overrightarrow{v_1}||\overrightarrow{v_2}|}\right)$$

21 GB worth of embeddings data



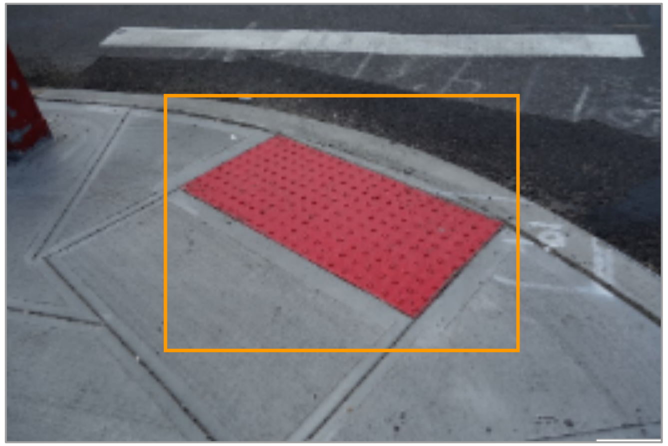
Angle
60
Polygon
Dataset
Tseries
images
bfar
comfar
facilf...
histdi...
resfar
yearb
age
income
noise
noise_...
crime
constr...
precip...
temper...
snow
noise
noise_...
crime
constr...
Select
Delete
Sort clusters by
time
ascending
Sort images by
time
ascending

Accessibility for older adults

Focuses on interventions that enable older adults to “Age in Place” and prevent outdoor fall.

Impact of the neighborhood environment on falls.

Accessibility: installation of tactile pavings

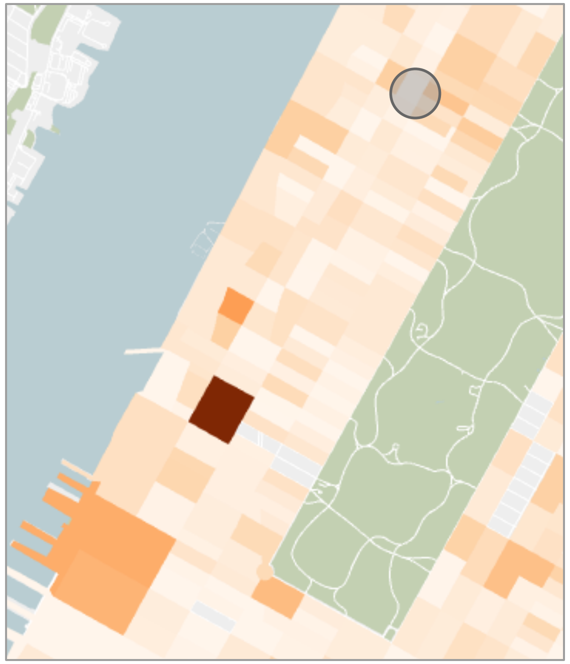


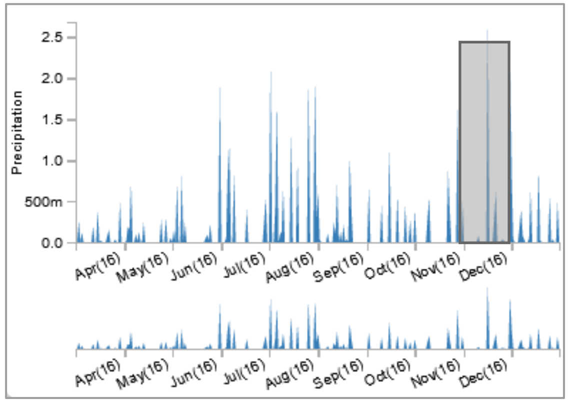
June

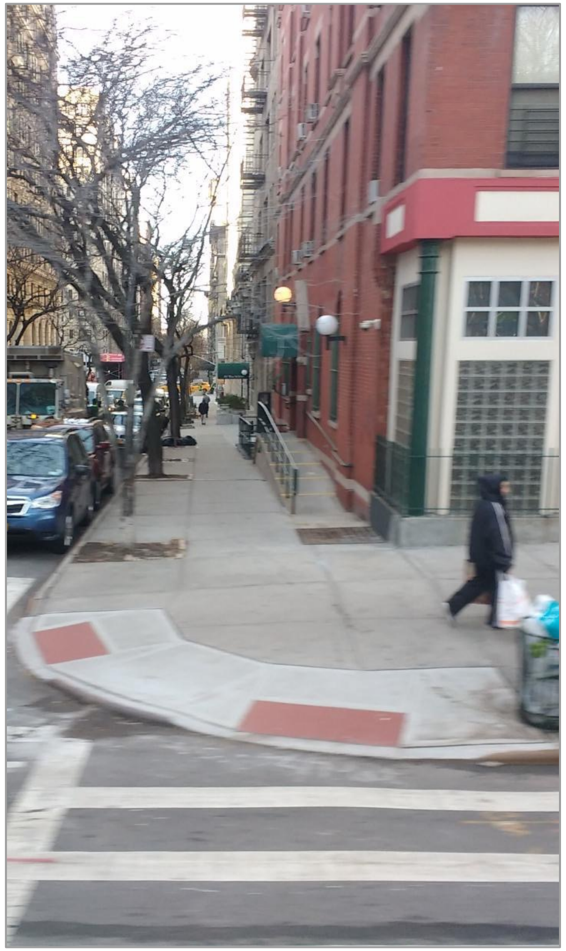
August

October

Accessibility: assessing hazards for older adults







Main takeaways

- General approaches might not scale.
- Big data visual analytics might require structures specifically designed for a given task.
- Understand the data, and how it can be transformed to best fit different computational resources and architectures.

Main takeaways