Visualization-Aware Sampling for Very Large Databases

Yongjoo Park
Michael Cafarella
Barzan Mozafari

University of Michigan, Ann Arbor
Prevalence of Viz-centric Data Analysis

Weather

Population

Brain

Genome
Big Data, Poor Latency

A part of 2 billion points took 2+ hours for 100M points using MathGL. Tableau crashed for 100 million points on a machine with 122GB memory (r3.4xlarge).
Big Data, Poor Latency

A part of 2 billion points
Big Data, Poor Latency

A part of 2 billion points

71 mins !! (matplotlib)
Big Data, Poor Latency

A part of 2 billion points

71 mins !! (matplotlib)

MathGL took 2+ hours for 100M points.
Big Data, Poor Latency

A part of 2 billion points

71 mins !! (matplotlib)

MathGL took 2+ hours for 100M points.

Tableau crashed for 100 million points on a machine with 122GB memory (r3.4xlarge).
Fast Response is Important

Five times more interactions as the response time was reduced. The most skilled user went from 800 interactions per hour with a 1.5-second response time up to 4,300 interactions per hour with a 0.4-second response time.

+500ms latency
50% less data exploration

6 out of 16 subjects did not report a noticeable difference in terms of system responsiveness.

A good reason for my coffee break.
Fast Response is Important

Five times more interactions as the response time was reduced [3].

"The most skilled user went from 800 interactions per hour with a 1.5-second response time up to 4,300 interactions per hour with a 0.4-second response time."
Fast Response is Important

Five times more interactions as the response time was reduced [3].

“\[
\text{The most skilled user went from 800 interactions per hour with a 1.5-second response time up to 4,300 interactions per hour with a 0.4-second response time.}
\]

+500ms latency $\rightarrow$ 50% less data exploration subconsciously [1]

“\[
\text{6 out of 16 subjects did not report a noticeable difference in terms of system responsiveness.}
\]
Fast Response is Important

Five times more interactions as the response time was reduced [3].

“The most skilled user went from 800 interactions per hour with a 1.5-second response time up to 4,300 interactions per hour with a 0.4-second response time.”

+500ms latency → 50% less data exploration subconsciously [1]

“6 out of 16 subjects did not report a noticeable difference in terms of system responsiveness.”

A good reason for my coffee break.
Why is Visualizing Big Data Slow?

We want to reduce computational efforts without affecting visual perception.
Why is Visualizing Big Data Slow?

We want to reduce computational efforts without affecting visual perception.
Why is Visualizing Big Data Slow?

We want to reduce computational efforts without affecting visual perception.
Why is Visualizing Big Data Slow?

We want to reduce computational efforts without affecting visual perception.
Why is Visualizing Big Data Slow?

We want to reduce computational efforts without affecting visual perception.
Why is Visualizing Big Data Slow?

We want to reduce computational efforts without affecting visual perception.
Why is Visualizing Big Data Slow?

We want to reduce computational efforts without affecting visual perception.
Why is Visualizing Big Data Slow?

We want to

Reduce computational efforts
Why is Visualizing Big Data Slow?

We want to

**Reduce** computational efforts

**Without** affecting visual perception
Two Approaches to Fast Viz

- **Pre-aggregation**
  - When $|D| < |S|$
  - Faster

- **Data Reduction**
  - Our Approach
  - Good when:
    1. You do not want pre-aggregations (no knowledge, lost info).
    2. You want no modifications on Viz Application/DB.
Two Approaches to Fast Viz

1. Pre-aggregation
   - You do not want pre-aggregations (no knowledge, lost info).
   - You want no modifications on Viz Application/DB.

2. Data Reduction
   - (# of bins < $|\mathcal{D}|$) \rightarrow faster

Pre-aggregation
Two Approaches to Fast Viz

Pre-aggregation

(Number of bins < |D|) $\rightarrow$ faster

Data Reduction

(|S| < |D|) $\rightarrow$ faster
Two Approaches to Fast Viz

Pre-aggregation

\((\# \text{ of bins } < |\mathcal{D}|) \rightarrow \text{faster}\)

Data Reduction

\((|S| < |\mathcal{D}|) \rightarrow \text{faster}\)

Our Approach. Good when:

1. You do not want pre-aggregations (no knowledge, lost info).
2. You want **no modifications** on Viz Application/DB.
Our Concrete Goal: Fast Viz of Scatter Plots

Why Scatter Plots?
Our Concrete Goal: Fast Viz of Scatter Plots

Why Scatter Plots?

1. Correlation
Our Concrete Goal: Fast Viz of Scatter Plots

Why Scatter Plots?

1. Correlation

2. Trend Analysis
Our Concrete Goal: Fast Viz of Scatter Plots

Why Scatter Plots?

1. Correlation

2. Trend Analysis

3. Geographic Info
Our Concrete Goal: Fast Viz of Scatter Plots

Why Scatter Plots?

1. Correlation
2. Trend Analysis
3. Geographic Info
4. Density Estimation
Simple Architecture if you have a Good Sample

Once you generate a sample offline, enjoy fast viz at query time just like B-tree used to speed up queries.

Question 1: What is a good Sample?
Question 2: How to obtain such a good Sample?
Simple Architecture if you have a Good Sample

Once you generate a sample offline → Enjoy fast viz at query time
Simple Architecture if you have a Good Sample

Once you generate a sample offline → Enjoy fast viz at query time

Just like B-tree used to speed up queries
Simple Architecture if you have a Good Sample

Once you generate a sample offline, enjoy fast viz at query time.

Just like B-tree used to speed up queries.

Question 1: What is a good Sample?
Simple Architecture if you have a Good Sample

Once you generate a sample offline, Enjoy fast viz at query time

Just like B-tree used to speed up queries

Question 1: What is a good Sample?

Question 2: How to obtain such a good Sample?
Existing Samplings Fail

Original (2 billion points, 71 mins)
Existing Samplings Fail

1. Uniform random sample
   (1 million, 3 secs)

Original (2 billion points, 71 mins)
Existing Samplings Fail

Original (2 billion points, 71 mins)

1. Uniform random sample
   (1 million, 3 secs)

2. Stratified sample
   (1 million, 3 secs)
Existing Samplings Fail

1. Uniform random sample
   (1 million, 3 secs)

2. Stratified sample
   (1 million, 3 secs)

3. Ours
   (1 million, 3 secs)

Original (2 billion points, 71 mins)
Three common visualization-driven tasks [4]:

1. Trend Analysis
2. Density Est.
3. Clustering

Other tasks: (1) shape viz, (2) classification, (3) hierarchy understanding, (4) community detection, etc.
Three common visualization-driven tasks [4]:

1. Trend Analysis
Three common visualization-driven tasks [4]:

1. Trend Analysis
2. Density Est.
Three common visualization-driven tasks [4]:

1. Trend Analysis
2. Density Est.
3. Clustering
Three common visualization-driven tasks [4]:

1. Trend Analysis
2. Density Est.
3. Clustering

Other tasks: (1) shape viz, (2) classification, (3) hierarchy understanding, (4) community detection, etc.
Three common visualization-driven tasks [4]:

1. Trend Analysis
2. Density Est.
3. Clustering

What we focus

Other tasks: (1) shape viz, (2) classification, (3) hierarchy understanding, (4) community detection, etc.
Outline

1. Motivation
2. Formal Definition of Good Sample
3. Approximation Algorithm to VAS Problem
4. Large-Scale User Study
5. Offline Runtime Analysis
What is a good sample \((S)\) of the original dataset \((D)\)?
Good Sample leads to High Fidelity

What is a good sample \((S)\) of the original dataset \((D)\)?

\[
\text{Loss}(S) = \int \text{subloss}(x) \, dx
\]

If the visual difference for every circle is small, two viz will look similar.

Our Goal
To minimize:

\[
\text{Loss}(S) = \int \text{subloss}(x) \, dx
\]

where \(\text{subloss}(x)\) is the viz distance for the circle centered at \(x\).

What function to use for \(\text{subloss}(x)\)?
What is a good sample ($S$) of the original dataset ($D$)?

\[ \text{Loss}(S) = \int \text{subloss}(x) \, dx \]

where \( \text{subloss}(x) \) is the viz distance for the circle centered at \( x \).

What function to use for \( \text{subloss}(x) \)?
What is a **good** sample \((S)\) of the original dataset \((D)\)?

\[
\text{subloss}(x) = \int \text{loss}(x) \, dx
\]

If the visual difference for every circle is small, two viz will look similar.

Our Goal

To minimize:

\[
\text{Loss}(S) = \int \text{subloss}(x) \, dx
\]

where \(\text{subloss}(x)\) is the viz distance for the circle centered at \(x\).
Good Sample leads to High Fidelity

What is a **good** sample \((S)\) of the original dataset \((D)\)?

\[ \text{subloss}(x) = D - S \]

If the visual difference for every circle is small, the two visualizations will look similar.

Our Goal: To minimize:

\[ \text{Loss}(S) = \int \text{subloss}(x) \, dx \]

where \(\text{subloss}(x)\) is the visualization distance for the circle centered at \(x\).

What function to use for \(\text{subloss}(x)\)?
What is a **good** sample \((S)\) of the original dataset \((D)\)?

\[
\text{subloss}(x) = \int \text{subloss}(x) \, dx
\]

If the visual difference for every circle is small, \(S\) and \(D\) will look similar.

Our Goal

To minimize:

\[
\text{Loss}(S) = \int \text{subloss}(x) \, dx
\]

where \(\text{subloss}(x)\) is the viz distance for the circle centered at \(x\).

What function to use for \(\text{subloss}(x)\)?
What is a **good** sample ($S$) of the original dataset ($D$)?

\[
\text{subloss}(x) = D - S
\]

If the visual difference for **every circle** is small, 
$\rightarrow$ two viz will look similar.
What is a **good** sample \((S)\) of the original dataset \((D)\)?

\[
\text{subloss}(x) = \int \text{subloss}(x) \, dx
\]

If the visual difference for **every circle** is small, 
→ two viz will look similar.

**Our Goal**

To minimize:

\[
\text{Loss}(S) = \int \text{subloss}(x) \, dx
\]

where \text{subloss}(x) is the viz distance for the circle centered at \(x\).
What is a **good** sample \((S)\) of the original dataset \((D)\)?

\[
\text{subloss}(x) = \|D| - S\|
\]

If the visual difference for **every circle** is small, → two viz will look similar.

**Our Goal**

To minimize:

\[
\text{Loss}(S) = \int \text{subloss}(x) \, dx
\]

where subloss\((x)\) is the viz distance for the circle centered at \(x\).

What function to use for subloss\((x)\)?
Our Suggestion for subloss

Let’s see what happens around the location $x$. $D_x S_1 x S_2 x S_3$

Initially, visual distance around $x$ reduces fast. Visual distance around $x$ NO LONGER reduces much.

Our Choice subloss $(x) = \sum_{s_i} S(x; s_i)$ where $s_i$ captures the proximity between two coordinates.
Our Suggestion for subloss

Let's see what happens around the location $x$.

Initially, visual distance around $x$ reduces fast. Visual distance around $x$ NO LONGER reduces much.

Our Choice subloss $(x) = \sum s_{i} S(x; s_{i})$ where $D$ captures the proximity between two coordinates.
Let’s see what happens around the location $x$.

Initially, visual distance around $x$ reduces fast. Visual distance around $x$ NO LONGER reduces much.

Our Choice subloss $(x) = 1 \sum_{s_i} S_s(x; s_i)$ where $S_s$ captures the proximity between two coordinates.
Our Suggestion for subloss

Let's see what happens around the location $x$.

Initially, visual distance around $x$ reduces fast. Visual distance around $x$ NO LONGER reduces much.

Our Choice subloss $(x) = \sum_{s_i} S_i(x; s_i)$

where $D$ captures the proximity between two coordinates.
Our Suggestion for subloss

Let's see what happens around the location $x$. Initially, visual distance around $x$ reduces fast. Visual distance around $x$ NO LONGER reduces much.

Our Choice subloss $(x) = \sum_{s_i} s_i$ where $D$ captures the proximity between two coordinates.
Our Suggestion for subloss

Let’s see what happens around the location \( x \).

\[
\text{Our Choice subloss}\,(x) = \sum_{s_i \in S} (x; s_i)
\]

where \( S \) captures the proximity between two coordinates.
Our Suggestion for subloss

Let’s see what happens around the location \( x \).

Initially, visual distance around \( x \) reduces fast.
Our Suggestion for subloss

Let’s see what happens around the location $x$.

Initially, visual distance around $x$ reduces fast.

Visual distance around $x$ NO LONGER reduces much.
Our Suggestion for subloss

Let’s see what happens around the location $x$.

Initially, visual distance around $x$ reduces fast.

Visual distance around $x$ NO LONGER reduces much.

Our Choice

$$\text{subloss}(x) = \frac{1}{\sum_{s_i \in S} \kappa(x, s_i)}$$

where $\kappa$ captures the proximity between two coordinates.
Given a budget $|S| = K$, we want

$$\arg\min_S \text{Loss}(S) \quad \text{s.t.} \quad S \subseteq \mathcal{D} \land |S| = K$$

$$\int \frac{1}{\sum_{s_i \in S} \kappa(x,s_i)} \, dx$$
Good Viz by Solving VAS Problem

Given a budget $|S| = K$, we want

$$\arg\min_S S \quad \text{s.t. } S \subseteq D \land |S| = K$$

$$\text{Loss}(S) \quad \int \frac{1}{\sum_{s_i \in S} \kappa(x, s_i)} \, dx$$

After some approximations and calculations,

**Problem VAS**

$$\arg\min_S \sum_{s_i, s_j \in S; i < j} \tilde{\kappa}(s_i, s_j) \quad \text{Loss}(S)$$

where

$$\tilde{\kappa}(s_i, s_j) = \int \kappa(x, s_i) \kappa(x, s_j) \, dx$$
Good Viz by Solving VAS Problem

Given a budget $|S| = K$, we want

$$\arg\min_S \quad \text{Loss}(S) \quad \text{s.t.} \quad S \subseteq \mathcal{D} \land |S| = K$$

After some approximations and calculations,

**Problem VAS**

$$\arg\min_S \quad \sum_{s_i, s_j \in S; i<j} \tilde{\kappa}(s_i, s_j) \quad \text{s.t.} \quad S \subseteq \mathcal{D} \land |S| = K$$

where

$$\tilde{\kappa}(s_i, s_j) = \int \kappa(x, s_i) \kappa(x, s_j) \, dx$$

NP-hard (evaluated later); How can we solve?
Outline

1. Motivation
2. Formal Definition of Good Sample
3. Approximation Algorithm to VAS Problem
4. Large-Scale User Study
5. Offline Runtime Analysis
Basic (Slow) Approach

Nemhauser, et al. [2] algorithm (comes with an error guarantee)
Basic (Slow) Approach

Nemhauser, et al. [2] algorithm (comes with an error guarantee)

As scanning over a dataset,
Nemhauser, et al. [2] algorithm (comes with an error guarantee)

As scanning over a dataset,

When $|S| < K$
Nemhauser, et al. [2] algorithm (comes with an error guarantee)

As scanning over a dataset,

When $|S| < K$

When $|S| = K$
Nemhauser, et al. [2] algorithm (comes with an error guarantee)

As scanning over a dataset,

When $|S| < K$

When $|S| = K$

Valid replacement: $\text{Loss}(S') < \text{Loss}(S)$
Basic (Slow) Approach

Nemhauser, et al. [2] algorithm (comes with an error guarantee)

As scanning over a dataset,

When $|S| < K$

New point

When $|S| = K$

Valid replacement: $\text{Loss}(S') < \text{Loss}(S)$

Testing for valid replacements is too Slow: $O(K^3)$ for every point
Our algorithm: Expand/Shrink operation
Our Fast Approach

Our algorithm: Expand/Shrink operation

When $|S| < K$
Our Fast Approach

Our algorithm: Expand/Shrink operation

When $|S| < K$

Expand

\[ |S''| = K + 1 \]

Shrink

\[ |S'| = K \]

When $|S| = K$

Expand/Shrink is Fast: $O(K)$ for every point vs $O(K^2)$ times faster!!

The result is exactly the same as the Nemhauser's algorithm.
Our Fast Approach

Our algorithm: **Expand/Shrink** operation

When $|S| < K$

Expand/Shrink is **Fast**: $O(K)$ for every point $\rightarrow O(K^2)$ times faster!!
Our Fast Approach

Our algorithm: **Expand/Shrink** operation

- **Expand**
  - When $|S| < K$
  - $|S''| = K + 1$

- **Shrink**
  - When $|S| = K$
  - $|S'| = K$

Expand/Shrink is **Fast**: $O(K)$ for every point $\rightarrow O(K^2)$ times faster!!

The result is exactly same as the Nemhauser’s algorithm.
1. Motivation
2. Formal Definition of Good Sample
3. Approximation Algorithm to VAS Problem
4. Large-Scale User Study
5. Offline Runtime Analysis
Large-scale User Study: Setup

Questions: User performance, Validity of VAS Problem

Tasks: trend analysis, density est, clustering

Questions #: 72 - 80 with different sample sizes

Datasets: 22 million GPS logs [5] Synthetic mechanical turk 40 unique
Large-scale User Study: Setup

Questions: User performance, Validity of **VAS Problem**
## Large-scale User Study: Setup

Questions: User performance, Validity of **VAS Problem**

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Questions</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>trend analysis, density est, clustering</td>
<td>#: 72 - 80 with different sample sizes</td>
<td>• 22 million GPS logs [5]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Synthetic</td>
</tr>
</tbody>
</table>
Large-scale User Study: Setup

Questions: User performance, Validity of VAS Problem

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Questions</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>trend analysis, density est, clustering</td>
<td>#: 72 - 80 with different sample sizes</td>
<td>22 million GPS logs [5] · Synthetic</td>
</tr>
</tbody>
</table>

Datasets

- 22 million GPS logs [5]
- Synthetic

Tasks

- trend analysis
- density estimation
- clustering

Questions

- #: 72 - 80 with different sample sizes

Datasets

- 22 million GPS logs [5]
- Synthetic

mechanical turk

× 40 unique
Where is the Densest?

In the following figure, density is indicated by either
- Putting more dots in denser areas, or
- Using larger dots for denser areas.

Given the map below, your task is to estimate the most or least dense area.
The left figure is without marks, and the right figure is with marks.

Choose the MOST dense area in the figure above:

- A
- B
- C
- D
- I am not sure.

Choose the LEAST dense area in the figure above:
Where is the Densest?

In the following figure, density is indicated by either

- Putting more dots in denser areas, or
- Using larger dots for denser areas.

Given the map below, your task is to estimate the most or least dense area.

The left figure is without marks, and the right figure is with marks.

Choose the MOST dense area in the figure above:

- [ ] A
- [ ] B
- [ ] C
- [ ] D
- [ ] I am not sure.

Choose the LEAST dense area in the figure above:

- [ ] A
- [ ] B
- [ ] C
- [ ] D
- [ ] I am not sure.
**Where is the Densest?**

In the following figure, density is indicated by either

- Putting more dots in denser areas, or
- Using larger dots for denser areas.

Given the map below, your task is to estimate the most or least dense area.

The left figure is without marks, and the right figure is with marks.

Choose the MOST dense area in the figure above:

- A
- B
- C
- D
- I am not sure.
Superior User Performance by VAS

Average user performance

1. Uniform Random Sampling
2. Stratified Sampling
3. Visualization-Aware Sampling (VAS)
Superior User Performance by VAS

Average user performance

1. Uniform Random Sampling
2. Stratified Sampling
3. Visualization-Aware Sampling (VAS)

Trend Analysis

success: if a user answers a question correctly.
Superior User Performance by VAS

Average user performance

1. Uniform Random Sampling
2. Stratified Sampling
3. Visualization-Aware Sampling (VAS)

Trend Analysis

Density

success: if a user answers a question correctly.
Superior User Performance by VAS

Average user performance

1. Uniform Random Sampling
2. Stratified Sampling
3. Visualization-Aware Sampling (VAS)

Trend Analysis
Density
Clustering

success: if a user answers a question correctly.
Strong Correlation (between User Performance and Loss)

Smaller Loss ($S$) implies More success (Spearman's rank) correlation coefficient = -0.85
Strong Correlation (between User Performance and Loss)

Smaller Loss($S$) $\rightarrow$ More successes
Strong Correlation (between User Performance and Loss)

Smaller Loss(S) $\rightarrow$ More successes

(Spearman’s rank) correlation coefficient = -0.85
1. Motivation

2. Formal Definition of Good Sample

3. Approximation Algorithm to VAS Problem

4. Large-Scale User Study

5. Offline Runtime Analysis
We tested Mixed Integer Programming (MIP) for exact solutions.

Exact approach (MIP) is prohibitively slow even for small data.
Ran our VAS algorithm over 2 billion points.
Ran our VAS algorithm over 2 billion points.

Minimizes Loss fast.
Ran our VAS algorithm over 2 billion points.

Minimizes Loss fast.
Ran our VAS algorithm over 2 billion points.

Minimizes Loss fast.
Fast Offline Sampling over Big Data

Ran our VAS algorithm over 2 billion points.

Minimizes Loss fast.

High-quality viz even at early steps
Ran our VAS algorithm over 2 billion points.

Minimizes Loss fast.

High-quality viz even at early steps

Gradually improves if more time is allowed
Also available at yongjoopark.com/vas
We formulated an important problem: **Sampling for Visualization**.

We proposed an efficient algorithm to the VAS problem. We demonstrated that users achieve superior performance with VAS.
We formulated an important problem: Sampling for Visualization.
Conclusion

We formulated an important problem: Sampling for Visualization.

We proposed an efficient algorithm to VAS problem.
Conclusion

We formulated an important problem: *Sampling for Visualization*.

We proposed an efficient algorithm to VAS problem.

We demonstrated users achieve *superior performance* with VAS.
Thank You!
Z. Liu and J. Heer.  
**The effects of interactive latency on exploratory visual analysis.**  
*TVCG*, 2014.

**An analysis of approximations for maximizing submodular set functions—i.**  

B. Shneiderman.  
**Response time and display rate in human performance with computers.**  
B. Shneiderman.
The eyes have it: A task by data type taxonomy for information visualizations.
In Symposium on Visual Languages, 1996.

Understanding mobility based on gps data.
In UbiComp, 2008.