Visualization Post-processing of Differentially Private Data

Diane Tam

University of Massachusetts Amherst
College of Information and Computer Sciences

Differential Privacy and Visualization

- Data visualization
- Differential privacy
- Challenges in visualizing private data
  - Visual utility
  - Visual artifacts
  - One-dimensional vs two-dimensional data

Problem Statement

- Using differentially private algorithms from DPBench:
  - Analyze visualization properties of varying algorithms
  - Explore baseline smoothing techniques for post-processing noisy data

Research Design and Methods

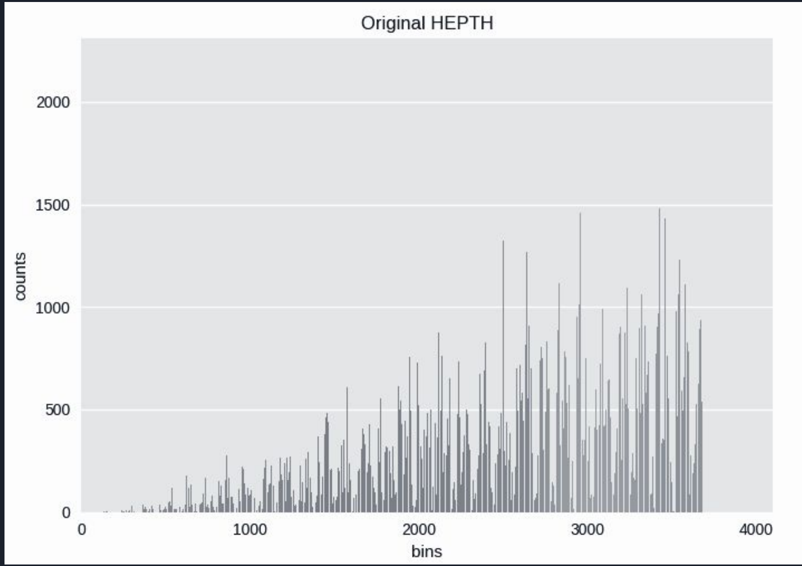
- DPBench code base (Python 2.7-> 3.4)
    - 1-D Experiments
        - Algorithms: Hb, Identity, MWEM, and DAWA
        - Datasets: BIDS-ALL and HEPTH
    - 2-D Experiments
        - Algorithms: Identity
        - Datasets: BJTaxi, US, GOWALLA, and BOS
- Seabon & Matplotlib 1.4.3
- OpenCV 3.3.1
    - Filters applied to 256px by 256px .png images

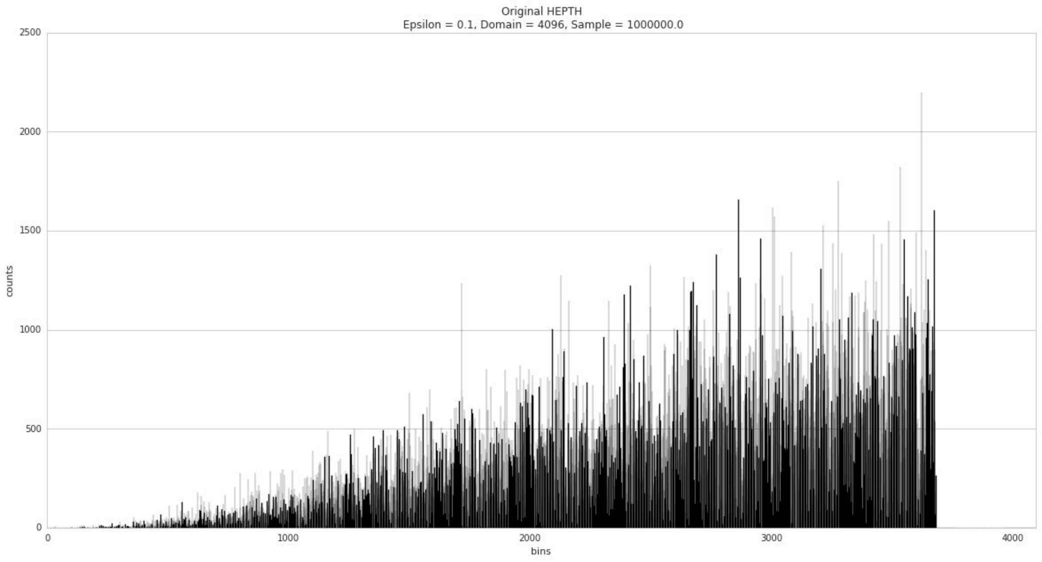
Research Design and Methods

- 1-D Experiment
  - Histogram and cumulative density function
  - Varying parameters:
    - Scale-epsilon pairing
    - Domain: 4096 or 1024
- Six noisy plots per choice of parameters
- Normalized non-negative rounding
  - Multiply positive bin counts by [original noisy data sum][sum without negative values]
  - Round negative bin counts to zero

Research Design and Methods

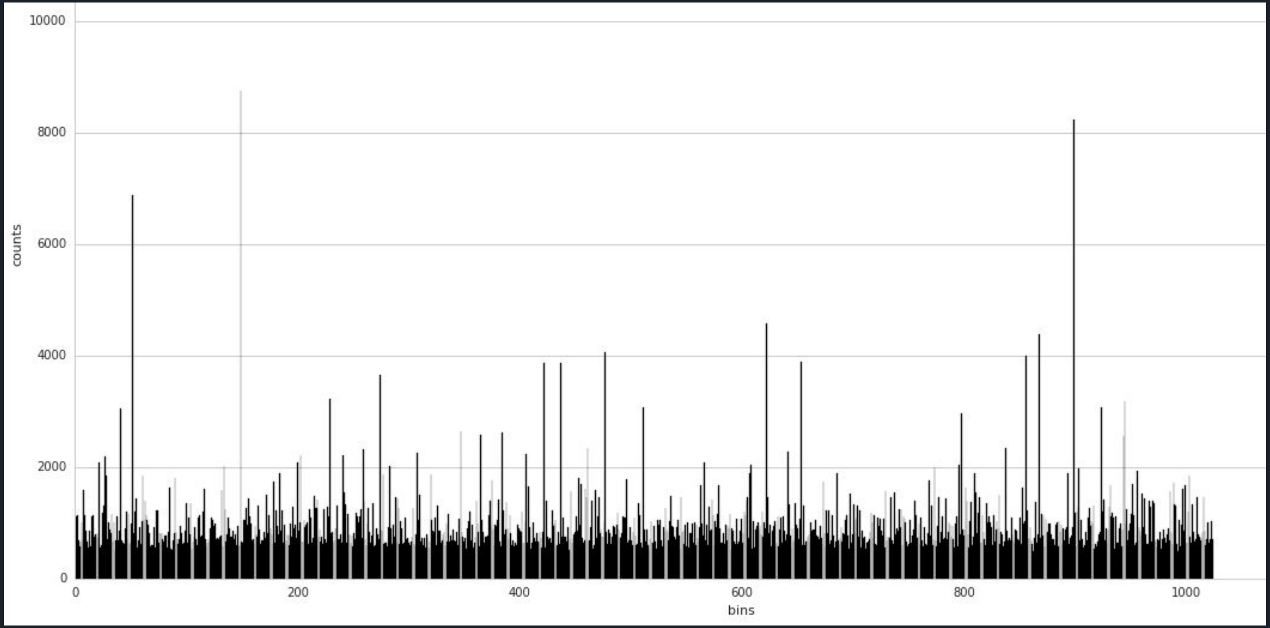
- Matplotlib 2.1.0 vs. matplotlib 1.4.3 (visual artifacts)
- Original HEPTH dataset w/ epsilon = 0.1, domain = 4096, sample = 10^6
  - Default bins seem to overlap and are clustered slightly differently





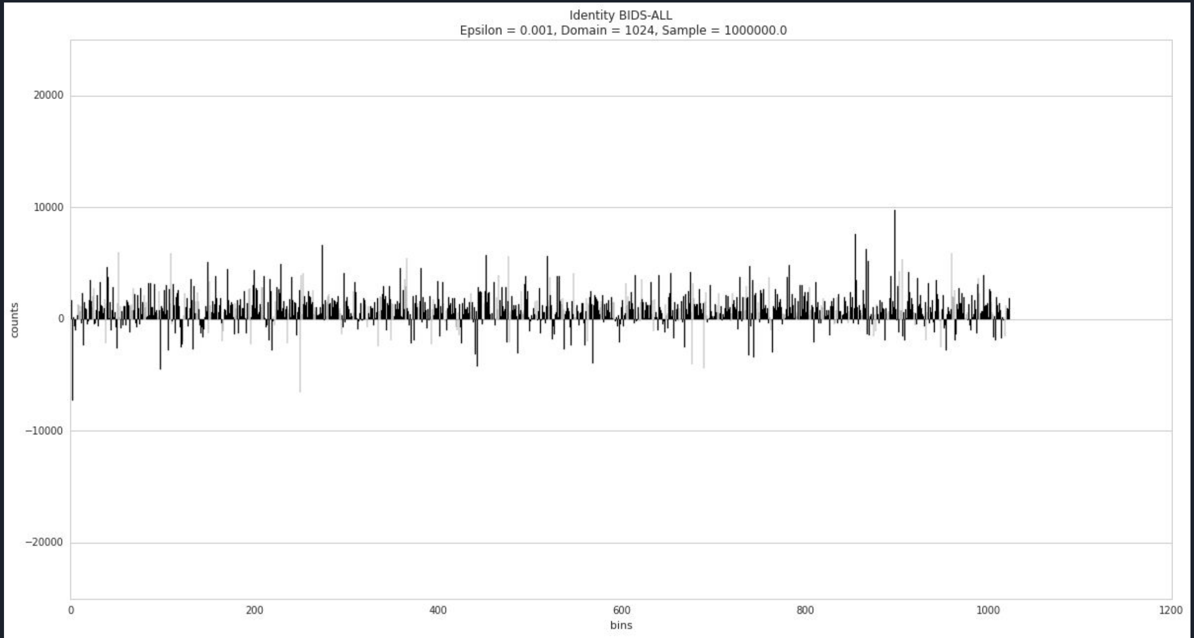
Findings (1 Dimensional Data)

Original BIDS-ALL at epsilon = 0.001, scale= 10^6



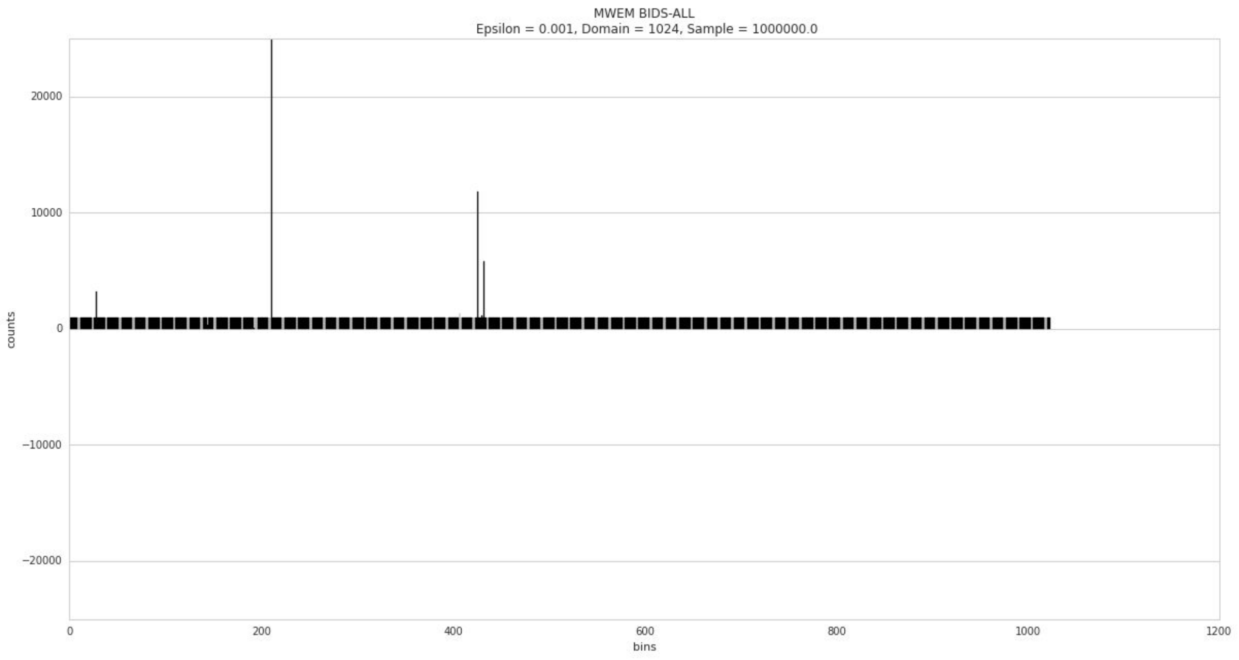
Findings (1 Dimensional Data)

Algorithms like H_b or Identity output negative counts which warp the scale of the visualization



Findings (1 Dimensional Data)

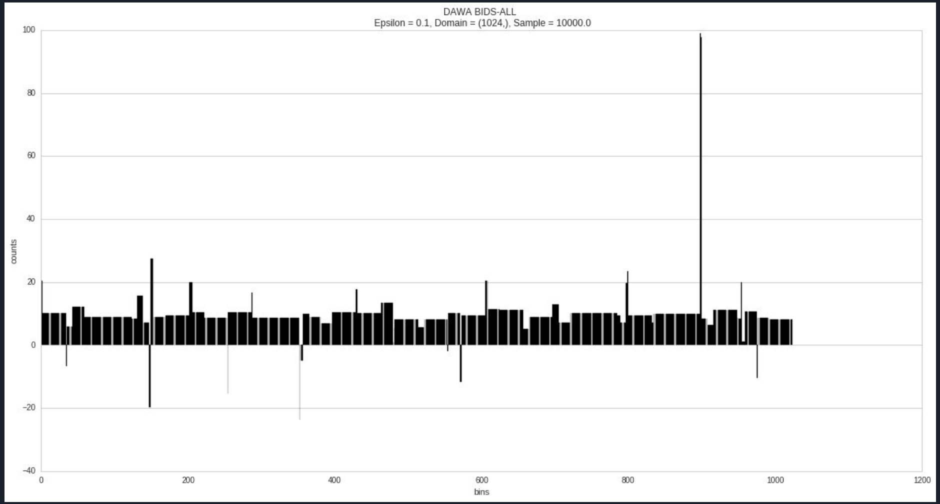
Data-dependent algorithms like DAWA or MWEM utilize workload queries to add noise which can cause “blocky” visualizations.

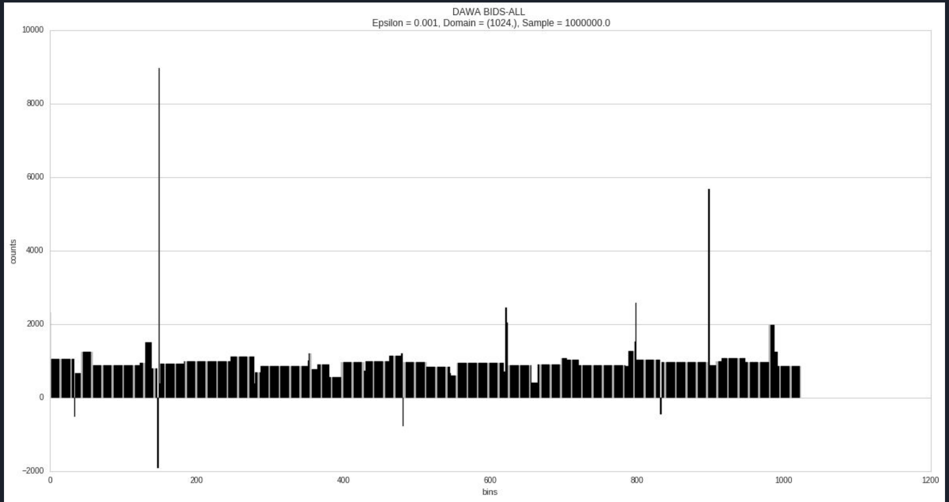


Findings (1 Dimensional Data)

Visualizations are not quite 100% scale-epsilon exchangeable

DAWA on BIDS-ALL with (left) epsilon = 0.1, scale = 10^4 (right) epsilon = 0.001, scale = 10^6

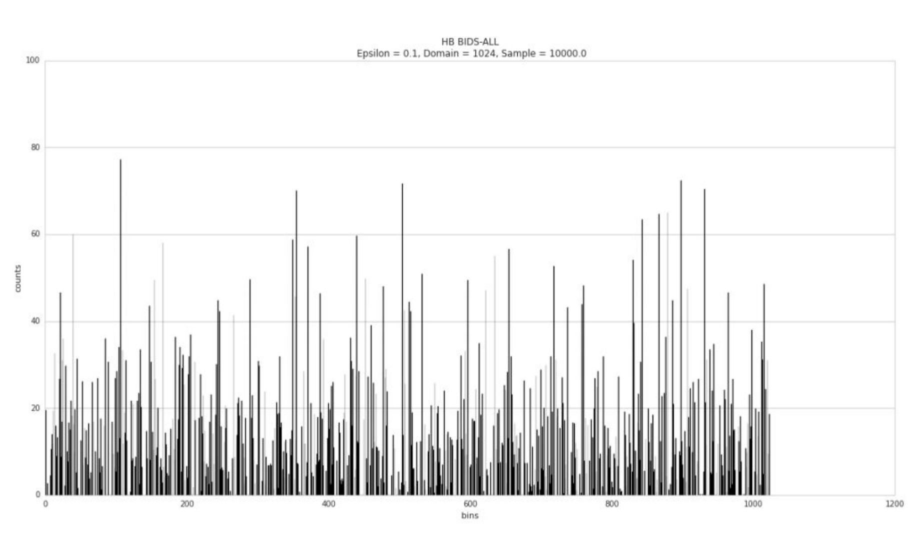


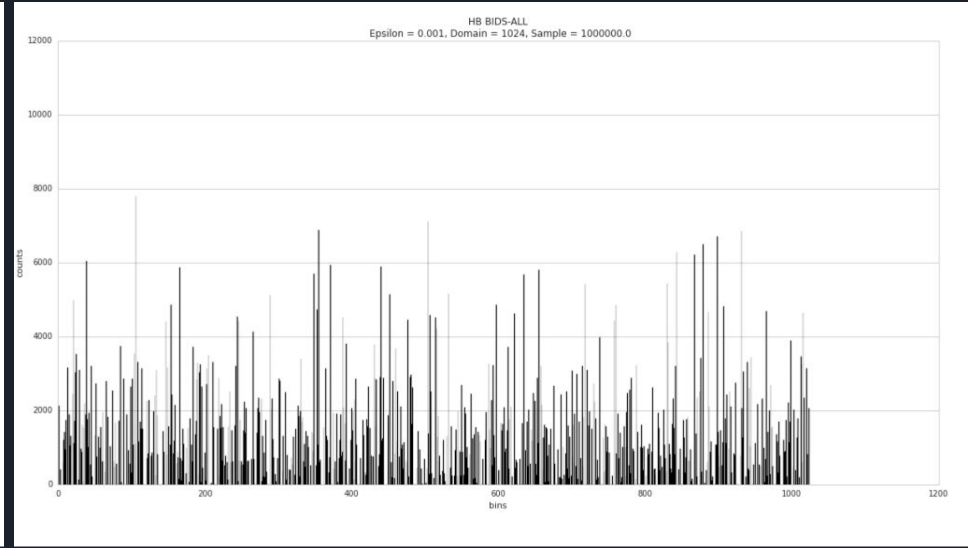


Findings (1 Dimensional Data)

Visualizations are considerably scale-epsilon exchangeable. Visual utility depends on the intended task

Hb on BIDS-ALL with (left) epsilon = 0.1, scale = 10^4 (right) epsilon = 0.001, scale = 10^6

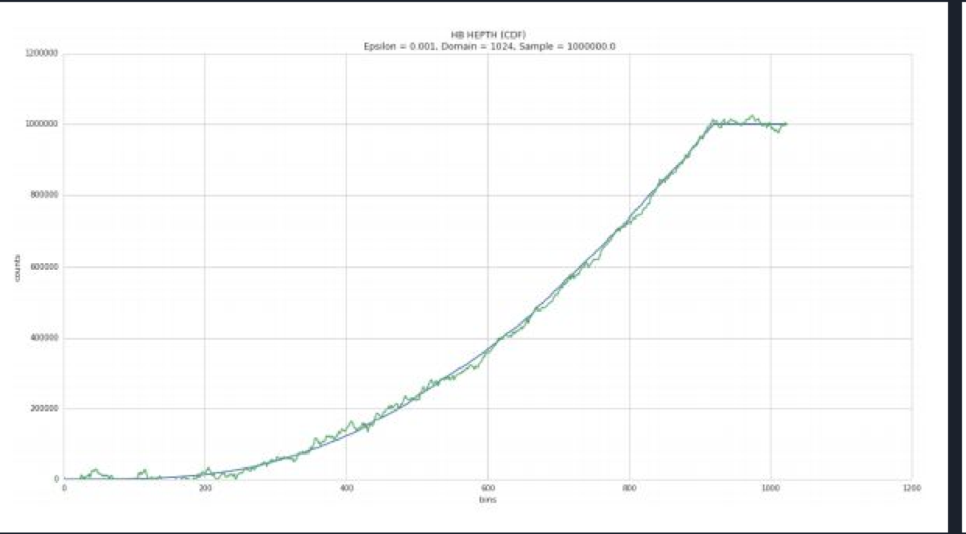


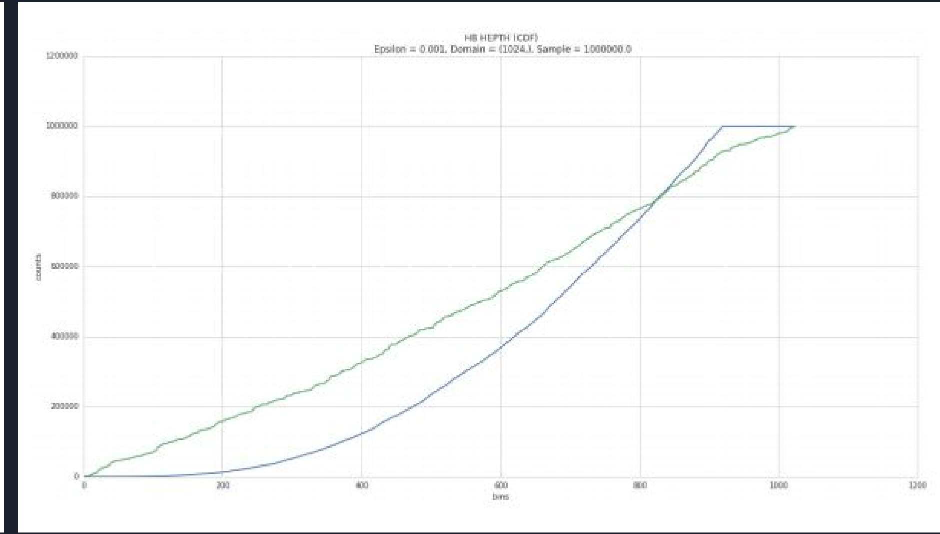


Findings (1 Dimensional Data)

Normalized non-negative rounding post-processing technique *improves* visual utility of PDF, *decreases* visual utility of CDF.

Hb on HEPTH epsilon = 0.001, scale = 10^6 (left) no post-processing (right) with post-processing





Research Design and Methods

- 2-D Experiment
    - Colored heatmap
        - Saved as .png of 256px by 256px
    - Scale-epsilon pairing
    - Apply OpenCV filter
        - Convolution filter
            - Kernel design
        - Bilateral filter
            - Diameter, SigmaColor, SigmaSpace
- Non-negative rounding to zero

Research Design and Methods

- Convolution filter

$$K = \frac{1}{K_{width} K_{height}} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

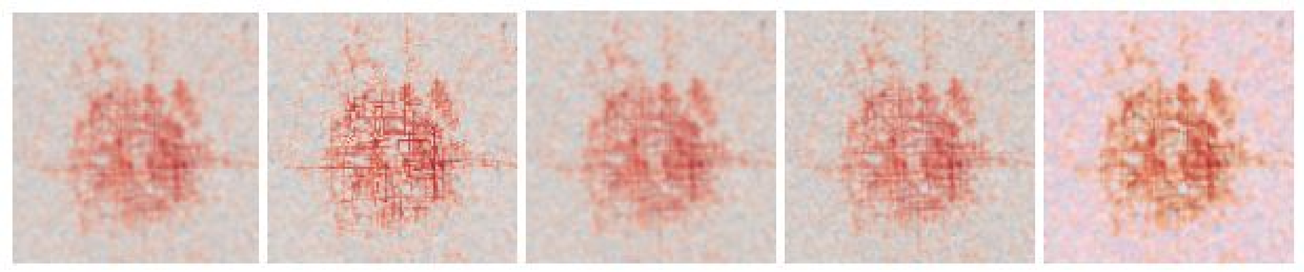
- Bilateral filter
    - Based on gaussian filter but accounts for pixel similarity and space (closeness)
    - Effective in removing noise and retaining sharp edges

Findings (2 Dimensional Data)

Five standard OpenCV filtering processes tested, bilateral and convolution filters best

Original BJTaxi with scale 10^6

Identity on BJTaxi with scale 10^6 with averaging, bilateral, blurred, gaussian blur, and median blur filters



Findings (2 Dimensional Data)

$$\mathrm{K} = \frac{1}{k^2 - 1} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

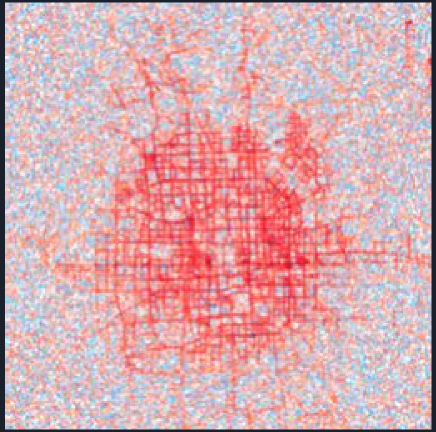
Top convolution filter kernel

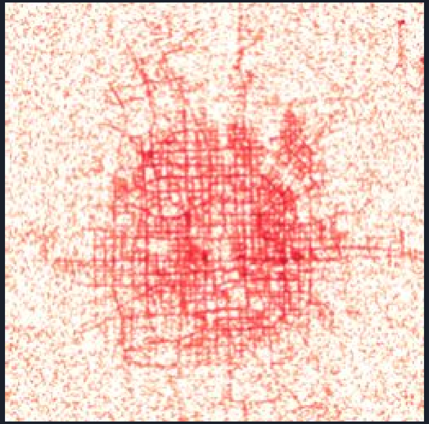
$k_{width} = k_{height} \Rightarrow k$

BJTaxi scale 10^8, epsilon 10^{-2}
(Left to right): Original, Averaging kernel, Top kernel

Findings (2 Dimensional Data)

Post-processing the noisy data by rounding all negative values to zero may benefit visual utility depending on the associated task





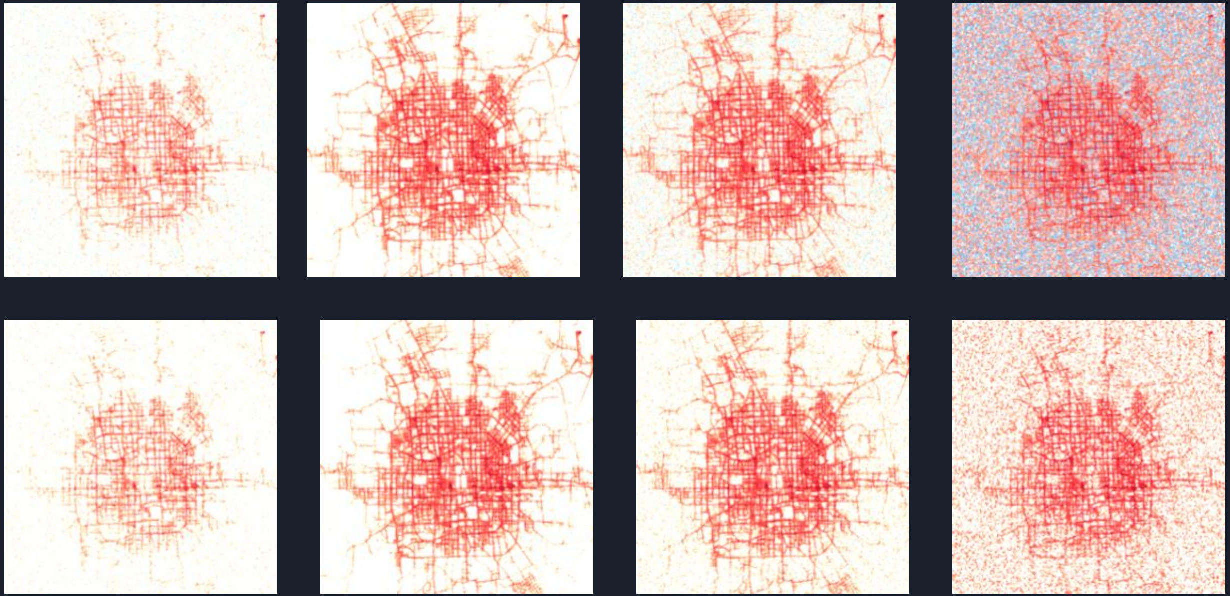
BJTaxi w/ top convolution filter scale 10^8, epsilon 10^{-3}
(Left to right): Original, No post-processing, Non-negative post-processing

Findings (2 Dimensional Data)

The degree of effectiveness in using this convolution filter as a baseline noise smoothing technique decreases as noise increases

No post-processing

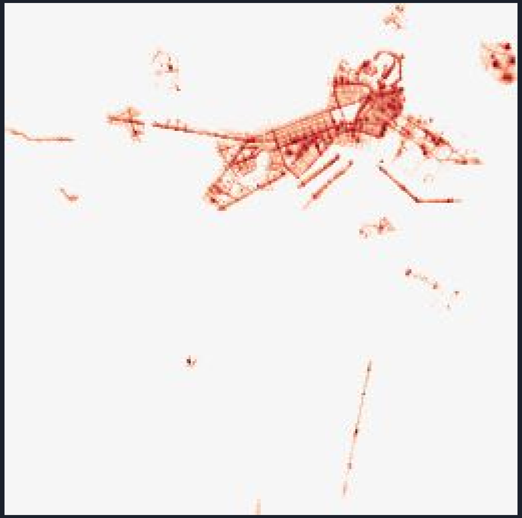
Non-negativity post-processing

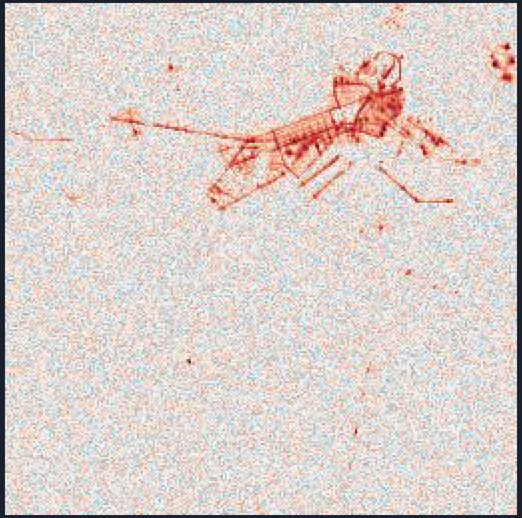


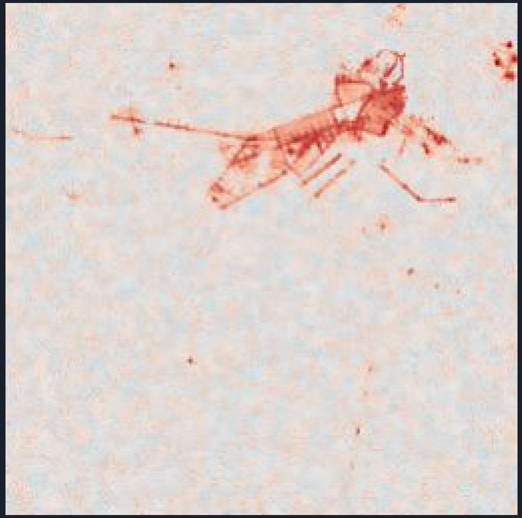
BJTaxi w/ Top Convolution filter: Increasing levels of noise from left to right

Findings (2 Dimensional Data)

Bilateral filter much less drastic improvements, but might be useful for tasks requiring detailed analysis of datasets with sparse points and low noise







BOS at scale 10^6, epsilon 0.1
(Left to right) Original, Noisy, Bilateral Filter d = 9, s = 70

Implications

- Different artifacts created by each DP algorithm and/or smoothing technique
    - Cannot confidently state all-in-one algorithm and smoothing technique for a reliable visualization for all tasks
- Previous visualization techniques rarely involve differential privacy
    - Initial exploration of how to categorize visual artifacts based on algorithm selection
    - Initial exploration of baseline smoothing techniques for DP algorithms that have similar problems with image processing

Conclusion

- Repeat with Amazon's Mechanical Turk
  - Psychology and human perception
- Computer vision image similarity concepts
  - Use CV to define "similar" image as human perceives visual utility of a noisy image
  - Help us determine good baselines and eliminate user bias from the rankings and conclusions
- Visualization concepts related to privacy
  - Different methods of presenting data
  - Color, shapes, etc.

Thank You & Questions

Committee Members:

Gerome Miklau, College of Information and Computer Sciences

Dan Zhang, College of Information and Computer Sciences