?ncprspc Rgjcq8 Understanding Big Data
with Tile-Based Visual Analytics

apertureTiles™

Goal

Provide new automated, **interactive**, web-based, easy-to-use visualization for **billions of data points**. Plot all the data.

Before testing hypotheses, conÿrmatory data analysis beneÿts from ÿrst examining the data to suggest hypotheses to be tested. This is known as exploratory data analysis (EDA).

Approach

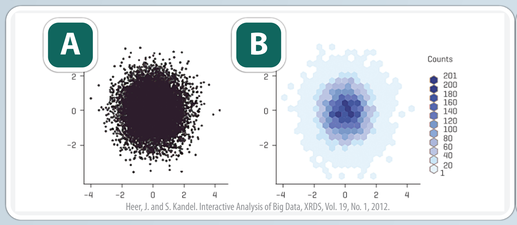
To assess the feasibility and utility of using tile-based rendering for big data scatter plots, to identify usability and user experience issues, and to develop requirements for further implementation, we are experimenting with a variety of datasets, including:

- **Kiva Microfinance** (500K points, 5.9 GB)
- **Bitcoin** (37M points, 3.6 GB)
- **AIS vessel data** (112M points, 29 GB)
- **Twitter** (292M points, 232 GB) (1B points, 146 GB)
- **Trace route data** (961M points, 157 GB)

Generate summary statistics of **all** the data, cross plots of every attribute against every attribute, box plots, geographic plots, frequency histograms of all attributes.

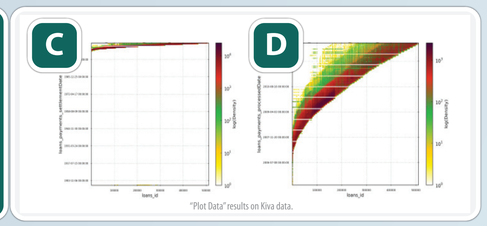
1 Scatter Plots for Big Data

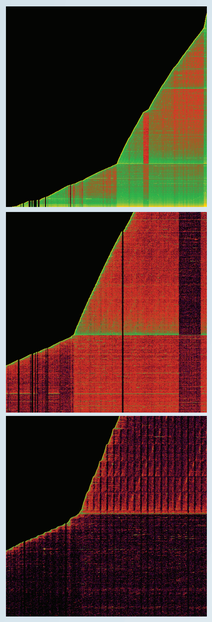
Scatter plots are an intuitive, easy to use, and widely understood tool for EDA. However, as the data plotted get larger, they su° er from **over plotting**, as pictured below, where true quantities and distributions become obscured.



Issue: Over plotting (A) addressed with **hexbinning** (B).

Issue: Showing all data with **extreme ranges** (C) **hides structure** of interest (D), but excluding them may hide important clues.

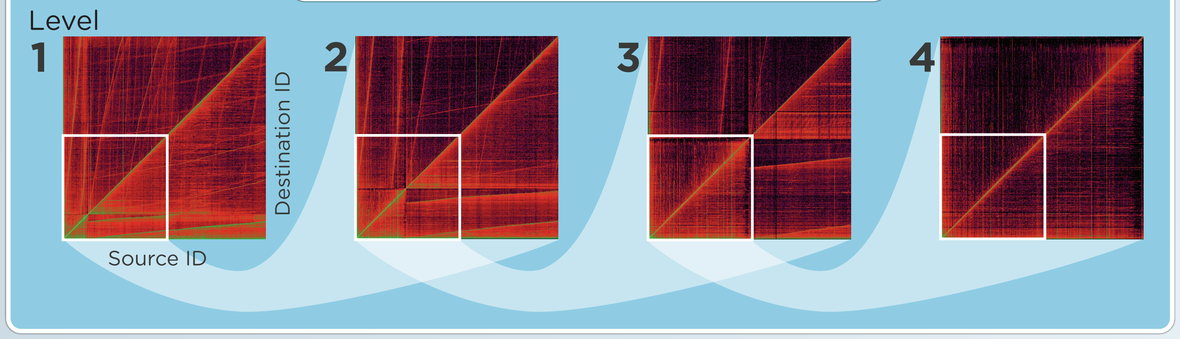




Hashtag occurences (vertical axis) over time (horizontal axis), with the hashtags ordered by time of ÿrst occurrence on the Y axis.

Zooming in (top to bottom) to knee in curve reveals variations in data collection frequency and methodology.

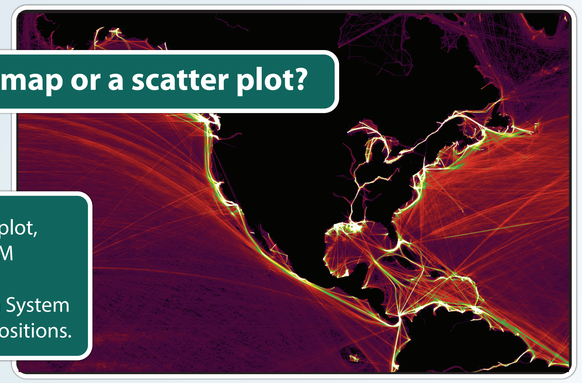
"Plot Data" results on *bitcoin* data.



2 Maps and Scatter Plots

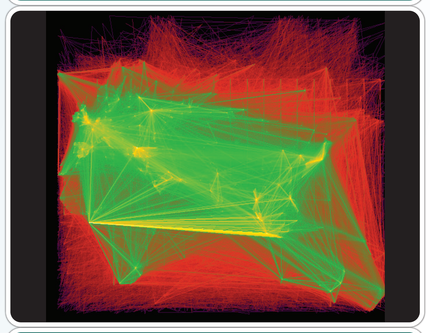
Maps share **much in common** with **scatter plots**. Both feature continuous data along two dimensions, use of layering and legends, axes and scales.

Is this a map or a scatter plot?



It is a scatter plot, showing 112M Automatic Identiÿcation System (AIS) vessel positions.

Scatter plots of geospatial data form maps. Figure shows geo-tagged tweet sequences, revealing travel patterns.



Anomalous visual artifacts: tweets at geo-location 0,0 (left), repeating vertical lines at 10 degree intervals (upper right).

3 Tile-based Visual Analytics for Situation Awareness

Aperture Tiles uses a multi-stage process to develop tile-based visual analytics. The ***aggregation*** stage projects and bins data into a predeÿned grid, such as the Tiling Map Service standard with a (z,x,y) coordinate system where z identiÿes the zoom level, and x,y identiÿes a speciÿc tile on the plot for the given zoom level. The ***summarization*** stage applies one or more summary statistics or other analytics to the data in each tile region storing the result in a data store. The ***rendering*** stage maps the summary to a visual representation, and renders it to an image tile or html at request time. Rendering on demand support interactions such as ÿltering, or rich dynamic interactions such as brushing and drill-downs if using client side rendering with html and JavaScript.

Interactive, configurable big data tiles encourages discovery and awareness.

We use Spark or R to calculate the tiles in a data view model. A high memory cloud conÿguration is used for computing, with 4 GB/ processor, 16 processors/ node and 20 nodes. Processing of a 4 GB dataset required 2.23 minutes to complete 42,191 tiles at level 9, and 4.62 minutes for 101,531 tiles at level 10.

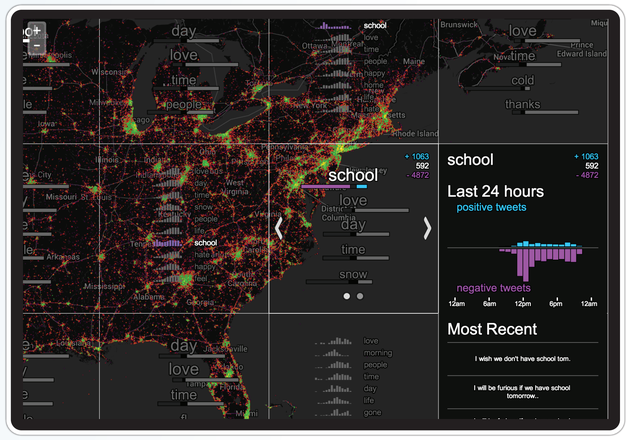
Next Steps

One next step will be to **compute feature extractors** (e.g. clusterers) and **annotate** the tiles with **additional layers**.

Pilot project deployment will further evolve the capabilities.

Tile-Based Alerts and Trends

Using multiple layers of tile data and tile carousels enables ¿pping through multiple per tile analytics. An "aggregate marker" layer provides an interactive carousel of visual analytic summaries of each tile region. Aggregate markers summarize and provide interactive visualizations that support analysis and integrate with advanced analytics.



Interactive Data Exploration with "Big Data Tukey Plots"
Schretlen P., Kronenÿeld N., McGeachie J., Hall E., Cheng D., Covello N. and Wright W., IEEE VisWeek, Oct 2013

Tile Based Visual Analytics for Twitter Big Data Exploratory Analysis
Cheng D., Schretlen P. and Wright W., IEEE Big Data Conference, Oct 2013

View a live demo at http://tiles.oculusinfo.com

http://tiles.oculusinfo.com

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Number FA8750-12-C-0317.
The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.

oculus
Oculus Info Inc