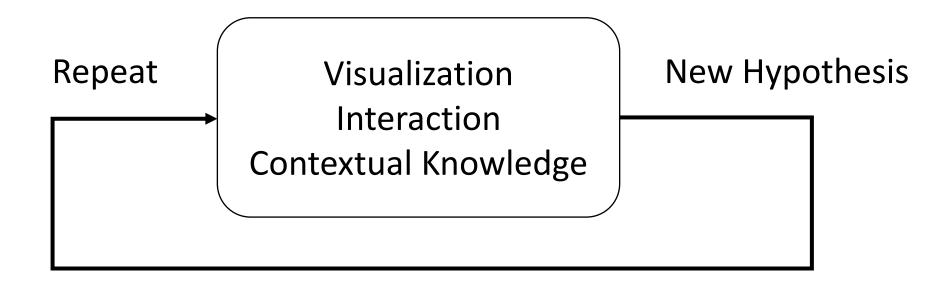
# Interactive Exploration of LEHD A Case Study in Knowledge Discovery

Saman Amirpour Amraii<sup>1,2</sup>, Michael Lewis<sup>2,3</sup>, Randy Sargent<sup>1</sup>, Illah Nourbakhsh<sup>1</sup>

- 1. CREATE Lab, Robotics Institute, Carnegie Mellon University
- 2. Intelligent Systems Program, University of Pittsburgh
- 3. School of Information Sciences, University of Pittsburgh

# Sensemaking Loop



This loop should happen fast, otherwise we lose our train of thought

## Sensemaking and Big Data

Data Production >>>> Data Consumption

How to visualize Petascale data?

How to visualize hundreds of dimensions?

How to make the process responsive? (delay cost)

What is a good question to start with?

### The Explorables Collaborative

http://explorables.cmucreatelab.org/

An effort to understand the challenges in visualizing, exploring, and analyzing large and complex data.

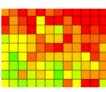
### **Explorables**

Explorables provide interactive and visual representations of large data sets, revealing patterns, encouraging discovery, and facilitating communication. The Explorables Collaborative, including CREATE Lab and SkyTruth, is dedicated to helping you make your information more impactful. Contact us!



### **Explorable Inequality**

Explorable visualization of the World Top Incomes Database illustrating global income inequality.



Speck Air Quality Test

Visualization of a Speck test showing how indoor air quality is effected by vacuuming, running an air purifier, and blowing out candles



### **FIRMS Timelapse**

Visualization of thirteen years of MODIS fire location vector data overlayed on an explorable global



### Suburban Sprawl

Video exploring the growth of suburban sprawl in Chesterfield County, Virginia using TimeMachine to show the full scope of changes.



Explorable visualization of AirNow PM 2.5 data showing changes in particle values and air quality across the United States



Time series visualization of over one million natural resource extraction wells in eight states spanning decades



### Fracking Earthquakes

Visualization of Landsat and USGS data, utilizing TimeMachine to explore the possible connection in Arkansas between earthquakes and natural gas drilling.

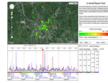


Multi-dimensional visualization of LEHD (Longitudinal Employer-Household Dynamics) data using EVA (Explorable Visual Analytics).



**Timelapse Story Telling** 

Visualization exploring landscape changes along Taiwan's coastline over two decades using TimeMachine



Visualization of air quality data collected by 64 Specks located in homes around the Pittsburgh area for a period of five days.



CREATE Lab | Carnegle Mellon University | Info@cmucreatelab.org

## Explorable Visual Analytics

http://eva.cmucreatelab.org

Goal: Improving <u>hypothesis</u> generation

How?

- Scaling: Interactive visualization of large datasets
- Navigation: Intuitive navigation in high dimensional space
- Responsiveness: Removing the delay between forming a hypothesis and seeing the visualization → Aiding our limited working memory

### EVA Demo<sup>↗</sup>

Data? large, complex, high spatial and temporal resolution, opportunities for real and meaningful discoveries

Census Longitudinal Employer-Household Dynamics (LEHD) <a href="http://lehd.ces.census.gov/">http://lehd.ces.census.gov/</a>

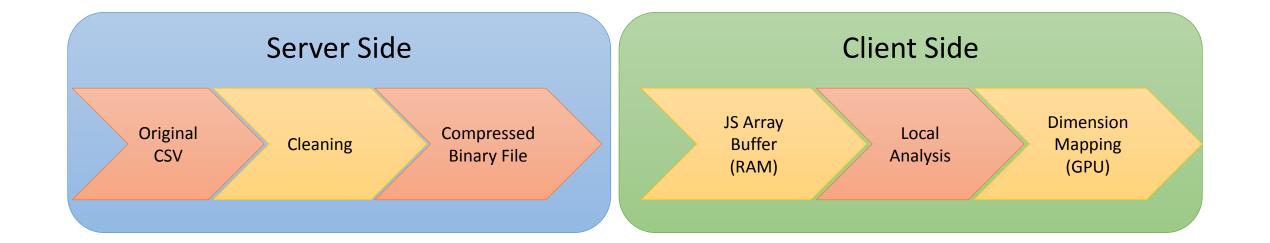
Pennsylvania section (~2.8 million data points with 45 dimensions), Census Blocks geospatial resolution, spanning over 10 years

## Technical Aspects

**Technologies:** Web-based, JavaScript, HTML, Three.js & WebGL (<a href="http://threejs.org/">http://threejs.org/</a>)

Open Source: <a href="https://github.com/nebeleh/EVA">https://github.com/nebeleh/EVA</a>

Current Capability:  $4^5$  M points with 10s of dimensions ( $1^2$  GB of RAM)



### Lessons Learned

**High Resolution:** Knowledge discovery is highly dependent on the amount of details a user can see.

**Explorability:** Seeing the data from multiple perspectives increases the chance of recognizing unexpected patterns. This can be beneficial in the formation of new hypotheses and possible new discoveries.

**Responsiveness:** The glue! Facilitates an uninterrupted train of thought.

All of these aspects improve hypothesis generation, leading to more chances for knowledge discoveries.

### Next Steps: Human Data Interaction

### Inaccurate but fast vs. accurate but slow:

screen-aware solutions which benefit from our cognitive limits

### Non-episodic interaction:

steering and active feedback, interactive query building

### Interactivity:

compensation for our inability to perceive high-dimensional space

# Thank You

samirpou@andrew.cmu.edu