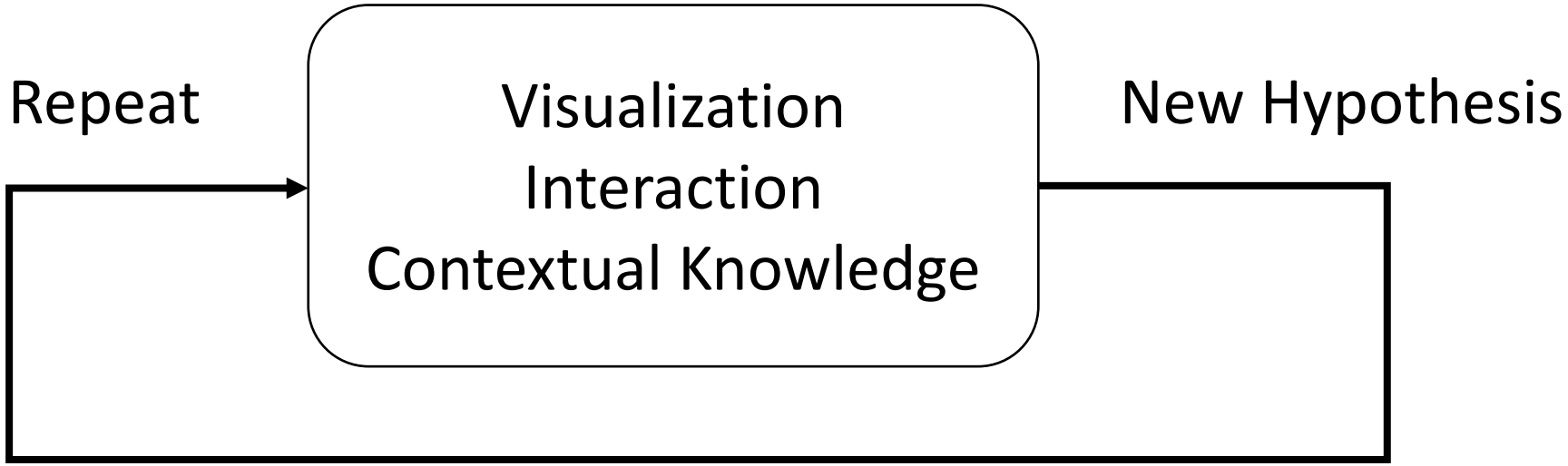
Interactive Exploration of LEHD
A Case Study in Knowledge Discovery

Saman Amirpour Amraii[1,2], Michael Lewis[2,3], Randy Sargent[1], Illah Nourbakhsh[1]

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?

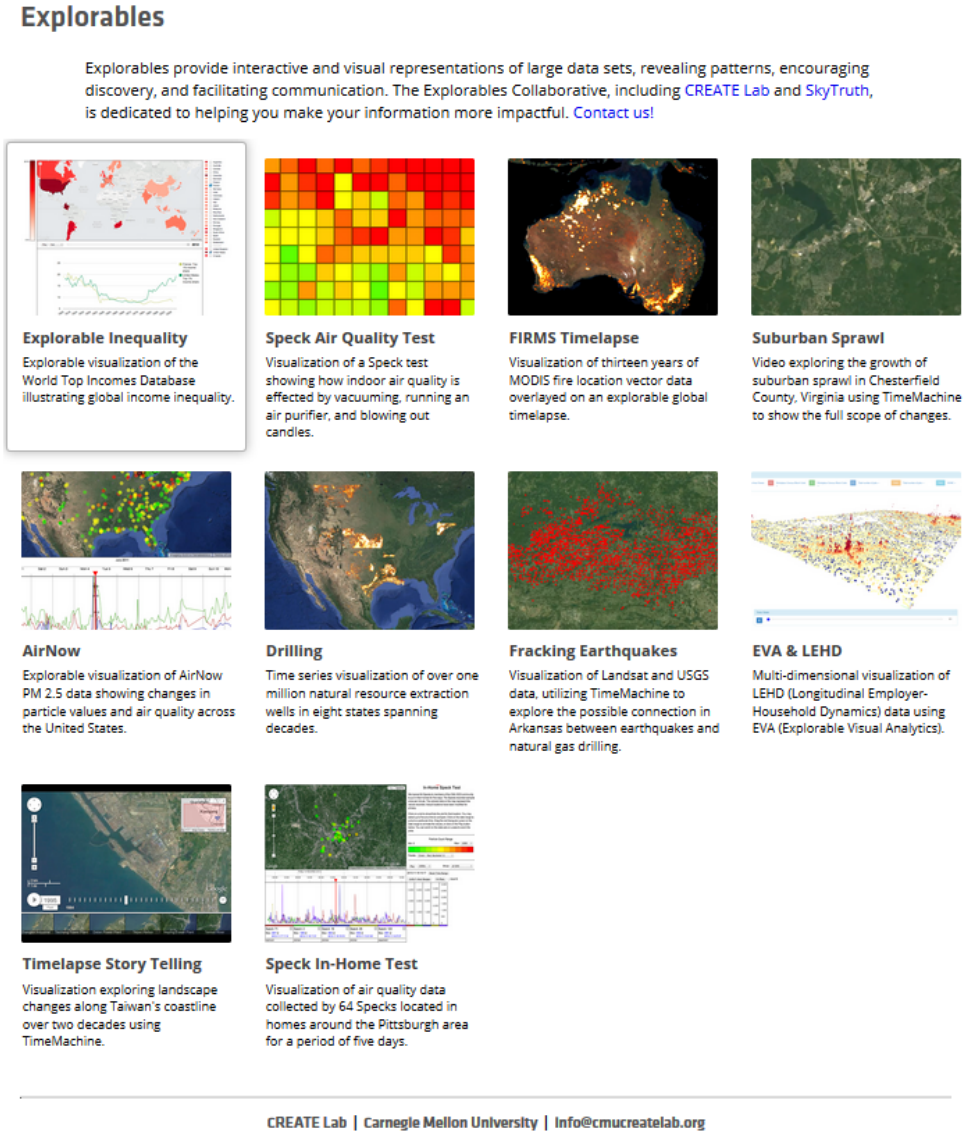
How to make the process responsive? (delay cost)

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.



Explorable Visual Analytics

http://eva.cmucreatelab.org

Goal: Improving hypothesis 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↗

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

Census Longitudinal Employer-Household Dynamics (LEHD)
http://lehd.ces.census.gov/

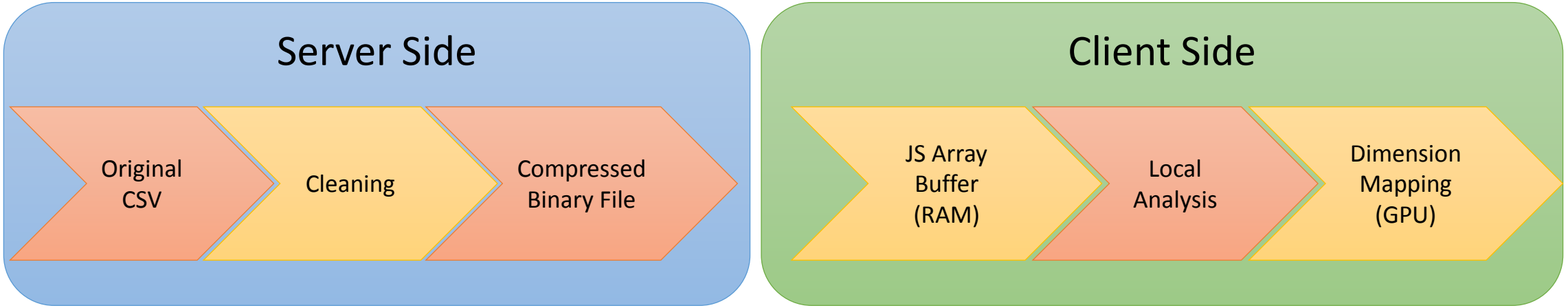
Pennsylvania section (~2.8 million data points with 45 dimensions), Census Blocks geospatial resolution, spanning over 10 years

Technical Aspects

Technologies: Web-based, JavaScript, HTML, Three.js & WebGL (http://threejs.org/)

Open Source: https://github.com/nebeleh/EVA

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