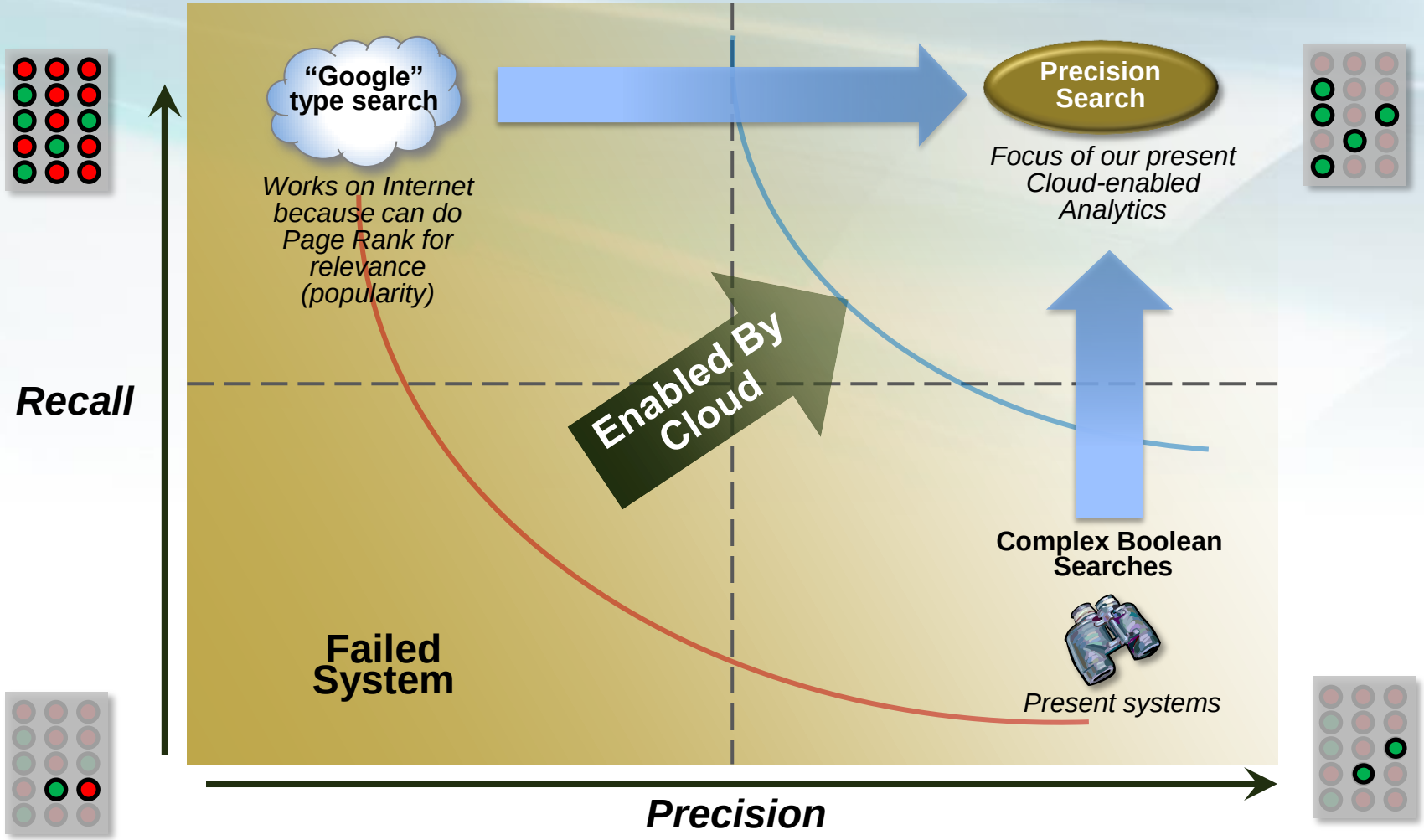
SOTERA DEFENSE SOLUTIONS

Data in the Aggregate: Discovering Honest Signals and Predictable Patterns Within Ultra Large Data Sets

Kathleen Lossau and Jonathan Larson

April 2013

Getting Precision Search from Big Data



Source: Dr. Russell Richardson – Chief Architect & Senior Science Advisor U.S. Army INSCOM

Sotera Defense Solutions -
Kathleen.Lossau@soteradefense.com

Why Data Science

- Forbes - How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did
- NYT - What Does Your Credit-Card Company Know About You?
  - Canadian Tire store (which sells electronics, sporting goods, kitchen supplies, automotive goods)
  - Credit ratings from purchasing activities
  - Chrome-skull hitch devices
  - Premium birdseed and “snow roof rakes” which helps protect pedestrian that walk by

His data indicated, for instance, that people who bought cheap, generic automotive oil were much more likely to miss a credit-card payment than someone who got the expensive, name-brand stuff. People who bought carbon-monoxide monitors for their homes or those little felt pads that stop chair legs from scratching the floor almost never missed payments. Anyone who purchased a chrome-skull car accessory or a “Mega Thruster Exhaust System” was pretty likely to miss paying his bill eventually.

http://www.nytimes.com/2009/05/17/magazine/17credit-t.html?pagewanted=all

As Pole’s computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

Sotera Defense Solutions -
Kathleen.Lossau@soteradefense.com

Analysis Process: How We Do It

SOTERA DEFENSE SOLUTIONS

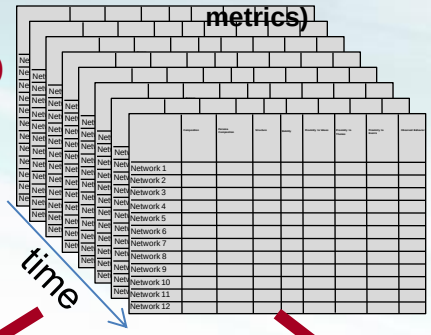
Raw Data: COMINT, OSINT, IMINT, …

1 Process raw data into features:

- SNA metrics (e.g., number of contacts, centrality)
- cultural metrics (e.g., proximity to themes, proximity to beliefs)
- Census features (e.g., age, ethnicity, education)
- Process over time

Features (e.g., metrics)

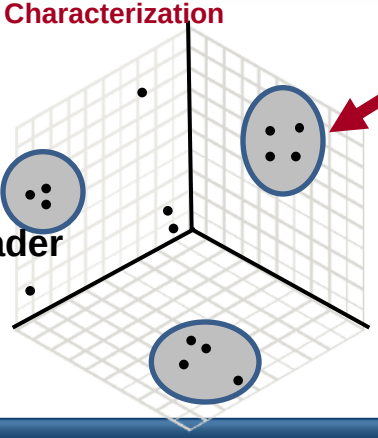
Metrics Over Time (Object Assessment)



2 Create aggregated signatures

- Features mapped to dimensions
- Entities mapped to points
- Similar behavior causes entities to be near each other in this space

Clustering & Characterization



Signature Clustering

3 Analyze and track aggregate personas

- Each cluster is a persona
- Some personas may be associated to known roles, e.g., cell leader, recruiter
- Some personas may be prominent, but not yet understood
  - Target these for further analysis

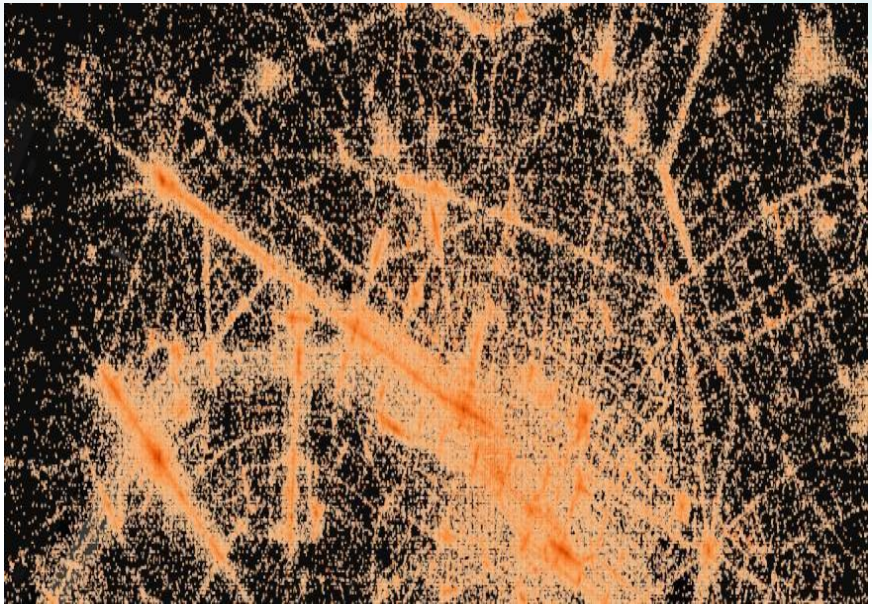
4 Identify / monitor salient personas

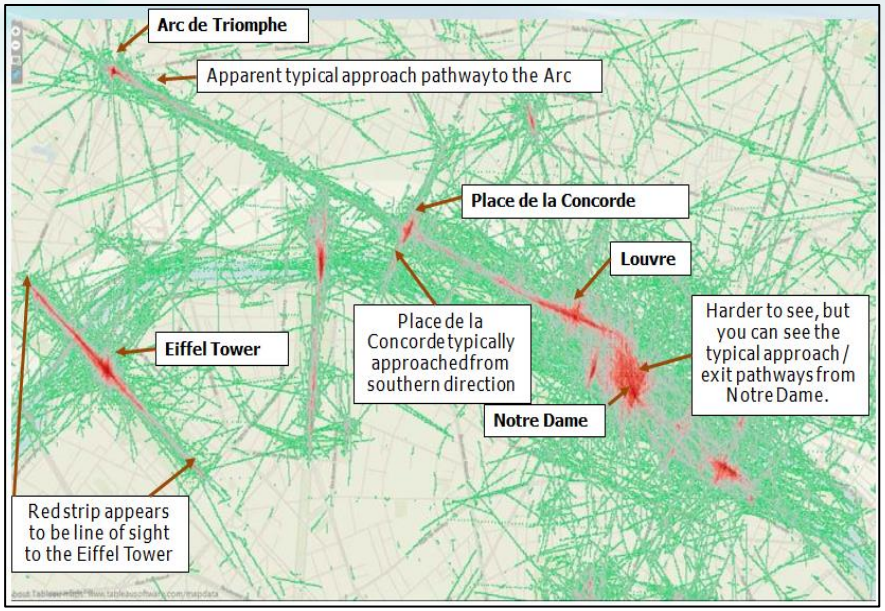
- Once persona types are known, system can detect entities that match this pattern
- Track the persona makeup of each network
- Identify individuals who pick up new personas or who's persona type changes
- Detect trends, identify entities at risk
- Provide situational awareness

Sotera Defense Solutions -
Kathleen.Lossau@soteradefense.com

Inferring Movement From Points

- How can we infer movement patterns from vast amounts of what appears to be just point data collected over time and associated with a distinct identifier (*e.g., a* user ID, bank account number)?
- *Aggregate Micro-paths* - Technique is applicable to Twitter, FourSquare and MANY other sources.





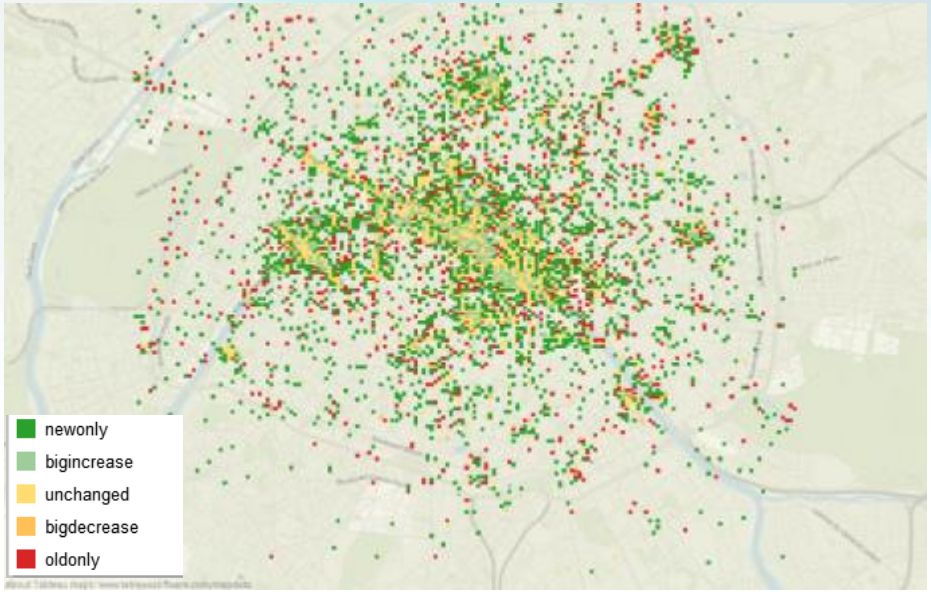
Volume plot of photos binned by area on log scale — Paris, France as seen from Flickr over all time.

Sotera Defense Solutions -
Kathleen.Lossau@soteradefense.com

Tool/Technique:

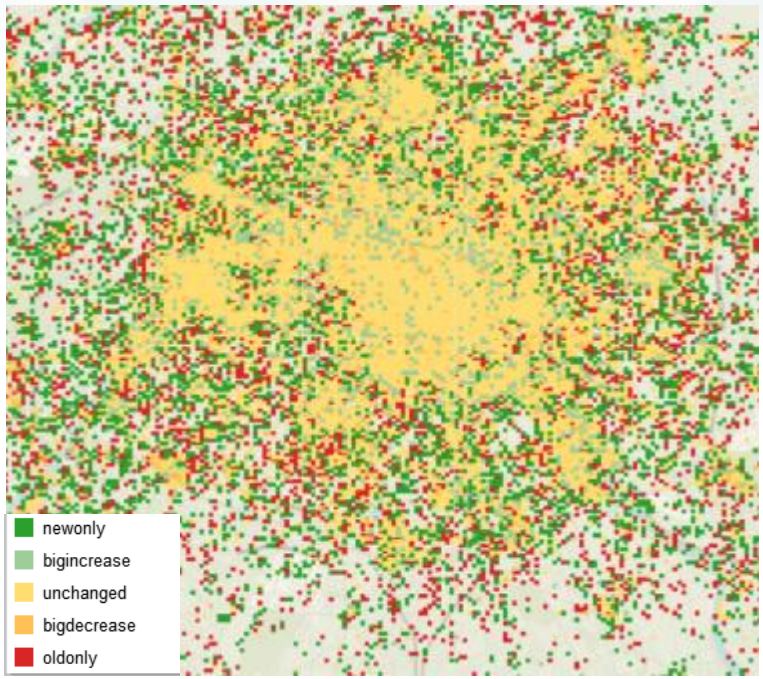
Temporal Difference Analysis

Flickr Paris 2004 changes vs 2005



Measuring the relative change of photographic activity year over year. Significant changes between years denoted by shades of green and red.

Flickr Paris 2011 changes vs 2010



Sotera Defense Solutions -
Kathleen.Lossau@soteradefense.com

Tool/Technique: 3D Hourly Tripline Blankets

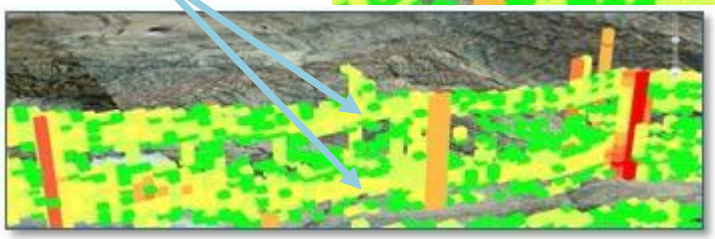
- All pings are assigned to a latitude / longitude cell and aggregated
- Additionally, all pings are binned to an hour of day on the Z-Axis **across all days**

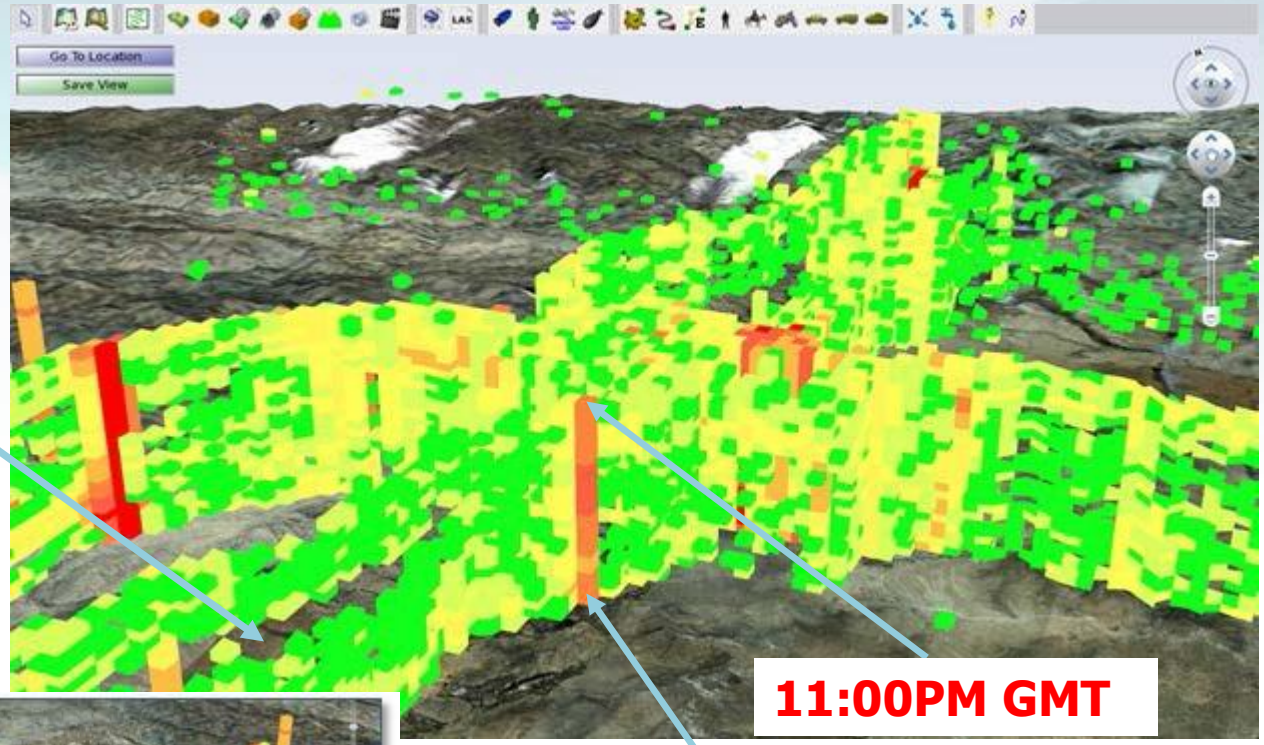
Legend (log scale):

Fewer Pings — More Pings

No pings on this road during the middle of the night (shown as absence of reading)

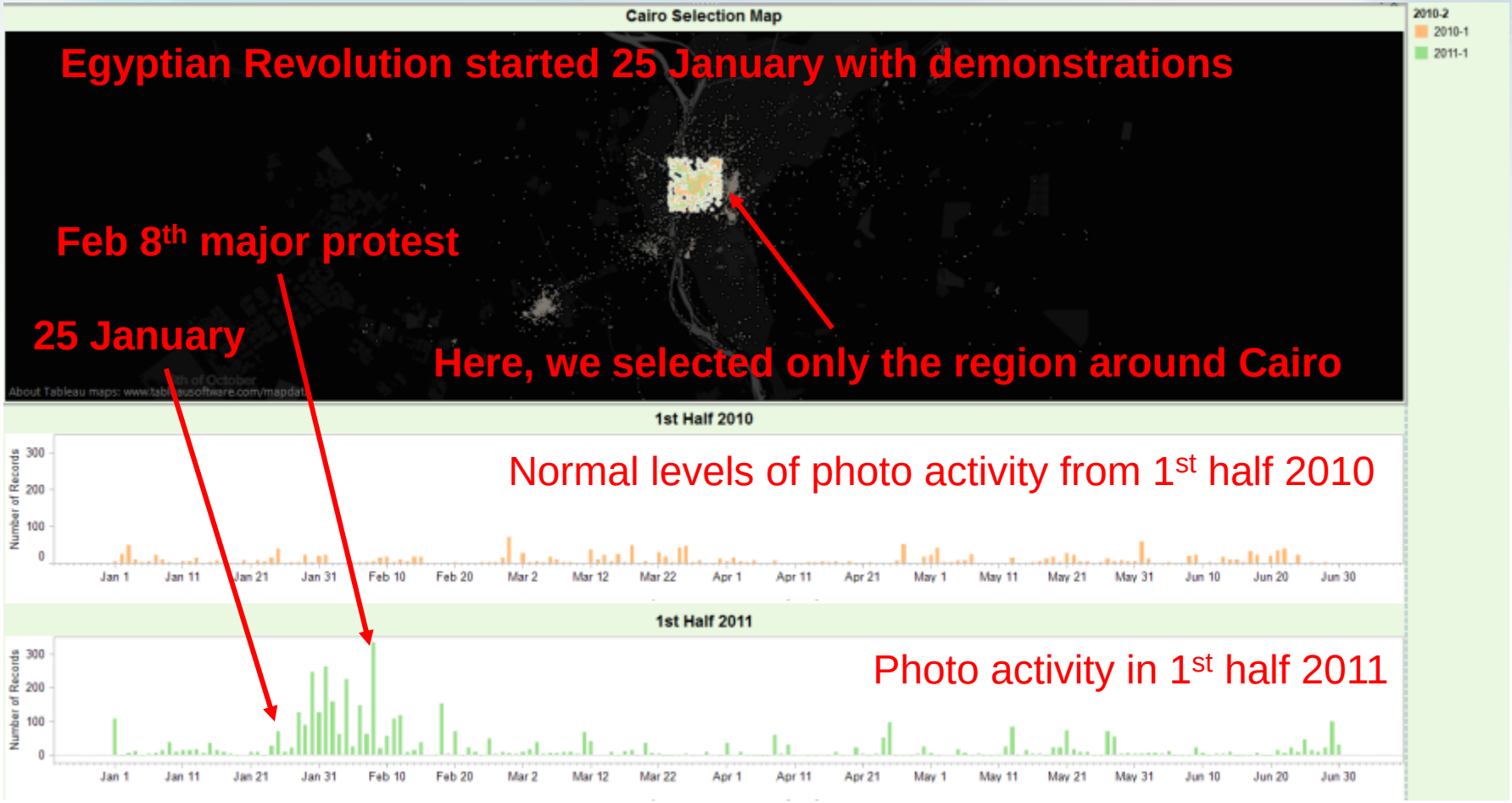
Notice the rush hour





Tool/Technique:

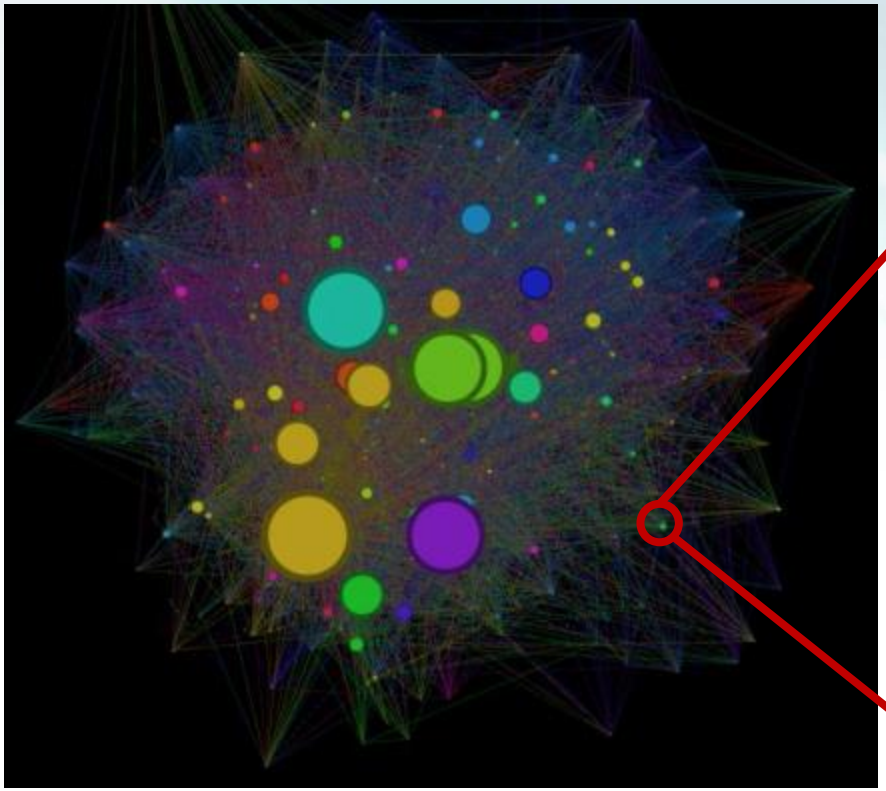
Time - Event Characterization

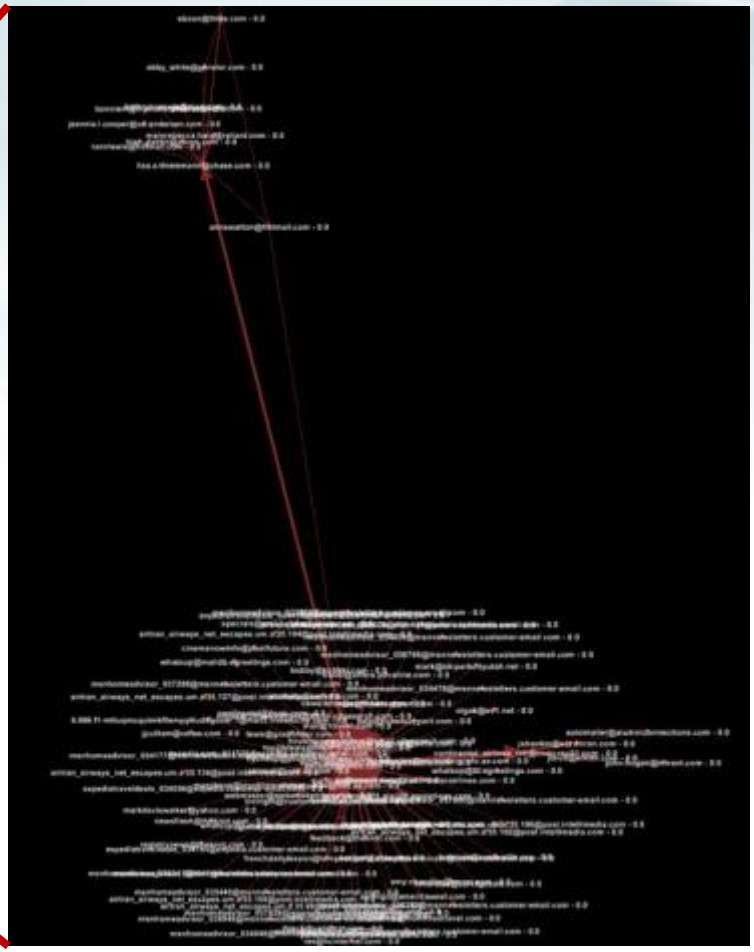


Sotera Defense Solutions -
Kathleen.Lossau@soteradefense.com

Zooming into a community

- This functionality allows actual browsing of enormous graphs





Zoomed snapshot of sscott5@enron.com's community

Sotera Defense Solutions -
Kathleen.Lossau@soteradefense.com

Conclusion

- Previously the job was to filter through large sources of data to find specific pieces of information that fused together tell a picture - Now it is the large data itself that is the product
  - Information fused on different levels reveals patterns and trends within a given slice of the data.
  - The challenge is in finding the right people to excavate the relevant dimensions within the data to create meaningful and relevant aggregated data products.

- Analysts will increasingly be looking at aggregated data products consisting of multiple sources of data fused together to provide an understanding of normal patterns of behavior.
  - Look at trends - products can be created on behaviors and other dimensions of communities, regions, large corporations, and ethnic, religious, cultural organizations.
  - Compare incoming (streaming?) data against known patters and trends to quickly find anomalies

Sotera Defense Solutions -
Kathleen.Lossau@soteradefense.com