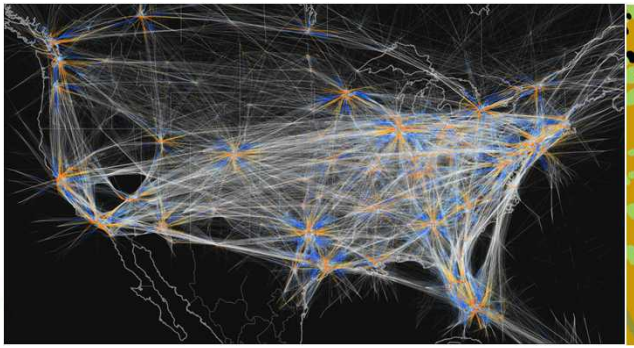
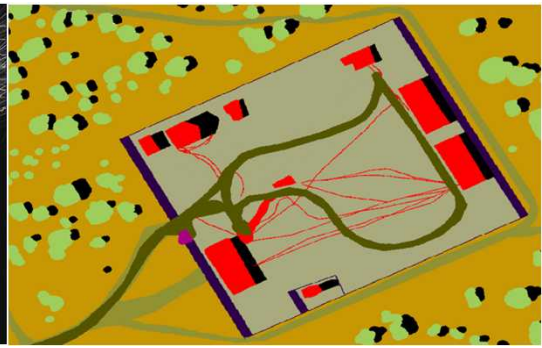
Exceptional service in the national interest

SAND2014-18850PE

Sandia National Laboratories

LDRD
LABORATORY DIRECTED RESEARCH & DEVELOPMENT





Previous UUR determinations
Derived from SAND #
SAND2014-17938 PE

Pattern Analytics

Kristina Rodriguez Czuchlewski, Ph.D.

Principal Investigator

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

The changing data landscape presents challenges… and opportunities

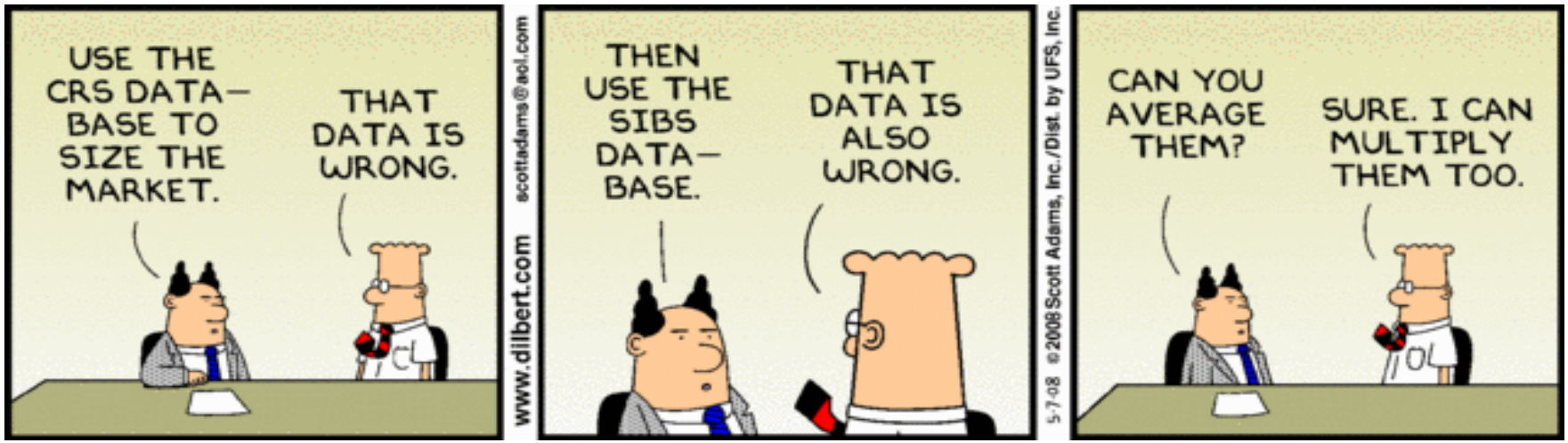
Sandia National Laboratories

- Many important phenomena below limit of human perception—in nearly every domain
- The phenomena are scaling much faster than the ability to observe and process them
- Key connections between observables cannot be made
- Overwhelmed operators struggle to use data for predictive and forensic purposes—especially in real time
- Data transmission and storage limitations

These problems demand multi-disciplinary scientific inquiry

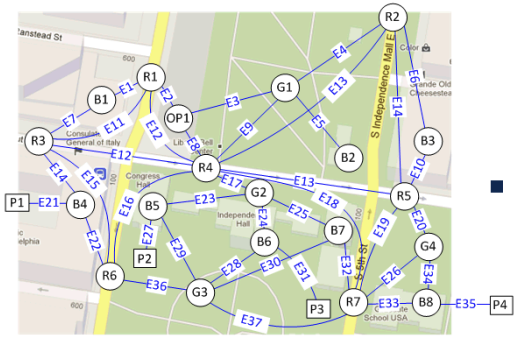
Contributors to current limitations:

- Only a small fraction of the data is ever examined by analysts. Workflows are labor intensive and devoid of effective computational tools.
- Systems do not exploit the relationship discovery potential of the data or identify **meaningful, defensible** trends and patterns.



Pattern **AN**alytics **T**o support **H**igh-performance **E**xploitation and **R**easoning

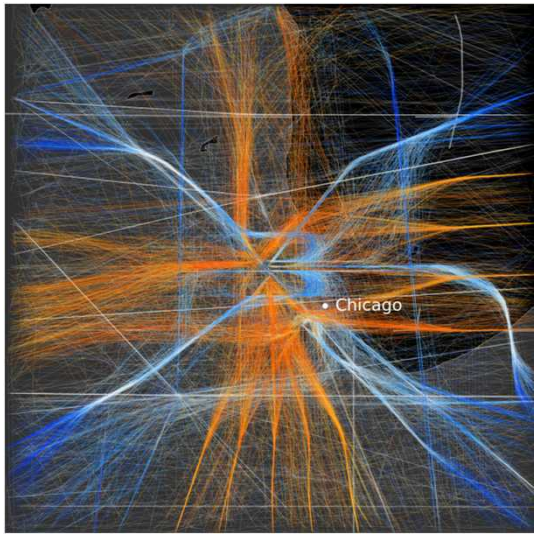
Enabling Search for Threat Signatures



- PANTHER **creates searchable features from imagery** to support analysts' temporal and geospatial pattern analysis demands. Existing approaches miss key threat signatures because they cannot exploit image features in a form compatible with relationship discovery or trending.
- PANTHER **discovers relationships under noise and uncertainty conditions** by developing a new geospatial-temporal semantic graph framework. Existing pattern analytic approaches are ineffective on noisy and intermittent sensor data.

Discovering Trajectory Outliers

PANTHER's geometric and temporal trajectory analyses **can now represent and compare tracks** efficiently and lightning-fast. Existing approaches for discrete motion are limited; they do not enable discovery of geospatial-temporal relationships or comparison of more than two tracks and trajectories.



Better knowledge & deeper insights from Big Sensor Data in minutes, not months; over months, not hours or minutes; covering hundreds, not 10s of km²

R&D Challenge Questions

Sandia National Laboratories

- Where are chemical processing plants? — Signature Search
- Where are active businesses?
- Did someone arrive in a car and enter a building? Which one(s)?
- Which aircraft flights are point-to-point? — Trajectory Analysis
- Are any aircraft flying search patterns?
- Are any aircraft flying search patterns over sensitive locations?
- Is there an activity surge? — Temporal Patterns
- Is an aircraft flight pattern departing from normal? What might it do next?
- What can we infer given limited data? — Limited Data
- For any of the above, what is the result confidence? — Uncertainty/Quality

Current Paradigm → PANTHER Goal

Sandia National Laboratories

	State of the Art	Sandia R&D	PANTHER Grand Challenge
Approach	Single frame or target recognition	Statistical, pixel-level processing for specific tactical missions	Identify patterns/ relationships and activities over multiple target classes and extend to strategic missions. Achieve 10^7 reduction in data volume in space & time.
Spatial	Single frame	Specific areas (10s of km^2)	Achieve 10^4 single frame reduction in data presented. Query and exploit 100s of km^2
Temporal	Minutes of data, real-time and offline.	Weeks of information, available in minutes	Pattern based query of up to 1 year's worth of data. Available in minutes.
Analysts	Eyes-on-pixel, image-by-image visual inspection, manual "fusion"	Behavioral studies capture implicit threat patterns. Product oriented.	Analyst Centric Design. Visual analytics supports data store overview, filtering-and-zooming, details and data provenance on demand.
Example	ATR, CCD, matched filters	Statistical Normalized Coherence	Object-based image analysis: features, entities, patterns and relationships. Find new answers to old/ new questions.
Will it work?	Limited, does not scale.	Limited, scales for specific missions.	On Track to demonstrate scalability & extensibility to other domains.

PANTHER's Status

Sandia National Laboratories

	PANTHER Grand Challenge	Signature Search	Aircraft and Maritime Trajectory Analysis*
Approach	Identify patterns/ relationships and activities over multiple target classes and extend to strategic missions. Achieve 10^7 reduction in data volume in space & time.	Expected FY15	Achieved.
Spatial	Achieve 10^4 single frame reduction in data presented. Query and exploit 100s of km^2	Achieved by NA-22 and PANTHER in FY14.	TBD calculation for appropriate spatial state of the art.*
Temporal	Pattern based query of up to 1 year's worth of data. Available in minutes.	Query possible FY15, availability a stretch.	Achieved.
Analysts	Analyst Centric Design. Visual analytics supports data store overview, filtering-and-zooming, details and data provenance on demand.	Expected FY15	Identify appropriate analyst domain; build technology transfer relationship.*
Example	Object-based image analysis: features, entities, patterns and relationships. Ask questions that cannot currently be answered.	Expected FY15	Achieved.
Will it work?	Unproven. Scalable & extensible to other domains.	Expected FY15	Yes.

*Note: analyst-centric performance calculation is TBD.

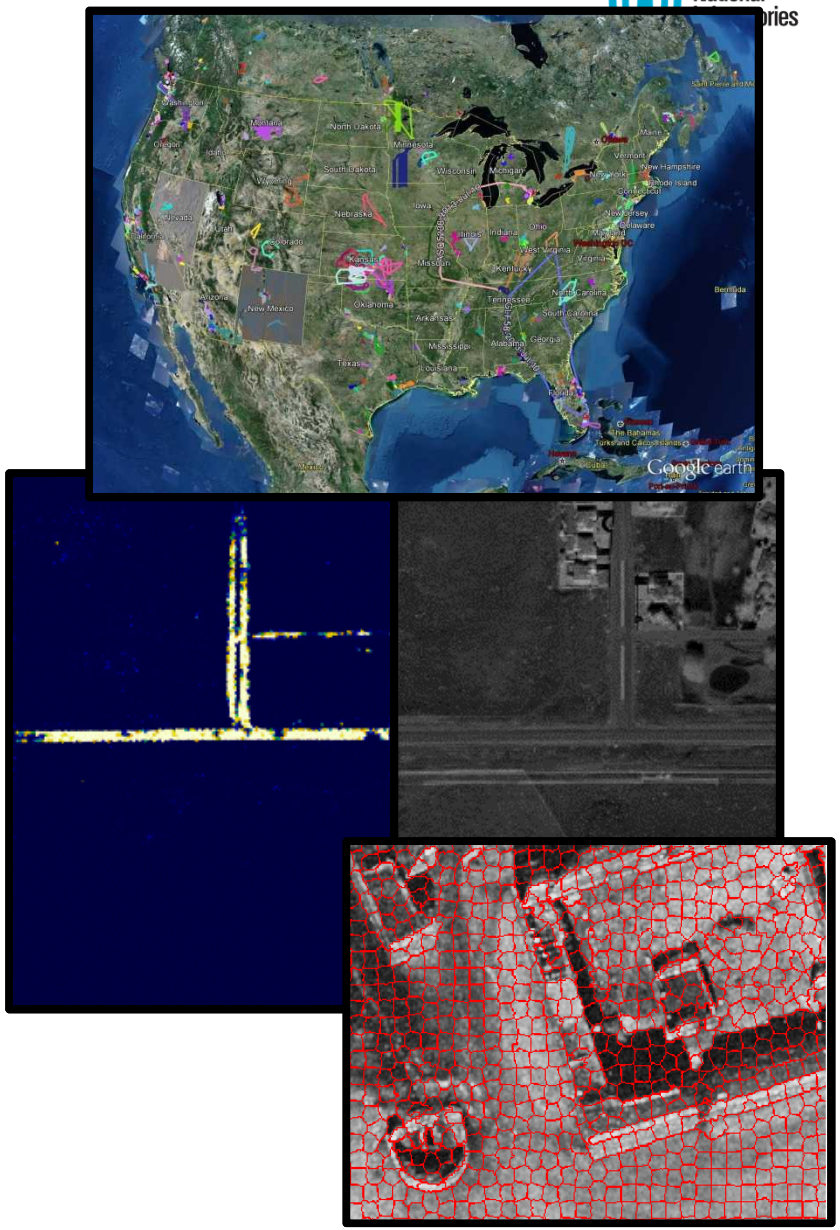
Approach and Metrics for Success

Sandia National Laboratories

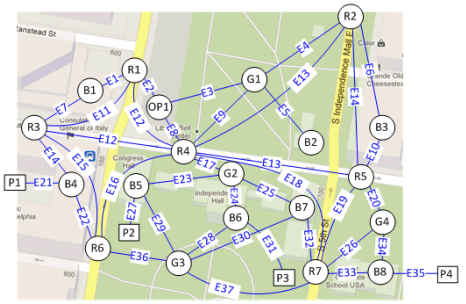
- We develop new mathematical approaches to this problem with a deep understanding of human capabilities and the information landscape
- Goals:
  - Foundational Scientific Research and Technology Development
    - Conference Publications and journal articles
    - Technical Advances (TA)/Patent Applications (PA)/ Software Copyright (CR) Assertions
    - Research Partnerships
  - Practical Technologies
    - PANTHER prototypes and demonstration systems
    - PANTHER System Performance (vs. current manual search, retrieval, and classification techniques)
  - Human Performance Studies
    - PANTHER Human-System Evaluation (vs. current manual search, retrieval, and classification techniques)

FY14 Highlights

- Unsupervised computational methods for detecting outliers (with no *a priori* knowledge) within a TB-scale database… in under 30 minutes.
  - New geometric feature vector enables comparisons.
  - Result: 44 seconds on our Netezza DB machine to compute a trajectory segmentation from points.
  - Result: 20 minutes of wall-clock time to turn 1.2 billion points into 15 million trajectories.
- Geospatial feature extraction and classification for signature relationship and temporal trending searches.
  - Pixel statistics extracted via new superpixel segmentation algorithm.
  - Unique temporal attributes of coherent changes exploited for labeling.



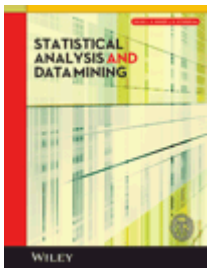
FY14 Highlights, cont.



- New, efficient graph algorithm formulation developed to enable geospatial-temporal topological search complexity.
  - Multi-source data search under one framework
  - Representation and search under heterogeneous temporal conditions (ephemeral and activity).
  - Complex sensor feature data “compressed” for analysis with pointers back to original sensor source.
  - Intermittency/ interruption nodes implemented.
- Eye-tracking experiments are enabling visual cognition/ search models – and eventual user interface design(s).
- Pattern match quality, statistical and probabilistic approaches are under investigation for characterizing uncertainty in geospatial temporal graph representations.

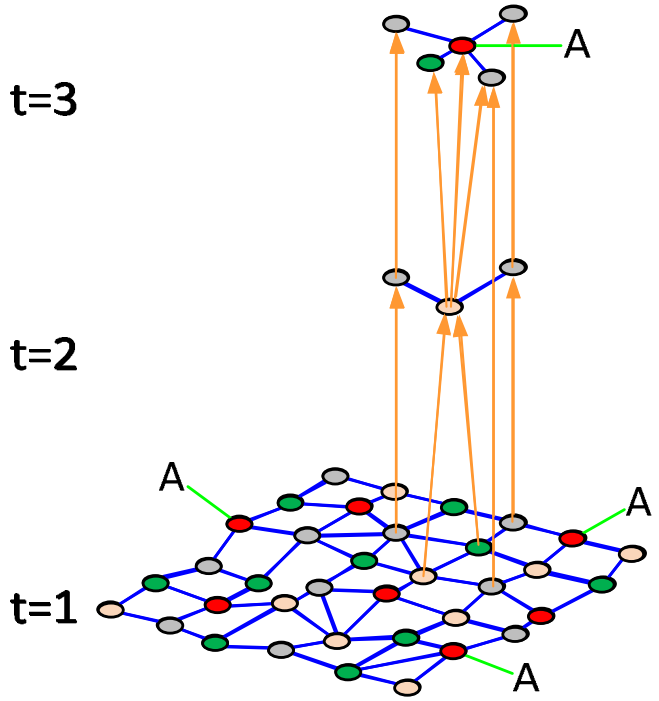
Intellectual Property

- 15 TAs, 3 Patent Applications
- 9 Peer Reviewed Publications
- 6 Conference Posters
- 7 Total SAND Reports/ Whitepapers
- 5 Manuscripts in Preparation
- 6 Presentations, 4 invited
- 8 Software Artifacts
- 9 Solidified R&D Relationships
- 8 R&D Projects, 4 funded so far
- 12 Service Roles for PANTHER Staff



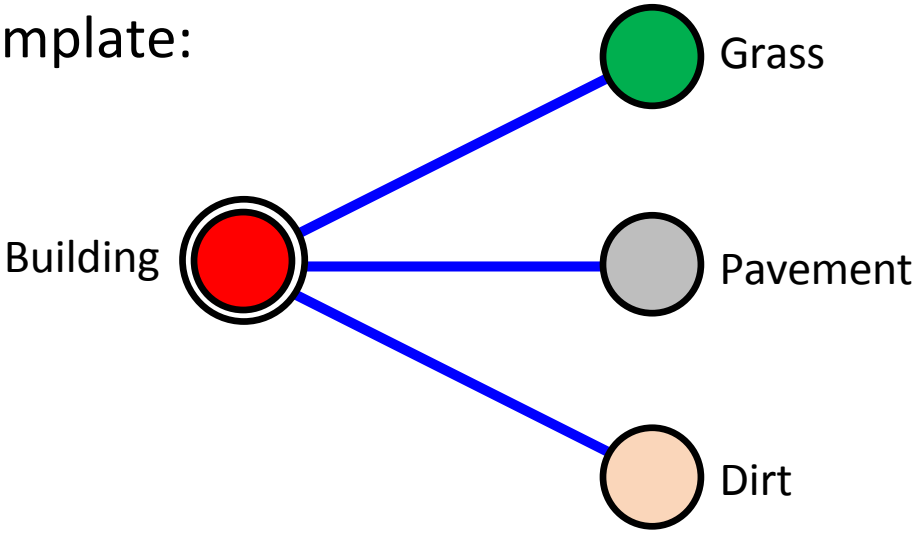
GeoSpatial Semantic Graph Representations of Features & Activity

- Graph includes activity:
  - @ t=1, the graph includes objects with location
  - From t=1 to t=2, the graph encodes change
  - Nodes for activity events.[1]
  - Node attributes include time observed.
  - No persistence expected.
  - Spatial and temporal relationship edges.[1]



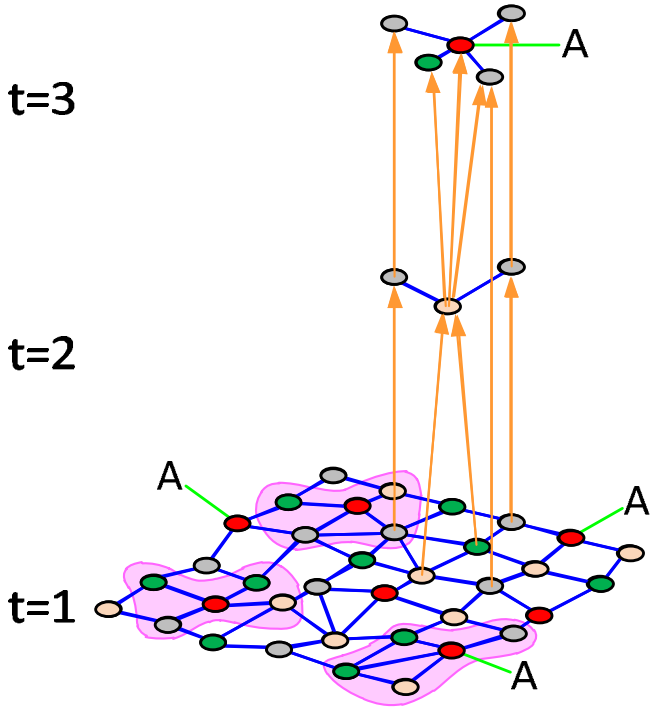
Signature Search

- A signature (over space/ time) encodes a desired question.
- For example, “Where are buildings with nearby grass, pavement, and dirt?
- Graph template:



Matches

- Graph search finds all matches to signature template.
  - In this case all red nodes with adjacent green, grey, and tan nodes.
- New approach to graph search, polynomial time complexity
- Searches are saved



Advantages of a Graph

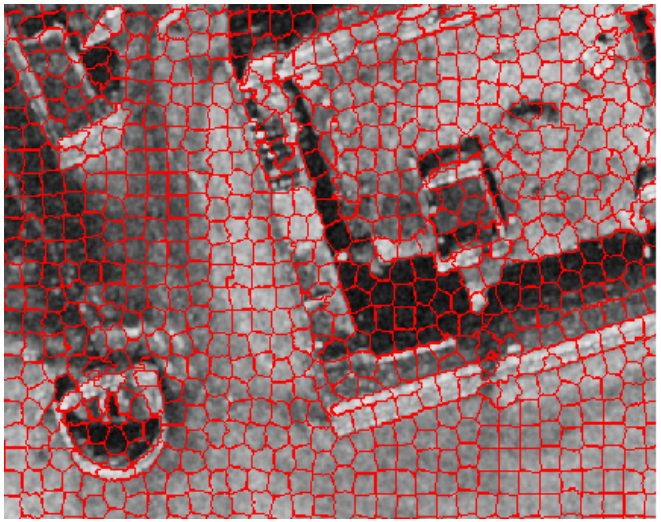
Sandia National Laboratories

- Efficient representations in time (only store change).
- Relationship, change, and temporal analysis and heterogeneous spatial ensembles in the same query.
  - Change detection.
  - Activity characterization.
- Combines direct analysis of remote sensing imagery, database query filtering, and graph search algorithms within one framework.
  (e.g, SQL queries cannot solve graph transitive closure operations.)
- Able to take full advantage of graph topology search, enhanced by geospatial-temporal semantics.
- Feature-based analysis:
  - Multi-modality, in a single search representation.
  - Sensor agnostic.

Extracting Statistical Distributions Using Superpixel Segmentation

Sandia National Laboratories

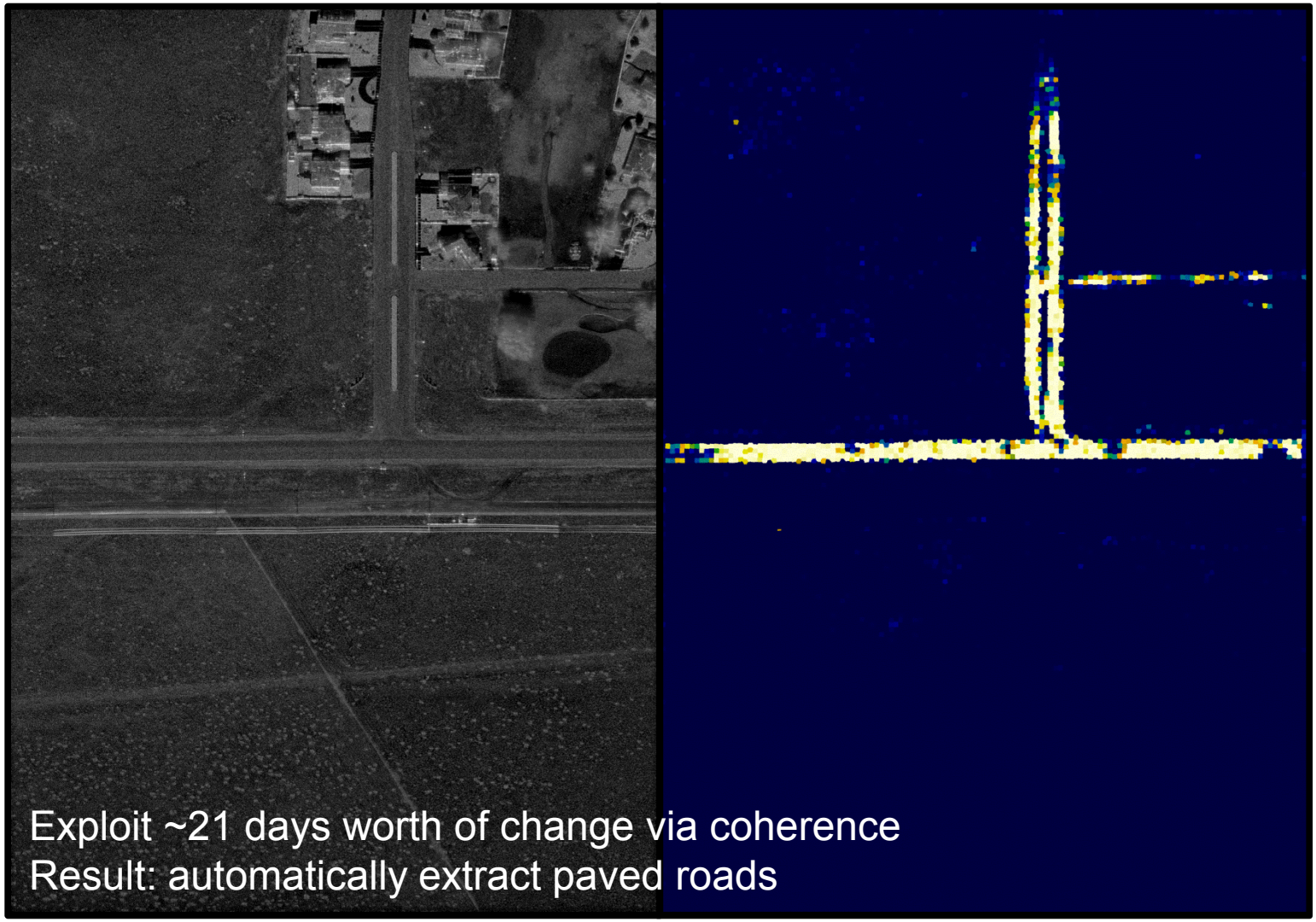
Superpixel Segmentation of SAR Image



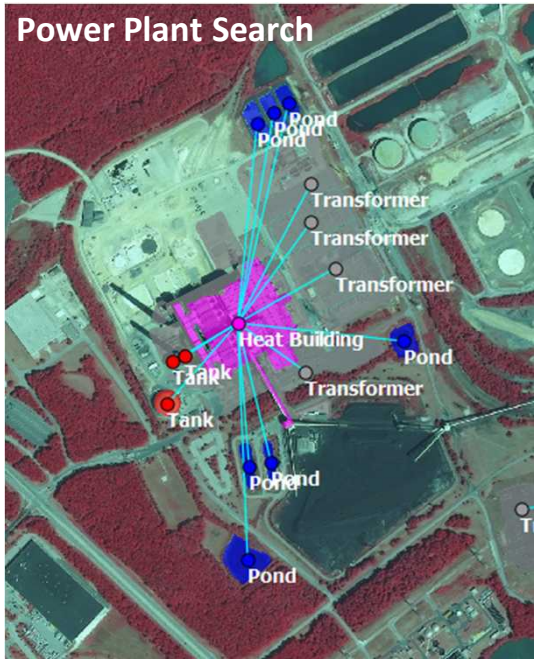
- **Superpixel Segmentation**
  - Divides image into compact regions containing pixels similar in spatial proximity and intensity
  - Derives from large body of research in optical image processing community extended to SAR imagery
- **PANTHER Approach**
  - Superpixel segmentation algorithms enable high quality segmentations and efficient execution
  - De-noising advances enable utilization of novel image processing techniques.

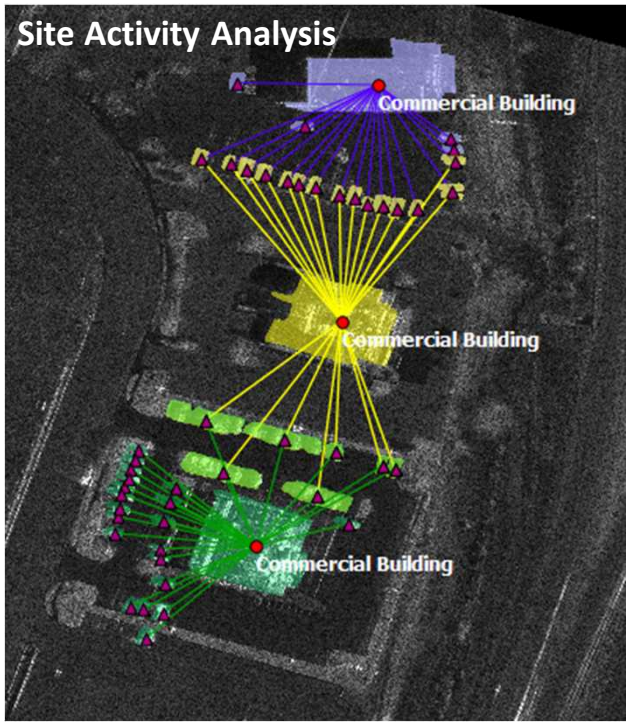
Initial Automated Static Feature Extraction Result

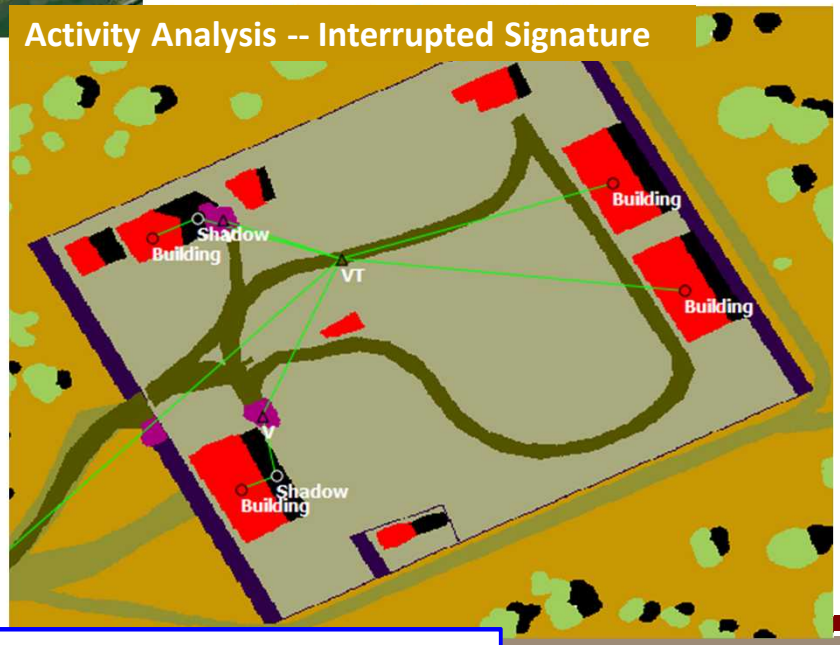
Sandia National Laboratories



Diversity of Problems



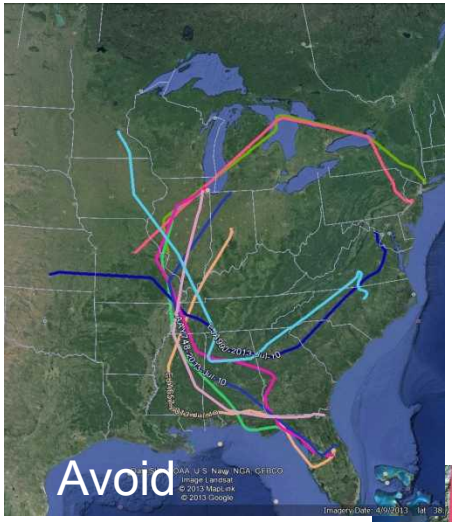


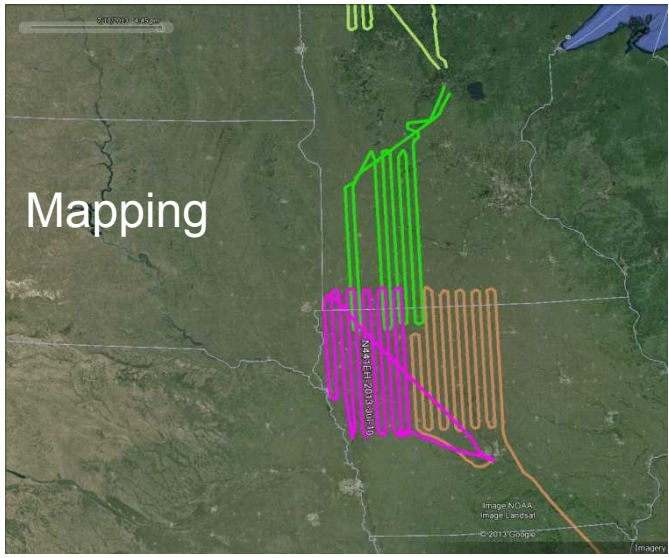


All of these were solved by the same code.

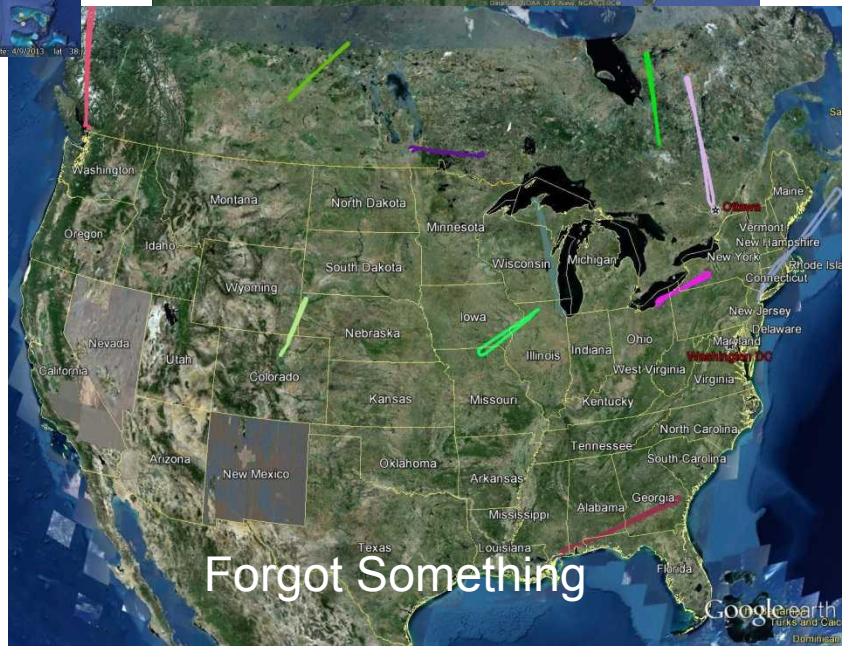
Discovery of Flight Patterns

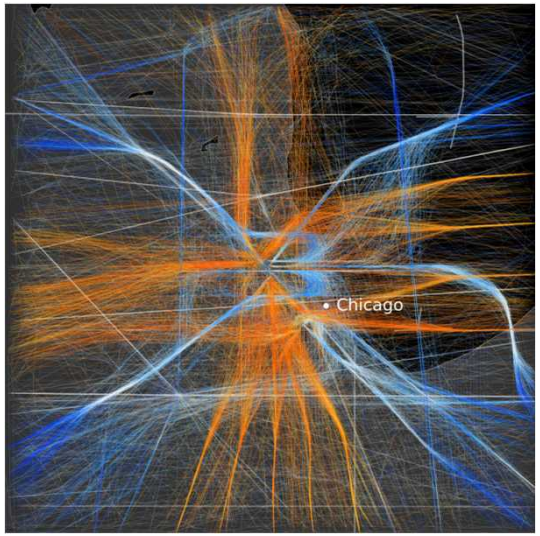
Sandia National Laboratories





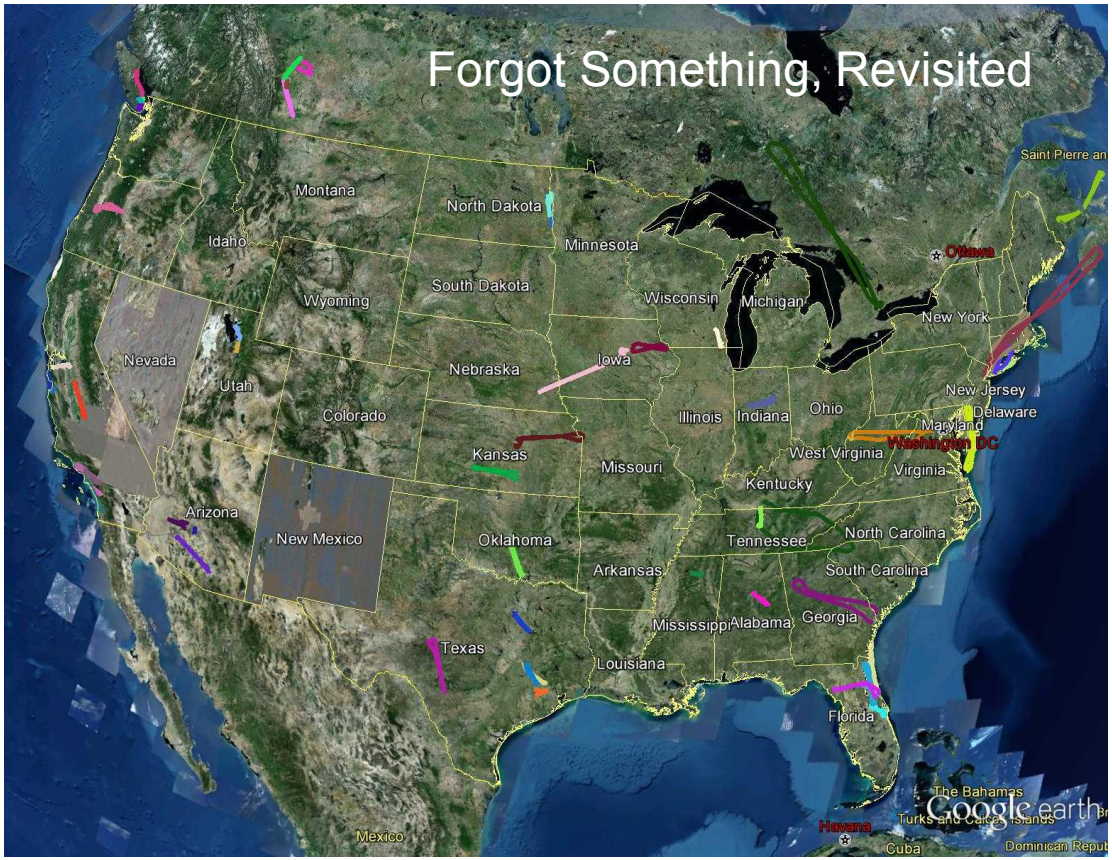
Collections of geometric descriptions can describe a trajectory. Extensions: impact of time.





Big Feature Space Advantage

- Clustering, leading to *unsupervised learning* techniques
- Previous examples showed searching the space for a specific pattern or a specific volume of the feature space
- But, with clustering, the computer can group the different patterns in the feature space without knowing *a priori* what they are.
- Perhaps most importantly, many clustering algorithms specifically identify outliers in the feature space that correspond to odd behaviors



Did not specify “find this,” only told routine to “make groups of similar flights.”
This was one of many clusters that had distinctive shapes

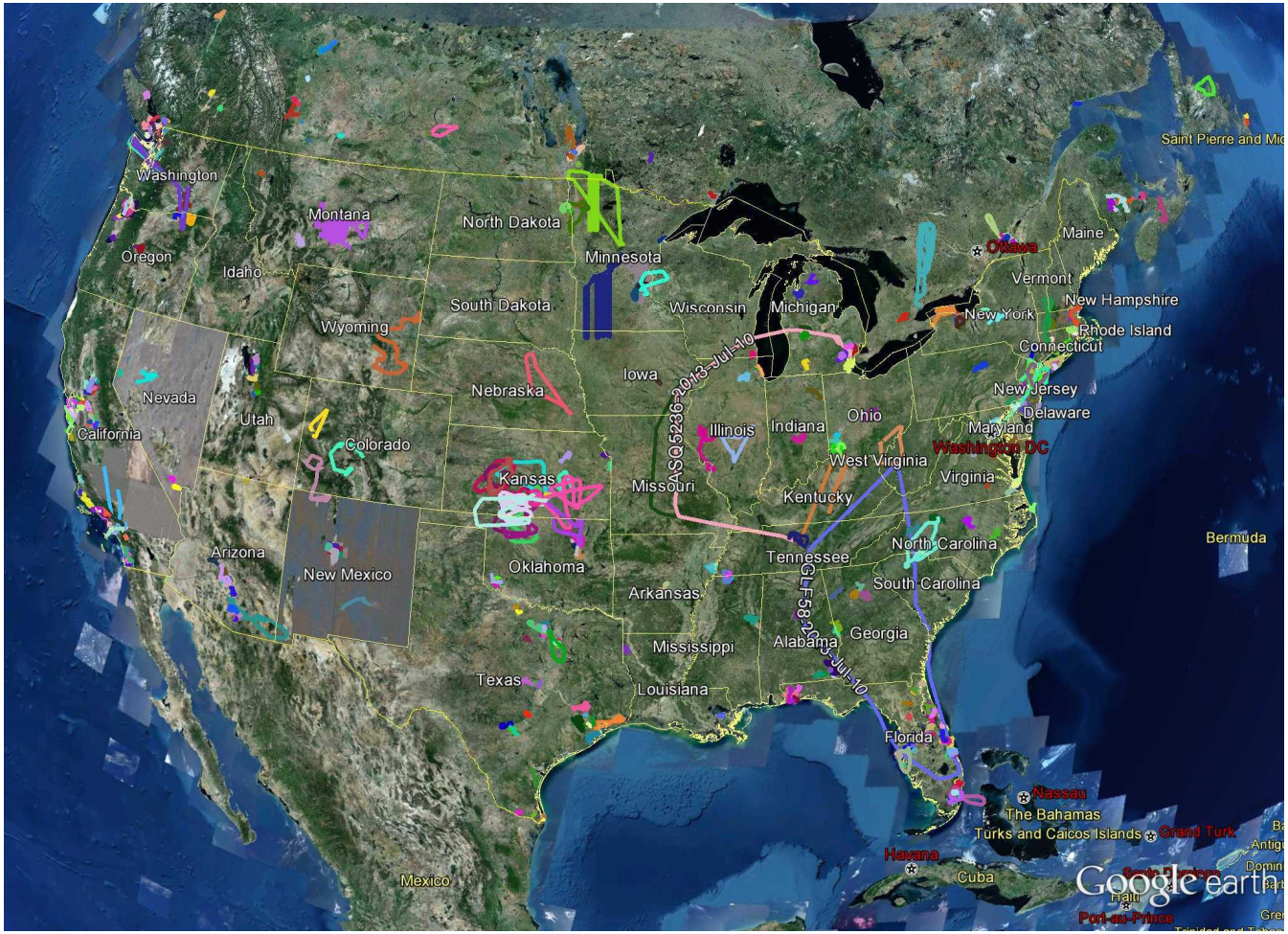
Discovery of Odd flights

Sandia National Laboratories

Clustering done based on geometric features

Many clusters found, but what remains is…

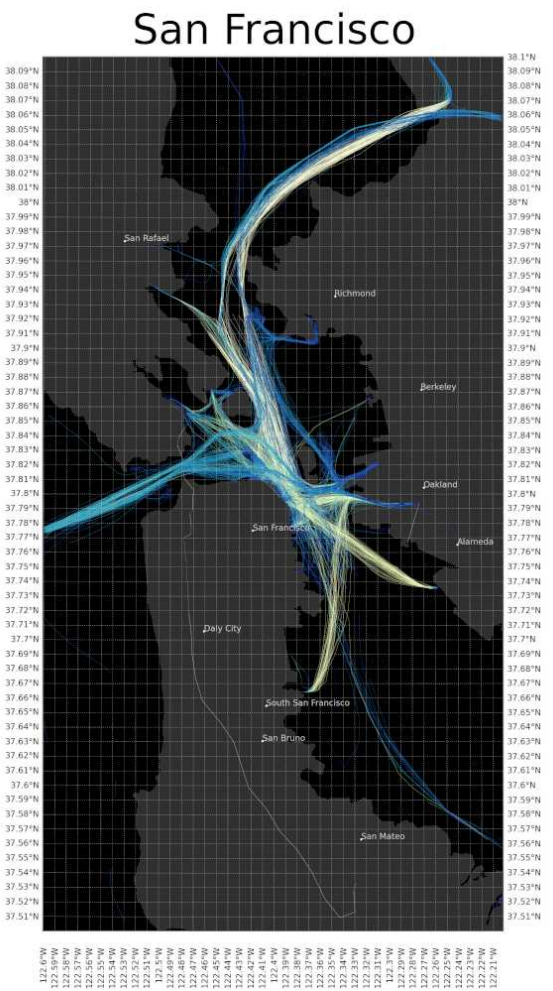
Note: we have ~5M points/day, ~1GB/day, currently >300GB

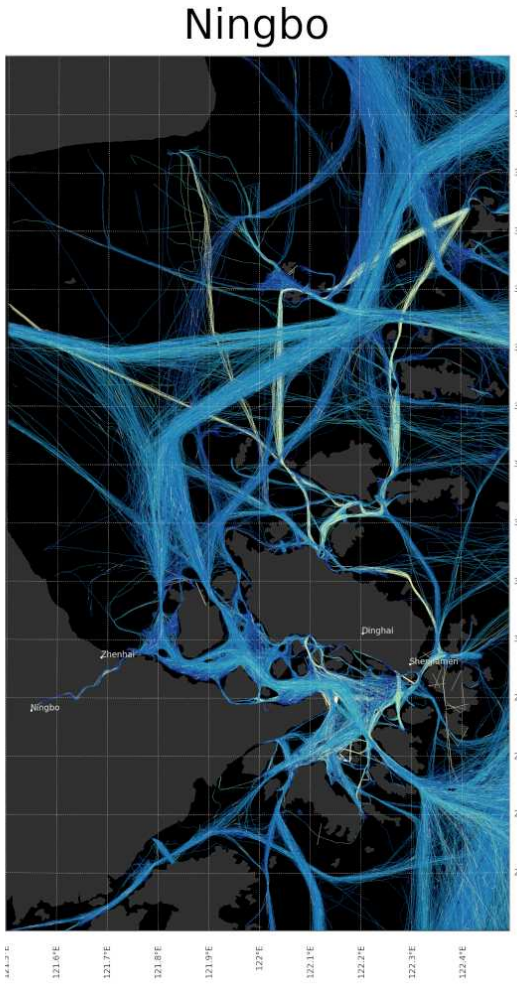


Represents approximately 700 out of a total of 50,000 flights from one day

Data Set 2: AIS Maritime Traffic

- Automatic Identification System
  - Modern collision avoidance
  - Required on int'l ships >300T, all passenger ships
  - Ship-to-ship, ship-to-shore
    - We have traffic within ~80km of land
  - Broadcast interval varies by size, speed
- Broad diversity of traffic and behavior
  - Lumbering herds of supertankers
  - Water ballet in and around ports

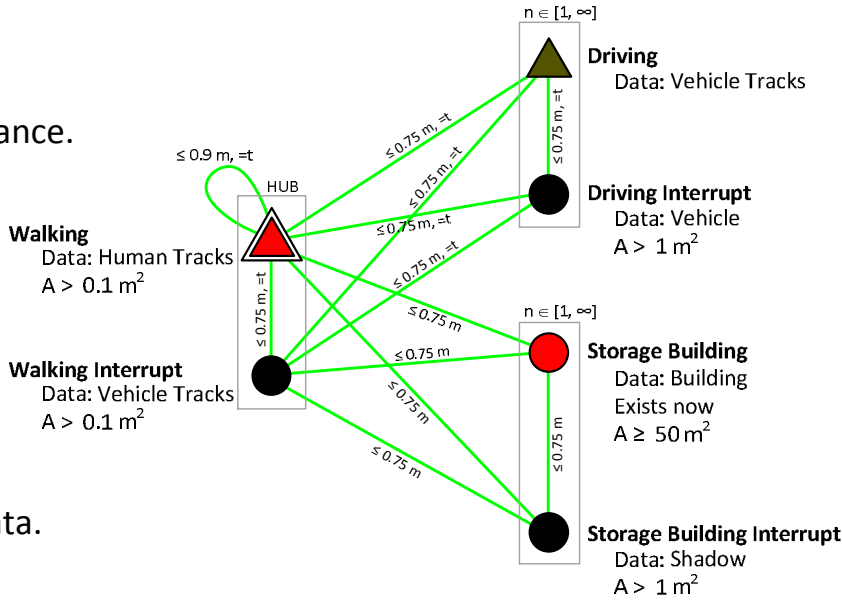




Integrated End State: Graphs

Strengthen Sandia's remote sensing search capability.

- Graph search:
  - Temporal search procedure.
  - Complexity and performance analysis.
  - Advanced disconnected signature.
  - Evaluation: Correctness, scope, complexity, performance.
- Working with real data.
  - Input file organization, provenance files.
  - Coping with noise.
- Ensemble nodes:
  - Allow reasoning at higher levels of abstraction.
  - Support uncertainty/quality calculation.
  - Support trajectory analysis, especially with limited data.
- Path partitioning, chaining.
- Uncertainty analysis.
- User application:
  - Code interface between GeoGraphy and user application.
- Support HA user studies.

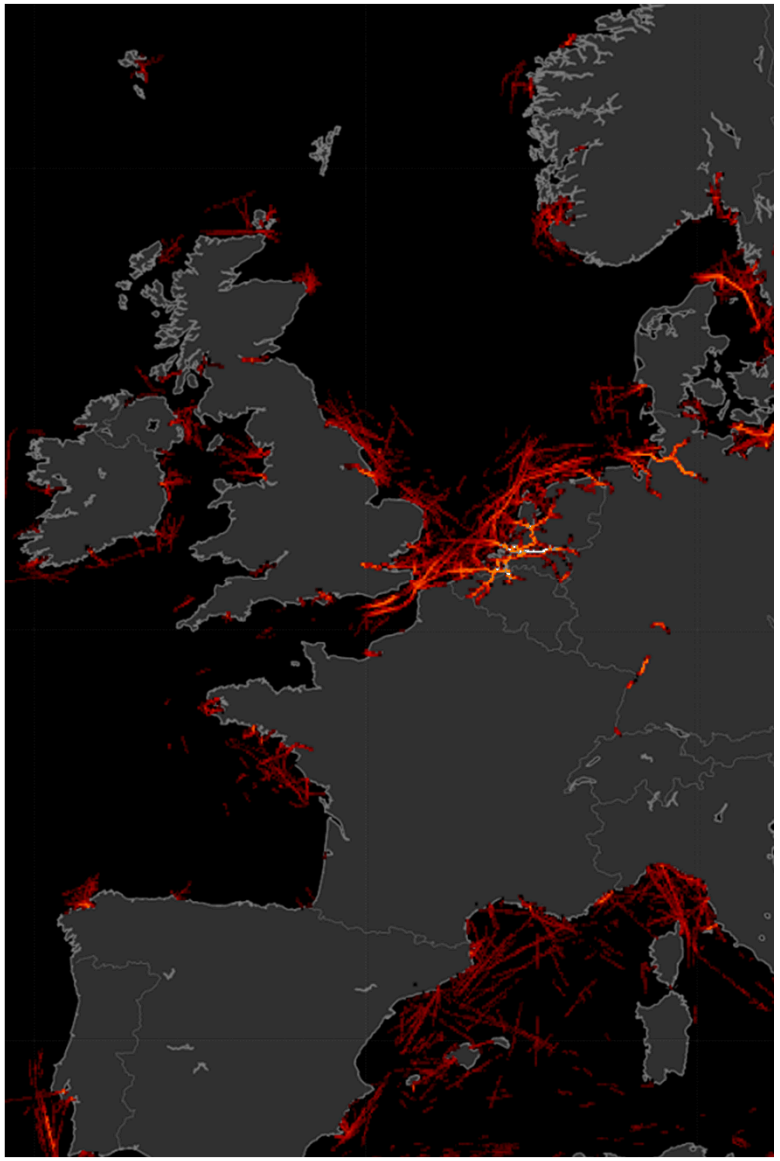


Integrated End State: Trajectories

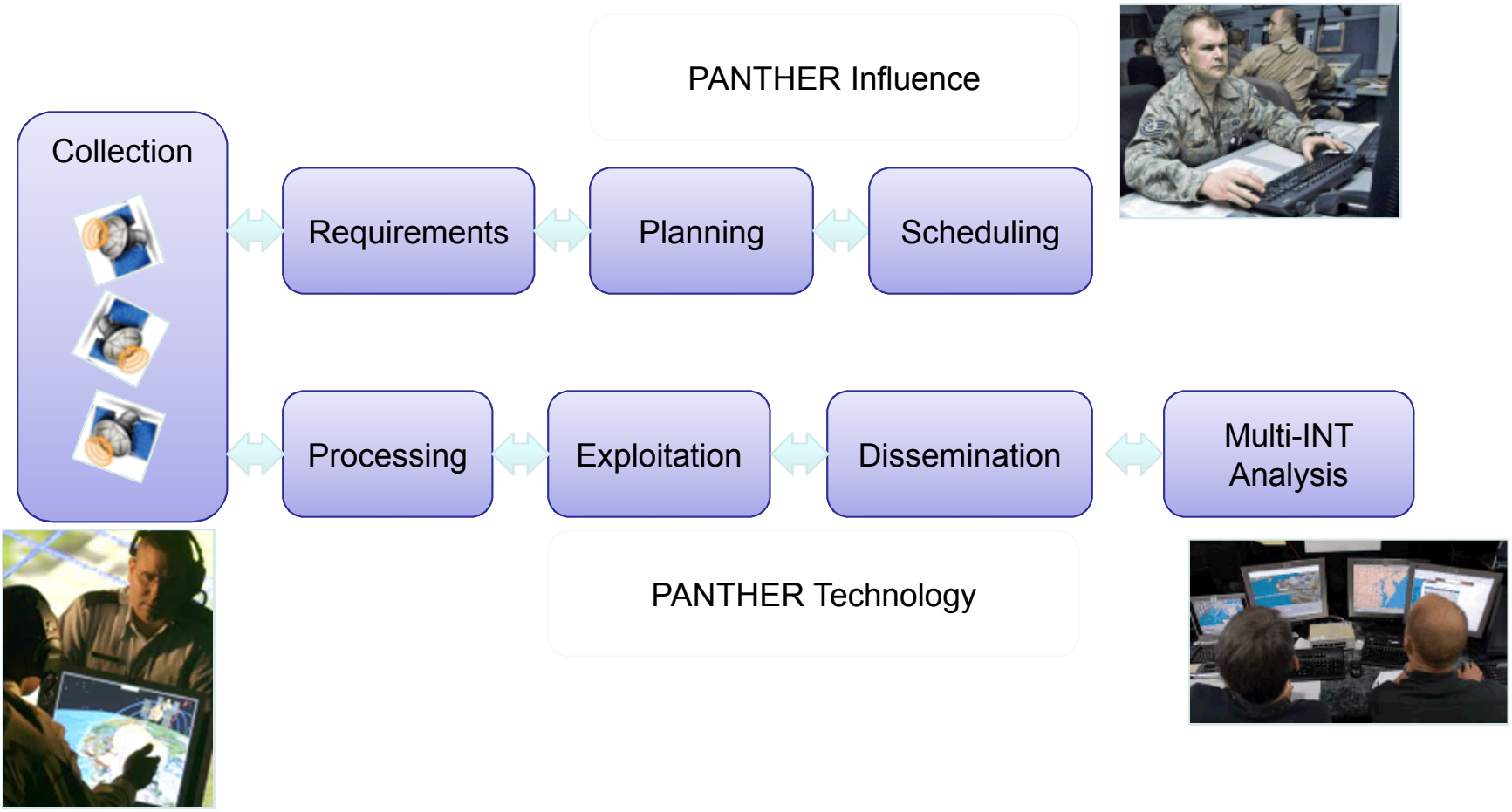
Sandia National Laboratories

Strengthen Sandia's expertise in mathematics, modern computing and visualization toward national impact.

- Algorithm development
  - Incorporation of sub-trajectory analysis
  - A covariance/correlation analysis of the information in the different features
  - Develop quantitative metrics to understand the effect of sparseness and intermittency in the data
- Software Engineering
  - Get a version 1.0 of TrackTable into the wild
  - Put a wrapper around parts of TrackTable that will enable customer usage
  - Incorporate trajectory work into GeoGraphy codebase
- Further IBM PureData development



Vision for PANTHER Technology



Sandia National Laboratories

Better knowledge & deeper insights
from Big Sensor Data
in minutes, not months;
over months, not hours or minutes;
covering hundreds, not 10s of km^2

PANTHER Spin-offs: R&D and Applications

Sandia National Laboratories

- Apply PANTHER trajectories to maritime & air domain missions.
- Apply PANTHER signature search to CBP and CuHd Missions.
- Apply PANTHER signature search to GEOINT system.
- Investigate single-pol SAR static feature algorithms and CCD utility for current systems. - *proposed*
- Investigate alternate data stores and query protocols for processing efficiency in signature search algorithms. - *proposed*
- Investigate implementation of geospatial temporal sematic graphs in a cloud environment.
- Image Interpretability Rating Scale for CCDs/ “Staircase” Visual Perception
- Modeling Imagery Analysis Processes in Theater Logs

Contrast National Security to Google

Sandia National Laboratories

	Google	NS Data to Decisions
Data Gathering	Highly Cooperative/dense, homogeneous (Android, Chrome, Google Searches, etc.)	Adversarial Environments/sparse/diverse (can't see everything all the time)
Data Analytics Environment	Large, homogeneous computing environment (cloud)	Diverse architectures, available throughput, bandwidth
Human Analysis	Consumer performing "buy" decisions	Analyst extracting actionable intelligence
Decision Consequences	Small – millions of consumers making decisions related to shopping	High – tactical and strategic, UQ and validation are critical