The Spectrum of Data Intensive Computing Activities at LANL

Francis J. Alexander*

Los Alamos National Laboratory

•With input from lots of folks from LANL!

LA-UR 10-02556

Developing *algorithms* for massive inference and learning problems:

Focus only on Algorithms –not architectures

Reducing computational complexity, enhancing efficiency, accuracy

Full understanding that the architecture matters

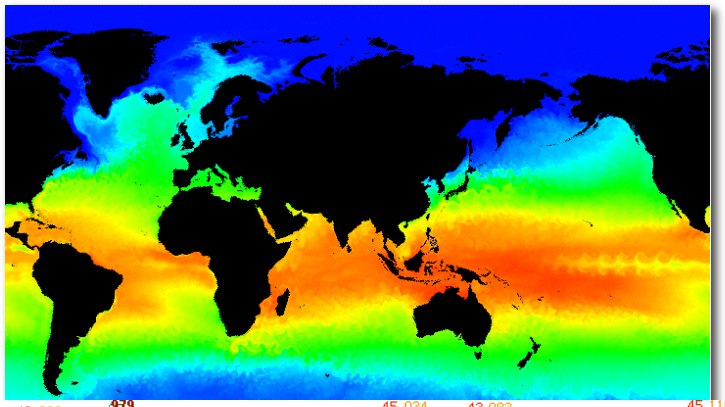
Hope to discuss that with many of you this week

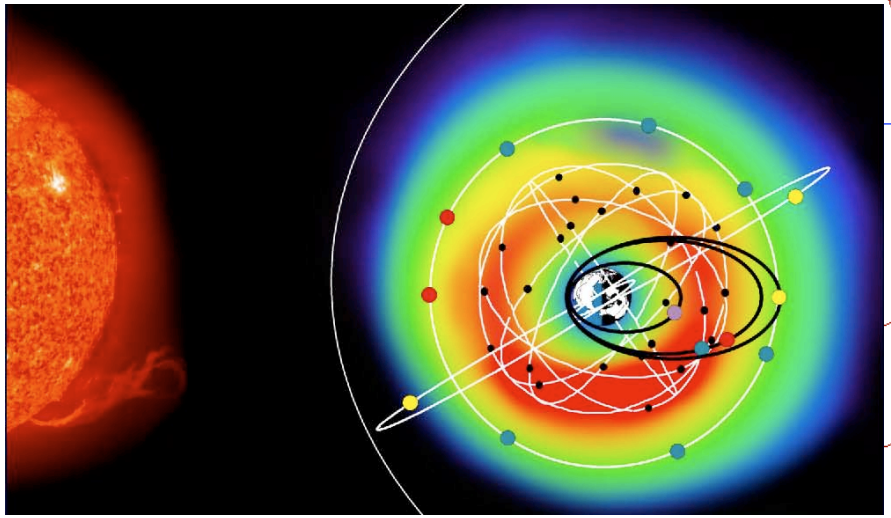
Outline

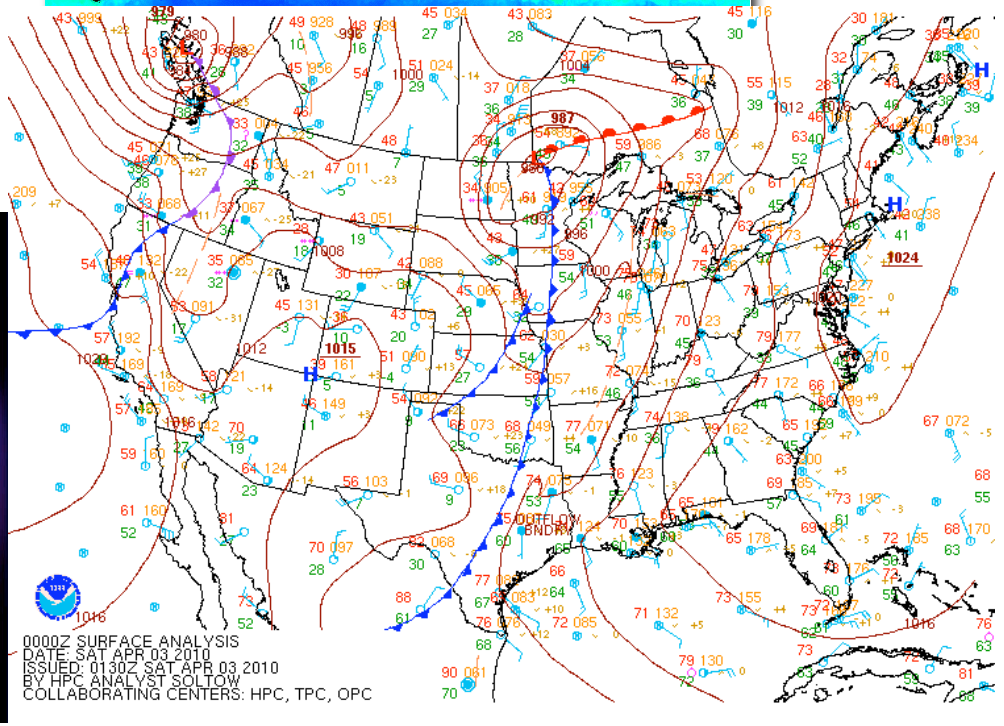
- Optimal Estimation with large data-sets
  - Data Assimilation
    - Motivation
    - Problem Statement
    - Algorithms
  - Graphical Models for Inference
    - Motivation
    - Problem Statement
    - Algorithms
    - Future Applications
- Statistical Learning with large data-sets
  - Game Theory / Machine Learning approach to Climate Modeling
  - Change Detection in Imagery / Large Scale Classification
  - Wide area video monitoring
- Other large-data problems at LANL
- Path Forward and Cautionary Remarks

Optimal Estimation: Motivation

- Examples
- (Terrestrial) Weather
- Ocean / Climate
- Space Weather







Optimal Estimation: Problem Statement

(one of the few equations you'll see here, and really the only one you might want to pay attention to…)

$$\mathbf{dx}(t) = \mathbf{f}(\mathbf{x}(t), t)dt + (2D)^{1/2}(\mathbf{x}, t)\,\mathbf{dW}(t)$$

$$\mathbf{x}(t_0) = \mathbf{x_0}.$$

$$\mathbf{y}(t_m) = \mathbf{h}(\mathbf{x}_m) + \boldsymbol{\epsilon}_m$$

$$\mathbf{x}_S(t) = E[\mathbf{x}(t)|\mathbf{y}_1, \ldots, \mathbf{y}_M]$$

Standard Approach

- Kalman/Bucy solved problems exactly for linear systems, Gaussian noise, additive Gaussian observation errors
- Scales badly with system size
- Ensemble Methods are popular– we're taking a different (complementary) approach
- Embrace the data
  - Use as much data as possible
  - Use complex dynamical models if necessary

Sampling Histories

- Idea is to Map to the following form for likelihood (not a new idea!)

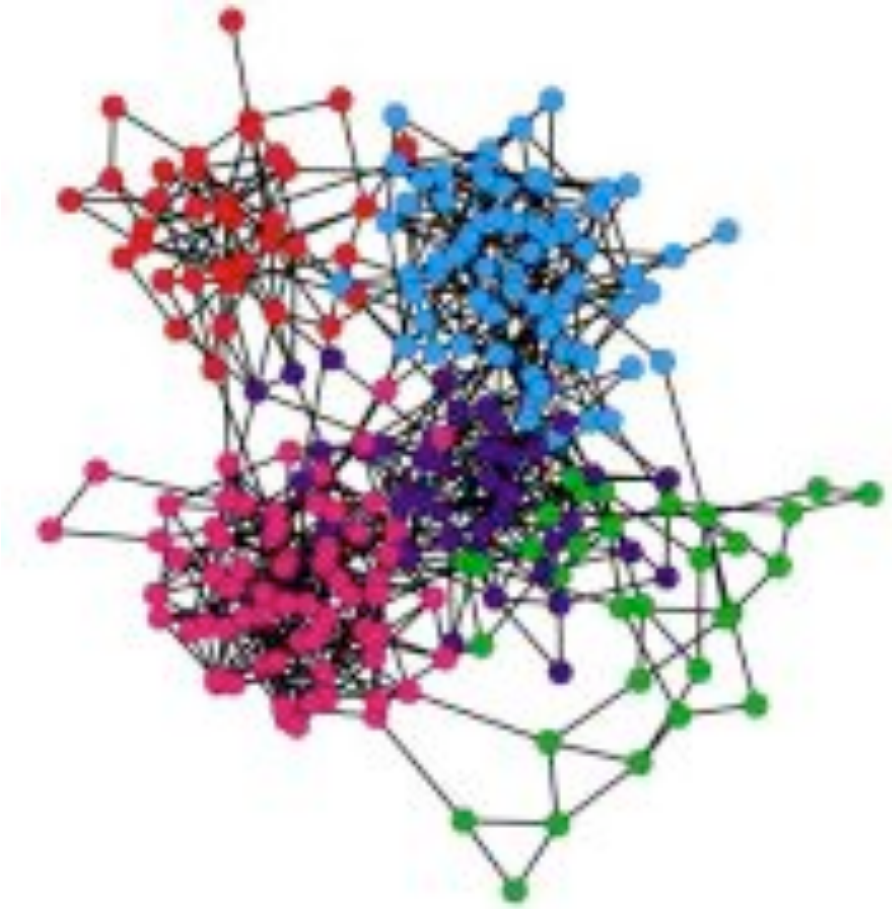
P(of a given history)= exp(-A(history))/Z

- Sampling techniques are new

Algorithms for Sampling Histories

- Multigrid Monte Carlo
- Langevin
- Fourier-Langevin
- Hybrid Monte Carlo
- Generalized Hybrid Monte Carlo
- Adjoint Driven
- Cluster Algorithms
- Variational Methods
- …

Inference on Graphs: Graphical Models



CyberSystems
Biological Systems
- Metabolic Networks
- Protein Networks
- …

Social Networks
Document Networks
Power Grid

…..

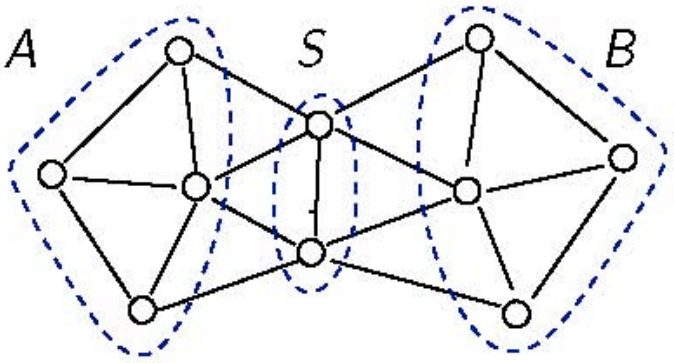
Graphical Models

A *graphical model* is a multivariate probability distribution whose definition is based on a graph G—nodes correspond to random variables and edges represent influences among the variables.

The joint probability distribution $P(x)$ is expressed in terms of potential functions defined on the edges (or, more generally, cliques) of the graph:

$$P(x) = \frac{1}{Z} \prod_{\{i,j\} \in G} \psi_{ij}(x_i, x_j) = \frac{1}{Z} \prod_{C} \psi_C(x_C)$$

It satisfies the *Markov* property with respect to separators of the graph:



$$P(x_A, x_B | x_S) = P(x_A | x_S) P(x_B | x_S)$$

Canonical Graphical Inference Problems

- Calculate partition function (high-dimensional sum/integral)

$$Z(\psi) = \sum_{x} \prod_{C} \psi_C(x_C)$$

- Calculate marginal distribution of each variable

$$P(x_i) = \frac{1}{Z} \sum_{x_{V \setminus i}} \prod_{C} \psi_C(x_C)$$

- Find most-likely joint state (estimation/optimization)

$$\hat{x} = \arg\max_{x} \prod_{C} \psi_C(x_C)$$

Many applications in image processing, remote sensing, information theory...

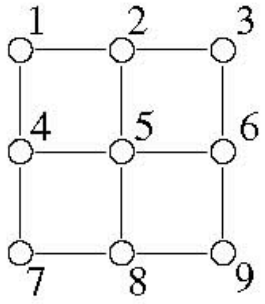
Optimal statistical inference in graphical models is generally computationally intractable $\Rightarrow$ need *tractable approximations*.

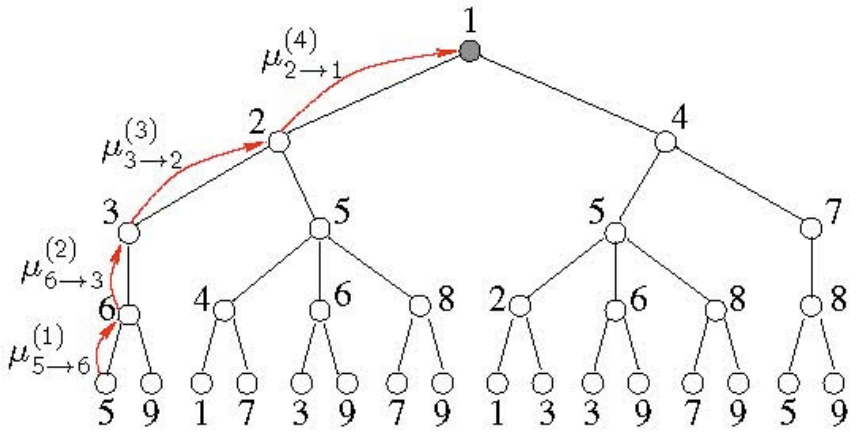
Belief Propagation

Message-passing algorithm that is exact in trees, but iterative and approximate for loopy graphs. Defined by the "sum-product" rule:

$$\mu_{i\to j}^{(t+1)}(x_j) = \sum_{x_i} \psi(x_j, x_i) \prod_{k\in\partial i\setminus j} \mu_{k\to i}^{(t)}(x_i)$$

It is interpreted as inference in the computation tree (universal cover) of the graph (Weiss & Freeman)





There is also a "max-product" form to solve discrete optimization problems and Gaussian elimination form for Gaussian graphical models.

Most problems do not live on trees--Belief Propagation is an approximation

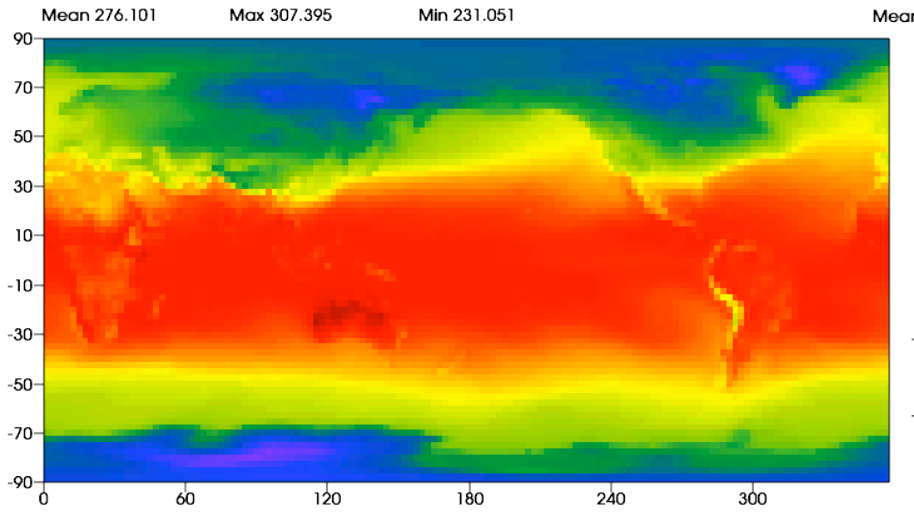
- Combinatorial Methods
  - loop-series analysis & correction of binary BP (Chertkov et al)
  - orbit-product analysis & correction of Gaussian BP (Johnson et al)
  - new approaches based on classical methods (Kac-Ward, Fisher-Kasteleyn) using *planar covers* of a graph
- Variational Methods
  - others have used tractable relaxations of Gibb's variational principle (Yedidia, Wainwright, Globerson, Jaakkola)
  - we are developing new methods exploiting monotonicity to obtain upper-bounds from tractable (e.g. planar) sub-graphs.
- Multiscale Methods
  - inspired by success of multiscale methods in linear algebra (multigrid) and statistical physics (renormalization group)
  - improve approximate inference by first reformulating problem in a multi-scale way, and then apply the approximation (e.g., BP).
  - developed generalization of Kadanoff's "bond moving" algorithm for approximate inference in lattice models—reformulated as lattice splitting method.

Statistical Learning with Large Data-sets

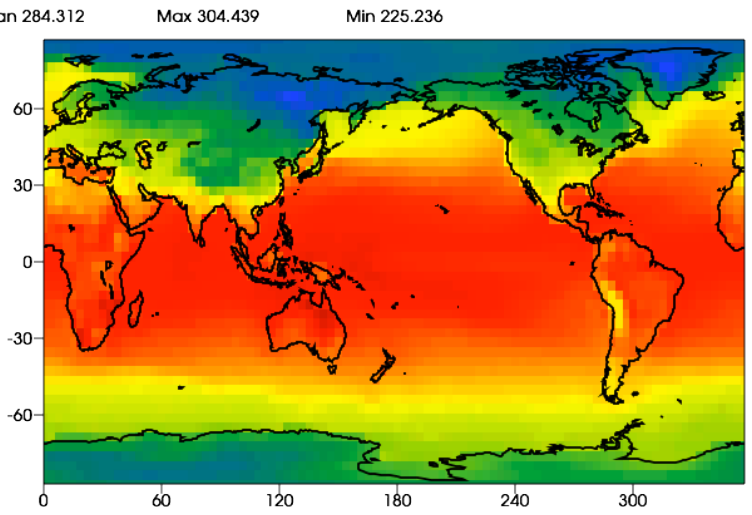
- Model Selection via Game Theory and Machine Learning
- Change Detection in Imagery / Anomaly Detection
- Wide Area Surveillance

Model Selection via Game Theory and Machine Learning

DATA SOURCES: OBSERVED & MODELED



Observed



Forecast

- Consolidated interpolated “Observed Data” from January 1960 to present.
- Monthly data available from 24 model simulations.
- GOAL: use algorithms from machine learning to determine best forecast.
- Play unique game for each region in the world.

BACKGROUND:
GAME THEORETIC ESTIMATION

- Instead of selecting a linear combination of simulations that results in the minimal misfit of the observed data with forecast data ($\min_{\alpha} ||(\mathbf{y}_k - H\sum_i \alpha^i \mathbf{x}^i_{k|k})||^2$) we optimize over regret.
- Regret, $R_{i,t}$ = how much better the algorithm would have done had it listened to i-th expert at time t.

Initially: Set $R_{i,0} = 0, p_{i,1} = 1/N$ for each i.
For $t = 1, 2, \ldots$

1. Each action i incurs loss $\ell_{i,t}$.
2. Learner incurs loss $\ell_{A,t} = \sum_{i=1}^{N} p_{i,t}\ell_{i,t}$.
3. Update cumulative regrets: $R_{i,t} = R_{i,t-1} + (\ell_{A,t} - \ell_{i,t})$ for each i.
4. Find $c_t > 0$ satisfying $\frac{1}{N}\sum_{i=1}^{N} \exp\left(\frac{([R_{i,t}]_+)^2}{2c_t}\right) = e$.
5. Update distribution for round $t+1$: $p_{i,t+1} \propto \frac{[R_{i,t}]_+}{c_t} \exp\left(\frac{([R_{i,t}]_+)^2}{2c_t}\right)$ for each i.

The Normal-Hedge Algorithm.
K. Chaudhuri, Y. Freund, and D. Hsu. A parameter free hedging algorithm. 2009.

- Regret of the algorithm bounded! $R_{A,T} = O\left(\sqrt{T\ln N + \ln^3 N}\right)$

Raw Data

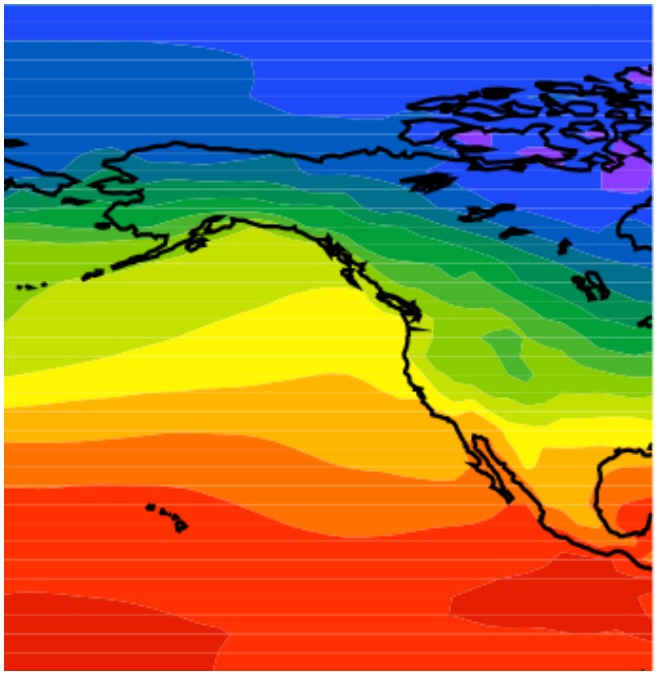
Region 7: Weights
Latitude = (0.0,90.0)
Longitude = (180.0,270.0)

Model	Organization	weights (1yr ave)	weights (4yr ave)	weights (11yr ave)
bccr_bcm2_0	BCCR, Norway	0	0	0
cccma_cgcm3_1	Canadian Center for Climate Modeling	0	0	0
cccma_cgcm3_1_t63	Canadian Center for Climate Modelling	0	0	0
cnrm_cm3	Centr. of Meteorological Research, France	0	0	0
csiro_mk3_0	CSIRO, Atmospheric Research, Australia	0	0	0
csiro_mk3_5	CSIRO, Atmospheric Research, Australia	0	0	0
giss_aom - run1	NASA, Goddard	0	0	0.0225729
giss_aom - run2	NASA, Goddard	0	0	0.0363424
giss_model_e_h	NASA, Goddard	0	0.00245193	0.0633354
giss_model_e_r	NASA, Goddard	1	0.951384	0.629234
iap_fgoals1_0_g - run1	Institute of Atmospheric Physics, China	0	0	0
iap_fgoals1_0_g - run2	Institute of Atmospheric Physics, China	0	0	0
iap_fgoals1_0_g - run3	Institute of Atmospheric Physics, China	0	0	0
inmcm3_0	Inst. for Numerical Mathematics, Russia	0	0.0461637	0.197001
ipsl_cm4	IPSL, France	0	0	0.0515145
mibu_echo_g	Many Groups, Germany/Korea	0	0	0
mri_cgcm2_3_2a	Meteorological Research Inst., Japan	0	0	0
ukmo_hadcm3	Hadley Centre, Met. Office, UK	0	0	0
ukmo_hadgem1	Hadley Centre, Met. Office, UK	0	0	0

Raw Data

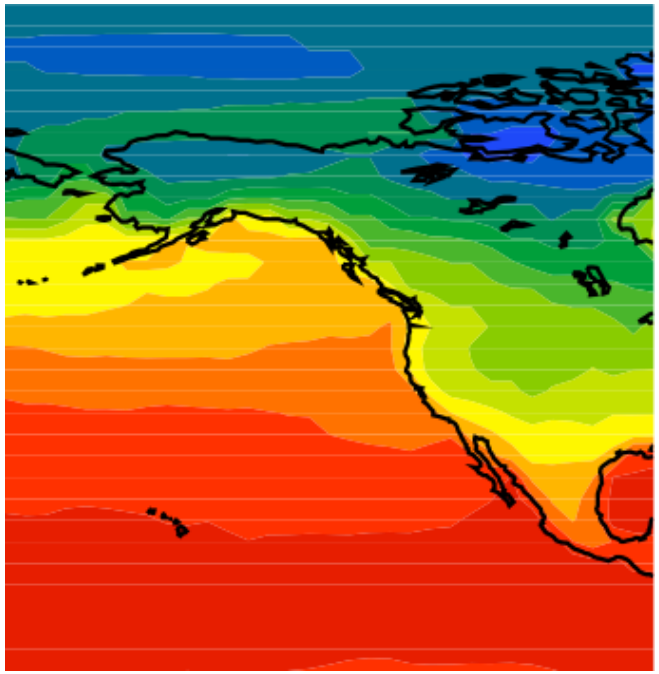
Region 7: Visualization
Latitude = (0.0,90.0)
Longitude = (180.0,270.0)

Truth



max: 301.438
mean: 283.448
min: 237.172

csiro_mk3_0 Estimate

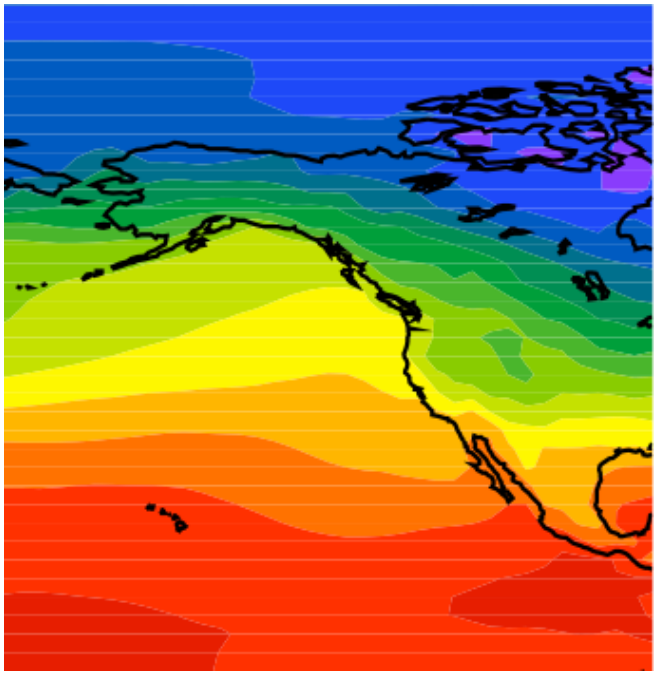


max: 299.723
mean: 279.981
min: 225.869

Raw Data

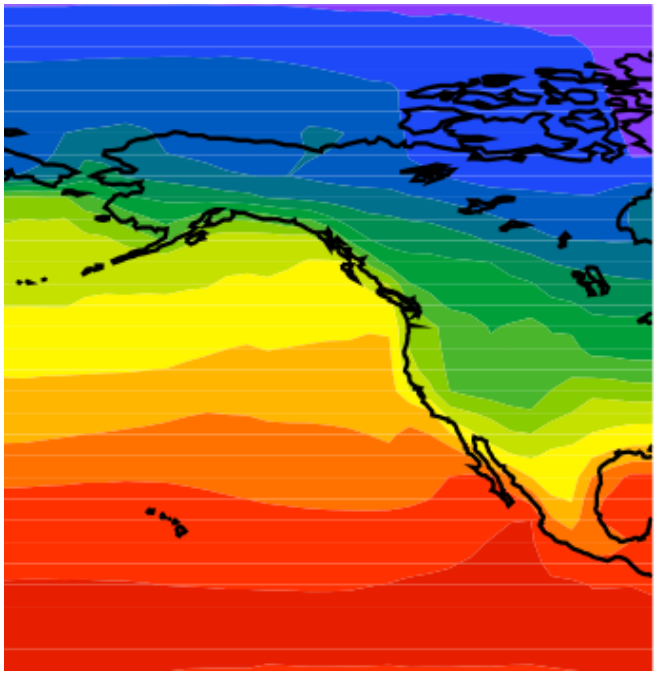
Region 7: Visualization
Latitude = (0.0,90.0)
Longitude = (180.0,270.0)

Truth



max: 301.438
mean: 283.448
min: 237.172

GT Reconstruction (11yr.)



max: 301.318
mean: 283.659
min: 236.252

Raw Data

Region 2: Weights
Latitude = (-90.0,0.0)
Longitude = (90.0,180.0)

Model	Organization	weights (1yr ave)	weights (4yr ave)	weights (11yr ave)
bccr_bcm2_0	BCCR, Norway	0	0	0.00312176
cccma_cgcm3_1	Canadian Center for Climate Modeling	0	0	0
cccma_cgcm3_1_t63	Canadian Center for Climate Modeling	0	0	0
cnrm_cm3	Centr. of Meteorological Research, France	0	0.0131017	0.0995652
csiro_mk3_0	CSIRO, Atmospheric Research, Australia	0	0	0
csiro_mk3_5	CSIRO, Atmospheric Research, Australia	0	0	0.0100237
giss_aom - run1	NASA, Goddard	0	0	0
giss_aom - run2	NASA, Goddard	0	0	0
giss_model_e_h	NASA, Goddard	0	0	0
giss_model_e_r	NASA, Goddard	0	0	0
iap_fgoals1_0_g - run1	Institute of Atmospheric Physics, China	0	0	0
iap_fgoals1_0_g - run2	Institute of Atmospheric Physics, China	0	0	0
iap_fgoals1_0_g - run3	Institute of Atmospheric Physics, China	0	0	0
inmcm3_0	Inst. for Numerical Mathematics, Russia	0	0	0
ipsl_cm4	IPSL, France	1	0.91845	0.547234
mibu_echo_g	Many Groups, Germany/Korea	0	0.0241407	0.118807
mri_cgcm2_3_2a	Meteorological Research Inst., Japan	0	0.0393476	0.151988
ukmo_hadcm3	Hadley Centre, Met. Office, UK	0	0.00495958	0.0692596
ukmo_hadgem1	Hadley Centre, Met. Office, UK	0	0	0

Change Detection in Large Sets of Images
What is interesting?

- Analysts face a glut of imagery

- Sometimes tasks are well defined
  - How many houses?
  - Are the roads passable?
  - Will this be a good year for soybeans?
  - Where are the golf courses?

- Some questions more open-ended
  - Is there any illegal activity?
  - Is there anything unusual?
  - What is interesting?

Quickbird, 26 March 2006

What has changed?

Ikonos, 23 May 2000

Quickbird, 26 March 2006

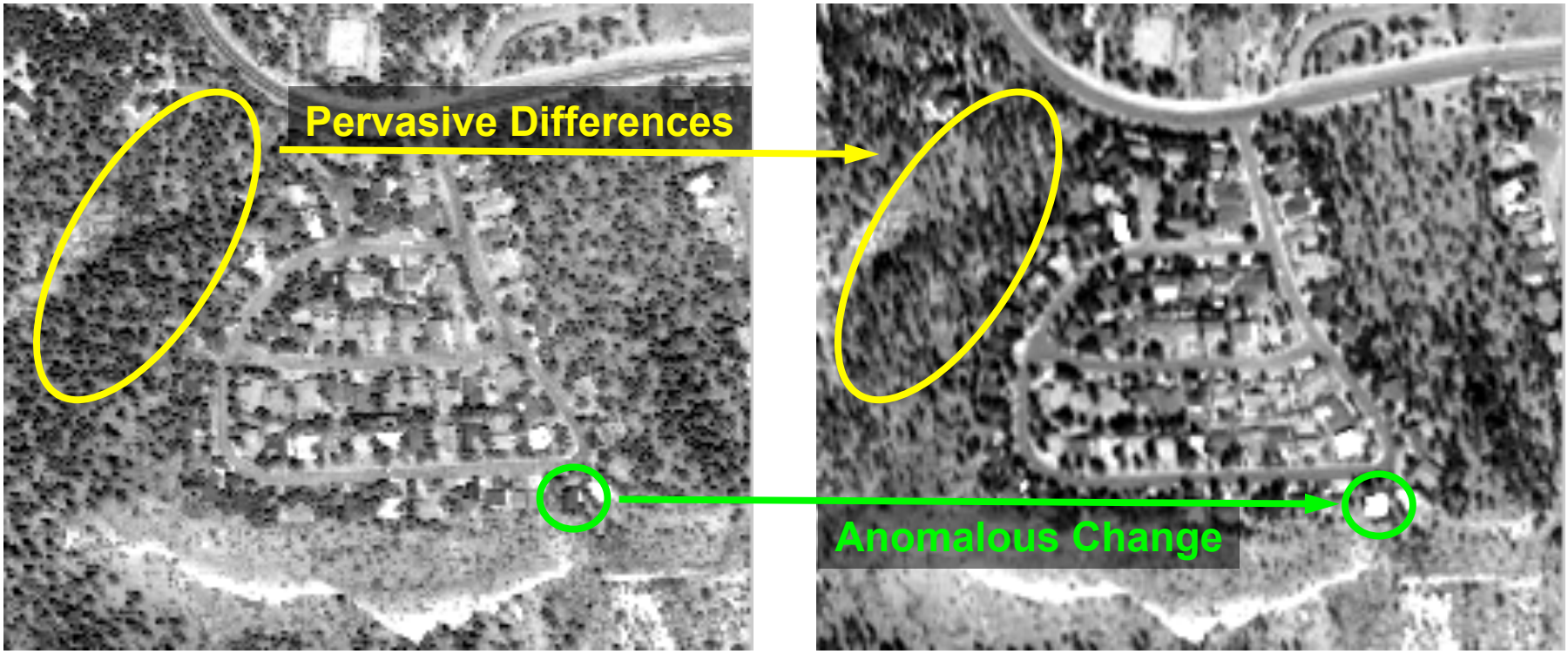
“Just because everything is different doesn’t mean anything has changed.”

-- Irene Peter

What is the *interesting* change? Developing new tools to answer this question

Ikonos, 23 May 2000

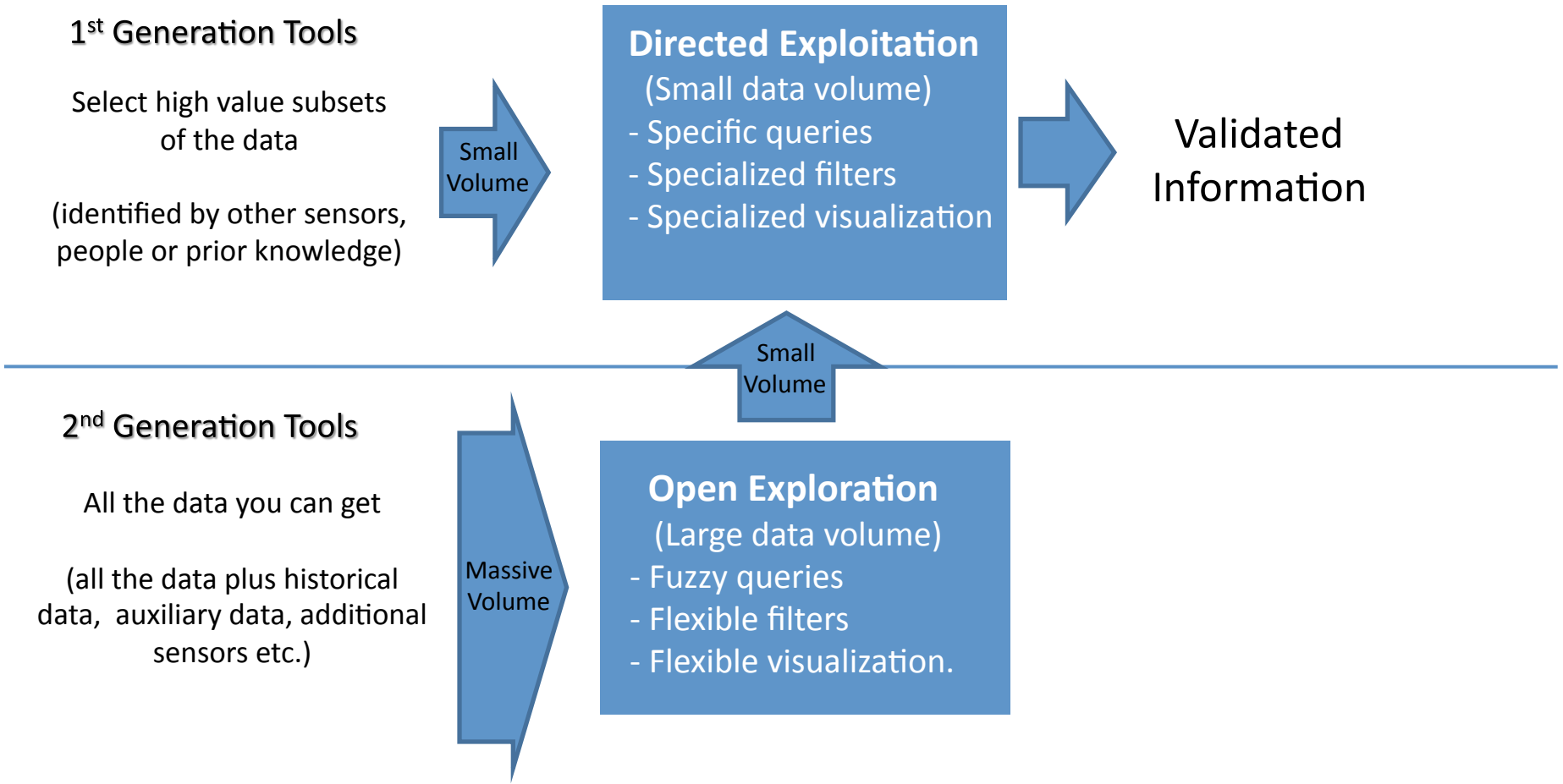
Quickbird, 26 March 2006



"Just because everything is different doesn't mean anything has changed."

-- Irene Peter

Large Volume Data Exploitation Tool Chains



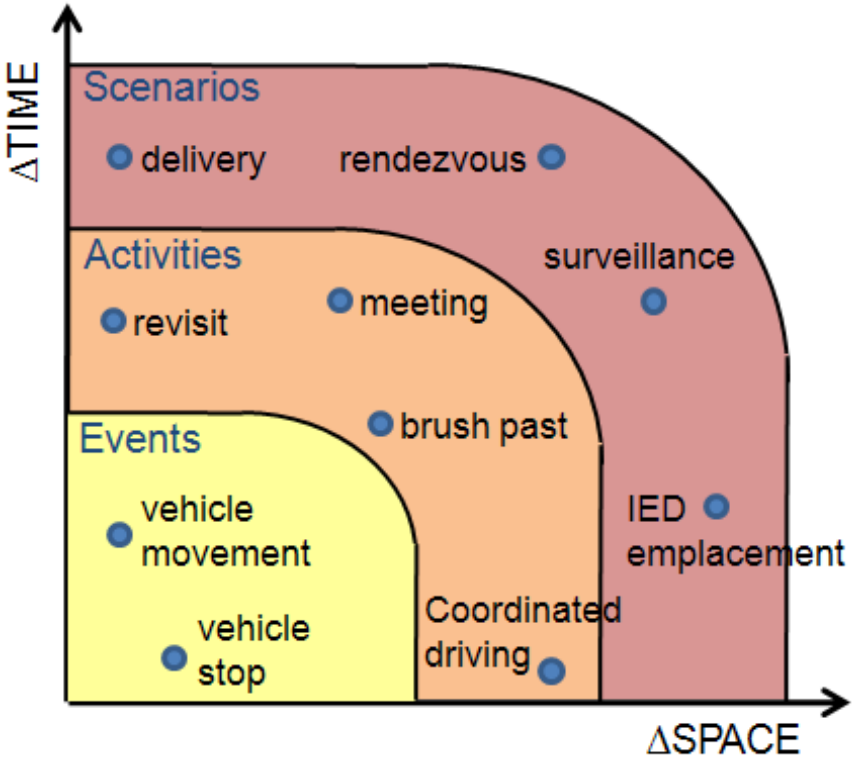
Scenario Extraction
2nd Generation Tools for Wide Area Motion Imagery

Wide-Area Motion Imagery

- *0.5 meters / pixel*
- *0.5 seconds / frame*
- *36 km^2 for 1-2 hours*

Scenario Extraction

Detecting and understanding the relationship between otherwise innocuous events in a complex urban scene



Activity Tool Use Example

1. Query

Meetings | Vehicle | Coordinated

3 vehicle meeting

2 vehicle meeting
3 vehicle meeting
4 vehicle meeting
Group meeting
Add new preset

Only search apix ROI

Edit

Number of Vehicles:

3 to 3

Interaction distance:

50 m 150 s

Observed:

Complete Include partial

Prioritize by: Weight

Probability At Location

No Priority

No Priority

Probability At Location
Duration
Direction
No Priority

Num Results: 20

Perform Query

2. Inspect

3. Triage

Current ROI

Rank: 1 / 20

Type: 3 vehicle meeting

X: 9823.72265 Y: 2892.8718 T: 3975.0000

Info:

A2(3975 3940) D3(7484 7556 7579)

Notes:

Save Delete

Delete all in view

Display

Tracks:

Single Arrival Departure

Left: ROIs Summary

Right: ROIs on right

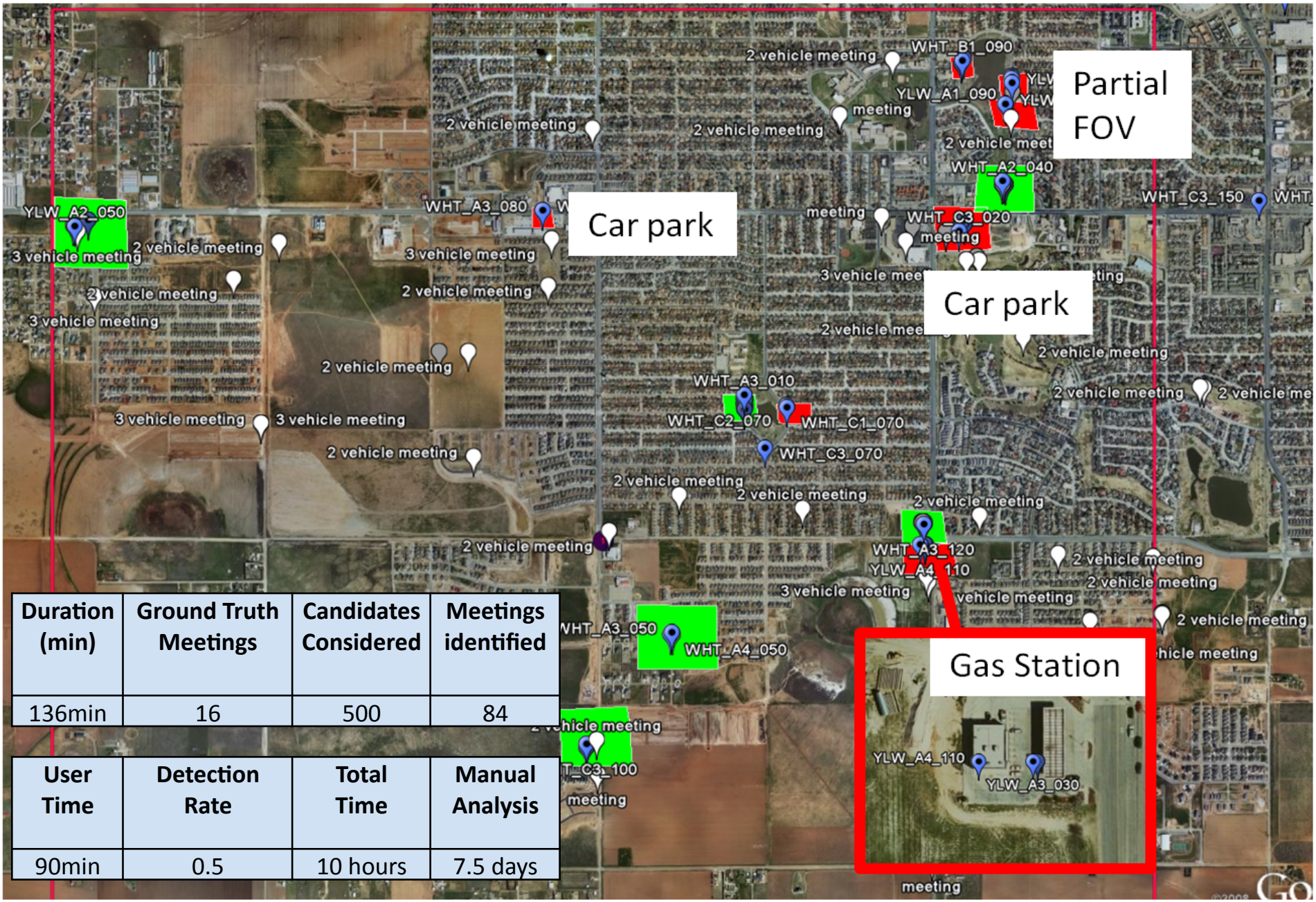
Query

Export User ROI

Import Recover

GeoCorrect Options

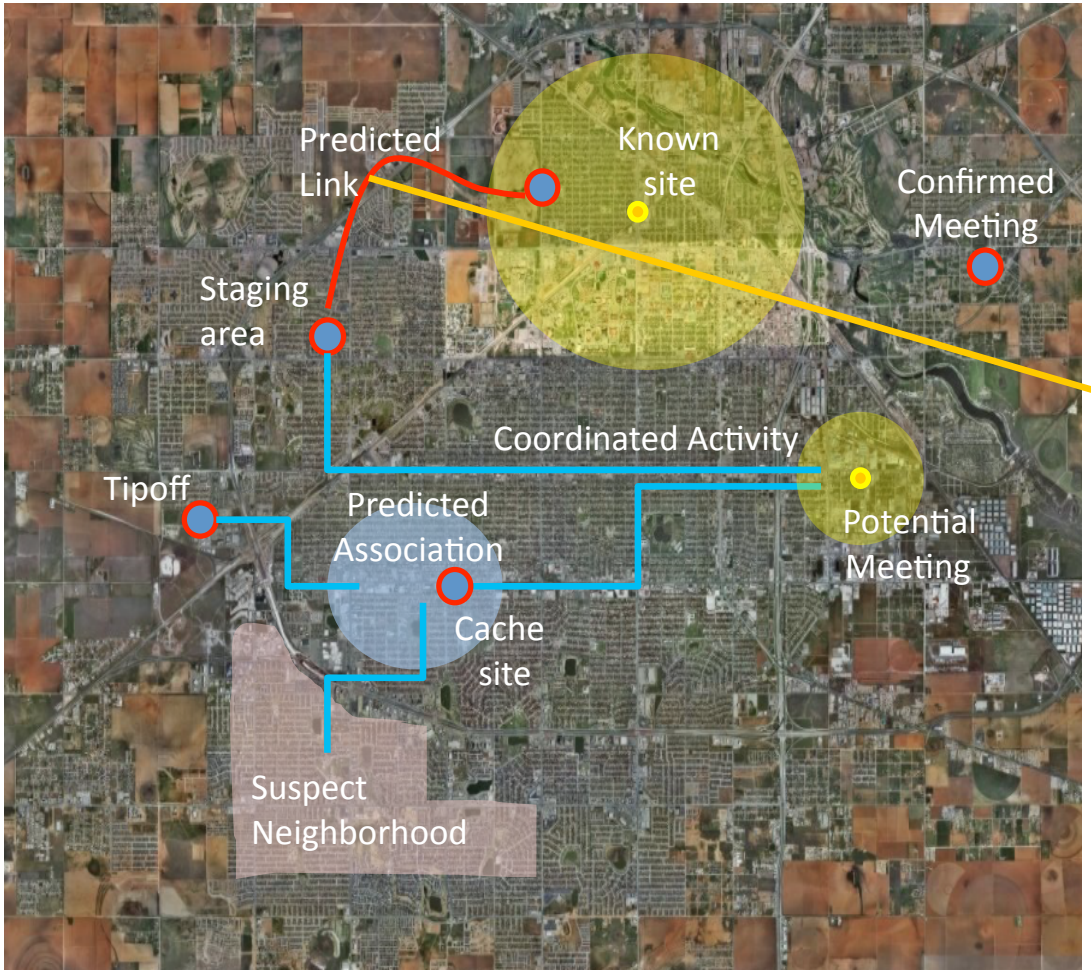
User Efficiency Improvement– 90mins versus 180hrs



Duration (min)	Ground Truth Meetings	Candidates Considered	Meetings identified
136min	16	500	84

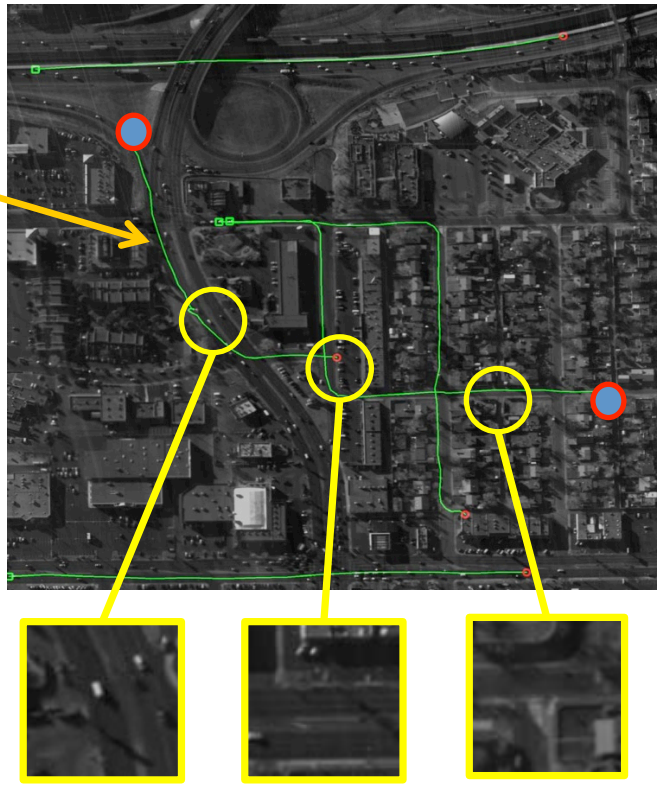
User Time	Detection Rate	Total Time	Manual Analysis
90min	0.5	10 hours	7.5 days

Towards Scenarios – Connecting Activities



- Query to find associated activities.

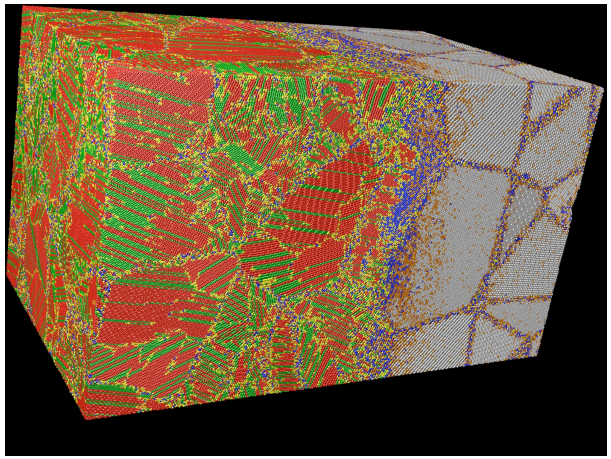
Example query result - potential link (60%)



User cues : confirm / deny

Other Large-Scale Data Problems at LANL

Quantifying Molecular Dynamics Studies with IS&T Tools



K. Kadau et al., *Phys. Rev. Lett.* (2006).

Shock-induced transformation in Fe predicted by large-scale molecular-dynamics simulations

Bcc(8nn, gray)->comp Bcc (10nn, blue)->hcp,fcc (12nn, red,green): simple algorithm

NEMD results confirmed experimentally by ultrafast (nanosecond) X-ray diffraction (shocks and high intensity X-rays produced by high energy laser systems OMEGA /Janus/Vulcan)

D.H. Kalantar et al., *Phys. Rev. Lett.* (2005).

Massive MD studies (SPaSM on Roadrunner): Trillions of atoms = several cubic microns

Tool to characterize sensitivity to polycrystalline morphology, realistic defect concentrations

However, massive data sets with limited storage capacity (1:1000 snapshots)

Identify "interesting" events in real time and save additional information in those areas: machine learning for anomaly detection and classification

More quantitative description of local geometry in terms of spherical harmonic expansions, enhanced classification of geometries

Group into classes of response behavior: phase transitions, slip planes, dislocation loops

Dominant modes for reduced order modeling: basis vectors for higher level models (e.g. plastic flow). What are the associated thermodynamic properties?

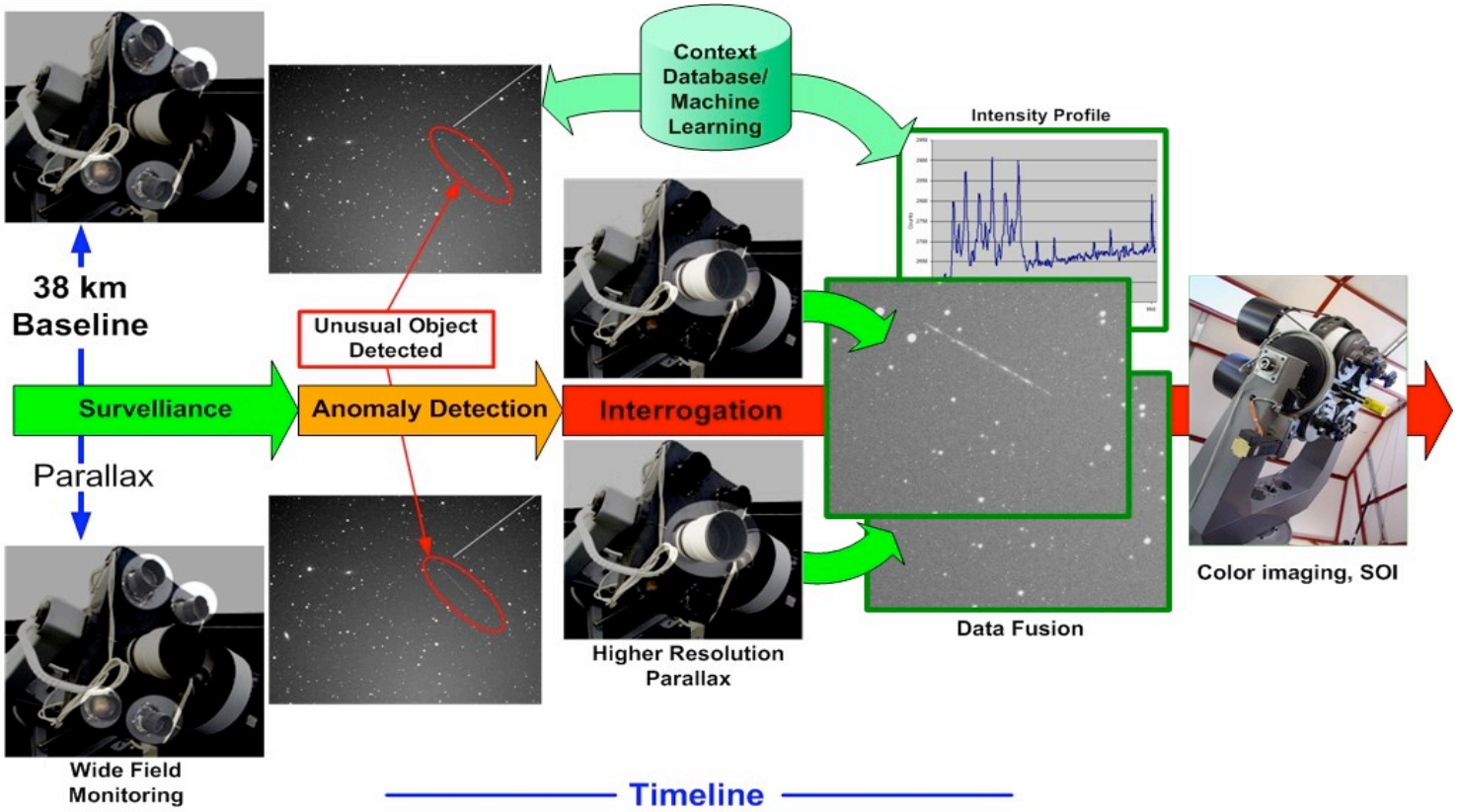
Sensitivity of results to the form and parameters of chosen potential functions.

Comparison and iteration with experimental results: diffraction, OIM

Kober, Barber, Steinwart

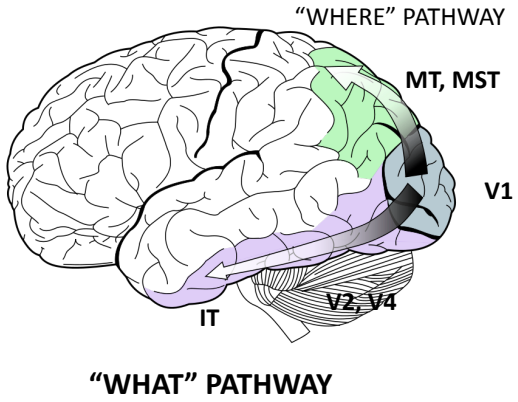
LANL's Thinking Telescopes:

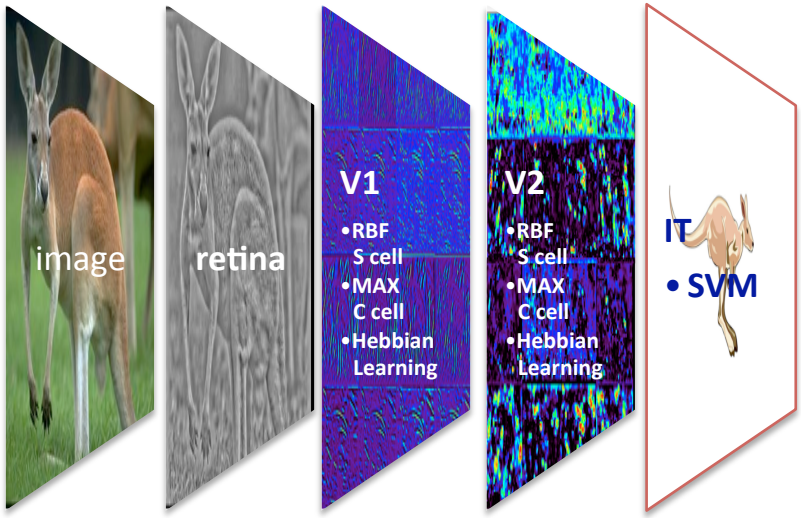
Autonomous, Real-Time, Robotic SSA Surveillance and Interrogation



UNCLASSIFIED

PANN Petascale Artificial Neural Network





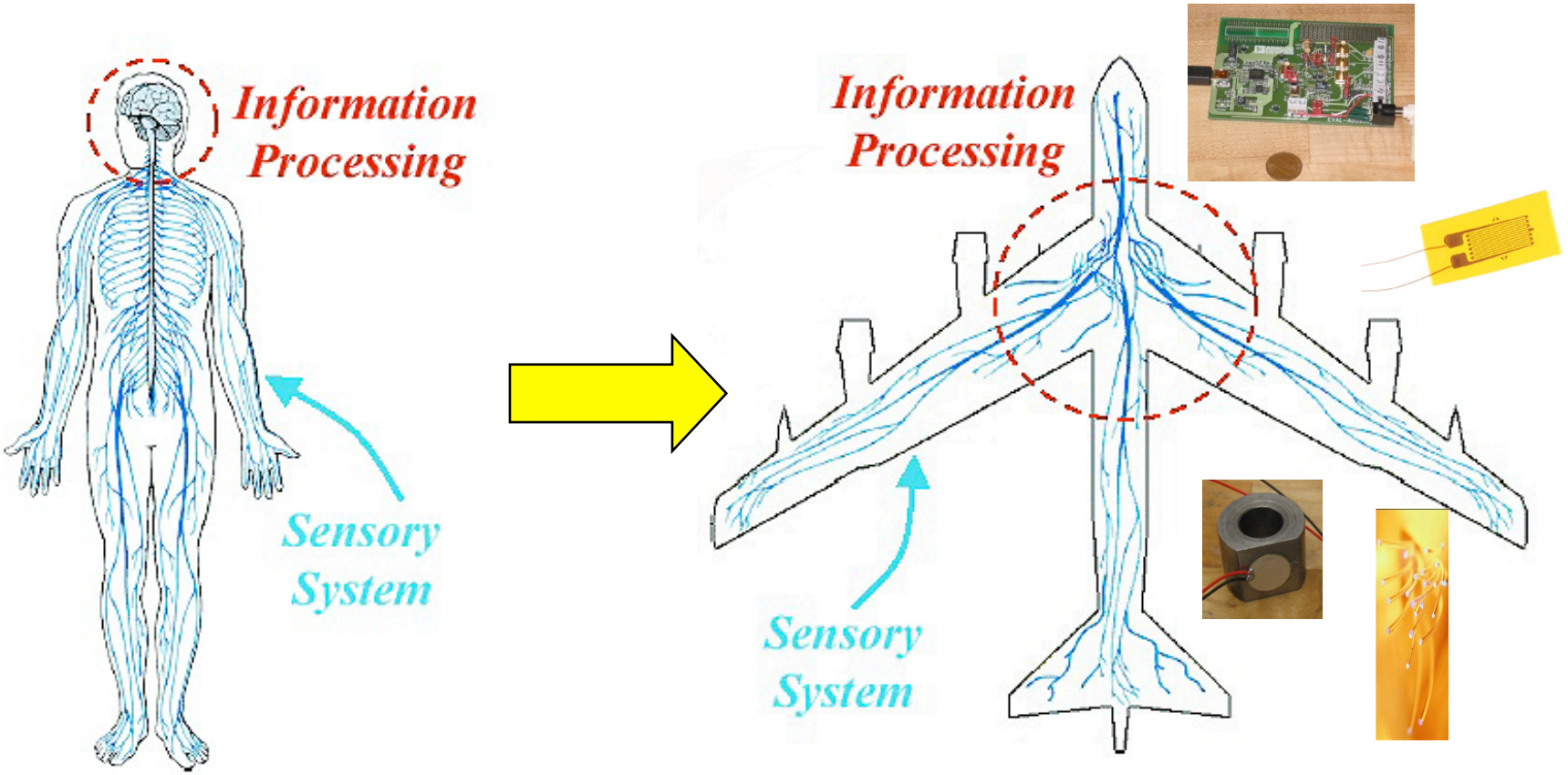
- Neocognitron/HMAX-type hierarchical feed-forward model of visual cortex **"what"** (ventral) pathway (V1 / V2 / IT).
- High performance parallel code using MPI, vector intrinsics, and Cell Broadband Engine.
- Can take as input any image format supported by open source GDAL library, or video format supported by open source FFMPEG library.
- On a cluster of 20 Opteron cores each with a dedicated Cell card (5 tri-blades), PANN processes "YouTube"-quality (200x200 pix) video in real time (> 20 fps).

- *PI:* Luis Bettencourt, T-5.
  http://synthetic-cognition.lanl.gov/

Animal / No animal binary classification task

Data Intensive Computing: Structural Health Monitoring

Structural Health Monitoring is the process of developing a damage assessment capability for aerospace, civil and mechanical infrastructure



Most Applications Generate Lots of Data

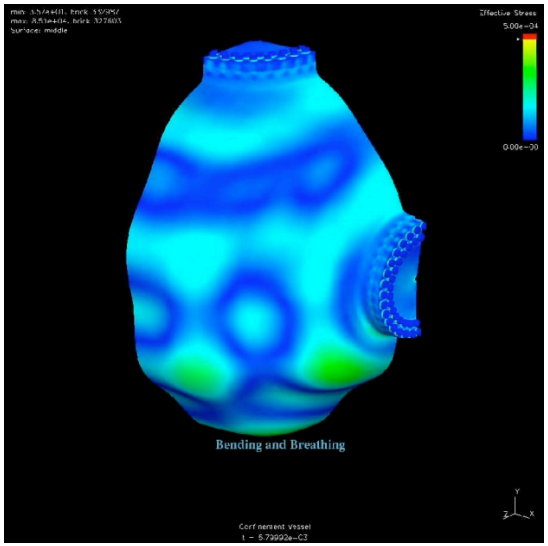
- Rotating Machinery
  - **100's of machines per plant**

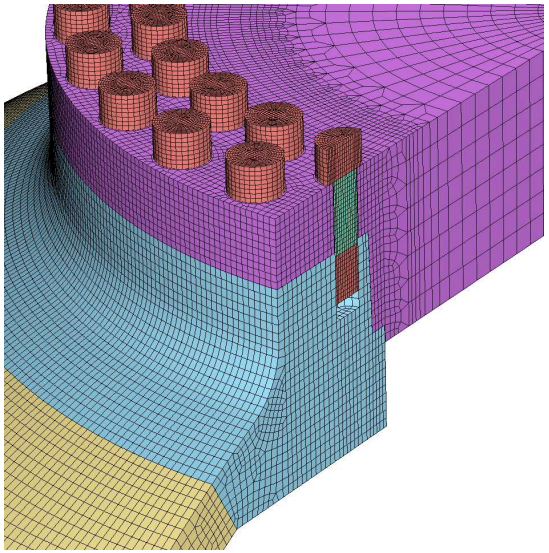
- Rotocraft Health and Usage Monitoring Systems (HUMS) :
  - **40 gbytes of data/hr of flight**

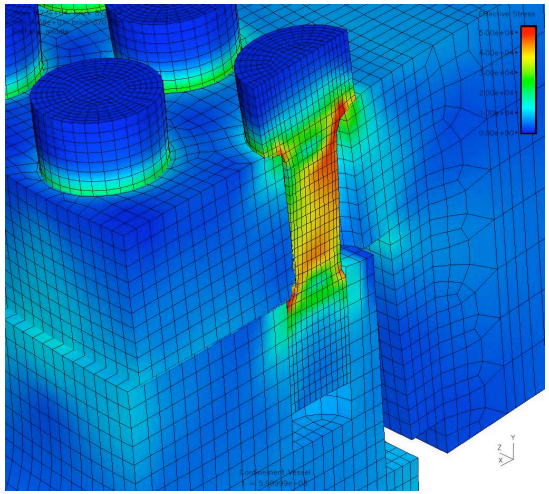
- **Civil Engineering Infrastructure**
  requires 1000's of sensors

SHM requires coupling between data & computationally intensive modeling

- The "Grand Challenge" for engineers in the 21st century: using SHM system state assessments as initial conditions to predictions of remaining system life







Time Scale of Damage Evolution Poses Challenges

- Long term SHM, periodically updates information regarding aging and degradation resulting from operational environments.
  - Maintenance staff continuity
  - Data archiving and retrieval
  - Maintenance vs construction budgets
  - **Sensor endurance and evolution**

- Rapid condition screening in near real time after extreme events, such as earthquakes or blast loading.
  - Robust sensing and processing hardware
  - Provide information to first-responders
  - Consequences of misdiagnosis are severe
  - **Must be integrated with control systems**

Cautionary Remarks

- Getting tons of data can be good but
  - Many "mundane" problems still left to solve (e.g., small sample size problem)
  - Shouldn't be enamored by just being able to measure more / simulate more
  - Make sure resources are devoted to turning data into knowledge (algorithms /theory/modeling)

Path Forward

Numerous challenges

Need for more accurate/faster
- methods for inference on graphs
- Filters/smoothers for nonlinear problems
- Anomaly detection algorithms (possibly single pass)
- Classification algorithms to deal with petabytes of data
- Optimal (or close to optimal) design of integrated HW/SW/Algorithms systems to solve targeted problem

Thank You!