Problem Day
NISS

Pacific Northwest National Laboratory
Statistical and Mathematical Sciences

Brent Pulsipher, brent.pulsipher@pnl.gov
Tom Ferryman, tom.ferryman@pnl.gov
Paul Whitney, paul.whitney@pnl.gov
Dale Anderson, dale.anderson@pnl.gov

http://www.pnl.gov/statistics

March 2004

Battelle

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

Location: Desert Part of Washington State

http://www.hanford.gov/top/maps.html

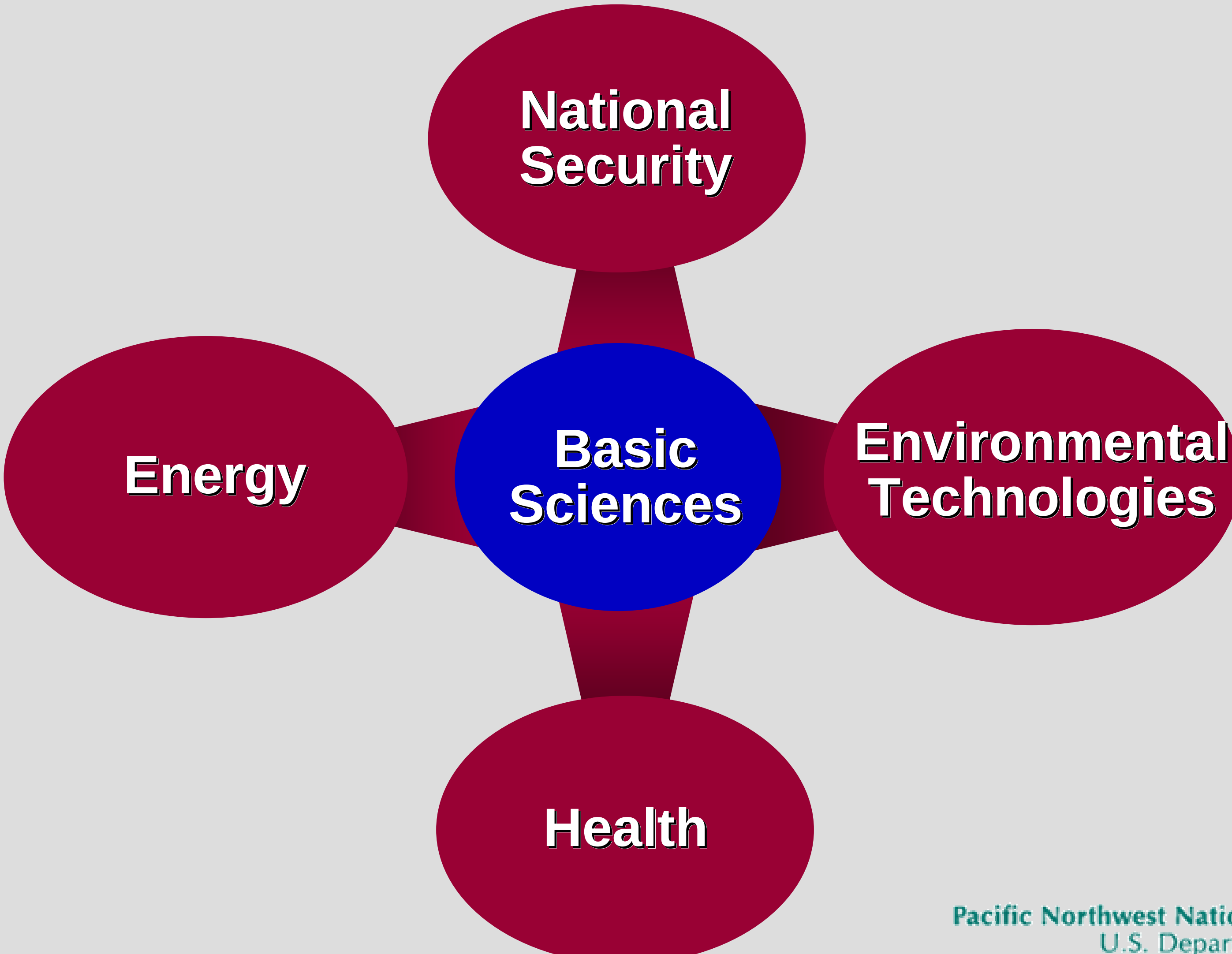
Outline

- Statistical and Mathematical Sciences Group
  - People
  - Business
  - Capabilities
- Project Examples of Capabilities
- Proposed Problem 1
- Proposed Problem 2
- Proposed Problem 3
- Closing Remarks

Pacific Northwest National Laboratory
U.S. Department of Energy
tom.ferryman@pnl.gov

What does PNNL Do?

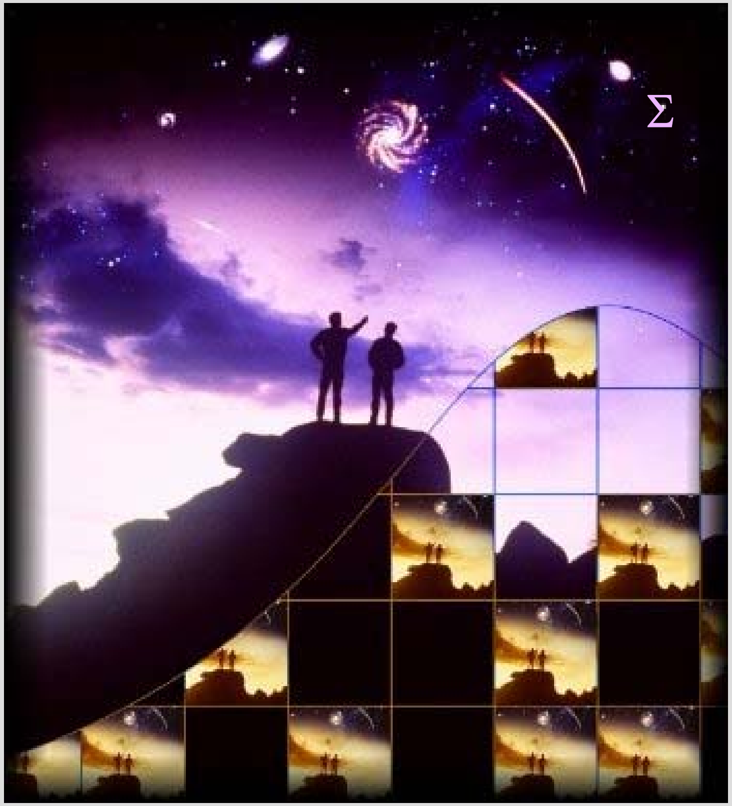
Statistical and Mathematical Sciences works in all areas



tom.ferryman@pnl.gov

STATISTICAL and MATHEMATICAL SCIENCES

- Use standard and/or novel data analysis methods
- Apply to simple or complex data sets
  - High dimensional
  - Large volume
  - Diverse data types
    - Numeric
    - Categorical
    - Text
    - Image
    - Spectra
    - Others
- Data Analysis and Tool Development
- Quantify uncertainties
- Validate models and simulations



Battelle

Pacific Northwest National Laboratory
U.S. Department of Energy
tom.ferryman@pnl.gov

PNNL Statistics and Quantitative Sciences

38 GREAT PEOPLE !!!

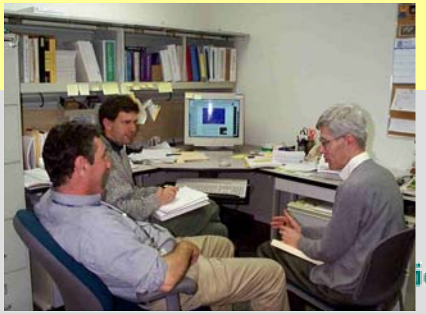
Sampling Design and DQOs | **Advanced Applied Math** | **Discovery Via Data Analytics/ Mining** | **Chemo-metrics Bio-informatics** | **Modeling and Simulation** | **Information Analytics** | **Experimental Design and Analysis**

- Brent Pulsipher, MS
- Rick Bates, MS
- Dick Gilbert, PhD
- Nancy Hassig, PhD
- Bob O'Brien, MS
- John Wilson, BS
- Denny Weier, PhD
- Alan Brothers, MS
- Brett Matske, MS
- Melissa Matske, MS

- Tom Ferryman, PhD
- Don Daly, PhD
- Kris Jarman, PhD
- Amanda White, MS
- Alan Willse, PhD
- Chad Scherrer, PhD
- Andrea Swickard, MS
- Ken Jarman, PhD
- Joel Malard, PhD

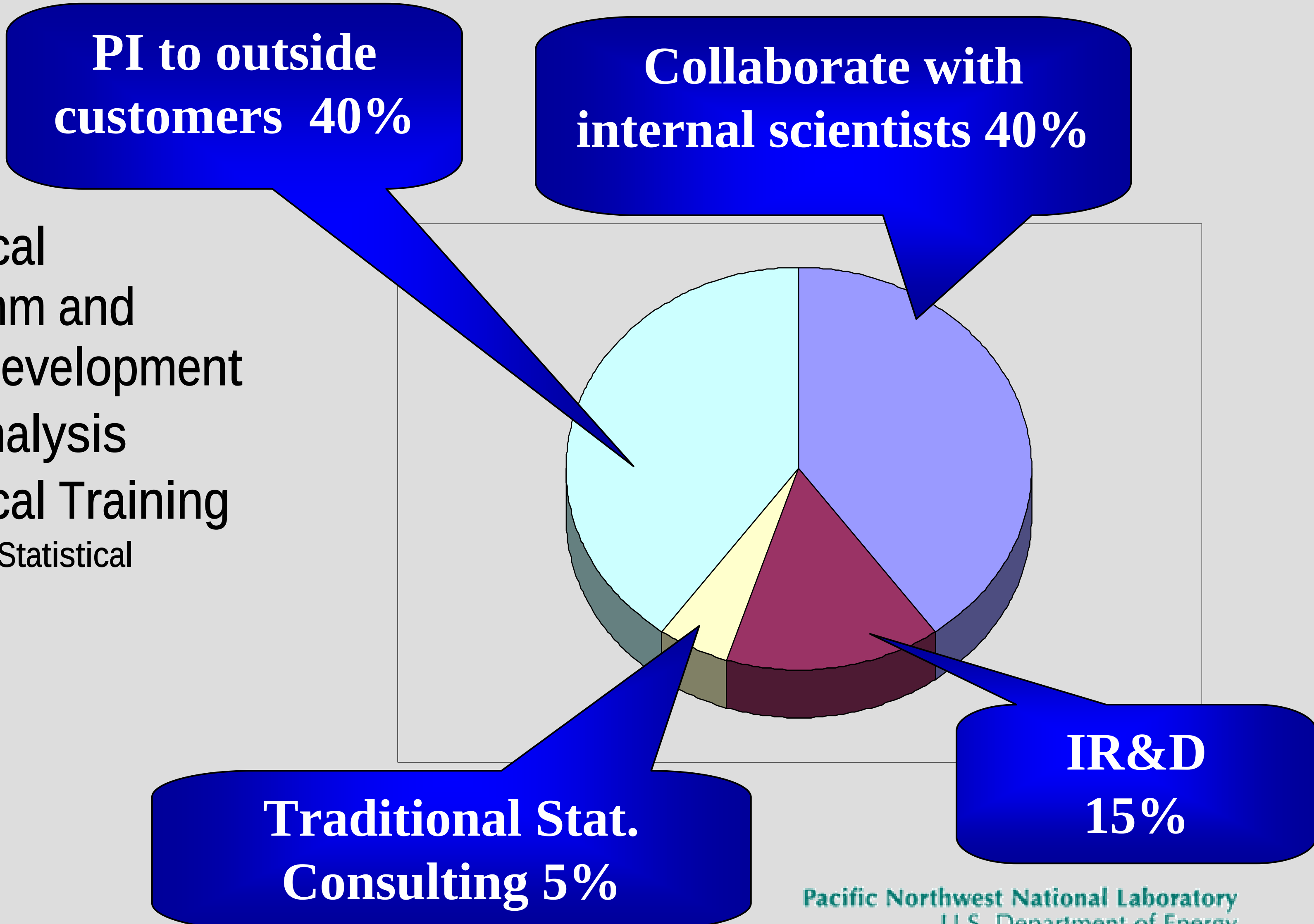
- Dale Anderson, PhD
- Kevin Anderson, PhD
- Dave Engel, MS
- Chuck LoPresti, MS
- Christian Posse, PhD
- Pat Heasler, MS
- Craig McKinstry, MS
- Al Liebetrau, PhD
- Nat Beagley, MS

- Paul Whitney, PhD
- Greg Piepel, PhD
- Brett Amidan, MS
- Sandra Thompson, PhD
- Stacey Hartley, MS
- Bobbi-Jo Webb-Robertson
- Scott Cooley, MS
- Deb Carlson, MS
- Alejandro Heredia-Langner, PhD



PNNL Statistical Products

- Statistical Algorithm and Tools Development
- Data Analysis
- Statistical Training
- Traditional Statistical Consulting

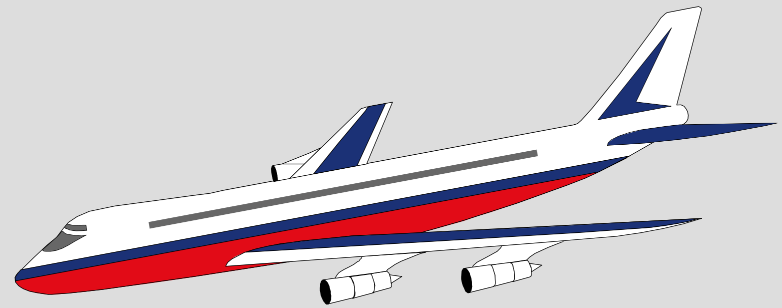


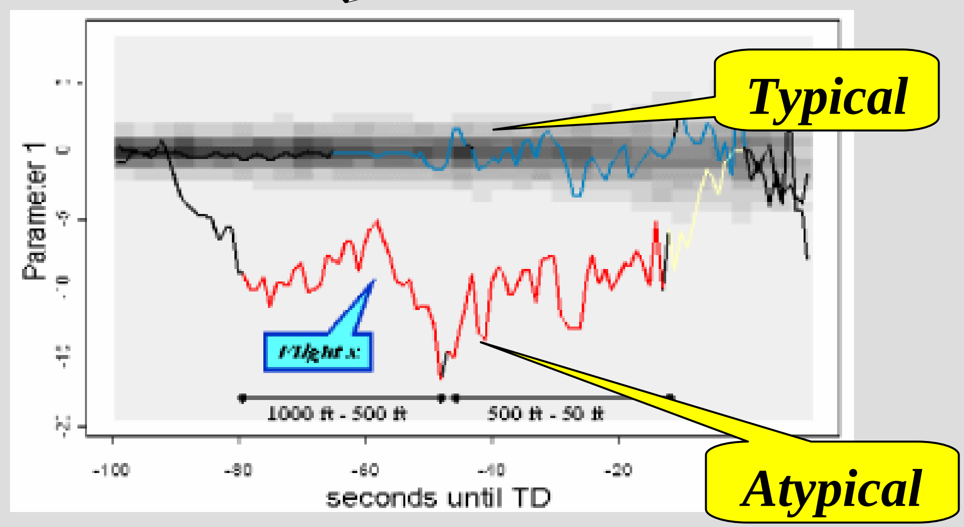
tom.ferryman@pnl.gov

A few examples of projects

In-flight Numerical and Categorical Data Analysis

- Goal: Build a pc-based workstation to allow individual airlines to automatically
  - Identify typical patterns
  - Find atypical flights
  - Find *unenvisioned relationships*
  - Investigate long term trends and cyclic patterns
- Data
  - Hundreds of flight variables measured every second on throughout a flight
  - Thousands of flights
  - Gigabytes of data
- Used by airlines today



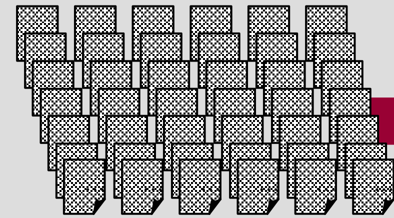


See the forest for the trees.
Find the needle in the haystack.

Analysis of Unstructure "dirty" Text

A Multi-Step, Multivariate Data Analysis Process

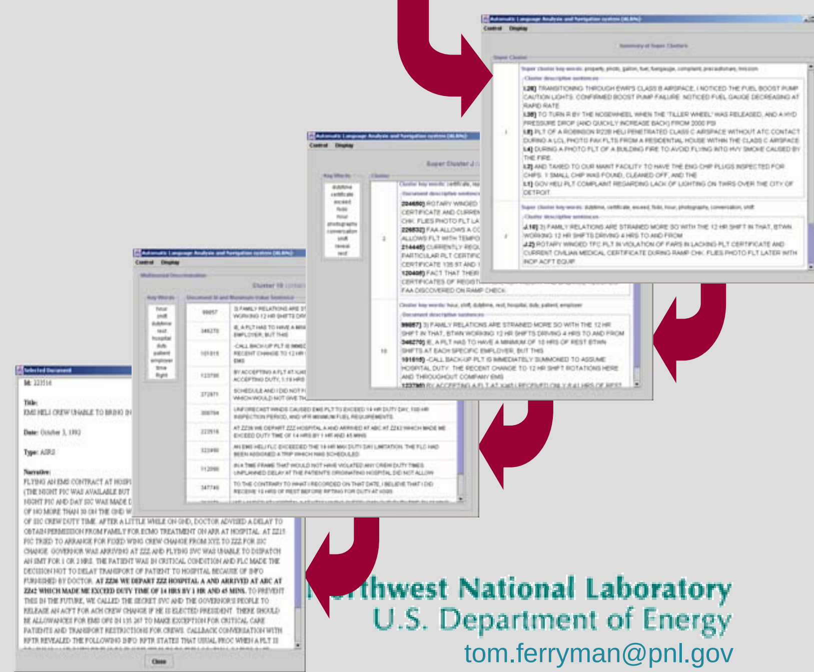
- Insight hidden in thousands of reports
  - Unstructured text
  - Numeric data
  - Categorical data
- Approach
  - Standardize the vocabulary.
  - Identify typical patterns, atypical reports
  - Retrieve by example capability
  - Display the analysis results in an intuitive and insightful manner.



Standardize vocabulary

Signature Generation

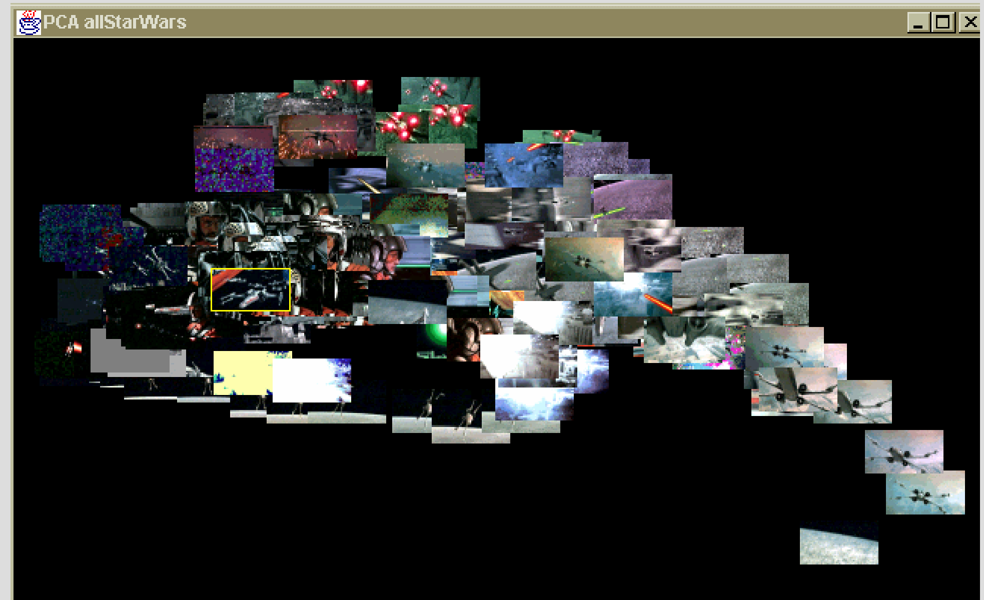
Multivariate clustering into Groups and Super-groups

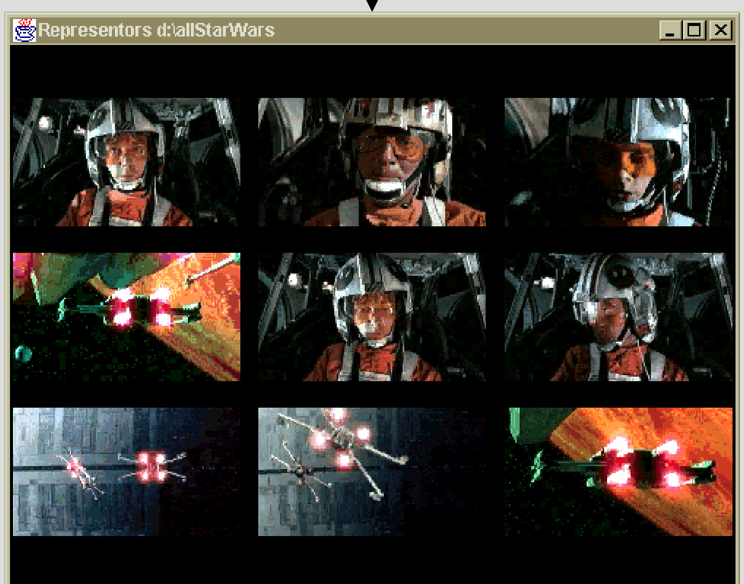


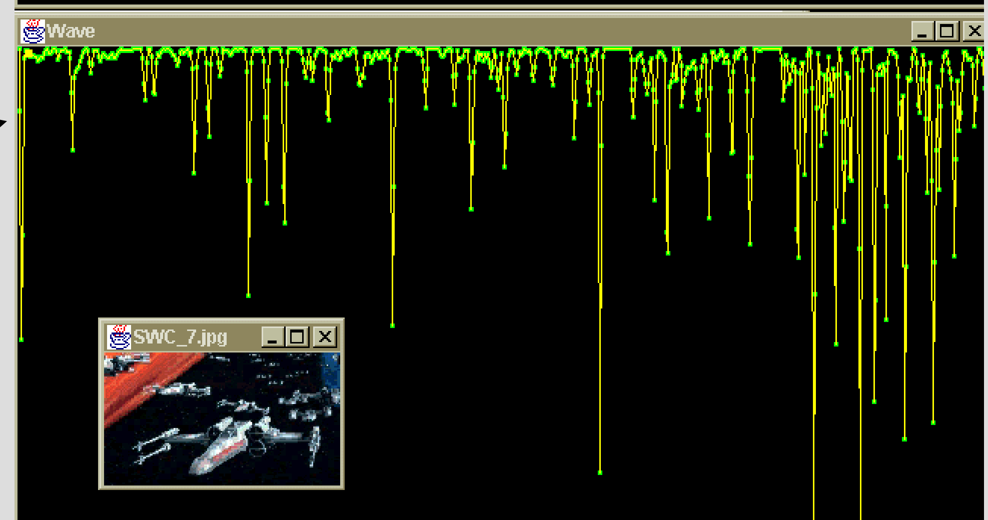
See the forest for the trees.
Find the needle in the haystack.

Video Clip Analysis: Segment and Summarize Sequential Images

- Sort image ensemble
- Estimate scene changes
- Calculated summary





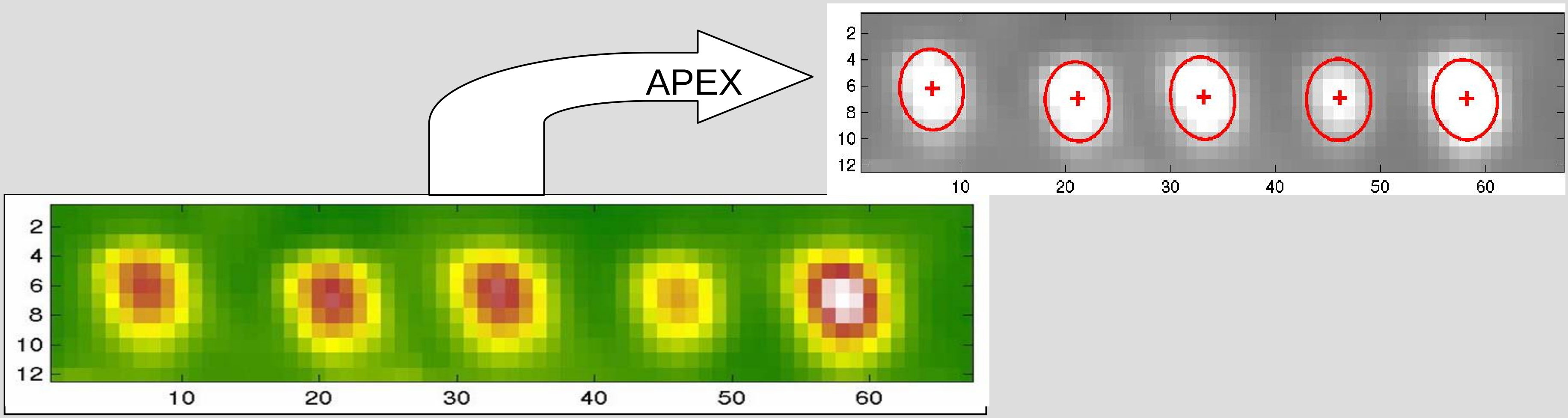


APEX Tool kit

Automated Peak Extraction for Mass Spec. data

Detect and Characterize Transient Features

- D Daly, K Jarman, K Anderson, K Wahl
- Stochastic foundation: Goodness Of Fit, uncertainty estimates
- US patent 6253162 B1 + Continuation in Part
- Peer-reviewed papers, tech. reports ...
- Licenses: 2+ com. + , 6+ gov. & univ.



tom.ferryman@pnl.gov

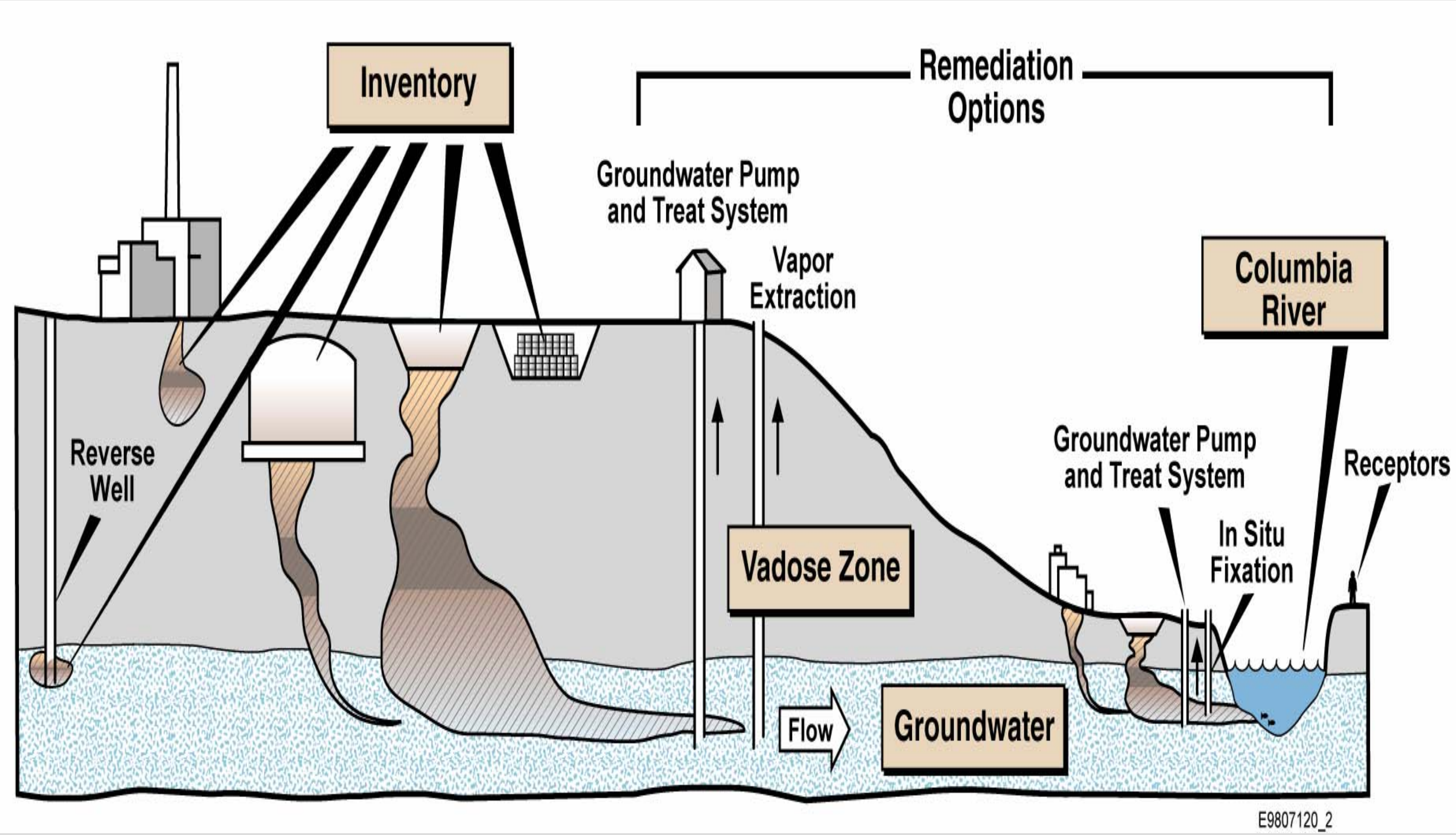
Scaling up to meet The challenges of the 21^{st} century

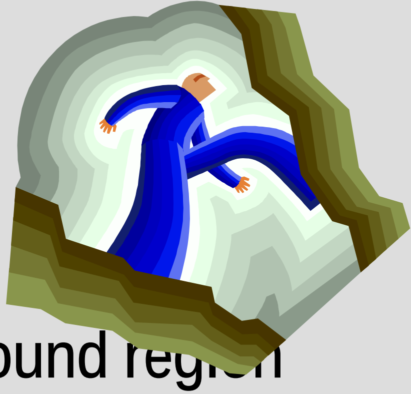
Big Big Problems

Hanford Site Integration Project

System Assessment Capability

Model Flow Schematic





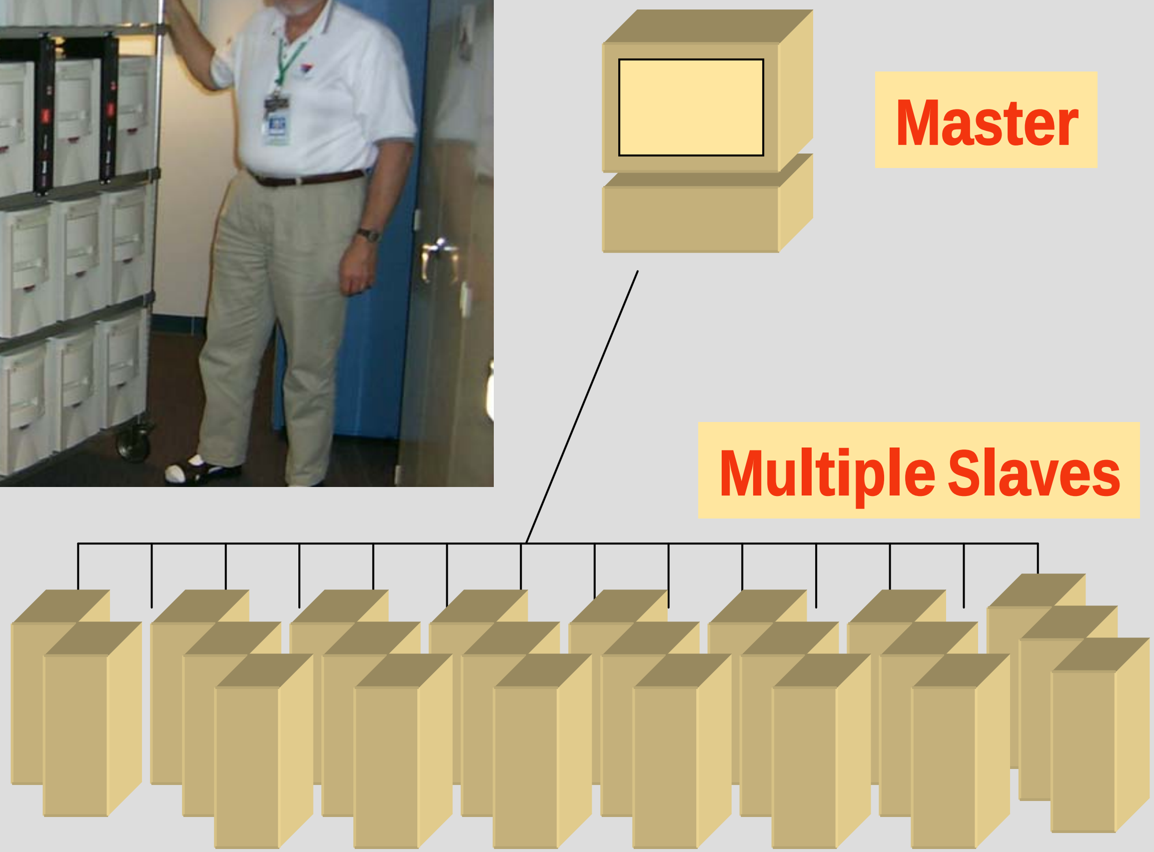
Computational Challenges

- Analysis Method: Simulate flow through the underground region
- Computational:
  - A single forward run currently requires ~3 hours clock time.
  - A full inverse run requires (as an example):
    - 20 parameters
    - 10 iterations
    - 20 attempts
    - 3 hours per simulation

      → 20*10*20*3 = 12000 hours = 1.4 years
- Very limited uncertainty analysis: 9 analytes, 25 Monte Carlos, takes 3 weeks on the 128 node parallel processing cluster.

Too big a problem
Too slow to really get insight

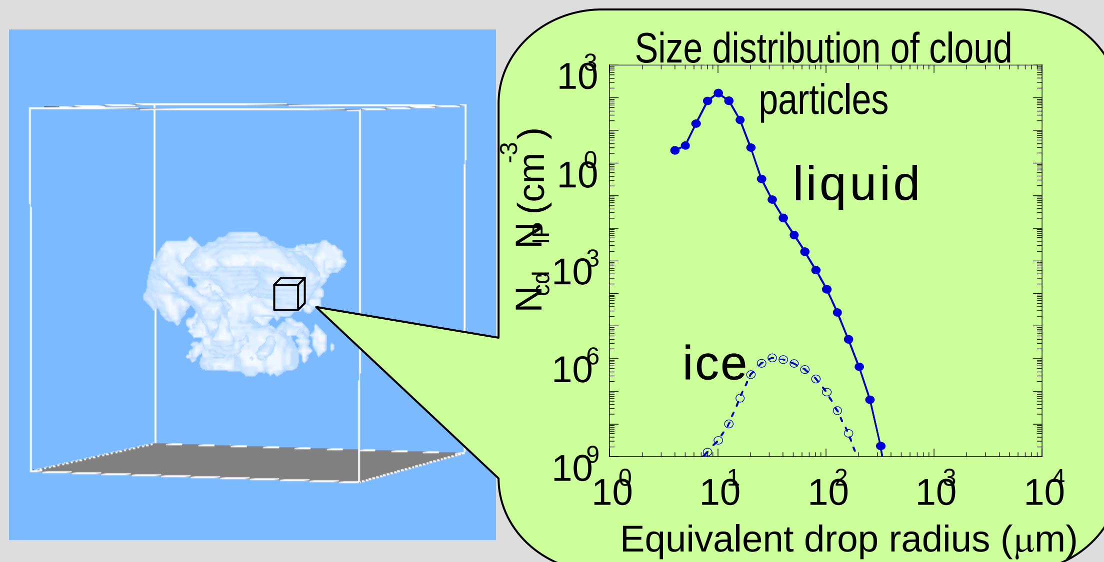
Distributed Computing Approach



Cloud model highlights

- 3-D nonhydrostatic dynamics
- Nonlinear interaction of thermodynamical, microphysical, and radiative processes
- 80+ variables: Dynamics & thermodynamics (*u, v, w, T, q, p*) and microphysics (cloud condensation nuclei - 12 size bins; liquid drops - 30 size bins; ice particles - 30 size bins)
- $75^3 \approx 500{,}000$ grid points

Processing takes 2 months on a huge parallel machine for one realization.



(After Ovtchinnikov and Kogan, *Journal of Atmospheric Science*, 2000)

21^{st} Century Science will investigate BIG problems. We need to find new solutions that will work.

- Explore large parameter space
  - 100 + parameters with just 5 levels = 7 x 10^{69}.
  - At one second per simulation this would need 2.5 x 10^{62} years.
- Quantify uncertainty about model
- Map the response surface
- Be fast
  - Allow scientist / computer interaction
  - Hypothesis explorations
  - "What if" investigations
- Handle huge data (tera-, peta- bytes)
- Robust to real world data: bad data, missing data

tom.ferryman@pnl.gov

Problem #1

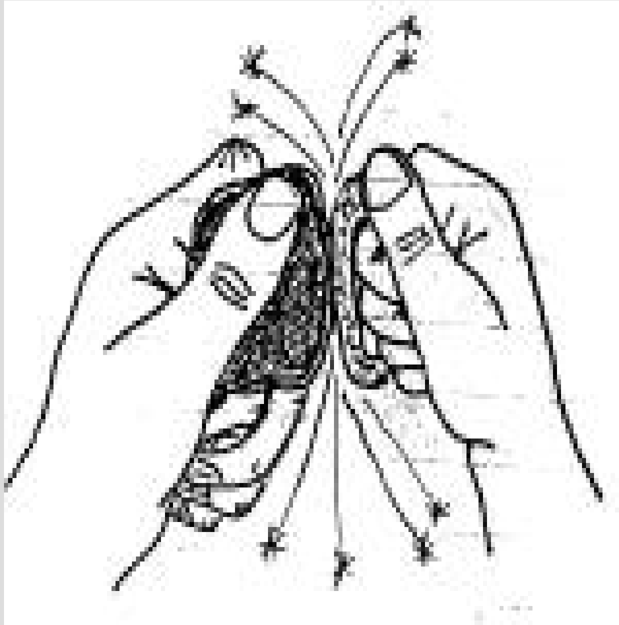
Big Computers (tera-, peta- flops)

Big Data (tera-, peta- bytes)

Can current algorithms handle this challenge?

Big Data Analysis – Prospects for Using Big Computers

- Data scales are readily encountered in which our typical tools fail due to the scale
- There's an increasing availability of multi-processor computers, and software



tom.ferryman@pnl.gov

Optimistic Vision

- Develop usable statistical analysis tools for big computing, with an eye towards these significant impacts
  1. *Many orders of magnitude increase in the scales of analyses that can be routinely addressed*
  2. An increase in the market for multiprocessor computers (more data analysts than computational chemists)
  3. Significant increase in capability can result in significant scientific discoveries with scientists using these improved tools.

Pacific Northwest National Laboratory
U.S. Department of Energy
tom.ferryman@pnl.gov

Optimistic Characteristic

- Existing analysis code/scripts could be compiled and run. Where possible, scaled up.
- Use familiar languages and user interfaces.
- Minimal overhead to convert from single processor to a suite of processors.

```
x <- rbind(matrix(rnorm(100000000, sd = 0.3), ncol = 2),
           matrix(rnorm(100000000, mean = 1, sd = 0.3),
                  ncol = 2))
cl <- kmeans(x, 2, 20)
plot(x, col = cl$cluster)
points(cl$centers, col = 1:2, pch = 8)
```


tom.ferryman@pnl.gov

Potential Resources

- Assorted multi-processor computers increasingly available
  - PNNL has some available
  - Many universities have some available
- Key support libraries exist for numerical computations
  - PNNL has made significant investments in the development of data management tools and specific application simulations.
  - PNNL has developed a beginning tool kit: Global Arrays
  - Others have similar seeds ready for use and refinements to mature
- Brain Power
  - Collaboration: Statistics, Mathematics, & Computer Science

tom.ferryman@pnl.gov

Big computing hardware at PNNL

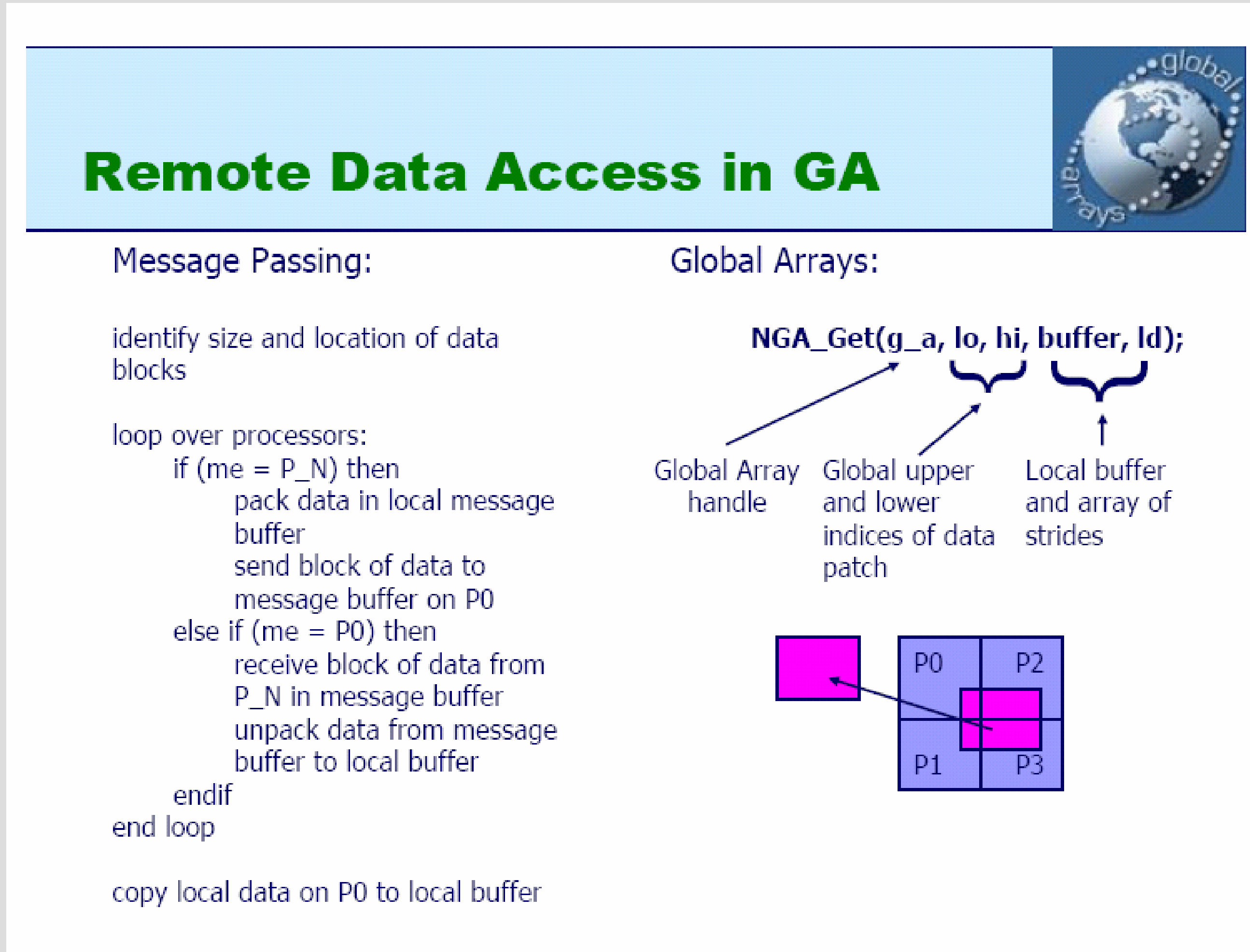
- Hewlett-Packard supercomputer
  - 11.8 teraflop system
  - 1400 processors
  - 3.8 terabytes RAM.

- Colony
  - 240-processor Linux cluster

tom.ferryman@pnl.gov

Big computing software at PNNL – Global Arrays

- Called from Fortran 77, C, C++, Python
- Provides support for data handling (abstracts memory management)
- Provides support for numerical analysis
- http://www.emsl.pnl.gov/docs/global/



tom.ferryman@pnl.gov

Next Steps?

- Formulate team
  - Who wants to play?
- Evaluate the market
  - Number of Big Computers
  - Dollar value of Big Computer sales
  - Typical job-types on Big Computers (e.g. computational chemistry)
  - Size of data analysis market that might be amenable to big computing
- Formulate technical approach and assess feasibility
- Plan a research program
- Go hunting for resources

Pacific Northwest National Laboratory
U.S. Department of Energy
tom.ferryman@pnl.gov

Problem #2
Quantifying Uncertainty in Modeling and Simulations

Beyond Monte Carlo

Quantifying Uncertainty in Complex Scientific Simulations

- Problem: Develop computationally efficient methods for local and global sensitivity and uncertainty analysis for complex computational scientific models with hundreds of uncertain input variables
- Application scientists need to be able to deal with increasing numbers of uncertain inputs, multiple conceptual models, model comparisons...

$$"\sigma^2_{\text{prediction}} = \sigma^2_{\text{model}} + \boxed{\sigma^2_{\text{input}}} + \sigma^2_{\text{numerical}} + \sigma^2_{\text{geometry}} + \cdots"$$

tom.ferryman@pnl.gov

Problem Drivers

Applications

- Subsurface: Contaminant fate
  - High variability in geologic properties, boundary conditions
- Biological Systems: Biochemical kinetics
  - Uncertain/unknown kinetic parameters, pathway structure
- Climate: Aerosols, high-resolution cloud models
  - Aerosol nucleation highly sensitive to microphysics
  - Large discrepancies in cloud model comparisons

Universal Challenges

- Quantifying confidence in predictions is critical to decision-making
- Advances in complex scientific simulations will require a leap in computational efficiency for uncertainty analysis
  - Current science poses problems for which uncertainty analysis is very limited or impossible
  - Dimensionality of uncertain parameters continually increases
  - Advancing scientific needs will exacerbate the challenge

Pacific Northwest National Laboratory
U.S. Department of Energy
tom.ferryman@pnl.gov

Sampling Input Space

- Standard simulation procedure – use Monte Carlos with many runs
  - Input variables are assigned joint probability distribution and sampled
  - Code is run many times to compute output for resulting input vector
  - Input distribution → output distribution
- Numerous ways to improve upon current practice
  - Improved sampling strategies, sensitivity analysis, screening, response surface modeling
  - Non-sampling methods for sensitivity and uncertainty estimates
- State of the art sampling designs will require far too many runs
- Future need:
  - Reduce reliance on Monte Carlo
  - Improve efficiency → deal with more uncertain variables
  - Improve global assessment of uncertainty

Pacific Northwest National Laboratory
U.S. Department of Energy
tom.ferryman@pnl.gov

Next Steps?

- Formulate team
  - Who wants to play?
- Evaluate the market
  - Identify research programs limited by current Monte Carlo techniques
    - Unable to explore full parameter space
    - Unable to estimate response surface variability
- Formulate technical approach and assess feasibility
- Plan a research program
- Go hunting for resources

Problem # 3

Missing Data,
Not at random

How to handle missing data?

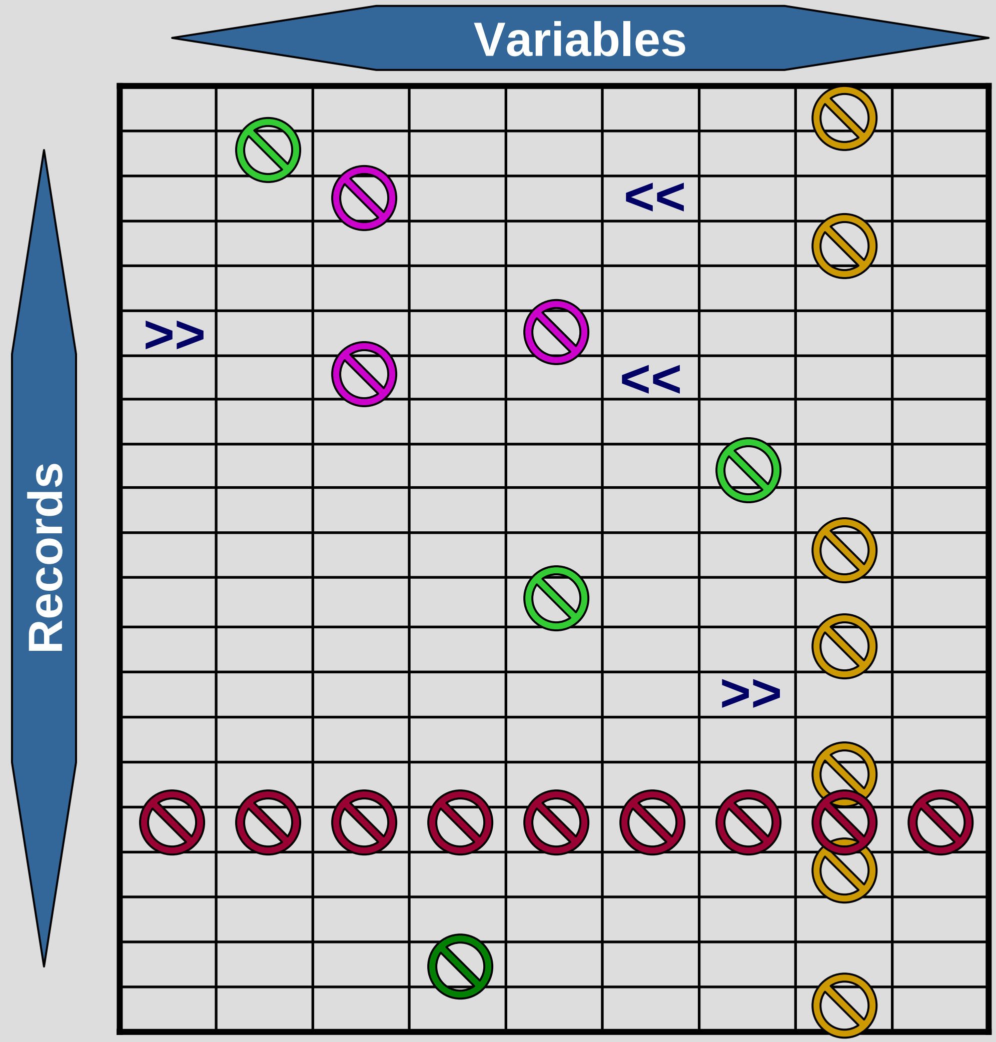
- Typical: assume data is missing at random and use EM algorithm or similar method
- Many applications can NOT make this assumption

- Data might be missing due to:
  - Identified as bad data
    - 20 millions parts per million in a chemical concentration assay.
    - 300,000 feet/sec altitude decrease in an aircraft that did not crash.
  - Not tested/collected due to:
    - Prior beliefs
    - Related variable values
    - Program budget or schedule constraints
    - Political, privacy, policy, legal decisions
    - …

tom.ferryman@pnl.gov

Missing data can have varying characteristics

- Data missing randomly in various cells
- Data missing in various cells but not believed to be random
- Data missing for entire record, or most of record
- Data censored, too high or too low
- Variables with low probability of being available



Battelle
Pacific Northwest National Laboratory
U.S. Department of Energy
tom.ferryman@pnl.gov

Can we find ways to do analysis without drastically reducing the available data?

- Currently, we use
  - EM algorithm or other method to impute values

  or
  - Data removal
    - Drop records with many missing variables
    - Drop variables with low probability of valid data
    - Iterate until data matrix is full

  or
  - Use a rather cumbersome conditional algorithm
- Do better methods exist now?
- Could we develop better methods?

Pacific Northwest National Laboratory
U.S. Department of Energy
tom.ferryman@pnl.gov

Next Steps?

- Formulate team
  - Who wants to play?
- Evaluate the market
  - Skip? This is ubiquitous.
- Formulate technical approach and assess feasibility
- Plan a research program
- Go hunting for resources

Closing Remarks

Closing Remarks
How can we collaborate?

- Conducting research
  - Remote collaboration
  - Professors: come work with us over the summers or take your sabbatical at PNNL
  - 3, 4, 5 year PhD students visits to PNNL (for 3 months or more)
  - Post-docs
- Joint pursuit of funding
  - Formulate joint research programs
  - Propose to funding agencies
    - Universities to NSF, DOE, DARPA, HSARPA, …
    - PNNL to DOE, DARPA, DOD, …

tom.ferryman@pnl.gov