Nonparametric Multivariate Anomaly Analysis in Support of HPC Resilience

G. Ostrouchov, T. Naughton, C. Engelmann,
G. Vallée, and S. L. Scott

Computer Science and Mathematics Division
Oak Ridge National Laboratory

Workshop on Computational e-Science at the
5th IEEE International Conference on e-Science

December 7-9, 2009
Oxford, UK

This work was partially funded by the Department of Defense and the Extreme Scale Systems Center at ORNL

Collaborators

Stephen W. Poole

Chief Scientist, Computer Science and Mathematics Division
LCF System Architect, National Center for Computational Sciences
Technical Director, Extreme Scale Systems Center
Oak Ridge National Laboratory

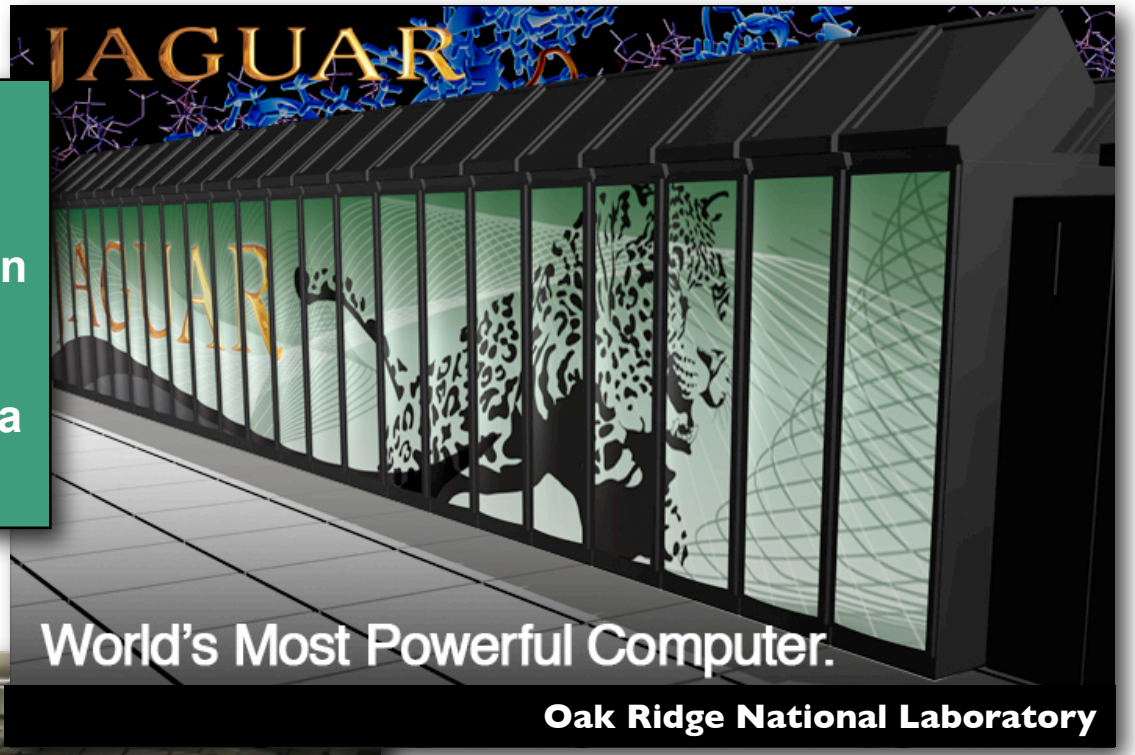
Chokchai (Box) Leangsuksun

Associate Professor of Computer Science
The SWEPCO Endowed Professor
Center for Entrepreneurship and Information Technology
Louisiana Tech University

Equipment for Computational Science: Component Count is Increasing

top500.org processor count:

- About three years ago the entire 500 list broke the million processor mark
- Now the top 7 add up to over a million



#1

#3

George Ostrouchov - ostrouchovg@ornl.gov

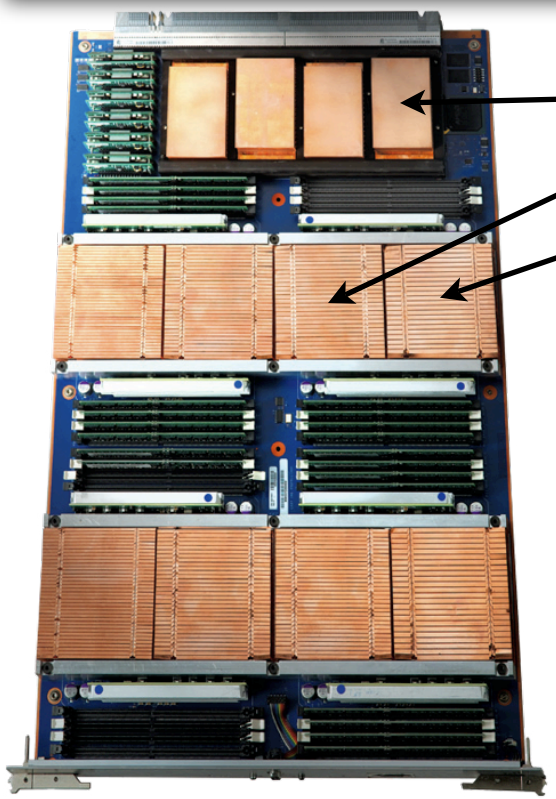
Jaguar Has Over Quarter Million Cores

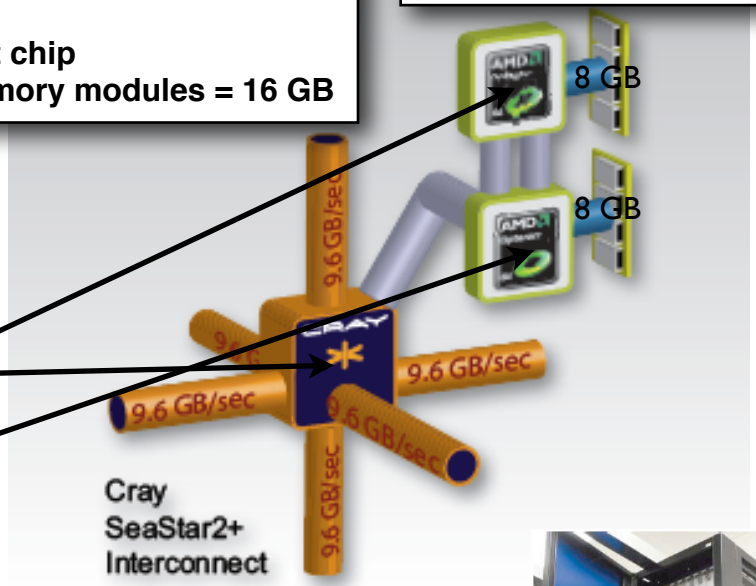
Blade = 4 Nodes
8 processors
48 cores
4 interconnect chips
16 (4 GB) memory modules = 64 GB
6 voltage converters

Node = 2 Processors
6 cores per processor

1 interconnect chip
4 x (4 GB) memory modules = 16 GB

Processor = 6 Cores
2 memory modules





Cabinet = 24 Blades
1152 cores
96 interconnect chips
384 memory modules (1.5 TB)
144 voltage converters
+ power supply, liquid cooling, etc.
Power 480V, ~40,000 Watt per cabinet

Jaguar = 284 cabinets (XT5 and XT4), ~ 6.5 Megawatts

Cray XT5

Reactive and Proactive Fault Tolerance

- **Reactive: Keeps applications alive through recovery from experienced failures**
  - **Checkpoint/restart**
  - **Message logging/replay**
  - **Effective until failures get too frequent**
  - **Timely failure reporting for restart**
  - **Root cause analysis for repair**
- **Proactive: Keeps applications alive by avoiding failures**
  - **Preemptive migration**
  - **Effectively decreases failure rate**
  - **Needs failure prediction**

George Ostrouchov - ostrouchovg@ornl.gov

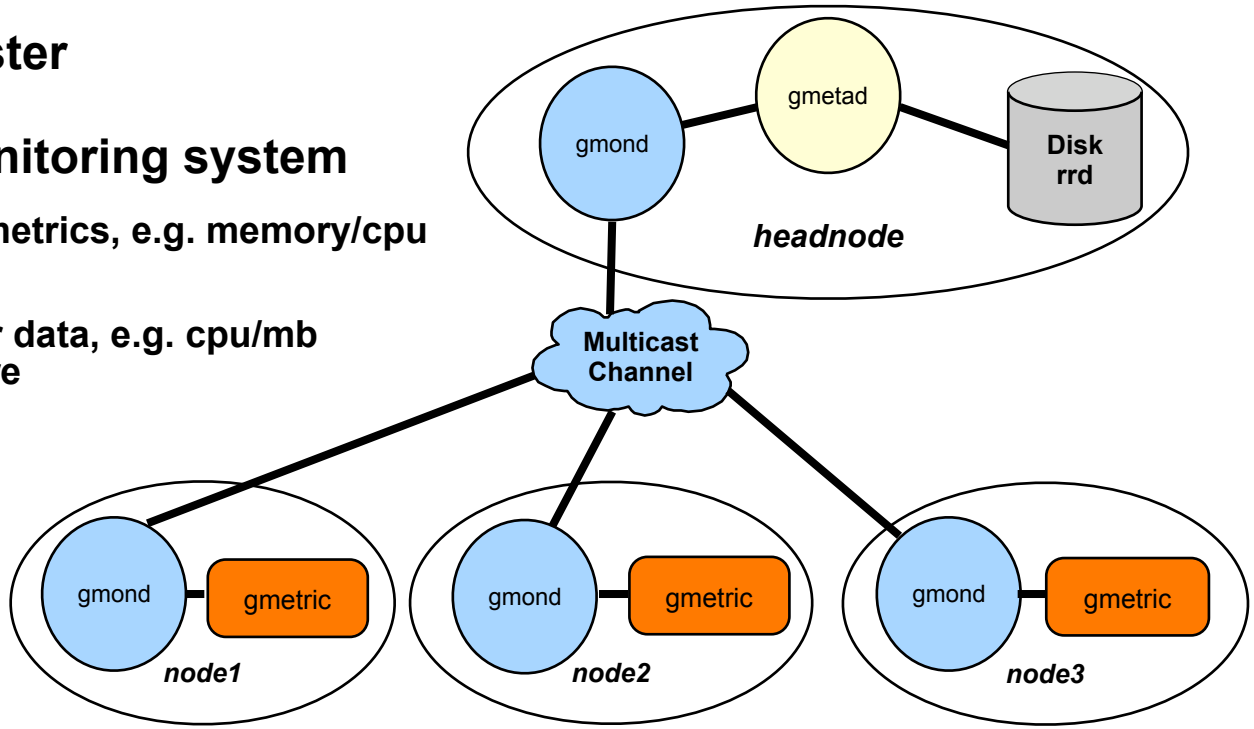
Reactive and Proactive Fault Tolerance

- **Reactive: Keeps applications alive through recovery from experienced failures**
  - **Checkpoint/restart**
  - **Message logging/replay**
  - **Effective until failures get too frequent**
  - **Timely failure reporting for restart**
  - **Root cause analysis for repair**
- **Proactive: Keeps applications alive by avoiding failures**
  - **Preemptive migration**
  - **Effectively decreases failure rate**
  - **Needs failure prediction**

George Ostrouchov - ostrouchovg@ornl.gov

Monitoring Prototype / Data Collection

- XTORC cluster
- Ganglia monitoring system
  - Standard metrics, e.g. memory/cpu utlization
  - LM_sensor data, e.g. cpu/mb temperature



- Leveraged RRD reader utilities from Ovis v1.1.1
  - RRD readers - read tge RRD databases
  - RRD starter - start instances of RRD readers on all nodes
  - RRD killer - kills all RRD readers
- Format metrics as single line per node

XTORC

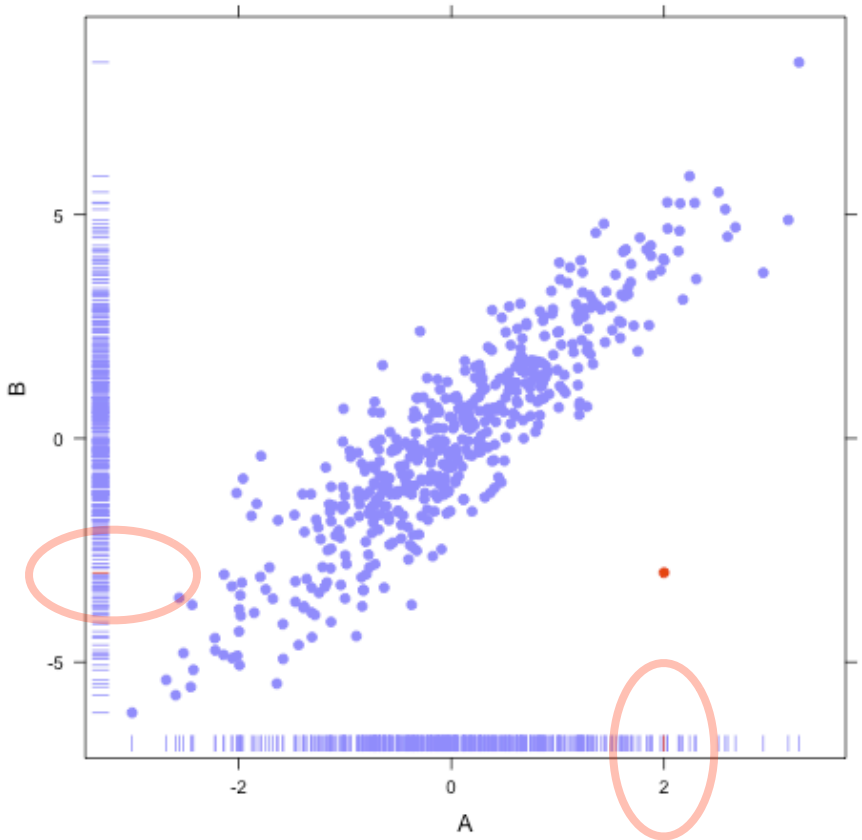
- **Hardware**
  - Compute nodes: ~45-60 P4 (2 Ghz)
  - Head node: 1 P4 (1.7Ghz)
  - Service/log server: 1 P4 (1.8Ghz)
  - 100 Mb Ethernet
- **Software configuration**
  - Spans RedHat 9, Fedora Core 4 & 5
    - RH9: node53
    - FC4: node4, 58, 59, 60
    - FC5: node1-3, 5-52, 61
  - RH9 is Linux 2.4
  - FC4/5 is Linux 2.6
  - NFS mounted /home

Anomaly Detection: Nonparametric Multivariate Analyzer

- Ability to view groups of components as statistical distributions
  - Identify anomalous components
  - Identify anomalous time periods
- Based on numeric data with no expert knowledge for grouping
  - Scalable approach, only statistical properties of simple summaries
  - Power from examination of high-dimensional relationships
  - Visualization utility used to explore data
- Implementation uses
  - *R* project for statistical computing
  - *GGobi* visualization tool for high-dimensional data exploration

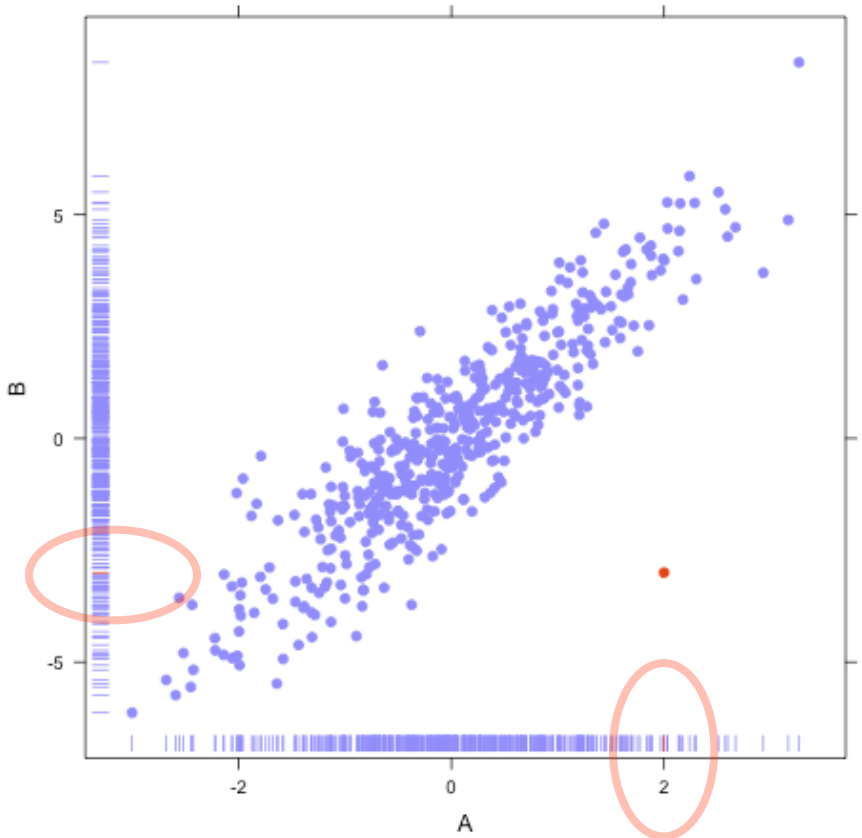
- With good failure data, could be used for failure prediction

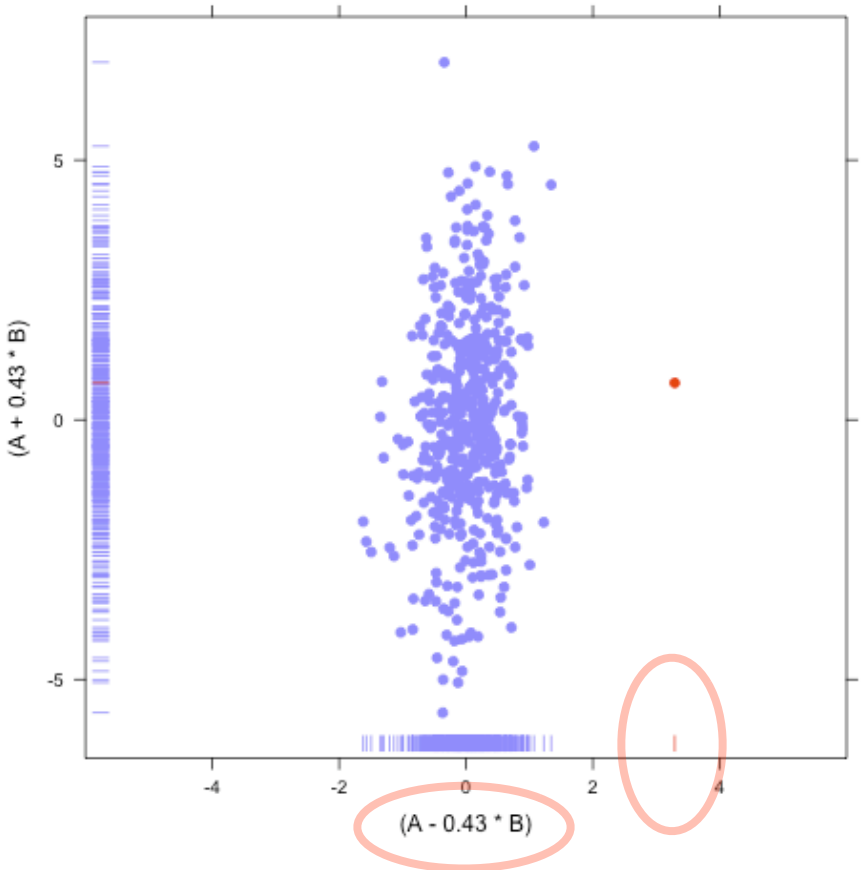
Why do Multivariate Methods Matter?



- **Outlier not visible by univariate methods**

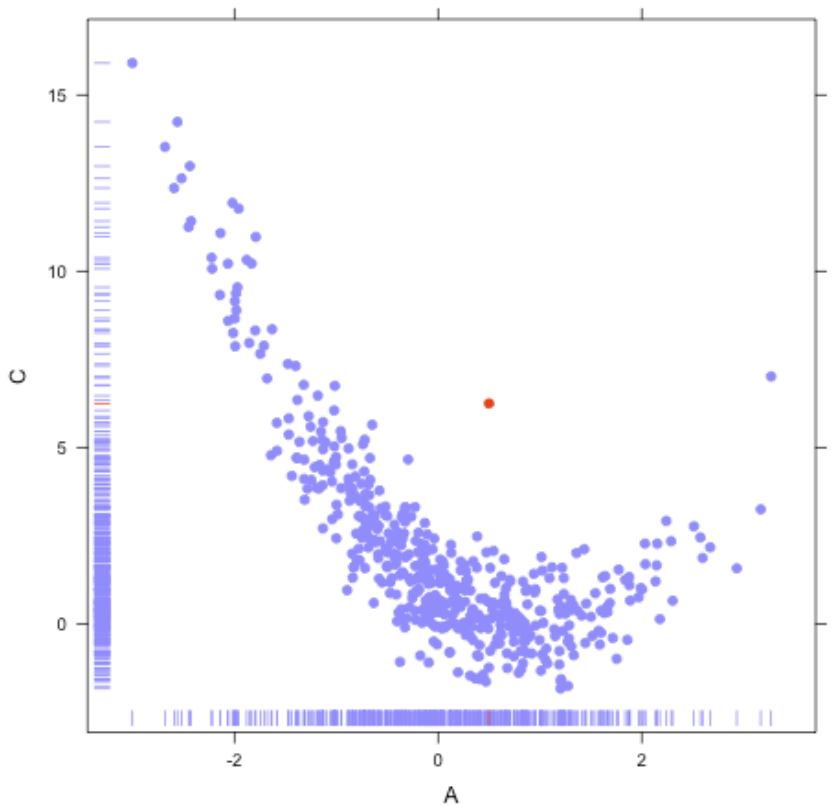
Why do Multivariate Methods Matter?





- **Outlier not visible by univariate methods**
- **Know the right feature to compute**

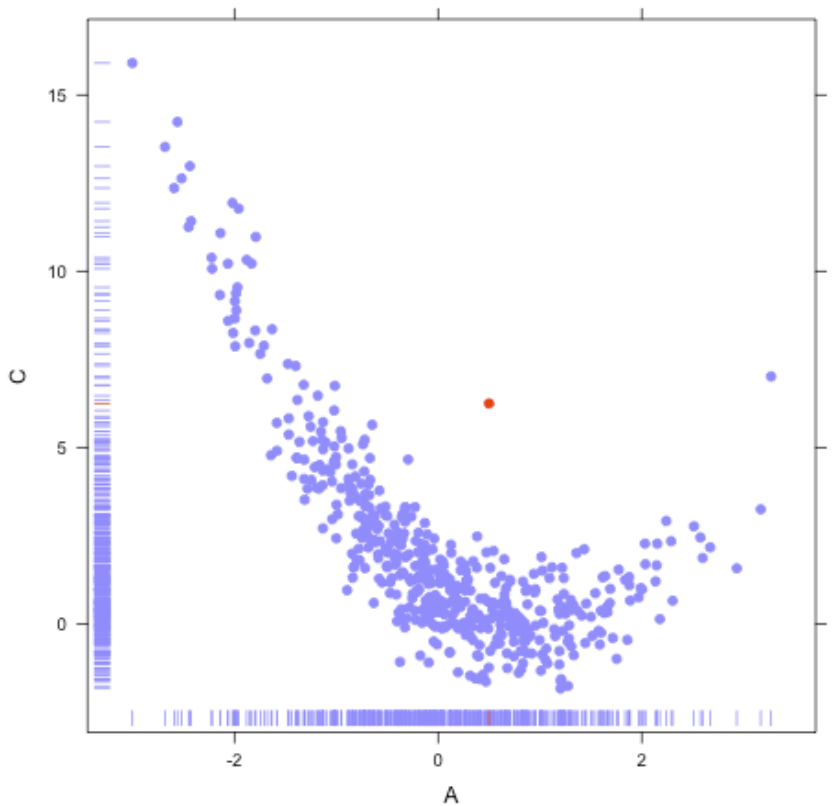
What if Nonlinear (Non-Gaussian) Data?



George Ostrouchov - ostrouchovg@ornl.gov

What if Nonlinear (Non-Gaussian) Data?

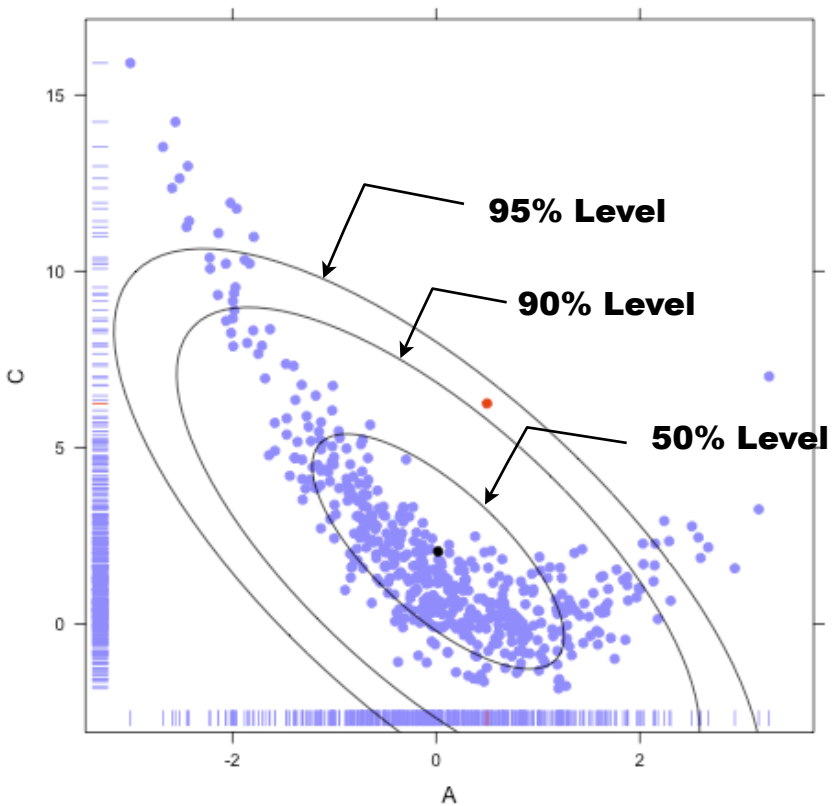
- **No linear transformation separates anomaly**



George Ostrouchov - ostrouchovg@ornl.gov

What if Nonlinear (Non-Gaussian) Data?

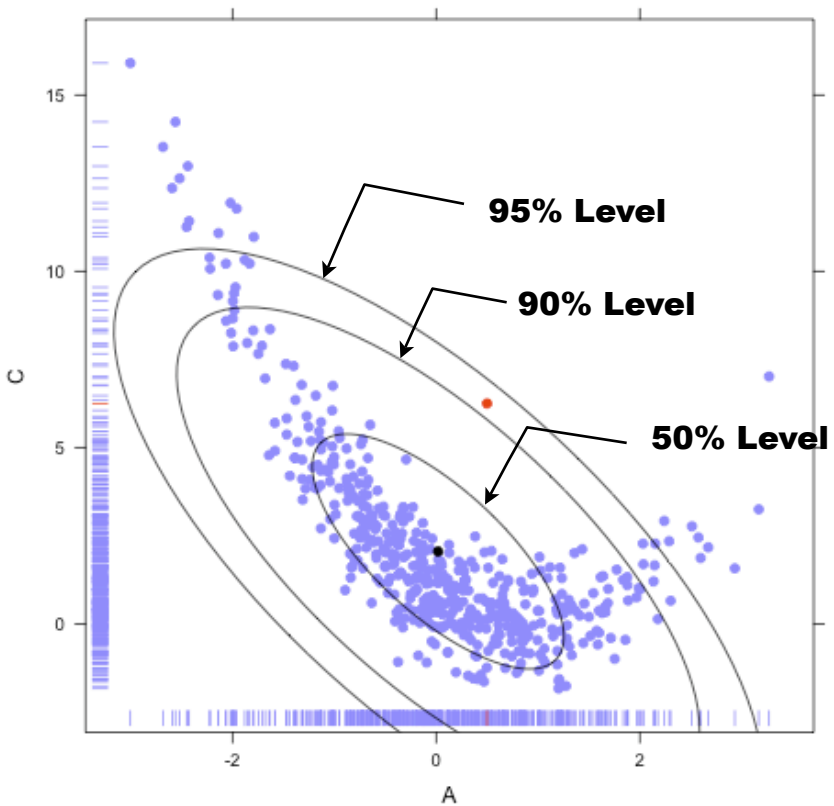
- No linear transformation separates anomaly
- Mahalanobis distance
  - Based on global Gaussian assumption
  - Extremes can have greater distance than anomalies



George Ostrouchov - ostrouchovg@ornl.gov

What if Nonlinear (Non-Gaussian) Data?

- No linear transformation separates anomaly
- Mahalanobis distance
  - Based on global Gaussian assumption
  - Extremes can have greater distance than anomalies
- Nearest neighbor distance
  - Related to nonparametric density estimation (Hartigan, 1975)
  - Single Linkage Clustering
  - Minimum Spanning Tree

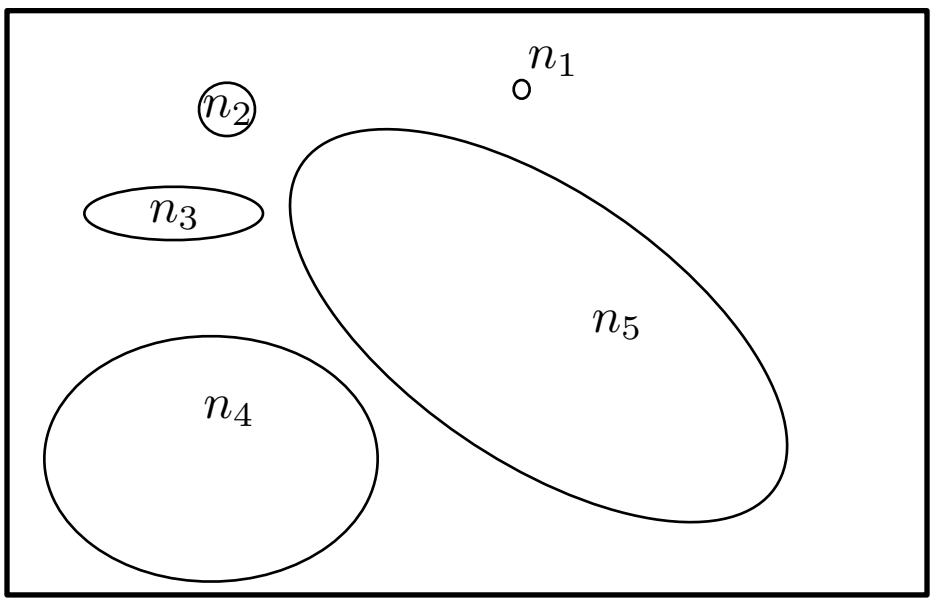


Nonparametric Multivariate Analyzer

Cluster Size Anomaly Measure

- **k clusters, sizes** $n_1, n_2, \ldots, n_k$
- **Anomaly measure inversely related to cluster size**

$$\alpha_i = \frac{1}{n_i}$$



- **Sampling from a multivariate density**
- **Density proportional to probability integral over cluster region**
- **Measure of cluster anomaly**
- **Sum over all equal or smaller clusters**

$$p_i = n^{-1} \sum_{n_j \leq n_i} n_j$$

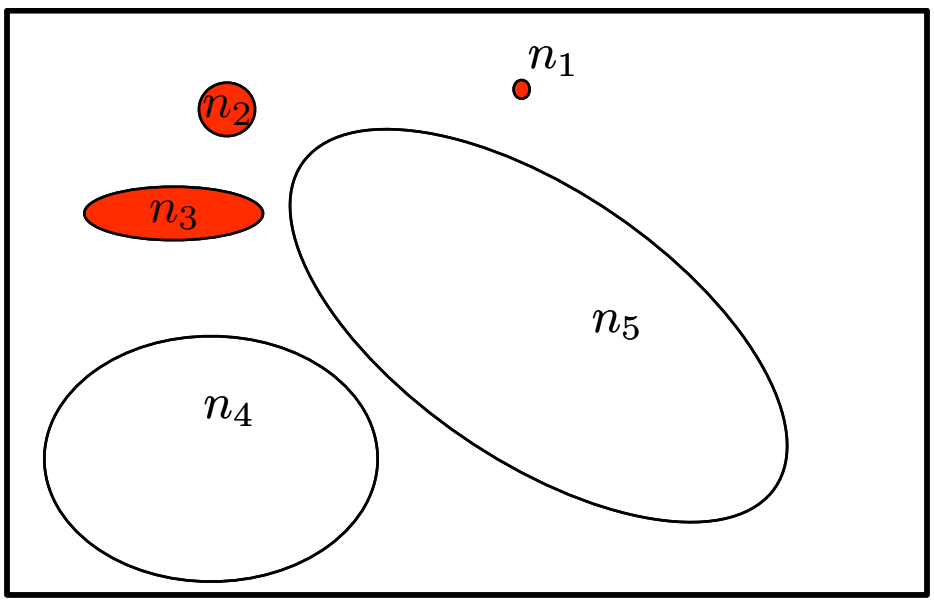
$$\alpha_i = 1 - p_i$$

- **Interesting to explore as we go down the dendrogram to more clusters**

Cluster Size Anomaly Measure

- **k clusters, sizes** $n_1, n_2, \ldots, n_k$
- **Anomaly measure inversely related to cluster size**

$$\alpha_i = \frac{1}{n_i}$$



- **Sampling from a multivariate density**
- **Density proportional to probability integral over cluster region**
- **Measure of cluster anomaly**
- **Sum over all equal or smaller clusters**

$$p_i = n^{-1} \sum_{n_j \leq n_i} n_j$$

$$\alpha_i = 1 - p_i$$

- **Interesting to explore as we go down the dendrogram to more clusters**

Applied to an Old Idle Cluster

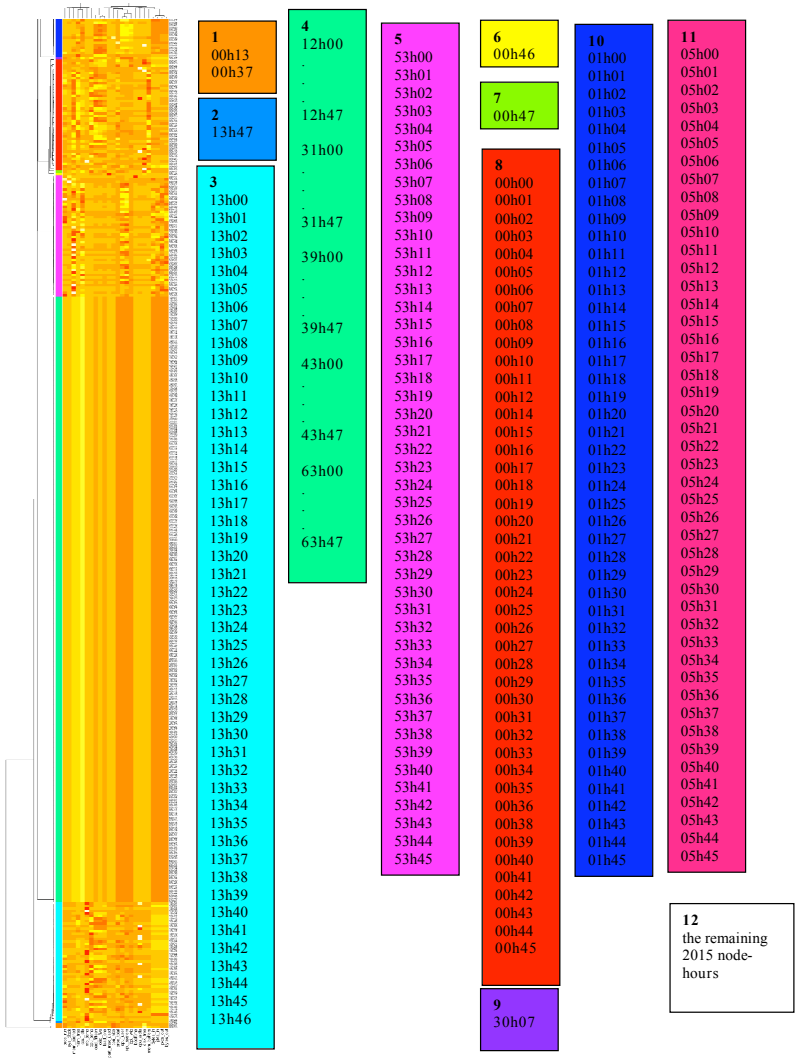
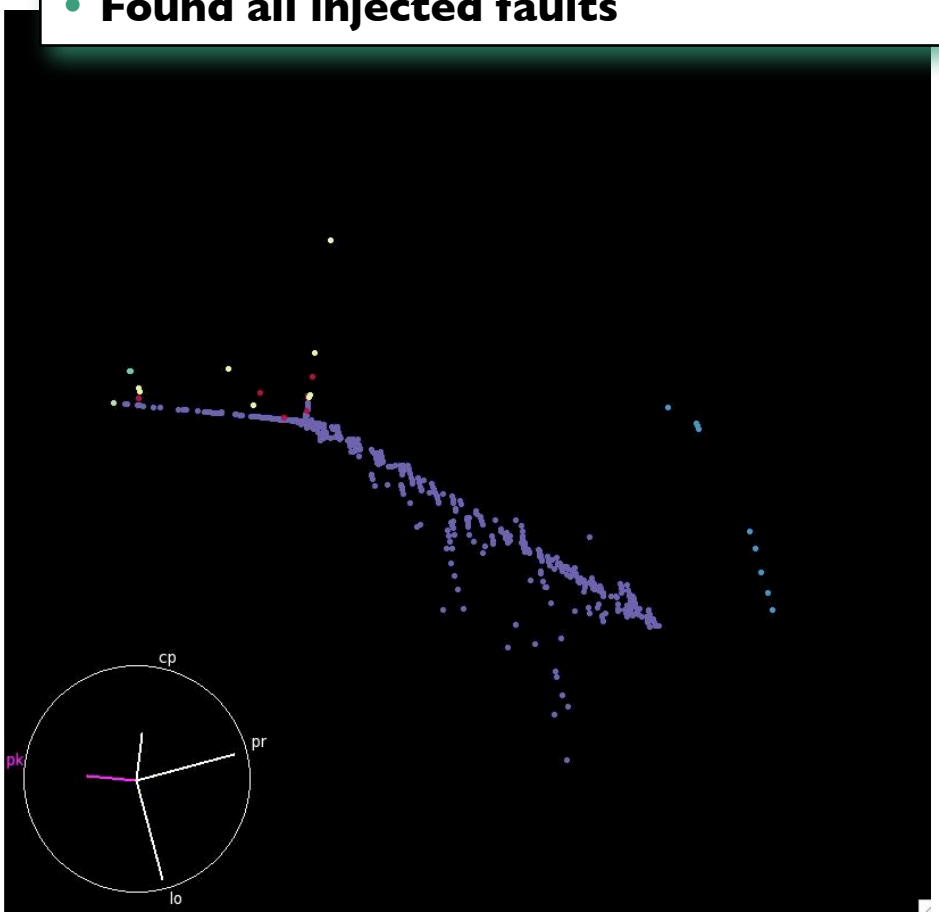


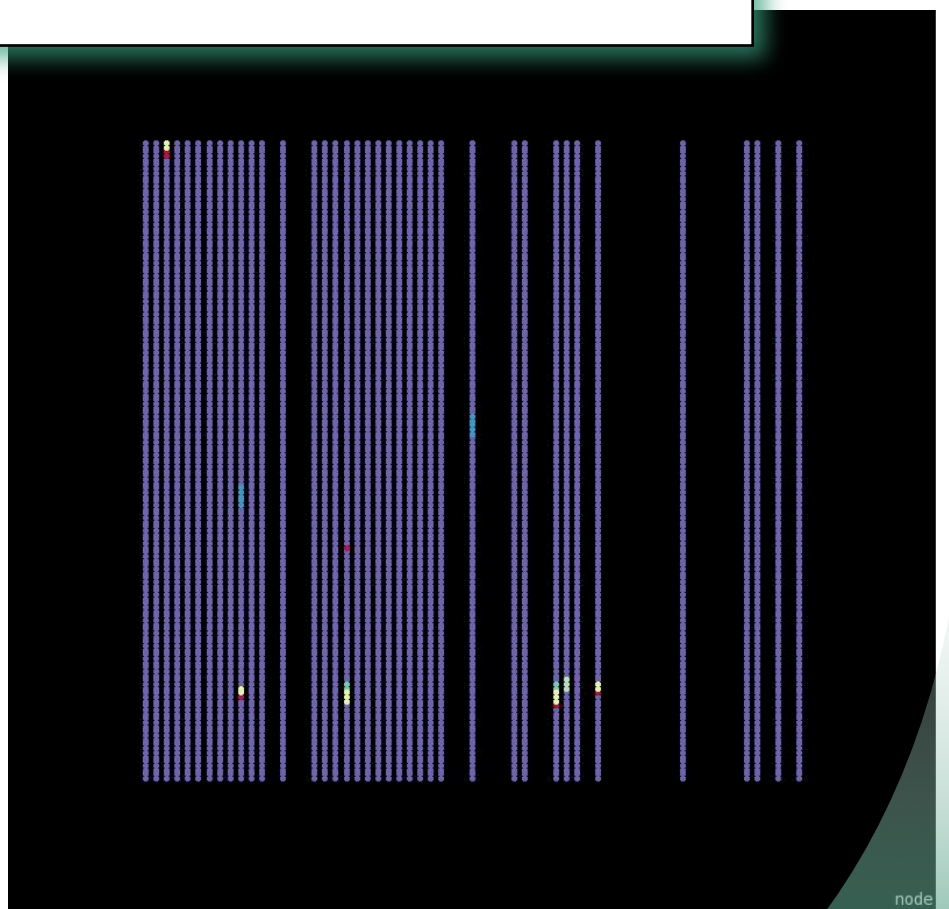
Figure 3 A cluster diagram on left shows the most interesting 400 (out of 2,496) node-hours as rows. Cluster lists are expanded on the right with matching colors.

- **Node 0 is the most different from the rest, particularly its hours 13, 37, 46, and 47. This is the head node where most services are running. Note that hours 13 and 37 are 24 hours apart, indicating a daily service.**
- **Node 53 (all hours) runs the older Red Hat 9 while the others run Fedora Core 4/5.**
- **Nodes 12, 31, 39, 43, and 63 were all down.**
- **Node 13 was distinct, particularly hour 47, for undetermined reasons.**
- **Node 1, Node 5, and Node 30 (hour 7) were also distinct for undetermined reasons.**

Injecting Faults on XTORC

- Clustering node-minute data (R/GGobi)
- Anomaly measure (purple-blue-yellow-red) proportional to cluster size
- Found all injected faults



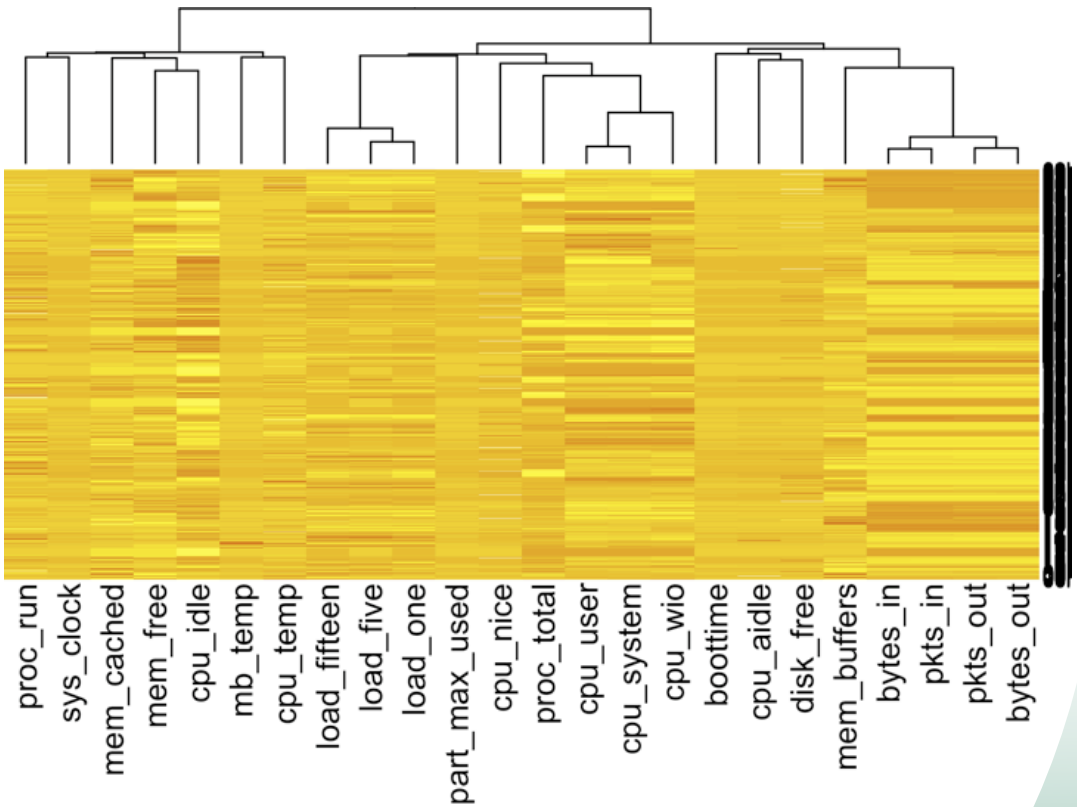


Identifying anomalies: first step to identifying failures and building a failure prediction capability

George Ostrouchov - ostrouchovg@ornl.gov

Attributes Can be Clustered for Similarity

- Three groups
  - temperature and memory
  - processor related
  - I/O related
- Finer groups
  - bytes/packets in
  - bytes/packets out
- Gives ordering of attributes



Automation!

Anomaly Detection - Next Steps

- Data
  - Reduce overhead in data gathering
  - Apply to data from Jaguar
  - Monitor more fields (ex. HDD)
  - Investigate methods to aid data interpretation
  - Identify significant fields for given workloads
  - Heterogeneous nodes
- Different workloads
  - Base (no/low work)
  - Loaded (benchmark/app work)
  - Loaded + Fault Injection
- Work toward links between anomalies and failures