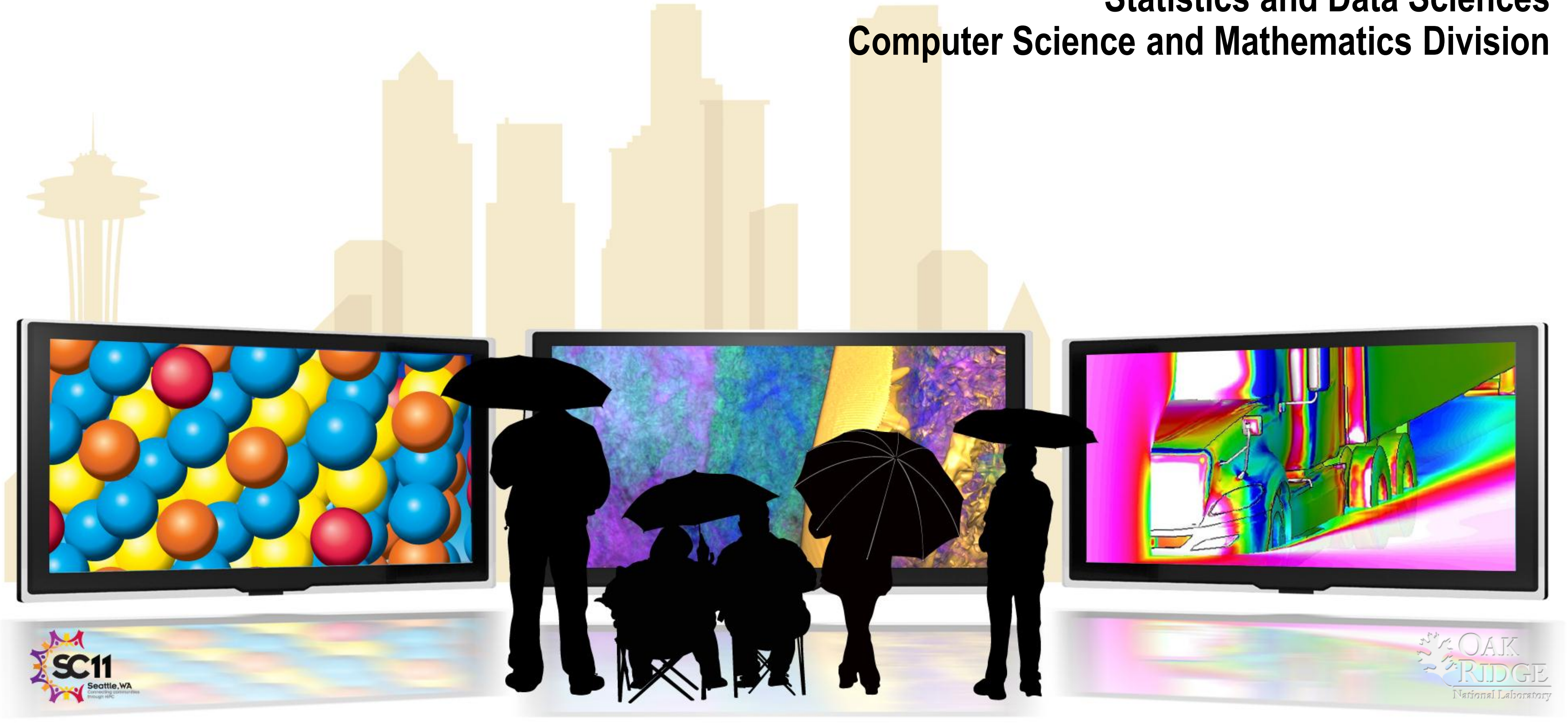
# Statistics at Scale

Presented by

**George Ostrouchov**

Statistics and Data Sciences
Computer Science and Mathematics Division

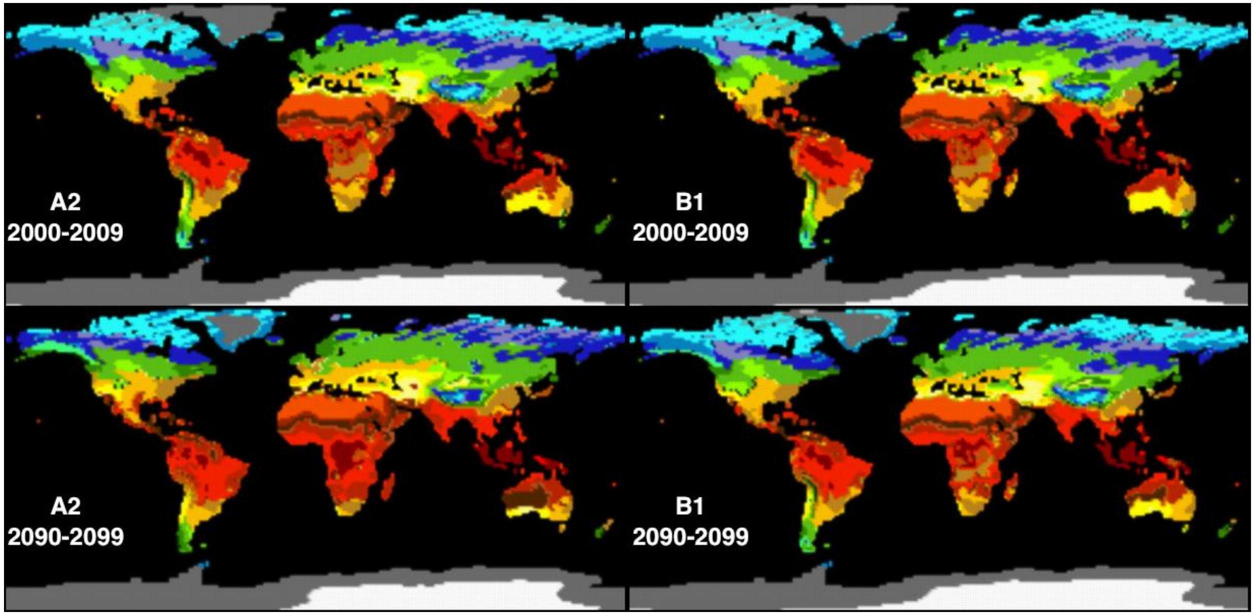


SC11 Seattle, WA Connecting communities through HPC

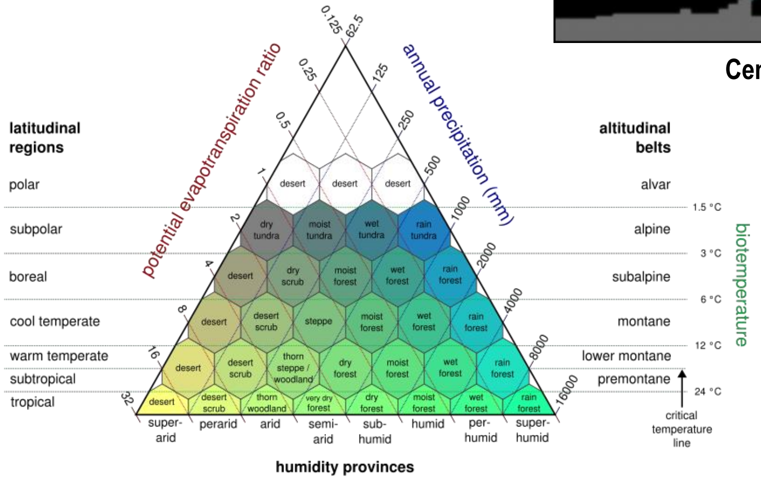
Oak Ridge National Laboratory

# Multivariate Comparison of Climate Simulations

- A multivariate classification capability as a VisIt plugin
- Demonstrated with Holdridge life zones to compare climate simulations



Century comparison of Holdridge life Zones under climate scenarios A2 and B1

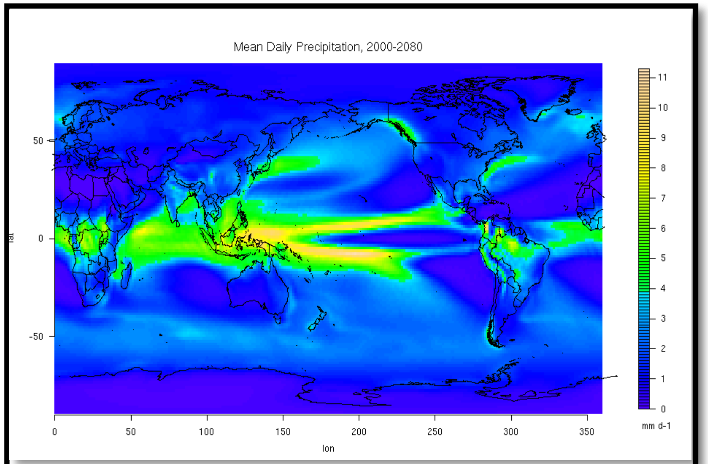


Holdridge Life Zones

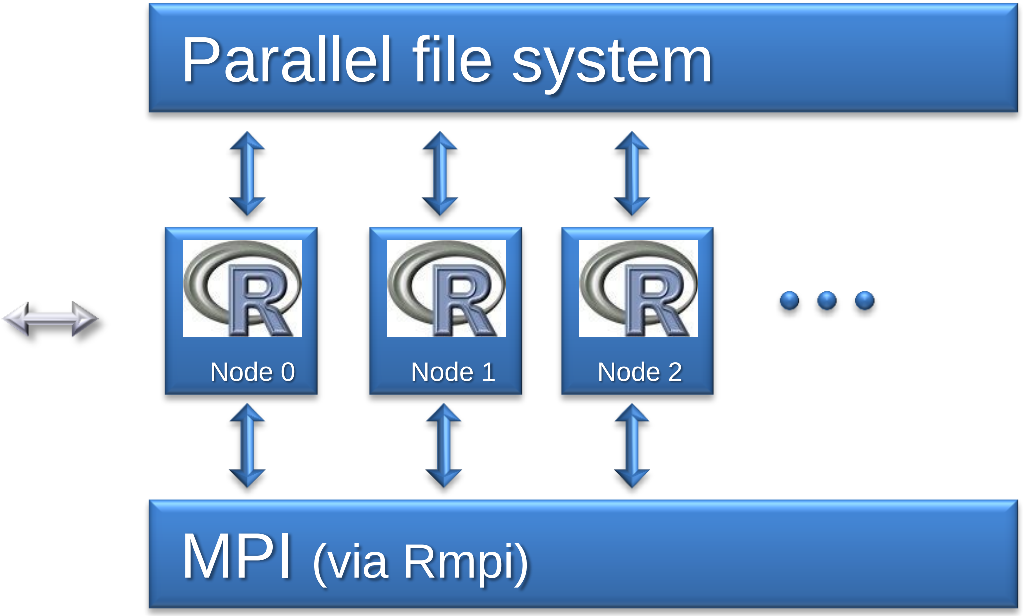
- Track changes in multivariate climate that drive changes in ecology
- Compare ecology impact of climate scenarios

R. Sisneros, J. Huang, G. Ostrouchov, and F. Hoffman (2011). *Procedia Computer Science,* Vol. 4, p1582-1591.

# Enabling R: to run data-paRallel

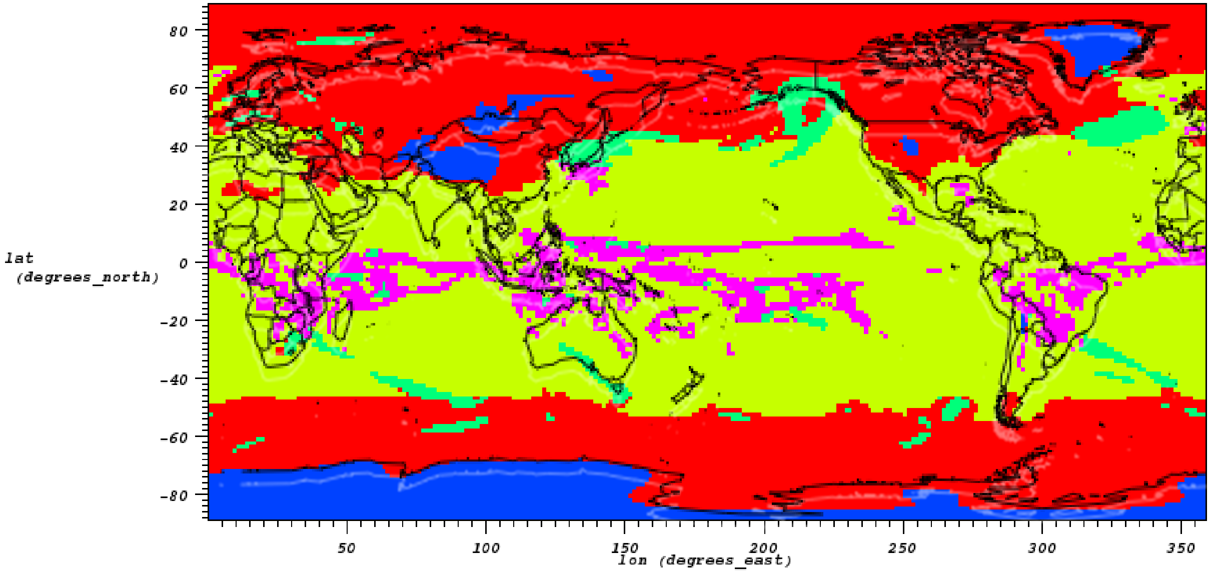


Command-line and graphics interaction with R on Node 0



- **Data readers to bring data from parallel file system**
- **Data-parallel analysis with full capability of R on every node**

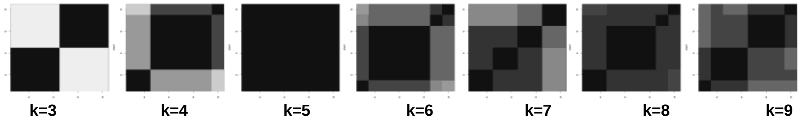
# Clustering in data-parallel R for automated extraction of climate events



**10 years of daily data:**
Lat x Lon x Day for 5 variables:
119,603,200 x 5 matrix (3 GB)

- **Cluster in R without** lat, lon**, and** time **information (semi-supervised)**
- **Play resulting clusters as** lat **by** lon **in** time **with VisIt**
- **Sampling reduces clustering time by order of magnitude (cluster model parameter uncertainty)**
- **Random start agreement selects number of clusters (classification uncertainty)**

**Pairwise agreement “max kappa correlation” of 8 starts**

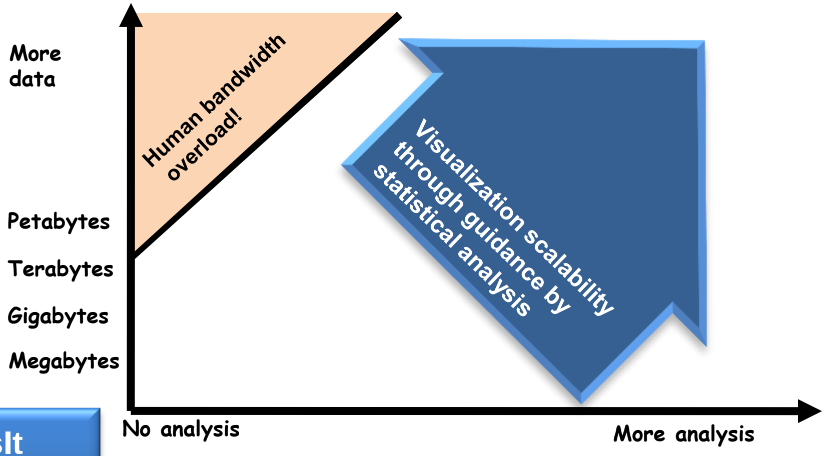


# Building R-VisIt statistical framework for visualization of massive data sets

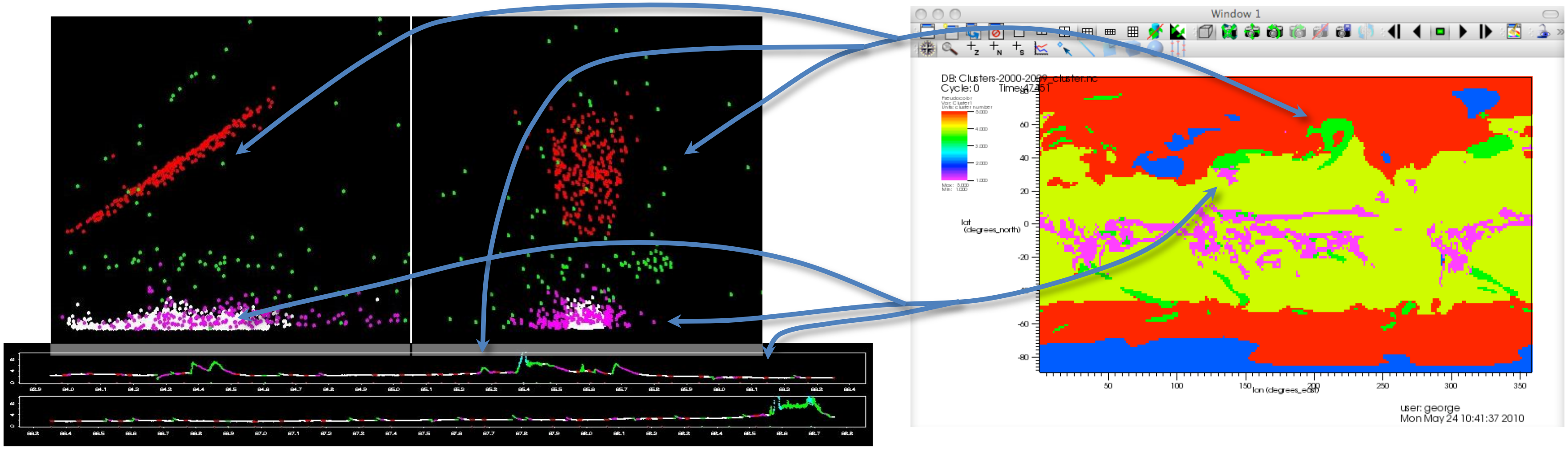
Browse a petabyte? Not humanly possible!

To view a petabyte at 100 MB/s takes 350 8-hour workdays! All year – only 8 weekends off!

Statistical analysis must select views or reduce to quantities of interest in addition to fast rendering


More data
Petabytes
Terabytes
Gigabytes
Megabytes
No analysis
More analysis
Human bandwidth overload!
Visualization scalability through guidance by statistical analysis

- Fit local models to segmented data with R inside VisIt
- R feature analysis in high-dimensional feature space
- Play statistically selected VisIt views of interest



# Common evolutionary steps:
## Experimental science and computational science

- **Mathematical Statistics** harnessed variability to bring rigor and efficiency to experimental science in the 20th century
  - Fusion of theory and data
  - Quantifying bias and uncertainty
  - Design of experiments and analysis of variance

- **Mathematical Statistics** can bring computational science to the rigor and efficiency standards of experimental science in the 21st century
  - Fusion of computational experiment (theory) and data
  - Quantifying bias and uncertainty at computational experiment scale
  - Statistical design of computational experiments
  - Methods to see through, examine, and classify variability in massive data
  - Hardware/software fault analysis and prediction
  - Fault tolerant estimation

# Contact

## George Ostrouchov

Statistics and Data Sciences
Computer Science and Mathematics Division
(865) 574-3137
ostrouchovg@ornl.gov