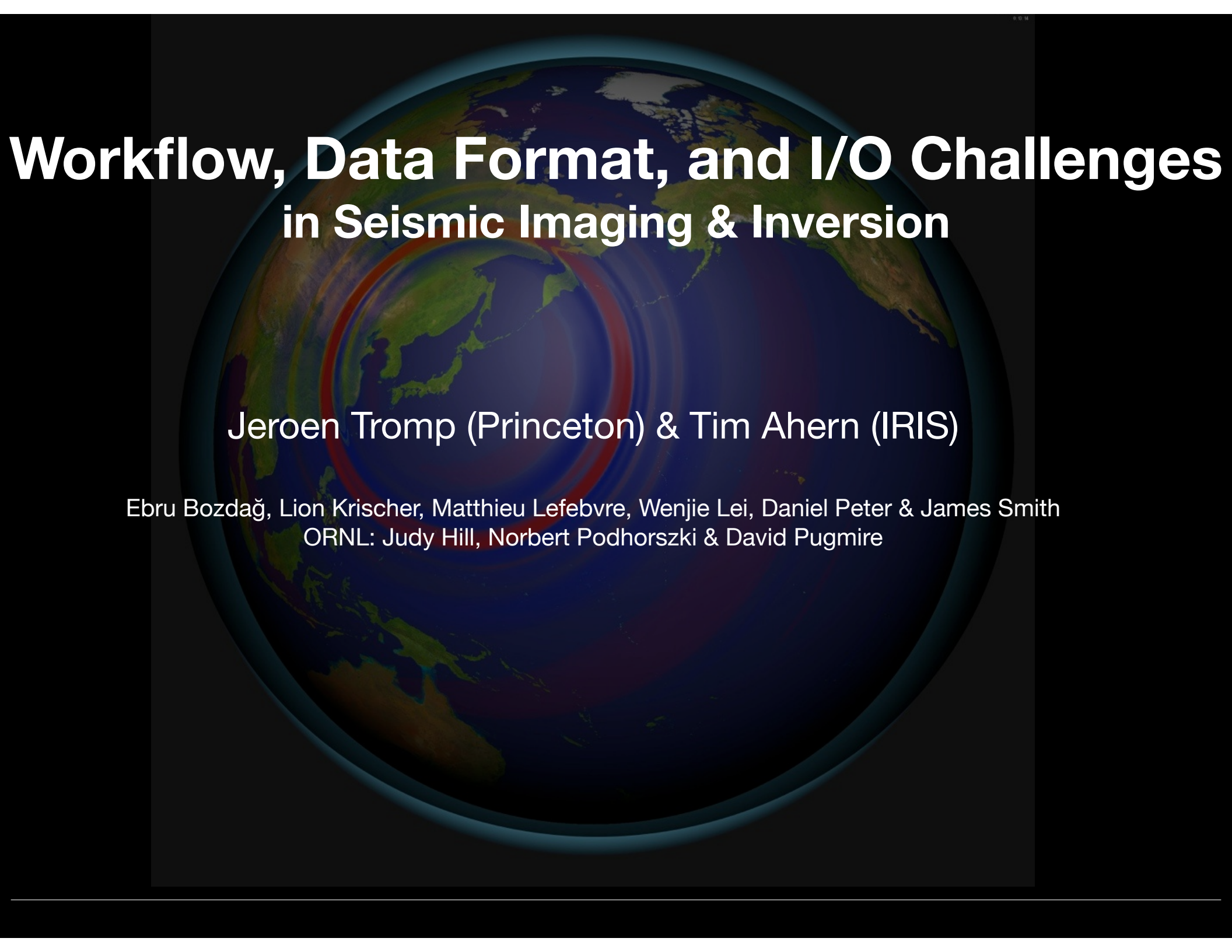


Workflow, Data Format, and I/O Challenges
in Seismic Imaging & Inversion

Jeroen Tromp (Princeton) & Tim Ahern (IRIS)

Ebru Bozdağ, Lion Krischer, Matthieu Lefebvre, Wenjie Lei, Daniel Peter & James Smith
ORNL: Judy Hill, Norbert Podhorszki & David Pugmire

Seismic Imaging & Inversion Challenges

- Cheap, abundant sensors
- Massive amounts of data
  - Industry data sets
  - Regional & global seismology data sets
  - Cross-correlation data sets for seismic interferometry
- On HPC systems, I/O is the bottleneck
- Adopt new data formats for fast parallel I/O (e.g., NetCDF, HDF5 & ADIOS)
- Data culling tools to reduce preprocessing time (e.g., MUSTANG)
- A standard for the exchange of Earth models (e.g., the IRIS NetCDF format)
- Adopt workflow management tools (e.g., Kepler, Pegasus & Swift)
- Tools for data mining, feature extraction, visualization & virtualization (e.g., ParaView, VisIt)

Taming Workflow Issues

Adjoint Tomography Workflow

1. Current data formats are inadequate for fast, parallel I/O

N selected earthquakes

Request observed data

Extract SEED files

Convert to ADIOS

N observed ADIOS data

Run mesher: I ADIOS mesh file

Run N forward simulations

N synthetic ADIOS data

Process data, select windows, make measurements & compute adjoint sources

Pre-processing (embarrassingly parallel)

N ADIOS adjoint source files

Run N adjoint simulations

N ADIOS kernel files

Post-processing (parallel)

Sum kernels: I ADIOS gradient file

Pre-condition & smooth the gradient

Determine step length

Update model: I ADIOS model file

Convergence?

Finish

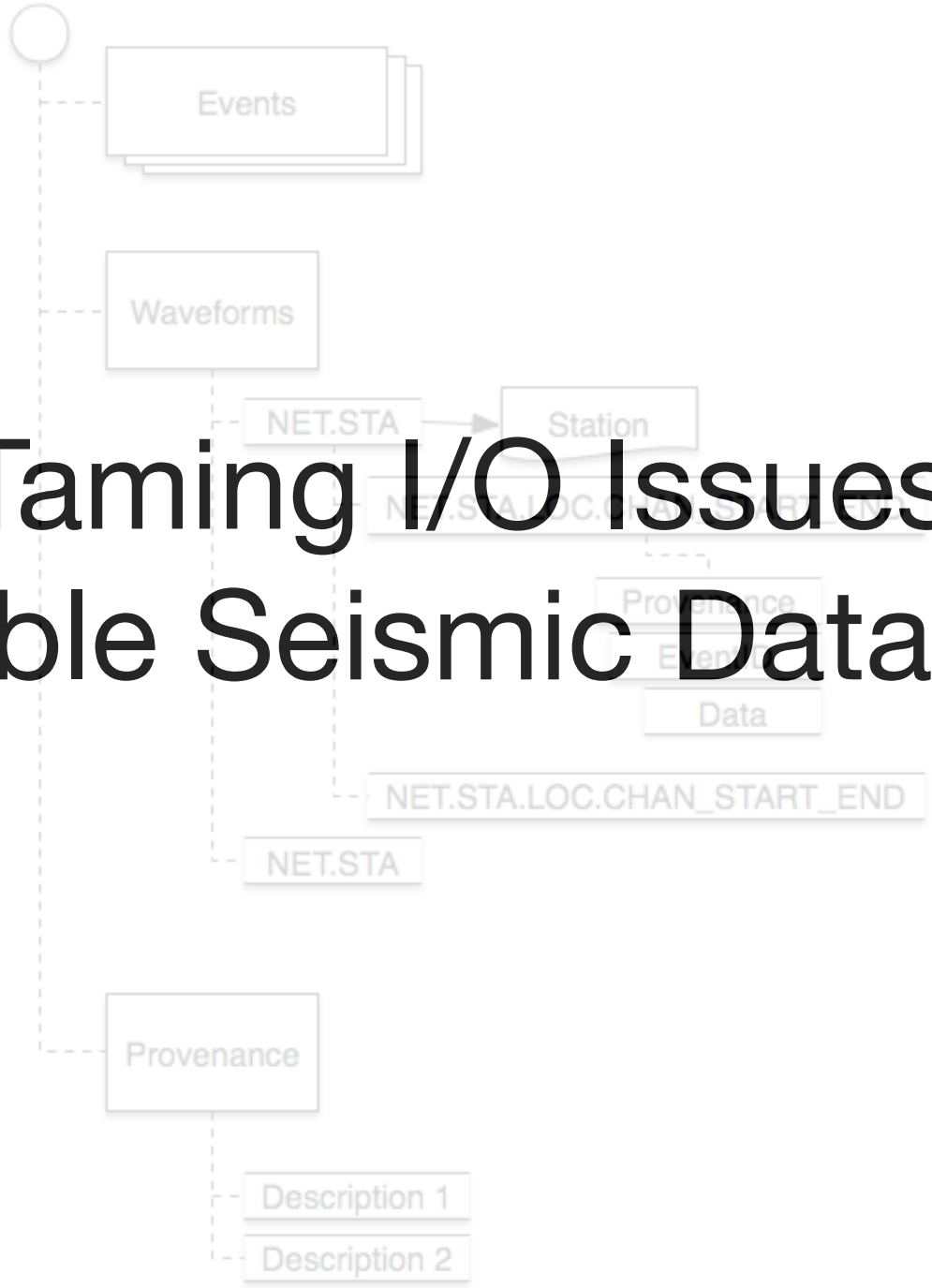
No

Iterate

L-BFGS
Quasi-Newton

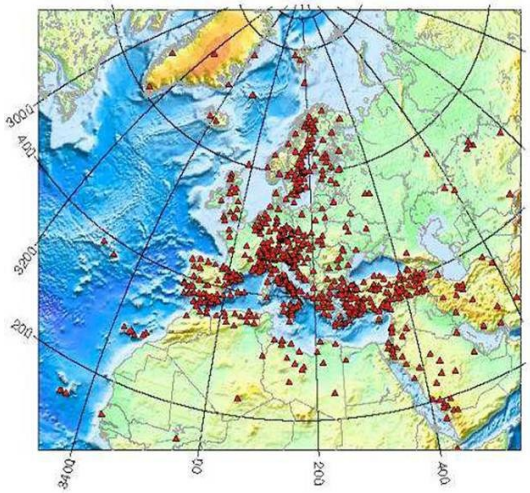
3. Workflow stabilization: Kepler, Swift & Pegasus

2. Visualization of models: ADIOS with VisIt

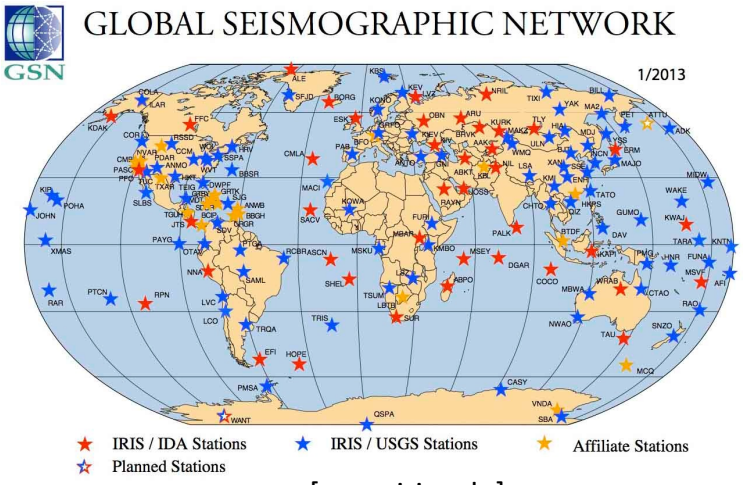


Taming I/O Issues: Adaptable Seismic Data Format

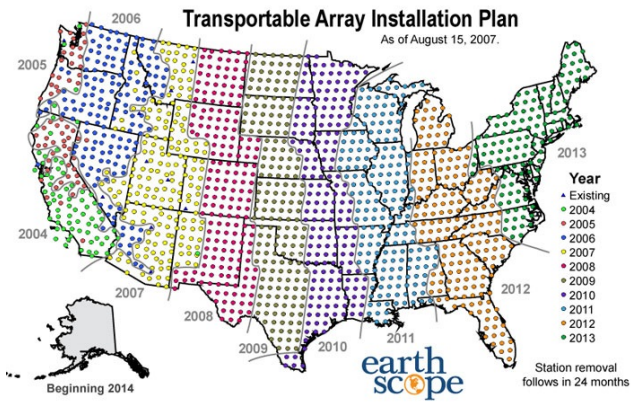
Data in Regional & Global Seismology



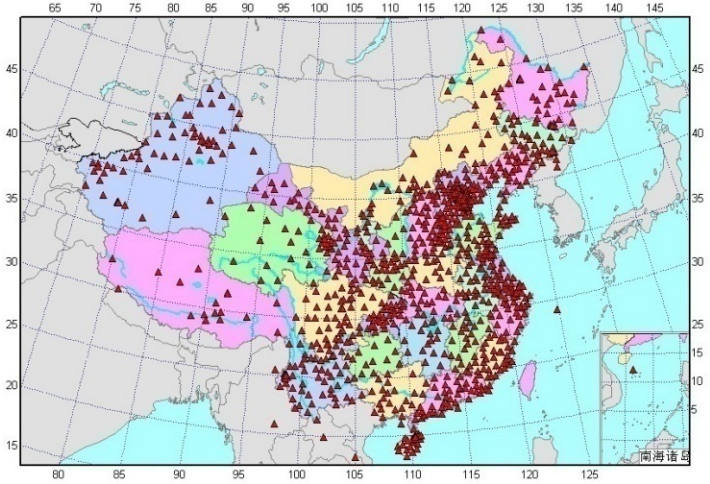
[www.geo.uib.no]



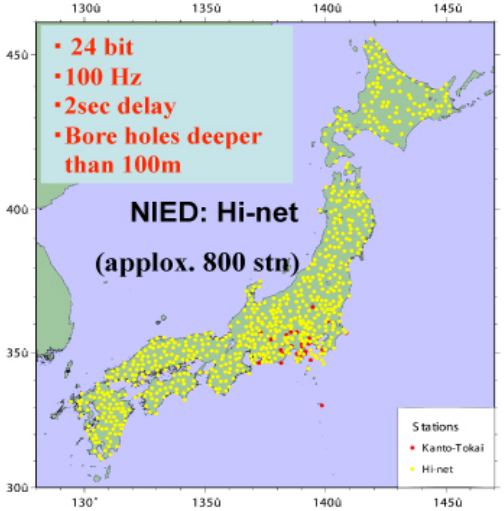
[www.iris.edu]



[web.mst.edu]

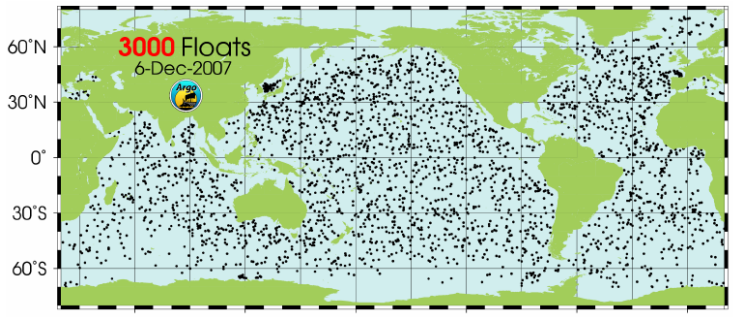


[data.earthquake.cn]



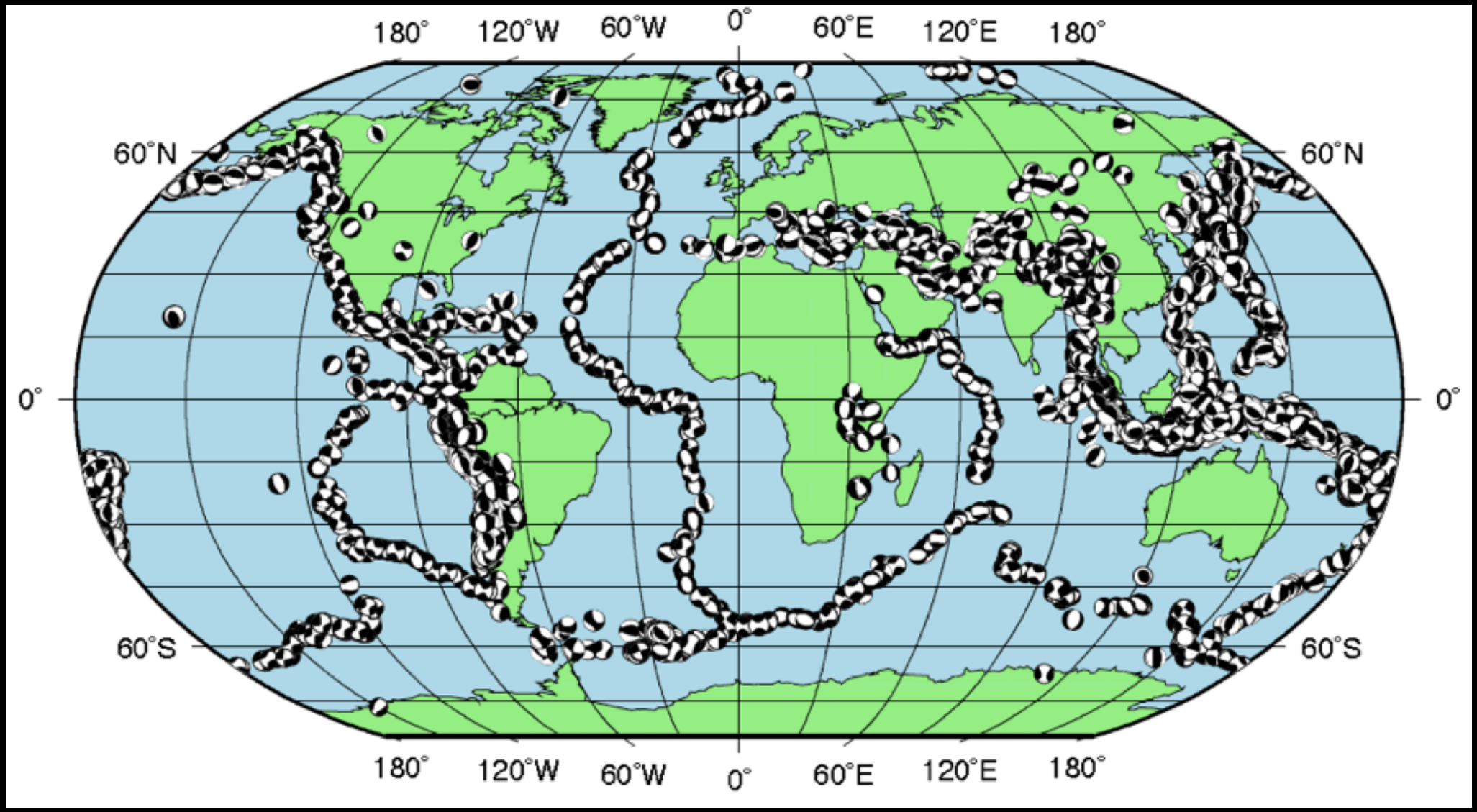
[drh.edm.bosai.go.jp]

MERMAID/MariScope



[Simons et al, 2006]

6,000 5.5 < M < 7 Events in Global Tomography



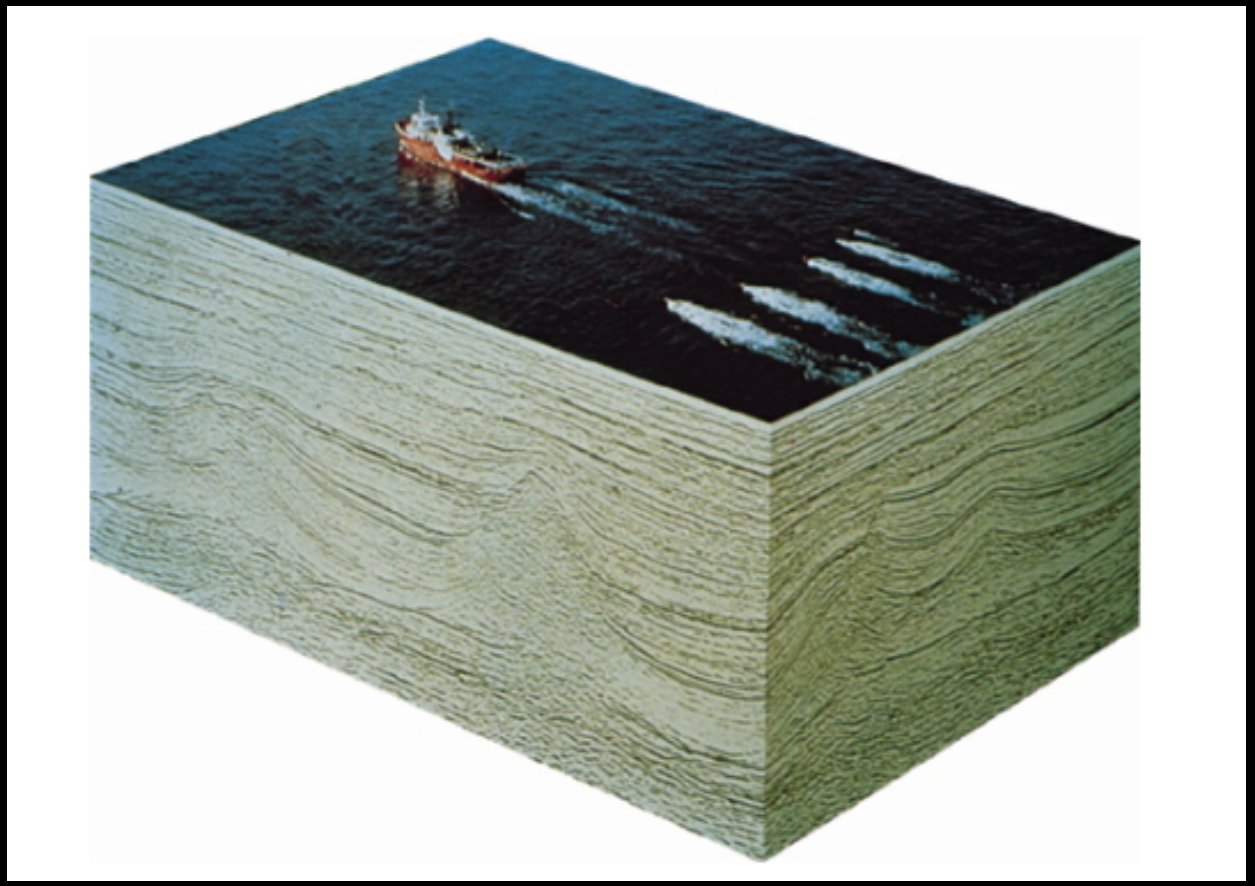
Assimilation of ~100 million data

Ebru Bozdag

Data in Exploration Seismology

3D marine survey can involve 5,000 shots and 50,000 recorders

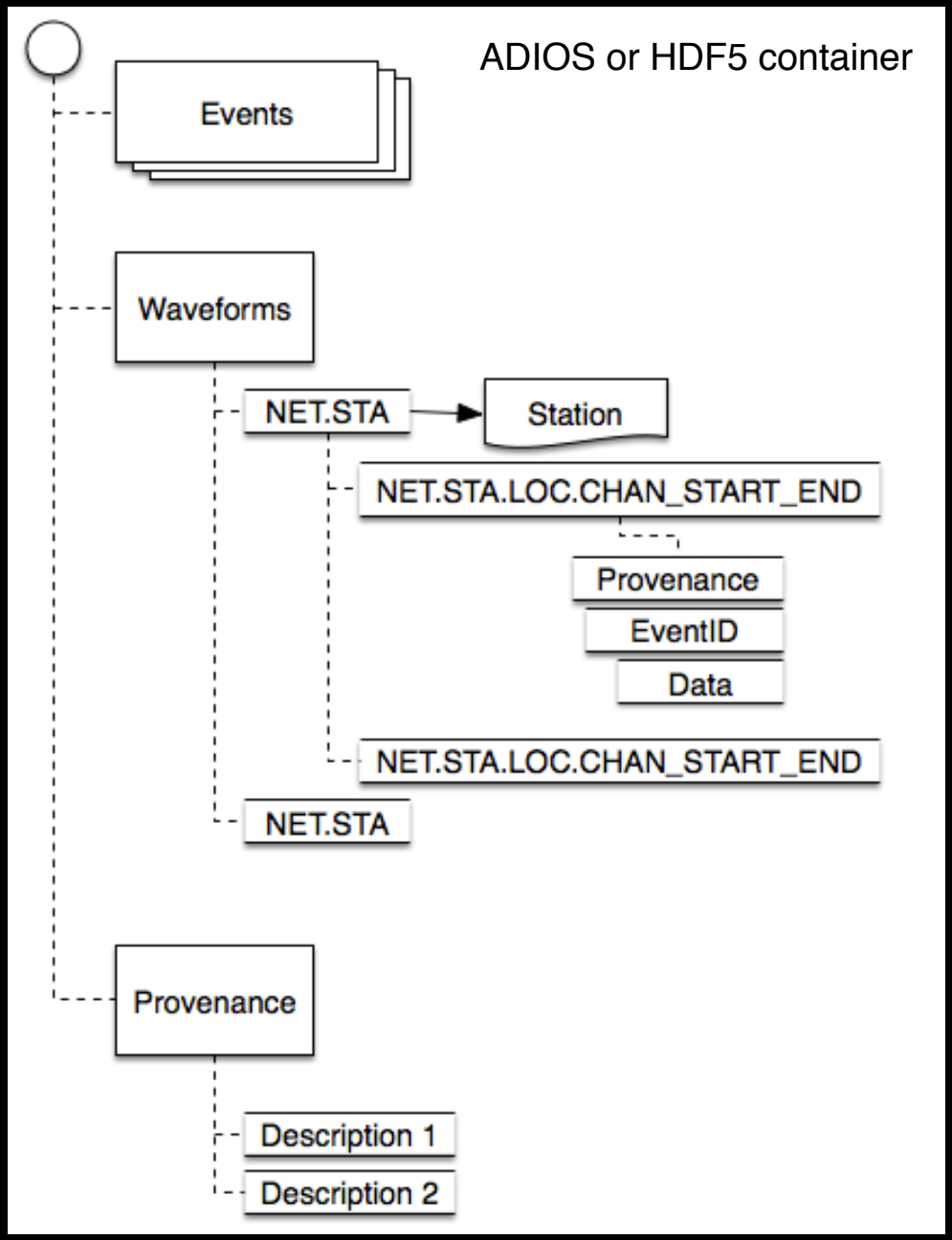
- Petabytes of data
- SEG-Y is the current standard
- Variable SEG-Y file structure
- SEG-Y programs do not always follow specifications



ASDF: an Adaptable Seismic Data Format

- Collaboration involving Princeton University, Munich University (ObsPy) and Oak Ridge National Laboratory
- Increase I/O performance by combining all the time series for a single shot or earthquake into one file
- Take advantage of parallel processing
- Use modern file format as container (e.g., HDF5 or ADIOS)
- Store provenance inside the file for reproducibility
- Use existing standards when possible (e.g., XML)
- Open wiki for development

ASDF Internal Structure

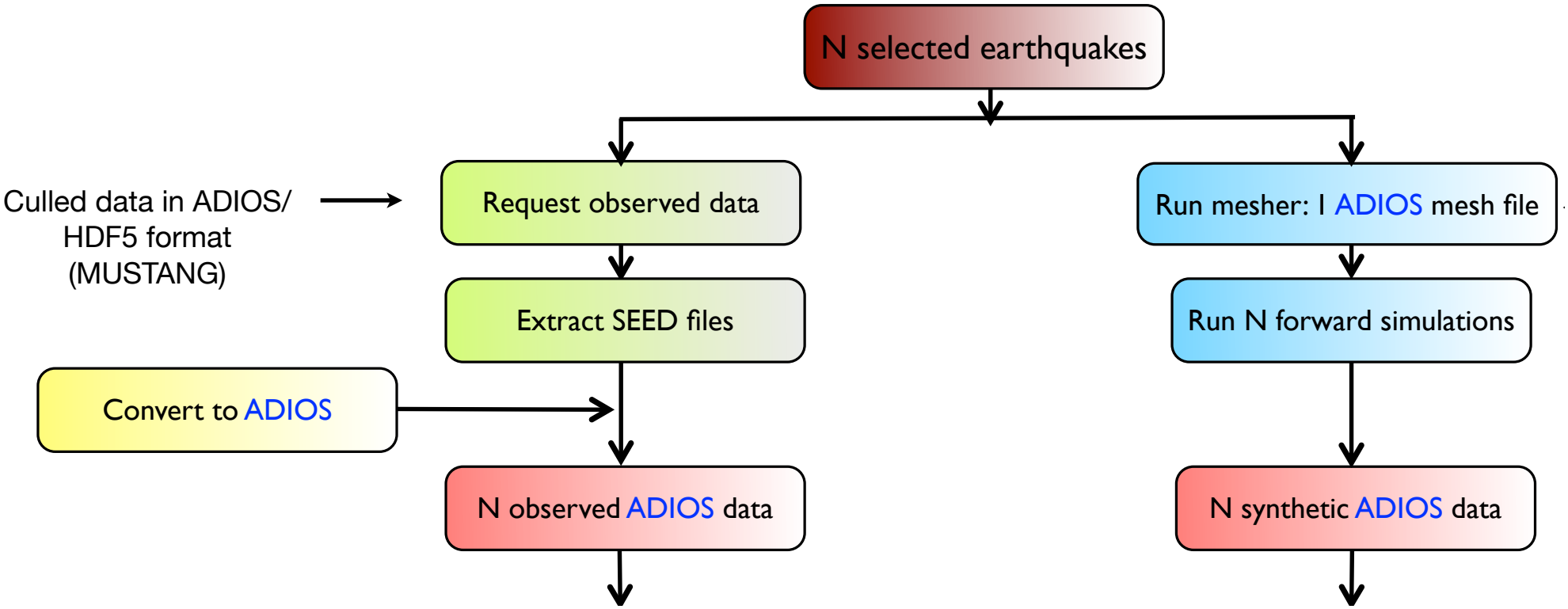


Lion Kirscher

ASDF and XML

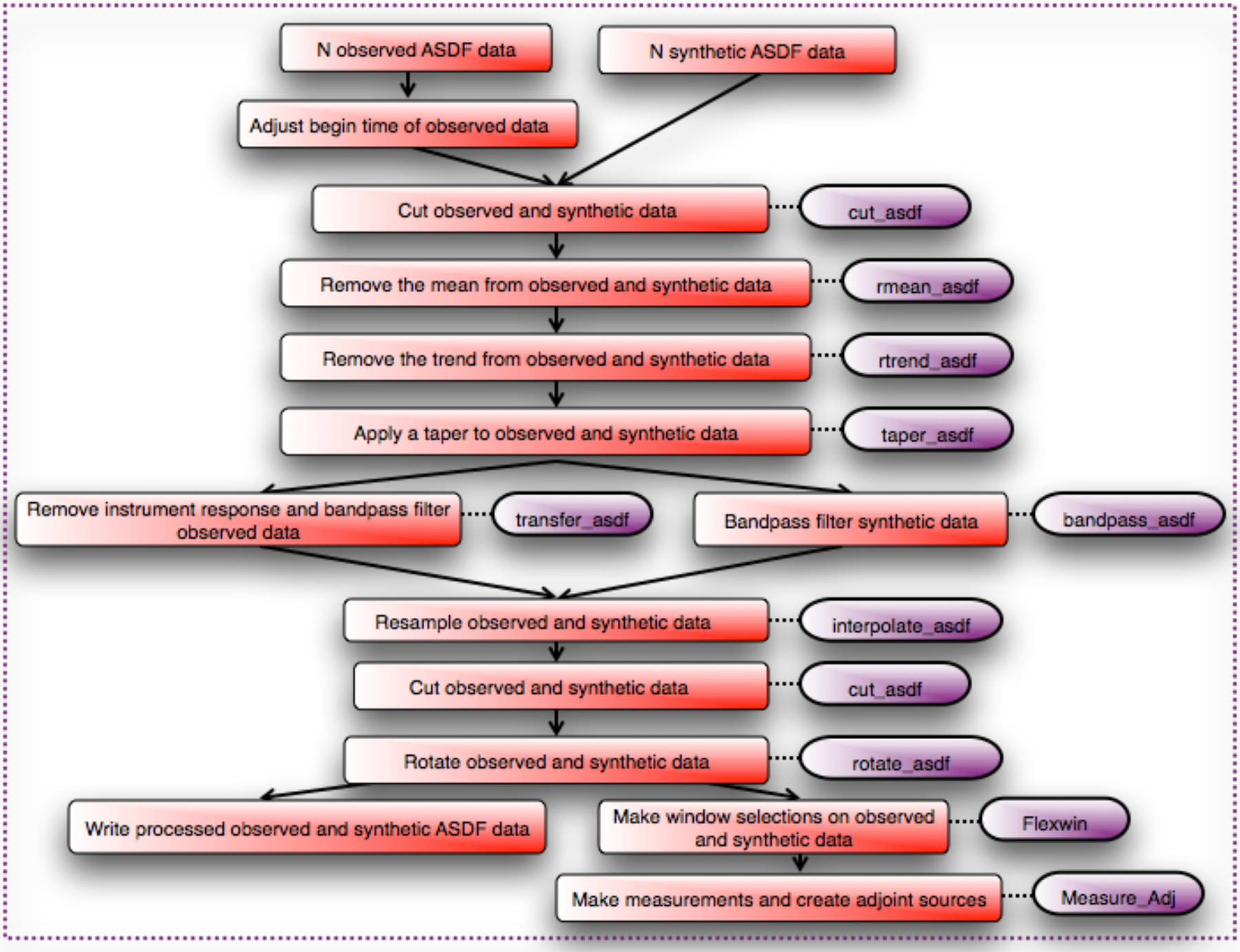
- XML is a flexible, platform-independent standard for defining the information content and structure of a file
- QuakeML is an XML representation of a seismological event which is intended to cover a broad range of fields of application in modern seismology
- StationXML is an XML representation of station information and includes the instrument response
- Provenance can be defined as an XML file where a chain of operations is defined, e.g., time series analysis or parameter settings of a numerical simulation

ASDF in Global Seismology



1000 Stations	Number of SAC Files	Number of ADIOS Files
255 Earthquakes	1,275,00	255
6,000 Earthquakes	30,000,000	6,000

Preprocessing ASDF Workflow

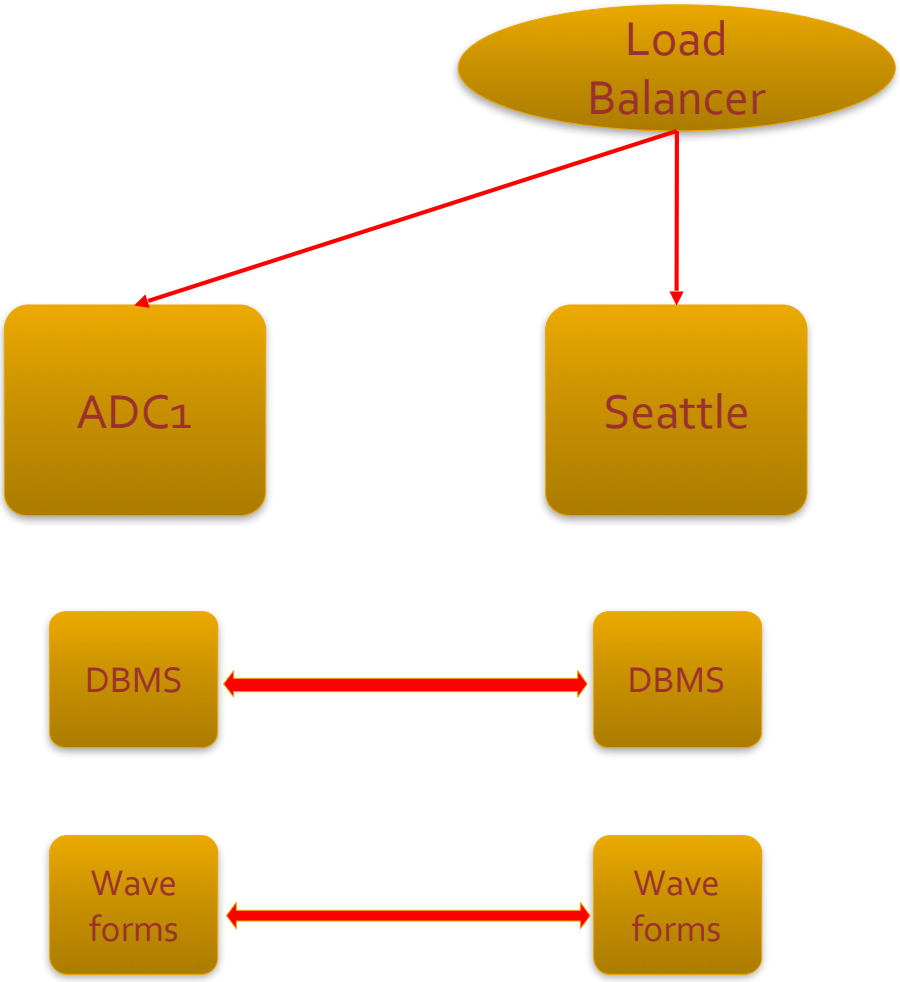


Auxilliary Data Centers to improve System Reliability

- Historically IRIS has operated a primary data center in Seattle, Washington
  - Backup system for redundant copies of data files, database files, software, etc.
  - Primarily for protecting assets in case of a major catastrophe
- IRIS currently operates a second facility in the San Francisco Bay Area near a High Performance Computing installation (LLNL)
- (Cycles Close to Data effort)

Tim Ahern

Multiple & Fully Functioning Data Centers



Ingestion
- BUD Real Time System
- File Ingestion System

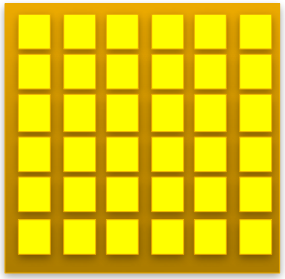
Web Services - Entire suite
- Breqfast
- WILBER3
- MUSTANG
- SeismiQuery
  - Breqfast Requests
- WebRequest

Tim Ahern

Links with High Performance Computing

Research Ready
Formatted for HPC
- ADIOS
- HDF5
- other

Event Products



Scriptable Event Extraction

Services With Research

LLNL

Seattle

DBMS ⟷ DBMS

Wave forms ⟷ Wave forms

Tim Ahern

Conclusions

To tame workflow and I/O issues in seismic imaging & inversion we should explore:

- Partnerships with Industry, National Labs & HPC Centers
  - Petroleum Industry collects, processes and utilizes vast 3D and 4D data sets
  - National Labs are developing tools for fast I/O, workflow management, visualization & virtualization
- Potential collaborations focused on:
  - Data formats for fast parallel I/O (e.g., NetCDF, HDF5 & ADIOS)
  - Standard for the exchange of Earth models
  - Cheap, abundant sensors
  - Full-waveform imaging & inversion
  - HPC workflow management tools (e.g., Kepler, Pegasus & Swift)
  - Data mining, feature extraction, visualization & virtualization (e.g., ParaView, VisIt)