Computation Institute

Social Informatics Data Grid

Cyberinfrastructure for Collaborative Research in the Neural, Social and Behavioral Sciences

Bennett I. Bertenthal
Indiana University
bbertent@indiana.edu

INDIANA UNIVERSITY

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

Infrastructure for Social and Behavioral Sciences

Goal:

Compare, measure and search for patterns in structured, semi-structured, and heterogeneous data sets.

Challenge:

Integrate information over time, place, and types of data

Needs:

(1) Data interface (shared datasets & databases)

(2) Service interface (shared tools for analysis)

(3) Intellectual interface (shared problems & theories)

An Example: What are S1 and S2 talking about?

S1 you know like those ¿fireworks?

S2 well if we're trying to drive'em / out her<r>e #
we need to put'em up her<r>e

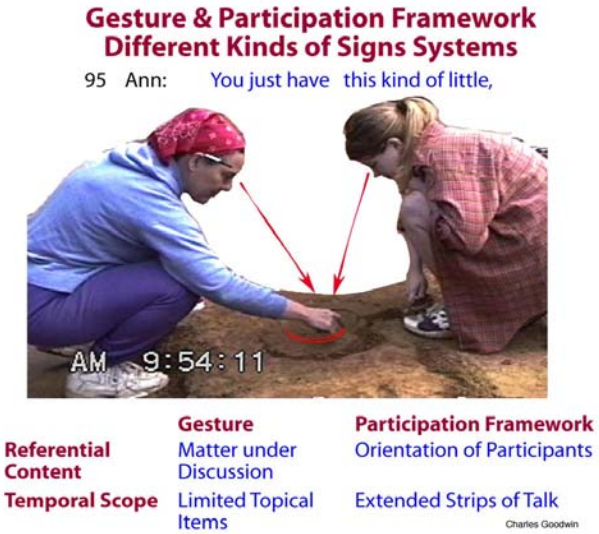
S1 yeah well what I'm saying is we should*

S2 in front

S1 we should do it*
we should make it a lin<n>e
through the room<m>s /
so that they explode
like here then here then here then here

Embodied Cognitive and Social Behavior

- **Social communication is not organized by speech alone**
- **Multiple forms of embodiment – speech, prosody, gesture, eye gaze, posture, facial expression**



- Opportunity: **Evaluate actions by which we construct in concert with each other the cognitive, social and cultural world**
- Challenge: **Lack tools to assess multiple measures at multiple levels simultaneously and to store and analyze these measures in a common database**

Why SIDGrid?

- Data-sharing
  - Understanding complex behavior requires sharing across labs and across disciplines
  - Sciences need an established empirical base
  - Sharing maximizes coverage
- Protocol, coding standardization
- Tool creation
- Multimedia window is crucial
  - Multiple time scales
  - Embodied cognition and linkage
  - Social embedding and linkage

Testbed for Research

Multiple Measures at Multiple Time Scales

- Multimodal Communication in Humans and Machines
- Cognitive and Social Neuroscience
- Neurobiology of Social Behavior in Humans and Animals

Primary Objectives

- Develop prototype of core facility for collecting multiple measures of time-synchronized data
- Develop integrated tools for storage, retrieval, annotation, and analyses of multiple data sets at different time scales
- Develop scripts for parallelizing code to run on grid clusters

SuperLab

Objectives of SuperLab

- To expand behavioral and biobehavioral models by **simultaneous** monitoring of physiological, psychological, and behavioral parameters
- To apply grid computing technologies to study multiple response system in behavioral and biobehavioral research
- To exploit the research potential of grid computing by facilitating “remote” and “virtual” collaborations
- To interface data collection and analysis with the SIDGrid

Sensor and Stimulus Systems

- Stimulus Display
- Audio Stimuli
- Video Recording
- Audio Recording
- Eye Tracking
- Motion Analysis
- IR Thermography
- Physiological Recording

- **Scene**
- **Face**
- **Full body**

- **Ambient array**
- **Personal boom mic**

- Multi-projector wall
- Transparent display
- High density display

- Audio feedback
- Spatialization

- **electrocardiogram**
- **electromyogram**
- **galvanic skin response**
- **respiration**
- **blood pressure**

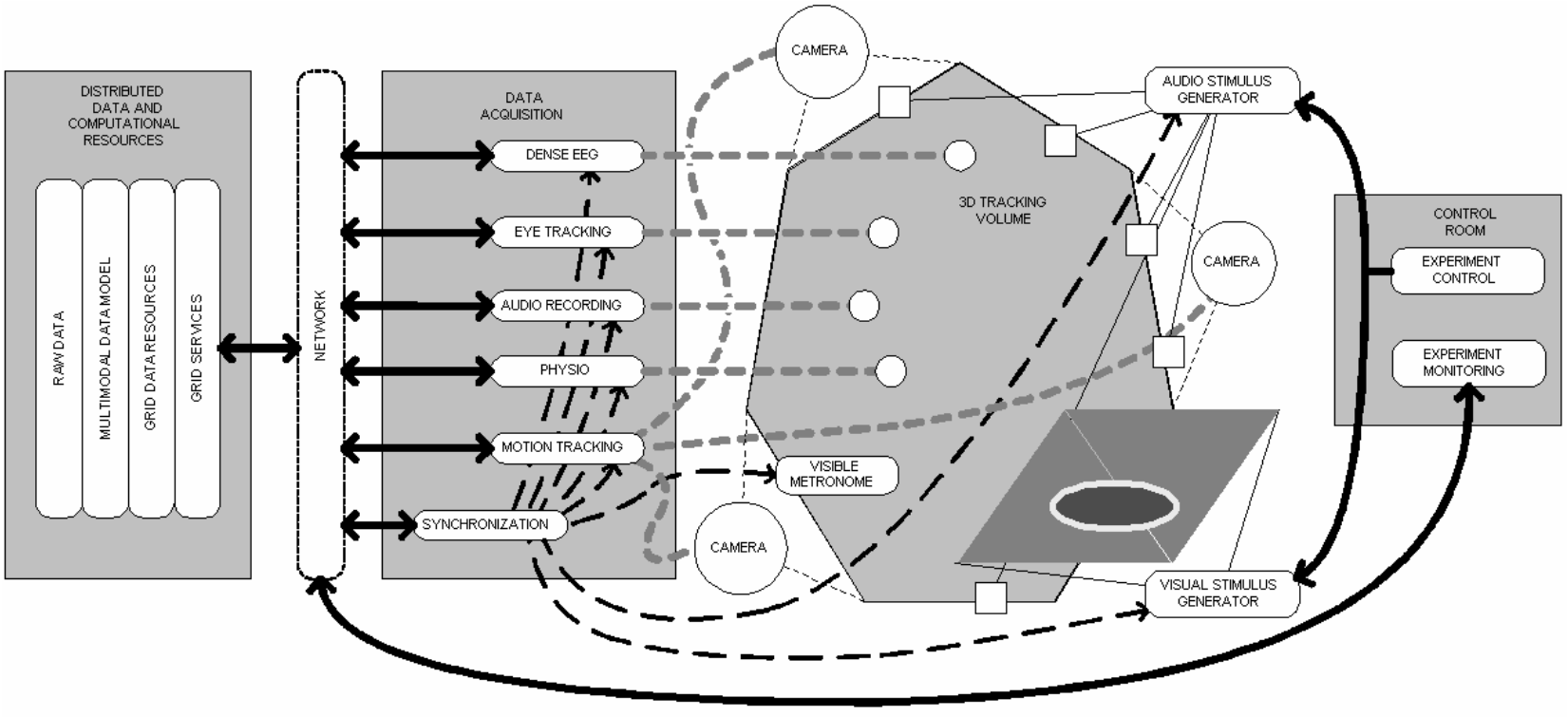
- **High density electroencephalogram**

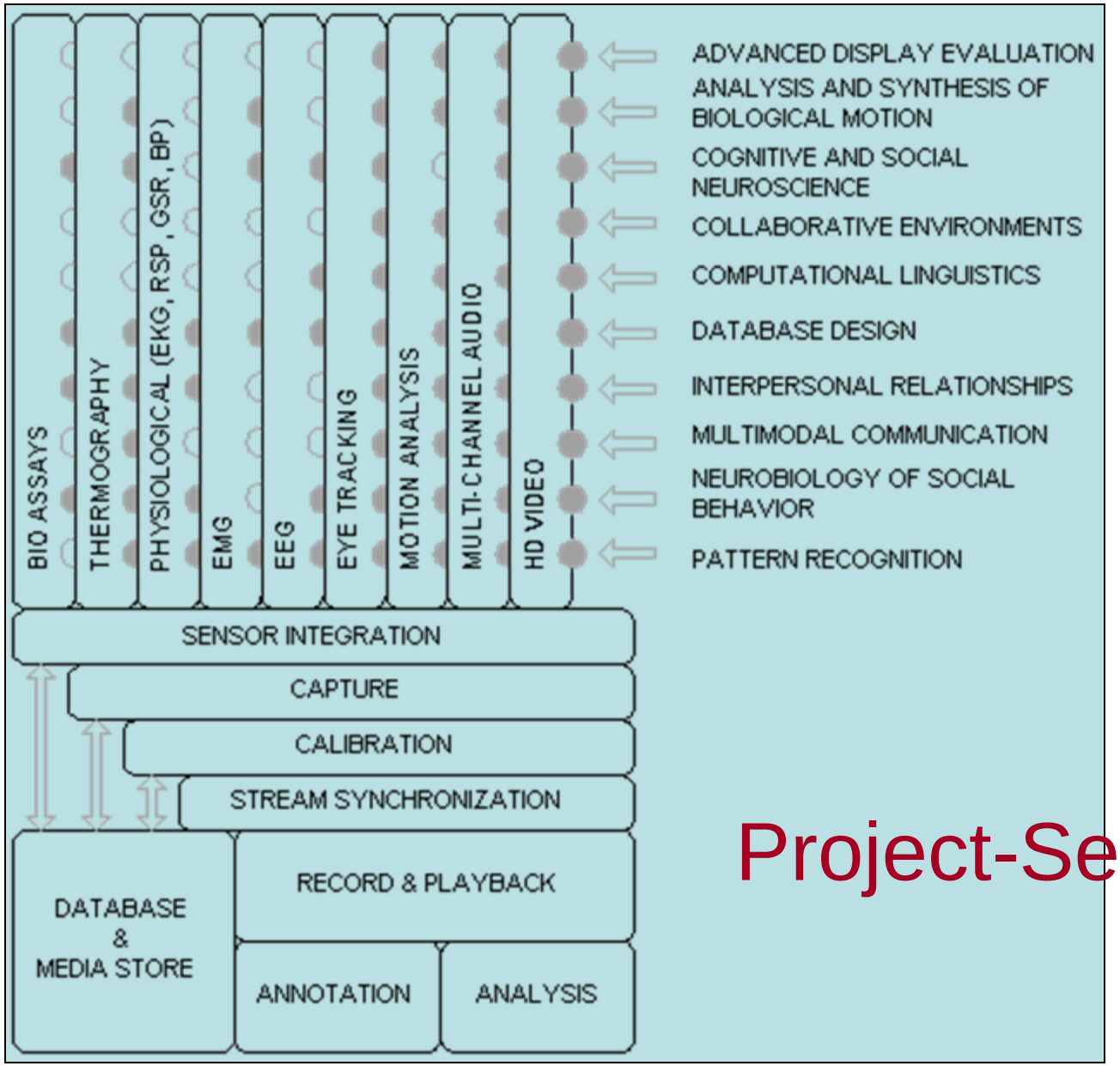
- **Head mounted**
- **"Remote optics"**

- **Visualeyez**
  - **512 active markers**
- **Mid-infrared 320x240 NTSC**


LifeShirt

SuperLab


DISTRIBUTED DATA AND COMPUTATIONAL RESOURCES
RAW DATA
MULTIMODAL DATA MODEL
GRID DATA RESOURCES
GRID SERVICES
NETWORK
DATA ACQUISITION
DENSE EEG
EYE TRACKING
AUDIO RECORDING
PHYSIO
MOTION TRACKING
SYNCHRONIZATION
CAMERA
3D TRACKING VOLUME
VISIBLE METRONOME
CAMERA
AUDIO STIMULUS GENERATOR
CAMERA
VISUAL STIMULUS GENERATOR
CONTROL ROOM
EXPERIMENT CONTROL
EXPERIMENT MONITORING

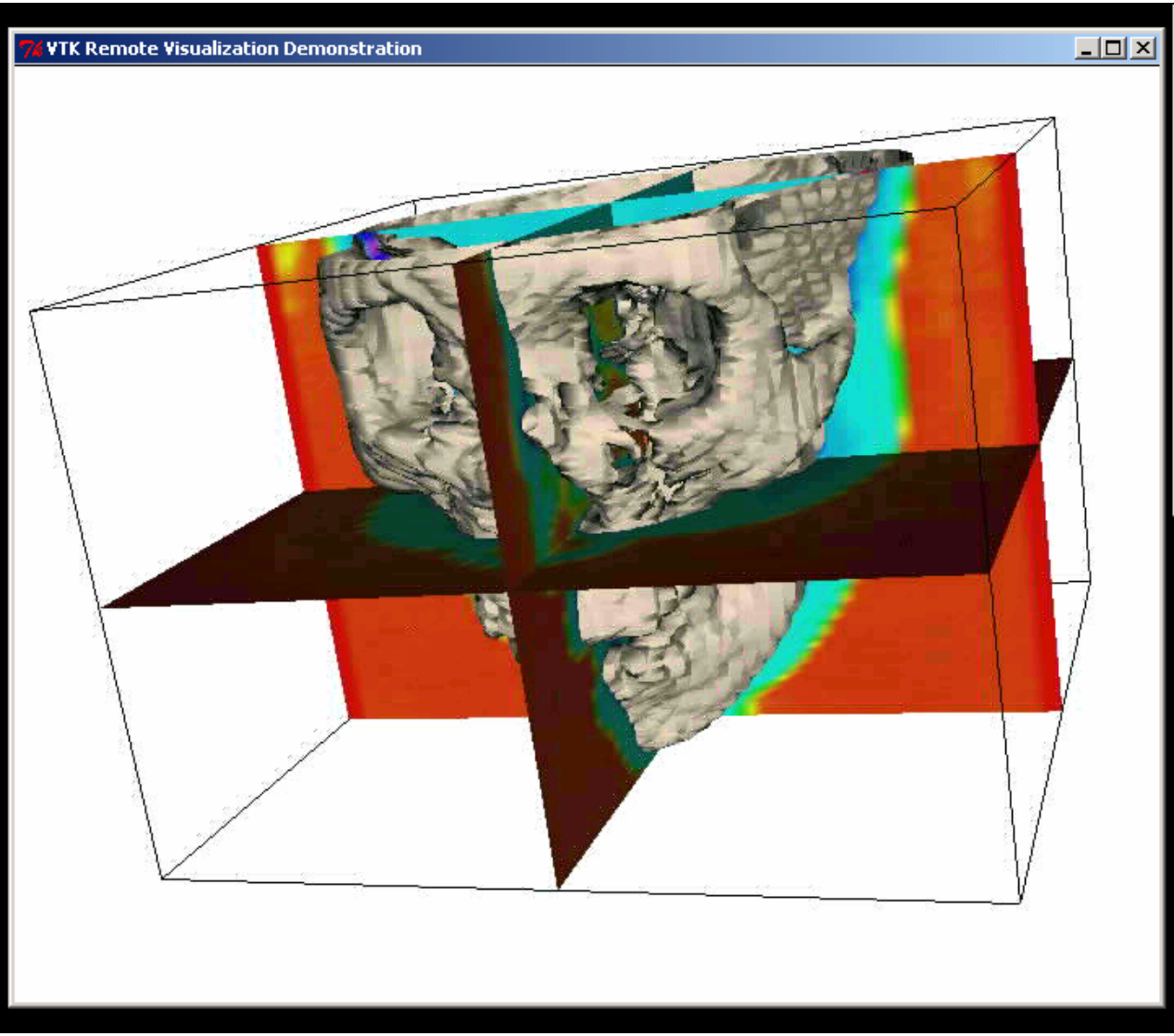


Project-Sensor Matrix

Features of the SuperLab

1. Capable of simultaneous and dynamic monitoring of several channels of behavioral (e.g.,audio, video) and physiological (e.g., autonomic activity, EMG, EEG, eye gaze, facial thermography) variables.
2. Synchronizing stimuli and with multimodal response systems.
3. The ability to monitor variables on at least two individuals simultaneously.
4. The ability for collaborators to contribute, observe, and collect data from remote sites connected to the SIDGrid.

Access Grid


VTK Remote Visualization Demonstration


ANL/FL Workshop MAIN

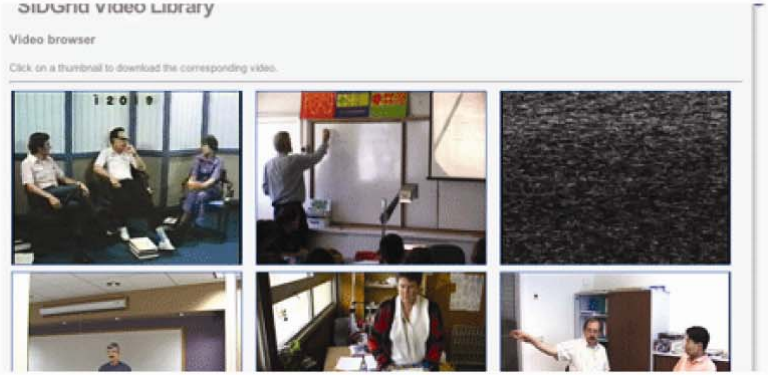

ANL/FL Workshop AUDIENCE

SIDGrid

ABOUT THE SIDGRid PROJECT

SIDGrid is a computing infrastructure that provides integrated computational resources for the processing of multimodal data in the social sciences.

The Social Informatics Data (SID) Grid enables researchers to collect real-time multimodal behavior at multiple time scales. Multimedia data (voice, video, images, text, numerical) is stored in a distributed data warehouse that employs Web and Grid services to support data storage, access, exploration, annotation, integration, analysis, and mining of individual and combined data sets. The SIDGrid includes the development of a user interface that supports access to the data regardless of the user's geographic location and can be coupled with existing Grid technology.



User Login

Username: *

Password: *

Log in

Create new account Request new password

USING SIDGrid

> Go to the getting started page to find out more about using SIDGrid and how to obtain a login account for the SIDGrid portal.

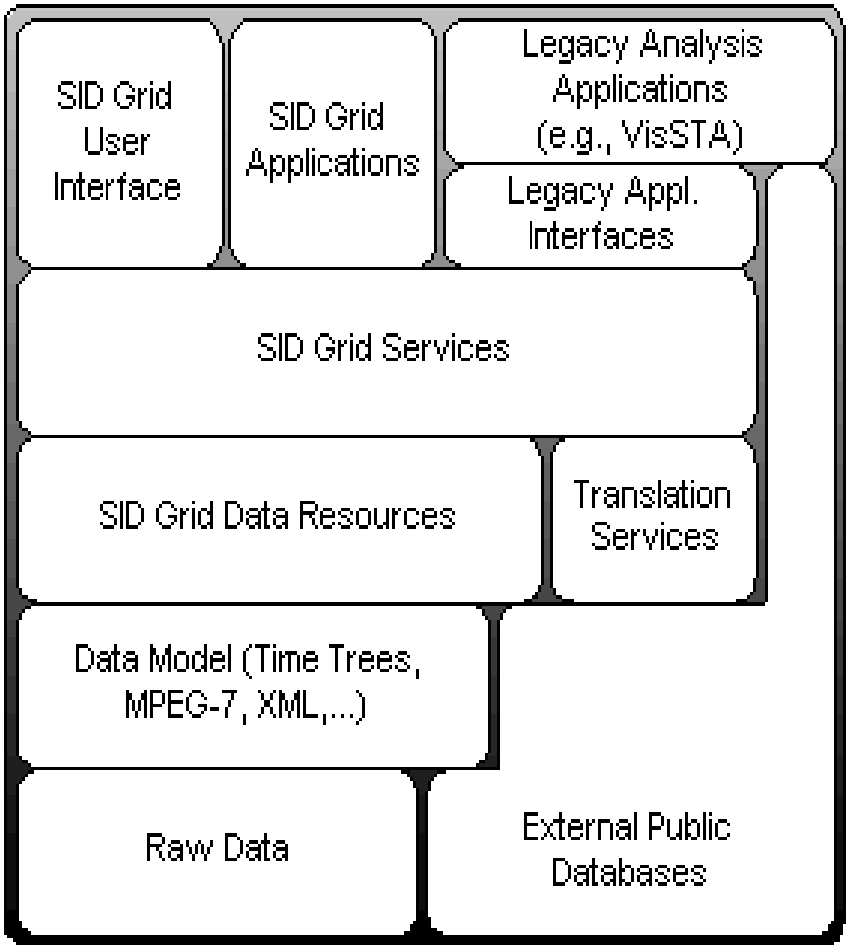
ABOUT THE TERAGRID

TeraGrid is an open scientific discovery infrastructure combining leadership class resources at nine partner sites to create an integrated, persistent computational resource.

[TeraGrid website][State of TeraGrid pdf]

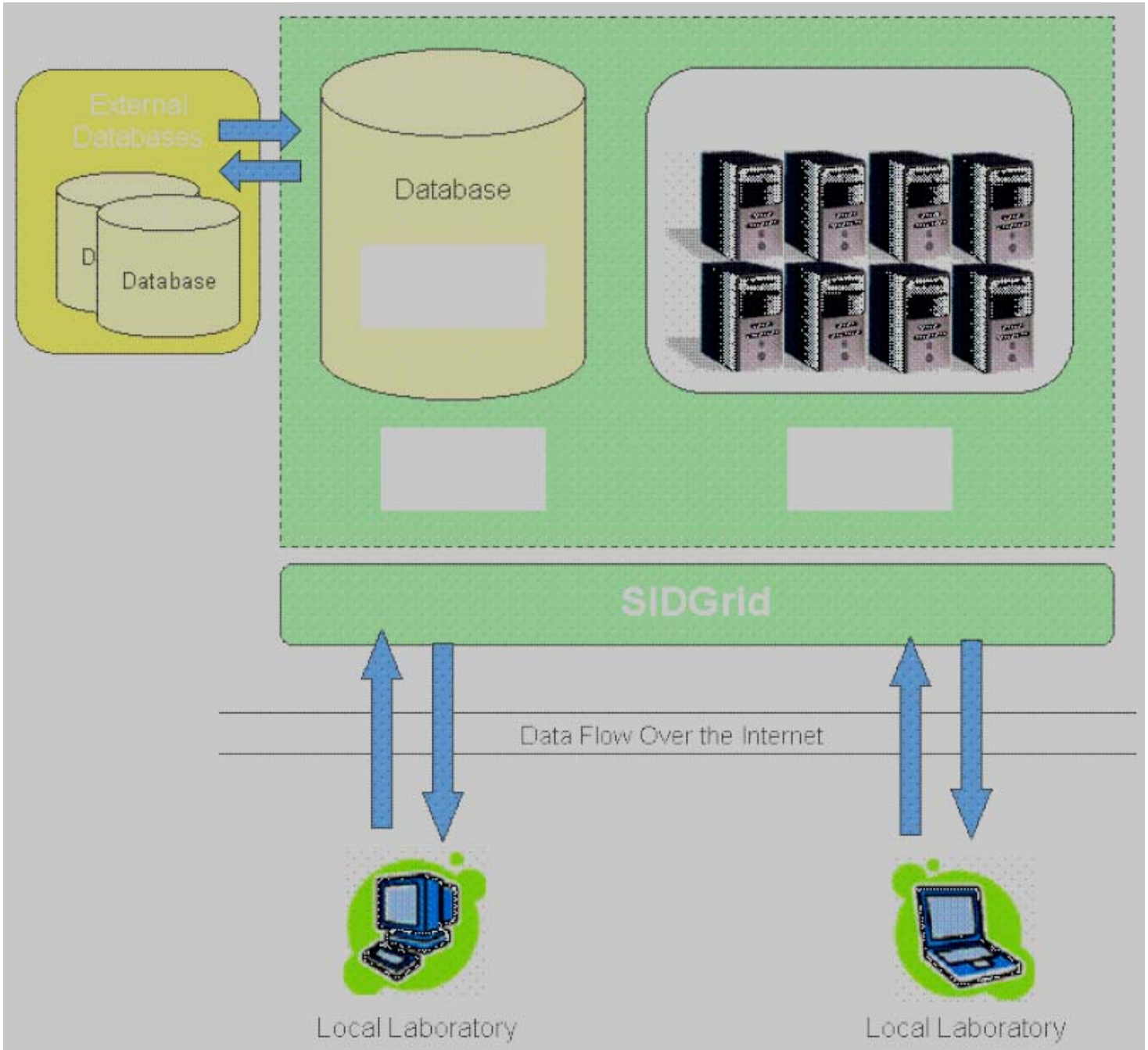
SITE ADMINISTRATION

Social Informatics Data Grid



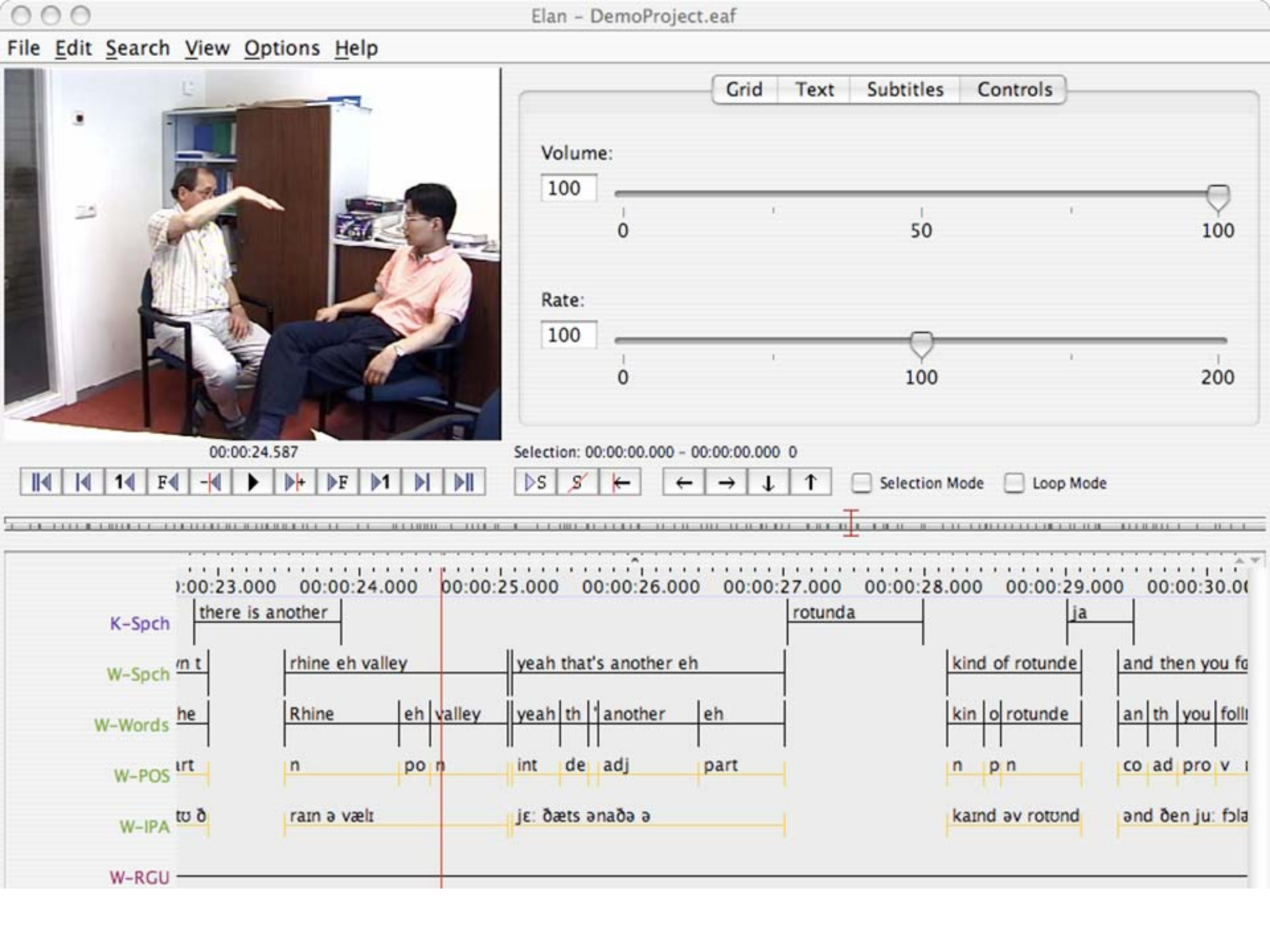
- **A *general purpose* architecture for streaming data applications (e.g., video, audio, time series)**
- **Built on well established database, multimedia and web and grid services standards**
- **Time alignment in distributed heterogeneous datasets**
  - Software and hardware based
  - Integrated with existing laboratory time stamping and registration techniques
- **Scalable**
  - Number of datasets
  - Types of data
  - Multiple end user applications

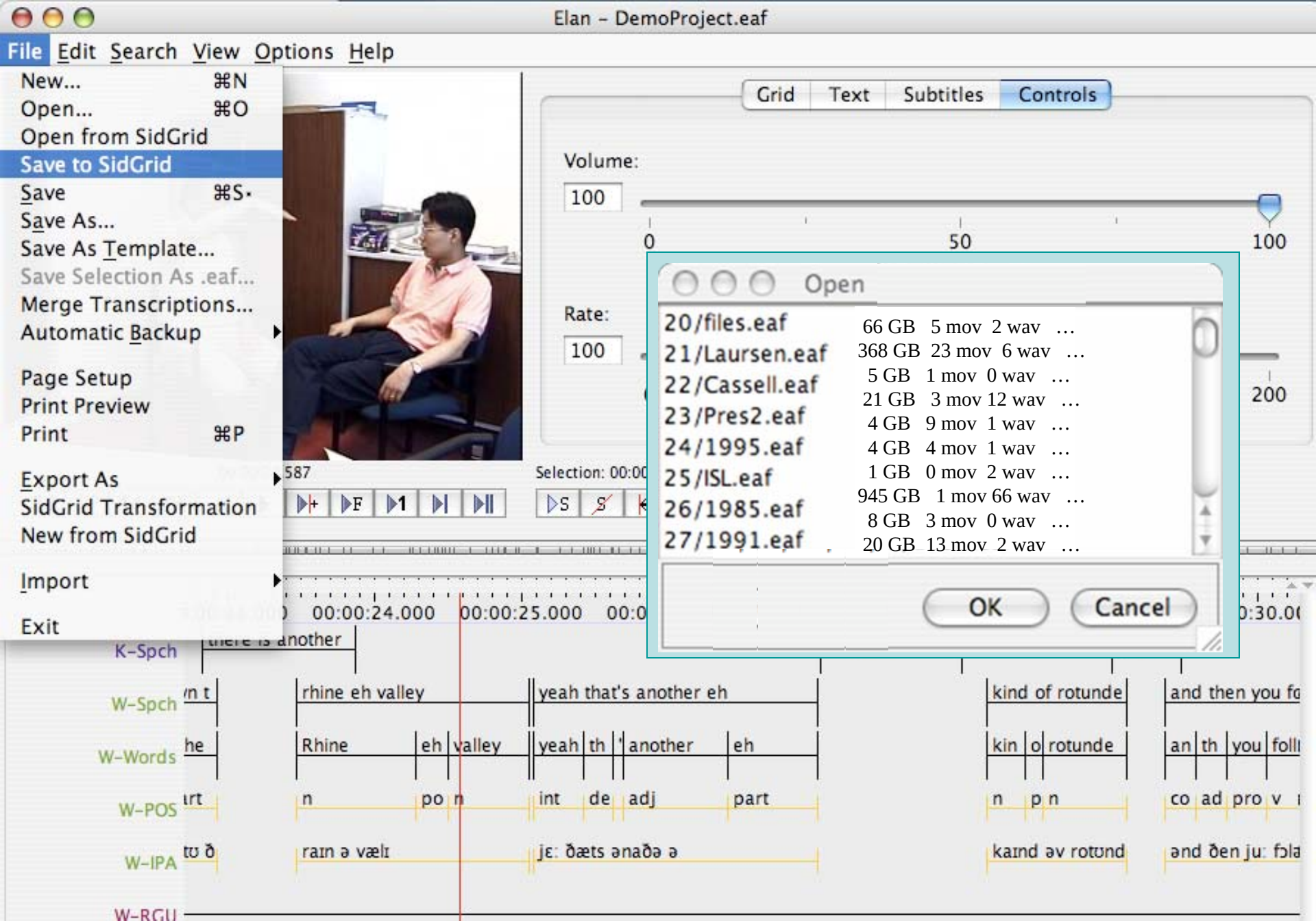
Server



Client

Client Side


Elan - DemoProject.eaf
File Edit Search View Options Help
Grid Text Subtitles Controls
Volume:
100
0 50 100
Rate:
100
0 100 200
00:00:24.587
Selection: 00:00:00.000 - 00:00:00.000 0
Selection Mode
Loop Mode
00:00:23.000 00:00:24.000 00:00:25.000 00:00:26.000 00:00:27.000 00:00:28.000 00:00:29.000 00:00:30.0
K-Spch
there is another
rotunda
ja
W-Spch
rhine eh valley
yeah that's another eh
kind of rotunde
and then you fo
W-Words
Rhine
eh
valley
yeah
th
another
eh
kin
o
rotunde
an
th
you
W-POS
n
po
n
int
de
adj
part
n
p
n
co
ad
pro
v
W-IPA
raɪn ə vælɪ
jɛ: ðæts ənaðə ə
kaɪnd əv rotʊnd
ənd ðen ju: fɔlə
W-RGU


Elan – DemoProject.eaf
File Edit Search View Options Help
New... ⌘N
Open... ⌘O
Open from SidGrid
Save to SidGrid
Save ⌘S
Save As...
Save As Template...
Save Selection As .eaf...
Merge Transcriptions...
Automatic Backup
Page Setup
Print Preview
Print ⌘P
Export As
SidGrid Transformation
New from SidGrid
Import
Exit
Grid Text Subtitles Controls
Volume:
100
0 50 100
Rate:
100
200
Open
20/files.eaf 66 GB 5 mov 2 wav …
21/Laursen.eaf 368 GB 23 mov 6 wav …
22/Cassell.eaf 5 GB 1 mov 0 wav …
23/Pres2.eaf 21 GB 3 mov 12 wav …
24/1995.eaf 4 GB 4 mov 1 wav …
25/ISL.eaf 1 GB 0 mov 2 wav …
26/1985.eaf 945 GB 1 mov 66 wav …
27/1991.eaf 8 GB 3 mov 0 wav …
OK
Cancel
Selection: 00:00
00:00:24.000
00:00:25.000
K-Spch
W-Spch
W-Words
W-POS
W-IPA
W-RGU
there is another
rhine eh valley
yeah that's another eh
kind of rotunde
and then you fo
Rhine
eh
valley
yeah
th
another
eh
kin
o
rotunde
an
th
you
n
po
n
int
de
adj
part
n
p
n
co
ad
pro
v
raɪn ə vælɪ
jɛ: ðæts ənaðə ə
kaɪnd əv rotʊnd
ənd ðen ju: fɔlə

Client Side

- Leveraging efforts for annotation and analysis of multimodal data
  - Familiarity and Interoperability
    - Elan (Max Planck Institute for Psycholinguistics, The Netherlands)
    - Talkbank (Carnegie Mellon University, US)
    - Digital Replay System (Nottingham University, UK)
  - XML, Java
  - Cross platform
- Adding SIDGrid functionality to Elan
  - Minimally intrusive
    - Avoid complicated co-development w/ELAN team
  - Browsing SIDGrid data
  - Additional data types
  - Upload / Download to SIDGrid server

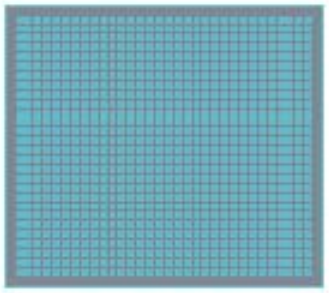
Current Status of CHILDES and TalkBank

- 5000 members of CHILDES, 1200 of TalkBank
- Over 2000 articles based on use of CHILDES, 100 based on TalkBank
- 3 million utterances in CHILDES
- 2 million utterances in TalkBank
- TalkBank consists mostly of linked media; CHILDES only has some (mostly transcriptions)

TalkBank Groups, Areas, Topics

- Child Language (CHILDES, PhonBank)
- Conversation Analysis (MOVIN, CA)
- SLA, Bilingualism, LIDES
- Legal (Supreme Court)
- AphasiaBank
- Classroom Discourse
- Linguistic Exploration
- Sociolinguistics (SLX)

Server Side



contact us

SIDGrid

home about people news & events partners portal log out

Welcome hereld

projects videos administrate

Search for Keywords like Search

my projects

all projects

GROUPS:

talkbank

ALL PROJECTS

NEXT ▶

	.mov	.wav	.eaf	GB	
files	10	0	0	45	2005-11-30 17:13:35
Cassell	4	30	0	20	2005-11-30 17:13:35
Pres2	2	2	1	3	2005-11-30 17:13:35
1995	12	100	9	200	2005-11-30 17:13:35
ISL	1	1	1	1	2005-11-30 17:13:35
1985	6	2	0	12	2005-11-30 17:13:35
1991	400	0	1	1001	2005-11-30 17:13:35
1989	0	666	1	312	2005-11-30 17:13:35
1992	0	0	13	0.1	2005-11-30 17:13:35
1986	0	0	0	0.0	2005-11-30 17:13:35
CLIPS	18	4	0	66	2005-11-30 17:13:35

Search for Keywords like Search

CLICK ON A THUMBNAIL TO DOWNLOAD THE CORRESPONDING VIDEO

NEXT ▶


group5.mov


2010CZ304.mov


dec19f.mov


7004JP405.mov


8-5.mov

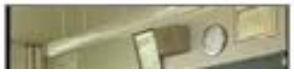

clip18.mov


dec21a.mov


site_administration_issues.mov


2011CZ406.mov


dec19e.mov



contact us

SIDGrid

home about people news & events partners portal log out

Welcome hereld

projects videos administrate

Search for Keywords like Search

projects > DemoProject

DemoProject

created by: susanne || updated

KEYWORDS:

Actions...

demo yoho testkey funfun

PROJECT FILES:

elan-example1.mpg

elan-example1.wav

ANNOTATION TIERS:

- K-Spch
- W-Spch
- W-Words
- W-POS
- W-IPA
- W-RGU
- W-RGph
- W-RGMe
- K-RGU
- K-RGph
- K-RGMe

Preview Export Selection Export All Export range from to (Ranges from -1 to 36180 milliseconds)

Process Data (select script and new experiment name)

New Experiment Name: Script: recon-all.freesurfer View Script Run Script

SIDGrid Preview

Color	Annotation Tier
	K-Spch
	W-POS

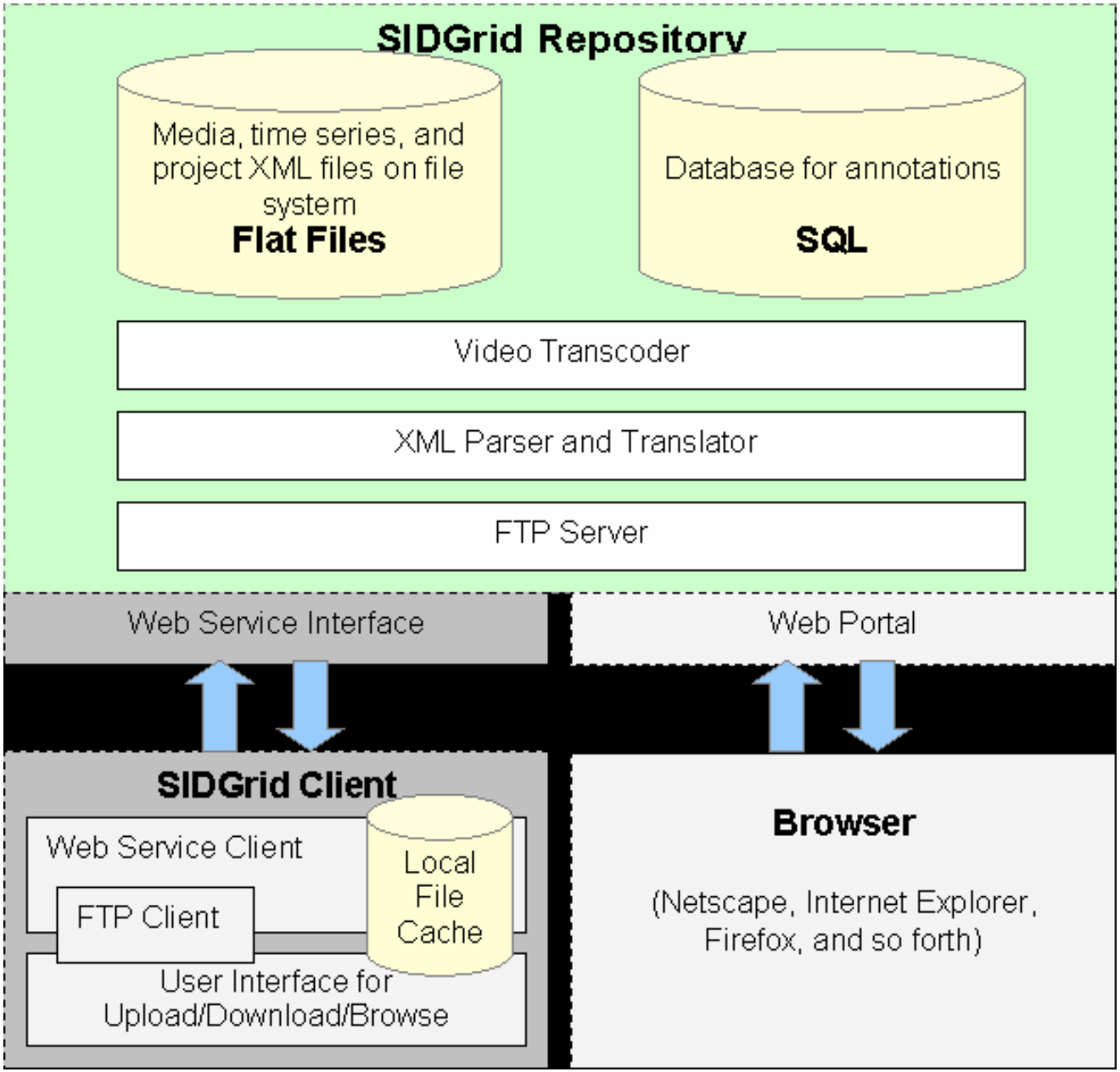
Search and Query (4,000 projects)

- Data Files
  - Names
  - Keywords
  - Attributes (keyword-value)
  - Date
  - Type (Elan, Chat)
- Contents of Files
  - Metadata
  - Tier
  - Annotations

Server Side

- Web services
  - Query
  - Data download / upload

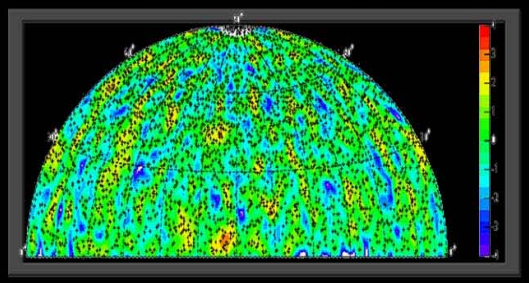
- Portal interface
  - Security
  - Data and metadata browsing
  - Preview
  - Tags, attributes
  - Projects
  - Groups
  - Search
  - Data transformation using grid resources

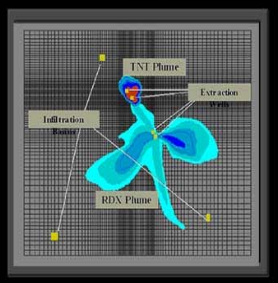

SIDGrid Repository
Media, time series, and project XML files on file system
Flat Files
Database for annotations
SQL
Video Transcoder
XML Parser and Translator
FTP Server
Web Service Interface
Web Portal
SIDGrid Client
Web Service Client
FTP Client
Local File Cache
User Interface for Upload/Download/Browse
Browser
(Netscape, Internet Explorer, Firefox, and so forth)

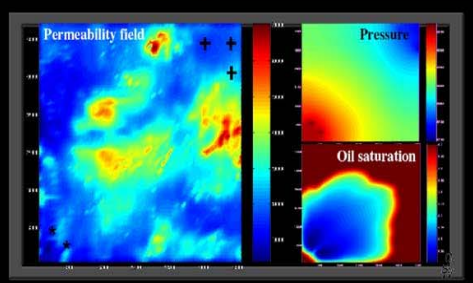
Science Gateway

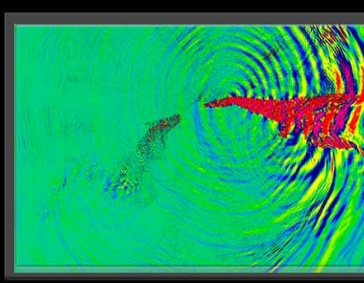
TERAGRID
Extensible Terascale Facility

ANL/UC
PSC
NCSA
PU
IU
ORNL
SDSC
TACC







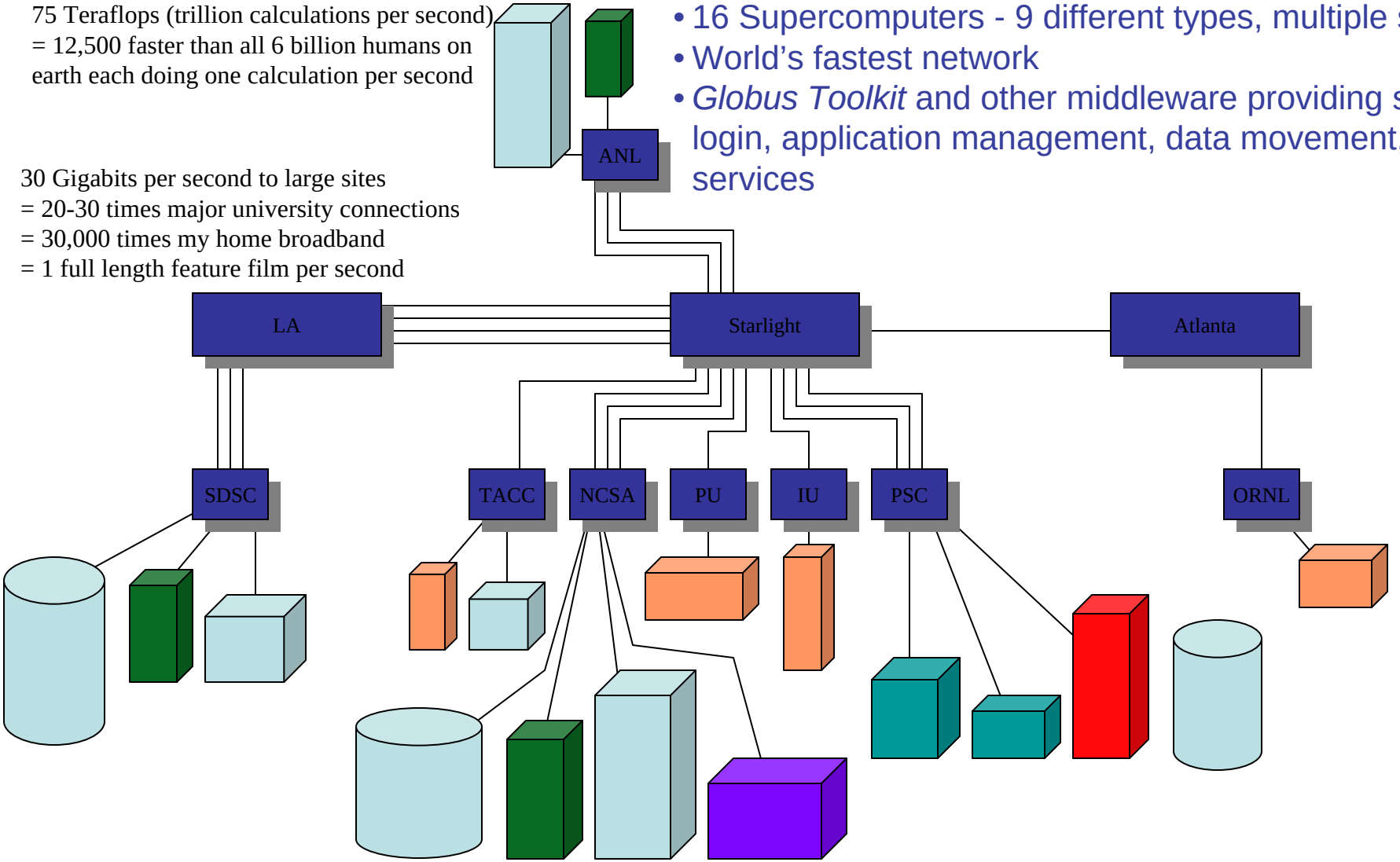


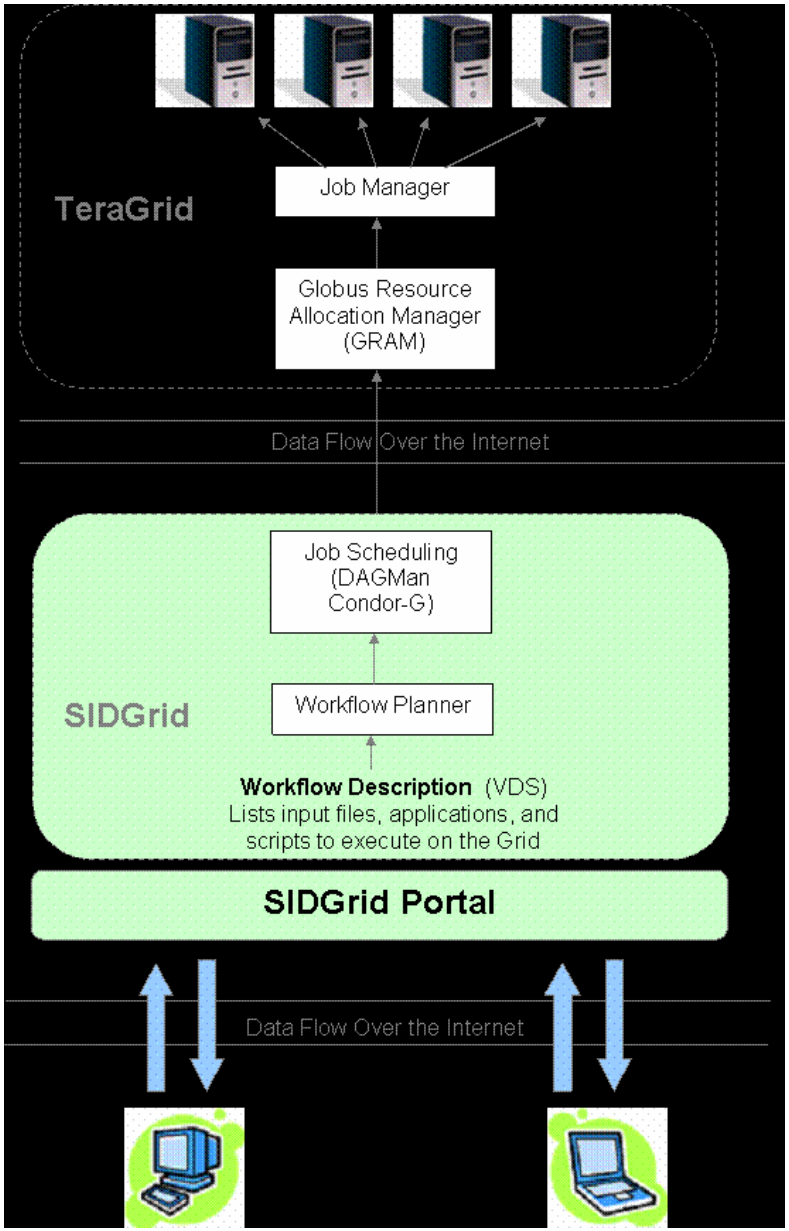
What Is The TeraGrid?

75 Teraflops (trillion calculations per second)
= 12,500 faster than all 6 billion humans on earth each doing one calculation per second

30 Gigabits per second to large sites
= 20-30 times major university connections
= 30,000 times my home broadband
= 1 full length feature film per second

- 16 Supercomputers - 9 different types, multiple sizes
- World's fastest network
- *Globus Toolkit* and other middleware providing single login, application management, data movement, web services




TeraGrid
Job Manager
Globus Resource Allocation Manager (GRAM)
Data Flow Over the Internet
Job Scheduling (DAGMan Condor-G)
SIDGrid
Workflow Planner
Workflow Description (VDS)
Lists input files, applications, and scripts to execute on the Grid
SIDGrid Portal
Data Flow Over the Internet

contact us

SIDGrid

home about people news & events partners portal log out

Welcome hereld

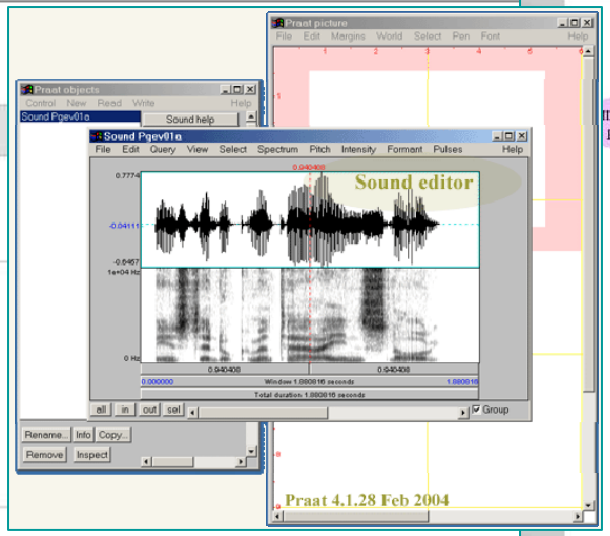
projects videos administrate

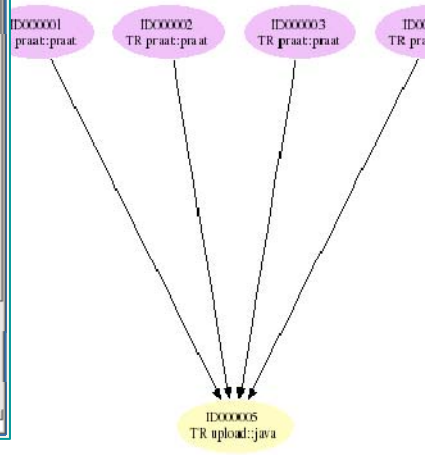
Search for Keywords like Search

job id	status
ID000001	in queue
ID000002	in queue
ID000003	in queue
ID000004	in queue
ID000005	in queue

projects > waverunner

waverunner





KEYWORDS:

Actions...

wave_test tagtest

PROJECT FILES:

- [x] his1.wav
- [x] his2.wav
- [x] his3.wav
- [x] his4.wav

Preview Export Selection Export All

Export range from to (Ranges from None to None milliseconds)

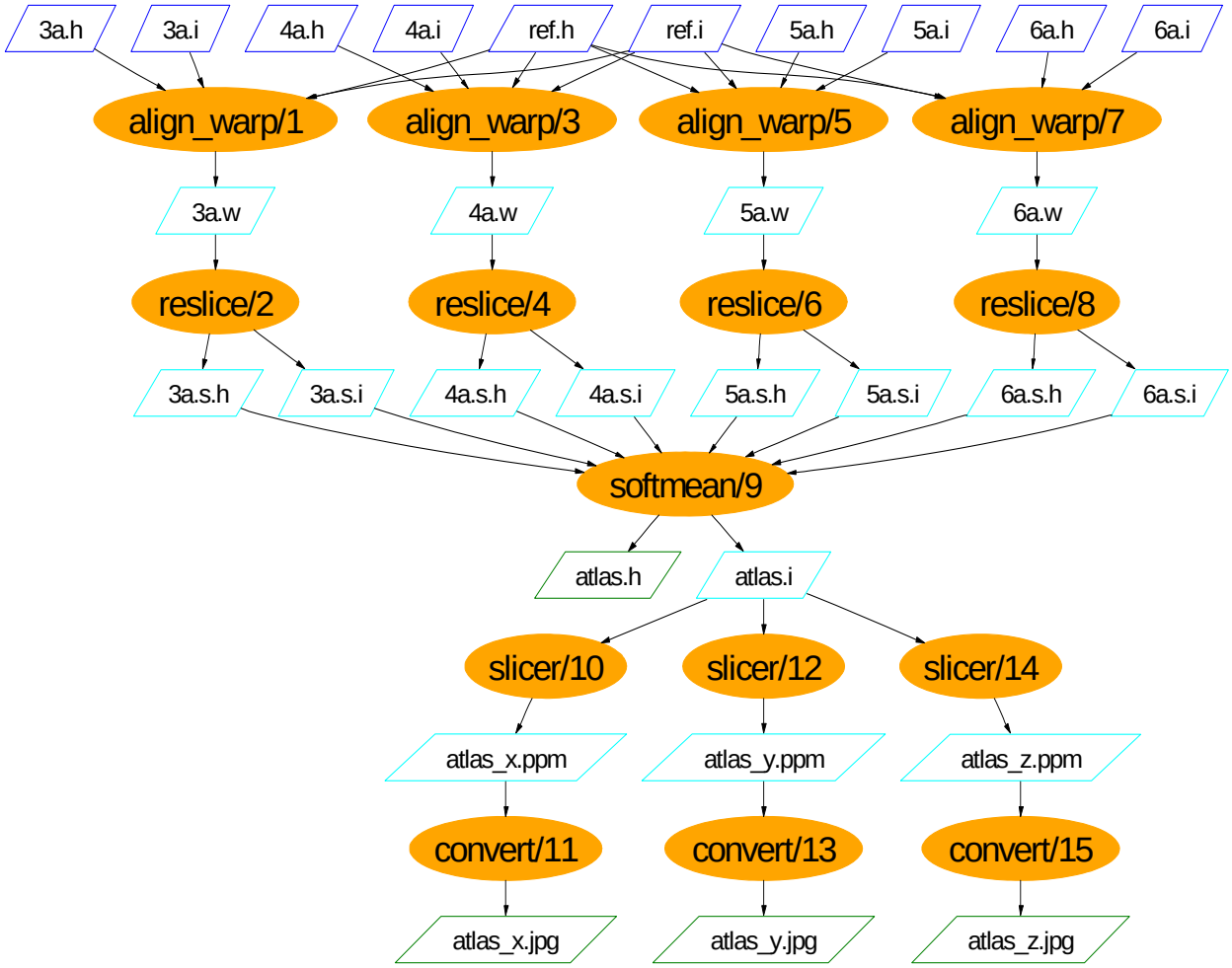
Process Data (select script and new experiment name)

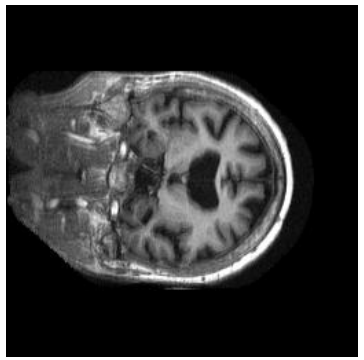
New Experiment Name: WACE2006test2 Script: pitch.praat View Script Run Script

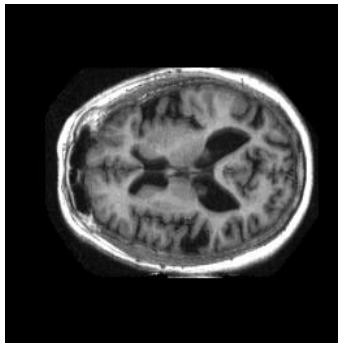
Grid-enabling Cognitive Neuroscience

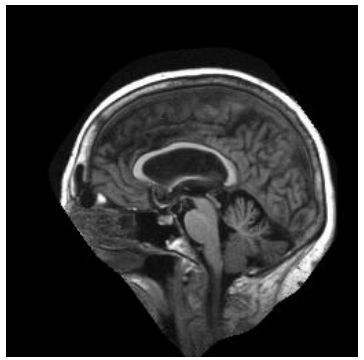
- Work done in collaboration with Jed Dobson, Dartmouth Brain Imaging Center
- Test problem: spatially normalize functional MRI volumes - prepare studies for analysis
- Goals:
  - Provide significant speedup of analysis runs
  - Reduce the manual effort (and errors) of ad-hoc methods
  - Benefits of provenance: share data and methods with remote researchers
  - Grid needs to be transparent to the users: Literally, “Grid as a Workstation”

fMRI Preprocessing



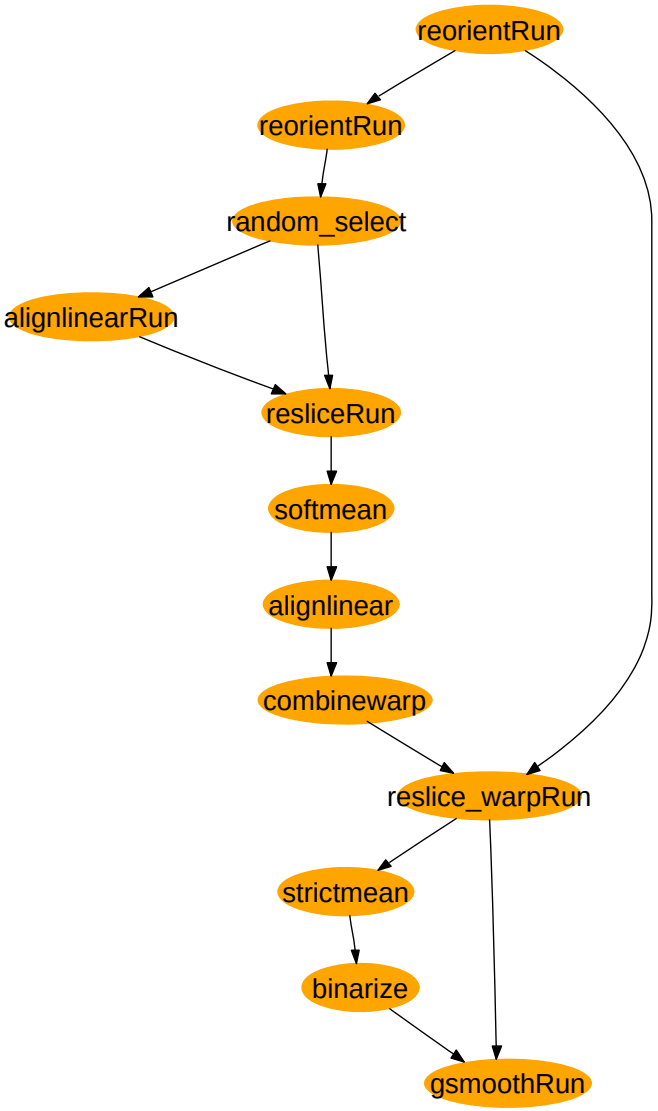




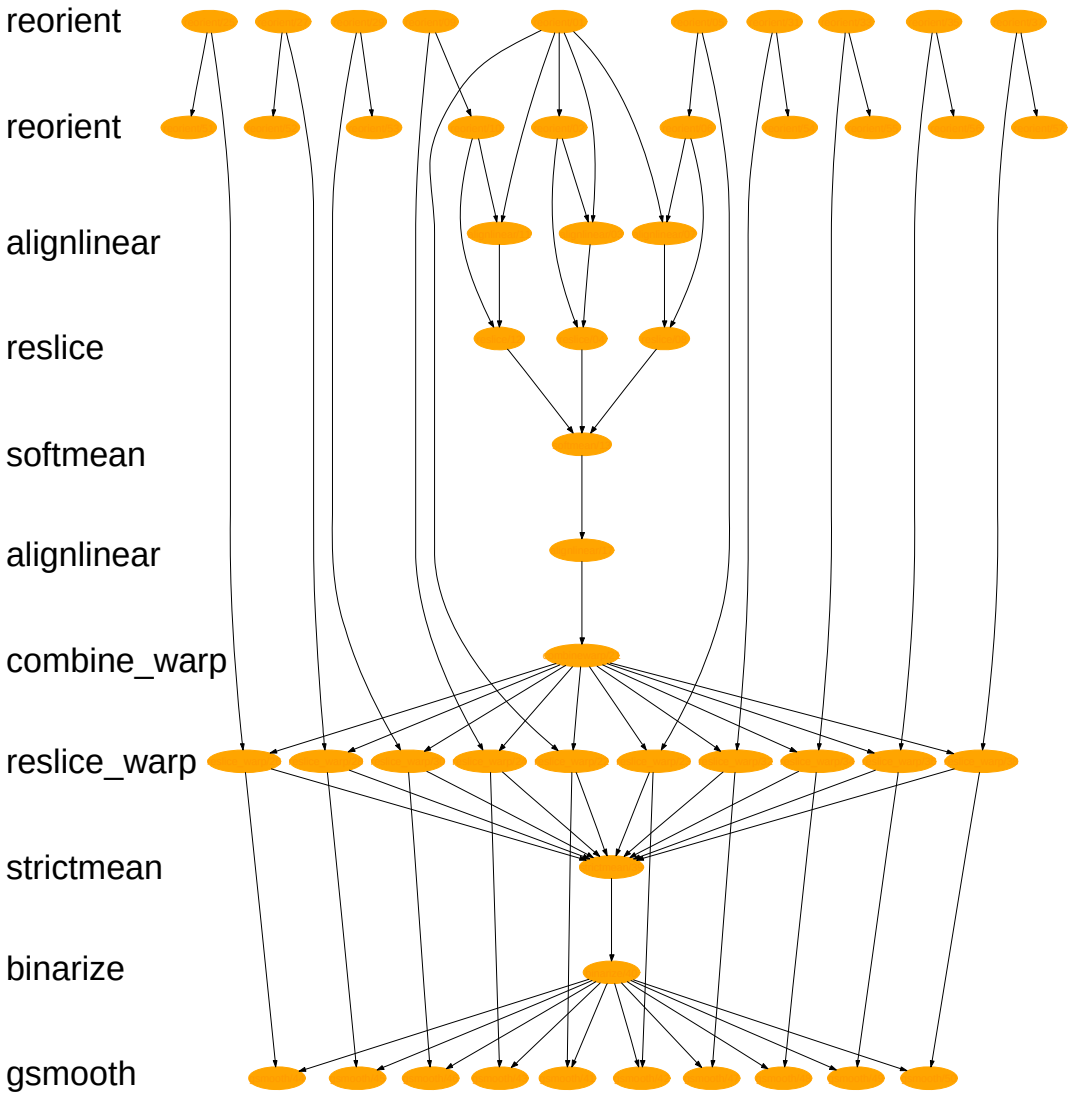


Workflow courtesy James Dobson, Dartmouth Brain Imaging Center

Spatial normalization of functional run



Dataset-level workflow

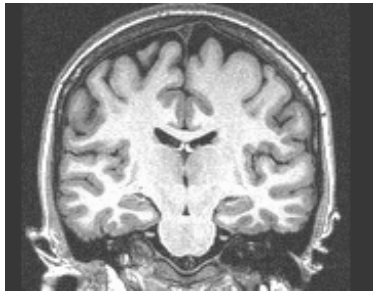


Expanded (10 volume) workflow

To create 2D cortical surface renderings from 3D anatomical images, typical modern cpu consumes ~40 hours to process one subject

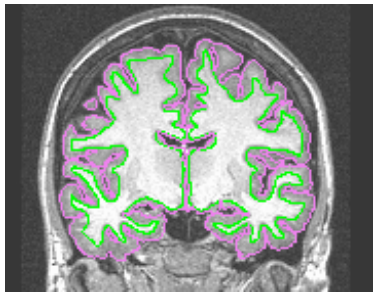
Typical neuroimaging studies have 15-20 subjects: implies more than 30 days to get surface renderings!!

Longitudinal studies: hundreds of brain scans!!!

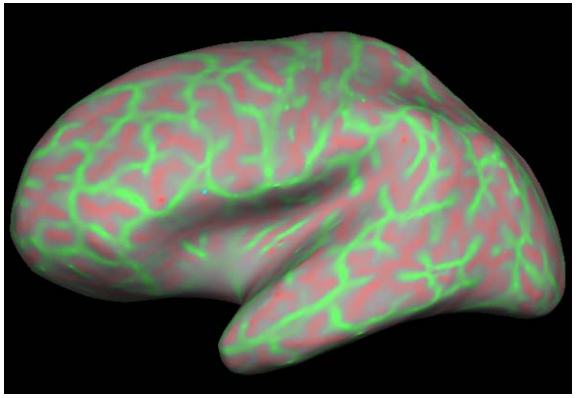


High resolution anatomical slice

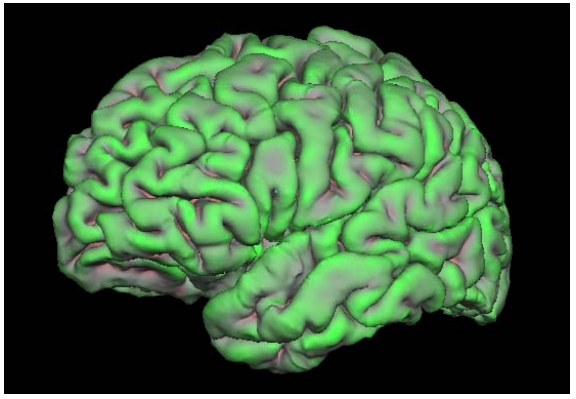
Freesurfer



Segmented grey (pink line) and white (green line) matter



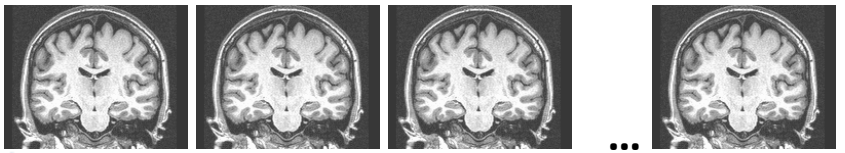
View: Inflated surface



View: Pial surface

SIDGrid

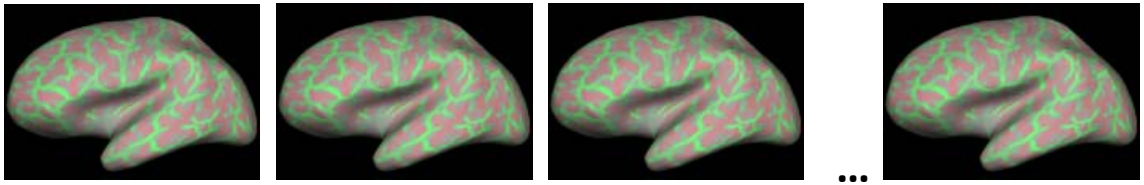
Generate cortical surface representations of whole study in 2 days!



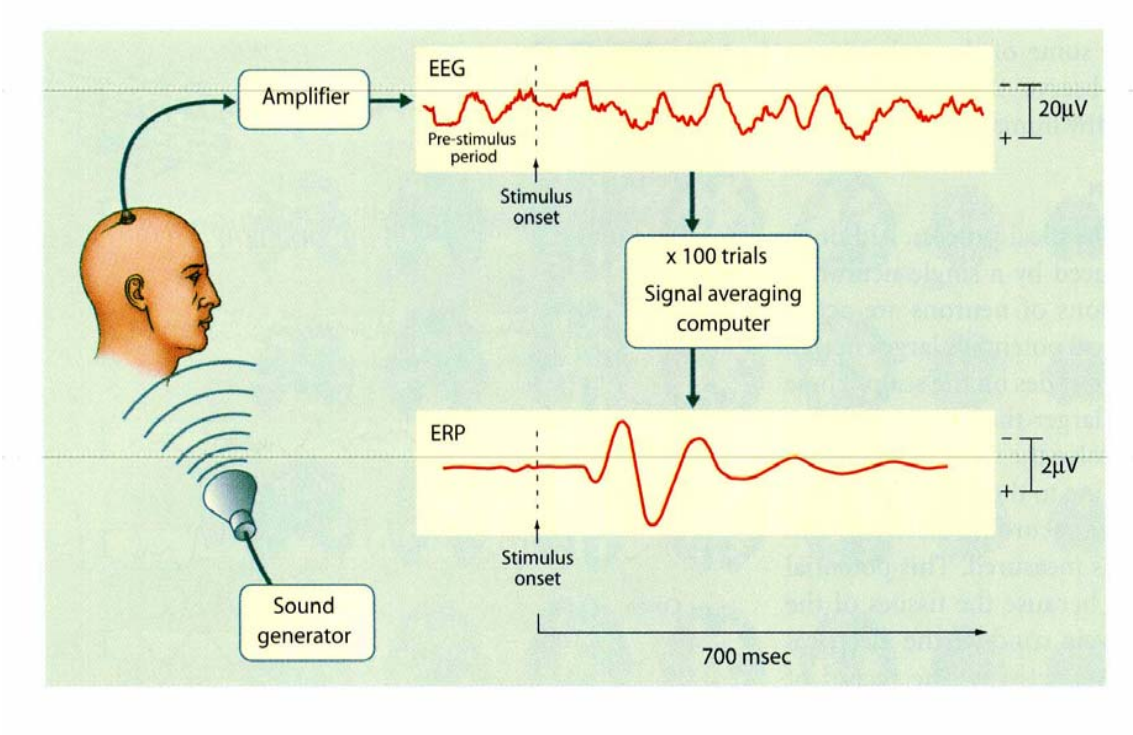
User CPU

Upload and run jobs

Download output files

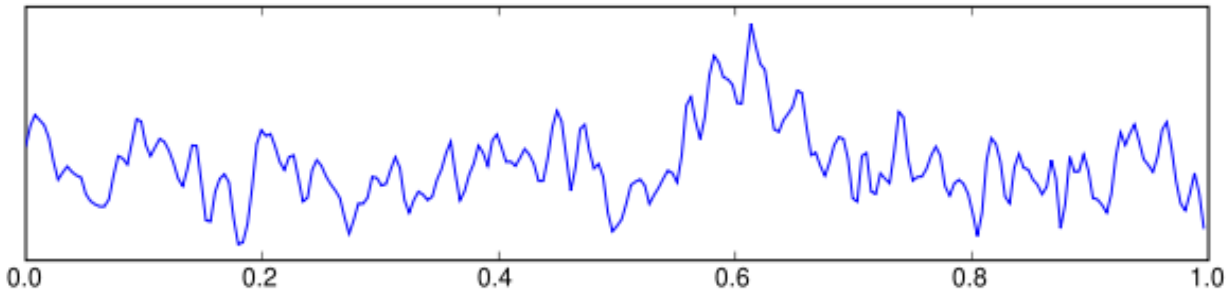


Modulation of Oscillatory Neural Synchronization



Electroencephalography

- Electromagnetic oscillations picked up by EEG electrodes are created by the somato-dendritic potentials of very large groups of ‘open field’ neurons.
  - These are neurons arranged in a parallel and structured fashion such that their membrane potentials superimpose
- Signal is filtered by the varying density and thickness of the skull and meninges.
- Raw signal is often analyzed as a superposition of frequency bands:
  - delta (.1-4hz), theta (4-8hz), alpha (8-12hz), beta (12-26hz), and gamma (26 hz+)



Functional Significance of Mu Rhythms
(Pineda, 2005, Brain Research Reviews)

- Mu rhythm is an EEG oscillation with dominant frequencies between 8-13 and 15-25 Hz
- Recent studies suggest mu rhythms reflect downstream modulation of motor cortex by prefrontal mirror neurons
  - Mu rhythms represent S-M transformations, specifically “seeing” and “hearing” into “doing”
- Mu power reduced in normal adults by self-initiated movement, imagined movement, and observed movement

Time-Frequency Analysis

Filtered Signal

Power Spectrum

Wavelet Transform

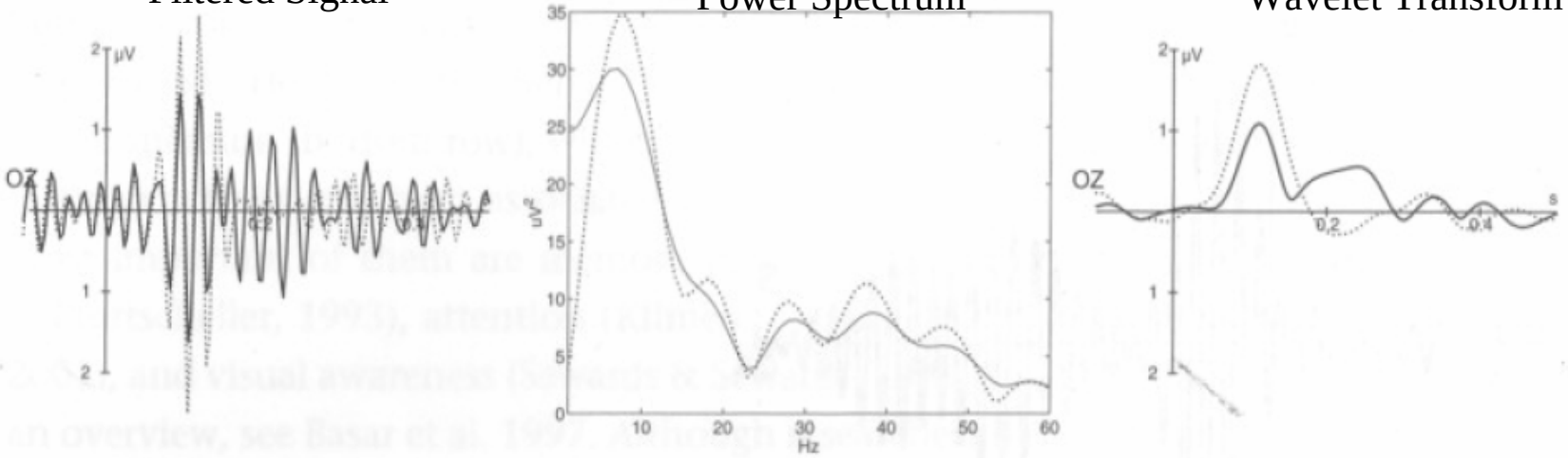
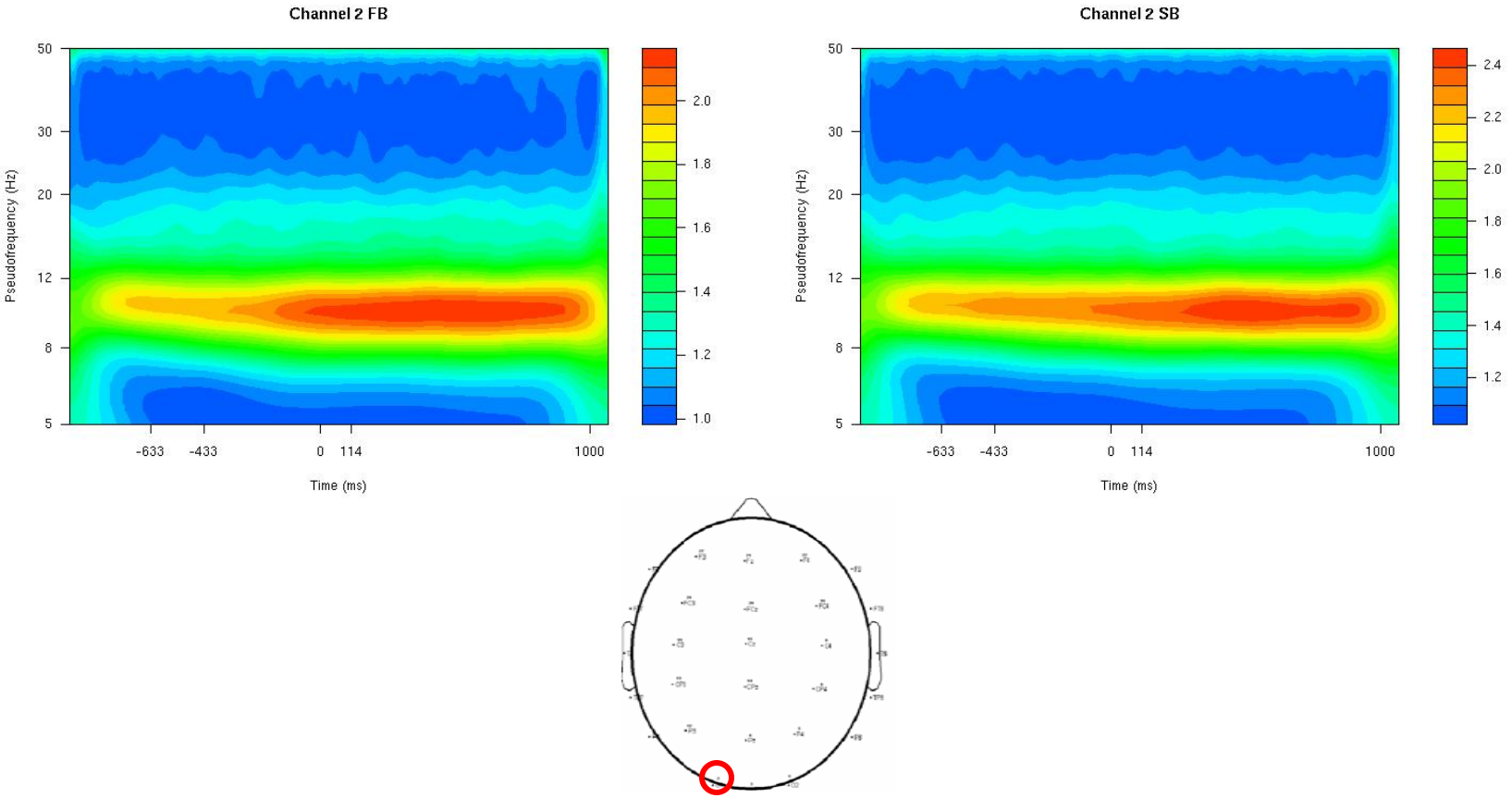
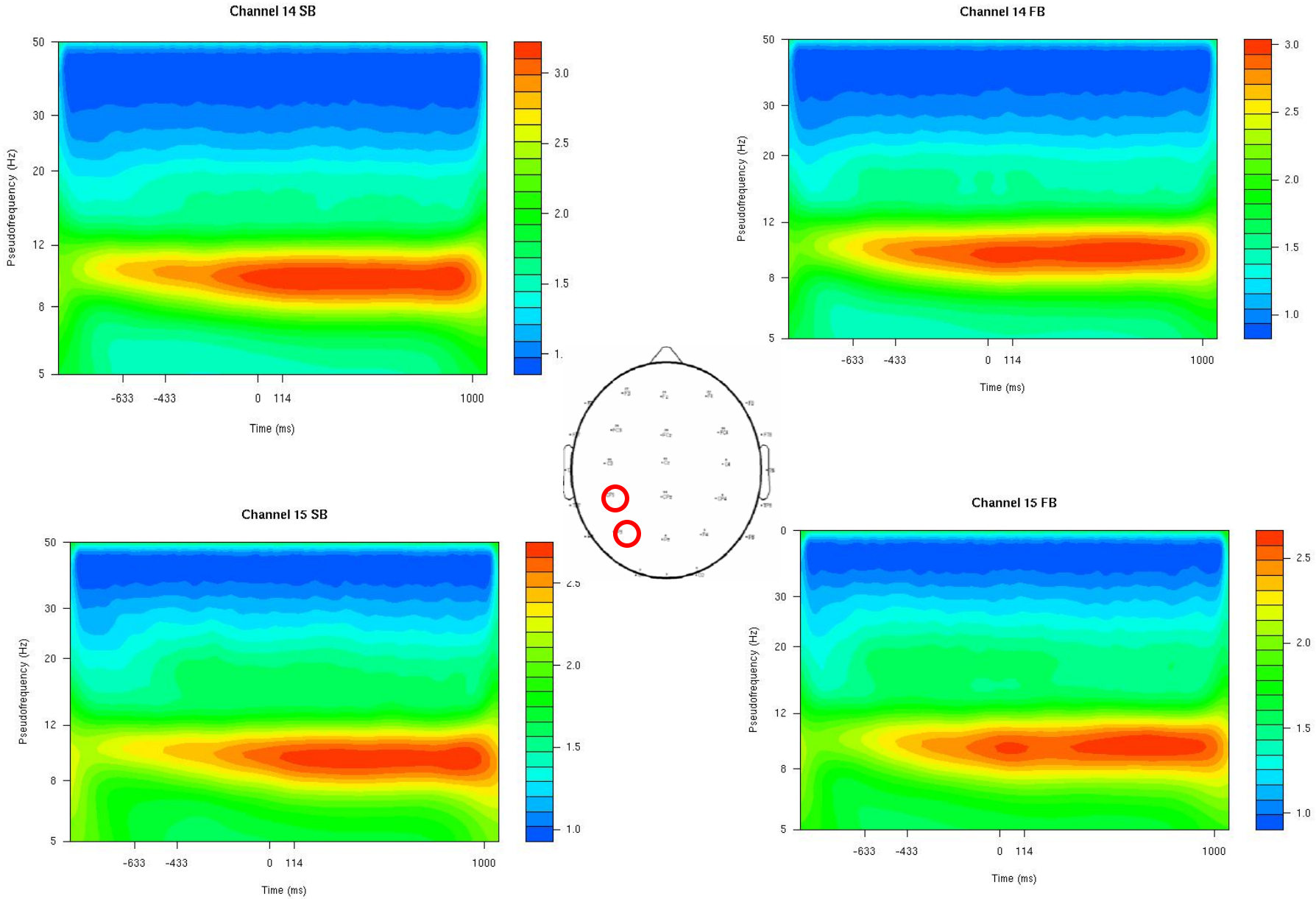
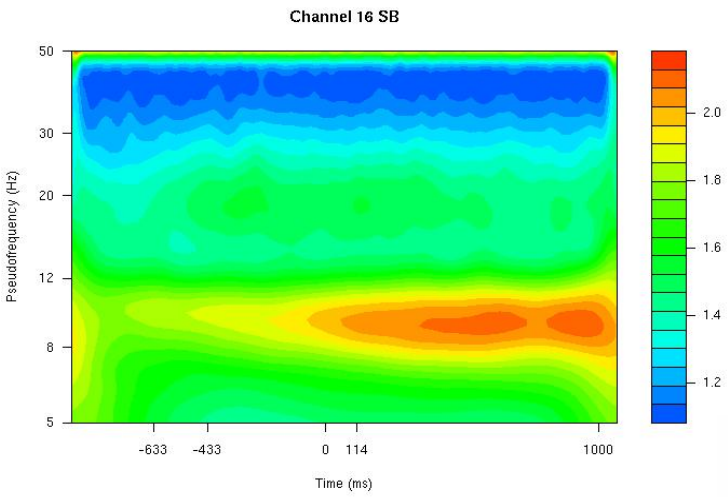


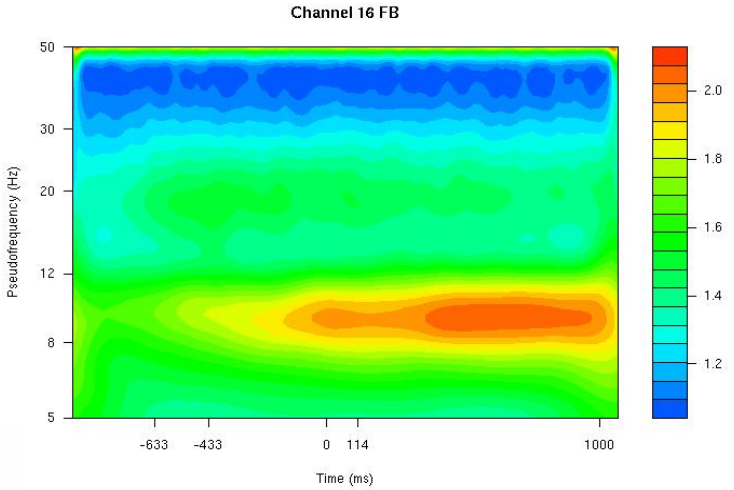
Figure 11.7

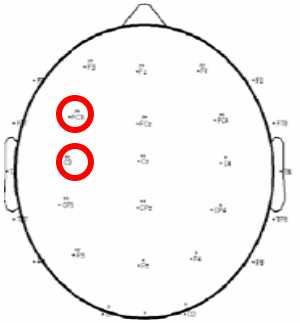
Three possibilities to extract frequency information from ERP data: two 35–45 Hz filtered ERPs (*left*), two FFT spectra of the ERPs (*middle*) and the wavelet transforms of the ERPs (*right*). Note that only the filtered signal and the wavelet transform still represent changes over time. The FFT spectra show the entire frequency range but no temporal information.

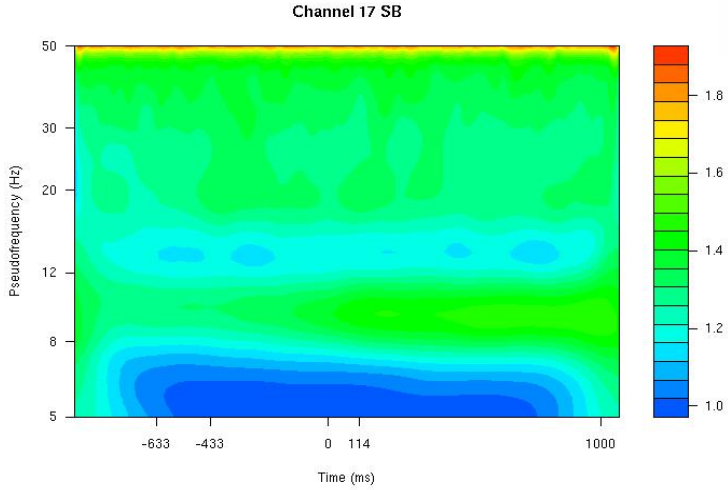

Channel 2 FB
Pseudofrequency (Hz)
50
30
20
12
8
5
-633
-433
0
114
1000
Time (ms)
2.0
1.8
1.6
1.4
1.2
1.0
Channel 2 SB
Pseudofrequency (Hz)
50
30
20
12
8
5
-633
-433
0
114
1000
Time (ms)
2.4
2.2
2.0
1.8
1.6
1.4
1.2

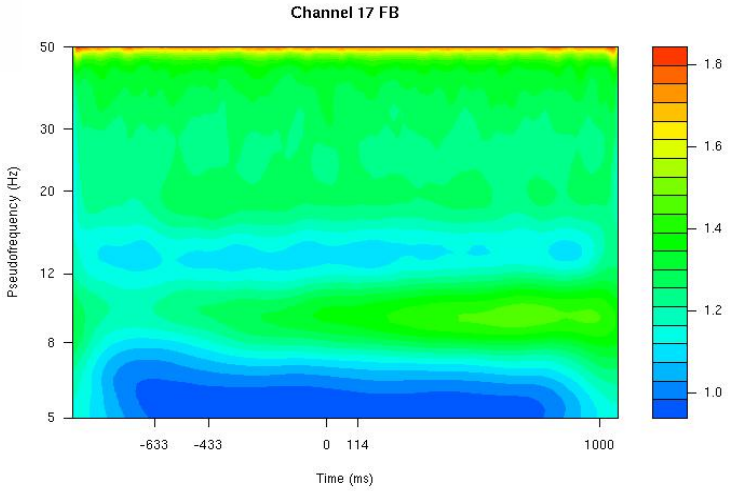

Channel 14 SB
Channel 14 FB
Channel 15 SB
Channel 15 FB
Pseudofrequency (Hz)
Time (ms)
-633
-433
0
114
1000


Channel 16 SB
Pseudofrequency (Hz)
50
30
20
12
8
5
-633
-433
0
114
1000
Time (ms)
2.0
1.8
1.6
1.4
1.2

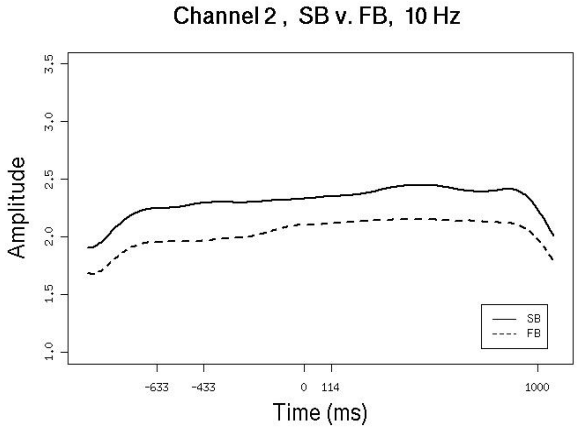

Channel 16 FB
Pseudofrequency (Hz)
50
30
20
12
8
5
-633
-433
0
114
1000
Time (ms)
2.0
1.8
1.6
1.4
1.2

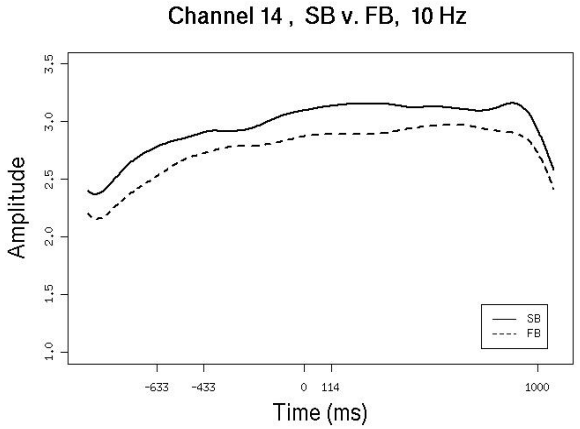


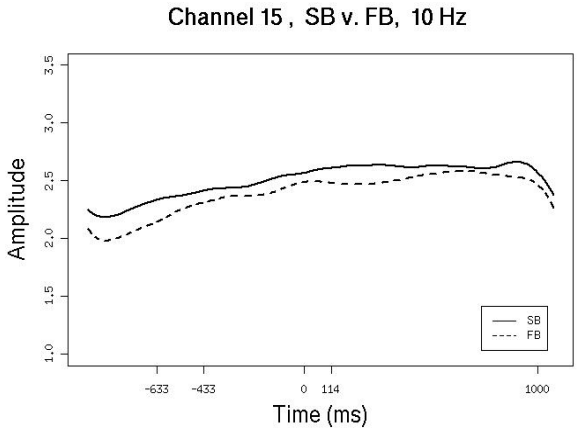

Channel 17 SB
Pseudofrequency (Hz)
50
30
20
12
8
5
-633
-433
0
114
1000
Time (ms)
1.8
1.6
1.4
1.2
1.0

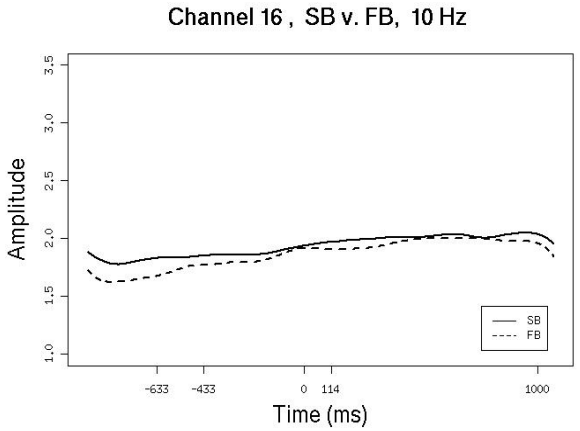

Channel 17 FB
Pseudofrequency (Hz)
50
30
20
12
8
5
-633
-433
0
114
1000
Time (ms)
1.8
1.6
1.4
1.2
1.0

Neural Synchronization at 10Hz







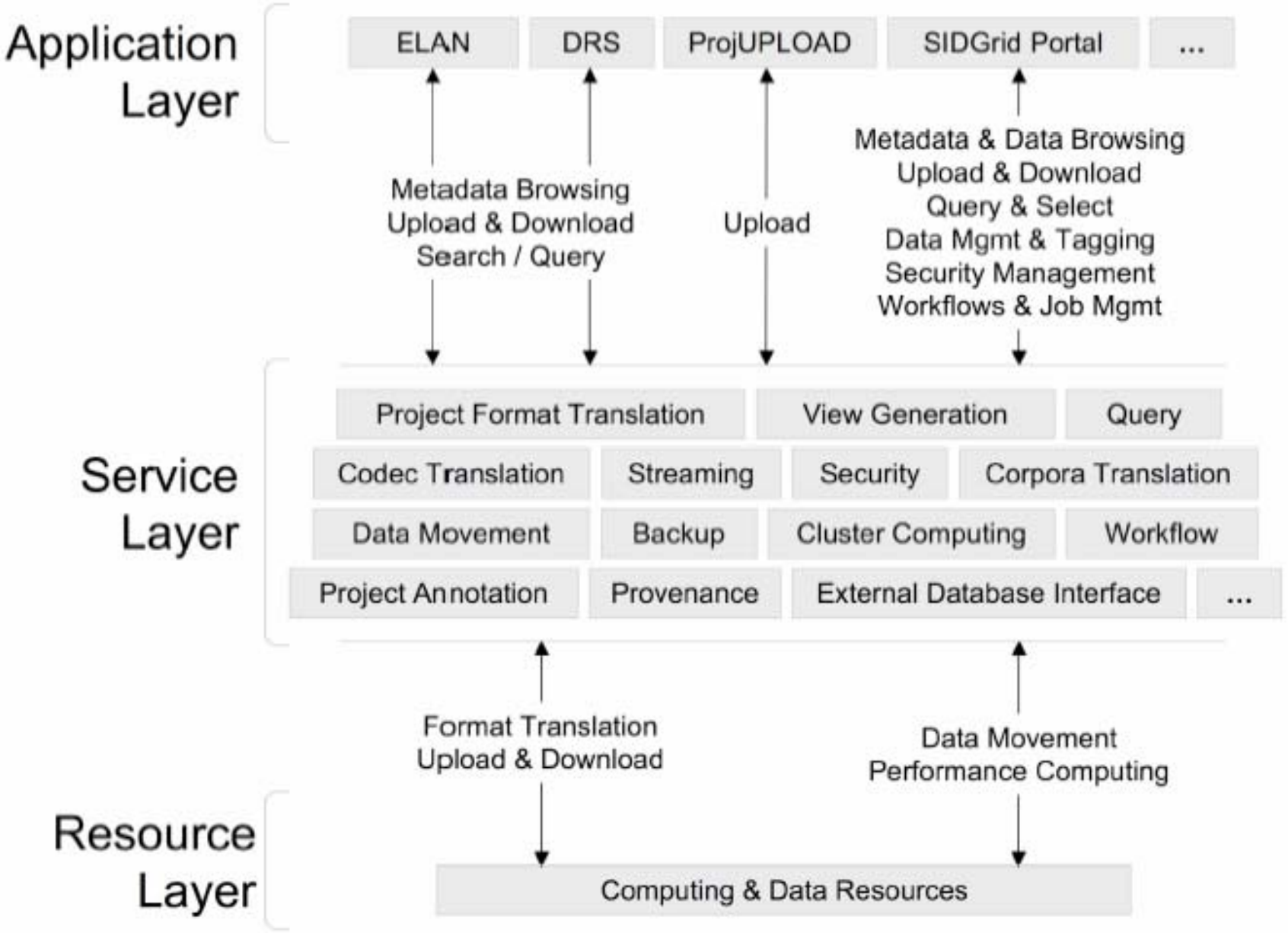


Scripts for Running Jobs on Grid

- Matlab (high-level language and interactive environment for peforming computationally intensive tasks)
- R (software environment for statistical computing and graphics)
- Praat (software for acoustic analysis)
- Free Surfer (automated tools for reconstruction of the brain's cortical surface from structural MRI data)
- AFNI (programs for processing, analyzing, and displaying FMRI data)
- SUMA (adds cortical surface based functional imaging analysis to the AFNI suite of programs)

Advantages of Grid Computing

- Vastly expanded computing and storage
- Reduced effort as needs scale up
- Improved resource utilization, lower costs
- Facilities and models for collaboration
- Sharing of tools, data, and procedures and protocols
- Recording, assessment and reuse of complex tasks


Application Layer
ELAN
DRS
ProjUPLOAD
SIDGrid Portal
...
Metadata Browsing
Upload & Download
Search / Query
Upload
Metadata & Data Browsing
Upload & Download
Query & Select
Data Mgmt & Tagging
Security Management
Workflows & Job Mgmt
Service Layer
Project Format Translation
View Generation
Query
Codec Translation
Streaming
Security
Corpora Translation
Data Movement
Backup
Cluster Computing
Workflow
Project Annotation
Provenance
External Database Interface
...
Format Translation
Upload & Download
Data Movement
Performance Computing
Resource Layer
Computing & Data Resources

Challenges and Opportunities

	Today	Tomorrow with SID Grid	Milestones
Theories & Models	Static Single cause Linear Component processes Specific time scale Single level of analysis (e.g., neural)	Dynamic Multiple causes Linear and Nonlinear Systems or networks Multiple time scales Multiple levels of analysis (e.g., neural, social)	Empirical tests of theories & models
Collaboration	Single labs Annotations by single investigators Local access only	Community of collaborators Collaborative annotation Remote & distributed access	Collaborative annotation tool
Query and Analysis	Standard statistical analyses Single stream Non-standard formats & coarse alignment of data streams Single location Stand alone application	Automated query, exploration, and analysis Multiple streams Tools to acquire, transform & align multiple data streams Multiple locations Extensible SID Grid application	Query & analysis services
Measurement & Annotation	Single measure Uni-modal Single time scale Manual coding	Multiple measures Multimodal Multiple time scales Automated coding	Multimodal data stream tool
Data Collection	Single investigator populating database on single workstation	Community of collaborators creating SID Grid data resources on grid	SuperLab Legacy datasets

Research Team

University of Chicago:

Bennett Bertenthal:	Professor, Psychological and Brain Sciences, Indiana University
David McNeill:	Professor Emeritus, Department of Psychology
Howard Nusbaum:	Chair, Department of Psychology, co-Director, Center for Cognitive & Social Neuroscience
Jean Decety:	Professor of Psychology, former Director of Research, CNRS, Lyons.
Gina Levow:	Assistant Professor of Computer Science

Argonne National Laboratory:

Ian Foster:	Distinguished Service Professor of Computer Science, Head, Distributed Systems Lab, Associate Division Director, Mathematics & Computer Science Division; Director, Computation Institute
Mark Hereld:	Experimental Systems Engineer, ANL, Fellow, Computation Institute
Michael Papka:	Acting Deputy Division Director, Mathematics & Computer Science Division, Fellow, Computation Institute
Rick Stevens:	Acting Chief Scientist, ANL, Professor of Computer Science, Director, Computation Institute, Director, Mathematics & Computer Science Division, ANL
Michael Wilde:	Software Architect, Mathematics & Computer Science Division, Fellow, Computation Institute

University of Illinois at Chicago:

Robert Grossman	Professor of Mathematics, Statistics and Computer Science, Director, National Center for Data Mining
Stephen Porges:	Professor of Psychology and Psychiatry, Director, Brain-Body Center
Sue Carter:	Professor of Psychiatry, Physiology, and Biophysics, co-Director, Brain-Body Center

Consultants:

Brian MacWhinney:	Professor of Psychology, CMU, Developer of CHILDES & TalkBank
Francis Quek:	Professor of Computer Science, VPI, Director, Center for Human-Computer Interaction

Programmers

- David Hanley
- Sarah Kenny
- Kavithaa Rajavenkateshwaran
- Thomas D. Uram
- Wenjun Wu

Questions