GAAIN Update

Arthur W. Toga
What is GAAIN?

A global cooperative of sharing, investigation and discovery for Alzheimer’s Disease Research

- Data federation platform enabling access to data across distributed studies
- Global network of analysis and workflow software and tools
- Global network of grid resources
Era of Big Data

- Exponential growth in data generation on every level
- Data is dispersed globally
- Data is heterogeneous
- Data is collected in non-standardized ways
- Google doesn’t do for searching databases, medical images or genetic data
- There is a huge amount of data out there that needs to be made accessible through a common framework
GAAIN Leadership

• Alzheimer’s Association
• Laboratory of Neuro Imaging (LONI)
• Istituto di Ricovero e Cura a Carattere Scientifico (N4U)
• Scientific Advisory Board
Scientific Advisory Board

- Paul Aisen, UC San Diego
- Rhett Alden, GE Healthcare
- Neil Buckholtz, NIH
- Enrique Castro-Leon, Intel
- Alon Halevy, Google
- William Klunk, Univ of Pittsburgh
GAAIN Partners

- Alzheimer’s Disease Neuroimaging Initiative (ADNI)
- Australian Imaging, Biomarkers and Life Study (AIBL)
- Coalition Against Major Diseases (CAMD)
- European Medical Information Framework (EMIF)
- NeuGRID
LONI Infrastructure

Data Center

• 208 servers
• 3328 cores
• 26 TB memory
• 2.43 PB usable storage
• 3.3 PB raw storage
• 136 TB tape back-up (expandable to double capacity)
GAAIN Challenges

• Subject privacy protection across international boundaries
• Complexity from cross-disciplinary, multi-source data collection and analysis
• Creating robust and compelling tools for searching, sharing, visualizing and analyzing federated data
• Data “ownership” and security
• Data analysis across distributed GAAIN infrastructure
• Ontologies, terminologies, standards
Early Progress

- Initial set of partners on-board
- Website deployed
- Additional partner applications received via online application form
- Initial set of standard terminologies selected
- Subset of data exchanged
- Mapping tool prototype in development
- Federated database infrastructure in development
- On-going communication with developers of CDISC terminologies
Partner Application

Global Alzheimer’s Association Interactive Network

PARTNERS

Network Partners in GAAIN will become part of a global effort to share and collaborate. GAAIN will link investigators, data, tools and infrastructure, creating a powerful and efficient environment for the study of Alzheimer’s Disease.

Benefits of GAAIN Network Partnership include:

• Prestigious membership in a global network of investigators committed to advancing a sustainable and federated Alzheimer’s Disease research infrastructure
• Access to a vast network of data and computer resources
• Access to a growing set of analytic tools spanning clinical, biologic, imaging and genetic data
• Data federation tools that support ontologies and metadata standards
• Online analytic tools allowing you to compare your data and analysis results with others.

Download Application Form (PDF)

Are you a GAAIN Partner?

Our expectations for new GAAIN partners include:

• Extensive, accessible and well organized set of Alzheimer’s Disease research data from imaging, behavior or genetics (300+ research subjects)
• Experience and interest in collaborative efforts
• Ability to share re-identified data
• Reliable computer resources and high speed internet access to support data federation
• Technical expertise sufficient to test and implement data federation tools
• Willingness to beta test new tools
• Proponent of ethical data sharing

Only data collected in a manner that allows sharing in de-identified form may be part of the GAAIN effort. Strong and responsive support from your institution’s ethics review committee is a plus.
Standard Terminologies

• Standard terminologies enable mapping of heterogeneous data into common terms
  – Supports searching across databases
  – Supports data federation

• CDISC Terminologies
  – CDISC AD v1.1 standard for Alzheimer’s Disease
  – On-going discussions with CDISC on the neuroimaging standard currently under development
Mapping Data

- Global Schema
- Global Terminologies
- Standard codes

- Local data
- Local terminology
- Local codes

- Local data
- Local terminology
- Local codes

- Global Schema
- Global Terminologies
- Standard codes
Mapping Tool Prototype

[Image of a mapping tool prototype with tables and data]

- Subject identifier: 12762, education level: 3, marital status: married, primary occupation: housewife
- Subject identifier: 12761, education level: 8, marital status: widowed, primary occupation: dealer
- Subject identifier: 12760, education level: 4, marital status: widowed, primary occupation: labourer
- Subject identifier: 12758, education level: 5, marital status: widowed, primary occupation: dealer
- Subject identifier: 12755, education level: 11, marital status: married, primary occupation: employee

Map Name: Occupation

- Occupation expression: Case: employee
- IF: 0
- THEN: employee
- ELSE: 0

[Table for another dataset]
Data Federation
# Inventory Search

<table>
<thead>
<tr>
<th>INVENTORY</th>
<th>ADNI</th>
<th>AIBL</th>
<th>DIAN</th>
<th>NAU</th>
<th>Partner X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuropsychological Assessments</td>
<td>☑</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>MMSE: Mini-mental State Examination</td>
<td>☑</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>BDI: Beck Depression Inventory</td>
<td>☑</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>ADAS: Alzheimer’s Disease Assessment Scale</td>
<td>☑</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>Clinical Dementia Rating</td>
<td>☑</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>Functional Assessments</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>Biological Markers</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>Physical Examinations</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>Neural Imaging</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
<tr>
<td>Genetics</td>
<td></td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td></td>
</tr>
</tbody>
</table>
Ontology-Directed Search
Ontology-Directed Search
Genetic Data

• Circa 2010 GWAS Data (per sample)
  • 620,000+ rows of data
  • ~81MB

• 2012: Full Genome Sequencing (per sample)
  • Standard output from Illumina – multiple files and formats
  • ~250GB per sample

  – Example (ADNI)
    • 800 subjects x 250GB = 195TB
    • Time to transfer 195TB:
      – High speed internet (90 Mbit/s): 26 days
      – DSL (45 Mbit/s): 59 days
      – Dial-up (56 kbit/s): 100+ years!
Reminder: The ADNI Data Use agreement prohibits unauthorized sharing of these data, posting to public databases and any attempt data to identify individuals using these data. By downloading these data you acknowledge to our terms & conditions. [Link to Data Use Agreement].

ADNI WGS: ALL

ADNI Whole Genome Sequencing (WGS) samples were genotyped using the illumina Omni 2.5M BeadChip and assembly performed using CASAVA-1.9.0a1_110909.90.

WGS Documentation

Data Source Reference for WGS Samples
Version: 1
.xlsx format

WGS Methods
Version: 1
.pdf format

Indels Data

Indels Data (1 of 7)
Indels Data (2 of 7)
Indels Data (3 of 7)
Indels Data (4 of 7)
Indels Data (5 of 7)
Indels Data (6 of 7)
Indels Data (7 of 7)

Indels Data .vcf format

SNPs Data

SNPs Data (1 of 28)
SNPs Data (1 of 28)

SNPs Data .vcf format
Lend Lease
# Neuroimaging Study Size (Typical)

<table>
<thead>
<tr>
<th>Year</th>
<th>Size</th>
<th>Equivalent to</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>54MB</td>
<td>20 copies of War and Peace</td>
</tr>
<tr>
<td>2005</td>
<td>67MB</td>
<td>24 copies of War and Peace</td>
</tr>
<tr>
<td>2012</td>
<td>531MB</td>
<td>193 copies of War and Peace</td>
</tr>
</tbody>
</table>
Image Data Growth Process

Each neuroimaging scan can spawn many derived image leading to exponential growth

ADNI Example:
One 22MB structural scan
  Five preprocessed images (176 MB)
  Eleven postprocessed images (222 MB)

22MB of raw data produces 420MB data for one small scan!