XXXVIII Reunião Anual da SBNeC

Introdução ao Armazenamento de Dados de Experimentos em Neurociência

Parte 1: Estratégias para o armazenamento de dados de experimentos em Neurociência – uma visão geral

Amanda S. Nascimento
DCC/UFOP
Kelly R. Braghetto
DCC- IME/USP

11 de setembro de 2014

Quem Somos

- Amanda S. Nascimento
  - Área de Pesquisa: Engenharia de *Software*
  - E-mail: anascimento@iceb.ufop.br

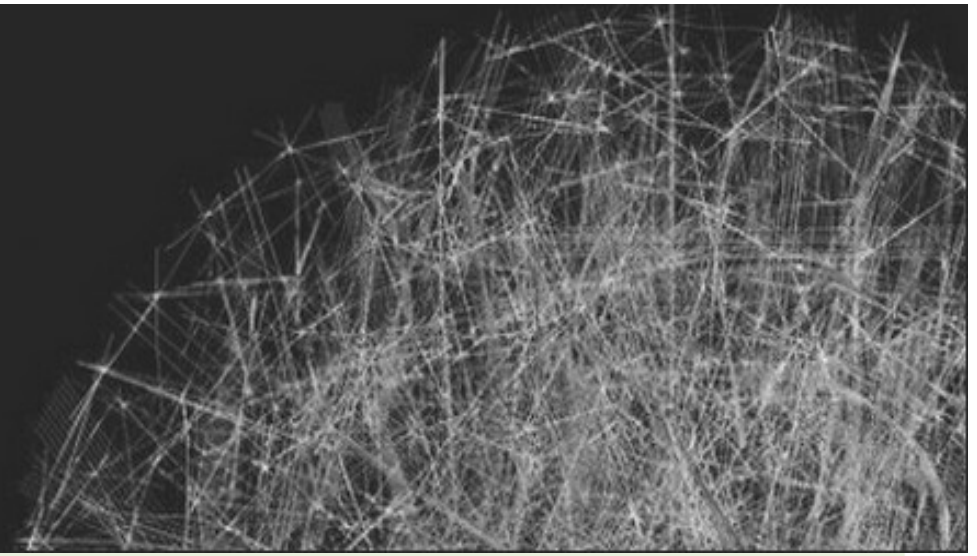
Universidade Federal
de Ouro Preto

Quem Somos

- Kelly R. Braghetto
  - Área de Pesquisa: Modelagem de Dados e Processos
  - E-mail: kellyrb@ime.usp.br

IME - Instituto de Matemática e Estatística

NeuroMat



Planeta by León Ferrari

Centro de Pesquisa, Inovação e Difusão científica (CePID) em Neuromatemática (NeuroMat)

FAPESP

Principal Objetivo

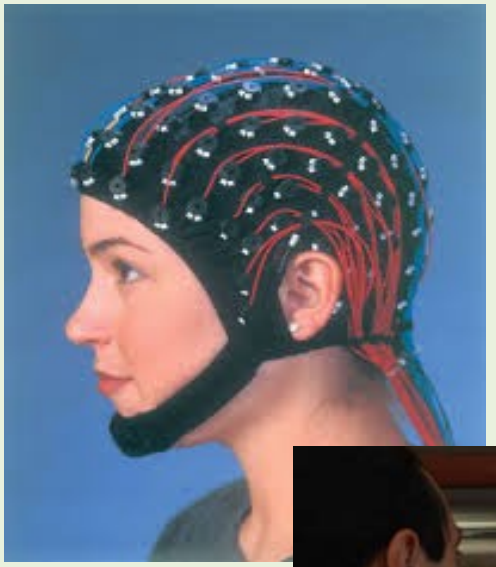
"Criação de um centro de matemática, integrando modelagem matemática com pesquisa básica e aplicada na fronteira da neurociência. A proposta responde à crescente importância da matemática na neurociência teórica."

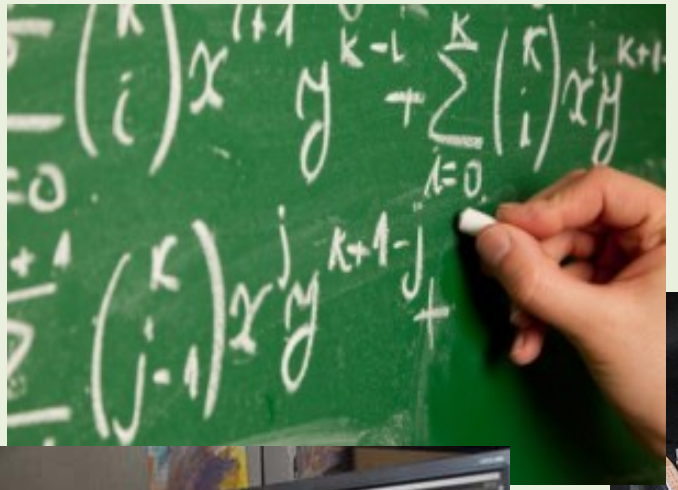
http://neuromat.numec.prp.usp.br/

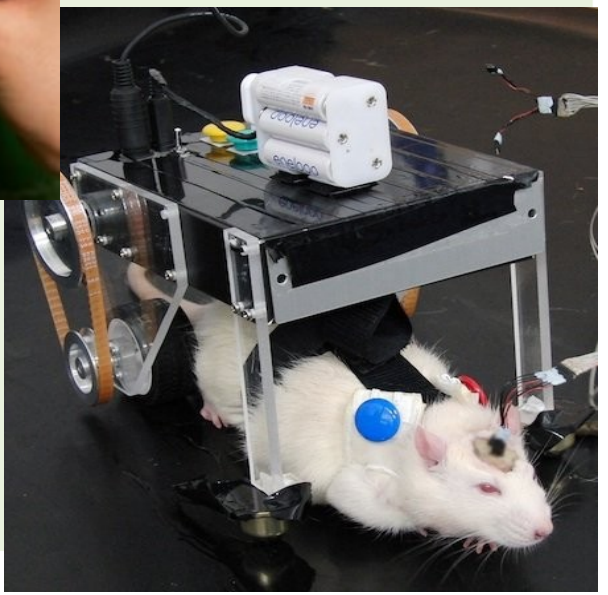
Planeta by León Ferrari

NeuroMat

Projeto Interdisciplinar







Sobre o Curso

- Material disponível em:

http://www.ime.usp.br/~kellyrb/sbnec2014/

Dados em Neurociência

- Em Neurociência, um mecanismo importante que os cientistas têm para estudar o funcionamento do cérebro humano são *experimentos*.

- Experimentos geralmente envolvem
  - a coleta,
  - o armazenamento e
  - a análise de **dados**.

Tipos de Dados

- **Dados "brutos"** – coletados em experimentos (*e.g.*, eletrofisiológicos, neuroimagens, comportamentais, moleculares, histopatológicos, etc.)

  → *Metadados*: informações sobre o protocolo experimental, equipamento de aquisição, configurações do equipamento, etc

- **Dados derivados** – gerados por meio de processamentos (*e.g.*, filtros, transformações, análises, etc.).

  → *Metadados*: Algoritmos aplicados e em qual sequência, parâmetros utilizados na execução de algoritmos e/ou métodos de análise.

Sobre os Dados Brutos

- Coleta “laboriosa”
- Pequeno volume (escassez de sujeitos)
- Alto custo
  - Tempo
  - Infraestrutura
- Pouco valorizados (usualmente não resultam em publicação)

→ Contradição: essenciais!

*"... because good research needs good data ..."**

- Quais valores é possível **agregar aos dados**?
  - Facilidade de uso (o que é o dado, o que ele significa?)
  - Facilidade de recuperação (como eu encontro a informação que preciso?)
  - Garantia de "vida-longa"* (o dado ainda é válido?)
  - **Registro de proveniência** (qual é a origem do dados?)
  - Facilidade de compartilhamento

* *Digital Curation Centre* (*DCC*) – http://www.dcc.ac.uk/

*"... because good research needs good data ..."**

- Quais valores é possível **agregar aos experimentos**?
  - Reprodutibilidade
  - Possibilidade de meta-análises

* *Digital Curation Centre* (*DCC*) – http://www.dcc.ac.uk/

Proveniência de Dados

- Assunto que vem sendo bastante discutido nos últimos anos
- **Objetivo –** responder questões frequentes dos cientistas:
  - *Quando*
  - *Onde*
  - *Como*
  - *Por quem*
  - *Por quê*

  … um dado foi gerado

Exemplos de Dados de Proveniência

- Experimento envolvendo coleta de sinais de EEG:
  - **Sistema de aquisição** (modelo do equipamento, fabricante, software, …)
  - **Configurações do equipamento** (taxa de aquisição do sinal, filtro amplificador, …)
  - **Sistema de posicionamento dos eletrodos** (sistema internacional 10-20, …)

Exemplos de Dados de Proveniência

- Experimento envolvendo coleta de sinais de EEG:
  - **Tamanho da toca de eletrodos** (P, M, G)
  - **Informações sobre o protocolo do experimento**
  - **Informações sobre quem conduziu o experimento** (afiliação, grupo de pesquisa, …)
  - **Informações sobre os sujeitos do experimentos** (sexo, idade, condição clínica, ...)

Reprodutibilidade

- Outro assunto que vem sendo bastante discutido nos últimos anos
- É importante para garantir **ciência de melhor qualidade**
  - Coíbe publicação de resultados falsos
  - Algumas revistas científicas já condicionam a submissão ou a publicação de um artigo à disponibilização de seus dados experimentais

Como Agregar Tais Valores aos Dados?

Organizando o armazenamento (= criando *bancos de dados*):

- Identificar e caracterizar os dados relevantes do experimento
- Buscar padrões que se apliquem a esses dados
- Definir a estrutura
  - Quais são as *entidades* e *atributos*?
  - Quais são os tipos, formatos e restrições dos dados?
- Definir políticas de segurança
  - Controle de acesso
  - Réplicas (*backup*)

Como os Dados são "Tradicionalmente" Armazenados e Compartilhados

Armazenamento

- Anotações em papel
- Planilhas eletrônicas
- Arquivos texto

Compartilhamento

- Troca de e-mails
- Dropbox
- Google Drive
- Unidades de armazenamento externo (pen-drive, HD)

Vantagens e Desvantagens

Armazenamento

- **Anotações em papel**
  - \+ Simplicidade (não requer conhecimentos específicos)
  - – Dificuldade para análise, recuperação, controle de acesso e *backup*
- **Planilhas eletrônicas**
  - \+ Facilidade de análise e de *backup*
  - – Dificuldade para recuperação e controle de acesso
- **Arquivos texto**
  - \+ *Backup*
  - – Dificuldade para análise, recuperação e controle de acesso

Vantagens e Desvantagens

Compartilhamento

- **Troca de e-mails**

  \+ Familiaridade no uso

  – Falta de privacidade (no caso de e-mails não institucionais), restrição de espaço e de tamanho de arquivo

- **Dropbox, Google Drive**

  \+ Facilidade no compartilhamento

  – Falta de privacidade

- **Unidade de armazenamento externo** (pen-drive, HD)

  \+ Grande espaço, sem "sobrecusto" de envio de dados pela internet

  – Dificuldade de compartilhamento

Behavioral Experiment Software

- São usados na execução de experimentos para:
  - Controlar a exibição dos estímulos visuais e sonoros aos sujeitos
  - Apresentar as tarefas aos sujeitos
  - Capturar respostas às tarefas (e.g., clique de *mouse*, teclas pressionadas, etc.)
  - Fazer a interface com outros dispositivos de coleta de dados brutos (ex.: sinais de EEG)

Behavioral Experiment Software

Por que não são suficientes para "guardar" dados?

- Só registram as informações necessárias para controlar a exibição dos estímulos e para a captura das respostas.
- Registram as informações em formatos proprietários, dificultando o reuso dos dados.

Behavioral Experiment Software

Por que não são suficientes para "guardar" dados?

- Não registram informações sobre o protocolo experimental completo:
  - Contextualização do experimento
  - Caracterização dos grupos de sujeitos
  - Configuração dos equipamentos usados

Behavioral Experiment Software

- **Código aberto** / gratuitas
  - OpenSesame (http://osdoc.cogsci.nl/2.8.3/)
  - PsyToolKit (http://psytoolkit.gla.ac.uk/)
  - DMDX (http://www.indiana.edu/~clcl/Q550_WWW/DMDX.htm)
- **Código fechad**o / pagas
  - Presentation (http://www.neurobs.com/)
  - SuperLab (http://www.superlab.com/)
  - e-Prime (http://www.pstnet.com/eprime.cfm)

Carência de Padrões de Dados em Neurociência

- A Neurociência não tem padrões para **representação** e **armazenamento** de dados de experimentos
  - Representação: quais "campos" são necessários para acomodar os dados?
  - Armazenamento: quais formatos de arquivos podem guardar os dados de forma mais eficiente?

Exemplo a ser seguido: **Bioinformática**

- **FASTA**: padrão para dados de sequências de genoma.

Alternativas para Gerenciar Dados (Digitais) de Experimentos

- Sistema (*software*) + banco de dados específico
  - Geralmente é desenvolvido para um único domínio
  - Acomoda tanto os dados brutos quanto os metadados
- **Sistemas de gerenciamento de questionários eletrônicos**
  - Solução de "propósito geral"
  - Acomoda bem metadados e alguns tipos de dados brutos
- **Sistemas "locais" de compartilhamento de arquivos**
  - Solução de "propósito geral"
  - Melhor para dados brutos de "grande porte"

Sobre Bancos de Dados na Neurociência

- A wikipédia tem uma lista dos mais conhecidos:

Database for Reaching Experiments And Models (DREAM) [16]	Reaching data. (Behavioral, generalization, adaptation, learning, spike, fMRI, uncertainty)	Human and monkey	Macroscopic	Kinematic, Spike, fMRI	Healthy
The fMRI Data Center [17]	fMRI datasets from published studies	Human	Macroscopic	fMRI datasets	Healthy
Invertebrate Brain Platform [18]	Photos of dissections of invertebrates nervous systems	Invertebrates (47 species in all)	Macroscopic	Photos	Healthy
Major Depressive Disorder Neuroimaging Database [19]	Meta-analysis and database of MRI studies	Human	Macroscopic	Descriptive, numerical	Major Depressive Disorder
Mouse Brain Library [20]	Atlas, stained sections from mouse brains	Mouse	Macroscopic	Images	Healthy
Neuromorpho.org [21]	3D models of real neurons	Human, Rat, Mouse, Monkey, others	Neuron	Images and 3D data	Healthy
NeuronDB [22]	Database of Neuron properties and classification	Human	Neuron	Descriptive	Healthy
Neuroscience Information Framework [23]	actually a meta database of neuroscience-relevant data incorporating over 100 databases like brainmaps, BREDE, ABA, MGI etc.	Human, Mouse, Rat, Worm	Microscopic, Macroscopic	Datasets	Healthy and Diseased

A maioria agrupa dados coletados no escopo de um projeto específico.

http://en.wikipedia.org/wiki/List_of_neuroscience_databases

Exemplos de BDs Abertos

OASIS

Home | Data access tools | Contact | About

What is OASIS?

The Open Access Series of Imaging Studies (OASIS) is a project aimed at making MRI data sets of the brain freely available to the scientific community. By compiling and freely distributing MRI data sets, we hope to facilitate future discoveries in basic and clinical neuroscience. OASIS is made available by the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN).

When publishing findings that benefit from OASIS data, please include the following grant numbers in the acknowledgements section and in the associated Pubmed Central submission: P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584.

People Publications Contact Data use agreement Data access tools

http://www.oasis-brains.org/app/template/Index.vm

Exemplos de BDs Abertos

Welcome to EEG.pl

EEG.pl is an open repository for software, publications and datasets related to the analysis of brain potentials: electroencephalogram (EEG), local field potentials (LFPs) and event related potentials (ERP), created to foster and facilitate Reproducible Research in these fields.

You can freely search the content of this and other thematic vortals linked via the inter-neuro initiative. As a registered user you can submit your article, data or model. Registration and submissions are free. You can also *comment and respond to comments* on any of the published items.

Service by Department of Biomedical Physics, University of Warsaw. Supported by polish funds for science 2001-2007 grants 4438/IA/115/2003 and 115/04/E-343/S/2006-3. Portal based exclusively upon Open Source tools: Plone, CMF and Zope. Implementation by CC, server running GNU/Linux.

http://eeg.pl/

Exemplos de BDs Abertos

View Edit History Print

Home
- News
- People
- Projects
- Publications

Resources
- Datasets
- Atlases
- Protocols
- Software

edit SideBar

IXI dataset

In this project we have collected nearly 600 MR images from normal, healthy subjects. The MR image acquisition protocol for each subject includes:

- T1, T2 and PD-weighted images
- MRA images
- Diffusion-weighted images (15 directions)

The data has been collected at three different hospitals in London:

- Hammersmith Hospital using a Philips 3T system (details of the scan parameters)
- Guy's Hospital using a Philips 1.5T system (details of the scan parameters)
- Institute of Psychiatry using a GE 1.5T system (details of the scan parameters not available at the moment)

The data has been collected as part of the project:

IXI - Information eXtraction from Images (EPSRC GR/S21533/02)

The images in NIFTI format can be downloaded from here:

- T1 images (all images)
- T2 images (all images)
- PD images (all images)
- MRA images (all images)
- DTI images (all images, bvecs.txt, bvals.txt)
- Demographic information (spreadsheet)

http://www.brain-development.org/

Sobre Bancos de Dados na Neurociência

- Alguns são “federações” de *data sets*
  - Os *data sets* são provenientes de diferentes projetos.
  - Os *data sets* podem possuir (e geralmente possuem!) estruturas de armazenamento diferentes.
  - Os *data sets* podem ter (e geralmente têm!) diferentes níveis de qualidade dos dados.

Problemas de Muitos BDs Abertos

- Ausência de documentação / documentação incompleta
- Dados de má qualidade
  - Inconsistentes
  - Sem informações de proveniência
- Dados não estruturados
- Dados desatualizados

Um banco de dados científico deve ser projetado de modo a servir como um instrumento para a geração de novos conhecimentos, e não apenas para exercer a função de um mero repositório de dados.

Problemas de Muitos BDs Abertos

- Apenas para exemplificar:

COMPUTER APPLICATIONS

Do brain image databanks support understanding of normal ageing brain structure? A systematic review

David Alexander Dickie · Dominic E. Job · Ian Poole · Trevor S. Ahearn · Roger T. Staff · Alison D. Murray · Joanna M. Wardlaw

Received: 25 October 2011 / Revised: 5 December 2011 / Accepted: 29 December 2011 / Published online: 22 February 2012

Questionários Eletrônicos

- São um meio fácil para se “alimentar” bancos de dados
- Padronizam as informações coletadas sobre os experimentos
- Garantem qualidade dos dados coletados
  - Campos obrigatórios
  - Domínio dos dados (tipo, formato e conjunto de valores válidos)

Questionários Eletrônicos

Outros benefícios:

- Software de apoio “poderoso”
  - Geração automática de estatísticas
  - Eficiência e segurança no armazenamento de dados
    - Diferentes perfis de acesso aos dados
  - Facilidade para consultar/filtrar dados

Neste Curso Veremos ...

- Como usar questionários eletrônicos para gerenciar dados de experimentos;
- Quais critérios usar na escolha de um *Sistema de Gerenciamento de Questionários Eletrônicos* que seja apropriado aos propósitos de um contexto de uso específico;
- Como usar sistemas de compartilhamento de arquivos.

Experiência no NeuroMat / INDC - UFRJ

Trabalho de desenvolvimento de um banco de dados para armazenar de forma padronizada e segura o conjunto de dados coletados no

Laboratório de Neurociência e Reabilitação (LabNeR)

do Instituto de Neurologia Deolindo Couto (INDC) da UFRJ, **facilitando o compartilhamento e reuso desses dados**.

Introdução ao Armazenamento de Dados de Experimentos em Neurociência

Parte 1: Estratégias para o armazenamento de dados de experimentos em Neurociência – uma visão geral

Dúvidas?