Open Source + Open Data = Open Science

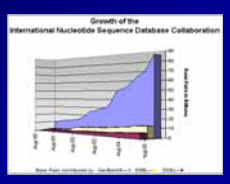
Terry S. Yoo, PhD
OHPCC / LHNCBC / National Library of Medicine
National Institutes of Health

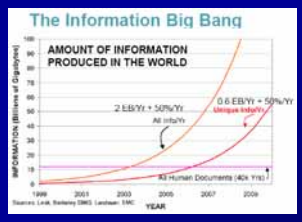
Open Science

What do you do when your customers are drowning in data?
What do you do when your engineers are starving for data?
What do you do when your scientists are reinventing the wheel?

- Infrastructure building.
- Validation, reproducible science.
- Open Source. Open Data.Transparency.

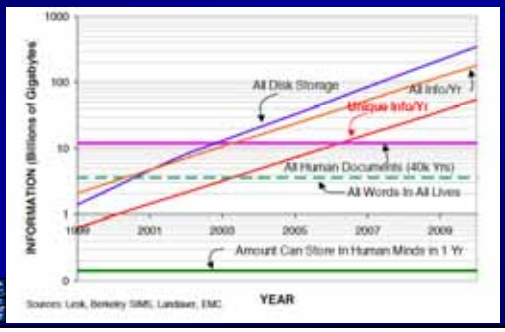
Do you recognize this problem?





* European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-Bank in Hinxton, UK)
* DNA Data Bank of Japan, had been launched at the National Institute of Genetics in Mishima

The Challenge of the Future: Information Big Bang



A Poverty of Attention

"A wealth of information creates a poverty of attention and a need to allocate it efficiently."

-Herbert Simon,

Mathematician
Noble Laureate, Economics

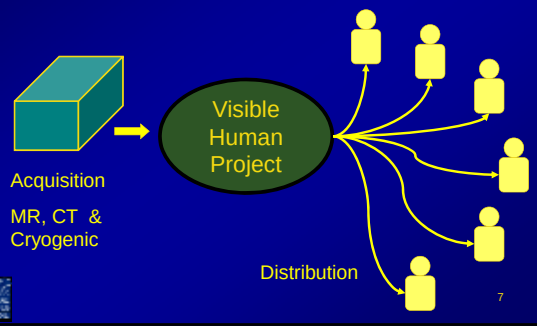
Current State of most Three-D Informatics Research

- Closed source code environment
- Protected proprietary data

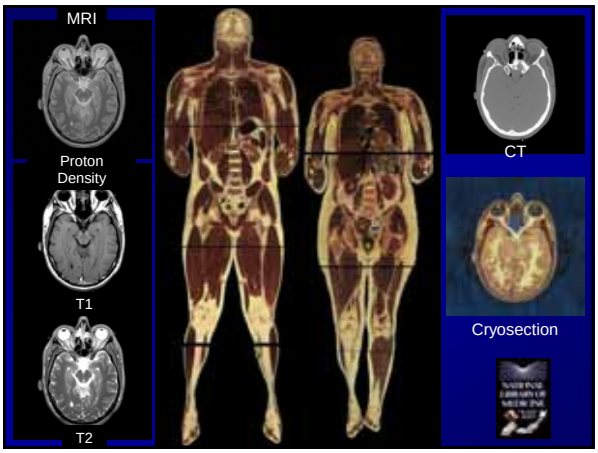
This is a two way street. If you beg, borrow, or steal someone's data, you should be willing to donate your code or data back to the community!

Open Source + Open Data = Open Science

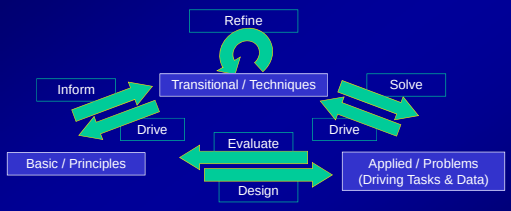
Where did we start ?



7



Research Cycles



- Moving beyond our current development trends.
- Balancing our concentrations of research effort.
- Closing the loop.
- Inherently interdisciplinary.

Open Source Initiatives

- Encourage high-level technical communication
- Provide conventions (vs. standards) for inter-operable software development
- Establish a baseline for improvement
- Opens the field to “beginners”
- Creates common ground for product development
  - example: the creation of HTML enabled Web-based internet development

10

Imaging Tools - “Insight”

The purpose of computing is insight, not numbers - Hamming

- A common platform to encourage communication and dissemination of research results.
- This initiative specifically does not include the development of visualization techniques or a user interface.
- Original release: September 2000.
- Current release: ITK 3.20.
- 45 countries.

11

Open Source

- Source code will be openly distributed.
- Consistent code style and data flow model.
- Regression tests will be included with all software.
- Free licensing for all software.

12

Software Development Program: Promoting Research

- Image Segmentation
  - multivalued (multimodal) data
- Image Registration
  - rigid and deformable registration
- Validation
  - Generation of mathematical models as test data
  - Acquisition of validation datasets from medical scanners

13

ITK Charter Sponsors

NATIONAL LIBRARY OF MEDICINE — THE VISIBLE HUMAN PROJECT®

The National Institute for Dental and Craniofacial Research

The National Science Foundation

NIMH National Institute of Mental Health

The National Institute of Neurological Disorders and Stroke

NIDCD

National Eye Institute

TATRC

NATIONAL CANCER INSTITUTE

14

The Insight Team (Charter Members)

"The purpose of computing is insight, not numbers." - *Hamming*

GE Corporate Research & Development — Kitware — Insightful — PENN — RUTGERS — COLUMBIA — Carnegie Mellon — MAYO CLINIC — Imperial College London — THE UNIVERSITY OF UTAH — COGNITICA — Georgetown UNIVERSITY

The Technology of the Insight Toolkit

Segmentation

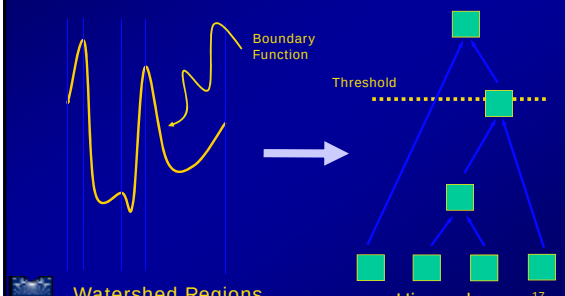
- Statistical, Fuzzy Logic, Markov Random Fields, Mixture Modeling, Parzen Windows, Nearest Neighbor, K-Means, …
- Level Set, Finite Element, Region Growing, Hybrid, Watershed, Connected Components, Parameterized Models, …

Registration

- Rigid, Similarity, Affine, Vector Field, Hierarchical, Quaternion, Versor, Parameterized Deformation, Euler, Perspective, ND/MD, …
- Mutual Information, Normalized Correlation, Demons, Mean Squared, Landmark, …

16

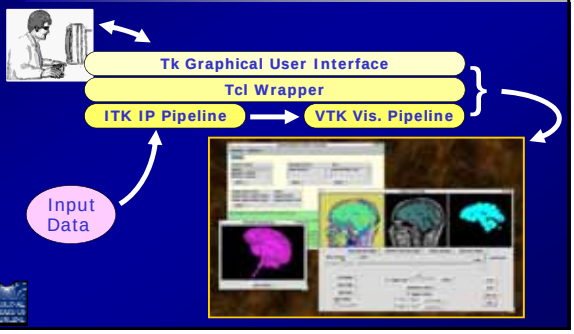
Watershed Algorithm



17

Watershed Segmentation Results

Watershed Segmentation User Interaction

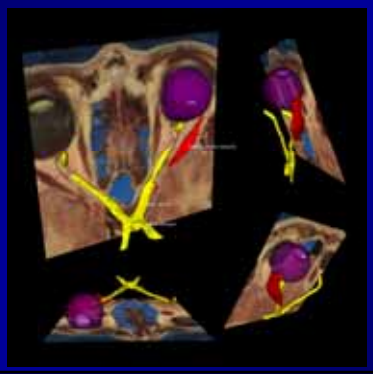


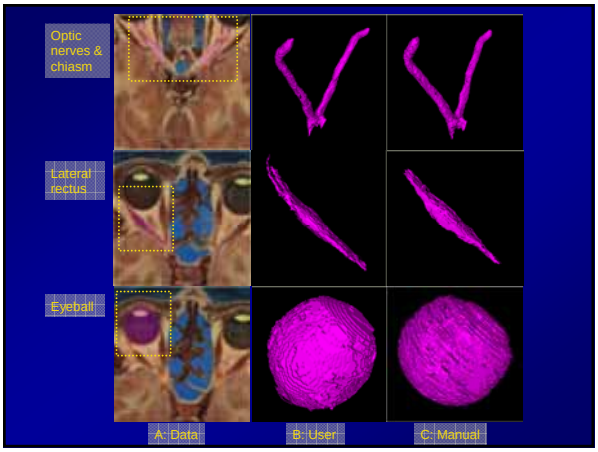
Watershed Segmentation

Interactive results (Visible Human Project female head)

20

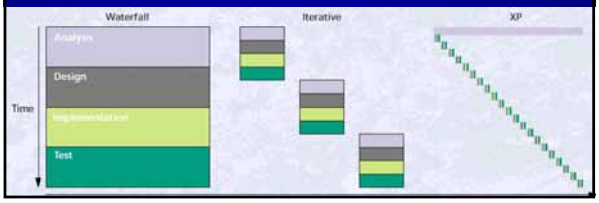
Segmentation Results





Extreme Programming

Compresses the standard analyze, design, implement, test cycle into a continuous process



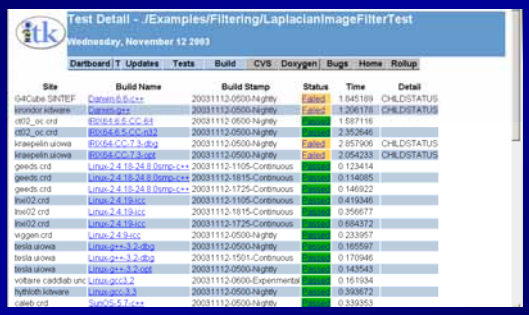
Insight - Open Source Products

Extreme Programming Daily Testing Is The Key

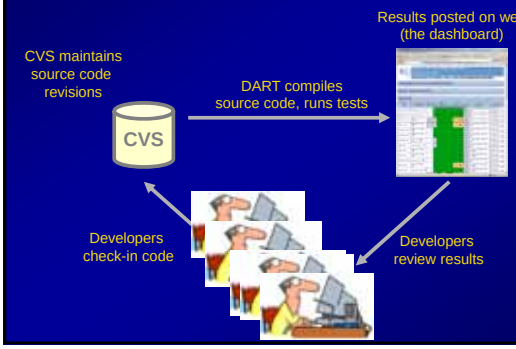
- Testing anchors and drives the development process (Dart)
- Opens up the development process to everyone
- Developers monitor the testing dashboard constantly
- Problems are identified and fixed immediately
- Developers receive e-mail if they "Break the Build"

- Norway
- The Netherlands
- France
- Sweden
- Belgium
- Spain
- Switzerland
- Germany
- China
- India

Multi-Platform Builds



How CDash Enables Collaboration



Achievements

- International collaboration of broad research interests and software developers.
- Approximately 4 companies, 7 medical schools, 5 engineering schools.
- Multi-agency $18.5M Initiative
  - $2.5M / year - 3 years.
  - $3.26M - year 4.

Achievements (continued)

- Comparable to other Centers and Cooperative Agreements.
- Shared software - deliverables.
- Open Source - extends beyond the consortium; a living contribution to research community.
- Consortium planned for growth.
- An affirmation of our ideals came in 2004.

Success

- IGSTK.
- NA-MIC.
- OSIRIX
- Mayo Clinic - Analyze
- MITK - Heidelburg, SCIRun, MeVis, VolView
- Users: Pfizer, Allen Brain Institute
- Linux K Development Environment
- Mayo BIR Collection

NA-MIC

The National Alliance for Medical Image Computing

- Committed to open source.
- Shared vision for brain research through imaging.
- Fast start, building from existing tools.
- Likely to accomplish the goal of making lives better.

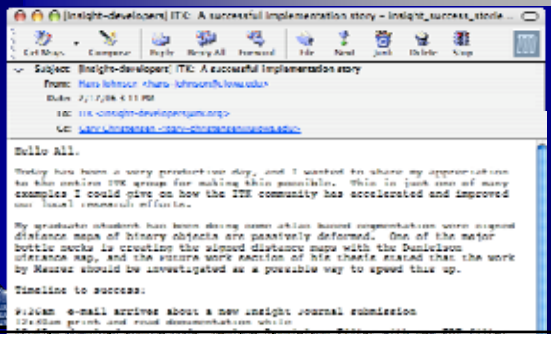
Isomics, Inc.

IGSTK

Image-Guided Surgery Toolkit

- Fast start, building from existing tools.
- STTR funded by NIBIB/NIH (Georgetown-Kitware) Grant 2R42EB000374-02
- The CADDLab at UNC joined the project and contributed SpatialObjects and the RF Ablation application
- 2005: Atamai Inc. joined the project and contributed the Tracker code
- March 2007: IGSTK 2.0.1

Success (continued)
The Insight Journal



Changes in NIH Culture

- 1987 - NHLBI/Mayo Analyze. NIH Image.
  - Monolithic, proprietary projects
- 1996 - Visible Human Project - open data
- 1999 - ITK: The Insight Toolkit - open source
- 2001 - NCI Lung Image Database Consortium
- 2003 - NCI CAD Open Source Software
- 2003 - NIH Roadmap
- 2004 - NIH BECON/BISTI Symposium
  - Biomedical Informatics for Clinical Decision Support
- 2004 - NIH makes four (4) NCBC awards

Principles from the First International Strategy Meeting on Human Genome Sequencing
(Bermuda, 25-28 February 1996)

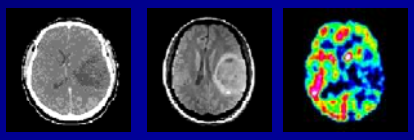
- Primary Genomic Sequence Should be in the Public Domain.
- Primary Genomic Sequence Should be Rapidly Released.
  - Sequence assemblies should be released as soon as possible; in some centres, assemblies of greater than 1 Kb would be released automatically on a daily basis.
  - Finished annotated sequence should be submitted immediately to the public databases.

Current NHGRI Policy for Release and Database Deposition of Sequence Data
(March 7, 1997)

At the Second International Strategy Meeting on Human Genome Sequencing (Bermuda, 1997), attendees affirmed the principle that was set out at the First (1996) International Strategy meeting, that primary genomic sequence should be rapidly released. …. NHGRI has determined, therefore, that its grantees engaged in large-scale genomic DNA sequencing should now be automatically releasing sequence assemblies of 2 kb or larger within 24 hours of their generation. … Any laboratory funded by NHGRI for large-scale human genomic sequencing must develop and submit to NHGRI a plan to implement such a data release program, which must be implemented within one month of its being approved by NHGRI. No non-competing or competing renewal will be funded until an acceptable plan has been approved. Mandatory data release as described above will be made a condition of the award for any grant funded by NHGRI for large-scale human sequencing.

Last Reviewed: August 2006

Retrospective Image Registration Evaluation Project (1995 – present)
J. Michael Fitzpatrick, P.I.



- The SAT test for rigid, non-deformable multimodal brain registration.
  - Separate training and test data. Provide tests under careful control.
  - Only scores are returned… no solutions.
- Funded across NIH over several installments.
  - Originally NINDS, then NCI, then NIBIB.

The 2010 Insight Team
"The purpose of computing is insight, not numbers." - Hamming

Carnegie Mellon
Georgetown University
PENN
MAYO CLINIC
THE UNIVERSITY OF UTAH

ITK Version 4 - Sponsors

NLM NATIONAL LIBRARY OF MEDICINE
THE VISIBLE HUMAN PROJECT®

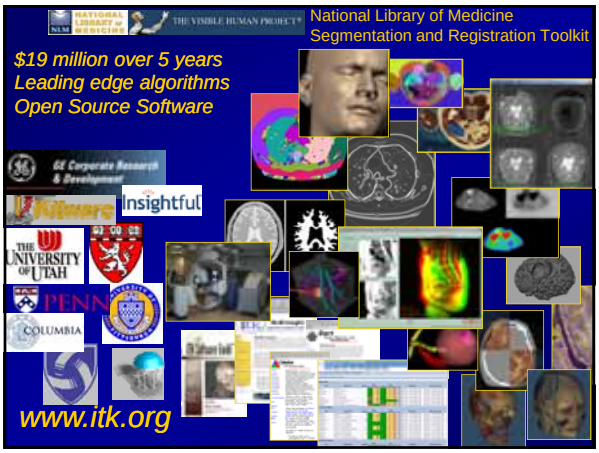
QuickTime™ and a decompressor are needed to see this picture.

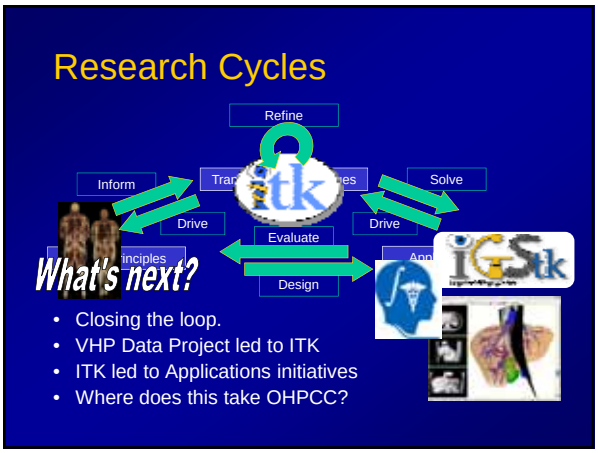
ITK-v4 and ITK-A2D2-2010

UPGRADE - v4
TEST and EVALUATE - A2D2
- Requested to sequential
- Grow the community
- Lower entry bar for new users
- seek new domains

Test - the architecture and the API

We know there's redundancy…
We're here to work it out.





Open Science

What do you do when your customers are drowning in data?
What do you do when your engineers are starving for data?
What do you do when your scientists are reinventing the wheel?

- Infrastructure building.
- Validation, reproducible science.
- Open Source. Open Data.Transparency.

45