Virtual Earthquake and seismology Research Community e-science environment in Europe
Project 283543 – FP7-INFRASTRUCTURES-2011-2 – www.verce.eu – info@verce.eu

# Welcome and Introduction

Malcolm Atkinson

Data-Intensive Research Group
University of Edinburgh

OSDC Workshop, UvA-CWI, Amsterdam, 9 June 2015

# Outline

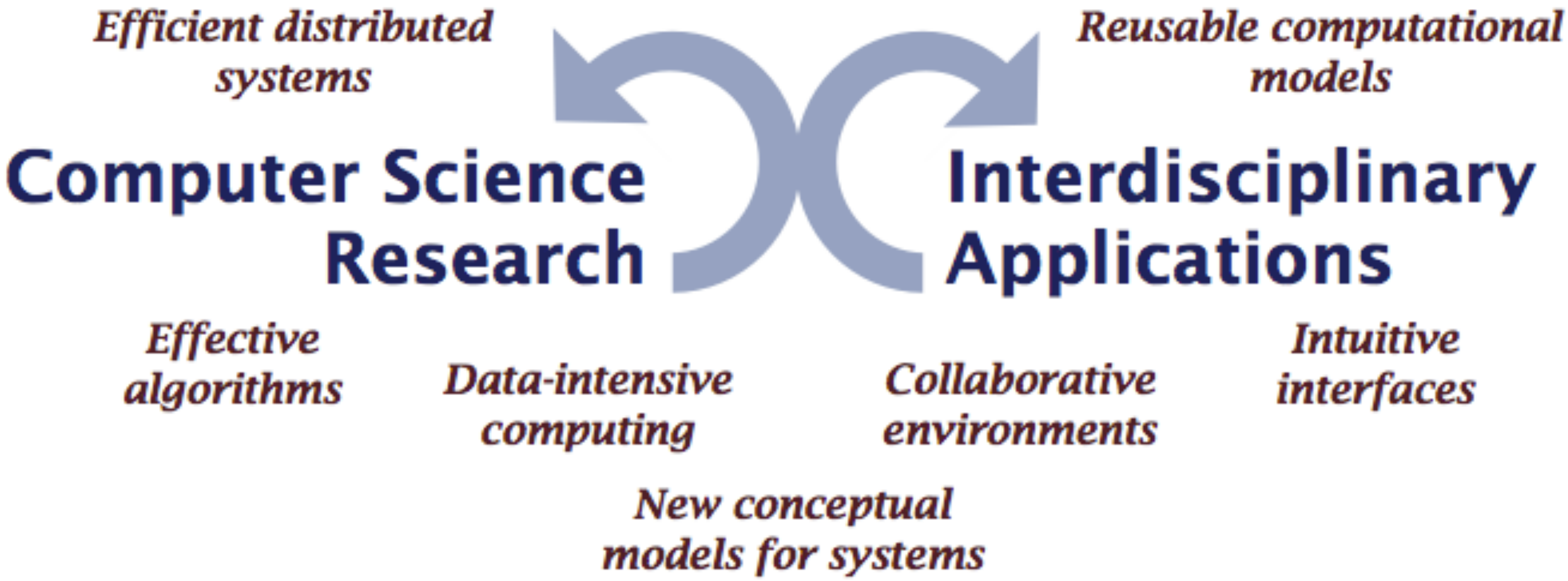
- Data-Intensive thinking
- Projects / alliances
- Data-Intensive methods
  - Principles
  - Strategy
  - Implementation

Cornish Coast Path

# Data-Intensive Thinking

# Data-Intensive Thinking

# Two rapidly changing worlds



*Efficient distributed systems*

*Reusable computational models*

**Computer Science Research**

**Interdisciplinary Applications**

*Effective algorithms*

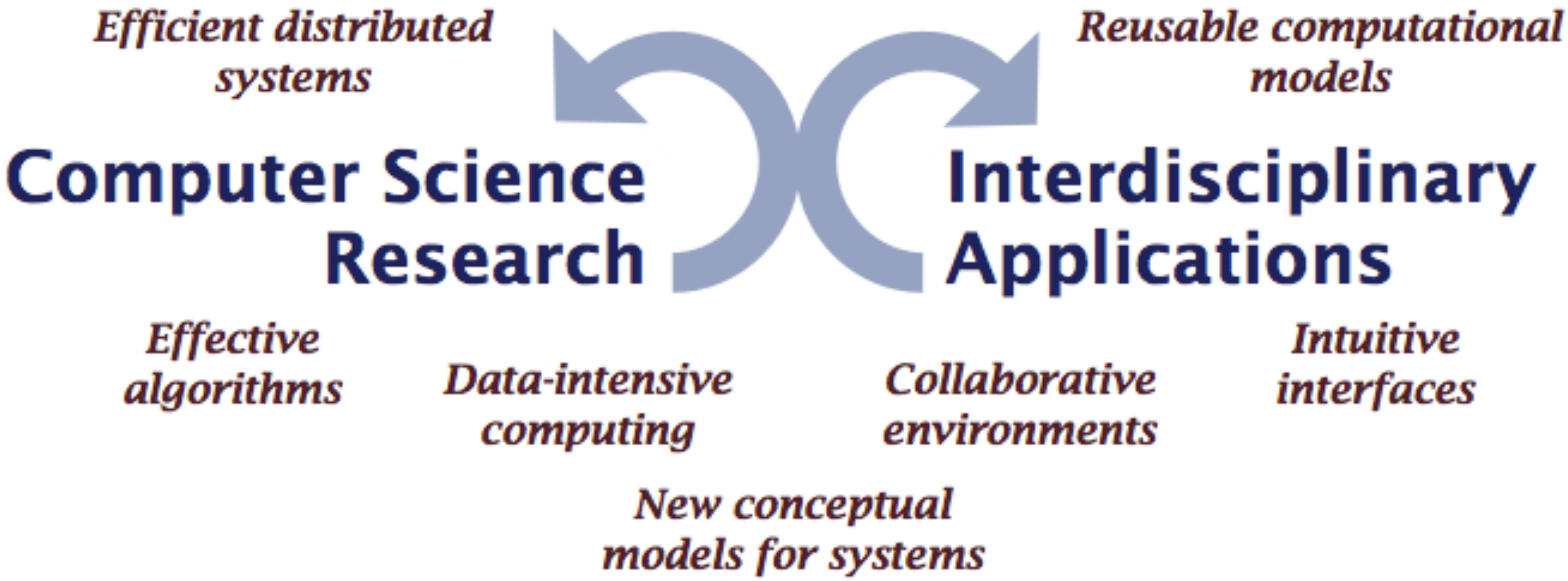
*Data-intensive computing*

*Collaborative environments*

*Intuitive interfaces*

*New conceptual models for systems*

# Two rapidly changing worlds



Efficient distributed systems

Reusable computational models

Computer Science Research

Interdisciplinary Applications

Effective algorithms

Data-intensive computing

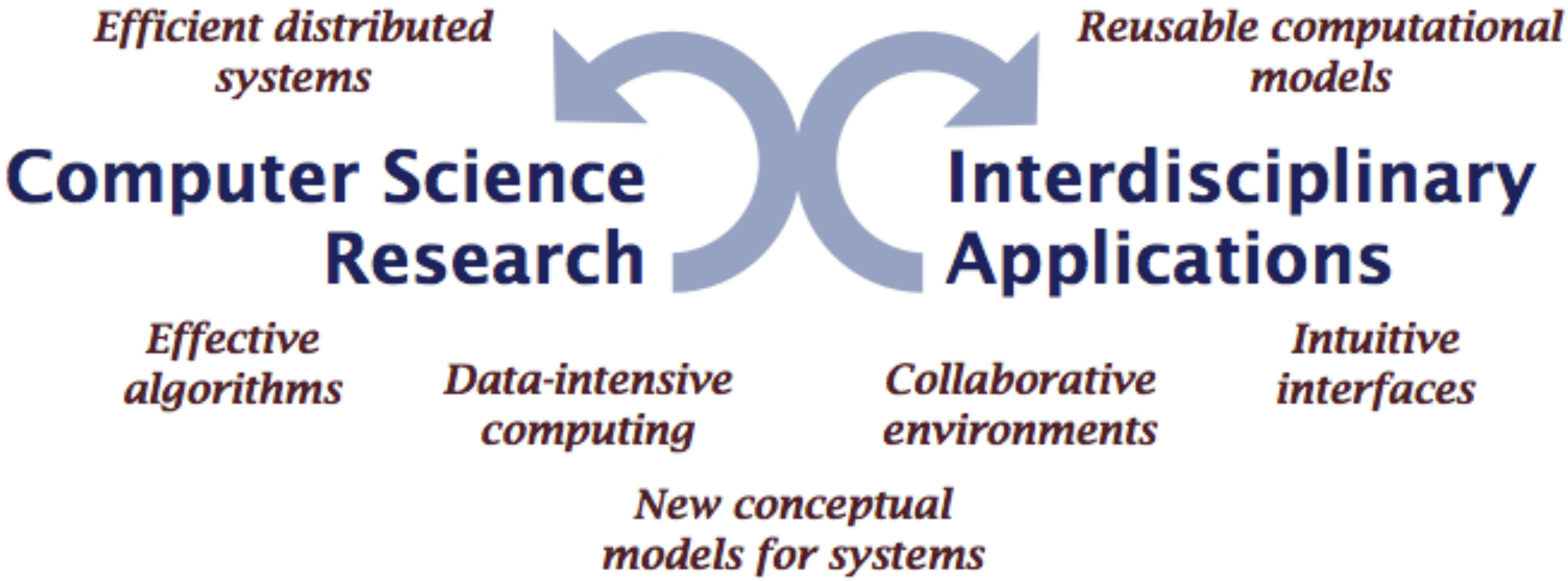
Collaborative environments

Intuitive interfaces

New conceptual models for systems

**Research is motivated by change and enables change**

# Two rapidly changing worlds



*Efficient distributed systems*

*Reusable computational models*

**Computer Science Research**

**Interdisciplinary Applications**

*Effective algorithms*

*Data-intensive computing*

*Collaborative environments*

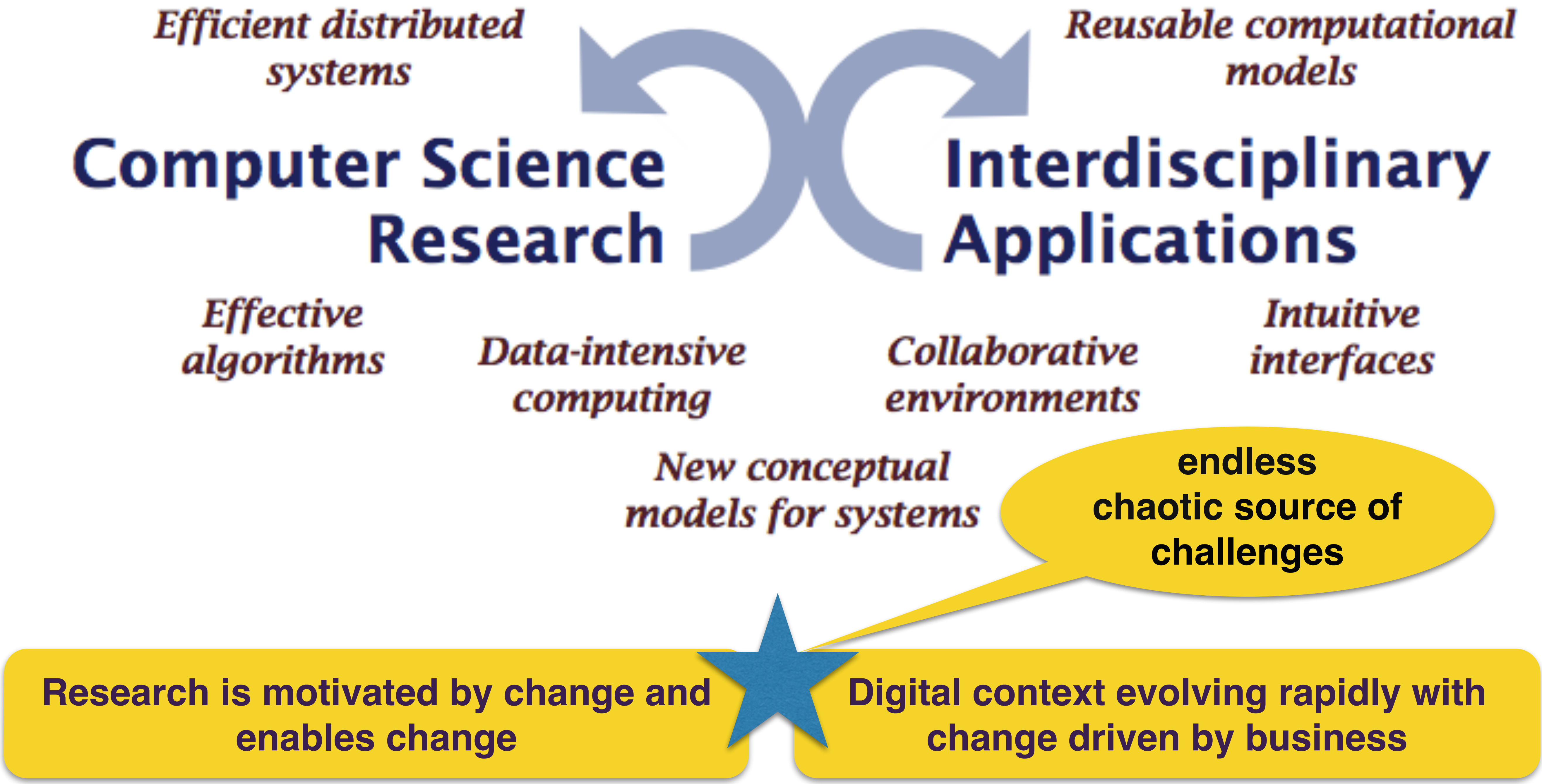
*Intuitive interfaces*

*New conceptual models for systems*

**Research is motivated by change and enables change**

**Digital context evolving rapidly with change driven by business**

# Two rapidly changing worlds



*Efficient distributed systems*

*Reusable computational models*

**Computer Science Research**

**Interdisciplinary Applications**

*Effective algorithms*

*Data-intensive computing*

*Collaborative environments*

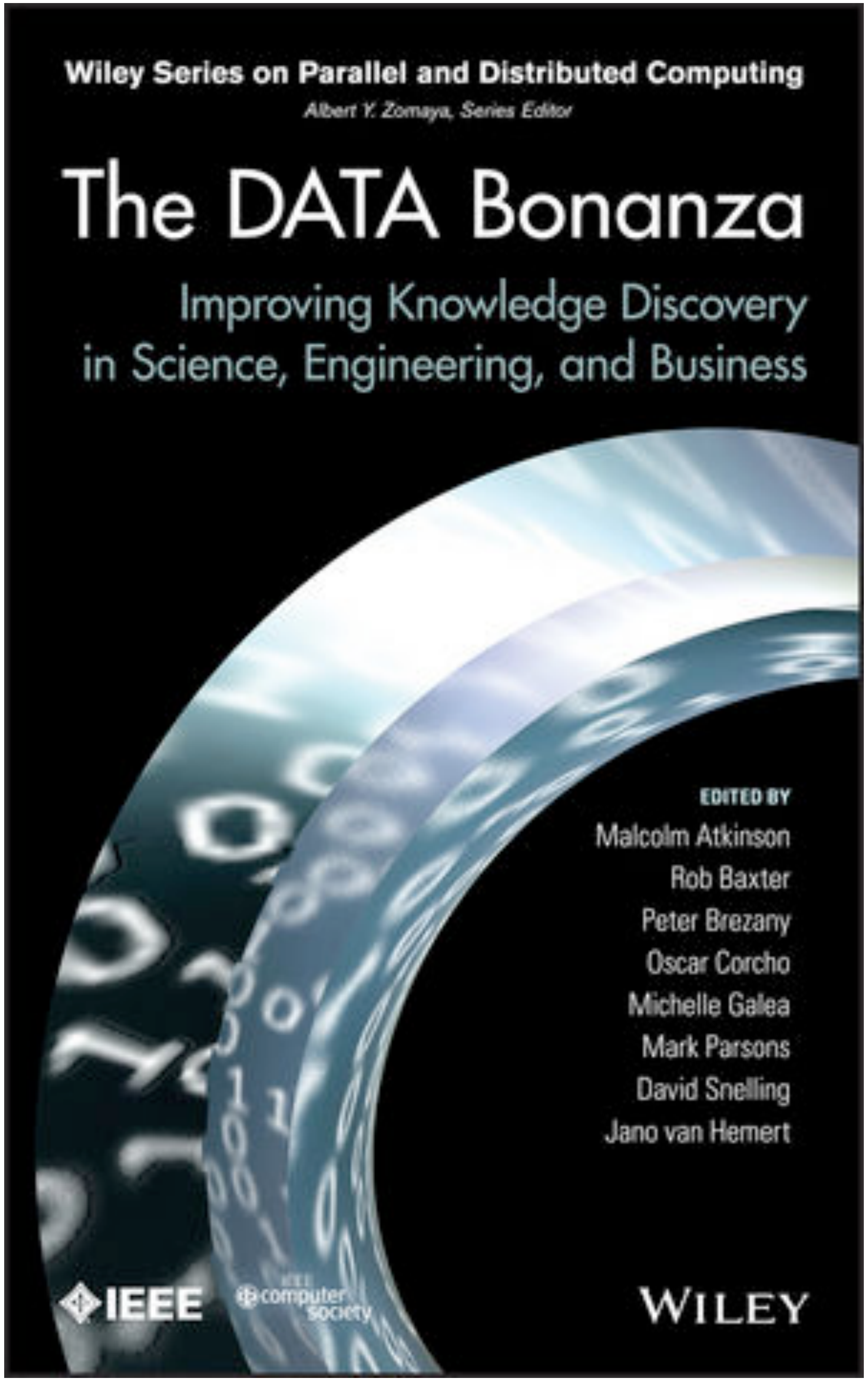
*Intuitive interfaces*

*New conceptual models for systems*

**endless chaotic source of challenges**

**Research is motivated by change and enables change**

**Digital context evolving rapidly with change driven by business**



Wiley Series on Parallel and Distributed Computing

Albert Y. Zomaya, Series Editor

The DATA Bonanza

Improving Knowledge Discovery in Science, Engineering, and Business

EDITED BY

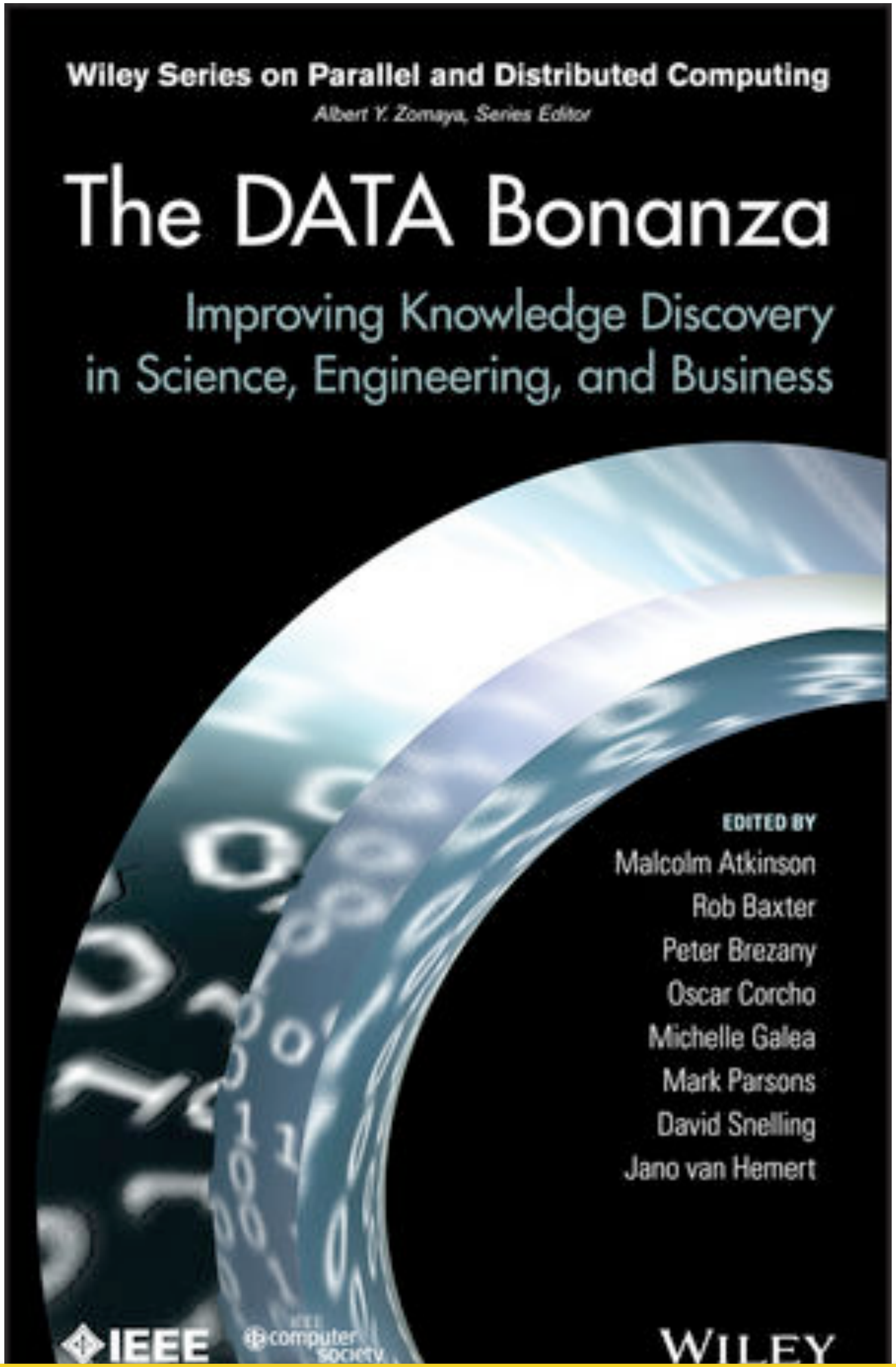
Malcolm Atkinson
Rob Baxter
Peter Brezany
Oscar Corcho
Michelle Galea
Mark Parsons
David Snelling
Jano van Hemert

IEEE computer society

WILEY

# Admire Project

- Model for Data Driven
  - science & research
  - engineering
  - business
- Abstraction
  - technical detail
- Longevity
  - as digital context evolves



Wiley Series on Parallel and Distributed Computing

Albert Y. Zomaya, Series Editor

The DATA Bonanza

Improving Knowledge Discovery in Science, Engineering, and Business

EDITED BY

Malcolm Atkinson
Rob Baxter
Peter Brezany
Oscar Corcho
Michelle Galea
Mark Parsons
David Snelling
Jano van Hemert

IEEE computer society

WILEY

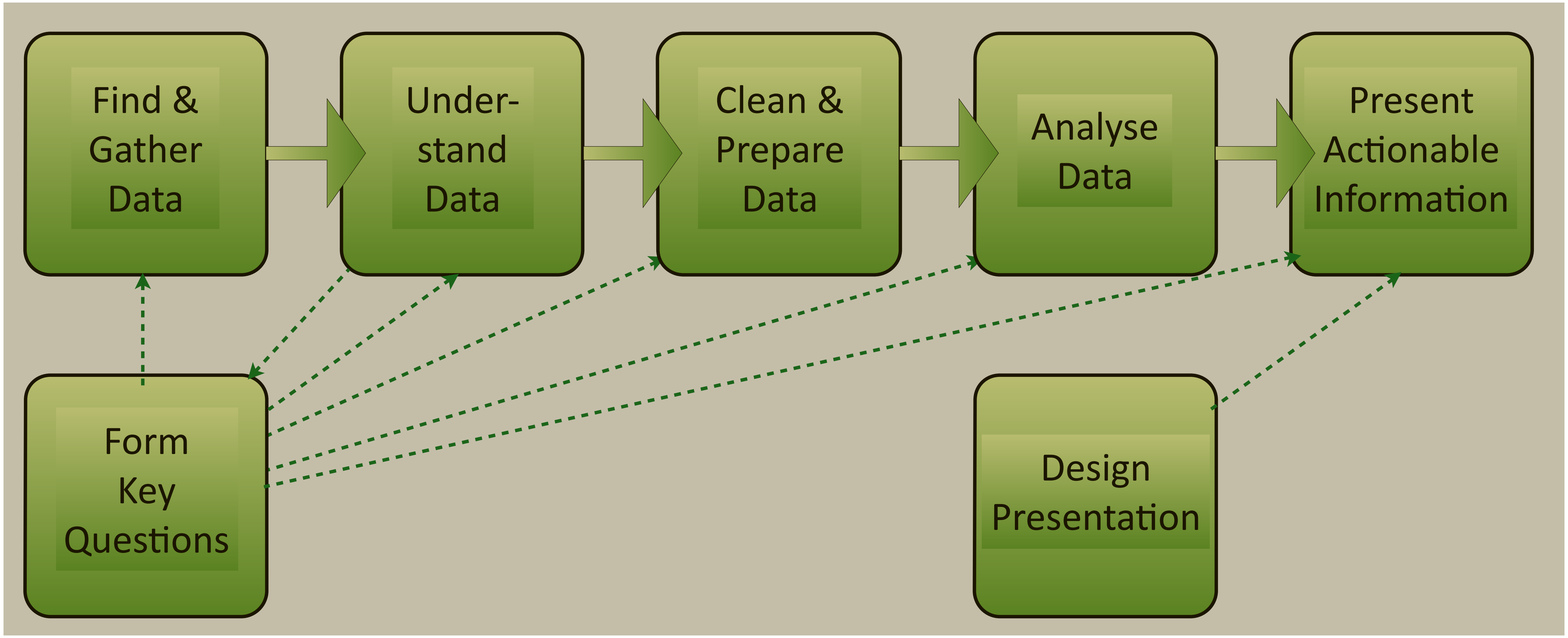
**Free** http://onlinelibrary.wiley.com/book/10.1002/9781118540343

# Admire Project

- Model for Data Driven
  - science & research
  - engineering
  - business
- Abstraction
  - technical detail
- Longevity
  - as digital context evolves

# *Three* Groups of Experts

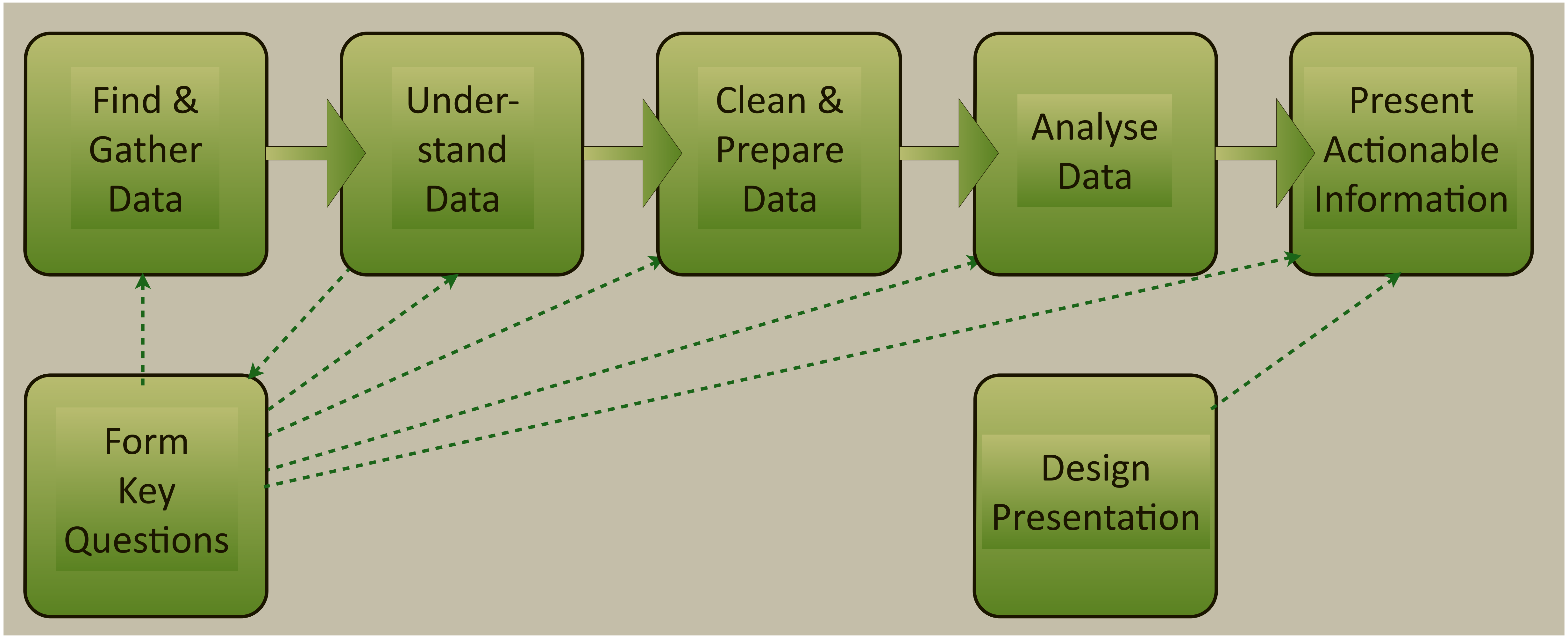
- **Domain expert**
- **Data-analysis experts**
- **Data-intensive engineers**



# *Three* Groups of Experts

- **Domain expert**
- **Data-analysis experts**
- **Data-intensive engineers**

**} Working together**



**Jim Gray's** legacy

- The Fourth Paradigm

# Domain Experts

**Jim Gray's** valediction
*"May all your problems be technical ones"*

- **Individuals**
  - > 90% doing day job delivering services & building the evidence base
  - <10% innovating: setting new goals & creating new methods
  - Big variation in ITC knowledge
  - different subdomains & different targets / changing
  - in groups, in projects, in organisations
    - cooperating, competing / allying & pulling in different directions
  - in organisational, in national & global cultures and communities
  - strongly held preferences for computer interaction
  - **Key primary issues**
    - *Formulating & refining* scientific methods - **Empower the scientists to do this themselves**
    - *Integrating* stages from *different* specialities - **Compose methods without understanding detail**
    - Drawing on packaged techniques from other viewpoints - **Well-defined boundaries and semantics**
    - *Demonstrable correctness* a **HUGE** challenge
    - *Sustained value* as the digital context evolves another **HUGE** challenge

**Jim Gray's** legacy

- The Fourth Paradigm

# Domain Experts

**Jim Gray's** valediction

*"May all your problems be technical ones"*

- **Individuals**
  - > 90% doing day job delivering services & building the evidence base
  - <10% innovating: setting new goals & creating new methods
  - Big variation in ITC knowledge
  - different subdomains & different targets / changing
  - in groups, in projects, in organisations
    - cooperating, competing / allying & pulling in different directions
  - in organisational, in national & global cultures and communities
  - strongly held preferences for computer interaction
  - **Key primary issues**
    - *Formulating & refining* scientific methods - **Empower the scientists to do this themselves**
    - *Integrating* stages from *different* specialities - **Compose methods without understanding detail**
    - Drawing on packaged techniques from other viewpoints - **Well-defined boundaries and semantics**
    - *Demonstrable correctness* a **HUGE** challenge
    - *Sustained value* as the digital context evolves another **HUGE** challenge

**abstraction**

**Jim Gray's** legacy

- The Fourth Paradigm

# Data-Analysis Experts

**Jim Gray's** valediction
*“May all your problems be technical ones”*

- **Individuals**
  - sub-specialists from mathematics and statistics to application-specific data-analysis
  - trade-offs between data/computational cost and reliability and certainty
  - favourite problem-solving environments
  - different subdomains & different targets / changing
  - in groups, in projects, in organisations
    - cooperating, competing / allying & pulling in different directions
  - in organisational, in national & global cultures and communities
  - strongly held preferences for computer interaction
  - **Key primary issues**
    - **Correctness** proven / tested ; clarity about scope of applicability / safety
    - **Usability** how easily can the domain specialists grasp how to use a technique
    - **Support** how much effort is there to sustain the technique and help get it used appropriately
    - **Credit and blame** how do we attribute these fairly
    - **Sustainability** dependencies and eInfrastructure independence
    - **Relationship** with data-intensive engineering

**Jim Gray's** legacy

- The Fourth Paradigm

# Data-Intensive Engineering

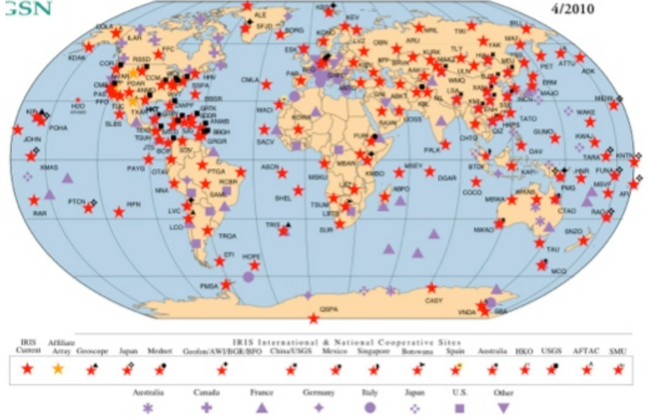
**Jim Gray's** valediction
*“May all your problems be technical ones”*

- **Individuals**
  - sub-specialists: data storage, data transport, data bases, data curation, …, computation, software & hardware architectures,…, requirements capture, …, human-machine interaction, …
  - software communities, language communities, development models, …
  - from demon coders to formalisation experts
  - in groups, in projects, in organisations
    - cooperating, competing / allying & pulling in different directions
  - in organisational, in national & global cultures and communities
  - strongly held preferences for interacting with computational systems
  - **Key primary issues**
    - **mapping** to existing and changing distributed computing platforms
    - **exploiting** systems, architectures and components near optimally
    - **Less energy** consumption
    - **Sustainability**, how long can the investment survive?
    - **Correctness** in the presence of diverse users and diverse infrastructures
    - **Support** enabling users of all kinds and colleagues to use what they build

# Common issues

- Diversity
- Composability
- Longevity
- Correctness
- Scalability
- Extensibility
- Avoiding change >90% + Innovators <10%
- Individuals, groups, organisations, projects, communities

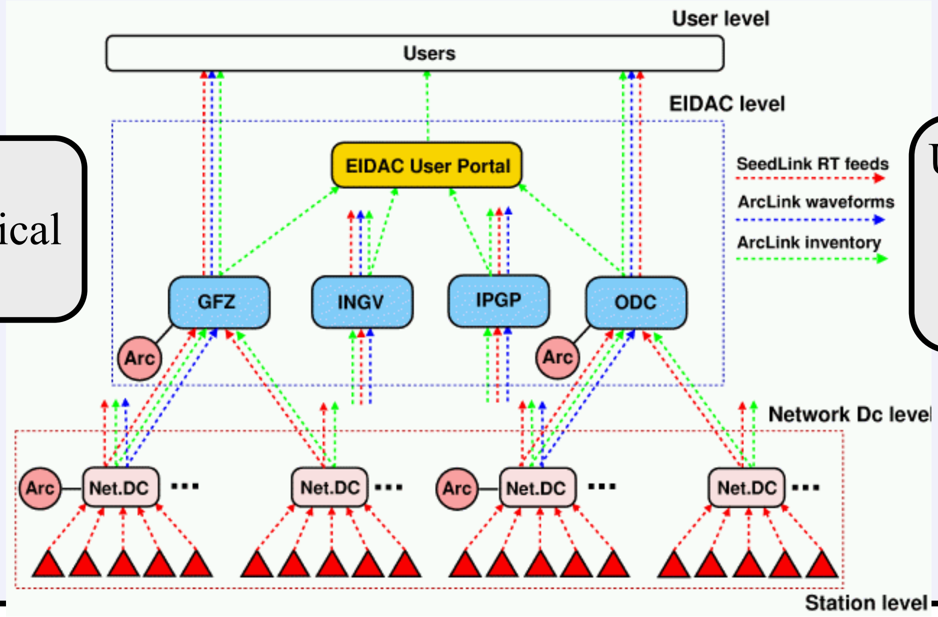
# Projects & Alliances


FDSN Global array

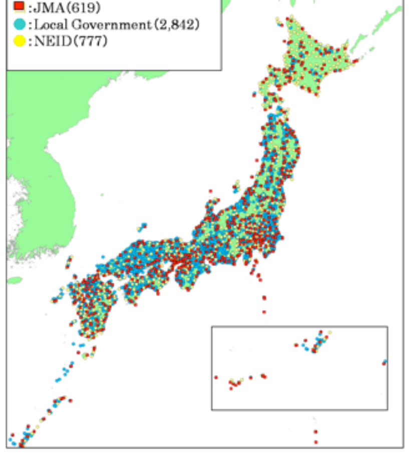
European array

## EUROPEAN INTEGRATED DATA ARCHIVES (EIDA)

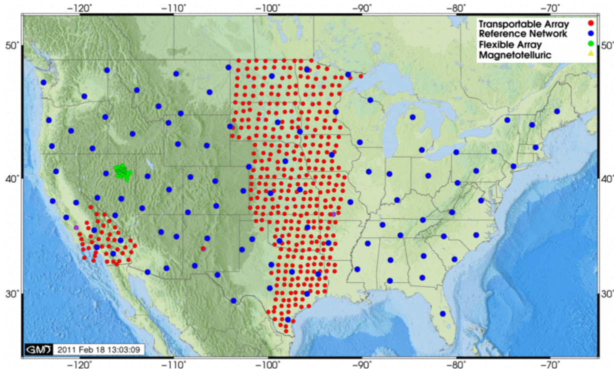
Japan Meteorological Agency



US Incorporated Research Institutions in Seismology



Japan array



US array

# Data Intensive Research

Visualization
Data analysis / Data mining
Simulation, inversion, HR imaging

# VERCE

*e-Science environment for data intensive research based on an extensive service-oriented architecture*

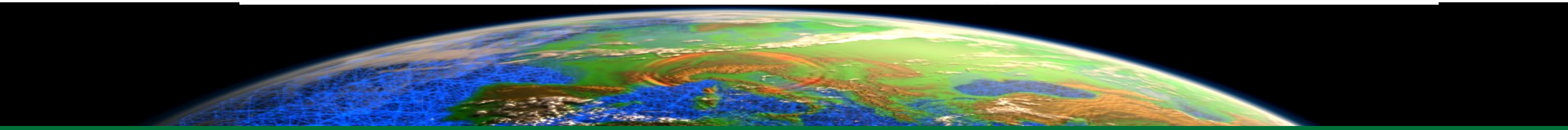
# HPC/GRID Infrastructures

Cloud

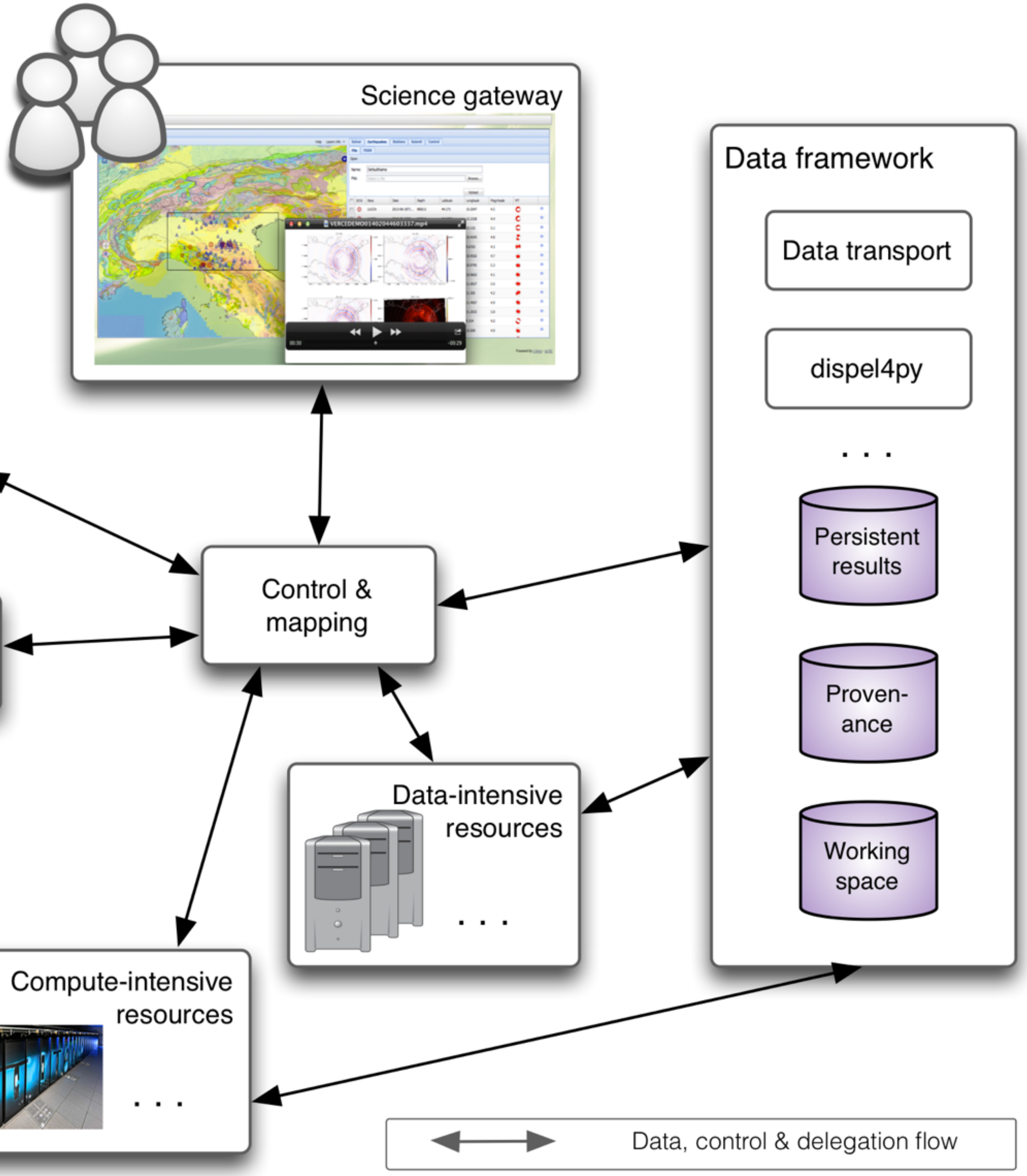
**Earth's interior imaging and dynamics: noise correlation, waveform analysis**

**Natural hazards: new tools for monitoring earthquakes, volcanoes, and tsunami**

**Interaction of solid Earth with Ocean and Atmosphere: environment, climate changes**



# VERCE architecture



Science gateway

Data framework

Data transport

dispel4py

. . .

Persistent results

Proven-ance

Working space

Domain services

Control & mapping

Generic services

Data-intensive resources

. . .

Simula-tion code

Compute-intensive resources

. . .

Data, control & delegation flow

*Virtual Earthquake and Seismology Research Community in Europe*

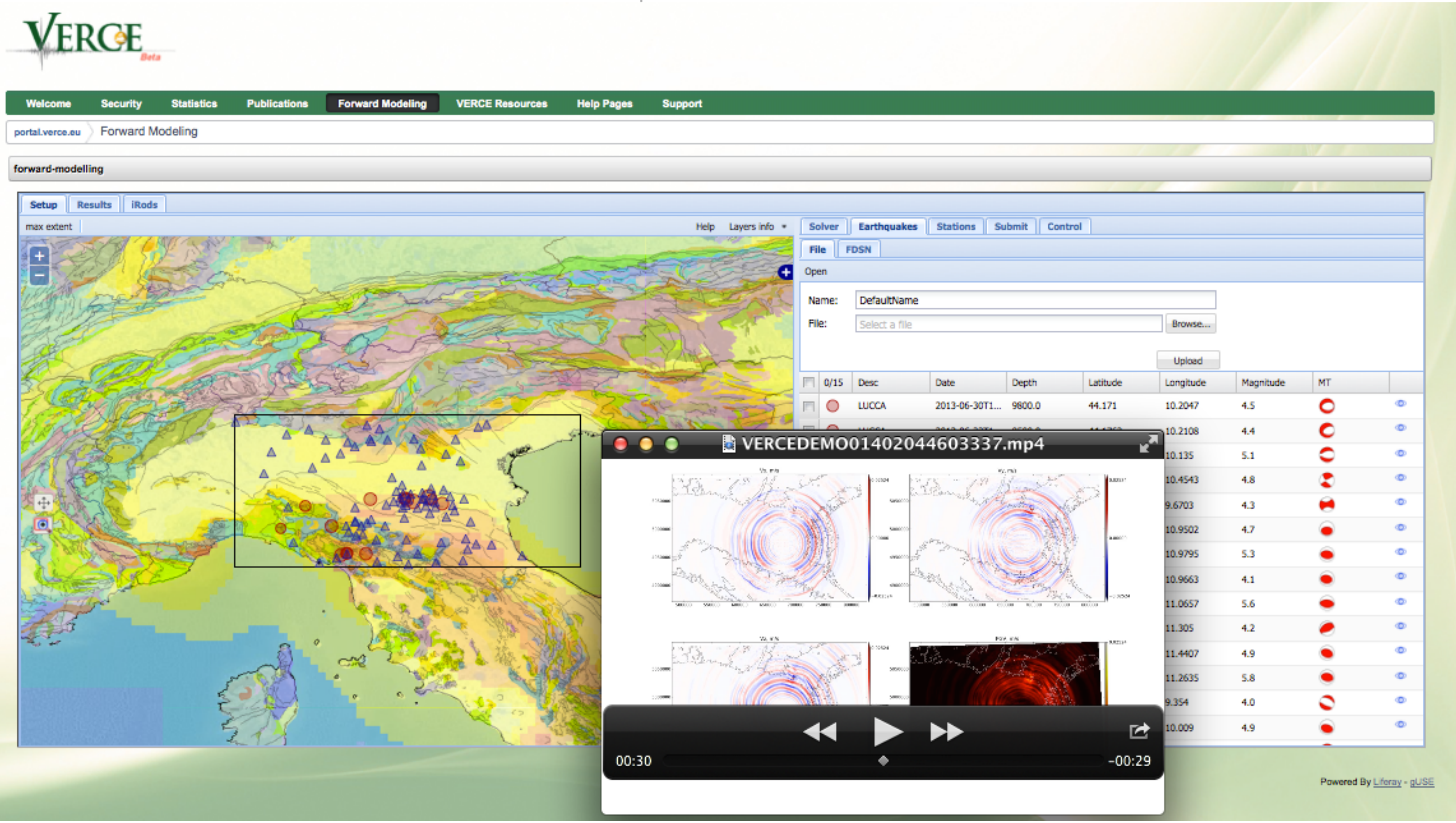
**Virtual Environment** for of **Earthquakes Simulations** and evaluation of **Earth Models**

**http://portal.verce.eu**

Combined access to **computing infrastructures** (EGI, PRACE, Local Clusters), for development and execution of large **HPC** computations

Access and use of **European data archives** and services adopting International standards (FDSN, GCMT, OneGeology, EFEHR, QuakeML)

**Adoption of Workflow Technologies, Data Management** and **Provenance System**



Name: DefaultName

File: Select a file

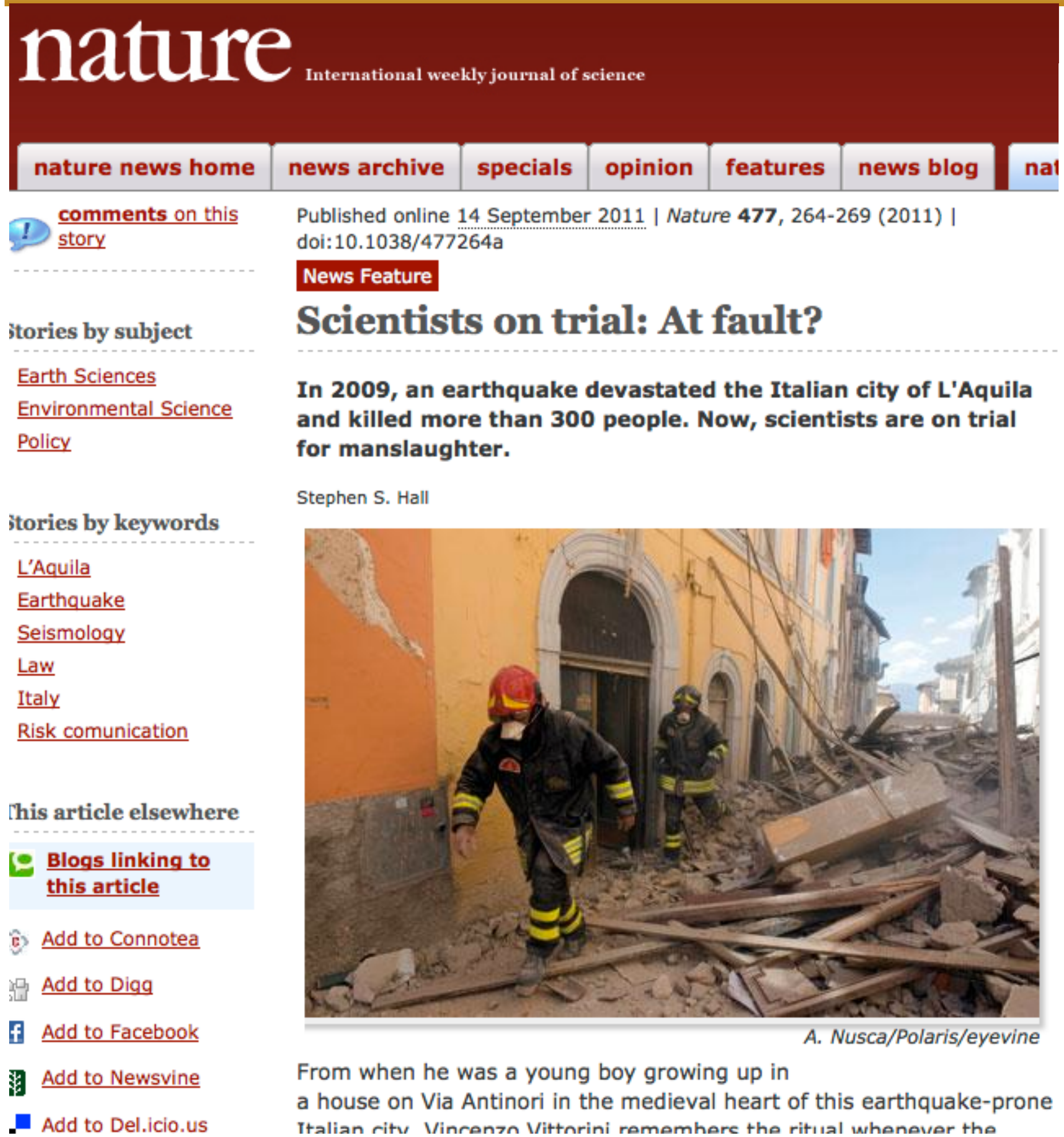
| □ | 0/15 | Desc | Date | Depth | Latitude | Longitude | Magnitude | MT |
|---|---|---|---|---|---|---|---|---|
| □ | ● | LUCCA | 2013-06-30T1... | 9800.0 | 44.171 | 10.2047 | 4.5 | ● |
| | | | | | | 10.2108 | 4.4 | ● |
| | | | | | | 10.135 | 5.1 | ● |
| | | | | | | 10.4543 | 4.8 | ● |
| | | | | | | 9.6703 | 4.3 | ● |
| | | | | | | 10.9502 | 4.7 | ● |
| | | | | | | 10.9795 | 5.3 | ● |
| | | | | | | 10.9663 | 4.1 | ● |
| | | | | | | 11.0657 | 5.6 | ● |
| | | | | | | 11.305 | 4.2 | ● |
| | | | | | | 11.4407 | 4.9 | ● |
| | | | | | | 11.2635 | 5.8 | ● |
| | | | | | | 9.354 | 4.0 | ● |
| | | | | | | 10.009 | 4.9 | ● |

VERCE Training

# Human issues!

http://www.nature.com/news/2011/110914/full/477264a.html

Nature 14 Sept. 2011



nature International weekly journal of science

comments on this story

Stories by subject

Earth Sciences
Environmental Science
Policy

Stories by keywords

L'Aquila
Earthquake
Seismology
Law
Italy
Risk comunication

This article elsewhere

Blogs linking to this article

Add to Connotea
Add to Digg
Add to Facebook
Add to Newsvine
Add to Del.icio.us

Published online 14 September 2011 | *Nature* **477**, 264-269 (2011) | doi:10.1038/477264a

News Feature

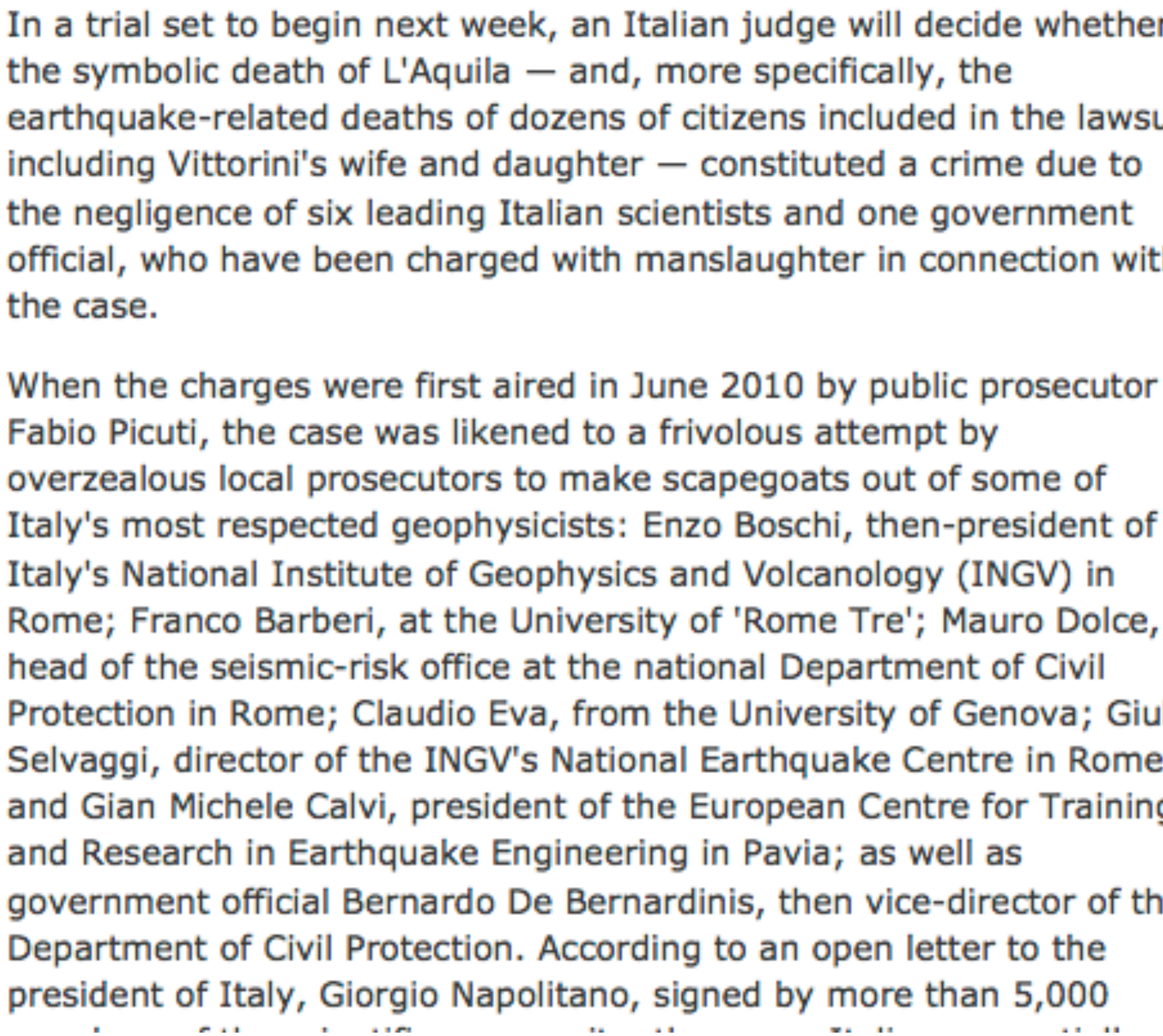
## Scientists on trial: At fault?

**In 2009, an earthquake devastated the Italian city of L'Aquila and killed more than 300 people. Now, scientists are on trial for manslaughter.**

Stephen S. Hall

A. Nusca/Polaris/eyevine

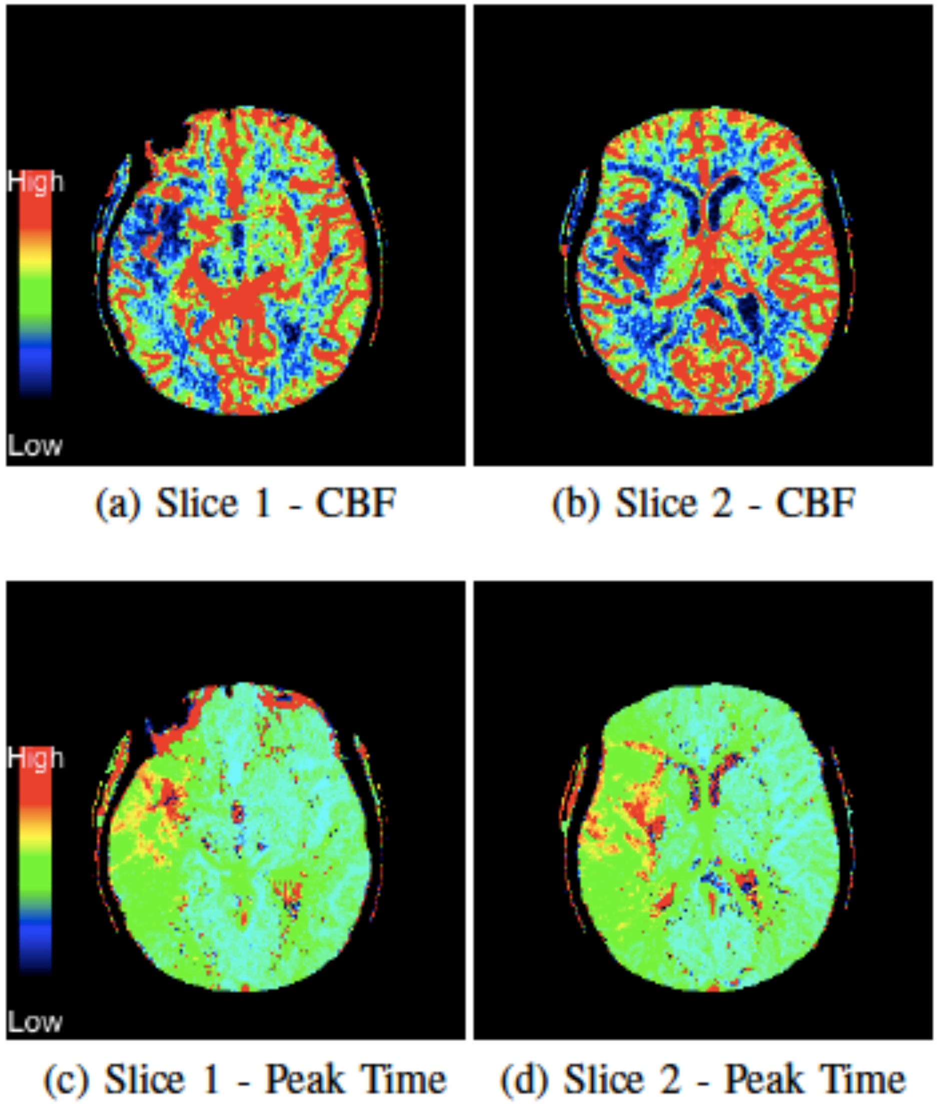
From when he was a young boy growing up in a house on Via Antinori in the medieval heart of this earthquake-prone Italian city, Vincenzo Vittorini remembers the ritual whenever the



In a trial set to begin next week, an Italian judge will decide whether the symbolic death of L'Aquila — and, more specifically, the earthquake-related deaths of dozens of citizens included in the lawsu including Vittorini's wife and daughter — constituted a crime due to the negligence of six leading Italian scientists and one government official, who have been charged with manslaughter in connection wit the case.

When the charges were first aired in June 2010 by public prosecutor Fabio Picuti, the case was likened to a frivolous attempt by overzealous local prosecutors to make scapegoats out of some of Italy's most respected geophysicists: Enzo Boschi, then-president of Italy's National Institute of Geophysics and Volcanology (INGV) in Rome; Franco Barberi, at the University of 'Rome Tre'; Mauro Dolce, head of the seismic-risk office at the national Department of Civil Protection in Rome; Claudio Eva, from the University of Genova; Giu Selvaggi, director of the INGV's National Earthquake Centre in Rome and Gian Michele Calvi, president of the European Centre for Trainin and Research in Earthquake Engineering in Pavia; as well as government official Bernardo De Bernardinis, then vice-director of th Department of Civil Protection. According to an open letter to the president of Italy, Giorgio Napolitano, signed by more than 5,000

# Brain image processing

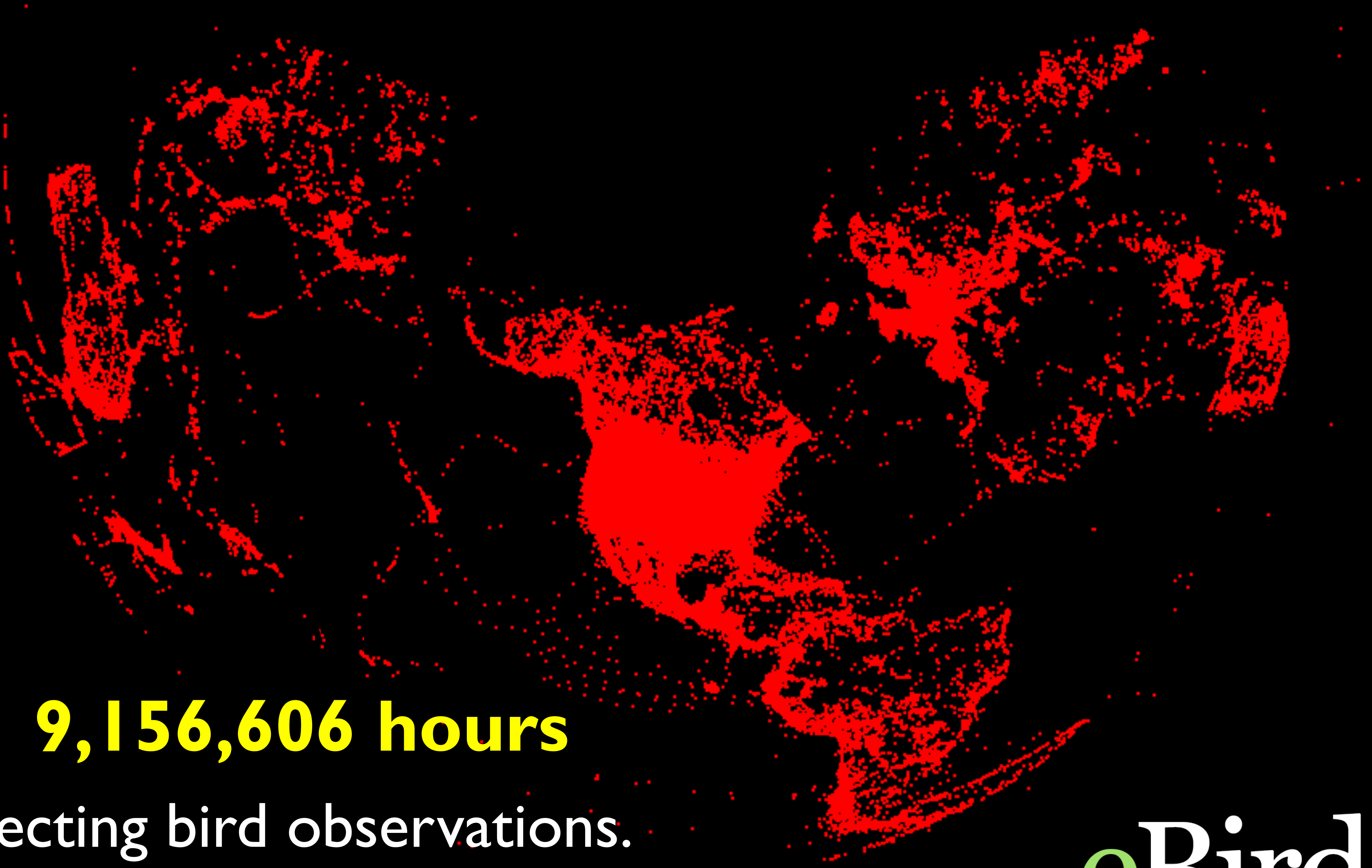


(a) Slice 1 - CBF (b) Slice 2 - CBF

(c) Slice 1 - Peak Time (d) Slice 2 - Peak Time

Lesion Area Detection Using Source Image Correlation Coefficient for CT Perfusion Imaging

Fan Zhu David Rodriguez Gonzalez Trevor Carpenter Malcolm Atkinson Joanna Wardlaw

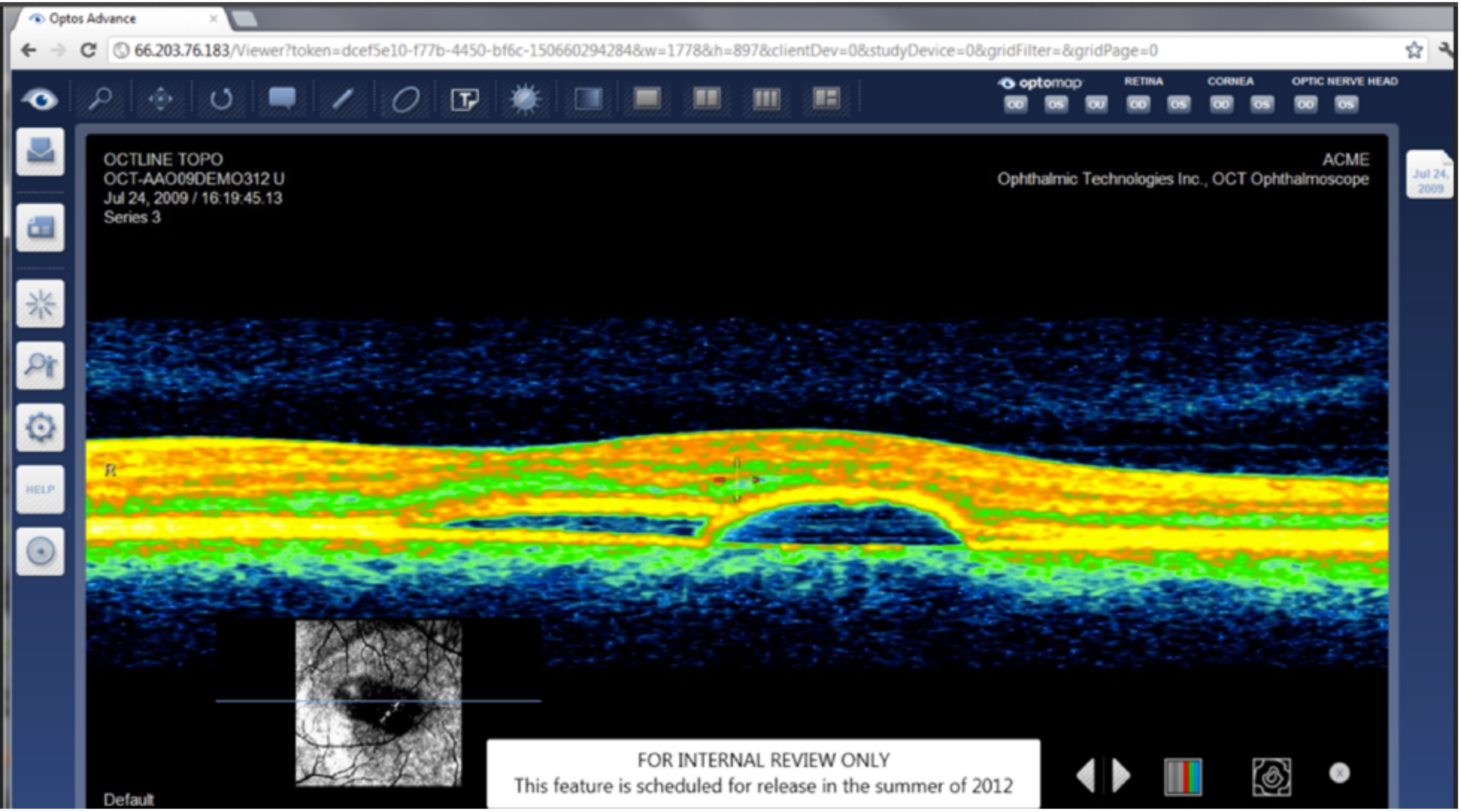


**9,156,606 hours**

collecting bird observations.

eBird

# Optos retinopathy diagnosis



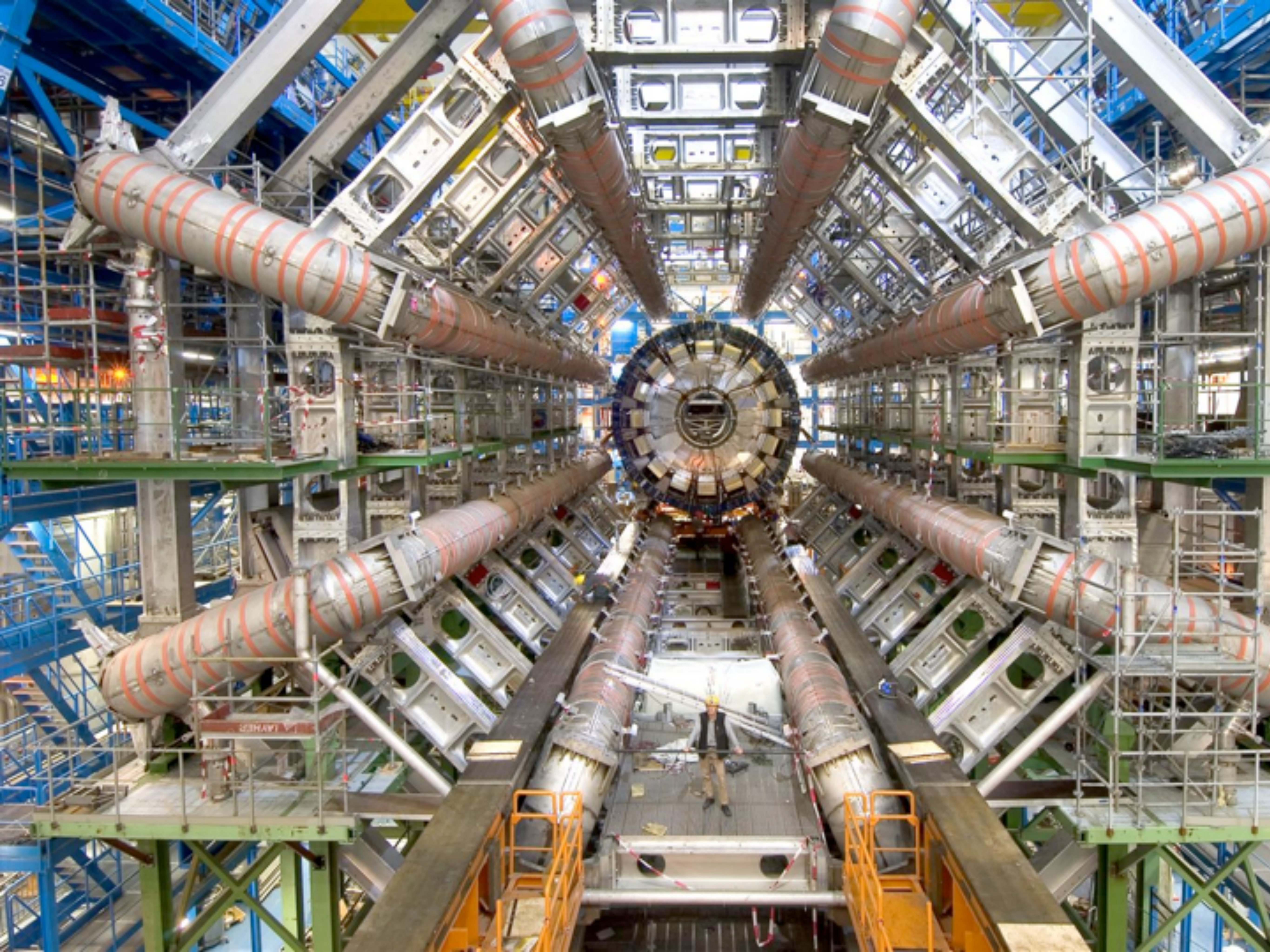
Optos Advance

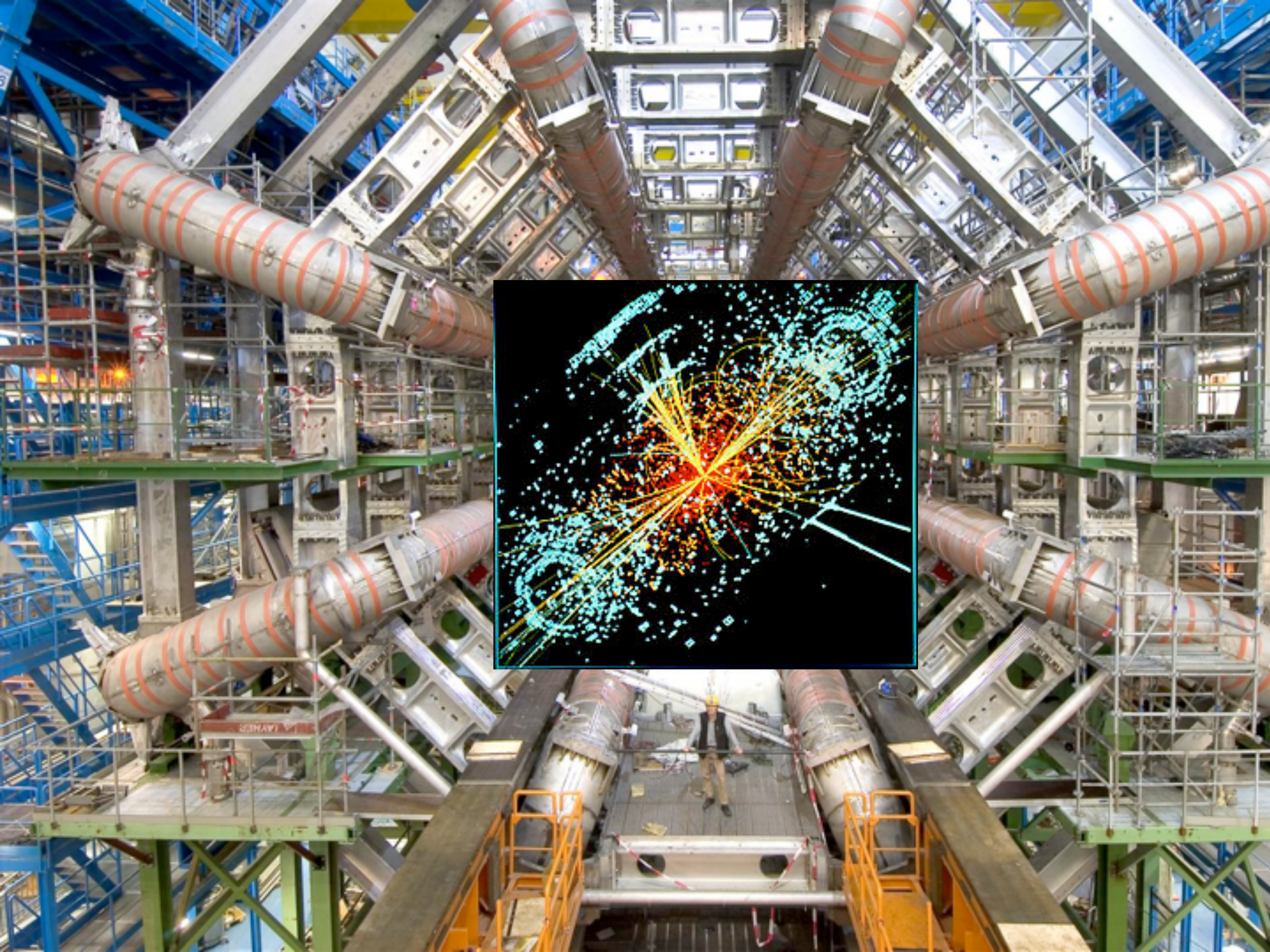
66.203.76.183/Viewer?token=dcef5e10-f77b-4450-bf6c-150660294284&w=1778&h=897&clientDev=0&studyDevice=0&gridFilter=&gridPage=0

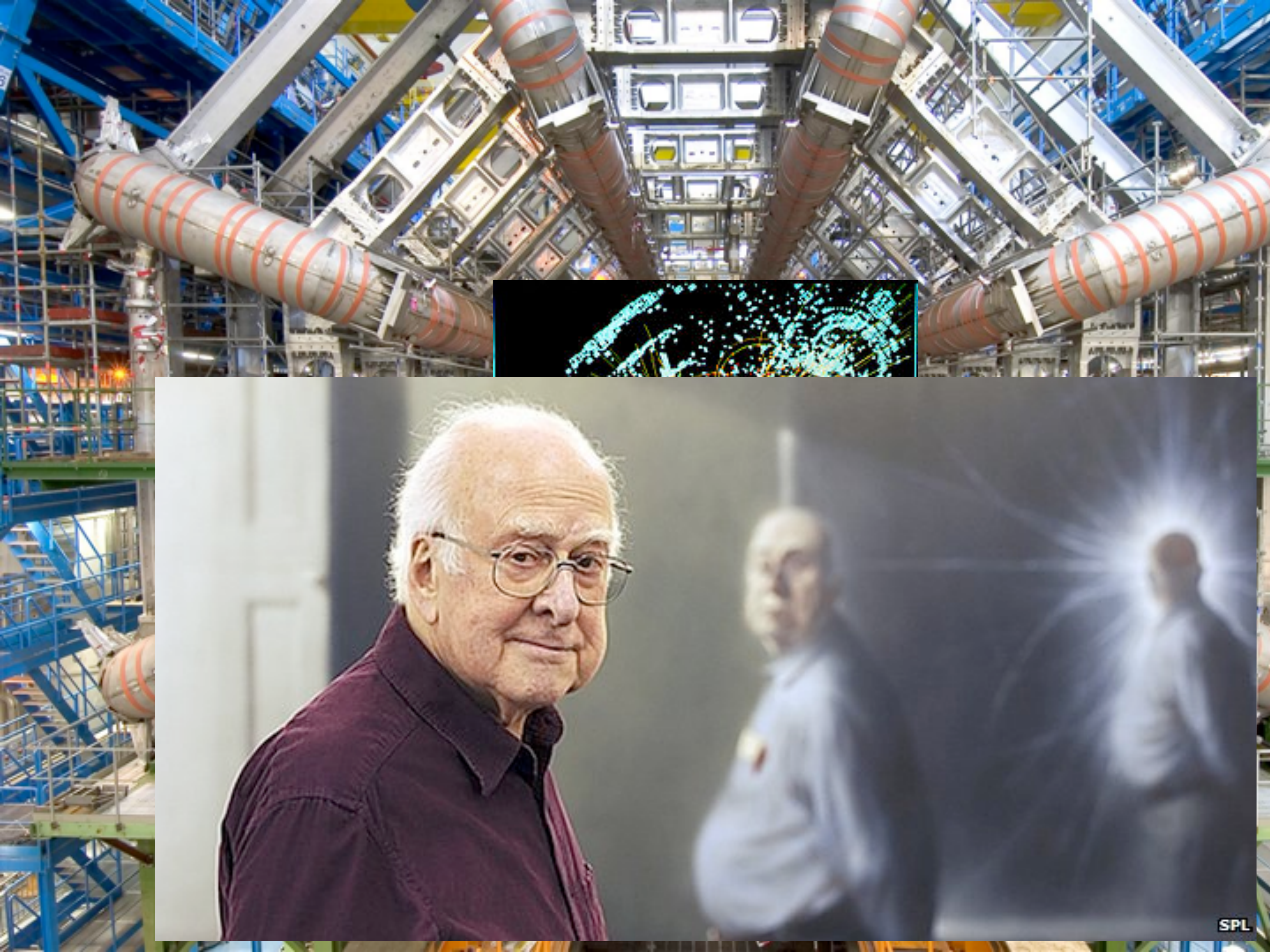
OCTLINE TOPO
OCT-AAO09DEMO312 U
Jul 24, 2009 / 16:19:45.13
Series 3

ACME
Ophthalmic Technologies Inc., OCT Ophthalmoscope

FOR INTERNAL REVIEW ONLY
This feature is scheduled for release in the summer of 2012







SPL

# Data is the catalyst

# Data is the catalyst

- Data as the messenger
  - coupling people via systems to people
  - coupling systems with systems
  - coupling organisations with organisations
  - joining processes, software and algorithms
  - output from creative work

# Data is the catalyst

- Data as the messenger
  - coupling people via systems to people
  - coupling systems with systems
  - coupling organisations with organis
  - joining processes, software and thms
  - output from creative work
- Data as a source
  - from specialised
  - from scattere ctions
  - from citi
  - as b cts of “data as a messenger”

**We are *already* dependent on data, its quality, movement & curation**

# Workflows as a DI strategy

## What is a workflow?

A **composition** of steps

to make a data-handling + data-analysis+ simulation journey

Many ways of forming steps

Require good libraries of ready made steps

Learn to add your own

Many ways of combining steps

Running in many computing environments

Recursive — a journey can be a step in another journey

# **Why use a workflow**?

Empower the *domain* experts

**Rapid prototyping and experiment**
**Saving labour and repeated drudgery**
**Reducing error rates**
Saving you from doing your own housekeeping
  Returning resources such as file space
  Gathering all your results
**Acceleration due to workflow optimisation**,
  e.g. parallelisation
**Sharing** & getting credit for methods
**Incrementally improving** methods
**Combining methods** developed by different experts

## There are many workflow languages - why invent **dispel**?

**Raising the level of discourse**

- Removing much technology specific information - technology changes
- Relieving users from concerns about optimisation

**Improving the logical description**

- Streams of data with auto-iteration over data units
- Multiple streams in & multiple streams out
- Behaviour, data interpretation & data representation

**Covering existing models**

- Distributed query
  - Optimisation based on avoiding IO & characterising operators
- Real-time processing
- Task-based batch processing

**Achieving scalability**

## **What is dispel4py good for**?

That is what you will learn today

Embedding Dispel in **Python** combines their strengths

Everything ….

but investment in libraries is needed for each new topic

plus common libraries for shared activities, such as data handling

Everything ….

but the dispel4py engineering team need to

make it perform at the scales you need

make it excel on the DCIs you use

- laptop to cloud via supercomputers & clusters

make it reliable

So I will hand you over to Rosa’s tender mercies

# Summary and Conclusions

# Exploiting the DATA Bonanza

# Exploiting the DATA Bonanza

- ***Educate*** to use data
  - The ***three* categories of expert**
  - Data literate managers, governmental officials & ...
  - A data savvy public

# Exploiting the DATA Bonanza

- ***Educate*** to use data
  - The ***three* categories of expert**
  - Data literate managers, governmental officials & ...
  - A data savvy public
- Long-term development of leadership
  - curated data
  - expert teams

# Exploiting the DATA Bonanza

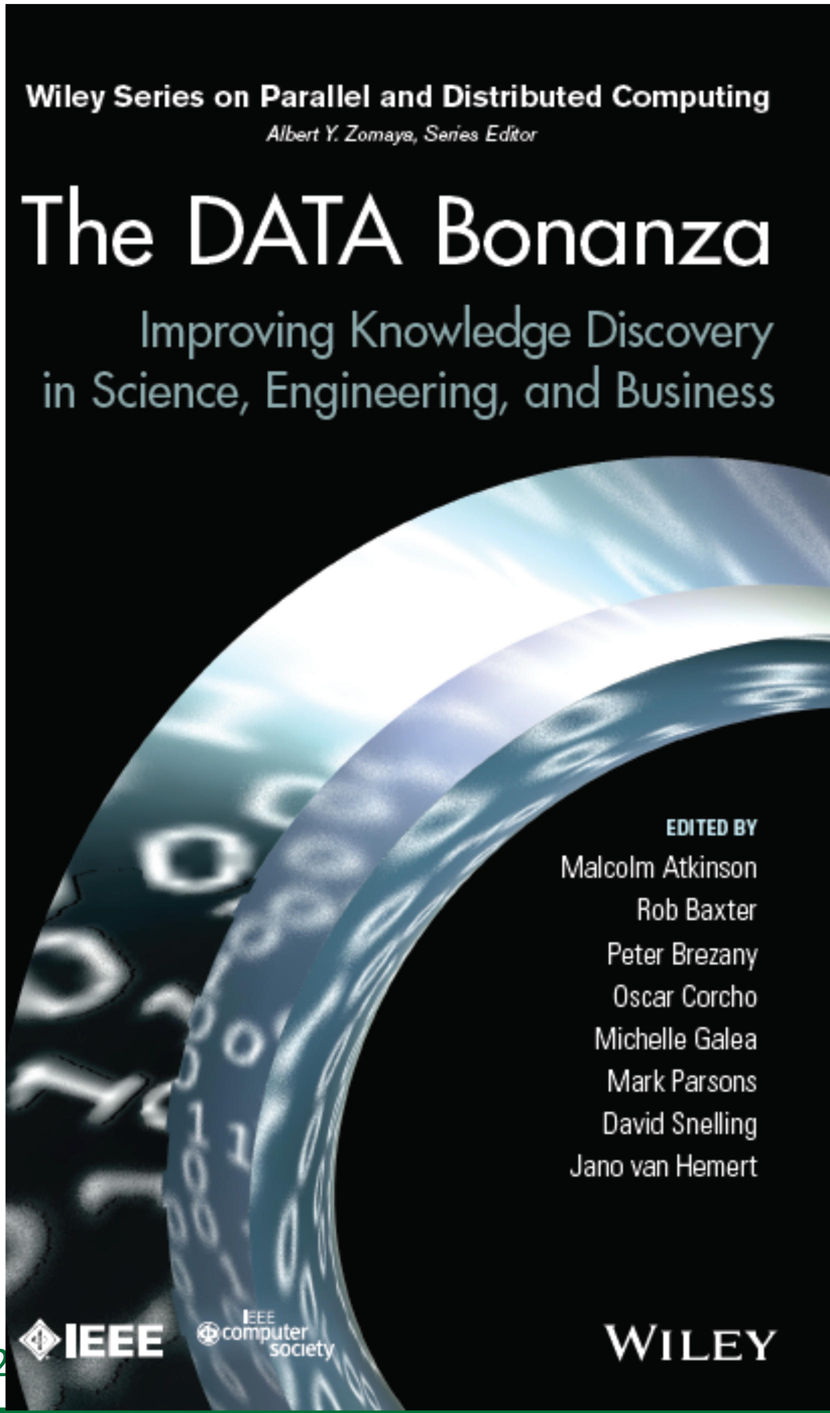
- ***Educate*** to use data
  - The ***three* categories of expert**
  - Data literate managers, governmental officials & ...
  - A data savvy public
- Long-term development of leadership
  - curated data
  - expert teams
- Balanced investment
  - from collection to *“final mile” of* information delivery

# Exploiting the DATA Bonanza

- ***Educate*** to use data
  - The ***three* categories of expert**
  - Data literate managers, governmental officials & ...
  - A data savvy public
- Long-term development of leadership
  - curated data
  - expert teams
- Balanced investment
  - from collection to *"final mile" of* information delivery
- Open data and processes
  - encouraging scrutiny, challenge & contribution

# Exploiting the DATA Bonanza

- ***Educate*** to use data
  - The ***three* categories of expert**
  - Data literate managers, governmental officials & ...
  - A data savvy public
- Long-term development of leadership
  - curated data
  - expert teams
- Balanced investment
  - from collection to *“final mile” of* information delivery
- Open data and processes
  - encouraging scrutiny, challenge & contribution



Wiley Series on Parallel and Distributed Computing
*Albert Y. Zomaya, Series Editor*

The DATA Bonanza

Improving Knowledge Discovery
in Science, Engineering, and Business

EDITED BY
Malcolm Atkinson
Rob Baxter
Peter Brezany
Oscar Corcho
Michelle Galea
Mark Parsons
David Snelling
Jano van Hemert