

eResearch: identifying the weak links

Martie van Deventer

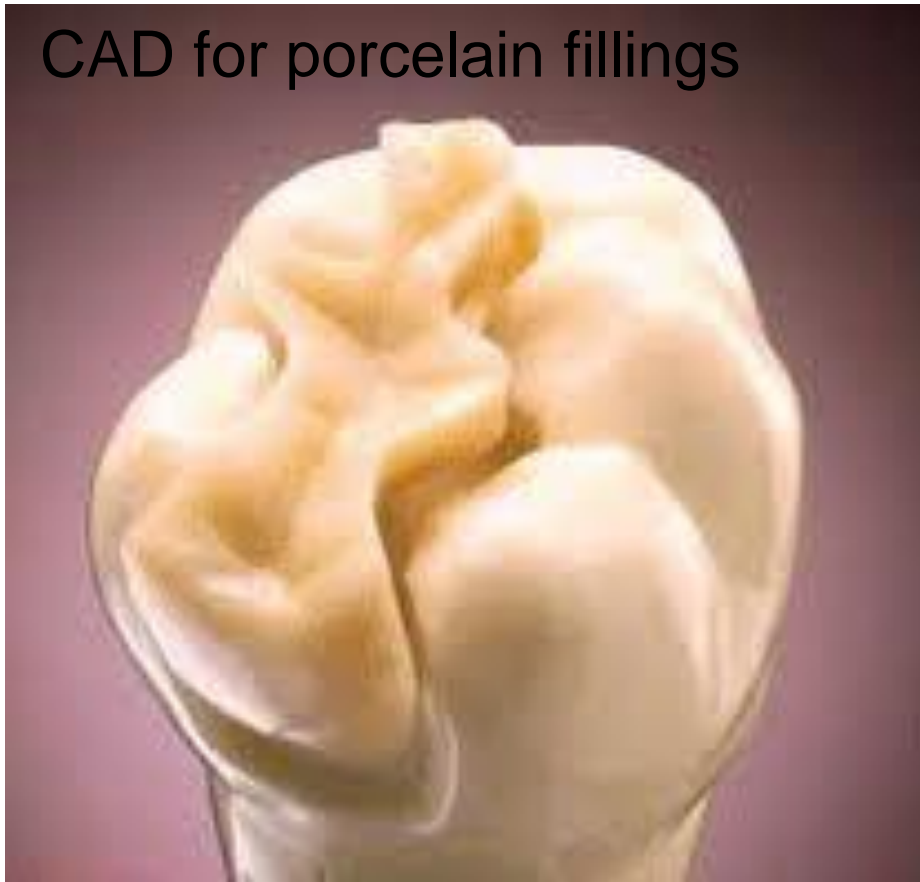
UP Seminar: 19 September 2012

The emergence of a networked knowledge society in the next twenty to thirty years is a major paradigm shift from the industrial model of the 19th and 20th century. This transition is of crucial importance in opening up new opportunities for education, social inclusion, and more efficient use of resources. Information and communication technologies are the effective tools of this transition.

© 8/2003 The Club of Rome

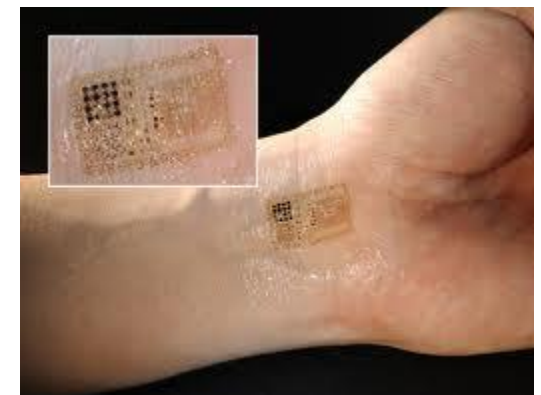
New products & services!

CAD for porcelain fillings



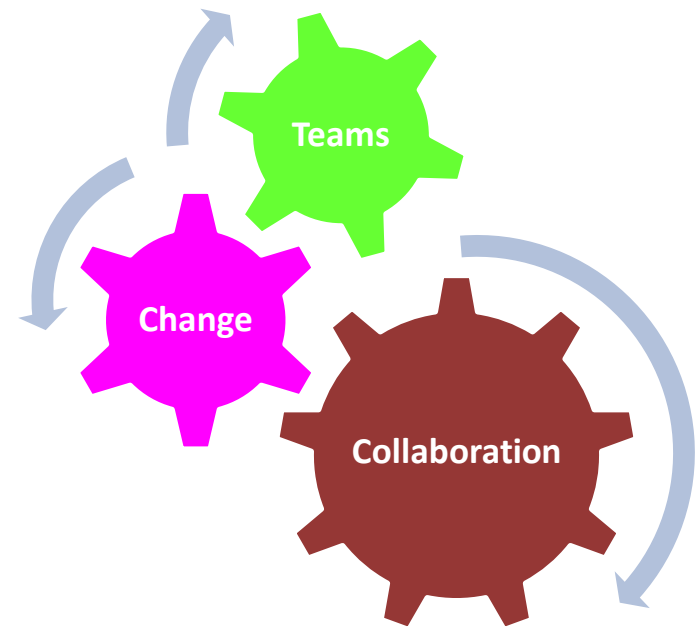
Before

After

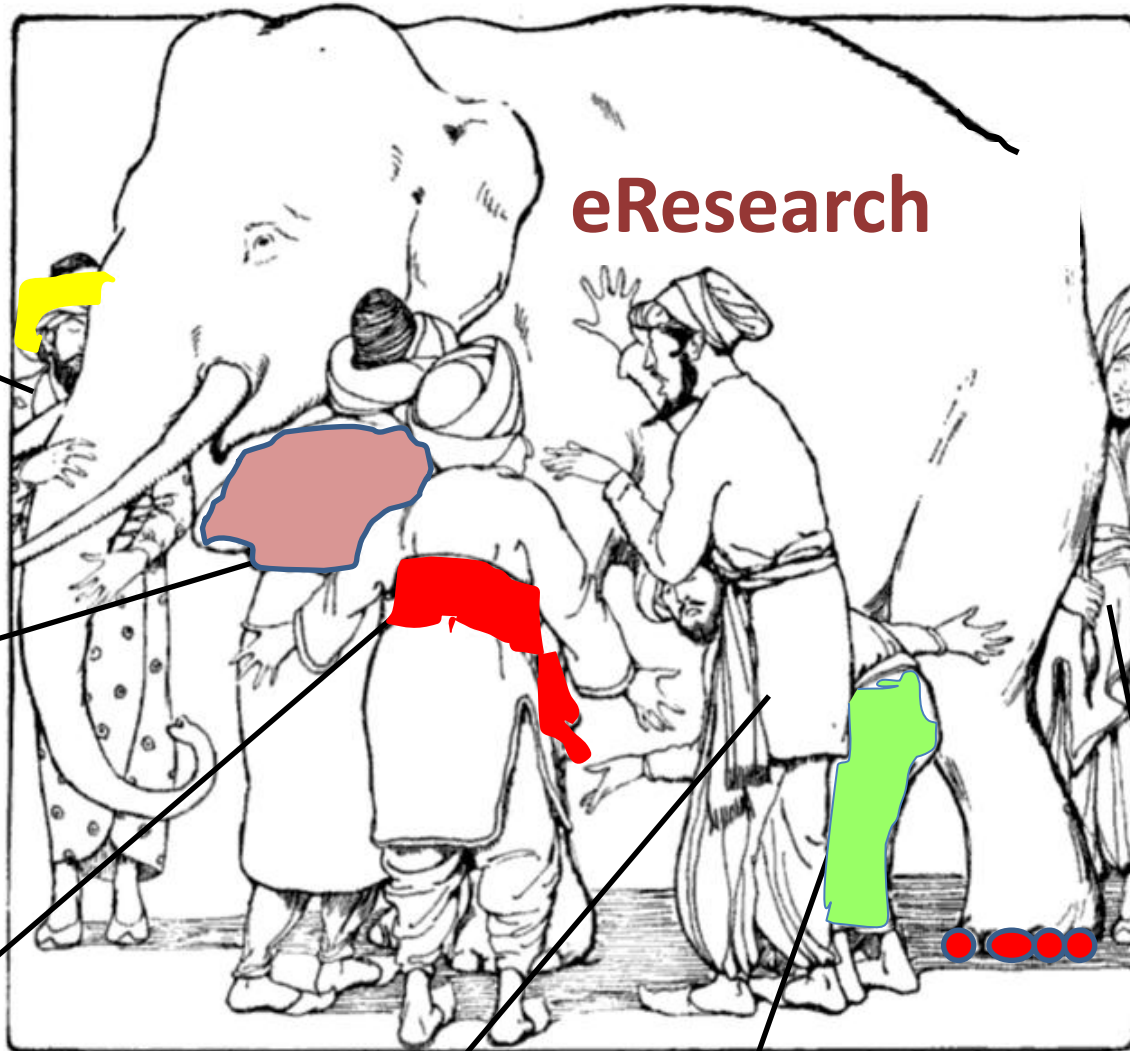


Roadmap: finding the pieces that do not fit comfortably and making them better





If eResearch is seen as bits and pieces
... you can regard that as a weak link



Workflow

*"flexibility;
web services;
integration"*

Databases

*"query processing;
data independence;
algebraic optimization;
needles in haystacks"*

Visualization

*"Exploratory science; mapping
quantitative data to intuition"*

Provenance

*"Reproducibility;
forensics;
sharing/reuse"*

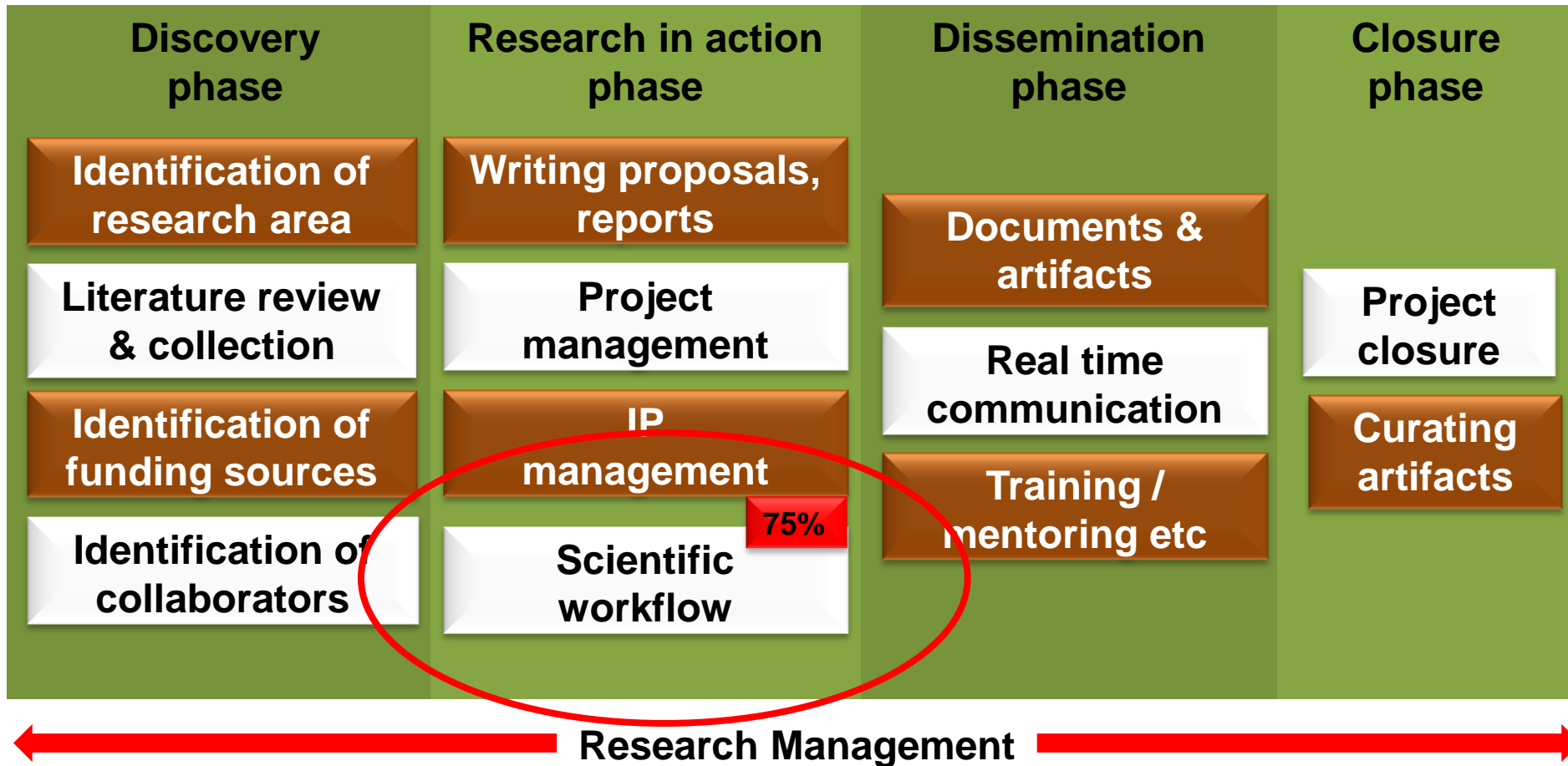
Cloud/Cluster

*"Massive data
parallelism"*

Mashups

*"Rapid Prototyping;
Simplified web
programming"*

Understand that the entire R&D process is affected ...





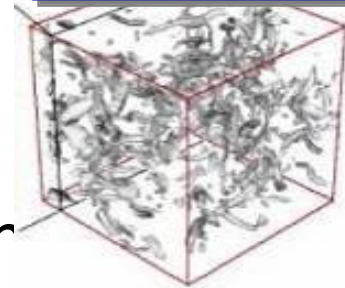
If we do not understand that things
have changed ... the link is VERY
weak!

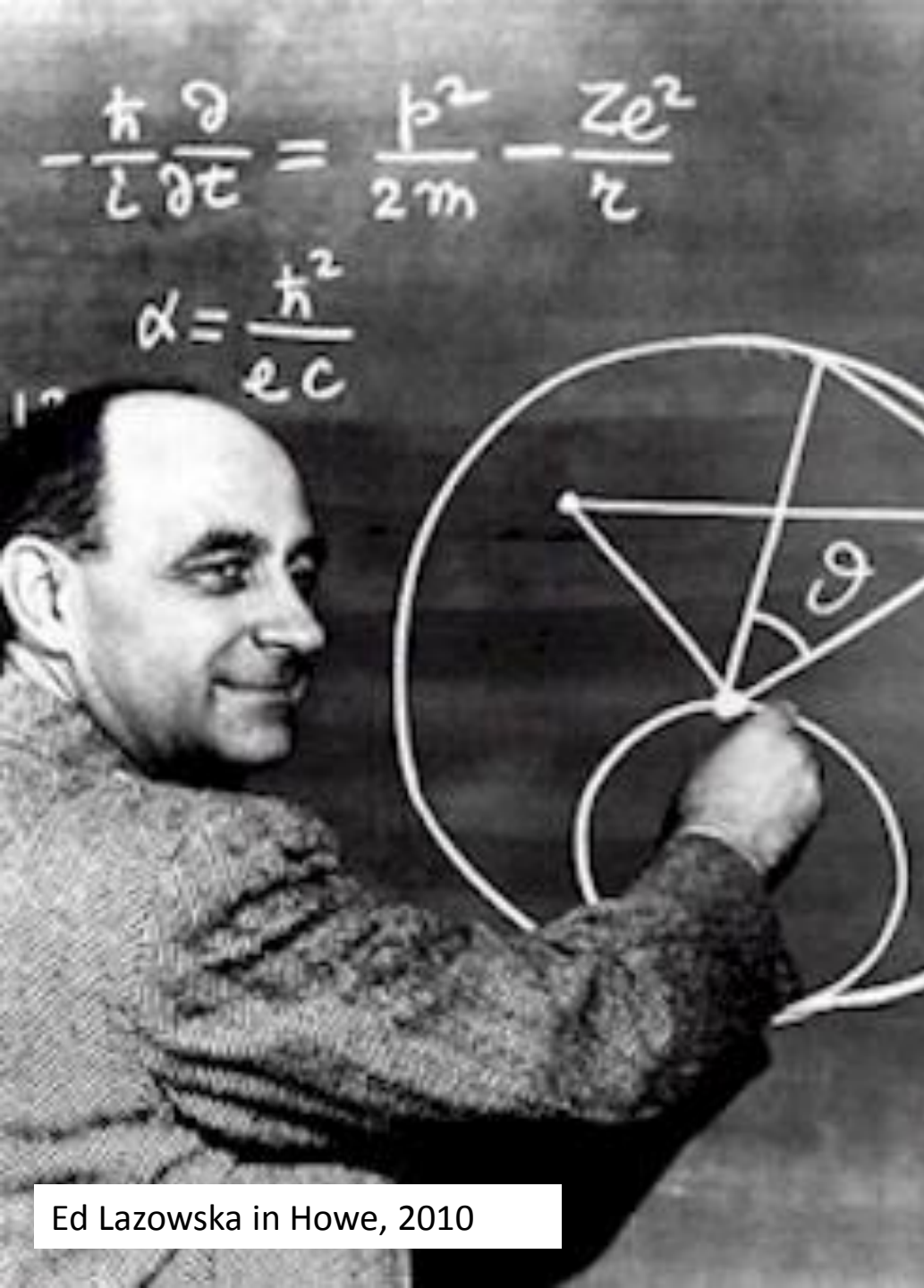
Science Paradigms



- Thousand years ago: science was **empirical**
describing natural phenomena
- Last few hundred years: **theoretical**
branch using models, generalizations
- Last few decades: a **computational** branch
simulating complex phenomena
- Today: **data exploration** (eScience/ eResearch)
unify theory, experiment, and simulation
 - Data captured by instruments Or generated by simulator
 - Processed by software
 - Information/Knowledge stored in computer
 - Scientist analyzes database / files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$






Ed Lazowska in Howe, 2010

Theory Experiment Observation



Ed Lazowska in Howe, 2010

Theory Experiment Observation



Theory
Experiment
Observation



Ed Lazowska in Howe, 2010

Theory
Theory
Experiment
Experiment
Observation
Computational
Observation
Science

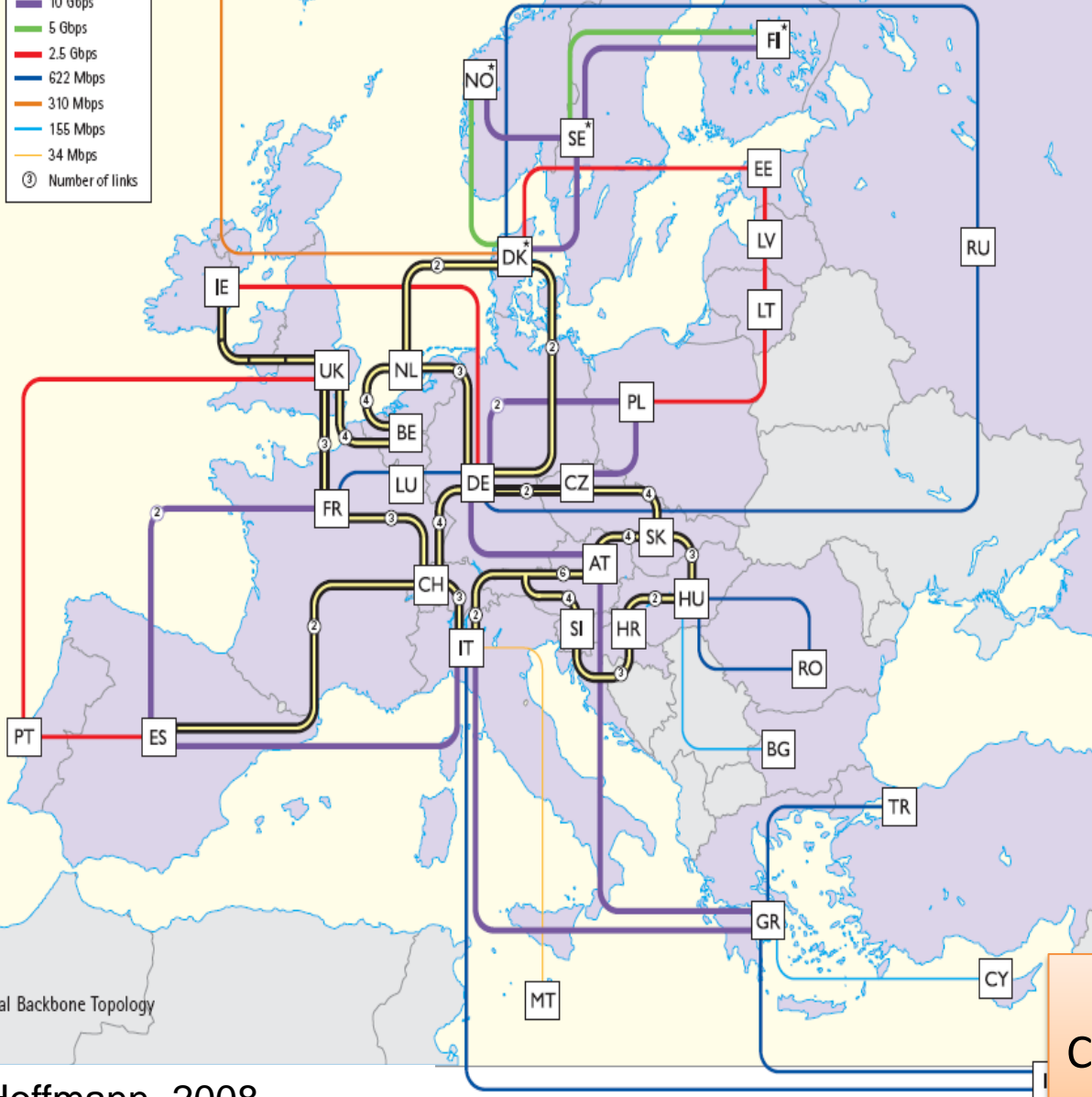
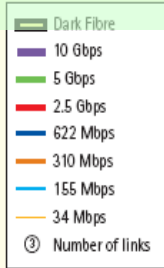




If we do not understand the eResearch needs reliable and extensive infrastructure ... which you should not try and build on your own it will become a weak link!

GÉANT 2 research network backbone, interconnect of NRENS

US similar, one country several funding agencies



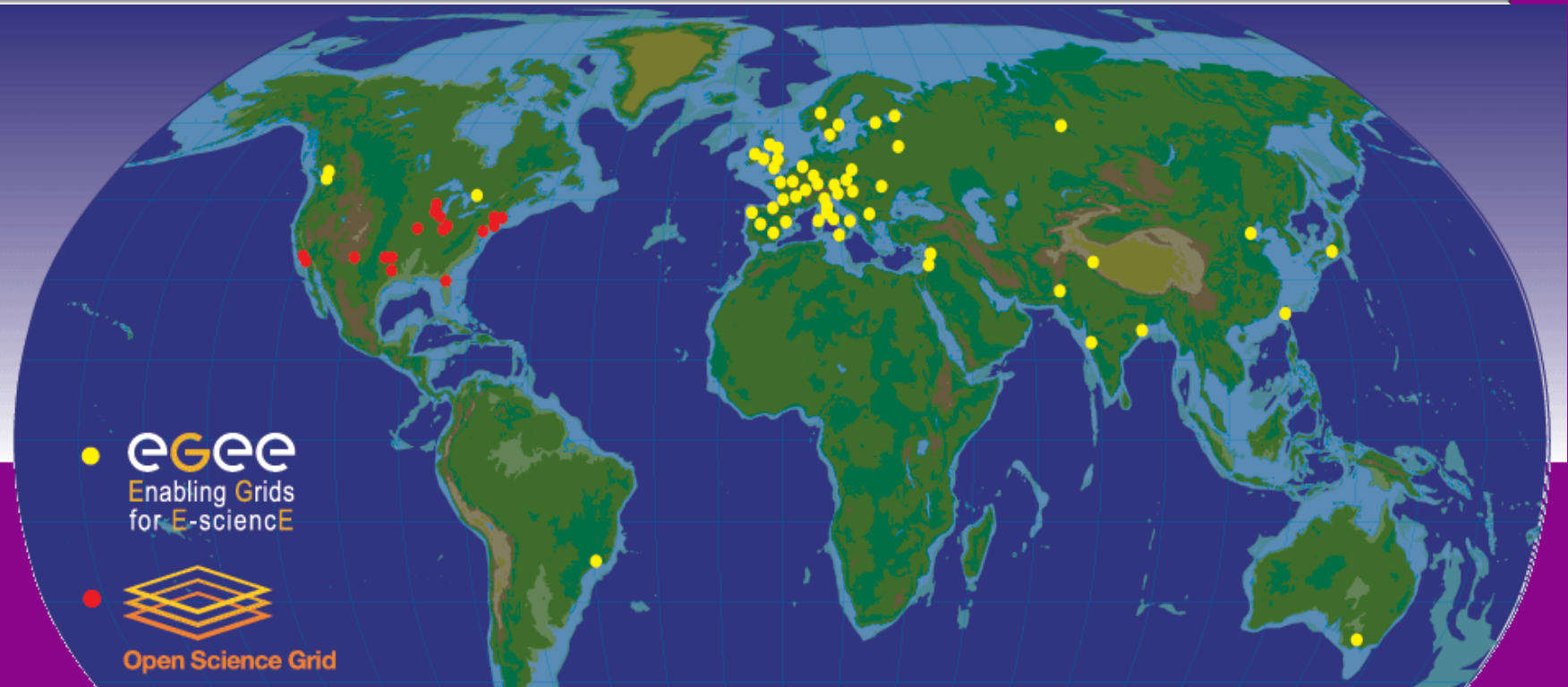
- Dark Fiber Connections Among 16 Countries:
- Austria
 - Belgium
 - Bosnia-Herzegovina
 - Czech Republic
 - Denmark
 - France
 - Germany
 - Hungary
 - Ireland
 - Italy,
 - Netherland
 - Slovakia
 - Slovenia
 - Spain
 - Switzerland
 - United Kingdom

CERN connectivity: >50GB/s

Initial Backbone Topology

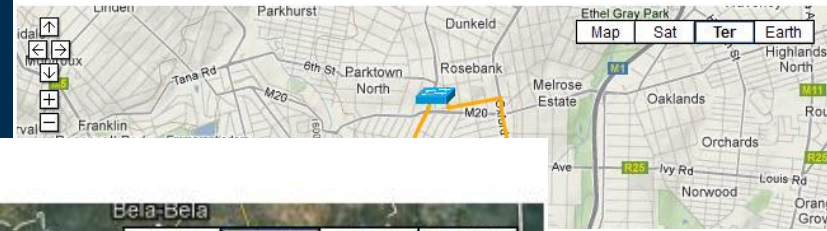
Grids: Centers around the world forming a **Global Computer System**

- The **EGEE** and **OSG** projects are the basis of the Worldwide LHC Computing Grid Project **WLCG**

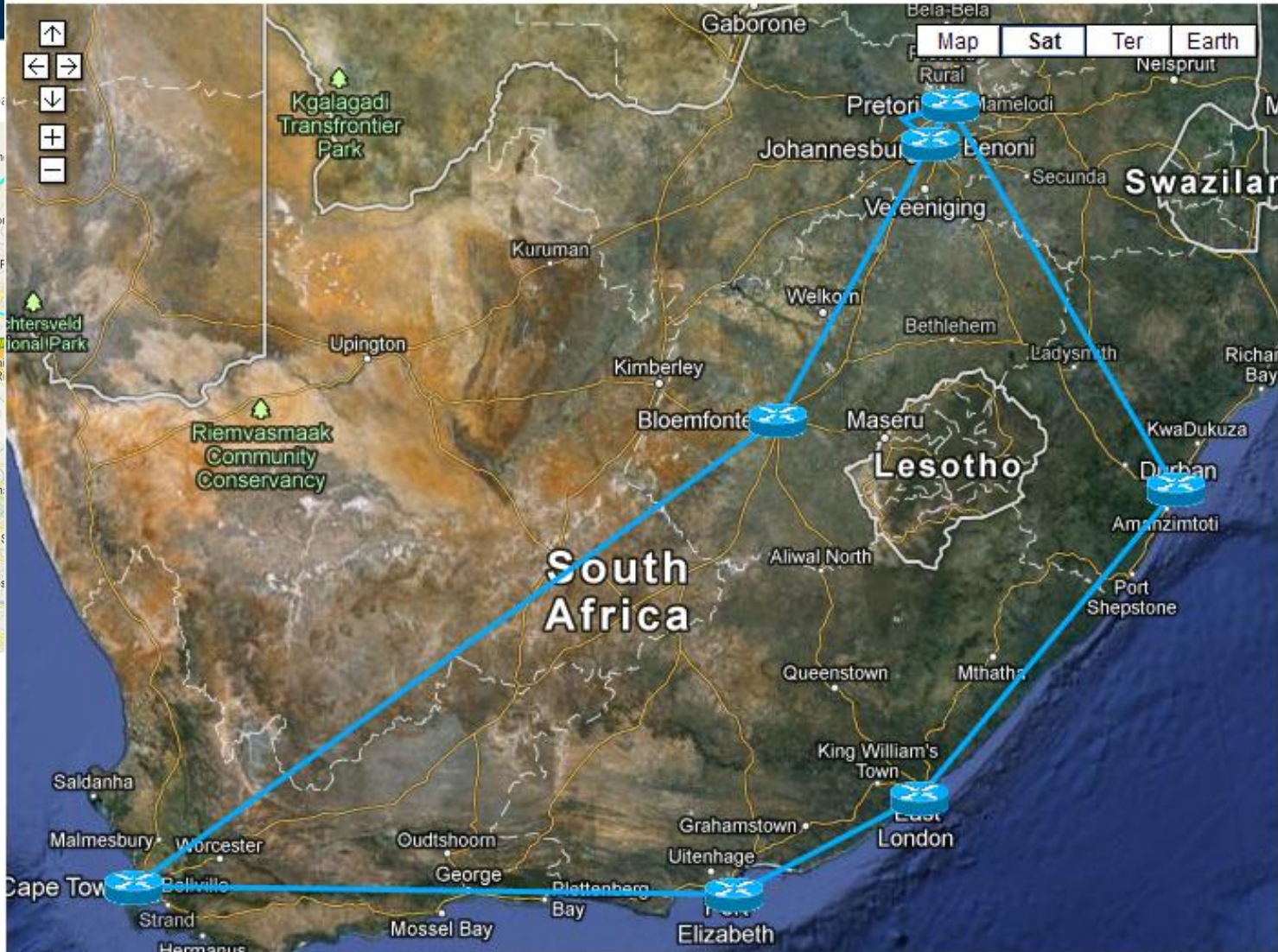


Inter-operation between Grids is working!

Download Johannesburg ring KML file directly [here](#).

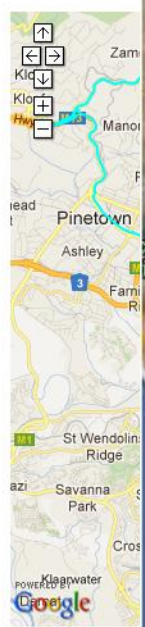


Map

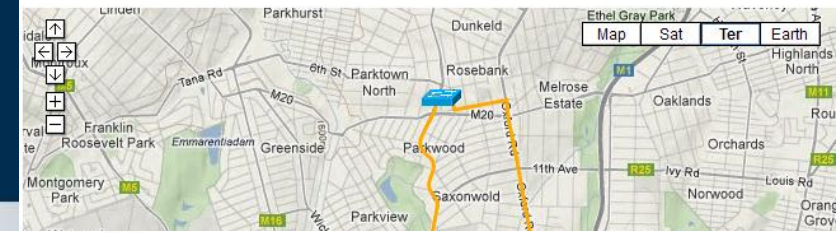


Durban

Download Durban

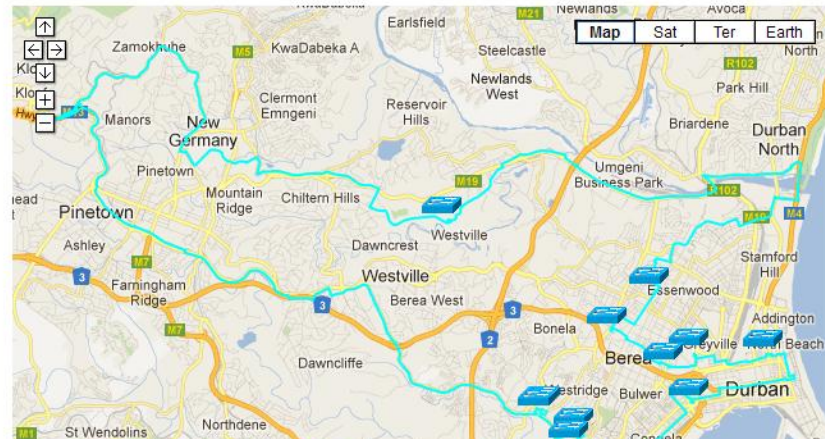


Download Johannesburg ring KML file directly [here](#).



Durban

Download Durban ring KML file directly [here](#).



Pretoria DWDM

Download Pretoria DWDM ring KML file directly [here](#).



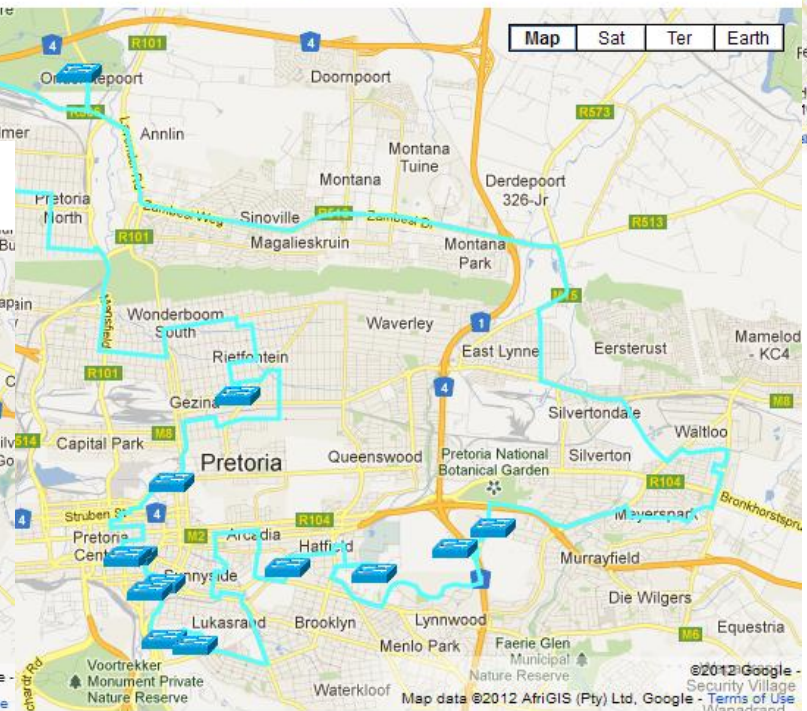
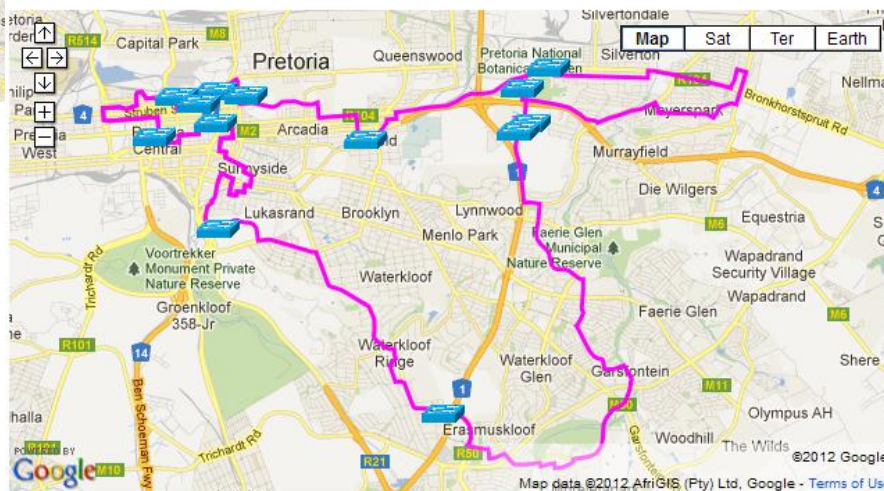
Pretoria North

Download Pretoria North ring KML file directly [here](#).



Pretoria South

Download Pretoria South ring KML file directly [here](#).



[View Larger Map](#)

Country	Operational NREN	NREN under development	Name of NREN/ Comments
Angola			No information
Botswana			No information
DRC		x	DRC REN in very early stages of formation
Lesotho			Universities connected by South Africa's TENET
Madagascar			World Bank project to promote REN
Malawi	x		MAREN- VSAT based REN
Mauritius			No information
Mozambique		x	MoRNet
Namibia		x	An educational ISP called EDUNET being created under the XNET initiative
South Africa	x		TENET- Most developed NREN. But also SANReN under development
Swaziland			Universities connected by TENET
Tanzania		x	TENET
Zambia		x	ZAMREN
Zimbabwe		x	ETNZ (Education and Tertiary Network of Zimbabwe)

Table 24- Status of NREN development in the SADC region

VREs to integrate technologies for work



PoI-SABINA Virtual Research Environment

You are not logged in. (Login)



Researchers of natural products enhancing food security and improving health!

Login

Username

Password

Login

[Create new account](#)

[Lost password?](#)

[Back to SABINA](#)

This digital workspace is home to a number of researchers concerned with the utilisation of African natural products - to enhance food security and improve human and animal health.

SABINA is a network that constitutes the following partners: University of Malawi, University of Namibia, University of Dar es Salaam, University of Pretoria, University of Witwatersrand, Council for Scientific and Industrial Research and Tea Research Foundation of Central Africa. SABINA partnership is focused on the building of capacity in natural product research through training of MSc and PhD students. The main goal of the SABINA network initiative is to implement proactive postgraduate programmes in the chemistry/biochemistry of natural products. This project is funded as a Carnegie Regional Initiative in Science and Education (see <http://sites.ias.edu/sig/ri/se/>).

The linked POL-SABINA initiative is mainly funded by EU ACP Science and Technology Programme and is supported through the Carnegie-RISE initiative. It focuses on the development of biological resources with application in medicine, health promotion, and

SciDev.Net

LATEST NEWS

on science, technology & the developing world.

Amazon Peruvians show protection against bat rabies

A study of two communities in northern Peru has revealed that a significant proportion have developed immunity to the rabies

virus. [More...](#)

3 August 2012 | Source:

Global initiative seeks to boost health innovation

Catalonia International

e-Science/ e-Research

... is about more than Networks, Grids, High Performance Computing... It is about **global collaboration** in key areas of science and the next generation of infrastructure that will enable it.

John Taylor, Director Research Councils, UK, 2000



SKA SOUTH AFRICA

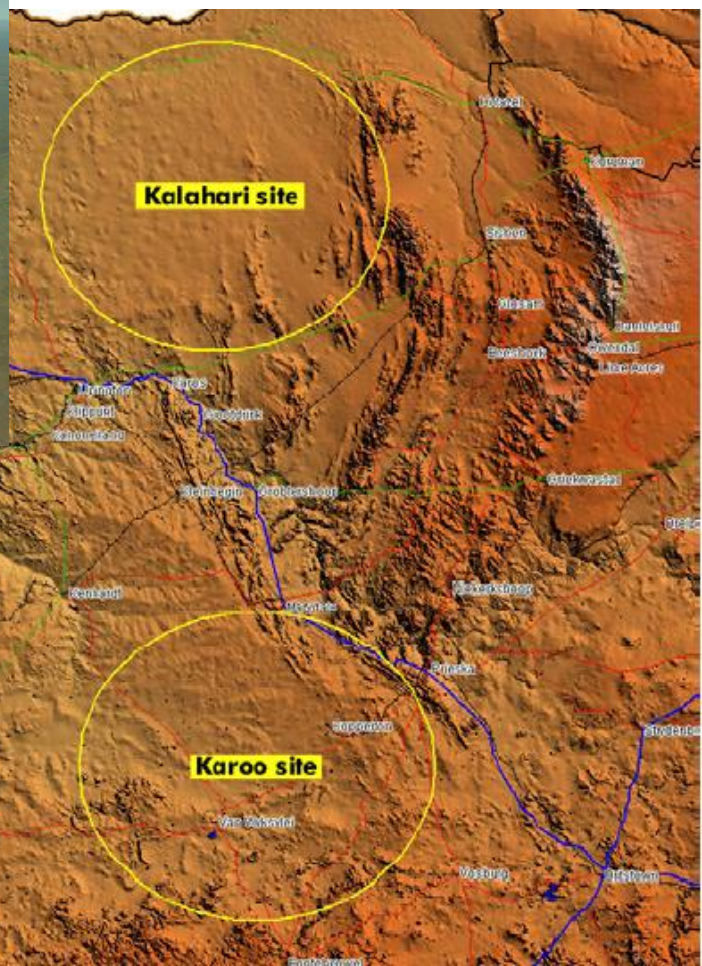
SQUARE KILOMETRE ARRAY



NORTHERN CAPE PROVINCE

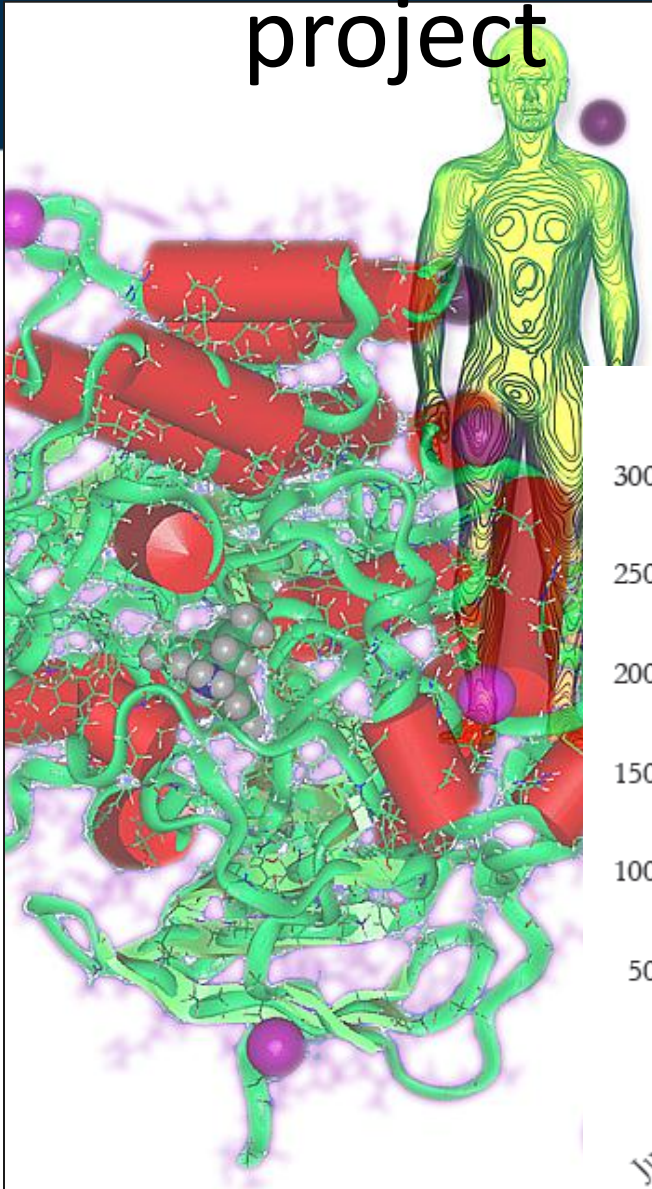
Northern Cape Province - South Africa
 Alternative Sites

Log spiral array map for the Namaqualand Site

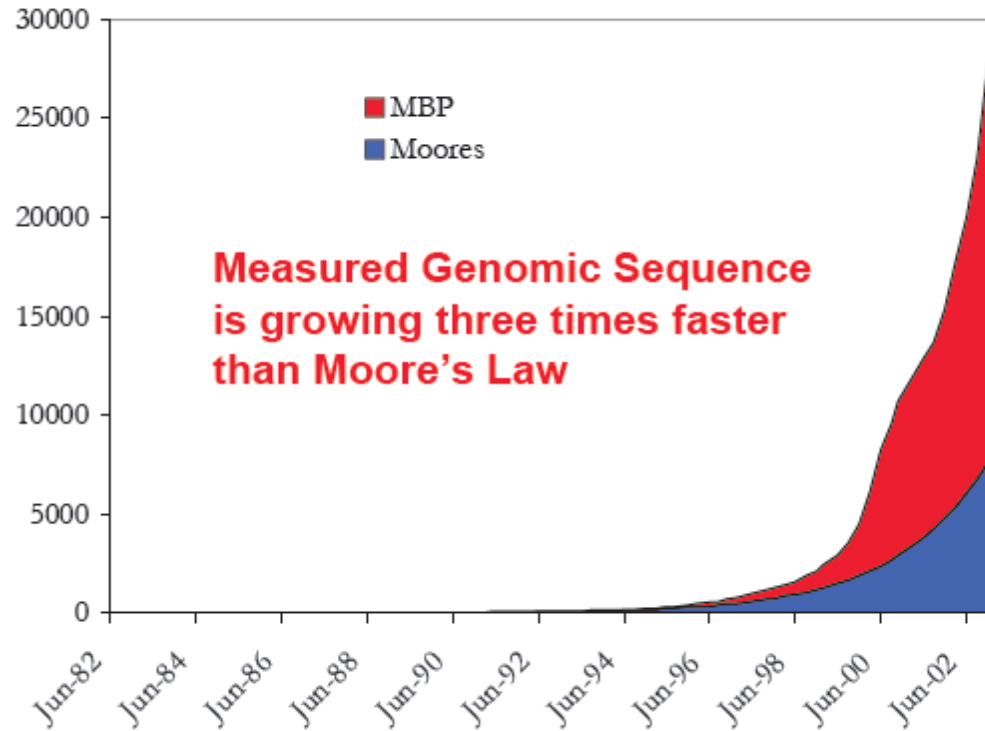


Human genome project

The Living Cell – A Grand Challenge For the Physical Sciences

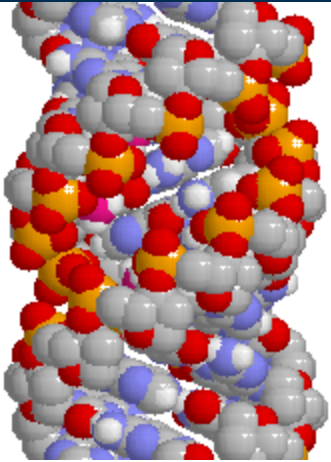


The Genomic Data Explosion

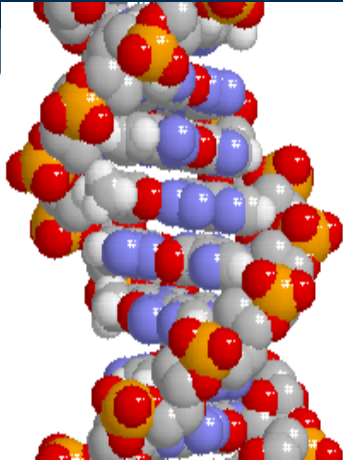


Courtesy: Graham Cameron

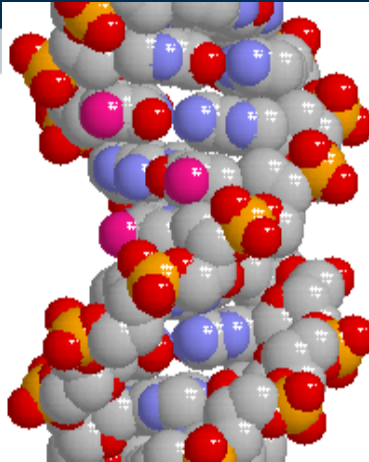
DNA conformation structures



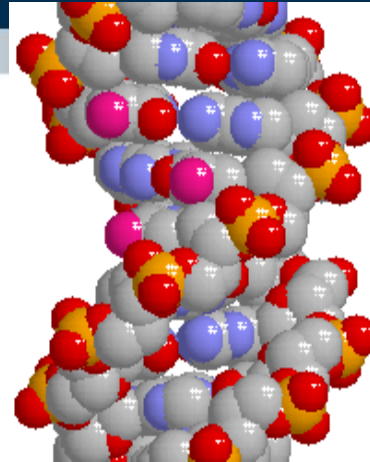
A-DNA
RH
11 bp/turn
pitch=28.2 Å



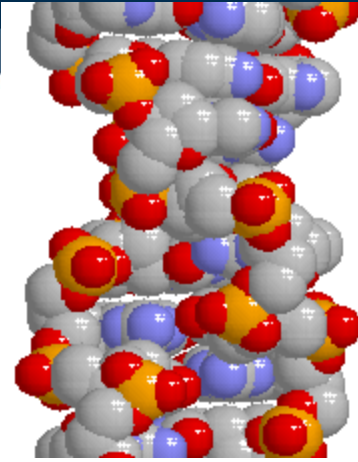
B-DNA
RH
10 bp/turn
pitch=34 Å



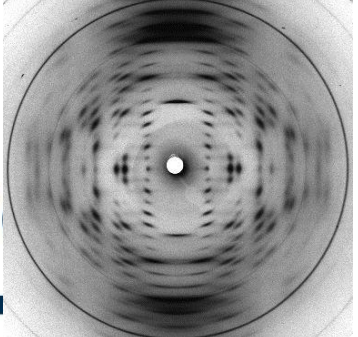
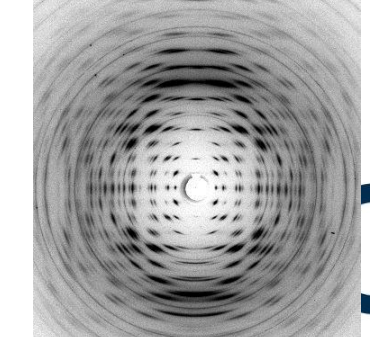
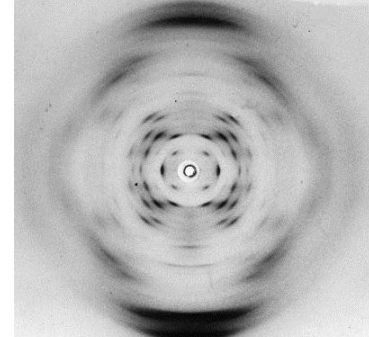
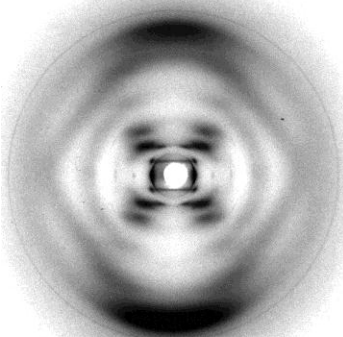
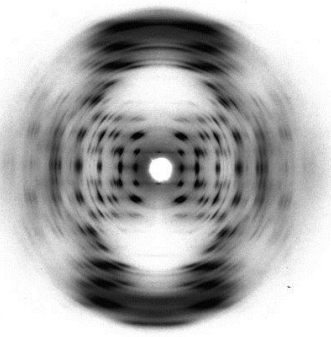
C-DNA
RH
9.33 bp/turn
pitch=31 Å



D-DNA
RH
8 bp/turn
pitch=24.2 Å



Z-DNA
LH
12 bp/turn
pitch=43 Å



Every datum counts!

Capitalising on small contributions to
the big dreams of mobilising
biodiversity information

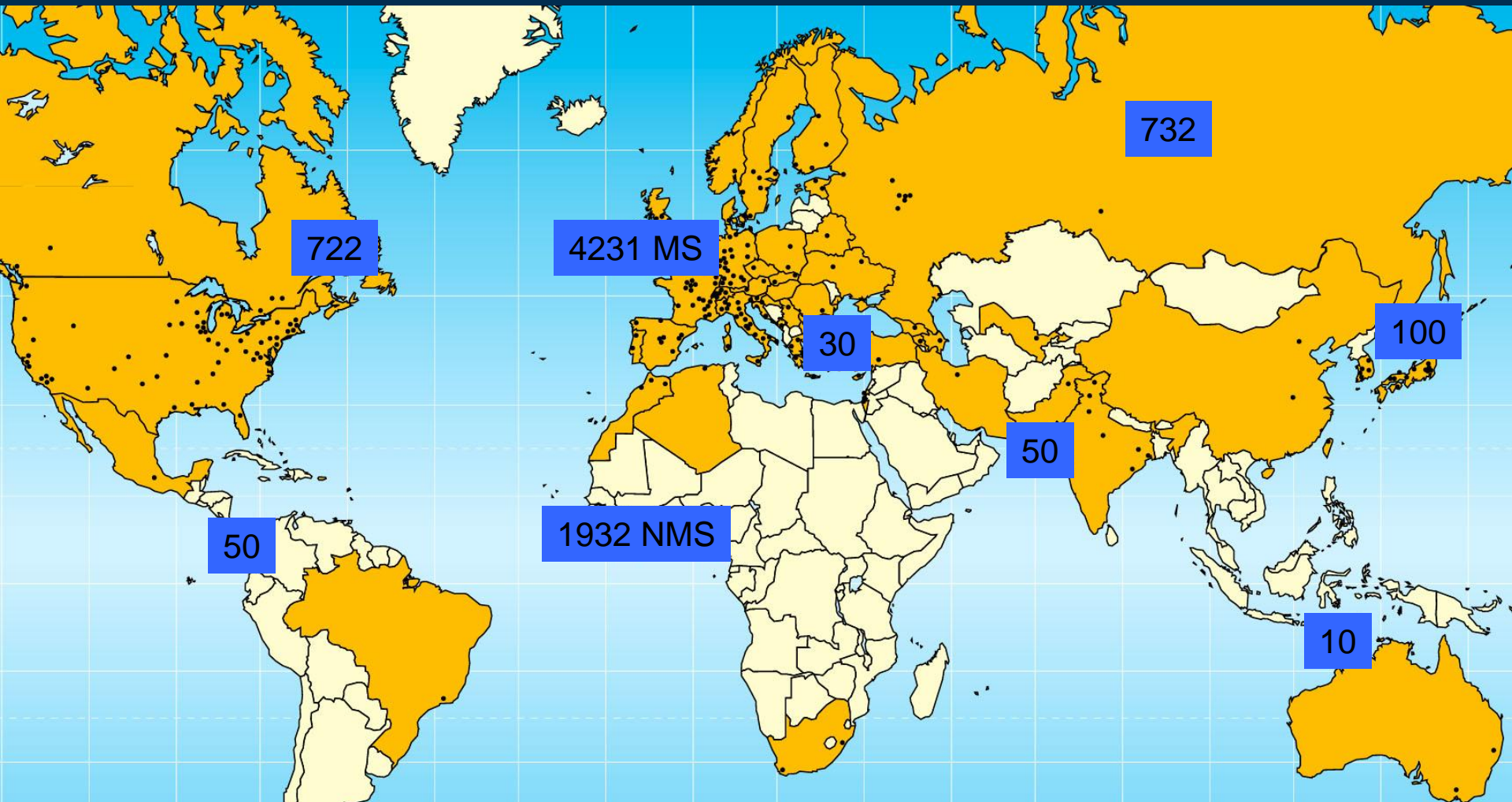


Vishwas Chavan, Eamonn O' Tuama,
Samy Gaiji, David Remsen and Nicholas King

We need to understand that eResearch is about people solving new/ different problems

- First and foremost it is about solving problems that could not be solved before
- Collaborating with the best minds available - across boundaries that could previously only be crossed with difficulty
- This sounds familiar but the scale is different ...


Particle physics: Network of universities around the world, virtual neighbourhood by ICT



CERN: **20 member states**; collaboration of **8000** scientists from **>500** institutes; **yearly** change over: **> 1200** (mostly young) scientists

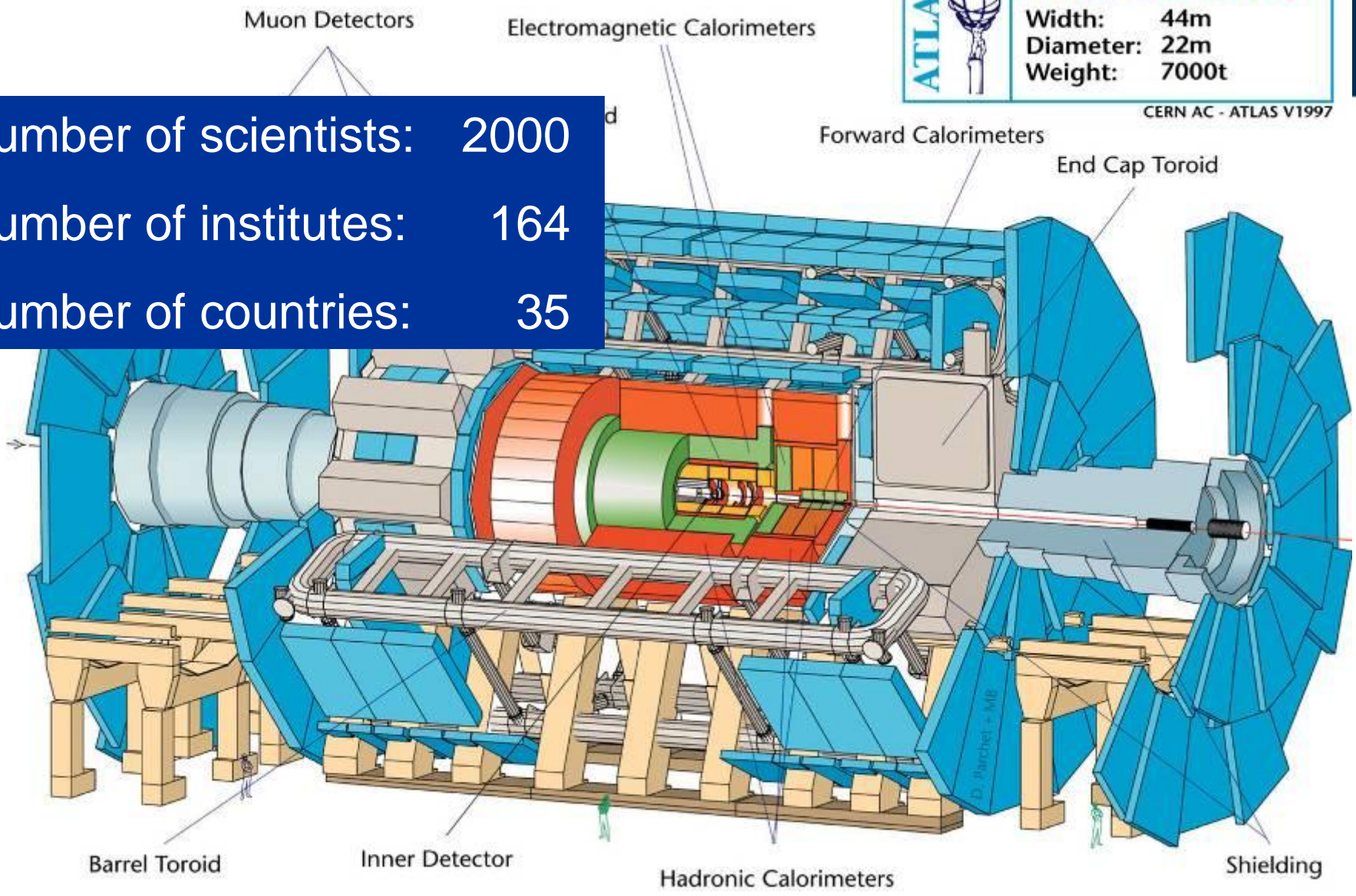
Hoffmann, 2008

ATLAS (spokesperson Peter Jenni)

ATLAS 	Detector characteristics	
	Width:	44m
	Diameter:	22m
	Weight:	7000t

CERN AC - ATLAS V1997

Number of scientists: 2000
Number of institutes: 164
Number of countries: 35



The CMS Detector

CALORIMETERS

SUPERCONDUCTING COIL

Number of scientists: 2350
Number of institutes: 180
Number of countries: 38

TRACKER

Silicon Microstrips
Pixels

Total weight : 12,500 t
Overall diameter : 15 m
Overall length : 21.6 m
Magnetic field : 4 Tesla

ECAL
Scintillating
PbWO₄ crystals

HCAL
Plastic scintillator/brass
sandwich

IRON YOKE

MUON BARREL

MUON
ENDCAPS

Drift Tube
Chambers

Resistive Plate
Chambers

Cathode Strip Chambers
Resistive Plate Chambers

our future through science



If we do not realise that eResearch
is

also about communication across
virtual organisations & distributed
teams ... the link is weak!

Services/utilities for virtual organisations: CERN

- Internet (video-) **telephony**
- Access Grid, EVO: **virtual meeting room** over internet
- **Agenda maker** with presentations, minutes (CERN: in 2007 ~14 000 meetings with 65 000 contributions for downloading)
- EDMS "life cycle and **configuration management** of complex, distributed apparatus"
- Other **book-keeping services** in a heterogeneous, distributed technical or resources environment
 - Individual notes
 - Logbooks
- Digital **Library**: Preprints, publications, web-enabled publications
- Object oriented, annotated, **curated data and data services**
(Semantic) search engines
- **Persistency guaranteed** only for referenced digital objects
- **Training**, tutorials, schools, summaries, educational material, seminars, colloquia, . .
- **Interconnect/share** within institution and with other disciplines for more comprehensive, integrated services

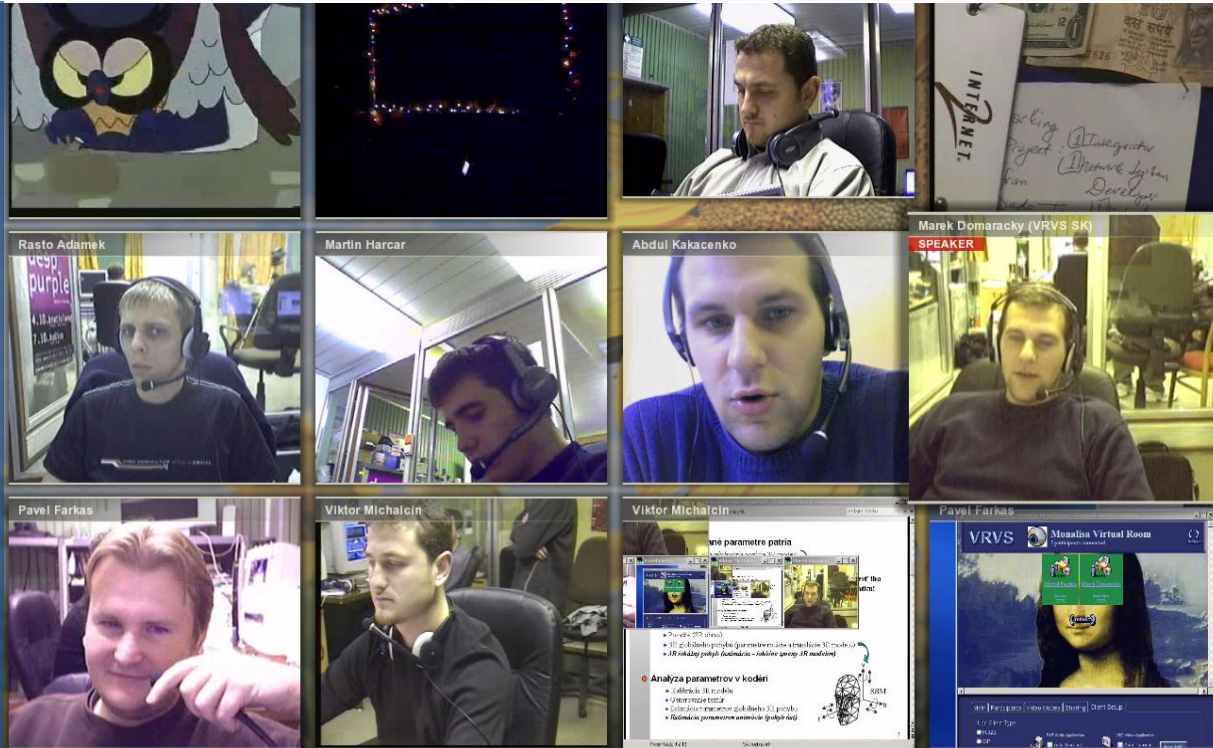
Do not underestimate video!

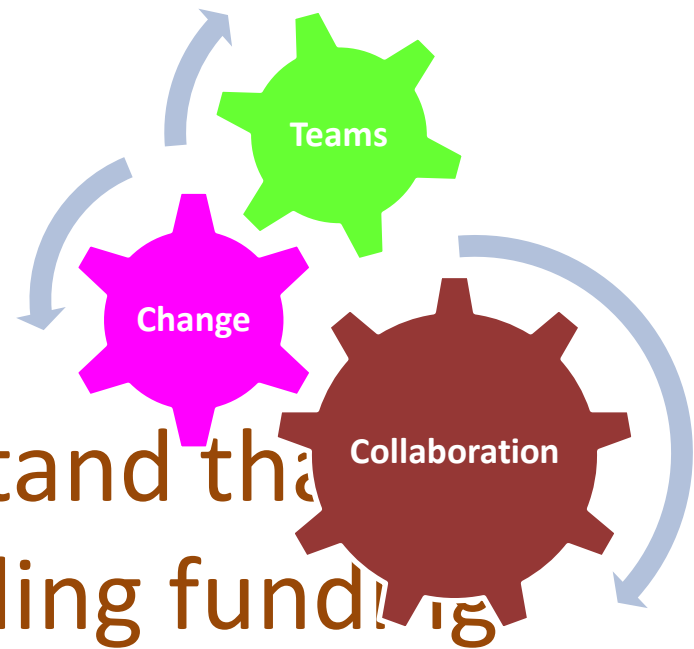


EVO – Enabling Virtual Organisations

EVO: a most flexible collaboration system,

After 10 years of experience with VRVS (>6000 users, ~60 countries)





Should we not understand that
eResearch is about spending funding
much more wisely ...

- This research is so expensive that it should not be duplicated
- The scale is too large
- Competition is dead ... long live collaboration!

Experiment budgets: $\frac{1}{4}$... $\frac{1}{2}$ towards software

Software for

- Instrument scheduling
- Instrument control
- Data gathering
- Data reduction
- Database
- Analysis
- Modeling
- Visualization

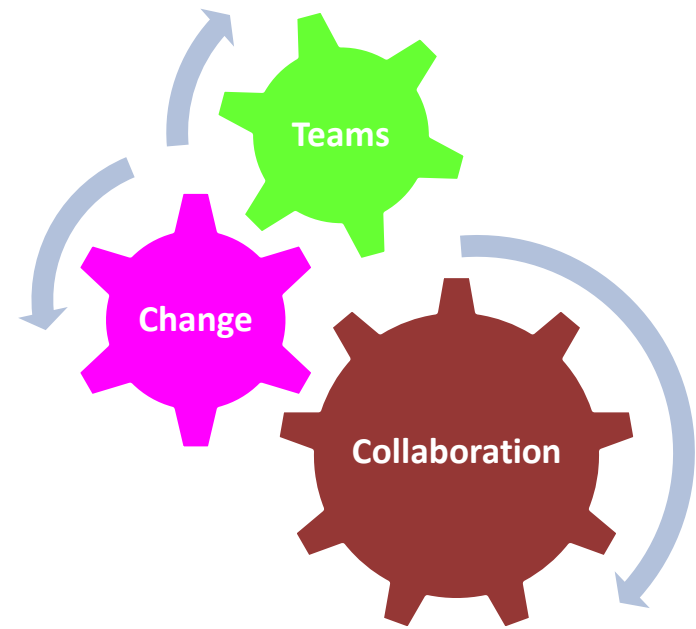
Millions of lines of code

- Repeated for experiment after experiment
- Not much sharing or learning

Building generic tools

- Workflow schedulers
- Databases and libraries
- Analysis packages
- Visualizers
- ...



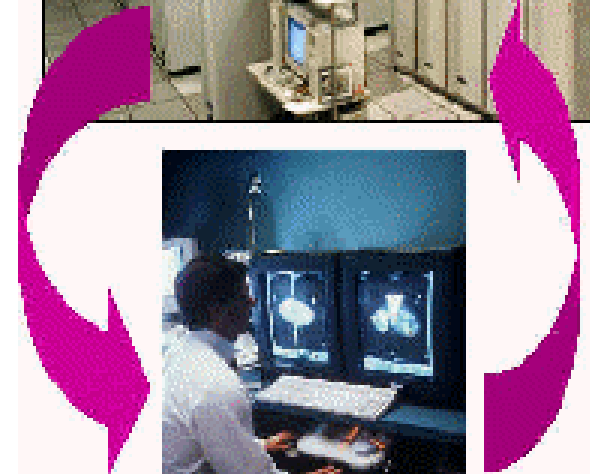
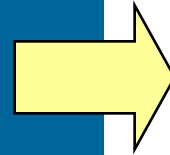
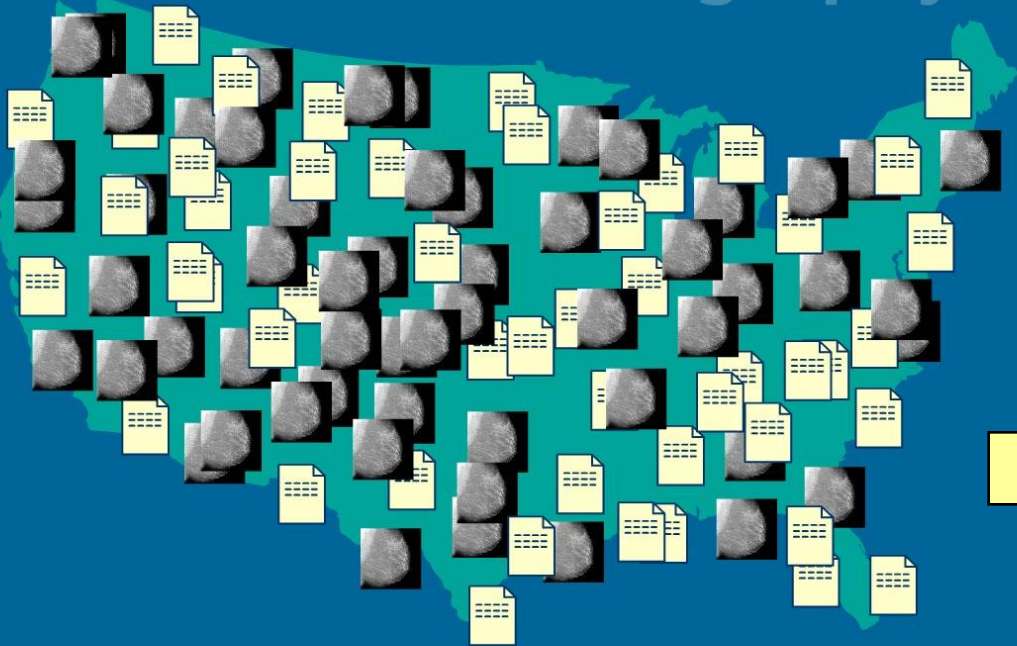


Which brings us to the research data ...
if we ignore the fact that eResearch is
ALL about data ... the link is very weak!

Think of Large-Scale Data

Hierarchical Storage and Indexing

DIGITAL Mammography



Scale of Digital Radiology data in the US

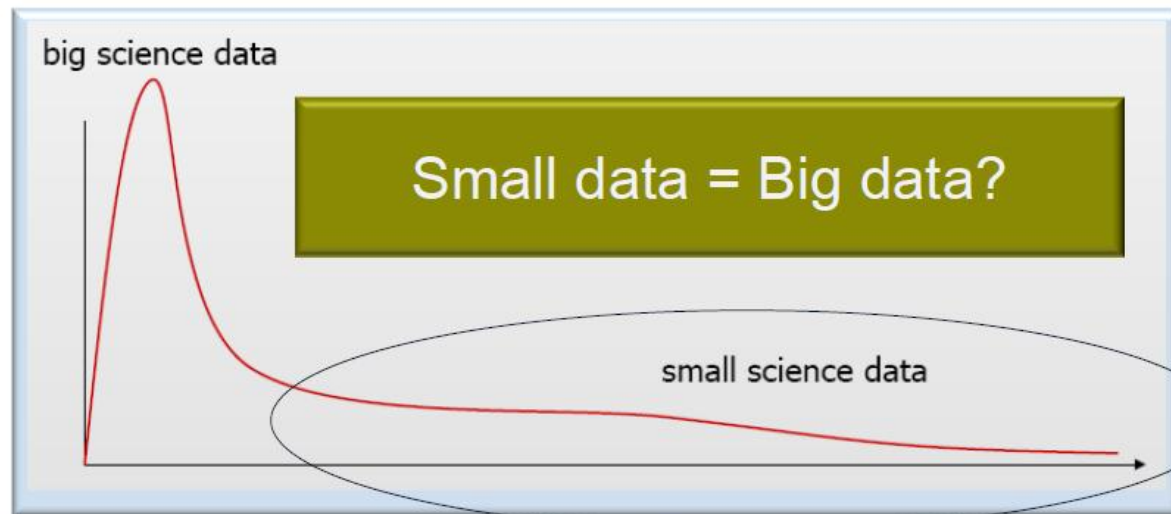
2000 Hospitals x 7 TB per year x 2 = **28 PetaBytes per year**



But also remember the small scale data

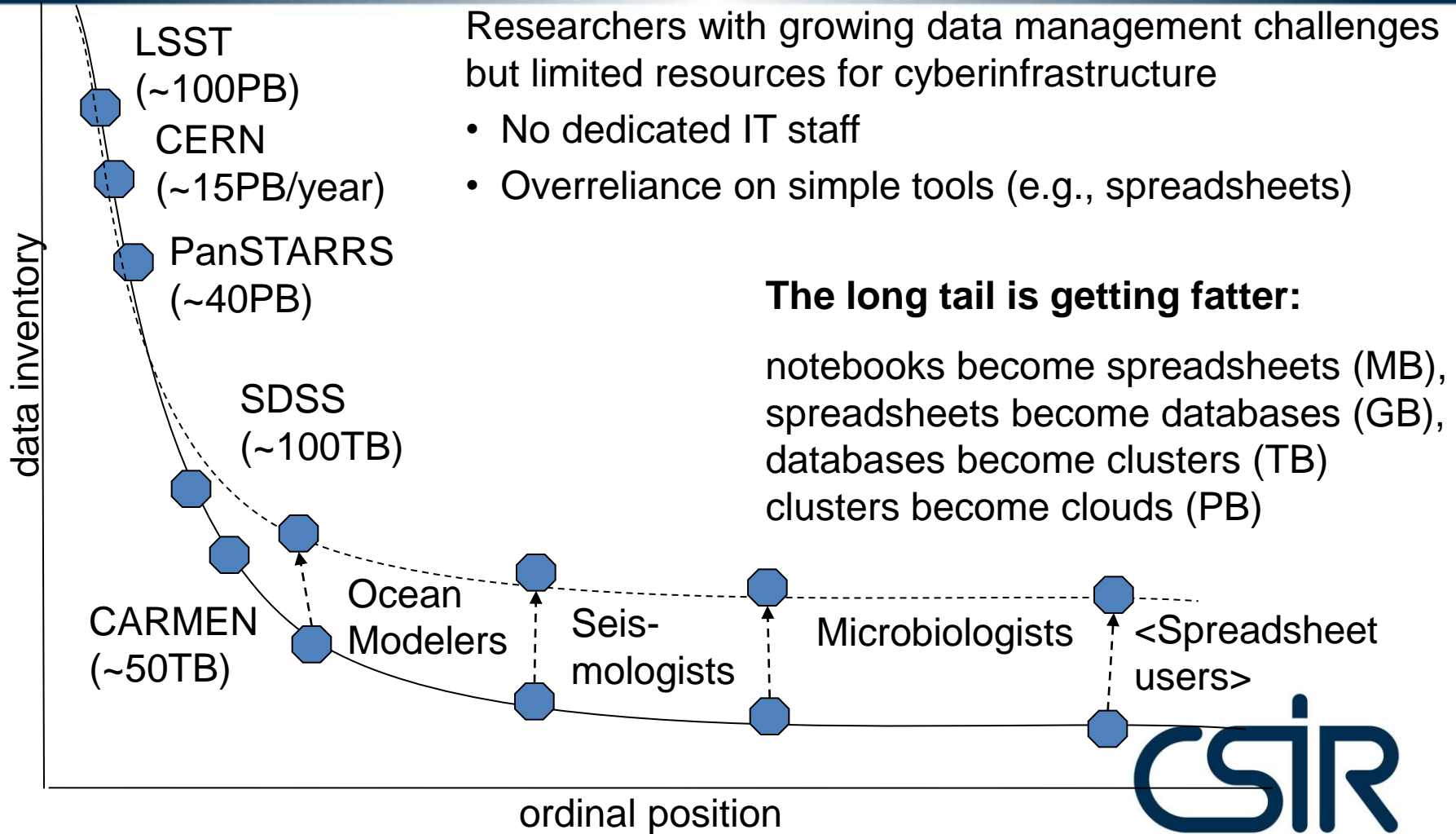


The dark tail



Palmer, C.L. (2008). *Contouring Curation for Disciplinary Difference and the Needs of Small Science*.
Sun PASIG Fall 2008 Meeting. 26 October.

The Long Tail



Mangling data ... a threat?

- Information overload is only a problem for manual curation.
- Google is not complaining about data deluge—they're constantly trying to get *more* data.
- The more data you collect, the better the filters will get.

Uhlir's 3 principles

- The same tool only a challenge in their use.

Public information wants to be free.

... it is an opportunity
Digital preservation does not happen by accident

- **Don't turn off the taps, build boats!**



our future through science



National Science Foundation
4201 Wilson Boulevard
Arlington, Virginia 22230

<http://www.nsf.gov/pubs/2012/nsf12058/nsf12058.pdf>

NSF 12-058

Dear Colleague Letter - Data Citation

DATE: March 29, 2012

Subject: **Data Citation in the Geosciences**

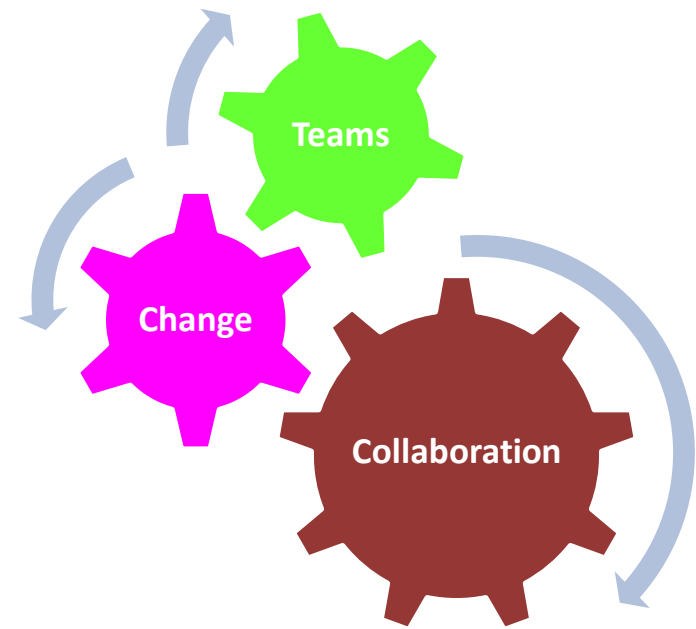
Facilitating open and equal access to data and data sets is a fundamental operating principle of the Directorate for Geosciences (GEO), and the National Science Foundation (NSF) as a whole. Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See [Award & Administration Guide \(AAG\) Chapter VI.D.4](#).

GEO believes that the benefit to the over-arching scientific enterprise from access to data and data sets far outweigh the burden of time and resources to an individual investigator and his or her host institution. GEO encourages data citation as a means to achieve the desired operating state for the geosciences with open and equal access to data available to all interested parties at a reasonable cost.

Principles of data citation are at various stages of maturity and adoption among scientific and engineering communities. In a 2009 report, for example, the American Meteorological Society (AMS) was urged by its *Ad Hoc* Committee on Data Stewardship Prospectus to "develop a plan for citing data referenced in publications and preserving data links for the long term." The American Geophysical Union (AGU) has taken the position that "the scientific community should recognize the professional value of data activities by endorsing the concept of publication of data, to be credited and cited like the products of any other scientific activity, and encouraging peer-review of such publications."

While many policy and practical challenges remain to be resolved and implemented, the Directorate for Geosciences encourages members of the community to lead an evolutionary transformation to establish data citation within the geosciences as the rule rather than the exception.

The Australian National Data Service lists many references to the benefits of and practices for data citation (http://ands.org.au/cite-data/resources.html#Data_Citation_Benefits). Benefits include the acceptance of research data as a legitimately citable contribution to the scientific record; permitting results to be verified and re-purposed for future study; and enabling data citation metrics to be tracked, as is done with

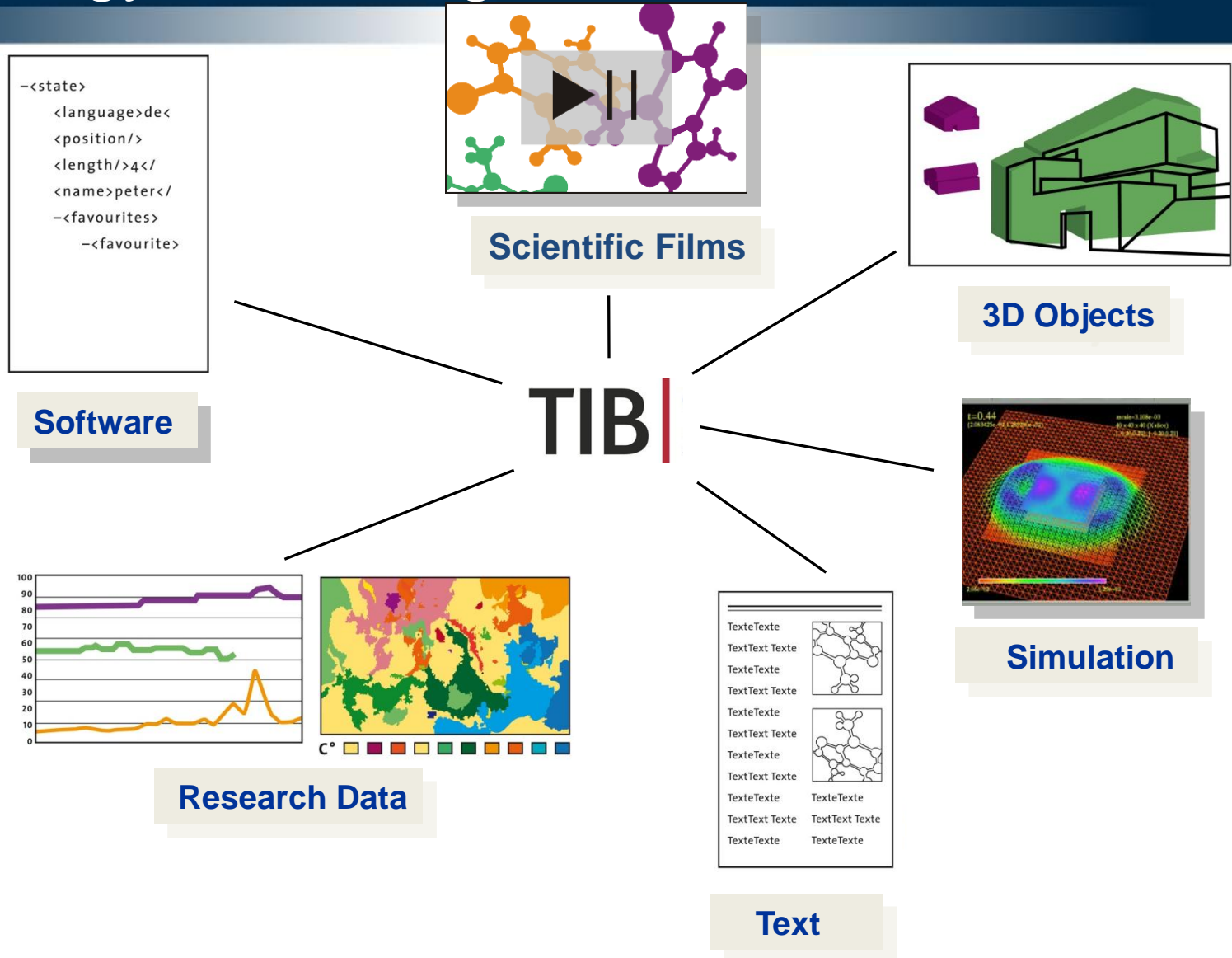


Understand that eResearch needs serious information management ...

Perhaps not the way we did IM in the past ...

- Simulators & experiments produce lots of data
- Projects have millions of files (or soon will) because in standard practice:
 - Each simulation run produces a file
 - Each instrument-day produces a file
 - Each process step produces a file
 - Files have descriptive names
 - Some files have similar formats (described elsewhere)
 - Different objects may require different formats
 - Using different languages
 - Using different standards
- No easy way to manage, integrate and/ or analyze all of this
- Highly unlikely that this will be done by hand or that
- Researchers will find managing content sexy!

German National Library of Science and Technology - including non-textual content



Extracting the information from the text

Chemical Names

The copper(I)-catalyzed 1,3-dipolar cycloaddition [33-38] of organic azides and alkynes (also called "click chemistry") resulting in the formation of 1,2,3-triazoles has become an increasingly attractive area [39]. According to the literature [33-38], the Cu(I) species can be used directly (e.g. CuI), or generated by oxidation of a Cu(0) or reduction of a Cu(II) species. Catalysis by the CuI is known to yield exclusively the 1,4-disubstituted regioisomer [33,34]. First, the N-(p-methoxyphenyl)-1-(trifluoromethyl)propargylamine was reacted with benzyl azide in the presence of CuI (10 mol%) and showed good reactivity with completion of the reaction within 24 h, whereas the use of CuSO₄/Na ascorbate afforded the cycloadduct in low yield. The reaction was then carried out with different propargylamines (N-(p-methoxyphenyl) and N-benzyl) and various azides at room temperature in acetonitrile within 24 h which afforded the compounds 2a-i with good yields (63-92%) after purification by column chromatography. The results are summarized in Table 1.

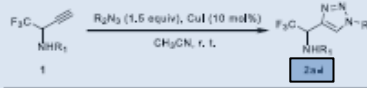
Linked entities from the table

As expected the new triazoles were formed in a fully regioselective manner affording the 1,4-regioisomer as highlighted from NOE experiments on compound 2c (Figure 1). A strong correlation was observed between the hydrogen H_a and H_b respectively. The structure of the other compounds 2a-i was assigned by analogy with 2c.

In our goal to study the influence of the CF₃ group on the conformation of peptidomimetics, we applied our strategy to the enantiopure trifluoromethyl-propargylamine **3** bearing the removable (*R*)-phenylglycinoil chiral auxiliary (Scheme 2) [30-32].

The reaction was carried out under the same condition with azidoacetic acid methyl ester and afforded the cycloadduct **4** in

Table 1: Copper(I)-catalyzed synthesis of 1,4-disubstituted triazoles

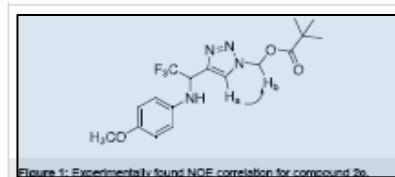


Entry	R ₁	R ₂	Product	Yield (%) ^b
1	-PMP*	-Bn	2a	82
2	-PMP	-CH ₂ CO ₂ Ph	2b	76
3	-PMP	-CH ₂ COOC(CH ₃) ₃	2c	73
4	-PMP	-CH ₂ CO ₂ CH ₃	2d	83
5	-PMP	-CH ₂ CH ₂ OH	2e	87
6	-Bn*	-Bn	2f	72
7	-Bn	-CH ₂ CO ₂ Ph	2g	63
8	-Bn	-CH ₂ CO ₂ CH ₃	2h	92
9	-Bn	-CH ₂ CH ₂ OH	2i	73

*PMP: p-methoxyphenyl, Bn: benzyl. ^bYield after flash purification.

Table with reaction scheme

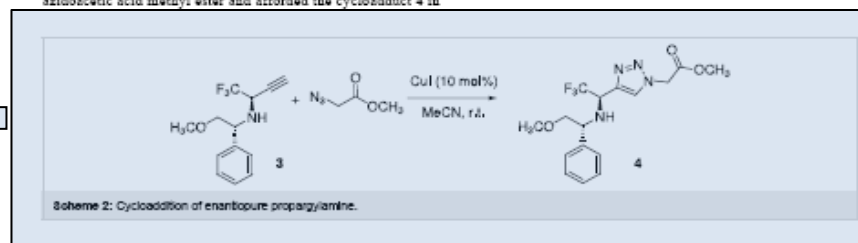
2a-i: Derivates from the reaction



Chemical structure

good yield (79%) and as a single isomer without any racemization. This compound can easily afford the free amino ester which is a promising trifluoromethyl building block for the synthesis of new triazole-based trifluoromethyl oligomers.

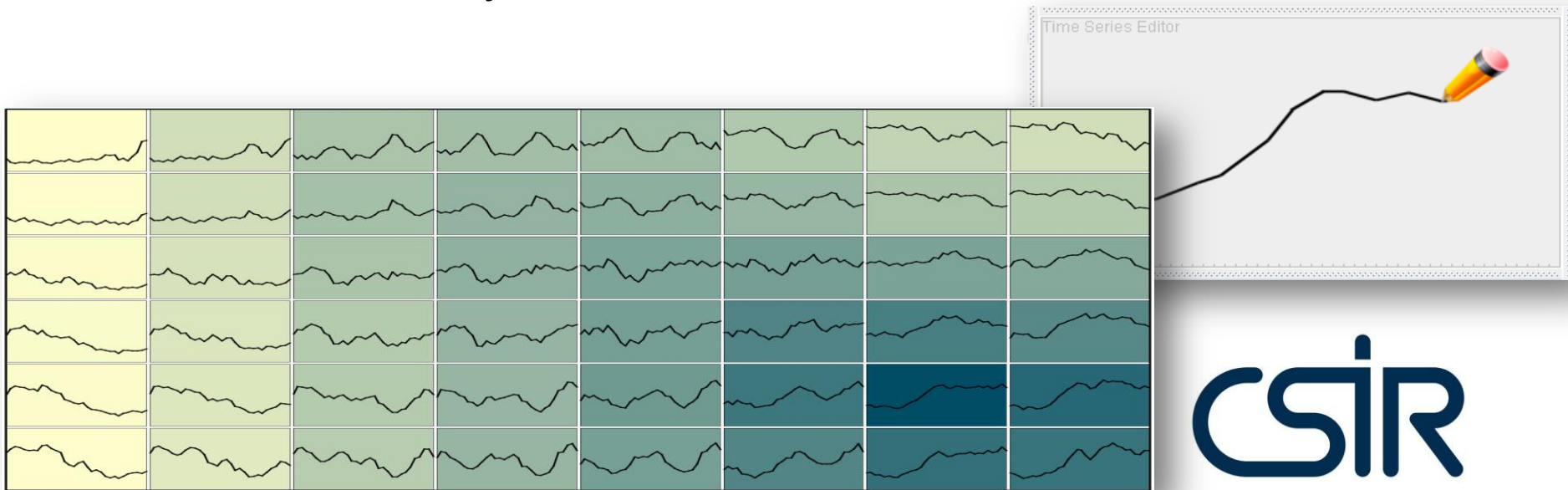
Reaction scheme



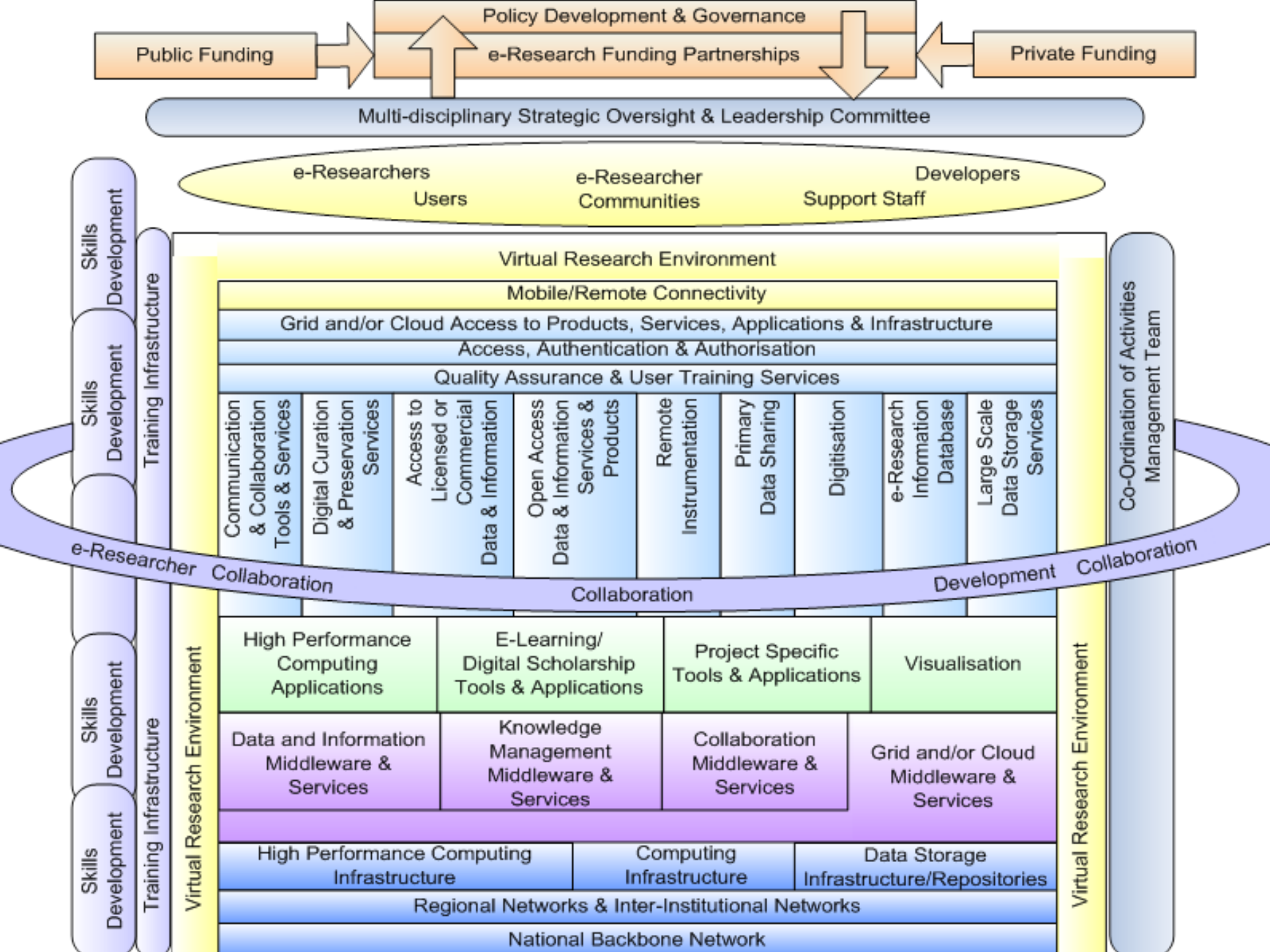
Visual search approach

Visual Search in Time series

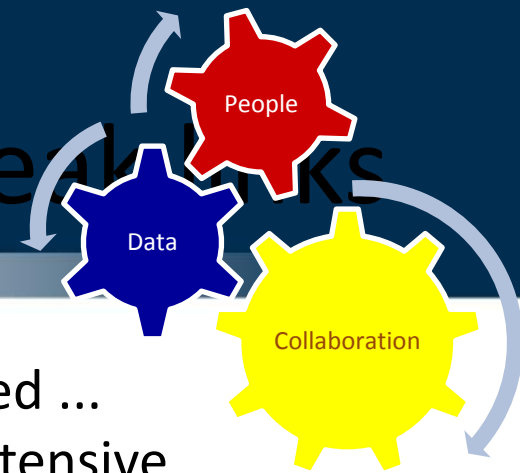
- Query-by-Example, Query-by-Sketch
- Visual Catalog as result list
- Colormaps for the indication of similarity



So ... let's put all of that in one model



Repeating the possible weak links



- Understand that the entire R&D process is affected ...
- Understand that eResearch needs reliable and extensive infrastructure - collaborate don't build your own
- Understand that eResearch is about people solving problems ...
- Realise that eResearch is about communication across virtual organisations & distributed teams – all experts in their own field!
- Understand that eResearch is about spending funding wisely
- Remember that eResearch adds the complexities of managing data together with a variety of other digital objects
- Understand that eResearch needs serious information management ... for which we too would need machine assistance
- Ultimately understand that all these eResearch cogs need to interact and run smoothly!

Some last words of wisdom

- e-Research: science of the 21st century, its students will be the entrepreneurs, innovators and teachers of the future
- Expand e-infrastructures, e-science and use them in common collaborative scientific projects for the development of all
- Open, comprehensive e-libraries, publications and data centres are essential: Science cannot do without! Develop these towards a Knowledge Utility!
- Persistent storage of relevant, curated data and derived knowledge and know how is a huge challenge for science and technology



Questions?

“On science and technology depend the standards of living of a nation ”

Abdus Salam, Pakistan

References

- Fernihough, S. 2011. e-Research (An Implementation Framework for South African Organisations). 4th African Conference for Digital Scholarship and Digital Curation, Pretoria: 16 May 2011. Available: <http://www.nedicc.ac.za/Conference/Upload%20Papaers/S%20Ferihough-eResearchImplementationFramework.pdf>
- Brase, J. 2011. Riding the Wave - Paradigm shifts in Information Access. Available: <http://www.slideshare.net/datacite/riding-the-wave-paradigm-shifts-in-information-access>
- Gray, J. n.d. eScience -- A Transformed Scientific Method. Available: <http://www.slideshare.net/dullhunk/escience-a-transformed-scientific-method>
- Hoffmann, H.F. 2008. From e-Science towards Knowledge Utilities. Presented at the 1st African Conference for Digital Scholarship and Curation, Pretoria. Available: <http://stardata.nrf.ac.za/nadicc/presentations/hoffmann.ppt>

References

- Howe, B. End-to-End eScience: Integrating Query, Workflow, Visualization, and Mashups at an Ocean Observatory Available: <http://www.slideshare.net/billhoweuw/endtoend-escience>
- Lötter, L. 2011. What can take the dark out of the long tail? Efforts to address the data management challenges of “small science”. Paper presented at the CHPC Conference, Pretoria. Available: http://www.chpcconf.co.za/Presentations/2011_08_L%20Lotter.pdf
- Uhler, P. 2008. Information Gulags, Intellectual Straightjackets & Memory Holes. 1st African Conference for Digital Scholarship and Curation, Pretoria. Available: http://stardata.nrf.ac.za/nadicc/presentations/uhler_pau.ppt