

# Microsoft Research Connections and Cloud Computing for Science

Dr. Fabrizio Gagliardi

Microsoft Research



# Introduction

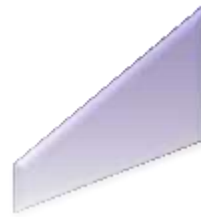


From 1975 till 2005: Computing Science at CERN ([www.cern.ch](http://www.cern.ch))

- Developing HPC distributed computing solutions for HEP
- Including EU-DataGrid and EGEE, foundation for the present LHC distributed computing Grid infrastructure ([www.eu-egee.org](http://www.eu-egee.org))
- Extending support to other scientific communities in the EU European Research Area context
- Among them **OGF-Europe** and the follow on **SIENA** project active of Grid and Cloud computing standards
- Deploying Cloud Computing for Science and Technology with **VENUS-C** ([www.venus-c.eu](http://www.venus-c.eu))

# Microsoft Research Connections

Work with the worldwide academic research community to speed research, improve education, and foster innovation



Collaborations to pursue scientific breakthroughs



Inspire emerging computer and research scientists



Accelerate scientific exploration with computing





-  *Microsoft Research Labs*
-  *External Research Groups*
-  *Technology Learning Labs*
-  *Collaborative Institutes and Centers*

# BSC-Microsoft Research Centre

Barcelona Supercomputing Centre: computer architecture, parallel programming models



MSRC expertise: programming language and operating system design & implementation

Research at the intersection of computer architecture, language implementation, and systems software

## Transactional memory (TM)

- Abstraction for scalable shared-memory data structures
- Research on using TM in real applications; game servers, recognition-mining-synthesis
- Debugging and profiling
- Major publications include PPOPP 09, MICRO 09, PPOPP

## Language runtime system

- Architecture support to accelerate synchronization and garbage collection
- "Dynamic filtering" support for GC read/write barriers (ASPLOS 10)
- H/W abstractions for fast and scalable locking

## Low-power vector processors

- New vectorization techniques for cloud computing and mobile applications
- Fusion of Edge and E2 with vector techniques

**More on:** <http://www.bscmsrc.eu/>



# The Microsoft Research-INRIA Joint Centre



- Founded by INRIA (the French National Research Institute for Computer Science and Applied Mathematics), Microsoft Corporation, and the Microsoft Research Laboratory Cambridge
- The Centre's objective is to pursue fundamental, long-term research in formal methods, software security, and the application of Computer Science research to the Computational Science.
- The Joint Centre benefits from the collaboration of 35 researchers from INRIA and other French academic institutions, 25 post docs and PHD students and 15 researchers from Microsoft Research.
- More on: <http://www.msr-inria.inria.fr/>



# The Microsoft Research - University of Trento Centre for Computational and Systems Biology (CoSBi)

- **Goals:** Perform computational system biology using latest HPC technology. Initially a MS HPC cluster was used to carry out simulations of complex biological systems modelled through the techniques developed at CoSBi.
- **Outcome-Impact:** Use of the cluster for time-consuming analyses, especially analyses and simulations that require multiple input data and with different parameters. It was impossible for CoSBi scientists to run this kind of programs before the introduction of the MS HPC cluster in their working environment.
- **Cloud Computing:** moving now to MS Azure cloud computing technology with EU FP7 Venus-C project.
- **More info on <http://www.cosbi.eu/>**



# KU-MSR-BSC

- Focus areas
  - Software engineering, reliability, verification
  - Multicore and multiprocessor systems
- Teams collaborate during the design process:
  - architecture (BSC)
  - systems (MSR)
  - software engineering (KU)
- Software engineering tools for
  - novel multicore architectures
  - novel concurrent programming approaches
- Verification tools early in the design process
  - Not as a late-stage debugging tool only.





# Collaborative Research in Computer Vision with MSU



Dr. Pushmeet  
Kohli,  
MSR Cambridge



Dr. Anton  
Konushin,  
MSU



Dr. Carsten Rother,  
MSR Cambridge



Dr. Olga Barinova.  
MSU



Dr. Victor Lempitsky,  
Yandex/MSU

Undergraduate and PhD students:



Mikhail Sindeev Elena Tretiak Sergey Milyaev Roman Shapovalov Tatiana Novikova

# 2011 Microsoft Computer Vision Summer School in Russia



## Facts & figures:

- 520+ registrations
- 70+ cities
- 80 students selected

The school offered students a unique opportunity to learn about fundamental and state of the art on Computer Vision from top scientists , including Andrew Blake, Andrew Fitzgibbon, Carsten Rother (Microsoft Research, UK), Andrew Zisserman (University of Oxford, UK).



# PhD Scholarship

- Goals
  - Encourage interdisciplinary research
  - Advance the state of the art
  - Create a community
  - Identify potential interns & employees
- Open & competitive
  - Application by research supervisors
  - Selection ratio 17%
  - Up to one year to find best possible students
- More than funding
  - Co-supervisions by MSR researchers
  - Internship
  - Summer School





# MSR Summer Schools

- Networking
  - Other students, MSRC researchers, Cambridge academics
- ‘Transferable skills’
  - Write paper, give talk, becoming an entrepreneur, applying for funding
- Research talks
- Poster sessions
- Social activity



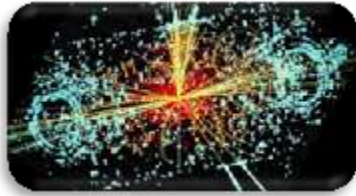


# The Future: an Explosion of Data

Experiments



Simulations



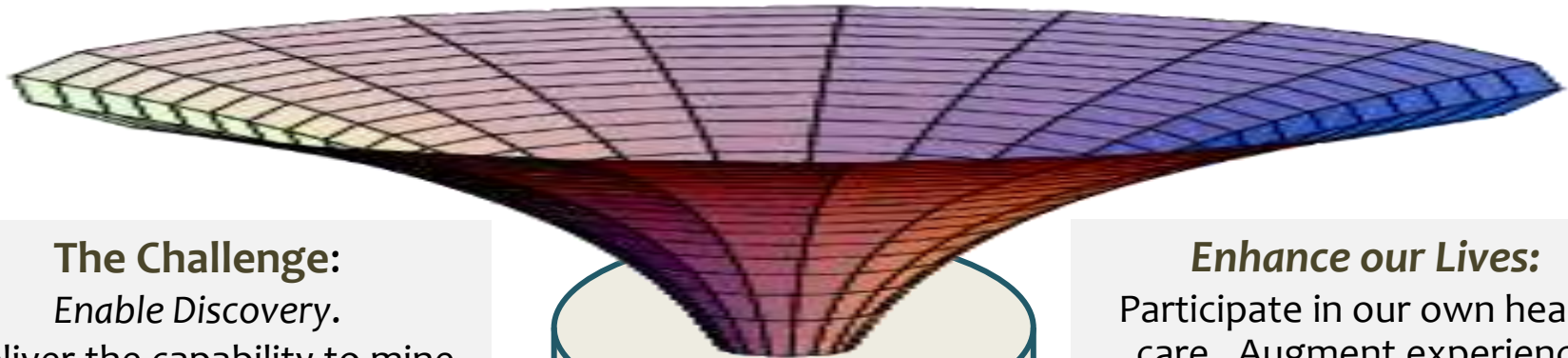
Archives



Literature



Instruments



## The Challenge:

*Enable Discovery.*

Deliver the capability to mine, search and analyze this data in near real time.

## Enhance our Lives:

Participate in our own health care. Augment experience with deeper understanding.

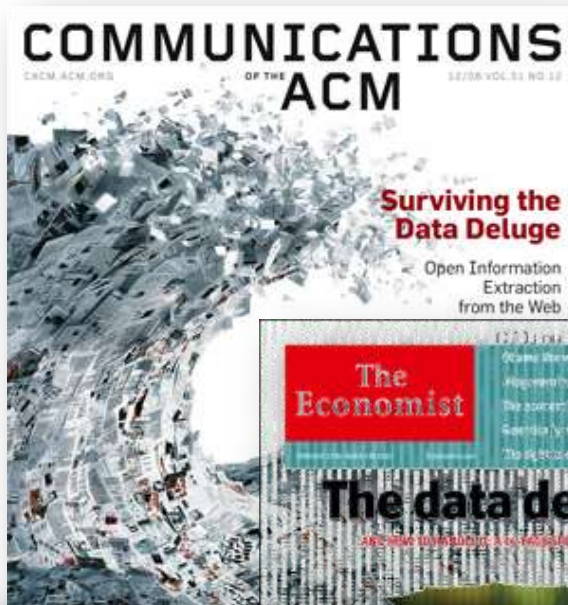
**By 2020, more than 1/3rd of all digital information created annually will either live in or pass through the cloud.**

(Source: EMC-sponsored IDC study)

## Petabytes

Digital information created annually will grow by a factor of 44 from 2009 to 2020

# A Tidal Wave of Scientific Data



# Emergence of a Fourth Research Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations...

Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

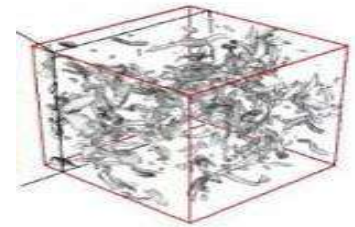
- Scientists overwhelmed with data sets from many different sources
  - Captured by instruments
  - Generated by simulations
  - Generated by sensor networks

eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination



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



*(With thanks to Jim Gray)*



# Changing Nature of Discovery

## Complex models

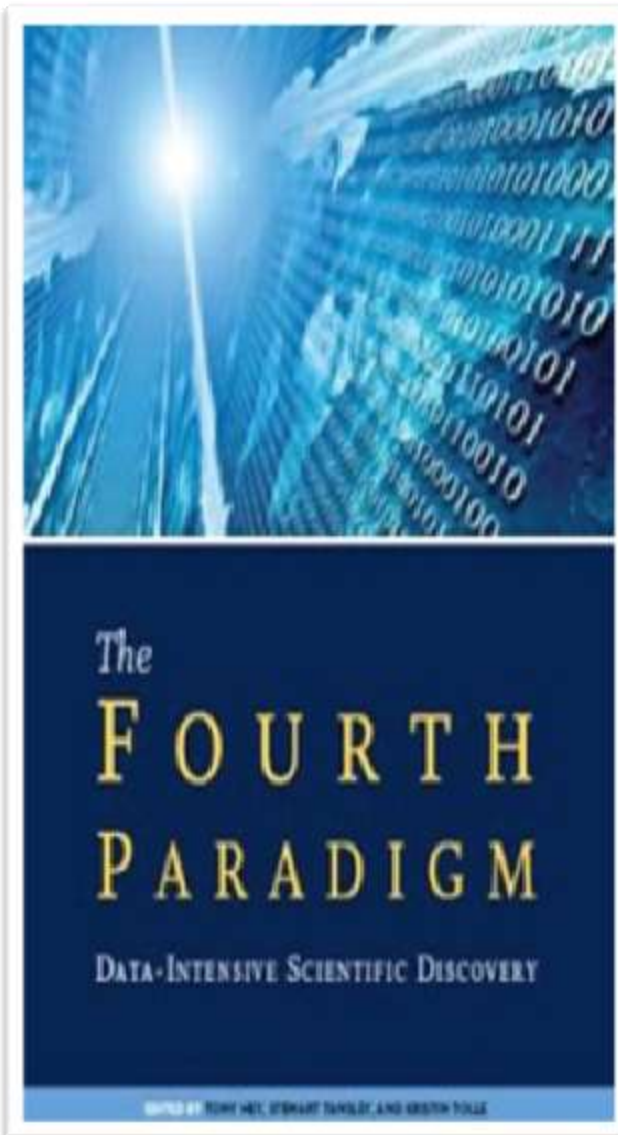
- Multidisciplinary interactions
- Wide temporal and spatial scales

## Large multidisciplinary data

- Real-time streams
- Structured and unstructured

## Distributed communities

- Virtual organizations
- Socialization and management





# Machine Translation: The Statistical Revolution

Instead of hand-coding rules

- Exploit large volumes of existing parallel text
- Learn how words, phrases, and structures translate in context

The Rosetta Stone

From Fort St Julien, el-Rashid (Rosetta), Egypt, Ptolemaic Period, 196 BC

A valuable key to the decipherment of hieroglyphs, the inscription on the Rosetta Stone is a decree passed by a council of priests. It is one of a series that affirm the royal cult of the 13-year-old Ptolemy V on the first anniversary of his coronation.

In previous years the family of the Ptolemies had lost control of certain parts of the country. It had taken their armies some time to put down opposition in the Delta, and parts of southern Upper Egypt, particularly Thebes, were not yet back under the government's control.

Before the Ptolemaic era (that is before about 332 BC), decrees in hieroglyphs such as this were usually set up by the king. It shows how much things had changed from Pharaonic times that the priests, the only people who had kept the knowledge of writing hieroglyphs, were now issuing such decrees. The list of good deeds done by the king for the temples hints at the way in which the support of the priests was ensured.

The decree is inscribed on the stone three times, in hieroglyphic (suitable for a priestly decree), demotic (the native script used for daily purposes), and Greek (the language of the administration). The importance of this to Egyptology is immense.

Soon after the end of the fourth century AD, when hieroglyphs had gone out of use, the knowledge of how to read and write them disappeared. In the early years of the nineteenth century, some 1400 years later, scholars were able to use the Greek inscription on this stone as the key to decipher them.

Thomas Young, an English physicist, was the first to show that some of the hieroglyphs on the Rosetta Stone wrote the sounds of a royal name, that of Ptolemy. The French scholar Jean-François Champollion then realized that hieroglyphs reproduced the sound of the Egyptian language and laid the foundations of our knowledge of ancient Egyptian language and culture.

Soldiers in Napoleon's army discovered the Rosetta Stone in 1799 while digging the foundations of an addition to a fort near the town of el-Rashid (Rosetta). On Napoleon's defeat, the stone became the property of the British under the terms of the Treaty of

Piedra Rosetta

Origen: Fuerte de San Julián, el-Rashid (Rosetta), Egipto  
Período ptolemaico, 196 a.C.

Pieza clave para descifrar jeroglíficos

El texto contenido en la Piedra Rosetta corresponde a un decreto dictado por un consejo de sacerdotes e integra una serie de decretos que ratifican el culto real de Ptolomeo V, de 13 años de edad, en el primer aniversario de su coronación.

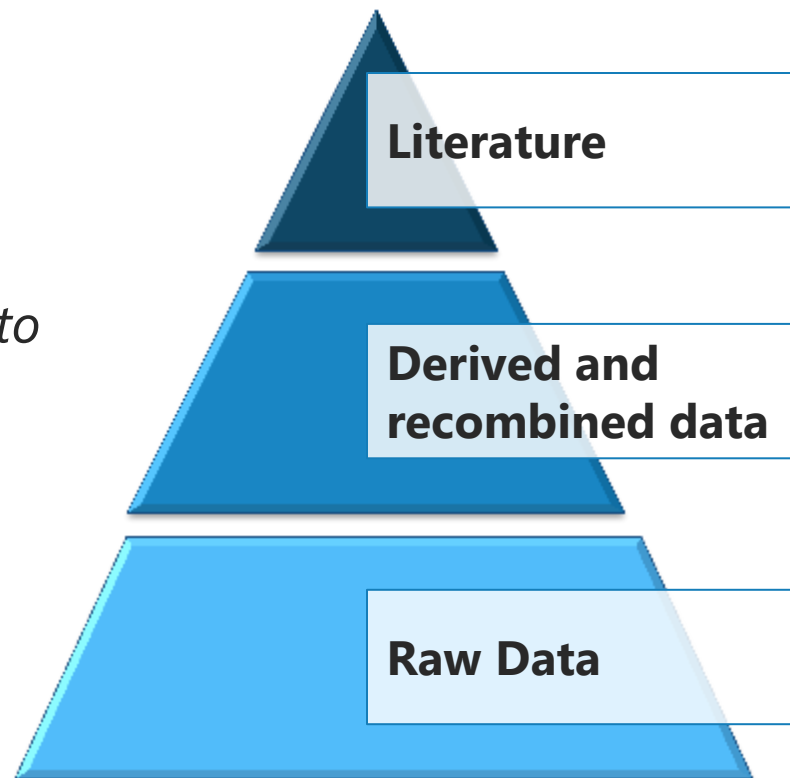
En años anteriores, la dinastía ptolemaica había perdido el control de ciertas zonas del país. Después de un largo tiempo, su ejército logró derrocar a la oposición en el Delta, pero la región sur del Alto Egipto. Tras un suceso, no había sido aún recuperada por el gobierno.

Antes de la era ptolemaica (hasta cerca del año 332 a.C.), el rey solía emitir decretos en jeroglíficos como el de esta pieza. Este dato da cuenta de cómo cambiaron las cosas desde los tiempos faraónicos, ya que los sacerdotes, las únicas personas que conocían la escritura jeroglífica, pasaron a emitir dichos decretos. La cantidad de actos reales condescendientes con los templos nos ilustra la forma en la cual se garantizaba el apoyo de los sacerdotes.

El decreto está escrito en la piedra por portada triple, en jeroglífico (acorde a un decreto sacerdotal), en demótico (la escritura nativa de uso diario) y en griego (el idioma del gobierno). Su importancia para la etimología es enorme. Al poco tiempo del final del s. IV a.C., cuando se dejaron de utilizar jeroglíficos, el conocimiento sobre cómo leerlos y escribirlos se perdió. A comienzos del s. XIX, unos 1400 años después, los científicos lograron descifrarlos utilizando

# All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips –  
For everyone, everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



*(From Jim Gray's last talk)*

# The Cloud

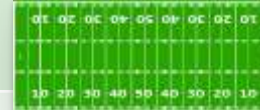
- A model of computation and data storage based on “pay as you go” access to “unlimited” remote data center capabilities
- A cloud infrastructure provides a framework to manage scalable, reliable, on-demand access to applications
- A cloud is the “invisible” backend to many of our mobile applications
- Historical roots in today’s Internet apps and previous DCI computing (Cluster, Grid etc.)



# The Cloud is built on massive data centers

Essentially driven by economies of scale

- Approximate costs for a small size center (1K servers) and a larger, 100K server center.

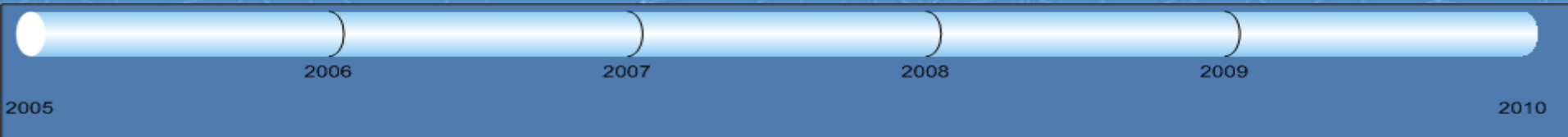


Technology	Cost in small-sized Data Center	Cost in Large Data Center	Ratio
Network	\$95 per Mbps/ Month	\$13 per Mbps/ month	7.1
Storage	\$2.20 per GB/ Month	\$0.40 per GB/ month	5.7
Administration	~140 servers/ Administrator	>1000 Servers/ Administrator	7.1

Each data center is  
**11.5 times**  
the size of a football field



# Microsoft's Datacenter Evolution



Datacenter Co-Location  
Generation 1

Quincy and San Antonio  
Generation 2

Chicago and Dublin  
Generation 3

Modular Datacenter  
Generation 4



**Facility PAC**

Deployment Scale Unit



**Server**

*Capacity*



**Rack**

*Density and Deployment*



**Containers**

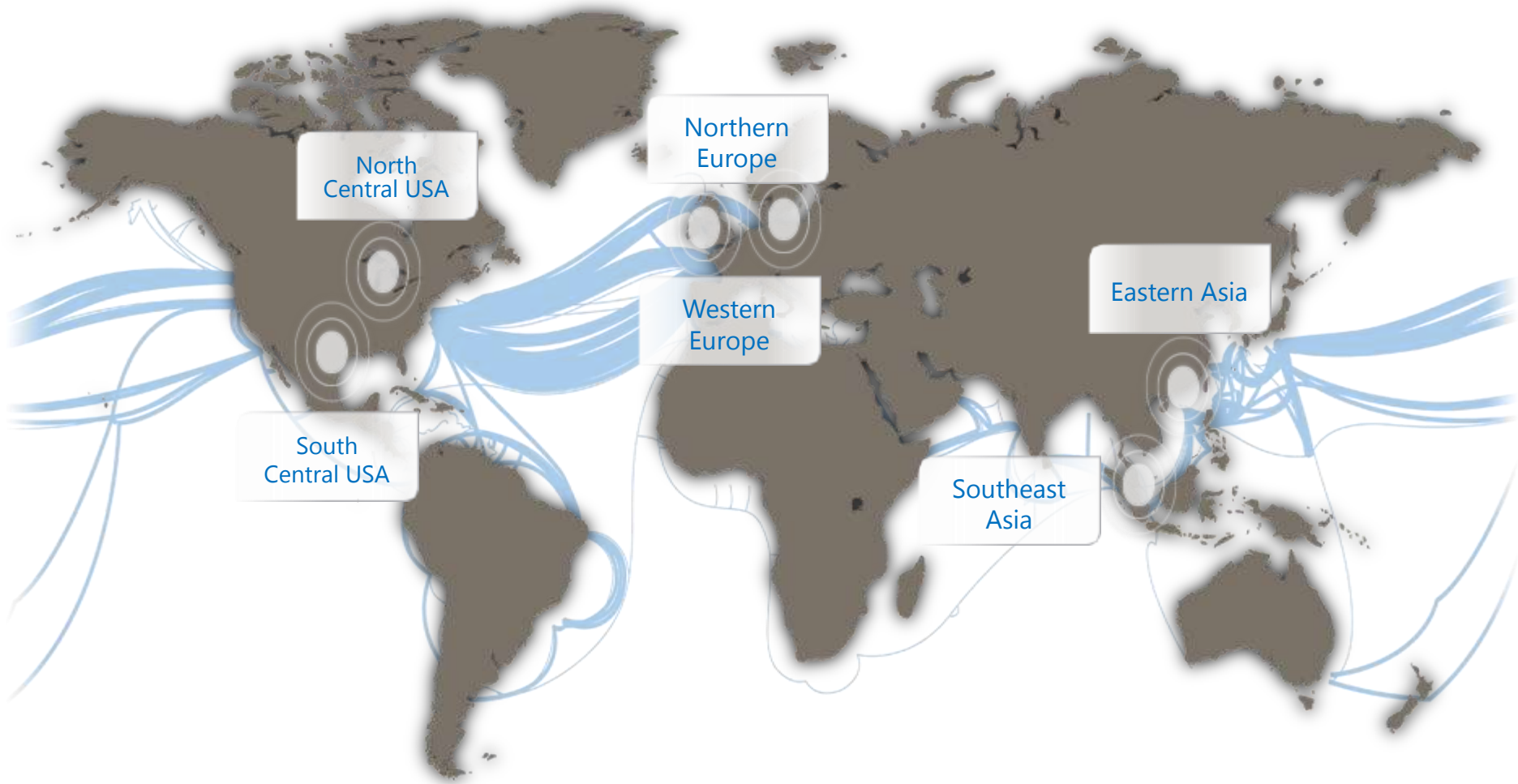
*Scalability and  
...Sustainability*



**IT PAC**

*Time to Market  
Lower TCO*

# Windows Azure Platform Availability



# Major Motivations

- Environmental responsibility
  - Managing energy efficiently
  - Adaptive systems management
- Provisioning 100,000 servers
  - Hardware: at most one week after delivery
  - Software: at most a few hours
- Resilience during a blackout/disaster
  - Service rollover for millions of customers
- Software and services
  - End-to-end communication
  - Security, reliability, performance, reliability

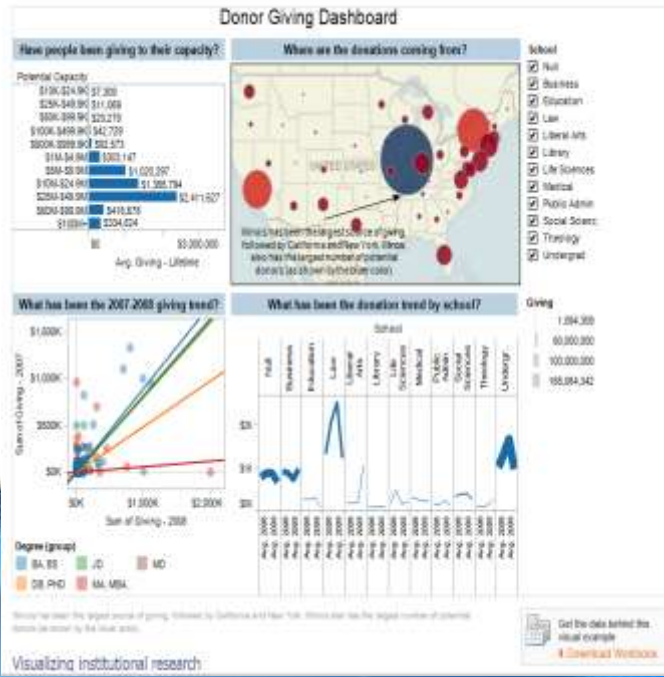




# Focus Client + Cloud for Research

## Seamless interaction

- Cloud is the lens that magnifies the power of desktop
- Persist and share data from client in the cloud
- Analyze data initially captured in client tools, such as Excel
  - Analysis as a service (think SQL, Map-Reduce, R/MatLab)
  - Data visualization generated in the cloud, display on client
  - Provenance, collaboration, other 'core' services...





# Simple Tools to Answer Complex Questions...

**Imagine: the client plus the invisible backend for problem solving**

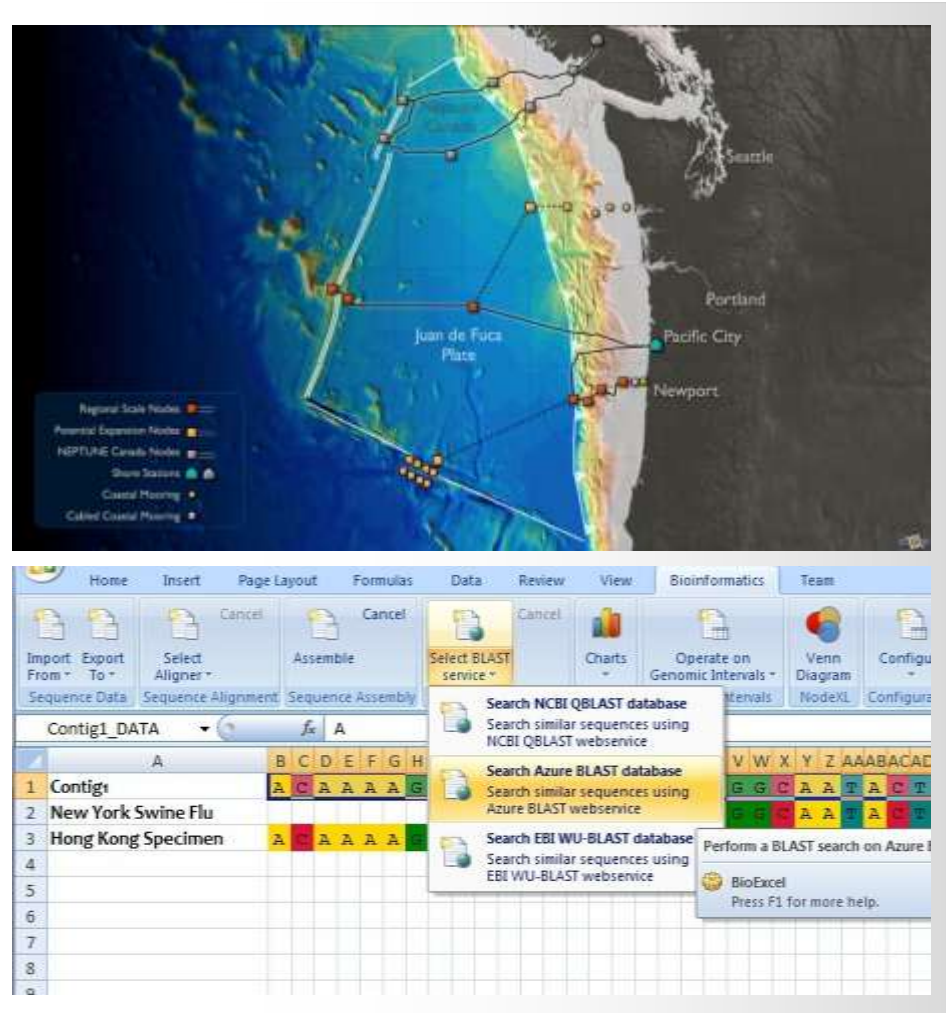
Give the standard science and engineering desktop tools a seamless extension

Use a spreadsheet to invoke genomic analysis tools running on 600 servers

Use a simple script to orchestrate data analytics and mining across 10000 MRI Images

Pull data from remote instruments for visualization on the desktop

**Create a revolution in scientific capability for everybody**



# Extend the research footprint

## Today

### Majority of Researchers

Use laptops and desktop computers

Overwhelmed by data

Finding analysis ever more difficult; sharing even harder

HPC users

Those with small clusters or servers

Majority of Researchers

## Tomorrow?

### Paradigm Shift

Powerful tools

Data and analysis tools in the cloud

Cycles, storage, support

Building communities around research results

The ability to marshal needed resources on demand  
Without caring or knowing how it gets done...

Accelerating discovery



VENUS-C



Virtual multidisciplinary EnviroNments USing Cloud infrastructures



# European Cloud Computing Strategy

## Three Pillars for Cloud

- *Legal frameworks*
- *Technical and commercial fundamental elements*
- *Development of the cloud market by supporting pilot projects of cloud deployments*



Vice-President Neelie Kroes,  
responsible for the Digital Agenda

Official opening of the Microsoft  
Cloud & Interoperability Center,  
March 2011



**Cloud Power**

**Neelie Kroes on international standardisation & open specifications**

*"I count here on the further support and commitment of Microsoft and all the other participants."*



# Industry contribution to the European Cloud Strategy

Building an industry-quality, highly scalable & flexible Cloud infrastructure

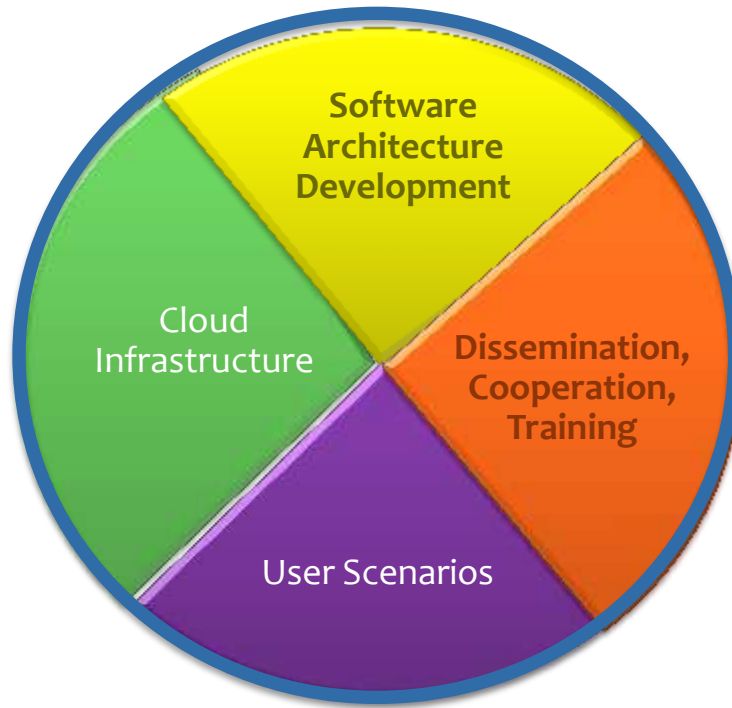
**ENGINEERING INGENIERIA INFORMATICA**

**Microsoft**  
EMIC-MICGR-MRL

**KTH**  
KTH VETENSKAP OCH TEKNIK

**BSC**

**OGF.eeig**  
European Federation of Open Grids



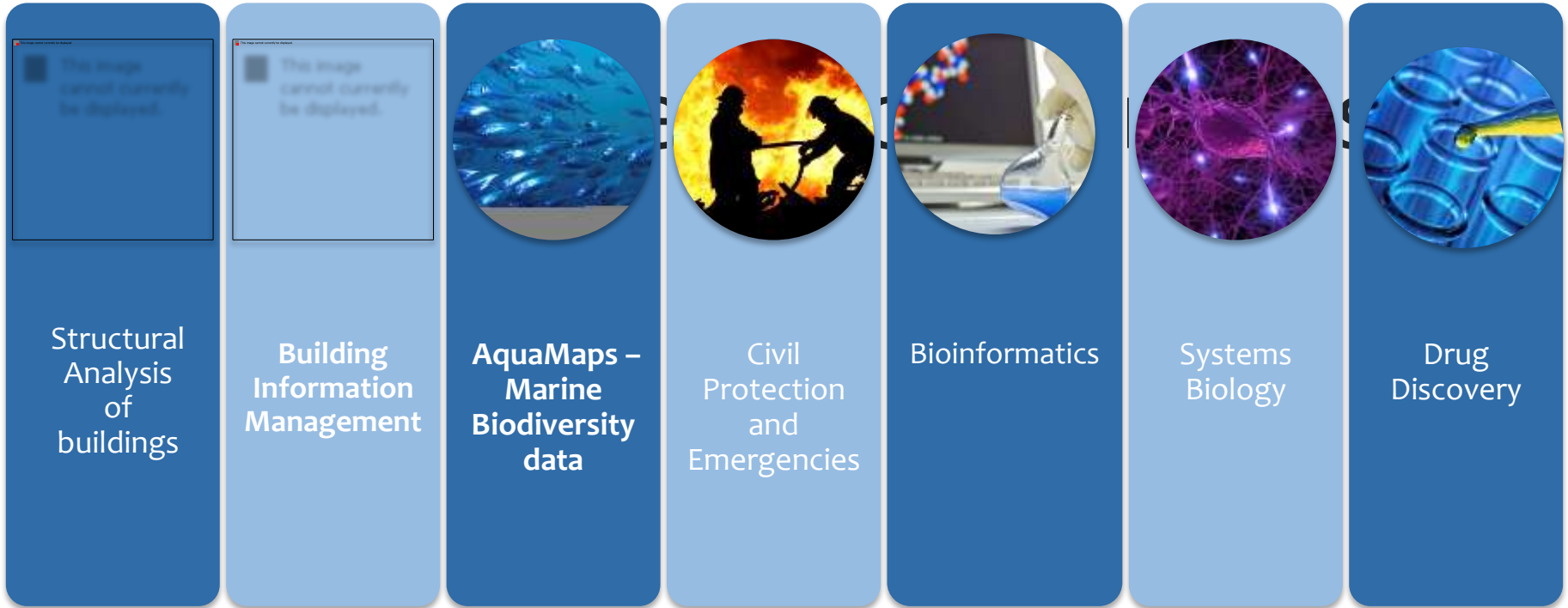
**UNIVERSIDAD POLITÉCNICA VALENCIA**

**COSBi**

Coordinated by Engineering – Investment in infrastructure provision & software development.  
 Microsoft invests in Azure resources & manpower through Redmond & its European data centres

# A user-centric Approach

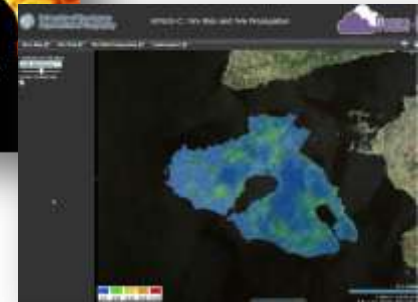
Building a Cloud Infrastructure with user needs interwoven  
Bringing about fundamental changes in scientific discovery & innovation



# Some Success Stories

- Interactive computation of fire risk and fire propagation estimation
- Access to burst-scalable cloud compute and storage
- Web-based GIS based on Bing Map

## [Wild Fire Demo](#)



- Collaboratorio & its new start-up Green Prefab
- Collaborative platform for the design of ecofriendly & affordable buildings
- Selected by INTESA SAN PAOLO Start-up initiative; expanding to US

Real-estate  
Investor

Designer

Engineer

Producer  
of building  
elements

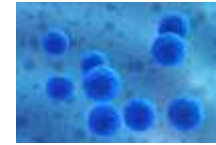
Contractor



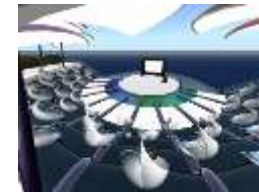
*“We feel like pioneers in the right direction to the still untouched gold mine,”* Furio Barzon

# Extending Cloud Usage - Pilots & Experiments

## Engineering & Science



Architecture & Civil  
Engineering  
Biology



## NEW DISCIPLINES

Earth Sciences, Healthcare, Maths, Mechanical Engineering, Physics,  
Social Media, Education

## Start-ups



*Computer resources can be scaled as required without committing to large capital purchases, which is critical to the success of our small business. **Molplex UK***



*DFRC is part of the EU Flagship project PERSEUS on maritime security. Scaling our platform with VENUS-C will enable us to support future growth in terms of vessels monitored in real time & usability by operators.*



# Value-add for eScience

- Distributing, managing and curating data is better served by a virtual, scalable and elastic infrastructure
- Economy of scale, energy costs and environmental impact are better addressed by Cloud computing
- Virtualisation of computing infrastructures can support funding agencies in developing new funding models:
  - Moving from CAPEX to OPEX
- Leading to more science per tax payer €
- Faster to deploy than conventional HPC in emerging scientific and business communities

**Thank you**

**?**

# Resources

- Microsoft Research
  - <http://research.microsoft.com>
  - Microsoft Research downloads:  
<http://research.microsoft.com/research/downloads>
- Microsoft External Research
  - <http://research.microsoft.com/en-us/collaboration/>
- Science at Microsoft
  - <http://www.microsoft.com/science>
- Scholarly Communications
  - <http://www.microsoft.com/scholarlycomm>
- CodePlex
  - <http://www.codeplex.com>



# References

- The Fourth Paradigm: Data-Intensive Scientific Discovery – Tony Hey (Editor), Stewart Tansley (Editor)
- How Will Astronomy Archives Survive the Data Tsunami? Communications of the ACM Vol. 54 No. 12
- Data-Intensive Science: A New Paradigm for Biodiversity Studies - Steve Kelling, Wesley M. Hochachka, Daniel Fink, Mirek Ried, Ewald Rich Caruana, Grant Ballard, Giles Hooker. Bioscience, 2009.
- Data-intensive e-science frontier research – HB Newman, MH Ellisman... - Communications of the ACM, 2003
- Data-intensive computing in the 21st century – I. Gorton, P Greenfield, A Szalay... - IEEE Computer, 2008