

# Data Intensive Science and Cloud Computing

Fares ZEKRI

Microsoft



Microsoft  
**Research** Connections

# Introduction



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

- Developing HPC distributed computing solutions for HEP
- Including 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 **EUMEDGrid** (2006-2007): where Morocco was represented by **MaGrid** ([www.magrid.ma](http://www.magrid.ma)), as explained by Prof. Rajaa C. El Moursli, and hosted 1er Workshop & Tutorial in Marrakech (December 2006) followed by **EUMEDGRID-Support** which recently held its EUMED4 event co-located with e-Age2011 in Amman December 2011

# Now: @ Microsoft Research Connections

## **Mission Statement:**

*Advancing multidisciplinary research worldwide by engaging and partnering with the Academic community, focusing on:*

**Breakthrough research and innovation;**

**Worldwide participation;**

**Community engagement;**

**Broad dissemination across;**

**Interoperability**





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

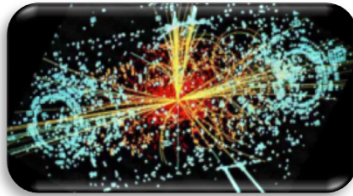


# The Future: an Explosion of Data

Experiments



Simulations



Archives



Literature



Instruments



## The Challenge:

*Enable Discovery.*

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

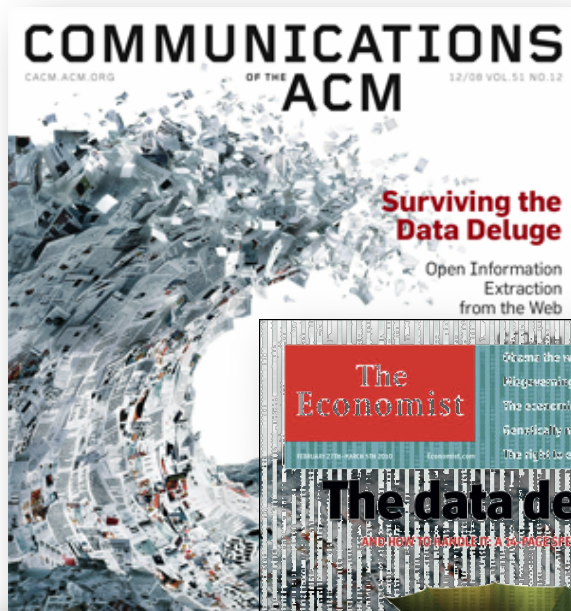
**By 2020, more than 1/3rd of all digital information created annually will either live in or pass through the cloud.**

(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**

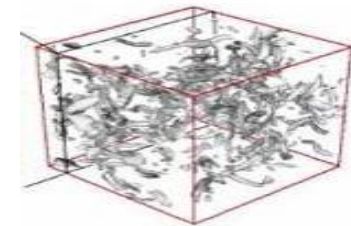
- Simulation of complex phenomena

Today – **Data-Intensive Science**

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



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



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



*(With thanks to Jim Gray)*





# Changing Nature of Discovery

## Complex models

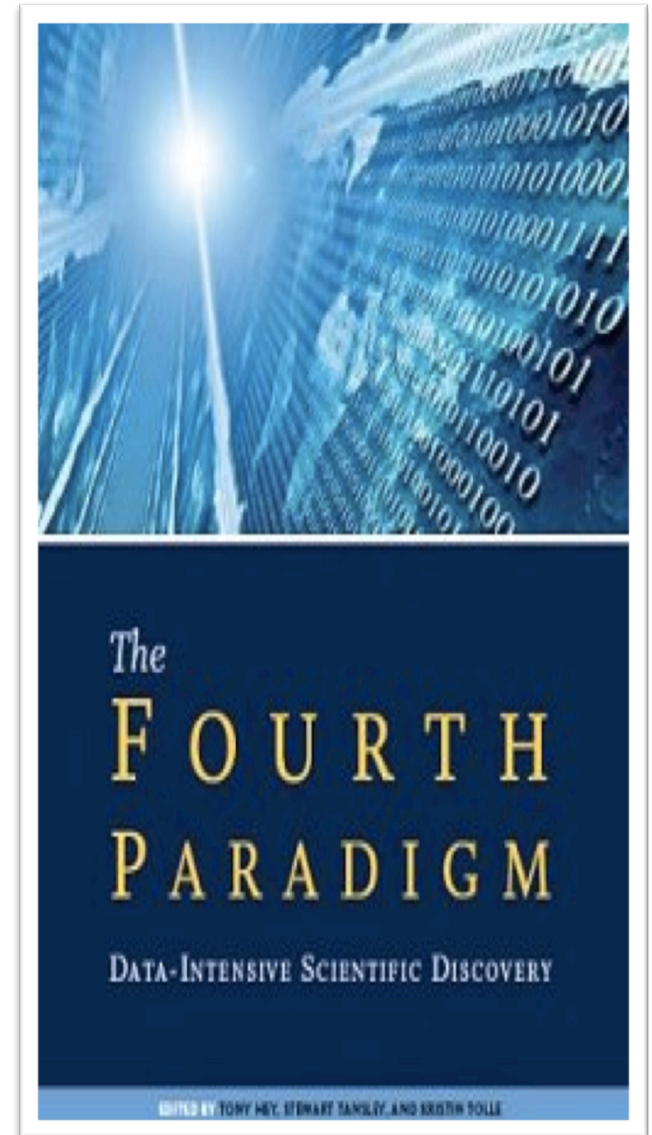
- Multidisciplinary interactions
- Wide temporal and spatial scales

## 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 > The British Museum

THE BRITISH MUSEUM

Home Visiting What's on **Explore** Research Learning The Museum Join in Shop online

Introduction Themes **Highlights** World cultures Online tours Galleries Young explorers

Home > Explore > Highlights

## 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 recorded 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

**On display**

8 4 Egyptian sculpture Room 4 View floorplan ▶

British Museum - Piedra Rosetta

THE BRITISH MUSEUM

Home Visiting What's on **Explore** Research Learning The Museum Join in Shop online

Introduction Themes Highlights World cultures Online tours Galleries Young explorers

Home > Explore > Highlights

## Explore / Highlights

English | **Français** | Italiano

## 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, Tebas en especial, no había sido aun 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 partida 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

**Larger image**  
**Use digital image**  
**Print record**

**Highlights**

Search:

Browse or search over 4,000 highlights from the Museum collection

**Related objects**

Busto colosal de Ramsés II

1 of 10 Objects See

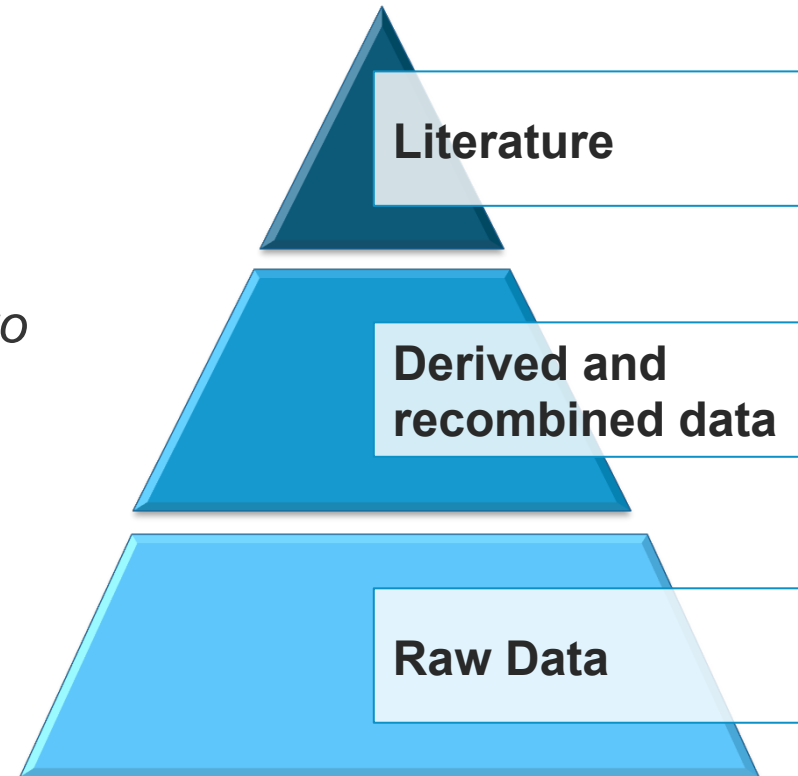
**Shop online**

Rosetta Stone plaque, €35.00

Visit the online shop

# 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.

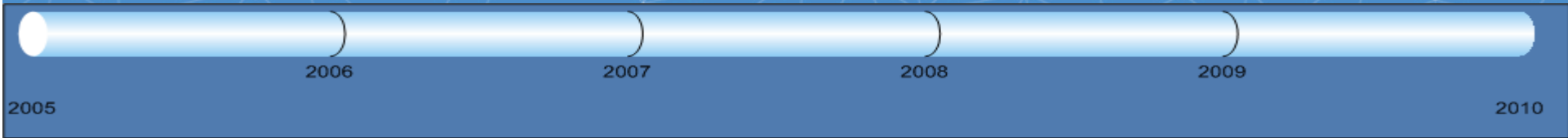
Technology	Cost in small-sized Data Center	Cost in Large Data Center	Ratio
Network	\$95 per Mbps/ Month	\$13 per Mbps/ month	7.1
Storage	\$2.20 per GB/ Month	\$0.40 per GB/ month	5.7
Administration	~140 servers/ Administrator	>1000 Servers/ Administrator	7.1



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

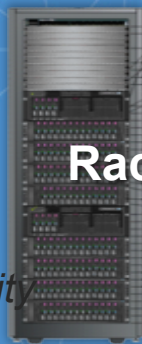


**Facility PAC**

Deployment Scale Unit



**Server**



**Rack**

*Capacity*

*Density and Deployment*



**Containers**

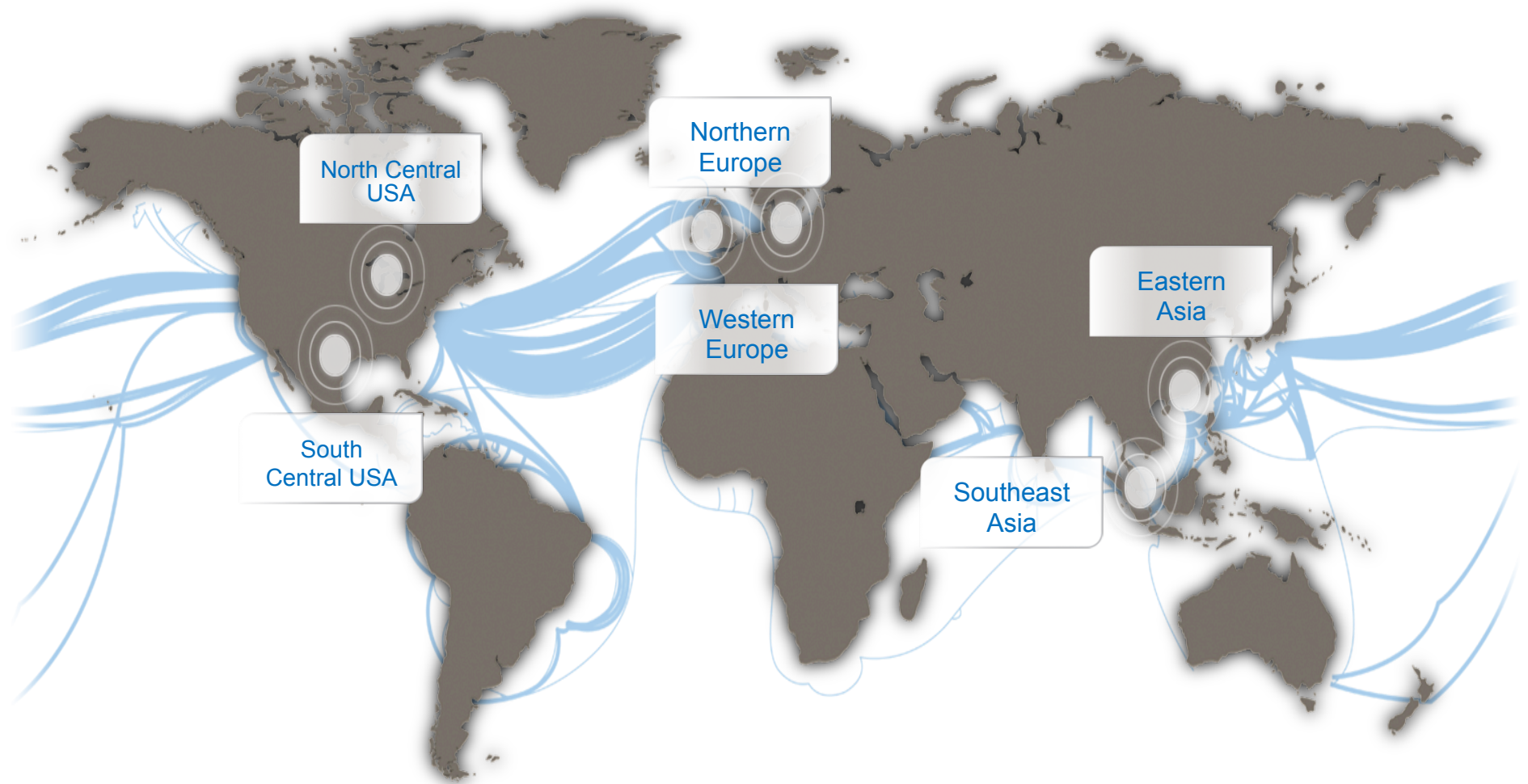
*Scalability and Sustainability* ...



**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

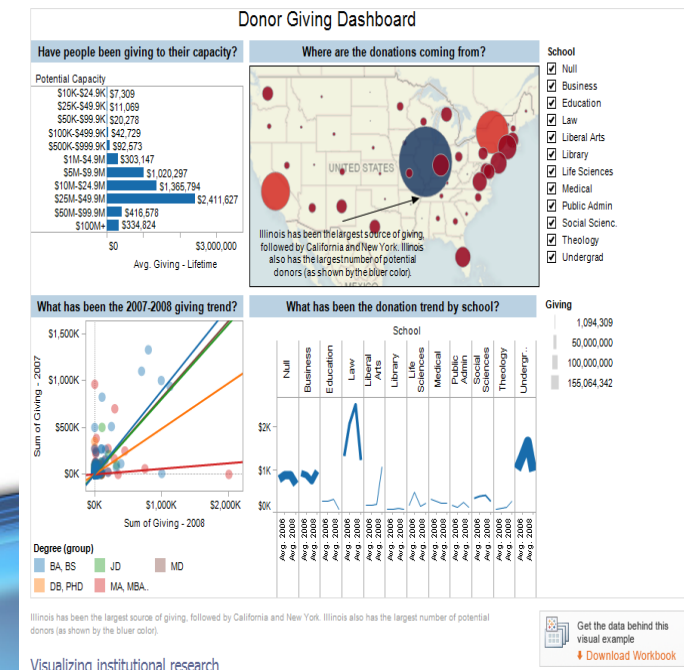
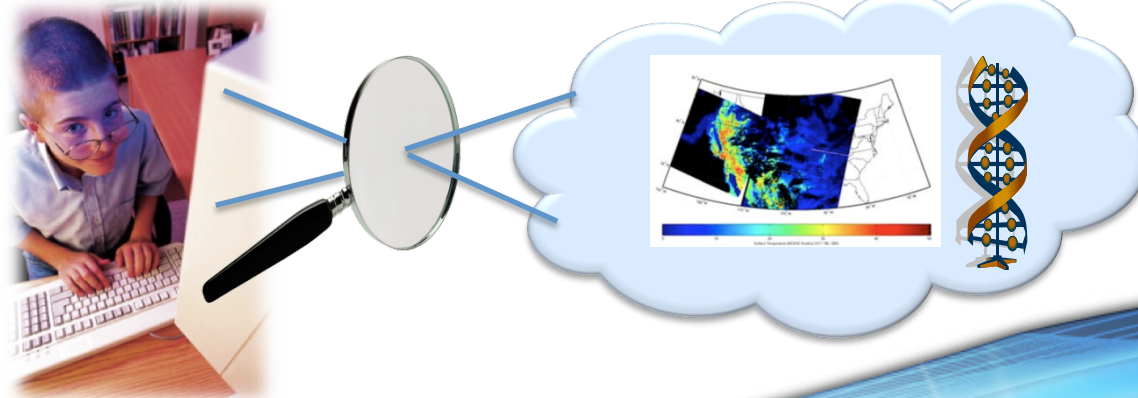




# Focus Client + Cloud for Research

## 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

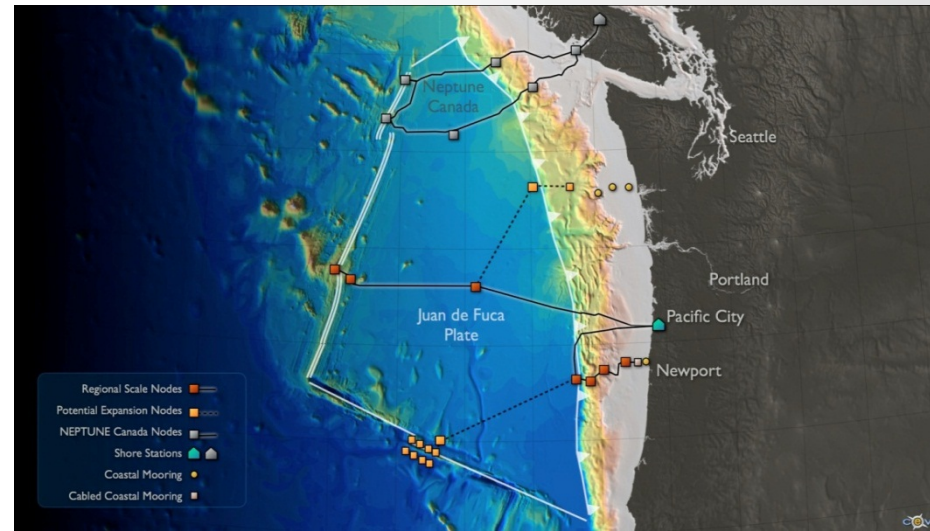
Give the standard science and engineering desktop tools a seamless extension

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



Home Insert Page Layout Formulas Data Review View Bioinformatics Team

Import From Export To Select Aligner Assemble Select BLAST service Charts Operate on Genomic Intervals Venn Diagram Configure

Contig1\_DATA fx A

	A	B	C	D	E	F	G	H
1	Contigi	A	C	A	A	A	A	G
2	New York Swine Flu							
3	Hong Kong Specimen	A	C	A	A	A	A	G
4								
5								
6								
7								
8								

Search NCBI QBLAST database  
Search similar sequences using NCBI QBLAST webservice

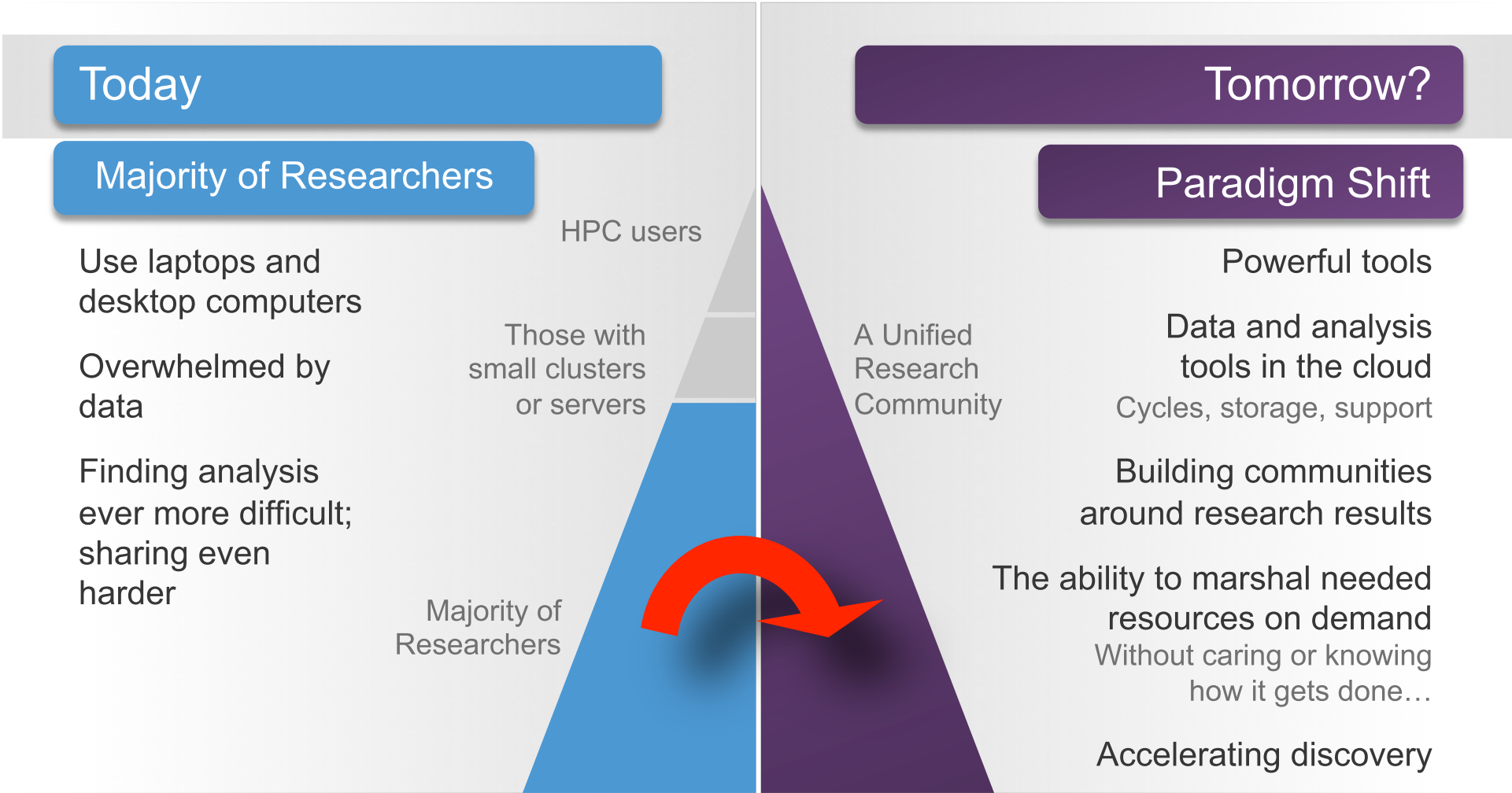
Search Azure BLAST database  
Search similar sequences using Azure BLAST webservice

Search EBI WU-BLAST database  
Search similar sequences using EBI WU-BLAST webservice

Perform a BLAST search on Azure I

BioExcel  
Press F1 for more help.

# Helping Democratisise Research



VENUS-C



Virtual multidisciplinary EnviroNments USing Cloud infrastructures

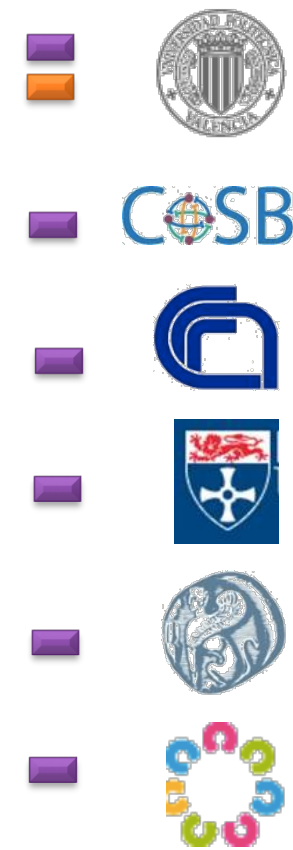
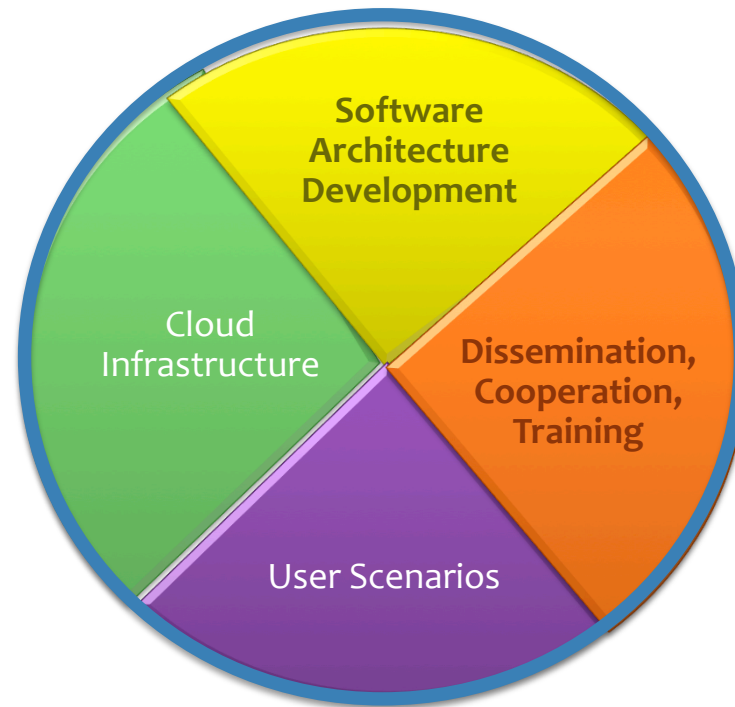


# Industry contribution to the European Cloud Strategy

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



Microsoft  
EMIC –  
MICGR -  
MRL

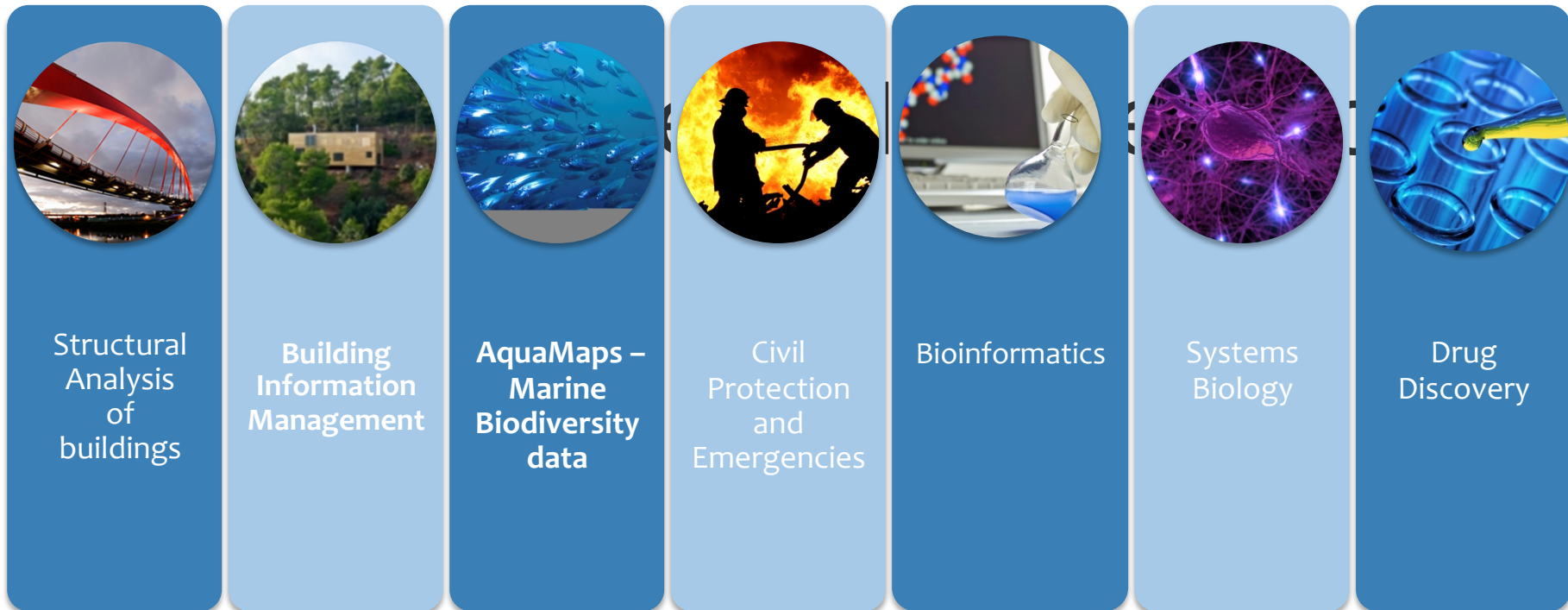


Coordinated by Engineering – Investment in infrastructure provision & software development.  
Microsoft invests in Azure resources & manpower through Redmond & its European data centres



# A user-centric Approach

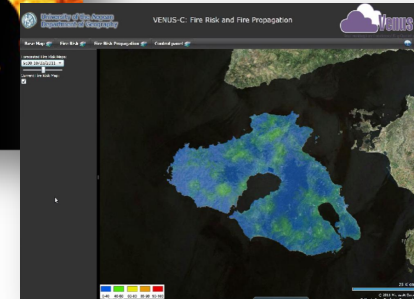
Building a Cloud Infrastructure with user needs interwoven  
Bringing about fundamental changes in scientific discovery & innovation



# 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

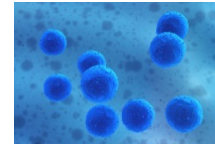


*“We feel like pioneers in the right direction to the still untouched gold mine,”* Furio Barzon

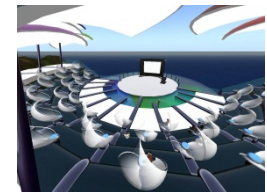
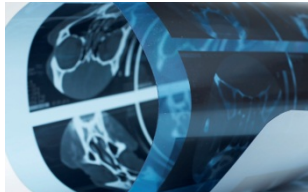


# Extending Cloud Usage - New Pilots & Experiments

## Engineering & Science



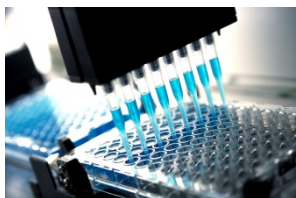
Architecture & Civil  
Engineering  
Biology



## 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 infrastructure and funding agencies support
- Leading to more science per tax payer €
- Faster to deploy than conventional HPC in emerging economy





- Happy to discuss how to move forward and explore this new Cloud computing approach for science in this region
- Please do not hesitate to contact Fabrizio Gagliardi at :

[fabrig@microsoft.com](mailto:fabrig@microsoft.com)



**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>

