Data Intensive Science and Cloud Computing

Fares ZEKRI

Microsoft

Microsoft Research Connections

Introduction

From 1975 till 2005: Computing Science at CERN (www.cern.ch)

- Developing HPC distributed computing solutions for HEP
- Including the present LHC distributed computing Grid infrastructure (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), 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;

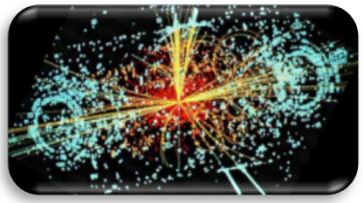
Interoperability

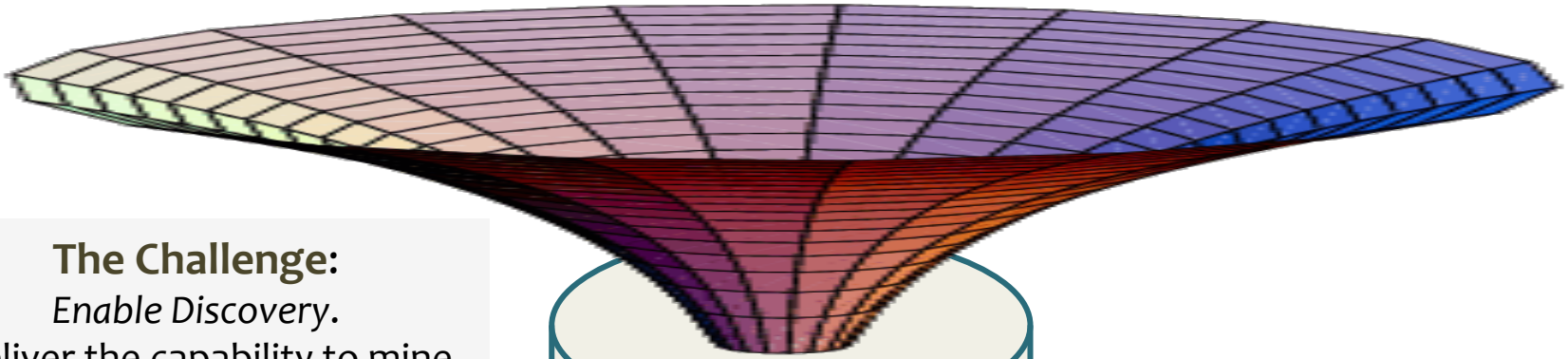
Microsoft Research

Microsoft

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…

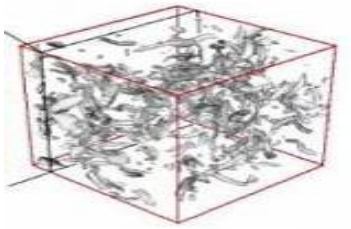
Last few decades – **Computational Science**

- 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

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



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

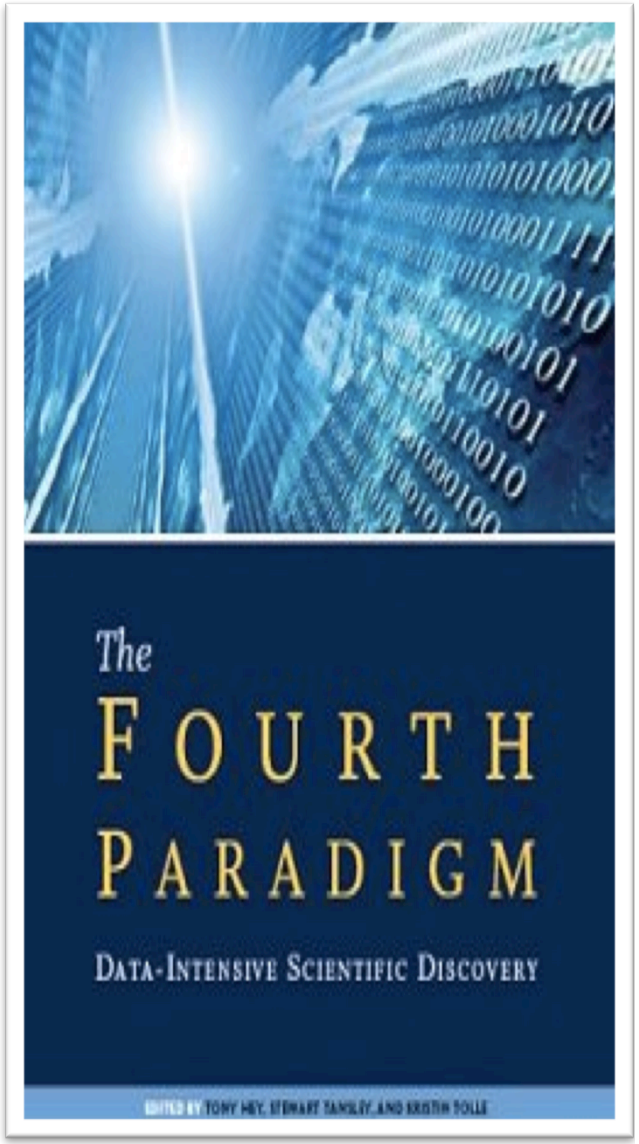
- Multidisciplinary interactions
- Wide temporal and spatial scales

Large multidisciplinary data

- Real-time steams
- Structured and unstructured

Distributed communities

- Virtual organizations
- Socialization and management

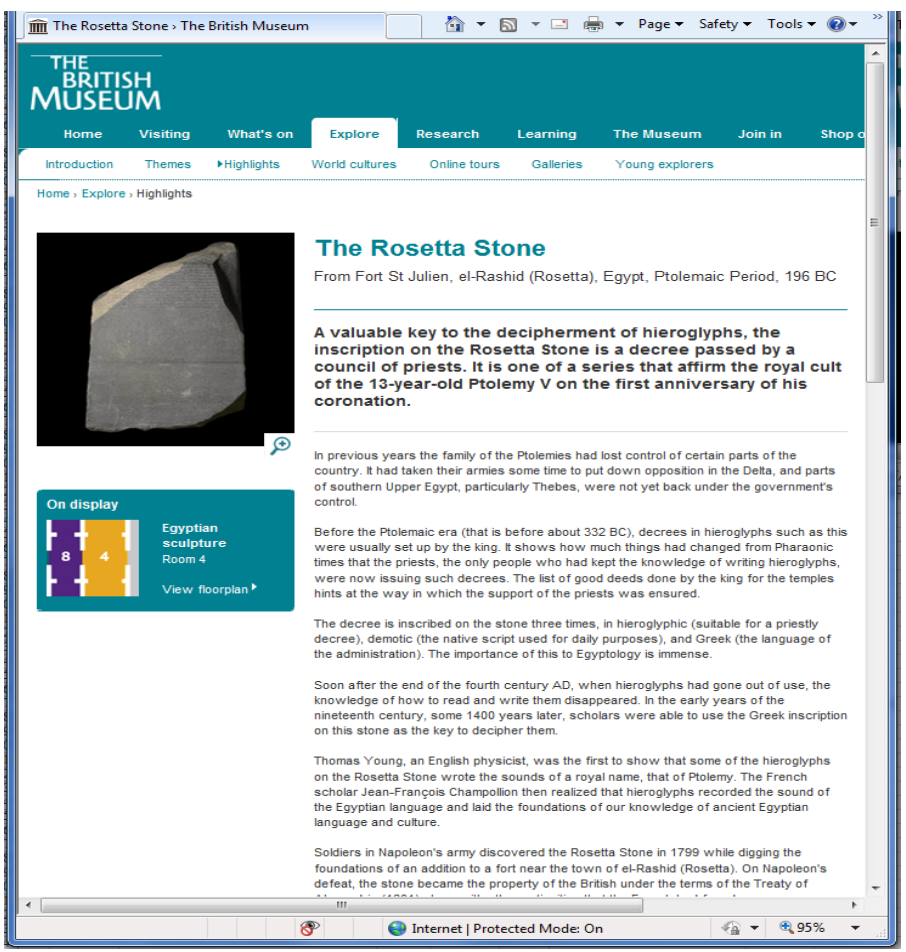


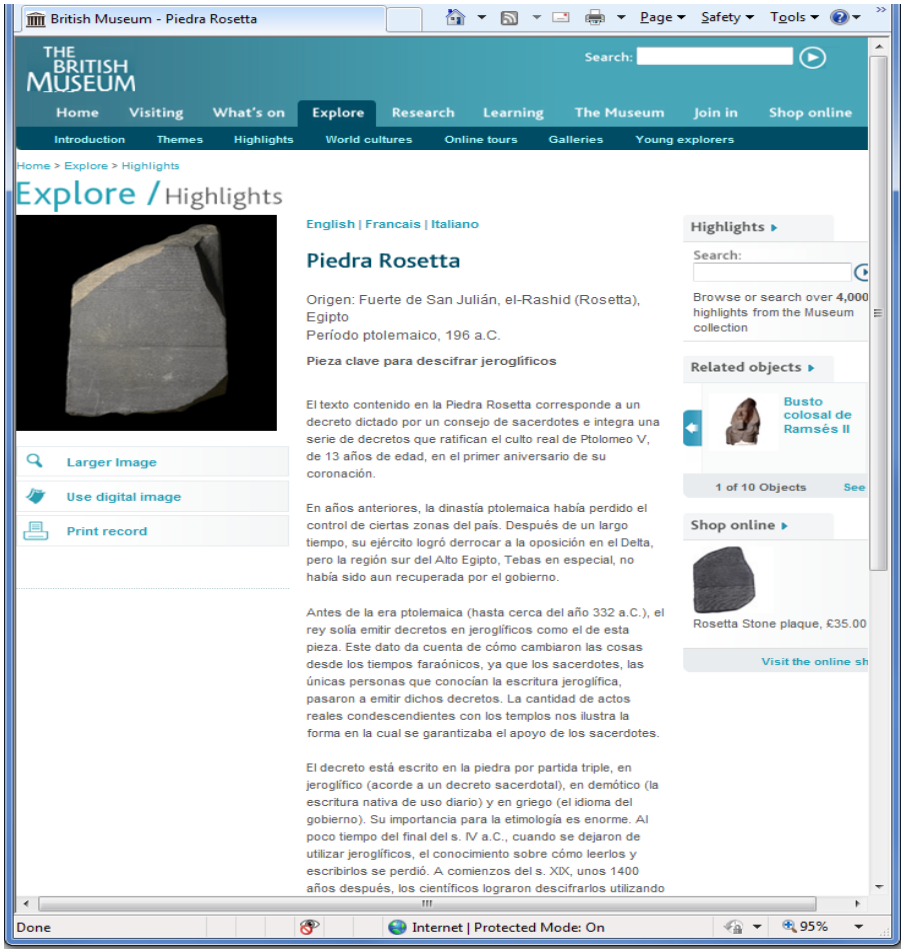
http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Machine Translation: The Statistical Revolution

Instead of hand-coding rules

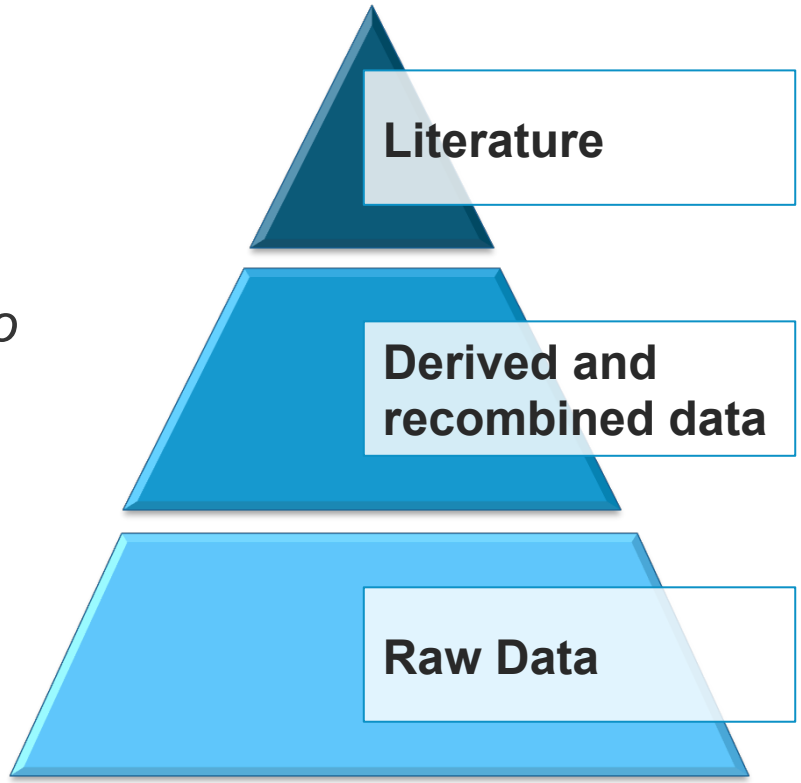
- Exploit large volumes of existing parallel text
- Learn how words, phrases, and structures translate in context





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

Microsoft's Datacenter Evolution

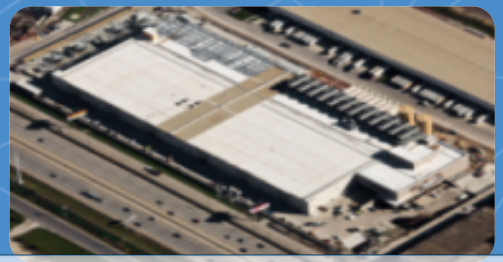
2005 2006 2007 2008 2009 2010

Datacenter Co-Location Generation 1

Quincy and San Antonio Generation 2

Chicago and Dublin Generation 3

Modular Datacenter Generation 4



Deployment Scale Unit

Capacity

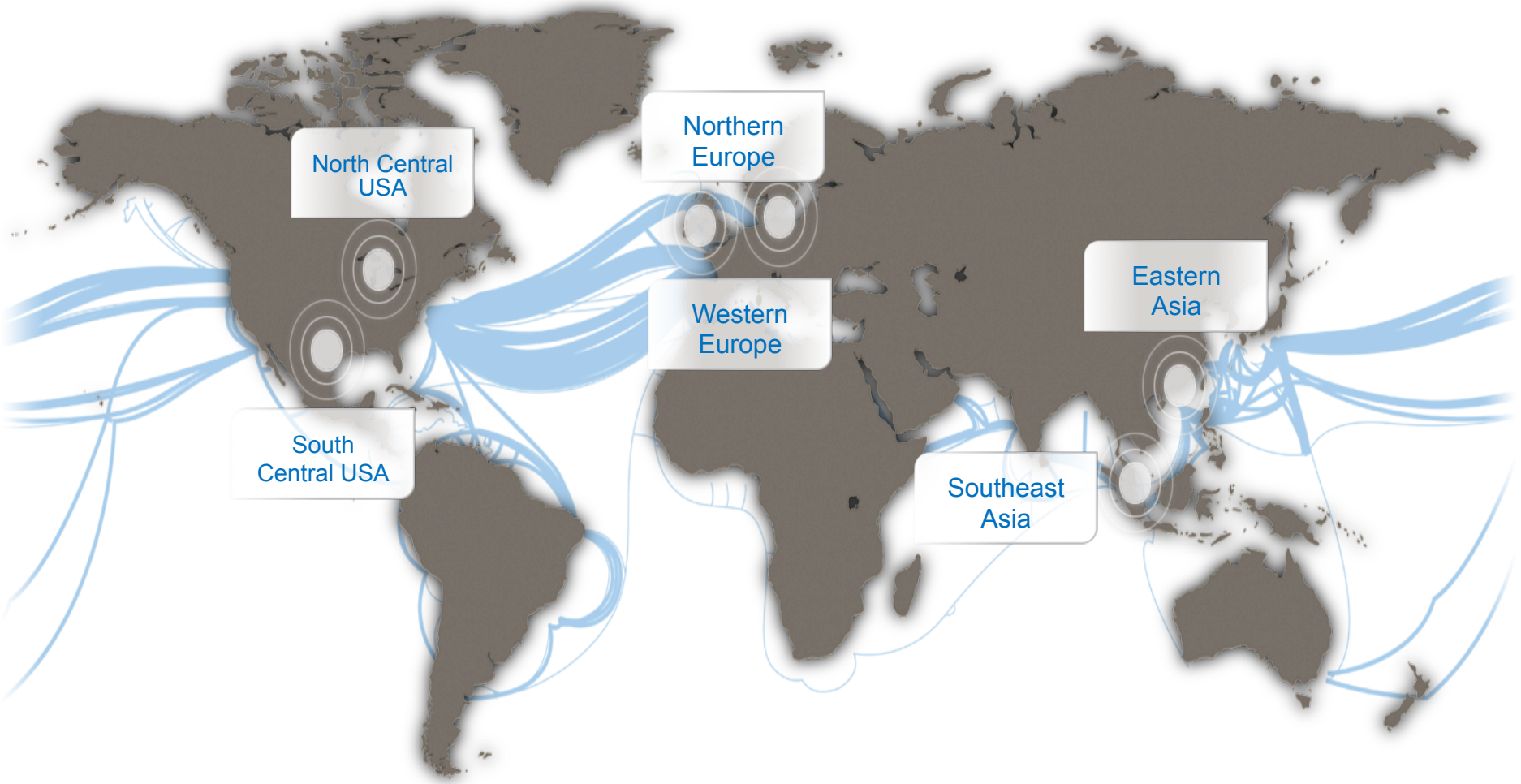
Density and Deployment

Scalability and Sustainability

…

Time to Market Lower TCO

Windows Azure Platform Availability



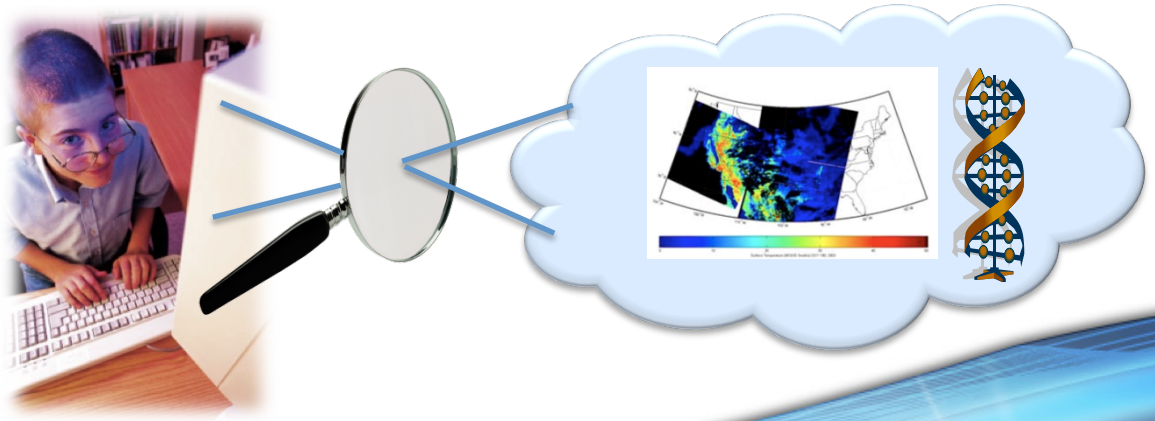
Major Motivations

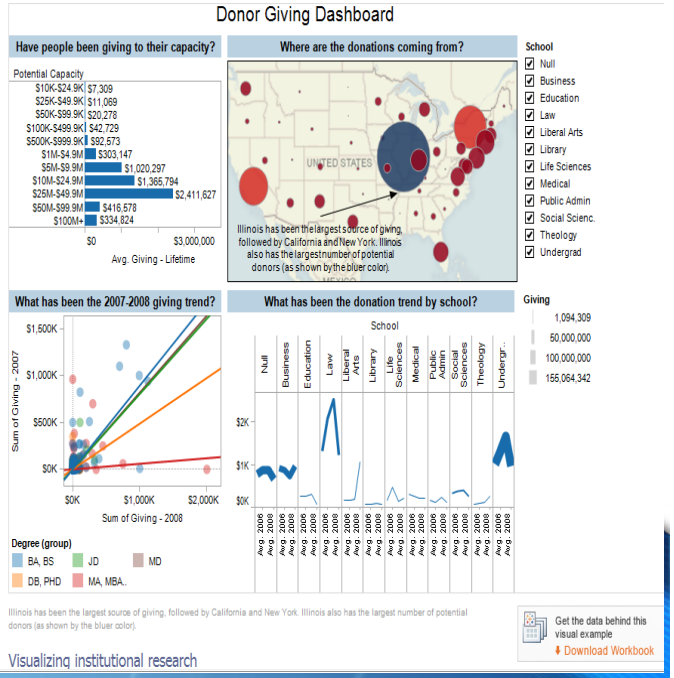
- Environmental responsibility
  - Managing energy efficiently
  - Adaptive systems management
- Provisioning 100,000 servers
  - Hardware: at most one week after delivery
  - Software: at most a few hours
- Resilience during a blackout/disaster
  - Service rollover for millions of customers
- Software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

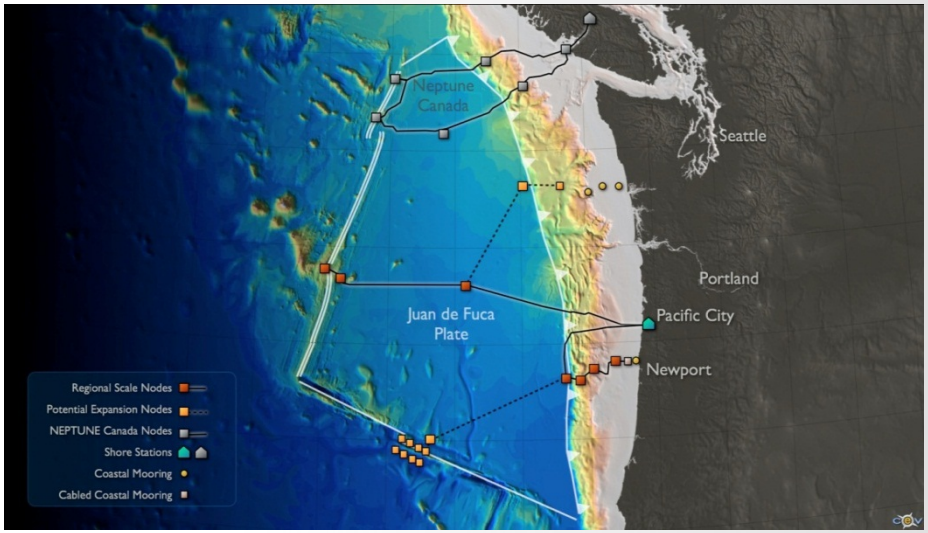
Give the standard science and engineering desktop tools a seamless extension

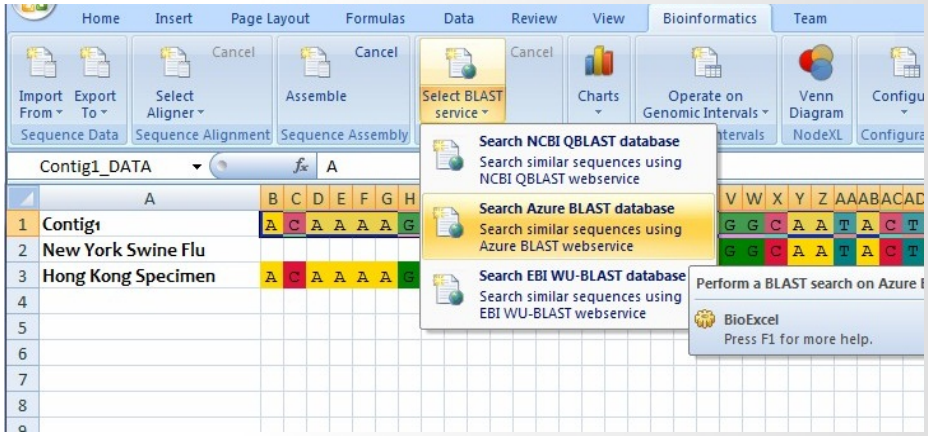
Use a spreadsheet to invoke genomic analysis tools running on 600 servers

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

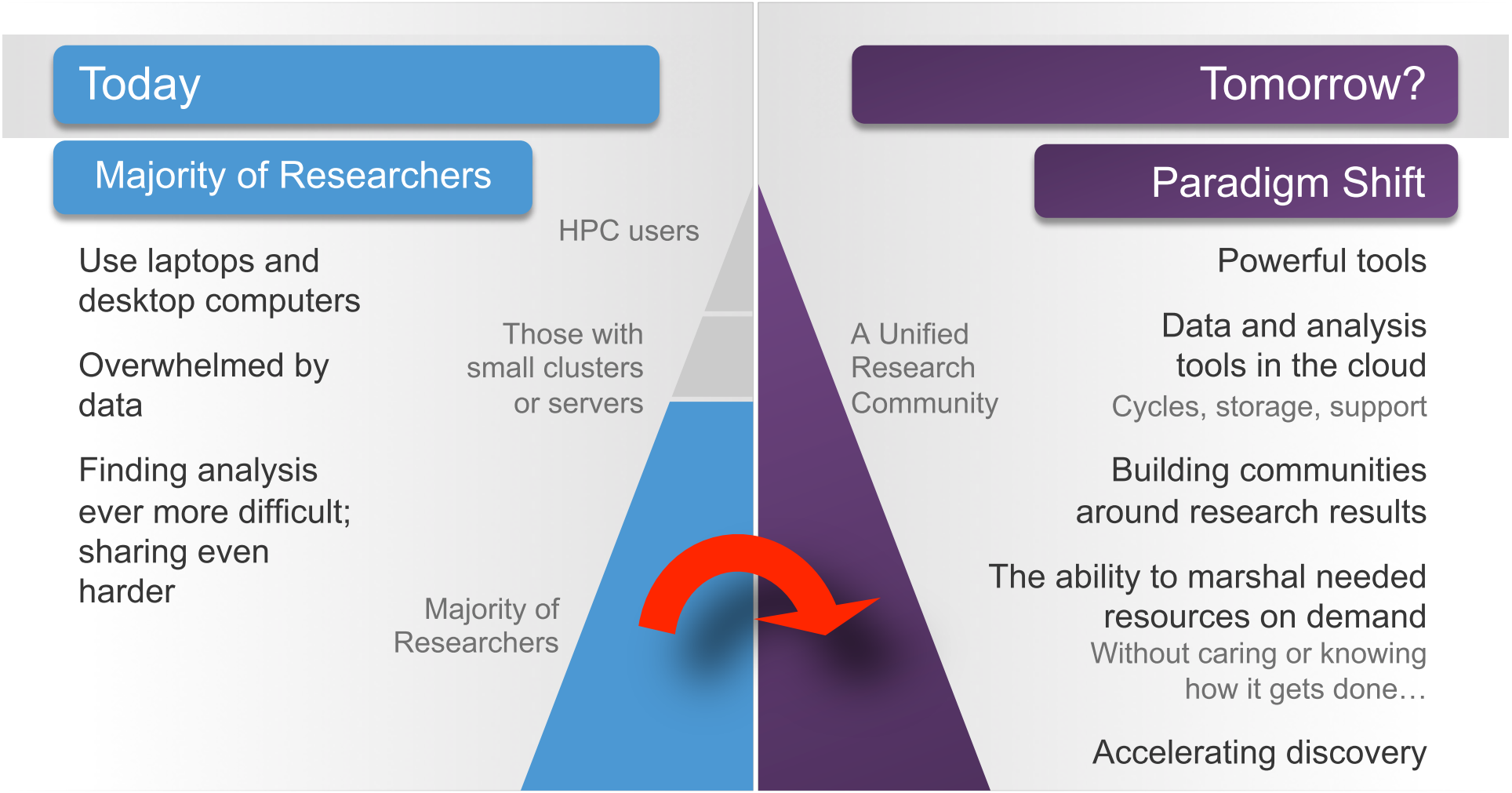
Pull data from remote instruments for visualization on the desktop

Create a revolution in scientific capability for everybody





Helping Democratise Research

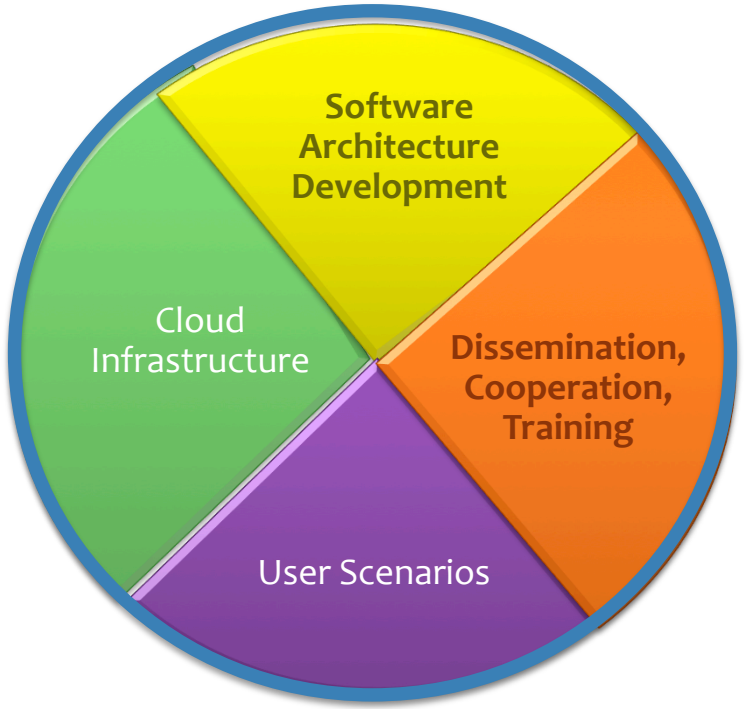


VENUS-C

Microsoft Research Connections

Industry contribution to the European Cloud Strategy

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



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

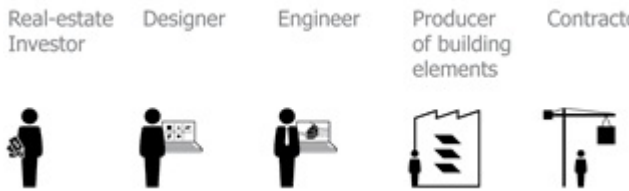
Some Success Stories

- 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
innovators in architecture

START UP

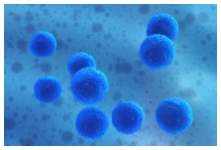


- Collaboratorio & its new start-up Green Prefab
- Collaborative platform for the design of ecofriendly & affordable buildings
- Selected by INTESA SAN PAOLO Start-up intitiative; expanding to US

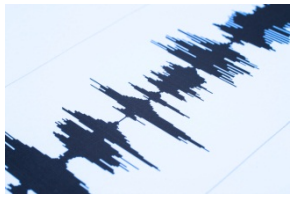
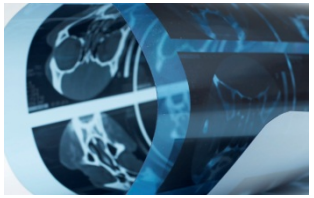
"We feel like pioneers in the right direction to the still untouched gold mine," Furio Barzon

Extending Cloud Usage - New Pilots & Experiments

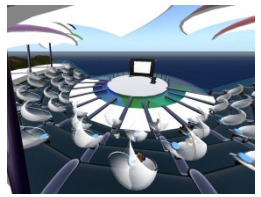
Engineering & Science

Architecture & Civil Engineering
Biology

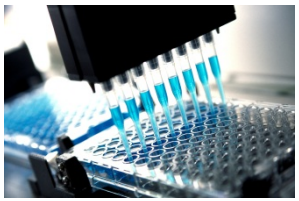


NEW DISCIPLINES

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

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