

Microsoft
Research



Microsoft Research Asia **Faculty Summit 2012**



The MSR Cloud For Science Project

Dennis Gannon

Director Cloud Research Strategy

Microsoft Research Connections



Outline

- The Challenge of the “long tail” of science
- Is there a sustainable financial model for scientific data?
- Data Centers powering commercial clouds
 - The Azure cloud in research
- Managing parallel data analysis from your desktop
- Opportunities now in China



The 4th Paradigm and the Revolution in Science





The data explosion is transforming science

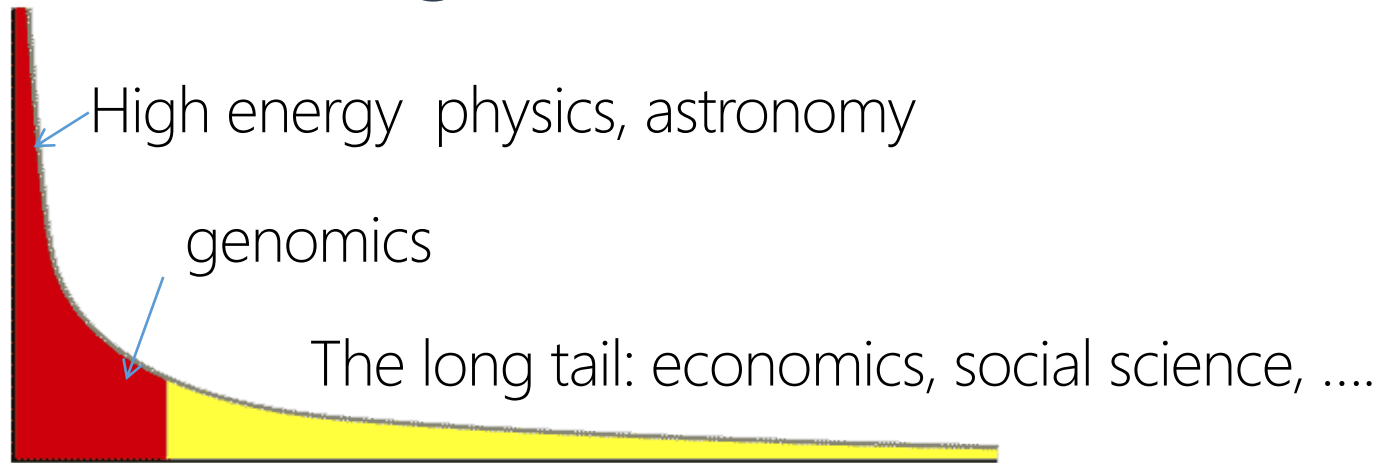


 Petabytes
Doubling & Doubling

- Every area of science is now engaged in data-intensive research
- Researchers need
 - Technology to publish and share data in the cloud
 - Data analytics tools to explore massive data collections
 - A sustainable economic model for scientific analysis, collaboration and data curation



The Long Tail of Science



Collectively “long tail” science is generating a lot of data
Estimated at over 1PB per year and it is growing fast.

Many research funding agencies now require all data be made public
US Universities are struggling with this new load
Data must be preserved
Data must be sharable, searchable, and analyzable



The Data for Science Sustainability Challenge

- Can we create a sustainable *economic* model for the long tail of science?
 - The government will not **directly** support an exponentially growing data collection.
- Our hypothesis
 - We can create an ecosystem that supports a **marketplace** of research tools and domain expertise
 - Allowing researchers to outsource special tasks to expert service providers
 - Funding will come from subscriptions from individual researchers, academic institutions and private sectors
 - Michigan's Inter-university Consortium for Political and Social Research (ICPSR) is a great model.



Data Centers and the Microsoft Cloud





The Microsoft Cloud is Built on Data Centers

~100 Globally Distributed Data Centers

Range in size from “edge” facilities to megascale (100K to 1M servers)



Quincy, WA



4 DCs





Cloud Properties

- Designed to Provide Information and Computation to Many Users
- Automatic Deployment and Management of Virtual Machine Instances
 - tens to thousands & dynamic scalability
- Dynamic Fault Recovery of Failed Resources
 - Cloud Services must run 24x7
- Automatic Data Replication
 - Geo-replication if needed
- Two levels of parallelism
 - Thousands of concurrent users
 - Thousands of servers for a single task.





Our Experience (so far) with Science in the Cloud





Microsoft Cloud Research Engagement Project

Work with international funding agencies to grant access to cloud resources to researchers. 90 projects world wide.

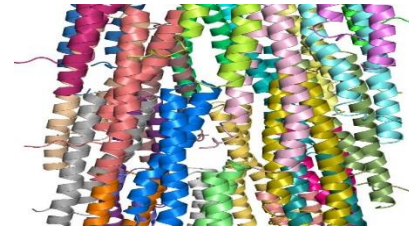




Sample Projects on Windows Azure

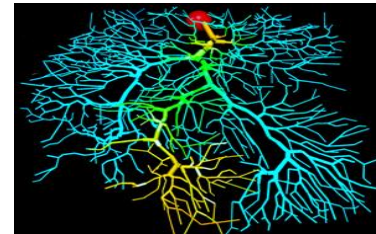
- Protein Folding

- The University of Washington is studying the ways proteins from salmonella virus inject DNA into cells. Used 2000 concurrent cores. PI: Nikolas Sgourakis, Baker Lab.



- Joint Genetic and Neuroimaging Analysis

- France's premier research institute INRIA is using 1000 cores of Azure to study large cohorts of subjects to understand links between genetic patterns and brain anomalies. PIs: Radu Marius Tudoran, Gabriel Antoniu IRISA INRIA France.



- Fire Risk

- This app from the University of Aegean estimates the fire risk probability using meteo and geo-data sources and calculates the so-called fire risk index. A client application for the fire and forest services as well as cloud services that allow access to real-time data from sensors and on-the-ground reporting. This service has been tested and validated with fire-fighting crews in both Mytilene and Thessaloniki, Greece PI: Kostas Kalabokidis





More Samples

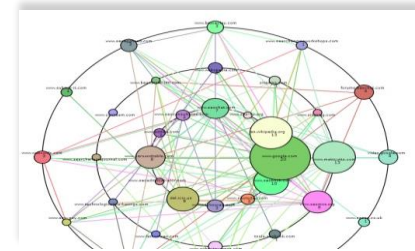
- **Drug Discovery**

- Researchers at Newcastle University in the U.K. are using Azure to model the properties (toxicity, solubility, biological activity) of molecules for potential use as drugs. This cloud solution is primarily aimed at domain scientists who do not have advanced IT skills. PI: Paul Watson



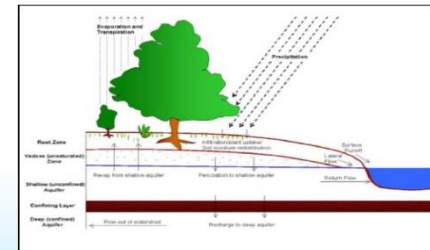
- **Predicate-Argument Structure Analysis**

- University of Kyoto team applied a predicate-argument structure analysis to a huge Japanese corpora consisting of about 20 billion web sentences, to improve the open-search engine infrastructure TSUBAKI, which is based on deep natural language processing. To achieve this goal 10,000 core on Windows Azure were used in a massively parallel computation that took about a week.



- **Model and Manage Large Watershed Systems**

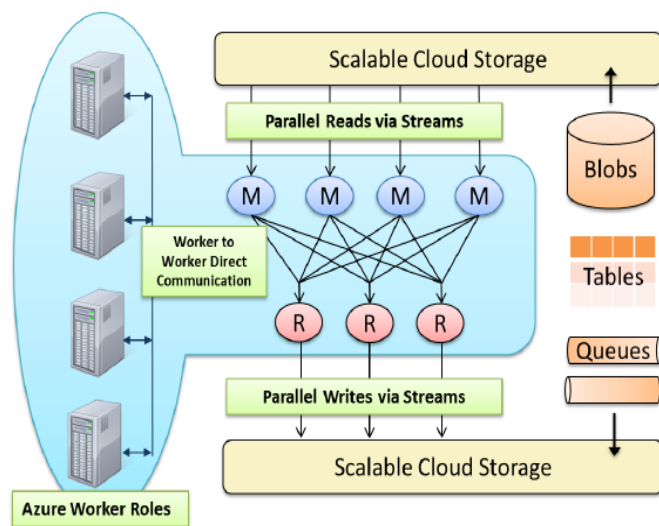
- Predict the impact of land use change and climate change on water resources. Going beyond modeling to include entire workflow from data collection to decision making. University of Virginia and South Carolina. Using HPC Scheduler on Azure to launch thousands of analysis jobs.



What have we learned?



1. Traditional communication-intensive MPI apps belong on supercomputers.
2. The cloud advantage
 - Applications that require sharing & web access
 - Massive “Map Reduce” data analytics on cloud resident data
 - Massive ensemble computations
3. The “scale-out as needed” model works.
 - Users prefer spending on pay-as-you-go cloud to buying and cluster hardware. (use 500 cores 2 days a week vs. 120 cores purchased)
 - Most researcher prefer to avoid maintaining cluster and data storage facilities.
 - Most users do not have access to supercomputers





Bringing Large scale data analytics to more people. Let Scientists Be Scientists...

Most scientists do not want to be system administrators
They don't want to learn to use supercomputers

They want to focus on their science

They use standard tools: spreadsheets,
statistical packages, desktop visualization
Programming = modifying a few
parameters in a trusted scripting language

They want to share experiments with their
collaborators





Cloud Science Stack

- **The challenge:** Design a platform for scientific data management and analysis that is
 - Open and extensible
 - Provides an economic sustainability model for data preservation and use
 - Is easily accessed by simple desktop/web analysis apps.
 - Encourages scientific collaboration
 - Leverages the capabilities of public clouds and on-campus resources
- Can we build a demonstration project to test the feasibility of this?
- Build it using the tools the community wants and uses.



Building a Research Marketplace

- Built on top of a general purpose framework for deploying apps and data as services
- Tools available now to support data curation in the cloud
 - DataUp (From MRC), DataVerse (Harvard), Duraspace (MIT), SQL Share (Bill Howe UW)
- More sophisticated large scale data movement tools.
- Should support access to HPC on Azure and evolving Big Data tools (Hadoop, etc.) and MSR and community ML libraries
- Tools to integrate services into re-usable pipelines must be available
 - In Genomics - eScience Central, Galaxy, ETA, etc
- It should be easy to deploy standard community tools as services or as stored VM images.



IPython Notebook on Windows Azure

Browser/OS independent

Math rendering

Notebooks stored in Azure Blobs

Interactive Parallel Computing & MPI

Engine on Azure via Windows or Linux VM

Support for R, LaTeX, PHP, ...

Inline graphics & Video

Collaborate with colleagues

IPython Dashboard x IPy spectrogram

127.0.0.1:8888/bed6aac-d118-4521-96e7-3f5054b6641d

IP[y]: Notebook spectrogram Last saved: Feb 23 5:19 PM

File Edit View Insert Cell Kernel Help

Markdown

Simple spectral analysis

An illustration of the [Discrete Fourier Transform](#)

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad k = 0, \dots, N-1$$

using windowing, to reveal the frequency content of a sound signal.

We begin by loading a datafile using SciPy's audio file support:

```
In [1]: from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin specgram routine:

```
In [2]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```

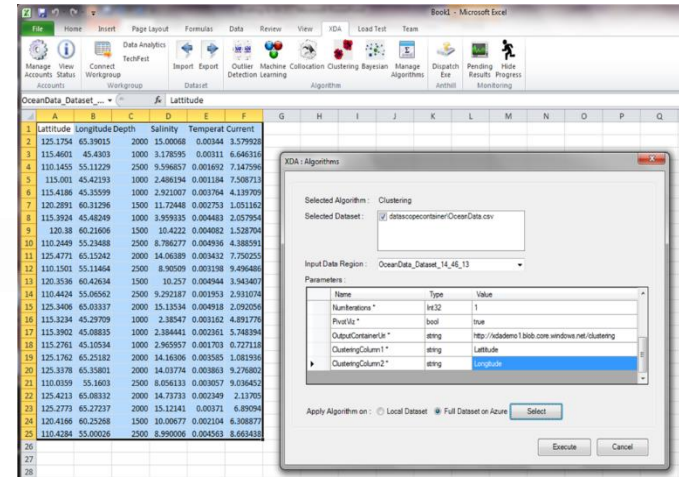
Raw audio signal

Spectrogram



Excel DataScope

Cloud Scale Data Analytics from Excel



Bringing the power of the cloud to the laptop

- **Data sharing in the cloud**, with annotations to facilitate discovery and reuse;
- **Sample and manipulate** extremely large data collections in the cloud;
- **Top 25 data analytics algorithms**, through Excel ribbon running on Azure;
- **Invoke models**, perform analytics and visualization to gain insight from data;
- **Machine learning** over large data sets to discover correlations;
- **Publish** data collections and visualizations to the cloud to share insights;

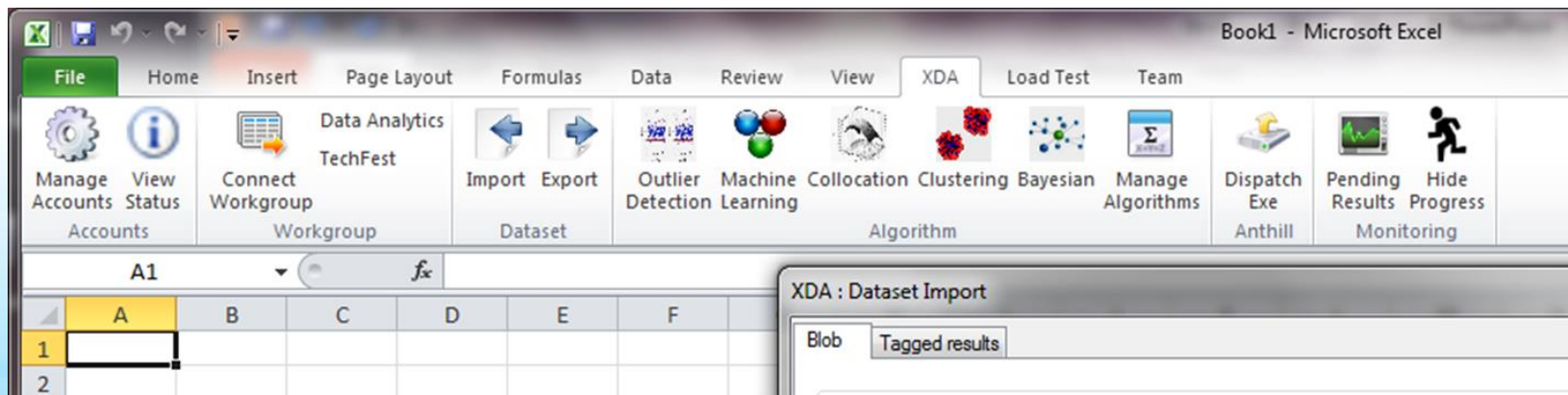
Researchers use familiar tools, **familiar but differentiated**.





Data Analytics Algorithms for Datascope

- Parallel algorithms for massively distribute data
 - ❑ Clustering: K-means, fuzzy clustering, canopy clustering;
 - ❑ Recommendation Mining: Log-Likelihood;
 - ❑ Prediction: SVM; trend prediction
 - ❑ Frequent Item Set Mining: Collocation, Outlier Detection;
 - ❑ Bayesian/Regression Toolkit (linear, non-linear, logistic);
 - ❑ Bayesian Net, Neural Nets, other Machine learning
- These algorithms are being built on top of the Daytona mapreduce engine





Opportunities for Cloud Research in China



- In 2012 CNIC and Microsoft announced an agreement to provide researchers with free access to cloud computing resources
 - A total of 5 million core hours of Windows Azure
 - 40 Terabytes of cloud storage and SQL Azure
- The Windows Azure Cloud is a powerful tool for
 - Hosting Web access to scientific data collections and tools
 - Large scale data analytics with MapReduce
 - Massive parameter sweep or ensemble computations
 - Hosting large distributed scientific collaborations
- Interested academic researchers please contact:
dennis.gannon@microsoft.com or Dr. Jianhui Li,
Director of Scientific Data Center of CNIC at
lijh@cnic.cn



Windows Azure[™]



Microsoft[®]
SQL Azure[™]



Questions