

Science as a Service

Data Analytics and Data Mining: The Approaching Tidal Wave

Dennis Gannon
Director Cloud Research Strategy

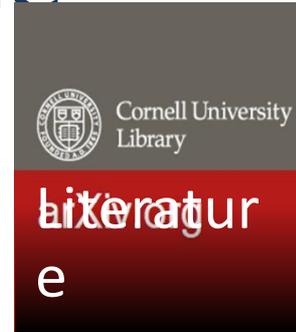


Talk Outline

- The 4th paradigm of science
- The genesis of “big data” analysis in the cloud :
searching the web
- The revolution in machine learning
- Examples
 - The n-gram and language translation
 - Recognizing images and spam
 - Predicting traffic flows
 - Hospital readmissions
 - Genome-wide association studies
- The challenges for the long tail of science



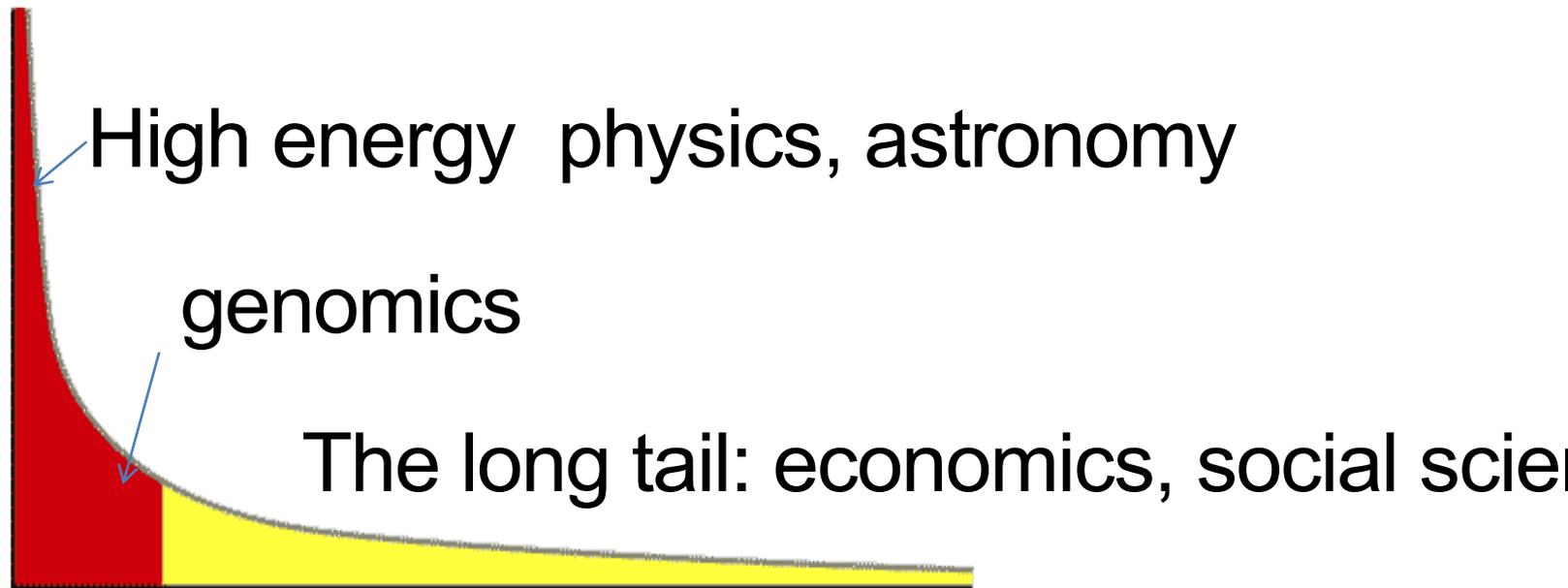
The data explosion is transforming science



 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.

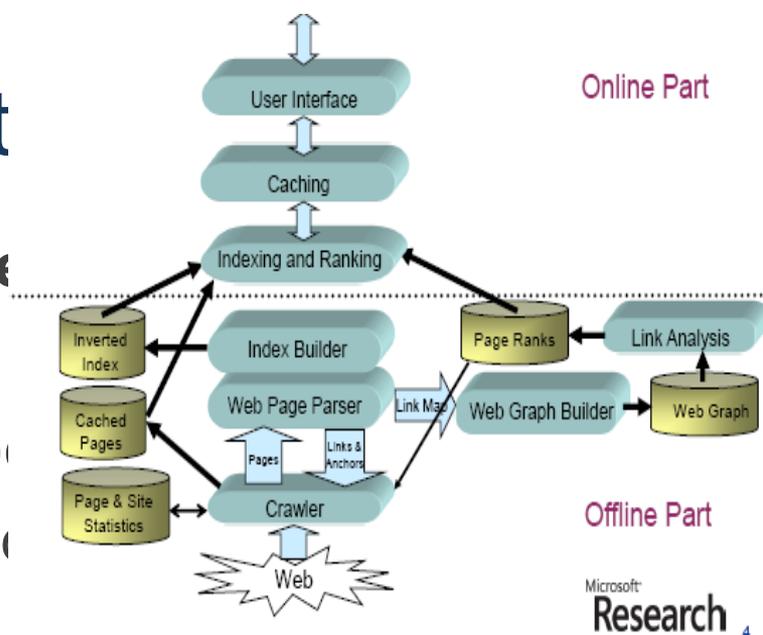
Many funding agencies now or soon will require all data be made public

US Universities are struggling with this new load

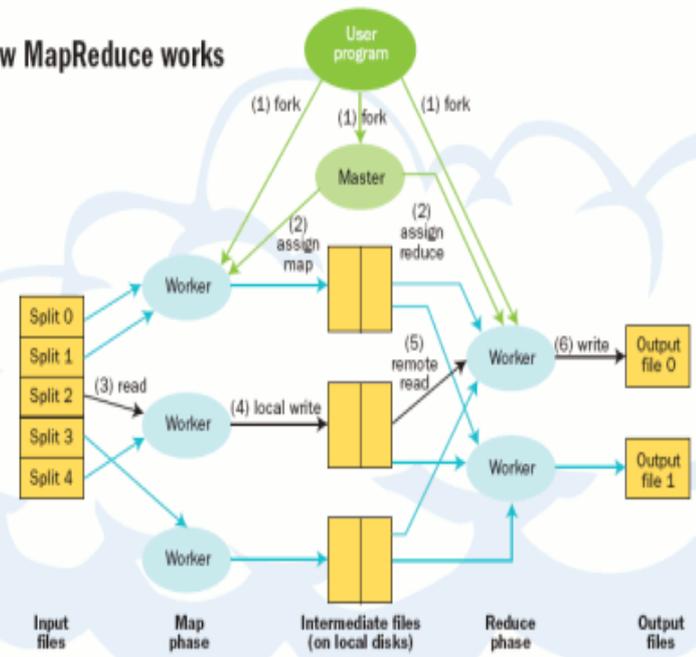
Data must be preserved

Origins of Big Data Analyt

- **Early Days – Building a index of the**
- **First challenges – make the search**
 - Distributed the data over 1000 no
 - Use MapReduce to build index and
- **Next steps – semantic challenges**
 - Query “theater classes in k
 - Hits: “Improv acting in San F
 - “Berkeley Rep School of Thea
 - Concept clustering and rel
 - Statistical models – K-mea
 - latent semantic indexing, SV
- **Now MR in many big data applications**



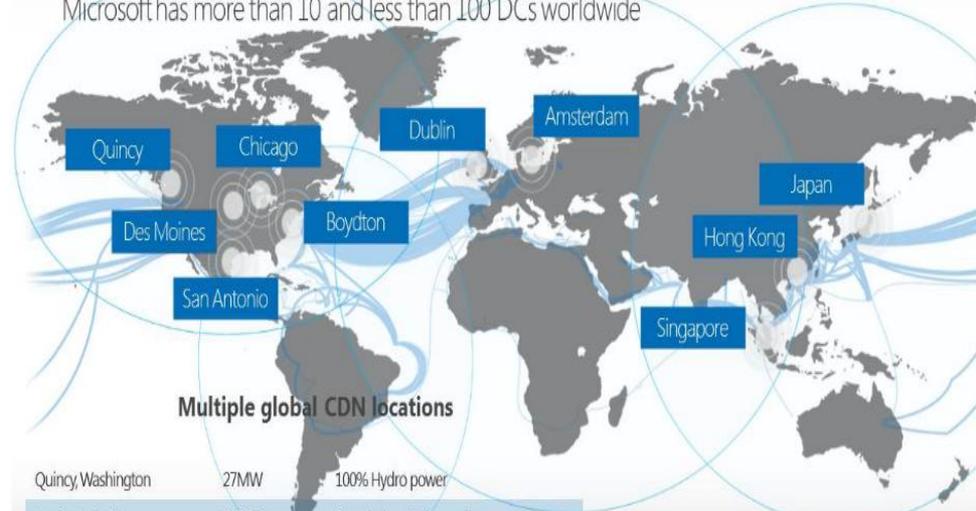
How MapReduce works



The Explosion of the Data Centers

Microsoft Data Center Scale

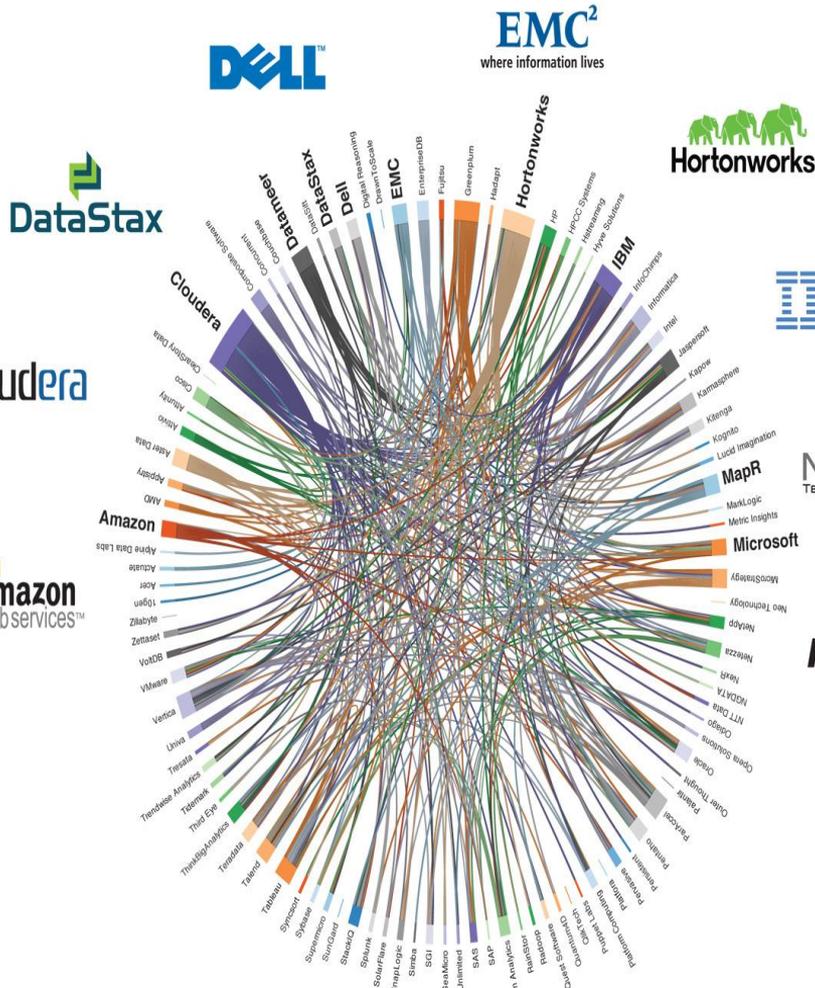
Microsoft has more than 10 and less than 100 DCs worldwide



Multiple global CDN locations

Quincy, Washington	27MW	100% Hydro power
San Antonio, Texas	27MW	*Recycled water for cooling
Chicago, Illinois	Up to 60MW	Water side economization, Containers
Dublin, Ireland	Up to 50MW	Outside air cooling, PODs

"Data Centers have become as vital to the functioning of society as power stations."
The Economist



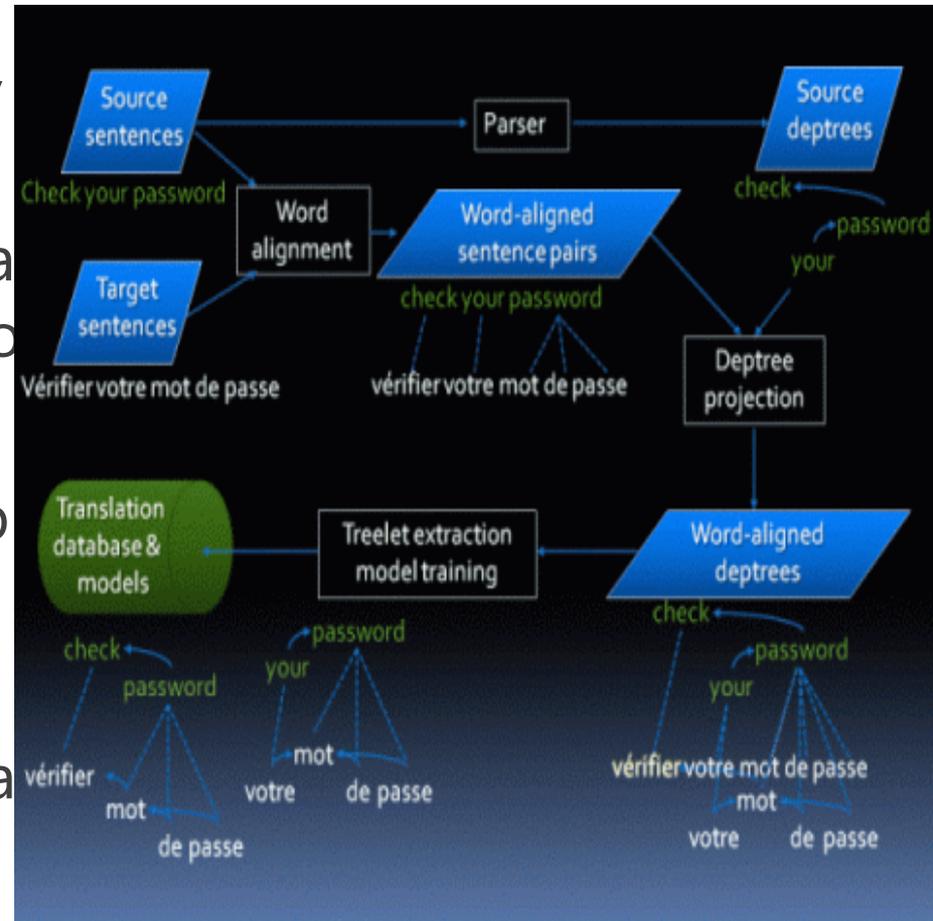
The Rise of the Hadoop Eco

<http://datameer2.datameer.com/blog>



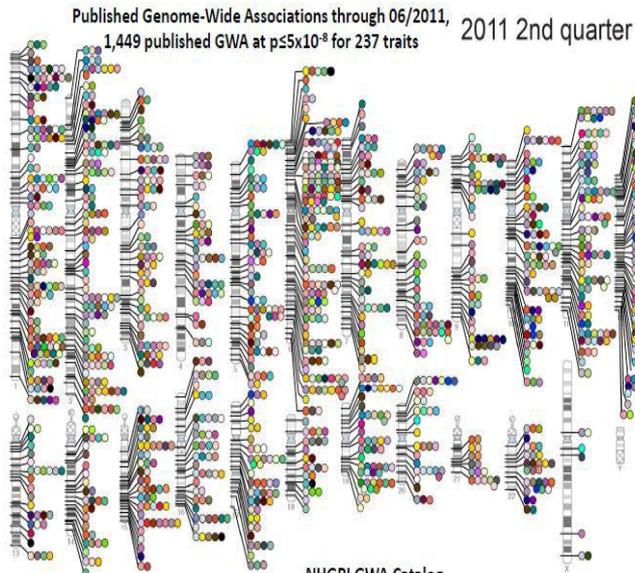
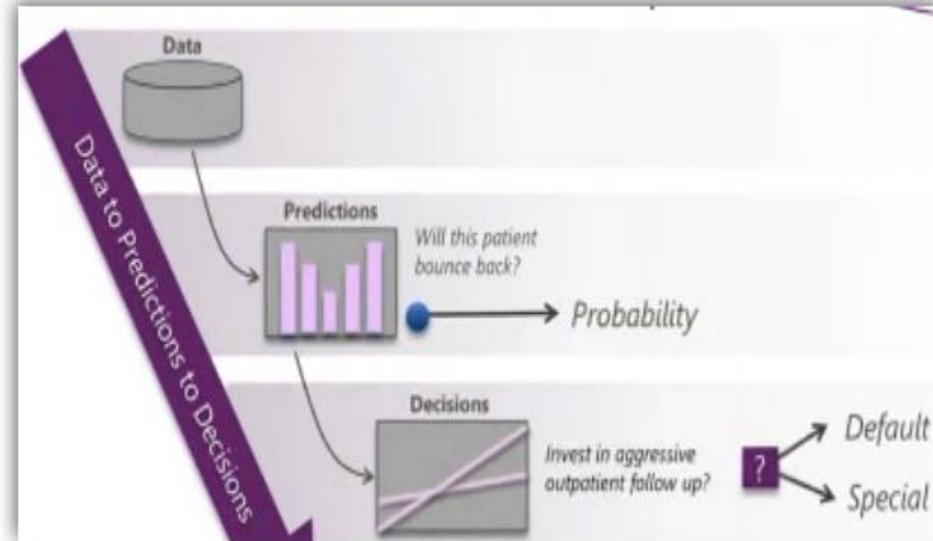
Natural Language Translation

- Given a sufficiently large collection of translated text we can “learn” to translate.
- Bing and Google have fairly on-line translators
- Both syntax-based and phrase statistical machine translation
- Other applications
 - N-grams for query completion
 - “I can’t get no ...”
 - ESL grammar assistant
 - Generate a summary of a



Big Data Analytics in Me

- Hospital Readmissions (from Eric Horvitz of MSR)
 - 20% of patients were rehospitalized within 30 days of their discharge from hospitals and that 35% of patients were rehospitalized within 90 days
 - Study of large multi-year data set of hospitalizations. Machine learning produced a predictive model that can accurately predict likelihood of a readmission given patient data.



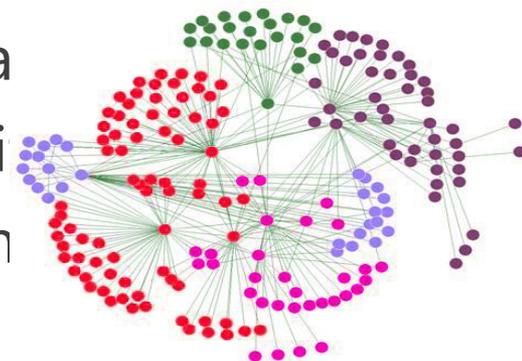
NHGRI GWA Catalog
www.genome.gov/GWAStudies

- The Genetic Causes of Disease (David Heckerman)
 - Use data from the Wellcome Trust for a GWAS for a large population looking for
 - Looking for causes for seven common diseases (bipolar, r. arthritis, coronary, hypertension,)
 - Confounding is a problem. Needed a new algorithm.
 - Ran on Azure cloud using 35,000 cores in 3 weeks.

Beyond MapReduce: Graph Computing



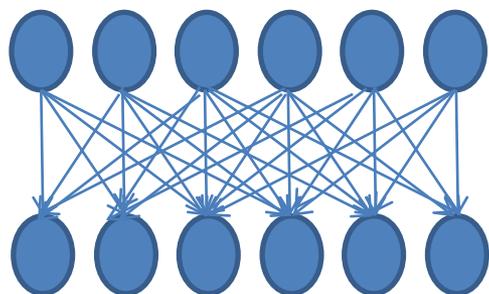
- GraphLab from CMU. (Carlos Guestrin)
- The concept: big data has connections and co-occurrences.
 - These are needed to make accurate predictions
- The social network:
 - Understanding the connections between people, what they like in common can be used for making suggestions (\$\$\$!)
- Well suited to massive asynchronous para
 - Similar to some network routing algorithm
 - Distributed graph nodes using random to avoid power law problems



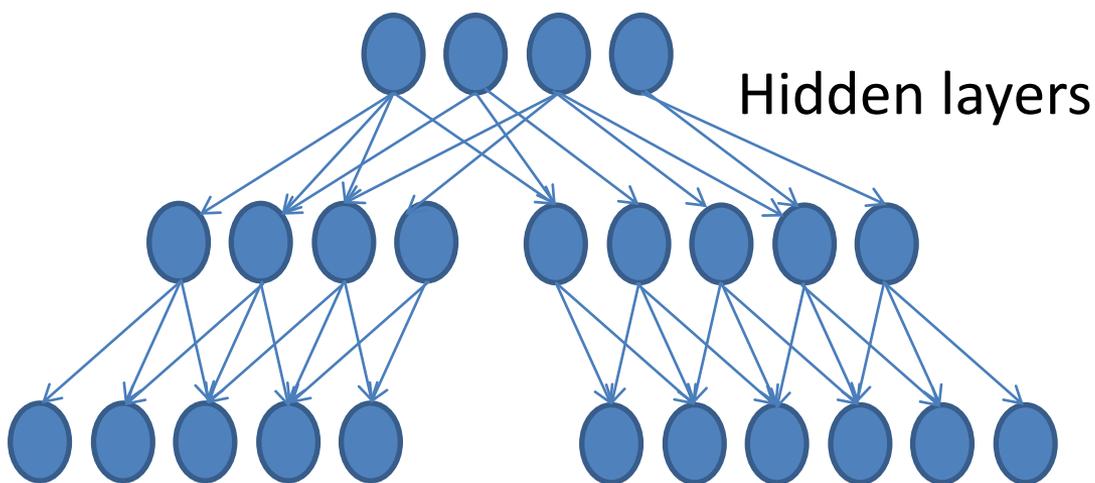
The Machine Learning Revolution

- **Big data and massive parallelism change the game.**
 - Supervised Machine Learning - inferring knowledge from labeled training data
 - Unsupervised – finding the hidden structure in data without labels

Inputs (training data)



Observations (outputs)

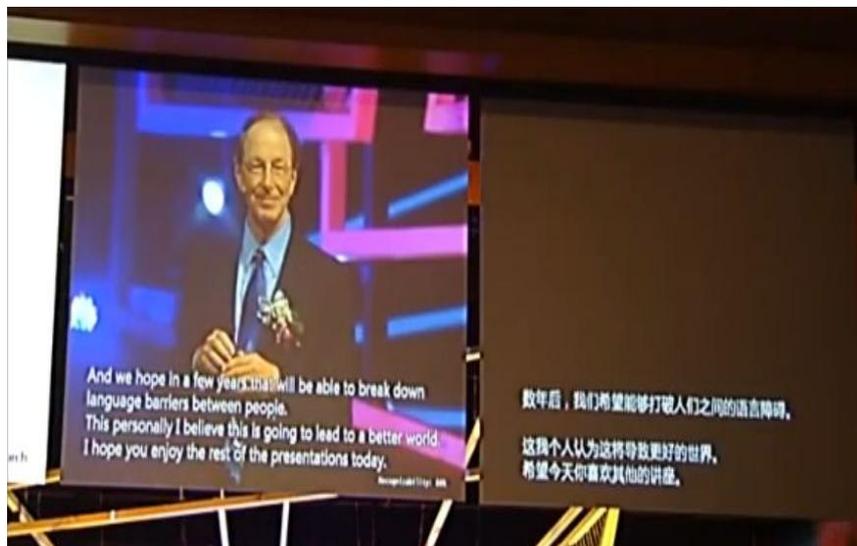
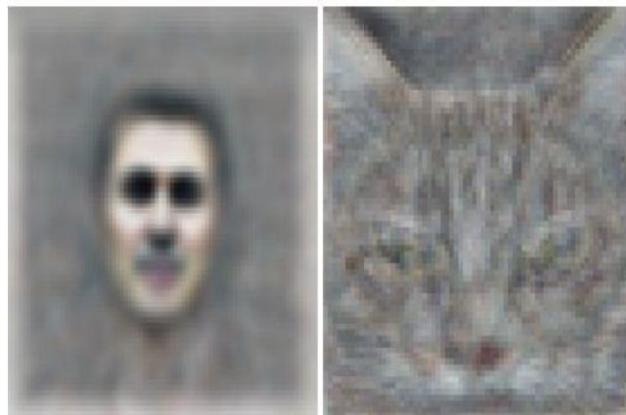


Output data

Input data

Deep Learning

- Deep neural network concepts pioneered by Geoffrey Hinton
 - *Building High-level Features Using Large Scale Unsupervised Learning*, Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeffrey Dean, and Andrew Y. Ng
 - Google cluster of 16,000 cores with a connections, 10 million unlabeled ima
- Microsoft Research demonstrates rea
ese tran

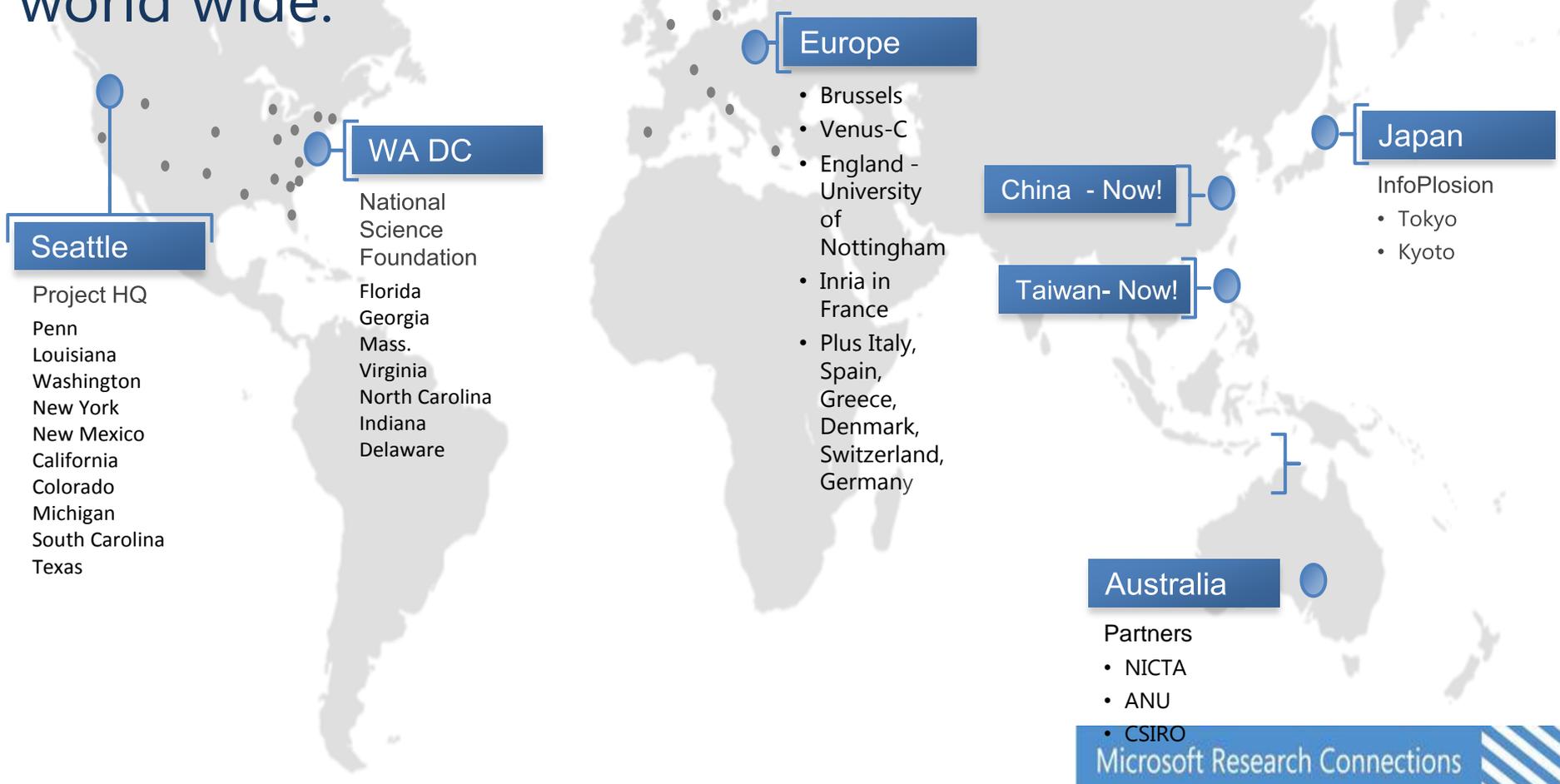


Our Experience (so far) with Science in the Cloud



Microsoft Cloud Research Engagement Project

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



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

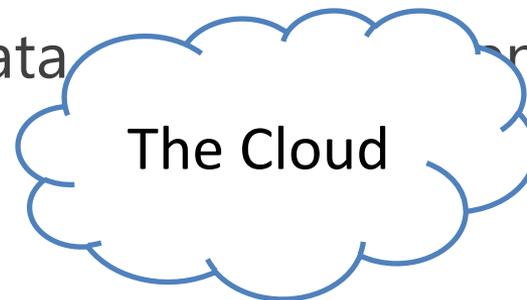
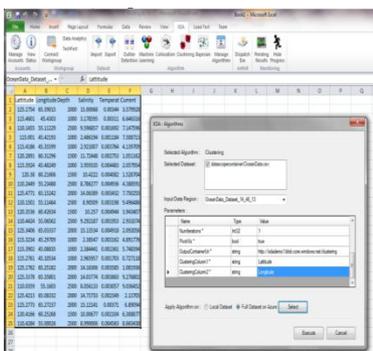
Programming = modifying a few parameters in a trusted scripting language

They want to share experiments with their collaborators



One Simple, Powerful Idea: DataUp

- Most scientists do data collection and analysis using spreadsheets.
- How to they share them? preserve them? generate metadata to store them?
- DataUp is an open source Excel plug in (or web tool) to help researchers document, manage, and archive their tabular data, DataUp operates within the scientist's workflow and integrates with Microsoft® Excel
- through basic metadata



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



Next Steps – Bringing Communities Together



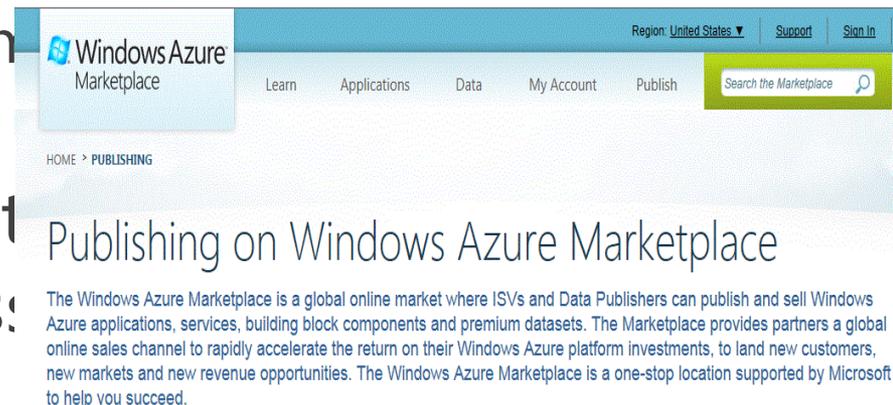
Internet2 and 13 University CIOs @ MS

- A meeting March 1 in Bellevue
 - Universities need to solve some problems
 - An effective way to use the cloud to address them
- Use standard authentication protocols
- Rational data costs and pricing
- The Research Genomics Challenge
 - A universal problem – analysis and storage of sequence data
 - A pilot project.
- The Rest – “The Long Tail of Science”
 - Many disciplines, each with unique data and analysis challenges



Next Goal: Build a Research Marketplace

- A place to host services for
 - advanced data analytics and machine learning libraries
 - Curated data collections (via dataverse or duracloud)
 - Data upload, curation and visualization tool (CDL project)
- A support platform for research challenge projects
 - such machine learning and n analysis
- Exploit Azure Marketplace to limited free and paid access:



The screenshot shows the Windows Azure Marketplace website. The top navigation bar includes the logo, 'Region: United States', 'Support', and 'Sign In'. Below the navigation bar, there are links for 'Learn', 'Applications', 'Data', 'My Account', and 'Publish'. A search bar is located on the right side of the navigation bar. The main content area shows a breadcrumb trail: 'HOME > PUBLISHING'. Below the breadcrumb trail, the heading 'Publishing on Windows Azure Marketplace' is displayed. The text below the heading describes the marketplace as a global online market where ISVs and Data Publishers can publish and sell Windows Azure applications, services, building block components and premium datasets. It mentions that the marketplace provides partners with a global online sales channel to rapidly accelerate the return on their Windows Azure platform investments, to land new customers, new markets and new revenue opportunities. It also states that the marketplace is a one-stop location supported by Microsoft to help you succeed.



© 2010 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.

The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

