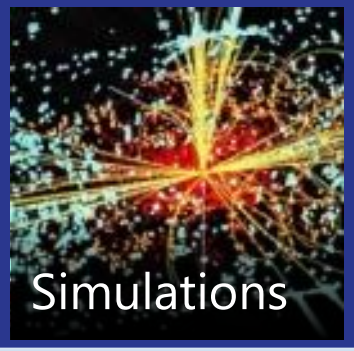
Sustaining Big Data

Dennis Gannon
Microsoft

The data explosion is transforming science

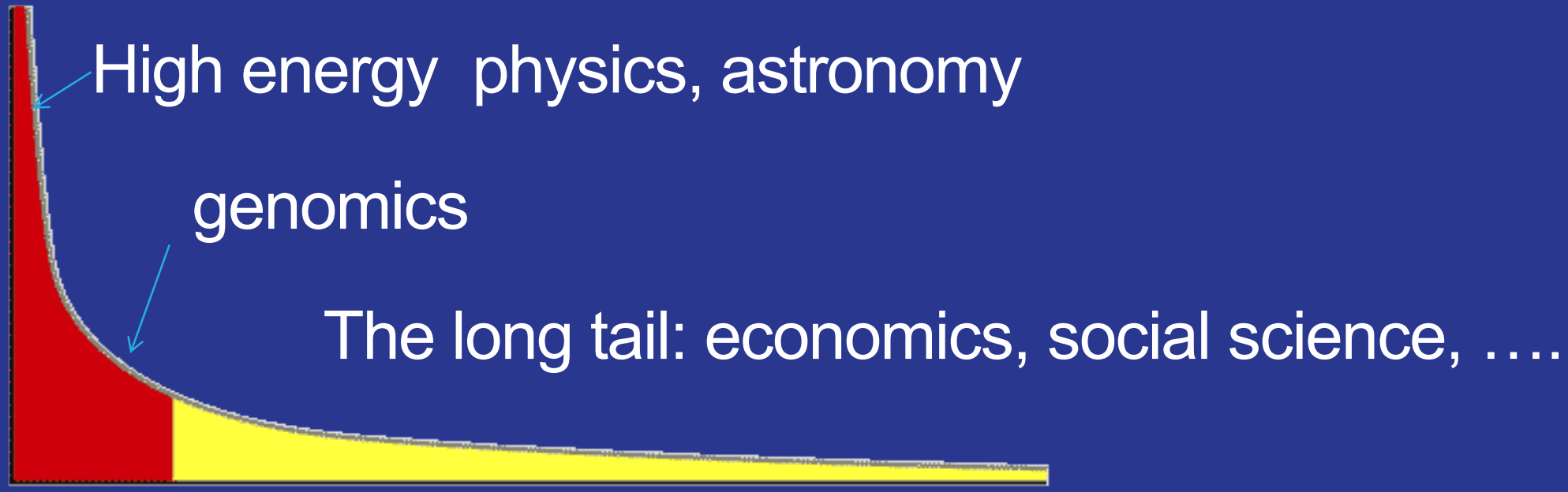


Petabytes
Doubling & Doubling

THE RESPONSE

- 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



- Can we create a sustainable model for the long tail of Science?
  - an ecosystem that supports a **marketplace** of research tools and domain expertise
    - Allowing researchers to outsource special tasks to expert service providers

Data Preservation and Sharing

- **Collectively "long tail" science is generating a lot of data**
  - Estimated at over 1PB per year and it is growing fast.
- **US National Science Foundation requires all data be made public**
  - US Universities are struggling with this new load
  - Data must be preserved
  - Data must be sharable, searchable, and analyzable
- **Is there a role for the commercial cloud provider?**

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

Chicago, IL

San **Antonio**, TX

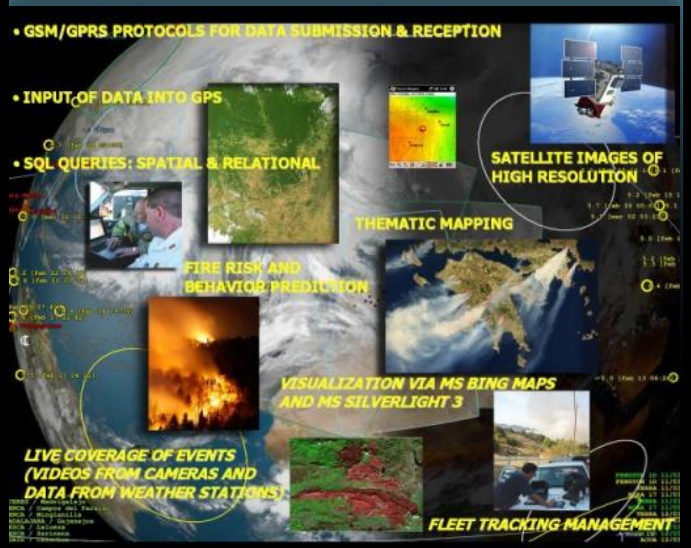
Dublin, Ireland

Generation 4 DCs

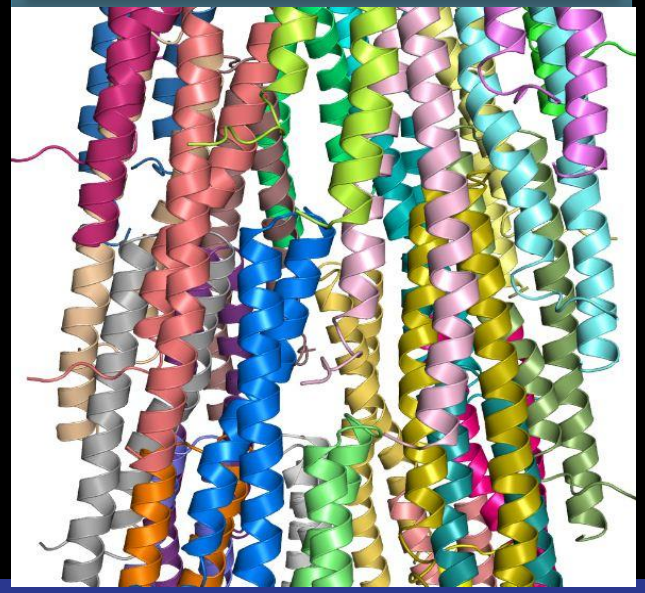
International Cloud Research Engagement Project

Demonstrate that the cloud is a powerful tool that can revolutionize academic research and collaboration.

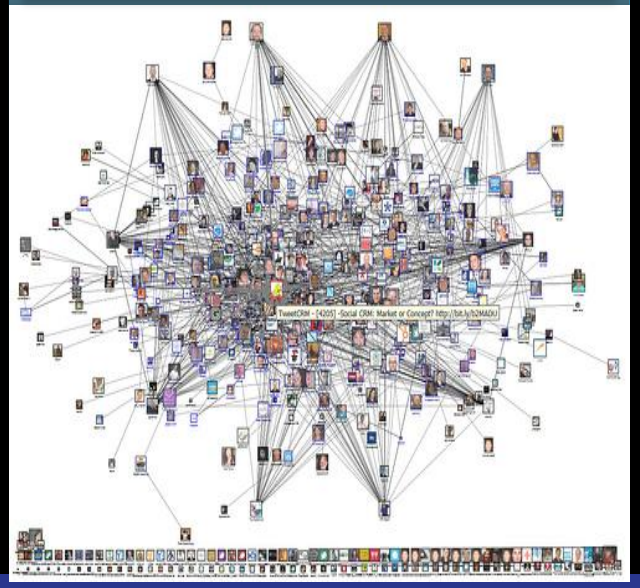
Civil Protection
& Emergencies



Protein Folding



Social Graph Analysis



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 - IU, OSU, UCD, Mich, Utah
    - Discussions starting.
- The Rest – “The Long Tail of Science”
  - Many disciplines, each with unique data and analysis challenges

Sample “Long Tail” Projects

- Transportation Research - UCD, Georgia Tech, USC, UW, UVA, …
  - UCD ULTRANS – modeling entire California transportation grid
  - Gtech – realtime analysis of Atlanta area traffic
- Social and Psychological Sciences – Virginia, Duke
  - Harvard’s Dataverse and UVA, UW, Harvard “project Implicit” social cognition
- Medical Sciences – UW and Utah
  - Imaging - CT, MR, PET / SPECT, and ultrasound and toxicology
- Maps and Geo Data – Oregon State
- Musical Composition and Performance Data – UCSD
- Plant Sciences – Texas and Arizona

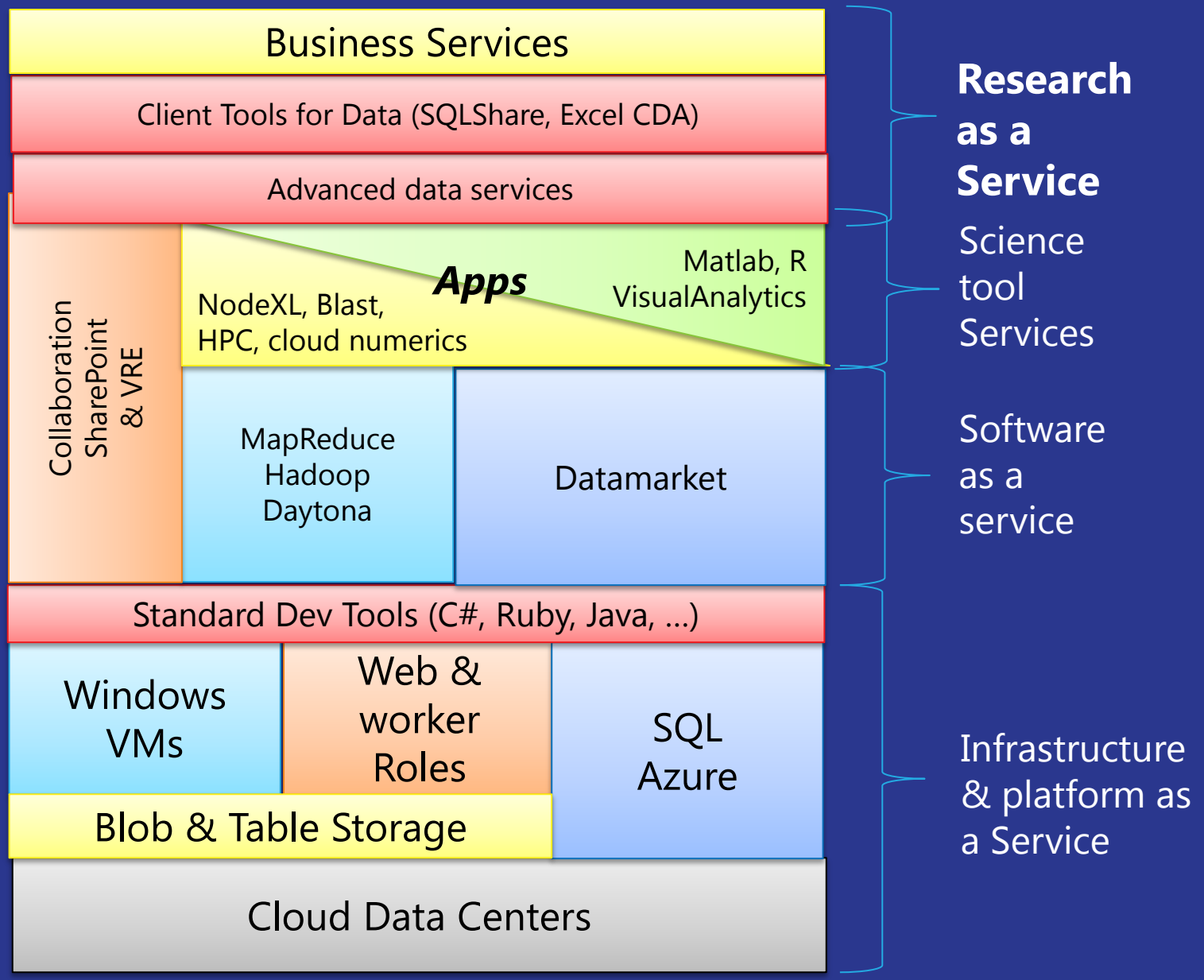
Next Steps

- We will convene a series of meetings to plan pilot projects in more detail
  - Genomics
  - One or two of the “long tail” topics where we can identify a critical mass of interested collaborators.
  - Coordinate between them and the product groups to accomplish something meaningful.

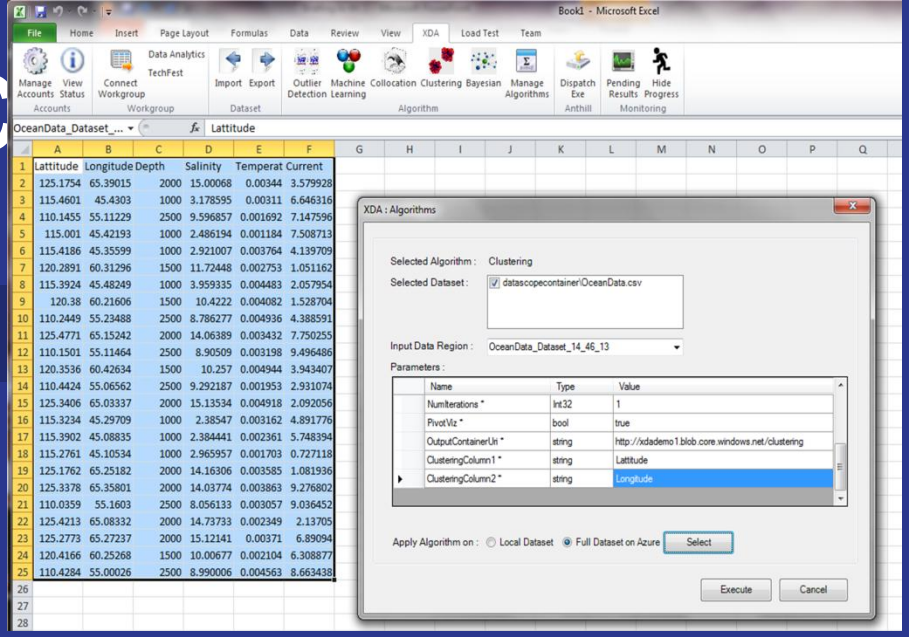
The Role of Commercial Clouds?

- Cloud data services from commercial providers can democratize access to big data.

- The cloud can support *research data services* that are
  - Open and extensible
  - Easily accessed by simple desktop/web analysis applications
  - Encourages scientific collaboration
  - Allows scientific analysis of massive data collections without requiring each researcher to acquire a private supercomputer


Business Services
Client Tools for Data (SQLShare, Excel CDA)
Advanced data services
Collaboration SharePoint & VRE
Apps
NodeXL, Blast, HPC, cloud numerics
Matlab, R VisualAnalytics
MapReduce Hadoop Daytona
Datamarket
Standard Dev Tools (C#, Ruby, Java, …)
Windows VMs
Web & worker Roles
SQL Azure
Blob & Table Storage
Cloud Data Centers
Research as a Service
Science tool Services
Software as a service
Infrastructure & platform as a Service

Excel Cloud Data Analytic



Bringing the power of the cloud to the laptop

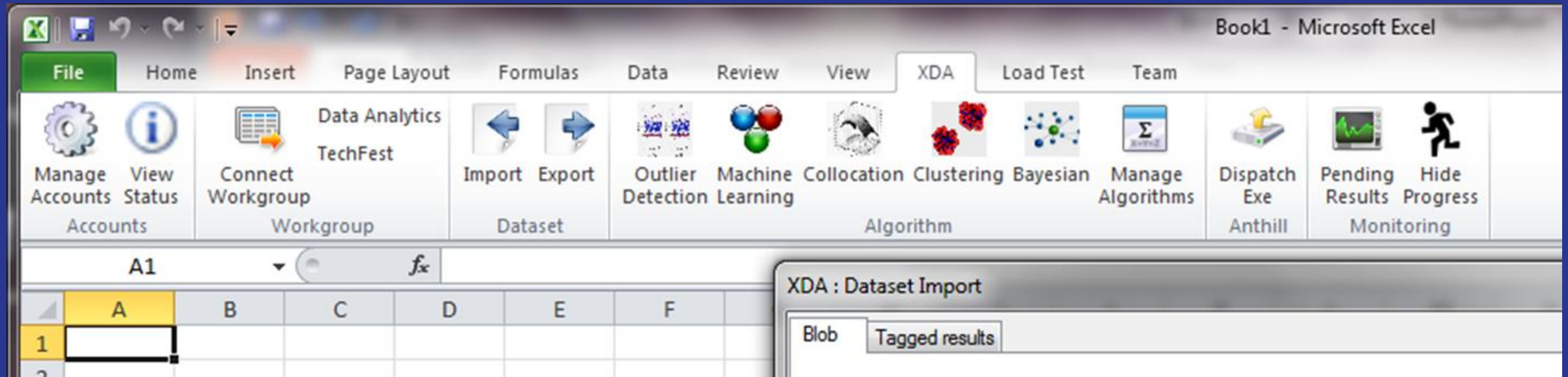
- Data sharing in the cloud, with annotations to facilitate discovery and reuse;
- Sample and manipulate extremely large data collections in the cloud;
- Top 25 data analytics algorithms, through Excel ribbon running on Azure;
- Invoke models, perform analytics and visualization to gain insight from 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 Excel

CDA

- 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



Whose Laws Apply to You and Your Data?

Germany

US

China

Kenya

Transnational Data Flows

- National sovereignty
- Economic competition
- Civil and criminal activity
- Social norms and behavior
- National security

Data Sovereignty

- Can I store my research data in a data center in another country?
  - For most research data this is not an issue
- For PII data we need **globally** harmonized data access and protection rules.
- There are technology solutions to protecting sensitive data
  - Based on Homomorphic Encryption techniques
  - Data owner can grant access to different entities for different uses
  - Cloud provider has no access and holds no keys.

Data Convergence: Opportunities & Risks

- The Internet of things
  - Streams of data from satellites, economic markets, weather, personal media, genomics and med data and geo sensors will converge in the cloud
- An unprecedented opportunity
  - Data mashup analytics will help track disease, warn of famine, optimize economic conditions on a global scale
- Risks
  - We need to prevent the possible abuses.
- We need basic research programs on privacy preserving data algorithms, collections and storage.