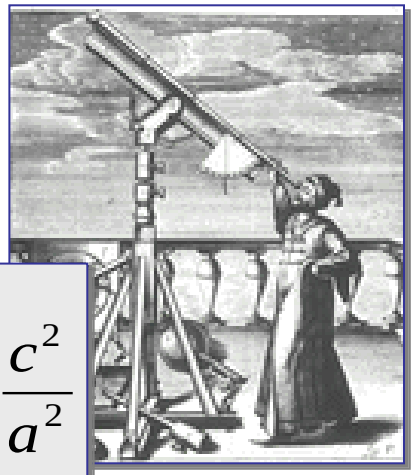
Data Science & Data Scientist

Yan Xu
Microsoft Research
yanxu@microsoft.com

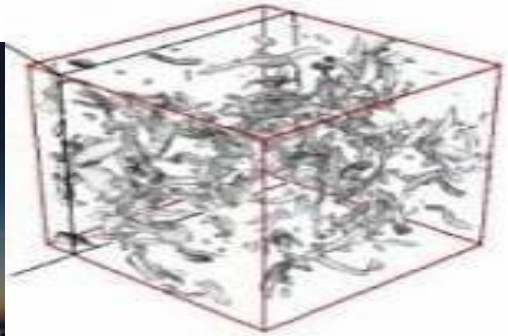
Microsoft Research Connections

Emerging of a Fourth Paradigm



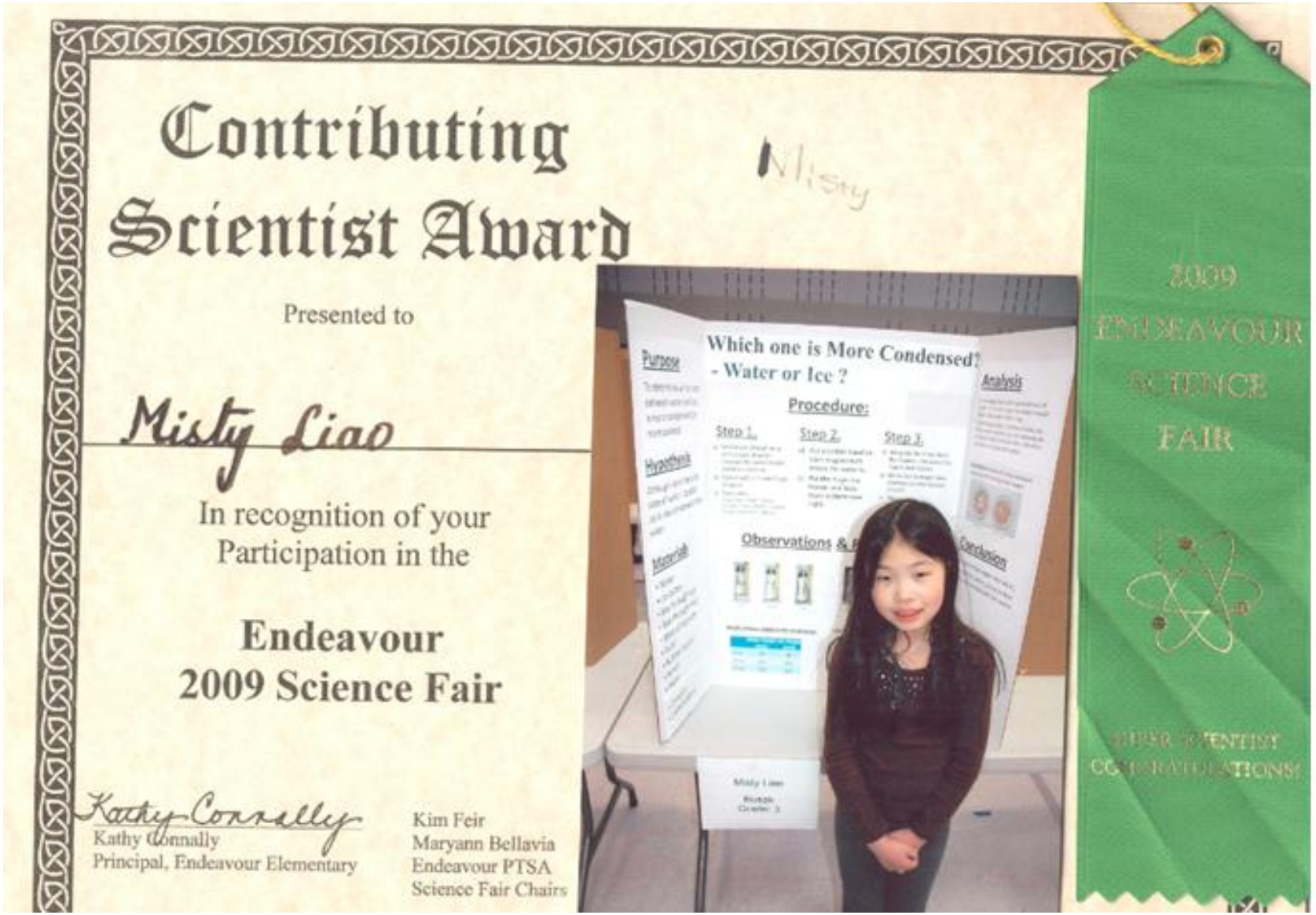
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

- 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
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks
- 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



脱颖而出的， 基于海量数据的，科研的第四范式

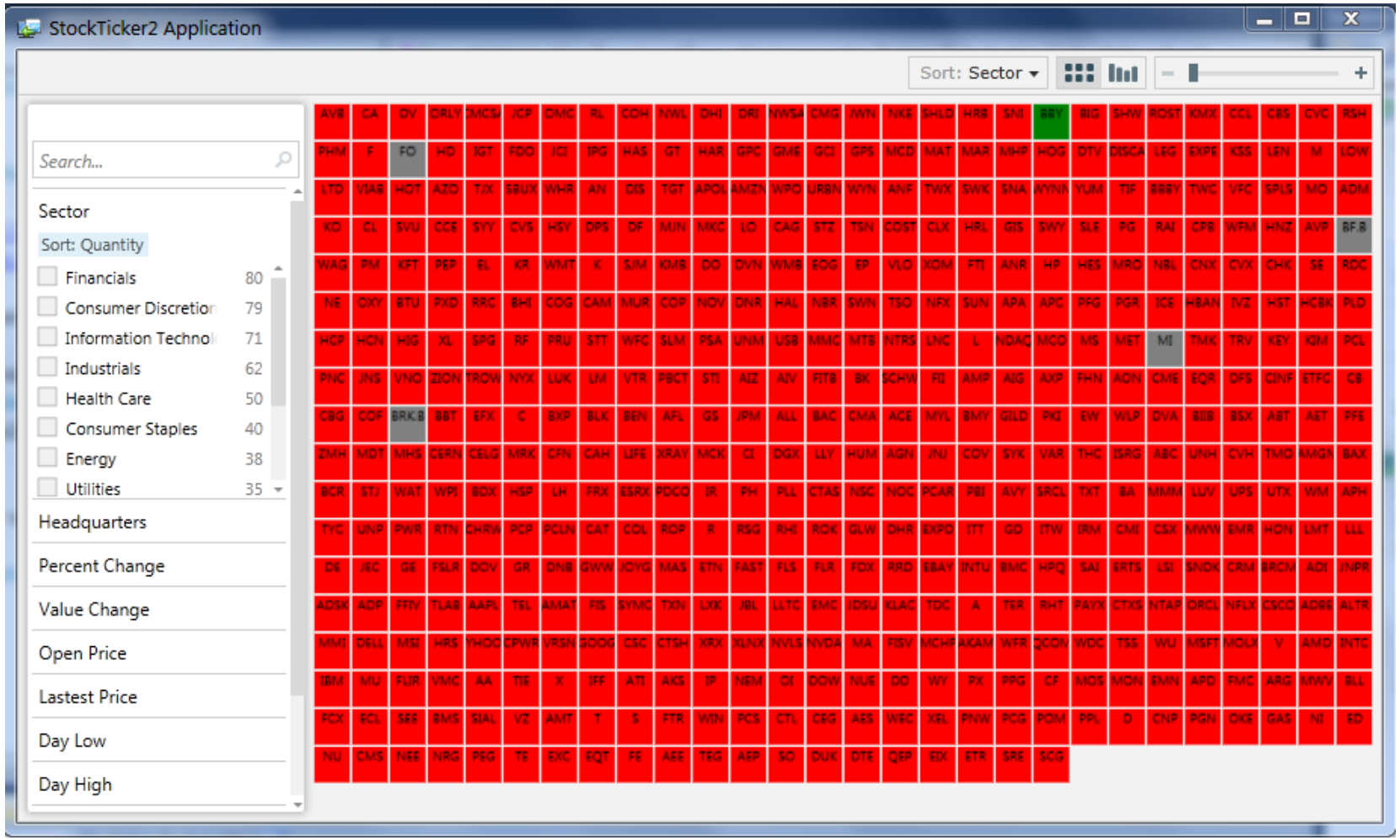
Data Science and Data Scientist



Data Science and Data Scientist

blind or guided data-mining

数据挖掘 - 盲目的，还是有的放矢



Microsoft Research (MSR)

- Founded in 1991
  - Staff of 850 + in 60+ disciplines
- International research teams
  - MSR Redmond, Cambridge, Asia, Silicon Valley, India, New England
- A “Safe house” for incubating technologies/ideas
  - Not bound to product cycles
  - Support long-term research in computer-science and eScience
- A environment for research collaboration
  - Sabbaticals, New Faculty Fellowships, Post-docs, interns, …
  - Microsoft Research Connections

微软研究院 – 研究计算机科学和计算科学的安全港

Earth, Energy, and Environment (E^3) at MSR

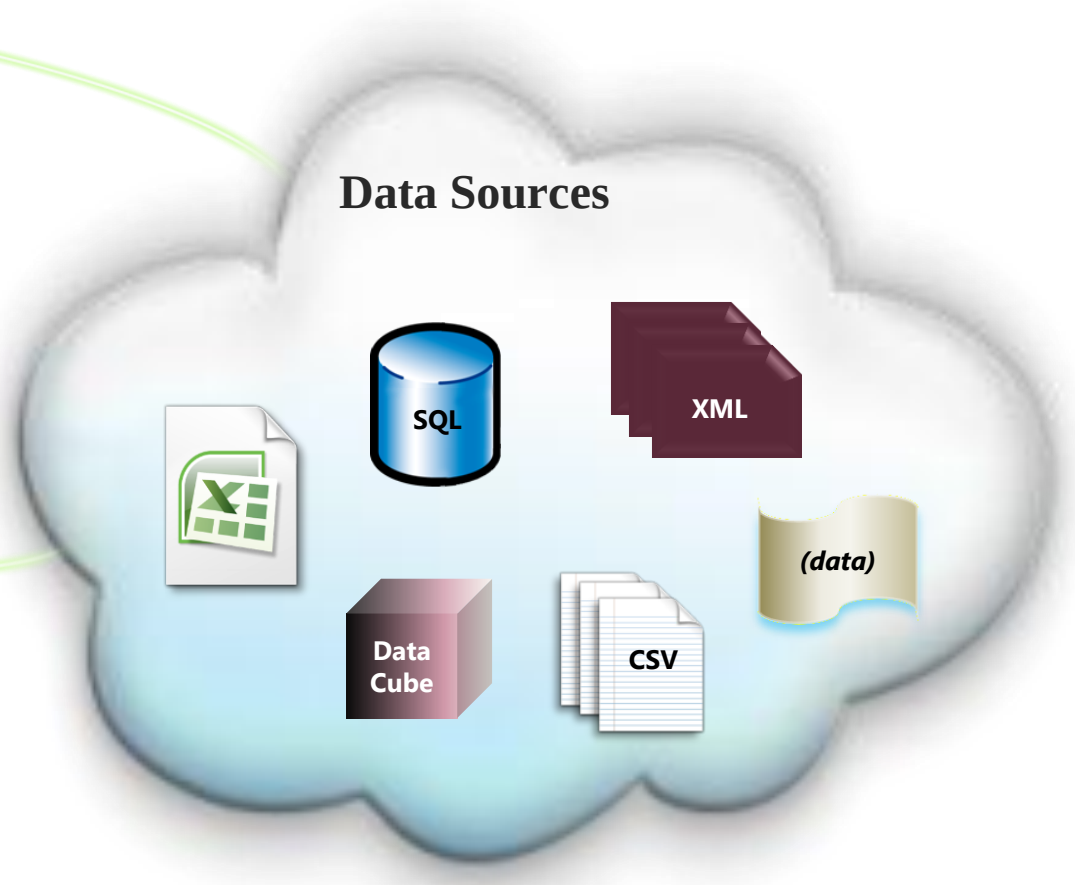
- Engage Microsoft technologies with scientific research
  - Astronomy and astrophysics
  - Environmental sciences
  - Biodiversity
  - Bioenergy
- Bridge gaps between research and higher education
  - Problem-based learning (PBL)
  - Computational thinking
- Innovatively support science outreach
  - e.g. WWT Community Beijing

E^3 工程 - 用微软的最新技术促进现代天文，地理，和环境的科研，科教，及科普

Informatics Framework

Common Problems with Data

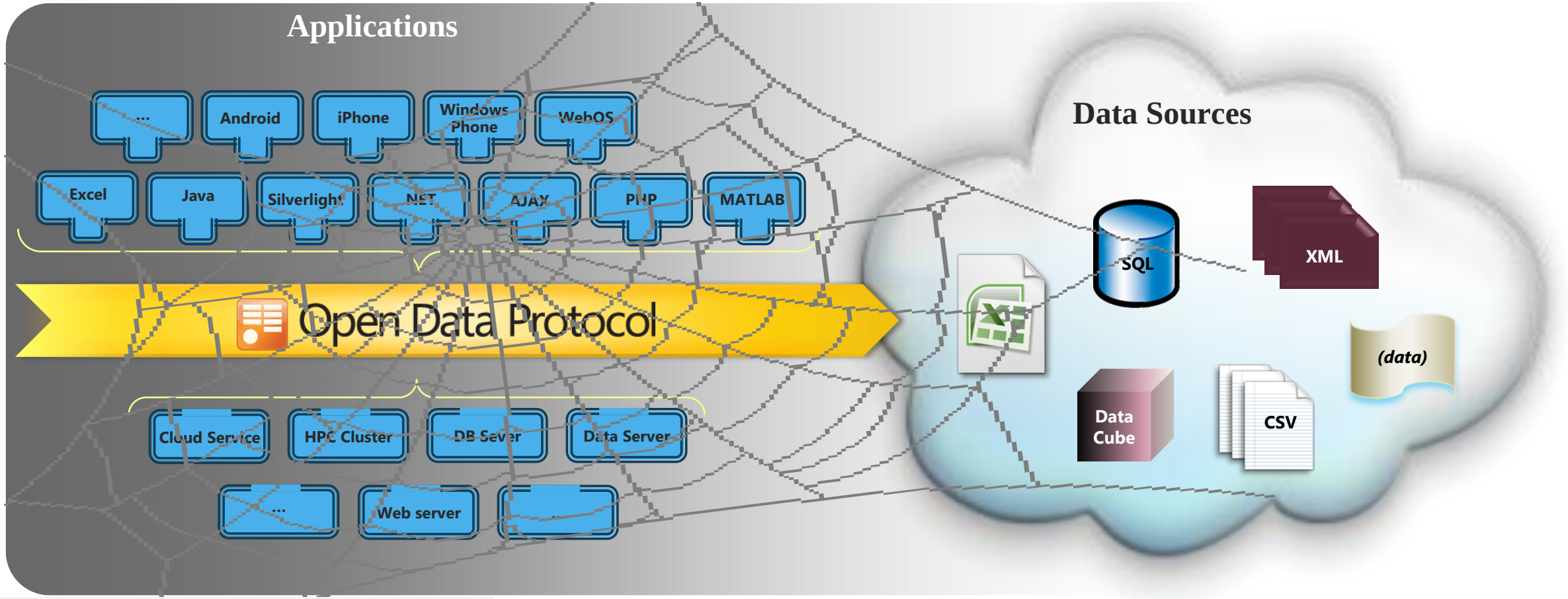
- To use data from different sources
  - Non-standard formats, scales, and units
  - Lack of data quality control
  - Lack of metadata
  - Difficult to repurpose data for different (my) tools
- To share data
  - Lack of incentive (no credit)
  - Need extra resources and tools
- Hidden problems, seldom addressed
  - Versioning
  - Provenance
  - Curation



信息学架构 – 对焦海量科学数据所带来的基本问题，即数据的分享，使用，及管理

Informatics Framework

Current State of Data Ecosystem



信息学架构 – 明辨目前科学数据的生态，采用OData解决数据共享的基本障碍

Informatics Framework

Advance data discoverability, accessibility, and consumability

Open Data Protocol (OData)

http://www.odata.org

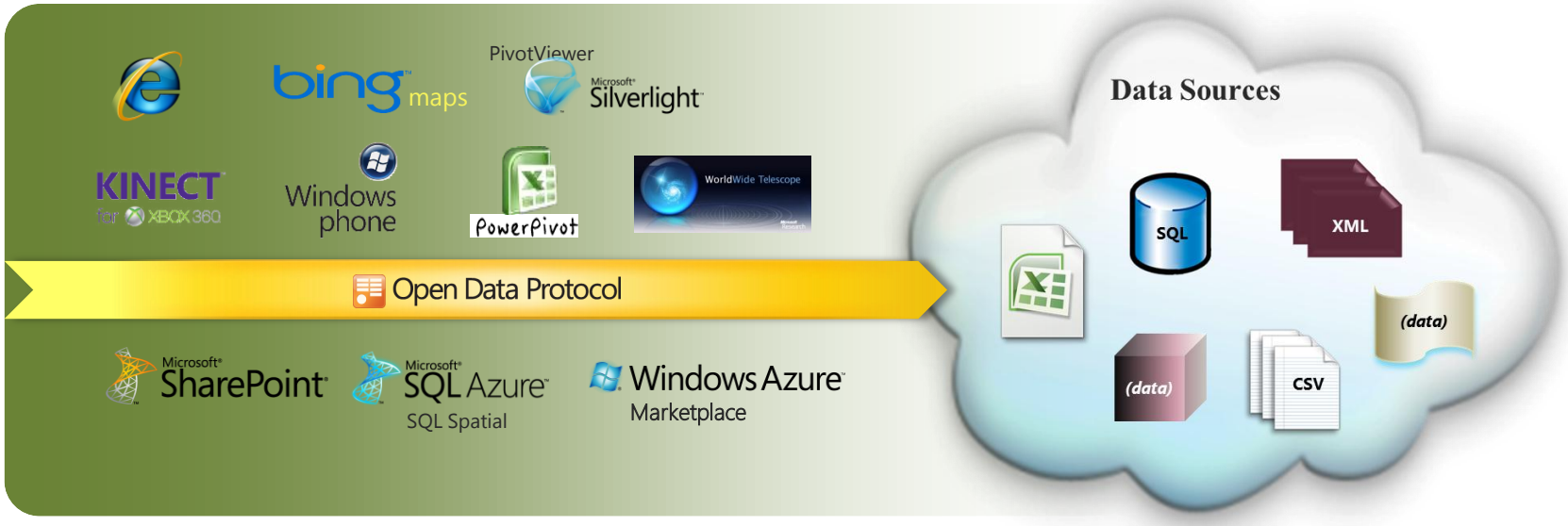
It allows you to form URLs based on what you know about the underlying data

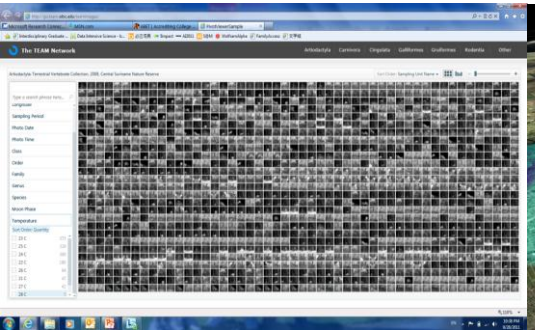
- A Web protocol for querying and updating data
  - provides a way to unlock your data and free it from data silos
  - does this by building upon Web technologies such as HTTP, Atom Publishing Protocol (AtomPub) and JSON to provide access to information from a variety of applications, services, and stores.
- In Open Source/Specifications Promise
- An application of a set of internet standards:
  - HTTP,
  - Atom (RFC 4287),
  - AtomPub (RFC 5023),
  - REST semantics
- Existing standards + easy data access API
- Adding **Geospatial data support** –
  - Feedback from the Community encouraged – www.odata.org

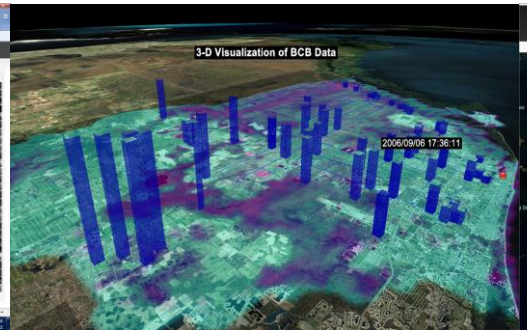
OData将数据的浏览变得象网页的浏览一样容易

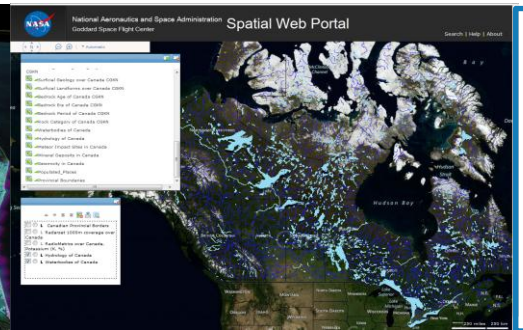
Informatics Framework

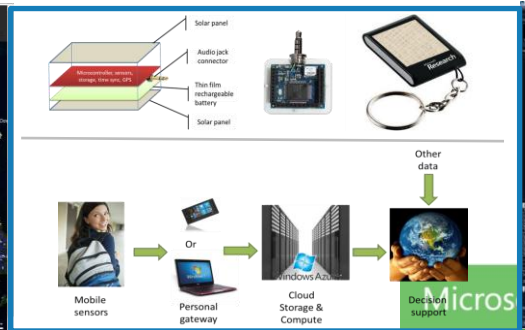
Advance data discoverability, accessibility, and consumability

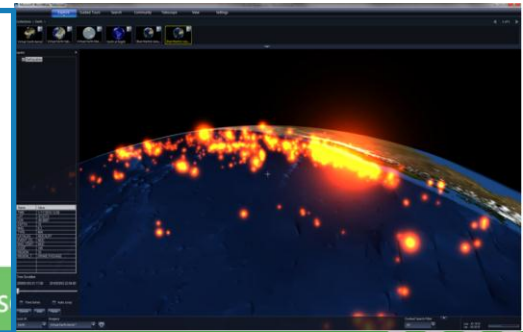












Microsoft®
Research Connections