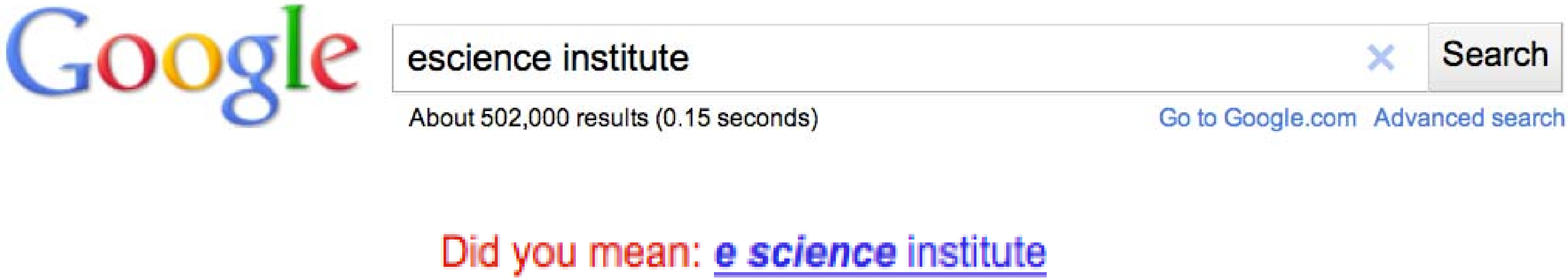
Speculating on Scientific Collaboration Futures

Bill Howe

eScience Institute

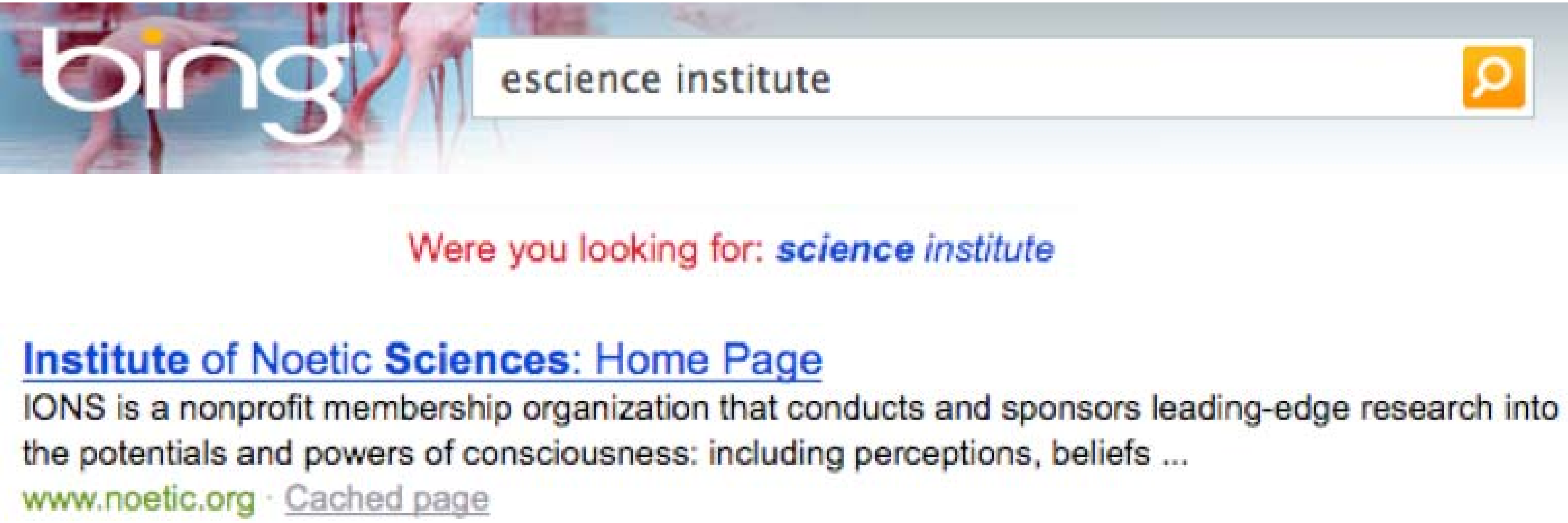
http://escience.washington.edu



escience institute

About 502,000 results (0.15 seconds)

Did you mean: ***e science*** institute



escience institute

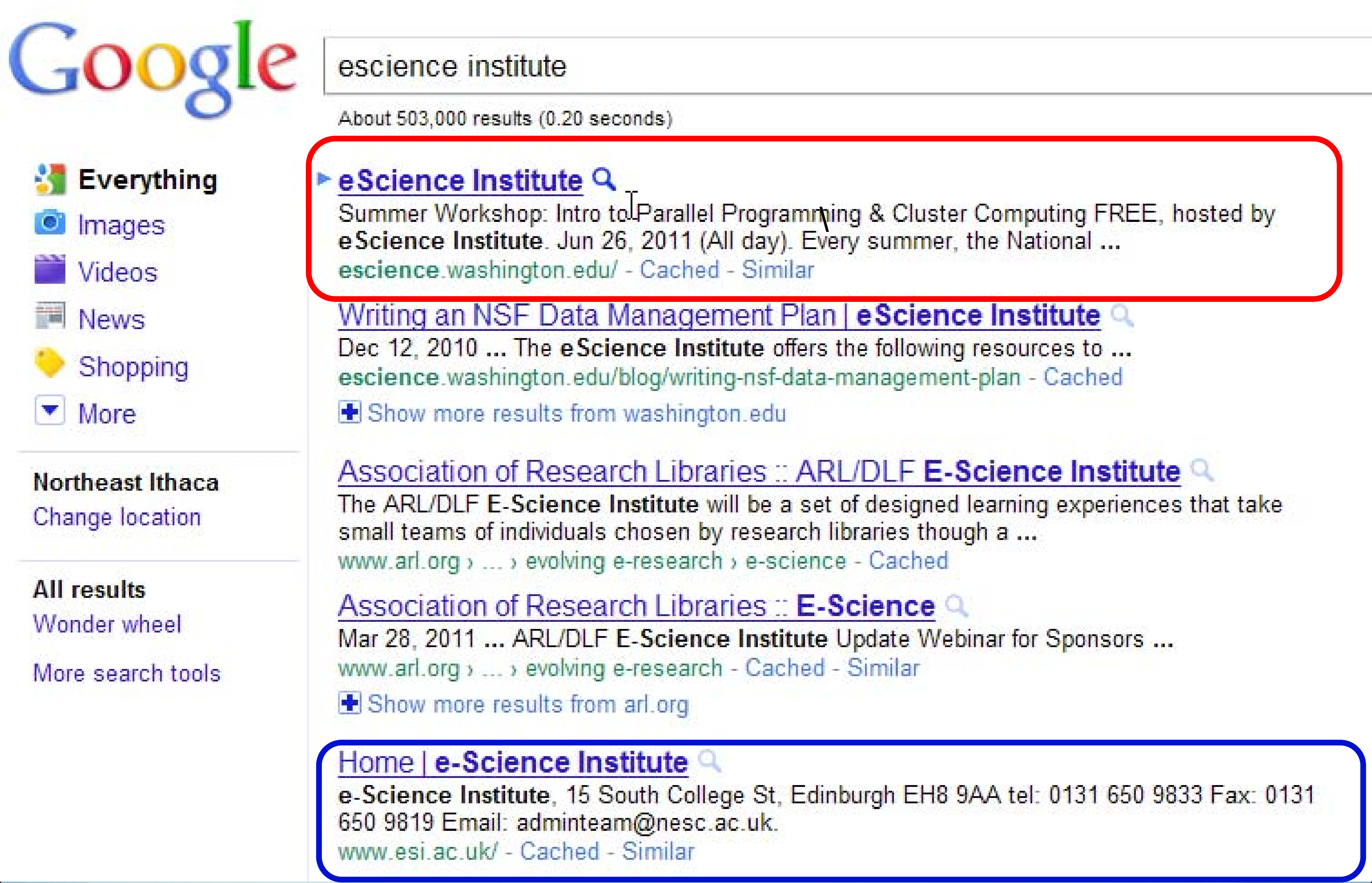
Were you looking for: ***science*** *institute*

Institute of Noetic **Sciences**: Home Page

IONS is a nonprofit membership organization that conducts and sponsors leading-edge research into the potentials and powers of consciousness: including perceptions, beliefs ...

www.noetic.org · Cached page

Viewed summer 2009



escience institute

About 503,000 results (0.20 seconds)

eScience Institute
Summer Workshop: Intro to Parallel Programming & Cluster Computing FREE, hosted by **eScience Institute**. Jun 26, 2011 (All day). Every summer, the National ...
escience.washington.edu/ - Cached - Similar

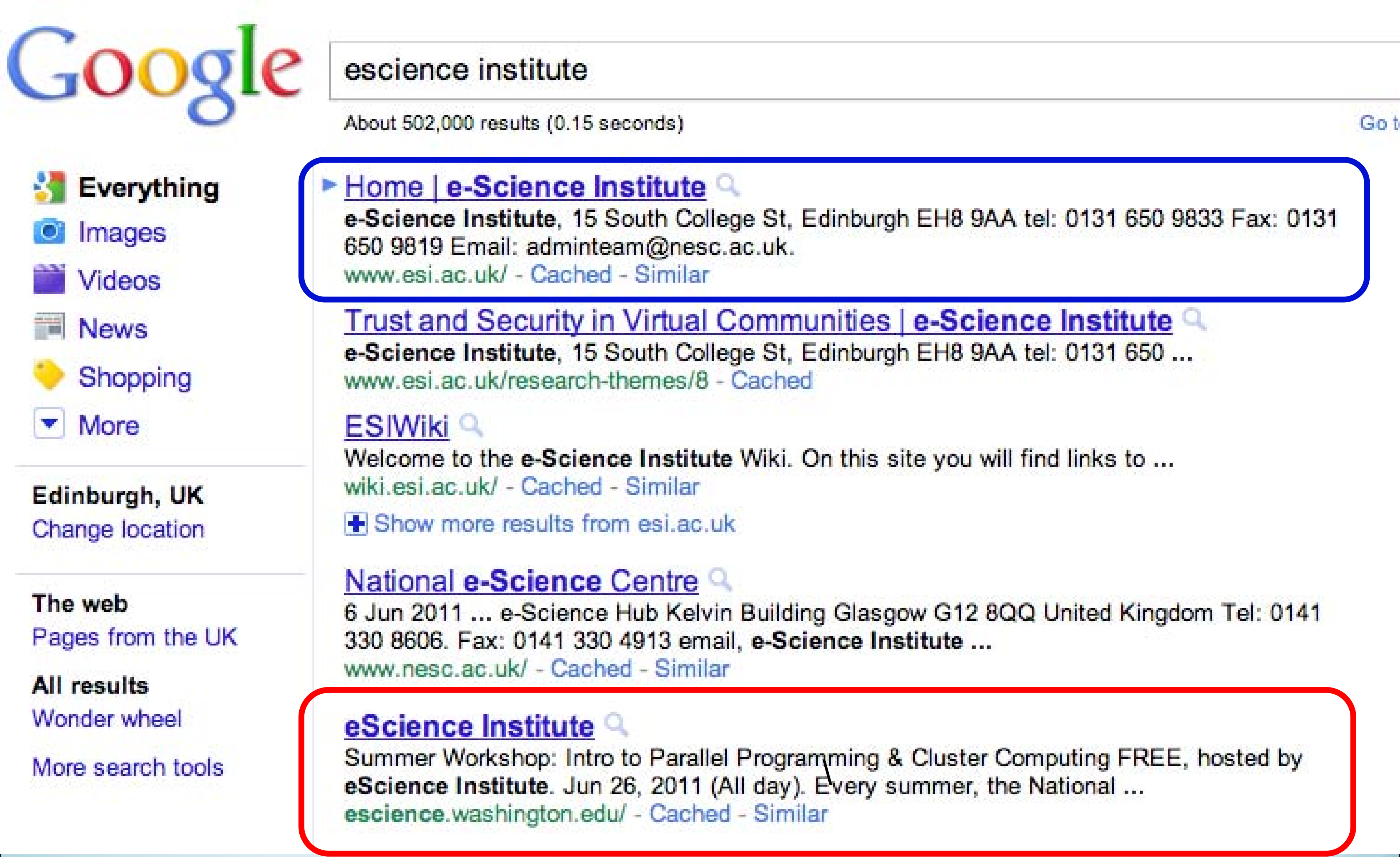
Writing an NSF Data Management Plan | **eScience Institute**
Dec 12, 2010 ... The **eScience Institute** offers the following resources to ...
escience.washington.edu/blog/writing-nsf-data-management-plan - Cached
Show more results from washington.edu

Association of Research Libraries :: ARL/DLF **E-Science Institute**
The ARL/DLF **E-Science Institute** will be a set of designed learning experiences that take small teams of individuals chosen by research libraries though a ...
www.arl.org › ... › evolving e-research › e-science - Cached

Association of Research Libraries :: **E-Science**
Mar 28, 2011 ... ARL/DLF **E-Science Institute** Update Webinar for Sponsors ...
www.arl.org › ... › evolving e-research - Cached - Similar
Show more results from arl.org

Home | **e-Science Institute**
e-Science Institute, 15 South College St, Edinburgh EH8 9AA tel: 0131 650 9833 Fax: 0131 650 9819 Email: adminteam@nesc.ac.uk.
www.esi.ac.uk/ - Cached - Similar

Viewed June 2011, Seattle, WA, USA



escience institute

About 502,000 results (0.15 seconds)

Home | e-Science Institute
e-Science Institute, 15 South College St, Edinburgh EH8 9AA tel: 0131 650 9833 Fax: 0131 650 9819 Email: adminteam@nesc.ac.uk.
www.esi.ac.uk/ - Cached - Similar

Trust and Security in Virtual Communities | e-Science Institute
e-Science Institute, 15 South College St, Edinburgh EH8 9AA tel: 0131 650 ...
www.esi.ac.uk/research-themes/8 - Cached

ESIWiki
Welcome to the **e-Science Institute** Wiki. On this site you will find links to ...
wiki.esi.ac.uk/ - Cached - Similar

Show more results from esi.ac.uk

National e-Science Centre
6 Jun 2011 ... e-Science Hub Kelvin Building Glasgow G12 8QQ United Kingdom Tel: 0141 330 8606. Fax: 0141 330 4913 email, **e-Science Institute** ...
www.nesc.ac.uk/ - Cached - Similar

eScience Institute
Summer Workshop: Intro to Parallel Programming & Cluster Computing FREE, hosted by **eScience Institute**. Jun 26, 2011 (All day). Every summer, the National ...
escience.washington.edu/ - Cached - Similar

Viewed June 2011, Edinburgh, UK

The University of Washington eScience Institute

- Rationale
  - The exponential increase in physical and virtual sensing tech is transitioning all fields of science and engineering from *data-poor to data-rich*
  - Techniques and technologies include
    - Sensors and sensor networks, data management, data mining, machine learning, visualization, cluster/cloud computing
  - If these techniques and technologies are not widely available and widely practiced, UW will cease to be competitive
- Mission
  - Help position the University of Washington and partners at the forefront of research both in modern eScience techniques and technologies, and in the fields that depend upon them.
- Strategy
  - Bootstrap a cadre of Research Scientists
  - Add faculty in key fields
  - Build out a “consultancy” of students and non-research staff
- Funding
  - $1M/year direct appropriation from WA State Legislature
  - augmented with soft money from NSF, DOE, Gordon and Betty Moore Foundation

eScience Data Management Group

**Bill Howe, Phd (databases, visualization, data-intensive scalable computing, cloud)

Staff

- **Garret Cole (cloud computing (Azure, EC2), databases, web services)
- Keith Grochow (visualization, earth science, graphics, cloud computing)
- Marianne Shaw, Phd (health informatics, semantic web, RDF, graph databases)
- Alicia Key (visualization, user-centered design, web applications)

Students

- Leilani Battle (undergrad), databases, performance evaluation
- Yuan Zhou (masters, Applied Math), machine learning, ranking, recommender systems

Partners

- **UW Learning and Scholarly Technologies (web applications, QA/support, release mgmt)
- **Cecilia Aragon, Phd, Associate Professor, HCDE (visualization, scientific applications)
- Magda Balazinska, Phd, Assistant Professor, CSE (databases, cloud, DISC)
  - YongChul Kwon Phd, databases, DISC, scientific applications (advisor: Balazinska)
  - Nodira Khoussainova, databases, machine learning (advisors: Balazinska, Suciu)
- Dan Suciu, Phd, Professor, CSE, (probabilistic databases, theory, languages)
  - Paraschos Koutris, theory, distributed computing

*** funded in part by eScience core budget*

What will scientific collaborations look like in 20 years?

Selected Characteristics of “The Computer”

- It’s never the bottleneck
  - No one ever swears at it
- How?
- All data addressable
- All operations composable
  - “Computer, apply X to Y”
- Zero latency
- Fancy Interfaces
  - Declarative interfaces for input (voice, NLP)
  - Intuitive visual interfaces for output

All data addressable

- One logical namespace
- Explicit data movement is never required
- Implicit data movement optimized appropriately

All operations composable

- Logical compatibility implies physical compatibility
  - No explicit typecasting file format conversions
- No distinction between “inside the DB” vs. “outside the DB”
  - “in situ” data [SciDB]
  - amortizing load cost [Ailamaki, Kersten]
- Incremental structuralization/schemafication
  - Extract Tables, Graphs, Trees, Arrays from files, incrementally
  - “Recognizers” to perform the information extraction
  - Pig (Yahoo), SCOPE (MS), [Ailamaki 2010]
- “Soft Schemas”
  - “Guess” the type, explore the consequences

Aside: There will always be data born “in the wild”

- No schema, certainly no ontology, weird format, shitty metadata
- There is no difference between debugging and formal experiments.
  - When it works, it's an experiment.
  - When it doesn't, it's debugging.
- “Free” trial and error is a beautiful property of computational science
  - Be conservative about limiting this freedom
- Need to embrace the chaos, not legislate it away

Zero latency

- “Semantic pre-fetching”
  - Choose an “important” and compatible pair (*f*, *X*)
  - Pre-generate *f*(*X*)
  - Solicit review from users
  - Incorporate feedback
  - “hypothesis generation”

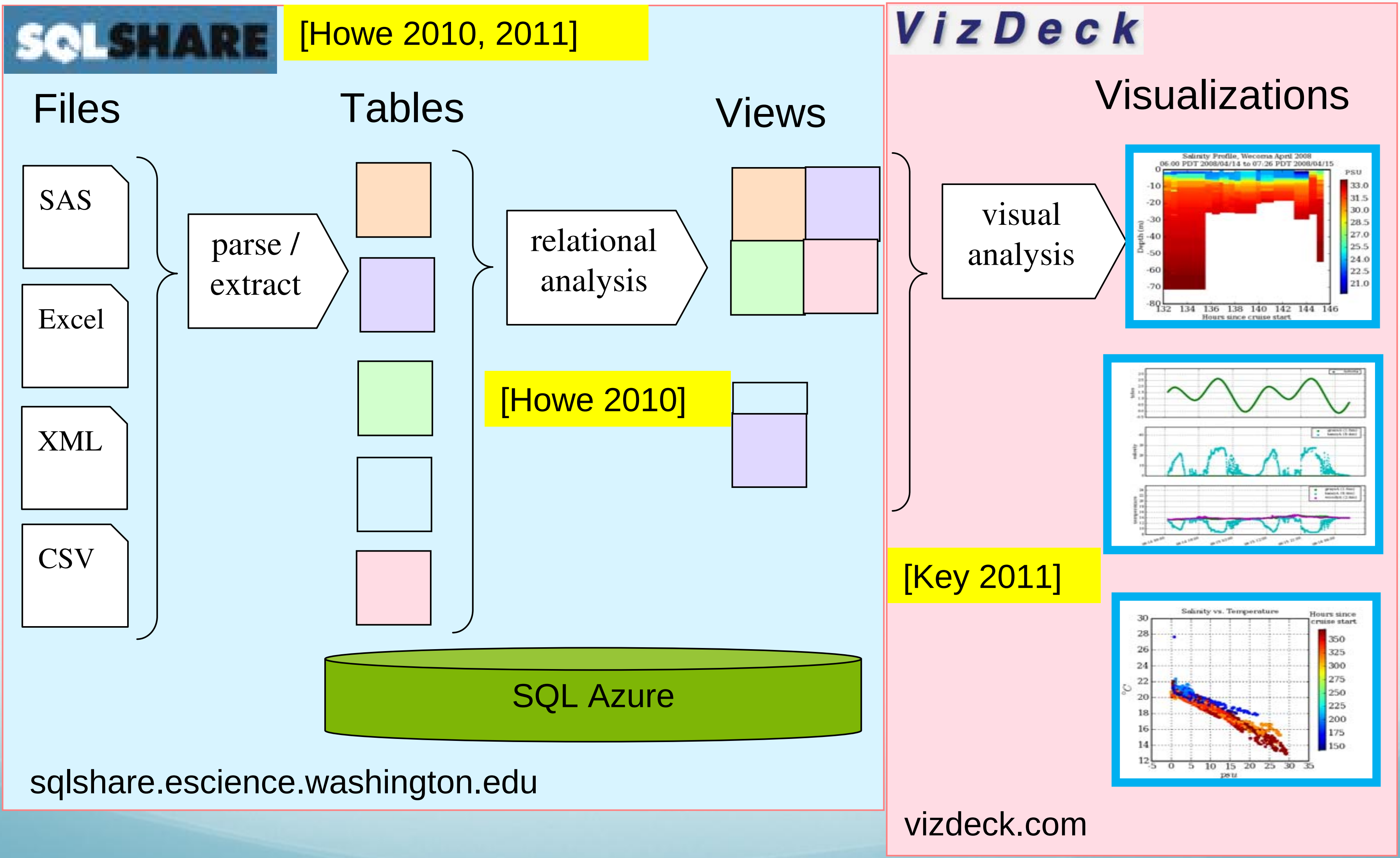
What breakthroughs are required?

- All data addressable
  - *Universal uptake of cloud computing; significant price reduction***
- All operations composable
  - *Soft schemas; in situ data; incremental structuralization*
- Zero latency
  - *Speculative, proactive execution*

*** All data import is now free; all new users get a free micro instance for a year; compute costs have dropped 80%; storage costs have dropped 50%*

Relevant Technologies

Where we are



SQLSHARE

[Howe 2010, 2011]

Files

Tables

Views

SAS

Excel

XML

CSV

parse / extract

relational analysis

[Howe 2010]

SQL Azure

sqlshare.escience.washington.edu

VizDeck

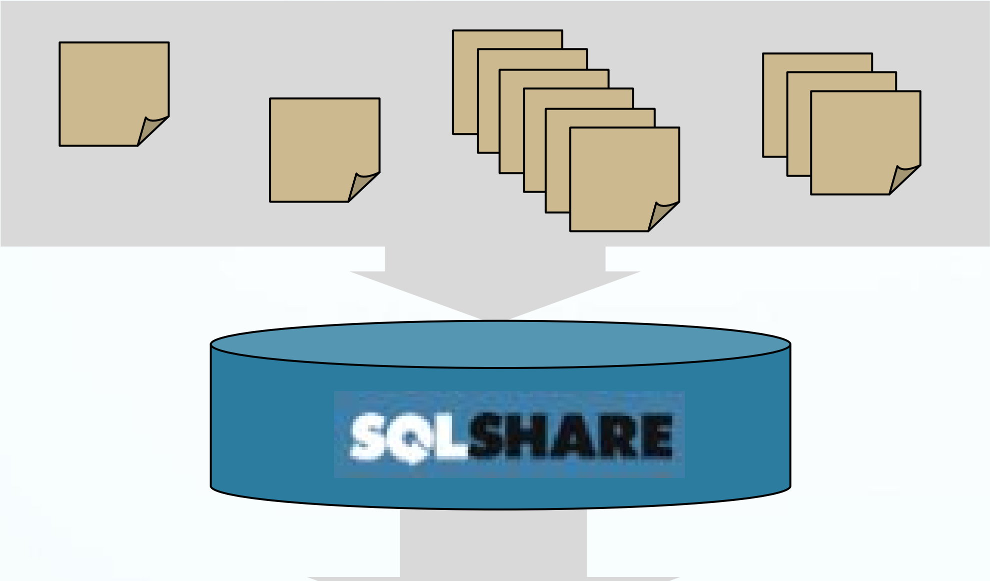
Visualizations

visual analysis

[Key 2011]

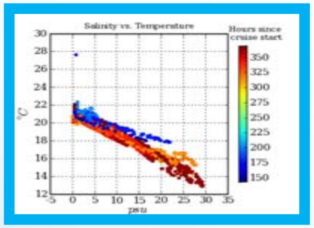
vizdeck.com

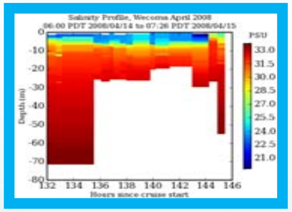
Where we’re headed

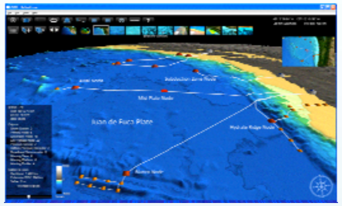


- 1000s of sources
- unknown structure
- unknown semantics
- unknown quality
- unknown relationships

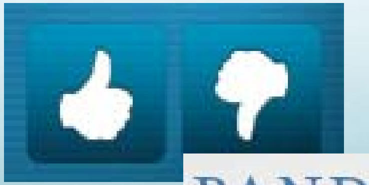
The only query that matters:
“show me what’s important”







Automatically suggest
- schemas
- queries
- visualizations
- predictive models



Reduce application design to a series of simple decisions

Takeaways

- All code and all data will be born, live, and die in the cloud
  - accessed through your tablet, phone, iDevice
  - *requires: nothing; it's already happening*
- Query and reason about the "derivation space"
  - i.e., everything that the system can potentially create
  - *requires: in situ data; soft schemas; incremental structuralization*
- Speculative, eager, proactive, automatic data mining
  - results presented to researchers for review and feedback
  - "Highlight reel" for unfamiliar data (trends and anomalies)
  - *requires: surplus computing resources; models of what's important*

The future is already here; it's just not very evenly distributed

-- William Gibson

PrePredict

- Same idea, but with machine learning
- Eagerly and proactively apply predictive algorithms to data in the database
- Emit results for review by humans
  - daily, weekly, whatever
- Learn from feedback
- incorporate explicit user interests
  - expressed as queries, hints, etc.
  - Many of the same signals search engines use, but applied to a search space with elements that don’t yet exist

Putting it together: Exploratory Analysis

- The only query is “What’s important here?”
- A narration of your data
- How?
  - Identify trends and anomalies
  - Generate candidate models, visualizations, queries
  - Show the best ones for review

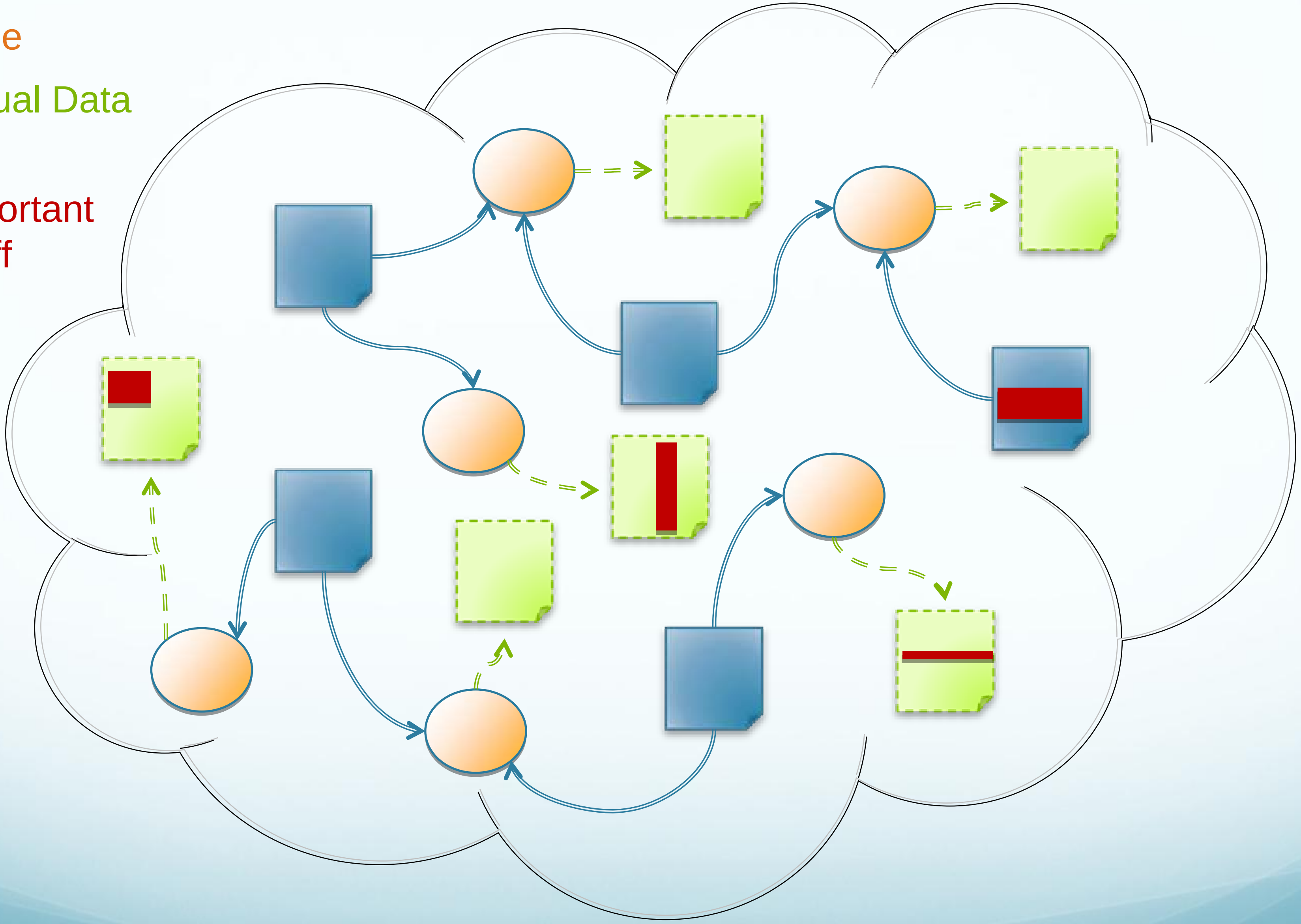
  - [Pandora, Tivo, Netflix]

Data

Code

Virtual Data

Important
Stuff



All code/data in the same logical space

What technologies do we need?

- Data “born” into the cloud
  - It never moves
  - Bring the computation to the data
- A rich and evolving suite of native services for manipulating the data available
  - MapReduce
  - SQL
  - etc.
- Virtual machines for new and custom operations
  - with some special support for parallelism