Breaking down Data Barriers to Drive Open Innovation

Evelyne Viegas

Microsoft Research

evelynev@microsoft.com

Predicting the future is hard

- Technology is rapidly changing
  - Old assumptions invalidated
- New approaches and markets can catch you by surprise
  - From consumers (Web 1.0) to producers (Web 2.0; Web 3.0)
- Hard to predict how people will react to technology
  - SMS in Europe & Asia vs US

What we do know

- Technology will continue to evolve quickly
  - Mainframe, Personal Computer, Devices and Data Centers
- Paradigms will continue to change
  - Software; Software+Services in a global world; Cloud; ?Sky; ?? Space
- Technology companies must continuously innovate

Data Intelligence

Vision – Enable the *Next Generation Internet* by working with Academia, stakeholders from industry, government, and internet consumers/innovators to build an Intelligent and Ubiquitous Safe Web, making sense of data via open innovation

DATA > INFORMATION > KNOWLEDGE

Challenge – How to make real-world, large-scale data available to researchers to nurture innovation and perform valid experimentation, while maintaining data privacy?

Surveys and Data Factoids

- Surveys
  - 66% of people surveyed use the internet as a vehicle for **completing tasks** and a source of information for **making decisions**
  - 75% of people say using search to inform a decision for product purchase

- Factoids (content and behaviours)
  - 2000% increase in the number of Web sites in 10 years
  - 100 thousand times the amount of content online in 12 years
  - 114% increase in video searches in 1 year
  - 43% people use the internet to research what ails them and how to treat it
    - cyberchondria “unfounded escalation of concerns about common symptomatology, based on the review of search results and literature on the Web.”

…

Why Data Intelligence?

- Data has become a natural resource
  - Consumed and **produced** by billions of people interacting with each other
  - Digital interactions are reshaping the communication infrastructure

- Data-driven research
  - **Innovations** – researchers can unveil *new* analysis or *research directions* and explore new questions
  - **Science** – *repeatability of experiments* can be performed and erroneous results avoided
  - **Training** – powerful tool for *training the next generation of researchers*

Roadblocks to Data-driven Research

- Data confidentiality can become a roadblock to open innovation in an information-based society
  - Lots of data in vaults of information industries
- Data regulations can slow down open innovation
- Data “consumers” & “producers” need to buy into an open data model for enabling research and drive innovation

A Face Is Exposed for AOL Searcher No. 4417749

The New York Times
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

"Search query data can contain the sum total of our work, interests, associations, desires, dreams, fantasies, and even darkest fears."

Scrubbed geo-location data not so anonymous after all

Anonymized data collected from GPS-enabled devices may not be as anonymous as you think, according to researchers who show that knowing someone's general home and work locations can be enough to identify an individual uniquely.

May 21, 2009, The Register

Multi-faceted solution to enable Data-driven Research

More effort needed on

- Technical solutions
- Legal frameworks
- Societal buy-in
- Businesses support

Find the stakeholders' incentives to contribute to enabling data-driven innovation

Accelerating Search in Academic Research

- Received over 200 proposals from 28 countries

- Search RFP Awards
  - Search assets (15 million search queries + click through)
    - Personal Identifiable Information (PII) (including inadvertent) removed
    - Provided under a limited data licensing agreement
  - Increased quota to the Search API

- Search Summit 2007
  - The Quest for Assets – the Good the Bad and the Wanted
    - Good: having data is good
    - Bad: more transparency in data
    - Wanted
      - More data, larger scale
      - Need userID
      - …

Beyond Search – Semantic Computing and Internet Economics

- Beyond Search – Semantic Computing and Internet Economics
  - Search and ad assets (100 million search queries + click through)
    - PII removed
    - Provided under a limited data licensing agreement
  - Increased quota to the Search API
- Virtual Earth Awards
  - Ground Images
    - Provided under a limited data licensing agreement

Enabling and Advancing Internet Research at Scale

- ata Confidentiality 2007
  - NSF supported, Co-sponsored by IBM, Microsoft, NSF
  - Participation from 13 federal agencies; 7 industries; 18 universities
    - (a) **better government**: reliable technology preventing agencies from accidentally compromising privacy
    - (b) **better science**: access to more and richer data through private data analysis
    - (c) **better industry**: personalised search; vendor analysis/testing/bug reporting

rivate Lives and Public Policies: Confidentiality and Accessibility of Government Statistics (Committee on National Statistics, NRC and the Social Science Research Council, National Academy Press, 1993)

xpanding Access to Research Data: Reconciling Risks and Opportunities (Committee on National Statistics, NRC and the Social Science Research Council, National Academy Press, 2005)

Towards Cloud Data Services for Research

- Driving Open Innovation with Large Scale Real World Data is a challenge
  - "anonymized" slice of data + limited licensing agreement
    - Maximizes privacy at the expense of research questions
    - Does not scale very well to the community of researchers

- Cloud Data Services
  - Interactive data access framework maximizing scientific research while preserving privacy (user, business)
  - Privacy-*preserving* data analysis with Differential Privacy (Dwork, 2006)

The Web as a Global Lab

- Summary
  - Data has become a natural resource, yet large portions still untouched
  - Need to understand incentives to enable data-driven research
  - More research needed on technical, legal, societal sand business solutions to access data safely

- Call to action
  - How does the web (content, interactions) contribute to the innovation process?
  - How can YOU extract value from web data?
    - Research questions?
    - Experiments to run?
    - …

contact me: evelynev at microsoft dot com