Advancing Science: OSTI's Current and Future Search Strategies

Jeff Given
IT Operations Manager
Computer Protection Program Manager

Office of Scientific and Technical Information
U.S. Department of Energy

November 14th, 2007

About OSTI

A U.S. Department of Energy program within the Office of Science

- Maintains appropriate public access to DOE research results

  All collections of scientific and technical information resulting from R&D activities generated from the facilities within the national DOE complex

- Provides stewardship for the Department's 60-year legacy of classified and unclassified scientific and technical reports
- Maintains an electronic repository of over 4 million DOE-produced R&D records dating to 1940s

About OSTI

OSTI accelerates the advancement of discovery by speeding access to R&D findings.

- Science.gov - 50 million pages of U.S. government science information from 17 US Government science organizations
- WorldWideScience.org - 200 million+ pages of international research information from the governments of 17 countries
- Science Accelerator - federated search of important DOE databases such as E-print Network (includes 1 million documents & 27,000 Web sites) and Information Bridge (includes over 145,000 DOE full text reports)

Overview

- Users and Search
- Current problems with search and retrieval
- OSTI strategy for overcoming problems
- Future and current work

Do You Know?

- How big is the web?
- How much of the worlds information is on the web?
- How similar are the major search engines in terms of search results?
- What percentage of a typical web site's functionality is actually used?

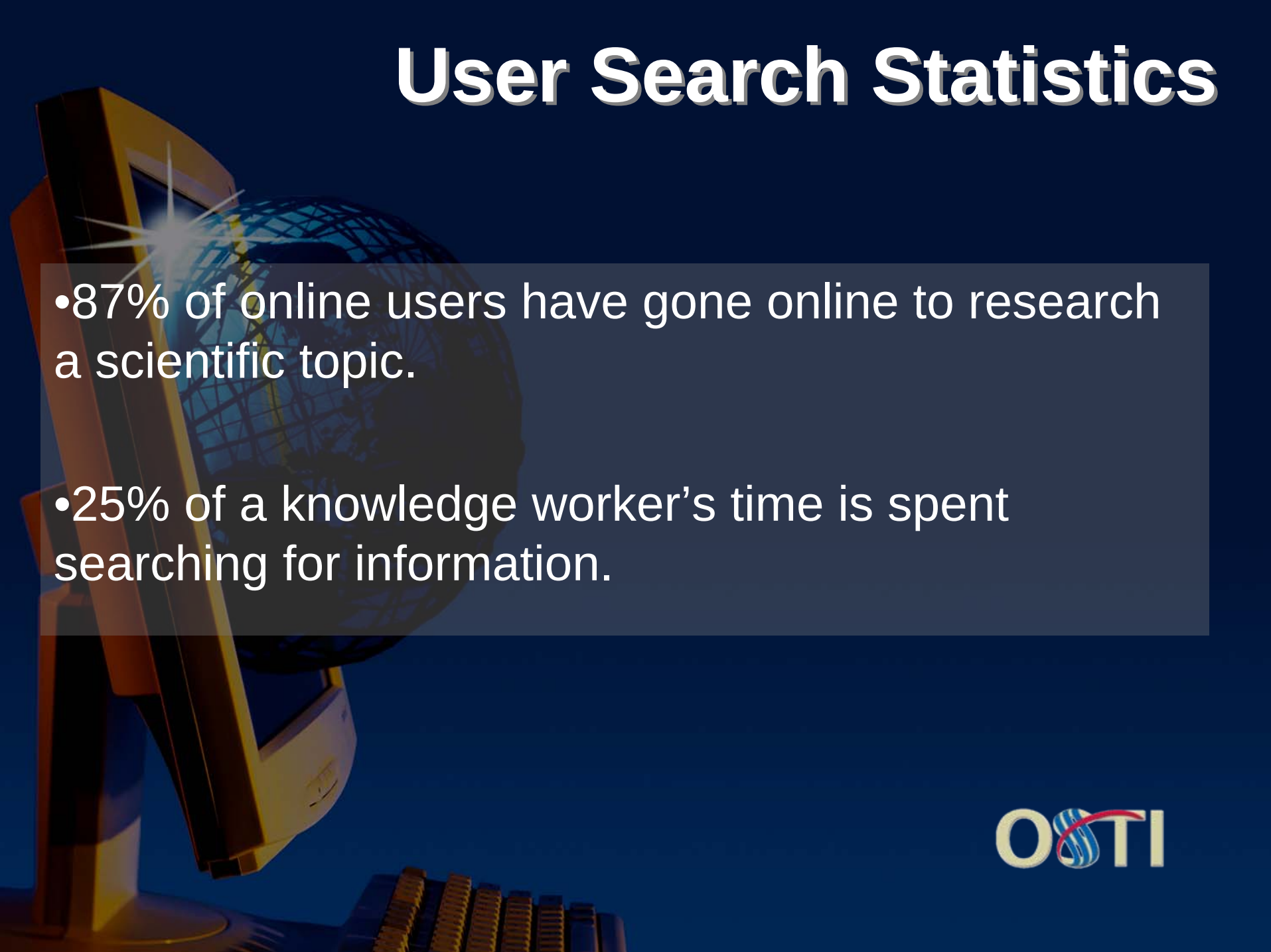
Users and Search

- User Goals:
  - Find authoritative and relevant information.
  - Users don't want to search, they want to get something done.
- Broad scope search engines
  - Google, Yahoo, MSN (GYM)
- Narrow scope search engines
  - Specialized, Topical, Vertical
  - PubMed, music.yahoo.com, Information Bridge

Search - Data Availability

- The web now encompasses over 100 million web sites (and a far larger number of pages).
- The deep web (non-Googleable) has been estimated to be several magnitudes greater than the surface web.
- Only about 5% of the world's total information is online today.
- Only 15% of DOE's R&D information is full text searchable on the internet.

User Search Statistics



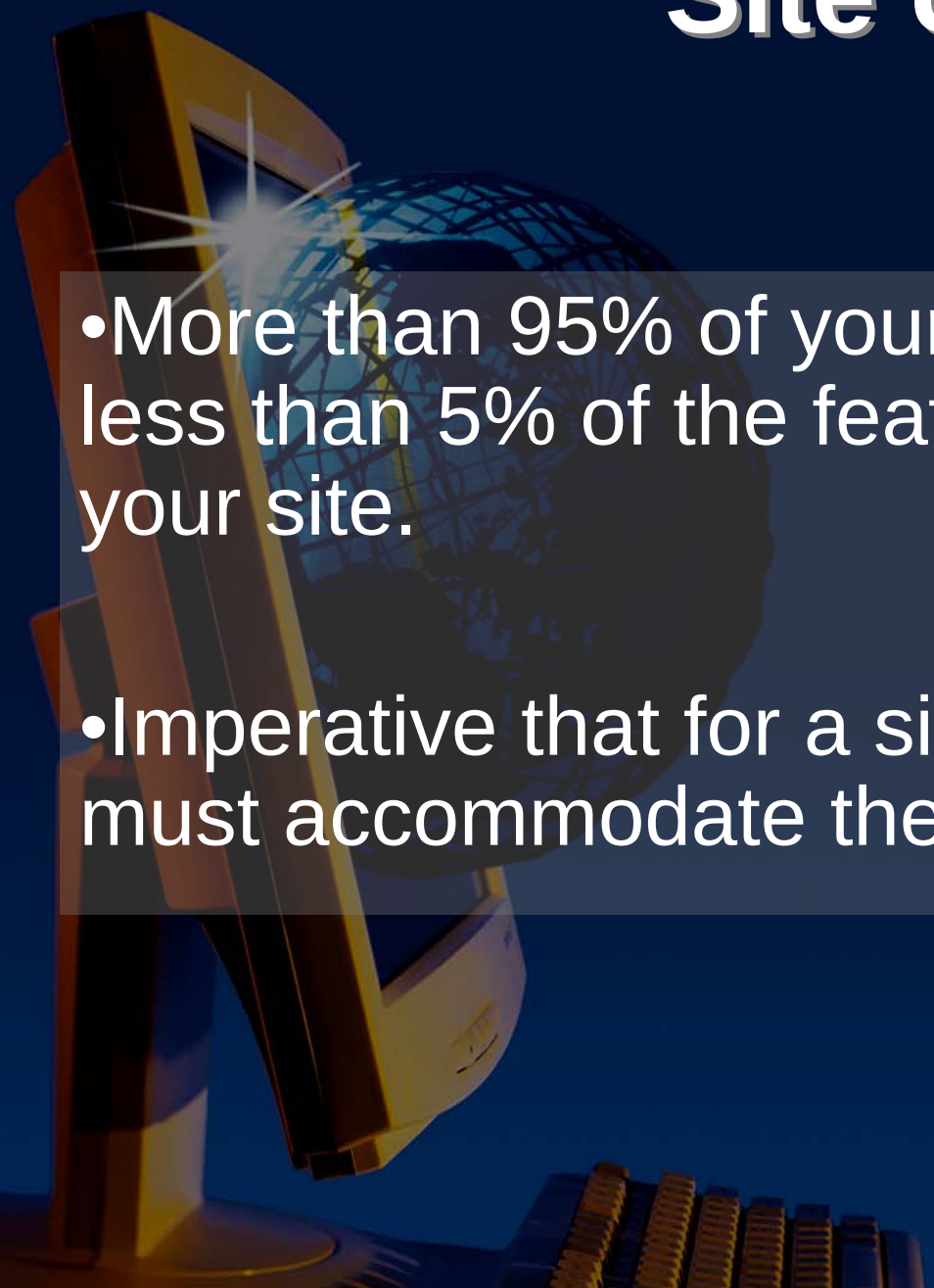
- 87% of online users have gone online to research a scientific topic.

- 25% of a knowledge worker's time is spent searching for information.

Relevancy Bias

- The conventional wisdom is that the major search engines serve up similar results.
- Survey participants reported ~70% overlap in the top 10 results on Google and Yahoo!.
- Using the 500 most popular search terms, on average, Google and Yahoo! share only 3.8 of their top 10 results.
- ~5% of searchers go beyond page #1

Site Usage Statistics



- More than 95% of your customers will use less than 5% of the features and functions of your site.

- Imperative that for a site to be successful it must accommodate the typical user.

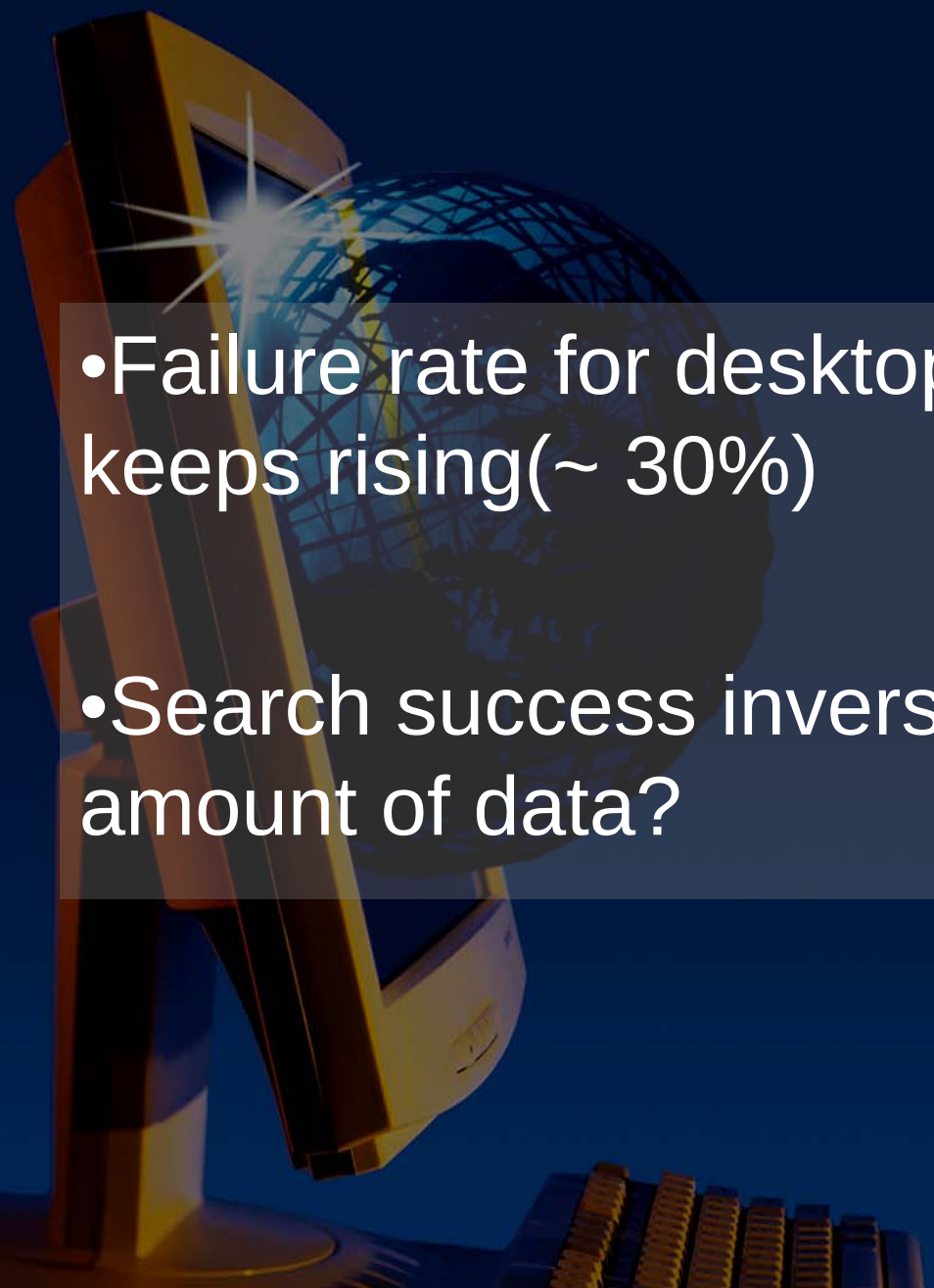
Users and Search Summary

- Users want authoritative, relevant information fast and easy
- Search is prevalent, information users spend a significant portion of their time searching
- Not all data is online, and not all information available online is included in GYM searches
- If relevancy rankings don't return "relevant information" on the first page – the data is not found most of the time

Problem Areas

- Users want authoritative, relevant information fast and easy
- Search is prevalent, information users spend a significant portion of their time searching
- Not all data is online, and not all information available online is included in GYM searches
- If relevancy rankings don't return "relevant information" on the first page – the data is not found most of the time

Problem Areas



- Failure rate for desktop information seekers keeps rising(~ 30%)

- Search success inversely proportional to amount of data?

OSTI's Focus

OSTI's focus has been and remains to make scientific and technical information searchable and retrievable.

OSTI Strategies

Distribution of DOE content to major search engines.

- –Sitemap Protocol – low development time, low maintenance, reduces amount of unnecessary repeated data requests from crawlers
- –Allows for nearly 100% coverage for each content source
- –~60% of October's traffic to Information Bridge were from Google referrals

OSTI Strategies

Enabling vertical search capabilities to authoritative, relevant Scientific and Technical Information (STI)

- –Federated search - Includes authoritative, subject-matter relevant searches of Deep Web Content
- –Web harvesting - Includes content harvested/crawled from authoritative, subject matter specific URLs

OSTI Strategies

- Development and maintenance of DOE STI data collections
  - Information Bridge
  - Energy Citations
  - DOE Patent Database

OSTI Strategies

- Attribution to source of data
  - –Makes users finding data via search engines aware of the source of data
  - –Users more likely to bookmark and re-visit high quality vertical search engines

OSTI Strategies - Overview

Content distribution via major search engines

+

Providing STI specific vertical search capabilities enabled via Federated Search and Web Harvesting

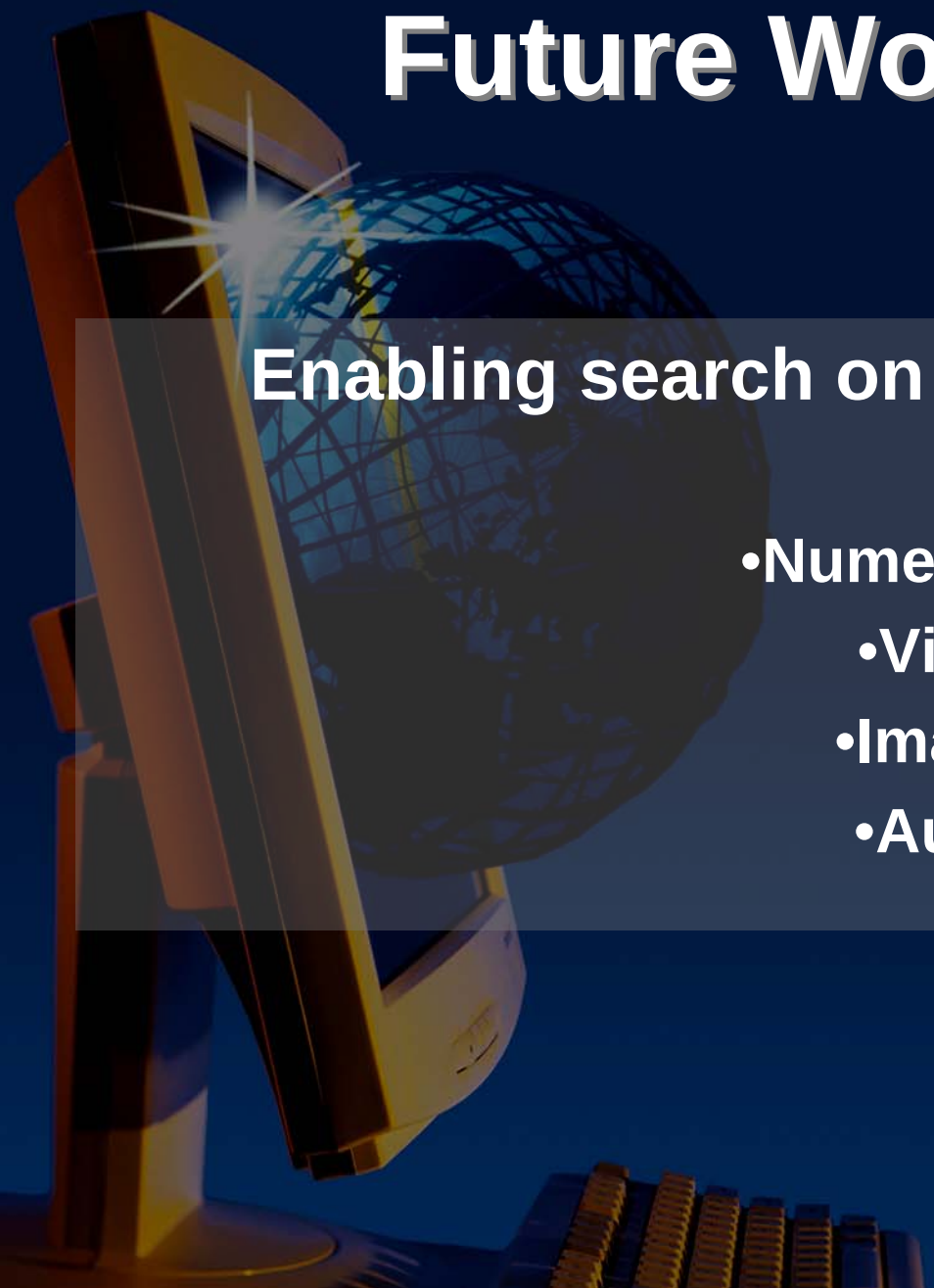
+

Increasing awareness of OSTI vertical search applications via attribution on search engine referrals

=

More users getting the most relevant results from swath of available internet

Future Work – Data Types



Enabling search on non-text information

- **Numeric Data**
- **Video**
- **Images**
- **Audio**

Future Work – Mobile

30% search failure rate tolerable for desktop, not necessarily true for mobile device searches

Ipsos Insight's 2005 "The Face of the Web" study shows significant increases in: ownership of mobile phones, mobile surfing by mainstream users, and adoption of wireless mobile technology by adults aged 35 and older.

Digital natives two thumb typing at incredible speeds (est. 1.5 digital natives in Japan can type at equivalent desktop speeds of 100 words / min)

Future Work – Visualization & Social Tools

- Visualization – identification of scientific communities (publishing groups) and cross over areas in scientific research
- Social Tools
  - –75% of a user's time spent on top news sites is spent reading user comments about the story, and only 25% on the story itself
  - –Over 60% of web content utilized by users age 25 and under is user generated

Future Work

- Utilize HCI labs and testing results to optimize web sites
- Expand reach of federated search by adding additional deep web content
- Add functionality to OSTI's federated vertical search engines

* *CompletePlanet.com – searchable directory of Deep Web sources*

Contact Information

Jeff Given
Office of Scientific and Technical Information
givenj@osti.gov
865.576.1146