NITRD and Big Data

George O. Strawn
NITRD

Caveat auditor

The opinions expressed in this talk are those of the speaker, not the U.S. government

Outline

- What is Big Data?
- Who is NITRD?
- NITRD's Big Data Research Initiative
- Big Data in general
- Big Data in Science and Business

What is Big Data?

- A term applied to data whose *size, velocity or complexity* is beyond the ability of commonly used software tools to capture, manage, and/or process within a tolerable elapsed time.

- What big data isn't (because we know how to do it): transaction processing and relational databases

Big Data includes *Data Intensive Science*

- The science community is a driving force for big data (and it's the NITRD focus)

- But it's often the case that developments in scientific computing have society-wide impact

- And, science often takes advantage of non-scientific computing developments (eg, gpu's, google's map-reduce)

NITRD
Networking and IT R&D

- A 21-year-old interagency program to enhance coordination and collaboration of the IT R&D programs of a number of Federal agencies

- Member agencies

- Areas of interest

NITRD Member Agencies

- DoC
  - NOAA
  - NIST
- DoD
  - OSD
  - DARPA
  - AFOSR, ARL, ONR
- DoE (SCI, NNSA, OE)
- DHS
- EPA
- HHS
  - AHRQ
  - NIH
  - ONC
- NARA
- NASA
- NRO
- NSA
- NSF (CISE, OCI)

NITRD PCAs
(program component areas)

- Cyber Security and Information Assurance
- High Confidence Software and Systems
- High-End Computing
- Human Computer Interaction and Info Mgmt
- Large Scale Networking
- Social, Economic, and Workforce Implications
- Software Design and Productivity

NITRD SSGs
(senior steering groups)

- Cybersecurity
- Health IT R&D
- Wireless Spectrum Efficiency
- CyberPhysical Systems
- *Big Data*

NITRD's
Big Data Initiative

- Core Technologies
- Domain Research Data
- Challenges/Competitions
- Workforce Development

Core Technology Research

- Solicitation: Core Technologies and Technologies for Advancing Big Data Science & Engineering (BIGDATA) NSF 12-499
- 9 NSF Directorates and 7 NIH Institutes
- http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf

Core Tech I: Collection, Storage and Management of Big Data

- Data representation, storage and retrieval
- New parallel data architectures, including clouds
- Data management policies, including privacy and access
- Communication and storage devices with extreme capabilities
- Sustainable economic models for access and preservation

Core Tech II: Data Analytics

- Computational, mathematical, statistical and algorithmic techniques for modeling high dimensional data
- Learning, inference, prediction and knowledge discovery for large volumes of dynamic data sets
- Data mining to enable automated hypothesis generation, event correlation and anomaly detection
- Information infusion of multiple data sources

Core Tech III: Data Sharing and Collaboration

- Tools for distant data sharing, real time visualization and software reuse of complex data sets
- Cross disciplinary model, information and knowledge sharing
- Remote operation and real time access to distant data sources and instruments

Domain Research Data

- Identify current projects that could benefit from cross-agency collaboration.
- Propose new cross-agency projects.
- Examples: Earth Cube, XCEDE, NIH Electronic health record Collaboratory, Climate Change and Health

Challenges/Competitions

- Series of Challenges: Ideation -> New Tool
- Workshops to decide the parameters
- NASA Center of Excellence for Collaborative Innovation - > Plan (April 2012)
- BD SSG Approval and Funding

Workforce Development

- 20 projects across 7 agencies that may be suitable for adapting to Big Data, (Grants, Fellowships, Summer Internships, Scholarships, etc.)
- Undergraduate, Doc, Post-Doc, Mid-Career
- Building a Data Science Community: Meetings at annual conferences, professional associations etc.

Why now for Big Data?

- Disk storage cost has gone from 25 cents per *byte* (IBM 305 Ramac in 1956) to 25 cents per *gigabyte* today. 25 cents per terabyte soon?

- Sensors: remote sensing, video surveillance, environmental sensing, scientific instruments, etc

- The Internet: Five billion gigabytes and counting (estimated by Eric Schmidt)

Big Data processing

- Phase 1 : Ingest
- Phase 2 : Store
- Phase 3 : Analyze (three options)
- Phase 4 : Visualize
- Phase 5 : Insight/Decide

Analyze phase options

- Distributed Memory Architecture for identify-needle-in-haystack applications; e.g., Hadoop
- Shared-Memory Non-Coherent Architecture
- Shared-Memory Coherent Architecture for connections-between-hay-in-stack analysis; e.g., DNA de novo assembly

The CAP Theorem

- Consistency, Accessibility, Partitionability
- Traditional Databases can have all three
- Big Date can have two out of three!

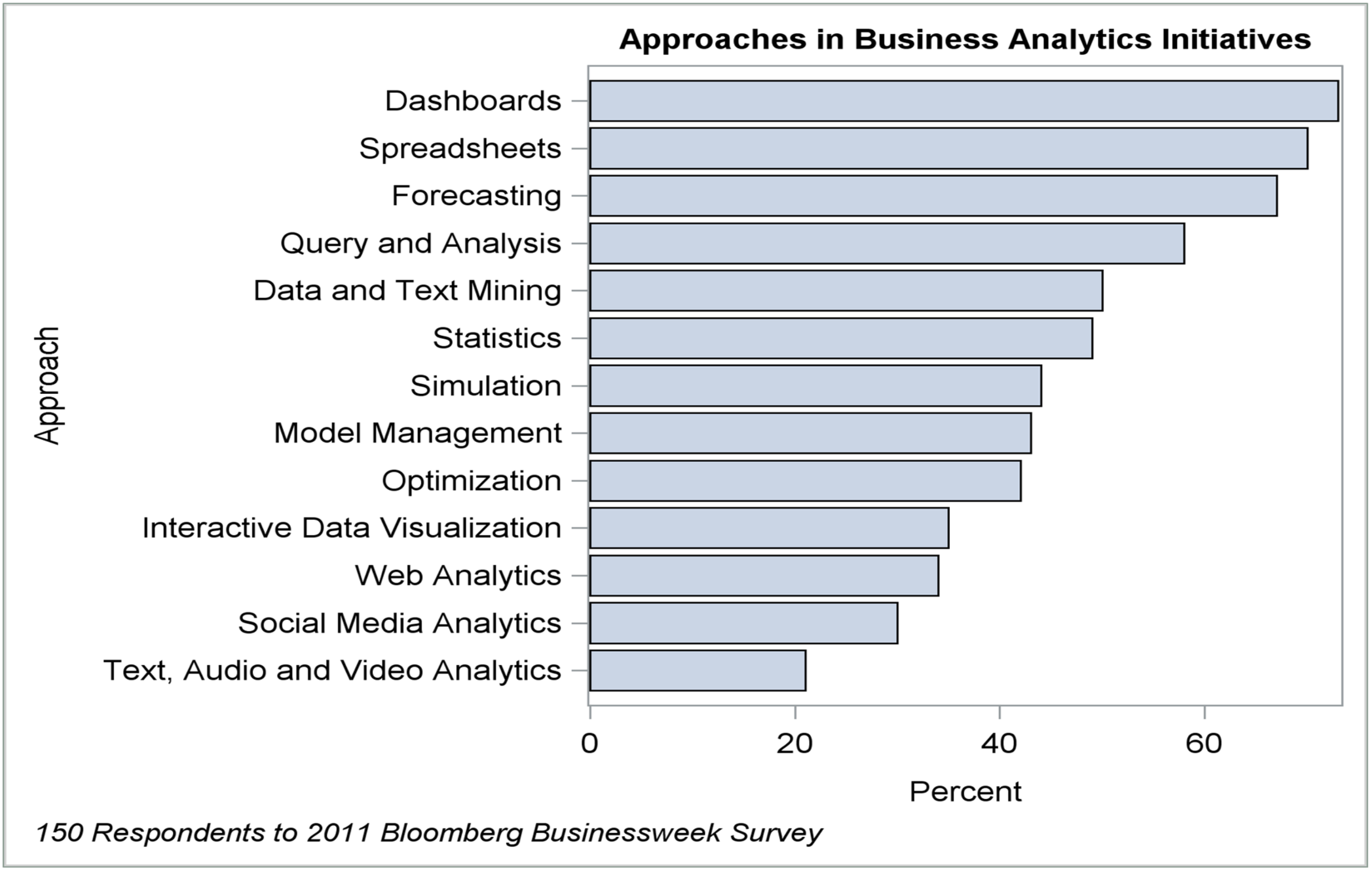
Big Data in Business and Government

- Business analytics
- Tools of business analytics
- Trends

Business Analytics

- The use of statistical analysis, data mining, forecasting, and optimization to make critical decisions and add value based on customer and operational data.
- Critical problems are often characterized by massive amounts of data and the need for rapid decisions and high performance computing
- Eg, modeling customer lifetime value in banks
- reducing adverse events in health care
- managing customer relationships in hospitality industry

Tools of Business Analytics



Trend 1: Bigger Data

- Volume, velocity, variety of big data keep increasing!
- Storage and compute capacity often less than needed for timely decision
- Basis for web-based businesses (Google, Facebook, …)
- Business sectors are leading the way in exploiting data about customers and transactions.
- Prevalent in pharmaceutical, retail, and financial sectors

Trend 2: Unstructured Data

- 70% of enterprise data is unstructured: images, email, documents

- Text analytics: linguistics, natural language processing, statistics

- Content categorization, sentiment analysis

- Text mining: statistical learning applied to a collection of documents

- Examples: discovery of adverse drug effects from patient notes; identification of fraudulent insurance claims; sentiment analysis based on Facebook posts; early warning from warranty and call center data

Trend 3: Distributed Data

- Terabyte-sized data are spread across multiple computers, and are increasingly held in distributed data stores that are amenable to parallel processing.
- Extraction into traditional computing environments chokes on data movement
- Challenge is to co-locate analysis with data
- Apache Hadoop is now widely used for Big Data applications

Trend 4: Distributed Computing

- Scaling our computational tools, algorithms and thinking: how do we apply parallel programming methods for processing data distributed on thousands of computers

- How do we acquire specialized programming skills?

- Where are the data located? What proportion of the work can be done in parallel by nodes?

- Do we understand the mechanisms that generate Big Data?

- What are useful models? How do we look further?

Trend 5: Analytical Software

Fraud detection
Credit and operational risk
Credit scoring
Warranty analysis
Customer retention
Markdown optimization

Big Data in Science

- Analyzing output from supercomputer simulations (eg, climate simulations)

- Analyzing instrument (sensor) output

- Creating databases to support wide collaboration (eg, human genome project)

- Creating *knowledge bases* from textual information (eg, Semantic Medline)

Scientific Data Analysis Today

- Scientific data is doubling every year, reaching PBs (CERN is at 22PB today, 10K genomes ~5PB)
- Data will never again be at a single location
- Architectures increasingly CPU-heavy, IO-poor
- Scientists need special features (arrays, GPUs)
- Most data analysis done on midsize BeoWulf clusters. Universities hitting the “power wall”
- Soon we cannot even store the incoming data stream
- Not scalable, not maintainable…

LHC tames big data?

- Produces a petabyte of info *per second*
- Saves for processing a petabyte per month
- This factor of 10**9 reduction in data is possible because i) the LHC is "smart" and ii) there is a "good model" of the data

Data in HPC Simulations

- HPC is an instrument in its own right
- Largest simulations approach petabytes--from supernovae to turbulence, biology and brain modeling
- Need public access to the best and latest through interactive numerical laboratories
- Creates new challenges in: how to move the petabytes of data (high speed networking); how to look at it (render on top of the data, drive remotely)
- How to interface (virtual sensors, immersive analysis)
- How to analyze (algorithms, scalable analytics)

Common Analysis Patterns

- Large data aggregates produced, but also need to keep raw data
- Need for parallelism; heavy use of structured data, multi-D arrays
- Requests enormously benefit from indexing (eg. rapidly extract small subsets of large data sets)
- Computations must be close to the data!
- Very few predefined query patterns
- Geospatial/locality based searches everywhere
- Data will never be in one place, and remote joins will not go away
- No need for transactions, but data scrubbing is crucial

Disk Needs Today

- Disk space, disk space, disk space!!!!
- Current problems not on Google scale yet:
- 10-30 TB easy, 100 TB doable, 300 TB hard
- For detailed analysis we need to park data for several months
- Sequential IO bandwidth--if analysis is not sequential for large data set, we cannot do it
- How to move 100TB within a University? 1Gbps --10 days; 10 Gbps--1 day (but need to share backbone); 100 pound box--few hours
- From outside? Dedicated 10Gbps or FedEx

Cloud vs Cloud

- Economy of scale is clear
- Commercial clouds are too expensive for Big Data--smaller private clouds with special features are emerging
- May become regional gateways to larger-scale centers
- The “Long Tail” of a huge number of small data sets (the integral of the “long tail” is big)
- Facebook brings many small, seemingly unrelated data to a single cloud and new value emerges. What is the science equivalent?

Science and Big Data

- Science is increasingly driven by data (large and small)
- Large data sets are here, COTS solutions are not
- From hypothesis-driven to data-driven science
- We need new instruments: “microscopes” and “telescopes” for data
- There is also a problem on the “long tail”
- Similar problems present in business and society
- Data changes not only science, but society
- A new, Fourth Paradigm of Science is emerging…

What the future may hold

- Data intensive science appears to be revolutionary science

- Data analytics and big data are major opportunities for business and government

- Big Data will continue to provide the basis for new services for citizens perhaps as important as the Web, Google and Facebook