Interoperability: the Key to Big Data Value

George O. Strawn, Director

NAS Board on Research Data and Information

Outline

- Preliminaries
- Big data and all data
- A previous interoperability problem and its solution
- FAIR data: interoperability++
- Analytics
- In conclusion, a caution

Preliminaries

A little about me

- Primarily a computer nerd from 1961 to 1991
- Primarily a networking nerd from 1991 to 2001
- Primarily a data nerd since 2001
- "A CIO should spend at least as much time on the I as on the T"

NIST Big Data Reference Architecture

1. Identify the high-level Big Data reference architecture (NBD-RA) key components

2. Define general interfaces between the NBD-RA components

3. Validate the NBD-RA by building Big Data general applications through the general interfaces.

Workshop Objectives

(a) NBDIF Version 2 Overview – consensus approach to establish an **interoperable** ecosystem for Big Data applications and analytics

(b) Big Data application challenges – how a standard ecosystem may overcome Big Data applications development challenges

(c) Big Data analytics challenges – how a standard ecosystem may provide reusable, deployable, and operational analytics tools

(d) Explore Exascale Big Data Analytics and Systems.

The Nist BD **Interoperability** Framework

Volume 1: Big Data Definitions
Volume 2: Big Data Taxonomies
Volume 3: Big Data Use Cases and Requirements
Volume 4: Big Data Security and Privacy
Volume 5: Big Data Reference Architectures White Paper Survey
Volume 6: Big Data Reference Architecture
Volume 7: Big Data Standards Roadmap
Volume 8: Big Data and Reference Arch Interface
Volume 9: Big Data Adoption and Modernization

Federal Big Data R&D Strategic Plan(1)

Strategy 1. Leverage emerging big data foundations, techniques and technologies

Strategy 2. Explore and understand trustworthiness of data

Strategy 3. Build and enhance research cyberinfrastructure

Strategy 4. Develop policies that promote the sharing and management of data

Strategy 5. Explore and understand big data privacy, security and ethics

Strategy 6. Improve big data training and education

Strategy 7. Create and enhance connections in the national big data innovation ecosystem

Big Data and *Data-intensive Science*

- The first scientific revolution, which began in the 17th century, was enabled by *printed information*
- The scientific revolution of the 21st century is enabled by *computed information*
- Interoperability enables *meaningful* data sharing, and science projects are then able to build on each other
- Interoperability is enabled by *metadata*

1st Gen Data Intensive Science

- Human Genome Project
- Proteomics Databases
- US Brain Project
- EU Brain Simulation Project
- Materials Genome Initiative
- Virtual Observatory

Big Data definitions

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

- Big Data consists of extensive datasets primarily in the characteristics of volume, variety, velocity, and/or *variability* that require a scalable architecture for efficient storage, manipulation, and analysis.

Why now for Big Data?

- Big Data is a child of the Internet, which has connected islands of information into a continent of (non-interoperable!) information
- "Moore's laws" for disks, sensors, networks and CPUs
- EG: sensors: cheap remote sensing, video surveillance, environmental sensing. IoT coming on strong!
- EG: disk storage cost has gone from a dollar per *byte* to a dollar per *25 gigabytes* today. A dollar per terabyte soon? Now cheaper to save than throw away? DNA storage on the way?(2)

Big Volume Data requires *big computing*

- *Supercomputers* are thousands to millions of tightly coupled computing elements
- *Server farms* are thousands to millions of loosely coupled computing elements
- All large scale computing involves parallel processing: supercomputers for fine-grain processing, server farms for coarse-grain

Big Volume Data requires new *data architectures*

- Relational database architecture doesn't scale
- NoSQL databases limit functionality and do scale
- Eg, Hadoop, BigTable, Document- and Column-oriented databases, Graph databases

Big Velocity Data

- Success with OLTP (*parallel* online transaction processing) such as google search and amazon ordering, but sensor input (Internet of Things) poses a bigger challenge
- Need "smart sensors" like the LHC, which generates a petabyte of data per second but "only" saves a petabyte per month (we need an Internet of *Smart* Things, not just an IoT)
- And/or micro clouds

Big Variety Data

- The *interoperability of heterogeneous data* is perhaps *the* major big data research challenge
- Eg, the "long tail" of the many small science data sets will require extensive metadata to enable interoperability
- Eg, computable literature such as *Semantic Medline* (Medline+ semantic metadata) portends a new mode of discovery from scientific text

Only big data?

All data has size (volume), rate of acquisition (velocity) or complexity (variety) is beyond the ability of ***novel and desired*** software tools to capture, manage, and/or process within a tolerable elapsed time.

A previous interoperability problem and its solution

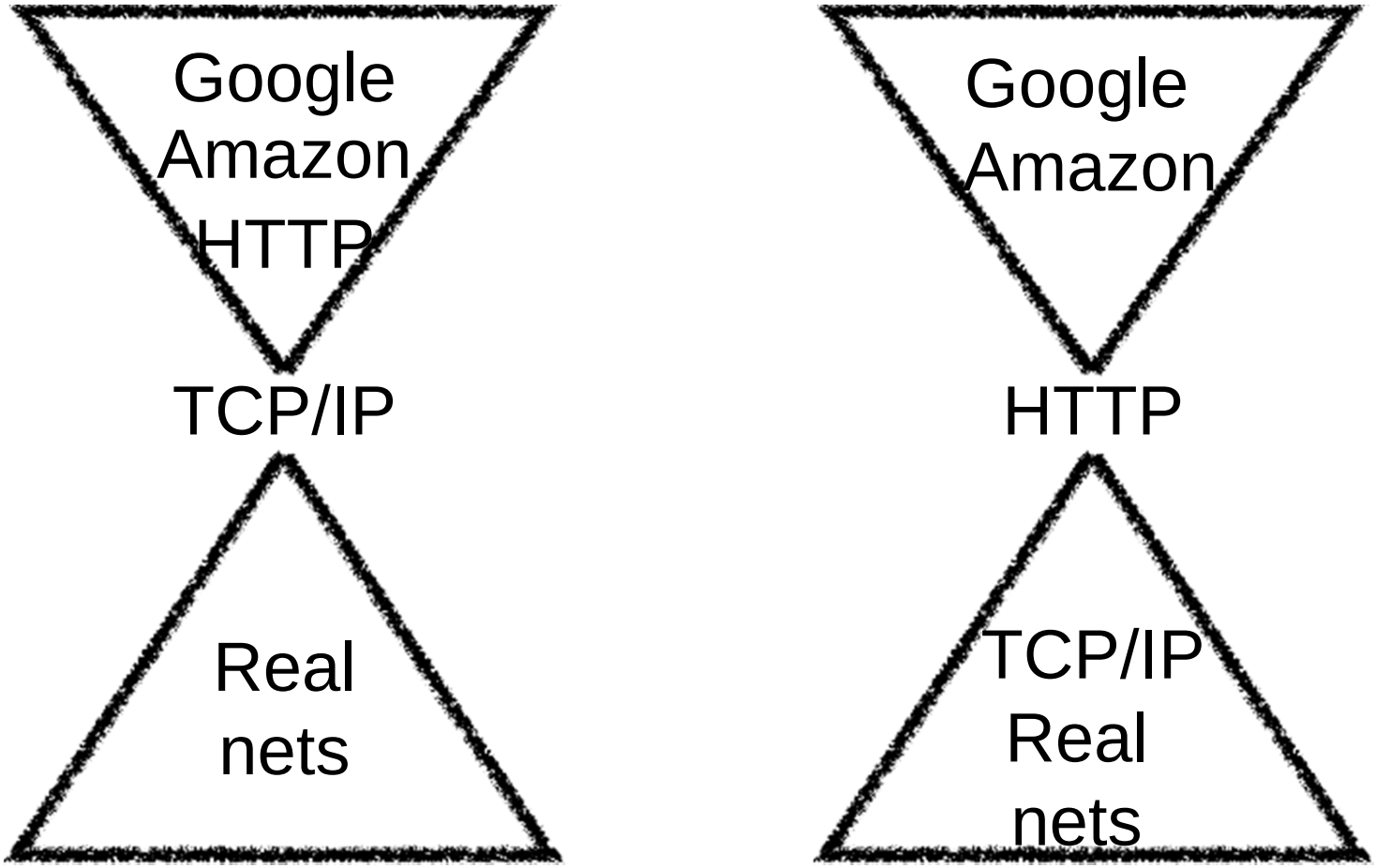
The Internet and Interoperability

- The Internet solved "the interoperability of heterogeneous networks" problem with a new layer of abstraction
- The TCP/IP protocols implemented a *virtual* network (at layer 3) that stitched together real networks (at layer 2)
- The ARPAnet and the follow-on NSFnet were *greenfields* projects (i.e., there was no industry)
- DARPA built the ARPAnet utilizing "rough consensus and running code" (hard to duplicate for big data)

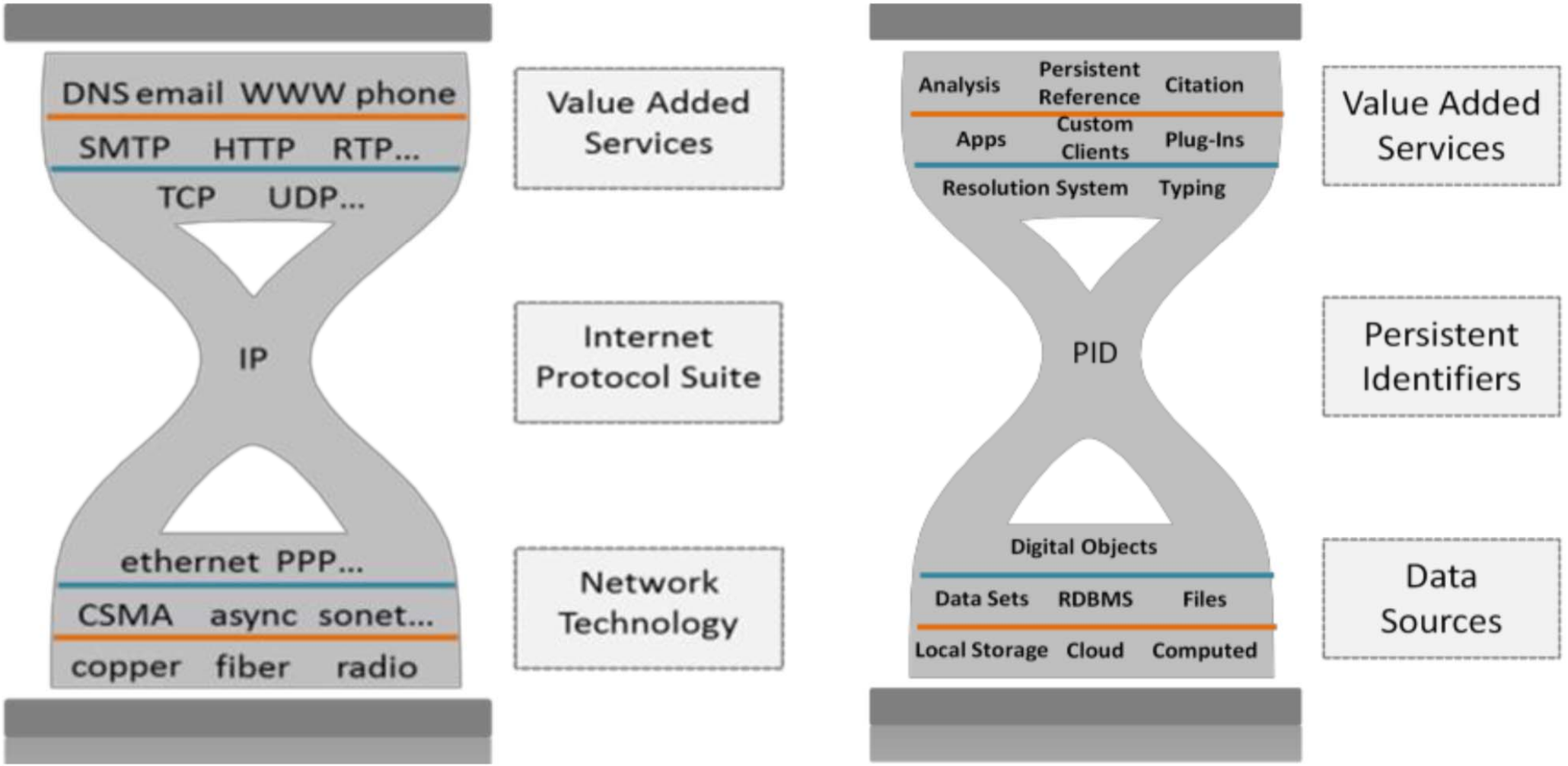
Does the Internet solution apply to Data?

- The "hourglass" view of the Internet shows TCP/IP at the narrow center with networks in the lower part and applications such as the Web in the upper
- The Web can be viewed this way with HTTP at the narrow center with TCP/IP in the lower part and applications such as Google search, Amazon sales and Facebook pages in the upper
- Might there be a general data protocol that would provide such an hourglass view of data?

Hourglass Views



Globally Interoperable IoT Identification and Data Processing(3)



FAIR data(4)

- Findable: keyword search now, semantic search soon
- Accessible: usually free, with exceptions for privacy, intellectual property, etc
- Interoperable: the key to value
- Reusable: minimal legal prohibitions

Analytics

Big data analytics is the process of examining large and varied **data** sets -- i.e., **big data** -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions. (http://whatis.techtarget.com)

1. Vision of near-/mid-/long-term Big Data analytics

2. How standards ecosystem can help in Big Data analytics in terms of deploy reusable analytics?

Big Data Capabilities

	Small era	Big era	Next generation
Goals	Answer a specific question, establish correlations	Flexible goals, possibly ill-posed questions, probabilistic prediction	Knowledge assimilation and reasoning, understanding causality
Location	One place	Highly distributed	Amorphous
Data structure & Content	Highly structured	Absorbs unstructured data from many sources	Differing in uncertainty and quality; combined with certified knowledge
Data preparation	By user or small group	Many sources, many people, possibly unconnected to users	Captured raw, ad hoc; combined w/ certified, standardized data
Longevity	Limited	Perpetual	Perpetual and reuseable
Reproducibility	Repeatable	Not necessarily repeatable	New data, information, knowledge continuously alters results
Analysis	All data analyzed together, all at once	Analyzed in incremental steps, distributed	Continuous processing within a context

Adapted from Berman, J.K.(2013*) Principles of Big Data,* New York; Elsevier

50 Years of Data Science(6)

- Data exploration and preparation
- Data representation and transformation
- Computing with data
- Data modeling
- Data visualization and presentation
- Science about data science

Data Mining

- The computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database

- It involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating

- Data mining is the analysis step of the "knowledge discovery in databases" process (eg, BD2K)

Machine Learning(7)

- Symbolists (inverse deduction)
- Connectionists (neural nets, deep learning)
- Evolutionaries (genetic programming)
- Baysians (dealing with uncertainty)
- Analogizers (support vector machine)

3. What does it take to get to Exascale computing and systems?

4. What Exascale computing and systems would bring to Big Data analytics?

5. How standards help to address Exascale computing and systems.

National Strategic Computing Initiative Strategic Plan(8)

- Objective 1. Achieve capable exascale computing by the mid-2020s
- Objective 2. Develop a coherent platform for modeling, simulation *and data analytics*
- Objective 3. Pursue R&D to move beyond CMOS and explore alternative paradigms
- Objective 4. Develop and adopt new approaches, technologies and software architectures to support application development, reusability, trustworthiness, sustainability and workforce development
- Objective 5. Broaden public-private collaboration

As exascale supersedes petascale

- Friendly tools and analytics will be developed for petascale
- But for exascale, the traditional maxim will hold: "Supercomputers aren't supposed to be friendly. They're supposed to be fast."

In conclusion, a caution

Dataism(9)

- *Dataism* declares that the universe consists of data flows and the value of any phenomenon or entity is determined by its contribution to data processing
- Dataism is the first movement since 1789 that has created a genuinely novel value: freedom of information
- Dataists believe in the invisible hand of the data flow
- Elections, political parties and parliaments might become obsolete because they can't process information efficiently enough
- By the time the cumbersome government bureaucracy makes up its mind about cyber regulation, the Internet will have morphed 10 times
- What will happen to society, politics and daily life when highly intelligent algorithms know us better than we know ourselves?

Know us too well?

The danger facing us is not Orwell, but Huxley. The combo of data collection and machine learning is too good at catering to human nature, seducing us and appealing to our worst instincts. We have to put controls on it. The algorithms are amoral; to make them behave morally will require active intervention.(10)

References

1. https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf

2. https://www.extremetech.com/extreme/218241-at-up-to-455-exabytes-on-a-single-gram-dna-storage-could-create-mankinds-permanent-record

3. https://www.rd-alliance.org/rda-eu-and-iot-forum-workshop-globally-interoperable-iot-identification-and-data-processing-6-june
4. The FAIR Guiding Principles for scientific data management and stewardship (https://www.nature.com/articles/sdata201618)

5. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information* by Jules J. Berman (2013)

6. https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf

7. *The Master Algorithm* by Pedro Domingos (2015) and http://www.dataversity.net/pedro-domingos-on-five-machine-learning-tribes/

8. https://www.whitehouse.gov/sites/whitehouse.gov/files/images/NSCI%20Strategic%20Plan.pdf

9. Homo Deus by Yuval Harari

10. http://www.nakedcapitalism.com/2017/05/notes-from-an-emergency.html?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+NakedCapitalism+%28naked+capitalism%29