Big Data in Context

George O. Strawn

Director, National Coordination Office for
Interagency Networking and IT R&D

Caveat auditor

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

Outline

- The Entrepreneurial State
- The USG and Big Data
- Big Data in Bioinformatics

Entrepreneurial State by Mazzacato

"Since its founding, the United States has always been torn between two traditions, the activist policies of Hamilton and Jefferson's maxim that 'the government that governs least, governs best'. With time and usual American pragmatism, this rivalry has been resolved by putting the Jeffersonian's in charge of the rhetoric and the Hamiltonians in charge of policy."

USG IT Entrepreneurialism

1840. Support for Morse and the telegraph (risk)

1890. Support for Hollerith and punch card data processing (use)

1940. Support for Mauchly and the electronic digital computer (defense)

1965. Support for Roberts, Kahn, Cerf and the Internet

Continuing Innovation in Information Technology

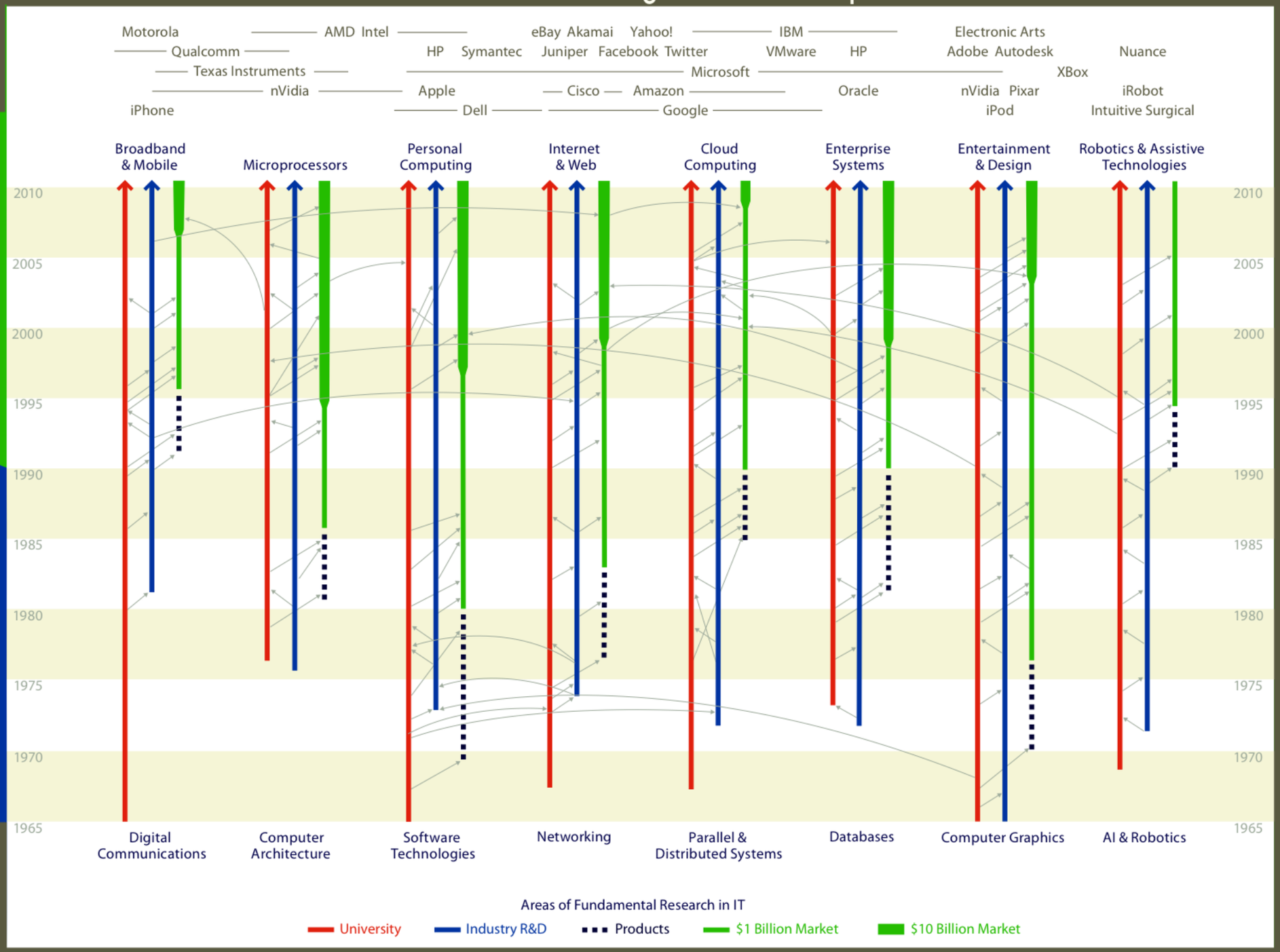
Fundamental research in IT underpins the creation of billion-dollar-plus IT market segments and a vital U.S. IT industry through a complex partnership between universities, industry, and government.

The first version of this figure was published in the 1995 report *Evolving the High Performance Computing and Communications Initiative to Support the Nation's Information Infrastructure*. The original figure, which was updated in 2002 and 2003, dispelled the assumption that the commercially successful IT industry is self-sufficient. It underscored the extent to which industry instead builds on government-funded university research—sometimes through long incubation periods of years and even decades.

As illustrated in this figure from the 2012 report *Continuing Innovation in Information Technology*, computing research and its impacts have since continued to evolve and blossom. The figure illustrates how fundamental research in IT, conducted in industry and universities, has led to the introduction of entirely new product categories that ultimately became billion-dollar industries. It reflects a complex research environment in which concurrent advances in multiple subfields have been mutually reinforcing, stimulating and enabling one another and leading to vibrant, innovative industries exemplified by top-performing U.S. firms. Such research often starts as a search for fundamental knowledge but time and again produces practical technologies that enable significant economic impact.

The gray lines illustrate the rich interplay between academic research, industry research, and products and indicate the cross-fertilization resulting from multi-directional flows of ideas, technologies, and people.

IT Sectors with Large Economic Impact



NITRD involvement

1995. Next Generation Internet

2005. Computer Security and Information Assurance

2010. Big Data

2012. Cyberphysical Systems

2014. Privacy

NITRD Networking and

ITRD IT R&D

CIC computing, info and comm

HPCC and communication

HPC high-performance computing

NITRD and the NCO

- NITRD: an interagency program to enhance coordination and collaboration of the IT R&D that is performed and supported by Federal agencies

- National Coordination Office: provides support for the NITRD Program, reports to OSTP, and interfaces for NITRD with OMB, GAO, Congress, etc.

NITRD Member Agencies

Department of Commerce (2)

Department of Defense (5)

Department of Energy (3)

Department of Health and Human Services (3)

Department of Homeland Security

Environmental Protection Agency

National Archives and Records Administration

National Aeronautics and Space Agency

National Reconnaissance Office

National Science Foundation

National Security Agency

NITRD PCAs
(program component areas)

- Cyber Security and Information Assurance
- High-End Computing (R&D and I&A)
- High Confidence Software and Systems
- Human Computer Interaction and Info Mgmt
- Large Scale Networking
- Social, Economic, and Workforce Implications
- Software Design and Productivity

NITRD SSGs
(senior steering groups)

- *Big Data* -> HCI&IM
- CPS -> HCSS
- Cybersecurity, Privacy -> CSIA
- Health IT R&D -> (cross-cutting)
- Wireless Spectrum R&D -> LSN

FY 2012 Budget Estimates

	HECia	HECrd	CSIA	HCIim	LSN	HCSS	SDP	SEW	Total
NSF	250	103	98	292	122	85	78	110	1,138
DoD	211	49	145	111	112	36	30		694
NIH	222	18		215	12	10	54	22	553
DOE	317	92	34		74	4	16	6	543
DARPA		75	223	138	53				489
NIST	14	5	47	15	8	6	4	1	100
NASA	61			14	1	18	9		103
DHS			43		1		3		47
AHRQ				25	1				26
NOAA	19				2		1		22
DOEnnsa	9	5						4	18
EPA	3			3					6
NARA				1					1
Total	1,107	347	590	814	385	158	196	143	3,739

USG and Data

- *Open Access* to usg data becomes the default (http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf)

- *Public Access* to Federally funded science results (journal articles *and* science data) required of all agencies funding more than $100M per year (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)

- Digital Accountability and Transparency Act of 2014 (S. 994; Pub.L. 113–101) is a law that aims to make information on federal expenditures more easily accessible and transparent.

Big Data

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.

Why now for Big Data?

- 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: disk storage cost has gone from a dollar per *byte* to less than a dollar per *ten gigabytes* today. A dollar per terabyte soon? Now cheaper to save than throw away?

- EG: sensors: cheap remote sensing, video surveillance, environmental sensing, scientific instruments (not necessarily cheap), etc

Volume: big data requires big computing

- These days, supercomputers aren't actually bigger: they're broader (thousands of *tightly* coupled cpu's)

- Server farms are *loosely* coupled cpu's (thousands of servers)

- Big volume data resides on supercomputers or server farms (or at least on clusters)

Volume: big data requires new database architectures

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

Velocity: fast big 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 (take the processing to the data if a good model exists)

- And/or micro clouds?

Variety: diverse big data

- The *interoperability of heterogeneous data* is a, perhaps *the*, major big data challenge

- The "long tail" of many small science data sets requires metadata to enable interoperability

- *Semantic Medline* (the creation and use of semantic metadata with Medline) portends the a new mode of discovery from scientific text

NITRD's
Big Data Initiative

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

Core Technologies

- Collection, Storage and Management, Visualization of Big Data
- Data Analytics
- Data Sharing and Collaboration

Domain Research Data

- Astronomy, Virtual Observatory
- data.gov
- Earth Observation Systems
- Genomics, Proteomics
- Materials Genome
- Nano S&T, Nanohub
- NSF projects such as DataOne, DataNet
- Particle Physics, LHC

Challenges/Competitions

- Engage a broader public

Workforce Development

- Data Science, Data degrees

- *Data scientist*: a computing person who knows more statistics than his/her colleagues or as tactician who know more about computing than his/her colleagues?

Next NITRD steps

- BRDI Big Data Day --10/23/14 https://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data_(BD_SSG)#title

- RFI already published (https://www.nitrd.gov/bigdata/rfi/02102014.aspx)

- Public workshop planned for early 2015

- An inter-agency strategic plan for Big Data R&D by mid-2015

Bioinformatics I

- Science quantified the motion of *matter* in the 1600s, the varieties of *energy* in the 1800s, and the nature of *information* in the 1900s

- Biology is *the* information-rich science (bio+cs in the future like physics+math in the past)

- The digital simulation of life forms has begun (and is under consideration for homo sapiens)

- The digital simulation of the brain in Europe and the BRAIN Interagency Initiative in the US

Bioinformatics II

- Open access to and interoperability of biomed research literature (Varmus + Semantic Medline)
- Open access to and interoperability of biomed research data (BD2K)
- Genomics and proteomics databases

Informatics, writ large?

- Information is physical, not ephemeral
- "Its from bits"?
- A replay of the steam engine leading to thermodynamics?
- A science of information to underly the technology of information?

What the IT future may hold

- Data intensive science appears to be revolutionary science, with bioinformatics leading the way

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

- Big Data will also be the basis of new services for people, perhaps as significant as the Web, Google and Facebook and as challenging for privacy