Big Data is Coming for You

W. Christopher Lenhardt
RENCI
DAARWG, Chair

renci

RESEARCH \ ENGAGEMENT \ INNOVATION

Outline

- A few words about RENCI
- Introduction: On the Nature of BIG
- Big Challenges
- Big Science Questions
- Big Data
- Other Big Trends
- Implications?

RENCI's Mission

- **Be a leader in cyber-infrastructure (CI) research and development**
- **Be an essential CI partner for:**
  - Triangle university research teams
  - Research Triangle area industries
  - State of NC and federal agencies
  - **IN ORDER TO address complicated multidisciplinary problems and research.**
- **Data is central to all we do. Underlying theme: *Data to Decisions***

renci

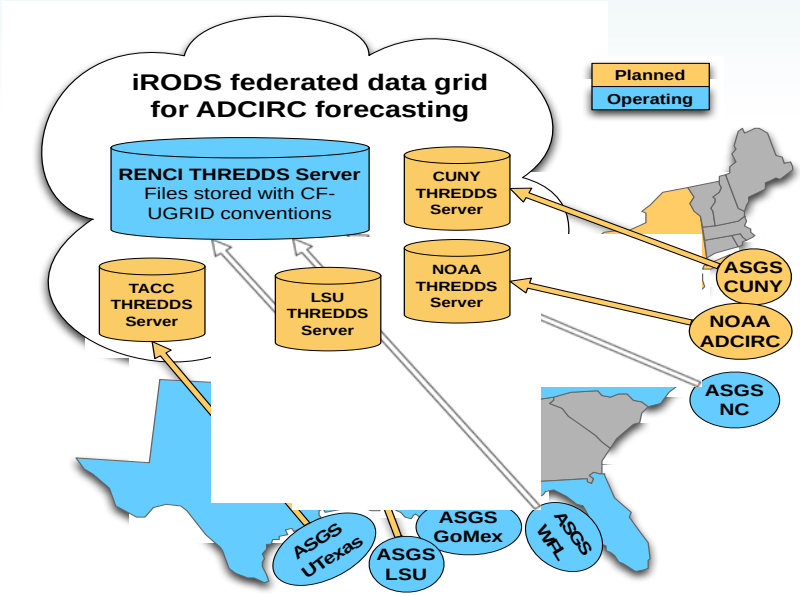
Collaboration: Key to Advance Data Science

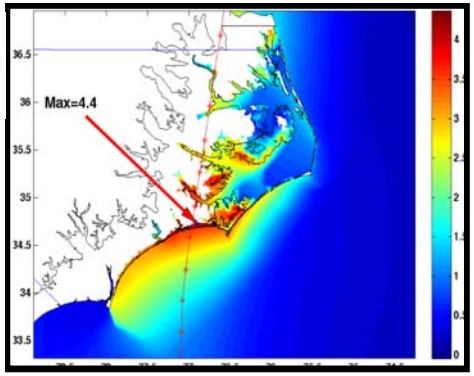
- **Almost all large-scale data projects:**
  - Include multiple producers of data
  - Data experts to curate and model the data
  - CI experts to store, manage, analyze and serve the data
  - Many end users who will use the information from the data for action

RENCI Big Data Snapshot

- **Physical and Human:** Compute cluster (**~30** teraflops peak performance); **2.7** petabytes spinning disk; experimental network connecting UNC, Duke and NCSU; experts in software, bioinformatics, HPC, geovisualization and domain sciences
- **Funded Data Projects**: **9** funded data projects, total funding of **$10.9M**
- **Partnership with UNC School of Med:** Data tech for genomics/medical research: HPC cluster, large disks for local computations, relational DB management system, software frameworks (iRODS, Secure Research Workspace, Hadoop, Pegasus)
- **Data Science Leadership**: **iRODS** is a leading data management framework for research; **E-iRODS** team at RENCI supports enterprise-quality iRODS releases; **National Consortium for Data Science** launched to put UNC and NC in Big Data leadership position.
- **Cyberinfrastructure Leadership**: ***ExoGENI***: building next-gen Internet for research in Big Data era; ***CI-BER***: CI for billions of EMRs; ***Secure Research Workspace:*** secure data handling for research data; ***Water Science Software Institute*:** transforming software needed for Big Data analysis; ***REACH-NC:*** searchable public database of NC expertise (LinkedIn for researchers)

Storm Surge Forecasting (ADCIRC)





Sandy (2012) and Irene (2011) flooding forecasts used by

- **National Hurricane Center in Miami**
- **US Coast Guard Atlantic Command**
- **Regional National Weather Service Offices**
- **State and local emergency managers**

- System uses **NSF/NARA** funded iRODS**, NOAA NOS** gauge data , **USGS** data, **DHS/FEMA** collected high-water mark, meteorological forecasts from **NOAA's NCEP** and **NHC**
- Very large pre-existing datasets; provides early guidance information, available about 10 minutes after official NHC forecast storm advisory
- **DHS-funded research activity** through the DHS Coastal Hazards Center of Excellence at the University of North Carolina at Chapel Hill
- **Winner, DHS Science & Technology Impact Award, 2012**

renci

RENCI helping to Host ESIP Summer Meeting

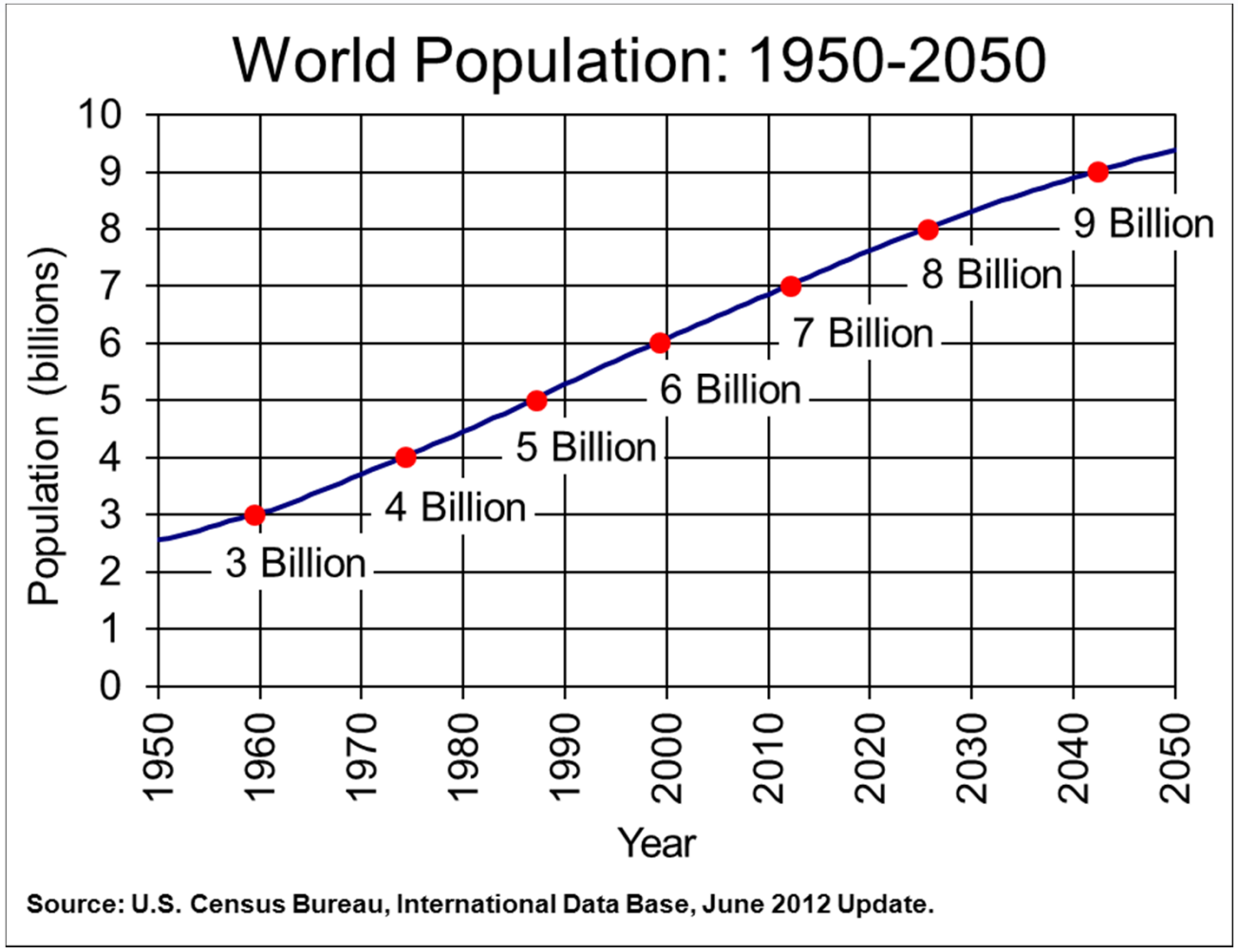
ESIP SUMMER MEETING – JULY 9–12, 2013 CHAPEL HILL, NC, HOSTED BY RENCI/UNC

We are now planning for the Summer Meeting. If you are interested in contributing to the planning, join the Visioneers telecons or add your ideas to the Summer Meeting wiki page. Thanks!

Date: July 9-12, 2013 in Chapel Hill, North Carolina

On the Nature of *BIG…*

World Population



http://www.census.gov/population/international/data/worldpop/images/worldpop.png

renci

May 9, 2013

SCRIPPS INSTITUTION OF OCEANOGRAPHY

ABOUT | NEWS | RESEARCH | EDUCATION | SHIPS | RESOURCES AND FACILITIES | GIVING | PEOPLE

/ Scripps News / Carbon Dioxide at Mauna Loa Observatory reaches new milestone: Tops 400 ppm

Carbon Dioxide at Mauna Loa Observatory reaches new milestone: Tops 400 ppm

Scripps, NOAA measurements cross threshold in same 24-hour period

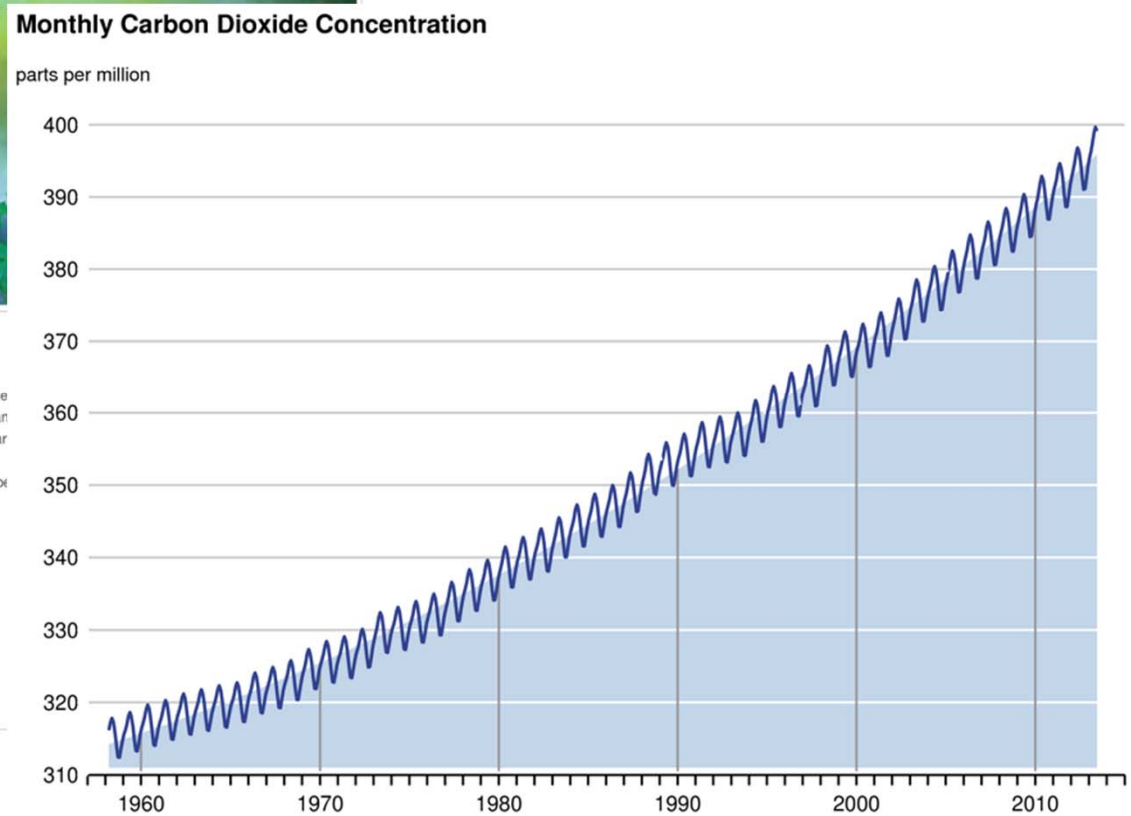
May 10, 2013

Media Contact: Robert Monroe | Phone: 858-534-3624 | Email: scrippsnews@ucsd.edu

Scripps Institution of Oceanography / University of California, San Diego

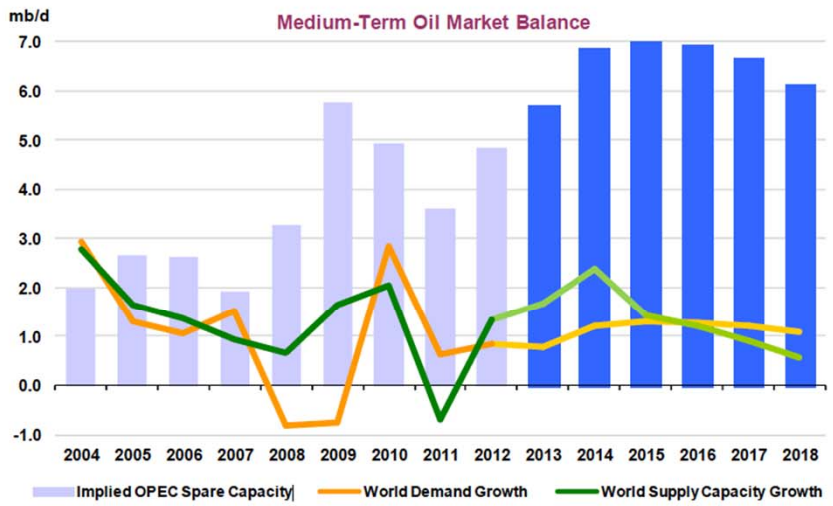
On May 9, 2013 the daily mean concentration of carbon dioxide in the atmosphe
million (ppm) for the first time since measurements began in 1958. This marks an
continuous carbon dioxide (CO_2) measurement station in the world, is the primar

NOAA instruments recorded a reading of 400.03 ppm for May 9. Instruments ope
Diego recorded a reading of 400.08 ppm for the same 24-hour period.



May 14, 2013

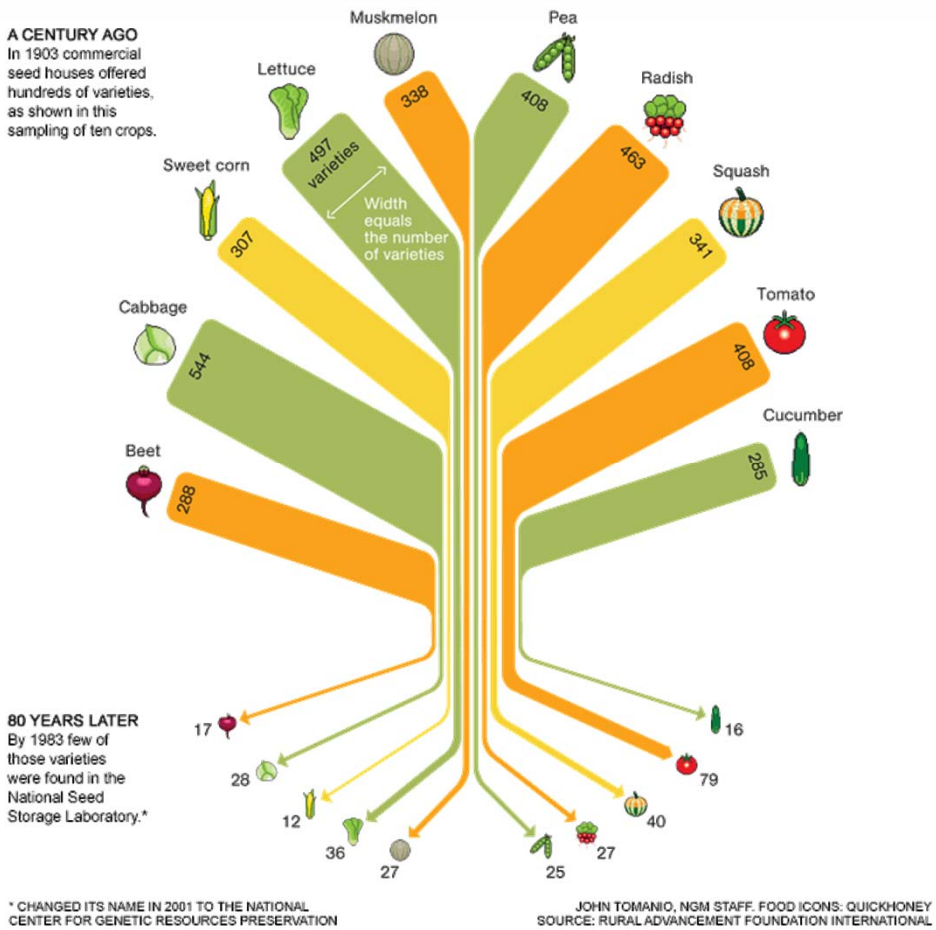
A pivotal time in oil markets

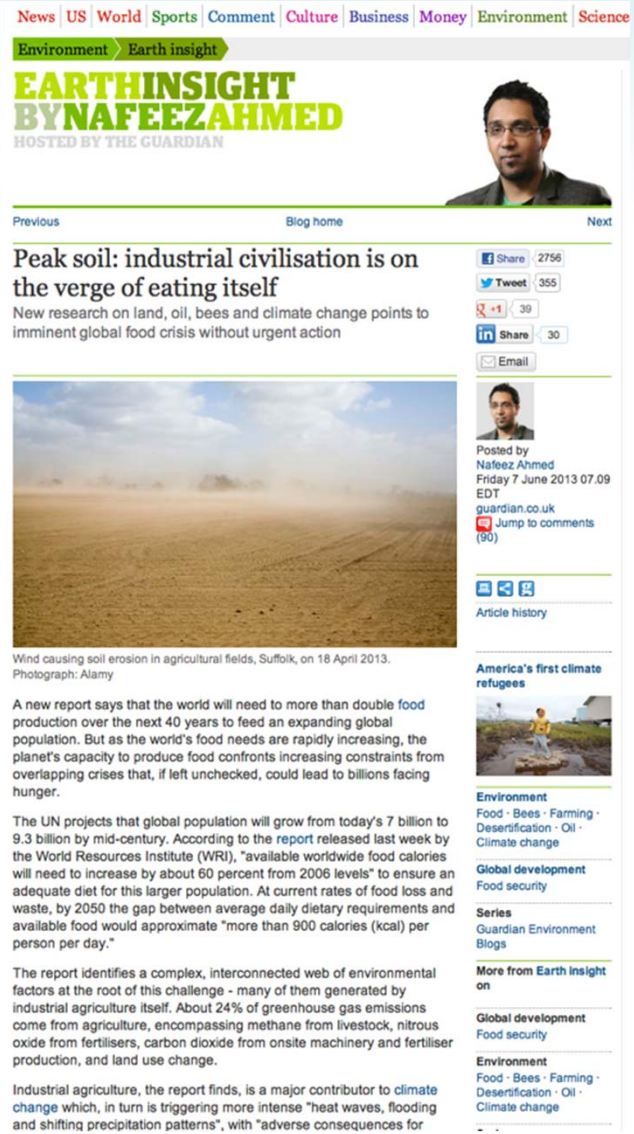


IEA Medium-Term Oil Market Report

http://www.iea.org/newsroomandevents/speeches/SlidePresentationMTOMRLaunch.pdf

Peak Soil?





News | US | World | Sports | Comment | Culture | Business | Money | Environment | Science

Environment > Earth insight

EARTHINSIGHT BYNAFEEZAHMED

HOSTED BY THE GUARDIAN

Previous | Blog home | Next

Peak soil: industrial civilisation is on the verge of eating itself

New research on land, oil, bees and climate change points to imminent global food crisis without urgent action

Wind causing soil erosion in agricultural fields, Suffolk, on 18 April 2013. Photograph: Alamy

A new report says that the world will need to more than double food production over the next 40 years to feed an expanding global population. But as the world's food needs are rapidly increasing, the planet's capacity to produce food confronts increasing constraints from overlapping crises that, if left unchecked, could lead to billions facing hunger.

The UN projects that global population will grow from today's 7 billion to 9.3 billion by mid-century. According to the report released last week by the World Resources Institute (WRI), "available worldwide food calories will need to increase by about 60 percent from 2006 levels" to ensure an adequate diet for this larger population. At current rates of food loss and waste, by 2050 the gap between average daily dietary requirements and available food would approximate "more than 900 calories (kcal) per person per day."

The report identifies a complex, interconnected web of environmental factors at the root of this challenge - many of them generated by industrial agriculture itself. About 24% of greenhouse gas emissions come from agriculture, encompassing methane from livestock, nitrous oxide from fertilisers, carbon dioxide from onsite machinery and fertiliser production, and land use change.

Industrial agriculture, the report finds, is a major contributor to climate change which, in turn is triggering more intense "heat waves, flooding and shifting precipitation patterns", with "adverse consequences for

Share 2756 | Tweet 355 | +1 39 | Share 30 | Email

Posted by Nafeez Ahmed Friday 7 June 2013 07.09 EDT
guardian.co.uk
Jump to comments (90)

Article history

America's first climate refugees

Environment
Food · Bees · Farming · Desertification · Oil · Climate change

Global development
Food security

Series
Guardian Environment Blogs

More from Earth insight on

Global development
Food security

Environment
Food · Bees · Farming · Desertification · Oil · Climate change

http://rafiusa.org/issues/reinvigorating-public-plant-and-animal-breeding/

http://www.guardian.co.uk/environment/earth-insight/2013/jun/07/peak-soil-industrial-civilisation-eating-itself

renci

Climate Change

Big Science Questions

- Water – too much, too little
- No topsoil, empty oceans
- What to do about climate change in a 500 ppm world?
- How to feed 9 billion w/o using any more land, water, or energy
- How to foster urban sustainability
- Connections between environmental change and conflict, human migration, economic well-being

Big Data:
From Compute-Centric to Data-Centric

From a Molehill To a Mountain

A timeline of digital data

Everyone knows that the amount of digital data has exploded in recent years. But it isn't always easy to put today's mountain of data in perspective.

To illustrate the data boom, here's a timeline with snapshots of some of the largest corporate data collections from each decade since the Univac computer was the hot new thing. —*Matthew Kassel*

2010s | **100 Petabytes**
Facebook Is Facebook's data hoard bigger than Google's today? Some say yes, some say no. According to Facebook, its user content makes up more than 100 petabytes of stored photos and video. And analyzing that data generates about 500 terabytes of new information every day—more than 2½ times the size of a '90s Wal-Mart data cache.

2000s | **25 Petabytes**
Google The explosion of the Internet in the '90s set the stage for Web-based companies to emerge in the following decade as the global leaders in big data.

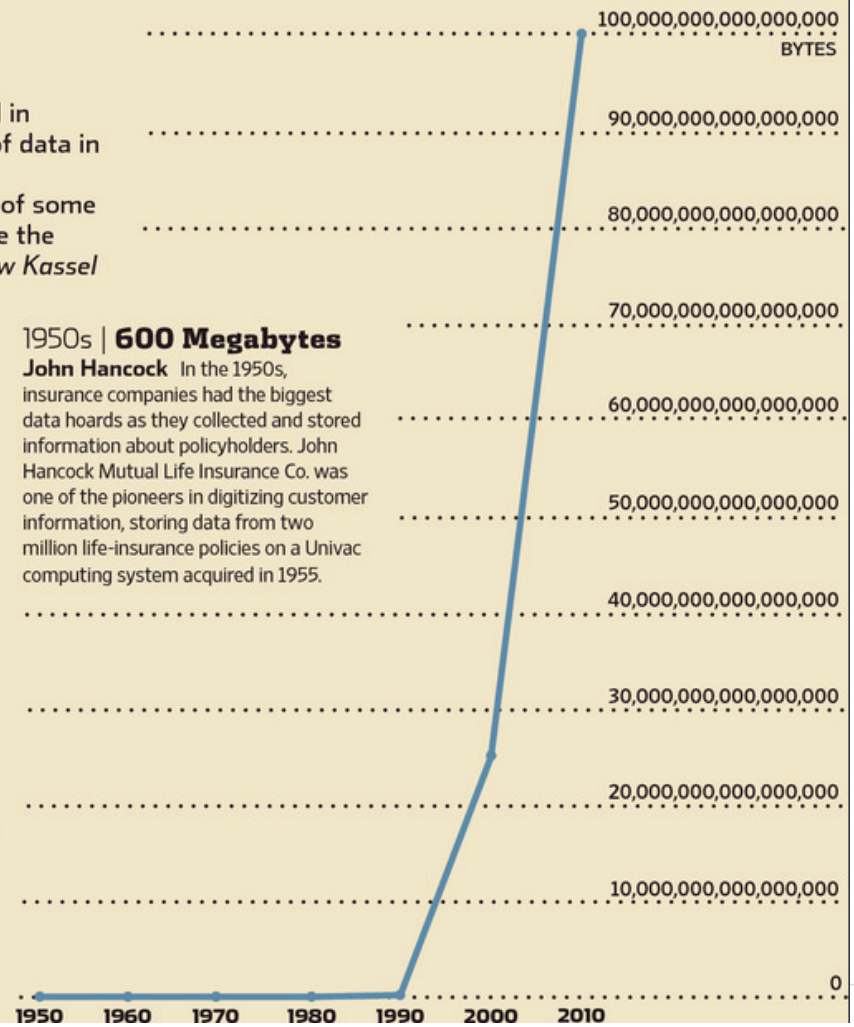
1990s | **180 Terabytes**
Wal Mart Wal-Mart in this decade became the largest American bricks-and-mortar retail operation, and, it is believed, had the biggest commercial data warehouse in the world. To put this number in perspective, experts have estimated that Amazon.com in the late '90s had bytes of stored data.

1980s | **450 Gigabytes**
CitiCorp's NAIB In the 1980s, banks were at the forefront of data growth, with ATMs coming into vogue and banks focusing on collecting and analyzing transaction data for all their businesses.

1970s | **80 Gigabytes**
Federal Express Cosmos FedEx's Cosmos system, introduced in the 1970s, allowed the company to scan and track its huge volume of packages being shipped around the world.

1960s | **807 Megabytes**
American Airlines Sabre In the 1960s, American Airlines developed Sabre, a flight-reservation system built around one of the largest IBM computing systems available. Sabre was one of the first online computing systems, allowing the airline to keep track of an immense matrix of reservations, flight schedules and seat inventories.

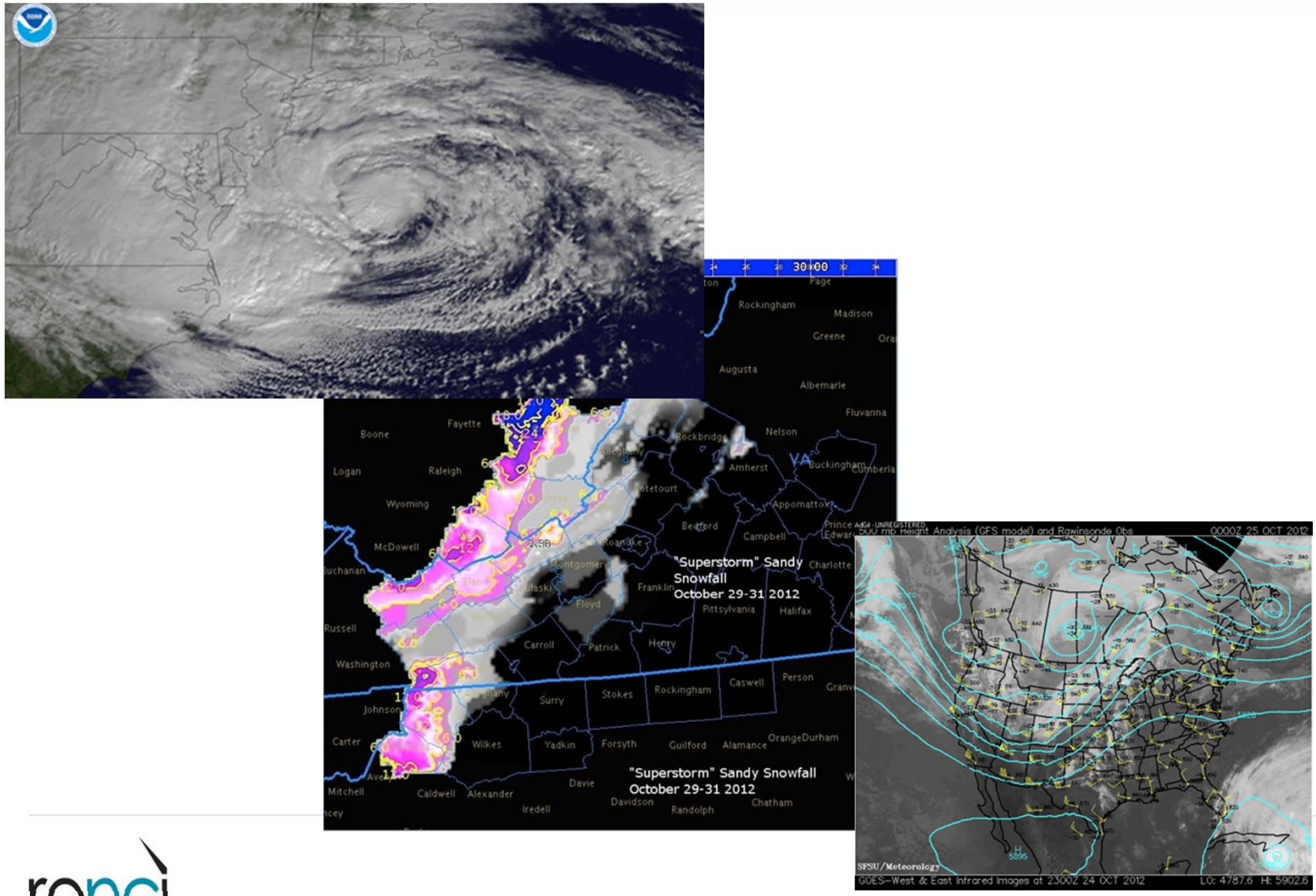
1950s | **600 Megabytes**
John Hancock In the 1950s, insurance companies had the biggest data hoards as they collected and stored information about policyholders. John Hancock Mutual Life Insurance Co. was one of the pioneers in digitizing customer information, storing data from two million life-insurance policies on a Univac computing system acquired in 1955.



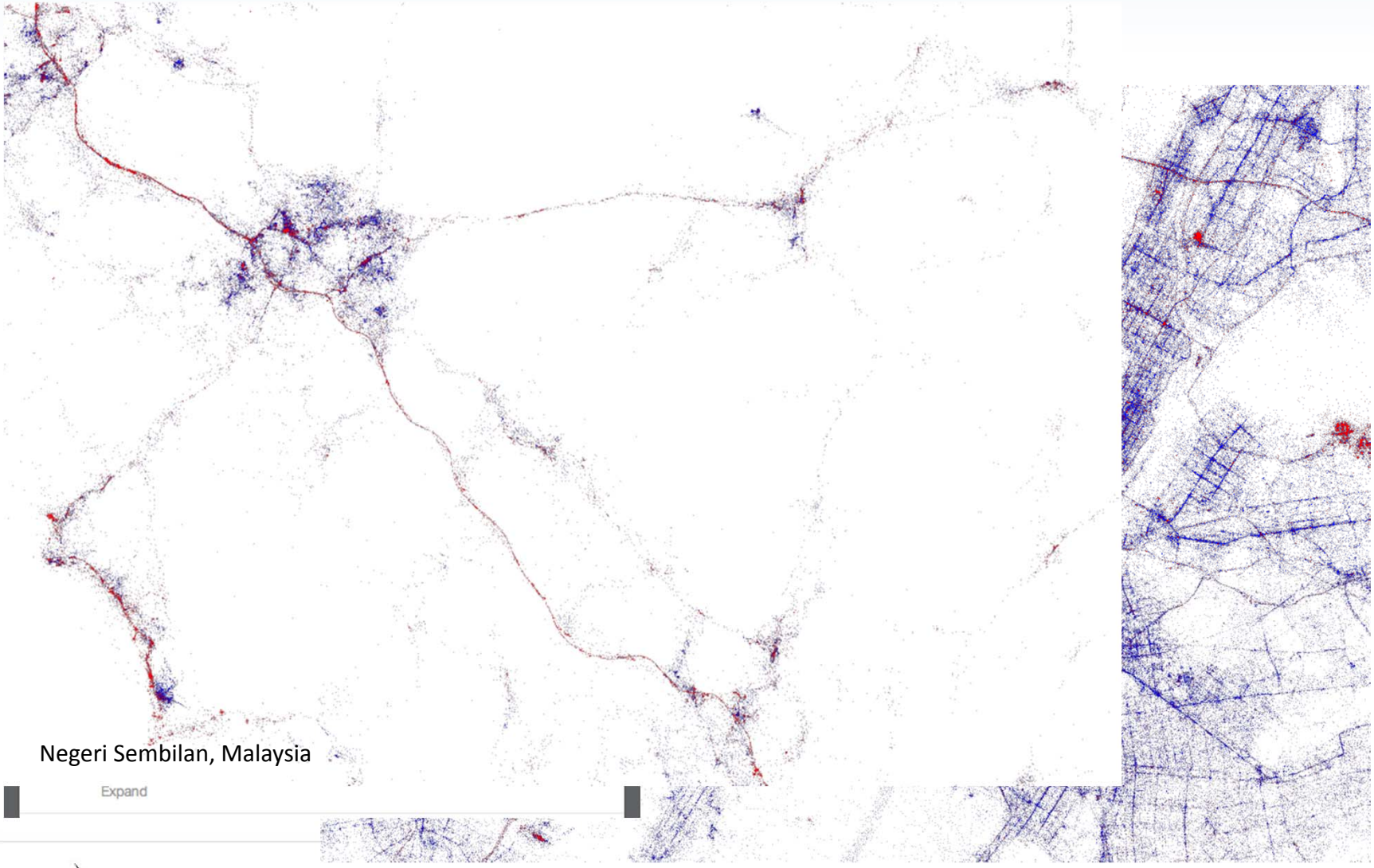
Close

Source: Wall Street Journal, Special Report on Big Data, March 11, 2013

What do we mean by data?



Transforming What We Mean by Data?



http://www.mapbox.com/blog/visualizing-3-billion-tweets/

Sensors are everywhere gathering data…

A cluster of three of the many sensors installed in the wire mesh of the new 35W bridge that measure and send data to civil engineering researchers. (Photo Courtesy of MnDOT)
Image 2 of 4 | < Prev Next >
CLOSE X

Other Big Trends

- Crowd Sourcing
- Open Source
- Open Data
- Open Science
- Store and Manage
- Move
- Policy

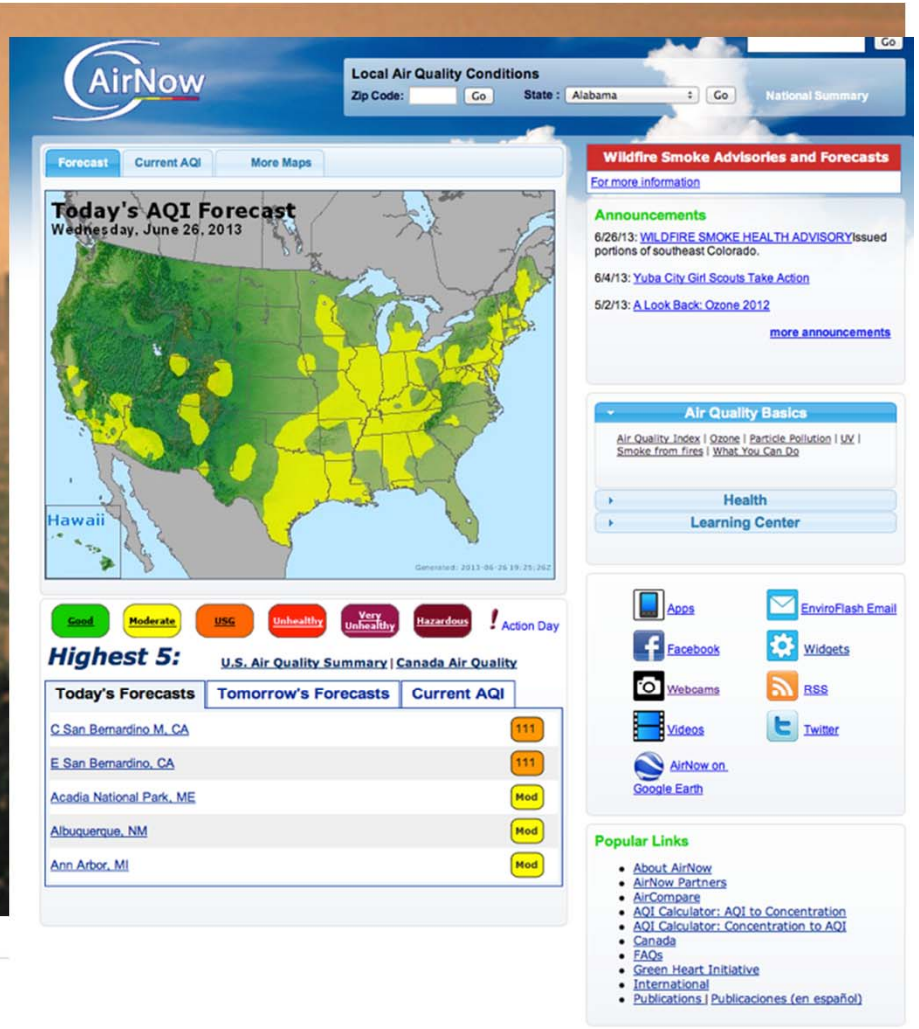
Crowd Sourcing, Open Sourcing, Open Data, Open Science

"A Silicon Valley startup is launching a fleet of imaging satellites that are cheap, small, and ultra-efficient. Their up-to-the-minute snapshots of the planet will give us data that could upend industries, transform economies—even help predict the future."

--Wired, 21.07

http://www.wired.com/wiredscience/2013/06/startup-skybox/

Environmental Health



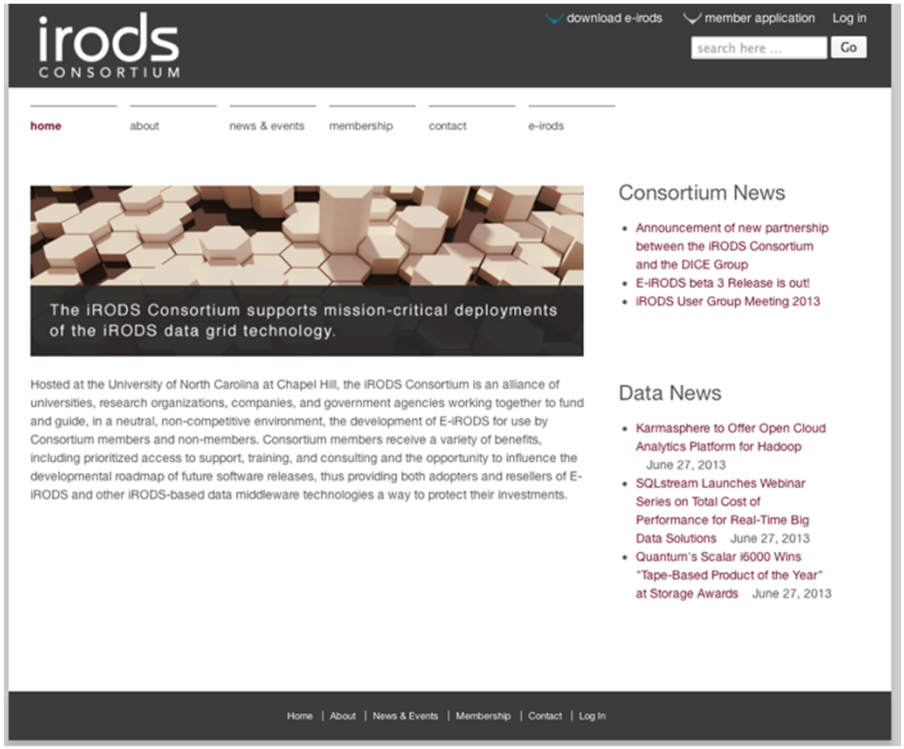
http://airnow.gov/index.cfm?action=airnow.main

Phenology

Yellow-throated Vireo

http://ebird.org/plone/ebird/about/occurrence-maps/yellow-throated-vireo

Alternatives to RDBMS…



http://irods-consortium.org

HPC wire

HPC In the Cloud

datanami

GREEN computing REPORT

HPC wire JAPAN

Tabor Communications Corporate Video

Top News from Leading Digital Manufacturing Solution Providers

Mellanox TECHNOLOGIES

CRAY THE SUPERCOMPUTER COMPANY

sgi

intel

sas THE POWER TO KNOW.

ANSYS

June 21, 2013

GE Looks to Hadoop for the Industrial Internet

Chelsea Lang, Managing Editor

With a name as big as "The Industrial Internet," it shouldn't come as a surprise that implementing the sensor-filled, waste-reducing, machine-optimizing network will require a multitude of research and development from its incubator, General Electric. While we've already learned about the vast array of sensor technologies and operational dashboards in place at GE's Schenectady battery plant, what's been missing thus far is the powerful analytics that would turn what is currently a reactive system into a predictive tool to circumvent failures on the factory floor.

This week, the industrial giant announced such a platform at the D: All Things Digital conference in San Francisco, California. The Hadoop-based big data and analytics platform will include expanded partnerships with Accenture and Pivotal, as well as a new partnership with Amazon Web Services, which will provide cloud storage. Together, they will serve as the infrastructure and analytics behind GE's push for the Industrial Internet.

Thus far the spotlight has fallen on GE's efforts to outfit equipment with the sensors and interconnects necessary to gather and transmit key data from the factory floor and beyond. But with this latest announcement, GE said it will provide "real-time data management, analytics, and machine-to-operations connectivity in a secure, closed-loop architecture so critical global industries can move from a reactive to a predictive industrial operating model."

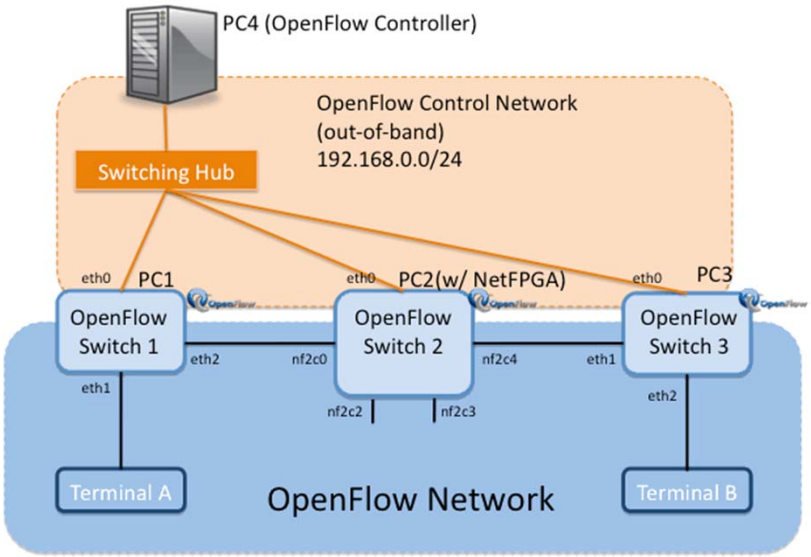
Specifically, the toolkit has been designed to optimize equipment for longevity, energy consumption, and throughput, as well as to predict when a part must be replaced in order to avoid failure, which was done by breaking analytics into two roles: asset health and process performance.

Brian Courtney, General Manager of GE Intelligent Platforms' Industrial Data Intelligence Software group, explained that the goal of asset health is to go beyond merely predicting equipment failure 90 minutes ahead of time to instead accurately forecast failures months in advance. For some industries this means significant reductions in downtime and a boost in productivity, but for others such as aerospace and energy, the power to predict failure in a jet engine or across the power grid could be life-saving.

http://www.digitalmanufacturingreport.com/dmr/2013-06-21/ge_looks_to_hadoop_for_the_industrial_internet.html?featured=top

renci

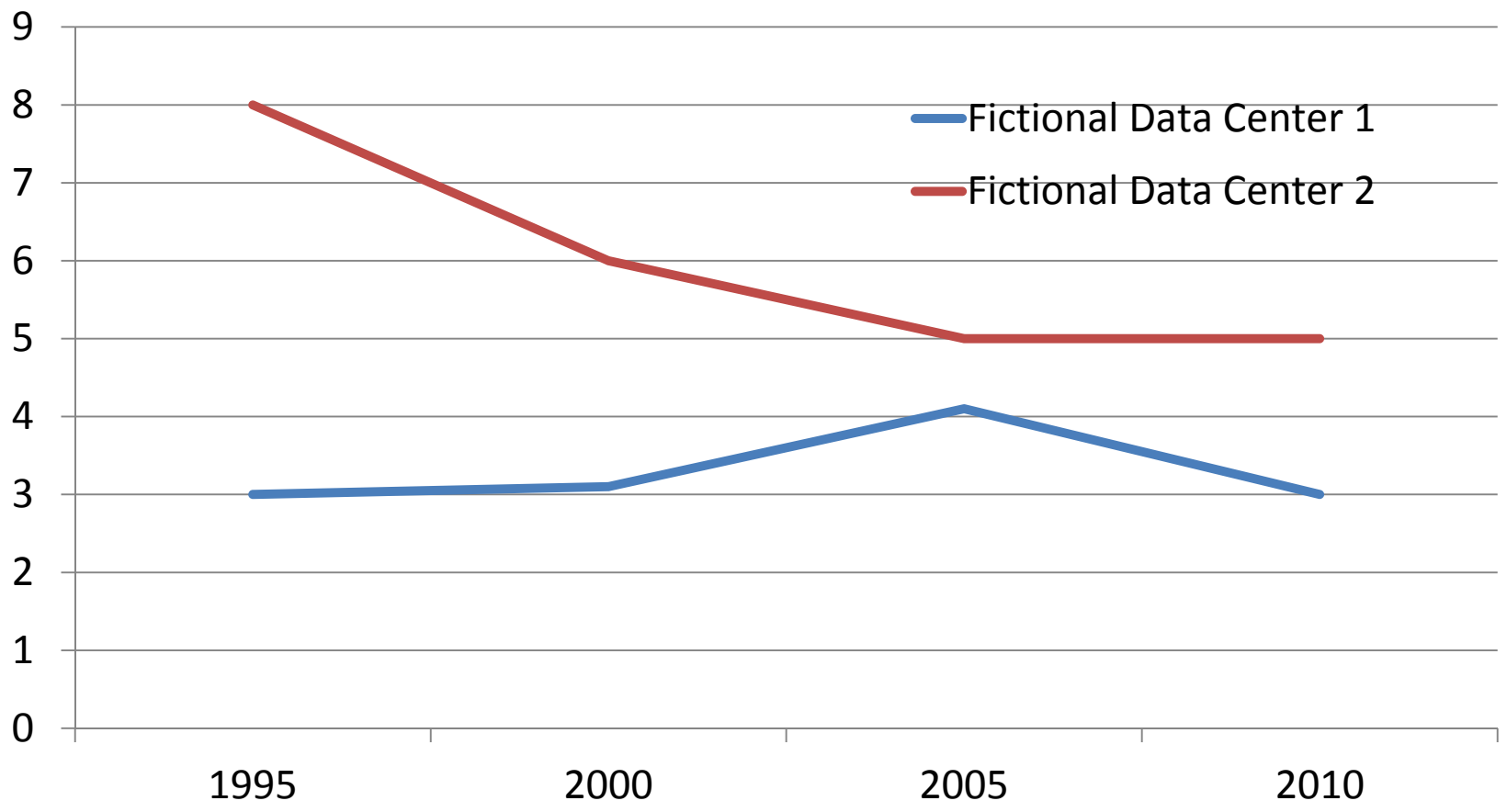
Software Defined Networks



http://www.openflow.org/wp/deploy-labsetup/

http://www.renci.org/news/releases/intercontinental-100-gbps-link-for-research-and-education

Data Center Budgets



9 May 2013 -- Executive Order

"To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the default state of new and modernized Government information resources shall be open and machine readable. Government information shall be managed as an asset throughout its life cycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable. In making this the new default state, executive departments and agencies (agencies) shall ensure that they safeguard individual privacy, confidentiality, and national security."

--President Barack Obama

creative commons | About | Licenses | Public Domain | Support CC | Projects | News

News

Obama issues Executive Order in support of open data

Timothy Vollmer, May 10th, 2013

Yesterday President Barack Obama issued an Executive Order requiring federal government information to be open and machine-readable by default. This Order is the latest in a series of actions going back to 2009 in support of increasing access to and transparency of government information.

In addition to the Executive Order, the White House released a Memorandum (PDF) explaining how federal government agencies will comply with the new open data policy.

> This Memorandum requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities. This includes using machine readable and open formats, data standards, and common core and extensible metadata for all new information creation and collection efforts. It also includes agencies ensuring information stewardship through the use of open licenses and review of information for privacy, confidentiality, security, or other restrictions to release.

It provides a forward-thinking set of guidelines for open data to be released by U.S. federal agencies:

Open data: For the purposes of this Memorandum, the term "open data" refers to publicly available data structured in a way that enables the data to be fully discoverable and usable by end users. In general, open data will be consistent with the following principles:

EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES

Executive Order on Open Data

Seal Of The Executive Office Of The President / Public Domain

What does this mean???

- Big Challenges, Big Science Questions
  - Will require new, innovative ways to solve
  - Require inter, intra, transdisciplinary science (this is what our users will be doing)
  - How to integrate non-traditional users
  - Link data from small scales to large scales; link models, observational data
  - Inter-agency collaboration?!
- Big Data Management
  - Will need to think creatively about how to solve the metadata challenge
  - Move beyond RDBMS
  - Need to re-think what the role of Data Centers is in this new environment
  - Examine the trade-offs between in-depth efforts to develop comprehensive metadata standards and making data available
  - How can technology facilitate addressing these challenges
  - How to effectively manage the data assets in a cost-effective manner

Role of Environmental Data Management

- Central to addressing these challenges
- Need to work hard to get out ahead of these challenges
- Try hard not to be reactive
- Enter into a candid dialog with program management about these issues
- Rethink traditional approaches to problem solving