X-Informatics Introduction: What is Big Data, Data Analytics and X-Informatics? Part III

July 6 2013

Geoffrey Fox
gcf@indiana.edu

http://www.infomall.org/X-InformaticsSpring2013/index.html

Associate Dean for Research, School of Informatics and Computing
Indiana University Bloomington

2013

Big Data Ecosystem in One Sentence

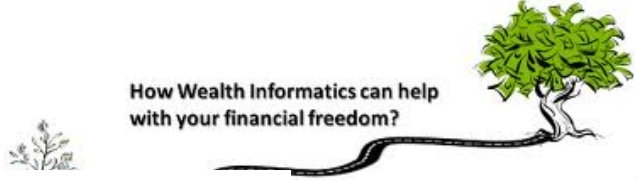
Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X)

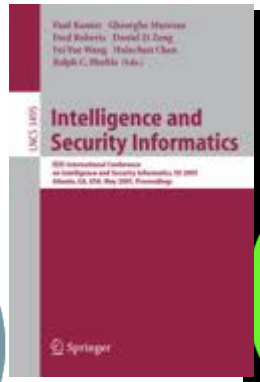
X = Astronomy, Biology, Biomedicine, Business, Chemistry, Climate, Crisis, Earth Science, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly

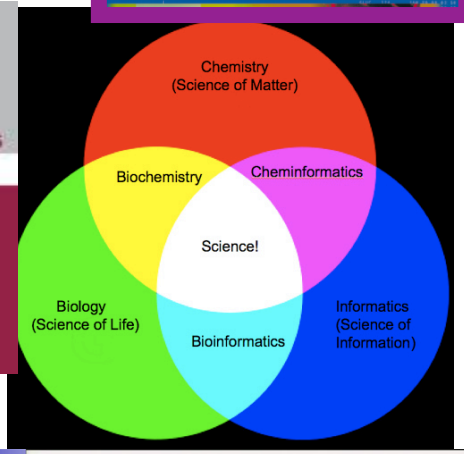
Spans Industry and Science (research)

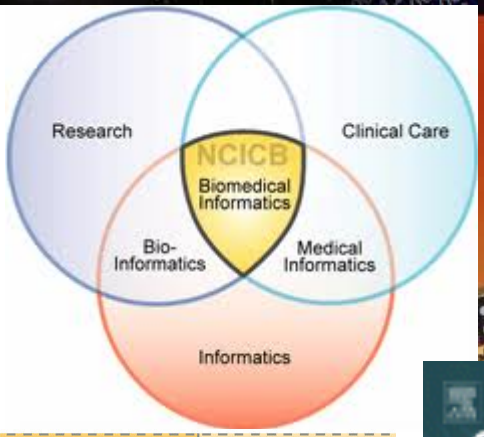
Education: Data Science see recent New York Times articles

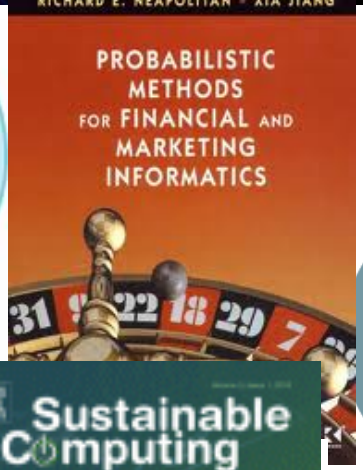
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

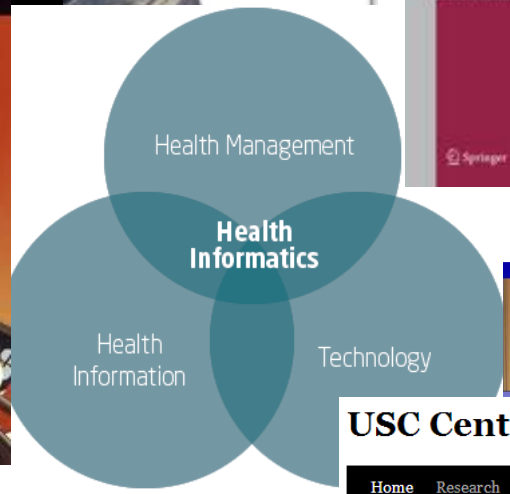


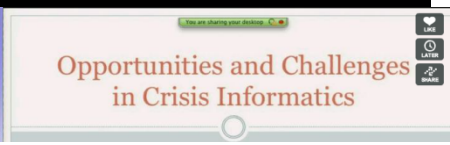


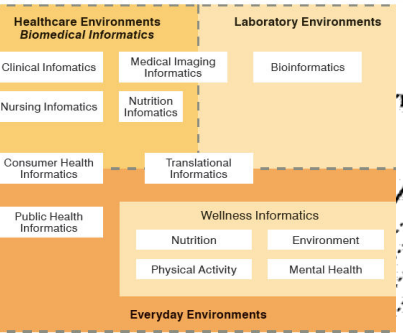


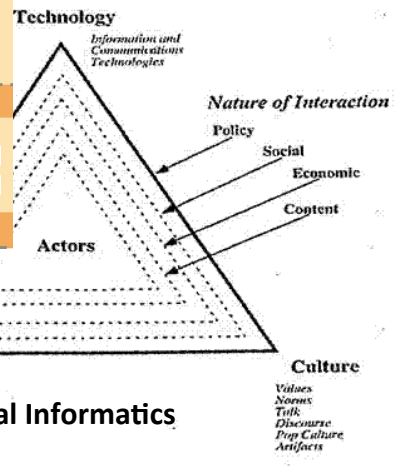








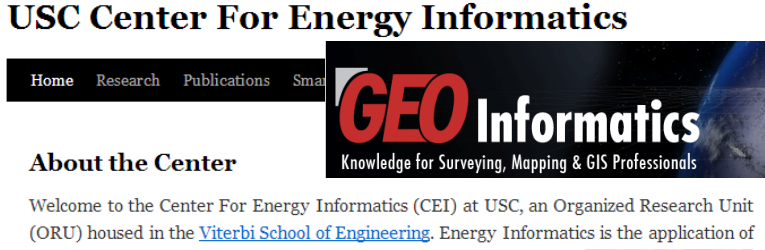


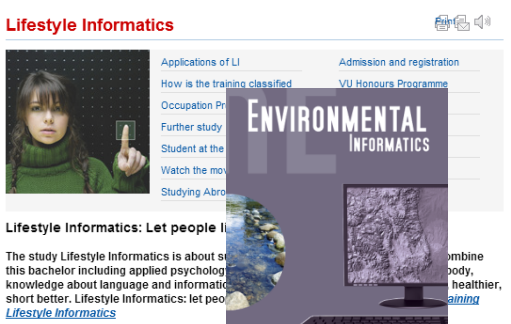


Noelia Penelope Greer (Ed.)

Business Informatics

Information technology, Management,





Clouds

The Microsoft Cloud is Built on Data Centers

~100 Globally Distributed Data Centers

Range in size from “edge” facilities to megascale (100K to 1M servers)

Quincy, WA

eneration 4 DCs

Data Centers Clouds & Economies of Scale I

Range in size from “edge” facilities to megascale.

Economies of scale

Approximate costs for a small size center (1K servers) and a larger, 50K server center.

2 Google warehouses of computers on the banks of the Columbia River, in The Dalles, Oregon

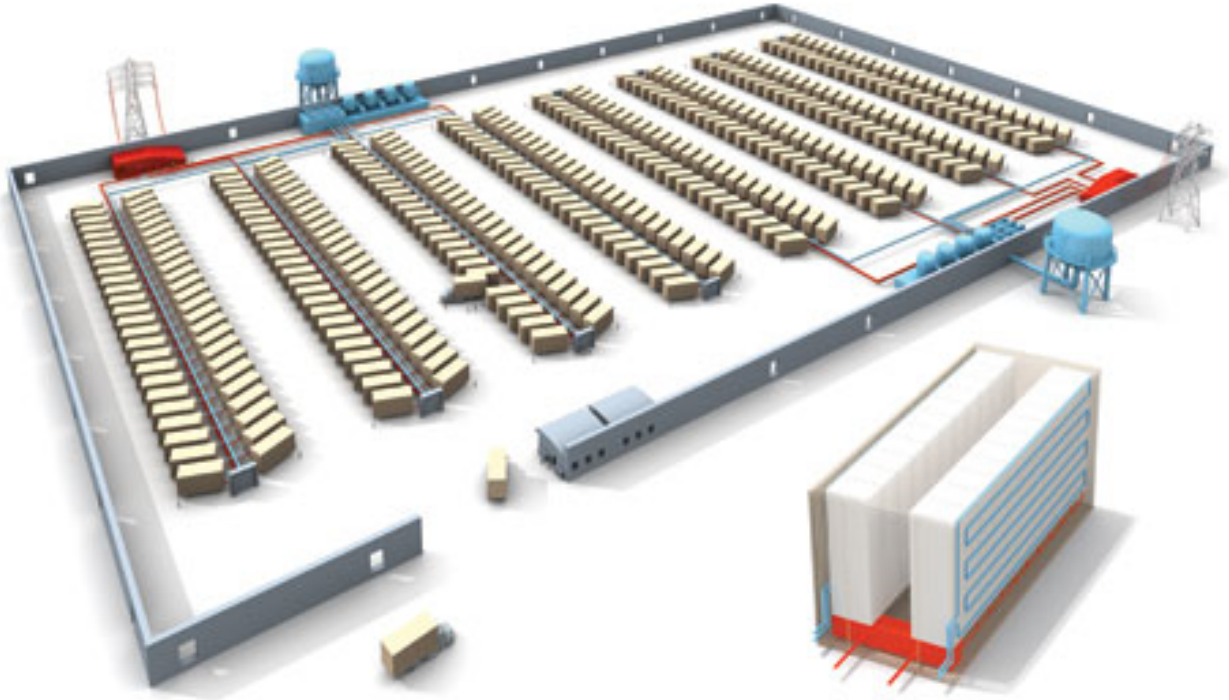
Such centers use 20MW-200MW (Future) each with 150 watts per CPU

Save money from large size, positioning with cheap power and access with Internet

a/sc09_cloudcomp_tutorial.pdf

Data Centers, Clouds & Economies of Scale II

- Builds giant data centers with 100,000's of computers; ~ 200-1000 to a shipping container with Internet access
- "Microsoft will cram between 150 and 220 shipping containers filled with data center gear into a new 500,000 square foot Chicago facility. This move marks the most significant, public use of the shipping container systems popularized by the likes of Sun Microsystems and Rackable Systems to date."



Green Clouds

- Cloud Centers optimize life cycle costs and power use

$$\mathrm{PUE} = \frac{\text{Total facility power}}{\text{IT equipment power}}$$

- http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pue-is-1-8/
- Average PUE = 1.8 (was nearer 3) ; Good Clouds are 1.1-1.2
- 4th generation data centers (from Microsoft) make everything modular so data centers can be built incrementally as in modern manufacturing
- http://loosebolts.wordpress.com/2008/12/02/our-vision-for-generation-4-modular-data-centers-one-way-of-getting-it-just-right/
- Extends container based third generation

Some Sizes in 2010

- http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2011finalversion.pdf
- 30 million servers worldwide
- Google had 900,000 servers (3% total world wide)
- Google total power ~200 Megawatts
  - < 1% of total power used in data centers (Google more efficient than average – **Clouds are Green**!)
  - ~ 0.01% of total power used on anything world wide
- Maybe total clouds are 20% total world server count (a growing fraction)

Some Sizes Cloud v HPC

- **Top Supercomputer** Sequoia Blue Gene Q at LLNL
  - 16.32 Petaflop/s on the Linpack benchmark using 98,304 CPU compute chips with 1.6 million processor cores and 1.6 Petabyte of memory in 96 racks covering an area of about 3,000 square feet
  - 7.9 Megawatts power
- **Largest (cloud)** computing data centers
  - 100,000 servers at ~200 watts per CPU chip
  - Up to 30 Megawatts power
  - Microsoft says upto million servers
- So **largest supercomputer** is around **1-2% performance of total cloud computing systems** with Google ~20% total

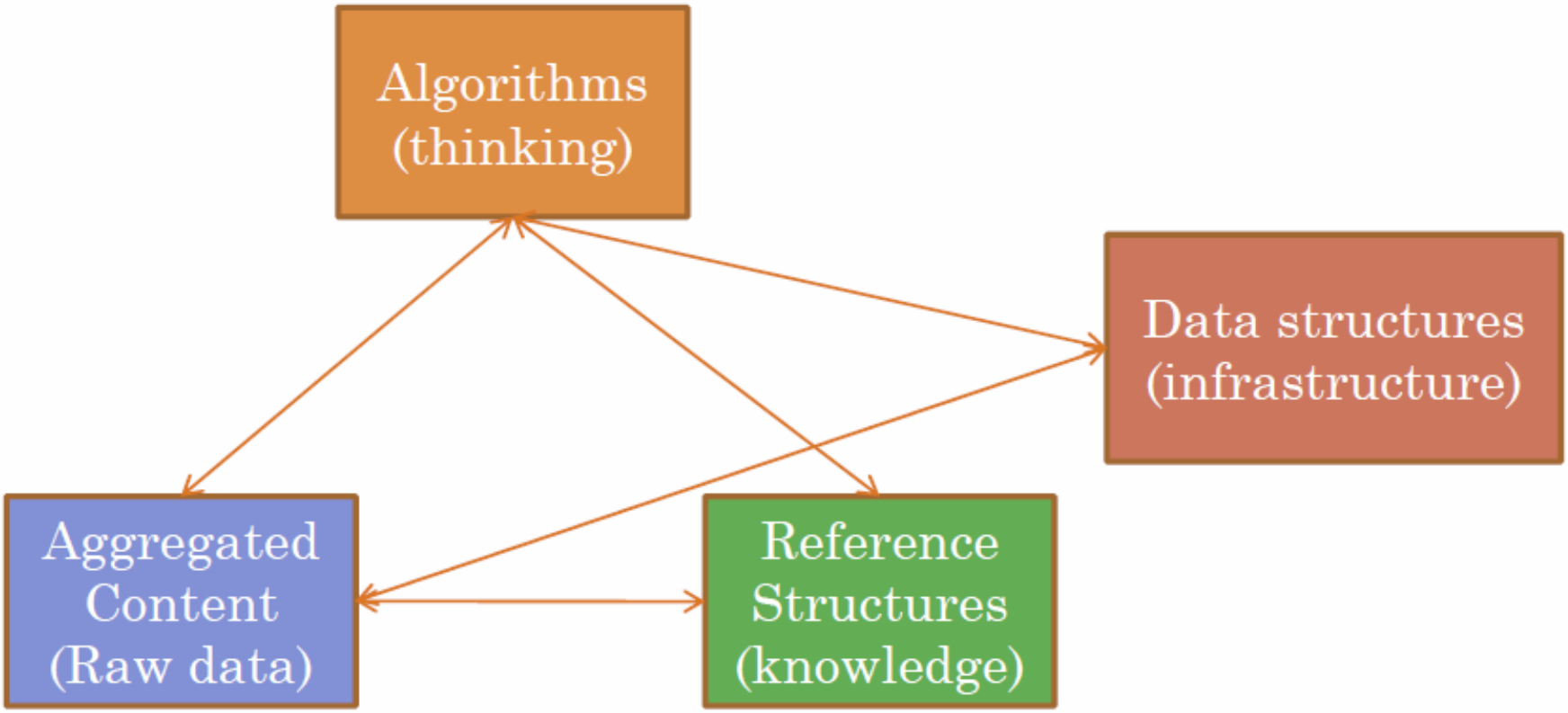
Clouds Offer From different points of view

- **Features from NIST:**
  - On-demand service (elastic);
  - Broad network access;
  - Resource pooling;
  - Flexible resource allocation;
  - Measured service
- **Economies of scale** in performance and electrical power **(Green IT)**
- Powerful new **software models**
  - **Platform as a Service** is not an alternative to **Infrastructure as a Service –** it is instead an incredible valued added
  - New programming ideas including MapReduce and new storage ideas like NOSQL, Bigtable etc. designed for Big Data as invented by Google and internet companies to solve Cloud applications like information retrieval, e-commerce and Social Media

Features of Data Deluge

Intelligence and Scale of Data

- Intelligence is a set of discoveries made by federating/processing information collected from diverse sources.
- Information is a cleansed form of raw data.
- For statistically significant information we need reasonable amount of data.
- For gathering good intelligence we need large amount of information.
- As the Fourth Paradigm (FP) book points out enormous amount of data is generated by the millions of experiments and applications.
- Thus intelligence applications are invariably data-heavy, data-driven and data-intensive.
- Very often the data is gathered from the web (public or private, covert or overt).


Algorithms
(thinking)
Data structures
(infrastructure)
Aggregated
Content
(Raw data)
Reference
Structures
(knowledge)

Basic Elements

Bina Ramamurthy
http://www.cse.buffalo.edu/~bina/cse487/fall2011/

- **Aggregated content**: large amount of data pertinent to the specific application; each piece of information is typically connected to many other pieces. Ex:
- **Reference structures**: Structures that provide one or more structural and semantic interpretations of the content. Reference structure about specific domain of knowledge come in three flavors: dictionaries, knowledge bases, and ontologies
- **Algorithms**: modules that allows the application to harness the information which is hidden in the data. Applied on aggregated content and some times require reference structure Ex: MapReduce
- **Data Structures**: newer data structures to leverage the scale and the WORM characteristics; ex: MS Azure, Apache Hadoop, Google BigTable

Semantic Web/Grid v. Big Data

- Original vision of Semantic Web was that one would annotate (curate) web pages by extra "meta-data" (data about data) to tell web browser (machine, person) the "real meaning" of page
- The success of Google Search is "Big Data" approach; one mines the text on page to find "real meaning"
- Obviously combination is powerful but the pure "Big Data" method is more powerful than expected 15 years ago

More data usually beats better algorithms

Here's how the competition works. Netflix has provided a large data set that tells you how nearly half a million people have rated about 18,000 movies. Based on these ratings, you are asked to predict the ratings of these users for movies in the set that they have not rated. The first team to beat the accuracy of Netflix's proprietary algorithm by a certain margin wins a prize of $1 million!

Different student teams in my class adopted different approaches to the problem, using both published algorithms and novel ideas. Of these, the results from two of the teams illustrate a broader point. Team A came up with a very sophisticated algorithm using the Netflix data. Team B used a very simple algorithm, but they added in additional data beyond the Netflix set: information about movie genres from the Internet Movie Database(IMDB). Guess which team did better?

Anand Rajaraman is Senior Vice President at Walmart Global eCommerce, where he heads up the newly created @WalmartLabs,

http://anand.typepad.com/datawocky/2008/03/more-data-usual.html

20120117berkeley1.pdf Jeff Hammerbacher

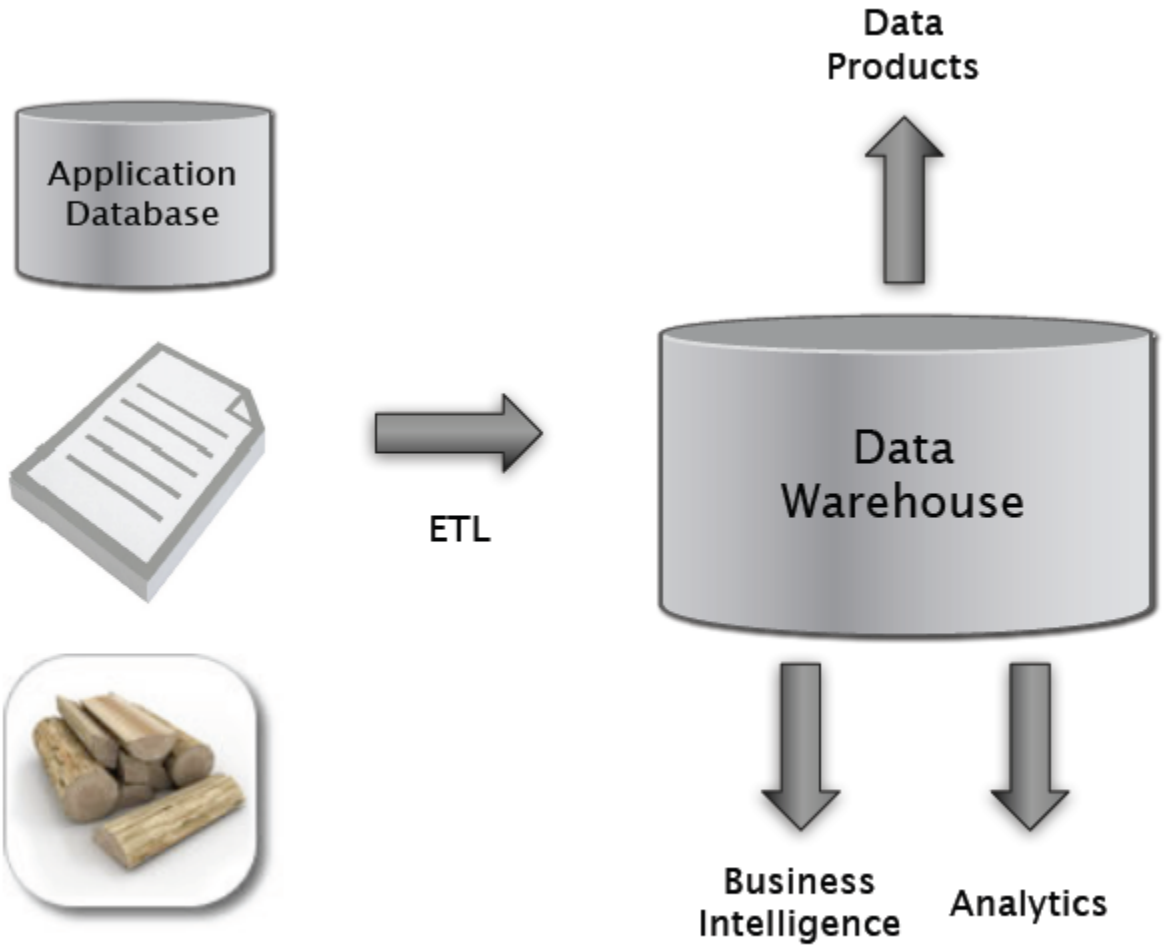
More intelligent data-intensive applications

- Social networking sites
- Mashups : applications that draw upon content retrieved from external sources to create entirely new innovative services.
  - http://www.ibm.com/developerworks/spaces/mashups
- Portals
- Wikis: content aggregators; linked data; excellent data and fertile ground for applying concepts discussed in the text
- Media-sharing sites
- Online gaming
- Biological analysis
- Space exploration

Types of Data Deluge

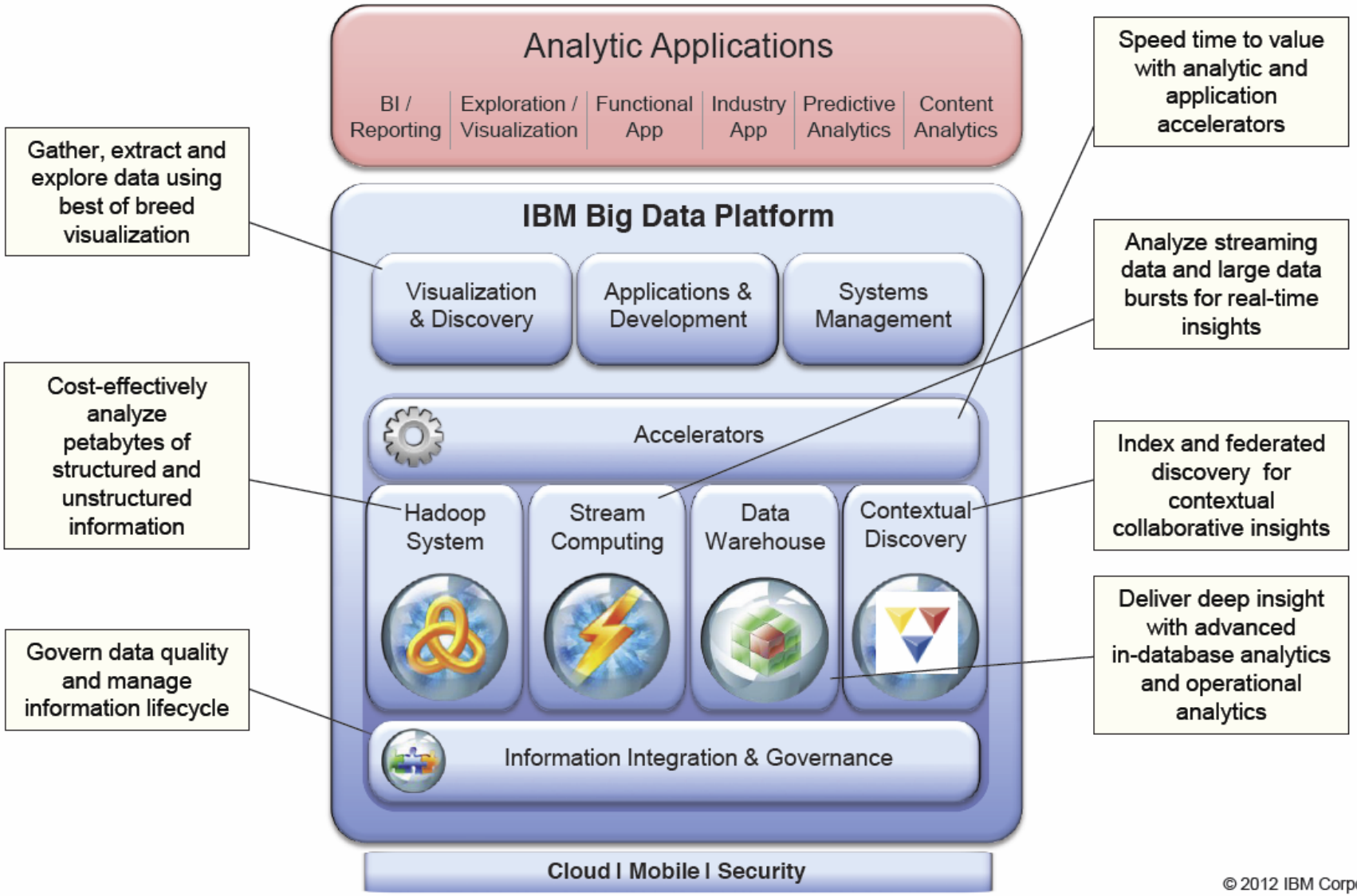
- Traditional business transaction data from credit cards to stock market
- Interaction data as in LinkedIn and Facebook
- Information retrieval with topic and language etc. enhancements
- Recommender and user customization systems
- Marketing information as produced in operation of Walmart
- Pervasive sensors in vehicles and people (health)
- Scientific Instruments such as satellites, telescopes, Large Hadron Colliders, Gene Sequencers
- Military sensors and satellites

Processing Big Data



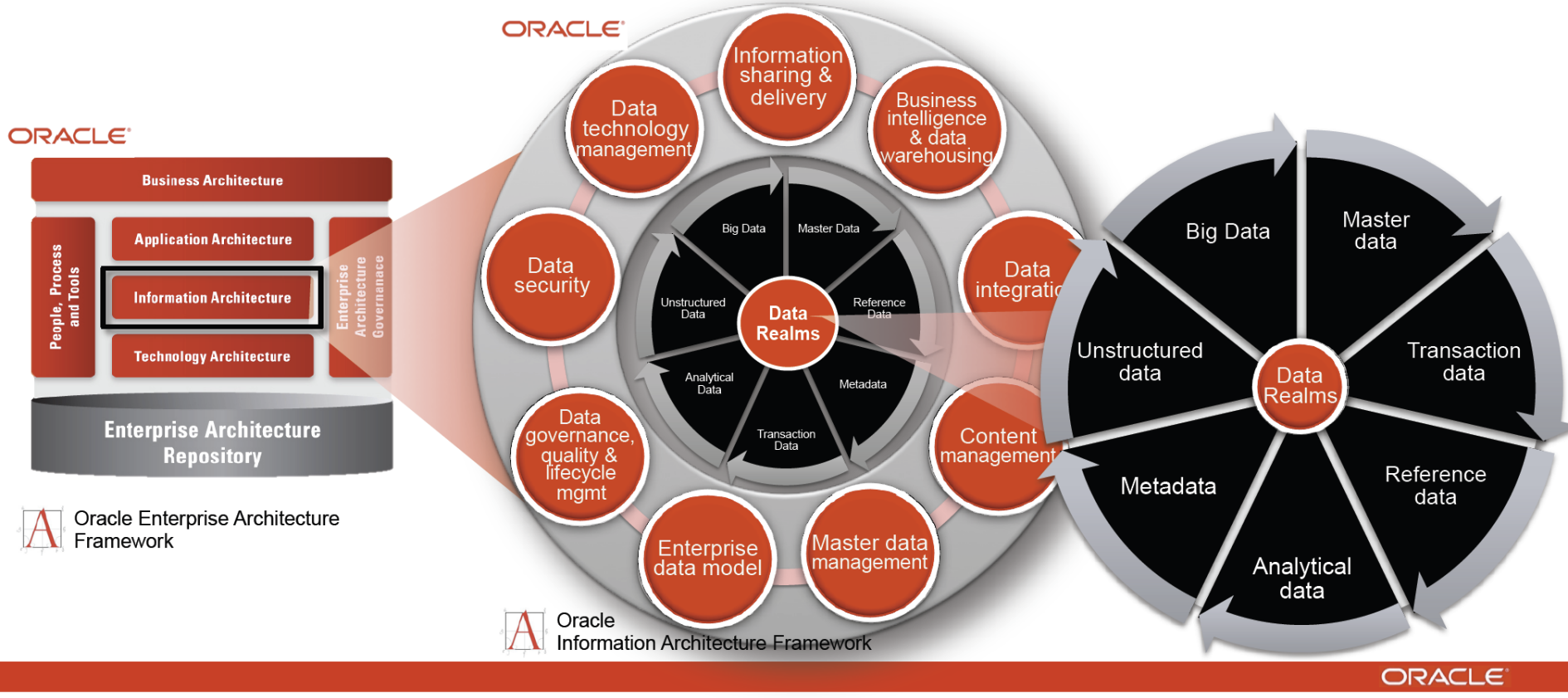
20120119berkeley.pdf Jeff Hammerbacher

Big Data Platform and Application Framework



© 2012 IBM Corporation

Information Architecture Capability Model



http://cs.metrostate.edu/~sbd/ Oracle

Hugh Williams
http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

ONE SIZE DOESN'T FIT ALL

- There isn't one solution for driving an organization with big data:
  - **Hadoop** is for:
    - Engineers, batch (asynchronous), map reduce (divide and conquer), unstructured, flexible problems
  - **HBase** is for:
    - Engineers, real-time, large data blob, unstructured, key lookup, flexible problems
  - **Teradata** (or another data warehousing solution) is for:
    - Analysts, real-time or batch, structured, flexible problems
  - **Cassandra** (or **MongoDB** or …) is for:
    - Engineers, real-time, smaller data blob, unstructured, key lookup, flexible problems
  - Some problems warrant specialized solutions

ebay

Hugh Williams
http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

Data Science Process

The Rise of the Data Scientist

Hybrids	• Half analytical, with modeling, statistics, and experimentation skills • Half focused on data management – extraction, filtering, sampling, structuring • Lots of programming skills – Python, Ruby, Hadoop, Pig, Hive
Scientific	• Experimental physicists • Computational biologists • Statisticians with dirty hands • Ecologists, anthropologists, psychologists, etc.
Impatient	• Try something and iterate • Don't wait for a data person to get your data • "We're a pain in the ass" • Job tenure is short
Ground-breaking	• "Nobody's ever done this before" • "If we wanted to deal with structured data, we'd be on Wall Street" • "Being a consultant is the dead zone – too hard to get things implemented" • "The output should be a product or a demo – not a report"

Tom Davenport Harvard Business School http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html Nov 2012

Is Davenport Correct?

- There are the 1.5 million decision makers/ managers of McKinsey report
- Up to 190,000 “nerds”
- Davenport appears to describe nerds not the larger 1.5M body of “generalists”

Jeff Hammerbacher's Process

1) Identify problem
2) Instrument data sources
3) Collect data
4) Prepare data (integrate, transform, clean, impute, filter, aggregate)
5) Build model
6) Evaluate model
7) Communicate results

20120117berkeley1.pdf Jeff Hammerbacher

Another Jeff Hammerbacher Process

1) Obtain
2) Scrub
3) Explore
4) Model
5) Interpret

20120117berkeley1.pdf Jeff Hammerbacher

Statistician Colin Mallows

1) Identify data to collect and its relevance to your problem
2) Statistical specification of the problem
3) Method selection
4) Analysis of method
5) Interpret results for non-statisticians

Ben Fry Data Visualization
http://en.wikipedia.org/wiki/Benjamin_Fry

1) Acquire
2) Parse
3) Filter
4) Mine
5) Represent
6) Refine
7) Interact

Peter Huber Statistics (UCB)

1) Inspection
2) Error checking
3) Modification
4) Comparison
5) Modeling and model fitting
6) Simulation
7) What-if analyses
8) Interpretation
9) Presentation of conclusions

Jim Gray Microsoft Database/ eScience

1) Capture
2) Curate
3) Communicate

Ted Johnson

1) Assemble an accurate and relevant data set
2) Choose the appropriate algorithm

20120117berkeley1.pdf Jeff Hammerbacher

Data Analytics

Big Data and Small Data Analytics – How Do They Compare?

Focus

- Big data is often external, small data often internal
- Big data is often part of a product or service, small data is used to manage

Relationships

- Big data and small data analysts require good relationships
- But relationships are different: product managers and customers for big data analysts; internal managers for small data analysts

Technologies

- Big data requires data management (Hadoop, Pig, Hive, Python)
  - Analysis in visual (Tableau, Spotfire), open-source (R) tools
- Small data requires less data management – SQL is sufficient
  - Analysis in BI (BO, Cognos, Qlikview) or statistical (SAS, SPSS) tools

Tom Davenport Harvard Business School http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html Nov 2012

General Remarks I

- **An immature (exciting) field:** No agreement as to what is data analytics and what tools/computers needed
  - Databases or NOSQL?
  - Shared repositories or bring computing to data
  - What is repository architecture?
- **Sources:** Data from observation or simulation
- **Different terms:** Data analysis, Datamining, Data analytics., machine learning, Information visualization
- **Fields:** Computer Science, Informatics, Library and Information Science, Statistics, Application Fields including Business
- **Approaches:** Big data (cell phone interactions) v. Little data (Ethnography, surveys, interviews)
- **Topics:** Security, Provenance, Metadata, Data Management, Curation

General Remarks II

- **Tools:** Regression analysis; biostatistics; neural nets; bayesian nets; support vector machines; classification; clustering; dimension reduction; artificial intelligence; semantic web
- **One driving force:** Patient records growing fast
- **Another:** Abstract graphs from net leads to community detection
- Some data in metric spaces; others very high dimension or none
- Large Hadron Collider analysis mainly histogramming – all can be done with MapReduce (larger use than MPI)
- **Commercial:** Google, Bing largest data analytics in world
- **Time Series:** Earthquakes, Tweets, Stock Market (**Pattern Informatics**)
- **Image Processing** from climate simulations to NASA to DoD to Radiology (Radar and Pathology Informatics – same library)
- **Financial decision support**; marketing; fraud detection; automatic preference detection (map users to books, films)

Algorithms for Data Analytics

- In simulation area, it is observed that equal contributions to improved performance come from increased computer power and better algorithms http://cra.org/ccc/docs/nitrdsymposium/pdfs/keyes.pdf
- In data intensive area, we haven't seen this effect so clearly
  - Information retrieval revolutionized but
  - Still using Blast in Bioinformatics (although Smith Waterman etc. better)
  - Still using R library which has many non optimal algorithms
  - Parallelism and use of GPU's often ignored

"Moore's Law" for fusion energy simulations

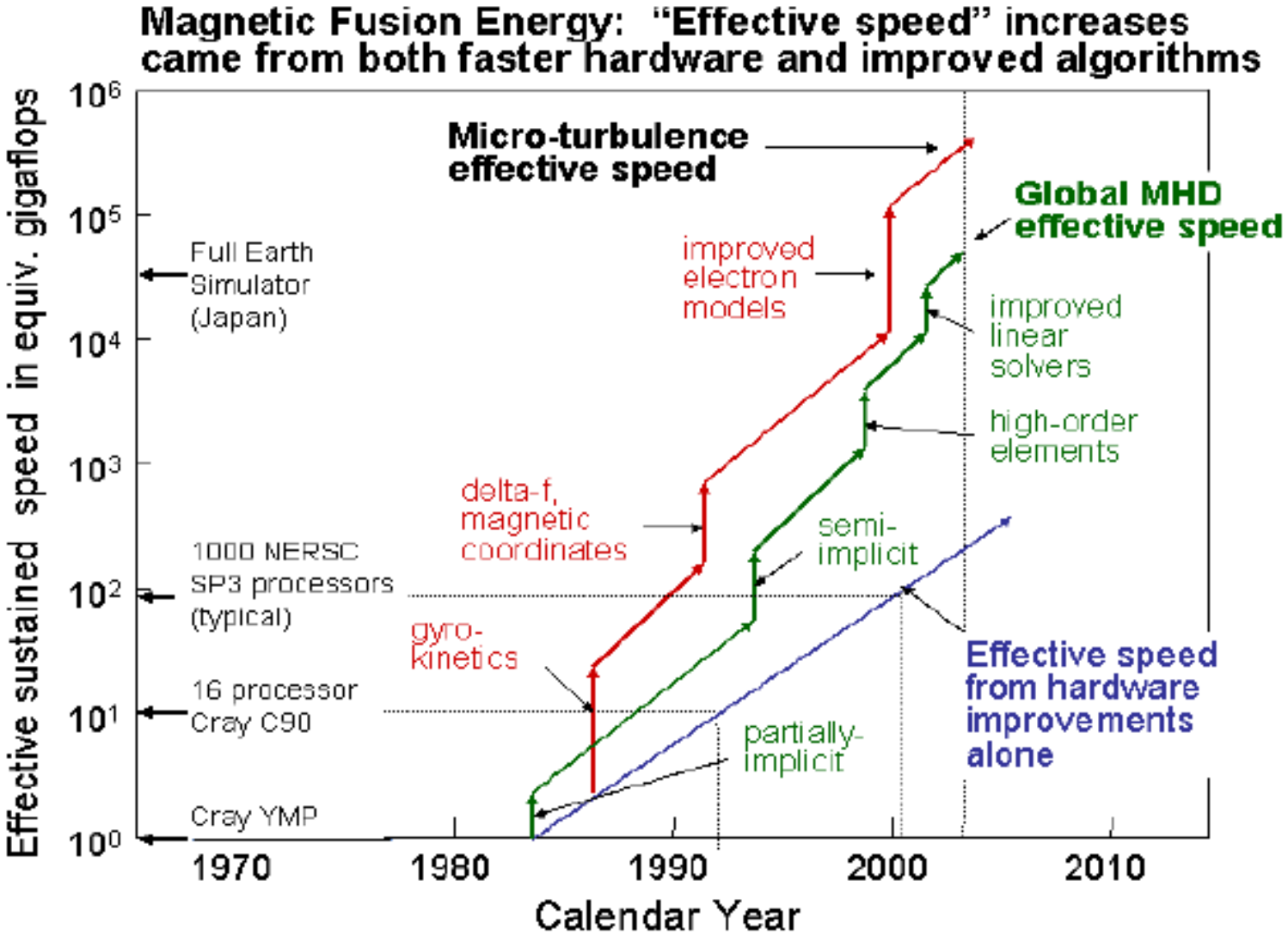


Figure from DOE "SCaLeS report" Volume 2 (Keyes et al., 2004)