X-Informatics Motivation: Big Data and the Cloud; Centerpieces of the Future Economy

July 7 2013

Geoffrey Fox
gcf@indiana.edu
http://www.infomall.org http://www.futuregrid.org

School of Informatics and Computing
Digital Science Center
Indiana University Bloomington

https://portal.futuregrid.org

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/

X-informatics
How Wealth Informatics can help with your financial freedom?
Xinformatics
Earth Science INFORMATICS
Climate Informatics network
Biomedical Informatics
Computer Applications in Health Care and Biomedicine
Journal of Pathology Informatics
AstroInformatics2012
Redmond, WA, September 10 - 14, 2012
Intelligence and Security Informatics
Chemistry (Science of Matter)
Biochemistry
Cheminformatics
Science!
Biology (Science of Life)
Bioinformatics
Informatics (Science of Information)
Research
Clinical Care
NCICB
Biomedical Informatics
Bio-Informatics
Medical Informatics
Informatics
RICHARD E. NEAPOLITAN · XIA JIANG
PROBABILISTIC METHODS FOR FINANCIAL AND MARKETING INFORMATICS
Health Management
Health Informatics
Health Information
Technology
Opportunities and Challenges in Crisis Informatics
Sustainable Computing
Informatics & Systems
USC Center For Energy Informatics
Home
Research
Publications
About the Center
Welcome to the Center For Energy Informatics (CEI) at USC, an Organized Research Unit (ORU) housed in the Viterbi School of Engineering. Energy Informatics is the application of
GEO Informatics
Knowledge for Surveying, Mapping & GIS Professionals
Healthcare Environments
Biomedical Informatics
Laboratory Environments
Clinical Infomatics
Medical Imaging Informatics
Bioinformatics
Nursing Infomatics
Nutrition Infomatics
Consumer Health Informatics
Translational Informatics
Public Health Informatics
Wellness Informatics
Nutrition
Environment
Physical Activity
Mental Health
Everyday Environments
Technology
Information and Communications Technologies
Nature of Interaction
Policy
Social
Economic
Content
Actors
Institutions
Societies
Markets
Social Communities
Organizations
Groups
Households
Processes
Procedures
Rules
Tasks
Culture
Values
Norms
Talk
Discourse
Pop Culture
Artifacts
Social Informatics
Noelia Penelope Greer (Ed.)
Business Informatics
Information technology, Management,
policy informatics network
ASU School of Public Affairs
ARIZONA STATE UNIVERSITY
Lifestyle Informatics
Applications of LI
How is the training classified
Occupation
Further study
Student at the
Watch the
Studying
Admission and registration
VU Honours Programme
Lifestyle Informatics: Let people
The study Lifestyle Informatics is about
this bachelor including applied psycholog
knowledge about language and informatic
short better. Lifestyle Informatics: let peo
Lifestyle Informatics
ENVIRONMENTAL INFORMATICS
BACHELOR-VOORLICHTINGSDAG
ZATERDAG 3 NOVEMBER
LOOP EEN DAG MEE MET EEN STUDENT

Introduction

Abstract

- There is an endlessly growing amount of data as we record every transaction between people and the environment (whether shopping or on a social networking site) while smart phones, smart homes, ubiquitous cities, smart power grids, and intelligent vehicles deploy sensors recording even more.
- Science with satellites and accelerators is giving data on transactions of particles and photons at the microscopic scale.
- This data are and will be stored in immense clouds with co-located storage and computing that perform "analytics" that transform data into information and then to wisdom and decisions; data mining finds the proverbial knowledge diamonds in the data rough.
- This disruptive transformation is driving the economy and creating millions of jobs in the emerging area of "data science".
- We discuss this revolution and its implications for universities and society

Some Trends

- **The Data Deluge** is clear trend from Commercial (Amazon, e-commerce) , Community (Facebook, Search) and Scientific applications
- **Smaller (INTEL/ARM/AMD) chips** drive
  - **Multicore (i.e. more computing) on shared servers**
  - **Smaller Light weight clients** from smartphones, tablets to sensors (i.e. more clients)
- **Clouds with cheaper, greener, easier to use** IT for applications
- **New jobs** associated with new curricula
  - **Clouds as a distributed system** (changing a classic CS course)
  - **Data Science** (new area)

Emerging Technologies Hype Cycle 2012

48 technologies are listed in this year's hype cycle which is the highest in last ten years. Year 2008 was the lowest (27)

Gartner Says: ***We are at an interesting moment — a time when the scenarios we've been talking about for a long time are almost becoming reality.***

Issues of Importance

- **Economic Imperative:** There are a lot of data and a lot of jobs
- **Computing Model:** Industry adopted clouds which are attractive for data analytics
- **Research Model:** 4th Paradigm; From Theory to Data driven science?
- **Research/Business** opportunities in advancing **computing** technologies and algorithms
- **Research/Business** opportunities in **X-Informatics**: applying 4th paradigm (more here!)
- Development in **Data Science Education**: opportunities at universities

Data Deluge

World's Content is Increasingly Findable + Shared + Tagged - Digital Info Created + Shared up 9x in Five Years

Amount of global digital information created & shared – from documents to pictures to tweets – grew 9x in five years to nearly 2 zettabytes in 2011, per IDC.*

Video = 100 Hours Per Minute Uploaded to YouTube, Up from 20 hours Six Years Ago

Source: YouTube.

Photos = 500MM+ Uploaded & Shared Per Day, Growth Accelerating, on Trend to Rise 2x Y/Y…

ı publicly disclosed company data.

http://cs.metrostate.edu/~sbd/ Oracle Future Grid https://portal.futuregrid.org

"Taming the Big Data Tidal Wave" 2012 (Bill Franks, Chief Analytics Officer Teradata)

- **Web Data** ("the original big data")
  - Analyze customer web browsing of e-commerce site to see topics looked at etc.
- **Auto Insurance** (telematics monitoring driving)
  - Equip cars with sensors
- **Text data** in multiple industries
  - Sentiment analysis, identify common issues (as in eBay lamp example), Natural Language processing
- **Time and location** (GPS) data
  - Track trucks (delivery), vehicles(track), people(tell them nearby goodies)
- Retail and manufacturing: **RFID**
  - Asset and inventory management,
- Utility industry: **Smart Grid**
  - Sensors allow dynamic optimization of power
- Gaming industry: Casino Chip tracking (RFID)
  - Track individual players, detect fraud, identify patterns
- **Industrial engines** and equipment: sensor data
  - See GE engine
- **Video games**: telemetry
  - This is like monitoring web browsing but rather monitor actions in a game
- Telecommunication and other industries: **Social Network** data
  - Connections make this big data.
  - Use connections to find new customers with similar interests

Scale of Industrial Internet

Social media versus electric generating power source

2012 Twitter Usage

Gas Turbine Compressor Blade Monitoring potential*

vs.

80 Gigabytes per day

enabling social connections

588 Gigabytes per day

enabling capital asset productivity

Data volume potential is 7x greater from a gas turbine than current Twitter usage

imagination at work

Ruh VP Software GE http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

eral Electric Company, 2012. All Rights Reserved.

50MW are equipped with Blade Health Monitoring capabilities

Value of Data & Analytics

Monitor fleet of ~25,000* engines ... 3.6MM flight records/month

B777

Prognostics

- ✓ Dispatch reliability
- ✓ Preventive maintenance
- ✓ Asset utilization

Prevent failures = customer efficiency

+

GE90

Asset Productivity

- ✓ Enhanced service offerings
- ✓ Airline cost structure
- ✓ Fuel performance

Streamline operations = increased airline productivity

=

DATA

90,000 flight records analyzed

~200 parameters per flight record

~18MM parameters per month

System & Optimization

- ✓ Time & space management
- ✓ Fuel efficiency
- ✓ Airspace capacity

Integrated systems = value-added services

Drives strong alignment with customers

Creates productivity in long-term service agreements

Value-added services fuels growth

MM = Million

imagination at work

© General Electric Company, 2012. All Rights Reserved.

is 50/50 JV with Pratt & Whitney

Ruh VP Software GE http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

Some Science/Technical Data sizes

- **LHC Particle Physics** 15 petabytes per year
- **Radiology** 69 petabytes per year
- **Square Kilometer Array Telescope** will be 0.5 zettabytes per year raw data in ~2022
- **Earth Observation** becoming ~4 petabytes per year
- **Earthquake Science** – few terabytes total today
- **PolarGrid** Radar studies of glaciers– 100's terabytes/year
- **Exascale simulation** data dumps – ~0.1 zettabyte per year

Cost per Genome

Need cost effective Computing!

Sequence every newborn by 2019 100 petabytes/year

NIH National Human Genome Research Institute

genome.gov/sequencingcosts

http://www.genome.gov/sequencingcosts/

The Long Tail of Science

Collectively "long tail" science is generating a lot of data
Estimated at over 1PB per year and it is growing fast.

80-20 rule: 20% users generate 80% data but not necessarily 80% knowledge

Data Intensive Activities

Bag=Space

- **Particle Physics** LHC (bag of events of particles)
- **Information Retrieval** or web search (bag of words)
- **e-commerce** (bag of items with properties or users with rankings)
- **Social Networking** (bag of people with links & properties)
- **Health Informatics** (bag of health records, gene sequences)
- **Sensors** – web cams, self driving cars etc. (bag of pixels)
- **Using**
- **Statistics** (Histograms, Chisq)
- **Deep Learning** (Machine Learning)
- **Image Analysis** (including internet uploaded images)
- **Recommender Engines** (Bag of Ratings or properties)
- **Patterns or Anomaly** detection in graphs (linked data)
- **On Clouds** using MapReduce etc.

Big Data Ecosystem in One Sentence

Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics educated in data science

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)

X-informatics
How Wealth Informatics can help with your financial freedom?
Xinformatics
Earth Science INFORMATICS
Climate Informatics network
Biomedical Informatics
Computer Applications in Health Care and Biomedicine
AstroInformatics2012
Redmond, WA, September 10 - 14, 2012
Journal of Pathology Informatics
Intelligence and Security Informatics
Chemistry (Science of Matter)
Biochemistry
Cheminformatics
Science!
Biology (Science of Life)
Bioinformatics
Informatics (Science of Information)
RICHARD E. NEAPOLITAN · XIA JIANG
PROBABILISTIC METHODS FOR FINANCIAL AND MARKETING INFORMATICS
Research
Clinical Care
NCICB
Biomedical Informatics
Bio-Informatics
Medical Informatics
Informatics
Health Management
Health Informatics
Health Information
Technology
Opportunities and Challenges in Crisis Informatics
Sustainable Computing
Informatics & Systems
USC Center For Energy Informatics
Home
Research
Publications
About the Center
Welcome to the Center For Energy Informatics (CEI) at USC, an Organized Research Unit (ORU) housed in the Viterbi School of Engineering. Energy Informatics is the application of
GEO Informatics
Knowledge for Surveying, Mapping & GIS Professionals
Healthcare Environments
Biomedical Informatics
Laboratory Environments
Clinical Infomatics
Medical Imaging Informatics
Bioinformatics
Nursing Infomatics
Nutrition Infomatics
Consumer Health Informatics
Translational Informatics
Public Health Informatics
Wellness Informatics
Nutrition
Environment
Physical Activity
Mental Health
Everyday Environments
Technology
Information and Communications Technologies
Nature of Interaction
Policy
Social
Economic
Content
Actors
Institutions
Societies
Markets
Social Communities
Organizations
Groups
Households
Processes
Procedures
Rules
Tasks
Culture
Values
Norms
Talk
Discourse
Pop Culture
Artifacts
Social Informatics
Noelia Penelope Greer (Ed.)
Business Informatics
Information technology, Management,
policy informatics network
ASU School of Public Affairs
ARIZONA STATE UNIVERSITY
Lifestyle Informatics
Applications of LI
How is the training classified
Occupation
Further study
Student at the
Watch the
Studying
Admission and registration
VU Honours Programme
Lifestyle Informatics: Let people
The study Lifestyle Informatics is about
this bachelor including applied psycholog
knowledge about language and informatic
short better. Lifestyle Informatics: let peo
Lifestyle Informatics
ENVIRONMENTAL INFORMATICS
BACHELOR-VOORLICHTINGSDAG
ZATERDAG 3 NOVEMBER
LOOP EEN DAG MEE MET EEN STUDENT

Jobs

Jobs v. Countries

http://www.microsoft.com/en-us/news/features/2012/mar12/03-05CloudComputingJobs.aspx

McKinsey Institute on Big Data Jobs

- There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.
- Informatics aimed at 1.5 million jobs. Computer Science covers the 140,000 to 190,000 http://www.mckinsey.com/mgi/publications/big_data/index.asp.

The Rise of Data Scientists and Analysts

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

Computer Science Job Opening Forecast = 2.4x # of Computer Science Graduates

ureau of Labor Statistics data.

Startups & Big Tech Companies Can't Hire Enough Engineers

Street View From Silicon Valley

Five High-Tech Companies Alone – IBM, Intel, Microsoft, Oracle And Qualcomm – Have Combined 10,000 Current Openings In The United States.

Image: View outside of Quixey's office on Castro St. in Mountain View, CA. Picture by Joshua Stanton-
http://www.cbsnews.com/news/ident and lawmakers, 3/13.
.30388692/Tech-CEO-letter

Industry Trends

Material Upside for Mobile Ad Spend vs. Mobile Usage

Meeker/Wu May 29 2013 Internet Trends D11 Conference

KPCB

Note: *Internet advertising reached $37B in USA in 2012 per IAB, Mobile advertising reached $4B per eMarketer. Print includes newspaper and magazine. $20B opportunity calculated assuming Internet and Mobile ad spend share equal their respective time spent share. Source: Time spent and ad spend share data based on eMarketer (adjusted to exclude outdoors / classified media spend), 12/12.

80% of Top Ten Global Internet Properties 'Made in USA'… 81% of Users Outside America

Source: comScore Global, 2/13.

Global Smartphone Operating Systems 'Made in USA'… 88% Share from 5% Six Years Ago

per Gartner, 2012 data per IDC.

Social Media – Facebook Leads, YouTube + Twitter + Google+ + Pinterest + Instagram + Tumblr Rising Fast

Source: Frank N. Magid Associates, "Facebook Fatigue - Fact or Fiction?", March 2013.
Following Social Media Do You Use?"
t and Instagram data from 9/12 / 4/12.

China –
Mobile Internet Access Surpassed PC, Q2:12

Source: CNNIC, 1/13.

Alibaba – Gross Merchandise Volume Surpassed Amazon.com + eBay, Q4:12

Source: Company data.

U.S. Postal Service Mail Volume Peaked in 2006 Owing to Email Rise… Profitability Plummeted

Data not available for 1914 - 1925.

Re-Imagination of Computing Operating Systems - iOS + Android = 60% Share vs. 35% for Windows

esearch, Gartner for 2012 data.

Computing Model

Industry adopted clouds which are attractive for data analytics

5 years Cloud Computing
2 years Big Data Transformational

Gartner Priority Matrix

years to mainstream adoption

benefit

less than 2 years | 2 to 5 years | 5 to 10 years | more than 10 years

transformational

high

moderate

2012
- Media Tablets
- Big Data
- Cloud Computing
- Gesture Control
- In-Memory Database Management Systems
- 3D Printing
- Automatic Content Recognition
- Autonomous Vehicles
- Complex-Event
- 3D Bioprinting
- Human Augmentation
- Internet of Things
- Mobile Robots
- Hosted Virtual D
- Predictive Analy
- Idea Managemer

2011
- "Big Data" and Extreme Information Processing and Management
- Cloud Computing
- 3D Printing
- Context-Enriched Services
- Internet of Things
- 3D Bioprinting
- Human Augmentation
- Mobile Robots
- E-Book Readers
- Hosted Virtual D
- Location-Aware Applications
- Mobile Applicati Stores
- Predictive Analy

2010
- Cloud Computing
- Cloud/Web Platforms
- Media Tablet
- 3D Printing
- Context Delivery Architecture
- Extreme Transaction
- Autonomous Vehicles
- Human Augmentation
- Mobile Robots
- Mobile Application Stores
- Predictive Analytic
- Consumer-Generat Media
- Pen-Centric Tablet

2009
- Web 2.0
- Cloud Computing
- Internet TV
- Public Virtual Worlds
- SOA
- 3-D Printing
- Context Delivery Architecture
- RFID (Case/Pallet)
- Huma
- Mobile
- Quant
- Corporate Blog

2008
- Web 2.0
- Cloud Computing
- Public Virtual Worlds
- SOA
- 3-D
- Con
- Arch
- RFID
- Electronic Paper
- Green IT
- Location-Aware Applications
- Service-Oriented Business Applications
- Solid-State Drives
- Beh

Amazon Cloud AWS making money

- It took Amazon Web Services (AWS) eight years to hit $650 million in revenue, according to Citigroup in 2010.
- Just three years later, Macquarie Capital analyst Ben Schachter estimates that AWS will top $3.8 billion in 2013 revenue, up from $2.1 billion in 2012 (estimated), valuing the AWS business at $19 billion.
- First public cloud computing supplier building on many cloud systems used to run Amazon, Google, Bing, eBay ….

Physically Clouds are Clear

- A bunch of computers in an efficient data center with an excellent Internet connection
- They were produced to meet need of public-facing Web 2.0 e-Commerce/Social Networking sites
- They can be considered as "optimal giant data center" plus internet connection
- Note enterprises use private clouds that are giant data centers but not optimized for Internet access

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

Build giant data centers with 100,000’s of computers;
~ 200-1000 to a shipping container with Internet access

Data Centers Clouds & Economies of Scale

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

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Virtualization made several things more convenient

- Virtualization = abstraction; run a job – you know not where
- Virtualization = use hypervisor to support "images"
  - Allows you to define complete job as an "image" – OS + application
- Efficient packing of multiple applications into one server as they don't interfere (much) with each other if in different virtual machines;
- They interfere if put as two jobs in same machine as for example must have same OS and same OS services
- Also security model between VM's more robust than between processes

Microsoft Server Consolidation

- http://research.microsoft.com/pubs/78813/AJ18_EN.pdf
- Typical data center CPU has 9.75% utilization
- Take 5000 SQL servers and rehost on virtual machines with 6:1 consolidation

60% saving

The Google gmail example

- http://www.google.com/green/pdfs/google-green-computing.pdf
- Clouds win by efficient resource use and efficient data centers

Business Type	Number of users	# servers	IT Power per user	PUE (Power Usage effectiveness)	Total Power per user	Annual Energy per user
Small	50	2	8W	2.5	20W	175 kWh
Medium	500	2	1.8W	1.8	3.2W	28.4 kWh
Large	10000	12	0.54W	1.6	0.9W	7.6 kWh
Gmail (Cloud)	∞	∞	< 0.22W	1.16	< 0.25W	< 2.2 kWh

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
  - Amazon is as much PaaS as Azure
- They are **cheaper than classic clusters** unless latter 100% utilized

IaaS Hardware e.g. Server
PaaS Systems Services e.g. MapReduce, Database
SaaS Applications e.g. Recommender System, Clustering
BPaaS Particular Application Set

High Growth Expected in Cloud Infrastructure Services

	CAGR (2011-2016)
IaaS	41.7%
PaaS	26.6%
SaaS	17.4%

Source: Public Cloud Services Forecast, 2Q12 Update

Gartner

BPM = Business Process management

Research Model

4th Paradigm; From Theory to Data driven science?

http://www.wired.com/wired/issue/16-07

September 2008

The End of Science

The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. Welcome to the Petabyte Age.

The 4 paradigms of Scientific Research

1. **Theory**
2. **Experiment or Observation**
   - E.g. Newton observed apples falling to design his theory of mechanics
3. **Simulation of theory or model** Supercomputers
4. **Data-driven** (Big Data) or The Fourth Paradigm: Data-Intensive Scientific Discovery (aka Data Science)
   - http://research.microsoft.com/en-us/collaboration/fourthparadigm/ A free book
   - More data; less models

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

Data Science Process

DIKW Process

- **Data** becomes
- **Information** becomes
- **Knowledge** becomes
- **Wisdom** or **Decisions**
  - Community acceptance of results or approach important here
  - Volume of bits&bytes decreases as we proceed down DIKW pipeline

Raw Data →
Data Deluge is also Information/Knowledge/Wisdom/Decision Deluge?
Another Grid
Another Grid
Another Service
Another Grid
Database
SS
fs
Filter Service
Filter Cloud
Discovery Cloud
Inter-Service Messages
Portal
Traditional Grid with exposed services
Compute Cloud
Storage Cloud
Sensor or Data Interchange Service

Example of Google Maps/Navigation

- **Data** comes from traditional maps (US Geological Survey), Satellites (overlays) and street cams
- **Information** is presented by basic Google Maps web page
- **Knowledge** is a particular optimized route
- **Decisions** (**Wisdom**) comes from deciding to drive a particular route

Physics-Informatics
Looking for Higgs Particle
with Large Hadron Collider LHC

http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf

This analysis raw data → reconstructed data → AOD and TAGS → Physics is performed on the multi-tier LHC Computing Grid. Note that every event can be analyzed independently so that many events can be processed in parallel with some concentration operations such as those to gather entries in a histogram. This implies that both Grid and Cloud solutions work with this type of data with currently Grids being the only implementation today.

ATLAS Expt

Note LHC lies in a tunnel 27 kilometres (17 mi) in circumference

Higgs Event

The LHC produces some 15 petabytes of data per year of all varieties and with the exact value depending on duty factor of accelerator (which is reduced simply to cut electricity cost but also due to malfunction of one or more of the many complex systems) and experiments. The raw data produced by experiments is processed on the LHC Computing Grid, which has some 200,000 Cores arranged in a three level structure. Tier-0 is CERN itself, Tier 1 are national facilities and Tier 2 are regional systems. For example one LHC experiment (CMS) has 7 Tier-1 and 50 Tier-2 facilities.

The inside of the RHIC (Relativistic Heavy Ion Collider) tunnel, a 2.4-mile high-tech particle racetrack at Brookhaven National Laboratory.

http://www.interactions.org/cms/?pid=1032811

http://www.quantumdiaries.org/2012/09/07/why-particle-detectors-need-a-trigger/atlasmgg/

Geoffrey Fox

Professor of Informatics and Computing and Physics

Cyberinfrastructure - Distributed Systems - Clouds - Parallel Computing - Particle Physics

Verified email at indiana.edu

Homepage

Google scholar

Search Authors

My Citations - Help

« Back to list | Edit | Export | Delete

Title	Observables for the Analysis of Event Shapes in e^{+} e^{-} Annihilation and Other Processes	[PDF] from stephenwolfram.com IU-Link
Authors	Geoffrey C Fox, Stephen Wolfram	
Publication date	1978/12/4	
Journal name	Physical Review Letters	
Volume	41	
Issue	23	
Pages	1581-1585	
Publisher	American Physical Society	
Description	FIG. 1. The distributions in H2 and. 03 for the processes e+ e qq (dotted lines), e+ e—qqG (full lines), and e+ e heavy resonance—GGG (dashed lines). The process e+ e+ qqG alone yields an infinite total cross section, but when added to e+ e—qq calculated through 0 (g) the combination of processes [denoted by e+ e qq (G)] gives a finite cross section. We have taken= 0. 25 for the e+ e-qqG distribution.	

Total citations: Cited by 1143

Citations per year

Scholar articles

Observables for the Analysis of Event Shapes in e^{+} e^{-} Annihilation and Other Processes
GC Fox, S Wolfram - Physical Review Letters, 1978
Cited by 1143 - Related articles - All 6 versions

☑ Show thumbnails in outline
Table 1
4. H→ZZ(□)→4ℓ channel
4.1. Event selection
4.2. Background estimation
Table 2

4.3. Systematic uncertainties
4.4. Results

Table 3

5. H→γγ channel
5.1. Event selection
5.2. Invariant mass reconstruction
5.3. Event categorisation
Table 4
5.4. Signal modelling
5.5. Background modelling
5.6. Systematic uncertainties
5.7. Results

6. H→WW(□)→eνμν channel
6.1. Event selection
6.2. Background normalisation and

ELSEVIER

Physics Letters B

Volume 716, Issue 1, 17 September 2012, Pages 1–29

http://www.sciencedirect.com/science/article/pii/S037026931200857X

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC ☆

Universally Available

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

ATLAS Collaboration *

G. Aad[48], T. Abajyan[21], B. Abbott[111], J. Abdallah[12], S. Abdel Khalek[115], A.A. Abdelalim[49], O. Abdinov[11], R. Aben[105], B. Abi[112], M. Abolins[88], O.S. AbouZeid[158], H. Abramowicz[153], H. Abreu[136], B.S. Acharya[164a, 164b], L. Adamczyk[38], D.L. Adams[25], T.N. Addy[56], J. Adelman[176], S. Adomeit[98], P. Adragna[75], T. Adye[129], S. Aefsky[23], J.A. Aguilar-Saavedra[124b, a], M. Agustoni[17], M. Aharrouche[81], S.P. Ahlen[22], F. Ahles[48], A. Ahmad[148], M. Ahsan[41], G. Aielli[133a, 133b], T. Akdogan[19a], T.P.A. Åkesson[79], G. Akimoto[155], A.V. Akimov[94], M.S. Alam[2], M.A. Alam[76], J. Albert[169], S. Albrand[55], M. Aleksa[30], I.N. Aleksandrov[64], F. Alessandria[89a], C. Alexa[26a], G. Alexander[153], G. Alexandre[49], T. Alexopoulos[10], M. Alhroob[164a, 164c], M. Aliev[16], G. Alimonti[89a], J. Alison[120], B.M.M. Allbrooke[18], P.P. Allport[73], S.E. Allwood-Spiers[53], J. Almond[82], A. Aloisio[102a, 102b], R. Alon[172], A. Alonso[79], F. Alonso[70], A. Altheimer[35], B. Alvarez Gonzalez[88], M.G. Alviggi[102a, 102b], K. Amako[65], C. Amelung[23], V.V. Ammosov[128, □], S.P. Amor Dos Santos[124a], A. Amorim[124a, b], N. Amram[153], C. Anastopoulos[30], L.S. Ancu[17], N. Andari[115], T. Andeen[35], C.F. Anders[58b], G. Anders[58a], K.J. Anderson[31], A. Andreazza[89a, 89b], V. Andrei[58a], M.-L. Andrieux[55], X.S. Anduaga[70], S. Angelidakis[9], P. Anger[44], A. Angerami[35], F. Anghinolfi[30], A. Anisenkov[107], N. Anjos[124a], A. Annovi[47], A. Antonaki[9], M. Antonelli[47], A. Antonov[96], J. Antos[144b], F. Anulli[132a], M. Aoki[101], S. Aoun[83], L. Aperio Bella[5], R. Apolle[118, c], G. Arabidze[88], I. Aracena[143], Y. Arai[65], A.T.H. Arce[45], S. Arfaoui[148], J.-F. Arguin[93], E. Arik[19a, □], M. Arik[19a], A.J. Armbruster[87], O. Arnaez[81], V. Arnal[80], C. Arnault[115], A. Artamonov[95], G. Artoni[132a, 132b], D. Arutinov[21], S. Asai[155], S. Ask[28], B. Åsman[146a, 146b], L. Asquith[6], K. Assamagan[25], A. Astbury[169], M. Atkinson[165], B. Aubert[5], E. Auge[115], K. Augsten[127], M. Aurousseau[145a], G. Avolio[163], R. Avramidou[10], D. Axen[168], G. Azuelos[93, d], Y. Azuma[155], M.A. Baak[30], G. Baccaglioni[89a], C. Bacci[134a, 134b], A.M. Bach[15], H. Bachacou[136], K. Bachas[30], M. Backes[49], M. Backhaus[21], J. Backus Mayes[143], E. Badescu[26a], P. Bagnaia[132a, 132b], S. Bahinipati[3], Y. Bai[33a], D.C. Bailey[158], T. Bain[158], J.T. Baines[129], O.K. Baker[176], M.D. Baker[25], S. Baker[77], P. Balek[126], E. Banas[39], P. Banerjee[93], Sw. Banerjee[173], D. Banfi[30], A. Bangert[150], V. Bansal[169], H.S. Bansil[18], L. Barak[172], S.P. Baranov[94], A. Barbaro Galtieri[15], T. Barber[48], E.L. Barberio[86], D. Barberis[50a, 50b], M. Barbero[21], D.Y. Bardin[64], T. Barillari[99], M. Barisonzi[175], T. Barklow[143], N. Barlow[28], B.M. Barnett[129], R.M. Barnett[15], A. Baroncelli[134a], G. Barone[49], A.J. Barr[118], F. Barreiro[80], J. Barreiro Guimarães da Costa[57], P. Barrillon[115], R. Bartoldus[143], A.E. Barton[71], V. Bartsch[149], A. Basye[165], R.L. Bates[53], L. Batkova[144a], J.R. Batley[28], A. Battaglia[17], M. Battistin[30], F. Bauer[136], H.S.

Recommender Systems

Overview of Many Informatics Areas

- In many cases, one needs personalized matching of items to people or perhaps collections of items to collections of people
- **People** to **products**: Online and Offline Commerce
- **People** to **People**: Social Networking
- **People** to **Jobs** or **Employers**: Job Sites
- **People+Queries** to the **Web**: Information Retrieval (search as in Bing/Google)

Recommender Systems in more detail

- A large number of online and offline commerce activities plus basic Internet site personalization relies on "**recommender systems**"
- Given real-time action by user, immediately suggest new actions (as in Amazon **buy recommendations** on web)
- Based on past actions of users (and others) suggest **movies** to **look** at, **restaurants** to eat at, **events** to go to, books and music to **buy**
- Based on mix of explicit user choice and grouping of internet sites, present customized **Google News** page
- Given sales statistics, decide on **discounts** at "real" **supermarkets** and **placement** of related (by analysis of buying habits) products
- Identify possible colleagues at Social Networking sites like **LinkedIn**
- Identify matches between employers and employees at sites like **CareerBuilder** and **Monster**

Everything is an Optimization Problem?

- **Fit** Model to Data
  - Higgs + Background
- **Match** User to Jobs or Books or Other Users?
- **Classification** is optimizing assignment of members of an ontology (list of categories) to data
- Typically minimize **some function** (or maximize negative of function)
  - Interesting feature of these problems is ingenious choice of function
  - Note Physics minimizes (free) energy
- Often involves thinking of people and/or items as **points in a space** (not always a traditional vector space)
  - Space called "**bag**" in "bag of words" model for information retrieval

http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial

Netflix on Personalization

Everything is personalized

Note: Recommendations are per household, not individual user

http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial

Netflix on Recommendations

Anatomy of a Personalization - Recap

- Everything is a recommendation: not only rating prediction, but also ranking, row selection, similarity…
- We strive to make it easy for the user, but…
- We want the user to be aware and be involved in the recommendation process
- Deal with implicit/explicit and hybrid feedback
- Add support/explanations for recommendations
- Consider issues such as diversity or freshness

NETFLIX

http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial

April 2013: The last two quarters have each brought more than 2 million new streaming subscriber signups. That gives Netflix a current total of nearly 29.2 million subscribers

Big Data @Netflix

- 25M+ subscribers
- Ratings: 4M/day
- Searches: 3M/day
- Plays: 30M/day
- 2B hours streamed in Q4 2011
- 1B hours in June 2012

NETFLIX

Netflix competition

http://www.ifi.uzh.ch/ce/teaching/spring2012/16-Recommender-

june 26, 2009 around three years after launch of netflix prize. BellKor Pragmatic Chaos (merge of three teams) broke the 10% barrier in improving accuracy on movie prediction. NetFlix data set: 100 million + ratings, 1.4 million used as test set. 1-5 stars. RMSE on test set.

http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial

Netflix on Data Science

Consumer (Data) Science

1. Start with a hypothesis:
   - *Algorithm/feature/design X will increase member engagement with our service, and ultimately member retention*
2. Design a test
   - Develop a solution or prototype
   - Think about dependent & independent variables, control, significance…
3. Execute the test
4. Let data speak for itself

Note Netflix and others run tests all the time on subsets of customers

NETFLIX

Distances in Funny Spaces I

- In **user-based collaborative filtering**, we can think of users in a space of dimension N where there are N items and M users.
  - Let *i* run over items and *u* over users
- Then each **user** is represented as a **vector $U_i(u)$** in "item-space" where ratings are vector components. We are looking for users u u' that are near each other in this space as measured by some distance between $U_i(u)$ and $U_i(u')$
- If u and u' rate all items then these are "real" vectors but almost always they each only rates a small fraction of items and the number in common is even smaller
- The "**Pearson coefficient**" is just one distance measure that can be used
  - Only sum over i rated by u and u'

Last.fm uses this for songs as does Amazon, Netflix

Distances in Funny Spaces II

- In item-based collaborative filtering, we can think of items in a space of dimension M where there are N items and M users.
  - Let *i* run over items and *u* over users
- Then each **item** is represented as a **vector $R_u(i)$** in "user-space" where ratings are vector components. We are looking for items i i' that are near each other in this space as measured by some distance between $R_u(i)$ and $R_u(i')$
- If i and i' rated by all users then these are "real" vectors but almost always they are each only rated by a small fraction of users and the number in common is even smaller
- The "**Cosine measure**" is just one distance measure that can be used
  - Only sum over users u rating both i and i'

Distances in Funny Spaces III

- In content based recommender systems, we can think of items in a space of dimension M where there are N items and M properties.
  - Let *i* run over items and *p* over properties
- Then each **item** is represented as a **vector $P_p(i)$** in "property-space" where values of properties are vector components. We are looking for items i i' that are near each other in this space as measured by some distance between $P_p(i)$ and $R_p(i')$
- Properties could be "reading age" or "character highlighted" or "author" for books
- Properties can be genre or artist for songs and video
- Properties can characterize pixel structure for images used in face recognition, driving etc.

Pandora uses this for songs (Music Genome) as does Amazon, Netflix

Do we need “real” spaces?

- Much of (eCommerce/LifeStyle) Informatics involves “points”
  - Events in LHC analysis
  - Users (people) or items (books, jobs, music, other people)
- These points can be thought of being in a “space” or “bag”
  - Set of all books
  - Set of all physics reactions
  - Set of all Internet users
- However as in recommender systems where a given user only rates some items, we don’t know “full position”
- However we can nearly always define a useful **distance d(a,b)** between points
- Always **d(a,b) >= 0**
- Usually **d(a,b) = d(b,a)**
- Rarely d(a,b) + d(b,c) >= d(a,c) **Triangle Inequality**

Using Distances

- The simplest way to use distances is "**nearest neighbor algorithms**" – given one point, find a set of points near it – cut off by number of identified nearby points and/or distance to initial point
  - Here point is either user or item
- Another approach is divide space into regions (topics, latent factors) consisting of nearby points
  - This is **clustering**
  - Also other algorithms like **Gaussian mixture models** or **Latent Semantic Analysis** or **Latent Dirichlet Allocation** which use a more sophisticated model

Web Search
Information Retrieval

"Web Data Analytics"

- Get the digital data (from web or from scanning)
  - Need to crawl web (? Solved "engineering" problem)
- Preprocess data to get searchable things (words positions)
- Form **Inverted Index** mapping words to documents
- Typically use **TF-IDF** (term frequency, Inverse Document frequency) to quantify importance of word match
- Rank relevance of documents: **PageRank**
- Lots of technology for advertising, "reverse engineering" "preventing reverse engineering"
- Clustering of documents into topics (as in Google News)

PageRank

Size of face proportional to PageRank

Modern Recommendation Systems (from Yahoo)

- Goal (Function to Optimize – Long Term dollars)
  - Serve the right item to a user in a given context to optimize long-term business objectives
- A scientific discipline that involves
  - Large scale Machine Learning & Statistics
    - Offline Models (capture global & stable characteristics)
    - Online Models (incorporates dynamic components)
    - Explore/Exploit (active and adaptive experimentation)
  - Multi-Objective Optimization
    - Click-rates (CTR), Engagement, advertising revenue, diversity, etc
  - Inferring user interest
    - Constructing User Profiles
  - Natural Language Processing to understand content
    - Topics, “aboutness”, entities, follow-up of something, breaking news,…

Recommend search queries

Recommend packages:
- Image
- Title, summary
- Links to other pages

Pick 4 out of a pool of K
- $K = 20 \sim 40$
- Dynamic

Routes traffic other pages

Recommend applications

Recommend news article

Some examples from content optimization

- Simple version
  - I have a content module on my page, content inventory is obtained from a third party source which is further refined through editorial oversight. Can I algorithmically recommend content on this module? I want to improve overall click-rate (CTR) on this module
- More advanced
  - I got X% lift in CTR. But I have additional information on other downstream utilities (e.g. advertising revenue). Can I increase downstream utility without losing too many clicks?
- Highly advanced
  - There are multiple modules running on my webpage. How do I perform a simultaneous optimization?

Cloud Applications in Research

Science Computing Environments

- **Large Scale Supercomputers** – Multicore nodes linked by high performance low latency network
  - Increasingly with GPU enhancement
  - Suitable for highly parallel simulations
- **High Throughput Systems** such as European Grid Initiative EGI or Open Science Grid OSG typically aimed at pleasingly parallel jobs
  - Can use "cycle stealing"
  - Classic example is **LHC data analysis**
- **Grids** federate resources as in EGI/OSG or enable convenient access to multiple backend systems including supercomputers
- Use **Services** (**SaaS**)
  - **Portals** make access convenient and
  - **Workflow** integrates multiple processes into a single job

Clouds HPC and Grids

- Synchronization/communication Performance
  Grids > Clouds > Classic HPC Systems
- **Clouds** naturally execute effectively **Grid** workloads but are less clear for closely coupled HPC applications
- **Classic HPC machines** as MPI engines offer highest possible performance on closely coupled problems
- The 4 forms of MapReduce/MPI
  1) **Map Only** – pleasingly parallel
  2) **Classic MapReduce** as in Hadoop; single Map followed by reduction with fault tolerant use of disk
  3) **Iterative MapReduce** use for data mining such as Expectation Maximization in clustering etc.; Cache data in memory between iterations and support the large collective communication (Reduce, Scatter, Gather, Multicast) use in data mining
  4) **Classic MPI!** Support small point to point messaging efficiently as used in partial differential equation solvers

What Applications work in Clouds

- **Pleasingly (moving to modestly) parallel** applications of all sorts with roughly independent data or spawning independent simulations
  - **Long tail** of science and integration of distributed sensors
- **Commercial and Science Data analytics** that can use MapReduce (some of such apps) or its **iterative** variants (most other data analytics apps)
- **Which science applications are using clouds**?
  - **Venus-C** (Azure in Europe): 27 applications **not using** Scheduler, Workflow or MapReduce (except roll your own)
  - 50% of applications on **FutureGrid** are from Life Science
  - Locally **Lilly** corporation is commercial cloud user (for drug discovery) but not IU Biology
- **But overall very little science use of clouds yet**

Parallelism over Users and Usages

- "**Long tail of science**" can be an important usage mode of clouds.
- In some areas like particle physics and astronomy, i.e. "**big science**", there are just a few major instruments generating now petascale data driving discovery in a coordinated fashion.
- In other areas such as genomics and environmental science, there are many **"individual" researchers** with distributed collection and analysis of data whose total data and processing needs can match the size of big science.
- **Clouds** can provide scaling convenient resources for this important aspect of science.
- Can be **map only** use of MapReduce if different usages naturally linked e.g. exploring docking of multiple chemicals or alignment of multiple DNA sequences
  - Collecting together or summarizing multiple "maps" is a **simple Reduction**

Internet of Things and the Cloud

- It is projected that there will be **24 billion devices** on the Internet by 2020. Most will be small sensors that send streams of information into the cloud where it will be processed and integrated with other streams and turned into knowledge that will help our lives in a multitude of small and big ways.
- The **cloud** will become increasing important as a controller of and **resource provider for the Internet of Things.**
- As well as today's use for smart phone and gaming console support, "Intelligent River" "smart homes and grid" and "ubiquitous cities" build on this vision and we could expect a growth in cloud supported/controlled **robotics**.
- Some of these "things" will be supporting science
- Natural parallelism over "things"
- "Things" are distributed and so form a Grid

Sensors (Things) as a Service

Output Sensor

A larger sensor .....

Sensor Processing as a Service (could use MapReduce)

https://sites.google.com/site/opensourceiotcloud/ Open Source Sensor (IoT) Cloud

A Car or a Computer on Four Wheels?

It's a Bird, It's a Plane, It's a Mini-Drone…

Use Cases of Low-Cost Drones

Agriculture

GPS-Enabled Photos Help Pinpoint Potential Crop Damage Early On

Sports / Entertainment

Unique Angle of View Provides Insights for Training + Broadcast Audience

Public Safety / Disaster Relief*

Providing Aerial Video Coverage / First-Aid / Other Supplies in Challenging Conditions

· Drone carrying cameras + other
thquake in Lushan, China (4/13).

QR Codes =
Scan & Be Scanned to Get Stuff…

SCAN

Quick Scan w/ Smartphone For Info on Nutrition / Product / Price…

BE SCANNED

Smartphone-Generated Codes For Boarding Pass / Ticket / Payment / Rewards…

Parallel Computing and MapReduce

Classic Parallel Computing

- **HPC:** Typically SPMD (Single Program Multiple Data) "maps" typically processing particles or mesh points interspersed with multitude of low latency messages supported by specialized networks such as Infiniband and technologies like MPI
  - Often run large capability jobs with 100K (going to 1.5M) cores on same job
  - National DoE/NSF/NASA facilities run 100% utilization
  - Fault fragile and cannot tolerate "outlier maps" taking longer than others
- **Clouds:** MapReduce has asynchronous maps typically processing data points with results saved to disk. Final reduce phase integrates results from different maps
  - Fault tolerant and does not require map synchronization
  - **Map only** useful special case
- **HPC + Clouds:** Iterative MapReduce caches results between "MapReduce" steps and supports SPMD parallel computing with large messages as seen in parallel kernels (linear algebra) in clustering and other data mining

MapReduce "File/Data Repository" Parallelism

Map = (data parallel) computation reading and writing data
Reduce = Collective/Consolidation phase e.g. forming multiple global sums as in histogram

Sam's Problem

http://www.slideshare.net/esaliya/mapreduce-in-simple-terms

- Sam thought of "drinking" the apple

- He used a to cut the and a to make juice.

Creative Sam

- Implemented a *parallel* version of his innovation

Each input to a map is a list of <key, value> pairs

A *list of <key, value>* pairs mapped into another *list of <key, value>* pairs which gets grouped by the key and reduced into a *list of values*

(<a, >, <o, >, <p, >, ...)

Each output of slice is a list of <key, value> pairs

(<a', > , <o', > , <p', >)

Grouped by key

The idea of Map Reduce in Data Intensive Computing

E ossibly a
li ing
mechanism)
e.g. <ao, (...)>

Reduced into a list of values

Data Science Education

Opportunities at Universities

see recent New York Times articles

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

Data Science Education

- Broad Range of Topics from Policy to curation to applications and algorithms, programming models, data systems, statistics, and broad range of CS subjects such as Clouds, Programming, HCI,
- Plenty of Jobs and broader range of possibilities than computational science but similar cosmic issues
  - What type of degree (Certificate, minor, track, "real" degree)
  - What implementation (department, interdisciplinary group supporting education and research program)

At Indiana University

- Have a proposal to set up certificates and Masters degree in **data science**
- Joint between 3 units in School of Informatics and Computing: Computer Science, Informatics, Information & Library Science, and Statistics department in COAS College
- Looking at version with Kelley with **Business data analytics** flavor
- Attractive to offer online as few universities have this and so a potentially large audience outside IU

Online Education Growing Rapidly =
32% of Students Taking at Least One Online Course, 2011

son Survey Research Group 1/13.

Education Being Democratized =
Fast Global MOOC (Massively Open Online Course) User Growth

& Coursera Company Data, 4/13.

Massive Open Online Courses (MOOC)

- MOOC's are very "hot" these days with Udacity and Coursera as start-ups; perhaps over 100,000 participants
- Relevant to Data Science as this is a new field with few courses at most universities
- Typical model is collection of short prerecorded segments (talking head over PowerPoint) of length 3-15 minutes
  - This is Boredom limit http://blog.coursera.org/post/49750392396/on-the-topic-of-boredom
- These "lesson objects" can be viewed as "songs"
- Google Course Builder (python open source) builds customizable MOOC's as "playlists" of "songs"
- **Tells you to capture all material as "lesson objects"**
- **We are aiming to build a repository of many "songs"; used in many ways – tutorials, classes …**

X-Informatics MOOC

Prof. Geoffrey Fox

http://x-informatics.appspot.com/course

Home Course FAQ Enroll

Overview

Prof. Geoffrey Fox examines how the explosion in worldwide data has spurred the need for Cloud computing, as well as transforming how industry and science operate.

Enroll *a Google account is required

Complete end of July

Instructors

Geoffrey Fox

Geoffrey Fox received a Ph.D. in Theoretical Physics from Cambridge University and is now professor of Informatics and Computing, and Physics at Indiana University where he is director of the Digital Science Center and Associate Dean for Research and Graduate Studies at the School of Informatics and Computing. He previously held positions at Caltech, Syracuse University and Florida State University. He has supervised the PhD of 65 students and published around 1000 papers in physics and computer science

Syllabus

- Unit 1 - Data Deluge Available Now
- Unit 2 - Scientific Method Available Now
- Unit 3 - Cloudburst Available Now
- Unit 4 - Leading the Charge Available Now
- Unit 5 - The Elusive Mr. Higgs Available Now
- Unit 6 - Number Theory Available Now
- Unit 7 - A Calculated Dice Roll Available Now

X-Informatics MOOC

Prof. Geoffrey Fox

Home Course FAQ Announcements My Profile gcfexchange@gmail.com | Logout Dashboard Admin

Course › Unit 2 › Lesson 1

Unit 2 - Scientific Method

2.1 Overview

2.2 Big Business

2.3 Every Day, in Every Way

2.4 Particle Pursuit Part I

2.5 Particle Pursuit Part II

2.6 The 4th Paradigm

2.7 Internet of Things

A brief overview of the course up 'til now is followed by a lesson in which several notable examples of Big Data producers, focusing on Ebay, are illustrated. The example then given uses Ebay to demonstrate how data collected is used to better tailor results from search queries. A look at data generated in the search for the Higgs-Boson particle follows, and the three classic paradigms of Scientific Research have a new member in the form of Data Science. The last lesson emphasizes how everyday devices are being used to create a wealth of data.

Next Page

- Seven ~10 minutes lesson objects in this lecture
- IU wants us to close caption if use in real course

Customizable MOOC's I

- We could teach one class to 100,000 students or 2,000 classes to 50 students
- The 2,000 class choice has 2 useful features
  - One can use the usual (electronic) mentoring/grading technology
  - One can customize each of 2,000 classes for a particular audience given their level and interests
  - One can even allow student to customize – that's what one does in making play lists in iTunes
- Both models can be supported by a repository of lesson objects (10-15 minute video segments) in the cloud
- The teacher can choose from existing lesson objects and add their own to produce a new customized course with new lessons contributed back to repository

Science Cloud MOOC Repository

http://iucloudsummerschool.appspot.com/preview

Unit ~1 hour with ~6 lessons, Total 115 lesson objects

Unit 1 - Introduction	Geoffrey Fox
Unit 2 - Biology in the Clouds	Michael Schatz
Unit 3 - Infrastructure Used: FutureGrid	Geoffrey Fox
Unit 4 - Virtualization on HPC	Thomas J. Hacker
Unit 5 - Running MapReduce in Non-Traditional Environments	Abhishek Chandra
Unit 6 - Virtual Clusters Supporting MapReduce in Cloud	Jonathan Klinginsmith
Unit 7 - Hadoop and HDFS	Jerome Mitchell
Unit 8 - HBase and Bigtable Storage	Thilina Gunarathne
Unit 9 - Data Mining with Twister Iterative MapReduce	Hui Li
Unit 10 - Twister Introduction	Judy Qiu
Unit 11 - Commercial IaaS/PaaS I: AWS: The Platform for Data Science	Jamie Kinney
Unit 12 - Commercial IaaS/PaaS II: Azure and Twister4Azure	Thilina Gunarathne
Unit 13 - Federating HPC, Cyberinfrastructure and Clouds using CometCloud I	Manish Parashar
Unit 14 - Scientific Workflows in the Cloud	Gideon Juve
Unit 15 - Cloud Technology: Virtual Private Clusters: Virtual Appliances and Networks in the Cloud	Renato Figueiredo
Unit 16 - Clouds and the Network	Martin Swany
Unit 17 - Applications of Cloud: DOE Systems Biology Knowledgebase	Rick Stevens
Unit 18 - Cloud Technology: Cloud Security: New Challenges and New Opportunities	XiaoFeng Wang
Unit 19 - Magellan: Evaluating Cloud Computing for Science	Lavanya Ramakrishnan
Unit 20 - Cloud Technology: GPU on Clouds	Andrew J. Younge

Customizable MOOC's II

- The 3-15 minute Video over PowerPoint of MOOC lesson object's is easy to re-use
- Qiu (IU)and Hayden (ECSU Elizabeth City State University – (a small HBCU Historically Black University) will customize a module
  - Starting with Qiu's cloud computing course at IU
  - Adding material on use of Cloud Computing in Remote Sensing (area covered by ECSU course)
- This is a model for adding cloud curricula material to wide set of universities where faculty not able to teach
- Defining how to support computing labs associated with MOOC's with clouds or VM's on clients
  - Appliances scale as download to student's client

Can of course build many different interfaces

Songs stored on YouTube

Songs prepared with Adobe Presenter on Laptop

http://cloudmooc.soic.indiana.edu/

Two limits where MOOC's are Compelling

- High volume courses (CS/Ph/Chem/Bio101…) where scalability of MOOC's make them attractive to reach a lot of students
- Niche areas where there is some student interest but either no faculty expertise or not enough students to justify traditional courses
  - Offer to many institutions simultaneously

Hermit's Cave Virtual University

- I proposed in 1999 misjudging pace of online education: One can place faculty in their favorite location (e.g. remote cave) and universities provide structure to give "credentials" while faculty teach
  - Maybe some populate repository; others actually deliver courses
- Note **Google Course Builder** has no need for local resources except for faculty client (laptop) – entire course stored in the cloud
- So we can radically change university system with a major cross institution virtual education and as well research component
- Enabled by cloud plus high performance reliable networking
- Lets set up community to define and build the virtual university MOOC repository and other needed activities

https:/

Mentoring / Grading

- We should aim at simplicity; attractive at moment is a mix of multi-topic forum plus more interactive Hangouts or equivalent (5-12 people)

https://www.youtube.com/watch?v=M3jcSCA9_hM

Re-Imagination of "Learning Tools" = Twitter / YouTube / Google Docs / Google / WordPress

- 582 learning professionals worldwide were asked to provide their top 10 tools used for learning in 2012 / 2009
- 'Learning tool' is defined as any tool that you could use to create or deliver learning content/solutions for others, or a tool you use for your own personal learning
- The 2012 survey compiled data from 582 learning professionals worldwide (55% working in education & 45% in non-educational organizations

C4LPT Ranking of Top Learning Tools

Company	2012 Rank	2009 Rank
Twitter	1	1
YouTube	2	3
Google Docs*	3	5
Google Search	4	8
WordPress	5	6
Dropbox	6	71
Skype	7	11
Powerpoint	8	13
Facebook	9	31
Wikipedia	10	17
Moodle	11	14
Evernote	12	27
Slideshare	13	7
Prezi	14	28
Blogger / BlogSpot	15	14

for Learning & Performance Technology (C4LPT) Top 100 Tools For Learning 2012.

Many Synergies Clouds and Universities

- Use clouds in faculty, graduate student and undergraduate research
  - Clouds for Scientific data analysis
  - Core cloud technologies (MapReduce, Virtualization)
- Teach clouds as it involves areas of Information Technology with lots of job opportunities
- Use clouds to support distributed learning and research environment
  - A cloud backend for course materials and collaboration as in MOOC repository
- Green environmentally friendly computing infrastructure

Conclusions

Conclusions

- **Clouds** are here to stay and one should plan on exploiting them
- **Data Intensive studies in business and research** continue to grow in importance
  - **Data Analytics:** Everything is an optimization problem in a funny space
- Growing **employment opportunities** in clouds and data related activities and so popular with students
  - Enabling many of the most important companies from Facebook/Google to General Electric
- Need community discussion of **data science education**
  - Agree on curricula; is such a degree attractive?
- **MOOC**'s interesting for
  - Disseminating new curricula
  - Managing course fragments that can be assembled into custom courses for particular interdisciplinary students

Big Data Ecosystem in One Sentence

Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics educated in data science

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)