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

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

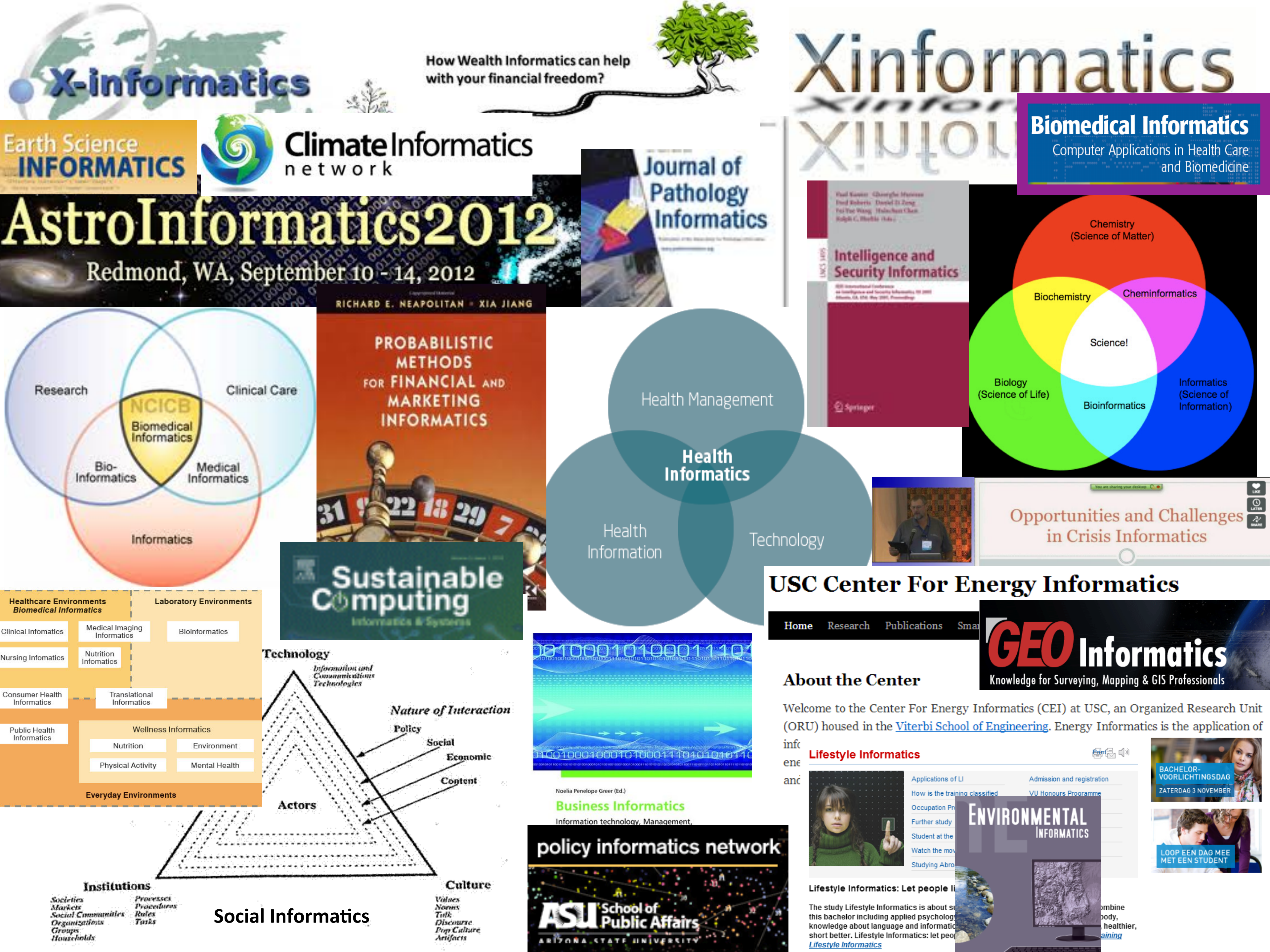
Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X)

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/

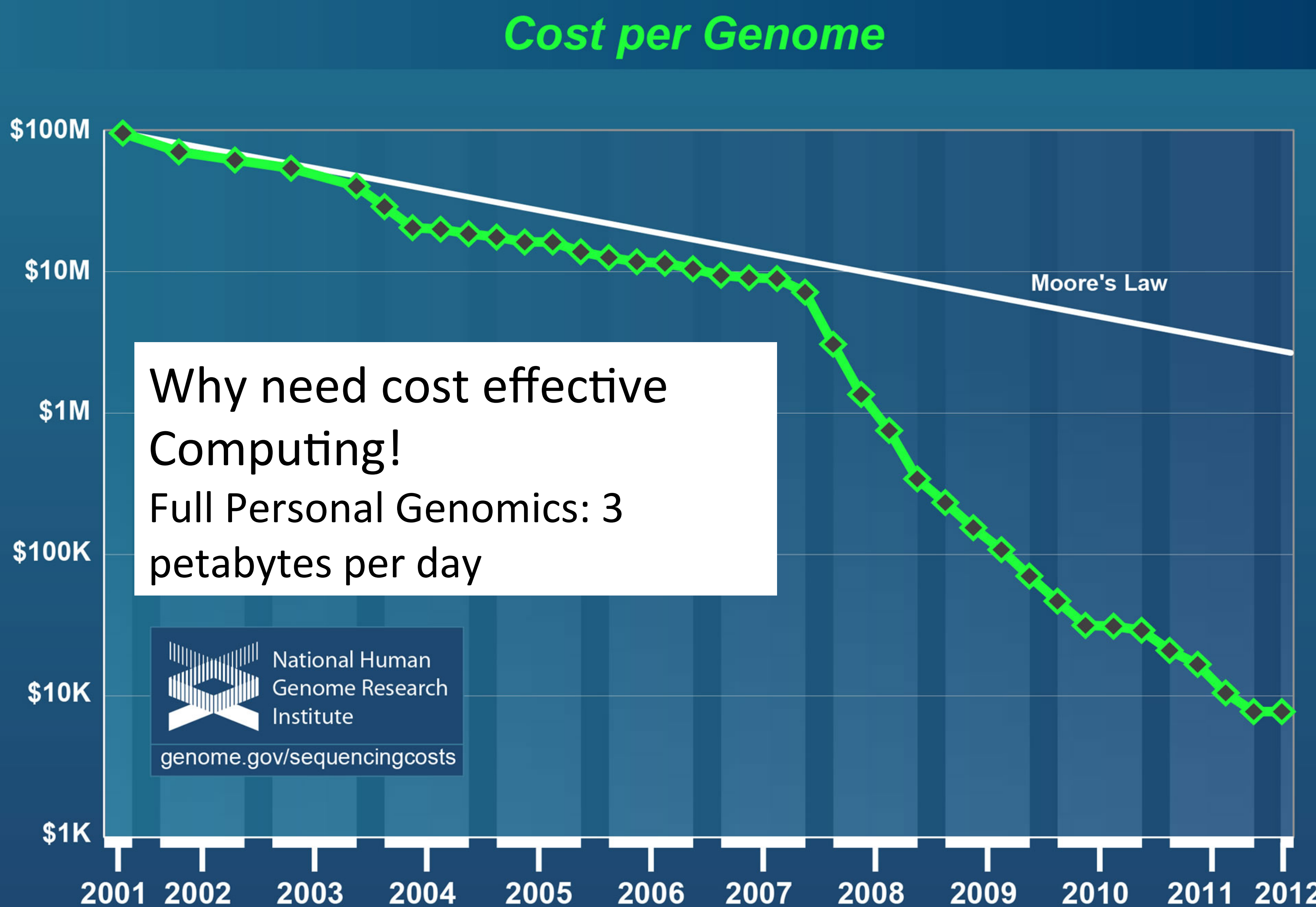


X-informatics
How Wealth Informatics can help with your financial freedom?
Xinformatics
Earth Science INFORMATICS
Climate Informatics network
Journal of Pathology Informatics
Biomedical Informatics
Computer Applications in Health Care and Biomedicine
AstroInformatics2012
Redmond, WA, September 10 - 14, 2012
Intelligence and Security Informatics
Springer
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
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
Sustainable Computing
Informatics & Systems
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
Admission and registration
VU Honours Programme
Further study
Lifestyle Informatics: Let people li
Lifestyle Informatics
ENVIRONMENTAL INFORMATICS
BACHELOR-VOORLICHTINGSDAG
ZATERDAG 3 NOVEMBER
LOOP EEN DAG MEE MET EEN STUDENT

Data Deluge
Science & Research

The data deluge: The Economist Feb 25 2010 http://www.economist.com/node/15579717

- WHEN the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. Now, a decade later, its archive contains a whopping 140 terabytes of information. A successor, the Large Synoptic Survey Telescope, due to come on stream in Chile in 2016, will acquire that quantity of data every five days.
- Such astronomical amounts of information can be found closer to Earth too. Wal-Mart, a retail giant, handles more than 1m customer transactions every hour, feeding databases estimated at more than 2.5 petabytes—the equivalent of 167 times the books in America's Library of Congress (see article for an explanation of how data are quantified).
- Facebook, a social-networking website, is home to 40 billion photos. And decoding the human genome involves analysing 3 billion base pairs—which took ten years the first time it was done, in 2003, but can now be achieved in one week.



Why need cost effective Computing!

Full Personal Genomics: 3 petabytes per day

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

Ninety-six percent of radiology practices in the USA are filmless and Table below illustrates the annual volume of data across the types of diagnostic imaging; this does not include cardiology which would take the total to over 10^9 GB (an Exabyte).

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

Modality	Part B non HMO	All Medicare	All Population	Per 1000 persons	Ave study size (GB)	Total annual data generated in GB
CT	22 million	29 million	87 million	287	0.25	21,750,000
MR	7 million	9 million	26 million	86	0.2	5,200,000
Ultrasound	40 million	53 million	159 million	522	0.1	15,900,000
Interventional	10 million	13 million	40 million	131	0.2	8,000,000
Nuclear Medicine	10 million	14 million	41 million	135	0.1	4,100,000
PET	1 million	1 million	2 million	8	0.1	200,000
Xray, total incl. mammography	84 million	111 million	332 million	1,091	0.04	13,280,000
All Diagnostic Radiology	174 million	229 million	687 million	2,259	0.1	68,700,000 **68.7 PETAbytes**

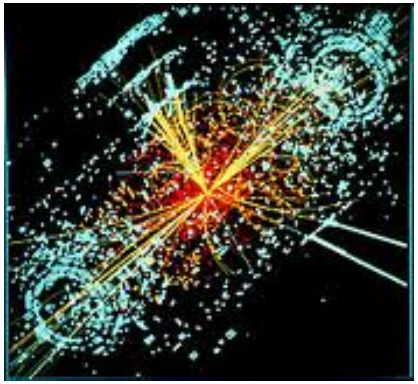
http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pd

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

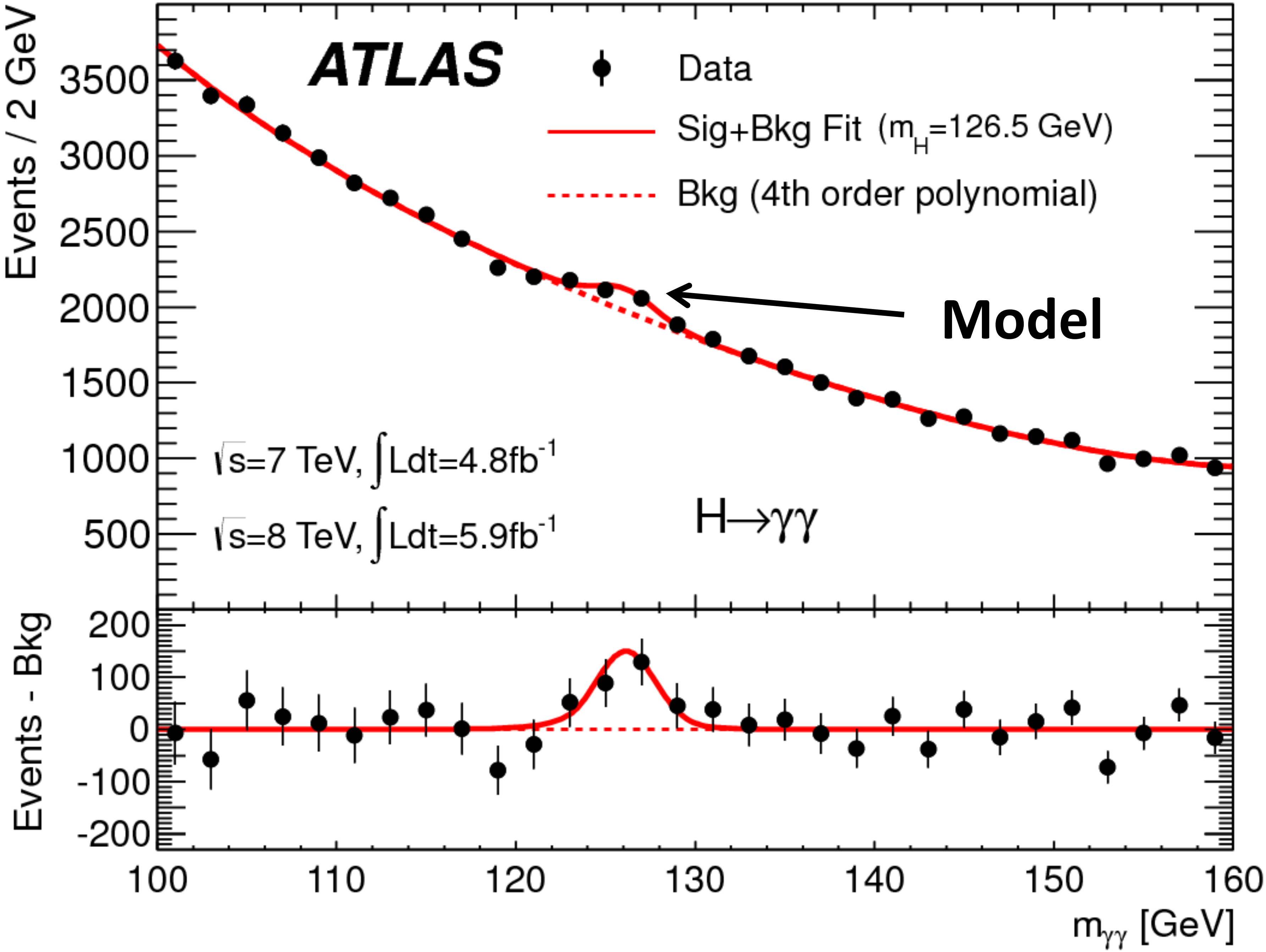
Note LHC lies in a tunnel 27 kilometres (17 mi) in circumference

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.

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



```
      SUBROUTINE ZERHPS(ICOM,IA)                                        ZERHPS   2
C     ZERO OUT HISTOGRAM SUM OR SCATTERPLOT STORED IN IA(1..INO)        ZERHPS   3
C     ICOM IS COMMAND NUMBER                                            ZERHPS   4
C                                                                       ZERHPS   5
C     STORAGE FOR INFORMATION USED BY DPLOT  -- POINTERS                KIO1     2
      COMMON/KIO1/NUMCOM(200),TYPCOM(200),IPTCOM(200),HPSCTD(200),NOCOM, KIO1     3
     1 NOTEST,NEED,MASK5,MASK6                                          KIO1     4
      INTEGER TYPCOM,HPSCTD                                             KIO1     5
C     STORAGE FOR INFORMATION USED BY DPLOT  --A) COMMUNAL ARRAYS       KIO1A    2
      COMMON/KIO1A/ISTCOM(1600),ASTCOM(400)                             KIO1A    3
C                                                                       ZERHPS   8
C     VARIABLES USED TO SPECIFY HPS BUT NOT USED BY DPLOT -- POINTERS   KIO2     2
      COMMON/KIO2/NOHPS,ONEHPS(150),ASSCOM(150)                         KIO2     3
      INTEGER ONEHPS,ASSCOM                                             KIO2     4
C     VARIABLES USED TO SPECIFY HPS BUT NOT USED BY DPLOT -- COMMUNAL ARRAYS KIO2A 2
      COMMON/KIO2A/IHPS(2000),AMPS(400)                                 KIO2A    3
C                                                                       ZERHPS  11
C     PRESET CONSTANTS                                                  KIOPRE   2
      COMMON/KIOPRE/NDLEN,NIDLEN,NJDLEN,MAXNM1,MAXNM2,MAXNM3,MOTEST,    KIOPRE   3
     1 MAXHPS,MAXNAM,MOCOM,MSTOR1,MSTOR2,MXHPS1,MXHPS2,MAXTIT,          KIOPRE   4
     2 NOCBIT,MACHIN,IDBFL,NOBLOC,IFILP,IRRFIL,PRETAR,                  KIOPRE   5
     3 NOMODE,ICORE(4),JCORE1,JCORE2,JCORE3,JSUMRY,IPERM,ISUMRY,        KIOPRE   6
     4 LCMONE,LCMTOT,PAPMOD,MAXCOL,MAXTYP,MAXCRD,ICRFIL,                KIOPRE   7
     5 NOUNSP,IPLOT1,IPLOT2,IPLOT3,IPLOT4,IPLOT5                        KIOPRE   8
      INTEGER WRDMOD(1),WRDUNS(5),PAPMOD,PRETAR                         KIOPRE   9
       EQUIVALENCE (WRDMOD,ICORE),(WRDUNS,IPLOT1)                       KIOPRE  10
C                                                                       ZERHPS  13
      COMMON/ZERCOM/AVAL                                                ZERHPS  14
      DIMENSION IIVAL(1)                                                ZERHPS  15
      EQUIVALENCE (IIVAL,AVAL)                                          ZERHPS  16
C                                                                       ZERHPS  17
      DIMENSION IA(1)                                                   ZERHPS  18
C                                                                       ZERHPS  19
      ITYPE= TYPCOM(ICOM)                                               ZERHPS  20
      IF(ITYPE.LE.3) GO TO 1                                            ZERHPS  21
      RETURN                                                            ZERHPS  22
C                                                                       ZERHPS  23
    1 J=IPTCOM(ICOM)                                                    ZERHPS  24
      INO=ISTCOM(J+2)                                                   ZERHPS  25
      I1=NUMCOM(ICOM)                                                   ZERHPS  26
      I1=ONEHPS(I1)                                                     ZERHPS  27
      KIN=0                                                             ZERHPS  28
      GO TO (11,12,13),ITYPE                                            ZERHPS  29
C                                                                       ZERHPS  30
C     HISTOGRAM                                                         ZERHPS  31
   12 IF(IHPS(I1+6).EQ.2) GO TO 11                                      ZERHPS  32
      KIN=INO-14*IDBFL                                                  ZERHPS  33
      GO TO 11                                                          ZERHPS  34
C                                                                       ZERHPS  35
C     SCATTERPLOT                                                       ZERHPS  36
   13 IF(IHPS(I1+6).EQ.2) GO TO 11                                      ZERHPS  37
      KIN=INO-30*IDBFL                                                  ZERHPS  38
C                                                                       ZERHPS  39
C     KIN INTEGER AND KFL FLOATING WORDS TO INITIALIZE                  ZERHPS  40
   11 KFL=(INO-KIN)/IDBFL                                               ZERHPS  41
      AVAL=0.                                                           ZERHPS  42
      L=0                                                               ZERHPS  43
```

- Newton’s laws s
Einstein’s specia
theory
- Physicists just d
whose existence
- Its search was h
model is needed
- A model is a hop
approach that a
that are fit to ex

[http://en.wikipedia.o
Simple_linear_regres](http://en.wikipedia.o)
macroeconomics is a
simple linear regress
dependent variable (
presumed to be in a
the changes in the ur

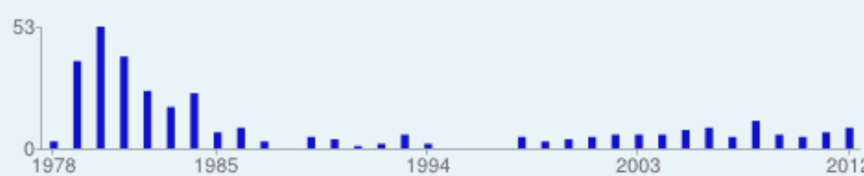
Geoffrey Fox

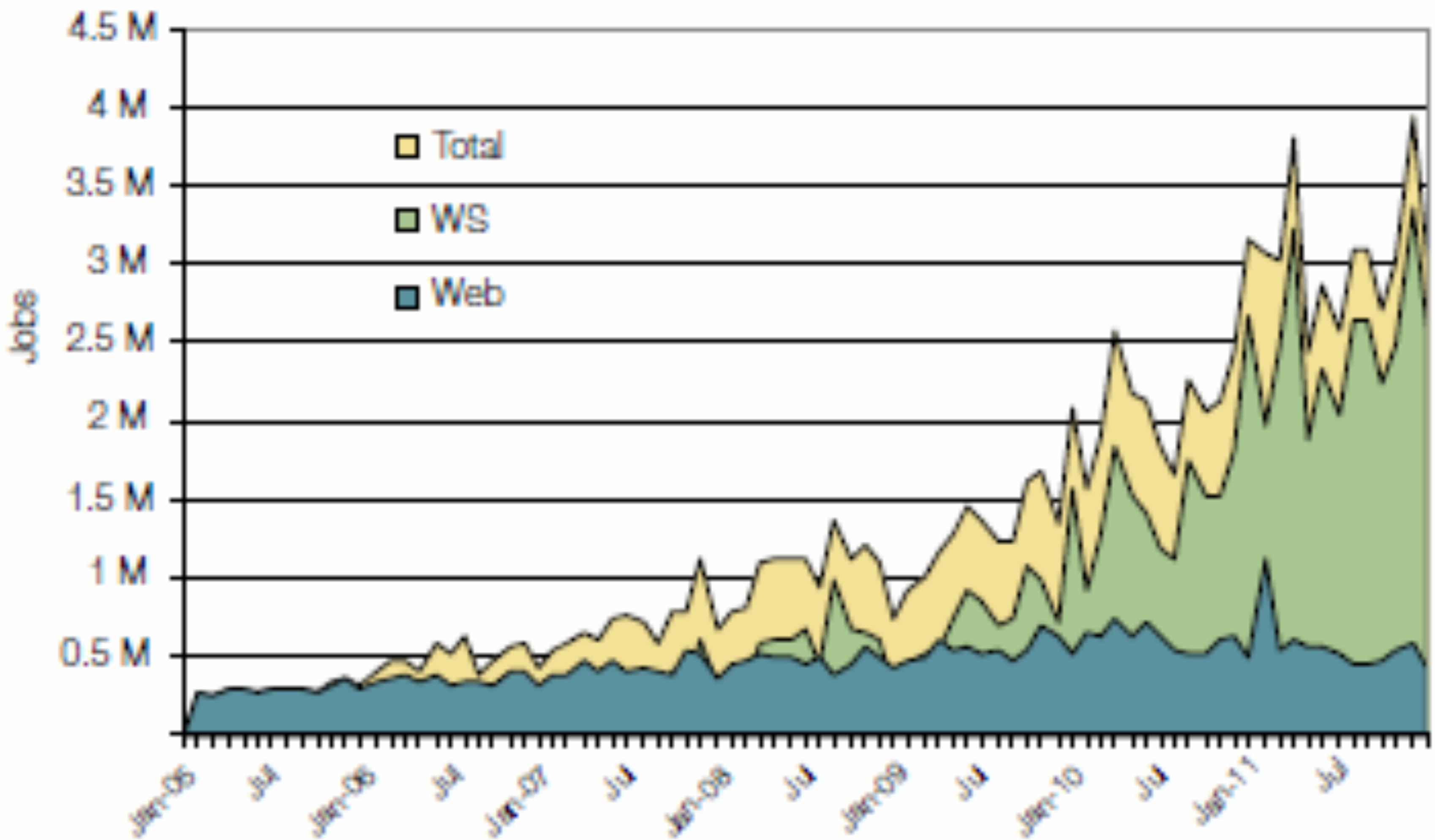
Professor of Informatics and Computing and Physics
Cyberinfrastructure - Distributed Systems - Clouds - Parallel Computing - Particle Physics
Verified email at indiana.edu
Homepage

Go
Searc
My Cita

« Back to list | Edit | Export | Delete

Title	Quantum-chromodynamic approach for the large-transverse-momentum production of particles and jets
Authors	RP Feynman, RD Field, GC Fox
Publication date	1978/11/1
Journal name	Physical Review D
Volume	18
Issue	9
Pages	3320
Publisher	American Physical Society
Description	I. INIODUCTION% e investigate whether the present experimental behavior of mesons with large transverse mo-mentum in hadron-hadron collisions is consistent with thetheory of quantum-chromodynamics(QCD) with asymptotic freedom, atleast as the theory is now partially understood. It is shown that if things behave more or less according to current theo-retical ideas, the experimental data at high P~ would be explicable with reasonable choices for currently unknown quantities (such as the dis-tribution of gluonsin the proton and the ...

Total citations	Cited by 373
Citations per year	
Scholar articles	Quantum-chromodynamic approach for the large-transverse-momentum production of particles and jets RP Feynman, RD Field, GC Fox - Physical Review D, 1978 Cited by 373 - Related articles - All 12 versions



2005-20011 Job request at European Bioinformatics Institute EBI for Web hits and automated services WS
http://www.ebi.ac.uk/Information/Brochures/

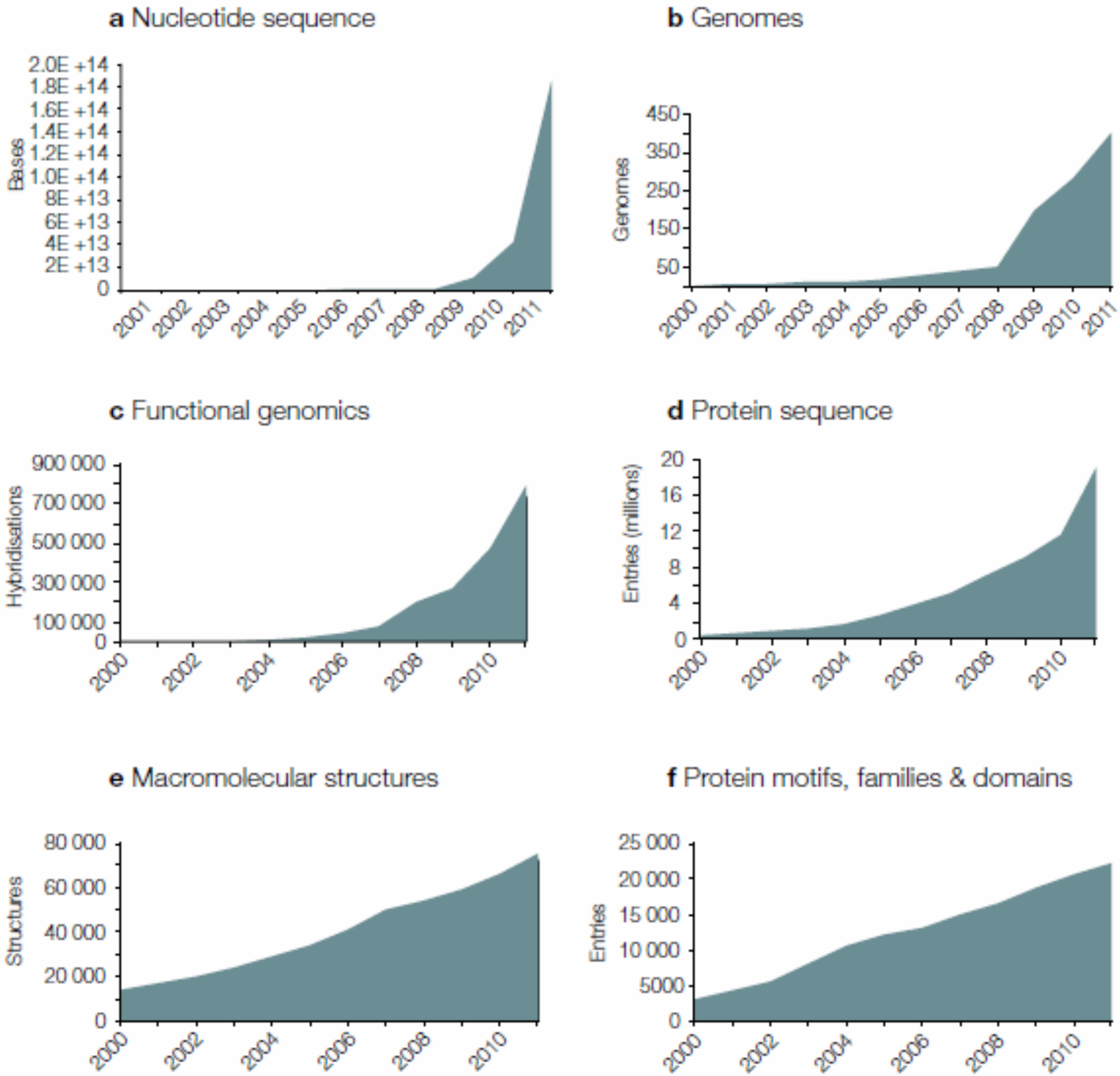


Figure 2. Growth of EMBL-EBI's core data resources from 2000 to 2011. (a) Nucleotide sequence (bases in the European Nucleotide Archive); (b) genomes (entire genomes in Ensembl plus Ensembl Genomes combined); (c) functional genomics (assays in the ArrayExpress Archive.); (d) protein sequence (protein sequences in UniParc); (e) macromolecular structures (structures in PDBe); (f) protein families, motifs and domains (entries in InterPro).

2005-20011 Data stored at European Bioinformatics Institute EBI

http://www.ebi.ac.uk/Information/Brochures/

Data Deluge Implications for Scientific Method

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

The 4 paradigms of Scientific Research

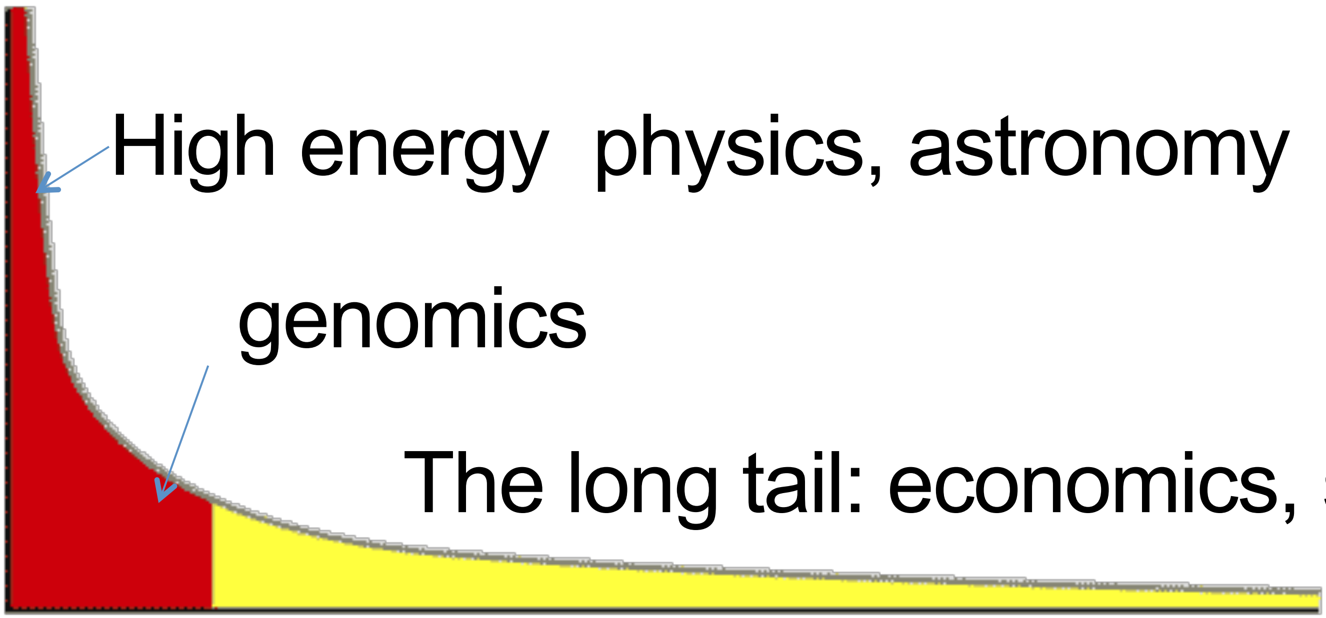
1. Theory
2. Experiment or Observation
   - E.g. Newton observed apples falling to design his theory of mechanics
3. Simulation of theory or model
4. Data-driven (Big Data) or The Fourth Paradigm: Data-Intensive Scientific Discovery (aka Data Science)
   - http://research.microsoft.com/en-us/collaboration/fourthparadigm/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

Another Personal Note

- In 1990, only methods 1 and 2 were recognized but due to increasing power of computers, method 3 (computation science) was being recognized
- I tried to persuade Caltech to adopt a “computational science curriculum” but failed
  - I left Caltech partly for this reason
- I now realize that perhaps not such a good idea as not huge numbers of jobs in area.
- However starting in 2005-2010, method 4 and data science emerges
  - There are lots of jobs in data science so curricula perhaps more interesting

Data Deluge
Long Tail of Science

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

Gannon Talk

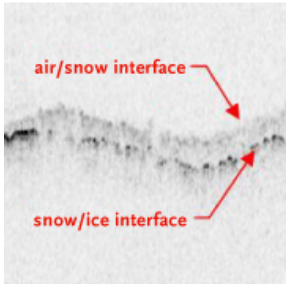
Data Deluge
Internet of Things

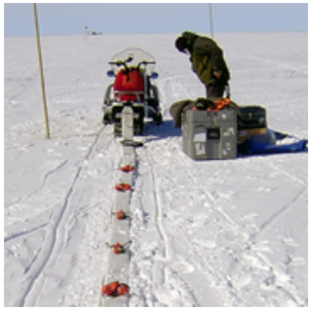
Internet of Things and the Cloud

- 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

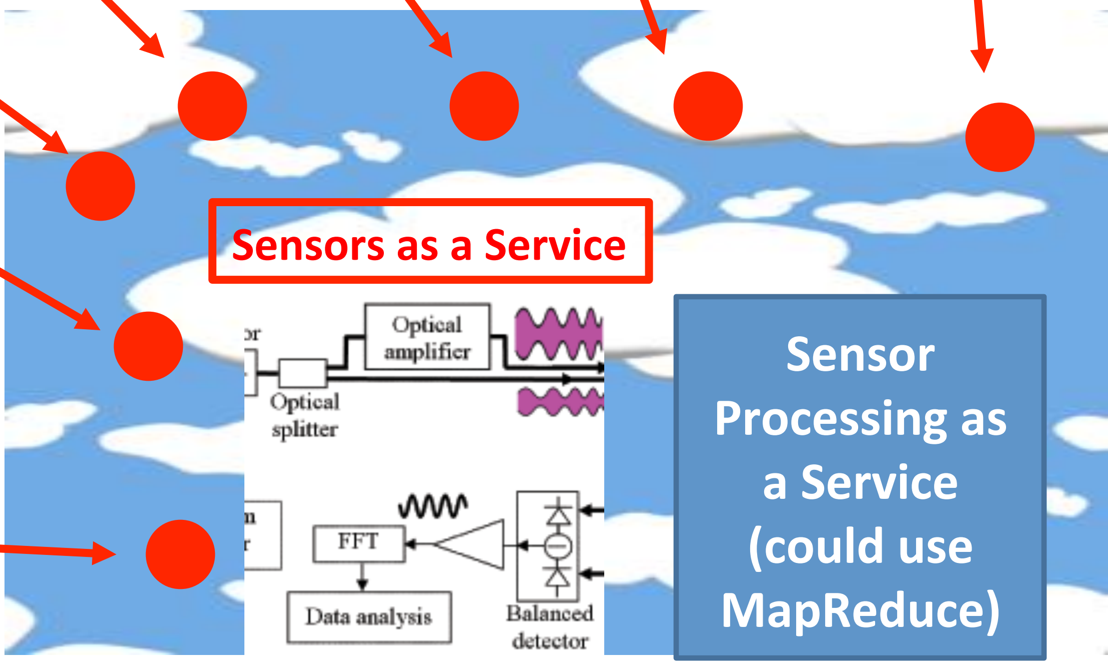
Sensors (Things) as a Service

Output Sensor





A larger sensor ....



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