Big Data and Clouds

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

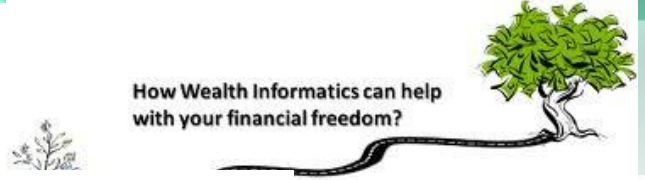
Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X)

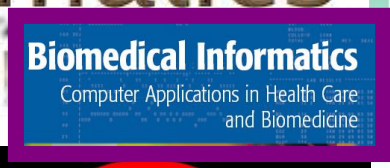
X = Astronomy, Biology, Biomedicine, Business, Chemistry, Climate, Crisis, Earth Science, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly

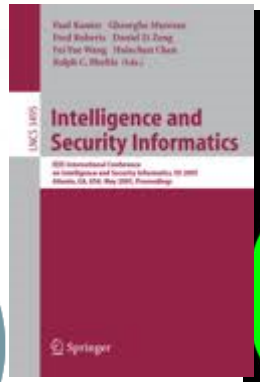
Spans Industry and Science (research)

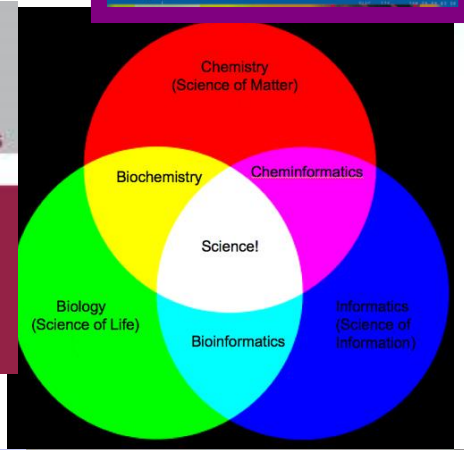
Education: Data Science see some New York Times articles
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

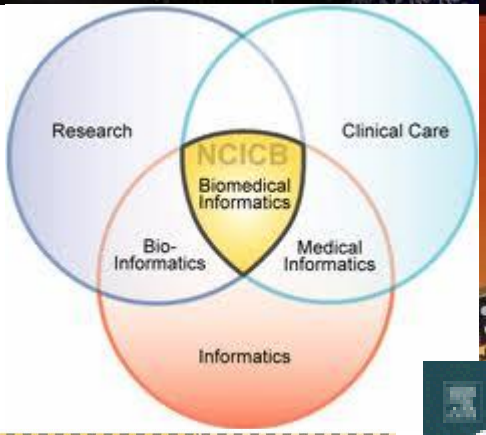
X-Informatics Class http://www.infomall.org/X-InformaticsSpring2013/
Big data MOOC http://x-informatics.appspot.com/preview

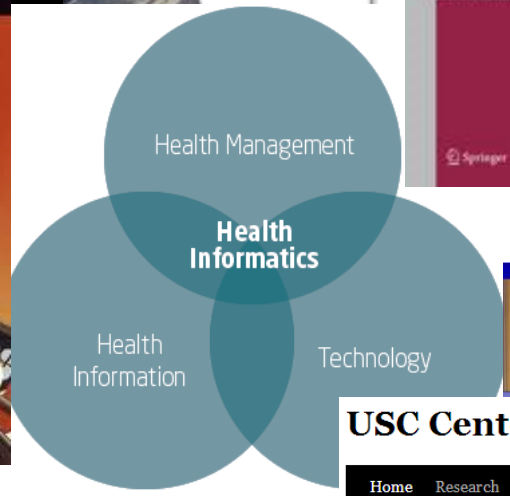


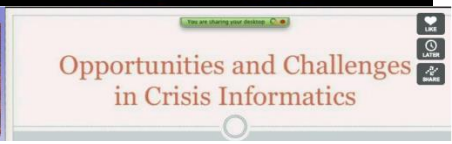


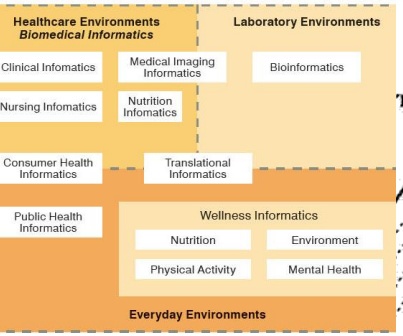


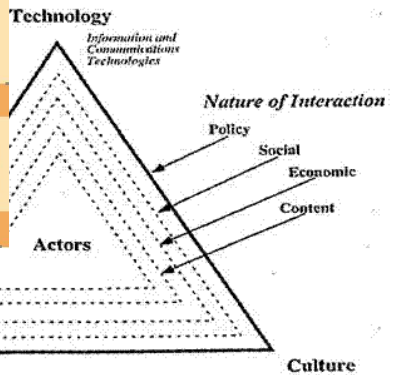




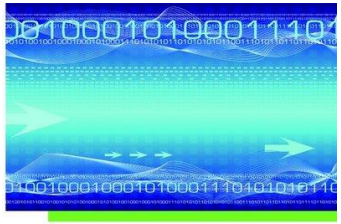








Social Informatics



Noelia Penelope Greer (Ed.)

Business Informatics

Information technology, Management,

USC Center For Energy Informatics

Home Research Publications Sma

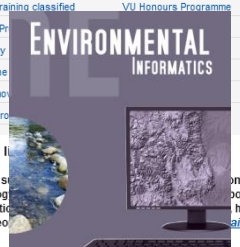
About the Center

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

Lifestyle Informatics

Applications of LI
How is the training classified
Occupation Pr
Further study
Student at the
Watch the mo
Studying Abro

Admission and registration
VU Honours Programme



Lifestyle Informatics: Let people l

The study Lifestyle Informatics is about s
this bachelor including applied psycholog
knowledge about language and informatic
short better. Lifestyle Informatics: let peo
Lifestyle Informatics

ombine
body,
healthier,
aining

Motivation

- In 2016 there exits 16 zettabytes of shared stored digital data with a zettabyte = 10^9 terabytes
  - A 2TB USB disk costs <$100 today
  - Today (late 2014) 1.8 Billion images are uploaded to cloud every day
- Cloud computing is exploding to handle exploitation of this data
- New industries and new research areas with new software and new algorithms
- There is an online course that gives an overview of big data from a use case (application) point of view noting that big data in field X drives the concept of X-Informatics
  - It covers applications, algorithms and infrastructure/technology (cloud computing)
- There is also a free MOOC with URL https://bigdatacourse.appspot.com/preview
- All lectures are offered online with a set of 5-15 minute lessons on YouTube containing video with content and talking head
- See http://www.infomall.org/cglmoocs for list of courses

Economic Imperative

There are a lot of data and a lot of jobs

Data Deluge

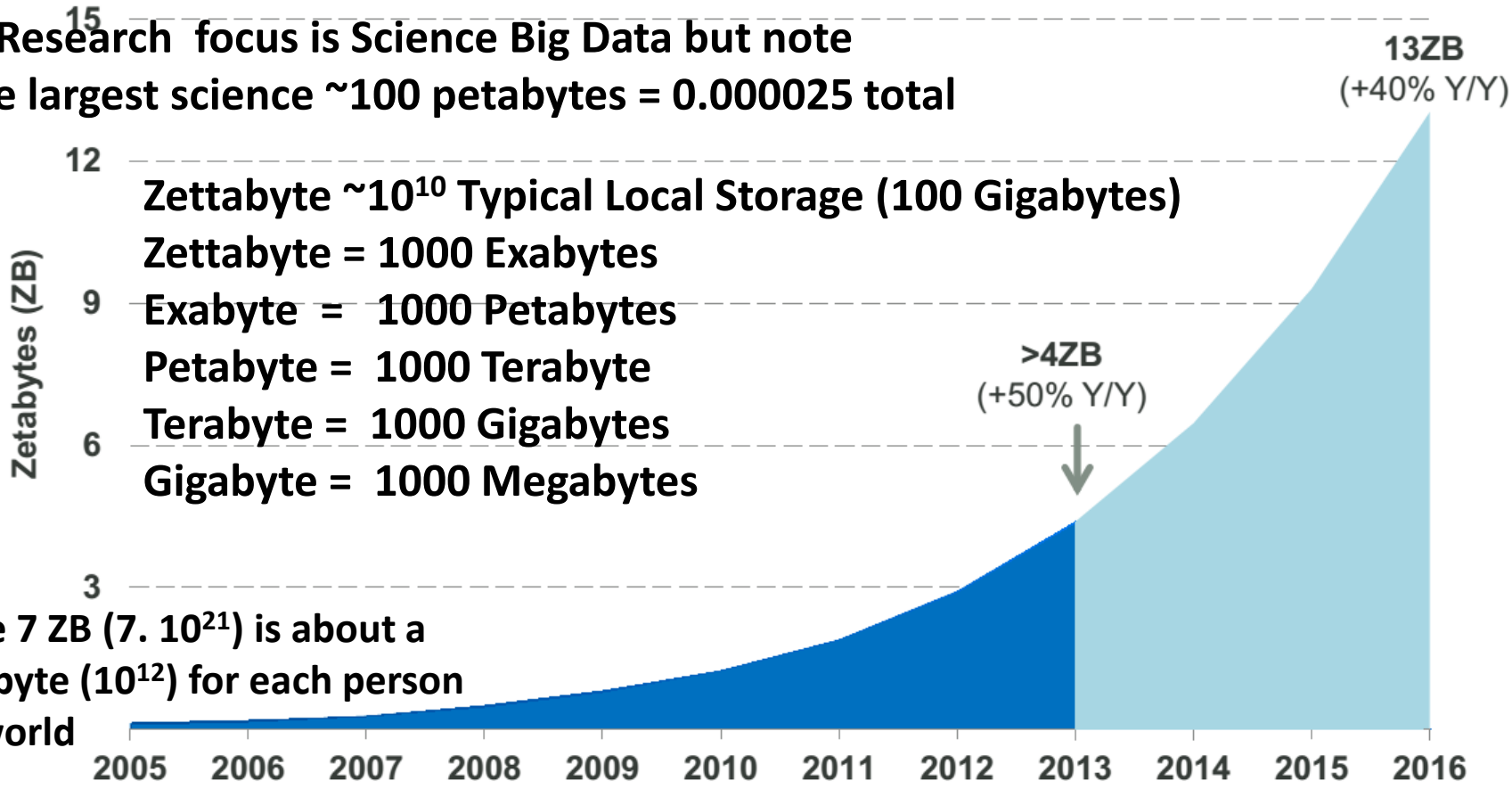
Some Trends

- **The Data Deluge** is clear trend from Commercial (Amazon, e-commerce) , Community (Facebook, Search) and Scientific applications
- **Light weight clients** from smartphones, tablets to sensors
- **Multicore** reawakening parallel computing
- **Clouds with cheaper, greener, easier to use** IT for (some) applications
- **New jobs** associated with new curricula
  - **Clouds as a distributed system** (classic CS courses)
  - **Data Analytics** (Important theme in academia and industry)
  - **Network/Web Science**

‘Digital Universe’ Information Growth = Robust... +50%, 2013

2/3rd's of Digital Universe Content = Consumed / Created by Consumers
...Video Watching, Social Media Usage, Image Sharing...

My Research focus is Science Big Data but note
Note largest science ~100 petabytes = 0.000025 total



Zettabyte ~10^{10} Typical Local Storage (100 Gigabytes)
Zettabyte = 1000 Exabytes
Exabyte = 1000 Petabytes
Petabyte = 1000 Terabyte
Terabyte = 1000 Gigabytes
Gigabyte = 1000 Megabytes

Note 7 ZB (7. 10^{21}) is about a terabyte (10^{12}) for each person in world

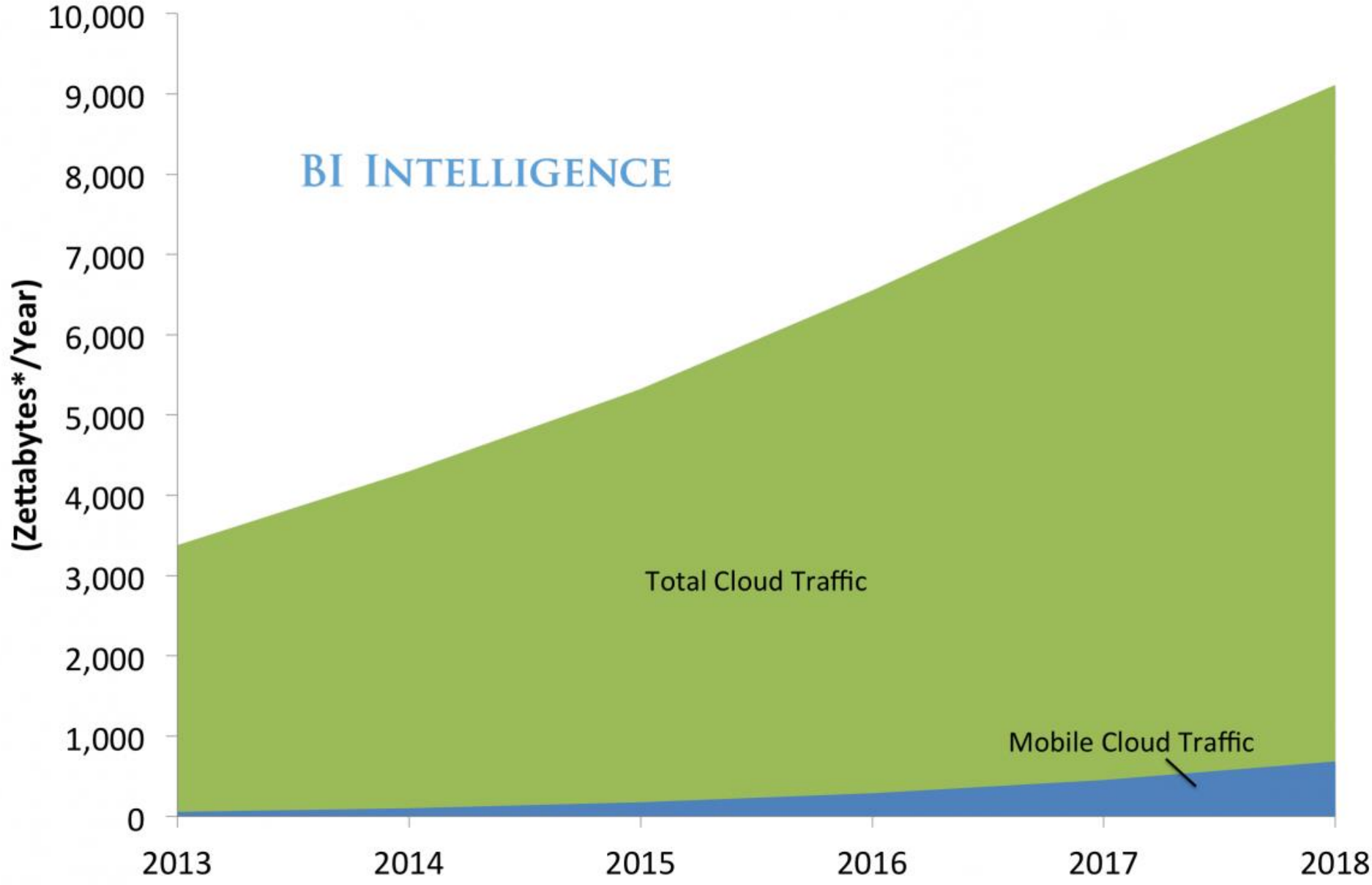
@KPCB

Note: 1 petabyte = 1MM gigabytes, 1 zetabyte = 1MM petabytes.
Source: IDC Digital Universe, data as of 5/14.

http://www.kpcb.com/internet-trends

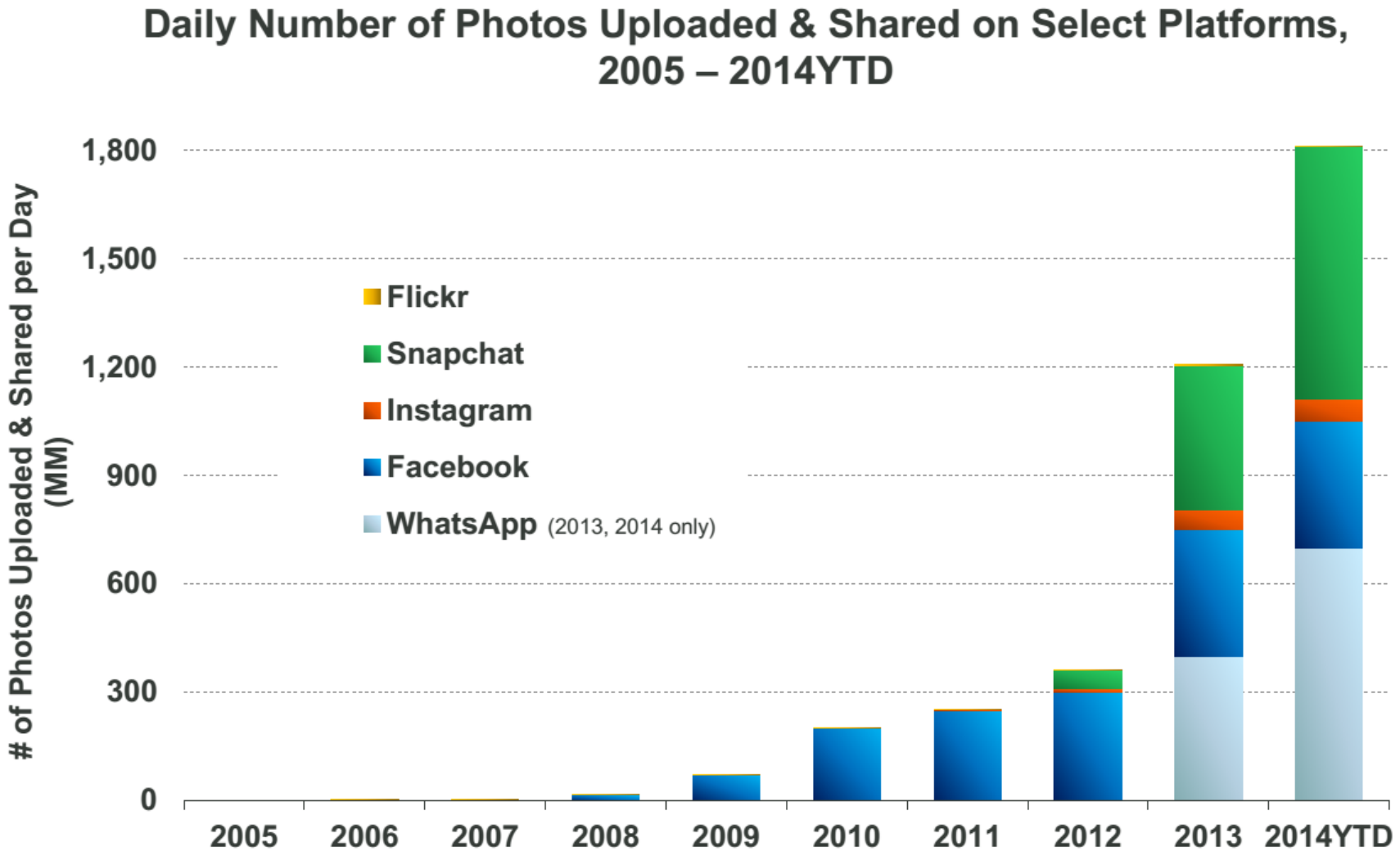
Global Cloud Traffic Forecast

Mobile Share Of Overall Cloud Traffic



*Source: Cisco, Mobidia, BI Intelligence Estimates; *1 Zettabyte = 1 Billion Terabytes*

Photos Alone = 1.8B+ Uploaded & Shared Per Day...
Growth Remains Robust as New Real-Time Platforms Emerge

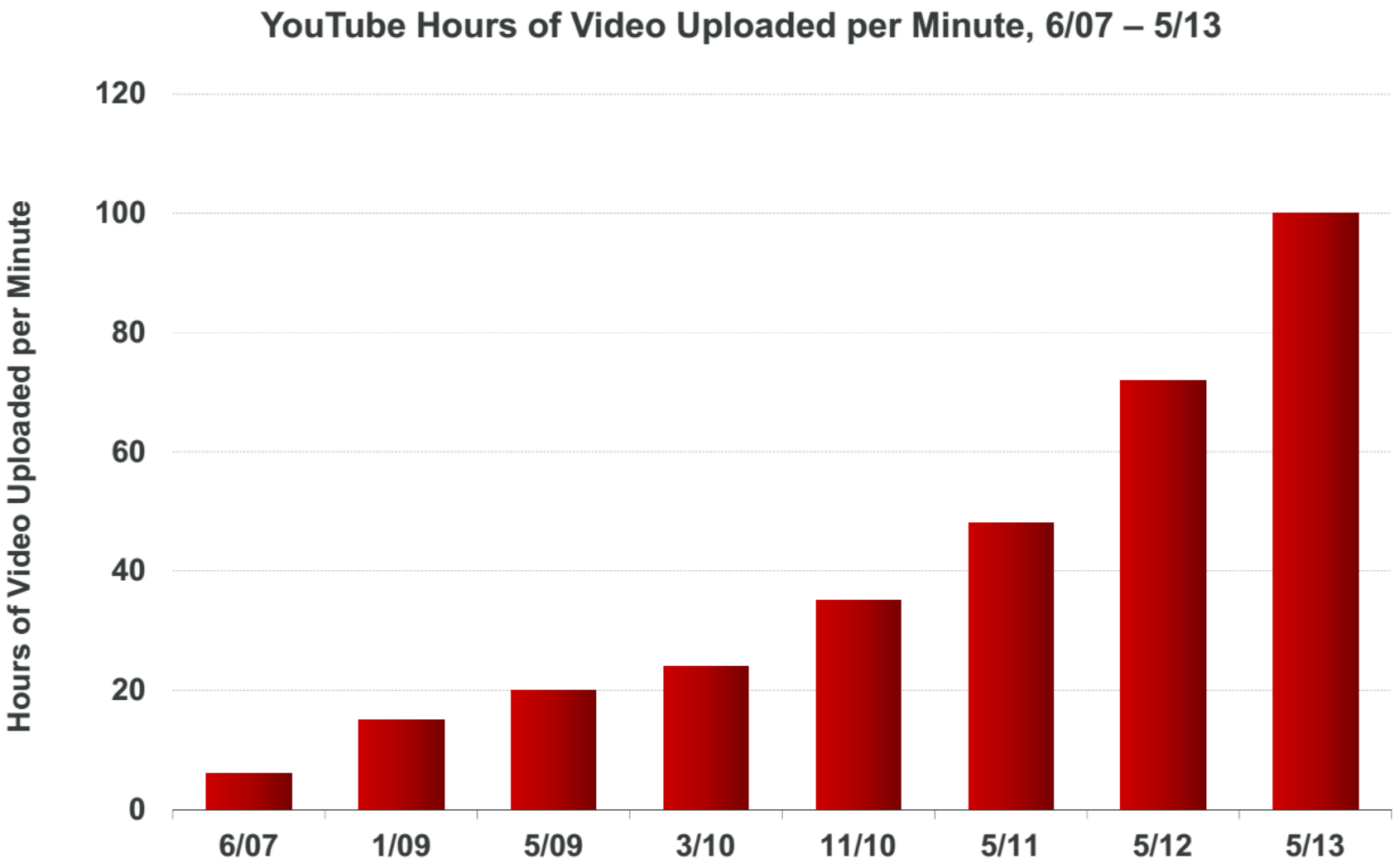


@KPCB

Source: KPCB estimates based on publicly disclosed company data, 2014 YTD data per latest as of 5/14.

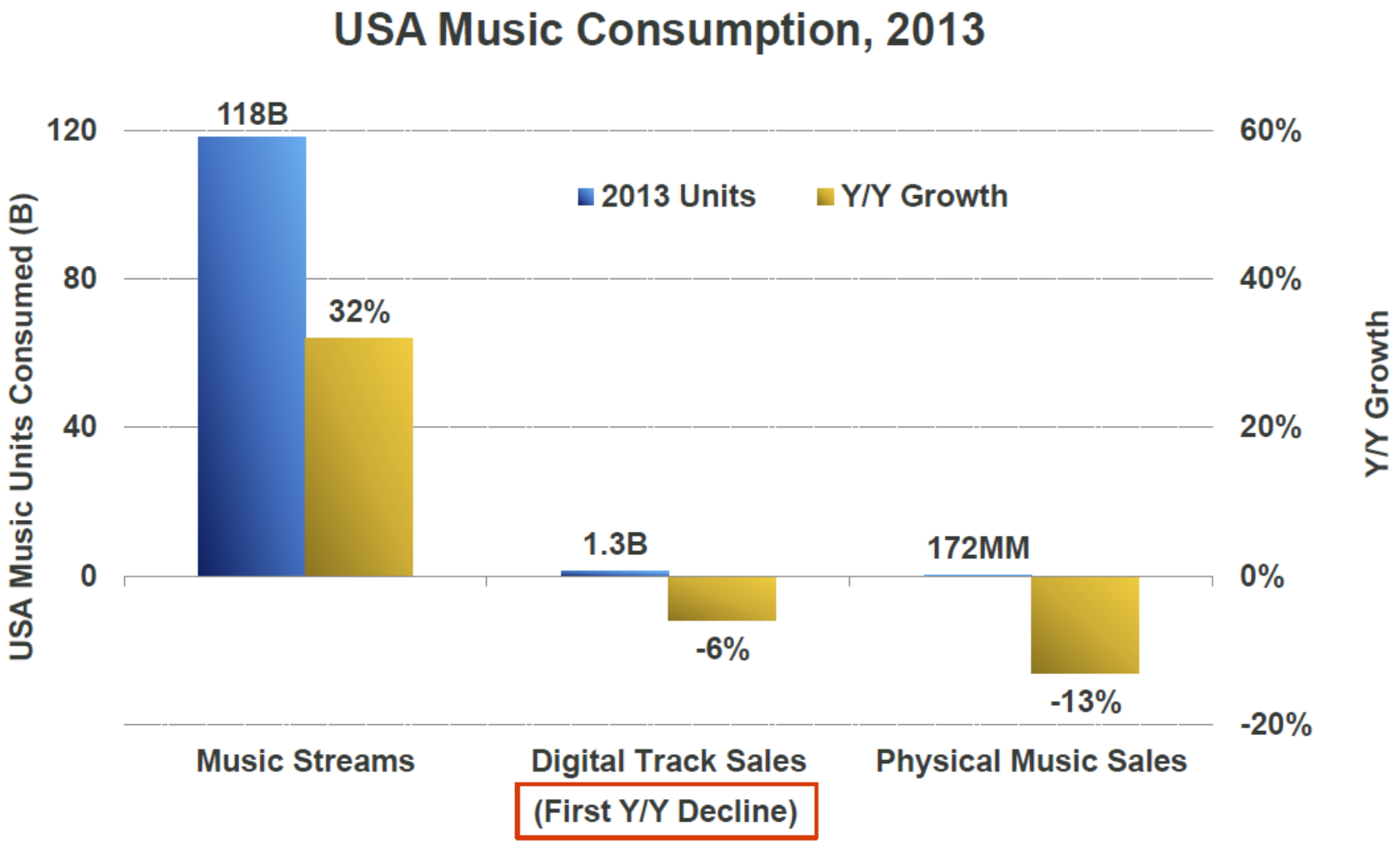
http://www.kpcb.com/internet-trends

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



rce: YouTube.

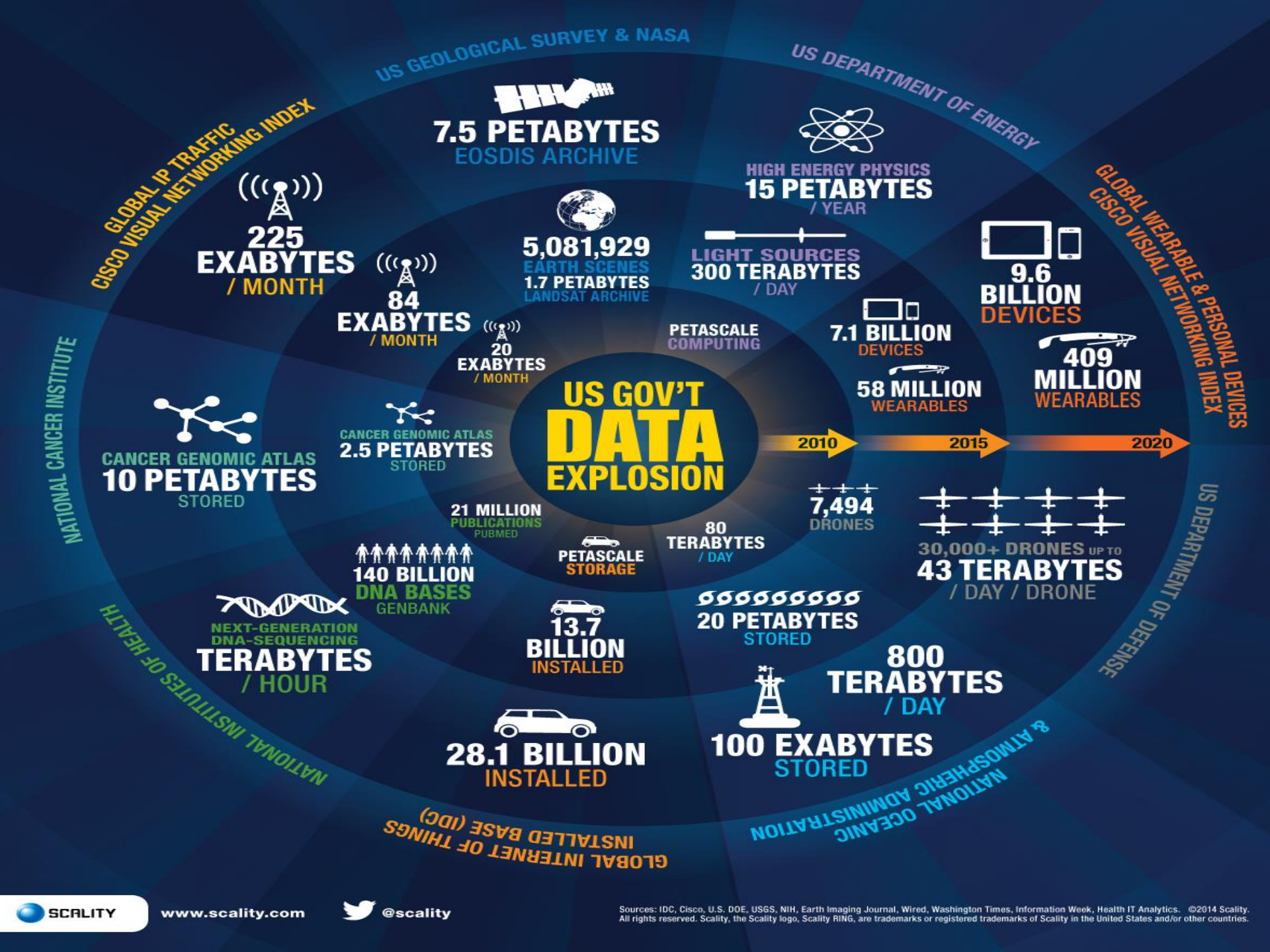
Re-Imagining Media (Music) Consumption = Streaming +32%, Digital Track Sales -6%



@KPCB

Source: Nielsen & Billboard 2013 US Music Report, 1/14. Note that absolute consumption comparisons are apples-and-oranges as tracks / physical sales are likely played multiple times but data is illustrative as growth rate is key indicator.


1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY
320+ NEW twitter ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
100+ NEW Linked in ACCOUNTS
IN 60 SECONDS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
50+ WORDPRESS DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
GO-Globe.com
Google
Google Search
Adobe Acrobat X Pro



"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

80 Gigabytes per day

enabling social connections

vs.

Gas Turbine Compressor Blade Monitoring potential*

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

* Note: Assumes operational gas turbines (generating units only) >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

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

is 50/50 JV with Pratt & Whitney

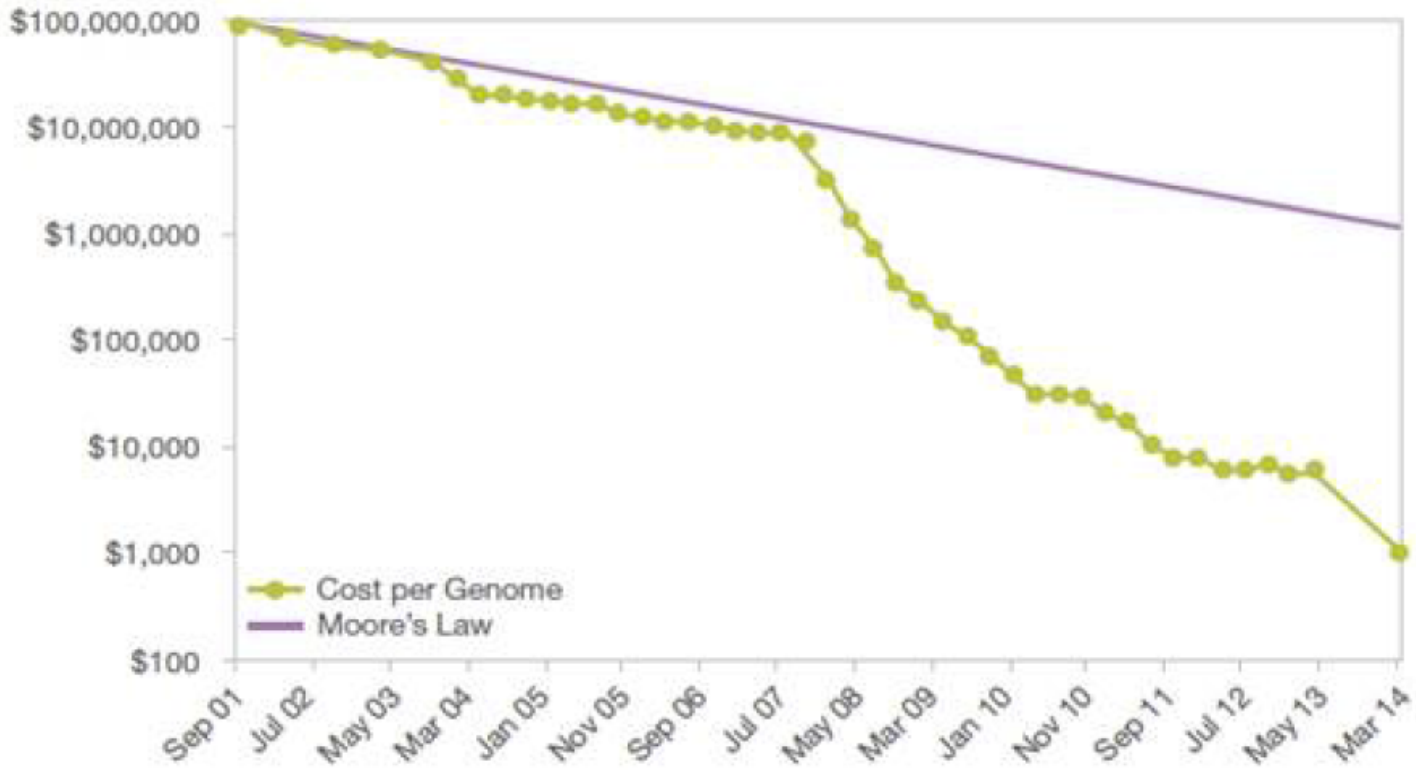
Some Science Data Sizes

- ~40 10^9 **Web pages** at ~300 kilobytes each = 10 Petabytes
- **Youtube** 48 hours video uploaded per minute;
  - in 2 months in 2010, uploaded more than total NBC ABC CBS
  - ~2.5 petabytes per year uploaded?
- **Radiology** 69 petabytes per year
- **Square Kilometer Array Telescope** will be 100 terabits/second
- **Earth Observation** becoming ~4 petabytes per year
- **Earthquake Science** – few terabytes **total** today
- **PolarGrid** – 100's terabytes/year
- **Exascale simulation** data dumps – terabytes/second

Cost / Time to Sequence Genome Down to $1,000 / 24 Hours – Treasure Trove of Patterns Will Rise Rapidly

*Accurate diagnosis is foundation for choosing right treatments for patients & clinical lab tests provide critical information health care providers use in ~70% of decisions**

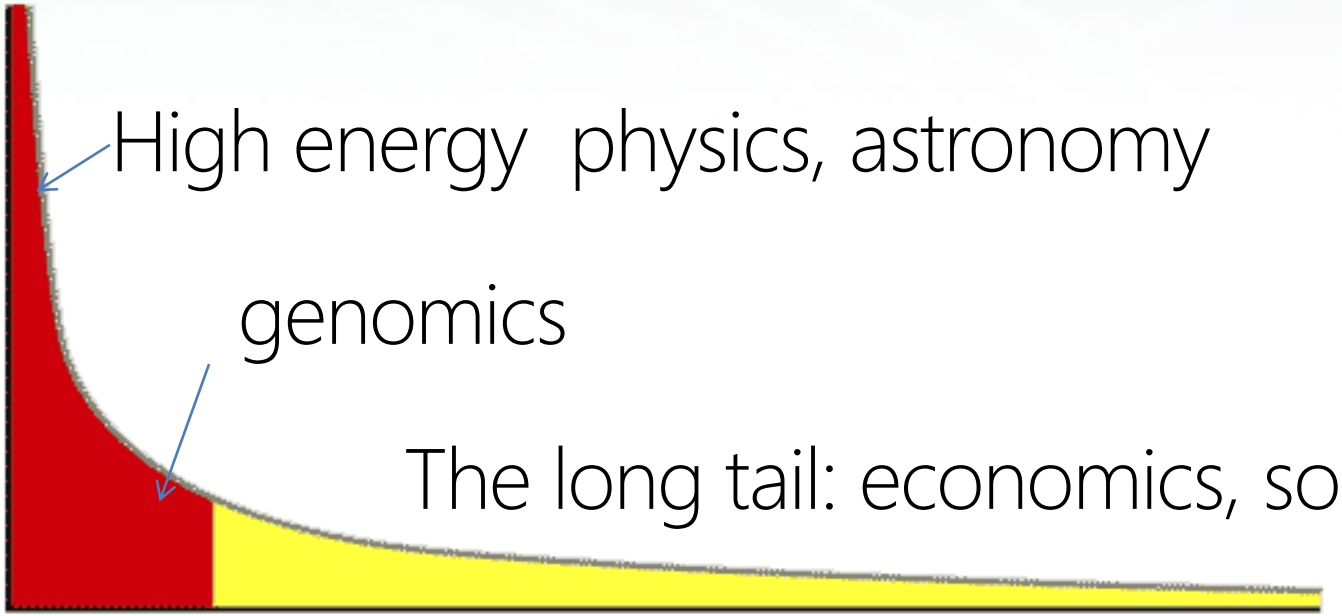
*Genetic & genomic testing can be at heart of a new paradigm of [precision] medicine that is evidence-based & rooted in quantitative science***



@KPCB

*UK Department of Health. ** American Clinical Laboratory Association / BattelleTechnology Partnership Practice. Image: Illumina.
Note: Genome sequencing data per Eric Schaldt. $1,000 cost is price of sequencing a genome at 30x coverage in the Mount Sinai Genome Core, 5/14.

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

From Dennis Gannon Talk

DATA INTENSIVE ACTIVITIES

- **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 (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

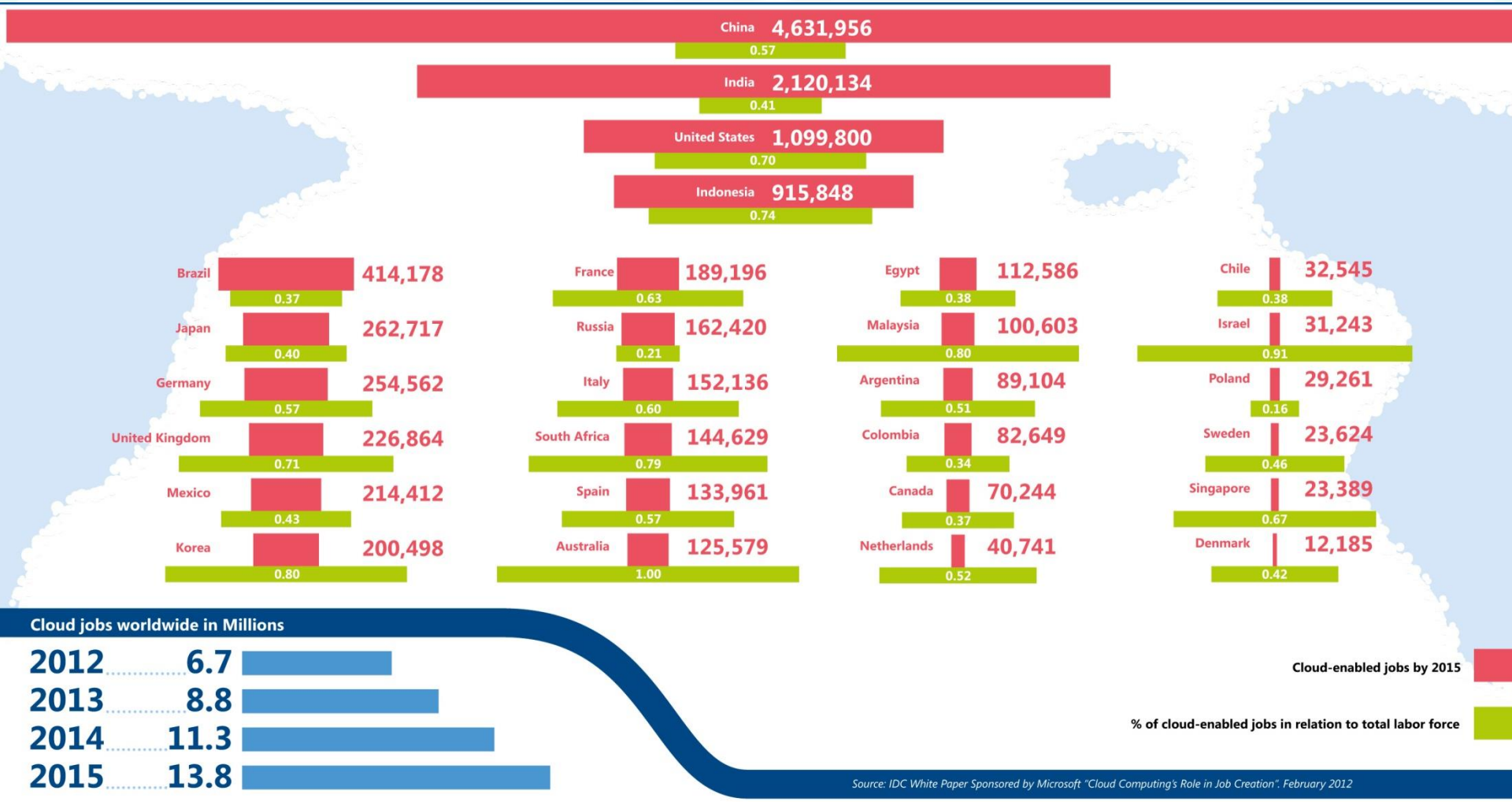
Spans Industry and Science (research)

Education: Data Science see some New York Times articles
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

X-Informatics Class http://www.infomall.org/X-InformaticsSpring2013/
Big data MOOC http://x-informatics.appspot.com/preview

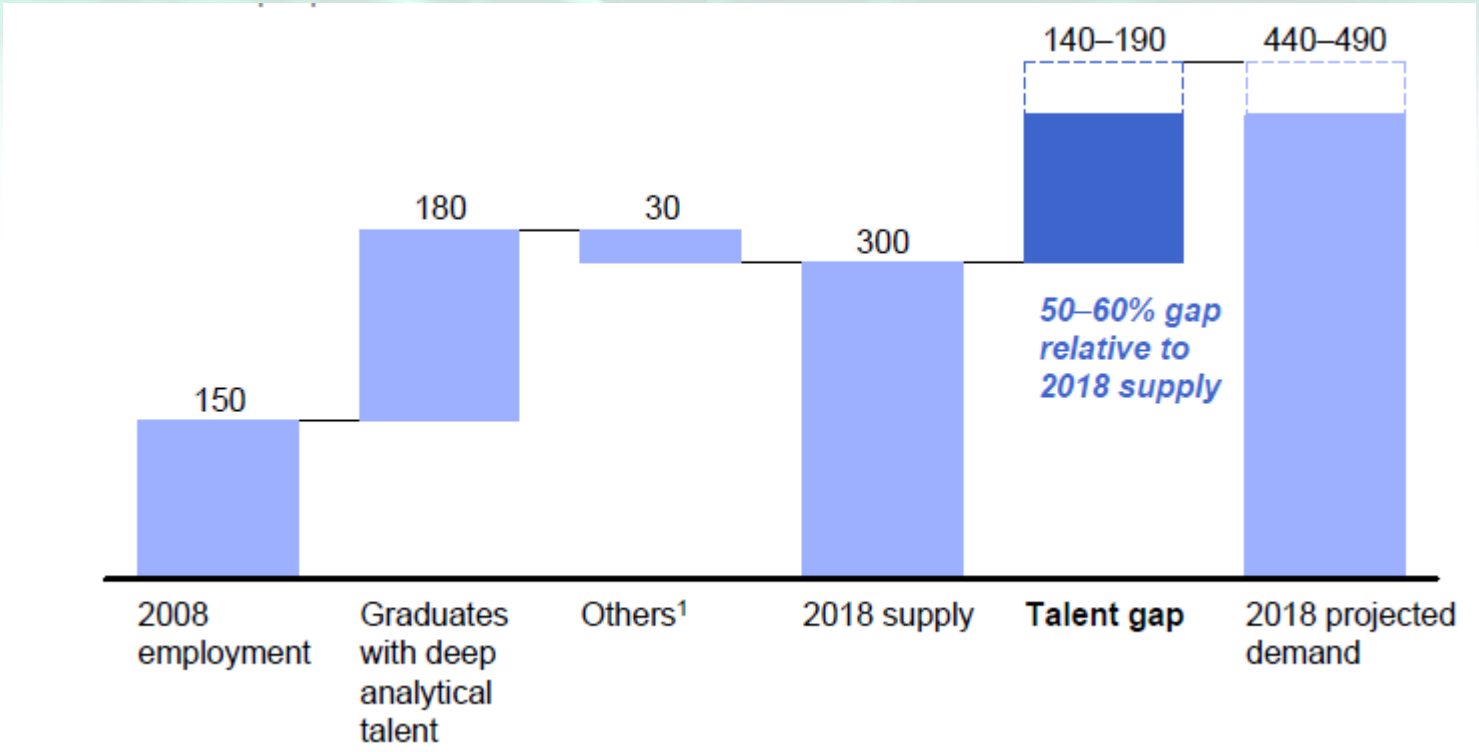
Jobs

Jobs v. Countries



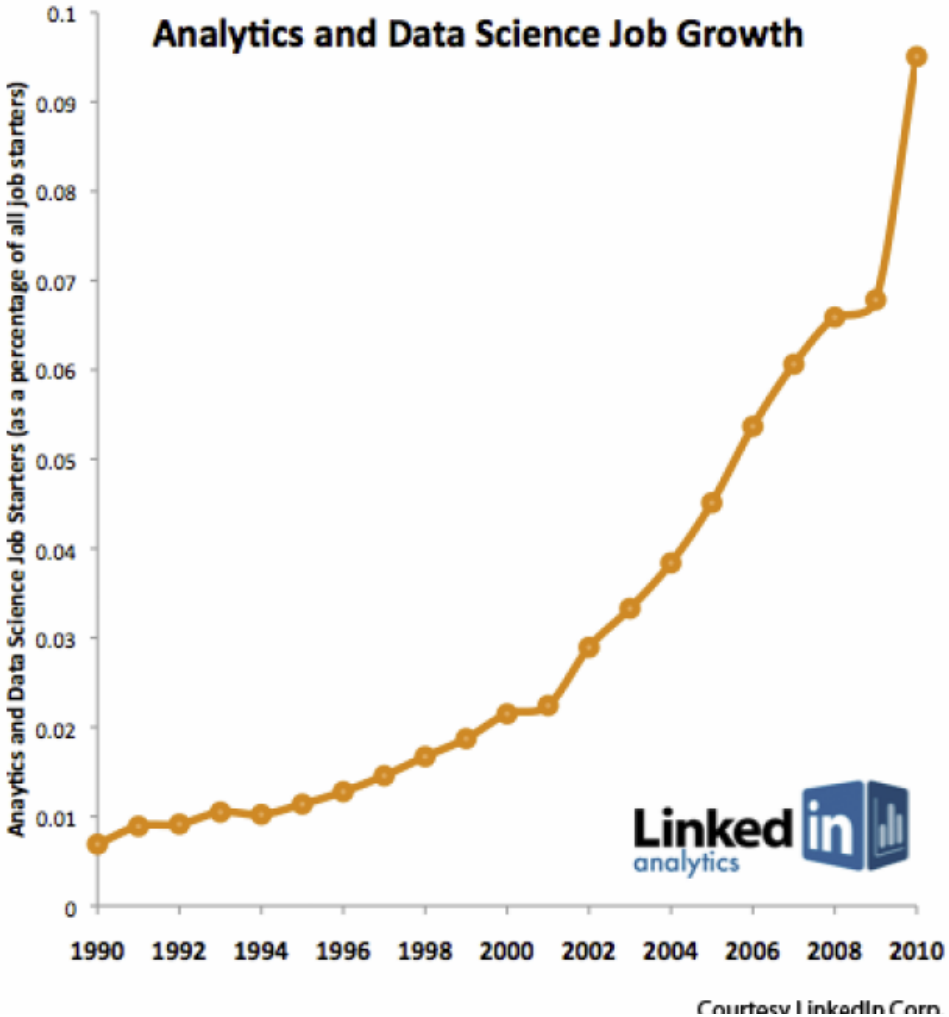
http://www.microsoft.com/en-us/news/features/2012/mar12/03-05CloudComputingJobs.aspx

McKinsey Institute on Big Data Jobs



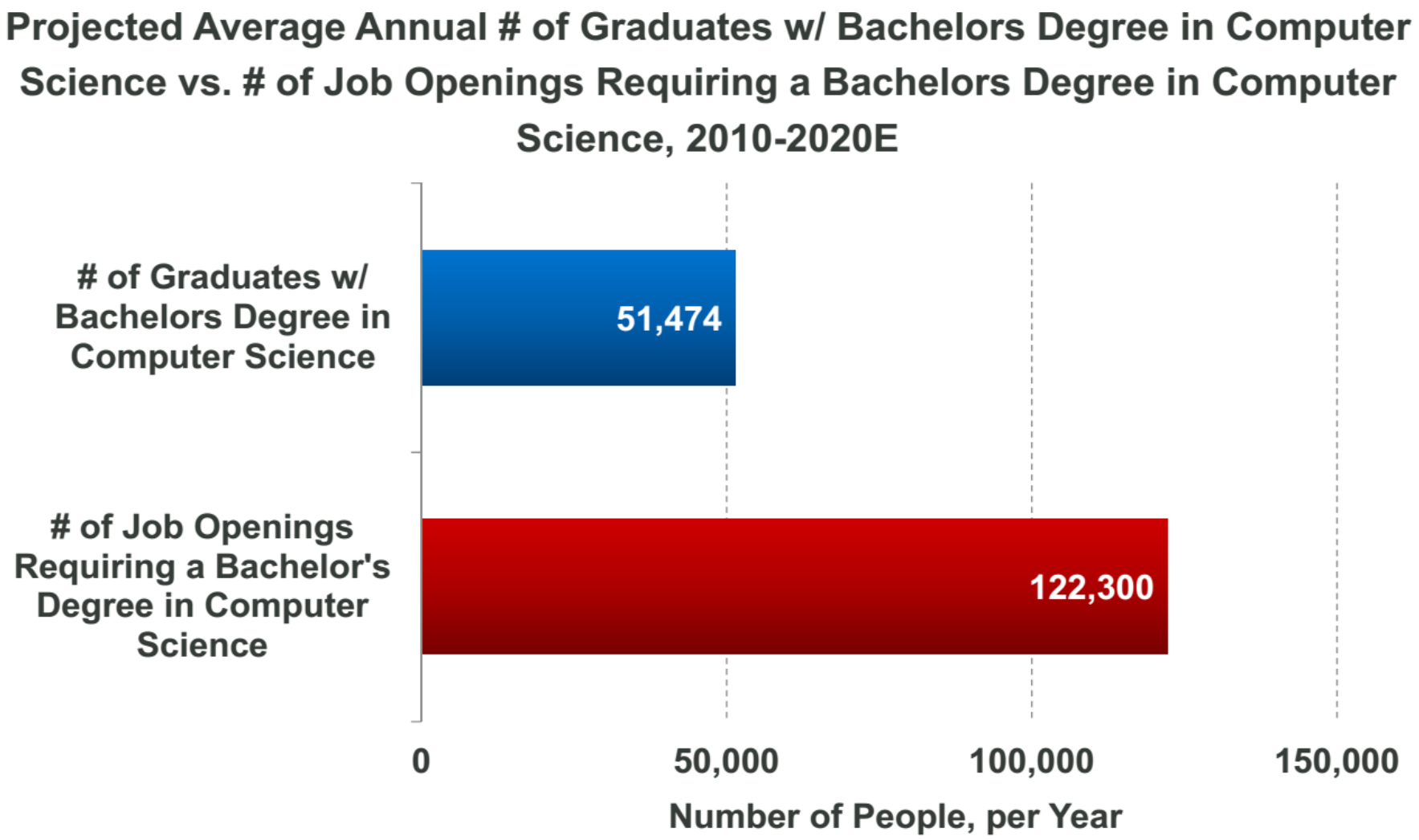
- There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data 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



KPCB

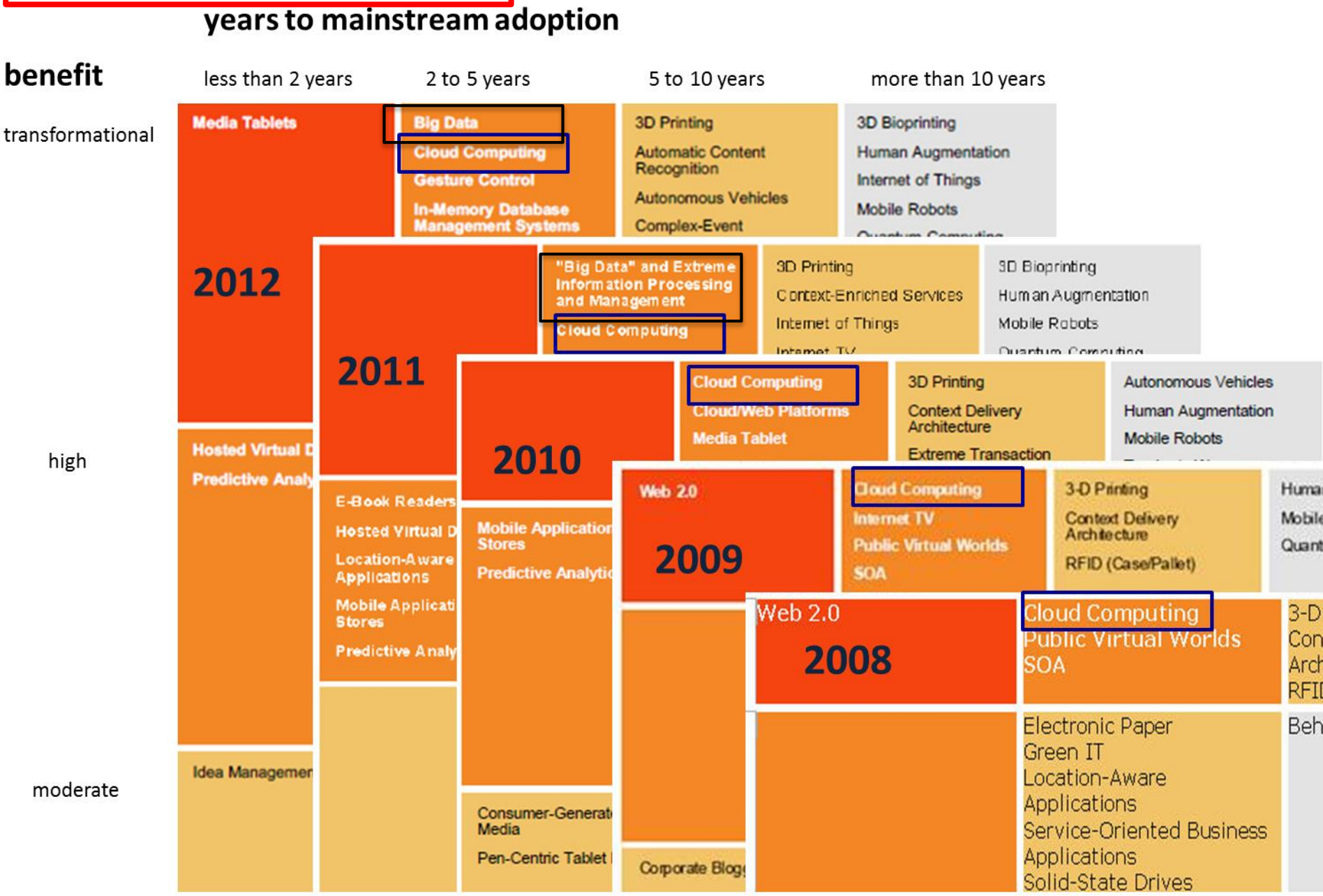
Source: Microsoft analysis (published in 2013) based on Bureau of Labor Statistics data.

Computing Model

Industry adopted clouds which are attractive for data analytics

5 years Cloud Computing
2 years Big Data Transformational

Gartner Priority Matrix



Amazon 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.

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

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

Clouds Offer From different points of view

- **Features:**
  - 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

Clouds in Research

2 Aspects of Cloud Computing: Infrastructure and Runtimes

- **Cloud infrastructure:** outsourcing of servers, computing, data, file space, utility computing, etc..
- **Cloud runtimes or Platform:** tools to do data-parallel (and other) computations. Valid on Clouds and traditional clusters
  - Apache Hadoop, Google **MapReduce**, Microsoft Dryad, Bigtable, Chubby and others
  - MapReduce designed for information retrieval but is excellent for a wide range of science data analysis applications
  - Can also do much traditional parallel computing for data-mining if extended to support iterative operations
  - **Data Parallel File system** as in HDFS and Bigtable

Clouds have highlighted SaaS PaaS IaaS

But equally valid for classic clusters

Software (Application Or Usage) **SaaS**
- **Education**
- **Applications**
- **CS Research Use e.g. test new compiler or storage model**

Platform **PaaS**
- **Cloud e.g. MapReduce**
- **HPC e.g. PETSc, SAGA**
- **Computer Science e.g. Compiler tools, Sensor nets, Monitors**

Infra structure **IaaS**
- **Software Defined Computing (virtual Clusters)**
- **Hypervisor, Bare Metal**
- **Operating System**

Network **NaaS**
- **Software Defined Networks**
- **OpenFlow GENI**

- Software Services are building blocks of applications
- The middleware or computing environment including **HPC, Grids** …
- Nimbus, Eucalyptus, OpenStack, OpenNebula CloudStack plus **Bare-metal**
- OpenFlow – *likely to grow in importance*

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