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

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

X-Informatics and its Motto

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
- **Exascale initiatives** will continue drive to high end with a simulation orientation
- **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)
  - **Social Media**

Some Terms

- **Data:** the raw bits and bytes produced by instruments, web , e-mail, social media
- **Information:** The cleaned up data without deep processing applied to it
- **Knowledge/wisdom/decisions** comes from sophisticated analysis of Information
- **Data Analytics** is the process of converting data to Information and Knowledge and then decisions or policy
- **Data Science** describes the whole process
- **X-Informatics** is use of Data Science to produce wisdom in field X

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

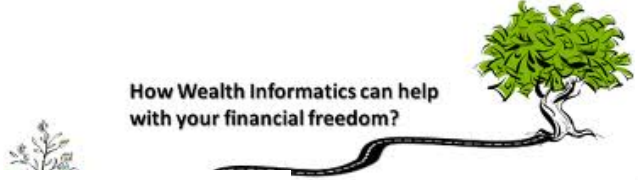
Spans Industry and Science (research)

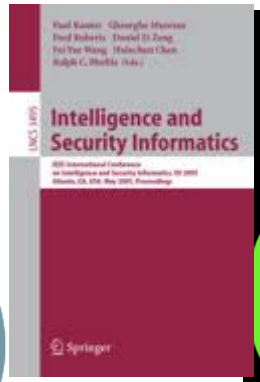
Education: Data Science see recent New York Times articles

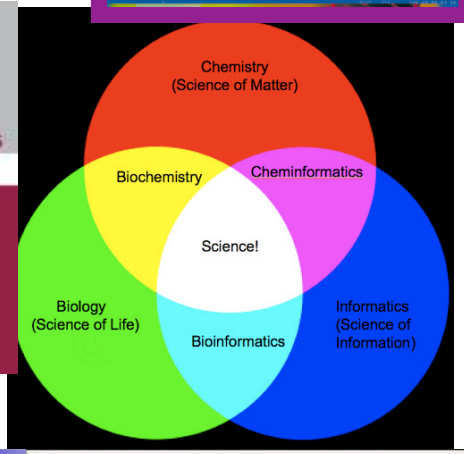
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

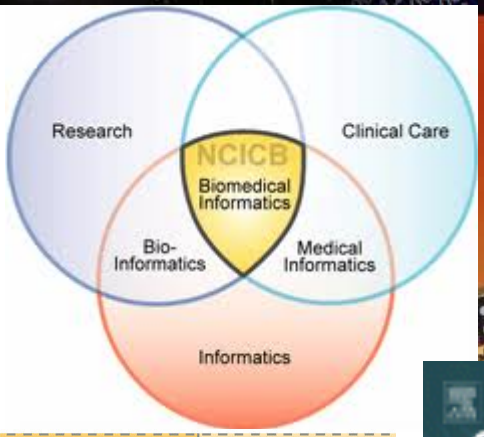
X-Informatics already Defined

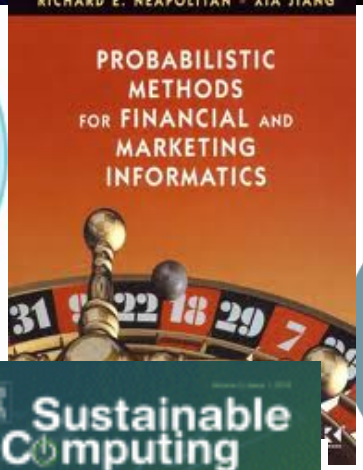
- Biomedical, Medical, Bio, Chem(istry), Health, Pathology, Wellness Informatics
- Life Style Informatics (from IT for Facebook to IT for life or Health – that's better Life Style Informatics))
- Astro(nomy), Climate, Earth Science, Energy, Environment, Radar, Sensor, Sustainability Informatics
  - Physics Informatics ought to exist but doesn't
- Social Informatics in our school
- Business, Wealth, Financial, Marketing Informatics
- Security (also in School), Crisis, Intelligence Informatics
- Policy Informatics (many X-Informatics impact policies)

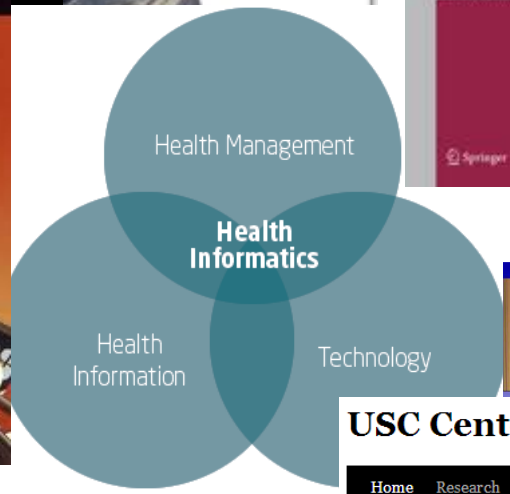


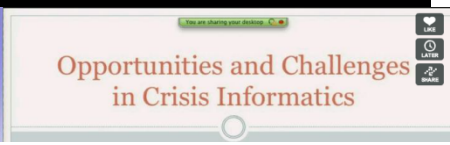


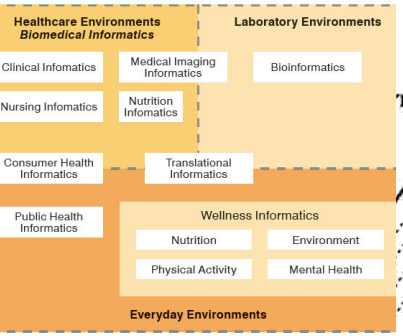


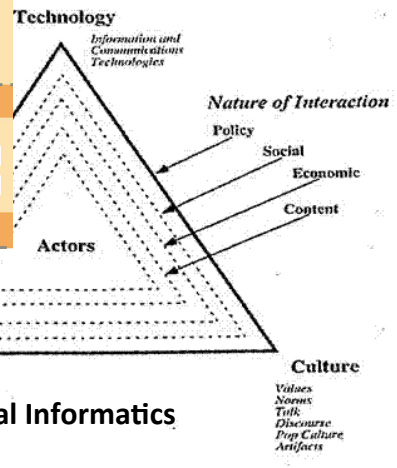








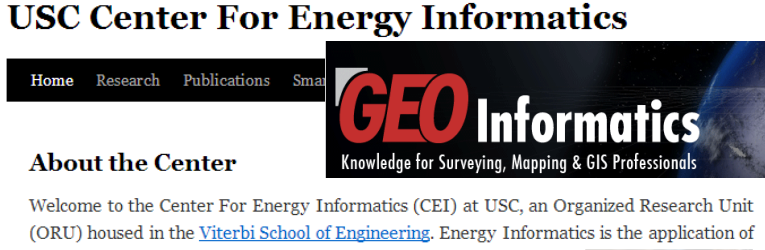


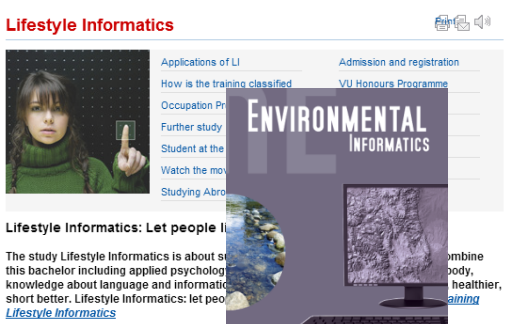


Noelia Penelope Greer (Ed.)

Business Informatics

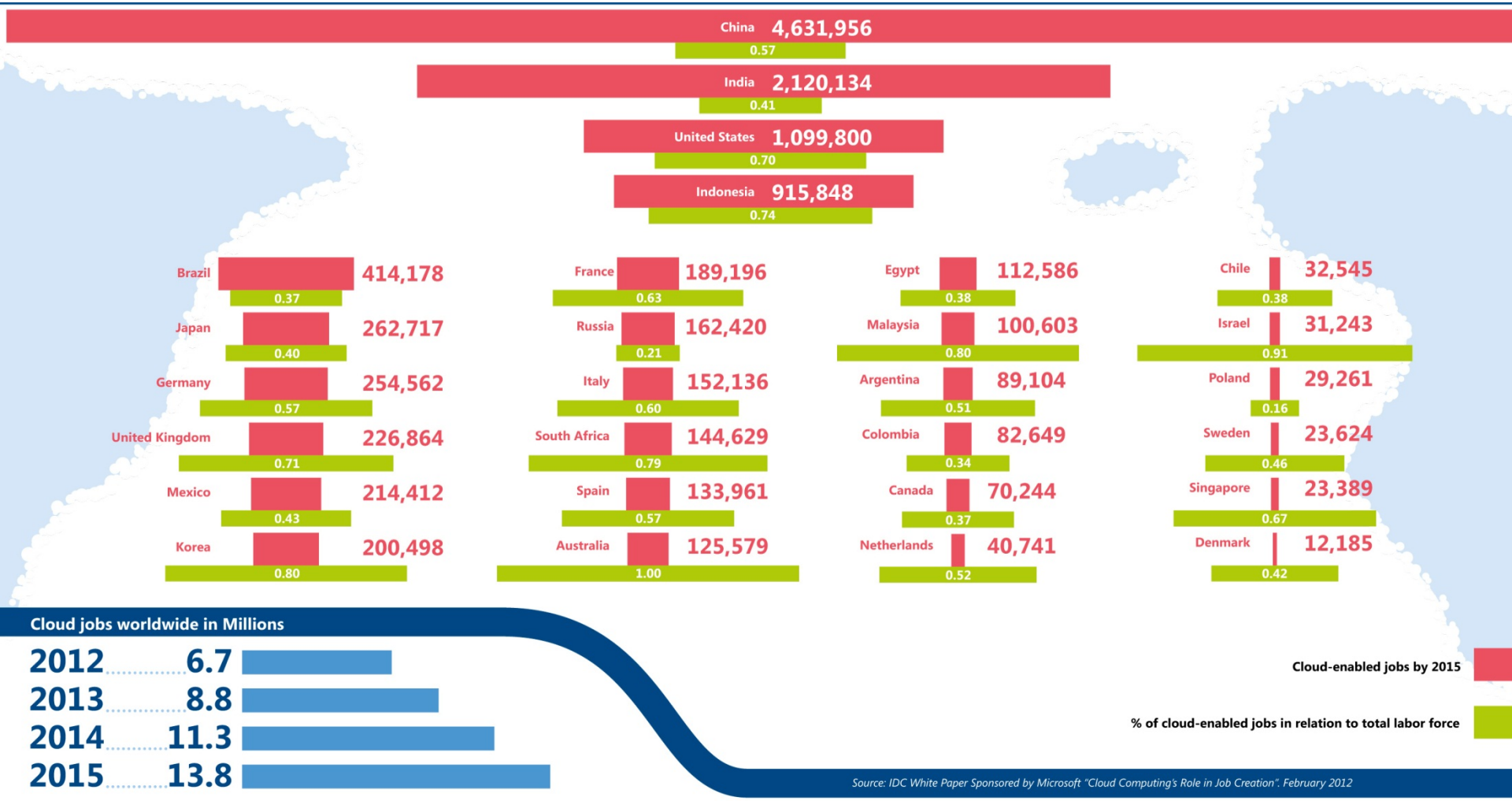
Information technology, Management,





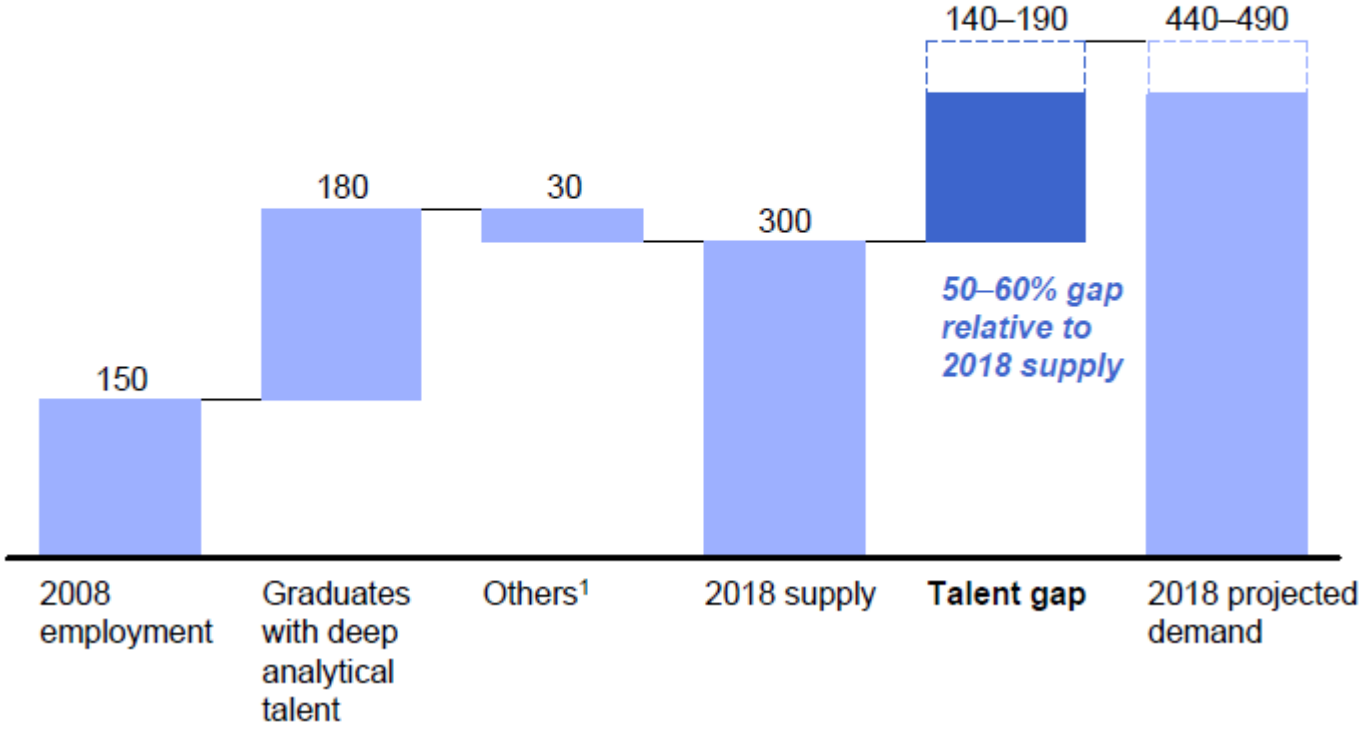
Jobs

Jobs v. Countries



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

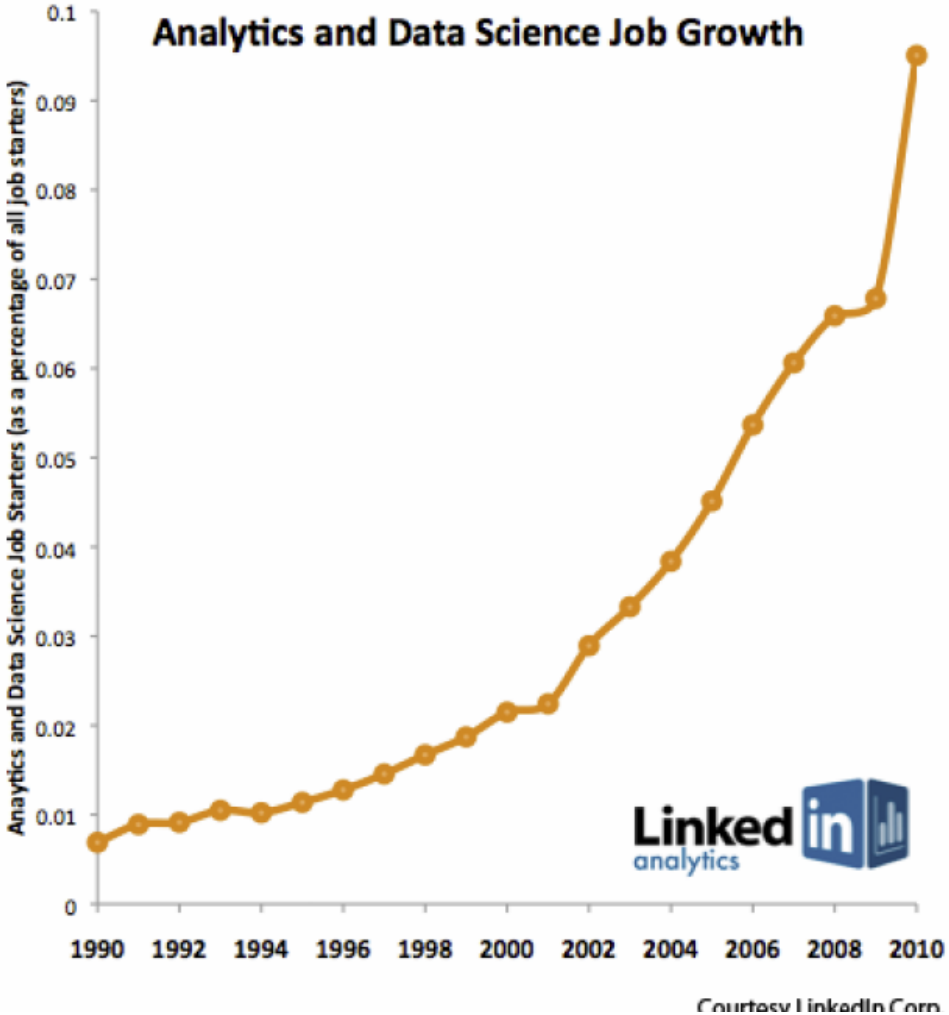
McKinsey Institute on Big Data Jobs



- 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.
- This course 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

Data Deluge
General Structure

Some 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?
- **LHC** 15 petabytes per year
- **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

Some Use Cases for Big Data

- Social media analytics – "People You May Know" at LinkedIn
- Voice analytics – Call center triage
- Text analytics – Voice of customer, sentiment analysis, warranty analysis
- Video analytics – Intelligence, policing, retail applications
- Genome data – what genetic profiles are associated with certain cancers?

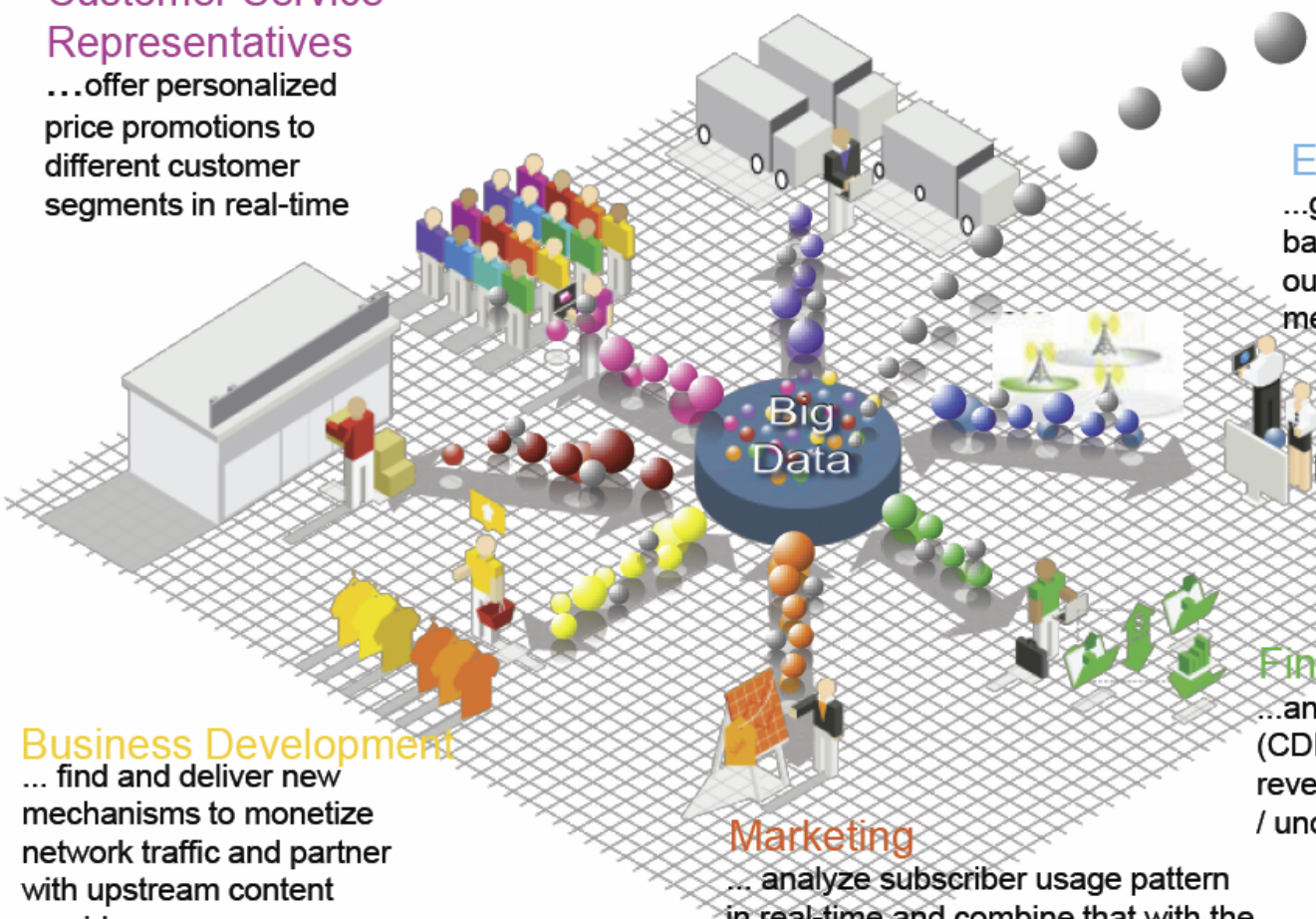
Tom Davenport Harvard Business School http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html Nov 2012

Anjul Bhambhri, VP of Big Data, IBM
http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

The new era of analytics delivers value across the enterprise

Network Operations
...identify network bottlenecks in real-time for faster resolution

Customer Service Representatives
…offer personalized price promotions to different customer segments in real-time



Google facebook YouTube twitter GPS
External Data

Executive Leaders
...get real-time reports and analysis based on data inside as well as outside the enterprise (web, social media etc.)

Business Analysts
... analyze social media buzz for the new services/offerings to gauge initial success and any course correction needed

Finance
...analyze all Call Detail Records (CDRs) to identify and reduce revenue leakage due to unbilled / underbilled CDRs

Business Development
... find and deliver new mechanisms to monetize network traffic and partner with upstream content providers

Marketing
... analyze subscriber usage pattern in real-time and combine that with the profile for delivering promotional or retention offers

The Rise of Big Data

What is it?	• Data that's too big (petabytes), too unstructured (not in rows and columns), or too diverse (mashups) to be stored and analyzed by conventional means (also relative)
Where does it come from?	• Internet/social media • Genomic analysis • Voice and video • Sensors everywhere
What is to be done with it?	• Structure, classify, and count it • Then analyze it (just as you would small data)

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

What's Different About Big Data?

The need for continuous flows of data, not stocks	• Stocks may be useful to develop models, but big data eventually requires a continuous process of analysis on moving data
Data scientists, not analysts	• IT "hacking" abilities in addition to the usual analyst attributes • Scientific and exploration focus • Closer to the product or process
New ways of deciding and acting on it	• It just keeps on coming, so have to establish ongoing processes to manage or decide on it

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

What's Different About Big Data? (cont.)

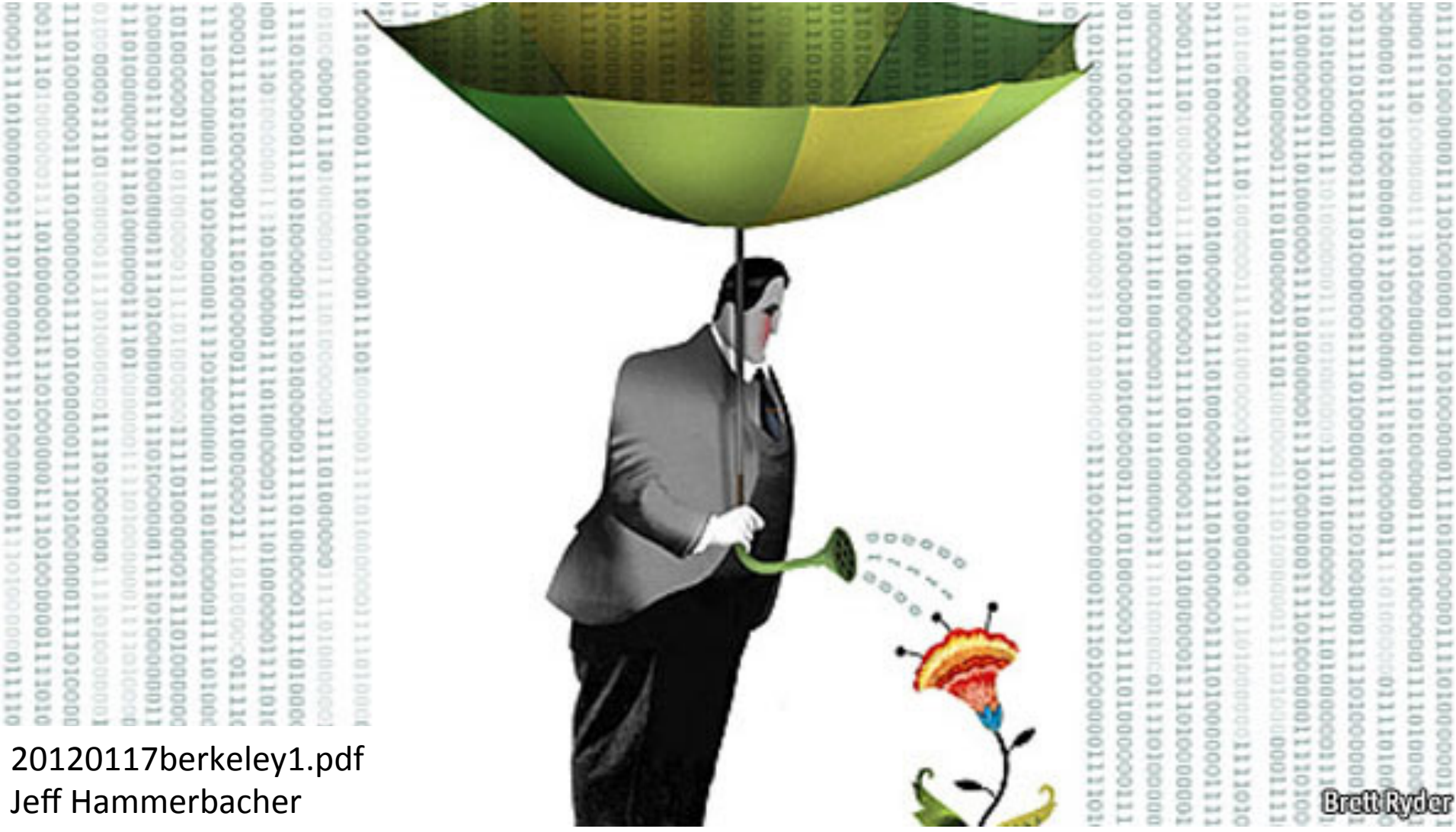
New technologies to manage it

- Filtering, structuring, and classification tools – MapReduce, Hadoop, etc.
- Content analytics tools--NLP
- Data redundancy management
- Cloud analytics
- Machine learning
- Open source everything, including R (it's capable, it's free, and that's what everybody coming out of school wants to use)

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

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

According to one estimate, mankind created 150 exabytes (billion gigabytes) of data in 2005. This year(2010), it will create 1,200 exabytes. Merely keeping up with this flood, and storing the bits that might be useful, is difficult enough. Analysing it, to spot patterns and extract useful information, is harder still. Even so, the data deluge is already starting to transform business, government, science and everyday life

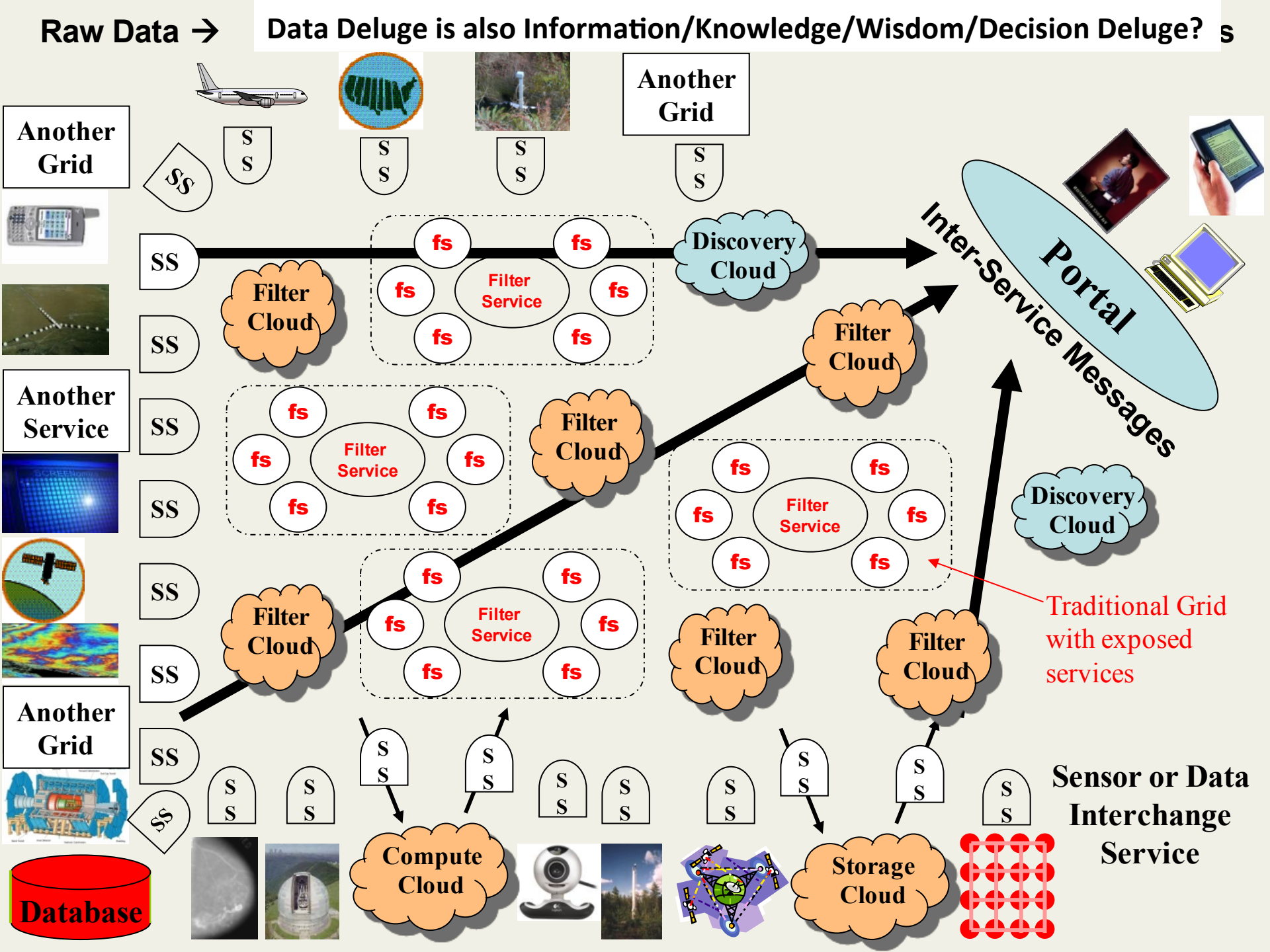


20120117berkeley1.pdf
Jeff Hammerbacher

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 Service
Another Grid
Another Grid
SS
Filter Cloud
fs
Filter Service
Discovery Cloud
Filter Cloud
Filter Cloud
Filter Cloud
Filter Cloud
Filter Cloud
Discovery Cloud
Portal
Inter-Service Messages
Traditional Grid with exposed services
Compute Cloud
Storage Cloud
Database
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

Data Deluge
Internet

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

Anjul Bhambhri, VP of Big Data, IBM
http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

Where is big data coming from?

12+ TBs of tweet data every day

? TBs of data every day

25+ TBs of log data every day

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

2+ billion people on the Web by end 2011

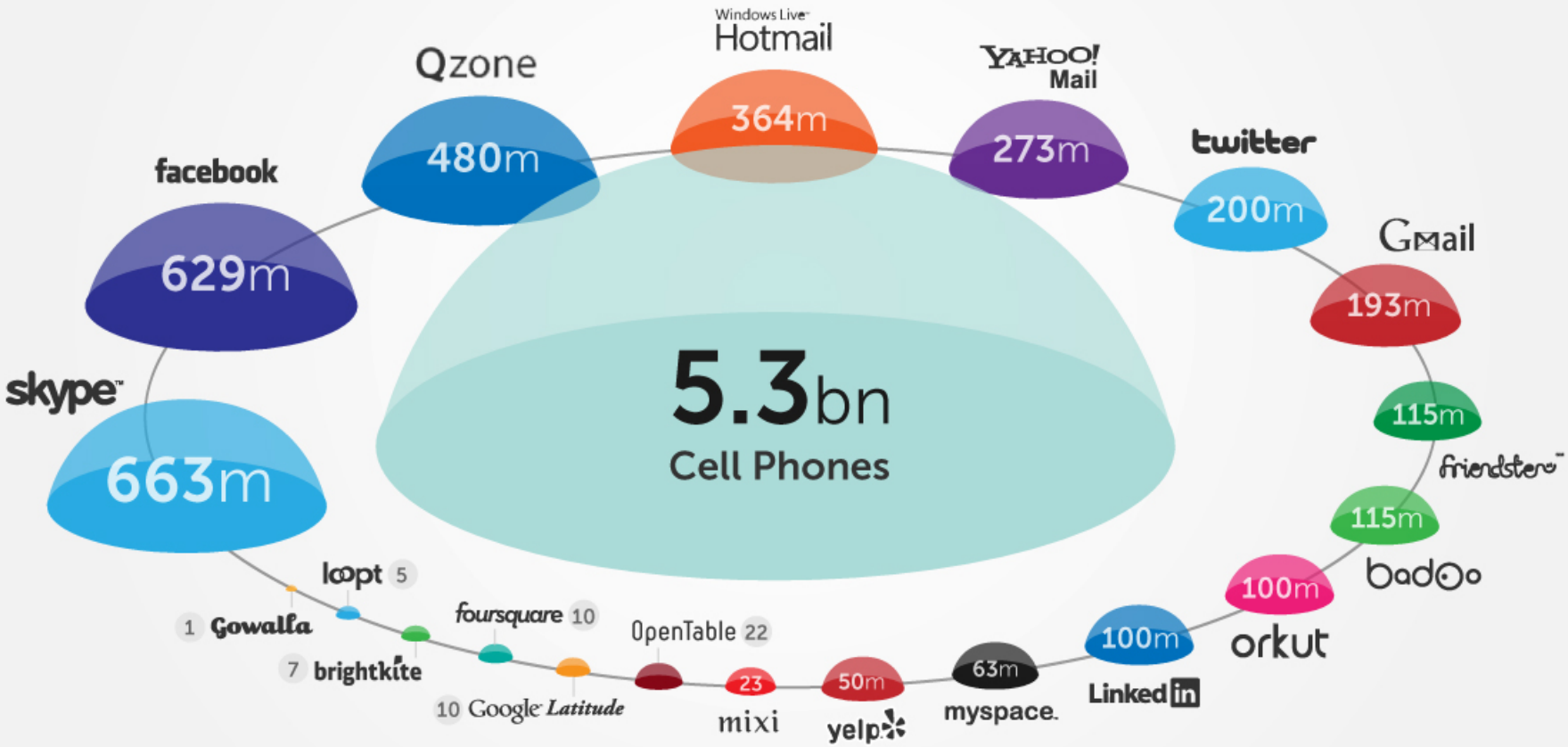
76 million smart meters in 2009… 200M by 2014

http://jess3.com/geosocial-universe-2/

the geosocial universe

MAY 2011

*Brought to you by **JESS3***



AMOUNT = ACTIVE USERS

Sources: TechCrunch | SocialMediaToday | Facebook | Wikipedia | Mashable | GeekoSystem | Daily Mail | LinkedIn | Loopt | SearchEngineLand | Brightkite | SocialTimes | Badoo | MobiThinking

Data Deluge Business

Vestas optimizes capital investments based on **2.5 Petabytes** of information.

- Model the weather to optimize placement of turbines, maximizing power generation and longevity.
- Reduce time required to identify placement of turbine from weeks to hours.
- Incorporate 2.5 PB of structured and semi-structured information flows. Data volume expected to grow to 6 PB.

Vestas

Anjul Bhambhri, VP of Big Data, IBM

Dublin City Centre Increases Bus Transportation Performance

Capabilities Utilized:

Stream Computing

- Public transportation awareness solution improves on-time performance and provides real-time bus arrival info to riders
- Continuously analyzes bus location data to infer traffic conditions and predict arrivals
- Collects, processes, and visualizes location data of all bus vehicles
- Automatically generates transportation routes and stop locations

Results:

- Monitoring 600 buses across 150 routes
- Analyzing 50 bus locations per second
- Anticipated to Increase bus ridership

Dublin City
Baile Átha Cliath

Anjul Bhambhri, VP of Big Data, IBM

Asian telco reduces billing costs and improves customer satisfaction.

Capabilities:

- Stream Computing
- Analytic Accelerators

Real-time mediation and analysis of **6B CDRs per day**

Data processing time reduced from **12 hrs to 1 sec**

Hardware cost reduced to 1/8th

Proactively address issues (e.g. dropped calls) impacting customer satisfaction.

CDR = Call Data Record

Anjul Bhambhri, VP of Big Data, IBM

Forces Shaping the Future

GE is a company that builds the machines that make the world work and has access to and deep understanding of the information that can make them work better

Internet of Things

1. Internet

Hyper-connectivity: a living network of machines data and people

Internet of things: more devices tap into the Internet than people on Earth to use them

2. Intelligent Machines

Increasing system intelligence through embedded software

Rise of machines: networked devices overtook the global population in 2011

3. Big Data

Democratization of data

Data overload: 2.5 quintillion bytes of data created every day

4. Analytics

Generating data-driven insights

Enhancing asset performance by detecting & predicting forecasts

Algorithms on installed base

10011010
10101010
10101010
11000101

imagination at work

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

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

0MW are equipped with Blade Health Monitoring capabilities

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

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

Trend: Analytical Software Solutions

- Integrated software solutions for business problems

Problem	Industry
Fraud detection	Banking
Credit and operational risk	Banking
Credit scoring	Financial services
Warranty analysis	Manufacturing
Customer retention	Telecommunications
Markdown optimization	Retail

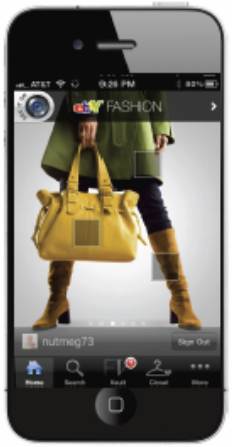
- What are the opportunities for statisticians?

http://www.hsph.harvard.edu/ncb2011/files/ncb2011-z03-rodriguez.pptx

http://fisheritcenter.haas.berkeley.edu/Big_Data/Williams.pdf

every
49 minutes
a Ford Mustang
is sold

every
5 seconds
a cell phone
is sold



every
6 seconds
a pair of shoes
is sold

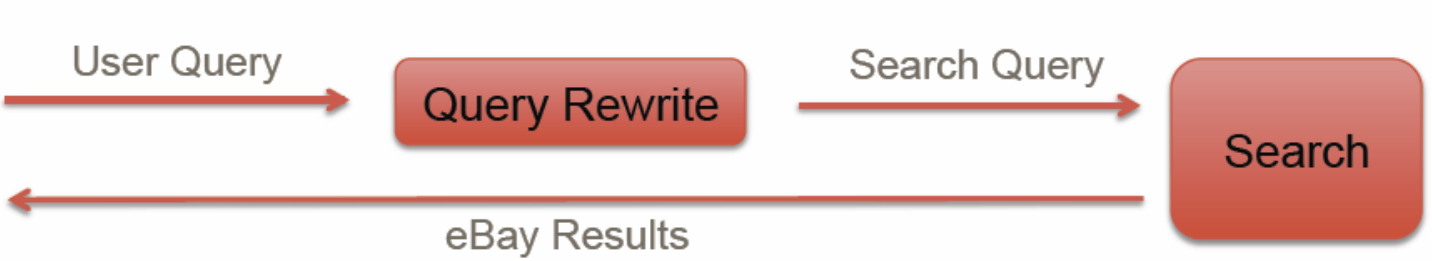
108+ million	Active buyers and sellers worldwide
250+ million	Queries every day to the eBay search engine
350+ million	Live global listings
20+ petabytes	Of data in our Hadoop and Teradata clusters
2 billion	Page views each day
75 billion	Database calls each day

BIG DATA IS TRANSFORMATIONAL

- Big data informs on:
  - *Patterns*
  - Anomalies and outliers
  - Generalizations
  - *Predictions*
  - *Relative performance*
  - An holistic customer picture
- Vast array of applications at eBay:
  - *Product development*, *A/B testing*, system performance, fraud and risk detection, *purchase prediction*, customer support, buyer demand, seller intelligence, financial performance, …

PATTERNS: QUERY REWRITES

- In 2010, our search engine was very literal: it matched exactly what you typed
- We're on a journey to make it more intuitive, so it does a great job of understanding user intent and finding all of the relevant results
- Idea: Mine our extreme data, look for patterns, and use these to map words in user queries to **synonyms** and **structured data** associated with items for sale at eBay



PATTERNS: QUERY REWRITES …

HOW DO BUYERS PURCHASE THE PILZLAMPE?

- It turns out, they try one (or more) of a few things:
  - Type pilzlampe, and purchase
  - Type pilzlampe, … , pilz lampe, and purchase
  - Type pilzlampe, … , pilzlampen, and purchase
  - Type pilz lampen, … , pilzlampe, and purchase
  - …

PATTERNS: QUERY REWRITES …

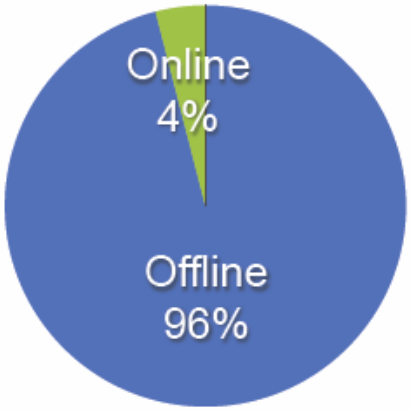
- From our big data mining:
  - We automatically discover that *pilz lampe* and *pilzlampe* are the same
  - We also discover that *pilz* and *pilze* are the same, and *lampe* and *lampen* are the same
- From these patterns, we rewrite the user's query *pilzlampe* as:

  pilzlampe OR "pilz lampe" OR "pilz lampen" OR pilzlampen OR "pilze lampe" OR pilzelampe OR "pilze lampen" OR pilzelampen

IT'S JUST COMMERCE

There's no longer online and offline

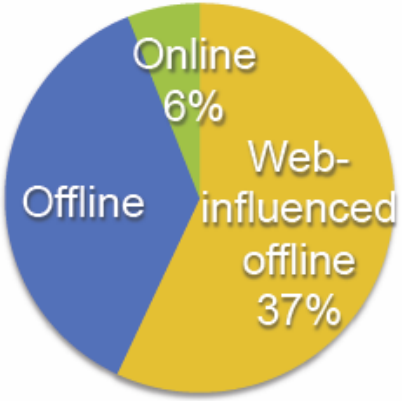
Yesterday



2008 = $325B

Source: Forrester, Euromonitor and Economist Intelligence Unit

Today



Source: Forrester

Tomorrow

"Commerce"
2013 = $10T

Source: Economist Intelligence Unit