Big Data: Introduction and Applications

August 20, 2015
HKU-HKJC ExCEL3 Seminar
Michael Chau, Associate Professor
School of Business, The University of Hong Kong

Over the past years…

- Data storage has grown exponentially
- Computation capacity has risen sharply
- Network bandwidth has increased greatly

Big Data

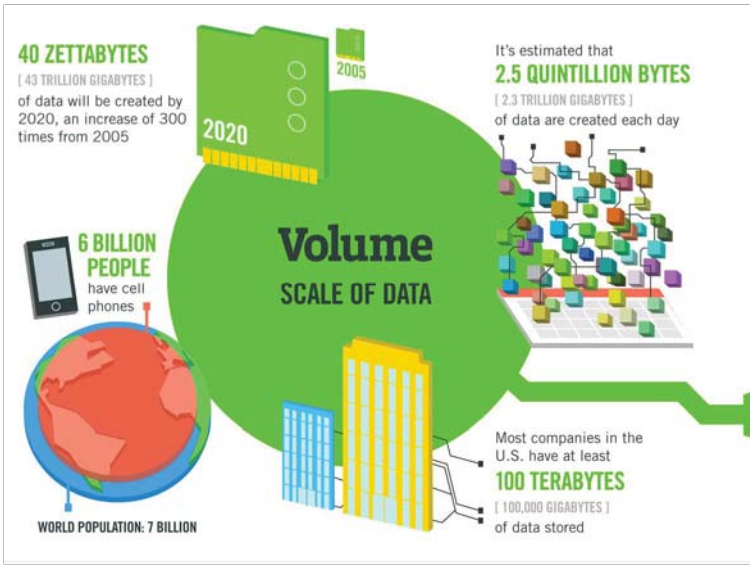
- Massive amount data are being generated at an unprecedented speed from various sources:
  - online transactions
  - mobile applications
  - sensors
  - images, audio, video
  - social media including blogs, weibos, facebook, and forums

Big Data

- Ample opportunities for business organizations and governments to provide better services and gain managerial and strategic insights by gathering, cleaning, and analyzing these "Big Data".

What is Big Data?

- It is not just big!
- Four Vs
  - Volume
  - Velocity
  - Variety
  - Veracity



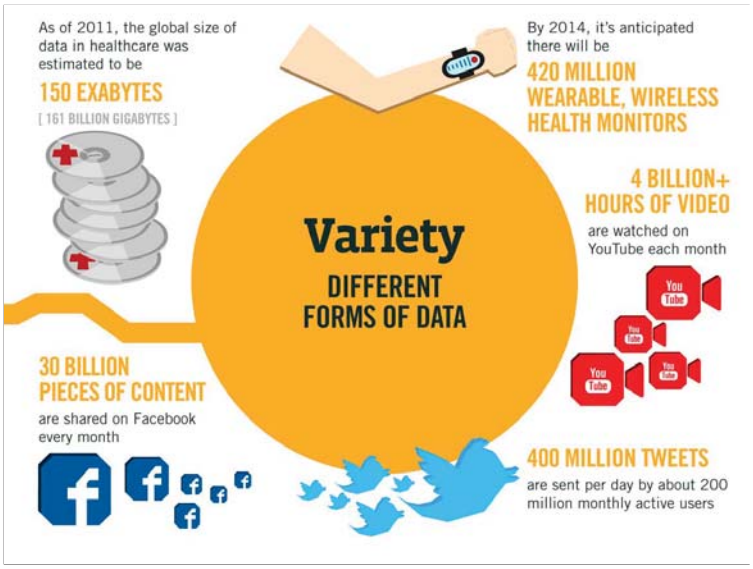
The Data Size Is Getting Big, Bigger…

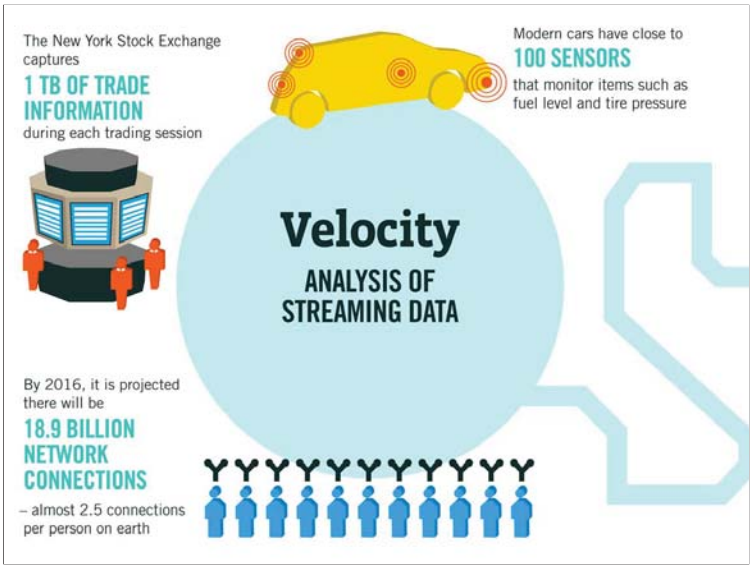
- Hadron Collider - 1 PB/sec
- Boeing jet - 20 TB/hr
- Facebook - 500 TB/day.
- YouTube – 1 TB/4 min.
- The proposed Square Kilometer Array telescope (the world's proposed biggest telescope) – 1 EB/day

Names for Big Data Sizes

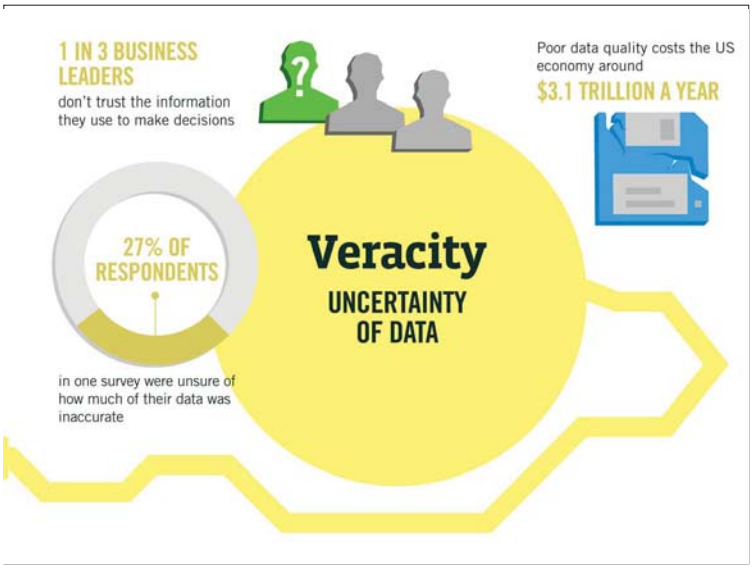
Name	Symbol	Value
Kilobyte	kB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^{12}
Petabyte	PB	10^{15}
Exabyte	EB	10^{18}
Zettabyte	ZB	10^{21}
Yottabyte	YB	10^{24}
Brontobyte*	BB	10^{27}
Gegobyte*	GeB	10^{30}

*Not an official SI (International System of Units) name/symbol, yet.





Google Analytics

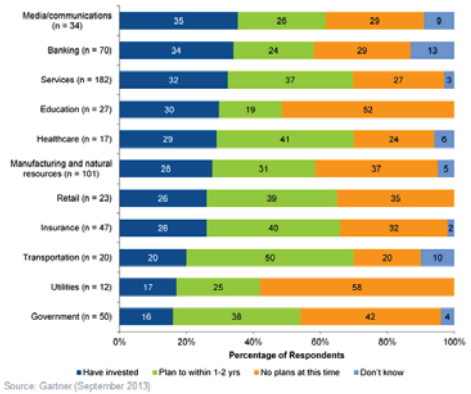


Mobile Devices

Social Media

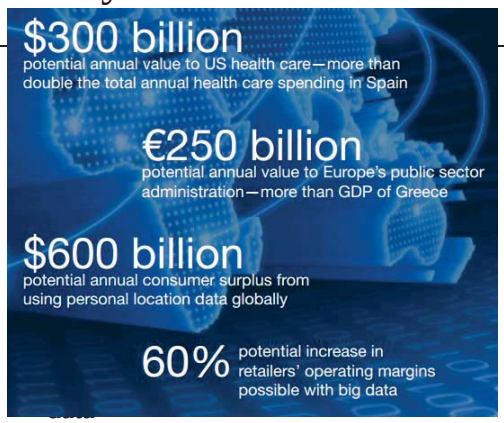
- An important component of Big Data
  - Social networking sites
  - Blogs, microblogs
  - Online reviews, discussion forums
  - Online news aggregator
- People reveal themselves on social media
  - Demographics, preferences, habits, family ties, social ties
  - Images, videos
- Unstructured but rich in content
- Increasingly used in marketing research

Big Data Investment by Industry



Source: Gartner (September 2013)

McKinsey estimates….



Big Data Investment by Region

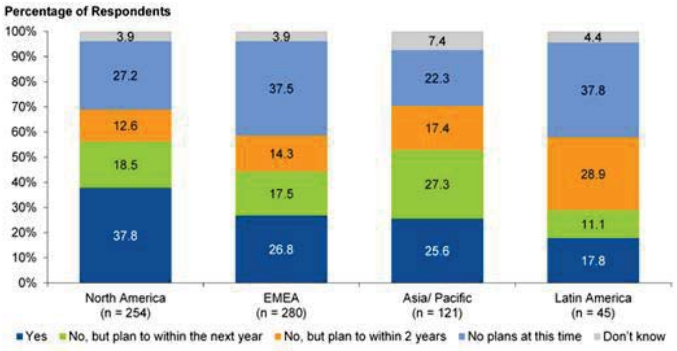
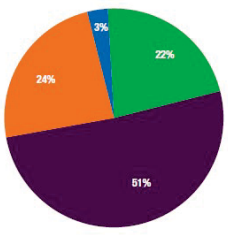
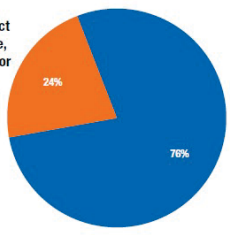


Figure 1: Investments in Big Data analytics are strong

1. What portion of your overall technology expenditure is made in Big Data analytics?



2. Do you expect this to increase, stay the same or decrease over the next year?



Source: Accenture "Industrial Internet Insights Report For 2015"

How governments see big data?

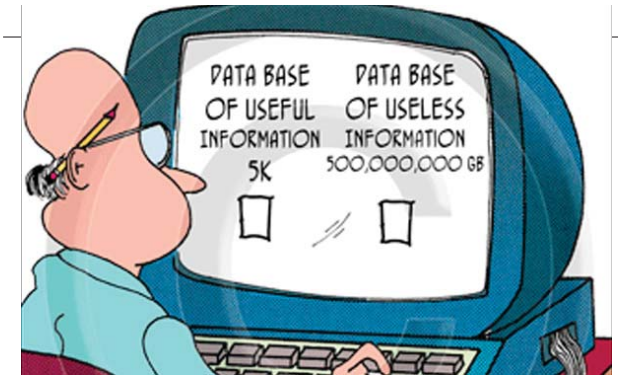
- USA: The White House Big Data R&D Initiative was launched in 2012
  - Extract knowledge and insights from large and complex collections of digital data
- UK: Formed the Alan Turing Institute for big data research
- South Korea: The Big Data Initiative was launched in 2011

How governments see big data?

- United Nations' Global Pulse project
  - Analyze social media and big data for sustainable development and humanitarian action

Big Data Considerations

- You can't process the amount of data that you want to because of the limitations of your current platform.
- You can't include new/contemporary data sources (e.g., social media, RFID, Sensory, Web, GPS, textual data) because it does not comply with the data schema.
- You need to (or want to) integrate data as quickly as possible to be current on your analysis.
- You want to work with a schema-on-demand data storage paradigm because the variety of data types.
- The data is arriving so fast at your organization's doorstep that your analytics platform cannot handle it.
- …



Source: The Storage Alchemist

Critical Success Factors for Big Data Analytics

- A clear business need (alignment with the vision and the strategy)
- Strong, committed sponsorship (executive champion)
- Alignment between the business and IT strategy
- A fact-based decision-making culture
- A strong data infrastructure
- The right analytics tools
- Right people with right skills

Fundamentals of Big Data Analytics

- Big Data by itself, regardless of the size, type, or speed, is worthless
- Big Data + "big" analytics = value (the 5th V)
- With the value proposition, Big Data also brought about big challenges
  - Effectively and efficiently capturing, storing, and analyzing Big Data
  - New breed of technologies needed (developed or purchased or hired or outsourced …)

Critical Success Factors for Big Data Analytics

Models and Technologies

- Traditional
  - Data mining, data warehouse
- Big Data Analytics
  - Unstructured data from multiple sources
  - Real-time analytics
  - Complex statistical analysis
  - Distributed computing

Big Data Vendors

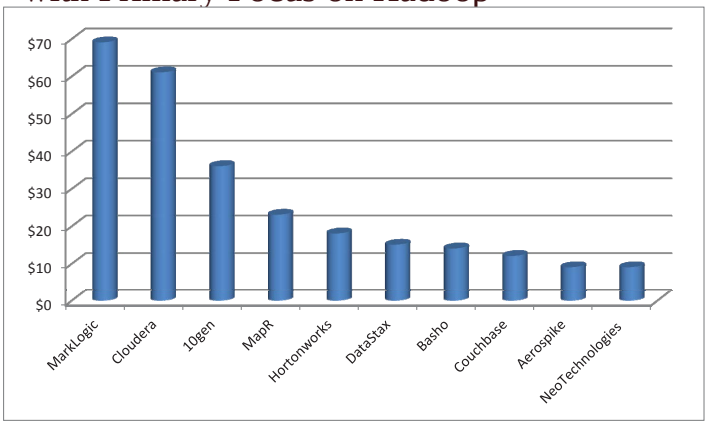
- Big Data vendor landscape is developing very rapidly
- A representative list would include
  - Cloudera - cloudera.com
  - MapR – mapr.com
  - Hortonworks - hortonworks.com
  - Also, IBM (Netezza, InfoSphere), Oracle (Exadata, Exalogic), Microsoft, Amazon, Google, …

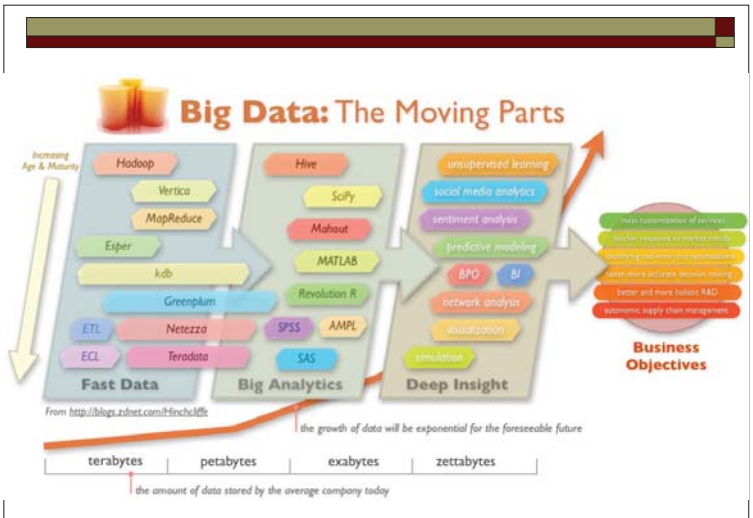
Software, Hardware, Service, ...

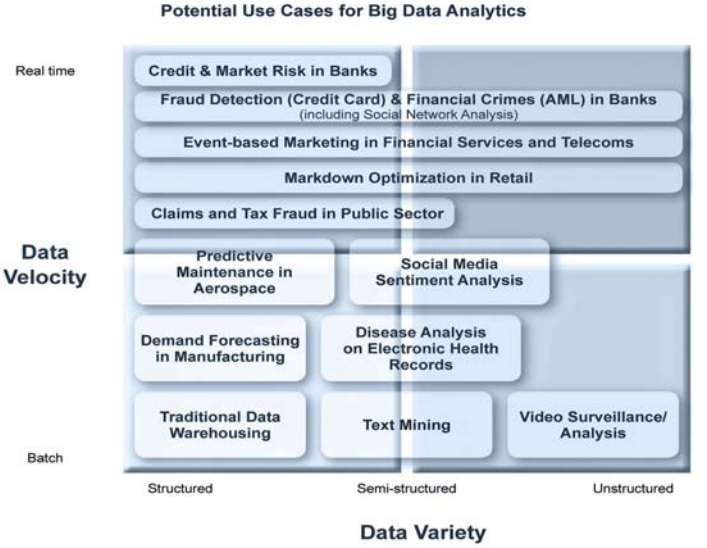
Models and Technologies

- Social media analytics
  - Text and sentiment analysis
  - Social network analysis
- Predictive modeling
  - Statistical method
  - Artificial intelligence/data mining
- Visualization

Top 10 Big Data Vendors with Primary Focus on Hadoop







Example: Amazon

- Predicts what customers will buy and manage their inventory
- "Customers who bought this item, also bought these…"
- Anticipatory Shipping: shipping an item to a customer in anticipation that this customer will order that product – will it work?

Example: Hewlett-Packard

- Analyzes data of 330,000 employees to predict who have a high risk of leaving the job
- Results in an estimated saving of $300 million

Example: Google Flu Trend

- Google Flu Trend: Predicts influenza breakout based on the occurrences of relevant search terms in search data
- Successfully predicts regional outbreaks of flu up to 10 days before they were reported by the CDC (Centers for Disease Control and Prevention).

Example: Germany Soccer

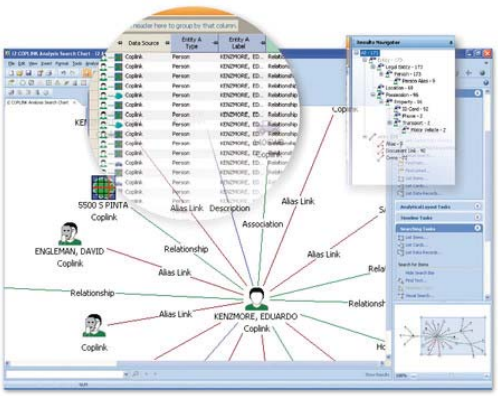
- Match Insights: Collects and analyzes massive amounts of player performance data (including video data)

Example: Macy's

- Analyzes a vast amount of customer data ranging from visit frequencies and sales to style preferences and personal motivations.
- Adjusts pricing in near-real time for millions of items, based on demand and inventory
- Sends targeted and customized direct mailings to customers

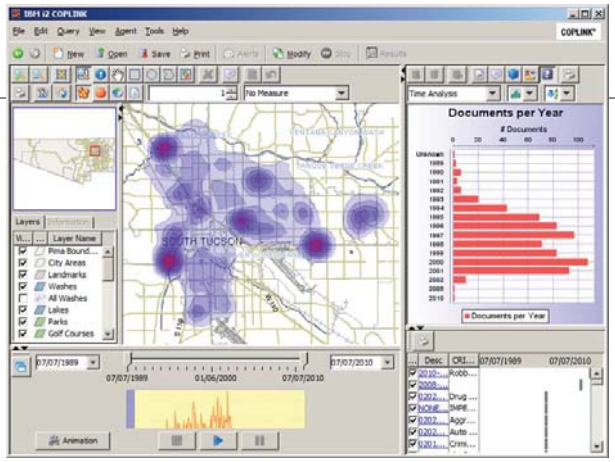
Example: US Law Enforcement

- Challenges
  - Isolated databases
  - Lack of analytics capability
- COPLINK system
  - Implemented in many police departments in the US
  - Database linking
  - Association mining



37

40



38

Social Media Analytics

- Social Media
  - Social interactions among people
  - People freely reveal their feelings
  - Highly dynamic

41

Example: US Law Enforcement

- Predictive policing
  - Predict time and location with high-probability of criminal activities
  - Prevent crimes before they happen
- Crime mapping
  - RAIDS Online
  - Connects law enforcement with the community to reduce crime and improve public safety

39

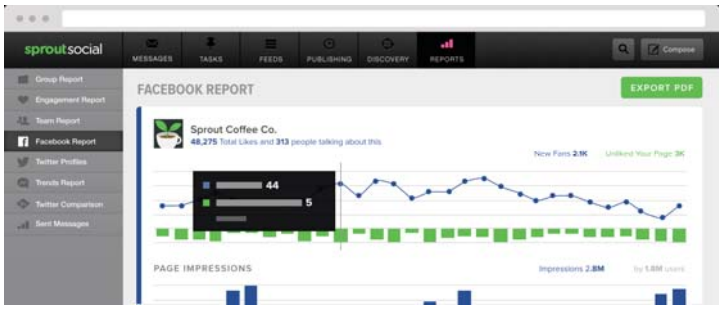
Different Types of Social Media

- Collaborative projects (e.g., Wikipedia, Wiktionary)
- Blogs and microblogs (e.g., Twitter, Weibo)
- Content communities (e.g., YouTube)
- Social networking sites (e.g., Facebook)
- Virtual game worlds (e.g., World of Warcraft)
- Virtual social worlds (e.g., Second Life)

42

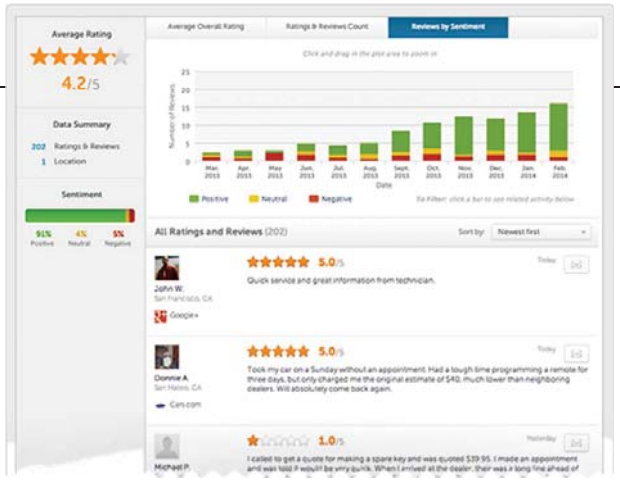
2015 Facebook usage

Social Media Analytics

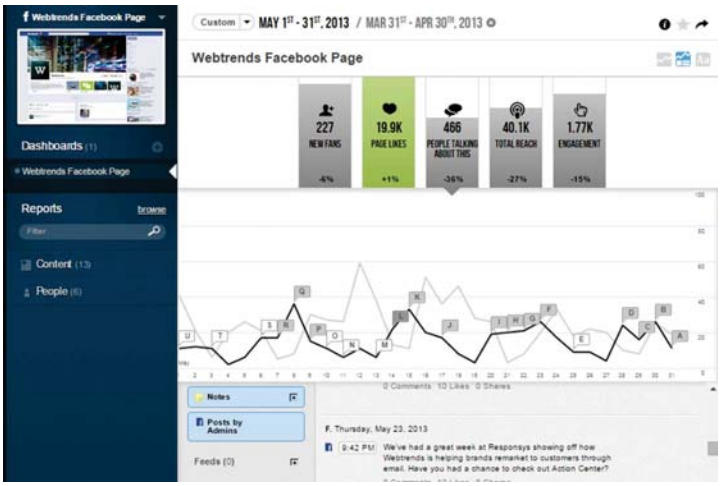


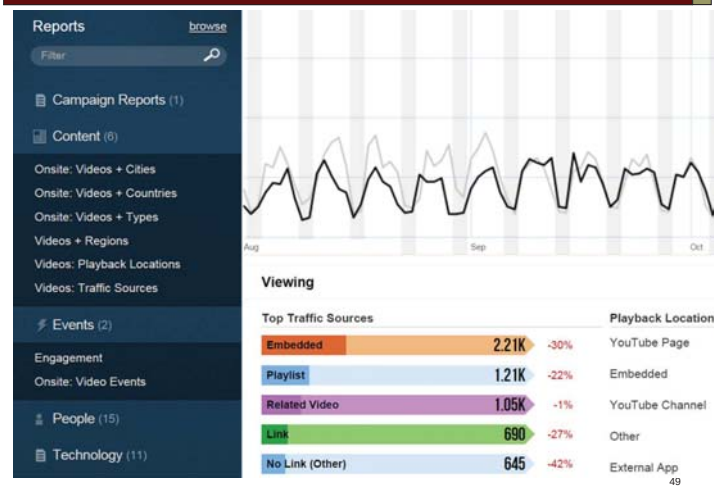
Social Media Analytics Tools and Vendors

ATTENSITY
reputation.com
salesforce
radian6
sproutsocial
CONVERSEON
Making Social Work
sysomos
SOSHIO
crimson hexagon
collective intellect
TwitSprout
webtrends
k-matrix



Social Media Analytics



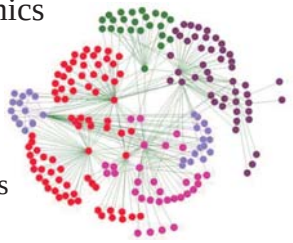


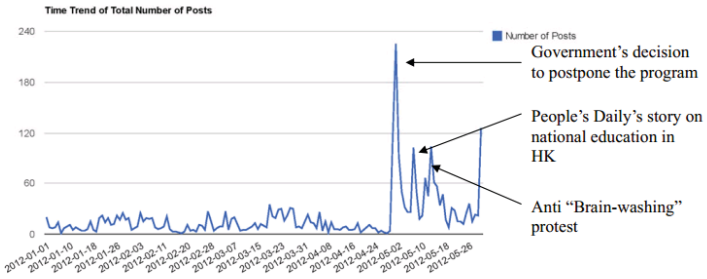
Application Case: HK Online Public Opinion

- Case Study: National education debate
- Controversy on the implementation of national education in secondary and primary schools
- Study period lasted from January 1, 2012 to May, 31, 2012.

Social Network Analysis

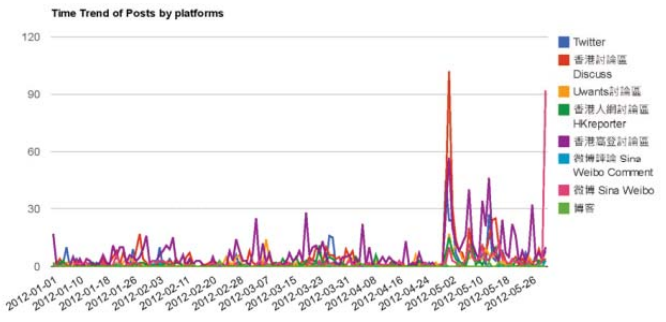
- Social Network - social structure composed of individuals linked to each other
- Analysis of social dynamics
- Identify
  - Opinion leaders
  - Bridges
  - Clusters and social circles

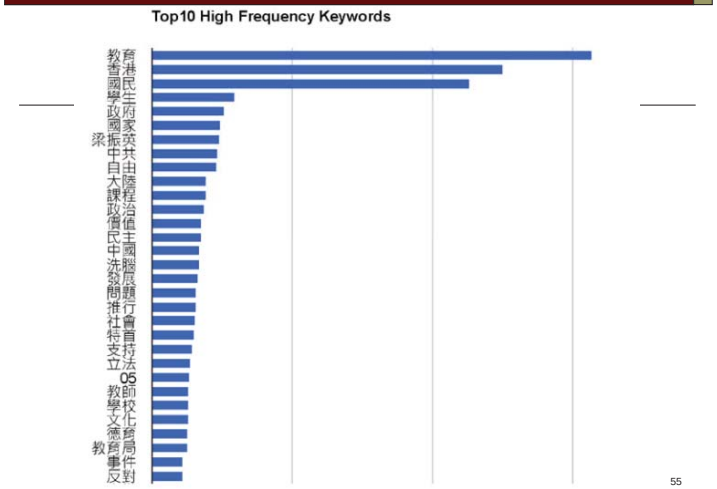


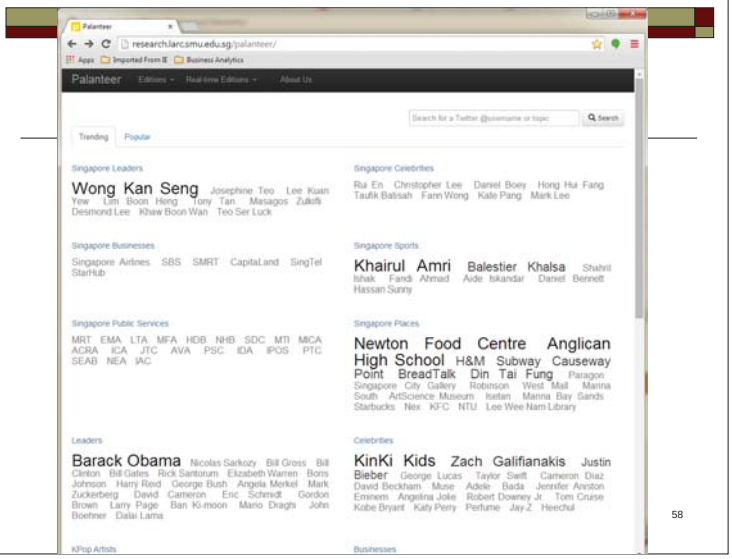


Application Case: HK Online Public Opinion

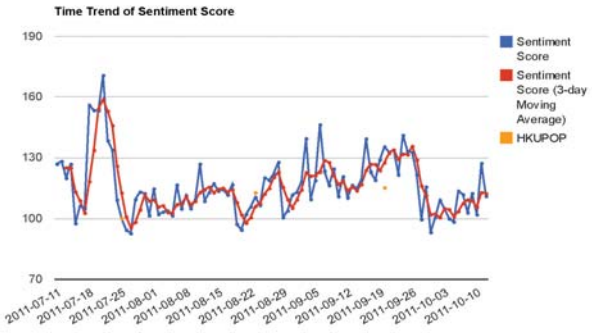
- Challenges
  - Online platforms become a important venue to understand public opinions
  - Too much information for analysis and monitoring
- Data collection from major Hong Kong based online opinion platforms
  - Discussion forums
    - Uwants, Discuss HK, Golden, HK Reporter
  - Twitter
  - Sina weibo
  - Blogs
  - Facebook pages, groups, and events

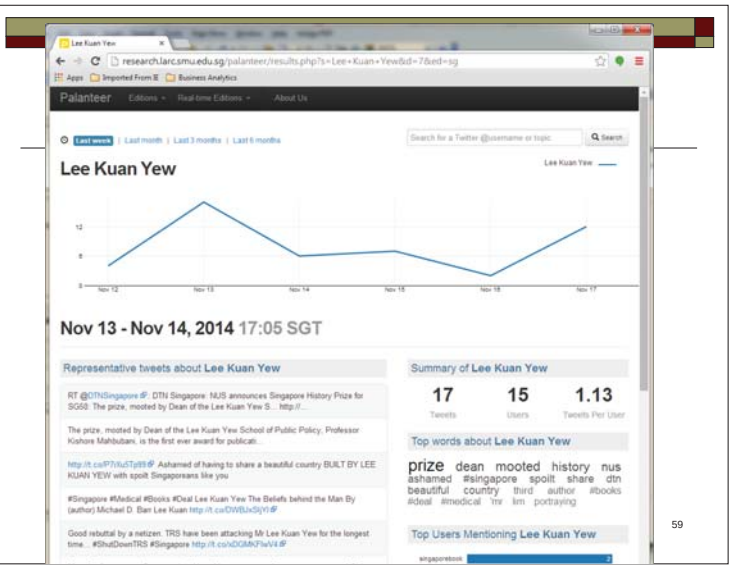






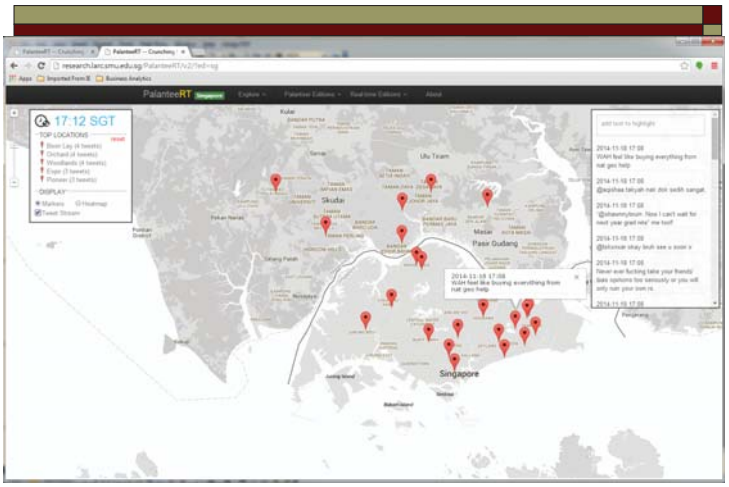
Prediction of public sentiment

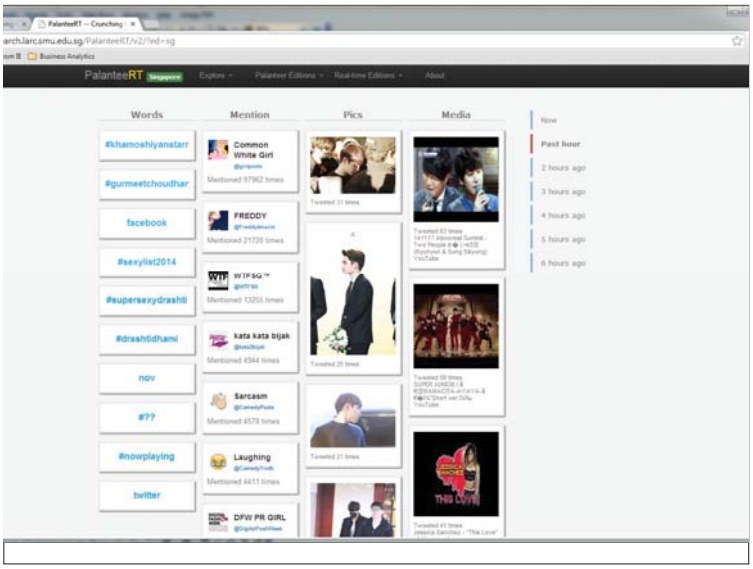


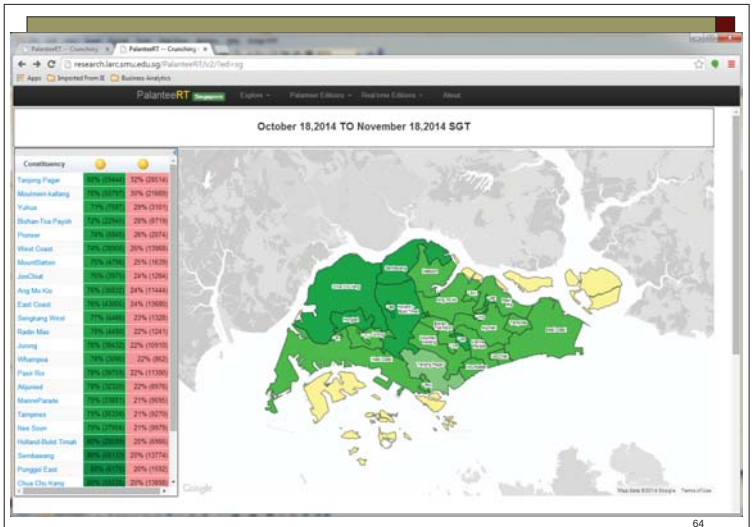


Application Case: Singapore Social Media Analytics

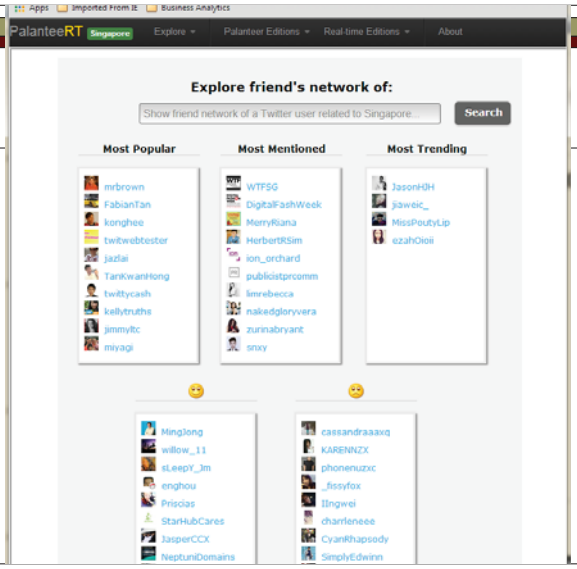
- Analyze and visualize social media
  - http://research.larc.smu.edu.sg/palanteer/
  - http://research.larc.smu.edu.sg/palanteert/







64

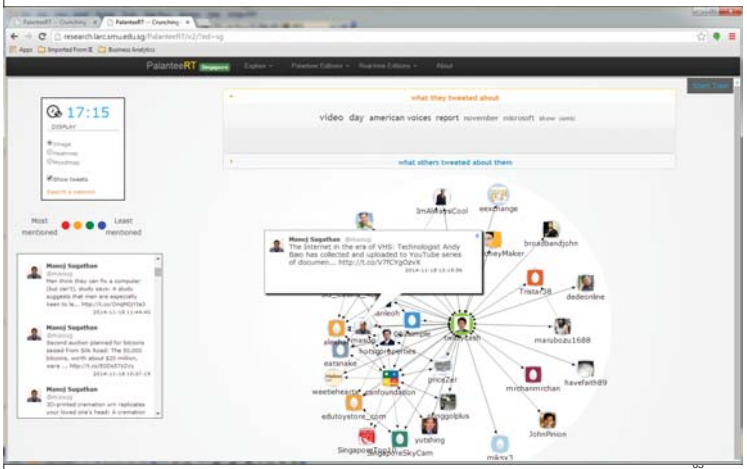


62

So, what big data can do?

- Reduce cost
- Improve services
- Improve design and planning

65



Questions and Discussions