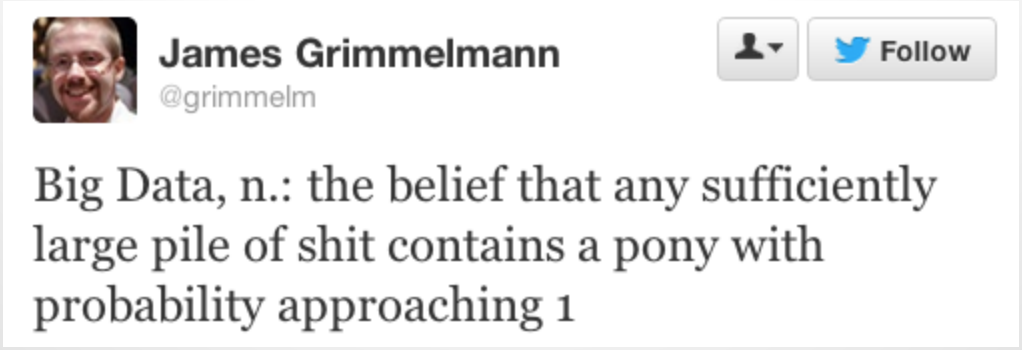
Big Data - the big picture

Group O



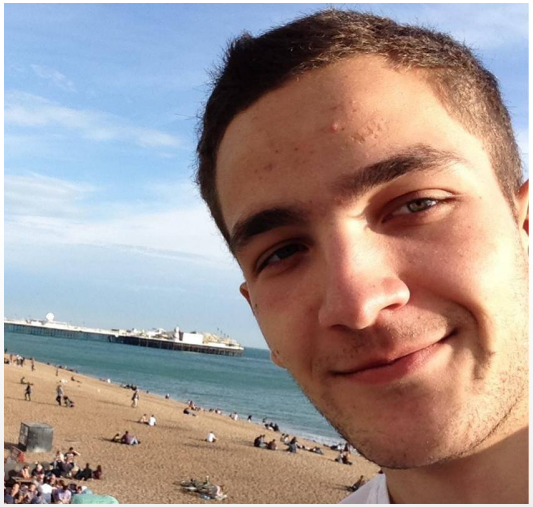
Us

Phill Raynsford

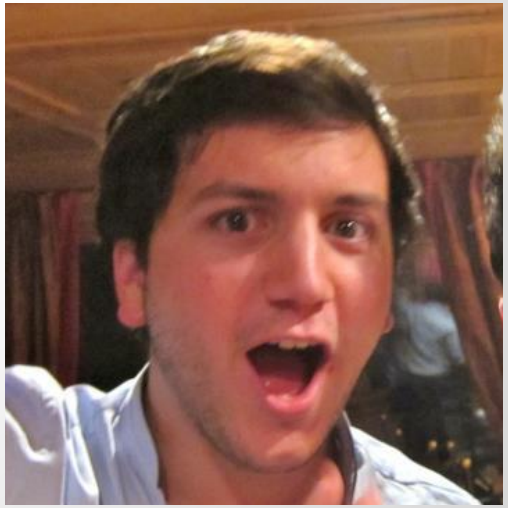
Callum McGregor

Giacomo Meanti

Nick Thuringer



Damyan Rusinov



Hector Goasguen

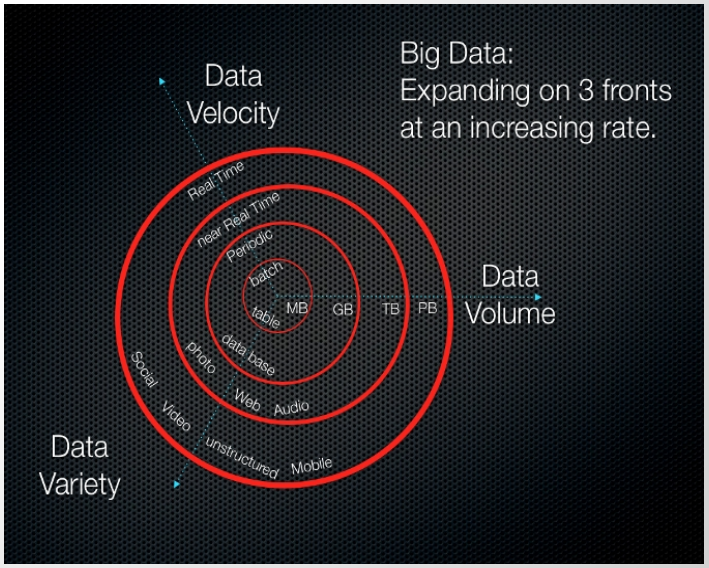
What is Big Data?

- **Big data is the term for a collection of data sets**
  - The data is so large and complex that it becomes hard to process using traditional database and software techniques. [1]
  - This data needs to be captured, converted, stored, shared, transferred, analysed and visualized. [2]
  - This term is not always used for the Data itself, but can sometimes be a reference to the technology used to collect the data or sort it.

What is Big Data?

- **There are 3 V's to Big Data [3]**
  - These are called dimensions, that show the
  - constant expansion of Big Data.
    - Volume: This is the amount of the data collected.
    - Variety: The types of data being collected.
    - Velocity: The rate at which the data is processed.

- There are also 2 potential additional “dimensions” [4]
  - Variability: How spread out or closely clustered a set of data is.
  - Complexity: The connection and correlation between the relationships and hierarchies of the multiple data linkages being sent.



Capitalizing on Big Data:

Strategies outperforming companies are taking to deliver results

Leaders are **166%** more likely to **make most decisions based** on data

And they are **2.2x** more likely to have **formal career path** for analytics

75% of Leaders cite **growth as the key source of value** from analytics

Leaders **measure the impact** of analytics investments

Leaders have **predictive analytics** capabilities

Leaders have some form of **shared analytics resources**

Join the conversation on Twitter at #ibmanalytics and follow @IBMIBV

Source: *Analytics: A blueprint for value - Converting big data and analytics insights into results*
IBM Institute for Business Value.

ibm.co/9levers

IBM

Social Media

Relationship between Big Data & Social Media

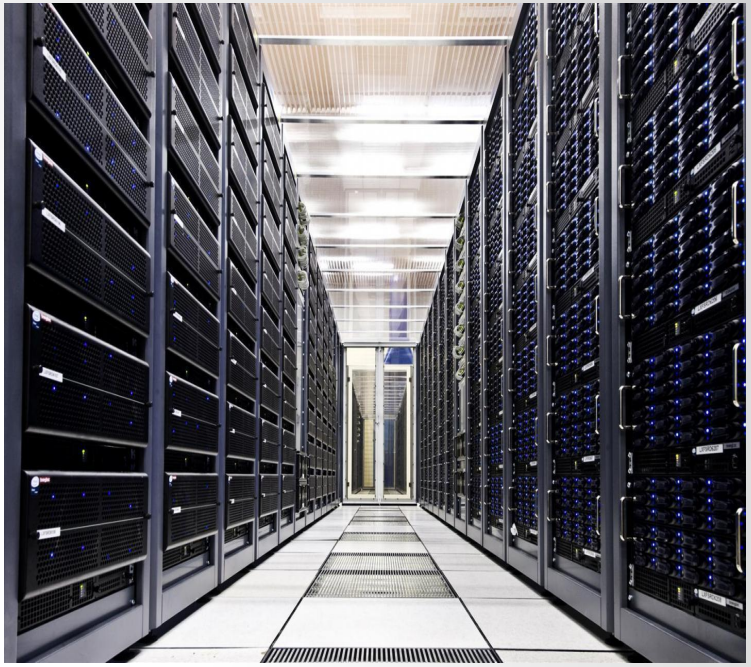
- Heaven for Business
- Social media generates tens of petabytes a day (1PB = 1000 terabytes)
- In comparison, just a few decades ago, all the information was just a few petabytes
- Builds of reputations

Story tell

- Unstructured
- Structured
- Hadoop
- petabytes of data
- SQL

Science

CERN and big data



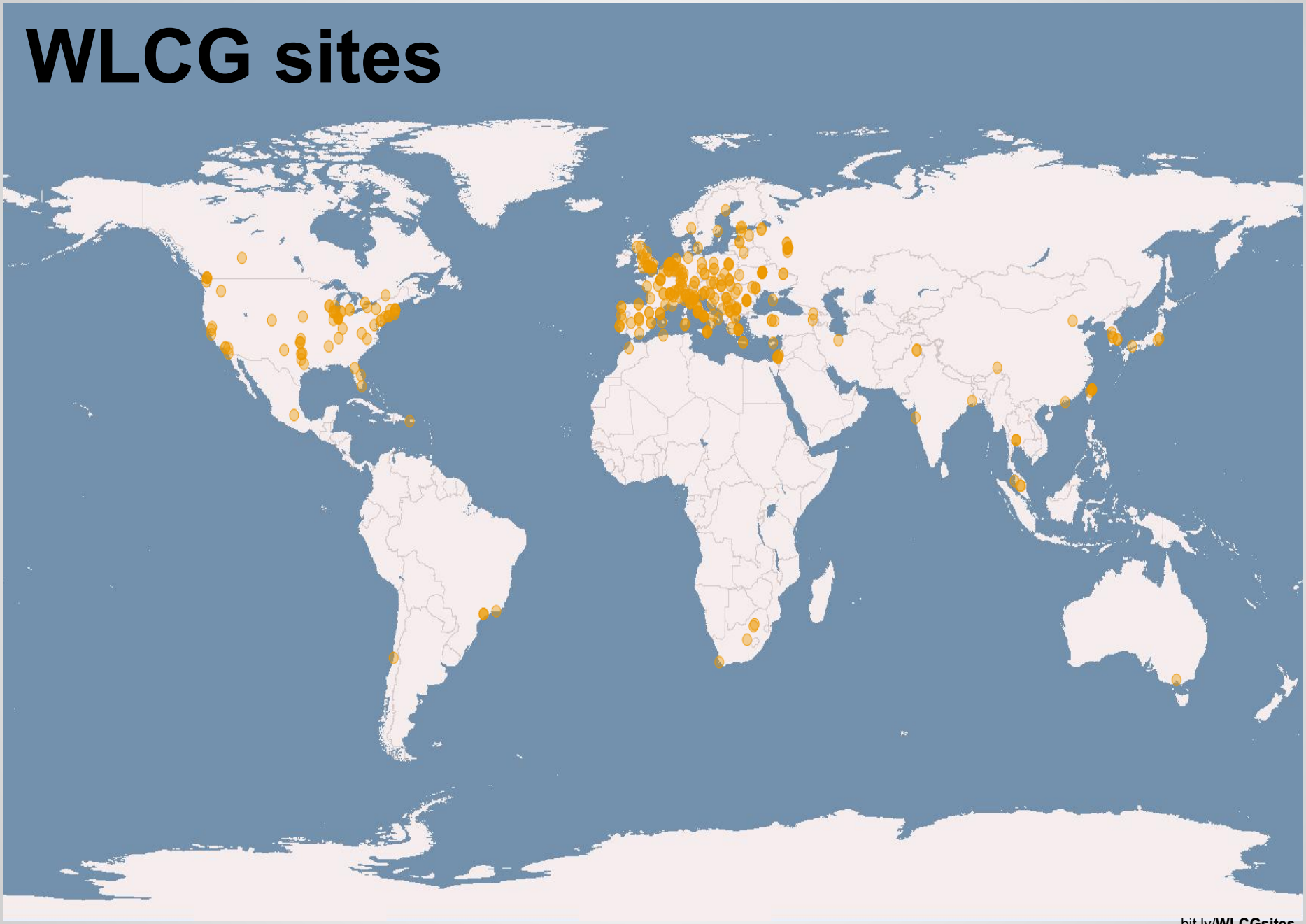
- millions of collisions every second ≈ one petabyte of data produced
- we can’t store that much! reduced to about 25 petabytes a year
- CERN Data Centre surpassed 100PB last year (75PB from LHC experiments)

Worldwide LHC Computing Grid

- provides computing power and backup storage
- more than 170 collaborating centres
- 36 countries
- 1.5 million jobs a day
- equivalent to a single computer running for 600 years

http://home.web.cern.ch/about/computing/worldwide-lhc-computing-grid

WLCG sites

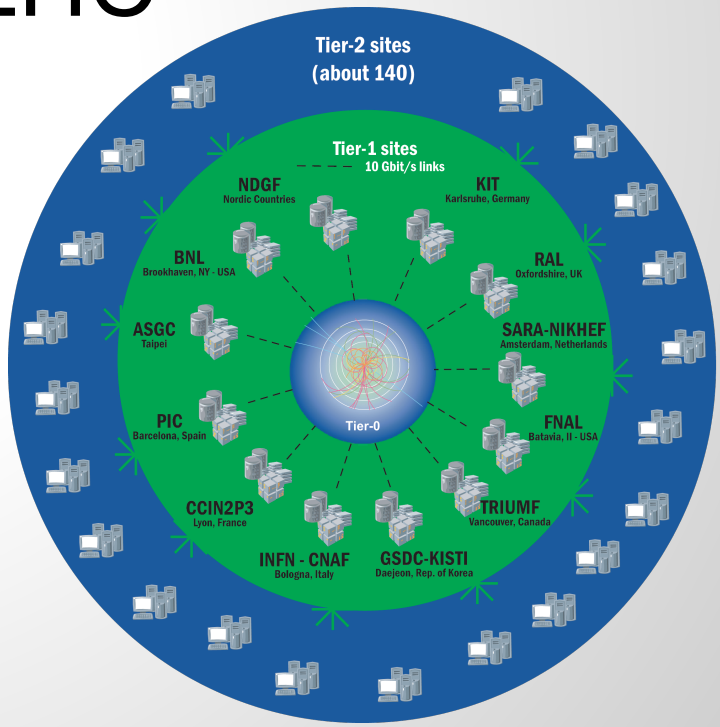


bit.ly/**WLCGsites**

http://gstat2.grid.sinica.edu.tw/gstat/gstat/geo/openlayers#/WLCG_TIER/ALL

A system of tiers

- Tier 0 - CERN data center
- Tier 1 - 11 computer centres large enough to store LHC data
- Tier 2 - around 140 universities and institutes
- Tier 3 - no formal engagements



http://home.web.cern.ch/about/computing/grid-system-tiers

http://wlcg.web.cern.ch/documents-reference

CERN Advanced Storage Manager

- priority is to capture, differing from immediate interactive responses (eg Facebook)
- Jobs are submitted to the manager, which then manipulates the data
- Storage is handled across both tape and disks

http://castor.web.cern.ch/
http://ercim-news.ercim.eu/en89/special/managing-large-data-volumes-from-scientific-facilities

Marketing

Big Data in Marketing

- Why big data matters to marketing:
  - Customer engagement.
  - Customer retention and loyalty.
  - Marketing optimization/performance.

Netflix

- Netflix use Big Data to make marketing decisions and make recommendations to customers.
- More than 30 million subscribers worldwide, and each on each visit, a customer will provide several data points.
- Playing, rating or searching for a video are events that can be captured and analysed.
- Netflix can also use the data to decide which series to fund next.

Amazon

- Amazon also uses Big Data in a similar way.
- Data gathered from customers on which products they purchase, as well as ratings and reviews can be used to make suggestions to customers based on which pages they visit.

Other uses of BIG data

Improving your business value

- Knowing yourself and your rivals
- Are you providing the most efficient interface for your users?
- Knowing the trends to adjust your services
- Connecting with seemingly unrelated partners

Dating?

- Big data is now used by dating websites to find you a better match, looks like size does matter after all!
- People lie on their profiles but you can access some data through social media for example
- Identifying you with people who have similar tastes : “other people who have liked Bob have also liked John”

And many more

- There’s data about everything
- Only limited by which data is relevant to you

Drawbacks

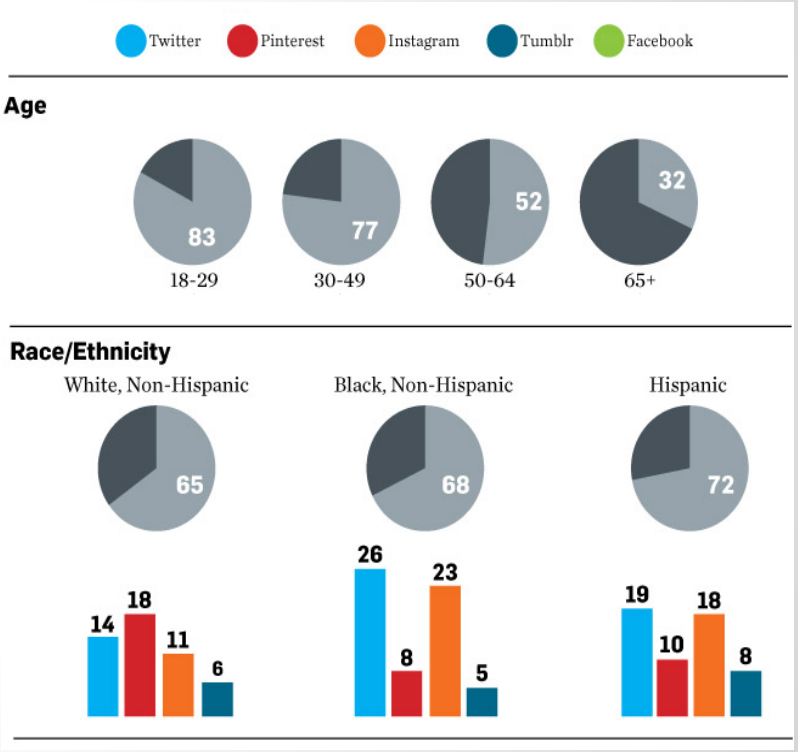
Can Big Data free us from theory?

http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory#

- In 2008 an article by Chris Anderson on Wired Magazine predicted "The end of theory" thanks to Big Data.
- Although this might seem tempting there have been lots of ups and downs in the ability of Big Data to build models.
- Google Flu Trends for example has successfully predicted flu epidemics for a couple of years by analyzing search trends.
- Recently though it has been shown that Google's results greatly overestimated the true impact of the disease.
- This has happened because there have been external events which have influenced the results.

Sampling bias and the 1936 U.S. elections

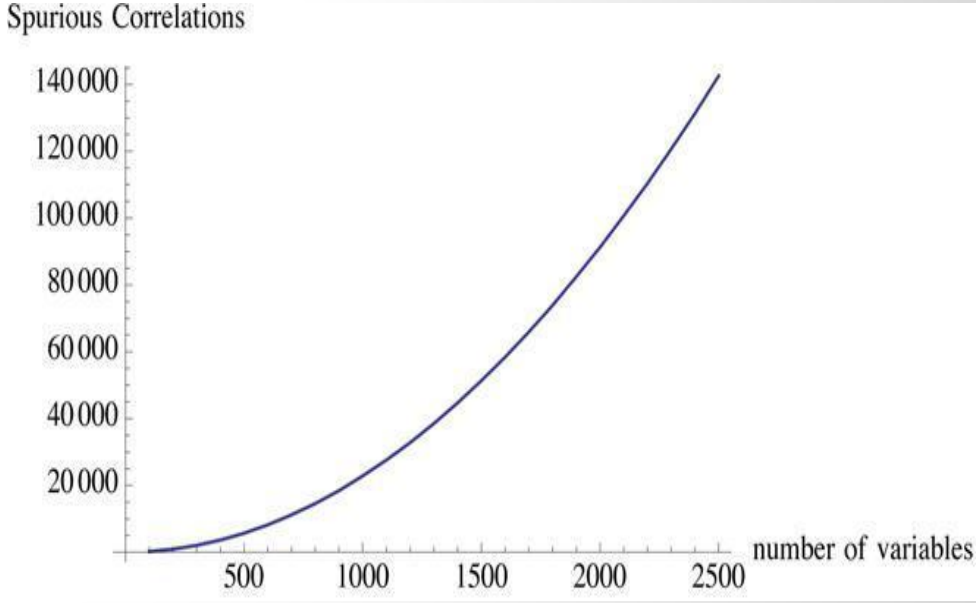
- Big Data leads researchers to think *'N=All'.*
- This is clearly an utopian goal.
- A very clear indication of how this may affect research is given by the 1936 U.S. elections (Landon vs .Roosevelt).
- Two voting polls were made, one had 2.4 million answers, another had 3000.
- Guess who got the closest to the election result?



http://www.pewinternet.org/2013/12/30/demographics-of-key-social-networking-platforms/

Beware of spurious correlations

- Spurious correlations are statistically sound correlations which appear by pure chance.
- They occur in any kind of statistical research, but are exacerbated when the amount of data increases.
- If researchers start taking them seriously we could be able to show virtually anything.
- For example a recently published paper showed that the amount of chocolate consumption is correlated with the number of Nobel Prize winners in a country.



http://www.wired.com/2013/02/big-data-means-big-errors-people/

Conclusion

- Big Data refers to the collection of large volumes of data
- It’s utilised in a large number of modern day areas, for Marketing, research and social networking
- The challenge is not so much gathering the data, but in effectively handling it
- Big Data is not perfect, it can lead to loose conclusions being made

References

- http://www.webopedia.com/TERM/B/big_data.html
- http://en.wikipedia.org/wiki/Big_data
- http://whatis.techtarget.com/definition/3Vs
- http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
- http://www.opallios.com/impact-of-big-data-in-social-media/
- http://www.exacttarget.com/blog/big-data-social-media/
- http://socialmediatoday.com/jayson-bowden/2066591/reasons-explore-big-data-social-media-analytics
- http://www.businessinsider.com/social-medias-big-data-future-2014-3
- http://www.ratioconsultants.nl/wp-content/uploads/2012/10/big-data-cloud1.png
- http://home.web.cern.ch/about/computing/worldwide-lhc-computing-grid
- http://home.web.cern.ch/about/computing/grid-system-tiers
- http://ercim-news.ercim.eu/en89/special/managing-large-data-volumes-from-scientific-facilities
- http://wlcg.web.cern.ch/documents-reference
- http://gstat2.grid.sinica.edu.tw/gstat/gstat/geo/openlayers#/WLCG_TIER/ALL
- http://castor.web.cern.ch/
- Chris Anderson, (2008), “The End of Theory: The Data Deluge Makes The Scientific Method Obsolete”, *Wired Magazine*, accessible at http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory, last accessed on 1/05/2014.
- Tim Harford ,(2014), “Big data: are we making a big mistake?”, *Financial Times*, accessible at http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz30DkRsjF4, last accessed on 1/05/2014.
- Maeve Duggan & Aaron Smith, (2013), “Demographics of key social networking platforms”, *PewResearch Internet Project*, accessible at http://www.pewinternet.org/2013/12/30/demographics-of-key-social-networking-platforms/, last accessed on 1/05/2014.
- Geoffrey Pullum, (2013), “Spurious Correlations Everywhere: the Tragedy of Big Data”, *The Chronicle of Higher Education*, accessible at http://chronicle.com/blogs/linguafranca/2013/03/04/spurious-correlations-everywhere/, last accessed on 1/05/2014.
- James R. Winters & Seán G. Roberts, (2012), “Chocolate Consumption, Traffic Accidents and Serial Killers”, accessible at http://replicatedtypo.com/wp-content/uploads/2012/11/ChocolateSerialKillers_WintersRoberts.pdf, last accessed on 1/05/2014.
- Messerli, F.H. (2012), “Chocolate consumption, cognitive function, and Nobel laureates”. *New England Journal of Medicine*.