Preserving Social Media

Big Data Network

Nathan Cunningham

Associate Director: Big Data
UK Data Archive, University of Essex

UK Data Service

DigitalPreservationCoalition

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

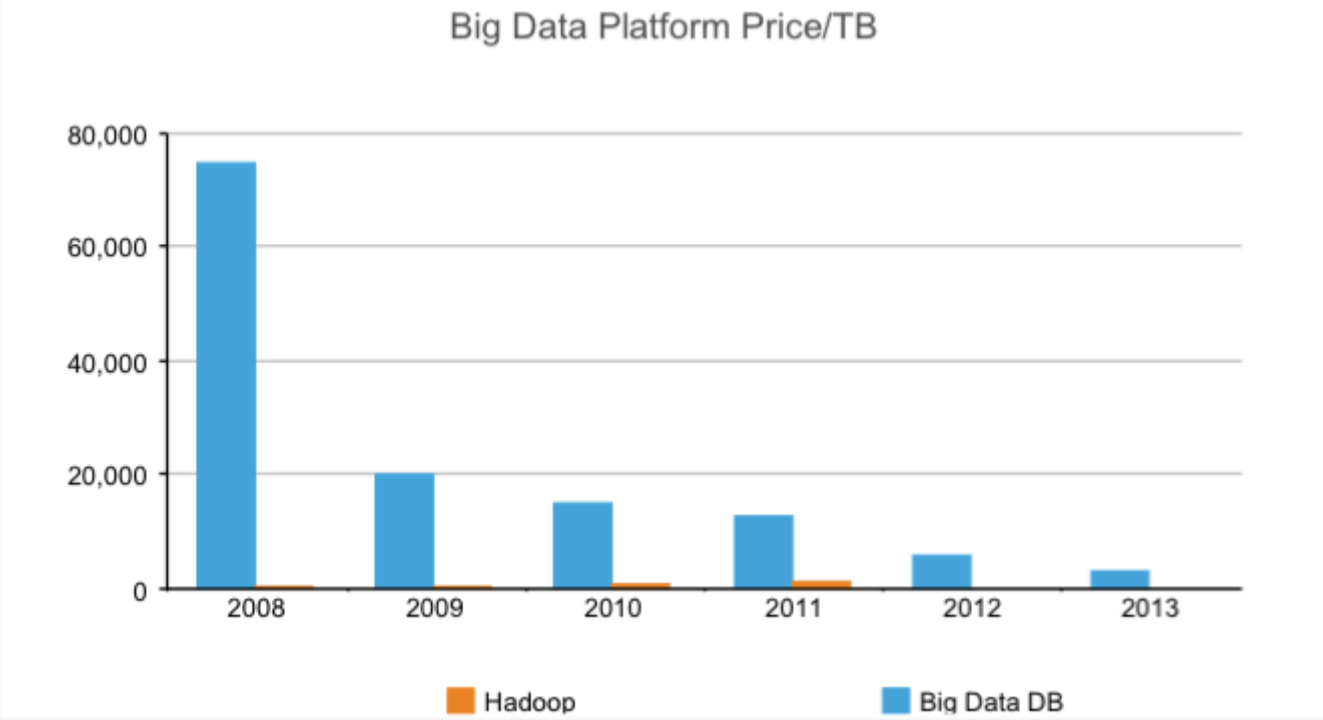
Outline

- What is Big Data
- Big Data Network
- Making Big Data useful for social science research
- UKDS Open Data Platform
- Unsafe data in a safe setting
- Challenges of Social Media

St Peters Square, Rome.

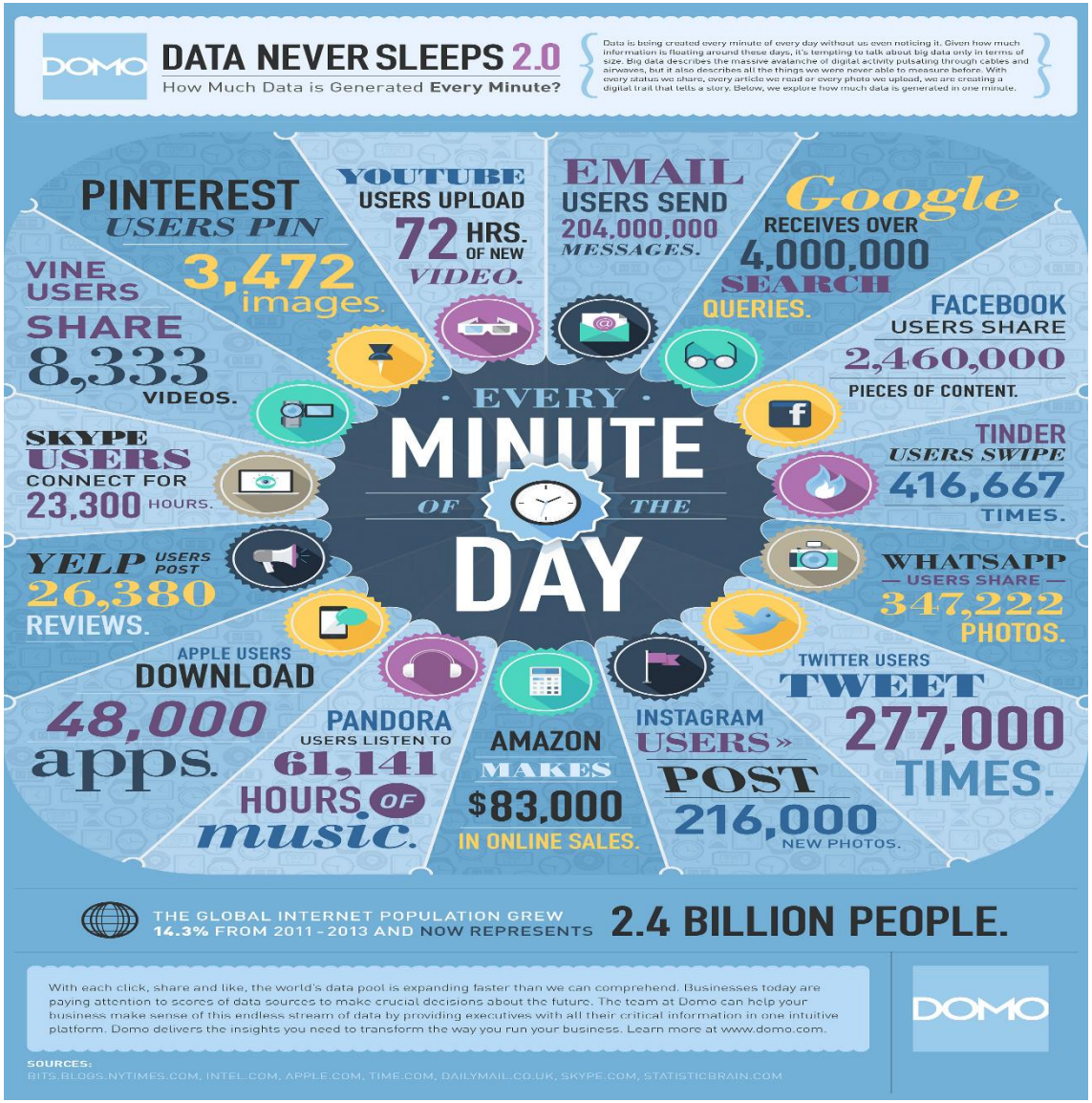
When data became a BIG deal

- Terrabytes of digital data are generated every second from computer systems around the world. They are now **accessible**.



(2014 Manoochehri – Data Just Right, Addison Wesley)

Data Never Sleeps (Infographic @ domo.com)



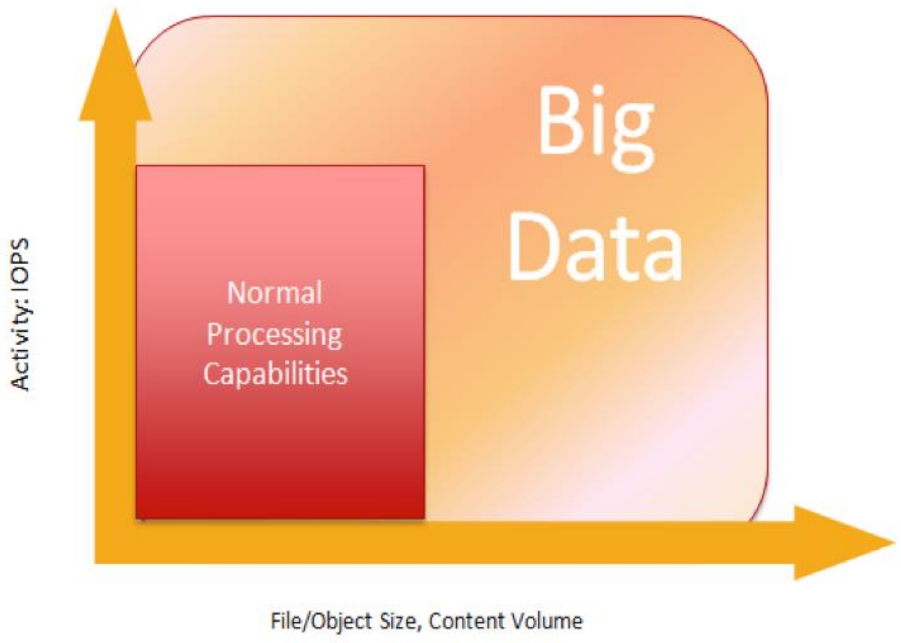
In 2012 the first infographic showed that Facebook users shared 684,478 pieces of content. Fast forward a couple of years to 2014 and that number has exploded to 2,460,000 pieces.

Big Data is about exploring new opportunities arising from our digital lives, technologies and services.

UK Data Service

A working definition of Big Data

Data sets that exceed the boundaries and sizes of normal processing capabilities, forcing you to take a non-traditional approach

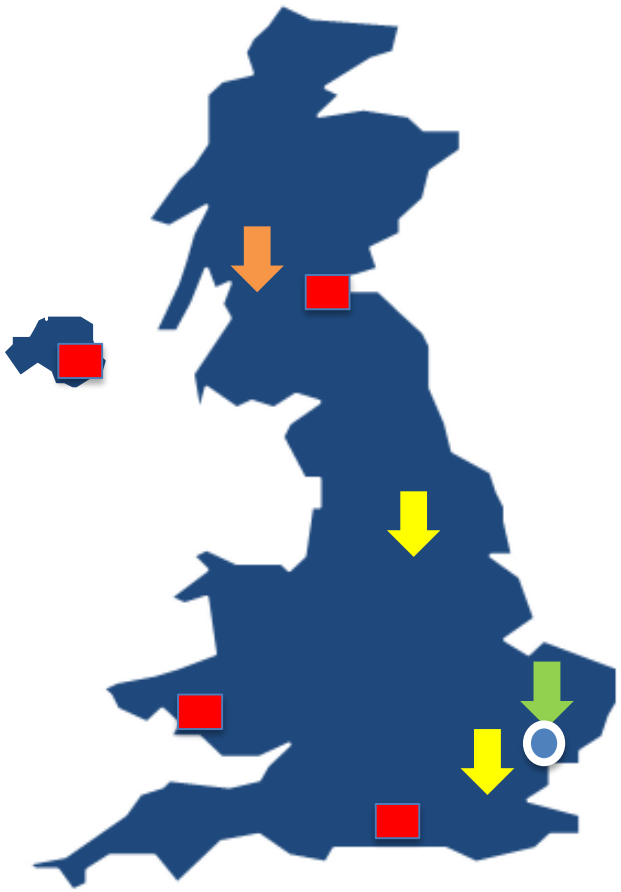


The most important V is **value**. My approach is to put the why before the how

Investment in Big Data

- Chancellor's Autumn Statement (2012) included £600m for science, research and innovation, £484m for RCUK
- Funds to support the development of innovative technologies across eight areas, including 'big data' - £189m for RCUK
- Draws from RCUK Strategic Framework for Capital Investment (published Nov 2012)
- April 2013 –ESRC earmarked £64m to support packages of activity within the 'big data' theme:
  - –Administrative Data Research Network
  - –Business and Local Government Data Research Centres
  - –Understanding Populations

ESRC £64 Million investment in Big Data



UKDS

ESRC Consumer Data Research Centre

ESRC Urban Big Data Research Centre

ESRC Data Research Centre for Business and Local Government

Administrative Data Research Network

UK Data Service

Big Data Network Support

- The UK Data Service is a comprehensive resource funded by the ESRC to support researchers, teachers and policymakers who depend on high-quality social and economic data.
- BDNS will support and coordinate activities between three dedicated Research Centres focusing on Business and Local Government Data.
- The Data Research Centres will make data, routinely collected by business and local government organisations, accessible for academics to undertake research that makes a difference: shaping public policies and making business, voluntary bodies and other organisations more effective as well as shaping wider society.
- Data will be made available by the owners in ways that prevent the identification of individuals.

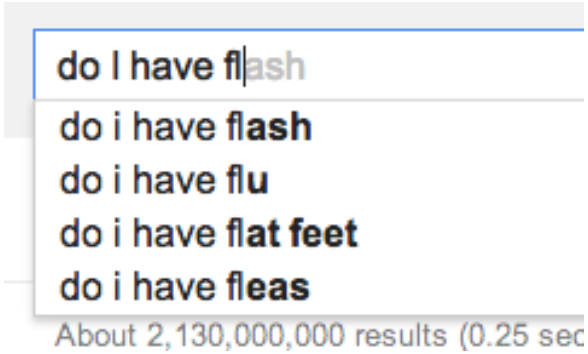
UK Data Service

The end of theory!

- Chris Anderson, editor-in-chief of Wired Magazine, wrote a provocative article entitled, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” (2008).
- He argued that hypothesis testing is no longer necessary with Google’s petabytes of data, which provides all of the answers to how society works. Correlation now “supercedes” causation.

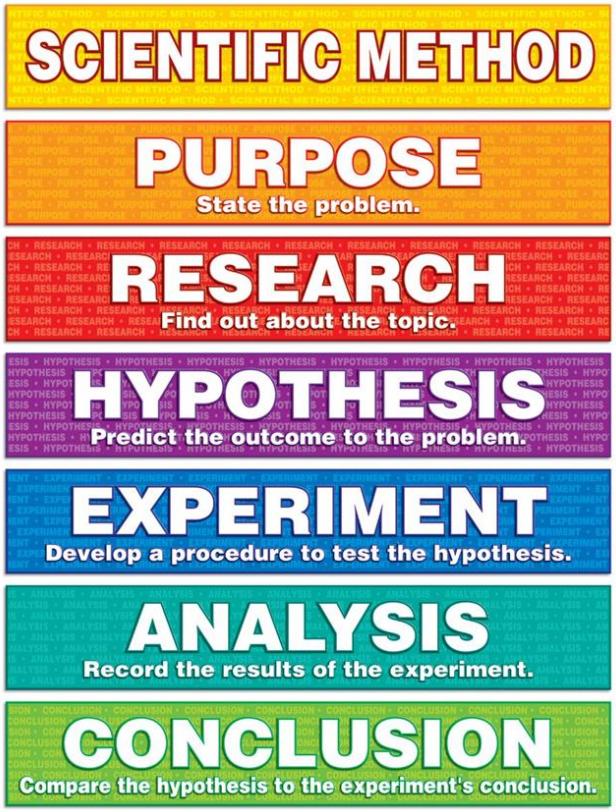
Google Flu

- Google Flu Trends is no longer good at predicting flu, scientists find
- Researchers warn of 'big data hubris' and the importance of updating analytical models, claiming Google has made inaccurate forecasts for 100 of 108 weeks.



Google's own autosuggest feature may have driven more people to make flu-related searches - and misled its Flu Trends forecasting system. Photograph: /Guardian

We are still doing science

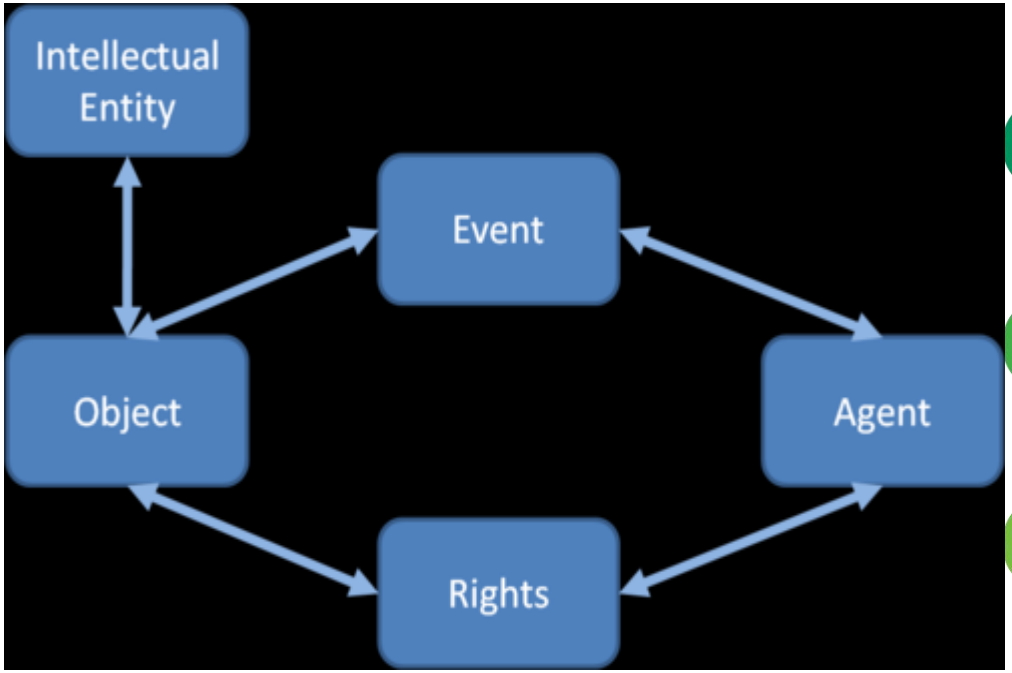


Pigliucci (2009:534) in response to Andersons Wired article:

“But, if we stop looking for models and hypotheses, are we still really doing science? Science, unlike advertising, is not about finding patterns—although that is certainly part of the process—it is about finding explanations for those patterns.”

Preserving Data

Intellectual Entities: a part of the content that can be considered as an intellectual unit for the management and the description of the content. This can be for example a book, a photo, or a database.



Object: a discrete unit of information in digital form, typically multimedia objects related to the intellectual entity.

Agent: a person, institution, or software application that is related to an event of an object or is associated to the rights of an object.

Event: An action that has an impact on an object or an agent.

Rights: description of one or more rights, permissions of an object or an agent

Making Big Data useful for social science research

The UKDS is implementing a Hadoop Data Lake to optimise a shared set of resources and implement a data security and governance.

Data Integrity	Reproducibility	Provenance	Quality
Curation	Preservation	Long term access and value.	Context
Ethics and legal frameworks	Publication and Citation	Licensing Conditions	

UK Data Service

Identifying Sources of Risk in Data

- The ‘five safes’ framework (Desai et al , 2014; see Camden, 2014, or Sullivan, 2011, for examples of use) is a way of identifying sources of risk in data access:
  1. Safe projects – whether the data use is lawful
  2. Safe people – whether the researchers can be trusted to hold and use the data appropriately
  3. Safe settings – whether the manner of accessing the data offers protection
  4. Safe data – whether there is any inherent protection in the data
  5. Safe outputs – whether the outputs from the research pose a disclosure risk

Ritchie, F. and Elliott, M. (2015) Principles- versus rules-based output statistical disclosure control in remote access environments. Working Paper. University of the West of England, Bristol. Available from: http://eprints.uwe.ac.uk/25376

UKDS-Open Data Platform

Modern platform standards are defined by open communities

Roadmap matches user requirements not vendor monetization requirements

For Hadoop, the ASF provides guidelines and a governance framework and the open community defines the standards for Hadoop.

Open Source Development Model accelerates innovation

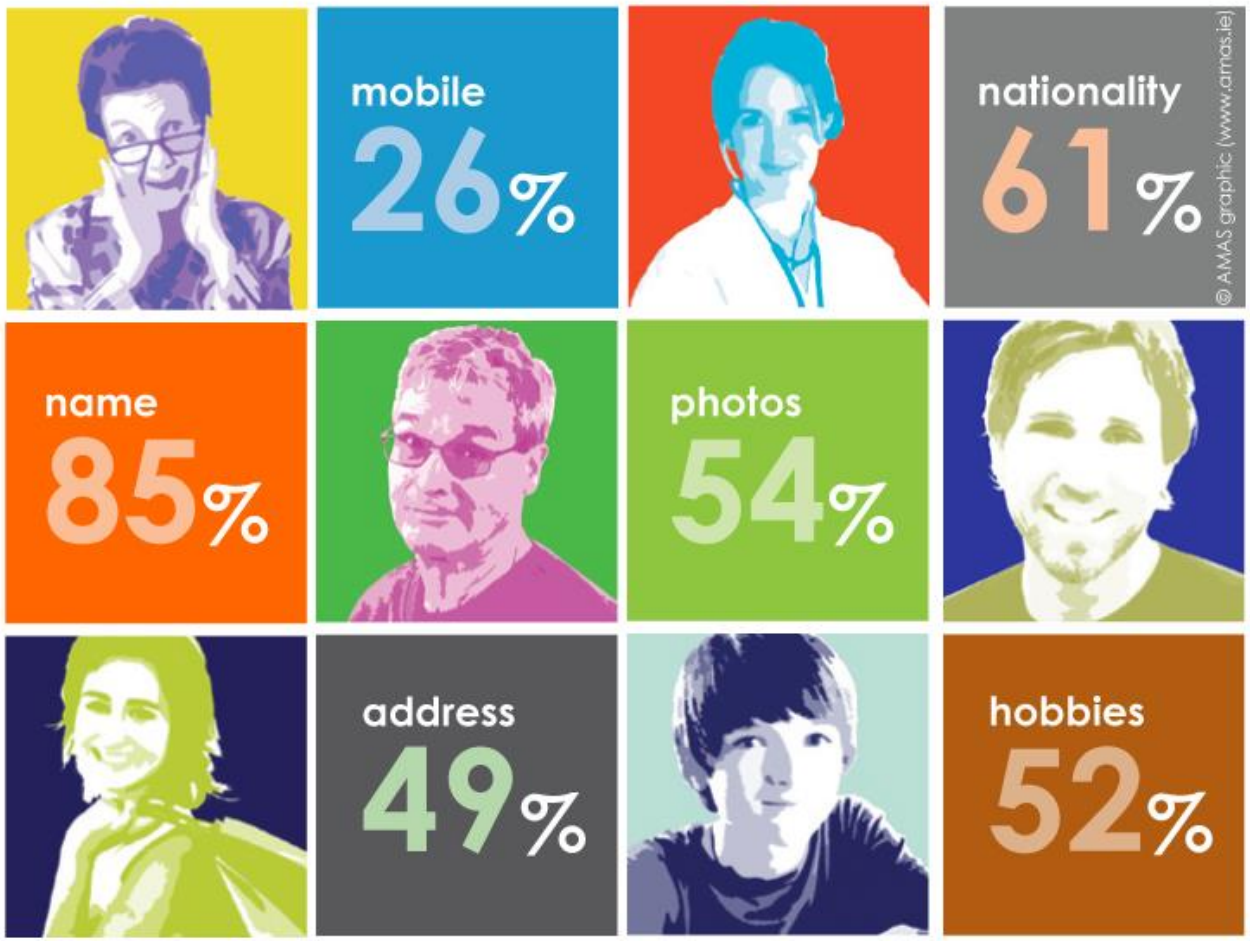
- Unconstrained number of developers under governance of ASF applied to problem
  - End users motivated to contribute to Apache Hadoop as they are consumers
  - Ecosystem motivated to align with Apache Hadoop to capture adjacent opportunities

Open Source Business Model lowers customer risk

- Business earned through ongoing value delivered, not one-time license sale
- Open platform eases integration with complementary systems
- Open community development reduces risk of vendor lock-in

UK Data Service

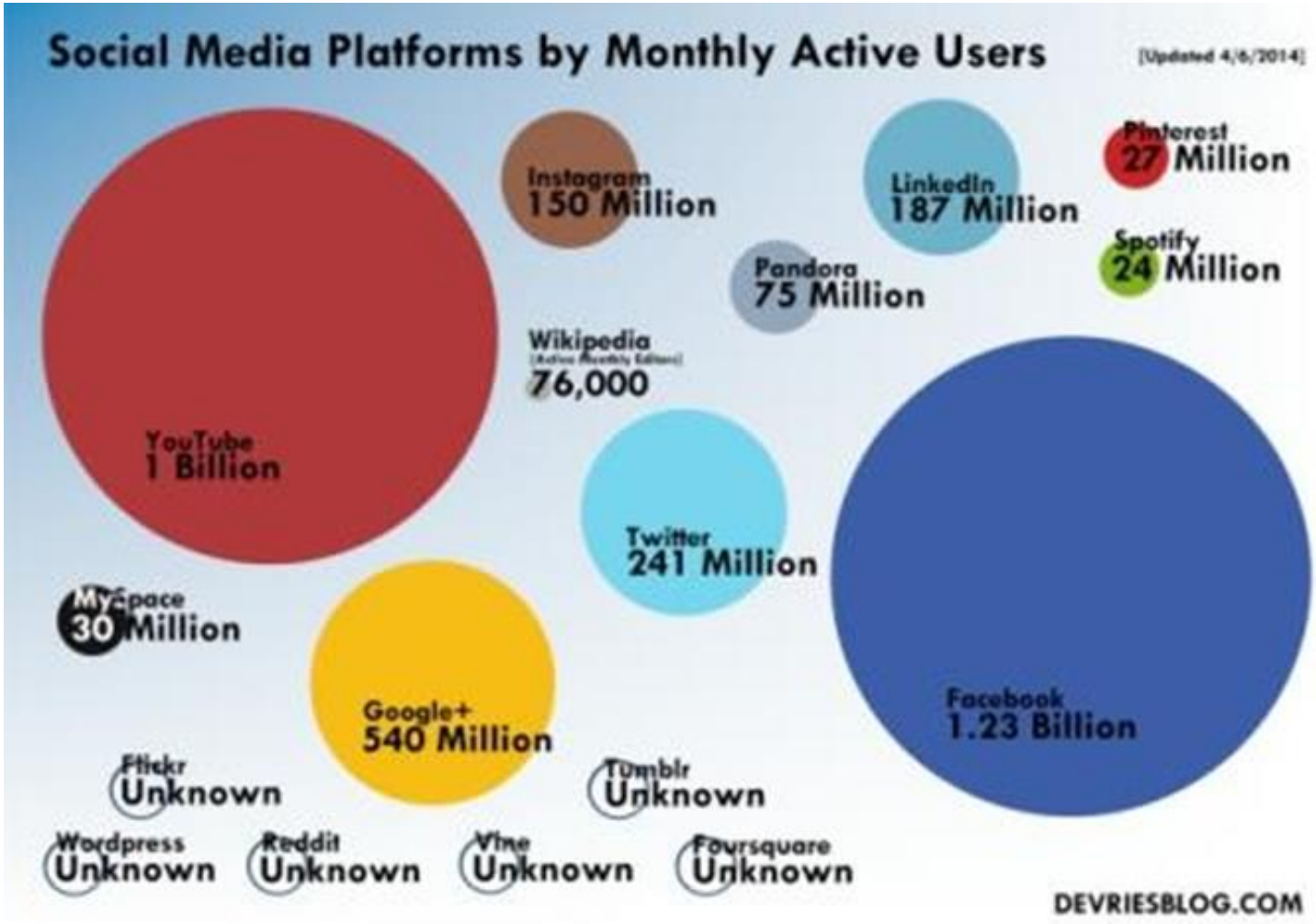
What we share



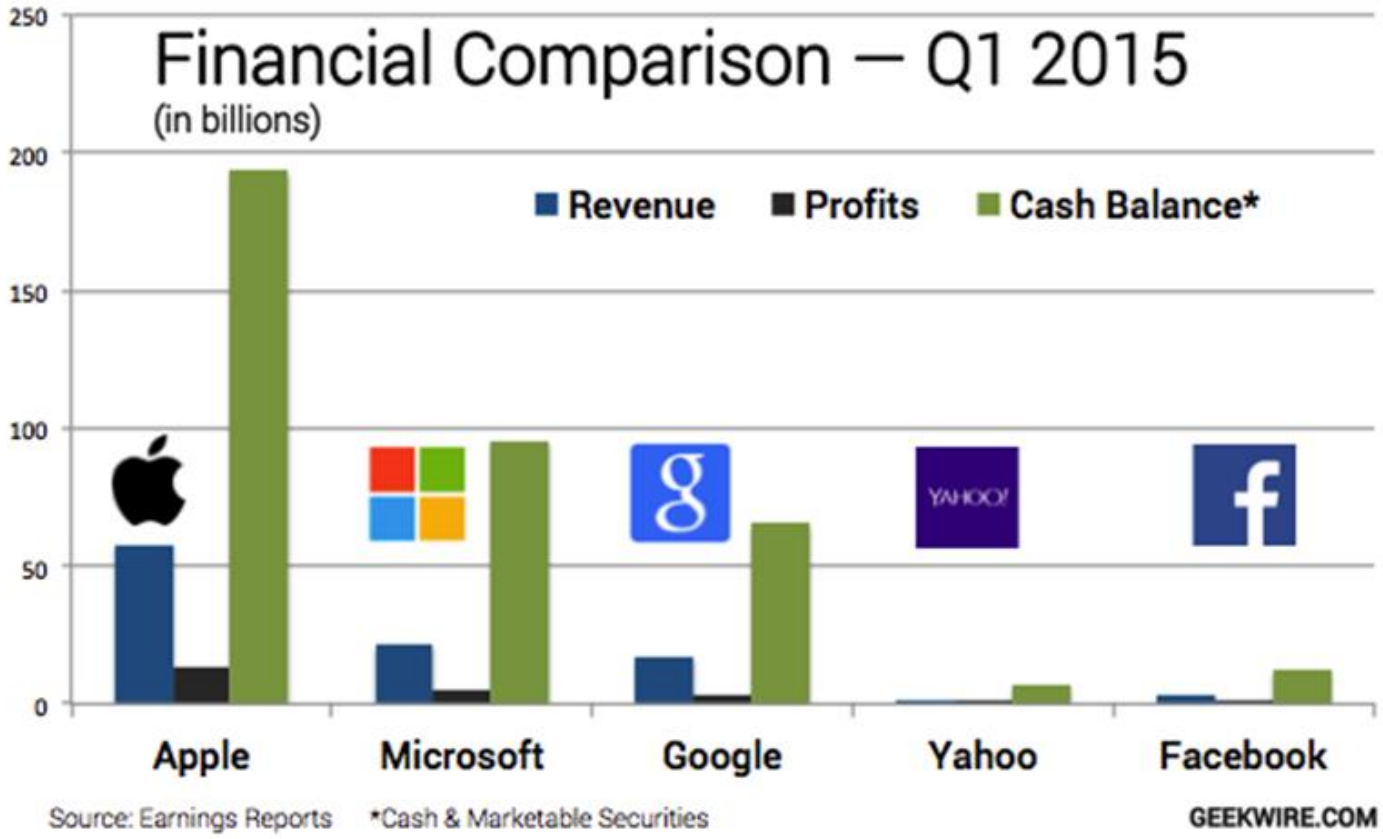
If the service is free, you are the product

http://amas.ie/online-research/state-of-the-net/state-of-the-net-issue-22-autumn-2011

Living the digital dream



Profits across many channels



The Explicit vs Implicit Digital You

	Twitter	LinkedIn	Amazon	Facebook	Apple	Google
Calendar	○	○	○	◑	●	●
Purchase history	○	○	●	○	●	●
Email	○	○	○	●	●	●
Viewing history	◑	◑	◑	●	●	●
Location	◑	◑	◑	●	●	●
Social graph	●	●	◑	●	◑	●
Address, phone number	◑	◑	●	●	●	●
Interests	◑	●	●	●	●	●

http://buytaert.net/winning-back-the-open-web

Questions

Nathan Cunningham

Big Data Network Support

njcunna@essex.ac.uk

http://ukdataservice.ac.uk/about-us/our-rd/big-data-network-support