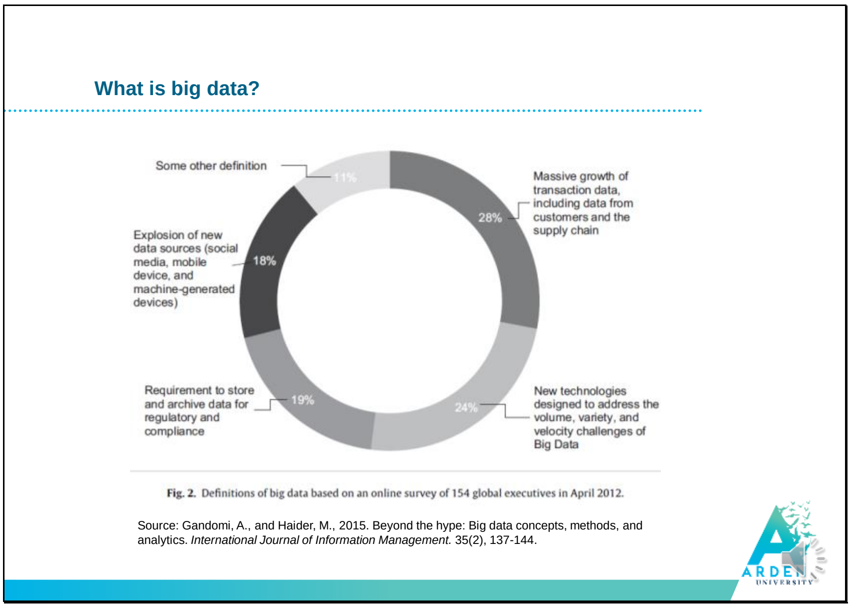
Slide 1

BIG DATA COLLECTION

What is “big”? And how does it apply to data collection?

Slide 2



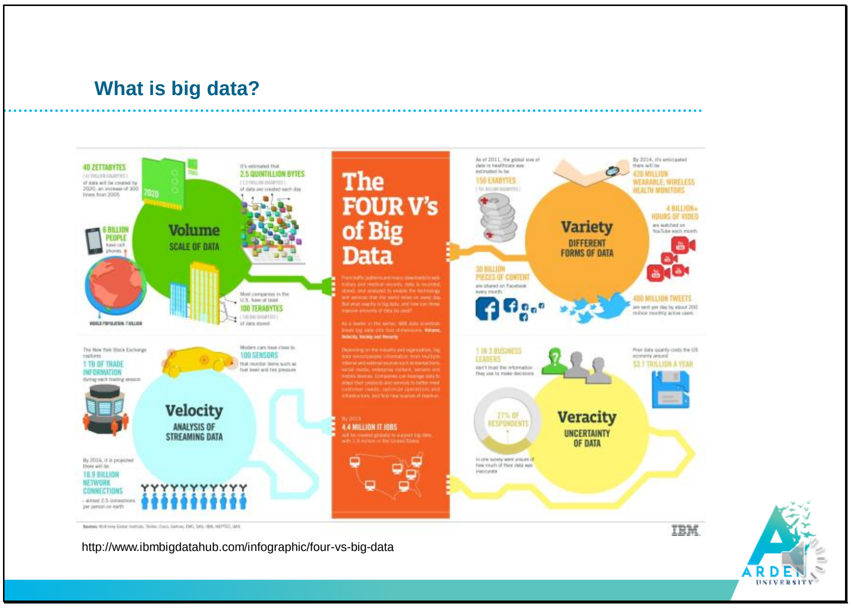
What is big data?

Fig. 2. Definitions of big data based on an online survey of 154 global executives in April 2012.

Source: Gandomi, A., and Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management.* 35(2), 137-144.

Big data obviously means different things to different people. But if you pay attention carefully to this diagram, which collected definitions from 154 executives, you will notice that they all address mostly data collection aspects, rather than, for example, data processing or modelling.

Slide 3



What is big data?

The FOUR V's of Big Data
Volume
SCALE OF DATA
40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005
6 BILLION PEOPLE have cell phones
WORLD POPULATION: 7 BILLION
It's estimated that 2.5 QUINTILLION BYTES of data are created each day
Most companies in the U.S. have at least 100 TERABYTES of data stored
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures 1 TB OF TRADE INFORMATION during each trading session
Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth
By 2015 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES
30 BILLION PIECES OF CONTENT are shared on Facebook every month
By 2014, it's anticipated there will be 420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO are watched on YouTube each month
400 MILLION TWEETS are sent per day by about 200 million monthly active users
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions
27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate
Poor data quality costs the US economy around $3.1 TRILLION A YEAR
IBM

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Most practitioners today use the definition of the 4 Vs of big data: Volume, Velocity, Variety and Veracity.

“Volume” represents the sheer size of the dataset. The size is considered due to both the large number of variables that is collected, as well as the size of the observations’ set for each variable.

“Velocity” reflects the speed at which these datasets are collected and analysed. It might be real-time collection from sensors, sales transactions or social media.

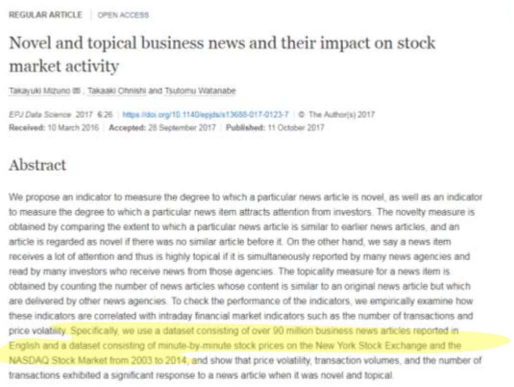
“Variety” represents the diverse structured and unstructured data sources and formats such as text, videos, networks, and images.

“Veracity” represents the inconsistency and unreliability of collected data. (George et al., 2016).

Simply put, big data is an amount and nature of data that we are uncomfortable (or less comfortable) to collect, visualize, store or analyse.

Slide 4

What is big data?



REGULAR ARTICLE | OPEN ACCESS

Novel and topical business news and their impact on stock market activity

Takayuki Mizuno, Takaaki Ohnishi and Tsutomu Watanabe

EPJ Data Science 2017 6:26 | https://doi.org/10.1140/epjds/s13688-017-0123-7 | © The Author(s) 2017
Received: 10 March 2016 | Accepted: 28 September 2017 | Published: 11 October 2017

Abstract

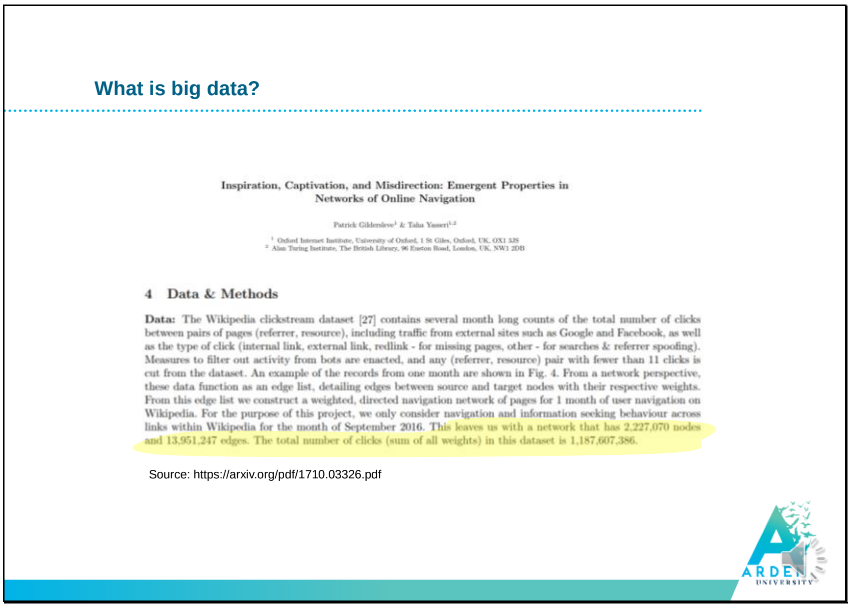
We propose an indicator to measure the degree to which a particular news article is novel, as well as an indicator to measure the degree to which a particular news item attracts attention from investors. The novelty measure is obtained by comparing the extent to which a particular news article is similar to earlier news articles, and an article is regarded as novel if there was no similar article before it. On the other hand, we say a news item receives a lot of attention and thus is highly topical if it is simultaneously reported by many news agencies and read by many investors who receive news from those agencies. The topicality measure for a news item is obtained by counting the number of news articles whose content is similar to an original news article but which are delivered by other news agencies. To check the performance of the indicators, we empirically examine how these indicators are correlated with intraday financial market indicators such as the number of transactions and price volatility. Specifically, we use a dataset consisting of over 90 million business news articles reported in English and a dataset consisting of minute-by-minute stock prices on the New York Stock Exchange and the NASDAQ Stock Market from 2003 to 2014, and show that price volatility, transaction volumes, and the number of transactions exhibited a significant response to a news article when it was novel and topical.

Source: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-017-0123-7

If only too few years ago, we were expecting our collected dataset to fit nicely into a spreadsheet, or in a typical relational database, the challenge of collecting data has completely changed, and changed with it our ability to make strong conclusions from it.

The study shown here, for example, is based on the collection of 90 million business articles, integrated with a dataset consisting of minute-by-minute stock prices from 2013 and 2014. This is far from just a sample.

Slide 5



What is big data?

Inspiration, Captivation, and Misdirection: Emergent Properties in Networks of Online Navigation

Patrick Gildersleve[1] & Taha Yasseri[1,2]

[1] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, UK, OX1 3JS
[2] Alan Turing Institute, The British Library, 96 Euston Road, London, UK, NW1 2DB

4 Data & Methods

Data: The Wikipedia clickstream dataset [27] contains several month long counts of the total number of clicks between pairs of pages (referrer, resource), including traffic from external sites such as Google and Facebook, as well as the type of click (internal link, external link, redlink - for missing pages, other - for searches & referrer spoofing). Measures to filter out activity from bots are enacted, and any (referrer, resource) pair with fewer than 11 clicks is cut from the dataset. An example of the records from one month are shown in Fig. 4. From a network perspective, these data function as an edge list, detailing edges between source and target nodes with their respective weights. From this edge list we construct a weighted, directed navigation network of pages for 1 month of user navigation on Wikipedia. For the purpose of this project, we only consider navigation and information seeking behaviour across links within Wikipedia for the month of September 2016. This leaves us with a network that has 2,227,070 nodes and 13,951,247 edges. The total number of clicks (sum of all weights) in this dataset is 1,187,607,386.

Source: https://arxiv.org/pdf/1710.03326.pdf

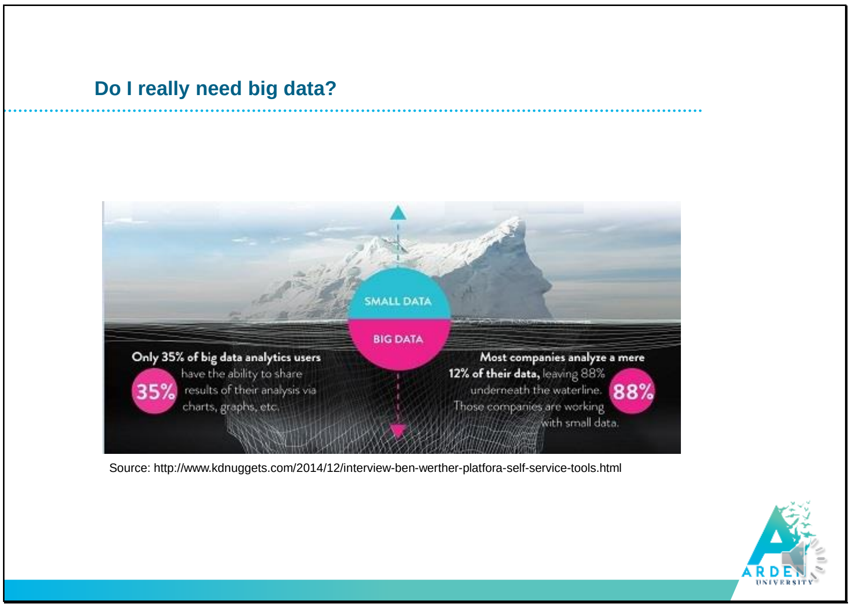
Big data is about collecting more data, that is being created more frequently, in a higher resolution (or granularity), and with much less ability to control its quality.

We were used to studies in which the unit of analysis is a human subject or a group of people, where data was collected using surveys and indirect observations.

In big data, however, the unit of analysis goes down to messages, clicks, or milliseconds of collected observations. This is yet another reason for which we must have automatic methods for data collection.

In this study, for example, the researchers collected more than one billion of clicks from Wikipedia during a single month.

Slide 6



Do I really need big data?

SMALL DATA

BIG DATA

Only 35% of big data analytics users have the ability to share results of their analysis via charts, graphs, etc.

35%

Most companies analyze a mere 12% of their data, leaving 88% underneath the waterline. Those companies are working with small data.

88%

Source: http://www.kdnuggets.com/2014/12/interview-ben-werther-platfora-self-service-tools.html

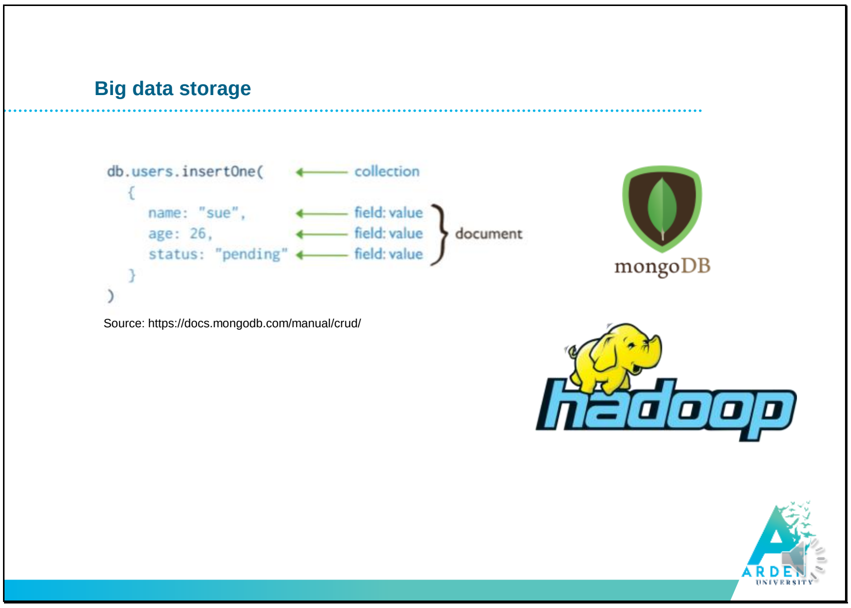
In reality, still, most of the time, we do not have the big data to collect about our business questions.

And when we do, we do not necessarily need it.

And when we do, sometimes its lack of structure, or level of quality, does not allow us to get to the bottom of things.

In Lesson 7 of this module, we will look into analysis methods, which are most often related to “small data”; that is, data which is sampled.

Slide 7



Big data storage

```
db.users.insertOne(
   {
      name: "sue",
      age: 26,
      status: "pending"
   }
)
```

collection
field: value
field: value
field: value
document
mongoDB
hadoop

Source: https://docs.mongodb.com/manual/crud/

Big data typically requires storage capacity which exceeds the capacity found in regular desktop computers or laptops.

In addition, the variety and veracity of big data is very large, which makes it hard to fit into a certain set of well-specified fields in a table.

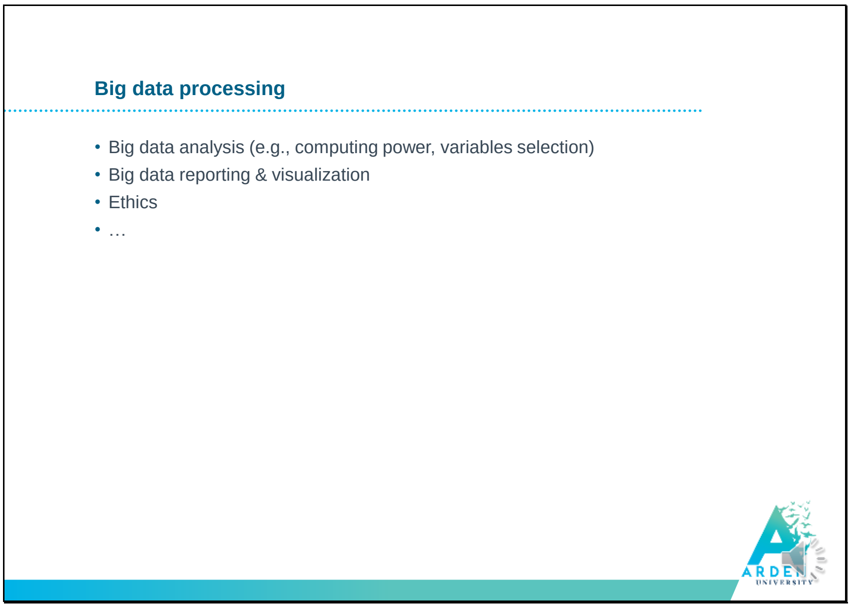
This lack of common structure of a large integrated dataset can be answered by NoSQL storage platforms (such as Hadoop's or MongoDB).

NoSQL databases can accommodate data that does not have a pre-defined structure.

Instead of a tabular (or relational) format, NoSQL formats are pairs of key-value (see in the image), graph-based, or document-based, all of which enable a dynamic structure of data.

This approach makes it more flexible in terms of different pre-defined schema, or format, of the collected data.

Slide 8



Big data processing

- Big data analysis (e.g., computing power, variables selection)
- Big data reporting & visualization
- Ethics
- …

Data collection is of course not the only aspect that needs consideration when it to comes to big data.

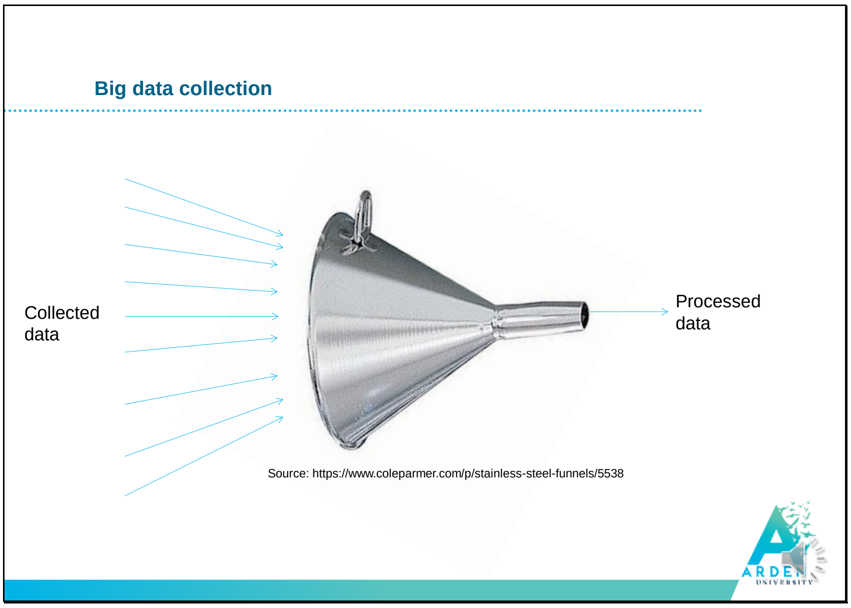
Big data analysis, for example, requires platforms that can handle large amounts of data. For example, social media analytics, video analytics and audio analytics.

Variable selection is another very typical issue. Big data may contain a large number of variables. If we are not sure in advance which exact variables we are going to need, we are going to need to use statistical methods for variable selection.

Other aspects exist, such as reporting, visualization and ethics.

However, the reality today is that much more emphasis is put on big data collection than on other aspects.

Slide 9



Big data collection

Collected data

Processed data

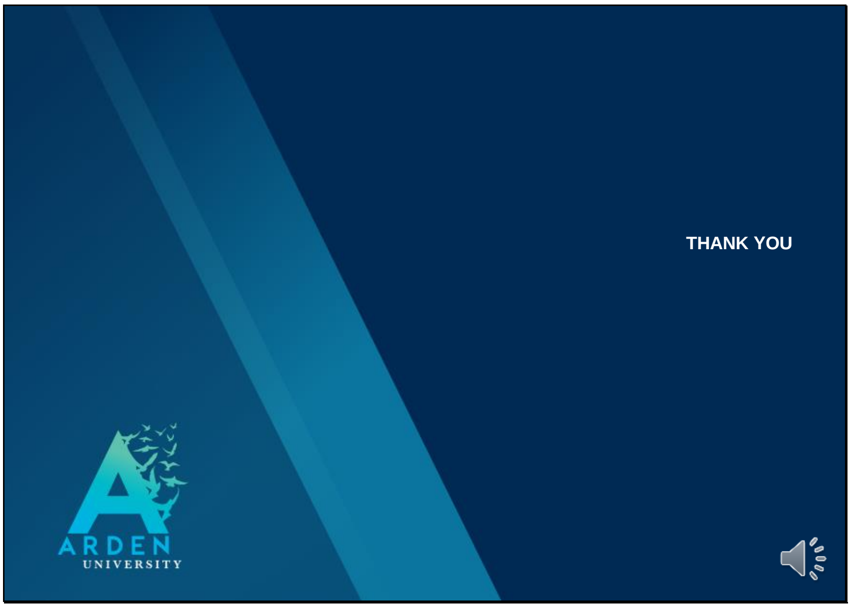
Source: https://www.coleparmer.com/p/stainless-steel-funnels/5538

The amount and variety of data that is collected significantly exceeds the amount of data that is asked about and eventually processed.

Unstructured data, for example (such as video or image data), is still very challenging to analyze, and requires large amounts of storage.

Thus, automatic data collection methods are of great significance, specifically when it comes to big data.

Slide 10



THANK YOU

Thank you.