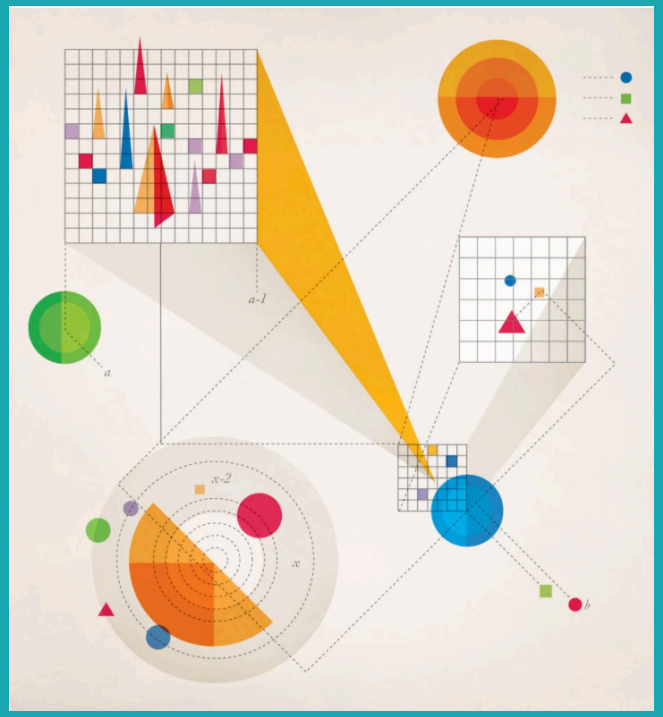
Big Data in R

ZOË KITCHEL

SPRING 2020

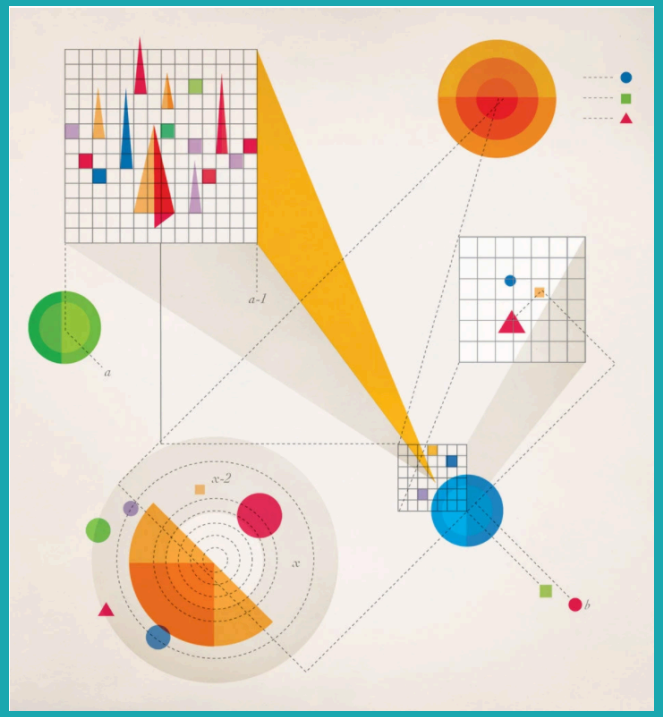
DATA CLUB

What is ‘**big** data?’



Hagen, NYTimes, 2012

What is ‘big data?’



Hagen, NYTimes, 2012

an amount of data that exceeds a petabyte—one million gigabytes

the exponential increase and availability of data in our world

there are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating (Forbes 2018)

over the past 2 years, 90% of the data in the world was created

can be transformed into:

- actionable insight
- improved decision making
- competitive advantage

“The culture has changed. There is this idea that numbers and statistics are interesting and fun. It’s cool now.”
-Andrew Gelman

applications of ‘big data’

applications of

‘big data’

“The culture has changed. There is this idea that numbers and statistics are interesting and fun. It’s cool now.”
-Andrew Gelman

Research | Open Access | Published: 01 December 2015

Applications of big data to smart cities

Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed & Jameela Al-Jaroodi

Journal of Internet Services and Applications **6**, Article number: 25 (2015) | Cite this article

54k Accesses | **181** Citations | **21** Altmetric | Metrics

applications of ‘big data’

“The culture has changed. There is this idea that numbers and statistics are interesting and fun. It’s cool now.”
-Andrew Gelman

Research | Open Access | Published: 01 December 2015

Applications of big data to smart cities

Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed & Jameela Al-Jaroodi

Journal of Internet Services and Applications **6**, Article number: 25 (2015) | Cite this article

54k Accesses | **181** Citations | **21** Altmetric | Metrics

ELSEVIER

Journal of Business Research
Volume 69, Issue 2, February 2016, Pages 897-904

Big Data consumer analytics and the transformation of marketing

Sunil Erevelles [a, 1], Nobuyuki Fukawa [b], Linda Swayne [a, 2]

Show more

https://doi.org/10.1016/j.jbusres.2015.07.001 Get rights and content

"The culture has changed. There is this idea that numbers and statistics are interesting and fun. It's cool now."
-Andrew Gelman

applications of '**big** data'

Research | Open Access | Published: 01 December 2015

Applications of big data to smart cities

Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed & Jameela Al-Jaroodi

Journal of Internet Services and Applications **6**, Article number: 25 (2015) | Cite this article

54k Accesses | **181** Citations | **21** Altmetric | Metrics

ELSEVIER

Journal of Business Research
Volume 69, Issue 2, February 2016, Pages 897-904

Big Data consumer analytics and the transformation of marketing

Sunil Erevelles [a, 1], Nobuyuki Fukawa [b], Linda Swayne [a, 2]

Show more

https://doi.org/10.1016/j.jbusres.2015.07.001 Get rights and content

FiveThirtyEight

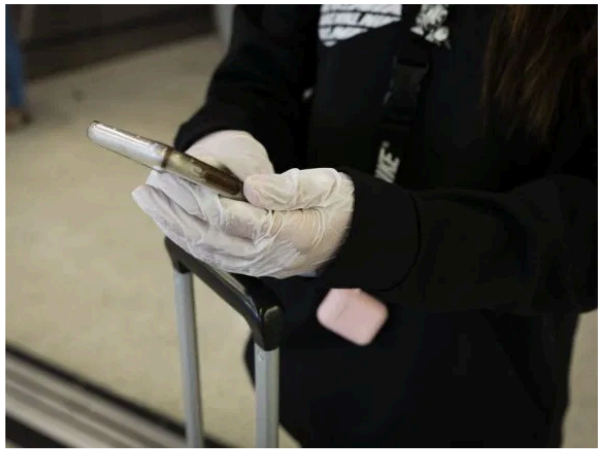
Politics Sports **Science & Health** Economics Culture

APR. 9, 2020, AT 7:00 AM

Big Data Is Helping Us Fight The Coronavirus — But At What Cost To Our Privacy?

By Neil Paine

Filed under Technology



Cellphone location data has been used to track whether people are staying home during the pandemic. But how much surveillance should we accept? EVA MARIE UZCATEGUI TRINKL / ANADOLU AGENCY / GETTY IMAGES

A.I. Is Helping Scientists Understand an Ocean's Worth of Data

Machine-learning applications are proving to be especially useful to the scientific community studying the planet's largest bodies of water.

NY Times 2020

"…in the ocean…there is both **so much data** — big surfaces, deep depths — and not enough data — it is too expensive and not necessarily useful to collect samples of any kind from all over"

> "**Climate change** makes machine learning that much more valuable, too: So much of the data available to scientists is not necessarily accurate anymore, as animals move their habitats, temperatures rise and currents shift. As species move, managing populations becomes even more critical.

NY Times 2020

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

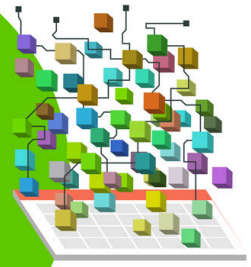
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

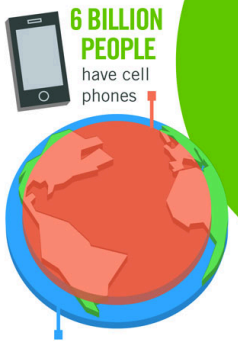
Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

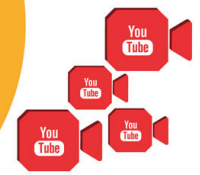
Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

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

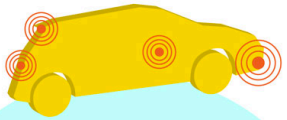
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users

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

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

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

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Doug Laney 2001 &

IBM

How do I know if my data are 'big?'

- 1 M records = good to go, 1 M – 1 B = can work in R with some extra help, > 1 B = need to be analyzed with map reduce algorithms with help from Hadoop etc.

More practically…

- If R doesn't work for you because you have too much data
- What can get more difficult when data is big?
  - The data may not load into memory
  - Analyzing data may take a long time
  - Visualizations get messy

Okay, my data are too ‘big,’ what now?

- Check if you’re using 64-bit version of R
- Allocate more memory (if you have it) to R
- Reduce # of objects stored in memory

Still too big?

- Make data smaller
- Get a bigger computer
- Access data differently
- Split up the dataset for analysis

Adapted from Big Data 2015 Columbia

Make data smaller

- Run your analyses on a smaller chunk of your overall dataset to make sure it is indeed a memory or data size issue

Adapted from Big Data 2015 Columbia

Get a bigger computer

- Convince your supervisor to buy you a new computer

When that fails

- Rutgers High-Performance Clusters
  - Amarel (DEENR Node)
  - Annotate (SEBS)
  - March 2nd tutorial with Ashley Trudeau

Adapted from Big Data 2015 Columbia

Access data differently

- Use `**data.table**` package
  - Good for very large data files
  - Behaves like a data frame
  - Offers fast subset, grouping, update, and joins
  - Makes it easy to turn an existing data frame into a data table

Adapted from Big Data 2015 Columbia

Split up analyses

- Do analyses on *x* MB at a time
- Combine results
- Can use computing clusters to parallelize analysis:
  - Farming out subtasks to independent processors
  - MapReduce algorithms

Resources

Wisconsin Data Science

Large Datasets and You

FasteR! HigheR! StrongeR! - A Guide to Speeding Up R Code for Busy People

Taking R to the Limit, Part II: Working with Large Datasets

R Bloggers: Big Data in R

CRAN-R: Intro to Data Table

CRAN-R: Keys in Data Table

CRAN-R: Assignment by Reference