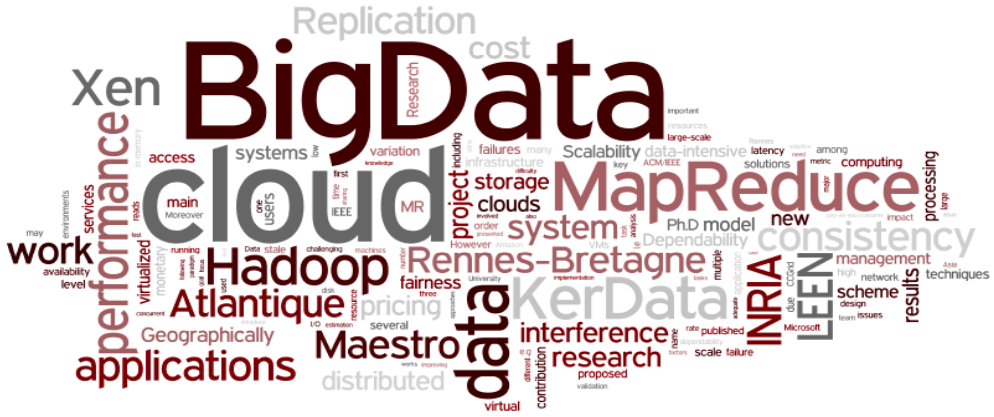
INVENTORS FOR THE DIGITAL WORLD

Cloud & Big Data



Shadi Ibrahim

Inria, Rennes - Bretagne Atlantique Research Center

Era of Big Data

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

f Big Data

Big Data

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

cnrs
n° 28
Quarterly
January 2013
international magazine
The
BIG DATA
REVOLUTION
In the News
Serge Haroche
2012 Nobel Laureate
Spotlight
Elysée Treaty
50th Anniversary

Big Data

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Meschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

01 Business
01business.com
LE MAGAZINE QUI MET L'ENTREPRISE À L'HEURE DU NUMÉRIQUE
LES PROMESSES DU BIG DATA
• À quoi ça sert
• Qui l'utilise
• Comment en tirer parti
P. 32
ACTUALITÉ
Le boom des Fab Labs donne un coup de fouet à notre R&D P. 20
ENQUÊTE
Dans quels cas se faire payer en bitcoins P. 44
CAS D'ENTREPRISES
Coca-Cola, Rue du Commerce, Accenture... P. 48

cnrs
international magazine
n° 28
Quarterly
January 2013
The BIG DATA REVOLUTION
In the News
Serge Haroche
2012 Nobel Laureate
Spotlight
Elysée Treaty
50th Anniversary

Big Data

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND Lawrence Meschler Questions the Cloud
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

01 Business
N°2175 17 OCTOBRE 2013
6,90€
01business.com
LE MAGAZINE QUI MET L'ENTREPRISE À L'HEURE DU NUMÉRIQUE
LES PROMESSES DU BIG DATA
ACTUALITÉ
Le boom des Fab Labs donne un coup de fouet à notre R&D P.20

cnrs
n° 28
Quarterly
January 2013
international magazine
The
BIG DATA
REVOLUTION
In the News
Serge Haroche
2012 Nobel Laureate
Spotlight
Elysée Treaty
50th Anniversary

4 September 2008 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

Big Data

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND Lawrence Weschler Questions the Cloud
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

01 Business
01business.com
LE MAGAZINE QUI MET L'ENTREPRISE À L'HEURE DU NUMÉRIQUE
LES PROMESSES DU BIG DATA
ACTUALITÉ
Le boom des Fab Labs donne un coup de fouet à notre R&D P.20

8-PAGE SPECIAL POSTGRADUATE SURVIVAL GUIDE
8TH BIRTHDAY ISSUE!
COSMOS
THE SCIENCE OF EVERYTHING
THE END OF VIOLENCE
Steven Pinker on the new peace p36
DEFEATING POLIO
Will politics jeapordise a cure? p68
FRAUDS AND FAKES
Science's biggest scams p24
GENIUS OF DOGS
Inside the canine brain p41
IS data THE NEW GOD?
How tracking your digital trail could predetermine your future – and why you'll benefit from today's data deluge. p46
GALAXIES AND NEBULAE ■ CANCER VACCINES ■ WHALES ■ FICTION ■ REVIEWS

cnrs
international magazine
n° 28
Quarterly
January 2013
The BIG DATA REVOLUTION
In the News
Serge Haroche
2012 Nobel Laureate
Spotlight
Elysée Treaty
50th Anniversary

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE BITER BI
Viral infections for viruses
TROPICAL CYCLONE
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

Big Data

POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

01 Business
01business.com
LE MAGAZINE QUI MET L'ENTREPRISE À L'HEURE DU NUMÉRIQUE
LES PROMESSES DU BIG DATA
ACTUALITÉ
Le boom des Fab Labs donne un coup de fouet à notre R&D P.20

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE BITER BI
Viral infections for viruses
TROPICAL CYCLONE
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

cnrs
international magazine
n° 28
Quarterly
January 2013
The BIG DATA REVOLUTION
In the News
Serge Haroche
2012 Nobel Laureate
Spotlight
Elysée Treaty
50th Anniversary

8-PAGE SPECIAL POSTGRADUATE SURVIVAL GUIDE
8TH BIRTHDAY ISSUE!
COSMOS
THE SCIENCE OF EVERYTHING
THE END OF VIOLENCE
Steven Pinker on the new peace
DEFEATING POLIO
Will politics jeopardise a cure?
FRAUDS AND FAKES
Science's biggest scams
GENIUS OF DOGS
Inside the canine brain
IS data THE NEW GOD?
How tracking your digital trail could predetermine your future – and why you'll benefit from today's data deluge.
GALAXIES AND NEBULAE

Pipeline | Volume 11 | Issue 3
Pipeline
The Unstructured Data Structure
GET SMARTER!
NEXT-GEN ANALYSIS
BIG DATA TO THE RESCUE
THE BIG DATA HUBRIS
BIG DATA & ANALYTICS
EXPERT TIPS for Call Center CEM
Box Fresh: Real-time Contextual Awareness
Big Data Takes the C-suite
Big Data Infrastructure Demands
Machine Learning and Big Data

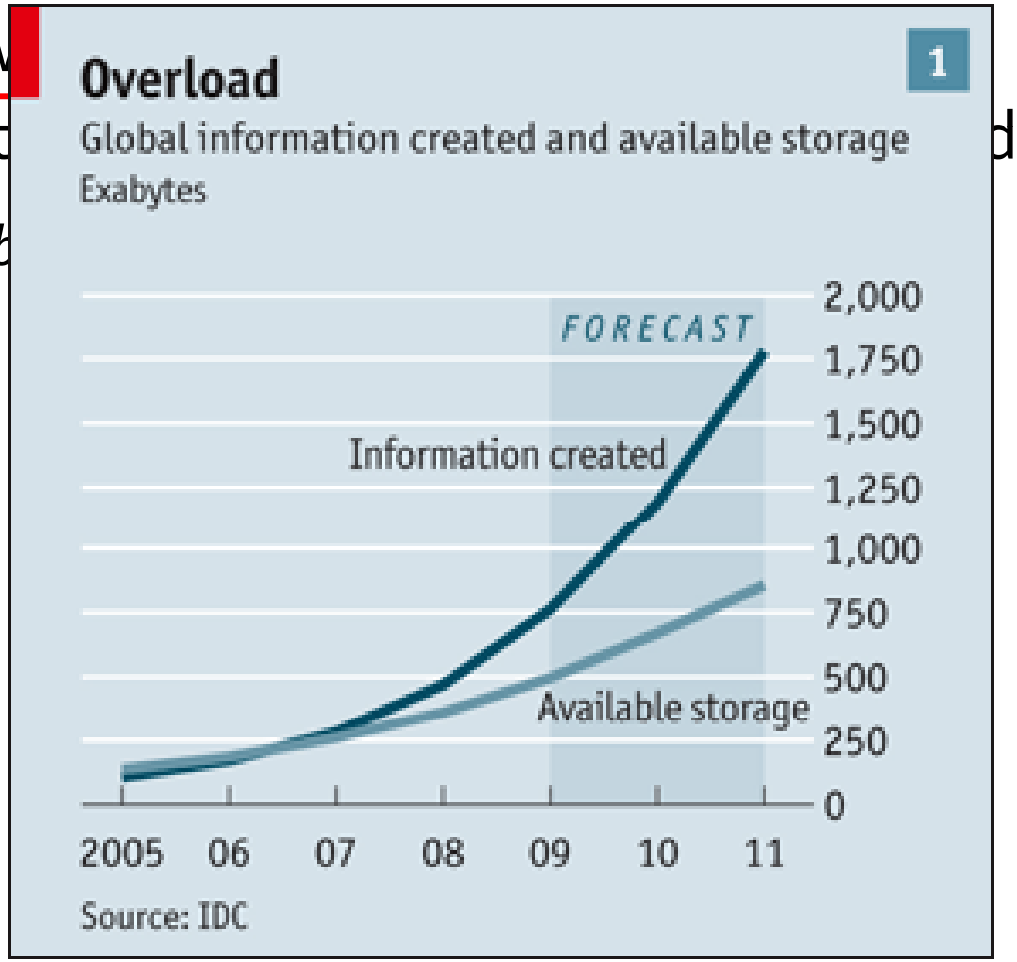
What is Big Data?

What is Big Data?

"Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." - *The McKinsey Global Institute, 2011*

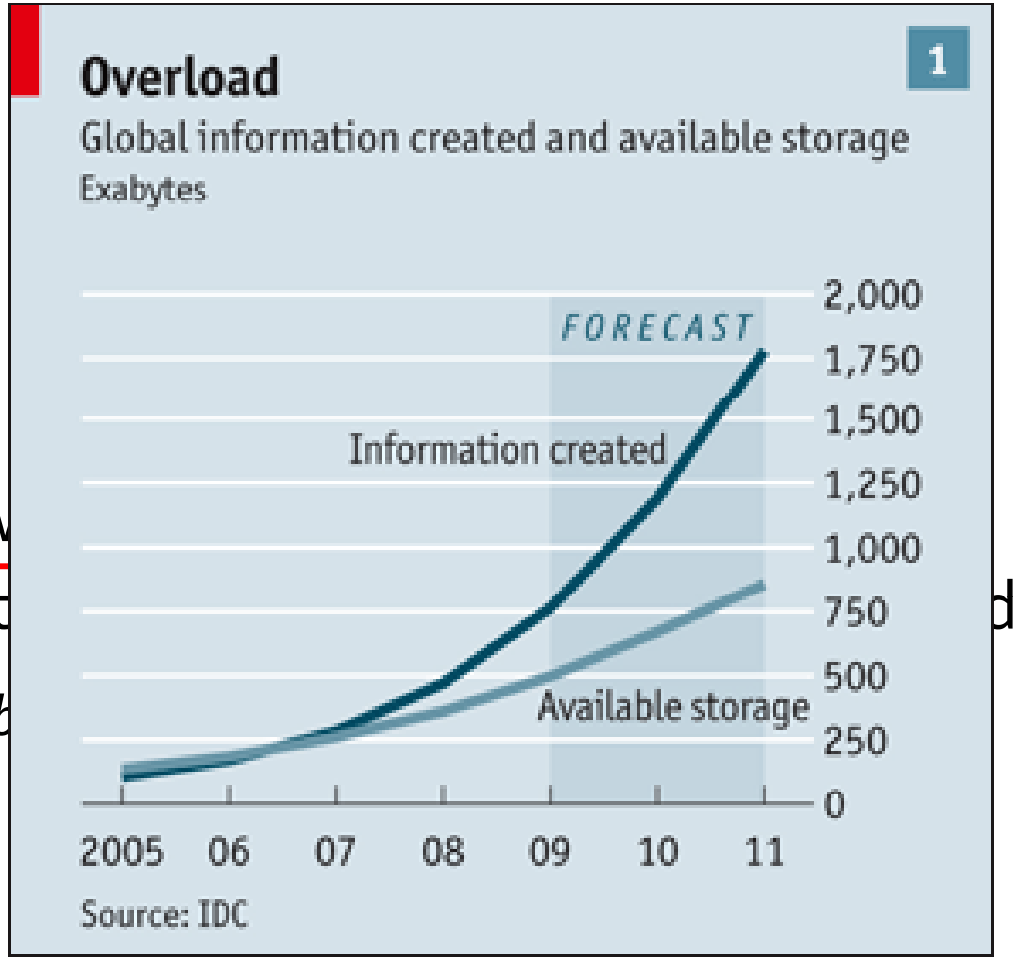
What is Big Data?

"Big data refers to data sets w
typical database software t
analyze." - *The McKinsey Glob*



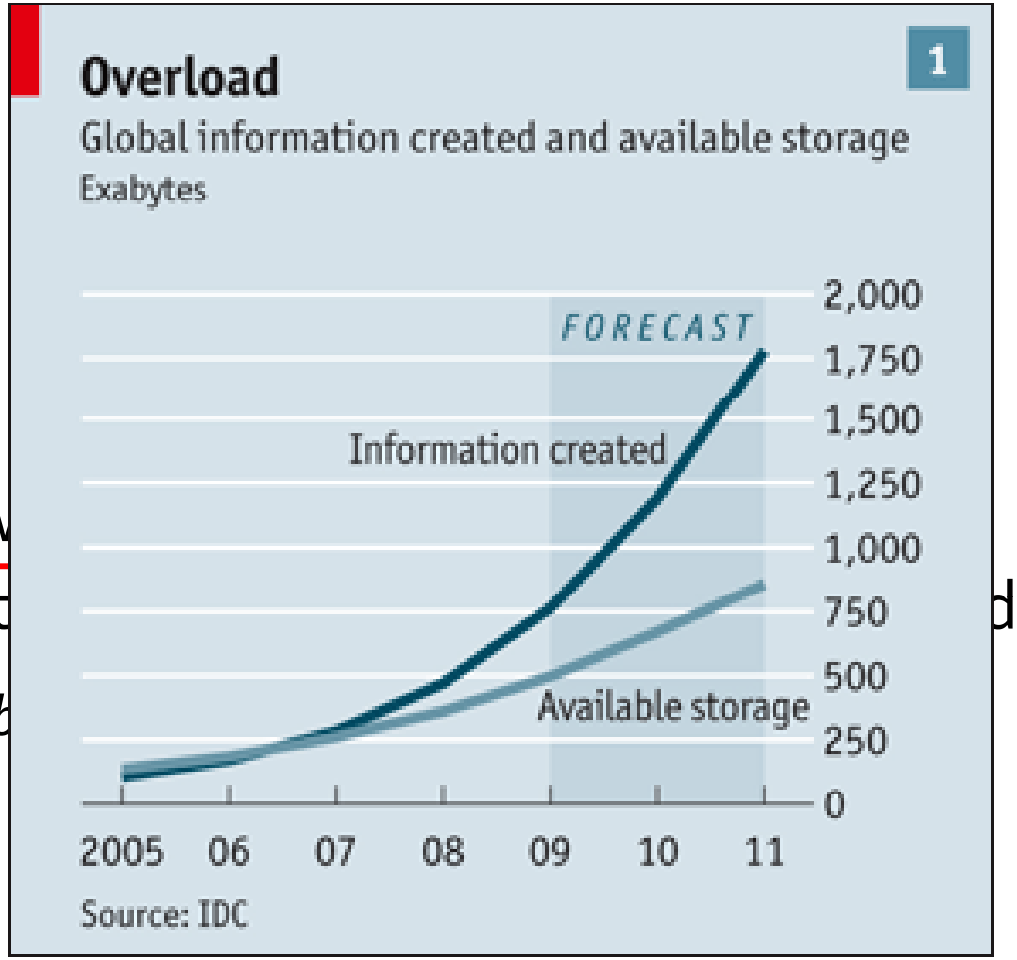
What is Big

"Big data refers to data sets w
typical database software t
analyze." - *The McKinsey Glob*



What is Big

"Big data refers to data sets w
typical database software t
analyze." - *The McKinsey Glob*



"Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. - *http://en.wikipedia.org/wiki/Big_data*

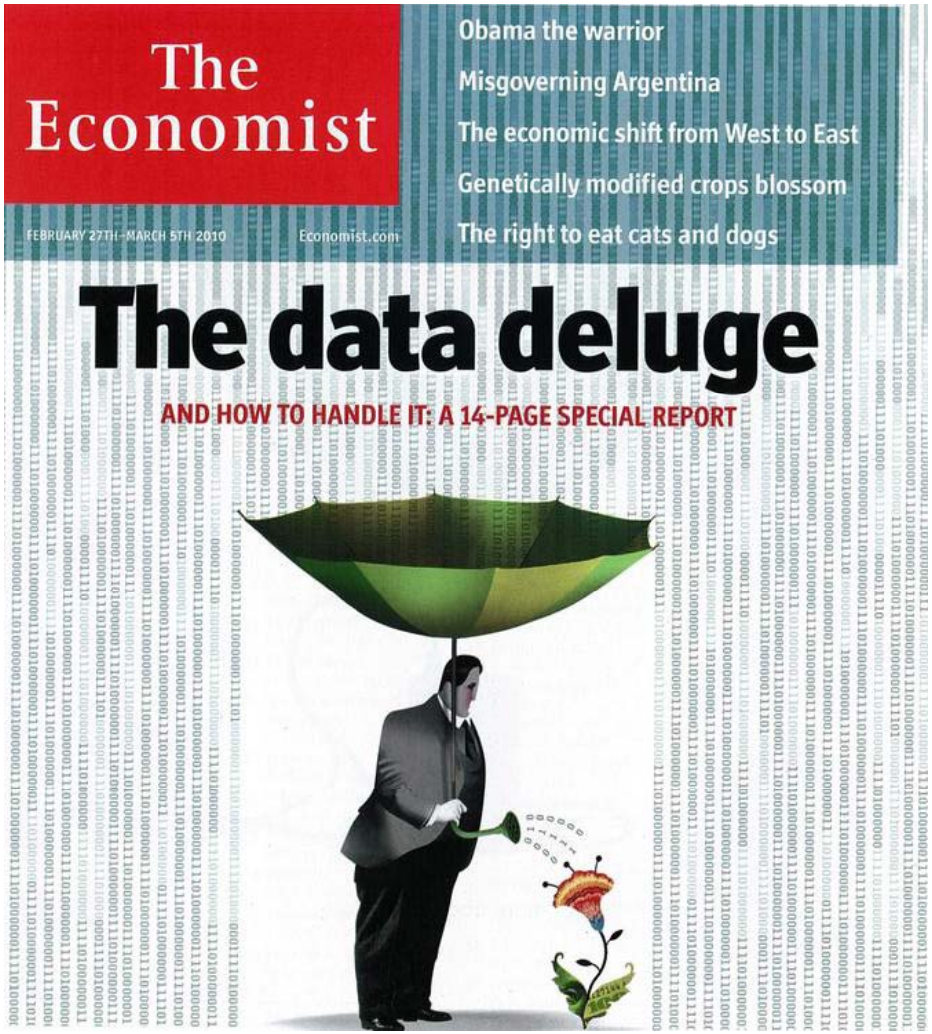
How *big* is Big Data ?

Earlier Berkeley studies estimated that by the end of 1999, the sum of human-produced information (including all audio, video recordings and text/books) was about 12 Exabytes of data.

Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003.

http://techcrunch.com/2010/08/04/schmidt-data/

Big picture of Big Data 1



Source: The economist Feb 25 2010

In 2010 The digital Universe was

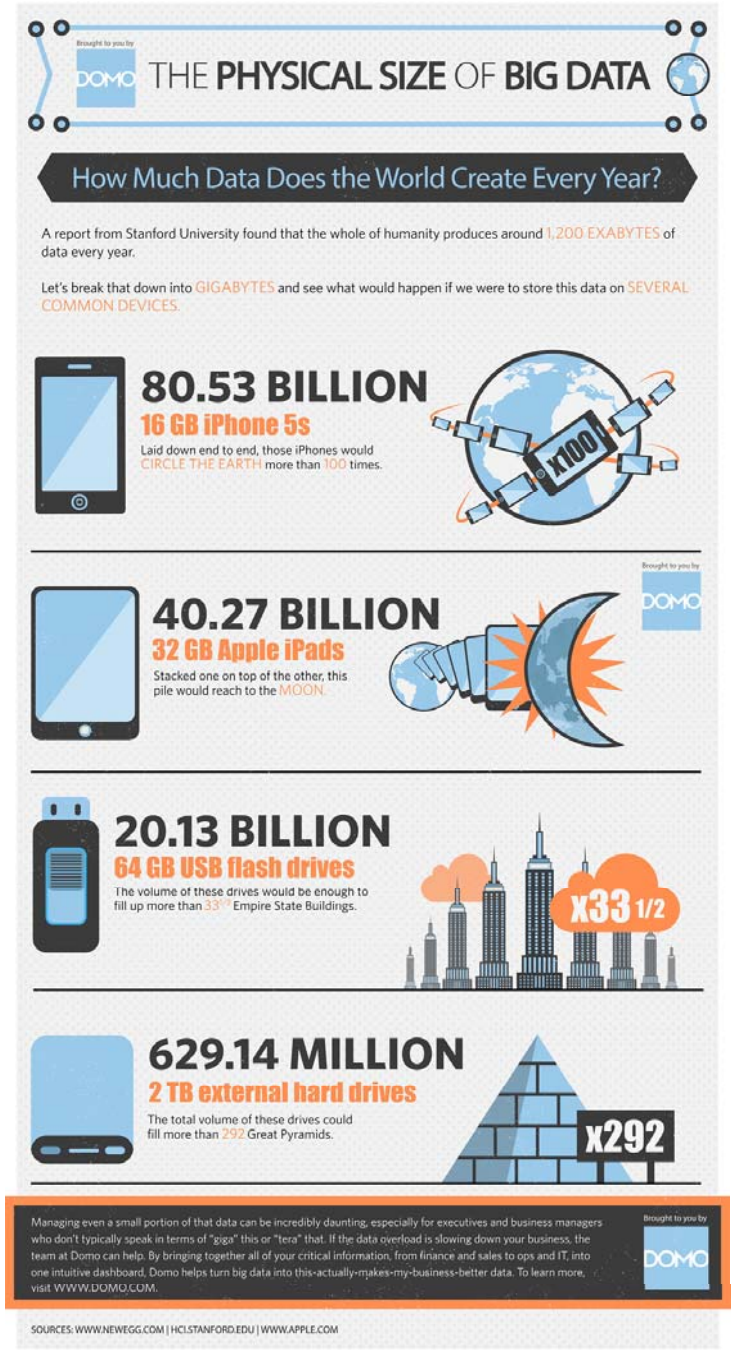
1.2 ZettaBytes

In a decade the Digital Universe will be

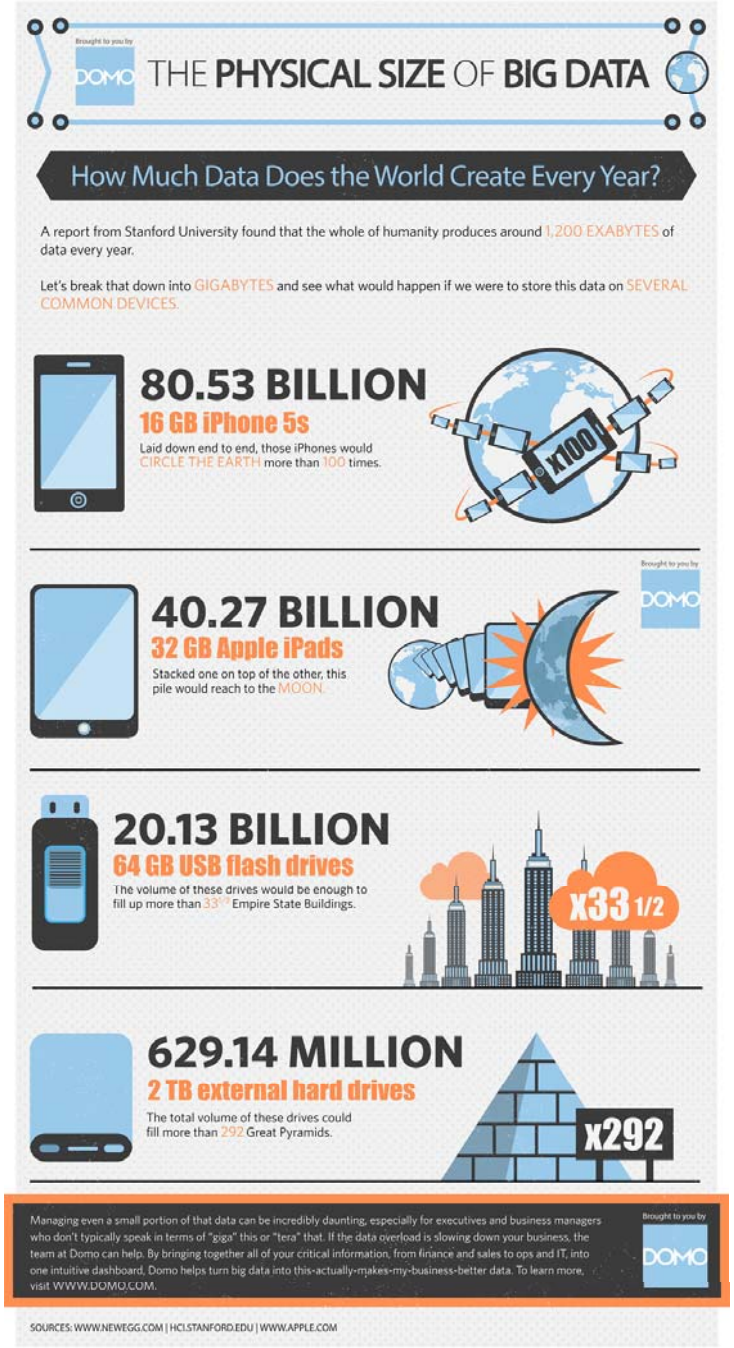
35 ZettaByte

1.2 Zettabyte?

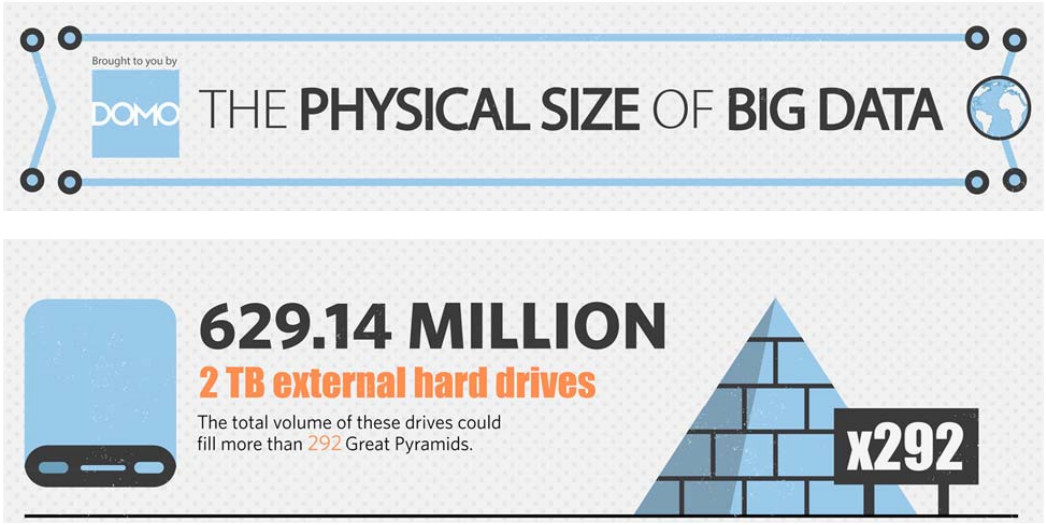
1.2 Zettabyte?

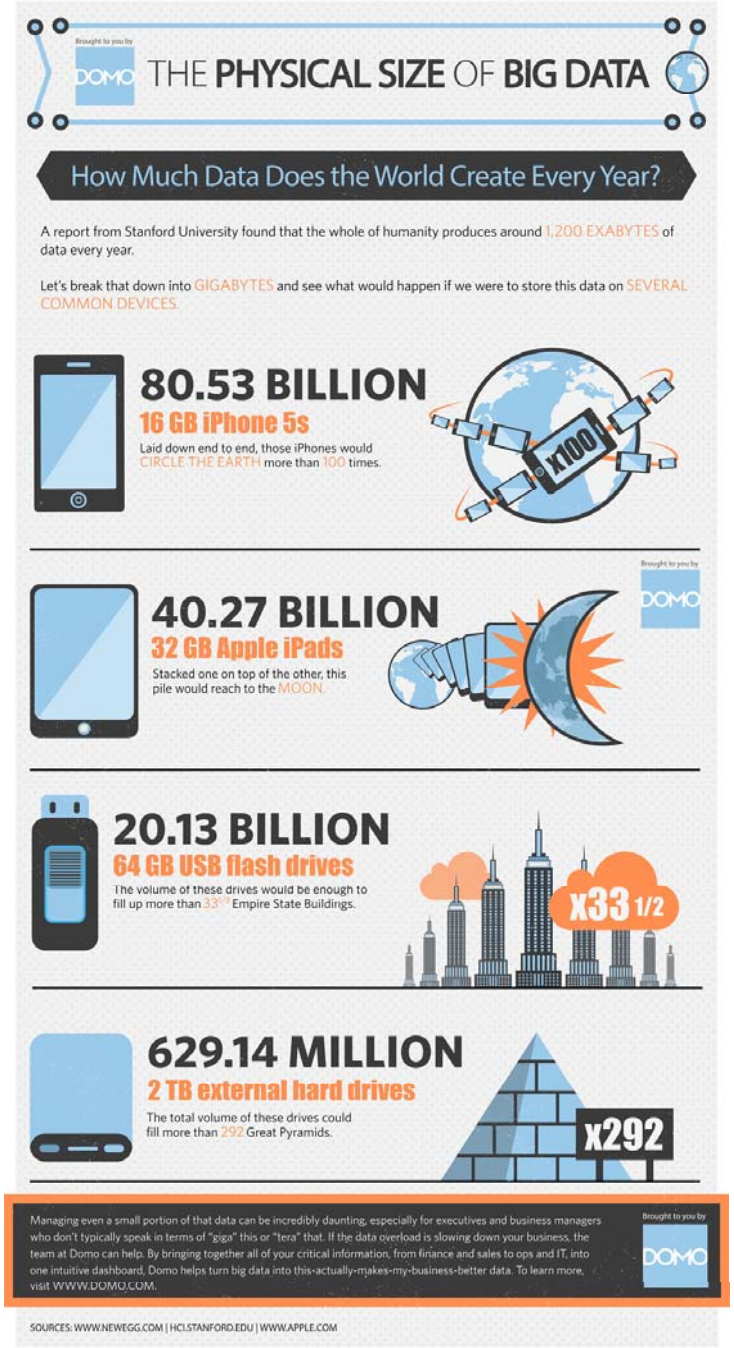


1.2 Zettabyte?

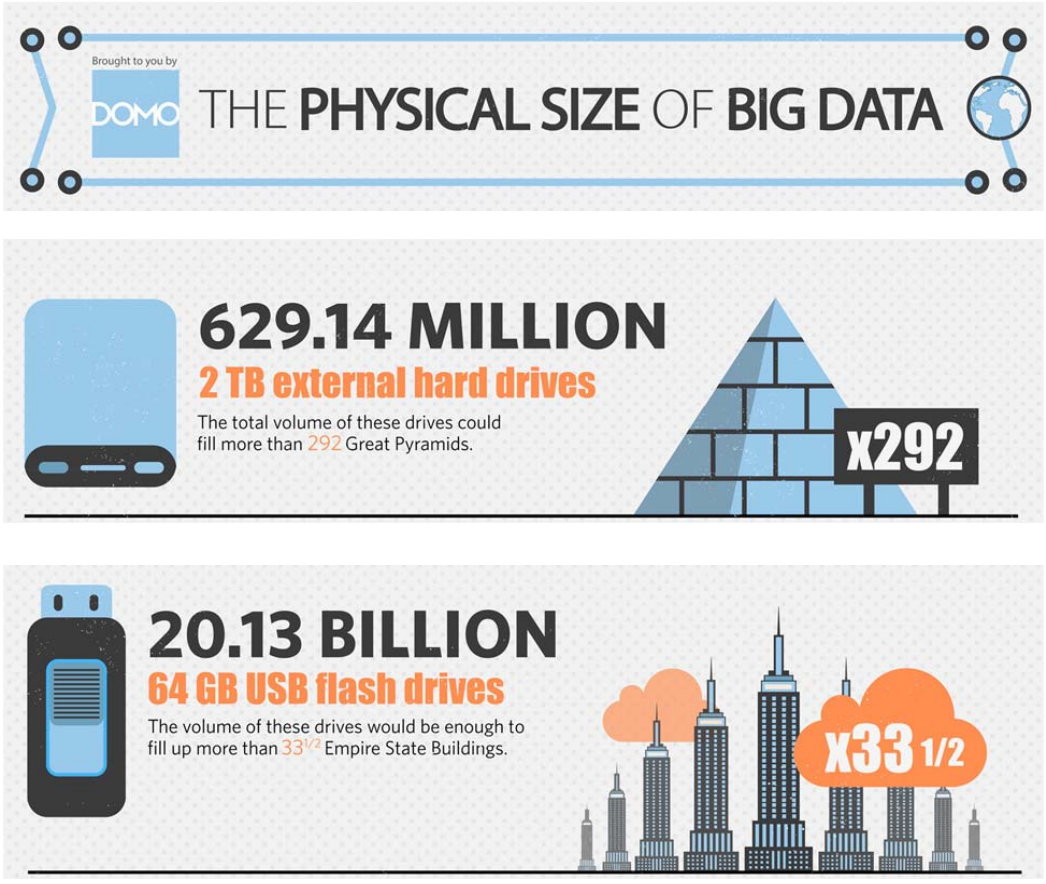


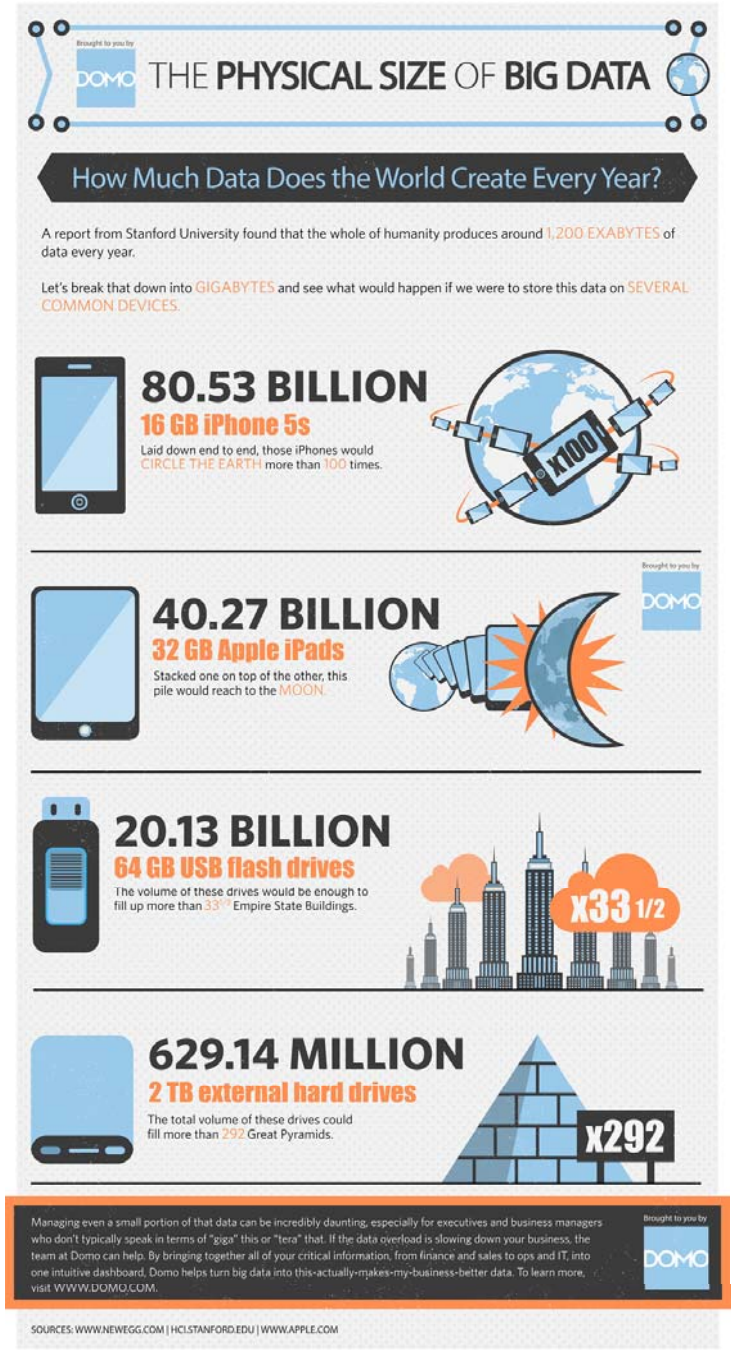
1.2 Zettabyte?





1.2 Zettabyte?





Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in G… One page of text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. … A typical pop song is … MB.
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek … GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in G… … in America's Library …
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered … … to around 5PB. Google …
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billi…
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of … … this year is forecast …
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to …

The prefixes are set by an intergovernmental … …es.

Sources: The Economist … Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Data inflation

2

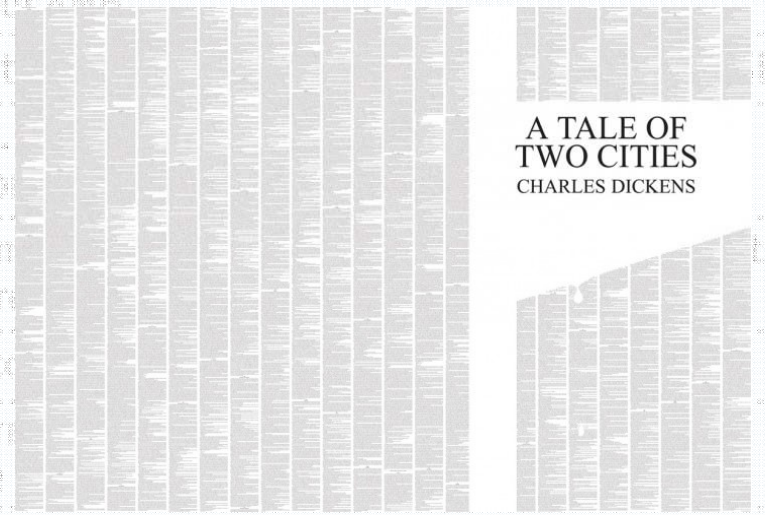
Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing

Variable Type	Keyword	Bytes Required	Range
Character	char	1	-128 to 127
Unsigned character	unsigned char	1	0 to 255
Integer	int	2	-32768 to 32767
Short Integer	short int	2	-32768 to 32767
Long Integer	long int	4	-2,147,483,648 to 2,147,438,647
Unsigned Integer	unsigned int	2	0 to 65535
Unsigned Short integer	unsigned short int	2	0 to 65535
Unsigned Long Integer	unsigned long int	4	0 to 4,294,967,295
Float	float	4	1.2E-38 to
Double	double	8	2.2E-308 to
Long Double	long double	10	3.4E-4932 to 1.1E+4932

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. in America's Library of Co
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by An to around 5PB. Google pr
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion c
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of infor this year is forecast to be
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine



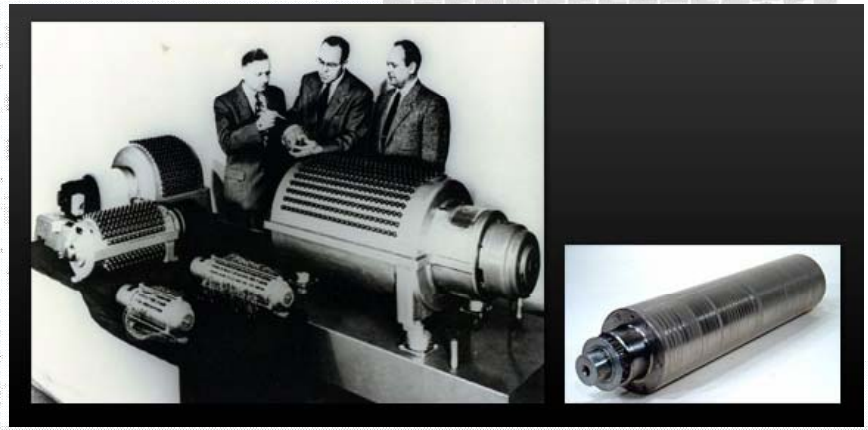
The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

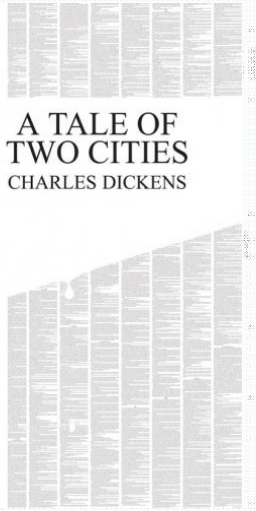
Source: http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB





Source: http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

Magnetic drum memory
A 16-inch-long drum from the IBM 650 computer. It had 40 tracks, 10 kB of storage space, and spun at 12,500 revolutions per minute.

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek.
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Gree in America's Library of
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by to around 5PB. Google
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of inf this year is forecast to
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to ima

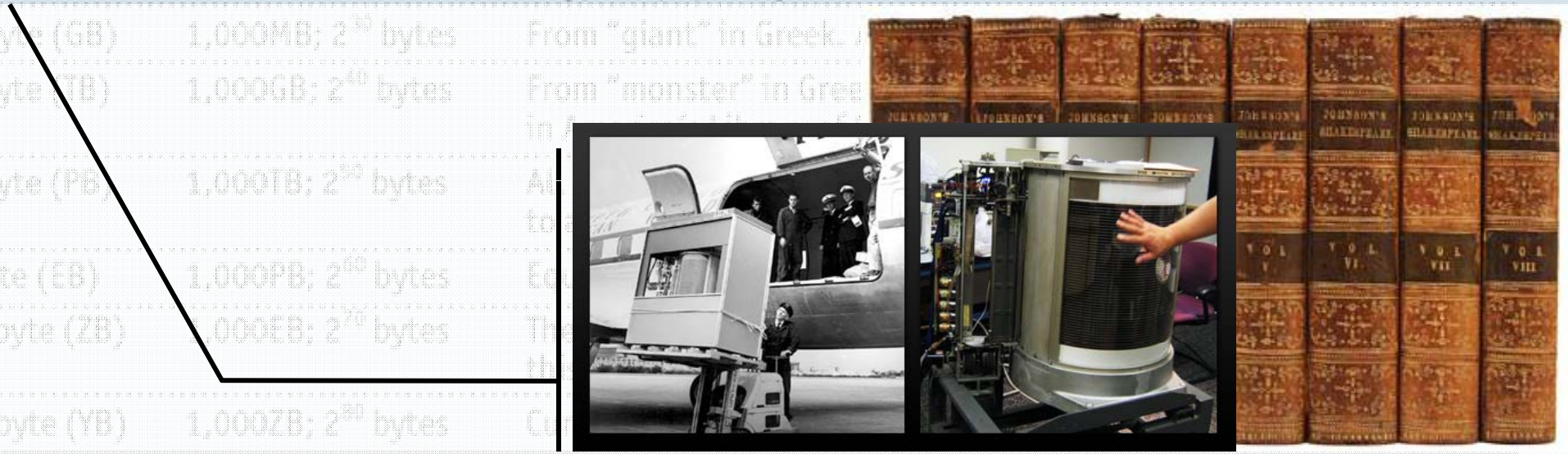
The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB



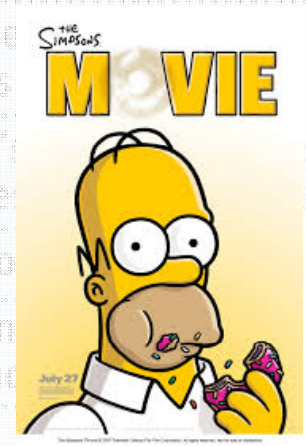
IBM Model 350, the first-ever hard disk drive.
5 Million Characters (5MB).

Source: http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the … in America's Library of Congress
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's … to around 5PB. Google processes …
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of …
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information … this year is forecast to be around …
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine



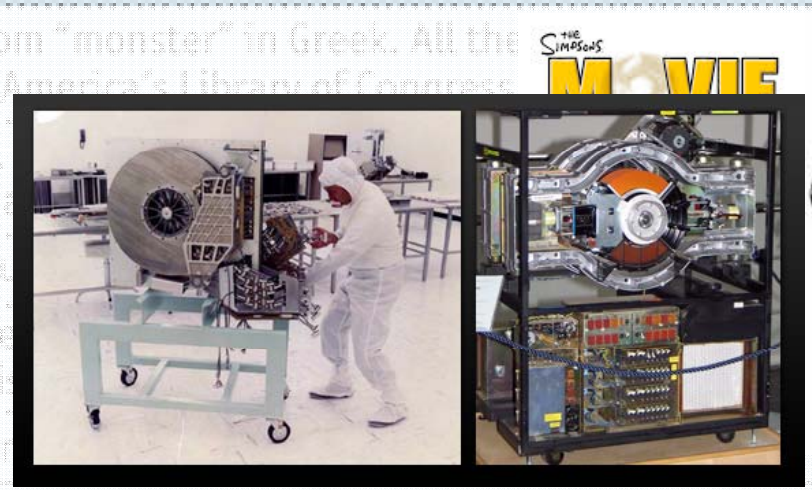
There are names by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB



Source: http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

A 250 MB hard disk drive from 1979

The IBM 3380 from 1980, the first gigabyte-capacity hard disk drive. (250KG, 81K$)

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google proces
Exabyte (EB)		Equivalent to 10 billion copie
Zettabyte (ZB		The total amount of informat this year is forecast to be aro
Yottabyte (YB		Currently too big to imagine

1 TB (500 $)

The prefixes are set by an intergovernmental group, the easures.
Source: The Economist Yotta and Zetta were added in 1991; term ublished.

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*

Next 500 TB / month	$0.0285 per GB
Next 4000 TB / month	$0.0280 per GB
Over 5000 TB / month	$0.0275 per GB

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures.

Source: *The Economist* Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	
Megabyte (MB)	1,000KB; 2^{20} bytes	
Gigabyte (GB)	1,000MB; 2^{30} bytes	
Terabyte (TB)	1,000GB; 2^{40} bytes	
Petabyte (PB)	1,000TB; 2^{50} bytes	
Exabyte (EB)	1,000PB; 2^{60} bytes	
Zettabyte (ZB)	1,000EB; 2^{70} bytes	
Yottabyte (YB)	1,000ZB; 2^{80} bytes	

1 EB =

8,192,000	128GB
217,872,340	DVD 4.7GB
1,497,965,714	700MB
728,177,777,778	1.44MB
500,000	2TB
4,772,185,884,444	0.21MB
9,162,596,898,133,330	0.12KB

The prefixes are s… Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: The Economist

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	
Megabyte (MB)	1,000KB; 2^{20} bytes	
Gigabyte (GB)	1,000MB; 2^{30} bytes	
Terabyte (TB)	1,000GB; 2^{40} bytes	
Petabyte (PB)	1,000TB; 2^{50} bytes	
Exabyte (EB)	1,000PB; 2^{60} bytes	
Zettabyte (ZB)	1,000EB; 2^{70} bytes	
Yottabyte (YB)	1,000ZB; 2^{80} bytes	

The prefixes are s... Yotta and Zetta were ...

Source: The Economist

Source: https://www.backupify.com/history-of-data-storage/

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: The Economist

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

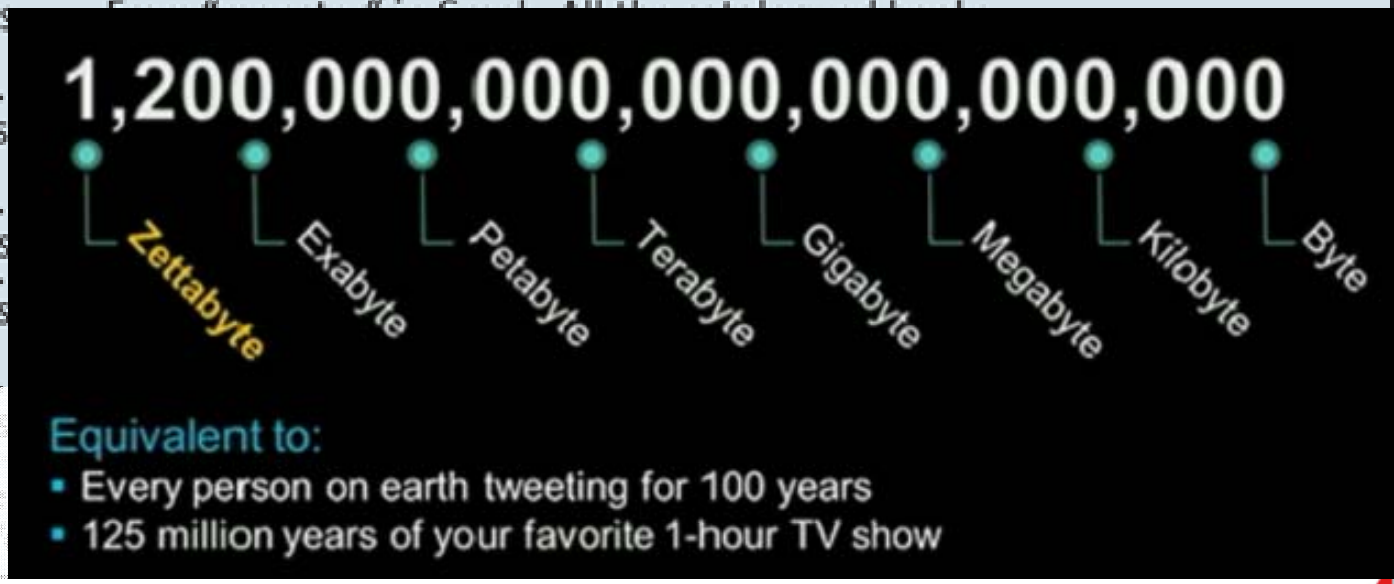
Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	
Petabyte (PB)	1,000TB; 2^{50} bytes	
Exabyte (EB)	1,000PB; 2^{60} bytes	
Zettabyte (ZB)	1,000EB; 2^{70} bytes	
Yottabyte (YB)	1,000ZB; 2^{80} bytes	

The prefixes are

Source: The Economist



Data inflation

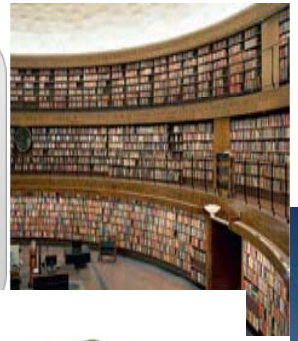
Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", af computers use to store an
Byte (B)	8 bits	Enough information to cre in computer code. It is the
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek.
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The A typical pop song is about
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A tw
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. in America's Library of Con
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by Am to around 5PB. Google pro
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

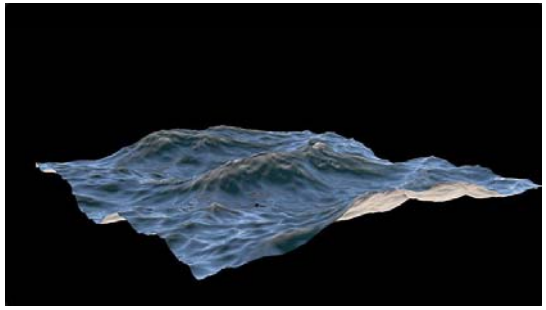
The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

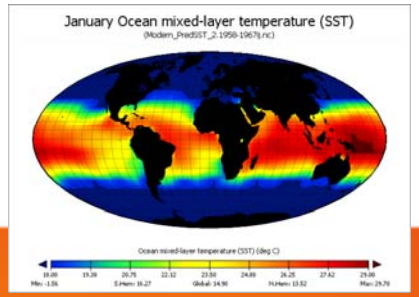
Source: *The Economist*

What are the sources of Big Data?

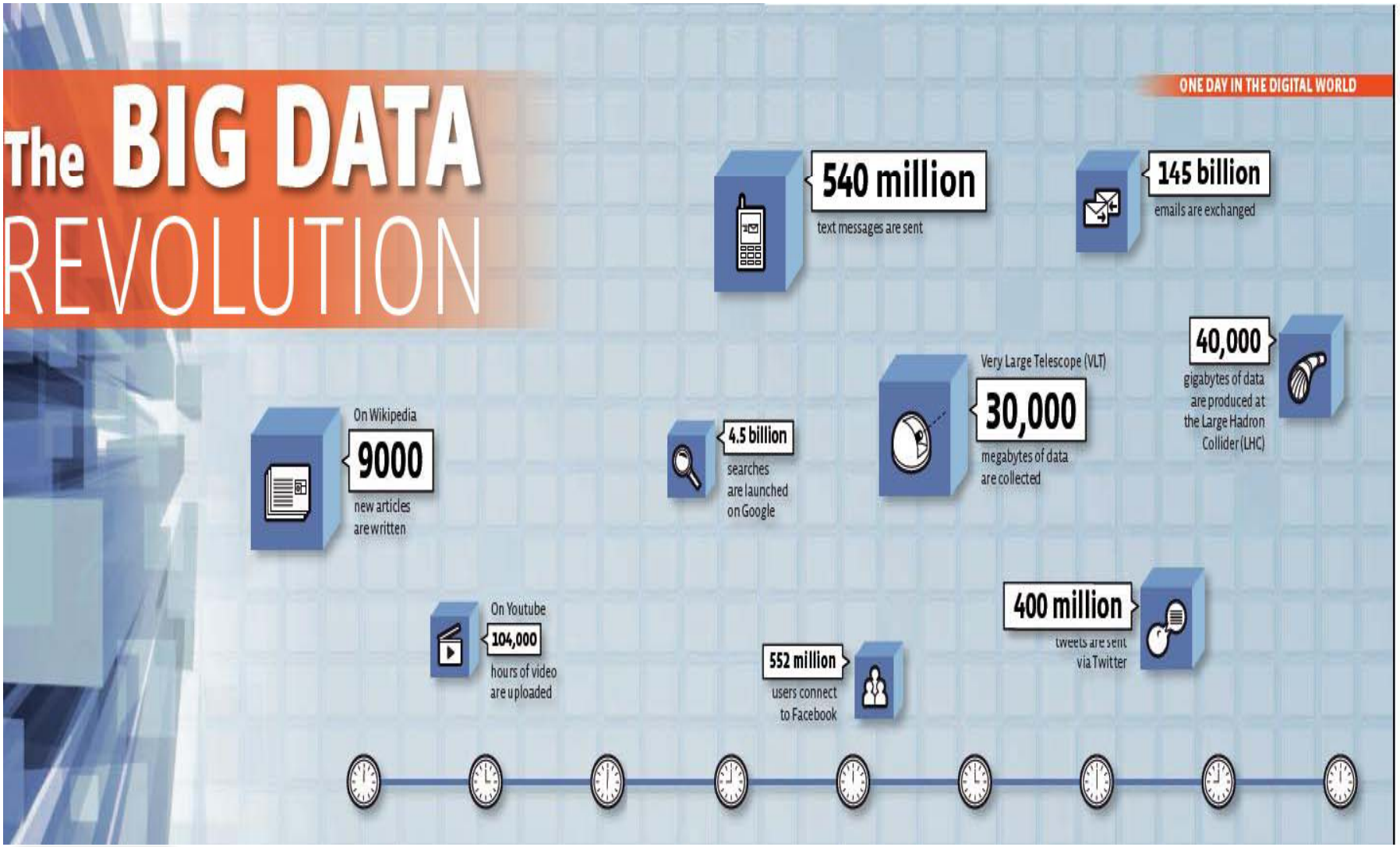
What are the sources of Big Data?





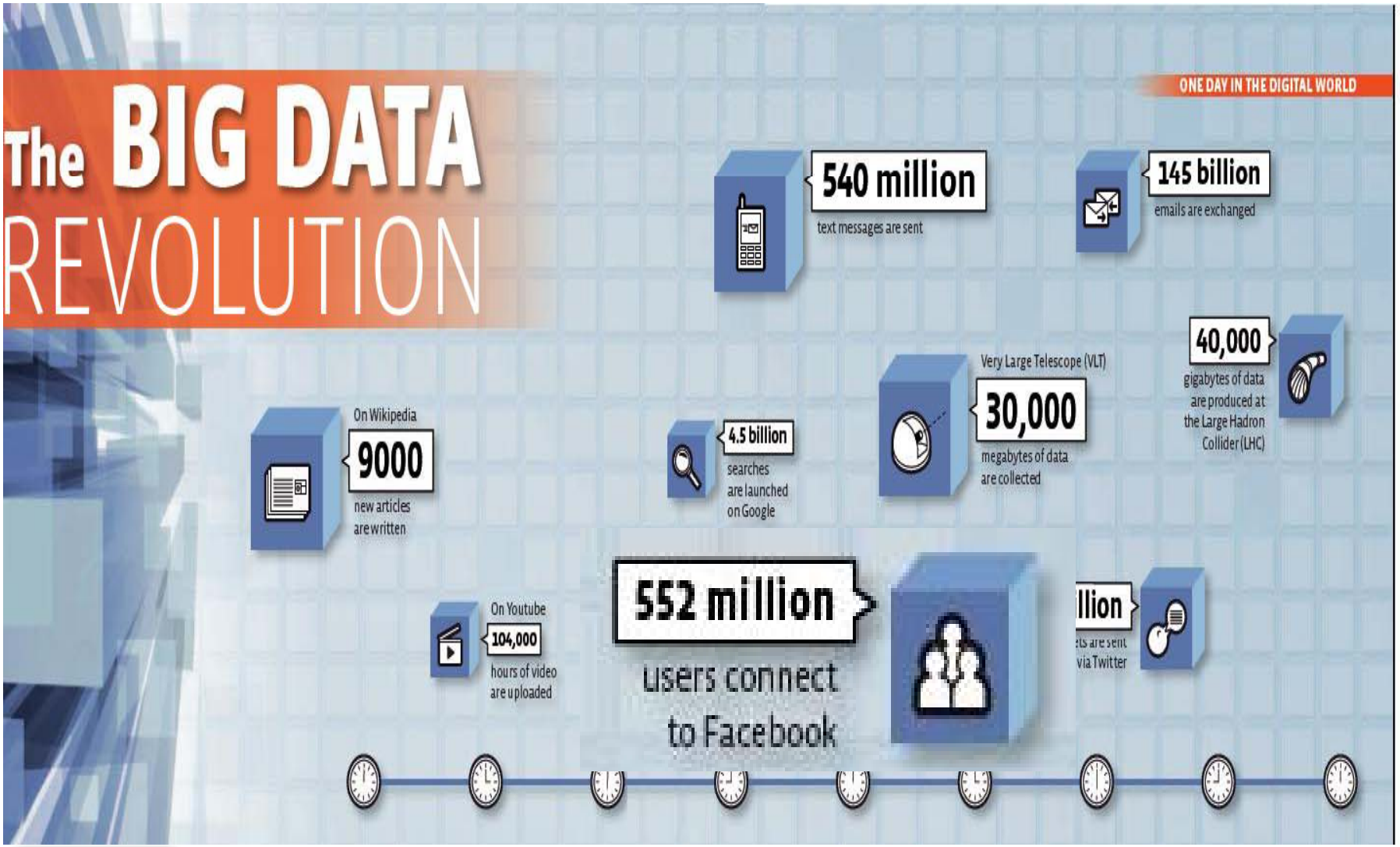


How much Data are we producing in a day ?



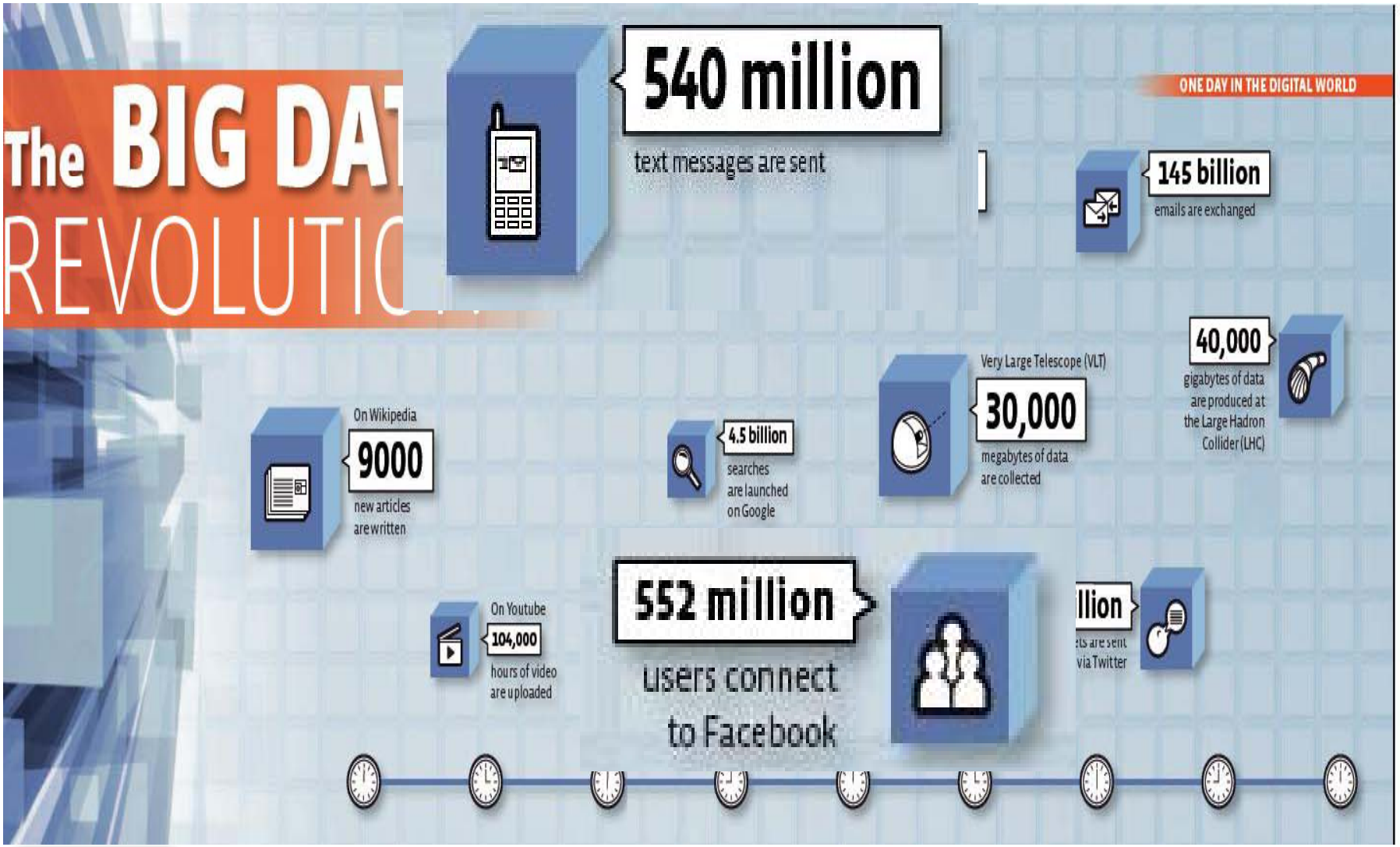
Source: CNRS Magazine 2013

How much Data are we producing in a day ?



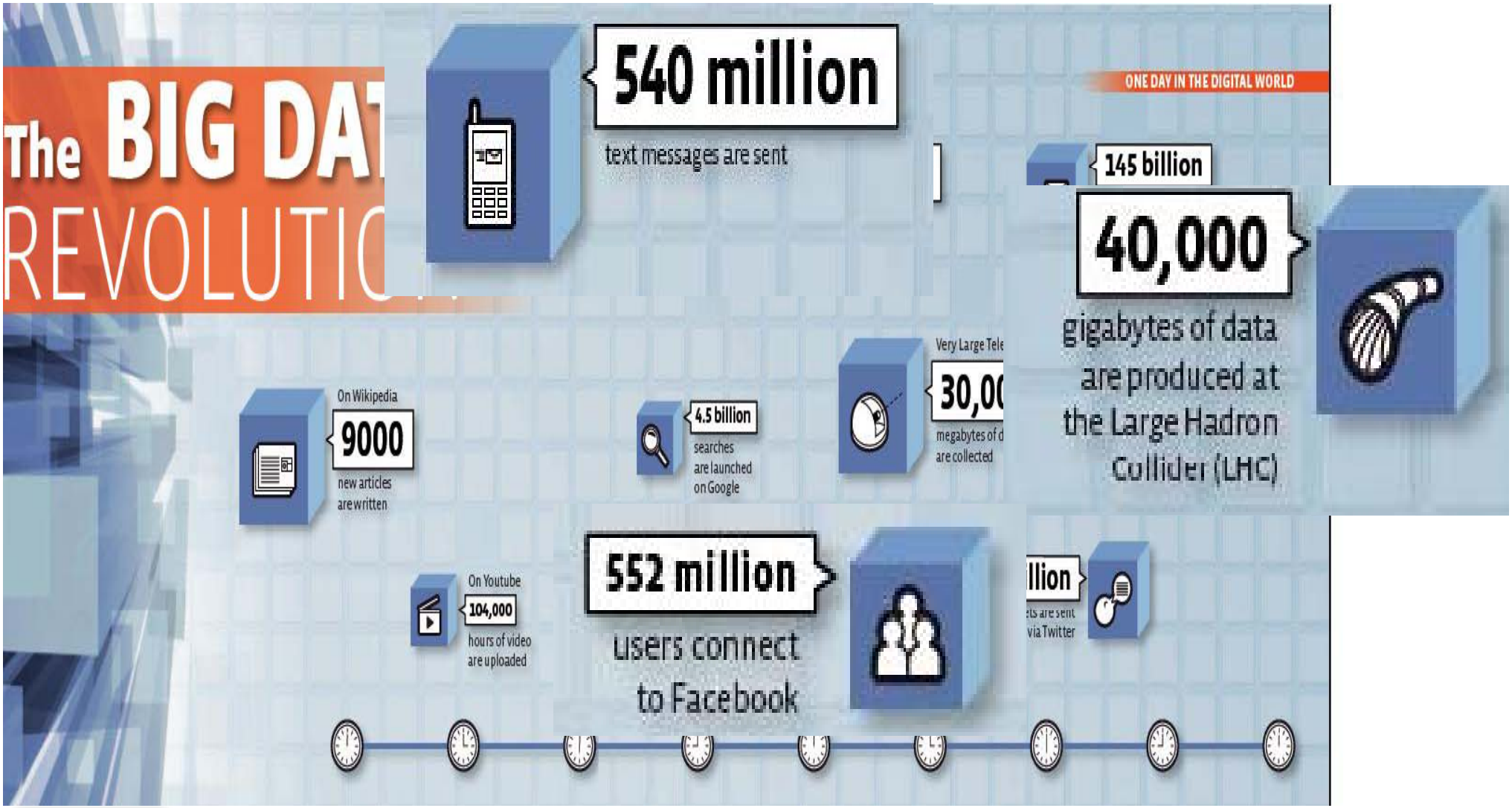
Source: CNRS Magazine 2013

How much Data are we producing in a day ?



Source: CNRS Magazine 2013

How much Data are we producing in a day ?



Source: CNRS Magazine 2013

IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WordPress DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
1,700+ Firefox DOWNLOADS
79,364 WALL POSTS
510,040 COMMENTS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
600+ NEW VIDEOS
25+ HOURS TOTAL DURATION
100+ Answers.com
40+ YAHOO! ANSWERS
QUESTIONS ASKED ON THE INTERNET...
20,000+ NEW POSTS ON tumblr.
13,000+ iPhone APPLICATIONS DOWNLOADED
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads
Google
Google Search
GO-Globe.com

Our Data-Driven world: **Science**

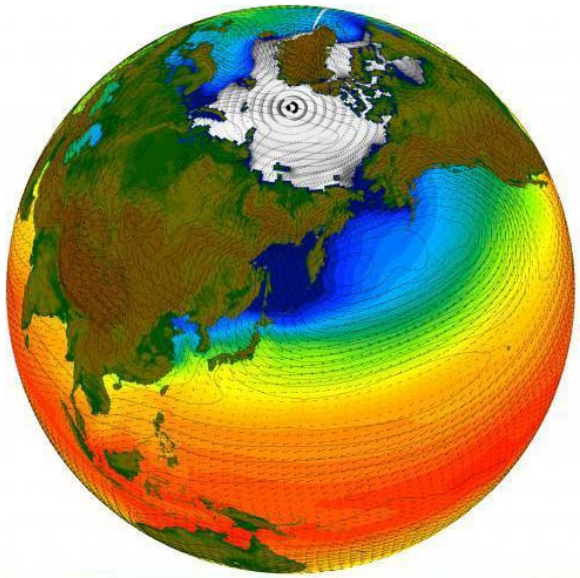
Particle accelerators in physics

*CERN's Large Hydron Collider (LHC) generates **15 PB** a year*

Astronomical instruments

*SQUARE KILOMETRE ARRAY (SKA) :The world's largest radio telescope) will collect **1 PB** a day ~ **400 PB** a year*

Climate Simulations



The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.

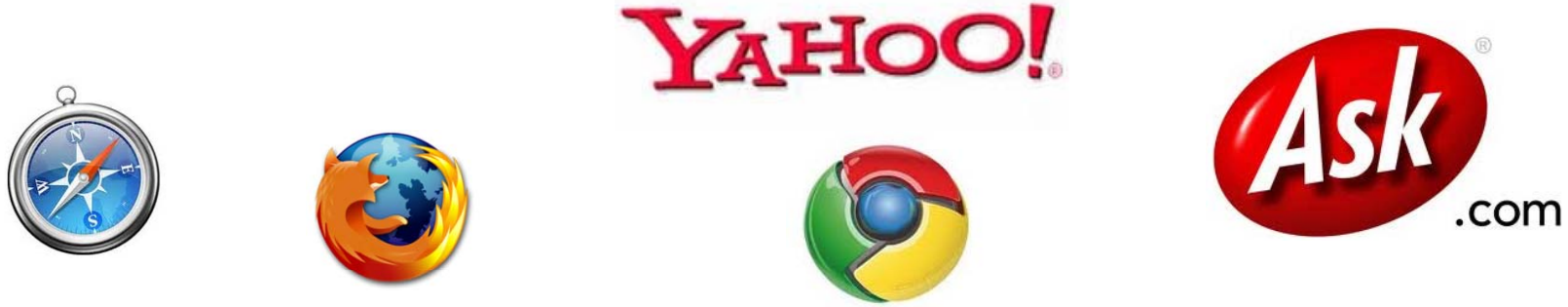
Genome sequencers in biology

Our Data-Driven world

Web Data

Our Data-Driven world

Web Data



- Google processes **20 PB** a day (2008)
- eBay has **6.5 PB** of user data + 50 TB/day (5/2009)

Our Data-Driven world

Social Networks

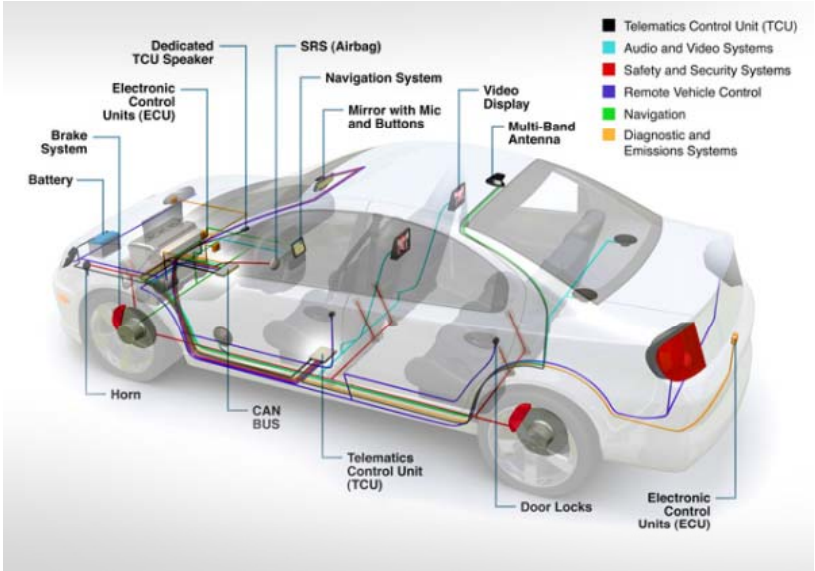
Our Data-Driven world

Social Networks

- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- Twitter Generate approximately 12 TB of data per day

Our Data-Driven world

Industry



AIRBUS

A single airplane engine generates more than 10 TB of data every 30 minutes.

Our Data-Driven world

Business & Commerce

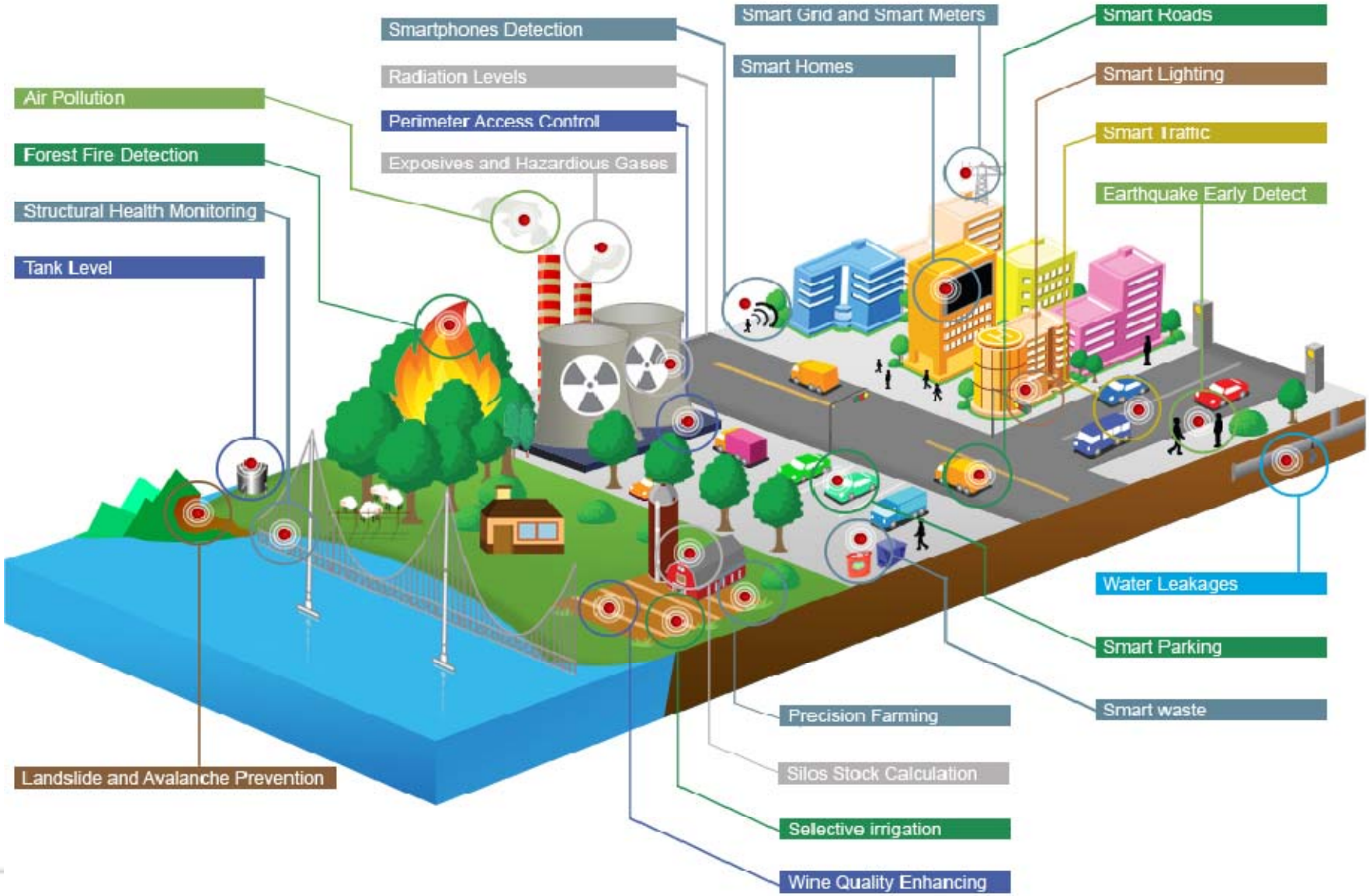
Our Data-Driven world

Business & Commerce

- New York Stock Exchange **1TB** of data everyday
- Walmart's customer transactions generate 2.5 petabytes of data every hour.

Our Data-Driven world

Internet of Things



More Examples

GPS Systems

Internet Service Providers

T-Mobile

orange™

at&t

Credit Card Companies

MasterCard

VISA

SFR

Our Data-Driven world ....and More

Phone Systems

Movie Rental Sites

HDTV players with internet connectivity

Recommendation Systems

What is Big Data used for?

What is Big Data used for?

- Harnessing scientific discoveries

What is Big Data used for?

- Harnessing scientific discoveries

- Initiating early warning of natural disasters (e.g., floods, volcanic eruptions, and earthquakes)

What is Big Data used for?

- Harnessing scientific discoveries

- Initiating early warning of natural disasters (e.g., floods, volcanic eruptions, and earthquakes)

- Reports
  - » Track business processes, transactions

What is Big Data used for?

- Diagnosis

- Decisions

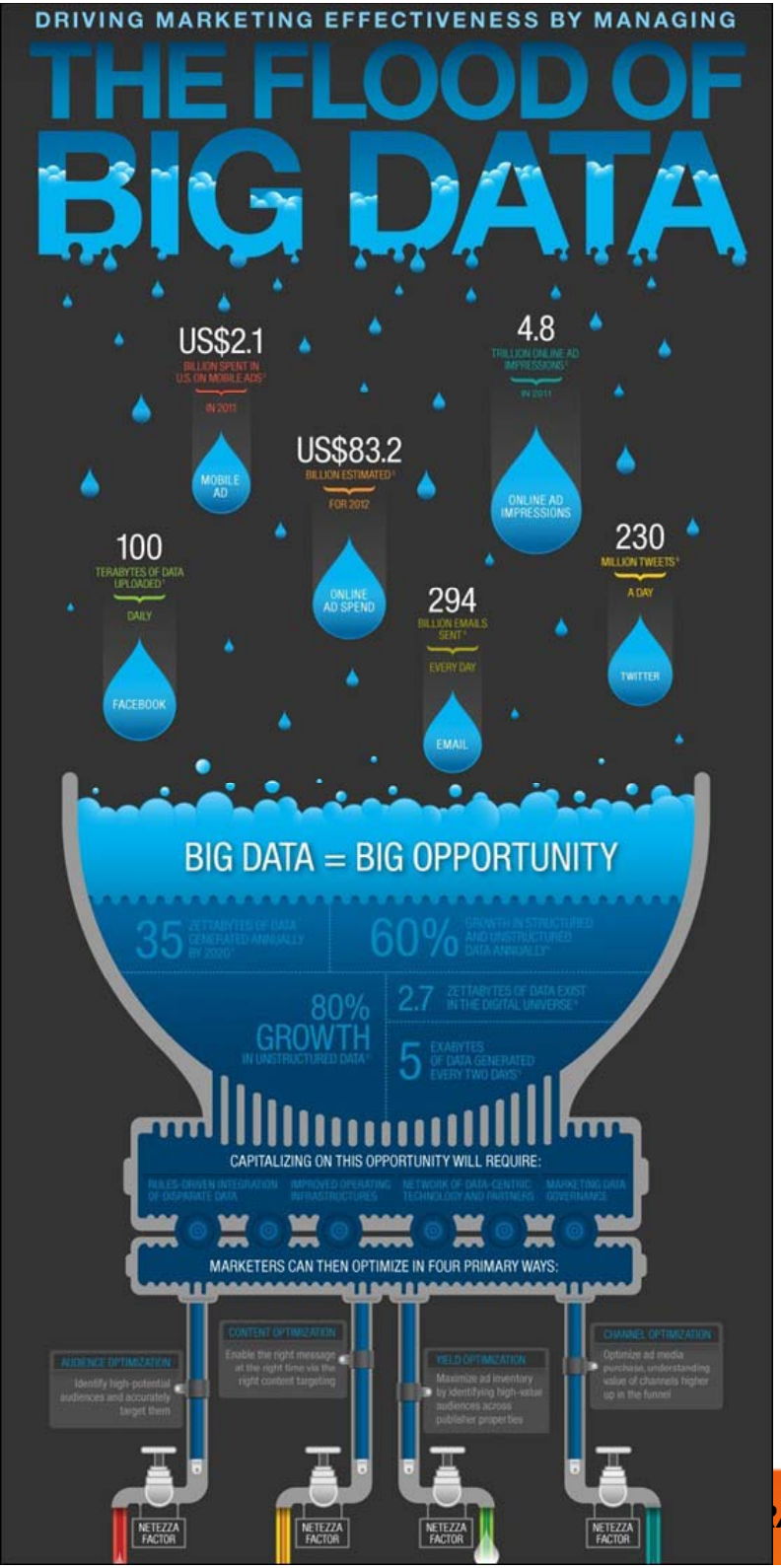
What is Big Data used for?

- Diagnosis
  - » Why is user engagement dropping?
  - » Why is the system slow?
  - » Prevent failures
  - » Detect spam, worms, viruses, DDoS attacks
- Decisions

What is Big Data used for?

- Diagnosis
  - » Why is user engagement dropping?
  - » Why is the system slow?
  - » Prevent failures
  - » Detect spam, worms, viruses, DDoS attacks
- Decisions
  - » Personalized medical treatment
  - » Decide what ads to show

5 Ways Big Data Will Change the World



Big Data @ Work

Organizations in all industries are under increasing pressure to capitalize on data

Healthcare

The average amount of data per hospital will increase from 167TB to 665TB in 2015, driven by the enormous growth of medical images and electronic medical records.[1]

With Big Data

Medical professionals can improve patient care and reduce costs by extracting relevant clinical information from vast amounts of data to better understand the past and predict future outcomes.

Customer Service

Today, 86% of consumers quit doing business with a company because of a bad customer experience, up from 59% four years ago.[2]

With Big Data

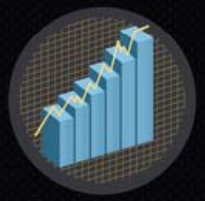
Service representatives can use data to gain a more holistic view of their customers, understanding their likes and dislikes in real-time in order to resolve a problem or capitalize on happy clients faster.

Insurance

Insurance companies and government agencies each gather fraud data related to their own individual missions. But the kind, quality and volume of data compiled varies widely.[3]

With Big Data

An insurance or citizen services provider can apply advanced analytics to data and detect fraud quickly, before funds are paid out.



Financial Services

Wall Street alone delivers 5 new research documents every minute. Dow Jones publishes upwards of 19,000 news items per day.[4]

With Big Data

Financial services professionals can better understand market changes through improved business insight from data, helping to anticipate performance gaps and more accurately assess investment alternatives.

Retail

$165 billion in total sales are missed each year because retailers don't have the right products in stock to meet customer demand.[5]

With Big Data

Retailers can better understand their customers by analyzing sales trends and incorporating more accurate forecasting, ultimately increasing customer loyalty and revenue.

Communications

5 billion global subscribers in the telco industry are demanding unique and personalized offerings that match their individual lifestyles.[6]

With Big Data

Communications providers can use data to create a more personalized customer experience and avoid losing customers to competitors.

*Information gathered by IBM

1 North American Health Care Provider Information Market Size & Forecast, Enterprise Strategy Group
2 Customer Experience Impact Report, Harris Interactive
3 Coalition Against Insurance Fraud
4 The Future of News in Trading, Dow Jones, June 2011
5 IHL Group, Dec 2011
6 IBM Global CFO Study 2010

© Copyright IBM Corporation 2012. All rights reserved.

Medicine

Aetna is using reams of data to try to get early diagnosis, **prevention and treatment of heart disease and diabetes**.

UCLA is using Big Data analysis to prevent **complications from brain injuries**.

The American Society of Clinical Oncology is using Big Data to help it **find the best treatments for cancer**.

http://insights.wired.com/profiles/blogs/5-ways-big-data-will-change-the-world#axzz3nHefVA1j

Security

There is a **pedometer** application that can actually **identify people based on their gait, how they walk**.

A new security firm called **Pindrop** is using Big Data analysis to help banks and other financial institutions **identify callers** to ensure the person on the other end of the line is who they say t..

Pindrop is able to listen to more than *100 different background sounds on a phone call* to tell where the call is coming from and whether it is a cell phone, land line, of VOIP. They can tell you if the person claiming to be in Nebraska is actually calling from Nigeria.

http://insights.wired.com/profiles/blogs/5-ways-big-data-will-change-the-world#axzz3nHefVA1j

Urban Planning

Tracking the movements of people and how that could impact urban planning.

Cities are using data discovery techniques to examine the myriad of ways small changes can impact a big urban centers.

The Urban Center for Computational Data talks about *computer models helping cities to figure out how things like a new bus line might impact crime, employment, and energy usage in parts of a city*. There is little question that how our cities are built and function will be changed by data analytics.

http://insights.wired.com/profiles/blogs/5-ways-big-data-will-change-the-world#axzz3nHefVA1j

Consumer Products

The tremendous rise in online shopping has created piles of data to better *understand what consumers want and how they shop*.

It even allows companies to customize their *pricing models based on who is shopping and when they want to buy*.

http://insights.wired.com/profiles/blogs/5-ways-big-data-will-change-the-world#axzz3nHefVA1j

Elections

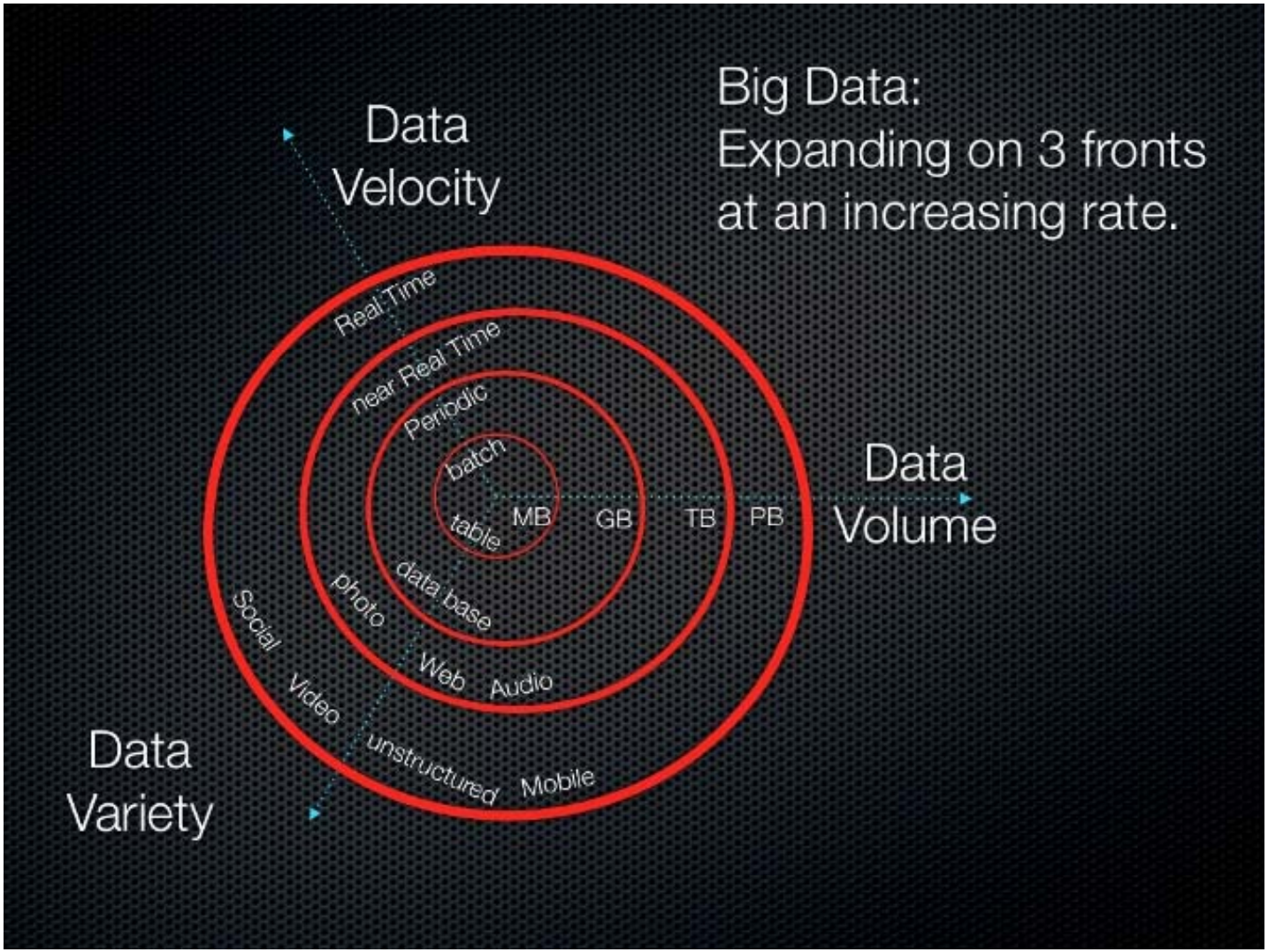
In the 2012 presidential election, the Obama Campaign made use of voter models on a scale never before seen.

They were able to *identify specific voters who would make a difference in the election and target messages to those voters*.

I am not talking about something general like, "**we need to appeal to soccer moms**," I am talking about true specifics like, "**the Johnson family Maple Lane in Columbus, Ohio will vote for us if they know our stance on social security**." It seems insane to think that presidential politics has gotten that local, but it has and it worked. There is little question that the Obama campaigns sophisticated methods of get out the vote and swing voter identification swung a very close election their way.

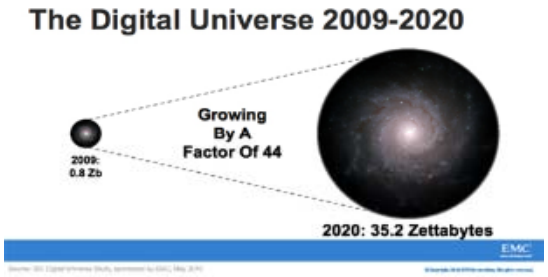
http://insights.wired.com/profiles/blogs/5-ways-big-data-will-change-the-world#axzz3nHefVA1j

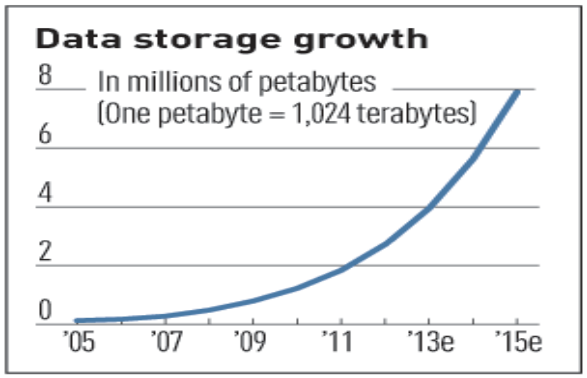
Big Data: Challenges



Volume (Scale)

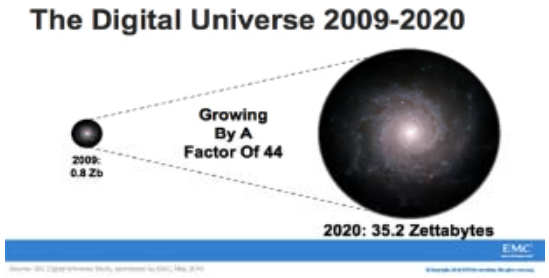
- **Data Volume**
  - » 44x increase from 2009 2020
  - » From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially

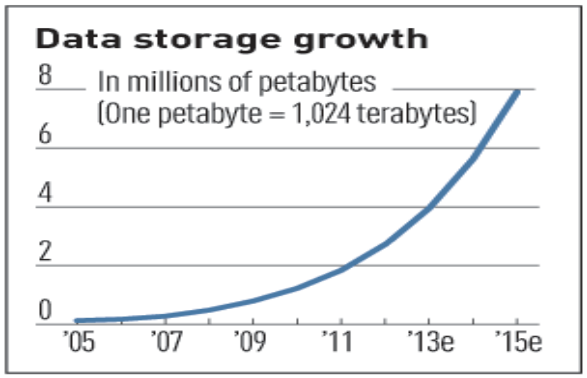




Volume (Scale)

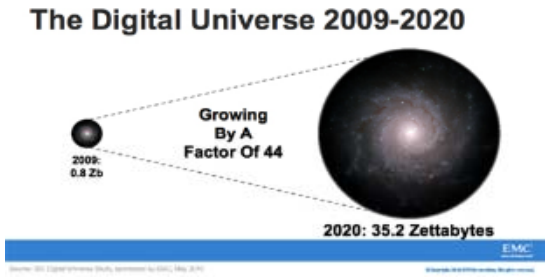
- **Data Volume**
  - » 44x increase from 2009 2020
  - » From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially

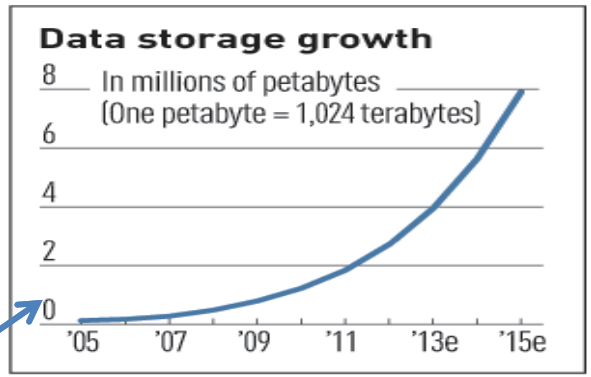


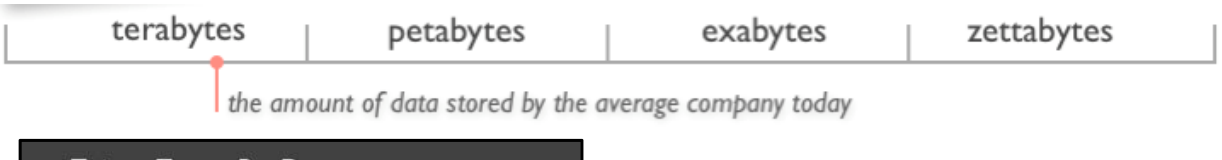


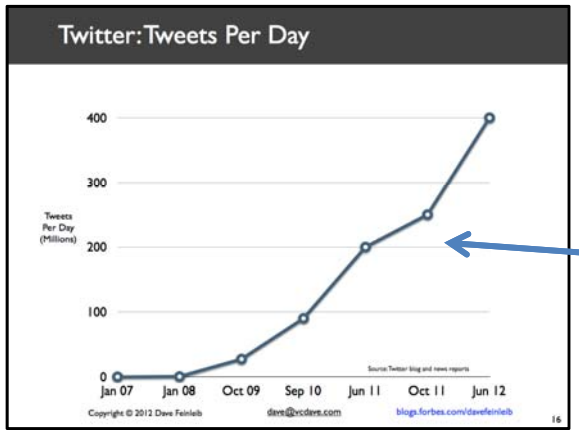
Volume (Scale)

- **Data Volume**
  - » 44x increase from 2009 2020
  - » From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially









Exponential increase in collected/generated data

12+ TBs
of tweet data
every day

12+ TBs
of tweet data every day

25+ TBs **of log data every day**

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs **of log data every day**

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs **of log data every day**

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs **of log data every day**

30 billion
RFID tags today
(1.3B in 2005)

4.6 billion
camera phones world wide

100s of millions of GPS enabled
devices sold annually

2+ billion
people on the Web by end 2011

76 million smart meters in 2009…
200M by 2014

Variety (Complexity)

Relational Data (Tables/Transaction/Legacy Data)

Text Data (Web)

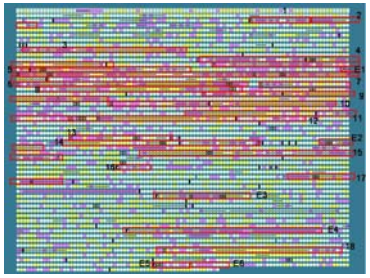
Semi-structured Data (XML)

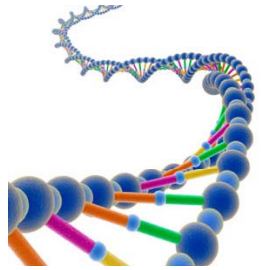
Graph Data
- » Social Network, Semantic Web (RDF), …

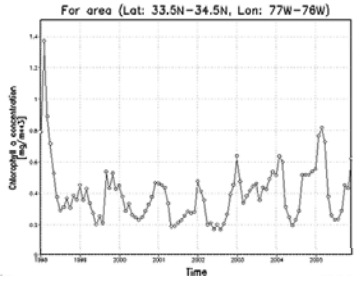
Streaming Data
- » You can only scan the data once

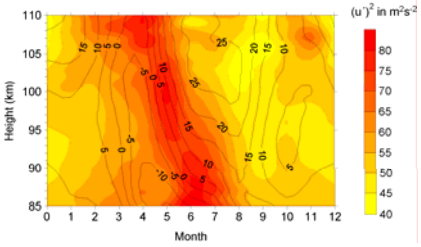
A single application can be generating/collecting many types of data

Big Public Data (online, weather, finance, etc)









Variety (Complexity)

Relational Data (Tables/Transaction/Legacy Data)

Text Data (Web)

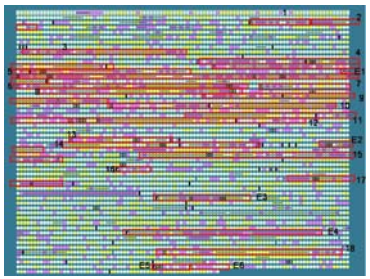
Semi-structured Data (XML)

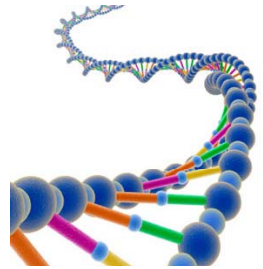
Graph Data
- » Social Network, Semantic Web (RDF), …

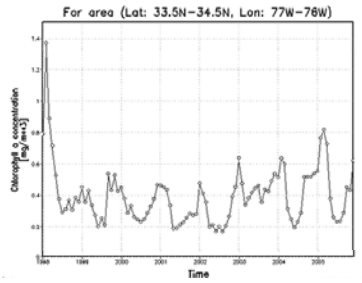
Streaming Data
- » You can only scan the data once

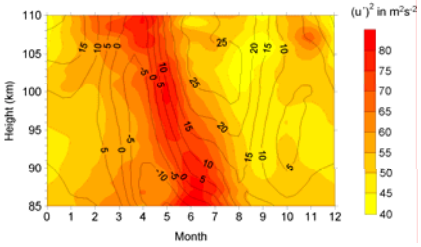
A single application can be generating/collecting many types of data

Big Public Data (online, weather, finance, etc)



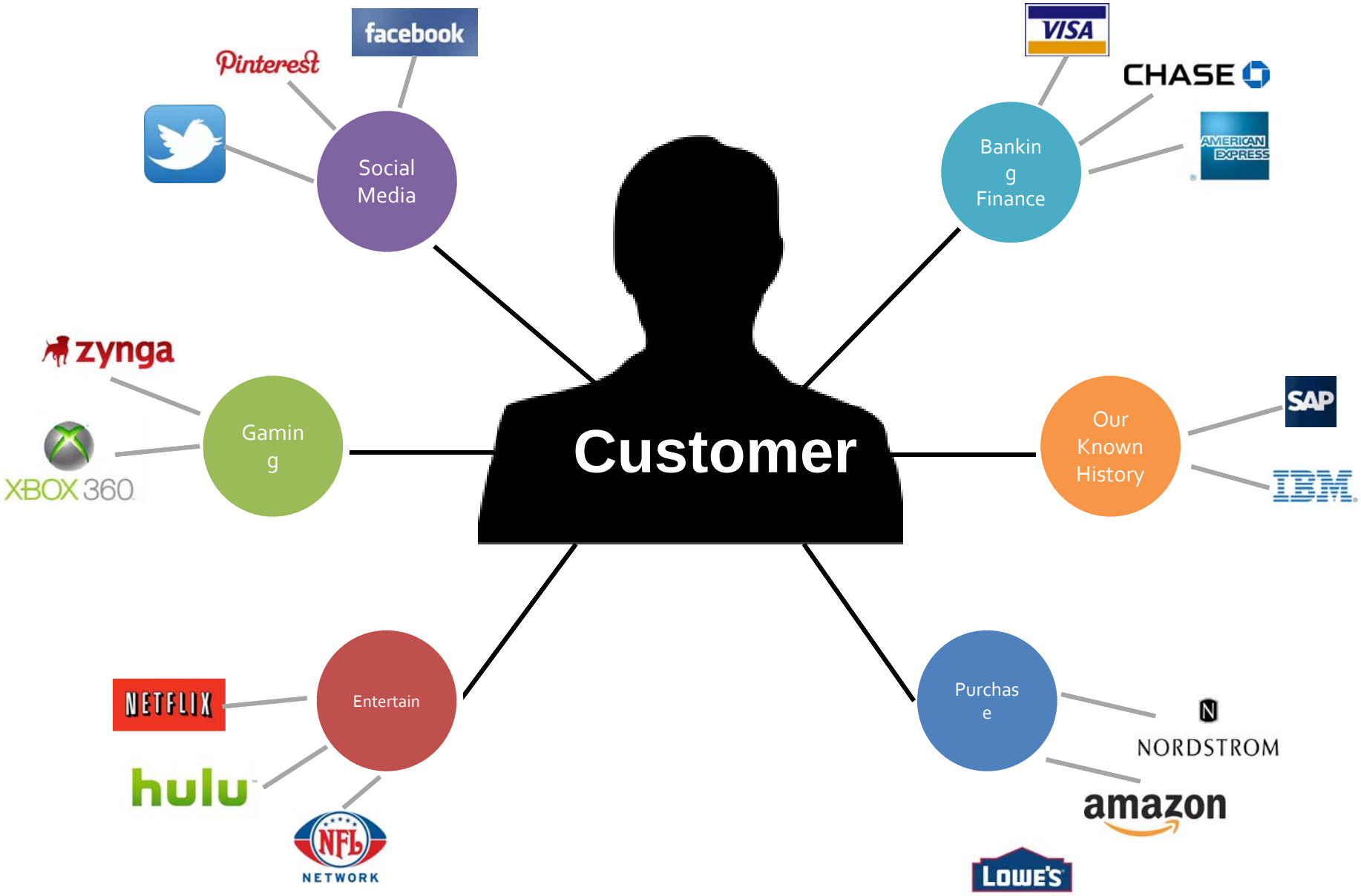






To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



Velocity (Speed)

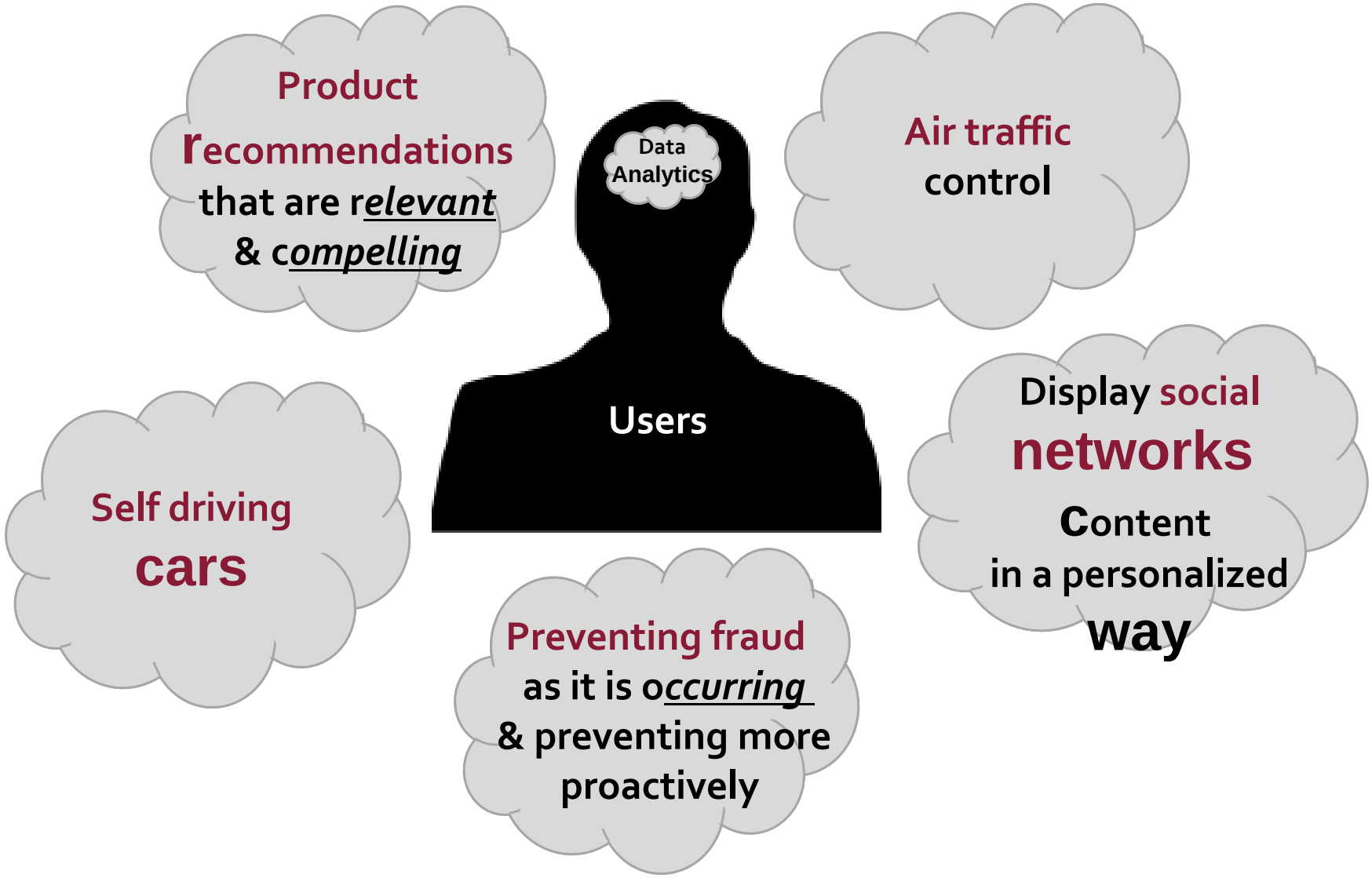
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

Velocity (Speed)

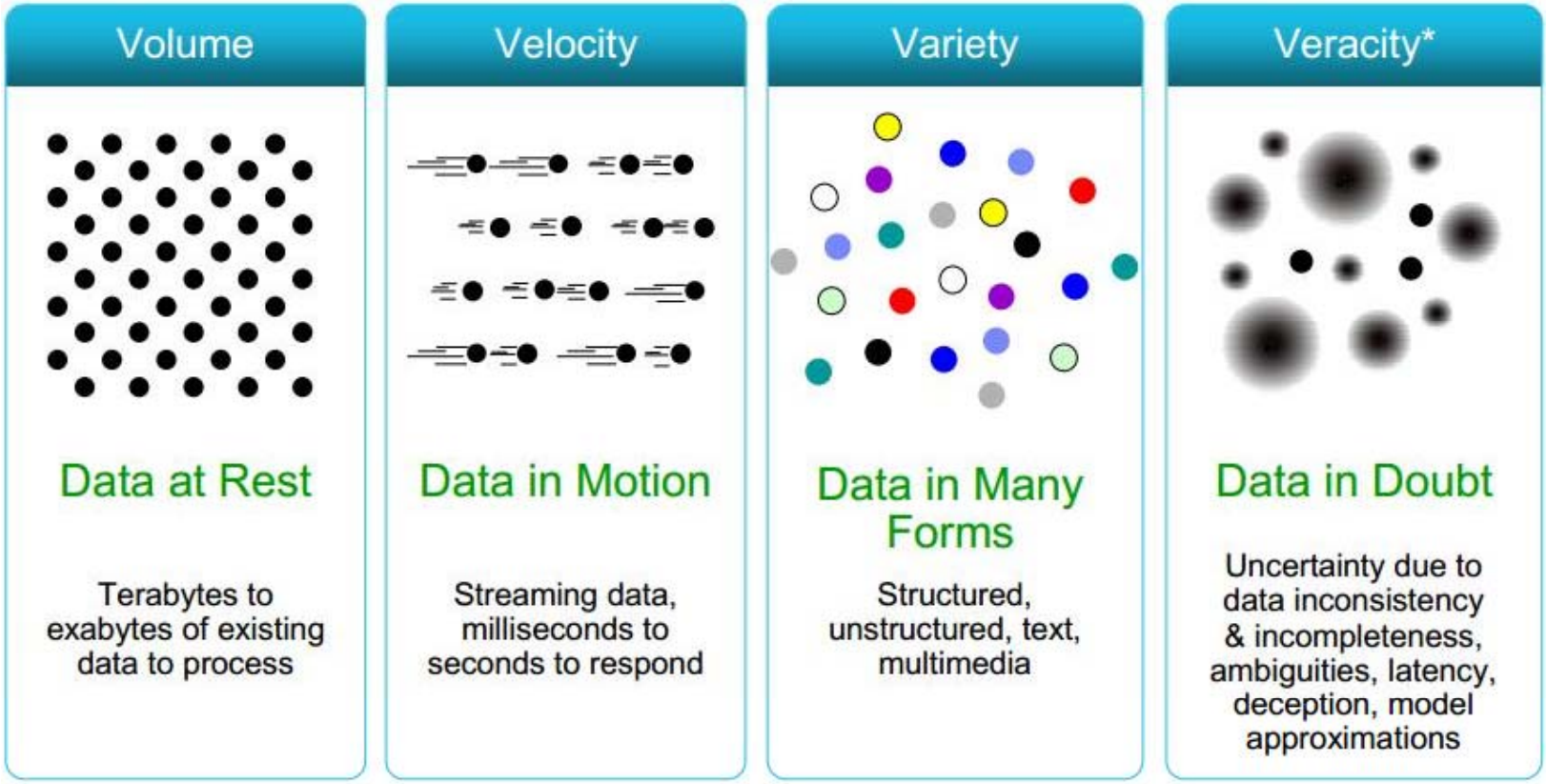
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

- **Examples**
  - » **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - » **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Velocity (Speed)

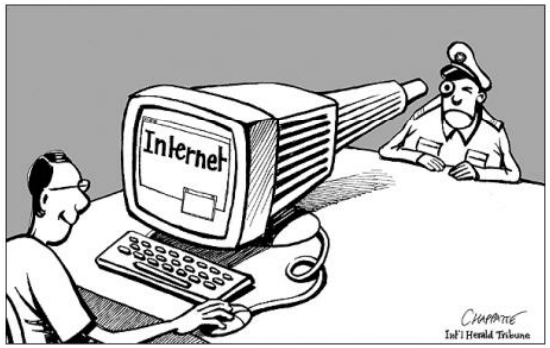


Some Make it 4V's



… and Privacy

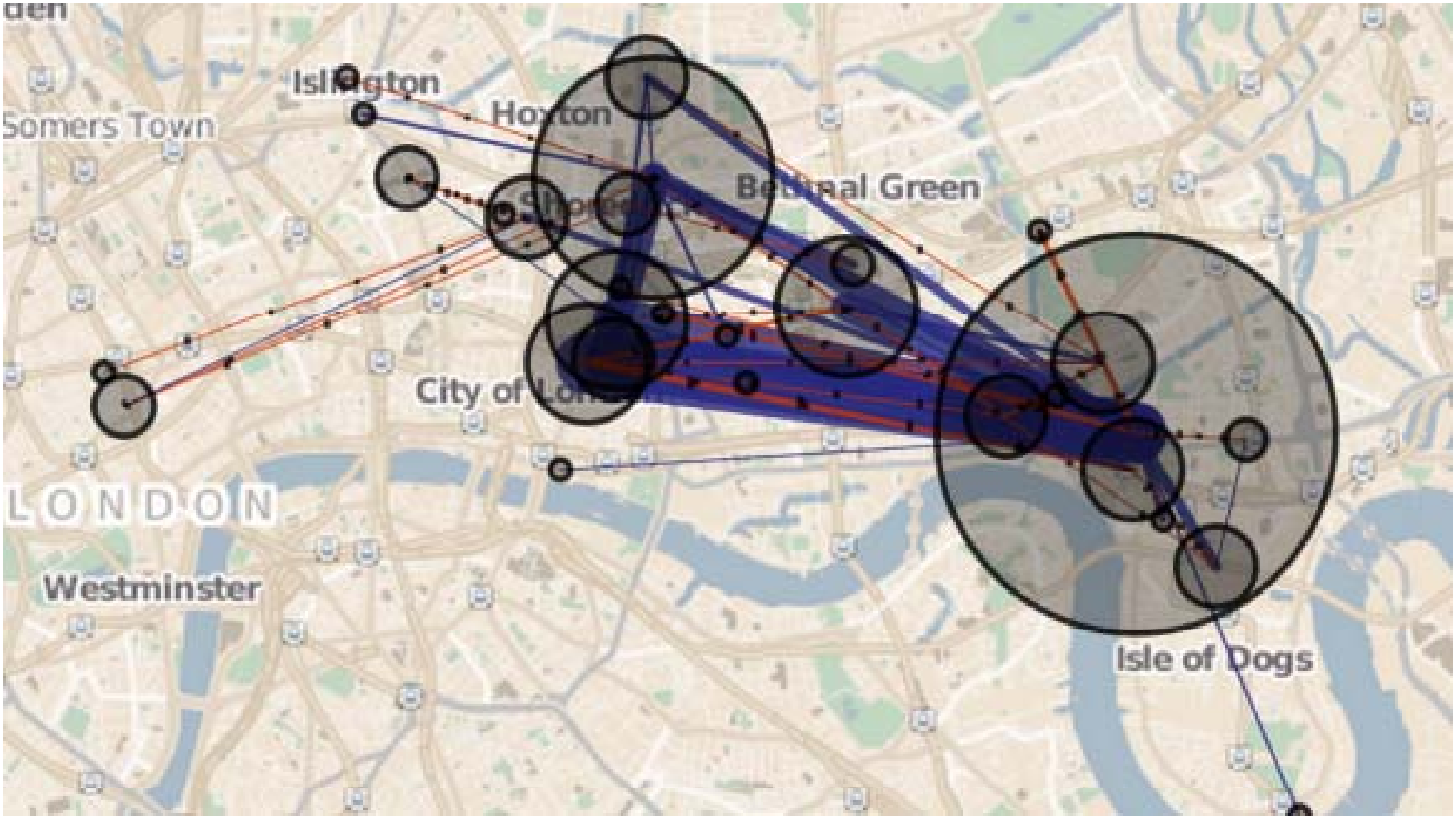
PRIVACY
PLEASE


Internet

DATA COLLECTION LISTS
CHEAP!
SOC. SEC. NUMBERS
DRIVER'S LICENSE NUMBERS
HOME OWNERSHIP
INCOME
MARITAL STATUS
INTERNET HABITS
AUTO
FAMILY STATUS
MEDICAL HISTORY
BANK RECORDS
INSURANCE RECORDS
FENCING STOLEN GOODS
KIRK

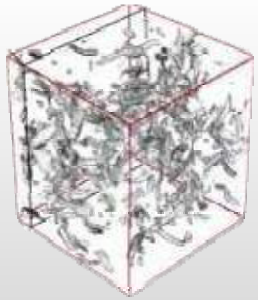
DOGBERT CONSULTS
CUSTOMER DATA IS AN ASSET THAT YOU CAN SELL.
IT'S TOTALLY ETHICAL BECAUSE OUR CUSTOMERS WOULD DO THE SAME THING TO US IF THEY COULD.
SOUNDS FAIR.
IN PHASE ONE, WE'LL DEHUMANIZE THE ENEMY BY CALLING THEM "DATA."
Dilbert.com DilbertCartoonist@gmail.com
10-13-10 © 2010 Scott Adams, Inc./Dist. by UFS, Inc.

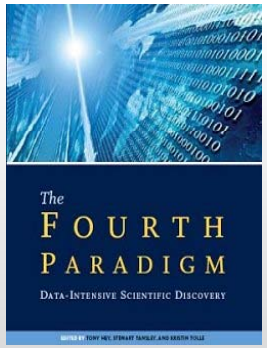
Goodbye Anonymity



The Data Science:
The 4th Paradigm for Scientific Discovery

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$





Experimental	Theoretical	Computational	The Fourth Paradigm
Thousand years ago	Last few hundred years	Last few decades	Today and the Future
Description of natural phenomena	*Newton's laws, Maxwell's equations…*	*Simulation of complex phenomena*	*Unify theory, experiment and simulation with large multidisciplinary Data* *Using data exploration and data mining (from instruments, sensors, humans…)* *Distributed Communities*

Crédits: Dennis Gannon

Big Data Science:
The art of understanding huge volumes of data

Data Science is not just data analysis.

Four main topics:

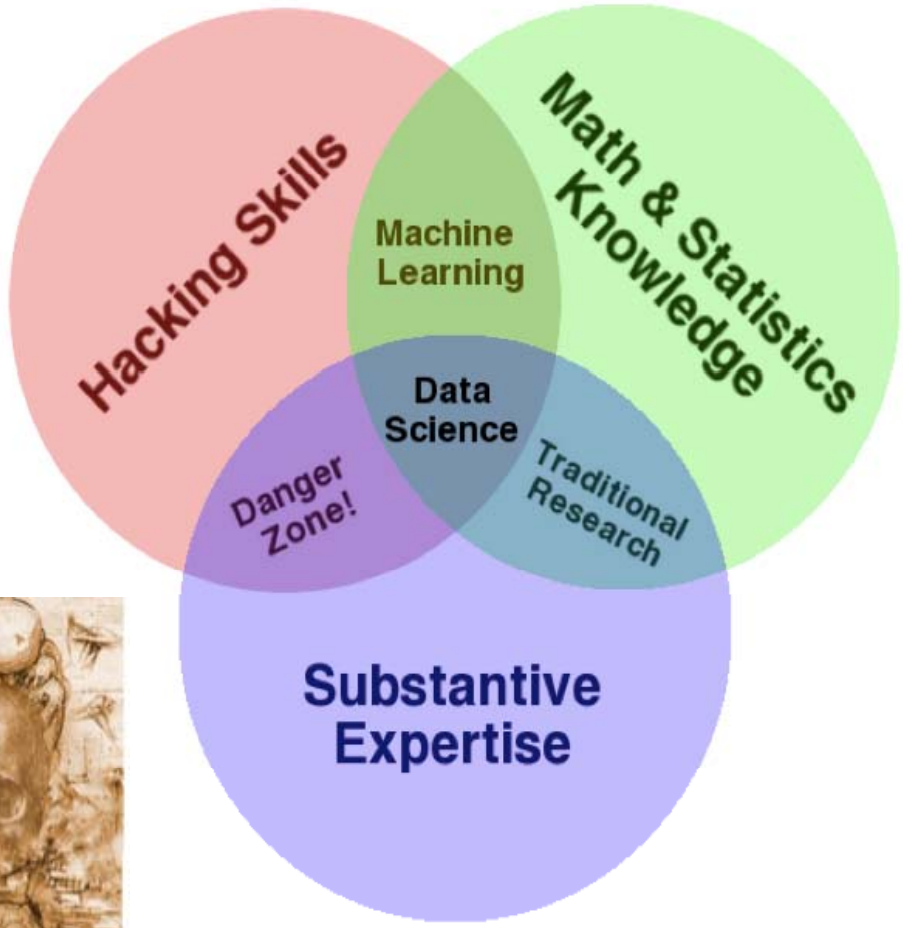
- **Data architecture:** how the data would need to be routed and organized to support the analysis, visualization and presentation of the data
- **Data acquisition:** how the data are collected, and, importantly, how the data are represented prior to analysis and presentation
- **Data analysis:** involves many technical, mathematical, and statistical aspects; still, the results have to be effectively communicated to the data user.
- **Data archiving:** preservation of collected data in a form that makes it highly reusable (data curation)

Data Scientist skills

Engineer collect & scrub disparate data sources manage a large computing cluster

Mathematician machine learning statistics

Artist visualize data beautifully, tell a convincing story

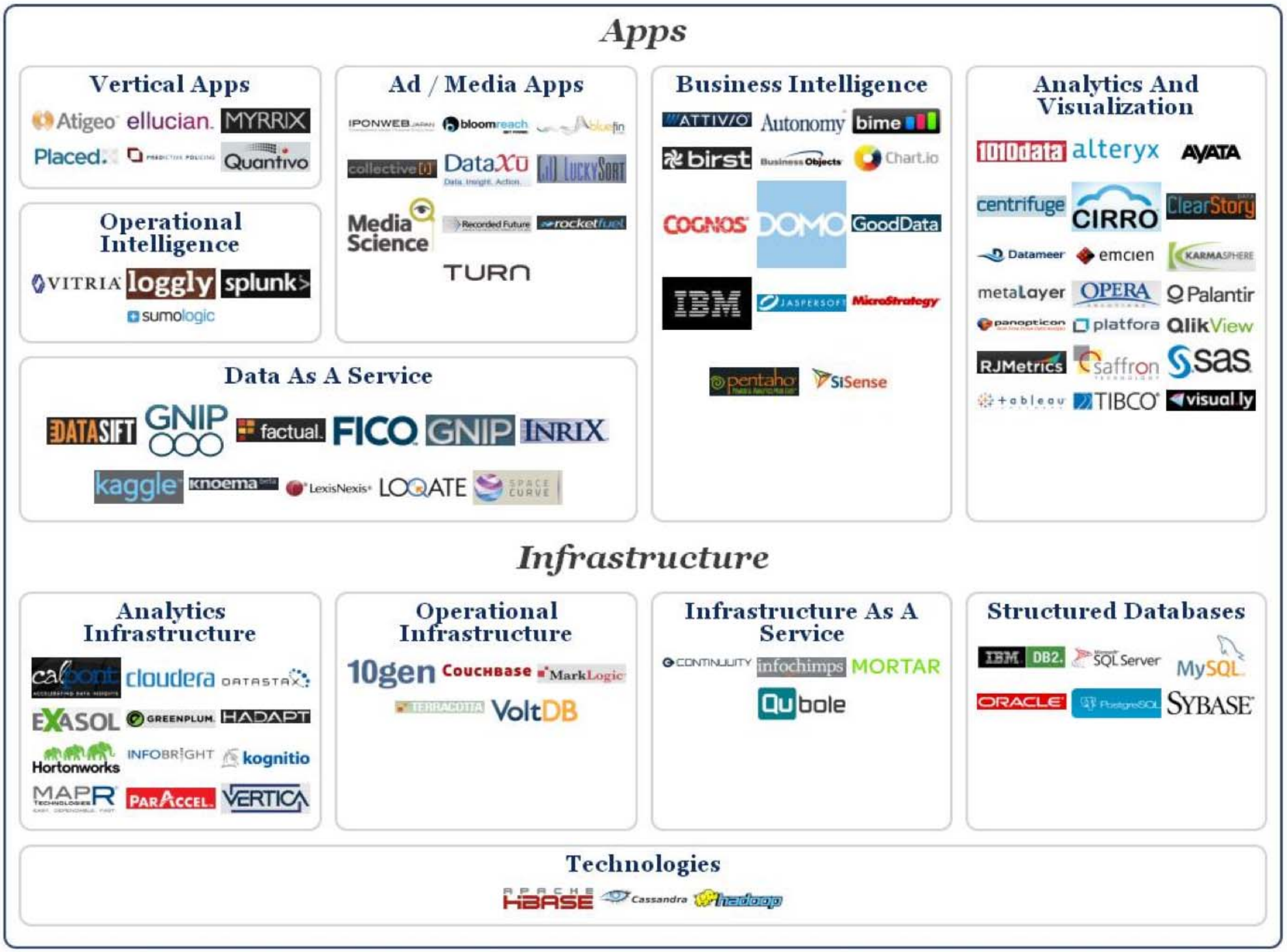


Data Scientist

"I keep saying the sexy job in the next ten years will be statisticians. The ability to take data - to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it."

Hal Varian, Google's chief economist

The Big Data Landscape



BIG DATA is GOING TO IMPACT US SOON !
CALL BRUCE WILLIS !!
DATA
www.nuxeo.com
Soulcié

Thank you!

Shadi Ibrahim

shadi.ibrahim@inria.fr