DALHOUSIE UNIVERSITY

Inspiring Minds

BIG-DATA & DATA PIPELINES

CSCI 5408:

Data Management, Warehousing, and Analytics

Prepared By: Suhaib Qaiser (suhaibqaiser@dal.ca)

Recap from last lecture…

Q1. How can we scale horizontally and vertically in Heroku?

Q2. What are Salesforce limitations as a platform?

Q3. How is Assignment 2 coming along?

Q4. What is the difference between Latency and Response delay?

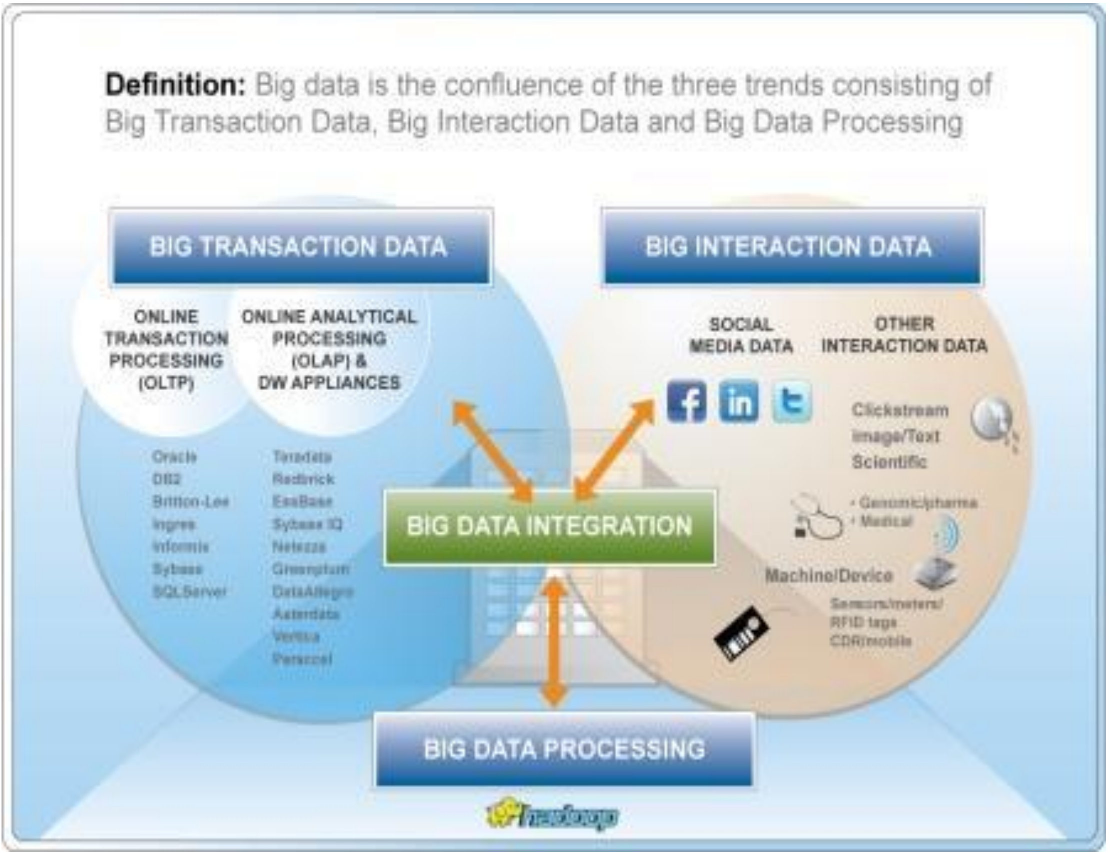
Q5. Is latency on Web Servers bigger than Database servers? Compare

BIG DATA

BIG DATA

DEFINITION

Big Data refers to data that because of its size, speed or format, that is, its volume, velocity or variety, cannot be easily stored, manipulated or analyzed with traditional methods like spreadsheets, relational databases or common statistical software

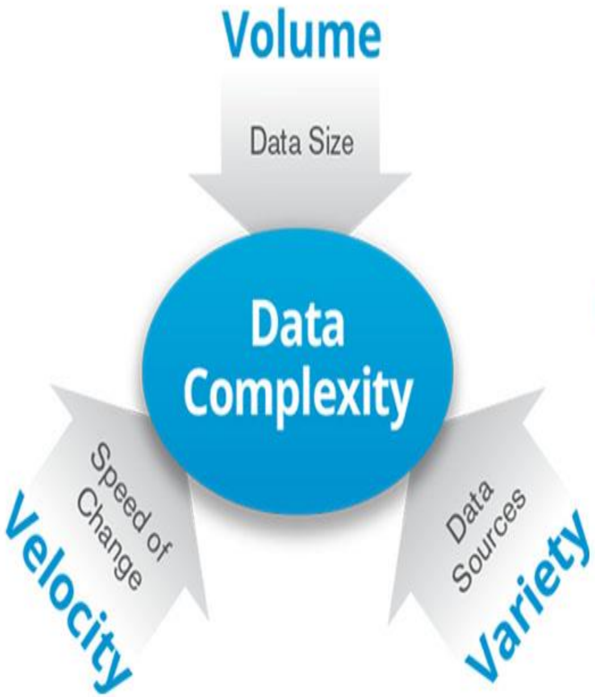


Definition: Big data is the confluence of the three trends consisting of Big Transaction Data, Big Interaction Data and Big Data Processing

BIGDATA

What is BIGDATA

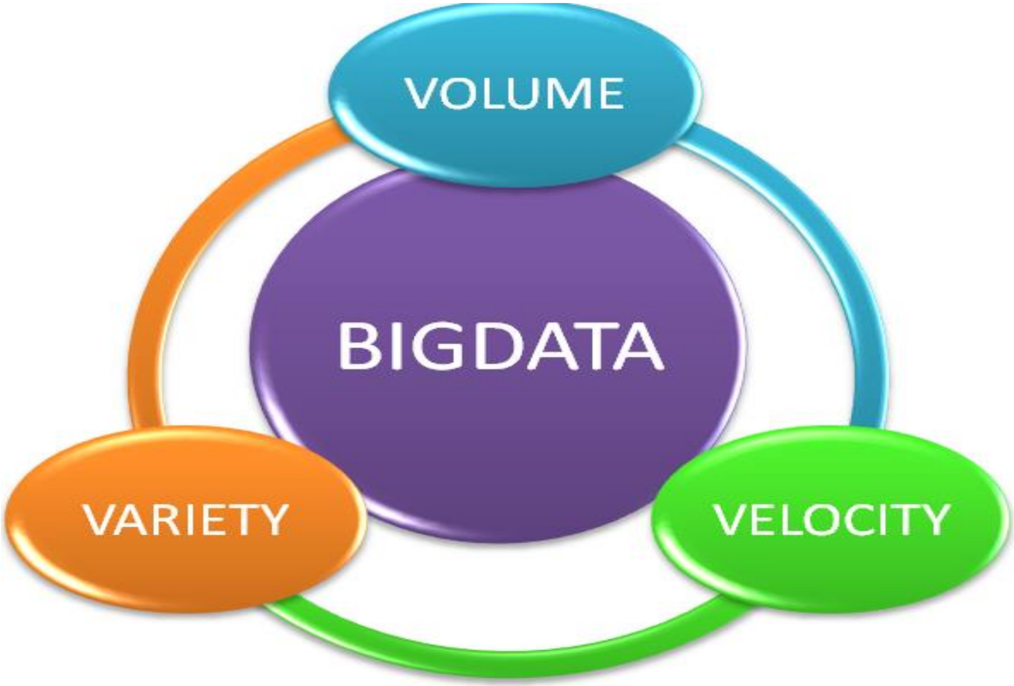
One way of describing big data is by looking at the three V's of volume, velocity, and variety. These come from an article written by Doug Laney in 2001, and they're taken as the most common characteristics of big data.

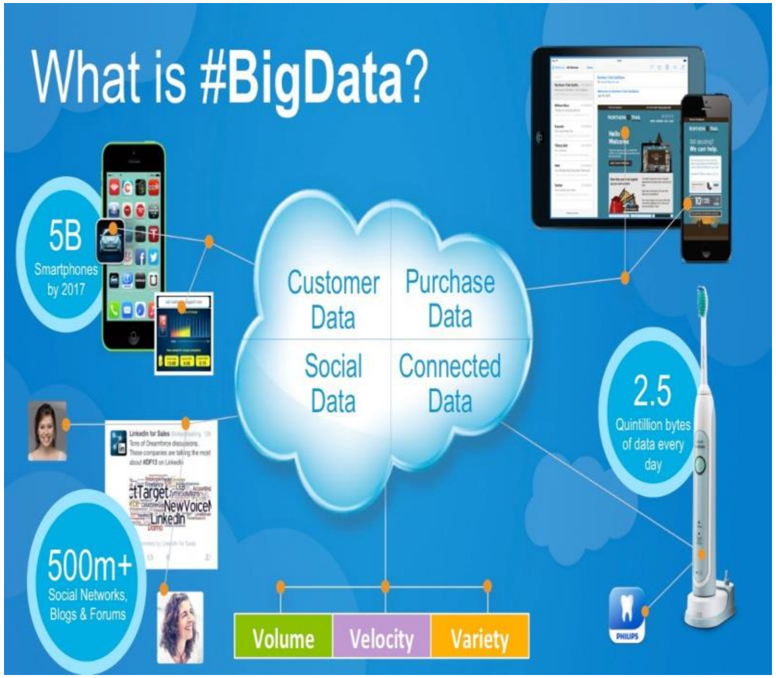


BIGDATA

What is BIGDATA

one way of describing big data is by looking at the three V's of volume, velocity, and variety. These come from an article written by Doug Laney in 2001, and they're taken as the most common characteristics of big data.



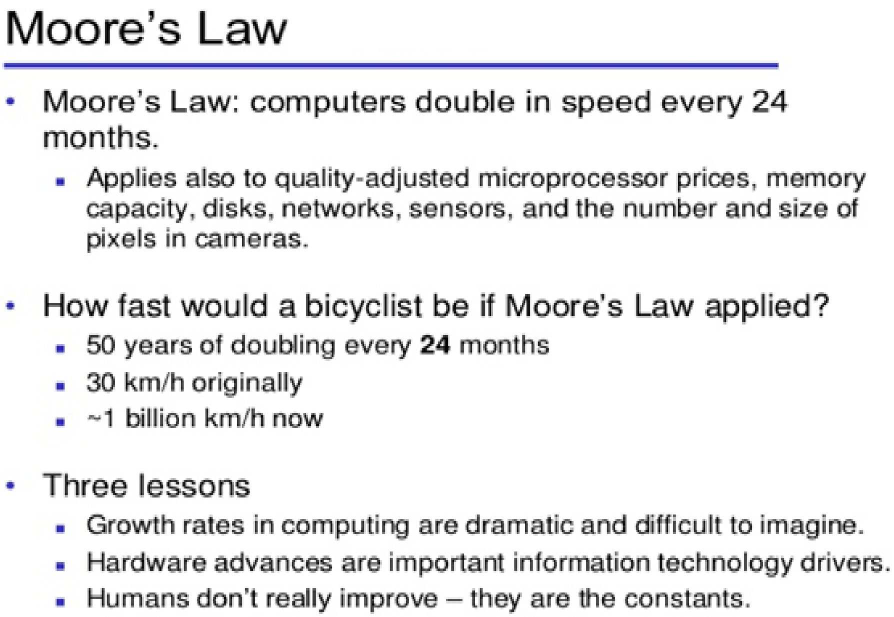


BIGDATA

What is BIGDATA

VOLUME

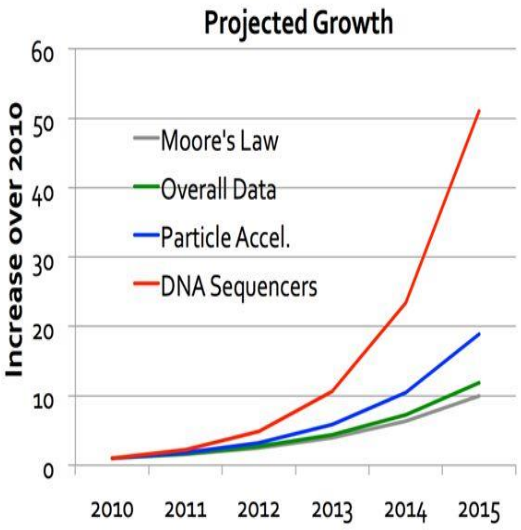
- big data is data that's just too big to work on your computer



Moore's Law

- Moore's Law: computers double in speed every 24 months.
  - Applies also to quality-adjusted microprocessor prices, memory capacity, disks, networks, sensors, and the number and size of pixels in cameras.
- How fast would a bicyclist be if Moore's Law applied?
  - 50 years of doubling every **24** months
  - 30 km/h originally
  - ~1 billion km/h now
- Three lessons
  - Growth rates in computing are dramatic and difficult to imagine.
  - Hardware advances are important information technology drivers.
  - Humans don't really improve – they are the constants.

Trends: Big Data



Data grows faster than Moore's Law!

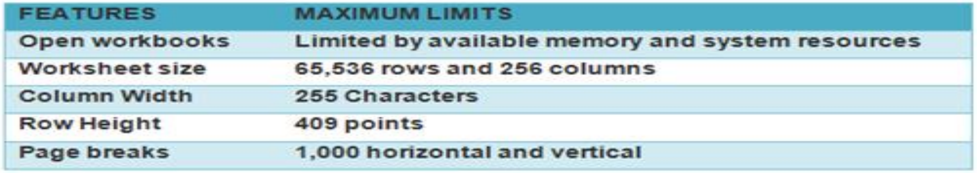
[Kathy Yelick LBNL, VLDBJ 2012, Dhruba Borthakur]

BIGDATA

What is BIGDATA

VOLUME

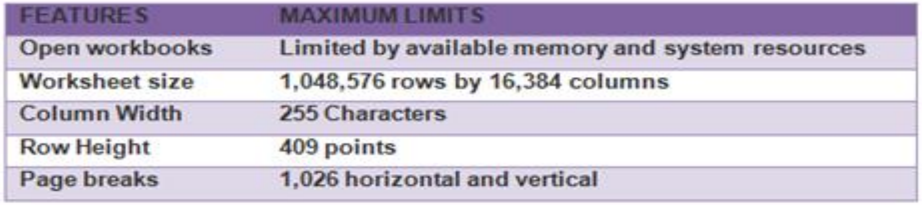
One way of describing big data is by looking at the three V's of volume, velocity, and variety. These come from an article written by Doug Laney in 2001, and they're taken as the most common characteristics of big data.



FEATURES	MAXIMUM LIMITS
Open workbooks	Limited by available memory and system resources
Worksheet size	65,536 rows and 256 columns
Column Width	255 Characters
Row Height	409 points
Page breaks	1,000 horizontal and vertical

Excel's Limit

Version 2003



FEATURES	MAXIMUM LIMITS
Open workbooks	Limited by available memory and system resources
Worksheet size	1,048,576 rows by 16,384 columns
Column Width	255 Characters
Row Height	409 points
Page breaks	1,026 horizontal and vertical

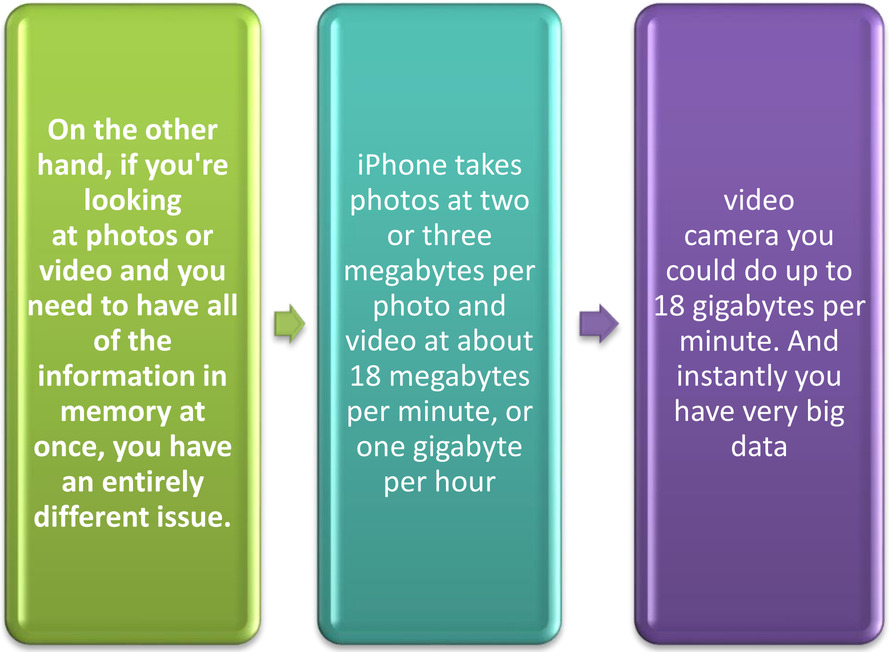
Excel's Limit

Version 2007

BIGDATA

What is BIGDATA

VOLUME



On the other hand, if you're looking at photos or video and you need to have all of the information in memory at once, you have an entirely different issue.

iPhone takes photos at two or three megabytes per photo and video at about 18 megabytes per minute, or one gigabyte per hour

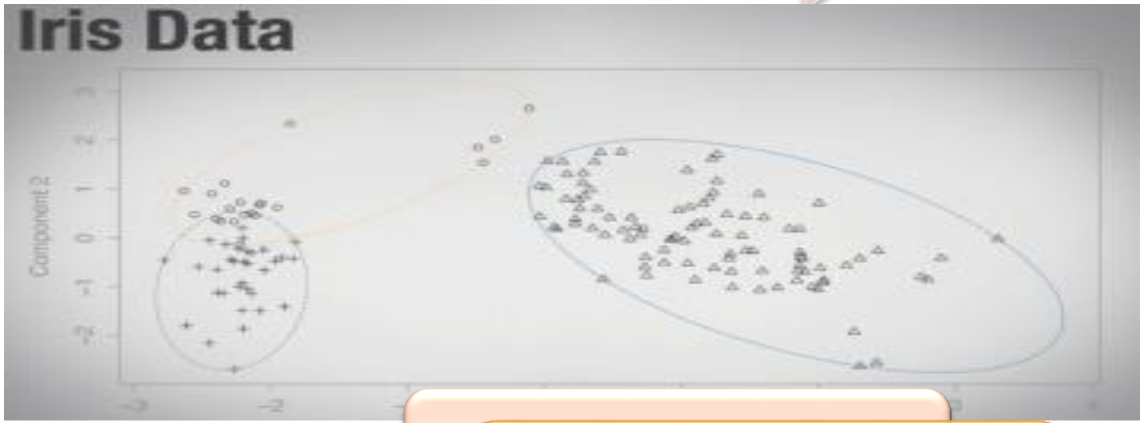
video camera you could do up to 18 gigabytes per minute. And instantly you have very big data

BIGDATA

What is BIGDATA

VELOCITY

- Velocity, this is when data is coming in very fast



the most familiar data set for teaching the statistical procedure, cluster analysis, is the Iris data collected by Edgar Anderson and analyzed by Ronald Fisher, both of whom published their papers in 1936

	6000 TWEETS PER SECOND
	5000 MILLION TWEETS PER DAY
	200 BILLION TWEETS PER YEAR

Data from a social media platform, like Twitter, you may have to deal with the so-called "fire horse“ That works out to 500,000,000 tweets per day and about 200,000,000,000 tweets per year.

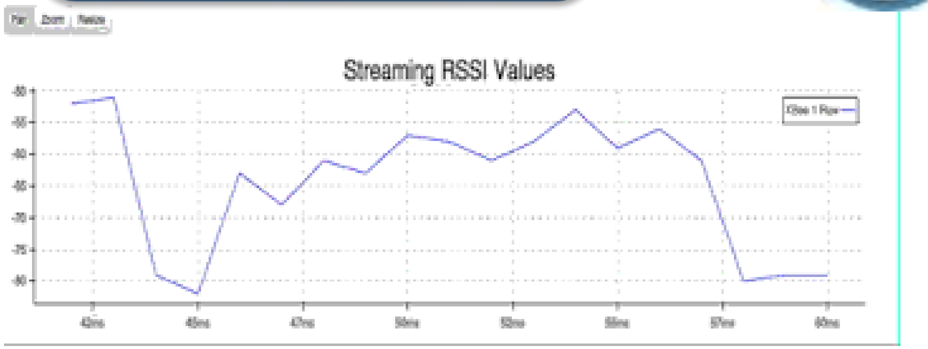
BIGDATA

What is BIGDATA

VELOCITY

- Velocity, this is when data is coming in very fast

The kind of constant influx of data, better known as streaming data, presents special challenges for analysis, because the data set, itself, is a moving target.



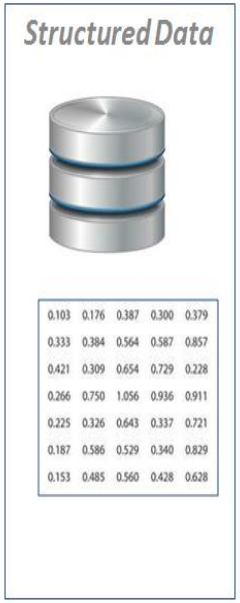
If you're accustomed to working with static data sets, in a program like SPSS or R, the demands and complexities of streaming data can be very daunting, to say the least.

BIGDATA

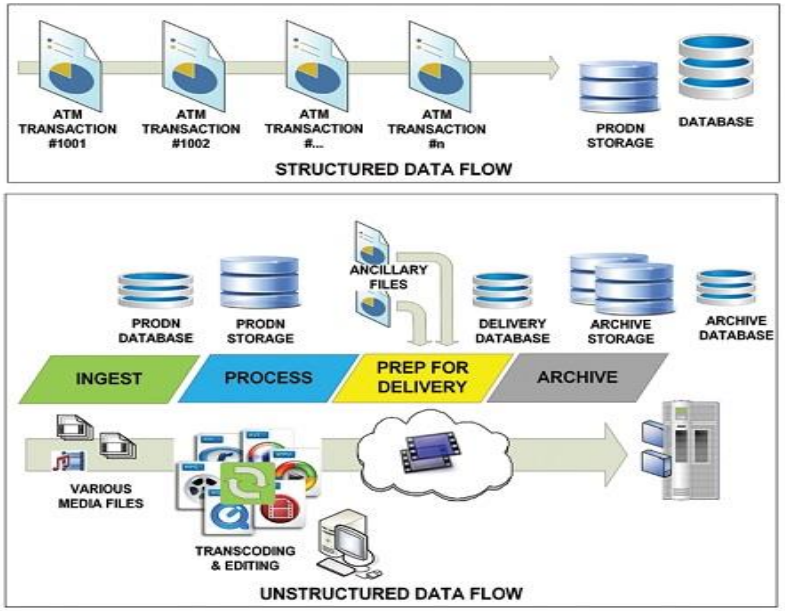
What is BIGDATA

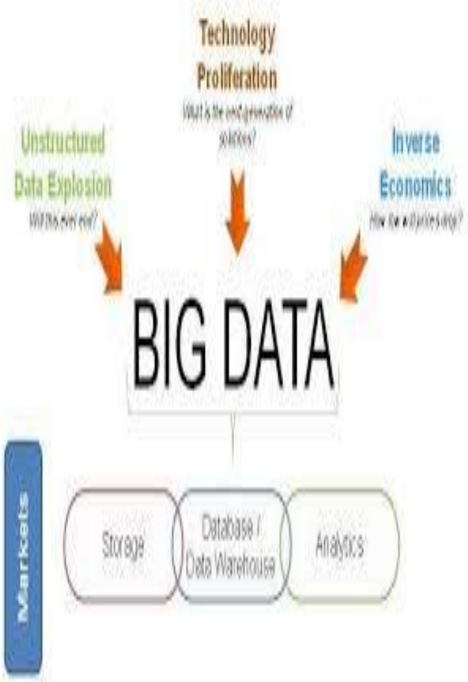
VARIETY

- The Structured Data that are rows and columns of a nicely formatted data set in a spread sheet, and many data sheets in many different formats. Data in unstructured text,like books and blog posts and comments on news articles and tweets. One researcher has estimated that 80 percent of enterprise data may be unstructured.



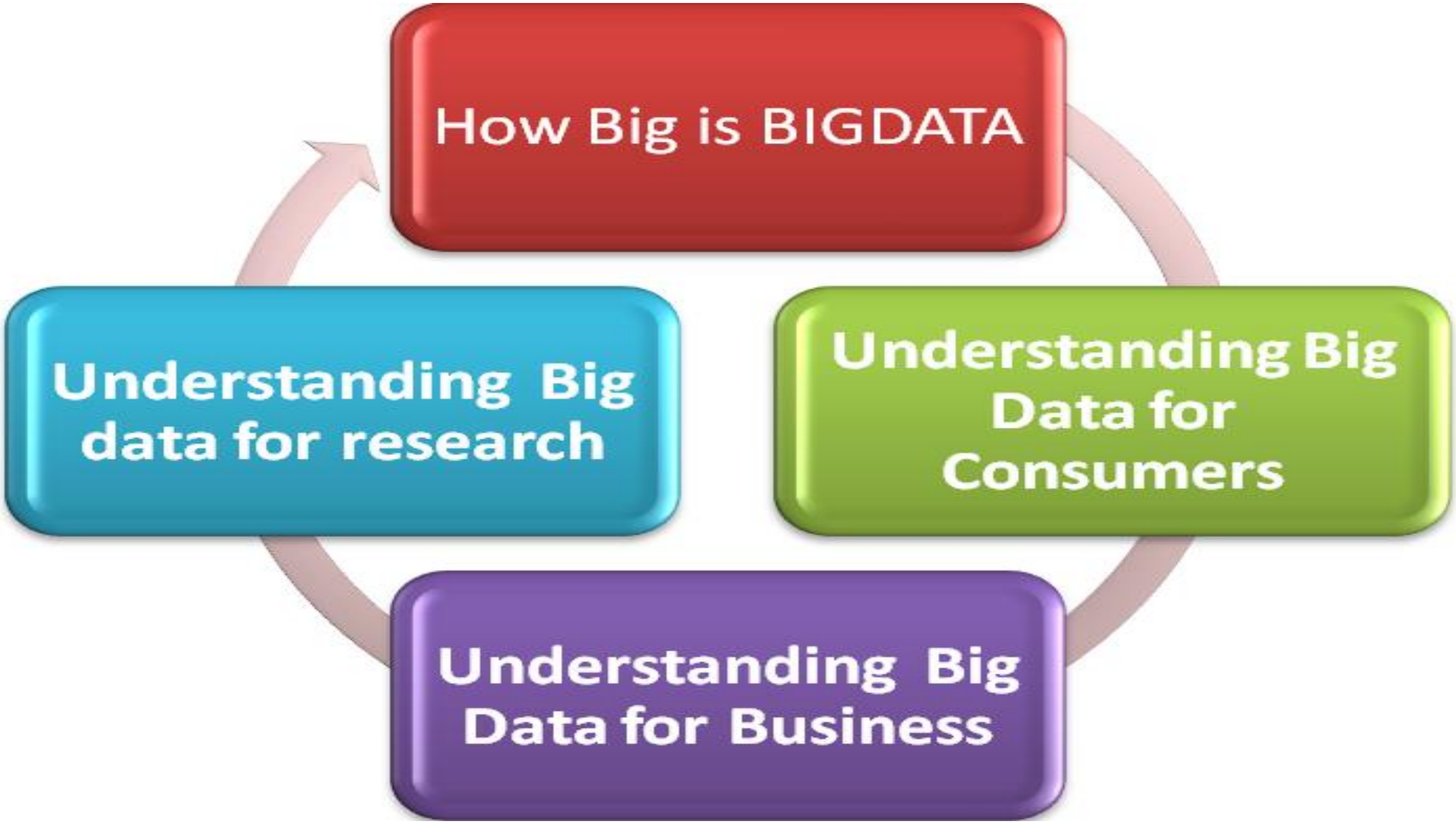
0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628




Technology Proliferation
What is the next generation of solutions?
Unstructured Data Explosion
Will this ever end?
Inverse Economics
How low will prices drop?
BIG DATA
Markets
Storage
Database / Data Warehouse
Analytics

BIGDATA

How Big is the BIGDATA



How Big is BIGDATA

Understanding Big Data for Consumers

Understanding Big Data for Business

Understanding Big data for research

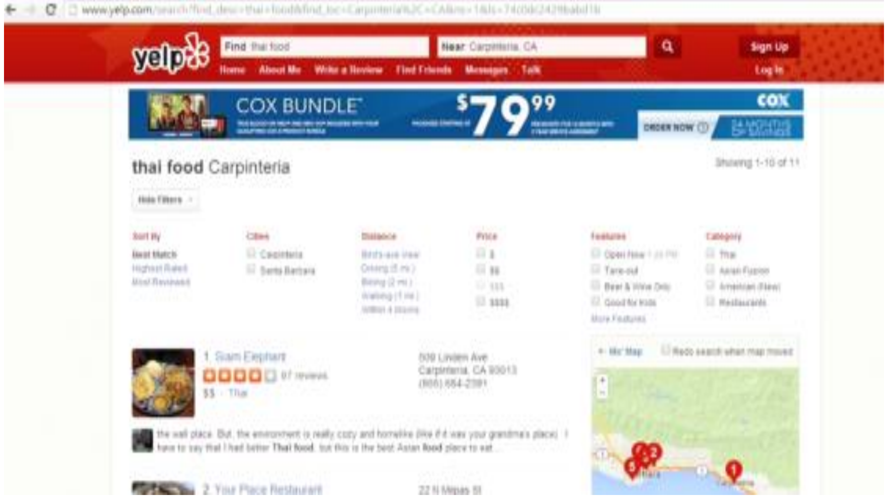
BIGDATA How big is Big Data

Understanding Big Data For Consumers

For Consumers big data plays an enormous role in providing valuable services big data can provide important conveniences and functionality for consumers

some common applications of big data for consumers that you may be using already without being aware of the sophistication of the big data analysis

Apple iPhone or iPad is what Siri can do. So for instance, aside from saying what's the weather like, and Siri actually knows what it is you mean, and where you are, and what time you're talking about, it can do things like look for restaurants of a particular kind of food and see if they have reservations available www.yelp.com/events



thai food Carpinteria

Showing 1-10 of 11

Hide Filters

Sort By	Cities	Distance	Price	Features	Category
Best Match	□ Carpinteria	Bird's-eye View	□ $	□ Open Now	□ Thai
Highest Rated	□ Santa Barbara	Driving (5 mi.)	□ $$	□ Take-out	□ Asian Fusion
Most Reviewed		Biking (2 mi.)	□ $$$	□ Beer & Wine Only	□ American (New)
		Walking (1 mi.)	□ $$$$	□ Good for Kids	□ Restaurants
		Within 4 blocks		More Features	

1. Siam Elephant — 97 reviews — $$ · Thai — 509 Linden Ave, Carpinteria, CA 93013, (805) 684-2391

the wall place. But, the environment is really cozy and homelike (like if it was your grandma's place). I have to say that I had better Thai food, but this is the best Asian food place to eat...

2. Your Place Restaurant — 22 N Milpas St

Mo' Map □ Redo search when map moved

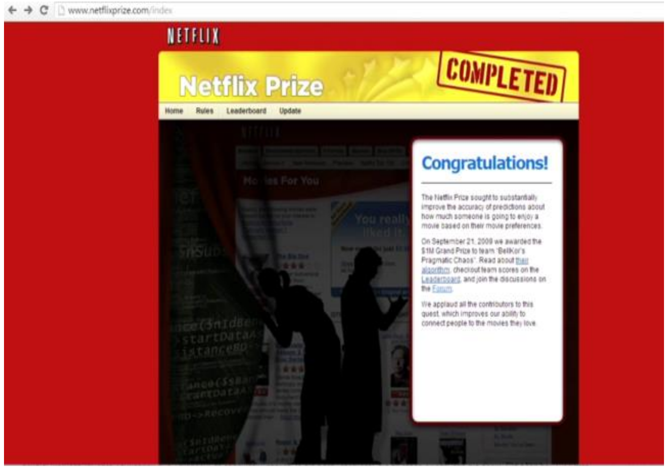
BIGDATA How big is Big Data

Understanding Big Data For Consumers

For Consumers big data plays an enormous role in providing valuable services big data can provide important conveniences and functionality for consumers

some common applications of big data for consumers that you may be using already without being aware of the sophistication of the big data analysis

Netflix makes specific suggestions for other movies you might like
www.netflix.com/browse



Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

BIGDATA

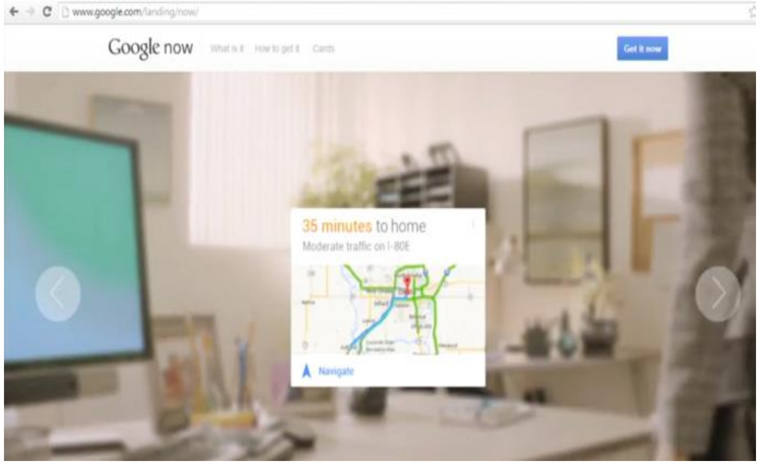
Understanding Big Data For Consumers

For Consumers big data plays an enormous role in providing valuable services big data can provide important conveniences and functionality for consumers

some common applications of big data for consumers that you may be using already without being aware of the sophistication of the big data analysis

Google Now makes recommendations before you ask for them:

https://play.google.com/store/apps/details?id=com.google.android.launcher&hl=en



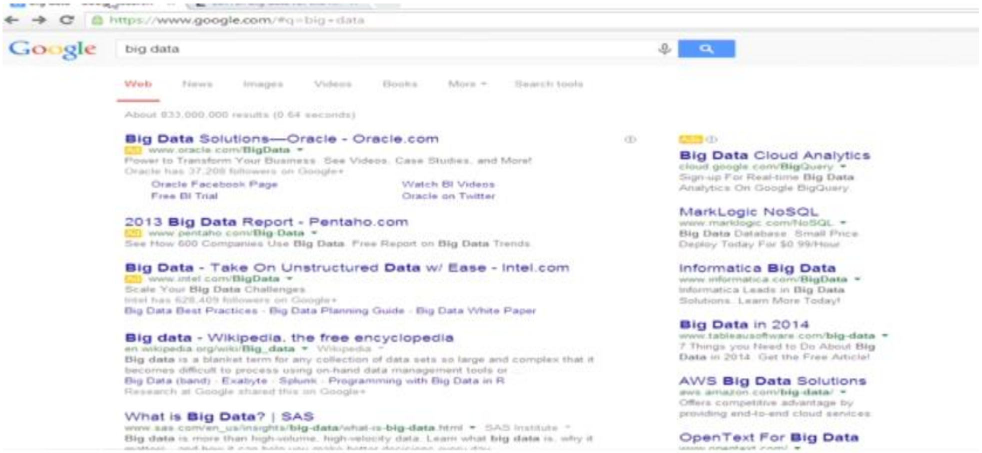
BIGDATA How big is Big Data

Understanding Big Data For Business

For the business world, big data is revolutionizing the way people do commerce

most people have encountered big data in commerce, and that's in the results for Google ad searches

Google Searches
www.google.com



https://www.google.com/#q=big+data

big data

About 833,000,000 results (0.64 seconds)

Big Data Solutions—Oracle - Oracle.com
Ad www.oracle.com/BigData
Power to Transform Your Business. See Videos, Case Studies, and More!
Oracle has 37,208 followers on Google+
Oracle Facebook Page · Watch BI Videos
Free BI Trial · Oracle on Twitter

2013 Big Data Report - Pentaho.com
Ad www.pentaho.com/Big-Data
See How 600 Companies Use Big Data. Free Report on Big Data Trends

Big Data - Take On Unstructured Data w/ Ease - Intel.com
Ad www.intel.com/BigData
Scale Your Big Data Challenges
Intel has 628,409 followers on Google+
Big Data Best Practices - Big Data Planning Guide - Big Data White Paper

Big data - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Big_data · Wikipedia
Big data is a blanket term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or ...
Big Data (band) · Exabyte · Splunk · Programming with Big Data in R
Research at Google shared this on Google+

What is Big Data? | SAS
www.sas.com/en_us/insights/big-data/what-is-big-data.html · SAS Institute
Big data is more than high-volume, high-velocity data. Learn what big data is, why it

Ads

Big Data Cloud Analytics
cloud.google.com/BigQuery
Sign-up For Real-time Big Data Analytics On Google BigQuery

MarkLogic NoSQL
www.marklogic.com/NoSQL
Big Data Database. Small Price. Deploy Today For $0.99/Hour

Informatica Big Data
www.informatica.com/BigData
Informatica Leads in Big Data Solutions. Learn More Today!

Big Data in 2014
www.tableausoftware.com/big-data
7 Things you Need to Do About Big Data in 2014. Get the Free Article!

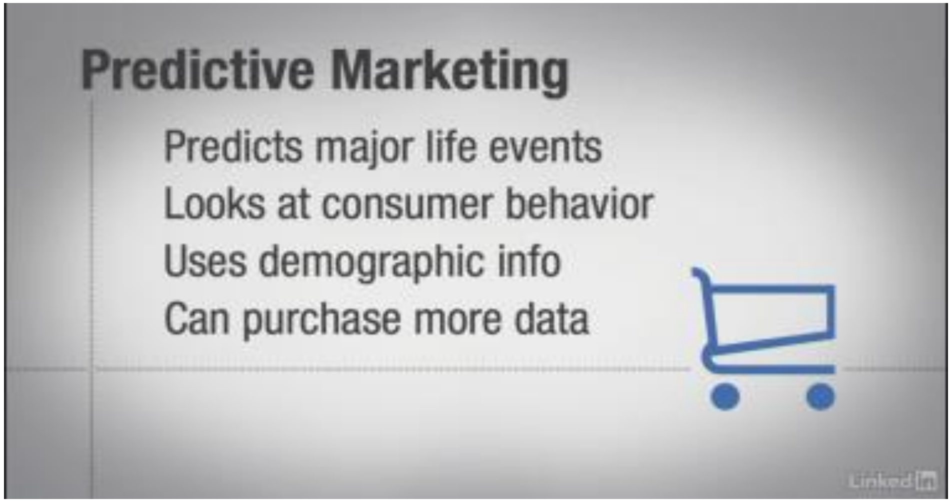
AWS Big Data Solutions
aws.amazon.com/big-data/
Offers competitive advantage by providing end-to-end cloud services.

OpenText For Big Data

BIGDATA

Understanding Big Data for Business

Another interesting place is what's called predictive marketing. This is when big data is used to help decide who the audience would be for something before they actually get there.



Predictive Marketing

- Predicts major life events
- Looks at consumer behavior
- Uses demographic info
- Can purchase more data

This is trying to predict, for example, major life events, like for instance graduating, or getting married, or getting a new job, or having a child, or any number of events that are often associated with a whole series of commercial transactions.

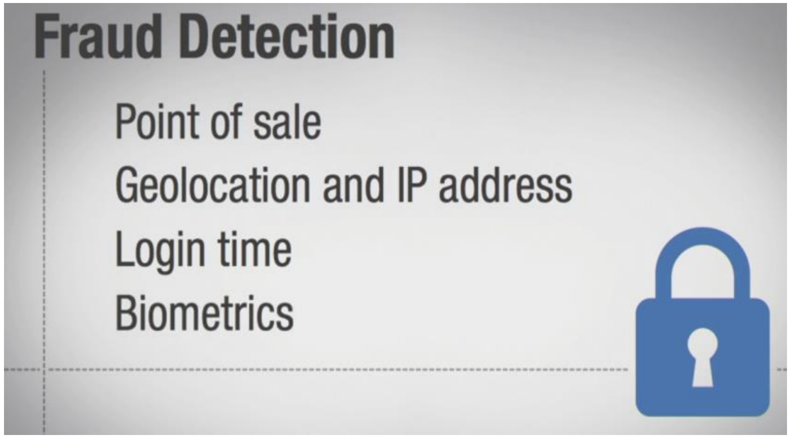
BIGDATA How big is Big Data

Understanding Big Data for Business

For the business world, big data is revolutionizing the way people do commerce

The use of big data in commerce is for fraud detection.

specifically means, how are you making the purchase? Are you online, and what website are you using? They can use geo location. Where are you physically located in the world? They can look at the IP address. What computer are you using to access the website. They can look at the log in time.



Fraud Detection

Point of sale

Geolocation and IP address

Login time

Biometrics

BIGDATA How big is Big Data

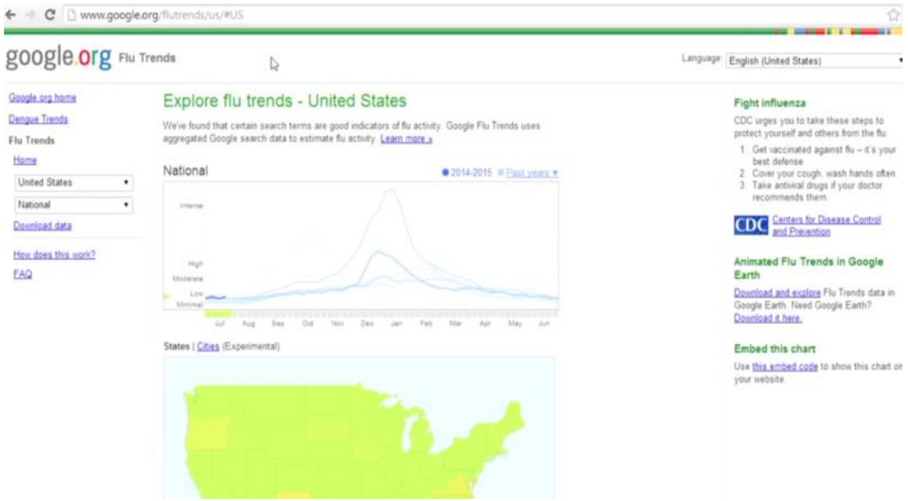
Understanding Big Data for Research

Big data has been revolutionizing aspects of scholarship and research

examples of where big data has influenced scientific progress..

The first one we want to look at is Google flu trends where they were able to find that search patterns for flu related words were actually able to identify outbreaks of the flu in the United States much faster than there search that the Center for Disease Control could do.

https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_



www.google.org/flutrends/us/#US

google.org Flu Trends

Language: English (United States)

Google.org home

Dengue Trends

Flu Trends

Home

United States

National

Download data

How does this work?

FAQ

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

States | Cities (Experimental)

Fight influenza

CDC urges you to take these steps to protect yourself and others from the flu.

1. Get vaccinated against flu – it's your best defense.
2. Cover your cough, wash hands often.
3. Take antiviral drugs if your doctor recommends them.

CDC Centers for Disease Control and Prevention

Animated Flu Trends in Google Earth

Download and explore Flu Trends data in Google Earth. Need Google Earth? Download it here.

Embed this chart

Use this embed code to show this chart on your website.

BIGDATA How big is Big Data

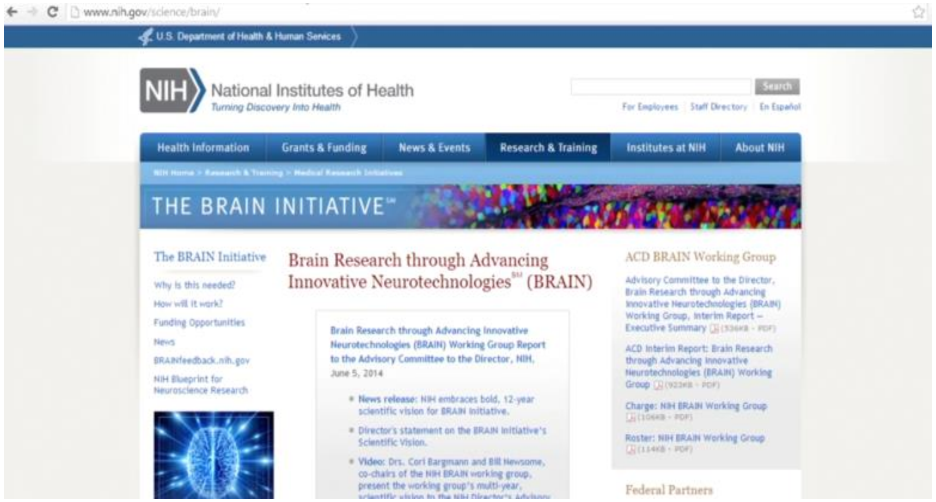
Understanding Big Data for Research

Big data has been revolutionizing aspects of scholarship and research

examples of where big data has influenced scientific progress..

The National Institutes of Health created the Brain Initiative as a way of taking enormous numbers of brain scans to create a full map of brain functioning.

https://www.nih.gov/research-training


www.nih.gov/science/brain/
U.S. Department of Health & Human Services
NIH
National Institutes of Health
Turning Discovery Into Health
Search
For Employees
Staff Directory
En Español
Health Information
Grants & Funding
News & Events
Research & Training
Institutes at NIH
About NIH
NIH Home > Research & Training > Medical Research Initiatives
THE BRAIN INITIATIVE℠

The BRAIN Initiative

- Why is this needed?
- How will it work?
- Funding Opportunities
- News
- BRAINfeedback.nih.gov
- NIH Blueprint for Neuroscience Research

Brain Research through Advancing Innovative Neurotechnologies℠ (BRAIN)

Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group Report to the Advisory Committee to the Director, NIH, June 5, 2014

- **News release:** NIH embraces bold, 12-year scientific vision for BRAIN Initiative.
- Director's statement on the BRAIN Initiative's Scientific Vision.
- **Video:** Drs. Cori Bargmann and Bill Newsome, co-chairs of the NIH BRAIN working group, present the working group's multi-year, scientific vision to the NIH Director's Advisory

ACD BRAIN Working Group

- Advisory Committee to the Director, Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group, Interim Report – Executive Summary (536KB - PDF)
- ACD Interim Report: Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group (923KB - PDF)
- Charge: NIH BRAIN Working Group (106KB - PDF)
- Roster: NIH BRAIN Working Group (114KB - PDF)

Federal Partners

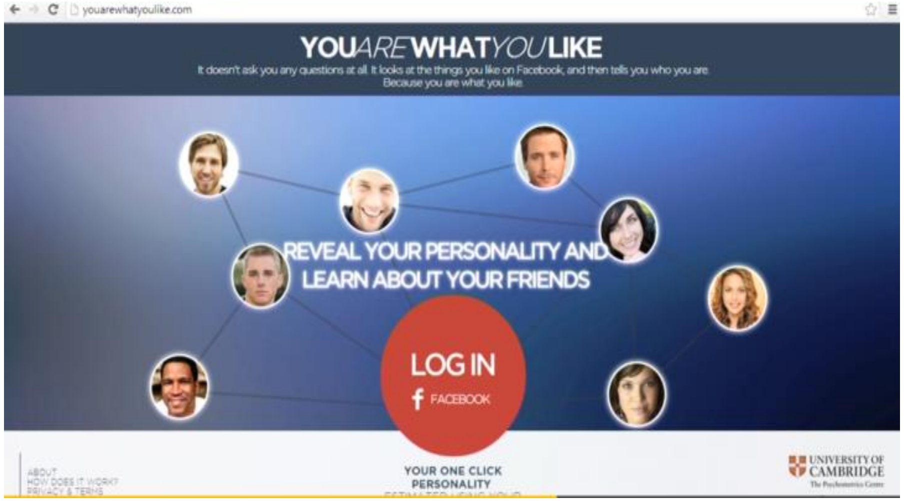
BIGDATA How big is Big Data

Understanding Big Data for Research

Big data has been revolutionizing aspects of scholarship and research

examples of where big data has influenced scientific progress..

A GROUP of researchers created an application on Facebook that used scientifically valid measure of personality https://applymagicsauce.com/demo.html



YOUAREWHATYOULIKE

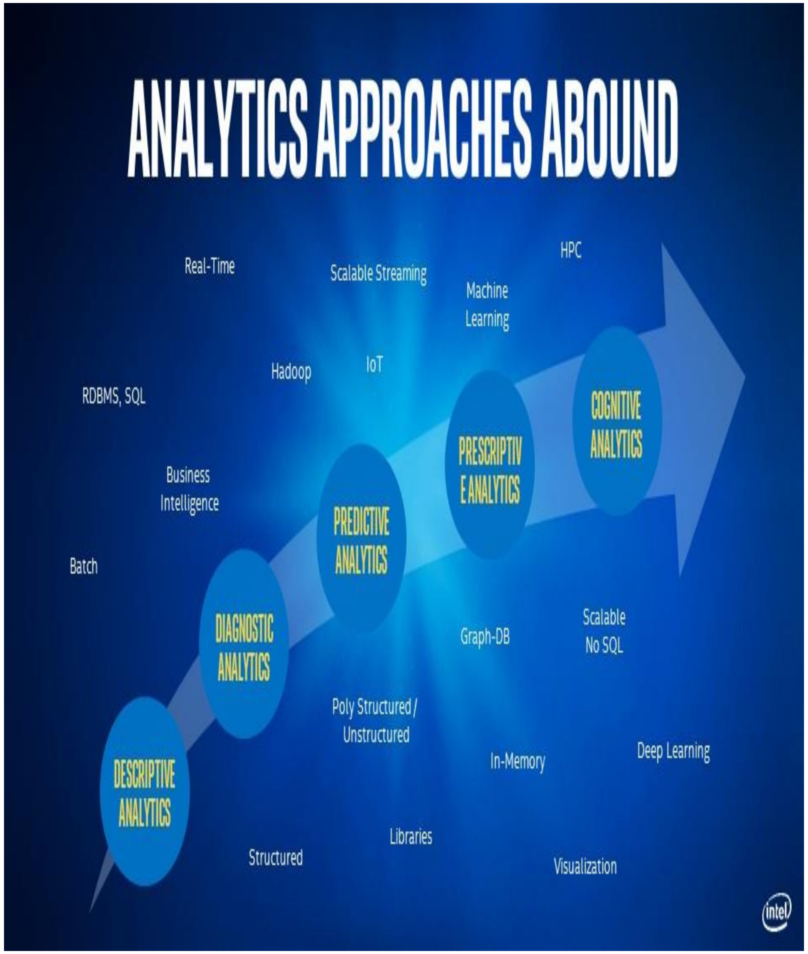
It doesn't ask you any questions at all. It looks at the things you like on Facebook, and then tells you who you are. Because you are what you like.

REVEAL YOUR PERSONALITY AND LEARN ABOUT YOUR FRIENDS

YOUR ONE CLICK PERSONALITY

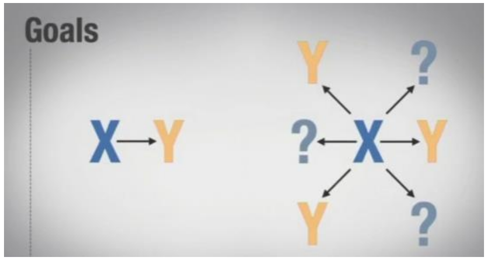
BIG DATA

BIG DATA AND DATA SCIENCES



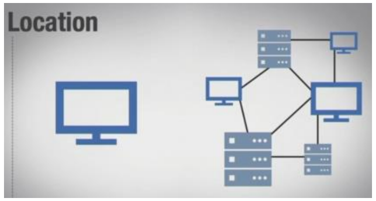
Big Data and Data Science

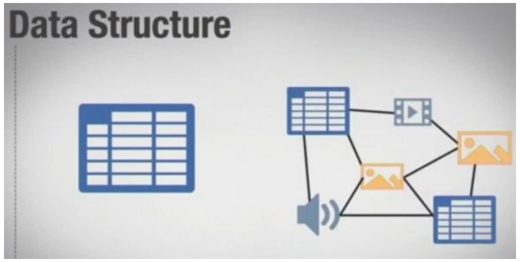
Ten Ways Big data is different from Small Data



The first is goals. Small data is usually gathered for a specific goal. Big data on the other hand may have a goal in mind when it's first started, but things can evolve or take unexpected directions.

The second is location. Small data is usually in one place, and often in a single computer file . Big data on the other hand can be in multiple files in multiple servers on computers in different geographic locations.





Third, the data structure and content.

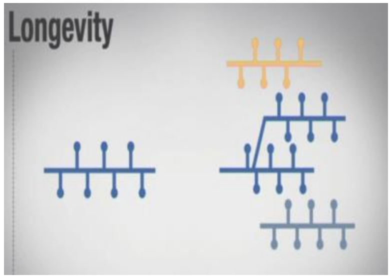
Small data is usually highly structured like an Excel spreadsheet, and it's got rows and columns of data. Big data on the other hand can be unstructured, it can have many formats in files involved across disciplines, and may link to other resources.

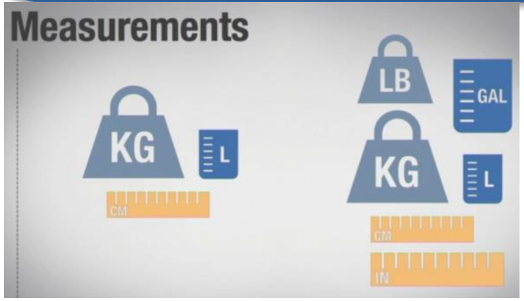
Big Data and Data Science

Ten Ways Big data is different from Small Data

Fourth, data preparation. Small data is usually prepared by the end user for their own purposes, but with big data the data is often prepared by one group of people, analyzed by a second group of people, and then used by a third group of people, and they may have different purposes, and they may have different disciplines. .

Fifth, longevity. Small data is usually kept for a specific amount of time after the project is over because there's a clear ending point. but with big data each data project, because it often comes at a great cost, gets continued into others, and things are going to stay there for a very long time. They may be added on to in terms of new data at the front, or contextual data of things or additional variables, or linking up with different files. So it has a much longer compared to a small data set.

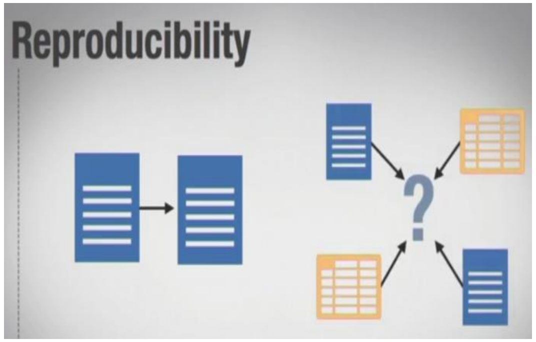




Small data is typically measured with a single protocol using set units and it's usually done at the same time. With big data on the other hand, because you can have people in very different places, in very different times, different organizations, and countries, you may be measuring things using different protocols, and you may have to do a fair amount of conversion to get things consistent

Big Data and Data Science

Ten Ways Big data is different from Small Data

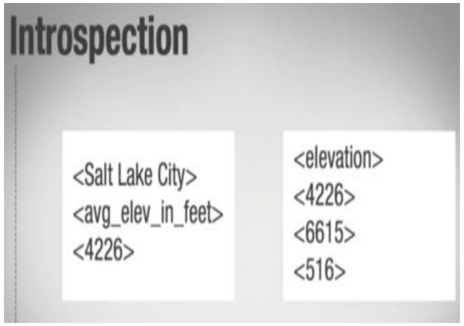


Number seven is reproducibility. Small data sets can usually be reproduced in their entirety if something goes wrong in the process. Big data sets on the other hand, because they come in so many forms and from different directions, it may not be possible to start over again if something's gone wrong. Usually the best you can hope to do is to at least identify which parts of the data project are problematic and keep those in mind as you work around them. .

Number eight is stakes. On small data, if things go wrong the costs are limited, it's not an enormous problem, but with big data, projects can cost hundreds of millions of dollars, and losing the data or corrupting the data can doom the project, possibly even the researcher's career or the organization's existence.

Big Data and Data Science

Ten Ways Big data is different from Small Data



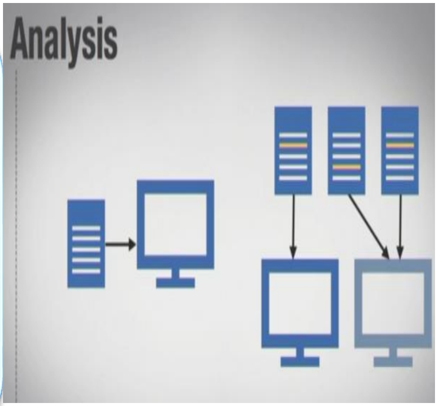
Introspection

<Salt Lake City>
<avg_elev_in_feet>
<4226>

<elevation>
<4226>
<6615>
<516>

The ninth is what's called introspection, and what this means is that the data describes itself in an important way.

The final characteristic is analysis. With small data it's usually possible to analyze all of the data at once in a single procedure from a single computer file. With big data however, because things are so enormous and they're spread across lots of different files and servers, you may have to go through extraction, reviewing, reduction, normalization, transformation, and other steps and deal with one part of the data at a time to make it more manageable, and then eventually aggregate your results.



QUIZ

Q1. Was dataset given to you in Assignment 2 an example of Big data?

Q2. Can Casandra hold Big Data?

Q3. What conventional tools fail to work with Big data applications?

Q4. What is the role of Hadoop in Big Data?

Q5. How can we use Casandra for Big Data?

ANY
QUESTIONS
?