

1/22 5:02:55 \*\*\*




# Big Data

"Go Big, Or Go Home!"





# Context (setting the stage)

The slides that follow are very general in nature – they present the 'big picture' concepts of Big Data. By nature, the content is relatively speaking, soft/squishy/fluffy.. 

It *\*is\** important to understand the context in which we will discuss data mining etc. in upcoming lectures, otherwise the material will seem dry/irrelevant.

# Big Data, Wordle (TM) summary :)



How many of **these** buzzwords do you know? :)

# What is 'Big Data'?

Big Data has indeed become somewhat of a catch-phrase/buzzword.

But, we can provide an operational definition: **Big Data is data that is 'too big' to be stored in a single machine, and/or processed by a single machine.** This definition is intentionally vague, to keep it relevant for the future as well.

# Another way to characterize Big Data

Big Data is data that has:

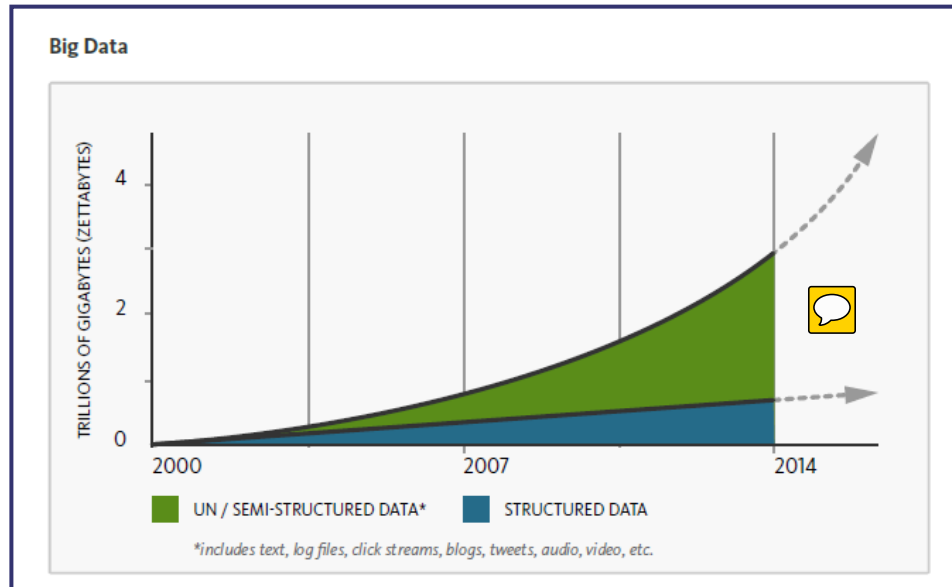
- variety 
- velocity 
- volume



In other words, it is data that is varied in nature (comprises diverse types), changes often, and comes in large quantities.






# More, MORE! (how big?)

Big Data is not only big, but getting bigger at a rapid rate..



# Sources of Big Data

## Big Data can result from:

- people: web browsing 'clickstreams' 
- people: purchasing etc. habits – shopping, dining..
- people: **social media** [communications]
- people: surveys, **polls**, census, court docs, employment histories, credit ratings.. 
- people etc: genomic data 
- people: entertainment/**education**.. 
- people + devices: medical, fitness etc. data
- people + devices: transportation [cameras, sensors, GPS..]
- devices: scientific instruments
- devices: sensors ("IoT") 



# Datafication


Wikipedia: Datafication is a modern technological trend turning many aspects of our life into computerised data and transforming this information into new forms of value. Examples of datafication as applied to social and communication media are how Twitter datafies stray thoughts or datafication of HR by LinkedIn and others.

In other words, it is the notion that people, our built environment (eg. number of freeways in the US), etc. can lead to data generation.

"Once we datafy things, we can transform their purpose and turn the information into new forms of value."

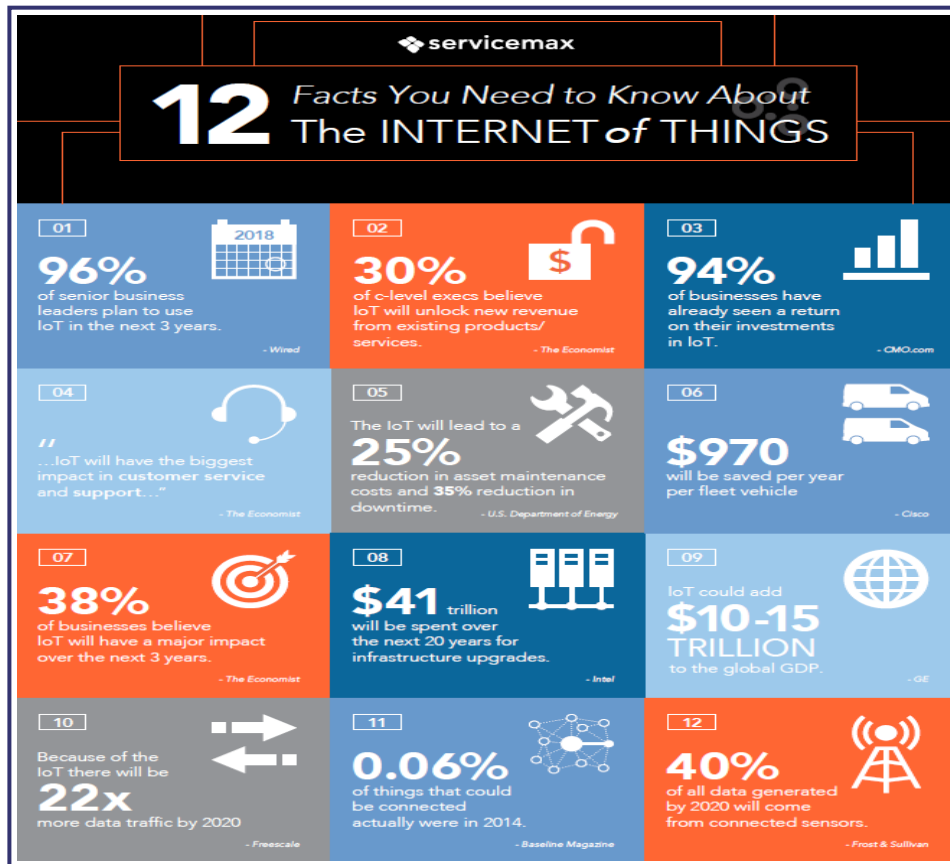


# IoT

IoT is the 'Internet of Things' - what if (almost) every lightbulb, tire, building, plane engine, bridge, fridge etc. had an IP address and a sensor, and transmits data periodically through a network? Among other things, it will lead to an \*explosion\* of data :) 

# IoT - infographic

## Interesting stats:



## Concern over security/privacy

Purchase history of products+services could reveal a lot.

Vehicle tracking: license plate pic capture is legal.

TACMA – OMG. And, More OMG.

Privacy and security are at odds at times: NSA large-scale surveillance, 'No Fly' List, real-time face tracking..

Are you being... spied on?

# Some data-related issues

Big Data can be quite useful if collected, analyzed and interpreted properly. Here are things that can be problematic:

- rigor (of data collection)
- timing
- comprehensiveness (Big Data modeling and analysis is not "value neutral")
- data standards, interoperability
- sampling
- utilization (by end users of the data)
- ...

# TMD (Too Much Data)!

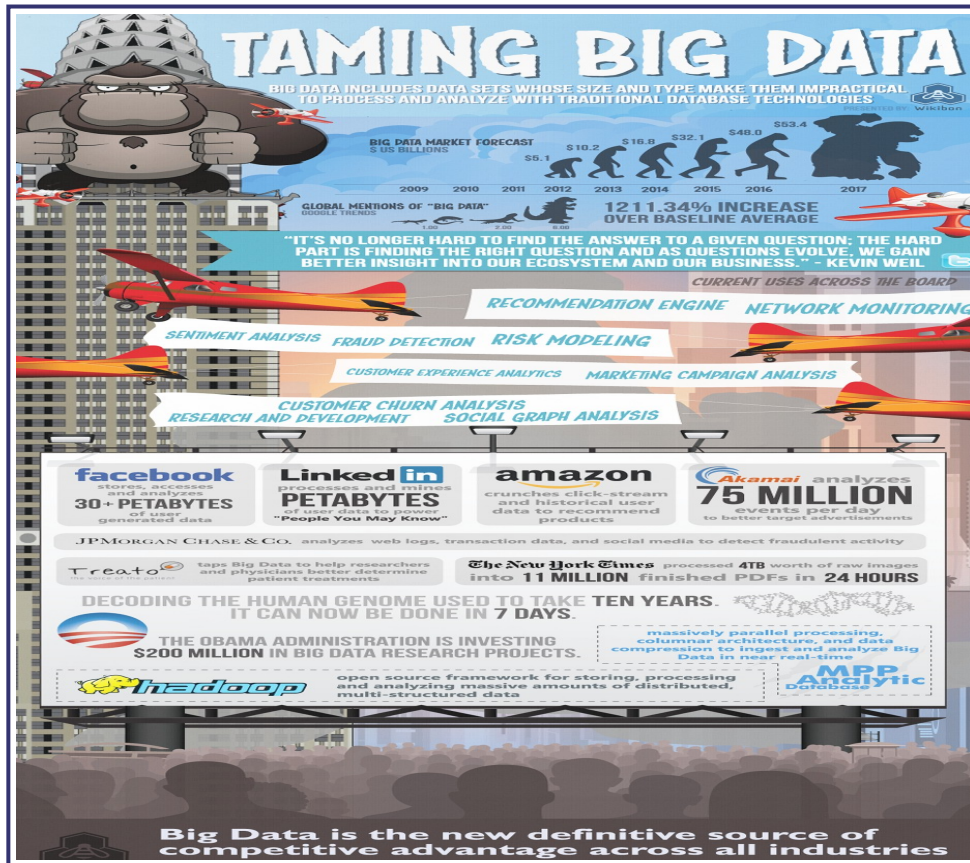
- The Large Hadron Collider at CERN generates 1 PB of data per second when it's operating and the upcoming Square Kilometer Array telescope will disgorge 1 EB of data (1,000 PB) per day. In both cases, only a small fraction of the data will be stored. Storing it all would break the budget.
- China's Internet-of-Things initiative is targeting a 10,000-fold data reduction to avoid having to confront 100ZB of data from home sensors alone by 2030. A major strategy of the effort, spearheaded by the Chinese Academy of Sciences, is to create a single mega-sensor that will do the work of the 40-50 device sensors that would otherwise be expected to inhabit the average Chinese home at that time.



Storing huge amounts of data costs time, money; retrieval could be problematic, analysis will cost as well – somewhat oblivious to such concerns, we are creating a data deluge! How come? Because sensors are ubiquitous, storage is cheap.

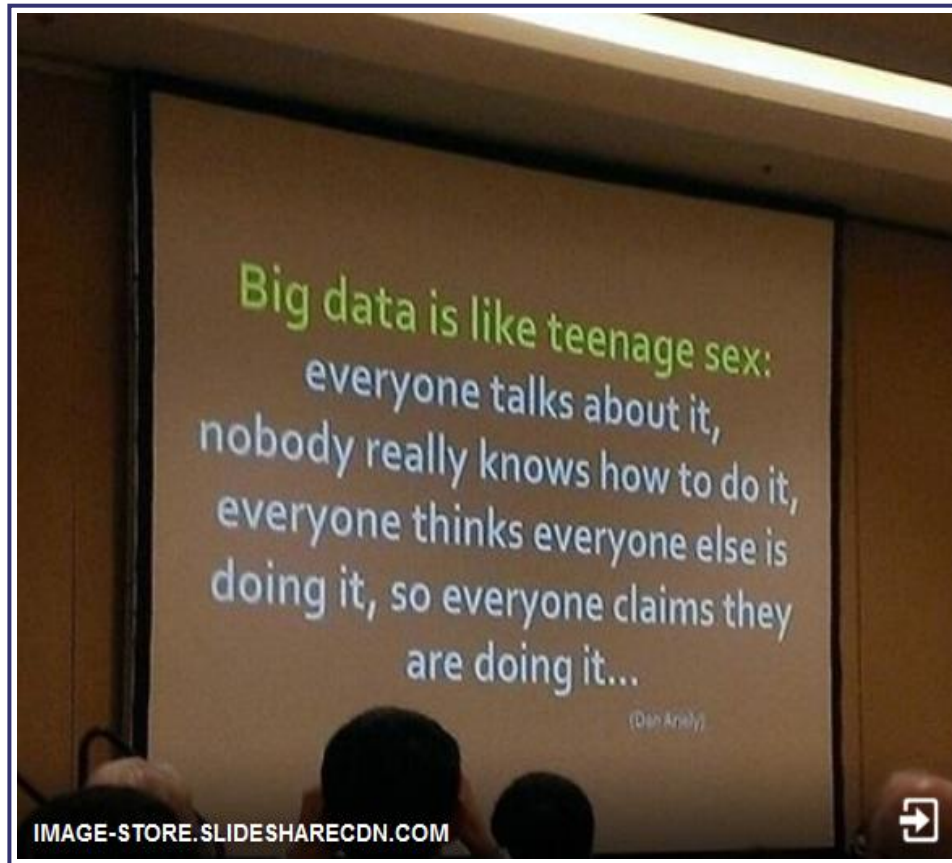
Maybe we need to be prudent: "The purpose of computing is insight, not numbers." – Richard Hamming (1962), in 'Numerical Methods for Scientists and Engineers'

# A Big Data infographic



# Who is 'doing it'?

Everyone, except you :





# Is it all (not) just hot air?

"Big data is the new oil. The companies, governments, and organizations that are able to mine this resource will have an enormous advantage over those that don't."

"Big data will generate misinformation and will be manipulated by people or institutions to display the findings they want."

So, which is it? BOTH!

# Big Data – redux

Again, these are Big Data's characteristics:



- ...> Encompasses large amounts of information
- ...> Consists of a variety of data types and formats
- ...> Generated by disparate sources
- ...> Retained for long periods
- ...> Utilized by new and innovative applications



# Why is this useful again?

What are things we can we do now, that we couldn't, before?

\* **combine multiple sources of data** (however small or seemingly insignificant) for a better 'bigger picture'



\* exploit unstructured data – voice, video, images, tweets, blog posts..

\* provide insights to [internal] frontline managers in near-real-time (to enable making more agile business decisions)

\* experiment with the marketplace (fluid price-setting) as often as needed!

So here's what is new: better insight, quicker action.



# How long will this be useful/relevant?

According to IEEE (and others), a **long** time.

# Conferences, organizations, LA user group



Here are some links:

- <http://www.big-dataservice.net/>
- <https://theinnovationenterprise.com/summits/big-data-innovation-summit-las-vegas-2017>
- <http://bigdata.ieee.org/>
- <http://www.meetup.com/Los-Angeles-Big-Data-Users-Group/>

# Summary



We are at the start of a transformative phase, fed by our relatively-new ability to collect, store, **analyze** and **benefit** from **MASSIVE** amounts of data from every walk of life.



# "Carpe Datem" - Seize the data!

