Lambton College | QUEEN'S COLLEGE of BUSINESS, TECHNOLOGY & PUBLIC SAFETY

# Introduction to the course

Sep 10, 2021

James Hong, PhD

# Bio and disclosure

- HBSc, University of Toronto, Biology and Computer Science
- PhD, University of Toronto, Neuroscience/Bioinformatics

- Post-doctoral Fellow, UHN
  *1) Deep learning on profiling and classifying neural stem cells*
  *2) Improving surgical outcomes of degenerative cervical myelopathy*

**Disclosure**

- Founder of Verismo Apps (www.verismo.xyz) and Verismo Health (www.verismohealth.com) harnessing machine learning and AI for applications that range from sales analytics (Terapeak @ eBay) to clinic management (computer vision for patient registration and integration into EMR)

# Theme of this Course

***Large-Scale Data Management***

***Big Data Analytics***

***Data Science and Analytics***

- How to manage very large amounts of data and extract value and knowledge from them

# Introduction to Big Data

***What is Big Data?***

***What makes data, “Big” Data?***

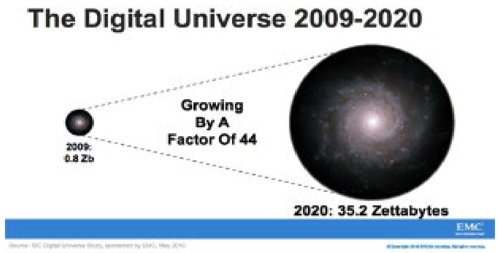
# Big Data Definition

- No single standard definition…

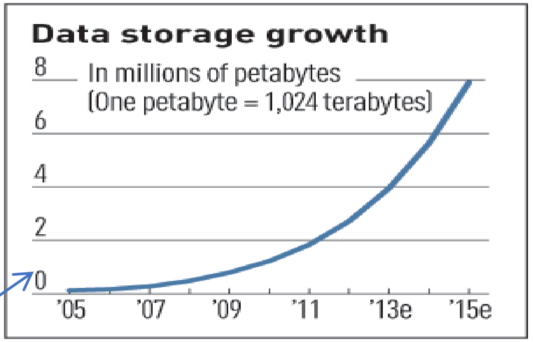
"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

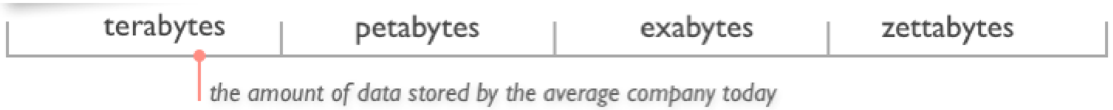
# Characteristics of Big Data: 1-Scale (Volume)

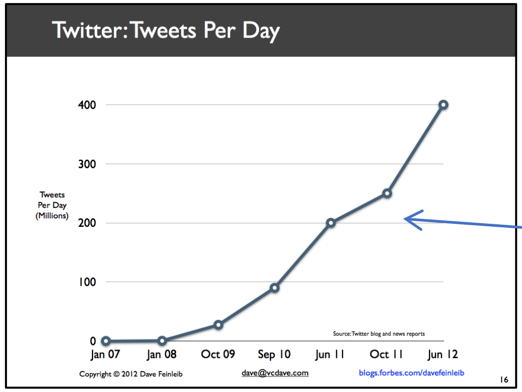
- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially


The Digital Universe 2009-2020

Source: IDC Digital Universe Study, sponsored by EMC, May 2010


Data storage growth




Twitter: Tweets Per Day

Source: Twitter blog and news reports

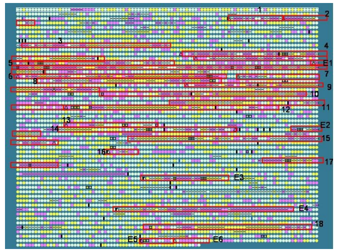
Copyright © 2012 Dave Feinleib dave@vcdave.com blogs.forbes.com/davefeinleib

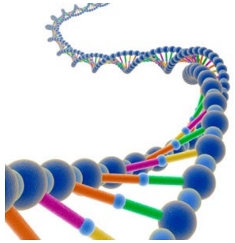
*Exponential increase in collected/generated data*

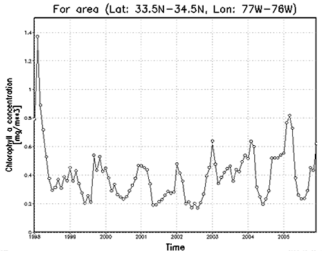
# Characteristics of Big Data:
# 2-Complexity (Varity)

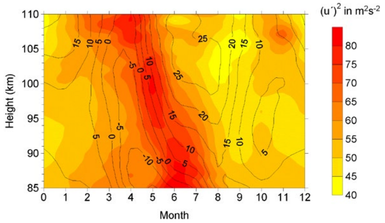
- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data

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





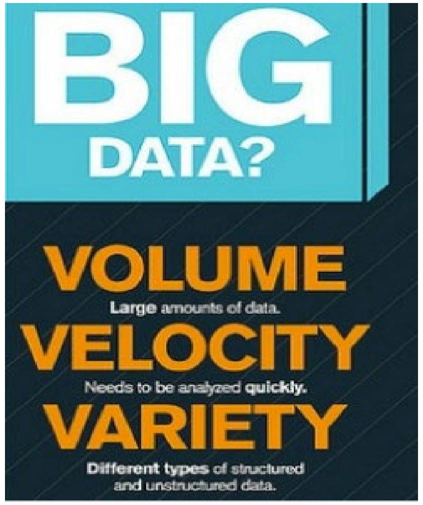

For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m**3]
Time

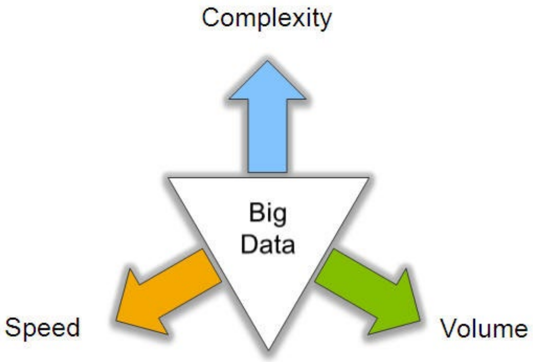


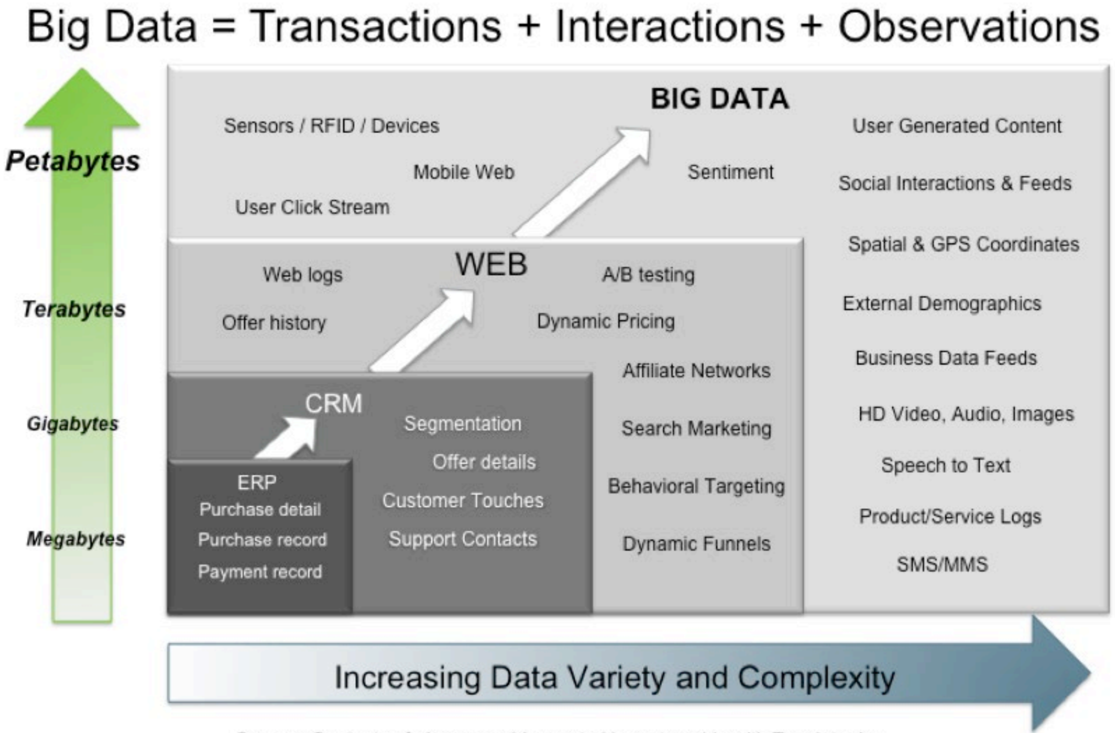
# Characteristics of Big Data:
# 3-Speed (Velocity)

- 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

# Big Data: 3V’s

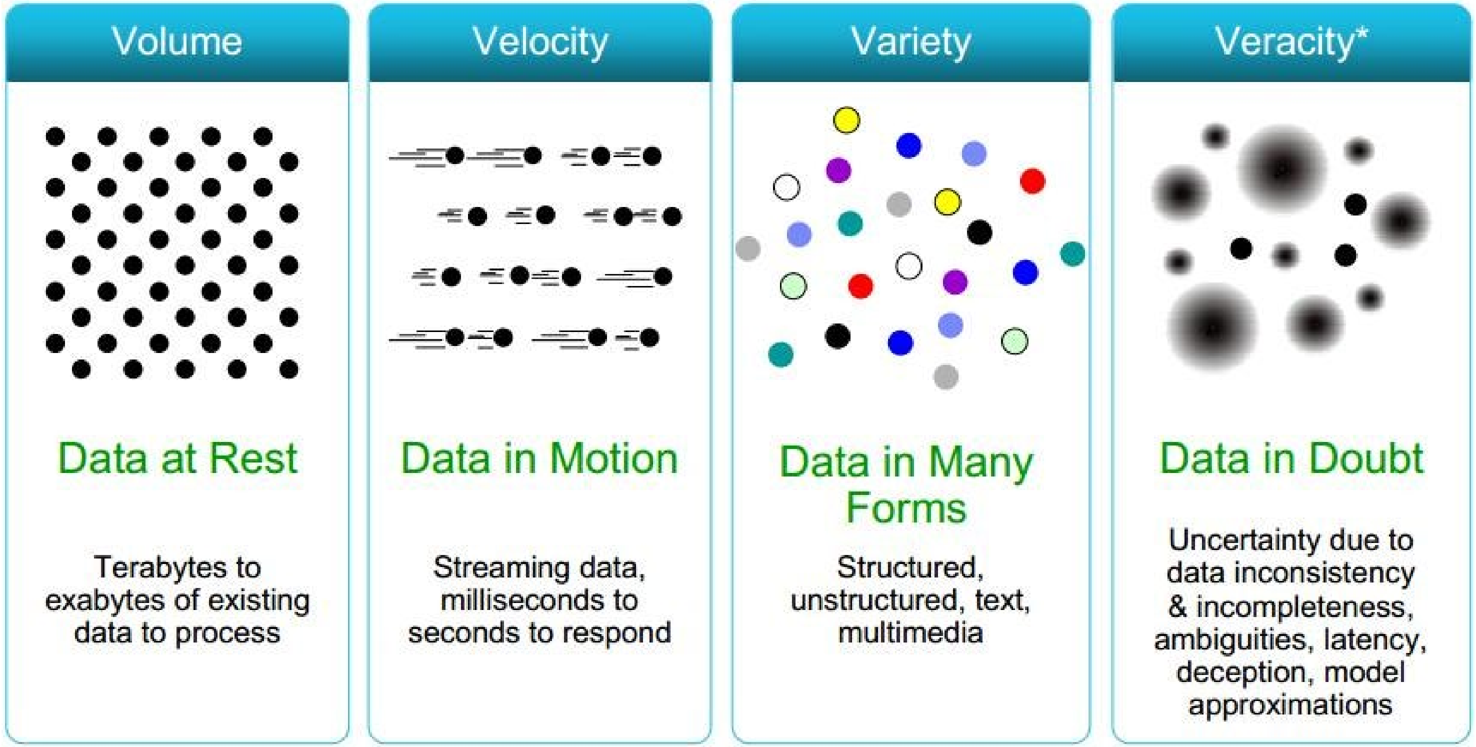

BIG
DATA?
VOLUME
Large amounts of data.
VELOCITY
Needs to be analyzed quickly.
VARIETY
Different types of structured and unstructured data.





**Source**: *Contents of above graphic created in partnership with Teradata, Inc.*

# Some Make it 4V's



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

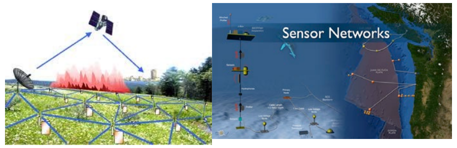
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

# Who’s Generating Big Data

**Social media and networks**
(all of us are generating data)

**Scientific instruments**
(collecting all sorts of data)

**Mobile devices**
(tracking all objects all the time)



**Sensor technology and networks**
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

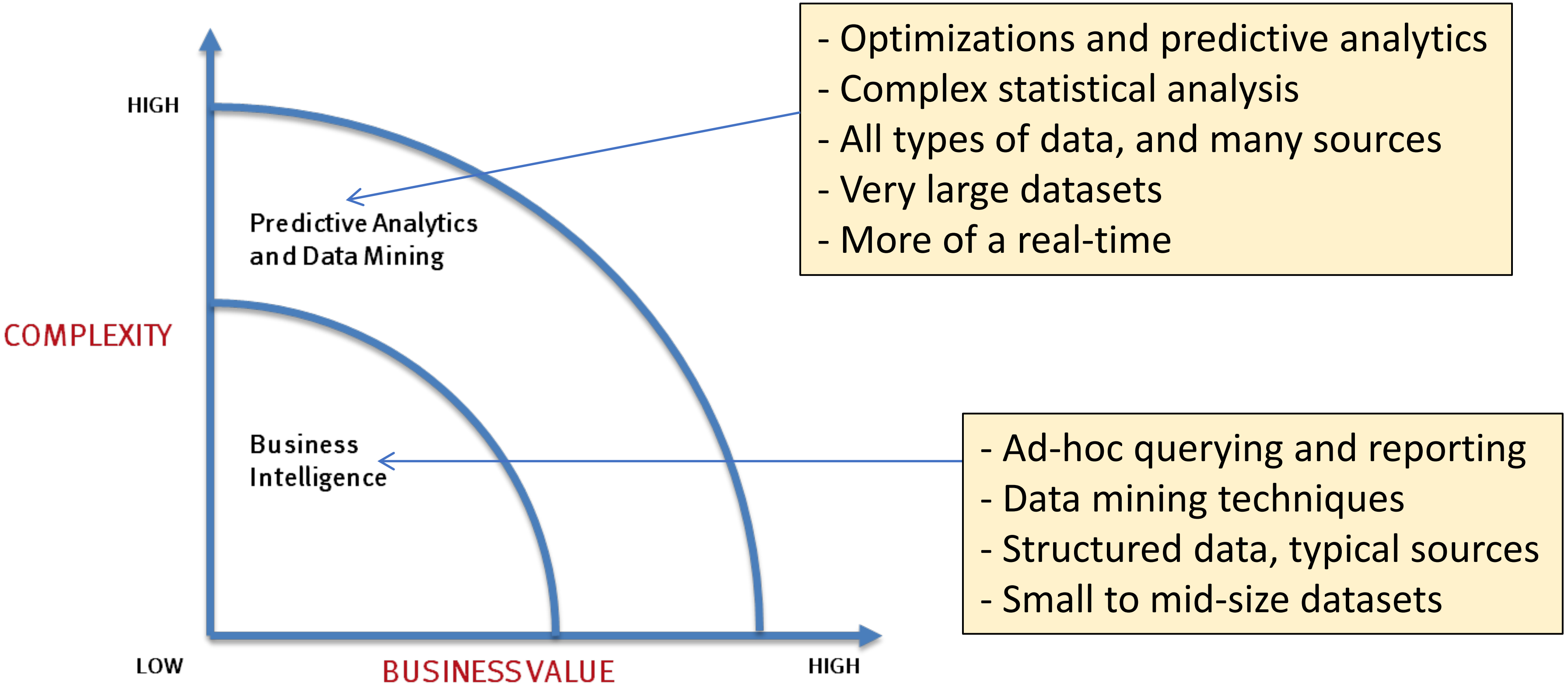
# The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

**Old Model:** Few companies are generating data, all others are consuming data

**New Model:** all of us are generating data, and all of us are consuming data

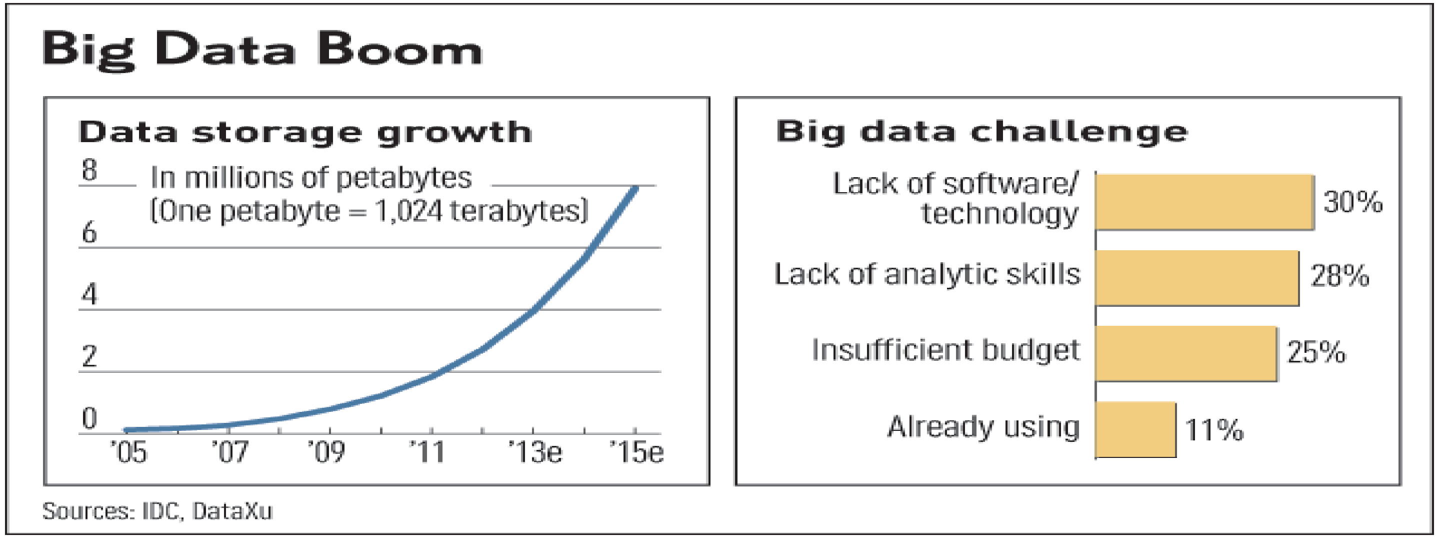
# What's driving Big Data



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

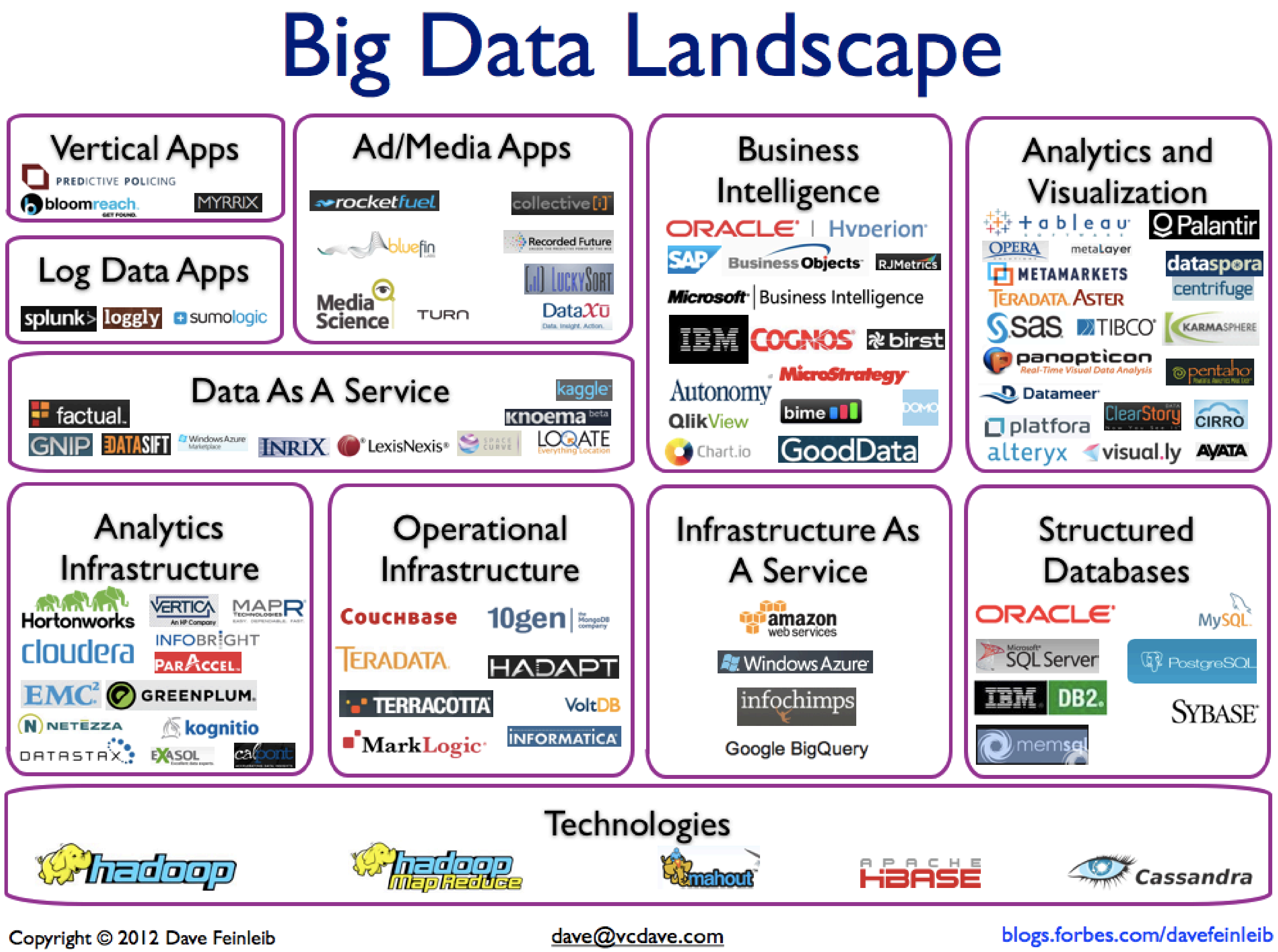
# Challenges in Handling Big Data



Sources: IDC, DataXu

- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

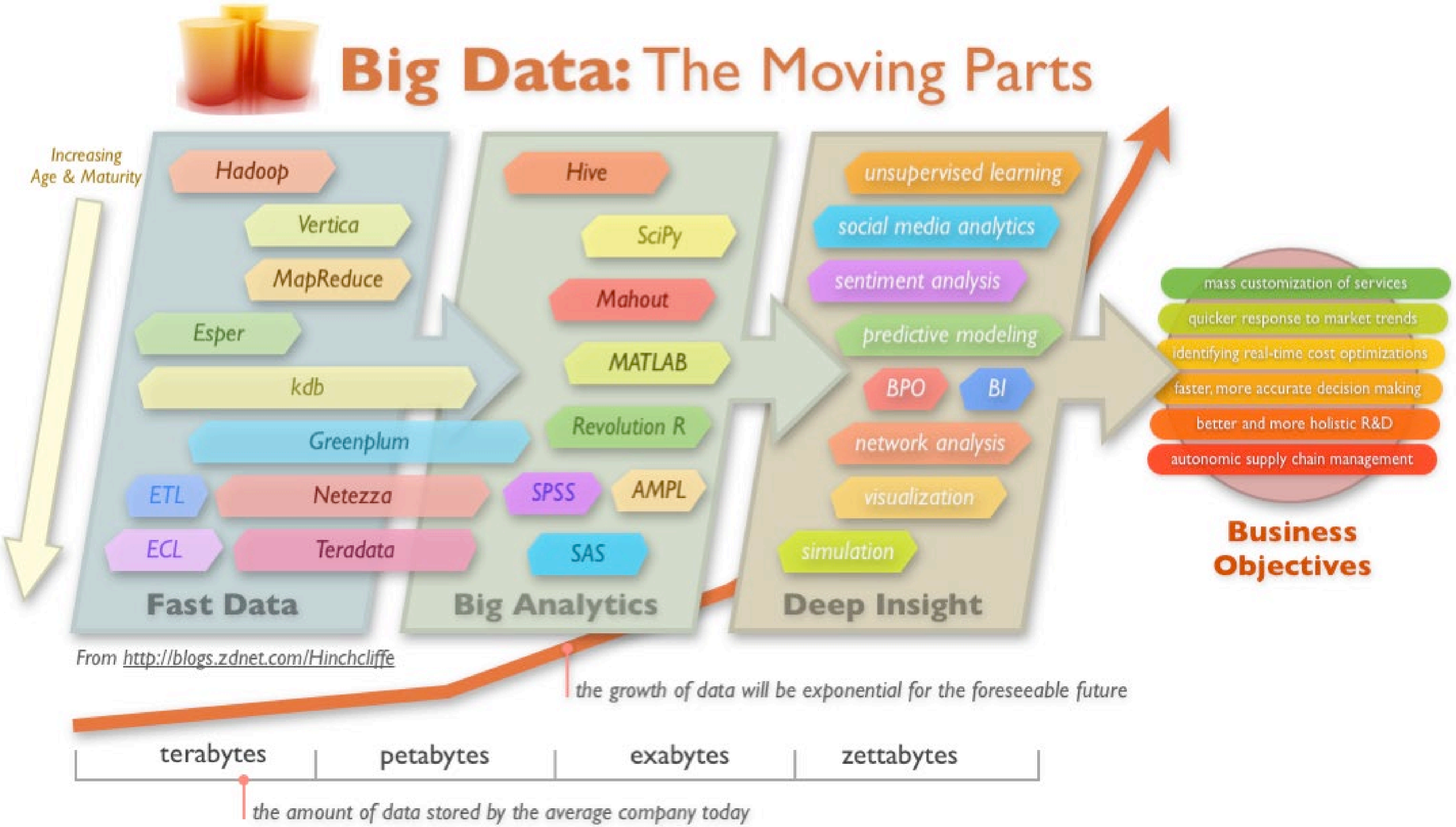
# What Technologies Do We Have For Big Data ?



Copyright © 2012 Dave Feinleib

dave@vcdave.com

blogs.forbes.com/davefeinleib



# Big Data: The Moving Parts

the growth of data will be exponential for the foreseeable future

the amount of data stored by the average company today

From http://blogs.zdnet.com/Hinchcliffe

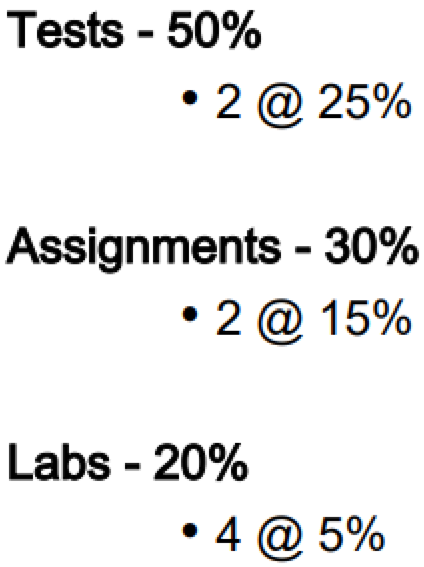
# What You Will Learn…

- The basics of data analytics and its applications
- We focus on ***Hadoop/MapReduce technology***
- **Learn the platform**
  - How big data are managed in a scalable, efficient way
- **Learn writing Hadoop jobs in different languages**
  - Programming Languages: Java, Python
  - High-Level Languages: Apache Pig, Hive

# Changes to the schedule

Lectures on Sep 18th and Sep 25th will be asynchronously delivered

# Course breakdown



**Tests - 50%**

- 2 @ 25%

**Assignments - 30%**

- 2 @ 15%

**Labs - 20%**

- 4 @ 5%