Big Data Technologies and Datasets

Lecturer: Dr. Erez Shmueli
shmueli@tau.ac.il

Teaching Assistant: Tamir Mendel
tamirmendel@mail.tau.ac.il

Course Prerequisites

- Programming (Python)
- Statistics
- Databases
- Information Systems Engineering

Grade components

- Exam – 60%

- HW – 40%
  (theory + coding, 2 HW assignments)

Course Schedule

Lecture	**Topic**	**HW**
1	Introduction	
2	Distributed RDBMS	
3	NoSQL DBMS	
4	Hadoop ecosystem & architecture	
5	Lab 1 (MongoDB, Redis)	
6	MapReduce	RDBMS + NoSQL MapReduce
7	Spark	
8	SQL for big data	
9	Lab 2 (Hadoop, Spark)	Hadoop + Spark + SQL
10	Lab 3 (Sqoop, Hive, Impala)	
11	Guest Lecture	
12	Machine learning for big data + Mllib	

Big Data Technologies and Datasets

Lecture 1 – Introduction

Heavily based on slides by Assaf Araki
(Intel Advanced Analytics Team)

Outline

- Analytics
- Big Data

Analytics Overview

Big Data Analytics Today

“My daughter got this in the mail!” he said. “She’s still in high school, and you’re sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?”

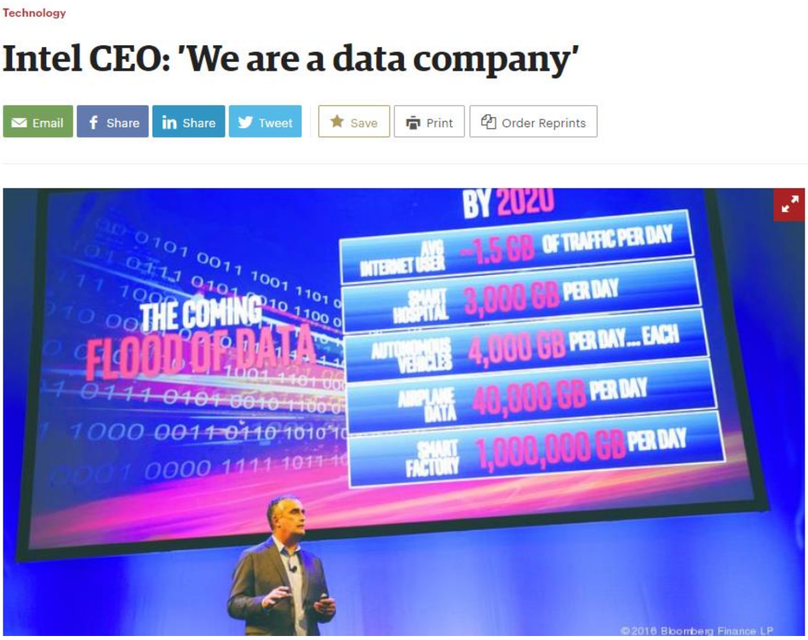
TARGET

Historical data + Algorithms = Hidden Patterns

Business Analytics

- Business analytics (BA) is the practice of iterative, methodical exploration of an organization’s data with emphasis on statistical analysis.
- BA is used to gain insights that inform business decisions and can be used to automate and optimize business processes.
- Data-driven companies treat their data as a corporate asset and leverage it for competitive advantage.

WhatIs.com

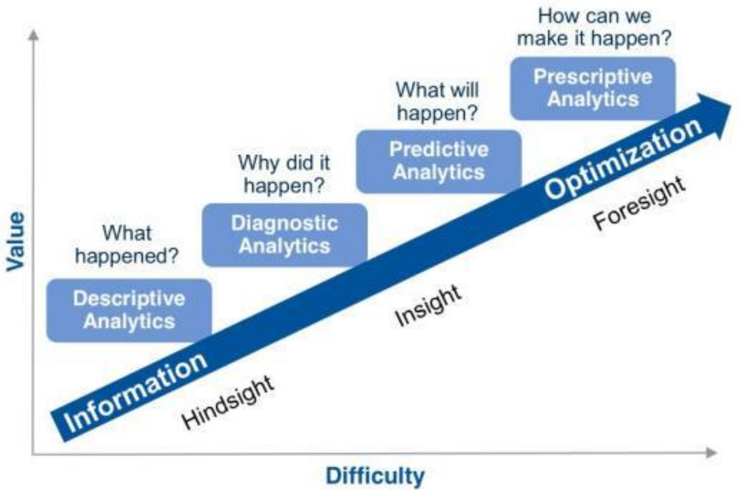


Technology

Intel CEO: 'We are a data company'

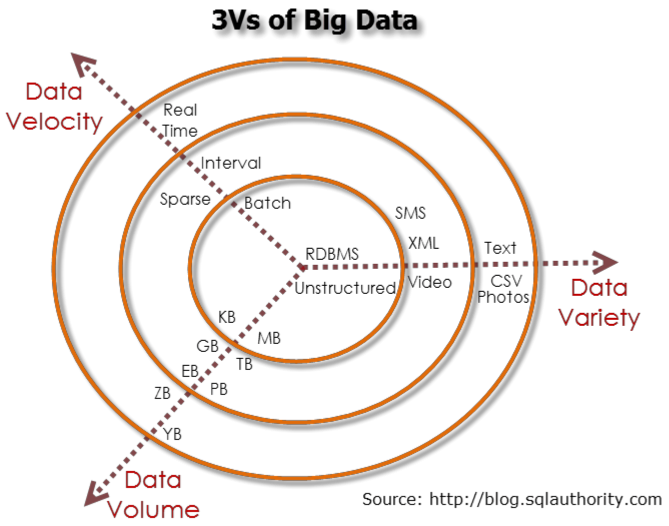
Email
Share
in Share
Tweet
Save
Print
Order Reprints
BY 2020
AVG INTERNET USER ~1.5 GB OF TRAFFIC PER DAY
SMART HOSPITAL 3,000 GB PER DAY
AUTONOMOUS VEHICLES 4,000 GB PER DAY... EACH
AIRPLANE DATA 40,000 GB PER DAY
SMART FACTORY 1,000,000 GB PER DAY
THE COMING FLOOD OF DATA
©2016 Bloomberg Finance LP

Levels of Analytics



Big Data Overview

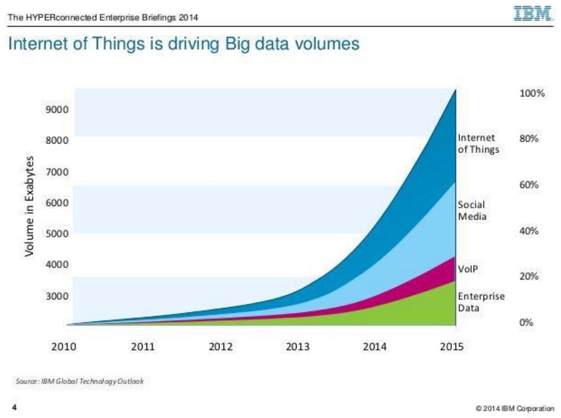
Data Characteristic

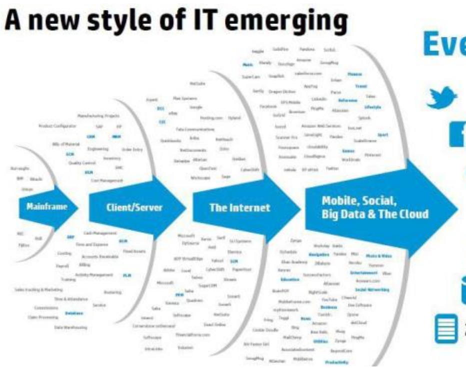


Source: http://blog.sqlauthority.com

Volume (Data Size)

Enterprise Data ->
Internet Data ->
Internet of Things Data




A new style of IT emerging
Mainframe
Client/Server
The Internet
Mobile, Social, Big Data & The Cloud

Every 60 seconds

98,000+ tweets

695,000 status updates

11million instant messages

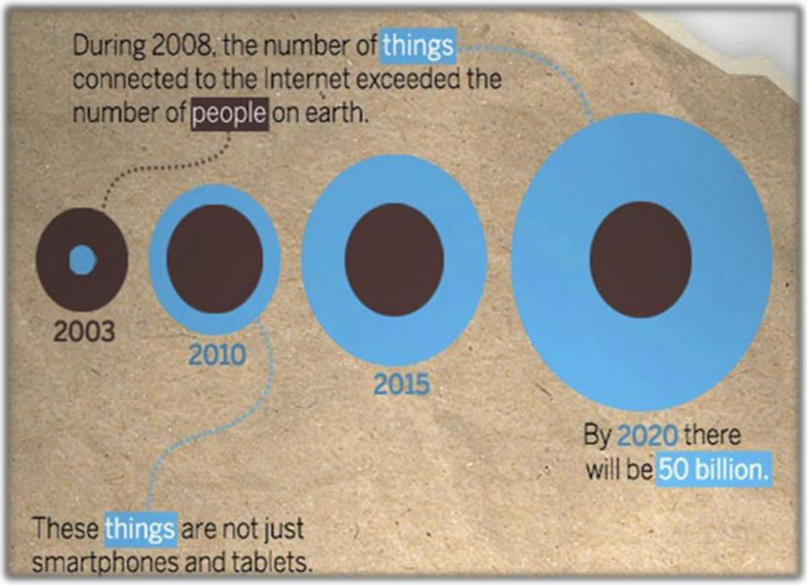
698,445 Google searches

168 million+ emails sent

1,820TB of data created

217 new mobile web users

Volume (Data Size) - Internet of Things (IoT)



During 2008, the number of things connected to the Internet exceeded the number of people on earth.

By 2020 there will be 50 billion.

These things are not just smartphones and tablets.

Velocity - 3 Layers of Analytics

Edge

Real-time

Event processing / pattern matching

Data filtering

Limited compute/storage

Stream

Near Real-time

In-memory analysis

Larger compute, Limited storage

Batch

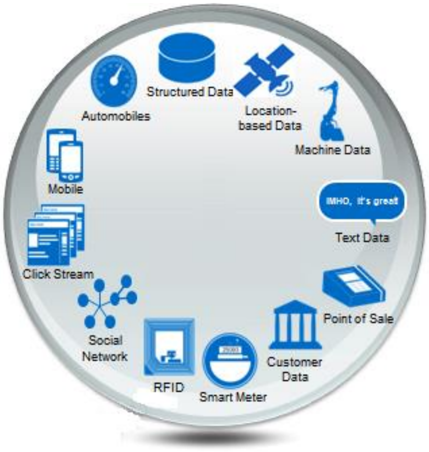
Ad-hoc

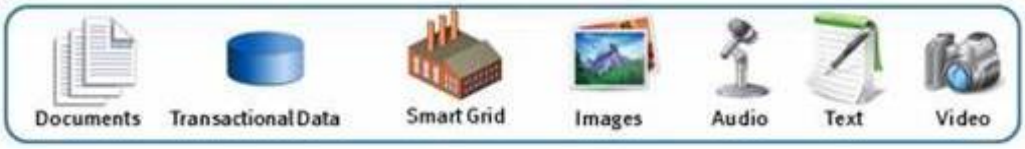
Identification of hidden patterns

Cross-device learning

Large compute/storage

Variety – sources and formats





What are BigData Platforms?

- Big Data Definitions
  - Wikipedia* - a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools
  - Forrester* / Gartner* – 3V's (Volume, Velocity & Variety) or 4V's (and Veracity)
  - Moore's Law of Big Data - "The amount of nonsense packed into the term BIG DATA doubles approximately every two years" (Mike Pluta)
  - "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..." (Prof. Dan Ariely)

- From a processing point of view:
  - Big Data is simply: A Cluster.

Why A Cluster ?

- Main Characteristics
  - Distributed – small portion of data reside on each node
  - Share nothing – each node is an autonomic unit of CPU, RAM and storage
  - Processing occurs where data reside (minimal data movement)
  - Scale Out vs. Scale Up

19

BIG DATA & AI LANDSCAPE 2018

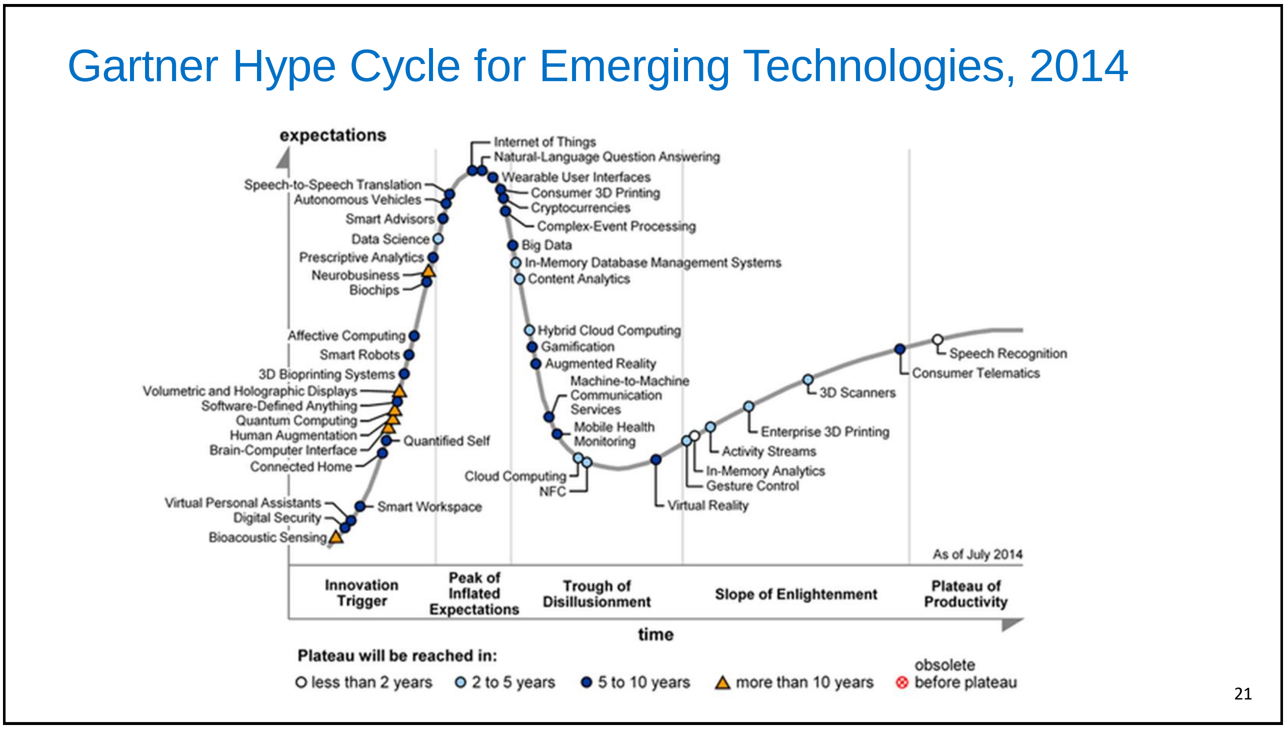
INFRASTRUCTURE
ANALYTICS
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
CROSS-INFRASTRUCTURE/ANALYTICS
OPEN SOURCE
DATA SOURCES & APIs
DATA RESOURCES

V1 – Last updated 6/19/2018

© Matt Turck (@mattturck), Demilade Obayomi (@demi_obayomi), & FirstMark (@firstmarkcap) mattturck.com/bigdata2018

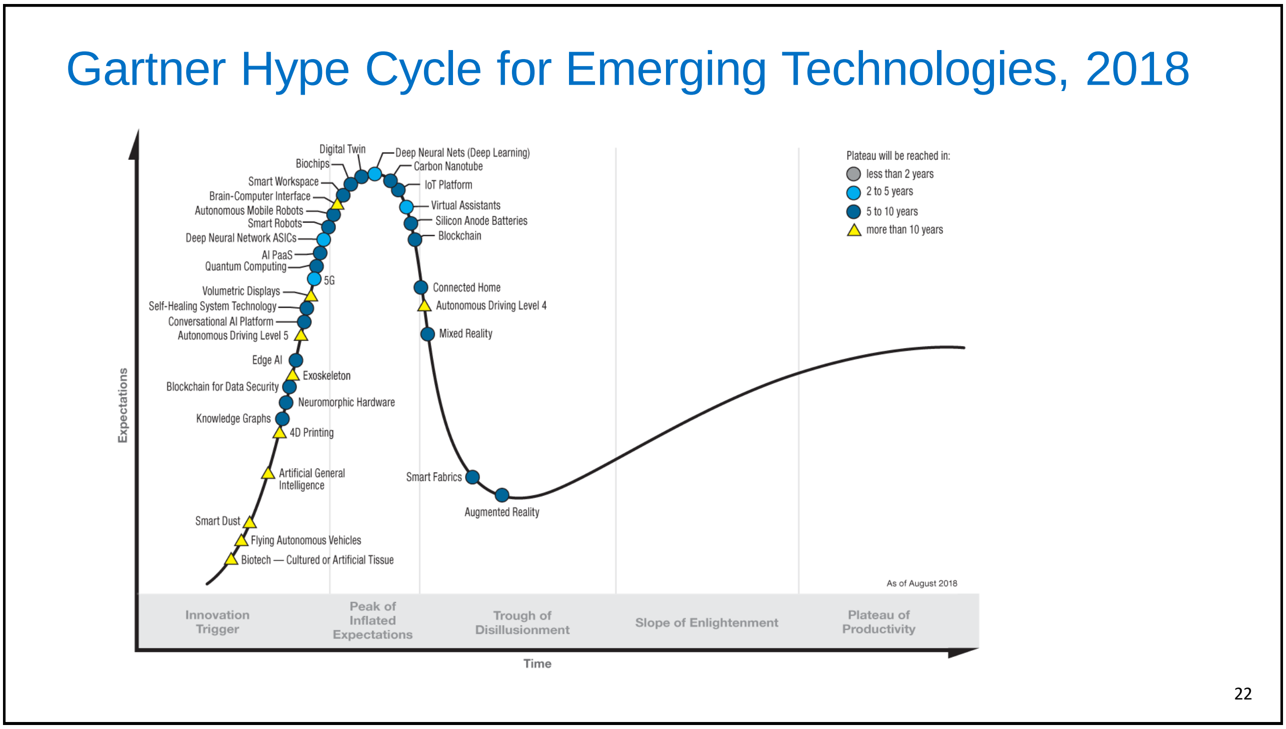
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

20



Gartner Hype Cycle for Emerging Technologies, 2014

21



Gartner Hype Cycle for Emerging Technologies, 2018

22

Big Data Limitations

- Limitations of Big Data Platforms
  - Not all platforms have same APIs (e.g. R, Java, Scala, Python etc.)
  - Most platform are for specific purposes (batch vs. real time, key-value vs. document, etc.)
  - Solutions are immature (lack of features, e.g. security)
  - …
- Limitations of Big Data Analytics - The Distribution Curse
  - Most algorithms written Sequential
  - It requires a change in the Data Scientist’s mind set to create & implement distributed algorithms
  - No cross platforms code – Different APIs and processing methods
  - Can’t leverage existing algorithms (e.g. R has more than 4000 packages)
  - …