

Big Data Systems (CS4545/CS6545)
Winter 2021

Introduction to Big Data Systems

Suprio Ray

UNB, Fredericton


About myself

- Research interests: <http://www.cs.unb.ca/~sray/>
 - Big Data and Database Systems
 - Scalable Systems for data science and analytics
 - Spatio-temporal and spatio-textual data management
- Education:
 - PhD: Univ. of Toronto
 - MSc: Univ. of British Columbia
- Work experience as a software engineer:
 - Oracle, Lucent (Bell Labs), Webtech Wireless, SAP
- Courses I am teaching during 2020 - 2021:
 - CS2545 (Data Science)
 - CS3543 (Database Systems and Administration)
 - CS4545/CS6545 (Big Data Systems)

Acknowledgement

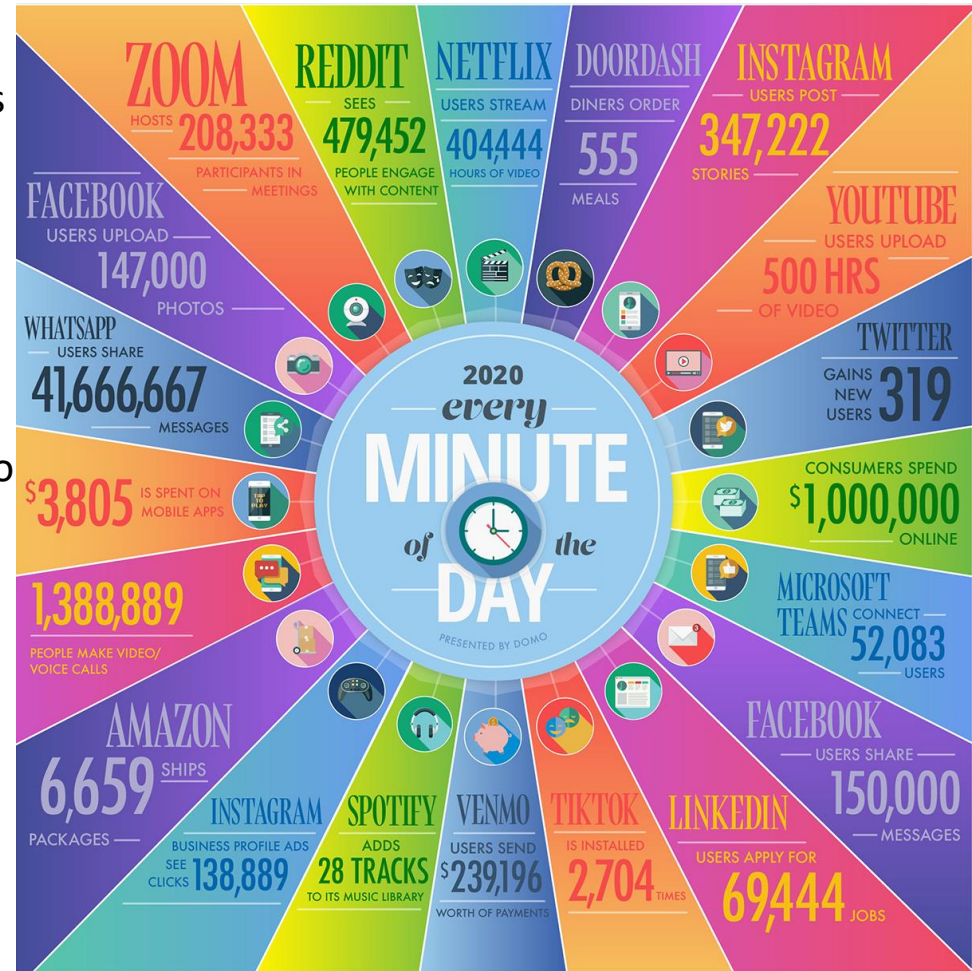
Thanks to N. Koudas, S. Babu, B. Mennecke, R. Jin, B. Zhang and Ramakrishnan and Gerhke for some of the materials in these slides. Also thanks to various articles and research papers.

Outline

- What is Big Data 
- Big Data dimensions
- Data Systems primer
- Origin of “Big” Data Systems
- Why Big Data Systems
- Types of Big Data Systems
- Course logistics

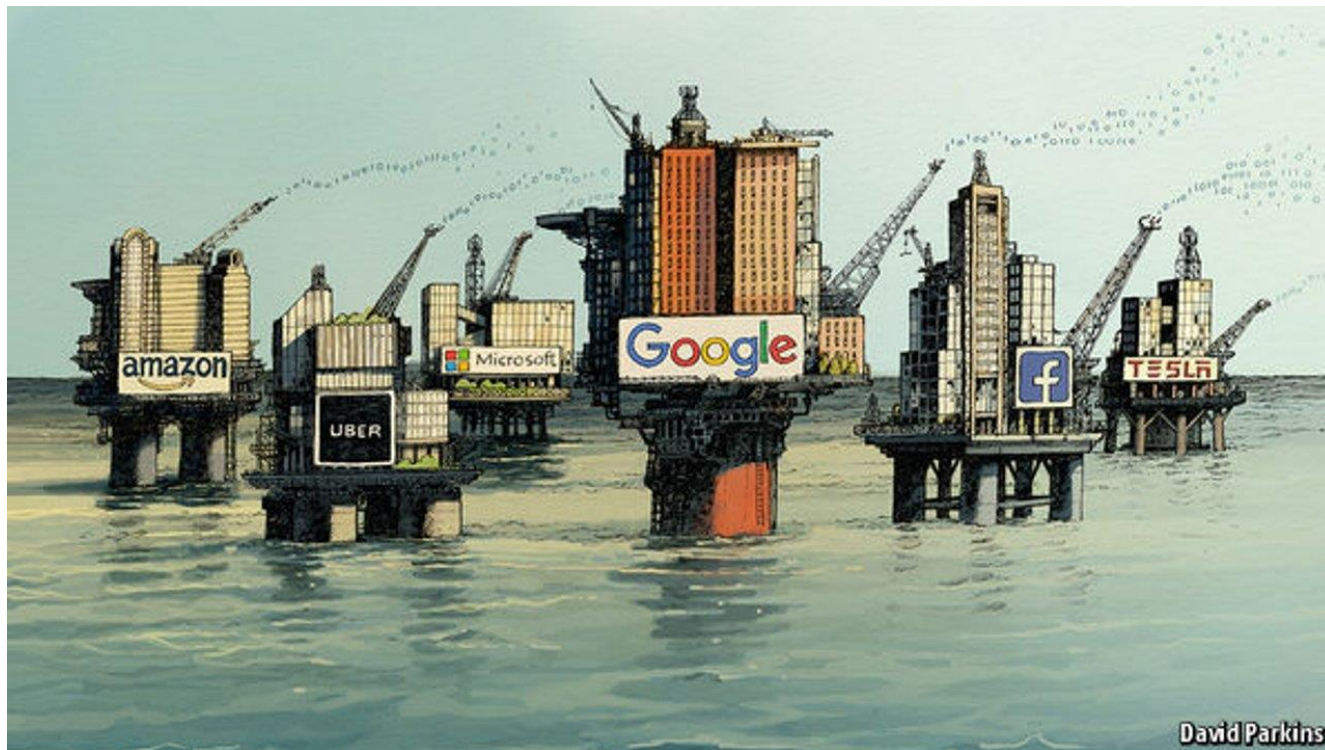
The World of Big Data

- Each minute in 2020
 - Whatsapp users share 41,666,667 messages
 - Google processes 3.8 million search queries
 - Netflix users stream 404,444 hours of video
 - On YouTube users upload 500 hours of video
 - Facebook users share 150,000 messages
 - Zoom hosts 208,333 persons in meetings
 - Amazon ships 6,659 packages
 - Consumers spent \$1,000,000 online



Is Big Data the new oil?

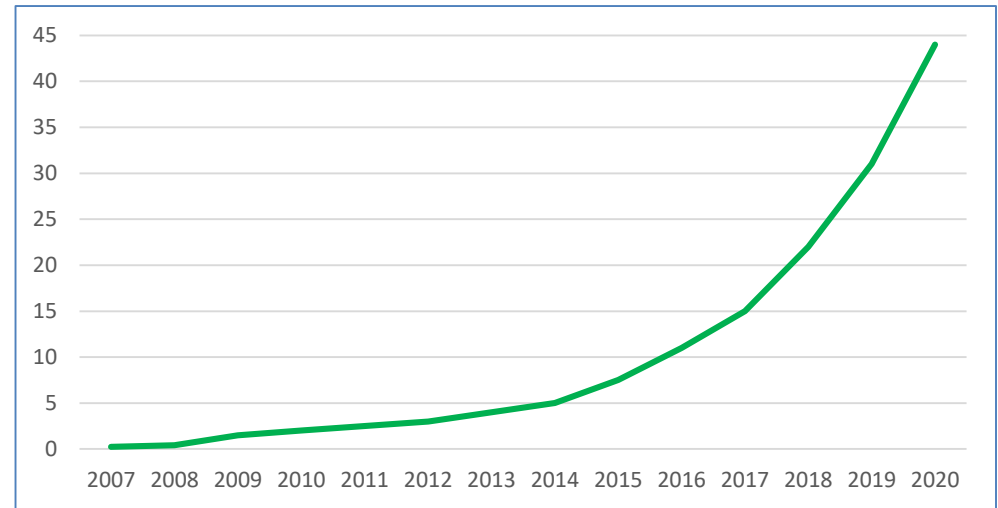
- The 5 most valuable companies in the world (Alphabet, Amazon, Apple, Facebook & Microsoft) have a total market capitalization of **US\$ 5 Trillion**
 - Amazon captures half of all dollars spent online in America.
 - Google and Facebook accounts for almost all the revenue growth in digital advertising in America.



Growing data volume

- Data deluge
 - 30 ZB in 2019
 - 44 ZB in 2020

In zettabytes (ZB)



Src: Oracle

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	a paragraph of a text document
Megabytes (MB)	1,000 Kilobytes	a small novel
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5th Symphony
Terabytes (TB)	1,000 Gigabytes	all the X-rays in a large hospital
Petabytes (PB)	1,000 Terabytes	half the contents of all US academic research libraries
Exabytes (EB)	1,000 Petabytes	about one fifth of the words people have ever spoken
Zettabytes (ZB)	1,000 Exabytes	as much information as there are grains of sand on all the world's beaches
Yottabytes (YB)	1,000 Zettabytes	as much information as there are atoms in 7,000 human bodies

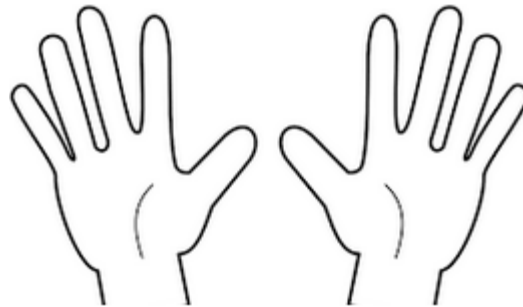
Src: Roy Williams, Caltech

What is Big Data

- Why is large data volume a problem?
- Big data is a broad term for **data sets** so **large** or **complex** that **traditional data processing** applications are **inadequate**.
(Wikipedia)
- **Data processing capacity** is a key factor

Big Data in the stone age

- People used their fingers to count stuffs



Src: giphy.com

- “Big data” for them?
 - Eleven!

Big Data in the Roman era

- Roman numerals
 - Uses letters to represent numbers, instead of digits

1	5	10	50	100	500	1000
I	V	X	L	C	D	M

- When a symbol appears after a larger symbol it is added
Example: VI = V + I = 5 + 1 = 6
- But if the symbol appears before a larger symbol it is subtracted
Example: IX = X - I = 10 - 1 = 9

Big Data in the Roman era

- Roman numerals
 - Uses letters to represent numbers, instead of digits

1	5	10	50	100	500	1000
I	V	X	L	C	D	M

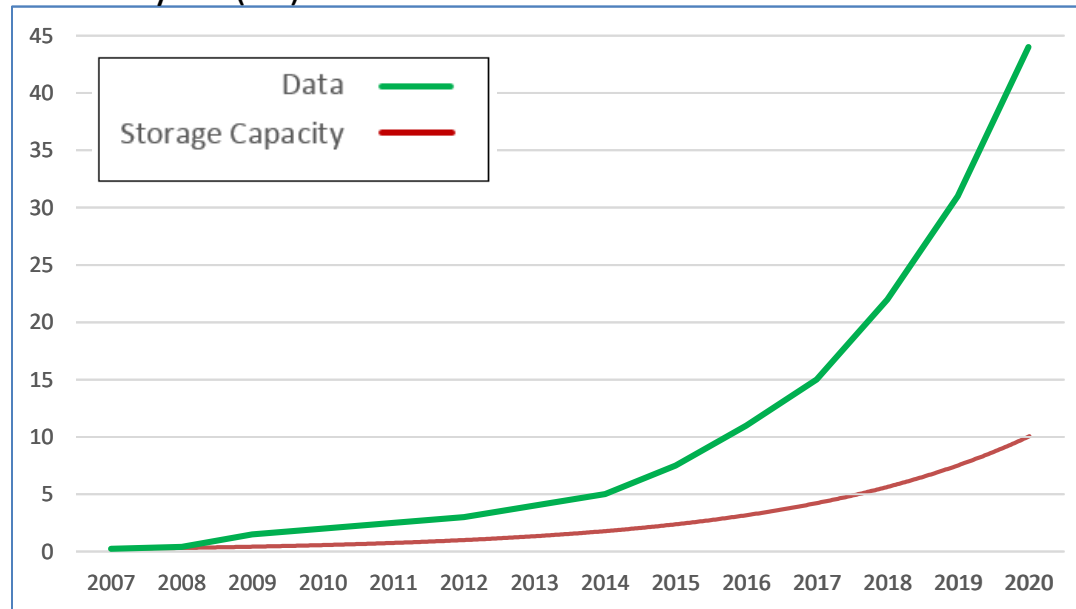
- When a symbol appears after a larger symbol it is added
Example: VI = V + I = 5 + 1 = 6
- But if the symbol appears before a larger symbol it is subtracted
Example: IX = X - I = 10 - 1 = 9

- Largest number in conventional (early) Roman numerals?
 - MMMCMXCIX (i.e. 3,999)
 - Later (in the middle ages) they invented a way to express larger numbers by placing a – above a letter (\overline{M} = 1,000,000)

Big Data today: growing data capacity gap

- Data deluge
 - 44 ZB by 2020
- Not enough capacity to store all generated data!
 - Let alone processing...

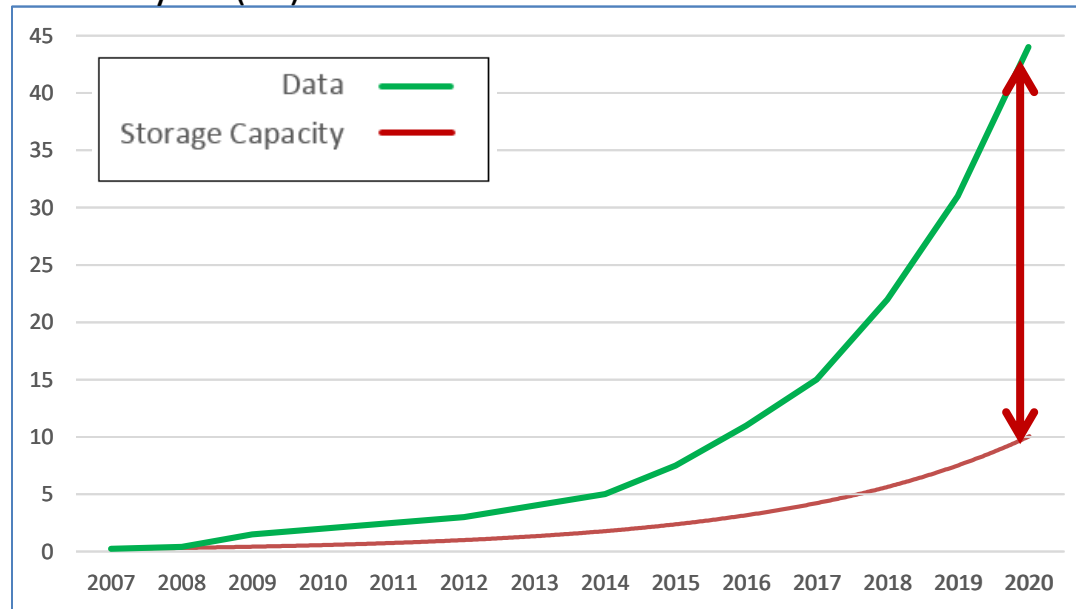
In zettabytes (ZB)



Big Data today: growing data capacity gap

- Data deluge
 - 44 ZB by 2020
- Not enough capacity to store all generated data!
 - Let alone process all the data...

In zettabytes (ZB)




Big Data - definition

- **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.
- The challenges include **capture**, **curation**, **storage**, **search**, **sharing**, **transfer**, **analysis**, and **visualization**.

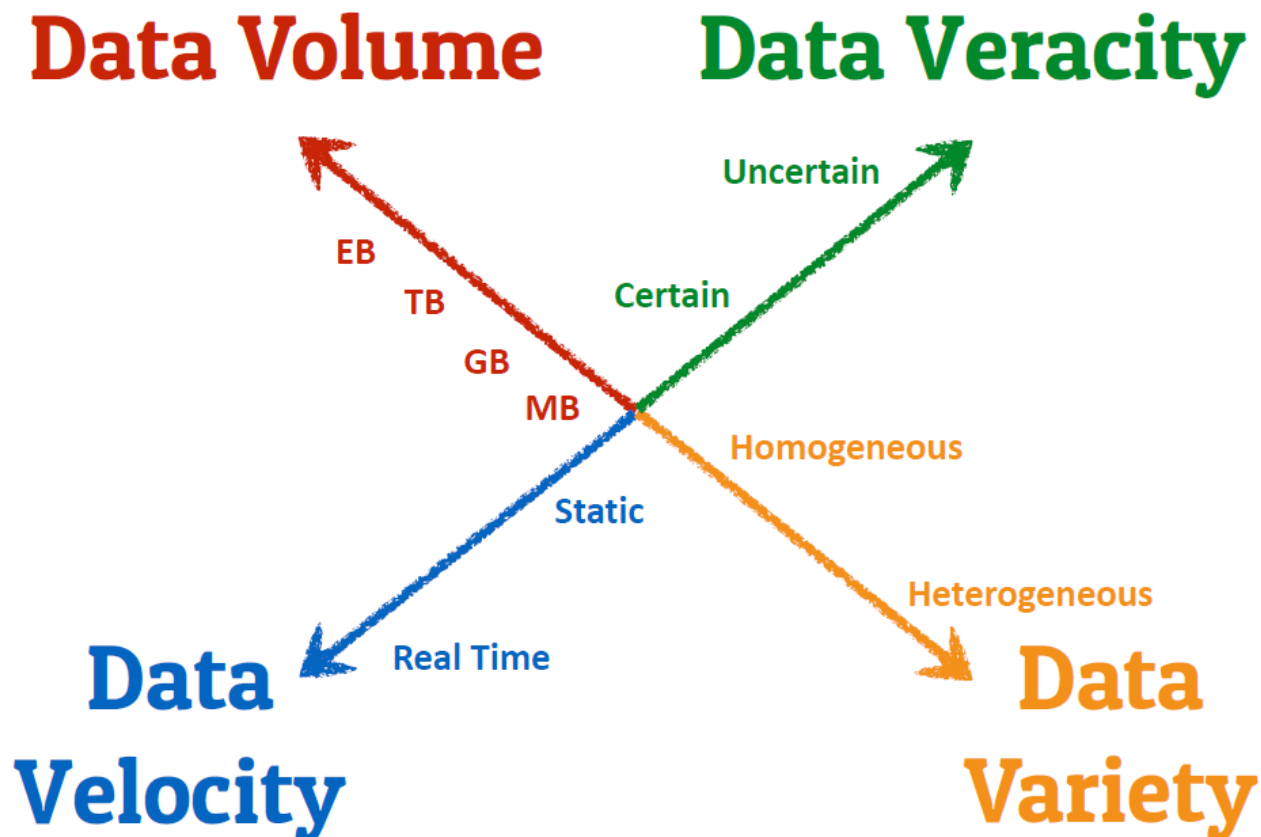
(Wikipedia)

Outline

- What is Big Data
- Big Data dimensions 
- Data Systems primer
- Origin of “Big” Data Systems
- Why Big Data Systems
- Types of Big Data Systems
- Course logistics

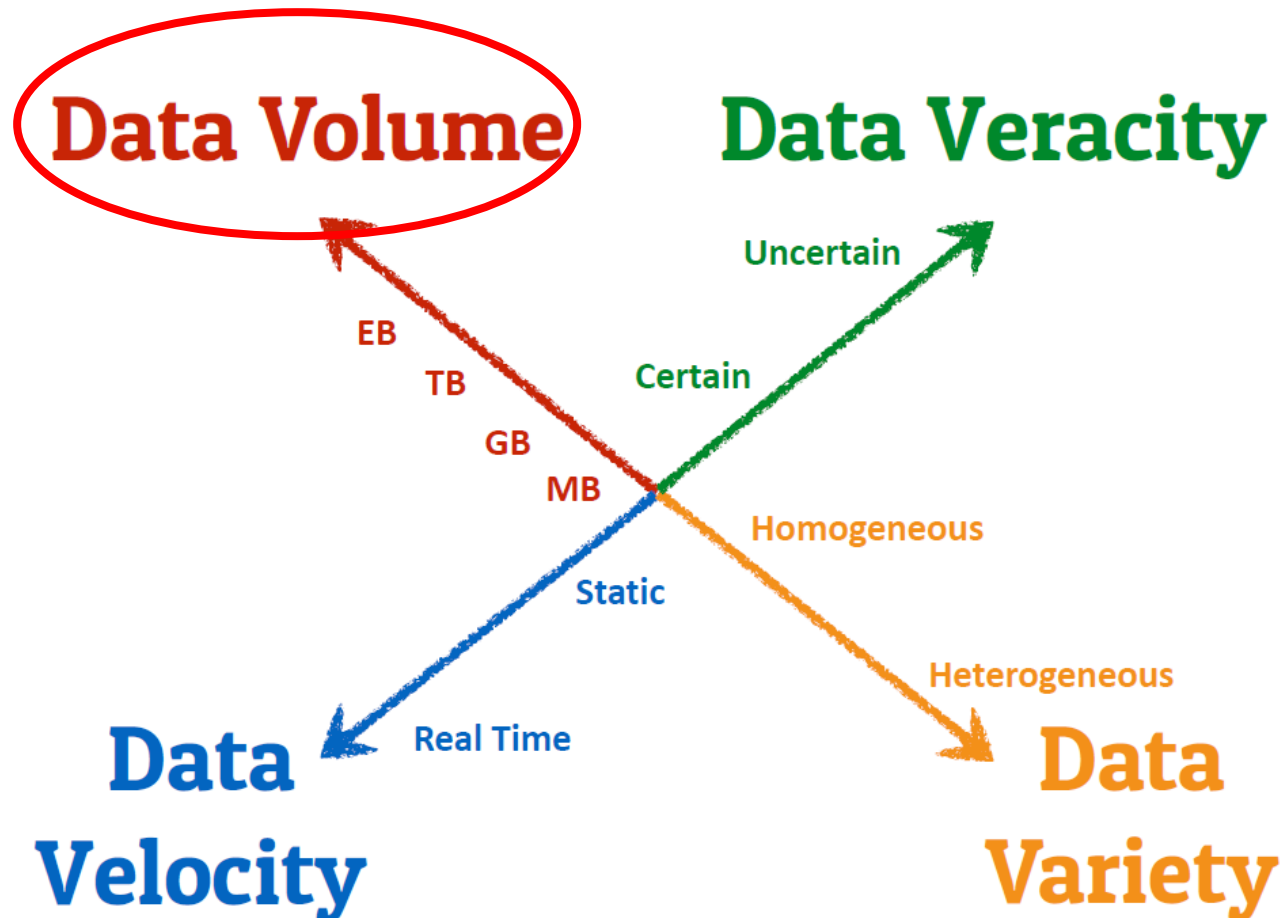
Dimensions of Big Data?

- There are many examples of "data", but what makes some of it "big"? The classic definition revolves around the **four Vs**.



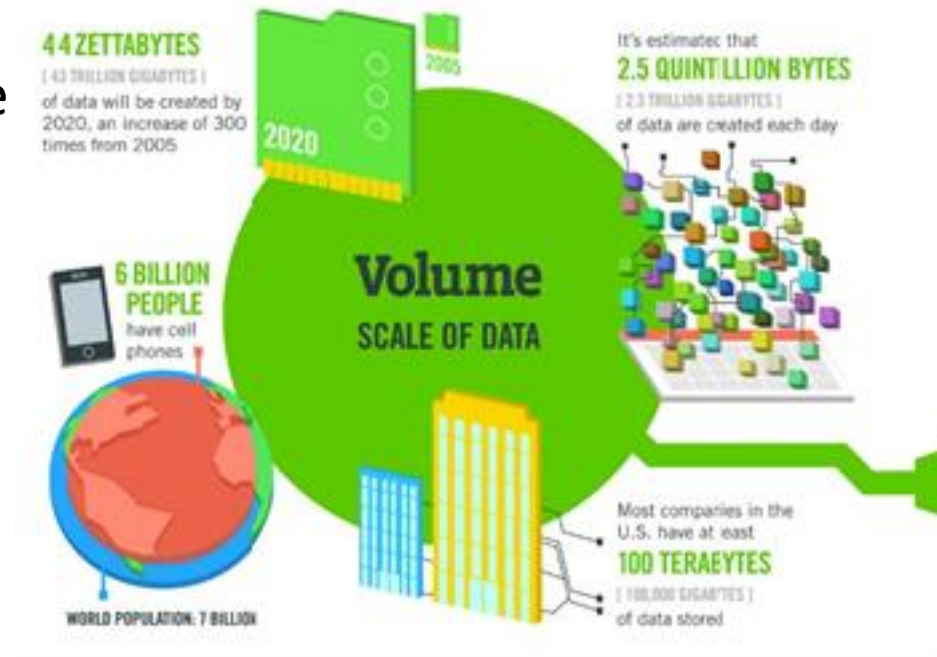
Dimensions of Big Data?

- There are many examples of "data", but what makes some of it "big"? The classic definition revolves around the **four Vs**.

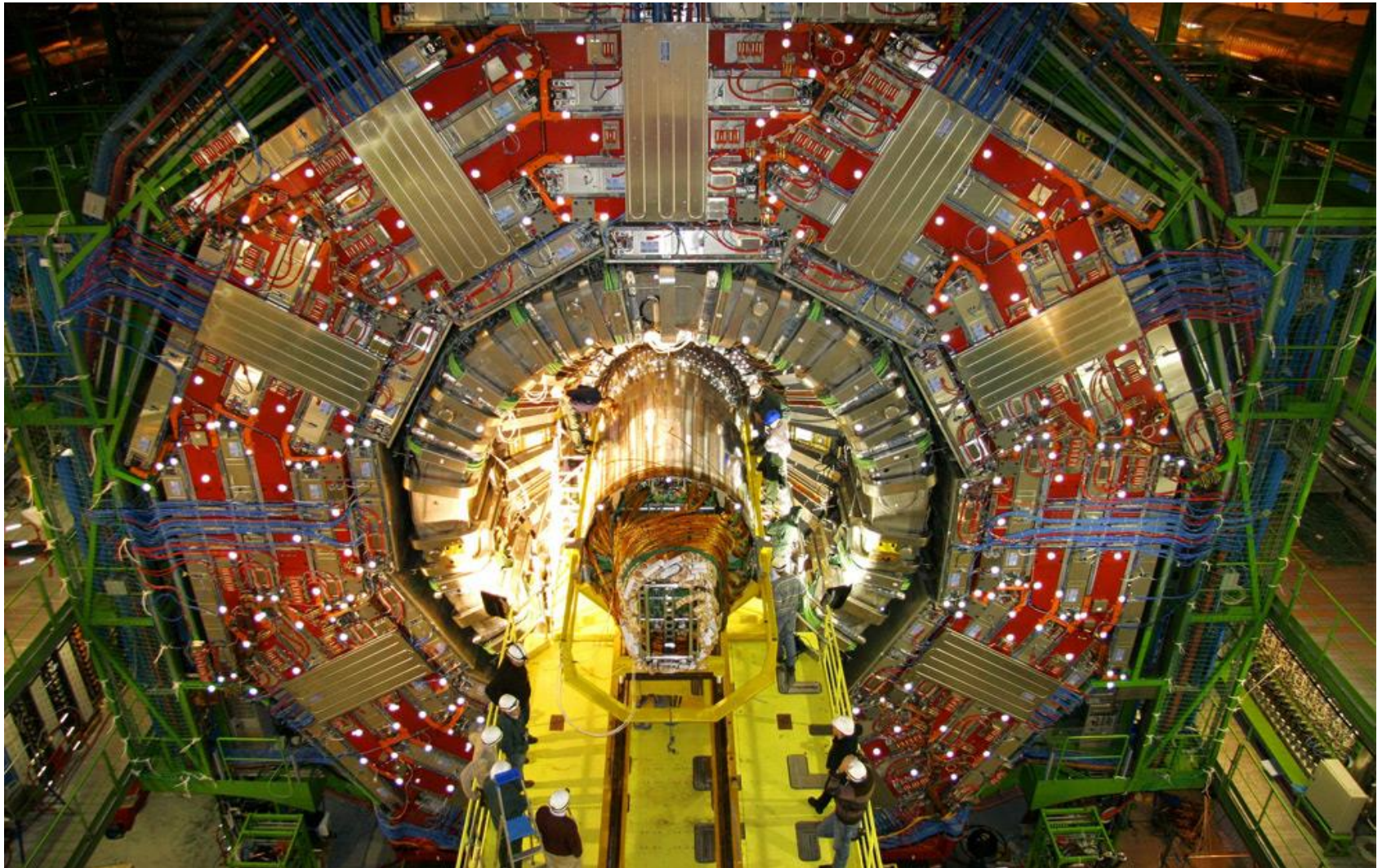


Dimensions of Big Data: Volume (scale)

- Over 2.5 quintillion bytes (exabytes or 10^{18} bytes) of data are generated every day!
- Most companies in the US have at least 100 terabytes of data
- Data come from many sources
 - Social media sites
 - Sensors
 - Digital photos
 - Business transactions
 - Location-based data



Growing data volume: sources of data

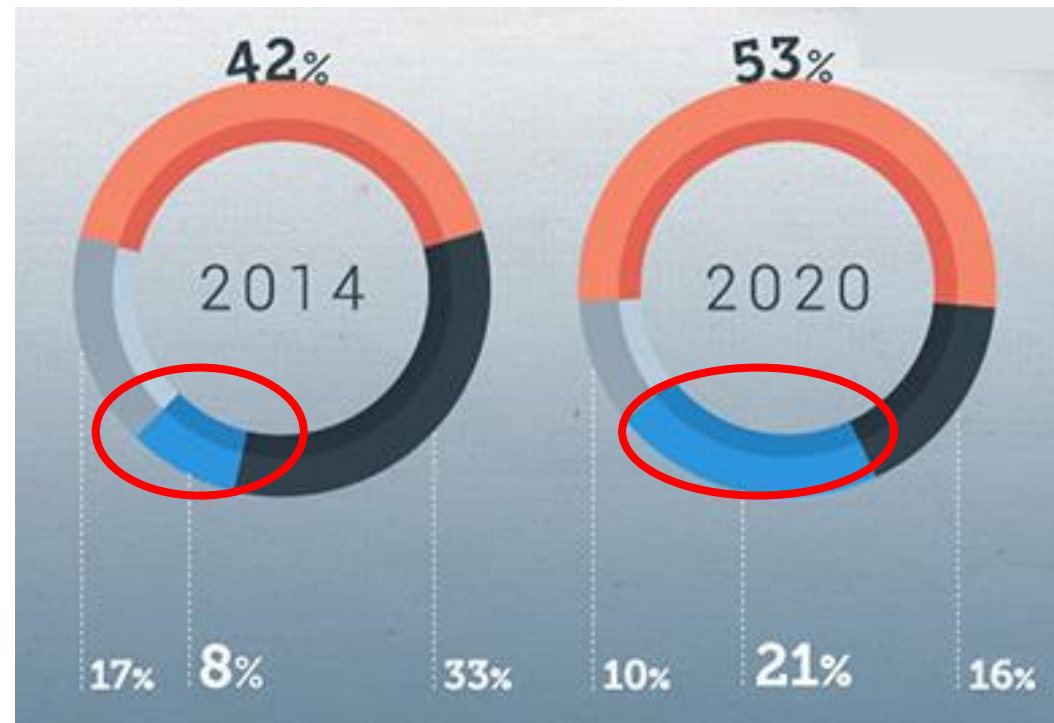
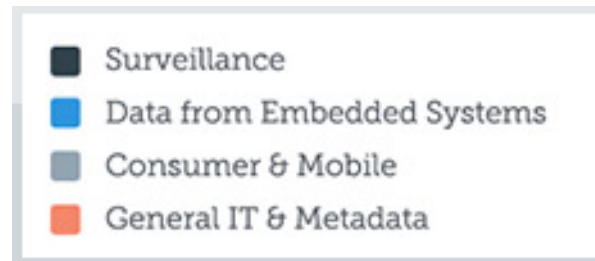


CERN's Large Hadron Collider (LHC) generates 15 PB a year

Growing data volume: sources of data

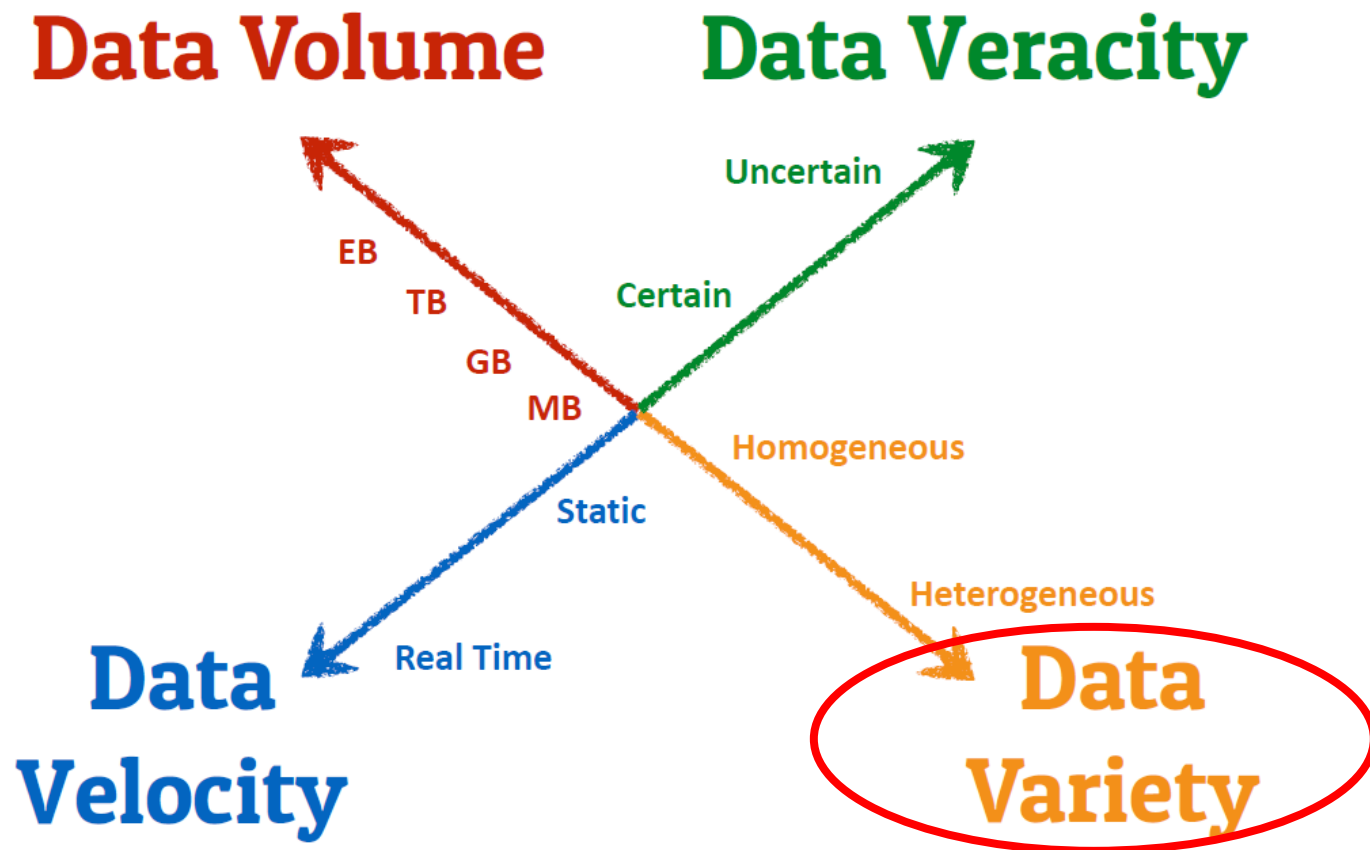
- Major sources of data
 - Surveillance
 - General IT and Metadata

- By 2020
 - Data from embedded systems (IoT) represents a larger %



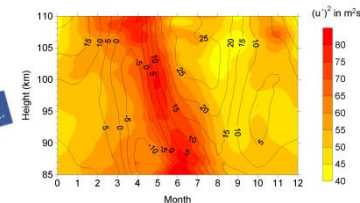
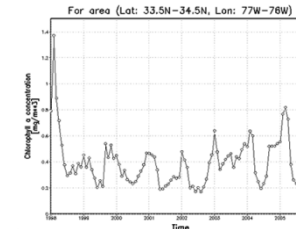
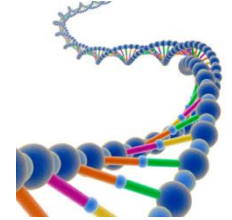
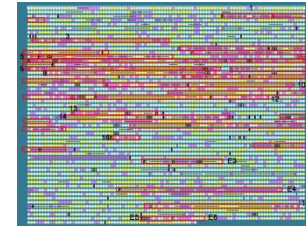
Dimensions of Big Data?

- There are many examples of "data", but what makes some of it "big"? The classic definition revolves around the **four Vs**.



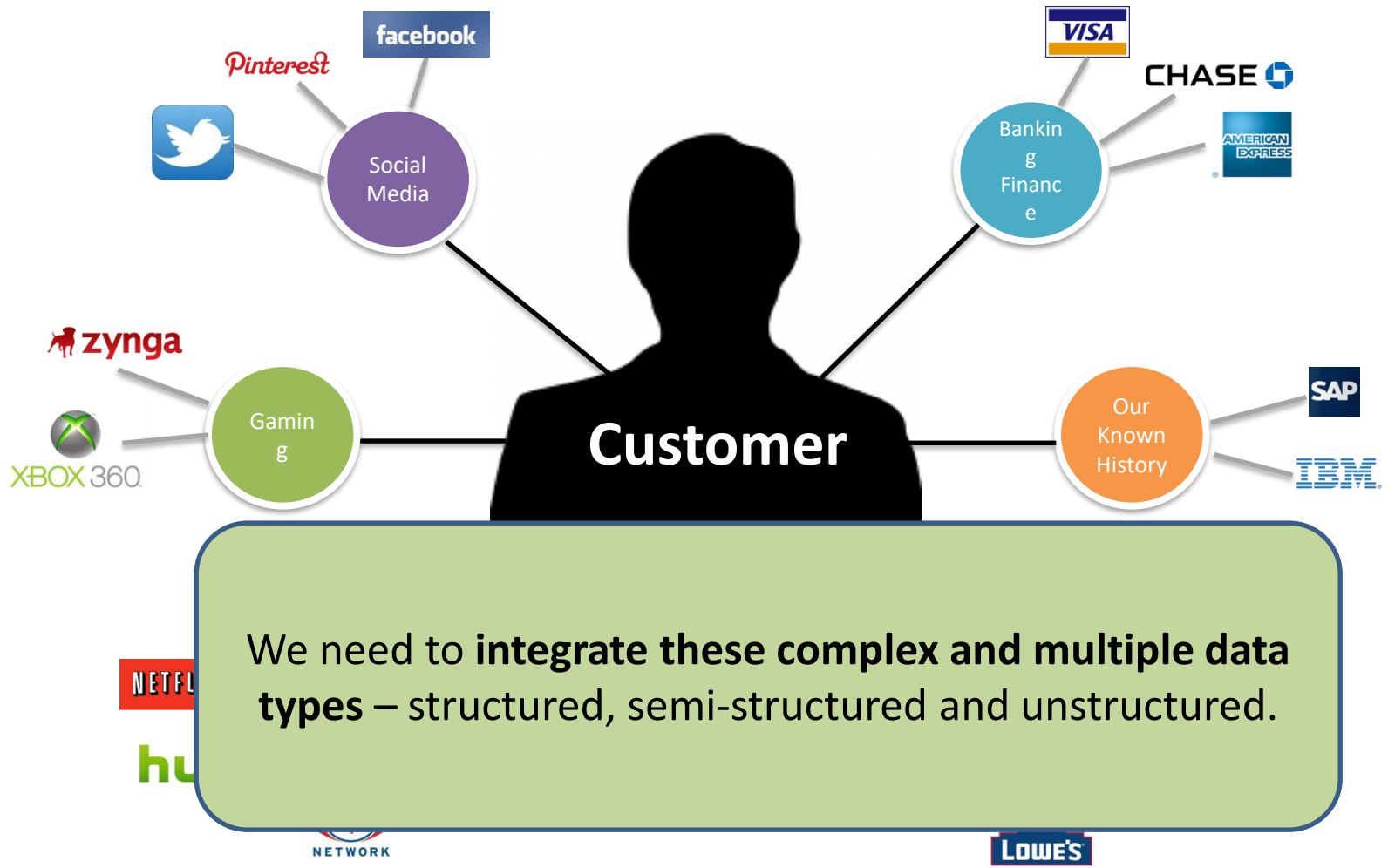
Dimensions of Big Data: Variety (complexity)

- Relational Tables, Spread-sheets (**Structured** Data)
- Text Data, E-mail, Web (**Un-structured** Data)
- XML (**Semi-structured** Data)
- Graph Data
 - Social Network, Semantic Web (RDF), ...
- Multimedia content
 - photos, audio, videos
- Streaming Data
 - You can only scan the data once
- A single application can be generating/collecting many types of data



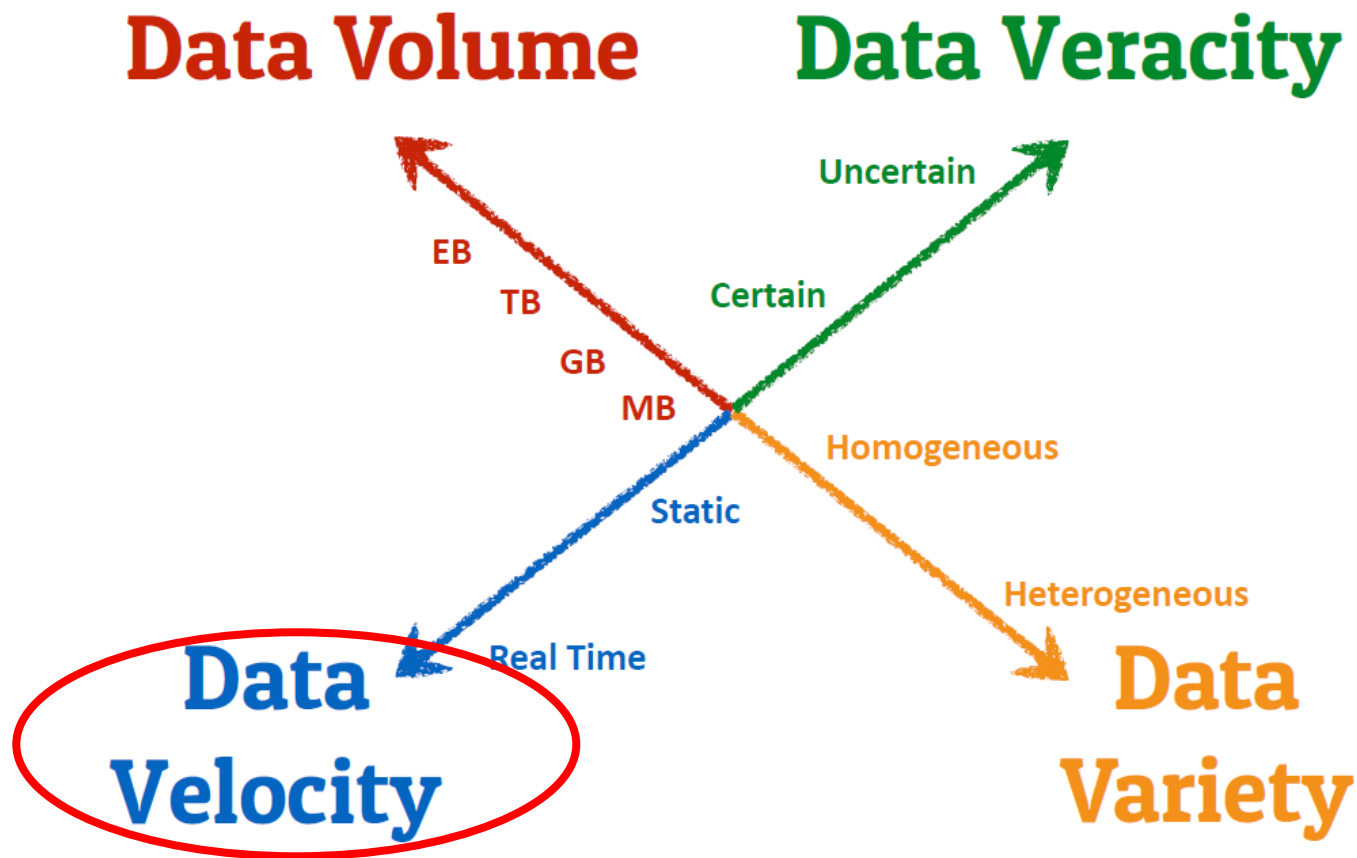
To extract knowledge → all these types of data need to be linked together

A Single View to the Customer



Dimensions of Big Data?

- There are many examples of "data", but what makes some of it "big"? The classic definition revolves around the **four Vs**.



Real-time/Fast Data

Sources of fast data: business processes, machines, networks and human interaction with social media sites, mobile devices, etc.



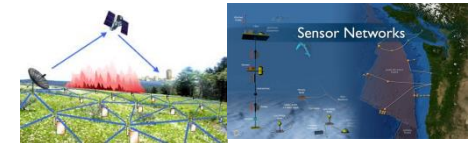
Social media and networks
(all of us are generating data)



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

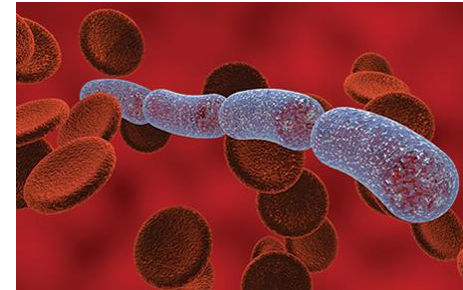
Real-Time Analytics/Decision Requirement

- Data is begin generated fast and need to be processed fast
 - Flow is continuous
- Online Data Analytics → strategic competitive advantages & ROI
- 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



Real-Time Analytics case study: continuous patient monitoring

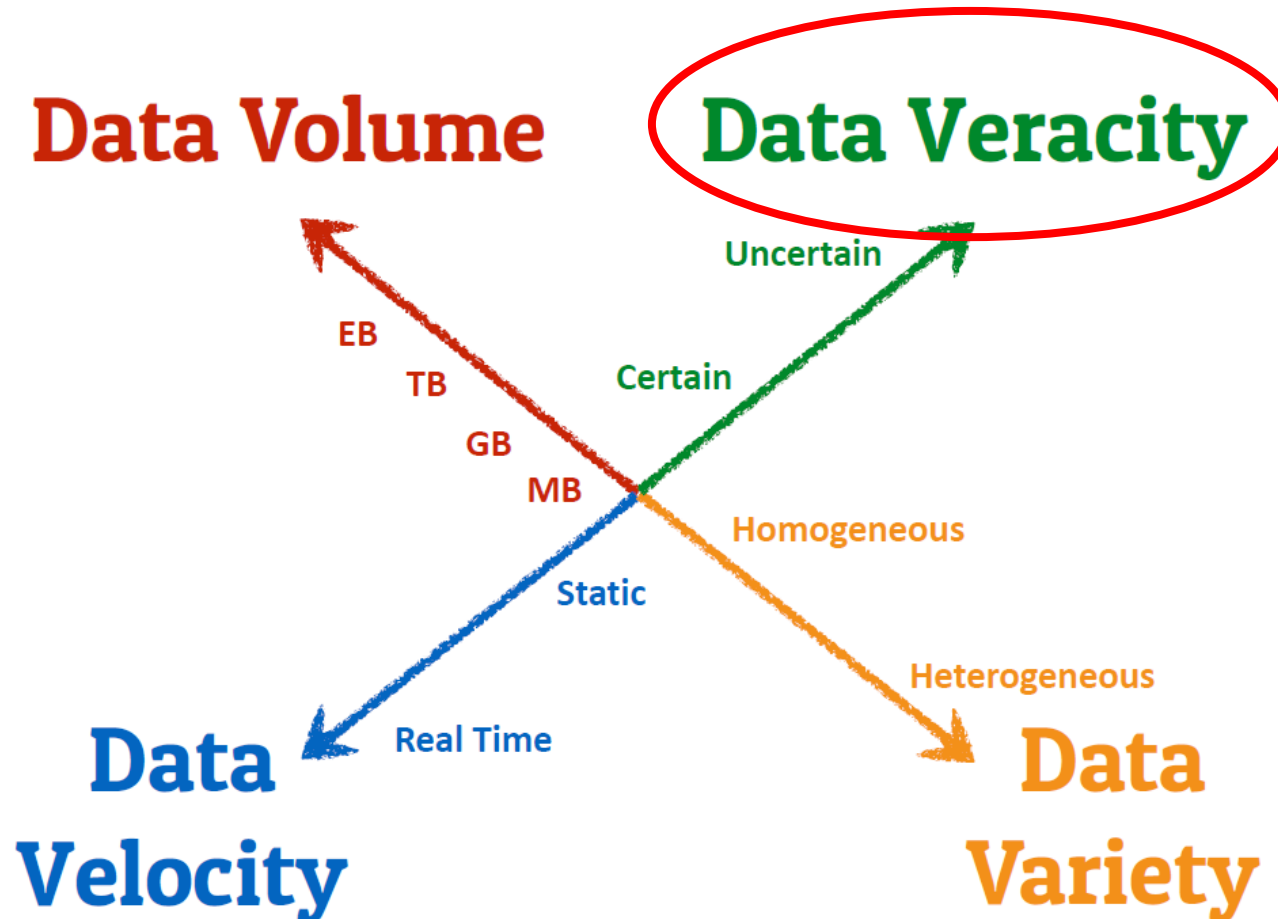
- Sepsis: over a million people each year are hospitalized
 - Nearly half of those cases result from infections in the lungs.
 - Mortality rate as great as 50%



- Mortality is significantly reduced if septic patients are identified and treated at early stages of the disease process.
- Real time monitoring of heart rate variability (HRV)
 - Researchers identified a distinctive V-shaped temporal pattern in HRV series before septic shock was occurred.

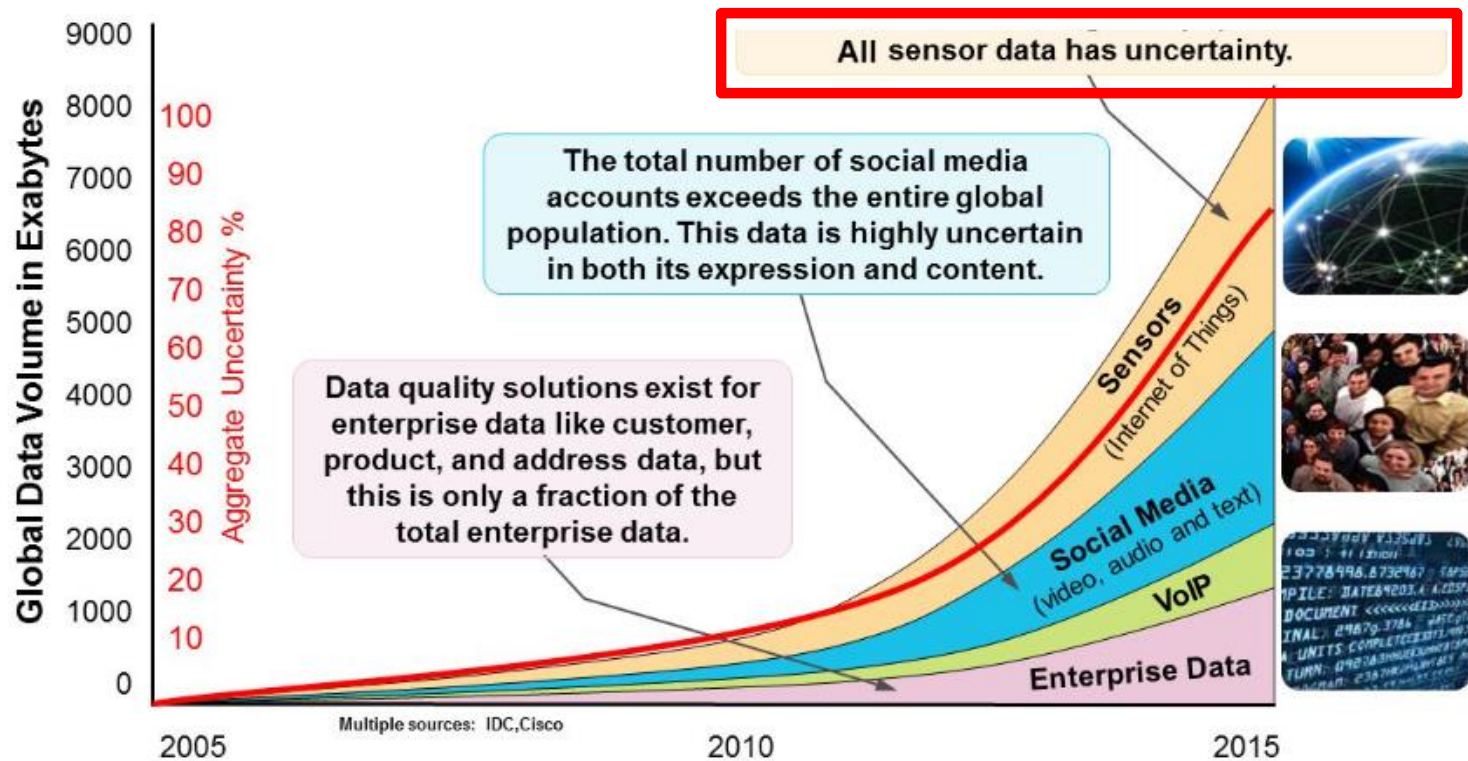
What is Big Data?

- There are many examples of "data", but what makes some of it "big"? The classic definition revolves around the **four Vs**.



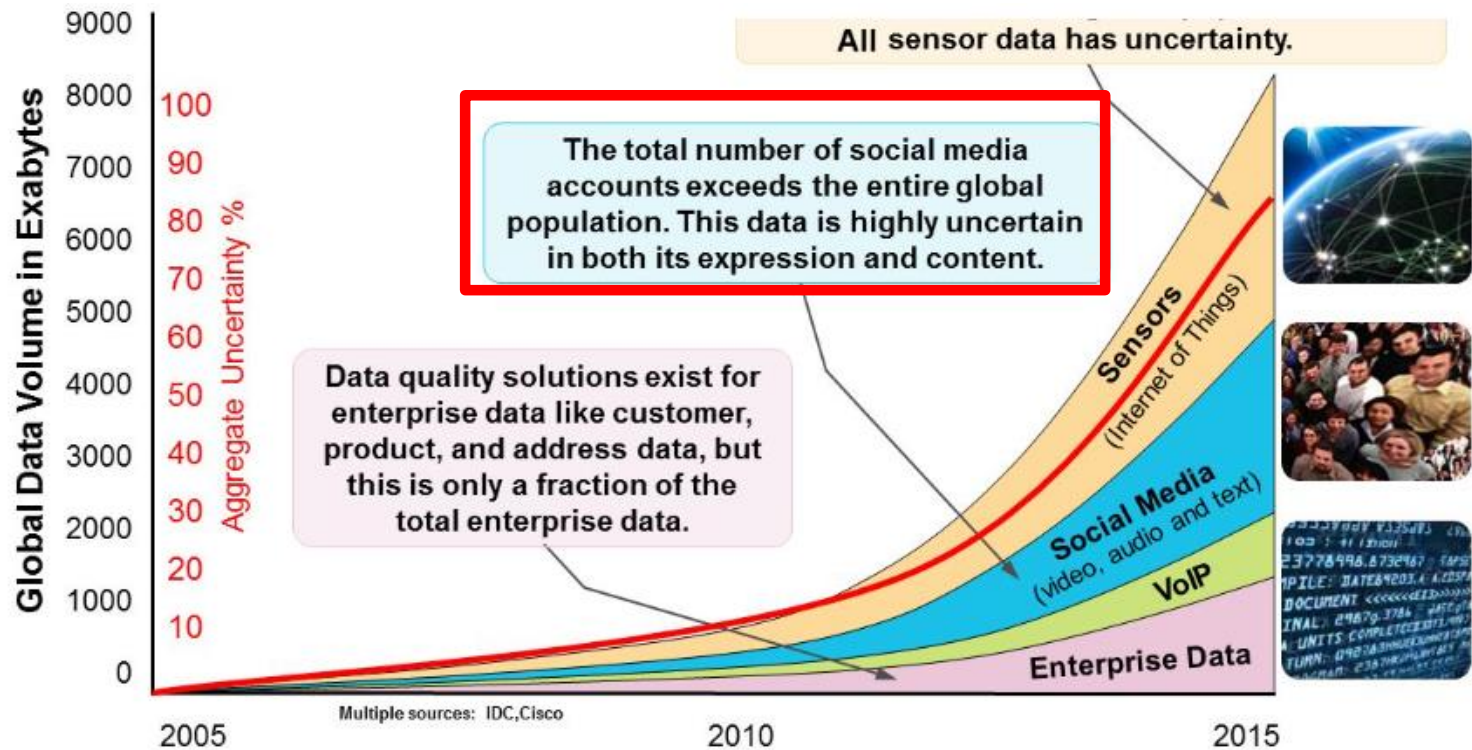
Dimensions of Big Data: Veracity (uncertainty)

- Uncertainty due to
 - Inconsistency, incompleteness, ambiguities and model approximations
- Since 2015, over 80% of all data is uncertain



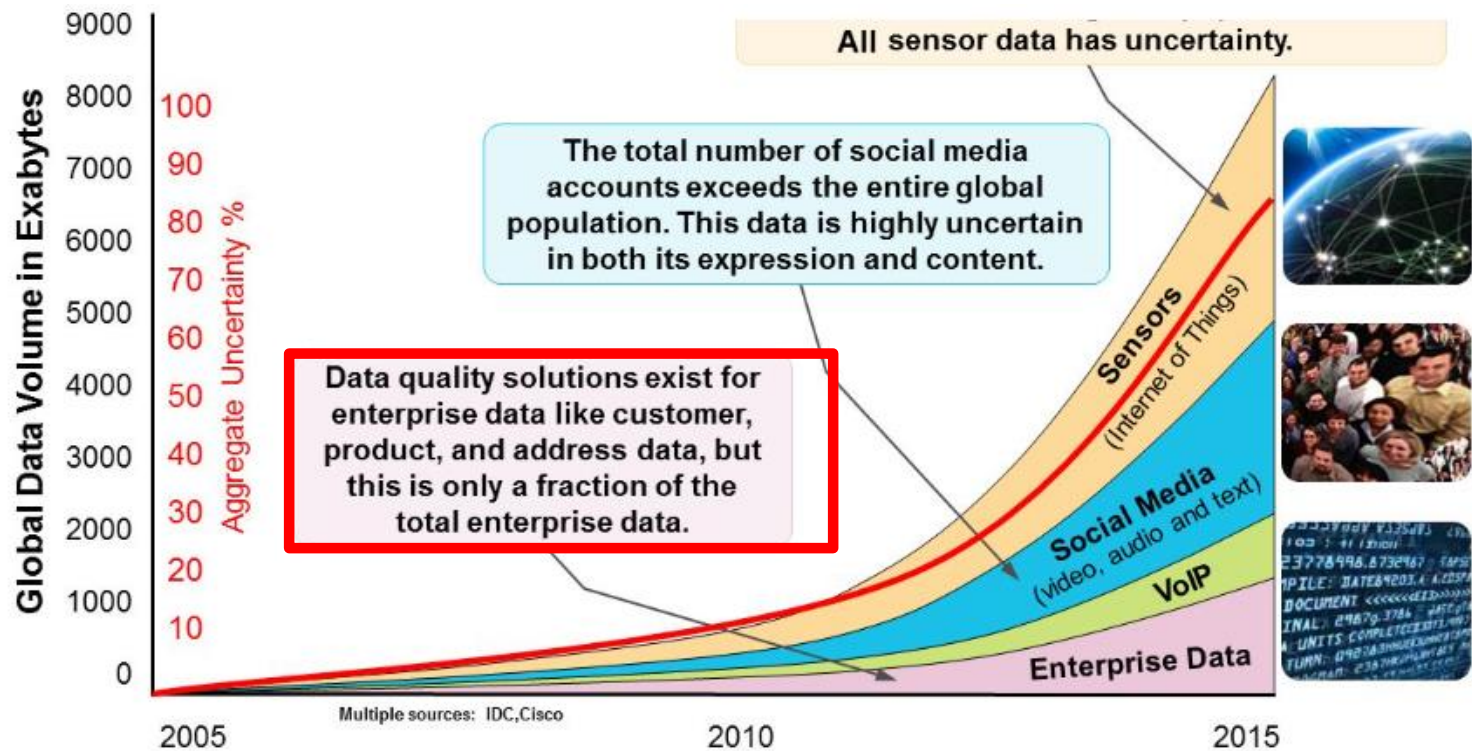
Dimensions of Big Data: Veracity (uncertainty)

- Uncertainty due to
 - Inconsistency, incompleteness, ambiguities and model approximations
- Since 2015, over 80% of all data is uncertain

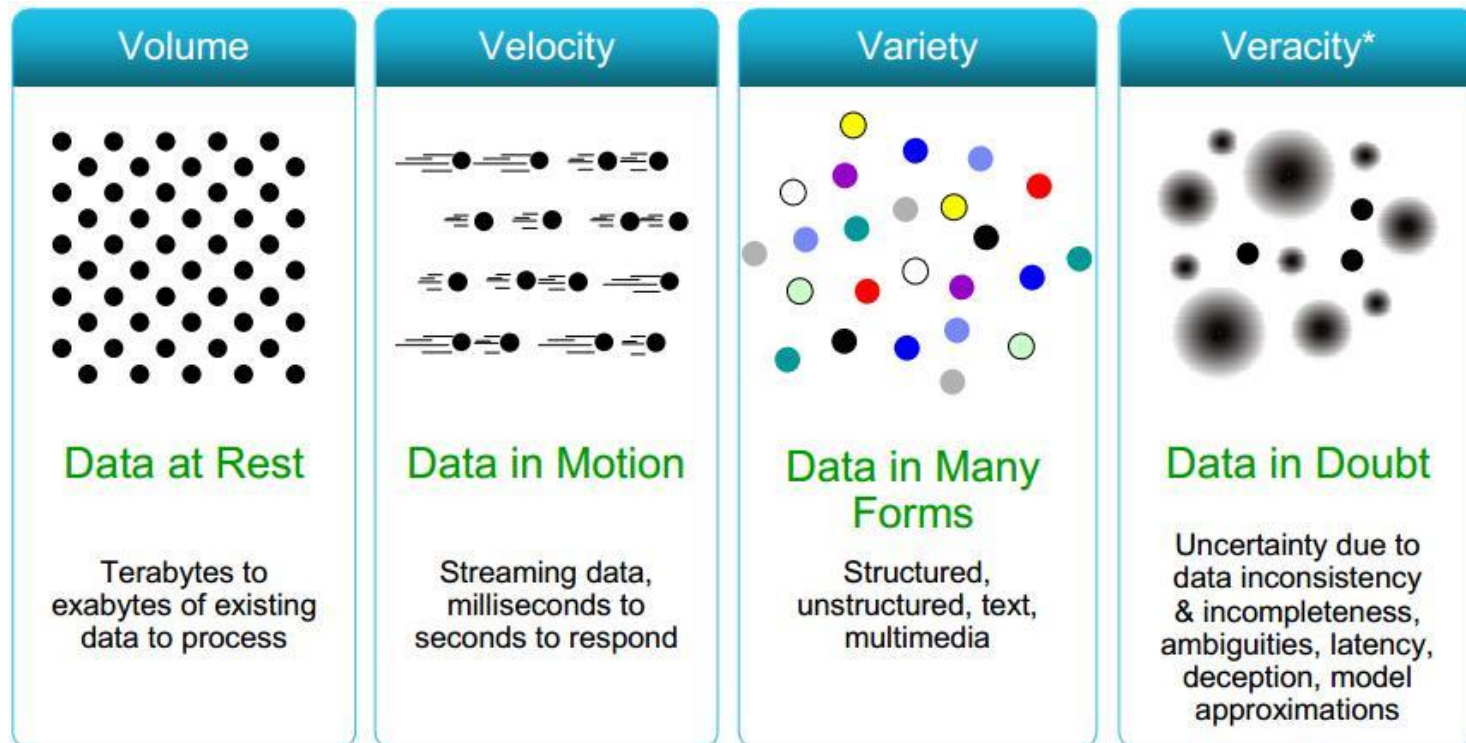


Dimensions of Big Data: Veracity (uncertainty)


- Uncertainty due to
 - Inconsistency, incompleteness, ambiguities and model approximations
- Since 2015, over 80% of all data is uncertain



Summary of the 4Vs

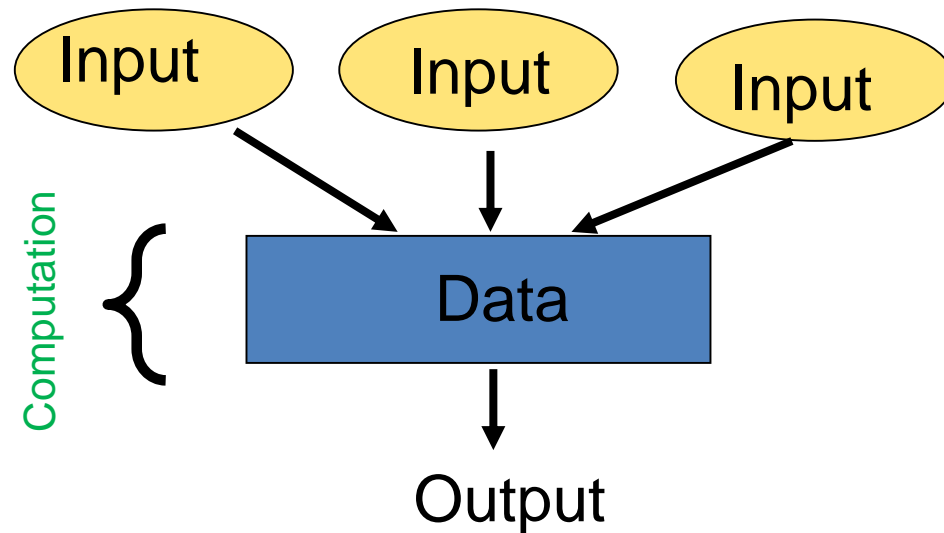


Outline

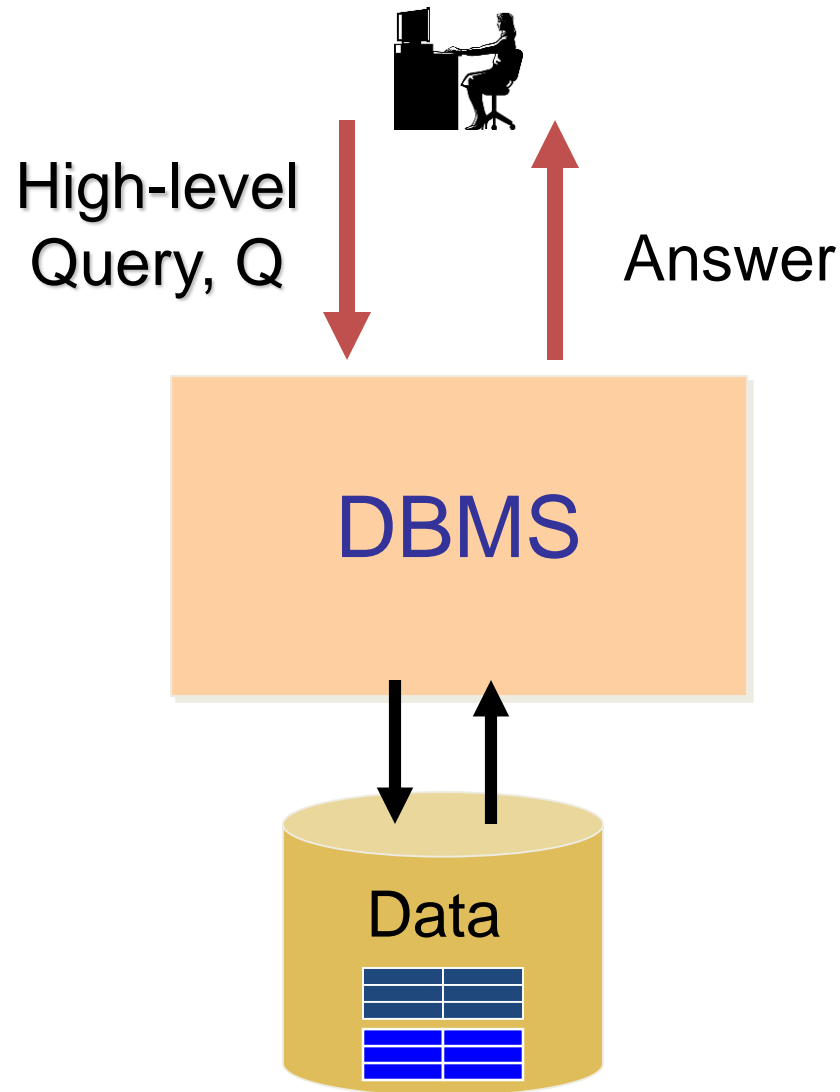
- What is Big Data
- Big Data dimensions
- Data Systems primer 
- Origin of “Big” Data Systems
- Why Big Data Systems
- Types of Big Data Systems
- Course logistics

Data Systems

- For a set of inputs produces a set of outputs by performing computations on a set of data
 - Scientific data processing systems
 - Information systems or DBMS
 - Data analysis systems



DBMS: the most popular Data System



Example of a Data System: At a Company

Query 1: Is there an employee named “Nemo”?

Query 2: What is “Nemo’s” salary?

Query 3: How many departments are there in the company?

Query 4: What is the name of “Nemo’s” department?

Query 5: How many employees are there in the
“Accounts” department?

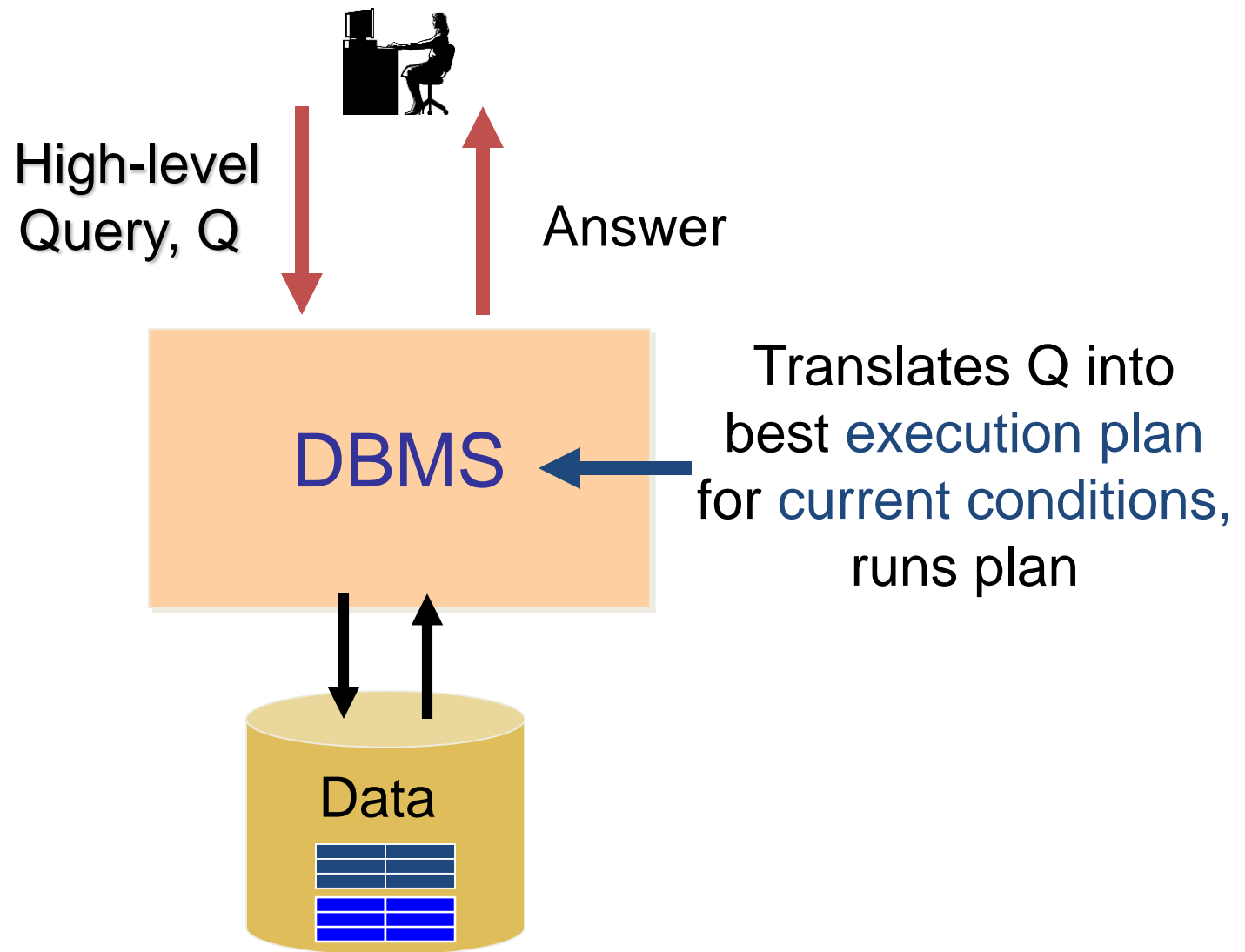
Employee

ID	Name	DeptID	Salary	...
10	Nemo	12	120K	...
20	Dory	156	79K	...
40	Gill	89	76K	...
52	Ray	34	85K	...
...

Department

ID	Name	...
12	IT	...
34	Accounts	...
89	HR	...
156	Marketing	...
...

DBMS: the most popular Data System



Example: Store that Sells Cars

Owners of
Honda Accords
who are \leq
23 years old

Make	Model	OwnerID	ID	Name	Age
Honda	Accord	12	12	Nemo	22
Honda	Accord	156	156	Dory	21

Join (Cars.OwnerID = Owners.ID)

Filter (Make = Honda and
Model = Accord)

Filter (Age \leq 23)

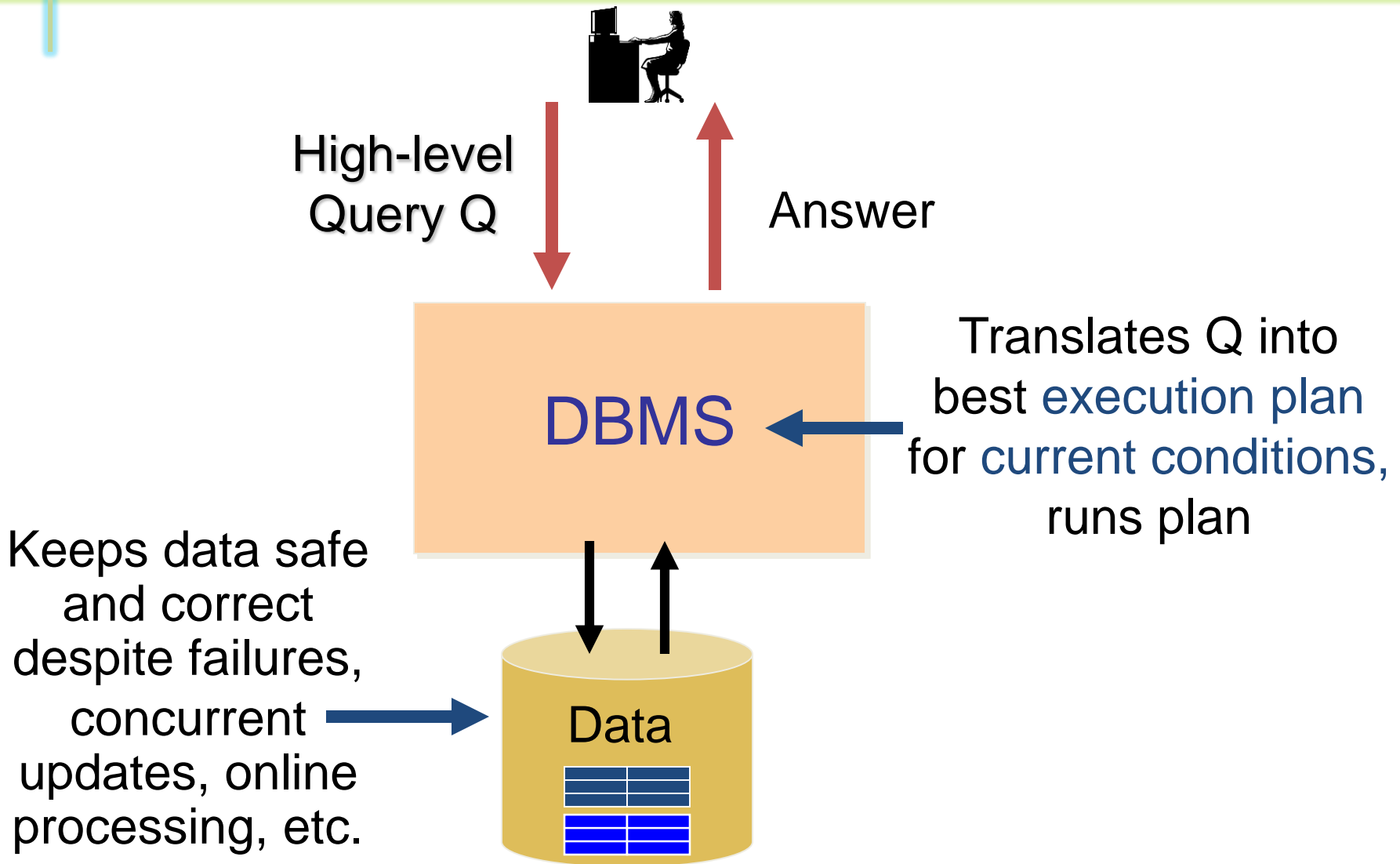
Cars

Make	Model	OwnerID
Honda	Accord	12
Toyota	Camry	34
Mini	Cooper	89
Honda	Accord	156
...

Owners

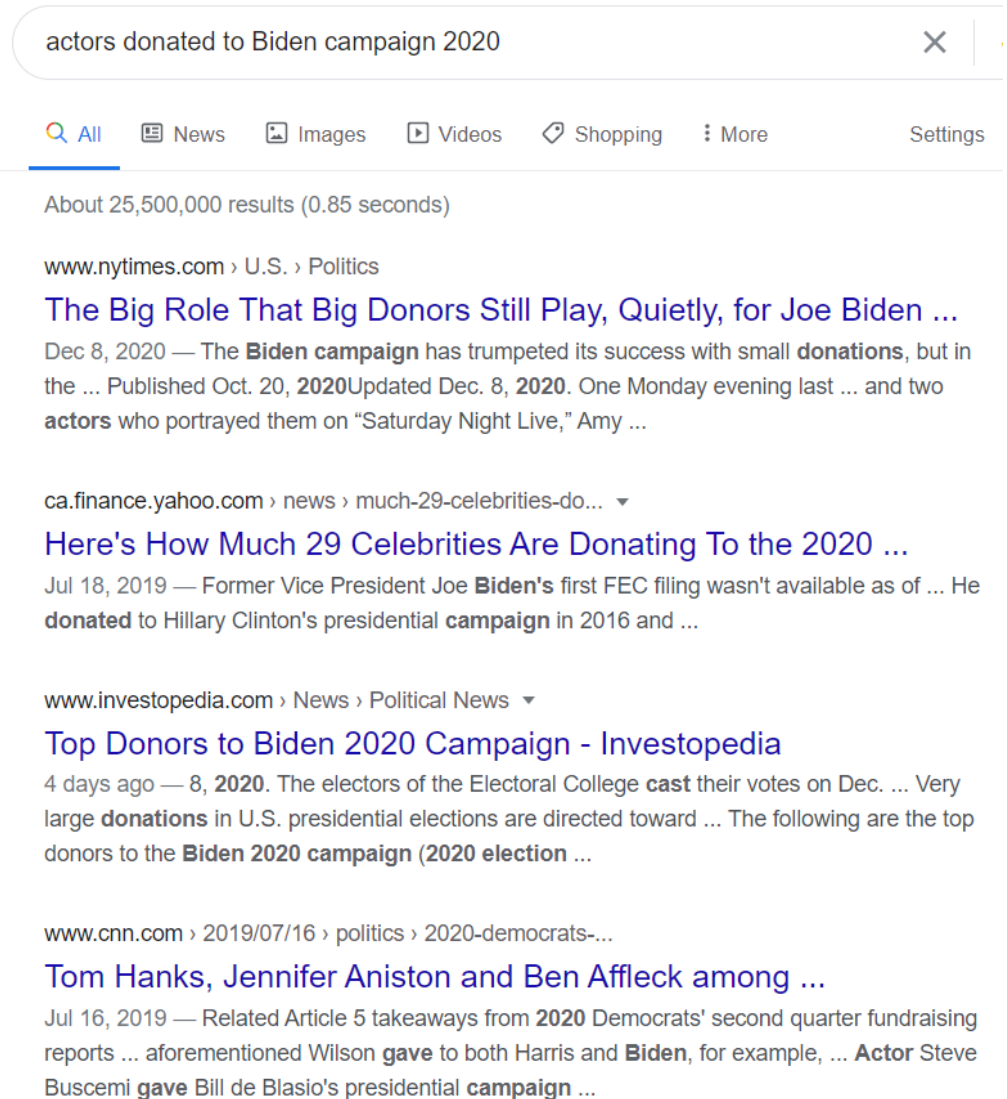
ID	Name	Age
12	Nemo	22
34	Ray	42
89	Gill	36
156	Dory	21
...

DBMS: the most popular Data System



“Search” vs. Query

- What if you wanted to find out which actors donated to Biden’s presidential campaign?
- Try “actors donated to Biden campaign 2020” in your favorite search engine.



actors donated to Biden campaign 2020

All News Images Videos Shopping More Settings

About 25,500,000 results (0.85 seconds)

www.nytimes.com › U.S. › Politics
The Big Role That Big Donors Still Play, Quietly, for Joe Biden ...
Dec 8, 2020 — The **Biden campaign** has trumpeted its success with small **donations**, but in the ... Published Oct. 20, 2020Updated Dec. 8, 2020. One Monday evening last ... and two **actors** who portrayed them on “Saturday Night Live,” Amy ...

ca.finance.yahoo.com › news › much-29-celebrities-do...
Here's How Much 29 Celebrities Are Donating To the 2020 ...
Jul 18, 2019 — Former Vice President Joe **Biden's** first FEC filing wasn't available as of ... He **donated** to Hillary Clinton's presidential **campaign** in 2016 and ...

www.investopedia.com › News › Political News
Top Donors to Biden 2020 Campaign - Investopedia
4 days ago — 8, 2020. The electors of the Electoral College **cast** their votes on Dec. ... Very large **donations** in U.S. presidential elections are directed toward ... The following are the top donors to the **Biden 2020 campaign** (2020 election ...

www.cnn.com › 2019/07/16 › politics › 2020-democrats-...
Tom Hanks, Jennifer Aniston and Ben Affleck among ...
Jul 16, 2019 — Related Article 5 takeaways from 2020 Democrats' second quarter fundraising reports ... aforementioned Wilson **gave** to both Harris and **Biden**, for example, ... **Actor** Steve Buscemi **gave** Bill de Blasio's presidential **campaign** ...

“Search” vs. Query

- “Search” can return only **what’s been “stored”**
- E.g., best match Google, Bing top ten

Tom Hanks, Jennifer Aniston and Ben Affleck among celebrities donating to 2020 Democrats



By David Wright, CNN
Published 5:29 PM EDT, Tue July 16, 2019




PHOTO: Getty

(CNN) — Hollywood has long been fertile fundraising ground for Democratic presidential candidates, and the [second quarter](#) of 2019 was no exception. Famous names were littered throughout [2020 Democrats'](#) Federal Election Commission reports.



[Pete Buttigieg](#), the overall leader in second quarter fundraising among Democrats, was a favorite of celebrity donors. Actress [Gwyneth Paltrow](#) hosted a fundraiser for Buttigieg at her home in Southern California in early May, with actor [Bradley Whitford](#) co-hosting alongside a number of famous attendees.

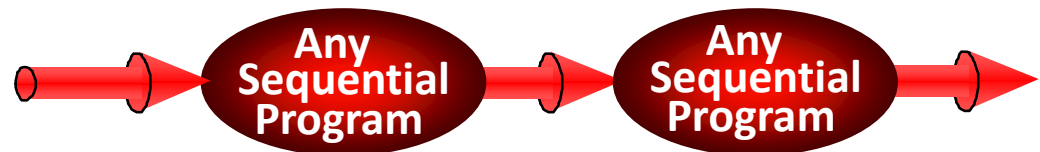
Outline

- What is Big Data
- Big Data dimensions
- Data Systems primer
- Origin of “Big” Data Systems 
- Why Big Data Systems
- Types of Big Data Systems
- Course logistics

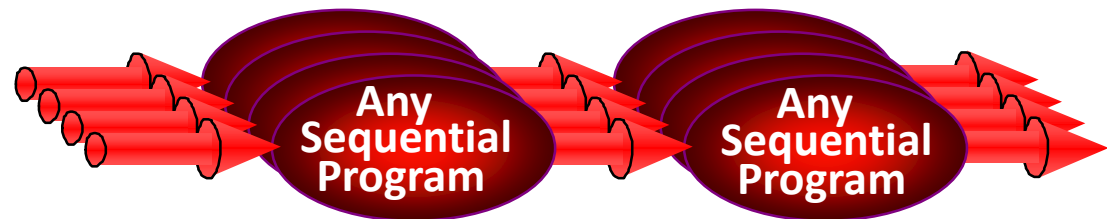
Parallel DBMS: *the original* big data systems

- Parallelism natural to DBMS processing
 - **Pipeline parallelism:** many processors, each doing one step in a multi-step process.
 - **Partition parallelism:** many processors doing the same thing to different pieces of data.
 - **Both are natural in DBMS!**

Pipeline



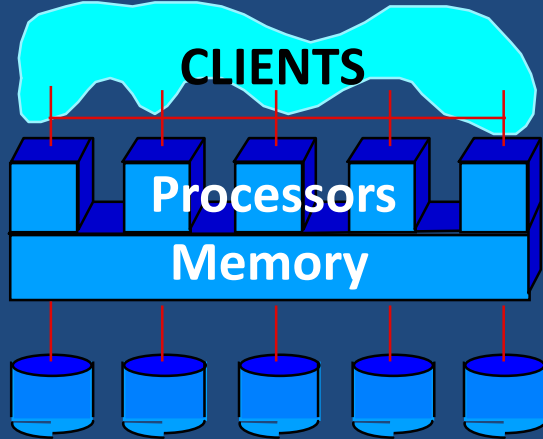
Partition



outputs split N ways, inputs merge M ways

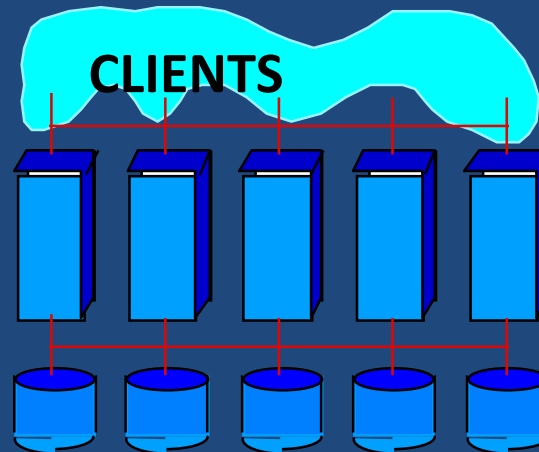
Parallel Databases: architecture

Shared Memory (SMP)

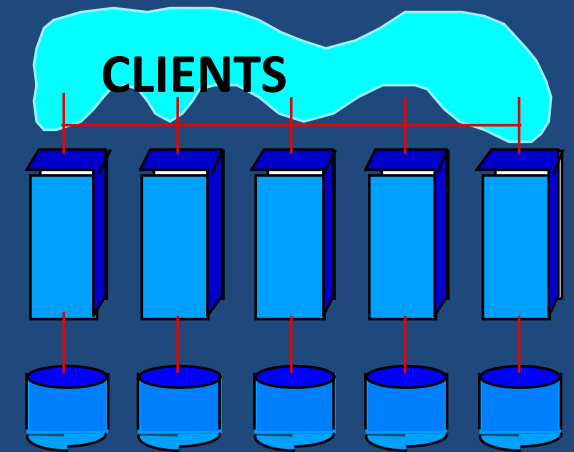


Easy to program
Expensive to build
Difficult to scaleup

Shared Disk



Shared Nothing (network)



Hard to program
Cheap to build
Easy to scaleup

Search engines in 1990s

MetaCrawler Parallel Web Search Service
by [Erik Selberg](#) and [Oren Etzioni](#)

Try the new [MetaCrawler Beta!](#)
If you're searching for a person's home page, try [Abow!](#)

[Examples](#) • [Beta Site](#) • [Add Site](#) • [About](#)

Search for:

☐ as a Phrase ☒ All of these words ☐ Any of these words

For better results, please specify:
Search Region: Search Sites:

Performance parameters:
Max wait: minutes Match type:

[[About](#) | [Help](#) | [Problems](#) | [Add Site](#) | [Search](#)]
webmaster@metacrawler.com
© Copyright 1995, 1996 Erik Selberg and Oren Etzioni

1996

excite [search](#) [reviews](#) [city.net](#) [new live!](#) [reference?](#)

[excite home](#) [maps](#) [news](#) [people finder](#)

Excite Search: twice the power of the competition.

What: [\[search\]](#)

Where: [\[Help\]](#) [\[Add URL\]](#)

Researching stocks?
Buying a car?
Planning a wedding?
[Check out ExciteSeeing Tours.](#)

[Bill Mitchell:](#)
[Satire that clicks!](#)

Excite Reviews: site reviews by the web's best editorial team.

Arts	Entertainment	Money	Regional
Business	Health	News & Reference	Science
Computing	Hobbies	Personal Pages	Shopping
Education	Life & Style	Politics & Law	Sports

1996

LYCOS It's amazing where Go Get It will get you.

Find:

[Enhance your search.](#)

[New Search](#) • [Top News](#) • [Sites by Subject](#) • [Top 5% Sites](#) • [City Guide](#) • [Pictures & Sounds](#)
[PeopleFind](#) • [Point Review](#) • [Road Maps](#) • [Software](#) • [About Lycos](#) • [Club Lycos](#) • [Help](#)

[Add Your Site to Lycos](#)

Copyright © 1996 Lycos™, Inc. All Rights Reserved.
Lycos is a trademark of Carnegie Mellon University.
[Questions & Comments](#)

1994

Wired Search Center

look for: [\[SEARCH\]](#)

for more options use [SuperSearch](#)

Date:

Country:

Include media type:
☐ Image ☐ Audio ☐ Video ☐ Download

Return Results:

[CLEAR FORM](#)

[Sandbox Entertainment](#)
[Shop WIRED Holiday Gift Guide](#)
[SOMETHING HAS SURVIVED.](#)
[Find more deals](#)
[Gap](#)
[Cybernet Outpost](#)
[Microsoft® Expedia™ Travel](#)
[ON SALE](#)

1997

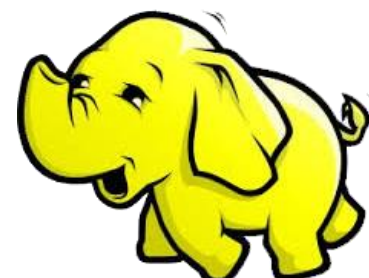
Google search engines



1998



2013





Is a DBMS good for Web Search?

- Large volumes of data
 - Number of web sites
 - In 1994: only 3000
 - In 2015: 935,951,027
 - 33 million percent increase in 20 years!!
 - 672 Exabytes of accessible data (2013)

Why didn't Google use Oracle for their search engine?

- In 1999, Google was fielding 3 million search queries per day
- In 2018, Google was serving more than 5.3 billion searches per day (3.7 million/minute)
- 48 Billion webpages indexed by Google
- Over 9,00,000 Servers - owned by Google, the Largest in the world
- Unstructured data
 - Webpages have no fixed schema
 - New standards like XHTML, HTML5...

Origins of “Modern” Big Data Systems

- Google's developed its own data systems infrastructure for **scalability and better control** of performance characteristics.

2003

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google*



2004

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.



2006

Bigtable: A Distributed Storage System for Structured Data

Fuy Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber

{fuy,jeff,sanjay,wilson,hsieh,deborah,wallach,mike,tushar,andy,robert}@google.com

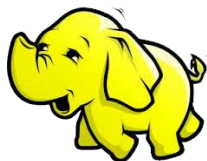
Google, Inc.



Abstract

Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large number of servers. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google Fused Location Platform. These applications place very different demands on Bigtable, both in terms of data size (from URLs to

achieved scalability and high performance, but Bigtable provides a different interface than such systems. Bigtable does not support a full relational data model; instead, it provides clients with a simple data model that supports dynamic control over data layout and format, and allows clients to reason about the locality properties of data represented in the underlying storage. Data is indexed using row and column names that can be arbitrary strings. Bigtable also treats data as uninterpreted strings.



Outline

- What is Big Data
- Big Data dimensions
- Data Systems primer
- Origin of “Big” Data Systems
- Why Big Data Systems
- Types of Big Data Systems
- Course logistics



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



Motivations behind Big Data Systems

- Data volume, different application domains
- Scalability and performance
- Graceful failure recovery
- Data format, manageability

Motivation1: volume; one size does not fit all

- OLTP
 - Amazon : 42 TB
 - Typical OLTP databases: less than a TB
- OLAP/Data Warehouse
 - SAP : 12.1 PB (Guinness Record holder)
 - Ebay: 1.4 PB (2015)
- Search engines (text)
 - Google : 850 TB
 - Youtube: 76 PB of video data/year (2014)
- Scientific
 - US Department of Energy (NERSC): 3.5 PB
- New application domains
 - Stream processing for IOT data
 - Social media



Motivation2: need for scale and performance

- Scaling up
 - “Join-pain”
 - Issues with scaling up when the dataset is just too big
 - RDBMS were not designed to be distributed
 - Cost effective strategy: ‘scaling out’ or ‘horizontal scaling’
- Some applications need very few database features; But need high scalability when traffic spike happens
 - “Shashdot effect”
 - SQL may be too heavy-weight
 - Does not need fancy indexing.
 - Just fast lookup by primary key








IT World Prediction...



What ad sites will crash during the Super Bowl?

IT service predicting which ad sites are likely to go kaput during the big game

By **Keith Shaw**  Add a new comment  Like  8  +1  2

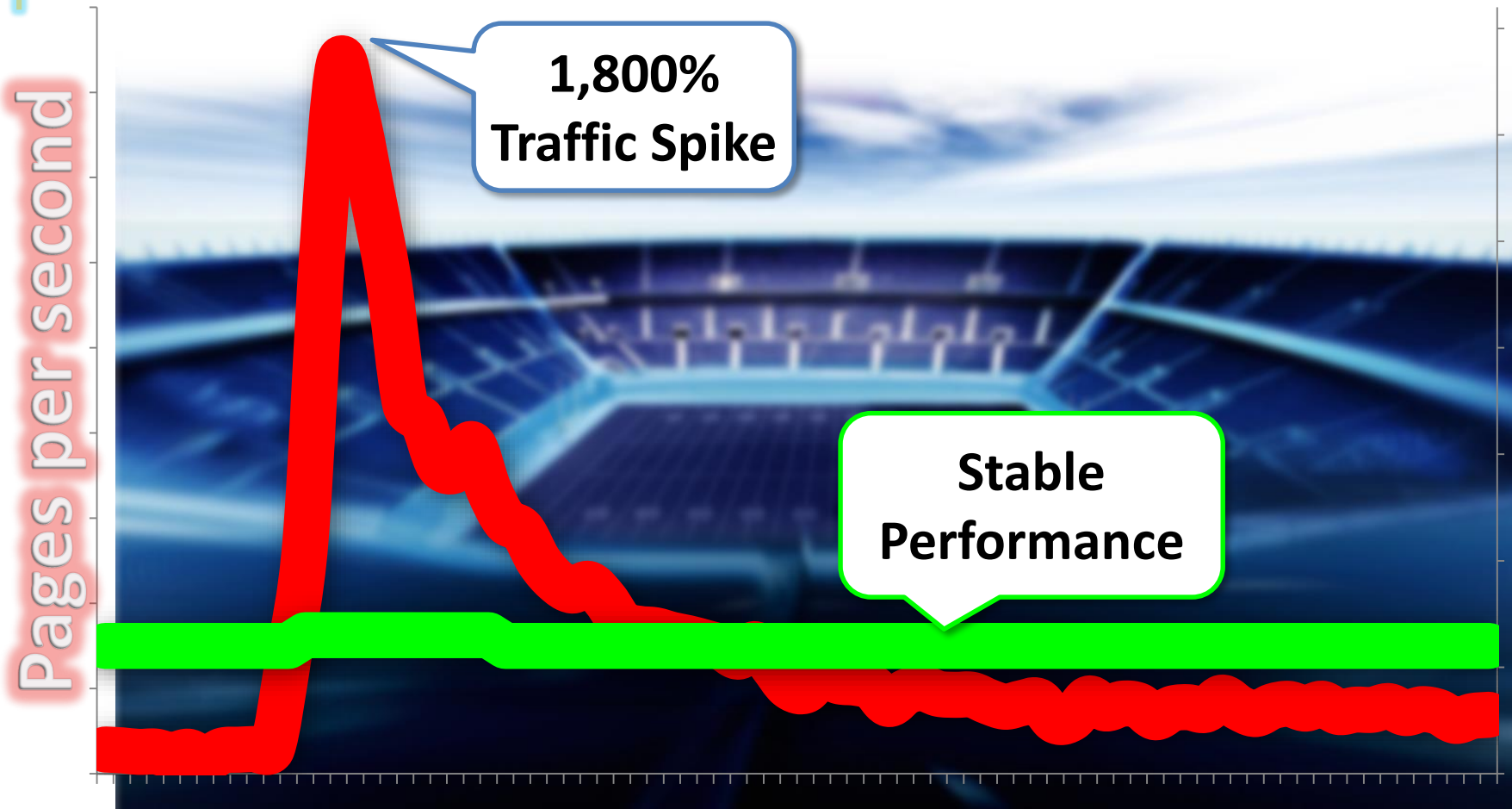
January 31, 2012, 3:55 PM — We get a lot of pitches from companies during the Super Bowl week, but this one intrigued us more than others - web site optimizer site [Yottaa](#) is predicting which Super Bowl advertiser sites are likely to crash or slow down significantly during the game.

Yottaa says there are many reasons why these sites will likely have performance issues, including "improperly integrated social media widgets, bloated images and the use of CDNs with variable performance histories". The company says it has reviewed the front-end performance (how many assets are loaded, size of Web pages, types of widgets used) of the Super Bowl advertiser's sites to make its predictions. The company says the top four sites that will crash during the Super Bowl include:

- #1: [Cars.com](#)
- #2: [GoDaddy.com](#)
- #3: [Taxact.com](#)
- #4: [History.com](#)

Source: <http://www.itworld.com/data-centerservers/246159/what-ad-sites-will-crash-during-super-bow>

Super Bowl traffic spike



**Commercial
Airs**

Motivation3: graceful failure recovery

- Dependence on Web services
 - We are addicted to Googling, Gmail, Google Map, Youtube, Facebook, Twitter, AWS...
- Graceful failure recovery
 - Need to continue to provide service
 - Cost of downtime



"You should check your e-mails more often. I fired you over three weeks ago."


The Cost of downtime

- The cost of unplanned outages has risen to of \$8,851 per minute
- The US recorded 3,526 outages which impacted 36.7 million people in 2017
- Amazon Web Services (AWS) US-East-1 region experienced 30 minutes downtime. (Jun 2018). Cost?
- Facebook was down for ~3 hours in Sep, 2010
 - \$1 million in lost ad revenues
- Rackspace was down due to power failure in Jun, 2009
 - was forced to pay ~\$3.5 million in service credits to customers
- Paypal was down due to network hardware failure in Aug, 2009
 - \$7.2 million in lost transactions in 4.5 hours
- Relational databases are not able to gracefully handle failures, especially with very large datasets. It may take hours to recover the after a disk failure

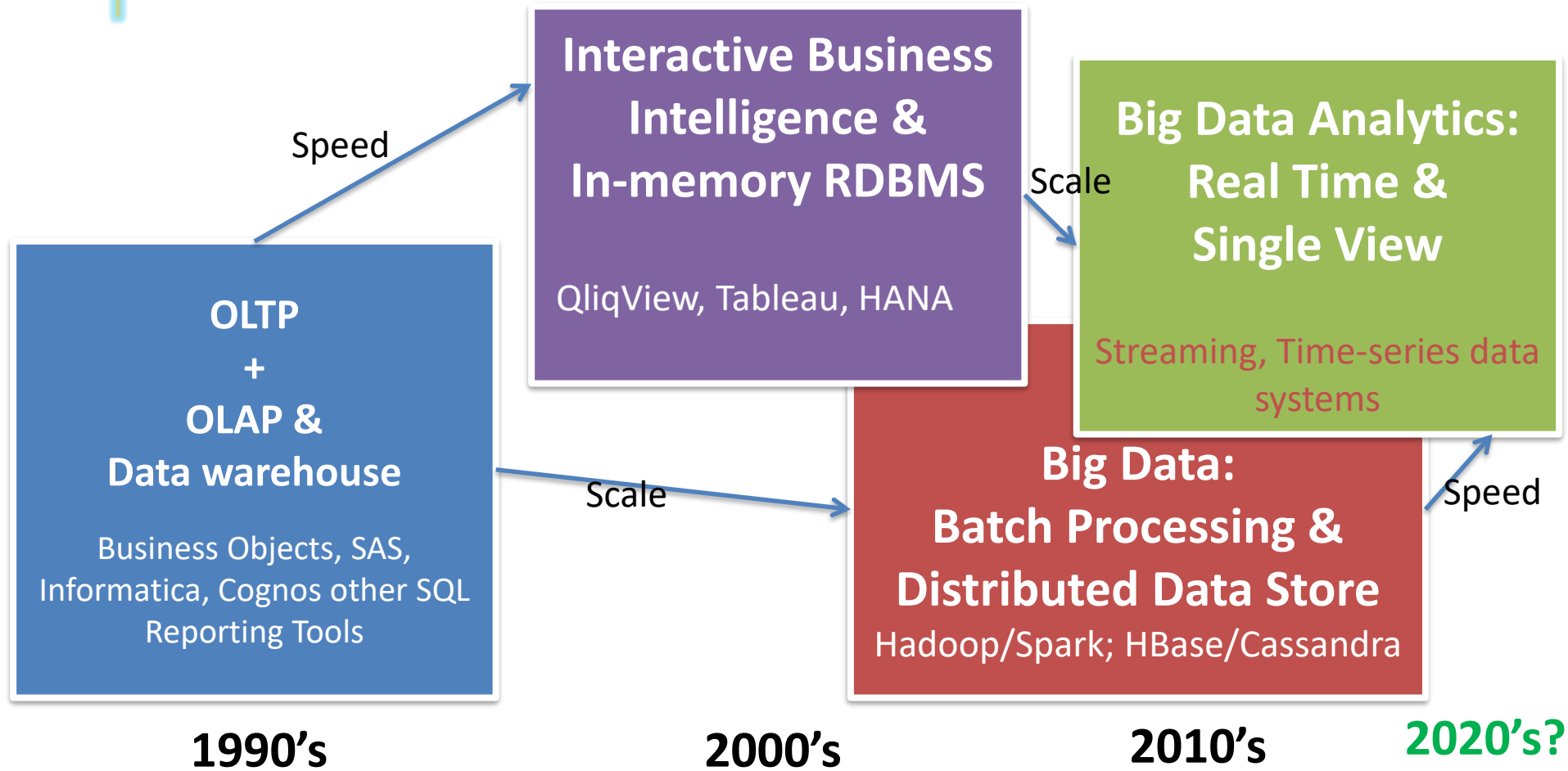
Motivation4: need for flexible schema

- Relational databases define the schema at **design time**
 - Rigid, no way to change dynamically...
 - Need a DBA
 - “Stop the world” to make any change
- Many applications don’t have any fixed schema
 - Log processing
 - Stream processing
 - Graph processing
- **Data model should not restrict data access**

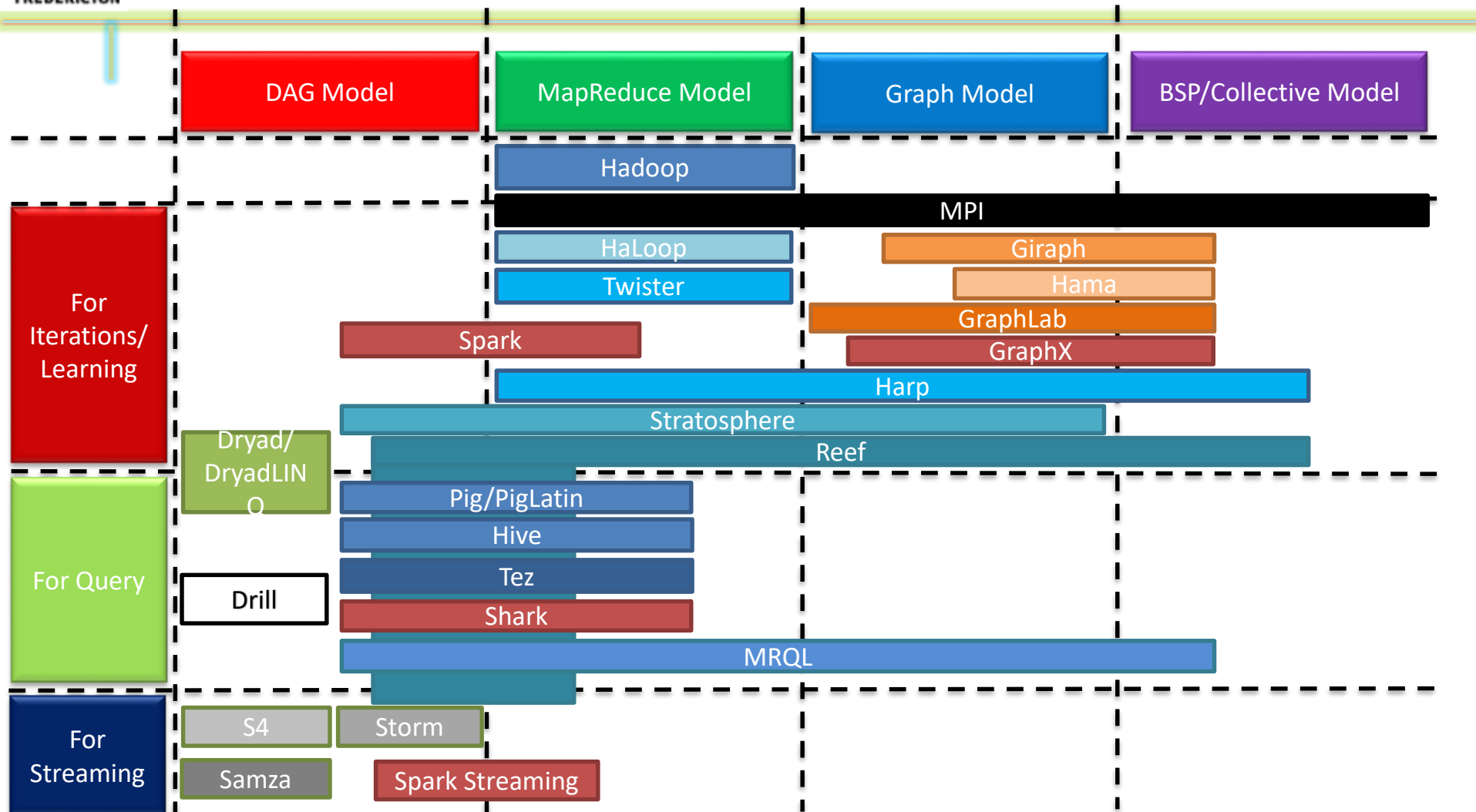
Outline

- What is Big Data
- Big Data dimensions
- Data Systems primer
- Origin of “Big” Data Systems
- Why Big Data Systems
- Types of Big Data Systems 
- Course logistics

Business Intelligence and Big data: The Evolution

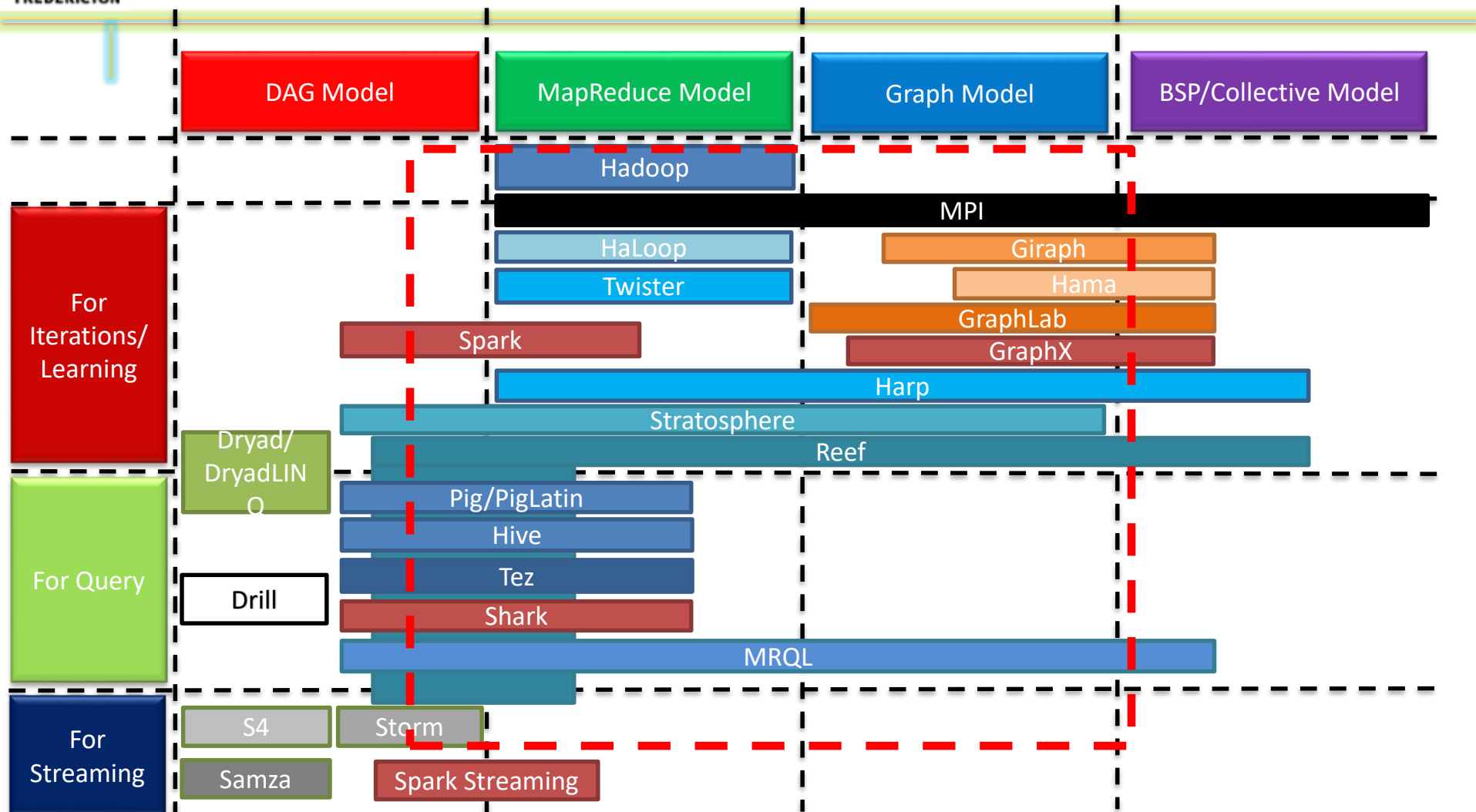


The World of Big Data Tools



Source: B. Zhang

The World of Big Data Tools



Source: B. Zhang

Some “Big Data” Systems: MapReduce Model

- Google MapReduce (2004)
 - Distributed computation on a large dataset can be boiled down to two kinds of computation steps - a map step and a reduce step
 - Jeffrey Dean et al. MapReduce: Simplified Data Processing on Large Clusters. OSDI 2004.



- Apache Hadoop (2005)
 - <http://hadoop.apache.org/>
 - <http://developer.yahoo.com/hadoop/tutorial/>
- Apache Hadoop 2.0 (2012)
 - Vinod Kumar Vavilapalli et al. Apache Hadoop **YARN: Yet Another Resource Negotiator**, SOCC 2013.
 - **Separation** between **resource management** and **computation model**.

Some “Big Data” Systems: Iterative MR Model



- Apache Spark (2010)
 - **Multi-stage in-memory primitives**
 - Matei Zaharia et al. Spark: Cluster Computing with Working Sets,. HotCloud 2010.
 - <http://spark.apache.org/>
 - **Resilient Distributed Dataset (RDD)**: logical collection of data partitioned across machines
 - RDD operations
 - MapReduce-like parallel operations (e.g. map, filter, reduce, join)
 - Simple collectives: broadcasting and aggregation

Some “Big Data” Systems: SQL-like Query processing


- Apache Hive (2007)
 - Facebook Data Infrastructure Team. Hive - A Warehousing Solution Over a Map-Reduce Framework. VLDB 2009.
 - <https://hive.apache.org/>
 - On top of Apache Hadoop
 - HiveQL: transparently converts queries to map/reduce
- Shark/Spark SQL (2012)
 - Reynold Xin et al. Shark: SQL and Rich Analytics at Scale. Technical Report. UCB/EECS 2012.
 - <http://shark.cs.berkeley.edu/>
 - Programming abstraction: DataFrames
 - Distributed SQL query engine



Some “Big Data” Systems: Graph Processing

- Persistent Stores
 - Neo4J
 - Sparksee
 - Titan
- Parallel graph processing systems
 - Pregel (2010)
 - Apache Giraph (2012)
 - GraphLab

Outline

- What is Big Data
- Big Data dimensions
- Data Systems primer
- Origin of “Big” Data Systems
- Why Big Data Systems
- Types of Big Data Systems
- Course logistics 

What we will cover in class

- **Foundations of access methods and query processing**
 - Data models: relational vs. NOSQL
 - Different indexing techniques
 - Query processing overview
 - Join processing
 - Main memory data and query processing
- **Parallel database**
 - Parallel algorithms
 - Partitioning
- **Batch processing frameworks**
 - MapReduce, Hadoop, HiveQL
- **Iterative processing frameworks**
 - Spark, SparkSQL
- **Update-intensive data processing frameworks**
 - Cassandra, HBase
- **Graph data processing**
 - Neo4J
- **Special topics in Big Data (*time permitting*)**
 - Spatio-temporal query processing
 - Blockchain data management