Introduction to Big Data

Welcome!

- Instructor: Ruoming Jin
  - Office: 264 MCS Building
  - Email: jin AT cs.kent.edu
  - Office hour: Mondays (4:30PM to 5:30PM) or by appointment
- TA: Xinyu Chang
  - Email: xchang AT kent.edu
- Homepage: http://www.cs.kent.edu/~jin/BigData/index.html

Topics

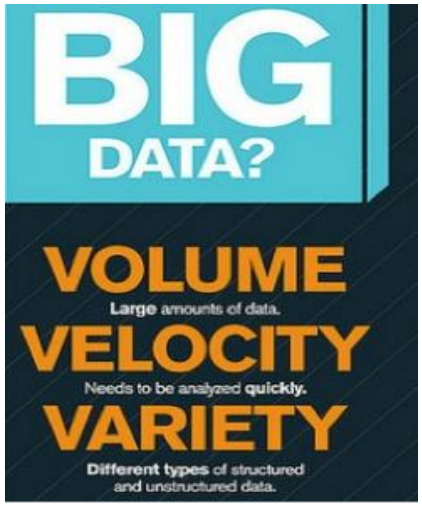
- Scope: Big Data & Analytics
- Topics:
  - Foundation of Data Analytics and Data Mining
  - Hadoop/Map-Reduce Programming and Data Processing & BigTable/Hbase/Cassandra
  - Graph Database and Graph Analytics

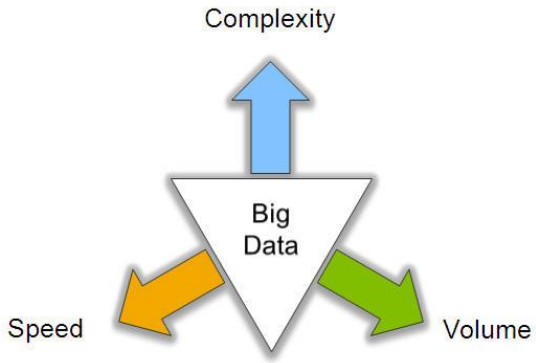
What's Big Data?

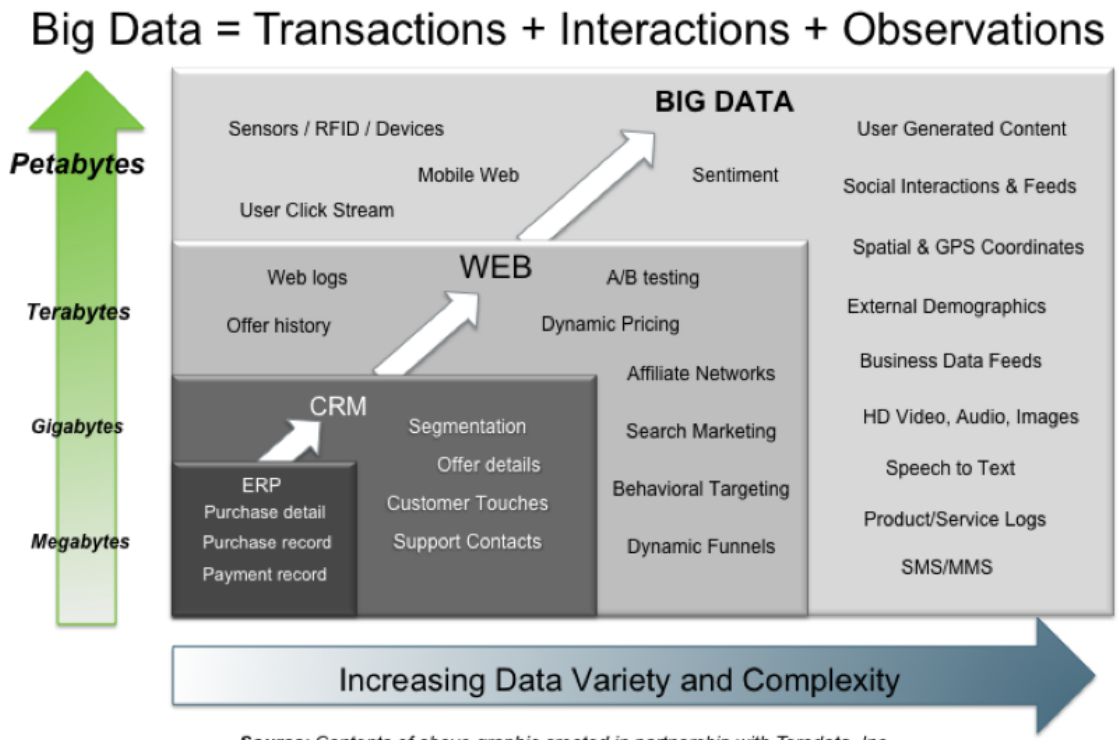
No single definition; here is from Wikipedia:

- **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.
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Big Data: 3V's



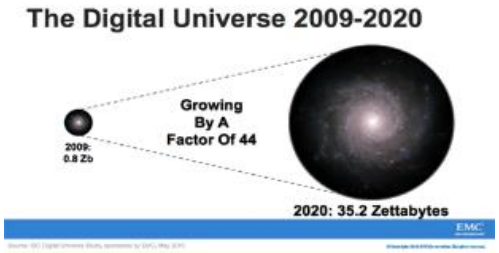


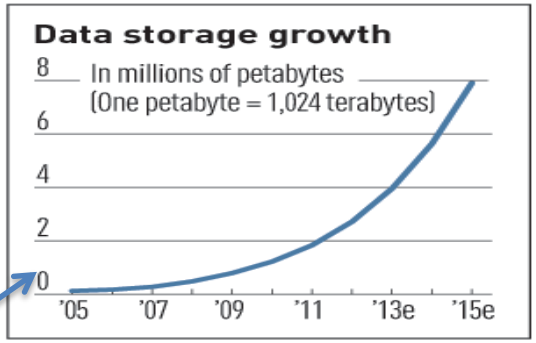


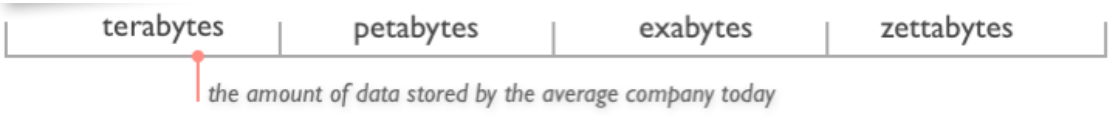
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

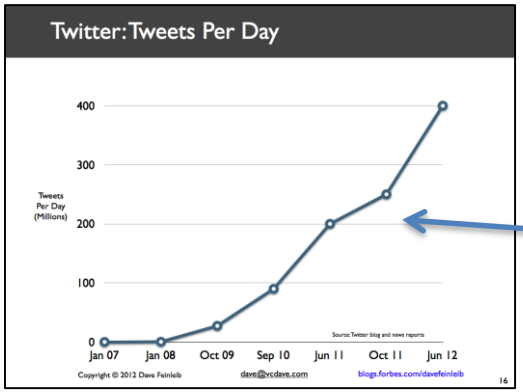
Volume (Scale)

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









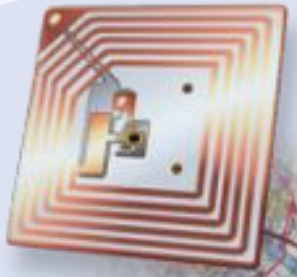
Exponential increase in collected/generated data

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs of **log data every day**

30 billion RFID tags today (1.3B in 2005)

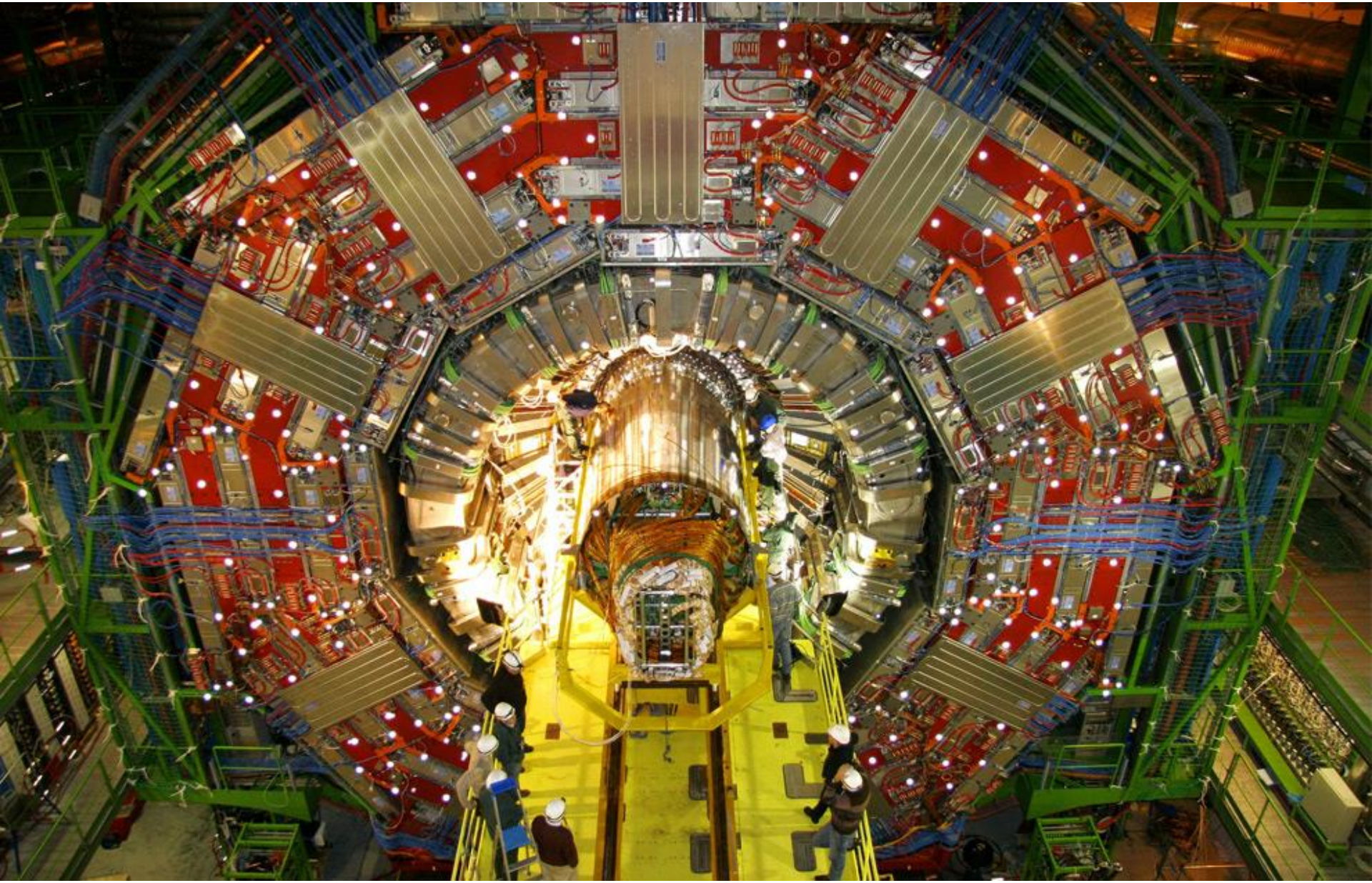


4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

76 million smart meters in 2009…
200M by 2014

2+ billion people on the Web by end 2011



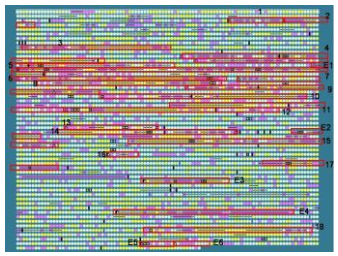
CERN’s Large Hydron Collider (LHC) generates 15 PB a year

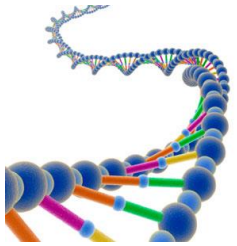
The Earthscope

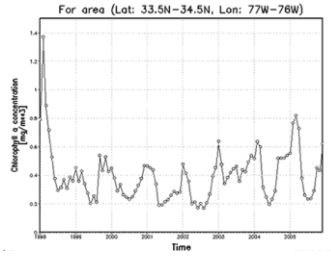
- The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more. (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI)

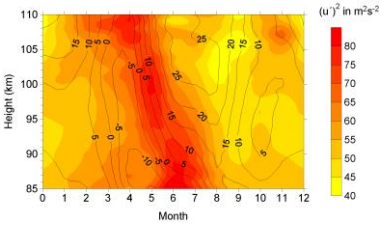
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)



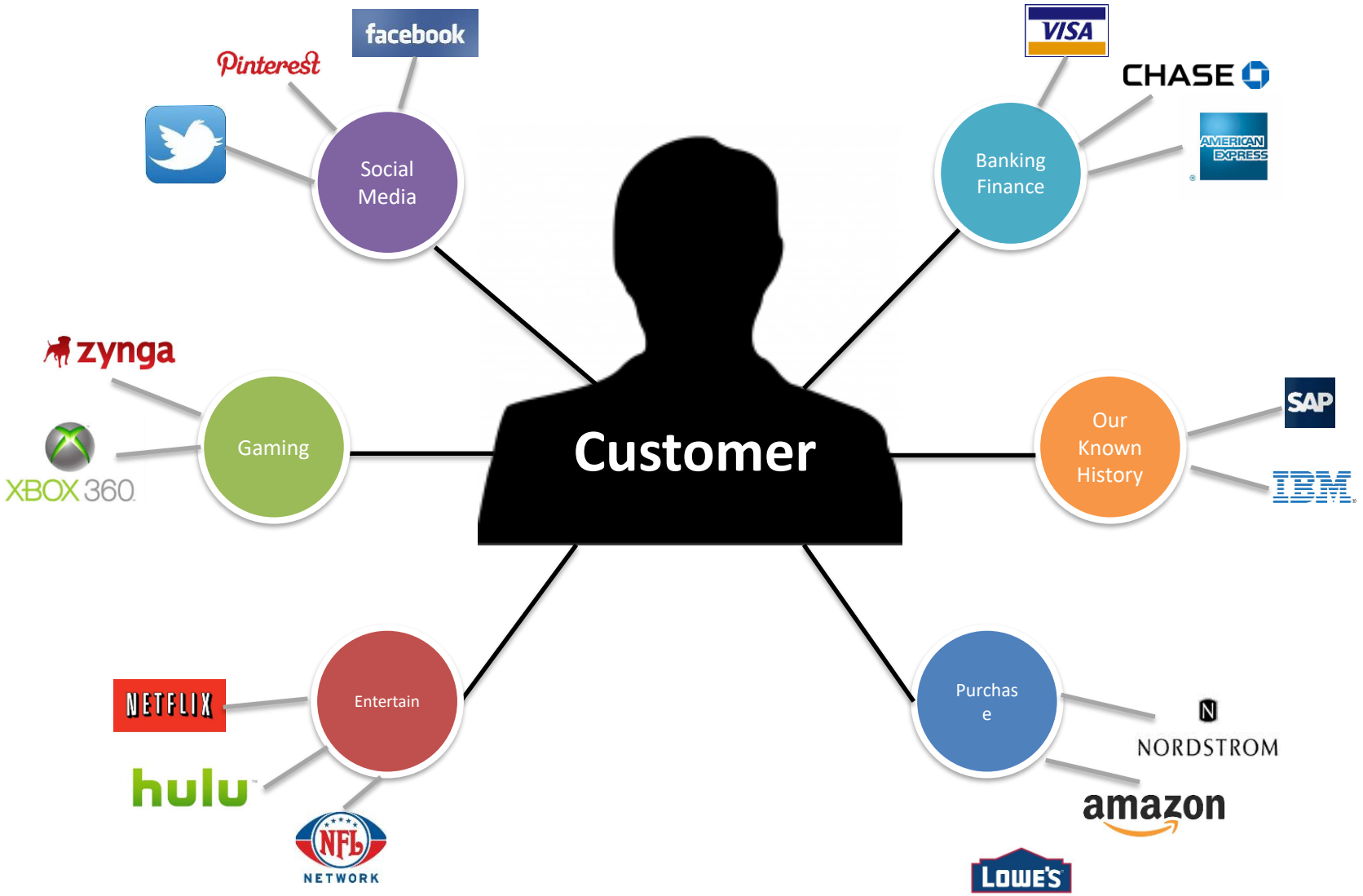






To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

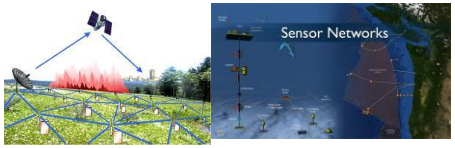
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

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

Scientific instruments
(collecting all sorts of data)

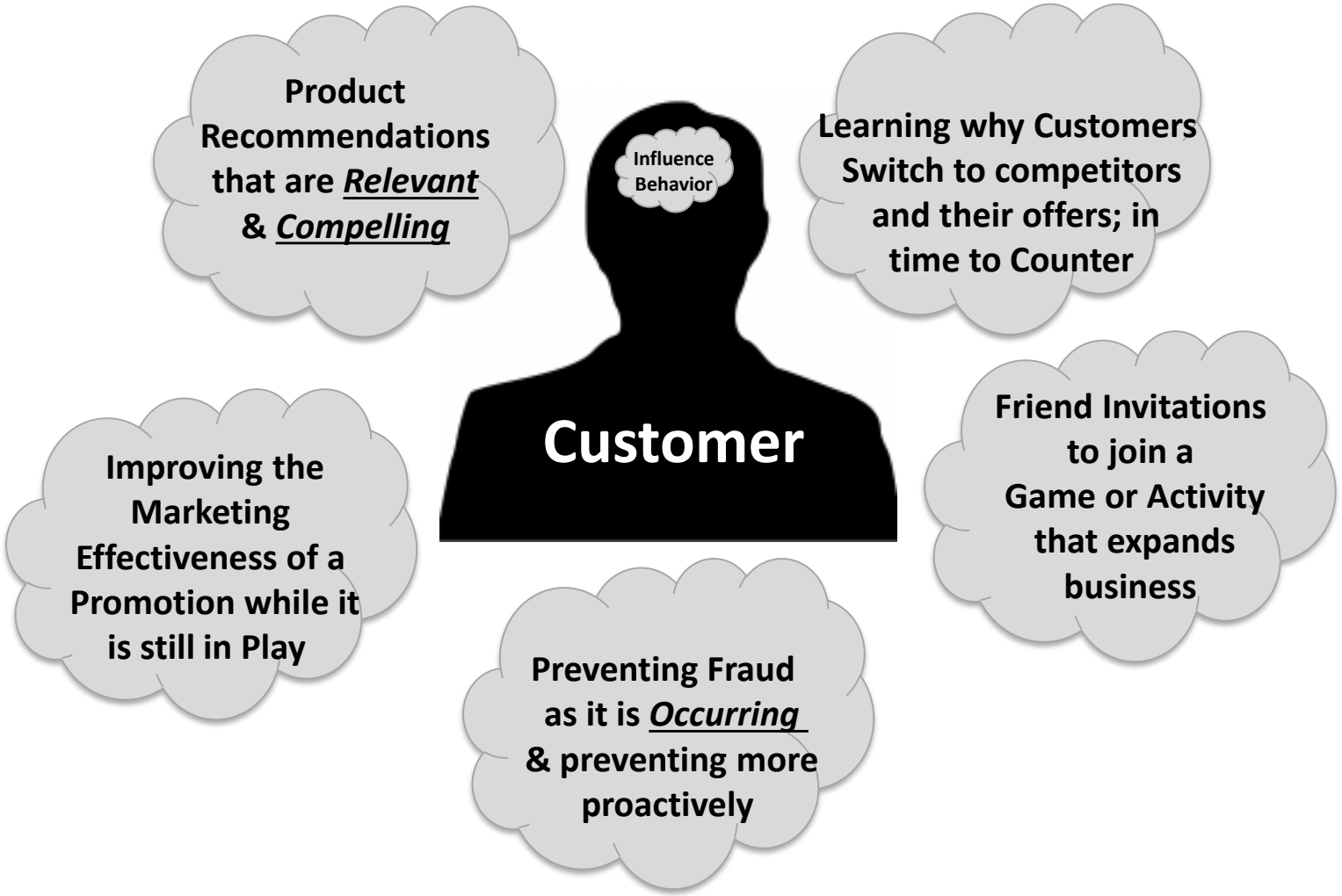
Mobile devices
(tracking all objects all the time)



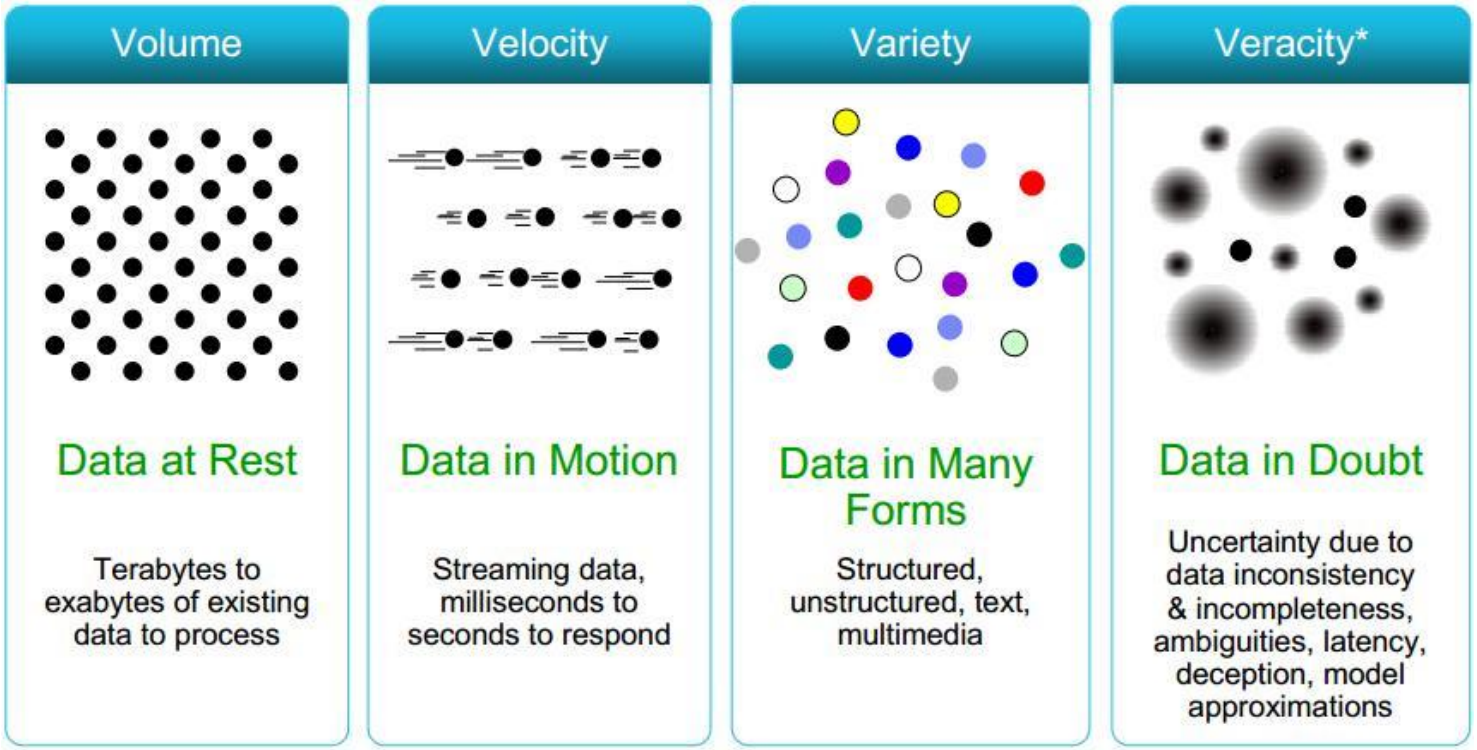
Sensor technology and networks
(measuring all kinds of data)

- 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

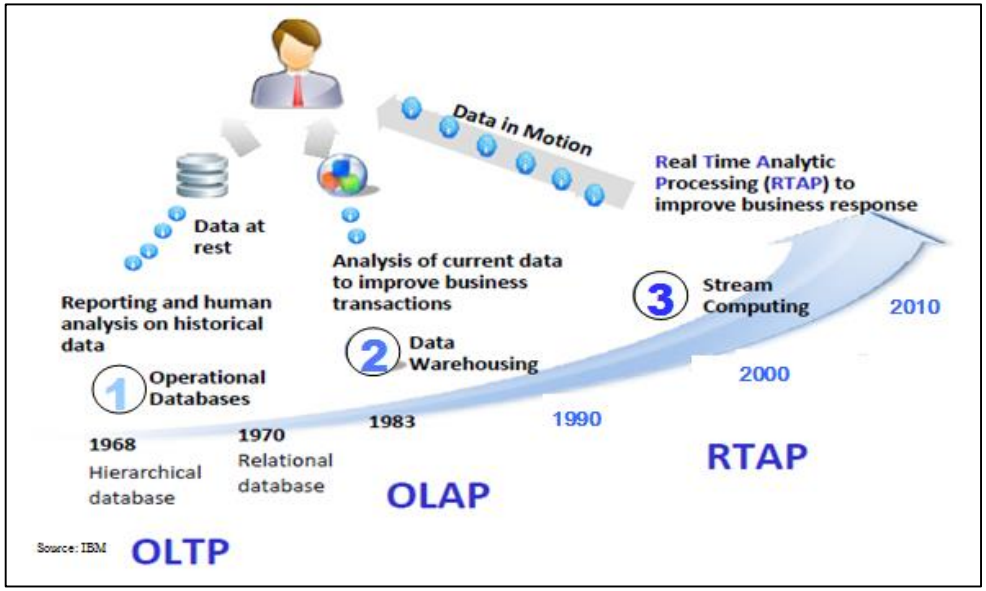
Real-Time Analytics/Decision Requirement



Some Make it 4V's



Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

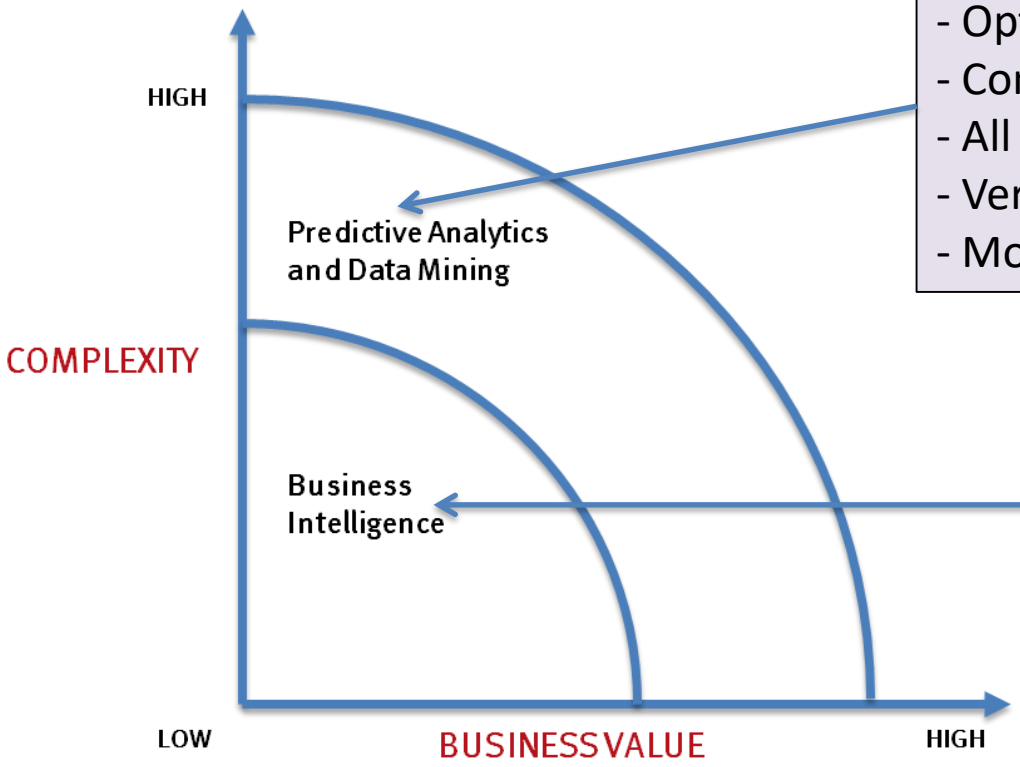
The Model Has Changed…

- **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

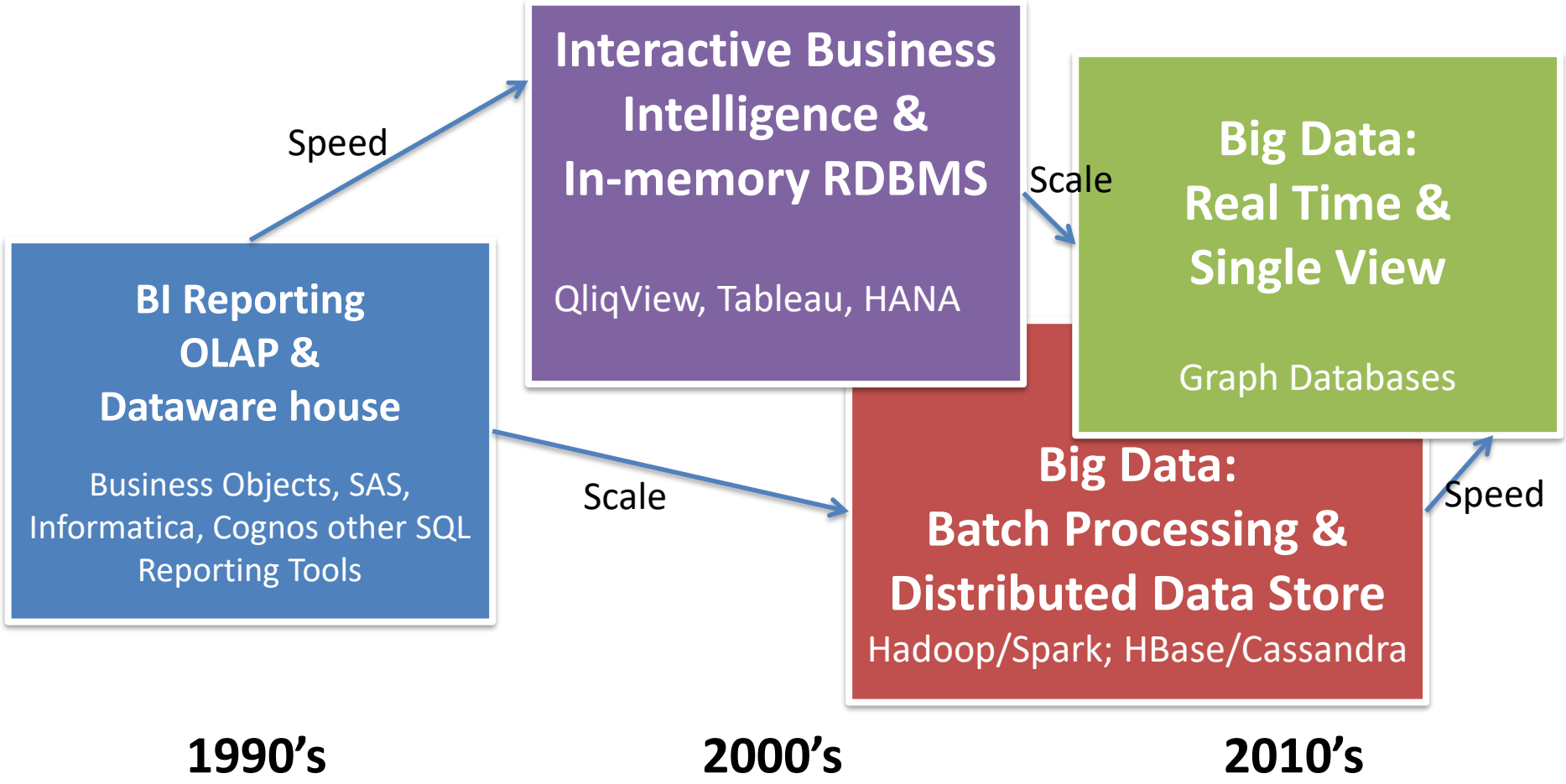
What's driving Big Data



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

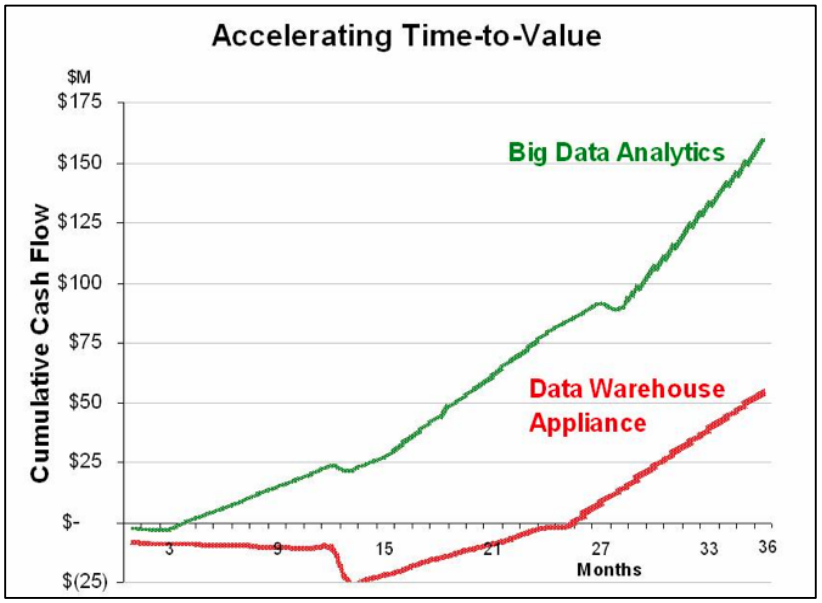
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

THE EVOLUTION OF BUSINESS INTELLIGENCE

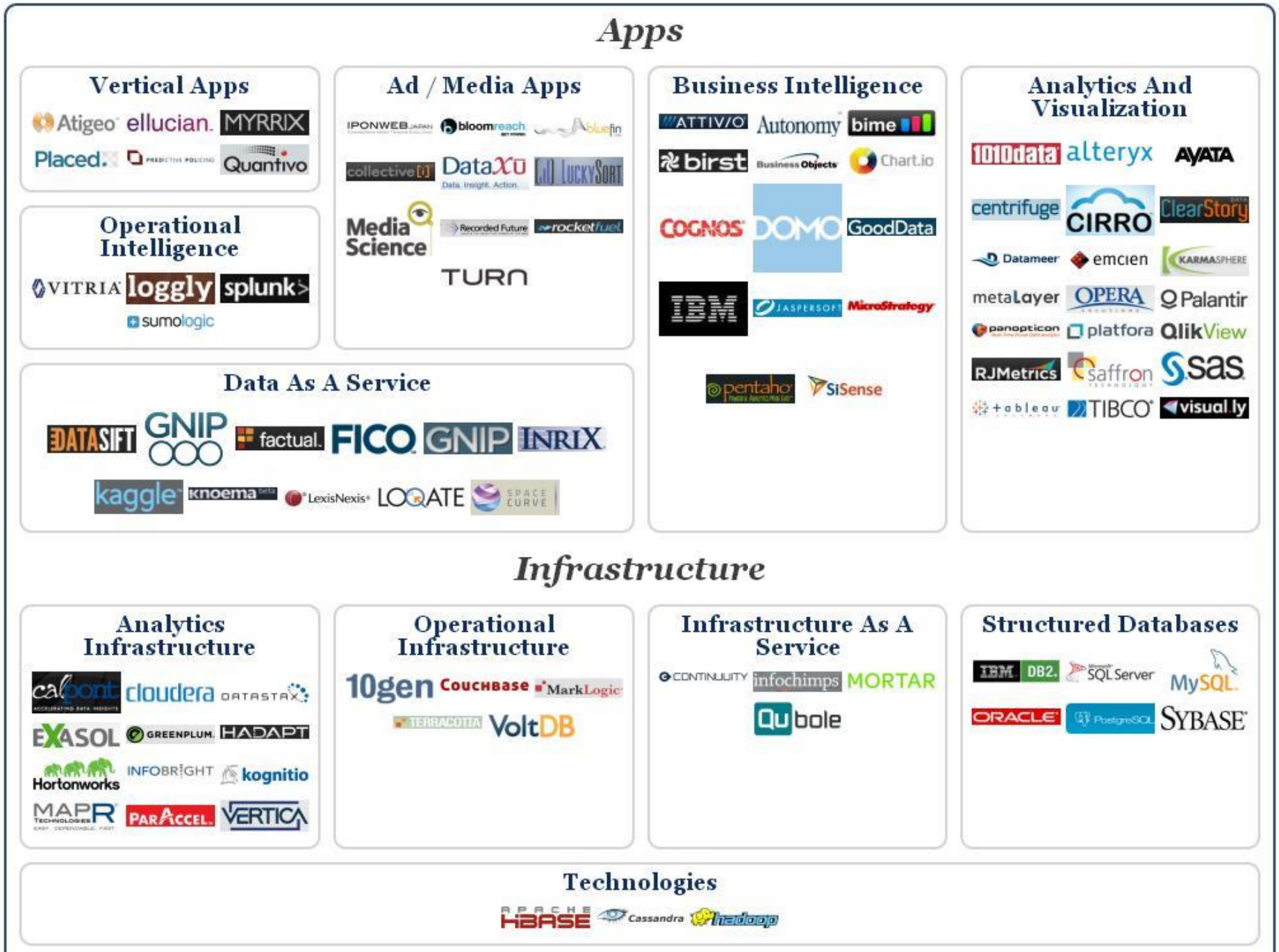


Big Data Analytics

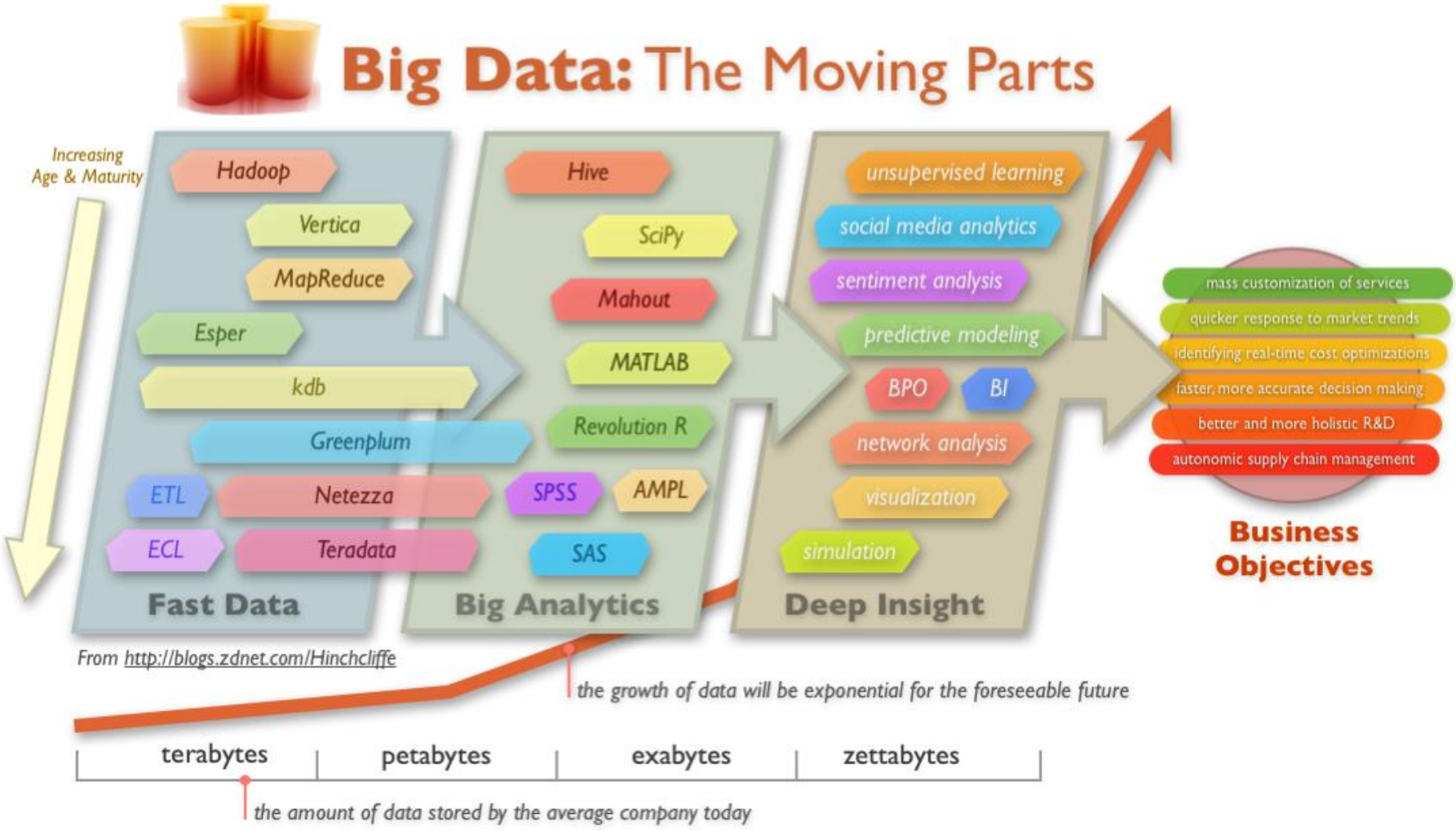
- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



The Big Data Landscape

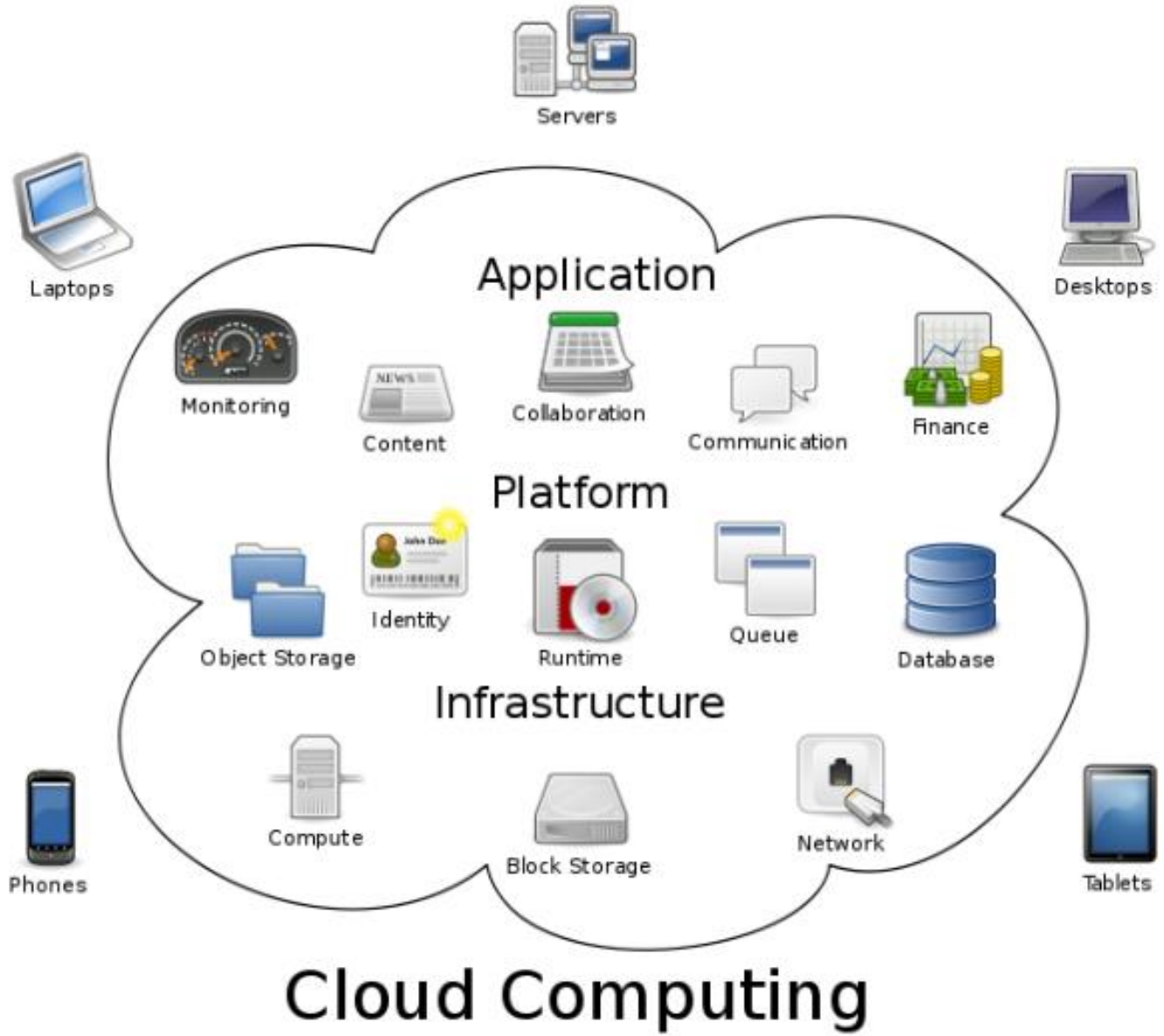


Big Data Technology



Cloud Computing

- IT resources provided as a service
  - Compute, storage, databases, queues
- Clouds leverage economies of scale of commodity hardware
  - Cheap storage, high bandwidth networks & multicore processors
  - Geographically distributed data centers
- Offerings from Microsoft, Amazon, Google, …



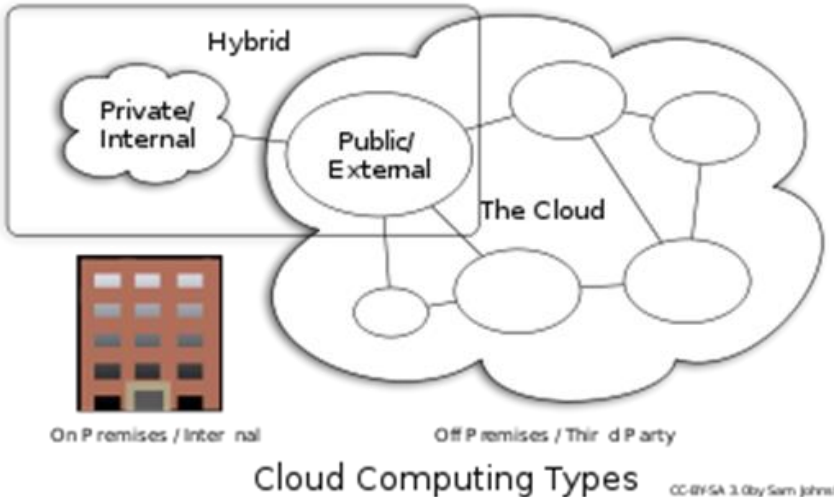
wikipedia:Cloud Computing

Benefits

- Cost & management
  - Economies of scale, “out-sourced” resource management
- Reduced Time to deployment
  - Ease of assembly, works “out of the box”
- Scaling
  - On demand provisioning, co-locate data and compute
- Reliability
  - Massive, redundant, shared resources
- Sustainability
  - Hardware not owned

Types of Cloud Computing

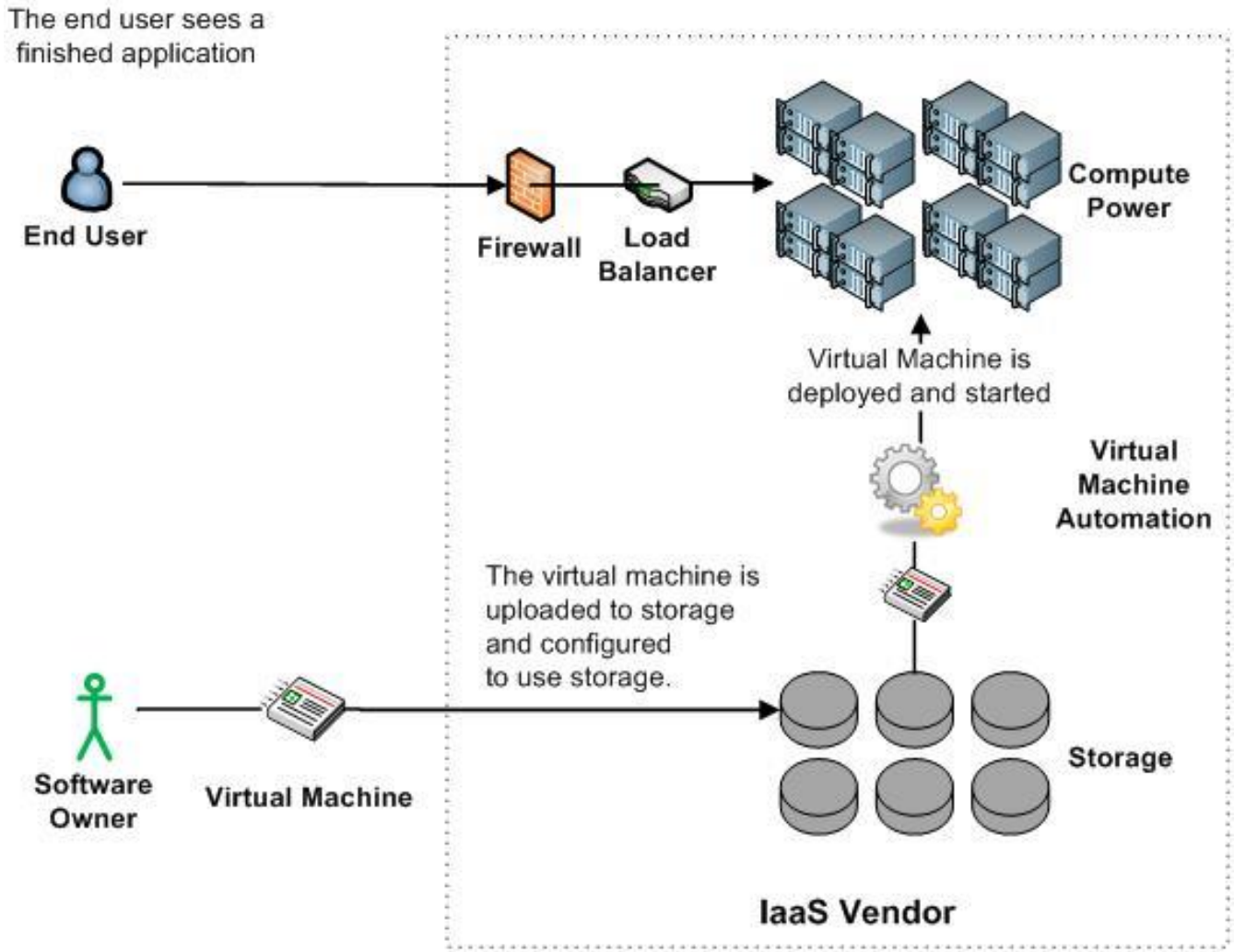
- **Public Cloud**: Computing infrastructure is hosted at the vendor's premises.
- **Private Cloud**: Computing architecture is dedicated to the customer and is not shared with other organisations.
- **Hybrid Cloud**: Organisations host some critical, secure applications in private clouds. The not so critical applications are hosted in the public cloud
  - **Cloud bursting**: the organisation uses its own infrastructure for normal usage, but cloud is used for peak loads.
- **Community Cloud**



Classification of Cloud Computing based on Service Provided

- Infrastructure as a service (IaaS)
  - Offering hardware related services using the principles of cloud computing. These could include storage services (database or disk storage) or virtual servers.
  - Amazon EC2, Amazon S3, Rackspace Cloud Servers and Flexiscale.
- Platform as a Service (PaaS)
  - Offering a development platform on the cloud.
  - Google's Application Engine, Microsofts Azure, Salesforce.com's force.com .
- Software as a service (SaaS)
  - Including a complete software offering on the cloud. Users can access a software application hosted by the cloud vendor on pay-per-use basis. This is a well-established sector.
  - Salesforce.coms' offering in the online Customer Relationship Management (CRM) space, Googles gmail and Microsofts hotmail, Google docs.

Infrastructure as a Service (IaaS)



More Refined Categorization

- Storage-as-a-service
- Database-as-a-service
- Information-as-a-service
- Process-as-a-service
- Application-as-a-service
- Platform-as-a-service
- Integration-as-a-service
- Security-as-a-service
- Management/ Governance-as-a-service
- Testing-as-a-service
- Infrastructure-as-a-service

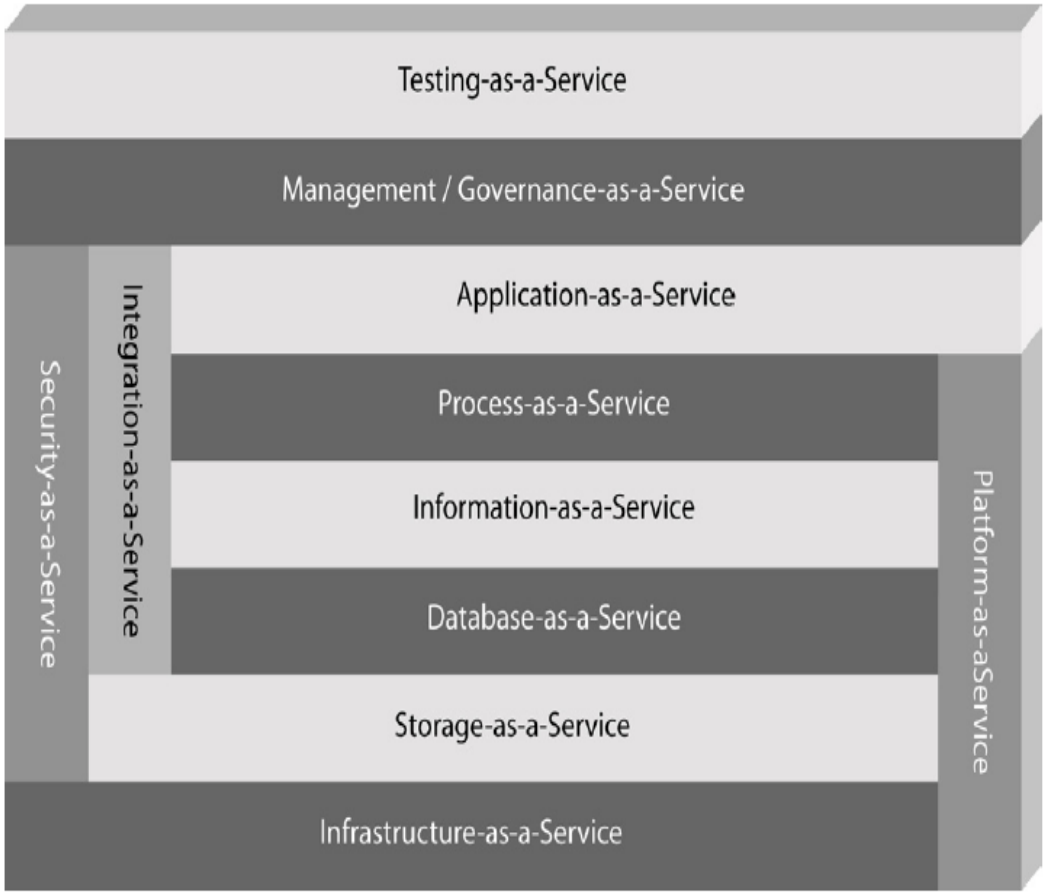


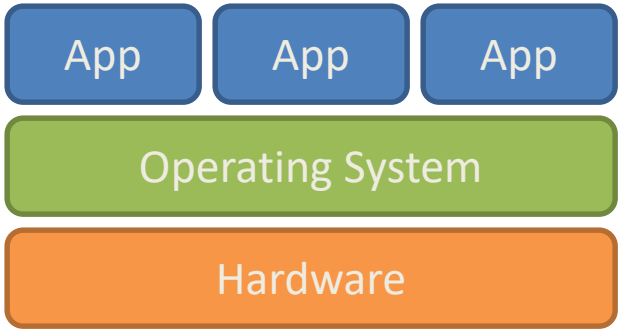
Figure 1: The patterns or categories of cloud computing providers allow you to use a discrete set of services within your architecture.

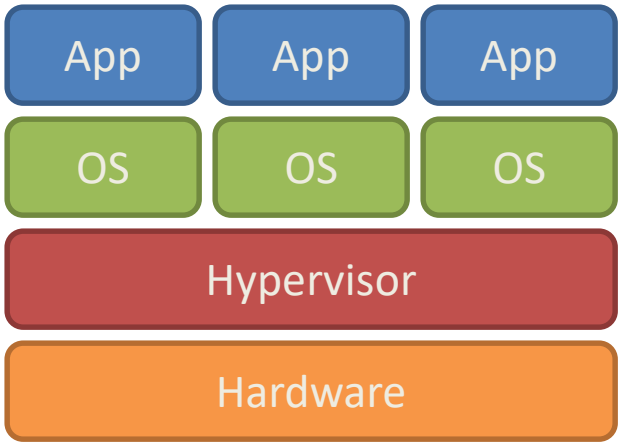
InfoWorld Cloud Computing Deep Dive

Key Ingredients in Cloud Computing

- Service-Oriented Architecture (SOA)
- Utility Computing (on demand)
- Virtualization (P2P Network)
- SAAS (Software As A Service)
- PAAS (Platform AS A Service)
- IAAS (Infrastructure AS A Servie)
- Web Services in Cloud

Enabling Technology: Virtualization





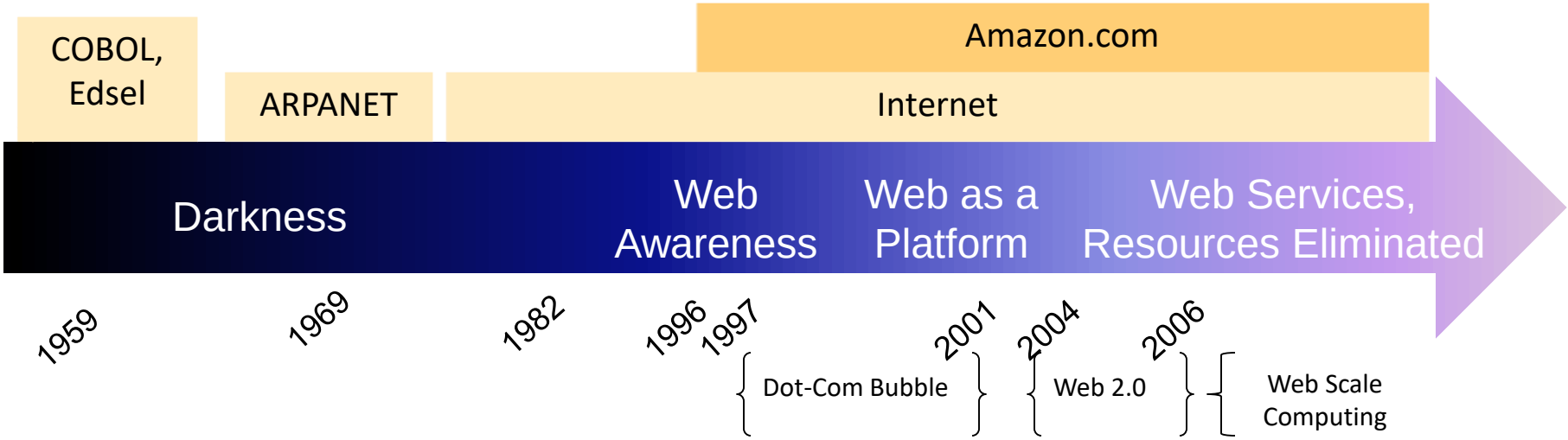
Everything as a Service

- Utility computing = Infrastructure as a Service (IaaS)
  - Why buy machines when you can rent cycles?
  - Examples: Amazon's EC2, Rackspace
- Platform as a Service (PaaS)
  - Give me nice API and take care of the maintenance, upgrades, …
  - Example: Google App Engine
- Software as a Service (SaaS)
  - Just run it for me!
  - Example: Gmail, Salesforce

Cloud versus cloud

- Amazon Elastic Compute Cloud
- Google App Engine
- Microsoft Azure
- GoGrid
- AppNexus

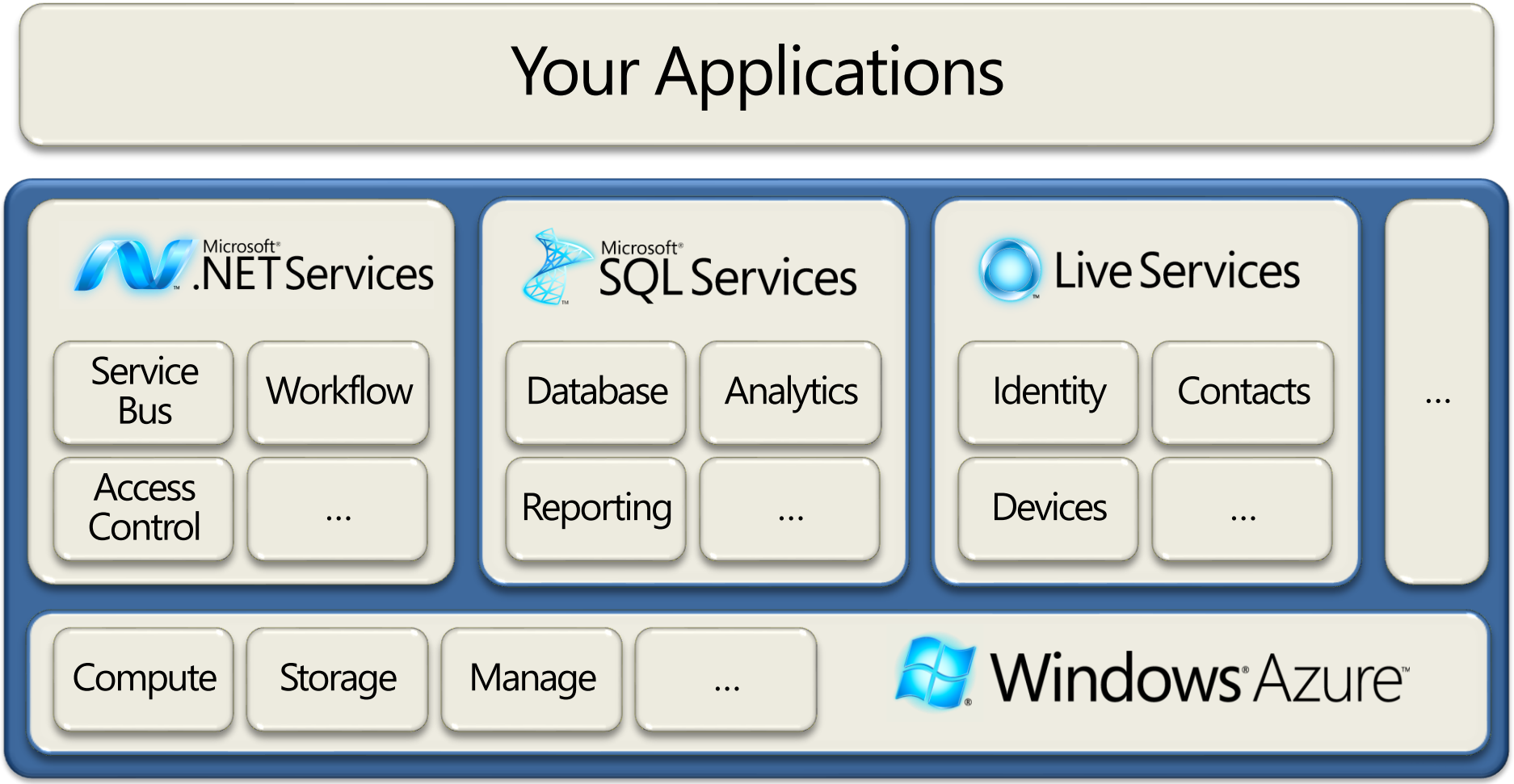
The Obligatory Timeline Slide (Mike Culver @ AWS)



AWS

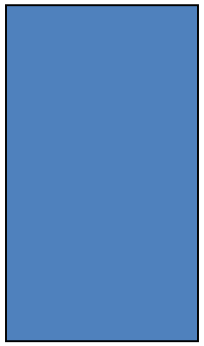
- Elastic Compute Cloud – EC2 (IaaS)
- Simple Storage Service – S3 (IaaS)
- Elastic Block Storage – EBS (IaaS)
- SimpleDB (SDB) (PaaS)
- Simple Queue Service – SQS (PaaS)
- CloudFront (S3 based Content Delivery Network – PaaS)
- Consistent AWS Web Services API

What does Azure platform offer to developers?



Google's AppEngine vs Amazon's EC2

Python
BigTable
Other API's



VMs
Flat File Storage

AppEngine:

- Higher-level functionality (e.g., automatic scaling)
- More restrictive (e.g., respond to URL only)
- Proprietary lock-in

EC2/S3:

- Lower-level functionality
- More flexible
- Coarser billing model

Topics Overview

Topic 1: Data Analytics & Data Mining

- Exploratory Data Analysis
- Linear Classification (Perceptron & Logistic Regression)
- Linear Regression
- C4.5 Decision Tree
- Apriori
- K-means Clustering
- EM Algorithm
- PageRank & HITS
- Collaborative Filtering

Topic 2: Hadoop/MapReduce Programming & Data Processing

- Architecture of Hadoop, HDFS, and Yarn
- Programming on Hadoop

- Basic Data Processing: Sort and Join
- Information Retrieval using Hadoop
- Data Mining using Hadoop (Kmeans+Histograms)
- Machine Learning on Hadoop (EM)

- Hive/Pig
- HBase and Cassandra

Topic 3: Graph Database and Graph Analytics

- Graph Database (http://en.wikipedia.org/wiki/Graph_database)
  - Native Graph Database (Neo4j)
  - Pregel/Giraph (Distributed Graph Processing Engine)
- Neo4j/Titan/GraphLab/GraphSQL

Textbooks

- No Official Textbooks
- References:
  - Hadoop: The Definitive Guide, Tom White, O'Reilly
  - Hadoop In Action, Chuck Lam, Manning
  - Data-Intensive Text Processing with MapReduce, Jimmy Lin and Chris Dyer (*www.umiacs.umd.edu/~jimmylin/**MapReduce**-book-final.pdf*)
  - Data Mining: Concepts and Techniques, Third Edition, by Jiawei Han et al.
- Many Online Tutorials and Papers

Cloud Resources

- Hadoop on your local machine
- Hadoop in a virtual machine on your local machine (Pseudo-Distributed on **Ubuntu**)
- Hadoop in the clouds with Amazon EC2

Course Prerequisite

- Prerequisite:
  - Java Programming / C++
  - Data Structures and Algorithm
  - Computer Architecture
  - Basic Statistics and Probability
  - Database and Data Mining (preferred)

This course is not for you…

- If you do not have a strong Java programming background
  - This course is not about only programming (on Hadoop).
  - Focus on “thinking at scale” and algorithm design
  - Focus on how to manage and process Big Data!
- No previous experience necessary in
  - MapReduce
  - Parallel and distributed programming

Grade Scheme

Homework	70%
Project	20%
Class Participation	10%

- Each Class will have a sign-in sheet
- Zero-Tolerance on plagiarism

Project

- Project (due April 24th)
  - One project: Group size <= 4 students
  - Checkpoints
    - Proposal: title and goal (due March 1st)
    - Outline of approach (due March 15th)
    - Implementation and Demo (April 24th and 26th)
    - Final Project Report (due April 29th)
  - Each group will have a short presentation and demo (15-20 minutes)
  - Each group will provide a five-page document on the project; the responsibility and work of each student shall be described precisely