# Big Data

Βάσεις Δεδομένων

2016-2017

National Technical University of Athens CSLab

# Τι είναι Big Data

- The basic idea behind the phrase **'Big Data'** is that everything we do is increasingly leaving a digital trace (or data), which we (and others) can use and analyse.

- Big Data therefore refers to our ability to make use of the ever-increasing volumes of data

- No single definition

# Wikipedia definition

- ***Big data*** is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.

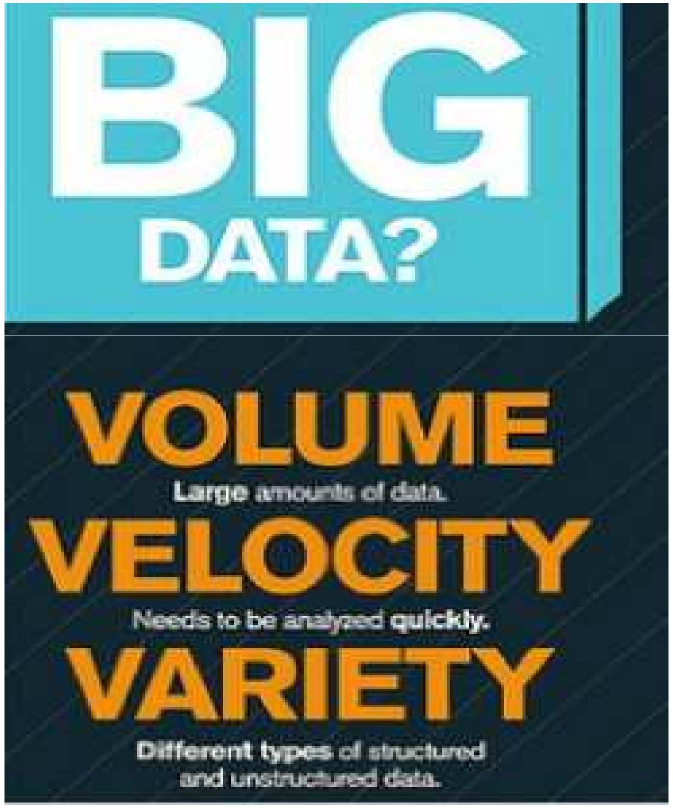
**From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days…and the pace is accelerating.**

**Eric Schmidt,**
***Executive Chairman, Google***

# Sources

- Activity Data
- Conversation Data
- Photo/Video Image Data
- Sensor Data
- The Internet of Things Data

# 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.

# Data Volume

- 90% των σημερινών δεδομένων δημιουργήθηκαν τα τελευταία 2 χρόνια
- Νόμος του Moore: Διπλασιασμός δεδομένων κάθε 18-24 μήνες
  - From 0.8 zettabytes in 2010 to 35zb in 2020

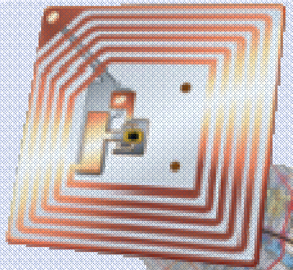
640K είναι αρκετά για όλους...

*12+ TBs*
**of tweet data every day**

*? TBs* of data every day

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

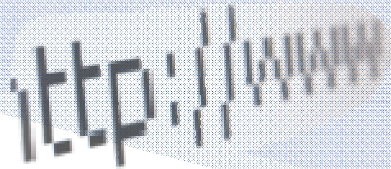
*30 billion* RFID tags today (1.3B in 2005)



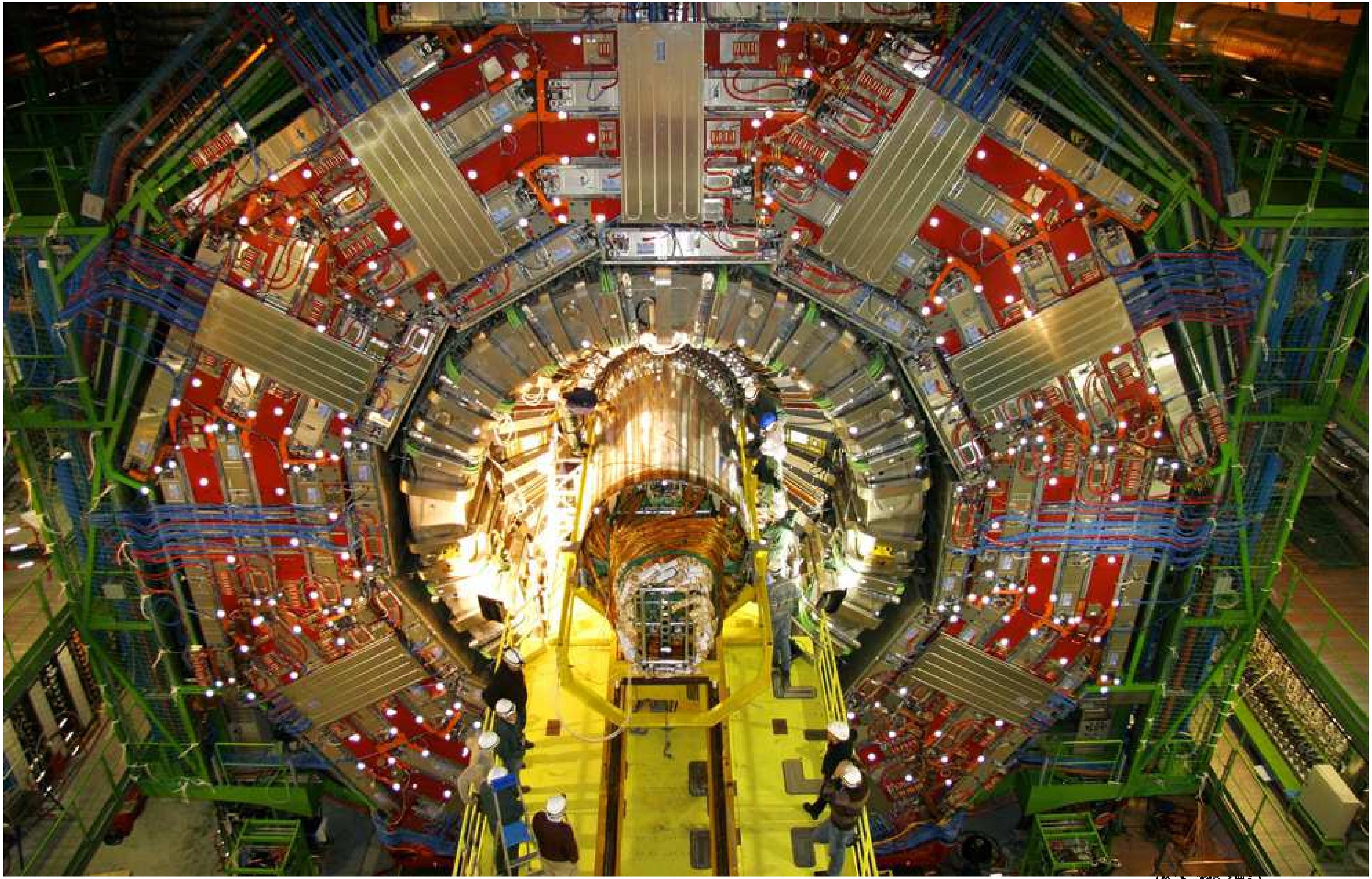
*4.6 billion* camera phones world wide

*100s of millions of GPS enabled* devices sold annually

*200 million* smart meters in 2014…



*2+ billion* people on the Web by end 2011



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

# Έκρηξη Δεδομένων

Byte : one grain of rice

Byte

# Έκρηξη Δεδομένων

| | |
|---|---|
| Byte | : one grain of rice |
| Kilobyte | : cup of rice |

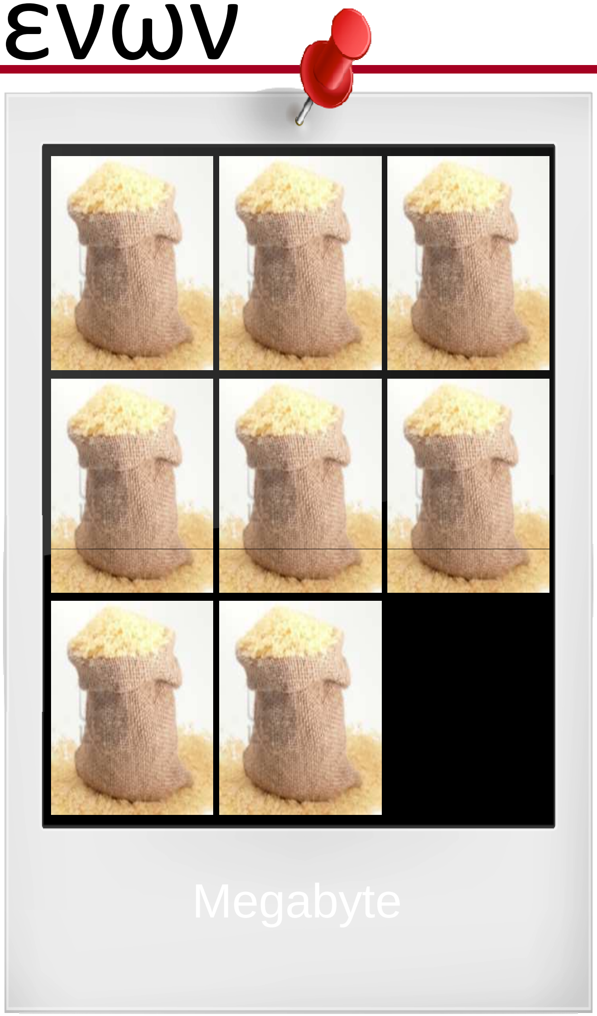
Kilobyte

# Έκρηξη Δεδομένων

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice



# Έκρηξη Δεδομένων

Byte : one grain of rice

Kilobyte : cup of rice

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Gigabyte

# Έκρηξη Δεδομένων

| | |
|---|---|
| Byte | : one grain of rice |
| Kilobyte | : cup of rice |
| Megabyte | : 8 bags of rice |
| Gigabyte | : 3 Semi trucks |
| Terabyte | : 2 Container Ships |

Terabyte

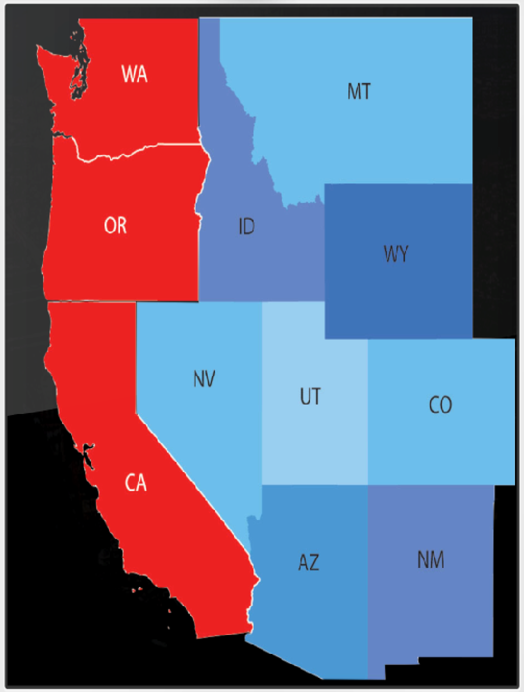
# Έκρηξη Δεδομένων

| | |
|---|---|
| Byte | : one grain of rice |
| Kilobyte | : cup of rice |
| Megabyte | : 8 bags of rice |
| Gigabyte | : 3 Semi trucks |
| Terabyte | : 2 Container Ships |
| Petabyte | : Blankets Manhattan |

Petabyte

# Έκρηξη Δεδομένων

Byte : one grain of rice
Kilobyte : cup of rice
Megabyte : 8 bags of rice
Gigabyte : 3 Semi trucks
Terabyte : 2 Container Ships
Petabyte : Blankets Manhattan
Exabyte : Blankets west coast states



Exabyte

# Έκρηξη Δεδομένων

Byte : one grain of rice
Kilobyte : cup of rice
Megabyte : 8 bags of rice
Gigabyte : 3 Semi trucks
Terabyte : 2 Container Ships
Petabyte : Blankets Manhattan
Exabyte : Blankets west coast states
Zettabyte : Fills the Pacific Ocean

Zettabyte

# Έκρηξη Δεδομένων

Byte : one grain of rice
Kilobyte : cup of rice
Megabyte : 8 bags of rice
Gigabyte : 3 Semi trucks
Terabyte : 2 Container Ships
Petabyte : Blankets Manhattan
Exabyte : Blankets west coast states
Zettabyte : Fills the Pacific Ocean
**Yottabyte : AN EARTH SIZE RICE BALL!**

Yottabyte

# Έκρηξη Δεδομένων

Byte : one grain of rice

Kilobyte : cup of rice

Hobbyist

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

Exabyte : Blankets west coast states

Zettabyte : Fills the Pacific Ocean

Yottabyte : AN EARTH SIZE RICE BALL!

# Έκρηξη Δεδομένων

| | |
|---|---|
| Byte | : one grain of rice |
| Kilobyte | : cup of rice |
| Megabyte | : 8 bags of rice |
| Gigabyte | : 3 Semi trucks |
| Terabyte | : 2 Container Ships |
| Petabyte | : Blankets Manhattan |
| Exabyte | : Blankets west coast states |
| Zettabyte | : Fills the Pacific Ocean |
| Yottabyte | : AN EARTH SIZE RICE BALL! |

Hobbyist

Desktop

# Έκρηξη Δεδομένων

Byte : one grain of rice

Kilobyte : cup of rice

Hobbyist

Megabyte : 8 bags of rice

Gigabyte : 3 Semi trucks

Desktop

Terabyte : 2 Container Ships

Petabyte : Blankets Manhattan

Internet

Exabyte : Blankets west coast states

Zettabyte : Fills the Pacific Ocean

Yottabyte : AN EARTH SIZE RICE BALL!

# Έκρηξη Δεδομένων

Byte : one grain of rice
Kilobyte : cup of rice

Hobbyist

Megabyte : 8 bags of rice
Gigabyte : 3 Semi trucks

Desktop

Terabyte : 2 Container Ships
Petabyte : Blankets Manhattan

Internet

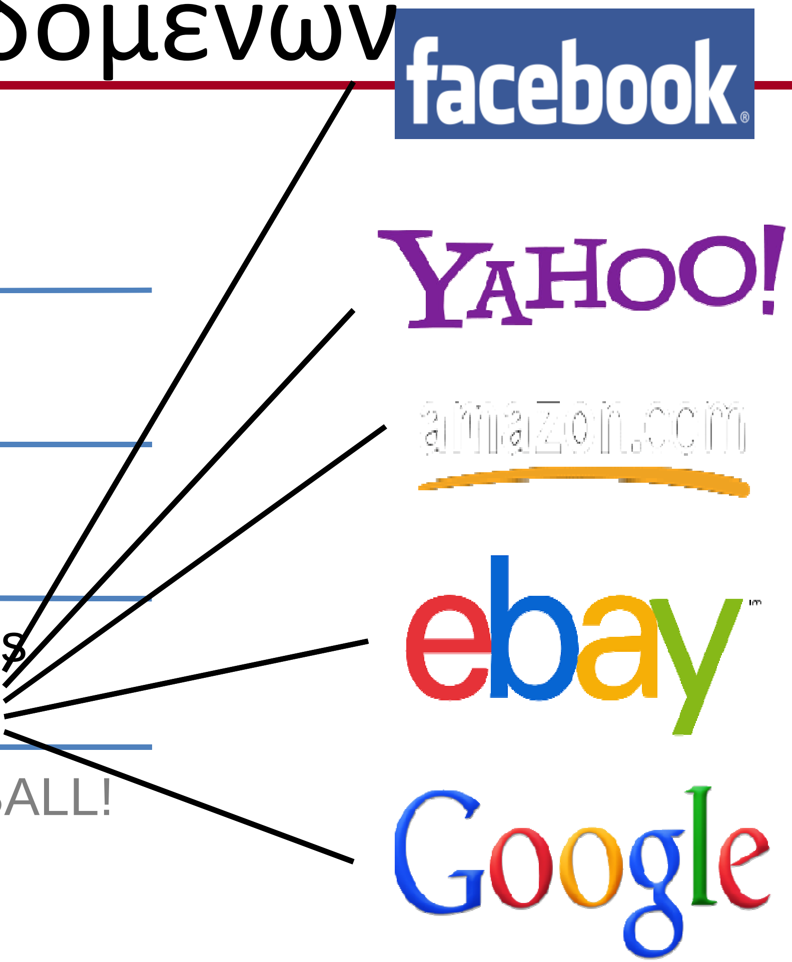
Exabyte : Blankets west coast states
Zettabyte : Fills the Pacific Ocean

Big Data

Yottabyte : AN EARTH SIZE RICE BALL!

# Έκρηξη Δεδομένων

| | |
|---|---|
| Byte | : one grain of rice |
| Kilobyte | : cup of rice |
| Megabyte | : 8 bags of rice |
| Gigabyte | : 3 Semi trucks |
| Terabyte | : 2 Container Ships |
| Petabyte | : Blankets Manhattan |
| Exabyte | : Blankets west coast states |
| Zettabyte | : Fills the Pacific Ocean |
| Yottabyte | : AN EARTH SIZE RICE BALL! |

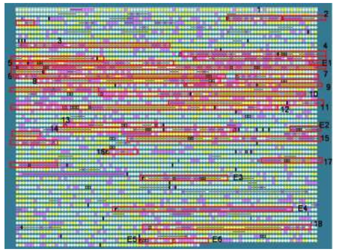


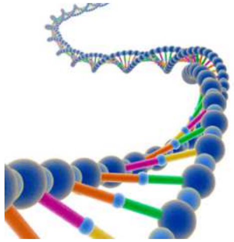
# Έκρηξη Δεδομένων

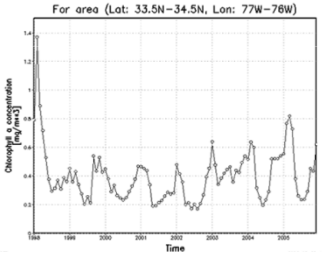
| | | | |
|---|---|---|---|
| Byte | : one grain of rice | | Hobbyist |
| Kilobyte | : cup of rice | | |
| Megabyte | : 8 bags of rice | | Desktop |
| Gigabyte | : 3 Semi trucks | | |
| Terabyte | : 2 Container Ships | | Internet |
| Petabyte | : Blankets Manhattan | | |
| Exabyte | : Blankets west coast states | | Big Data |
| Zettabyte | : Fills the Pacific Ocean | | |
| **Yottabyte** | **: AN EARTH SIZE RICE BALL!** | | *The Future?* |

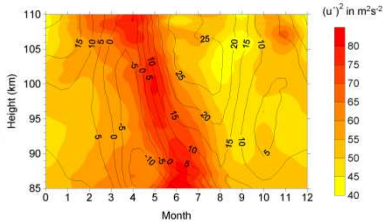
# Variety (Complexity)

- 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)



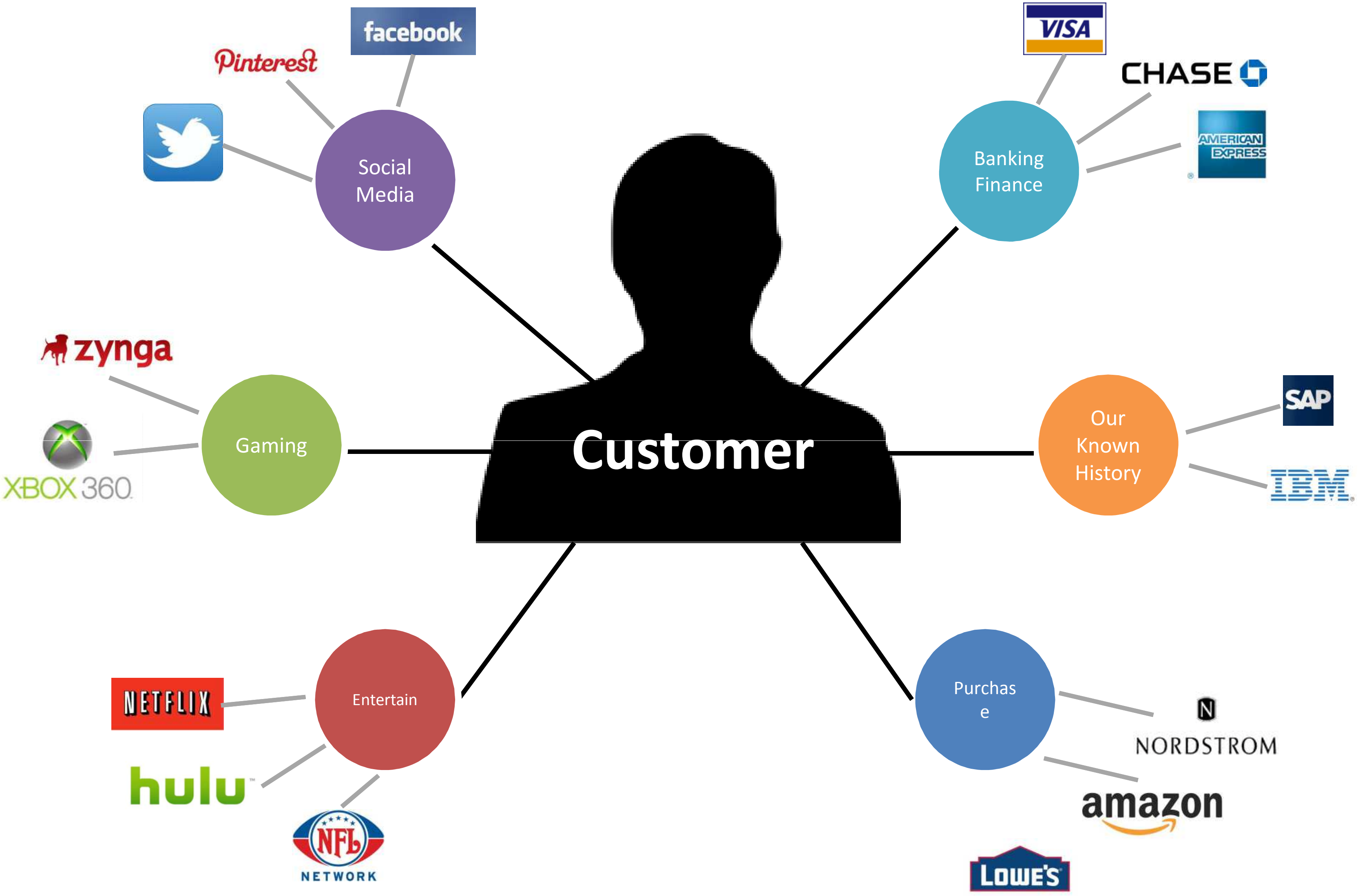



For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m**3]
Time



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)

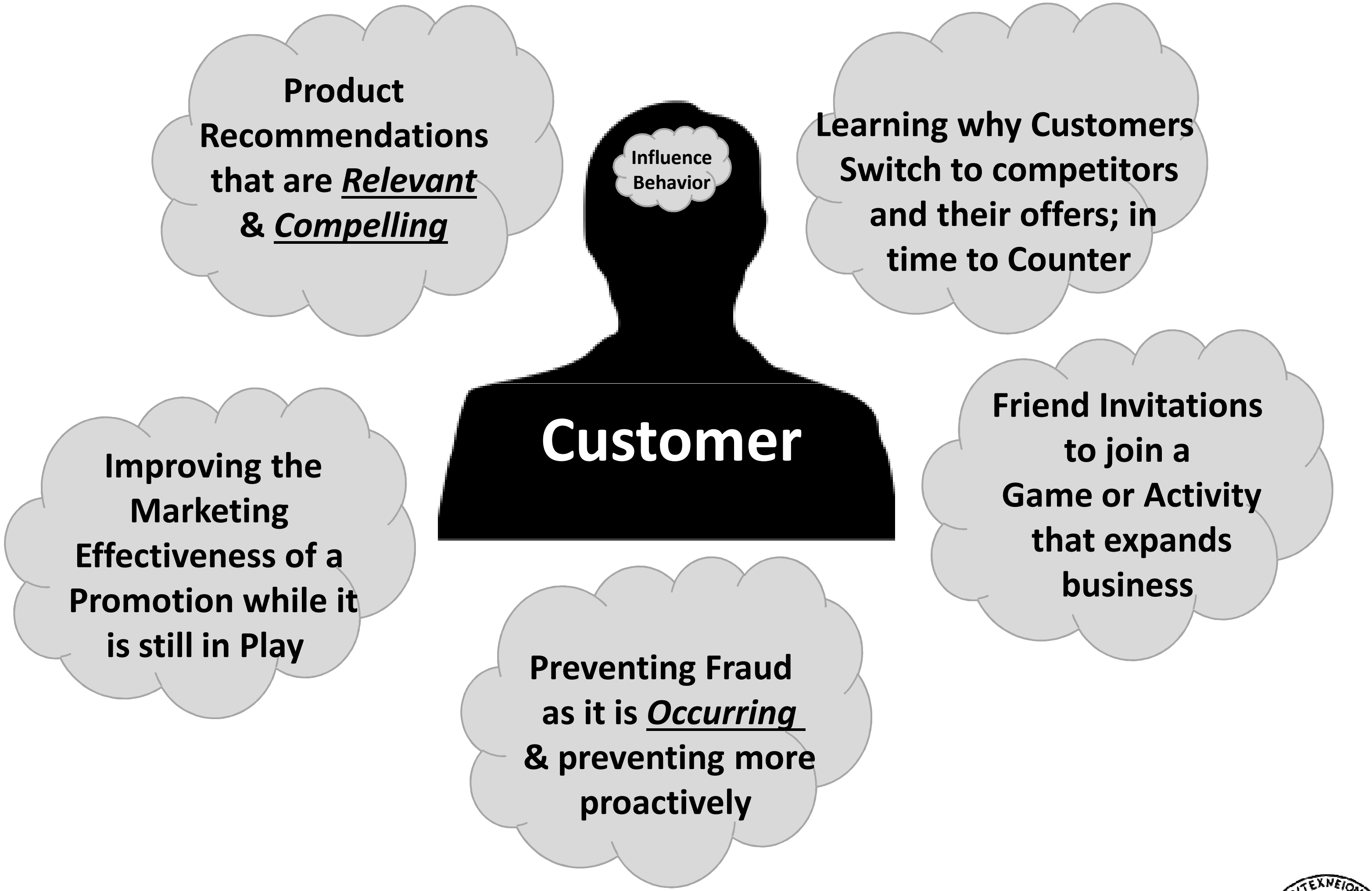
**Scientific instruments**
(collecting all sorts of data)

**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



Product Recommendations that are ***Relevant*** & ***Compelling***

Influence Behavior

Learning why Customers Switch to competitors and their offers; in time to Counter

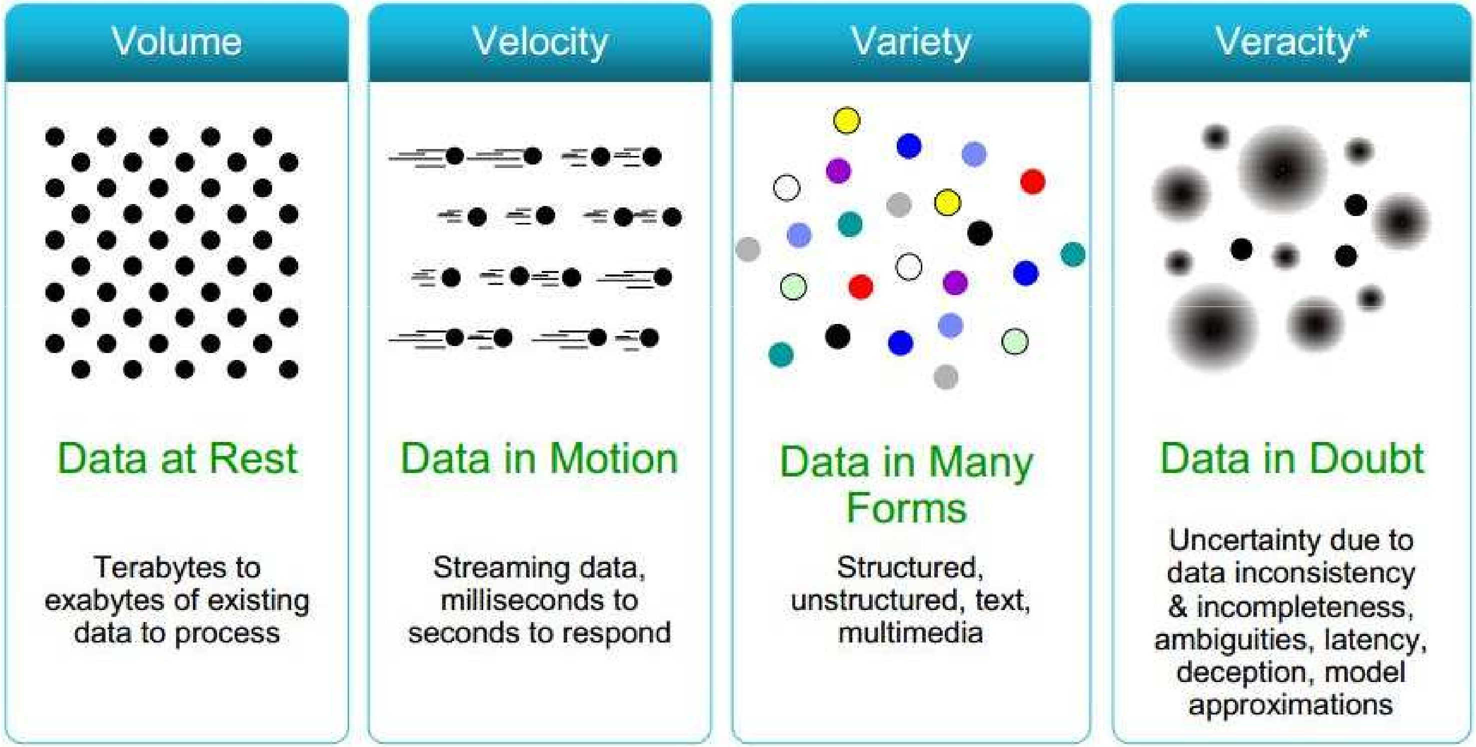
Customer

Improving the Marketing Effectiveness of a Promotion while it is still in Play

Friend Invitations to join a Game or Activity that expands business

Preventing Fraud as it is ***Occurring*** & preventing more proactively

# 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

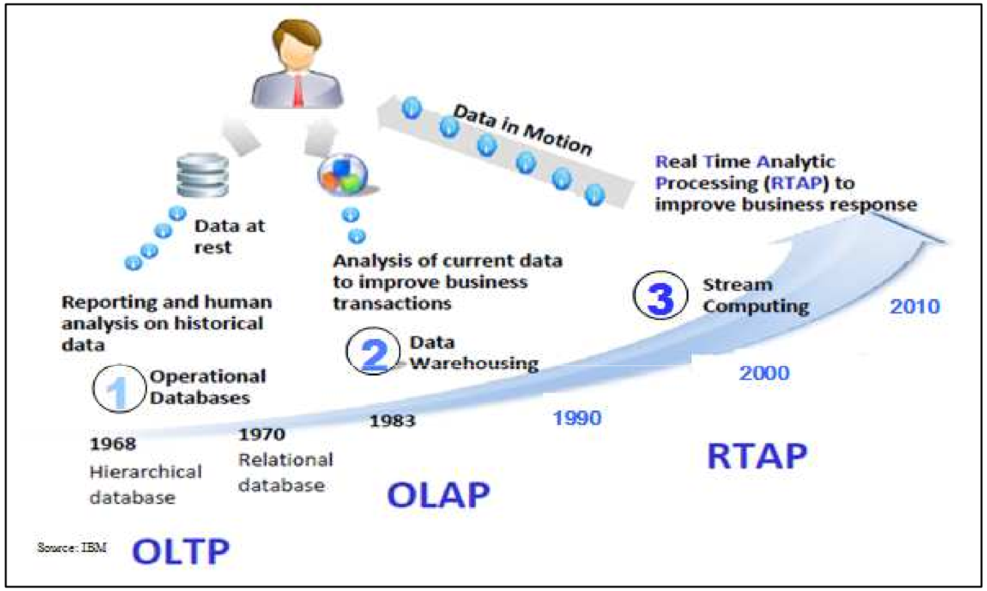
Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

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

# Harnessing Big Data


Data in Motion
Real Time Analytic Processing (RTAP) to improve business response
Data at rest
Analysis of current data to improve business transactions
Reporting and human analysis on historical data
3 Stream Computing
2010
2 Data Warehousing
2000
1 Operational Databases
1983
1990
1968
1970
Hierarchical database
Relational database
RTAP
OLAP
OLTP

Source: IBM

- **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

# How is Big Data actually used? Example 1

## Better understand and target customers:

To better understand and target customers, companies expand their traditional data sets with social media data, browser, text analytics or sensor data to get a more complete picture of their customers. The big objective, in many cases, is to create predictive models. Using big data, Telecom companies can now better predict customer churn; retailers can predict what products will sell, and car insurance companies understand how well their customers actually drive.

# How is Big Data actually used? Example 2

## Understand and Optimize Business Processes:

Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictive models generated from social media data, web search trends and weather forecasts. Another example is supply chain or delivery route optimization using data from geographic positioning and radio frequency identification sensors.

# How is Big Data actually used? Example 3

## Improving Health:

The computing power of big data analytics enables us to find new cures and better understand and predict disease patterns. We can use all the data from smart watches and wearable devices to better understand links between lifestyles and diseases. Big data analytics also allow us to monitor and predict epidemics and disease outbreaks, simply by listening to what people are saying, i.e. “Feeling rubbish today - in bed with a cold” or searching for on the Internet, i.e. “cures for flu”.

# How is Big Data actually used? Example 4

## Improving Security and Law Enforcement:

Security services use big data analytics to foil terrorist plots and detect cyber attacks. Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data analytics it to detect fraudulent transactions.

# How is Big Data actually used? Example 5

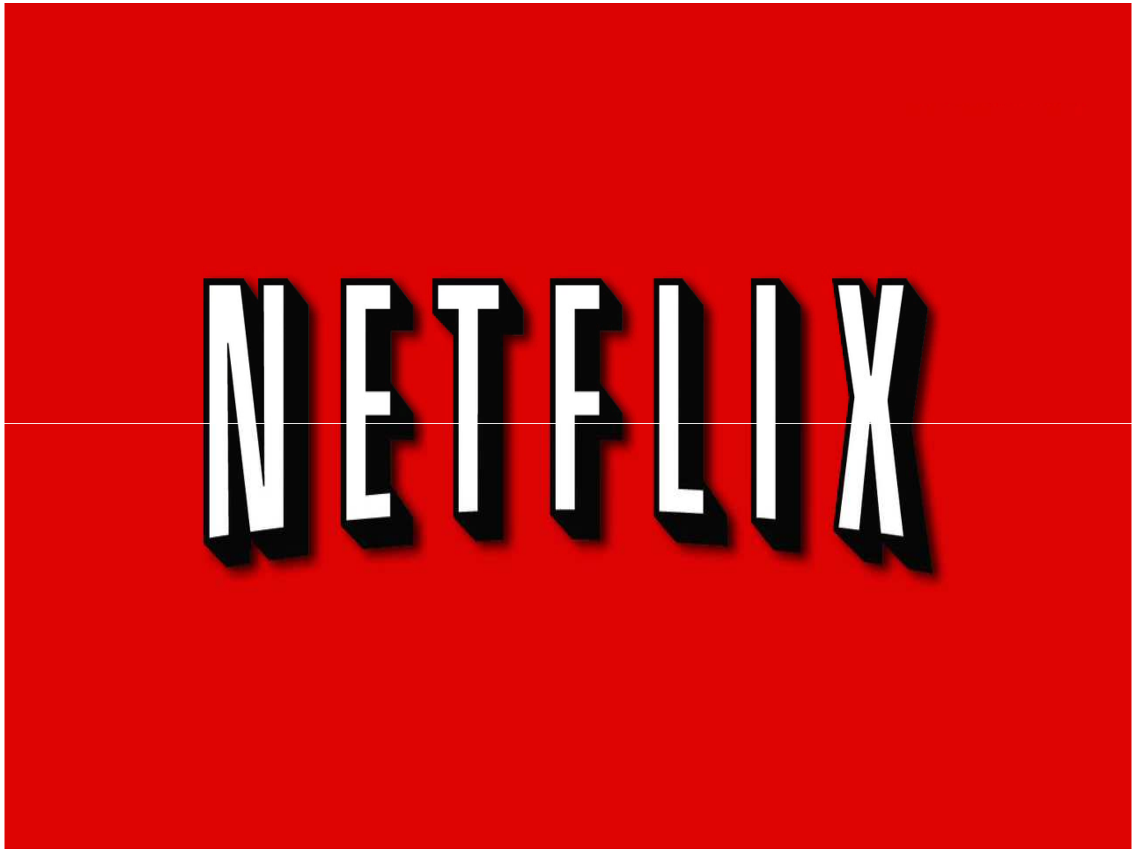
## Improving Sports Performance:

Most elite sports have now embraced big data analytics. Many use video analytics to track the performance of every player in a football or baseball game, sensor technology is built into sports equipment such as basket balls or golf clubs, and many elite sports teams track athletes outside of the sporting environment – using smart technology to track nutrition and sleep, as well as social media conversations to monitor emotional wellbeing.

# How is Big Data actually used? Example 6

## Improving and Optimizing Cities and Countries:

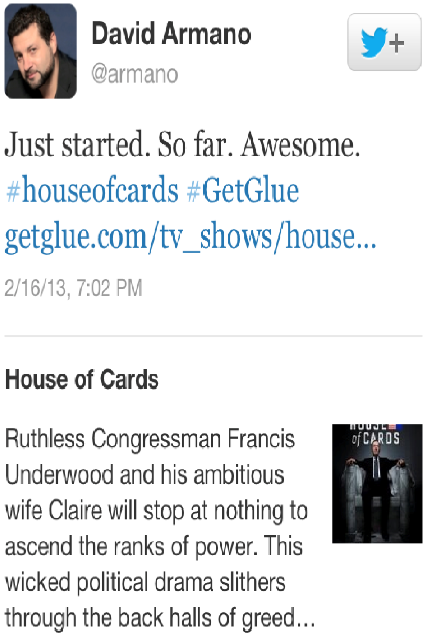
Big data is used to improve many aspects of our cities and countries. For example, it allows cities to optimize traffic flows based on real time traffic information as well as social media and weather data. A number of cities are currently using big data analytics with the aim of turning themselves into Smart Cities, where the transport infrastructure and utility processes are all joined up. Where a bus would wait for a delayed train and where traffic signals predict traffic volumes and operate to minimize jams.



NETFLIX

A NETFLIX ORIGINAL SERIES
HOUSE of CARDS

*"House of Cards" is one of the first major test cases of this Big Data-driven creative strategy. For almost a year, Netflix executives have told us that their detailed knowledge of Netflix subscriber viewing preferences clinched their decision to license a remake of the popular and critically well regarded 1990 BBC miniseries. Netflix's data indicated that the same subscribers who loved the original BBC production also gobbled down movies starring Kevin Spacey or directed by David Fincher. Therefore, concluded Netflix executives, a remake of the BBC drama with Spacey and Fincher attached was a no-brainer, to the point that the company committed $100 million for two 13-episode seasons.*



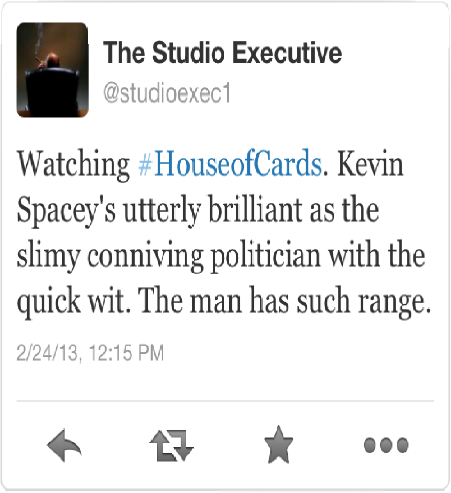
**David Armano**
@armano

Just started. So far. Awesome. #houseofcards #GetGlue getglue.com/tv_shows/house...

2/16/13, 7:02 PM

**House of Cards**

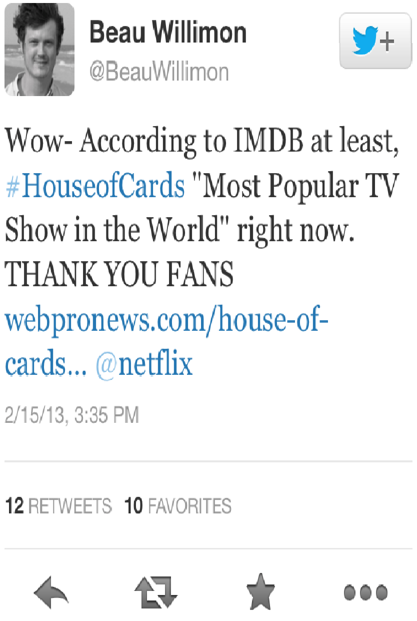
Ruthless Congressman Francis Underwood and his ambitious wife Claire will stop at nothing to ascend the ranks of power. This wicked political drama slithers through the back halls of greed...



**The Studio Executive**
@studioexec1

Watching #HouseofCards. Kevin Spacey's utterly brilliant as the slimy conniving politician with the quick wit. The man has such range.

2/24/13, 12:15 PM

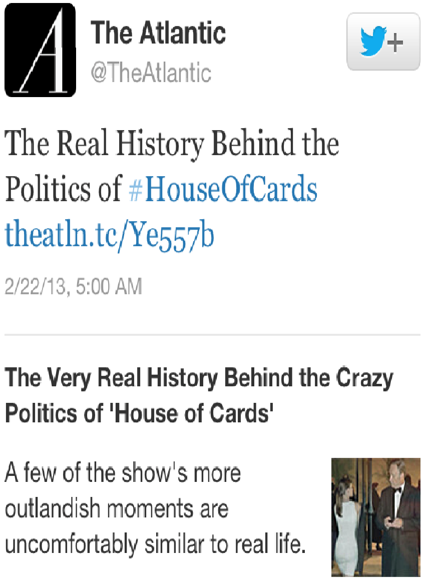


**Beau Willimon**
@BeauWillimon

Wow- According to IMDB at least, #HouseofCards "Most Popular TV Show in the World" right now. THANK YOU FANS webpronews.com/house-of-cards... @netflix

2/15/13, 3:35 PM

12 RETWEETS 10 FAVORITES



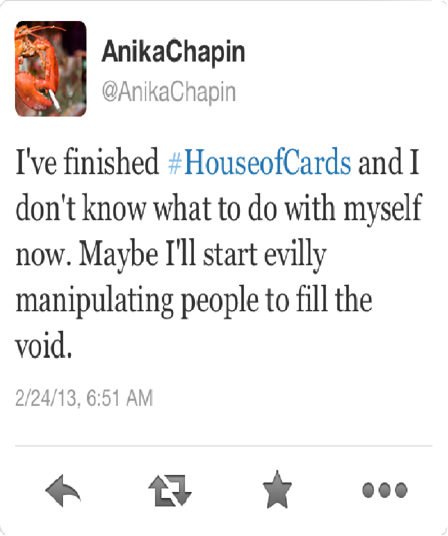
**The Atlantic**
@TheAtlantic

The Real History Behind the Politics of #HouseOfCards theatln.tc/Ye557b

2/22/13, 5:00 AM

**The Very Real History Behind the Crazy Politics of 'House of Cards'**

A few of the show's more outlandish moments are uncomfortably similar to real life.



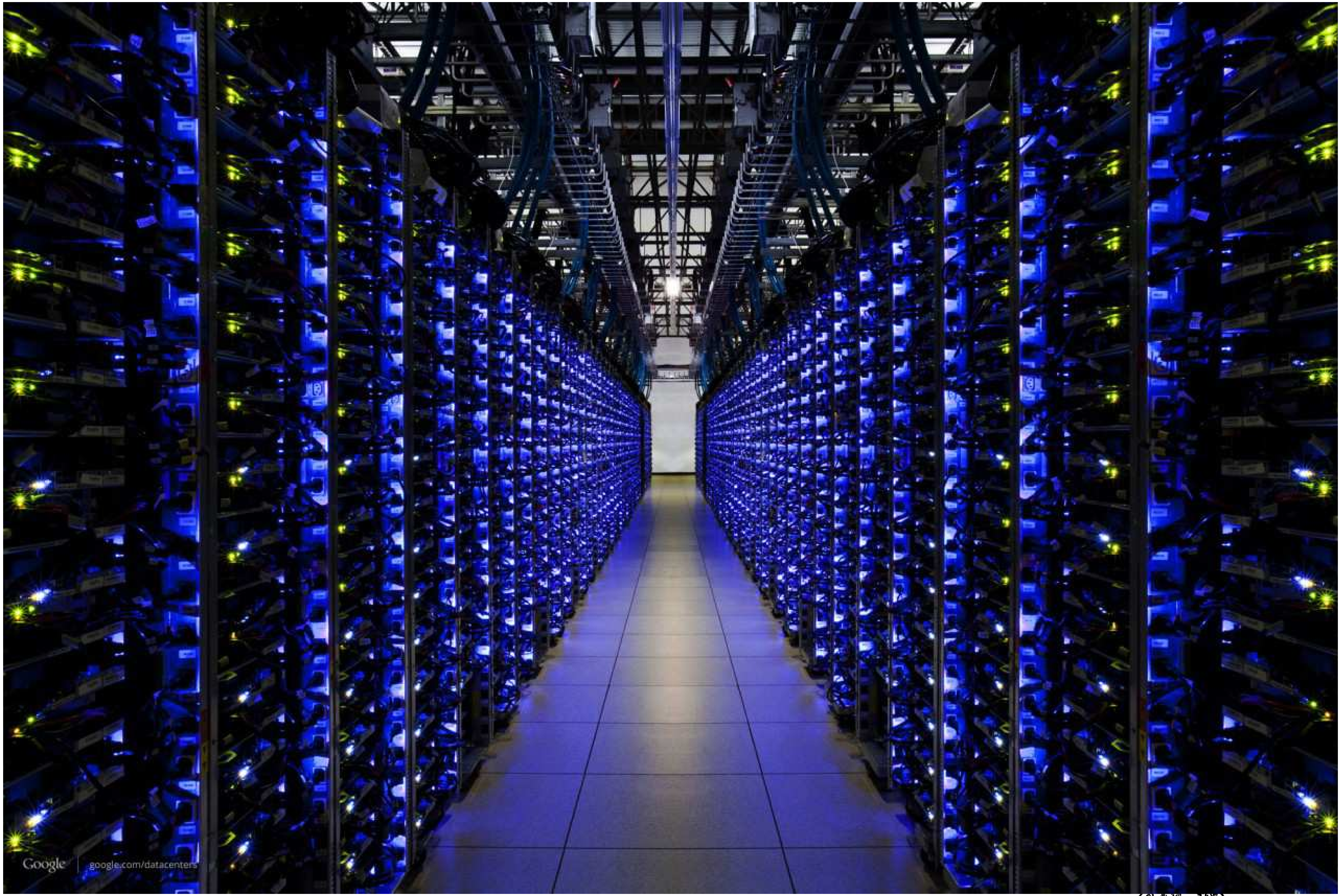
**AnikaChapin**
@AnikaChapin

I've finished #HouseofCards and I don't know what to do with myself now. Maybe I'll start evilly manipulating people to fill the void.

2/24/13, 6:51 AM

# Πώς;

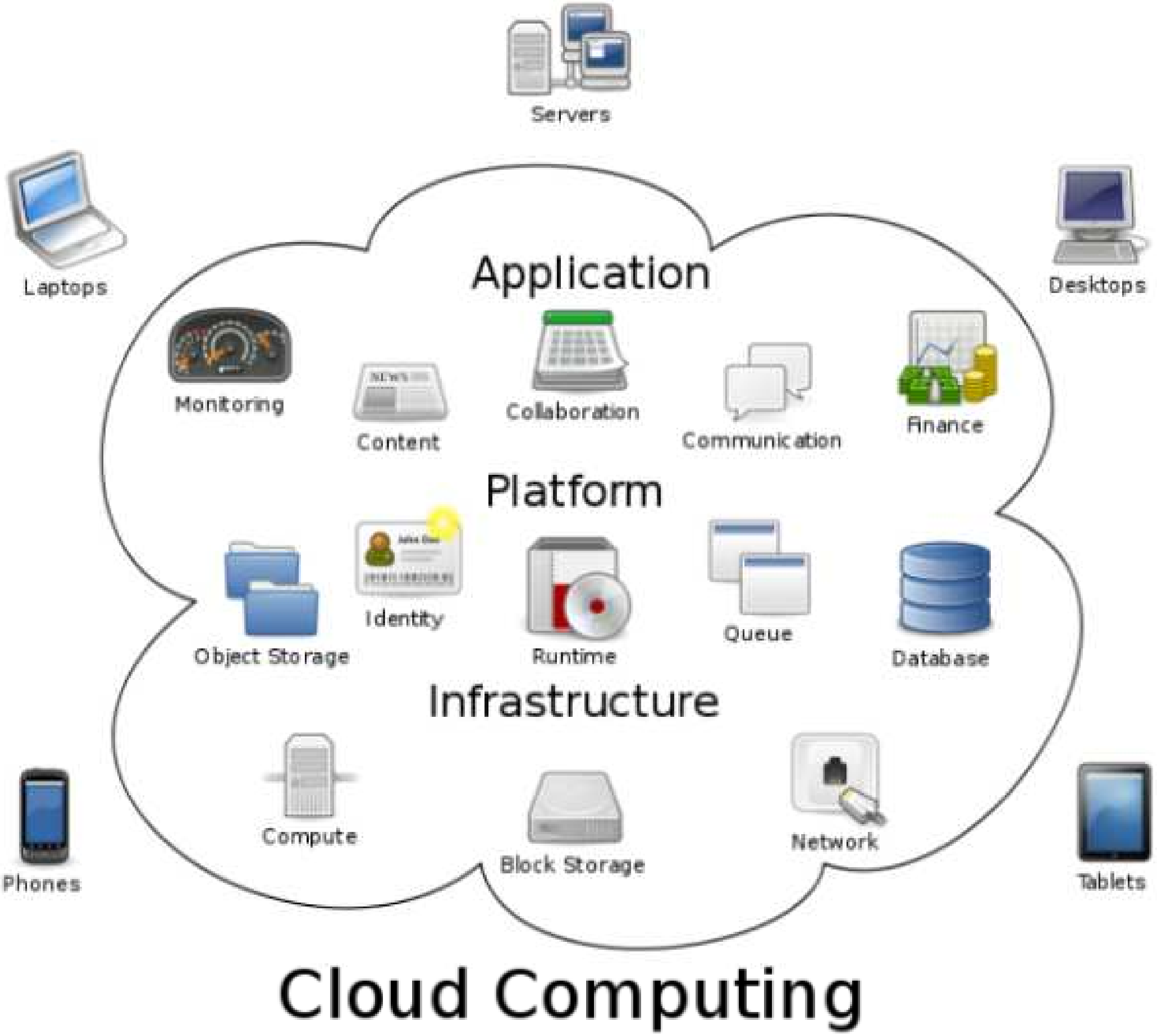
Κλιμακωσιμότητα



Source: Wikipedia (IBM Roadrunner)

# 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, …



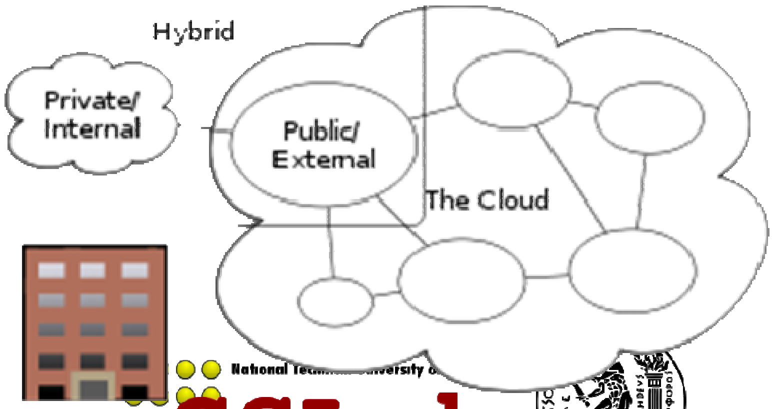
wikipedia:Cloud Computing

# 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

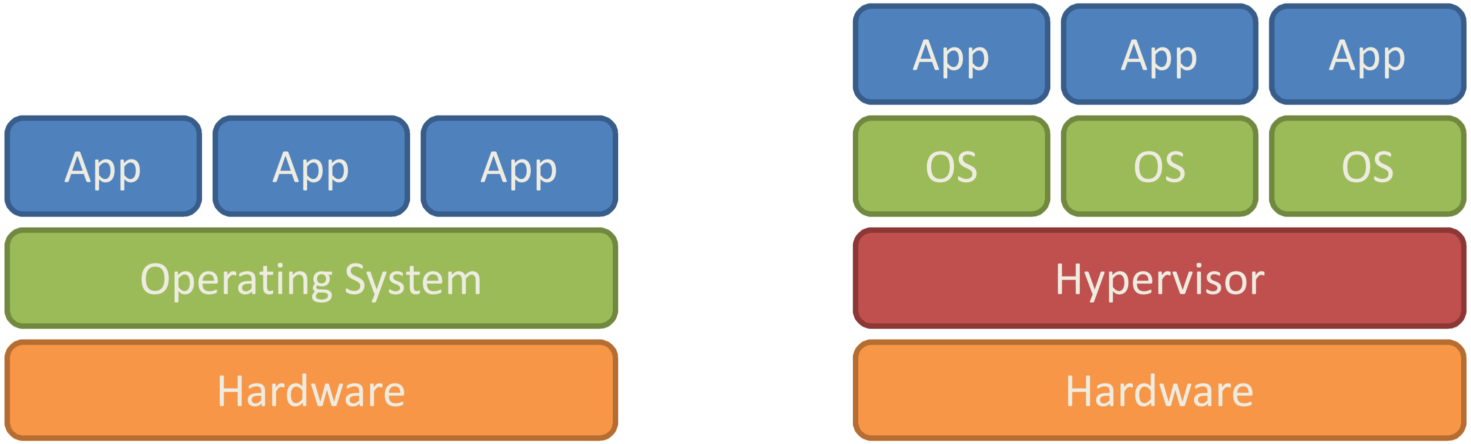
- **Public Cloud**: Computing infrastructure is hosted at the vendor’s premises.
- **Private Cloud**: Computing architecture is dedicated to the customer and is not shared with other organisations.
- **Hybrid Cloud**: Organisations host some critical, secure applications in private clouds. The not so 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.



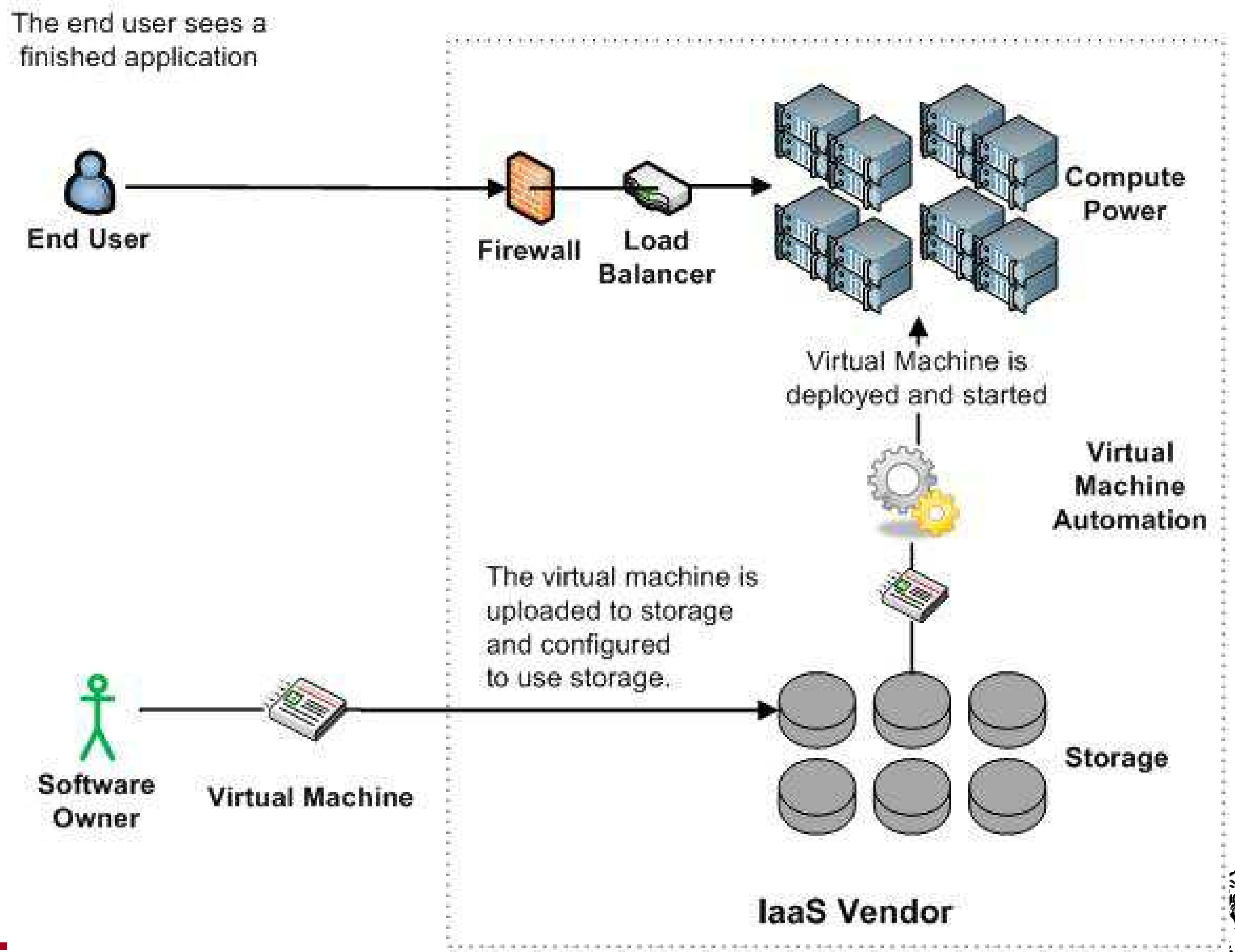
# Classification of Cloud Computing based on Service Provided

- 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.

# Enabling Technology: Virtualization



# Infrastructure as a Service (IaaS)


The end user sees a finished application
End User
Firewall
Load Balancer
Compute Power
Virtual Machine is deployed and started
Virtual Machine Automation
The virtual machine is uploaded to storage and configured to use storage.
Software Owner
Virtual Machine
Storage
IaaS Vendor