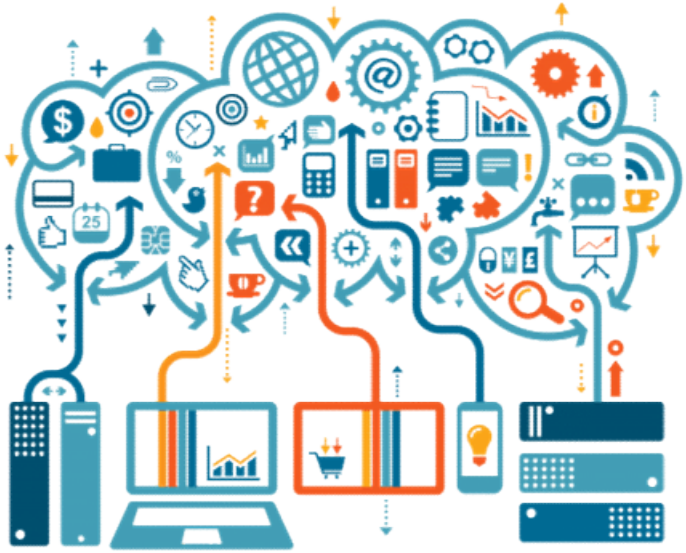
Big Data Storage and Management

(CDS502)



Course Lecturers:
Dr. Mohd. Adib Haji Omar
Ts. Dr. Chew XinYing

Course Lecturer

- Senior Lecturer, School of Computer Sciences, USM
- Academic Qualification:
  - BIT. (Hons.), UKM
  - PhD., USM
- Specialization :
  - Statistical Quality Control, Advanced Analytics
- Certified trainer with Human Resources Development Fund (HRDF).
- Instructor for Data Science Certification Program (Industry).
- Trainer for MDEC-Intel AI Academy Program.
- Trainer for SAP Next-Gen Program.
- Previously worked in Advanced Analytics Research Team.
- Adjunct Research Fellow of Swinburne University of Technology, Sarawak Campus.
- Professional Technologist (PT19100021), Malaysia Board of Technologists (MBOT).
- Certified Scrum Master, Scrum Alliance.

Ts. Dr. Chew XinYing

https://cs.usm.my/index.php/faculty-member/174-chew-xin-ying-dr

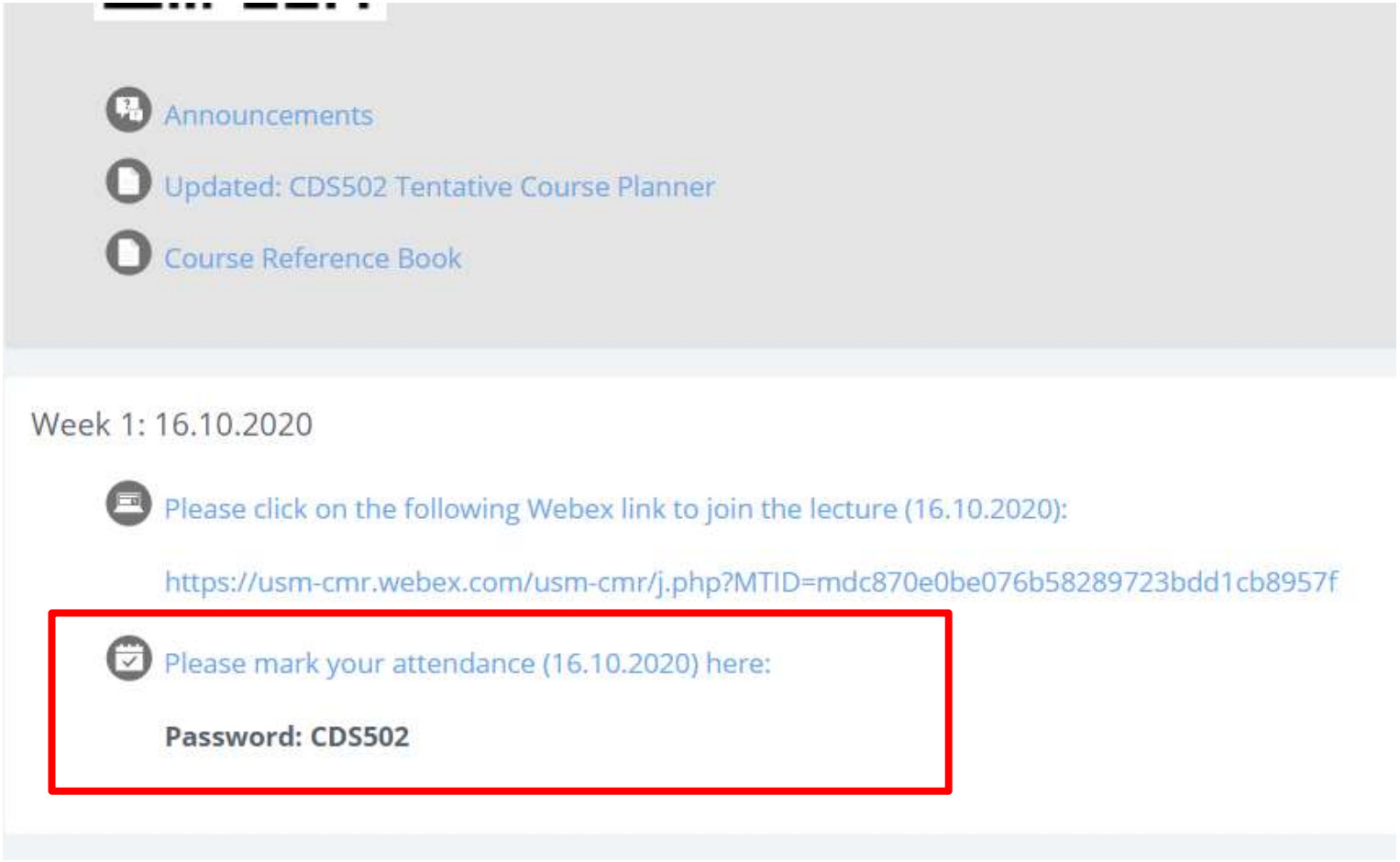
Course Lecturer

Dr. Mohd Adib Haji Omar

- Senior Lecturer, School of Computer Sciences, USM

https://cs.usm.my/index.php/faculty-member/186-mohd-adib-haji-omar-dr

Mark your Attendance



WhatsApp Group (CDS502 Sem 1 2021/2022)

Click on the link / Scan the QR code to join the WhatsApp Group for CDS502 Sem 1, 2021/2022:
https://bit.ly/502_21_22

Lets Get to Know Each Other

Introduction

- Photo (Breakout Sessions)
- Name
- Previous field of study
- Programming experience
  - Hadoop experience?
  - Python experience?
  - SQL experience?
  - Other programming experience?
- Experience on Big Data (If any)
- Industrial Experience / Position (If any)
- Fun facts about yourself

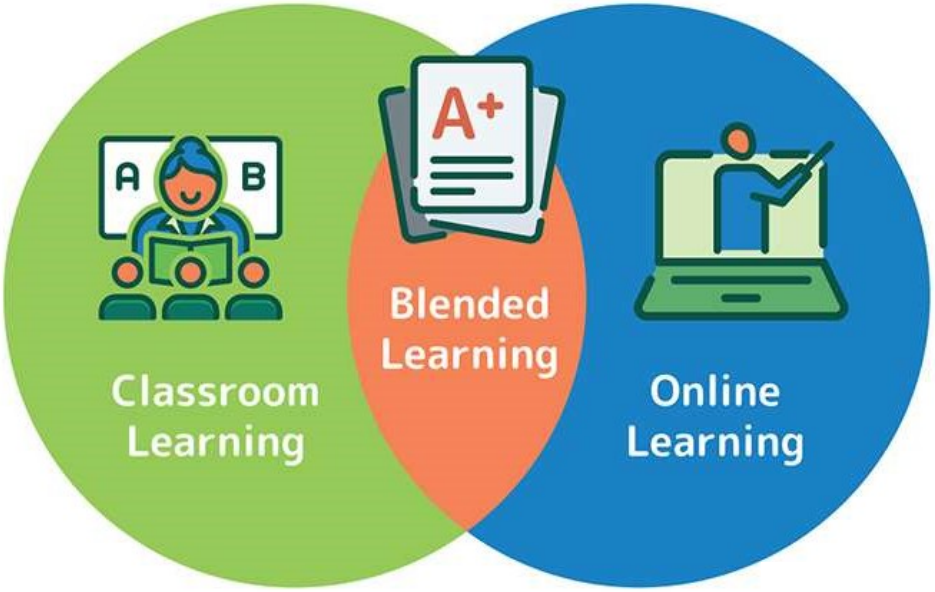
Post on Padlet Wall

https://padlet.com/XYChew/cds502

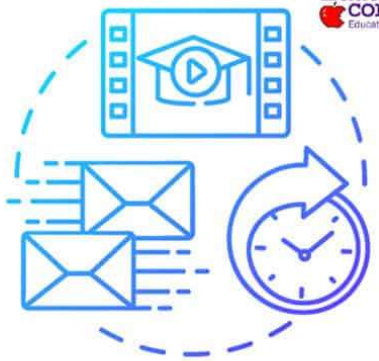
Planner Walkthrough

Course Syllabus & Planner

Blended Learning



SYNCHRONOUS LEARNING



ASYNCHRONOUS LEARNING

Fill the Survey:

shorturl.at/fxDEV

Google Docs

Intel AIAcademy

USM
We lead

Intel® AI DevCloud

Course Title & Code: Big Data Storage and Management (CDS502)

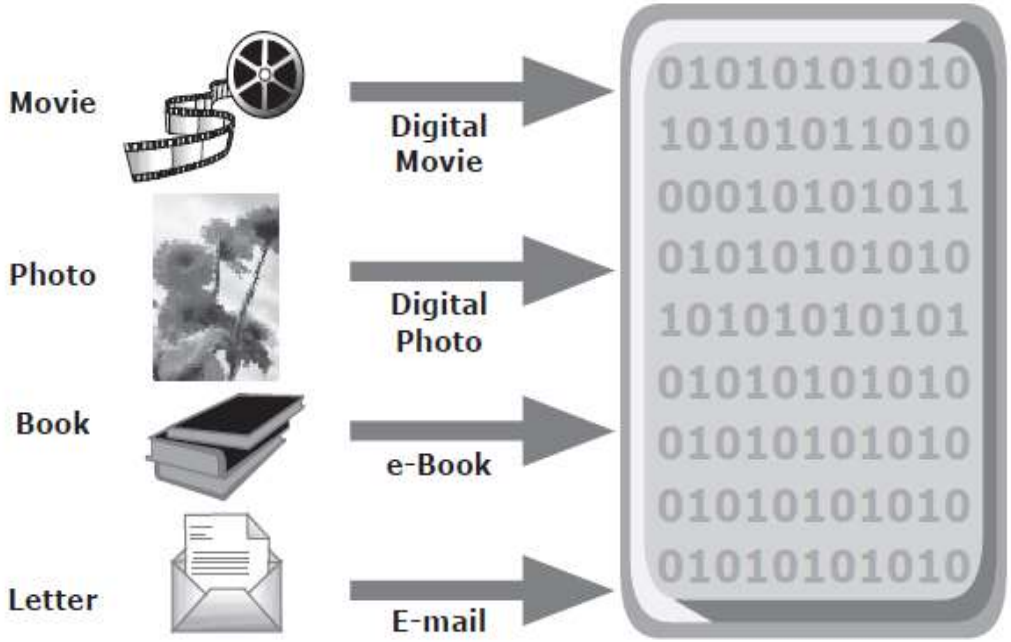
Introduction to Big Data Storage and Management

Course Lecturers:
Dr. Mohd. Adib Haji Omar
Ts. Dr. Chew XinYing

Data and Information

- Data is a collection of raw facts from which conclusions might be drawn.
- Past: paper and film.
- Current: e-mail message, e-book, digital image, digital movie.
- Data can be generated using a computer and stored as strings of binary numbers (0s and 1s).

Digital data: Accessible by the user only after a computer processes it.



Factors Contributed to the Growth of Digital Data

- **Increase in data-processing capabilities:** Modern computers provide a significant increase in processing and storage capabilities.
Convert various types of content and media: Conventional → Digital formats.

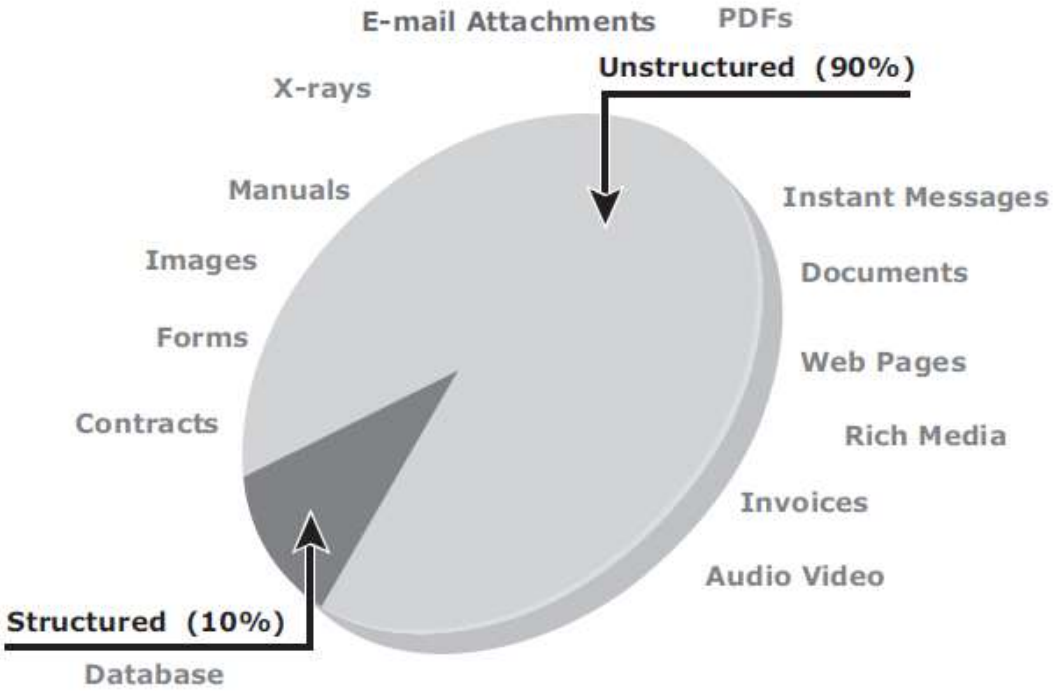
- **Lower cost of digital storage:** Technological advances and the decrease in the cost of storage devices → low-cost storage solutions. This cost benefit has increased the rate at which digital data is generated and stored.

- **Affordable and faster communication technology:** The rate of sharing digital data is much faster than traditional approaches.
Handwritten: need weeks to reach destination; email: a few seconds to recipient.

- **Proliferation of applications and smart devices:** Smartphones, tablets, and newer digital devices, along with smart applications, have significantly contributed to the generation of digital content.

Types of Data

- Data → structured or unstructured based on how it is stored and managed
- Structured data:
  - Organized in rows and columns in a rigidly defined format
  - Applications can retrieve and process it efficiently
  - Typically stored using a database management system (DBMS)
- Unstructured Data:
  - Its elements cannot be stored in rows and columns
  - Difficult to query and retrieve by applications



Evolution of Storage Architecture

Centralized computers (mainframes) and **information storage devices** (tape reels and disk packs)

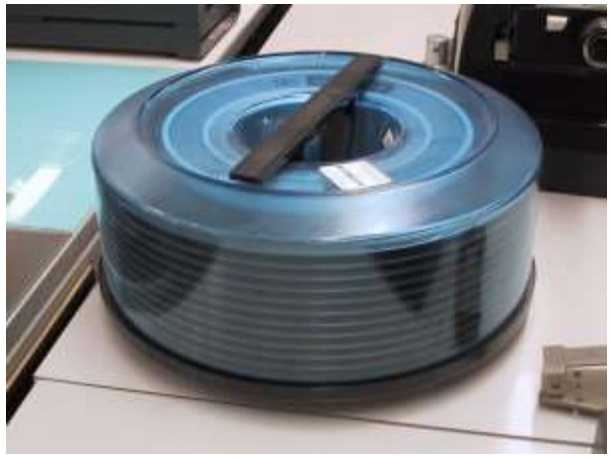
↓

Open systems:

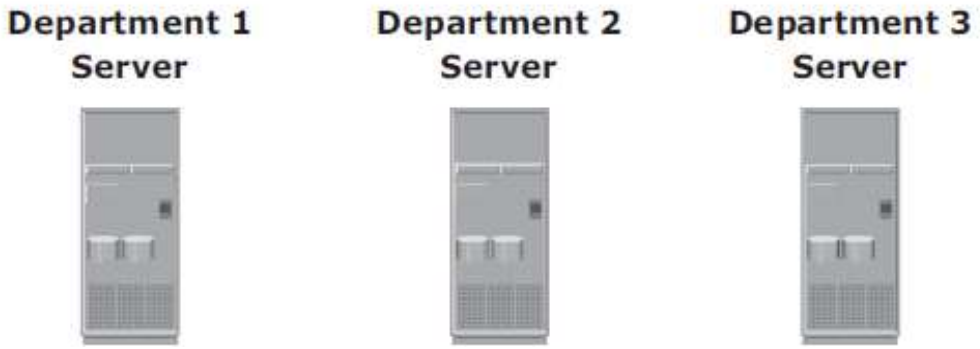
- Affordability, and ease of deployment → business units / departments have their own servers and storage
- Storage was typically internal to the server, storage devices could not be shared with any other servers
- Referred to as **server-centric storage architecture**

↓

Storage evolved from server-centric to **information-centric architecture**



Server-Centric Storage Architecture



- Each server has a **limited** number of storage devices.

- Any administrative tasks, such as **maintenance** of the server or increasing storage capacity, might result in **unavailability** of information.

- The **proliferation** of departmental servers in an enterprise resulted in unprotected, unmanaged, fragmented islands of information and increased capital and operating expenses.

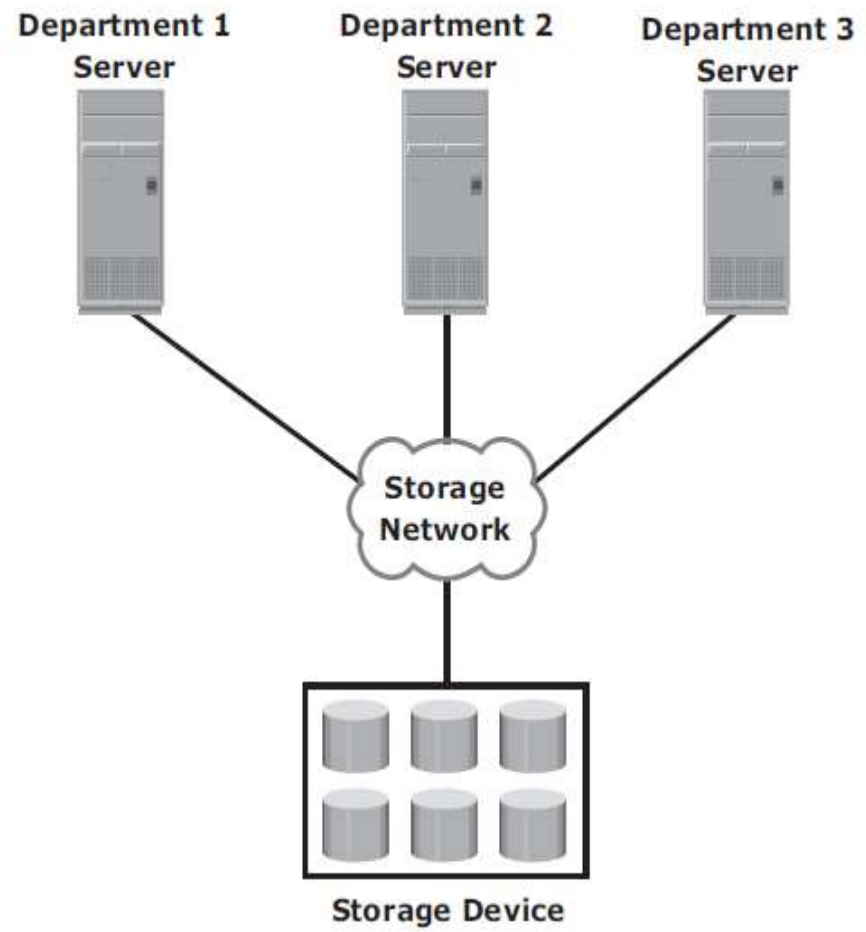
Information-Centric Storage Architecture

• Storage devices are managed **centrally** and **independent** of servers.

• Centrally-managed storage devices are **shared** with multiple servers.

• The capacity of shared storage can be increased dynamically by adding more storage devices **without** impacting information availability.

• Information management is **easier** and **cost-effective**.



Core Elements of a Data Center

Five core elements are essential for the functionality of a data center:

- **Application**: A computer program that provides the logic for computing operations.

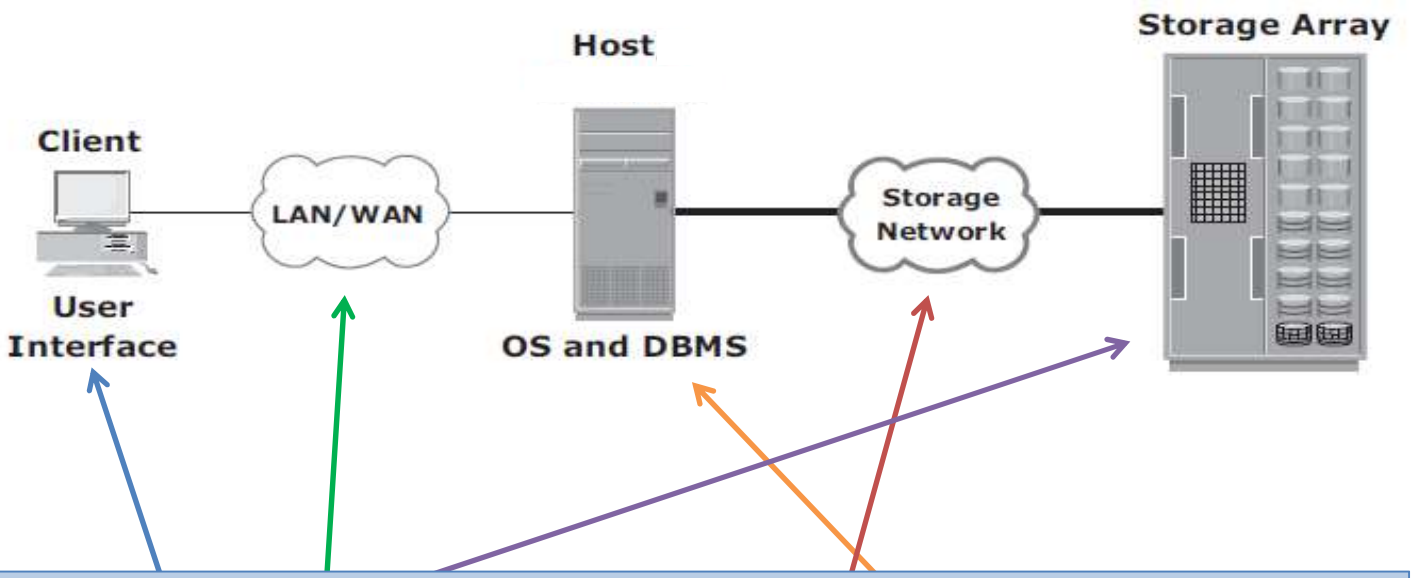
- **Database management system (DBMS)**: Provides a structured way to store data in logically organized tables that are interrelated.

- **Host**: A computing platform (hardware, firmware, and software) that runs applications and databases.

- **Network**: A data path that facilitates communication among various networked devices.

- **Storage**: A device that stores data persistently for subsequent use.

Example of an Online Order Transaction System



A customer order is entered via the Application User Interface on a client.

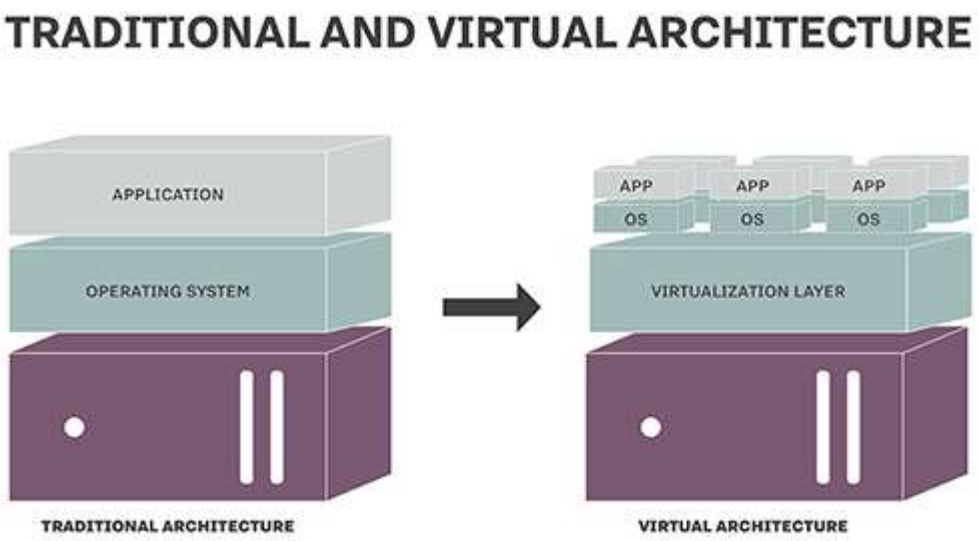
physical disks.

Virtualization

• Virtualization is the ability to run multiple operating systems on a single physical system and share the underlying hardware resources.

• It is the process by which one computer hosts the appearance of many computers.

• Virtualization is used to improve IT throughput and costs by using physical resources as a pool from which virtual resources can be allocated.



Cloud Computing

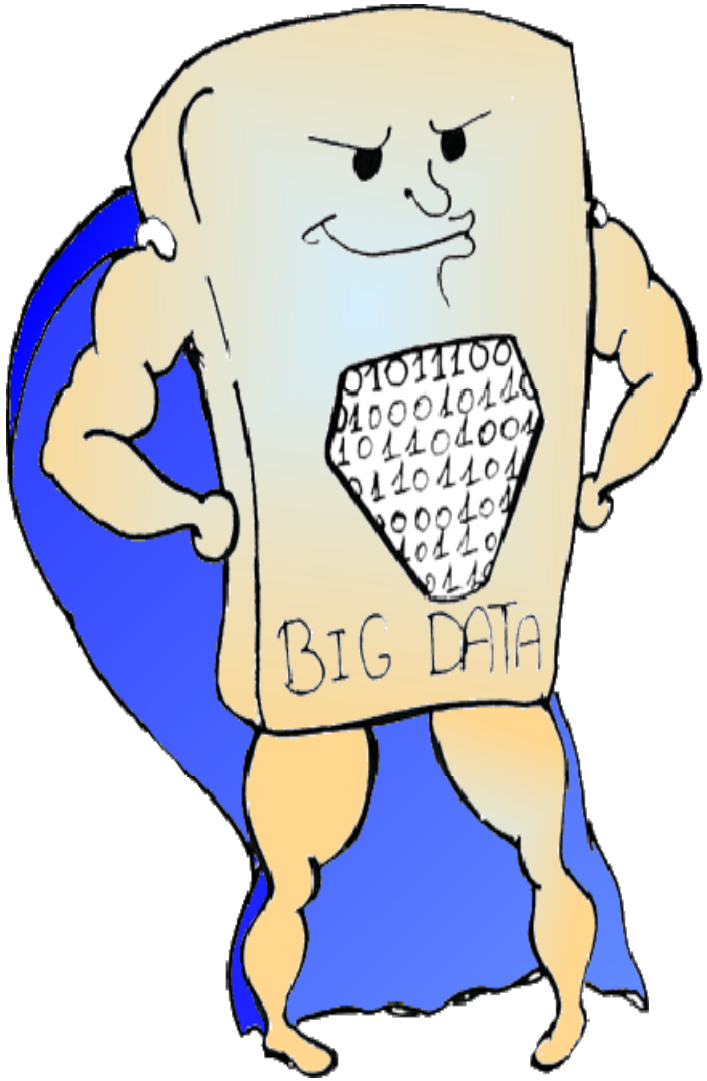
- Cloud computing enables individuals or businesses to use IT resources as a service over the **network**.

- Provides highly **scalable** and **flexible** computing that enables provisioning of resources on demand.

- Users can scale up or scale down the demand of computing resources, including storage capacity, with **minimal** management effort or service provider interaction.

- Cloud computing empowers self-service requesting through a fully automated request - **fulfilment process**.

- Cloud infrastructure is usually built upon **virtualized** data centers, which provide resource pooling and rapid provisioning of resources.



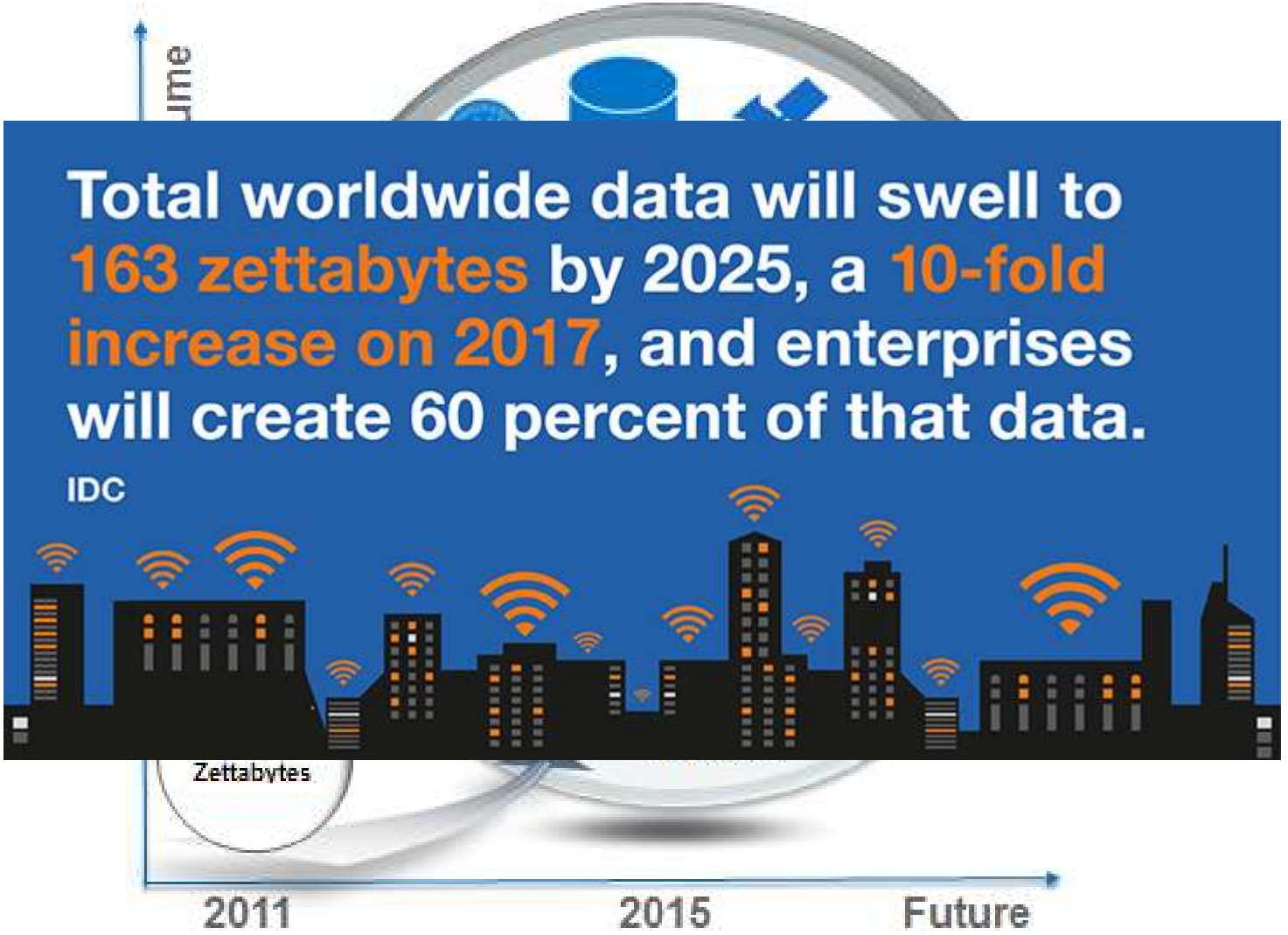
What is **Big Data**?

Defining Big Data

Big Data is data whose **scale, distribution, diversity**, and/or **timeliness** require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.

- ✓Requires new architectures
- ✓New tools
- ✓New analytical methods
- ✓Massively parallel processing (MPP) data platforms
- ✓Requires multiple skills into the role of data scientist

Big Data is Everywhere



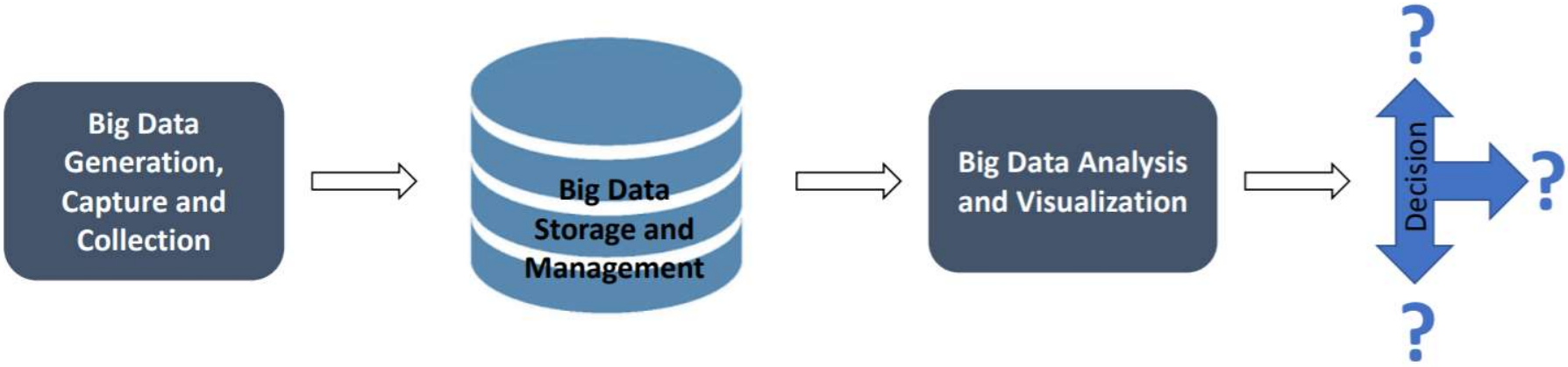
Four Domains of Big Data in 2025

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLOS Biology 13(7): e1002195. https://doi.org/10.1371/journal.pbio.1002195

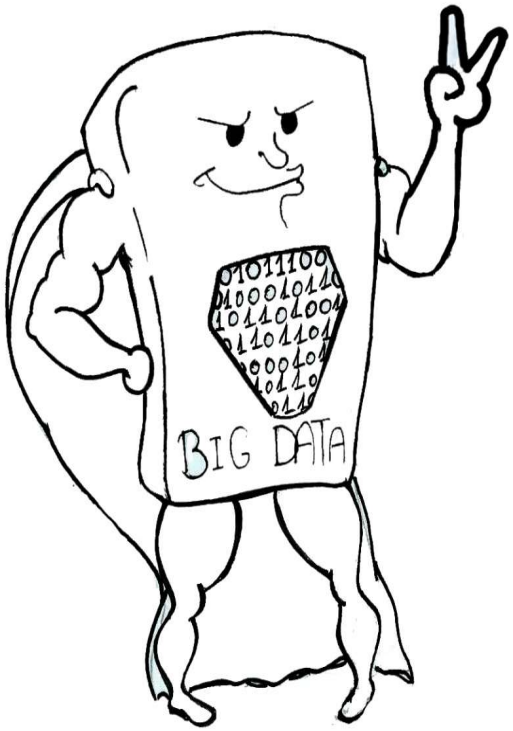
Main Phases of a Big Data Process

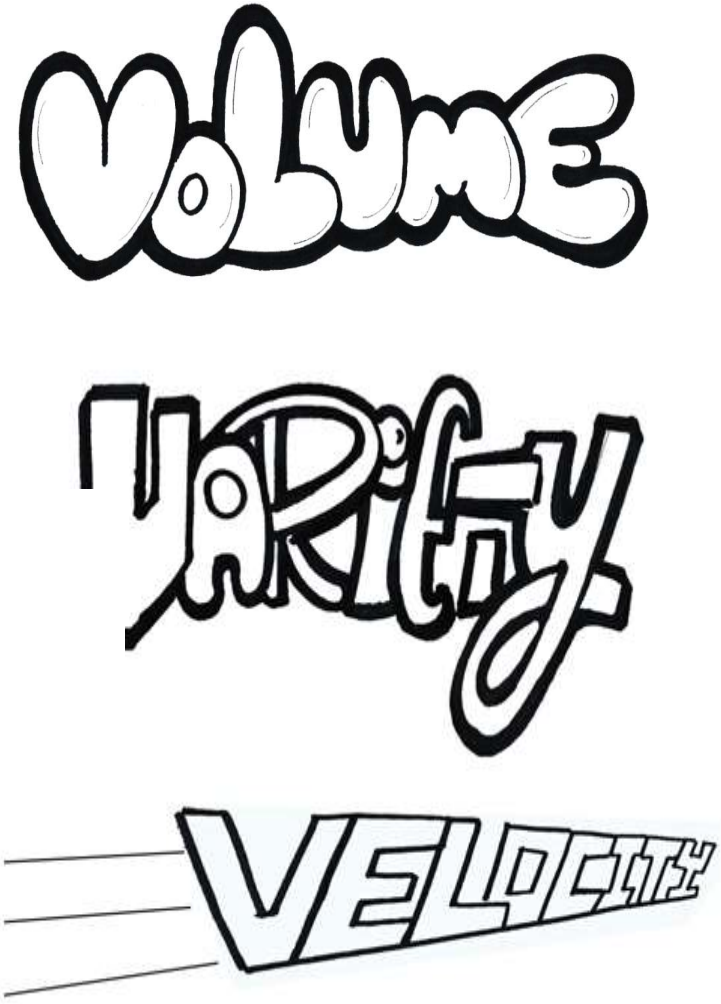


Dhaenens, C., & Jourdan, L. (2016). Metaheuristics for big data. ISTE.

The “Three Vs” of Big Data

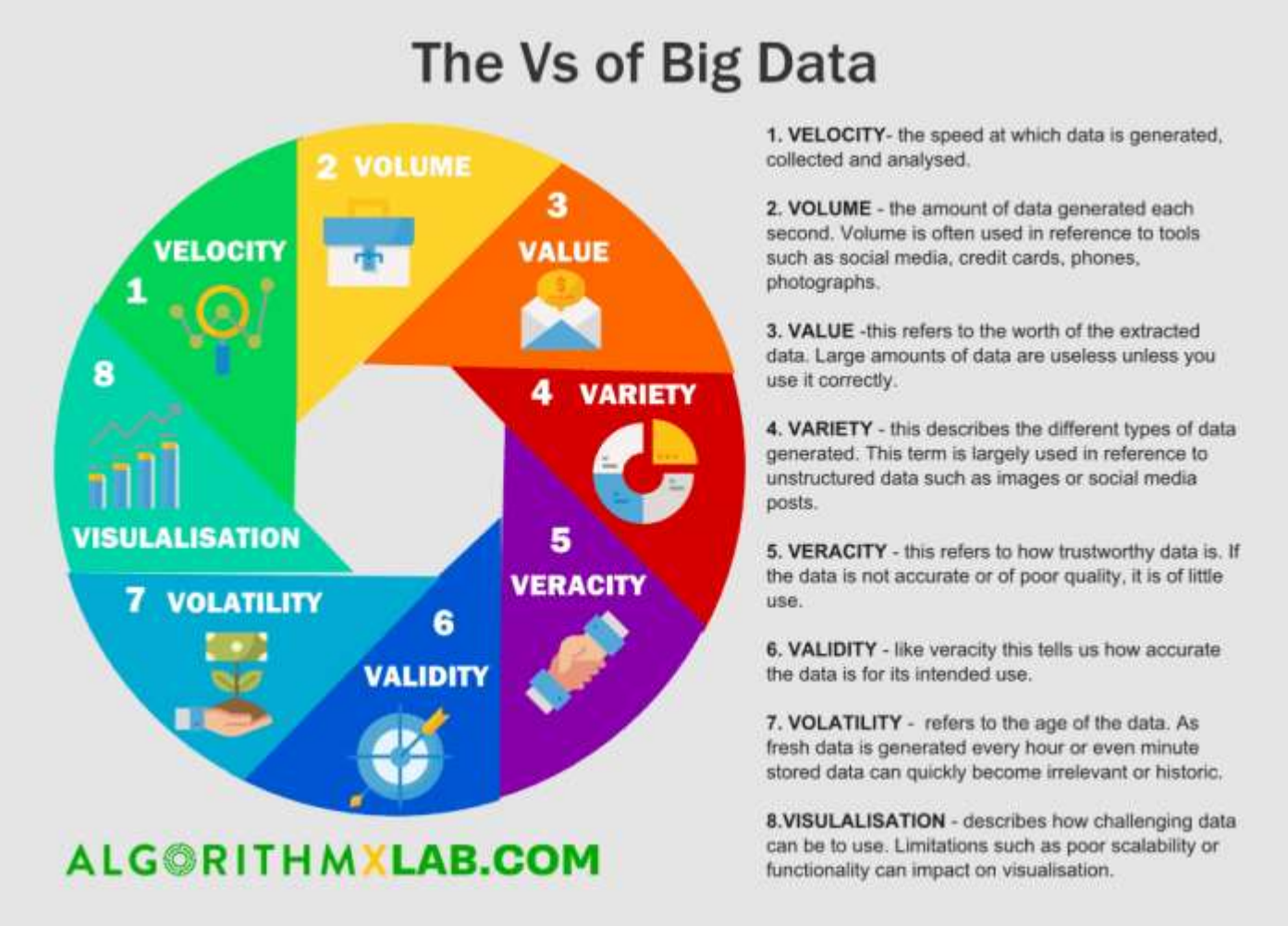
In 2001, industry analyst Doug Laney defined the “Three Vs” of big data





Beyond the Big Three Vs

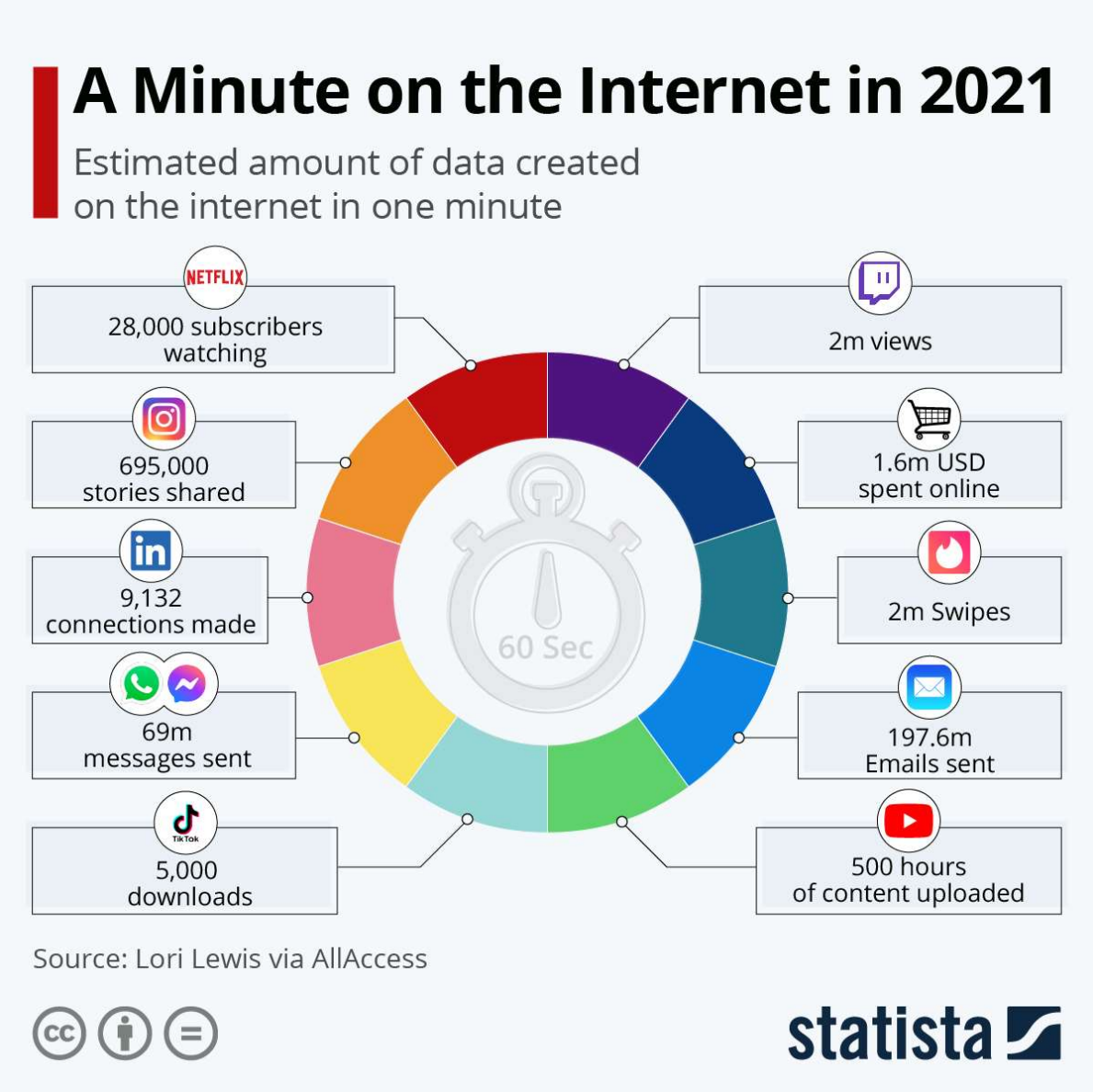
More recently, big-data practitioners and thought leaders have proposed additional Vs:



https://algorithmxlab.com/blog/big-data/

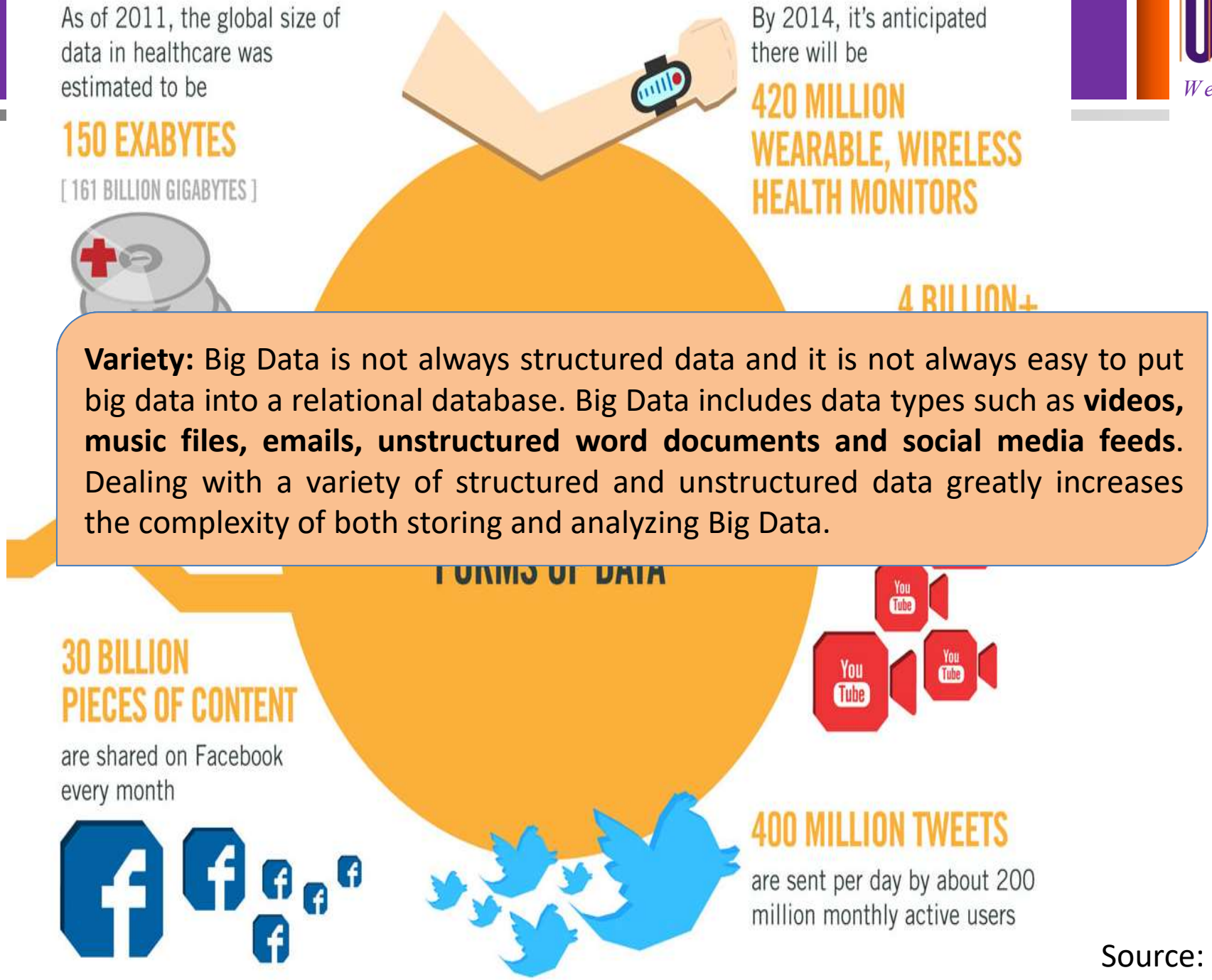
How much Data is Generated Every Minutes?

Volume



Volu
of B
not
anal

hen they think
ent times. It is
, and typically



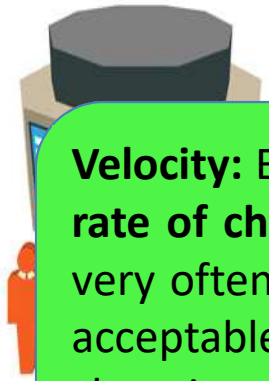
Variety: Big Data is not always structured data and it is not always easy to put big data into a relational database. Big Data includes data types such as **videos, music files, emails, unstructured word documents and social media feeds**. Dealing with a variety of structured and unstructured data greatly increases the complexity of both storing and analyzing Big Data.

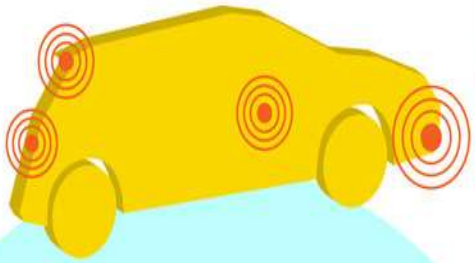
Source: IBM

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session





Modern cars have close to

100 SENSORS

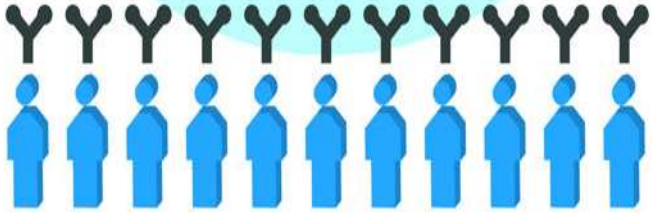
that monitor items such as fuel level and tire pressure

Velocity: Big Data is not just about the volume though. Just as important is the **rate of change of the data**. For a large volume of data which doesn't change very often, analysis that takes a number of hours or days to complete may be acceptable, but if the dataset is growing by terabytes per day, or the data is changing at a high rate of speed, the processing time of analysis becomes much more important.

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

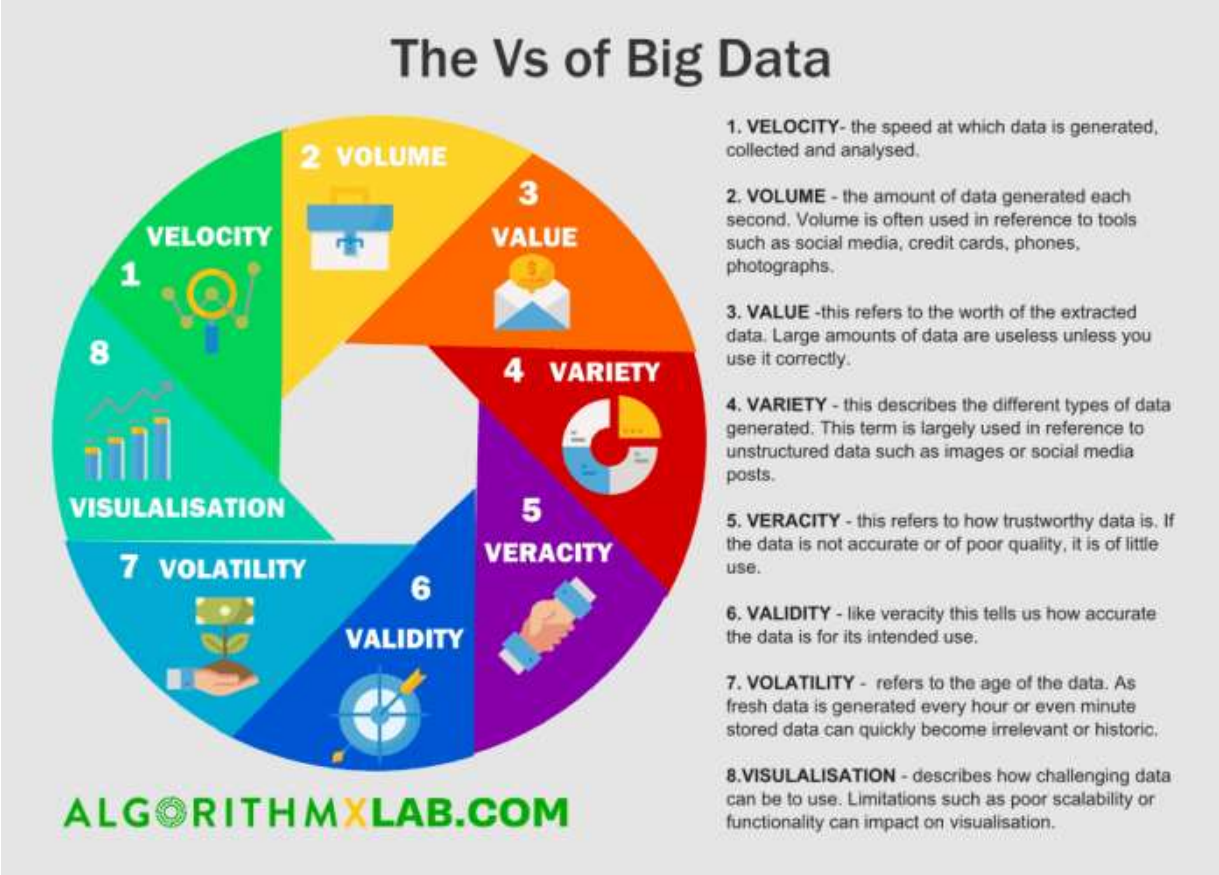
– almost 2.5 connections per person on earth



Source: IBM

Beyond the Big Three Vs

No matter how many Vs you prefer in your big data, one thing is sure: Big data is here, and it's only getting bigger. Every organization needs to understand what big data means to them and what it can help them do. The possibilities really are endless.



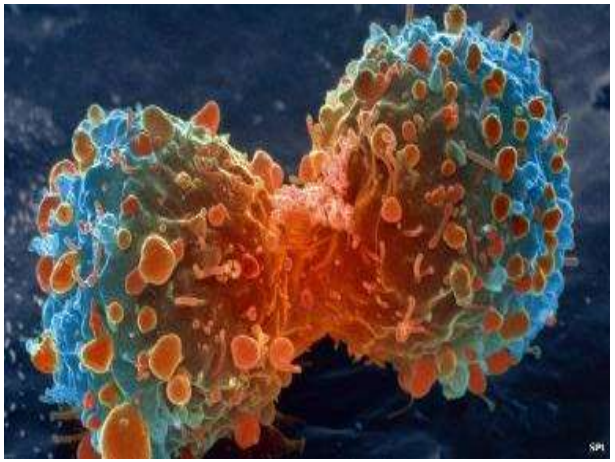
https://algorithmxlab.com/blog/big-data/

BIG DATA EXAMPLES

Computer Scientists May Have What It Takes to Help Cure Cancer

USM
We lead

- AMP Technology
- Goal: Sequencing the genome of cancer tumors
  - Understanding mutations and changes in DNA that generate diversity in cancer tumors
  - To be able to prescribe therapy and personalized medicines based on the genetics of the patient
- 20,000 genomes can be stored in 5 Petabytes

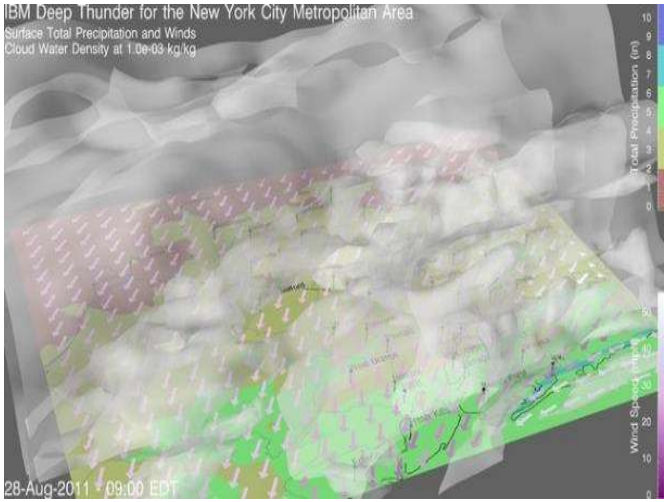


How Big Data is Beating Ebola

- Goal: Model the way the virus spreads
  - Create synthetic population models:
    - Demographics, family structures, travel patterns, activities, etc
    - Based on actual data
- Technology:
  - High performance computing (HPC) system, Shadowfax
  - Cluster: 2500 cores and 1 Petabyte storage
  - Running a simulation for all US takes 12 seconds
  - 7 billion people → take around 6 minutes
  - Ten years ago it would take us over an hour to simulate just one city

How Big Data Can Boost Weather Forecasting

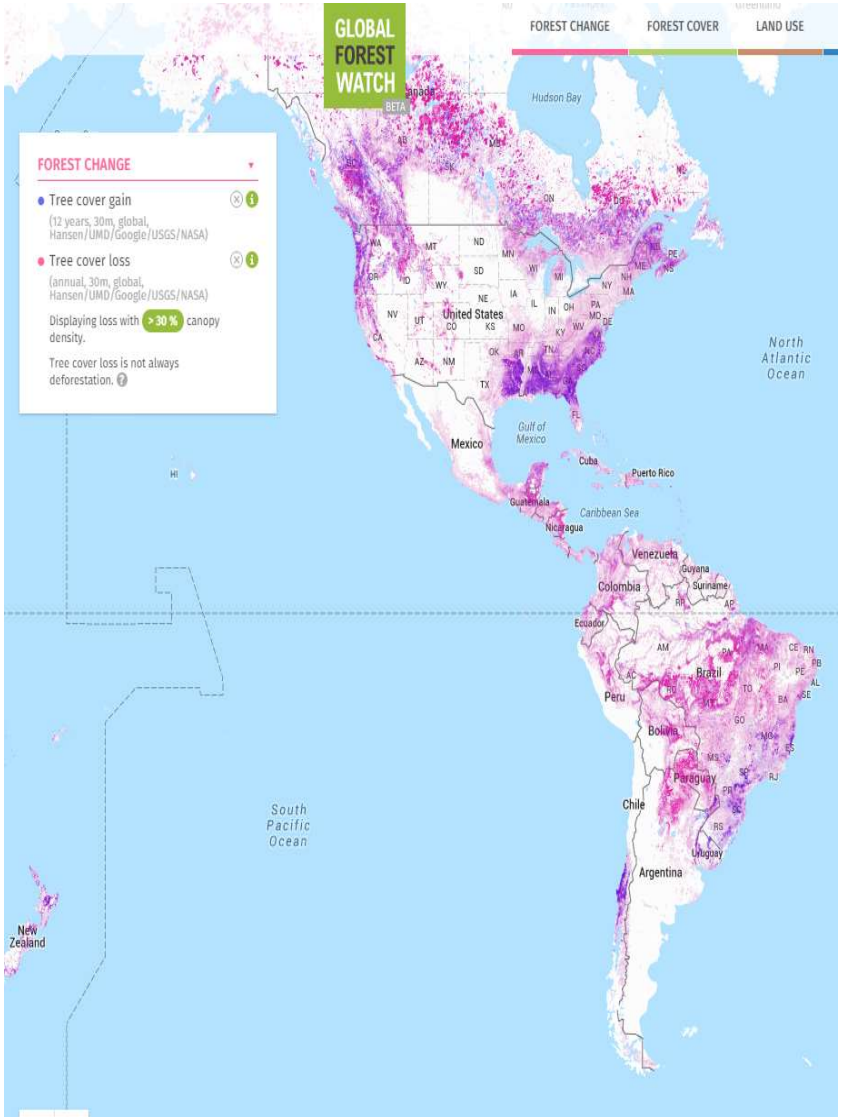
- Korean Meteorological Administration increases the agency's data storage capacity by nearly 1000% to 9.3 Petabytes
- IBM Deep Thunder Project:
  - Three-dimensional models of a space
    + prediction algorithms
    + computational power
  - Run calculations that produce very precise weather forecasts for a particular locale



- Examples:
  - Predicted snowfall in northeastern US in Feb 2014
  - Prediction of rains in Rio de Janeiro, up to 40 hours before an event with 90% certainty

Global Forest Watch

- Goal: To be able to monitor, in almost real time, deforestation-reforestation at a global level
- Uses more than 700,000 high-resolution satellite images from NASA
  - More government data, non-profit organizations, research institutions, etc.

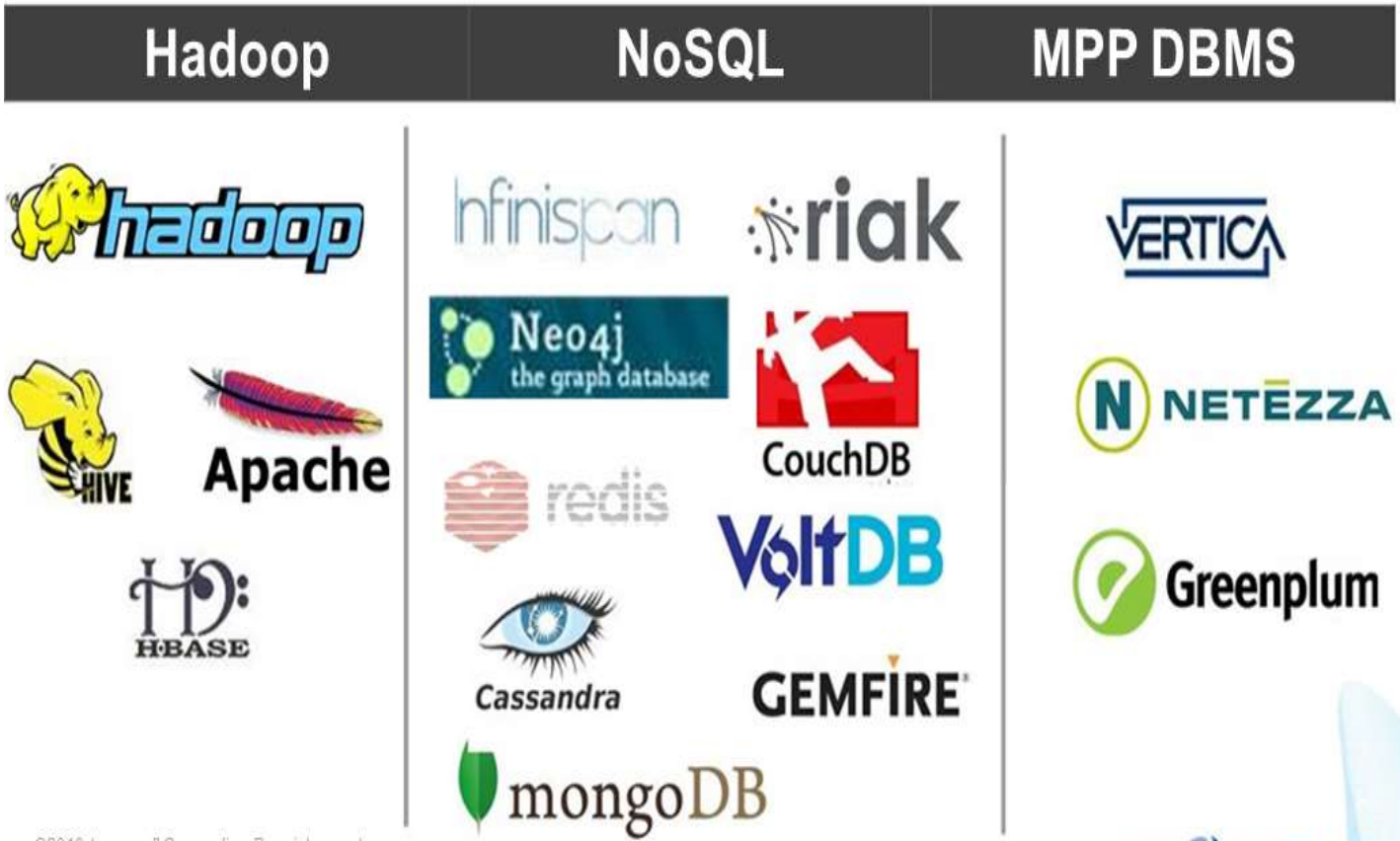


How Big Data is Going to Help Feed Nine Billion People by 2050

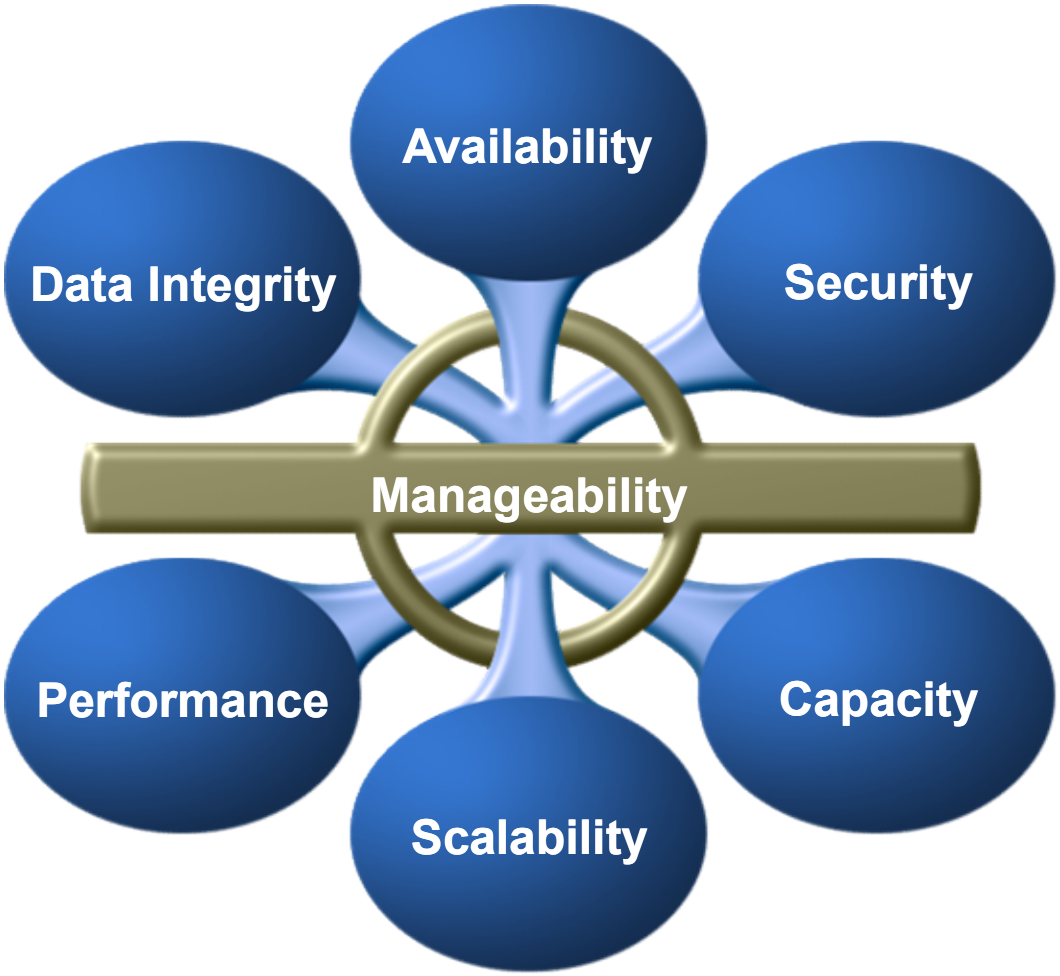
USM
We lead

- Goal: Precision Agriculture
  - Measurement and collection of data related to the production and yield of a crop
    - Using satellites, GPS, sensors on farms and equipment, etc.
  - Idea is to be able to respond and act based on the data
- Studies have shown increases in yield, cost reduction (seeds, fertilizers, chemicals, water)

Big Data - Technologies



Key Requirements for Data Center Elements



Constraints to Meeting the Requirements

Constraints include:

- Cost
- Physical environment
- Maintenance and support
- Compliance – regulatory and legal
- Hardware and software infrastructure
- Interoperability and compatibility

Management Activities

Managing a modern, complex storage environment involves many tasks. Some of the key management activities include:

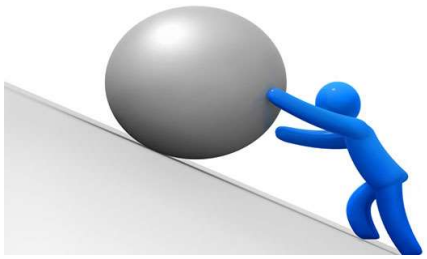
- Provisioning / Capacity / Resource Planning - resource allocation, proactive planning and provisioning for anticipated increases in capacity have to be performed
- Monitoring - array performance, data security and availability have to be continually monitored
- Reporting - periodic reporting on performance, capacity utilization, internal chargeback for cost recovery should be performed

Challenges in Managing Information

- **Exploding digital universe**
  - Creating copies of data to ensure high availability and repurposing has contributed to the **multifold** increase of information growth.

- **Increasing dependency on information**
  - The **strategic use of information** plays an important role in determining the success of a business and provides competitive advantages in the marketplace.

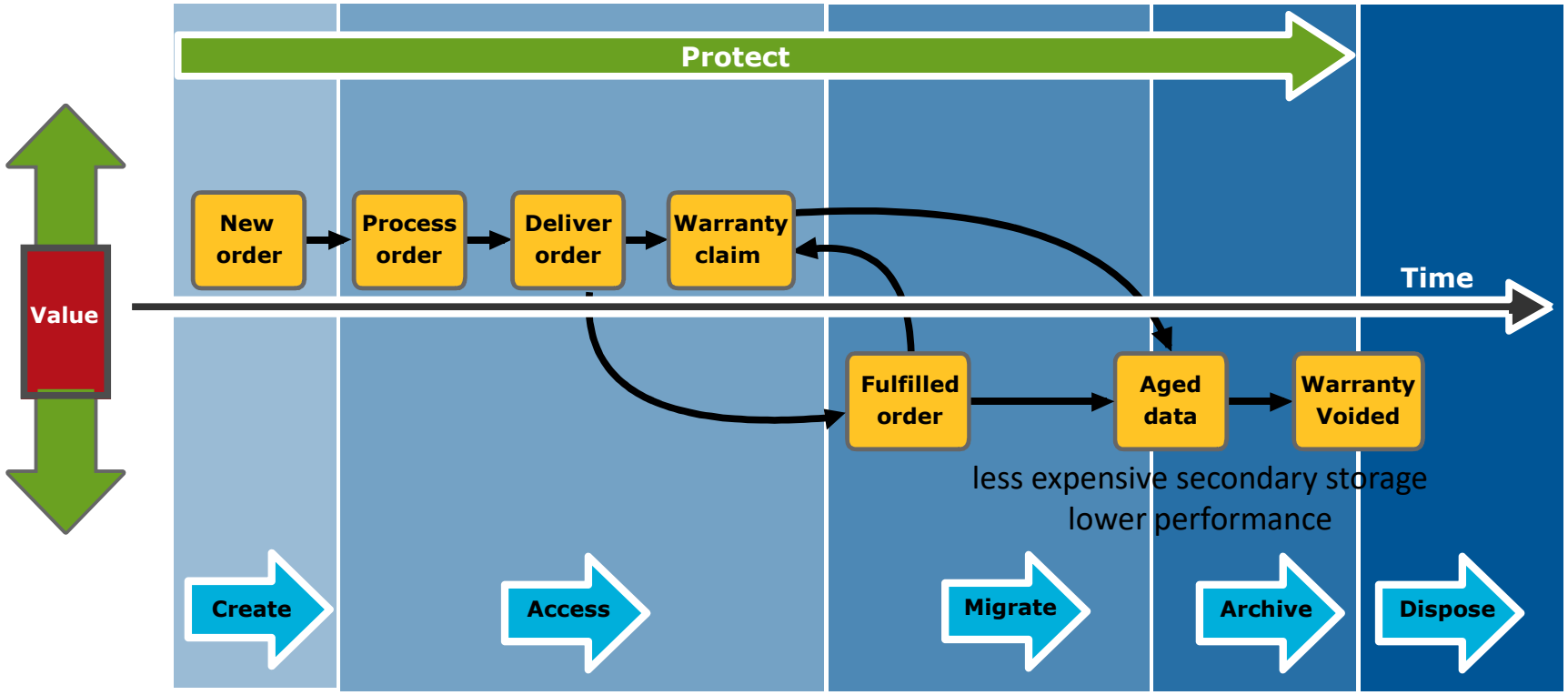
- **Changing value of information**
  - Information that is valuable today may become less important tomorrow.



Information Lifecycle Management (ILM)

We lead

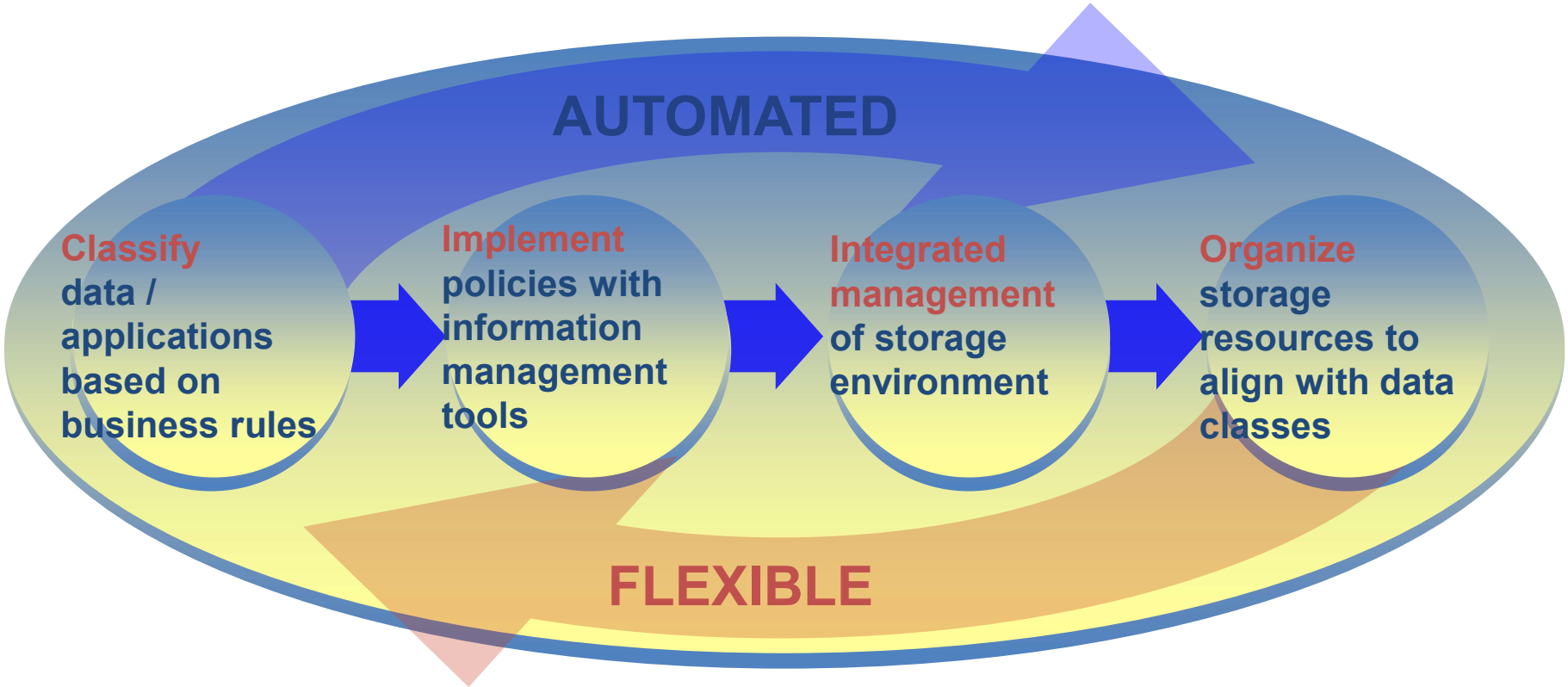
Example: Sales order application



ILM - A proactive strategy that enables an IT organization to effectively manage the data throughout its lifecycle

Information Lifecycle Management Process

Policy-based Alignment of Storage Infrastructure with Data Value



Benefits of Implementing ILM

- Optimized utilization
  - Tiered storage platforms
- Simplified management
  - By integrating process steps and interfaces with individual tools and by increasing automation
- Simplified backup and recovery
  - A wider range of options to balance the need for business continuity
- Lower Total Cost of Ownership
  - By aligning the infrastructure and management costs with information value

Introduction to Python

- Easy to learn, powerful programming language
- Great interactive environment
- Version – 2.5.x, 2.6.x, 3.x (Latest 3.10.0)

https://www.python.org/downloads/

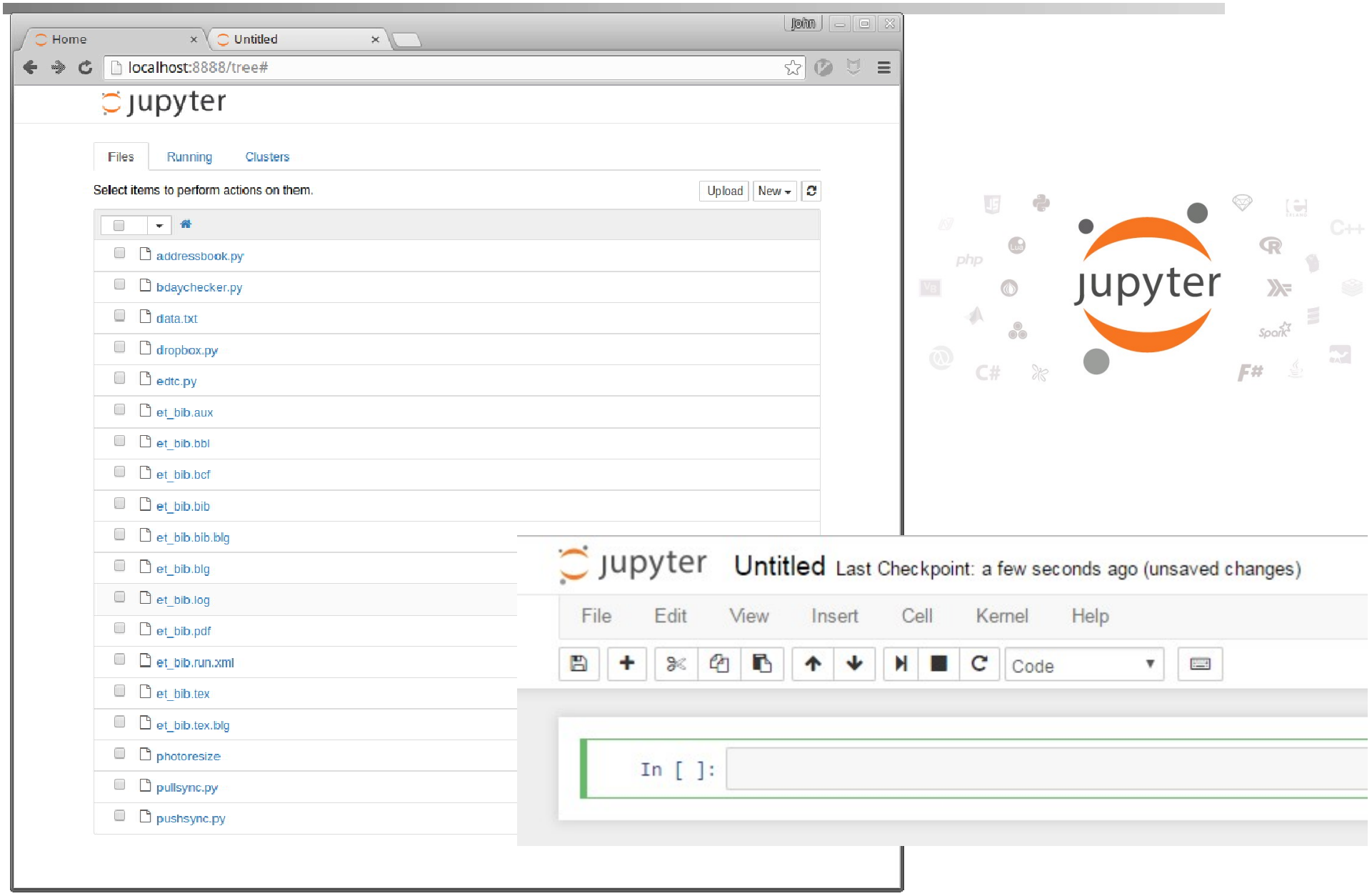
Introduction to Jupyter Notebook

- Way to combine text (& math) and code (that can be run) in one document that is rendered in a browser.

- Notebook is stored as text file in JSON format.

- Jupyter can run over 40 different languages, originally conceived for Julia, Python and R.

Introduction to Jupyter Notebook



Introduction to Anaconda

- Anaconda is a Python distribution that is particularly popular for data analysis and scientific computing:
  - Open source project developed by Continuum Analytics, Inc.
  - Available for Windows, Mac OS X and Linux
  - Includes many popular packages: NumPy, SciPy, Matplotlib, Pandas
  - Includes Spyder, a Python development environment
  - Includes conda, a platform-independent package manager

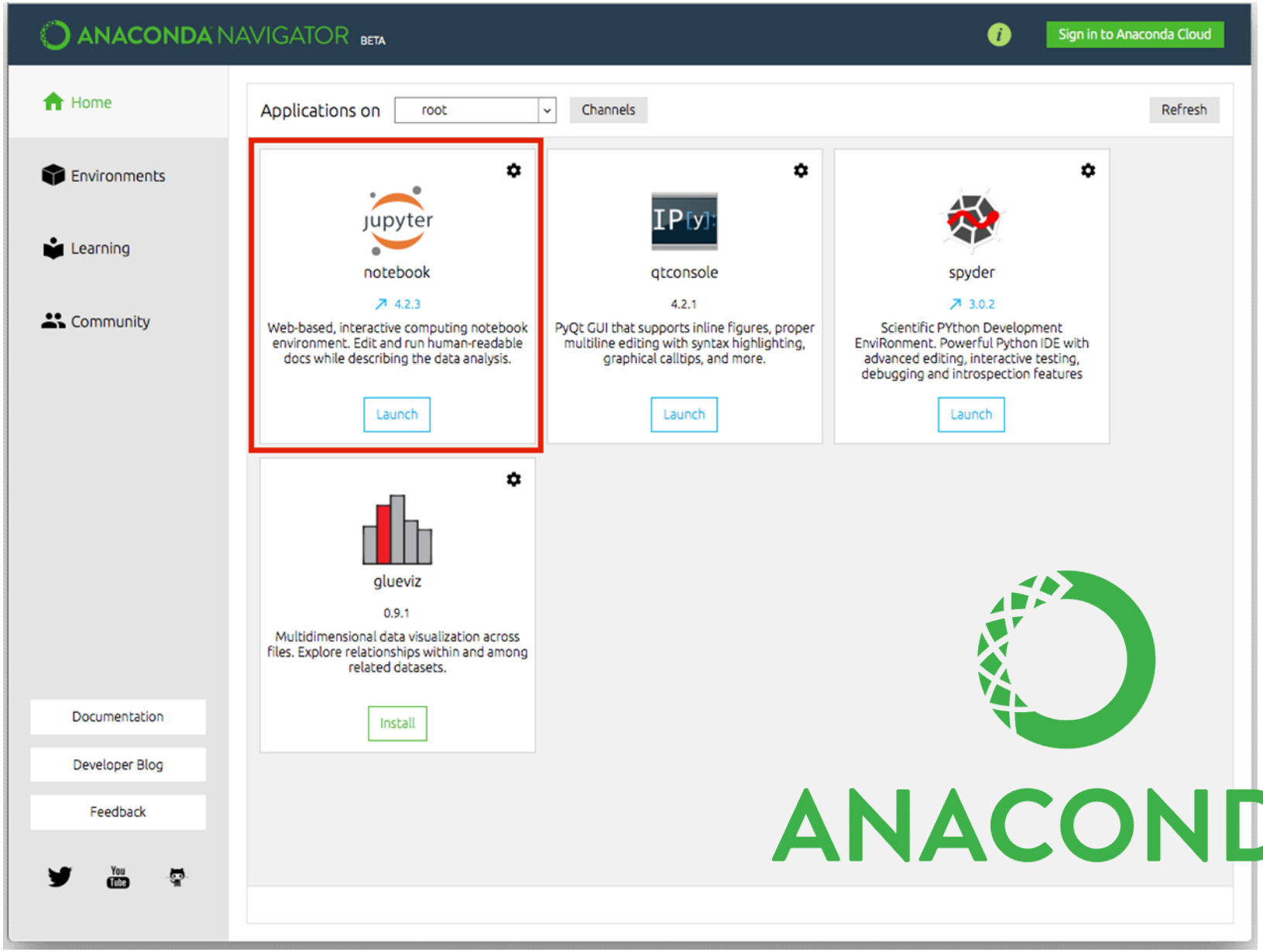
- Simplifies installation of Python packages

- Platform-independent package manager

- Provides “virtual environment” capabilities

- Many channels exist that support additional packages

Introduction to Anaconda



Tools / Language

Python / R
- Scalable
- Ease Integration
- Allow customization

https://www.python.org/downloads/
https://www.anaconda.com/distribution/
https://www.rstudio.com/products/rstudio/download/

https://softfamous.com/weka/

Weka / Rapid Miner
- Minimum Coding
- User Friendly

https://docs.rapidminer.com/latest/studio/installation/

Google Colab
- Free Jupyter notebook environment
- Supports Python 2.7 and 3.6
- Equipped with many ML & DL libraries
- Free-of-charge 12GB-RAM GPU
- 13GB RAM

https://colab.research.google.com/notebooks/welcome.ipynb

Intel AI DevCloud
- Supports Python
- Equipped with many ML & DL libraries
- 96GB DDR4 RAM

https://software.intel.com/en-us/devcloud

Download & Install

Anaconda Navigator

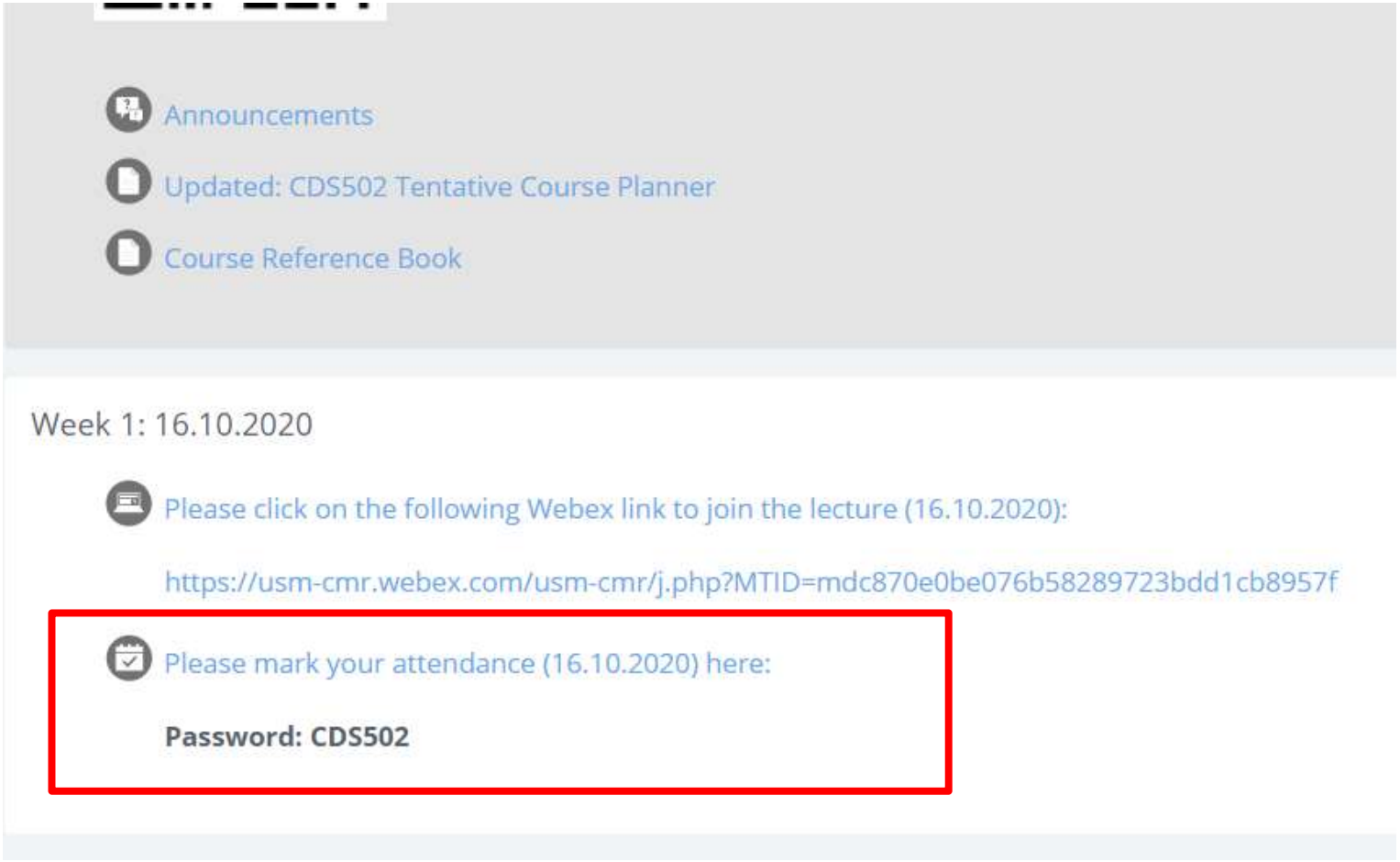
Installation Steps:

These setup steps might take up to 1 hour to complete, hence it is highly recommended to have these software & libraries installed prior to the lab sessions.

1) Download and install Anaconda at:
https://www.anaconda.com/distribution/#download-section
Alternative: You can also access this link for Anaconda installer: **https://bit.ly/2UW1VKh**

2) Ensure that your Anaconda is able to start up upon installation completion. Go to Windows, Search for "Anaconda Navigator", run it. It will take 1 – 2 minutes to boot up. If the Anaconda window is shown, then you are good to go.

Mark your Attendance



https://quizizz.com/join

Thank You

Prepared & Presented by:
Ts. Dr. Chew XinYing
School of Computer Sciences