SWS3004: Cloud Computing with Big Data

School of Computing Summer Workshop: 12 July to 1 August 2018

Teo Yong Meng
Department of Computer Science
National University of Singapore
email: teoym@comp.nus.edu.sg
url: www.comp.nus.edu.sg/~teoym

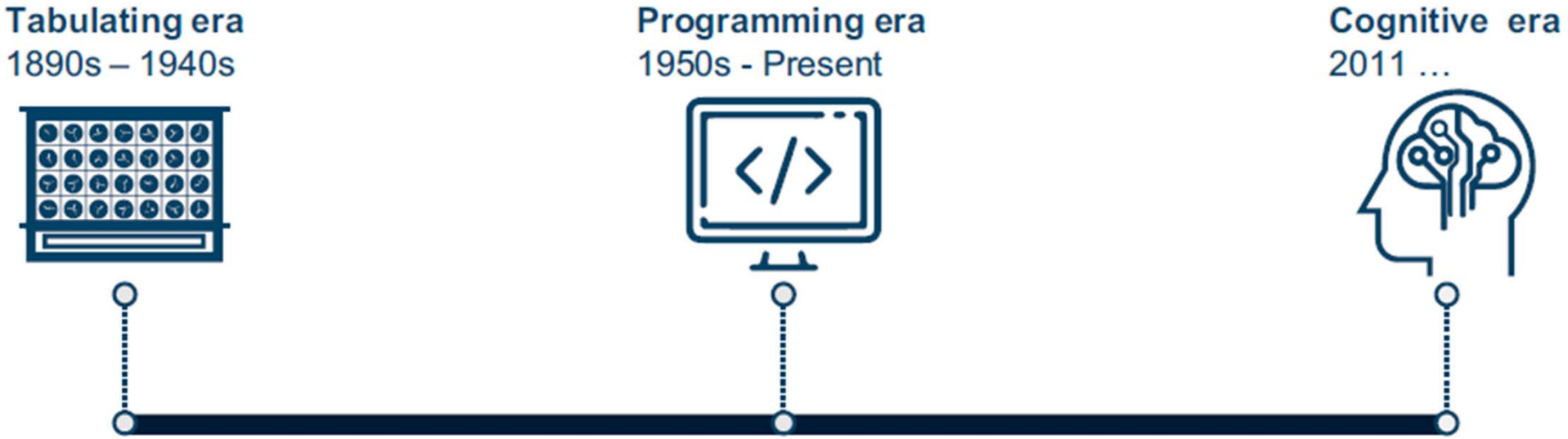
course page: www.comp.nus.edu.sg/~teoym/sws3004-18

userid: **sws3004** password: **2018sws**

L01: Introduction

1. Three Eras of Computing
2. Cloud Computing
3. Cloud-enabled Data Analytics

1. Three Eras of Computing



Three Eras of Computing

1. Tabulating (1890s – 1940s)
   - Mechanical tabulators to count and summarize information
   - limited to a single task

2. Programming (1950s – present)
   - programmable systems to perform different and multiple tasks

3. Cognitive (2011 – future)
   - Extend the boundaries of human cognition
   - learn and adapt over time

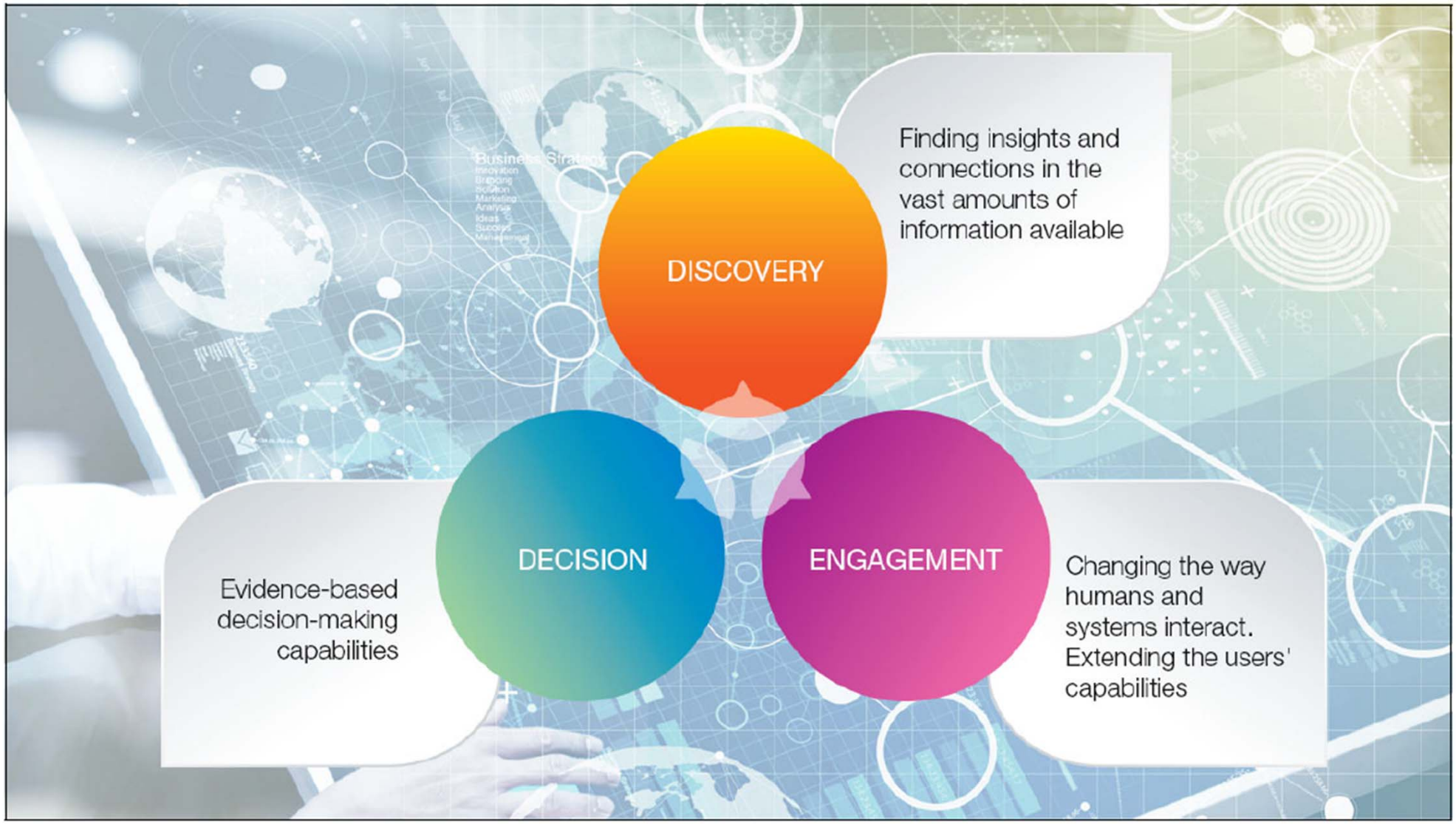
Building cognitive applications with IBM Watson Services: Volume 1, IBM Redbooks, June 2017.

Cognitive Computing

- cognition – act of thinking
- cognitive computing – “systems that learn at scale, reason with purpose and interact with human naturally”

  J.E. Kelly, Computing, cognition and the future of knowing: how humans and machines are forging a new age of understanding, 2015.
- explosion of (unstructured) data leads to the development of cognitive systems
- humans and cognitive systems are complementary
- broad capability of cognitive systems: discovery, engagement and decision

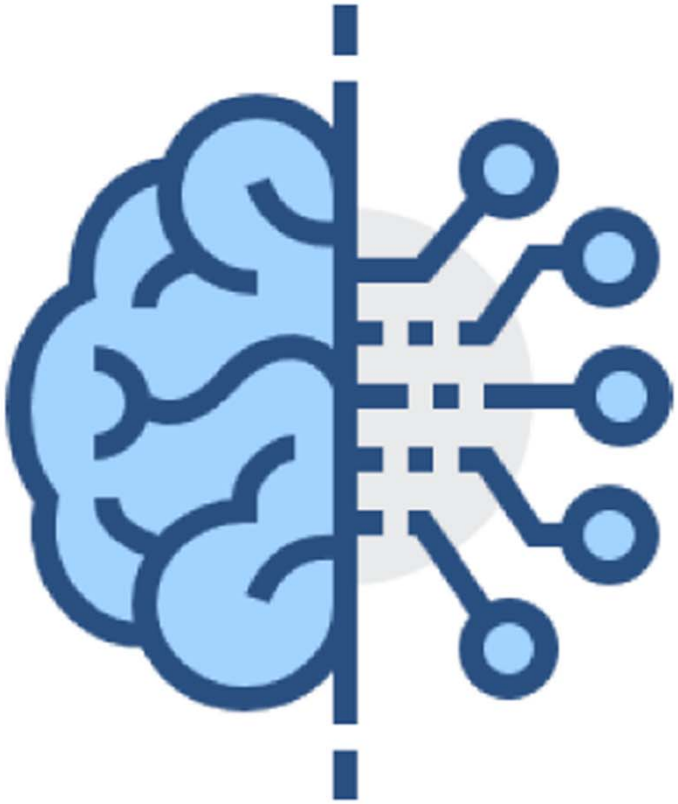
Broad Capability Areas of Cognitive Systems



Complementary Human and Cognitive Systems

Humans Excel at:

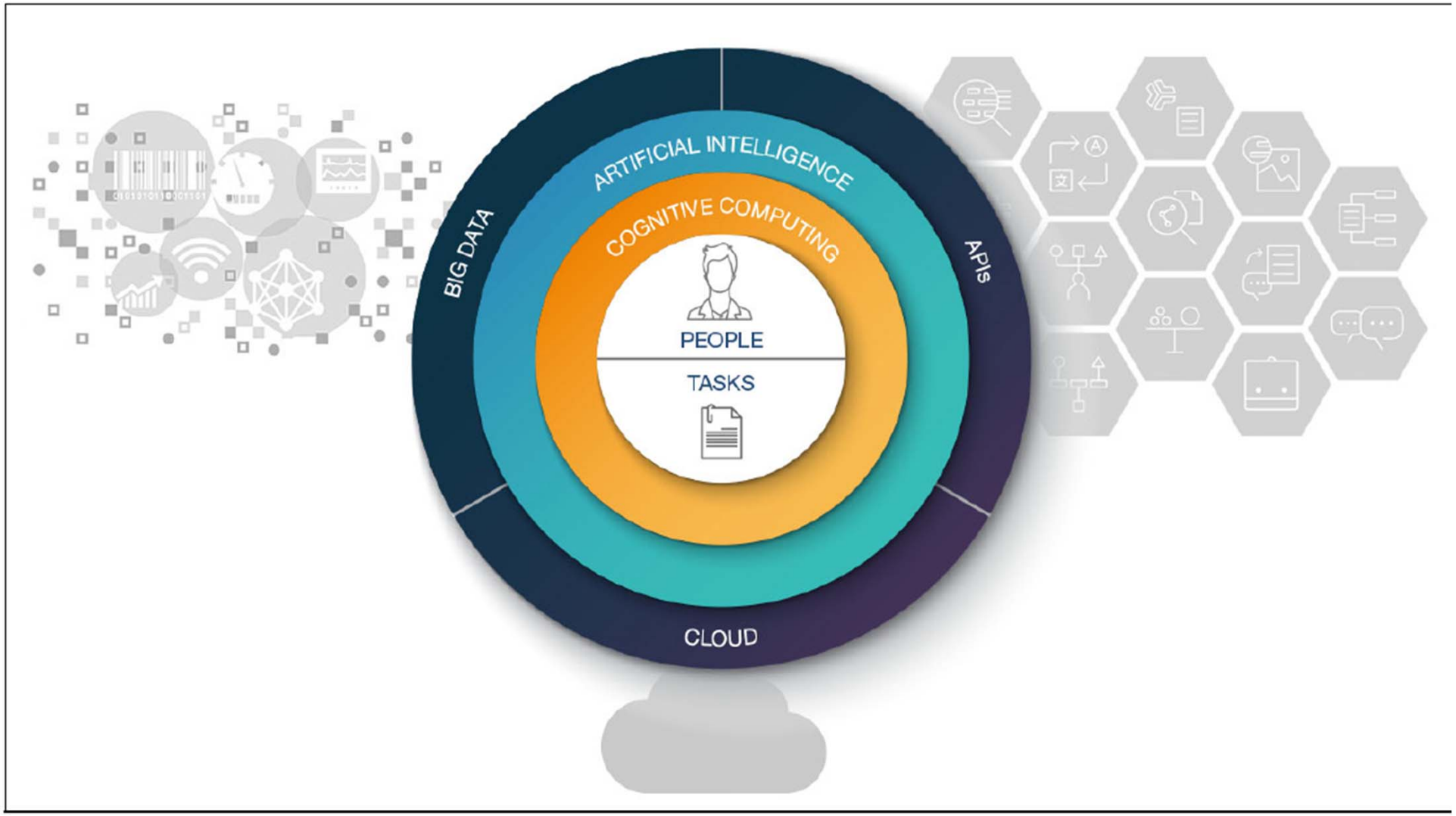
Common sense
Morals
Imagination
Compassion
Abstraction
Dilemmas
Dreaming
Generalization



Cognitive Systems Excel at:

Locating Knowledge
Pattern Identification
Natural Language
Machine Learning
Eliminate Bias
Endless Capacity

Cognitive Computing Model



Cloud Computing

- Delivers cognitive services

- Examples of cloud platforms:
  - IBM Watson cognitive services over Bluemix
  - Amazon AI services over AWS
  - Microsoft AI tools over MS Azure
  - Google AI services on Google Cloud

2. CLOUD COMPUTING

- What and Why
- Cost Model
- History
- Key Business Drivers

- Basic Concepts and Terminology
- Goals and Benefits
- Technical and Non-Technical Challenges
- Summary

Key Terms

1. Elastic resource
2. Availability
3. Capacity planning (resource provisioning)
4. Scaling (horizontal, vertical)
5. Cloud-based IT resources
6. Cloud service
7. Trust boundary

Definitions

- Gartner Report
  ".. a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a **service** to external customer using Internet technologies."

- Forrester Research
  ".. a standardized IT capability (services, software, or infrastructure) delivered via Internet technologies in a pay-per-use, self-service way."

- NIST 2011
  "Cloud computing is a model for enabling *ubiquitous*, *convenient*, ***on-demand*** *network access* to a shared pool of **configurable computing resources** (e.g., *networks, servers, storage, applications and services*) that can be **rapidly provisioned** and released **with minimal management effort or service provider interaction**."

elastic resource + on-demand service

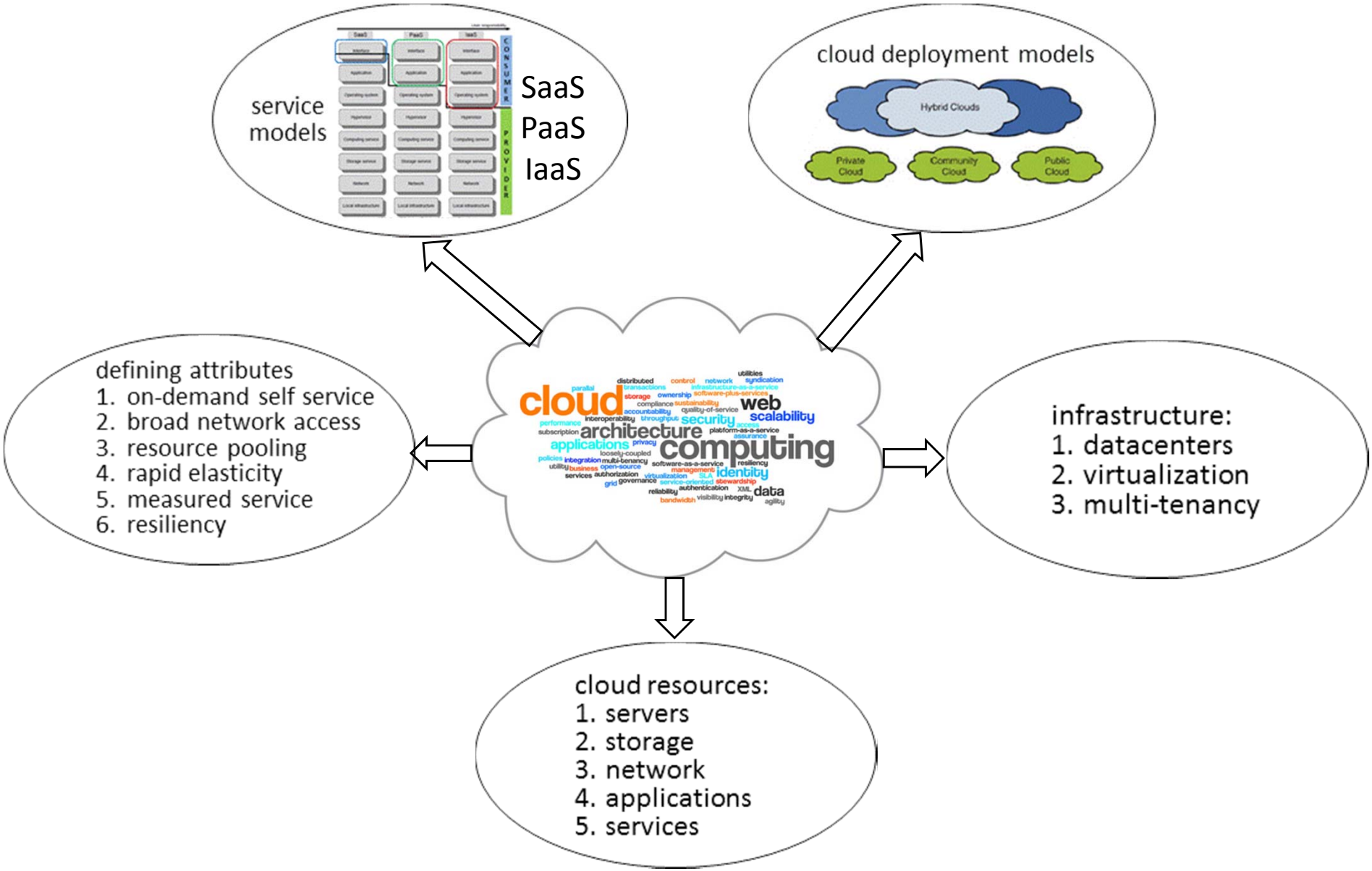
Why Cloud Computing?

- **resource demand**
  - elastic user demand: peak/trough, business cycles
  - mismatch between elastic resource demand and fixed (in-house) resource capacity
  - loss in business opportunity and higher business cost

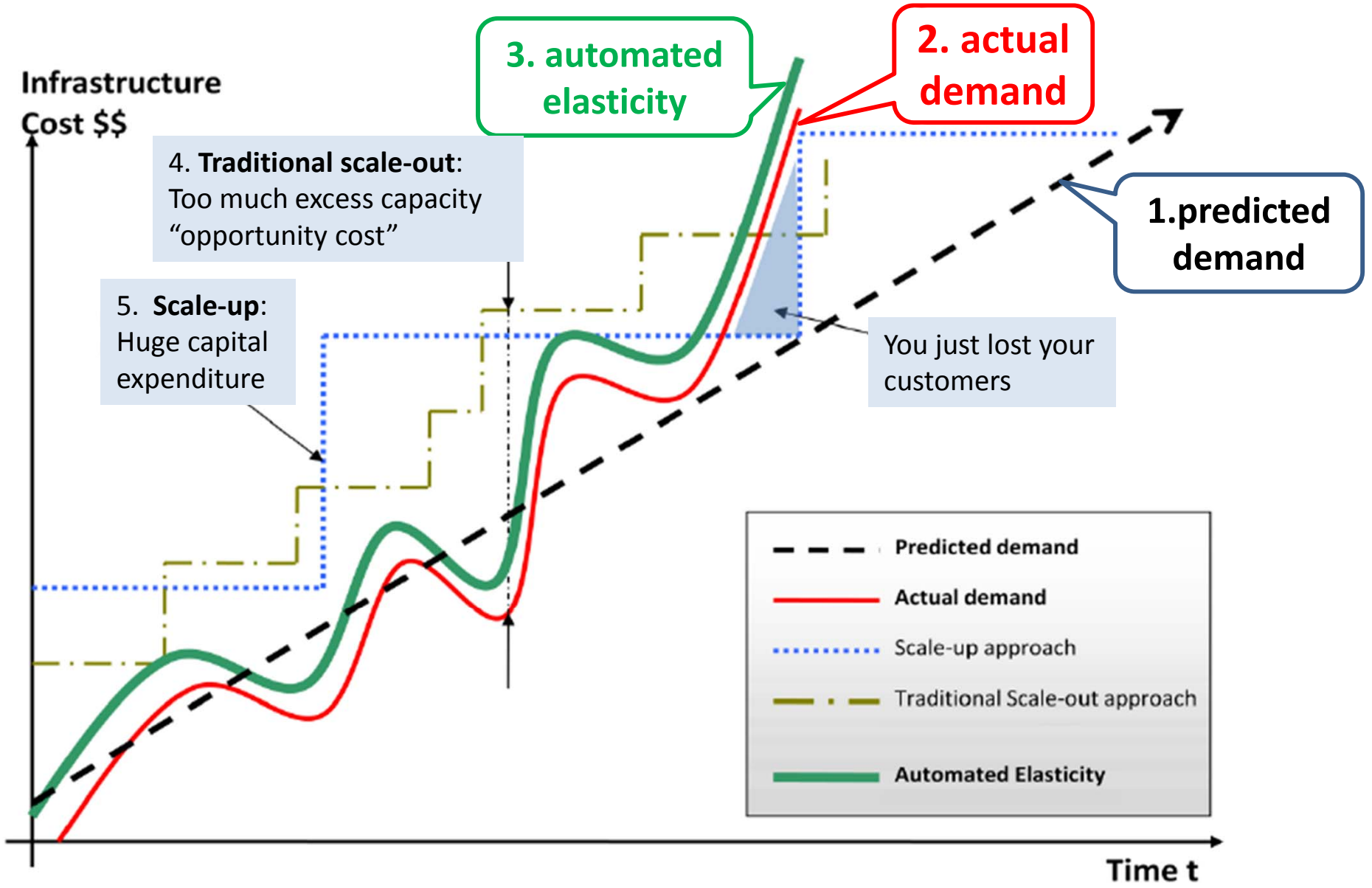
- **computing resource**
  - commoditized and shared -> economy of scale
  - Internet (latency and bandwidth)

- cloud improves the match between elastic resource demand and elastic computing resource -> reduces business cost

What is Cloud Computing?

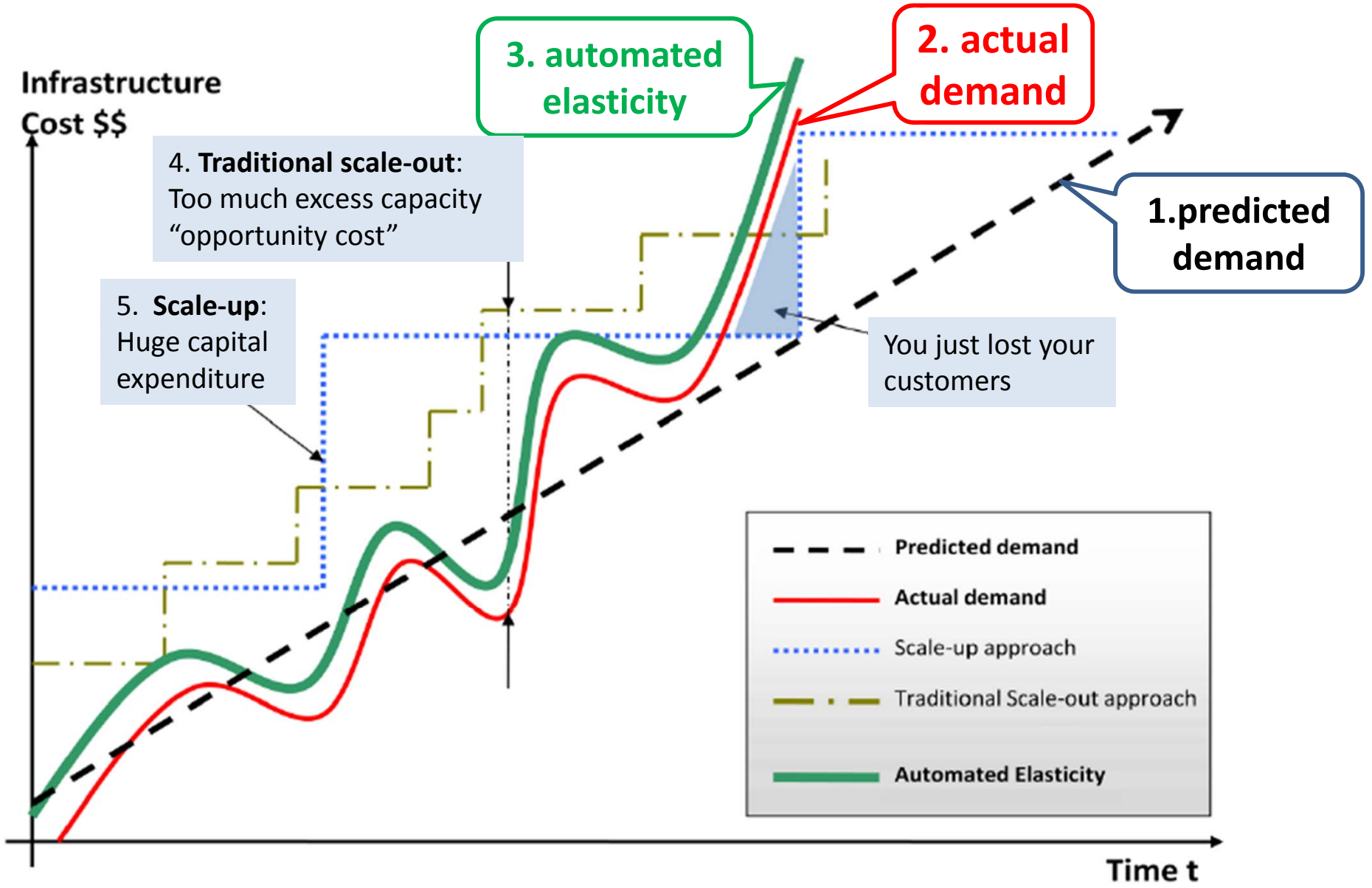


Why Cloud Computing?



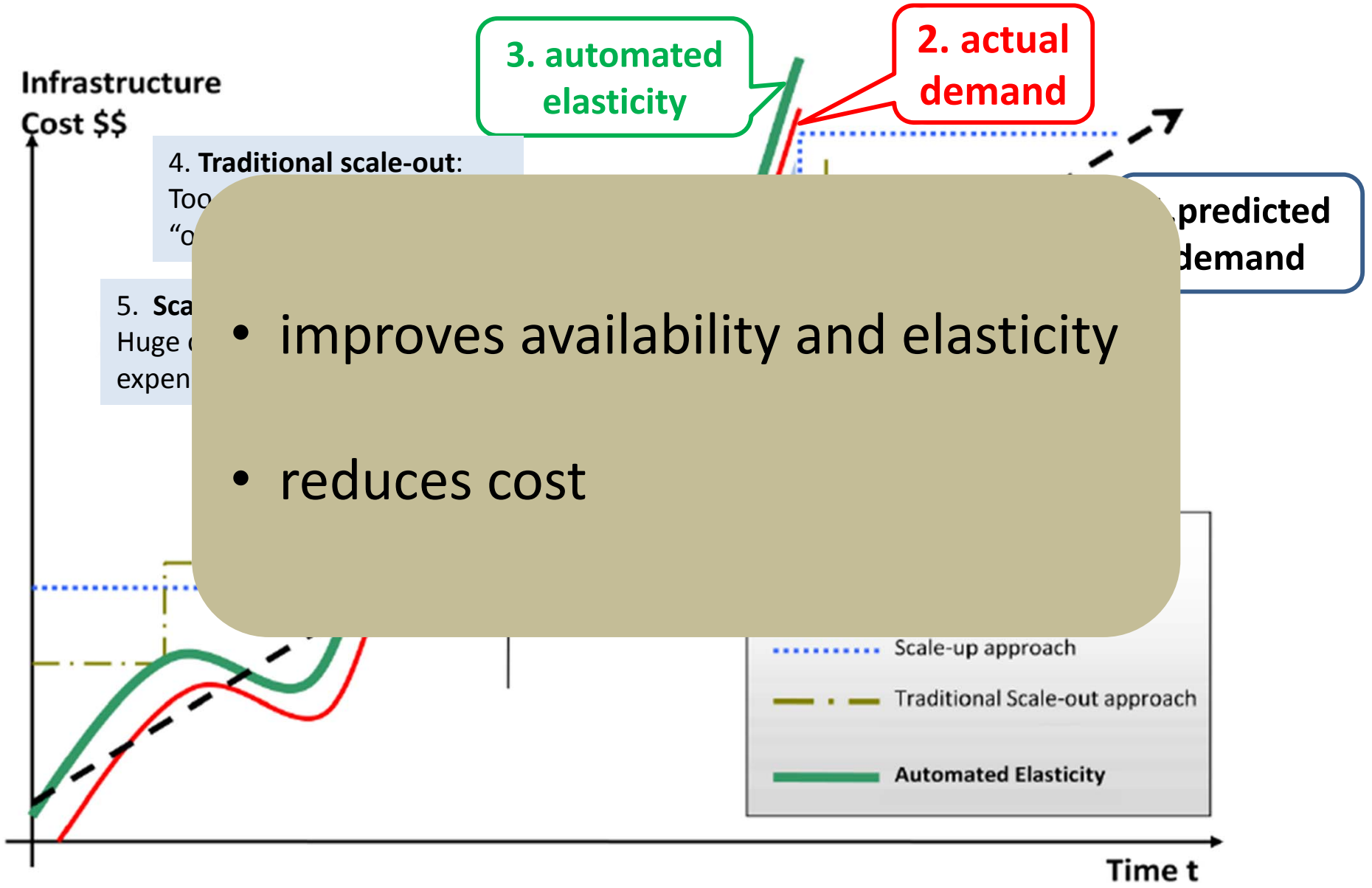
adapted from http://aws.amazon.com/economics

Why Cloud Computing?



adapted from http://aws.amazon.com/economics

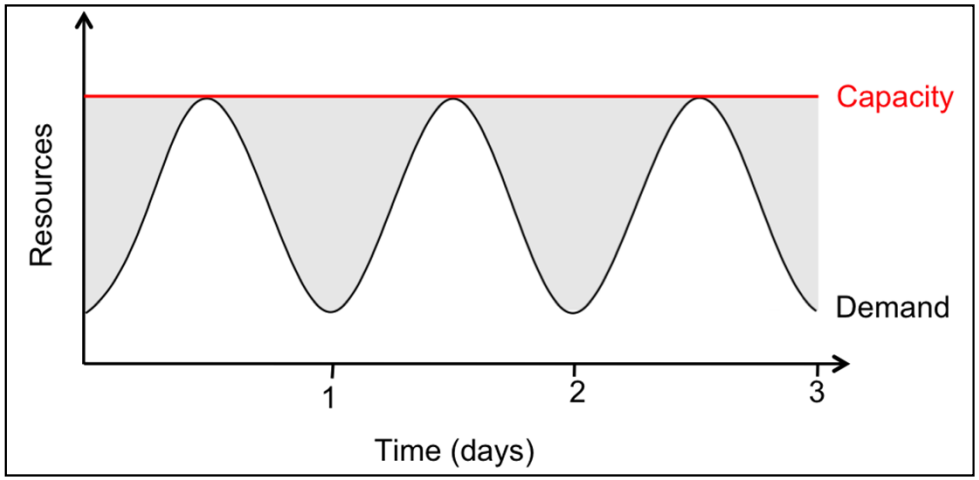
Why Cloud Computing?



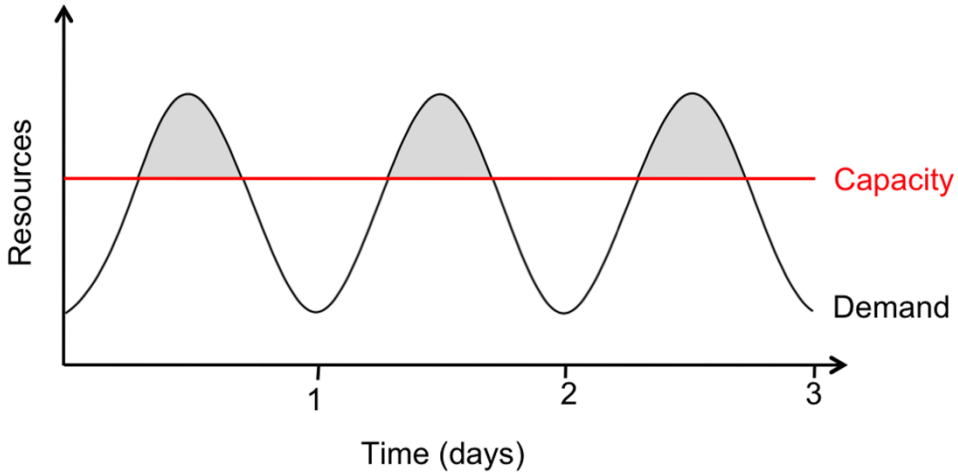
- improves availability and elasticity
- reduces cost

adapted from http://aws.amazon.com/economics

Resource Provisioning



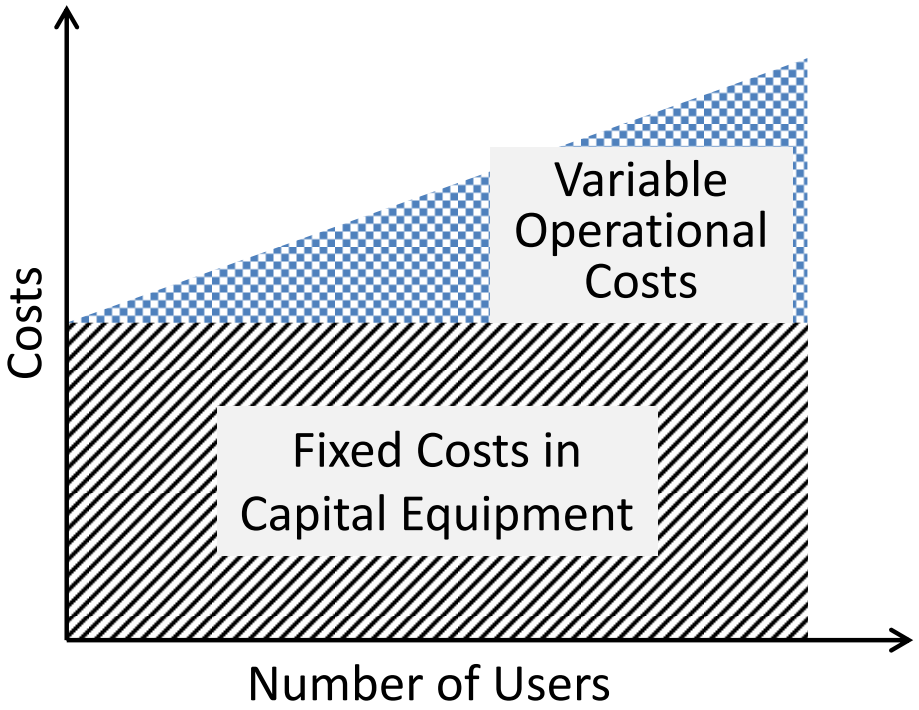
a. Provisioning for Peak Load
Even if peak load can be correctly anticipated, resources are wasted (shaded area) during non-peak times



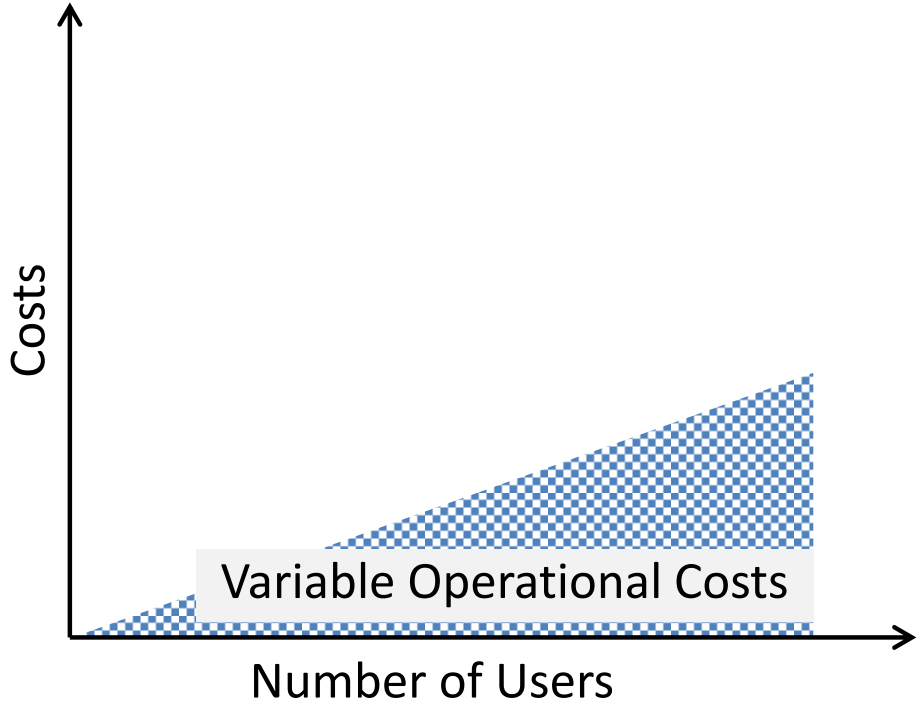
b. Under-provisioning
loss of potential revenue from users not served (shaded area)

Reference: Above the Clouds: A Berkeley View of Cloud Computing, 2009.

Cost Model



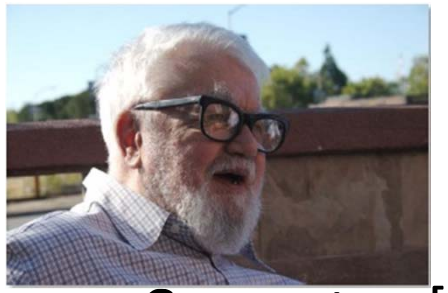
a. **Traditional IT**



b. **Cloud Computing**

HISTORY

- **1961** "computation may someday be organized as a public utility" – John McCarthy (1927-2011)



- **1996** – "cloud computing" was coined at Compaq Computer [MIT Technology Review]
- **1999 Salesforce.com** – pioneered the concept of delivering enterprise applications via a simple website
- **Jul 2002 Amazon Web Services (AWS)** – provided a suite of cloud-based services including computation, storage and even human intelligence through the Amazon Mechanical Turk
- **Mar 2006 Amazon Simple Storage Service (S3)** – "pay-per-use" storage

HISTORY

- **Aug 2006 Amazon Elastic Compute Cloud (EC2)** – a commercial web service (IaaS) that allows small companies and individuals to rent computers to run their own computer applications

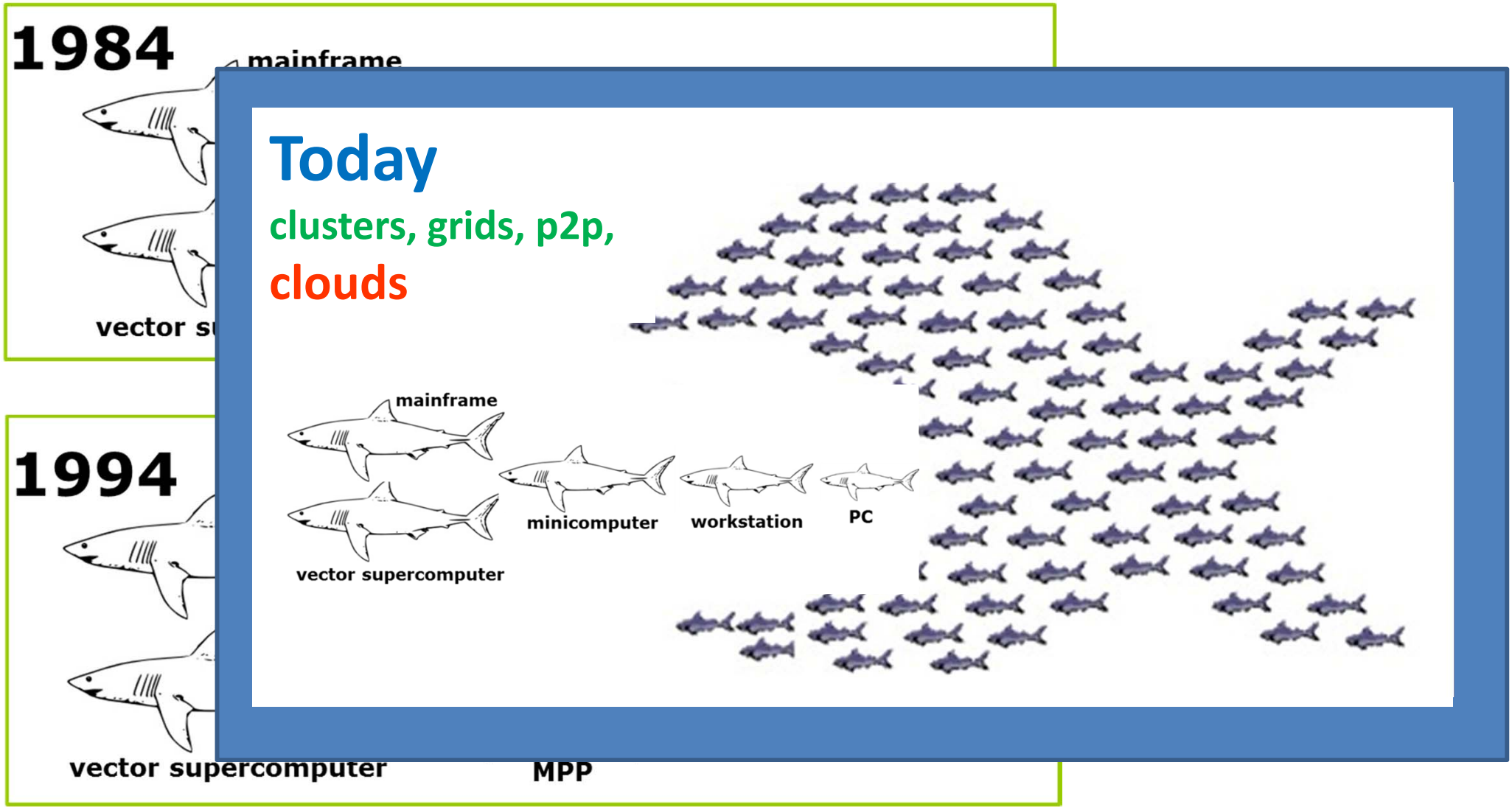
- **Apr 2008 Google App Engine** - Google's PaaS: Bigtable and GFS for storage, MapReduce, …

- **Nov 2009 Microsoft Windows Azure** - an operating environment "designed to manage extremely large pools of computational resources"; customers run Windows-based applications over the Internet using Microsoft's data centers, with Azure organizes resources and handles spikes in demand

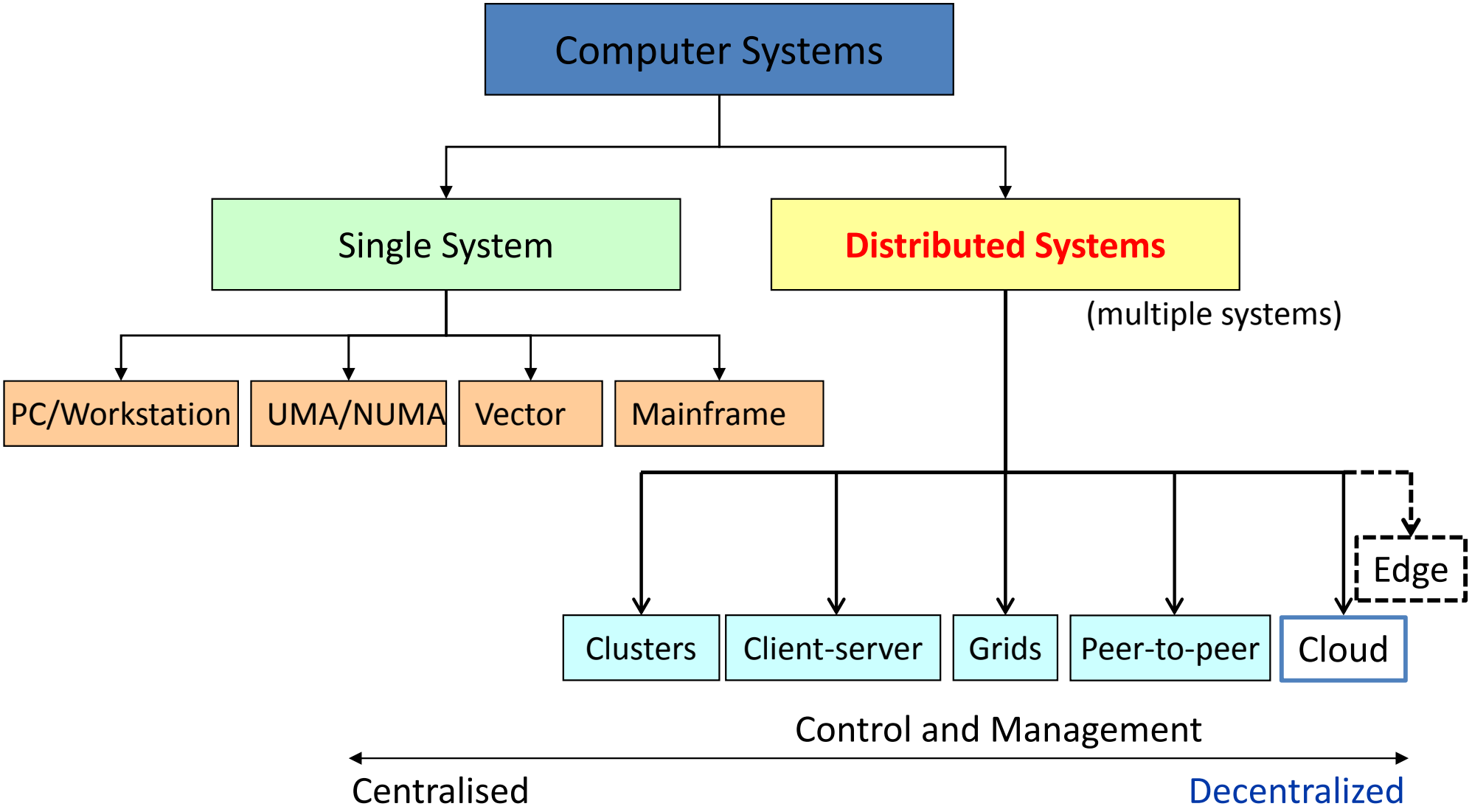
- **Apr 2011** - IBM SmartCloud, IBM Bluemix Beta (2014?)

Computer Systems Landscape

(Computer Food Chain)



Types of Computer Systems



KEY BUSINESS DRIVERS

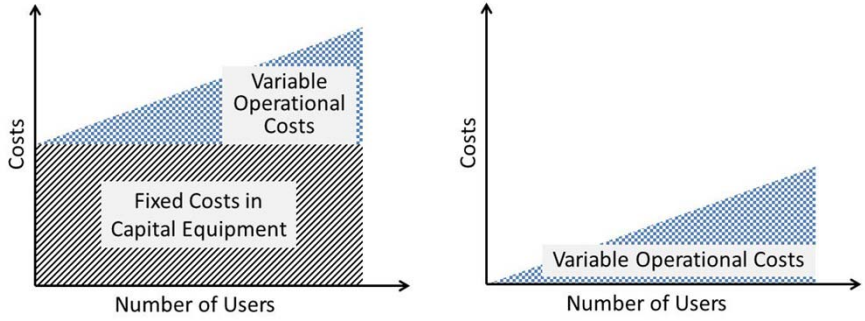
- Capacity planning
- Cost reduction
- Organization agility

Capacity Planning

- Process of determining and fulfilling future demands of an organization's IT resources, products and services
- Challenges: usage load fluctuations, peak usage, cost of provisioning, …
- Strategies:
  - *lead strategy*: add capacity in anticipation of demand
  - *lag strategy*: add capacity when resources reaches its full capacity
  - *match strategy*: add capacity in small increments as demand increases
- Cloud: on-demand self-service, rapid elasticity, measured service

Cost Reduction

- Difficult to directly align IT costs and business performance
- IT department is a **cost center** in an enterprise
- Costs: up-front investment (ownership) costs (capital expenditure) + operational costs
- Operational costs: technical staff (manpower), utility bills (power and cooling), security and access control to protect infrastructures, administration/account staff (software licenses, etc.)



a. Traditional IT b. Cloud Computing

Organizational Agility

- A measure of an organization's responsiveness to change
- *Elastic IT resources* to response to business cycles beyond what was previously predicted or planned for
- Upfront investments and infrastructure ownership costs may be prohibitive

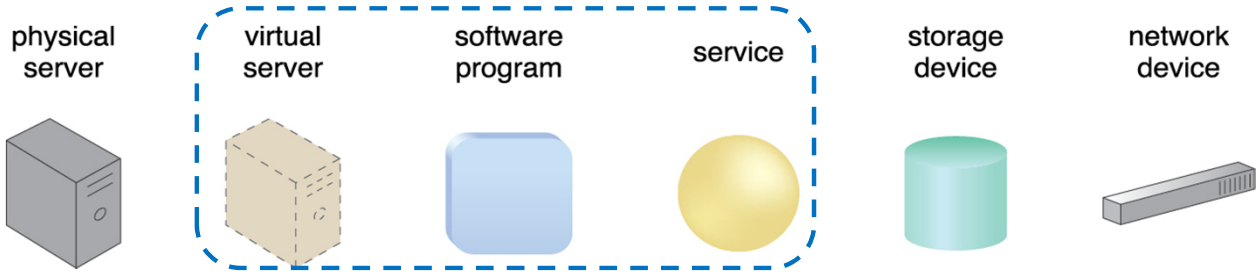
Technology Innovations

- Clustering – group of independent IT resources interconnected to work as a single system
- Grid computing – sharing of resources organized into logical pools to provide high performance computing; system is loosely coupled and distributed
- Virtualization – abstraction of IT resources to improve ease of use and facilitates the creation of virtual instances of IT resources
- Technology innovations vs enabling technologies (cover in L03) – virtualization, datacenters, multi-tenancy, …
- …

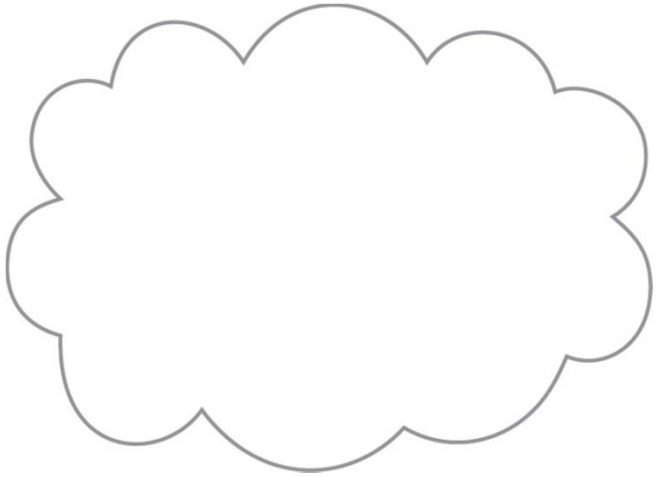
Key Terms

1. Elastic resource
2. Availability
3. Capacity planning (resource provisioning)
4. Scaling (horizontal, vertical)
5. Cloud-based IT resources
6. Cloud service
7. Trust boundary

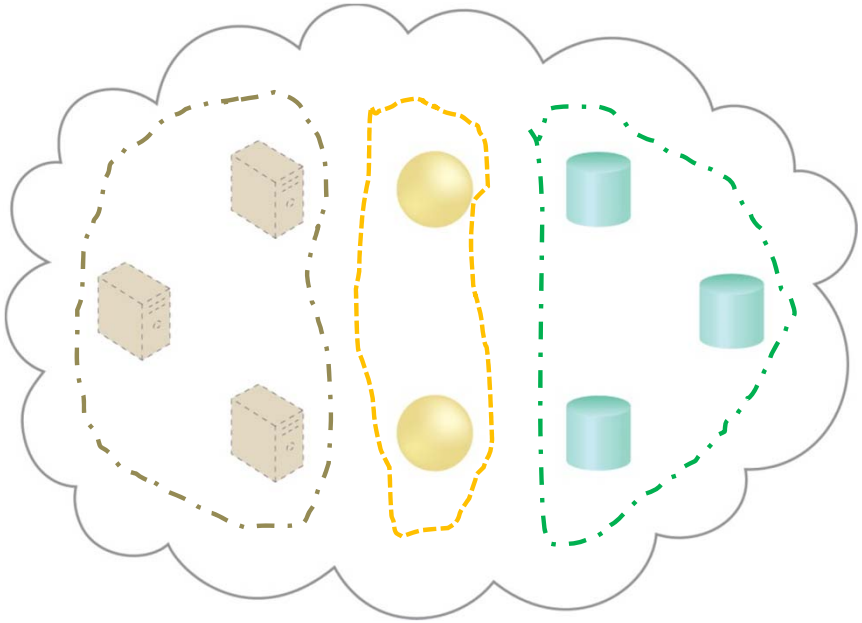
BASIC CONCEPTS AND TERMINOLOGY



Examples of common Cloud-based IT resources and their corresponding symbols



Symbol denotes the Boundary of a Cloud Environment

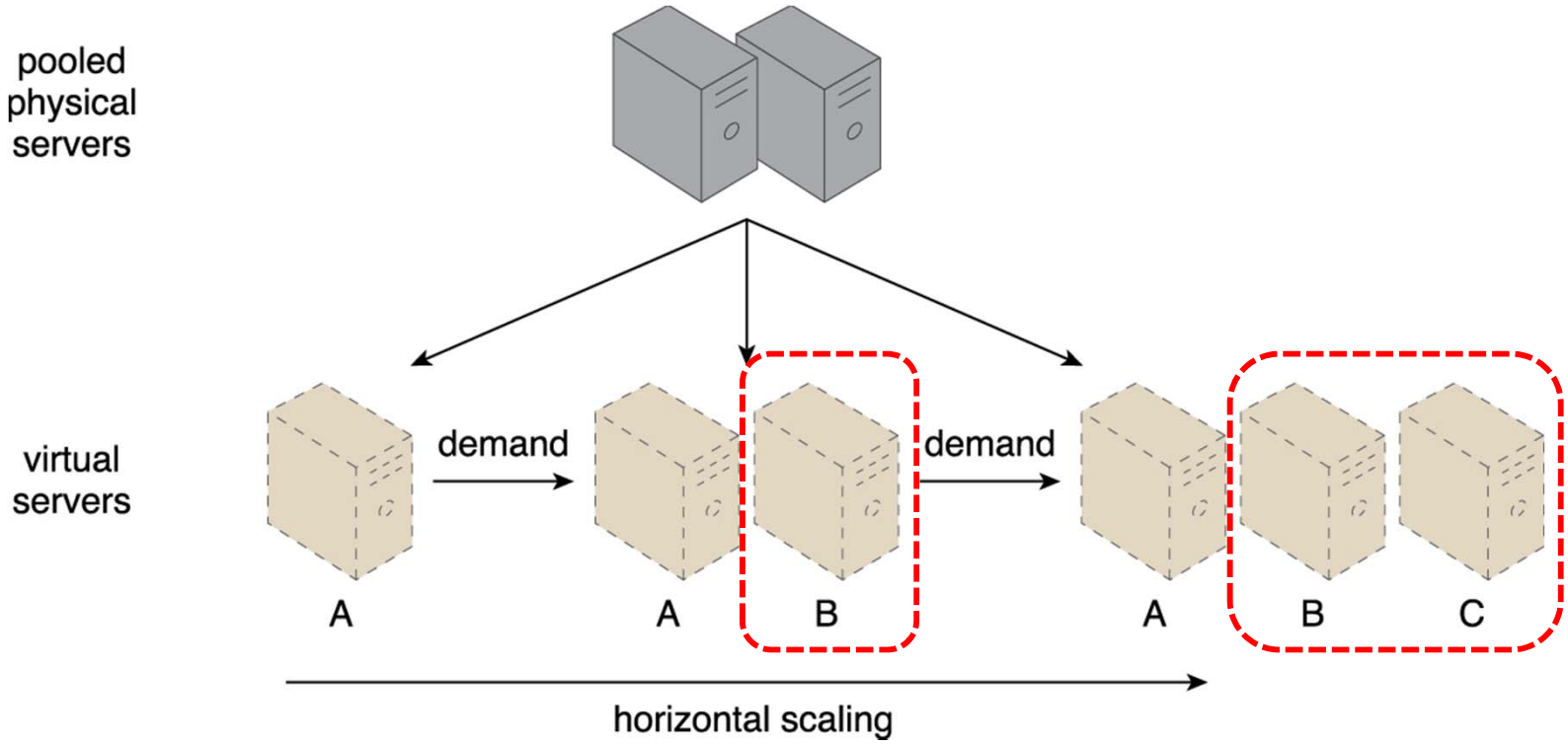


Example: A cloud hosting 8 IT resources: 3 virtual servers, 2 **cloud services**, and 3 storage devices

Scaling

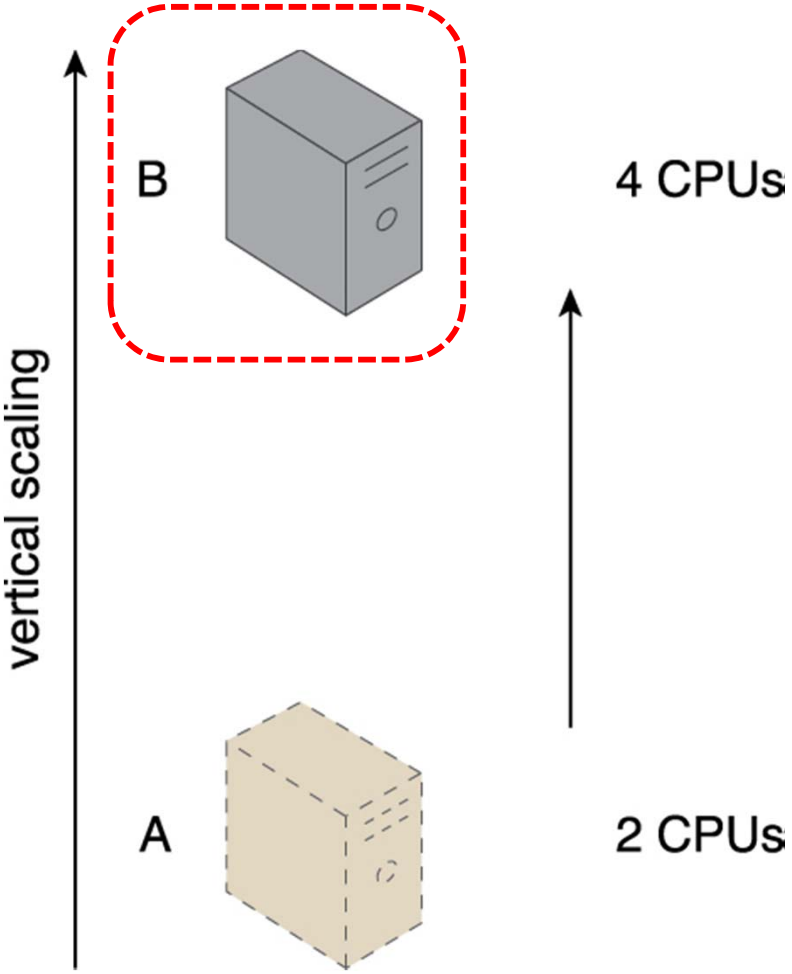
- ability of the IT resource to handle increased or decreased usage (business) demands
- **Horizontal scaling** (out and in) – same resource type, i.e., add more nodes
- **Vertical scaling** (up and down) – replace resource with higher or lower capacity or add resources to a single node

Horizontal Scaling



Example: An IT resource (Virtual Server A) is scaled out by adding more of the same IT resources (Virtual Servers B and C).

Vertical Scaling



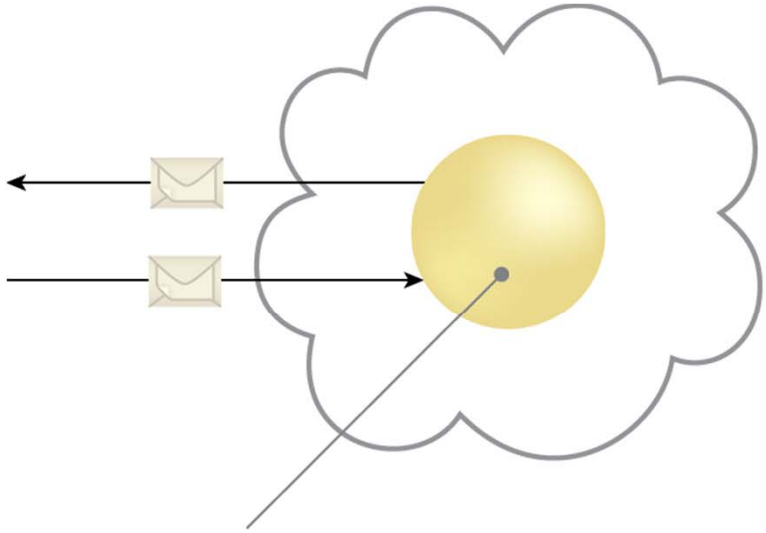
Example: An IT resource (a virtual server, A, with two CPUs) is scaled up by replacing it with a more powerful IT resource with increased capacity for data storage (a physical server, B, with 4 CPUs).

Scaling Comparison

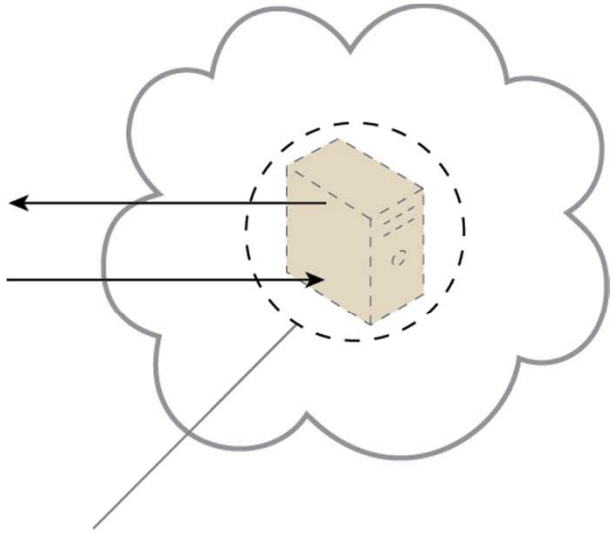
Horizontal	Vertical
Less expensive using commodity hardware	More expensive using specialized hardware
Resources instantly available	Resources normally instantly available
Resource replication and automated scaling	May need additional setup
Additional IT resources needed	No additional IT resources needed
Not limited by hardware capacity	Limited by maximum hardware capacity

Cloud Service

- Any IT resource made remotely accessible via a cloud



remotely accessed Web service acting as a cloud service



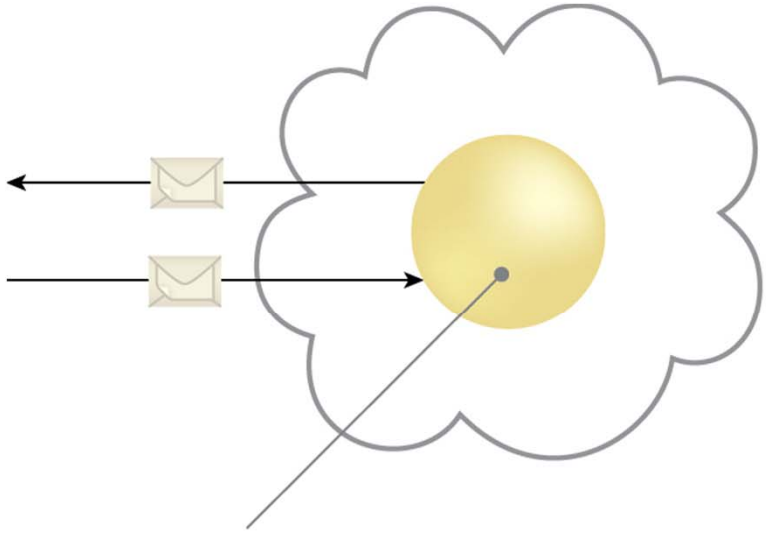
remotely accessed virtual server acting as a cloud service

A cloud service with a published technical interface is being accessed by a consumer outside of the cloud (left). A cloud service that exists as a virtual server is also being accessed from outside of the cloud's boundary (right). The cloud service on the left is likely being invoked by a consumer program that was designed to access the cloud service's published technical interface. The cloud service on the right may be accessed by a human user that has remotely logged on to the virtual server.

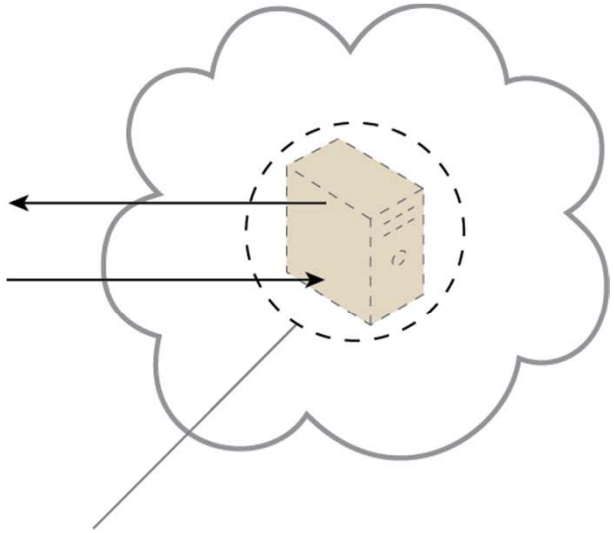
- A multitude of service usage models: Infrastructure-as-a-Service (IaaS), PaaS (Platform-as-a-Service), SaaS (Software-as-a-Service) (cover in L02)

Cloud Service

- Any IT resource made remotely accessible via a cloud



remotely accessed Web service acting as a cloud service



remotely accessed virtual server acting as a cloud service

A cloud service with a published technical interface is being accessed by a consumer outside of the cloud (left). A cloud service that exists as a virtual server is also being accessed from outside of the cloud's boundary (right). The cloud service on the left is likely being invoked by a consumer program that was designed to access the cloud service's published technical interface. The cloud service on the right may be accessed by a human user that has remotely logged on to the virtual server.

- A multitude of service usage models: Infrastructure-as-a-Service (IaaS), PaaS, SaaS (cover in L02)

Cloud Service

- Any IT resource made remotely accessible via a cloud

A cloud service with a published technical interface is being accessed by a consumer outside of the cloud (left). A cloud service that exists as a virtual server is also being accessed from outside of the cloud’s boundary (right). The cloud service on the left is likely being invoked by a consumer program that was designed to access the cloud service’s published technical interface. The cloud service on the right may be accessed by a human user that has remotely logged on to the virtual server.

- A multitude of service usage models: Infrastructure-as-a-Service (IaaS), PaaS, SaaS (cover in L02)

Cloud Service Consumer

- A runtime role assumed by a software program to access a cloud service

Examples: Depending on the nature of a given diagram, an artifact labeled as a cloud service consumer may be a software program or a hardware device (in which case it is implied that it is running a software program capable of acting as a cloud service consumer).

- Actors in cloud computing (in Lecture #02)

Goals and Benefits

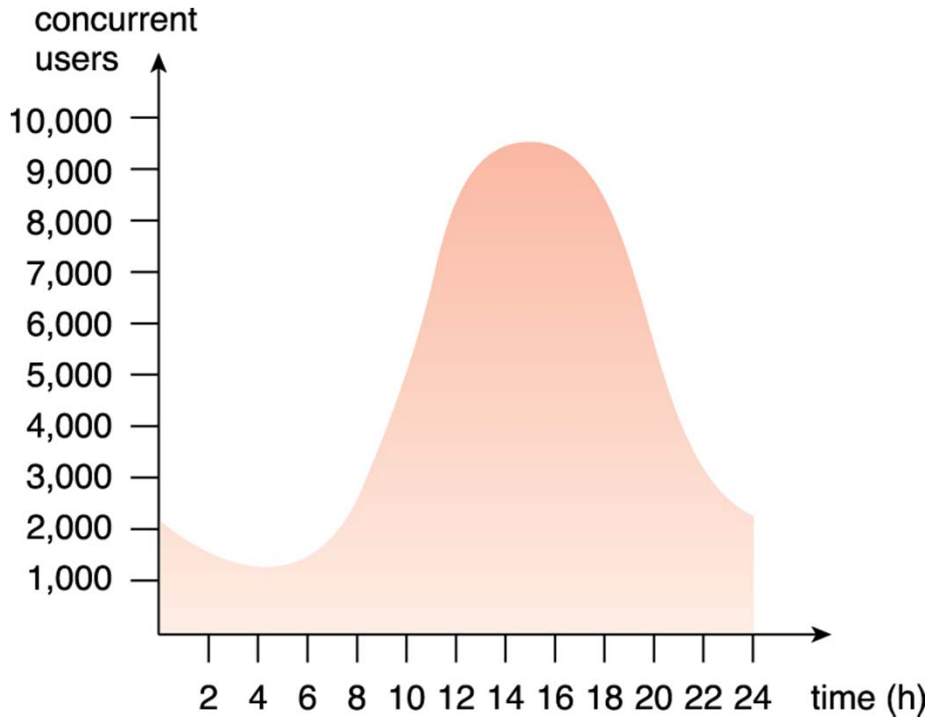
- Reduced investments and proportional costs
- Increased scalability (elasticity)
- Increased availability and reliability

Reduced Investments and Proportional Costs

- Reduction or elimination of upfront capital expenditures – hardware, software and ownership costs
- Proportional costs – cloud measured usage allows measured operational expenditures (directly related to business performance) to replace anticipated capital expenditures
- Cost reduction allows capital to be directed to the core business of the company
- Benefits: pay-as-you-go on-demand access, unlimited computing resources, add or remove IT resources at a fine-grain level, applications are not locked into devices or locations – can be easily moved

Increased Scalability (Elasticity)

- Empower cloud consumers to scale cloud resources to accommodate workload fluctuations and peaks



- Meets unpredictable usage demands and avoid potential loss of business

Increased Availability and Reliability

- Tangible business benefits: service outages result in loss revenue

- Service outages:
  - Gmail: Feb 2011 – 120,000 of 150m users without email over the weekend
  - Paypal: Nov 2010 – offline for 4.5 hours (network equipment failure)
  - Facebook: Sep 2010 – offline for 2.5 hours (database overload)
  - Microsoft Azure Cloud: Mar 2008 – down for 22 hours
  - Dropbox: Feb, 2013 – down for 48 hours (200m users)

- Availability: Gmail [99.984%(2010), 99.983%(2012)]

  99.984% = ? downtime per year

Technical Challenges

- Programming is tricky but improving
- Tools are continuously evolving
- Moving large data is still expensive
- Security
- Quality of service
- Green computing
- Internet dependence
- …

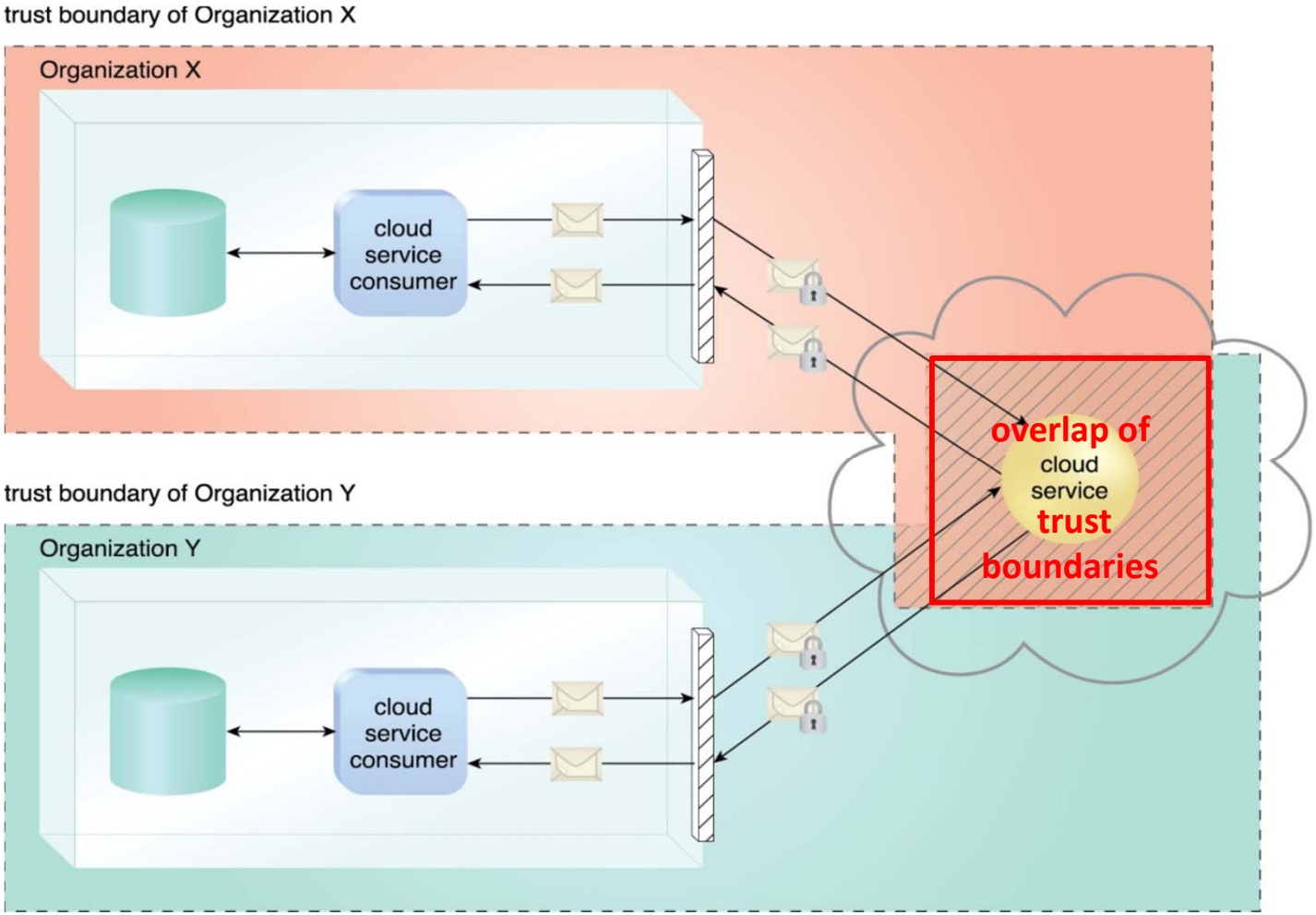
Non-Technical Challenges

- Increased security vulnerabilities
- Reduced operational governance control
- Limited portability between cloud providers
- Multi-regional compliance and legal issues
- Service-level agreements
- …

Increased Security Vulnerabilities

- Responsibility over data security becomes shared with cloud providers
- Remote usage requires expansion of trust boundary and introduced new vulnerabilities
- Shared IT resources across different cloud consumers introduce overlapping trust boundaries
- Overlapping trust boundaries introduce opportunities for malicious cloud consumers to steal or damage business data

Overlapping Trust Boundaries

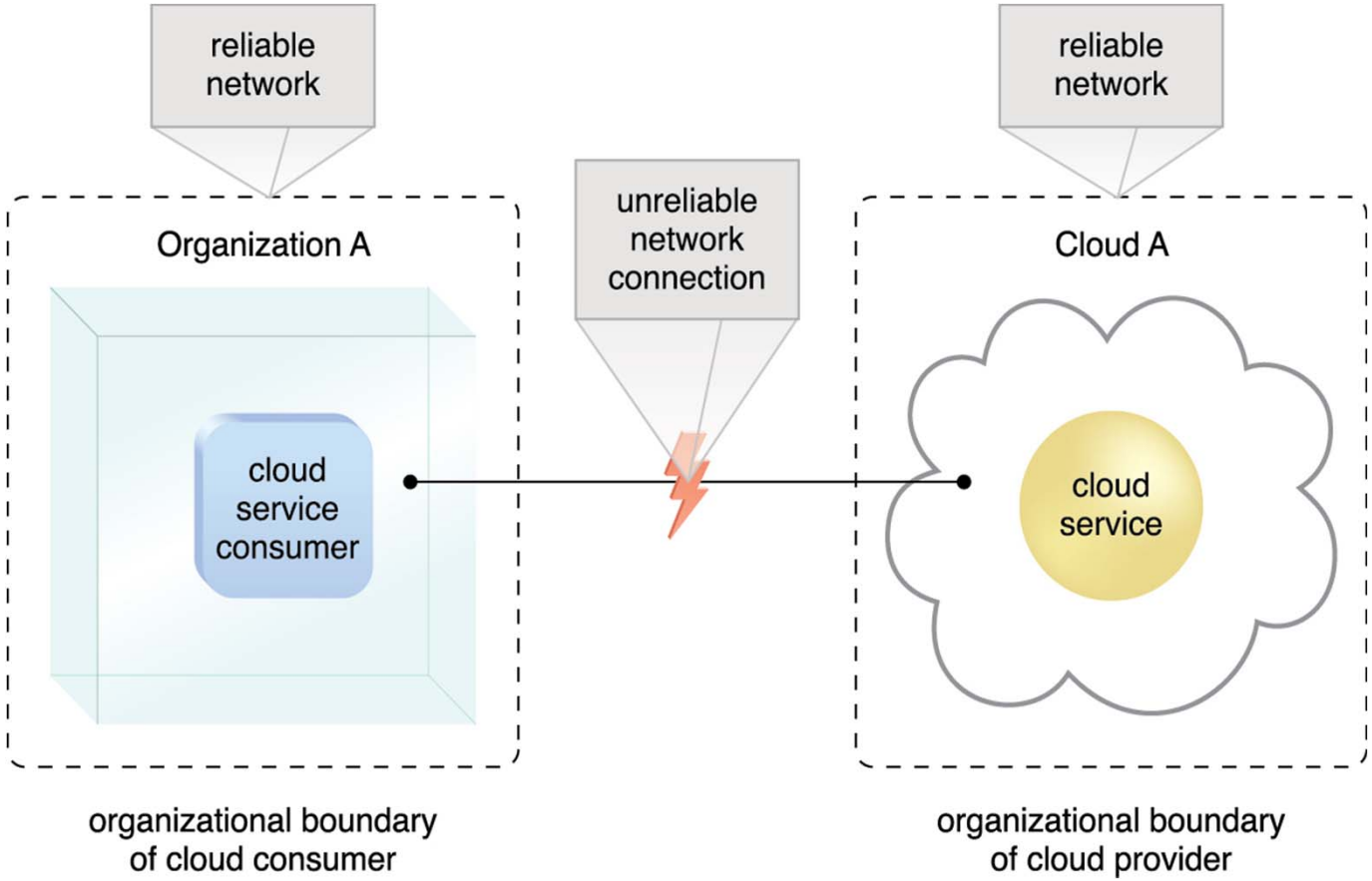


Shaded area with diagonal lines indicates the overlap of trust boundaries between Organization X and Organization Y

Reduced Operational Governance Control

- Cloud consumers are usually allocated a level of governance lower than on-premise IT resources
- An unreliable cloud provider may not maintain the SLA guarantees
- Longer geographical distance between cloud consumer and provider may introduce additional fluctuating latency and bandwidth constraints
- Mitigate by legal contracts combined with SLAs, technology inspections and monitoring

Reduced Operational Governance Control

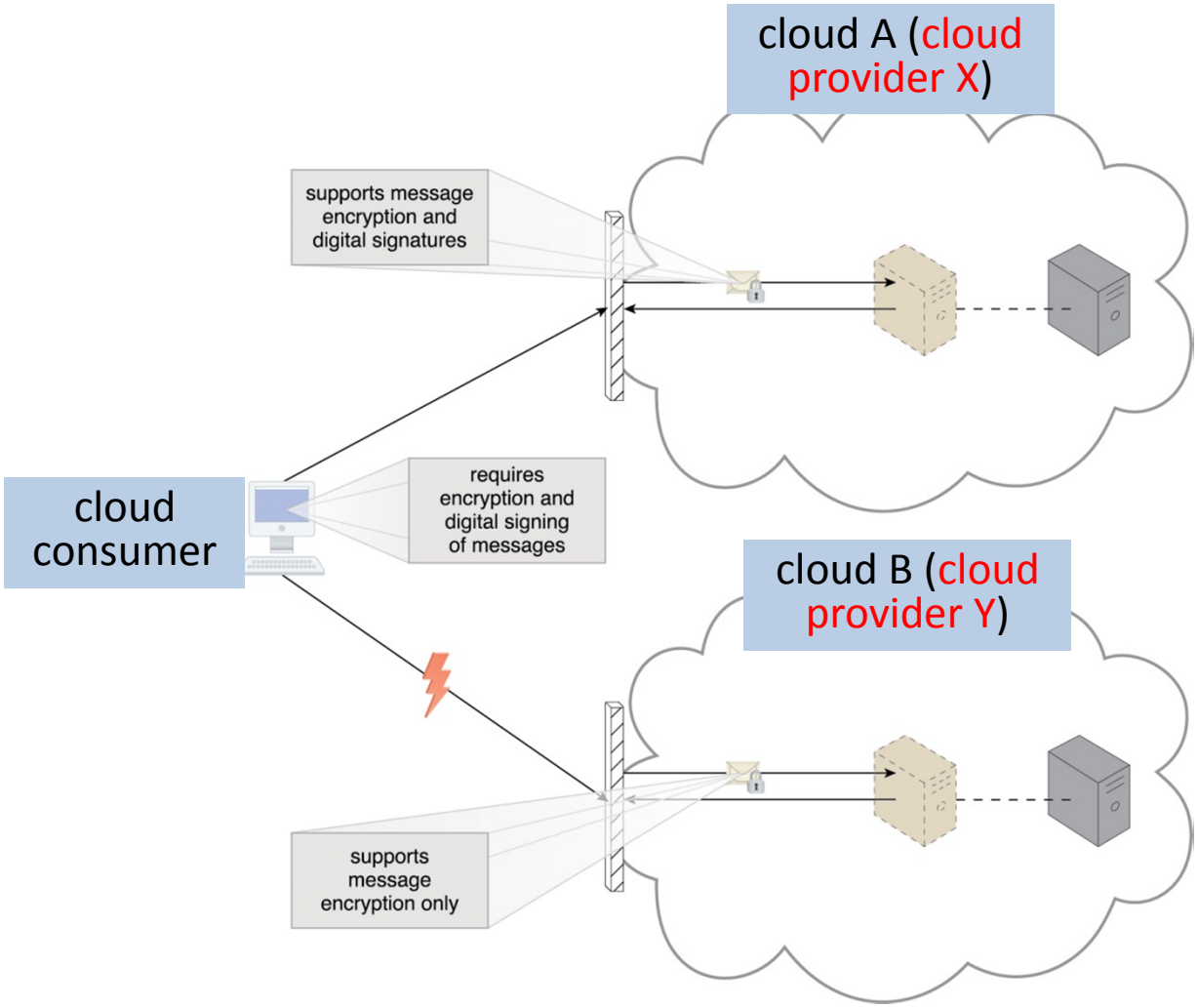


An unreliable network connection compromises the quality of communication between cloud consumer and cloud provider environments.

Limited Portability between Cloud Providers

- Public clouds are commonly proprietary due to the lack of established industry standards -> vendor lock-in
- Challenge of moving custom-built solution with environment dependencies across cloud providers

Limited Portability between Cloud Providers



A cloud consumer's application has a decreased level of portability when assessing a potential migration from Cloud A to Cloud B, because the cloud provider of Cloud B does not support the same security technologies as Cloud A.

Multi-regional Compliance and Legal Issues

- Data centers are setup in affordable or convenient geographical locations
- Industry or government regulations on data privacy or storage policies, e.g., personal data belonging to UK citizens be kept within UK
- Legal issues pertaining to accessibility and disclosure of data, e.g., laws require data to be disclosed to government agency

Summary

- Understanding cloud computing
- Cost model
- Main business drivers
- Elasticity and on-demand computing
- Types of scaling - horizontal and vertical
- Technical and non-technical challenges

Key Terms

1. Elastic resource
2. Availability
3. Capacity planning (resource provisioning)
4. Scaling (horizontal, vertical)
5. Cloud-based IT resources
6. Cloud service
7. Trust boundary

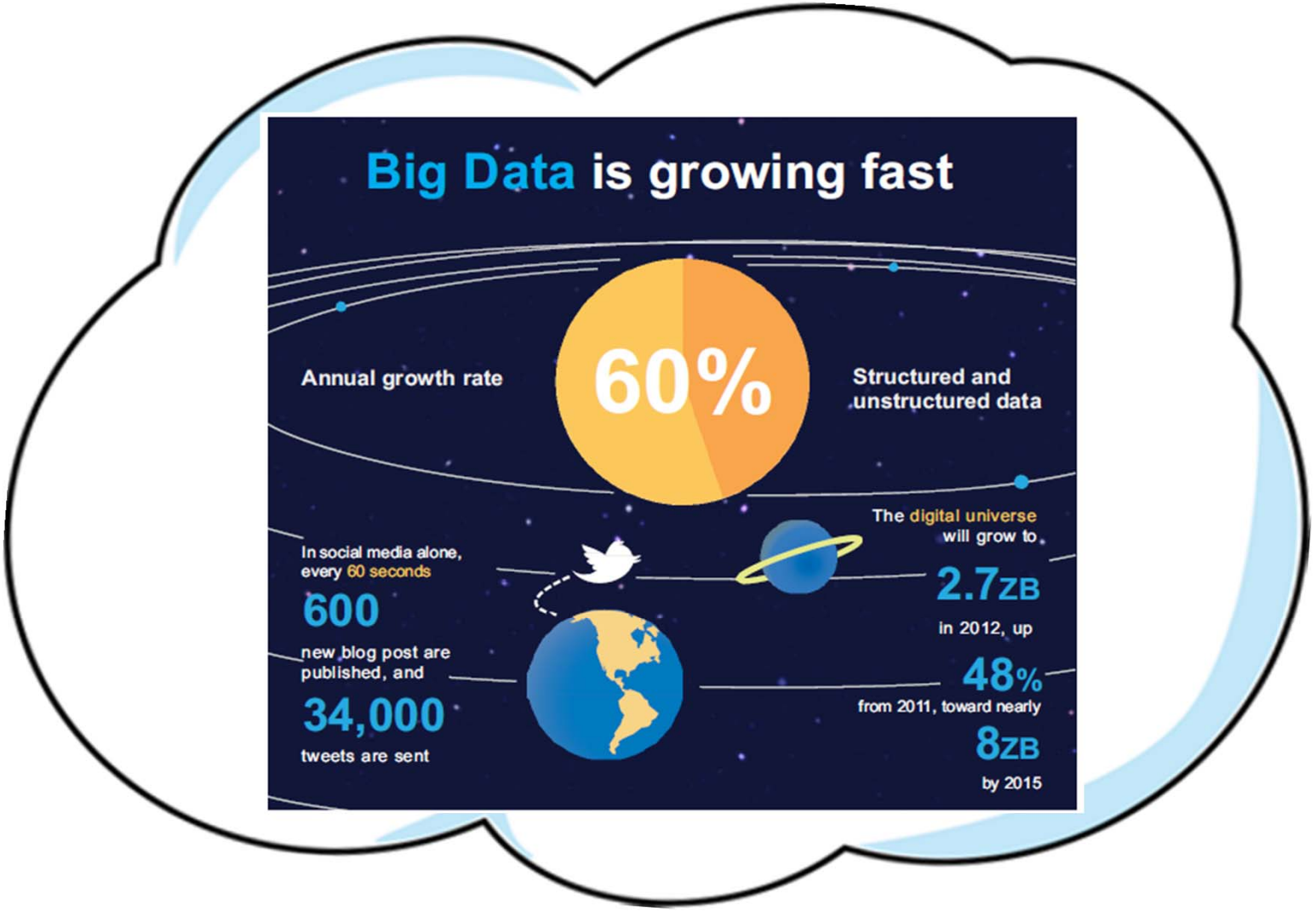
References

- Chapter 3, Cloud Computing: Concepts, Technology and Architecture, Thomas Erl, Zaigham Mahmood and Ricardo Puttini, Prentice-Hall, 2013.

- Above the Clouds: A Berkeley View of Cloud Computing, 2009.

- The NIST Definition of Cloud Computing, NIST Report, 2011.

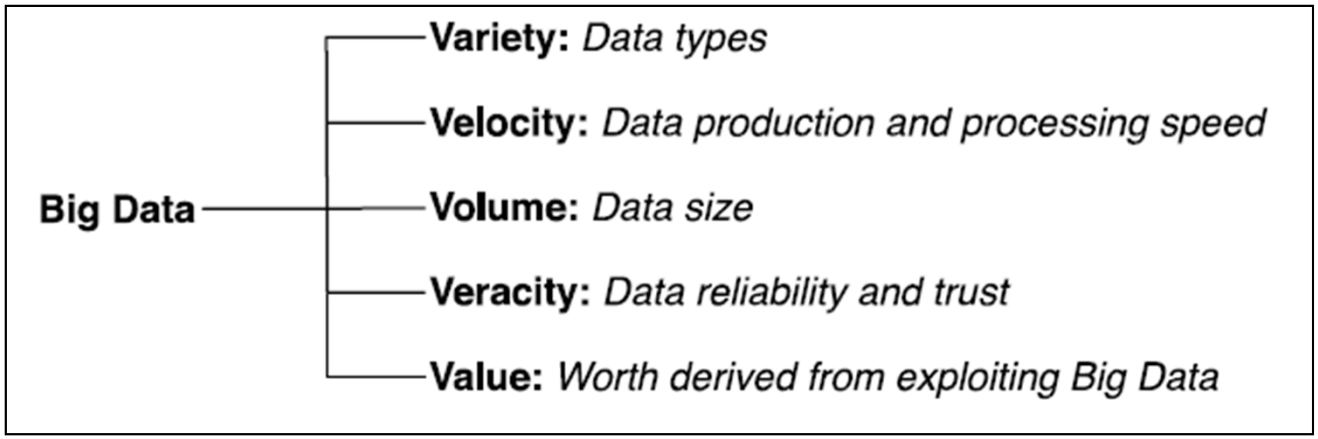
3. Cloud-enabled Data Analytics

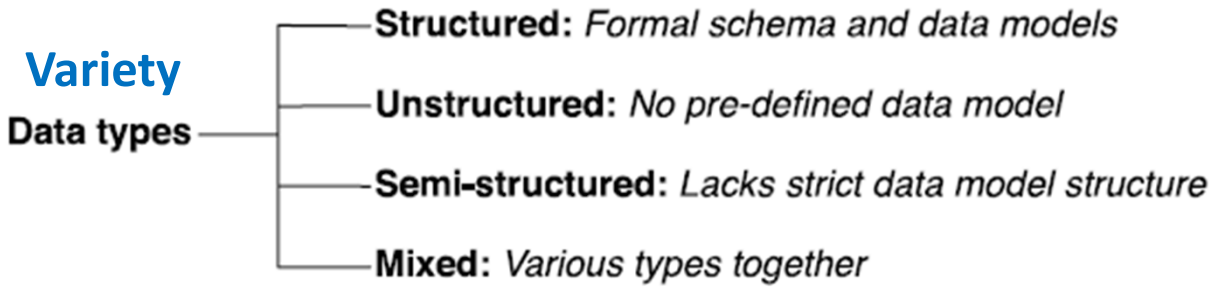


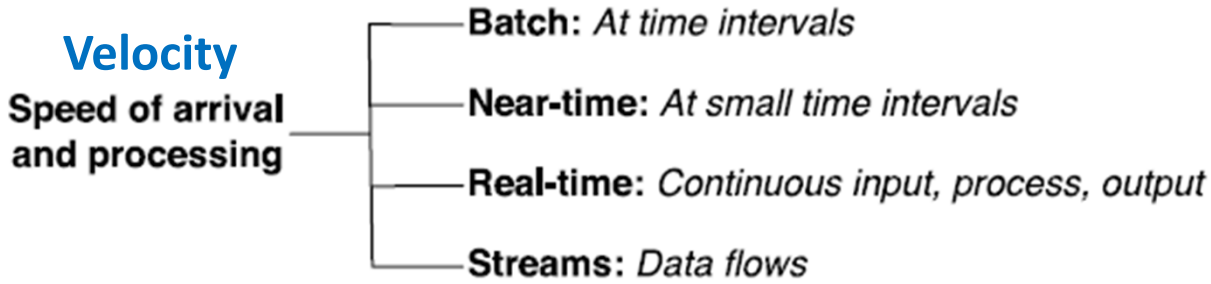
Outline

- Cloud analytics
- Analytics workflow for big data
- 4 Key Areas
  - Data Management
  - Model Building and Scoring
  - Visualization and User Interaction
  - Business Models
- Summary

Big Data


Big Data
Variety: Data types
Velocity: Data production and processing speed
Volume: Data size
Veracity: Data reliability and trust
Value: Worth derived from exploiting Big Data

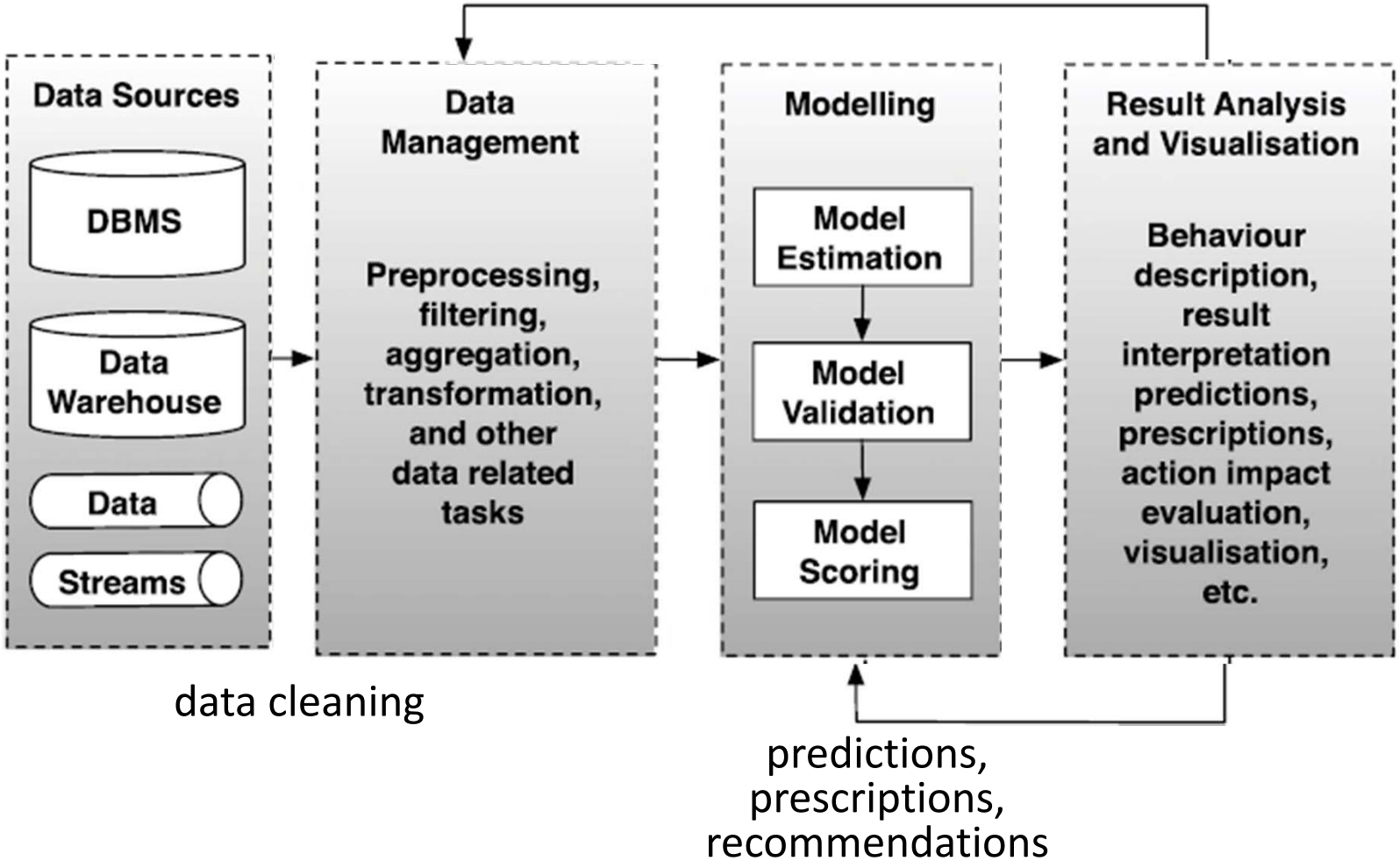

Variety
Data types
Structured: Formal schema and data models
Unstructured: No pre-defined data model
Semi-structured: Lacks strict data model structure
Mixed: Various types together


Velocity
Speed of arrival and processing
Batch: At time intervals
Near-time: At small time intervals
Real-time: Continuous input, process, output
Streams: Data flows

Cloud Analytics

- Purpose: exploits insights from data – a key competitive advantage
- Consumer's needs and demands => Big data + data analytics using **cloud resources** (reduces IT capital and operational expenditures)
- Analytics - mine structured and unstructured data from **private data** and **public data** on the web (tweets, blogs, product evaluations, social networks, ..)
- Cloud computing provides a (service) **delivery model** for hosting analytics solutions and used of these services
- Cloud analytics issues: data management, integration and processing; tuning of models; security and privacy; data quality; data currency

Analytics Workflow for Big Data



Analytics Solutions

- Descriptive: model past behavior
- Predictive: forecasts based on available data
- Prescriptive: assess actions, assist decision making

Four Key Areas

- Data management
- Model building and scoring
- Visualization and user interaction
- Business models

DATA MANAGEMENT

- Cloud deployments – data + analytics models
- Challenges

Cloud Deployments: Data + Analytics Models

1. Data and models are **private**
2. Data is **public**, models are **private**
3. Data and models are **public**
4. Data is **private**, models are **public**

Challenges in Big Data Management

1. Data *variety*
2. Data *storage*
3. Data *integration* solutions
4. Data *processing* and *resource management*

Model Building and Scoring

Work	Goal	Service model	Deployment model
Guazzelli et al. [64]	Predictive analytics (scoring)	IaaS	Public
Zementis [138]	Data analysis and model building	SaaS	Public or private
Google Prediction API [59]	Model building	SaaS	Public
Apache Mahout [11]	Data analysis and model building	IaaS	Any
Hazy [88]	Model building	IaaS	Any

Challenges:

1. Techniques to explore cloud rapid elasticity
2. “prediction and analytics as services” providers lock in

VISUALIZATION AND USER INTERACTION

- Analysis needs good visualization support
- Considers quality of data and presentation to facilitate navigation
- Typically, data analytics tasks (processing) used the **batch-job** model, big data visualization is not real-time due to computational complexity of analytics operations

BUSINESS MODELS - EXAMPLES

1. Hosting customer analytics jobs in a shared platform
2. full service stack to provide customers with end-to-end solution
3. Expose analytics models as hosted services

Summary

- 5 V's
- Big data analytics workflow
- Four important areas

Reference

- M. Assuncao et al., Big Data Computing and Clouds: Trends and Future Directions, Journal of Parallel & Distributed Computing, 79-80 (2015), 3-15.

L01: Introduction

1. Three Eras of Computing
2. Cloud Computing
3. Cloud-enabled Data Analytics

Key Terms

1. Elastic resource
2. Availability
3. Capacity planning (resource provisioning)
4. Scaling (horizontal, vertical)
5. Cloud-based IT resources
6. Cloud service
7. Trust boundary

Exercise 1

What is total downtime per year for the following availability?

1. 99% = ?
2. 99.9% =
3. 99.95% =
4. 99.99% =
5. 99.999% =