JOHNS HOPKINS
CAREY BUSINESS SCHOOL

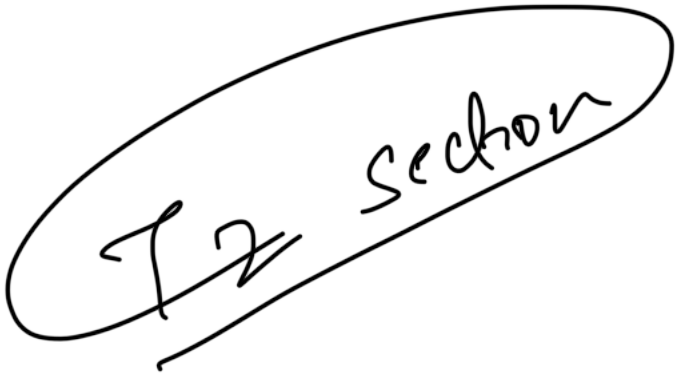
Lecture 1

BU.330.740 Large Scale Computing on the Cloud

Minghong Xu, PhD.
Associate Professor

Today's Agenda

- Course Overview
  - Requirements and evaluation
- Introduction to Cloud Computing and AWS
- Alumni Connection
  - Tuesday afternoon: **Mengting Chu**, MS BARM '23
  - Tuesday evening: **Yevhenii Lukianchuk**, MS IS '21
  - Wednesday afternoon: **Renfei Qiao**, MS IS '24
  - Wednesday evening: **Muhao Feng**, MS BARM '23



About the instructor

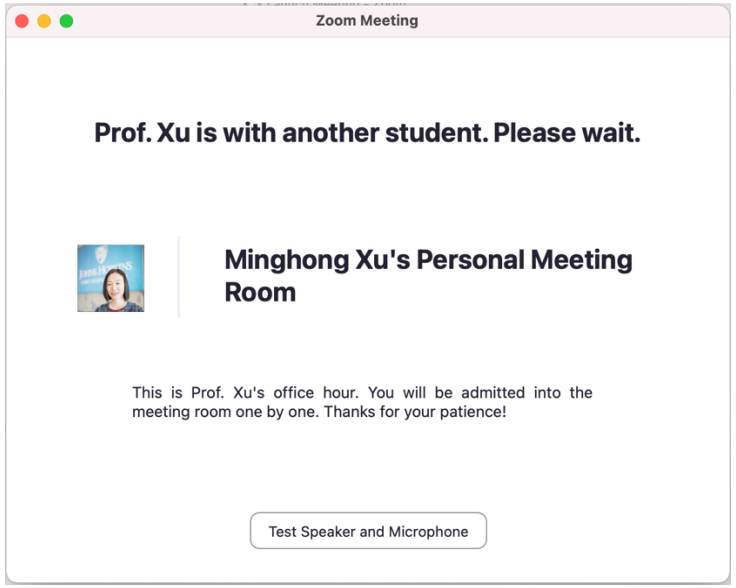
- Minghong Xu, Ph.D.
- Email: xu.minghong@jhu.edu
  - Best way to connect
  - **Please include your section # in your email**
- Affiliation: Center for Digital Health and Artificial Intelligence (CDHAI)
- Research and teaching: Artificial Intelligence and Big Data

Offices Hours

- Monday and Thursday 11:30am-1:30pm, and by appointment

https://jhucarey.zoom.us/j/4658557490?pwd=Y2NvL0M0RjdFb3RpUjlVOFBSSkFLZz09

- Link available on Canvas
- Waiting room enabled, admit one by one
- First come first served
  - unless book ahead of time



Teaching Assistants

- **Boxi Jiao** bjiao1@jh.edu
- MS IS, Carey 2024

- **Zibo Lin** zlin71@jh.edu
- MS BARM, Carey 2024

- **Biao Xiang** bxiang2@jh.edu
- MS BARM, Carey 2024

Course Overview

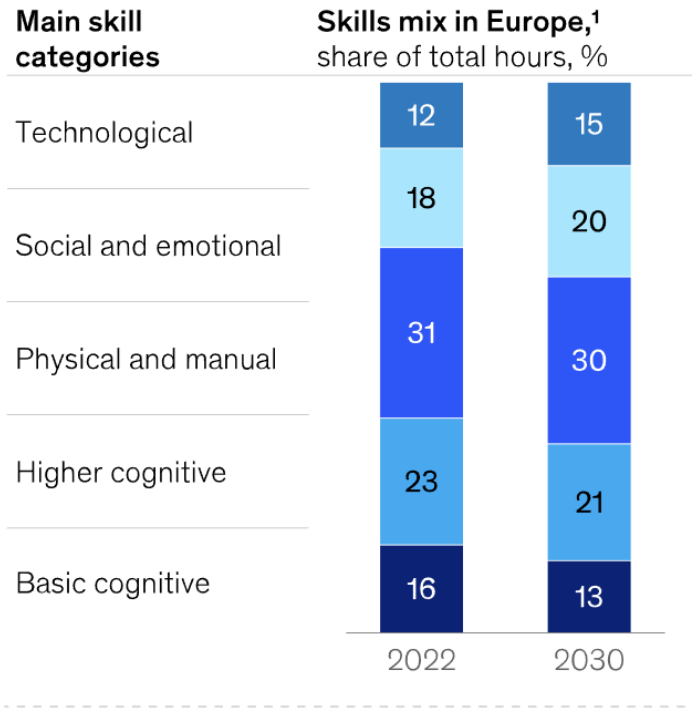
Cloud for AI and GenAI

- Large-scale AI models, such as ChatGPT , are typically trained on distributed systems/cloud computing platforms
- Cloud computing the critical skill for AI
  - To deploy AI models at scale
  - 17 Critical Skills for the AI "Techie"

- *"If the AI can replace my work, then I don't think I'm doing a good job"* - Generative AI Is About To Reset Everything, And, Yes It Will Change Your Life | Forbes

“Businesses will need a major skills upgrade”

- Demand for technological and social and emotional skills could rise
- Demand for physical and manual and higher cognitive skills stabilizes



Teaming Up (as of Jan 2025)

AI good at	Human good at
- Data input, storage, analysis - information processing - Language, read and write - Programming, machine languages	- physical and manual - knowledge, which requires higher level of cognition - Social and interpersonal - Tool usage, especially complex tool usage

Tips

- Raise your course learning to the knowledge level
- Acquire domain knowledge
  - Alumni connection sessions to boost you for job hunting
  - Resume polishing tools, e.g. customGPTs
- Explore more complex tools
  - We will gain extensive experience with cloud services
- Polish social skills
  - Keep improving your presentation through team project
- Keep exercising, stay physically active

Changes This Year

- Less theoretical framework, move towards a “tool” course
- Less coding, more practice on AWS
  - Only pySpark as core
  - Hive and Apache Pig move to the optional part
  - Email me if you would like more on this part
- Longer lab time, including 15 minutes buffer time
  - Extension materials available for you to explore more
- Increased importance of team project
  - Encourage more business ideas

Course Highlights

- Hands-on learning using Amazon AWS
- Create awareness of the technologies
- Explore breadth over depth

- Two Canvas sites: JHU and AWS Academy
  - JHU Canvas BU.330.740: Main LMS for all teaching materials, assignments submission and grading
  - **Please turn on JHU canvas notification**
  - AWS Academy: two courses on it

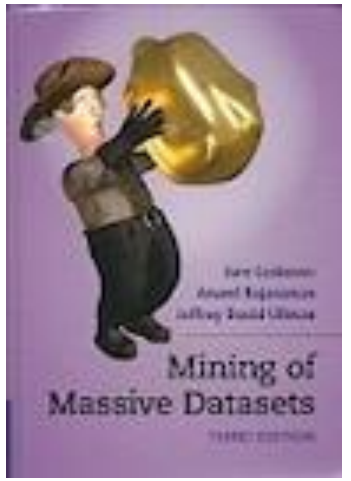
Two AWS Academy Courses

- AWS Academy Learner Lab [104575]
  - Mainly used in lab sessions, and for completing assignments and/or project
  - $50 credit per student to use some AWS services
- AWS Academy Cloud Foundations [104574]
  - Completely optional!
  - Help you to prepare for the AWS Certified Cloud Practitioner exam

- Invitation to join AWS Academy will be sent out on Friday
  - Titled “Course Invitation”
  - Use “@jh.edu”

Recommended Text for Data Mining

- **Mining of Massive Datasets**, by Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman; Cambridge University Press; 3rd edition (February 13, 2020), ISBN-13: 978-1108476348
- http://www.mmds.org
- The book is based on Stanford Computer Science courses Mining Massive Datasets and Data Mining
- Recommended if you want to go more technical



Requirements

Assignment	Weight
Attendance and participation in class discussion	5%
Homework	40%
Project	30%
Final Exam	25%
Total	100%

Class Participation (5%)

- **Please use your name tent**

- Class participation is an important part of learning
- Discussion questions every week

- Expect cold calls
  - First row exempted
  - Last row having the highest chance

Assignments (40%)

- 4 weekly assignments, each of 10%
- Due the next week before class
- Subject to TA interviews
- Responsible for your submission

Project (30%)

- Important assessment of this course
  - Business ideas become more important when AI can automate your work
- 4 students form a group, team roster due in week 2 before class
  - Email member names to TAs
- Business proposals using big data and cloud computing
- Open data on AWS: https://registry.opendata.aws
- I will prepare you in terms of data, programing and tools
  - Data mining examples every week, tools every week
  - AI-assisted programming in week 6

Project (Cont.)

- Detailed rubrics available on Canvas
- Present in week 7, final report due before class
- Report coverage
  1. Business problem or opportunity
  2. Big data
  3. Tools and method
  4. Preliminary results and findings

Final Exam (25%)

- In week 8, closed-book
  - More details will be announced later
- Administered via Respondus LockDown Browser
- Install LockDown Browser from https://download.respondus.com/lockdown/download.php?id=123533816
- Sample test will be posted on Canvas for you to test

Tentative Schedule

Week	Topic	Hands-on Learning	Due
1	Course introduction Overview of cloud computing and AWS		
2	Big Data framework: MapReduce Frequent Itemset mining and its business usage	Lab 1: AWS S3, EC2 and EMR	Project team due
3	Advanced Big Data framework: Spark Data engineering	Lab 2: pySpark in Google Colab	HW 1 due
4	Scaled machine learning pipeline Recommender system	Lab 3: AWS SageMaker	HW 2 due
5	Mining of massive dataset Computer vision and business applications	Lab 4: AWS Rekognition	HW 3 due
6	Natural language processing on cloud AI-assisted programming Project preview and final review	Lab 5: AWS Q Developer	HW 4 due
7	Project presentation		Project report due
8	Final exam		

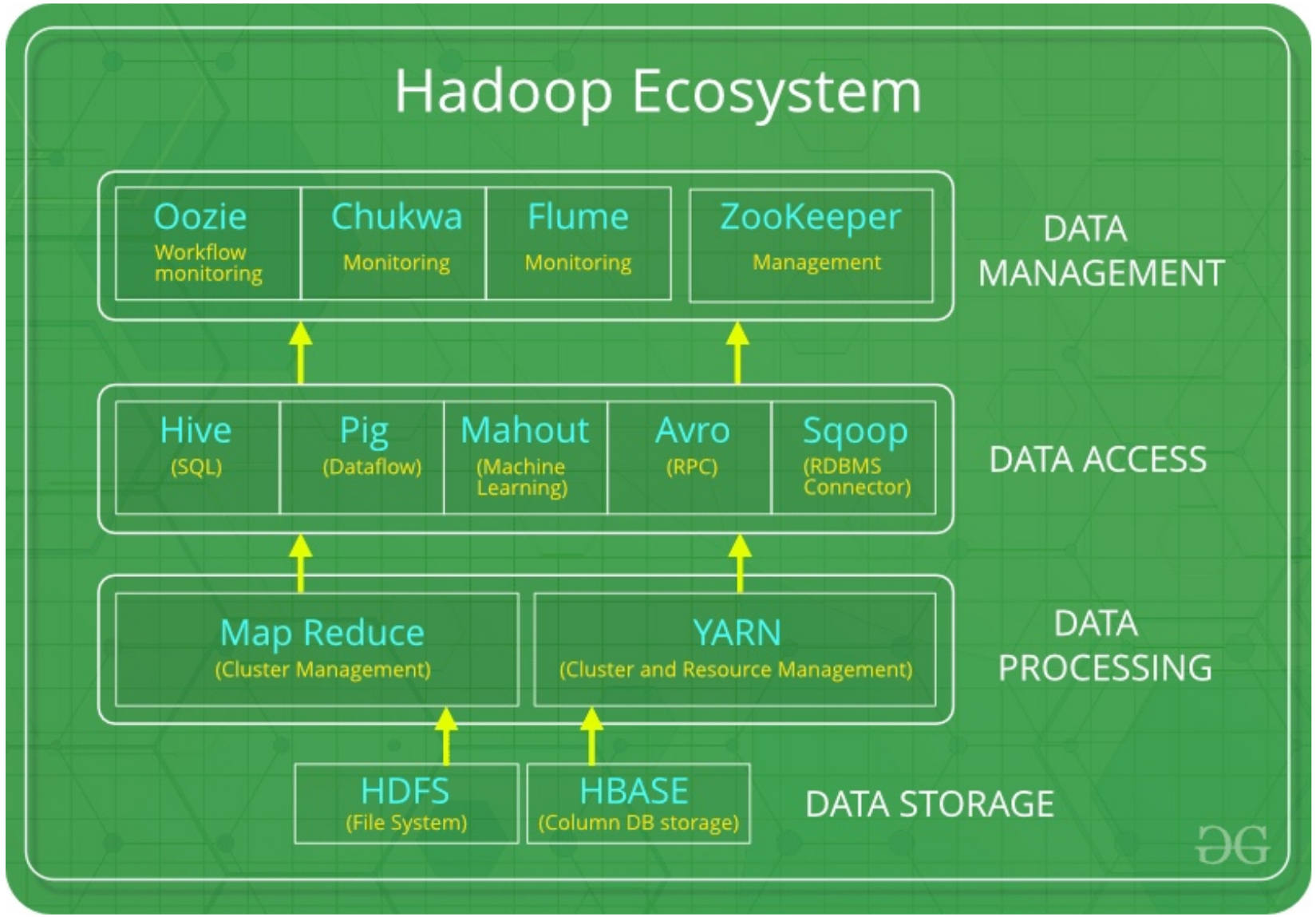
40%

10%

25%

Big data frameworks: are software platforms designed to store, process, analyse vast and complex datasets.

Examples: ① Apache Hadoop
② Apache Spark


Hadoop Ecosystem
Oozie
Workflow monitoring
Chukwa
Monitoring
Flume
Monitoring
ZooKeeper
Management
DATA MANAGEMENT
Hive
(SQL)
Pig
(Dataflow)
Mahout
(Machine Learning)
Avro
(RPC)
Sqoop
(RDBMS Connector)
DATA ACCESS
Map Reduce
(Cluster Management)
YARN
(Cluster and Resource Management)
DATA PROCESSING
HDFS
(File System)
HBASE
(Column DB storage)
DATA STORAGE

Policy Page

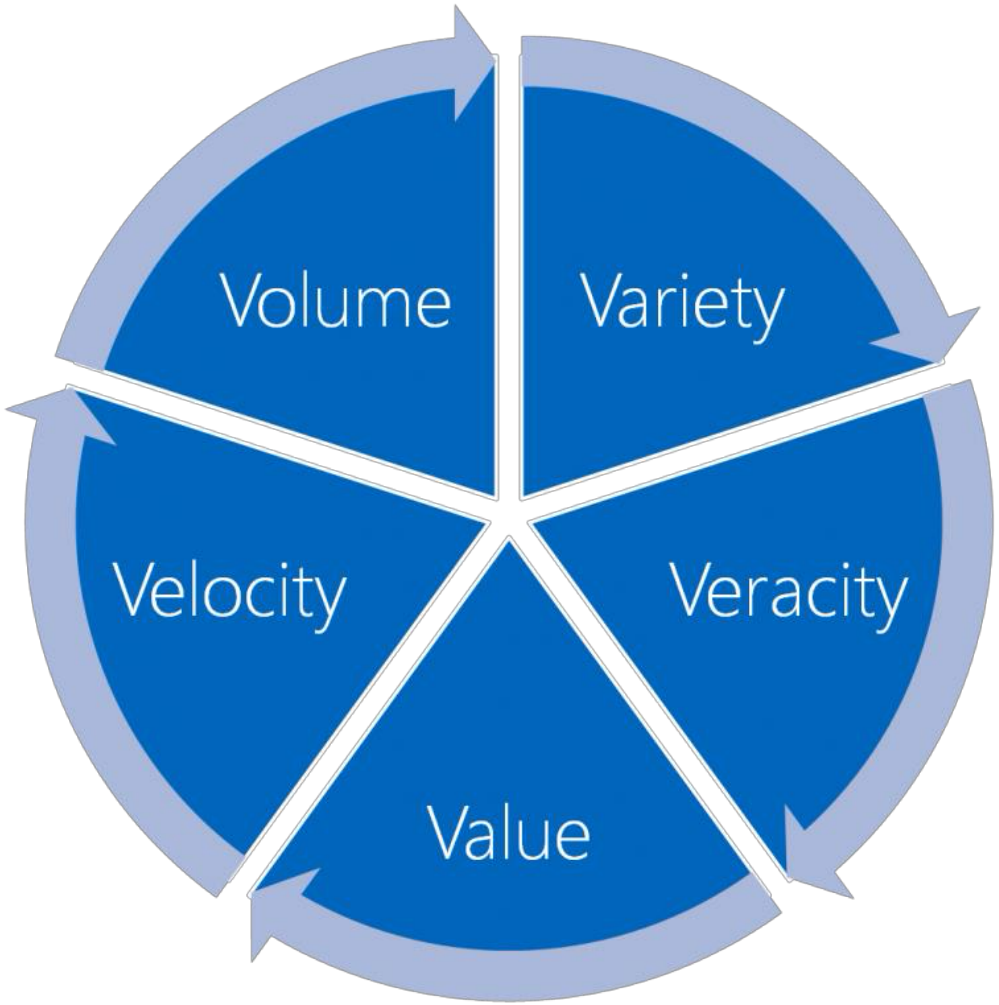
- No cellphones
- No excessive talking
  - Questions during lecture should be addressed to me
  - Suggest to sit closer to me
- Academic honesty (important!)
- Rectifying your score
  - Discuss with TA or me your concerns on grading asap, within 1 week after scores are posted
  - Request after 1 week may not be entertained

Final Grades

- I may curve up or curve down at the end

- A/A-: 10%
- B+: 40%
- B: 40%
- B- and below: 10%

What is Big Data and Large Scale Computing?



We see increasing volume of data, that grow at exponential rates

- Volume refers to the vast amount of data generated every second.
- *Question: How big a data set needs to be before it qualifies as "big data"?*
  - For some, any data set that won't fit in a spreadsheet
  - For others, if it won't fit on a single computer
  - **Or, if you actually need to use big data techniques to manage data across many connected computers** (*Geertsema, P. 2023. Machine Learning for Managers*)
- Distributed System

We see increasing velocity (or speed) at which data changes, travels, or increases

- Velocity refers to the speed at which new data is generated and the speed at which data moves around.

- Think of social media messages going viral in seconds.
- Technology now allows us to **analyze** the data while it is being generated (sometimes referred to as it in-memory analytics), without ever putting into databases.

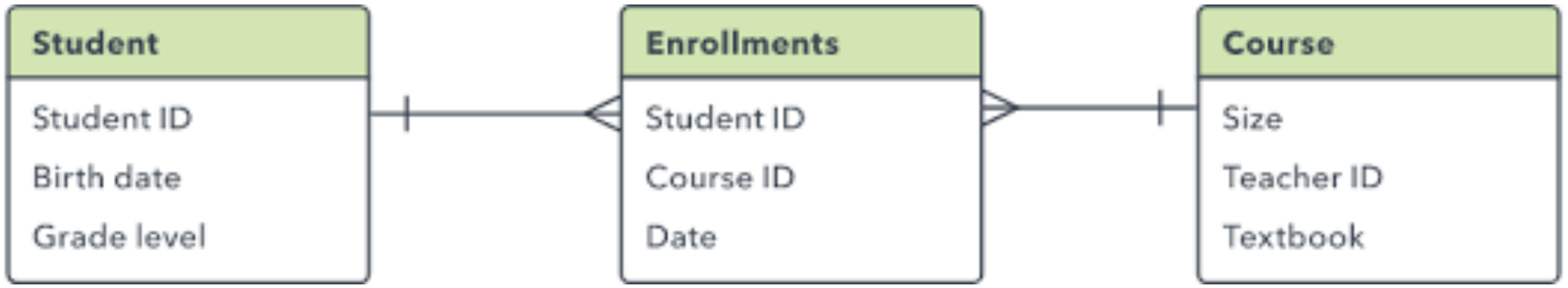
We see increasing variety of data types

- Variety refers to the different types of data we can now use.

- In the past we only focused on structured data
- In fact, 80% of world's data is unstructured
  - Our focus in this course

What is Structured Data?

- Well defined schema
- Usually in relational databased (RDBMS)
- Easily searchable



- Details in Database Management course

What is Unstructured Data?

- *"everything else"*
- Still have some internal structures
- But not structured via schema
- Stored in non-relational database like NoSQL
- Format: text, audio, video, social media postings
- "Analysts at Gartner estimate that upward of 80% of enterprise data today is unstructured" (*Forbes*)

Examples of Unstructured Data

- Text files: WSJ articles, 10-K analysis report, logs
- Mobile: text, locations
- Emails: school emails, colleague communications
- Social media: data from
  - Facebook, Twitter, LinkedIn, YouTube, Instagram
- Satellite imagery: weather, military movements
- Sensor data: traffic, production line

Differences between Structured and Unstructured

- Applications that use them
- Ease of analyzing

- We need analytical tools for mining unstructured data, some of them are developing

- With big data technology we can now analyze and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

We see increasing veracity (or accuracy) of data

- Veracity refers to messiness or trustworthiness of data.

- With many forms of big data quality and accuracy are less controllable (just think Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but technology now allows us to work with this type of data.

Value – The most important V of all!

- There is another V to take into account when looking at big data: Value.
- Having access to big data is no good unless we can turn it into value. Companies are starting to generate amazing value from their big data.
- *Goal? technology to store, process, analyze the Big Data*

Case: Go Game

- AlphaGo: 1202 CPUs and 176 GPUs
- AlphaZero: 5,000 first-generation TPUs and 64 second-generation TPUs in parallel

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Cloud Computing

What is Cloud Computing?

- Cloud computing is the on-demand delivery of compute power, database, storage, applications, and other IT resources via the internet with pay-as-you-go pricing.
  - Owning a car v.s. Ride sharing

- *Why "cloud"?*
- *So where is your computing?*
- https://www.youtube.com/watch?v=XZmGGAbHqa0

Discussion Time

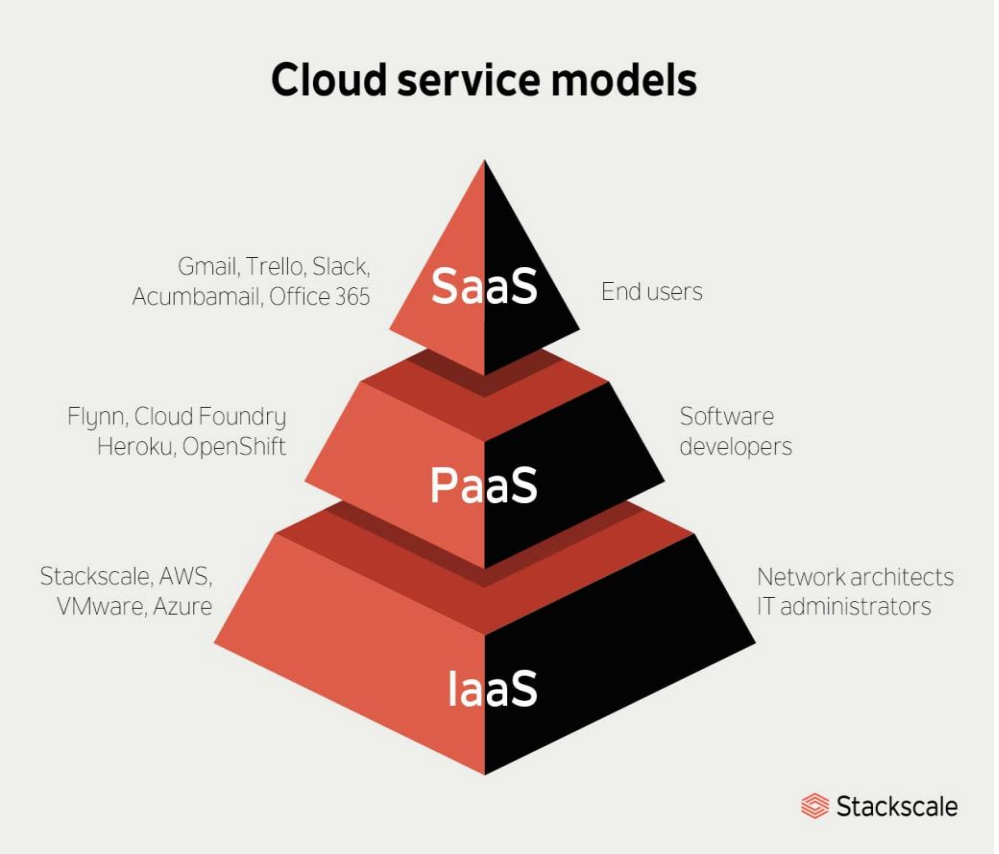
- *What are the advantages of using Cloud Computing?*
- *What are the disadvantages?*

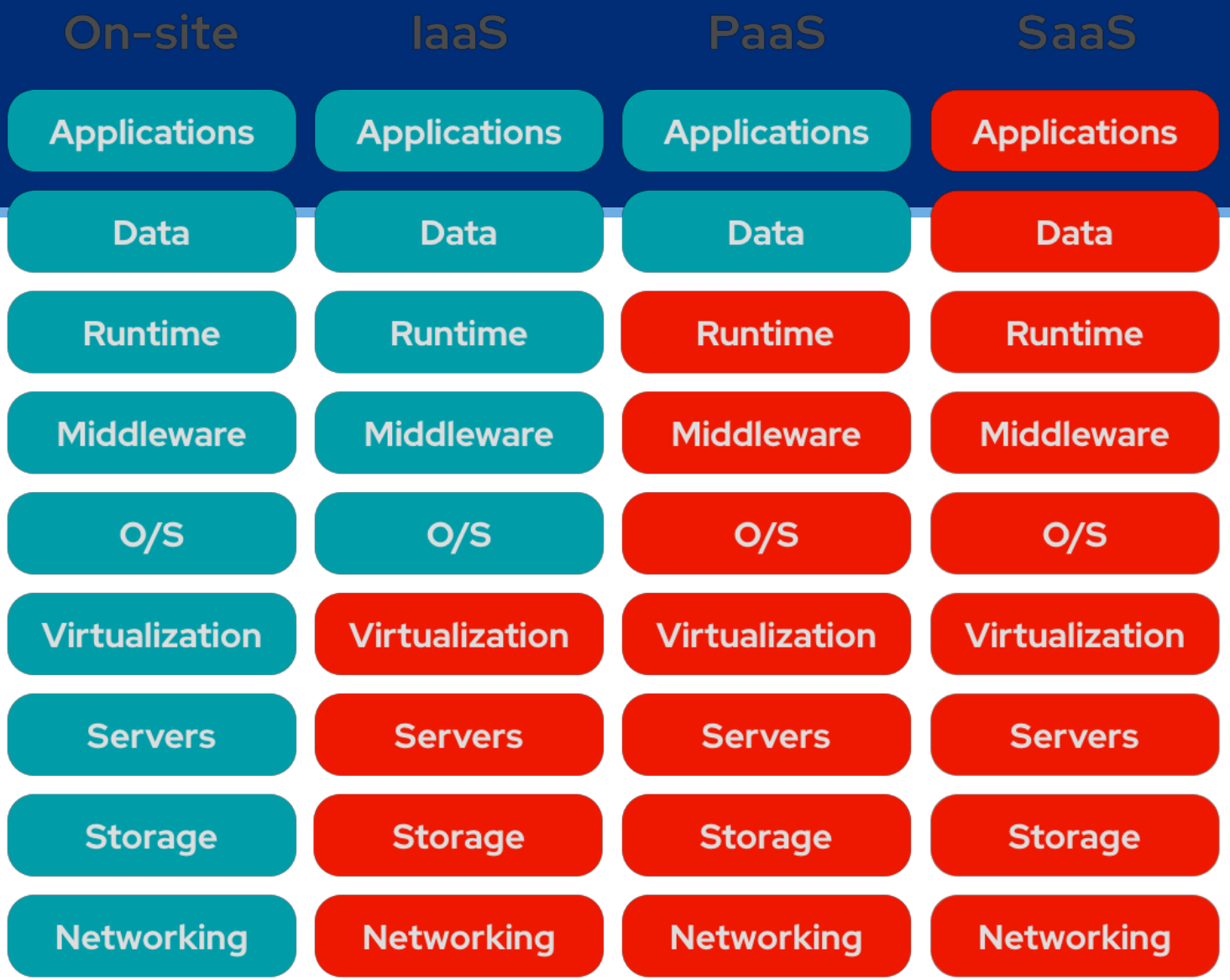
Advantages of Cloud Computing

- Trade capital expenses for variable expense
  - Pay only for the amount you consume
- Massive economies of scale
- Stop guessing capacity
  - Overestimate vs Underestimate vs Scale-on-demand
- Increase speed and agility
- Stop spending money on running and maintaining data centers
  - Focus on business and customers

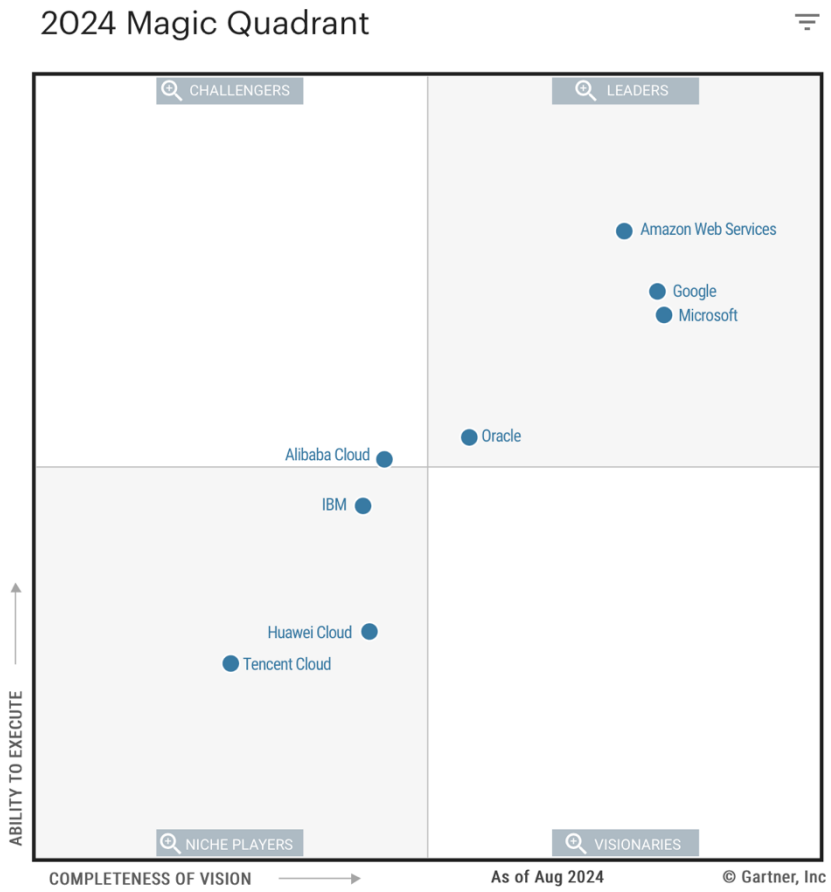
Three Subsets

- SaaS: Software as a Service
  - Developers manage and develop applications for enterprises
- PaaS: Platform as a Service
  - Provider makes its hardware and basic application suite available to developers and enterprises that create add-on apps
- IaaS: Infrastructure as a Service
  - Developers create apps on basic computing and storage infrastructure




On-site
IaaS
PaaS
SaaS
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
You manage
Service provider manages

Cloud Competition



"Over the past two years, the race for preeminence in generative AI has resulted in a deluge of new GenAI services from all providers, based on both NVIDIA and proprietary hardware designs."

https://databases.library.jhu.edu/databases/subject/Information%20Technology%20+%20Security

Case Studies

- Netflix:
  - https://www.youtube.com/watch?v=XrWII4ewrXA
- Expedia:
  - https://www.youtube.com/watch?v=zgzKTEAzddI&list=PLhr1KZpdzukfZxFGkA796dKaUufAgnU4c&index=13
- Google Cloud Disruption:
  - https://www.bloomberg.com/news/articles/2018-07-17/google-cloud-has-disruption-bringing-snapchat-spotify-down?leadSource=uverify%20wall

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Alumni Connection

Next Week

- **Project Team Due**
- MapReduce Framework
- Frequent Itemset Mining

References

- Kunpeng Zhang's notes (2019)
- AWS Academy Cloud Foundations
- A new future of work: The race to deploy AI and raise skills in Europe and beyond, McKinsey 2024