INFORMATION TECHNOLOGY UNIVERSITY

DATA SCIENCE LAB

Big Data Analytics

Dr. Faisal Kamiran

TAs

- Abdur Rehman Ali
  - abdurrehman.ali@itu.edu.pk
- Muhammad Ahmad Ehsan
  - msds18008@itu.edu.pk
- TA hours would be announced soon.

Classroom

- Please join the classroom for this course using the following code.

tftsbfu

Tentative Course Modules

1. Data Modeling with Relational Databases
2. Data Modeling with NoSQL Databases
3. Data Warehousing
4. Cloud Computing
5. Big Data Tools (Spark, Hive, Hbase, Hadoop, Map-reduce)
6. Data Wrangling with Spark
7. Debugging and Optimization
8. Streaming Data
9. Data Pipelines with Apache Airflow

What is Big Data Analytics?

- Big Data
  - Buzz Word
  - Datasets too large for modern relational databases.
  - Semi-Structured/ Unstructured Datasets
- Analytics
  - How to measure?
  - Identifying patterns in data.

Big Data Timeline

Brief History

- In 1970s traditional relational database systems, such as IBM's databases, enabled SQL and increased the adoption of data processing by wider audiences.
- SQL expanded the number and type of applications relevant to data processing such as business applications, analytics on average rates, average basket size, year-on-year growth figures, etc.
- In 1980s, Decision Support Systems, Information Systems emerged.
- In 1990s, Business Intelligence Systems start getting popular for BI reporting and business alalytics.

THE BIRTH OF THE INTERNET

1989 — Birth of Internet

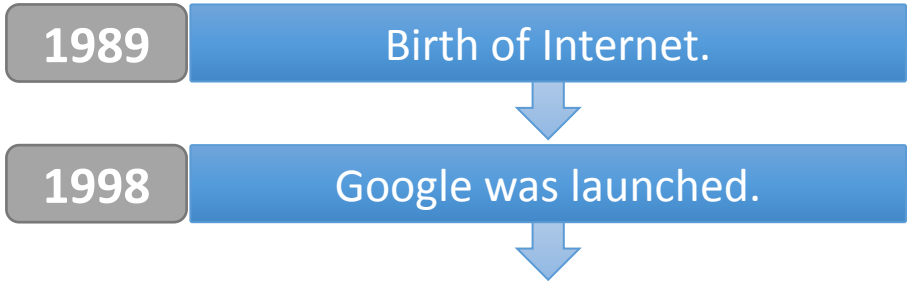
Courtesy of Sir Tim Berners-Lee (best known as the inventor of WWW), data and information was posted online for the first time in 1989.

https://en.wikipedia.org/wiki/Tim_Berners-Lee

The creation of data was spurred as more and more devices connected to the internet

GOOGLE WAS LAUNCHED

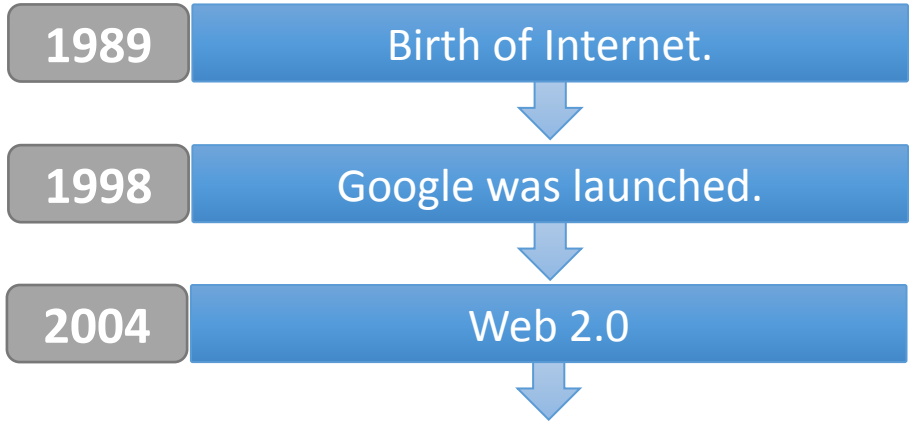
- The Google search engine was launched in 1998.
- Today, over 40,000 Google searches are completed every second, or over **7 billion** in a single day.
- https://www.internetlivestats.com/google-search-statistics/



Web 2.0

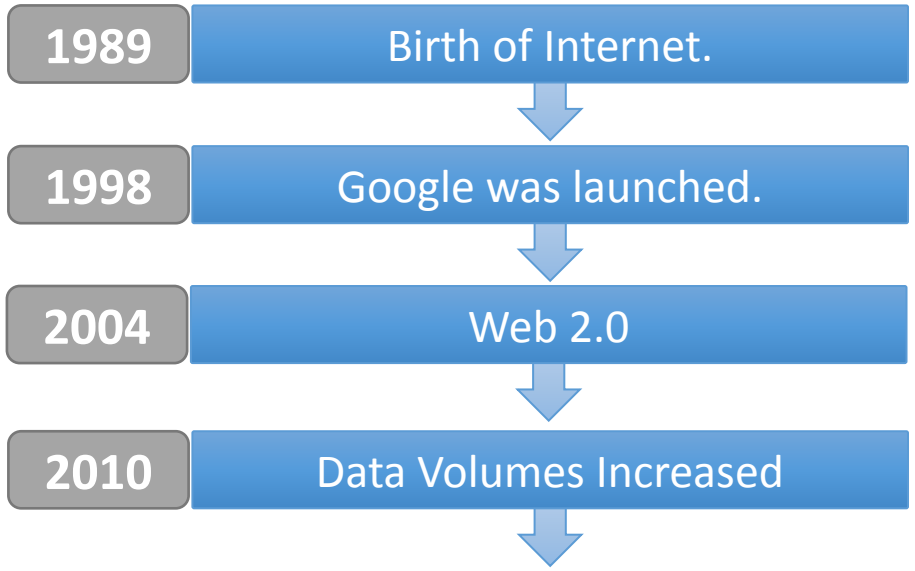
(Evolution of Unstructured Data)

- Second generation of the World Wide Web was introduced in 2004.
- Web 2.0 technologies provided a level of user interaction that was not available before.
- Web 2.0 provided a base for unstructured and user-generated data that included blogs, Wikis, social media platforms, etc.
- More and more data was created on a daily basis as many social networks appeared on the horizon.
  - MySpace, Facebook, etc.



DATA VOLUMES INCREASED

- In 2010, Eric Schmidt (CEO Google) while speaking at Technology conference in Lake Tahoe in California stated that "there were 5 exabytes of information created by the entire world between the dawn of civilization and 2003. Now that same amount is created every two days.“
- i.e. 5 x 10^9 GBs of data every two days



https://techcrunch.com/2010/08/04/schmidt-data/

RADIO FREQUENCY IDENTIFICATION - 2011

Over 12 million RFID (radio frequency identification) tags had been sold to monitor and track goods worldwide by 2011.

GLOBAL MARKET - 2013

The global Big Data market was said to be worth **$10bn** in 2013, and at that time, was predicted to rise to **$54bn** by 2017.

MOBILE DEVICES CAME OUT ON TOP - 2014

The number of people accessing the internet using mobile devices such as phones and tablets surpassed desktop users for the first time.

DAILY DATA - 2020

A whopping 1.14 Exabytes bytes of data was being created each day.

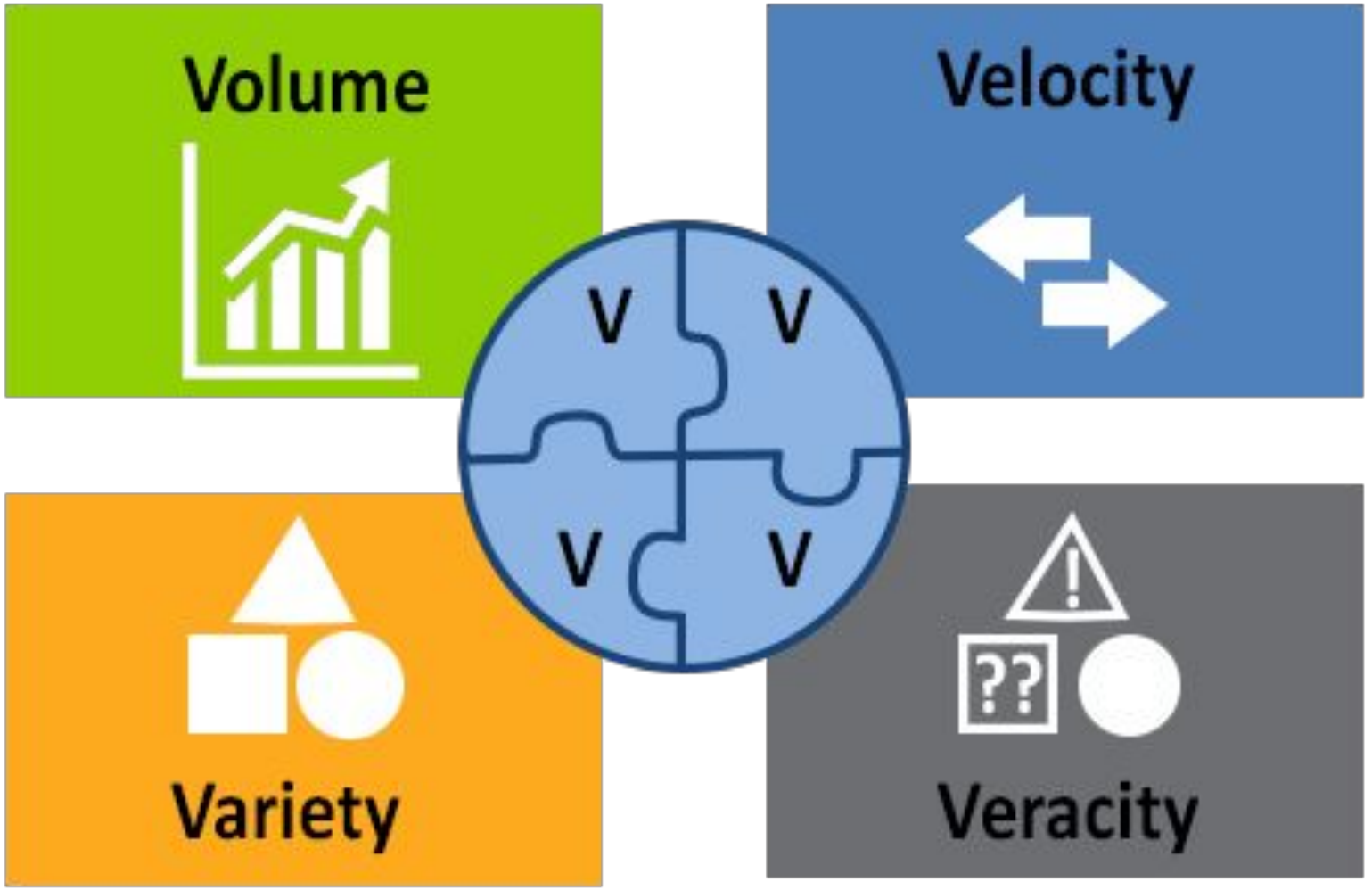
1.14 million terabytes
=
1.14 trillion megabytes.

IT'S ONLY GETTING BIGGER

According to **International Data Corporation as of April 4, 2019** – Worldwide revenues for big data and business analytics (BDA) solutions was worth $189.1 billion in 2019 with an increase of 12.0% over 2018.

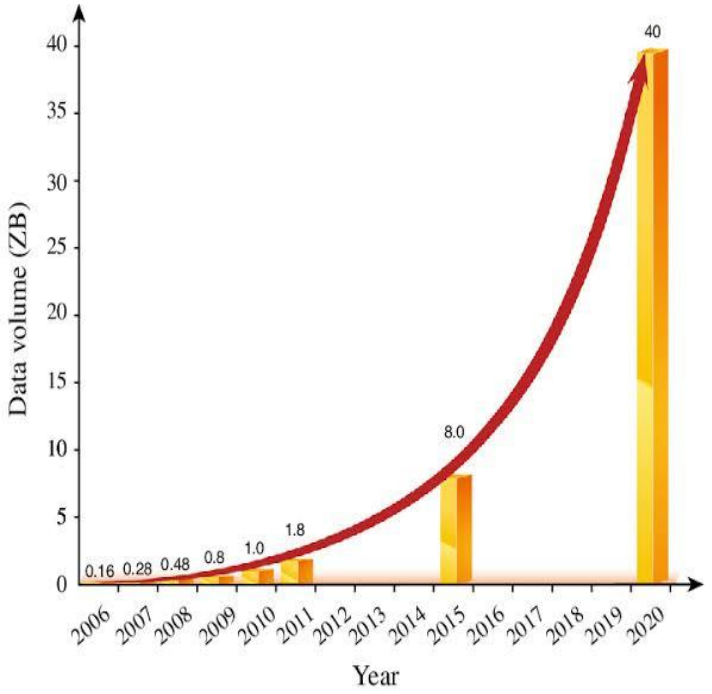
- The IDC forecasts that the big data and analytics market will reach $274 billion by the end of 2022.

Big Data Using 4 Vs



Data Volume

- A typical PC might have had 10 gigabytes of storage in 2000.
  - Today, WhatsApp users exchange up to 65 billion messages every day.
- More than 120 professionals join LinkedIn every minute.
- 88,000 YouTube videos are viewed every second.
- Twitter users send over 528,780 tweets every minute.
- Instagram users post 60,740 photos every minute.
- More than 300 million photos get uploaded per day on Facebook.
- Every minute there are 510,000 comments posted and 293,000 statuses updated
- Internet users generate about **2.5 quintillion bytes** of data each day.

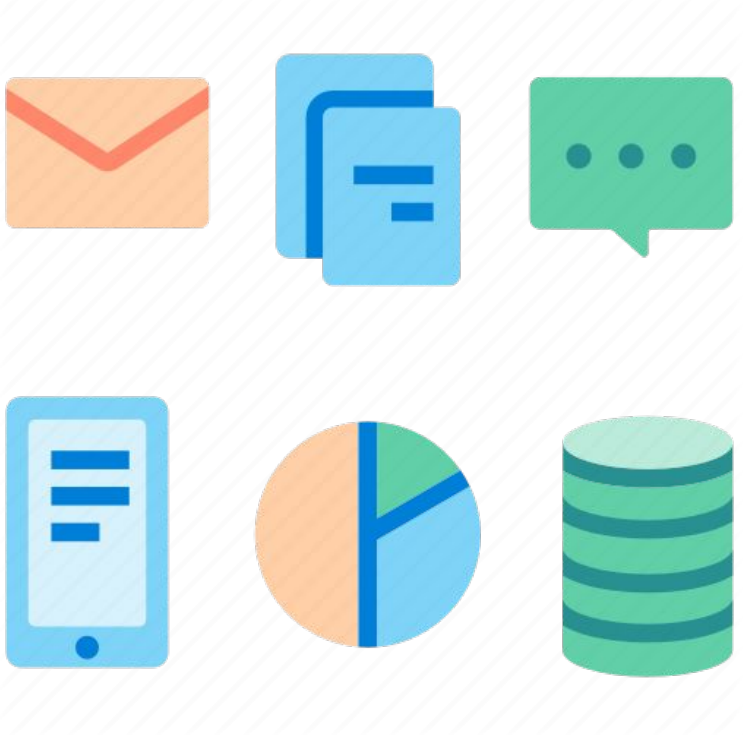


Data Velocity

- Clickstream and ad impressions capture user behavior at millions of events per second.
- High-frequency stock trading algorithms reflect market changes within microseconds.
- Machine to machine processes exchange data between billions of devices.
- Infrastructure and sensors generate massive log data in real-time.
- Online gaming systems support millions of concurrent users, each producing multiple inputs per second.

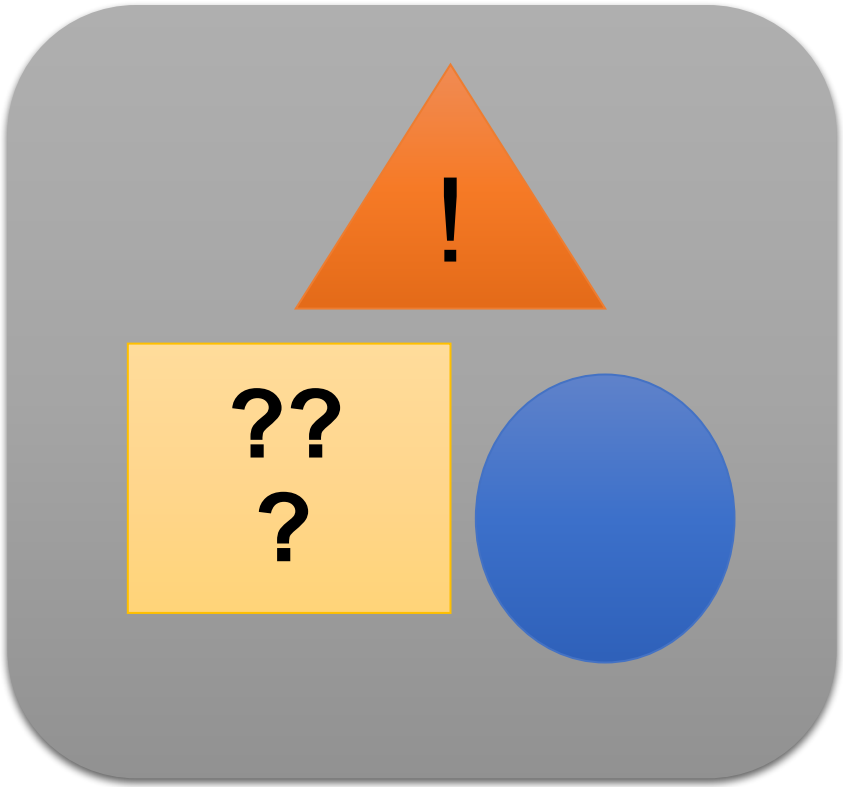
Data Variety

- Big Data isn't just numbers, dates, and strings. Big Data is also
  - Geospatial data
  - 3D data
  - Audio and video
  - unstructured text, including log files and social media.
- Traditional database systems were designed to address smaller volumes of structured data.

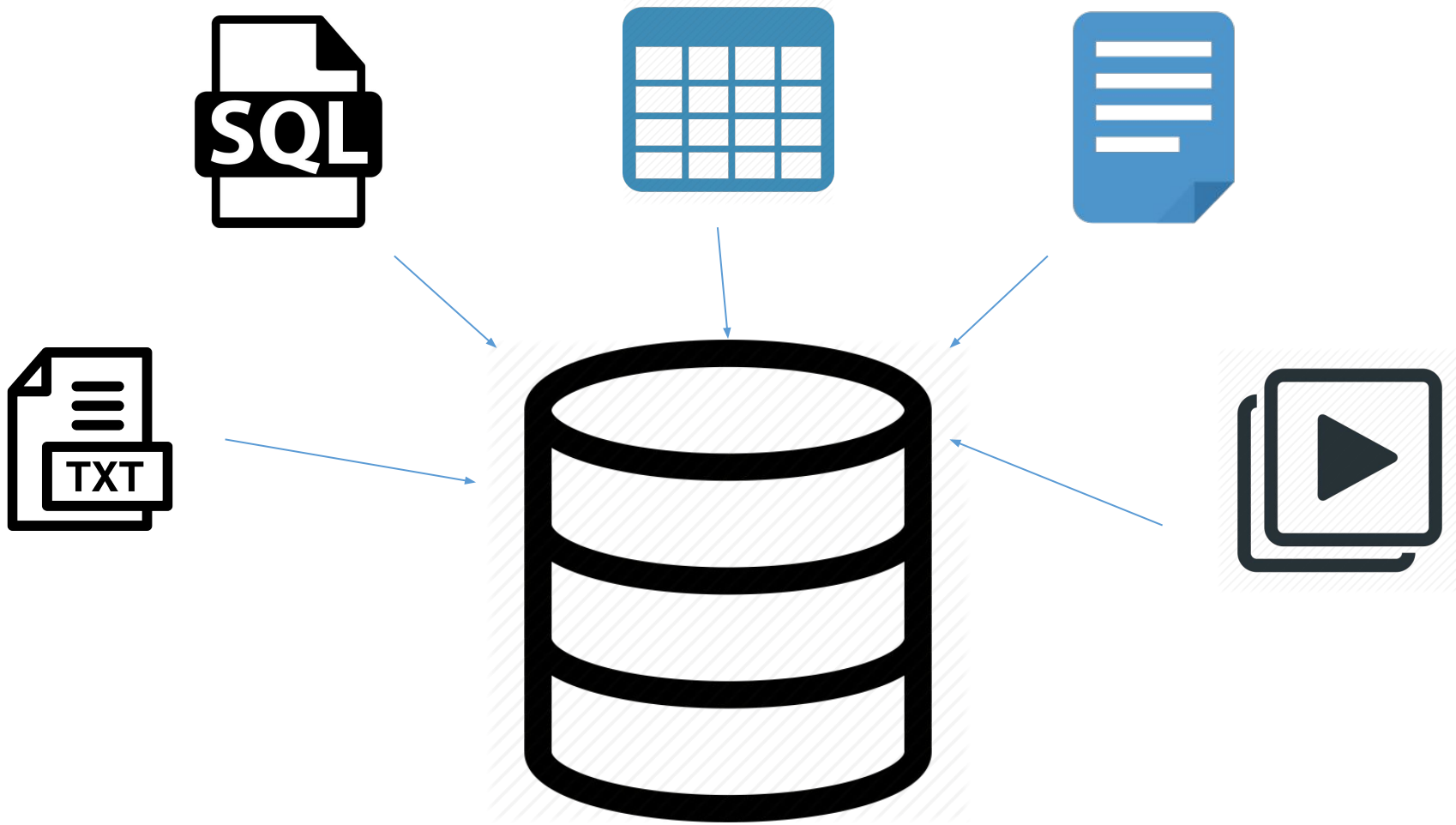


Data Veracity

- Biases, noise and abnormality in data
- Is the data that is being stored, and mined meaningful to the problem being analyzed?
- Need to have your team and partners work to help keep your data clean



New Approach to Data

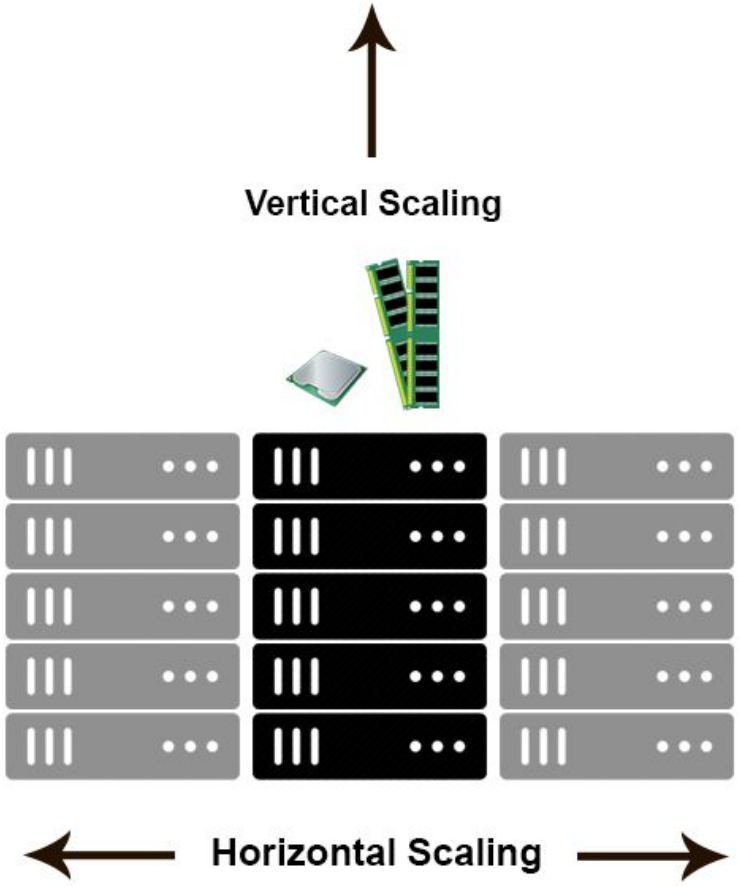


Big Data Analytics

- Where processing is hosted?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is stored?
  - Distributed Storage (e.g. Amazon S3)
- What is the programming model?
  - Distributed Processing (e.g. MapReduce)
- How data is stored & indexed?
  - High-performance schema-free databases (e.g., MongoDB)
- What operations are performed on data?
  - Analytic / Semantic Processing

Vertical Scaling Vs. Horizontal Scaling?

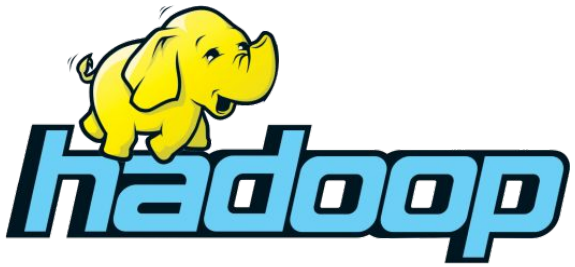
- **Horizontal scaling** means that you scale by adding more machines into your pool of resources
- **Vertical scaling** means that you scale by adding more power (CPU, RAM) to an existing machine.



DATA SCIENCE LAB

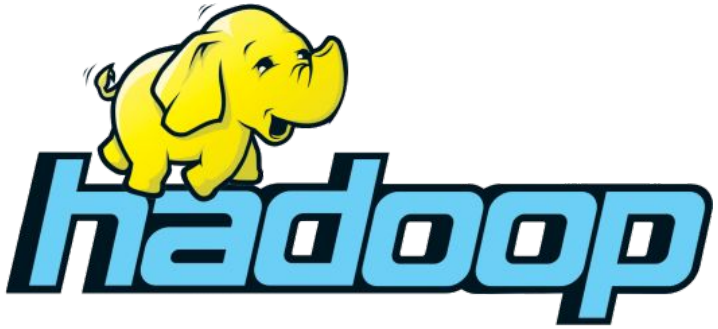
INFORMATION TECHNOLOGY UNIVERSITY

Big Data Platforms



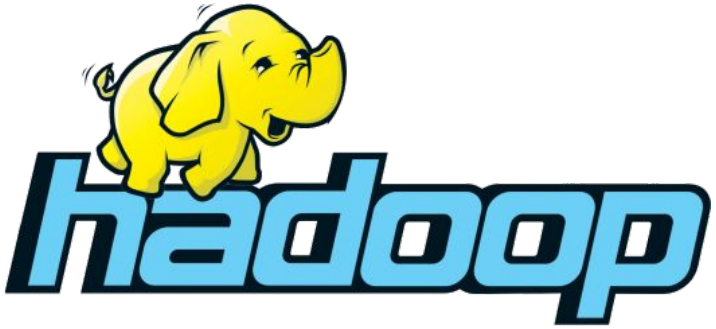
Big Data Platforms

- Hadoop
  - HDFS
  - Map-reduce
- Spark
  - RDD

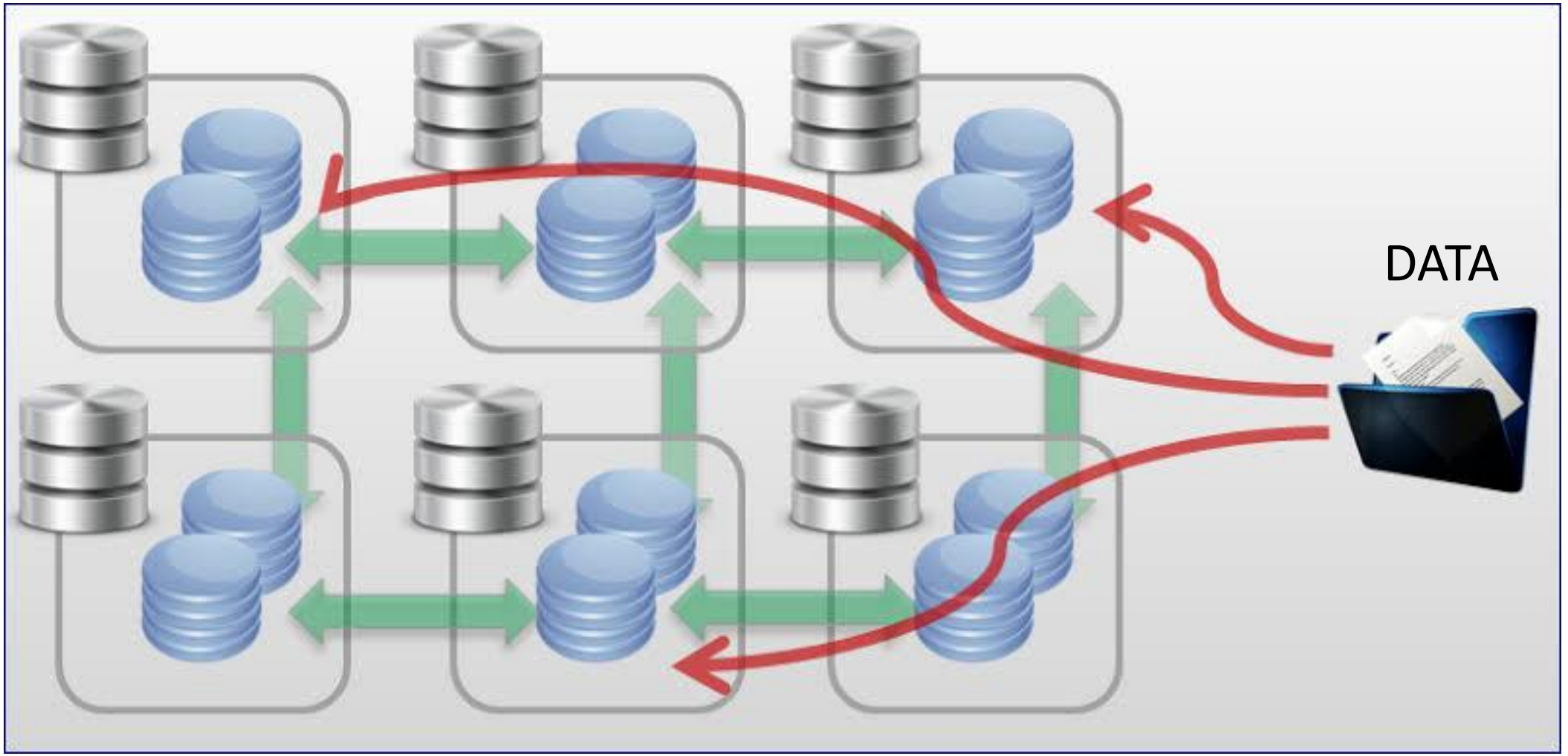


What is Hadoop?

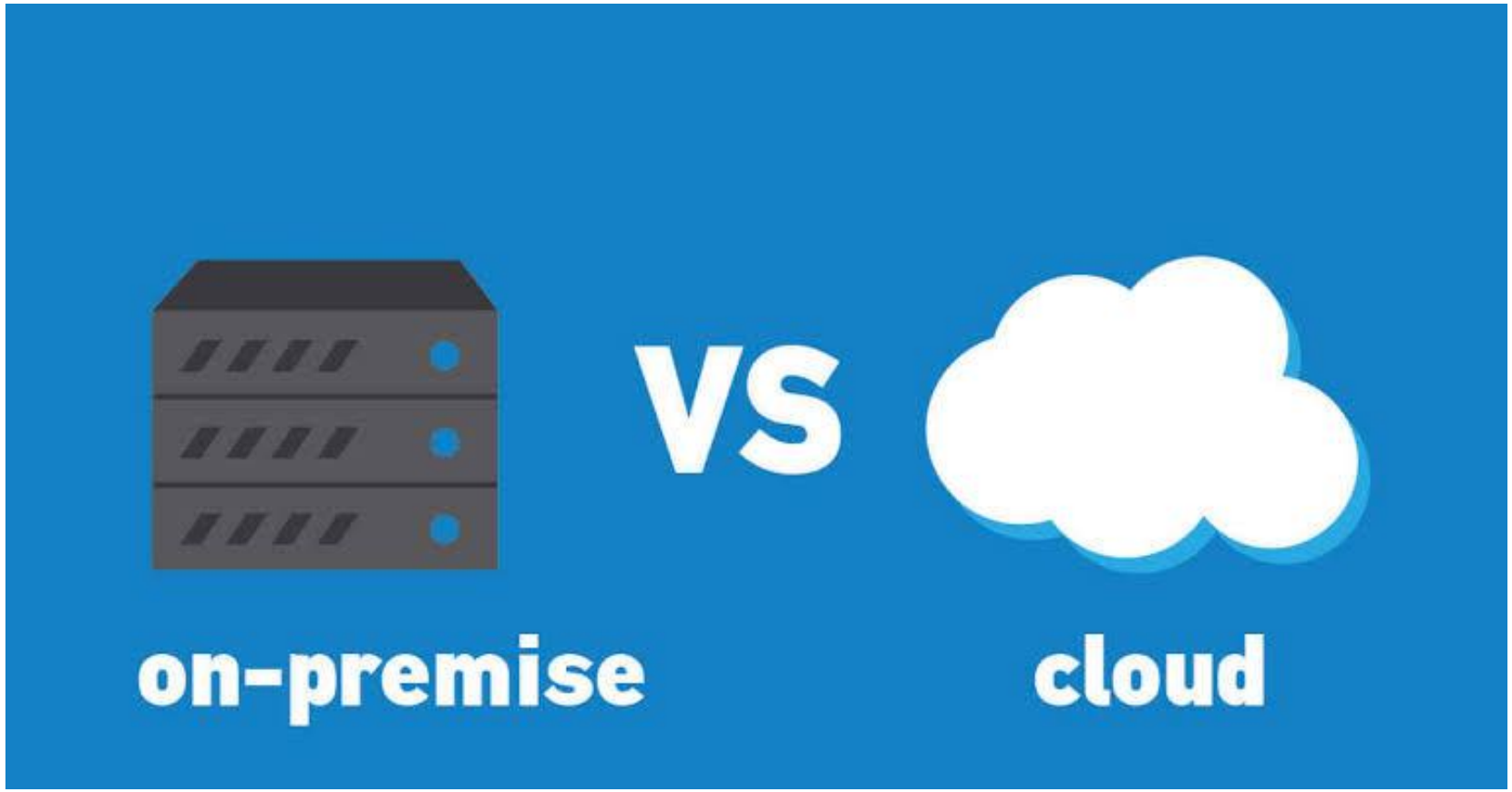
Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware.



Moving Computation to Data



Source: UC San Diego, Big Data


VS
on-premise
cloud

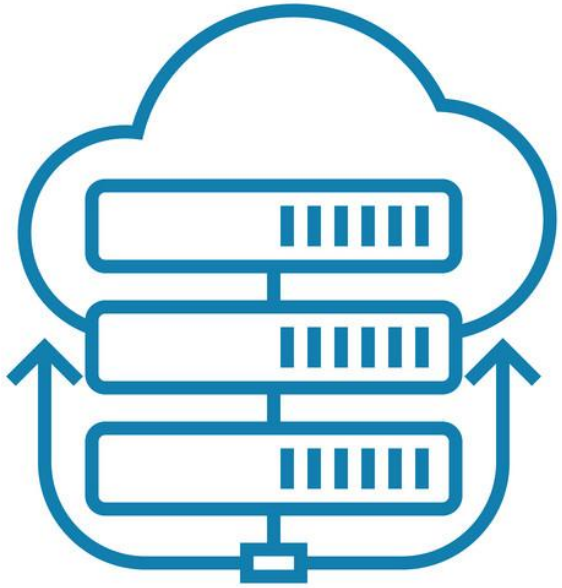
On Premises

- Software and technology
- located within the physical confines of an enterprise
- Often in the company's data centers – as opposed to running remotely on hosted servers or in the cloud.

On Premise

Cloud Computing

- Differs from on-premises software in one critical way.
- Third-party provider hosts all that for you.
- Allows companies to pay on an as-needed basis.
- Allows companies to effectively scale up or down depending on overall usage, user requirements, and the growth of a company.



Data Engineering

What is Data Engineering?

"Data engineering comprises all engineering and operational tasks required to make data available for the end-user, whether for the purpose of analytics, model building or app development etc."

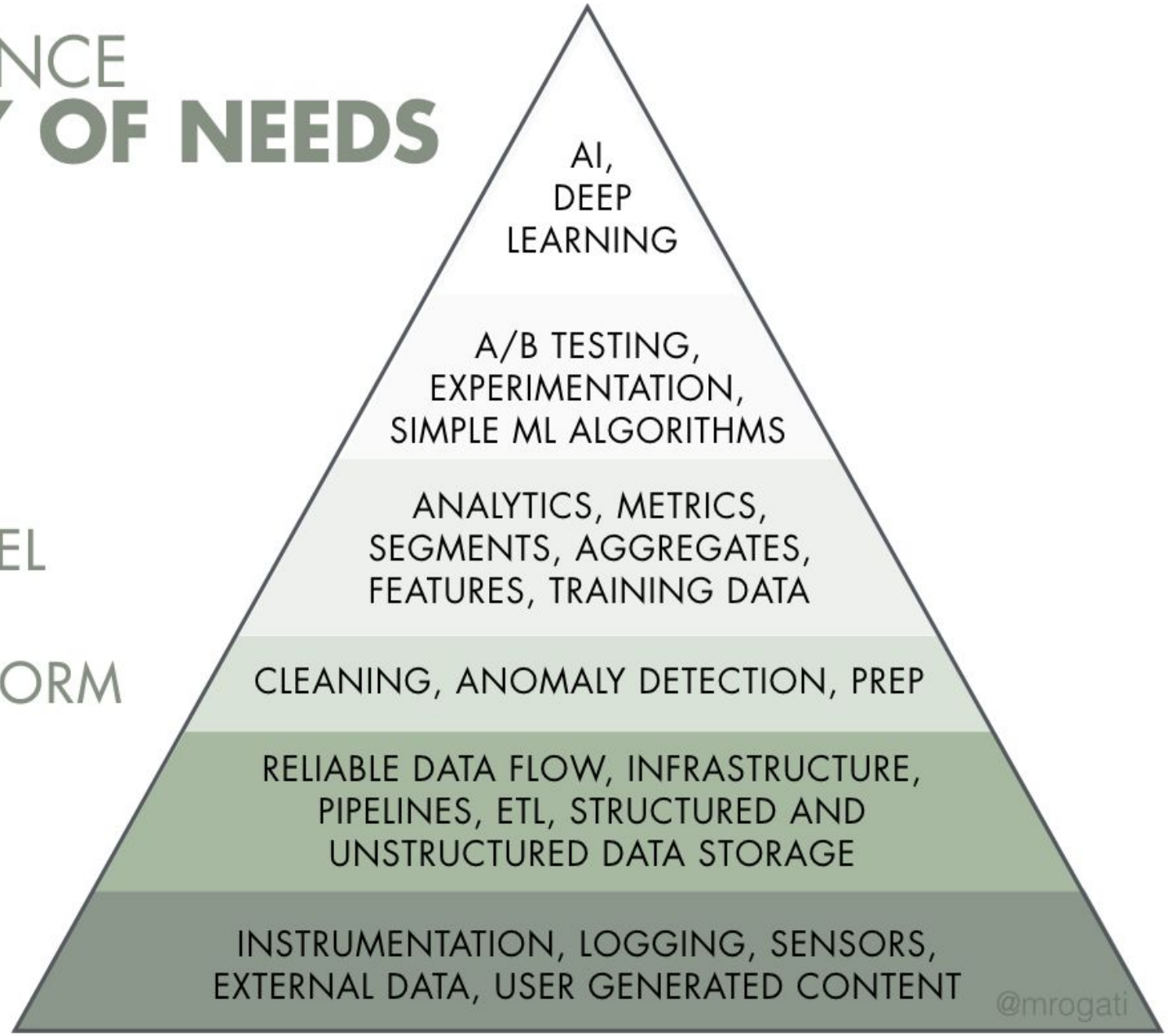
3 step process in layman's terms.

1. Taking raw data.
2. Doing a bunch of work to it.
3. Delivering a clean dataset of database.

What do Data Engineers do?

- Model data
- Build production ready data warehouses and data lakes.
- Tools that process massive amount of data.
- Automate data pipelines.

THE DATA SCIENCE **HIERARCHY OF NEEDS**



LEARN/OPTIMIZE

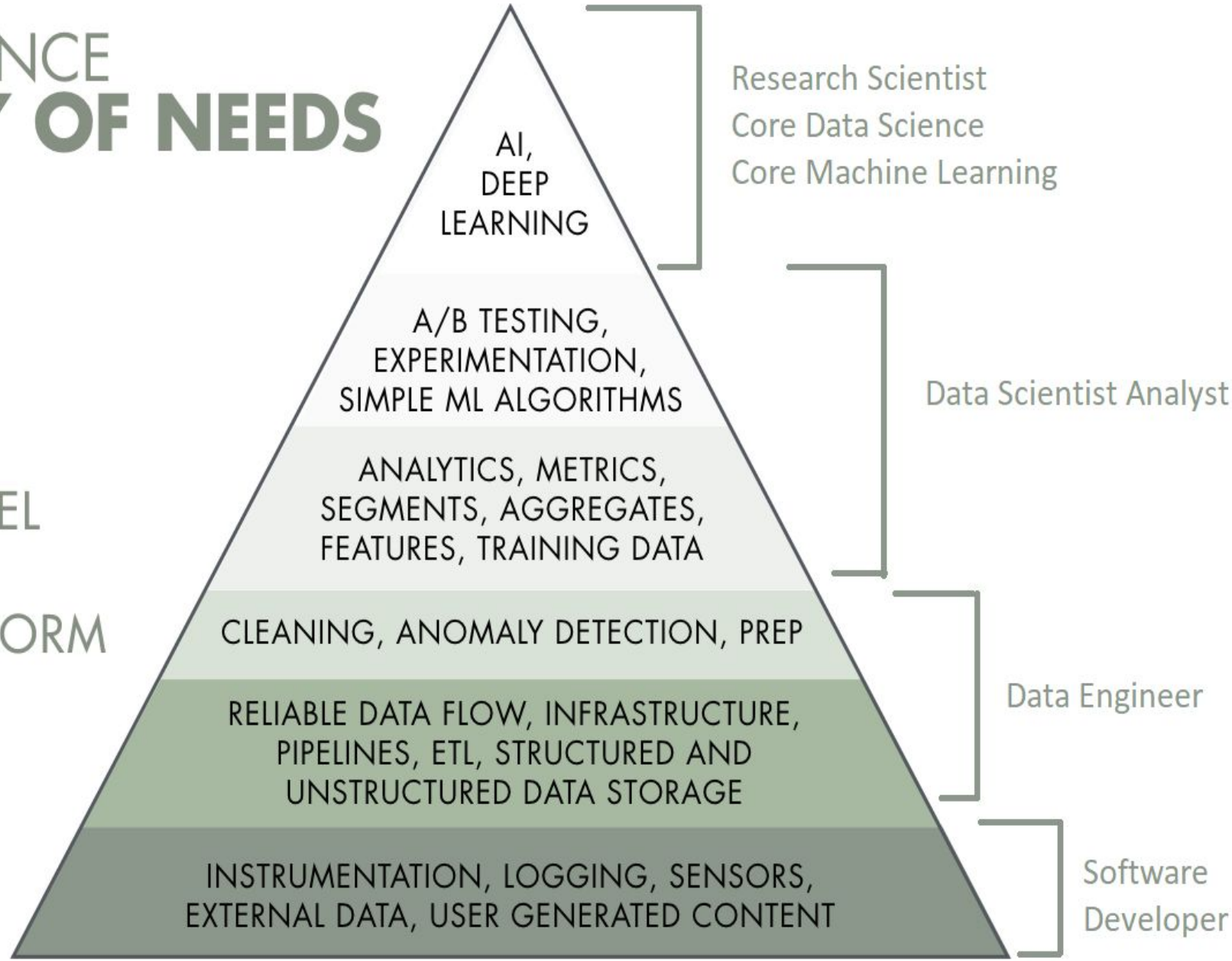
AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

THE DATA SCIENCE **HIERARCHY OF NEEDS**



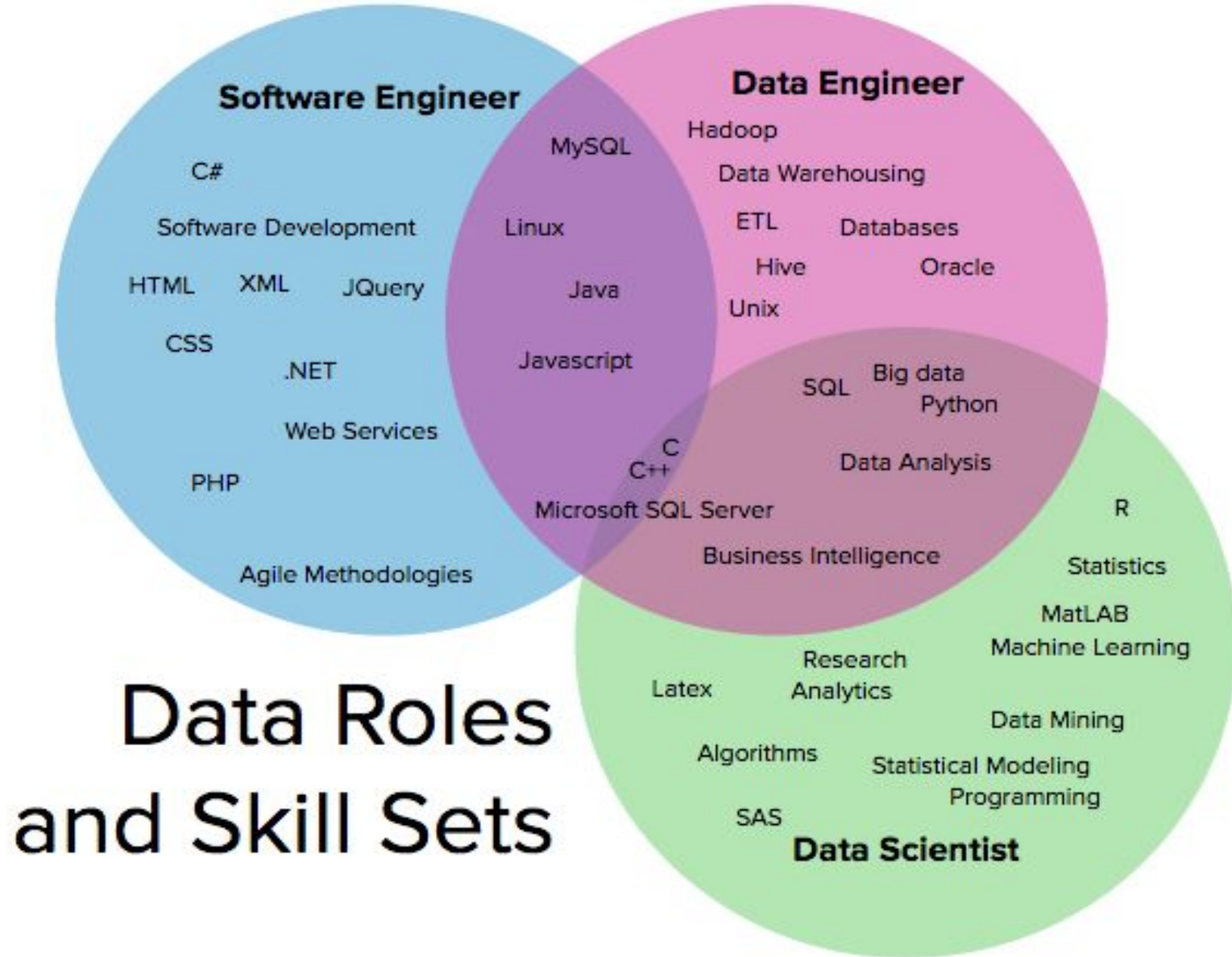
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



Data Roles and Skill Sets

Common data engineering activities.

1. Ingest data from a data source.
2. Build and maintain data warehouse.
3. Create a data pipeline.
4. Create an analytics table for a specific use case.
5. Migrate data to the cloud.
6. Schedule and automate pipelines.
7. Debug data quality issues.
8. Optimize Queries.
9. Design a database.

Questions?