Cloud Computing & Big Data

PARALLEL & SCALABLE MACHINE LEARNING & DEEP LEARNING

Prof. Dr. – Ing. Morris Riedel
Associated Professor
School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
Research Group Leader, Juelich Supercomputing Centre, Forschungszentrum Juelich, Germany

@Morris Riedel @MorrisRiedel @MorrisRiedel

LECTURE 1

Cloud Computing & Big Data Introduction

September 10, 2020
Online Lecture

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

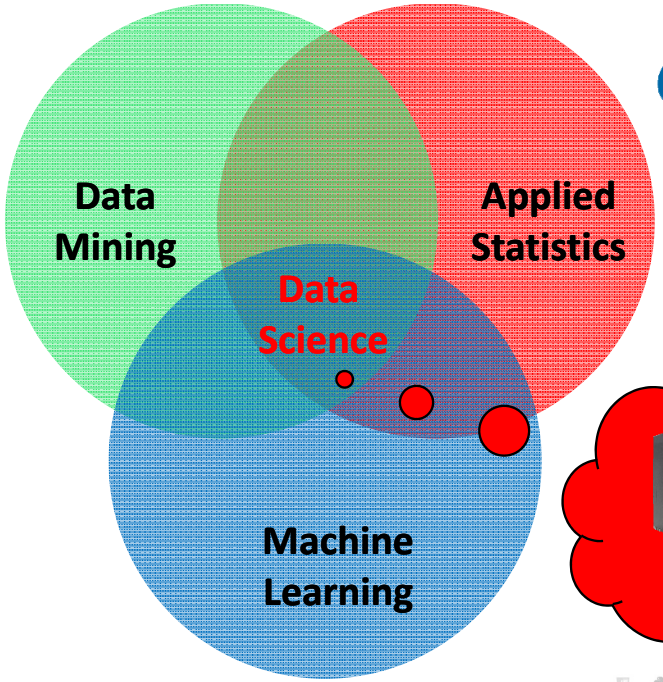
JÜLICH Forschungszentrum
JÜLICH SUPERCOMPUTING CENTRE

DEEP Projects

HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

Review of Practical Lecture 0.1 – Short Introduction to Python & Jupyter

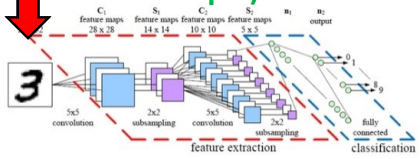
- Data Science applications often use Python

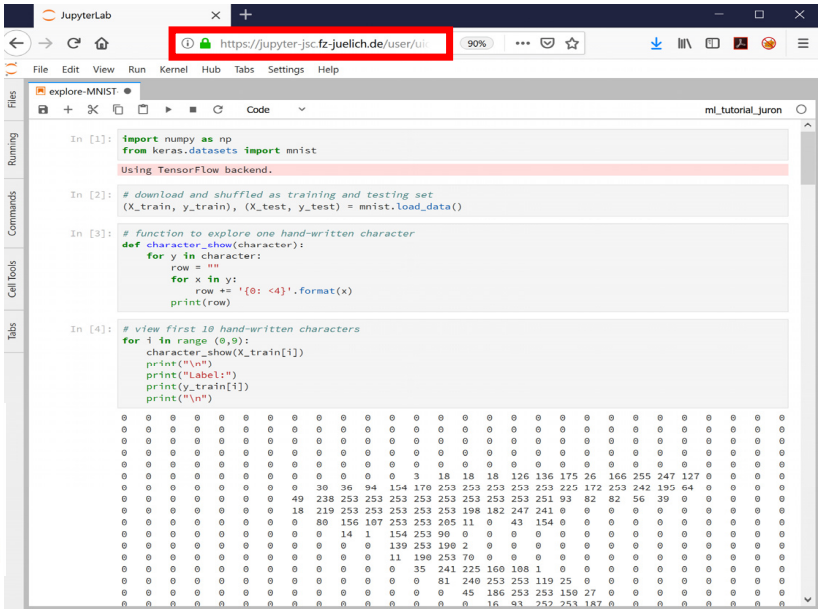


- Jupyter Toolset for Cloud & HPC Access

(remote cloud & HPC access)

(data preprocessing example)





[7] A. Rosebrock

[1] Python [2] Jupyter [3] Amazon Web Services [4] Microsoft Azure [5] Google Cloud [6] Jupyter @ JSC

Outline of the Course

1. Cloud Computing & Big Data Introduction
2. Machine Learning Models in Clouds
3. Apache Spark for Cloud Applications
4. Virtualization & Data Center Design
5. Map-Reduce Computing Paradigm
6. Deep Learning driven by Big Data
7. Deep Learning Applications in Clouds
8. Infrastructure-As-A-Service (IAAS)
9. Platform-As-A-Service (PAAS)
10. Software-As-A-Service (SAAS)
11. Big Data Analytics & Cloud Data Mining
12. Docker & Container Management
13. OpenStack Cloud Operating System
14. Online Social Networking & Graph Databases
15. Big Data Streaming Tools & Applications
16. Epilogue

+ additional practical lectures & Webinars for our hands-on assignments in context

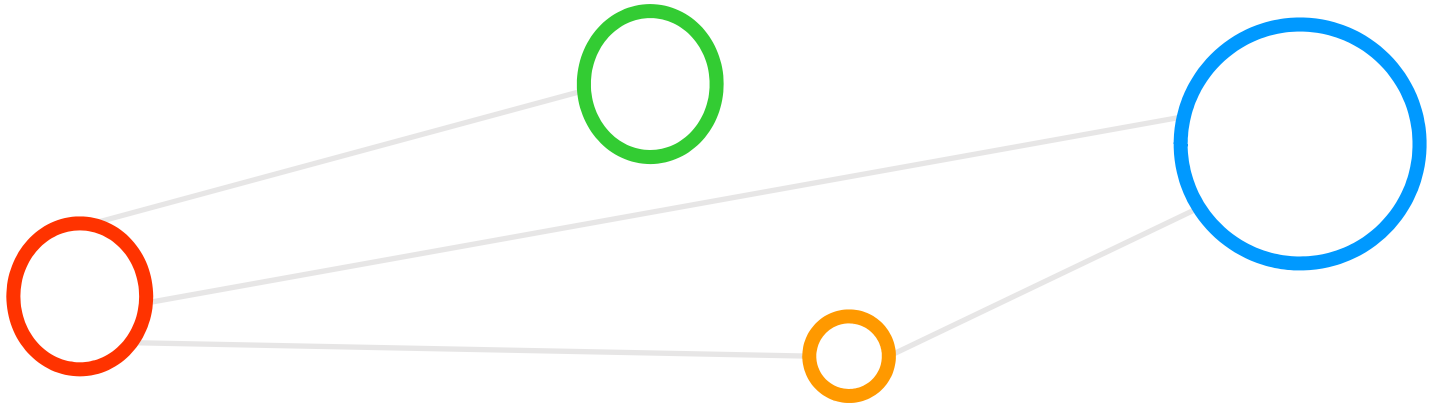
- Practical Topics
- Theoretical / Conceptual Topics

Outline

- Foundations of Cloud Computing
  - Parallel & Distributed Computing Evolution
  - Internet Cloud Systems & Distributed Computing Examples
  - Technology Advances & Parallel Computing
  - Multi-core CPUs & Many-core GPUs Technology Foundations
  - Motivation for *-as-a-Service Approach & Examples

- Scalability driven by Big Data
  - What is Big Data & Key Challenges
  - Evolutions with Limitations using Memory & Disk Storage
  - Examples of Cloud Storages & Cloud Scalability Approach
  - Wide Area Networks & Distributed Programming Models
  - Using Big Data Analytics & Machine/Deep Learning in Clouds

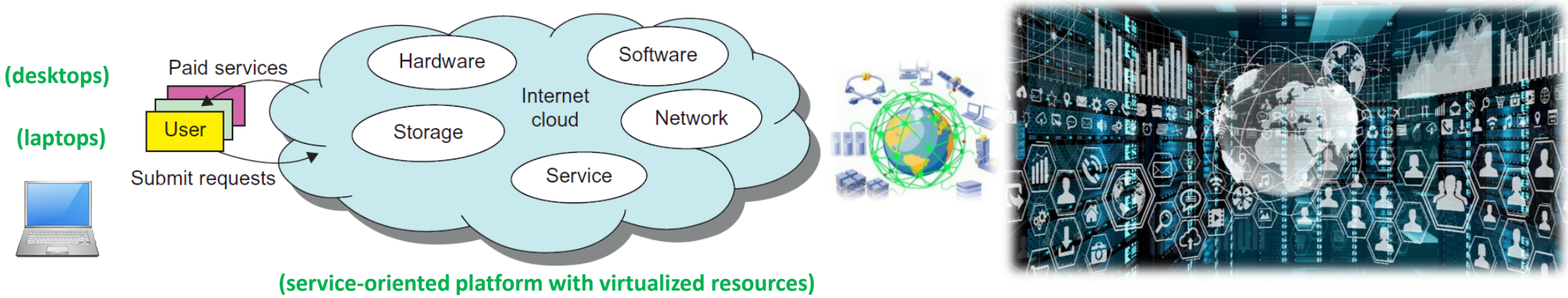
Foundations of Cloud Computing



What is Cloud Computing from 10.000 ft?

- Data Centres
  - Provide virtualized resources to form an Internet cloud
  - Provisioned with hardware, software, storage, network, and services
  - End user pay to run their applications or services after submitting requests

[8] Distributed & Cloud Computing Book



(service-oriented platform with virtualized resources)

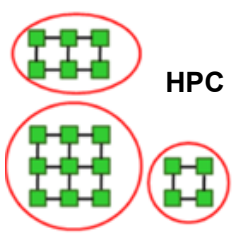
- **Cloud computing moves desktop computing and laptop computing via the Internet to a service-oriented platform using remote large server clusters and massive storages to data centres**
- **Virtualization has enabled the cost-effectiveness and simplicity of cloud computing solutions and enabled services for multiple users**

➢ **Lecture 4 provides more details about the underlying virtualization technology and its relevance for large data centers & clouds today**

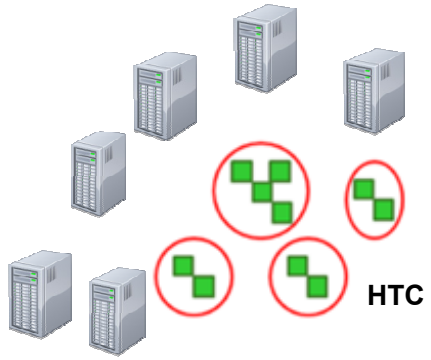
Evolutions towards Cloud Computing

- Many evolutions in parallel and distributed computing
  - Over the past 30 years
  - Driven by applications with variable workloads and large big data sets
- Established computing paradigms
  - High Performance Computing (HPC)
  - High Throughput computing (HTC)
- Increase use of parallel computers
  - Computer clusters, service oriented architectures, computational Grids, peer-to-peer networks, Internet Clouds, Internet of Things, …

[8] Distributed & Cloud Computing Book



HPC

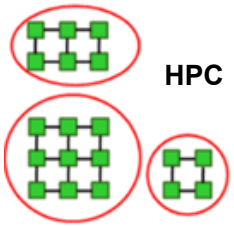


HTC

- **Technology achievements of parallel and distributed computing over 30 years fuels many technologies behind cloud computing today**
- **Instead of using a centralized computer to solve computational problems, a parallel and distributed computing system uses multiple computers to solve large-scale problems over the Internet**
- **Cloud Computing is based on computing paradigms known as High Performance Computing (HPC) and High Throughput Computing (HTC)**

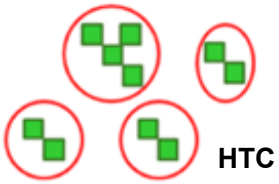
High Performance Computing (HPC) vs. High Throughput Computing (HTC)

- **High Performance Computing (HPC) is based on computing resources that enable the efficient use of parallel computing techniques through specific support with dedicated hardware such as high performance cpu/core interconnections.**

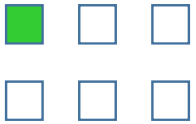


(network connection
very important & costly)

- **High Throughput Computing (HTC) is based on commonly available computing resources such as commodity PCs and small clusters that enable the execution of 'farming jobs' without providing a high performance interconnection between the cpu/cores.**

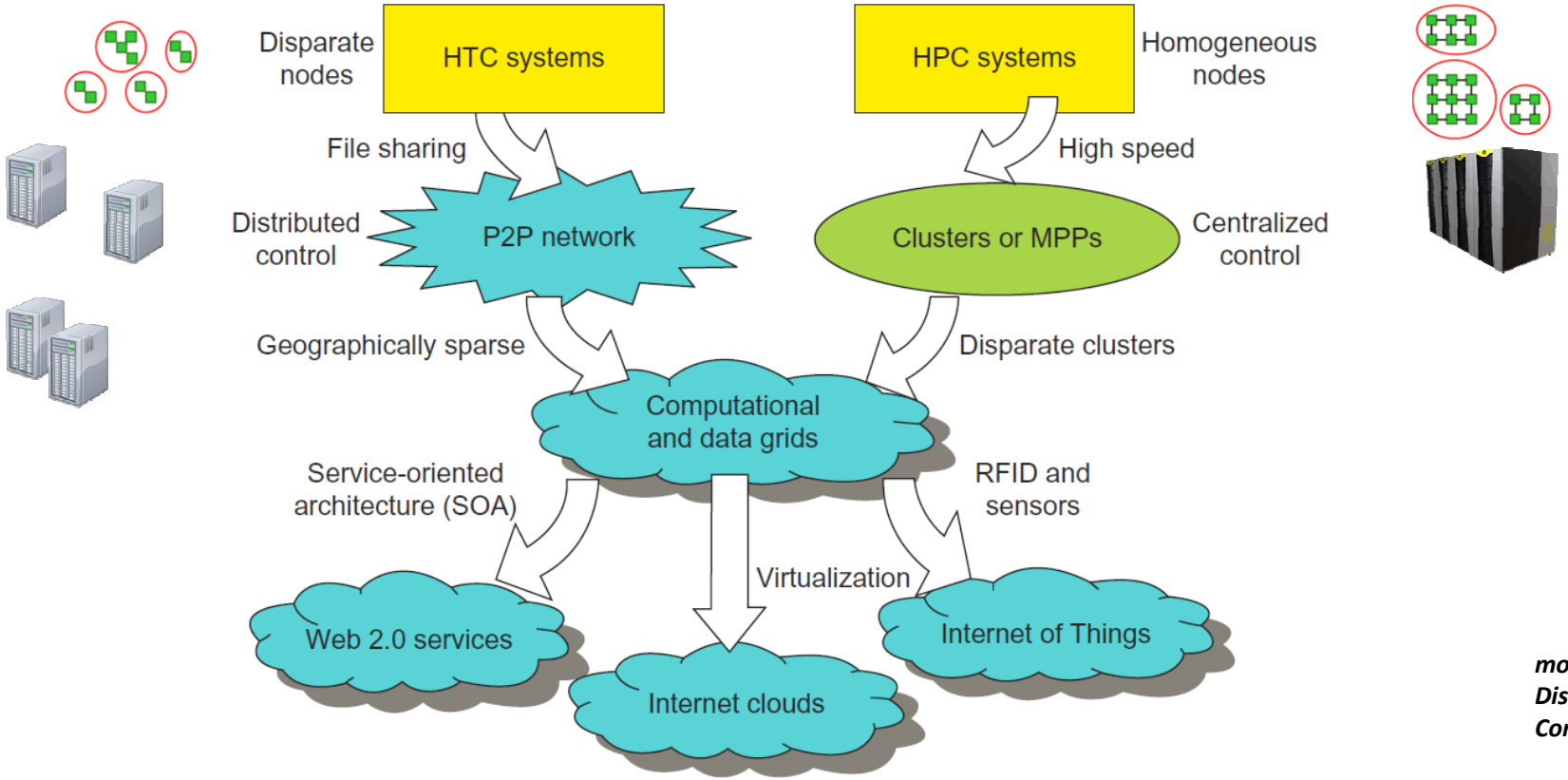


(network connection
less important)



➢ **This course is often using Cloud resources while the general techniques and algorithms can also work on pure HPC resources**

Evolution over time



modified from [8] Distributed & Cloud Computing Book

➢ **Lecture 4 provides more details about the computing paradigms HPC and HTC and their relevance to cloud computing systems today**

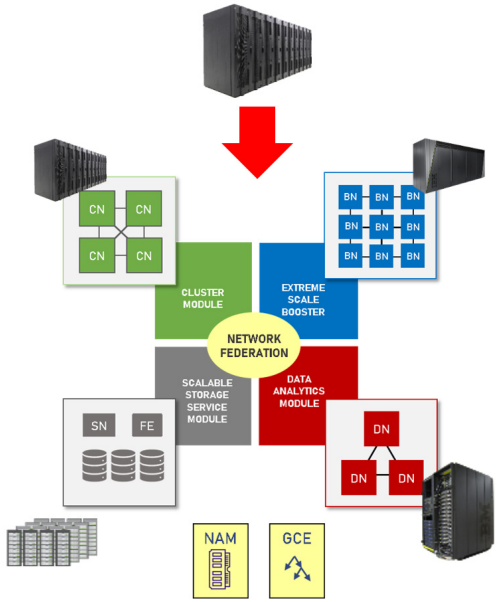
Internet Cloud Systems – Examples from Every Day Life

- Selected Cloud Systems (aka 'Clouds') known today
  - Google Cloud → massive computing/storage/applications
  - Amazon Web Service → massive computing/storage/services
  - Microsoft Azure → massive computing/storage/toolsets
  - Facebook → online social networking & advertisement
  - SalesForce.com → customer relationship management
  - Rackspace → managed cloud provider & hosting
  - IBM Bluemix → cloud platform
  - Enomaly → elastic computing cloud
  - European Open Science Cloud → computing & storage services for research
  - Uber Cloud → specialized computing & storage services for engineers

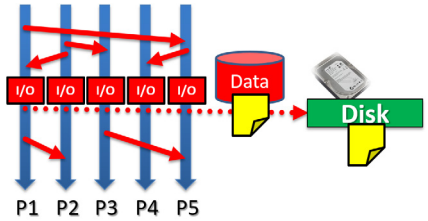
- **Cloud systems play an increasingly important role in upgrading traditional Web services and Internet applications to modern scalable online services**
- **Large-scale Internet applications driven by cloud computing have significantly enhanced the quality of life in society today**
- **In one form or another are cloud computing applications requiring large amounts of computing capability or storage capacity**

Terminologies: Centralized Computing & Parallel Computing

- Centralized Computing
  - All computer resources are centralized in one physical system
  - All resources (e.g. processors, memory, storage, etc.) fully shared and tightly coupled within one integrated operating system
  - Many data centers & supercomputers are centralized systems
  - Trend moves towards the use of modular supercomputing
- Parallel Computing
  - All processors can be tightly coupled using shared memory
  - All processors can be loosely coupled using distributed memory
  - Interprocessor communication via shared memory or message passing
  - Computer systems capable of parallel computing is a parallel computer
  - Programs running in a parallel computer are called parallel programs
  - Process of writing parallel programs is referred to as parallel programming



[9] DEEP Projects Web Page



➢ **The complementary HPC course teaches parallel computing with insights into shared memory and distributed memory programming**

Terminologies: Distributed Computing & Cloud Computing

- Distributed Computing ('parallel computing across computers')
  - Distributed systems consist of multiple autonomous computers (each having its own memory and stroage; communication via network)
  - Field of computer science/engineering that studies distributed systems
  - Information exchange in a distributed system is done via message passing
  - Program that runs in a distributed system is a distributed program
  - Process of writing distributed programs is known as distributed programming

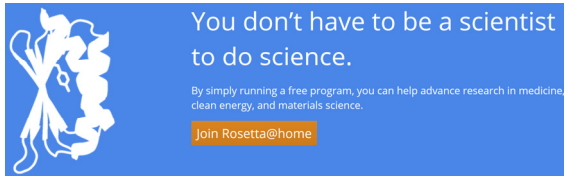
[8] Distributed & Cloud Computing Book

links to Internet of Things (IoT)

- **An Internet Cloud of resources can be either a centralized or a distributed computing system**
- **Clouds apply parallel or distributed computing or a combination of both**
- **Clouds are using physical or virtualized resources over large centralized/distributed data centers**

Distributed Computing – Early Cloud-like BOINC Example for Research

- Tool for distributed computing
  - BOINC is one middleware tool in distributed computing, e.g. Rosetta@Home
  - Provides functionalities to work with compute and big data sets
  - Different geographically distributed nodes in a distributed system consist of a large number of heterogenous nodes used for computing
  - Architecture consists of a large number of rather ordinary desktop computers
- Unique selling proposition
  - BOINC implements concept of using unused resources
  - Use ‘free’ unused computing power or storage during the night or during longer inactive usage periods

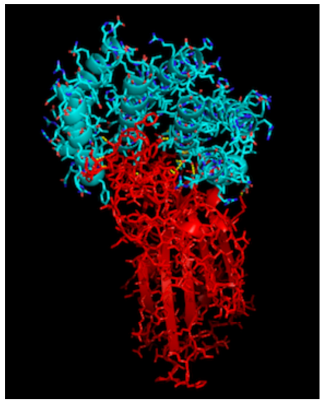


[8] Distributed & Cloud Computing Book

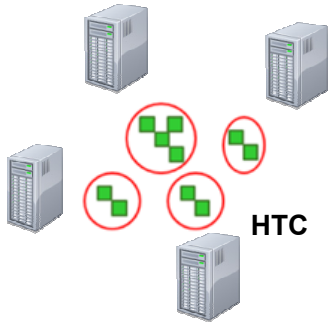
[10] BOINC middleware tool

- **Berkely Open Infrastructure for Network Computing (BOINC) is distributed computing framework**
- **BOINC implements CPU scavenging that means using unused resources in distributed computing**

Help in the fight against COVID-19!



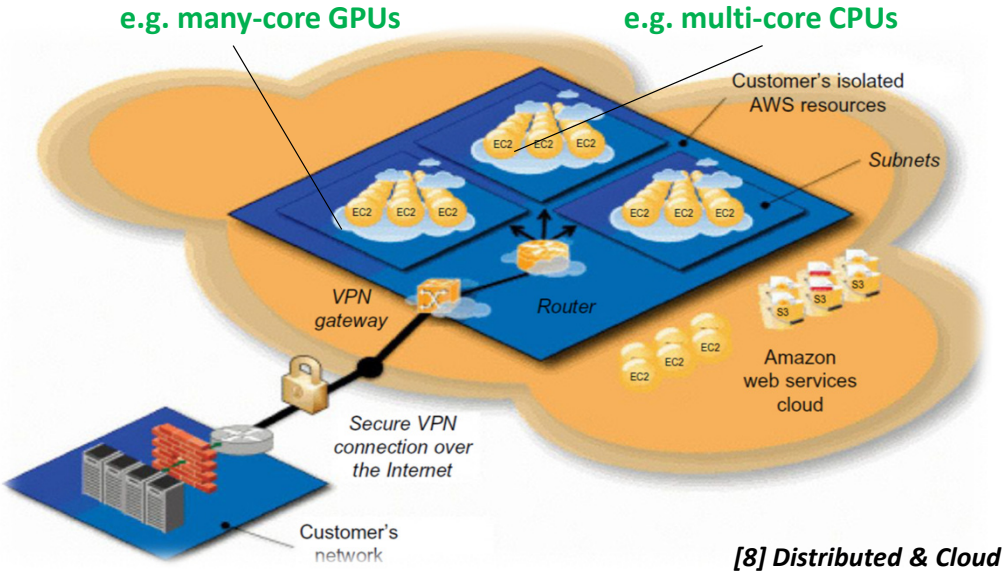
With the recent COVID-19 outbreak, R@h has been used to predict the structure of proteins important to the disease as well as to produce new, stable mini-proteins to be used as potential therapeutics and diagnostics, like the one displayed above which is bound to part of the SARS-CoV-2 spike protein.



Amazon Web Services – Modern Infrastructure-as-a-Service (IAAS) Example

- Amazon EC2 provides an elastic compute cloud (EC2)
  - Elastic load balancing services and so-called auto-scaling
  - E.g. great during peak times in business (e.g. x-mas shopping, etc.)
  - Ensures that a sufficient number of EC2 instances are provisioned to meet expected performance
  - E.g. New York Times use it to quickly retrieve pictorial information from millions of articles
- Amazon Web Services (AWS)
  - Offers infrastructure used for Amazon shopping also for computing customers
  - Ideal situation for Amazon
  - Offers high number of resources & services

[3] Amazon Web Services

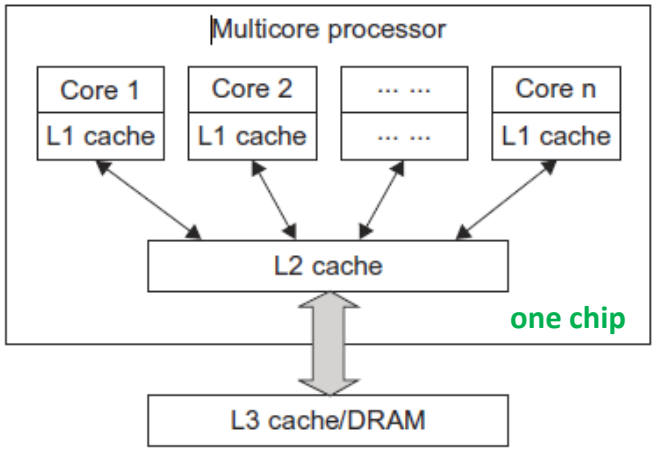


[8] Distributed & Cloud Computing Book

➢ **Lecture 8 provides more details about Amazon Web Services and its Infrastructure-as-a-Service (IAAS) models & various cloud services**

Cloud & HPC Building Blocks – Multi-Core CPUs

- Significant advances in CPU (or microprocessor chips)
  - Multi-core architecture with dual, quad, six, or n processing cores
  - Processing cores are all on one chip
- Multi-core CPU chip architecture
  - Hierarchy of caches (on/off chip)
  - L1 cache is private to each core; on-chip
  - L2 cache is shared; on-chip
  - L3 cache or Dynamic random access memory (DRAM); off-chip

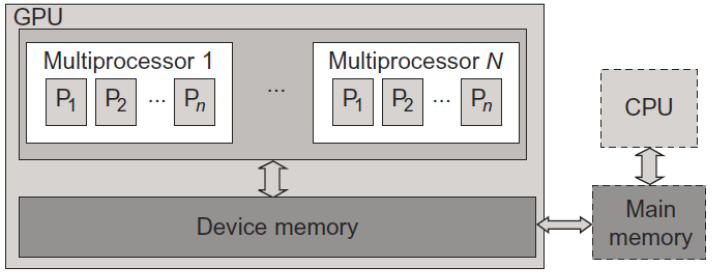


[8] Distributed & Cloud Computing Book

- **Clock-rate for single processors increased from 10 MHz (Intel 286) to 4 GHz (Pentium 4) in 30 years**
- **Clock rate increase with higher 5 GHz unfortunately reached a limit due to power limitations / heat**
- **Multi-core CPU chips have quad, six, or n processing cores on one chip and use cache hierarchies**

Cloud & HPC Building Blocks – Many-core GPGPUs

- Use of very many simple cores
  - High throughput computing-oriented architecture
  - Use massive parallelism by executing a lot of concurrent threads slowly
  - Handle an ever increasing amount of multiple instruction threads
  - CPUs instead typically execute a single long thread as fast as possible



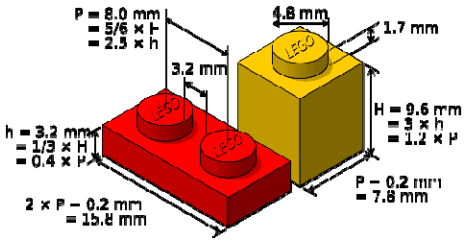
[8] Distributed & Cloud Computing Book

- Many-core GPUs are used in large clusters and within massively parallel supercomputers today
  - Named General-Purpose Computing on GPUs (GPGPU)
  - Different programming models emerge

- **Graphics Processing Unit (GPU) is great for data parallelism and task parallelism**
- **Compared to multi-core CPUs, GPUs consist of a many-core architecture with hundreds to even thousands of very simple cores executing threads rather slowly**

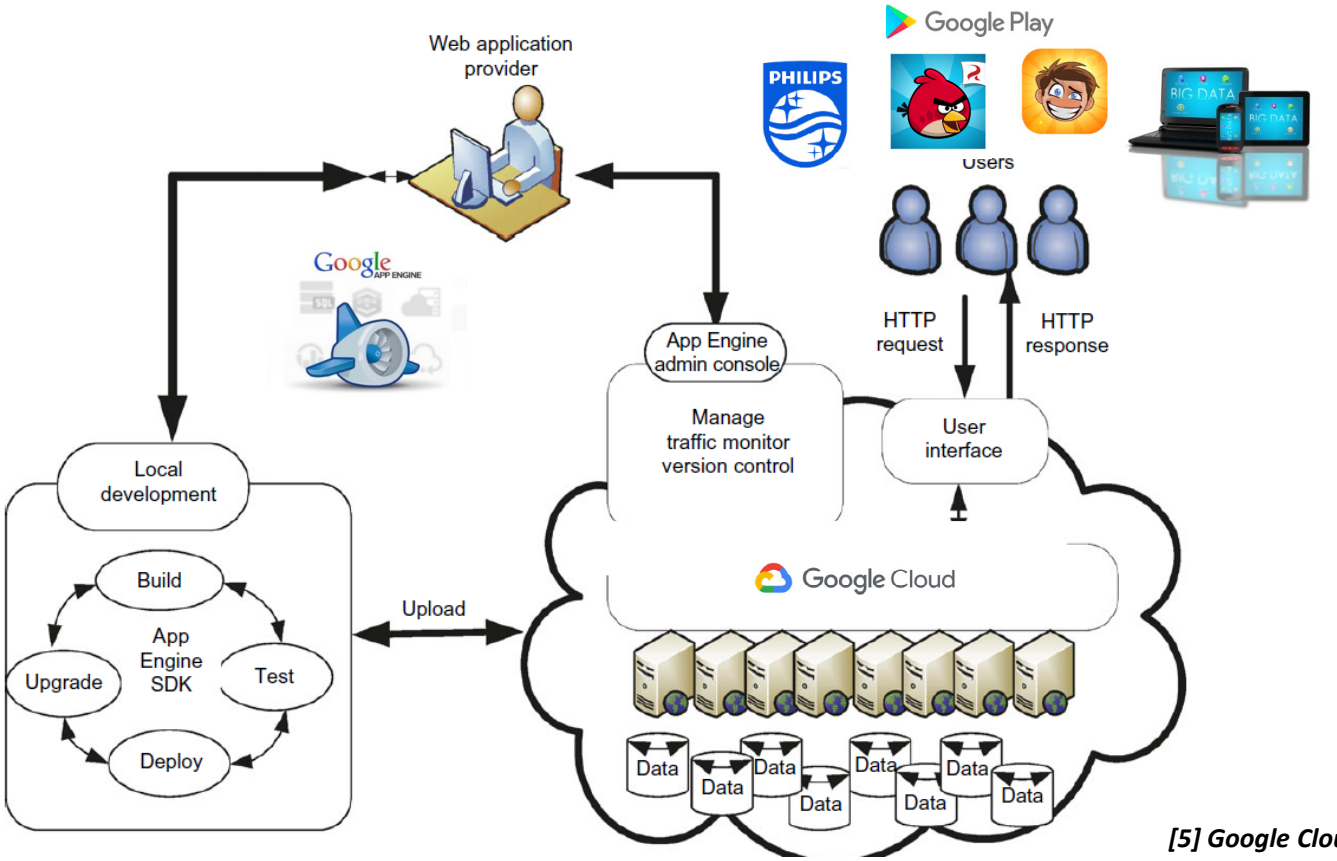
Google Cloud – Modern Platform-as-a-Service (PAAS) Example

- Simple idea
  - Abstract from underlying computing and storage infrastructure by using a platform, e.g. Google Cloud
  - Platform provides easy-to-use ‘lego bricks‘ (aka ‘service bricks‘) to create online services & Apps



[12] Lego Bricks

Modified from [8] Distributed & Cloud Computing Book



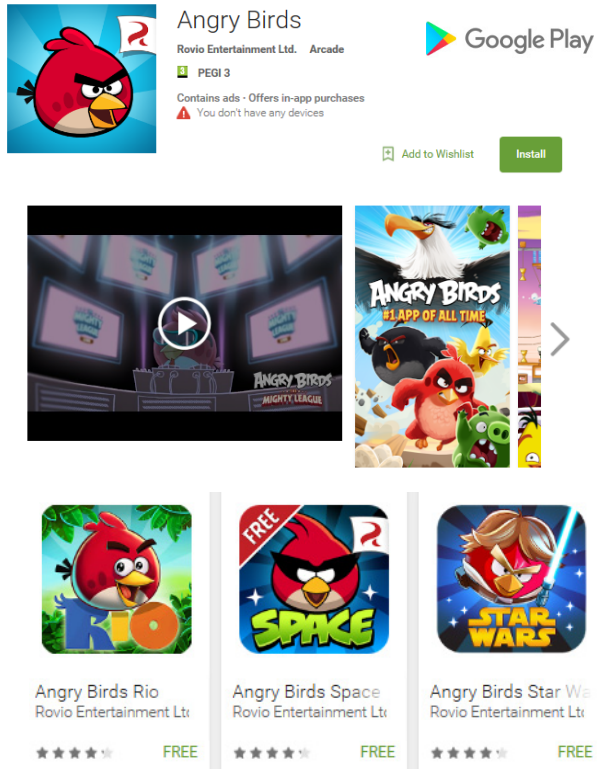
[5] Google Cloud

➢ **Lecture 9 provides more details about Google Cloud services and its Platform-as-a-Service (PAAS) models & various cloud services**

Rovio Games Example – Angry Birds Game App

- Business Case
  - Angry Birds destroy different pigs
  - Each bird has unique talents
  - Revenue: In-app purchases & ads
- Challenges
  - More than 140 million downloads & users
  - Robust capabilities and scalability to deliver a superior user experience for various apps
- Approach
  - Own infrastructure would be to costly
  - Services that are required already implemented in Google Cloud
  - Rovio Web games tend to be popular immediately (no scaling over time)
  - Use Google Cloud Platform: GAE, Memcache API, Google Cloud Datastore

ROVIO



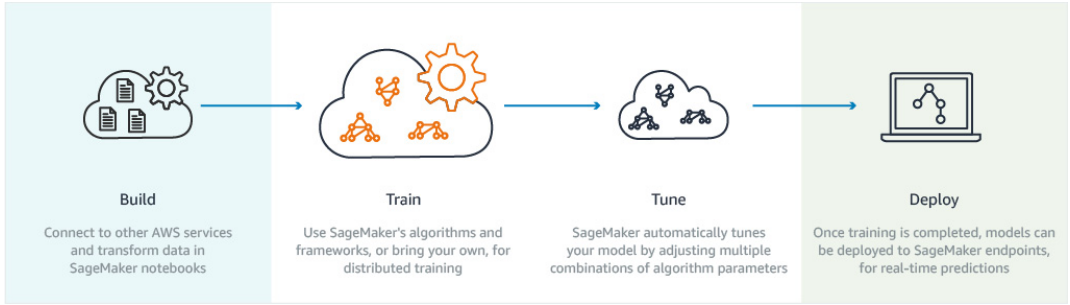
[13] Google Play Angry Birds

[14] Rovio Case Study

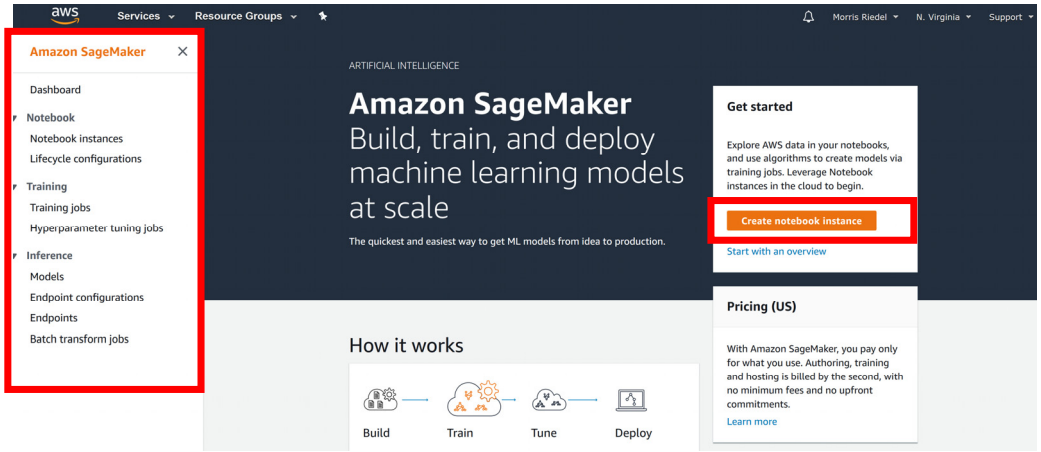
➢ **Lecture 9 provides more details about Google Cloud services and its Platform-as-a-Service (PAAS) models & various cloud services**

Amazon Web Services – Modern Software-as-a-Service (SAAS) Example

- AWS Cloud – Amazon Sagemaker
  - Fully managed service that enables quick & easy machine & deep learning applications
  - Avoids time-consuming manual installation of many required software frameworks
  - Builds on-top of various IAAS & PAAS services



(SAAS solutions often abstracts away completely underlying resources)



- **AWS Amazon Sagemaker is a SAAS oriented service that provides fully managed instances running Jupyter notebooks that include examples training & tuning various machine and deep learning models**
- **Offers Amazon SageMaker Studioas a fully integrated development environment (IDE) for machine learning in the AWS cloud**
- **SAAS services are usually not free and often require a subscription**

[2] Jupyter Web page

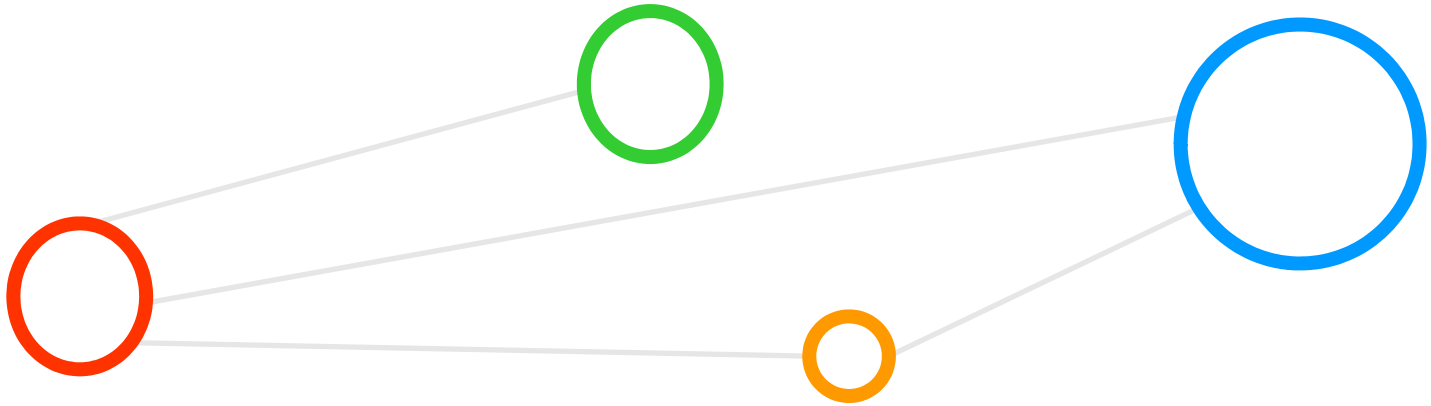
[15] AWS – Amazon Sagemaker

➢ **Lecture 10 provides more details about AWS Cloud services and its Software-as-a-Service (SAAS) models & other SAAS cloud services**

[Video] Cloud Computing Explained

[11] YouTube video, The three ways to Cloud compute

Scalability driven by Big Data



What is Big Data?

- Buzzword in science and engineering – what does this mean?
  - When does ‘big data’ start – with hundreds of MB / GB / TB / PB … EB?
  - Exact definition (in terms of volume) of big data is hard to find…
  - We have to look on concrete examples to find answers in Cloud context
  - Initially referred to VVV (Volume, Velocity, Variety)
  - Being constantly extended to n ‘Big Data Vs‘ (Veracity, Validity, …)

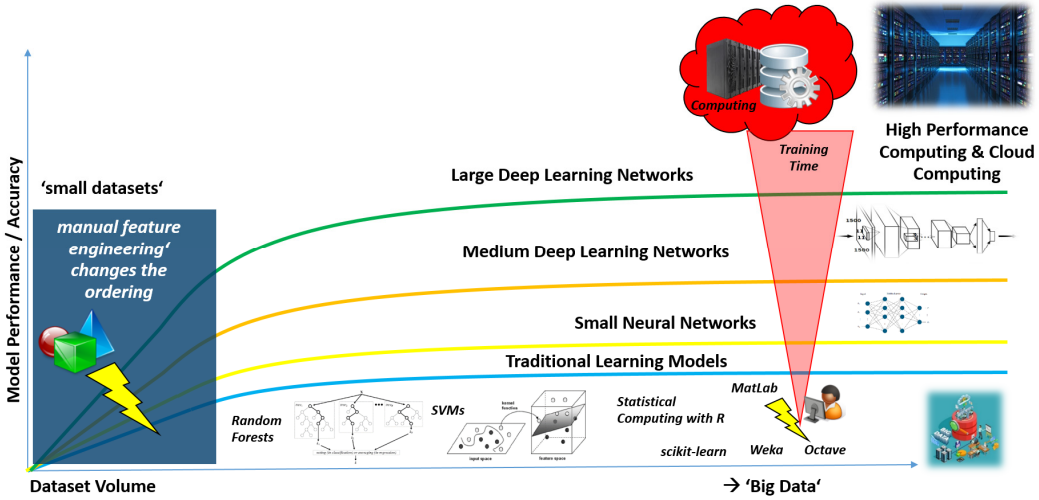
- Selected attempts of definitions for ‘Big Data’ :

> **‘Big Data’ is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications**

[17] Wikipedia ‘Big Data’ Online

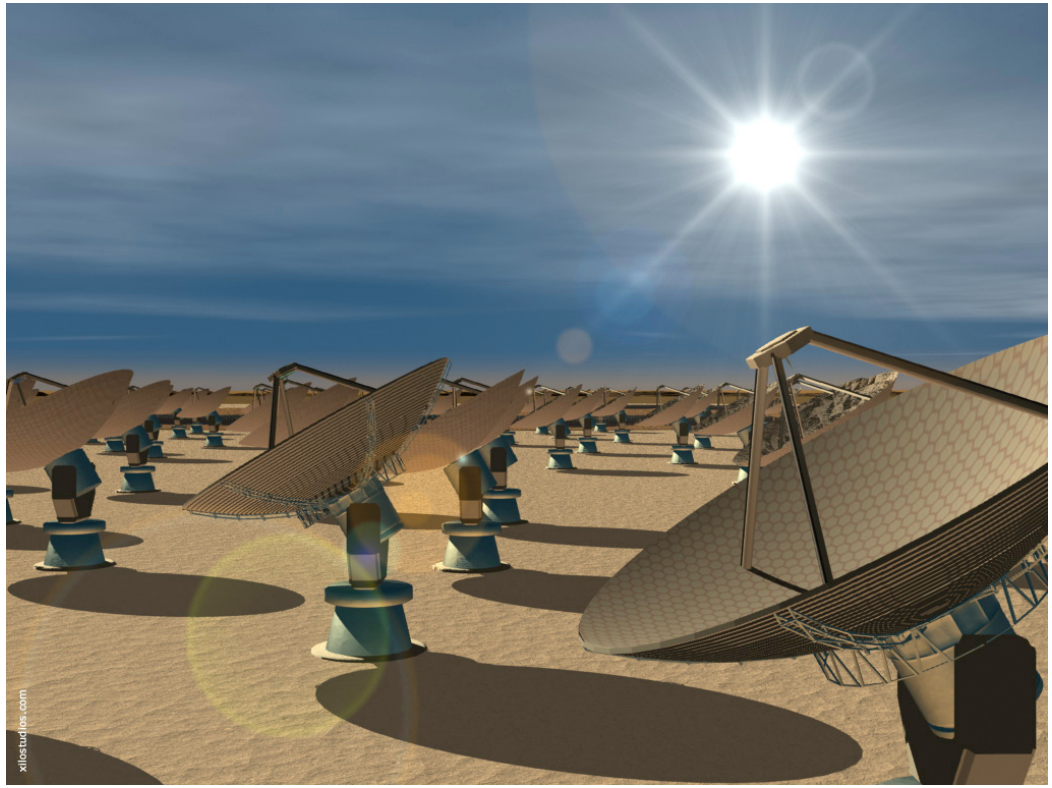
> **‘Big Data’is data that becomes large enough that it cannot be processed using conventional methods.’**

[18] O’Reilly Radar Team, ‘Big Data Now: Current Perspectives from O'Reilly Radar’



[16] www.big-data.tips

Search for Concrete ‘Big Data’– Examples & Challenges in Science & Engineering

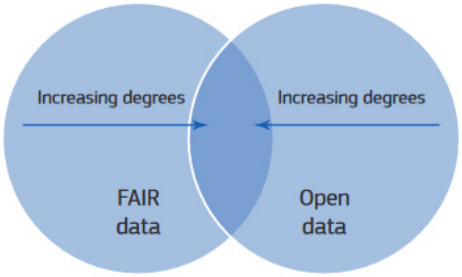


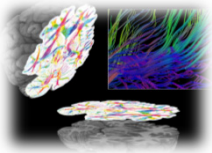
- In science environments the term ‘Big Data’ is often related to one concrete scientific experiment: e.g. square kilometre array → 1 PB / 20 seconds

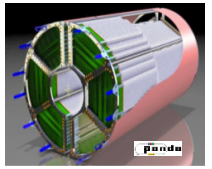
- In commercial environments Big Data is all about Volume, Variety, Velocity, V…, but concrete ‘sizes’ are rarely given

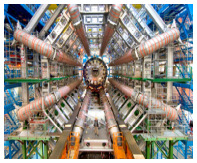
- ‘Big data’ and all data should be FAIR: (F)indable; be (A)ccessible; be (I)nteroperable; and be (R)eusable

[21] EC Expert Group on FAIR data final report

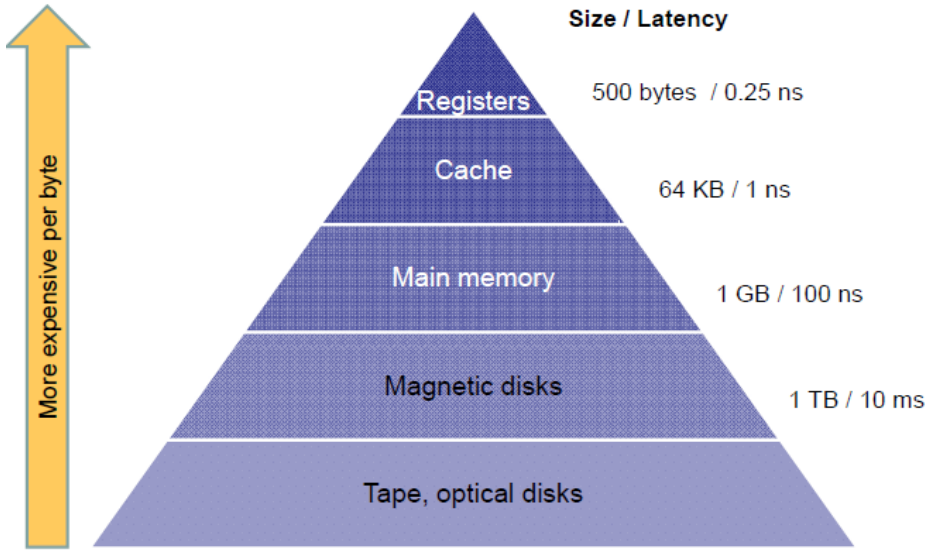








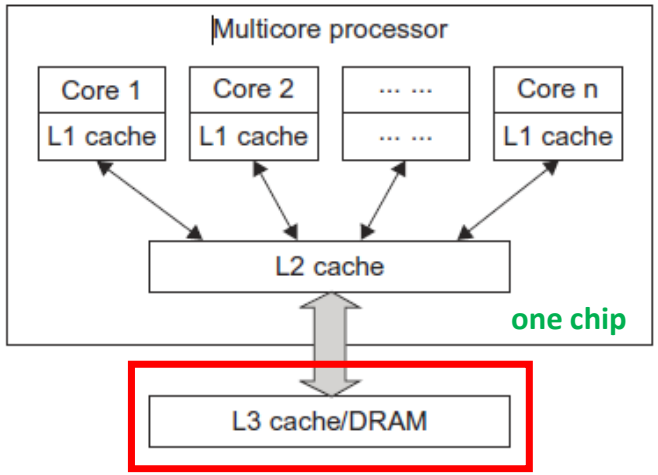
Storage Devices as Storage Hierarchy & Terminologies



- **Data can be stored in various different storage devices and technologies that all have different capacities (i.e., size) and different speeds (i.e., latency)**
- **Clouds use a wide variety of storage devices that can be categorized as a storage hierarchy**
- **The storage hierarchy gets more expensive per byte in each layers that are from bottom to up defined as: (1) tape & optical disks; (2) magnetic disks; (3) main memory; (4) cache; (5) registers**
- **Fast memory chip capacities are measured in Kilobyte (KB), Megabyte (MB), Gigabyte (GB)**
- **Slower disk storage capacities are measured in Terabyte (TB), Petabyte (PB), and Exabyte (EB)**

Dynamic Random Access Memory (DRAM) & Memory Wall

- Trends
  - Far in the past: computing expensive and limited
  - Today: computing is getting cheaper and widely available
  - Challenge: Fast storage capacity close to CPUs to provide data
- Fast memory used for data or program code
  - Computer processor needs memory to function
  - Volatile memory that loses data when power is removed
- Towards 'memory wall problem'
  - Faster CPU processor speeds
  - Larger memory capacity but access time not much better
  - Result is a wide gap between them

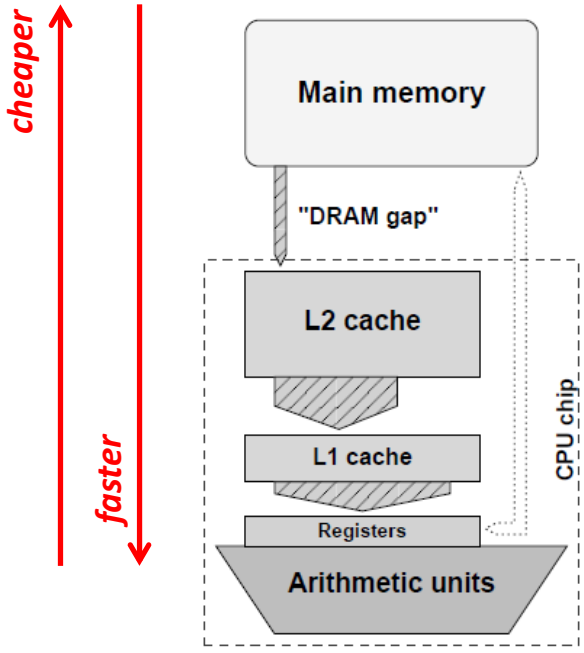


[8] Distributed & Cloud Computing Book

- **Dynamic Random Access Memory (DRAM)**
- **Huge growth of DRAM chip capacity in the last 30 years; 16 KB (1976) to 64 GB (2011) and much faster today**
- **DRAM access time did not improve much contributing to 'memory wall problem' with increasingly better & cheaper CPUs**

Increasing Key Challenge with ‘Big Data‘ Today: Fast Data Access for Processing

too slow



- Processor core elements & Memory
  - Compute: floating points or integers
  - Arithmetic units (compute operations)
  - Registers (feed those units with operands)
- ‘Data access‘ for application/levels
  - Registers: ‘accessed w/o any delay‘
  - L1D = Level 1 Cache – Data (fastest, normal)
  - L2 = Level 2 Cache (fast, often)
  - L3 = Level 3 Cache (still fast, less often)
  - Main memory, DRAM, slow, but larger in size than caches
  - Too slow: storage like harddisk, solid state disks, tapes, etc.

- **The DRAM gap is the large discrepancy between main memory and cache bandwidths**

[17] Introduction to High Performance Computing for Scientists and Engineers

Storage Technologies

- Slower disk storage used for data used by program code
  - Magnetic disks in the past
  - Rapid growth of Solid State Disks (SDDs)
  - But SDDs are still quite expensive
  - Flash memory is a (solid-state) non-volatile storage for persistent data

array of disks used within cloud data centres

- Huge disk storage capacity growth
  - Capacity increase last 30 years and continues for disk arrays in the future at a much faster pace
  - Cloud data centres work with 'big data' in PBs, EBs, and above
  - Derived usage models by application users: 'never delete data, it is better to keep it and costs are ok'

- **Huge growth of disk capacity in the last 30 years; 260 MB (1981), 250 GB (2004), 3 TB (2011), now even with massive more speeds**
- **Rapid growth of flash memory and solid state disks (SDDs) impacts cloud computing storages and pricing models**

➢ **Lecture 4 provides more details about storage technologies relevant for large data center designs that operate multi-user clouds today**

European Open Science Cloud (EOSC) – Storage and 'Big Data' Service Examples

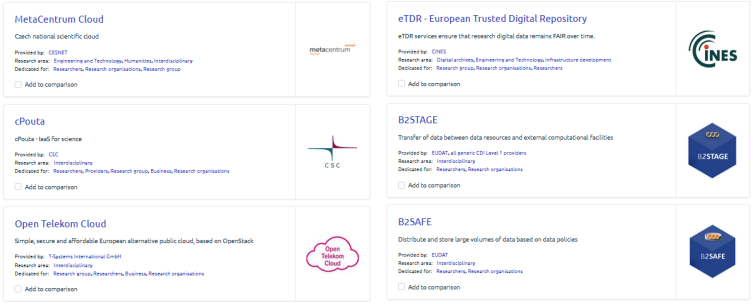
- European Open Science Cloud (EOSC)
  - EU EOSC-Nordic project in Iceland: provisioning of a couple of data services for selected application communities in Iceland
  - Many different service providers for 'big data'

EUROPEAN OPEN SCIENCE CLOUD

[19] EOSC Web page

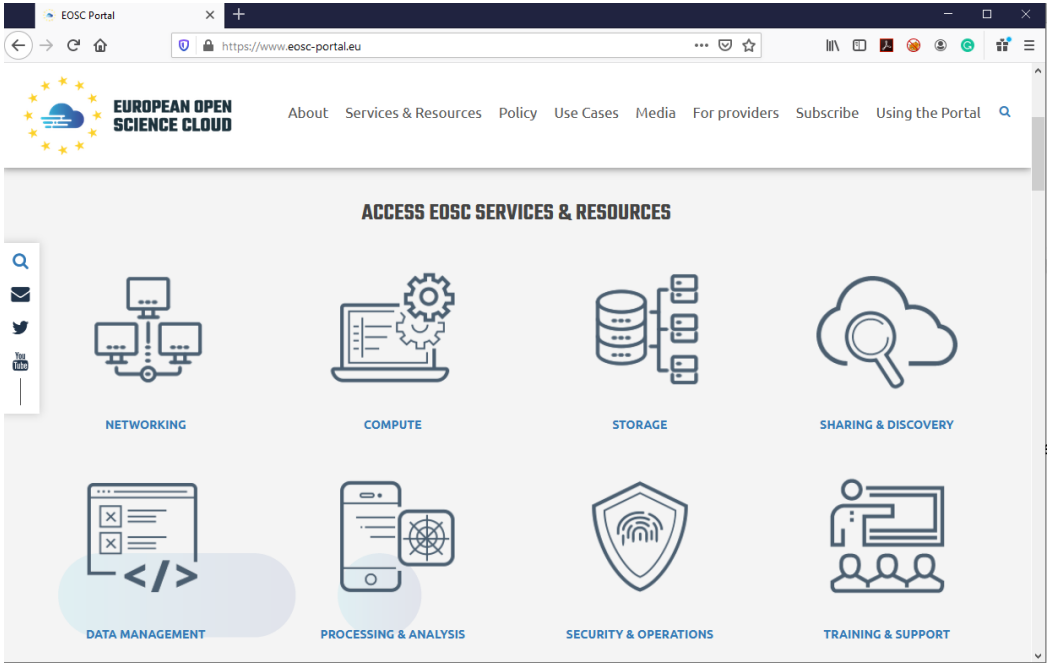
EOSC NORDIC

[20] EOSC-Nordic



[21] EC Expert Group on FAIR data final report

modified for [22] Go-Fair Initiative



- **The European Open Science Cloud (EOSC) provides services and tools for large-scale datasets (aka 'big data') for European researchers**
- **Increasing momentum for EOSC that is centrally based on making data (F)indable, (A)ccessible, (I)nteroperable, and (R)eusable (FAIR) like the Internet of FAIR Data and Services (IFDS) that provides services where machines and people can find and reuse each other's research objects under optimal and well-defined conditions**

➢ **Lecture 10 provides more details about the EOSC service landscape offering Software-as-a-Service (SAAS) models for EU researchers**

Amazon Web Services – Broadly used Storage-as-a-Service (S3) Example

- S3 is 'storage as a service' with a Web messaging interface
  - Using API with Representational State Transfer (REST)
  - Using API with Simple Object Access Protocol (SOAP)
- Remote object storage
  - Data considered objects to be named by end users
  - Objects alongside metadata are stored in bucket containers
  - Buckets enable the organization with namespace for user identification & accounting
  - (Automatically) scalable

Expedia airbnb NETFLIX Allergan NASA JPL CATHAY PACIFIC

SCALABLE

Amazon S3 automatically scales to high request rates. It is simple to increase your read or write performance exponentially.

DURABLE

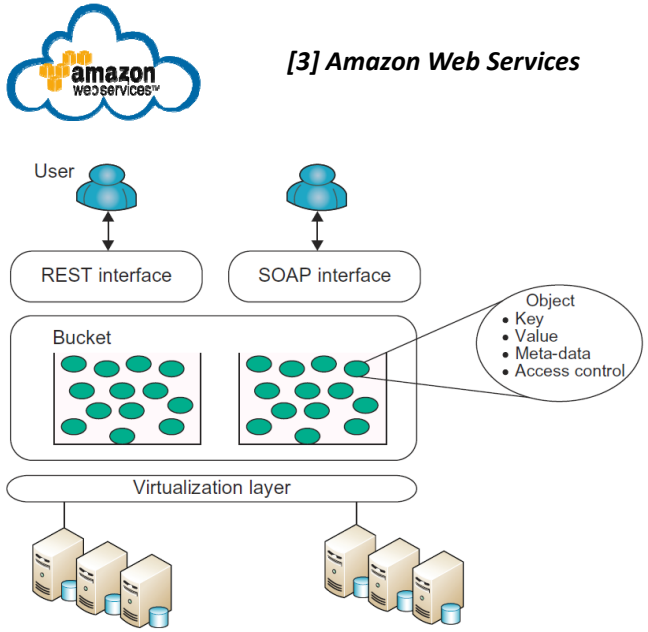
Amazon S3 runs on the world's largest global cloud infrastructure, and is designed to deliver 99.999999999% of durability.

SECURE

Amazon S3's highly durable, secure, global infrastructure offers a robust disaster recovery solution designed to provide superior data protection.

EASY MIGRATION

You can choose from the widest range of options to transfer your data into (or out of) Amazon S3. Our simple APIs make it easy to transfer data over the Internet.



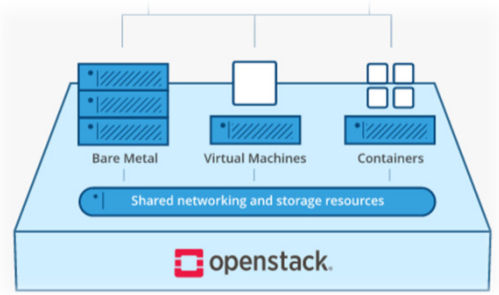
[3] Amazon Web Services

[8] Distributed & Cloud Computing Book

➢ **Lecture 8 provides more details about Amazon Web Services and its Infrastructure-as-a-Service (IAAS) models & various cloud services**

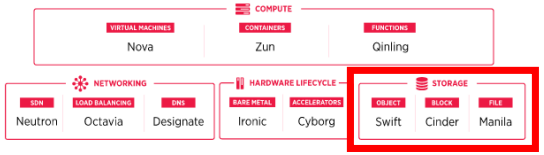
Build your own Cloud – OpenStack Storage Service Examples

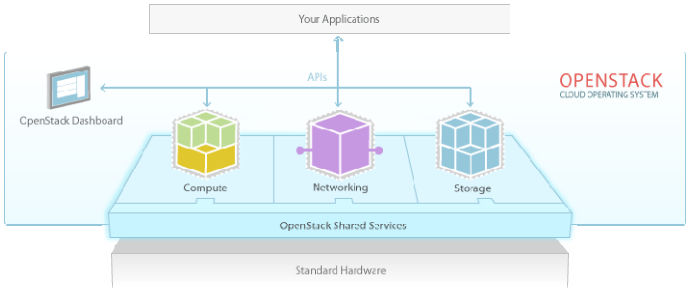
- OpenStack Cloud Operating System
  - Manages and controls cloud resources
  - One of the top 3 most active open source projects and manages 10 million compute cores today



[8] OpenStack Web page

- Swift service
  - Manages and provides object storage as distributed system platform
  - Includes scale-out storage and highly fault tolerant features
  - E.g. storage tasks such as backup, archiving, data retention, etc.
- Cinder service
  - Manages and provides block storage
- Manila service
  - Manages and provides file access





➢ **Lecture 13 provides in-depth insights into features of the OpenStack cloud operating system that enables the creation of own clouds**

‘Big Data‘ Examples – Online Social Networking working with Graph Databases

1 | Facebook
3 - eBizMBA Rank | **1,100,000,000** - Estimated Unique Monthly Visitors | 3 - Compete Rank | 3 - Quantcast Rank | 2 - Alexa Rank | *Last Updated* January 1, 2017.
The Most Popular Social Networking Sites | eBizMBA

2 | YouTube
3 - eBizMBA Rank | **1,000,000,000** - Estimated Unique Monthly Visitors | 4 - Compete Rank | 2 - Quantcast Rank | 3 - Alexa Rank | *Last Updated:* January 1, 2017.
The Most Popular Social Networking Sites | eBizMBA

3 | Twitter
12 - eBizMBA Rank | **310,000,000** - Estimated Unique Monthly Visitors | 21 - Compete Rank | 8 - Quantcast Rank | 8 - Alexa Rank | *Last Updated* January 1, 2017.
The Most Popular Social Networking Sites | eBizMBA

4 | LinkedIn
18 - eBizMBA Rank | **255,000,000** - Estimated Unique Monthly Visitors | 25 - Compete Rank | 19 - Quantcast Rank | 9 - Alexa Rank | *Last Updated* January 1, 2017.
The Most Popular Social Networking Sites | eBizMBA

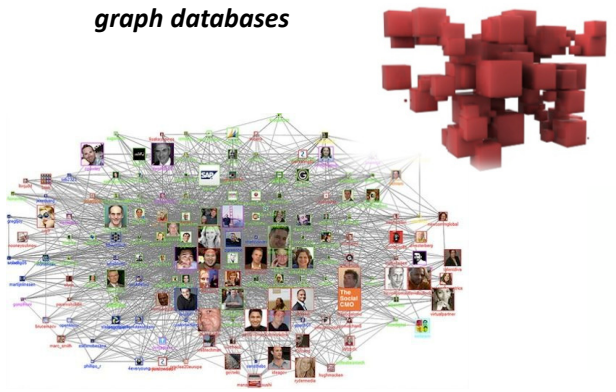
5 | Pinterest
22 - eBizMBA Rank | **250,000,000** - Estimated Unique Monthly Visitors | 27 - Compete Rank | 13 - Quantcast Rank | 26 - Alexa Rank | *Last Updated* January 1, 2017.
The Most Popular Social Networking Sites | eBizMBA

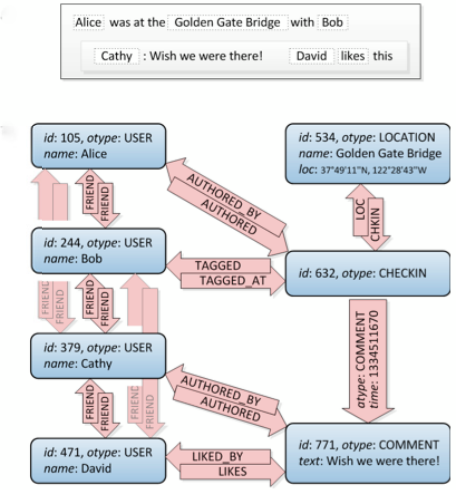
6 | Google Plus+
30 - eBizMBA Rank | **120,000,000** - Estimated Unique Monthly Visitors | *32* - Compete Rank | *28* - Quantcast Rank | NA - Alexa Rank | *Last Updated* January 1, 2017.
The Most Popular Social Networking Sites | eBizMBA

7 | Tumblr
34 - eBizMBA Rank | **110,000,000** - Estimated Unique Monthly Visitors | 55 - Compete Rank | *13* - Quantcast Rank | 34 - Alexa Rank | *Last Updated* January 1, 2017.
The Most Popular Social Networking Sites | eBizMBA

[23] Top 15 most popular social networking sites

[24] www.mashable.com, graph databases





[25] Facebook Engineering, TAO graph database

- **Most of the popular Onlince Social Networking (OSN) Web sites are installed with a client/server architecture where a large number of servers form a data center or using several data centres**
- **OSN key technologies include the use of graph databases**
- **Graph theoretical analysis of OSNs can be used, for instance, for social graph traversal along specific social links or networks in order to understand user behaviour & influence & trends**

➢ **Lecture 14 provides insights into online social networking systems & how they use cloud and big data with graph database technologies**

Role of Scalability & Wide-Area Networks

- High-bandwidth networking
  - Increases the capability of building massively distributed systems
  - Enables elastic computing and scalability beyond one server or one data center
  - Interconnected cloud data centers & servers raise demands for 'perfect networks'
- 'Data locality' as major Cloud & 'Big Data Analytics' movement
  - Requirements for scalable programming models and tools
  - CPU speed has surpassed IO capabilities of existing cloud resources
  - Data-intensive clouds with advanced analytics and analysis capabilities
  - Considering moving compute task to data vs. moving data to compute
  - E.g., map-reduce paradigm had a major impact in cloud adoptions

(does it make sense to transfer TBs/PBs again and again for applications?)

Google *[26] MapReduce: Simplified Dataset on Large Clusters, 2004*

[8] Distributed & Cloud Computing Book

- **We observe tremendous price/performance ratio of commodity hardware that is driven by the desktop, notebook, and tablet computing markets today**
- **Price/performance ratio has also driven the adoption and use of commodity hardware in large-scale and distributed computing including high-bandwidth network increases**

➢ **Lecture 5 provides more details on how the map-reduce paradigm works and how it enables a variety of cloud applications today**

Big Data Analytics vs. Data Analysis

- Data Analysis supports the search for 'causality'
  - Describing exactly WHY something is happening
  - Understanding causality is hard and time-consuming
  - Searching it often leads us down the wrong paths
  - Focus on the findings in the data and not on enabling infrastructure

- Big Data Analytics focussed on 'correlation'
  - Not focussed on causality – enough THAT it is happening
  - Discover novel patterns and WHAT is happening more quickly
  - Using correlations for invaluable insights – often data speaks for itself
  - Often includes technologies and enabling infrastructure (e.g., scalability)

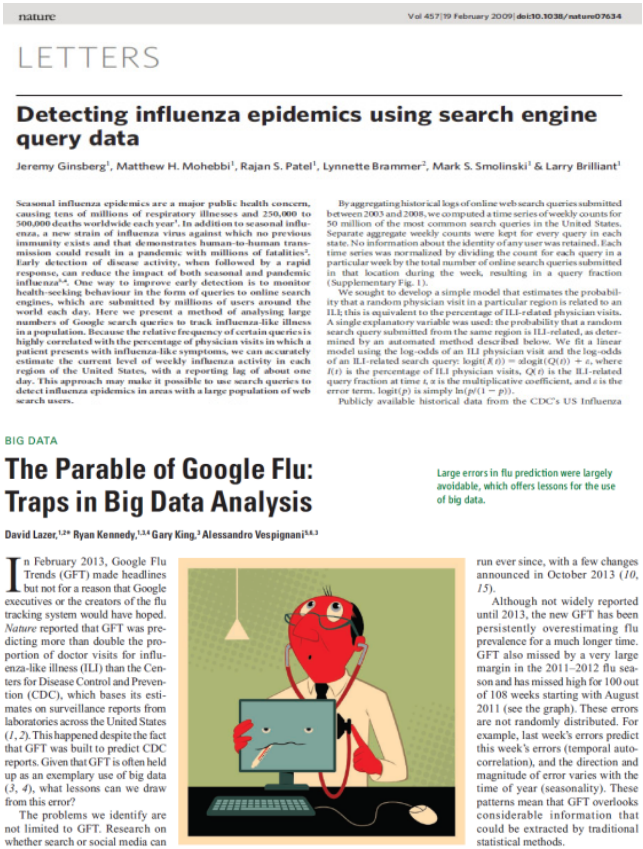
- **'Big Data Analytics' are powerful techniques to work on large data including the enabling infrastructure to work with 'big data' using Clouds**
- **Data Analysis is the in-depth interpretation of research data that is often part of a concrete 'Big Data Analytics' approach**

➢ **Lecture 5 provides more details on how the map-reduce paradigm works and how it enables a variety of cloud applications today**

Big Data Analytics Applications – Example

- ~2009 – H1N1 Virus Made Headlines
  - Google using ‘logged big data‘ → search queries
  - Explains how Google is able to predict fast winter flus
  - Not only on national scale, but down to regions
- ~2014 – The Parable of Google Flu
  - Large errors in flu prediction & lessons learned

- **Large errors are possible when working with ‘big data‘ to infer insights with ‘statistical data mining‘ methods**
- **(1) Dataset: Transparency & Replicability impossible**
- **(2) Study the algorithm since they keep changing in Google: making reproducability impossible**
- **(3) It‘s not just about the size of the data: elements like data quality and many other factors play a role too**



nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities[2]. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza[3,4]. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha \text{logit}(Q(t)) + \varepsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t, α is the multiplicative coefficient, and ε is the error term. $\text{logit}(p)$ is simply $\ln(p/(1-p))$.

Publicly available historical data from the CDC's US Influenza

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,4,5]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can

run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

[30] Jeremy Ginsburg et al., ‘Detecting influenza epidemics using search engine query data’, Nature 457, 2009

[31] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, ‘The Parable of Google Flu: Traps in Big Data Analysis’, Science Vol (343), 2014

Big Data Analytics Frameworks

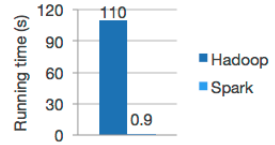
- Distributed Processing
  - ‘Map-reduce via files‘: Tackle large problems with many small tasks
  - Advantage of ‘data replication‘ via specialized distributed file system
  - E.g. Apache Hadoop

[27] Apache Hadoop

- In-Memory Processing
  - Perform many operations fast via ‘in-memory‘
  - Enable tasks such as ‘map-reduce in-memory‘
  - Needs hardware systems that offer large memory
  - E.g. Apache Spark, Apache Flink

[28] Apache Spark

(does it make sense to transfer TBs/PBs again and again for applications?)



Logistic regression in Hadoop and Spark

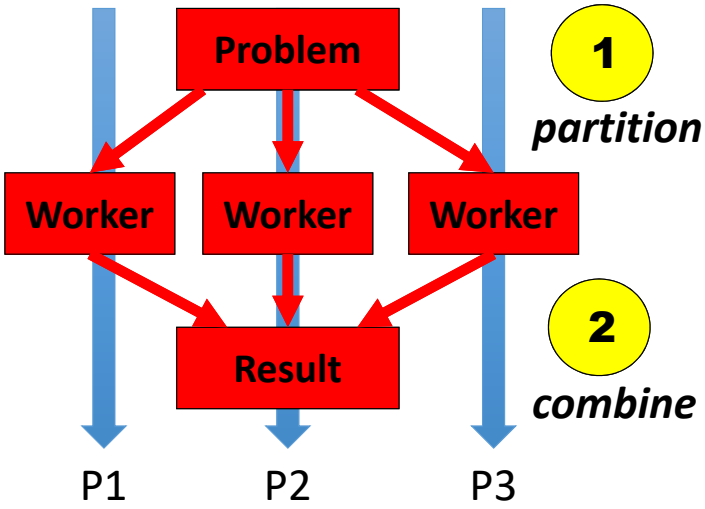
- **Big Data analytics frameworks shift the approach from ‘bring data to compute resources‘ into ‘bring compute tasks close to data‘ including infrastructure technologies (e.g., Apache Hadoop and/or Apache Spark)**

➢ **Lecture 3 provides more details on how the Apache Spark system works and how it is used in cloud computing applications today**

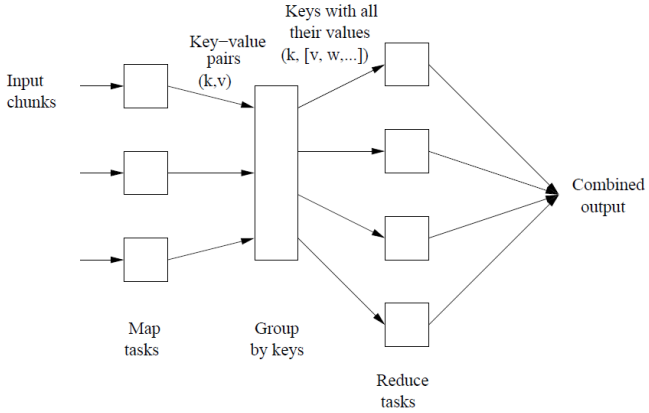
Map-Reduce Computing Paradigm & Open Source Implementation

- Idea not completely new
  - Derived from traditional divide & conquer strategy

Divide & Conquer



- Map-Reduce Paradigm
  - Follows divide & conquer approach
  - Injects a key element between this process: short/shuffle/group
  - Open Source Implementation: Apache Hadoop



Google

[26] MapReduce: Simplified Dataset on Large Clusters, 2004

hadoop

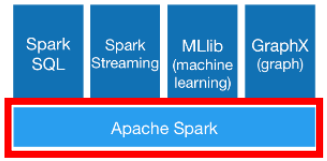
[27] Apache Hadoop

- **Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models**
- **Apache Hadoop is an open source implementation of the map-reduce computing paradigm**
- **Apache Hadoop is designed to scale up from single servers to thousands of machines, each offering local computation and storage**

➢ **Lecture 5 provides more details on how the map-reduce paradigm works and how it is implemented in the Apache Hadoop software**

Networking & Big Data Impacts on Cloud Computing

- Requirements for scalable programming models and tools
  - CPU speed has surpassed IO capabilities of existing cloud resources
  - Data-intensive clouds with advanced analytics and analysis capabilities
  - Considering moving compute task to data vs. moving data to compute
  - E.g., MS Azure HDInsight is only one concrete example of many cloud solutions
  - E.g., clouds often use open source implementations like Apache Spark & Apache Hadoop & related technologies



[28] Apache Spark

Microsoft Azure

[32] Microsoft Azure HDInsight Service

HDInsight

- **Apache Spark is a fast & general engine for large-scale data processing optimized for memory usage**
- **Open source & compatible with cloud data storage systems like Amazon Web Services (AWS) S3**
- **Apache Spark is often used as cloud service offerings in commercial clouds like MS Azure, AWS, or Google Cloud**

- Requirements for Reliable Filesystems
  - Traditional parallel filesystems need to prove their 'big data' feasibility
  - Emerging new forms of filesystems that assume hardware error constantly
  - E.g. Hadoop distributed file system (HDFS)

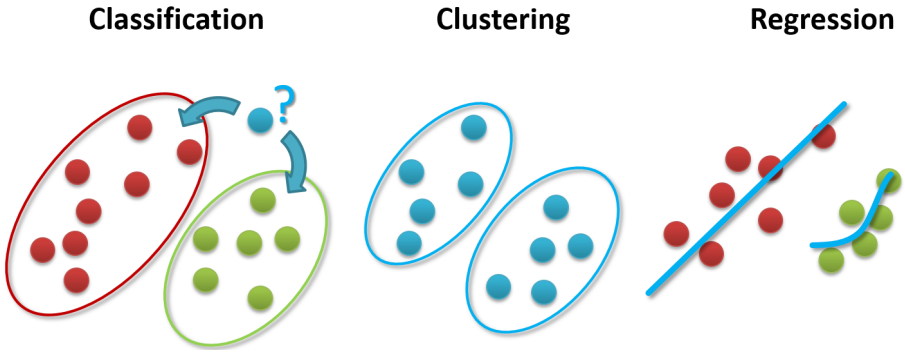
[27] Apache Hadoop

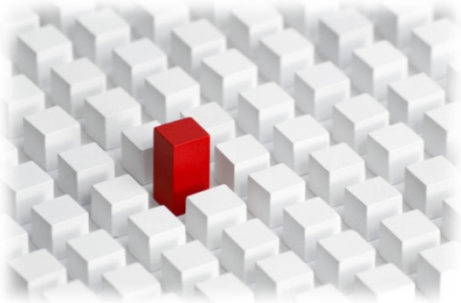
➢ **Lecture 3 provides more details on how the Apache Spark system works and how it is used in cloud computing applications today**

Big Data Motivation: Intertwine Clouds & Machine/Deep Learning

- Rapid advances in data collection and storage technologies in the last decade
  - Extracting useful information is a challenge considering ever increasing massive datasets
  - Traditional data analysis techniques cannot be used in growing cases (e.g. memory, speed, etc.)
- Machine Learning & Statistical Data Mining
  - Traditional statistical approaches are still very useful to consider
  - Deep Learning tools become effective and are available in Clouds today

- **Machine learning / Data Mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data (aka 'big data')**
- **Machine Learning / Data Mining is the process of automatically discovering useful information in large data repositories ideally following a systematic process**

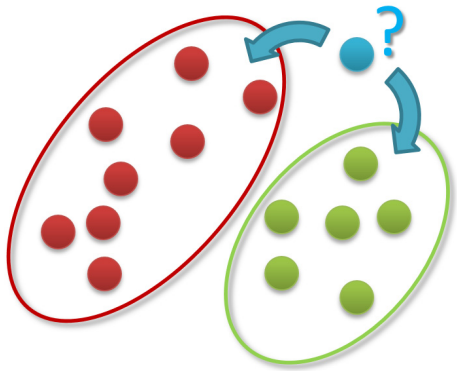




➢ **Lecture 2 offers a moderate introduction to machine learning models with an emphasize on their usage in this course and clouds**

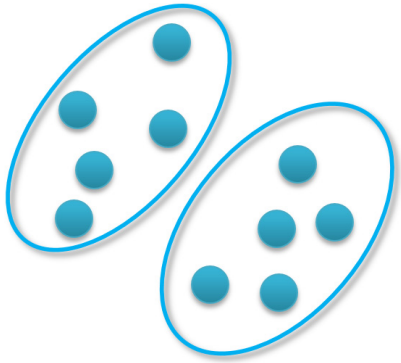
Machine Learning Models – Short Overview

Classification



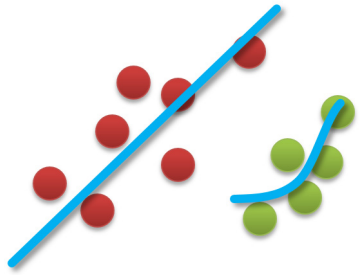
- Groups of data exist
- New data classified to existing groups

Clustering



- No groups of data exist
- Create groups from data close to each other

Regression



- Identify a line with a certain slope describing the data

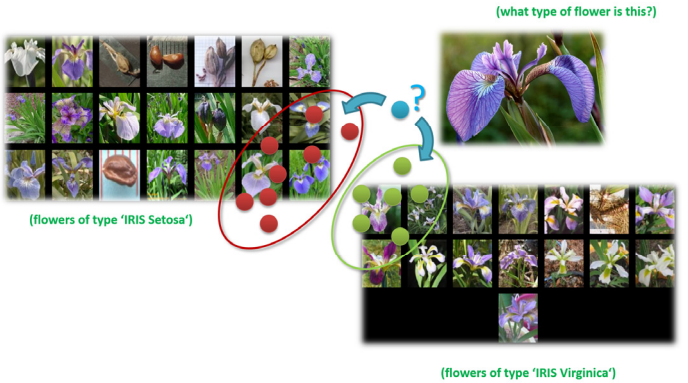
- **Machine learning methods can be roughly categorized in classification, clustering, or regression augmented with various techniques for data exploration, selection, or reduction – despite the momentum of deep learning, traditional machine learning algorithms are still widely relevant today**

➢ **Lecture 2 offers a moderate introduction to machine learning models with an emphasize on their usage in this course and clouds**

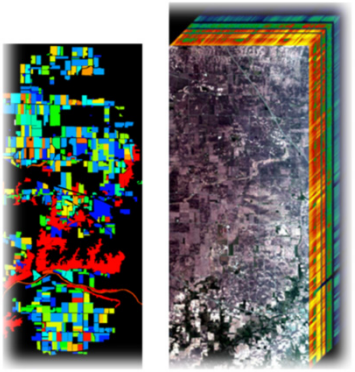
Learning Approaches – What means Learning from data?

- **The basic meaning of learning is 'to use a set of observations to uncover an underlying process'**
- **The three different learning approaches are supervised, unsupervised, and reinforcement learning**

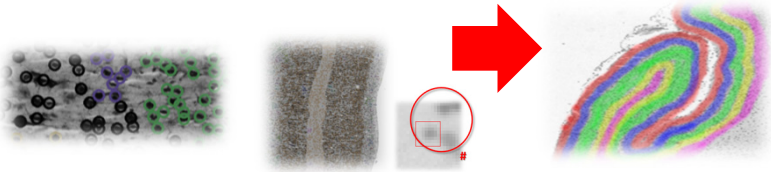
- Supervised Learning
  - Majority of methods follow this approach in this course
  - Example: credit card approval based on previous customer applications



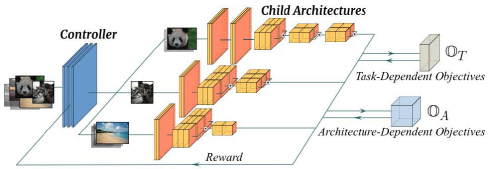
[34] Image sources: Species Iris Group of North America Database, www.signa.org



- Unsupervised Learning
  - Often applied before other learning → higher level data representation
  - Example: Coin recognition in vending machine based on weight and size



[35] A.C. Cheng et al., 'InstaNAS: Instance-aware Neural Architecture Search', 2018



- Reinforcement Learning
  - Typical 'human way' of learning
  - Example: Toddler tries to touch a hot cup of tea (again and again)

➢ **Lecture 2 offers a moderate introduction to machine learning models with an emphasize on their usage in this course and clouds**

Cloud Computing Approach with Microsoft Azure HDInsight for Machine Learning

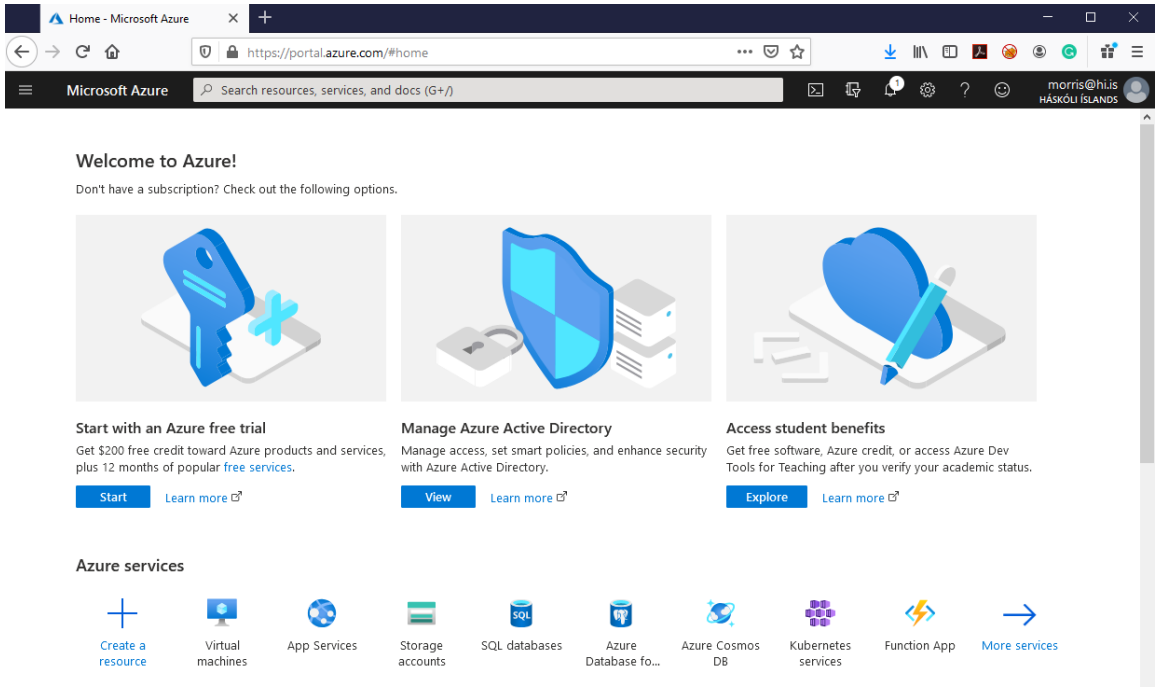
- Microsoft Azure Cloud
  - Wide variety of different cloud-based services & resources for many application areas
  - Managed via Microsoft Azure Portal
  - Needs a Microsoft Azure account
  - Provides Apps to monitor Cloud usage
- Microsoft Azure HDInsight Service
  - Open source frameworks in MS Cloud

Apache Hadoop, Apache Spark, Apache Kafka, Apache HBase, Apache Hive LLAP, Apache Storm, Machine Learning

  - Accessible via Jupyter *[2] Jupyter*

Microsoft Azure HDInsight *[32] Microsoft Azure HDInsight Service*

[33] Azure Portal Hub



Reasoning for using Cloud Computing & Step-wise Approach with HDInsight

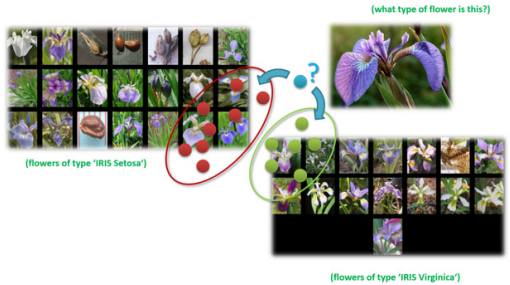
- Big Data Challenges
  - Jupyter, Anaconda and scikit-learn are great – but only for small data sets
  - Issues: Laptop too slow runtime (less CPUs/GPUs) and/or errors due to memory problems/limits

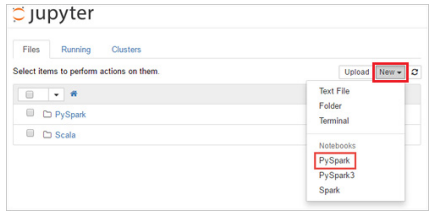
- Cloud Computing Step-wise Approach
  1. Check and/or create subscription (setup pay-per-use of resources, free in course)
  2. Deploy a Spark Cluster in HDInsight (prepare computing infrastructure) using a pre-configured Azure Resource Manager (ARM) template
  3. Startup Jupyter notebook using a PySpark kernel to use Apache Spark
  4. Create Spark session & check application dataset
  5. Initial data analysis, visualization & modeling
  6. Perform machine learning model on Spark Cluster
  7. Machine learning model evaluation & refinement

Microsoft Azure HDInsight

[32] Microsoft Azure HDInsight Service

[2] Jupyter



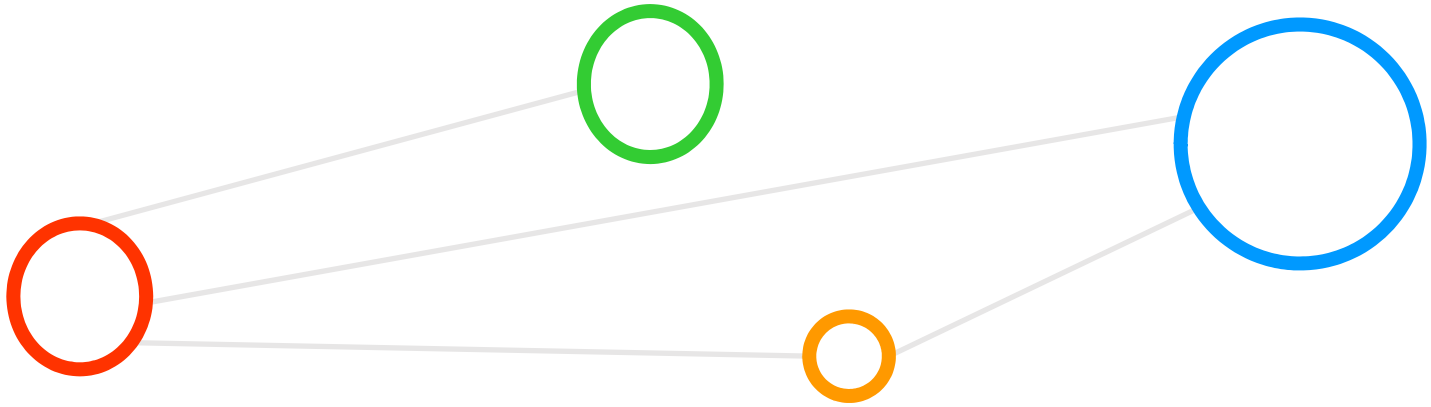


[34] Image sources: Species Iris Group of North America Database, www.signa.org

[Video] Big Data Analytics in Banking Industry

[29] IBM Big Data and Analytics

Lecture Bibliography



Lecture Bibliography (1)

- [1] Python Programming Language, Online: https://www.python.org/
- [2] Project Jupyter, Online: https://jupyter.org/
- [3] Amazon Web Services Web page, Online: https://aws.amazon.com/
- [4] Microsoft Azure, Online: https://azure.microsoft.com/en-us/
- [5] Google Cloud, Online: https://cloud.google.com/
- [6] Jupyter @ Juelich Supercomputing Centre, Online: https://jupyter-jsc.fz-juelich.de/
- [7] A. Rosebrock, 'Get off the deep learning bandwagon and get some perspective', Online: http://www.pyimagesearch.com/2014/06/09/get-deep-learning-bandwagon-get-perspective/
- [8] K. Hwang, G. C. Fox, J. J. Dongarra, 'Distributed and Cloud Computing', Book, Online: http://store.elsevier.com/product.jsp?locale=en_EU&isbn=9780128002049
- [9] DEEP Series Projects Web Page, Online: http://www.deep-projects.eu/
- [10] BOINC Middleware Tool, Online: https://boinc.berkeley.edu/
- [11] YouTube Video, 'The Three Ways to Cloud Compute', Online: https://www.youtube.com/watch?v=SgujaIzkwrE

Lecture Bibliography (2)

- [12] Lego Bricks, Online: https://commons.wikimedia.org/wiki/File:Lego_dimensions.svg
- [13] Google Play Angry Birds, Online: https://play.google.com/store/apps/details?id=com.rovio.angrybirds&hl=en
- [14] Google Cloud Platform Case Study, 'Rovio', Online: https://cloud.google.com/customers/rovio
- [15] AWS Amazon Sagemaker Service, Online: https://aws.amazon.com/sagemaker
- [16] Big Data Tips – Big Data Mining & Machine Learning, Online: http://www.big-data.tips/
- [17] Introduction to High Performance Computing for Scientists and Engineers, Georg Hager & Gerhard Wellein, Chapman & Hall/CRC Computational Science, ISBN 143981192X
- [18] OpenStack Web page, Online: https://www.openstack.org/software/
- [19] European Open Science Cloud (EOSC) Web page, Online: https://www.eosc-portal.eu/
- [20] EU EOSC-Nordic Project Web page, Online: https://www.eosc-nordic.eu/
- [21] European Commission Expert Group on FAIR Data Final Report, Online: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf
- [22] Go-Fair Initiative, Online: https://www.go-fair.org/resources/faq/go-fair-relation-to-eosc-and-ifds/

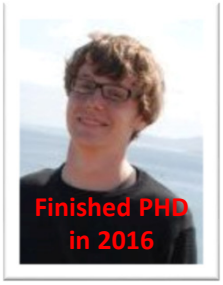
Lecture Bibliography (3)

- [23] Top 15 Most Popular Social Networking Sites, Online: http://www.ebizmba.com/articles/social-networking-websites
- [24] www.mashable.com, 'Graph Databases: The New Way to Access Super Fast Social Data', Online: http://mashable.com/2012/09/26/graph-databases/
- [25] Facebook Engineering, 'TAO – The Power of the graph', Online: https://www.facebook.com/notes/facebook-engineering/tao-the-power-of-the-graph/10151525983993920/
- [26] J. Dean, S. Ghemawat, 'MapReduce: Simplified Data Processing on Large Clusters', OSDI'04: Sixth Symposium on Operating System Design and Implementation, December, 2004
- [27] Apache Hadoop Web page, Online: http://hadoop.apache.org/
- [28] Apache Spark Web page, Online: http://spark.apache.org/
- [29] 'Demo: IBM Big Data and Analytics at work in Banking', YouTube Video, Online: https://www.youtube.com/watch?v=1RYKgj-QK4I
- [30] Jeremy Ginsburg et al., 'Detecting influenza epidemics using search engine query data', Nature 457, 2009
- [31] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, 'The Parable of Google Flu: Traps in Big Data Analysis', Science Vol (343), 2014
- [32] Microsoft Azure HDInsight Service, Online: https://azure.microsoft.com/en-us/services/hdinsight/
- [33] Microsoft Azure Portal Hub, Online: https://portal.azure.com/#home

Lecture Bibliography (4)

- [34] Species Iris Group of North America Database, Online: http://www.signa.org
- [35] Cheng, A.C, Lin, C.H., Juan, D.C., InstaNAS: Instance-aware Neural Architecture Search, Online: https://arxiv.org/abs/1811.10201

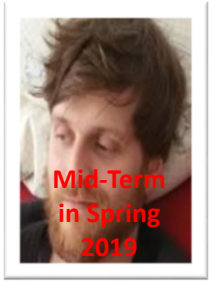
Acknowledgements – High Productivity Data Processing Research Group



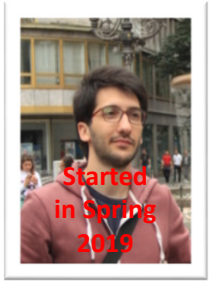
PD Dr.
G. Cavallaro

Senior PhD
Student A.S. Memon

Senior PhD
Student M.S. Memon



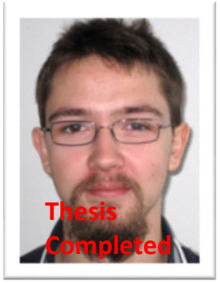
PhD Student
E. Erlingsson



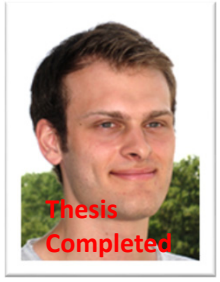
PhD Student
S. Bakarat

PhD Student
R. Sedona

Dr. M. Goetz
(now KIT)



MSc M.
Richerzhagen
(now other division)



MSc
P. Glock
(now INM-1)

MSc
C. Bodenstein
(now
Soccerwatch.tv)

MSc Student
G.S. Guðmundsson
(Landsverkjun)

This research group has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 763558 (DEEP-EST EU Project)


services
measurement
within
fusion
new
funding
Policy-based
concepts
device
Analysis
forms
cross-disciplinary
resources
Centre
climate
Computer
expertise
Approaches
construction
Fusion
year
like
Computational
Test
E.g
Computing
disciplines
results
Enable
different
Cross-Disciplinary
methods
High
big
nodes
PB
community
Environment
Modelling
EU
Science
centers
Key
scientific
important
files
model
etc
DKRZ
Use
storage
basis
Infrastructure
international
simulations
ENES
technologies
SMAQ
engineering
References
formats
brain
increasing
via
EUDAT
physical
IPCC
Big
Project
Data
processing
Common
TB
MapReduce
create
Methods
computing
Using
European
databases
GB
Supercomputing
Work
images
codes
analysis
stored
Resources
analyze
Large
Summary
data-intensive
Service
Scientific
levels
often
use
DLCL
Level
external
JSC
computational
Research
Matters
Concrete
performance
research
e.g
science
HPC
datasets
acquisition
large-scale
Juelich
system
Landscape
management
compute
System
Climate
modelling
AR5
Hardware
cores
access
hundreds
Nutshell
Services
Understanding
Technologies
Earth
distributed
Structure
current
network
DLCLs
structures
directory
project
General
Germany
Simulation
Provide
NASA
Energy
systems
day
Processing
Health