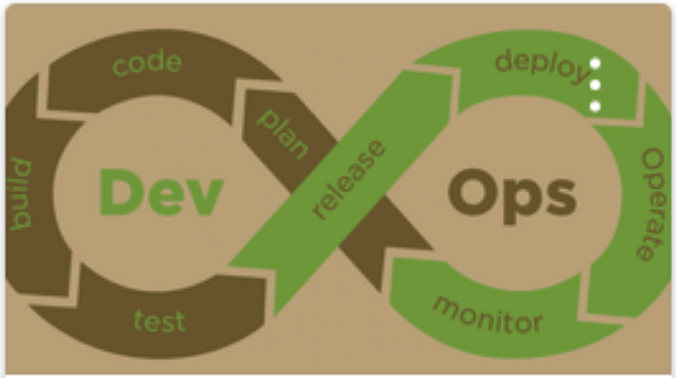
UPPSALA UNIVERSITET

# Data Engineering II

## 1TD076 62016




Salman Toor
salman.toor@it.uu.se

UPPSALA UNIVERSITET

# Teachers 2025

- Salman Toor (salman.toor@it.uu.se) (Distributed Computing Infrastructures, Applied Machine Learning)

- Teaching Assistant
  - Li Ju (li.Ju@it.uu.se)

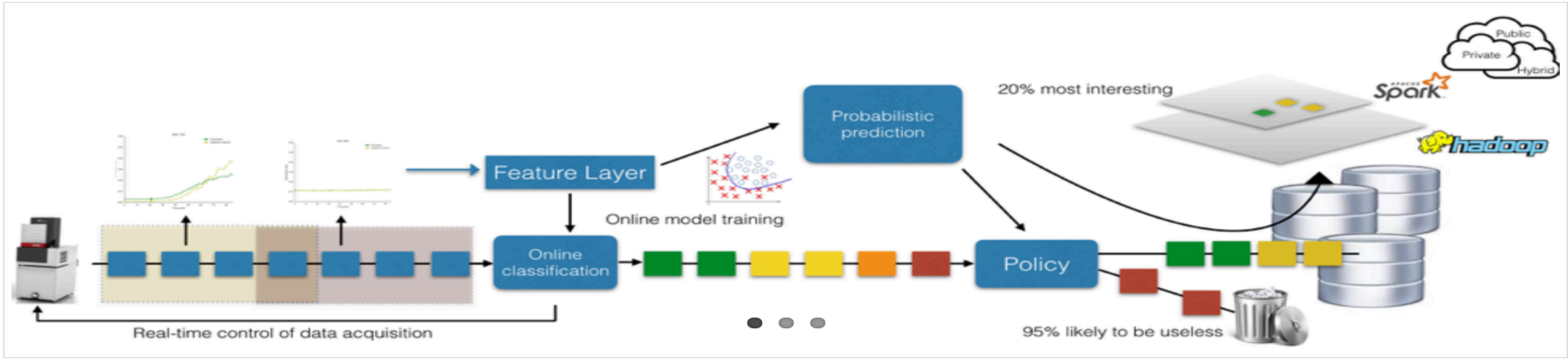
  - Anand Mathew Muthukulam Simon (anand-mathew.muthukulam-simon.6015@student.uu.se )
  - Xiong Luo (xiong.luo.7609@student.uu.se )
  - Ella Johanna Schmidtobreick (ella-johanna.schmidtobreick.4283@student.uu.se )
  - Finn Vaughankraska (finn.vaughankraska.2674@student.uu.se )

UPPSALA UNIVERSITET

SciML UU

https://sciml.se/

# Research Group




## Integrative Scalable Computing Laboratory

ISCL is a research group at the Department of Information Technology at Uppsala University. PIs Andreas Hellander and Salman Toor.

### Stochastic simulation

We often use stochastic descriptions to model complex systems. Many of our projects involve kinetic Monte Carlo, agent-based models and multiscale modeling. A reoccurring theme is how to leverage distributed e-infrastructure for simulations and how to use machine learning to construct approximations.

### Artificial intelligence

A core theme in the group is the use of machine learning to make scientific computing software and infrastructure more efficient, interactive and scalable. We also do disciplinary research in specific areas of ML, such as likelihood-free inference and privacy-preserving federated machine learning.

### Distributed computing

Our research in distributed computing and data engineering sciences ranges from development of new ways to manage large and fast data to design and development of massively parallel, interactive, cloud native applications operating in cloud, fog and edge infrastructure.

UPPSALA UNIVERSITET

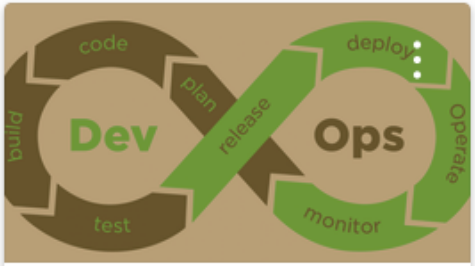
# Data Engineering I


- M1 - Use of distributed infrastructures
- M2 - Data analysis frameworks, Hadoop and Spark
- M3 - Tools to build analysis pipelines

UPPSALA UNIVERSITET

# Data Engineering II

- Course will be based on three modules
  - M1 - Contextualization and containers
  - M2 - Data stream processing
  - M2 - Distributed infrastructures and workflow management
  - M3 - Distributed machine learning



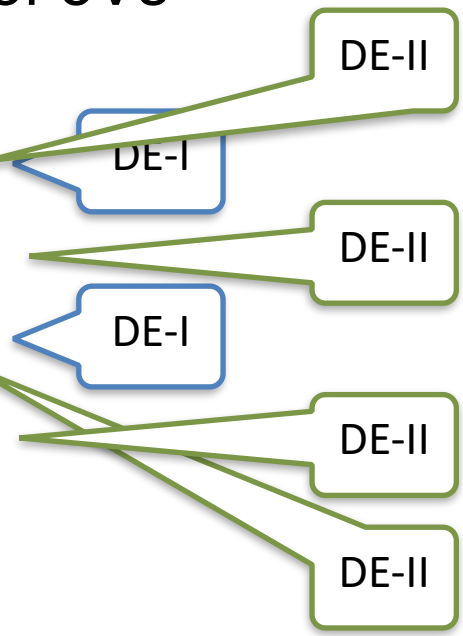

Data Engineering II 1TD076 6201...
1TD076 62016
VT2025

Why are these three areas important?

UPPSALA UNIVERSITET

# World of Big Data

- The power of 5Vs
  - Volume
  - Velocity
  - Variety
  - Veracity
  - Value



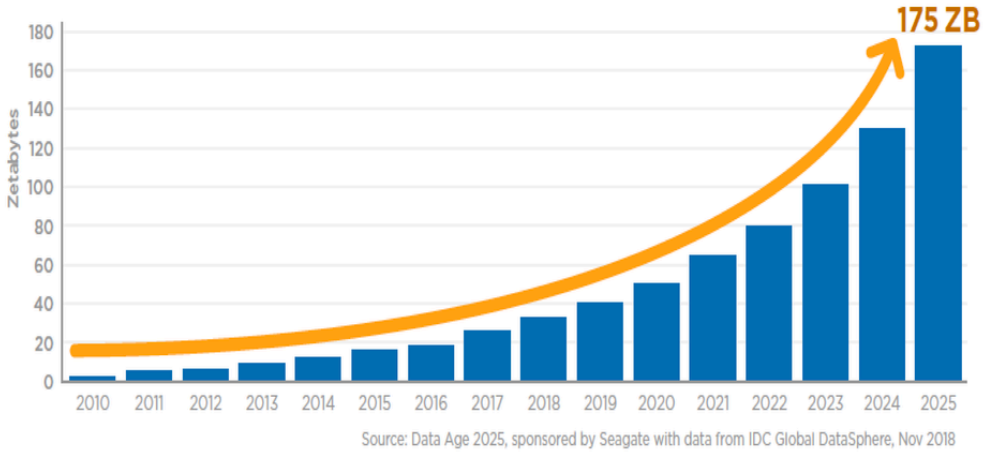


https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d

UPPSALA UNIVERSITET

# World of Big Data

- Volume



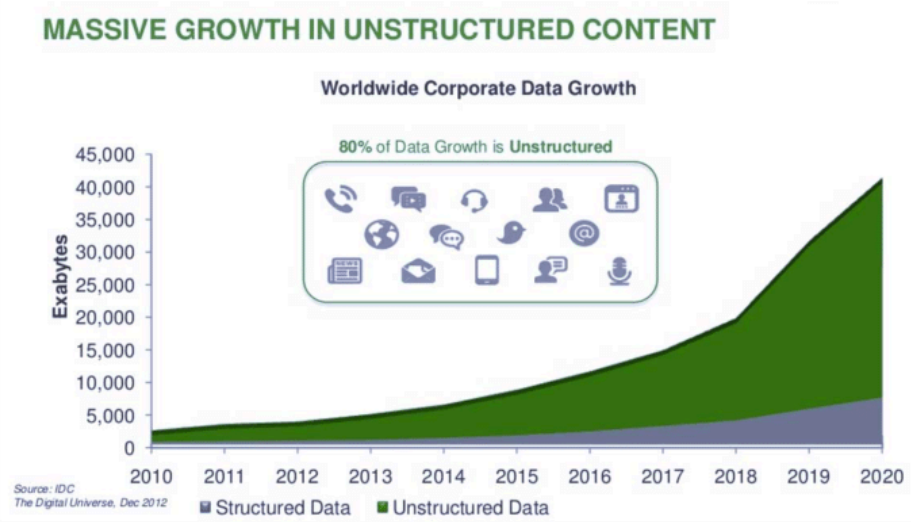

- Velocity
  - In the year 2000, Google was receiving 32.8 million searches per day.
  - In the year 2018, Google search requests increased to 5.6 billion searches per day
  - In 2022, Google processes over 8.5 billion searches per day

https://www.oberlo.com/blog/google-search-statistics#:~:text=Google%20processes%20over%208.5%20billion,most%20searched%20keyword%20on%20Google.

UPPSALA UNIVERSITET

# World of Big Data

- Variety




- Veracity
  - assurance of quality/integrity/credibility/accuracy of the data

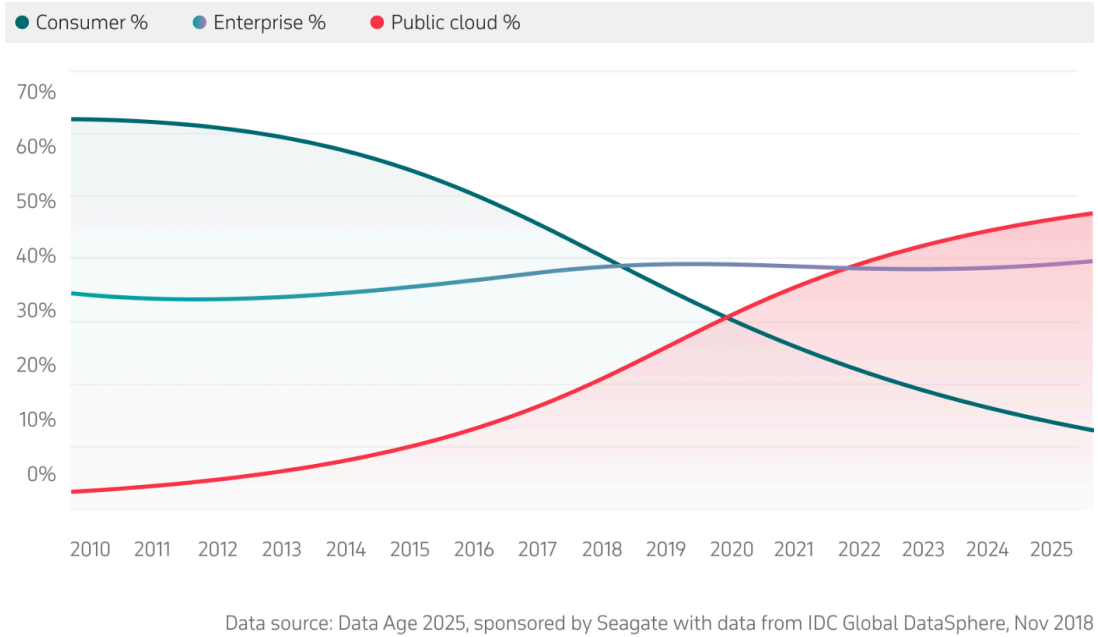
- Value
  - refers to the usefulness of data in decision making

https://medium.com/swlh/if-data-is-the-new-oil-then-most-of-it-is-still-in-the-ground-4a0e41d95031

UPPSALA UNIVERSITET

# Management of Big Data using distributed infrastructures

- Main benefits
  - less expensive
  - no maintenance
  - high availability
  - scalability




Data source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

- Challenges
  - less control
  - not suitable for sensitive data
  - vendor lock-in

UPPSALA UNIVERSITET

# Data stream processing

## Case study of Walmart


245 million customers visiting 10,900 stores and 10 active websites across the globe–Walmart is a name to reckon in the retail sector.

Walmart sees close to **300,000** social mentions every week.

Walmart made a move from the experiential 10 node Hadoop cluster to a **250 node** Hadoop cluster in 2012.

Walmart collects **2.5 petabytes** of unstructured data from 1 million customers every hour.

The analytics systems at Walmart analyse close to 100 million keywords on daily basis to optimize the bidding of each keyword.

Walmart Labs analyses every clickable action on Walmart.com–

1) What consumers buy in-store and online?
2) What is trending on Twitter?
3) Local events such as San Francisco giants winning the World Series?
4) How local weather deviations affect the buying patterns?

https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109

UPPSALA UNIVERSITET

# Data stream processing

*Walmart acquired a small startup Inkiru based in Palo Alto, California to boost its big data capabilites. Inkiru Inc. helps in targeted marketing, merchandising and fraud prevention. Inkiru's predictive technology platform pulls data from diverse sources and helps Walmart improve personalization through data analytics. The predictive analytics platform of Inkiru incorporates machine learning technologies to automatically enhance the accuracy of algorithms and can integrate with diverse external and internal data sources.*

*To fulfil the need for a general purpose real time stream processing platform which can tackle issues like performance and scalability, Walmart developed Mupd8 for Fast Data. With Mupd8, stream processing applications could emphasize on the quality of generated data. Mupd8 does for fast data, what hadoop mapreduce computational model does for big data.*

*For example, an application can be written to subscribe to the Twitter firehose of every tweet written; such an application can analyse the tweets to determine Twitter's most influential users, or identify suddenly prominent events as they occur. Alternatively, an application can be written to subscribe to a log of all user activity on a Web site; such an application can detect service problems users' face as they occur, or compute suggestions for users' next steps based on up-to-the-moment activity.*

https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109

UPPSALA UNIVERSITET

# World of big data and artificial intelligence

*Amazon changes product prices 2.5 million times a day, meaning that an average product listed on Amazon changes prices every 10 minutes.*

*Amazon's pricing update model is fifty times more often than Walmart and Best Buy!*

https://www.businessinsider.com/amazon-price-changes-2018-8?r=US&IR=T#:~:text=Amazon%20changes%20product%20prices%202.5,change%20about%20every%2010%20minutes.

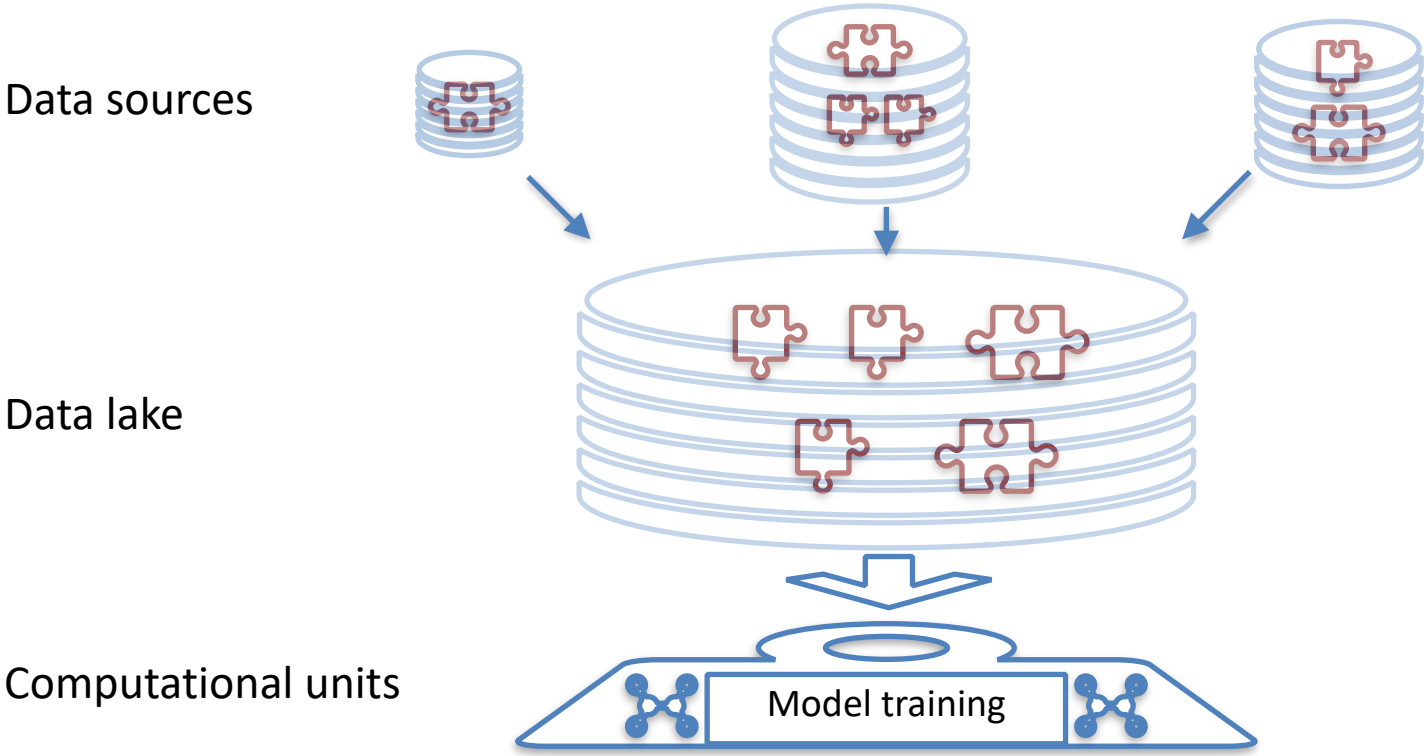
*According to the recently updated International Data Corporation (IDC) Worldwide Artificial Intelligence Systems Spending Guide, spending on AI systems will reach $97.9 billion in 2023, more than two and one half times the $37.5 billion that will be spent in 2019. The compound annual growth rate (CAGR) for the 2018-2023 forecast period will be 28.4%.*

https://www.idc.com/getdoc.jsp?containerId=prUS46794720

UPPSALA UNIVERSITET

# World of artificial intelligence

- Classical machine learning requires
  - collection of a complete dataset at one place
  - enough computational resources on a single site to train a model



UPPSALA UNIVERSITET

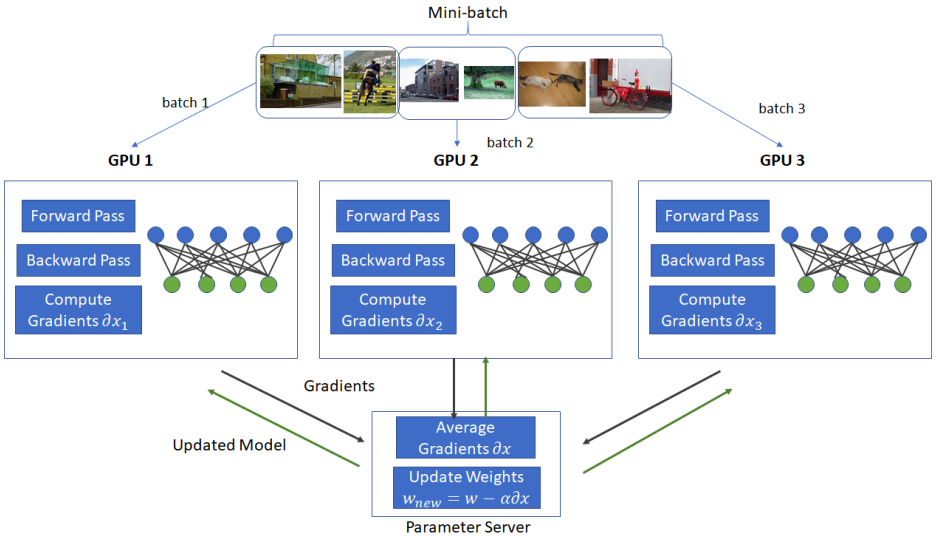
# Classical machine learning

- Benefits
  - comprehensive view of the datasets
  - complete control
  - possibility to try different model training approaches
- Challenges
  - required resources for data transfer
  - cannot share sensitive data
  - data ownership issues
  - efficiency and performance challenges
- Possible solution -> distributed and/or federated machine learning

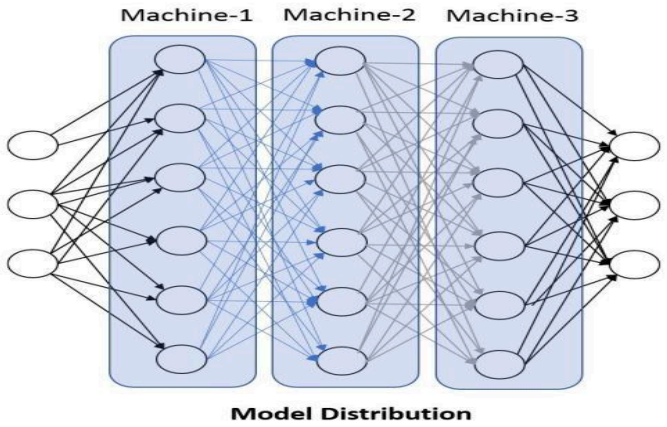
UPPSALA UNIVERSITET

# Distributed machine learning

- Data parallelism




- Model parallelism

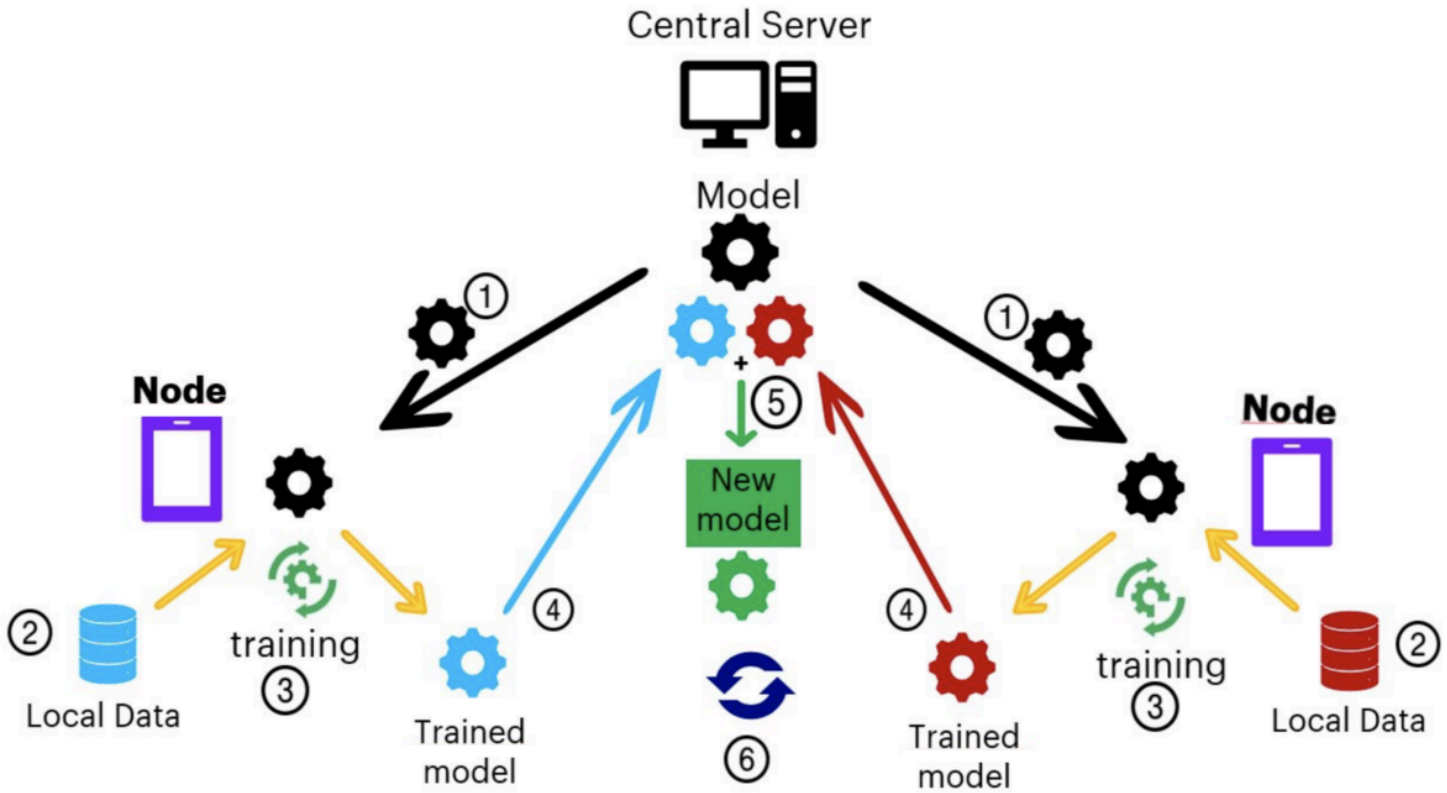



https://www.telesens.co/2017/12/25/understanding-data-parallelism-in-machine-learning/

http://track-n-test.com/machine-learning/top-libraries-for-distributed-deep-learning/

UPPSALA UNIVERSITET

# Federated machine learning

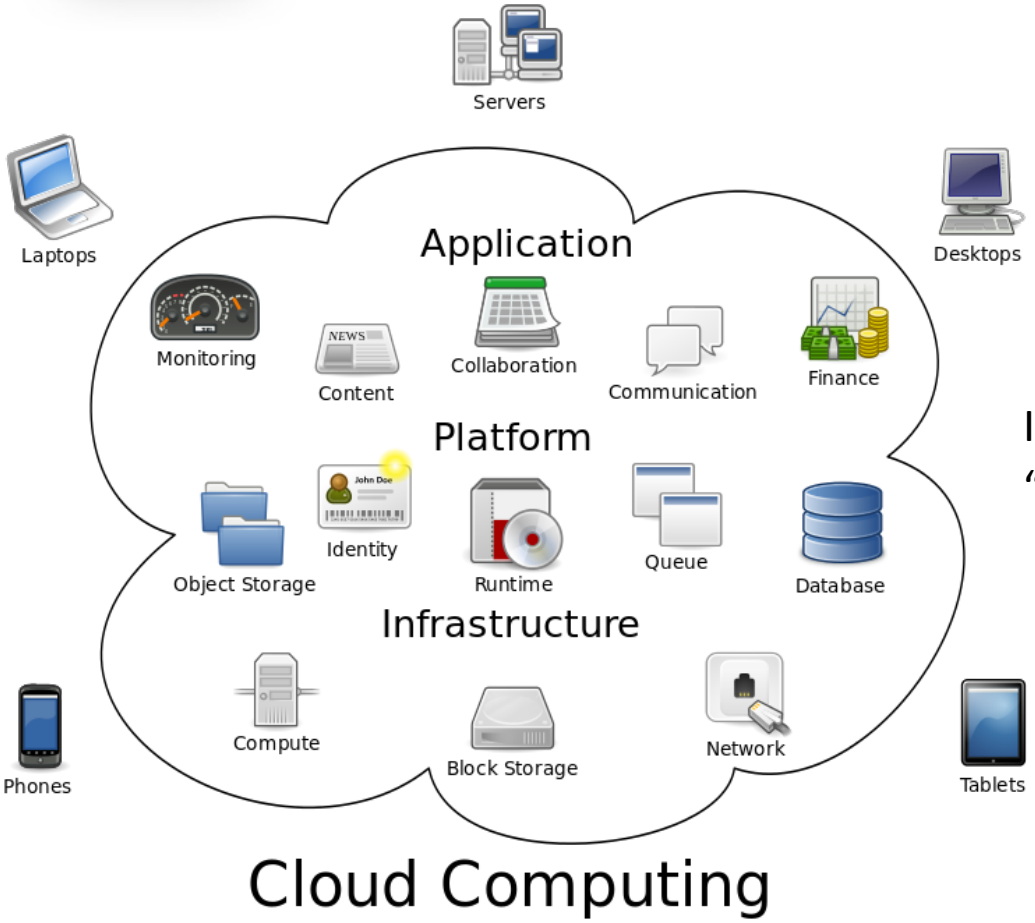





https://medium.com/accenture-the-dock/instilling-responsible-and-reliable-ai-development-with-federated-learning-d23c366c5efd

UPPSALA UNIVERSITET

# Distributed infrastructures

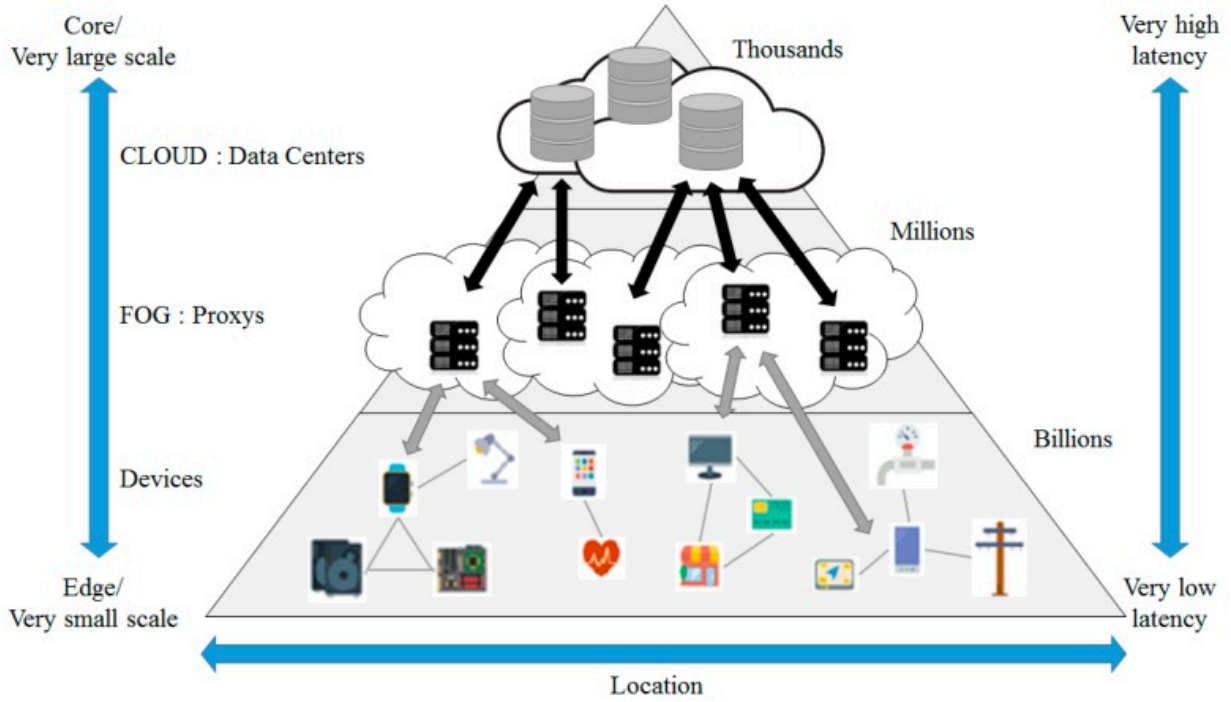



IT-infrastructure and services
"obscured by a a cloud"

UPPSALA UNIVERSITET

# Distributed infrastructures

- Cloud Computing
- Fog Computing
- Edge Computing

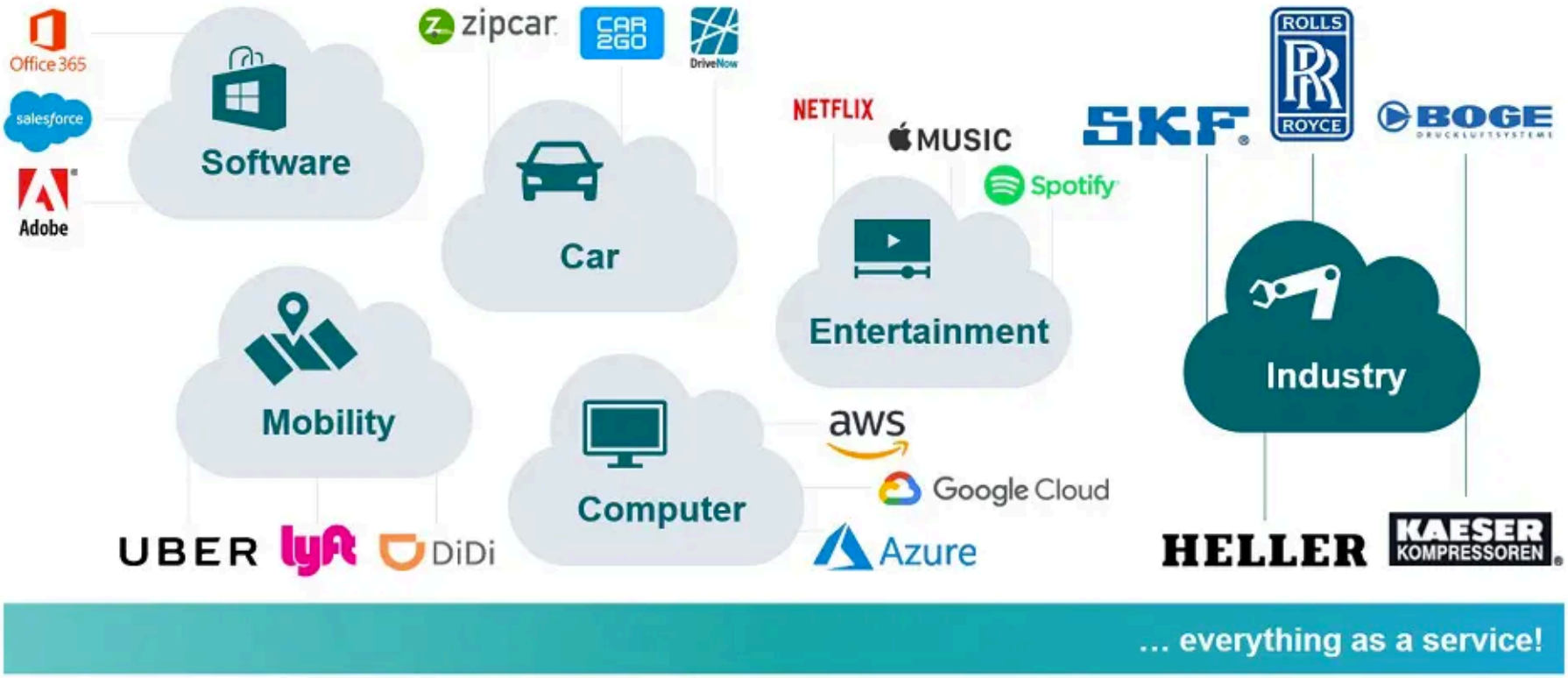



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6263423/

UPPSALA UNIVERSITET

# Distributed infrastructures

- Everything as a Service (*aaS)



UPPSALA UNIVERSITET

# Distributed infrastructures

## AI enabled distributed infrastructures

UPPSALA UNIVERSITET

# Ten years (2008 - 2018)

- BigData
- Connectivity
- Smart Infrastructures
- Artificial Intelligence

## Largest Global companies in 2018 vs 2008

| 2018 | | | | 2008 | | | |
|---|---|---|---|---|---|---|---|
| Rank | Company | Founded | USbn | Rank | Company | Founded | USbn |
| 1. | Apple | 1976 | 890 | 1. | PetroChina | 1999 | 728 |
| 2. | Google | 1998 | 768 | 2. | Exxon | 1870 | 492 |
| 3. | Microsoft | 1975 | 680 | 3. | General Electric | 1892 | 358 |
| 4. | Amazon | 1994 | 592 | 4. | China Mobile | 1997 | 344 |
| 5. | Facebook | 2004 | 545 | 5. | ICBC (China) | 1984 | 336 |
| 6. | Tencent (China) | 1998 | 526 | 6. | Gazprom(Russia) | 1989 | 332 |
| 7. | Berkshire | 1955 | 496 | 7. | Microsoft | 1975 | 313 |
| 8. | Alibaba (China) | 1999 | 488 | 8. | Royal Dutch Shell | 1907 | 266 |
| 9. | J&J | 1886 | 380 | 9. | Sinopec (China) | 2000 | 257 |
| 10. | JP Morgan | 1871 | 375 | 10. | AT&T | 1885 | 238 |

Source: Bloomberg, Google

UPPSALA UNIVERSITET

# Course overview: Advanced concepts of Data Engineering

- L1, -> Course Introduction
- L2, -> (M1) Contextualization and containers
- C1, -> Docker based container orchestration exercise. (Online)
- C1, -> Docker based container orchestration exercise. (Online)
- L4, -> (M2) Data Stream Processing (Part-1)
- L5, -> (M2) Data Stream Processing (Part-2)

UPPSALA UNIVERSITET

# Course overview: Advanced concepts of Data Engineering

- C2,  -> A practical introduction to one of the popular Data stream processing framework Apache Pulsar.  (Online)
- C2, -> A practical introduction to one of the popular data stream processing frameworks Apache Pulsar. (Online)

- L6, -> (M3) Distributed Computing Infrastructures
- L7, -> (M3) Continuous Integration + Model serving

- C3, -> Design and implementation of computational workflows … (Online)
- C3, -> Design and implementation of computational workflows … (Online)

UPPSALA UNIVERSITET

# Course overview: Advanced concepts of Data Engineering

- L8, -> Announcement of the projects and literature seminars

- L9, -> (M4) Distributed Machine Learning
- L10, -> (M3) Distributed Machine Learning

- C4, -> Federated and distributed machine learning. (Online)
- C4, -> Federated and distributed machine learning. (Online)

- S, -> Literature Seminars (Onsite)
- S, -> Literature Seminars (Onsite)
- S, -> Literature Seminars (Onsite)

- L11, -> Guest Lecture (Onsite)

UPPSALA UNIVERSITET

# Course overview: Advanced concepts of Data Engineering

- CX, -> Extra lab session (Online)

- Project presentations (Onsite)
- Project presentations (Onsite)
- Project presentations (Onsite)

UPPSALA UNIVERSITET

# Course overview: Advanced concepts of Data Engineering

Computer Lab 1
- Announcement date: 28-03-2025
- Deadline: 08-04-2025

Computer Lab 2
- Announcement date: 07-04-2025
- Deadline: 24-04-2025

Computer Lab 3
- Announcement date: 24-04-2025
- Deadline: 06-05-2025

Computer Lab 4
- Announcement date: 06-05-2025
- Deadline: 15-05-2025

Literature Seminar
- Announcement date: 28-04-2025
- Deadline: 12-05-2025

Mini-project
- Announcement date: 28-04-2025
- Deadline: 28-05-2025

UPPSALA UNIVERSITET

# Assessment summary

Assessment is based on:

1. Obligatory and Optional Computer Lab assignments
2. Participation in Literature Seminar
3. Completion of a mini-project

UPPSALA UNIVERSITET

# Computer labs

- The computer assignments have a mandatory as well as an optional part.
- They should be presented in a (brief report), uploaded to the studium.
- All labs should be conducted ***individually***, but you may of course discuss concepts with peers.
- Successful completion of the non-optional part gives points that count toward higher marks

Hypothesis: Time spent with fingers on the keyboard is strongly correlated with learning the concepts.

UPPSALA UNIVERSITET

# Computer labs

- The computer labs **should not** be expected to take only the 2h in the lab session
- The lab session should be seen as a teacher-assisted **introduction** to the assignment
- Budget significantly more time for completing the Lab assignments (10-20h) - they are a very important part of the learning process.
- Get additional assistance in the forum in Studium

*We will use a private cloud infrastructure for all the labs.*

UPPSALA UNIVERSITET

# Research paper seminar

- **S:** Literature Seminar
- In your project teams, you will read and discuss research papers and use them to answer discussion questions both in writing (hand-in) and in a seminar.
- More information about papers and seminar groups will follow later (after registration closes).

UPPSALA
UNIVERSITET

# Guest lecture

UPPSALA UNIVERSITET

# Project

- Completed in groups of 4 students.
- Assessed by a written course paper.
- The course paper is to be written on the format of a short scientific paper.
- Grading criteria and general advice for the report writing will be provided in the Studium.
- We will provide the group divisions in a few weeks time (when we know more precisely who will follow the course)

UPPSALA UNIVERSITET

# Grading criteria (summary)

- 3
  - Shows a basic understanding of key concepts
  - Can use key technologies to develop cloud software
- 4
  - Same as 3, and in addition
  - Shows a deepened understanding of key concepts
  - Can independently use key technologies to implement cloud software
- 5
  - Same as 4, and in addition
  - Can independently plan, analyze, implement and present software based on key technologies from the course.
  - Can critically evaluate key technologies w.r.t. a given application
  - Is acquainted of current research in data engineering

*More detailed document will appear in the Studium*

UPPSALA UNIVERSITET

# Grading based on points

| | Pt1 | Pt2 |
|---|---|---|
| C1 | 1 | 1 |
| C2 | 1 | 1 |
| C3 | 1 | 1 |
| C4. | 1 | 1 |
| S | 1 | |
| Mini-project: | 1 | 3 |

(Pass) 3: 6 of which at least 1 point on the mini-project

4: 9 of which at least 2 points on the mini-project

5: 12 of which at least 2 points on the mini-project

UPPSALA UNIVERSITET

# Programming language?

- We have not set a particular prerequisite programming language.
- We expect maturity when it comes to programming (i.e. you can quickly adopt to different languages and APIs).
- The assignments will mostly use Python.
- For the mini-project, you will be working largely independently, consuming the APIs you need to complete the project. Use whatever language you seem fit to get the job done.
- Even if your programming skills are rusty, you will be OK, but get prepared to work hard and don't save things to the last minute.

UPPSALA UNIVERSITET

# Operating system for the clients

- You can of course work on any computer system you want and are familiar with, but many things (ssh, scp, automation scrips…) will be much easier if you use Linux or OSX. In particular, we can offer very limited assistance if you choose to work on Windows.

- Virtual Machines will be based on Linux-flavors (mostly Ubuntu/Debian) — you will have to pick up a base level of Linux-admin skills (cmp.”DevOps”)

- If you have a Windows laptop, consider installing VirtualBox to run an Ubuntu VM locally.

UPPSALA UNIVERSITET

# **Important**: register in SUPR

For access to the the private cloud infrastructure SSC, you have to register a user in SUPR, the account management system of SNIC. It only takes a few minutes. Please fill in this online form: https://supr.snic.se/person/register/new/?

**Do this as soon as possible (i.e. today!), in good time before Lab 1.**

**Please state you university email address whenever possible (and check it). You have to accept the end-user agreement**

**Once registered in SUPR, apply for the project "____-"**

After the above steps have been completed, accounts will be created for you (you will receive an e-mail to the address you stated above)

UPPSALA UNIVERSITET

# Practical advice 1: Learn Python

- Python is being established as one of the major languages for data analysis
- Python is getting a wide-spread adoption in the CSE community
- **Python is a major language in cloud computing and Web programming**

**Resources:**

https://www.python.org/about/gettingstarted/

https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks#general-python-programming

UPPSALA UNIVERSITET

# Advice 2: Learn to use Git

- Using a version control system is **absolutely essential** for software development
- Git is widely used (more popular in new projects than SVN)
- Learn to work with the “Forking Workflow” model (ideal for Open Source projects)

https://www.atlassian.com/git/tutorials/comparing-workflows/gitflow-workflow

- A “real life” skill, essential part of a (software)engineers toolbox

UPPSALA UNIVERSITET

# Advice 2: Learn to use it now

- Easiest way to start is by using GitHub

  www.github.com

or BitBucket

  www.bitbucket.org

- The GitHub/BitBucket WebUIs add lots of nice tools for collaborative code development

**Importantly**: You are required to use a GitHub or BitBucket private repository for developing the software for your mini-project. If need be, we will provide code feedback and reviews directly in the respective WebUI. In GitHub, only public repositories are free. If you are not comfortable with displaying your code in public, BitBucket has free private repositories. *We will check commit-logs/Pull requests/code contributions.* Use the "ticket" system!

UPPSALA UNIVERSITET

# Virtual Machines are not for ever

- Another important reason to always use versioning for your assignment codes is that you will likely develop them on running VMs. VMs should be treated as volatile “software components”. Always assume that they can die/go away at any moment. Plan accordingly both in your software (resilience) and for your work — push anything important (that you can’t easily reproduce) out of the VM to a permanent location (for example a remote code repository)

- Automate things when possible so you can recreate after failure (assume that this will occur)

- You will most likely learn this the hard way during this course…