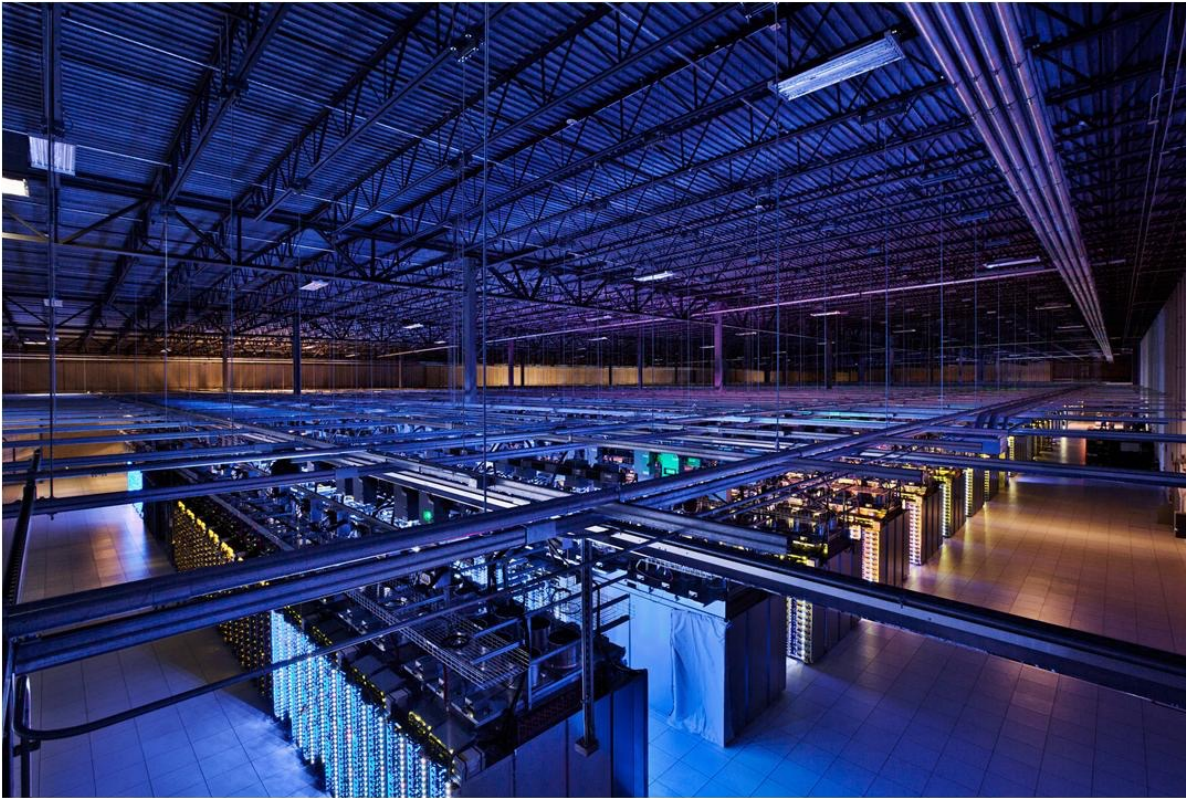


# Large-Scale Data Engineering

Intro to LSDE,

Intro to Big Data & Intro to Cloud Computing



# Administration

- Canvas Page
  - Announcements, also via email (pardon html formatting)
  - Turning in practicum assignments, Check Grades
- Contact: Slack

The screenshot shows a Canvas LMS course page for 'Large Scale Data Engineering' (course ID X\_405116). The page is viewed from a user's perspective, showing a navigation sidebar on the left with options like Account, Dashboard, Courses, Calendar, Inbox, Commons, and Help. The main content area includes a 'Recent announcements' section with a link to an event page, a 'Course status' section with 'Published' selected, and a 'Course objective' section with a photo of a server room. A table at the bottom right shows the course's weighted assignments.

Course status

Group	Weight
Assignments	0%
Assignment 1: Birthday Marketing Query	30%
Presentation for your data science project	20%
Assignment2: data science project	50%
<b>Total</b>	<b>100%</b>

# Goals & Scope

- The goal of the course is to gain **insight** into and **experience in** using hardware infrastructures and software technologies for **analyzing 'big data'**.

**This course delves into the practical/technical side of data science understanding and using large-scale data engineering to analyze big data**

# Goals & Scope

- The goal of the course is to gain **insight** into and **experience in** using hardware infrastructures and software technologies for **analyzing 'big data'**.

## Confronting you with the problems

### method: struggle with assignment 1

- Confronts you with some data management tasks, where
  - naïve solutions break down
  - problem size/complexity requires using a cluster
- Solving such tasks requires
  - insight in the main factors that underlie algorithm performance
    - access pattern, hardware latency/bandwidth
  - these factors guided the design of current Big Data infrastructures
    - helps understanding the challenges

# Goals & Scope

- The goal of the course is to gain **insight** into and **experience in** using hardware infrastructures and software technologies for **analyzing 'big data'**.

**Learn technical material about large-scale data engineering**  
**material: slides, scientific papers, videos**

- Understanding the concepts

hardware

- What components are hardware infrastructures made up of?
- What are the properties of these hardware components?
- What does it take to access such hardware?

software

- What software layers are used to handle Big Data?
- What are the principles behind this software?
- Which kind of software would one use for which data problem?

# Goals & Scope

- The goal of the course is to gain **insight** into and **experience in** using hardware infrastructures and software technologies for **analyzing 'big data'**.

**Obtain practical experience by doing a big data analysis project**

**method: do this in assignment 2**

**(code, report, 2 presentations, visualization website)**

- Analyze a large dataset for a particular question/challenge
- Use state-of-the-art tools in the cloud
  - (kind of) unlimited resources
  - Interactive notebooks, scheduled jobs
  - huge machines, many machines, GPUs, ...



# Your Tasks

- Interact in class, and in the slack channel (always)
- Start working on Assignment 1:
  - Register your github account on Canvas (now). Open one if needed.
  - 1a: Implement a 'query' program that solves a marketing query over a social network (deadline next week Monday night)
  - 1b: Implement a 'reorg' program to reduce the data and potentially store it in a more efficient form (deadline, one week after 1a).
- Read the papers in the reading list as the topics are covered (from next week on)
- Form practicum groups of three students (1c, deadline one week after 1b)
- Practice Spark on the Assignment1 query (in three weeks)
- Pick a unique project for Assignment 2 (in three weeks), FCFS in leaderboard order
  - Write a project plan & peer review others (in four weeks)
  - Conduct the project on Amazon Web Services
  - Give a 10min final presentation (in six weeks)
  - Submit code, Project Report and Visualization (deadline end of October)

# “Big Data”



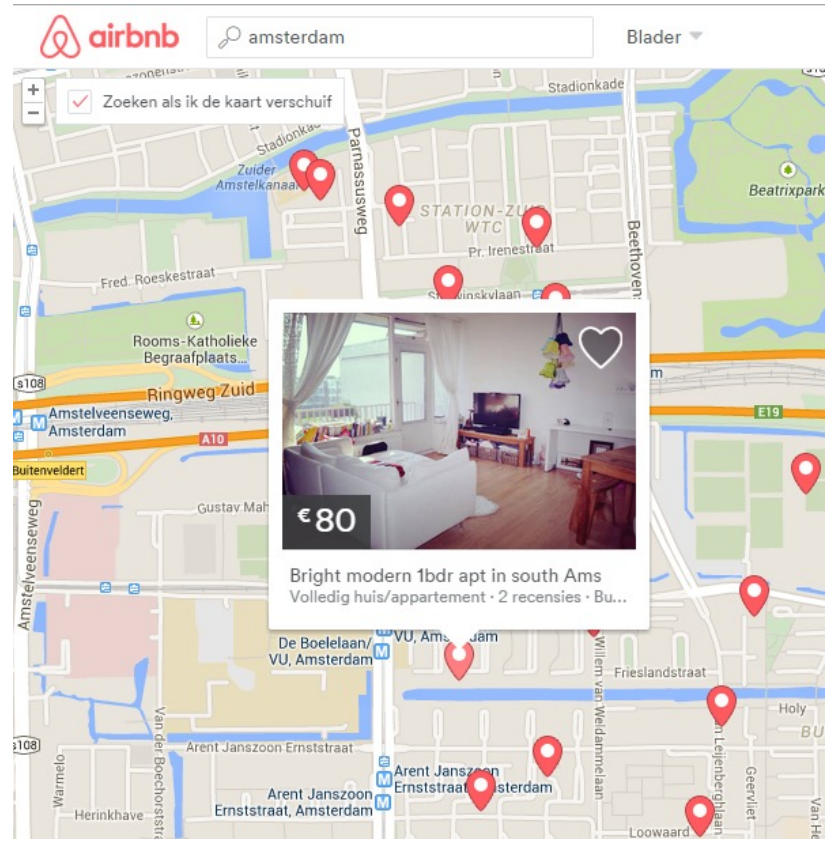
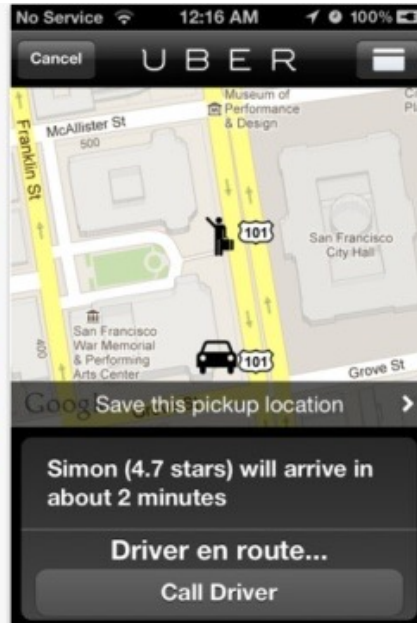
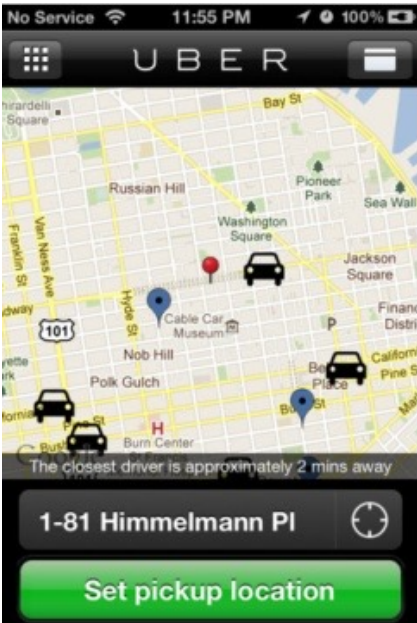


# Disruptions by the Data Economy



# U B E R

EVERYONE'S PRIVATE DRIVER™

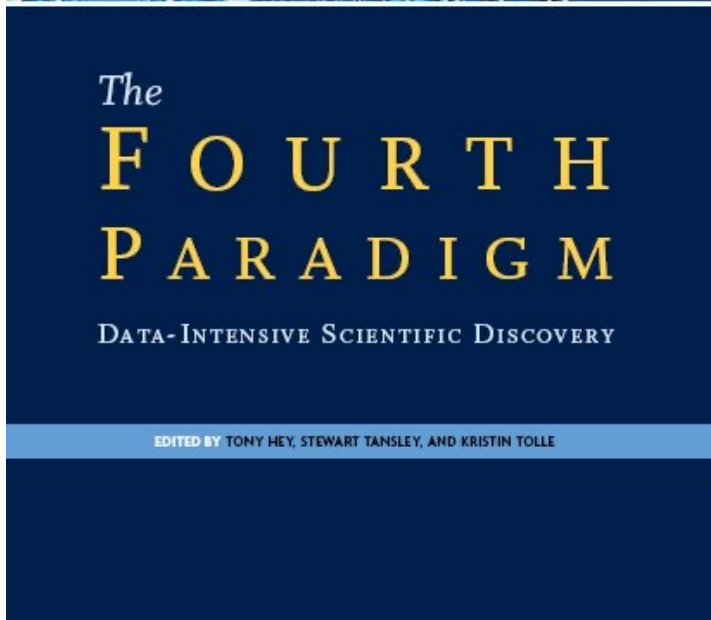


# Data Disrupting Science



Scientific paradigms:

1. Observing
2. Modeling
3. Simulating
4. **Collecting and Analyzing Data**



# Data Driven Science

## International LOFAR Telescope (ILT)



Chilbolton



Opening van de LOFAR telescoop door koningin Beatrix in juni 2010

Dutch stations

LOFAR Core (NL)

Jülich

Effelsberg

Unterweilenbach

Tautenburg

Potsdam

ASTRON

Netherlands Institute for Radio Astronomy

Onsala

Norderstedt

Baldy

raw data rate  
30GB/sec  
per station  
=  
1 full disk drive  
per second

# Big Data

- Big Data is a relative term
  - If things are breaking, you have Big Data
  - Big Data is not always Petabytes in size
  - Big Data for Informatics is not the same as for Google
- Big Data is often hard to understand
  - A model explaining it might be as complicated as the data itself
  - This has implications for Science
- The game may be the same, but the rules are completely different
  - What used to work needs to be reinvented in a different context



# Power laws



- This is **not** the 80/20 rule of skewed data (“80% of the sales are the 20% of the products”)
  - In a power law distribution, 80% of the data sales could well be in the “long tail” (the model is as large as the data).
- Modelling the head is easy, but may not be representative of the full population
  - Dealing with the full population might imply Big Data (e.g., selling all books, not just block busters)
- Processing Big Data might reveal power-laws
  - Most items take a small amount of time to process, individually
  - But there may be very many relevant items (“products”) to keep track of
  - A few items take a lot of time to process

# Big Data Challenges (1/3)

- **Volume** → data larger than a single machine (CPU, RAM, disk)
  - Infrastructures and techniques that scale by using more machines
  - Google led the way in mastering “cluster data processing”
- Velocity
- Variety

# Cloud Computing vs Cluster Computing

- Cluster Computing
  - Solving large tasks with more than one machine
    - Parallel database systems (e.g. Teradata, Vertica)
    - noSQL systems
    - Hadoop / MapReduce
- Cloud Computing



# Cloud Computing vs Cluster Computing

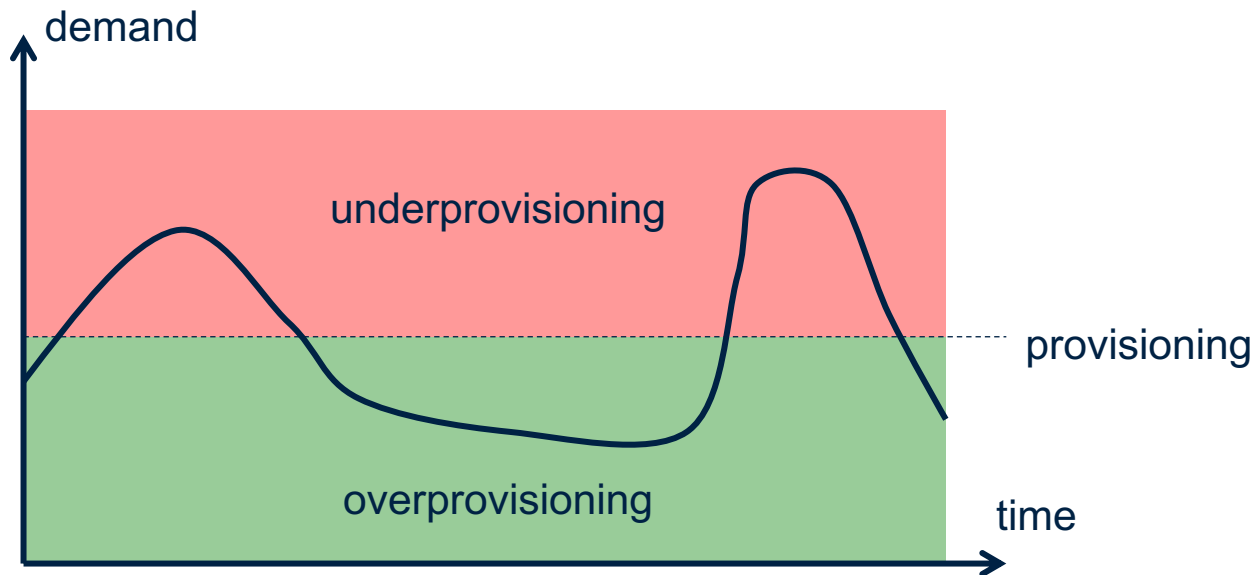
- Cluster Computing
- Cloud Computing
  - Machines operated by a third party in large data centers
    - sysadmin, electricity, backup, maintenance externalized
  - Rent access by the hour
    - Renting machines (Linux boxes): **Infrastructure as a Service**
    - Renting systems (Redshift SQL): **Platform-as-a-service**
    - Renting an software solution (Salesforce): **Software-as-a-service**
- {Cloud,Cluster} are independent concepts, but they are often combined!
  - We will do so in the practicum (Hadoop on Amazon Web Services)

# Economics of Cloud Computing

- A major argument for Cloud Computing is pricing:
  - We could own our machines
    - ... and pay for electricity, cooling, operators
    - ...and allocate enough capacity to deal with peak demand
  - Since machines rarely operate at more than 30% capacity, we are paying for wasted resources
- Pay-as-you-go rental model
  - Rent machine instances by the hour
  - Pay for storage by space/month
  - Pay for bandwidth by space/hour
- No other costs
- This makes computing a commodity
  - Just like other commodity services (sewage, electricity etc.)

# Cloud Computing: Provisioning

- We can quickly scale resources as demand dictates
  - High demand: more instances
  - Low demand: fewer instances
- Elastic provisioning is crucial



# Cloud Computing: some rough edges

- Some provider hosts our data
  - But we can only access it using proprietary (non-standard) APIs
  - **Lock-in** makes customers vulnerable to price increases and dependent upon the provider
  - Local laws (e.g. privacy) might prohibit externalizing data processing
- Providers may control our data in unexpected ways:
  - Amazon remotely remove books from Kindles
  - Twitter prevents exporting tweets more than 3200 posts back
  - Paying customers forced off Picasa towards Google Plus
  - Etc..
- Anti-terror laws mean that providers have to grant access to governments
  - This privilege can be overused

# Privacy and security

- People will not use Cloud Computing if trust is eroded
  - Who can access it?
    - Governments? Other people?
    - Snowden is the Chernobyl of Big Data
  - Privacy guarantees needs to be clearly stated and kept-to
- Privacy breaches
  - Numerous examples of Web mail accounts hacked
  - Many many cases of (UK) governmental data loss
  - TJX Companies Inc. (2007): 45 million credit and debit card numbers stolen
  - Every day there seems to be another instance of private data being leaked to the public

# Big Data Challenges (2/3)

- Volume
- **Velocity** → endless stream of new events
  - No time for heavy indexing (new data keeps arriving always)
  - led to development of data stream technologies
- Variety

# Big Streaming Data

- Storing it is not really a problem: disk space is cheap
- Efficiently accessing it and deriving results can be hard
- Visualising it can be next to impossible
- Repeated observations
  - What makes Big Data big are repeated observations
  - Mobile phones report their locations every 15 seconds
  - People post on Twitter > 100 million posts a day
  - The Web changes every day
  - Potentially we need unbounded resources
    - Repeated observations motivates streaming algorithms



# Big Data Challenges (3/3)

- Volume
- Velocity
- **Variety** → Dirty, incomplete, inconclusive data (e.g. text in tweets)
  - Semantic complications:
    - AI techniques needed, not just database queries
    - Data mining, Data cleaning, text analysis (AI techniques)
  - Technical complications:
    - Skewed Value Distributions and “Power Laws”
    - Complex Graph Structures → Expensive Random Access
    - Complicates cluster data processing (difficult to partition equally)
    - Localizing data by attaching pieces where you need them makes Big Data even bigger

# Big Data Challenges (3/3)

- Volume
- Velocity
- **Variety** → Dirty, incomplete, inconclusive data (e.g. text in tweets)
  - Semantic complications:
    - AI techniques needed, not just database queries
    - Data mining, Data cleaning, text analysis (AI techniques)
  - Technical complications:
    - Skewed Value Distributions and “Power Laws”
    - Complex Graph Structures → Expensive Random Access
    - Complicates cluster data processing (difficult to partition equally)
    - Localizing data by attaching pieces where you need them makes Big Data even bigger

# Summary

- Introduced the notion of Big Data, the three V's
  - Volume, Velocity, Variety
- Explained differences between Super/Cluster/Cloud computing

## Cloud computing:

- Computing as a commodity is likely to increase over time
- Cloud Computing adaptation and adoption are driven by economics
- The risks and obstacles behind it are complex
- Three levels:
  - Infrastructure as a service (run machines)
  - Platform as a service (use a database system on the cloud)
  - Software as a service (use software managed by other on the cloud)