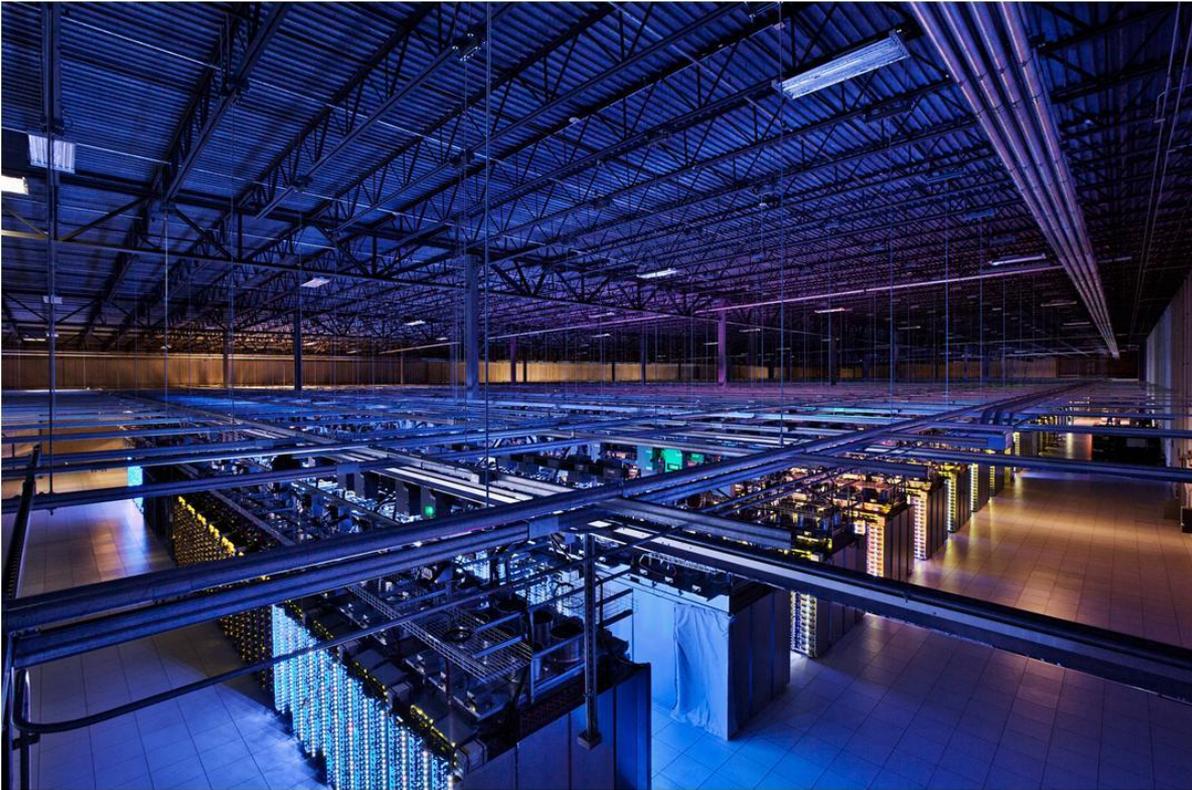# Large-Scale Data Engineering
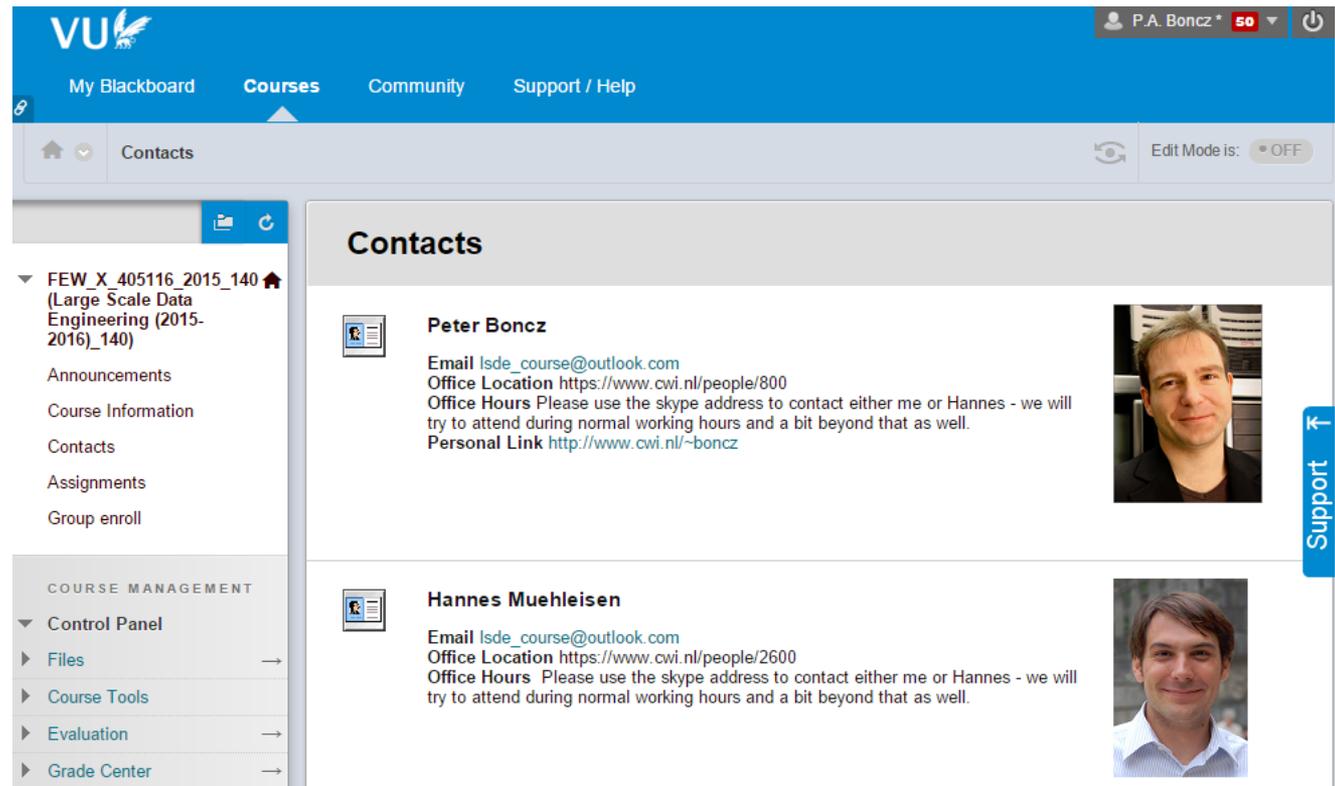
Intro to LSDE,
Intro to Big Data & Intro to Cloud Computing

Administration

- Blackboard Page
    - Announcements, also via email (pardon html formatting)
    - Turning in practicum assignments, Check Grades
- Contact: Email & Skype lsde_course@outlook.com

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 infratructures
    - 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, books, videos, magazine articles

- 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 (+talk +report)

- Analyze a large dataset for a particular question/challenge
- Use the SurfSARA Hadoop cluster (90 machines) and appropriate cluster software tools

Your Tasks

- Interact in class (always)

- Start working on Assignment 1 (now)
  - Form couples via Blackboard
  - Implement a 'query' program that solves a marketing query over a social network (and optionally also a 'reorg' program to store the data in a more efficient form).
  - Deadline within 2.5 weeks. Submit a *short* PDF report that explains what you implemented, experiments performed, and your final thoughts.

- Read the papers in the reading list as the topics are covered (from next week on)

- Pick a unique project for Assignment 2 (in 2.5 weeks)
  - 20min in-class presentation of your papers (last two weeks of lectures)
    - We can give presentation feedback beforehand (submit slides 24h earlier)
  - Conduct the project on a Hadoop Cluster (DAS-4 or SurfSARA)
    - write code, perform experiments
  - Submit a Project Report (deadline wk 13)
    - Related work (papers summary), Main Questions, Project Description, Project Results, Conclusion

The age of Big Data

1500TB/min =
1000 full drives
per minute

= a stack of
20meter high

4000 million
TeraBytes
=
3 billion
full disk drives
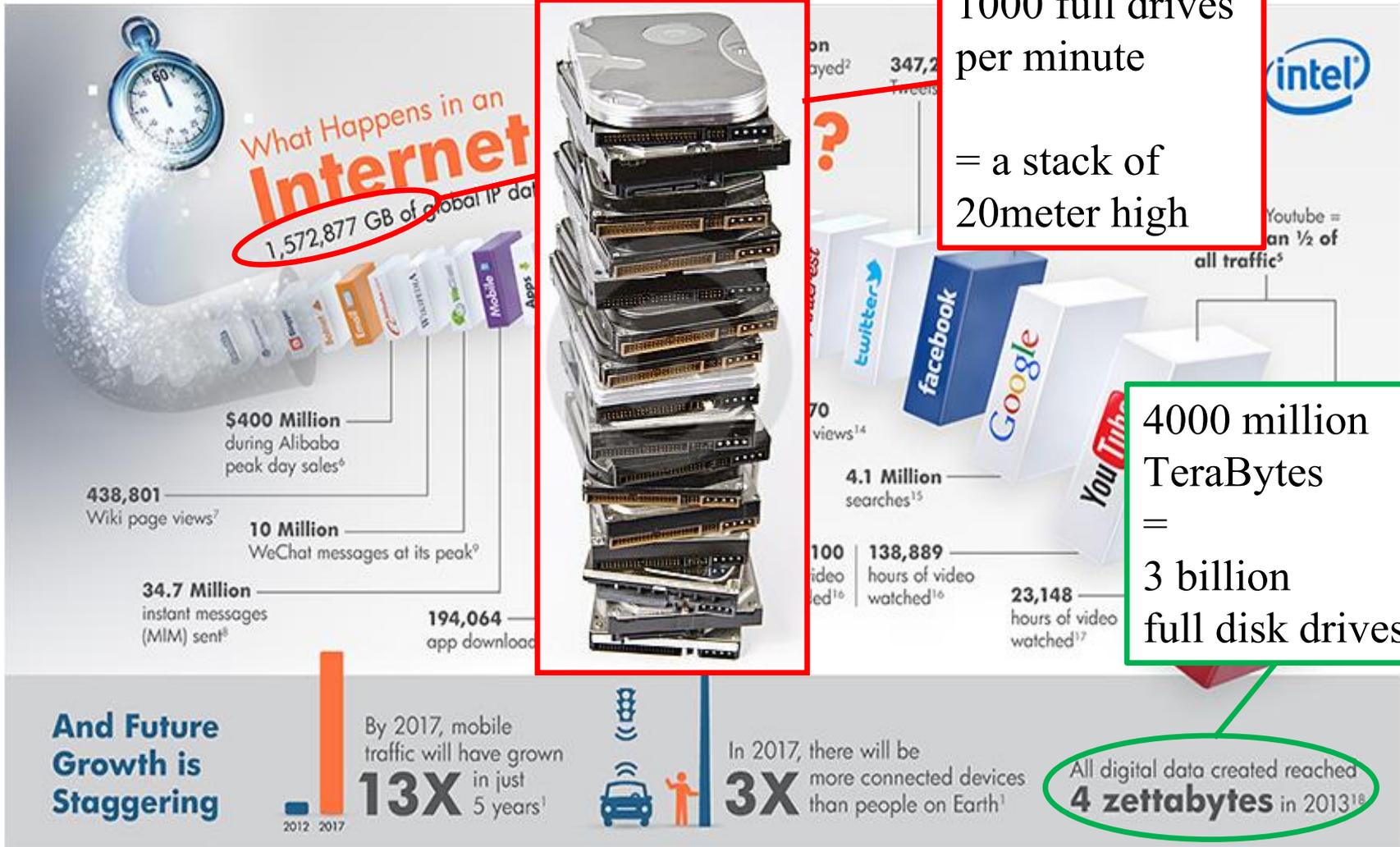
"Big Data"

The Data Economy
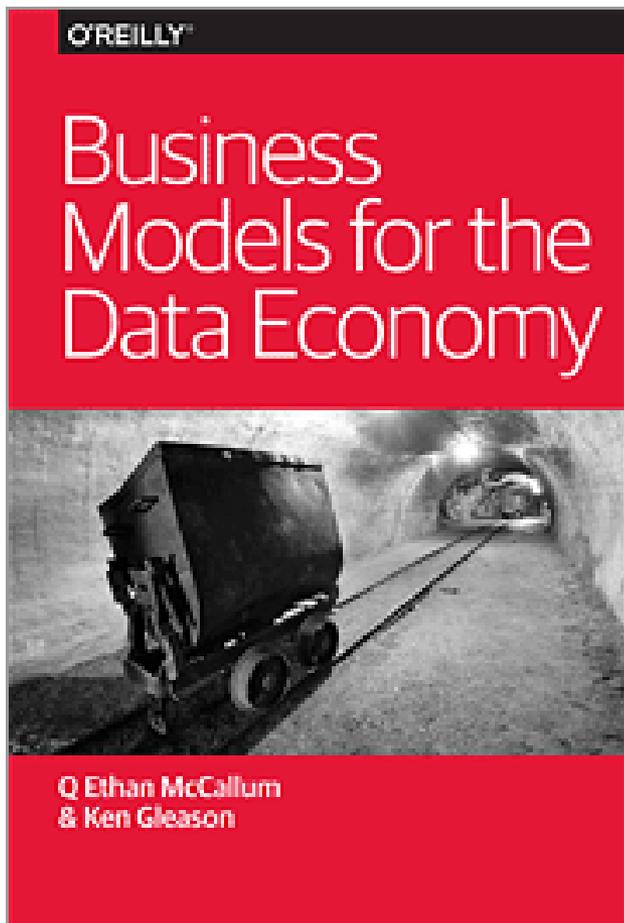
Disruptions by the Data Economy

UBER

EVERYONE'S PRIVATE DRIVER™
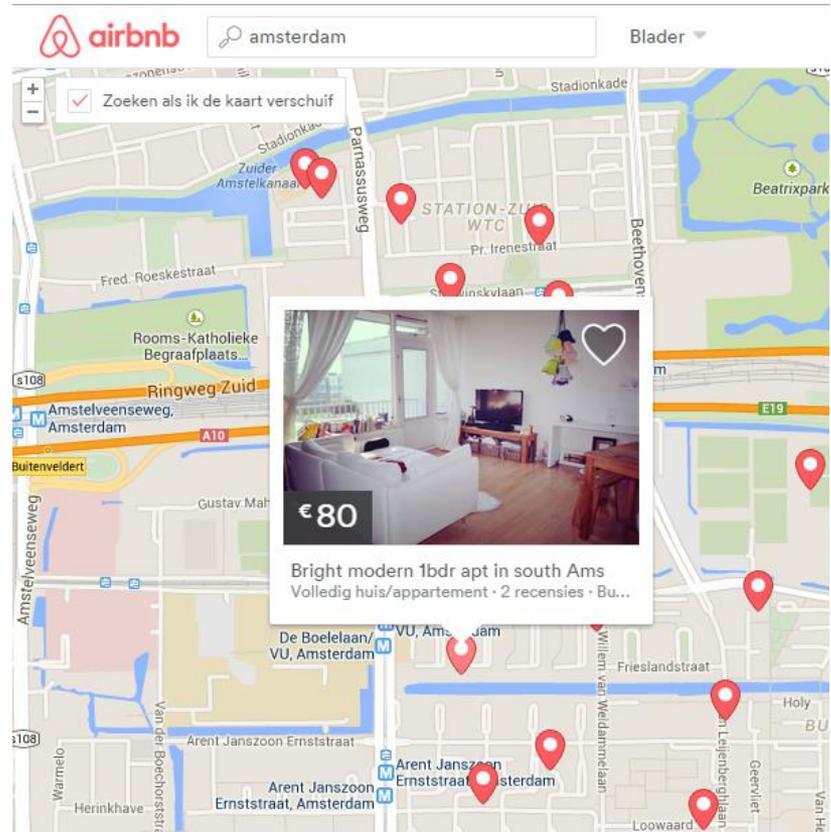
Data Disrupting Science
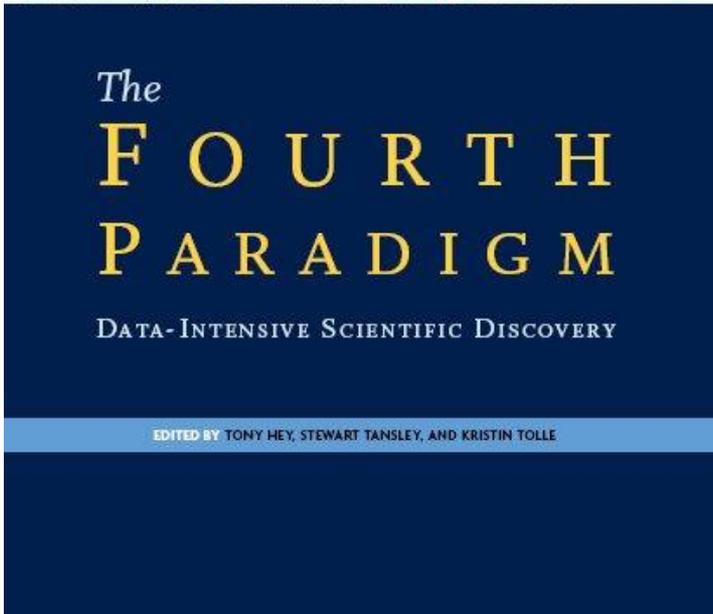
Scientific paradigms:

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

Data Driven Science

International LOFAR Telescope (ILT)

Onsala
Dutch stations
Norderstedt
Chilbolton
LOFAR Core (NL)
Bałdy
Potsdam
Jülich
Borówiec
Effelsberg
Tautenburg
Łazy
Unterweilenbach

Opening van de LOFAR telescoop door koningin Beatrix in juni 2010

ASTRON
Netherlands Institute for Radio Astronomy

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

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

Supercomputers?

- Take the top two supercomputers in the world today
    - Tiahne-2 (Guangzhou, China)
        - Cost: US$390 million
    - Titan (Oak Ridge National Laboratory, US)
        - Cost: US$97 million
- Assume an expected lifetime of five years and compute cost per hour
    - Tiahne-2: US$8,220
    - Titan: US$2,214
- This is just for the machine showing up at the door
    - Not factored in operational costs (e.g., running, maintenance, power, etc.)

Let's rent a supercomputer for an hour!

- Amazon Web Services charge US$1.60 per hour for a large instance
  - An 880 large instance cluster would cost US$1,408
  - Data costs US$0.15 per GB to upload
    - Assume we want to upload 1TB
    - This would cost US$153
  - The resulting setup would be #146 in the world's top-500 machines
  - Total cost: US$1,561 per hour
  - Search for (first hit): LINPACK 880 server

Supercomputing vs Cluster Computing

- Supercomputing
  - Focus on performance (biggest, fastest).. At any cost!
    - Oriented towards the [secret] government sector / scientific computing
  - Programming effort seems less relevant
    - Fortran + MPI: months do develop and **debug** programs
    - GPU, i.e. computing with graphics cards
    - FPGA, i.e. casting computation in hardware circuits
  - Assumes high-quality stable hardware
- Cluster Computing
  - use a network of many computers to create a 'supercomputer'
  - oriented towards business applications
  - use cheap servers (or even desktops), unreliable hardware
    - software must make the unreliable parts reliable
  - focus on economics (bang for the buck)
    - programming effort counts, a lot! No time to lose on debugging..

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

- Target (US retailer) uses Amazon Web Services (AWS) to host `target.com`
  - During massive spikes (November 28 2009 –"Black Friday") `target.com` is unavailable

- Remember your panic when Facebook was down?

event.cwi.nl/lsde
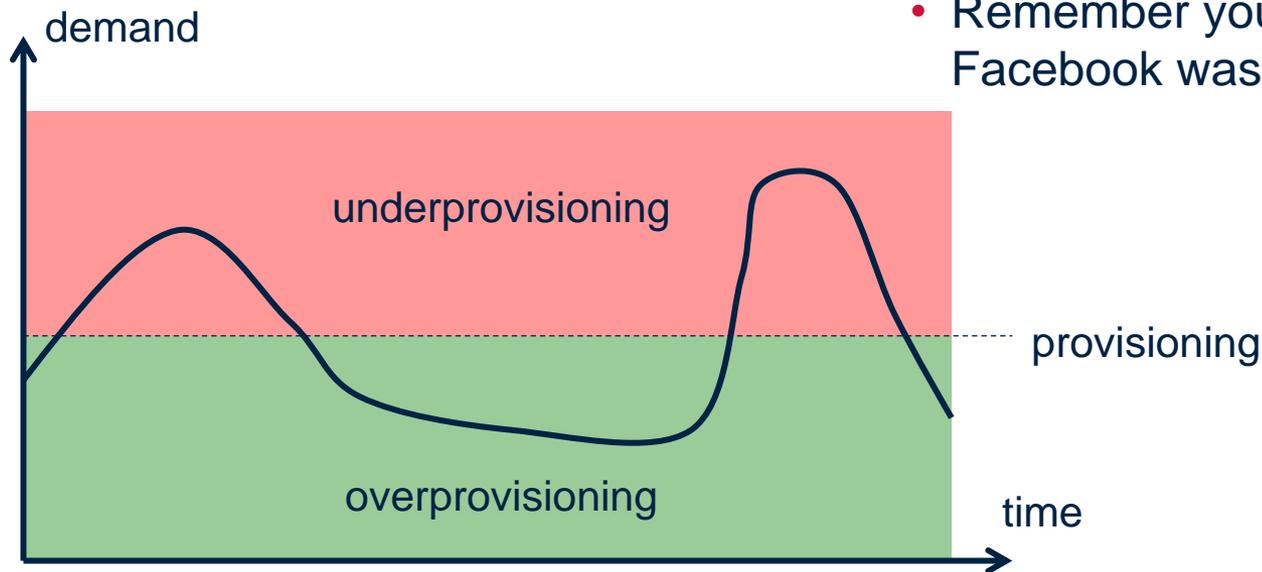
Cloud Computing: some rough edges

- 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:
    - July 2009: Amazon remotely remove books from Kindles
    - Twitter prevents exporting tweets more than 3200 posts back
    - Facebook locks user-data in
    - Paying customers forced off Picasa towards Google Plus
- 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

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

Skewed Data

- Distributed computation is a natural way to tackle Big Data
  - MapReduce encourages sequential, disk-based, localised processing of data
  - MapReduce operates over a cluster of machines
- One consequence of power laws is uneven allocation of data to nodes
  - The head might go to one or two nodes
  - The tail would spread over all other nodes
  - All workers on the tail would finish quickly.
  - The head workers would be a lot slower
- Power laws can turn parallel algorithms into sequential 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

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)