

# Big Data Analytics

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)  
Institute for Computer Science  
University of Hildesheim, Germany

# Outline

1. What is Big Data?
2. Overview
3. Organizational Stuff

# Outline

1. What is Big Data?

2. Overview

3. Organizational Stuff

# What is Big Data?



# What is Big Data?

*“Big data is like teenage sex:*

*— Dan Ariely*

# What is Big Data?

*“Big data is like teenage sex:  
everyone talks about it,*

*— Dan Ariely*

# What is Big Data?

*“Big data is like teenage sex:  
everyone talks about it,  
nobody really knows how to do it,*

*— Dan Ariely*

# What is Big Data?

*“Big data is like teenage sex:  
everyone talks about it,  
nobody really knows how to do it,  
everyone thinks everyone else is doing it,*

*— Dan Ariely*

# What is Big Data?

*“Big data is like teenage sex:  
everyone talks about it,  
nobody really knows how to do it,  
everyone thinks everyone else is doing it,  
so everyone claims they are doing it.”*

*— Dan Ariely*

# What is Big Data?

Some definitions:

- ▶ “data sets that are so **voluminous and complex** that traditional data processing application software are inadequate to deal with them.”

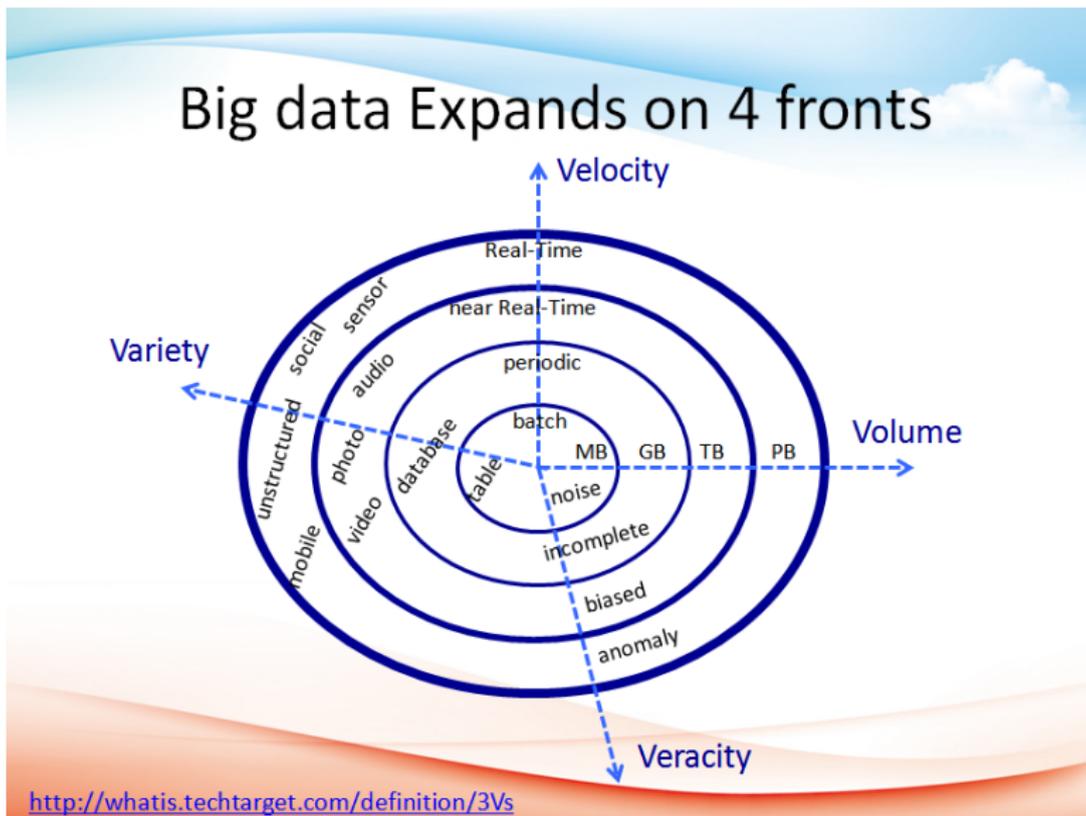
[[http://en.wikipedia.org/wiki/Big\\_data](http://en.wikipedia.org/wiki/Big_data)]

- ▶ “Big data is **high-volume, high-velocity** and/or **high-variety** information assets that demand cost-effective, innovative forms of information processing for **enhanced insight and decision making.**”

[[www.gartner.com/it-glossary/big-data/](http://www.gartner.com/it-glossary/big-data/)]

Note: The “3 Vs” go back to Laney [2001]. Often a 4th V “veracity” and a 5th V “value” is used.

# Big Data — Dimensions (the “4 Vs”)



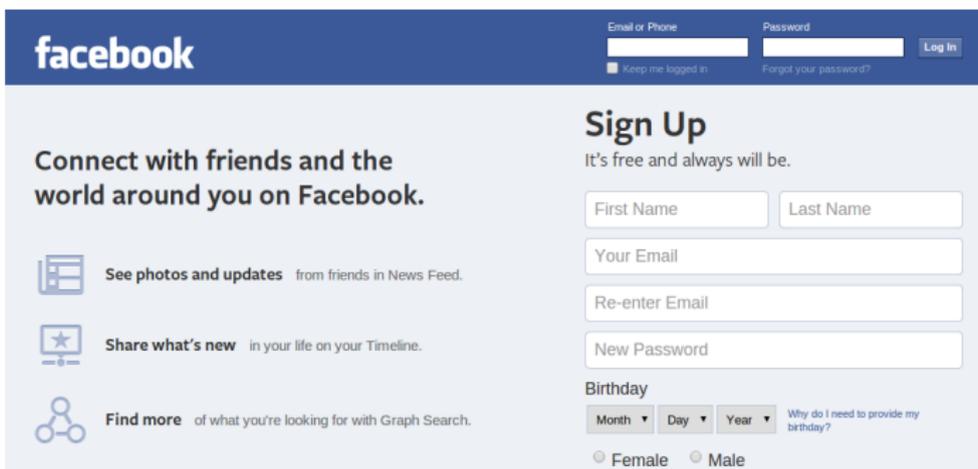
# What is Big Data?

Big Data is about:

- ▶ Storing and accessing
  - ▶ large amounts of
  - ▶ (complex/unstructured) data
- ▶ Processing high volume data streams
- ▶ Making sense of the data
- ▶ Making predictions based on the data

Note: Unstructured data in this context means, data that is not already a vector. Some of this data, e.g., relational data and graph data, confusingly often also is called “structured”.

# Where to find Big Data?



The image shows the Facebook login and sign-up interface. At the top, there is a blue header with the Facebook logo on the left and login fields on the right. The login fields include 'Email or Phone' and 'Password', both with input boxes, and a 'Log In' button. Below the login fields are two checkboxes: 'Keep me logged in' and 'Forgot your password?'. The main content area is divided into two sections. On the left, there is a heading 'Connect with friends and the world around you on Facebook.' followed by three icons and their corresponding text: 'See photos and updates from friends in News Feed.', 'Share what's new in your life on your Timeline.', and 'Find more of what you're looking for with Graph Search.'. On the right, there is a 'Sign Up' section with the text 'It's free and always will be.' followed by four input fields: 'First Name', 'Last Name', 'Your Email', and 'Re-enter Email'. Below these is a 'New Password' field. At the bottom of the sign-up section, there is a 'Birthday' section with three dropdown menus for 'Month', 'Day', and 'Year', and a link 'Why do I need to provide my birthday?'. At the very bottom of the sign-up section are two radio buttons for 'Female' and 'Male'.

- ▶ 1.52 billion daily active users  
(2.32 billion monthly active users, Dec. 2018)
- ▶ size of user data stored by Facebook: 300 Petabytes
- ▶ average amount of data that Facebook takes in daily: 600 terabytes
- ▶ size of Facebook's graph search database: 700 Terabytes

[source: <https://newsroom.fb.com/company-info/>; online source for points 2-4 vanished]

# Where to find Big Data?

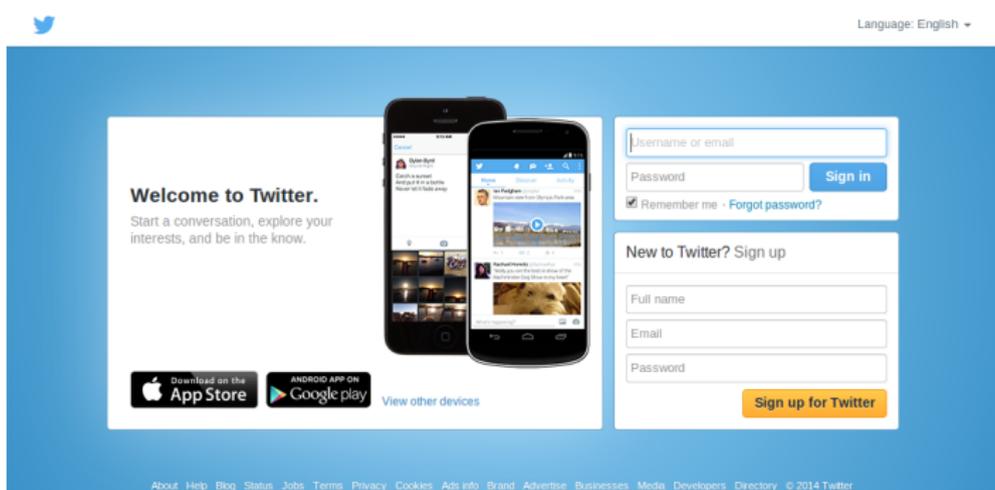


- ▶ 3.3 billion searches per day (on average)<sup>1</sup>
- ▶ 30 trillion unique URLs identified on the Web<sup>1</sup>
- ▶ 20 billion sites crawled a day<sup>1</sup>
- ▶ In 2008 Google processed more than 20 Petabytes of data per day<sup>2</sup>

<sup>1</sup><http://searchengineland.com/google-search-press-129925> (2012)

<sup>2</sup>Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113.

# Where to find Big Data?

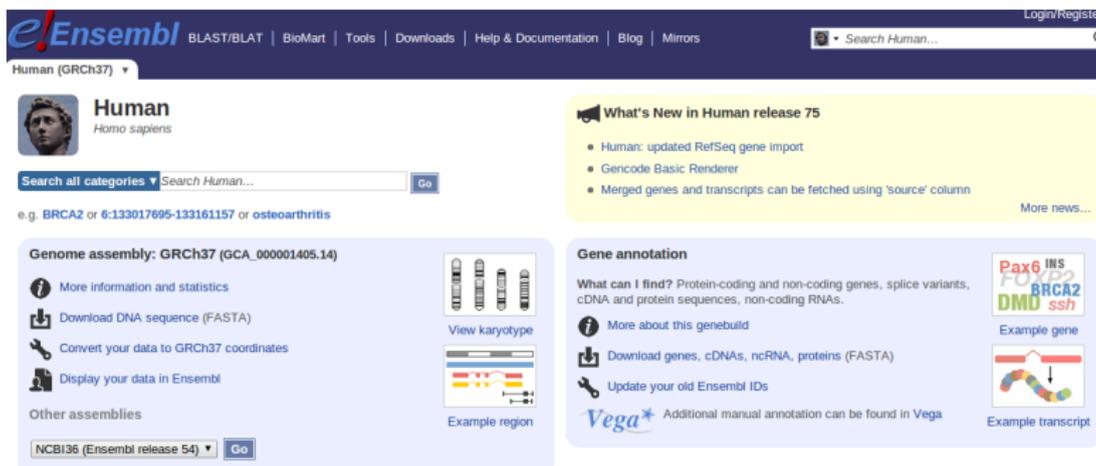


The image shows the Twitter sign-up page. On the left, there is a 'Welcome to Twitter' message with a description: 'Start a conversation, explore your interests, and be in the know.' Below this are buttons for 'Download on the App Store' and 'ANDROID APP ON Google play', along with a link to 'View other devices'. In the center, two smartphones display the Twitter mobile app interface. On the right, there is a sign-in form with fields for 'Username or email' and 'Password', a 'Remember me' checkbox, and a 'Forgot password?' link. Below the sign-in form is a 'New to Twitter? Sign up' section with fields for 'Full name', 'Email', and 'Password', and a 'Sign up for Twitter' button. At the top right, there is a language selector set to 'English'. At the bottom, there is a footer with various links like 'About', 'Help', 'Blog', etc., and a copyright notice for 2014 Twitter.

- ▶ tweets per day: 58 million<sup>1</sup>
- ▶ Twitter search engine queries per day: 2.1 billion<sup>1</sup>
- ▶ registered/active Twitter users: 695 million / 342 million<sup>1</sup>

[<sup>1</sup><http://www.statisticbrain.com/twitter-statistics/>] (9/2016)

# Where to find Big Data?



Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Human (GRCh37)

Human  
*Homo sapiens*

Search all categories Search Human... Go

e.g. BRCA2 or 6:133017695-133161157 or osteoarthritis

**Genome assembly: GRCh37 (GCA\_000001405.14)**

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh37 coordinates
- Display your data in Ensembl

Other assemblies

NCBI36 (Ensembl release 54) Go

View karyotype

Example region

**Gene annotation**

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download genes, cDNAs, ncRNA, proteins (FASTA)
- Update your old Ensembl IDs

Vega Additional manual annotation can be found in Vega

Example transcript

What's New in Human release 75

- Human: updated RefSeq gene import
- Gencode Basic Renderer
- Merged genes and transcripts can be fetched using 'source' column

More news...

- ▶ Ensembl database contains the genome of humans and 50 other species
- ▶ “only” 250 GB<sup>1</sup>

[<sup>1</sup><http://www.ensembl.org/>]

# Where to find Big Data?



- ▶ CERN Large Hadron Collider has collected data from over 300 trillion proton-proton collisions
- ▶ Approx. 25 Petabytes per year

# Big Data — Public Datasets

1000 Genomes Project	DNA of 1700 humans	200 TB
Common Crawl Corpus	5G web pages	81 TB
Wikipedia / Freebase	1.9G subject/predicate/object triples	250 GB
Million Song Dataset	audio features of 1M songs	280 GB
OpenStreetMap	a map of earth	90 GB
2000 US Census	US census data	200 GB
PubChem library	biological activities of small molecules	230 GB
NCDC weather data	daily measurements from 9000 stations	20 GB
Open Library	metadata of 20M books	7 GB
Twitter	1.6G tweets	0.6 GB

CD 700 MB, DVD 4.7–17 GB, Blu-ray 25–100 GB, hard disc: 10 TB.

# How Large is 1 Petabyte

- ▶ 1 PB = 1000 TB =  $10^{15}$  B

# How Large is 1 Petabyte

- ▶ 1 PB = 1000 TB =  $10^{15}$  B
- ▶ 35.7M years counted one byte per second,

# How Large is 1 Petabyte

- ▶ 1 PB = 1000 TB =  $10^{15}$  B
- ▶ 35.7M years counted one byte per second,
- ▶ 254 years listening to music stored in CD quality (500MB/h)

# How Large is 1 Petabyte

- ▶ 1 PB = 1000 TB =  $10^{15}$  B
- ▶ 35.7M years counted one byte per second,
- ▶ 254 years listening to music stored in CD quality (500MB/h)
- ▶ 25 years watching DVDs
  - ▶ 223.000 DVDs (a 4.7 GB)

# How Large is 1 Petabyte

- ▶ 1 PB = 1000 TB =  $10^{15}$  B
- ▶ 35.7M years counted one byte per second,
- ▶ 254 years listening to music stored in CD quality (500MB/h)
- ▶ 25 years watching DVDs
  - ▶ 223.000 DVDs (a 4.7 GB)
  - ▶ but there are only 74.000 TV movies on IMDB !  
(1.8M including TV episodes)

# How Large is 1 Petabyte

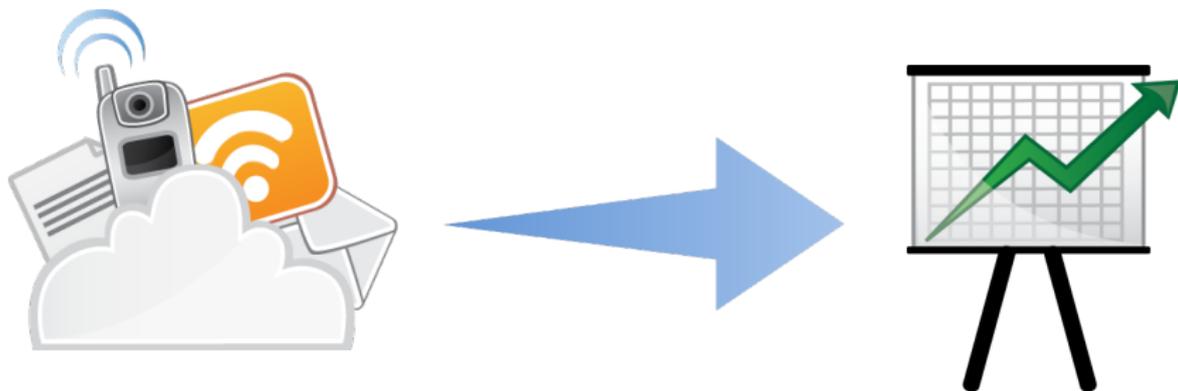
- ▶ 1 PB = 1000 TB =  $10^{15}$  B
- ▶ 35.7M years counted one byte per second,
- ▶ 254 years listening to music stored in CD quality (500MB/h)
- ▶ 25 years watching DVDs
  - ▶ 223.000 DVDs (a 4.7 GB)
  - ▶ but there are only 74.000 TV movies on IMDB !  
(1.8M including TV episodes)
- ▶ can be stored on 100 harddisks à 10 TB/300 € (30,000 €)

# How Large is 1 Petabyte

- ▶ 1 PB = 1000 TB =  $10^{15}$  B
- ▶ 35.7M years counted one byte per second,
- ▶ 254 years listening to music stored in CD quality (500MB/h)
- ▶ 25 years watching DVDs
  - ▶ 223.000 DVDs (a 4.7 GB)
  - ▶ but there are only 74.000 TV movies on IMDB !  
(1.8M including TV episodes)
- ▶ can be stored on 100 harddisks à 10 TB/300 € (30,000 €)
- ▶ 96 days to read from standard harddisks sequentially (1030 MBits/s)

# What to do with Big Data?

We do not want to know things but to understand them!



# What to do with Big Data? Leasing Vehicle Return @VWFS

initial data generating process:



expert



240 €

target process:



customer



automatic  
service



240 €

# What to do with Big Data? - Case Studies

## ▶ **T-Mobile USA:**

- ▶ Integrated Big Data across multiple IT systems to combine customer transaction and interaction data in order to better predict customer defections
- ▶ By leveraging social media data along with transaction data from CRM and billing systems, customer defections have been cut in half in a single quarter.

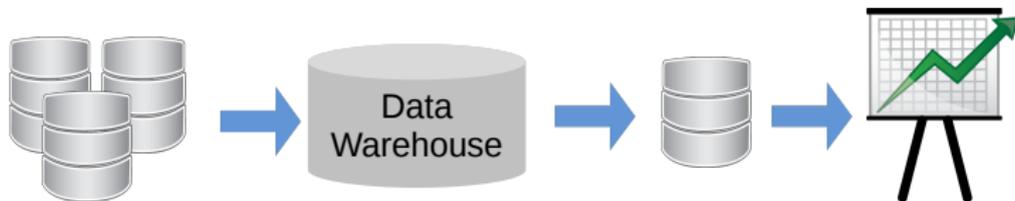
## ▶ **US Xpress:**

- ▶ Collects data elements ranging from fuel usage to tire condition to truck engine operations to GPS information
- ▶ Optimal fleet management

## ▶ **McLaren's Formula One racing team:**

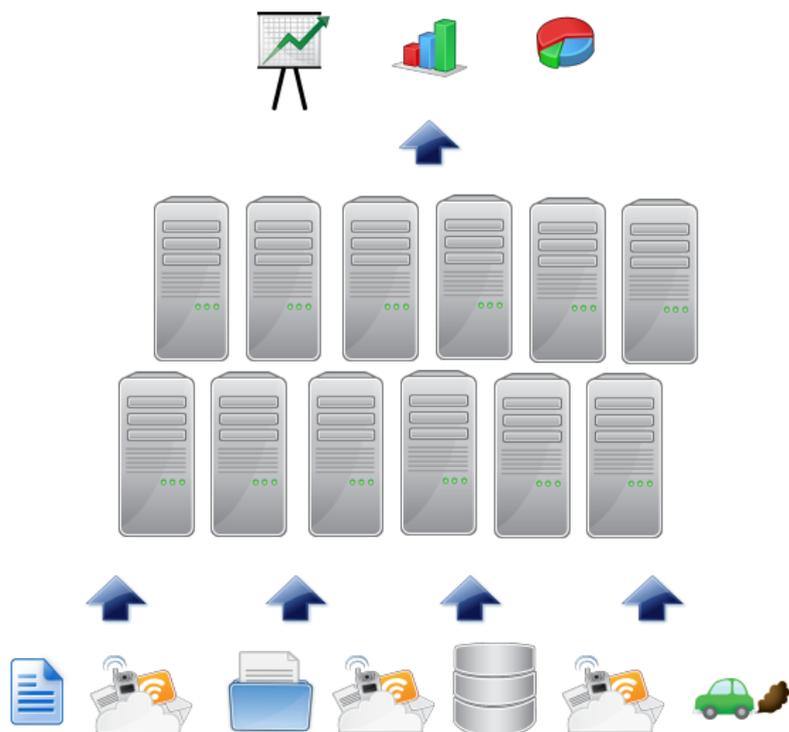
- ▶ Real-time car sensor data during car races
- ▶ Real-time identification of issues with its racing cars

# What to do with Big Data? - The BI Approach



- ▶ Static databases
- ▶ Structured data
- ▶ Centralized approaches

# What to do with Big Data?



- ▶ Heterogeneous data sources
- ▶ Unstructured data
- ▶ Data streams
- ▶ Massive Parallelism

# What to do with Big Data?

## Application examples:

- ▶ Online personalized advertising
- ▶ Sentiment analysis and behavior prediction
- ▶ Detecting adverse events and predicting their impact
- ▶ Automatic Translation
- ▶ Image Classification and object recognition
- ▶ Intelligent public services

# Challenges Posed By Big Data

- ▶ Effectively **store and query** large amounts of data in a distributed environment
- ▶ **Parallel and distributed execution environments** / programming models
- ▶ **Distributed and scalable machine learning techniques** to learn from the data
- ▶ **Distributed and scalable data visualization techniques**

# Outline

1. What is Big Data?
- 2. Overview**
3. Organizational Stuff

# Technology Stack

*Part D*

**Distributed Machine  
Learning Algorithms**

*Part C*

**Distributed Execution Environments**

*Part B*

**Distributed Storage**

*Part A*

**Parallel/Distributed Computing**

# Technology Stack

*Part D*

**Distributed Machine Learning Algorithms**

*Part C*

**Distributed Execution Environments**

*Part B*

**Distributed Storage**

*Part A*

**Parallel/Distributed Computing**

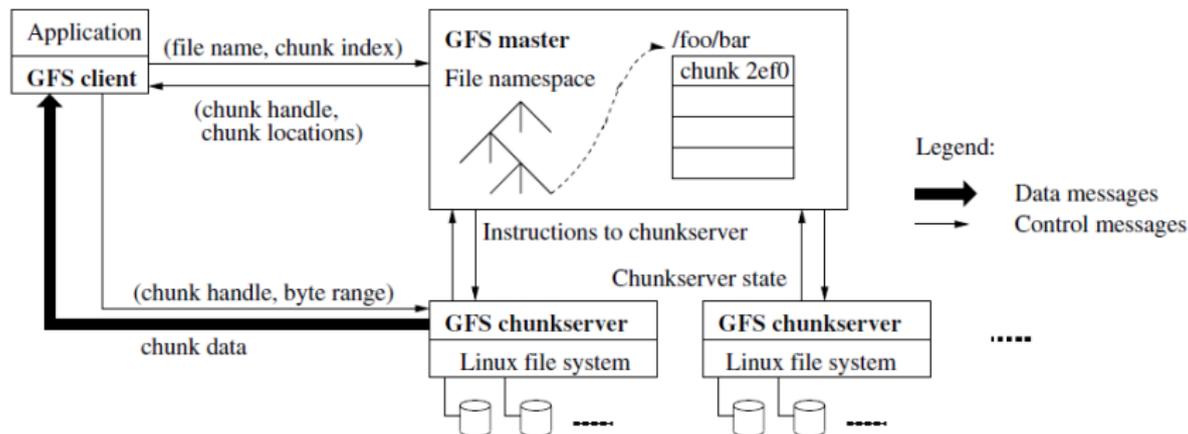
# Storing

In a distributed environment the data storing mechanisms should address the following issues

- ▶ Parallel Reading and Writing
- ▶ Data node Failures
- ▶ High Availability

# Distributed File Systems

## The Google File System Architecture



# Databases

Databases are needed for

- ▶ Querying and indexing
- ▶ transaction processing

## **State-of-the-art:** Relational Databases

For processing big data one needs a database which:

- ▶ Supports high level of parallelism
- ▶ Supports analytical processing
- ▶ Has a flexible data model to deal with unstructured data sources

# Databases for Big Data - NoSQL

## NoSQL - “Not only SQL”

- ▶ Wide variety of database technologies
- ▶ Dynamic Schema
- ▶ sharded indexing
- ▶ horizontal scaling
- ▶ support columnar storage

# NoSQL Databases

## key-value

Amazon  
DynamoDB (Beta)

ORACLE  
BERKELEY DB 11g

 redis

## graph

 Neo4j  
the graph database

 InfiniteGraph

 sones

## column

 HBASE

 riak

 Cassandra

## document

 CouchDB  
relax

 mongoDB

 terrastore

# Technology Stack

*Part D*

**Distributed Machine Learning Algorithms**

*Part C*

**Distributed Execution Environments**

*Part B*

**Distributed Storage**

*Part A*

**Parallel/Distributed Computing**

# Accessing

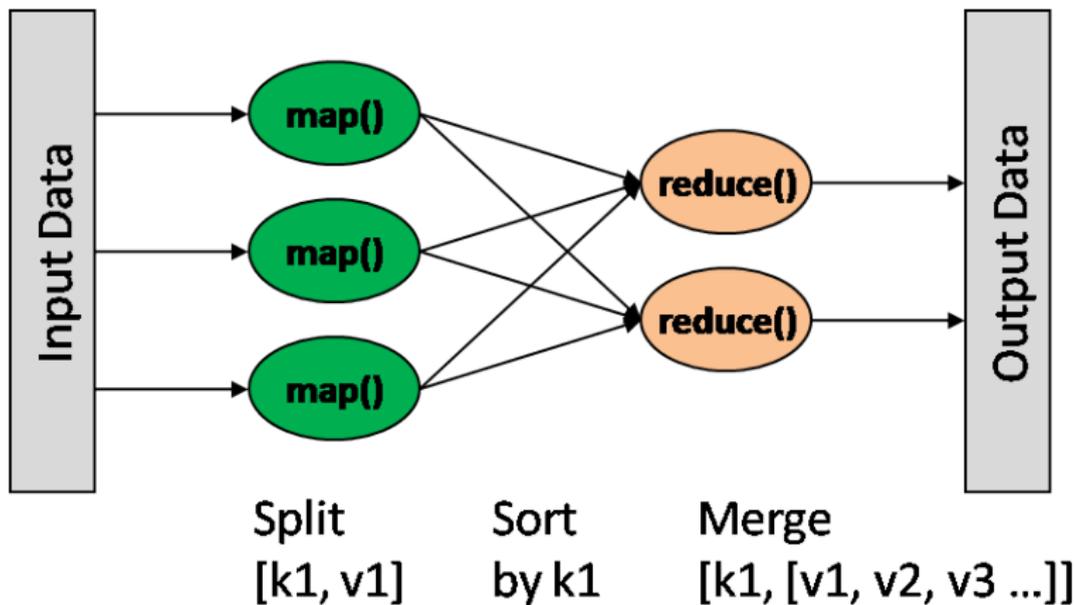
A **distributed execution environment** / **computational model** is needed to:

- ▶ Provide a set of useful computational primitives
- ▶ Hide the complexity of distributed and parallel programming
- ▶ Ensure Fault Tolerance

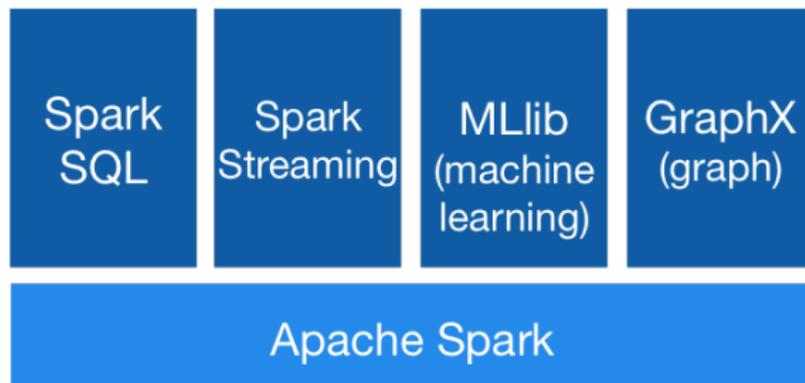
Examples:

- ▶ MapReduce
- ▶ GraphLab
- ▶ Pregel
- ▶ Apache Spark
- ▶ TensorFlow

# MapReduce



# Apache Spark



# Technology Stack

*Part D*

**Distributed Machine Learning Algorithms**

*Part C*

**Distributed Execution Environments**

*Part B*

**Distributed Storage**

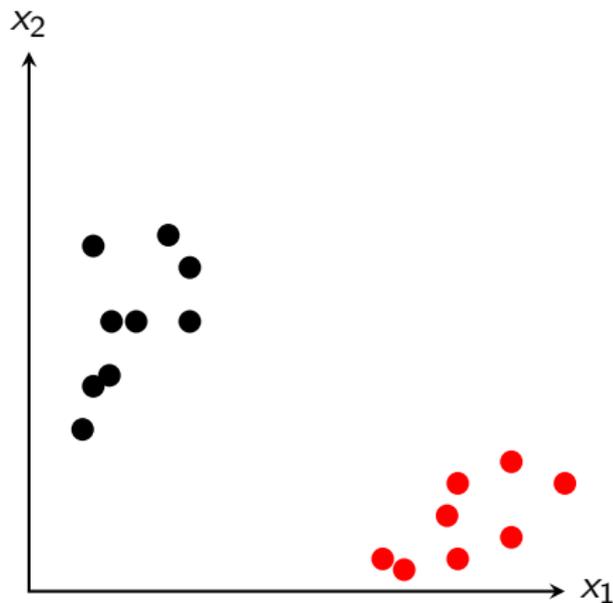
*Part A*

**Parallel/Distributed Computing**

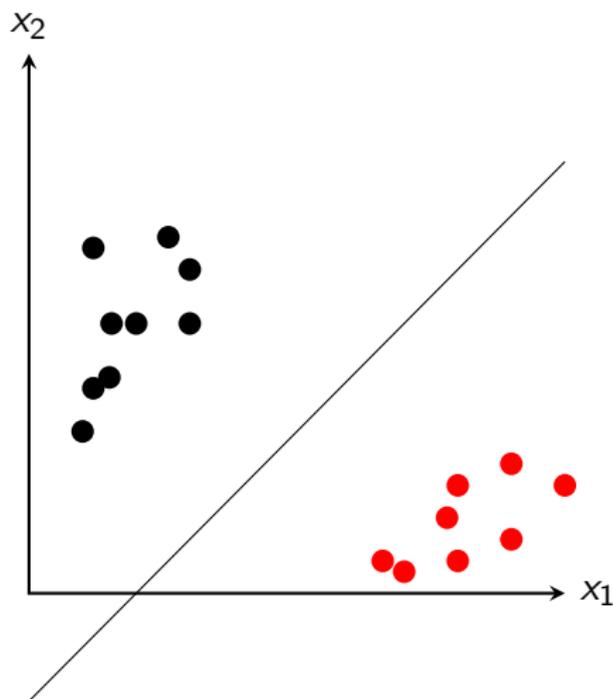
# Learning from the data

- ▶ Linear and Non Linear Models for classification and regression
  - ▶ Scalable learning algorithms (e.g. Stochastic Gradient Descent)
  - ▶ Distributed Learning Algorithms (e.g. ADMM)
  
- ▶ Models for Link Prediction and link analysis
  - ▶ Factorization models
  - ▶ Distributed Learning Schemes (e.g. NOMAD, FPSGD)

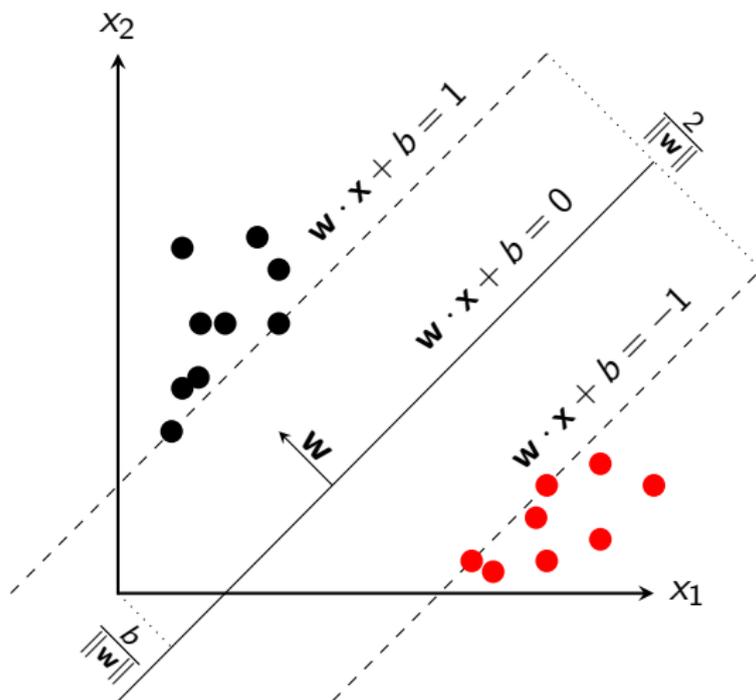
# Classification



# Classification



# Classification



# Recommender Systems

## Heutige Empfehlungen für Sie

Hier sind einige der Ihnen empfohlenen Artikel. Klicken Sie hier, um [alle Empfehlungen anzuzeigen](#).

Seite 1 von 10



**The Mentalist - Die komplette...** DVD  
- Simon Baker  
★★★★☆ (28) EUR 17,98  
[Diese Empfehlung korrigieren](#)



**Two and a Half Men: Mein coo...** DVD  
- Charlie Sheen  
★★★★☆ (105) EUR 9,95  
[Diese Empfehlung korrigieren](#)



**Monk - 1. Staffel (4 DVDs)** DVD -  
Tony Shalhoub  
★★★★☆ (34) EUR 13,98  
[Diese Empfehlung korrigieren](#)



**Bones - Season 1 (6 DVDs)** DVD -  
David Boreanaz  
★★★★☆ (48) EUR 20,11  
[Diese Empfehlung korrigieren](#)



**Dr. House - Season 1, 2. Episod...**  
DVD - Hugh Laurie  
★★★★☆ (1) EUR 12,99  
[Diese Empfehlung korrigieren](#)




[Bros](#)

You're using an experimental YouTube homepage. [Feedback?](#)

[Back to classic homepage](#)



Idrumond

2

All activity

Subscription uploads

### Recommended Videos (6 hours ago)



**SHAMAN NEW LIVE**  
DVD

68,869 views



**Egberto Gismonti - O**  
Sonho

16,033 views



**Edu Ardanuy -**  
Improviso 1

19,985 views



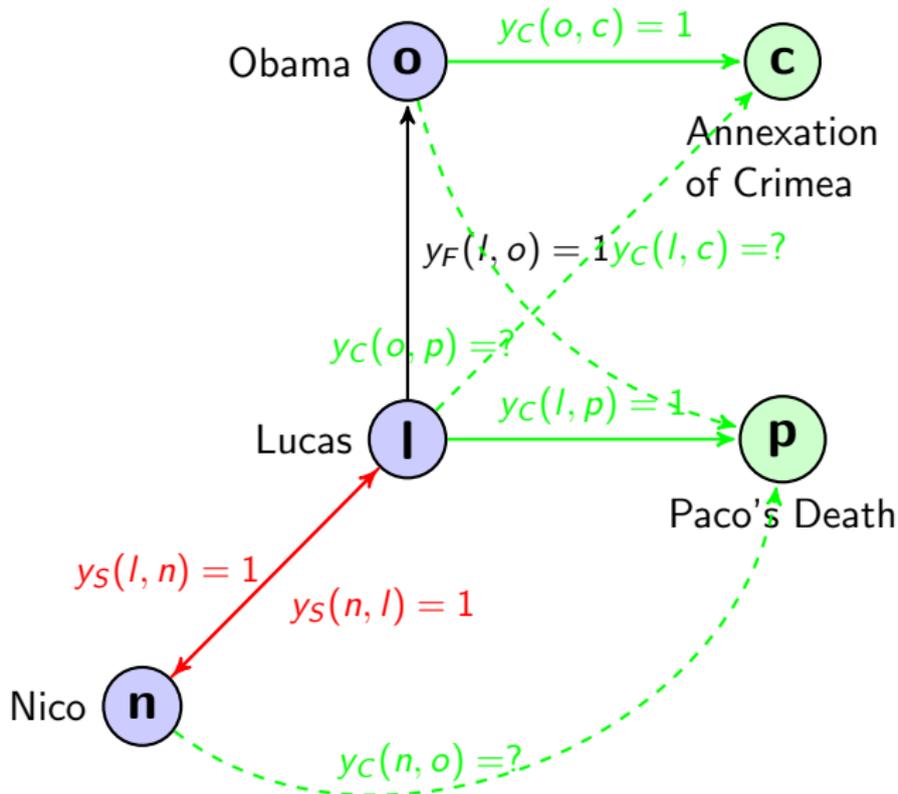
**Egberto Gismonti 03 -**  
Bachianas ...

9,517 views

[See More](#)



# Graph Analysis



# Syllabus

- |            |      |   |
|------------|------|---|
| Tue. 9.4.  | (1)  | 0. Introduction                                   |
|            |      | <b>A. Parallel Computing</b>                      |
| Tue. 16.4. | (2)  | A.1 Threads                                       |
| Tue. 23.4. | (3)  | A.2 Message Passing Interface (MPI)               |
| Tue. 30.4. | (4)  | A.3 Graphical Processing Units (GPUs)             |
|            |      | <b>B. Distributed Storage</b>                     |
| Tue. 7.5.  | (5)  | B.1 Distributed File Systems                      |
| Tue. 14.5. | (6)  | B.2 Partitioning of Relational Databases          |
| Tue. 21.5. | (7)  | B.3 NoSQL Databases                               |
|            |      | <b>C. Distributed Computing Environments</b>      |
| Tue. 28.5. | (8)  | C.1 Map-Reduce                                    |
| Tue. 4.6.  | —    | — <i>Pentecoste Break</i> —                       |
| Tue. 11.6. | (9)  | C.2 Resilient Distributed Datasets (Spark)        |
| Tue. 18.6. | (10) | C.3 Computational Graphs (TensorFlow)             |
|            |      | <b>D. Distributed Machine Learning Algorithms</b> |
| Tue. 25.6. | (11) | D.1 Distributed Stochastic Gradient Descent       |
| Tue. 2.7.  | (12) | D.2 Distributed Matrix Factorization              |
| Tue. 9.7.  | (13) | Questions and Answers                             |

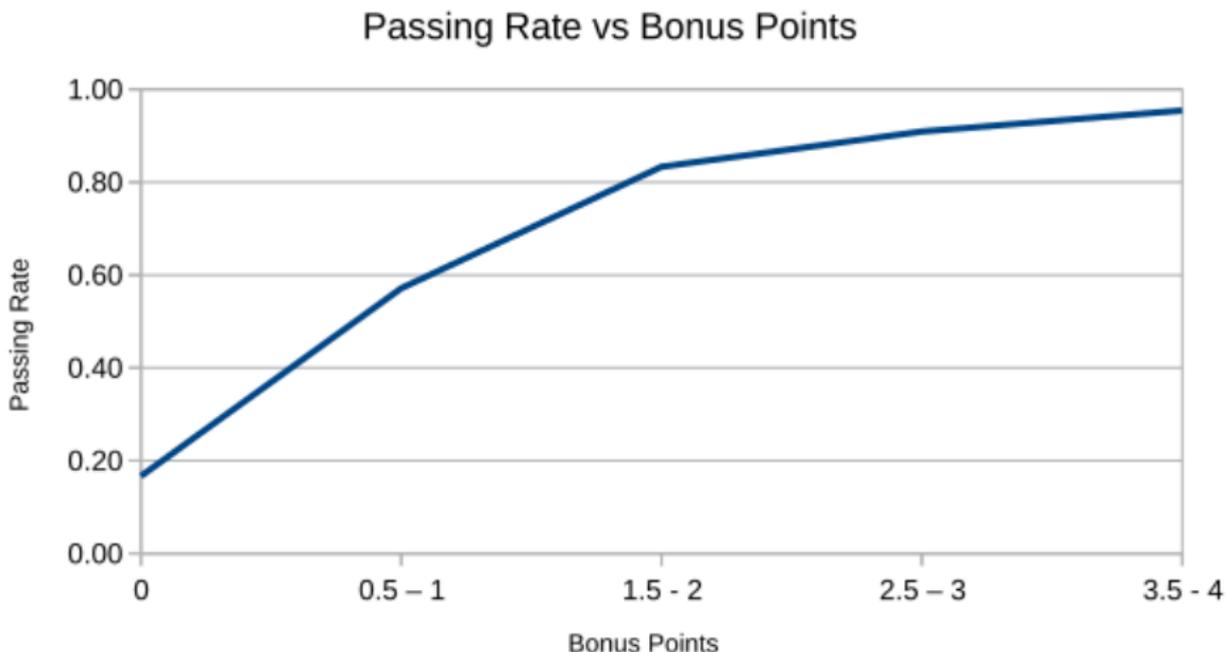
# Outline

1. What is Big Data?
2. Overview
3. Organizational Stuff

# Exercises and tutorials

- ▶ There will be a weekly sheet with two exercises handed out **each Tuesday, 12pm** after the lecture.
  - ▶ 1st sheet will be handed out Thur. 11.4. (exception),  
2nd on Tue. 23.4.
- ▶ Solutions to the exercises can be submitted until **next Monday 8:00 am**.
  - ▶ 1st sheet is due Thur. 18.4., 8 am (exception),  
2nd on Mon. 29.4., 8 am.
- ▶ Exercises will be corrected
- ▶ Tutorials
  - ▶ each **Tuesday 8-10** (beginners). — 1st tutorial at Tuesday 16.4.
  - ▶ each **Thursday 14-16** (advanced). — 1st tutorial at Thursday 18.4.
- ▶ Successful participation in the tutorial gives up to 10% bonus points for the exam.
  - ▶ group submissions are OK (but will yield no bonus points)

# Attend and Work the Tutorials!



[ML exam 2018/19]

# Exams and credit points

- ▶ There will be a written exam at the end of the term
  - ▶ 2h, 4 problems
- ▶ The course gives 6 ECTS
- ▶ The course can be used in
  - ▶ International Master in Data Analytics / Obligatory
  - ▶ Angewandte Informatik MSc. / Informatik / Gebiet KI & ML
  - ▶ IMIT MSc. / Informatik / Gebiet KI & ML
  - ▶ Wirtschaftsinformatik MSc / Informatik / Gebiet Business Intelligence

## Some books

- ▶ Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman (2014):  
*Mining of massive datasets*,  
Cambridge University Press.  
Available online:  
<http://infolab.stanford.edu/~ullman/mmds.html>
- ▶ Gautam Shroff (2014):  
*The Intelligent Web: Search, smart algorithms, and big data*,  
Oxford University Press.

# References

Doug Laney. 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70), 2001.