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: 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:

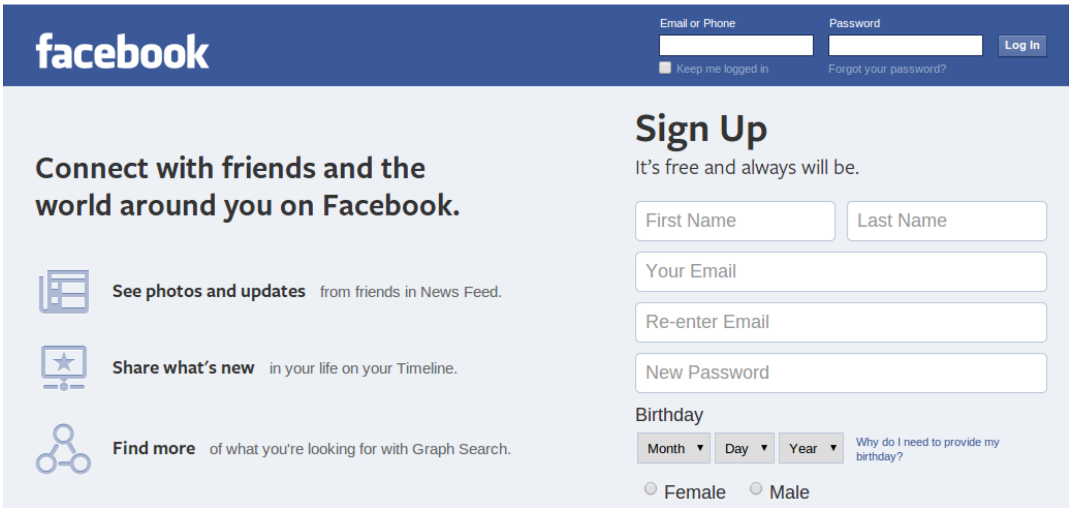
- "A collection of data sets so **large and complex** that it becomes difficult to process using on-hand database management tools or traditional data processing applications."
  `http://en.wikipedia.org/wiki/Big_data`
- "Big data is **high-volume, high-velocity and high-variety** information assets that demand cost-effective, innovative forms of information processing for **enhanced insight and decision making**."
  `www.gartner.com/it-glossary/big-data/`

What is Big Data?

Big Data is about:

- Storing and accessing large amounts of (unstructured) data
- Processing high volume data streams
- Making sense of the data
- Predictive technologies

Where to find Big Data?



Connect with friends and the world around you on Facebook.

See photos and updates from friends in News Feed.

Share what's new in your life on your Timeline.

Find more of what you're looking for with Graph Search.

Sign Up

It's free and always will be.

Birthday

○ Female ○ Male

- 1.28 billion users (1.23 billion monthly active in January 2014)
- 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: `http://allfacebook.com/orcfile_b130817`

Where to find Big Data?

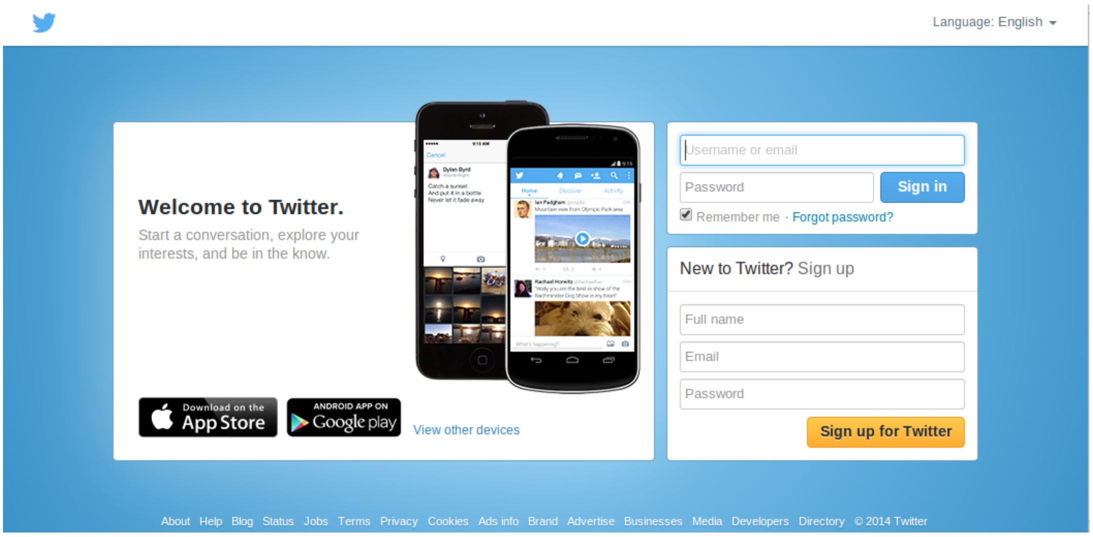
Google

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

[1] `http://searchengineland.com/google-search-press-129925`

[2] 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?



Language: English
Download on the App Store
ANDROID APP ON Google play
View other devices
Username or email
Password
Sign in
☑ Remember me · Forgot password?
Full name
Email
Password
Sign up for Twitter
About Help Blog Status Jobs Terms Privacy Cookies Ads info Brand Advertise Businesses Media Developers Directory © 2014 Twitter

Welcome to Twitter.

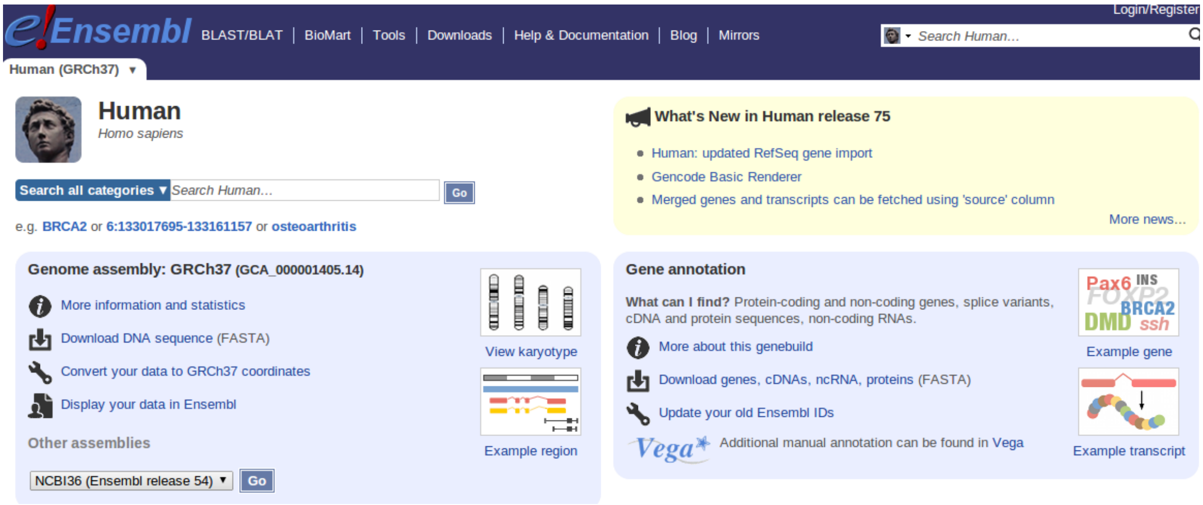
Start a conversation, explore your interests, and be in the know.

New to Twitter? Sign up

- Average number of tweets per day: 58 million[1]
- Number of Twitter search engine queries every day: 2.1 billion[1]
- Total number of active registered Twitter users: 645,750,000[1]

[1] `http://www.statisticbrain.com/twitter-statistics/`

Where to find Big Data?



Human

Homo sapiens

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

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...

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

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)

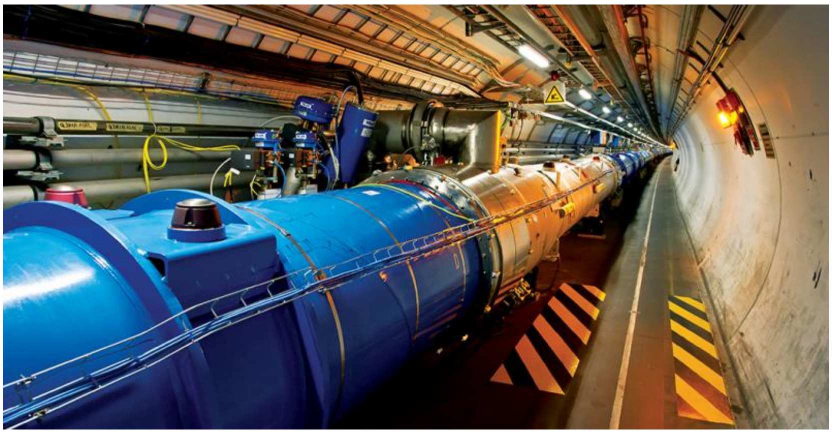
Update your old Ensembl IDs

Additional manual annotation can be found in Vega

- Ensembl database contains the genome of humans and 50 other species
- "only" 250 GB[1]

[1] `http://www.ensembl.org/`

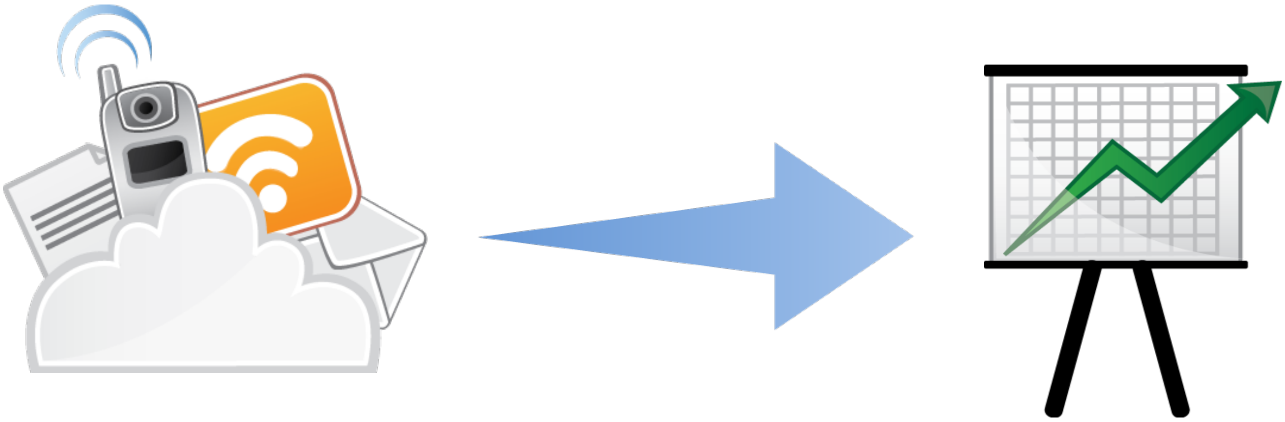
Where to find Big Data?



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

What to do with Big Data?

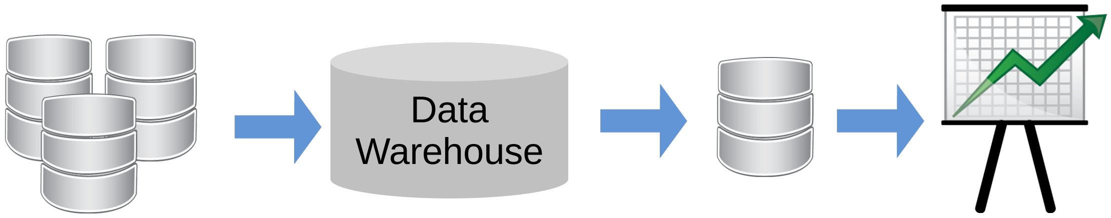
We don't want to know things but to understand them!



What to do with Big Data? - Case Studies

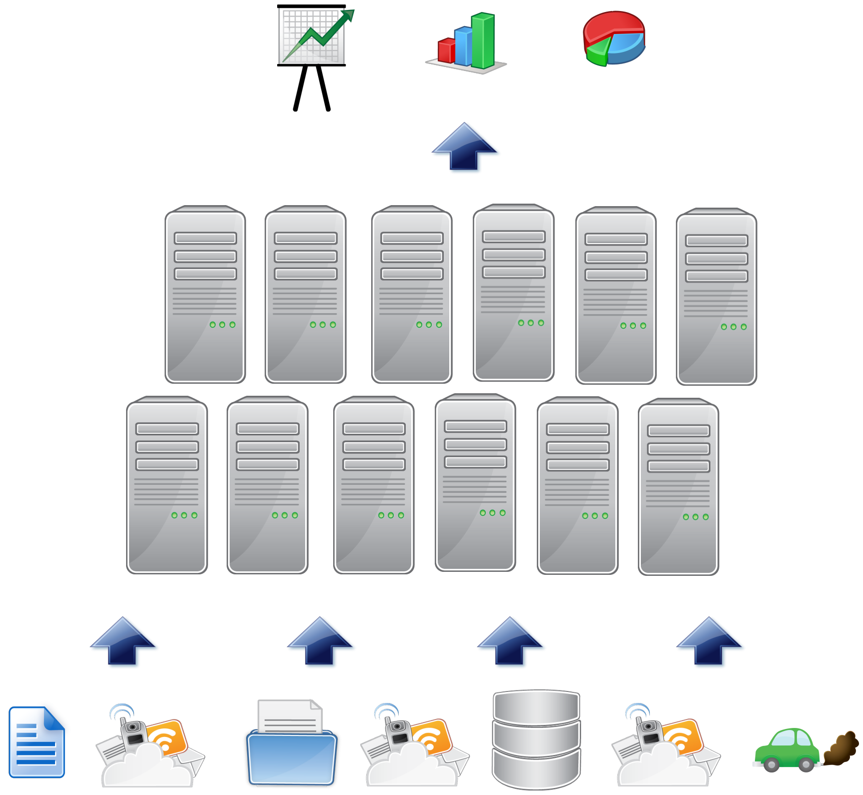
- **T-Mobile USA:** integrated Big Data across multiple IT systems to combine customer transaction and interactions data in order to better predict customer defections
  - By leveraging social media data along with transaction data from CRM and Billing systems, customer defections has been cut in half 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?



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

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

How?

In order to deal with large volumes of data we need to address the following challenges:

- Effectively store and large amounts of data in a distributed environment
- Query distributed databases
- Parallel and distributed programing models
- Data Mining and machine learning techniques to make sense of the data
- Effective data visualisation techniques

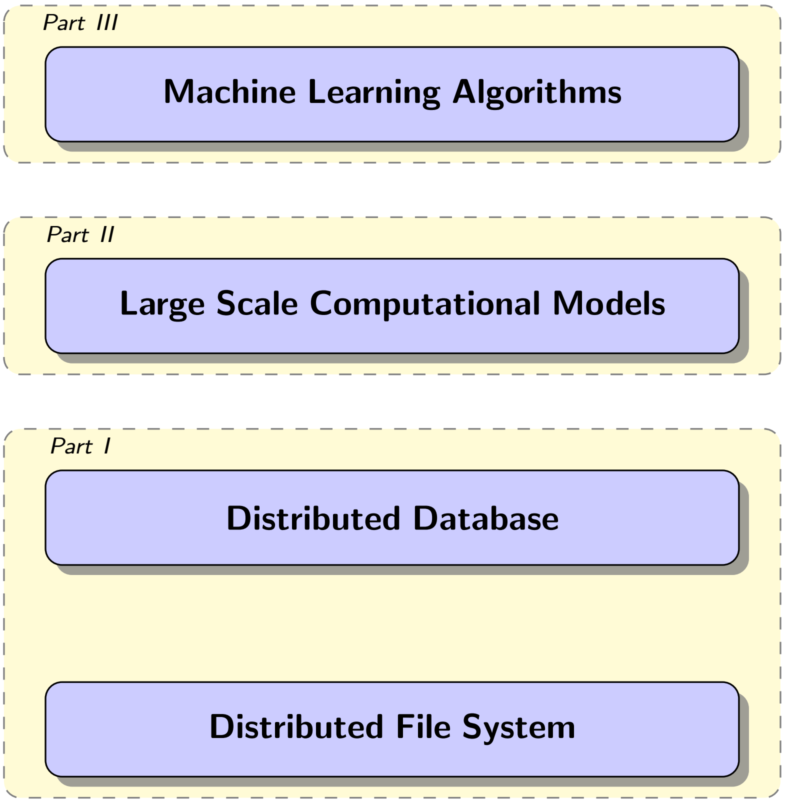
Outline

1. What is Big Data?

2. Overview

3. Organizational Stuff

Overview



Part III

Machine Learning Algorithms

Part II

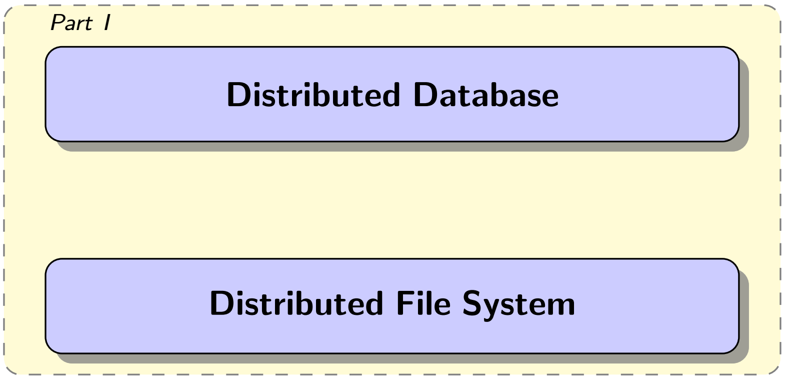
Large Scale Computational Models

Part I

Distributed Database

Distributed File System

Overview



Part I

Distributed Database

Distributed File System

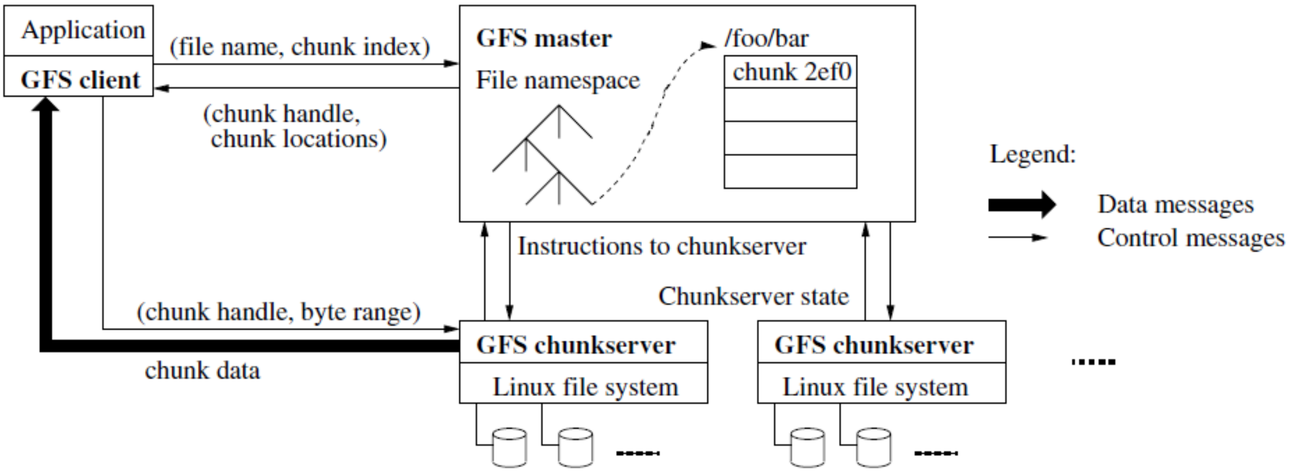
Storing

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 procesing

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

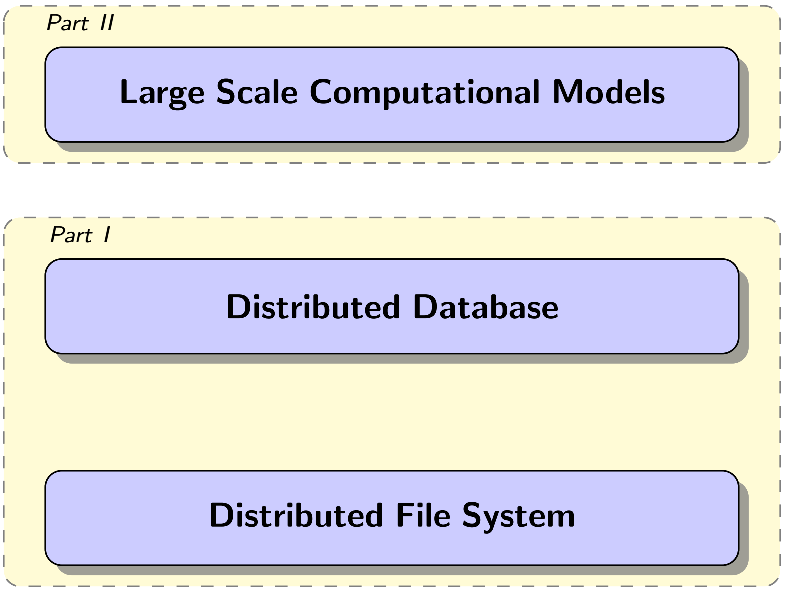
document

CouchDB relax

mongoDB

terrastore

Overview



Part II

Large Scale Computational Models

Part I

Distributed Database

Distributed File System

Accessing

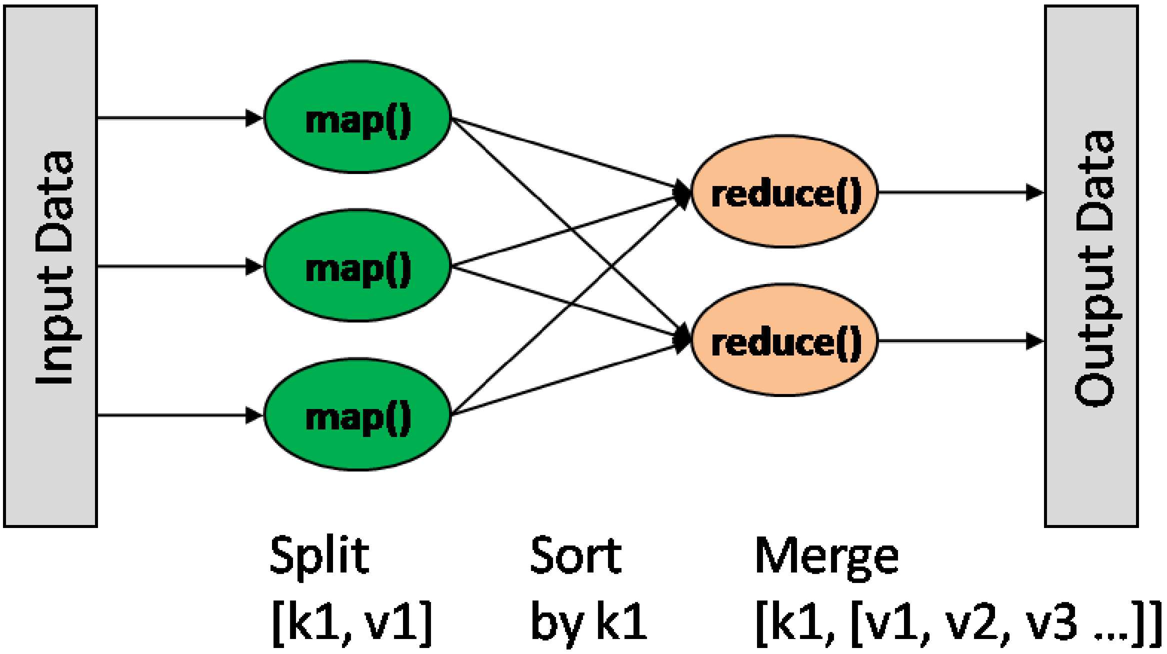
A computational model is needed to:

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

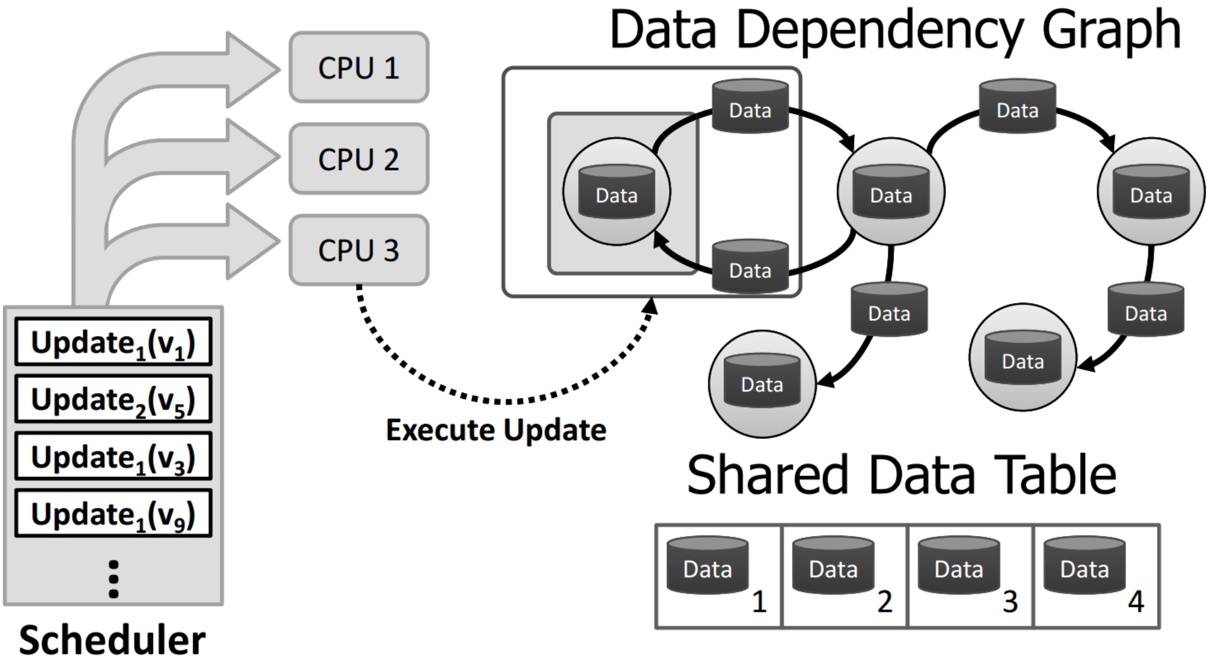
Examples:

- MapReduce
- GraphLab
- Pregel
- Apache Spark

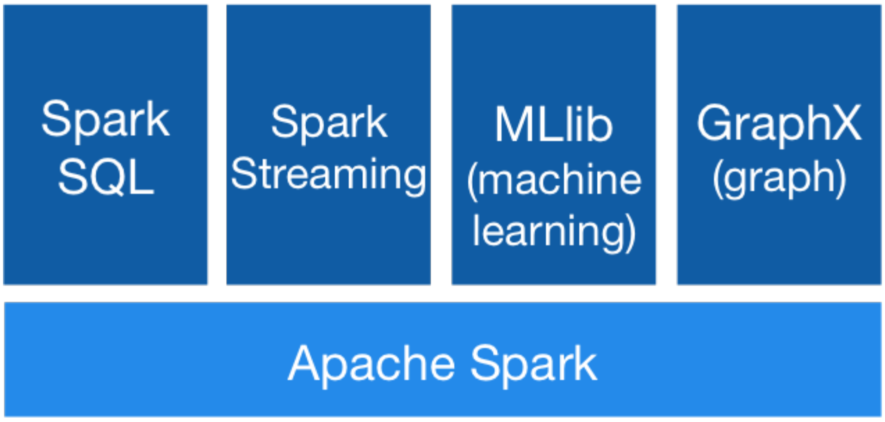
MapReduce



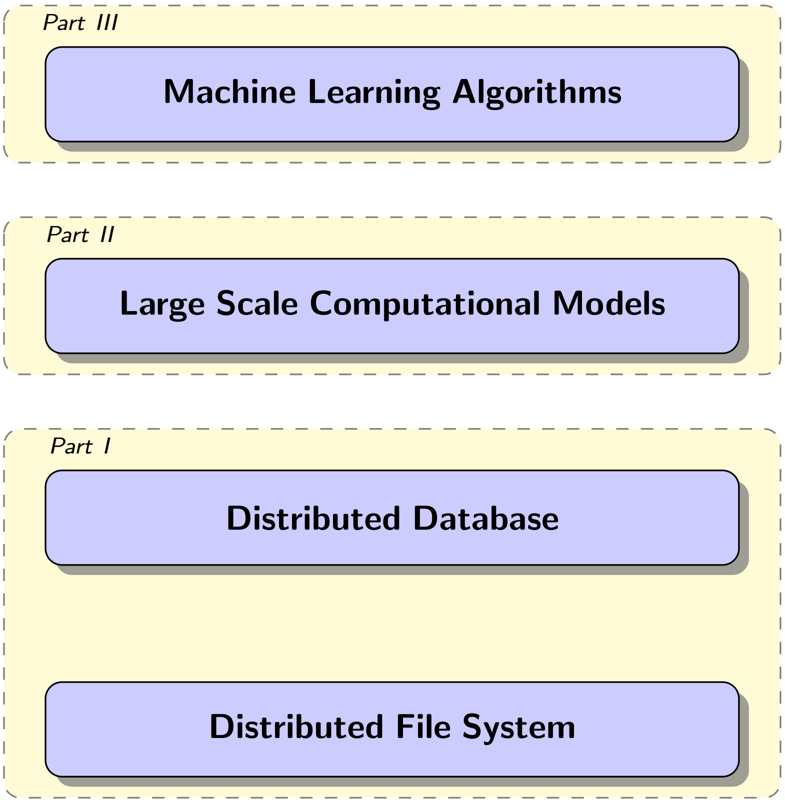
GraphLab



Apache Spark



Overview



Part III

Machine Learning Algorithms

Part II

Large Scale Computational Models

Part I

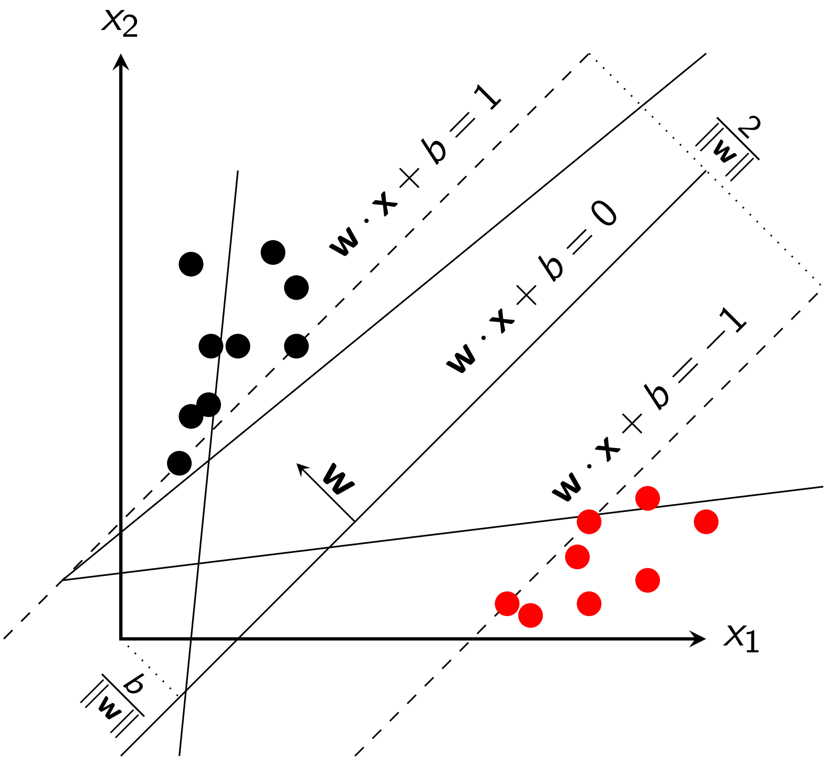
Distributed Database

Distributed File System

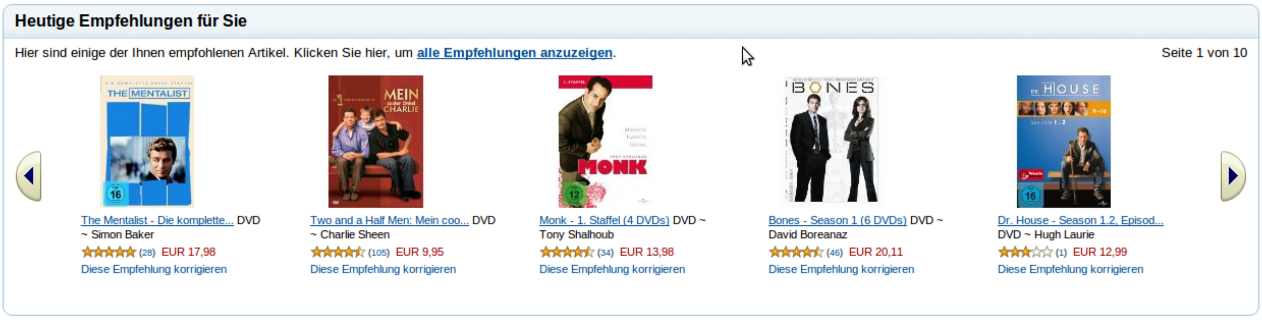
Making sense of the data

- 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


x_2
x_1
$\mathbf{w} \cdot \mathbf{x} + b = 1$
$\mathbf{w} \cdot \mathbf{x} + b = 0$
$\mathbf{w} \cdot \mathbf{x} + b = -1$
$\mathbf{w}$
$\frac{2}{\|\mathbf{w}\|}$
$\frac{b}{\|\mathbf{w}\|}$

Recommender Systems



Heutige Empfehlungen für Sie

Hier sind einige der Ihnen empfohlenen Artikel. Klicken Sie hier, um alle Empfehlungen anzuzeigen.

Seite 1 von 10

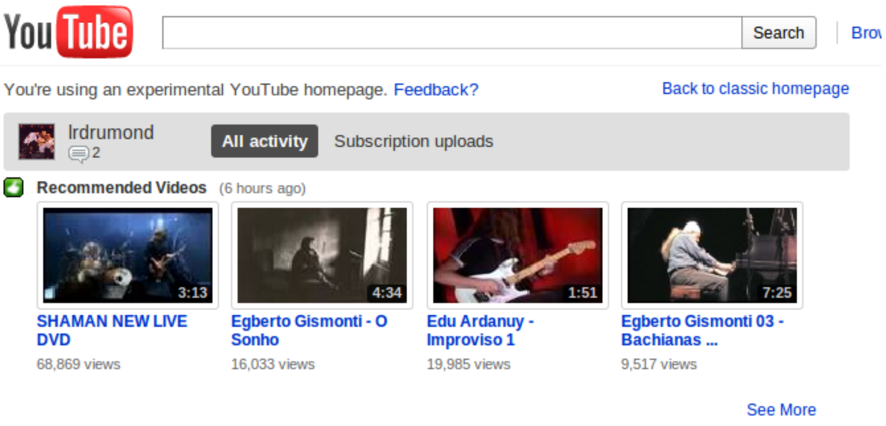
The Mentalist - Die komplette... DVD ~ Simon Baker (28) EUR 17,98 Diese Empfehlung korrigieren

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 (46) EUR 20,11 Diese Empfehlung korrigieren

Dr. House - Season 1.2, Episod... DVD ~ Hugh Laurie (1) EUR 12,99 Diese Empfehlung korrigieren



You're using an experimental YouTube homepage. Feedback? Back to classic homepage

lrdrumond 2 All activity Subscription uploads

Recommended Videos (6 hours ago)

SHAMAN NEW LIVE DVD 3:13 68,869 views

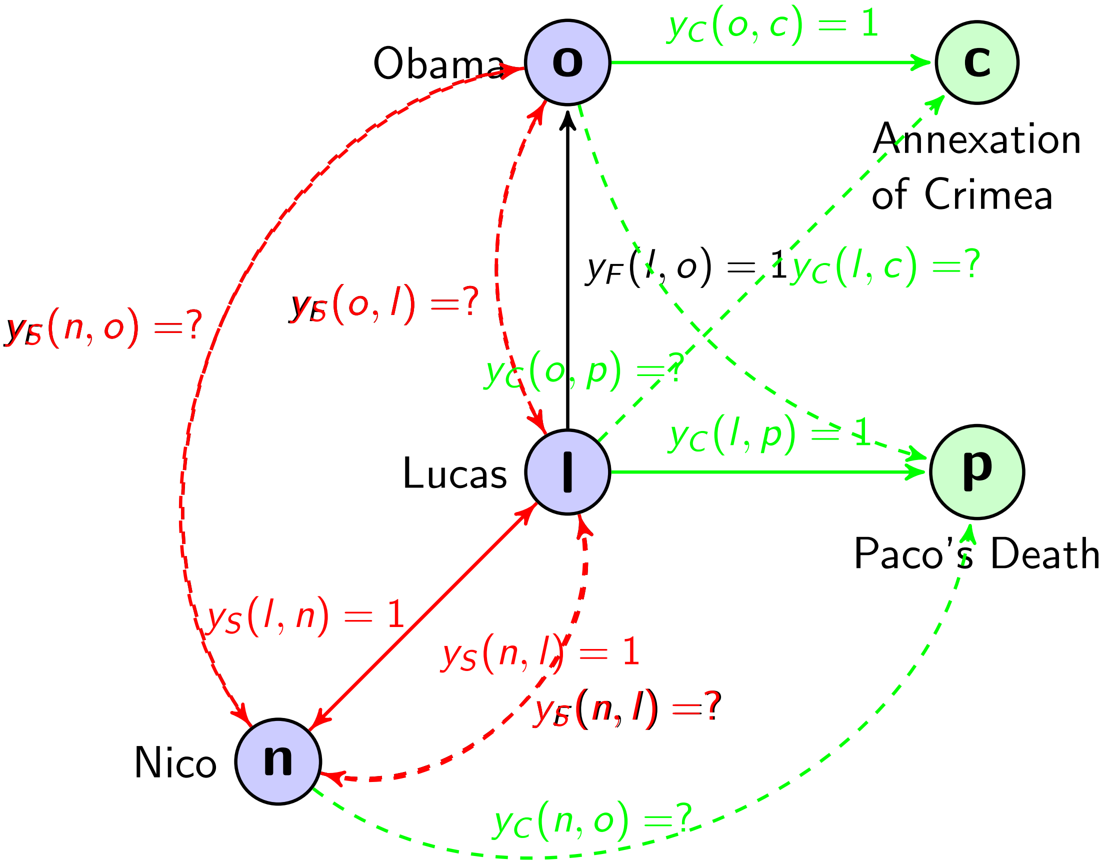
Egberto Gismonti - O Sonho 4:34 16,033 views

Edu Ardanuy - Improviso 1 1:51 19,985 views

Egberto Gismonti 03 - Bachianas ... 7:25 9,517 views

See More

Graph Analysis



Course Overview

Main goal: predictive analytics from large scale data!

- Introduction (1 Lecture)
- Machine Learning problems afflicted by Big Data (3 Lectures)
- Distributed Learning algorithms (3 Lectures)
- Parallel and distributed programing models (4 Lectures)
- Large scale storage and retrieval mechanisms (1 Lecture)

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 Wednesday** in the lecture.
  1st sheet will be handed out Wed. 13.4
- Solutions to the exercises can be submitted until **next Wednesday before the lecture**.
  1st sheet is due Wed. 20.4.
- Exercises will be corrected
- Tutorials each Thursday 14-16.
  1st tutorial at Friday 7.4
- Successful participation in the tutorial gives up to 10% bonus points for the exam.

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
  - IMIT MSc. / Informatik / Gebiet KI & ML
  - Wirtschaftsinformatik MSc / Informatik / Gebiet KI & ML

Some books

- Anand Rajaraman, Jure Leskovec, and Jeffrey Ullman: "Mining of massive datasets" Available online: http://infolab.stanford.edu/ ullman/mmds.html
- Gautam Shroff: “The Intelligent Web: Search, smart algorithms, and big data”