Big Data Analytics

COMP336 / COMP529

Dom Richards

University of Liverpool

December 1, 2024

Table of Contents

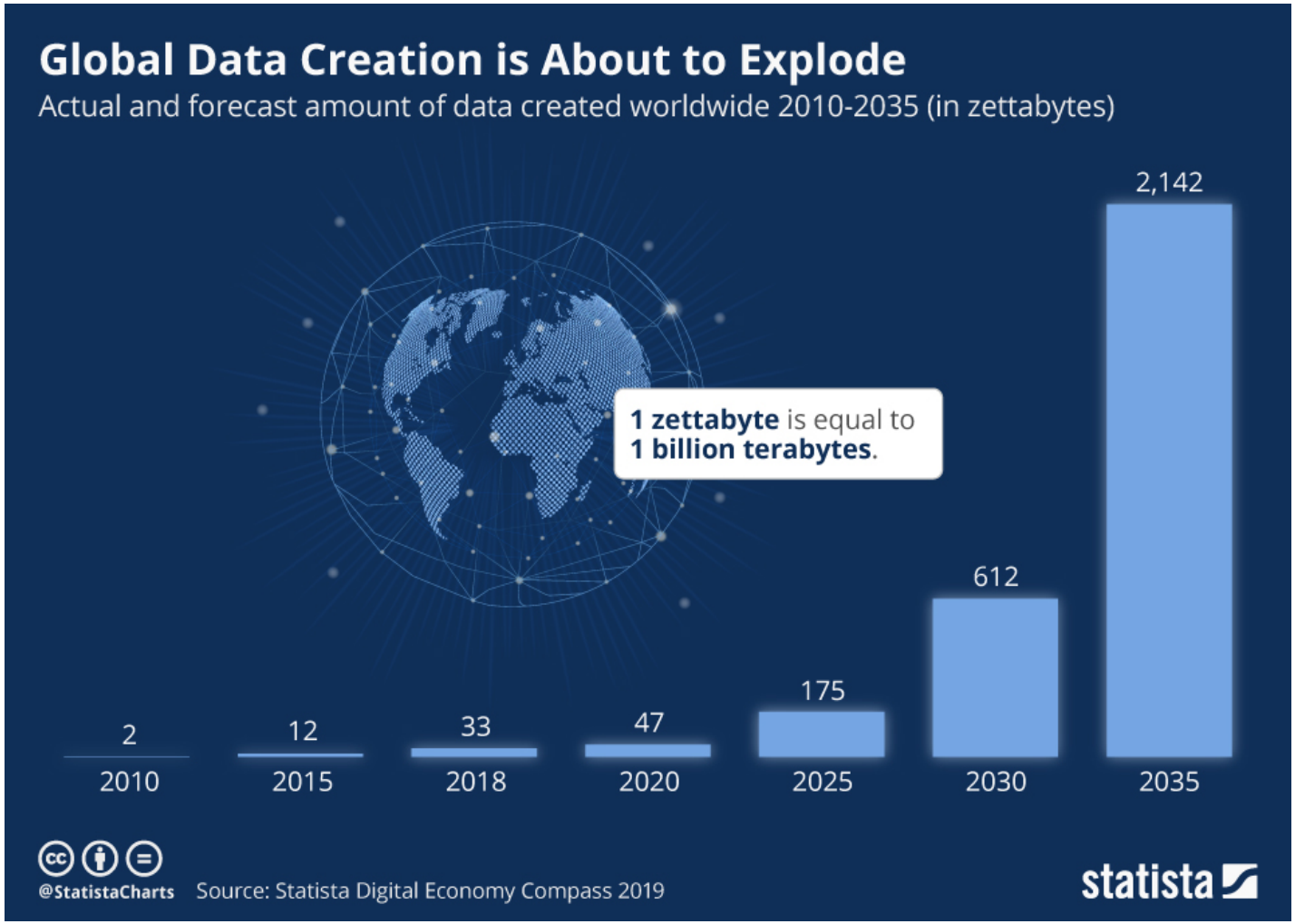
Introduction

Attendance: 154797

Introduction

- Welcome to Big Data Analytics!
- Over the next 12 weeks, we'll explore the field of big data
- It's concerned with the storage and analysis of datasets that are too large to fit on a computer, and hence typically require a *data center* to store and process
- 'Big' datasets (also referred to as big data) are often very valuable, and they crops up everywhere: written text; social media posts; financial transaction records; industrial sensors; scientific research data
- By the end of this course, you'll have a solid understanding of big data technologies and how to apply them to real-world problems
- We'll introduce a range of techniques to extract value from big datasets including: supervised and unsupervised machine learning; statistical analysis; natural language processing; network science; signal processing

How Common is Big Data?



Source Statista report on global data creation forecasts.

Example Datasets

- Search Engine Data: googlebot
- Social Media Data: Posts, comments, and reactions on platforms like Facebook and Twitter
- E-commerce Data: Online shopping behaviours, product views, and purchases
- Healthcare Data: Patient records, genetic sequences, and medical device data
- IoT Data: Sensor data from connected devices like smart homes and wearables

Attendance: 154797

Understanding Big Datasets

- In the next two sections, we'll explore the characteristics of big datasets and the challenges they present
- In this section, we'll discuss the 6 Vs of big data: volume; velocity; variety; veracity; value; and variability
- In the next section, we'll take a look at how big data is represented for the purpose of analysis

Characteristics of Big Data

- Big data comes in many shapes and sizes, but it can be characterised in terms of “6 Vs”:
  - Volume: the amount of data in a given dataset
  - Velocity: Data can be generated at a range of possible speeds
  - Variety: Data comes in different formats, including structured, semi-structured, and unstructured
  - Veracity: Data quality and accuracy can vary
  - Value: which derives from the ability to transform big data into actionable insights
  - Variability: Data can be inconsistent or change rapidly

Volume

- Volume refers to the quantity of data
- Ranges from GBs upwards

Velocity

- Velocity refers to the speed at which data is being generated, processed, and analysed
- For some big data applications, data arrives continuously from multiple sources in real time
- Example: Credit card transactions need to be validated quickly to prevent fraud
- Rapid decision-making is crucial in fields like:
  - Stock trading platforms, where delays can result in losses
  - Autonomous vehicles, which rely on sensor data being processed in real-time

Variety

- Variety refers to the wide range of data types and formats
- Traditional systems only handled structured data, typically found in relational databases (e.g., SQL)
- Today, data comes in a variety of formats:
  - Structured data: Databases, spreadsheets, CSV, Parquet
  - Unstructured data: Emails, videos, social media posts, images

Veracity

- Veracity refers to the accuracy, quality, and trustworthiness of data
- Big data often includes noisy, incomplete, or inconsistent data
- Example challenges:
  - Social media data may include misleading or biased information
  - Sensor data can contain errors due to faulty devices
- Techniques to improve veracity:
  - Data cleaning: Removing inaccuracies or noise
  - Data validation: Ensuring data is correct before analysis

Value

- Value refers to the insights and actionable information extracted from big data
- Big data is often valuable if it can be transformed into meaningful insights
- In this course, we'll see a range of techniques to extract value from big data, including: machine learning; statistical analysis; natural language processing; network science; signal processing.

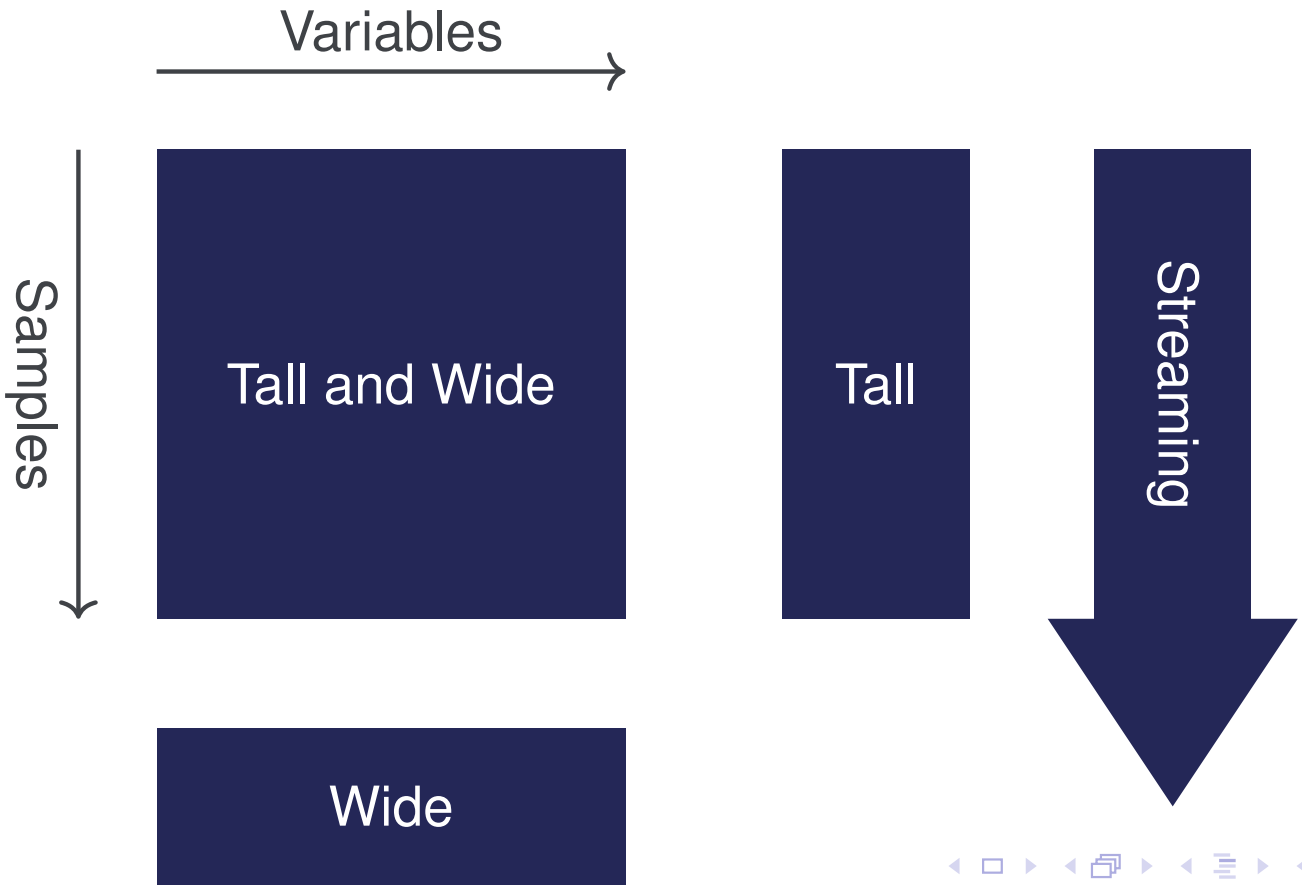
Variability

- Variability refers to the changing nature of data and the inconsistency in its flow
- Data can be generated at unpredictable rates and can vary in its meaning and relevance
- E.g. regime switching:
  - Social media sentiment analysis can shift rapidly based on trending topics or events
  - IoT sensor data might fluctuate based on environmental changes or technical issues
- Handling variability requires adaptive algorithms and real-time monitoring to respond quickly to data changes

Attendance: 154797

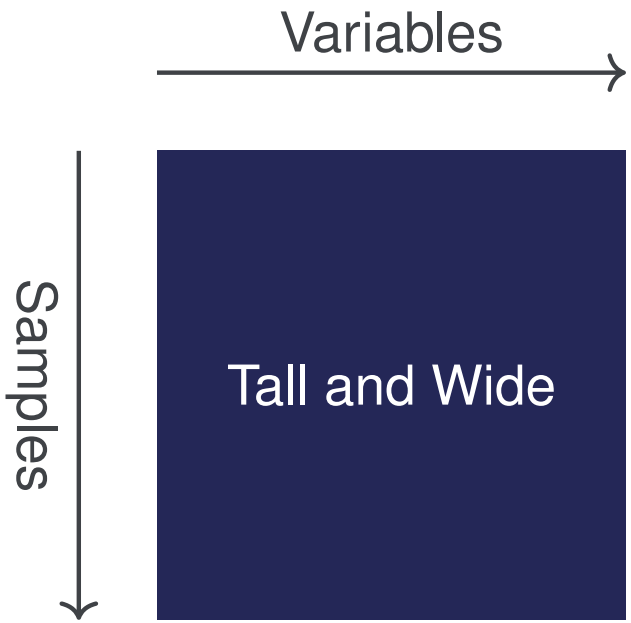
Dataset Types

Big datasets come in various shapes and sizes. We tend to think about matrices of variables and samples, even when the data isn't originally presented to us like this (e.g. text documents or images)



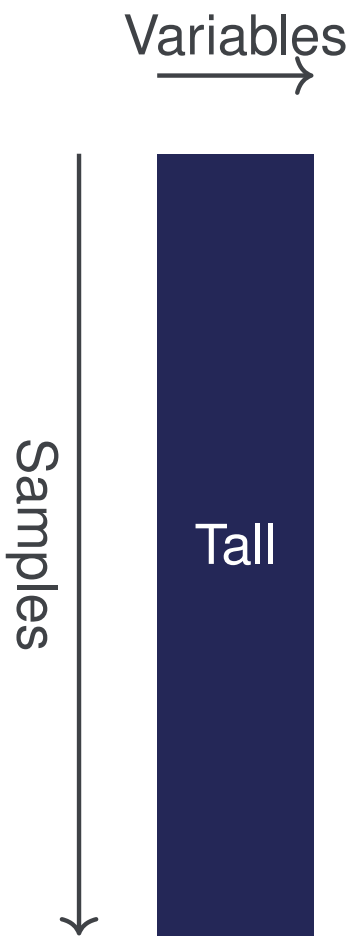
Tall and Wide Datasets

- These datasets can be very large:
  - Millions to billions of variables and samples
- Typical example is networks:
  - Social networks, contagion networks, etc.



Tall Datasets

- Hundreds to thousands of variables
- Thousands to millions of samples
- Often suitable for machine learning
- Examples:
  - NLP (Natural Language Processing)
  - Image databases
  - Scientific computing (e.g., solutions of fluid dynamics equations)
  - Reinforcement learning

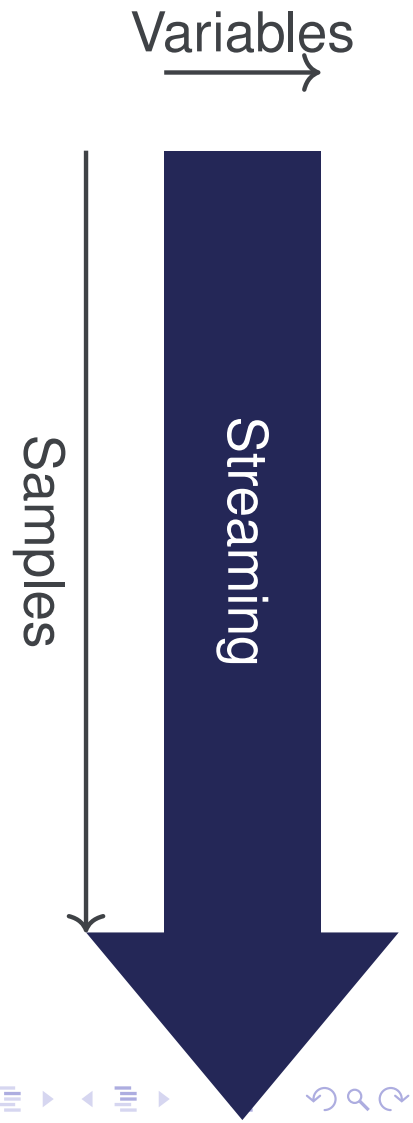


Wide Datasets

- Thousands to millions of variables
- Hundreds to thousands of samples
- Variables $\gg$ samples
- Examples:
  - Genomics datasets
  - Document modelling

Streaming Datasets

- Samples continually created in real time
- Examples:
  - Sensor data
  - Social media feeds
  - Financial transactions



Dense vs Sparse Matrices

- Data matrices can be dense or sparse
- Dense matrix:
  - Most matrix elements are non-zero
  - All elements are stored, including the zeroes
  - Memory usage is proportional to the total number of elements ($variables \times samples$)
- Sparse matrix:
  - Majority of elements are zero
  - Only non-zero elements are stored, often as a coordinate list, which stores triplets (row, column, value)
  - Memory usage is proportional to the number of non-zero elements

Course Overview

2 Course Overview
- Big Data Middleware
- Network Science
- Unsupervised Learning
- Supervised Learning
- Natural Language Processing
- Bayesian Statistics and Filters

Attendance: 154797

Attendance: 154797

Working with Big Data

- In weeks 2, 3 and 11, you'll learn about big data middleware
- Datasets in big data are typically too large to be processed on a single computer, so software has been developed to process such datasets across multiple computers in a data centre
- This software is called *big data middleware*
- It is installed onto a cluster of compute nodes in a data centre, and allows users to interact with the entire cluster as if it were a single computer
  - Storage is distributed across the cluster, and the user is presented with a single file system
  - Processing is distributed across the cluster, allowing the user to operate on all of the data by writing a single program which is executed on all the cluster nodes in parallel
- You'll learn about the two most popular open source big data middleware frameworks: Hadoop and Spark

Hadoop

- Next week we'll see Hadoop.
- Hadoop was the first major open-source big data middleware framework
- It was based on Google's MapReduce and Google File System papers (which are required reading for next week)
- The core components of Hadoop:
  - HDFS (Hadoop Distributed File System): A scalable and fault-tolerant file storage system
  - MapReduce: A programming model for processing large datasets in parallel. Performs fairly simple operations (map and reduce) at huge scales

Storing Data with HDFS

- The Hadoop Distributed File System (HDFS) provides scalable, fault-tolerant data storage
- HDFS splits large files into blocks and distributes them across multiple machines for parallel storage and retrieval

Spark

- In weeks 3 and 11, we'll look at Spark
- Spark was developed after Hadoop to address some of its perceived limitations
- It's often faster than Hadoop (due to a more sophisticated use of RAM on cluster nodes), and can perform a broader range of tasks
- You'll learn about Spark in week 3
- In week 11, you'll see how Spark can perform advanced data analysis, including machine learning and graph processing:
  - Spark's MLLib library supports a wide range of machine learning algorithms
  - Spark's GraphX library enables complex graph computations

Attendance: 154797

Network Science

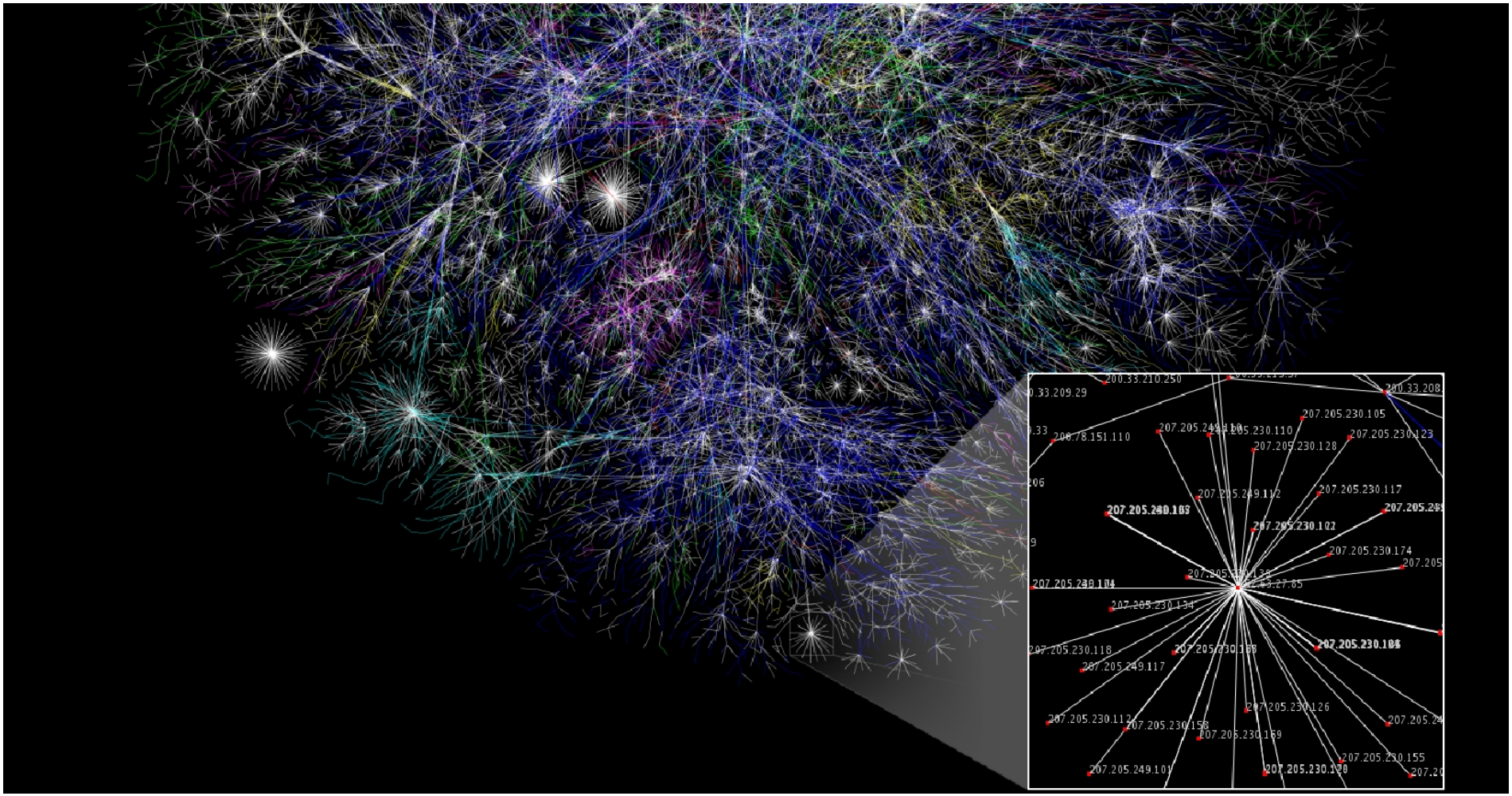
- In weeks 4, you'll move on to the task of extracting value from big datasets
- You'll learn about Network Science, and it's use in processing network (AKA graph) based big datasets
- A network consists of nodes (or vertices) connected by edges (or links)
- Examples of networks in big data include social networks, biological networks, and transportation systems
- Network Science studies the structure, dynamics, and behaviour of networks

Facebook Connections

Facebook Connections Map the World

- The image shows a map of Facebook connections as of 2010
- The lines on the map don't represent physical features like coasts or borders, but instead reflect real human relationships and social networks
- The map also highlights areas with minimal Facebook presence c. 2010, such as China and central Africa, where few connections exist

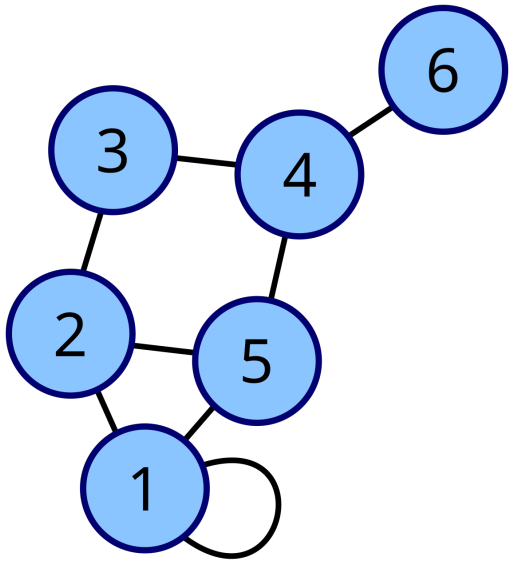
Internet Map



Internet Map Visualisation

- This image shows a map of the Internet, visualising the global structure of interconnected networks
- Each node represents a network, and the edges represent the connections between these networks
- Such visualisations help understand the topology and connectivity of the Internet infrastructure

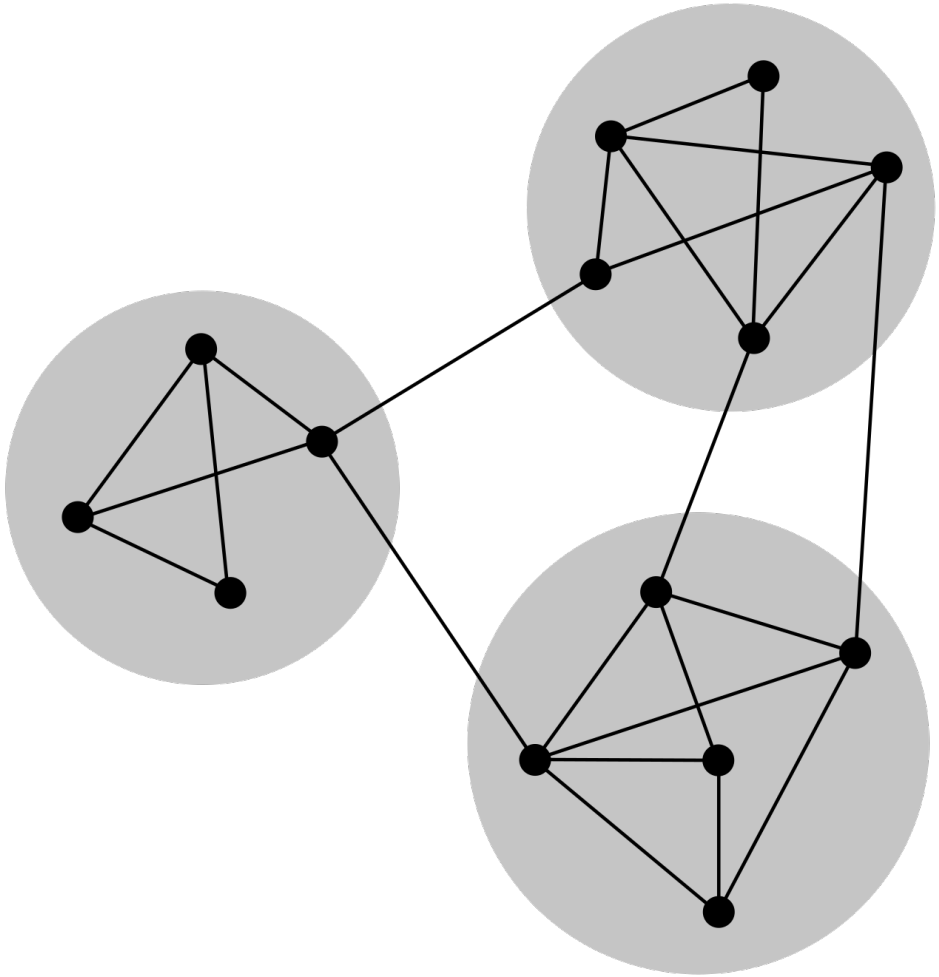
Network Representations



$$\begin{pmatrix} 2 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

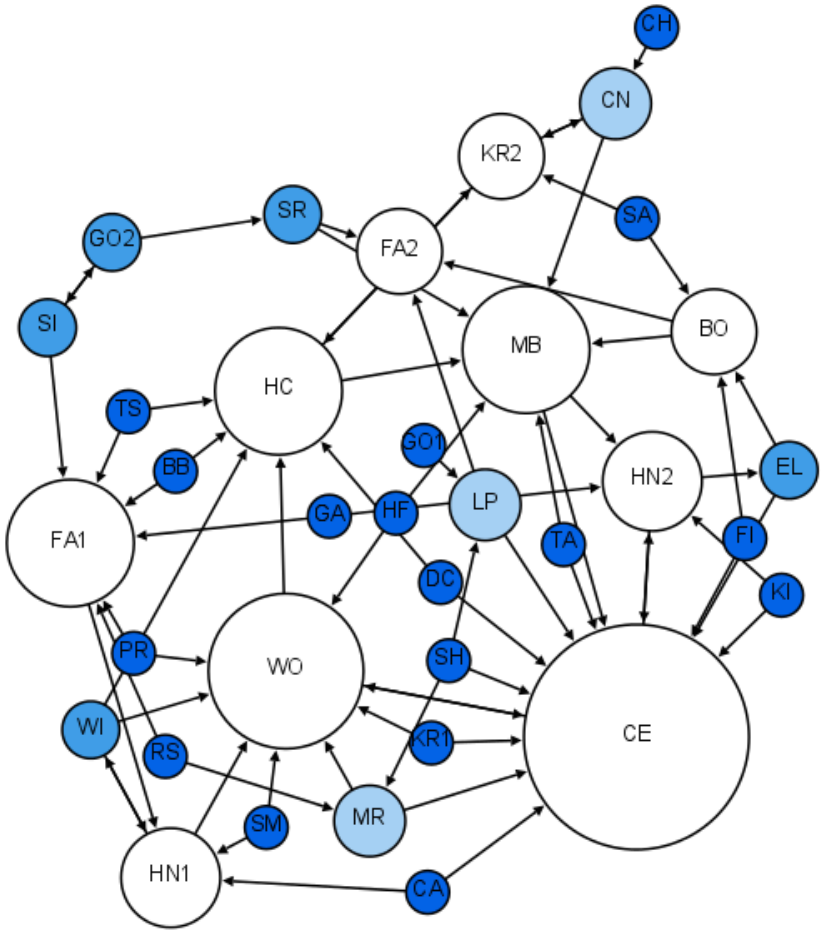
Figure: Comparison between graph and its adjacency matrix.

Network Community Structures



Preferential Attachment in Networks

The Moreno sociogram of a 1st grade class:



Scale-Free Networks

- **Scale-free networks** are a type of network characterised by a power-law degree distribution
- In these networks, most nodes have few connections, while a few nodes (called "hubs") have many connections
- They have a **Power-law distribution**, as you'll see in week 4
- Examples of scale-free networks include the internet, social networks, and biological systems like protein interactions

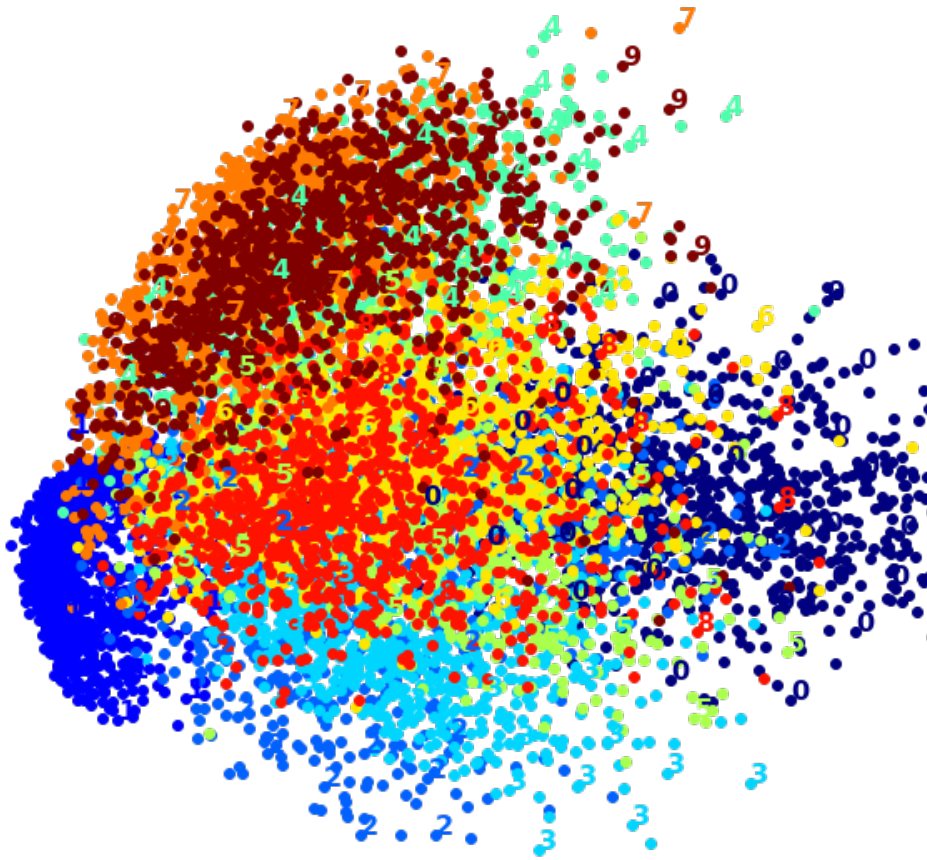
Attendance: 154797

Unsupervised Learning: PCA and Clustering

- In weeks 5 and 6, you'll learn about unsupervised learning
- Unsupervised learning deals with unlabelled data, where the goal is to find hidden patterns or structure in the data
- You'll see two of the most common techniques in unsupervised learning:
  - **Principal Component Analysis (PCA)** (week 5):
    - A dimensionality reduction technique that projects data onto the principal components—directions of maximum variance
    - PCA helps simplify datasets by reducing features while retaining most of the important information
  - **Clustering** (week 6):
    - Clustering algorithms, such as K-means and hierarchical clustering, group similar data points into clusters based on their features
    - The goal is to partition data into distinct groups where points within a group are more similar to each other than to those in other groups

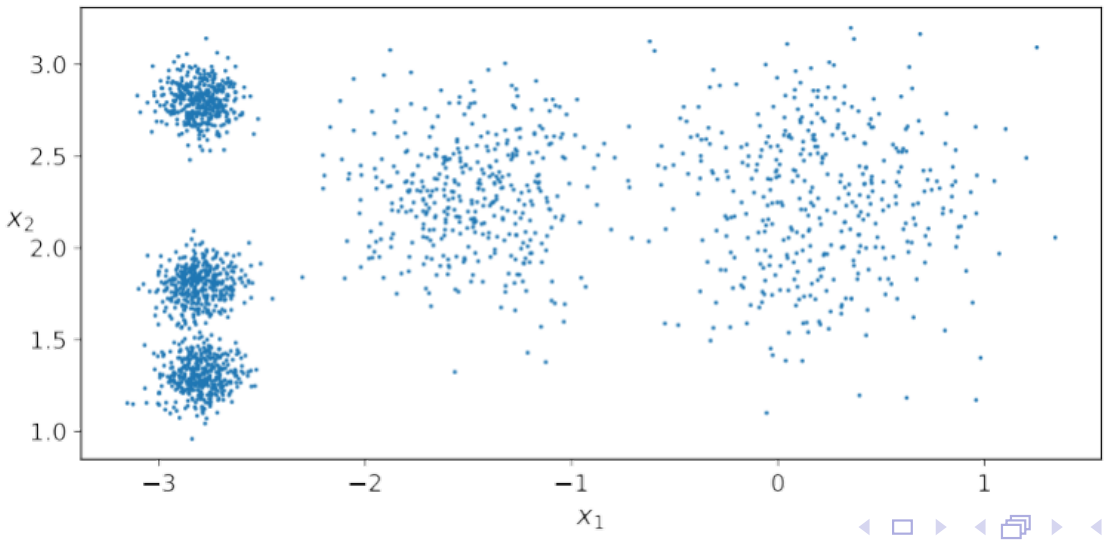
PCA for Exploratory Data Analysis

- We can use PCA to reduce the dimensionality of data all the way down to 2 or 3 dimensions
- At that point, the data can be visualised
- E.g., the plot shows the first two principal components for the handwritten digits dataset
- This often yields insights that can be exploited with further application of machine learning



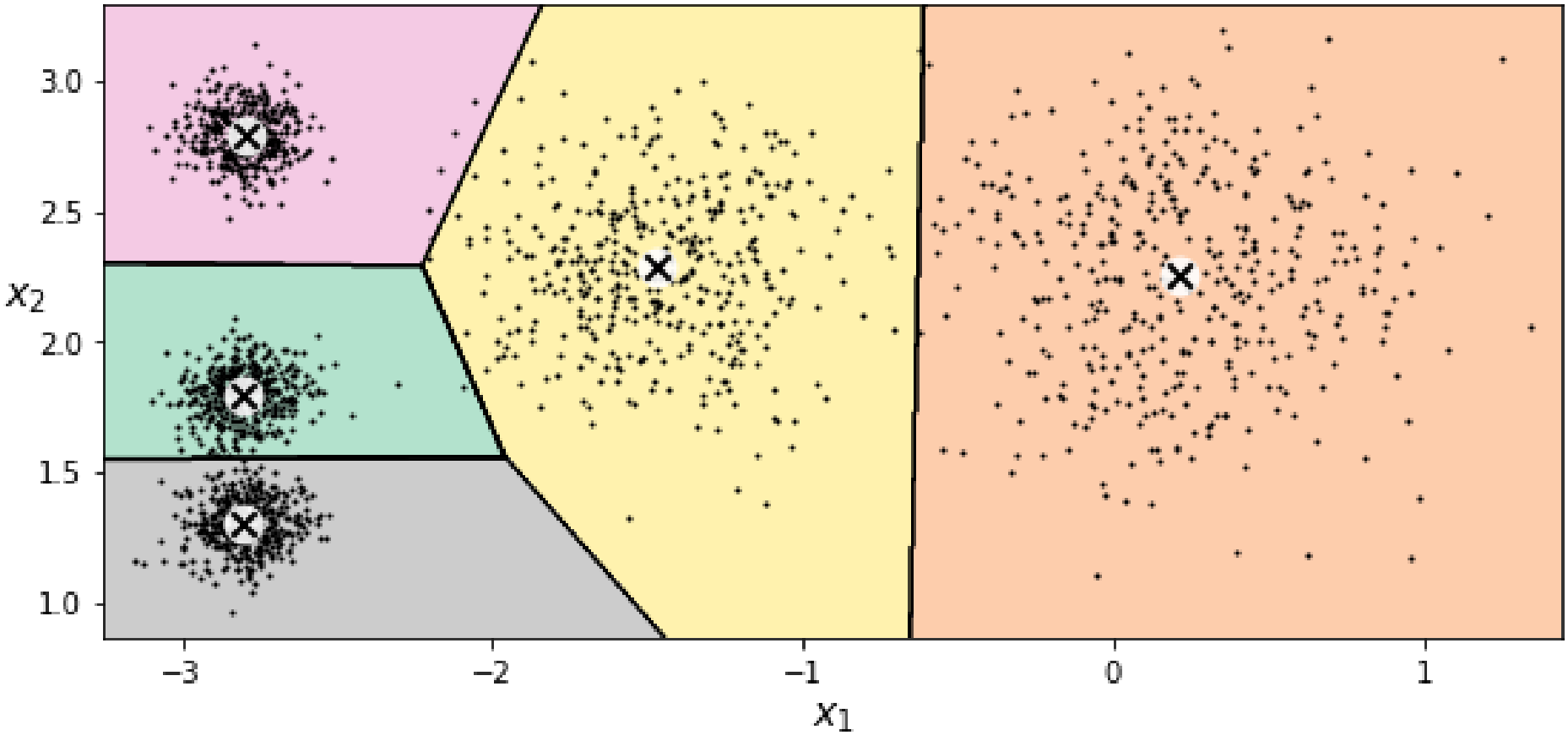
Clustering

- Clustering is an unsupervised learning technique that groups similar data points into distinct clusters
- Clustering allows us to group data points based on their similarity or distance from each other
- It's useful for exploratory data analysis, pattern recognition, and anomaly detection
- Popular algorithms include K-Means, DBSCAN, and Hierarchical Clustering, all of which you'll see in week 6



Clustering with K-Means

Here's the result of applying the K-Means clustering algorithm to the above dataset:



Clustering Faces

Cluster 30

36 36 36

Cluster 31

15 15

Cluster 32

23 23 23

Cluster 33

21 21 21

Cluster 34

26 26 26 26 26

Cluster 35

9 34 34

Cluster 36

31 31

Cluster 37

33 33 33

Cluster 38

6 6

Cluster 39

15 15

Cluster 40

6 6 6

Cluster 41

36 36 36 36

Cluster 42

16 16 16

Cluster 43

16 16 16

Cluster 44

35 35 35

Clustering-Based Image Compression

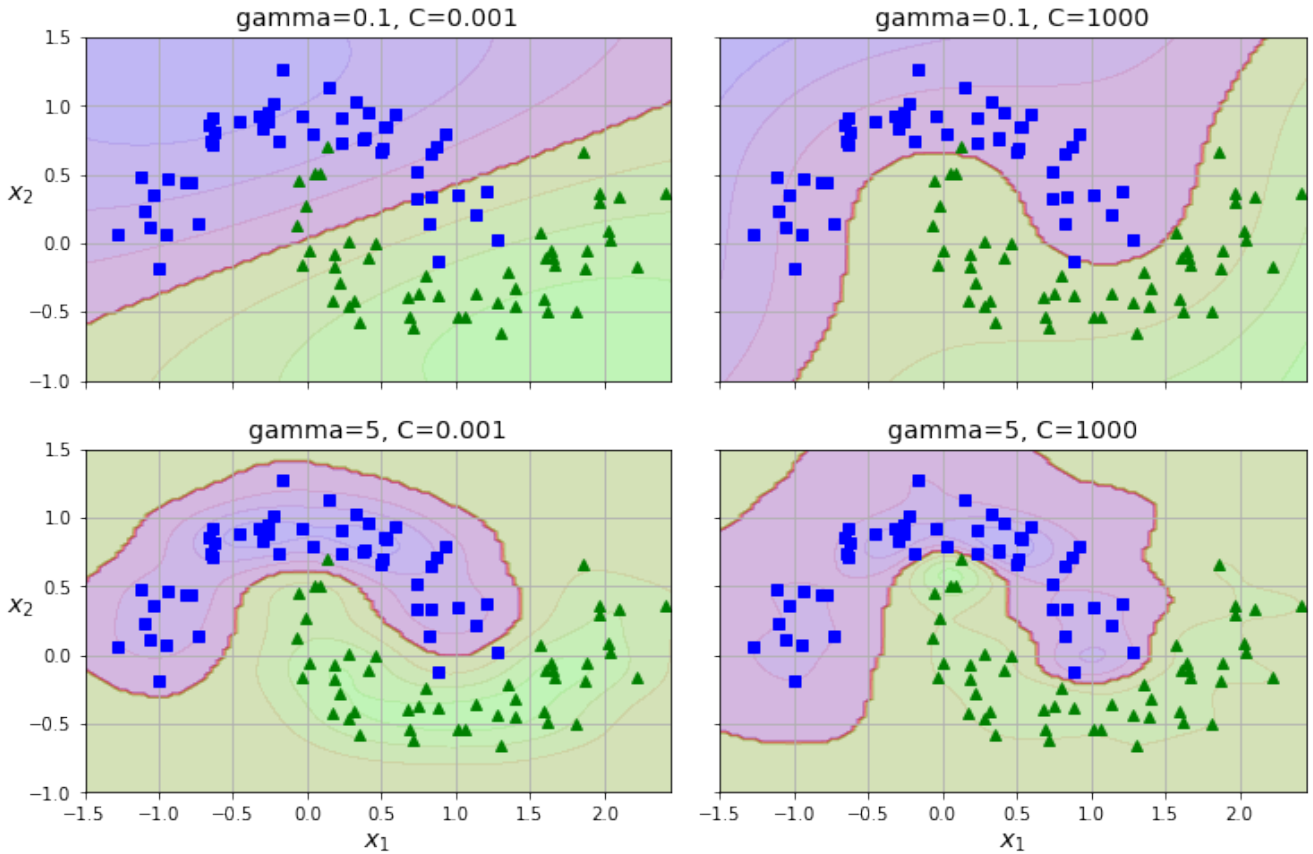
Attendance: 154797

Supervised Learning

- In week 8, we'll look at supervised learning
- Supervised learning is a type of machine learning where the model is trained on labelled data
- The model learns to map input data to output data based on the input-output pairs provided during training
- We'll look at several supervised learning algorithms, including k-Nearest Neighbours and Support Vector Machines

Half Moon Classification

The plot illustrates the use of Support Vector Machines (SVM) on a dataset



Attendance: 154797

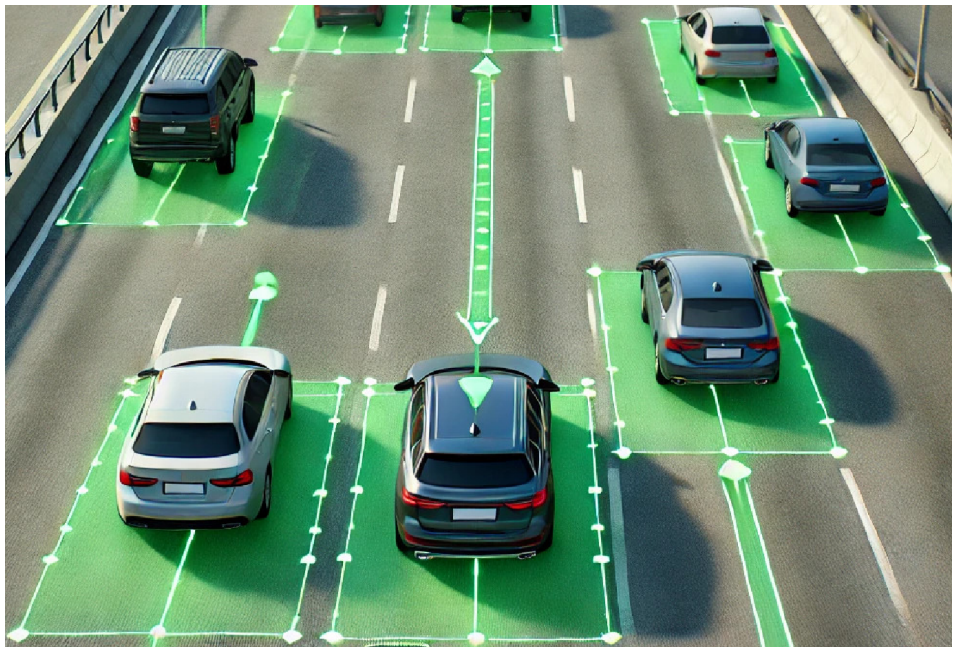
Natural Language Processing

- In week 8, we'll also look at Natural Language Processing (NLP), which enables computers to understand, interpret, and respond to human language
- We'll focus on the task of document search, where the goal is to retrieve relevant documents from a corpus, usually based on a user's query
- We'll look at "bag of words" style document reprentations, culminating in an approach called Term Frequency-Inverse Document Frequency (TF-IDF):

Attendance: 154797

Bayesian Statistics and Filters

- In weeks 7, 9 and 10, you'll learn about Bayesian statistics and filtering
- This will culminate with an introduction to the Kalman filter - a very widely-used algorithm that uses a series of measurements observed over time to estimate the current state of a system and predict it's future evolution
- Applications include vehicle tracking, robotics and navigation



Course Outline

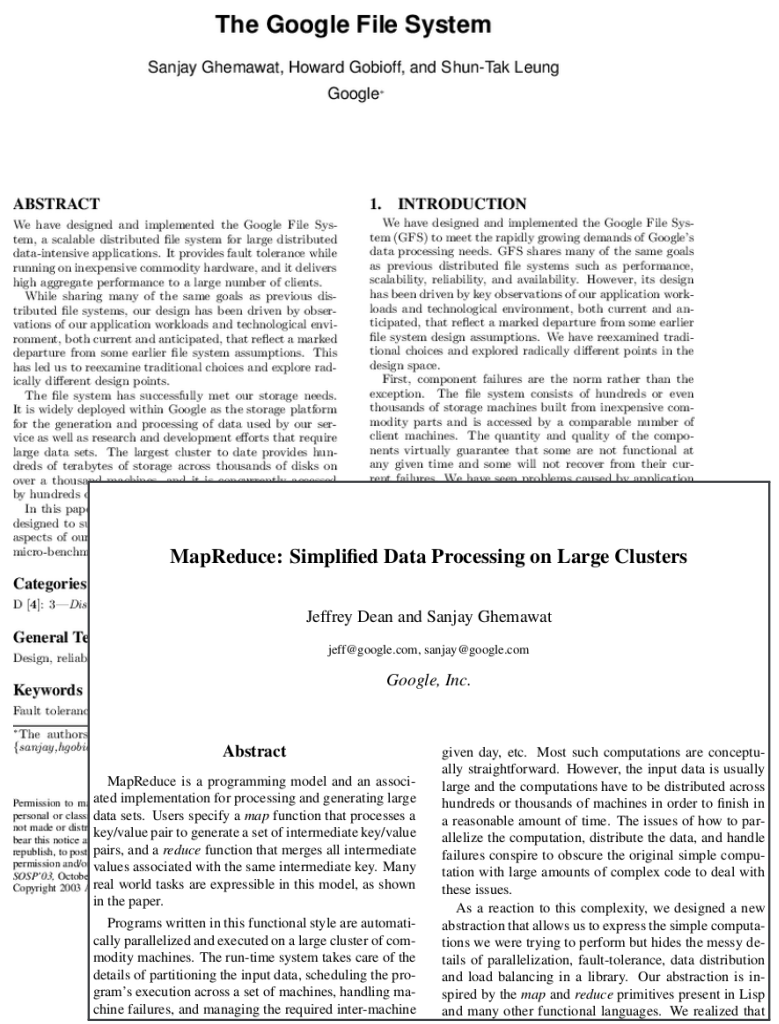
1. Introduction and Course Overview
2. Hadoop
3. Spark
4. Network Science
5. Linear Algebra Approaches
6. Clustering
7. Probabilistic Modelling
8. Classification and Real World Applications
9. Parallel Computing, Bayesian Modelling and Real World Applications
10. Parallel Computing, Bayesian Modelling and Real World Applications
11. Advanced Spark
12. Advanced Approaches

Prep for Next Week

Please read through Google's original white papers on Google File System (GFS) and MapReduce:

- **The Google File System**, 2003
- **MapReduce: Simplified Data Processing on Large Clusters**, 2004

They're very accessible, and they laid the groundwork for Hadoop and HDFS



The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

Google*

ABSTRACT

We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.

While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore radically different design points.

The file system has successfully met our storage needs. It is widely deployed within Google as the storage platform for the generation and processing of data used by our service as well as research and development efforts that require large data sets. The largest cluster to date provides hundreds of terabytes of storage across thousands of disks on over a thousand machines, and it is concurrently accessed by hundreds

In this pap
designed to s
aspects of our
micro-benchm

Categories

D [4]: 3—*Dis*

General Te

Design, reliab

Keywords

Fault tolerand

*The authors
{*sanjay,hgobi*

Permission to m
personal or class
not made or distr
bear this notice a
republish, to post
permission and/o
SOSP'03, Octobe
Copyright 2003 *

1. INTRODUCTION

We have designed and implemented the Google File System (GFS) to meet the rapidly growing demands of Google's data processing needs. GFS shares many of the same goals as previous distributed file systems such as performance, scalability, reliability, and availability. However, its design has been driven by key observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system design assumptions. We have reexamined traditional choices and explored radically different points in the design space.

First, component failures are the norm rather than the exception. The file system consists of hundreds or even thousands of storage machines built from inexpensive commodity parts and is accessed by a comparable number of client machines. The quantity and quality of the components virtually guarantee that some are not functional at any given time and some will not recover from their current failures. We have seen problems caused by application

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new abstraction that allows us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library. Our abstraction is inspired by the *map* and *reduce* primitives present in Lisp and many other functional languages. We realized that

Distributed File Systems and MapReduce

Attendance: 154797

Big Data Middleware

- Today and in weeks 3 and 11, we'll focus on middleware for big data analysis
- Many datasets (or collections of datasets) are too big to fit onto a single computer
  - We'll see a few examples today
- These datasets are stored in a distributed way across different servers in a data centre
- Big data middleware sits on top of data centre hardware. It:
  - Handles the distributed storage of data
  - Provides a computational abstraction (AKA programming model) to make the storage and subsequent processing of these datasets more tractable

Big Data Middleware

- We'll start with the original big data middleware: Google File System (GFS) and Google MapReduce, which were the first big data middleware solutions
  - Google published papers on them in 2003 and 2004 respectively (and implemented them internally prior to this)
  - We'll introduce them with the original motivating examples from Google
  - We'll see that "map reduce" is a very simple computational paradigm, but you can do a lot of useful big processing with it
  - Google unfortunately hasn't released the source code for GFS and MapReduce, so you can't run it

Big Data Middleware

- We'll then look at the first open source big data middleware solution, which was Hadoop MapReduce and the Hadoop Distributed File System (HDFS)
- All of this will lay the groundwork for introducing Apache Spark next week, which is a fast, cutting edge generalisation of the MapReduce concept

Required Reading for Today's Lecture

- **The Google File System**, 2003
- **MapReduce: Simplified Data Processing on Large Clusters**, 2004

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

Google*

ABSTRACT

We have designed and implemented the Google File System, a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and it delivers high aggregate performance to a large number of clients.

While sharing many of the same goals as previous distributed file systems, our design has been driven by observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system assumptions. This has led us to reexamine traditional choices and explore radically different design points.

The file system has successfully met our storage needs. It is widely deployed within Google as the storage platform for the generation and processing of data used by our service as well as research and development efforts that require large data sets. The largest cluster to date provides hundreds of terabytes of storage across thousands of disks on over a thousand machines, and it is concurrently accessed by hundreds

In this pap
designed to s
aspects of our
micro-benchm

Categories

D [4]: 3—*Dis*

General Te

Design, reliab

Keywords

Fault toleranc

*The authors
{*sanjay,hgobi*

Permission to m
personal or class
not made or distr
bear this notice a
republish, to post
permission and/o
SOSP'03, Octobe
Copyright 2003 *

1. INTRODUCTION

We have designed and implemented the Google File System (GFS) to meet the rapidly growing demands of Google's data processing needs. GFS shares many of the same goals as previous distributed file systems such as performance, scalability, reliability, and availability. However, its design has been driven by key observations of our application workloads and technological environment, both current and anticipated, that reflect a marked departure from some earlier file system design assumptions. We have reexamined traditional choices and explored radically different points in the design space.

First, component failures are the norm rather than the exception. The file system consists of hundreds or even thousands of storage machines built from inexpensive commodity parts and is accessed by a comparable number of client machines. The quantity and quality of the components virtually guarantee that some are not functional at any given time and some will not recover from their current failures. We have seen problems caused by application

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new abstraction that allows us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library. Our abstraction is inspired by the *map* and *reduce* primitives present in Lisp and many other functional languages. We realized that

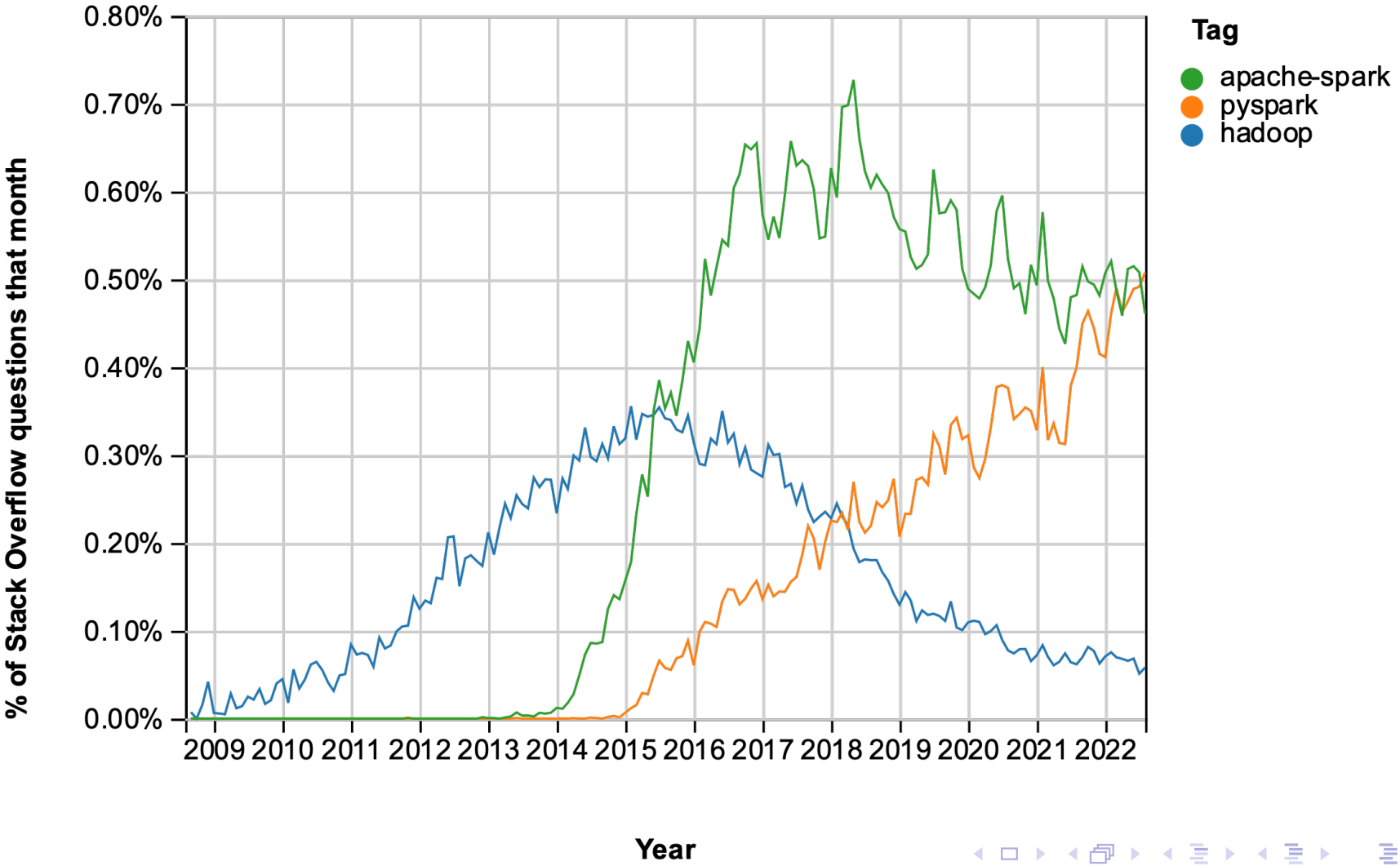
Big Data Middleware at Google

- The original example of big data middleware was Google's proprietary MapReduce and Google File Store:
  - Described by Google in 2003 and 2004 research papers, and implemented at Google some years before that
  - It provided a computational abstraction: simple distributed processing + result aggregation, as we'll see
  - It was designed for use by Google's search engine
- Google's MapReduce software is proprietary and has never been released to the public, but shortly after the 2003 and 2004 publications, Hadoop and HDFS emerged as a free open source equivalent, which use the same MapReduce computational abstraction

Hadoop and Spark

- We'll focus on Spark in week 3, and we'll see advanced Spark algorithms in week 11
- Spark is a newer alternative to Hadoop
  - It can perform the same MapReduce tasks as Hadoop, often with a speed-up of 10x to 100x
  - It also performs a more general set of computations which generalise MapReduce, whilst maintaining the same high performance

Hadoop and Spark Popularity Over Time



Attendance: 154797

Google's Datasets

- Google builds its search engine by using a piece of software known as a "web crawler" to explore the entire web continually in order to find pages, which are then added to a giant index
  - As you might imagine, this generates a lot of data
- The web crawler is called Googlebot
- For more information, see:
  - https://developers.google.com/search/docs/fundamentals/how-search-works
  - https://developers.google.com/search/docs/crawling-indexing/googlebot

Datasets and Data Centre Node Specs

- From Google's MapReduce paper, as of 2004:
  - Google had multiple data centres
  - Each data centre had hundreds or thousands of servers
  - Hard disk capacity per server: 320GB
  - Memory per server: 4GB
  - Network bandwidth per server: 100-200Gbps
  - Average file size: 113GB
- Data was largely generated by the Googlebot web crawler, as we'll see
- Note that individual files were usually small enough to fit onto a single server (113GB vs 320GB).
  - However, the entire file store was far too large to fit onto a single server

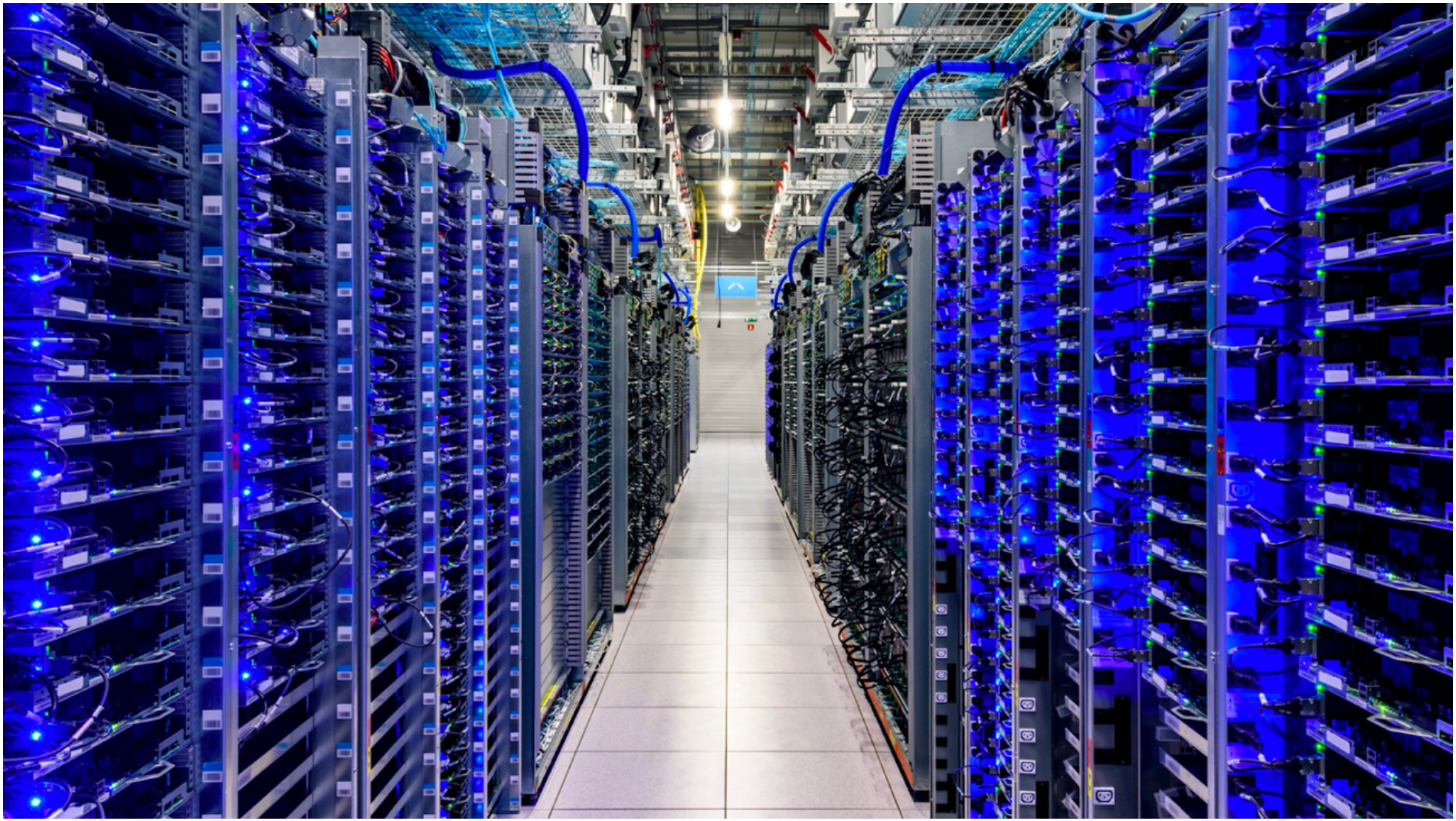
Datasets and Data Centre Node Specs

Number of jobs	29,423
Average job completion time	634 secs
Machine days used	79,186 days
Input data read	3,288 TB
Intermediate data produced	758 TB
Output data written	193 TB
Average worker machines per job	157
Average worker deaths per job	1.2
Average map tasks per job	3,351
Average reduce tasks per job	55
Unique *map* implementations	395
Unique *reduce* implementations	269
Unique *map/reduce* combinations	426

Table 1: MapReduce jobs run in August 2004

AWS Data Centre External View

Google Data Centre Inside View

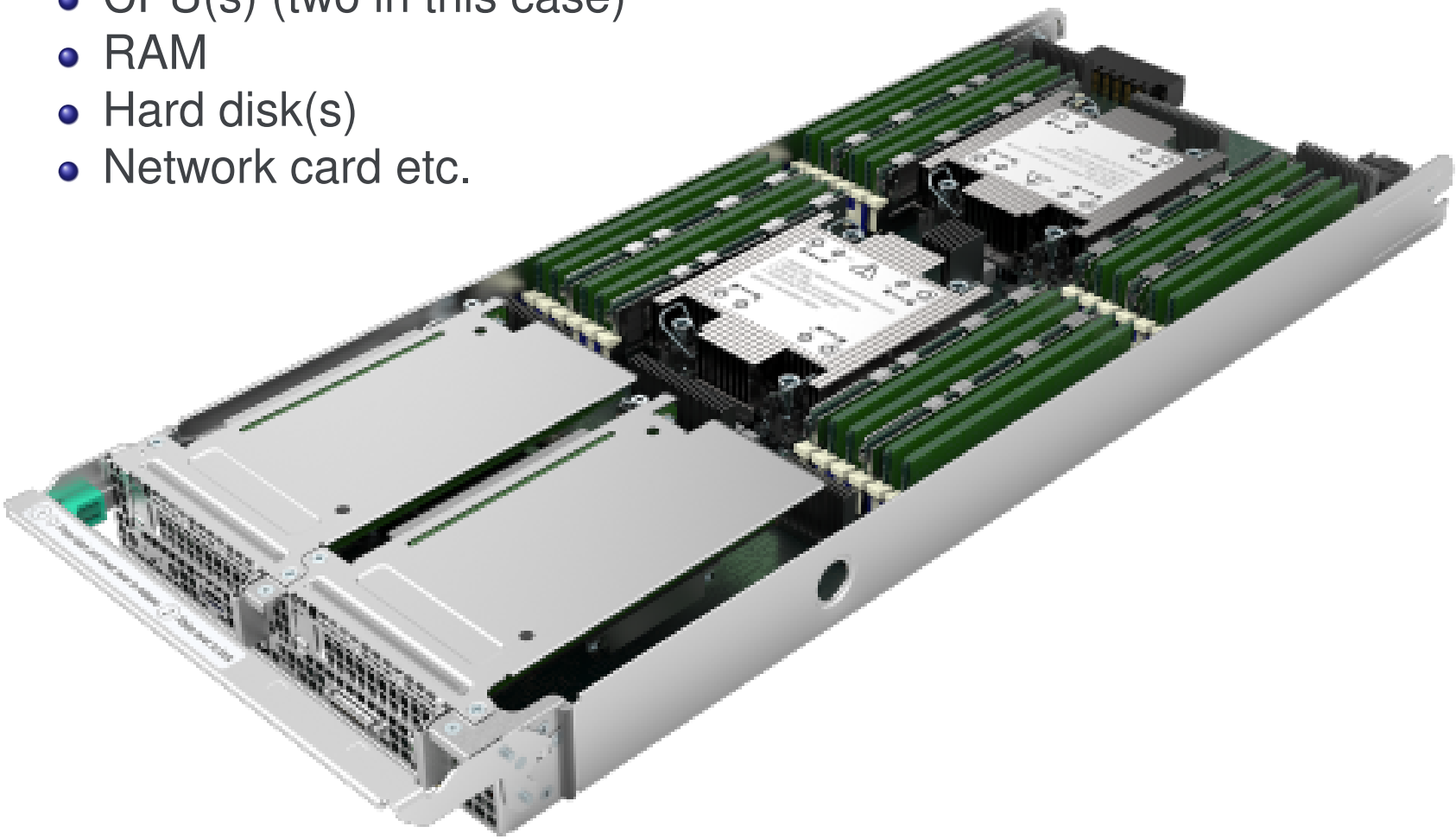


Server Rack

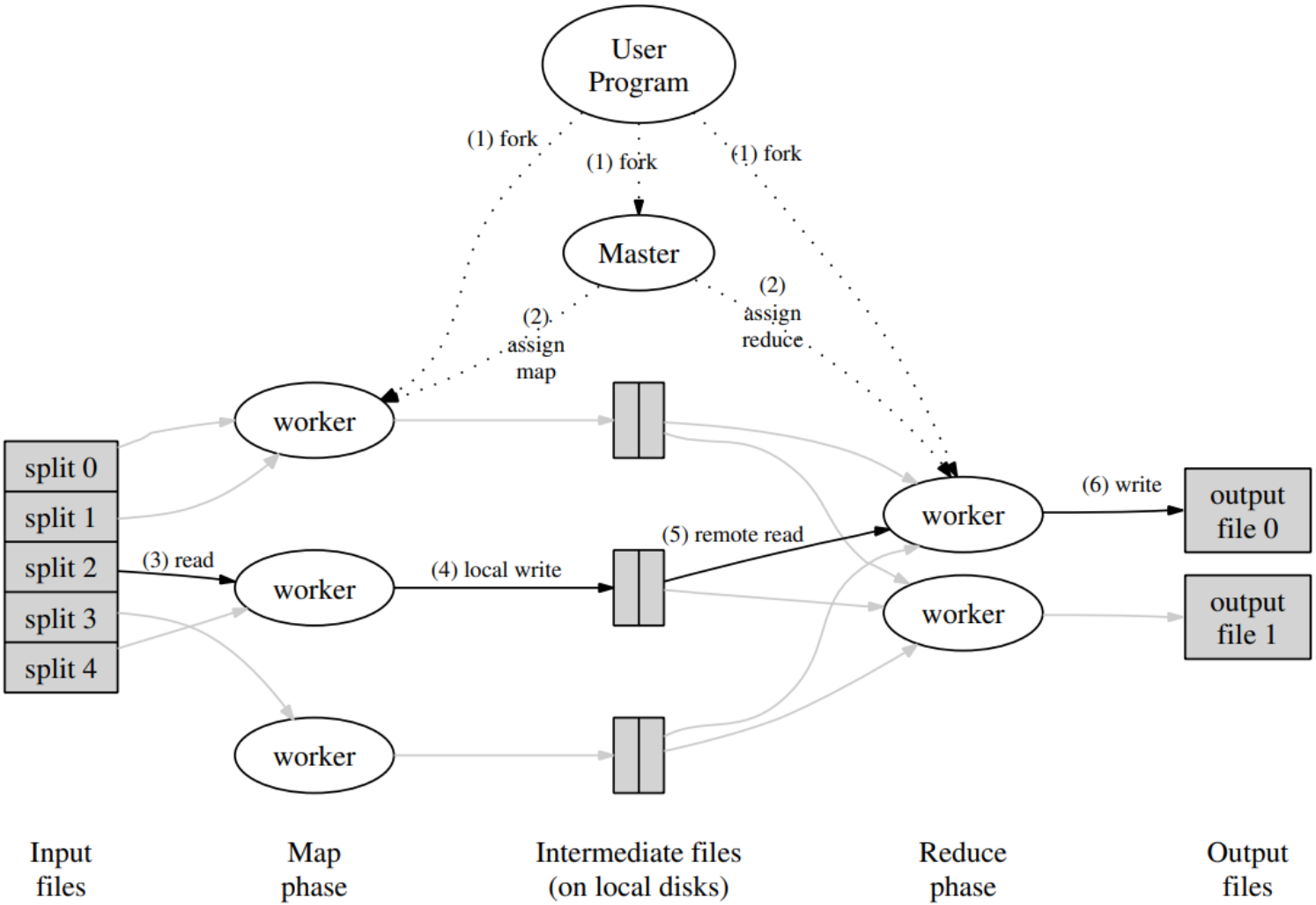
- Server racks hold servers, storage units, and switches

Rack-Mounted Server

- Contains inexpensive "commodity" hardware, similar to a desktop PC:
  - CPU(s) (two in this case)
  - RAM
  - Hard disk(s)
  - Network card etc.



The MapReduce Programming Model



Google File System

- Google's MapReduce is supported by a distributed file system called Google File System (GFS)
- When data is loaded onto the system it is divided into blocks
- Each block is stored on multiple different servers (typically 3)
- Files in GFS are write-once only, although data can be appended
- GFS is optimised for streaming reads of large files, rather than random access
- MapReduce tasks are divided into two phases:
  - Map tasks which are done concurrently on each block of data in a file, with each block being mapped over using the CPU of one of the servers on which it's stored
  - Reduce tasks then combine data to produce the final output

Advantages of Google File System

Advantages of Google's storage approach:

- It's resilient to server failure, because each block is stored on multiple servers
  - This is important, because failures will be frequent with thousands of servers in a cluster
- It allows parallel processing of the data, because the blocks of a single file are stored on different servers, each with its own CPU(s)
- It has huge capacity, because the file system uses hundreds or thousands of servers, each with hundreds of TBs of hard disk space

Advantages of Google File System

- It allows data sharing, because any user with access to GFS has access to the data, and can analyse it with MapReduce
- It's scalable – it scales linearly to handle larger data by adding more servers to the cluster
- Provides a clean abstraction for parallel programming:
  - MapReduce isn't essential for operating on big data, but it makes it easier
  - Google developed MapReduce to simplify big data tasks that engineers were previously implementing in more manual, *ad hoc*, ways.

Using MapReduce for Smaller Datasets

- In Google's original white paper, individual files were usually small enough to fit onto a single server (113GB vs 320GB)
  - However, the entire file store was far too large to fit onto a single server
- Big data middleware is useful for large files that can still technically fit onto a single server.
  - You still get the advantages of resilience, parallel execution of MapReduce, data sharing, etc.
- However, it's less well-suited for very small files:
  - GFS block size is 64MB, so files smaller than this will be stored as a single block, meaning that you'll get no parallel processing speed-up when using MapReduce

Attendance: 154797

Motivating Examples

- In this part, we'll get an idea of just how useful MapReduce is for applying data processing to big data.

Word Counting

- Consider the problem of counting the number of occurrences of each word in a large collection of documents
  - This is an important step for the natural language processing that we'll do in week 9
  - We'll use these word counts to create document representations, and these will form the basis of our search engine designs

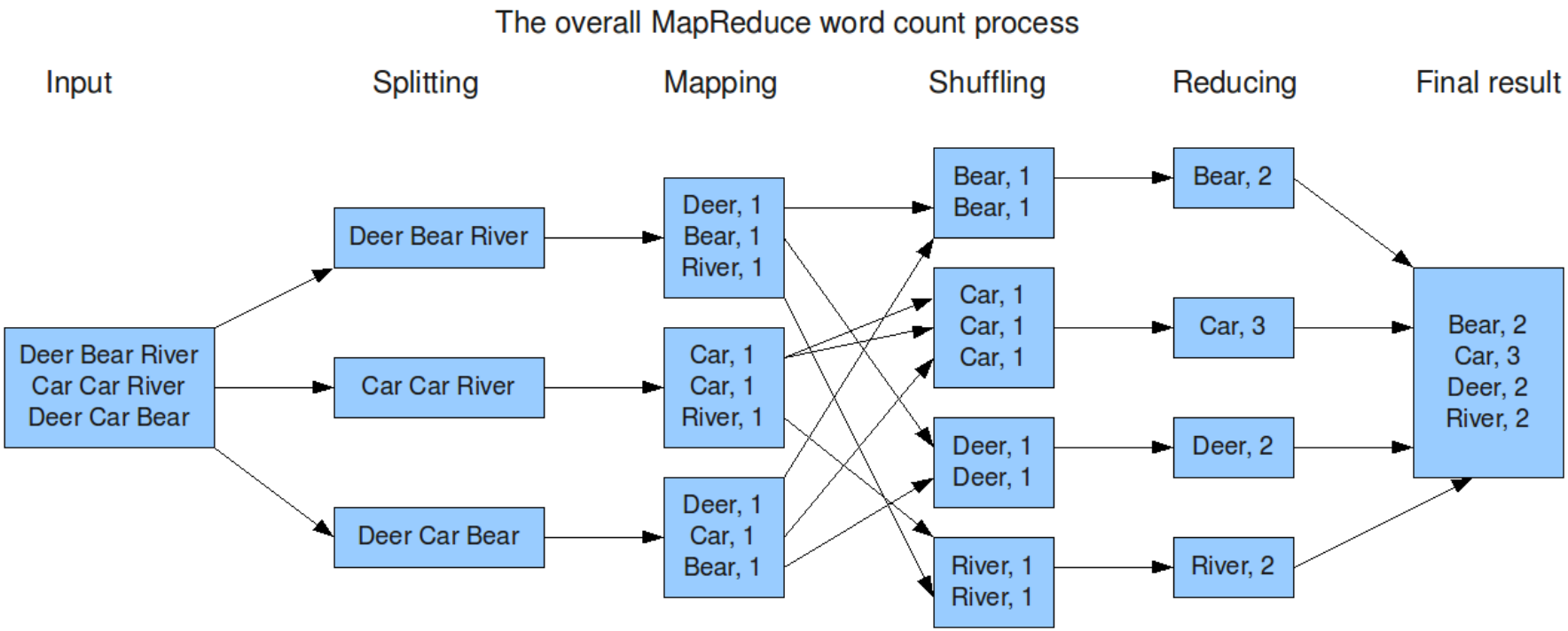
Word Counting

- MapReduce pseudo-code for this problem:

```
map(String key, String value):
  // key: document name
  // value: document contents
  for each word w in value:
    EmitIntermediate(w, "1");

reduce(String key, Iterator values):
  // key: a word
  // values: a list of counts
  int result = 0;
  for each v in values:
    result += ParseInt(v);
  Emit(AsString(result));
```

MapReduce: Word Count



Count of URL Access Frequency

- Dataset: a log of web page requests
- Output: a count of the number of requests for each web page
- MapReduce implementation:
  - The map function processes logs of web page requests and outputs (URL, 1) for each web request
  - The reduce function adds together all values for the same URL and emits a (URL, total count) pair

Reverse Web-Link Graph

- Dataset: a file of web pages and the URLs that each one links to
- Output: a mapping from each web page to the list of the web pages that link to it
- MapReduce implementation:
  - The map function outputs (target, source) pairs for each link to a target URL found in a page named source.
  - The reduce function concatenates the list of all source URLs associated with a given target URL and emits the pair: (target, list[source])

Term-Vector per Host

- Dataset: A collection of text files
- Output: A term vector summarises the most important words that occur in a document or a set of documents as a list of `<word, frequency>` pairs.
- MapReduce implementation:
  - The map function emits a `<hostname, term vector>` pair for each input document (where the `hostname` is extracted from the URL of the document).
  - The reduce function is passed all per-document term vectors for a given host. It adds these term vectors together, throwing away infrequent terms, and then emits a final `<hostname, term vector>` pair.

Inverted Index

- Dataset: A collection of text files
- Output: A mapping from each word to the list of documents that contain the word
- MapReduce implementation:
  - The map function parses each document, and emits a sequence of (word, document ID) pairs.
  - The reduce function accepts all pairs for a given word, sorts the corresponding document IDs and emits a (word, list(document ID)) pair.

Distributed Grep

- Dataset: a collection of text files
- Output: for a given regular expression, the set of lines that match it
- MapReduce implementation:
  - The map function emits a line if it matches a supplied pattern.
  - The reduce function is an identity function that just copies the supplied intermediate data to the output.

Attendance: 154797

Hadoop

- Open source software framework, providing the functionality of Google's MapReduce and GFS
- Created by Doug Cutting and Mike Carafella in 2005, based on Google's 2003 and 2004 white papers
- Cutting named the program after his son's toy elephant.

Who Uses Hadoop?

Attendance: 154797

HDFS

- HDFS is Hadoop's implementation of the file system concept outlined by Google in their 2003 GFS white paper
- As with GFS:
  - HDFS works best with large files
    - Typically GBs per file
  - Files in HDFS are write-once only, but can be appended to
  - HDFS is optimised for streaming reads of large files and not random reads
- Files are split into 64MB blocks (for Hadoop v1) or 128MB blocks (for Hadoop v2)
- Each block is replicated across several servers (default 3)

HDFS Implementation Details

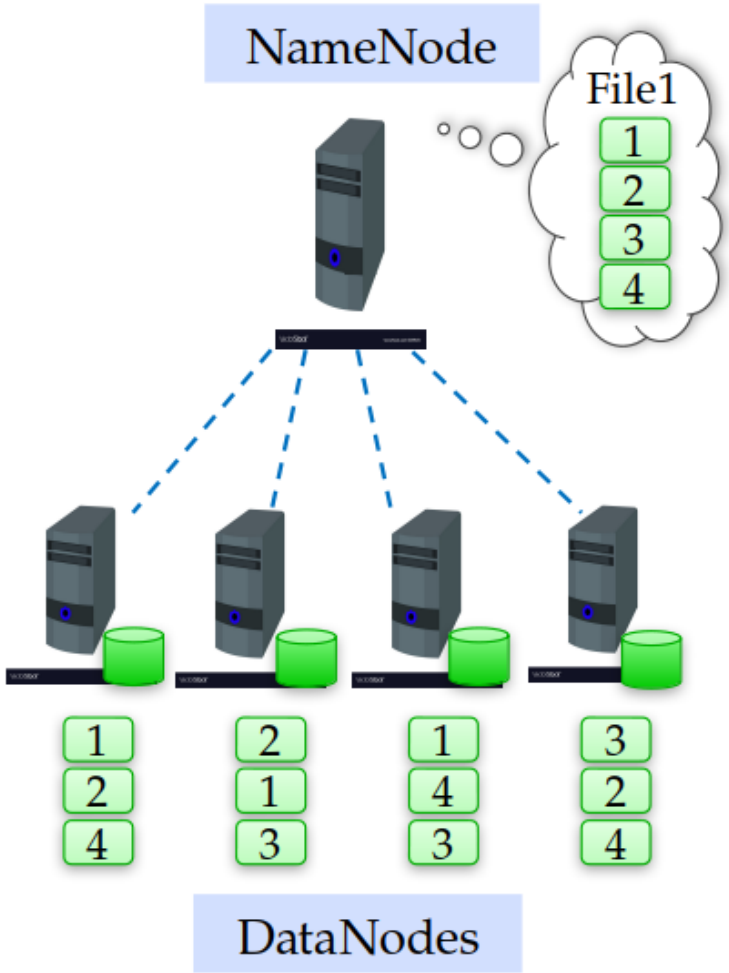
Hadoop clusters include the following daemons (i.e., programs) which work together to maintain HDFS:

- **NameNode**:
  - Runs on the (unique) master server.
  - Acts as the bookkeeper for HDFS.
  - Keeps track of where blocks constituting copies of each file are stored.
  - Monitors the health of the distributed file system.
  - Is a single point of failure for the cluster (so it is commonly on a dedicated machine).

HDFS Implementation Details

- **DataNodes**:
  - Runs on each worker node.
  - Stores the actual blocks (the client can upload files to DataNodes).
  - Informs the NameNode of any local changes to files.
  - Polls the NameNode for any required changes to blocks.
- **Secondary NameNode**:
  - Takes snapshots of the NameNode's metadata.
  - Purpose is to minimise the impact of NameNode failure.

HDFS Implementation Details



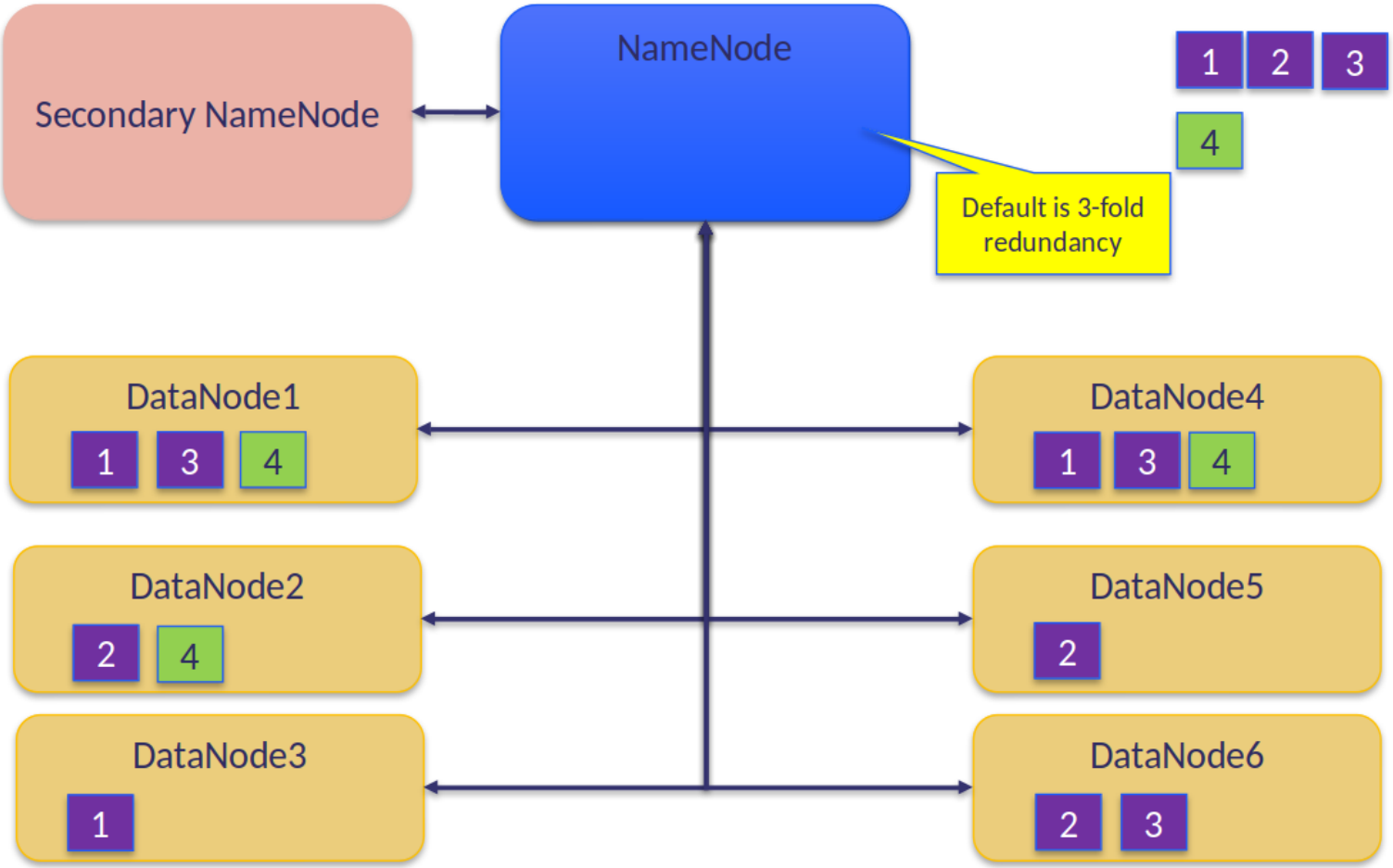
The NameNode

- Stores the HDFS file system information in the FsImage.
  - FsImage is a file stored on the OS filesystem that contains the complete directory structure (namespace) of the HDFS, with details about the location of the data on the Data Blocks and which blocks are stored on which node.
  - This file is used by the NameNode when it is started.
- Updates to the file system (add/remove blocks) do not change the FsImage.
  - They are instead written to a log file.
- When starting, the NameNode loads the FsImage file and then applies the changes in the log file.

The Secondary NameNode

- Not a backup for the NameNode
- Periodically reads the log file and applies the changes to the FsImage file, bringing it up to date
- Allows the NameNode to restart faster when required

HDFS Storage Example



Data Retrieval

- When a client wants to retrieve data:
  - Communicates with the **NameNode** to determine which blocks make up a file and on which data nodes those blocks are stored
  - Then communicates directly with the data nodes to read the data
  - So, no data passes through the master server

Hadoop MapReduce

- Written in Java
- Performs the MapReduce operation on HDFS data

The Mapper

- Reads data as key/value pairs
  - The key is often discarded
- Outputs zero or more key/value pairs
- E.g., for word count:

```
public static class TokenizerMapper
    extends Mapper<Object, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context
                    ) throws IOException, InterruptedException {
        StringTokenizer itr = new StringTokenizer(value.toString());
        while (itr.hasMoreTokens()) {
            word.set(itr.nextToken());
            context.write(word, one);
        }
    }
}
```

Shuffle and Sort

- Shuffle: copies the output from each Mapper using HTTP across the network.
- Sort: merge sorts Reducer inputs by key(s) (since different Mappers may have output the same key).
- As a result:
  - Output from the Mapper is sorted by key
  - All values with the same key go to the same worker server

The Reducer

- Called once for each unique key
- Gets a list of all values associated with a key as input
- The reducer outputs zero or more final key/value pairs
  - Usually just one output per input key
- E.g., for word count:

```
public static class IntSumReducer
    extends Reducer<Text,IntWritable,Text,IntWritable> {

    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values,
                       Context context) throws IOException, InterruptedExceptio
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        result.set(sum);
        context.write(key, result);
    }
}
```

Attendance: 154797

Orchestration of Hadoop MapReduce

- "Orchestration" refers to the process of implementing MapReduce computations on a Hadoop cluster
- Hadoop refers to each MapReduce call as a "job":
  - This "job" is broken down into multiple "tasks"
  - "Tasks" are then performed in parallel

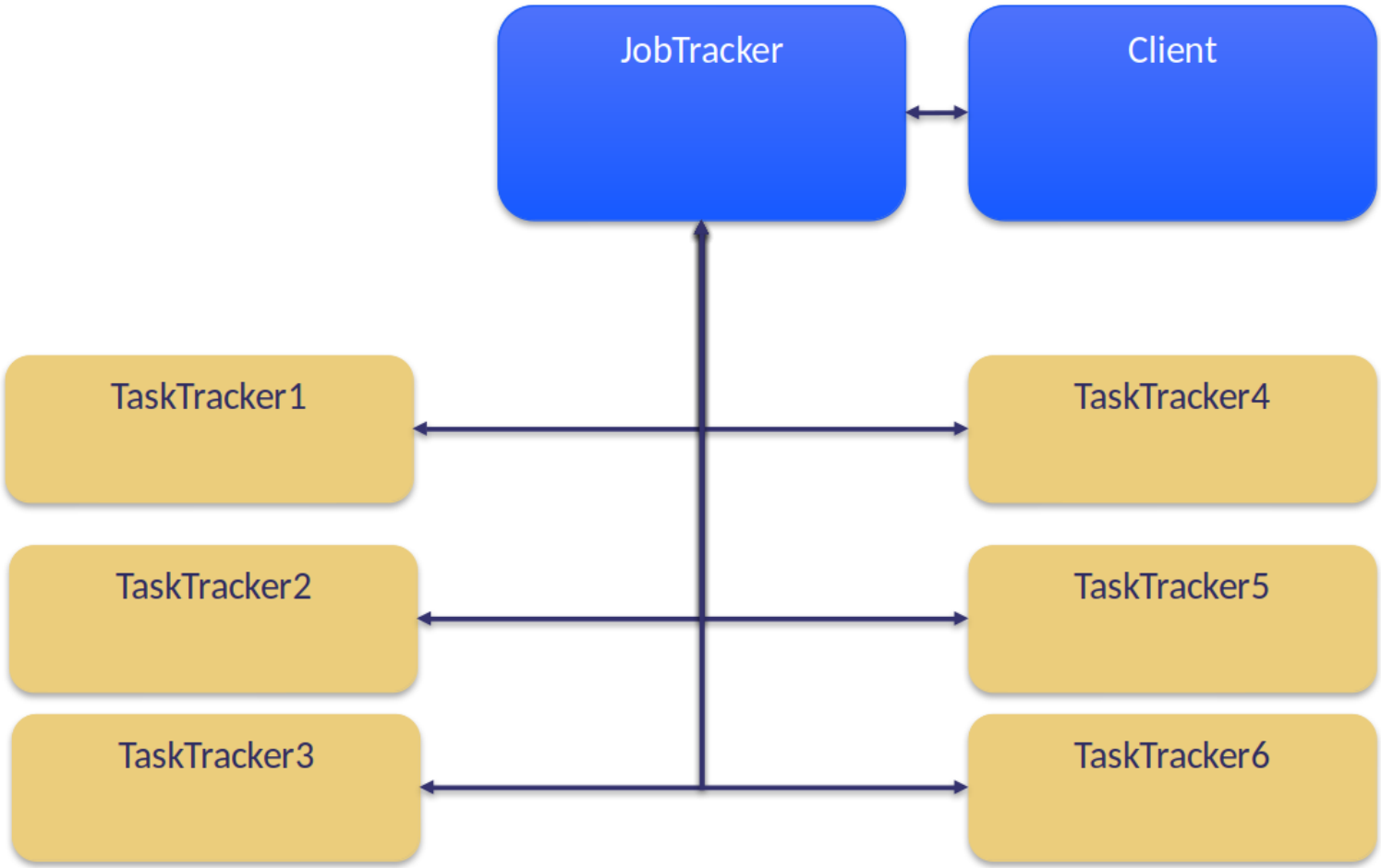
Orchestration of Hadoop MapReduce

- Hadoop clusters running **MapReduce v1** include the following daemons:
- **JobTracker**:
  - Runs on the single master server
  - The interface to clients wishing to run MapReduce jobs
  - Determines which computers do what and when
  - Will reassign (map and reduce) “tasks” to computers if tasks fail
    - Including if it stops hearing the heartbeat signal
- **TaskTrackers**:
  - Run on each worker server
  - Spawns (multiple) Java Virtual Machines (JVMs) to execute tasks
  - Maintains a heartbeat with JobTracker

Orchestration of Hadoop MapReduce

- In **MapReduce v2**:
  - **JobTracker** was replaced by Resource Manager and Application Master
  - **TaskTracker** was replaced by Resource Manager, Application Manager, and Application Master

JobTracker and TaskTracker



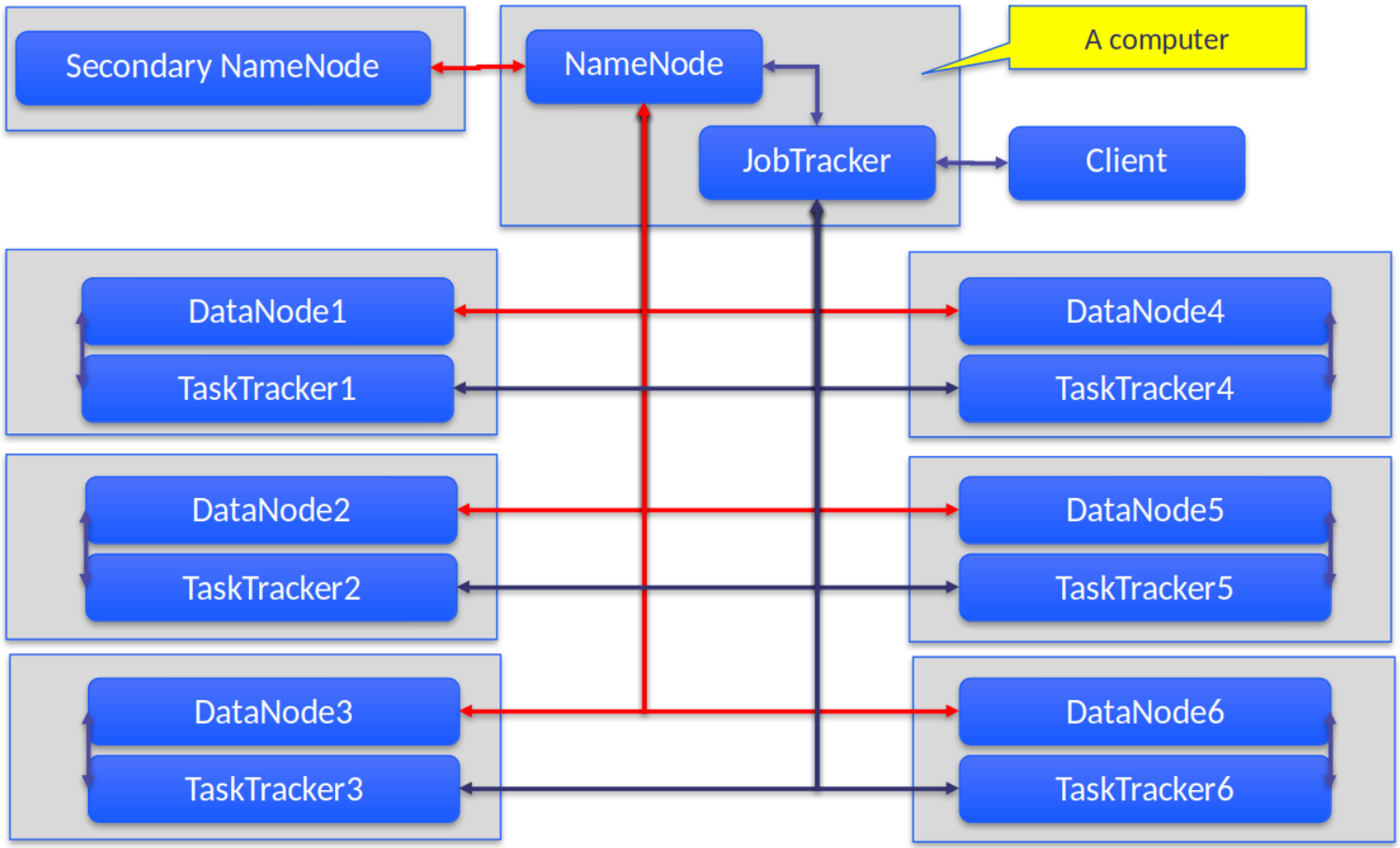
Execution Modes

- Hadoop can run in three execution modes:
  1. Fully distributed
  2. Pseudo-distributed
  3. Local

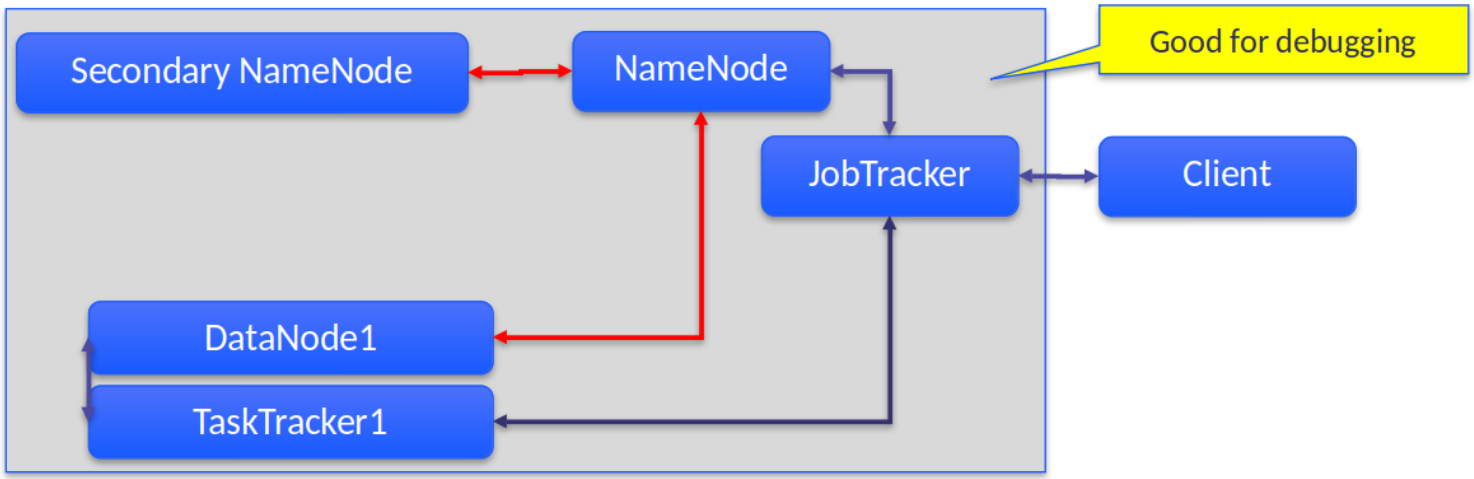
Fully Distributed Mode

- The following diagram represents a fully distributed mode in Hadoop.
  - The NameNode and Secondary NameNode handle metadata.
  - DataNodes store the actual data.
  - JobTracker and TaskTrackers manage the computational tasks.
- Key Components:
  - NameNode: Manages file system namespace and access.
  - DataNodes: Store data and report to NameNode.
  - JobTracker: Schedules and tracks jobs.
  - TaskTrackers: Execute tasks assigned by JobTracker.

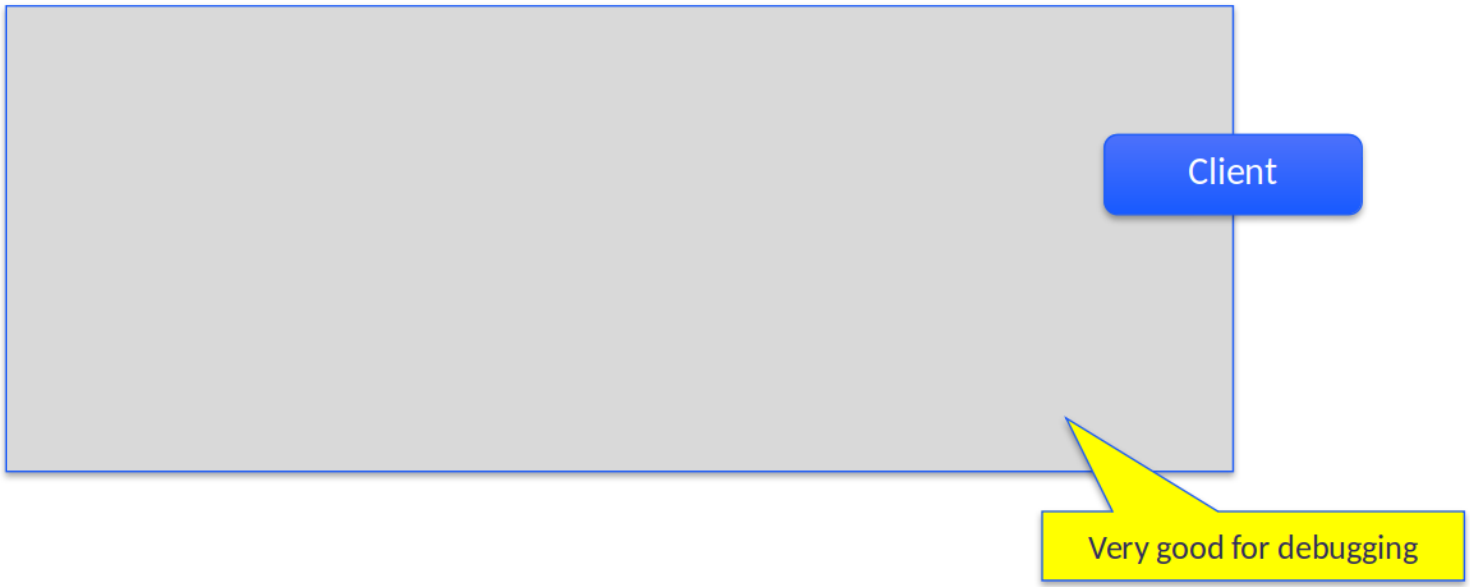
Fully Distributed Mode



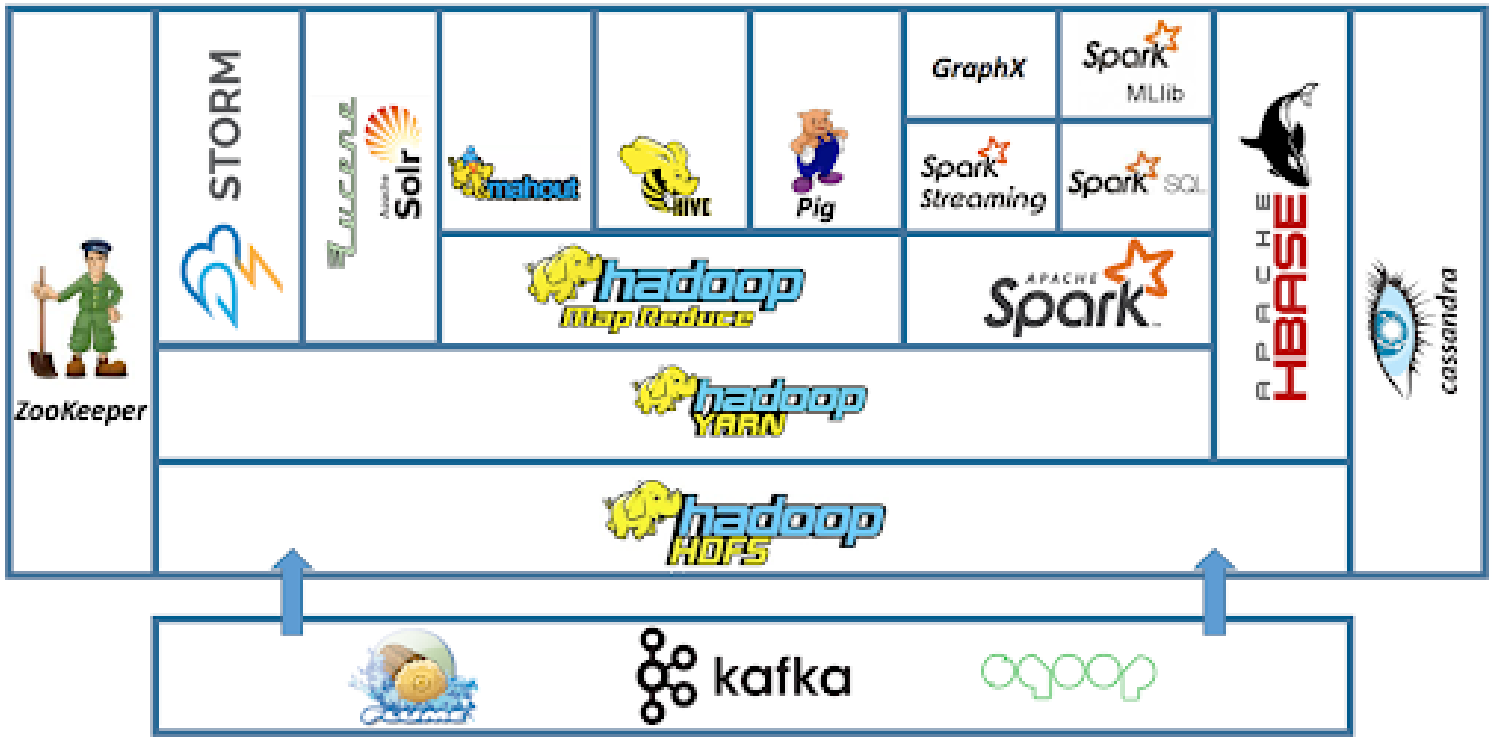
Psuedo-Distributed Mode



Local Mode



The Hadoop V2 Ecosystem



Spark

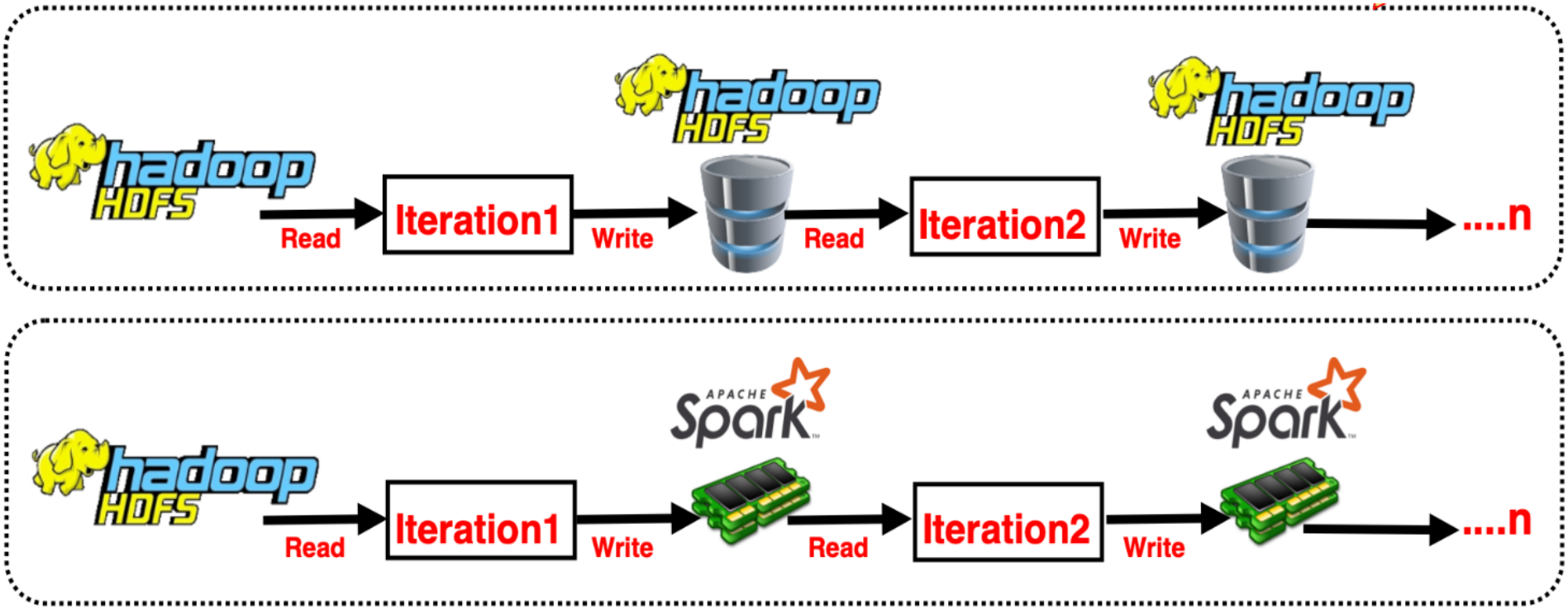
Attendance: 154797

Spark

- Developed as an improvement of Hadoop MapReduce.
- Works with any Hadoop-supported storage system (HDFS, S3, Avro, etc.), but not a modified version of Hadoop.
- Generally faster than Hadoop MapReduce:
  - Spark can perform in-memory processing, while Hadoop MapReduce has to read from/write to a disk (100x faster)
- Supports a broader set of computations than Hadoop
- APIs in Java, Scala, Python, and R (often 2-10x less code vs Hadoop)

Hadoop Vs Spark

- Spark keeps data in-memory while Hadoop MapReduce writes to disk after each map and reduce operation.
- This makes Spark 10x-100x faster than Hadoop for MapReduce computations



Word Count in Spark and Hadoop

Spark also has a concise API, often resulting in far less code than Hadoop:

```java
public class WordCount {
  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context
                    ) throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
      }
    }
  }

  public static class IntSumReducer
       extends Reducer<Text,IntWritable,Text,IntWritable> {
    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values,
                       Context context
                       ) throws IOException, InterruptedException {
      int sum = 0;
      for (IntWritable val : values) {
        sum += val.get();
      }
      result.set(sum);
      context.write(key, result);
    }
  }

  public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();
    if (otherArgs.length < 2) {
      System.err.println("Usage: wordcount <in> [<in>...] <out>");
      System.exit(2);
    }
    Job job = new Job(conf, "word count");
    job.setJarByClass(WordCount.class);
    job.setMapperClass(TokenizerMapper.class);
    job.setCombinerClass(IntSumReducer.class);
    job.setReducerClass(IntSumReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    for (int i = 0; i < otherArgs.length - 1; ++i) {
      FileInputFormat.addInputPath(job, new Path(otherArgs[i]));
    }
    FileOutputFormat.setOutputPath(job,
      new Path(otherArgs[otherArgs.length - 1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
  }
}
```

Figure: Word Count in Hadoop

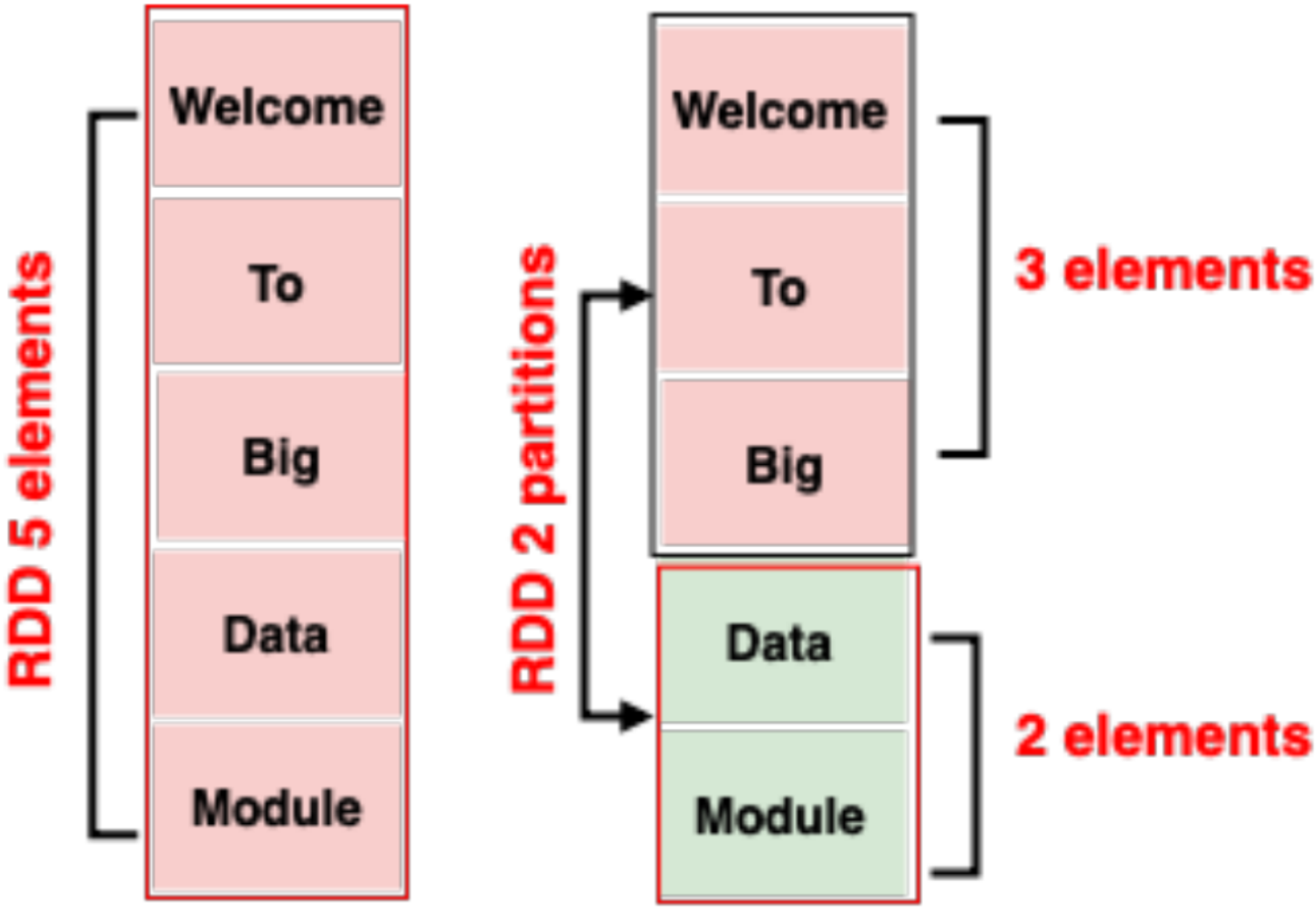
```scala
val f = sc.textFile(inputPath)
val w = f.flatMap(l => l.split(" ")).map(word => (word, 1)).cache()
w.reduceByKey(_ + _).saveAsText(outputPath)
```

Figure: Word Count in Spark

Resilient Distributed Datasets

- Data is represented in spark as **Resilient Distributed Datasets** (RDDs)
- RDDs are immutable collections of items/objects
- As we saw last week with HDFS, RDDs are are split into partitions and can be stored on the worker nodes of the Spark cluster
- RDD operations can be performed on partitions in parallel
  - **Resilient** because each RDD partition is stored on multiple nodes of a Spark cluster
  - **Distributed** because different RDD partitions are spread across the cluster nodes, and can be processed in parallel

RDD Partitions



Immutability

- RDDs are **immutable** meaning that once they're created they can's be modified.
- This means that every RDD transformation creates a new RDD (just like lists in Haskell)

RDD1

```
input_file = sc.textFile("/usr/local/spark/input.txt")
```

RDD2

```
map = input_file.flatMap(lambda line: line.split("
")).map(lambda word: (word, 1))
```

RDD3

```
counts = map.reduceByKey(lambda a, b: a + b)
```

```
counts.saveAsTextFile("/path/to/output/")
```

RDD4

Creating RDDs

There are two ways to create an RDD:

1. Parallelising an existing collection in your driver program:

```python
# Parallelize in Python
wordsRDD = sc.parallelize(["fish", "cats", "dogs"])
```

```scala
// Parallelize in Scala
val wordsRDD= sc.parallelize(List("fish", "cats", "dogs"))
```

```java
// Parallelize in Java
JavaRDD<String> wordsRDD = sc.parallelize(Arrays.asList("fish", "cats", "dogs"));
```

2. Reading from an external storage system or filesystem:

```
# Read a local txt file in Python
input_file = sc.textFile("/usr/local/spark/input.txt")
```

Installing and Using PySpark

Spark is very easy to run locally with Python using PySpark:

- Install PySpark using pip:

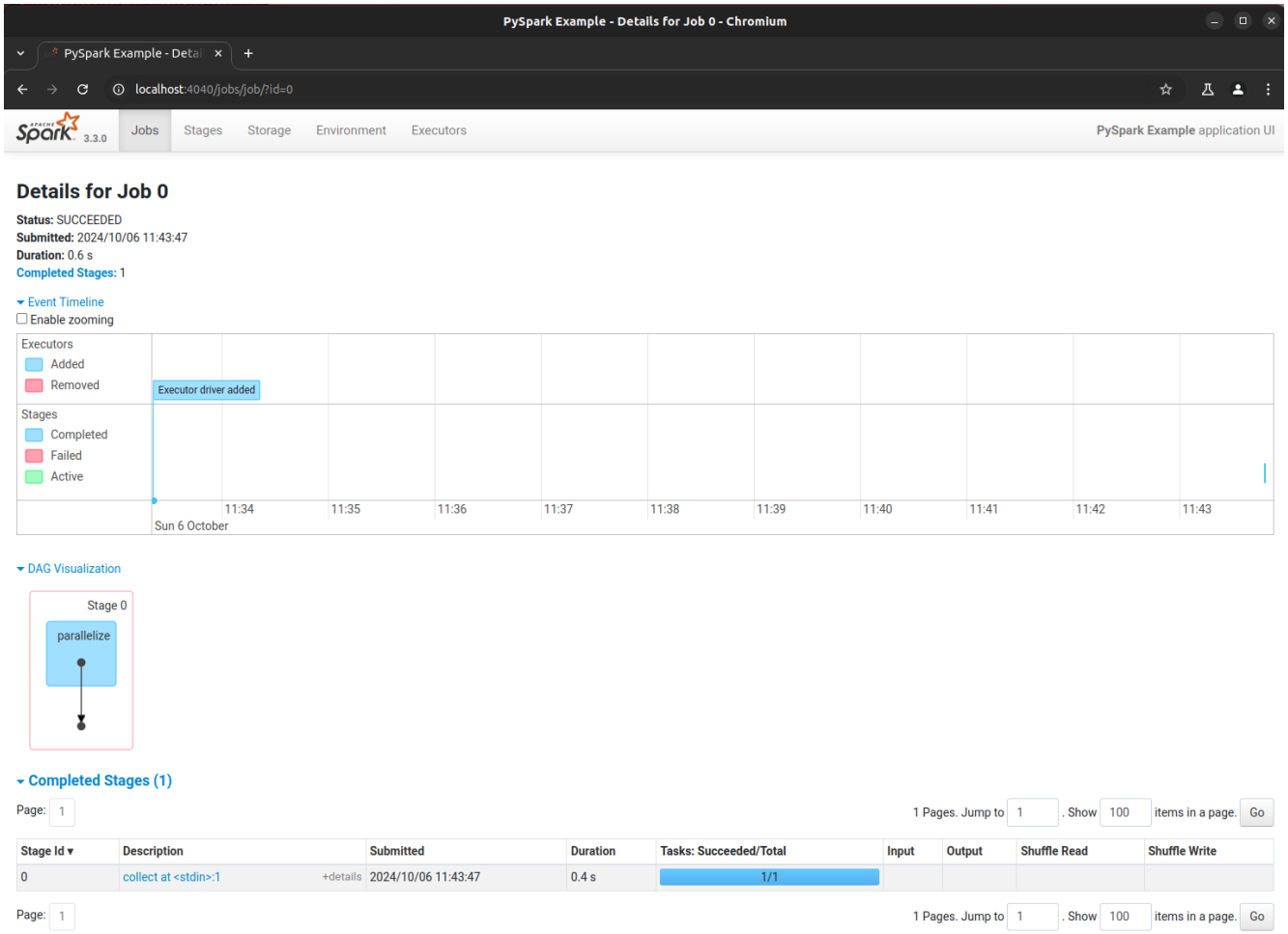
```
$ pip install pyspark
```

- Import PySpark, initialize a SparkContext and run some commands:

```
$ python
>>> from pyspark import SparkContext
>>> sc = SparkContext("local", "PySpark Example")
>>> data = [1, 2, 3, 4, 5]
>>> rdd = sc.parallelize(data)
>>> squared_rdd = rdd.map(lambda x: x * x)
>>> result = squared_rdd.collect()
>>> print(result) # Output: [1, 4, 9, 16, 25]
[1, 4, 9, 16, 25]
>>> sc.stop()
```

The Spark Application UI

Visiting **http://localhost:4040/**, we see:



The Spark Stack

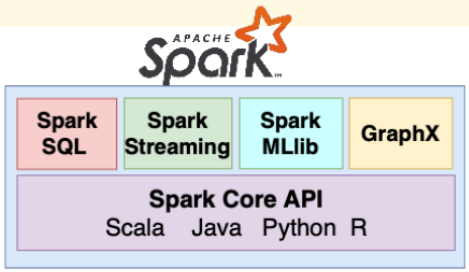
Spark Core, this is the heart of Spark, and is responsible for management functions such as task scheduling.

Spark SQL is a Spark module for structured data processing.

Spark Streaming is a framework for ingesting real-time streaming data from various sources.

Spark MLlib is a library of machine-learning algorithms (e.g., logistic regression, naïve Bayes classification, SVMs, decision trees, random forests, linear regression, and k-means clustering).

GraphX are data structures comprising vertices and the edges for graph and graph parallel computations.



Spark MLLib in Context

- A word of caution: Spark isn't always the best choice for a given machine learning application.
- Here are the alternatives:
  - **Spark MLLib:** Designed specifically to handle large-scale data processing across distributed computing environments. Ideal for big data applications.
  - **Python scikit-learn AKA sklearn:** A python library for machine learning. We'll use this in this course.
  - **R:** a language for statistics and machine learning.
  - **TensorFlow/PyTorch:** frameworks for designing and training deep neural networks.

Comparison of Machine Learning Frameworks

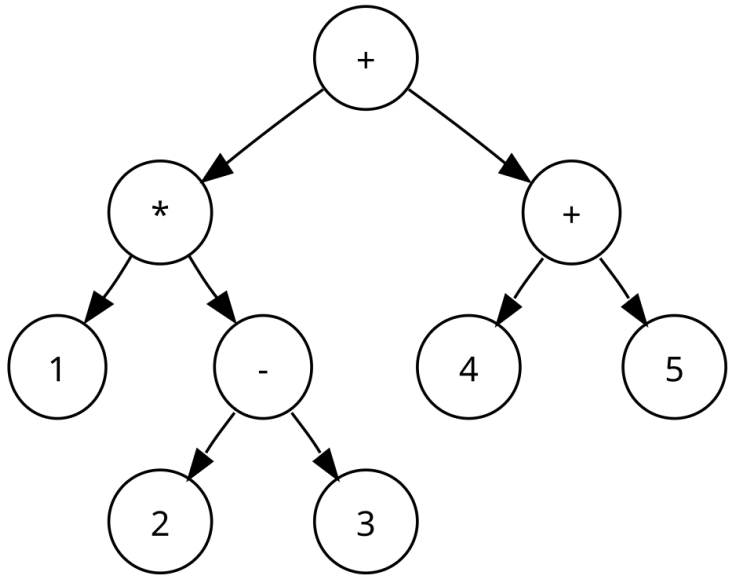
Framework	Dataset Size	ML Techniques
Spark	Large-scale	Basic
sklearn	Small to Medium	Basic, Intermediate
R	Small to Medium	Basic, Intermediate, Advanced
TF/PyTorch	Small to Large-scale	Deep Learning

- R is the gold standard for stats and machine learning on small to medium sized datasets (with the exception that deep learning is mostly done in Python with TensorFlow or PyTorch)
- We don't use it in this course because there's a (small) learning curve due to the fact that it provides its own programming language
- However, the language is very easy to learn, so R is ideal for anything most stats and ML projects

Attendance: 154797

Directed Acyclic Graphs and Data Flows

- Spark programming model is based on parallelizable operators, which can be composed together to form a **directed, acyclic graph (DAG)**
  - This is a directed graph with no directed cycles
  - E.g. here's a DAG for the trivial computation $1 * (2 - 3) + 4 + 5$:

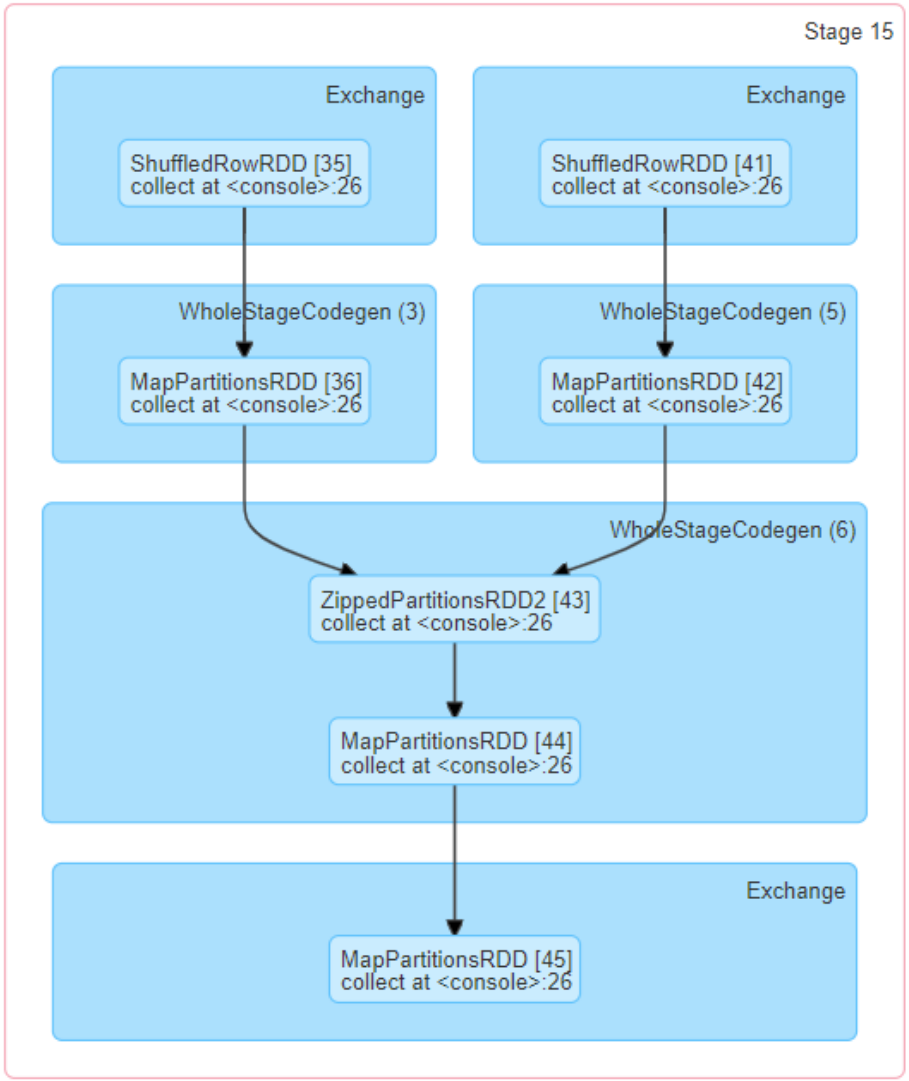


- A **data flow** is composed of any number of data sources, operators, and data sinks. This is done by connecting their inputs and outputs.

DAGs in Spark and Lazy Evaluation

- In Spark, DAGs are used to represent a sequence of computations performed on data, where
  - Each **node** in the graph is an RDD
  - Each **edge** in the graph is a transformation that creates a new RDD from an existing one
- The DAGs are evaluated **lazily** (just like functions in Haskell)
- This allows Spark to optimise the execution plan by reordering and combining transformations.
- This optimisation leads to efficient in-memory execution, which **explains why Spark is so much faster than Hadoop**

Directed Acyclic Graphs in Spark



Spark RDD Operations

Spark RDD support two types of operations: transformations and actions.

i. **Transformations** can be applied to an RDD to create a new RDD
   - These are the DAG edges
   - E.g. `x.map(f)` creates a new RDD by applying `f` to each element in x.

ii. **Actions** compute a result based on an RDD and either return a value or save it to an external storage system (e.g., HDFS).
   - E.g. `x.saveAsTextFile(path)` writes write one RDD element per line to `path`

Spark RDD Operations

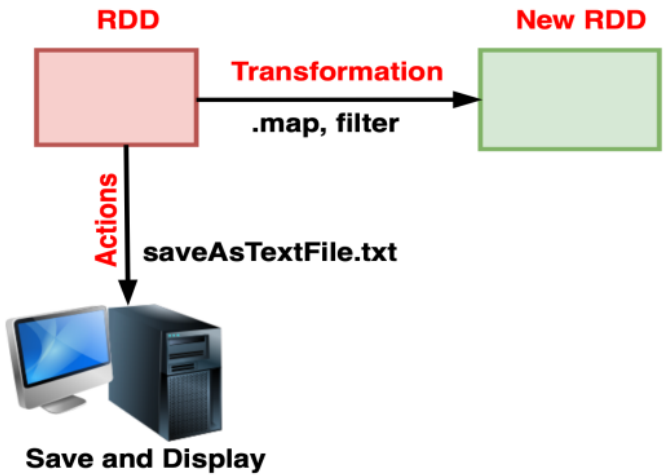
Transformation

```
input_file = sc.textFile("/usr/local/spark/input.txt") map
 = input_file.flatMap(lambda line: line.split("
")).map(lambda word: (word, 1))

counts = map.reduceByKey(lambda a, b: a + b)
```

```
counts.saveAsTextFile("/path/to/output/")
```

Action

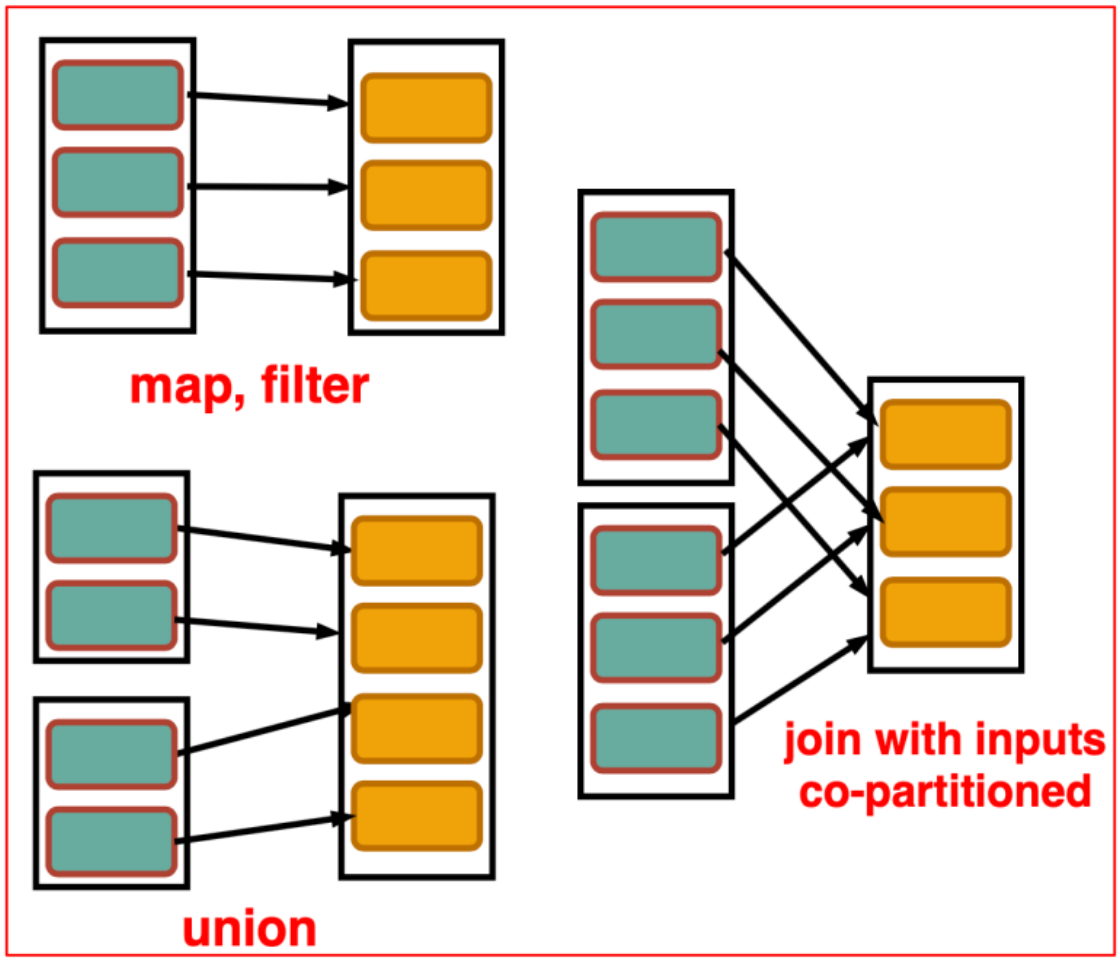


Narrow Vs Wide Transformation

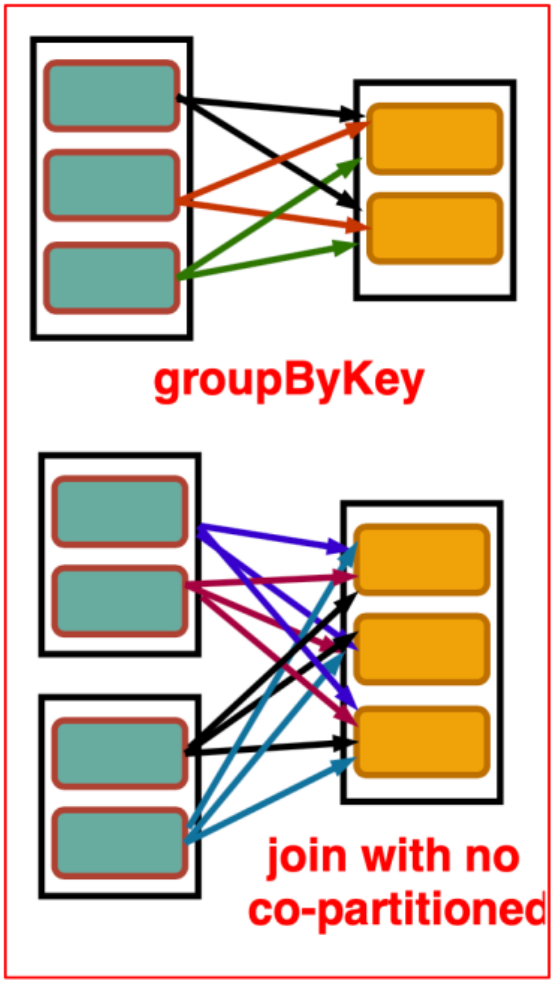
Transformations can be narrow or wide:

- **Narrow Transformation**: each partition of the parent RDD is used by at most one partition of the child RDD
- **Wide (AKA shuffle) transformations**: multiple child RDD partitions may depend on a single parent RDD partition

Narrow Transformations



Wide Transformations



RDD Operations in Spark

	General	Math / Statistical	Set Theory / Relational	Data Structure / I/O
Transformations	• map • filter • flatMap • mapPartitions • mapPartitionsWithIndex • groupBy • sortBy	• sample • randomSplit	• union • intersection • subtract • distinct • cartesian • zip	• keyBy • zipWithIndex • zipWithUniqueID • zipPartitions • coalesce • repartition • repartitionAndSortWithinPartitions • pipe
Actions	• reduce • collect • aggregate • fold • first • take • forEach • top • treeAggregate • treeReduce • forEachPartition • collectAsMap	• count • takeSample • max • min • sum • histogram • mean • variance • stdev • sampleVariance • countApprox • countApproxDistinct	• takeOrdered	• saveAsTextFile • saveAsSequenceFile • saveAsObjectFile • saveAsHadoopDataset • saveAsHadoopFile • saveAsNewAPIHadoopDataset • saveAsNewAPIHadoopFile

Highlighted operations are hard

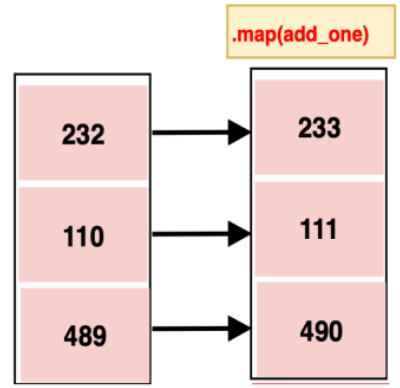
Attendance: 154797

map and flatMap

- **map:** takes as input a function which is applied to each element of the RDD and map each input item to another item

```
# transforming each element through a function

nums = sc.parallelize([232, 110, 489])
big_no = nums.map(lambda x: [n + 1])
```



If we want to map every record 1, .map(lambda x: [x,1]

(232, 1), (110,1), (489,1)

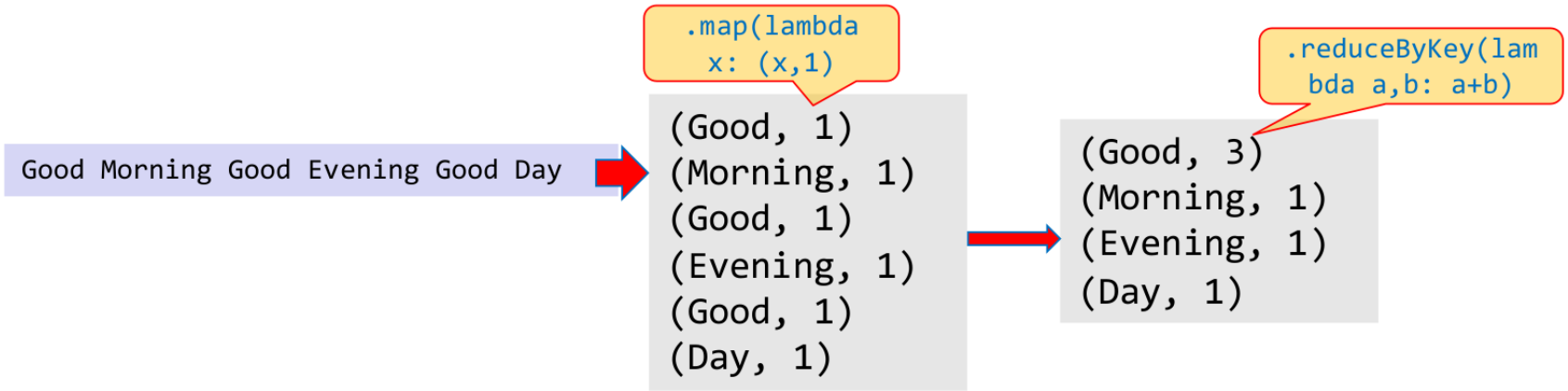
- **flatMap:** generates a new RDD by filtering the source dataset using the specified function

```
Good Morning Good Evening Good Day
Happy Birthday Happy New Year
```

```
lines.flatMap(lambda line: line.split()).collect()
['Good', 'Morning', 'Good', 'Evening', 'Good', 'Day', 'Happy',
'Birthday','Happy', 'New', 'Year']
```

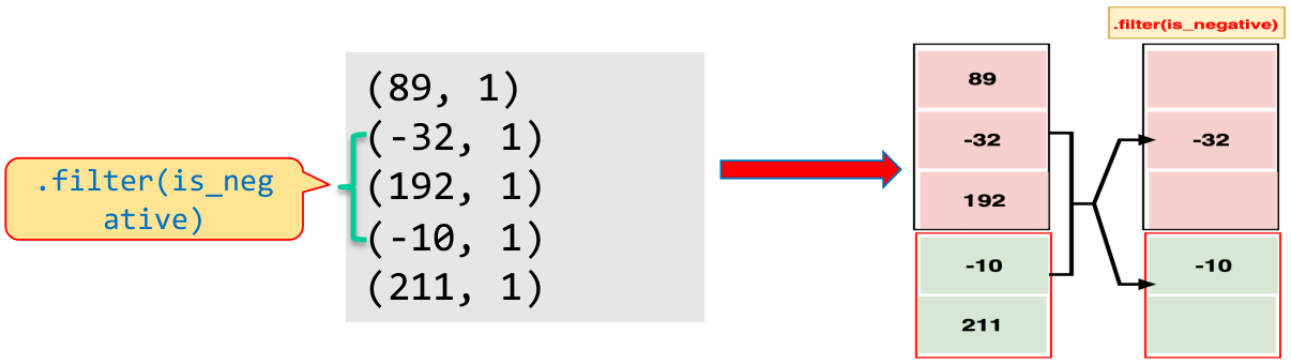
filter

filter generates a new RDD by filtering the source dataset using the specified function:



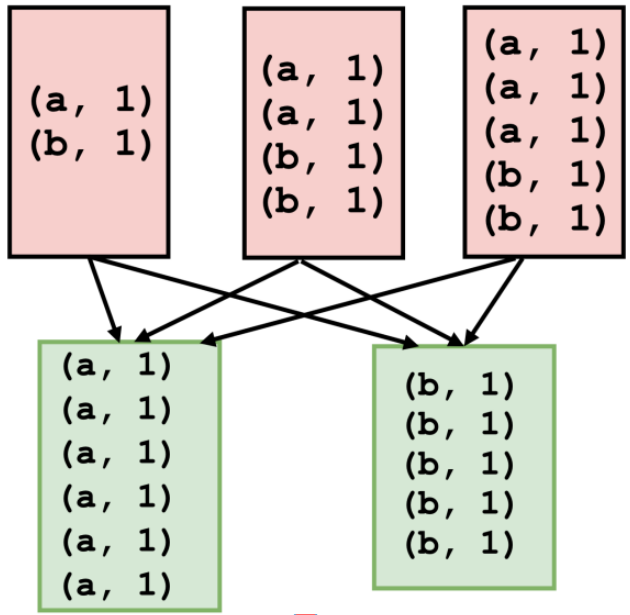

```
rdd6 = rdd5.filter (lambda x : 'a' in x[1])
```

="(1, 'Day')" that has a value 'a'.



groupByKey

groupByKey groups the data based on a key and returns an RDD of grouped data:



Note
For an aggregation (e.g., **sum or average**) over each key, using reduceByKey or aggregateByKey will provide much better performance.

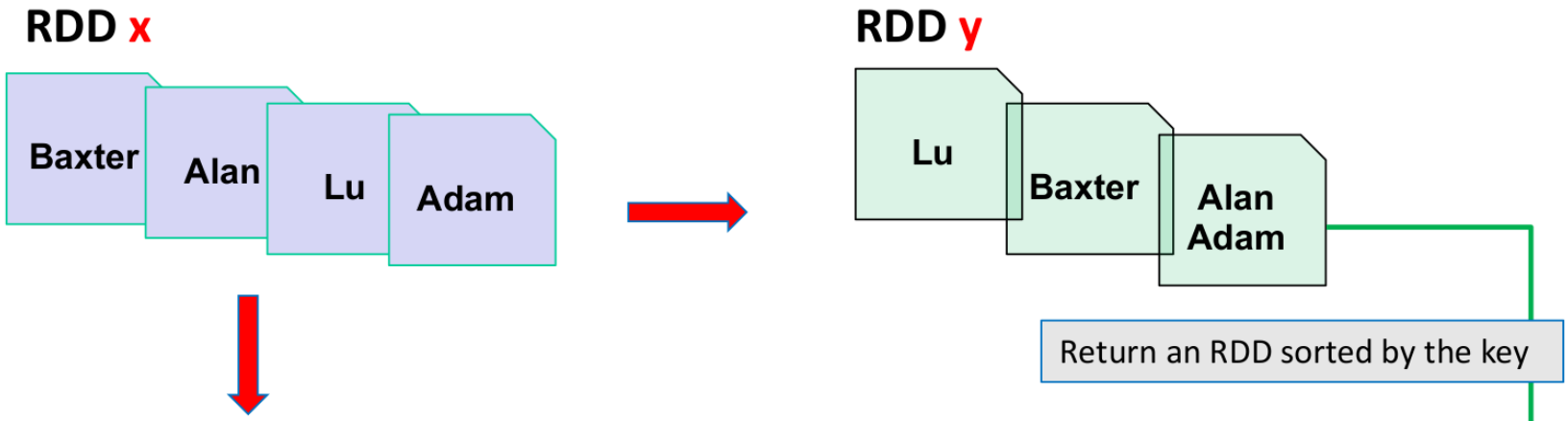
Group values with the same key

```
rdd = sc.parallelize([("a", 1), ("b", 1), ("a", 1), ("a", 1), ("b", 1), ("b", 1), ("a", 1), ("a", 1), ("a", 1), ("b", 1), ("b", 1)])
```

```
sorted(rdd.groupByKey().mapValues(list).collect()) [('a', [1, 1, 1, 1, 1, 1]), ('b', [1, 1, 1, 1, 1])]
```

groupBy

groupBy groups the data based on a user-provided function that is applied to each element in the RDD:



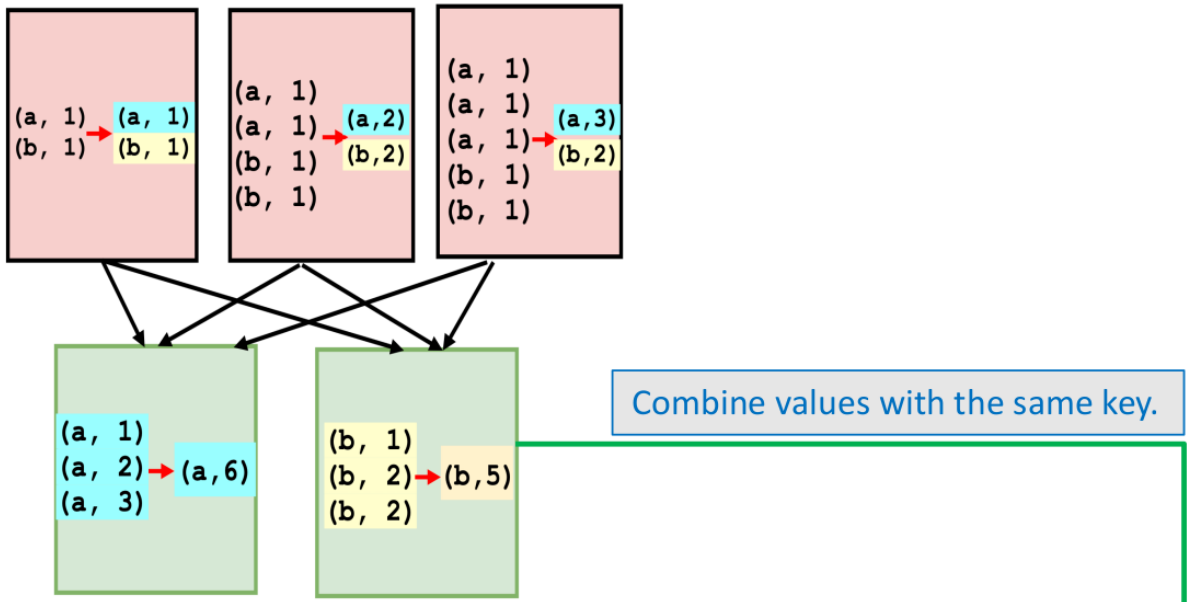

```
x = sc.parallelize(['Baxter', 'Alan', 'Lu', 'Adam'])
y = x.groupBy(lambda w: w[0])

print [(k, list(v)) for (k, v) in y.collect()]
```

```
x: ['Baxter', 'Alan', 'Lu', 'Adam']
y: [('A',['Alan, 'Adam']),('B',[Baxter']),('L',['Lu'])]
```

reduceByKey

reduceByKey aggregates elements of the RDD using a specified associative and commutative function:

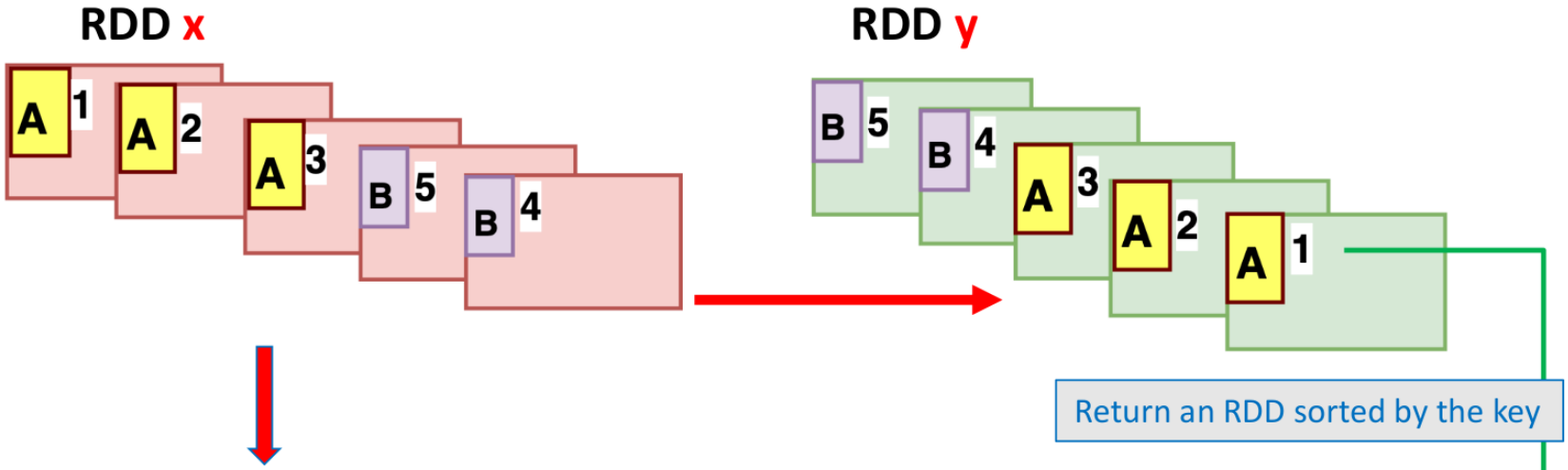



```
from operator import add
rdd = sc.parallelize([("a", 1), ("b", 1), ("a", 1), ("a", 1), ("b",
1), ("b", 1), ("a", 1), ("a", 1), ("a", 1), ("b", 1), ("b", 1)])
```

```
sorted(rdd.reduceByKey(add).collect()) [('a', 6), ('b', 5)]
```

sortByKey

sortByKey sorts elements by key:

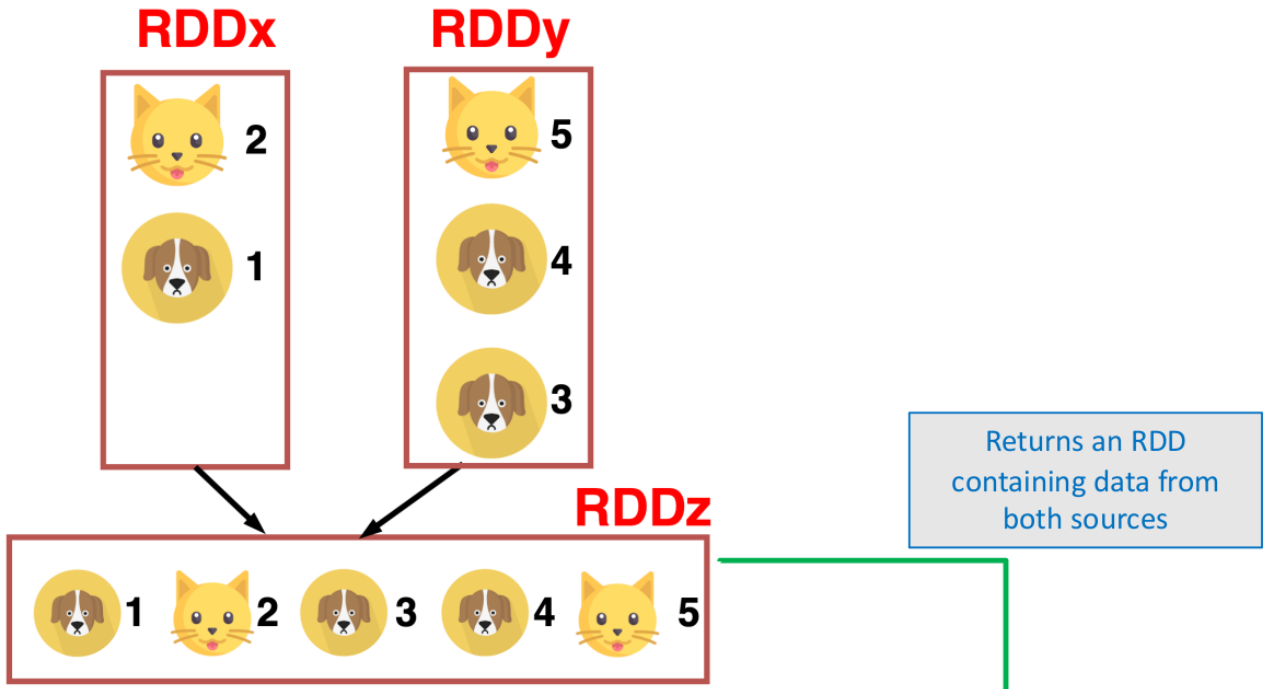



```
x = sc.parallelize([('B',4),('B',5),('A',3),('A',2),('A',1)])
y = x.sortByKey()
print(x.collect())
print(list((j[0], list(j[1])) for j in y.collect()))
```

```
x: [('B', 4),('B', 5),('A', 3),('A', 2),('A', 1)]
y: [('A', 1), ('A', 2), ('A',3), ('B', 4), ('B', 5)]
```

union

union returns the union of two RDDs:

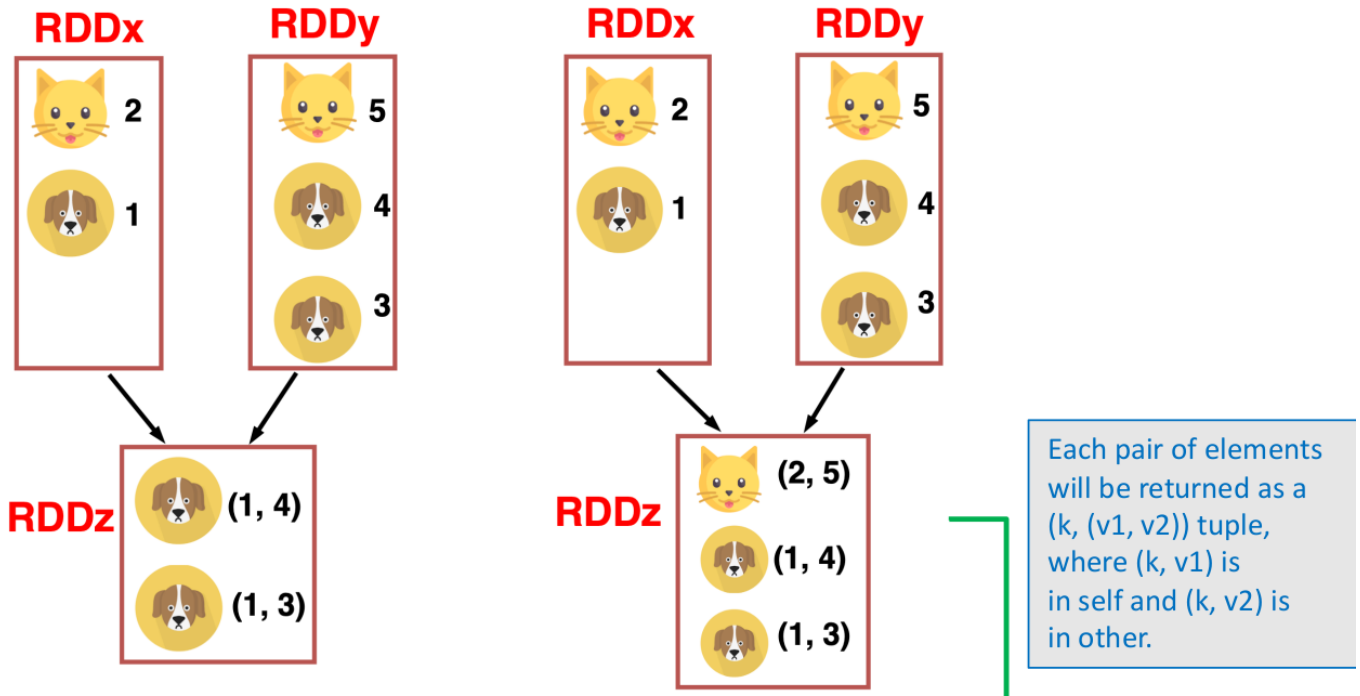



```
x = sc.parallelize([("dog",1), ("cat",2)])
y = sc.parallelize([("dog",3), ("dog",4), ("cat", 5)])
```

```
z =([("dog",1), ("cat",2), ("dog",3), ("dog",4), ("cat", 5)])
```

join

join returns the inner join of two RDDs:

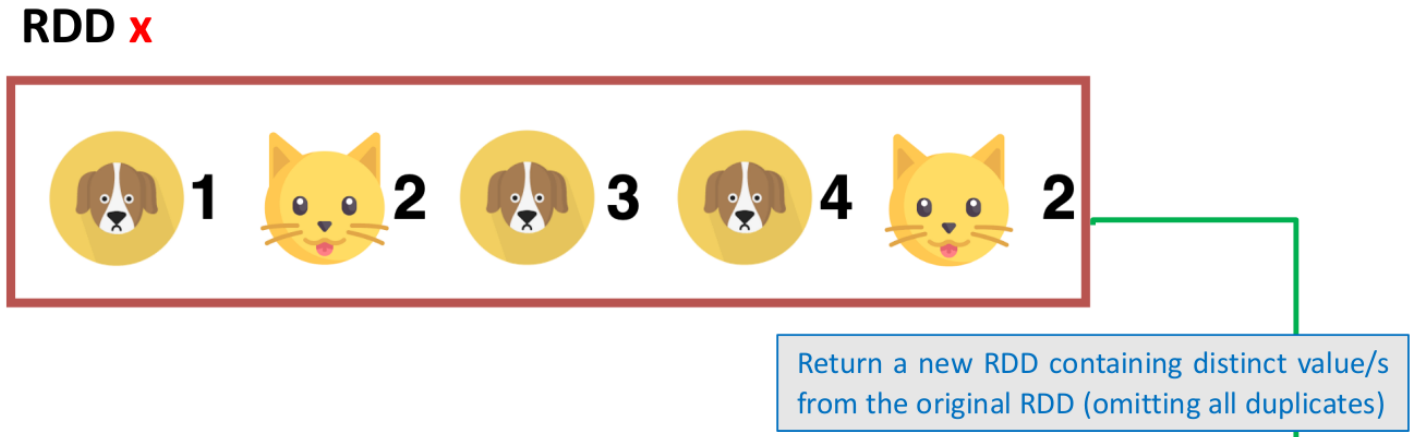



```
x = sc.parallelize([("dog", 1), ("cat", 2)])
y = sc.parallelize([("dog", 3), ("dog", 4), ("cat", 5)])
sorted(x.join(y ).collect())
```

```
z: [('dog', (1,3)), ('dog', (1,4), ('cat', (2, 5))]
```

distinct

distinct returns an RDD with distinct elements:




```
x = sc.parallelize([("dog", 1), ("cat", 2), ("dog", 3), ("dog", 4),
("cat", 2)])
y = x.distinct()
print(y.collect())
```

```
z = ([("dog", 1), ("cat", 2), ("dog", 3), ("dog", 4)])
```

Attendance: 154797

Actions

- **reduce** AKA **aggregate** applies a binary operator to an initial state and all elements in an RDD, and reduces this to a single state. The final state is converted into the final result by applying a finish function.
  - Note that reduceByKey is a transformation because it returns an RDD of keys and transformed values.
  - However, reduce is an action because it returns a single value rather than an RDD
- **fold** is similar to reduce. It aggregates the elements of each partition, and then the results for all the partitions. In contrast to reduce, it uses a neutral "zero value".
- **collect**: Returns all the elements of the dataset as a list (NB: avoid using with large datasets because it will dump the entire dataset into your local computer's memory).

Actions

- **first** returns the first element in this RDD.
- **take(n)** returns the first n elements from the dataset.
- **forEach(func)** applies a function to all elements of the RDD, generally used for side effects (e.g., saving data to a database). Similar to map, except that is is an action, so it does not produce a new RDD.
- **top(n)** returns the top n elements of the RDD, based on the default or specified ordering.

Actions

- **count** return the number of elements in this RDD.
- **takeSample(withReplacement, num, [seed])** returns a fixed-size sampled subset of this RDD.
- **max** returns the maximum element in the RDD (must be comparable).
- **min** returns the minimum element in the RDD (must be comparable).
- **sum** computes the sum of elements in the RDD.
- **histogram(buckets)** computes a histogram of the elements in the RDD, dividing them into buckets.
- **mean** returns the mean of the dataset.
- **variance** returns the variance of the dataset.
- **stdev** returns the standard deviation of the dataset.

Attendance: 154797

DataFrames

DataFrames are distributed collections of rows under named columns.

- Introduced to Spark in 2015
- Similar to Excel tables or Python DataFrames
- Build on top of RDDs
- Immutable in Nature: You can create a DataFrame, but you will not be able to change it. A DataFrame can be transformed just like an RDD.
- Lazy Evaluations: A task is not executed until an action is performed.
- Distributed: DataFrames, just like RDDs, are distributed in nature.

Creating a DataFrame in PySpark

```
>>> from pyspark.sql import SparkSession
>>> spark = SparkSession.builder \
...     .appName("PySpark Example") \
...     .getOrCreate()
>>> data = [
...     ("Alice", 25, "Data Scientist"),
...     ("Bob", 30, "Software Engineer"),
...     ("Catherine", 29, "Doctor")
... ]
>>> columns = ["Name", "Age", "Occupation"]
>>> df = spark.createDataFrame(data, columns)
>>> df.show()
+---------+---+-----------------+
|     Name|Age|       Occupation|
+---------+---+-----------------+
|    Alice| 25|   Data Scientist|
|      Bob| 30|Software Engineer|
|Catherine| 29|           Doctor|
+---------+---+-----------------+
```

How to Create DataFrames?

- **Using different data formats:** Data can be loaded from JSON, CSV, RDBMS, XML, or Parquet.
- **Loading from an existing RDD:** A DataFrame can be created from an already existing RDD.
- **Programmatically specifying a schema:** You can specify the schema programmatically to create a DataFrame.

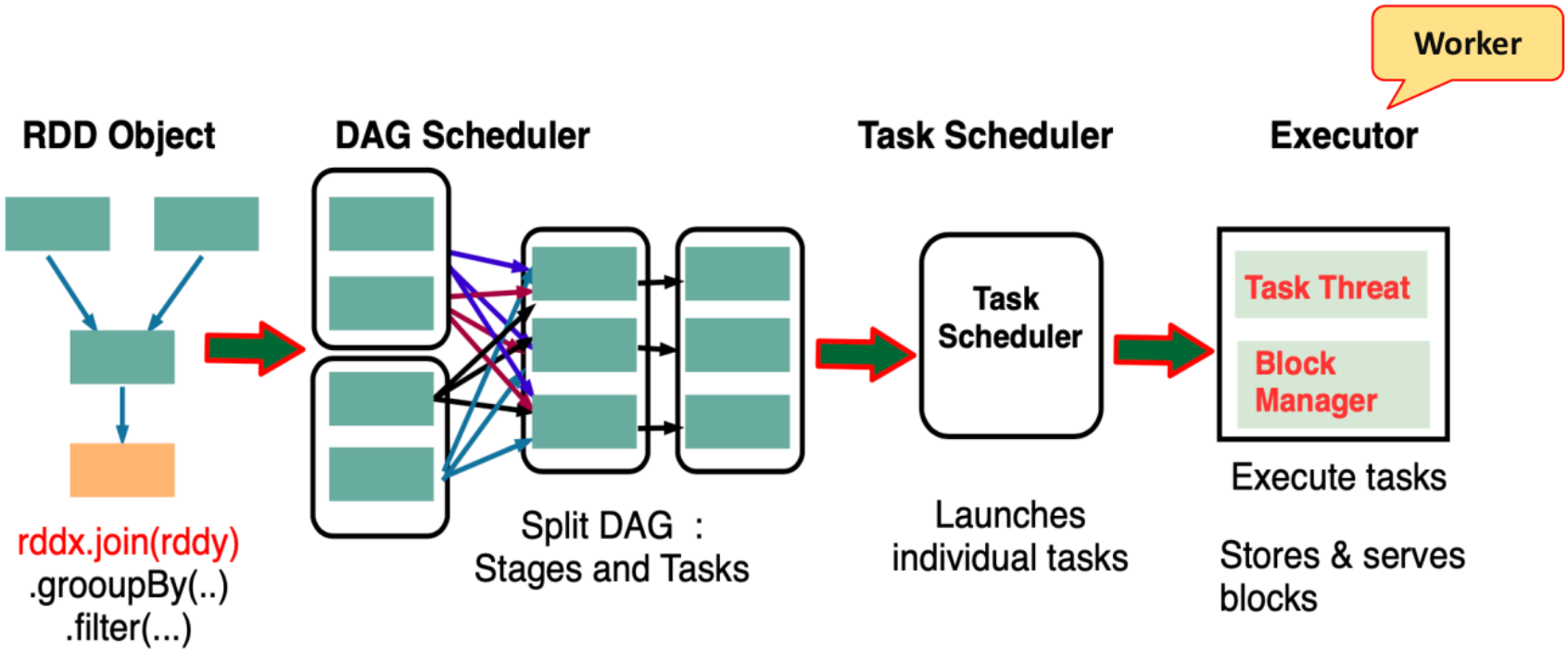
Attendance: 154797

Spark Components

- **Task:** Individual unit of work sent to one executor over a sequence of partitions.
- **Job:** Set of tasks executed as a result of an action.
- **Stage:** Set of tasks in a job that can be executed in parallel at the partition level.
- **RDD:** Parallel dataset with partitions.
- **DAG:** Logical graph of RDD operations.

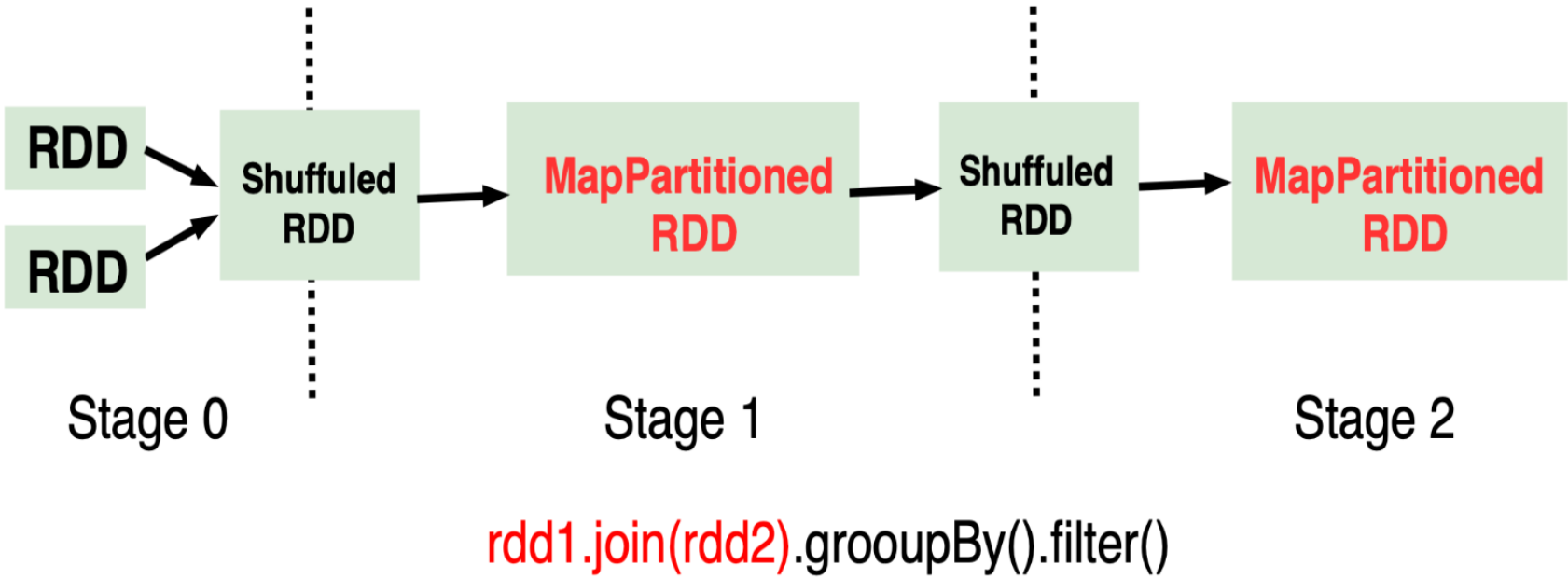
Job Scheduler

The job scheduler transforms a logical execution plan (i.e., RDD lineage of dependencies built using RDD transformations) to a physical execution plan (using stages).

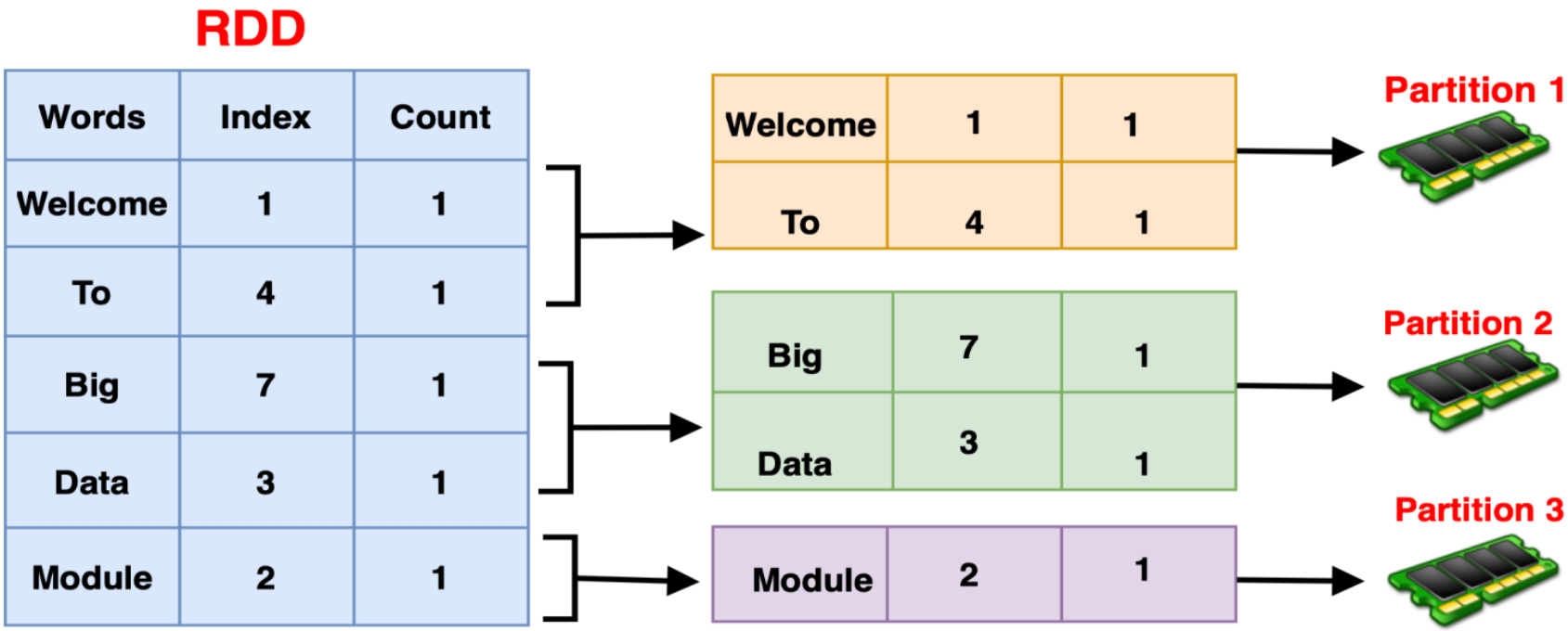


Abstraction 2: RDD Stages

- Each **Spark stage** will dynamically allocate executors.
- Stages contain many neighbouring (in the lineage graph) transformations and actions.



RDD


RDD
Words
Index
Count
Welcome
1
1
To
4
1
Big
7
1
Data
3
1
Module
2
1
Welcome
1
1
To
4
1
Big
7
1
Data
3
1
Module
2
1
Partition 1
Partition 2
Partition 3

Attendance: 154797

Spark Architecture

Spark architecture is made up of a **driver** program and a set of **executor** programs.

Driver:

- Runs the `main()` function, located on a node in the cluster.
- Responsible for:
  - Maintaining information about the Spark applications.
  - Responding to a user's program or input, analyzing, distributing.
  - Scheduling work across the executors.

Executors:

- Perform all the data processing as assigned by the driver.
- Read from and write data to external sources.
- Store the computation results in-memory, cache, or on hard disk drives.
- Interact with the storage systems.

Attendance: 154797

Spark: Standalone Cluster

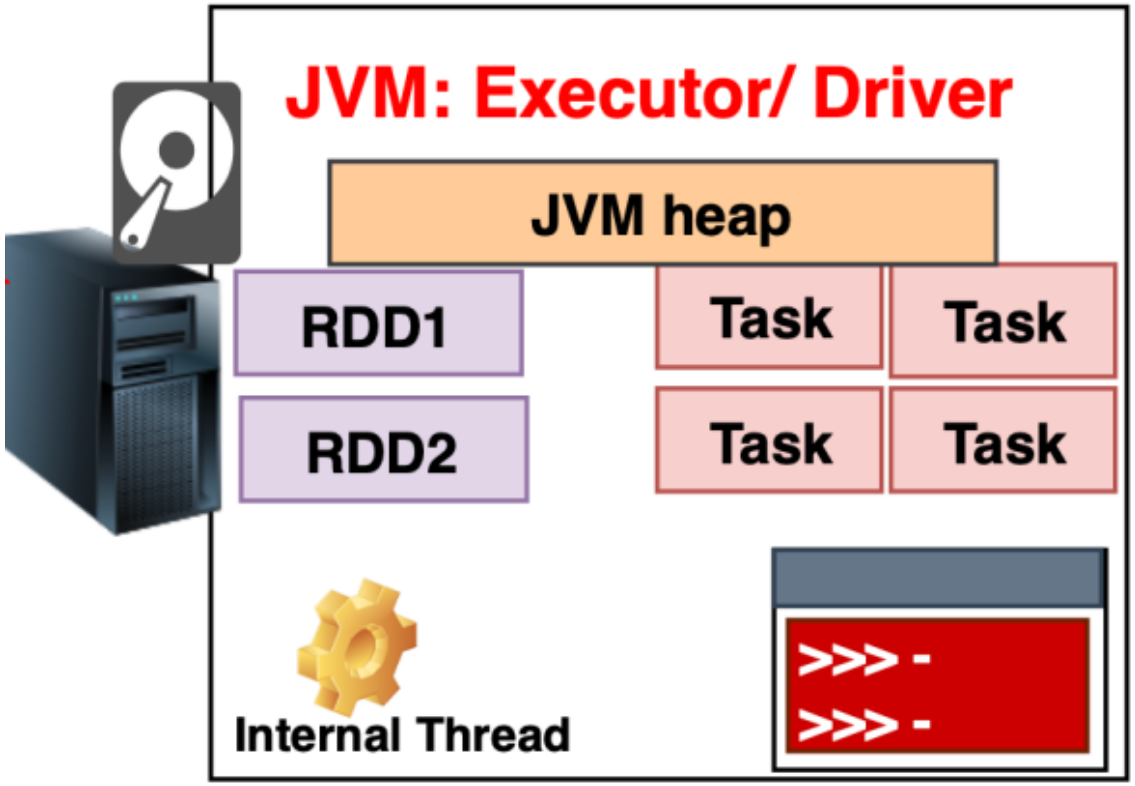
- Spark **local mode** and Spark **local cluster mode** are special cases of running Spark on a single machine.
  - **Local mode:** Runs Spark with one worker thread.
  - **Local[N]:** Runs Spark with N worker threads (like a thread pool).
  - **Local[*]:** Runs Spark with as many worker threads as there are logical cores on your machine.
- This mode is useful for running Spark applications on a local machine for testing or small workloads.

SparkContext Example in Local Mode

```
import pyspark
sc = pyspark.SparkContext('local[4]')
txt = sc.textFile('file:///usr/share/doc/python/copyright')
python_lines = txt.filter(lambda line: 'python' in line.lower())
print(python_lines.count())
```

- The SparkContext is created in local mode with 4 worker threads.
- The `textFile` method reads a file.
- The `filter` operation filters the lines containing the word "python."
- The `count` operation returns the number of matching lines.

Spark Standalone Cluster

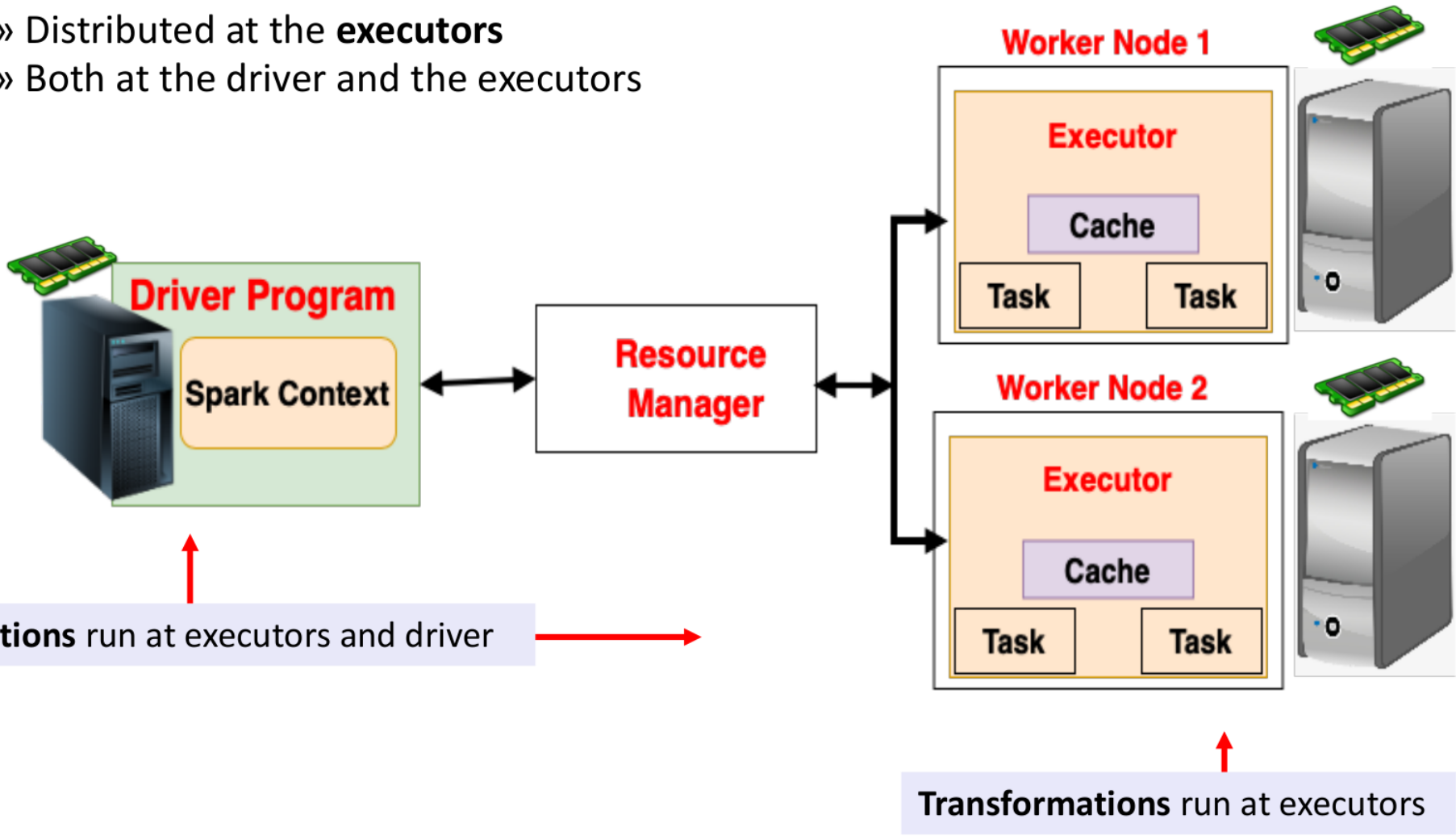


Cluster Mode

Where does code run?

» Distributed at the **executors**

» Both at the driver and the executors



References

- Feng, W. (2019). Learning Apache Spark with Python
- https://databricks.com/glossary/spark-api
- https://realpython.com/pyspark-intro/
- https://www.cs.princeton.edu/courses/archive/fall13/cos518/papers/spark.pdf

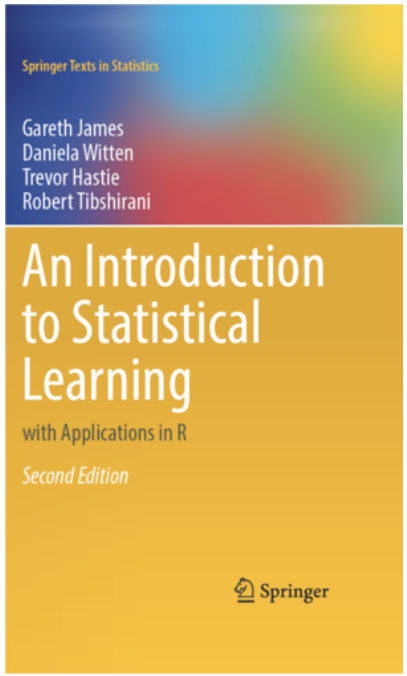
Classification

5 Classification
- Maths for Classifiers

Attendance: 154797

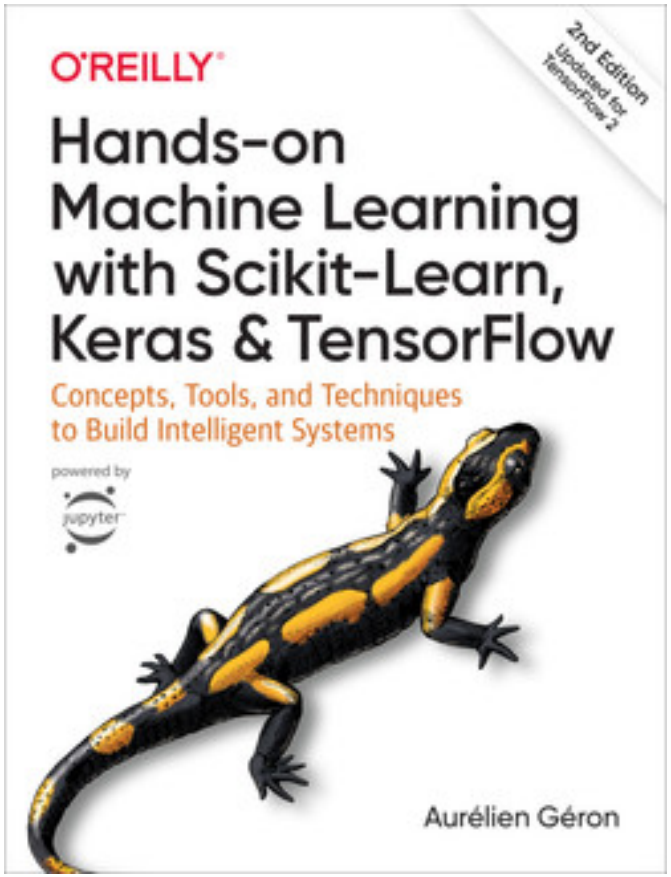
Recommended Reading for Classification

- This lecture is self-contained, but the following might also be useful:
- **An Introduction to Statistical Learning**
  - See chapter 9 for the maths in today's lecture
  - Examples in R (not used in this course)
  - Free online: https://www.statlearning.com/
  - Accompanying lectures: https://web.stanford.edu/~hastie/lectures.htm



Recommended Reading for Classification

- **Hands-on Machine Learning, 2nd Edition 2019**
  - Available online from the university library: https://libguides.liverpool.ac.uk/online
  - Chapter 5 introduces classifiers without maths
  - Examples in Python
  - Code for today's lecture came from Hands-on ML



Supervised Versus Unsupervised Learning

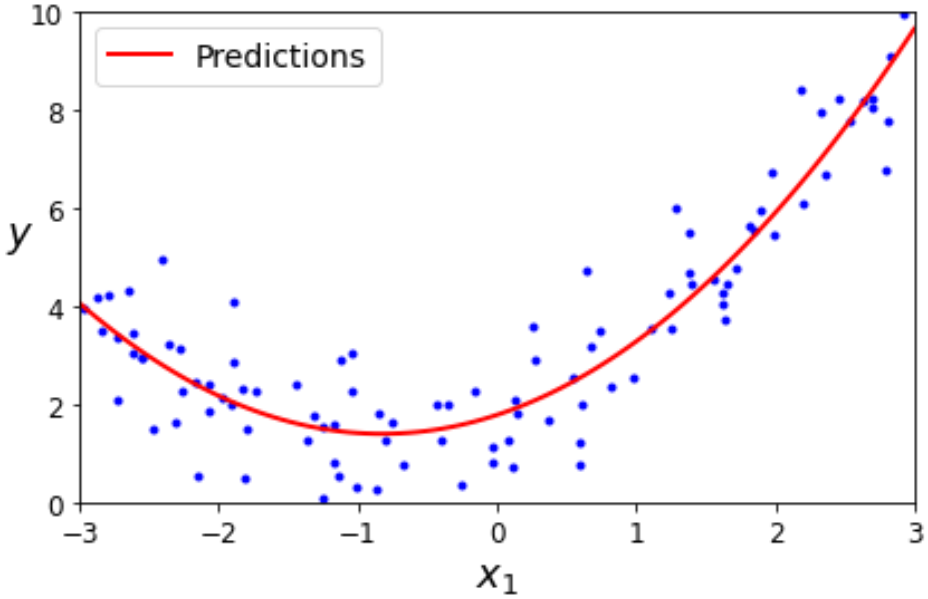
- You've seen a few machine learning algorithms so far: PCA, clustering and linear regression.
- How do these fit into the broader space of ML algorithms?
- These are three types of machine learning algorithms: unsupervised learning, supervised learning and reinforcement learning
- **For unsupervised learning:**
  - The training data is a design matrix X, representing observations of a system
  - The aim is to find structure in this data
  - Examples:
    - PCA (week 5)
    - Clustering (week 6)

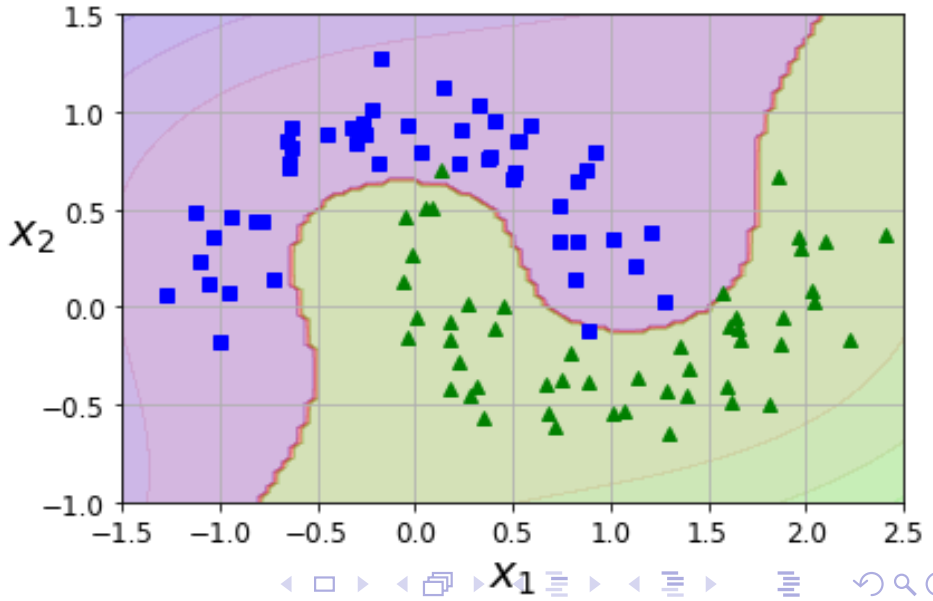
Supervised Versus Unsupervised Learning II

- **For supervised learning:**
  - The training data is a design matrix X and a vector (or matrix) of responses y
  - The aim is to train a model to predict the responses in y given the observations in X
  - This model is then used to predict unseen responses from new observations
  - Examples:
    - Linear regression (week 7)
    - K-nearest neighbours (today)
    - Linear support vector classifiers (today)
    - Support vector machines (today)
- Reinforcement learning is outside the scope of this course

Supervised Learning

- There are two types of supervised learning:
  - **Regression:**
    - The responses are real numbers or vectors thereof.
    - Regression model outputs an expected value.
    - E.g., linear regression (week 7).
  - **Classification:**
    - The responses are integers representing membership in a class.
    - Classification model outputs a class.
    - E.g., K-Nearest Neighbours, SVMs, linear SVCs.



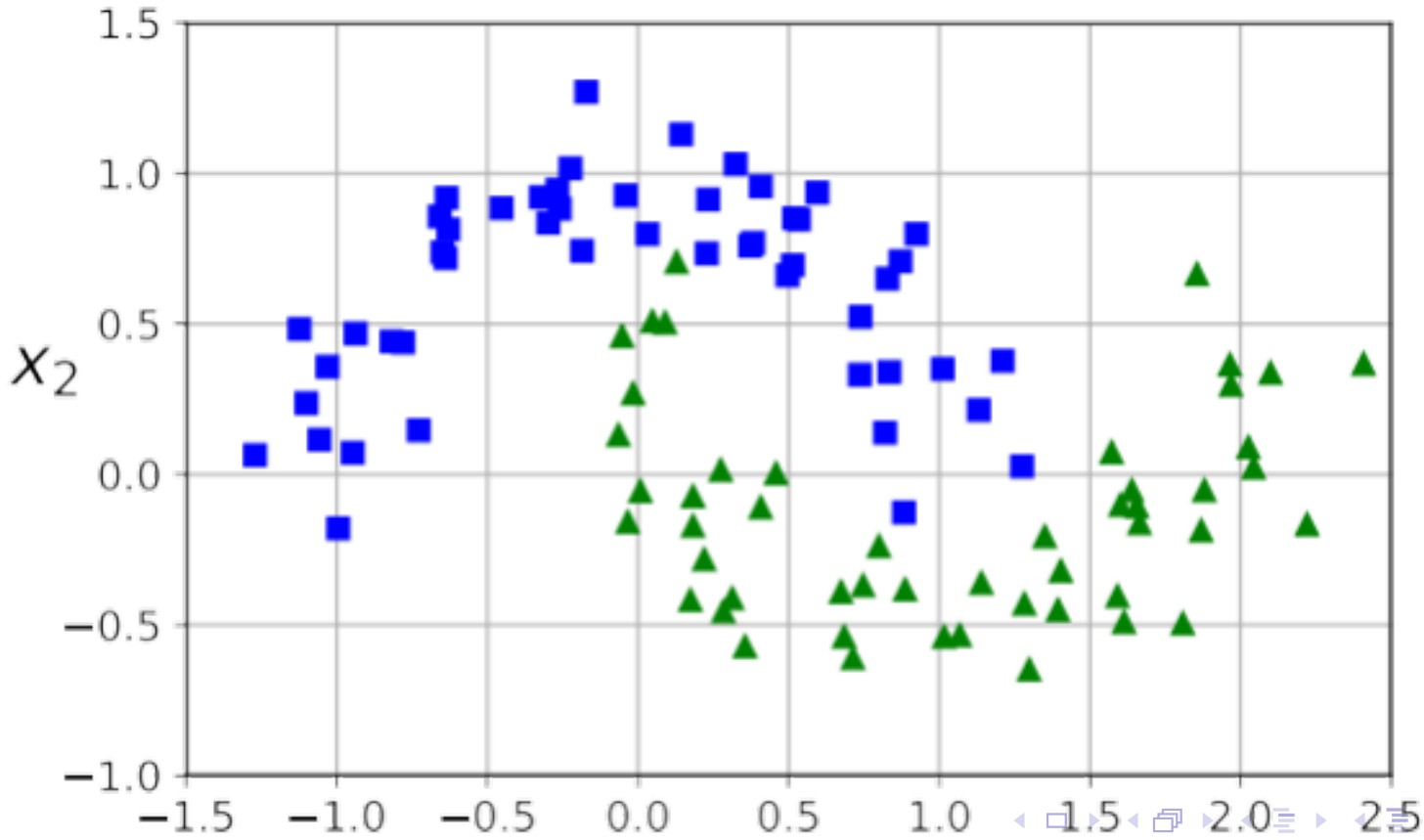


Classification Versus Clustering (1/2)

- Classification is similar to the clustering problem we saw in week 6:
  - We have a set of training data, and we wish to construct a decision function that will associate new data with sub-groups of the training data.
- The difference with classification is that the training data has already been separated into classes, and we need to generate a decision function that is consistent with these classes whenever possible, and generalises well to unseen data.
- The fact that we have class labels makes classification a supervised learning problem, whereas clustering is an unsupervised learning problem.

Classification Versus Clustering (2/2)

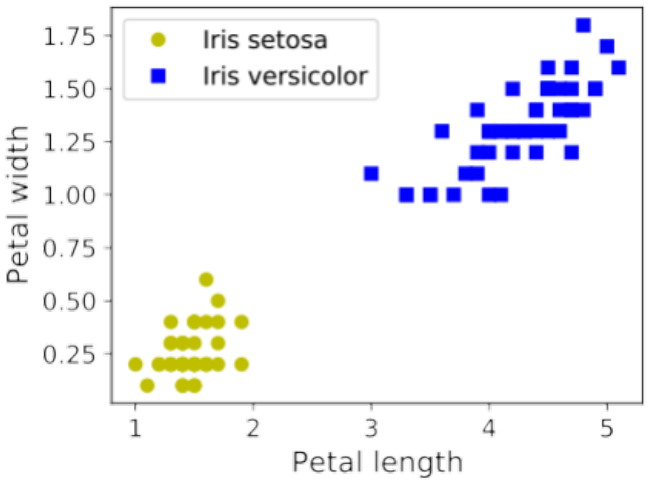
- The figure shows an example dataset for classification.
- It's the familiar half-moon dataset from week 6, except we now have class labels for the points in each half moon.

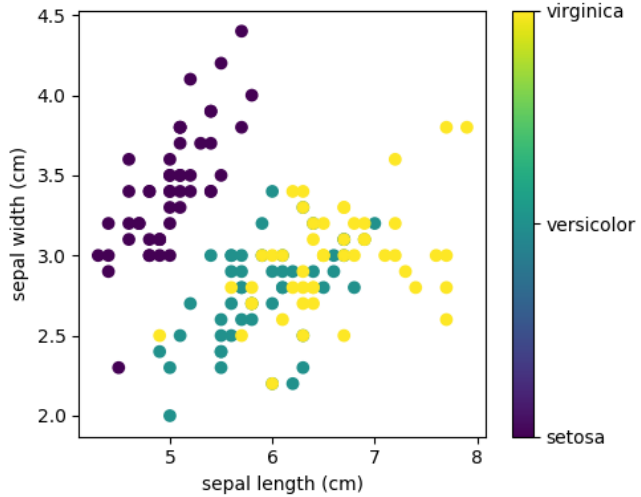


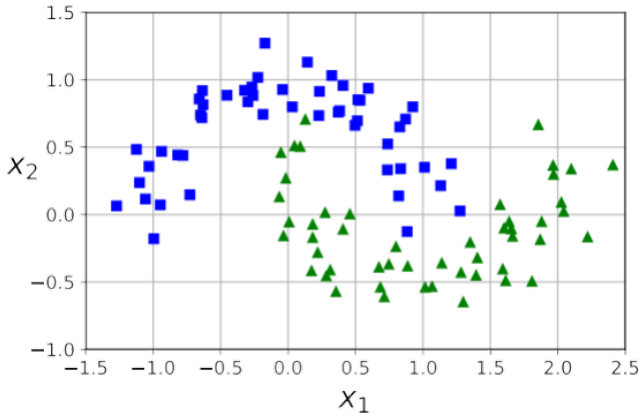
Types of Classification Datasets

- Classification datasets can present different challenges. . .
- The simplest datasets to work with are ones in which the classes are linearly separable.
- Sometimes classes can only be separated with non-linear boundaries.
- Sometimes classes aren't separable with realistic boundaries.

Classification Dataset Examples







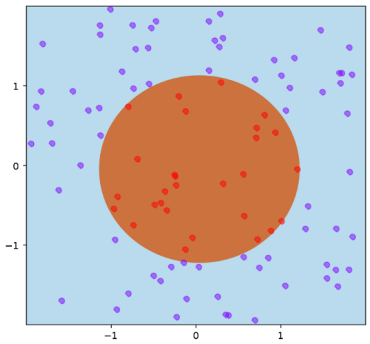


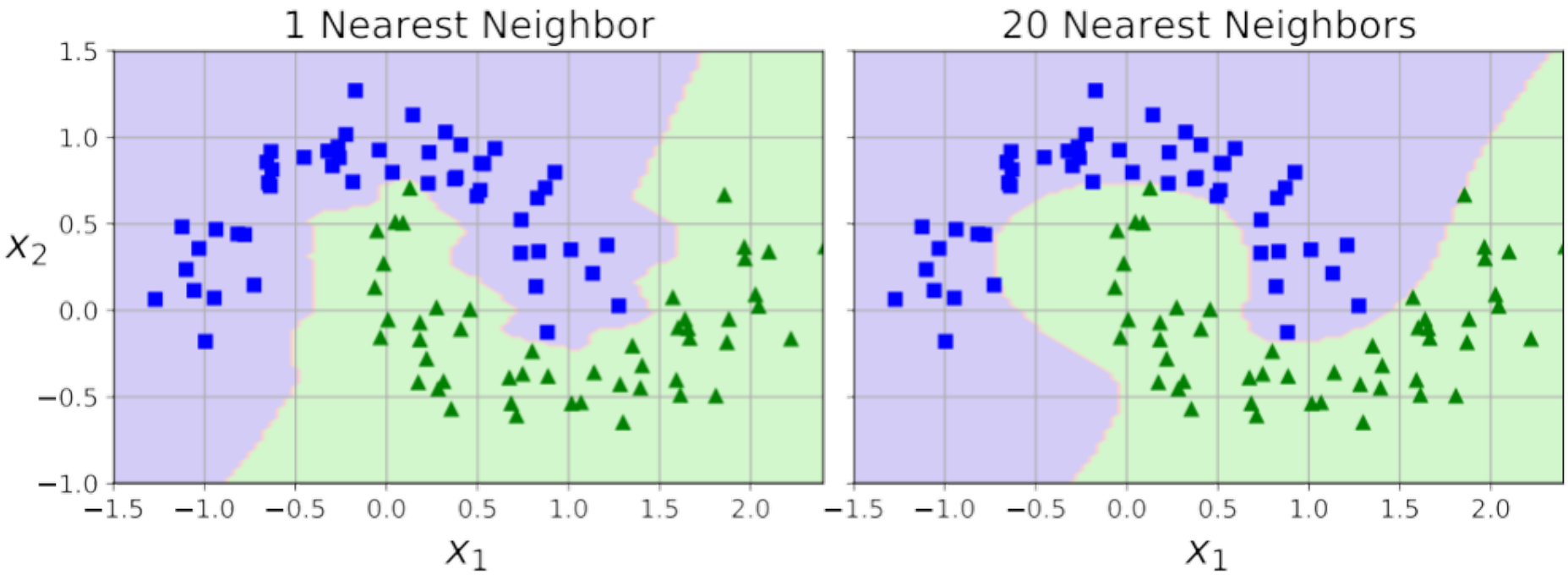
Figure: Examples of linearly separable, non-linearly separable, and inseparable

Decision Boundaries

- Classification algorithms work by constructing decision boundaries that split the sample space into regions corresponding to the different classes.
- Different classification algorithms will build different decision boundaries.
  - The plots show decision boundaries for the half moon dataset, created using KNN and SVM classifier (which we'll see in later slides).
- For a given classification algorithm (e.g. KNN or SVM), different choices of the algorithm's parameters will give rise to different decision boundaries (as seen in the plots).
  - Choosing the best values for these parameters is called model selection. It's out of scope for us, but see Hands-on Machine Learning if you're interested.

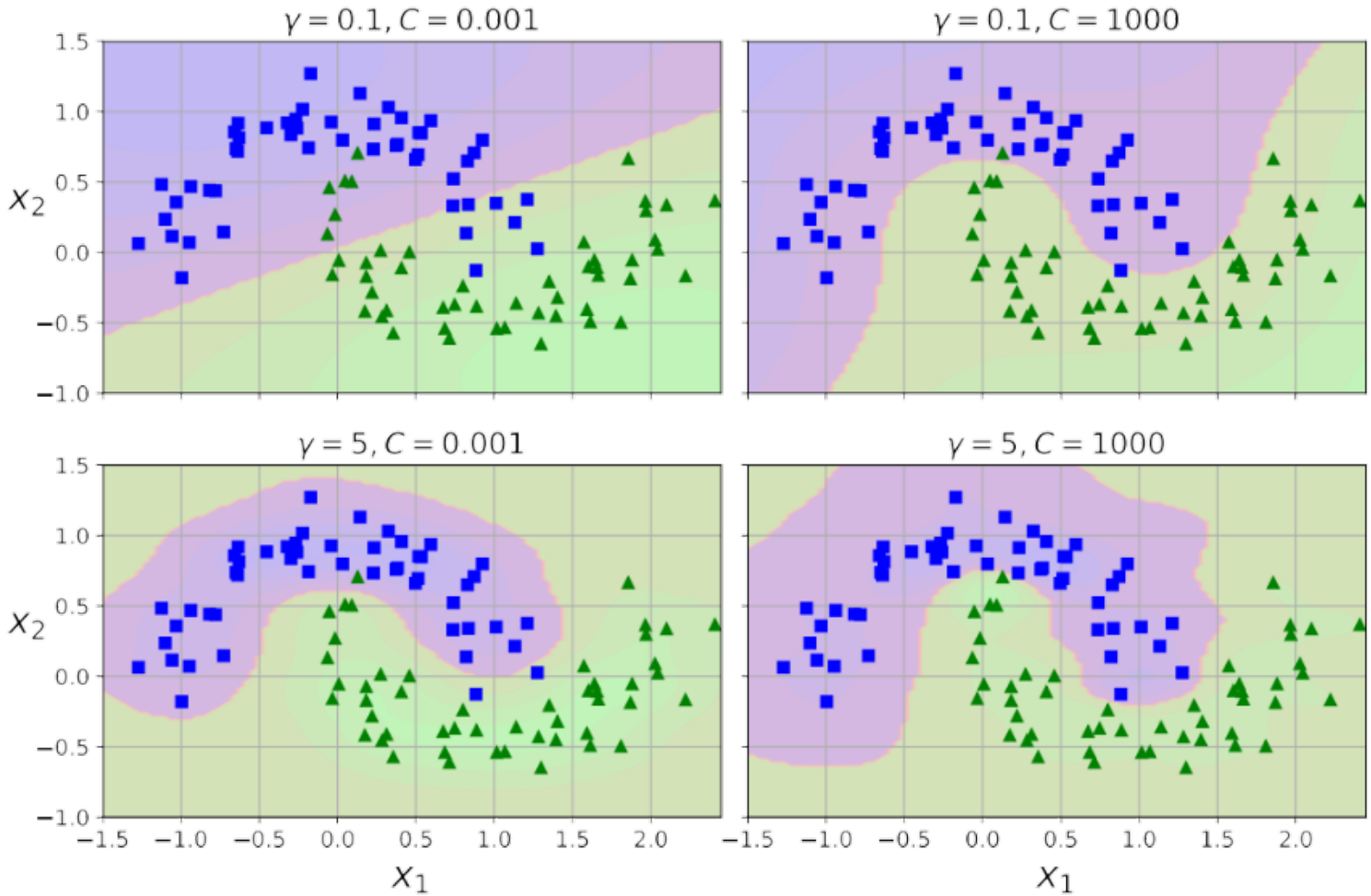
KNN Decision Boundaries

- Decision boundaries created with KNN classifiers.



SVM Decision Boundaries

- Decision boundaries created with SVM classifiers.

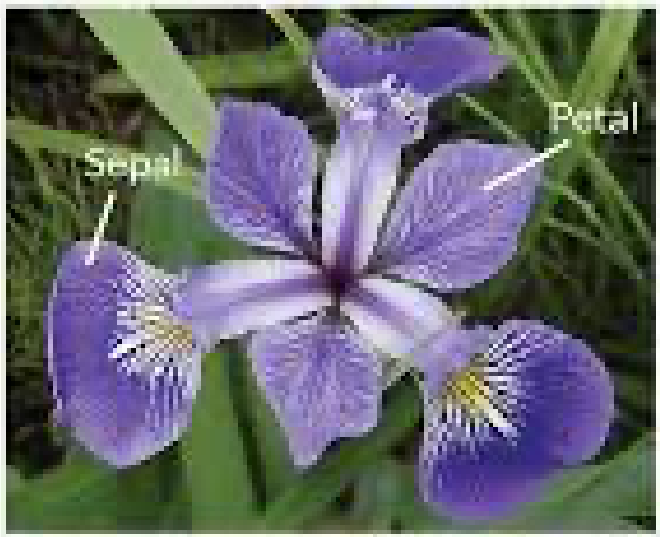


Two-Class Versus Multi-Class

- Two-class classification problems are easier to solve than multi-class, because only one decision boundary is needed.
- We'll mostly focus on two-class classifiers today, but we'll briefly introduce multi-class classifiers later in the lecture.

Iris Dataset

- Classic dataset for classification
- Classes are 3 sub-species of the iris flower
- Measurements are dimensions of different parts of the flower
- Good demonstration of separable and non-separable classes
- We'll use it as a running example

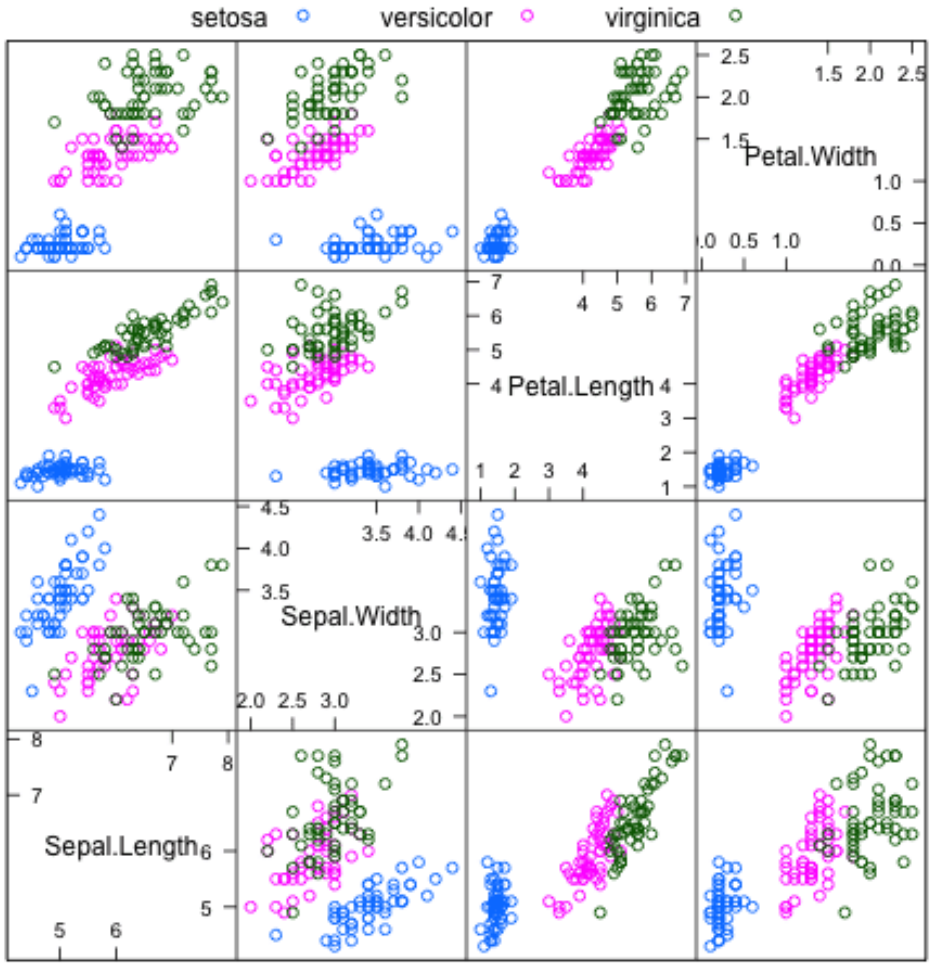


Iris Versicolor

Iris Setosa

Iris Virginica

Iris Dataset Scatter Plot

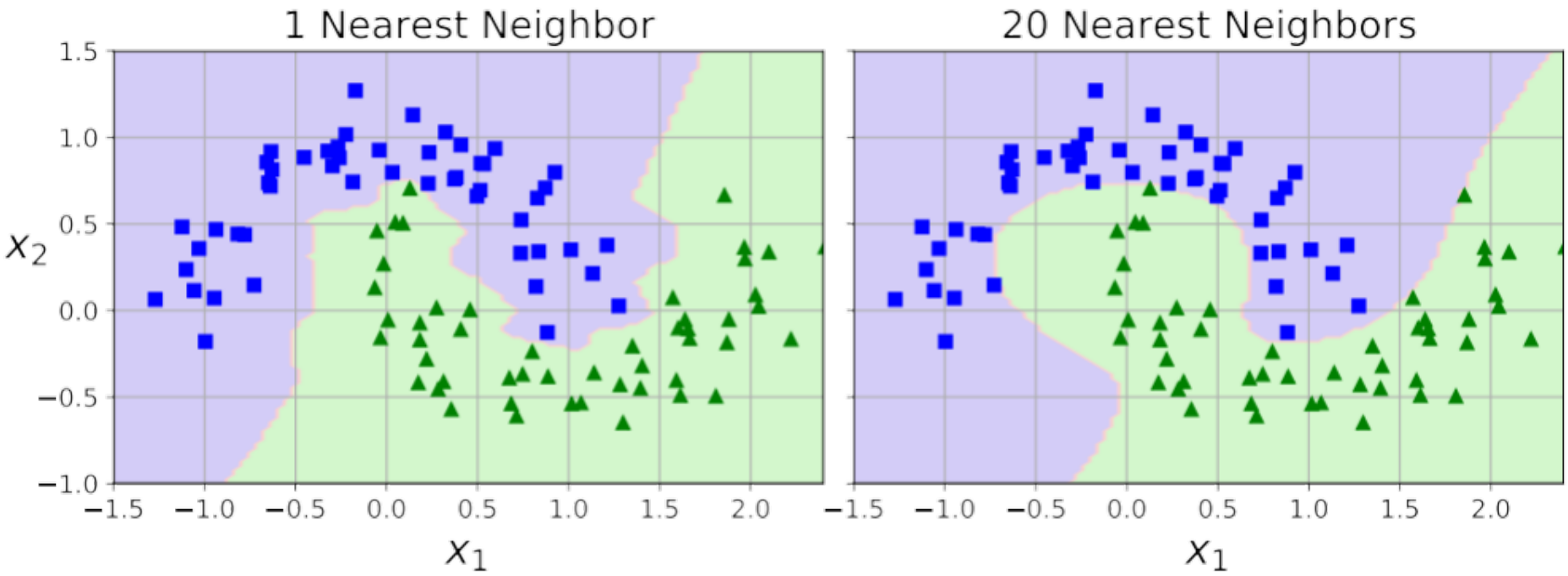


Scatter Plot Matrix

K-Nearest Neighbours

- K-nearest neighbours (**KNN**) is simple, intuitive decision classifier
- It assigns new points to the most common class among their k-nearest neighbours, for some *k*
- This often works very well for low dimensional data (when the width of the design matrix is ≤ 3)
  - **KNN** is said to be *the best classifier for about 1/3 of real-world classification problems*
- Larger values of *k* give smoother decision boundaries, as shown in the plot
- **Exercise:** What's the computational complexity for this classifier?
- K-nearest neighbours unfortunately doesn't work well for high dimensional data (when the width of the design matrix is > 3)
  - The reason is outside the scope of this course. Interested students can look into the "curse of dimensionality" (see either of the recommended textbooks).

KNN Decision Boundaries

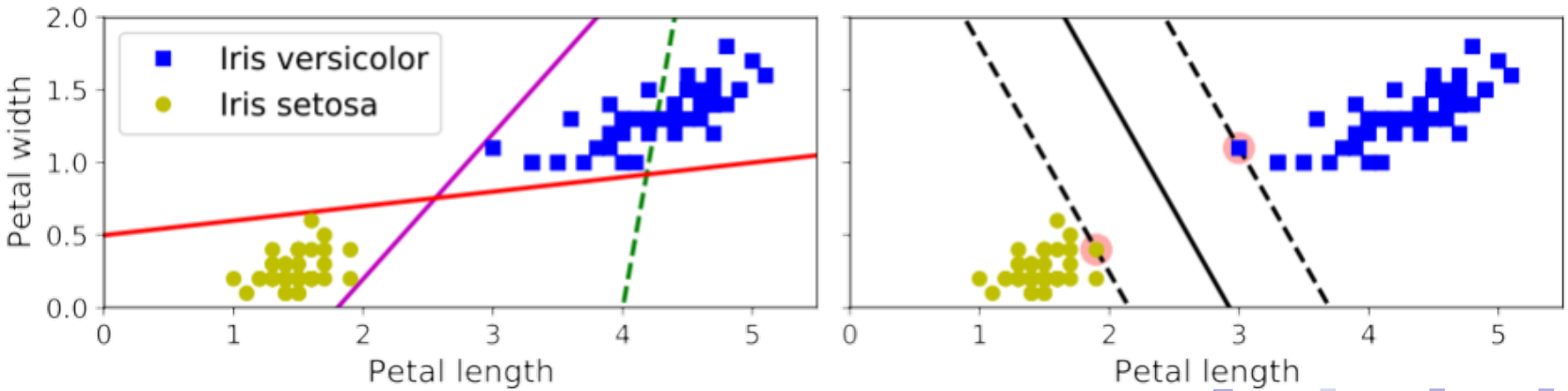


SVMs and Linear SVCs

- K-nearest neighbours doesn't work well for high dimensional data (when the width of the design matrix is large)
  - For these datasets, we need a more sophisticated approach.
- Two popular classification methods for high dimensional data are linear support vector classifiers (**SVCs**) and (nonlinear) support vector machines (**SVMs**)
- **Linear SVC** supports linear decision boundaries, and has time complexity of (almost) $O(n)$
  - Stochastic gradient descent can be used, allowing linear **SVCs** to be applied to very large datasets
- **SVM** supports non-linear decision boundaries and has time complexity of $O(n^2)$ to $O(n^3)$, depending on the specific implementation
- We'll focus on these methods for the rest of the lecture

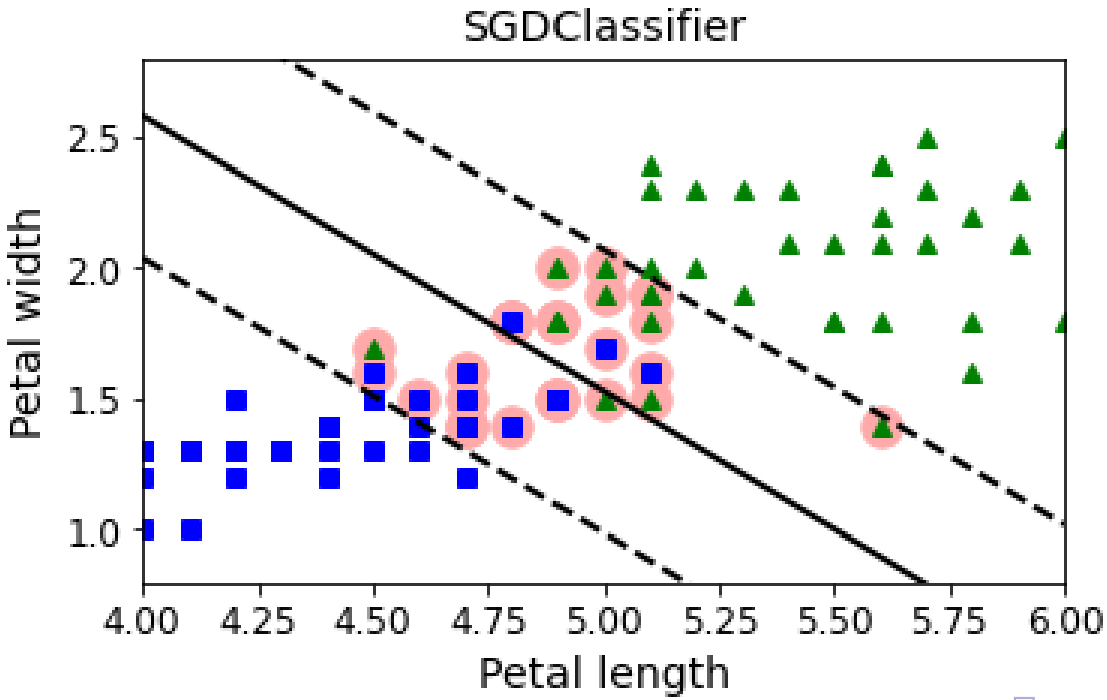
Classifiers for Linearly Separable Data

- For separable classes, it's easy to build a classifier: just draw a straight line between the two classes and treat this as the decision boundary.
- But which straight line should we choose?
  - There are infinitely many of them.
  - Two are shown in the plot below.
- One solution is to choose the separating line that permits the largest margin surrounding it which is free of observations (*e.g.*, see the right hand plot).
- The classifier that uses this line is called the *maximal-margin classifier*, and it's used by linear **SVCs** and **SVMs**.



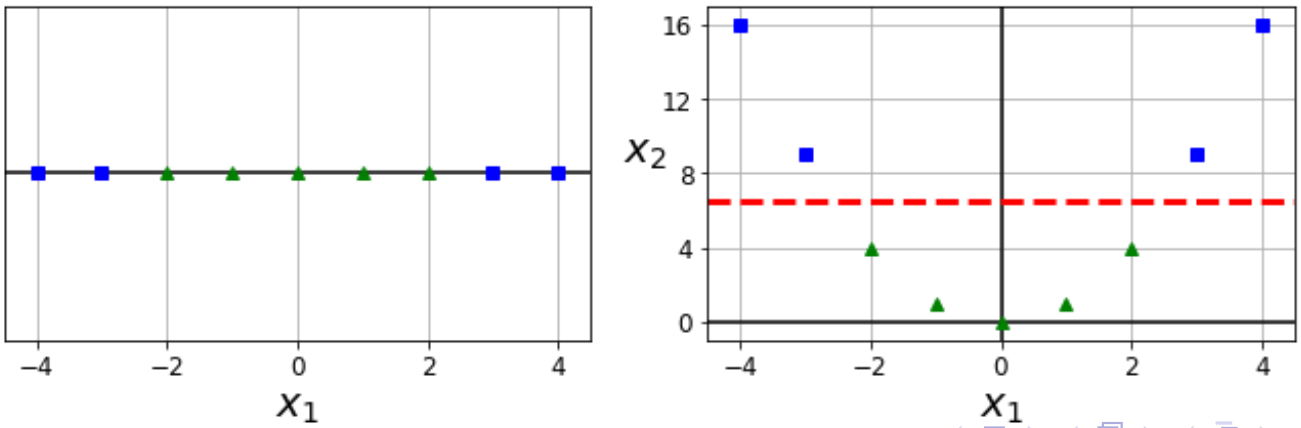
Classifiers for Linearly Separable Data

- What if the classes aren't linearly separable?
- In this case, there's no (straight) line that separates the classes.
- However, we can still use the maximal-margin classifier, if we allow a small number of observations to cross the margin.
- This classifier is called a *linear support vector classifier*.



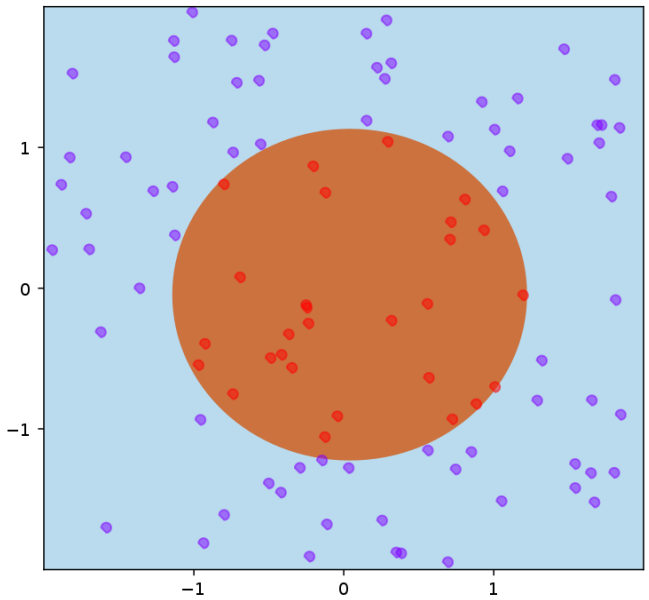
Basis Expansion

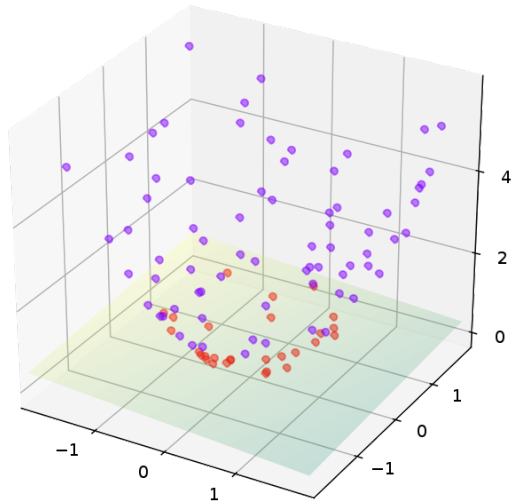
- Some classification problems just can't be handled with linear decision boundaries.
- Basis expansion is one way to deal with these datasets:
  - Augment the data with extra variables.
  - Below we turned a 1D dataset into a 2D dataset, where the second value for a given observation is set to $x_2 = x_1^2$.
- In this example, the enriched data can now be separated with a straight line in the 2D space.



Basis Expansion II

- Below is another example in which basis expansion can create a linearly separable dataset.
- This time, we added a third dimension to a 2D dataset.
- The third value for a given observation is set to $x_3 = x_1^2 + x_2^2$.
- In this case, the 3D dataset can be separated with a linear plane.

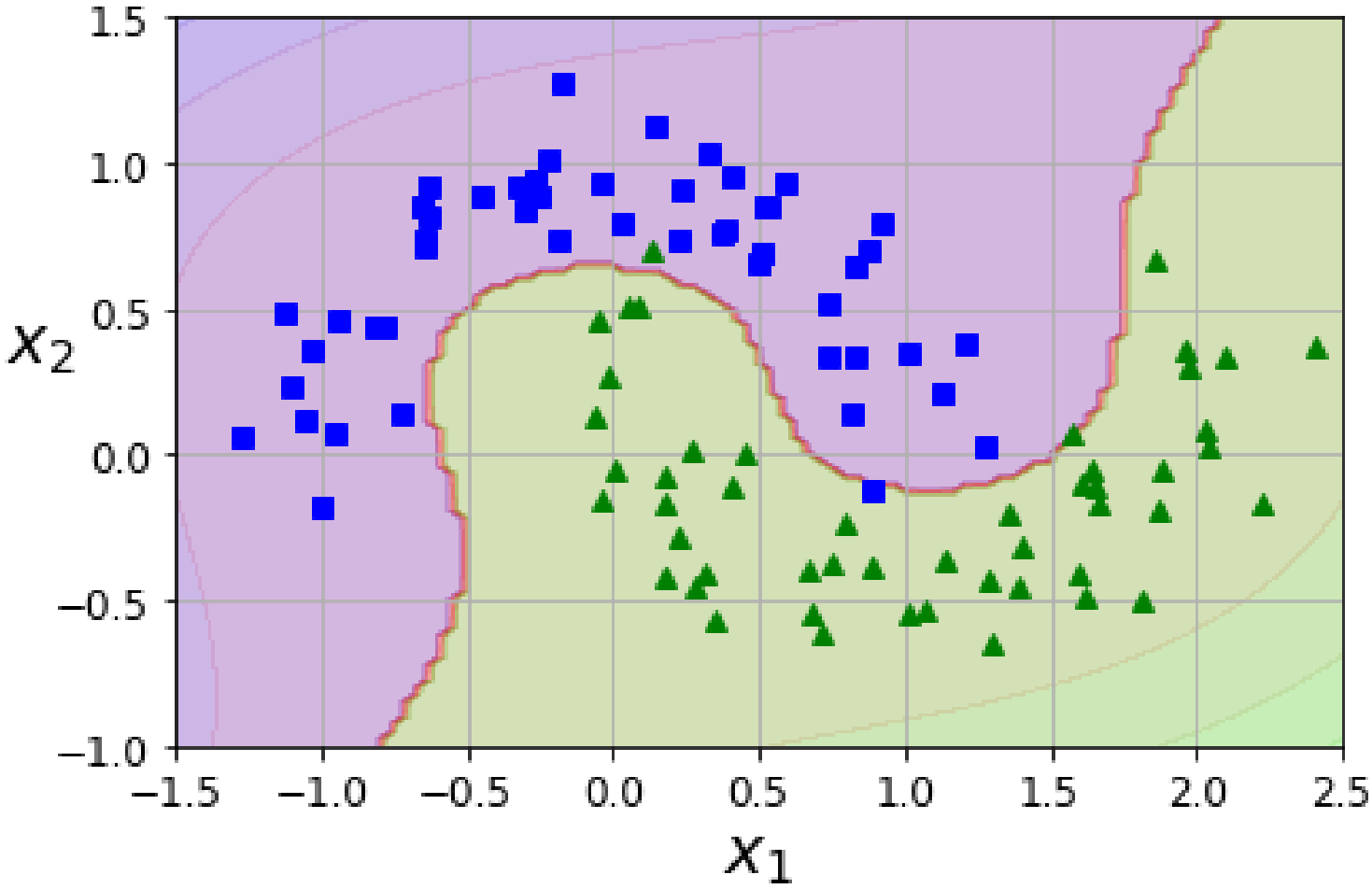




Non-Linear Decision Boundaries and SVMs

- We've seen that some classification problems just can't be handled with linear decision boundaries.
- Basis expansion works well for some, but not all, datasets.
- Enter the support vector machine. . .
- **SVMs** extend the linear support vector classifier to handle non-linear decision boundaries in a very general way.
- They use a technique called *the kernel trick* (which we'll see shortly):
  - This is a generalisation of the basis expansion approach that we saw in previous slides, and it can fit very flexible decision boundaries (*e.g.*, bendy or circular).
- We'll see more about the kernel trick after we cover the *maths* for linear support vector classifiers.

Non-Linear Decision Boundaries and SVMs



Multi-Class Classification

- Linear **SVCs** and **SVMs** are binary classifiers. However, they can be used to build multi-class classifiers as follows:
- **One-versus-the-rest (OvR)** (also called one-versus-all)
  - For an m-class dataset, fit m binary classifiers to the whole dataset, with each classifier trying to identify a single class.
  - To classify new data, run all m binary classifiers, and choose the class that achieved the highest score on its classifier.
- **One-versus-one (OvO)**
  - For an m-class dataset, fit $\frac{m(m-1)}{2}$ binary classifiers. For each pair of classes (i, j), remove all data that isn't either class i or class j. For this reduced dataset, fit a binary classifier.
  - To classify new data, see which class wins the most 1-versus-1 contests.

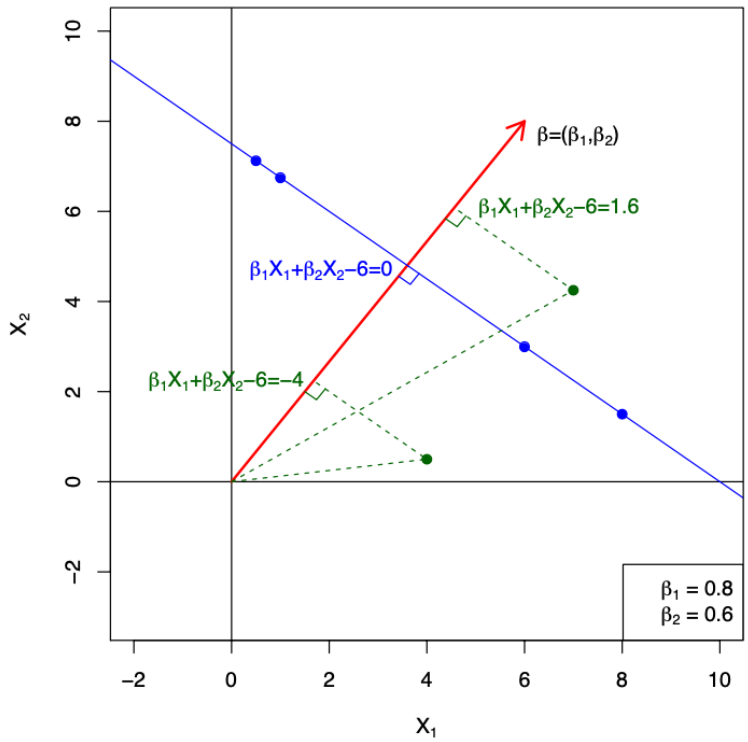
5 Classification

- Maths for Classifiers

Attendance: 154797

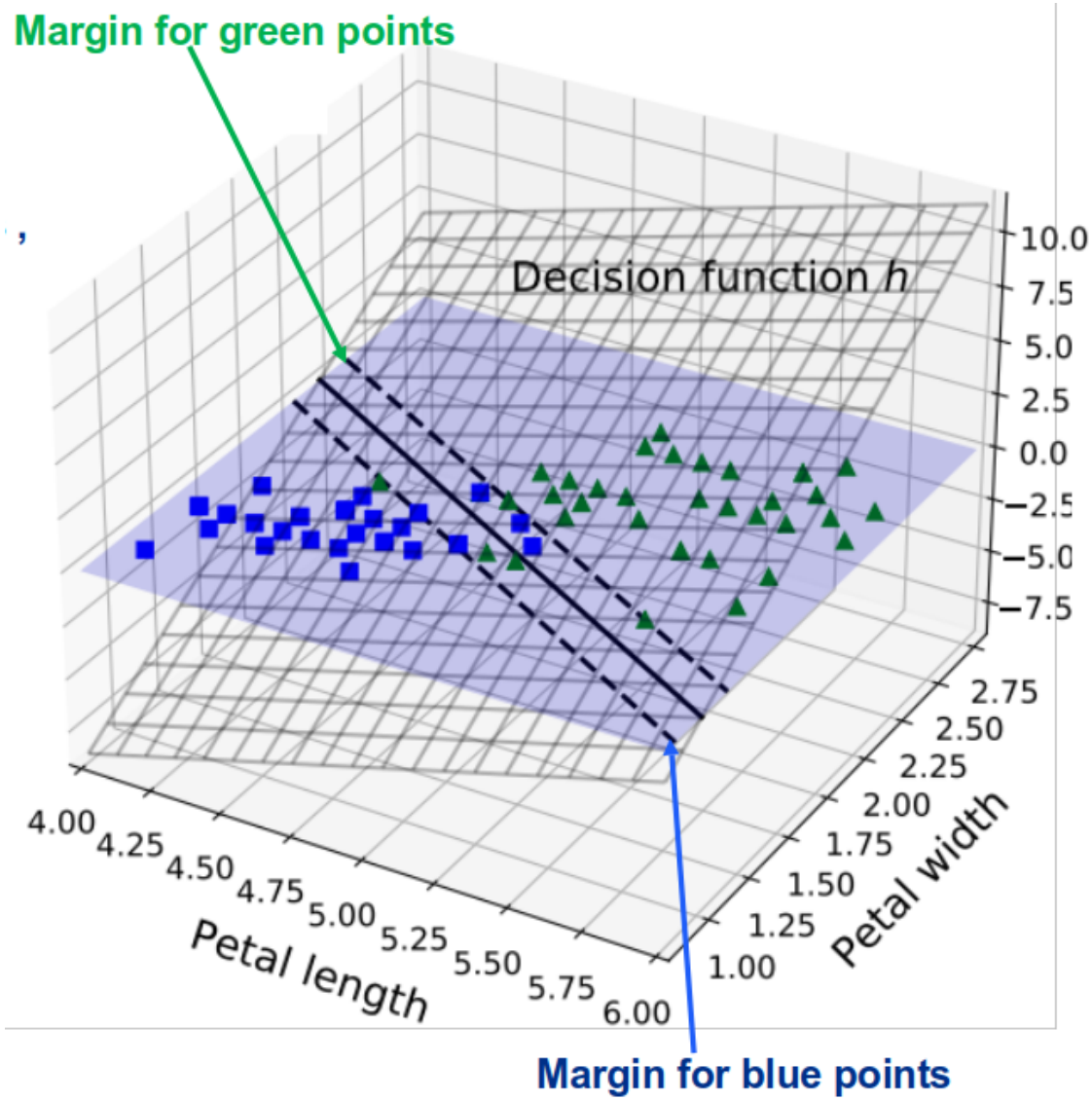
Hyperplanes

- So far, we've seen that 2D spaces can be separated with a straight line and 3D spaces can be separated with a plane.
- In general, an N-dimensional space can be separated with an (N-1) dimensional *hyperplane*, which generalises the concept of a plane in 3D space.
- A hyperplane is defined by **a vector and an offset**, as shown in the plot:
  - We have a vector $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix}$ and an offset $\beta_0 = 6$.
  - We define a hyperplane as the set of all vectors $\mathbf{x}$ for which $\mathbf{x}^T\boldsymbol{\beta} = \beta_0$.
  - In this case, $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$, meaning that $\mathbf{x}^T\boldsymbol{\beta}$ is the projection of $\mathbf{x}$ onto the red line.
- (Exercise: use $\mathbf{x}^T\boldsymbol{\beta} = |\mathbf{x}||\boldsymbol{\beta}|\cos\theta$)



Decision Functions

- For binary classification with class labels ± 1, we can use the value of $h(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta} - \beta_0$ as a *decision function*.
  - We decide on the class based on the sign of $h(\mathbf{x})$.
  - i.e., class$(\mathbf{x})$ = sign$(h(\mathbf{x}))$.
- The plot shows this decision function for iris data:
  - (The dotted lines show the **SVC** boundaries.)



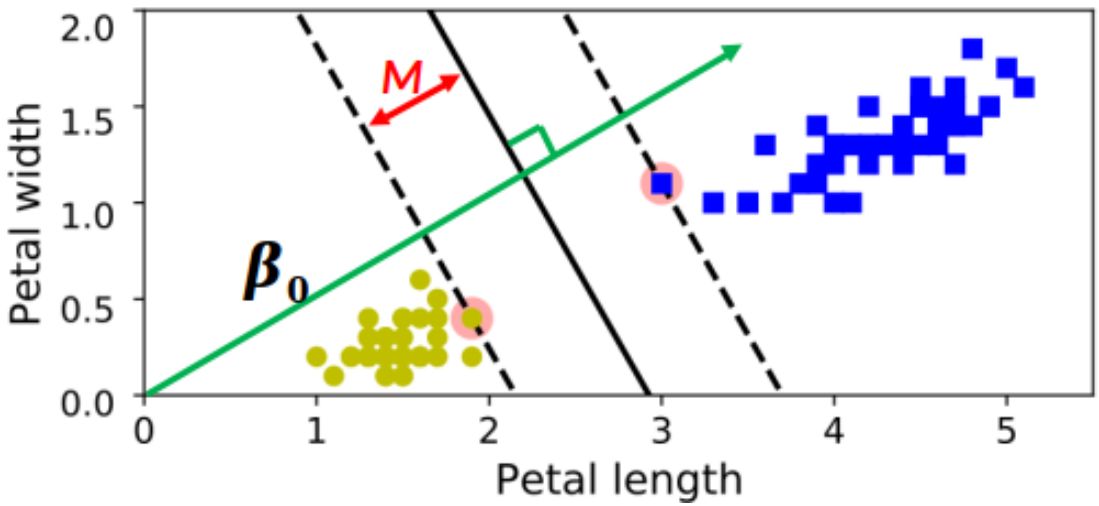
Maximal Margin Classifier

- In the slide *"Classifiers for Linearly Separable Data"*, we saw the maximal-margin classifier.
- We can find the maximal-margin classifier by solving an optimisation problem:
  - Optimisation problems were introduced in week 5.
- The maximal margin classifier is specified by a unit vector $\boldsymbol{\beta}$ (i.e., $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$) and offset β_0, which solve the following optimisation problem:

$$\begin{aligned} &\text{maximise} \quad M \\ &\text{subject to} \quad \boldsymbol{\beta}^T\boldsymbol{\beta} = 1 \\ &\qquad y_i(\boldsymbol{\beta}^T\mathbf{x}_i - \beta_0) \geq M \\ &\text{for all } i = 1, \ldots, N \end{aligned}$$

- This can be rephrased as a convex quadratic program and solved efficiently (outside the scope of this course, but it's discussed in *Elements of Statistical Learning*).

Maximal Margin Classifier



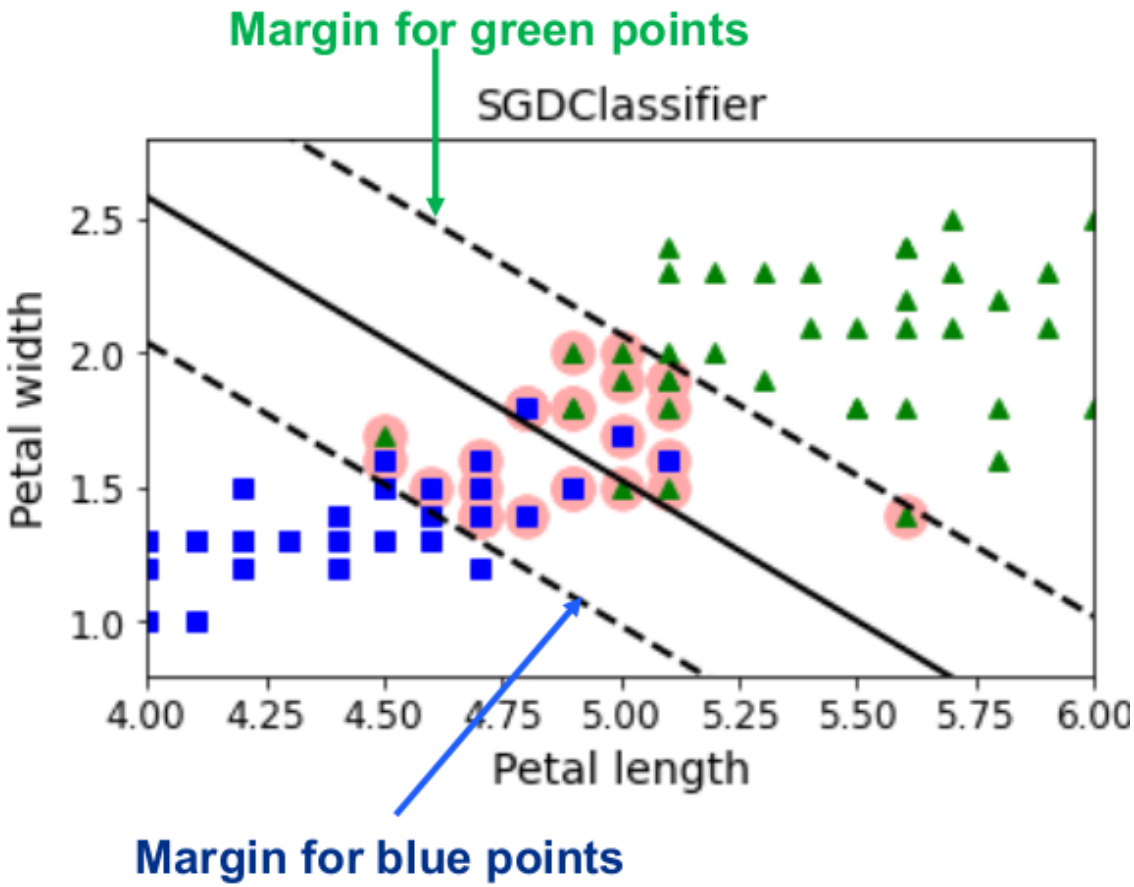
Linear Support Vector Classifier

- The Linear Support Vector Classifier requires a **small change** to the optimisation problem that we just saw.
- We allow some points to cross the separating margin (shown in **orange** in the plot on the next slide) but constrain the total distance that all points collectively pass their margins to a fixed value C:

$$\begin{aligned} &\text{maximise} \quad M \\ &\text{subject to} \quad \boldsymbol{\beta}^T\boldsymbol{\beta} = 1 \\ &\quad y_i(\boldsymbol{\beta}^T\mathbf{x}_i - \beta_0) \geq M(1-\epsilon_i) \\ &\quad \epsilon_i \geq 0, \quad \sum_{i=1}^{N} \epsilon_i \leq C \end{aligned}$$

- Again, this problem can be rephrased as a convex quadratic program and solved efficiently (outside the scope of this course, but it's discussed in *Elements of Statistical Learning*).

Linear Support Vector Classifier



Support Vectors

- It turns out that the optimal $\boldsymbol{\beta}$ for a linear SVC is a linear combination of the observations:

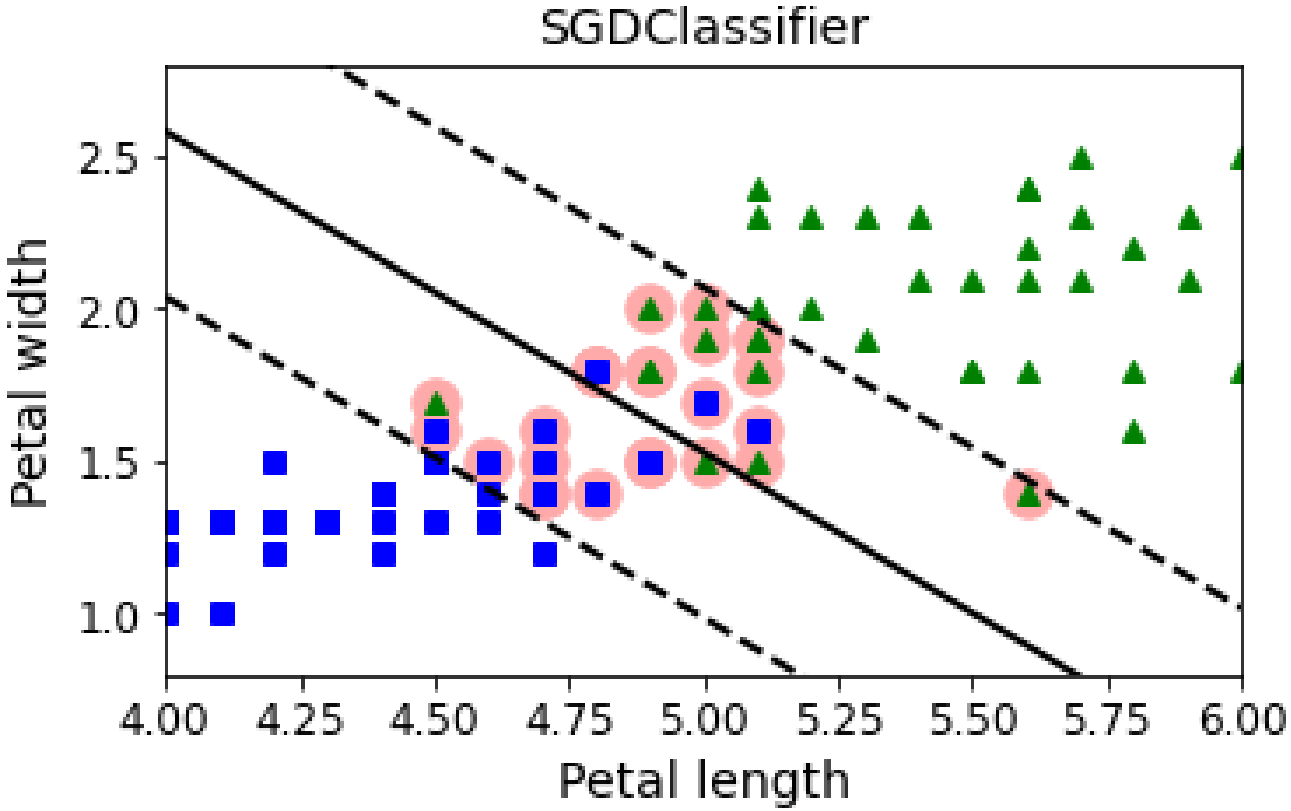
$$\boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i$$

- Hence, the decision function can be written as:

$$h(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x} - \beta_0 = \left(\sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x} \right) - \beta_0$$

- Furthermore, the α_i turn out to be non-zero only for $\mathbf{x}_i$ that are either:
  - Touching or
  - On the wrong side of their respective margins.
- These $\mathbf{x}_i$ are referred to as the *support vectors*, shown in orange in the plot.
- The proof of this is outside the scope of the course (and quite tricky). It's given in *Elements of Statistical Learning*, which is a more advanced book from the same authors as *An Introduction to Statistical Learning*.

Support Vectors



The Kernel Trick

- We saw in the last slide that the decision function for linear support vector classifiers can be written as:

$$h_{\text{SVC}}(x) = \left(\sum_{i=1}^{N} \alpha_i x_i^T x \right) - \beta_0$$

  where the α_i are non-zero only for x_i that are support vectors.
- We can make a small modification to this decision function, by replacing the expression $x_i^T x$ by a different function called a kernel:

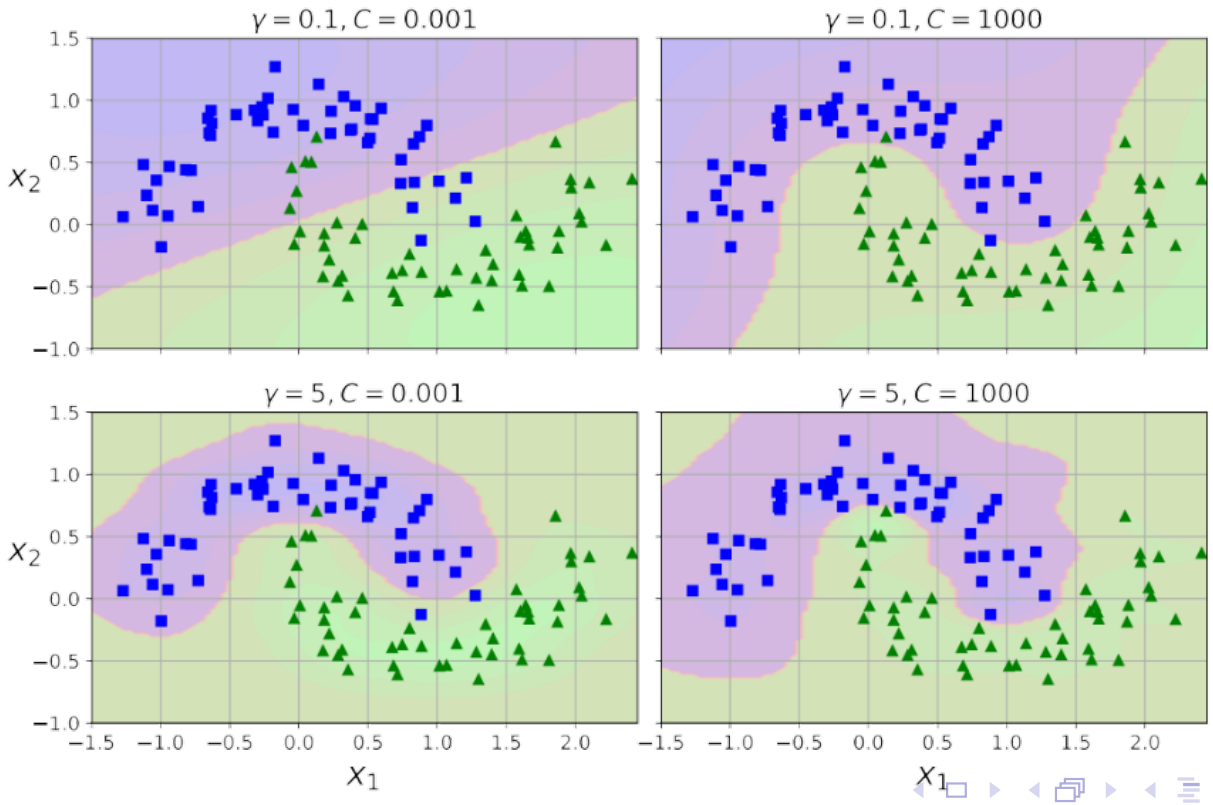
$$h_{\text{SVM}}(x) = \left(\sum_{i=1}^{N} \alpha_i K(x, x_i) \right) - \beta_0$$

- This small modification to the linear SVC decision function allows non-linear decision boundaries.
  - We call the resulting classifier a *Support Vector Machine*.

SVM Kernels

- A popular choice of kernel function $K(x, x_i)$ is the *radial basis function* (see plot):

$$K(x, x_i) = \exp\left(-\frac{(x - x_i)^T(x - x_i)}{2\sigma^2}\right)$$



Complexity of Linear SVCs and SVMs

- Calculating the computational complexity for linear SVCs and SVMs is outside the scope of this course (because the algorithms for solving the associated optimisation problems are outside scope).
- **Time complexity:**
  - Training complexity of nonlinear SVM is generally between $O(n^2)$ and $O(n^3)$ with n the number of training instances.
  - For the linear case, the algorithm used in LinearSVC by the liblinear implementation is much more efficient than its libsvm-based SVC counterpart and can scale almost linearly to millions of samples and/or features.
- **Space complexity** depends on the implementation. Naive implementations have $O(n^2)$ complexity, because the entire kernel matrix is stored.
- Computational efficiency of the linear SVC can be improved with stochastic gradient descent (next slide).

Stochastic Gradient Descent

- Stochastic Gradient Descent (SGD) is a very efficient approach to fitting linear SVCs.
  - Can also be used to speed up linear regression from week 6 (see the sklearn class `SGDRegressor`).
- The sklearn class `SGDClassifier` implements a stochastic gradient descent for linear SVCs.
- SGD time complexity is $O(npI)$, where:
  - n is the number of samples (i.e., height of the design matrix),
  - p is the sample dimensionality (i.e., width of the design matrix),
  - I is the number of iterations required for convergence (which is usually small).

Machine Learning for Text Analysis

6 Machine Learning for Text Analysis
- Vector Space Document Representations
- PCA for Document Corpora
- Measuring Document Similarity
- Document Classification
- Clustering Documents

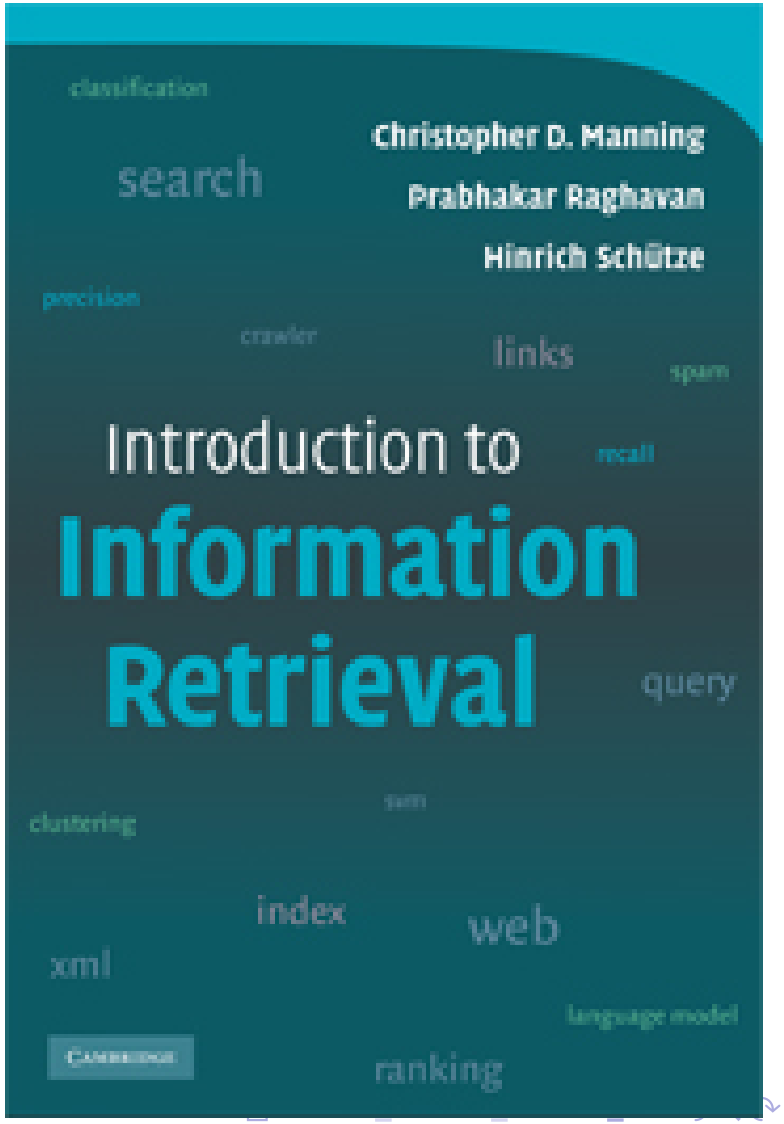
Attendance: 154797

Lecture Outline

- Today we'll look at applications of machine learning in text analysis.
- We'll apply **PCA** (week 5), clustering (week 6), and classification (week 8).
- We'll also see how to build a simple search engine.
- In order to apply machine learning for text analysis, we first need to represent text documents numerically.
  - For this, we'll look at *vector space representations* of text documents. In particular:
    - **Bag of words** document representation
    - **Term log-frequency** document representation
    - **Term frequency, inverse document frequency (TF-IDF)** representation
- We'll measure the similarity of documents using their **TF-IDF** representations with a distance measure called *cosine distance*, as well as the usual *Euclidean distance*.

Recommended Reading

- This lecture is self-contained. However, the following might be useful:
  - **Introduction to Information Retrieval**, Manning et al.
  - Free online
  - Also see Christopher Manning's YouTube lectures on NLP
  - In particular, lectures 105 – 111



Attendance: 154797

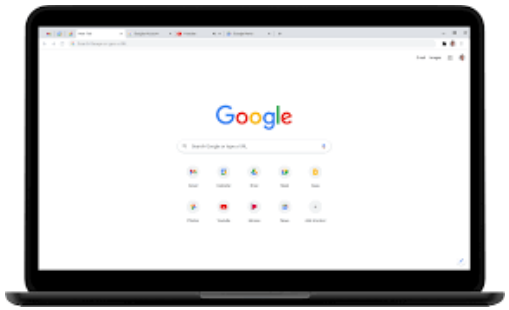
The Data

- Training data for text analysis is typically a **corpus** (i.e., a collection, plural *corpora*) of text documents

Emails, Messages, Social Media

Movie Reviews

News Articles



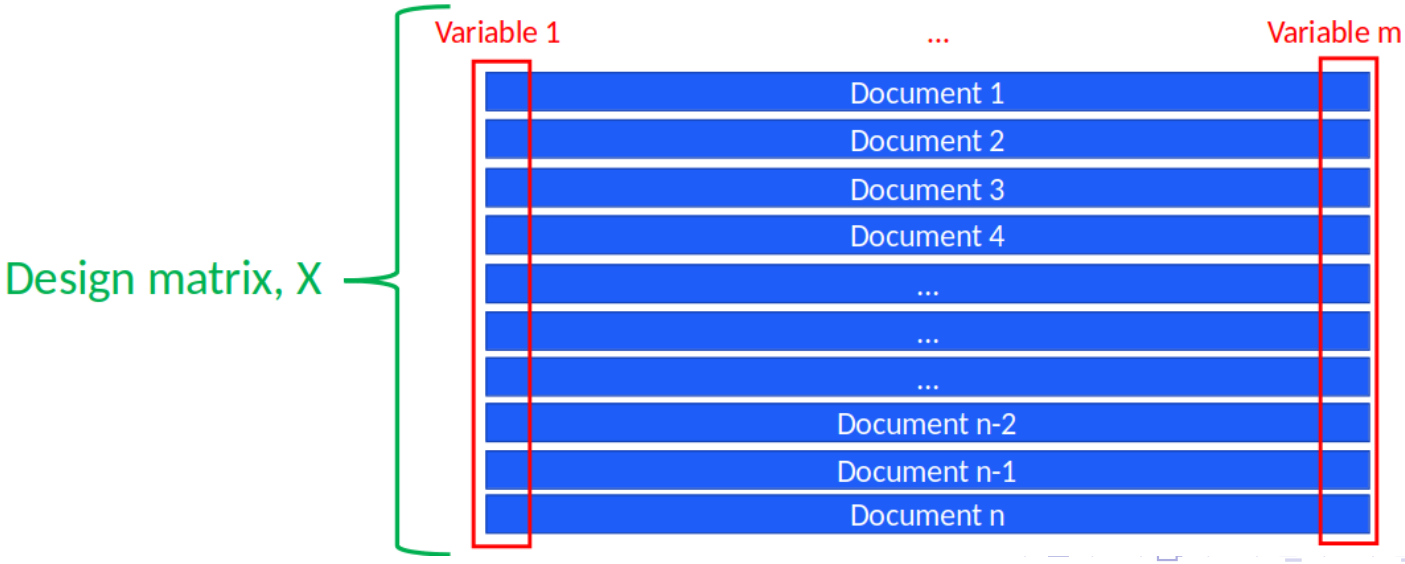
Webpages

The Goals

- How might we use these corpora?
- **Unsupervised learning:**
  - Principal component analysis
  - Topic discovery (AKA latent semantic analysis, latent semantic indexing)
- **Clustering**
  - Search enhancement
  - Keyword-free search
  - Faster search
  - Enriching search results
- **Supervised Learning:**
  - Classification
    - Sentiment analysis
    - Spam detection
    - Standing queries

Numerical Representations of Documents

- So, the training data for NLP is typically a **corpus** (i.e., a collection) of documents
- In order to use them for machine learning, we need to build a design matrix, X (week 1)
- For this, each document must be represented as a **vector** of m real variables (for some m)
- We refer to these representations as *vector space representations* of documents



Bag of Words

- One idea is to represent each document as a **"bag of words"**.
  - For each document, form a vector of the counts for each of the (e.g.) 20,000 most commonly occurring words in English.
  - The count of occurrences of a given word t (t stands for *term*) in a document d is referred to as the **raw term frequency**, $tf_{t,d}$.

Bag of Words

- This might seem like an overly-simplified representation:
  - We discard the *ordering* of words in a document.
  - E.g., “John is quicker than Mary” has the same representation as “Mary is quicker than John”.
- In practice, bag of words representations still retain a lot of usable information, despite throwing away so much structure.
- Hence, they’re used extensively in natural language processing.
- The design matrix is:

$$X = \begin{pmatrix} tf_{0,0} & \dots & tf_{20000,0} \\ \vdots & \ddots & \vdots \\ tf_{0,n} & \dots & tf_{20000,n} \end{pmatrix}$$

Zipf's Law and Log-Frequencies

- The bag of words representation has a drawback:
  - Common words tend to occur exponentially more often in document collections than less common words.
- This phenomenon is known as **Zipf's law**:
  - If t_1 is the most common term in the collection, t_2 is the next most common, and so on, then the collection frequency cf_i of the ith most common term is proportional to $\frac{1}{i}$.
- Hence, very common words such as "the" and "a" will tend to dominate bag of words representations.

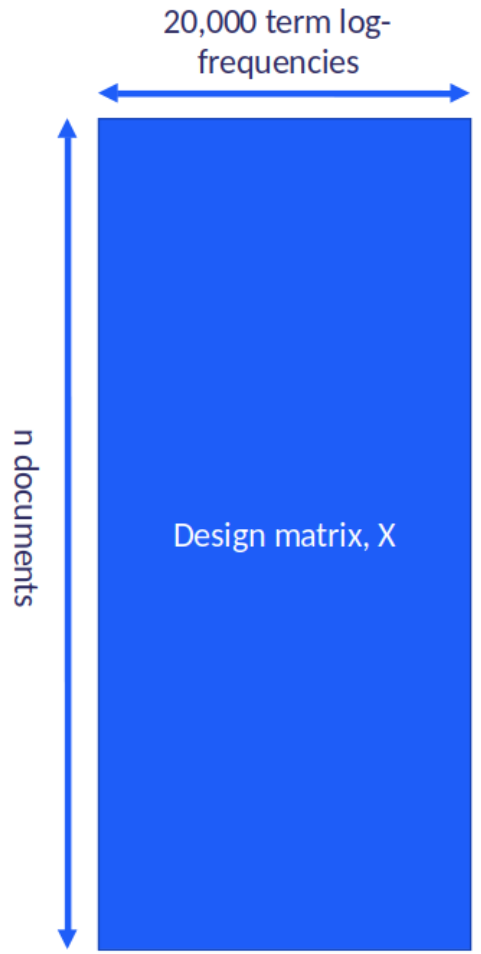
Zipf's Law and Log-Frequencies

Hence, it's common to apply a **log-frequency weighting** to the raw term frequencies:

$$w_{t,d} = \begin{cases} 1 + \log_{10}(tf_{t,d}) & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- The "otherwise" part is necessary because $\log_{10}(0) = -\infty$.
- The design matrix is:

$$X = \begin{pmatrix} w_{0,0} & \cdots & w_{20000,0} \\ \vdots & \ddots & \vdots \\ w_{0,n} & \cdots & w_{20000,n} \end{pmatrix}$$



Inverse Document Frequency

- Log-frequencies are a step in the right direction, but we can do better still...
- Rare terms tend to be more informative than frequent terms, so we'd like to make them more prominent in the document representation.
- We want to give high weights for rare terms and lower rates for frequent terms.
- We can measure the commonality of a word with **document frequency** df_t:
  - The document frequency of a term is the number of documents that contain the word at least once.

Inverse Document Frequency

- In practice, we use the so-called **inverse document frequency**:

$$idf_t = \log\left(\frac{N}{df_t}\right)$$

  - N is the number of documents in the corpus.
  - $\frac{df_t}{N}$ is the fraction of documents in which term t appears.
  - Log is used to dampen the effect of idf_t.
  - Using $\frac{N}{df_t}$ itself would be too strong a factor.

TF-IDF

- Combining the log frequency weighting with the inverse document frequency weighting, we get a value called the *term frequency, inverse document frequency* (TF-IDF):

$$w_{t,d} = \begin{cases} (1 + \log tf_{t,d}) \times \log\left(\frac{N}{df_t}\right) & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Again, the design matrix is:

$$X = \begin{pmatrix} w_{0,0} & \cdots & w_{20000,0} \\ \vdots & \ddots & \vdots \\ w_{0,n} & \cdots & w_{20000,n} \end{pmatrix}$$

- This is the most commonly-used vector space representation for documents.

Attendance: 154797

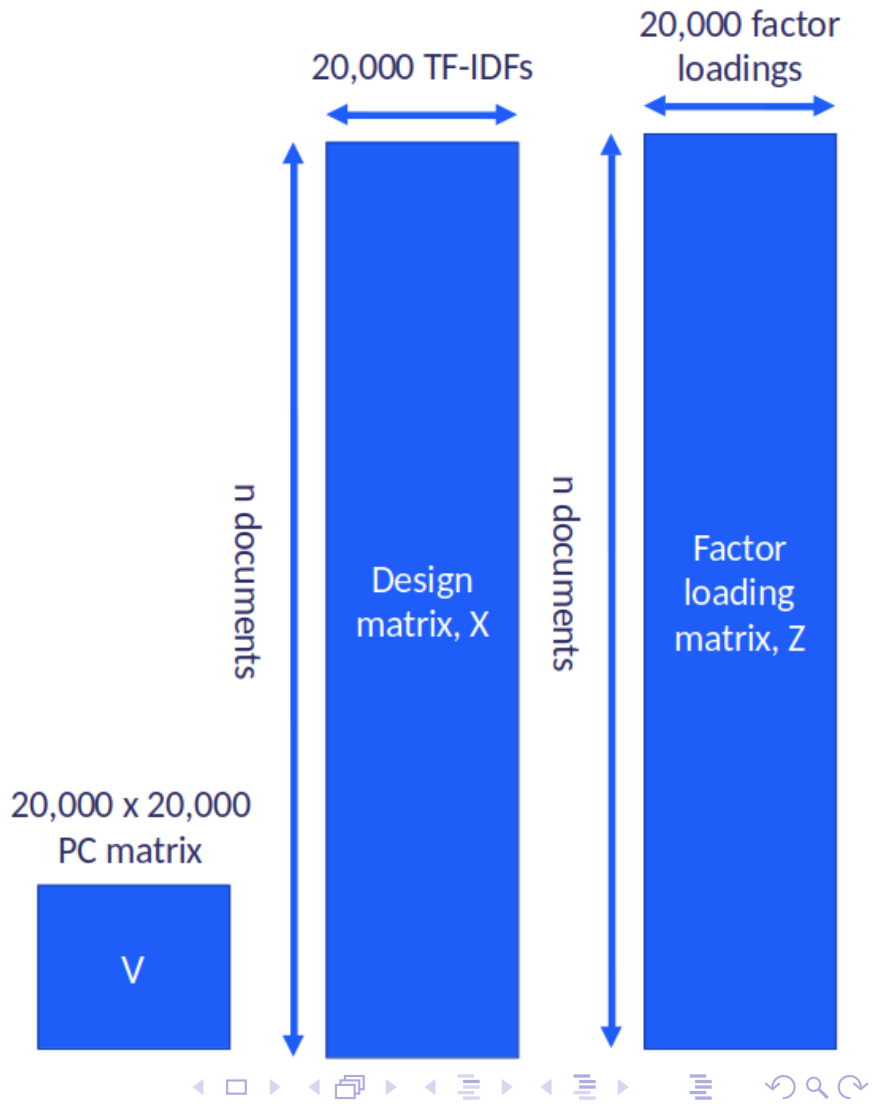
Latent Semantic Analysis

- We can apply PCA to "term log-frequency" design matrices.
  - In the jargon of natural language processing, this is called **latent semantic analysis (LSA)**, or sometimes **latent semantic indexing (LSI)**.
- When we do this, the principal components have an interesting interpretation...
  - Recall from weeks 5 and 6 that applying PCA to images of faces gave principal components that represented things like glasses, smiles, cheekbones, face width, nose length.
  - Applying PCA to term log-frequency design matrices gives principal components which represent document topics.

Latent Semantic Analysis

- Recall that the PC matrix is $\mathbf{V} = (\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_{20,000})$, where $\mathbf{v}_i$ is the ith principal component.
- We saw in week 5 that $\mathbf{X}$, $\mathbf{V}$ and $\mathbf{Z}$ are related by the equations:

$$\mathbf{Z} = \mathbf{XV}, \quad \mathbf{X} = \mathbf{ZV}^T$$



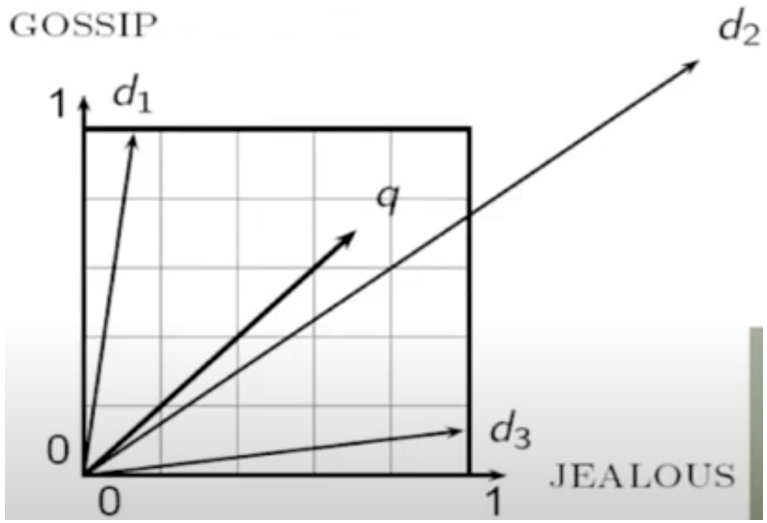
Attendance: 154797

Euclidean and Cosine Distances

- So far, we've seen a few vector space representations of documents, and we saw that the TF-IDF vector space representation is a particularly sensible one.
- How should we measure the distance between documents using their TF-IDF representations?
- E.g., imagine that we have a document q that reads "jealous and gossip", and we'd like to compare it to three other documents:
  - d_1 is a short document mostly about gossip.
  - d_2 is a long document about both jealousy and gossip.
  - d_3 is a short document mostly about jealousy.

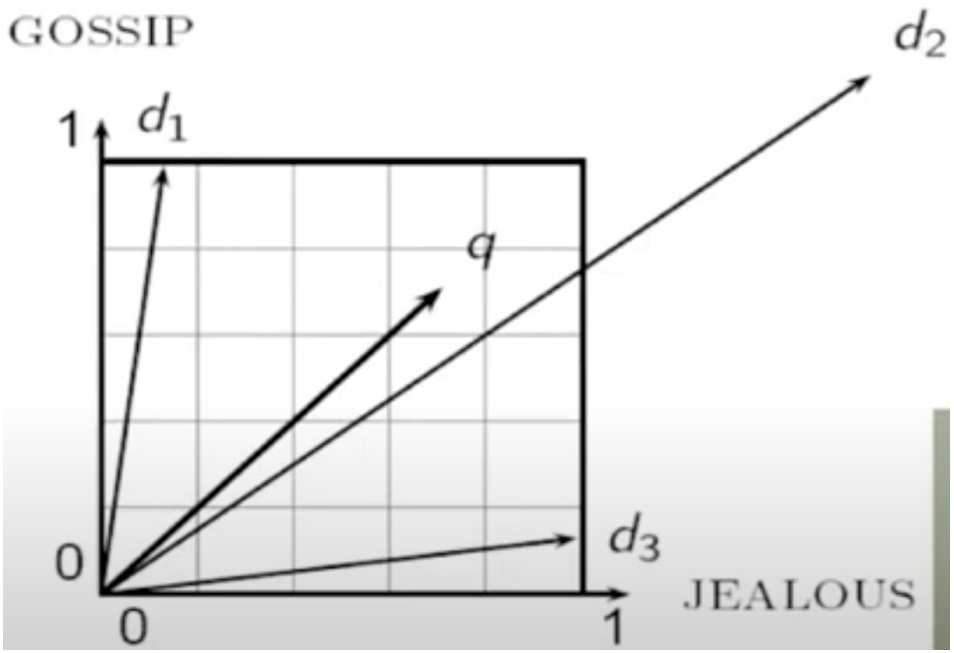
Euclidean and Cosine Distances

- q is closer in Euclidean distance to d_1 and d_3 than d_2.
- So, Euclidean distance isn't a great measure of similarity in this case.
- In this instance, a better measure of similarity would be the angle between vectors.
- We call this **cosine similarity**.
- **Exercise:** Consider the Euclidean and cosine distances between any document d and the document formed by appending d to itself. Which would be a more sensible measure of distance in this case?



Search Engines

- Cosine distance between TF-IDF document representations is one way to build a search engine.
- E.g., imagine that q from the previous example were a search term and d_1, d_2, and d_3 were webpages.



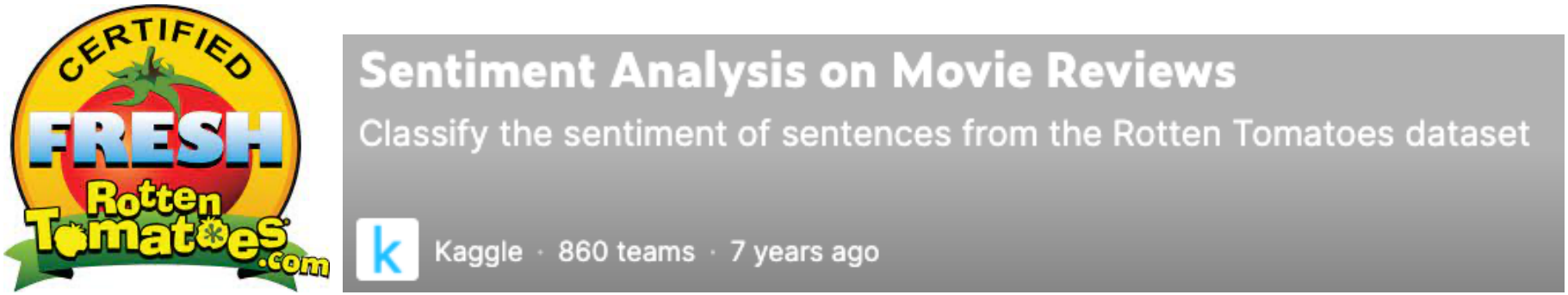
Attendance: 154797

Classification and the Contiguity Hypothesis

- In week 8, we saw a few algorithms for classification:
  - K-nearest neighbours
  - Linear support vector classifiers
  - Support vector machines
- For classification of documents, one should use a document representation that approximately obeys the **contiguity hypothesis**:
  - Documents in the same class form a contiguous region, and regions of different classes do not overlap.
  - TF-IDF generally provides good contiguity.
  - See Manning for more details if you're interested.
- Both SVMs and linear SVCs use Euclidean distance for classification, so for these algorithms, one would use Euclidean distance for TF-IDF vectors.
- KNN could support either Euclidean distance or cosine distance between TF-IDF vectors.

Sentiment Analysis on Movie Reviews

- In order to train a classifier, we need a body of documents which already has class labels.
- One standard example is a collection of movie reviews:
  - E.g., the Rotten Tomatoes movie review dataset on Kaggle.



Attendance: 154797

Clustering Documents

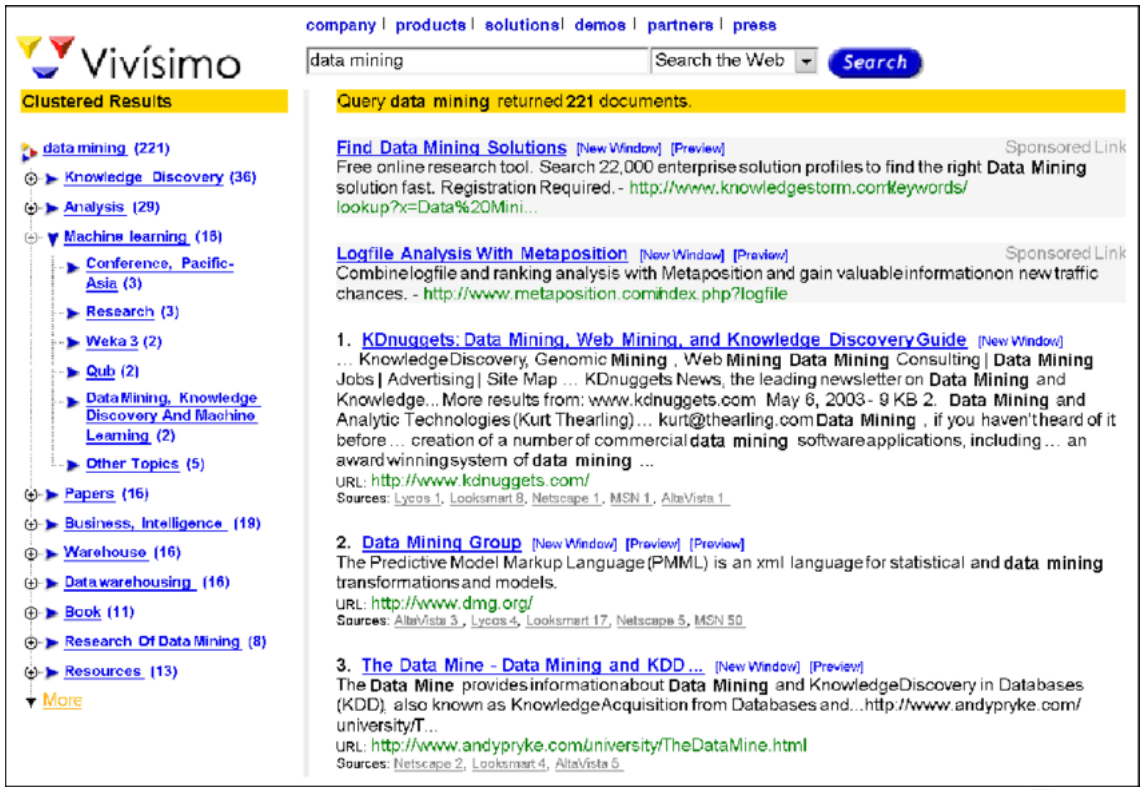
- We saw in week 6 that clustering can be applied to any dataset for which we have a dissimilarity measure.
  - We also saw that k-means only supports one dissimilarity measure: Euclidean distance.
- Today we have seen 2 ways to measure the dissimilarity between documents:
  - Euclidean distance between TF-IDF vectors.
  - Cosine distance between TF-IDF vectors.
- Hence, we can use these dissimilarity measures to cluster text documents with the usual clustering algorithms such as k-means, HAC, and DBSCAN.

Cluster Hypothesis for Information Retrieval

- So, we can cluster text documents. But why would we want to?
- Clustering can be useful for document search.
- This application is motivated by a truism called the **Cluster Hypothesis for Information Retrieval**:
  - *"Documents in the same cluster behave similarly with respect to information needs"*.
  - E.g., documents in the same cluster are relevant to the same search query.
  - See Manning for more discussion if you're interested.

Search Result Clustering

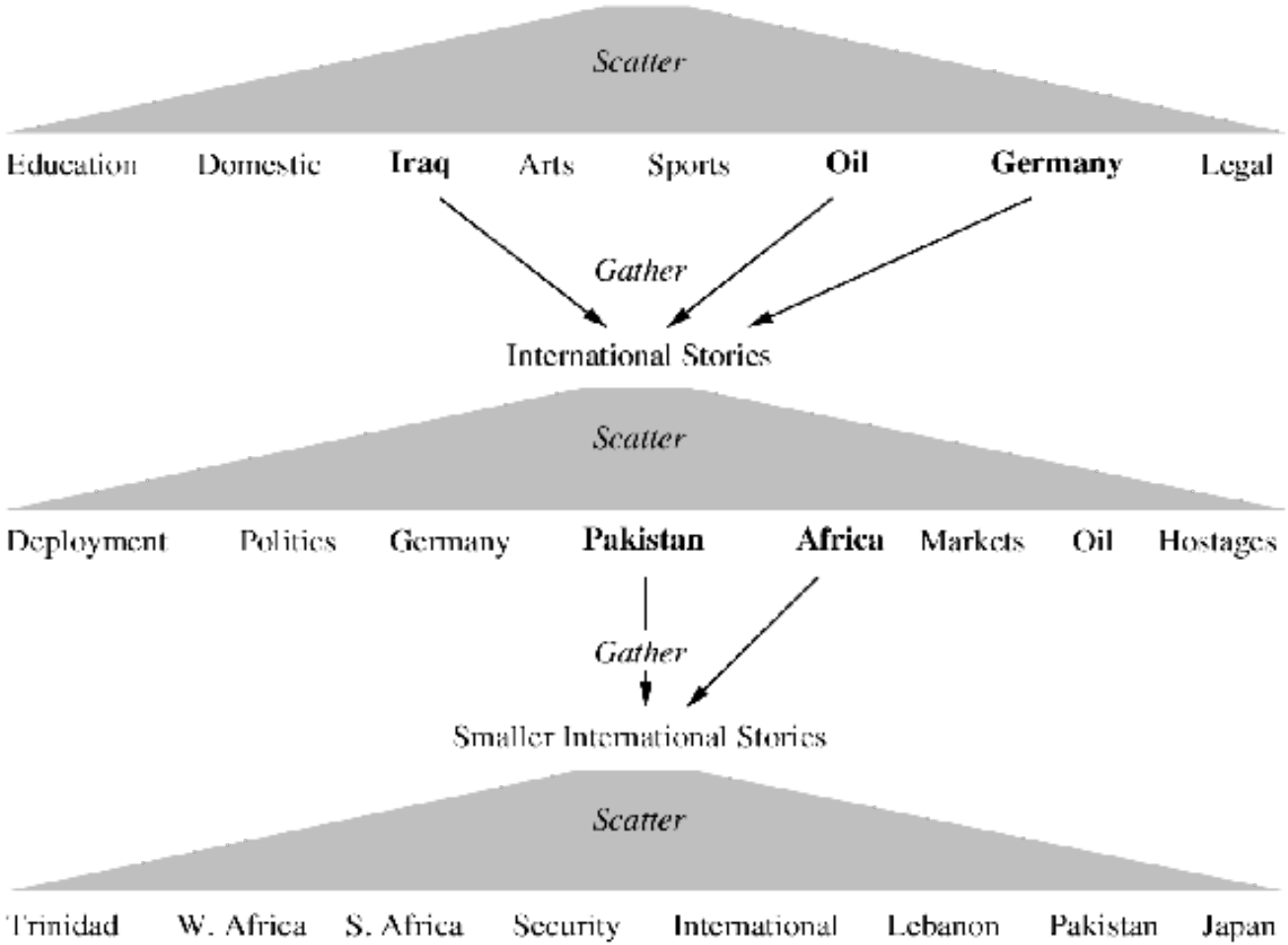
- Web search engines often retrieve more webpages than a user can realistically view.
- Clustering can be applied to these webpages to help the user.
- E.g., the Vivisimo search engine:



Scatter-Gather

- Provides an alternative to keyword search for navigating large bodies of text.
- The scatter-gather algorithm begins by splitting the document corpus into clusters.
- The user then picks a subset of the clusters.
- The user-selected clusters are gathered together and clustered again.
- This process is repeated until a cluster of interest is found.

Scatter-Gather



Collection Clustering

- Clustering can be used to enrich the results produced by a traditional search engine.
- E.g., imagine you have created a search engine that selects documents based on the cosine similarity between the TF-IDF representations of the user's search term and the documents in your corpus.
- For documents that were selected by your TF-IDF search engine, documents from the same cluster can also be added to the search results.
- This will enrich the search result by adding documents which are related to the search query but don't contain exactly the same words.

Cluster-Based Retrieval

- Clustering can be used to speed up search.
- Rather than searching the whole document corpus, find the clusters that most closely match the query and only consider the documents in those clusters.
- Offers a good speedup for a small decrease in the expected search quality.

Advanced Spark

Attendance: 154797

Recap

- In weeks 2 and 3, we looked at big data middleware
- **Hadoop:**
  - Provides a simple computational abstraction: map followed by reduce over distributed collections of (key, value) pairs
- **Spark:**
  - Extends the computational abstraction of Hadoop to include list programming functions (map, reduce, filter, fold, . . .) over **Resilient Distributed Datasets** (RDDs)
  - Also provides a performance boost over Hadoop

Today's Lecture

- Today, we'll see that Spark generalises the big data computational model further still.
- We'll see data structures and algorithms for:
  - **Machine learning** (Spark MLlib)
  - **Graph algorithms** (Spark GraphX)
  - **Streaming data** (Spark Streams)
- These packages are built on top of the core computational abstraction of Spark (which is the RDD).
- They also provide further performance gains.
- (Spark also supports SQL queries via the Spark SQL library. However, we'll skip this because databases are outside the scope of this module.)

Recommended Reading

- This lecture is self-contained.
- However, if you'd like to know more, example code and documentation for today can be found here:
  - https://spark.apache.org/docs/3.5.3/ml-classification-regression.html
  - https://spark.apache.org/graphx/
  - https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html
  - https://graphframes.github.io/graphframes/docs/_site/user-guide.html

Attendance: 154797

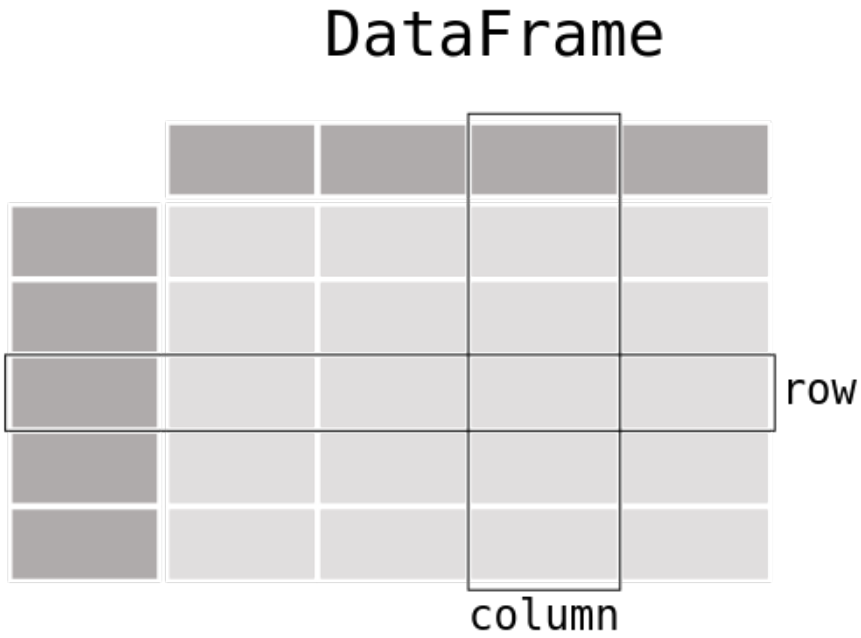
MLLib Overview

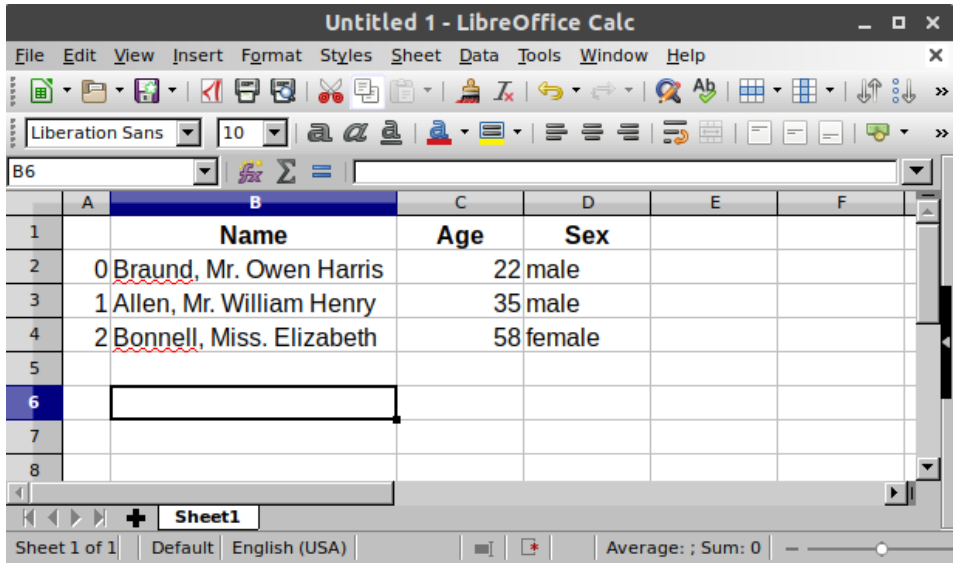
- MLLib is Spark's machine learning library.
- It provides:
  - Algorithms for supervised and unsupervised learning.
  - Tools for constructing end-to-end ML pipelines.
  - Tools for saving and loading algorithms, models, pipelines etc.
  - Utility functions for linear algebra, statistics, data handling etc.

Spark DataFrames

- Before getting into the nuts and bolts of MLLib, we'll lay some groundwork by introducing DataFrames...
- All the packages we'll see today depend on DataFrames.
- It is conceptually equivalent to a data frame in R/Python (or a table in a relational database).
- DataFrames are built on top of RDDs, and are distributed collections of rows with named columns.
- As with RDDs, they are:
  - **Immutable** (i.e. you can't change a DataFrame once it has been created).
  - **Lazily evaluated** (i.e. a function on RDDs is not executed until its result is required).
  - **Distributed** over nodes in a Spark cluster.
- DataFrames can be created in the following ways:
  - Loading data stored in various formats (e.g. JSON, CSV, XML, Parquet).
  -

DataFrames


DataFrame
row
column


Untitled 1 - LibreOffice Calc
File Edit View Insert Format Styles Sheet Data Tools Window Help
Liberation Sans 10
B6
A B C D E F
1 Name Age Sex
2 0 Braund, Mr. Owen Harris 22 male
3 1 Allen, Mr. William Henry 35 male
4 2 Bonnell, Miss. Elizabeth 58 female
5
6
7
8
Sheet1
Sheet 1 of 1 Default English (USA) Average: ; Sum: 0

Other MLLib Data Structures

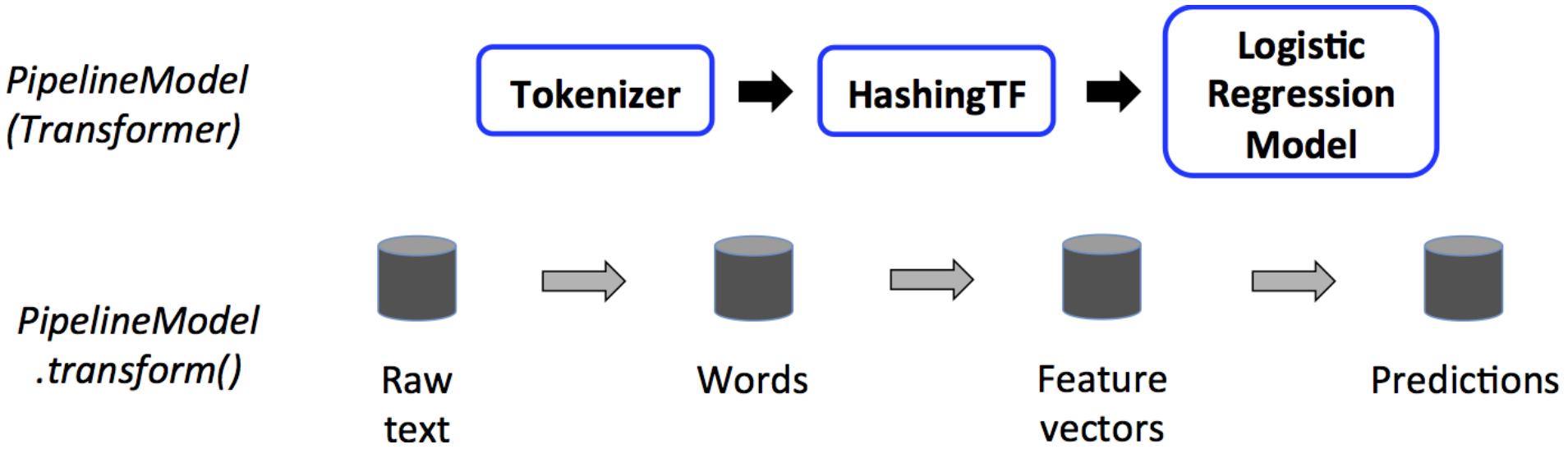
- MLLib also supports some other data structures:
- **Vectors:**
  - These can be dense or sparse, and local to a single spark node or distributed across several nodes
- **Labelled points**
  - A local vector, dense or sparse, which is associated with a response
- **Local matrices**
- **Distributed matrices, of which there are several types:**
  - **RowMatrix**
    - Backed by an RDD of rows
  - **IndexedRowMatrix**
    - Backed by an RDD of indexed rows
  - **CoordinatedMatrix**
    - Backed by an RDD of its individual entries
  - **Block matrices**
    - Backed by an RDD of matrix blocks - (int, int, matrix)

Machine Learning Pipelines

- In machine learning, it is common to run a sequence of steps to process and learn from data.
- E.g., a simple text document processing workflow might include several stages:
  - Split each document's text into words.
  - Convert each document's words into a numerical feature vector.
  - Learn a prediction model using the feature vectors and labels.
- More generally, the usual ML pipeline proceeds as follows:
  1. Load and Clean Data
  2. Extract Features
  3. Train Model
  4. Evaluate Model
  5. Repeat as necessary

MLLib Pipelines

- MLLib represents such a workflow as a Pipeline, which consists of a sequence of PipelineStages to be run in a specific order.
- PipelineStages can be *Transformers* and *Estimators*, as we'll see in the next few slides.



Transformers

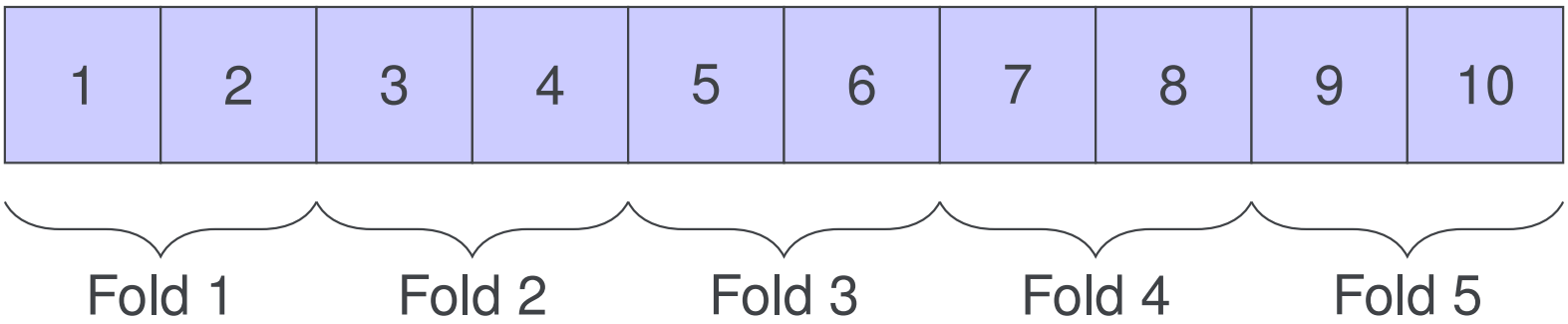
- Transformers are pipeline stages that transform DataFrames in some way:
  - E.g. feature transformation, model-based estimation etc.
- In Spark, a Transformer implements the method `transform()`, which converts one DataFrame into another, generally by appending one or more columns. For example:
  - A feature transformer might take a DataFrame, read a column (e.g., text), map it into a new column (e.g., feature vectors), and output a new DataFrame with the mapped column appended.
  - A learning model might take a DataFrame, read the column containing feature vectors, predict the label for each feature vector, and output a new DataFrame with predicted labels appended as a column.

Evaluator

- Evaluators will evaluate the performance of a model, based on a certain performance metric:
  - E.g. mean squared error between predicted responses and actual responses
- Evaluator classes have an `evaluate()` method, which takes a DataFrame and returns a double representing the evaluation metric.
- Examples:
  - `BinaryClassificationEvaluator`
  - `CrossValidator`

Introduction to Cross-Validation

- **Cross-Validation:** is a technique used to evaluate the performance of a machine learning model.
- **Purpose:** To assess how the model will generalize to an independent dataset.
- **Process:**
  1. Split the dataset into k equal-sized folds.
  2. For each fold:
     - Use the fold as the test set.
     - Use the remaining $k-1$ folds as the training set.
     - Train the model on the training set and evaluate it on the test set.
  3. Calculate the average performance across all k folds.



Cross-Validation

- **Benefits:**
  - Provides a more reliable estimate of model performance.
  - Reduces the risk of overfitting.
  - Utilizes the entire dataset for both training and testing.
- **Common Variants:**
  - **K-Fold Cross-Validation:** The dataset is divided into k folds.
  - **Leave-One-Out Cross-Validation (LOOCV):** Each data point is used as a test set once.

Automated Model Tuning

- An important task in ML is model selection, when one attempts to find the best model or parameters for a given task. (This is also called tuning.)
- In MLLib, tuning may be done for individual Estimators such as `LogisticRegression`, or for entire Pipelines which include multiple algorithms, feature extraction, etc.
  - Hence, users can tune an entire Pipeline at once, rather than tuning each element in the Pipeline separately.
- MLLib supports model selection using tools such as `CrossValidator` and `TrainValidationSplit`. These tools require the following items:
  - An *Estimator*, which contains an algorithm or Pipeline to tune.
  - A set of *ParamMaps*, specifying the parameters to choose from, sometimes called a "parameter grid" to search over.
  - An *Evaluator*, which implements the evaluation metric which should be measured in order to assess how well a fitted Model is performing.

Automated Model Tuning

- Spark's model selection tools work as follows:
  1. Split the input data into separate training and test datasets.
  2. For each (training, test) pair, iterate through the set of ParamMaps:
     - For each ParamMap, fit the Estimator using those parameters, get the fitted Model, and evaluate the Model's performance using the Evaluator.
  3. Select the Model produced by the best-performing set of parameters.
- You'll see more about model tuning next week.

K-Means Data

kmeans_data.txt

```
0.0 0.0 0.0
0.1 0.1 0.1
0.2 0.2 0.2
9.0 9.0 9.0
9.1 9.1 9.1
9.2 9.2 9.2
```

K-Means Clustering

Code Example

```
from pyspark.ml.clustering import KMeans
from pyspark.ml.evaluation import ClusteringEvaluator

dataset = spark.read.format("libsvm")\
    .load(f"{data_dir}/mllib/sample_kmeans_data.txt")

print(f"The type of 'dataset' is: {type(dataset)}")
dataset.show(10, False)

# Trains a k-means model.
kmeans = KMeans().setK(2).setSeed(1)
model = kmeans.fit(dataset)

# Make predictions
predictions = model.transform(dataset)

# Evaluate clustering by computing Silhouette score
evaluator = ClusteringEvaluator()
silhouette = evaluator.evaluate(predictions)
print("Silhouette with squared euclidean distance = " + str(silhouette))

# Shows the result.
centers = model.clusterCenters()
print("Cluster Centers: ")
for center in centers:
    print(center)
```

K-Means Clustering

Output

```
The type of 'dataset' is: <class 'pyspark.sql.dataframe.DataFrame'>
+-----+--------------------+
|label|features            |
+-----+--------------------+
|  0.0| (3,[],[])          |
|  1.0| (3,[0,1,2],[0.1... |
|  2.0| (3,[0,1,2],[0.2... |
|  3.0| (3,[0,1,2],[9.0... |
|  4.0| (3,[0,1,2],[9.1... |
|  5.0| (3,[0,1,2],[9.2... |
+-----+--------------------+

Silhouette with squared euclidean distance = 0.9997530305375207
Cluster Centers:
[9.1, 9.1, 9.1]
[0.1, 0.1, 0.1]
```

Sample Linear Regression Data

sample_linear_regression_data.txt

```
-9.490009878824548 1:0.4551273600657362 2:0.36644694351969087 3:-0.38256108933468047 4:-0.4458430198517267 5:0.33109790358914726 6:0.8
0.2577820163584905 1:0.8386555657374337 2:-0.1270180511534269 3:0.499812362510895 4:-0.22686625128130267 5:-0.6452430441812433 6:0.188
-4.438869807456516 1:0.5025608135349202 2:0.14208069682973434 3:0.16004976900412138 4:0.505019897181302 5:-0.9371635223468384 6:-0.284
-19.782762789614537 1:-0.0388509668871313 2:-0.4166870051763918 3:0.8997202693189332 4:0.6409836467726933 5:0.273289095712564 6:-0.261
-7.966593841555266 1:-0.06195495876886281 2:0.6546448480299902 3:-0.6979368909424835 4:0.6677324708883314 5:-0.07938725467767771 6:-0.
-7.896274316726144 1:-0.15805658673794265 2:0.26573958270655806 3:0.3997172901343442 4:-0.3693430998846541 5:0.14324061105995334 6:-0.
-8.464803554195287 1:0.39449745853945895 2:0.817229160415142 3:-0.6077058562362969 4:0.6182496334554788 5:0.2558665508269453 6:-0.0732
2.1214592666251364 1:-0.005346215048158909 2:-0.9453716674280683 3:-0.9270309666195007 4:-0.032312290091389695 5:0.31010676221964206 6
1.0720117616524107 1:0.7880855916368177 2:0.19767407429003536 3:0.9520689432368168 4:-0.845829774129496 5:0.5502413918543512 6:-0.4423
-13.772441561702871 1:-0.3697050572653644 2:-0.11452811582755928 3:-0.807098168238352 4:0.4903066124307711 5:-0.6582805242342049 6:0.6
...
```

Linear Regression

Code Example

```
from pyspark.ml.regression import LinearRegression

# Load training data
training = spark.read.format("libsvm")\
    .load(f"{data_dir}/mllib/sample_linear_regression_data.txt")

print(f"The type of 'training' is: {type(training)}")
training.show()

lr = LinearRegression(maxIter=10, regParam=0.3, elasticNetParam=0.8)

# Fit the model
lrModel = lr.fit(training)

# Print the coefficients and intercept for linear regression
print("Coefficients: %s" % str(lrModel.coefficients))
print("Intercept: %s" % str(lrModel.intercept))

# Summarize the model over the training set and print out some metrics
trainingSummary = lrModel.summary
```

Output

```
The type of 'training' is: <class 'pyspark.sql.dat
+--------------------+--------------------+
|               label|            features|
+--------------------+--------------------+
|  -9.490009878824548| (10,[0,1,2,3,4,5...|
|  0.2577820163584932| (10,[0,1,2,3,4,5...|
|  -4.438869807455616| (10,[0,1,2,3,4,5...|
| -19.782762789614537| (10,[0,1,2,3,4,5...|
|  -7.965938481555266| (10,[0,1,2,3,4,5...|
+--------------------+--------------------+

Coefficients: [0.0,0.3229251667740594,...]
Intercept: 0.15989368442397356
```

Linear Support Vector Classifier

Code Example

```
from pyspark.ml.classification import LinearSVC

# Load training data
training = spark.read.format("libsvm")\
    .load(f"{data_dir}/mllib/sample_libsvm_data.txt")

print(f"The type of 'training' is: {type(training)}")
training.show()

lsvc = LinearSVC(maxIter=10, regParam=0.1)

# Fit the model
lsvcModel = lsvc.fit(training)

# Print the coefficients and intercept for linear SVC
print("Coefficients: " + str(lsvcModel.coefficients))
print("Intercept: " + str(lsvcModel.intercept))
```

Output

```
The type of 'training' is: <class 'pyspark.sql.dataframe.DataFrame'>
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  0.0|(692,[127,128,129...|
|  1.0|(692,[158,159,160...|
|  1.0|(692,[124,125,126...|
|  1.0|(692,[152,153,154...|
|  1.0|(692,[151,152,153...|
|  0.0|(692,[129,130,131...|
|  1.0|(692,[158,159,160...|
|  1.0|(692,[99,100,101,...|
|  0.0|(692,[154,155,156...|
|  0.0|(692,[127,128,129...|
|  1.0|(692,[154,155,156...|
|  0.0|(692,[153,154,155...|
|  0.0|(692,[151,152,153...|
|  1.0|(692,[129,130,131...|
|  0.0|(692,[154,155,156...|
|  1.0|(692,[150,151,152...|
|  0.0|(692,[124,125,126...|
|  0.0|(692,[152,153,154...|
|  1.0|(692,[97,98,99,12...|
|  1.0|(692,[124,125,126...|
+-----+--------------------+
only showing top 20 rows
Coefficients: [0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
Intercept: 0.5232286178786085
```

Attendance: 154797

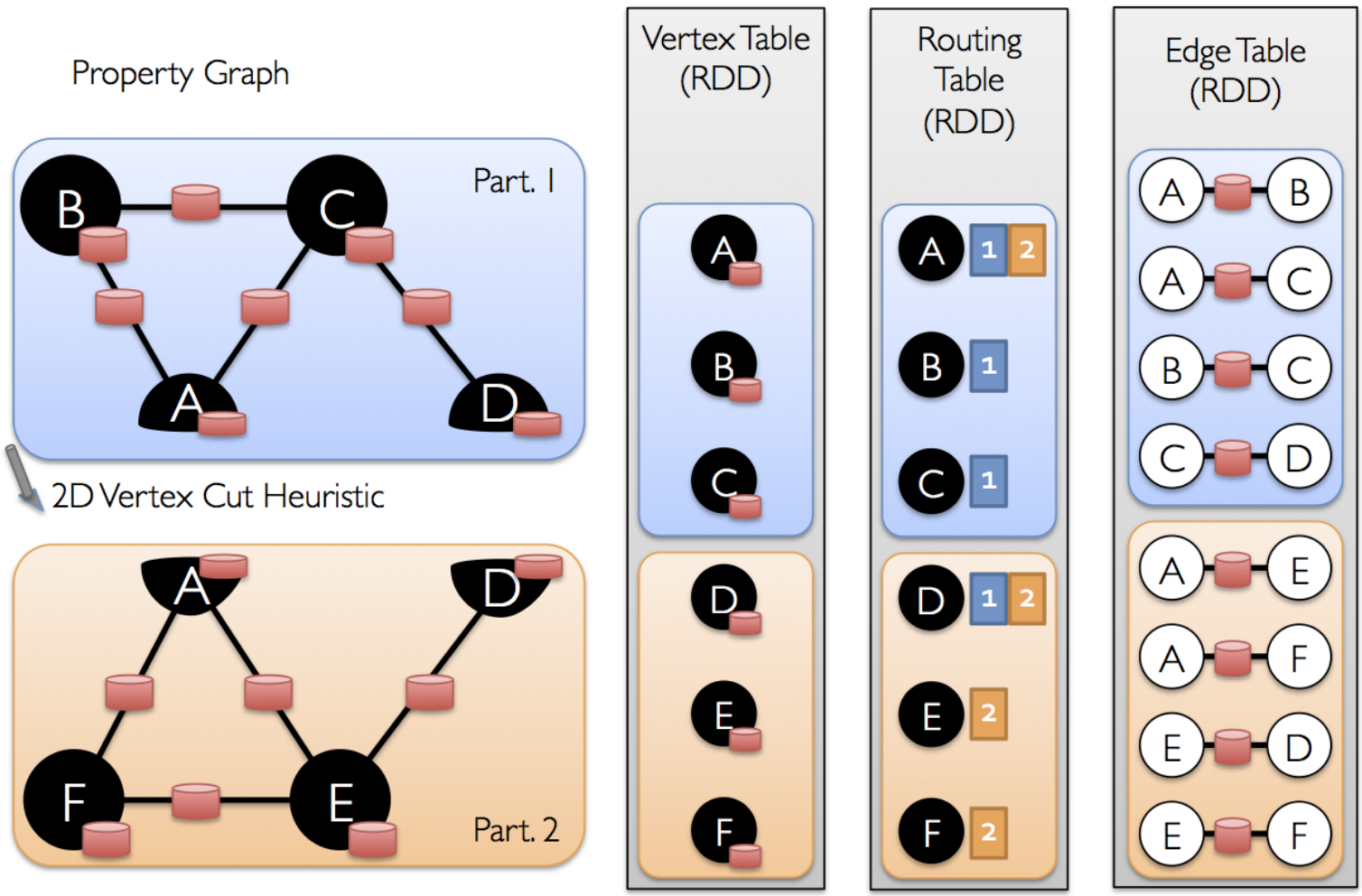
Introduction

- GraphX is a library for graph-based computation in Spark.
- It is based around directed multigraphs:
  - Directed graphs have directions associated with the edges.
  - Multigraphs can have multiple parallel edges going between the same vertices.
- Specifically, GraphX uses **property graphs**, which are directed multigraphs with user-defined objects attached to each vertex and edge.
- The ability to support parallel edges simplifies modeling scenarios where there can be multiple relationships (e.g., co-worker and friend) between the same vertices.
- Each vertex is keyed by a unique 64-bit long identifier (`VertexId`).
- Similarly, edges have corresponding source and destination vertex identifiers.

Data Structures for Property Graphs

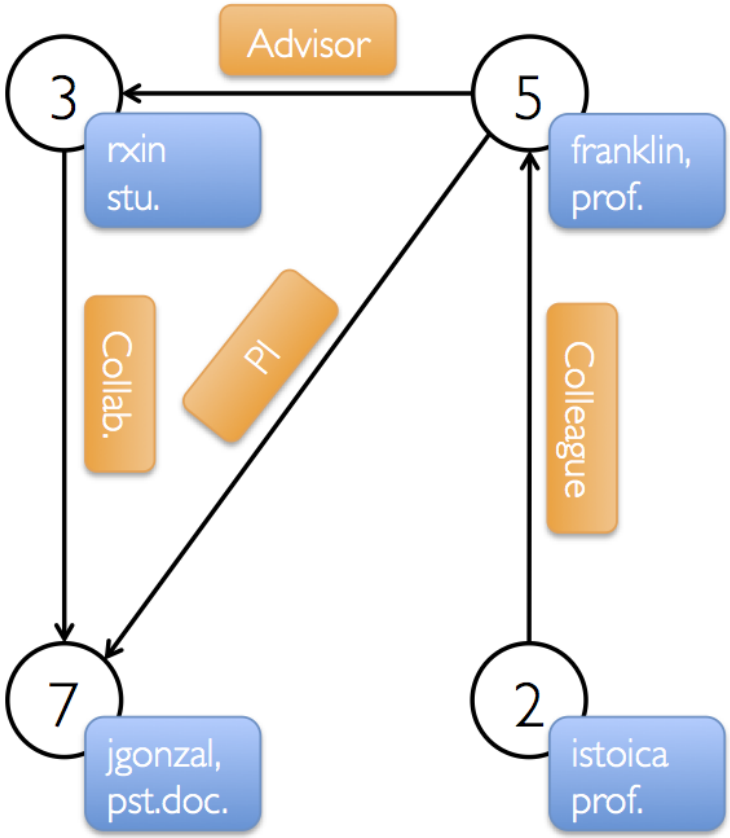
- Property graphs are represented with the following RDDs:
  - **Vertex table:**
    - Stores the vertices and their properties
  - **Edge table:**
    - Stores the edges and their properties
  - **Routing Table:**
    - Stores information about which partitions messages should be sent to
  - **Edge triplets:**
    - Derived from the above data structures
    - Stores the graph's edges, along with their adjacent vertex properties
    - Triplets are a 3-way join of edges with the vertices on each side

Data Structures for Property Graphs II



Property Graphs Example

Property Graph



Vertex Table

Id	Property (V)
3	(rxin, student)
7	(jgonzal, postdoc)
5	(franklin, professor)
2	(istoica, professor)

Edge Table

SrcId	DstId	Property (E)
3	7	Collaborator
5	3	Advisor
2	5	Colleague
5	7	PI

GraphX Algorithms

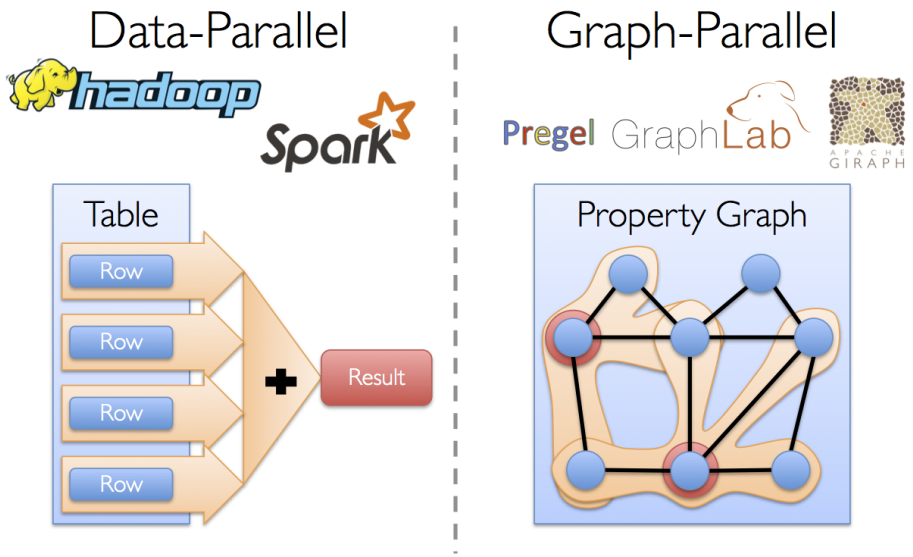
- GraphX algorithms are based around distributed message passing inside multigraphs.
- To support this, GraphX exposes an operator called `aggregateMessages`.
  - As we'll see, `aggregateMessages` is analogous to MapReduce, except it runs on graphs instead of (key, value) collections.
- `aggregateMessages` takes two user-defined functions:
  - **sendMsg:**
    - Analogous to the map function in map-reduce.
    - Takes an `EdgeContext`, which exposes the source and destination attributes along with the edge attribute and functions to send messages to the source and destination attributes (called `sendToSrc` and `sendToDst`).
  - **mergeMsg:**
    - Analogous to the reduce function in map-reduce.
    - Takes two messages destined to the same vertex and yields a single message.

GraphX Algorithms

- `aggregateMessages` works by:
  - Scanning the triplets on each edge partition
  - Running the `sendMsg` function for each of those triplets
  - Aggregating the messages to get a result
- The `aggregateMessages` operator returns a `VertexRDD[Msg]` containing the aggregate message (of type `Msg`) destined to each vertex. Vertices that did not receive a message are not included in the returned `VertexRDD`.

Distribution: The Graph Parallel Pattern

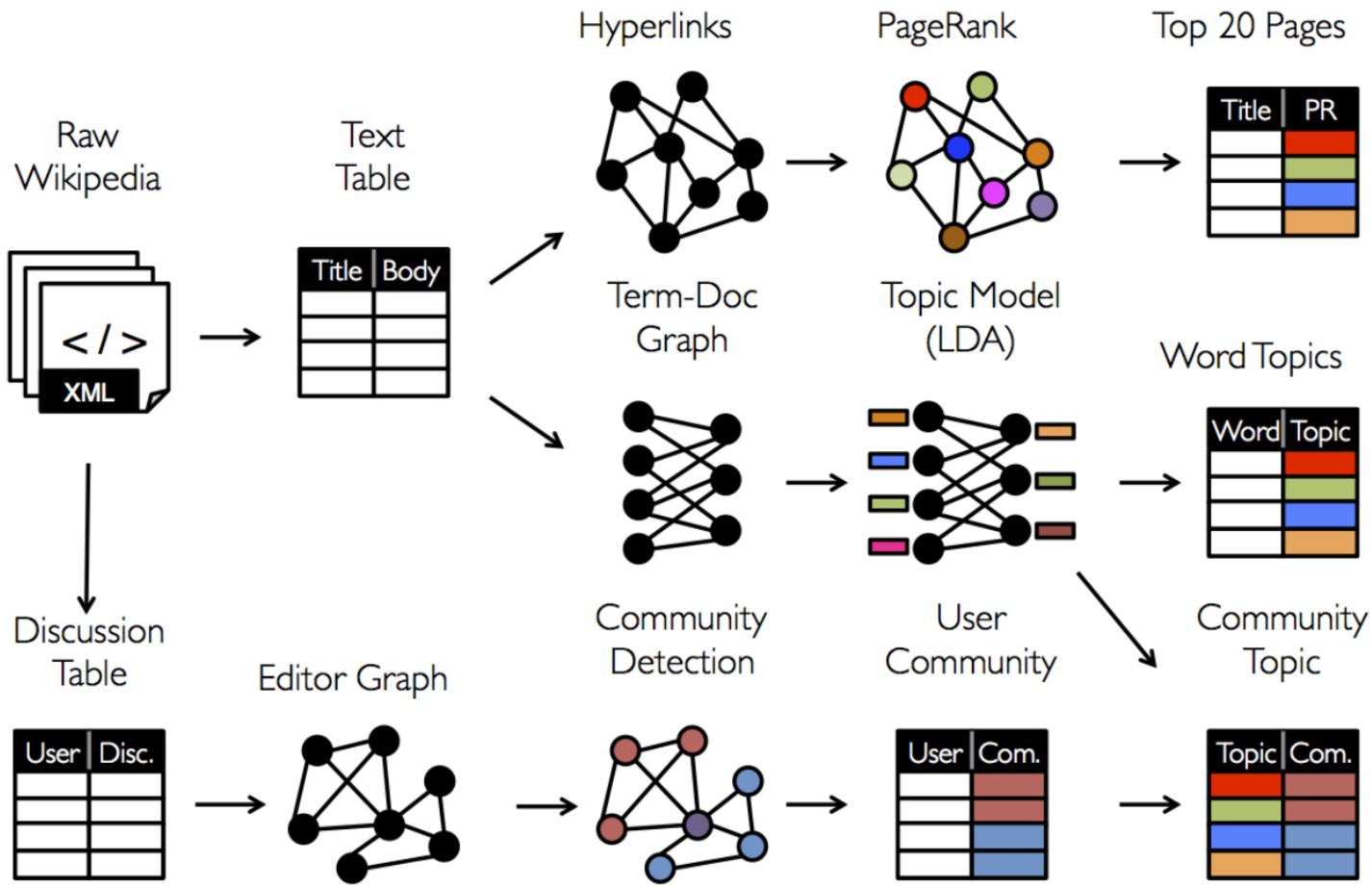
- Message passing graph algorithms can be parallelised because each node will only receive messages from nodes which are close to it in the graph.
- GraphX relies on this fact to distribute graph algorithm computations across the nodes in a Spark cluster.
- In fact, GraphX was created to express this pattern.
  - It's called the **graph parallel pattern**.



Motivating Examples

- Examples of graph problems in big data:
  - Social networks: community cohesiveness; page rank
  - Website links: page rank
  - Wikipedia links: graphs of top pages and top editors; community detection for editors
- Page rank
  - Uses links as a vote of importance
  - Link counts for more if it's coming from an important page
- Triangle counting
  - Measures the cohesiveness of communities
  - Strong communities have more triangles – e.g., for social networks, a person's friends tend to know each other
- Message passing is core to these and other graph algorithms, and hence they can be implemented in GraphX, as we shall see...

Graph-Based Examples for Web Data



Basic Graph Operations

Code Example

```
from graphframes import *
from graphframes.examples import Graphs

# Create a Vertex DataFrame with unique ID column "id"
v = sqlContext.createDataFrame([
    ("a", "Alice", 34),
    ("b", "Bob", 36),
    ("c", "Charlie", 30),
], ["id", "name", "age"])

# Create an Edge DataFrame with "src" and "dst" columns
e = sqlContext.createDataFrame([
    ("a", "b", "friend"),
    ("b", "c", "follow"),
    ("c", "b", "follow"),
], ["src", "dst", "relationship"])

# Create a GraphFrame
g = GraphFrame(v, e)
g.vertices.show()
g.edges.show()

# Query: Get in-degree of each vertex.
g.inDegrees.show()
```

Output

```
+---+-------+---+
| id|   name|age|
+---+-------+---+
|  a|  Alice| 34|
|  b|    Bob| 36|
|  c|Charlie| 30|
+---+-------+---+

+---+---+------------+
|src|dst|relationship|
+---+---+------------+
|  a|  b|      friend|
|  b|  c|      follow|
|  c|  b|      follow|
+---+---+------------+

+---+--------+
| id|inDegree|
+---+--------+
|  b|       2|
|  c|       1|
+---+--------+
```

A Larger Graph

Code Example

```
from graphframes import *
from graphframes.examples import Graphs

g = Graphs(sqlContext).friends() # Get example graph

# Display the vertex and edge DataFrames
g.vertices.show()
g.edges.show()
g.inDegrees.show()

# Find the youngest user's age in the graph.
# This queries the vertex DataFrame.
g.vertices.groupBy().min("age").show()
```

Output 1

```
+---+-------+---+
| id|   name|age|
+---+-------+---+
|  a|  Alice| 34|
|  b|    Bob| 36|
|  c|Charlie| 30|
|  d|  David| 29|
|  e| Esther| 32|
|  f|  Fanny| 36|
+---+-------+---+

+---+---+------------+
|src|dst|relationship|
+---+---+------------+
|  a|  b|      friend|
|  b|  c|      follow|
|  c|  b|      follow|
|  f|  c|      follow|
|  e|  f|      follow|
|  e|  d|      friend|
|  d|  a|      friend|
+---+---+------------+
```

Output 2

```
+---+--------+
| id|inDegree|
+---+--------+
|  b|       2|
|  c|       2|
|  f|       1|
|  d|       1|
|  a|       1|
+---+--------+

+--------+
|min(age)|
+--------+
|      29|
+--------+
```

Shortest Paths and Triangle Count

- A vertex is part of a triangle when it has two adjacent vertices with an edge between them.
- GraphX implements a triangle counting algorithm in the `TriangleCount` object that determines the number of triangles passing through each vertex, providing a measure of clustering.
- We compute the triangle count of the social network from the previous slide.

Shortest Paths and Triangle Count

Code Example

```
results = g.shortestPaths(landmarks=["a", "d"])
results.select("id", "distances").show()

results = g.triangleCount()
results.select("id", "count").show()
```

Output

```
+---+----------------+
| id|       distances|
+---+----------------+
|  f|              {}|
|  e|{a -> 2, d -> 1}|
|  d|{a -> 1, d -> 0}|
|  c|              {}|
|  b|              {}|
|  a|       {a -> 0} |
+---+----------------+

+---+-----+
| id|count|
+---+-----+
|  a|    0|
|  b|    0|
|  c|    0|
|  d|    0|
|  e|    0|
|  f|    0|
+---+-----+
```

PageRank

- PageRank measures the importance of each vertex in a graph, assuming an edge from u to v represents an endorsement of v's importance by u.
  - For example, if a Twitter user is followed by many others, the user will be ranked highly.
- GraphX comes with static and dynamic implementations of PageRank as methods on the `PageRank` object.
- Static PageRank runs for a fixed number of iterations, while dynamic PageRank runs until the ranks converge (i.e., stop changing by more than a specified tolerance).

PageRank

Code Example

```
results = g.pageRank(resetProbability=0.01, maxIter=20)
results.vertices.select("id", "pagerank").show()
```

Output

```
+---+------------------+
| id|          pagerank|
+---+------------------+
|  f|0.014950000000000028|
|  e|0.010000000000000018|
|  d|0.014950000000000028|
|  c|  2.965195        |
|  b|  2.9701          |
|  a|0.024800500000000048|
+---+------------------+
```

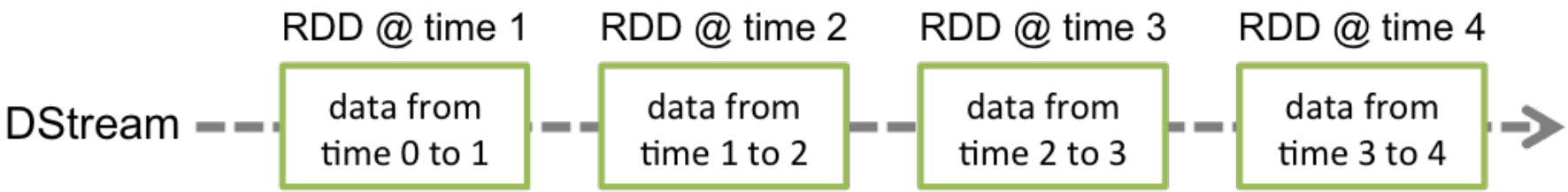
Attendance: 154797

Spark Streaming Introduction

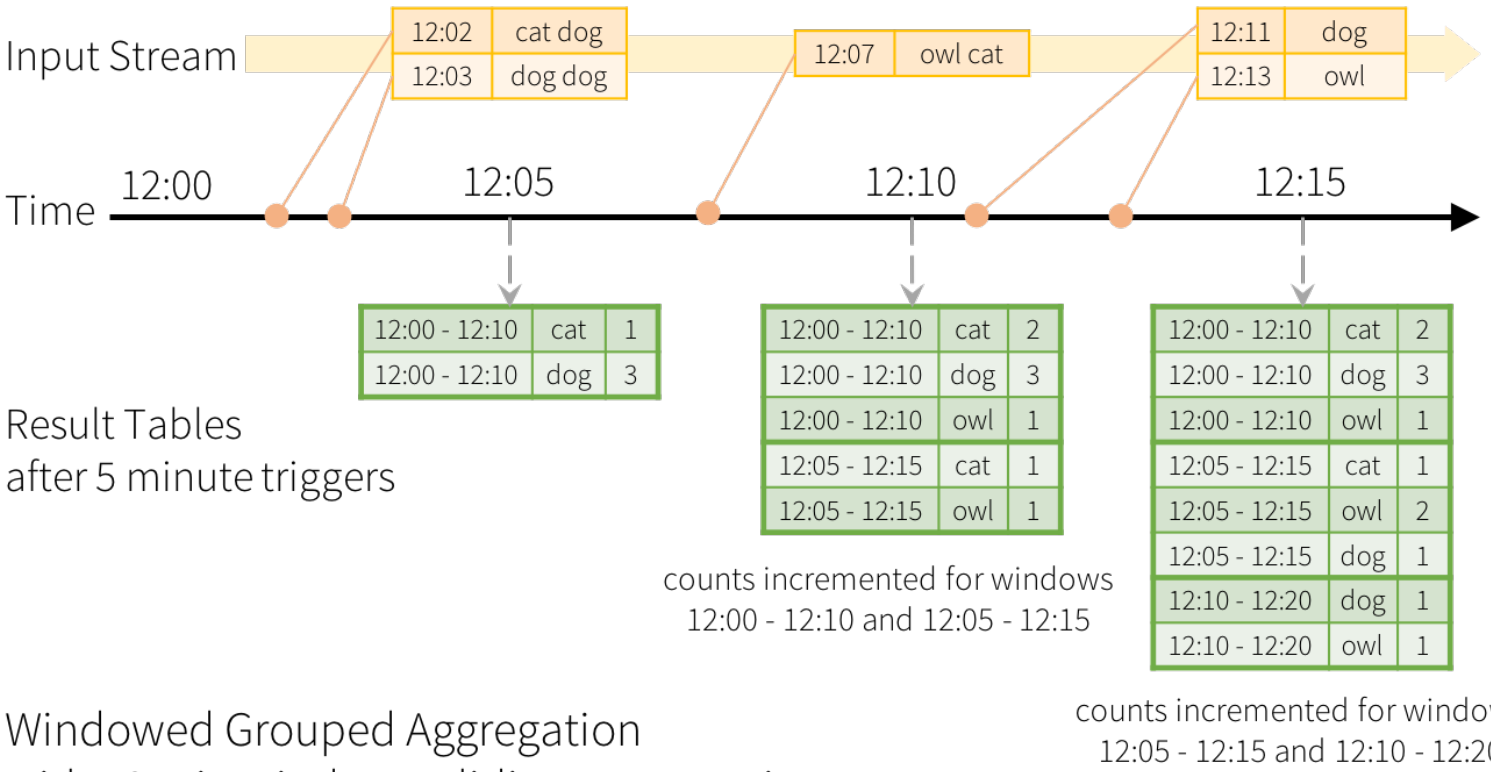
- Spark Streaming enables scalable, high-throughput, fault-tolerant stream processing of live data streams.
- Data can be ingested from many sources, e.g., Kafka, Kinesis, or TCP sockets.
- Streaming data can be processed using functions such as map, reduce, join, and window (which we'll see shortly).
- Processed data can be directed to filesystems, databases, and live dashboards.
- You can also apply Spark's machine learning and graph processing algorithms to data produced by Spark Streaming.

Spark Streaming Data Structures

- DStream (i.e., Discretized Stream) is the basic abstraction provided by Spark Streaming.
- It represents a continuous stream of data, either the input data stream received from a source, or the processed data stream generated by transforming the input stream.
- Internally, a DStream is represented by a continuous series of RDDs.
- Each RDD in a DStream contains data from a certain interval, as shown in the following figure.



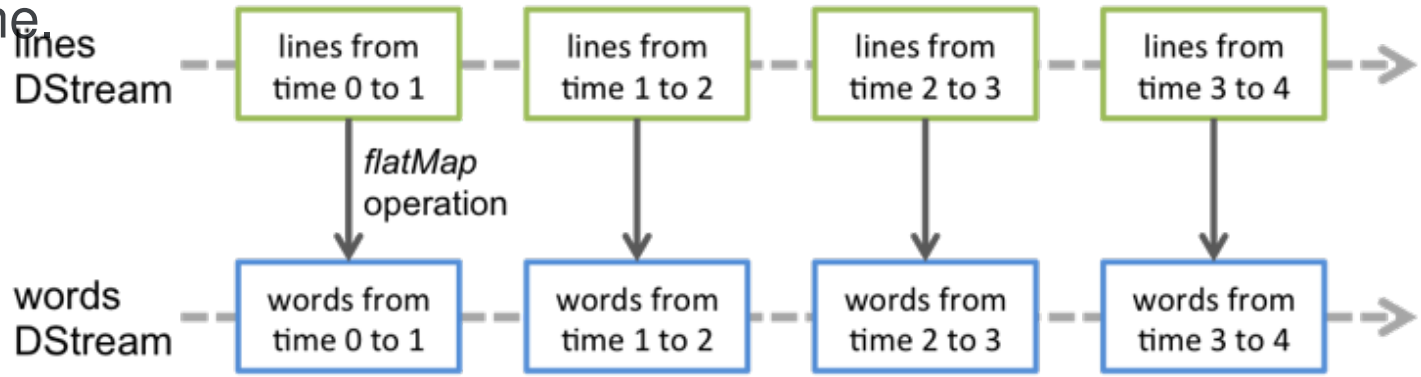
Spark Streaming Data Structures



Windowed Grouped Aggregation
with 10 min windows, sliding every 5 mins

Spark Streaming Algorithms

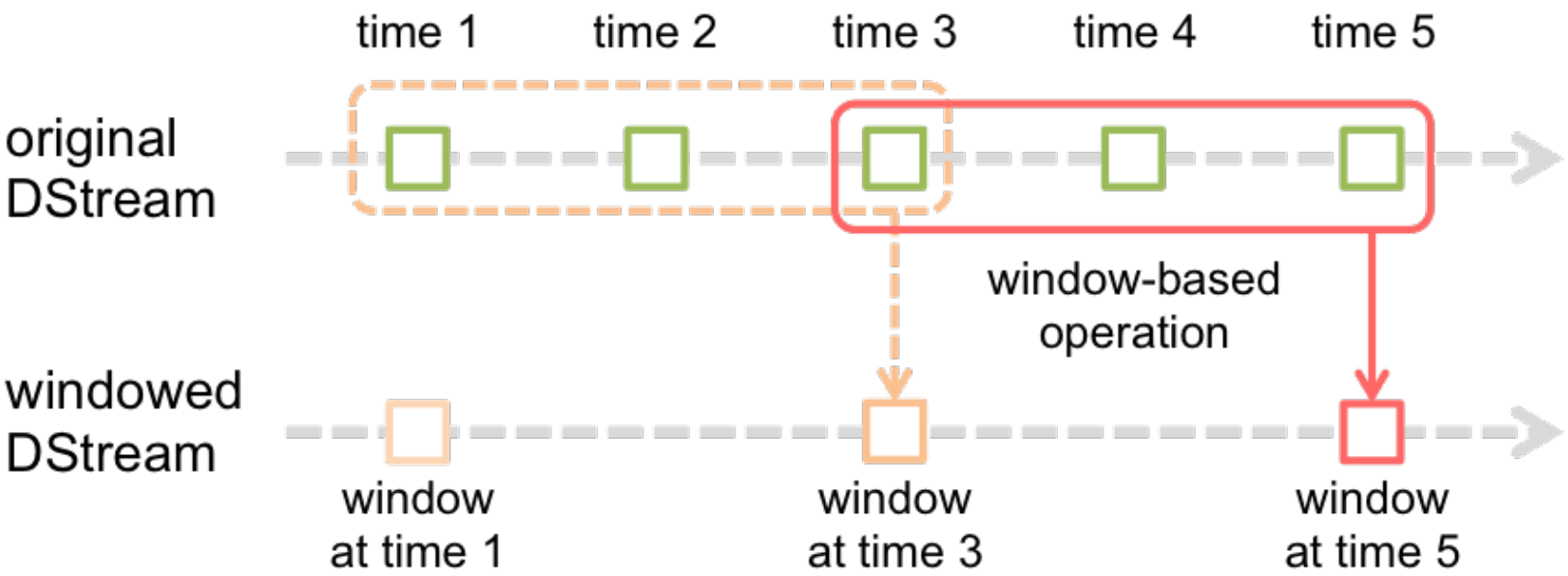
- Spark Streaming receives live input data streams and divides the data into batches.
- These batches are processed by the Spark engine to generate the final stream of results in batches.
- DStreams are created either from input data streams from sources such as Kafka, or by applying high-level operations on other DStreams.
- Any operation applied on a DStream translates to operations on the underlying RDDs.
  - This is shown in the figure below.
- These underlying RDD transformations are computed by the Spark engine.



Window Operations

- Spark Streaming also provides windowed computations, which allow you to apply transformations over a sliding window of data.
- Every time the window slides over a source DStream, the source RDDs that fall within the window are combined and operated upon to produce the RDDs of the windowed DStream.
- In this specific case, the operation is applied over the last 3 time units of data, and slides by 2 time units.
- Any window operation needs to specify two parameters:
  - **Window length** - The duration of the window (3 in the figure).
  - **Sliding interval** - The interval at which the window operation is performed (2 in the figure).
- These two parameters must be multiples of the batch interval of the source DStream (1 in the figure).

Window Operations



Streaming Example: Word Counting

Code Example

```
from pyspark import SparkContext
from pyspark.streaming import StreamingContext

sc = SparkContext(appName="PythonStreamingNetworkWordCount")
ssc = StreamingContext(sc, 1)

lines = ssc.socketTextStream("localhost", 9999)
counts = lines.flatMap(lambda line: line.split(" "))\
    .map(lambda word: (word, 1))\
    .reduceByKey(lambda a, b: a + b)
counts.pprint()

ssc.start()
ssc.awaitTermination()
```

Input and Output

```
> nc -lk 9999
hello world

-------------------------
Time: 2022-12-04 18:05:36
-------------------------
('hello', 1)
('world', 1)

-------------------------
Time: 2022-12-04 18:05:37
-------------------------
```