VE472 Lecture 1

Jing Liu

UM-SJTU Joint Institute

Summer

Course Information

- Course Description:

  Over the last decade, we have seen the emergence of "big data". This course is to introduce students to the basics of big data methods and tools. The course begins with various issues that we need to address when comes to big data. The course then introduce the current big data ecosystem/infrastructural technologies. Students are expected to obtain the most cutting-edge big data technologies and frameworks, such as Hadoop, Spark and Drill.

- Who should take this class?

  The prerequisite for this class is computer/programming knowledge at the level of Ve482 (or above), and statistics knowledge at the level of Ve406 (or above). Both undergraduates and graduate ECE students are welcome to take the course.

Contact Information

- Instructor:

  Manuel Charlemagne
  Jing Liu

- Lectures:

  Tuesday (12:10pm – 1.50pm) Online
  Thursday (12:10pm – 1.50pm) Online

- Office Hours:

  Tuesday (09:00am – 11:00am) Online
  Thursday (09:00am – 11:00am) Online

- Email:

  charlem@sjtu.edu.cn and stephen.liu@sjtu.edu.cn

- Teaching Assistant/s:

  See Canvas for his/her contact information

- To improve communication between the students and the teaching team please observe the following guidelines:
  - Any student facing a special situation likely to impact his studies, such as serious illness or full time work, is expected to contact the instructor as early as possible in order to discuss it and see if any solution can be found.
  - When sending an email related to this course please include the tag [VE472] in the subject e.g. Subject: [VE472] special request
  - When contacting an instructor for a grade issue or any other major problem send a carbon copy (cc) to the other instructor. Not doing it might result in omissions, not up-to-date grades etc. If such problem occurs and there is no record of the issue the request will be **automatically rejected**.
  - Never attach a large file ($>$ 2 MB) to an email, use a Dropbox type of service instead and only include a link in the email.
  - Keep in touch with the teaching team, feedbacks and suggestions will be much appreciated.

Grading Policy

- Quiz:

  20% It will focus on various ideas covered in class.

- Lab:

  10% It will focus on various ideas in terms of tools.

- Assignment:

  15% It will focus on various ideas in terms of methods.

- Project:

  30% You need to form a group of 3 people for the project.

- Exam:

  25% Midterm 15% + Final 10%

- For this course, the grade will be curved to achieve a median grade of "B".

Honour Code

- Honesty and trust are important. Students are responsible for familiarising themselves with what is considered as a violation of honour code.
- Assignments/projects are to be solved by each student individually. You are encouraged to discuss problems with other students, but you are advised not to show your written work to others. Copying someone else's work is a very serious violation of the honour code.
- Students may read resources on the Internet, such as articles on Wikipedia, Wolfram MathWorld or any other forums, but you are not allowed to post the original assignment question online and ask for answers. It is regarded as a violation of the honour code.
- Since it is impossible to list all conceivable instance of honour code violations, the students has the responsibility to always act in a professional manner and to seek clarification from appropriate sources if their or another student's conduct is suspected to be in conflict with the intended spirit of the honour code.

Textbook

- Textbook:

  White. (2015) , Hadoop: the definitive guide.

- Some Additional Material:

  Sitto and Presser (2015), Field Guide to Hadoop.

  Holmes (2014) Hadoop in Practice.

  Hastie et al. (2008), The element of Statistical Learning

  Tan et al. (2014), Introduction to Data Mining

  Leskovec et al. (2014), Mining of Massive Datasets

- Other course related materials will be available on Canvas.

Teaching Schedule Part I (Jing)

Week	Topics	Others
1	Intro, overview and basics Least squares fit as projection	
2	Stability as data size increases Gauss-Markov and Shrinkage	
3	Dimension reduction Case study	A1 due
4	Time series Case study	A2 due
5	Classification Case study	A3 due
6	Matrix multiplication for big data Random projection (Optional)	A4 due

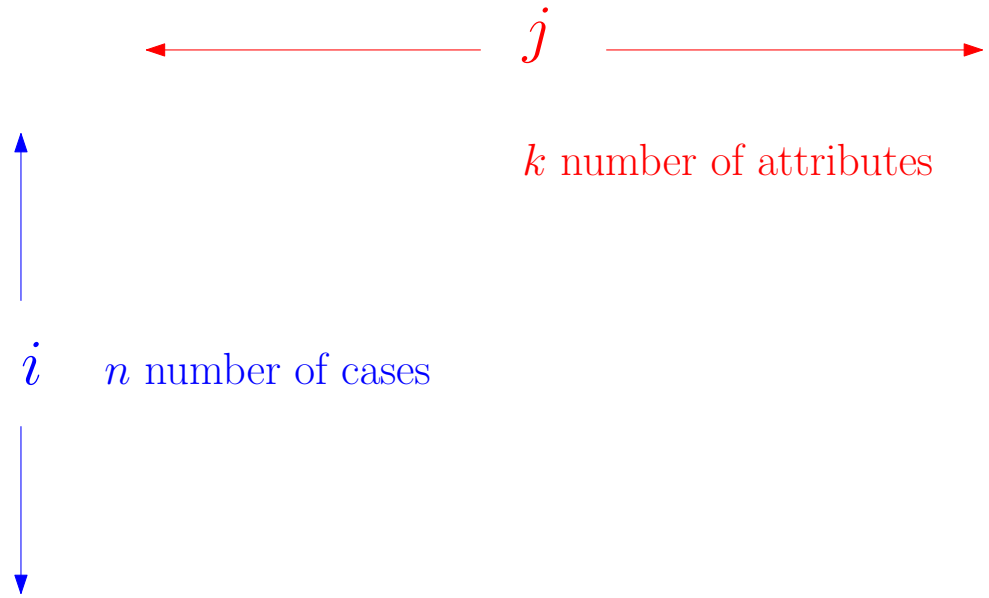
Teaching Schedule Part II (Manuel)

7	Basics on distributed systems	A5 due
	Toward big data	
8	**Midterm**	
	Hadoop overview	
9	Hadoop ecosystem	
	HDFS	
10	YARN	
	MapReduce	
11	Apache Drill	
	Apache Spark	
12	More sophisticated Hadoop uses	
	Alternatives to Hadoop	Proj due
13	**Final Exam**	

Q: What does the term "big data" actually refer to?

Big data is a term refers to datasets that are too large or complex for classical data analysis to deal with using a personal computer.

- Being "large" is often referred to the dataset having a large number of cases.



- Being "complex" is often referred to having a large number of attributes.
- In a tabular form, it refers to having too many rows or too many columns.

- Of course, not all data are in a tabular form, and can be treated as matrices,

- Source

  e.g. existing databases, surveys, experiments, internet, and etc.

- Type

  e.g. numbers, strings, sound, images, and etc.

- Algorithms

  e.g. array, tree, graphs, and etc.

- Nevertheless if we ignore efficiency from compute science point of view, it is often possible to transform data into matrices, and process/model data with matrix manipulations to provide insight into a large number of questions.
- We will focus on those type of questions, and study the relevant methods under the framework of matrices as they become bigger and bigger.

- Before going further, it is useful to categorise the scale we will consider

- Tiny

  A dataset is tiny if we can store and process it without a computer.

- Small

  A dataset is small if we can store and process it using naive methods on any computer, i.e. a computer that might be older than you.

- Medium

  A dataset is medium-sized if it can be fitted "comfortably" into the RAM of a modern computer and processed in a reasonable length of time.

- Big

  A dataset is big if it cannot fit into the RAM of a modern computer or it takes too long to conduct a typical data analysis procedure.