Foundations of Data Science

Lecture 0: Course introduction

MING GAO

SE & DASE @ ECNU
(for course related communications)
mgao@sei.ecnu.edu.cn

Sep. 18, 2016

Outline

1. Textbooks and references
2. Requirements and assessment
3. Office hour and contact information
4. Overview of this course
   - What is data science?
   - Course schedule
5. Take-aways

Required sources

Required sources

- John Hopcroft and Ravindran Kannan, Foundations of Data Science.
- Anand Rajaraman and Jeffrey D. Ullman, Mining of Massive Datasets.

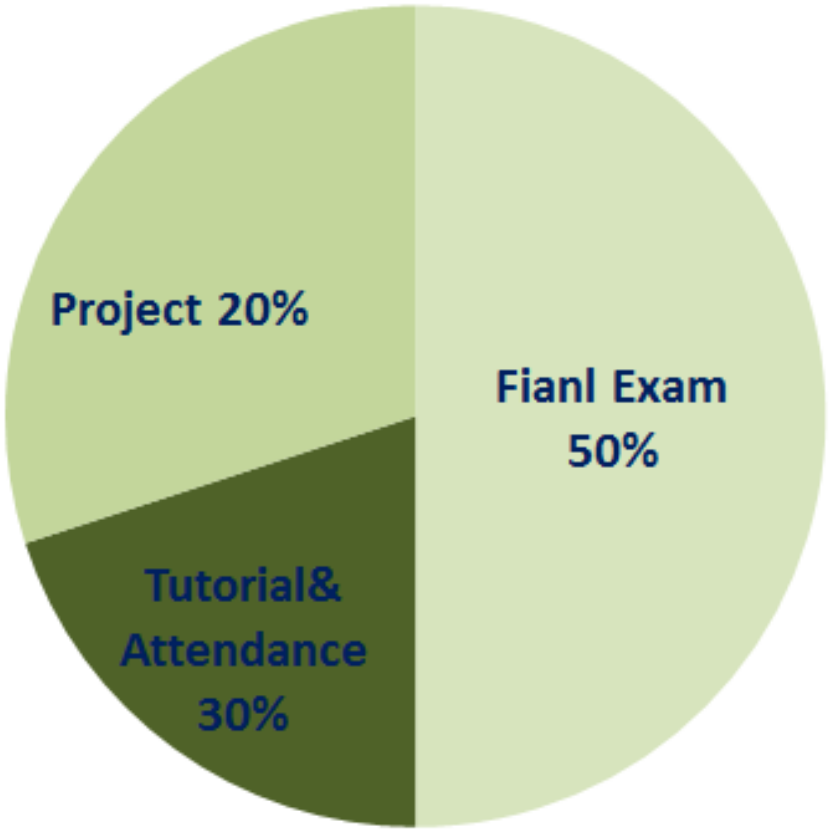
References

- Daphne Koller and Nir Friedman, Probabilistic Graphical Models: Principles and Techniques.
- Gilbert Strang, Linear Algebra and Its Applications(Fourth Edition).
- Fan Chung Graham, Spectral Graph Theory.

Requirements

1. Slides will be posted 1-2 days before lecture, but
2. Students are expected to
   - take notes during lecture (no lecture notes will be provided)
   - read the assigned readings before and after the lecture
   - think through the answers of tutorial (a set of questions) every week before the lecture
3. Implement a technique published in the top venues, such as KDD, ICDM, SIGMOD, VLDB, ACL, etc. (honestly and independently)

Assessment



Contact information

Lecturer: GAO Ming—- 高明

- Office: Rm. East 115, Math. Building
- Phone: 6223 2061
- Mobile: 189 1694 3299
- Email: mgao@sei.ecnu.edu.cn
- Course homepage: http://dase.ecnu.edu.cn/mgao/teaching/DataSci_2016_Fall/DS.html
- Research focus:
  - Social data ming
  - User profiling
  - Knowledge graph
  - Streaming data management and mining

What to be taught in this course?

Aim at helping students to build up data thinking

1. What is the course of data science?
2. Its components and available technologies
3. How to become a data scientist?
4. Some basic knowledge for a data scientist
5. Some advanced technologies that are related, or would be integrated into DS in the future

Data science and big data

- How to understand big data?
  - Volume: 100PB and 20PB data daily processing for Baidu and Google, respectively; Alibaba and Tecent have data more than 100PB.
  - Velocity: Large Hadron Collider generates PB data in seconds; many streaming such as clickstream, log, RFID, Twitter, etc. #Trans. is almost 100,000 per second in Tabao during "Double 11".
  - Variety: structured, semi-structured and non-structured, including text, logs, video, voice and image etc.
  - Value: interests, behaviors, trustworthiness, and preference, etc.
- Fragmentation of information:
  - Telecom
  - E-commerce
  - Social media
  - ...

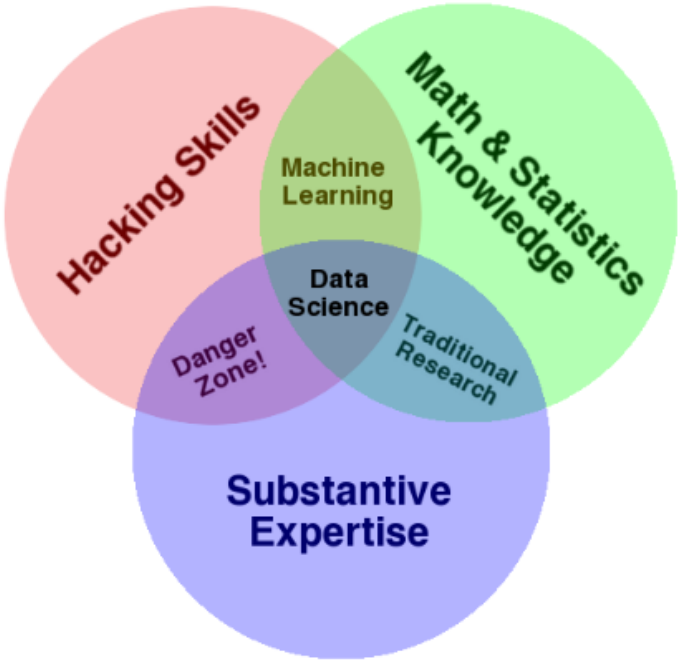
Birth of data science

- Reasons
  - Challenges of 4V
  - Hardware updating
  - Open sources, including Hadoop, Spark, Storm, and so on.
  - Applications, such as E-commerce, sharing economy, industry 4.0, smart city, and intelligent education, etc.

What is data science?

Definition

Data science is an interdisciplinary field, which is a continuation of some of the data analysis fields such as mathematics, statistics, machine learning, data mining, and parallel computing, similar to Knowledge Discovery in Databases (KDD).



Objective

Data science goals to:

- extract knowledge
- insight from data in various forms, either structured or unstructured
- help users to understand massive data

DS co-evolution

- Data science was mentioned by John W. Tukey in 1962 ("The Future of Data Analysis").
- Data science was defined by Peter Naur in 1974 ("Concise Survey of Computer Methods")
- Many data mining approaches were proposed in the 1980s of the 20th century.
- In 1996, international federation of classification societies issue set up a conference, namely Data Science, Classification and Related Methods.
- In June 2009, Nathan Yau published a paper talking about the rising of data science.
- Data scientist is the sexiest job in the 21st century (Hal Varian on Sep. 2012).

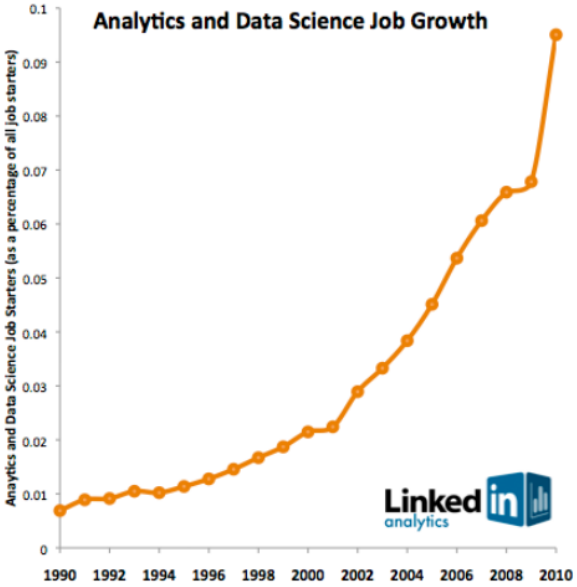
Types of data scientists

- Data developer: data acquisition, organization and management.
- Data researcher: statisticians, social scientist, computer scientist, etc.
- Data creatives: experts in machine learning, data mining, and programming, etc., contributor in open-source community,
- Data businessmen: project manager, Chief Data Officer (CDO)
- Mixed/Generic type: deep-understand in business, professional in technology, good at programming, etc.

Four paradigms of scientific research

- Experimental science
- Theoretical science
- Computational science
- Data science?
  - It was firstly proposed by Jim Gray (a database researcher) in 2009.
  - The Forth Paradigm: Data-Intensive Scientific Discovery was wrote by Tony Hey (vice president of Microsoft) et al. in 2009.
  - Thus, the capability for big data processing is important to scientific researchers.

The shortage of data scientists



Schedule

Background

DS overview

Probability and Statistics

- Random variable, distribution and expectation
- Estimation, hypothesis testing
- Markov chain, Bayesian network, and Markov random field
- Sampling and probabilistic inequality

Schedule

Algebra

- Vector, Matrix, and operation
- Eigenvalues and eigenvectors
- Laplacian and spectral analysis
- Matrix factorization

Graph

- Centrality and similarity
- Community detection
- Information propagation
- Link prediction and recommendation

Schedule

Streaming Data

- Count-min sketch, frequent item mining, and moment estimation
- Bloom filter and LSH
- Clustering

Take-aways

Course homepage

http://dase.ecnu.edu.cn/mgao/teaching/OS_2015_Fall/OS.html

Advices to learning DS

- Not a reading course.
- More than a programming course, though it is project-heavy
- No *standard answers*