Undergraduate Course –

Introduction to Data Science (Block course)

Prospective Course Outline

Institute for Interdiscplinary Information Sciences (IIS), Tsinghua University

C. Schommer, University of Luxembourg
ILIAS Laboratory, ilias.uni.lu

Summer Term 2015

Contents

1 Overview

1.1 Leitmotif

We believe in a cascade of *data*, *information*, *insights*, *knowledge* and back to *data*. We work on the different aspects to discover valuable insights in *data (masses)*, where *data* is the 'nutrient solution' for any kind of processing. It exists everywhere, but is depending on changes of the environment (for example, through valuable *insights* or *knowledge*).
The term *information* is a well-discussed term and is, therefore, understood differently. In this (data science) context, however, it refers to something about the data, for example a structure or a pattern. *Information* per se is a neutral term, it neither is worth nor worthless. As a step further, the term *insights* refers to *information that has a certain worth and importance*, for example a reliable model to predict anomalies or outliers, a topic engine to identify financial terms, or a conversational pattern within texts, which explains a characteristic feature/situation. *Insights* are different to information, since *information* implies structures/patterns that are potentially redundant, worthless, or suboptimal whereas insights is the result of an (human) interpretation (and, therefore, subjective as well, of course).
Discovered *insights*, on the other side, affect a behavioral change with respect to the data, quasi an inherent adaptation (and optimization). We call systems, which support such an adaptive cascade: intelligent. An inherent data (and information) complexity can be reduced to its essence and temporal changes encountered.
Finally, *knowledge* refers to discovered insights, which have proven a worth (over time).

1.2 The place of *Data Science*

It is still rather unclear in literature, academia, and business, what the definition of *Data Science* is. However, it turns out that – in the presence of masses of data – and understanding of the data and its structures as well as the discovery of patterns (if existing) seems to be the important part of it.

1.3 Organisation & Learning Outcomes

Modern computer machines and algorithmic intelligence have made it possible that large amounts of data can be collected, retrieved, and analyzed from a variety of sources. This offers many challenges but raises also questions, for example: how can data efficiently be managed, processed, and retrieved? Which algorithms should be used (and why) to extract knowledge out of it? Are the extracted patterns reliable, durable, but justifying a privacy? Do the patterns motivate a causal relationships? Beside a technical comprehension, the course is aimed at the understanding of the matter.

Figure 1: Data Science in the sense of targeting different disciplines.

In this course, we discuss *selected* aspects of the concept of data, related techniques and aspects, and future achievements in the area of **Data Science**. The course is organized with 24 lecture units (each unit counts for 45 minutes) and 8h exercises. It closes with a written examination.

The course intends to motivate students by fundamental (selected) aspects of *Data Science*. The successful participant will learn about the *concept of data* and to think about typical problems that are associated with it: how to manage data, how to search for information in data, how to discover patterns in data, how to visualise data, and what about concepts to make data secure.

The outcome of the course should not only be a knowledge transfer but much more a student's awareness regarding the justified need for *data science*, particularly, due to the masses of data. Each participant should get an understanding. I attach importance to the 'why' rather than to the 'how', encouraging students to look for alternatives, to be curious and critical, and to understand problems from different perspectives.

1.4 Preliminary Schedule

The course is organised as lecture (U1–U24; 24 units) plus exercises (E1–E8; 8 units):

: August 24, 2015 (Monday):

- U1, U2 (13h30–15h05)
- U3, **E1** (15h20–16h55)

: August 25, 2015 (Tuesday):

- U4, U5 (09h50–11h25)
- U6, **E2** (13h30–15h05)

: August 26, 2015 (Wednesday):

- U7, U8 (13h30–15h05)

: August 27, 2015 (Thursday):

- **E3**, U9 (09h50–11h25)
- U10, U11 (13h30–15h05)

: August 28, 2015 (Friday):

- U12, **E4** (13h30–15h05)
- U13, U14 (15h20–16h55)

: August 29, 2015 (Saturday):

- **E5**, U15 (09h50–11h25)

: August 31, 2015 (Monday):

- U16, U17 (13h30–15h05)
- **E6**, U18 (15h20-16h55)

: September 1, 2015 (Tuesday):

- U19, U20 (09h50–11h25)
- **E7**, U21 (13h30-15h05)

: September 2, 2015 (Wednesday):

- U22, U23 (13h30-15h05)

: September 3, 2015 (Thursday):

- **E8**, U24 (09h50–11h25)
- Examination (15h20–16h50)

: Day 12 (Fri):

The time frame is from Monday, August 24, 2015 – Friday, September 4, 2015

1.5 Exercises

Exercise sheets will be given to the participants one day before (except for exercise 1). However, there will be no corrections.

1.6 Examination

A written examination is made at the end of the course (Day 10).

2 Contents

2.1 Lectures

U1 Data Science: Course Overview

We begin with an overview of the subject and outline *Data Science* as a discipline, which explains *data* as its centric core. Goal: get an understanding of the field, in particular why the selected disciplines have its worth.

- The centric role of Data.
- Big Data and Small Data.
- Brief Overview of what the course content is.

U2, U3, U4 Data Management

Standard management systems base on the relational model, which was originally introduced by Edgar Codd. We will take a look at it and discuss its principles and technical structure. The chapter closes with a discussion regarding the advantages and the disadvantages of this approach.

- Relational System and SQL.
- Alternative Data Stores: NoSQL.

U5, U6 Data Quality.

Data Mining denotes the attempt to explore the data and to discover unknown, non-trivial, but useful patterns. It stands in contrast to processes, which only verify a given hypothesis. In this matter, an important point is that reliable, correct, and stable findings afford some kind of a data quality. In this chapter, therefore, we concentrate on a diverse number of data preparation techniques.

- Descriptive statistics.
- Data Cleaning.
- Data Normalisation.
- Data Reduction.
- Data Transformation.
- Data Discretization and Concept Hierarchies.

U7 Data Security.

The management of data must guarantee a consistency and correctness, otherwise we risk that data becomes corrupted and invalid. SQL offers a diverse number of techniques to handle such problems, for example the idea of having access rights or a diverse number of integrity mechanisms. Beside, we will discuss specific scenarios, which may occur if processes run in parallel.

- Triggers and Assertions.
- Data Types.
- Access Control.

U8 Data Privacy The existence of data requires solid ways of protection, particulary in the presence of private data. In this chapter, we study a diverse number of elementary techniques, which may support a deeper insight into privacy.

- Overview
- Data Publishing: k-anonymity
- Data Publishing: l-diversity
- Data Publishing: t-closeness
- Other methods (cryptographic, utility-based, query auditing)

U9, U10, U11, U12 Data (Information) Retrieval.

The (verificative) search for content is a central concern. How do we search, how should we rank the results, and how do we evaluate the results sets? Even more, the user's role must be considered as well: how to repair given search requests and how imply a feedback, in particular regarding a satisfaction with regard to the result set?

- Linguistic aspects: terms, documents, collections.
- Similarities: n-grams, Jaccard, Euclidean (Minkowski), angle and cosine, levenshtein, soundex.
- Boolean Retrieval; term lists; positional indexes in posting lists.
- Dictionaries.
- Vector Space Model:
- Ranking: tf-idf; cosine, normalisation.
- Evaluation: Precision and Recall.
- Feedback: implicit, explicit.

U13, U14 Big Data (Prof. de Melo)

- MapReduce, Distributed Processing.
- Big Data-algorithms (e.g., Bloom Filters, Locality-Sensitive-Hashing, Count-Min-Sketch)

U15, U16, U17, U18, U19, U20 Data Mining

This chapter continues the previous chapter, but raises the questions of how we can find interesting patterns? Given masses of data, which explorative intelligence can be used to pursue and attain this goal?

- Problem Statement and Definition.
- Verification vs. Exploration.
- Process of Data Mining; Application scenarios.
- Association Discovery: Apriori.
- Sequence Mining: FP Growth.
- Classification: Decision Trees.
- Clustering: K-means, kNN.
- Applications.

U21, U22, U23 Data Visualization

An understanding of data can be supported if data is shown expressively, effectively, and correctly. for example by suitable visualization procedures. However,

- What makes a good visualization: Effectiveness, Appropriateness, Expressiveness. No Lie Factor.
- The visualization process: Filtering (data preparation), Mapping (selection of the visoualisation method), and Rendering (production).
- Influences and Cognitive Aspects for Visualisation:
  - Red-Green Blindness
  - Lengths; Colors; Shapes; Textures; Sharpness, et cetera.
- Fundamental techniques, particularly for multi-dimensional data, FOR EXAMPLE:
  - *Geometric*: **Scatterplots**, Landscapes, Projection Techniques, Prosection Views, **Hyperslice**.
  - *Graph-based*: **simple graphs**, special graphs (direct acyclic graph, symmetric graph), systems (z.B. Tom Sawyer, Hy+, SeeNet, Narcissus)
  - *Icon-based*: **Chernoff Faces**, **Stick Figures**, Shape-Coding, Color-Icons, TileBars.
  - *Pixel-based*: Recursive Pattern Techniques, **Circle Segments** Technique, Spiral & Axes Techniques.
  - *Hierarchical*: Dimensional Stacking, **Worlds-within-worlds**, **Treemaps**, **Cone Trees**, **InfoCube**.
  - *Parallel Coordinates*
  - *Hybrid* (combination of some of the techniques given above

U24 Course Summary

- Summary of the course.
- About the examination.

3 Contact

Dr. Christoph Schommer, Associate Professor. **Working address:** Dept. of Computer Science and Communication (CSC), ILIAS Lab, University of Luxembourg, Luxembourg. **Phone**: +352-466644-5228. **Web**: ilias.uni.lu **Email**: christoph.schommer@uni.lu.

References

[a] R. Elmasri, S. Navathe: *Fundamentals of Database Systems.* Pearson. Book chapters are available online.

[b] A. Silberschatz, H. Korth, S. Sudarshan: *Database Systems Concepts.* McGraw Hill. Book chapters are available online.

[b] L. Sweeney: *k-anonymity - a model for protecting pricacy.* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[d] A. Machanavajjhala, J. Gehrke, D. Kifer: *l-diversity - Privacy Beyond k-Anonymity.*

[e] N. Li, T. Li, S. Venkatasubramanian: *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.*

[f] A. Rajaraman, J. Lescovec, J. Ullman: *Mining of Massive Data Sets. Stanford University.*

[g] I. Witten, E. Frank, M. Hall: *Data Mining - Practical Machine Learning Tools.* Morgan Kaufman.

[h] J. Han, M. Kamber: *Data Mining - Concepts and Techniques.* Morgan Kaufman.

[i] C. Ware: *Information Visualization - Perception for Design. Morgan Kaufmann. Book chapters are available online.*

[j] C. Manning, P. Raghavan, H. Schütze: *Introduction to Information Retrieval. Cambridge. Book chapters are available online.*