# Big Data Seminars Course: State-of-the-Art

## 1. Motivation

Data Science / Big Data are fairly broad fields and one of the main skills identified to be a good practitioner is autonomy and capacity for researching on your own. Thus, one of the objectives in this master is learning to do research on your own in a rigorous manner. This will be helpful when you will need to complement the fundamental concepts learned in the master with specific knowledge you might need during your professional career.

For this project, we will propose some topics to learn about, all of them extremely relevant for Data Science / Big Data. You may choose topics in other fields but in case of selecting a topic not included in this list, please, contact the master coordinator first to get green light. In general, any topic related with the semester is acceptable.

## 2. Project Objectives

In this project you must choose one of the topics listed in the event or propose a new one related to the topics covered during this semester. Importantly, the project must be conducted in groups of two people.

Thus, start by selecting a topic. In Learn-SQL you will find a list of topics and a brief summary of what is it about and some initial pointers (not to be considered exhaustive). Once you made a choice enrol, together with your partner, to the group with that name. If you want to propose a topic, contact the BDS responsible lecturer and if accepted, a new group with the name of your topic will appear. Join it then. As thumb rule, we will not accept two different groups to work on the same topic, so the first group enrolling on a topic will get it.

The objective of this project is twofold:

- Learn first-hand about a hot topic in Data Science,
- Build up on your autonomy and capacity for doing rigorous research on your own, as well as your capacity for summarizing your findings and mapping them to a common framework that will facilitate your understanding of a given problem.

## 3. What is a state-of-the-art (or literature survey)?

A state of the art is a document that covers a certain field (e.g., data discovery) and presents the current status of that field. Thus, for conducting a state-of-the-art you must:

- Find relevant papers. There are two main manners to look for papers (complementary to each other):
    - Keyword search: Starting from the seminal papers we propose, you must be able to identify relevant keywords related to that topic. Use them to explore specialized repositories. The main open repositories to search for scientific literature are Google Scholar and DBLP. In both, you may search by keywords but also by authors. This is important because at some point you will realize there are some big names related to a topic and you may save time directly searching for such representative gurus and accessing their publications.

You may also perform keyword search directly in the main publisher websites (Springer, Elsevier, IEEE, ACM). However, Google Scholar and DBLP index most of them.

- Snowballing: once you have found some papers, snowballing consists in searching, transitively, by means of the bibliography section of a paper or a book. Whenever you reach an interesting paper, it is likely you will find more relevant papers in its bibliography and so on so forth from those found. Snowballing refers to the transitive nature of this kind of search. Last but not least, snowballing traditionally looks into the past, but Google Scholar has a nice feature to snowball forward. Whenever you access a paper in Google Scholar, realize there is an option to query papers citing it. Thus, Google Scholar allows you to snowball forward too and therefore, find the last trends.

*Hint#1*: Focus on high quality venues (i.e., conferences and journals) that guarantee you do not waste time with publications of uncertain quality. Measuring the quality of a conference or a journal is, however, controversial. Nevertheless, there are some rankings that may facilitate identifying top venues. For conferences, you may check the GII-GRIN-SCIE Ranking. For journals, the JCR index.

*Hint#2*: Google Scholar, DBLP and the GII-GRIN-SCIE Ranking can be accessed for free. All the other portals require you access to them through eBIB (the UPC system to access scientific literature). To access the JCR index is however a bit trickier. Select *Federation of Spain by FECYT* as institution. Next, at the FECYT website, select *Universitat Politècnica de Catalunya*. Finally, use your UPC login / password to access JCR.

*Hint #3*: The number of citations of a paper is a good measure of its quality. However, it does not work for recent papers. For these ones, the only quality criteria you may follow is the authors and venues where they are published.

- Organize and read the papers found. You may have found quite a few papers. Therefore, it is important to organize them. You may use Excel, Notion or whatever other tool. This is just an advice, you may organize them as you feel better.

*Hint #4*: You will need to go over the papers several times as you get into the topic and realize about subtle details as long as you dig in. That is why it is a good idea to attach metadata to your papers (e.g., main idea, title, authors, keywords, etc.) when organizing them.

*Hint #5*: Do not read all the whole paper first. Researchers usually follow a two (some talk about three) read passes. In the first pass, read the title, abstract, conclusions and skim the paper. This will help you to understand the basics about it. Try to get the big picture and then narrow down the list of interesting papers for you. For those, go for a second pass reading the paper but avoiding proofs or formal methods. Leave this for a third potential pass just in case you want to get a very deep understanding of that paper. Check the *how_to_read_a_paper* paper in Learn-SQL for more details. However, be aware that it is thought for thorough searches. At the master level, it is probably fine to stay with a two-pass read.

- Choose a research question. After doing a first pass to your papers, you will realize that any of the topics we suggest, is complex and broad enough. Therefore, it is fine and advisable to focus on a specific point of interest within that topic. For example, if you

have chosen data discovery, you may want to focus on scalable data discovery and then, focus on specific algorithms and analyse their scalability. Thus, your research question is narrowed and you should focus on your second pass only on those papers fitting it. Selecting a proper research question is crucial for a smooth project.

*Hint #6*: Even if all the steps presented before look kind of sequential, this is more a kind of iterative process that at some point starts converging when you do not find new relevant papers or representative researchers of the field and you set a proper research question. Thus, you might feel overwhelmed at the beginning. That is normal and that is why it is important to organize well so that you can extract conclusions and start narrowing your search and set the research question.

- Sketch the main conclusions drawn from your work. After reading the relevant papers, you need to communicate your conclusions. For this, it is important to identify and extract common data / findings. It is important to realise that different papers will have different notations and formalisations, so the most difficult part is to harmonise notation / discussions based on a common backbone. Identifying these commonalities is pure synthesis, which is known to be harder than analysis.

*Hint #7*: To draw strong conclusions, try to avoid following the papers notation. Each of them have their own and explain what is beneficial for them. Many times, the most relevant aspect is what they do not discuss or say at all. For example, choose the notation / formalism of a paper that can be generalized to include the others. Then, think of dimensions or aspects from which to analyse your relevant works and be able to compare them (e.g., with a summary table). Just think of the definition of a graph. You can define it in a mathematical way or based on data structures. At some point you will need to choose your reference formalism and extract conclusions mapping to your reference backbone (e.g., graphs as adjacency matrices).

- Sum up and draw conclusions: Your document will be successful if someone not familiar with the topic, but majoring in CS, is able to follow and understand your conclusions. The more specific, the better.

*Hint #8*: Depending on your research question, you may decide to run some tools and play with them. That is perfectly fine. This project has some room to adapt it to your taste. You may go for a purely theoretical discussion of the papers, or you may want to check tools and compare them empirically.

*Hint #9*: Back up your main claims with either content from the papers (by citing them properly; see the attached resource at Learn-SQL for this matter) or with experiments from your empirical tests (e.g., in Github). Nevertheless, the experiments conclusions must be drawn in the document (e.g., with charts).

## 4. Deliverables

The outcome of this project is a document of, at most, 10 pages of content (not counting bibliography and cover page).

Also, if you decided to run and compare tools empirically, you are suggested to create a Github project and share its link. Just be sure to properly explain how to understand and follow the code there.

## 5. Evaluation

The project will be evaluated according to the following criteria:

**Conciseness**

The document fits in 10 pages and successfully introduces the reader into the topic at the right level of detail.

**Understandability**

There is a framework of reference that facilitates the read and you provide enough details as to assess your solution.

**Soundness**

There are no contradictions about the conclusions drawn and the inherent advantages / disadvantages identified of the underlying theory / tools analysed.

**Maturity**

The document is able to provide a detailed discussion of the topic and does not read shallow.