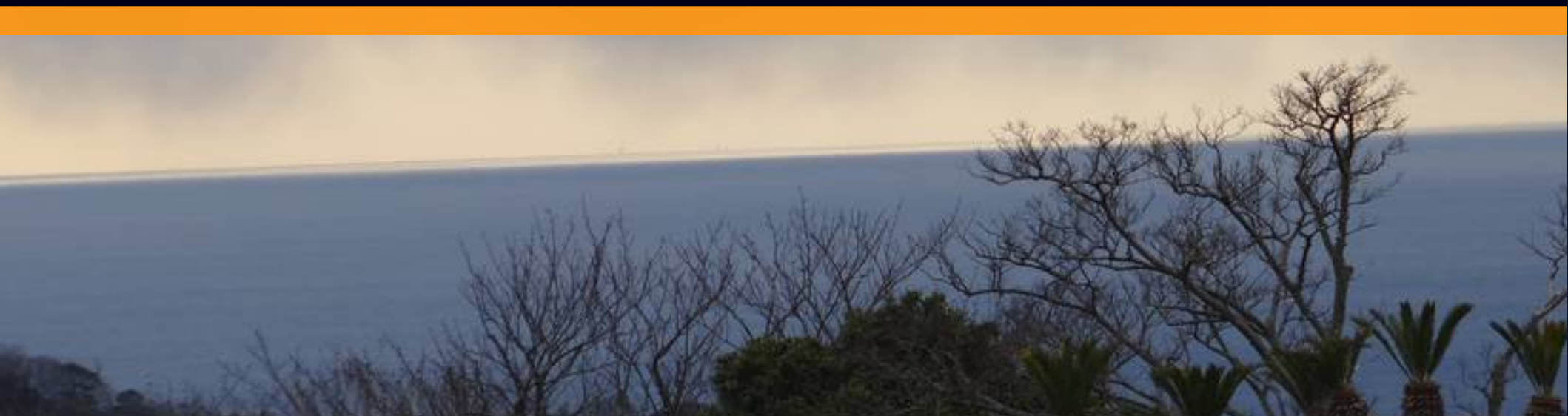


Big Data Analytics and Text Mining: Course Project

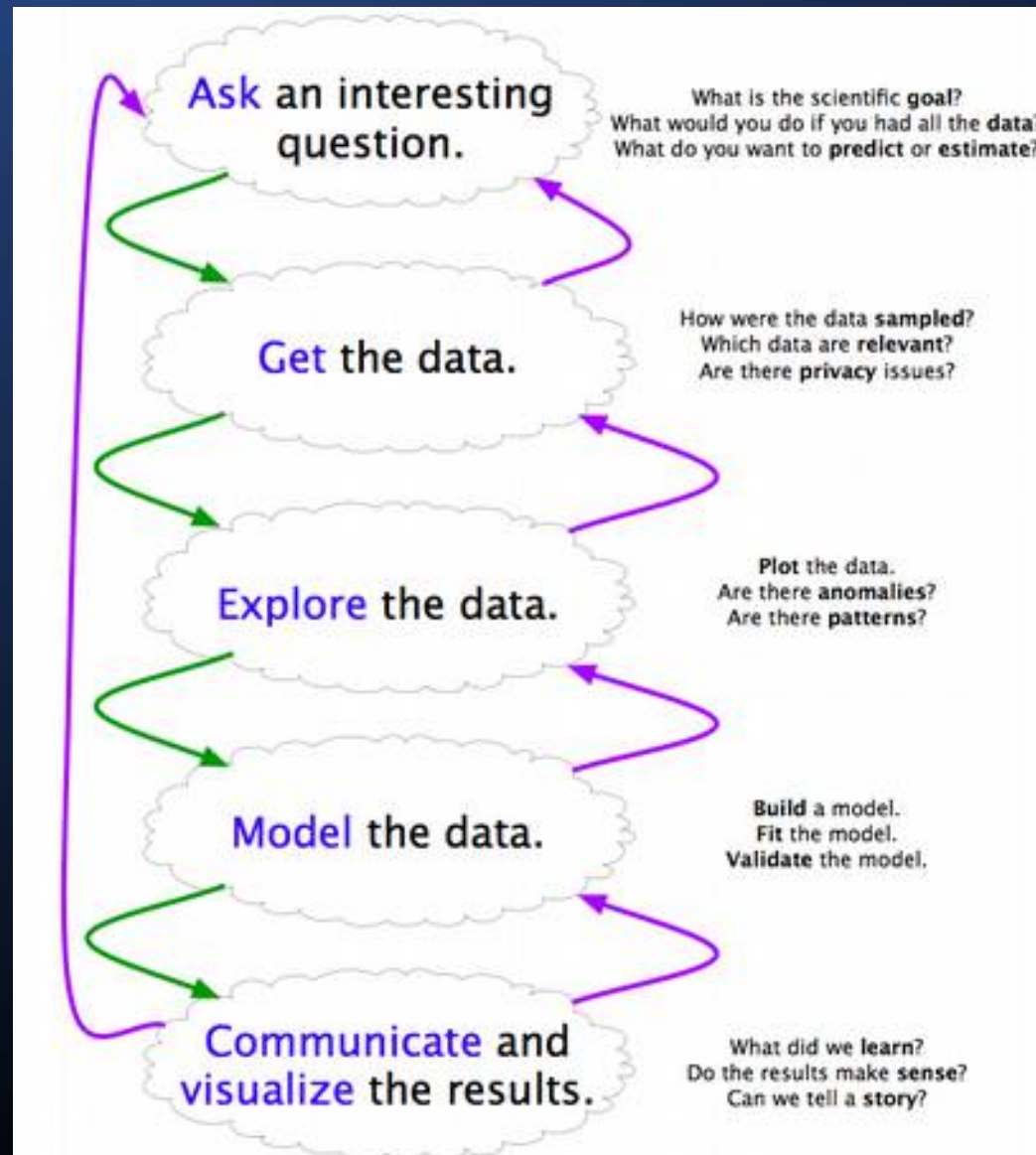
Gerard de Melo

<http://gerard.demelo.org>

Rutgers University



Course Project



Course Project

Goal

Work on a Big Data Analytics/Text Mining project of your choice.

Main requirements:

- 1. Analysis of data yielding interesting insights.**
- 2. Use of machine learning (2nd part)**

Optional: Data acquisition, Visualization, Interactive Demo, Tool that can be applied to new data (e.g.

We will provide some pointers for inspiration.

Course Project

Step 1: Short Project Proposal (by Feb. 20)

Just a short one paragraph description of what you are planning to do and hoping to achieve.

This can still be changed later, with approval from us.

Course Project

Milestone: Intermediate Report (by March 26)

1. Describe project goals and why it is interesting
2. Describe data collection/source of data, data format, data preprocessing.
3. Describe contents of data in detail
(use Spark to analyse it, preferably visualize it)
4. Describe possible applications of this data, Including your ideas for the next phase.

Course Project

Final Report (by May 1)

1. Improve on intermediate report.

The final report supersedes the intermediate report, so all crucial results from the intermediate one should be repeated.

Note: You can also analyse multiple related datasets.

2. Conduct machine learning experiments on your data. Ideal goal: Practical application.

3. Explain and evaluate your results, numerically or via visualizations. Should show how well your method works, or what insights have been gained.

Course Project

Final Report (by May 1)

4. Describe related work

Cite related research papers.

5. Conclusion section

What insights did you gain? What worked, what didn't work?

What else would you do if you had more time (or could start over)?

6. Acknowledgments section

Mention libraries used, third-party material used.

Course Project

Short Project Presentations (April 24/26)

Very short (less than 5 minute) presentation of your work

May use slides or interactively demonstrate your system.

Course Project

Teams

Team Size: 1 or 2

**(exceptions only with prior approval,
for particularly large/challenging projects)**

**Teams normally cannot be changed after
submitting the proposal.**

Grade: Equal for all team members

**However, we reserve the right to deviate from this
if the contributions were particularly unequal.**

Course Project

Rules

You may use any external libraries, as long as you explicitly mention this in an “Acknowledgments” section in your report.

Any third-party material used, even if modified or translated from a different programming language, must be mentioned in the “Acknowledgments” section in your report. Clearly indicate the extent of your own contribution.

All deadlines refer to 11:59pm Eastern time.

Late submissions at discretion of instructor, but with grade penalty.

Course Project

Report Format

Option 1) PDF + Source Code ZIP

Written like an academic technical report, typically at least 4 pages.

Recommendation: ACM or ACL 2017 LaTeX stylesheets.

Option 2) Spark Notebook:

Notebook with detailed descriptions integrated

Important: Provide both PDF and SNB files!

Additional code/data can be attached as well.

Course Project

Report Contents

Problem Description

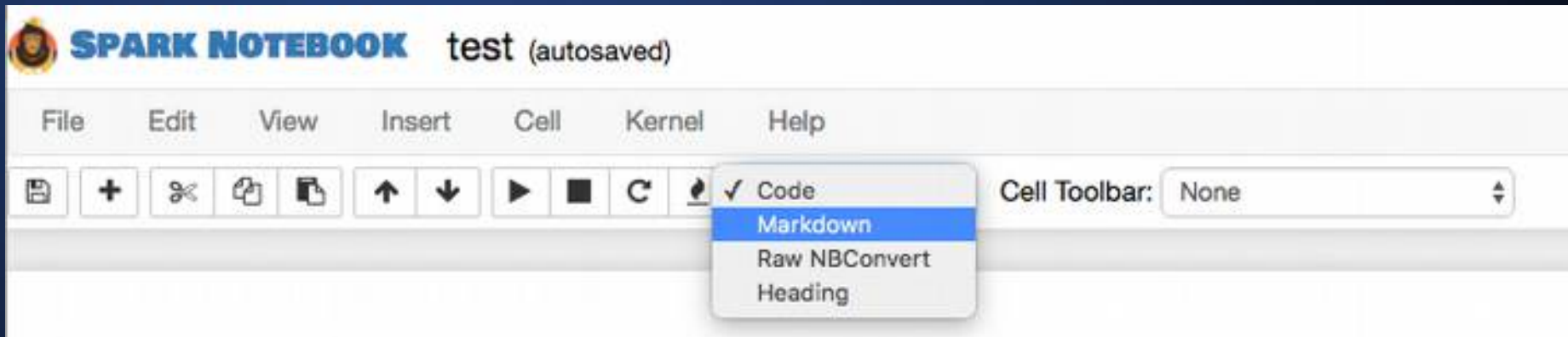
What are you working on? Why is it interesting?

Data

What data are you going to use?

Methods/Algorithms

Course Project



First-Level

Second-Level

Regular text

Your Spark Notebook file must be a report, so it must include text, not just code.

To do this, change a cell from “code” to “markdown”. Use “#” for headings.