CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Big Data Processing

L01: Module Introduction

Dr. Ignacio Castineiras
Department of Computer Science

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Presentation

- Ignacio Castiñeiras.
  - Lecturer at the Department of Computer Science.
  - Email: Ignacio.Castineiras@cit.ie
  - Office Room: C131
  - Telephone: +353 21 433 5857

- Qualification:
  - PhD. in Computer Science: 2014.
  - MEd. in Computer Science: 2011.
  - MSc. in Computer Science: 2009.
  - BSc. in Computer Science: 2007.

Dr. Ignacio Castineiras Department of Computer Science

Research Experience

[2018 -] Cork Institute of Technology
Lecturer at Dept. Computer Science
- Research Group Ríomh

[2015 - 2018] Cork Institute of Technology
Assistant Lecturer at Dept. Computer Science
- Research Group Ríomh

[2014 - 2015] University College Cork
Postdoc at Insight Centre for Data Analytics
- EU FP7 Project GENiC

[2007 - 2014] Complutense University of Madrid
PhD. & MSc. at Declarative Programming Group
- Spanish National Projects FAST & MERIT

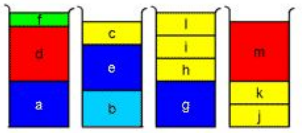
Background: Optimisation and decision analytics.
Application of **Constraint Programming** to real-life
Constraint Satisfaction and Optimisation Problems.

Research Background

- Constraint satisfaction and optimisation problems:
  - Examples: Manufacturing & service industries:
  - Feasible/optimal allocation/scheduling of company resources.
  - Challenge: Combinatorial nature.
- Constraint Programming:
  - Subfield of Artificial Intelligence.
  - High-level declarative problem formulation.
  - Problem solving: Inference process + search on top of it.

PhD. Research Experience

- Tackle real-life problems with Constraint Programming
  - Employee Timetabling Problem.
  - Bin Packing Problem.



- Comparison among multiple paradigms and solvers.
  - Algebraic - Object Oriented - (Functional) Logic Programming
  - C++, Python, SICStus Prolog, Haskell, TOY, etc.
- Implementation of constraint solvers:
  - Adapt object-oriented solver library to a logic programming environment.
  - Extend solver with high-level user defined search strategy specification.

Postdoc Research Experience

- GENiC: Globally Optimised Energy Efficient Data Centres.
  - Green computing.
  - Sustainable DCs.
  - Renewable energy sources.

CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

TU/e Technische Universiteit Eindhoven University of Technology

IBM

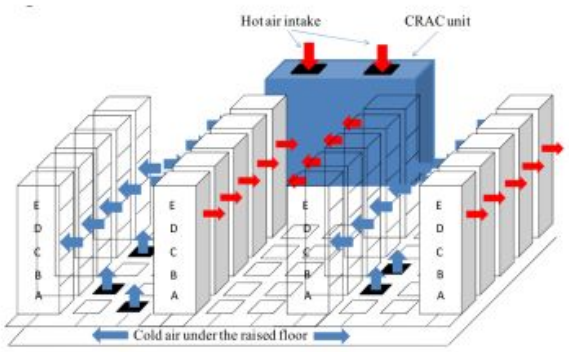
United Technologies Research Center

Atos

acciona
Infraestructuras

Postdoc Research Experience

- Develop scalable decision support tools:
  <u>Optimise workload allocation of single and distributed DCs</u>
  - Single DC: Reduce power consumption.
  - Geographically Dcs: Reduce overall energy consumption.



Lecturer Research Experience

Ríomh: Intelligent Secure Systems Group.

Research Areas:

- Future Networks & Internet of Things
- Virtualisation Technologies
  - Cloud Computing
  - Network and Information Security
- Data Analytics
  - Machine Learning
  - Optimisation Techniques

Contact us: Donna.Oshea@cit.ie (Head)

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Learning Outcomes

Module Descriptor:
https://courses.cit.ie/index.cfm/page/module/moduleId/13442

- LO1: Appraise how the velocity, volume and variety of data will impact how data is stored, managed and analysed.
- LO2: Survey the different tools that constitute a big data framework.
- LO3: Process large-scale temporal, geospatial, text and graph datasets using descriptive and analytical tools.
- LO4: Design and develop a machine learning algorithm for performing large scale distributed computation.

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Syllabus Week Plan

Week 1: September 9th – September 15th.

Lectures

- L01. Module Introduction.
- L02. Big Data.

Lab

- Lab01. Sequential-solving: Programming Exercise.

Big Data Mindset.

- *Introductory example of a Big-Data driven society.*

Syllabus Week Plan

Week 2: September 16th – September 22nd.

Lectures

- L03-04. Distributed Programming.

Lab

- Lab02. Divide-Map-Reduce Lab Demonstration.

Big Data Mindset.

- *The thinking or mental shift big data requires: Sampling => All data.*

Syllabus Week Plan

Week 3: September 23rd – September 29th.

Lectures

- L05. Functional Programming.
- L06. Gentle Introduction to Apache Spark.

Lab

- Lab03. Divide-Map-Reduce Solving: Programming Exercise.

Big Data Mindset.

- *The thinking or mental shift big data requires:*
  Clean experiments => Messy experiments.

Syllabus Week Plan

Week 4: September 30th – October 6th.

Lectures

- L07-08. Spark Model of Parallel Computing: RDDs.

Lab

- Lab04. Functional Programming Exercises.

Big Data Mindset.

- *The thinking or mental shift big data requires: Causation (Why?) => Correlations (What?)*

Syllabus Week Plan

Week 5: October 7th – October 13th.

Lectures

- L09-10. Anatomy of the Execution of a Spark Program.

Lab

- Lab05. Inside Spark Lab Demonstration.

Big Data Mindset.

- *Datification (or the art of extracting data from the most surprising places).*

Syllabus Week Plan

Week 6: October 14th – October 20th.

Lectures

- L11-12. Spark Core API.

Lab

- Lab06. Assignment 1 - Part 1

Big Data Mindset.

- *Data Reuse: Data's multiple lives.*

Syllabus Week Plan

Week 7: October 21st – October 27th.

Lectures

- L13. Spark Core API.
- L14. Spark SQL: RDDs vs DataFrames/Datasets. Catalyst and Tungsten.

Lab

- Lab07. Assignment 1 - Part 2

Big Data Mindset.

- *Data regulations: Data ownership and its accountability.*

Syllabus Week Plan

Week 8: November 4th – November 10th.

Lectures

- L15-16. Spark SQL API.

Lab

- Lab08. Assignment 1 - Part 3

Big Data Mindset.

- *The dark side of big data: I know who you are. I guess what would you do.*

Syllabus Week Plan

Week 9: November 11th – November 17th.

Lectures

- L17. Spark SQL API.
- L18. Spark Streaming: Concepts and Infrastructure.

Lab

- Lab09. Assignment 1 - Part 4

Big Data Mindset.

- *Big data industry revolution: Education as a use-case: Get to know students better.*

Syllabus Week Plan

Week 10: November 18th – November 24th.

Lectures

- L19-20. Spark Streaming API.

Lab

- Lab10. Assignment 2 - Part 1

Big Data Mindset.

- *Big data industry revolution: Education as a use-case: Adaptative learning.*

Syllabus Week Plan

Week 11: November 25th – December 1st.

Lectures

- L21. Structured Streaming: Concepts and Infrastructure.
- L22. Structured Streaming API.

Lab

- Lab11. Assignment 2 - Part 2

Big Data Mindset.

- *Big data industry revolution: Education as a use-case: The dark side again.*

Syllabus Week Plan

Week 12: December 2nd – December 8th.

Lectures

- L23. Structured Streaming API.
- L24. Module Wrap-Up

Lab

- Lab12. Assignment 2 - Part 3

Big Data Mindset.

- *Big data: What do you think?*

Syllabus Week Plan

Week 13: December 9th – December 15th.

Lectures

Lab

- Lab13. Assignment 2 - Part 4

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Methodology

- 1h lecture (twice per week):
  - Concepts explanation.
  - Application via code examples.
  - Put together to extract conclusions.

- 2h lab session (once per week):
  - Reinforce the concepts seen in the lectures.
  - Assignments split by weeks.
  - Weekly parts: Solve a bunch of exercises and submit.
  - Assignment demo after submission.

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Evaluation

Module Descriptor:
https://courses.cit.ie/index.cfm/page/module/moduleId/13442

Assignment 1:

- Use Spark Core and Spark API to perform descriptive analytics of an real-world open source dataset.
- Compare and contrast the efficiency and expressiveness of both approaches.

Marks: 50
Deadline: Week 9 - November 17th - 11.59pm

Evaluation

Module Descriptor:
https://courses.cit.ie/index.cfm/page/module/moduleId/13442

Assignment 2:

- Use Spark Streaming and Spark Structured Streaming to perform offline and online analytics of an real-world open source dataset.
- Compare and contrast the efficiency and expressiveness of both approaches.

Marks: 50
Deadline: Week 13 - December 15th - 11.59pm

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Motivation

Why is important to study this module?

For any module I teach I usually take some minutes during lecture 1 to justify/motivate why to study the module.

In this case, the motivation has grown so much that it has become part of the indicative content: Big Data Mindset.

But, in a single point: Why to study big data?
Because it is transforming our society.

Outline

1. Presentation.
2. Learning Outcomes.
3. Syllabus Week Plan.
4. Methodology.
5. Evaluation.
6. Motivation.

Thank you for your attention!