Academic year 2019/2020

Big Data
Presentation of the course

Prof. Riccardo Torlone

Università Roma Tre

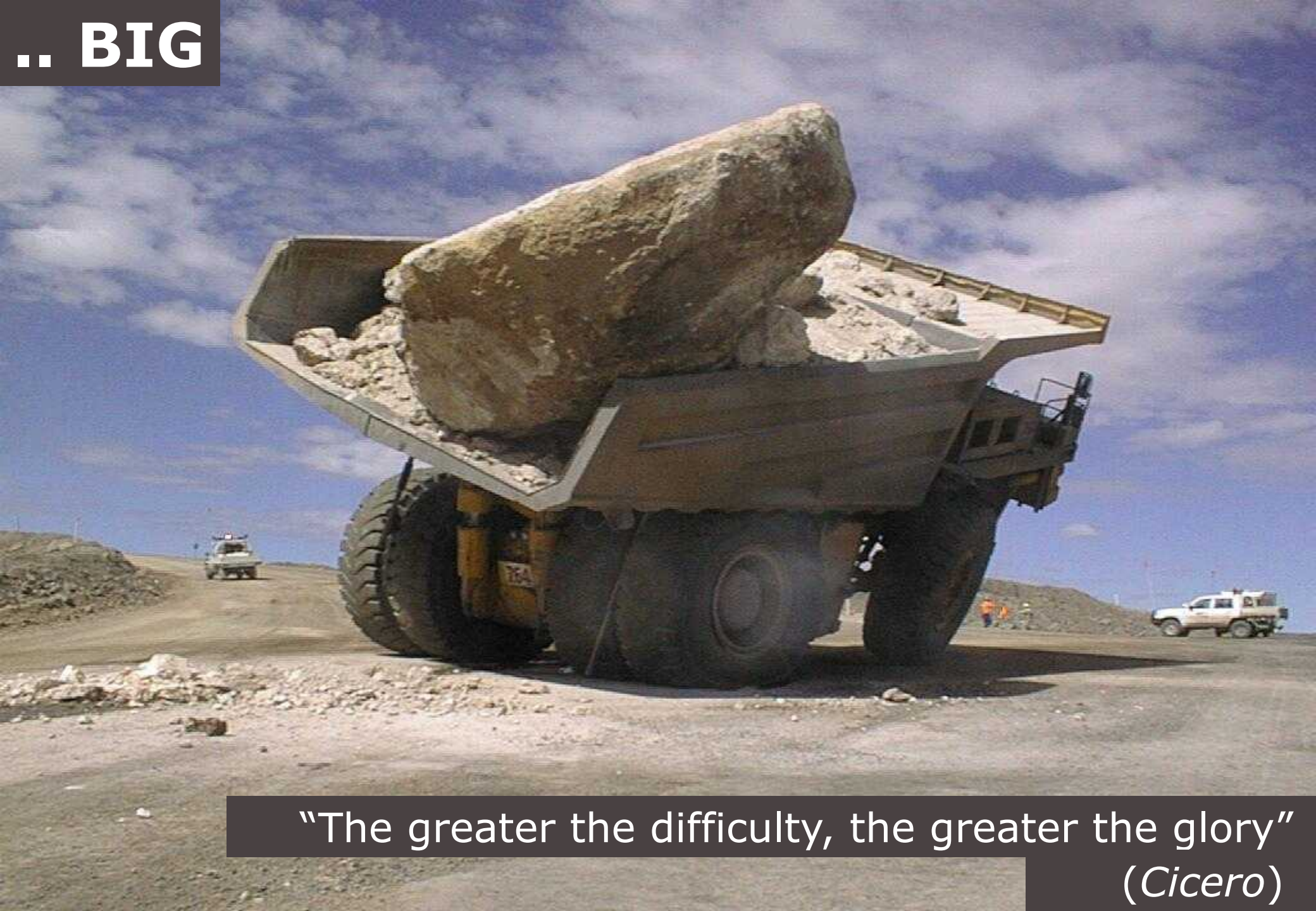
A modern course

- Introduced recently at Roma Tre
- First university course on Big Data in Italy
- We will experiment together some technologies
- We will take advantage of advanced infrastructures
- We will know research and applicative projects on Big Data
- We will meet people from industry working on Big Data
- In conclusion, we will face an adventure..

Big Data? Why?

Well, because they are..

.. BIG

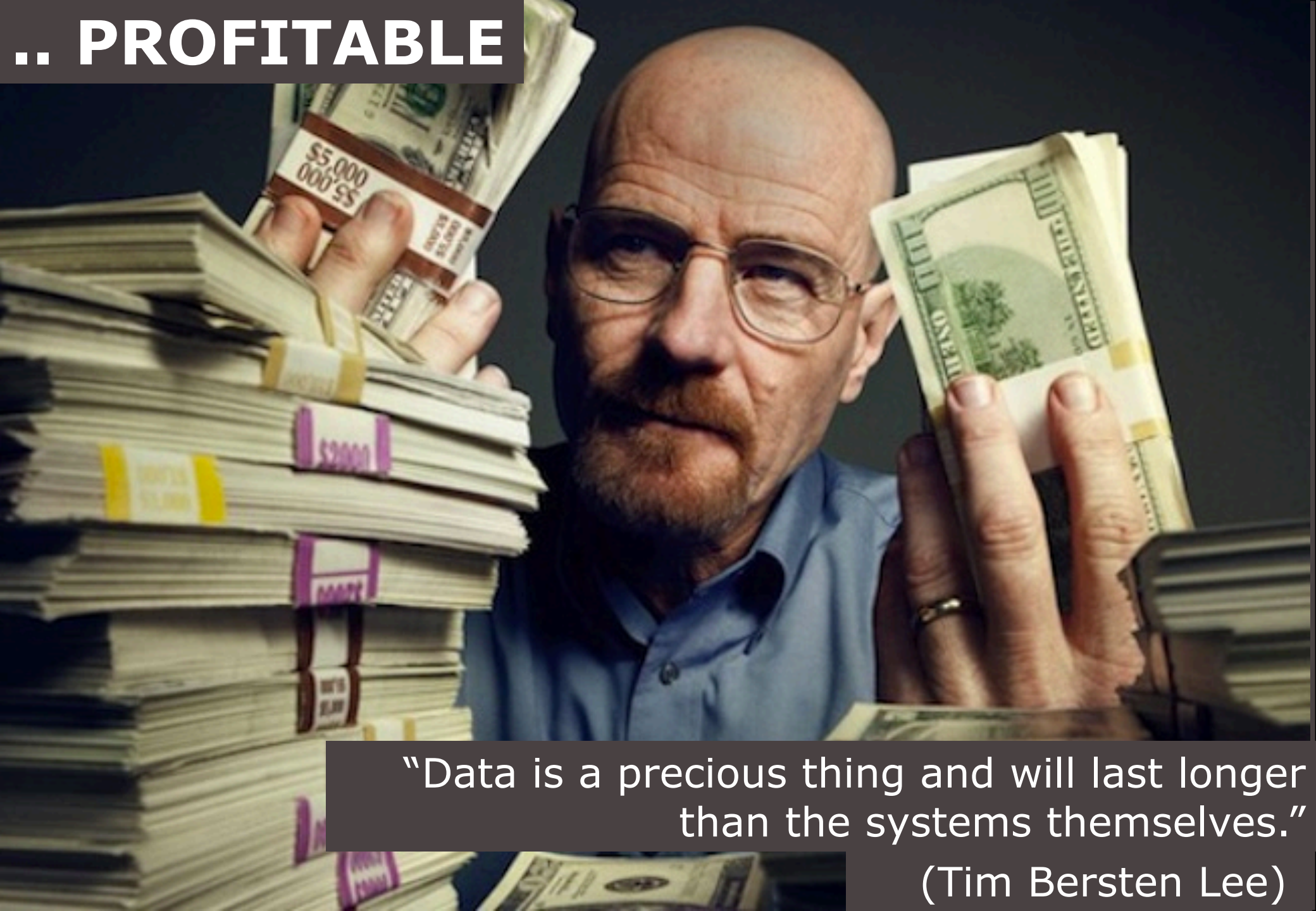


"The greater the difficulty, the greater the glory"
(*Cicero*)

.. CHALLENGING

"It always seems impossible until it is done."
(*Nelson Mandela*)

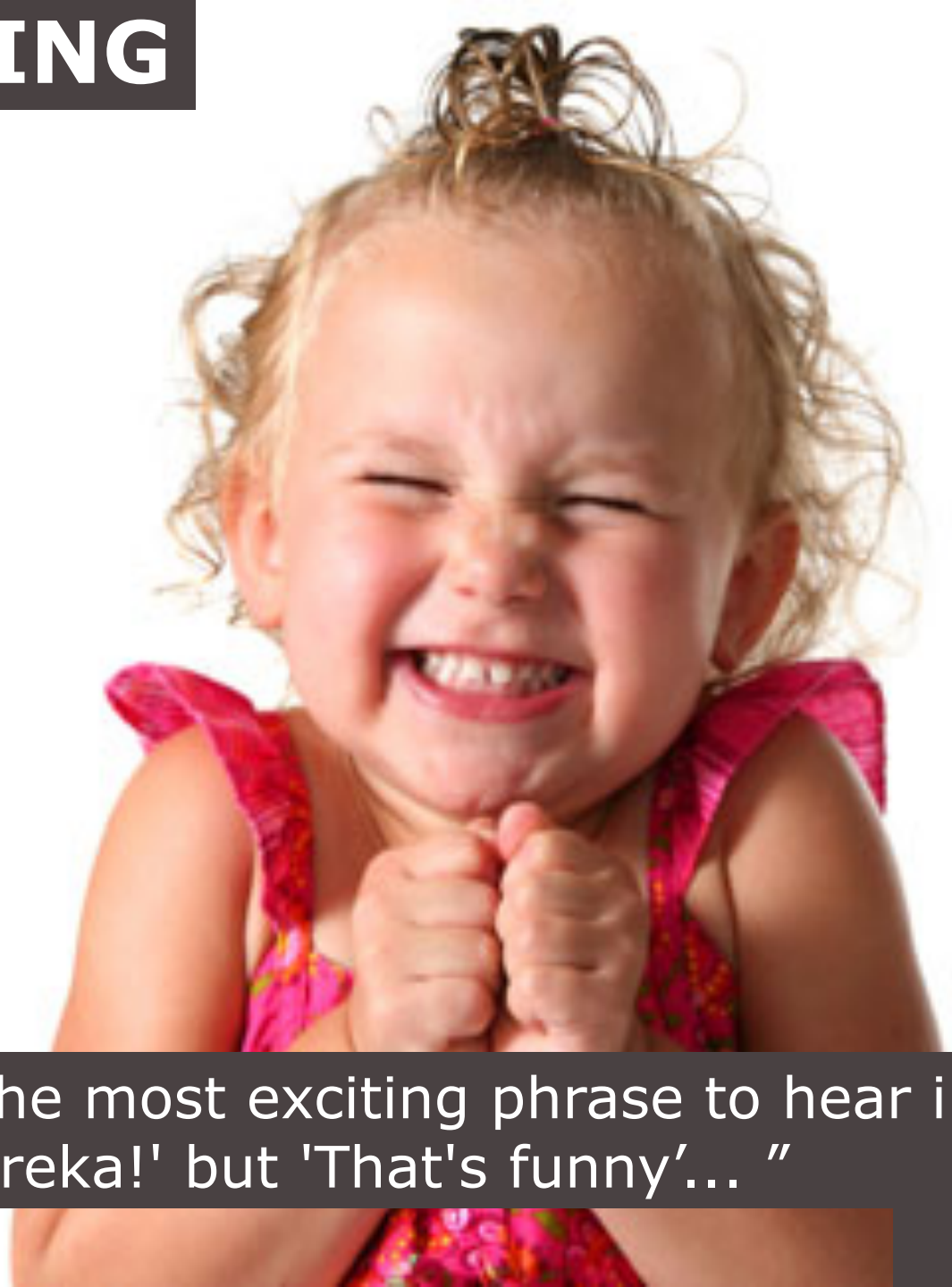
.. PROFITABLE



"Data is a precious thing and will last longer than the systems themselves."

(Tim Bersten Lee)

.. EXCITING



" The most exciting phrase to hear in science, is not 'Eureka!' but 'That's funny'... "

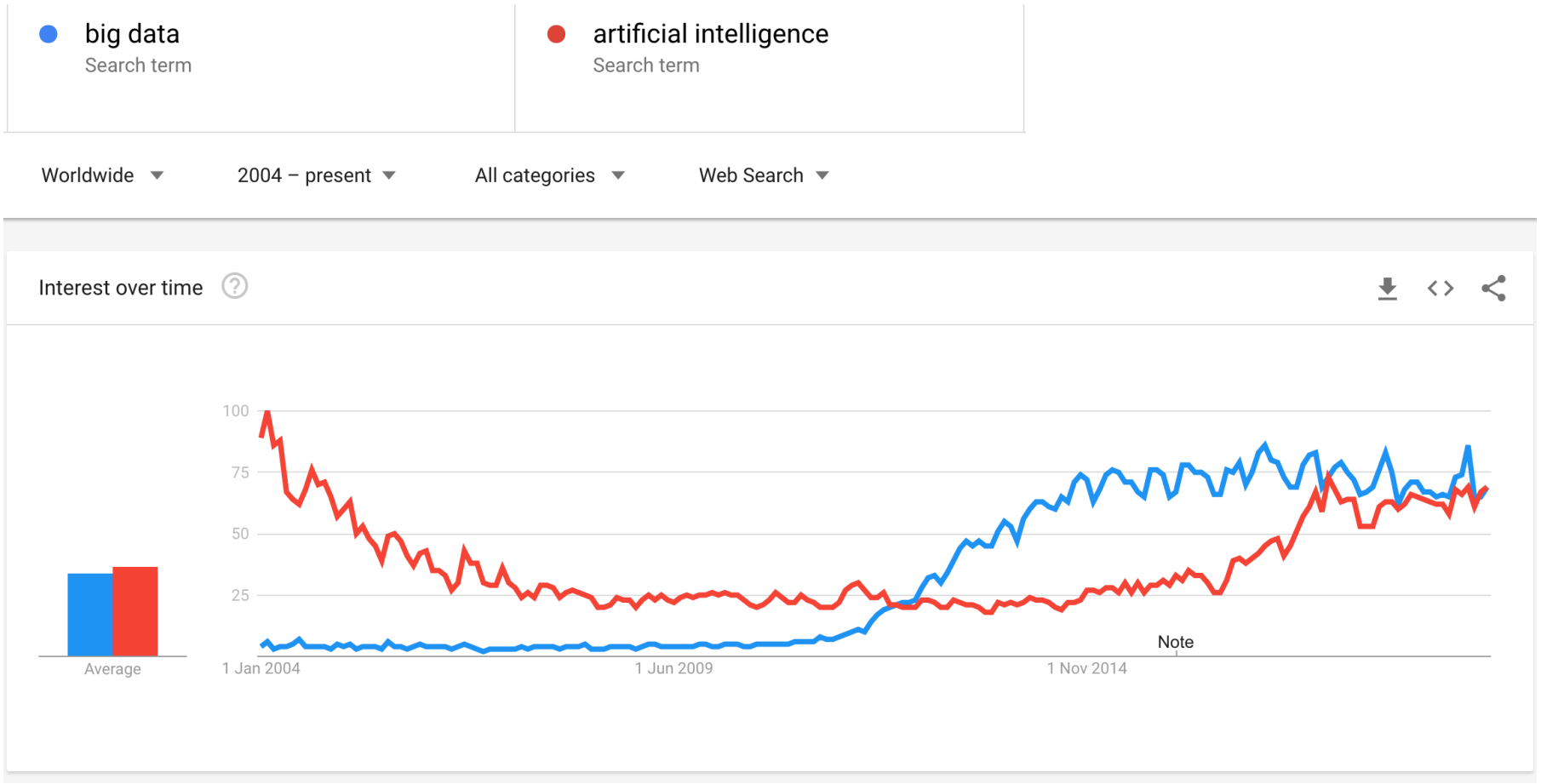
(*Isaac Asimov*)

.. FASHIONABLE

Fashion is about dreaming and making other people dream

Donatella Versace

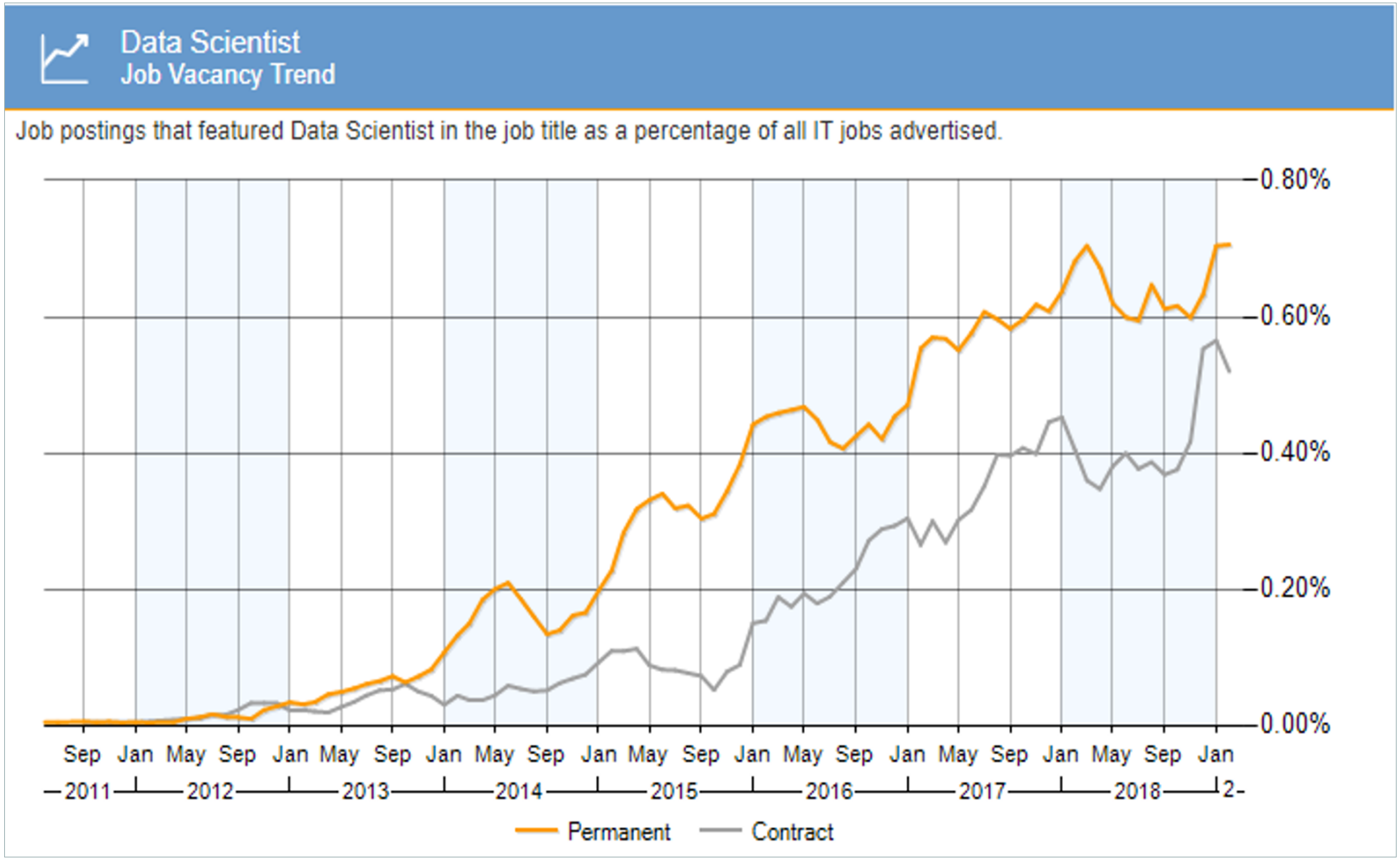
Topic trend


big data
Search term
artificial intelligence
Search term
Worldwide
2004 – present
All categories
Web Search
Interest over time
100
75
50
25
Note
Average
1 Jan 2004
1 Jun 2009
1 Nov 2014

Data scientist: a new profession

- Data Scientist: The Sexiest Job of the 21st Century [Harward Business Review 2013]
- Data scientist? A guide to 2015's hottest profession [Mashable 2015]
- "It's official – data scientist is the best job in America" [Forbes, 2016]

Opportunities for Data Scientists today



Some of them...

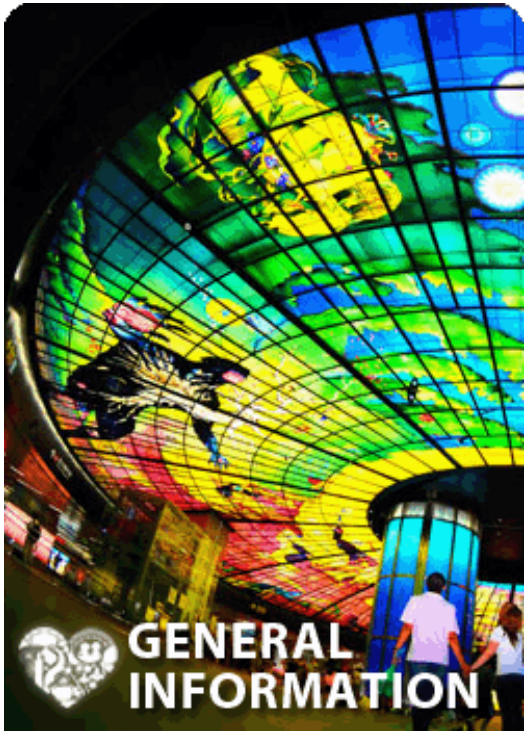
- Chiara Bartalotta (Unicredit)
- Edoardo Basili (Amazon)
- Davide Morgagni (BNL)
- Amir Salama (Bip)
- Andrea D'Amelio (Data Reply)
- Luca Massuda (Engineering)
- Costanza Brachetti (Data Reply)
- Roberto Fenaroli (Lottomatica)
- Caterina Mordente (BNL)
- Marco Ventirini (AMIGO)
- Fabio Scanu (Farfetch)
- Matteo Amadei (Enel)
- Pierluigi Pirro (Be)
- Andrea Alessi (BNL)
- Bernardo Marino (Engineering)
- Marco Santoni (Brembo)
- Luca Pasquini (Engineering)
- Marco Pavia (Altran)
- Simone Brundu (CERN)
- Miriana Mancini (Bridgestone)
- Leonardo Tilomelli (N26)
- Andrea Salvoni (KPI6)
- Nicholas Tucci (Big Telematics)
- Marco Faretra (NTT Data)
- Emanuele Rellini (Sogei)
- Marco De Leonardis (Banca d'Italia)
- Daniel Morales (KI Labs)
- Giulio Dini (Acea)
- David Santucci (Cloud Academy)
- Luca Dell'Anna (Qi4M)
- Enrico Petrachi (HCL)
- Marco Pavia (Altran)
- Angelo Del Re (Iconsulting)
- Carlo Loffredo (AbInitio)

After this course

"So you want to hire me as a Data Scientist for Intelligent Virtualized Deep Machine Learning Real-time Big Data in the Cloud for Social Networks? Ok, but if you also want Hadoop, increase my salary by 50%."

General information

- Teacher
  - Prof. Riccardo Torlone
  - Email: torlone@dia.uniroma3.it
- Office hours:
  - Wednesday, 14.00-16.00
  - Via Vasca Navale 79 – 2° floor – room 209
- Course Web site
  - http://torlone.dia.uniroma3.it/bigdata/
- Moodle page (projects)
  - https://moodle1.ing.uniroma3.it/
  - You must register!!
- A "social" course!
  - Facebook: https://www.facebook.com/groups/bigdataroma3/
  - Twitter: #bigdataroma3
- Lectures
  - Monday and Wednesday 11:00-12:30 (N13)
  - Pause: Easter holidays



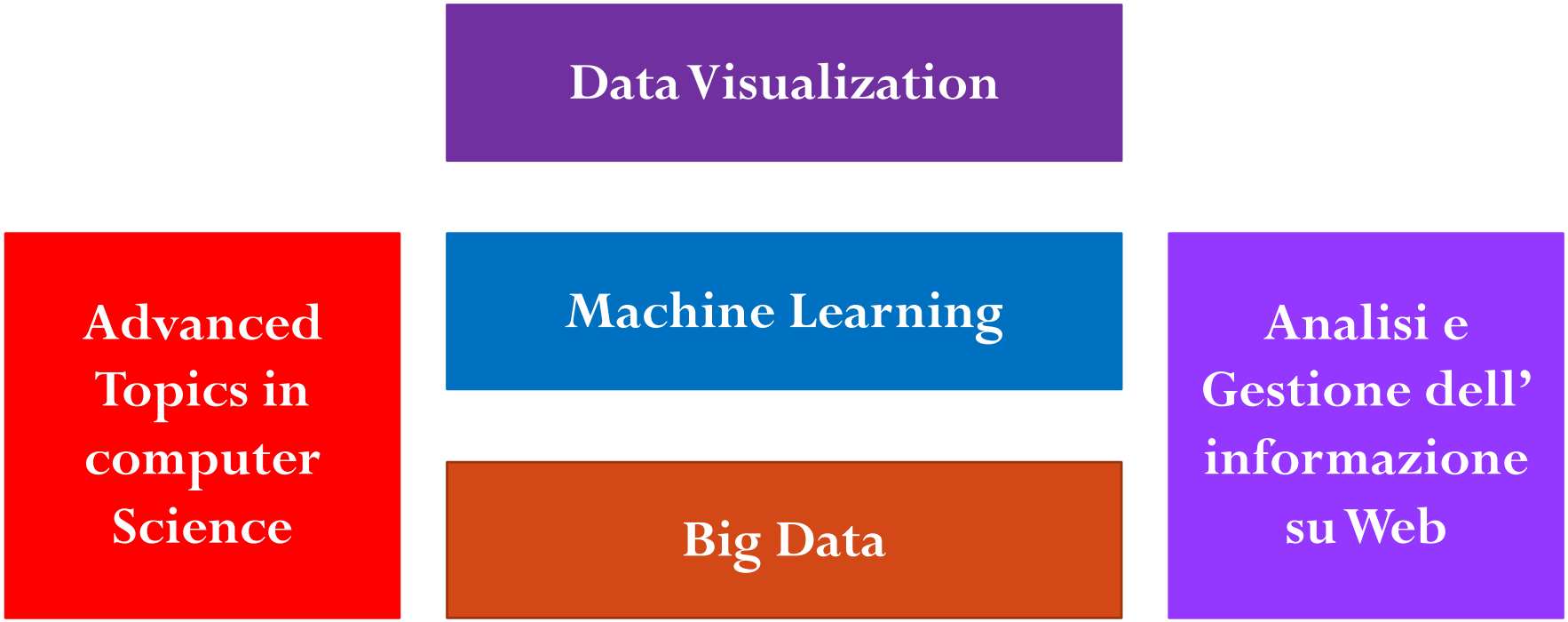
Goals

- *The course aims at illustrating tools and methods for the management of big data, i.e. massive amounts of unstructured data whose size exceed the capacity of conventional database management systems to capture, store, manage and analyze.*
- Focus on:
  - The requirements of modern applications
  - The problems of storing and processing big data
  - The hardware and software solutions
- Strategy:
  - Coverage of both methods and tools
  - Exercises with real systems
  - Practical projects
  - Guest lectures on Big Data use cases
  - Business seminars

Contents (provisional)

- Introduction
  - Terminology, main aspects and examples of applications.
- Infrastructures and programming paradigms for big data
  - Hadoop;
  - MapReduce;
  - Cloud computing;
- Big data processing
  - Hive;
  - Spark;
  - Kafka;
  - Beyond Spark.
- NoSQL systems
  - Introduction and data models
  - Sharding, replication and consistency
  - Implementation
- Big data analytics
  - Methods and techniques for data analysis.
- Applications
- Business seminars
- Challenges

Relationship with other courses



Past Business Seminars

An big event linked to the course!

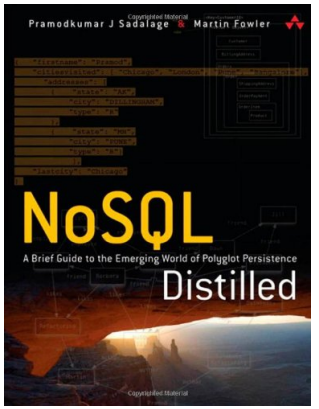
- An international summit focused on Technological, Economic, Legal and Social perspectives on Big Data
- Summit: October, 2020 co-located with

ROME
Maker Faire®
THE EUROPEAN EDITION

- Location: Fiera di Roma
- https://2019.datadriveninnovation.org/

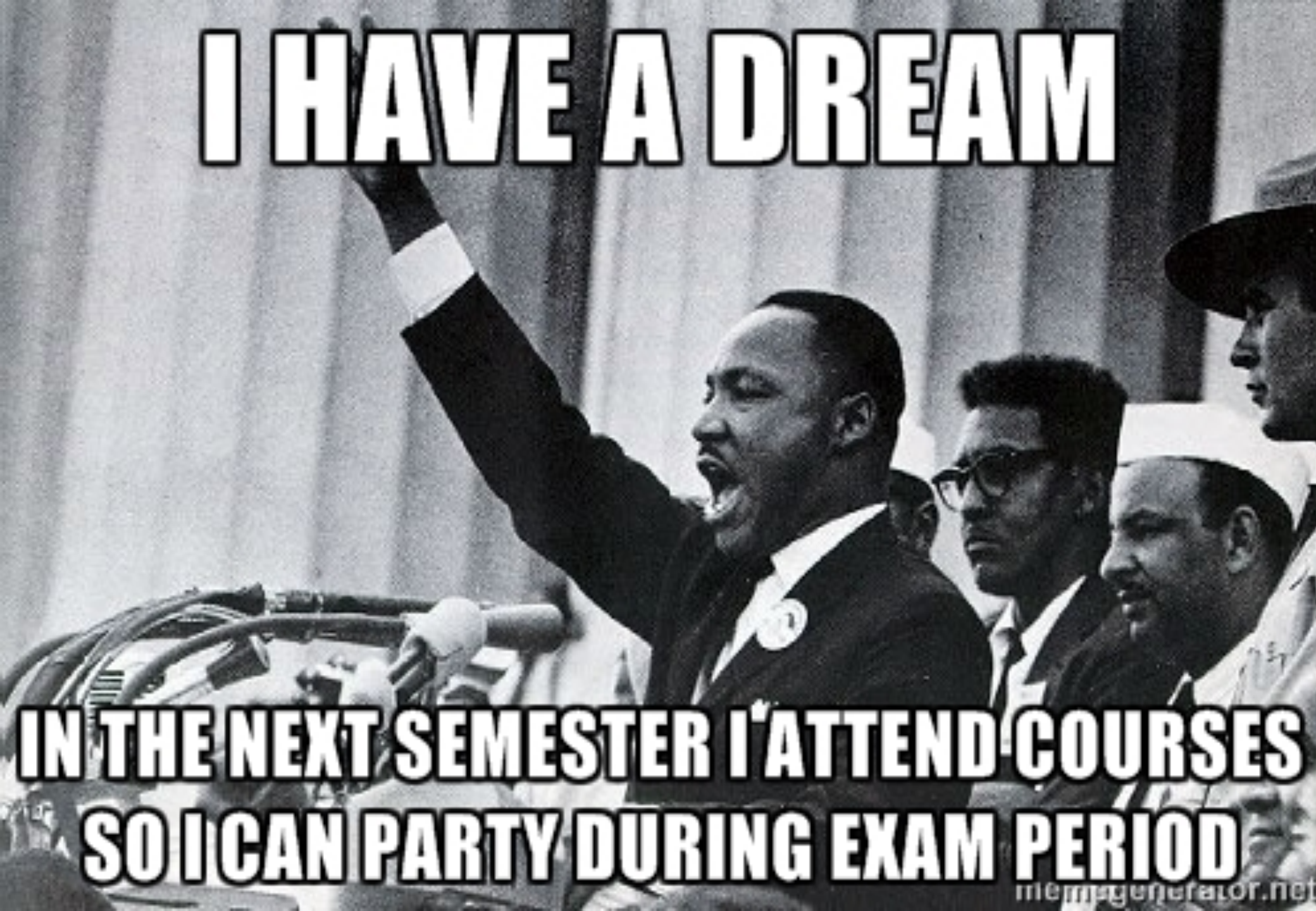
Material

- Books and papers
  - Teacher slides (available on the Web side of the course)
  - NoSQL systems:
    - Martin J. Fowler, Pramodkumar J. Sadalage. "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley, 2013.
  - Scientific papers and book chapters
    - To be published on the Web site of the course
- Software
  - Hadoop
  - PySpark
  - NoSQL systems
  - Others..
- Infrastructures
  - Amazon Web Services
  - Server Blade @ Roma3



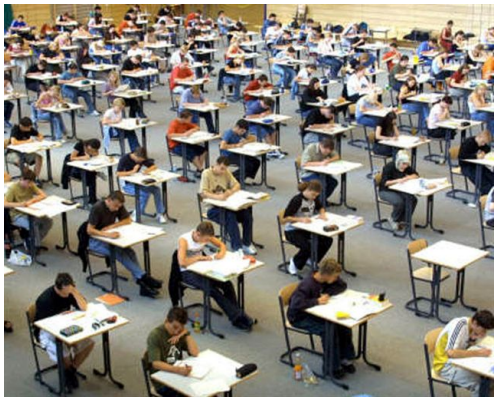
Exams..

- I have a dream..


I HAVE A DREAM
IN THE NEXT SEMESTER I ATTEND COURSES
SO I CAN PARTY DURING EXAM PERIOD
memegenerator.net

Exam modalities

- For those attending the course:
  - 2 projects to be done by groups of **1, 2**, max 3 students <u>with the same background</u>
    - Common project, deadline: mid April, weight:30%
    - Given project, deadline: before the exam, weight:40%
  - A written test: around 45 minutes, date of the exam, weight:30%
- For the other students:
  - Individual project, assigned by the teacher
  - A written test: 3 hours
- Rules:
  - Similar to all the other exams
  - Three chances: July 2020, September 2020, February 2021



Main project

- Goals
  - To solve a problem of Big data
  - To experiment new technologies
- Steps:
  - Find challenges and data
  - Choose an approach to analyze data
  - Choose suitable technologies
  - Implement the approach
  - Testing of the system

Statistiche