Programming for Data Science at Scale

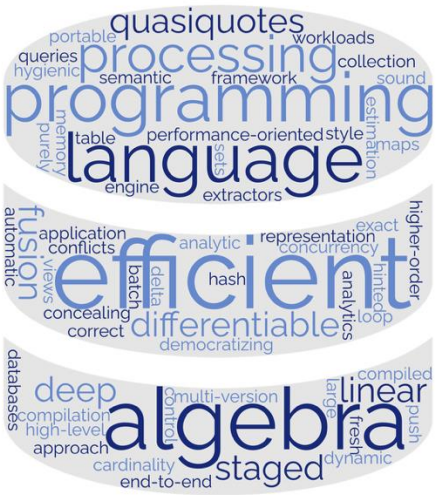
Introduction to Large-Scale Data Processing

THE UNIVERSITY *of* EDINBURGH

Amir Shaikhha, Fall 2025

Lecturer

- Amir Shaikhha
  - Reader
  - https://amirsh.github.io
  - Interests
    - Programming Languages
    - Database Systems
    - Compilers
    - Domain-Specific Languages



Essentials

- Webpage: http://course.inf.ed.ac.uk/pdss
- Piazza: https://piazza.com/class/mf7ce4fxdmg4yq
- Learn: https://www.learn.ed.ac.uk/ultra/courses/_127073_1/outline

Course Timetable

- Lectures (**UPDATED**):
  - Thursdays 16:10 – 17:30
  - Week 1: Lecture Theatre G.04 - Robson Building (Monday 10:10 – 11:30)
  - Weeks 2, 5, 10: Lecture Theatre B - 40 George Square
  - Weeks 3, 4, 6, 7, 8, 9: Lecture Theatre G.03 - 50 George Square
- Labs:
  - Mondays 14:00 – 15:30
  - Fridays 15:00 – 16:30
  - Appleton Tower, 6.06

Course assessment

- 100% coursework → No Exam
- CW1: 70%
- CW2: 30%

CW1: Group Coursework

- Goal: Learn to design, implement, optimize, evaluate, and document a large-scale data science system
- Group size: 3 students (formed by your own)
- Stage 1: Design, Implement, Optimize, Evaluate
- Stage 2: Write a paper
  - Template will be provided

CW2: Individual Coursework

- Goal: Learn to assess a large-scale data processing system
- Write a review for the paper and the code
  - Template will be provided

Coursework Schedule

<table>
<tr><td>Week 1 (Sep 15)</td><td></td><td>Week 7 (Oct 27)</td><td rowspan="3">CW1</td></tr>
<tr><td>Week 2 (Sep 22)</td><td></td><td>Week 8 (Nov 3)</td></tr>
<tr><td>Week 3 (Sep 29)</td><td></td><td>Week 9 (Nov 10)</td></tr>
<tr><td>Week 4 (Oct 6)</td><td rowspan="3">CW1</td><td>Week 10 (Nov 17)</td><td rowspan="2">CW2</td></tr>
<tr><td>Week 5 (Oct 13)</td><td>Week 11 (Nov 24)</td></tr>
<tr><td>Week 6 (Oct 20)</td><td></td><td></td></tr>
</table>

Labs

- Start: Week 3
- End: Week 10
- 3 Lab Sessions
  - Weeks 3, 4, 6
  - Will help you with coursework
  - Not graded
- Rest of the weeks
  - Work on group coursework with your peers
- 2 sessions of 2 hours per week
  - You need to only attend 1 session

Preferred Prerequisites

- Programming Languages
  - Strong programming skills
    - Java
    - Scala
    - C++
    - Python

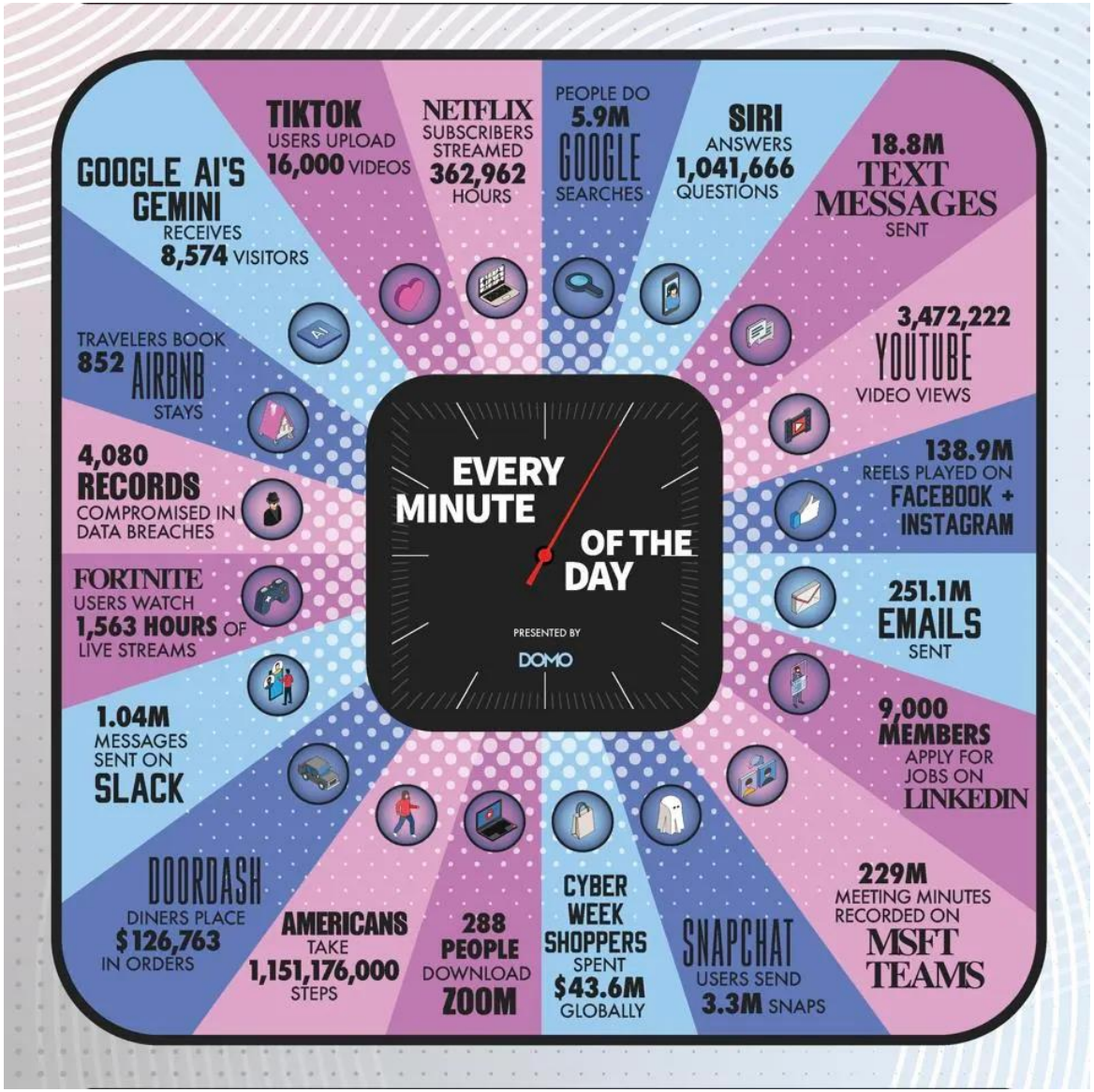
Acknowledgements

The lecture slides draw on notes by several folks to which I am grateful, in particular:

- P. Bhatotia (formerly Univ. of Edinburgh, now TUM)
- M. Odersky (EPFL)
- C. Koch (EPFL)
- H. Miller (CMU)
- M. Zaharia (Berkeley & DataBricks)
- The many researchers whose work I will mention in the slides (I will give pointers to their research papers)

COURSE OVERVIEW

Internet in 2025



https://bondhighplus.com/blog/what-happens-in-an-internet-minute/

Mainstream Languages for Data Scientists

Mainstream Languages for Data Scientists (cont.)

Pros

✓Rapid Development

✓Large community

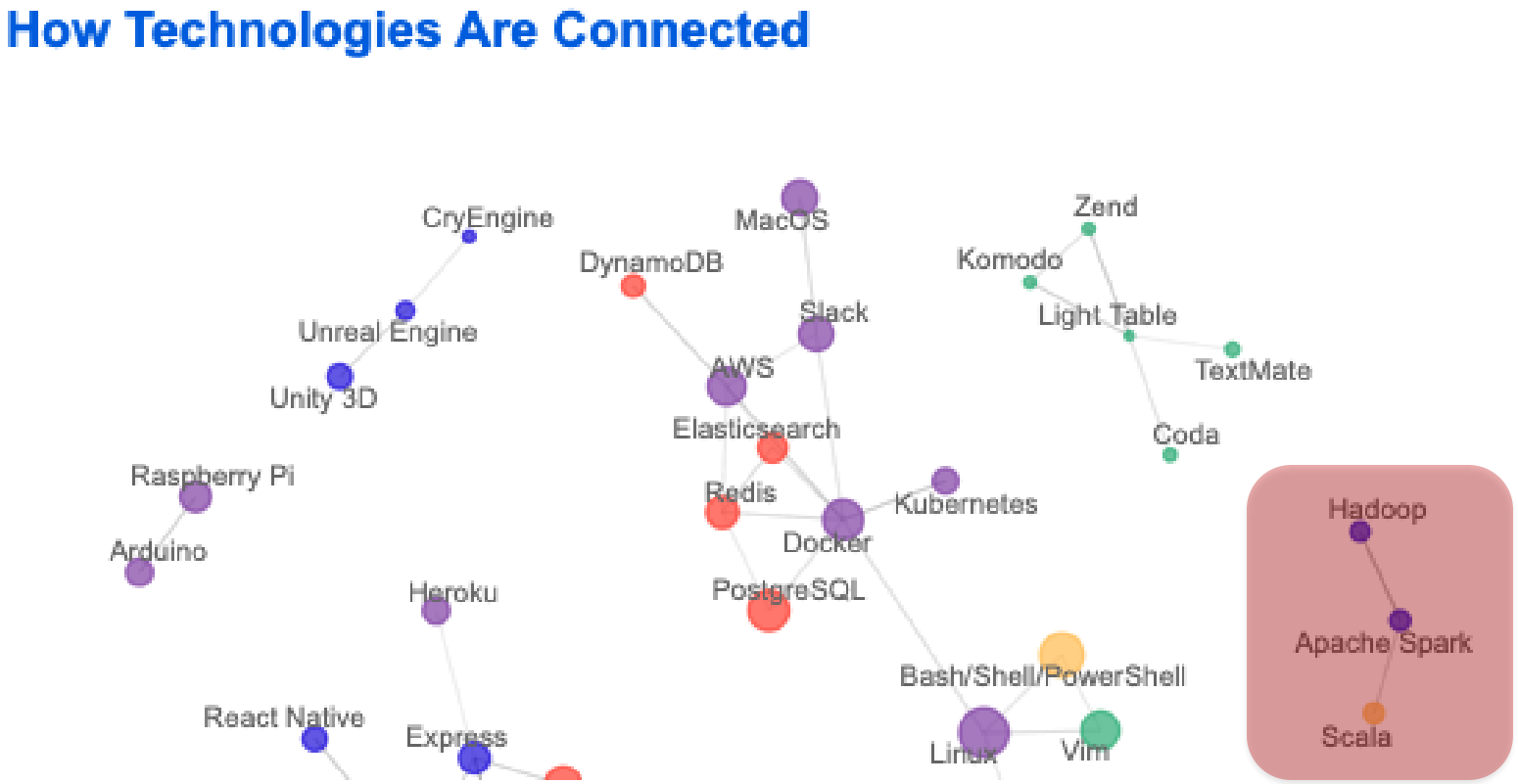
Cons

❖What to do with large datasets?

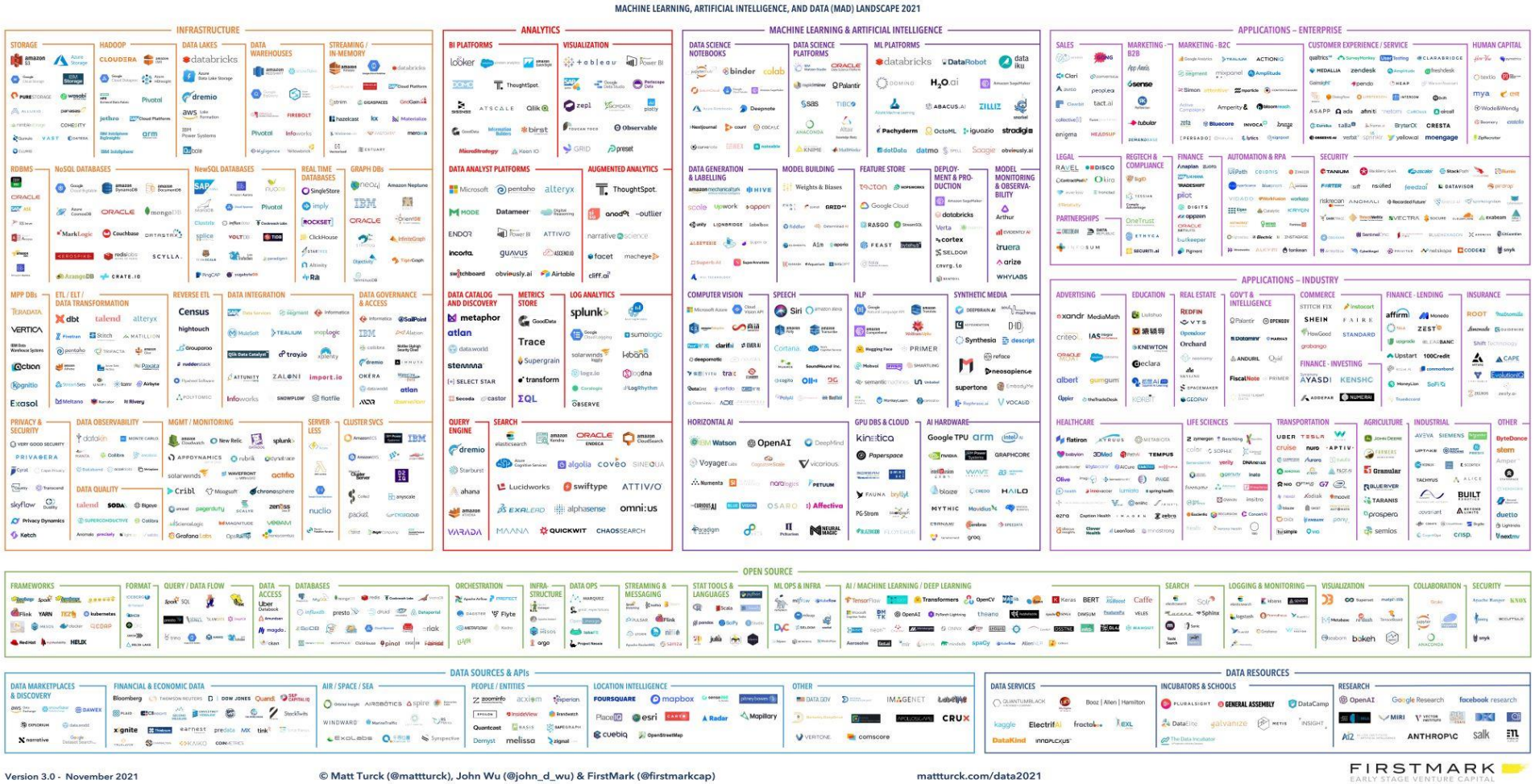
Rewrite from scratch ☹

Is there any language without this issue?

Why Scala is related to BigData?



https://insights.stackoverflow.com/survey/2019


MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021
INFRASTRUCTURE
ANALYTICS
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
OPEN SOURCE
DATA SOURCES & APIs
DATA RESOURCES
Version 3.0 - November 2021
© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)
mattturck.com/data2021
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Mainstream Big Data models

How to store, manage and process Big Data by harnessing large clusters of commodity nodes

- MapReduce family: simpler, more constrained

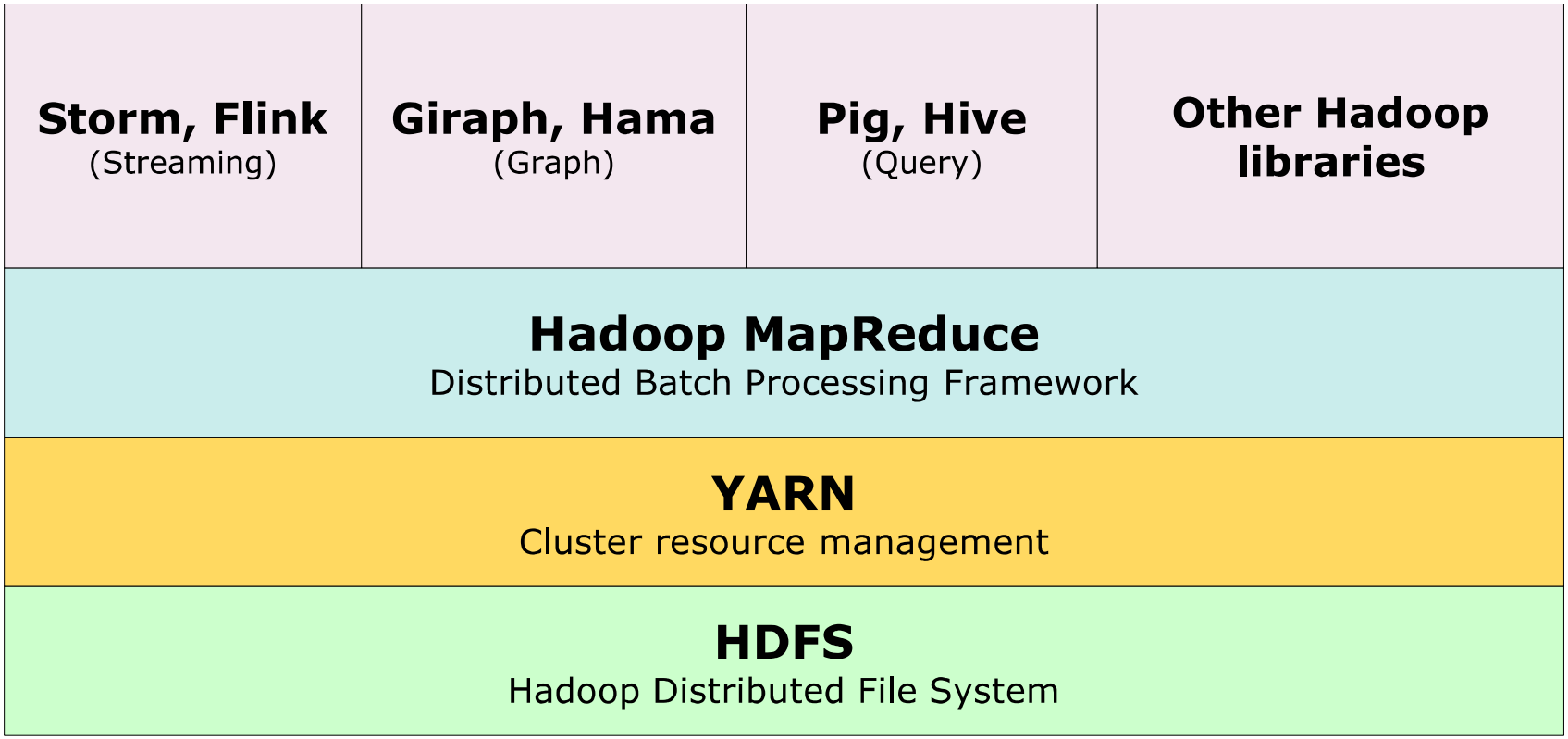
HadoopDB

- Dataflow family: enables more complex processing & data, optimization opportunities

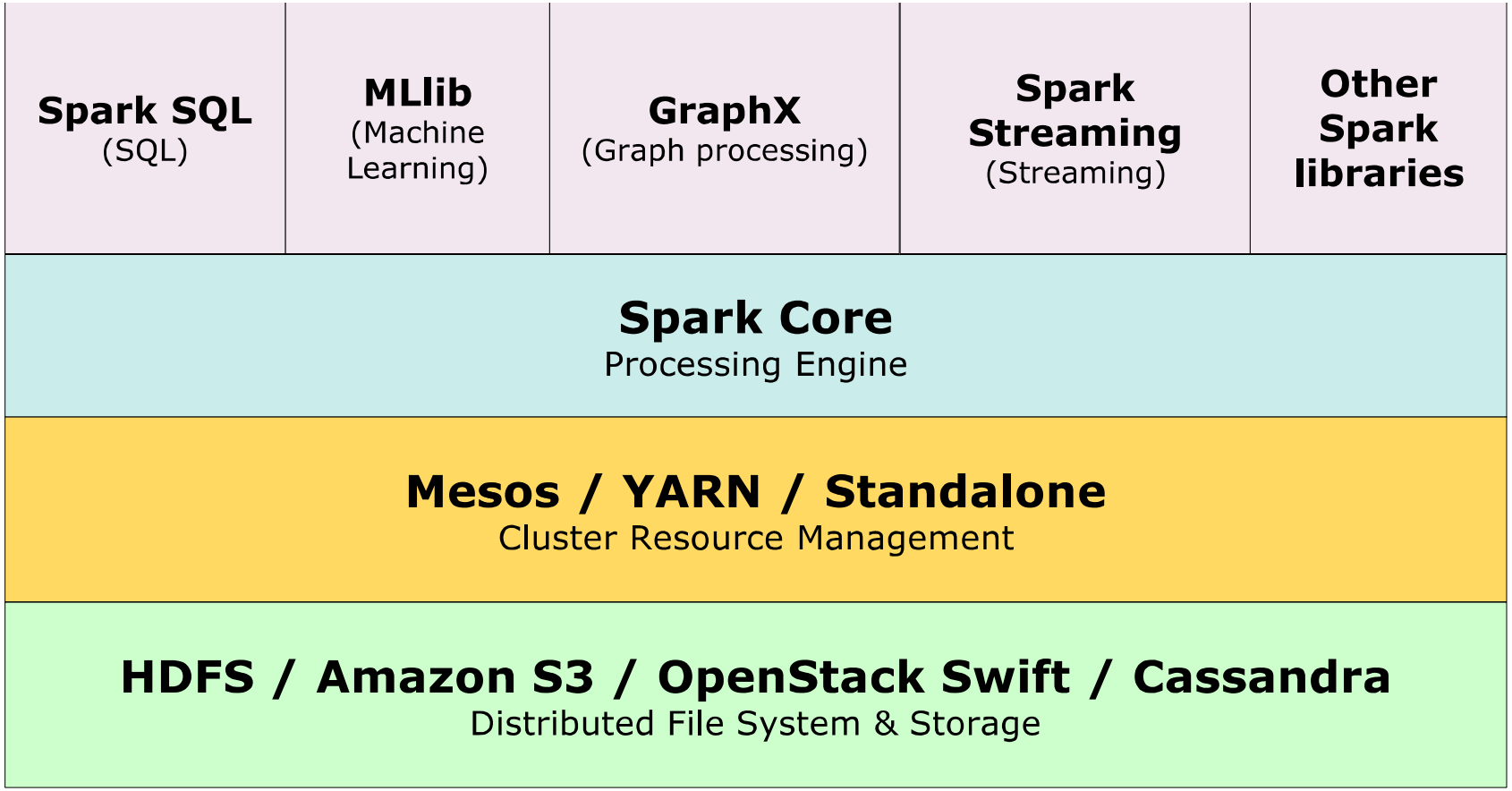
Google Pregel

Microsoft Dryad

The Hadoop Ecosystem



Spark Software Stack



Syllabus

- Data-Parallel Programming
- Functional Collections
- Distributed Data-Parallel Programming
- Distributed Key-Value Processing
- Optimizing Distributed Data Processing
- Distributed Query Processing
- Distributed Graph Processing

Guest Lecture

- Week 10
- Dr. Manos Karpathiotakis

QUESTIONS?