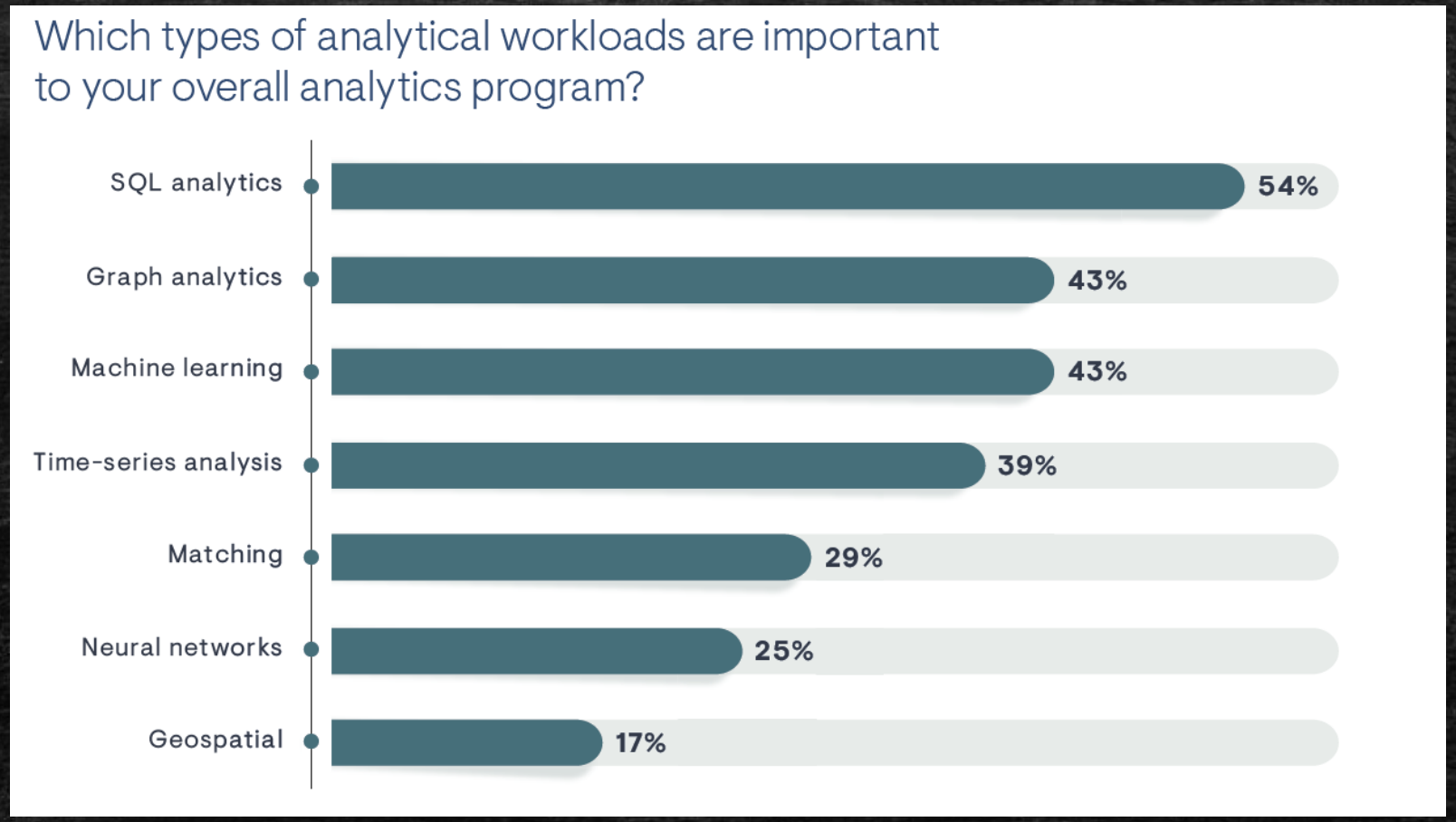
DSE 203 (Fall 2021)

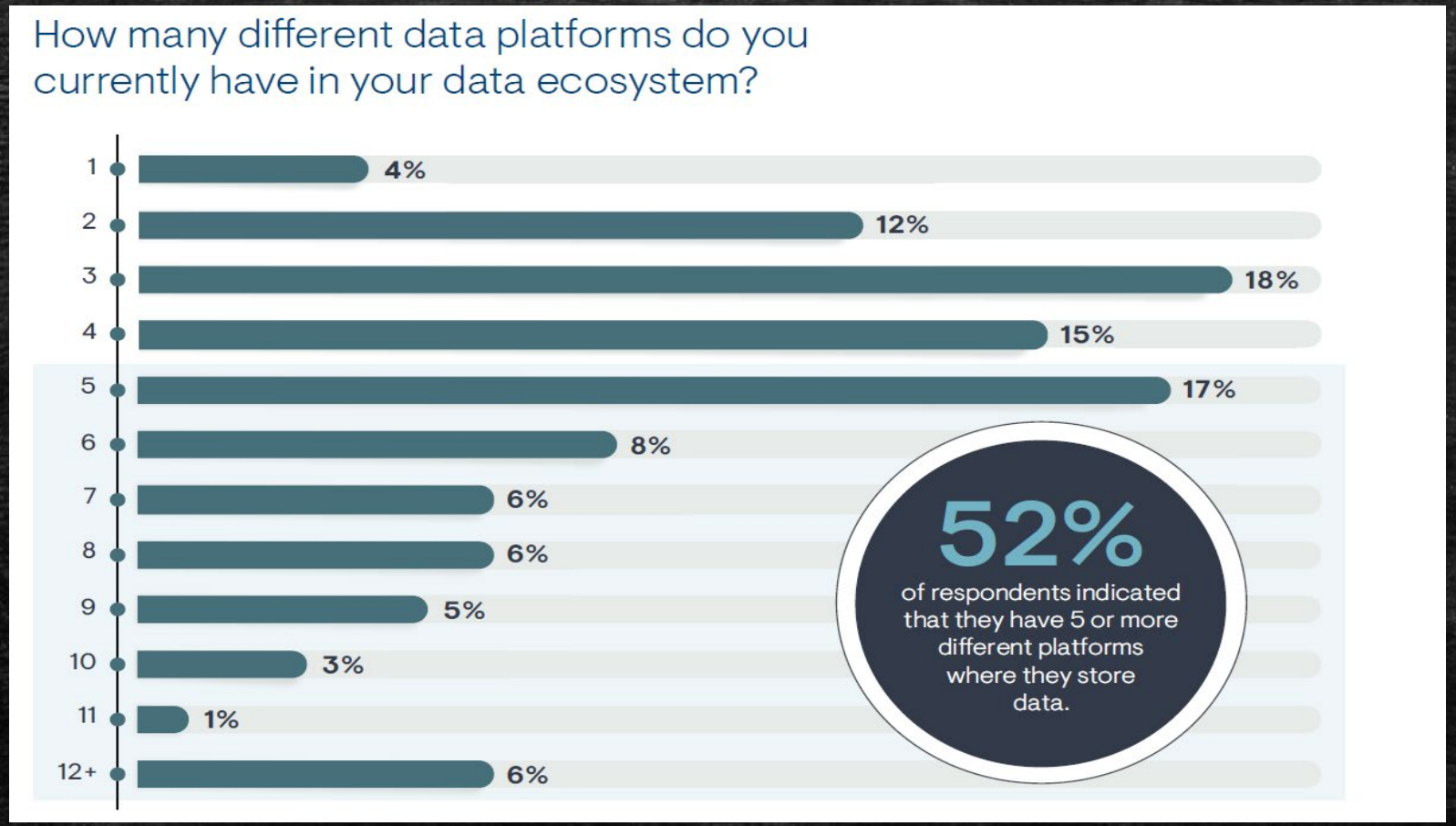
Class Logistics, Assignment and Project

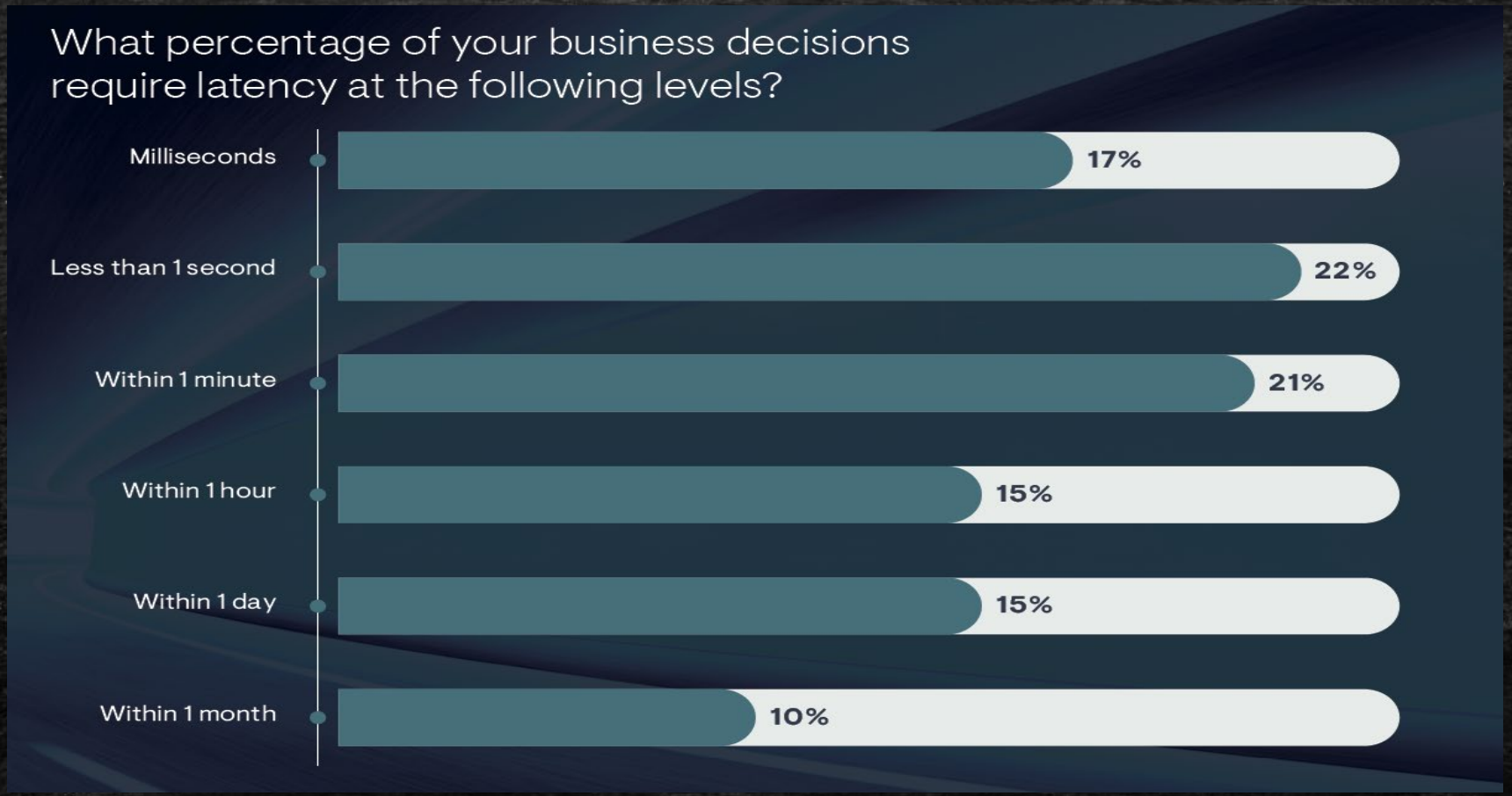
Instruction

- Instructor: Amarnath Gupta
  - Office: SDSC E-312 (on Zoom) – office hours by appointment
  - Email: a1gupta@ucsd.edu
- TAs
  - Xiuwen Zheng
    - Office: SDSC E-309
    - Email: xiz675@eng.ucsd.edu
  - Megha Agarwal
    - Office: SDSC E-309
    - Email: meagarwal@ucsd.edu
- Canvas: https://canvas.ucsd.edu/courses/29595
- Piazza: https://piazza.com/class/ktvotkzflm65ye Access Code: DSE203

Why this Course?
Results from a 2021 Survey on Data Analytics




How many different data platforms do you currently have in your data ecosystem?
1
4%
2
12%
3
18%
4
15%
5
17%
6
8%
7
6%
8
6%
9
5%
10
3%
11
1%
12+
6%
52%
of respondents indicated that they have 5 or more different platforms where they store data.


What percentage of your business decisions require latency at the following levels?
Milliseconds
17%
Less than 1 second
22%
Within 1 minute
21%
Within 1 hour
15%
Within 1 day
15%
Within 1 month
10%

The Data Pipeline Dilemma

Digital success requires rapid responses to business events, and data pipelines are needed to process and deliver insight. Unfortunately, data pipelines can take too much time to develop and place into production, creating a backlog of data and preventing real-time decisions.

Considering that many data pipelines are created by developers without the right tools or platforms to speed and automate development, 60% of respondents said that they take more than a business day to develop a data pipeline, with 27% in the three days to two months range.

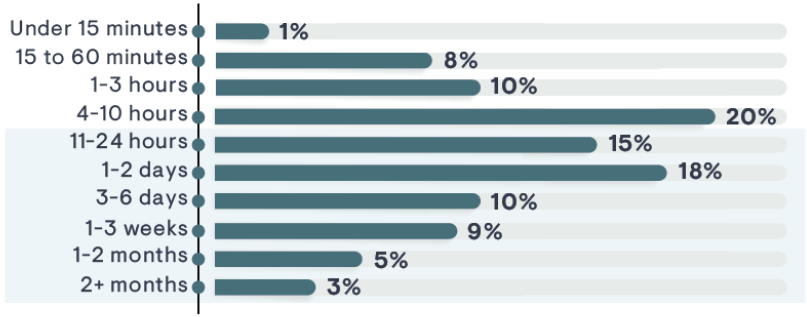
Further Delays in Data Pipeline Operationalization

In addition, making a data pipeline operational takes even more time. A full two-thirds of respondents said it takes at least another full business day to get new data pipelines into production, with 24% indicating that production takes more than a week.

The Difficulties of Data Pipelines

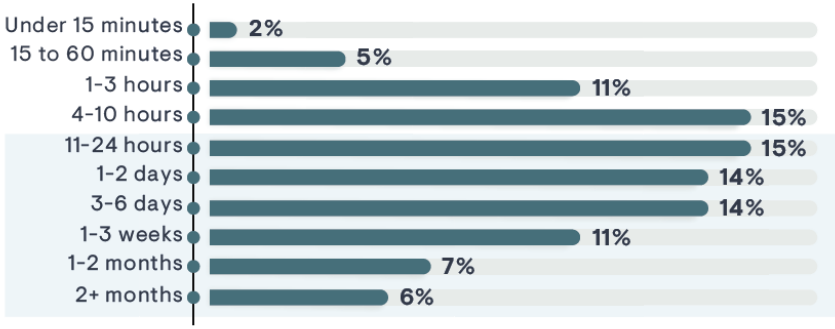
What is it that makes data pipelines so difficult? While the number-one answer given by participants was combining data in motion with data at rest, the next four answers give the real story. Data pipelines are inherently complex, and most companies are using too much manual coding. As a result, they have difficulty deploying error-free data pipelines and end up spending excessive time on fixing broken pipelines.

On average, how long does it take to develop a data pipeline?



(Sample Size = 402)

On average, how long does it take to make a data pipeline operational in production?



(Sample Size = 402)

What are the biggest challenges you face in building and deploying data pipelines?

(Sample Size = 402, Valid Cases = 402, Total Mentions = 850)

Top Trends in Data and Analytics For 2021

Accelerating Change

1. Smarter, Responsible, Scalable AI
2. Composable Data and Analytics
3. Data Fabric is The Foundation
4. From Big to Small and Wide Data

Operationalizing Business Value

5. XOps
6. Engineering Decision Intelligence
7. D&A as a Core Business Function

Distributed Everything

8. Graph Relates Everything
9. The Rise of the Augmented Consumer
10. D&A At the Edge

Source: Gartner

729348_C

Syllabus

- 5 Class Sessions and a Project Presentation Session

1. Introduction and the Value Matching Problem
2. Entity Matching, Machine Learning, and User-in-the-loop Matching
3. Data Integration Systems and Architectures
4. Handling Unstructured Data
5. Advanced Data Integration Techniques

- More Details on Canvas

Evaluation

- No mid-term and final exam
- 3 Programming Assignments (20 X 3 = 60 points)
  - Assignment 1
    - Using data sets given to you and a library of different distance functions,
      - Find the best value matching strategy
      - Implement a similarity join method
  - Assignment 2
    - Given a relational data set
      - Solve an entity resolution problem
      - Solve the same problem using an active learning library
  - Assignment 3
    - Given a JSON data set and a relational data set
      - Implement a workflow that will create a combined JSON data

Evaluation: Project

- Goal: Integrating structured, semistructured and unstructured data (40 points)
  - Groups of 3
  - Each group will
    - Get 3 different data sets and a taxonomy from instructors
    - Create a knowledge graph
    - Load the graph in Neo4J
    - Answer a set of queries on the knowledge graph
    - Some of the queries would have analytical functions
- Deliverables
  - Project presentation with a functional demo
  - The steps of integration as Jupyter Notebook
  - The final Neo4J graph as a zipped file

Last Year's Project

- Each group chose a set of data sources
  - Some needed my help
- They formed a design team with me and the Tas to validate their data integration and Knowledge Graph Problem
- Sometimes they used Wikipedia and other Knowledge Sources for Background data
- With this information and information other sources they produced, they created a knowledge graph to answer queries like
  - Find pairs of companies that compete in some areas and cooperate in other areas. Find these areas.
  - Which companies have acquired new companies to start a new product or service line?

We will hear from a student from the last cohort today