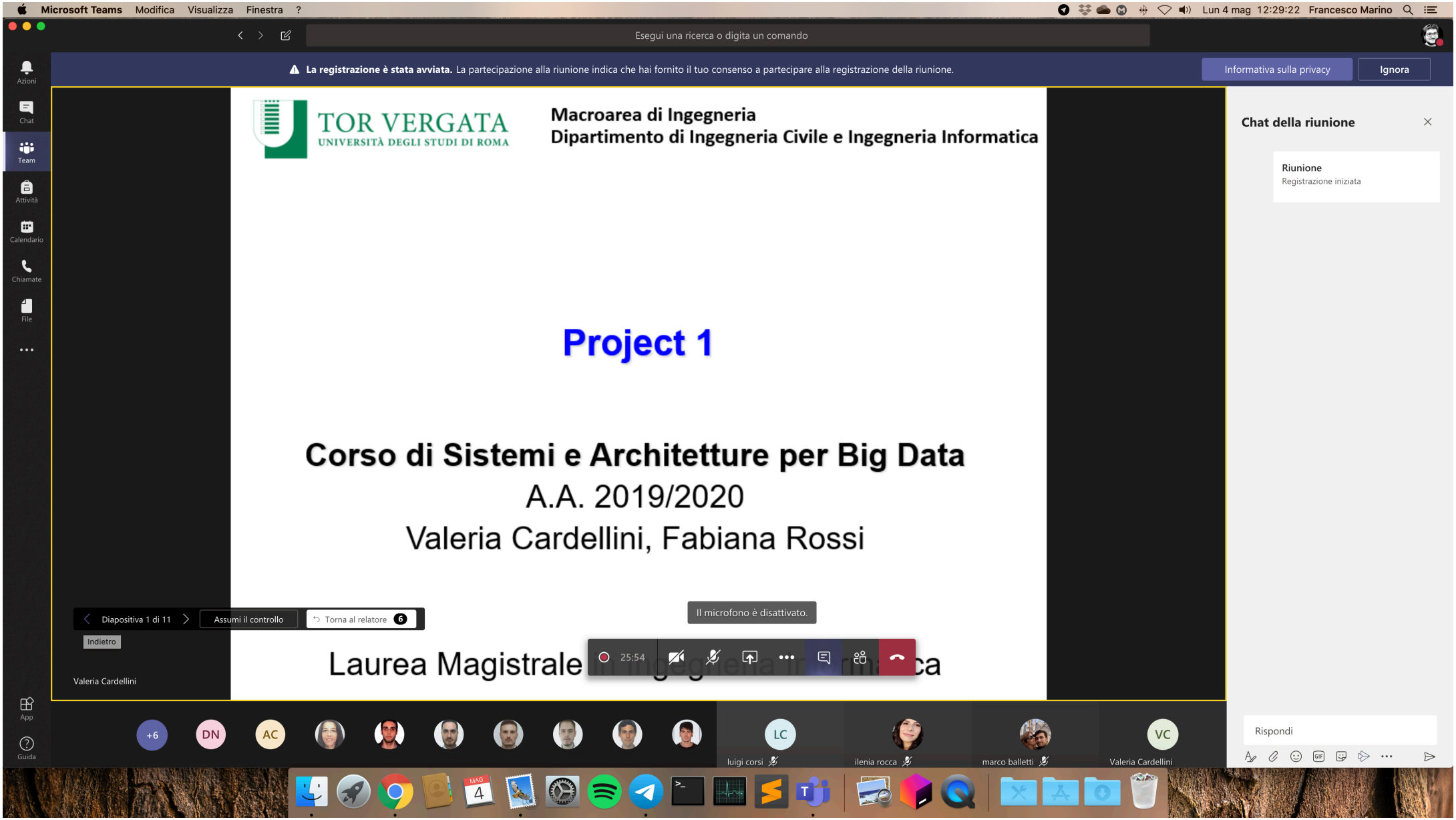
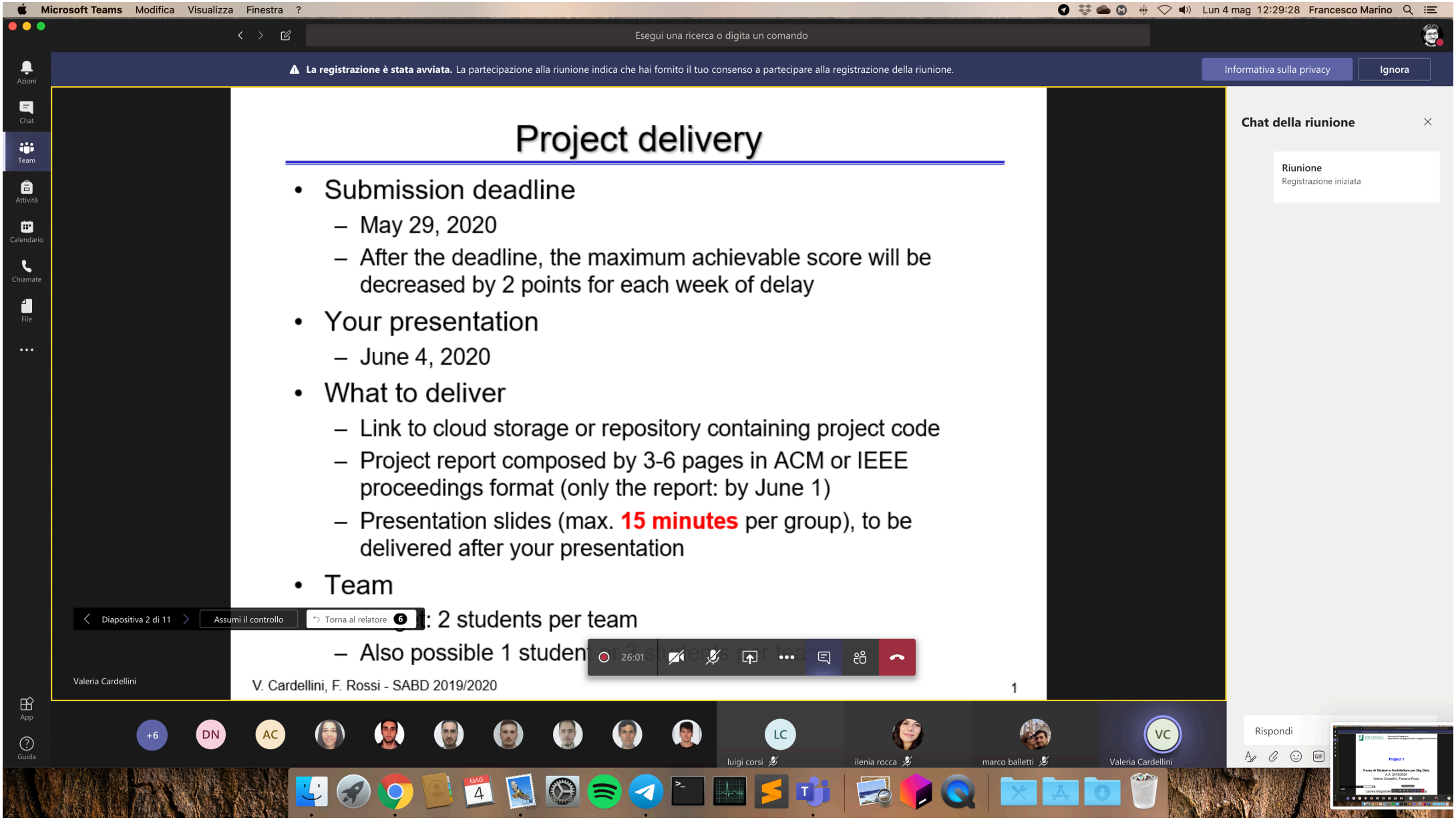
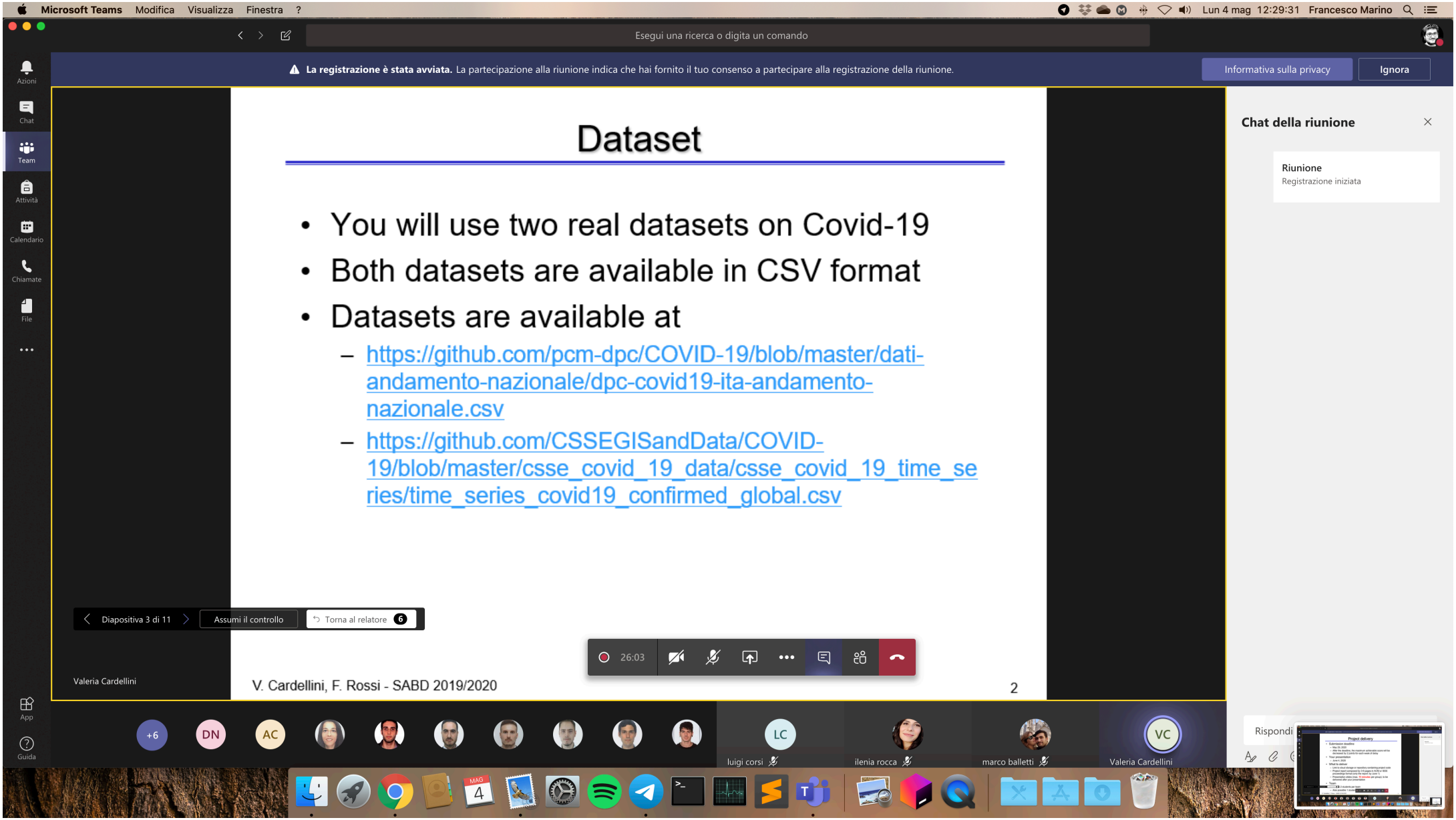
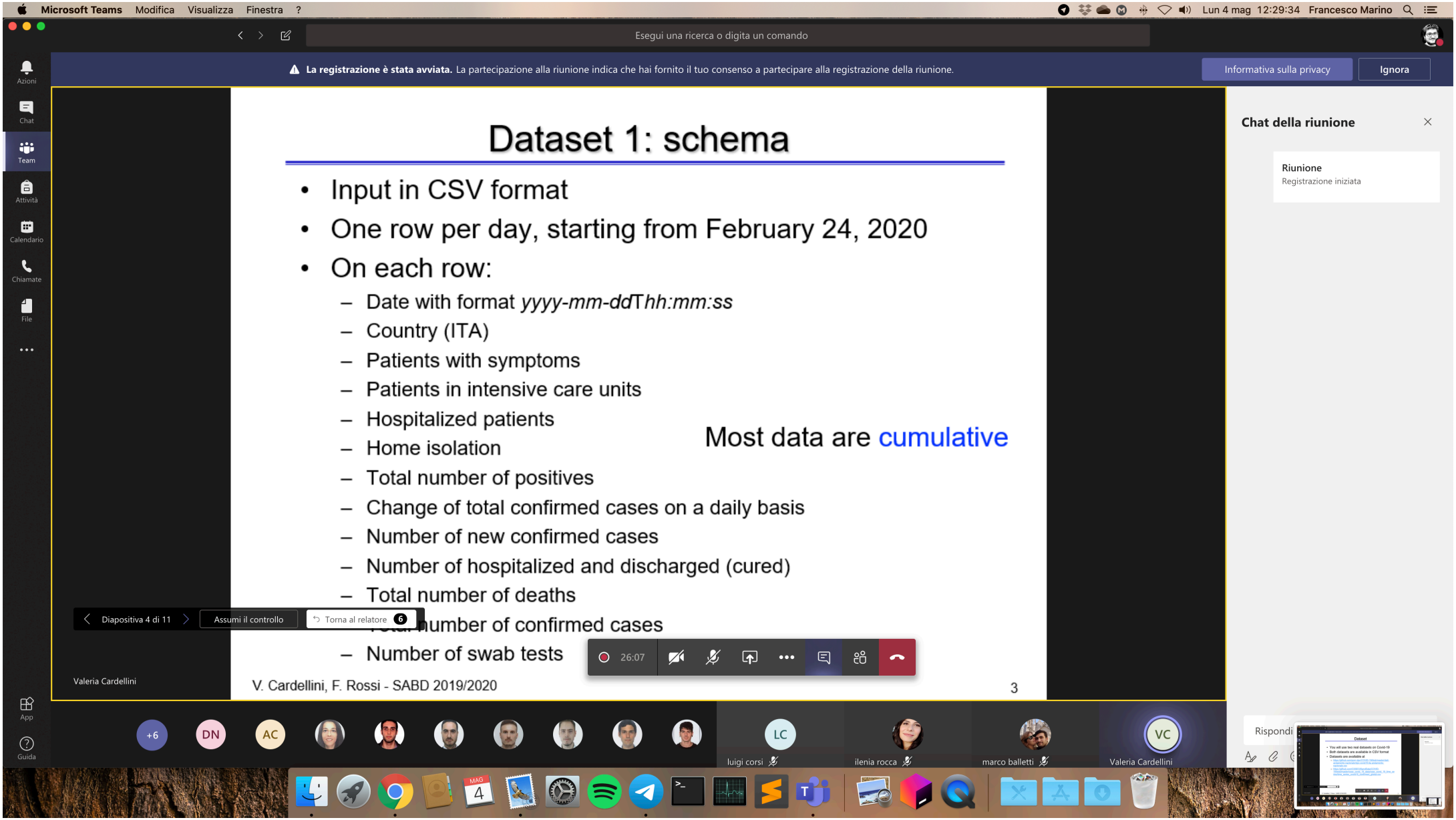
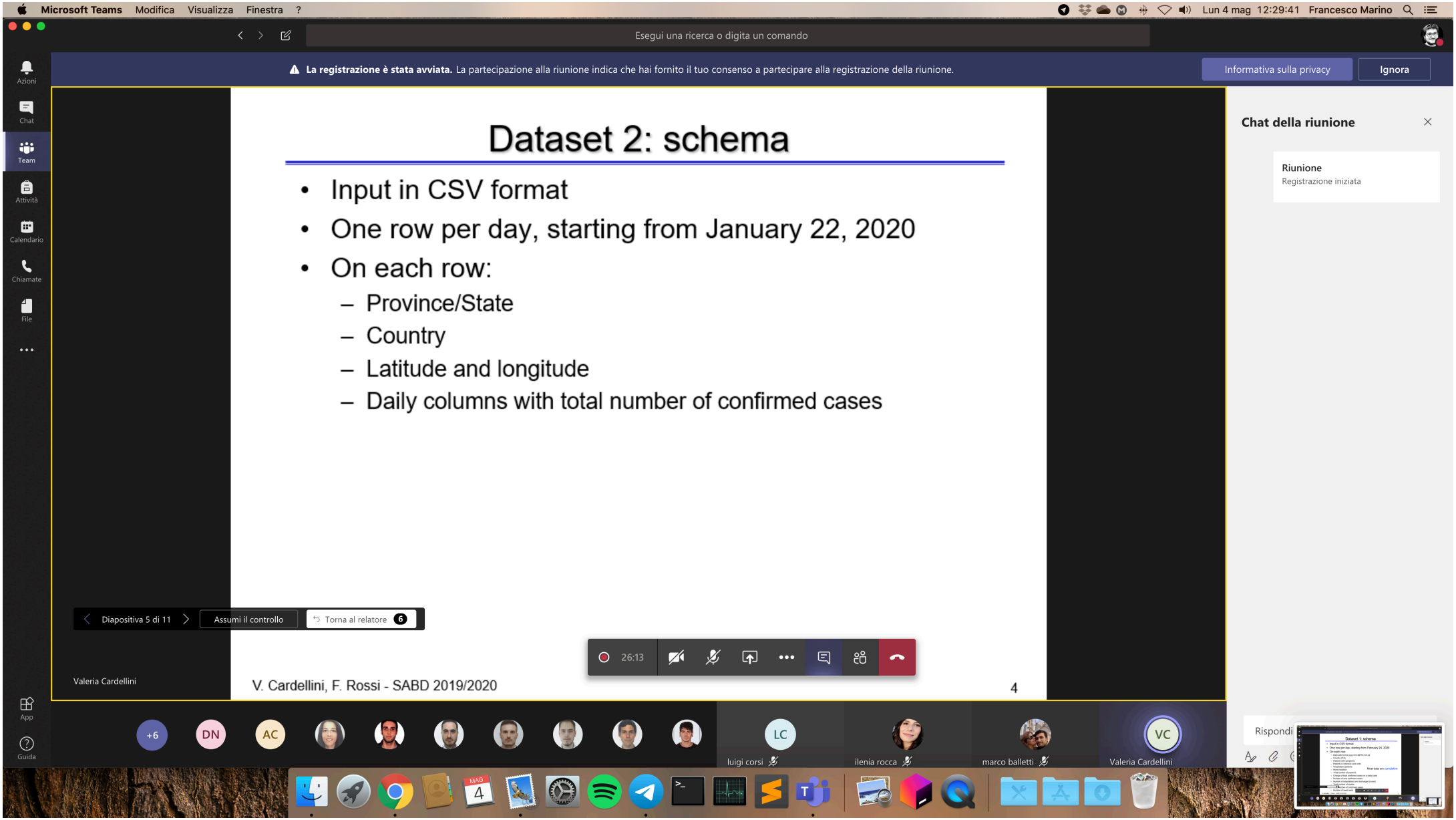
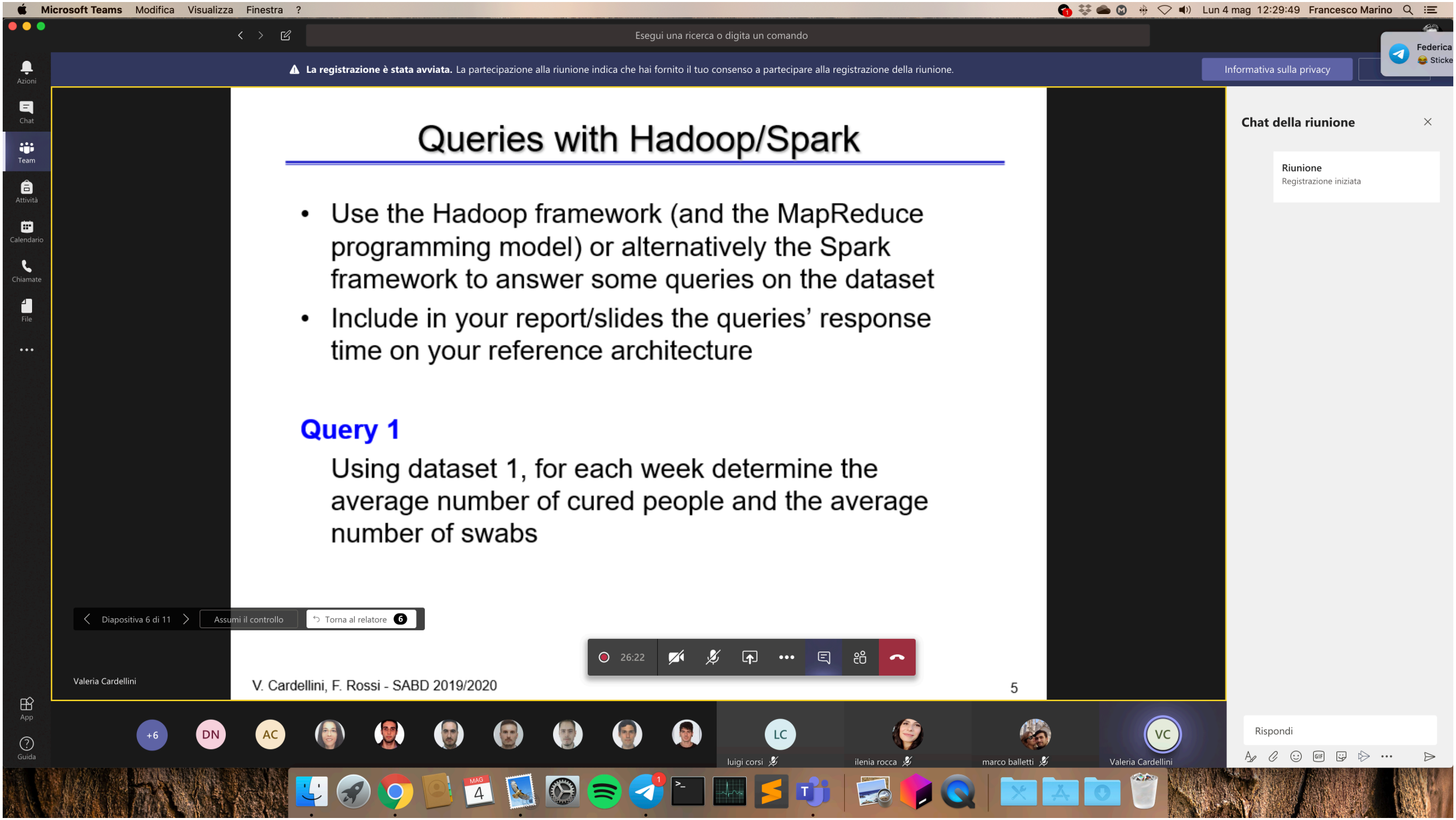

Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:29:22 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
App
Guida
TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA
Macroarea di Ingegneria
Dipartimento di Ingegneria Civile e Ingegneria Informatica
Project 1
Corso di Sistemi e Architetture per Big Data
A.A. 2019/2020
Valeria Cardellini, Fabiana Rossi
Laurea Magistrale
Diapositiva 1 di 11
Assumi il controllo
Torna al relatore
6
Indietro
Valeria Cardellini
Il microfono è disattivato.
25:54
Chat della riunione
Riunione
Registrazione iniziata
Rispondi
+6
DN
AC
LC
luigi corsi
ilenia rocca
marco balletti
VC
Valeria Cardellini

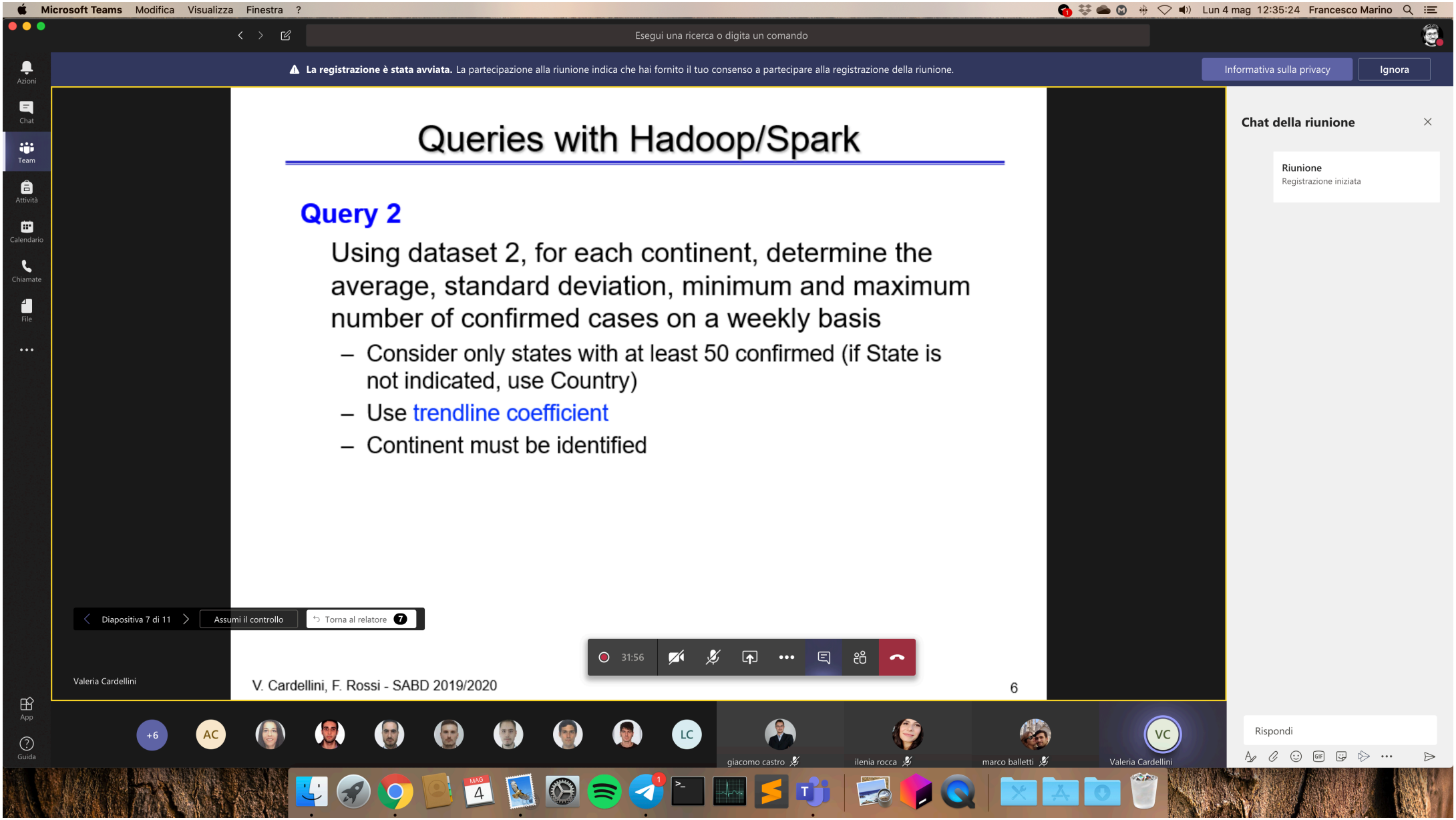

Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:29:28 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
App
Guida
Project delivery
• Submission deadline
– May 29, 2020
– After the deadline, the maximum achievable score will be decreased by 2 points for each week of delay
• Your presentation
– June 4, 2020
• What to deliver
– Link to cloud storage or repository containing project code
– Project report composed by 3-6 pages in ACM or IEEE proceedings format (only the report: by June 1)
– Presentation slides (max. 15 minutes per group), to be delivered after your presentation
• Team
– : 2 students per team
– Also possible 1 studen
Diapositiva 2 di 11
Assumi il controllo
Torna al relatore 6
26:01
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
1
Chat della riunione
Riunione
Registrazione iniziata
Rispondi
+6
DN
AC
LC
luigi corsi
ilenia rocca
marco balletti
VC
Valeria Cardellini

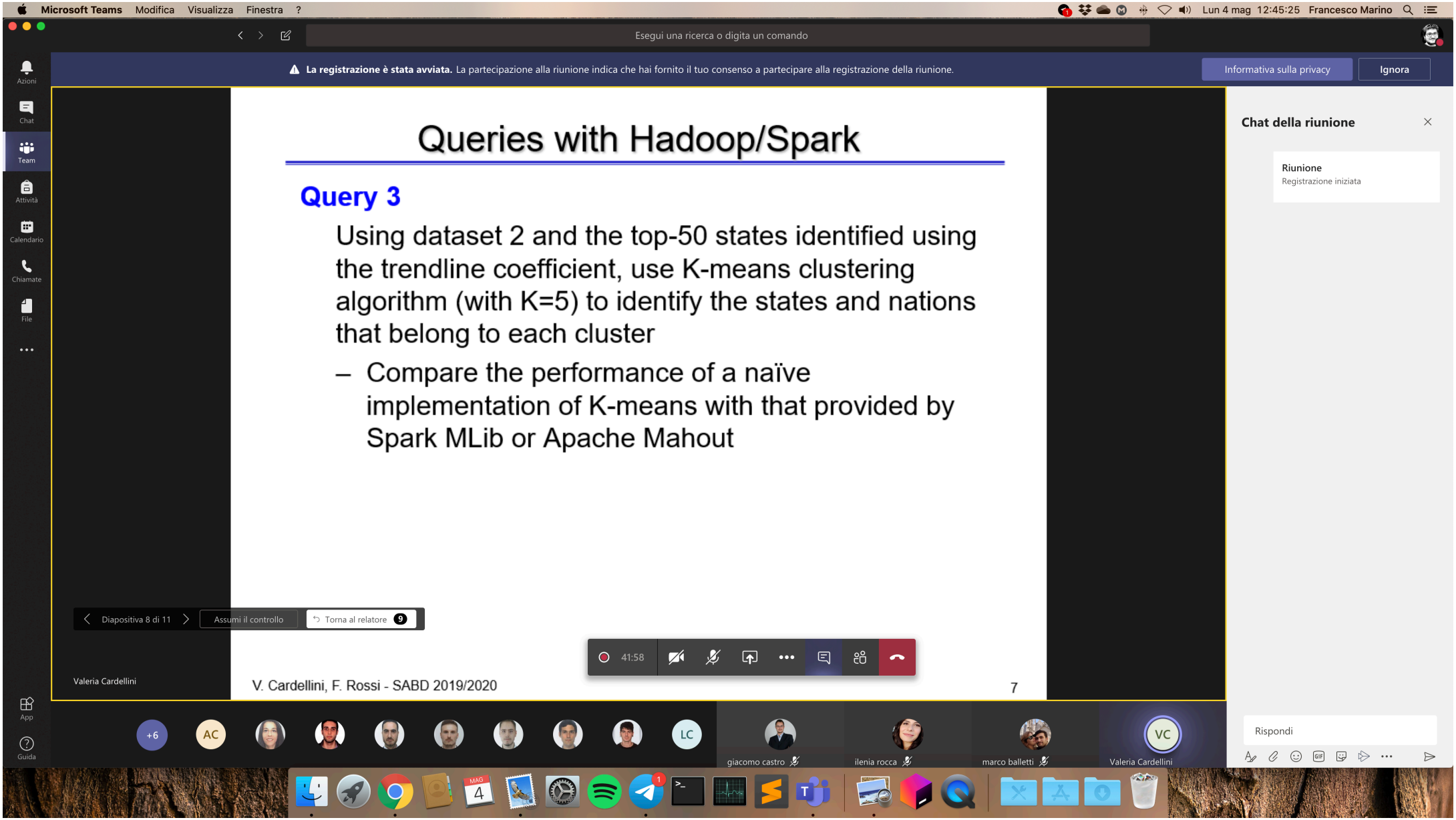

Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:29:31 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
App
Guida
Dataset
You will use two real datasets on Covid-19
Both datasets are available in CSV format
Datasets are available at
https://github.com/pcm-dpc/COVID-19/blob/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv
https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
Diapositiva 3 di 11
Assumi il controllo
Torna al relatore 6
26:03
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
2
Chat della riunione
Riunione
Registrazione iniziata
+6
DN
AC
luigi corsi
ilenia rocca
marco balletti
Valeria Cardellini
Rispondi

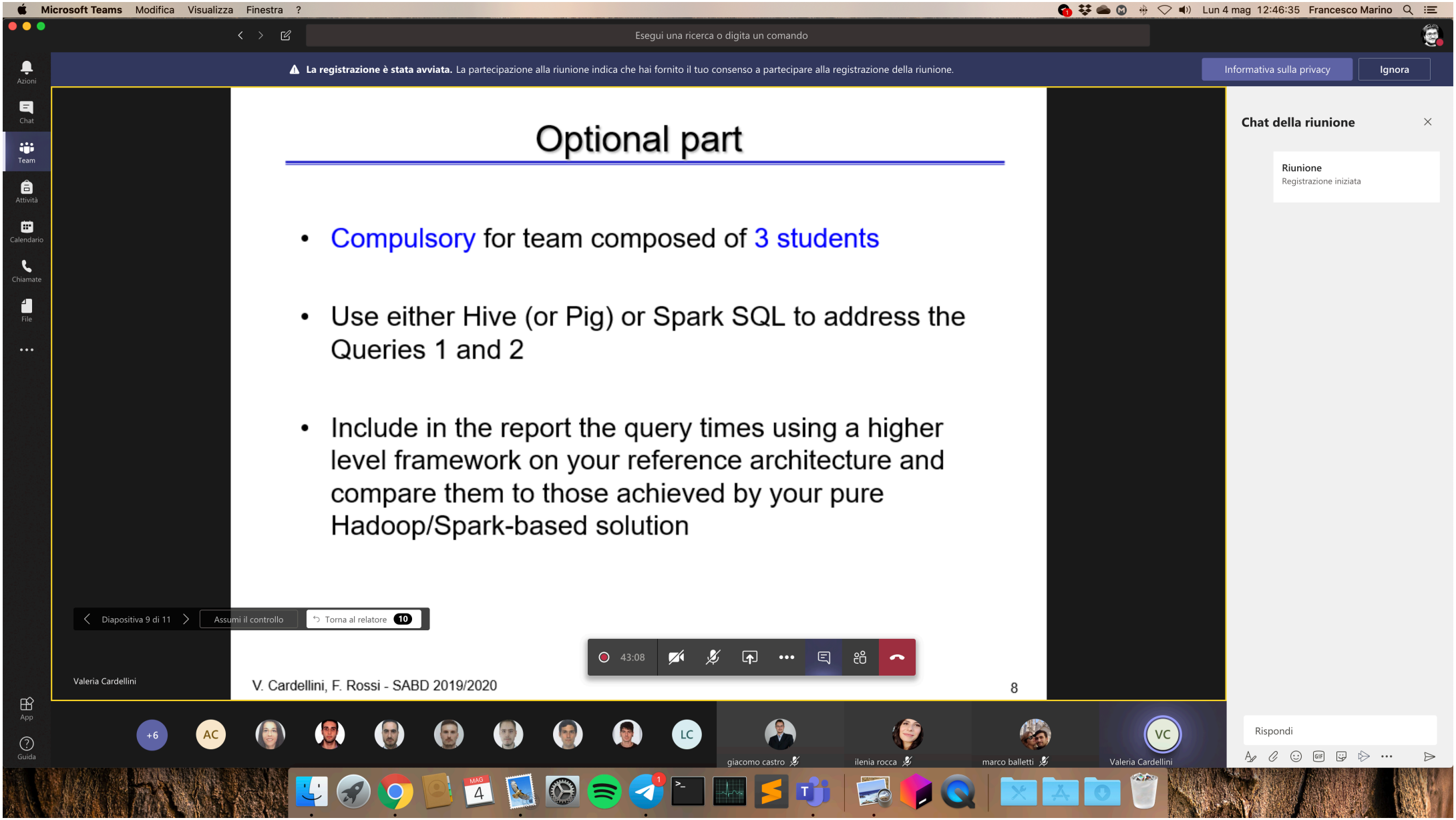

Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:29:34 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
Dataset 1: schema
Input in CSV format
One row per day, starting from February 24, 2020
On each row:
Date with format yyyy-mm-ddThh:mm:ss
Country (ITA)
Patients with symptoms
Patients in intensive care units
Hospitalized patients
Home isolation
Total number of positives
Change of total confirmed cases on a daily basis
Number of new confirmed cases
Number of hospitalized and discharged (cured)
Total number of deaths
Total number of confirmed cases
Number of swab tests
Most data are cumulative
Diapositiva 4 di 11
Assumi il controllo
Torna al relatore 6
26:07
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
3
Chat della riunione
Riunione
Registrazione iniziata
+6
DN
AC
LC
luigi corsi
ilenia rocca
marco balletti
VC
Valeria Cardellini
Rispondi
App
Guida

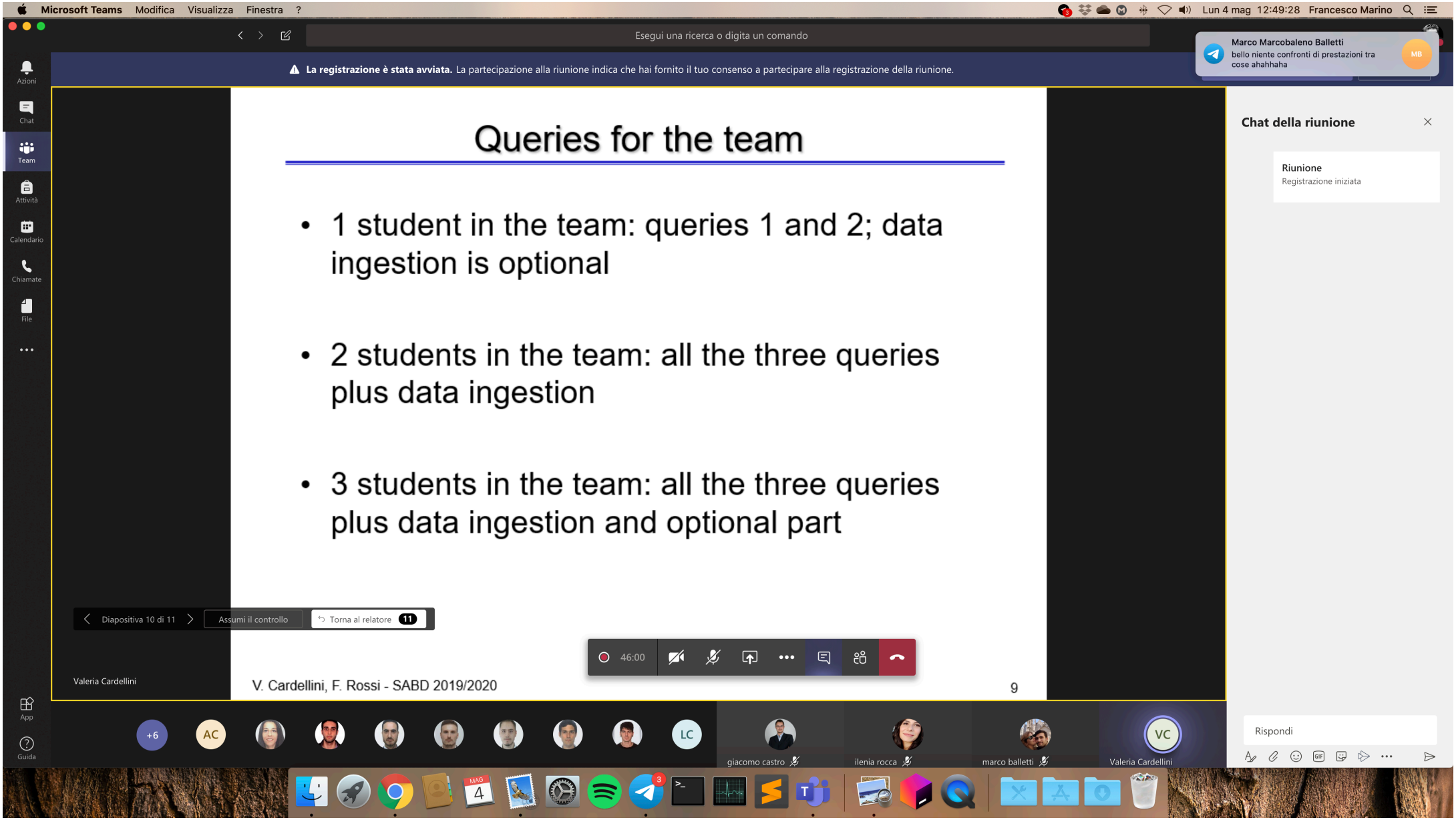

Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:29:41 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
App
Guida
Dataset 2: schema
• Input in CSV format
• One row per day, starting from January 22, 2020
• On each row:
– Province/State
– Country
– Latitude and longitude
– Daily columns with total number of confirmed cases
Diapositiva 5 di 11
Assumi il controllo
Torna al relatore 6
26:13
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
4
Chat della riunione
Riunione
Registrazione iniziata
+6
DN
AC
LC
luigi corsi
ilenia rocca
marco balletti
VC
Valeria Cardellini
Rispondi

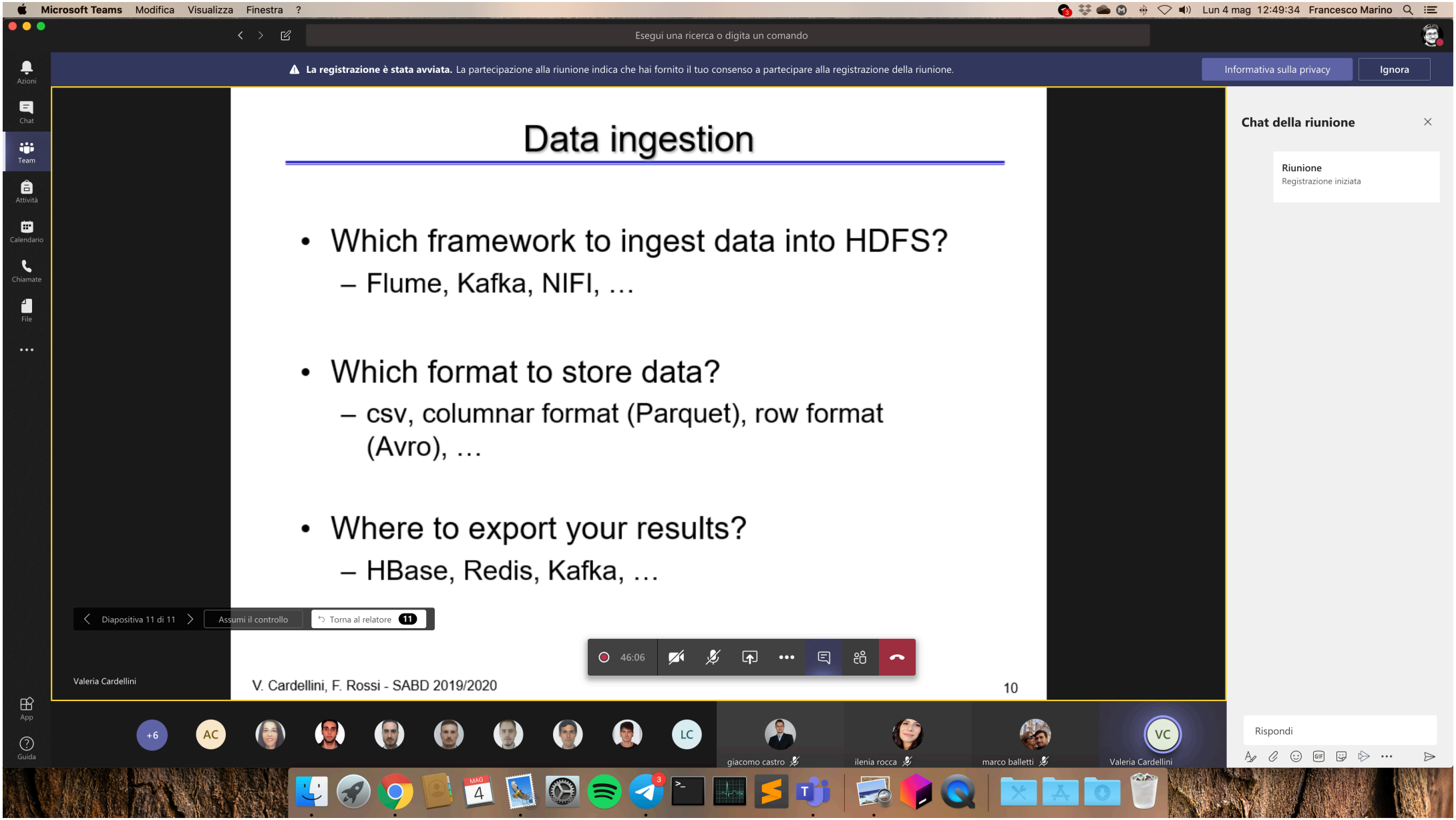

Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:29:49 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
App
Guida
Queries with Hadoop/Spark
• Use the Hadoop framework (and the MapReduce programming model) or alternatively the Spark framework to answer some queries on the dataset
• Include in your report/slides the queries' response time on your reference architecture
Query 1
Using dataset 1, for each week determine the average number of cured people and the average number of swabs
Diapositiva 6 di 11
Assumi il controllo
Torna al relatore 6
26:22
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
5
Chat della riunione
Riunione
Registrazione iniziata
Rispondi
+6
DN
AC
LC
luigi corsi
ilenia rocca
marco balletti
VC
Valeria Cardellini


Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:35:24 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
Queries with Hadoop/Spark
Query 2
Using dataset 2, for each continent, determine the average, standard deviation, minimum and maximum number of confirmed cases on a weekly basis
– Consider only states with at least 50 confirmed (if State is not indicated, use Country)
– Use trendline coefficient
– Continent must be identified
Diapositiva 7 di 11
Assumi il controllo
Torna al relatore 7
31:56
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
6
Chat della riunione
Riunione
Registrazione iniziata
Rispondi
+6
AC
LC
giacomo castro
ilenia rocca
marco balletti
VC
Valeria Cardellini
App
Guida


Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:45:25 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
Queries with Hadoop/Spark
Query 3
Using dataset 2 and the top-50 states identified using the trendline coefficient, use K-means clustering algorithm (with K=5) to identify the states and nations that belong to each cluster
– Compare the performance of a naïve implementation of K-means with that provided by Spark MLib or Apache Mahout
Diapositiva 8 di 11
Assumi il controllo
Torna al relatore 9
41:58
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
7
Chat della riunione
Riunione
Registrazione iniziata
Rispondi
+6
AC
LC
giacomo castro
ilenia rocca
marco balletti
VC
Valeria Cardellini
App
Guida


Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:46:35 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
App
Guida
Optional part
• Compulsory for team composed of 3 students
• Use either Hive (or Pig) or Spark SQL to address the Queries 1 and 2
• Include in the report the query times using a higher level framework on your reference architecture and compare them to those achieved by your pure Hadoop/Spark-based solution
Diapositiva 9 di 11
Assumi il controllo
Torna al relatore 10
43:08
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
8
Chat della riunione
Riunione
Registrazione iniziata
Rispondi
+6
AC
LC
giacomo castro
ilenia rocca
marco balletti
VC
Valeria Cardellini


Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:49:28 Francesco Marino
Esegui una ricerca o digita un comando
Marco Marcobaleno Balletti
bello niente confronti di prestazioni tra cose ahahhaha
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
Queries for the team
• 1 student in the team: queries 1 and 2; data ingestion is optional
• 2 students in the team: all the three queries plus data ingestion
• 3 students in the team: all the three queries plus data ingestion and optional part
Diapositiva 10 di 11
Assumi il controllo
Torna al relatore 11
46:00
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
9
Chat della riunione
Riunione
Registrazione iniziata
+6
AC
LC
giacomo castro
ilenia rocca
marco balletti
VC
Valeria Cardellini
Rispondi
App
Guida


Microsoft Teams Modifica Visualizza Finestra ?
Lun 4 mag 12:49:34 Francesco Marino
Esegui una ricerca o digita un comando
La registrazione è stata avviata. La partecipazione alla riunione indica che hai fornito il tuo consenso a partecipare alla registrazione della riunione.
Informativa sulla privacy
Ignora
Azioni
Chat
Team
Attività
Calendario
Chiamate
File
App
Guida
Data ingestion
• Which framework to ingest data into HDFS?
– Flume, Kafka, NIFI, …
• Which format to store data?
– csv, columnar format (Parquet), row format (Avro), …
• Where to export your results?
– HBase, Redis, Kafka, …
Diapositiva 11 di 11
Assumi il controllo
Torna al relatore
11
46:06
Valeria Cardellini
V. Cardellini, F. Rossi - SABD 2019/2020
10
Chat della riunione
Riunione
Registrazione iniziata
Rispondi
+6
AC
LC
giacomo castro
ilenia rocca
marco balletti
VC
Valeria Cardellini