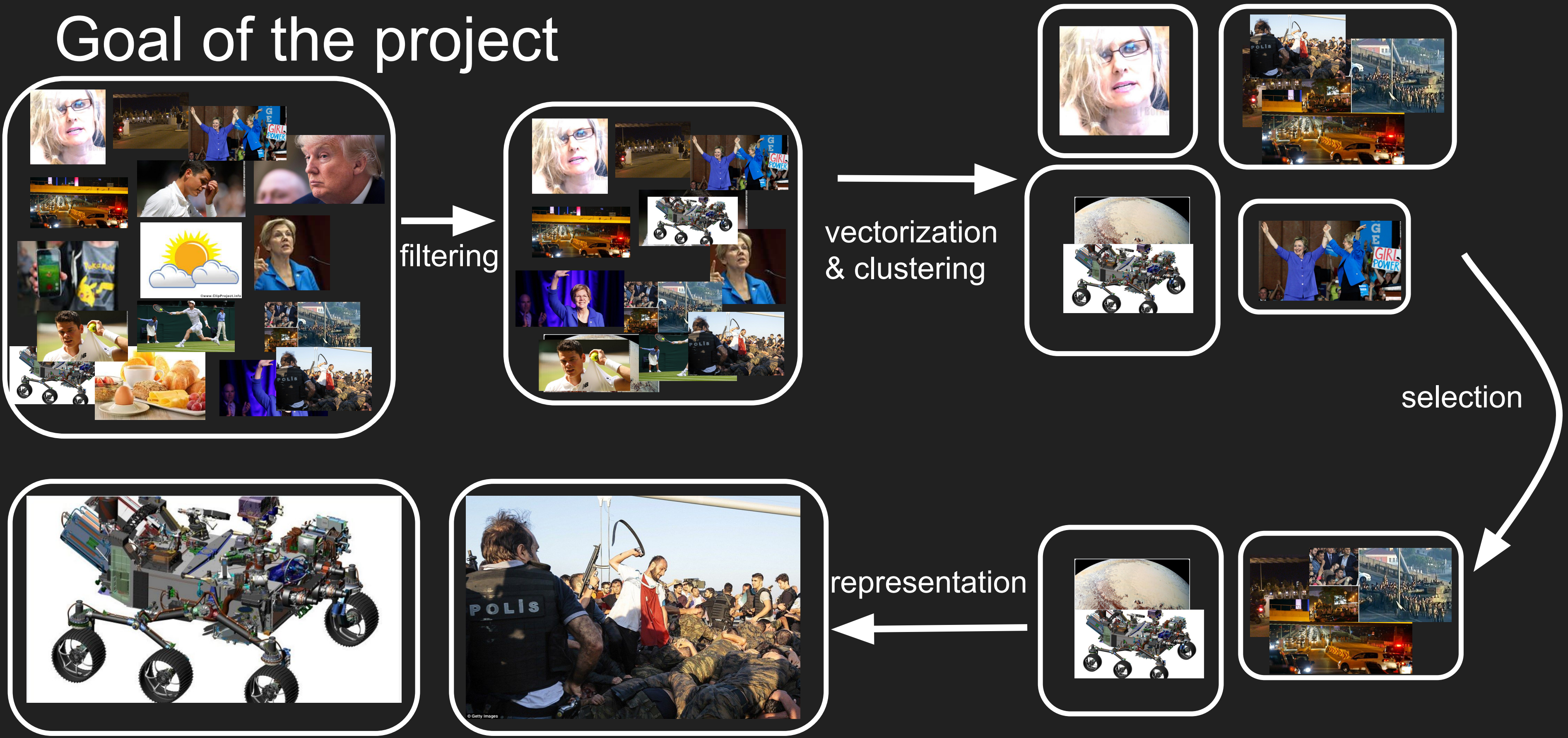
Twitter News Feed

Final presentation

Goal of the project



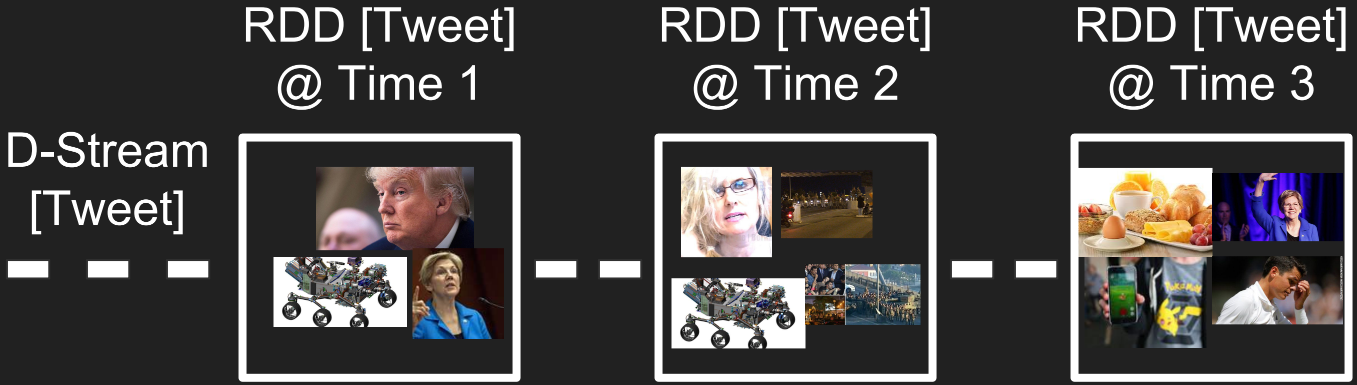
filtering

vectorization & clustering

selection

representation

Apache Spark Streaming



RDD [Tweet] @ Time 1

RDD [Tweet] @ Time 2

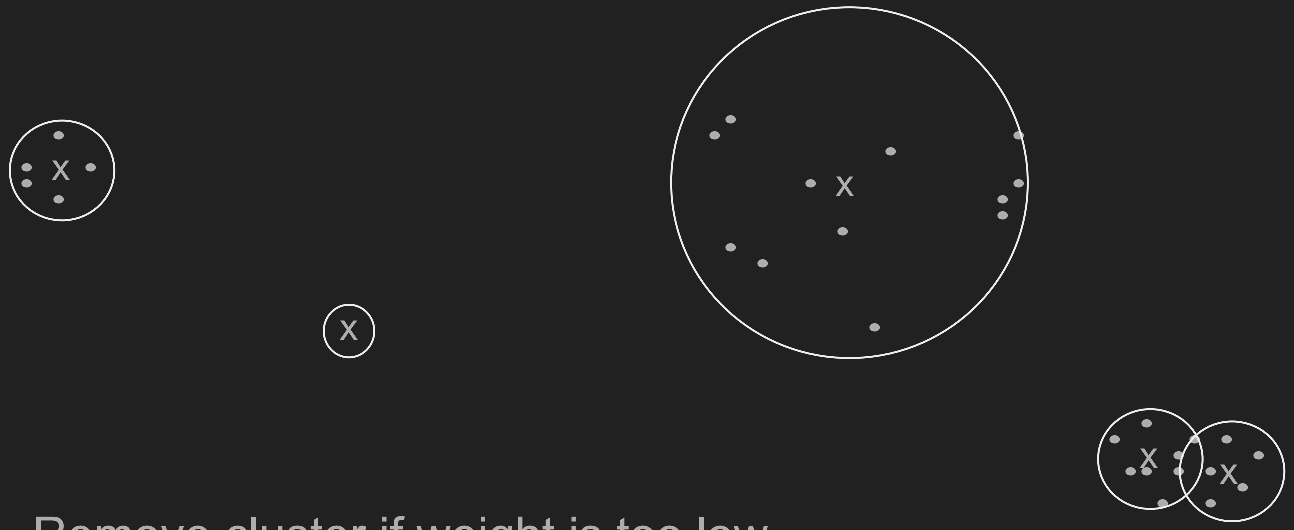
RDD [Tweet] @ Time 3

D-Stream [Tweet]

Clustering: Streaming k-Means

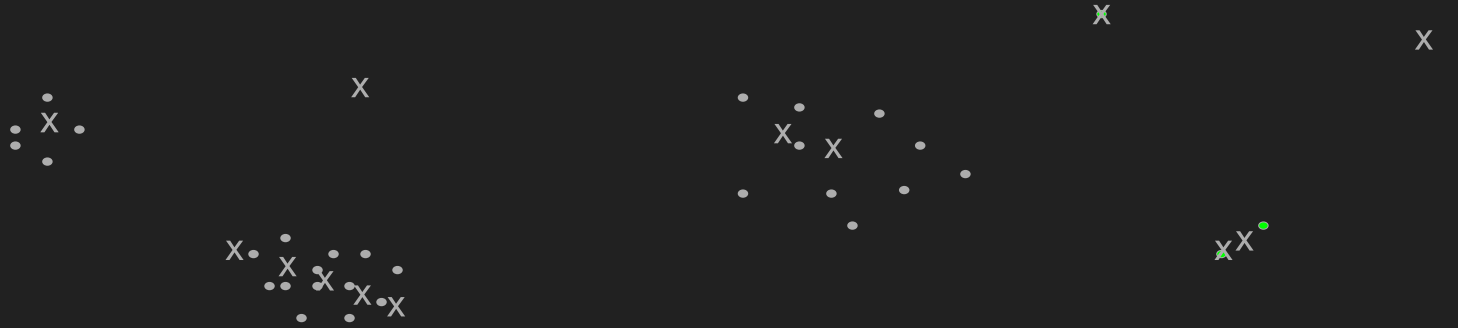
- Continuous clustering with batches
- For every batch, new centroids are calculated
- Centroids and cluster size are saved
- Cluster quality problems due to:
    - Randomly generated initial centroids
    - Static k not based on cluster results

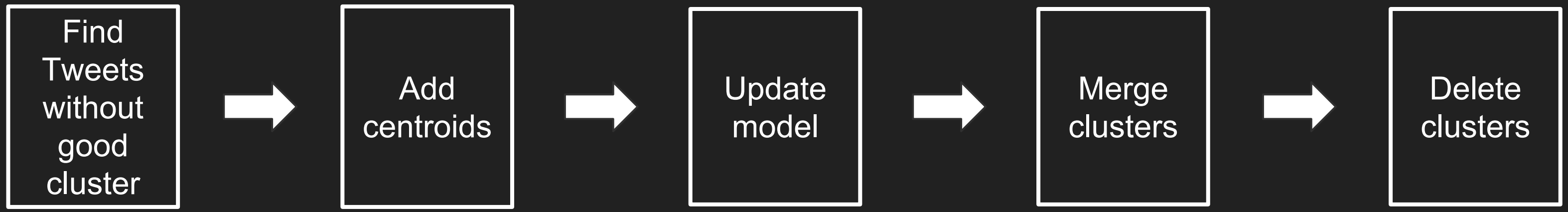
Dynamic k-Means (1/3)



- Remove cluster if weight is too low
- Add a cluster if tweet is too far away from closest centroid
- Merge clusters which are close to each other

Dynamic k-Means (2/3)



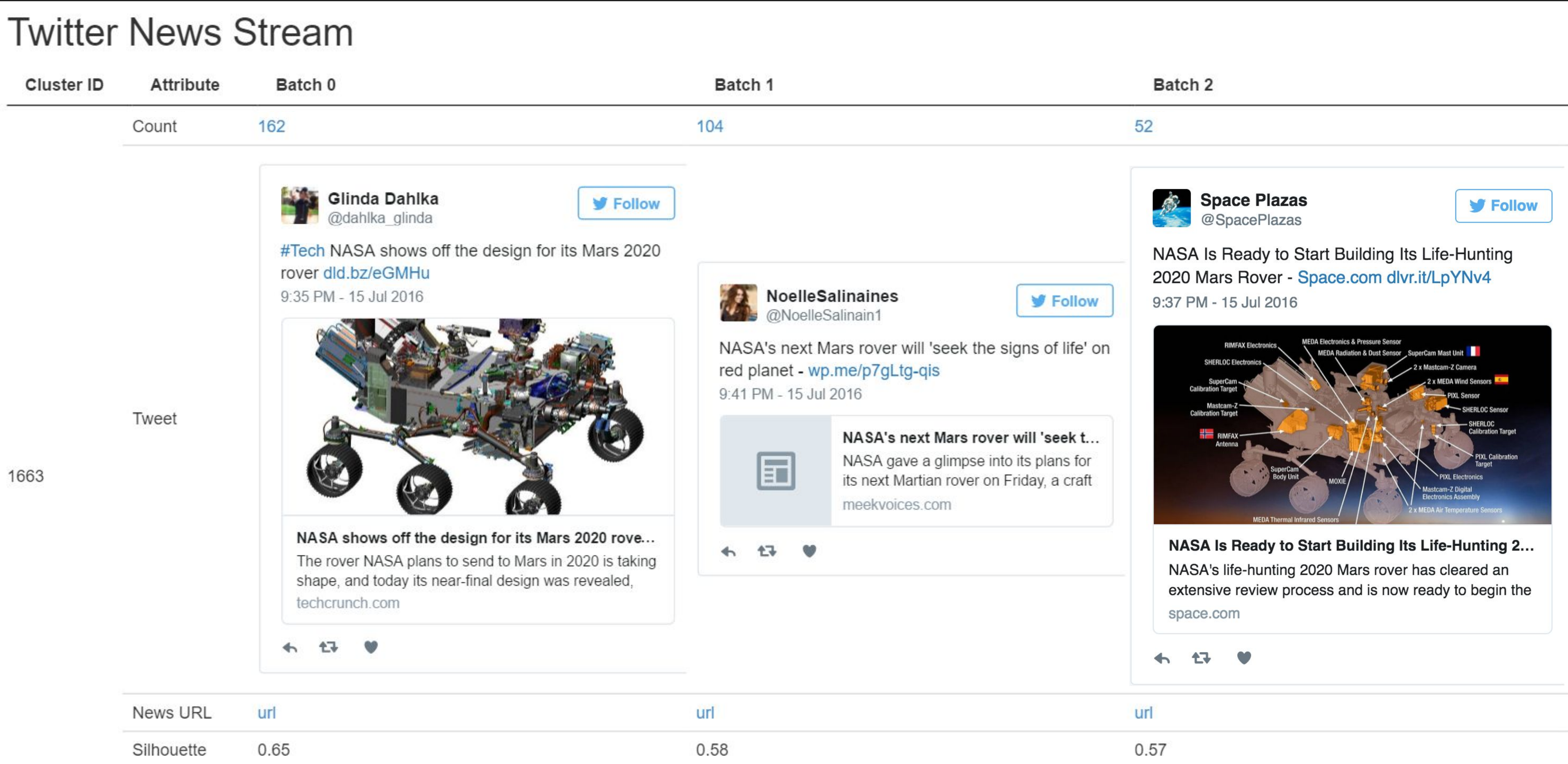


Find Tweets without good cluster → Add centroids → Update model → Merge clusters → Delete clusters

Dynamic k-Means (3/3)

- Advantages:
  - No need to find a *k* that works “okay” for arbitrary input
  - No need to choose initial centroids
  - Easy to parallelize

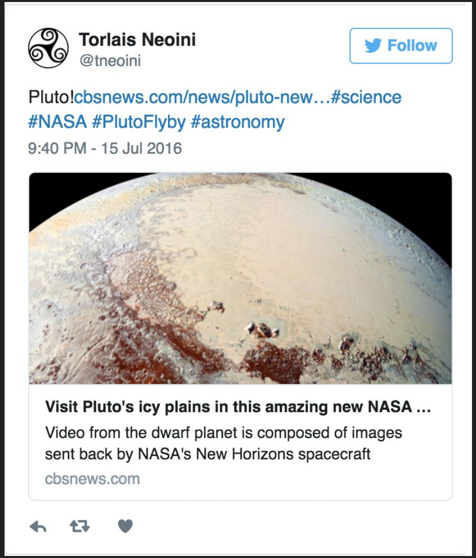
- Disadvantages:
  - More computationally expensive
  - New thresholds to define



Twitter News Stream

Cluster ID	Attribute	Batch 0	Batch 1	Batch 2
1663	Count	162	104	52
	Tweet	Glinda Dahlka @dahlka_glinda Follow #Tech NASA shows off the design for its Mars 2020 rover dld.bz/eGMHu 9:35 PM - 15 Jul 2016 NASA shows off the design for its Mars 2020 rove... The rover NASA plans to send to Mars in 2020 is taking shape, and today its near-final design was revealed, techcrunch.com	NoelleSalinaines @NoelleSalinain1 Follow NASA's next Mars rover will 'seek the signs of life' on red planet - wp.me/p7gLtg-qis 9:41 PM - 15 Jul 2016 NASA's next Mars rover will 'seek t... NASA gave a glimpse into its plans for its next Martian rover on Friday, a craft meekvoices.com	Space Plazas @SpacePlazas Follow NASA Is Ready to Start Building Its Life-Hunting 2020 Mars Rover - Space.com dlvr.it/LpYNv4 9:37 PM - 15 Jul 2016 NASA Is Ready to Start Building Its Life-Hunting 2... NASA's life-hunting 2020 Mars rover has cleared an extensive review process and is now ready to begin the space.com
	News URL	url	url	url
	Silhouette	0.65	0.58	0.57

RIMFAX Electronics
MEDA Electronics & Pressure Sensor
MEDA Radiation & Dust Sensor
SuperCam Mast Unit
SHERLOC Electronics
2 x Mastcam-Z Camera
SuperCam Calibration Target
2 x MEDA Wind Sensors
PIXL Sensor
Mastcam-Z Calibration Target
SHERLOC Sensor
SHERLOC Calibration Target
RIMFAX Antenna
PIXL Calibration Target
SuperCam Body Unit
MOXIE
PIXL Electronics
Mastcam-Z Digital Electronics Assembly
2 x MEDA Air Temperature Sensors
MEDA Thermal Infrared Sensors



Torlais Neoini
@tneoini

Follow

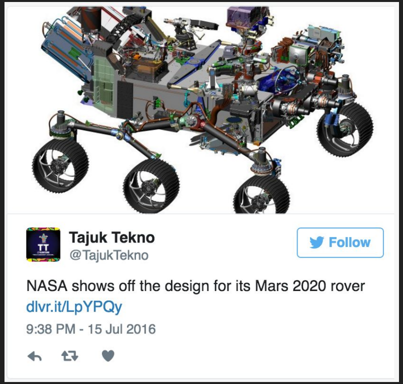
Pluto!cbsnews.com/news/pluto-new…#science #NASA #PlutoFlyby #astronomy

9:40 PM - 15 Jul 2016

Visit Pluto's icy plains in this amazing new NASA …

Video from the dwarf planet is composed of images sent back by NASA's New Horizons spacecraft

cbsnews.com

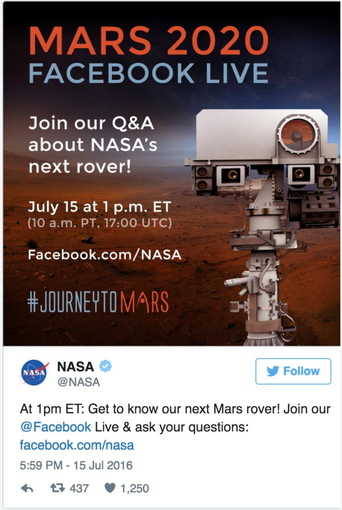


Tajuk Tekno
@TajukTekno

Follow

NASA shows off the design for its Mars 2020 rover dlvr.it/LpYPQy

9:38 PM - 15 Jul 2016



MARS 2020
FACEBOOK LIVE

Join our Q&A about NASA's next rover!

July 15 at 1 p.m. ET
(10 a.m. PT, 17:00 UTC)

Facebook.com/NASA

#JOURNEYTOMARS

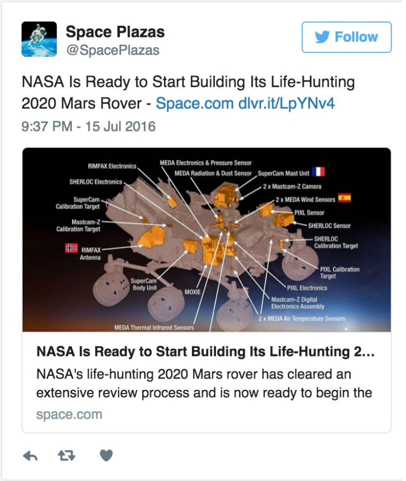
NASA
@NASA

Follow

At 1pm ET: Get to know our next Mars rover! Join our @Facebook Live & ask your questions: facebook.com/nasa

5:59 PM - 15 Jul 2016

437 1,250



Space Plazas
@SpacePlazas

Follow

NASA Is Ready to Start Building Its Life-Hunting 2020 Mars Rover - Space.com dlvr.it/LpYNv4

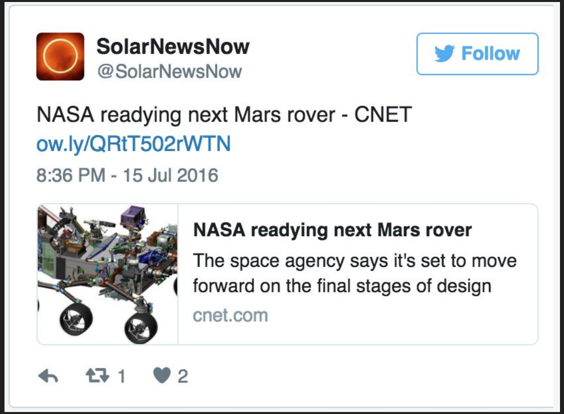
9:37 PM - 15 Jul 2016

RIMFAX Electronics
MEDA Electronics & Pressure Sensor
MEDA Radiation & Dust Sensor
SuperCam Mast Unit
SHERLOC Electronics
2 x Mastcam-Z Camera
SuperCam Calibration Target
2 x MEDA Wind Sensors
PIXL Sensor
Mastcam-Z Calibration Target
SHERLOC Sensor
SHERLOC Calibration Target
RIMFAX Antenna
PIXL Calibration Target
SuperCam Body Unit
MOXIE
PIXL Electronics
Mastcam-Z Digital Electronics Assembly
2 x MEDA Air Temperature Sensors
MEDA Thermal Infrared Sensors

NASA Is Ready to Start Building Its Life-Hunting 2…

NASA's life-hunting 2020 Mars rover has cleared an extensive review process and is now ready to begin the

space.com



SolarNewsNow
@SolarNewsNow

Follow

NASA readying next Mars rover - CNET ow.ly/QRtT502rWTN

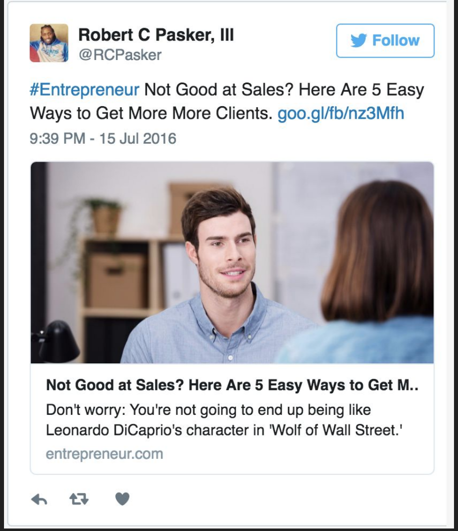
8:36 PM - 15 Jul 2016

NASA readying next Mars rover

The space agency says it's set to move forward on the final stages of design

cnet.com

1 2



Robert C Pasker, III
@RCPasker

Follow

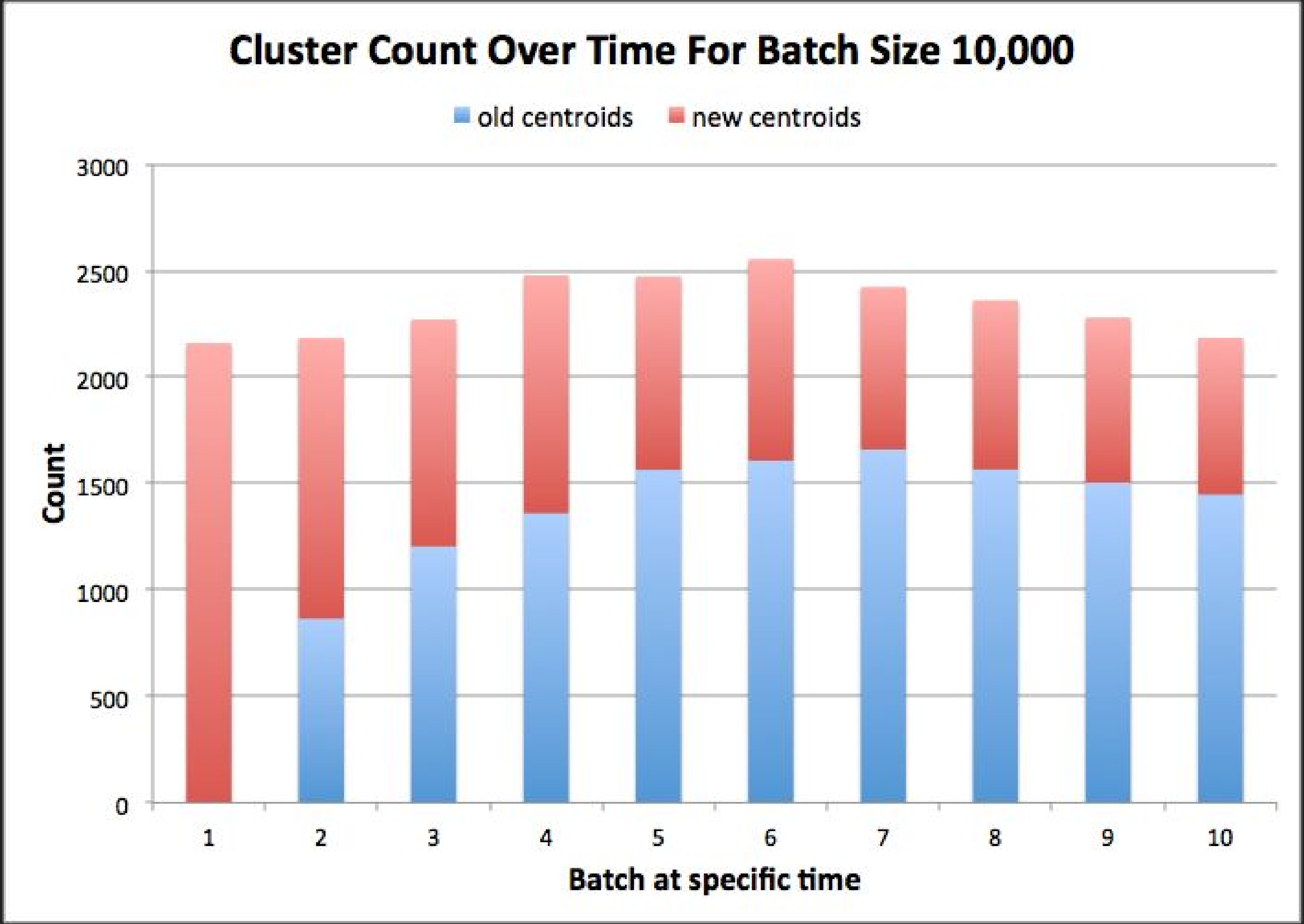
#Entrepreneur Not Good at Sales? Here Are 5 Easy Ways to Get More More Clients. goo.gl/fb/nz3Mfh

9:39 PM - 15 Jul 2016

Not Good at Sales? Here Are 5 Easy Ways to Get M..

Don't worry: You're not going to end up being like Leonardo DiCaprio's character in 'Wolf of Wall Street.'

entrepreneur.com


Cluster Count Over Time For Batch Size 10,000
🟦 old centroids
🟥 new centroids
3000
2500
2000
1500
1000
500
0
Count
1
2
3
4
5
6
7
8
9
10
Batch at specific time

Performance Evaluation

Hardware: HPI-Cluster

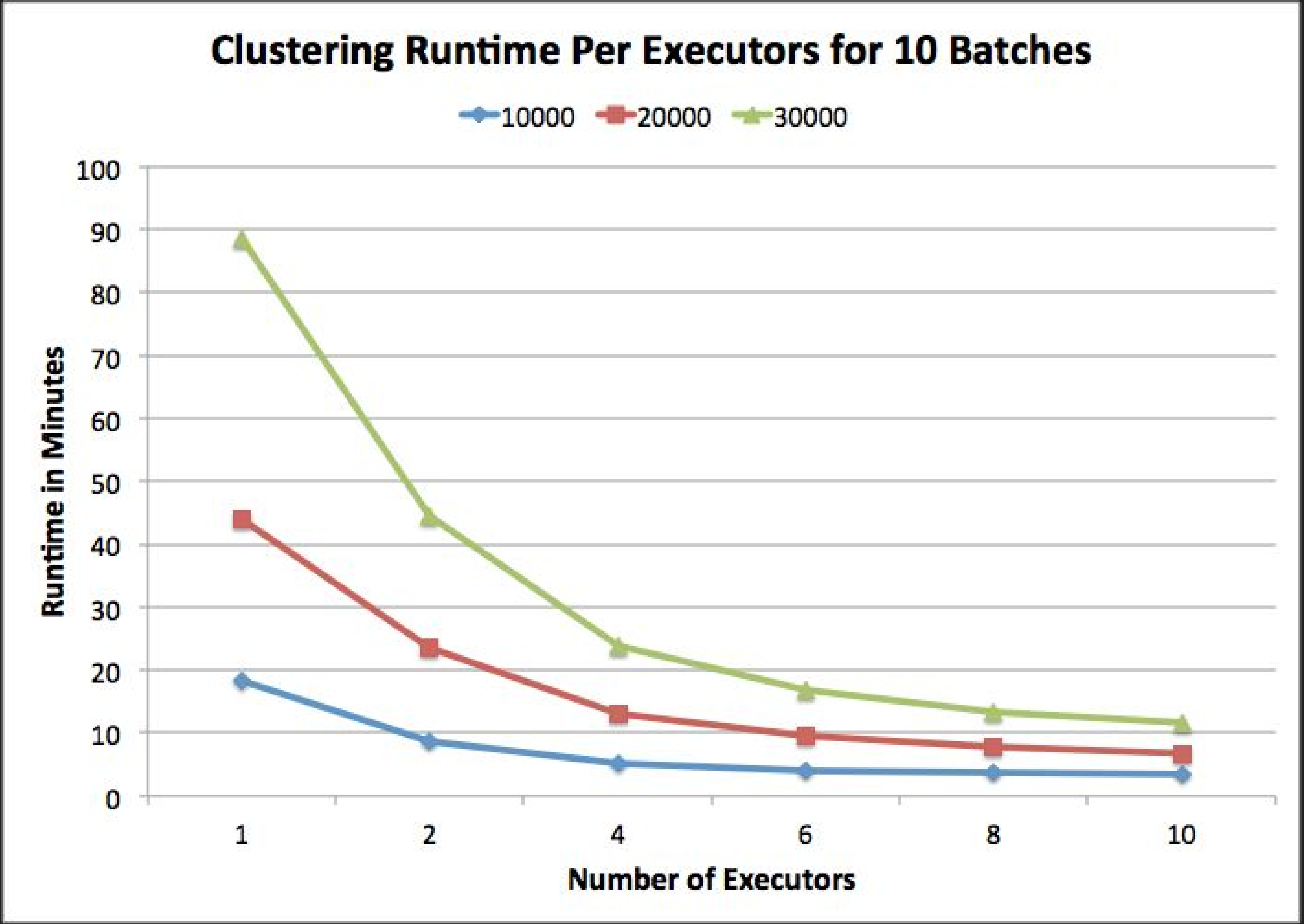
- 10 Server à 2 Cores x 2.67 GHz CPU, 4 GB RAM

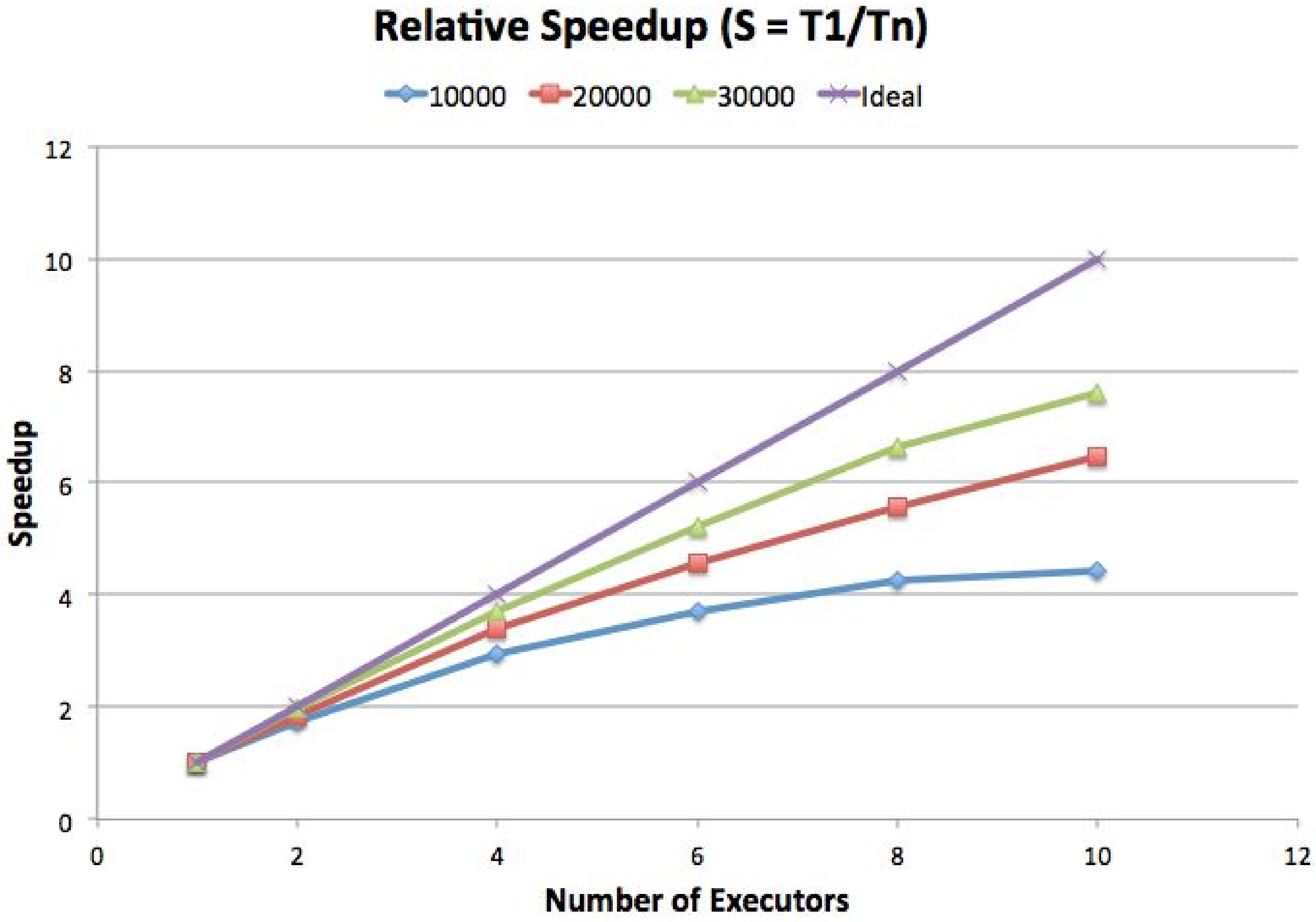
Input

- 300.000 Tweets 15.07.16 19:30 - 23:30 Uhr

Execution

- 3 times per setting
- 1 Core per Executor

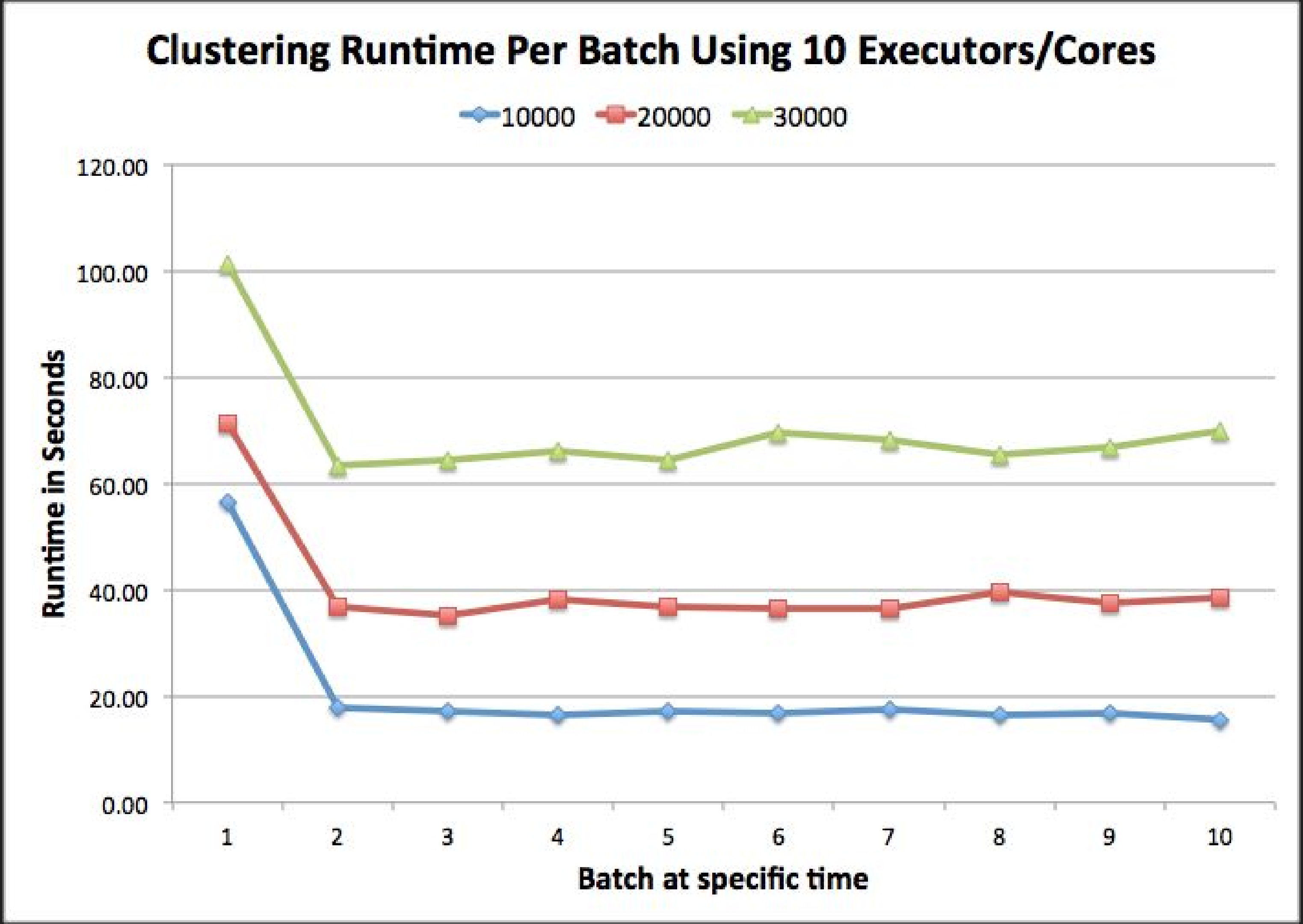

Clustering Runtime Per Executors for 10 Batches
10000
20000
30000
Runtime in Minutes
100
90
80
70
60
50
40
30
20
10
0
1
2
4
6
8
10
Number of Executors


Relative Speedup (S = T1/Tn)
10000
20000
30000
Ideal
Speedup
0
2
4
6
8
10
12
Number of Executors
0
2
4
6
8
10
12

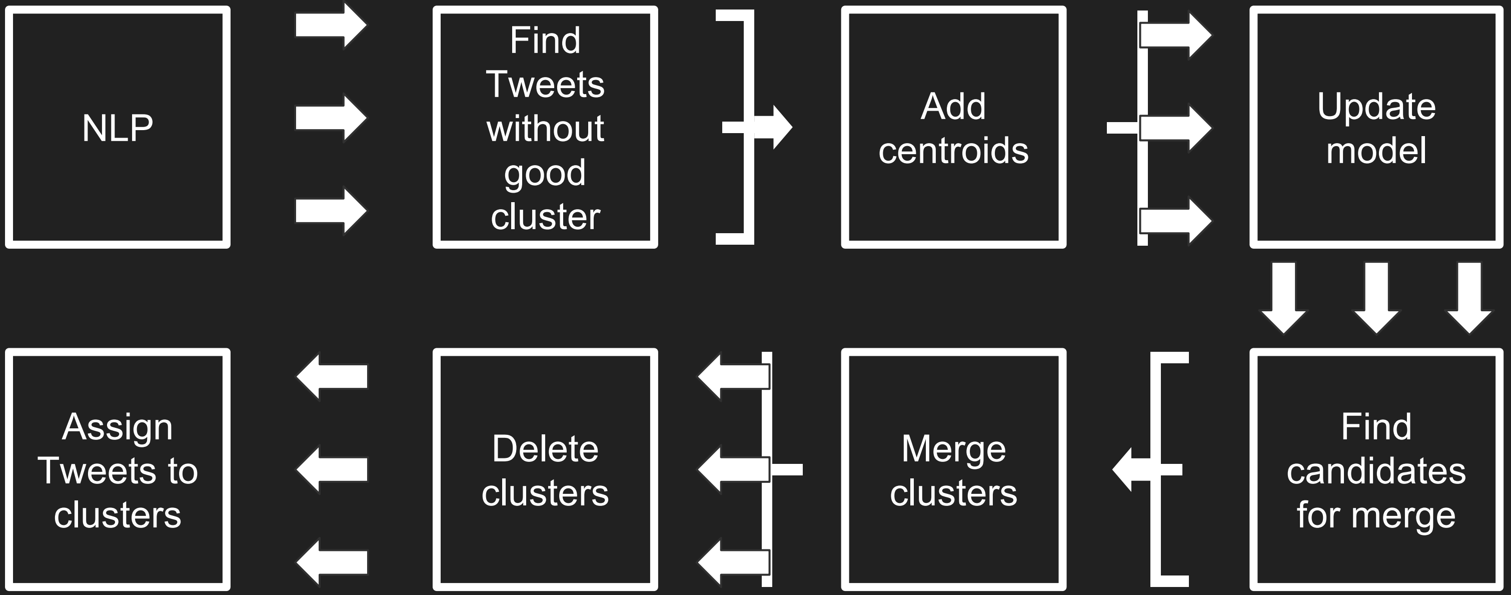
Conclusion

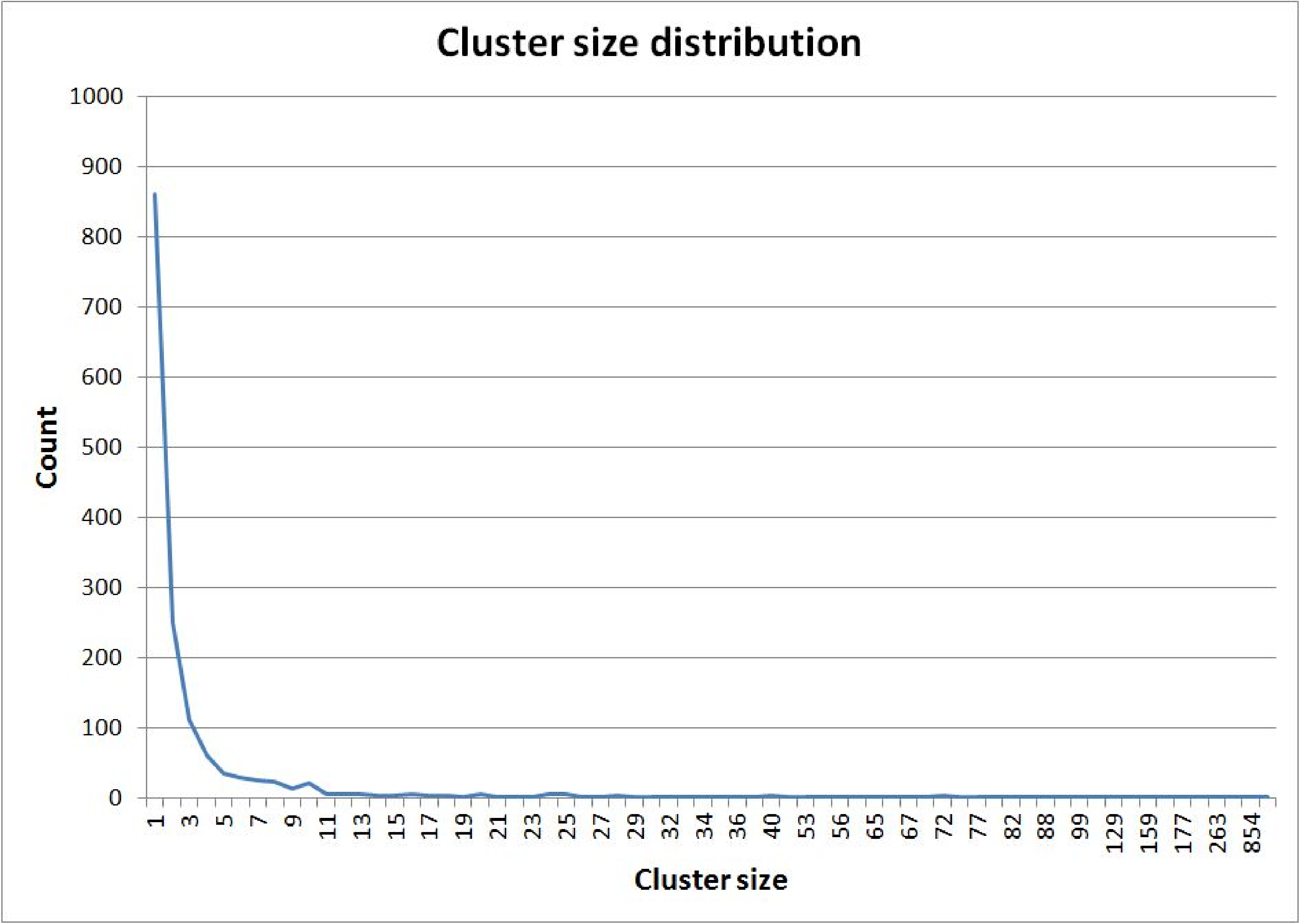
- We can process and cluster tweets live
- Getting high quality results for clustering is very tough

Backup


Clustering Runtime Per Batch Using 10 Executors/Cores
10000
20000
30000
Runtime in Seconds
120.00
100.00
80.00
60.00
40.00
20.00
0.00
1
2
3
4
5
6
7
8
9
10
Batch at specific time

Dynamic k-Means: Parallelization


NLP
Find Tweets without good cluster
Add centroids
Update model
Find candidates for merge
Merge clusters
Delete clusters
Assign Tweets to clusters


Cluster size distribution
Count
1000
900
800
700
600
500
400
300
200
100
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
32
34
36
40
53
56
65
67
72
77
82
88
99
129
159
177
263
854
Cluster size