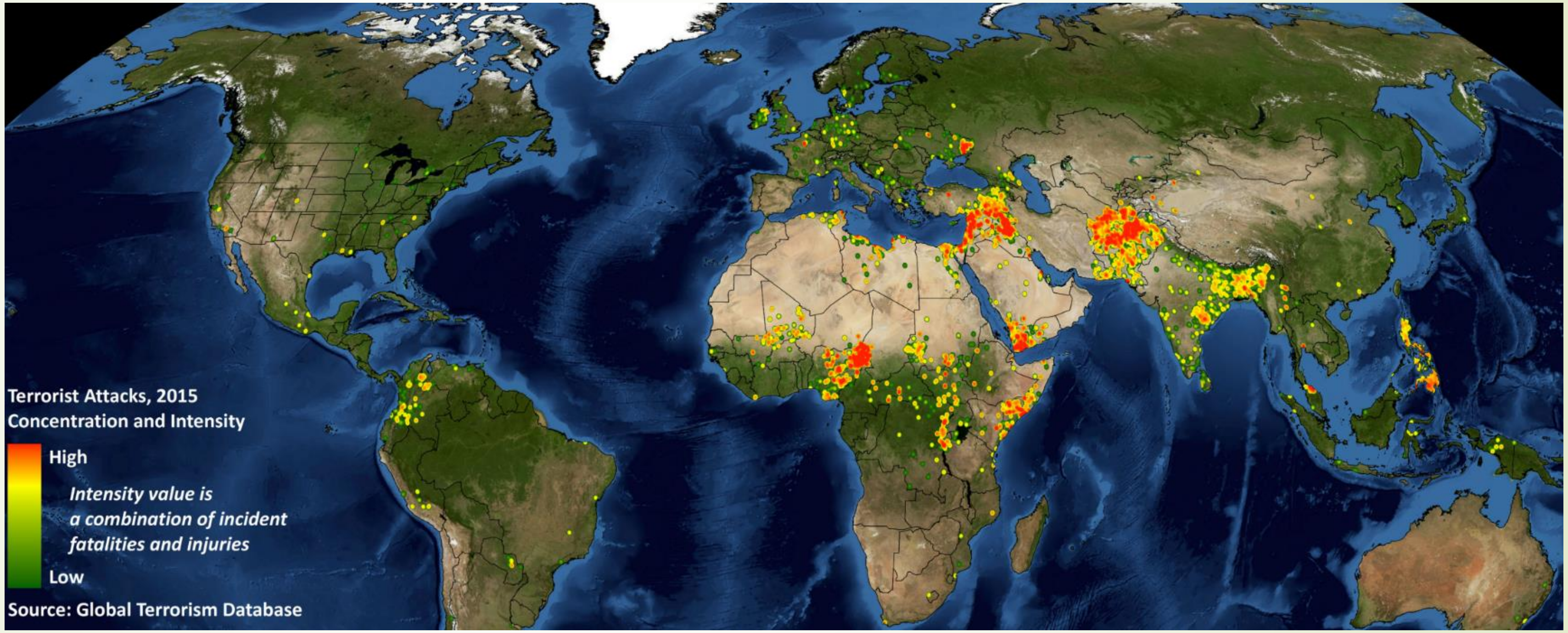
Dangerous events forecasting

Sergei Dmitriev,
Student Springboard

27 September 2016

Description of the problem

- Terrorist attacks, conflicts, mass violence have a great negative impact on society.
- These events spread around the world.


Terrorist Attacks, 2015
Concentration and Intensity
High
Intensity value is a combination of incident fatalities and injuries
Low

Source: Global Terrorism Database

Figure 1

Can we predict dangerous events?

Datasets

- Datasets with description of events from around the world. Source: GDELT event files :
  - about 1200 files for period 2013-04-01 – 2016-09-03
  - total amount of records is about 200 000 000
- Each record includes:
  - event represented as «Actor1 performed an action upon Actor2»
  - location of an event
  - AvgTone and other characteristic
  - hyperlink
  - a date an event was added to a database
- Example of record:

454094223,20140801,201408,2014,2014.5781,CAN,CANADIAN,CAN,,,,,,,,,,,,,,,,,,0,111,111,11,3,-2.0,36,1,36,4.6908315565032,1,Australia,AS,AS,-27.0,133.0,AS,0,,,,,,,1,Australia,AS,AS,-27.0,133.0,AS,20150801,http://www.jewellermagazine.com/Article.aspx?id=5184&h=Showcase-Jewellers-shifts-focus&vc=1032

Datasets (2)

- Event characteristic:
  - EventCode
  - QuadClass
  - GoldsteinScale
  - NumMentions
  - AvgTone - the score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This score, calculated automatically, can be used for measuring of the importance of an event. For example, an event like terrorist attack has a AvgTone less than -15
- Dangerous event:
  - EventCode, QuadClass, NumMentions – any
  - GoldsteinScale = -10
  - AvgTone < -15

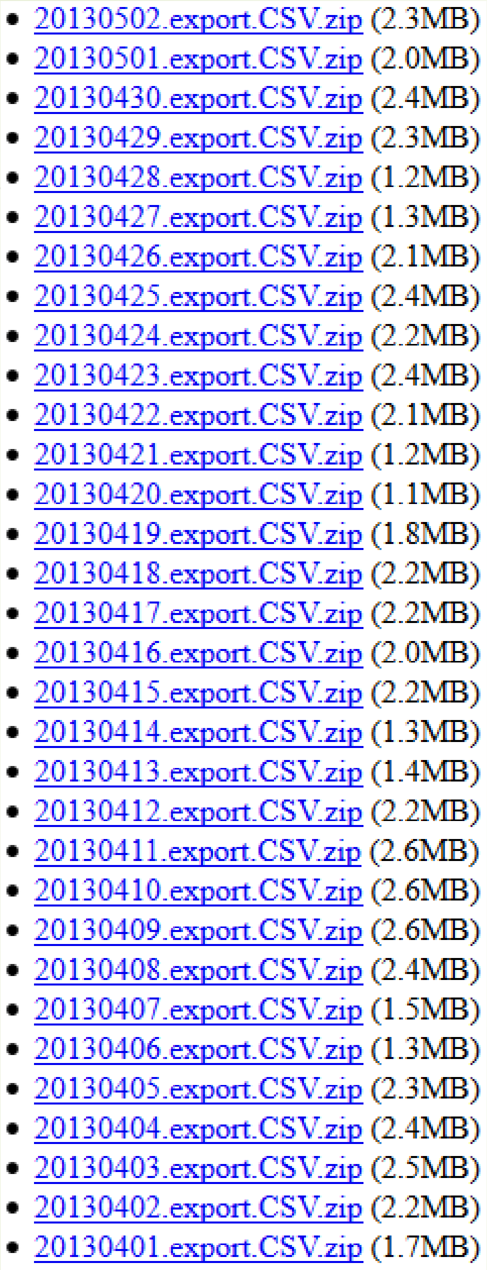
Approach to the problem

- France and time period 2013-04-01 – 2016-09-03
- Building time series from AvgTone
- There are many events during a certain day
  - If there are no a dangerous events during a day, tone is average of tones these events.
  - If there are dangerous events during a day, tone is equal tone of dangerous event with minimum of tone.

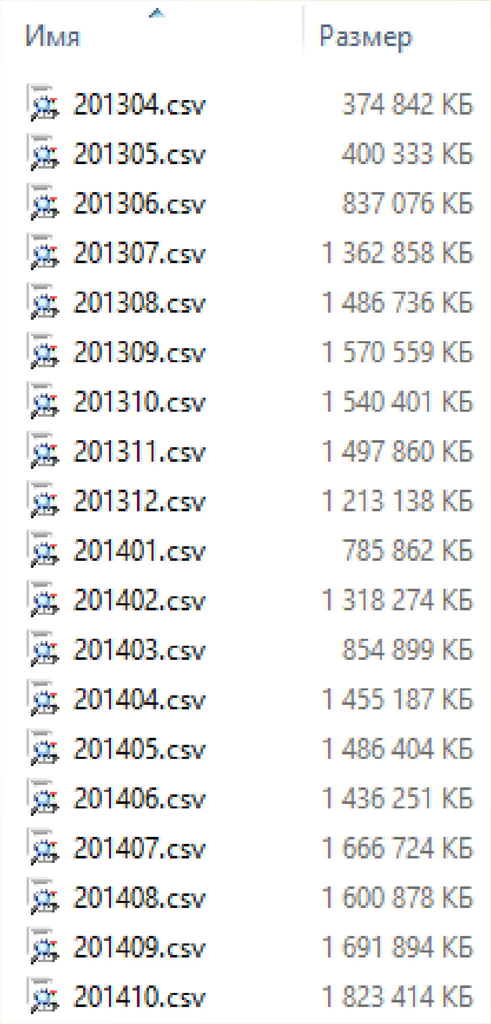
Approach to the problem (2)

- Loading and merging datasets
- Cleaning and transforming data
- Visualizing data
- Applying neural network

Loading and merging datasets



- 20130502.export.CSV.zip (2.3MB)
- 20130501.export.CSV.zip (2.0MB)
- 20130430.export.CSV.zip (2.4MB)
- 20130429.export.CSV.zip (2.3MB)
- 20130428.export.CSV.zip (1.2MB)
- 20130427.export.CSV.zip (1.3MB)
- 20130426.export.CSV.zip (2.1MB)
- 20130425.export.CSV.zip (2.4MB)
- 20130424.export.CSV.zip (2.2MB)
- 20130423.export.CSV.zip (2.4MB)
- 20130422.export.CSV.zip (2.1MB)
- 20130421.export.CSV.zip (1.2MB)
- 20130420.export.CSV.zip (1.1MB)
- 20130419.export.CSV.zip (1.8MB)
- 20130418.export.CSV.zip (2.2MB)
- 20130417.export.CSV.zip (2.2MB)
- 20130416.export.CSV.zip (2.0MB)
- 20130415.export.CSV.zip (2.2MB)
- 20130414.export.CSV.zip (1.3MB)
- 20130413.export.CSV.zip (1.4MB)
- 20130412.export.CSV.zip (2.2MB)
- 20130411.export.CSV.zip (2.6MB)
- 20130410.export.CSV.zip (2.6MB)
- 20130409.export.CSV.zip (2.6MB)
- 20130408.export.CSV.zip (2.4MB)
- 20130407.export.CSV.zip (1.5MB)
- 20130406.export.CSV.zip (1.3MB)
- 20130405.export.CSV.zip (2.3MB)
- 20130404.export.CSV.zip (2.4MB)
- 20130403.export.CSV.zip (2.5MB)
- 20130402.export.CSV.zip (2.2MB)
- 20130401.export.CSV.zip (1.7MB)



Имя	Размер
201304.csv	374 842 КБ
201305.csv	400 333 КБ
201306.csv	837 076 КБ
201307.csv	1 362 858 КБ
201308.csv	1 486 736 КБ
201309.csv	1 570 559 КБ
201310.csv	1 540 401 КБ
201311.csv	1 497 860 КБ
201312.csv	1 213 138 КБ
201401.csv	785 862 КБ
201402.csv	1 318 274 КБ
201403.csv	854 899 КБ
201404.csv	1 455 187 КБ
201405.csv	1 486 404 КБ
201406.csv	1 436 251 КБ
201407.csv	1 666 724 КБ
201408.csv	1 600 878 КБ
201409.csv	1 691 894 КБ
201410.csv	1 823 414 КБ

Cleaning and transforming data

- Delete duplicates with the same URL
- Delete duplicates with the same characteristic:
  - FractionDateN'
  - 'QuadClass'
  - 'GoldsteinScale'
  - 'AvgToneN'
  - 'Actor1CountryCode'
  - 'Actor2CountryCode'

Visualizing data

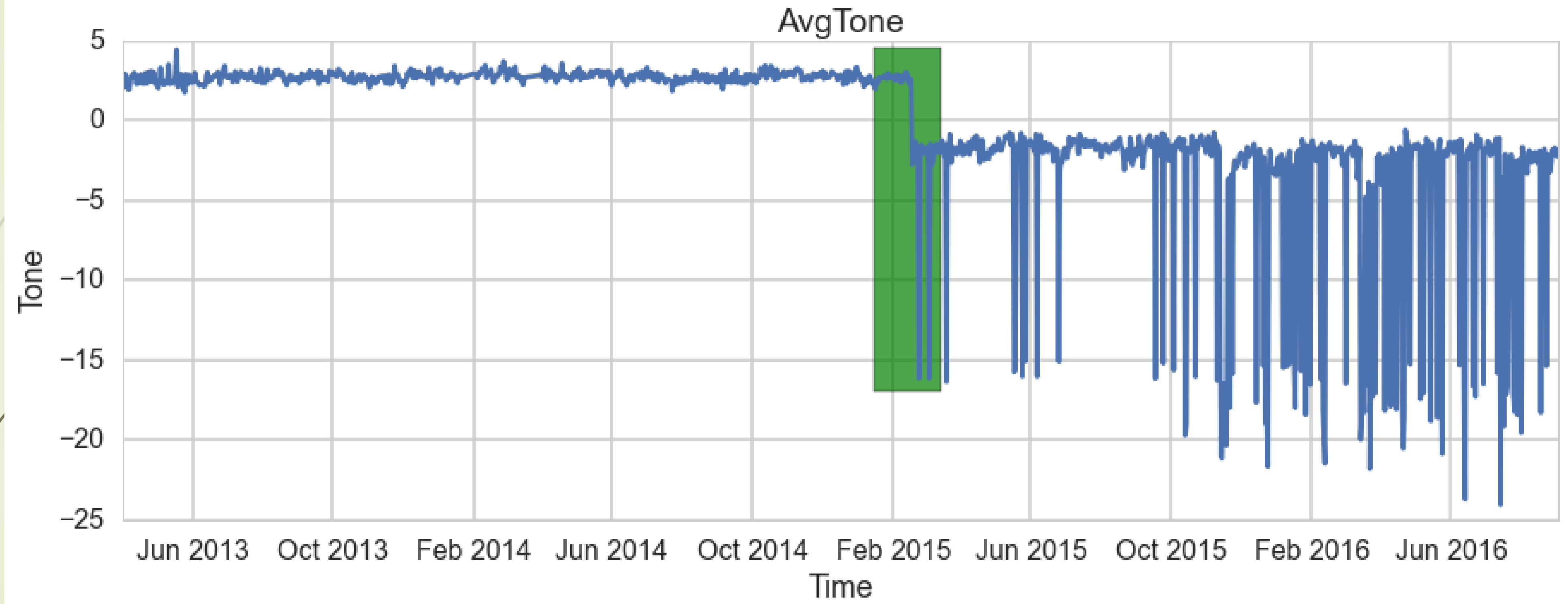


Figure 2. France 2013-2016

Visualizing data (2)

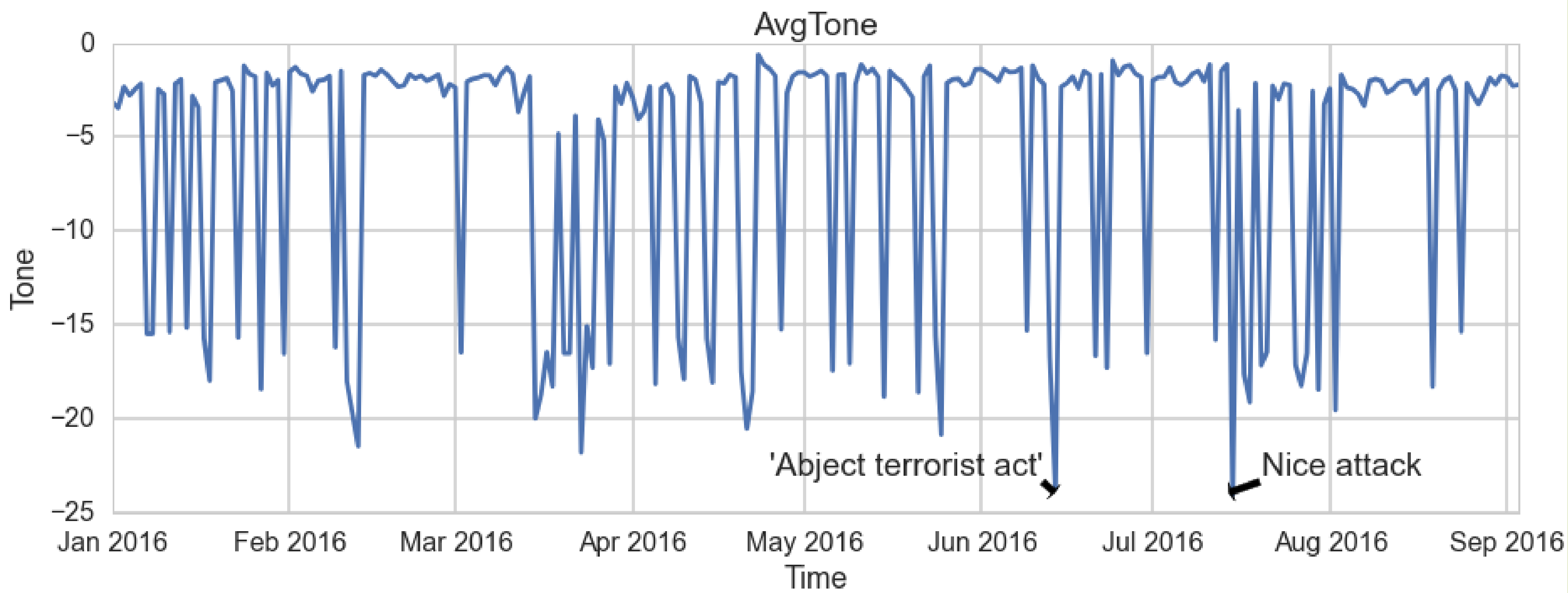


Figure 3. France 2016

Testing stationarity

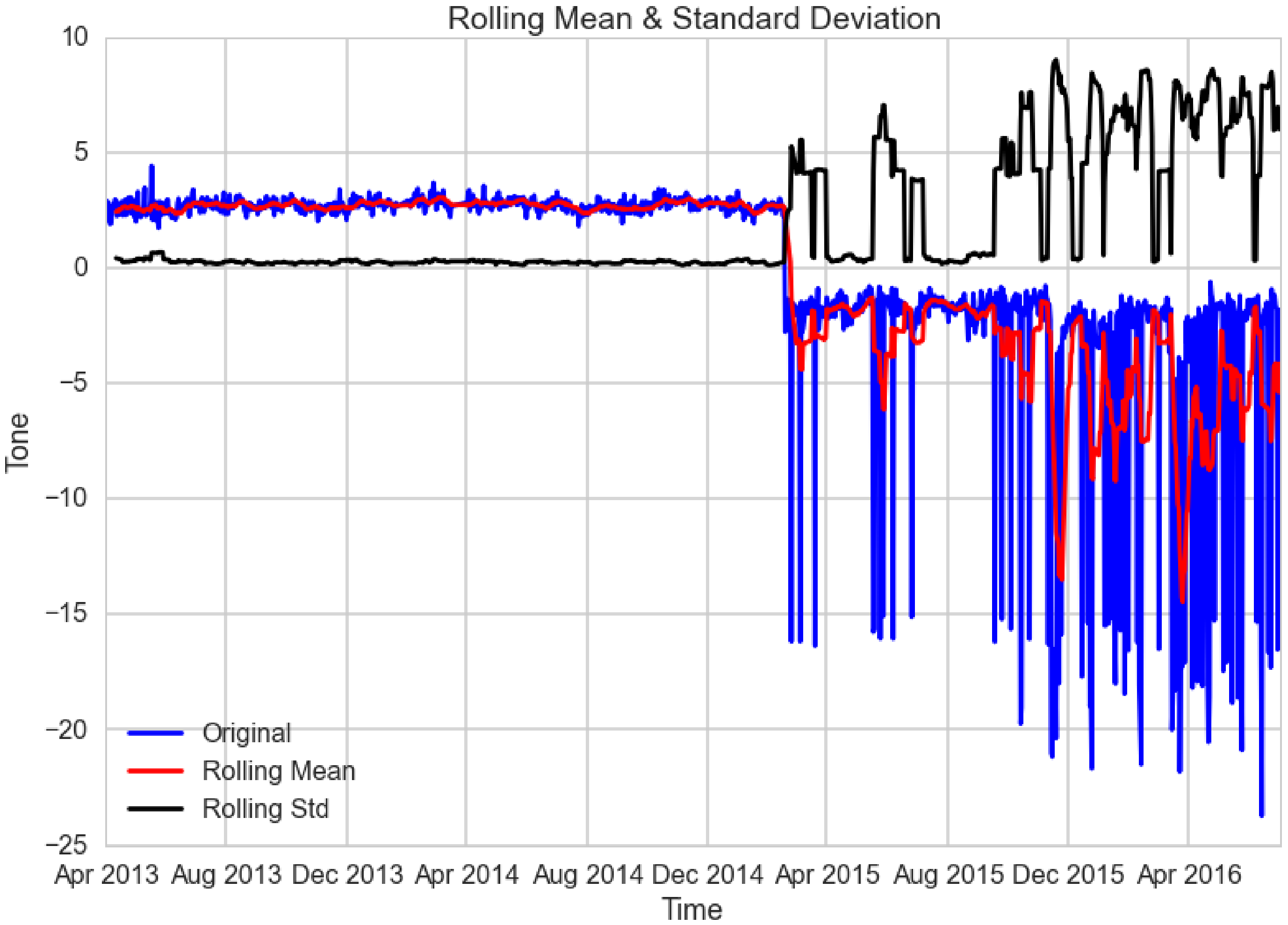


Figure 4. Testing stationarity

Applying neural network

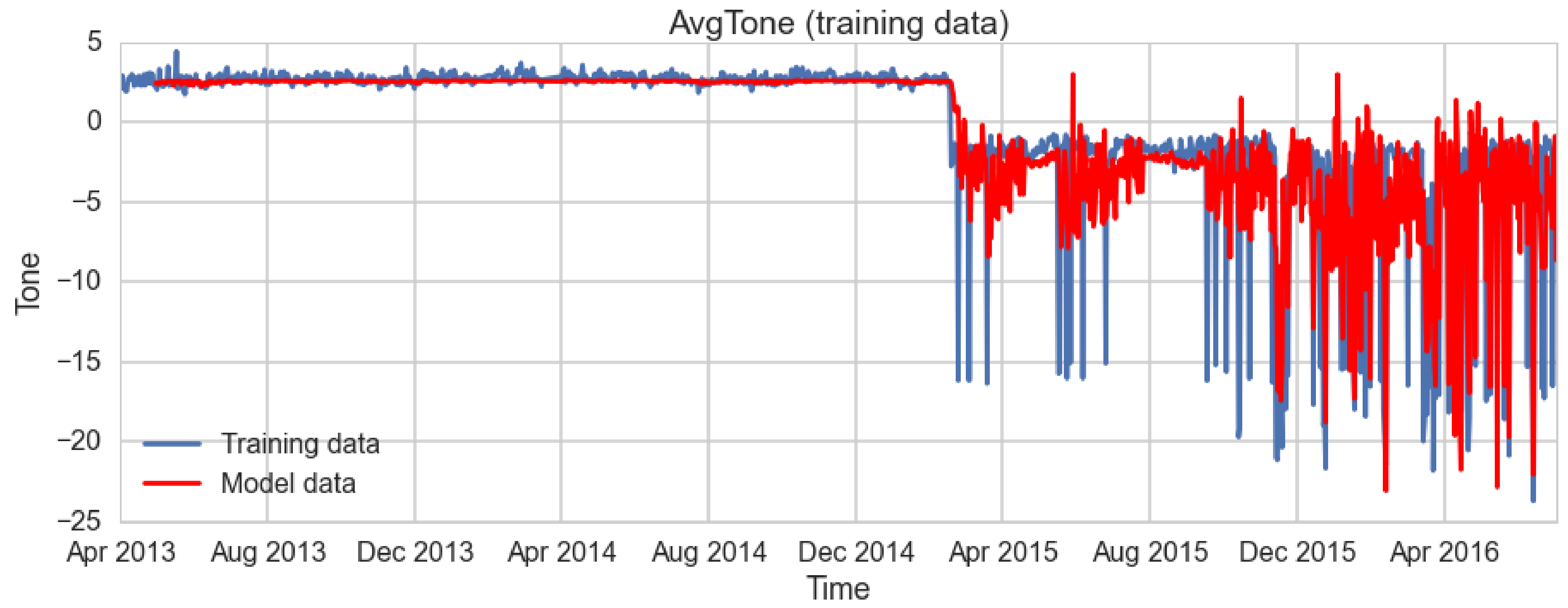


Figure 5. Training and model data (after implementation Long Short-Term Memory Networks)

Applying neural network (2)

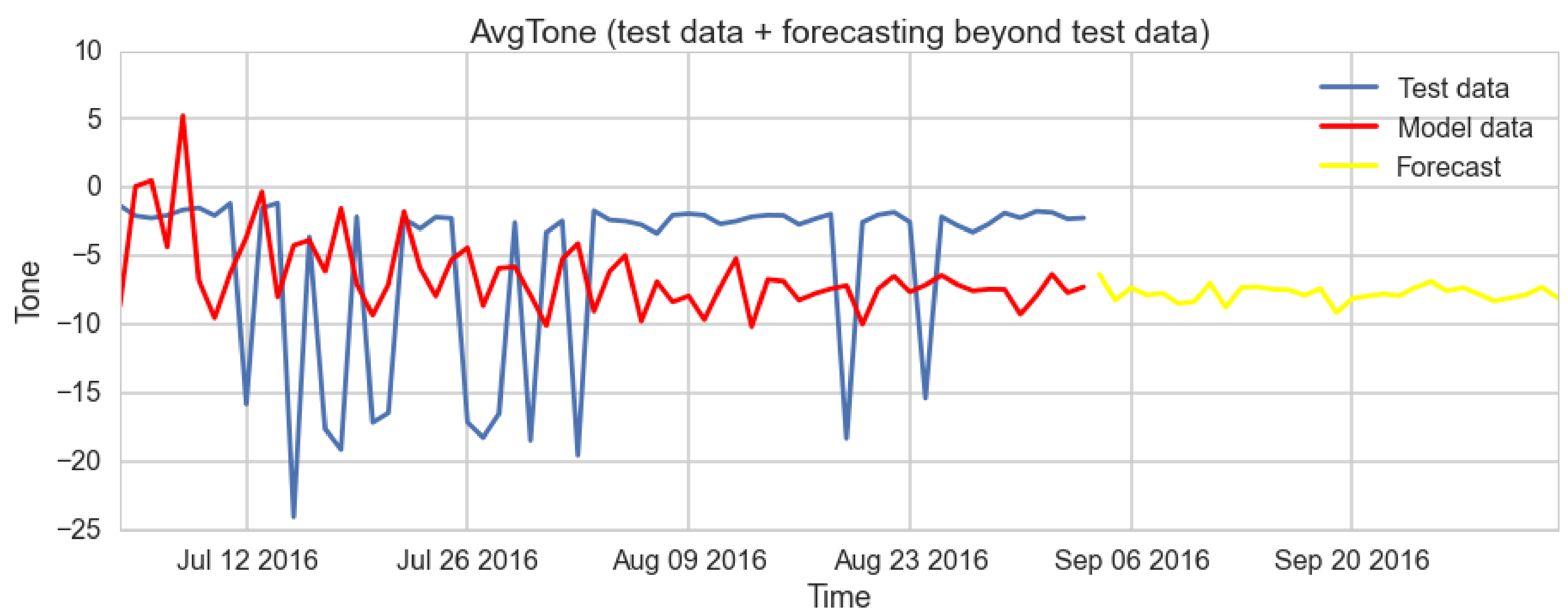


Figure 6. Test and model data (after implementation Long Short-Term Memory Networks)

Conclusion

- This project represents one of the biggest task in Data Science – prediction continues variable.
- Approach with representing events as a time series of tone and implementing NN requires some improvements:
  - find appropriate transformation for input data
  - play with parameters of NN
  - use genetic algorithm instead of NN
  - build a model for forecasting based on not only time, but also some other independent variables
- The most important step is the last improvement.

Thank you for your time

Q&A

Sergei Dmitriev,
Student Springboard