# FTS log analysis

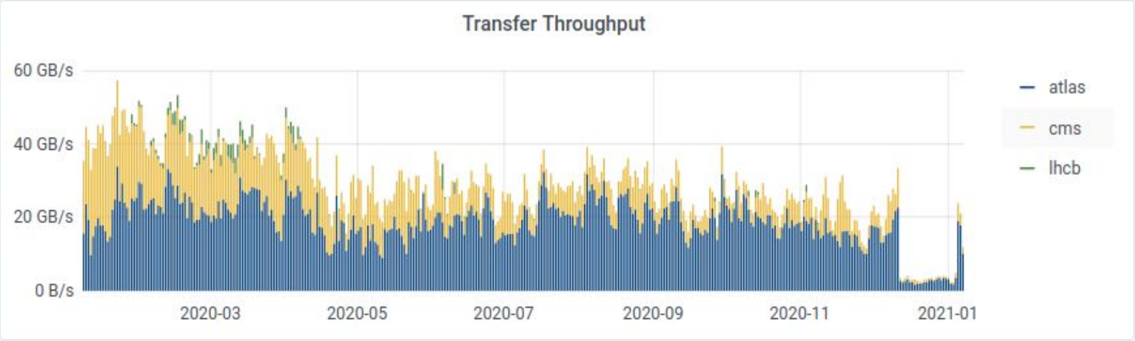
Use case ML INFN

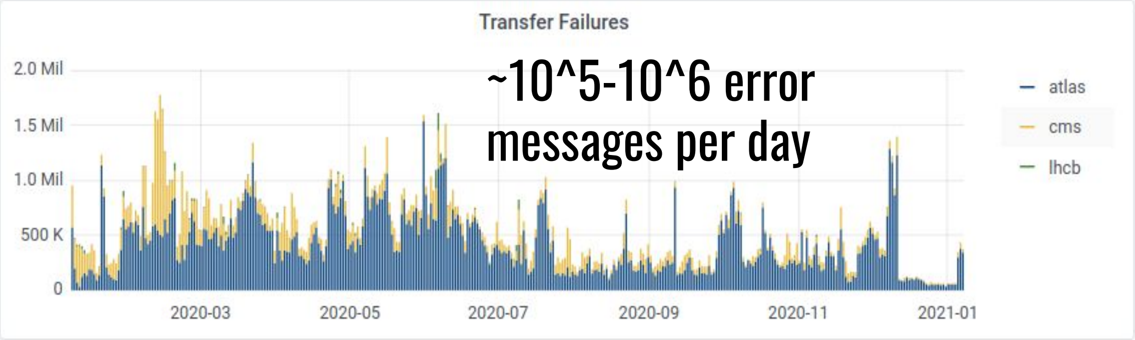
***Federica Legger,*** *Micol Olocco*

Istituto Nazionale di Fisica Nucleare

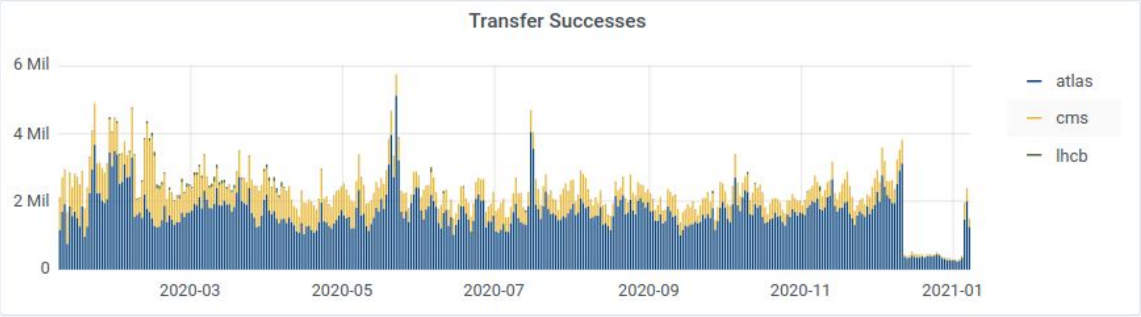
# FTS

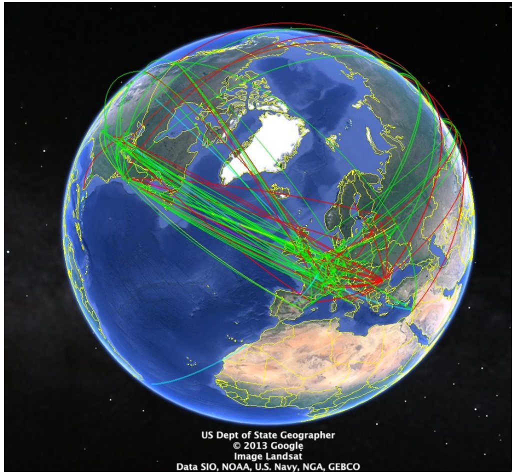
**FTS** (File Transfer Service) is the service responsible for globally distributing the **multiple petabytes of LHC data** across the WLCG infrastructure.





~10^5-10^6 error messages per day




US Dept of State Geographer
© 2013 Google
Image Landsat
Data SIO, NOAA, U.S. Navy, NGA, GEBCO

# Motivation

- For each FTS transfer:
    - Timestamp, source and destination hosts, protocol, filename
    - Error message in case of failed transfers
        - Often a concatenation of messages coming from different steps in the pipeline (handshake, data transfer, finalisation)
    - Error category (from GFAL)
        - FTS developers advise not to use it
- DDM Shifters/Operators address transfer failures by opening tickets to sites, escalating to experts, …
    - Huge manual effort, typically only biggest problems are addressed

- **Aim of this work: analyse error messages with ML**
    - Clusterize similar messages -> work in progress
    - Anomaly detection -> project with Google

# FTS error messages

```
TRANSFER ERROR: Copy failed with mode 3rd pull, with error: copy 0)
Could not get the delegation id: Could not get proxy request: Error
404 fault: SOAP-ENV:Server [no subcode] HTTP/1.1 404 Not Found Detail:
<!DOCTYPE HTML PUBLIC -//IETF//DTD HTML 2.0//EN> <html><head>
<title>404 Not Found</title> </head><body> <h1>Not Found</h1> <p>The
requested URL /gridsite-delegation was not found on this server.</p>
</body></html> .
```

```
DESTINATION OVERWRITE srm-ifce err: Communication error on send, err:
[SE][srmRm][] httpg://svr018.gla.scotgrid.ac.uk:8446/srm/managerv2:
CGSI-gSOAP running on lcgfts08.gridpp.rl.ac.uk reports Error reading
token data header: Connection closed
```

# Technique

- **WORD EMBEDDING:** mapping of a word into a numeric vector space. Similarity of two words given, for example, by the cosine of the relative angle.
- **PRE-PROCESSING PHASE**
  - **Data preparation: Cleaning** of the messages from particular (meaningless) information before vectorizing (hostnames, usernames, etc).
  - **tokenization:** example connection timed out during ssl handshake → [connection,timed,out,during,ssl,handshake]
- **PROCESSING PHASE**
  - **Vectorization**: Transformation of the textual information into numeric: Word2Vec+Sent2Vec.
  - **Clustering**: Grouping of the numerical representations : DBSCAN

# Preprocessing

**Which part of the error message do we want to define the clustering?**

```
A.  TRANSFER ERROR: Copy failed with mode 3rd pull, with error: Transfer
    failed: failure: rejected GET: 404 Not Found

B.  [gfalt_copy_file][perform_copy] TRANSFER [gfal_http_copy] ERROR: Copy
    failed with mode 3rd pull, with error: [davix2gliberr] Transfer failed:
    failure: Remote copy failed with status code 0: transfer closed with
    1486840628 bytes remaining to read
```

**TASK:** Cleaning implementation to select **nucleus messages**

⚠ **ENCAPSULATED SYSTEM**

# Corpus creation

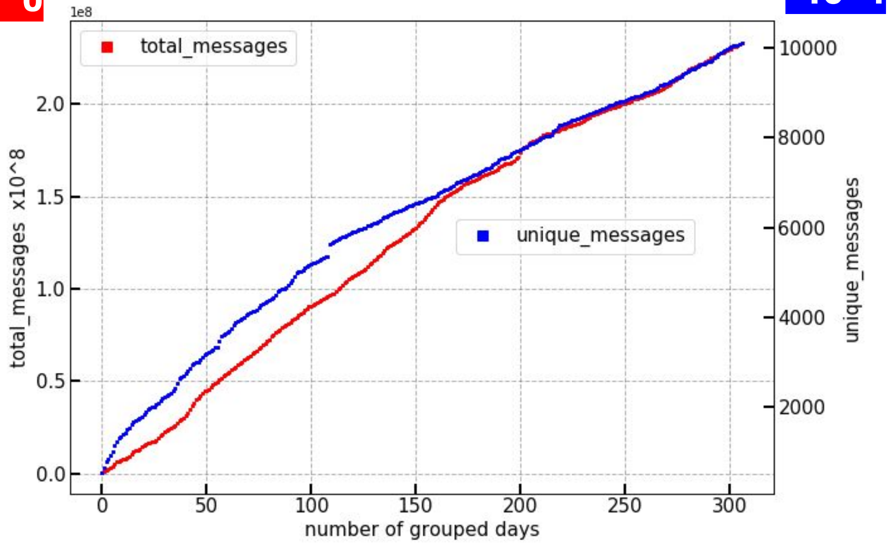
How to Generate a Good Word Embedding?

**CORPUS:** collection of documents given as input for training the model

- **SIZE:** training on a large corpus ***generally*** improves the quality of word embeddings
- **TYPE:** training on an in-domain corpus can ***significantly*** improve the quality of word

~10^8

~10^4



number of total messages (left axis) and unique messages (right axis) over time

**Corpus domain prevails over corpus size → No update of existing pre-trained models**

# Details

- Code on github https://github.com/leggerf/TestsINFNCloud/blob/master/test_clusterLogs/NLP_example.ipynb
- Small input file message_example.zip also in github
- Tested also with input file on minio (minio.cloud.infn.it/)
- Runs on any system with JupyterHub (your machine, INFN Cloud, Google colab)
- Description in confluence https://confluence.infn.it/display/MLINFN/9.+FTS+log+analysis+with+NLP

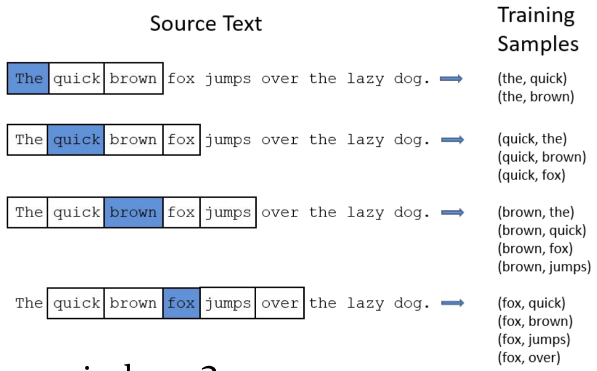
# Word2vec

Word2Vec: algorithm to perform **numerical representations for words** that capture their *meanings* and *semantic relationships* → if vectors, computers can handle them.

Two possible architectures:

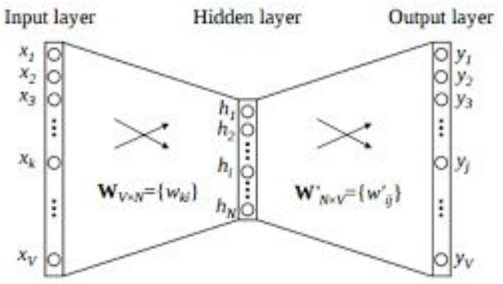
- CBOW predicts the current word based on the context
- Skip-gram predicts surrounding words given the current word.

**GOAL:** to learn the weights of the hidden layer

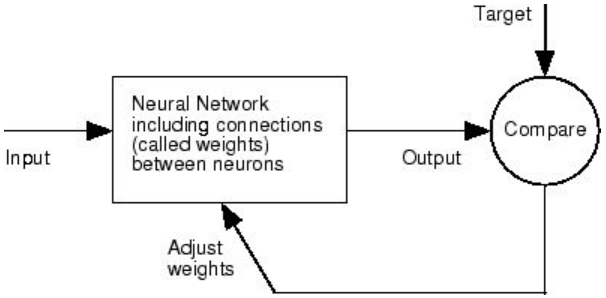


| Source Text | Training Samples |
|---|---|
| The quick brown fox jumps over the lazy dog. | (the, quick)<br>(the, brown) |
| The quick brown fox jumps over the lazy dog. | (quick, the)<br>(quick, brown)<br>(quick, fox) |
| The quick brown fox jumps over the lazy dog. | (brown, the)<br>(brown, quick)<br>(brown, fox)<br>(brown, jumps) |
| The quick brown fox jumps over the lazy dog. | (fox, quick)<br>(fox, brown)<br>(fox, jumps)<br>(fox, over) |

window=2



word2vec model architecture



Target
Neural Network including connections (called weights) between neurons
Input
Output
Compare
Adjust weights

The network is going to learn the statistics from the number of times each pairing shows up

visualize word2vec demo

# DBSCAN

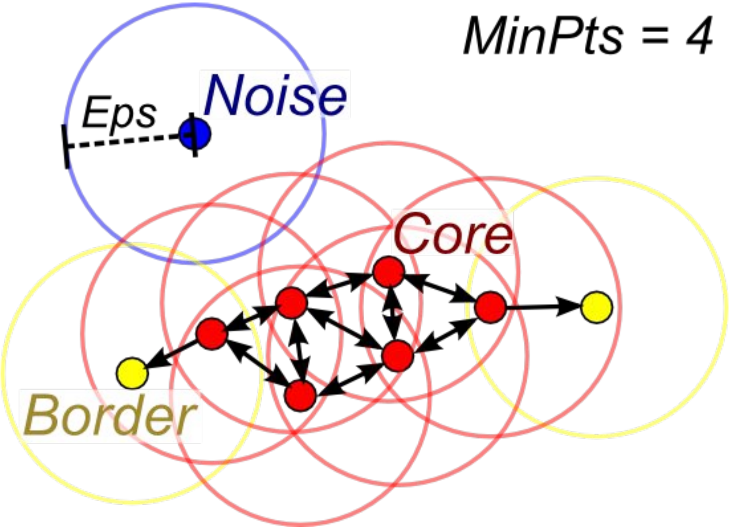
**DBSCAN - <u>Density-Based Spatial Clustering of Applications with Noise</u>**

Two parameters, **min_samples and eps**, which define formally *density* concept:

- **eps=**The maximum distance between two samples for one to be considered as in the neighborhood of the other.

- **min_samples=**The number of samples in a neighborhood for a point to be considered as a core point. This includes the point itself.

Higher min_samples or lower eps indicate higher density necessary to form a cluster.


MinPts = 4
Eps
Noise
Core
Border