

Privacy Preserving Pandemic Monitoring with Big Data Analytics

Dr. Thomas Win

University of Gloucestershire

twin@glos.ac.uk

Overview

- 1 Big Data
- 2 Federated Learning
- 3 Setting the Scene
- 4 Applications in pandemic monitoring
- 5 Challenges
- 6 Conclusion

Overview

- Refers to large datasets
- Challenging for traditional data processing approaches
- Obtained from a wide variety of sources
- Typically in the order of GBs

Categories of Big Data

- **Structured**

- Organised in a pre-defined *schema*

- **Semi-structured**

- Consist of both structured/unstructured characteristics

- **Unstructured**

- Does not consist of a pre-defined structure

- **Note:** We refer to the *contents* of the data rather than its *metadata*

Architecture

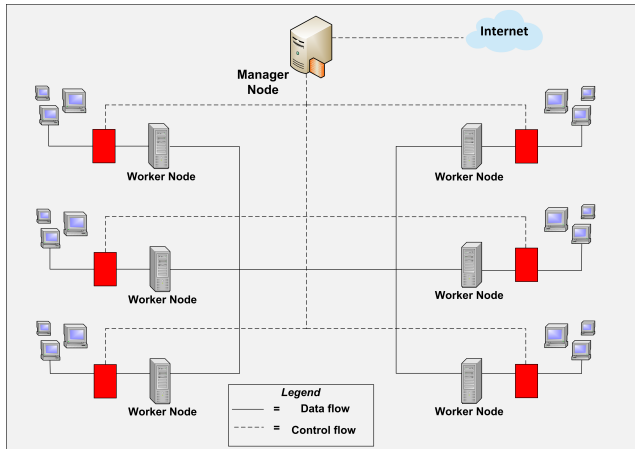


Figure: Big Data Analytics architecture

Overview

- Decentralised means of data analytics and machine learning [1]
- Training done locally using the local data
- Model trained is shared
- Does not require storing data centrally

Architecture

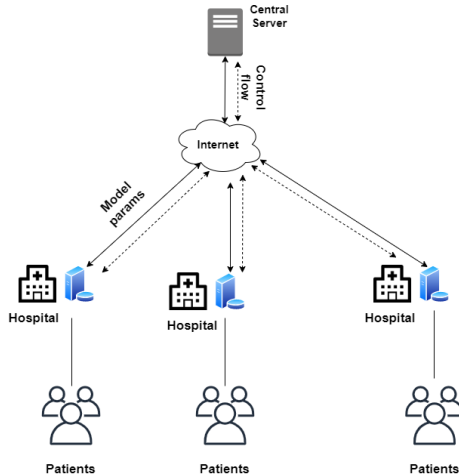
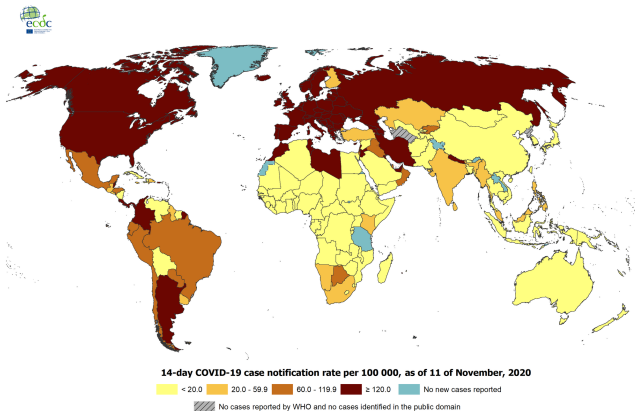


Figure: Federated Learning overview

Setting the Scene



Administrative boundaries: © EuroGeographics © UN-FAO © Turkstat.
The boundaries and names shown on this map do not imply official endorsement or acceptance by the European Union.

Date of production: 11/11/2020

Figure: COVID-19 global spread over a 14-day period, from *European Centre for Disease Prevention and Control* [2]

Use of BDA-based FL in pandemic monitoring

- Distributed nature of patient data has seen widespread application of FL in pandemic detection
- Different approaches to this end:
 - Outbreak prediction
 - Tracking pandemic spread
 - Treatment and drug discovery

Outbreak prediction

- Features the use of geographically-distributed datasets
- Used for predicting and visualising effects of pandemic spread
- Prediction accuracy depends on different factors (e.g., population size, infected cases, etc)
- Publicly-available datasets available on sites such as [data.world](#) *data.world* [3] and [John Hopkins University Centre for Systems Science and Engineering](#) *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University* [4]

Tracking pandemic spread

- Features the use of both *structured* and *unstructured* data
- For instance, Zhao, Liu, and Li [5] used air flight data and destination population to estimation pandemic spread
- Led to the development of a macroscopic growth law

Treatment and drug discovery

- Given the rate of infection has resulted in the generation of genome data
- Used for identifying possible vaccine through genome analysis, such as the work done by Ortea and Bock [6]
- Leverages the distributed nature of federated learning





Research challenges

- Regulations
 - Varying prevention measures makes data collection a challenge
- Lack of standard datasets
 - Datasets are varied both in terms of structure and detail
- Privacy and security issues
 - Need for compliance with various governmental privacy regulations (e.g., GDPR, HIPAA)

Bringing it all together

- The COVID-19 pandemic has had a significant effect on a global scale
- Federated Learning and BDA leverage to identify cure and mitigate its spread
- Still an open research problem with venues to explore

References I

-  Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 1273–1282.
-  *European Centre for Disease Prevention and Control*.
<https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases>. Accessed: 12-11-2020.
-  *data.world*. <https://data.world/datasets/covid-19>. Accessed: 10-11-2020.
-  *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*.
<https://github.com/CSSEGISandData/COVID-19>. Accessed: 10-11-2020.

References II



Xumao Zhao, Xiang Liu, and Xinhai Li. “Tracking the spread of novel coronavirus (2019-nCoV) based on big data”. In: *medRxiv* (2020).



Ignacio Ortea and Jens-Ole Bock. “Re-analysis of SARS-CoV-2 infected host cell proteomics time-course data by impact pathway analysis and network analysis. A potential link with inflammatory response.”. In: *BioRxiv* (2020).

Q & A