Red Team
The more we are together,
the smarter the solutions!
Members:
- *Kseniaa VASYLIEVA from Passau*
- *Salim HAFID from Lyon*
- *Davide BALICE from Milano*
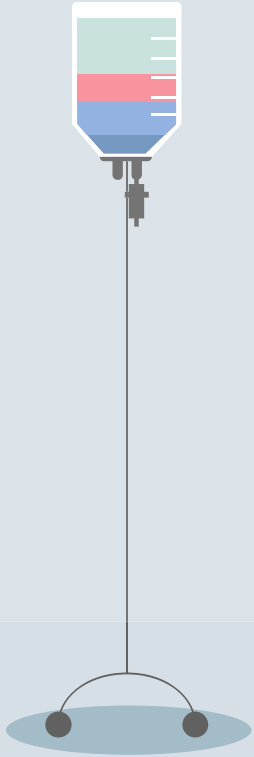- *Alexandru BREBU from Bucharest*

SEMINAR 1
22-26.02.2021

PASSAU

2021 - 02
Passau

2021 - 06
Lyon

2021 - 10
Milan

2022 - 02
Bucharest

Data-driven sustainable development, possible or utopic?

Salim Hafid
Kseniia Vasylieva
Davide Balice
Andrei-Alexandru Brebu

TABLE OF CONTENTS

01

Introduction

Overview of Big Data in sustainable development

02

Focus Topic : Healthcare

1. How can Big Data help improve healthcare :

- Use-case 1: Medical Diagnosis
- Use-case 2: COVID-19

2. What are the dangers that come with using Big Data in healthcare : The case of Privacy
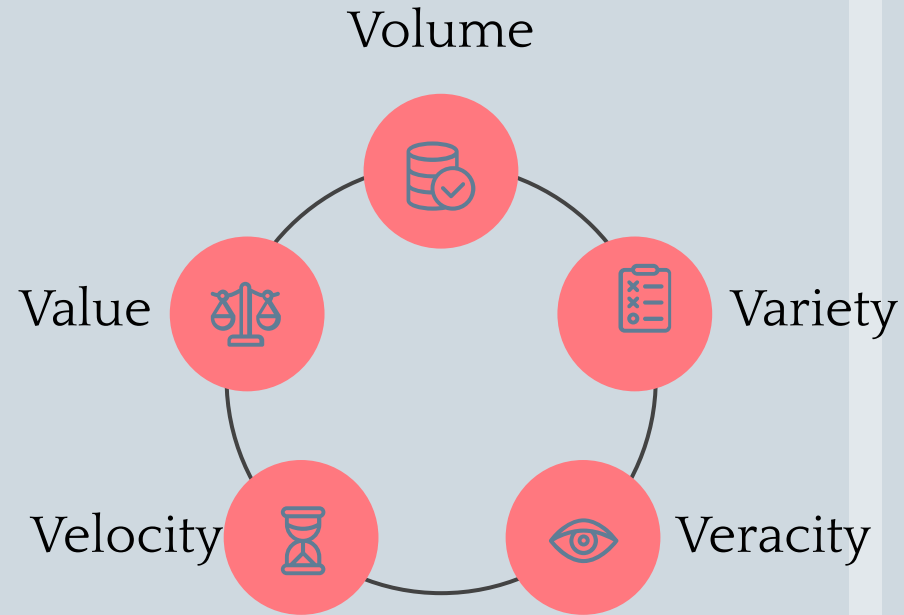
03

Conclusion

Should Big Data be used to improve healthcare ? Challenges & future directions of research

Big Data: Definition
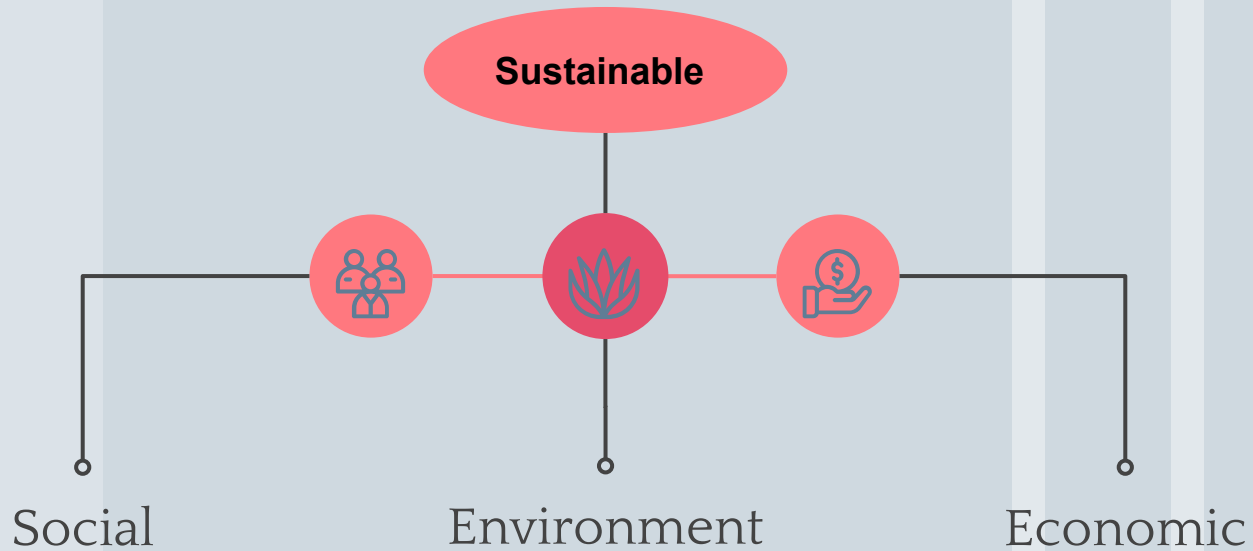

Any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications

image source:https://www.btelligent.com/themen/big-data/

The Concept of Sustainable
Development

Sustainable

Social

Environment

Economic

Sustainable development
Goals

image source: https://www.unicef.org/georgia/sustainable-development-goals

How can data help
Use-case 1: Diagnosis

EARLY DIAGNOSIS

CLINICAL DIAGNOSIS
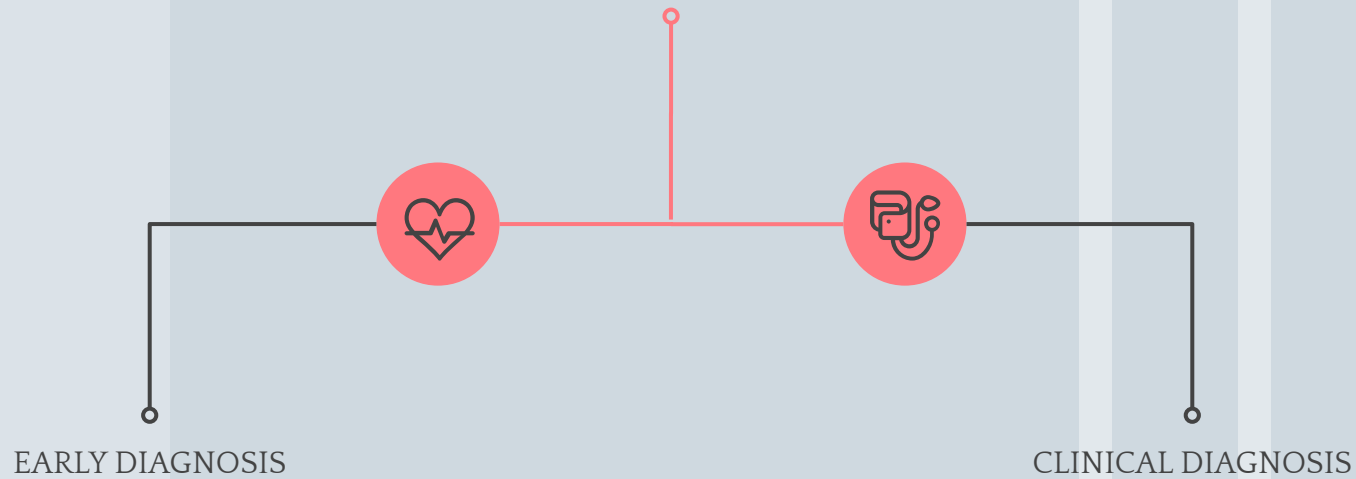
Early Diagnosis
The case of neurological diseases

PREVALENCE

- More than 600 different diseases of the nervous system [17]

- An estimated 6.8 million people die every year as a result of neurological disorders [17]

TREATMENTS & GENERATED DATA

- Current diagnosis technologies produce huge quantities of data for detection, monitoring and treatment

- Medical image data range anywhere from a few megabytes to hundreds of megabytes per study [17]

36.5 m
cases of dementia worldwide [18]

Clinical Diagnosis
The case of heart diseases

PREVALENCE

– World's deadliest diseases (over 8 million deaths in 2019 [19])

TREATMENTS & GENERATED DATA

– Generated data are high in volume and can carry non-linear relationships

Attribute	Data type	Attribute description
Age	Real	Age (in years)
Sex	Binary	0-Female, 1-male
Cp	Nominal	Chest pain type (1-typical angina, 2-atypical angina, 3-nonangina, 4-asymptomatic
Restbps	Real	Resting blood pressure (mm of hg)
Chol	Real	Serum cholesterol (mg/dl)
Fbs	Binary	Fasting blood sugar > 120 mg/dl (1-true, 0-false)
Restecg	Nominal	Resting electrocardiographic results (0-normal, 1-having ST-T wave normality, 2-probable/defined left ventricular hypertrophy)
Thalach	Real	Maximum record heart rate
Examg	Binary	Angina induced by exercise (1-yes, 0-false)
Odpeak	Real	ST depression tempted by workout comparative to rest
Slope	Nominal	Slant of the peak exercise ST segment (1-upsloping, 2-flat, 3-downsloping)
Ca	Real	Major vessels colored by fluroscopy
Thal	Nominal	3-normal, 6-fixed defect, 7-reversible defect
Calss	Binary	Represent present or absence of heart disease (1-present, 2-absence)

Attributes used for a heart disease prediction model - source : [19]

How can data help
Computer-aided-diagnosis (CAD)

What's it for

Decision support system to the health experts

What it does

Assists experts in accurately interpreting medical big data

Benefits to diagnosis

+ improves accuracy
+ improves consistency
+ increases treatments' success rate

sources : [17]

How can data help
Use-case 2: COVID-19

Other directions

e.g. noisy adaptive group testing
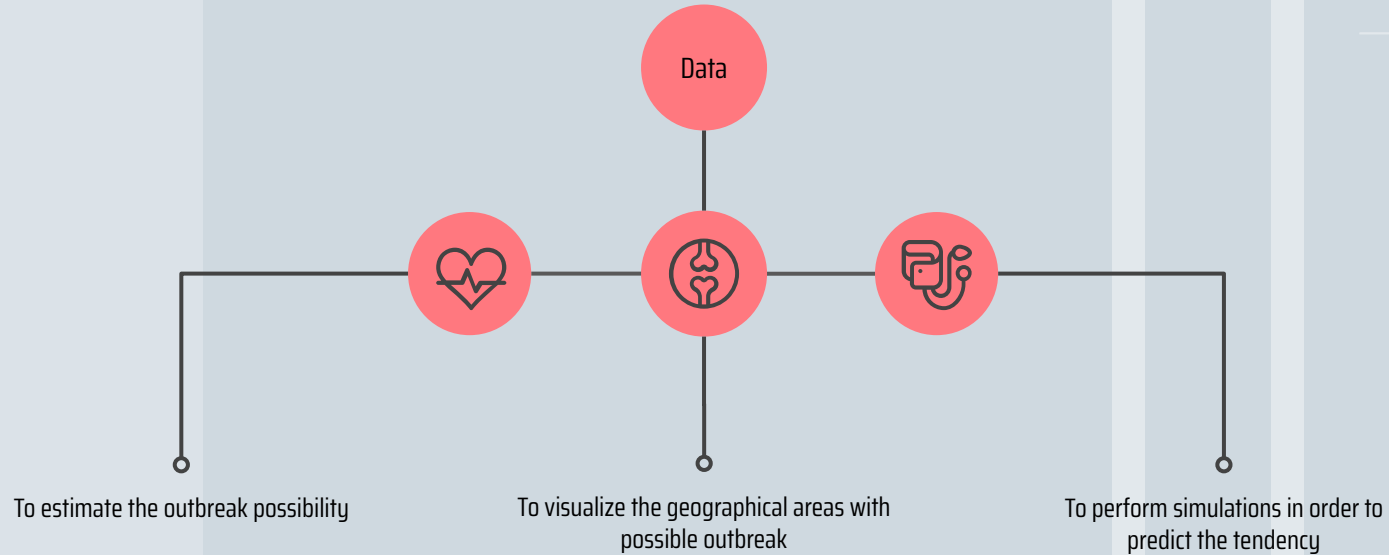strategy with prior information
on patients [4]

Outbreak
Prediction

Spread
Tracking

COVID-19
Diagnosis

Vaccine/Drug
Production

Main resource used for covering this use-case [24]

How can data help - COVID-19
Outbreak Prediction

Data

To estimate the outbreak possibility

To visualize the geographical areas with possible outbreak

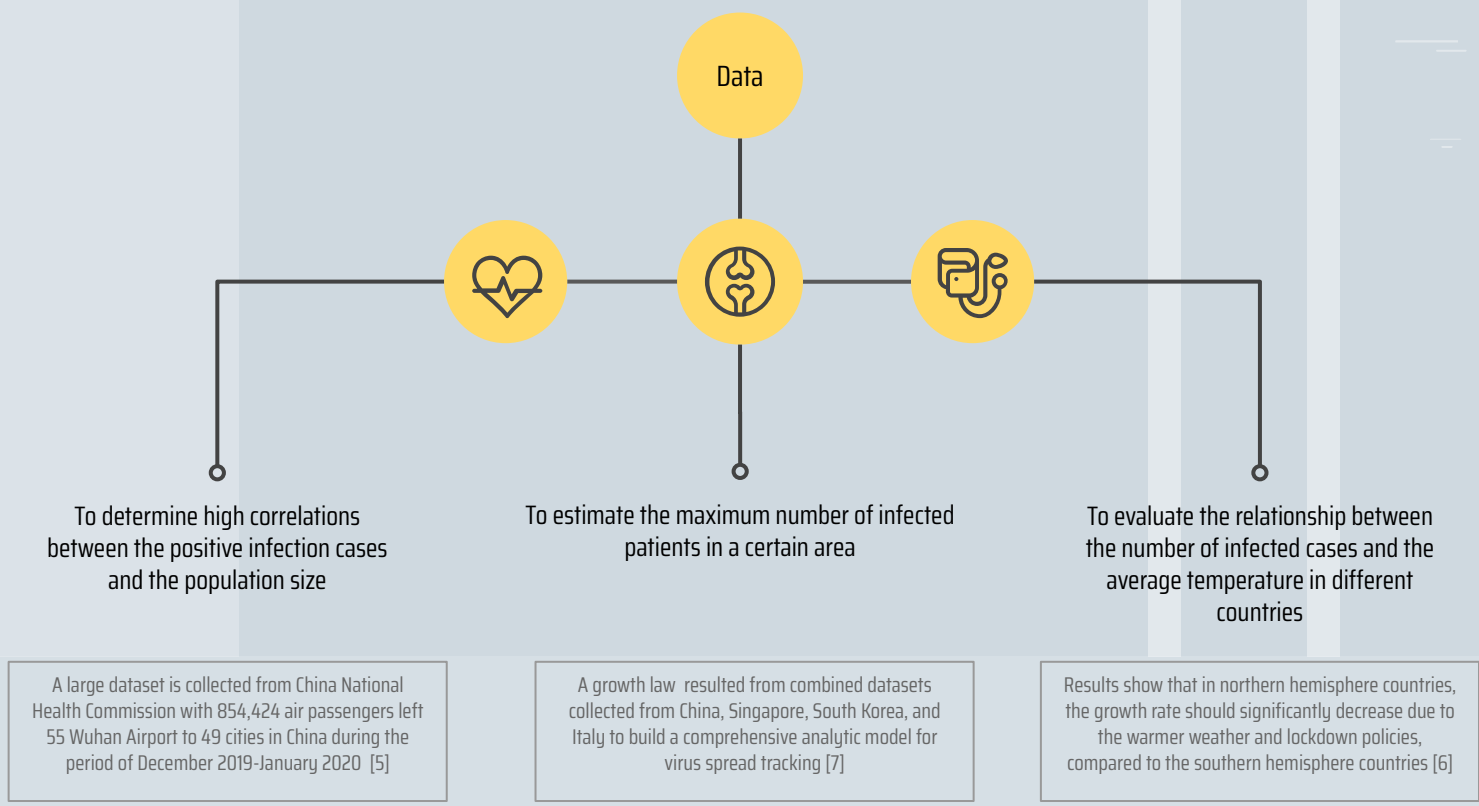To perform simulations in order to predict the tendency

The authors developed complex models that can accurately formulate the dynamics of the pandemic based on huge data sets from Italian Civil Protection sources. [2]

A first trial is an investigation in Wuhan, aiming to monitor the people migration from and to Wuhan city so that the health agencies can predict the population infected with COVID-19 for quarantine. [1]

Trial implemented in India indicating that the proposed data analytic can estimate the outbreak in short-term intervals, i.e. two-week duration [3]
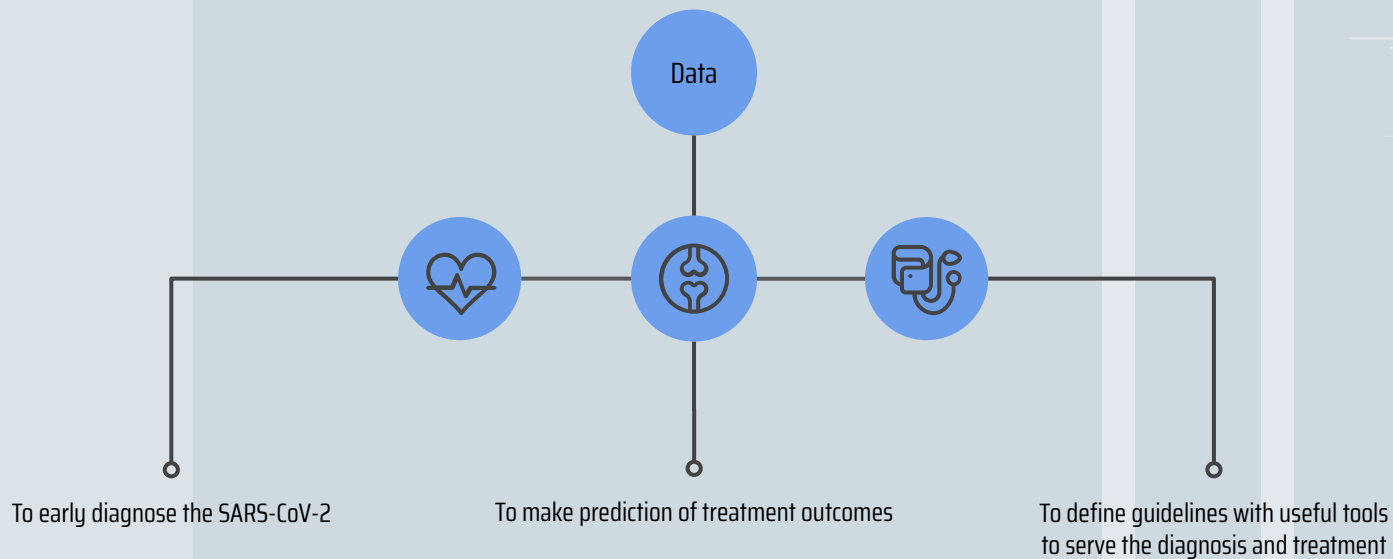
How can data help - COVID-19
Spread Tracking

Data

To determine high correlations between the positive infection cases and the population size

To estimate the maximum number of infected patients in a certain area

To evaluate the relationship between the number of infected cases and the average temperature in different countries

A large dataset is collected from China National Health Commission with 854,424 air passengers left 55 Wuhan Airport to 49 cities in China during the period of December 2019-January 2020 [5]

A growth law resulted from combined datasets collected from China, Singapore, South Korea, and Italy to build a comprehensive analytic model for virus spread tracking [7]

Results show that in northern hemisphere countries, the growth rate should significantly decrease due to the warmer weather and lockdown policies, compared to the southern hemisphere countries [6]

How can data help - COVID-19
COVID-19 Diagnosis

Data

To early diagnose the SARS-CoV-2

To make prediction of treatment outcomes

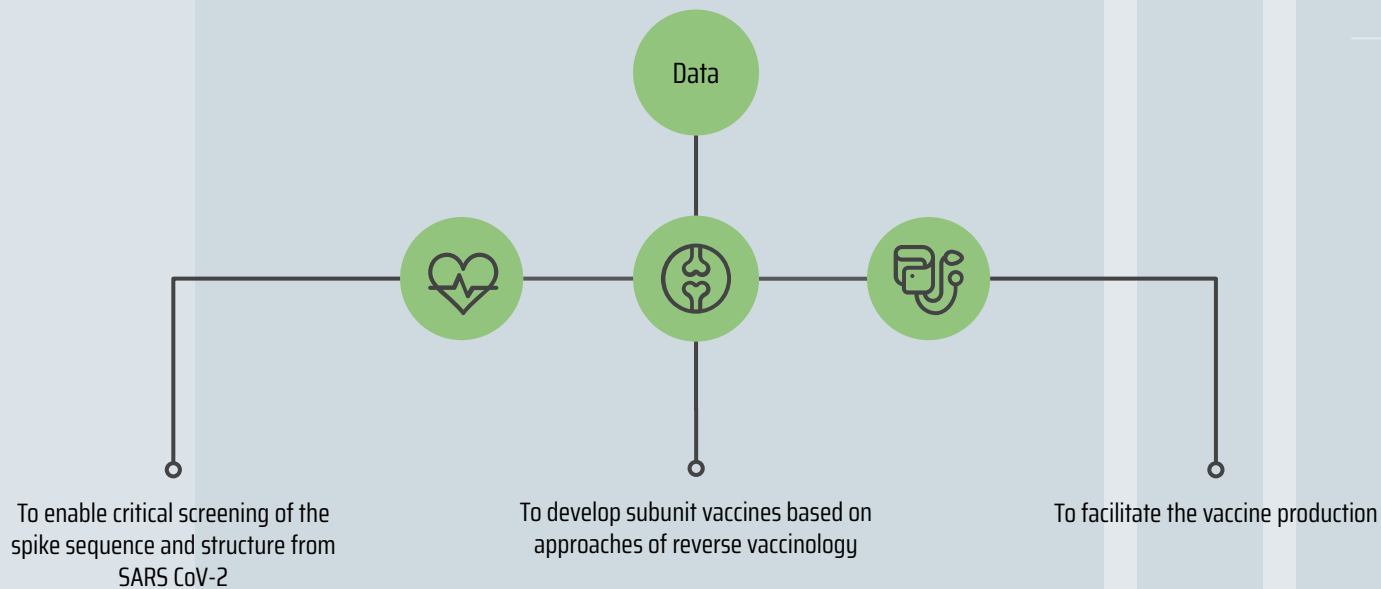To define guidelines with useful tools to serve the diagnosis and treatment

A study in [11] introduced a robust, sensitive, specific and highly quantitative solution based on multiplex polymerase chain reactions that are able to diagnose the SARS-CoV-2

A large-scale data of Zhongnan Hospital of Wuhan University has been analysed from a collection system where 11,500 persons were screened, and 276 were identified as suspected infectious victims, and 170 were diagnosed. [12]

How can data help - COVID-19
Vaccine/Drug Production

Data

To enable critical screening of the spike sequence and structure from SARS CoV-2

To develop subunit vaccines based on approaches of reverse vaccinology

To facilitate the vaccine production

The work in [14] used the GISAID database (www.gisaid.org/CoV2020/) to extract the amino acid residues

The strain of the SARS-CoV-2 has been selected by reviewing numerous entries of the online database of the National Center for Biotechnology Information [15]

A huge dataset has been also collected from the National Center of Biotechnology Information for facilitating vaccine production [16]. Different peptides were proposed for developing a new vaccine against COVID-19.

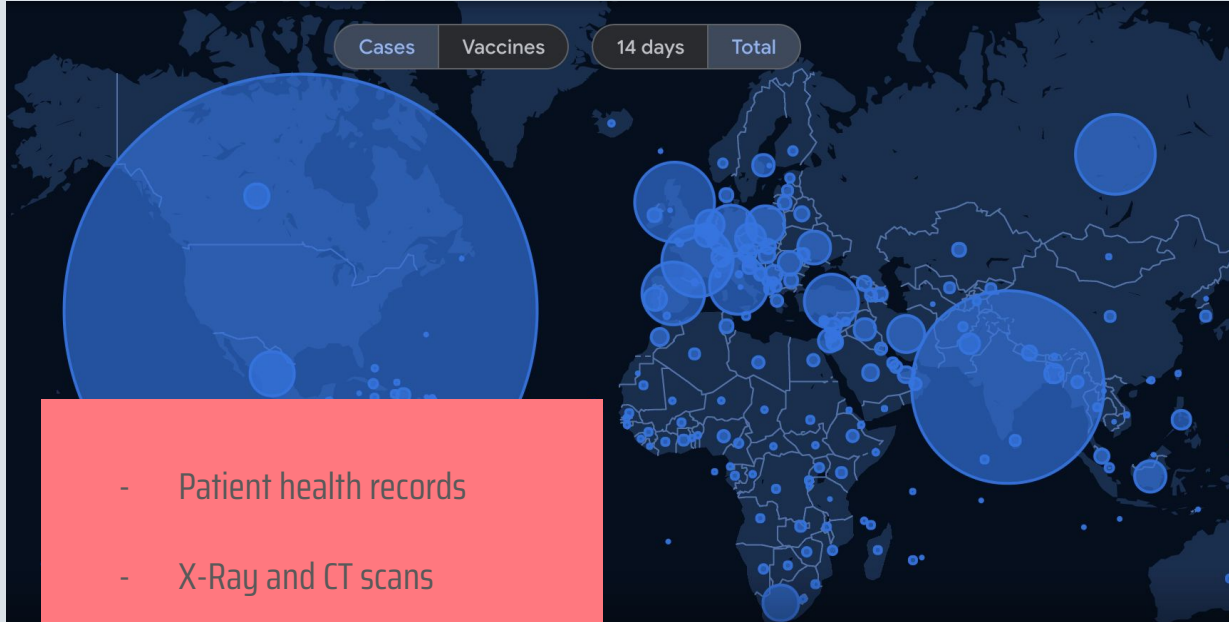# How can data help - COVID-19
Relevant questions

What does the data used for fighting COVID-19 look like?

What does the intersection of Big Data and COVID-19 look like?

How can data help - COVID-19
What does the data used for fighting COVID-19 look like?

- Patient health records

- X-Ray and CT scans

- Infected case history
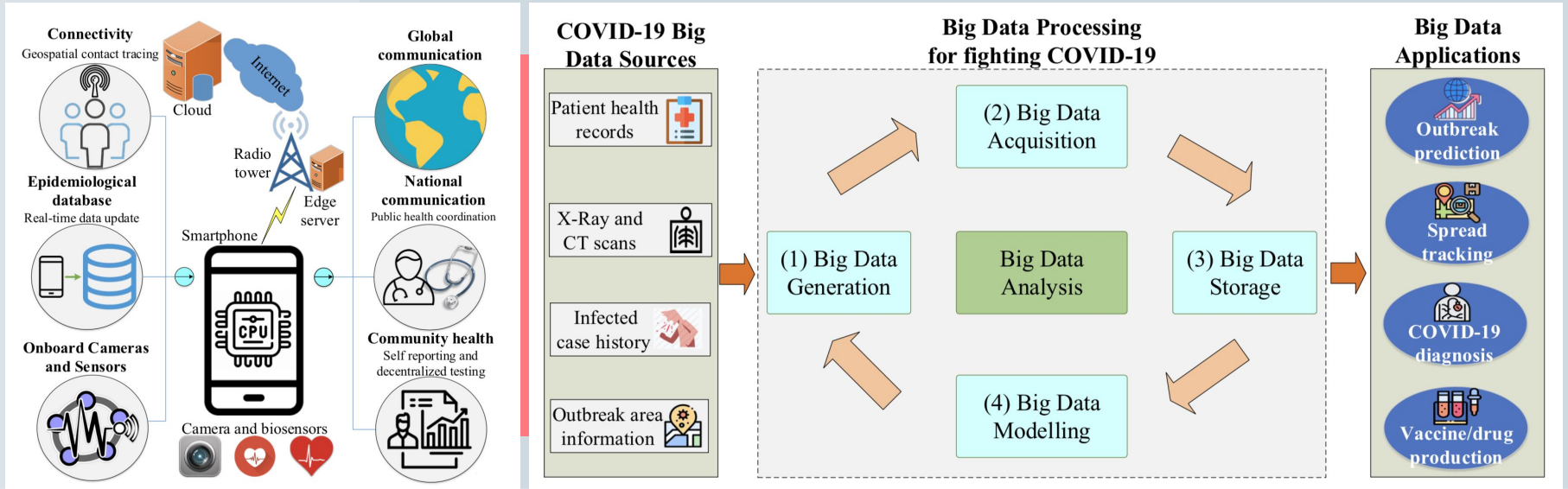
- Outbreak area information

How can data help - COVID-19
What does the intersection of Big Data and COVID-19 look like?

A framework for COVID-19 diagnosis and surveillance [24]   *Big data and its applications for fighting COVID-19 pandemic* [24]

Why should we protect health data?

Patients' trust is mainly
towards their physician

Health records are worth money

How can data be dangerous
Privacy

SSN	NAME	DoB	SEX	ZIP	MARITAL STATUS	DISEASE
🚫	🚫	64/09/13	F	94141	DIVORCED	STOMACH ULCER
🚫	🚫	64/04/15	F	94139	MARRIED	CHEST PAIN
🚫	🚫	63/03/13	M	94139	MARRIED	GASTRITIS
🚫	🚫	63/03/18	M	94139	MARRIED	SHORT BREATH
🚫	🚫	64/09/27	F	94138	SINGLE	SHORT BREATH

How information can be disclosed:

- Linking Attack
- Similarity Attack
- Skewness Attack

Example of published de-identified microdata
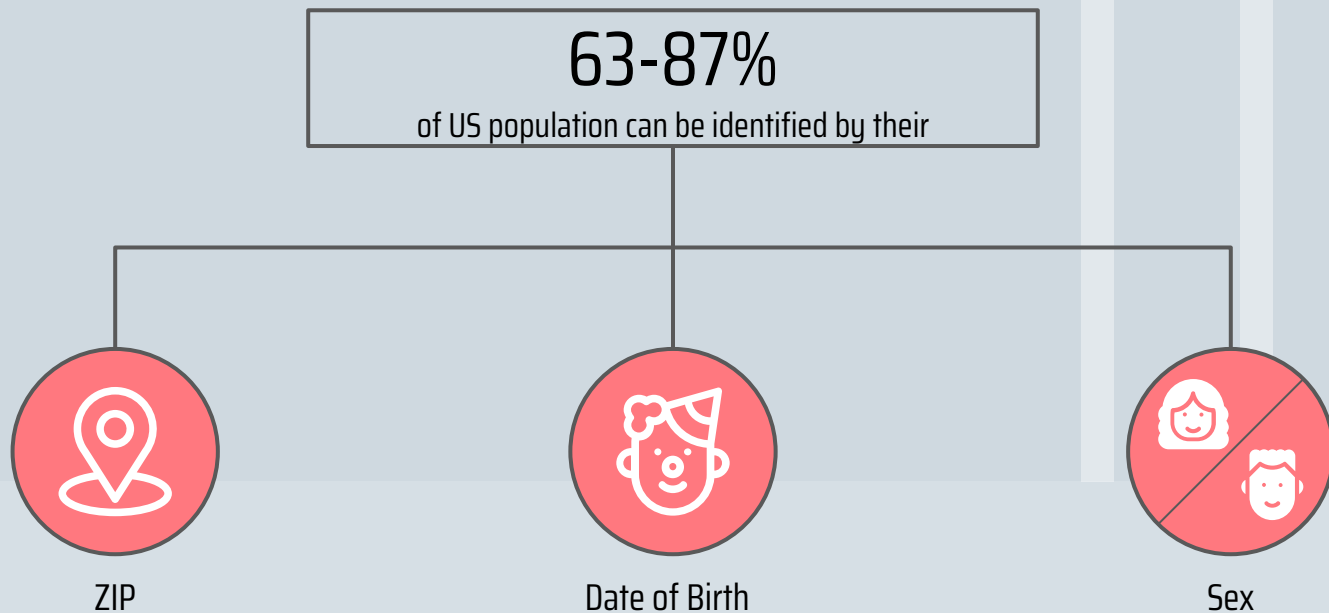
Ending Discussion
Challenges and Solutions

RED TEAM

The lack of standard datasets

Many AI algorithms and big data platforms have been proposed, but they are not tested using the same dataset.

For example, we cannot decide which algorithm is better for the virus detection since two datasets may have different numbers of samples or noise could be present, etc

MR. NOBODY

"What do you think are the major challenges that need to be discussed and addressed in the future based on what we're experiencing now ?"

RED TEAM

Privacy and security challenges

New laws protecting user privacy have been enacted (GDPR) and others will be (ePrivacy) by the European Union. For this reason it is necessary to enforce these regulations to protect the privacy of users, but further work must be done to make this happen

THANK YOU

REFERENCES

- [1] B. Chen, M. Shi, X. Ni, L. Ruan, H. Jiang, H. Yao, M. Wang, Z. Song, Q. Zhou, and T. Ge, "Visual data analysis and simulation prediction for COVID-19," arXiv preprint arXiv:2002.07096, 2020.

- [2] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, and M. Colaneri, "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy," Nature Medicine, pp. 1–6, 2020.

- [3] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, "Epidemic analysis of COVID-19 in China by dynamical modeling," arXiv preprint arXiv:2002.06563, 2020.

- [4] Marco Cuturi, Olivier Teboul, Quentin Berthet, Arnaud Doucet, Jean-Philippe Vert, Noisy Adaptive Group Testing using Bayesian Sequential Experimental Design arXiv:2004.12508v6

- [5] X. Zhao, X. Liu, and X. Li, "Tracking the spread of novel coronavirus (2019-ncov) based on big data," medRxiv, 2020.

- [6] A.Notari,"Temperature dependence of COVID-19 transmission,"arXiv preprint arXiv:2003.12417, 2020.

- [7] P. Castorina, A. Iorio, and D. Lanteri, "Data analysis on coronavirus spreading by macroscopic growth laws," arXiv preprint arXiv:2003.00507, 2020.

- [8] "Almost one zettabyte of mobile data traffic in 2022-cisco," 2020. [Online]. Available: https://telecoms.com/495666/almost-one-zettabyte-of-mobile-data-traffic-in-2022-cisco/

- [9] G.Manogaran, D.Lopez, C.Thota, K.M.Abbas, S.Pyne, and R.Sundarasekar, "Big data analytics in healthcare internet of things," in Innovative healthcare systems for the 21st century. Springer, 2017, pp. 263– 284.

- [10] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K. M. Abbas, and R. Sundarsekar, "Big data knowledge system in healthcare," in Internet of things and big data technologies for next generation healthcare. Springer, 2017, pp. 133–157.

- [11] C. Li, D. N. Debruyne, J. Spencer, V. Kapoor, L. Y. Liu, B. Zhou, L. Lee, R. Feigelman, G. Burdon, J. Liu et al., "High sensitivity detection of coronavirus SARS-CoV-2 using multiplex PCR and a multiplex-PCR-based metagenomic method," bioRxiv, 2020.

- [12] Y.-H. Jin, L. Cai, Z.-S. Cheng, H. Cheng, T. Deng, Y.-P. Fan, C. Fang, D. Huang, L.-Q. Huang, Q. Huang et al., "A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-ncov) infected pneumonia (standard version)," Military Medical Research, vol. 7, no. 1, p. 4, 2020.

REFERENCES

- [13] J. R. Lon, Y. Bai, B. Zhong, F. Cai, and H. Du, "Prediction and evolution of B cell epitopes of surface protein in SARS-CoV-2," bioRxiv, 2020.

- [14] S. F. Ahmed, A. A. Quadeer, and M. R. McKay, "Preliminary identifica- tion of potential vaccine targets for the COVID-19 coronavirus (SARS- CoV-2) based on SARS-CoV immunological studies," Viruses, vol. 12, no. 3, p. 254, 2020.

- [15] B. Sarkar, M. A. Ullah, F. T. Johora, M. A. Taniya, and Y. Araf, "The essential facts of wuhan novel coronavirus outbreak in China and Epitope-based vaccine designing against 2019-nCoV," BioRxiv, 2020.

- [16] M.I.Abdelmageed,A.H.Abdelmoneim,M.I.Mustafa,N.M.Elfadol, N. S. Murshed, S. W. Shantier, and A. M. Makhawi, "Design of a Multiepitope-Based Peptide Vaccine against the E Protein of Human COVID-19: An Immunoinformatics Approach," BioMed Research Inter- national, vol. 2020, 2020.

- [17] Siuly, S., Zhang, Y. Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis. *Data Sci. Eng.* **1,** 54–64 (2016). https://doi.org/10.1007/s41019-016-0011-3

- [18] World Health Organisation, Dementia : https://www.who.int/news-room/fact-sheets/detail/dementia Consulted on February 23rd 2021

- [19] Thanga Selvi, R., Muthulakshmi, I. An optimal artificial neural network based big data application for heart disease diagnosis and classification model. *J Ambient Intell Human Comput* (2020). https://doi.org/10.1007/s12652-020-02181-x

- [20] Mulligan E, Braunack-Mayer A. Why protect confidentiality in health records? A review of research evidence. Aust Health Rev. 2004;28(1):48-55. doi: 10.1071/ah040048. PMID: 15525250.

- [21] Selck-Paulsson D., Morris C., Ileana C. and Bijman D, "Follow-up study: Data breaches in the health sector", Orange Cyberdefense https://orangecyberdefense.com/global/blog/healthcare/follow-up-study-data-breaches-in-the-health-sector/

- [22] Ciriani, Valentina & Vimercati, Sabrina & Foresti, Sara & Samarati, Pierangela. (2009). Theory of privacy and anonymity.

- [23] Rothstein, Mark A. "Is deidentification sufficient to protect health privacy in research?." The American journal of bioethics : AJOB vol. 10,9 (2010): 3-11. doi:10.1080/15265161.2010.494215

- [24] Q. Pham, D. C. Nguyen, T. Huynh-The, W. Hwang and P. N. Pathirana, "Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts," in IEEE Access, vol. 8, pp. 130820-130839, 2020, doi: 10.1109/ACCESS.2020.3009328.