International Semantic Intelligence Conference (ISIC 2022)

Keynote

19th May 2022

Leveraging Artificial Intelligence and Machine Learning in Pandemics using COVID-19 as a Case Study

Professor Dr. rer. nat. habil. Sven Groppe
https://www.ifis.uni-luebeck.de/index.php?id=groppe

Born in COVID-19 Pandemic

- International Semantic Intelligence Conference (ISIC)
  - in 2021: planned in-person, but **hybrid** (MERI College, New Delhi, India)
  - in 2022: planned hybrid (Georgia Southern University, USA), but **online**
- International Health Informatics Conference (IHIC 2022)
  - in 2022: **hybrid** (Sri Sri University, Cuttack, Odisha, India)
- Before pandemic:
  - online and hybrid conferences do not have any good reputation
- During (and after?) pandemic:
  - "All" inclusive top ranked conferences go online or hybrid
  - no traveling necessary
    - less time consumed $\rightarrow$ more people participating even those not able before
  - often registration fees are enormously lowered and some events even do not have any fees at all
    - researchers with low budget can participate
  - more choices, but also more picking of sessions and presentations by participants

Worldwide: COVID-19 Confirmed Cases Daily

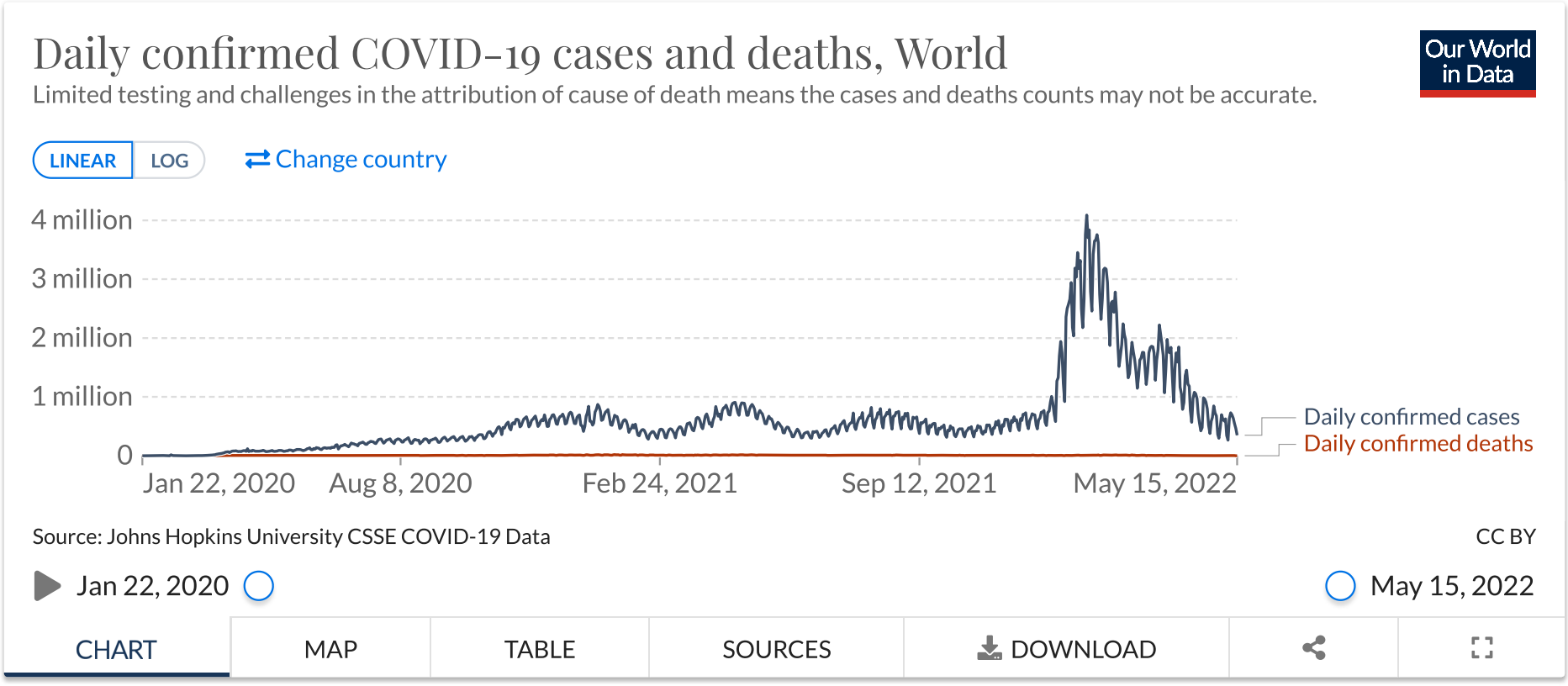
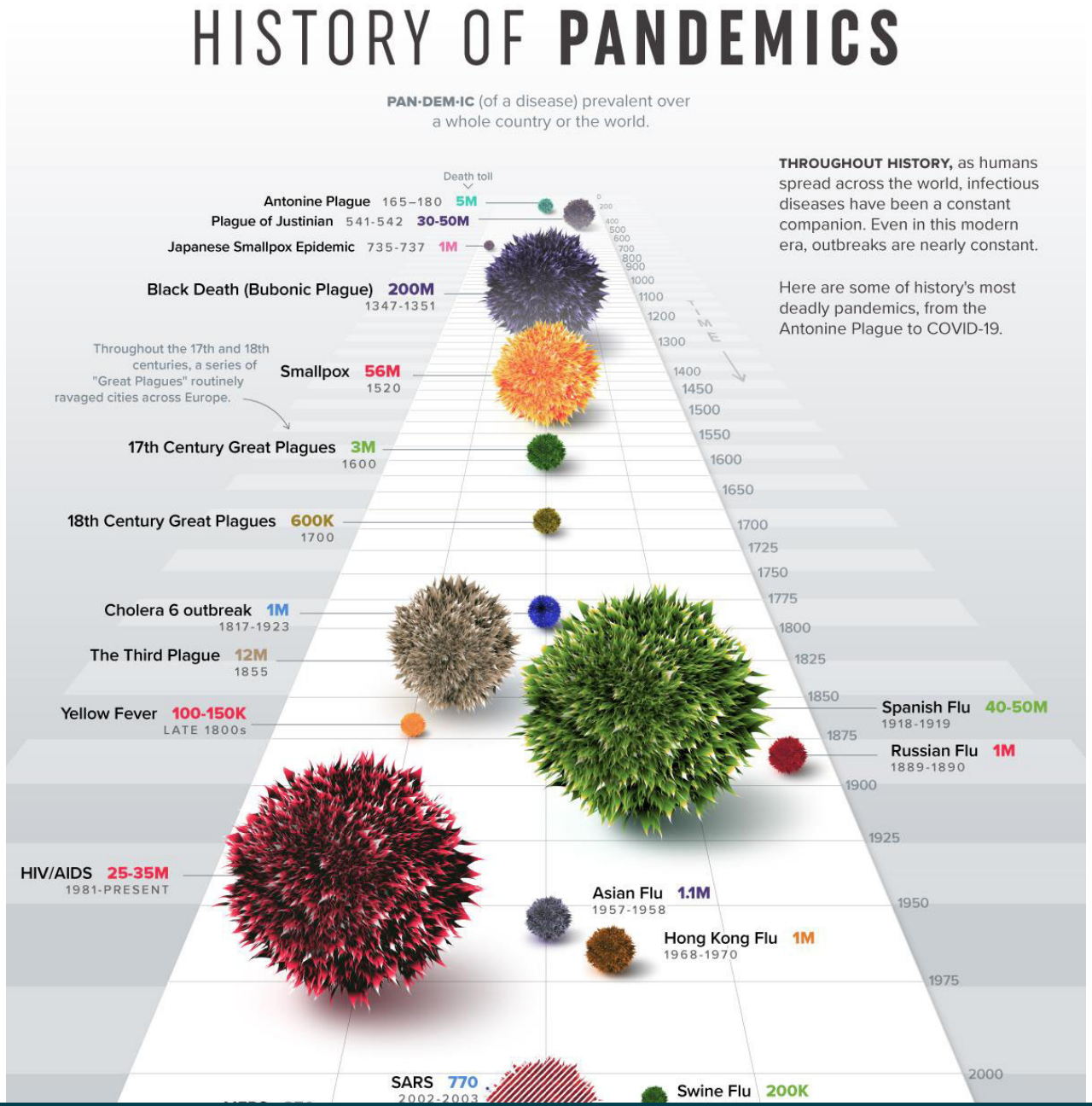


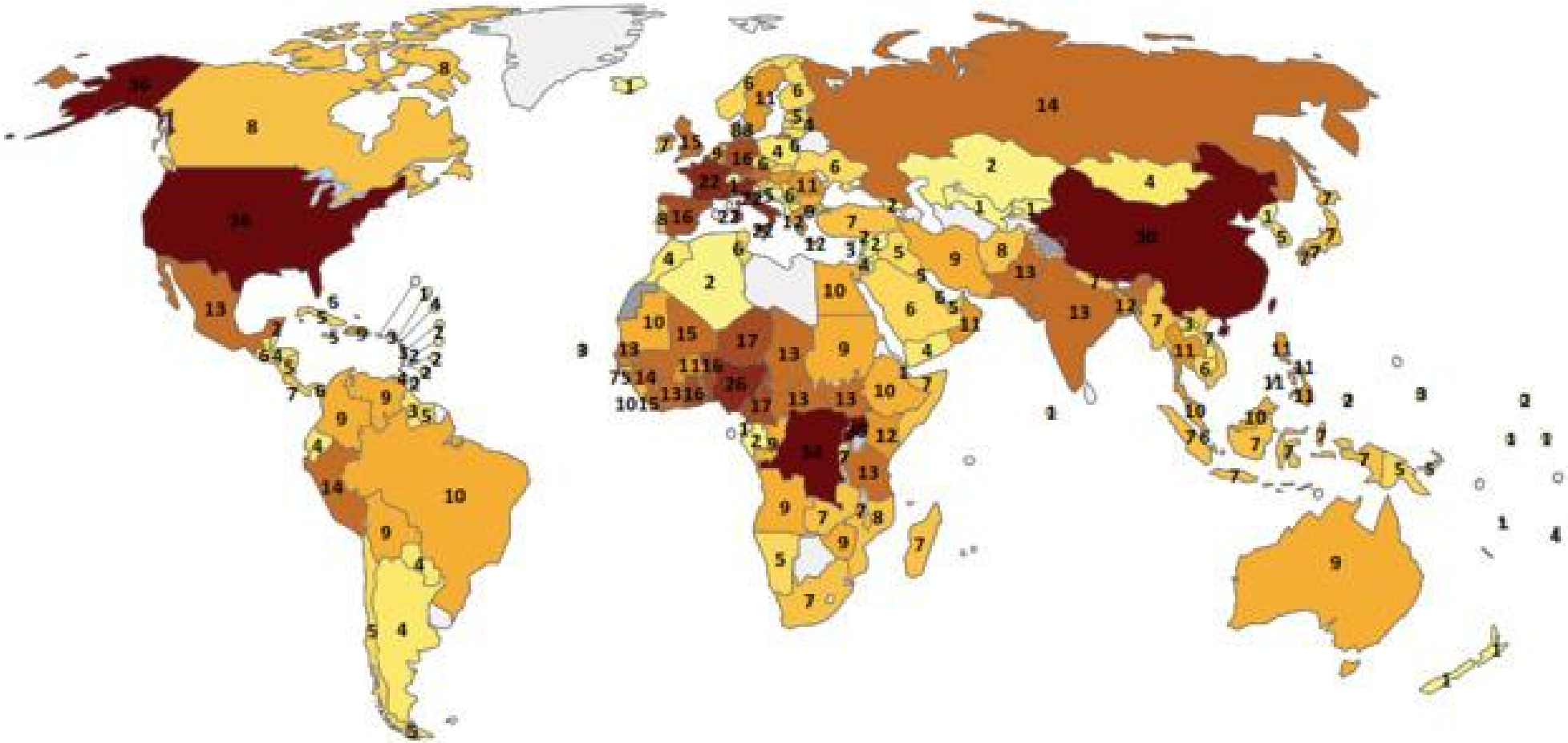
Chart of Mortality, Great Plague in London (1665 to 1666)



Source: https://www.visualcapitalist.com/history-of-pandemics-deadliest/

Global Epidemic Events

- Burden of epidemics: illustrations: epidemic events* globally, 2011–17: a total of 1307 epidemic events in 172 countries.

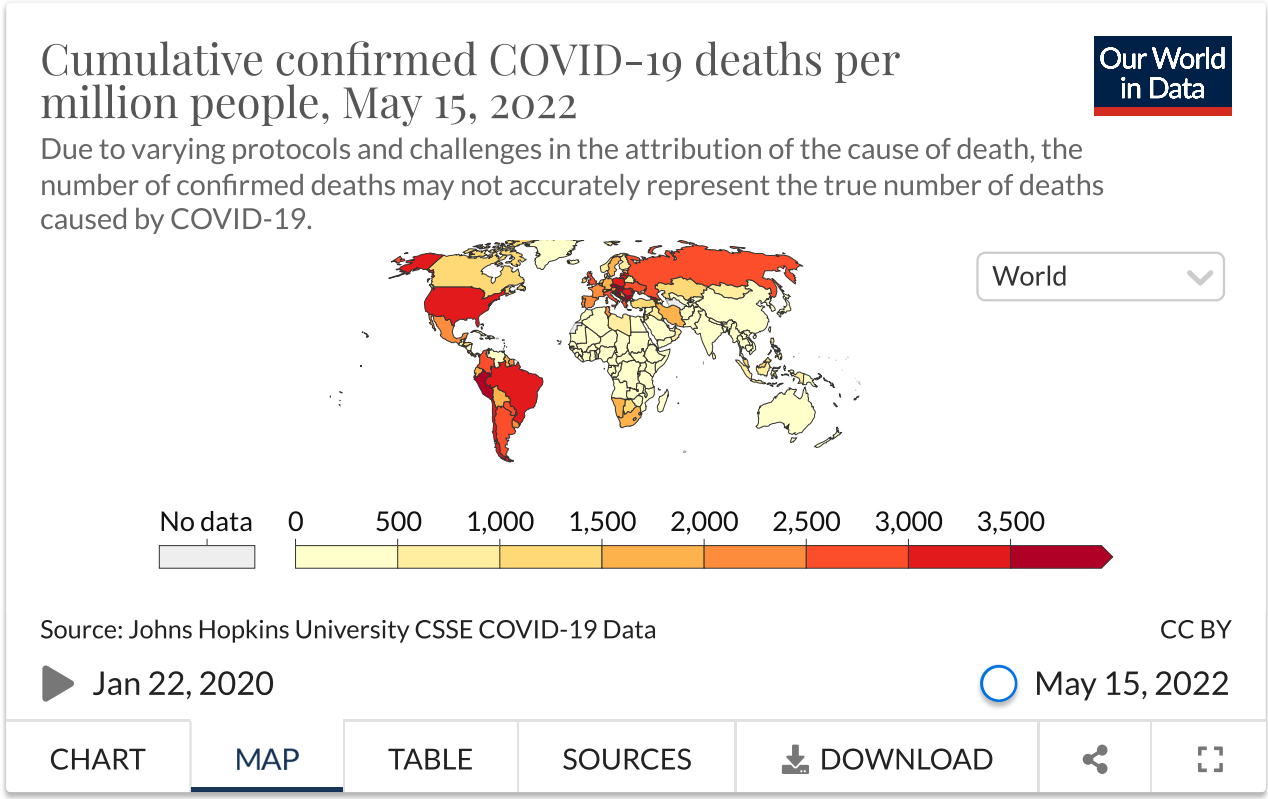


Courtesy: Managing epidemics, Key facts about major deadly diseases, Version 1, World Health Organization, 2018.

Global Health Security Index

of countries with population of more than 5 million

Cumulative confirmed COVID-19 deaths per million people



COVID-19 Tech Timeline 1/2

Date	COVID-19 Discovery	Tech Trend
January 7, 2020	Novel coronavirus identified	Virus recognized as coronavirus within weeks of the first-identified cases of 'pneumonia of unknown cause' thanks to supercomputers
January 12, 2020	Genome sequenced	Supercomputers and big data allowed researchers to analyze the genetic sequences of COVID-19 patients and SARS-CoV-2 mutations at population scale
January 16, 2020	Diagnostic reagents optimized	SARS-CoV-2's genetic sequence $\rightarrow$ development of testing kits in a matter of weeks
January 30, 2020	Person-to-person transmission confirmed	Contact tracing technologies
February 19, 2020	'Spike' protein mapped	The mRNA vaccine type fast-tracks research and development by requiring significantly less data be sent to human cells (compared to traditional vaccines)

COVID-19 Tech Timeline 2/2

Date	COVID-19 Discovery	Tech Trend
October 5, 2020	Airborne transmission termed possible	Amidst waves of outbreaks, governments used big data and artificial intelligence technologies to predict transmission in their communities and across their borders. The data was digitally collected in real time from sources like mobile phones, mobile payment applications, and social media platforms.
December 2, 2020	Vaccine first authorized	One vaccine found approval (in UK) and others entered stage three clinical trial evaluations.
Since December 2020	Common variants produced global concern	Researchers used supercomputers to determine the sequences of variants' genomes for global mutation surveillance and vaccine efficacy assessments.

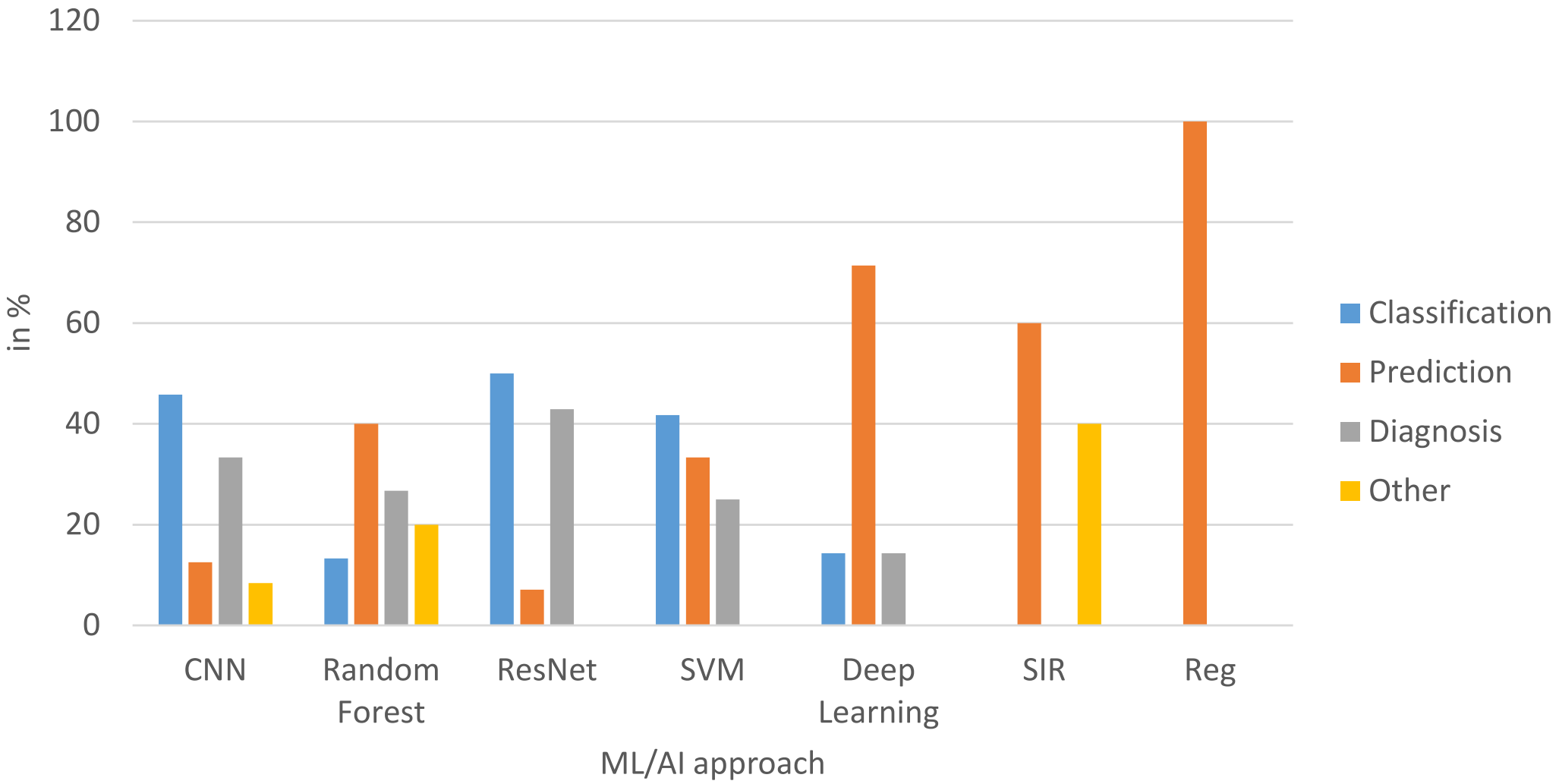
Where Computers can help... with topics discussed so far

- Prediction of incidence rates
  - considering COVID-19 confinements and other contexts
    - Simulation
    - Machine learning approaches
- Management of physical contacts, e.g., at events and restaurants
- Warning public and single persons
- Software within Health Systems, e.g.
  - Patient registration and status in hospitals
  - Databases of confirmed COVID-19 cases
  - Detection and diagnosis of COVID-19 patients based on computer tomography scans

ML/AI approaches in COVID-19 Scientific Literature 1/2

- Systematic Survey of 264 papers

ML/AI approaches in COVID-19 Scientific Literature 2/2



Contact Tracking - Privacy Risks

- Health Status Privacy
  - leak the identities of users infected by COVID-19 (or who have been in contact with them)
  - should remain accessible only to the infected users and the health authority
- Location Privacy
  - leak a user's mobility traces
    - Geolocation-based: require location to infer proximity
    - Bluetooth-based: co-location information and local Bluetooth sniffing stations
- Social Graph Privacy
  - leak user's social graph
    - through proximity data between users (for Bluetooth-based systems),
    - based on location data (for location-based systems)
  - no need of a global social graph to perform contact tracing (but only the contacts between infected users and other users (proximity/local graph))
  - Knowing the social graphs of a significant number of users $\Rightarrow$ de-anonymize these users (by comparing with e.g. social graph in social networks)

Contact Tracking - International Apps

	Launch Date	Cell-phone location data	Proximity-based solutions			
			Legacy Bluetooth Low Energy	Dongle for the Elderly (without smartphone)	Apple-Google Exposure Notification	DP-3T
Israel	18.3.2020	✓				
Singapore	20.3.2020		✓			
Singapore	28.6.2020			✓		
Austria (outdated)	1.4.2020		✓			
Austria	26.6.2020				✓	
Australia	26.4.2020		✓			
Italy	1.6.2020				✓	
France	2.6.2020		✓ centralised			
Germany	16.6.2020				✓	
Switzerland	25.6.2020				✓	✓

Contact Tracking - Evaluation

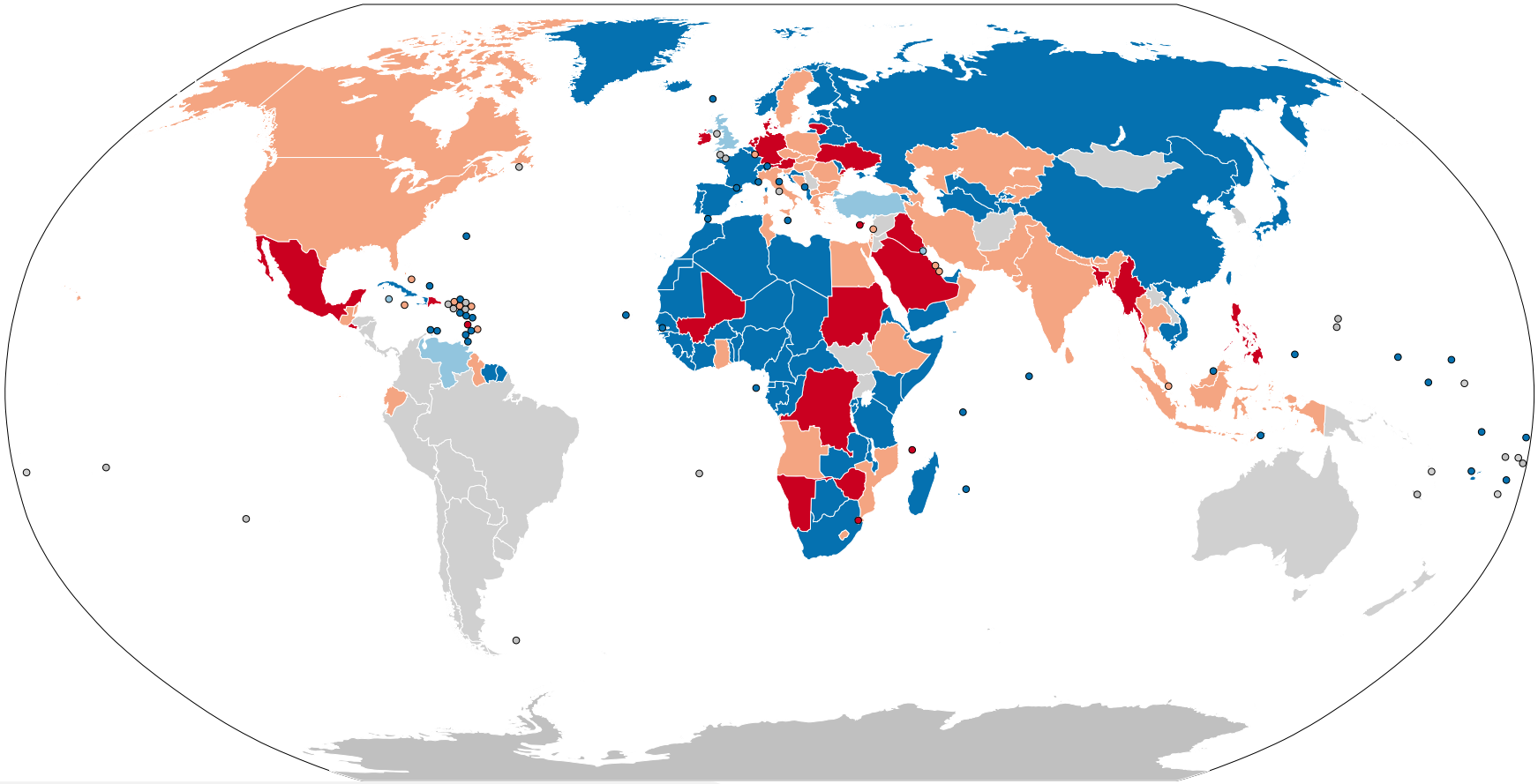
	Mobile phone Tracking	GPS Tracking App (SafePath)	Bluetooth Centralised Tracking App (France)	Bluetooth Decentralised Tracking App (Apple-Google, DP-3T)
Efficiency/accuracy (precision and notification time)	3 (')	2	2 (")	3
Privacy	1 (')	2	3	4
Cybersecurity	4	2	4	3
Battery efficiency	5	3	3	4 ("")
Adoption likelihood	-	1	2	3
OVERALL SCORE	3.3	2	2.8	3.4

(') These numbers relate to the Israelian approach, not Swisscom's Mobility Insights.
(") The centralised approach has a lower score despite its advantages mentioned earlier (i.e. pandemic oversight and faster notification) because it can only rely on legacy BLE, which has shown to be very unreliable for peer discovery on iPhones compared to the dedicated Exposure Notification offered by Apple-Google.
("") Apple-Google's Exposure Notification is expected to be more power-efficient than other solutions using the legacy Bluetooth API since it is handled at the OS level taking advantage of a phone's duty cycles.

AI-powered facial recognition to track COVID-19 cases

- Pilot project in South Korea becomes operational in January
  - rolled out in Bucheon (population: > 800,000 people)
- AI algorithms & facial recognition technology to analyze > 10,820 CCTV cameras and track an infected person's movements, anyone they had close contact with, and whether they were wearing a mask
  - South Korea already has an aggressive, high-tech contact tracing system that harvests credit card records, cellphone location data and CCTV footage, among other personal information.
  - manual work currently takes ½ to 1 hour to track 1 person
  - new system can simultaneously track up to 10 people in 5 to 10 minutes
  - opposition in South Korea compares system with 'Big Brother'
- China, Russia, India, Poland, Japan & several U.S. states have rolled out or at least experimented with facial recognition systems for tracking COVID-19 patients

Learners affected by school closures caused by COVID-19 as of February 2021



https://en.wikipedia.org/wiki/Impact_of_the_COVID-19_pandemic_on_education

COVID-19 as trigger for research

- Google Scholar Search

Year(s)	Number of Search Results for "COVID-19" in title	Number of Search Results for "COVID-19" anywhere in document
2019	758	≈ 74.600
2020	≈ 278.000	≈ 387.000
2021	≈ 150.000	≈ 208.000
2022	≈ 1.740	≈ 9.130
any time	≈ 314.000	≈ 4.440.000

Search on 6.12.2021 via Google Scholar

COVID-19 Datasets

Dataset	Content, Remarks
COVID-19 Open Research Dataset Challenge (CORD-19)	• > 500K scholarly articles, including > 200K with full text, about COVID-19, SARS-CoV-2, and related coronaviruses • 17 tasks like "What is known about transmission, incubation, and environmental stability?" • > 1.6K Notebooks on Kaggle
Novel Corona Virus 2019 Dataset	• daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus • 7 tasks like "Can We Correlate weather conditions and Corona virus Spread through Data?" • > 1.5K Notebooks on Kaggle
Open COVID-19 Data Working Group	• cases of a novel coronavirus • > 71M confirmed cases worldwide

COVID-19 Data Repository by Johns Hopkins University

- daily updates of confirmed COVID-19 cases and deaths, active and recovered patients, incident rates, number of people hospitalized and hospitalization rate per nation
- basis for their visual **dashboard**
- Aggregated data sources, e.g.:
  - World Health Organization (WHO)
  - European Centre for Disease Prevention and Control (ECDC)
  - DXY.cn. Pneumonia. 2020
  - US CDC
  - BNO News
  - WorldoMeters
  - 1Point3Arces
  - COVID Tracking Project
  - Los Angeles Times
  - The Mercury News

COVID-19 Knowledge Graphs (KGs)

- According to [P17], KGs
  - mainly describe real world entities and their interrelations,
  - define possible classes and relations of entities in a schema
  - allow for potentially interrelating arbitrary entities with each other and
  - cover various topical domains (here COVID-19 related topics)

Application	Authors
Surveillance in primary care	COVID-19 Surveillance Ontology
Infectious disease domain	Infectious Disease Ontology (IDO) IDO Virus IDO-Covid-19
Literature search	Steenwinckel et al., Wise et al., Cernile et al., Michel et al.
Drug repurposing	Stebbing et al., Wang et al., Domingo-Fernandez et al., Hsieh et al., Zhou et al.
Multi-purpose (e.g. phenotype, vaccines, drugs, COVID-19 response, SARS-CoV-2 virus–host interaction mechanisms)	WikiData, Chen et al., Reese et al., Ostaszewski et al., He et al., Dutta and DeBellis
Risk factor discovery	Bettencourt-Silva et al.
Case Statistics	CovidGraph

COVID-19 Knowledge Graphs (KGs) - Examples

	CIDO (He et al. 2020)	CODO (Dutta and DeBellis 2020)
Classes	82	51
Object Properties (Relations)	15	61
Data Properties (attributes)	-	45
Individuals	82	56
Logical Axioms	90	463
Further Statistics		over 71K patients, $\approx$ 5M triples (DeBellis and Dutta 2021)

CODO Competency Questions

- Find all instances of the class Patient
- Find all people diagnosed with Covid [who are in family relations]
- Gives all the patients who have contracted the virus from another
- Gives all the patients who have passed the virus to 2 (or more) possible patients
- Find the cities patients have travelled to
- Give all patients where we know the reason they caught the virus
- Find all clusters and the patients in them
- Count all the patients
- List all the patients where we know their city
- List all the patients between 18 and 30
- List all patients who have a diagnosis
- How many people recovered from COVId-19 in place p until date t?
- How many people died in country c?
- Give me the travel history of patient p?
- Give me the COVID-19 patients and their relationship, if any.
- ...

Search for COVID-19 classes on BioPortal

- Search URL: https://bioportal.bioontology.org/search?q=COVID-19

BioPortal information about ontologies

COVID-19 Ontologies on BioPortal

Which ontology to choose for your COVID-19 application?

- Not all COVID-19 ontologies will survive
  - Many ontologies will not maintained any more in the future
    - funded project ends $\longrightarrow$ no money any more for maintenance
    - some ontologies are not widely used
- Choose COVID-19 ontology with high impact
  - $\longrightarrow$ higher probability for being maintained in the future
  - $\longrightarrow$ higher interoperability with other applications

Which ontology to choose for your COVID-19 application?

- Not all COVID-19 ontologies will survive
  - Many ontologies will not maintained any more in the future
    - funded project ends $\longrightarrow$ no money any more for maintenance
    - some ontologies are not widely used
- Choose COVID-19 ontology with high impact
  - $\longrightarrow$ higher probability for being maintained in the future
  - $\longrightarrow$ higher interoperability with other applications

How to measure impact of COVID-19 Ontologies?

How to measure impact of COVID-19 Ontologies? Number of reuses in ontologies

- High number of reuses in other ontologies
  - is a sign for the popularity of the reused ontology and its classes,
  - let ontology users stumple over the reused ontology when applying the reusing ontology, and
  - makes more reusings more likely
  - Resulting Ranking:
    - Class COVID-19 of the Human Disease Ontology (DOID): 6 reuses
    - Class COVID-19 of the Mondo Disease Ontology (MONDO): 2 reuses
    - Class COVID-19 of COVID-19 Surveillance Ontology (COVID19), of Medical Subject Headings (MESH) and of National Cancer Institute Thesaurus (NCIT): 1 reuse

How to measure impact of COVID-19 Ontologies? Number of usages in projects

- Number of usages in projects
  - obvious measure for the ontology impact
  - in BioPortal is incomplete
    - more a sign for a good maintenance of the ontology & motivated ontology developers and project members pointing out these usages
    - remains as a good metric for the ontology impact
- Resulting Ranking:
  1. 32 projects: SNOMED CT
  2. 17 projects: NCIT
  3. 15 projects: MESH
  4. 11 projects: MedDRA
  5. 10 projects: DOID
  6. 8 projects: LOINC
  7. 5 projects: NIFSTD, EFO, ICD10CM
  8. 4 projects: MS
  9. 3 projects: VO
  10. 2 projects: MONDO, MEDLINEPLUS
  11. 1 project: COVID-19 Surveillance Ontology, CODO, VANDF

How to measure impact of COVID-19 Ontologies? Number of direct and indirect usages in projects

- Number of direct and indirect usages in projects
  - **direct usage:** the usage of a given ontology in projects
  - **indirect usage:** the project usage of an ontology reusing the given ontology
- Resulting Ranking:
  1. 32 projects: SNOMED CT
  2. 21 projects: NCIT
  3. 15 projects: MESH
  4. 15 projects: DOID
  5. 11 projects: MedDRA
  6. 8 projects: LOINC
  7. 7 projects: MONDO
  8. 5 projects: NIFSTD, EFO, ICD10CM
  9. 4 projects: MS
  10. 3 projects: VO
  11. 2 projects: MEDLINEPLUS
  12. 1 project: COVID-19 Surveillance Ontology, CODO, VANDF

How to measure impact of COVID-19 Ontologies? Weighted Combinations of Number of Reuses and Projects

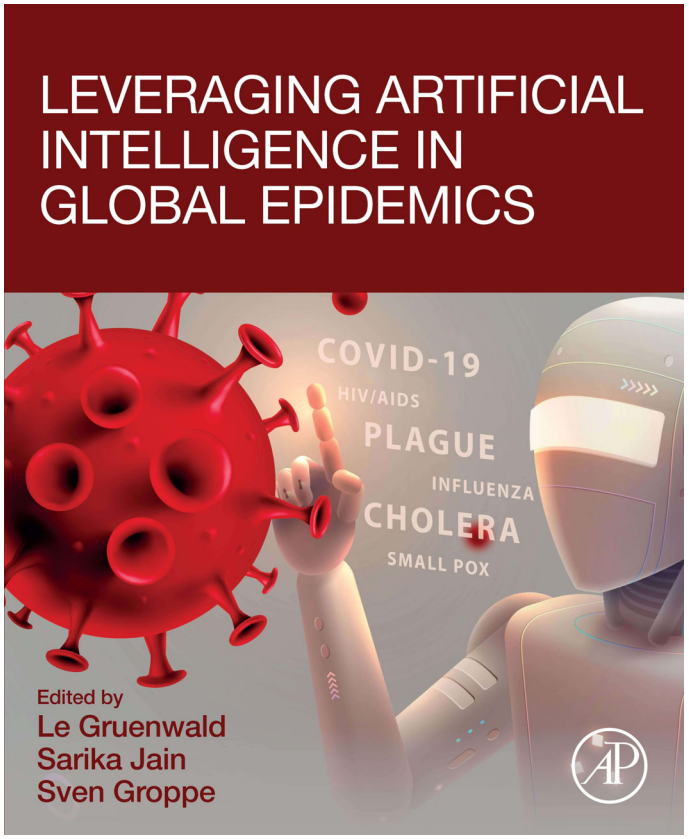
- Both metrics (#reuses and projects) are independent in theory, but
  - a high number in using projects often results in an increased number of reusing ontologies and vice versa in practice
  - BioPortal is incomplete $\rightarrow$ #reuses and projects are sometimes extremely different
- Idea: Calculating a balanced metric for these extreme cases
  - Finding a good balanced metric based on rigorous analysis open challenge for future work

How to measure impact of COVID-19 Ontologies? Open Challenges and Future Work

- Datasets (other than BioPortal) for usages of ontologies and projects
- Metrics based on other properties
  - Searches in ontology search engines
  - Number of instances of ontology classes in knowledge graphs
  - Number of applications using these ontologies
  - Number of accesses to instances of ontologies in applications
  - ...
- Impact of knowledge graphs and datasets

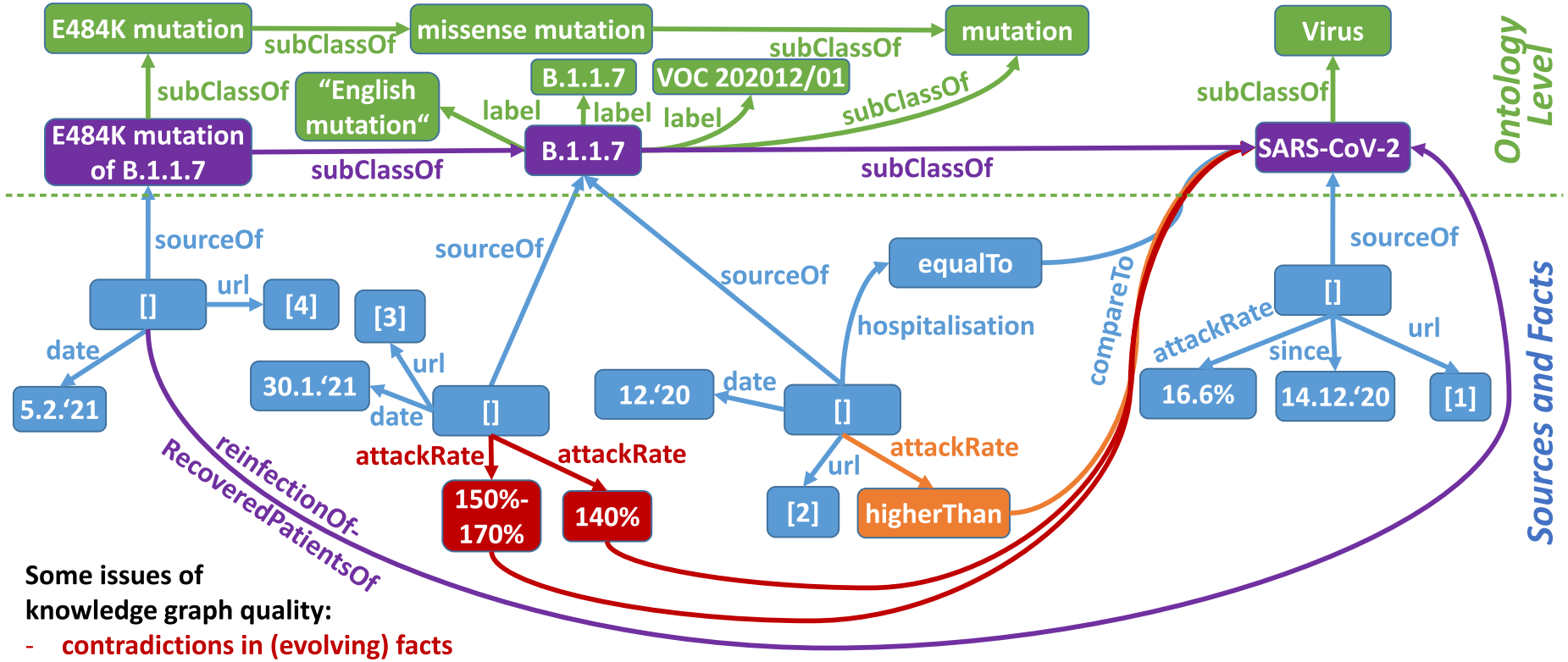
Further Reading

Table of Contents



COVID-19 KG Open Challenges:

For example: Vague Formulations ↔ NLP



Some issues of knowledge graph quality:

- **contradictions in (evolving) facts**
- **checking vague formulations and compare them with other given information**
- **errors with ambiguity: E484K mutation of B.1.1.7 is a SARS-CoV-2 virus, but reinfection with E484K is only possible for patients recovered from SARS-CoV-2 virus not mutating E484K (→ introduce class "SARS-CoV-2 without E484K mutation")**

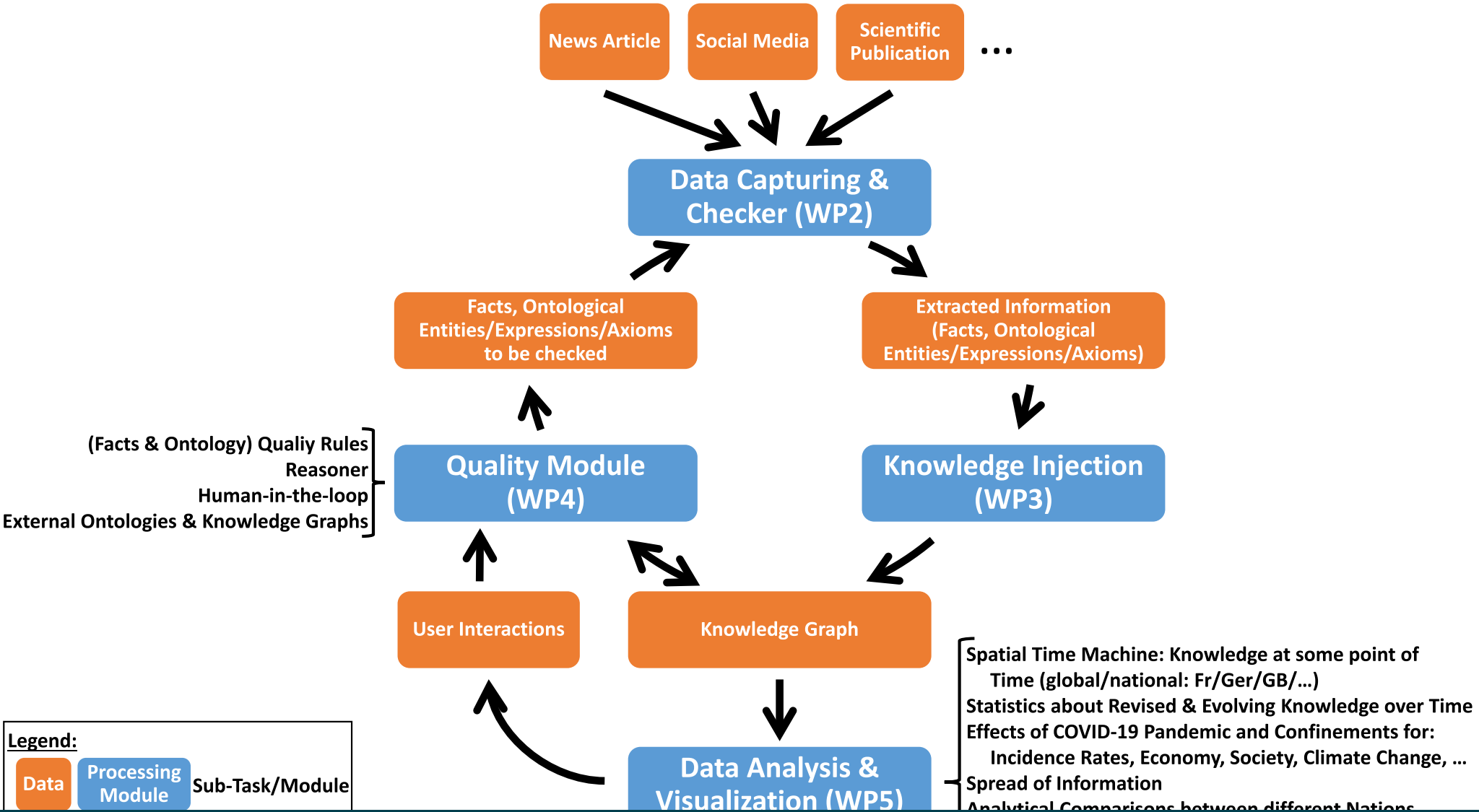
[1] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2774102
[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959361/Technical_Briefing_VOC202012-2_Briefing_2.pdf
[3] https://www.ruhr24.de/service/corona-britische-mutation-neue-studie-mutante-virus-toedlicher-sterberate-deutschland-90184403.html
[4] https://www.bmj.com/content/372/bmj.n359

High Quality Knowledge Graphs from recent English, French and German Emergent Trends with the example of COVID-19
Main Objectives:

- project starting soon (universities of Paris, Toulouse and Lübeck)
- Generate high-quality Kowledge Graph for emergent English, French and German trends with the example of COVID-19
  - and make the resulting graph publicly available
- Compare the facts extracted from different data sources
  - e.g., last year scientific publications, news articles and headlines, social media like twitter, existing KGs
- identify conflicting assertions as well as complementary ones
- Investigate differences between En/Fr/Ger data sources
- An extensive data analysis & visualization of research findings based on the time machine
- Evaluate the quality of the KG throughout the process of KG enrichment and its querying

This work is jointly funded by the DFG – Project-ID 490998901, and the French National Agency QualityOnt ANR-21-CE23-0036-01

High Quality Knowledge Graphs from recent English, French and German Emergent Trends with the example of COVID-19



This work is jointly funded by the DFG – Project-ID 490998901, and the French National Agency QualityOnt ANR-21-CE23-0036-01

Expected Results

- Data analysis and visualization tools in order to deal with the following issues:
  - extensive statistics about the COVID-19 pandemic especially with focus on facts evolving over time and differences of knowledge in different nations like:
    - number of contradicting and revised facts,
    - number of changes of COVID-19 confinements,
    - calculate effects (using machine learning approaches) of COVID-19 pandemic and confinements for incidence rates, economy, society and climate change.
  - a visualization tool for visualizing the results obtained by the statistics module by:
    - an easy-to-use tool, but
    - which offers flexible ways for querying the data in order to support sophisticated analysis,
    - with satisfactorily answering queries with consideration of users' requirements on data quality as well as the fitness of data to meet those requirements

This work is jointly funded by the DFG – Project-ID 490998901, and the French National Agency QualityOnt ANR-21-CE23-0036-01

Summary - COVID-19 Pandemic

- Statistics
  - Health: incidence rates, global health security index
- Timeline of discoveries and tech trends
- Predictions of incidence rates and other COVID-19 data in time series
- COVID-19 ontologies, knowledge graphs & data sets
  - Overview over existing ontologies, knowledge graphs and data sets
  - Quality assessment
  - Knowledge graph construction
- Further reading