DB Group @ unimore

UNIVERSITAS STUDIORUM MUTINENSIS ET REGIENSIS 1175

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Big Data Integration for E-Health

Data4SmartHealth 2021 October 27 2021
NOI Techpark Bozen-Bolzano

Prof. Sonia Bergamaschi
Department of Engineering "Enzo Ferrari"
sonia.bergamaschi@unimore.it
www.dbgroup.unimore.it

Outline

- Internet of Medical Things (IoMT)
- Big Data
- Big Data Technologies: Relational or NoSQL systems?
- Big Data Integration
- Data Science & Data Driven AI for Big Data
- Big Data Integration & Data Science in E-Health
- GDPR and Ealth Data

Internet of Medical Things (IoMT)

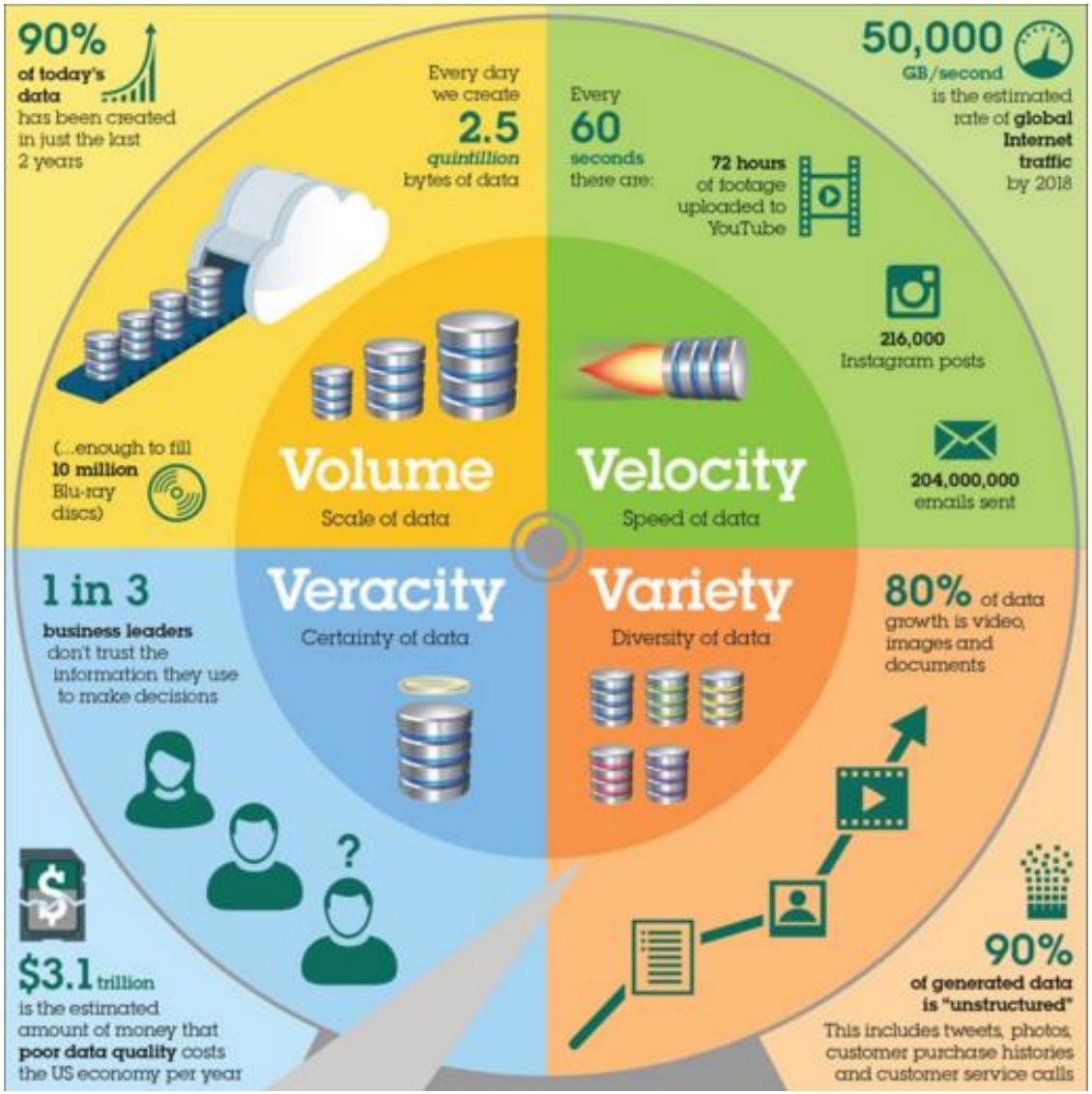
- The Internet of Medical Things (IoMT) is the collection of medical devices and applications that connect to healthcare systems through online computer networks
- IoMT allows to easily collect and monitor patients' parameters (e.g. by using wearable devices) increasing the quality of the healthcare services
  - Remote patient monitoring (telemedicine)
  - Use of NFC/RFID tags to monitor the stocks of medical supplies
  - Real-time patient location to reduce the intervention time
  - Integrate collected data to discover new knowledge
  - Use of Data Driven Artificial Intelligence to extract value from collected data

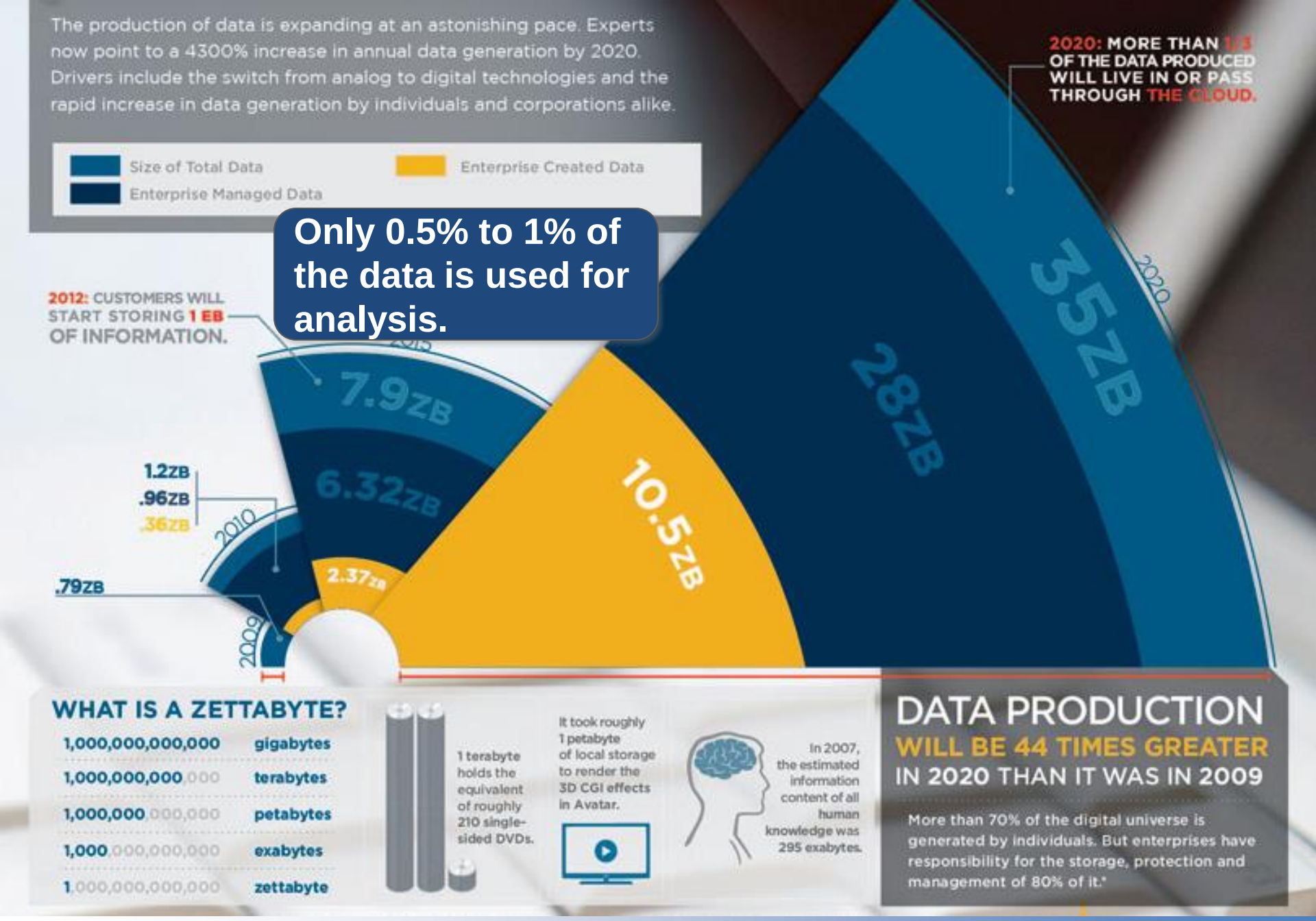
IoMT devices generate a huge amount of data, how to deal with?

Big Data - Full faith in the power of data

The quest for knowledge used to begin with grand theories.
Now it begins with massive amounts of data. **Welcome to the Petabyte Age**!

The FOUR V's of Big Data





http://www.csc.com/insights/flxwd/78931-big_data_growth_just_beginning_to_explode
http://www.guardian.co.uk/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume

Value – The most important V of all!

- Then there is another V to take into account when looking at Big *Data: Value!*
- Having access to big data is no good unless we can turn it into value.
- What technologies?

Technologies for Big Data

- Big Data Management
- Big Data Integration
- Big Data Science

Database Philosophy

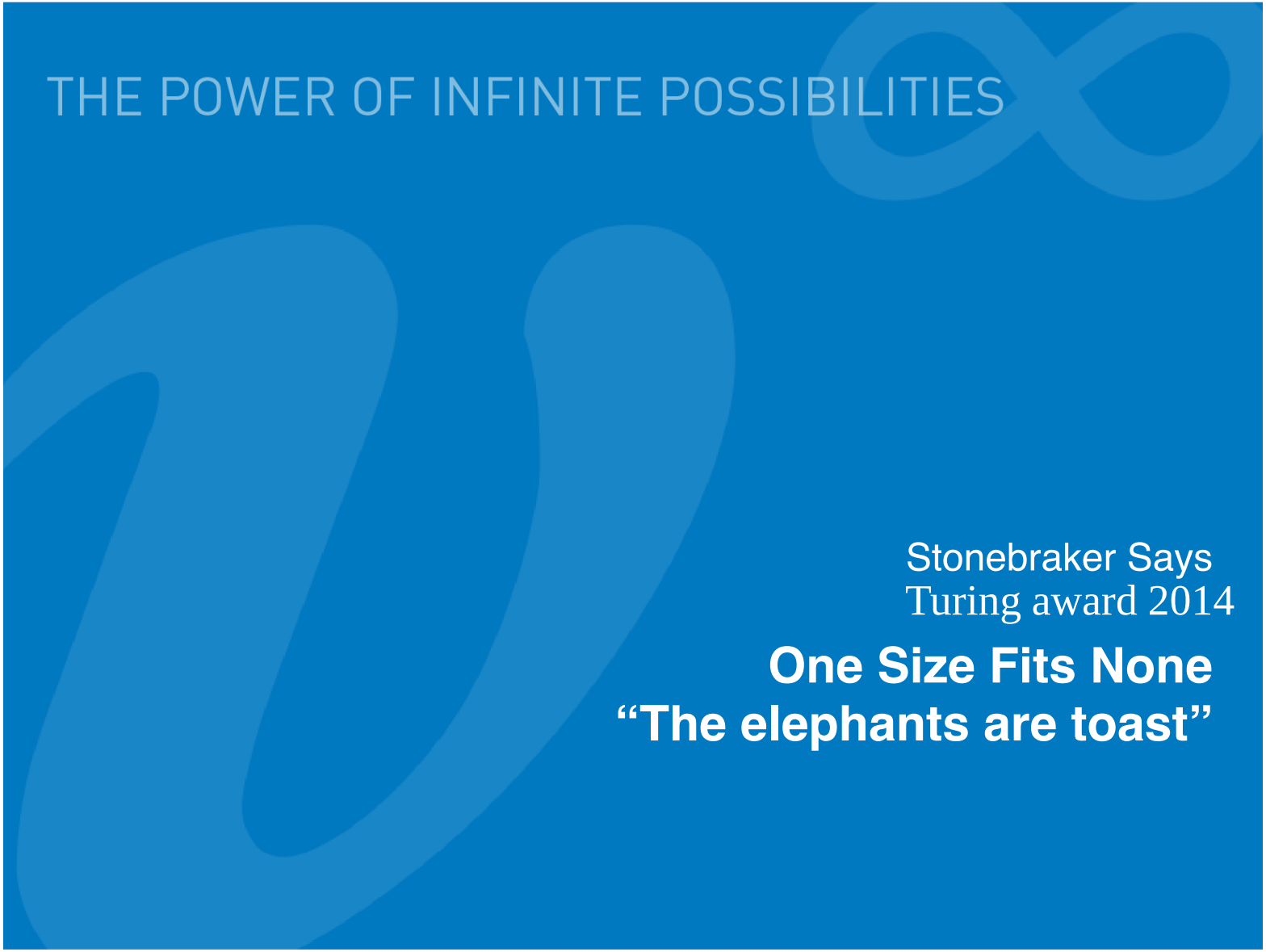
God made integers,
all else is the work of man.

(Leopold Kronecker, 19th Century Mathematician)

Codd made relations,
all else is the work of man.

(Raghu Ramakrishan, DB text book author)

Traditional RDBMS



At This Point, RDBMS is “long in the tooth”'

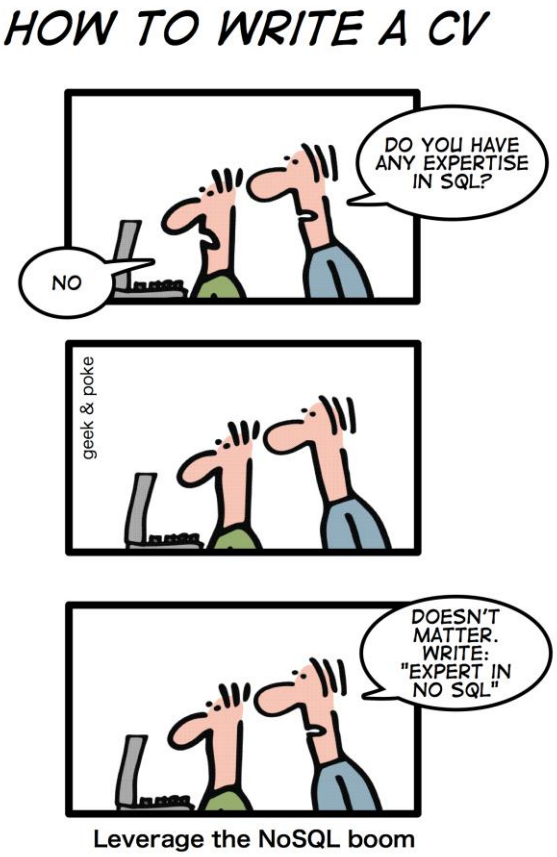
There are at least 6 (non trivial) markets where a row store can be clobbered by a specialized architecture !

Warehouse (Vertica, Red Shift, Sybase IQ, DW Appliances)
OLTP (VoltDB, HANA, Hekaton)
RDF (Vertica, et. al.)
Text (Google, Yahoo, …)
Scientific data (R, MatLab, SciDB)
Data Streaming (Storm, Spark Streaming, InfoSphere)

An emerging "movement" around non-relational software for Big Data

- NOSQL stands for "Not Only SQL" (but is not entirely agreed upon), where SQL doesn't really mean the query language, but instead it denotes the traditional relational DBMS.

- Google **Bigtable & Mapreduce**, **Memcached**, and Amazon's **Dynamo** are the "proof of concept" that inspired many of the NOSQL systems:
  - Memcached demonstrated that in-memory indexes can be highly scalable, distributing and replicating objects over multiple nodes
  - Dynamo pioneered the idea of *eventual consistency* as a way to achieve higher availability and scalability
  - BigTable demonstrated that persistent record storage could be scaled to thousands of nodes & Mapreduce introduces parallel computation for distributed data platforms.



Challenges (1) – Selection of the Big Data Technology

- **Volume, Velocity**

  Calling for new **Big Data systems:**

  – **Big Data Management Systems: *NOSQL & more***

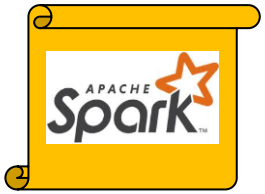
  

  Many more…

  – **Big Data Analysis Systems:**

  - **Batch + Streaming**

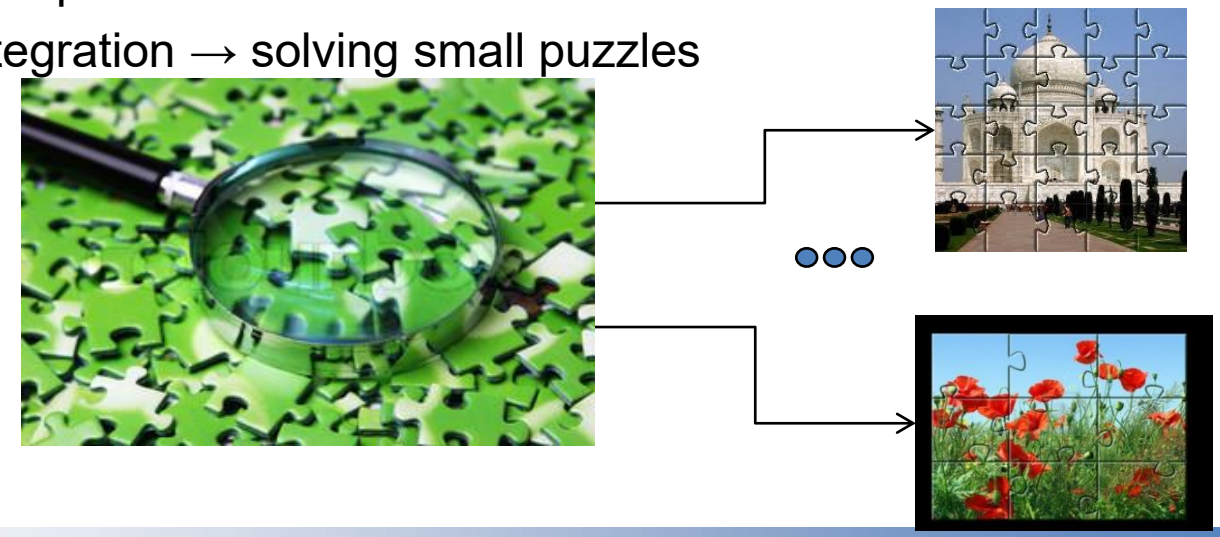
  

  Many more…

  Not only Relational Database Management Systems and Business Intelligence

Data integration as a new technological solution

- The discipline of data integration comprises the practices, architectural techniques and tools that ingest, transform, combine and provision data across the spectrum of information types in the enterprise and beyond in order to meet the data consumption requirements of all applications and business processes.
- Applications of Data Integration
  - Business, Science, Government, The Web, Health, Pretty much everywhere
- Small Data integration = solving lots of puzzles
  - Each puzzle (e.g., Taj Mahal) is an **integrated entity**
  - Each piece of a puzzle comes from some **source**
  - Small data integration → solving small puzzles



Data integration as a new commercial software

- ✓ **According to Gartner:**
  Gartner estimates that the data integration tool market generated more than $2.7 billion in software revenue (in constant currency) at the end of 2016.

- ✓ A projected five-year compound annual growth rate of 6.32% will bring the total market revenue to around $4 billion in 2021 (see "Forecast: Enterprise Software Markets, Worldwide, 2014-2021, 2Q17 Update")
- ✓ ***$3.3 billion software revenue in 2020.***

Market Overview:

- ✓ The biggest change in the market from 2016 is the pervasive yet elusive demand for metadata-driven solutions.

- ✓ Consumers are asking for hybrid deployment not just in the cloud and on-premises but also across multiple data tiers throughout broad deployment models, plus the ability to blend data integration with application integration platforms (which is metadata driven in combination with workflow management and process orchestration) and a supplier focus on product and delivery initiatives to support these demands.

Data integration in the research community

- The research community has been investigating data integration for more than 30 years: different research communities (database, artificial intelligence, semantic web) have been developing and addressing issues related to data integration:
  - Definitions, architectures, classification of the problems to be addressed
  - Different approaches have been proposed and benchmarks developed
- **Open issues**
  - Uncertainty, **Provenance**, and Cleaning
  - Lightweight Integration
  - Visualizing Integrated Data
  - Integrating Social Media
  - **Big Data Integration**

MOMIS
Mediator environment for Multiple Information Sources

BDI: Why is it Challenging?

- Data integration = solving lots of puzzles
  - Big data integration → **big messy** puzzles
  - E.g., missing, duplicate, damaged pieces

Big Data Science

Data Analysis (Business Intelligence, Statistics, Data Mining, Math)

+

DataDriven Artificial Intelligence)

Data Driven AI

- From the Big Data era people do not focus on improvement the quality of data, but just add more data to overcome errors from noisy and poor-quality information;
- From a recent talk Andrew Ng (https://it.wikipedia.org/wiki/Andrew_Ng) states that 99% of the papers were model-centric;
- As results many models do not work well on real data;
- A recent paper from Google researchers analyzes the work of 53 AI practitioners reports that *"data cascades—compounding events causing negative, downstream effects from data issues—triggered by conventional AI/ML practices that undervalue data quality… are pervasive (92% prevalence), invisible, delayed, but often avoidable."*

Model-Centric	Data-Centric
- Collect as much data as possible - Iteratively **improve the model** to deal with the noise in the data	- Hold the model fixed - Iteratively **improve the quality of the data** to obtain good results

Data Driven AI

- **Data education lack** of adequate training on AI data quality, collection, and ethics. AI courses focuses on toy datasets with clean values, but AI in practice required the creation of data pipelines, often from scratch, going from groundtruth to model maintenance.

- We have to define a **systematic pipeline** to improve the quality of data;
- Systematic **improvement of data quality** on a basic model is better than using the state-of-the-art models with low-quality data;
- In his recent talk Andrew Ng states that **good data** for ML/AI is:
  - Defined consistently (the label definition is unambiguous)
  - Cover important cases (good coverage of inputs)
  - Has a feedback from the production data
  - Sized appropriately

Big Data Integration in Healthcare

Data Analysis + Data Driven AI

Mobile devices

Information systems

Web Applications

Medical devices

Diagnostic systems

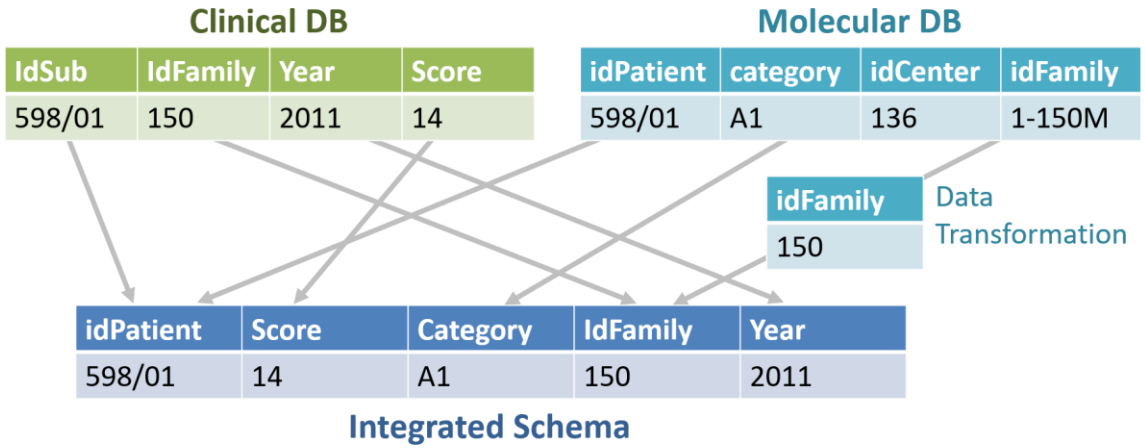
NFC/RFID *tags*

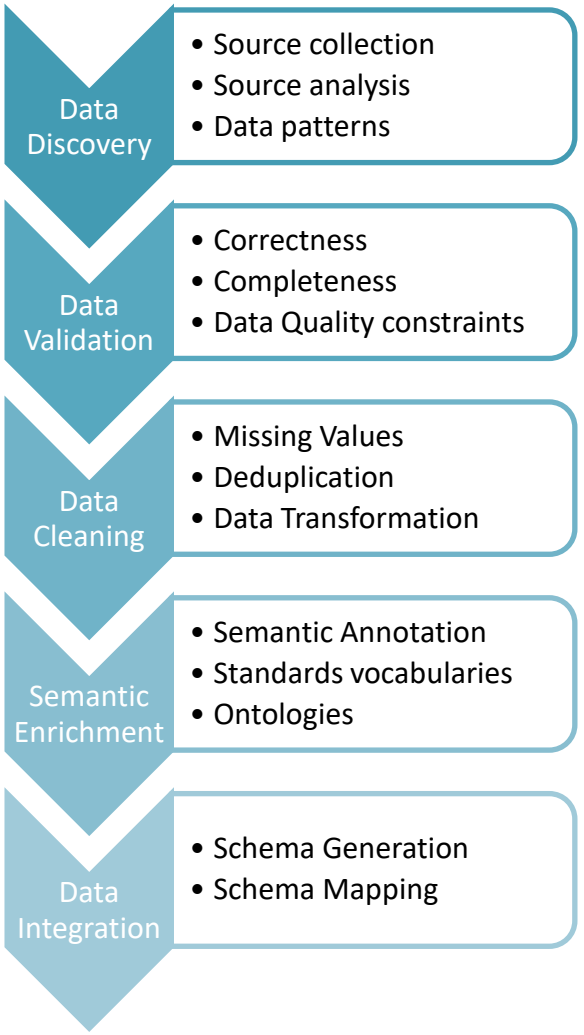
Wearable devices

MOMIS
Mediator envirOnment for Multiple Information Sources

Data Integration with MOMIS

MOMIS is a data integration system able to aggregate data from heterogeneous (structured and semi-structured) and distributed sources (EHR, PHR, eCRF, ePRO, medical devices) in a semi-automatic way, exploiting the **semantic relationships** existing in the data sources (available as open source by Datariver www.datariver.it).





DataRiver
contract research organization

- 2009 – Founded as Spin-Off of the University of Modena and Reggio Emilia
- 2011 – Accreditation as **Industrial Research Lab of the High Technology** Network of the Emilia-Romagna Region
- 2017 Self-certification as **CRO** (Contract Research Organization) by Italian Medicines Agency (AIFA) for **Data Management** and **Statistical Analysis**
- 2017 **Digital Innovation Award in Health** - Politecnico of Milano Observatory
- 2019 **EU INNOLABS Acceleration Programme Award**
- 2021 **EU INNO4COV-19 Funding**
- 2021 Certification as **EHDEN Certified-SME**

DataRiver
contract research organization

Mission

- Provide innovative solutions for Clinical Trials, Patient Support Programs (PSP), pathology and rare disease registries to Pharma and Biotech companies, Clinical research institutes and Hospitals

- Specialized in designing and developing Web and Mobile software solutions for **Clinical Data Management, Big Data Integration & Analytics, Internet of Medical Things (IoMT)**.

MyHealth is a Web and Mobile IoTM platform for monitoring and **improving quality of life of individuals**

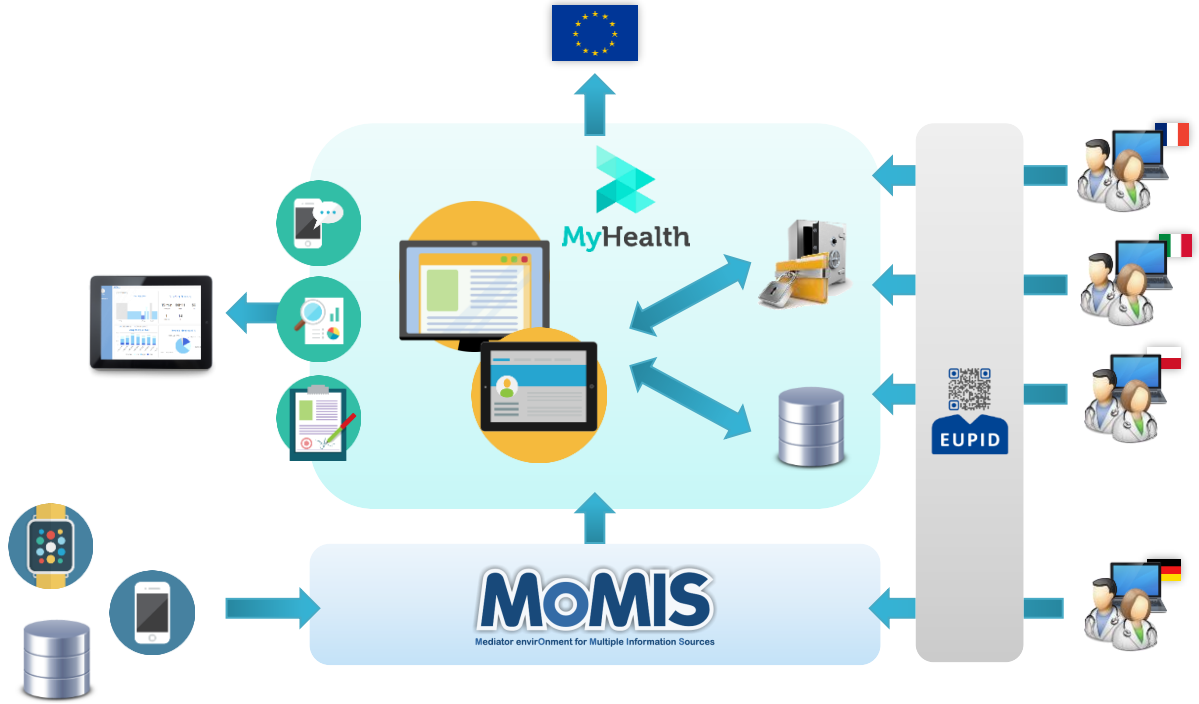
- MyHealth active projects:
  - Patients with **specific pathologies**
    12 Clinical trials, 4 Patient Support Programs (PSP)
  - Monitoring **elderly people**
    16 residential care & nursing homes
  - **Promoting multi-aging physical activity**
    1 research project on open air recreational activities
  - MyHealth-COV **EU INNO4COV-19 Funding**

INNO4 COV-19

PARTNER – Paediatric Rare Tumors Network

MOMIS system for virtual data integration of heterogeneous, fragmented and not standardized data from **Biobank databases, existing registries and other distributed data sources**.

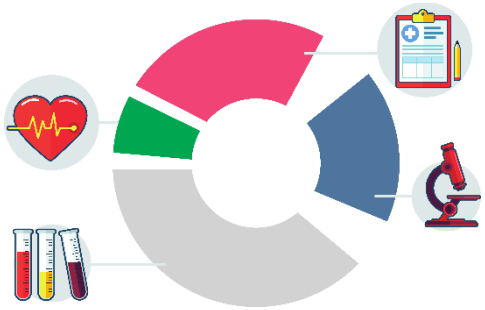
- ✓ secure access to the information **without migration of data**;
- ✓ Elaboration of **indexes** that process the patient outcomes;
- ✓ increase **information** and **knowledge** about diseases and patients.



Partner
Paediatric Rare Tumours Network - European Registry

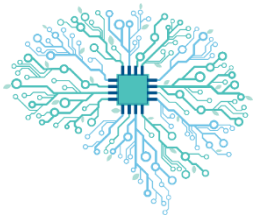
Decision support tool

AI and ML algorithm applied to the integration of radiological images with clinical and laboratory data to **identify new information and relationships between data** that can subsequently be refined and extended to further sources to expand and increase the **prognostic value** of the platform.



Predictive models

Artificial Intelligence & Machine Learning Algorithm

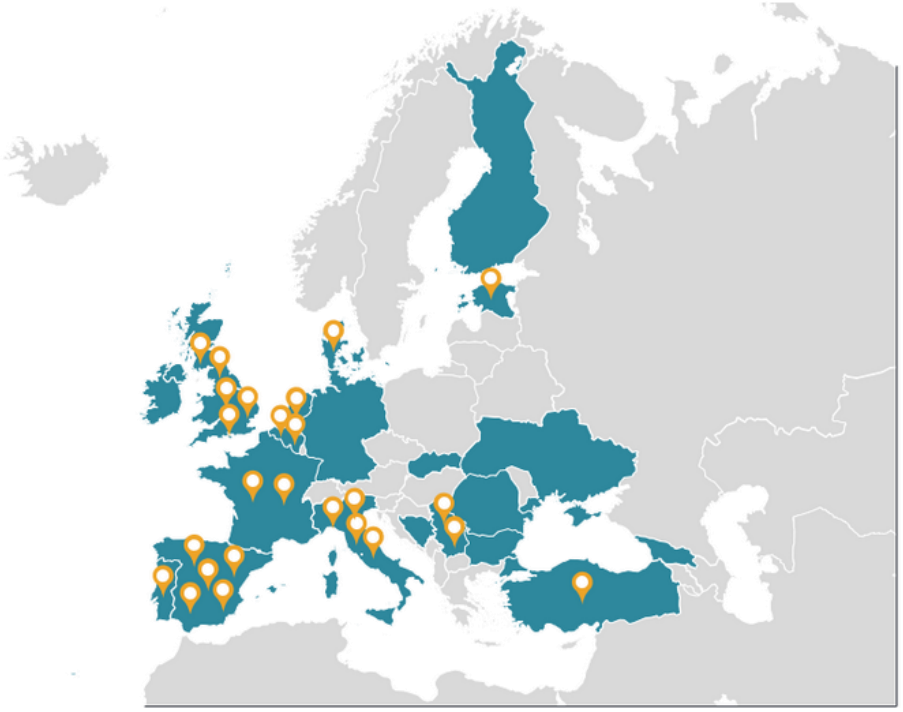


Radiomic data

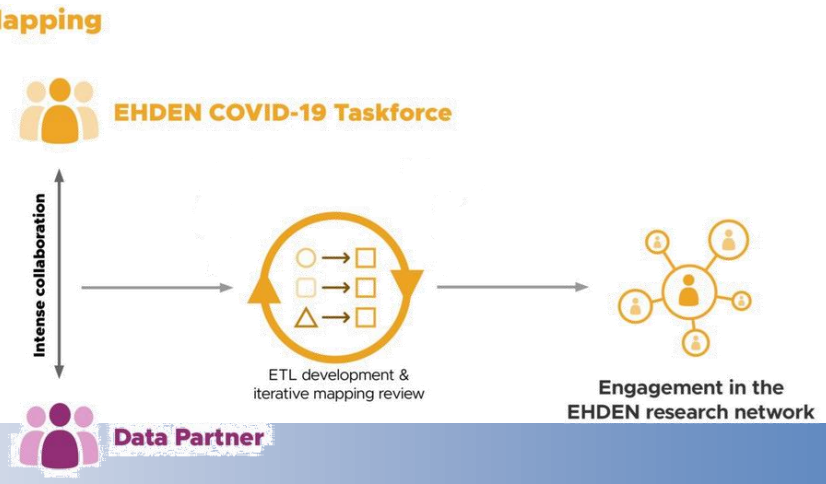
Clinical data

Probio platform

EHDEN – *COVID19 Rapid Collaboration Call*



- ✓25 Data Partners
- ✓Over 1 million SARS-COV-2-tested patients
- ✓228,000 of whom tested positive



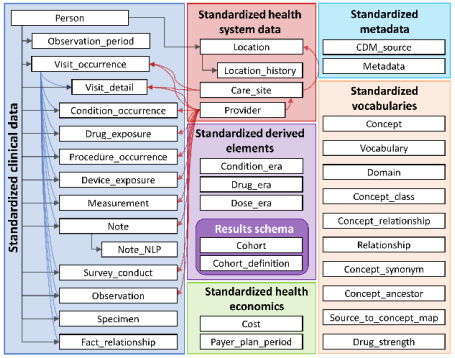
EHDEN – *COVID19 Rapid Collaboration Call*

Data from Hospitals' departments:

- *Infectious Diseases*
- *COVID*
- *Intensive and Sub-intensive Care Units*

MoMIS

Mediator envirOnment for Multiple Information Sources

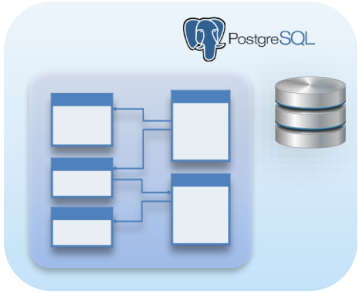


OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

OMOP Common Data Model (OHDSI) Semantic Integration

Data and metadata repository



Common interface to share data and metadata

EHDEN

EUROPEAN HEALTH DATA & EVIDENCE NETWORK

MyHealth-COV

Assist home care patients with COVID-19 and during the follow-up providing a remote and continuous monitoring of their parameters and tele support, improving the intensive care and fight the pandemic.

- **Telemonitoring system**: patient remote monitoring using wearables devices to collect patient physiological parameters and a mobile App for ePRO questionnaires;
- **AI data analysis** for multidimensional monitoring to engage and support patients at home at the right time;
- **Vocal assistant tool** for patient support and engagement;
- **Televisit system** for medical staff-patient video visits;
- **Web and mobile App** for doctor-patient communication.

MyHealth-COV

Videovisit and Tele support

Patients Mobile App
Patient Reported Outcomes (ePRO)

MyHealth-COV

Medical devices to collect physiological data

Electronic Health Records (EHR)

Doctor-patient communication Vocal Assistant & ChatBot

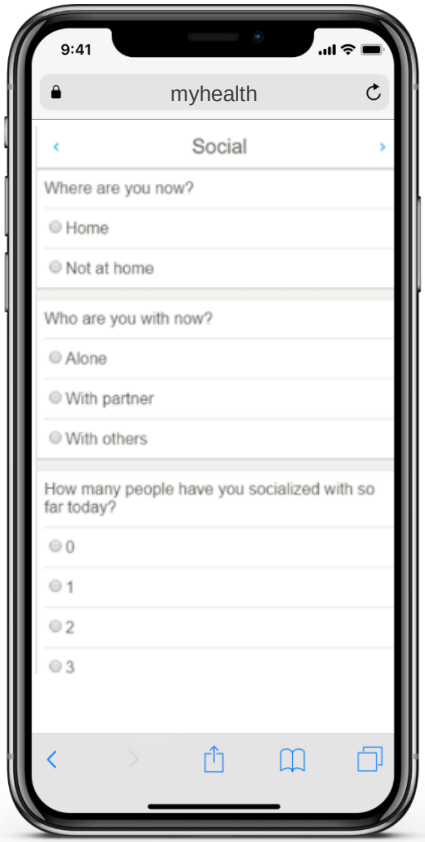
INNO4 COV-19

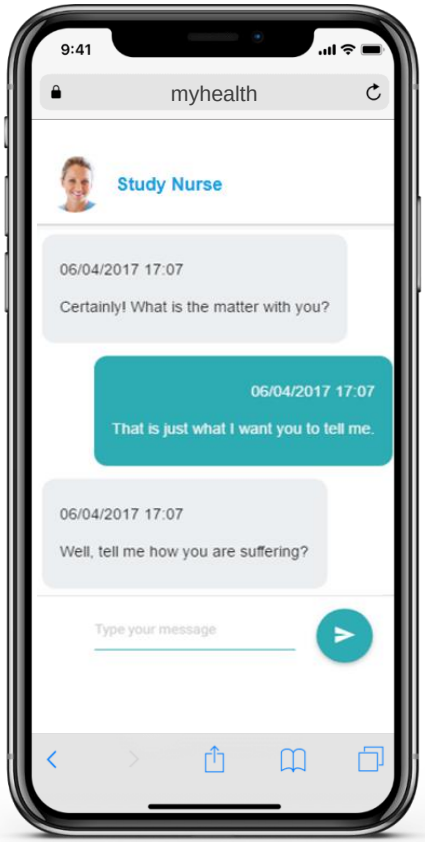
Devices data

- Oxigen Saturation (SpO2), Breathing
- Diastolic and systolic blood pressure (mmHg)
- ECG and Detection of atrial fibrillation (AFib)
- Heart rate (BPM avg, max, min)
- Duration and quality of sleep, sleep apnea (n. of awakenings, deep and light sleep)
- Glycemia
- Physical Activity (walk, run, sleep, cycling), number of steps, Calories burned

MyHealth-COV

User Mobile App





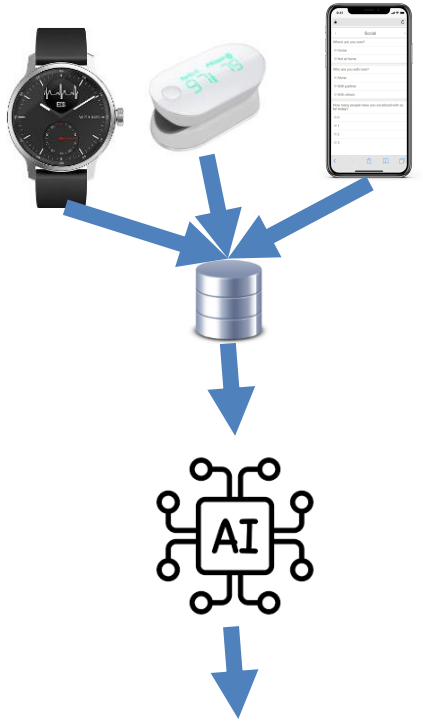
INNO4 COV-19

Clinician/Nurse Web Application

Data driven AI

MyHealth-COV

a) *Personalized support for each patient*



- **Input**: integrated data collected by wearable devices and mobile app's forms

- **Output**: personalized feedback for each user:
  - i. Personalized contents:
    - » Level of details based on the patient's profile
    - » Action Recommendation based on her status

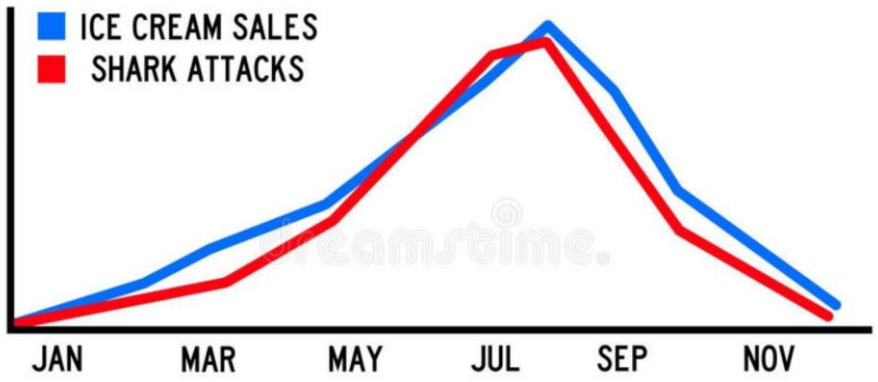
  - ii. Personalized notifications:
    - » customized time, frequency

Data driven AI

MyHealth-COV

AI is not the silver bullet

- **Some tasks <u>require human-in-the-loop</u> to read the data correctly**



b) Prioritization of the patients for clinical users

- **<u>Input</u>**: integrated data about all the monitored patients
- **<u>Output</u>**: priority of patients and best channel of interaction
  i. <u>Priority for each clinical user</u>: what patient's data to review so to take action

GDPR & Ealth Data

GDPR and health data

- Personal data **concerning health** should include all data regarding to the health status of a subject such as:
  - **Information about the person** collected for provisioning health case services;
  - **Clinical** and **biometric data** derived from tests or examination of the patient, including biological samples;
  - Any information on e.g., disease, disabiliy, medical history, clinical treatment, disease risks, etc.
- These kind of data are considered **sensitive data** by the General Data Protection Regulation (GDPR art. 51-52).

GDPR and Secondary Use of Health Data

- Access to health data is crucial for medical research
  - Large volume of data is more likely to **provide compelling and robust evidence**;
  - The **integration** of more data sources helps the **improvement** of **disease diagnostic** techniques and treatments;
  - Exchange of data in the context of multi-national clinical trials ensures that **conclusions are valid** for different groups of people, avoid misleading conclusions;
- GDPR impose **limitations** on sharing and reuse of health data for research purpose;
- **Processing personal health data is lawful** only when:
  - The data subject given the **explicit consent** to the processing of personal data for one or more specified purposes;
  - **vital interests** of the data subject are protected;
  - Reasons of **public interest in the area of public health exist (**for instance ensuring high standards of quality and safety of health care and of medicinal products or medical devices).

GDPR – Data anonymization

- **Pseudonymization** means that personal data is processed in a manner that it can no longer be attributed to a specific data subject without the use of **additional information**;
- Personal data which have undergone pseudonymization is still **considered personal data**, meaning that are still protected by GDPR;
- **Pseudonymization ≠ Anonymization**
  - An information is anonymized when does not relate to an identified or identifiable person or regards to personal data rendered anonymous in such a manner that the **data subject is not or no longer identifiable**.
  - **GDPR does not concern anonymized information** which can be used for statistical research purposes.

Anonymization techniques

- **Data masking**: hiding data with altered values. E.g. by using shuffling, encryption, character substitution. Data masking makes reverse engineering impossible.
- **Generalization**: removes some of the data to make it not identifiable. E.g. removes the house number in an address. The purpose is to eliminate identifiers while retaining data accuracy.
- **Shuffling and permutation**: rearrange the attribute values so they do not correspond with the original record.
- **Data perturbation**: applying techniques that round numbers and add a random noise. A small noise may lead to weak anonymization while a large one can reduce the utility of the dataset.
- **Synthetic data**: algorithmically manufactured information that has no connection to real events. The process involves creating statistical models based on patterns found in the original dataset. E.g. by using standard deviations, medians, linear regression or other statistical techniques.

GDPR – Data Retention and Transparency

- Personal data should be kept (in a form which permits identification) for **no longer than is necessary for the purposes of the processing**
  - A time limit should be established for **erasure/ anonymization** of the data
- GDPR states that personal data should be processed **lawfully, fairly and in a transparent way.** Transparency means that any information regarding the processing of personal data must be:
  - Easily **accessible**;
  - Easy to **understand**;
  - **Clear** and written in **plain language**.

GDPR – Rights of data subjects

- Right of **access** to personal data which have been collected
  - For example, medical records such as examinations results, diagnoses, treatments, etc.
- Right to **rectification**
  - Wrong or missing data must be fixed without delays;
- Right to be **forgotten**
  - Personal data must be deleted, for example if data subject withdrawn her consent to data processing.

How to protect privacy while doing Data Integration?

- European Patient Identity Management
  - prevent duplicate registration of patients
  - avoid creating a transparent universal patient ID but
  - provide distinct pseudonyms for patients in different contexts
  - preserve the possibility for re-identification by a trusted third party
  - keep a protected link between the different pseudonyms in the background
  - which supports creating merged, datasets for secondary use