CODATA-RDA Schools for Research Data Science

- Very large open datasets for ML in Personalized Medicine

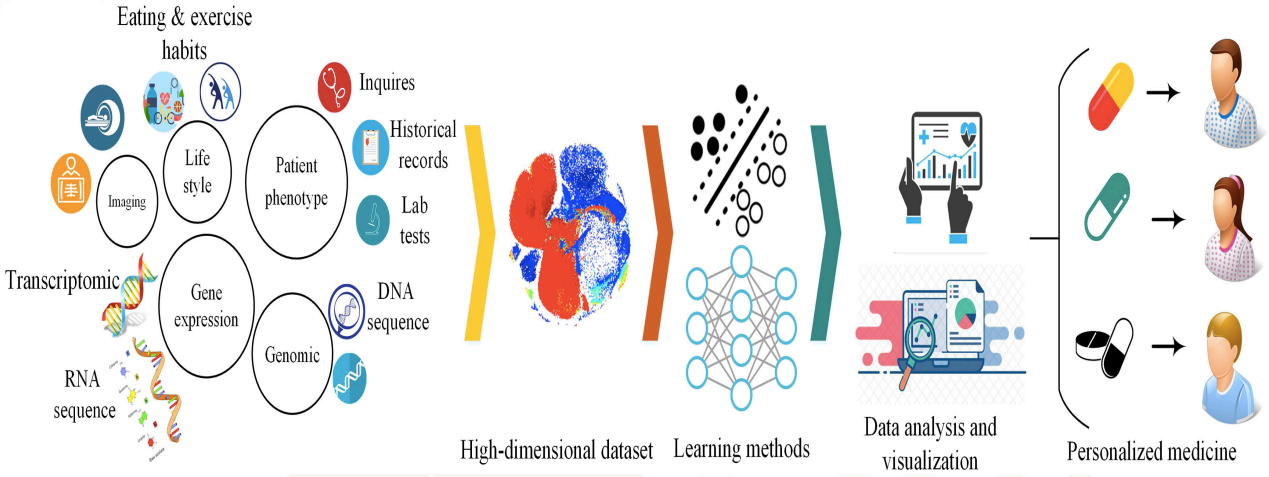
“The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no 101017536””.

Isabelle Perseil, EOSC Future Domain Ambassador

Roadmap

- AI for Personalized medicine landscape
- Some research data issues in Personalized medicine
- Open science benefits for Personalized medicine
- RDA actions
- EOSC Future Domain Ambassador (RDA open call)
- Altogether
- Next steps

AI for Personalized Medicine landscape



AI for Personalized Medicine landscape

- It is now possible
  - to rapidly sequence the entire DNA (a genome)
  - to measure tens of thousands of biomolecules in the human body
  - to use sensory sensors to continuously monitor an individual's physical activity
  - to characterize the communities of microbiota that colonize his or her gut.

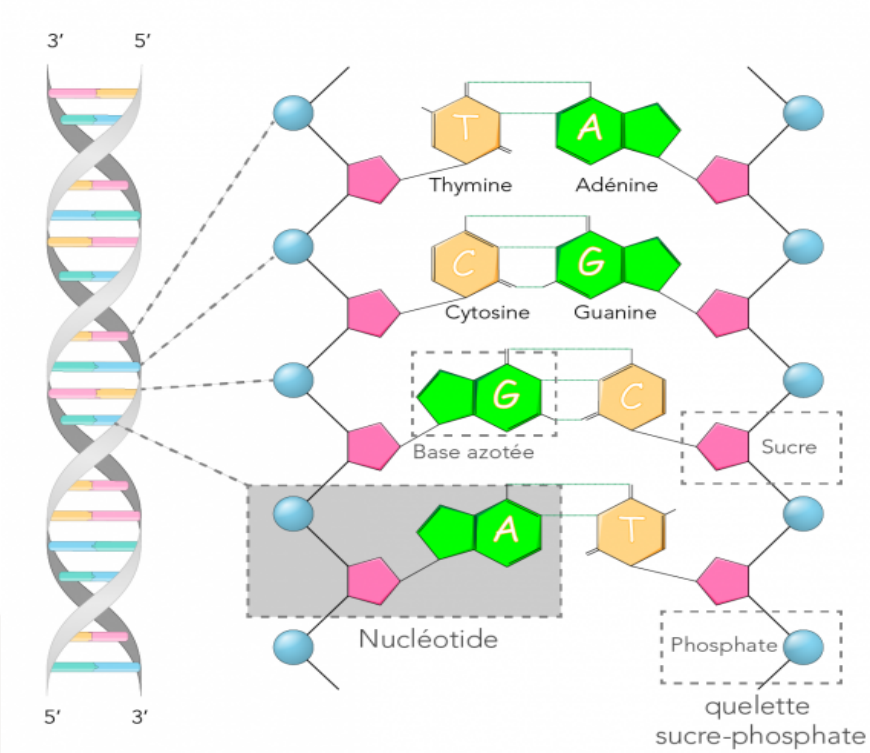
All these measurements provide **specific information** on a person's health. The data sets collected must be properly integrated and interpreted.

AI for Personalized Medicine landscape

- The prospects are that all this information gathered will be able to guide diagnosis (as well as lifestyle habits) based on experiments carried out on individuals **with comparable characteristics.**

- There is genomic information that we know how to interpret, and other information that we don't, and interpretations that unfortunately turn out to be wrong over **time**.

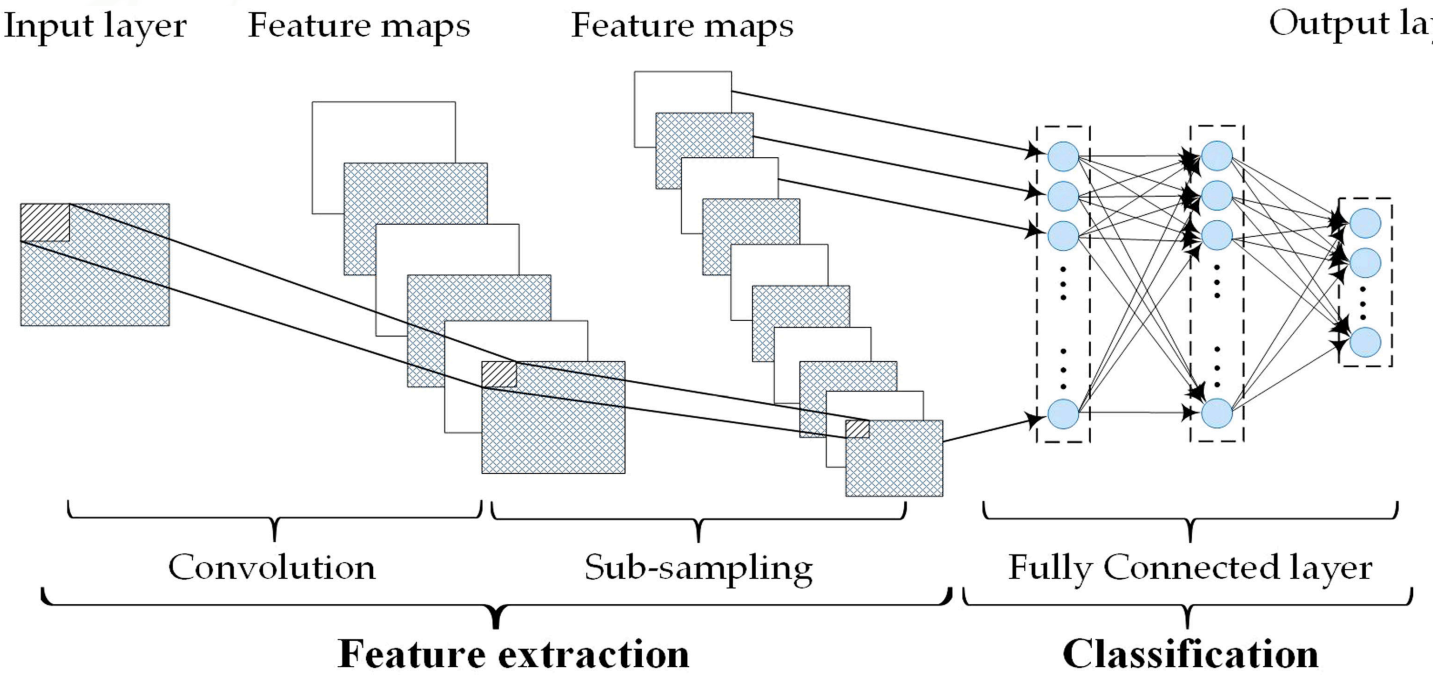
AI for Personalized Medicine landscape



AI for Personalized Medicine landscape

- Deep learning (DL) is a representation learning method, which enables the system to discover the features needed for a specific task, from raw data, by building multiple layers of processing with non-linear operations to learn representations at different levels of abstraction.
- **CNN models** are suitable for processing visual and other two-dimensional data such as images and speech.
- Inspired by animal visual cortex, the standard building blocks of CNNs comprise one or more layers divided into three types:
  - convolution layers with a set of learning filters to generate two-dimensional feature maps, each responding to a specified local model of the input,
  - one or more pooling layers for subsampling and to make the representation translation-invariant,
  - and one or more fully connected layers attached to the end of the network for high-level reasoning.

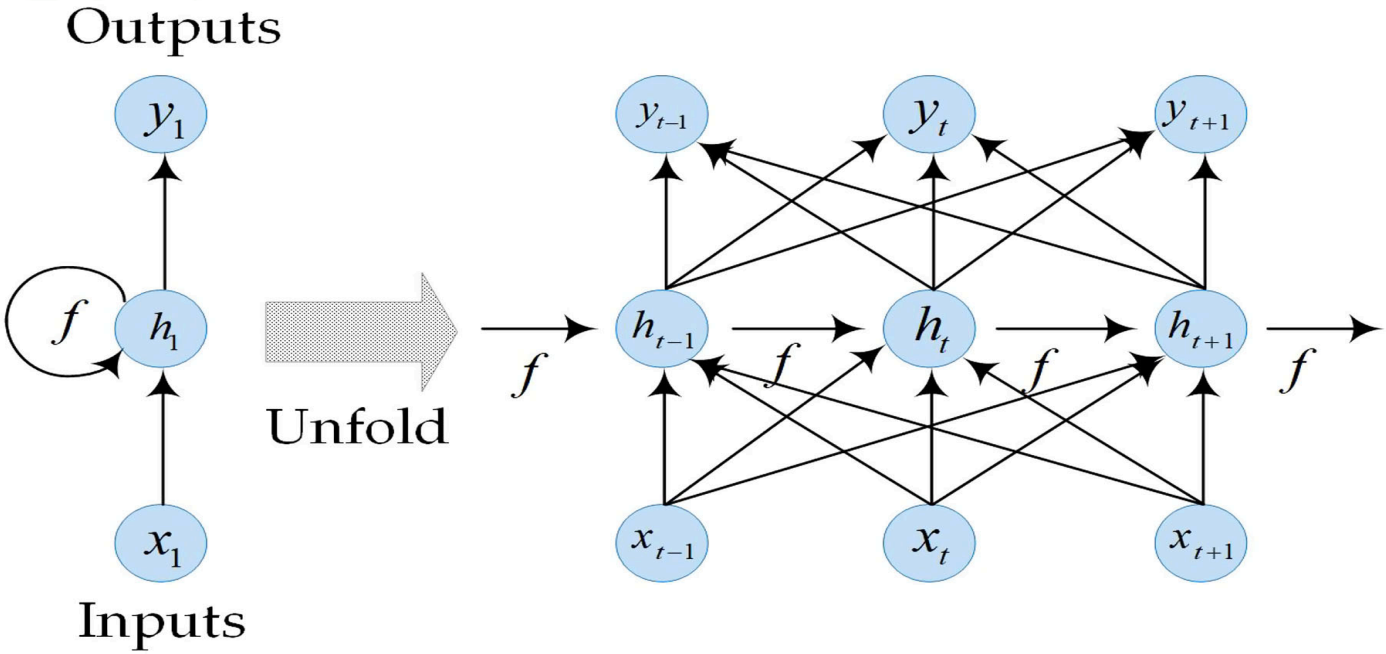
AI for Personalized Medicine landscape



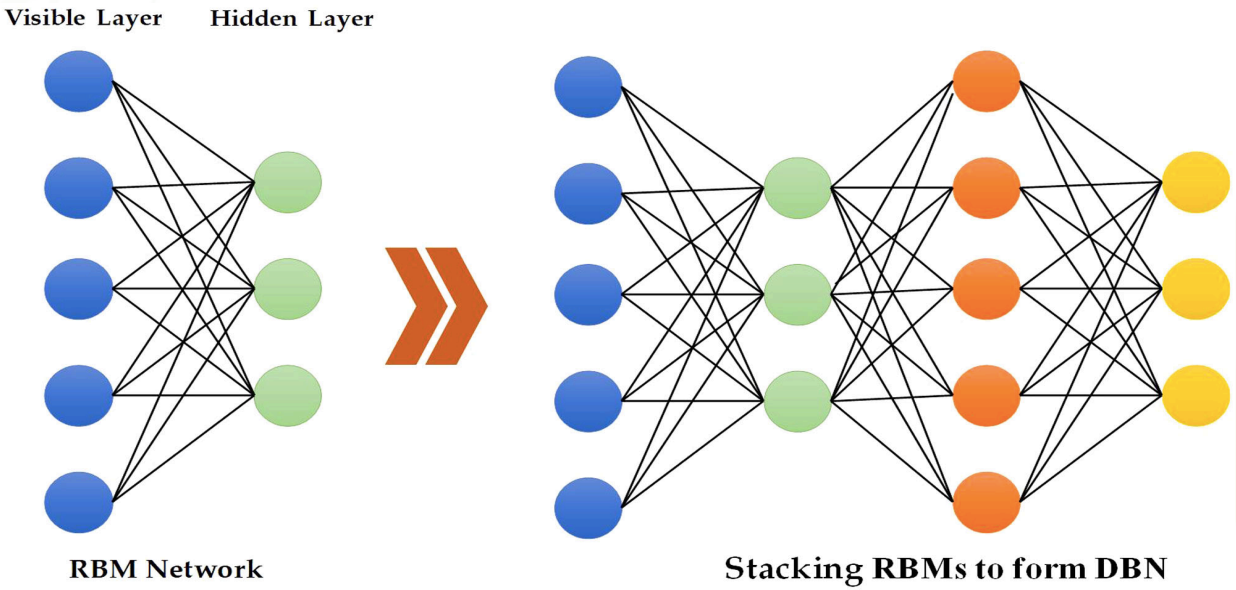
AI for Personalized Medicine landscape

- RNNs have shown superior results in sequential data such as NLP.RNNs differ from other feedforward networks in their cyclic connections, which form feedback loops in the hidden layers, allowing historical information to be stored in the hidden states of the RNN.
- Consequently, RNNs are particularly suited to processing temporal data and using sequential information.
- However, the ability of conventional RNNs to process sequences over a long period of time is rather limited due to gradient explosion (combinatorial) problems.
- LSTM and GRU networks excel at processing very large sequential data sets, which are effective at capturing long-term dependencies.

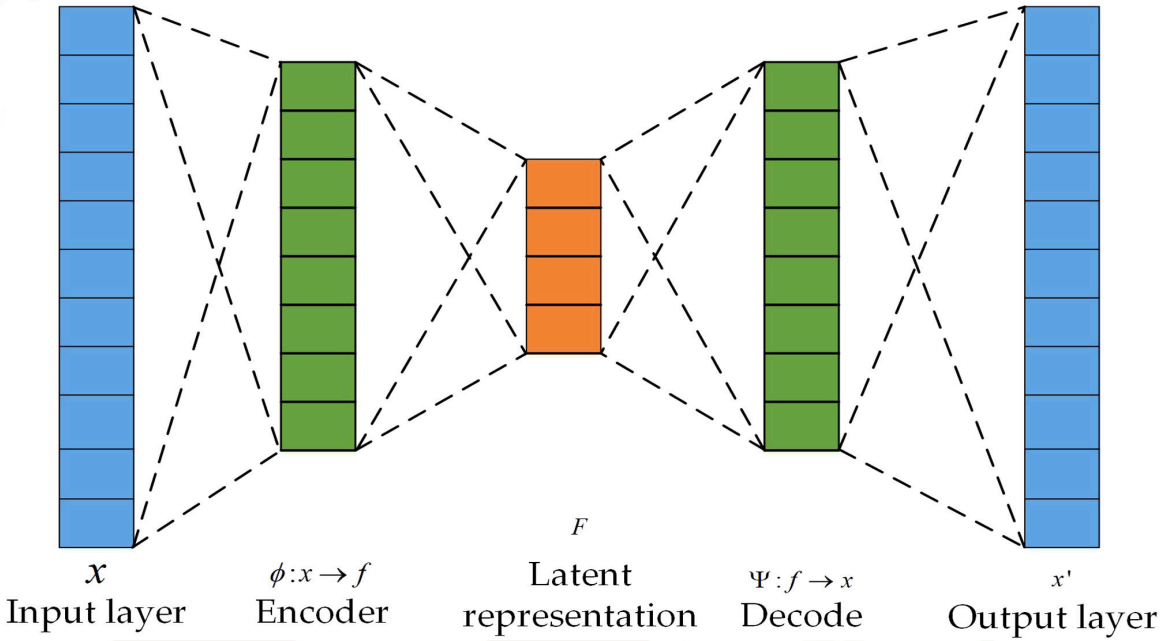
AI for Personalized Medicine landscape



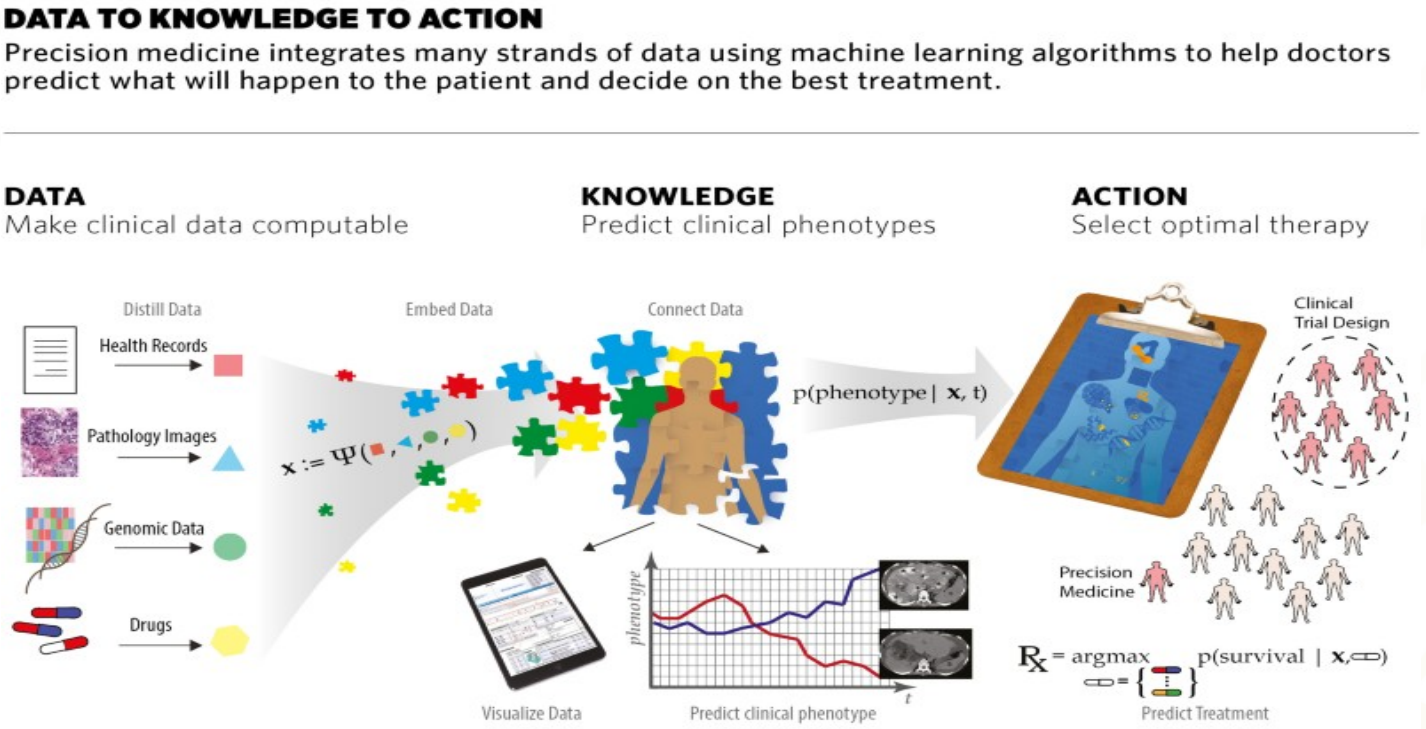
AI for Personalized Medicine landscape



AI for Personalized Medicine landscape



Some research data issues in Personalized Medicine we want to solve



Some research data issues in Personalized Medicine we want to solve

Sensitive data

- *Information that is regulated by law due to possible risk for plants, animals, individuals and/or communities and for public and private organisations.*
- There are a number of barriers which need to be overcome before sensitive data can be utilised safely and to its best advantage. One major challenge is that not all sensitive data is alike, with significant disciplinary variation in how sensitive data is defined, linked, managed, stored, and reused
- Sensitive data requires careful stewarding such that it can be disseminated in an ethically and culturally appropriate way

Some research data issues in Personalized Medicine we want to solve

Policies harmonization

- Promote and improve pan-European policies, standards and risk analyses to ensure full regulatory compliance of data (within the EOSC), especially sensitive or classified data.
- Propose and improve standardised solutions for working with multi-source (partitioned) environments, where datasets are not directly available for integration and metadata may not even be disclosed in detail.
- Propose solutions enabling secure federated storage, access and sharing of sensitive research or health data.
- Promote the integration and knowledge of privacy-enhancing technologies.

Some research data issues in Personalized Medicine we want to solve

Semantic interoperability

- Ability to share data between systems and ensure understanding at the concept's level of the domain
- Different health standards aim to enable data sharing among healthcare organizations. However, the adoption of standards still presents several challenges to achieving interoperability at the semantic level.
- Therefore Task forces are working at interoperability recommendations, based on minimum (meta)data set and interoperability indicators and are providing some tools to ensure this interoperability

Some research data issues in Personalized Medicine we want to solve

FAIRification

- What are the requirements needed to make Medical research data and infrastructures Findable, Accessible, Interoperable and Reusable according to the FAIR Principles
- There is an urgent need for a service, like FAIRsharing, which enhances the information available on the evolving constellation of heterogeneous standards, databases, repositories and knowledge bases,
  - that guides users in the selection of these resources, educates policy makers, such as publishers and funders, to recommend the relevant resources to their authors and awardees,
  - and works with developers and maintainers of these resources to foster collaboration and promote harmonization.

Some research data issues in Personalized Medicine we want to solve

Provenance

- In the context of the life sciences, provenance can be perceived as comprehensive documentation that describes the whole scientific process, from the collection, generation, processing and analysis of biological material to respective data derivation, integration and analysis
- provenance information must be regarded crucial to the implementation of FAIR principles. Provenance information details the procedures, techniques, equipment, materials and actors involved in collecting, accessioning, processing, handling, storing, shipping and disposing of biological samples.

Open Science benefits for Personalized Medicine

- **European policies harmonization → data standards**
- Standards for outcomes measures
- Standards for data structures and syntax
- Standards for data transport
- Standards for data semantics
- Standards for metadata

Open Science benefits for Personalized Medicine

Semantic interoperability

- EOSC TF is working on
  - Interoperability recommendations (Minimum (meta)data set and interoperability indicators)
  - Inventory of tools for interoperability (Crosswalks, services, methods and formal languages)
  - Long-term sustainability (Recommendations for governance and processes for preservation and maintenance of semantic artefacts)
  - A survey of semantic artefact catalogues
  - Some semantic interoperability case studies and use cases

Open Science benefits for Personalized Medicine

Sensitive data RDA Interest group objectives (https://www.rd-alliance.org/groups/sensitive-data-interest-group):

- Developing a clear understanding of different levels of sensitivity of the data for designing governance frameworks to manage sensitive data.
- Creating adaptable protocols for researchers to collect, analyse, store, share and re-use data with different levels of sensitivity.
- Produce guidelines for balancing F.A.I.R. principles with sensitive data management principles (e.g. C.A.R.E. principles)
- Formulating strategies for overcoming barriers to utilizing, sharing and reusing sensitive data for deriving maximum knowledge out of sensitive data assets
- Developing guidelines for ethical sharing, use and re-use of sensitive data assets
- Exploring consent models governing the primary and secondary uses of sensitive data

Open Science benefits for Personalized Medicine

FAIRification → requirements for medical research datasets

- requirements for FAIR repositories that support the increase of FAIRness of the metadata and data hosted on them:
  - Provide globally unique and persistent identifiers for both metadata and data.
  - Define a metadata schema that includes relevant content to support findability and reuse.
  - This metadata schema should be extensible so different domains and applications can add more metadata items that are relevant for them;
  - Include in the metadata record the identifier of the data it describes using a well-defined predicate/property;
  - Index the datasets' metadata record allowing clients (human and computational) to search for datasets based on the metadata records' items;
  - Provide an accessibility method to both metadata and data based on an open, free and universally implementable protocol;

Open Science benefits for Personalized Medicine

- **Provenance → Requirements on Common Provenance Model**
- DR.1 The provenance information model should capture, in a computable (machine-readable and processable) and reproducible way, all the events connected to the physical operations performed on the biological material and all the details of the data generation and data processing workflows, in order to allow the tracking and the backward reconstruction of the history related to sample processing and/or data generation and/or data processing.
- DR.2 Provenance model should have "institution" entity, in order to capture institutional responsibility and also to support resolving distributed provenance. The model should, in a computable computable manner, define responsibility of institution and responsibility of individual persons (and possible delegation of responsibility from institution to a particular person).
- DR.3 The provenance information model shall specify clear serialization guidelines as well as implementation guidelines to achieve interoperability of applications producing/consuming the provenance information.

Open Science benefits for Personalized Medicine

Very large datasets are delt through Life science data infrastructures

- With classical services: Configuration, Administration, Availability and maintenance, Performance, Storage, Server resources
- Large-scale sustainable and interoperable data management and storage methods that allow secure and easy access to and reuse of these highly complex data.
- Simultaneously, as omics-focussed life science research projects increasingly depend on more than one type of measurement, ability to integrate different data types. Phenotyping, bioimaging and biosample management can all be considered part of life science data, and omics-focussed life science data systems will need to interoperate with these

RDA actions (the Machinery)

Act !

RDA Individual, Organisational & Regional Membership

Working Groups
implementable, impactful outcomes

Interest Groups
idea generation, maintenance, …

Communities of Practice
discipline | research domain coordination

Do (advise)

Technical Advisory Board
socio-technical vision and strategy

Organisational Advisory Board
needs, adoption, business advice

Regional Advisory Board
regional vision and strategy

Plans and decides

Council
organisational vision and strategy

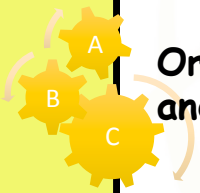
Secretariat
administration and operations

Secretary General

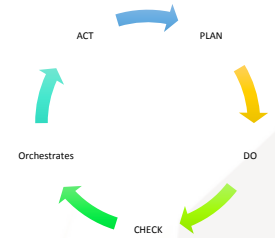
RDA Foundation

RDA Funders Forum

supports



Orchestrates and check !



RDA in action (the groups Matrix)

- FAIRSharing Registry
- FAIR4ML
- Life Science Data Infrastructures
- Blockchain Applications in Health
- Discipline Guidance for Data Manangement Plan
- Neuroimaging Data
- Raising FAIRness in Health data and health research performing organisation
- Reproducible Heath Data Services
- Health Data
- Sensitive Data

EOSC Future Domain Ambassador (RDA open call)

EOSC Future (https://eoscfuture.eu/)

- is an EU-funded H2020 project that is implementing the European Open Science Cloud (EOSC). EOSC will give European researchers access to a wide web of FAIR data and related services.
- Is a fully operational web of data and related services founded on FAIR protocols, principles and standards for accessing interoperable datasets. In practice, we will work with key stakeholders to ensure a smooth user experience, developing:
  - EOSC core, the set of enabling services needed to operate the EOSC
  - EOSC exchange registering resources and services from research infrastructures, other EOSC projects and science clusters to the EOSC and integrating them with the EOSC core functionalities
  - the EOSC interoperability framework will provide guidelines for providers that want to integrate services or data into EOSC

EOSC Future Domain Ambassador (RDA open call)

- **Domain ambassador: domain expert and skilled communicator to promote data sharing and open science practices in their disciplines**

- **RDA/EOSC Future is building a network of ambassadors to support activities designed to build awareness around the work and outputs of EOSC Future from a disciplinary perspective.**

Altogether !

- **Main actors**
  - ICPerMeD
  - The European Medicines Agency (EMA)
  - IHI (Innovative health initiative)
  - ERAPERMED
  - EOSC-Life Research Infrastructures
  - ECRIN
  - NIH
- **Citizen engagement**

Next steps

- **ESOC Symposium (domain ambassadors outputs)**

- **IDW 2023 / RDA P21: Presentation of a “BoF” (Birds of a Feather) session**

https://www.rd-alliance.org/very-large-open-datasets-machine-learning-personalized-medicine

with the aim to assess interest in the BoF’s topic. The expected outcome from the session would be a decision by the participants to either convert the idea to an IG or a WG