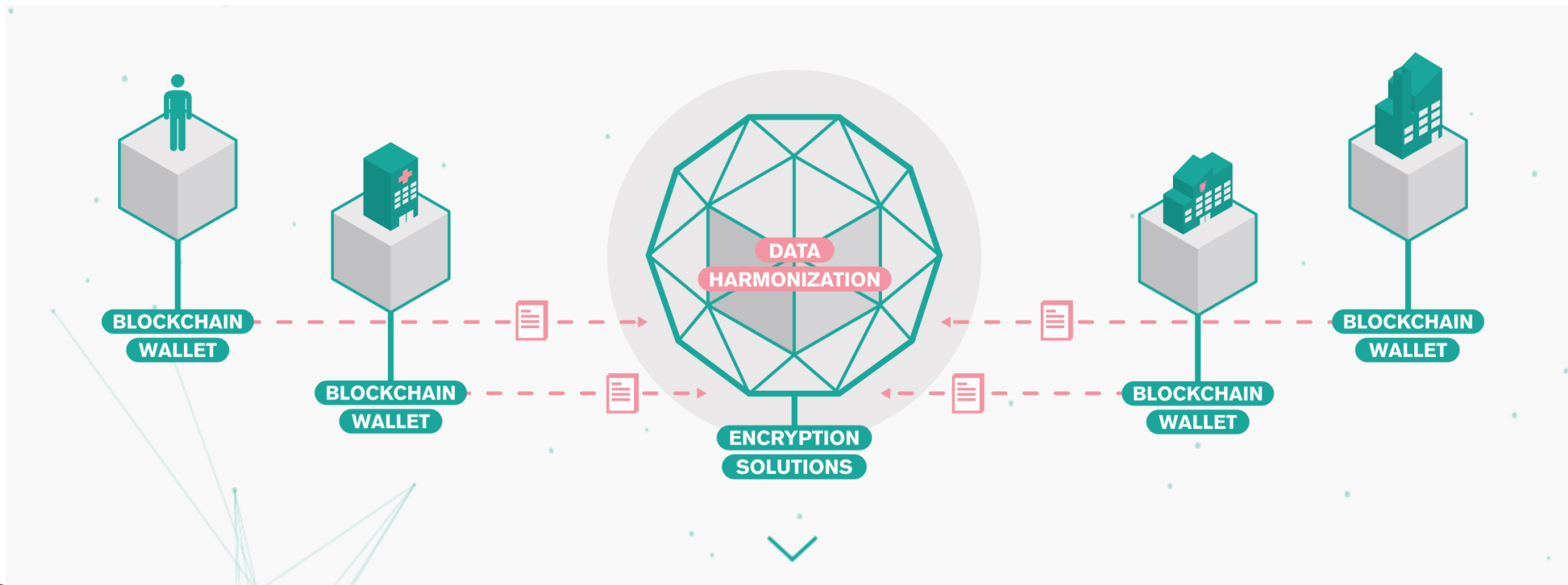


My Health, My Data **(and other related projects)**

Yannis Ioannidis
ATHENA Research Center & University of Athens

My Health, My Data!

- ▶ 1 / 11 / 2016 - 30 / 10 / 2019
- ▶ ~3M€ (~420K€ for ARC)



A NEW PARADIGM IN HEALTHCARE DATA PRIVACY AND SECURITY

CONSORTIUM

LYNKEUS .



National Research Council of Italy



Institute of Electronics,
Computer and
Telecommunication Engineering



gnúbila

Hes·SO

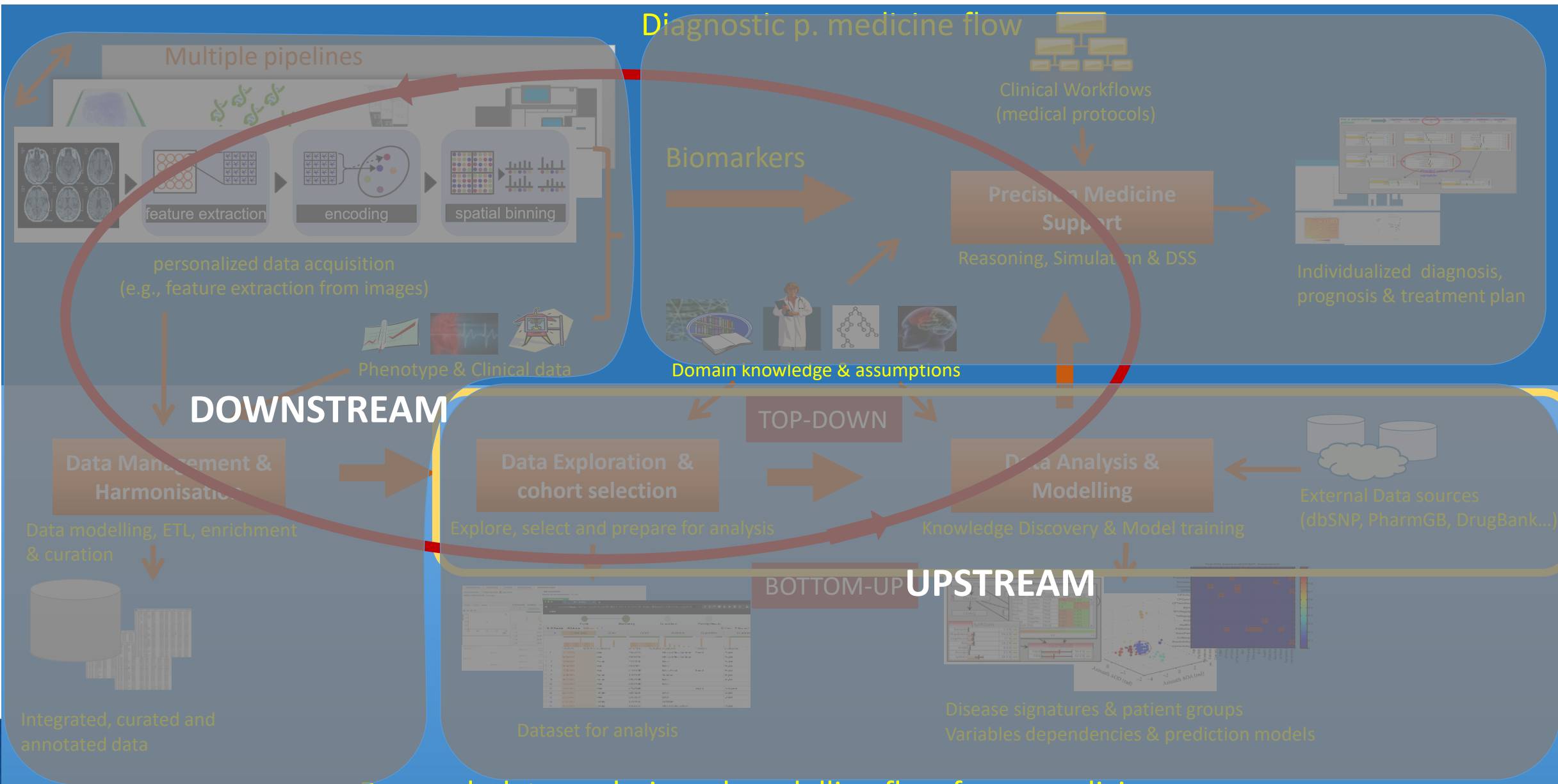
Haute Ecole Spécialisée
de Suisse occidentale

HWC
www.hwcomms.com

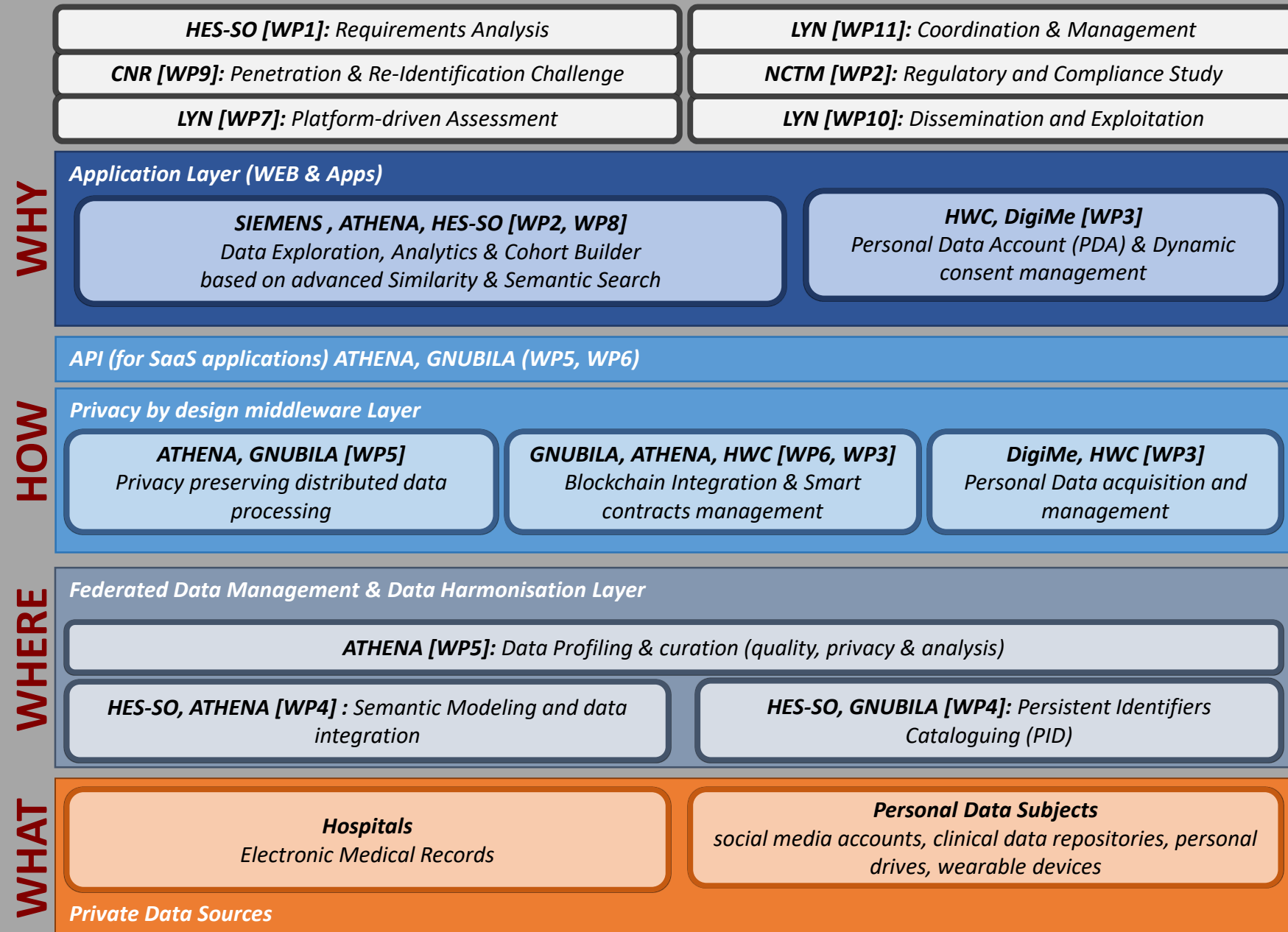


SIEMENS
Healthineers





Research data analysis and modelling flow for p. medicine



WHY

HES-SO [WP1]: Requirements Analysis	LYN [WP11]: Coordination & Management
CNR [WP9]: Penetration & Re-Identification Challenge	NCTM [WP2]: Regulatory and Compliance Study
LYN [WP7]: Platform-driven Assessment	LYN [WP10]: Dissemination and Exploitation

Application Layer (WEB & Apps)

SIEMENS , ATHENA, HES-SO [WP2, WP8] Data Exploration, Analytics & Cohort Builder based on advanced Similarity & Semantic Search	HWC, DigiMe [WP3] Personal Data Account (PDA) & Dynamic consent management
---	--

API (for SaaS applications) ATHENA, GNUBILA (WP5, WP6)

HOW

Privacy by design middleware Layer

ATHENA, GNUBILA [WP5] Privacy preserving distributed data processing	GNUBILA, ATHENA, HWC [WP6, WP3] Blockchain Integration & Smart contracts management	DigiMe, HWC [WP3] Personal Data acquisition and management
--	---	--

WHERE

Federated Data Management & Data Harmonisation Layer

ATHENA [WP5]: Data Profiling & curation (quality, privacy & analysis)	
HES-SO, ATHENA [WP4] : Semantic Modeling and data integration	HES-SO, GNUBILA [WP4]: Persistent Identifiers Cataloguing (PID)

WHAT

Private Data Sources

Hospitals Electronic Medical Records	Personal Data Subjects social media accounts, clinical data repositories, personal drives, wearable devices
--	---

Data Collection and Management

- ▶ Data collection / origin
 - Pseudonymised (de-identified) clinical (routine) data
 - Personal data including machine-generated data from Internet of Things (IoT)
 - Derived data related to the usage and the processing of the data
- ▶ Data storage & preservation
 - Federated data management for clinical data
 - ETL, pre-processing and pseudo-anonymization flow
 - DIGI.me Personal Data Account (PDA) application
 - retrieve personal data to an encrypted local library, which the users can then add to a personal cloud
- ▶ Data Modelling, Harmonisation, Cataloguing and Integration
 - Global dynamic Subjective-Objective-Assessment-Plan (SOAP) model
 - Use biomedical taxonomies and ontologies such as LOINC, SNOMED CT, ICD-10-CM, CPT, MESH
 - Persistent Identifiers (PIDs)
- ▶ Secure data access, sharing and processing in line with GDPR legislation

Hospitals



Human Brain Project

OPBG - Vatican

UCL/GOSH – London

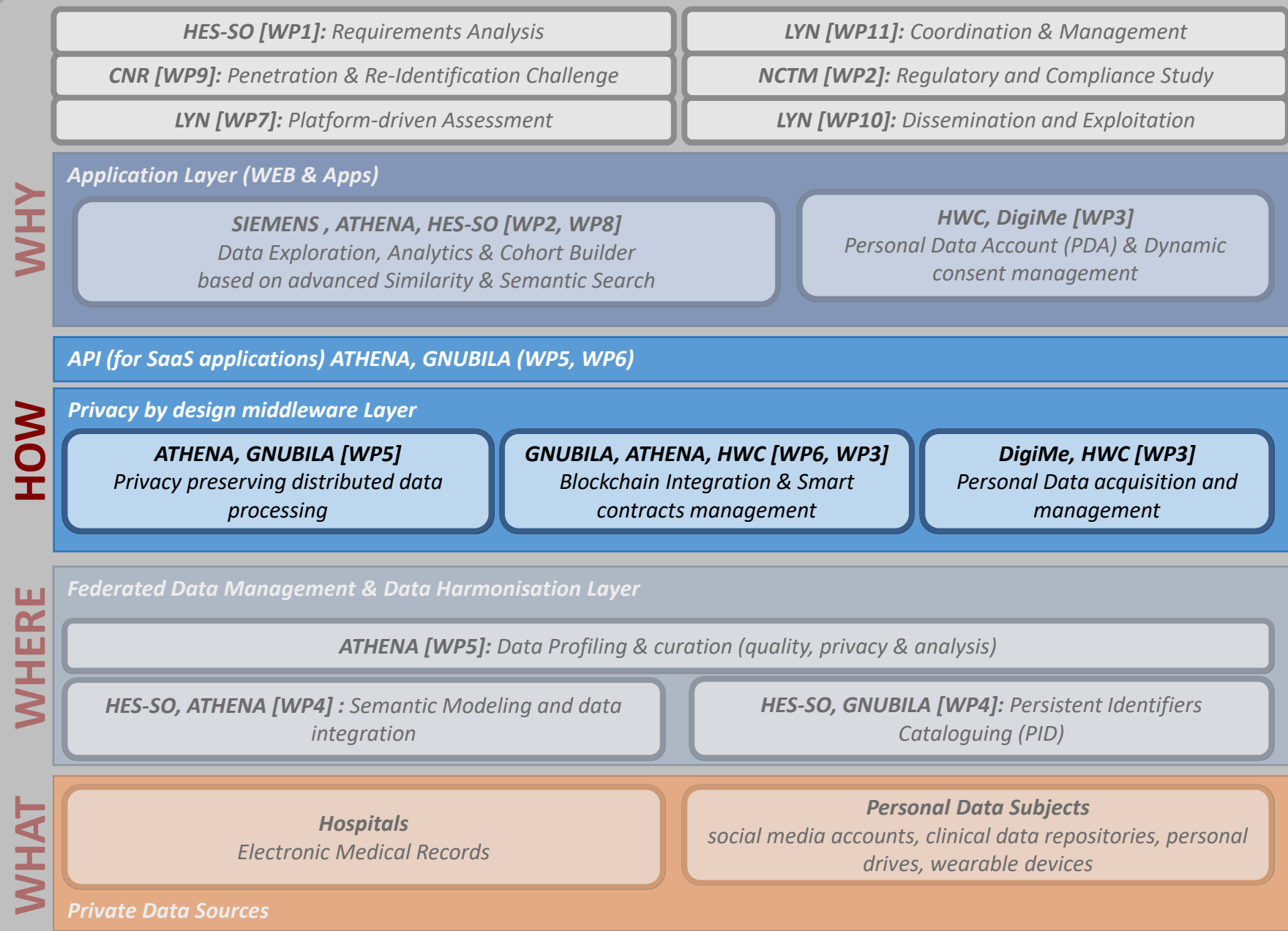
DH – Berlin

IGG – Genova

KU - Leuven

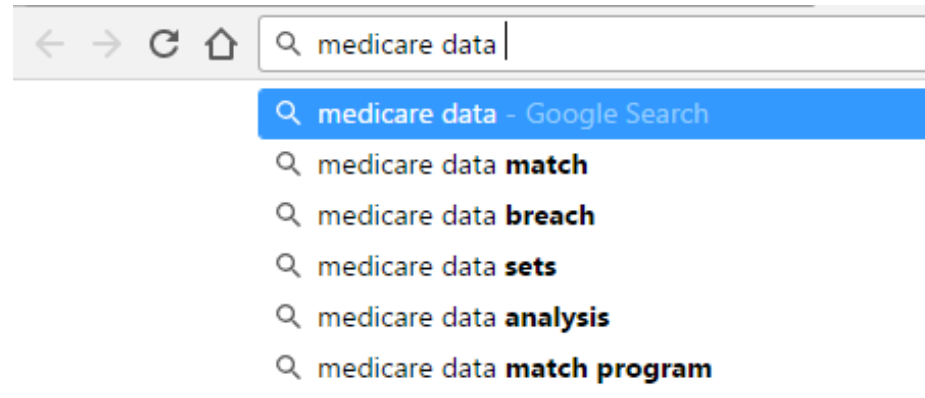
CHUV – Lausanne

...

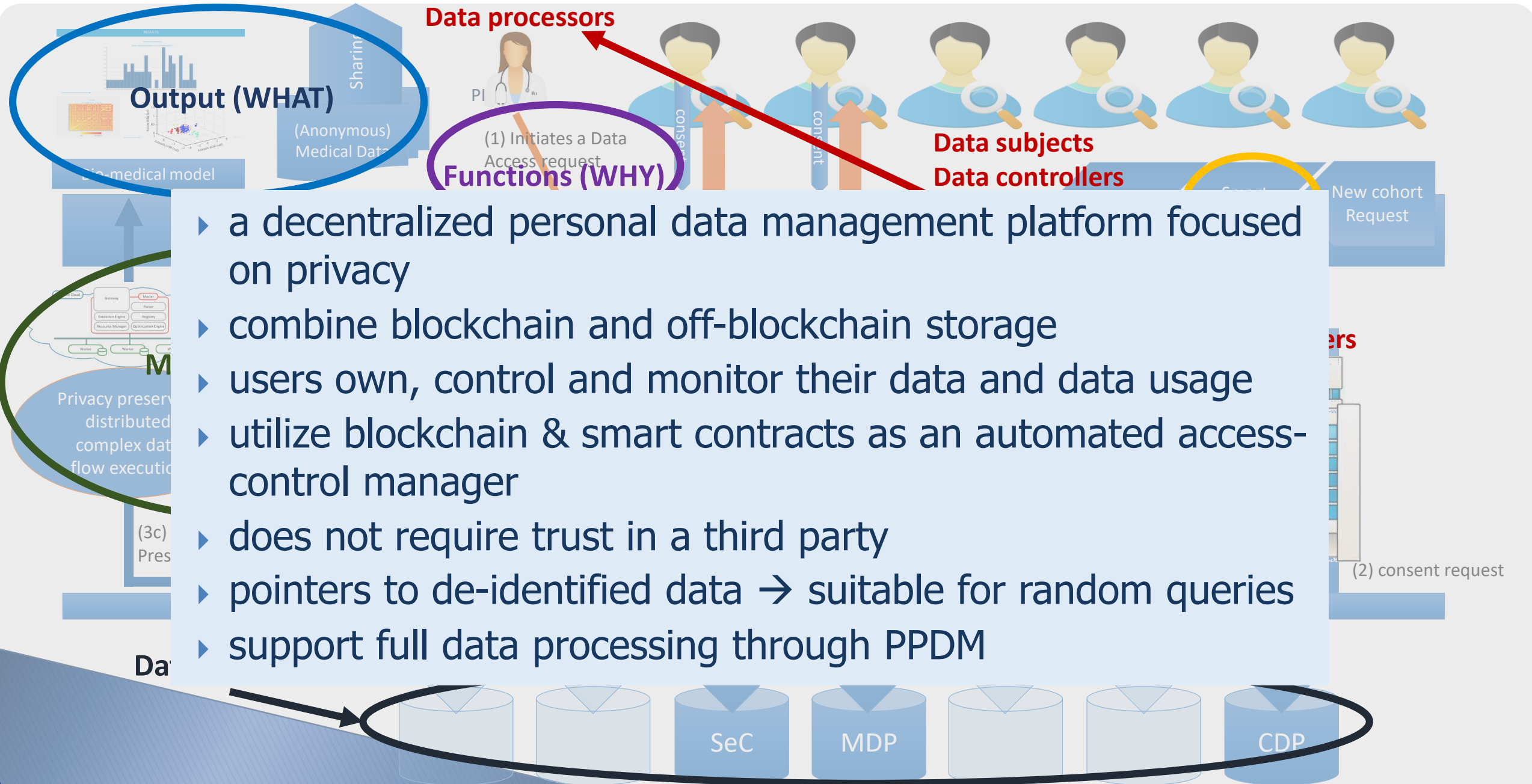


Data access & Privacy preservation

- ▶ Security / privacy breaches:
 - avoid a single point of failure (i.e., datawarehouse, TTP): decentralize data (transactions, patient data) and control using federation and blockchain
 - offer multiple levels of privacy preservation
- ▶ Ownership: Users should control their data, easily join or leave
- ▶ Transparency: Users should audit the usage of their data
- ▶ Privacy is important



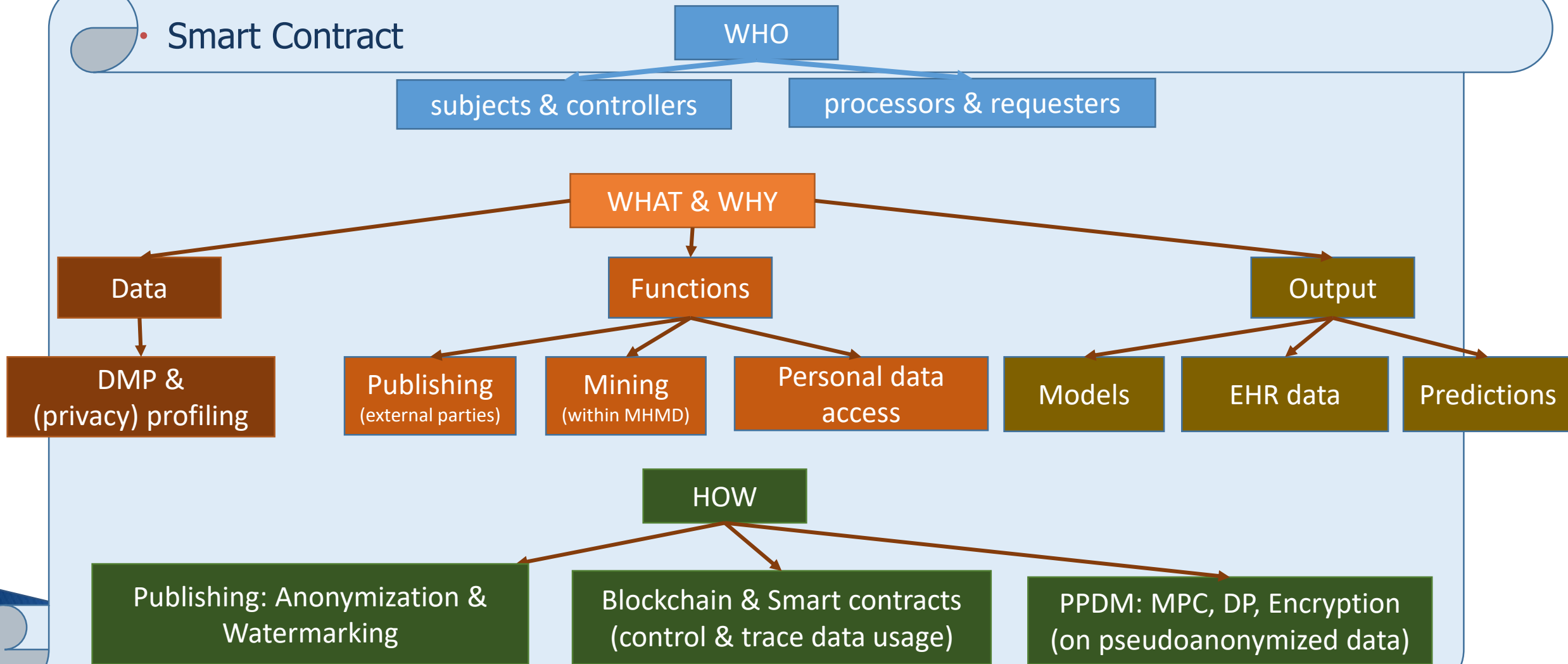
Blockchain integration @ MHMD



- ▶ a decentralized personal data management platform focused on privacy
- ▶ combine blockchain and off-blockchain storage
- ▶ users own, control and monitor their data and data usage
- ▶ utilize blockchain & smart contracts as an automated access-control manager
- ▶ does not require trust in a third party
- ▶ pointers to de-identified data → suitable for random queries
- ▶ support full data processing through PPDM

Blockchain integration

• Smart Contract



Encryption and privacy preserving policies

Three main use cases:

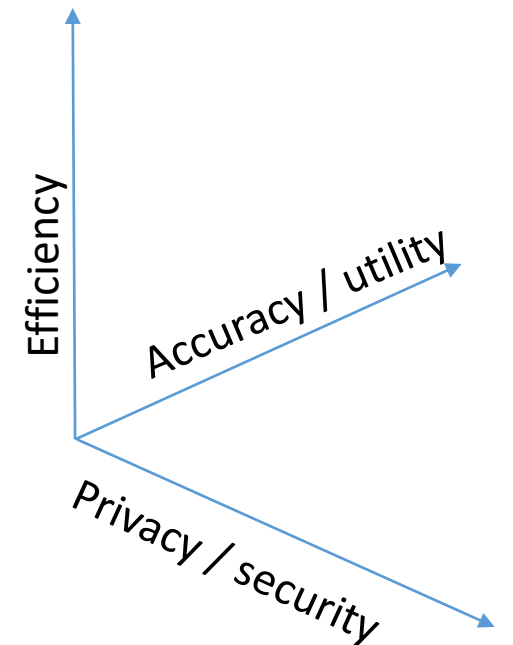
- ▶ Personal Data Access (no privacy)
 - Patient accessing his/her EHR
- ▶ Static Data publishing
 - Research VS other purposes
 - Anonymization requirements (AMNESIA)
 - Watermarking
- ▶ Privacy Preserving Data Mining (within platform)
 - Move data (authorized applications get and process the data i.e., MDP / Cardioproof)
 - Move computation to data: secure multiparty computation (SMC, DP) on federated data / distrustful parties (MHMD, HBP)
 - Other encryption techniques (homomorphic)

Encryption and privacy preserving policies

- ▶ **static data publishing:** “Sanitization” (Anonymization)
- ▶ **secure multi party computation:** Only overall aggregated data are transferred between nodes
- ▶ **interactive anonymization:** Differential Privacy & Crowd-Blending privacy
- ▶ **encryption:** Fully/Partially Homomorphic Encryption (FHE)
- ▶ **decentralization:** Use Blockchain to Protect Personal Data

Encryption and privacy preserving policies

- ▶ Privacy & Sensitivity Data Profiling:
 - Define privacy profiles per data type & usage scenario
- ▶ Trade-offs among efficiency, accuracy & privacy
- ▶ Define a formal methodology to describe “privacy budget” in terms of expected accuracy
- ▶ Automate privacy preserving method selection based on privacy & sensitivity profile and efficiency / accuracy trade-offs

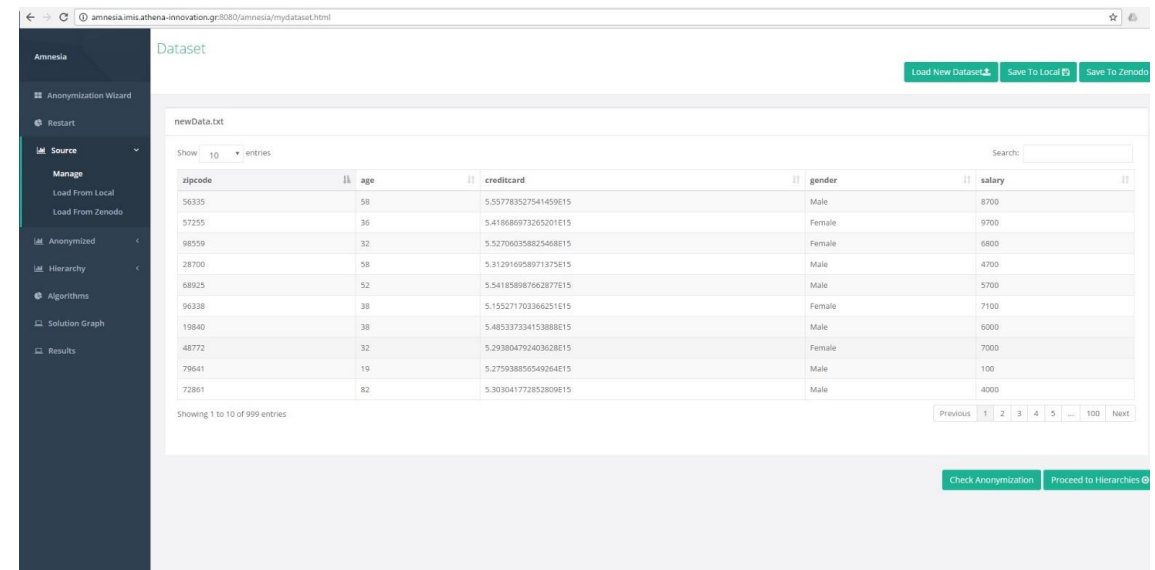


Secure Data publishing

- ▶ “Sanitization” (Anonymisation) hiding individual information (ensuring k-anonymity) but preserving aggregated (sufficient) statistics
- ▶ **Different dangers**
 - Identity leakage
 - Attribute leakage
 - Participation leakage
- ▶ **Different transformations**
 - Generalization
 - Suppression
 - Perturbation
 - Partitioning
 - Noise addition

Secure Data publishing

- ▶ Amnesia anonymization tool
 - It offers several versions of k-anonymity
 - It allows the user to select and customize possible solutions
 - It offers graphical tools that allow the user to analyze the anonymized dataset
 - It is scalable and uses all available CPU cores in the anonymization process
- ▶ Watermarking techniques



The screenshot displays the Amnesia web interface. On the left is a dark sidebar with navigation options: Annesia, Anonymization Wizard, Restart, Source (Manage, Load From Local, Load From Zenodo), Anonymized, Hierarchy, Algorithms, Solution Graph, and Results. The main area is titled 'Dataset' and shows a table named 'newData.txt'. The table has columns for 'zipcode', 'age', 'creditcard', 'gender', and 'salary'. Below the table, it indicates 'Showing 1 to 10 of 999 entries' and includes pagination controls. At the top right of the main area are buttons for 'Load New Dataset', 'Save To Local', and 'Save To Zenodo'. At the bottom right are buttons for 'Check Anonymization' and 'Proceed to Hierarchies'.

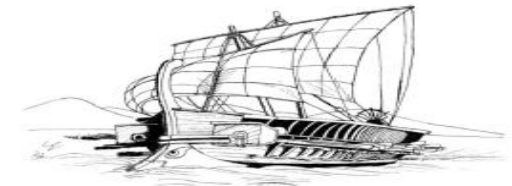
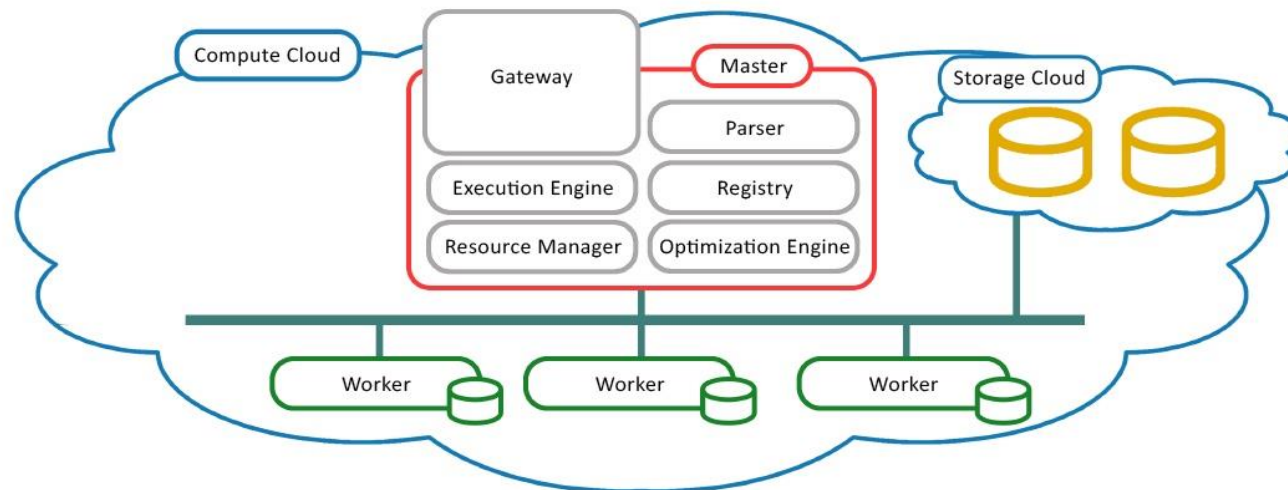
zipcode	age	creditcard	gender	salary
56335	58	5.557783527541459E15	Male	8700
57255	36	5.418686973265201E15	Female	9700
98559	32	5.527060358825468E15	Female	6800
28700	58	5.312916958971375E15	Male	4700
68925	52	5.541858987662877E15	Male	5700
96338	38	5.155271703366251E15	Female	7100
19840	38	5.48533734153888E15	Male	6000
48772	32	5.293804793403628E15	Female	7000
79641	19	5.275938856549264E15	Male	100
72881	82	5.303041772852809E15	Male	4000

Privacy Preserving Data Mining

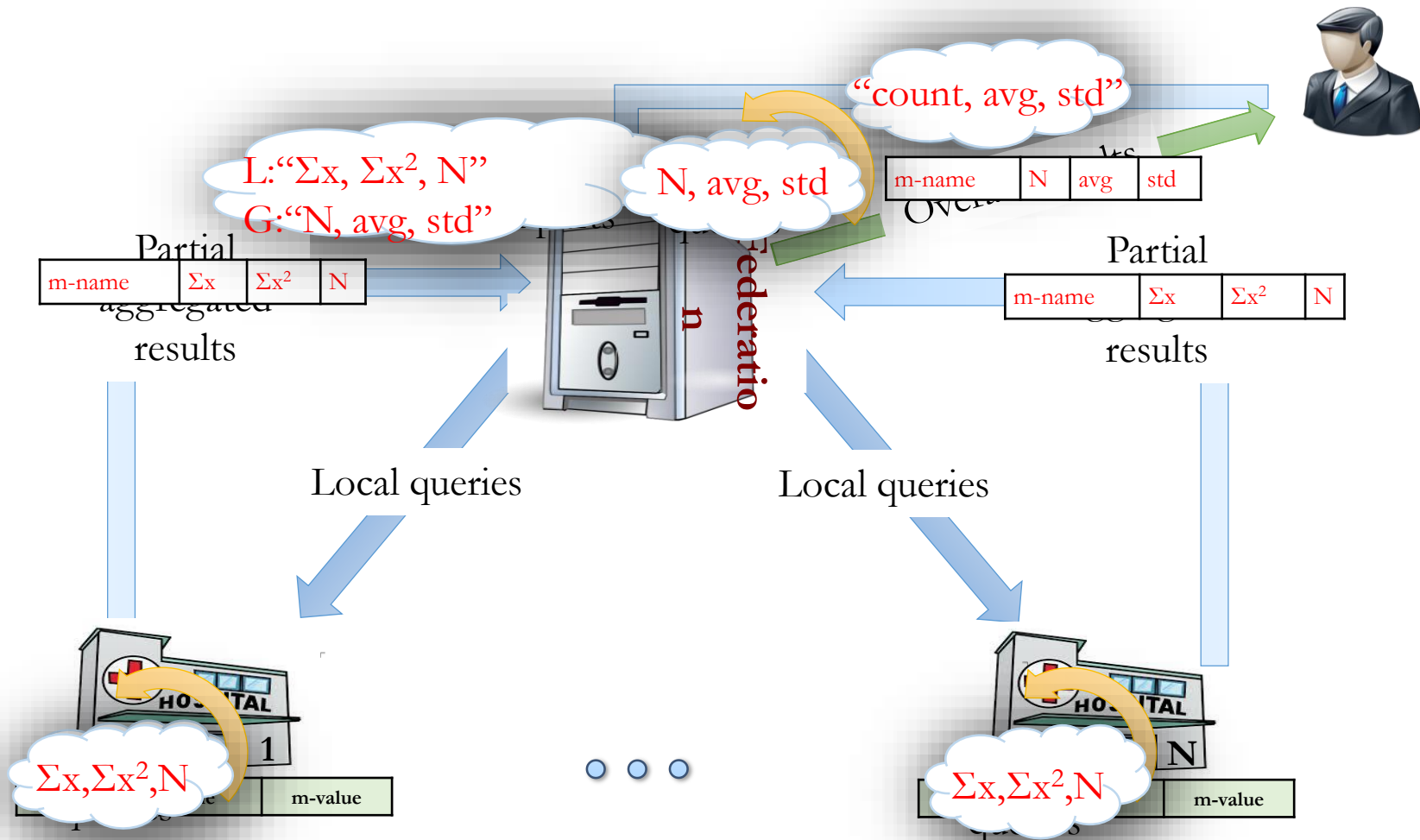
- ▶ **The setting:** Data is horizontally distributed at different sites on a Private Data Network (PDN) of mutually distrustfully parties
- ▶ **The aim:** Compute the data mining algorithm on the data so that nothing but the output is learned
 - Use secure computation using SMPC, encryption, DP etc
 - Assume Semi-honest types of adversaries that follow the protocol
 - Makes sense where the participating parties really trust each other (e.g., hospitals)
- ▶ **Training (learning) vs Reasoning:** different requirements and privacy related issues
 - training: needs access to patient records
 - reasoning: needs only the model and new data subjects but...
 - Inference from the results: One can break privacy using well specified queries and analyzing the results

Distributed Privacy Preserving Data Mining: EXAREME

- ▶ Distributed elastic execution
- ▶ Iterative dataflow execution: Support ML algorithms
- ▶ Powerful data programming paradigm: SQL with User Defined Functions
- ▶ **Privacy-aware query processing**



Dataflow Execution Example





MY HEALTH
MY DATA



A NEW PARADIGM IN HEALTHCARE DATA PRIVACY AND SECURITY

THANK YOU