Infrastructures for secure data analytics

Wessel Kraaij

TNO & Leiden University

EU: Digital technology will redefine health and care

- Care: transition to Value Based Health Care
  - PROMs
  - Aggregated cost of care paths
- Patient: self management of health, patient science
  - N=1 , personalized health
  - Find comparable health trajectories
- Combining the right data may lead to new insights
  - BUT Data storage is fragmented
  - BUT The GDPR limits the combination of data sets
- Challenges:
  - Provide a trusted environment, individual control on data access and sharing
  - Supporting secure and legal data analytics for combined datasets

Some barriers for statistical analysis and ML

- Data is horizontally partitioned

ID	age	income	sex
1	55	70000	M
2	45	60000	F

ID	age	income	sex
3	20	25000	F
4	22	20000	M

- Distributed learning
  - Personal Health Train (J van Soest)

- Data is vertically partitioned

ID	Age	sex
1	55	M
2	45	F
3	20	F
4	22	M

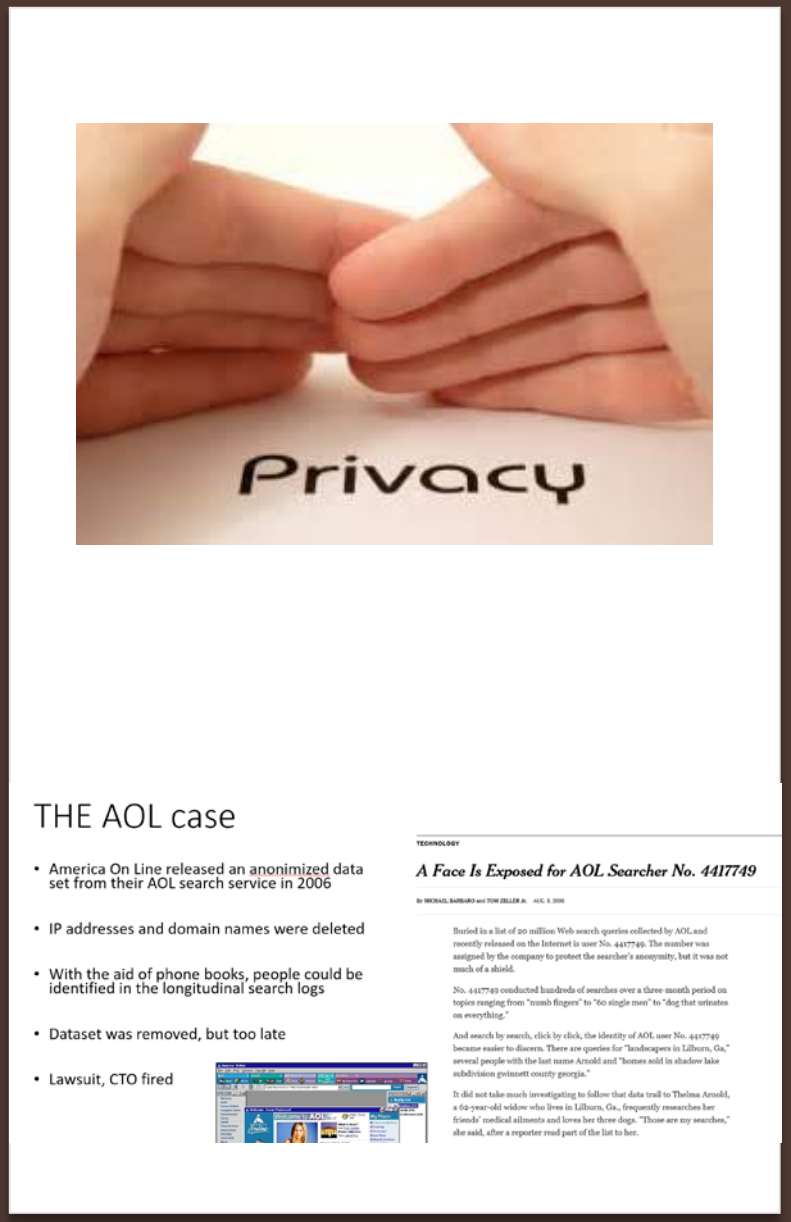
ID	income
1	70000
2	60000
3	25000
4	20000

- Existing practice: Trusted 3rd party (TTP)
- Health Data Cooperative (Midata)
- Prana Data (example of secure multiparty computation)

Traditional solution – trusting a third party

- For research:
  - create anonymized/ pseudonymized datasets, possibly using a **trusted** third party
  - Anonymization: remove identifiers

- Export anonymized dataset for research (academic/ commercial)

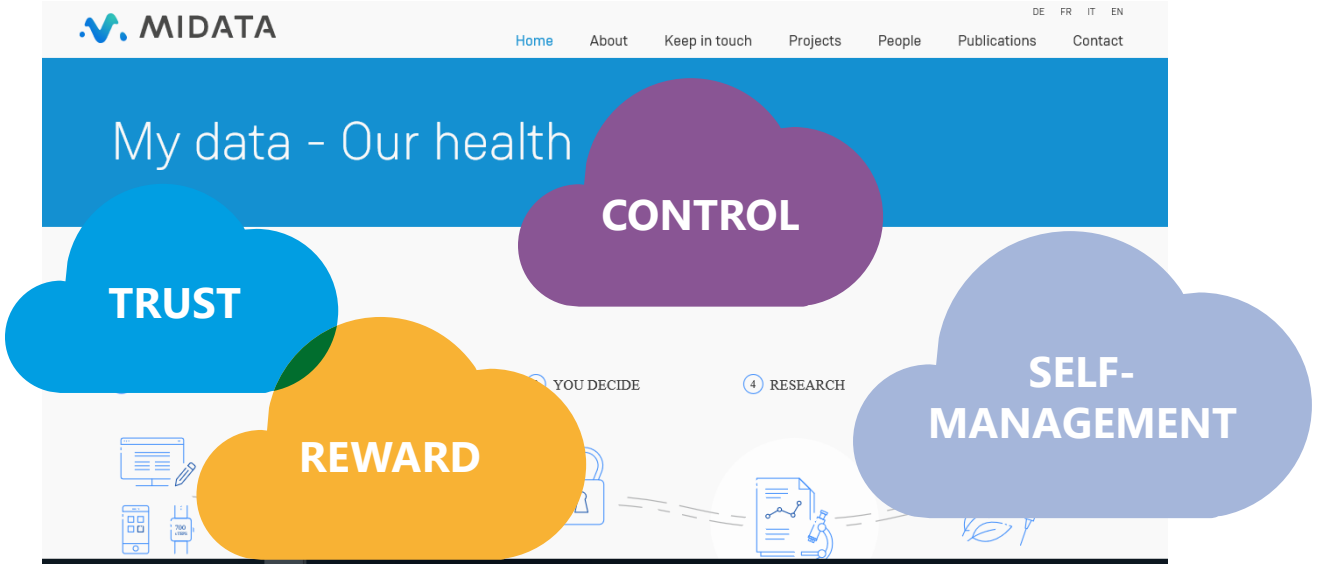


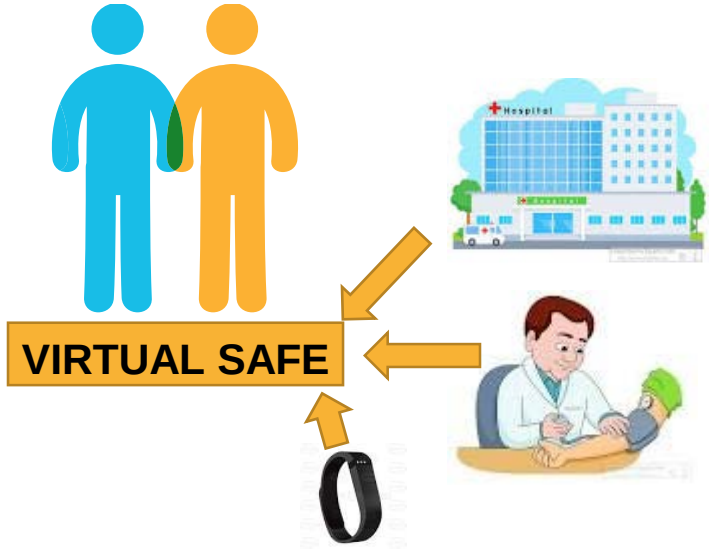
Assessment

- The more personal data is combined, the easier it is to re-identify a profile and the more difficult the anonymization process is
  - E.g. personal well-being data (Fitbit, smartphone apps) are quite personal and could lead to re-identification using external data.
- Personal data may seem innocent, but can lead to valuable insights

- Strict regulation on data processing and storage
  - Data leaks can lead to substantial fines
  - Professionals are reluctant to use big data approaches

HOLLAND HEALTH DATA COOPERATIVE:





ROADMAP NEXT ECONOMY

THE GAMECHANGER

Towards a citizen driven healthcare economy:

> Citizen's together form a **Cooperative and a Community**

> The cooperative delivers the platform and governance structure

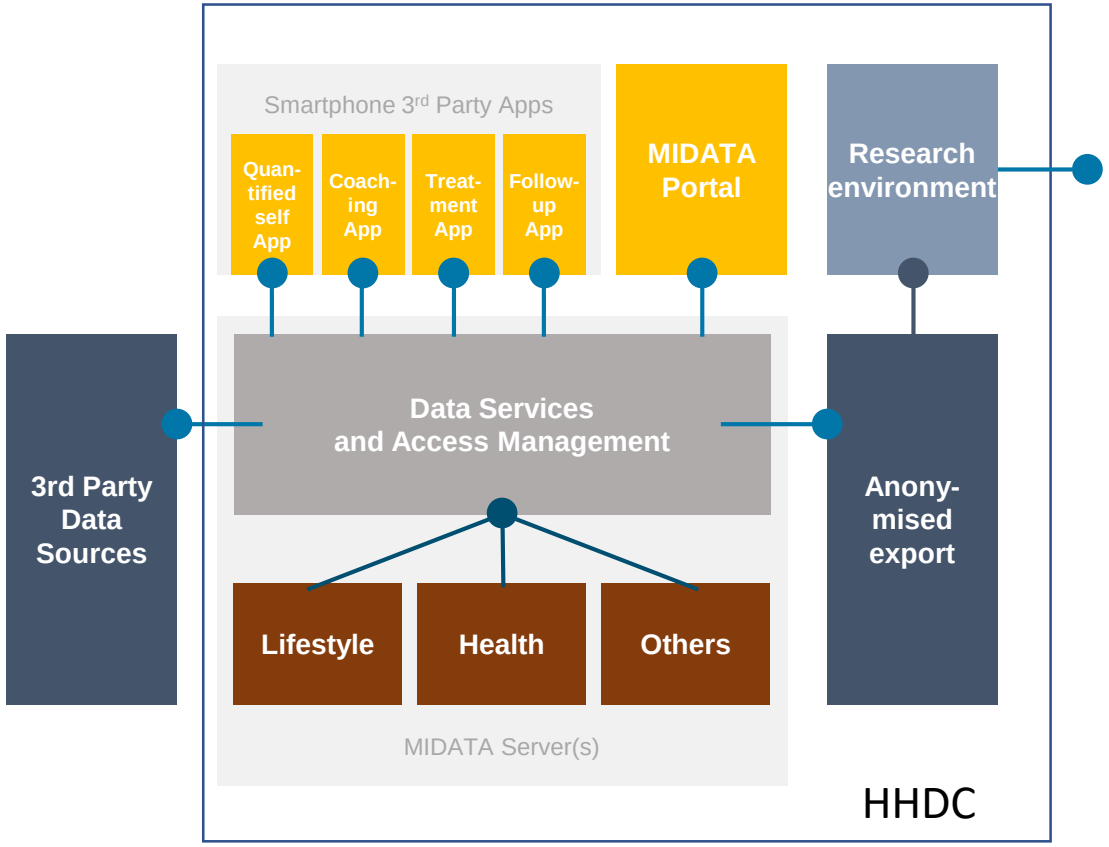
> Enables individuals to collect their data (medical and lifestyle)

> Provides services for members and delivers services to customers

> Data is controlled by citizen and patients themselves

> Rely on their cooperative for support

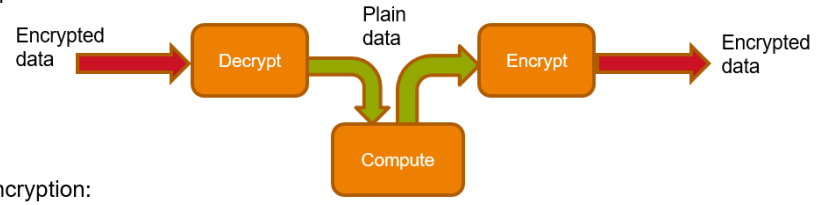
HHDC architecture



PRANA DATA (WWW.PRANADATA.NL)

- Privacy preserving analyse of sensitive data
- Results:
  - Proof of principle: secure ‘babies like mine’ , homomorphic encryption, for predictive mean matching (characteristics of babies with similar growth curve)
  - User study
- White paper
- Collaboration with Personal health train, PEP

- Traditional encryption:



- Homomorphic encryption:

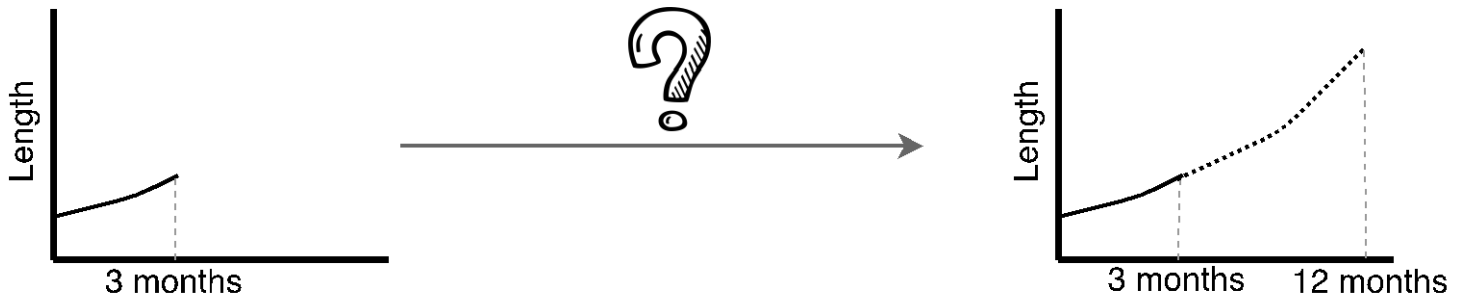
PRANA Advisory board

Follow up in H2020 BigMedilytics (ErasmusMC, Achmea, TNO and VWDATA

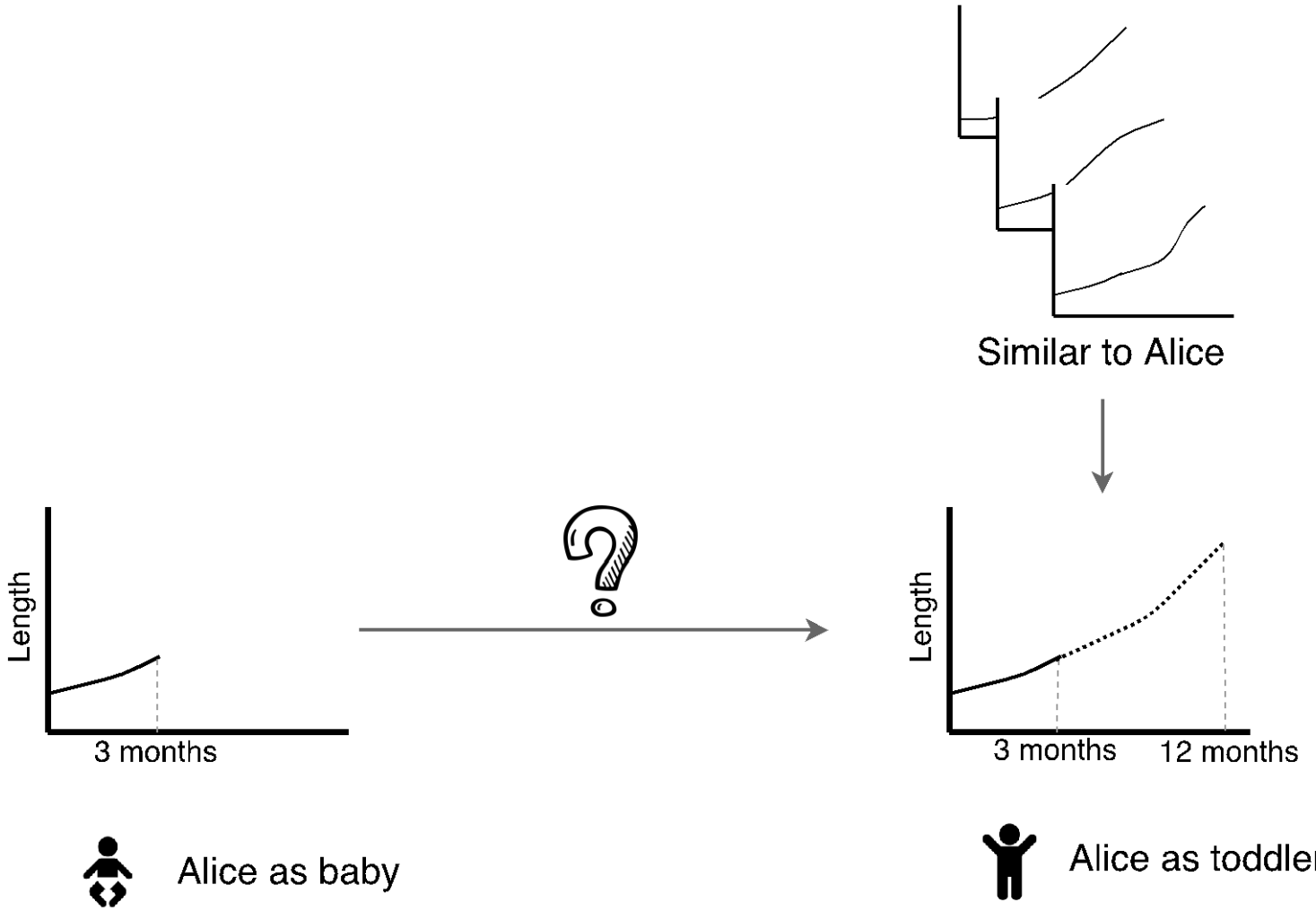
PREDICT EVOLVEMENT USING PATIENT MATCHER

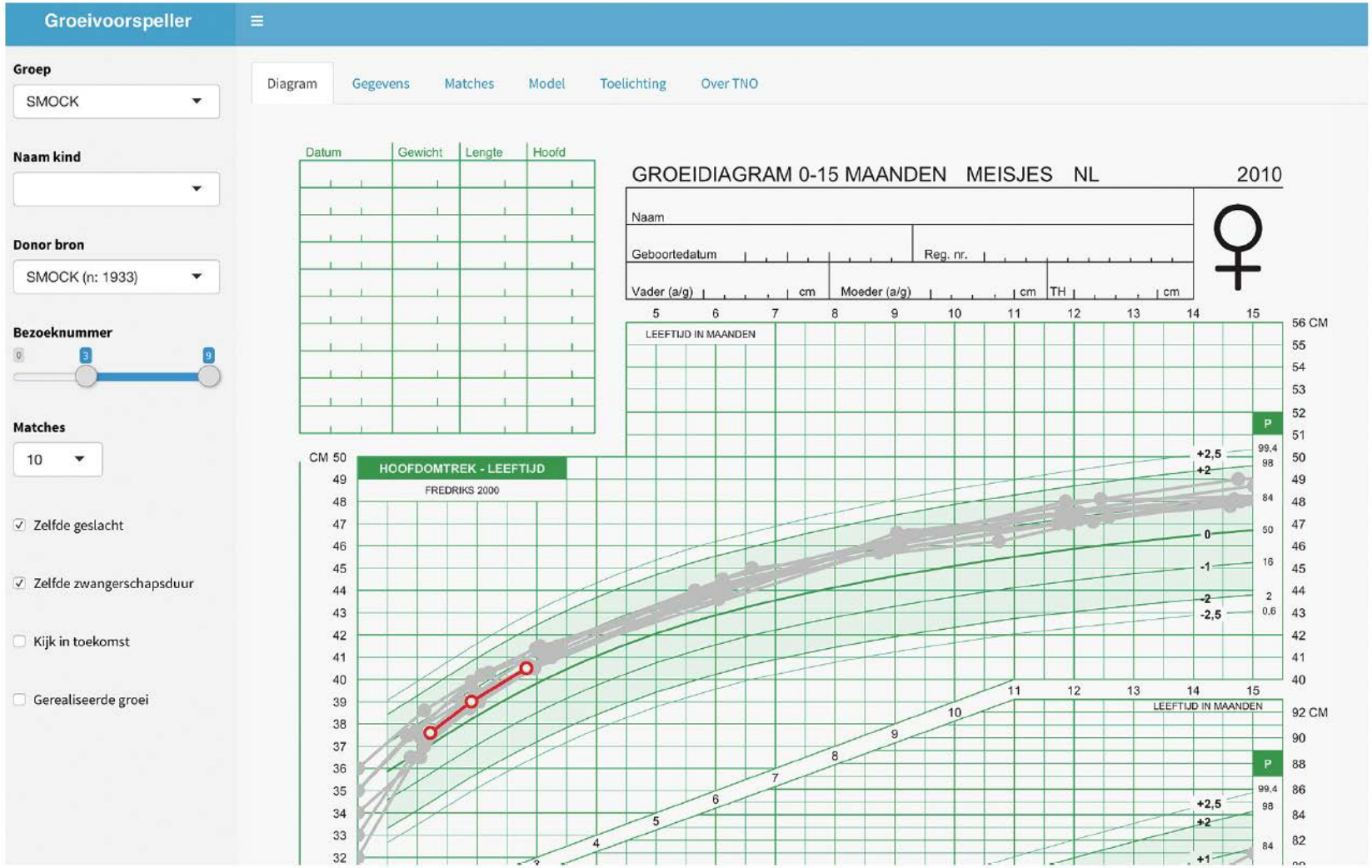
Alice as baby

PREDICT EVOLVEMENT USING PATIENT MATCHER



PREDICT EVOLVEMENT USING PATIENTS MATCHER





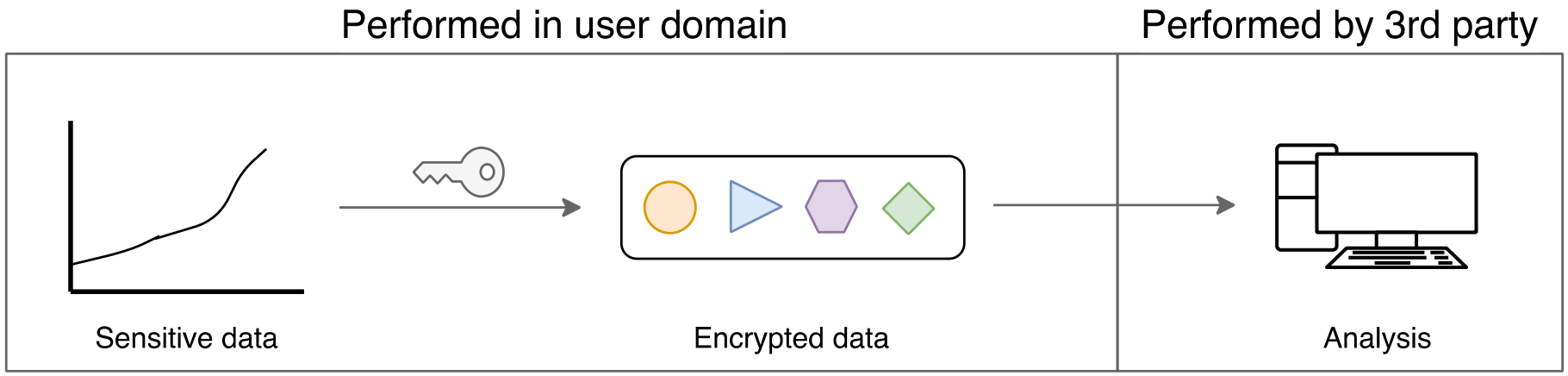
FIGUUR Groeidiagram voor de hoofdomtrek van meisjes van 0-15 maanden. Hierin is de al gerealiseerde hoofdomtrek van het doelkind (—○—) en de hoofdomtrek van 10 kinderen die op basis van de groeivoorspeller lijken op dit doelkind (—●—) weergegeven.
SMOCK = Sociaal Medisch Onderzoek Consultatiebureau Kinderen.

Predictie van groei vanaf jonge leeftijd: 'curve matching' met de TNO groeivoorspeller

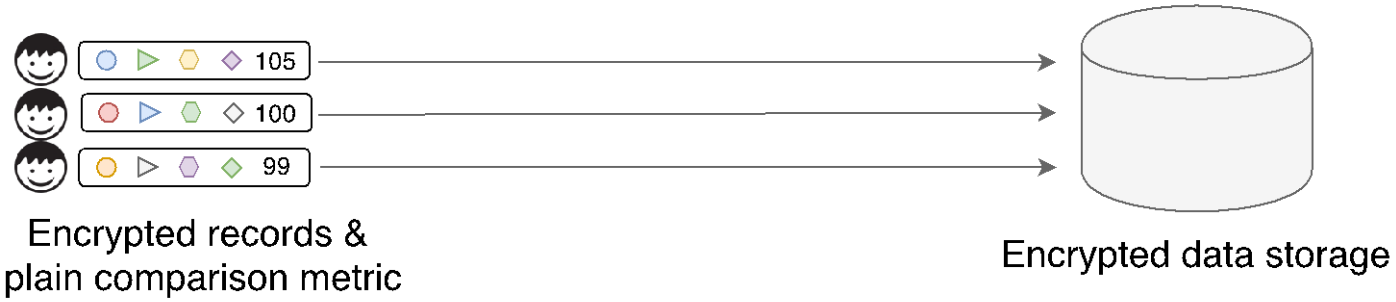
van Buuren, Stef; Bezemer, RA; Reijneveld, Sijmen; L'Hoir, MP

Published in:
Nederlands Tijdschrift voor Geneeskunde

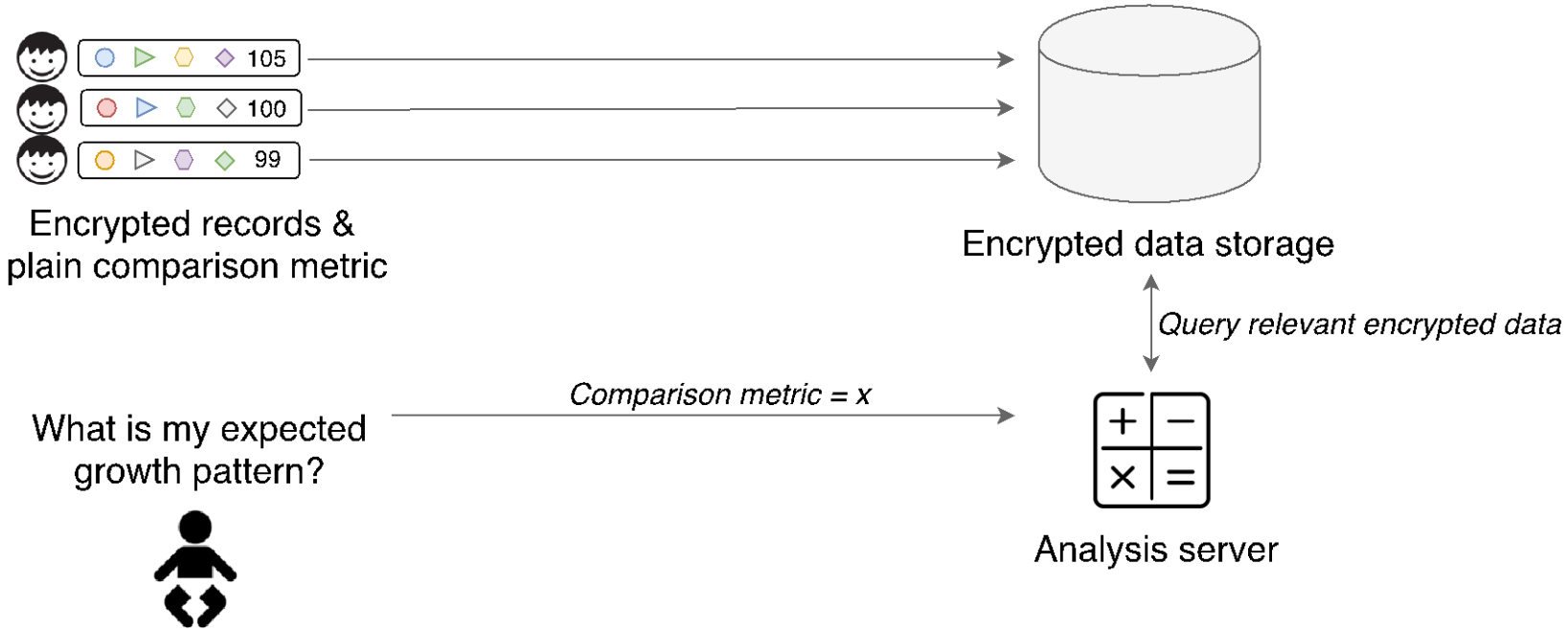
GOAL: ANSWERING CHILD DEVELOPMENT QUESTIONS WITHOUT SHARING SENSITIVE DATA



SECURE PATIENTS LIKE ME: 1/3 DATA INGEST

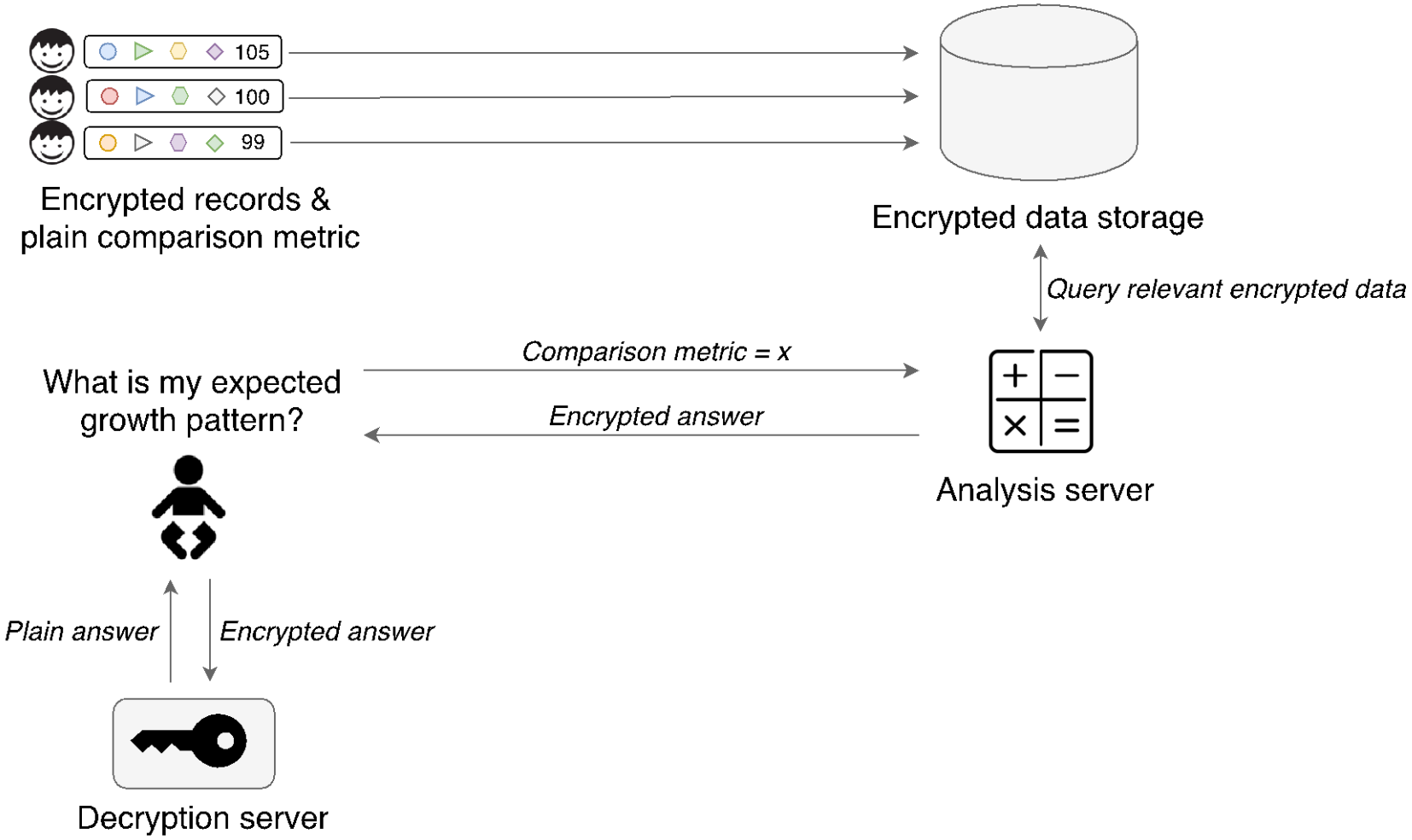


SECURE PATIENTS LIKE ME: 2/3 ANALYSIS



Apply ‘predictive mean matching’ in the encrypted domain

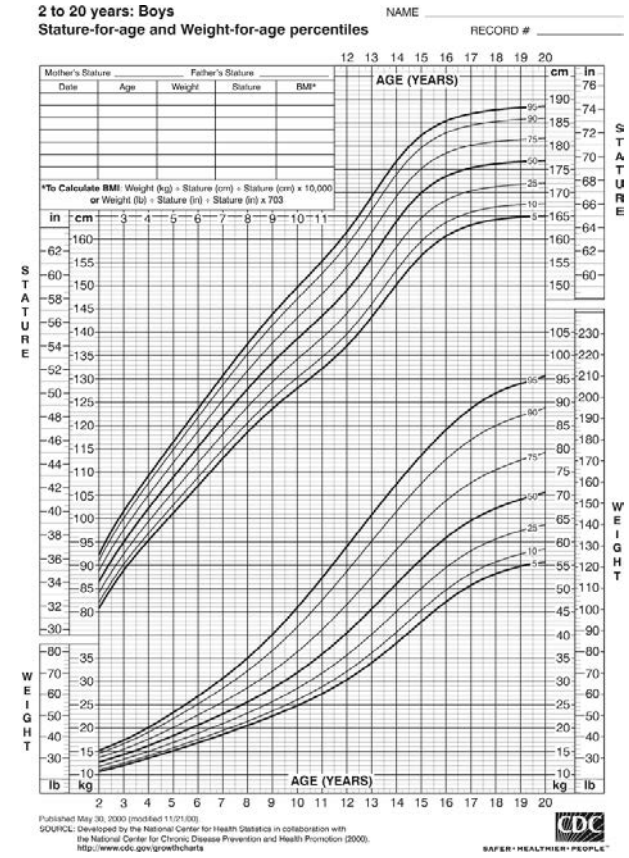
SECURE PATIENTS LIKE ME: 3/3 DECRYPTION



QUESTIONS ANSWERED? PRIVACY PRESERVED?

Alice's parents/doctor can:

- Plot expected growth curve
- Spot growth issues that might occur
- Obtain benchmarks about parent's age, doctor visits, obesity, etc.



Privacy preserved:

- Sensitive records nor the aggregations are learned by the system
- Analysis server learns distribution of *comparison metric*
  - The distribution of child lengths is not sensitive