Healthcare Transformation from Data and System Perspectives

Beng Chin OOI

www.comp.nus.edu.sg/~ooibc

Contents

- **Healthcare Problems**
- Challenges
- Our Healthcare Data Analytics Stack
  - GEMINI
    - Cleaning, De-biasing, Regularizing
  - ForkBase
    - Storage Engine for Collaborative Analytics and Forkable Applications
  - Foodlg / Foodhealth
    - Pre-diabetes app
  - MediLOT
    - A blockchain solution
- Conclusions

FIRST OPINION

An Obamacare success: financial penalties reduce hospital readmission rates

By JASON H. WASFY, FRANCESCA DOMINICI, *and* ROBERT W. YEH / DECEMBER 27, 2016

The Mnistry Of Health (MOH) Office for Healthcare Transformation (MOHT) (formed in 2018) aims to shape the future of healthcare in Singapore. This is done by identifying, developing and experimenting with game-changing systems-level concepts and innovations in the key areas of health promotion, illness prevention and the delivery of care.

AI in Health Grand Challenge (Ongoing large grant call by AI.SG – 3 x5 mil in the first phase and 1 x 20 mil in the second phase)

"How can Artificial Intelligence (AI) help primary care teams stop or slow disease progression and complication development in 3H – *Hyperglycemia (diabetes), Hypertension (high blood pressure) and Hyperlipidemia (high cholesterol)* patients by 20% in 5 years?"

3H Problems: Where/what Can We Contribute?

Life Style

Hyperglycemia

Hypertension

Hyperlipidemia

Drug Compliance +
Pharmacogenomics

Eye (DME, retinopathy, glaucoma, …)

Kidney (AKI, ESRF …)

Cardiac (AMI)

Stroke (AF, fall…)

Limb Salvage/amputation

Personal Health Coach

Chatbot +
Behavior …

Sensors + Cameras

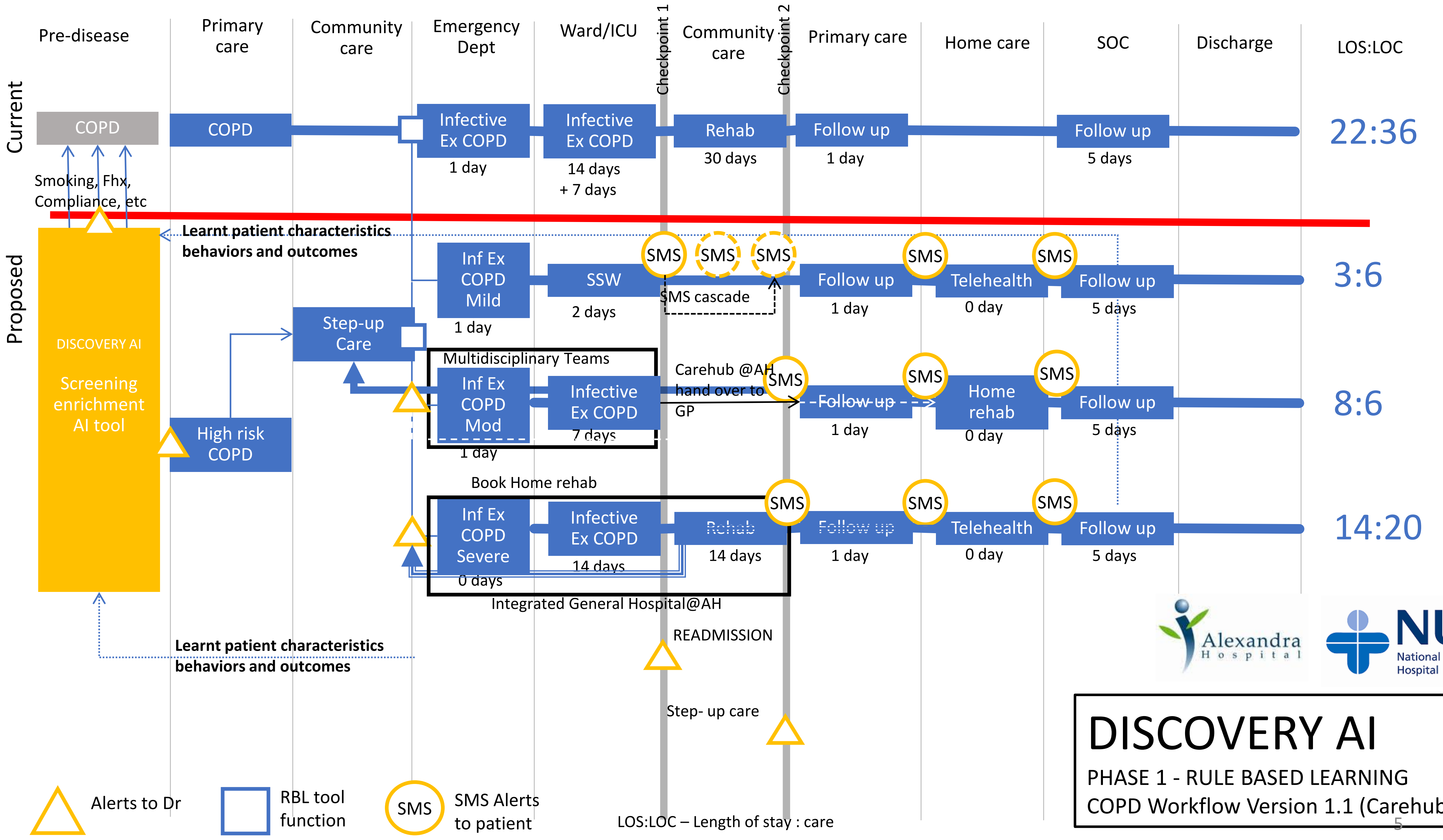
Telemedicine

Hospital System

Healthcare Analytics

Primary Care

Secondary Care ++



DISCOVERY AI

PHASE 1 - RULE BASED LEARNING

COPD Workflow Version 1.1 (Carehub)

Healthcare System/AI's Objective

PubMed

A unified end-to-end engine to integrate all available data sources and provide a holistic view of medical data, from where we support all sorts of medical applications.

- Increase the accuracy of diagnoses
- Improve preventive medicine
- Optimize insurance product costs
- Better understand the needs for medications
- Cut costs on healthcare facility management etc

This is beyond typical database query processing

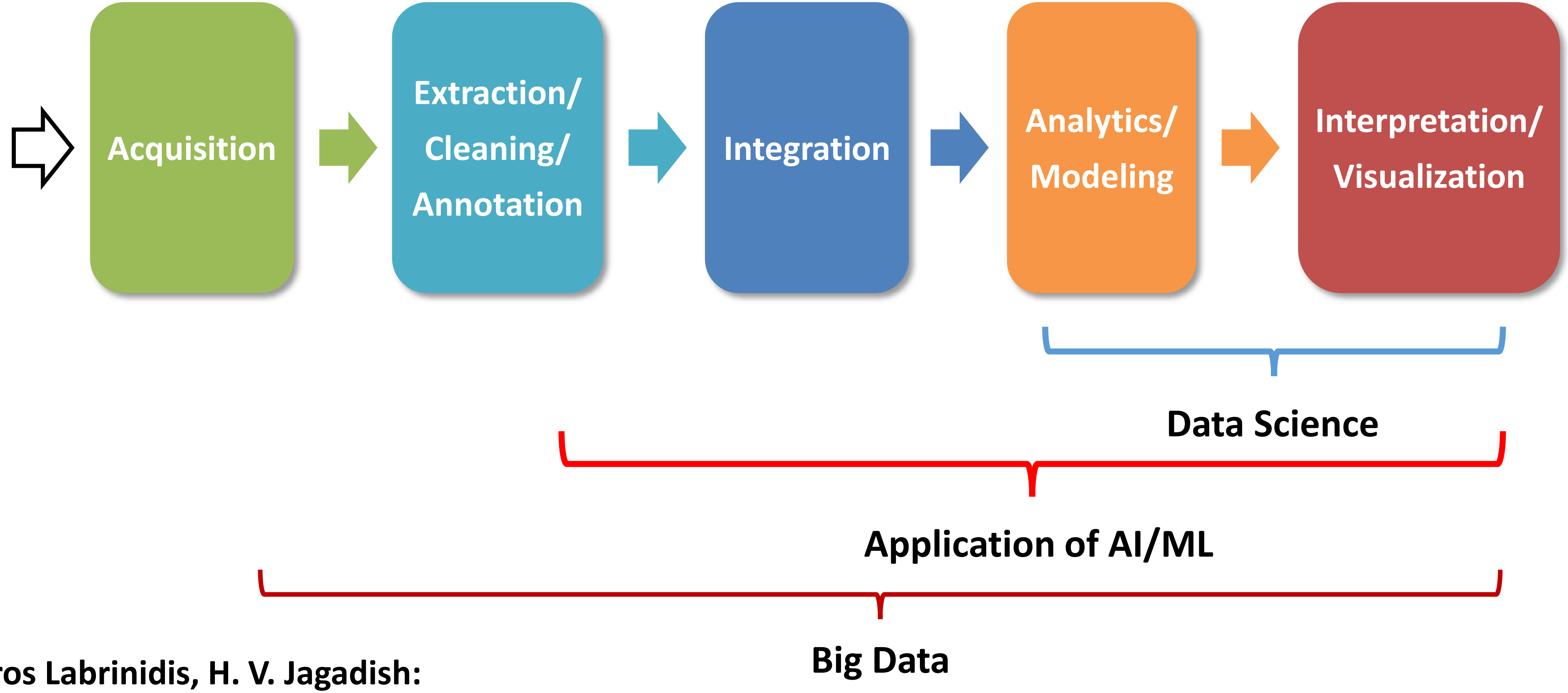
The Reality of Exploiting AI

- The actual implementation of the ML algorithm is usually **less than 5%** lines of code in a real, non-trivial application
- The main effort (i.e. those 95% LOC) is spent on:
  - Data cleaning & annotation
  - Data extraction, transformation, loading
  - Data integration & pruning
  - Parameter tuning
  - Model training & deployment
  - … …

These are what we have been doing!

- This blurs the line between DB and "non-DB" processing, and calls for better integration

The BIG Data Analytics Pipeline*



***Alexandros Labrinidis, H. V. Jagadish:**
Challenges and Opportunities with Big Data. PVLDB 5(12): 2032-2033 (2012)

Challenges

Identifying Common Challenges

Readmi-ssion | **DPM** | **Radio-logy** | **App** | **Prediabetes Prev.** | …

…

Support

GEMINI Platform

Research

Clinical Needs

- ✓ Readmission
- ✓ Disease Progression Modelling (DPM)
- ✓ …

NUHS
National University Health System

NUH
National University Hospital

Jurong Health
Ng Teng Fong General Hospital

SingHealth
Defining Tomorrow's Medicine

Tan Tock Seng HOSPITAL

铭之慧科技

MEDILOT

Singapore General Hospital
SingHealth

KK Women's and Children's Hospital
SingHealth

China Healthcare Providers/Hospitals

...... more

Challenges

Time-consuming data extraction
- Different storage formats
- Unstructured data

Difficult data cleaning
- Missing data
- Duplications
- Different coding standards

Doctors-in-the-loop data annotation (medical expertise)
- Missing code filling
- Standardized diagnoses

Bias in observation data
- Observation data is biased from the actual conditions of the patients

Complexity of medical features
- Numerous concepts
- Heterogeneous data
- Complex relations

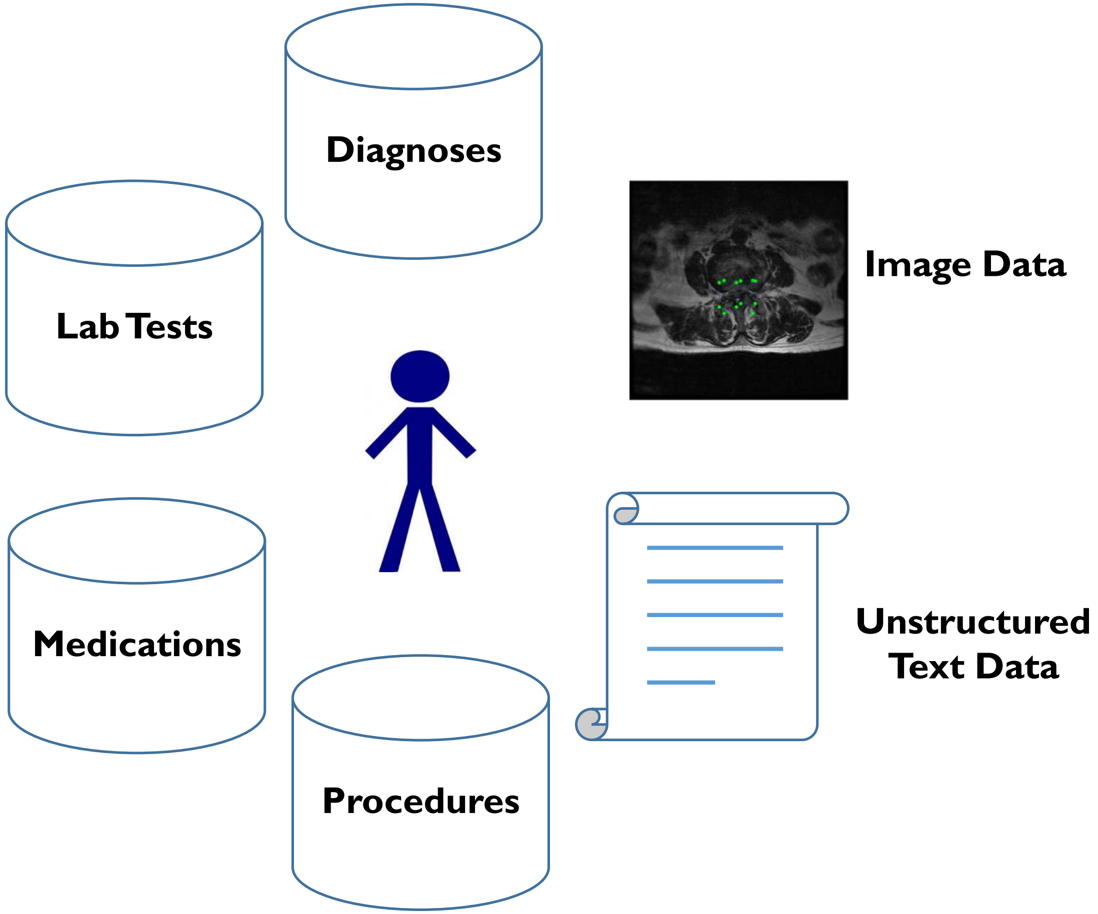
Demanding data storage requirements
- Multi-source and heterogeneous data formats
- Reuse of datasets
- Provenance

Challenge 1： Data Preprocessing

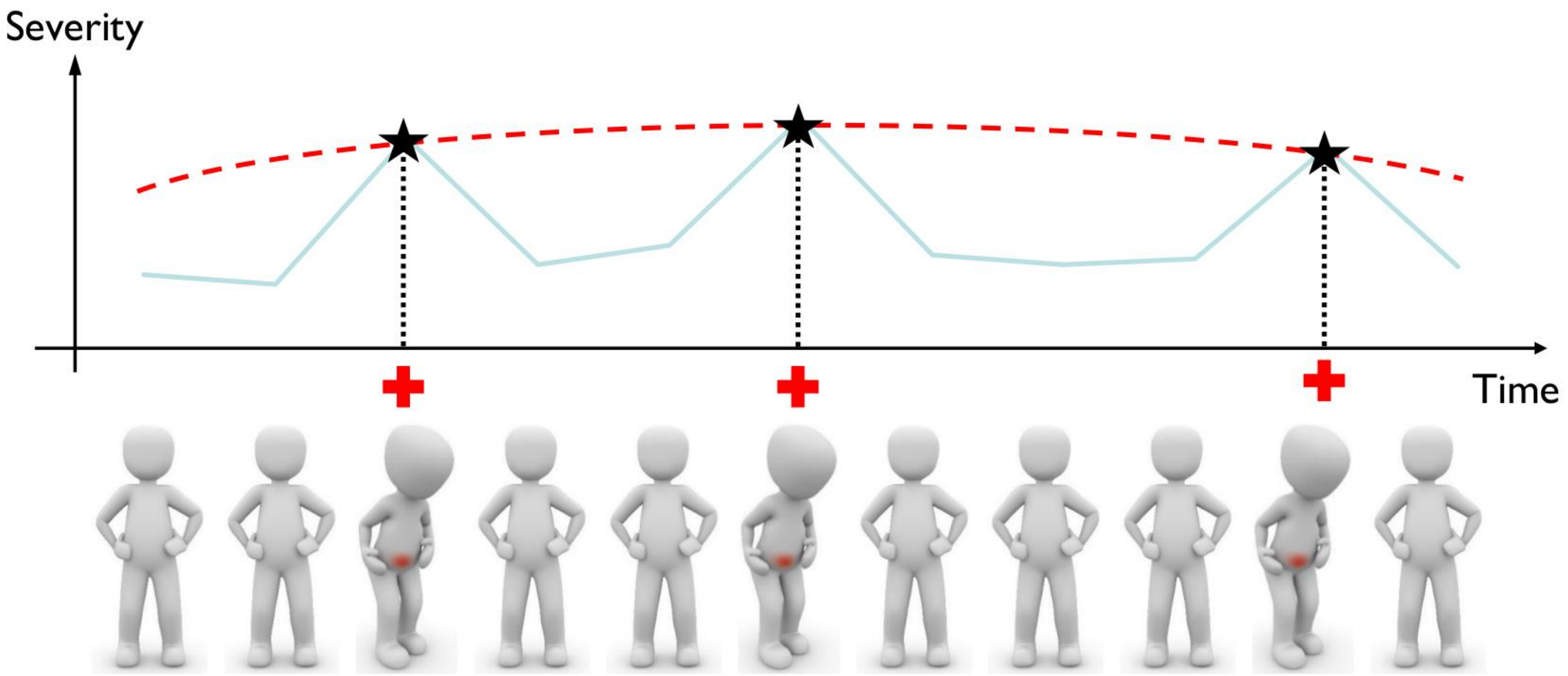
time-consuming data extraction
different storage formats, un-structured data

difficult and expensive data cleaning
missing data, duplications, different coding standards

medical expertise required for data annotation
standardizing diagnoses, missing code filling



Challenge 2： Bias in EMR Data



Challenge 3：Complex Features Relations

Numerous Concepts

UMLS consists of over 2.97 million concepts and 10+ million terms.

Multi-source and Heterogeneous Data

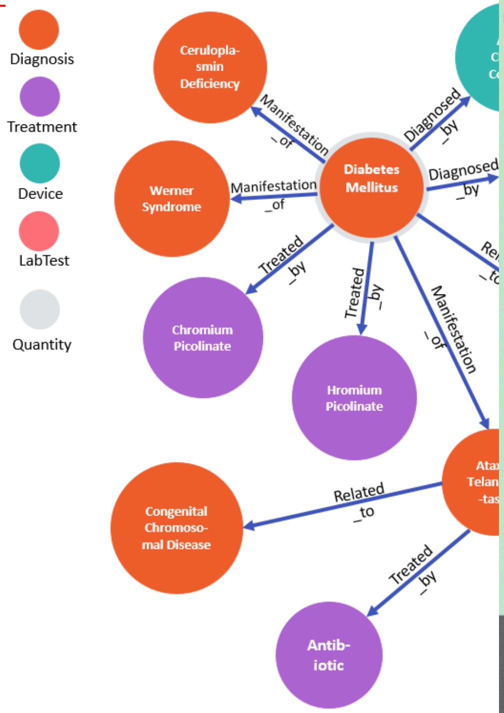
Medical data consists of diagnoses, lab tests, procedures, etc.

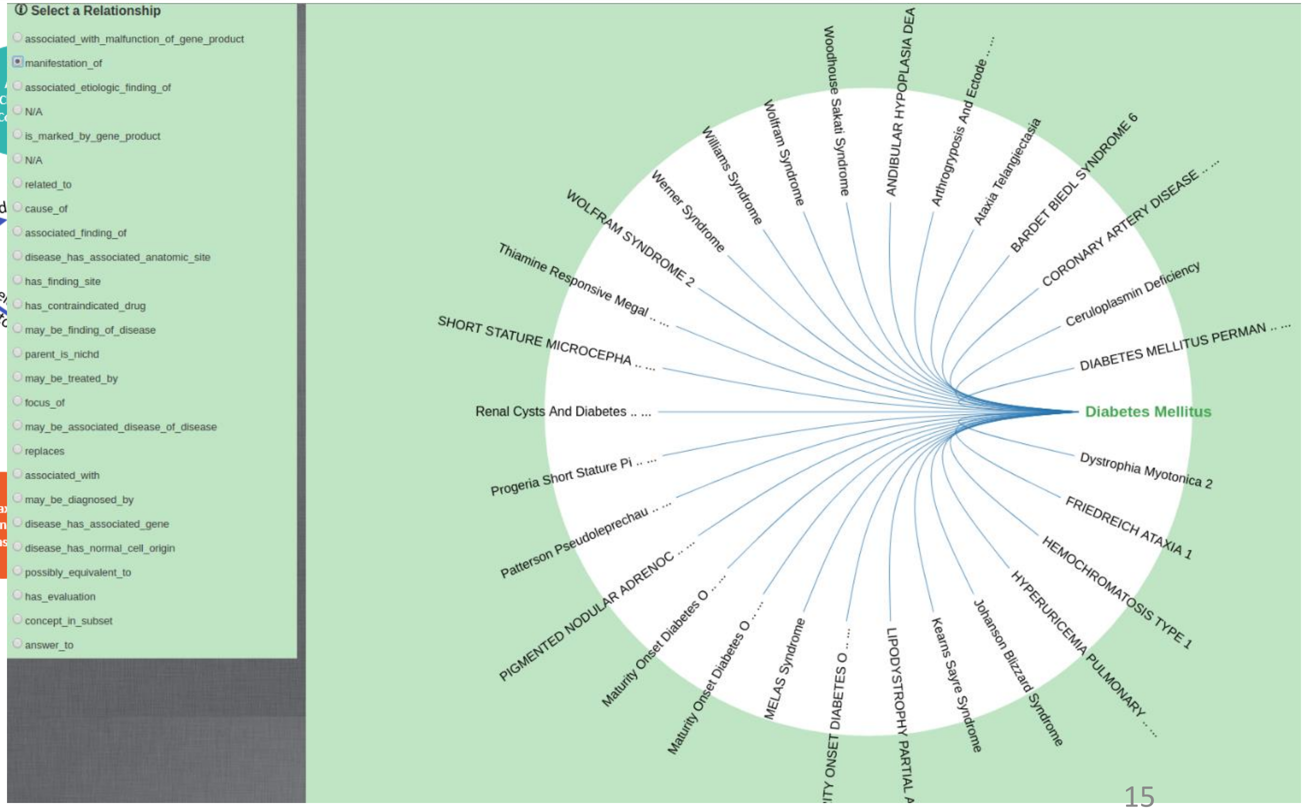
Complex Relations

Complex relations among different sources of medical data

NUH surgery dataset:
22987 medical features

12319 diagnosis codes
2335 lab test codes
6932 medication names
1401 procedure codes
8 demographic features
(BirthYear, Gender etc)





Challenge 4：Dataset Management in Healthcare

- Dataset Cleansing
  - Track evolution history to ensure correctness
- Dataset Transformation
  - Save different formats for future reuse
- Dataset Sharing/redundancy
  - Avoid data redundancy to reduce storage overhead
- Dataset Security
  - Impose access control to healthcare data

Challenge 5: Data Prior

- Existing ML algorithms work well for image classification and sequence prediction, but not healthcare problems

- Images are not random pixels
  - Neighbor pixels are most corelated --> CNN
  - Color channel prior --> haze removal/super-resolution
- Sequences are not random numbers/words
  - Latent state at each time point --> RNN LSTM

- Prior for healthcare?
  - How to find and formulate?
  - How to create algo/model to utilize them?

Matching Data and Model/Algorithm

- No Free Lunch Theorem [1997]
- Checklist for useful AI：
  - Lots of data
  - Flexible models
  - Efficient system and algorithm design
  - Powerful priors that can defeat the curse of dimensionality
- Opportunities come from utilizing data distribution information
  - Can we learn prior from data? (Domain-specific AutoML)

Development Pipeline

- Parameterize existing data processing solutions to meet the characteristics of healthcare data

Data Acquisition:

Hospital Data
Genome Data
Medical KB
CT/MRI Images

Integration& Augmentation:

AE/D Data Cleaning
Collaborate Analytics
KB Data Enrichment
Image Augmentation

Understanding& Interpretation:

EMR Bias Resolving
EMR Imputation
EMR Embedding
EMR Pattern Mining

Application Deployment:

Standard Model Pool
Adaptive Regularizer
KB Hashing Model
Bagging & Evaluation

Extensive Raw Data

Cleaned Data with Rich Semantics

Extracted Effective Feature Sets

Medical Insights

Enabling Global Optimization

- SINGA – RAFIKI (MLaaS) -- PANDA mainly for healthcare

Apache SINGA
A General Distributed
Deep Learning Platform

	PANDA Healthcare	Current AI systems
Aim	**Defining new AI problems**	**Optimizing for existing AI problems**
Iteration	**Doctors take part in the development circle**	**Data scientists as the agent**
Key Techs	**Efficient declarative interaction**	**ML model and platform**
Domain Knowledge	**Instilled by doctors**	**Understood by data scientists**
Delivery	**Explored together with doctors**	**Plain model outputs**

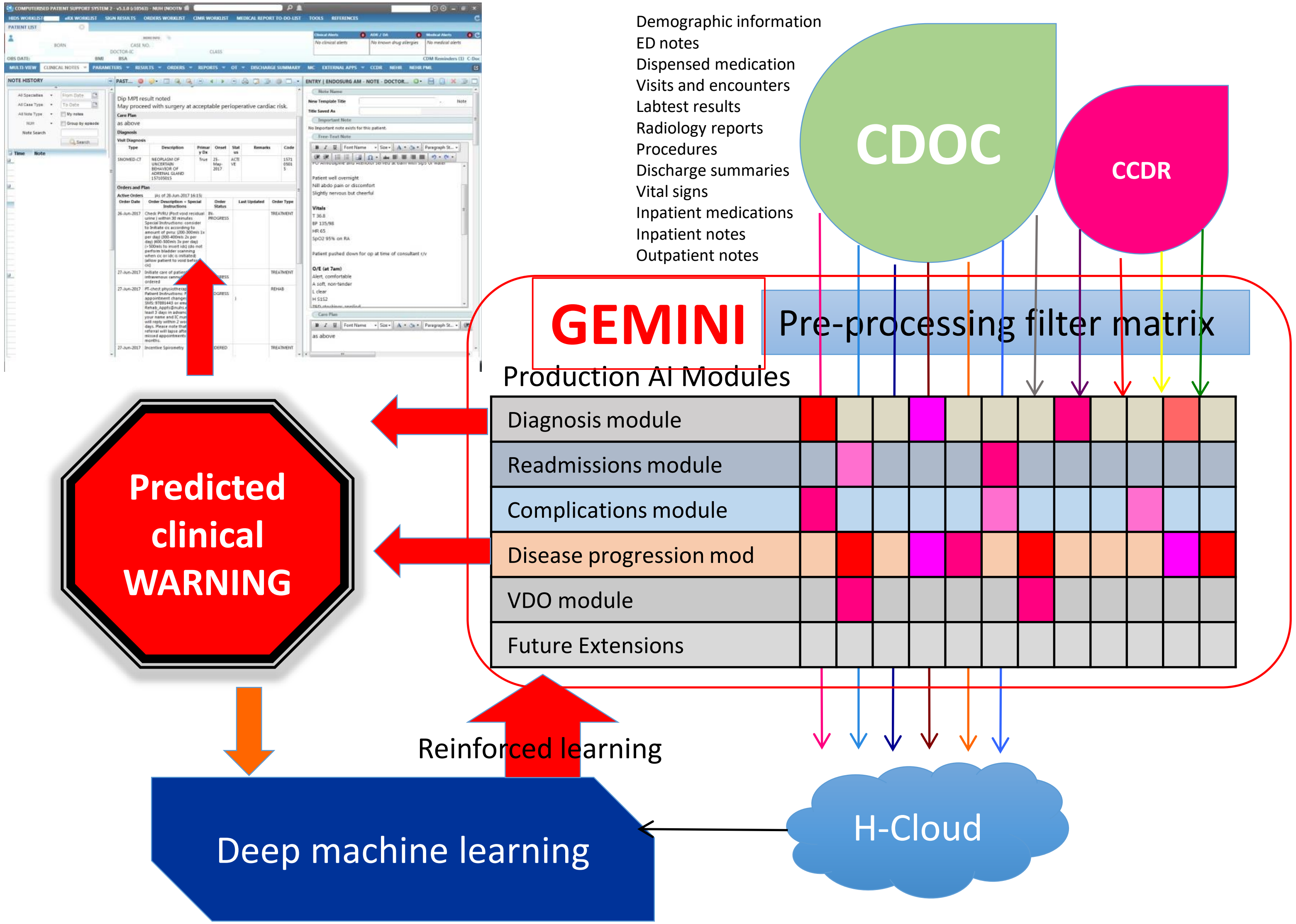
J. Gao, W. Wang, M. Zhang, G. Chen, H.V. Jagadish, G. Li, T.K. Ng, B.C. Ooi, S. Wang, J. Zhou: [PANDA: Facilitating Usable AI Development.](https://arxiv.org/pdf/1804.09997.pdf) https://arxiv.org/pdf/1804.09997.pdf 2018.
W. Wang, S. Wang, J. Gao, M. Zhang, G. Chen, T.K. Ng, B.C. Ooi, J. Shao: Rafiki: Machine Learning as an Analytics Service System. 2018

Healthcare Data Analytics Stack

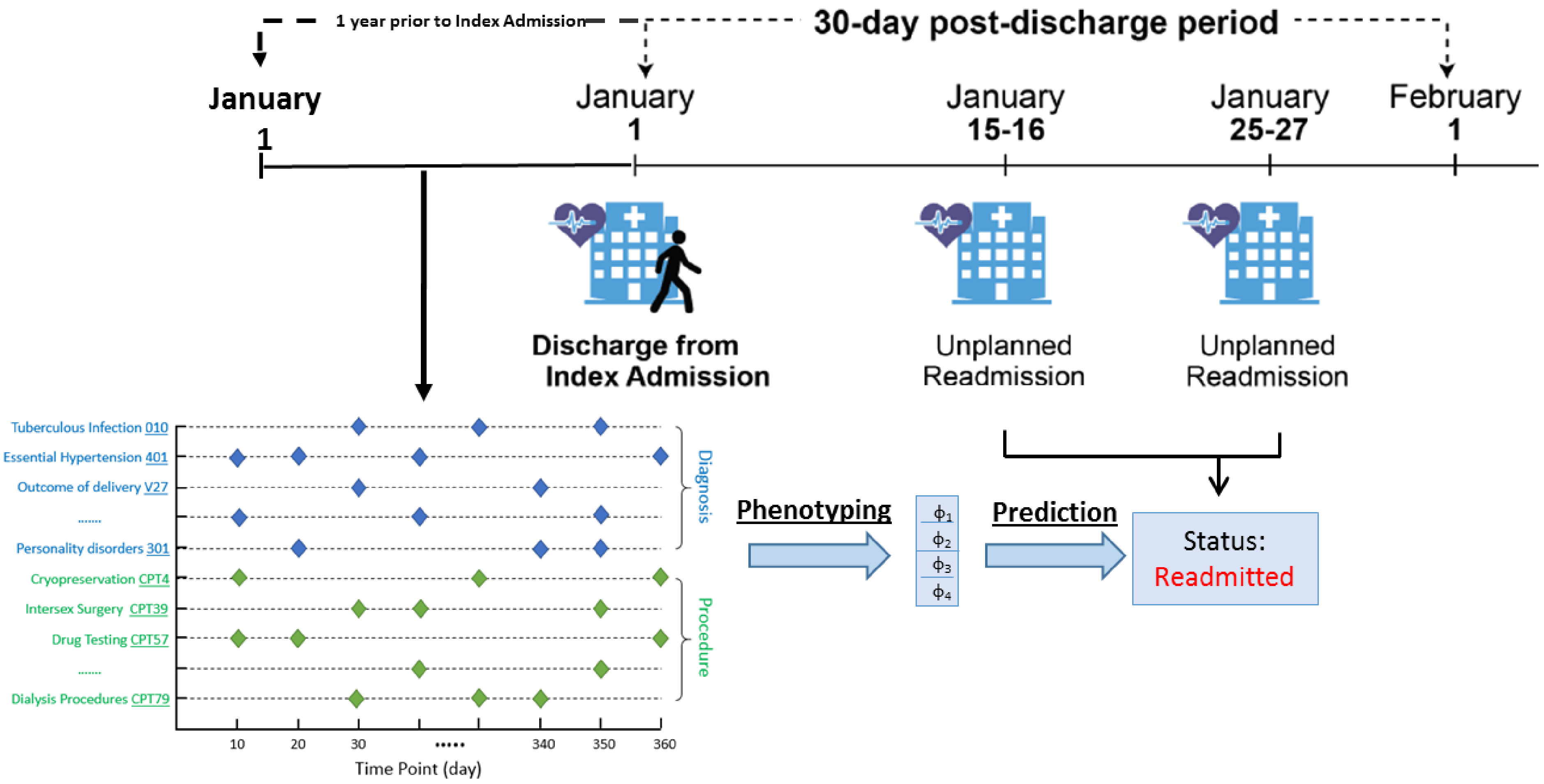
GEMINI (*GEneralisable Medical Information* aNalysis and *Integration* platform)

Z.J. Ling, Q.T. Tran, J. Fan, G.C.H.Koh, T. Nguyen, C.S. Tan, J.W.L. Yip and M. Zhang. GEMINI: An Integrative Healthcare Analytics System PVLDB 7(13): 1766-1771, 2014.

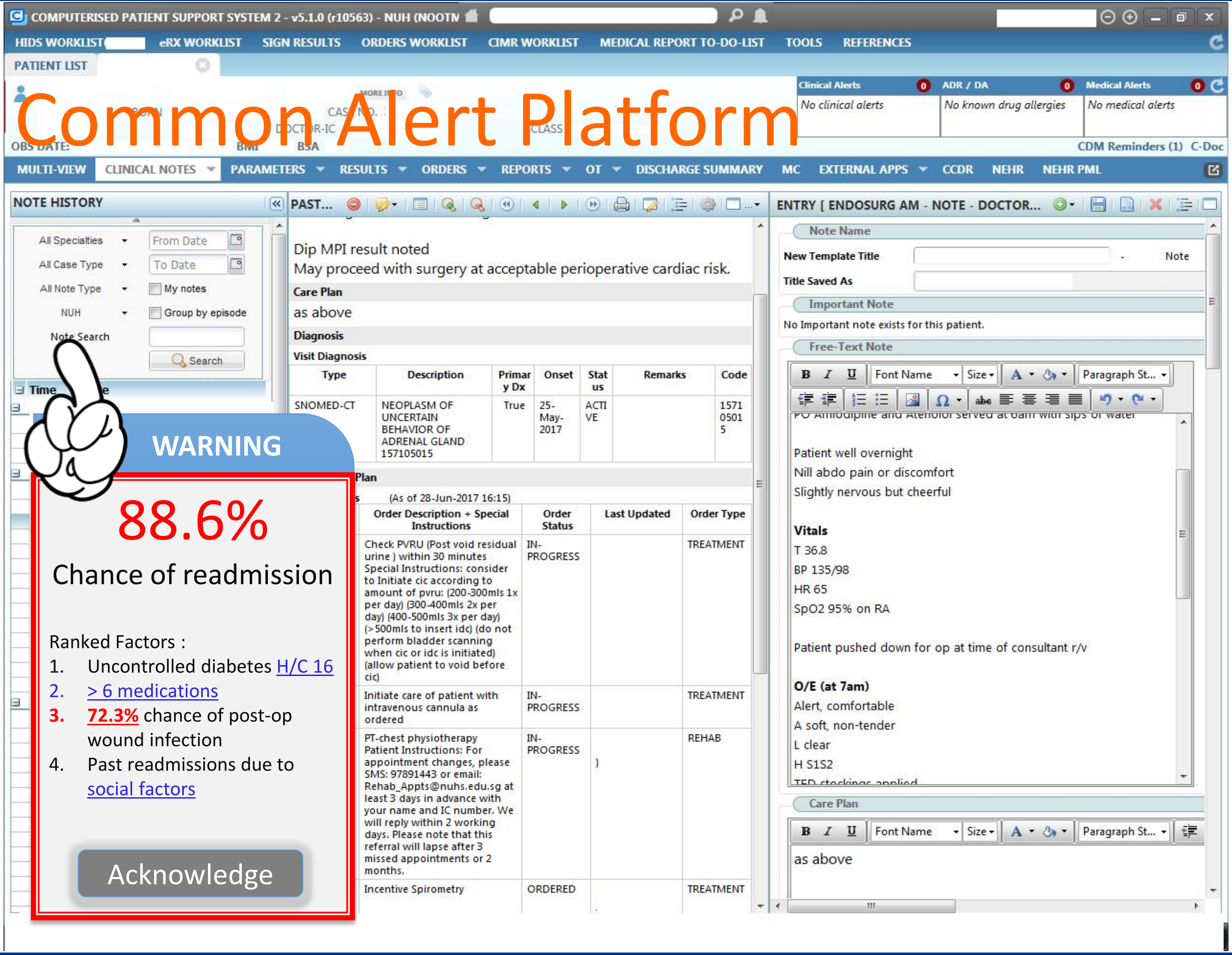
AI Implementation at NUH



Example: Readmission Prediction

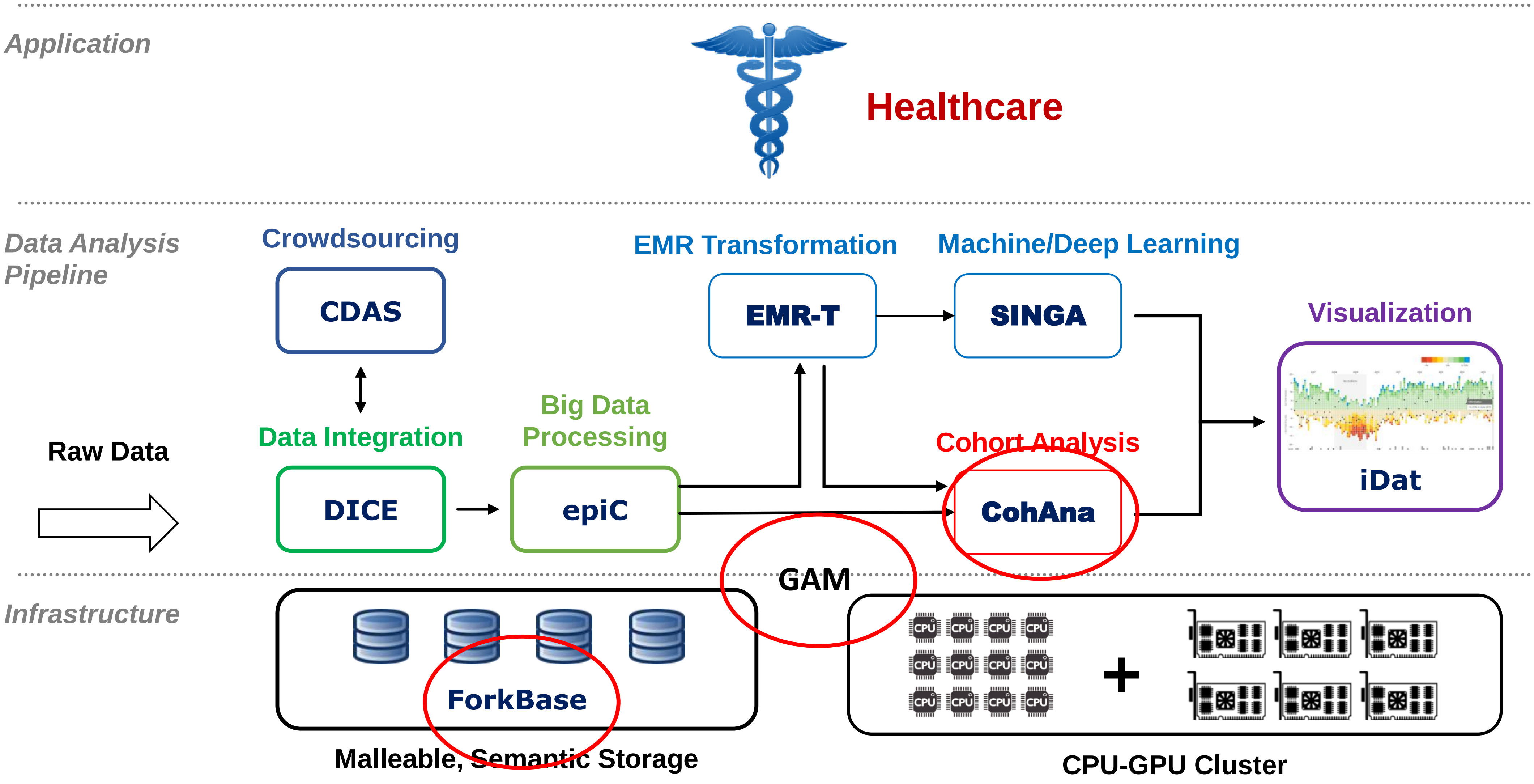


Common Alert Platform



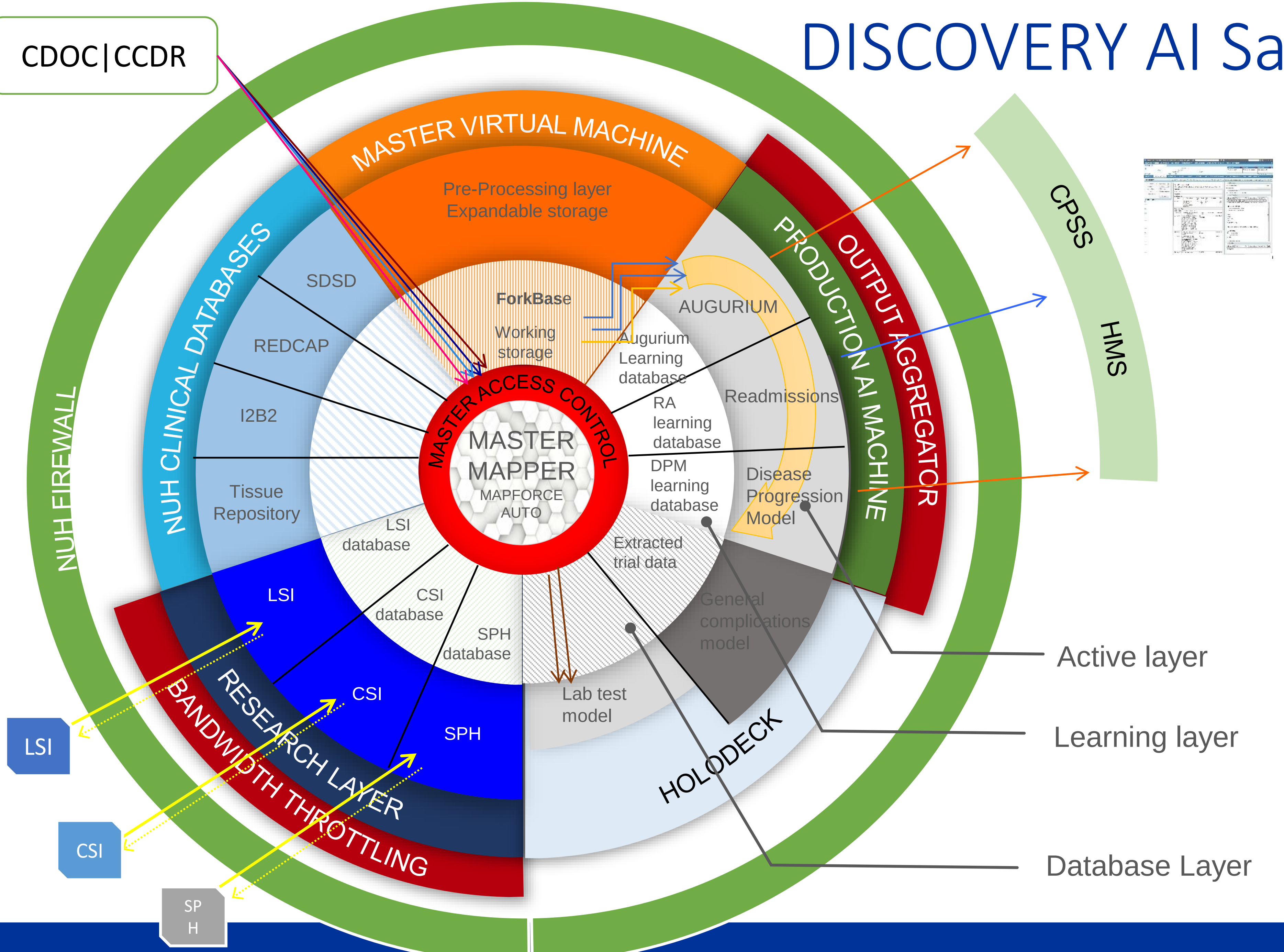
NUH National University Hospital

GEMINI Platform (2011 -)



DISCOVERY AI SandBox

CDOC|CCDR



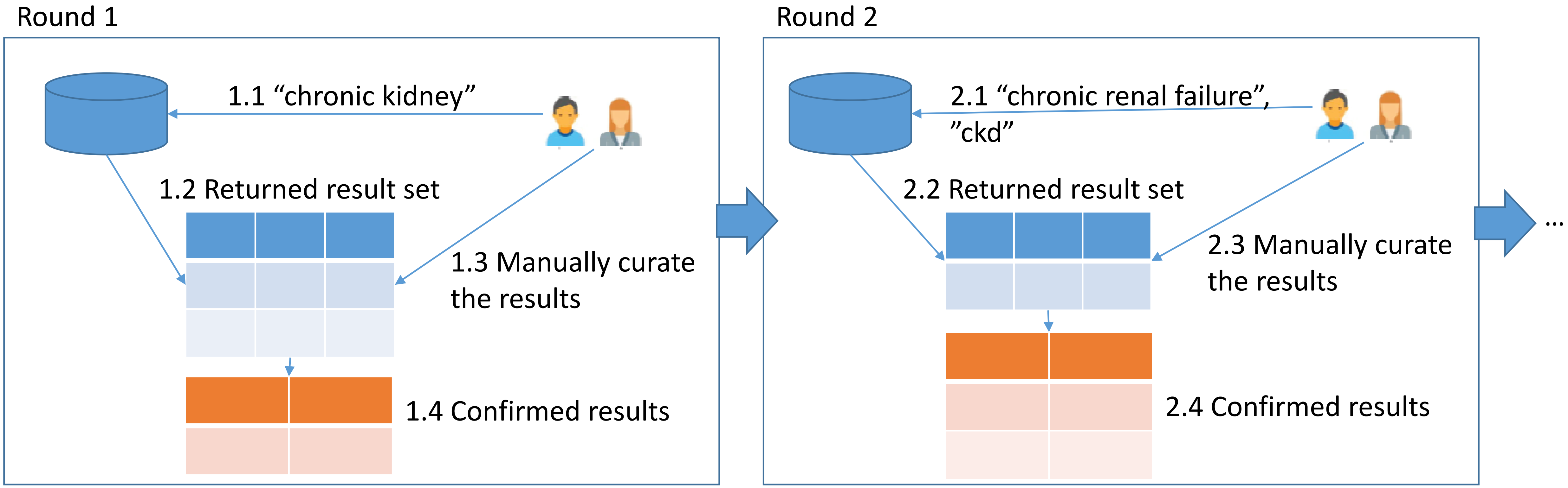
Making Healthcare Data Usable

J. Dai, M. Zhang, G. Chen, J. Fan, K.Y. Ngiam, B.C. Ooi: Fine-grained Concept Linking using Neural Networks in Healthcare. ACM SIGMOD 2018

X. Cai, J. Gao, K. Y. Ngiam, B. C. Ooi, Y. Zhang, and X. Yuan. Medical concept embedding with time- aware attention. IJCAI 2018.

Healthcare Data Usability

If a doctor wants to analyze the medical records related to “chronic kidney disease” …



Healthcare Data Usability

- Two reasons cause the healthcare data usability.
  - Different writing styles.

Real-world healthcare data
2 recent cva
posterior circulation transient ischaemic infarct
multi infarct cva with dementia
massive ischemic stroke with hemorrhagic conversion
acute stroke infarct
2 rt sided cva with gd recovery 1994 5
r groin hematoma
cerebellar stroke
acute left pontine cva
acute cva   left ic laci
acute cva left sided weakness
basal ganglion infarct

refer to →

concept code	Canonical description
I63.50	Cerebral infarction due to unspecified occlusion or stenosis of unspecified cerebral artery

Healthcare Data Usability

- Two reasons cause the healthcare data usability.
  - Different writing styles.
  - Different medical standards.

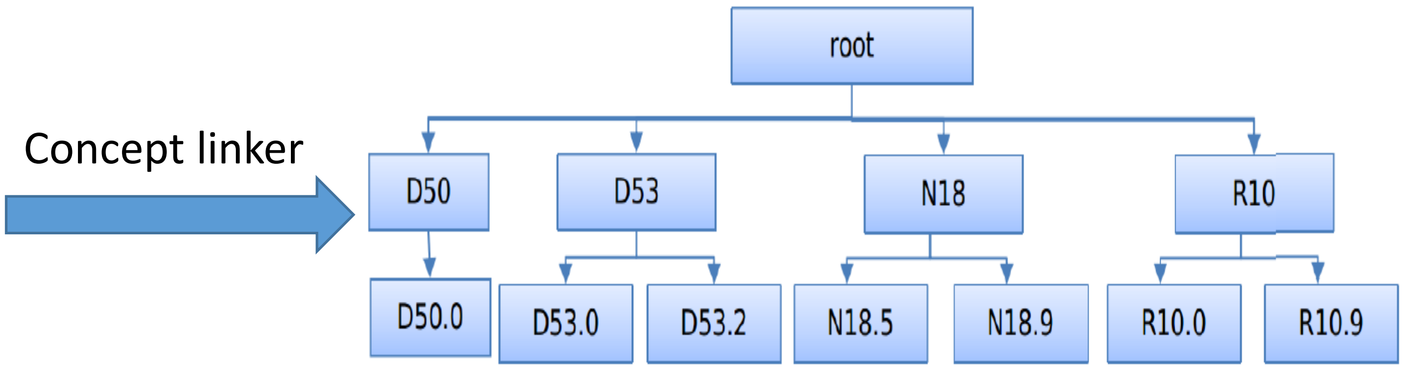
Standard	Concept code	Canonical description
ICD-10-CM	K64.2	Third degree hemorrhoids
ICD-9-CM	455.0	Internal hemorrhoids without mention of complication
ICD-9-CM	455.1	Internal thrombosed hemorrhoids
ICD-9-CM	455.2	Internal hemorrhoids with other complication
ICD-9-CM	455.5	External hemorrhoids with other complication
ICD-9-CM	455.6	Unspecified hemorrhoids without mention of complication
ICD-9-CM	455.7	Unspecified thrombosed hemorrhoids
ICD-9-CM	455.8	Unspecified hemorrhoids with other complication

Real-world healthcare data
internal haemorrhoid prolapsed
haemorrhoid bleeding ligated
3 degree pile
prolapsed haemorrhoid
3rd degree prolasped piles, not thrombosed
thrombosed internal haemorrhoid
3rd degree pile x 1
haemorrhoid
3rd degree external hemorrhoids
hemorrhoids prolapsing piles
haemorrhoids no complication
prolapsed and thrombosed haemorrhoid at 4 clock

Healthcare Data Usability

- Two reasons cause the healthcare data usability.
  - Different writing styles.
  - Different medical standards.
- To improve the healthcare data usability, we need a linker that is able to automatically link a medical record to a unified concept ontology.

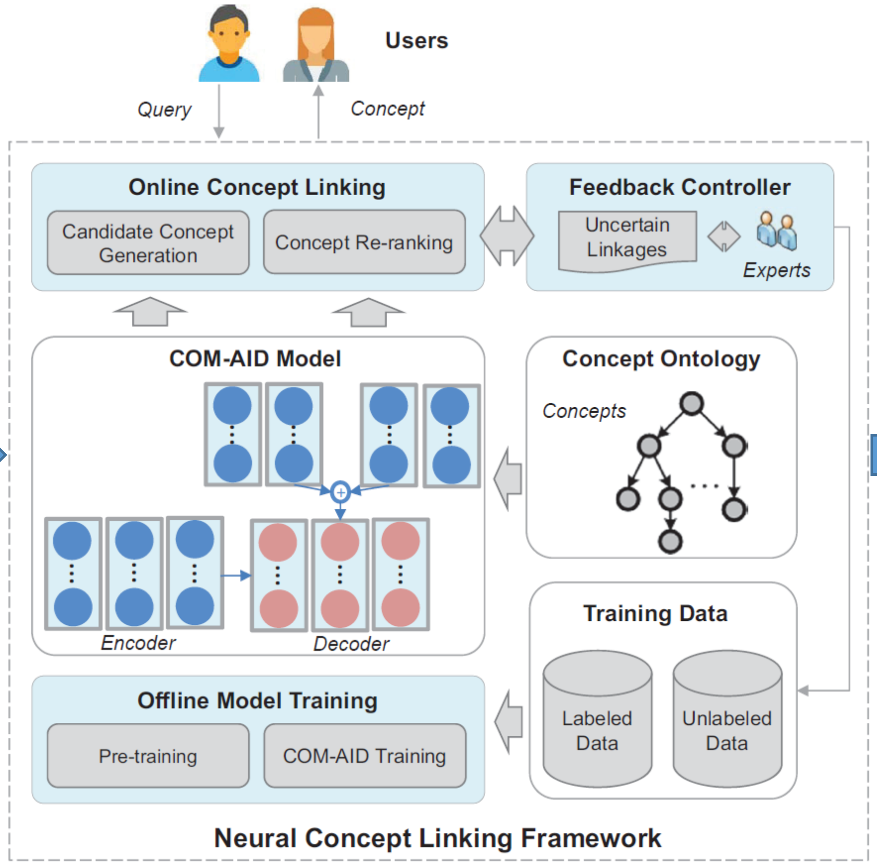
ckd 5
abdomen pain
ckd secondary to dm nephropathy
chr iron deficiency anemia
symptomatic anemia from menorrhagia



Neural Concept Linking

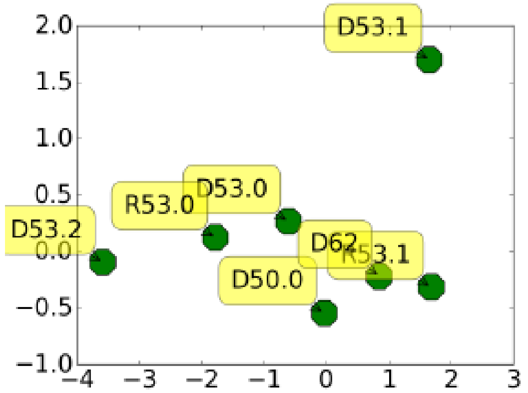
- We have developed a neural concept linking framework to accomplish the healthcare concept linking.

ckd 5
abdomen pain
ckd secondary to dm nephropathy
chr iron deficiency anemia
symptomatic anemia from menorrhagia

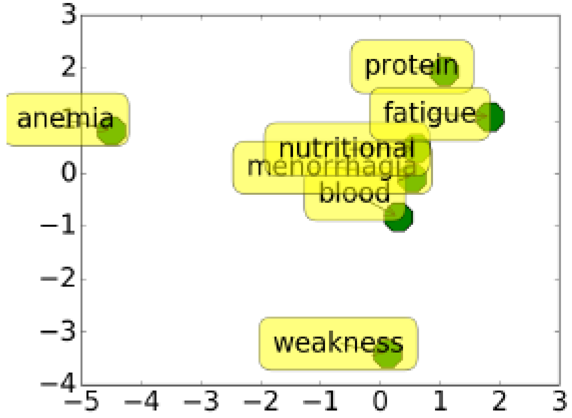


ckd 5	N18.5
abdomen pain	R10.9
ckd secondary to dm nephropathy	N18.9
chr iron deficiency anemia	D50.0
symptomatic anemia from menorrhagia	D50.0

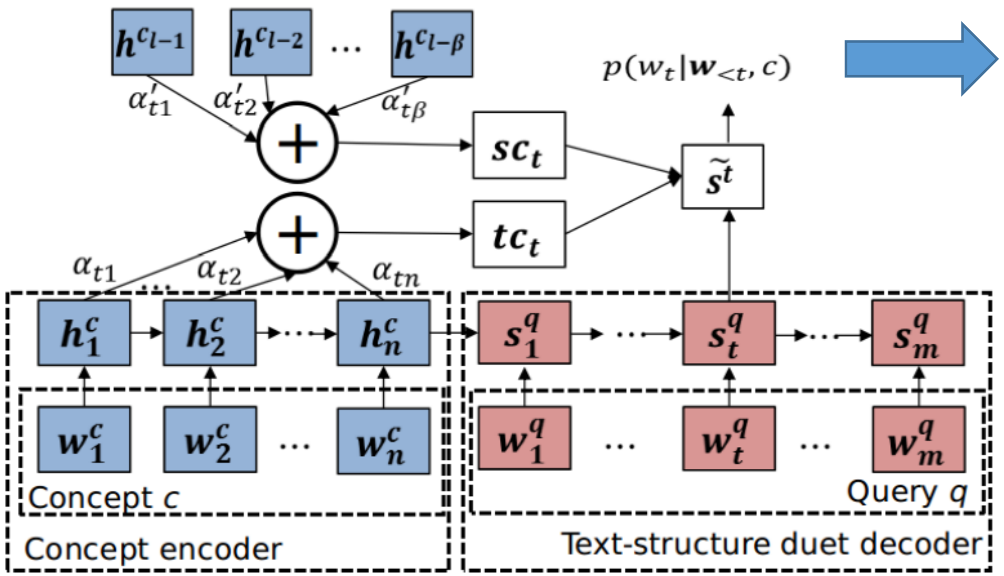
Neural Concept Linking



Concept representations



Word representations

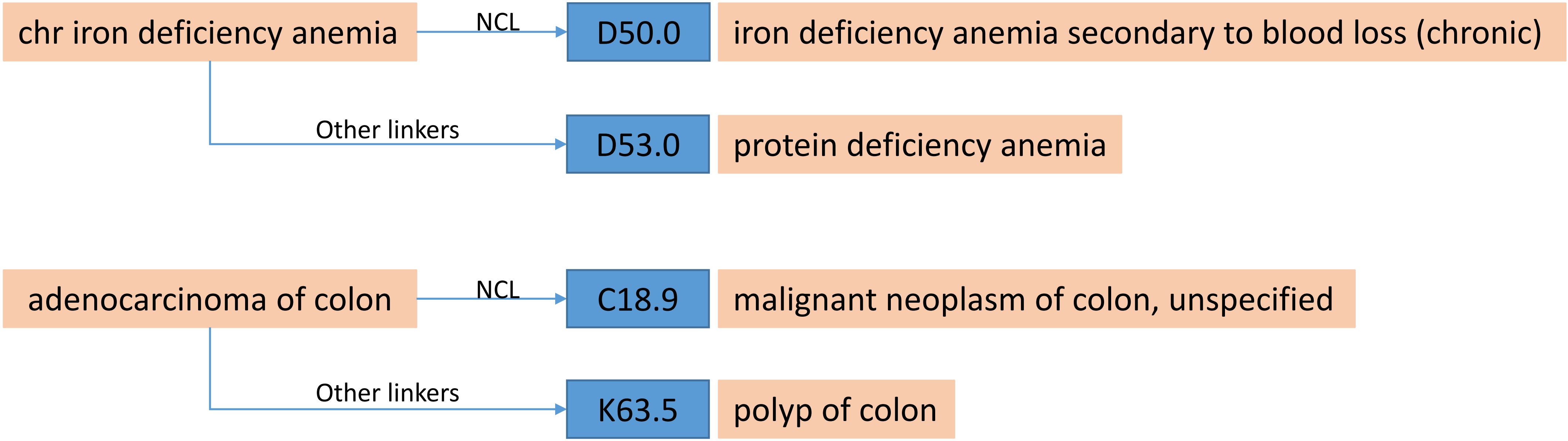


N18.9

p(s|c)=0.016

ckd secondary to dm nephropathy

Example Results



We cleaned 13 years of NUHS data – 90 % done by machine, 10% done by human

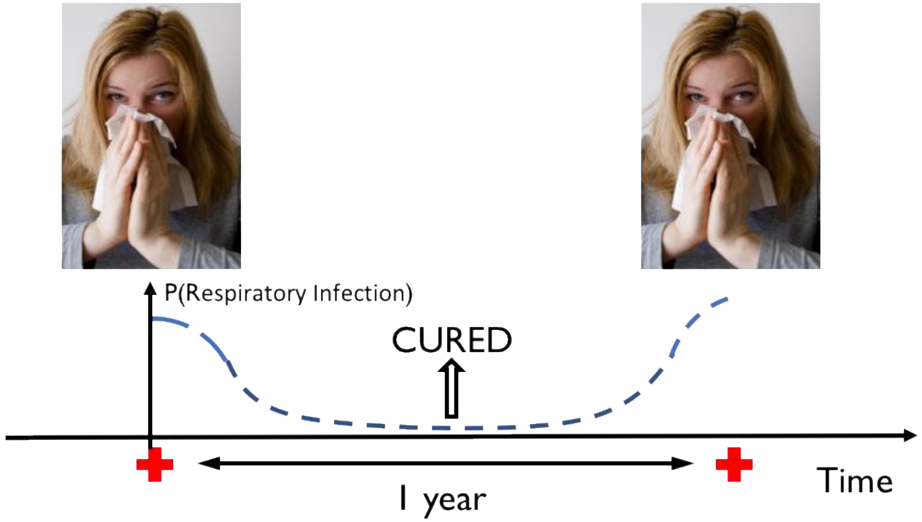
Resolving “bias”

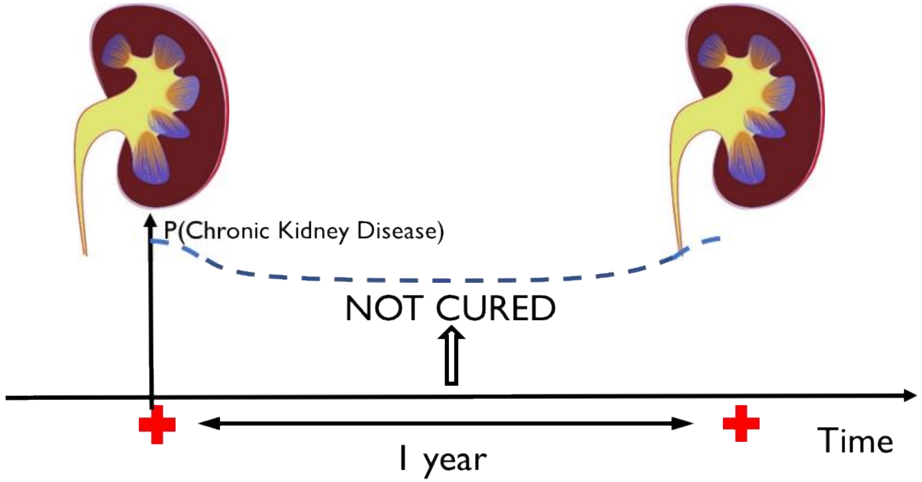
K. Zheng, J. Gao, K. Y. Ngiam, B. C. Ooi and W.L.J. Yip: Resolving the Bias in Electronic Medical Records. ACM KDD, 2017.

Adaptive Lightweight Regularization Tool for Complex Analytics. Z. Luo, S. Cai, J. Gao, M. Zhang, K.Y. Ngiam, G. Chen and W. Lee. ICDE, 2018.

Knowledge Driven Regularization. K. Yang, Z. Luo, J. Gao, J. Zhao, B.C. Ooi, B. Xie. 2019

Similar Pattern and yet Different Results

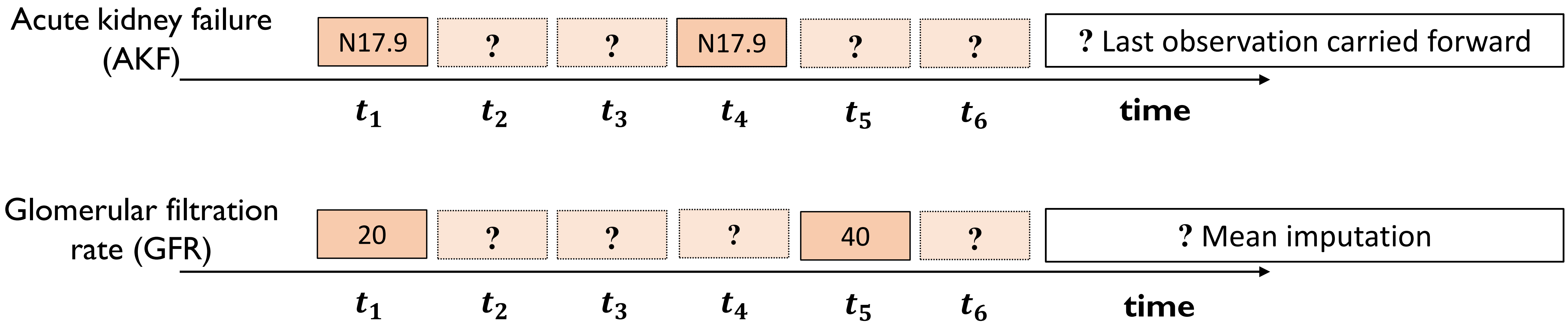




- Patient1 always visits hospital due to respiratory infection
  - Can we conclude that Patient1 has respiratory infection every day?
- Patient2 always visits hospital due to chronic kidney disease
  - Can we conclude that Patient2 has chronic kidney disease every day?
- What is the difference?

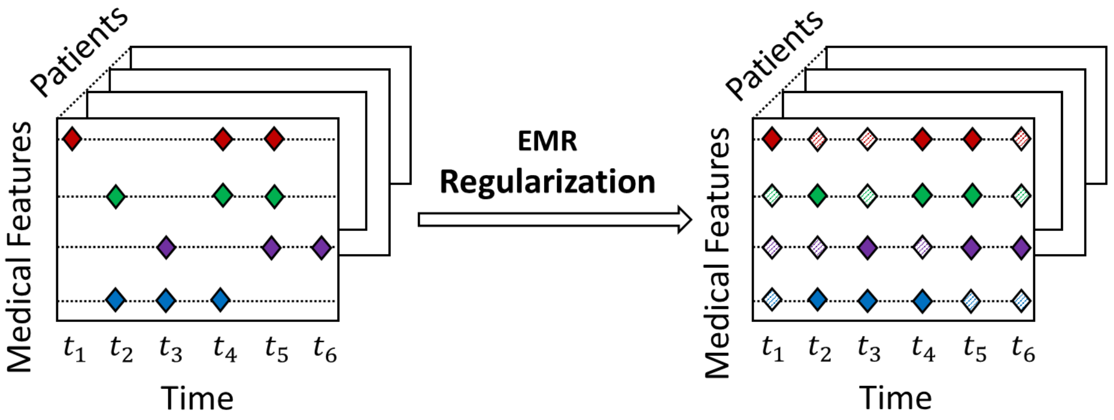
Bias in EMR Data

- If a doctor or analyst want to analyze the EMR data with missing values, they may employ traditional imputation methods directly
- → Misinterpretation

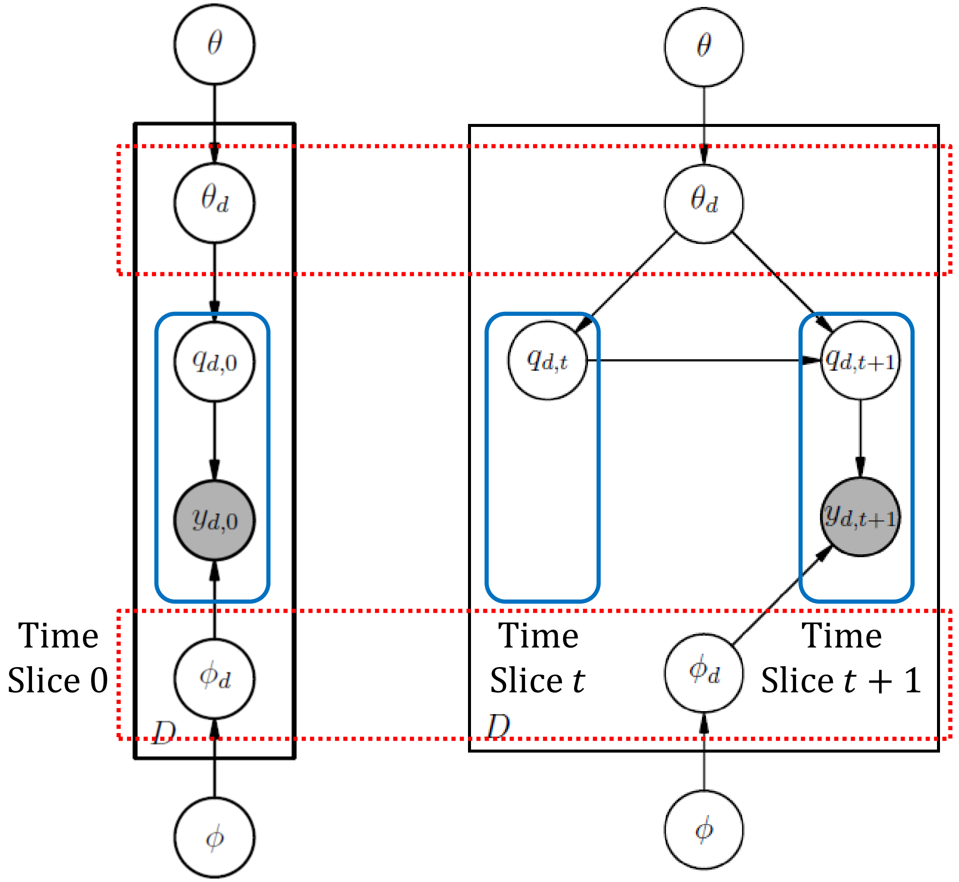


Bias in EMR Data

- Bias – recorded EMR series is different from patients' actual hidden conditions
  - Patients tend to visit hospital more often when they feel sick
  - Doctors tend to prescribe the lab examinations that show abnormality
- ***To Solve Bias Challenge – EMR Regularization***
  - Transform the biased EMR series into unbiased EMR series



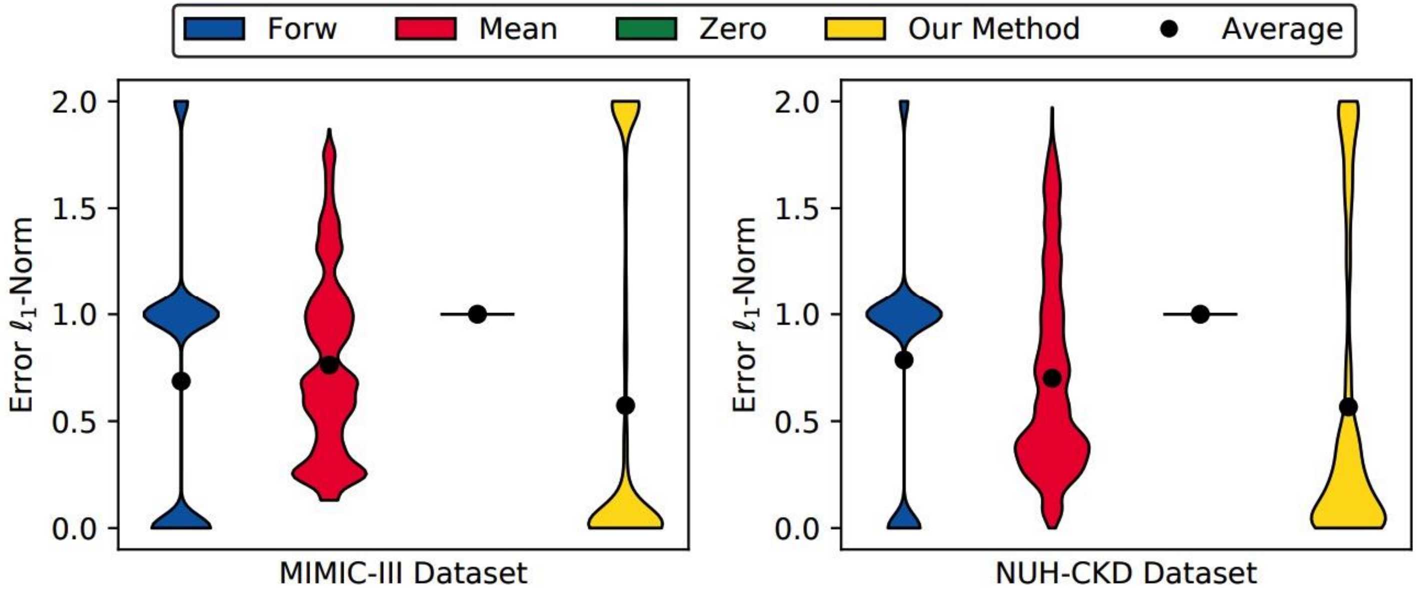
Resolving Bias in EMR Data



- Condition Change Rate (CCR)
  - measures how a medical feature is likely to change from its condition in the previous observation
- Observation Rate (OR)
  - measures the probability that a medical feature is exposed at a time point based on its actual condition at that time point

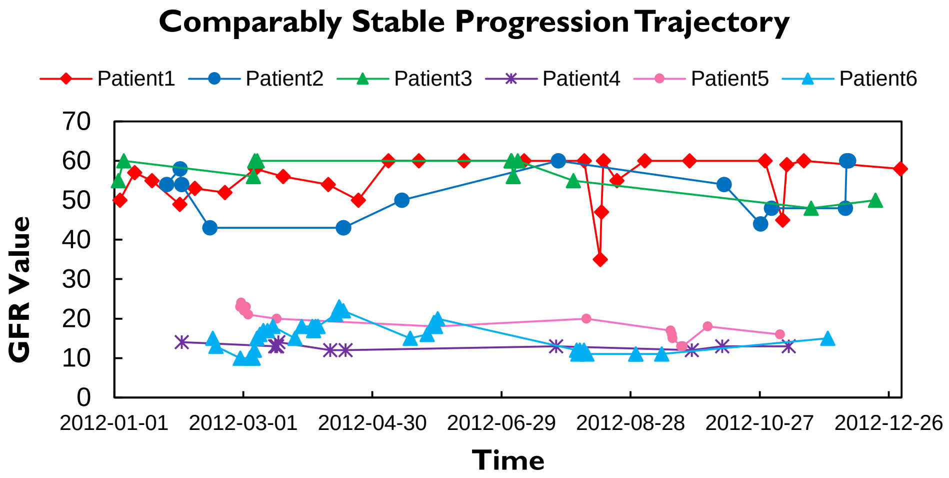
Resolving Bias in EMR Data

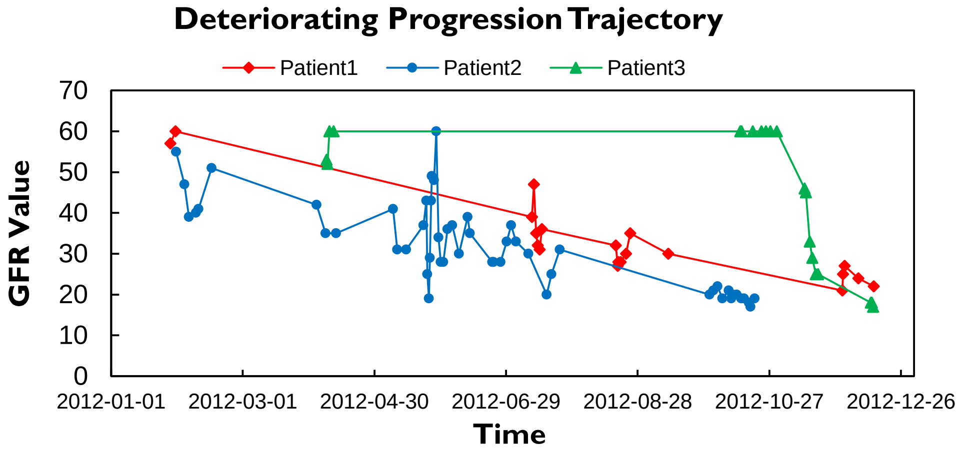
- Imputation accuracy evaluation

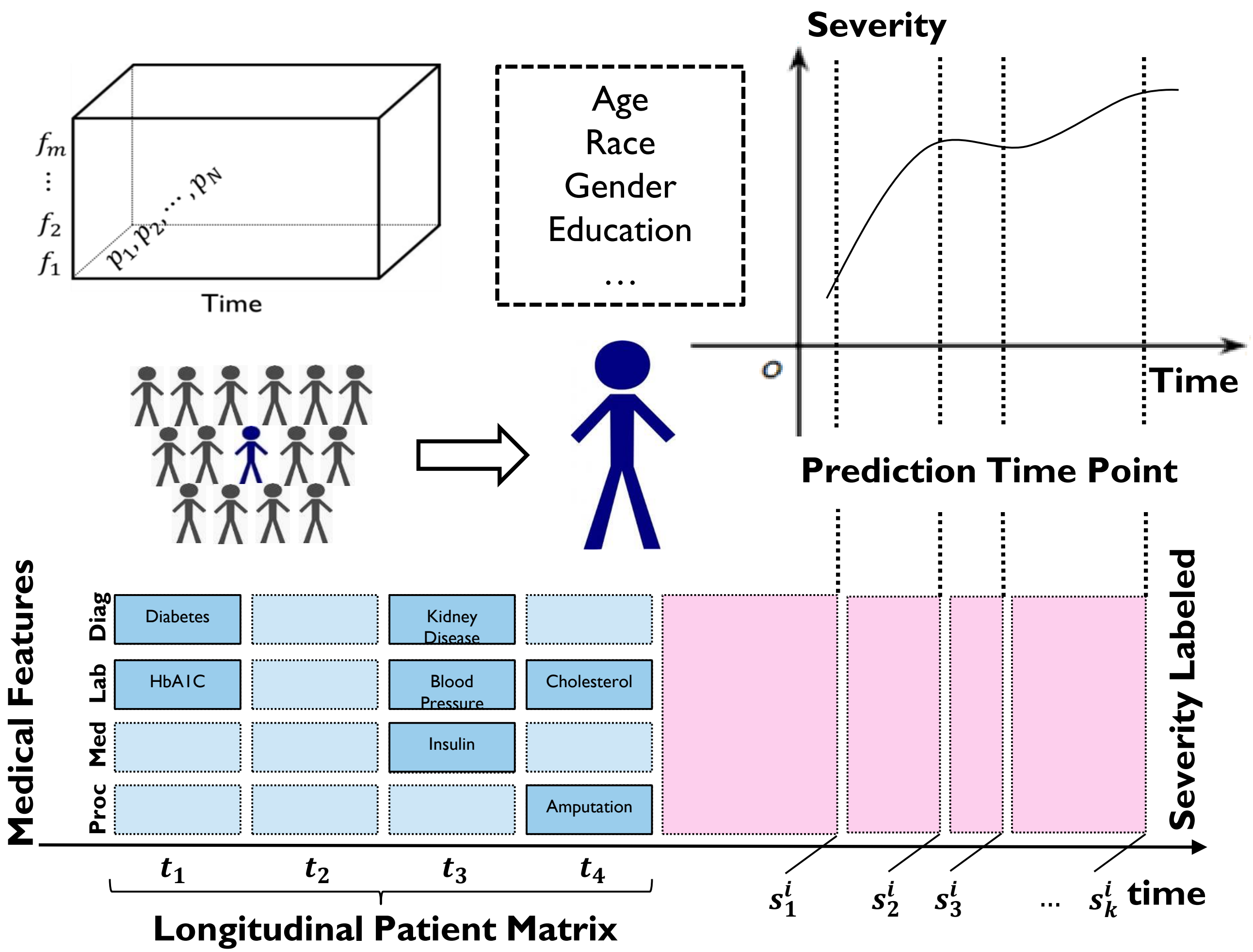


- Benefits for analytic tasks
  - In-hospital mortality prediction, Diagnosis by category prediction
  - Disease progression modelling

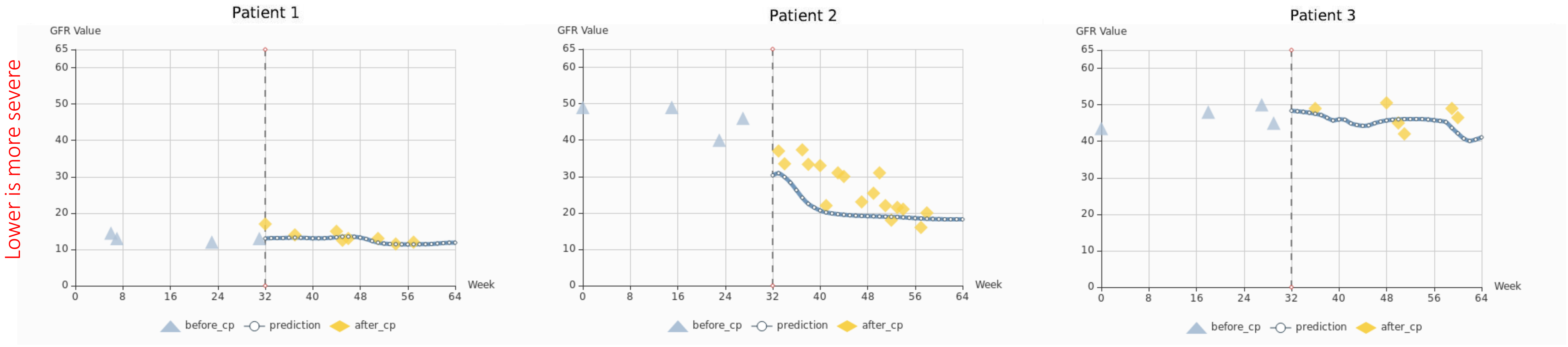
Disease Progression Modeling







Advice to Doctors on Intervention



Powered by GEMINI

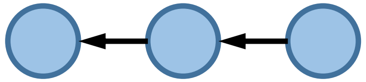
- Our model would suggest to guarantee the monitoring for Patient 1 → may need dialysis or kidney transplant
- Our model would suggest healthcare workers to provide more aggressive interventions to Patient 2 in advance
- Our model would suggest to guarantee the monitoring for Patient 3

Facilitating Data Sharing and Provenance

S. Wang, T. T. A . Dinh, Q. Lin, Z. Xie, M. Zhang, Q. Cai, G. Chen, B.C. Ooi, P. Ruan: ForkBase: An Efficient Storage Engine for Blockchain and Forkable Applications. VLDB 2018

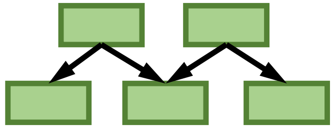
ForkBase Designs

Versioning &
Tamper Evidence



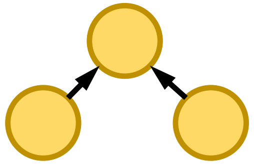
Merkle DAG

Indexing &
Deduplication



SIRI indexes

Collaboration
Workflows

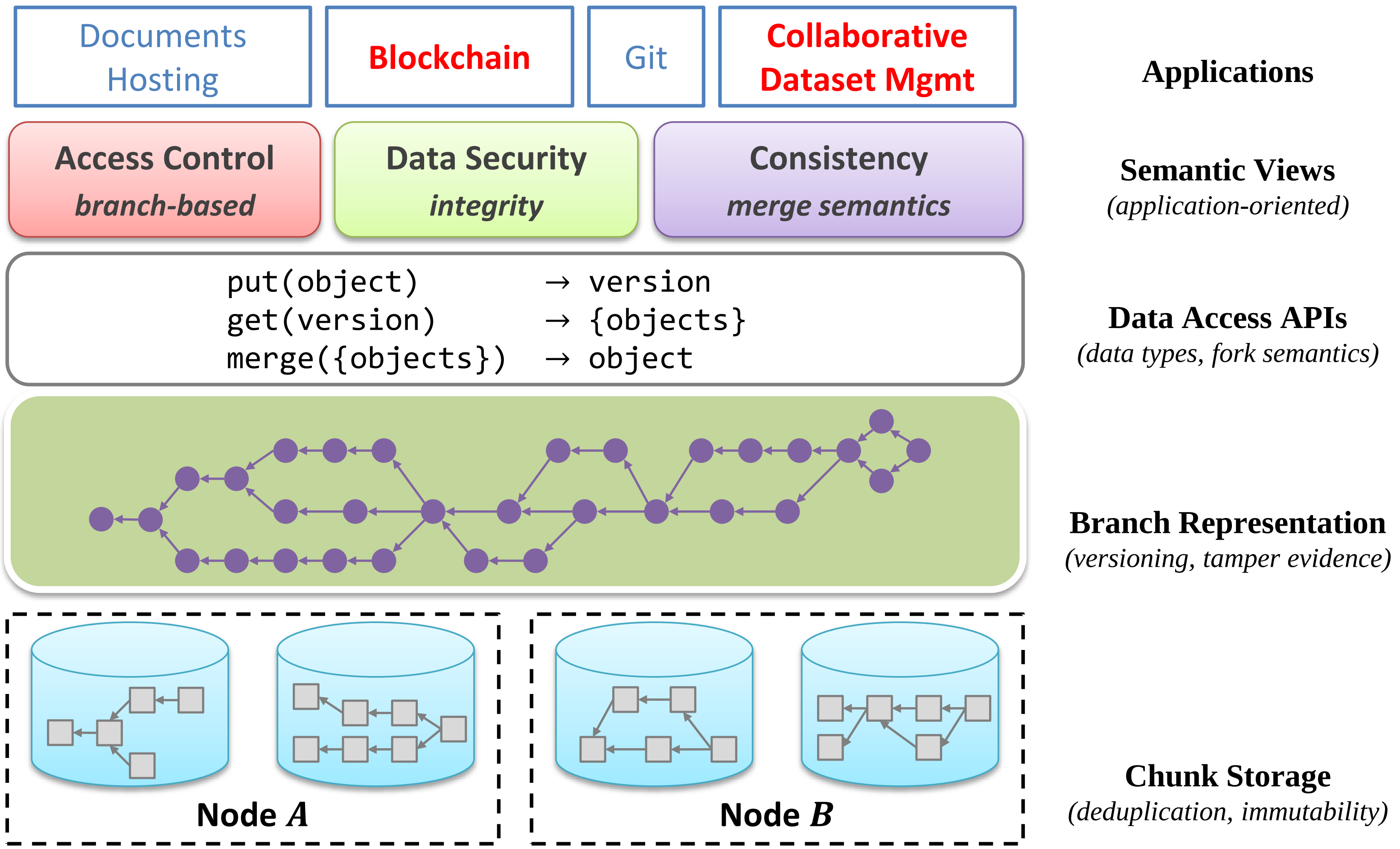


Fork Semantics

ForkBase

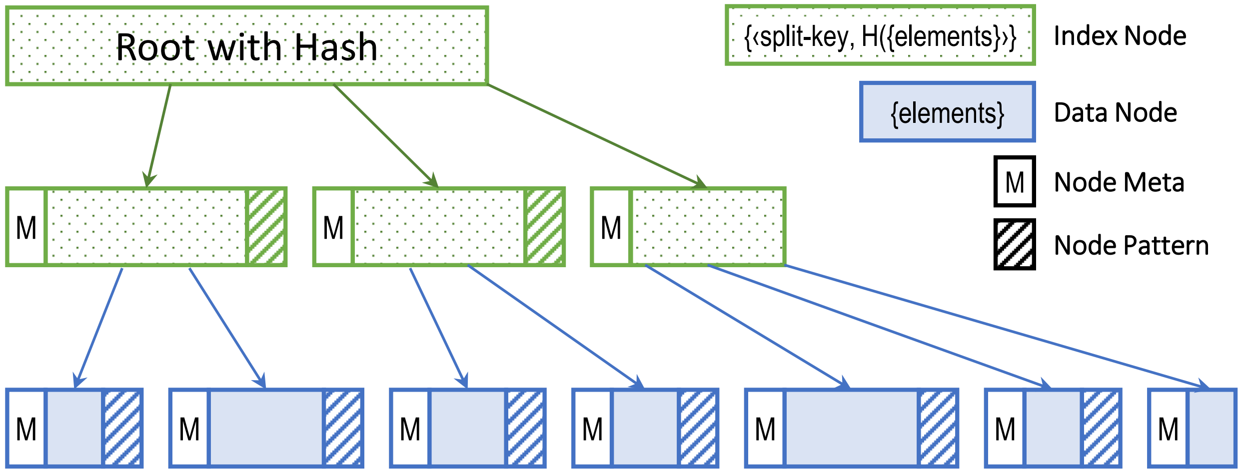
git (versioning) + **database** (query) + **blockchain** (integrity)

ForkBase Storage Stack



SIRI Indexes & POS-tree

- An Index Class: Structurally-Invariant Reusable Indexes
  - Structurally Invariant, Recursively Identical, Universally Reusable …
- An Implementation: Pattern-Oriented-Split Tree



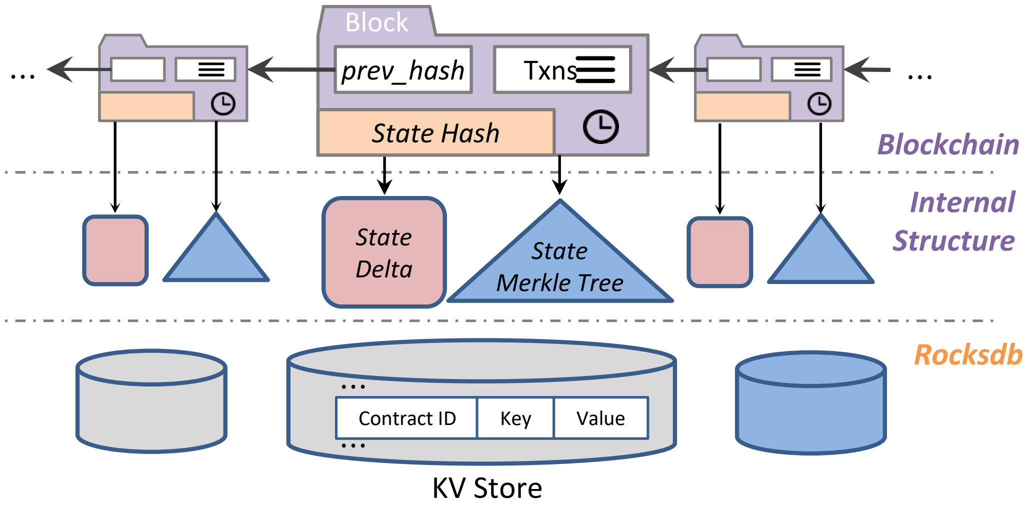
Content-determined Structure
(-> Deduplication)

Native Merkle Tree
(-> Tamper Evidence)

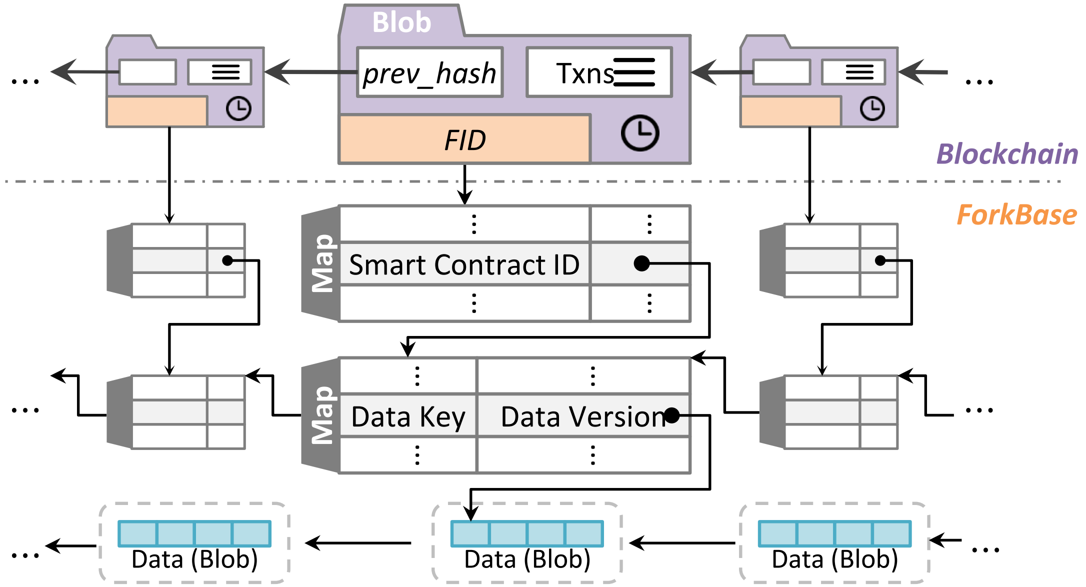
Probabilistically Balanced Tree
(-> Query Efficiency)

Blockchain Data Model in ForkBase

- KV Store
  - Customized structures
    - Linked block
    - State Merkle tree
    - State delta
    - …
  - Hard to implement

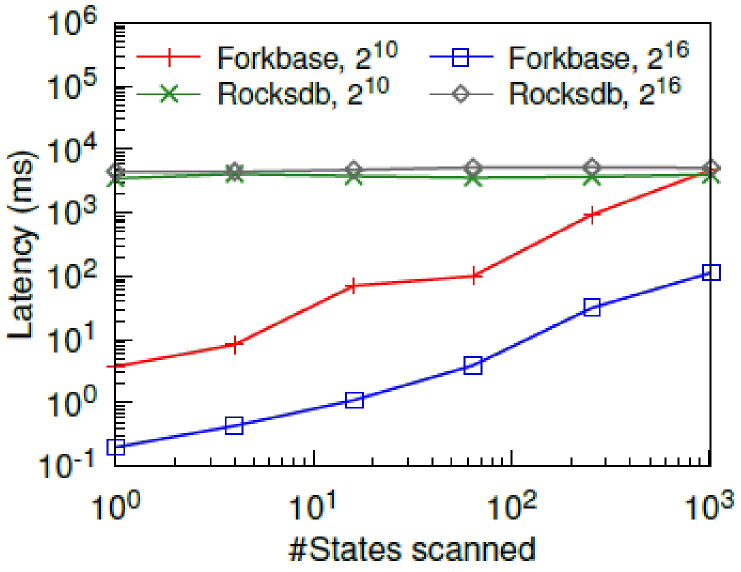


- ForkBase
  - Achieve with built-in types
    - UBlob
    - UMap
    - …
  - Easy to maintain
    - 10+ lines for each structure

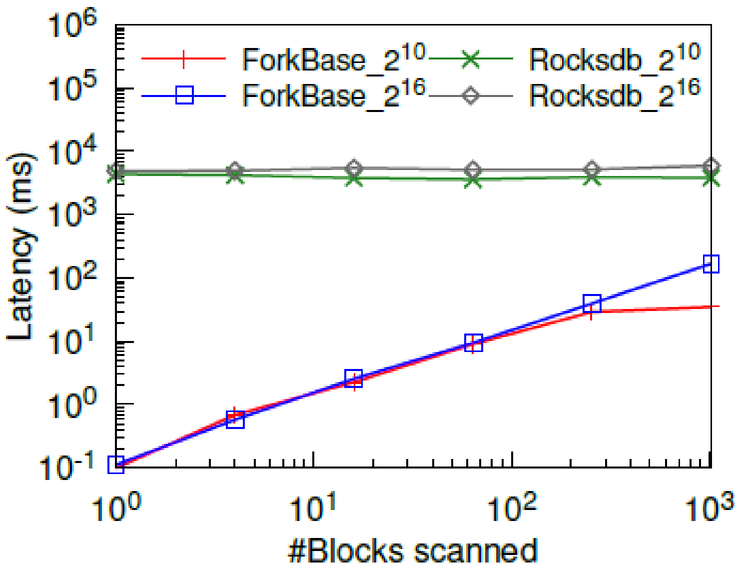


Analytic-Ready Blockchain Backend

- Analytic on blockchain is expensive
  - Need to scan whole block history to extract information
- Built-in data types in ForkBase to support fast analytics



State Scan Query



Block Scan Query

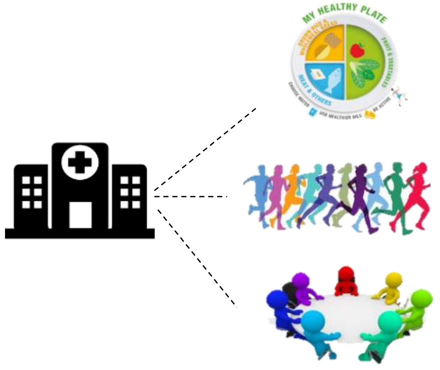
Prevention is Better Than Cure

L. Long, W. Wang, J. Wen, M. Zhang, Q. Lin, B.C. Ooi: Object-Level Representation Learning for Few-Shot Image Classification. arXiv preprint arXiv:1805.10777. 2018

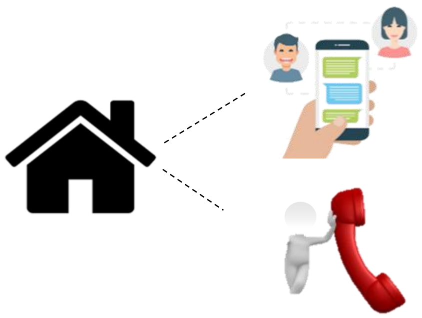
Lifestyle InterVENtion Programme (LIVEN)

The effect of a behaviour-based lifestyle change program using combined face and remote sessions on weight, diet intake and physical activity level in people at-risk of diabetes: a Randomised Controlled Trial

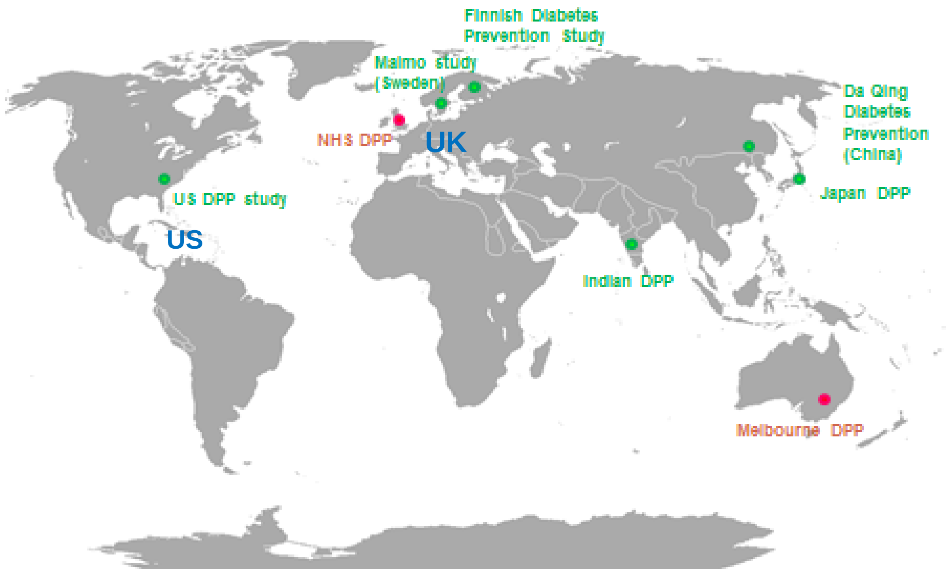
Face to Face Sessions



Remote Sessions



Diabetes Prevention Programme



Effecting Behavioral Change

Snap

- Quick and Easy way to record dietary intake
- A deep learning image-based food recognition for a faster, closest food match and handy recording

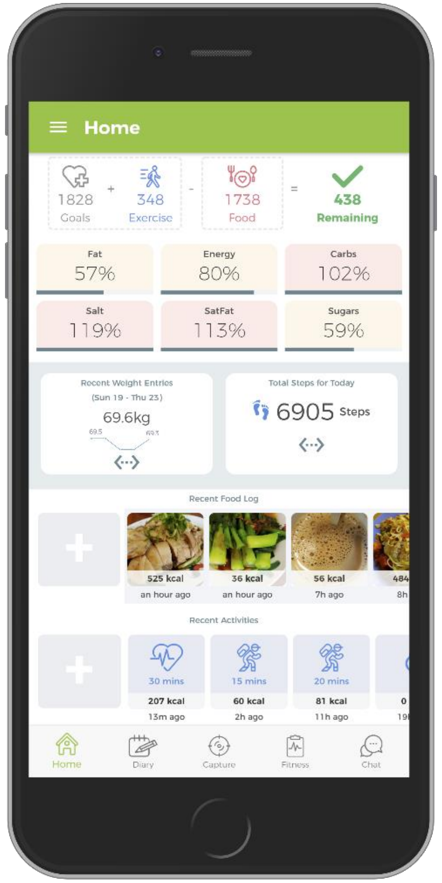
Track

- Self-monitoring with pre-set goals and intuitive nutrition information
- Peer-to peer monitoring of dietary and physical activity goals
- Daily and weekly reports of progress

Feedback

- Remote monitoring by healthcare professionals for timely and meaningful feedback

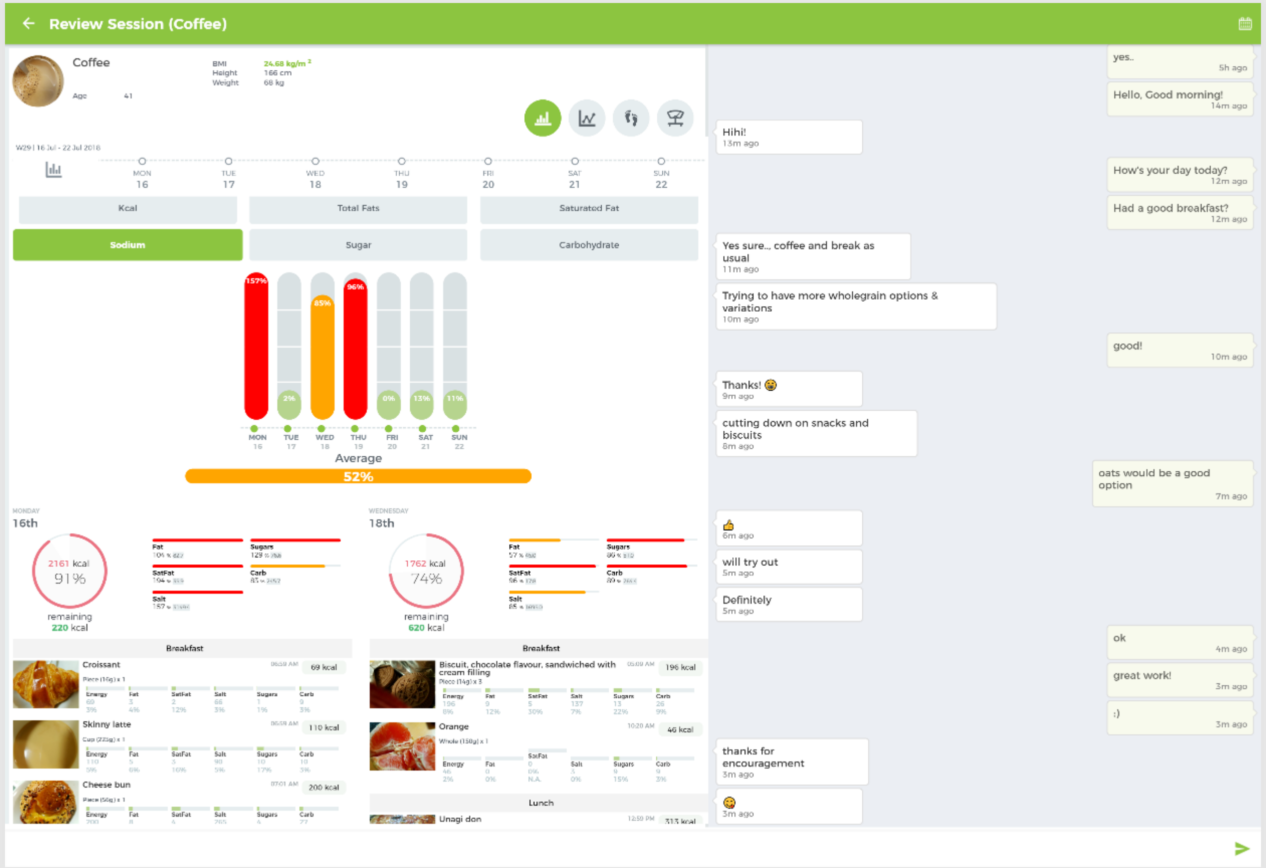
Diabetes Prevention



Healthy Diet + Exercise

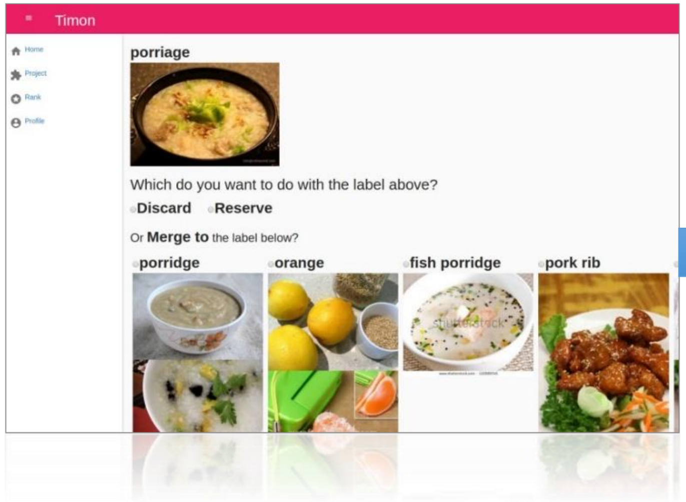
Administrator/Dietician Portal

- **Dietary Review + Chat**
  - Review user's weekly meal (photo) history



Realtime Chat with Dietician
provides instant feedback to users

Foodhealth/Foodlg



STEP 1

Collect training images from heterogeneous sources and label them via crowdsourcing

STEP 2

Train deep learning models for food recognition

STEP 3

Food recognition and health analysis using images and other information from the Foodlg app

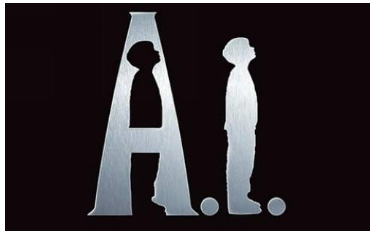
Off-line

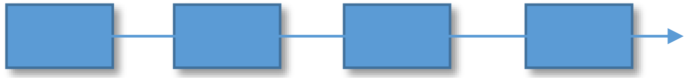
On-line

Personalizing and Decentralizing Healthcare

AI + BlockChain + Cloud + big Data

Analytics/
DataScience





BigData/
DBMS

Objectives:

1. Transparency
2. Accountability
3. Auditability
4. Governance
5. Security
6. …

BlockChain enabled Healthcare

- BlockChain (BC) acts as a tamper-evident storage for archiving Healthcare Records from different healthcare providers
- BlockChain acts a “Central Healthcare Record Repository”
- It enables Data Provenance, Data Analytics, and Medical-care everywhere based on patient’s preference
- It may help transform Healthcare management and research

NUHS
National University Health System

山东省远程医学中心
TELEMEDICINE CENTER OF SHANDONG PROVINCE

杭 州 市 萧 山 区 中 医 院
浙江中醫藥大學附屬江南医院

KINGCOME
SHENZHEN KINGCOME HOSPITAL
深 圳 静 康 医 院

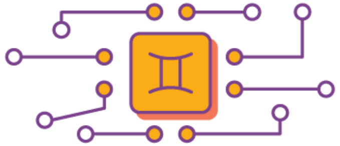
The MediLOT Solution

1. Holistic

Every patient will have a complete longitudinal health record: their own health story that they can access at any institution

2. Patient-centric

The patient holds his/her own private key and has fine control over who can view their medical records



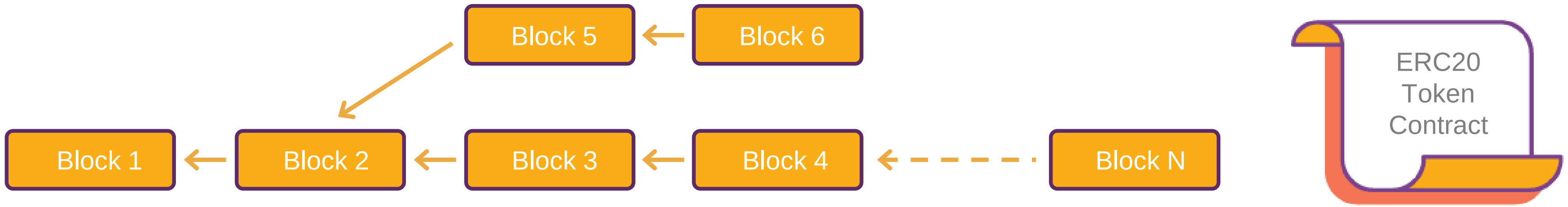
3. Personalised

Using an advanced analytics overlay (**GEMINI**), MediLOT facilitates personalised treatment strategies

4. Decentralised

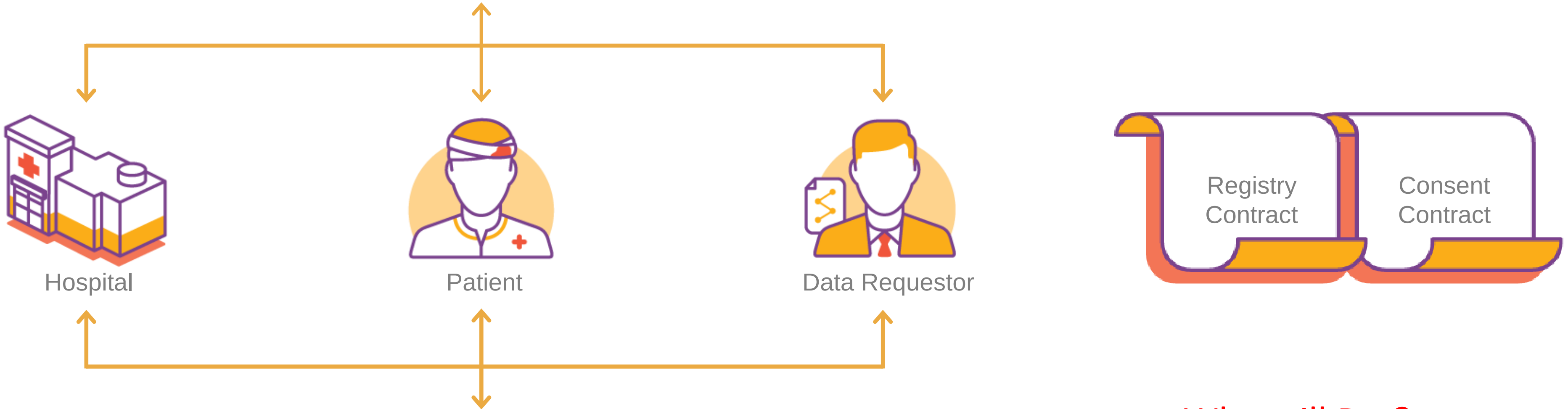
Patients’ data is stored in different locations, eliminating the risk of a single catastrophic breach

Dual BlockChain Schema



Public (Ethereum)

Allows for transfer and crediting of ERC20 LOT tokens (MediLOT utility token)

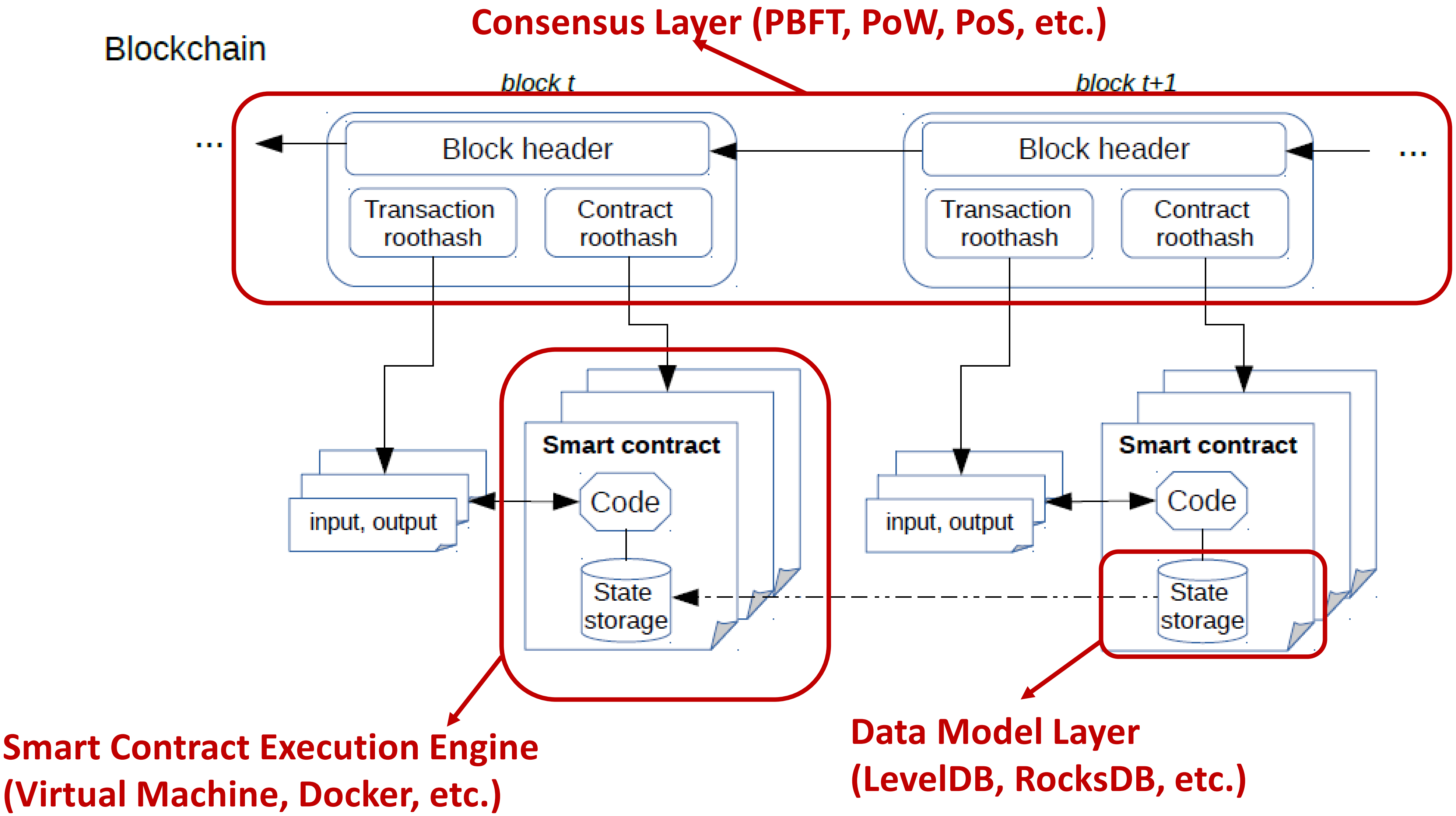


Permissioned (Hyperledger++)

Who will Pay?

Responsible for aggregation of patient EHR

On-Chain Scalability



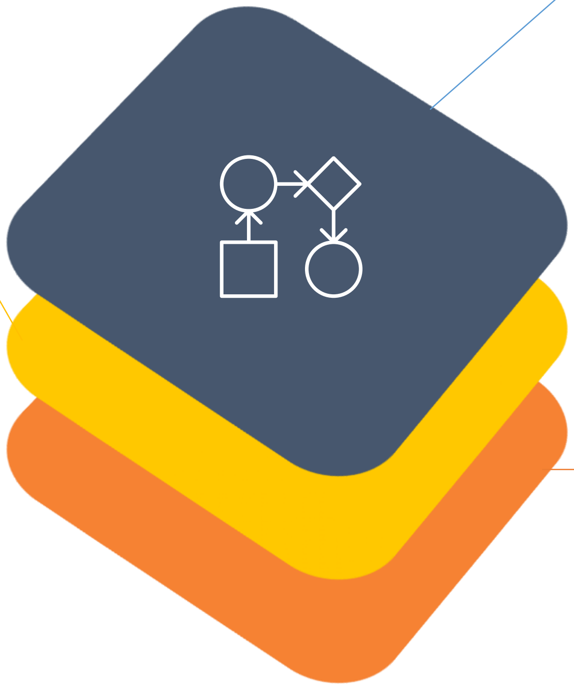
Dinh, J. Wang, G. Chen, R. Liu, B. C. Ooi, K.-L. Tan: BLOCKBENCH: A Framework for Analysing Private Blockchains. ACM SIGMOD 2017
A. Dinh, R. Liu, M. Zhang, G. Chen, B.C. Ooi, J. Wang: Untangling Blockchain: A Data Processing View of Blockchain Systems. IEEE TKDE, 2018.

MediLOT's Technologies

Dual Blockchain

Ethereum & Hyperledger++

- Enhanced Hyperledger with scalable consensus and sharding
- Throughput up by 15x



Analytics

GEMINI

The underlying healthcare suite that supports big data analytics and personalised medicine

Data Storage

ForkBase

Proprietary storage with rich semantics, immutability and data sharing, Blockchain optimised native storage system

Conclusions

Minority Report In Healthcare?

- Healthcare is a complex but impactful/meaningful Application
  - Domain Knowledge
  - Verification and Validation – a tedious process
- A good (example) application that calls for better integration of AI/ML and Database technologies, and possibly Blockchain technologies
- We have addressed some of the challenges, and have implemented:
  - GEMINI (DICE, CDAS, epiC, Apache SINGA, ForkBase) is being used by 2 major hospitals in Singapore
  - Foodhealth (foodlg) is used by 3 hospitals in Singapore
  - MediLOT is in testnet phase and used by hospitals in China
- Objectives:
  - To predict, prevent/pre-empt, personalize for more effective healthcare
- Be Good. If you can’t, be Safe. Live well …

Acknowledgements

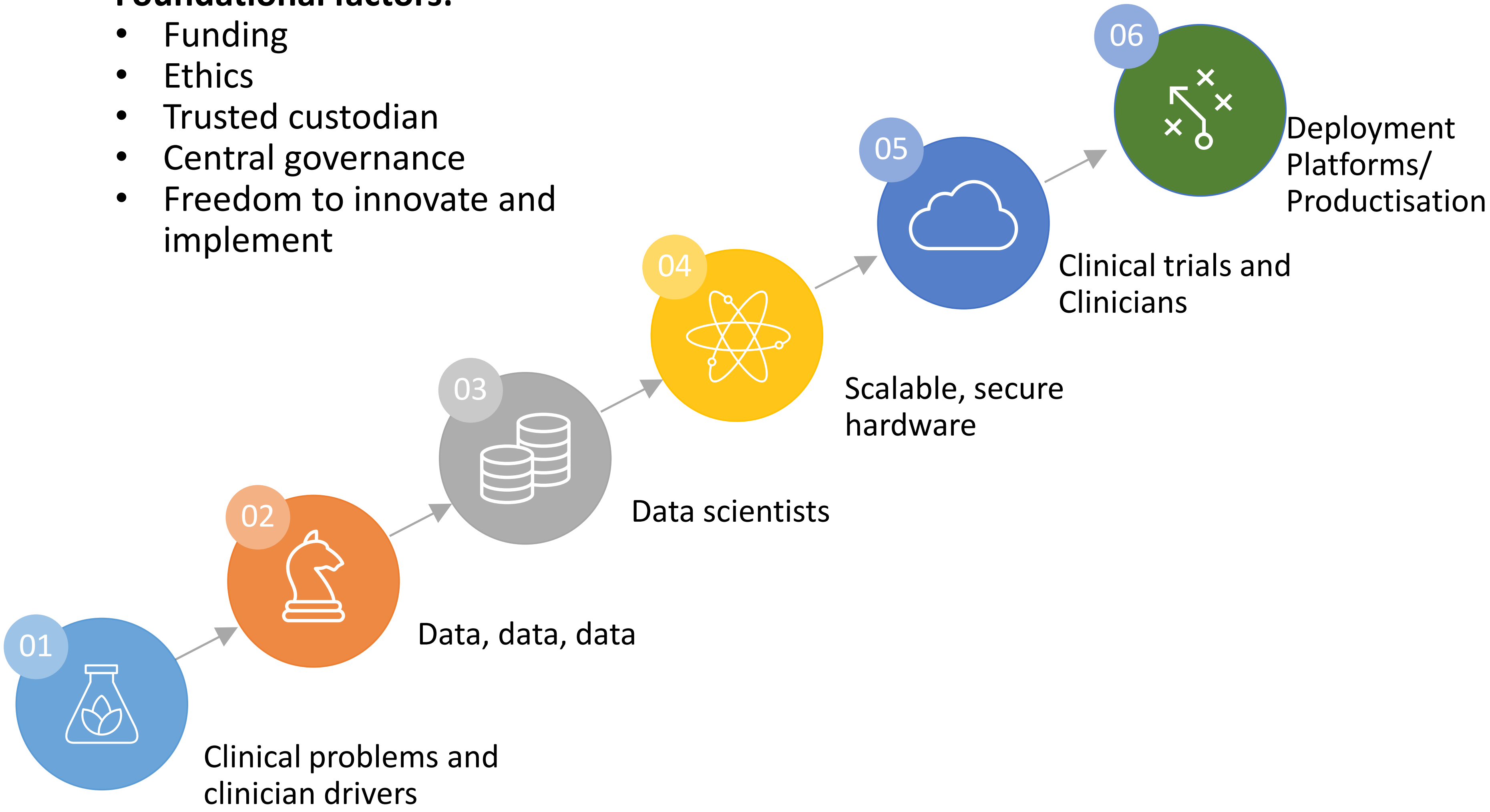
- Collaborators: Gang Chen, H.V. Jagadish, Kee Yuan Ngiam, James Yip++
- Collaborators (ex-students): Meihui Zhang, Wei Wang, Jinyang Gao, Chang Yao
- **Visitors**: Divy Agrawal, H.V. Jagadish, Dave Maier, Renée Miller, Tamer Özsu, Amit Sheth, Wang-Chien Lee, Wang-Chew Tan, Ju Fan, ++
- Current set of 6-10-10 bosses: Zhaojing Luo, Kaiping Zheng, Jian Dai, Sheng Wang, Shaofeng Cai, Lei Zhu, Qian Lin, Pingcheng Ruan, Qingchao Cai, Anh Dinh, Zhongle Xie, Piaopiao Feng ++
- Ex-Research Fellows and RAs/Engineers/Students: ....

Healthcare AI Success Factors

NUHS
National University Health System

Foundational factors:

- Funding
- Ethics
- Trusted custodian
- Central governance
- Freedom to innovate and implement



NUS National University of Singapore

ZHEJIANG UNIVERSITY 1897

铭之慧科技

MEDILOT

Thanks!