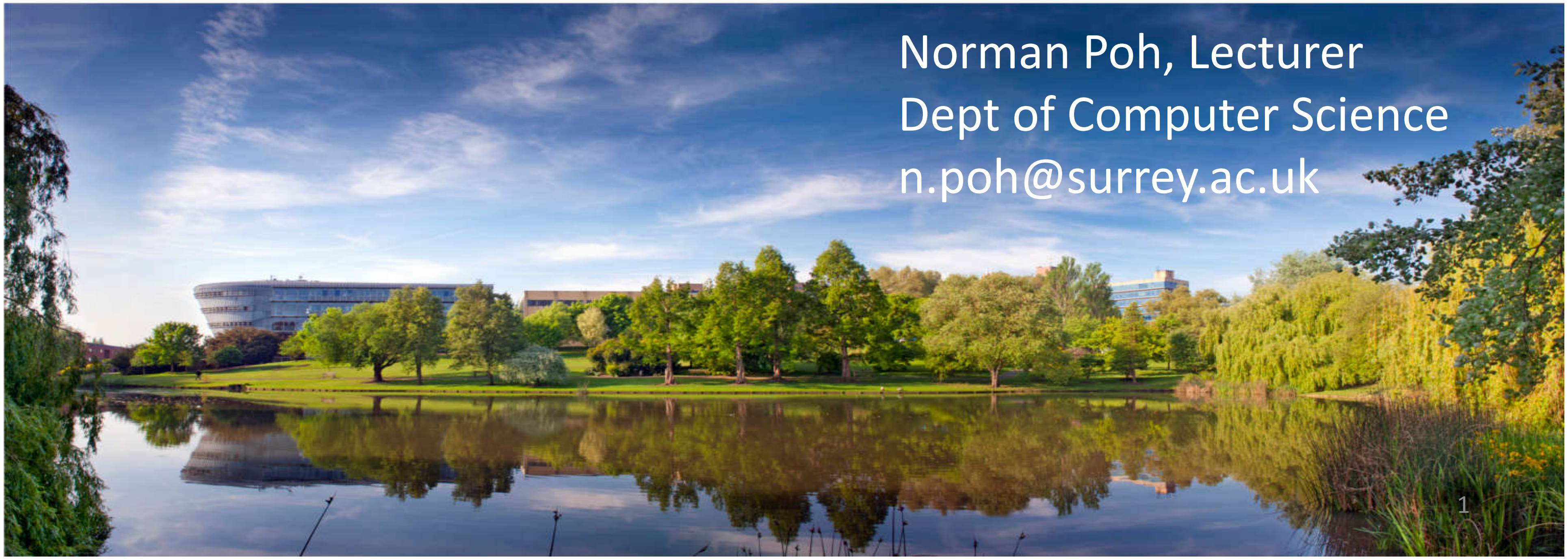
What could we learn from modelling millions of patient records?

A machine learning perspective

UNIVERSITY OF SURREY

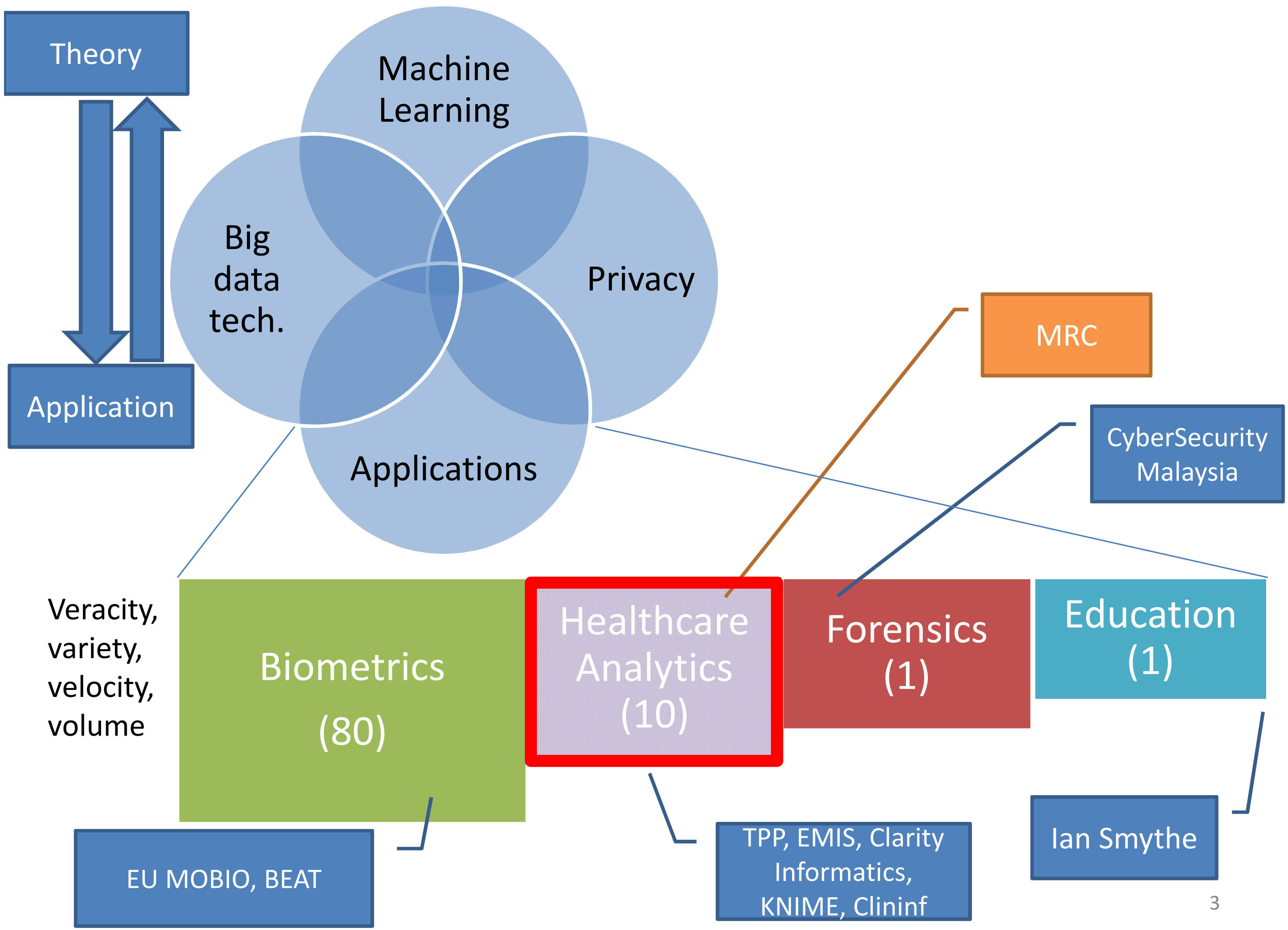


Norman Poh, Lecturer
Dept of Computer Science
n.poh@surrey.ac.uk

40 minutes from London

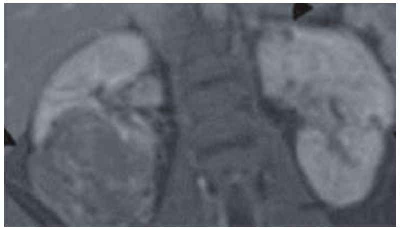
Day/Short-term visits

(n.poh@surrey.ac.uk)

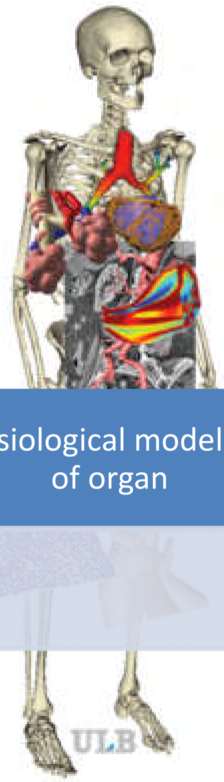


Theory
Application
Machine Learning
Big data tech.
Privacy
Applications
MRC
CyberSecurity Malaysia
Veracity, variety, velocity, volume
Biometrics (80)
Healthcare Analytics (10)
Forensics (1)
Education (1)
EU MOBIO, BEAT
TPP, EMIS, Clarity Informatics, KNIME, Clininf
Ian Smythe

Where machine learning is applicable



Biomedical imaging – computer vision and image processing



Physiological modelling of organ

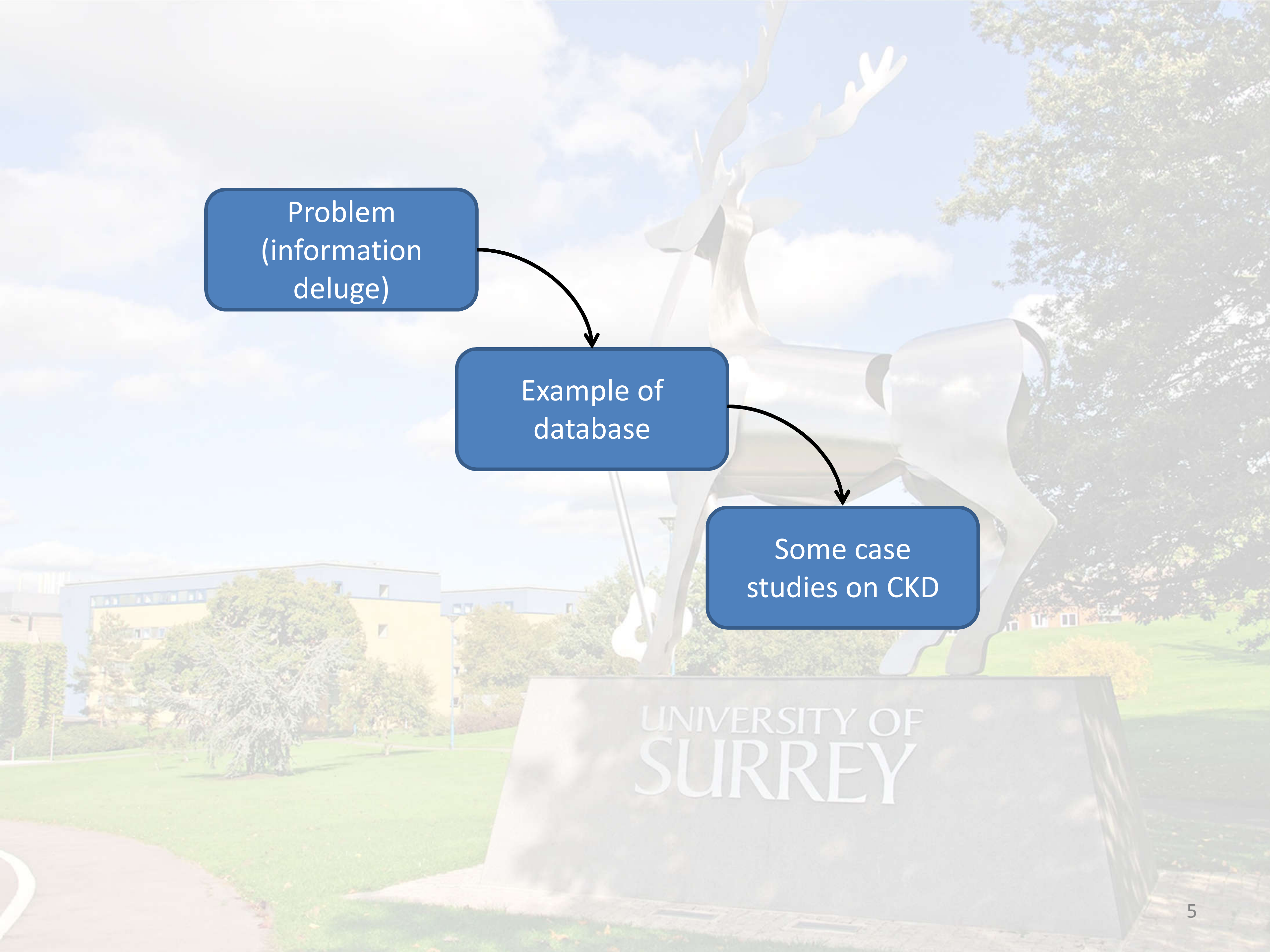
Bioinformatics

Electronic medical records (Epidemiology) – massive data

Problem (information deluge)

Example of database

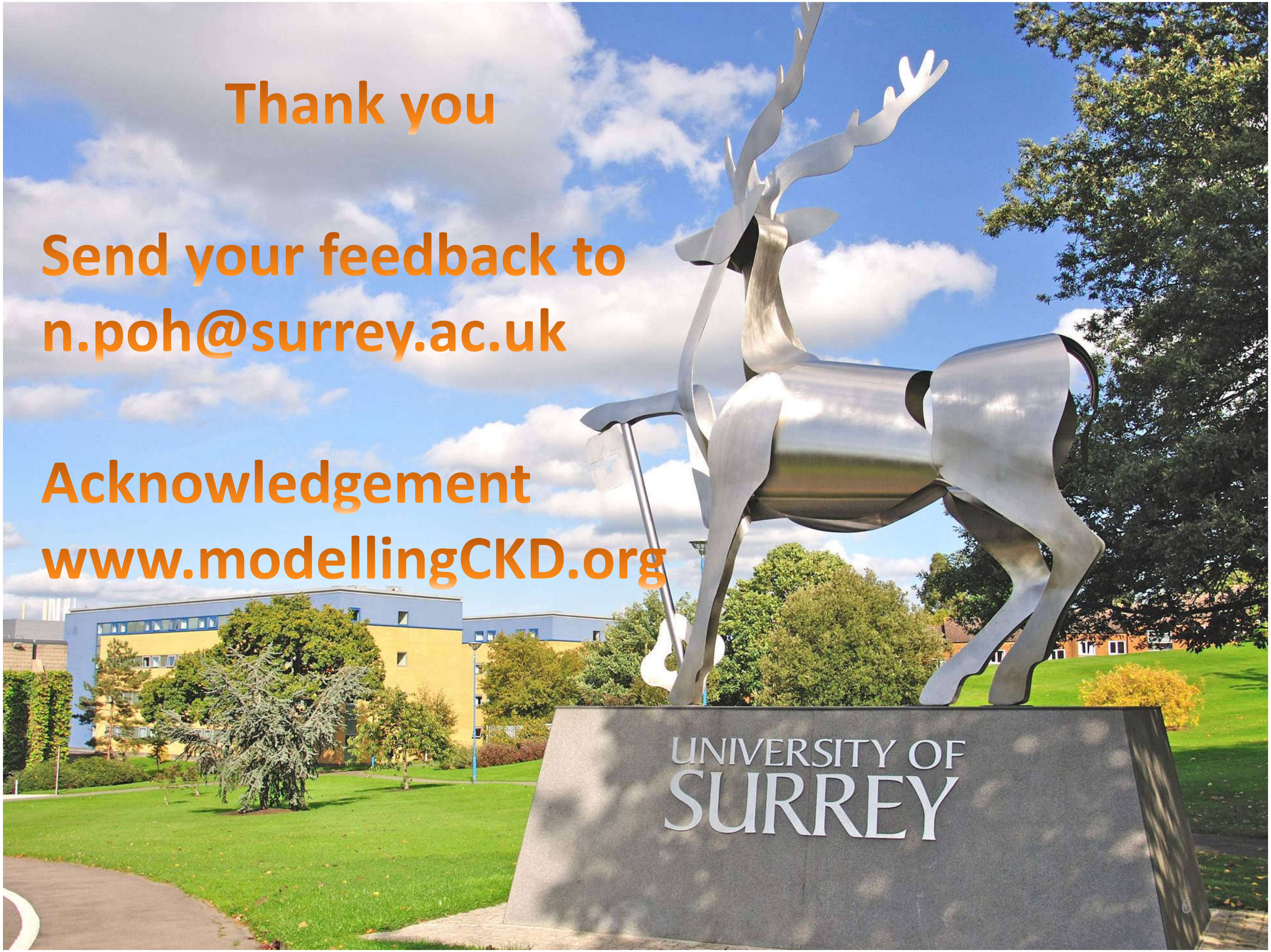
Some case studies on CKD



Thank you

Send your feedback to
n.poh@surrey.ac.uk

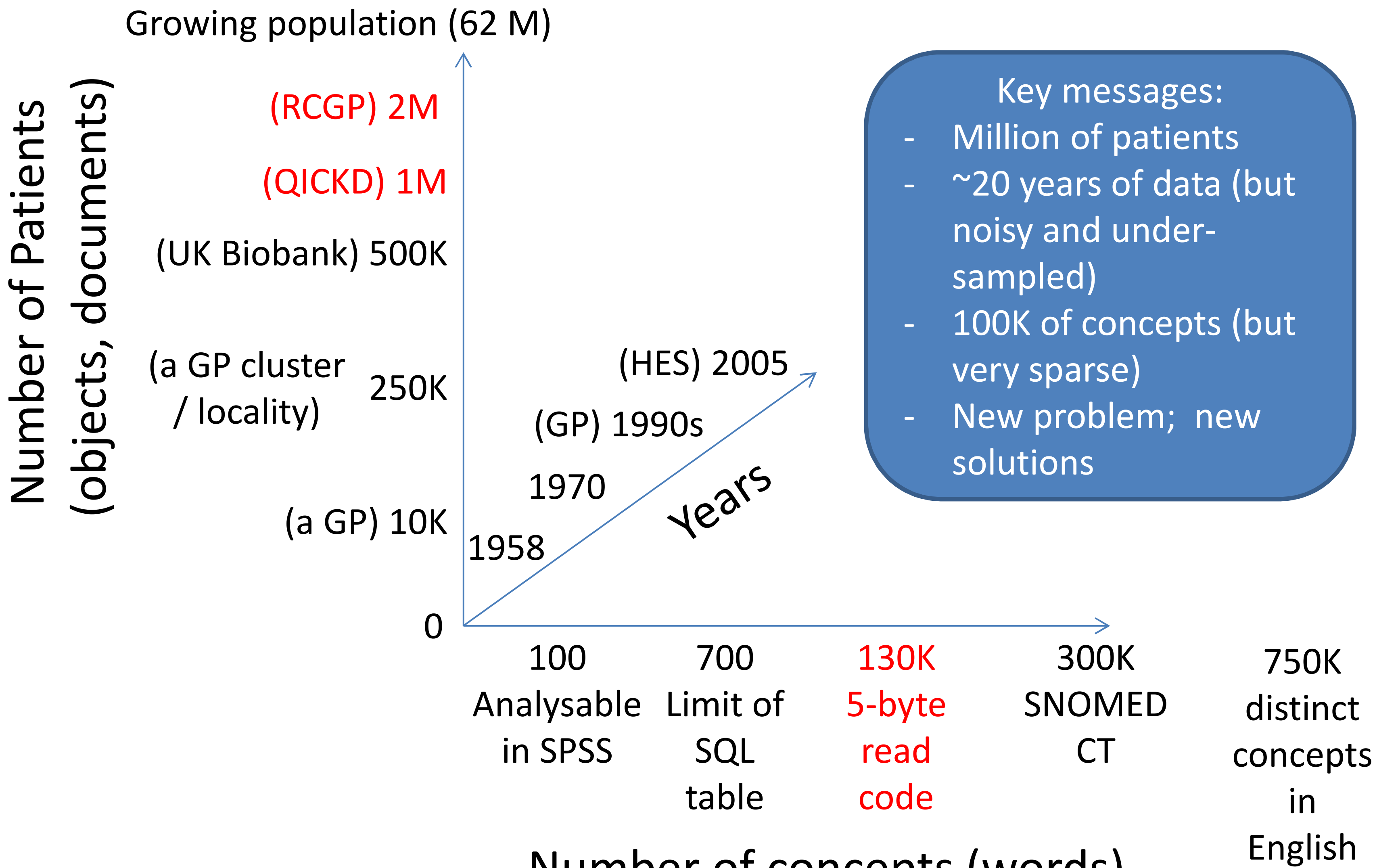
Acknowledgement
www.modellingCKD.org



Part I

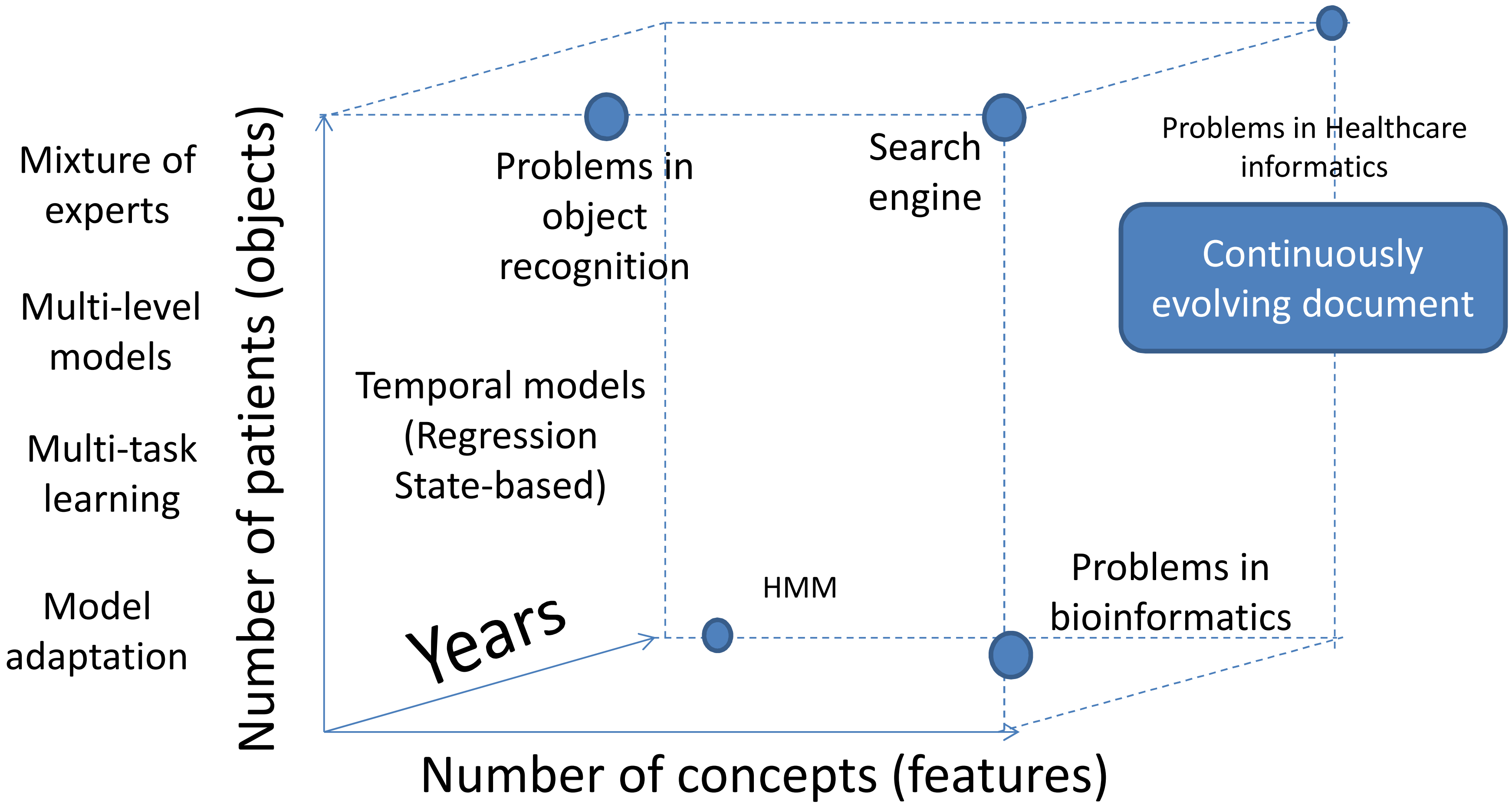
DELUGE OF INFORMATION IN HEALTHCARE

Deluge of information in healthcare



A "data-engineering" problem

… and not (just) a clinical problem



Theme-topic models, Ontology
Sparse representation, Compressive sensing
Feature selection, AdaBoost, Naïve Bayes, SNoW

Royal College of GP (RCGP)

About us | News | Publications | Jobs | eLearning | Revalidation ePortfolio | Trainee ePortfolio | Bookshop | (0 in basket) | Login to My RCGP

Training | Revalidation | RCGP Learning | Making a difference | Clinical | Membership | RCGP local | RCGP nations

RC GP Royal College of General Practitioners

Search RCGP website

Home ›› Clinical ›› Our programmes ›› Research and Surveillance Centre

Research and Surveillance Centre

The RCGP Research and Surveillance Centre (RSC) is part of the RCGP Clinical Innovation and Research Centre (CIRC). It is an internationally renowned source of information, analysis and interpretation, dedicated to research the onset patterns, prevalence and trends over time of morbidity in primary care.

Established in 1957, the RSC is an active research and surveillance unit which collects and monitors data, in particular influenza and other diseases, and monitors vaccine effectiveness.

Research and Surveillance Centre – a cohort profile

The RSC is a representative network, having only small differences with the national population, which have now been quantified and can be assessed for clinical relevance for specific studies. With twice weekly data extractions, the dataset is one of the most up to date in the UK.

The RSC is pleased to announce that an article, describing the network and the usefulness of our practices' data has been published in the BMJ Open. The Centre is keen to hear about new opportunities for collaboration and this free to access paper is a great source of information for anyone unfamiliar with the dataset.

The article describes the first 650,000 patients processed through our new hub established in March 2015. We now have over 1,000,000 patients in the annual report, which is around 1.5% of the English population. We plan to continue to expand the network until we cover around 2% of the national population.

Find courses & events

Enter keyword(s)

Topic | Region

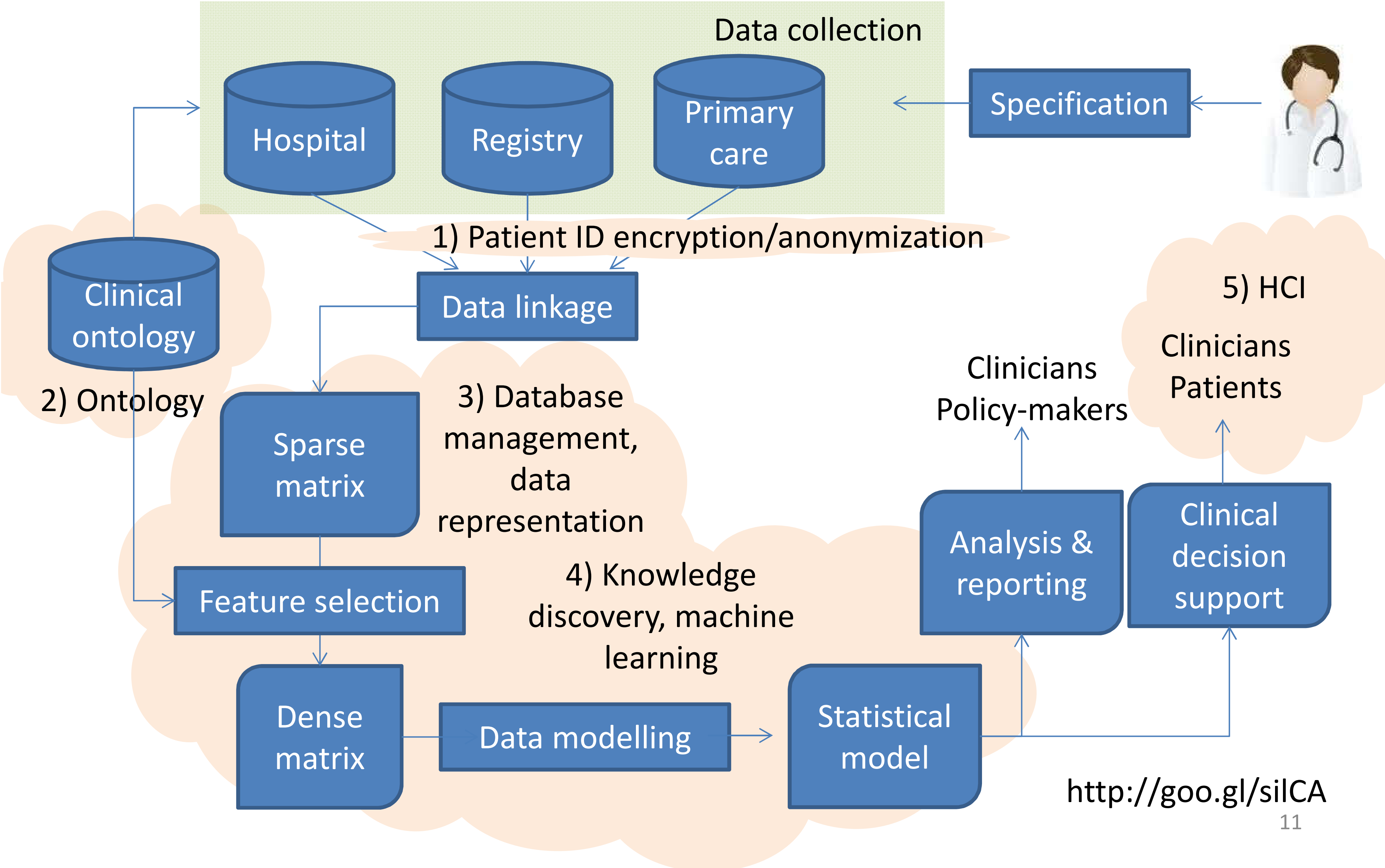
From | To

Date | Date

Advanced search | Find

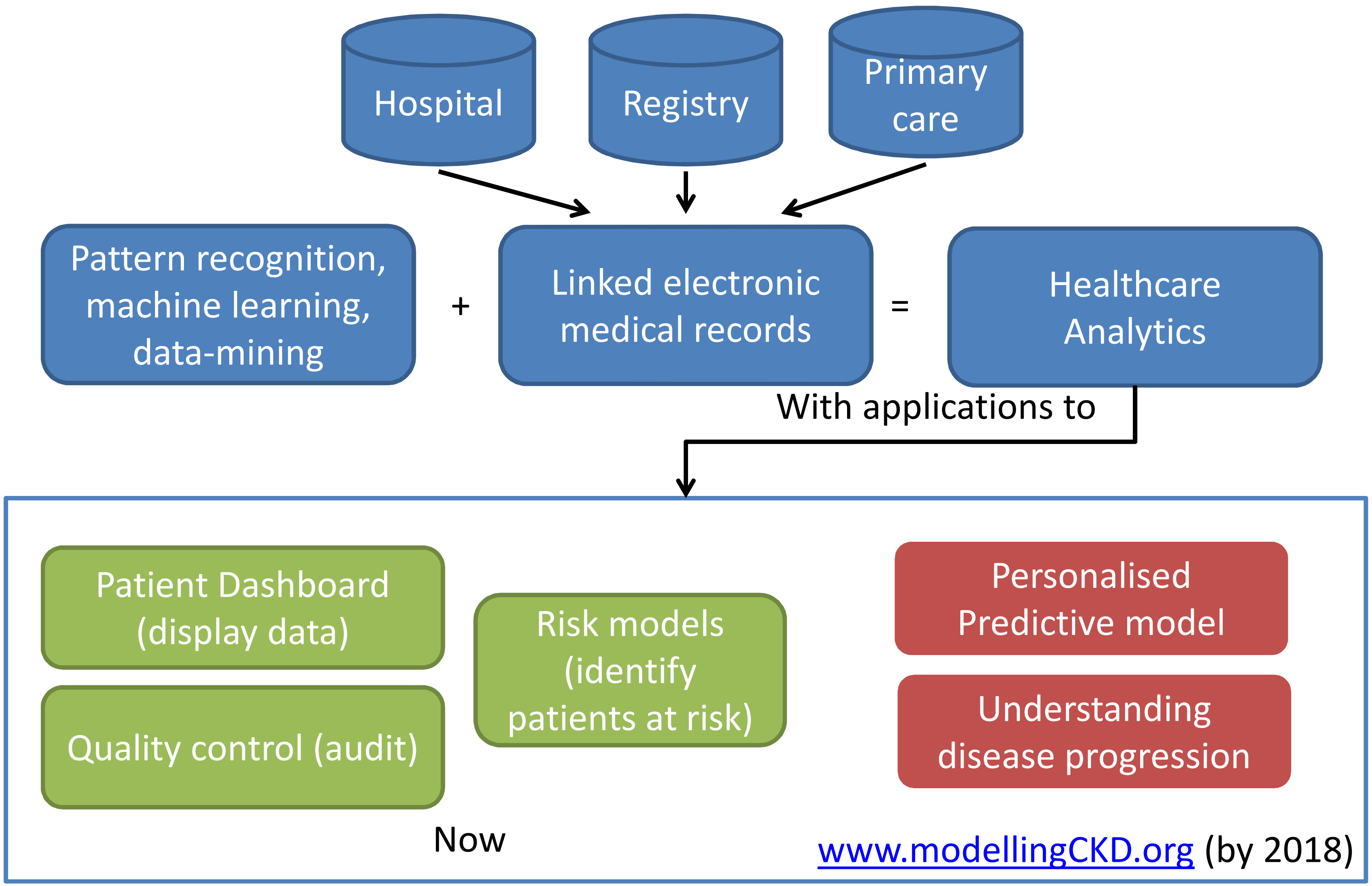
http://www.rcgp.org.uk

Where innovative algorithms are needed?



http://goo.gl/silCA

Our goal


Hospital
Registry
Primary care
Pattern recognition, machine learning, data-mining
+
Linked electronic medical records
=
Healthcare Analytics
With applications to
Patient Dashboard (display data)
Quality control (audit)
Risk models (identify patients at risk)
Now
Personalised Predictive model
Understanding disease progression
www.modellingCKD.org (by 2018)

What does healthcare analytics promise?

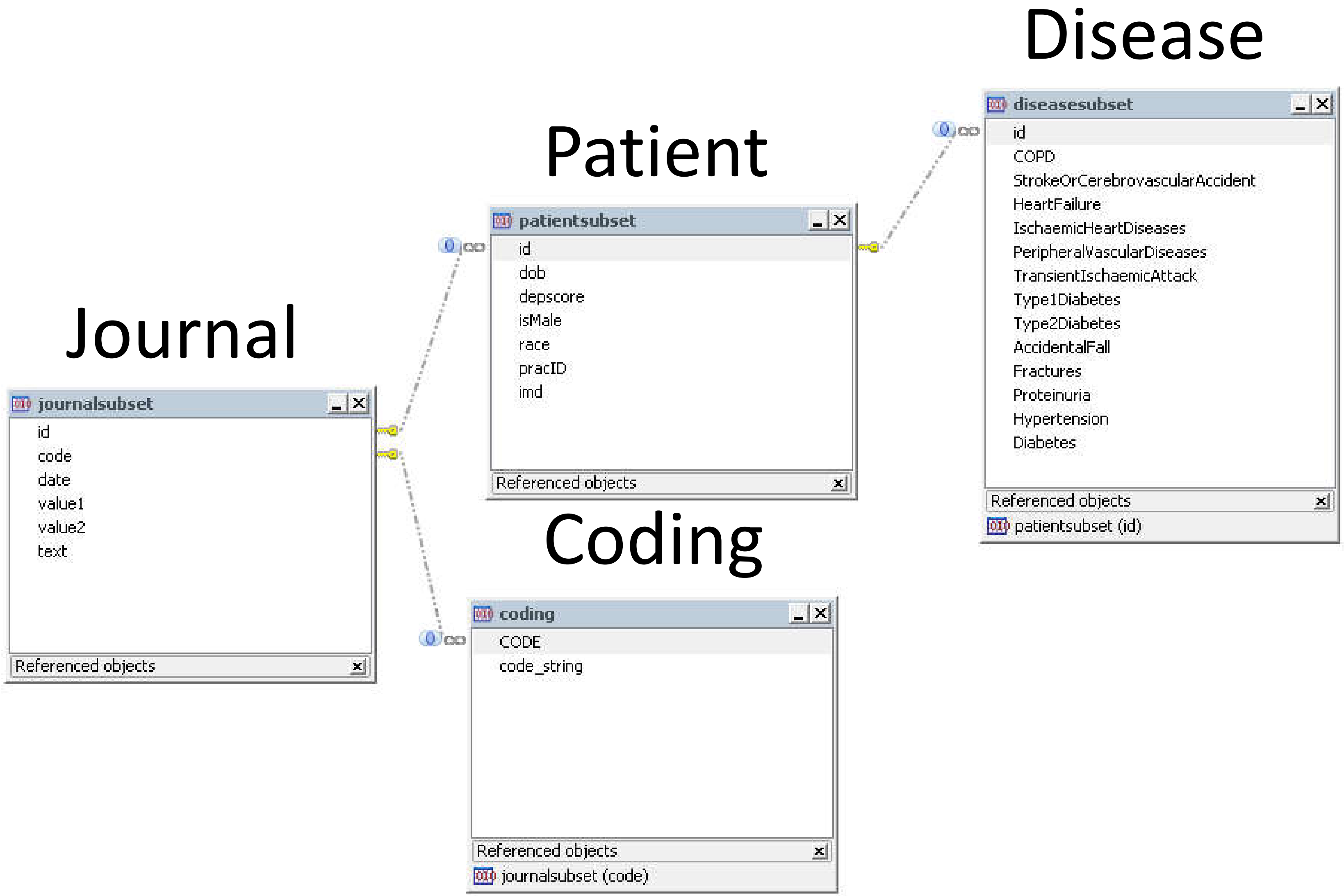
Readmission: Reduce unplanned admission to hospital	Triage: Estimate risk of complications	High cost patients: 5% patients – 50% cost
Adverse events: renal failure, infection, adverse drugs	Decompensation: Real time monitoring of vitality sign	Diseases affecting multiple organ systems

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, *33*(7), 1123-1131.

Topic

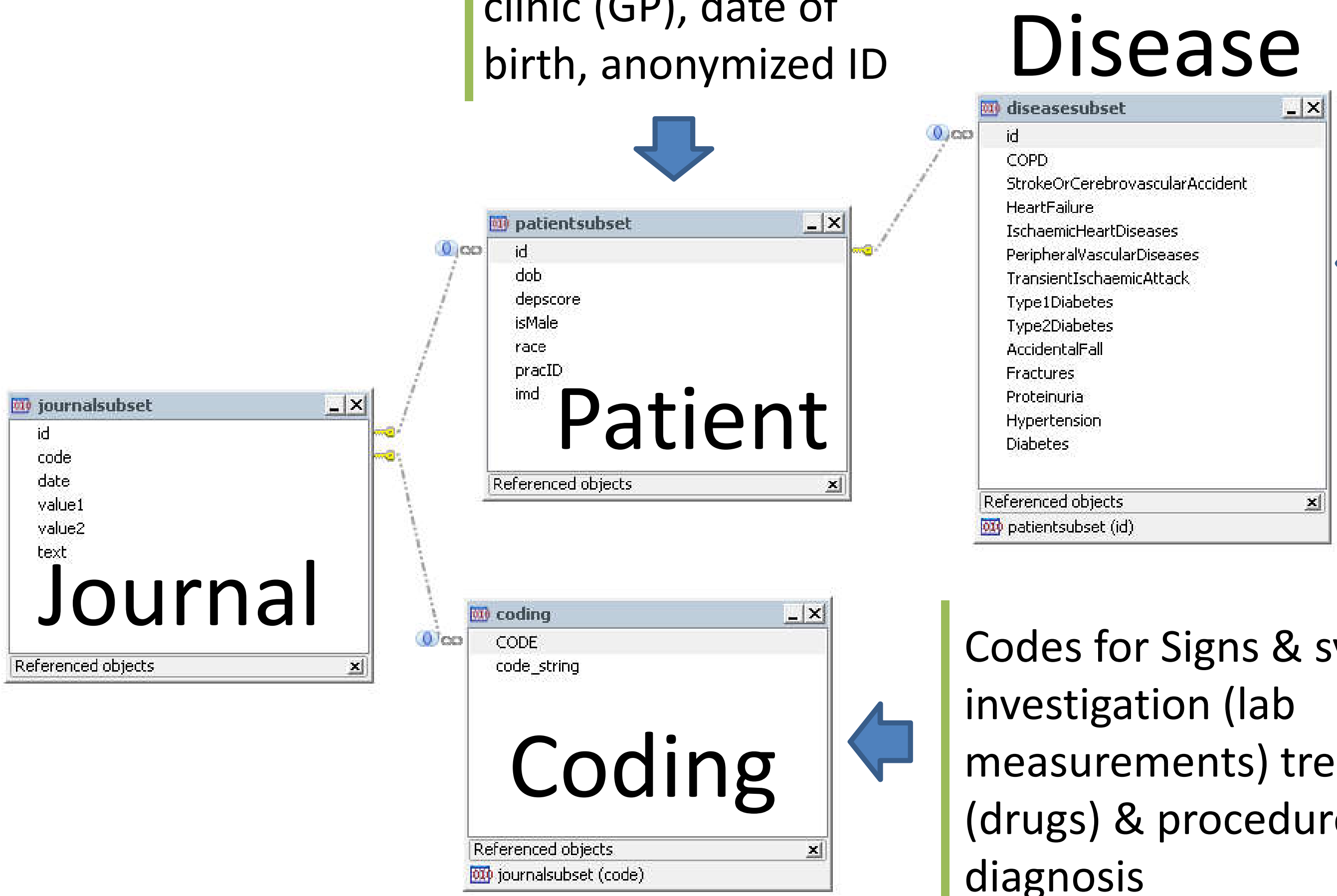
EXAMPLE OF DATABASE: QUALITY IMPROVEMENT CKD (QICKD)

Structure of a database



What information is available?

Ethnicity, age, gender, clinic (GP), date of birth, anonymized ID



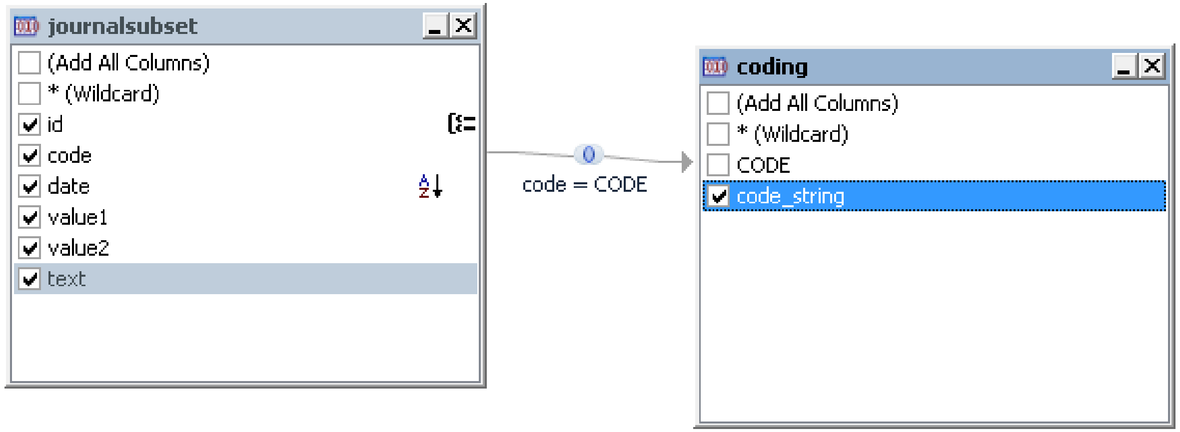
Derived diagnosis based on data in journal

Codes for Signs & symptoms, investigation (lab measurements) treatment (drugs) & procedures, diagnosis

Patient table

id	dob	depscore	isMale	race	pracID	imd
26015	1967	15.32	0	1	56	5
26016	1937	23.66	1	6	56	7
26017	1976	23.66	1	7	56	7
26018	1947	23.66	0	7	56	7
26019	1920	5.59	1	7	56	1
26020	1957	4.00	1	7	56	1
26021	1932	8.08	0	7	56	2
26022	1931	8.68	0	7	56	3
26023	1975	23.66	1	7	56	7
26024	1950	23.66	1	7	56	7
26025	1954	23.66	0	7	56	7
26026	1981	23.66	0	7	56	7
26027	1972	10.27	0	7	56	3
26028	1961	5.85	1	7	56	2
26029	1964	6.47	0	1	56	2
26030	1925	23.66	0	1	56	7
26031	1962	6.47	1	7	56	2

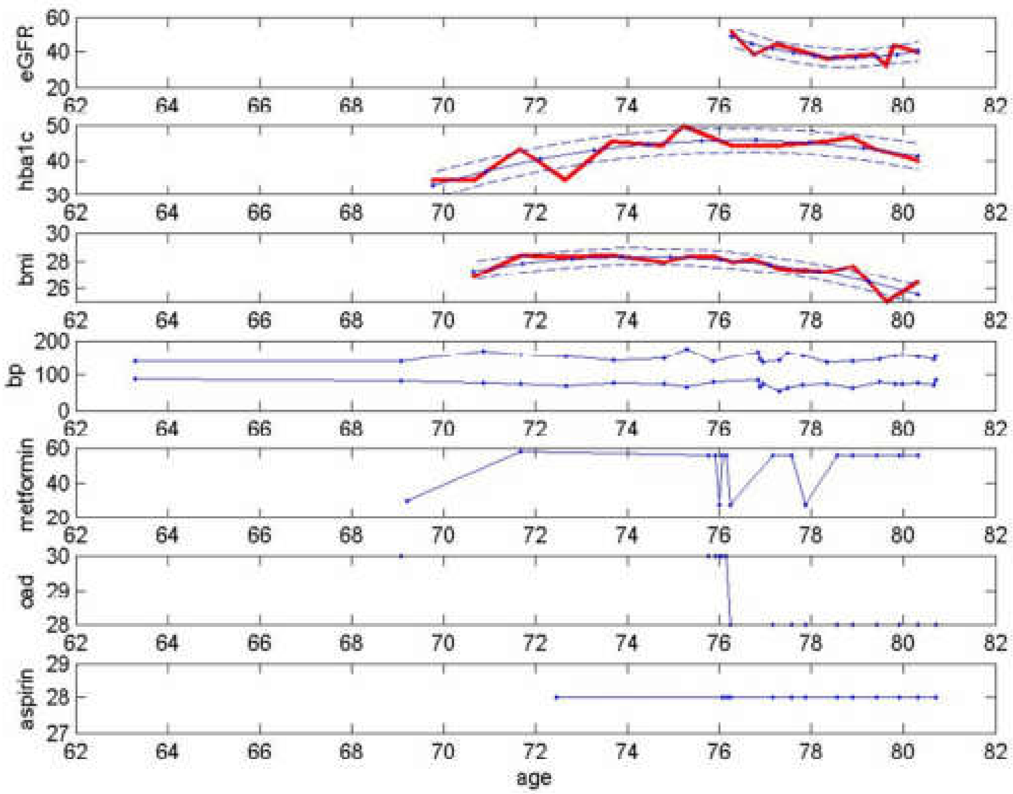
Journal and Coding table




```sql
SELECT journalsubset.id,
       journalsubset.code,
       coding.code_string,
       journalsubset.`date`,
       journalsubset.value1,
       journalsubset.value2,
       journalsubset.`text`
  FROM qickd2.journalsubset journalsubset
       INNER JOIN qickd2.coding coding ON
(journalsubset.code = coding.CODE)
GROUP BY journalsubset.id
ORDER BY journalsubset.`date` ASC
```

View file

Key difficulties



- How to represent irregularly sampled time-series with real and binary values?
- How to model the trajectories?
- Millions of patients; each of which has few observations (multi-task learning? Hierarchical model?)
- Can we develop models that can explain its reasoning, yet flexible enough to fit the data and generalize to unseen data?

Topic

CASE STUDIES BASED ON CHRONIC KIDNEY DISEASE

Some basic medical terminology

English	French
Chronic Kidney Disease (**CKD**)	l'insuffisance rénale chronique
Acute Kidney Injury (**AKI**)	l'insuffisance rénale aiguë
estimated Glomerular Filtration Rate (**eGFR**)	Débit de filtration glomérulaire (DFG)

Classiquement, on distingue l'insuffisance rénale aiguë de l'insuffisance rénale chronique.

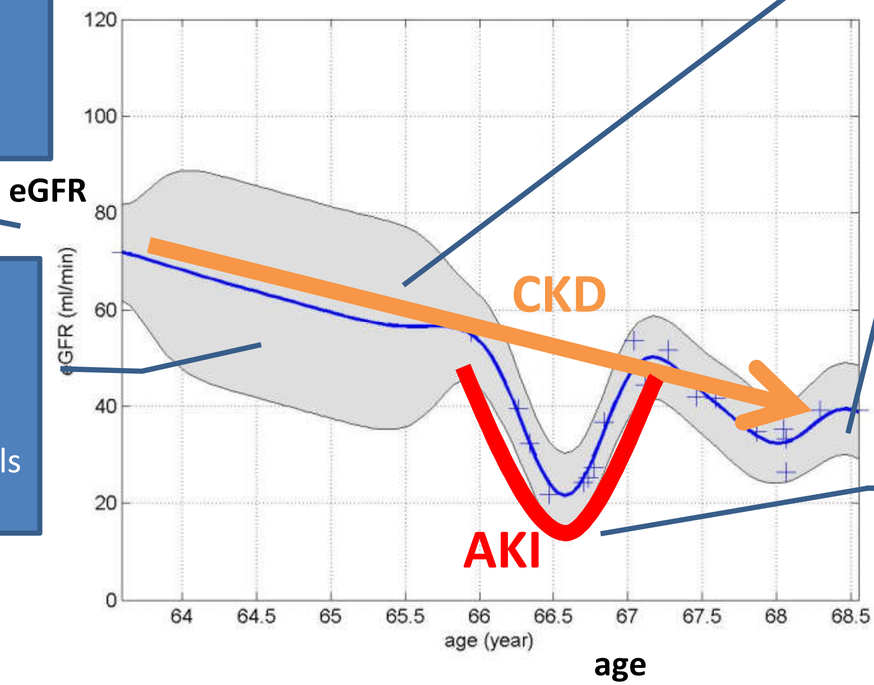
Globalement, une insuffisance rénale se caractérise par une diminution de la fonction, et du nombre des néphrons (unités de base constituant le rein et servant à débarrasser le sang des toxines qu'il contient, en élaborant l'urine primitive).

L'insuffisance rénale aiguë, contrairement à l'insuffisance rénale chronique, est généralement réversible et guérit le plus souvent. Elle consiste en une privation brutale de l'organisme de sa fonction rénale (fonctionnement des reins).

Challenges in modelling and predicting CKD

Different formulas (e.g., eGFR-MDRD, CKD-EPI , Cystatin C) cause structural bias when modelling the trend

Commonly used algorithms (e.g., logistic & Cox regression) are not adequate; new models not yet tested



CKD is a long-term but asymptomatic condition

Comorbid with diabetes, hypertension, CVD

Progressive risk, leading to costly end stage renal disease (dialysis & kidney transplant); so needs timely and *accurate* referral from primary to secondary care

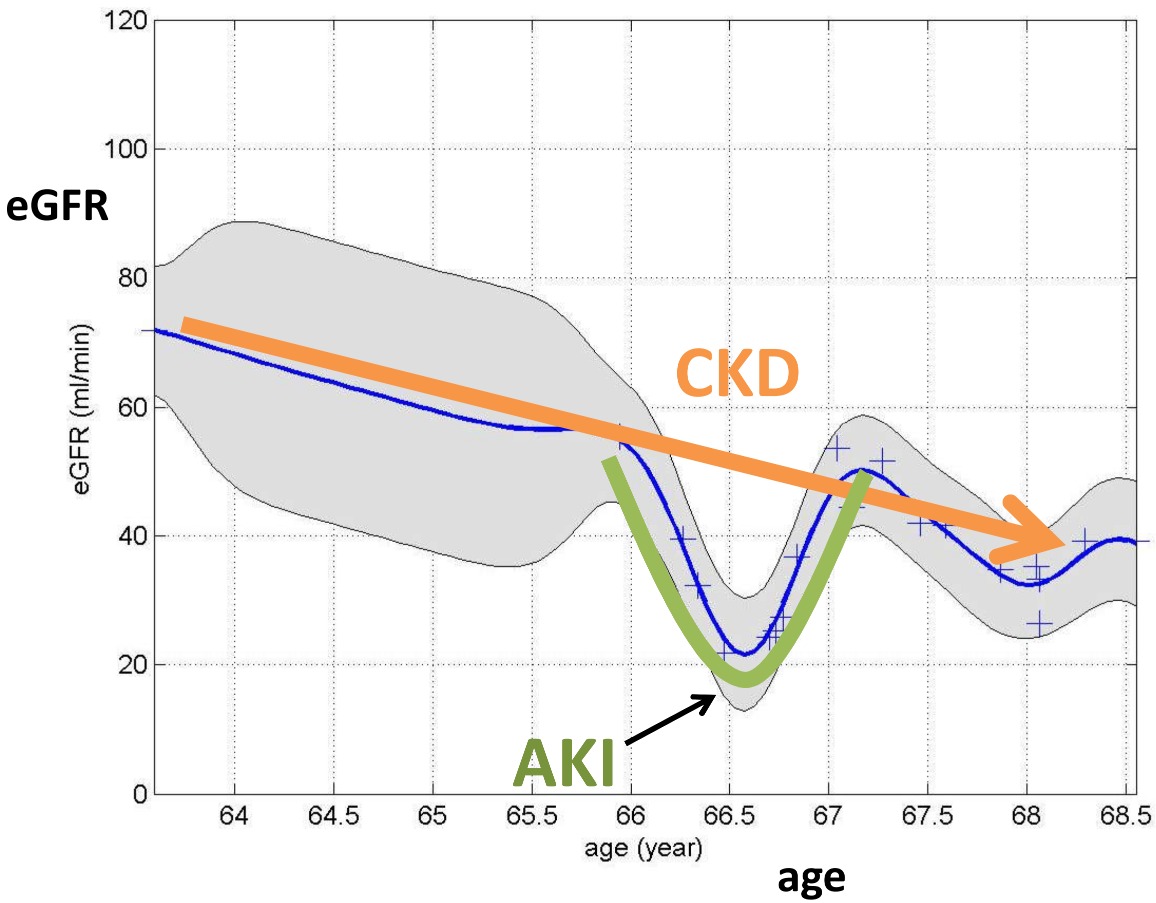
Acute Kidney Injury (AKI) can temporarily disrupt the trend

$p(g|a)$ = Gaussian Process Regression

CKD and AKI

Regression

$p(g|a)$ = Gaussian Process Regression

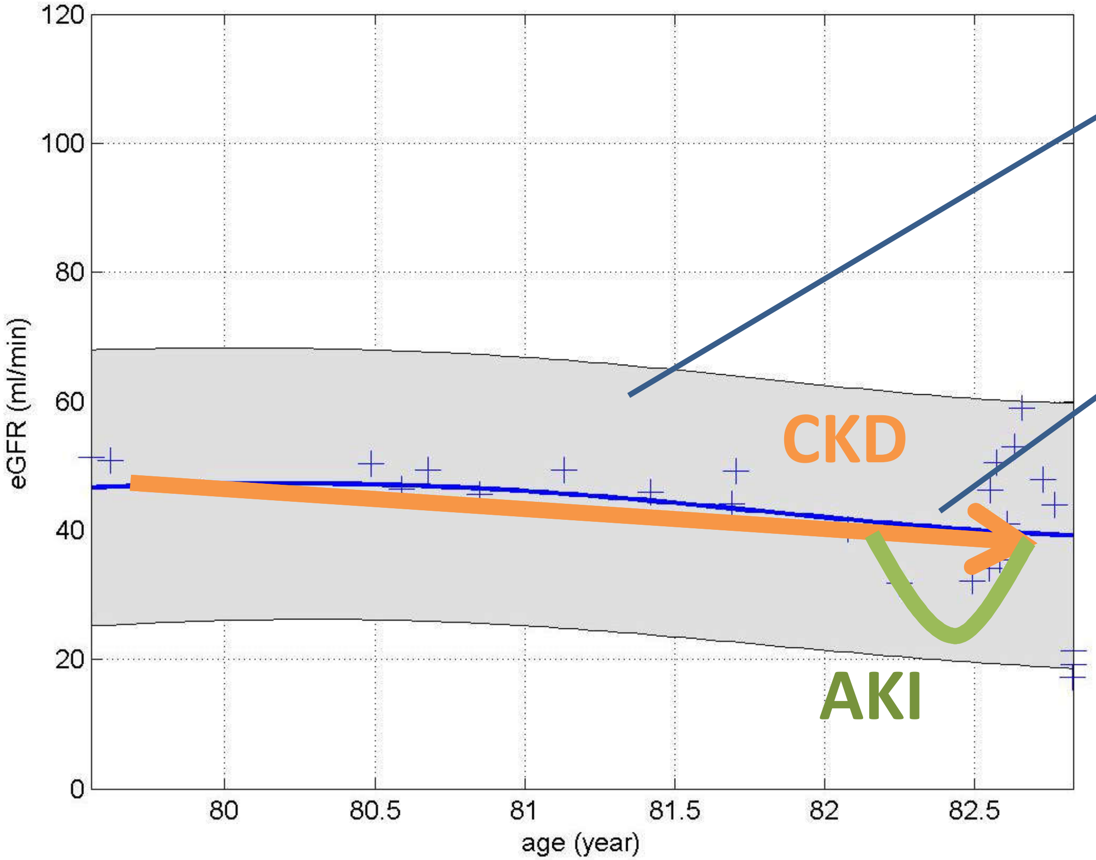


Classical regression does not work

Non-parametric regression works some times but not a guarantee

AKI not always modelled

$p(g|a)$ = Gaussian Process Regression



Learning to tune GPR hyper-parameters

Another solution: Mixture of experts, mixture of (2) regressions

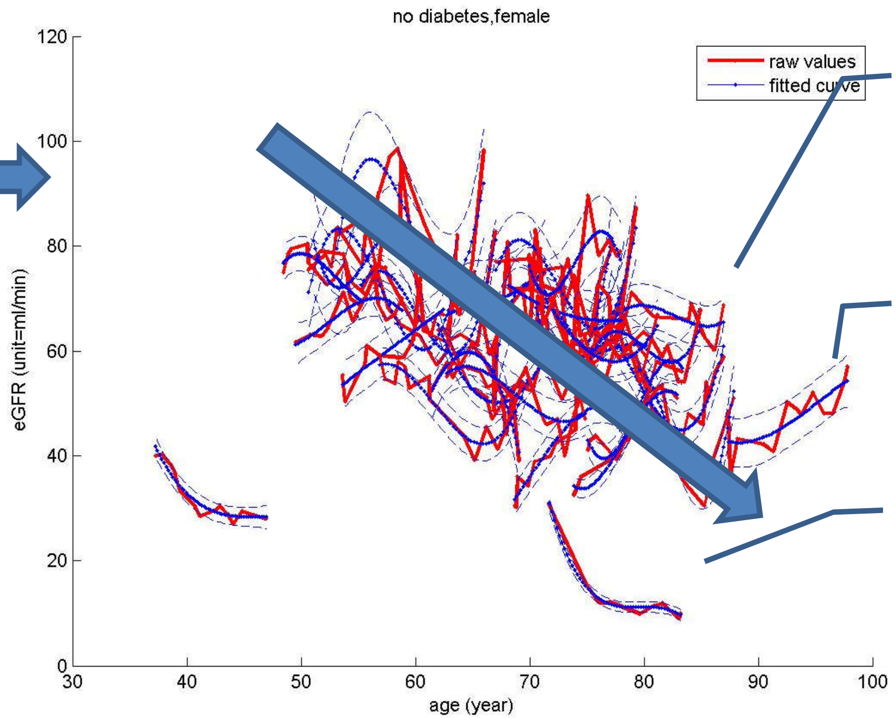
Methodology

Data + Machine Learning
+ Expert feedback
= Clinically useful models

Using routinely collected data: QICKD, Qresearch, ResearchOne, East Kent, ASSIST-CKD, RCGP data sets

Objectives:

1. Predict eGFR
2. Stratify patients
3. Predict AKI



Obj 1: Model & predict individual eGFR

Understand factors contributing to renal progression & regression

(renal regeneration not normally expected)

Hierarchical regression model

Obj 2: Model AKI risk

AKI risk modelling

Mixture of regression

Obj 3. Stratify patients

t-stochastic Neighbourhood Embedding (t-SNE), deep neural network

Hierarchcical clustering, constrained clustering

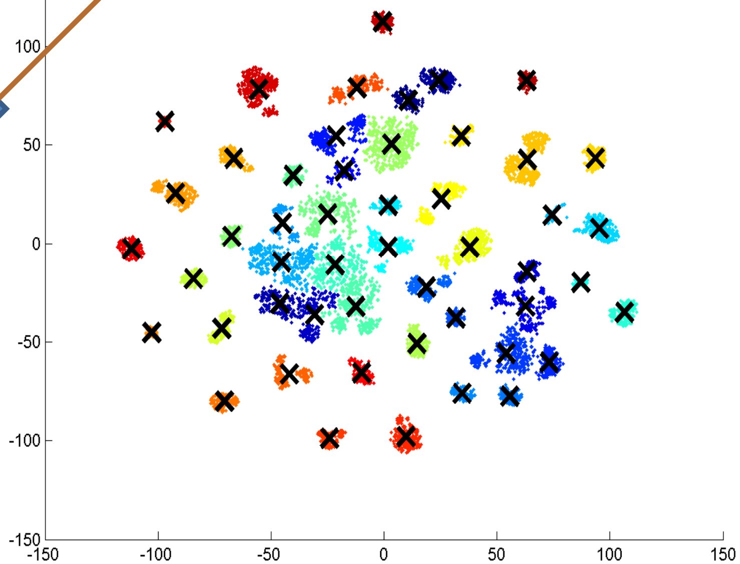
Using routinely collected data: QICKD, Qresearch, ResearchOne, East Kent, ASSIST-CKD, RCGP data sets

20,000 clinical codes

(1) Dimensionality reduction

(2) Clustering

10,000 patients with chronic kidney disease



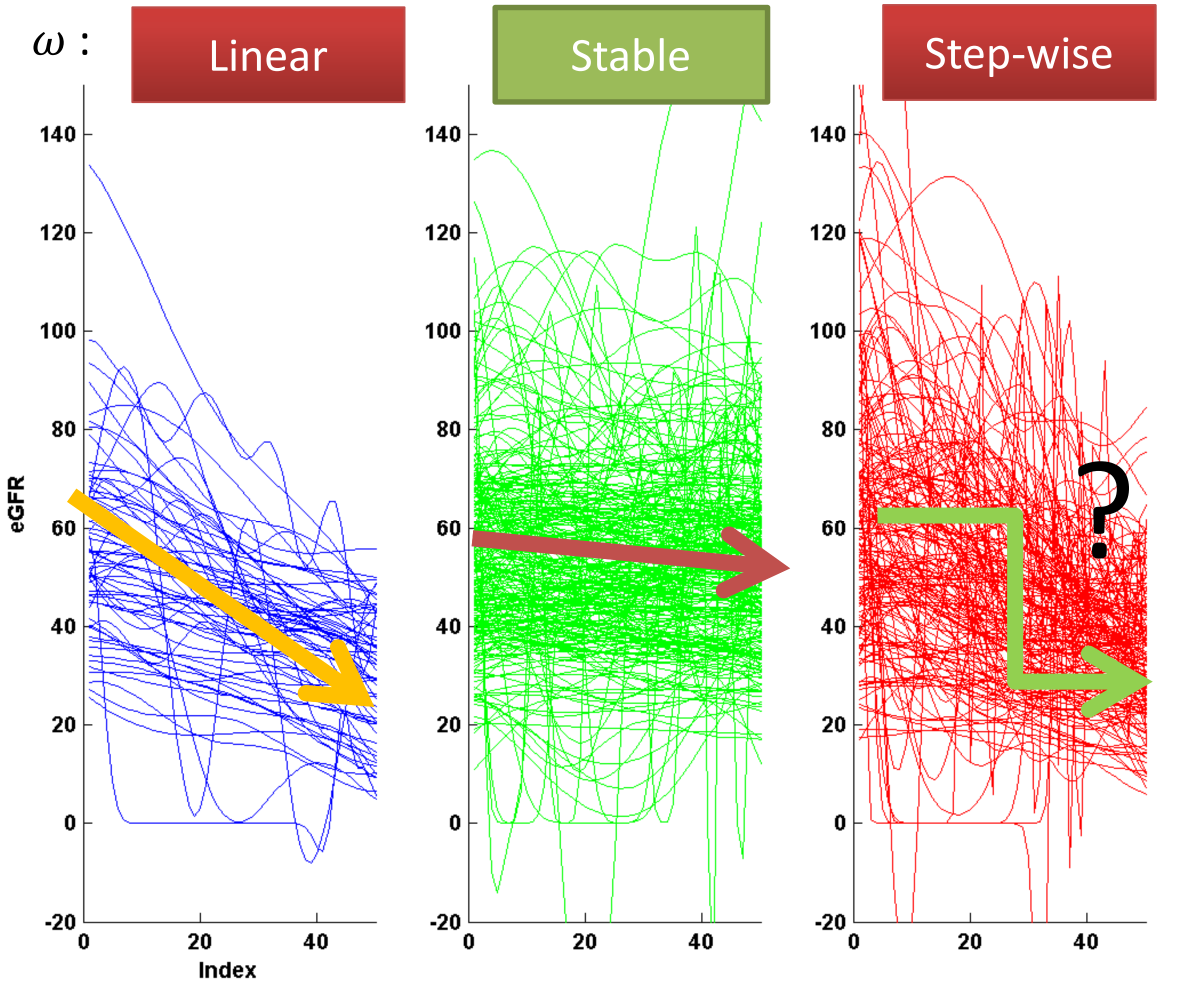
Contributions:

1. Proactive CKD management by eGFR prediction
2. Ability to profile CKD care pathways
3. Novel AKI risk model for community care

MRC (Medical Research Council) project grant: MR/M023281/1| 2015-2018 | www.modellingCKD.org

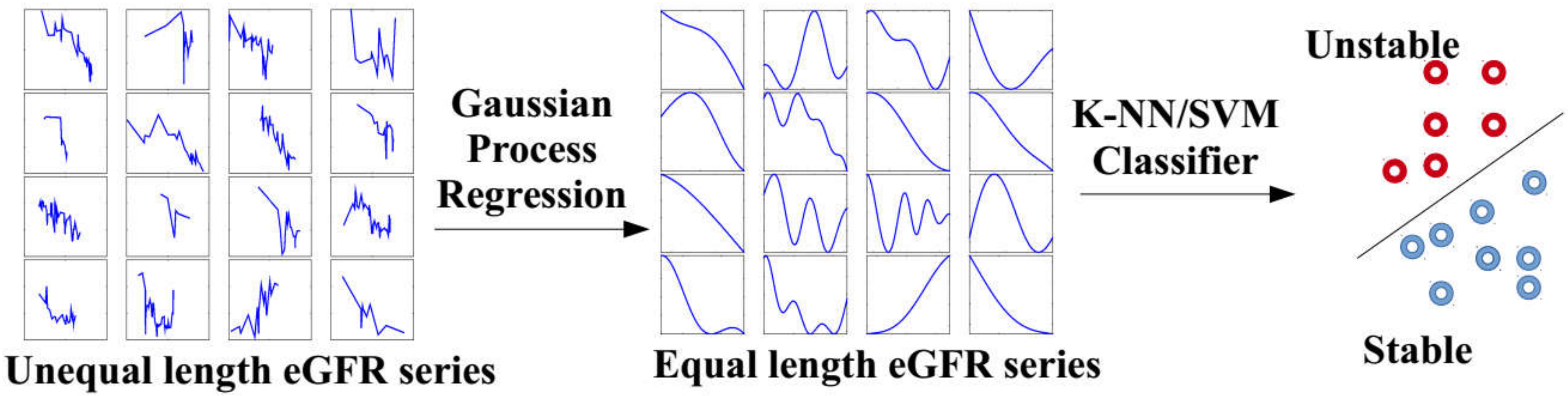
Automatic classification of eGFR trends

$$P(\omega|\{g\}_1^T)$$



Classification of irregular time-series

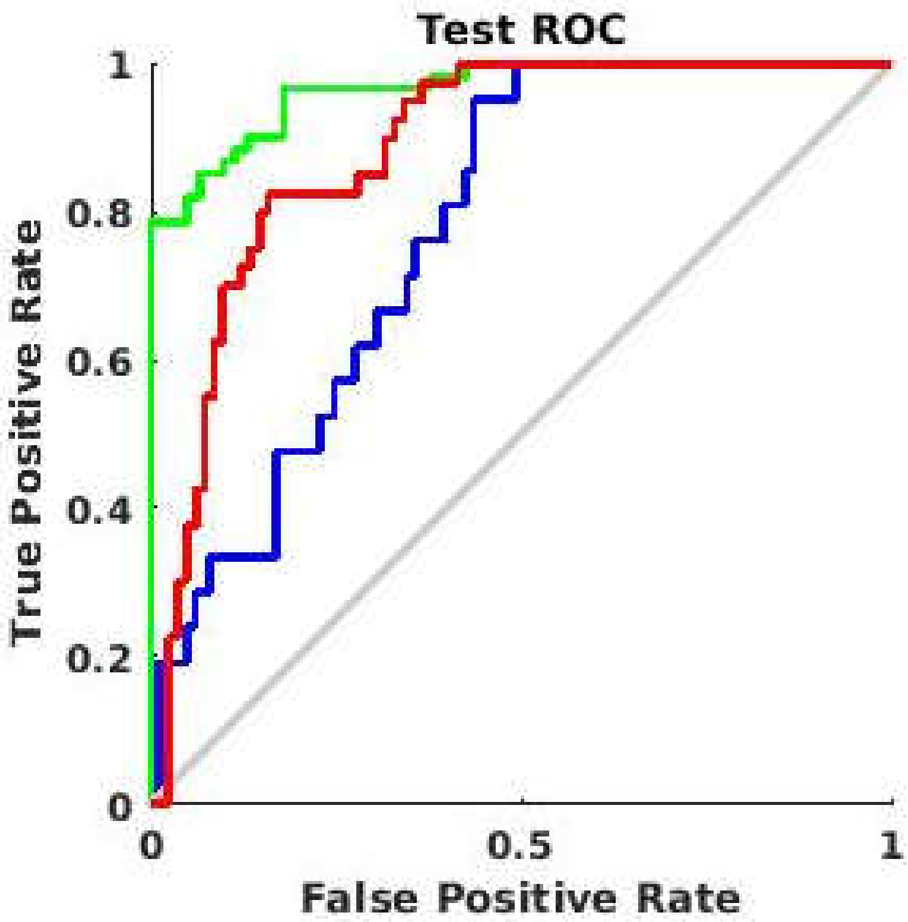
Approach



Unequal length eGFR series
Gaussian Process Regression
Equal length eGFR series
K-NN/SVM Classifier
Unstable
Stable

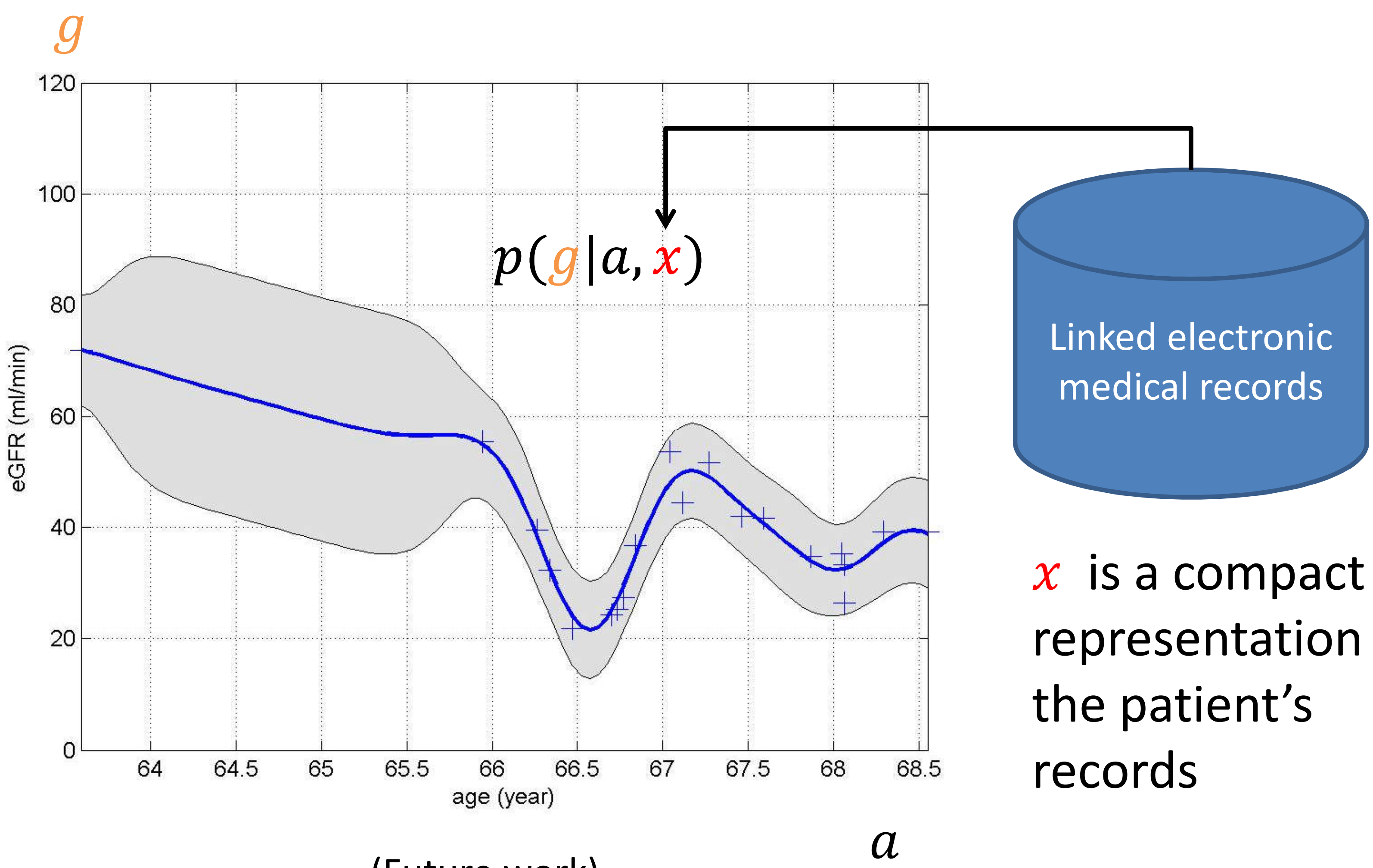
Some preliminary results

Linear stable step-wise



http://arxiv.org/pdf/1605.05142.pdf

Prediction with additional information



(Future work)

x is a compact representation of the patient's records