
UKRI
Science and Technology Facilities Council
Hartree Centre
Welcome

Science and Technology Facilities Council

Hartree Centre

Big Data Week 9

Real World Applications and Examples

Dr Simon Goodchild
Data Science group leader, Hartree Centre

Dr Dominic Richards
Artificial Intelligence group leader, Hartree Centre

Science and Technology Facilities Council

Hartree Centre

INTELLECTUAL PROPERTY RIGHTS NOTICE:

The User may only download, make and retain a copy of the materials for their use for non-commercial and research purposes. If you intend to use the materials for secondary teaching purposes it is necessary first to obtain permission.

The User may not commercially use the material, unless a prior written consent by the Licensor has been granted to do so. In any case, the user cannot remove, obscure or modify copyright notices, text acknowledging or other means of identification or disclaimers as they appear.

For further details, please email us: hartreetraining@stfc.a.c.uk

Lecture overview

- Hartree Centre data science case studies
  - Roberts Vain Wilshaw (big data preparation and automated processing)
  - Liverpool CCG (machine learning with logistic regression and random forests)
  - Southwest Water (time series analysis)
- Applications of Artificial Intelligence

Science and Technology Facilities Council

Hartree Centre

Hartree Centre data science case studies

Science and Technology Facilities Council

Hartree Centre

Big Data processing

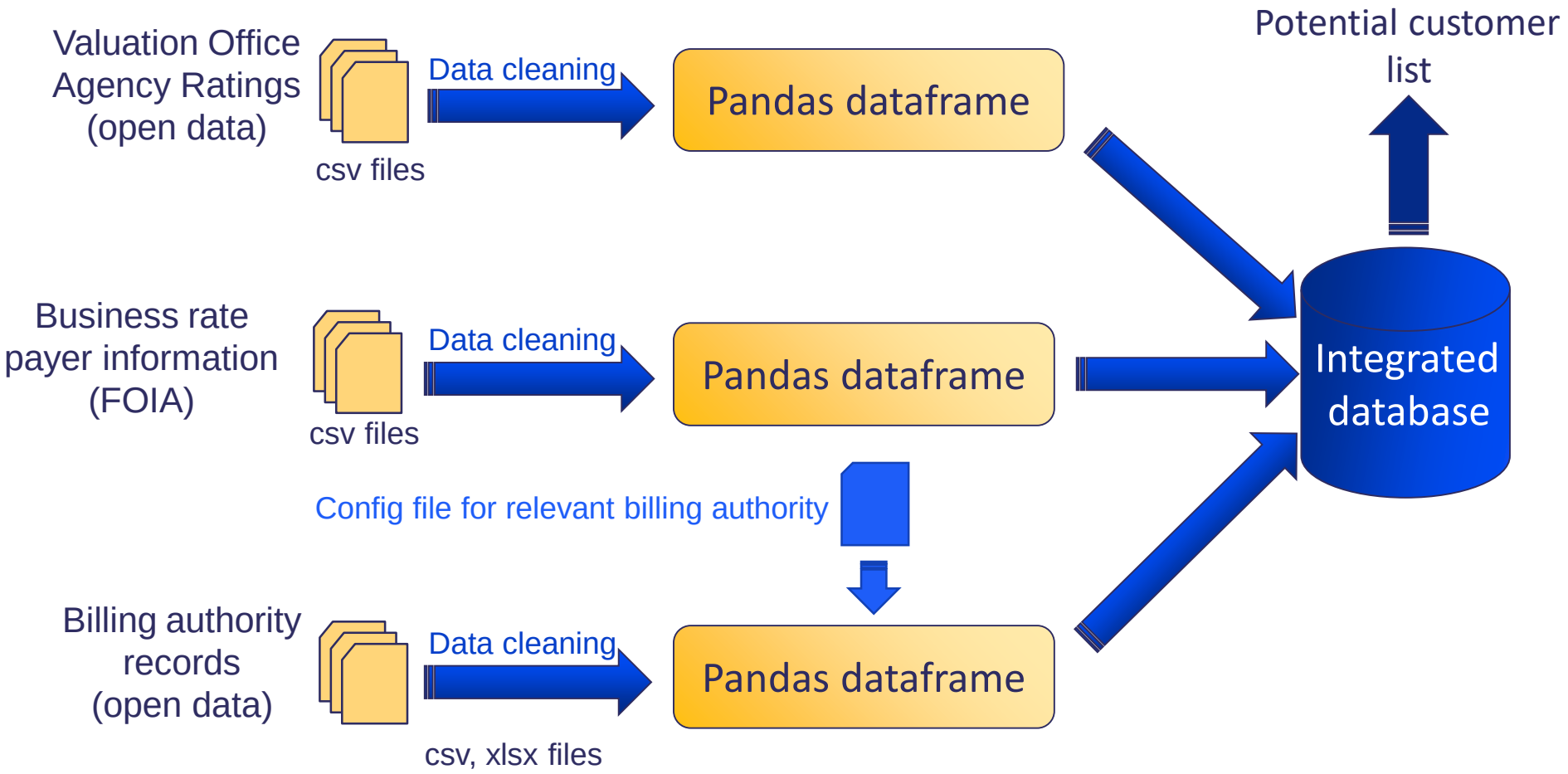
‘Data preparation accounts for about 80% of the work of data scientists’

Forbes magazine (2016)
https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

‘Respondents reported that on average 45% of their time is spent getting data ready (loading and cleansing) before they can use it to develop models and visualizations.’

Anaconda, *The State of Data Science,* (2020)
https://www.anaconda.com/state-of-data-science-2020

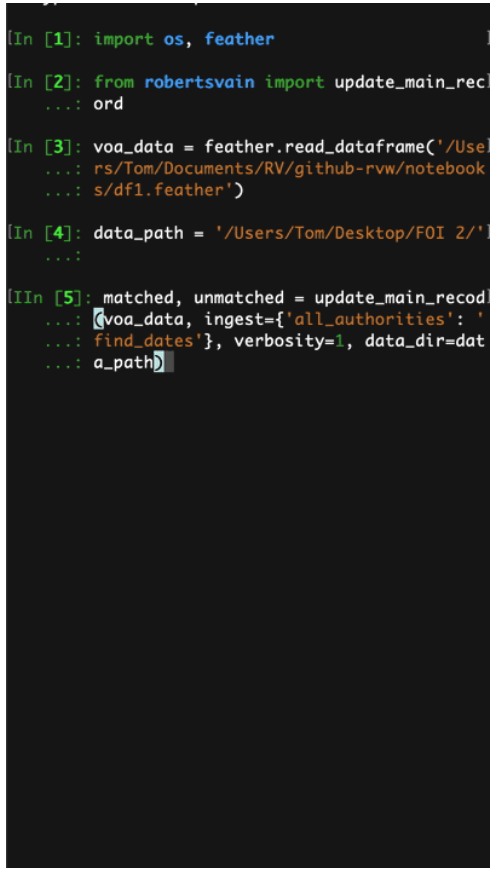
Big Data Processing



Science and Technology Facilities Council
Hartree Centre

Roberts VainWilshaw CHARTERED SURVEYORS

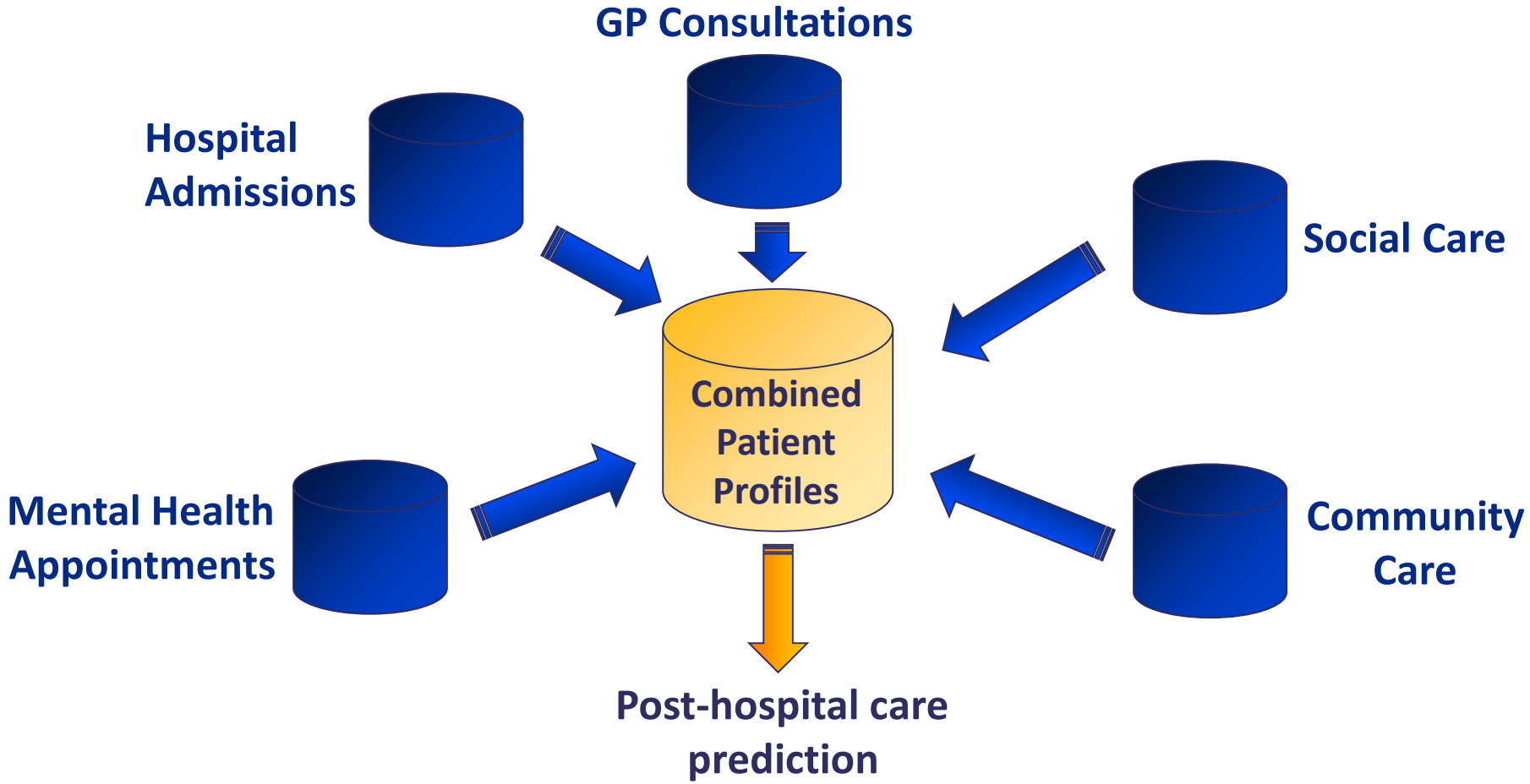
Big Data Processing



Science and Technology Facilities Council

Hartree Centre

Predictive modelling



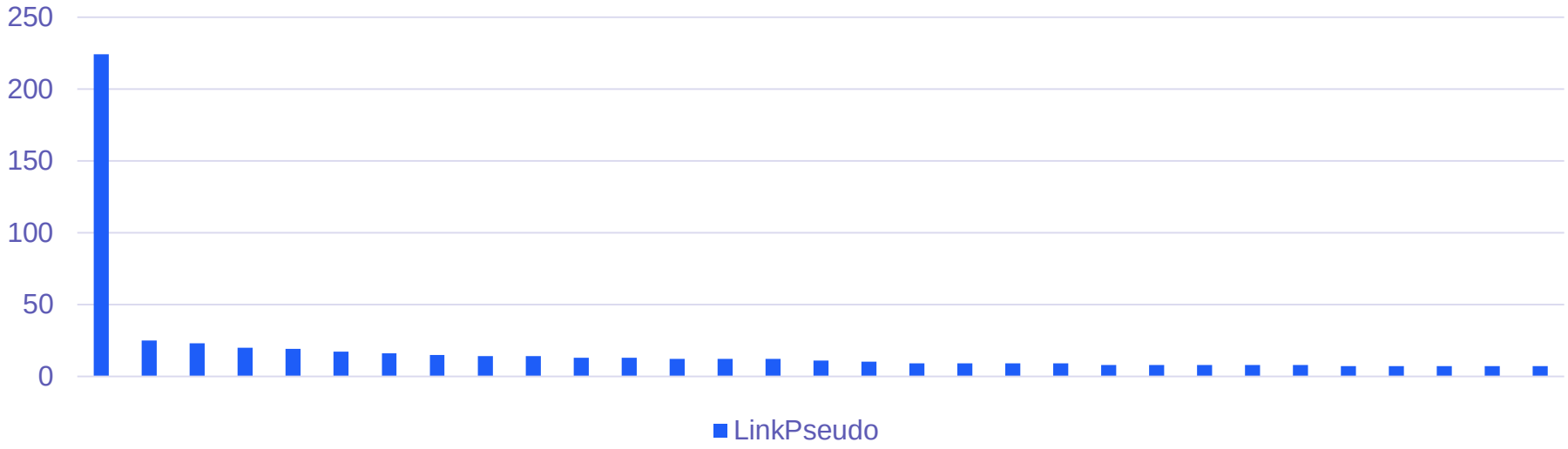
Predictive modelling

Data processing

Anonymised NHS patient records:

NHS number 1234567890 → Hashing → LinkPseudo 35cc4c9488db26c

25% of records due to one LinkPseudo



NHS number "" → Hashing → LinkPseudo 571e94cd349e1f0

Predictive modelling

Data processing

Map

Data line → (key, value)

e.g.
35cc4c9488db264c,L14,2020-04-21,LTC_Asthma,… → (35cc4c9488db264c, LTC_Asthma)

35cc4c9488db264c,L14,2020-03-17,LTC_Depression,… → (35cc4c9488db264c, LTC_Depression)

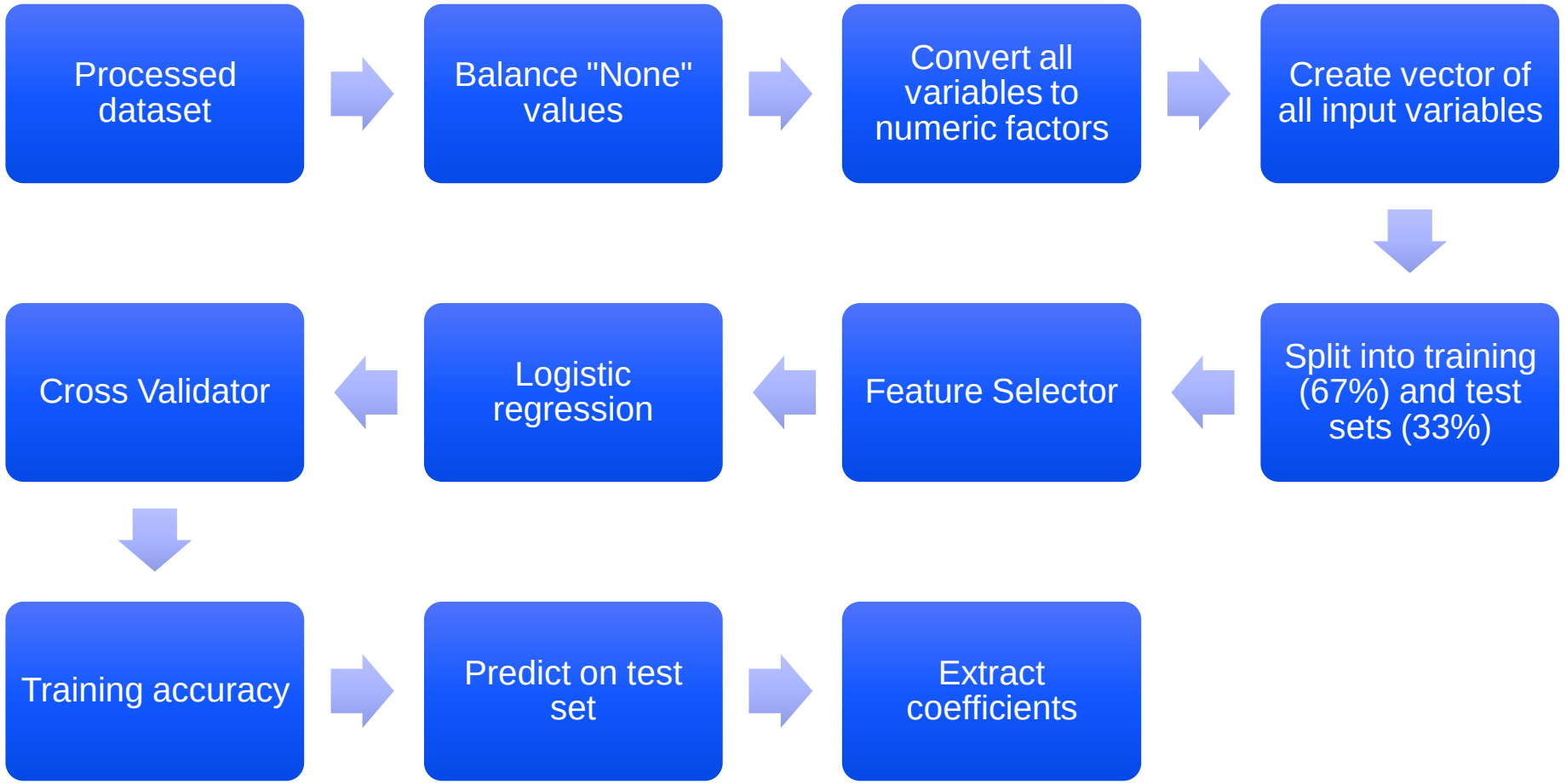
Reduce

{(key, value)} → (key, result)

(35cc4c9488db264c, LTC_Asthma)
(35cc4c9488db264c, LTC_Depression) → (35cc4c9488db264c, LTC_Asthma, LTC_Depression)

3.1 million rows → 55,000 usable records

Predictive modelling



Multivariate logistic regression

Univariate logistic regression:
a set of n observations with p independent variables $\{y_i, x_{ij}\}$ where y is 0 or 1

$$z_i = \sum_{j=1}^{p} x_{ij}\beta_j \qquad \mathrm{P}(y_i = 1) = \frac{e^{z_i}}{1 + e^{z_i}}$$

Multivariate logistic regression:
now y can be any one of m classes $y_i = \{0, 1, 2, 3, \ldots, m-1\}$

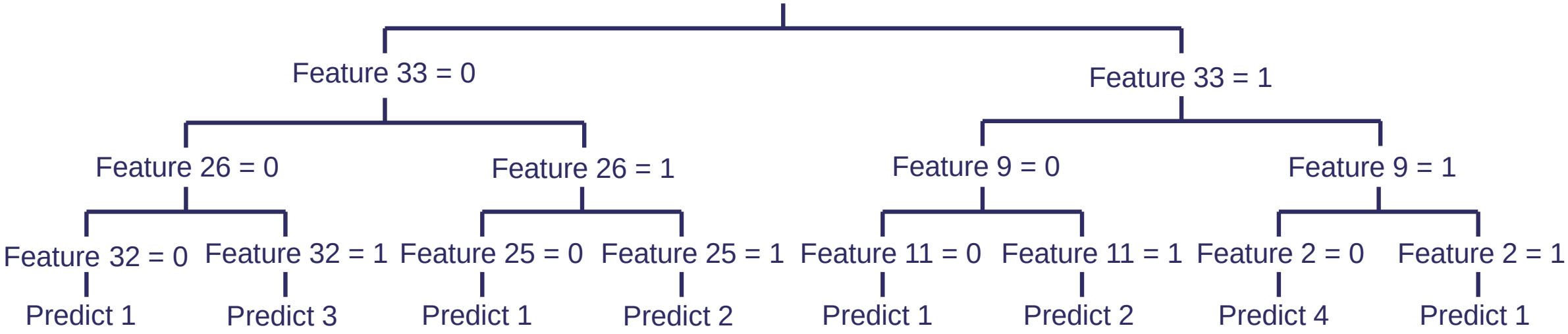
$$z_{ik} = \sum_{j=1}^{p} x_{ij}\beta_{jk} \qquad \mathrm{P}(y_i = k) = \frac{e^{z_{ik}}}{1 + \sum_{k=1}^{m-1} e^{z_{ik}}}$$

$$\mathrm{P}(y_i = 0) = \frac{1}{1 + \sum_{k=1}^{m-1} e^{z_{ik}}}$$

(one of the constants is fixed by the requirement that all probabilities sum to 1)

Decision Trees

Decision tree



This tree has depth 3

Decision Trees

- Decision trees always make a binary decision at each split
- Create a decision tree by splitting the data according to an impurity measure.
  - This describes how similar the two classes are.
  - Want a split which separates two similar classes
  - Common measure is the *Gini impurity: m*easures how often a label would be mis-classified if its label was picked at random from all the others on that branch of the tree.

$$I_G = 1 - \sum_{i-1}^{m} p_i^2$$

- Limit the depth of a tree to avoid over-fitting
  - Can always classify the training data perfectly with a tree, but this model will be useless for anything else!

Random Forest

***Random**:* create extra ‘data sets’ by re-sampling the original data with replacement to create a new data set of the same size. This is called *bagging*, for ‘bootstrap aggregating’, and is intended to reduce over-fitting.

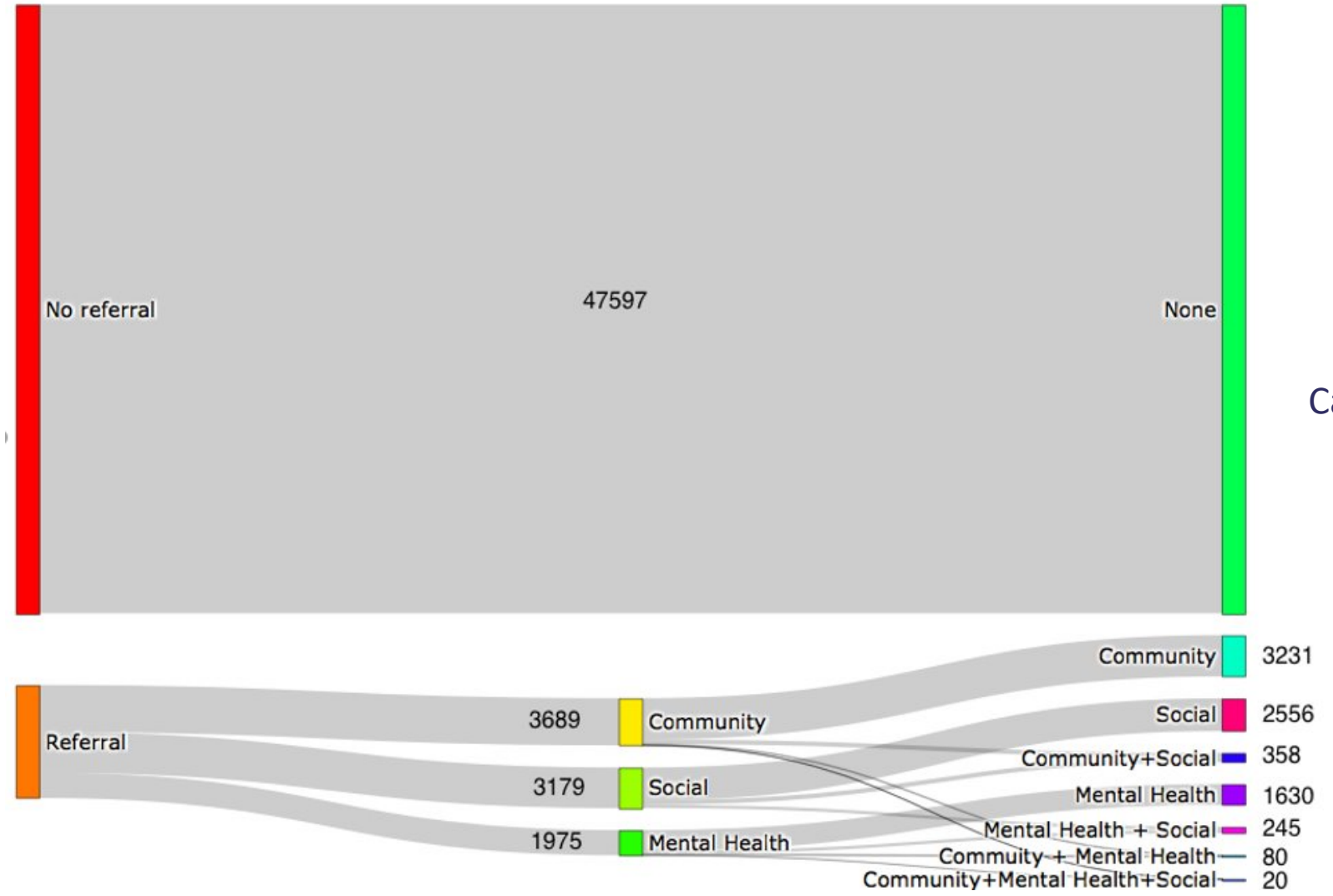
Out-of-bag error: use the data points that were left out of the sample to estimate the model error.

Also uses a random sample of the features at each decision point, not all the features.

***Forest**:* create several different trees and use the most popular class for each data point.

Hyperparameters: number of trees, depth of each tree.

Predictive modelling



Care pathways for patients

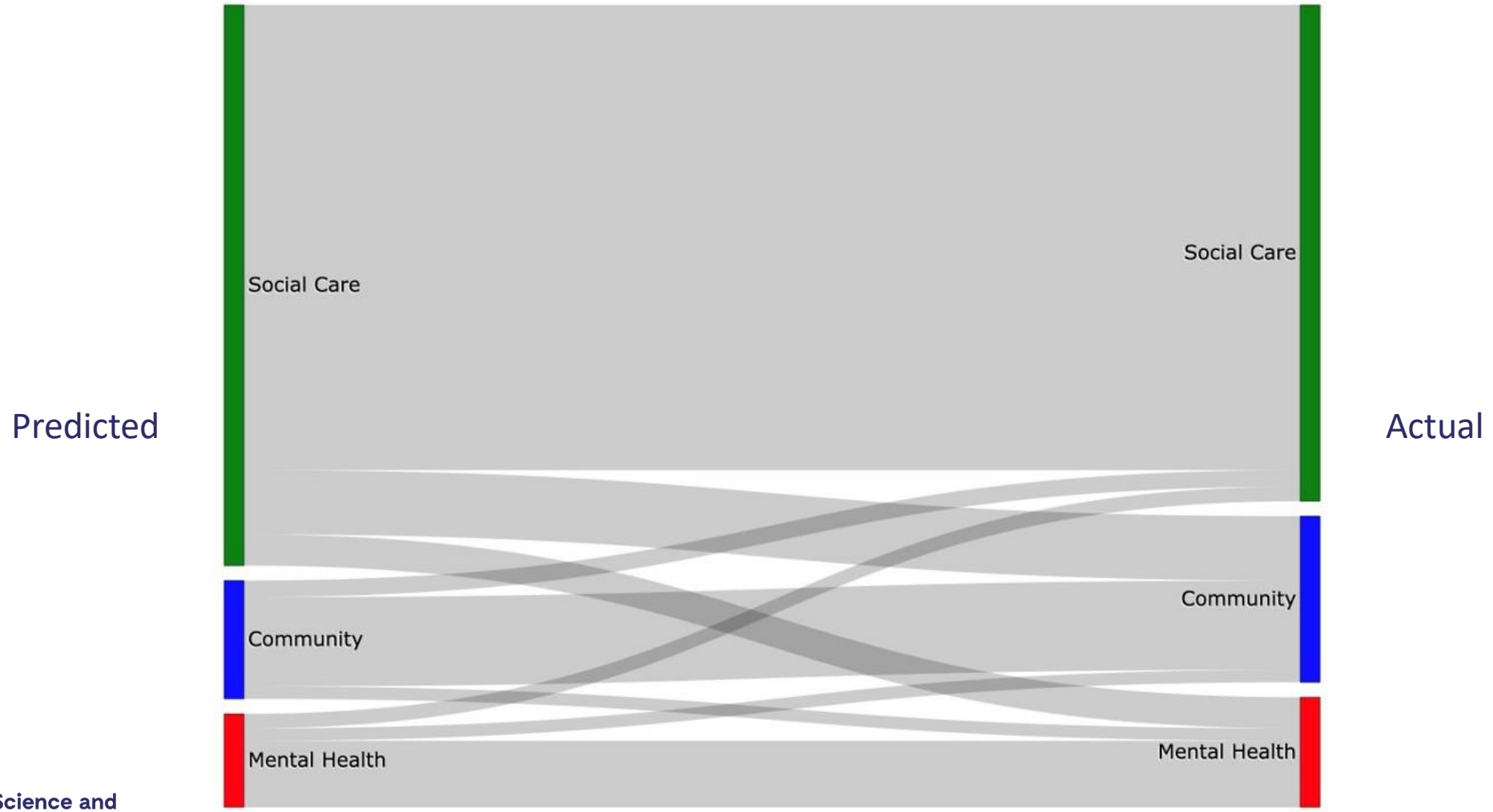
Science and Technology Facilities Council

Hartree Centre

Predictive modelling

	None	Other
None	1236	9
Other	219	34

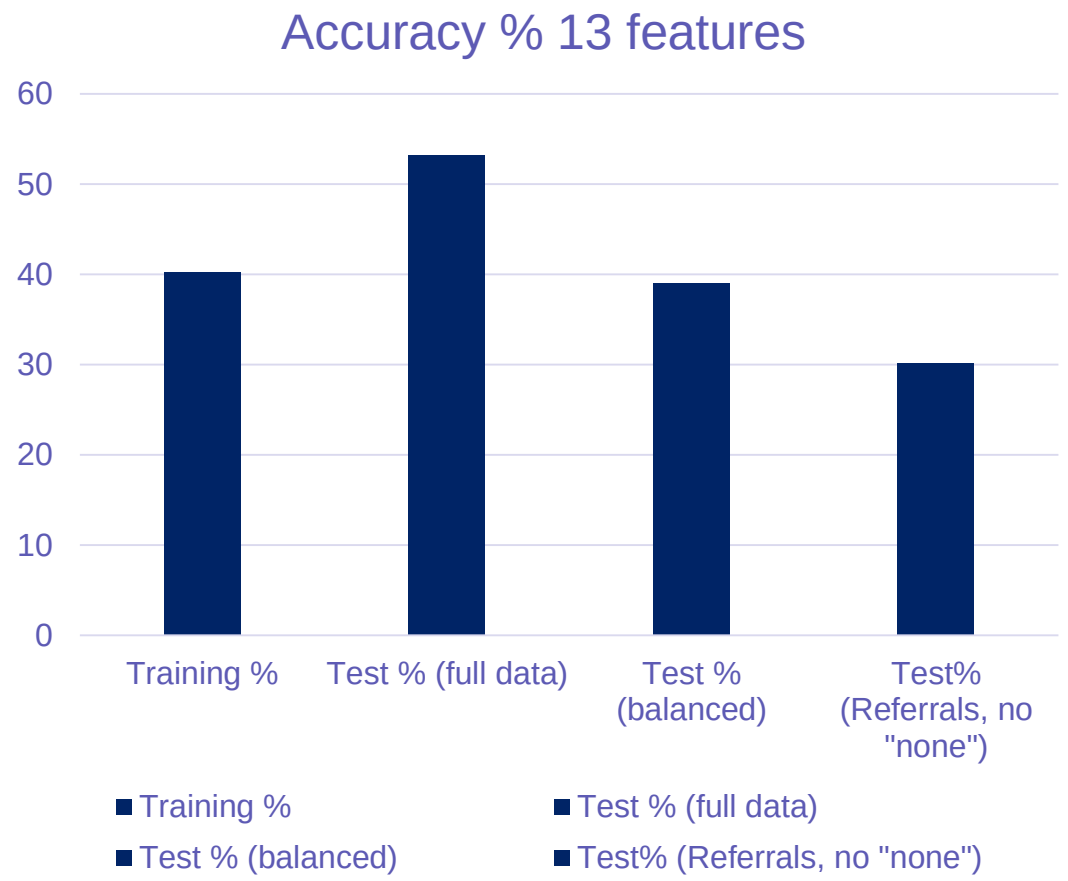
Predictive modelling



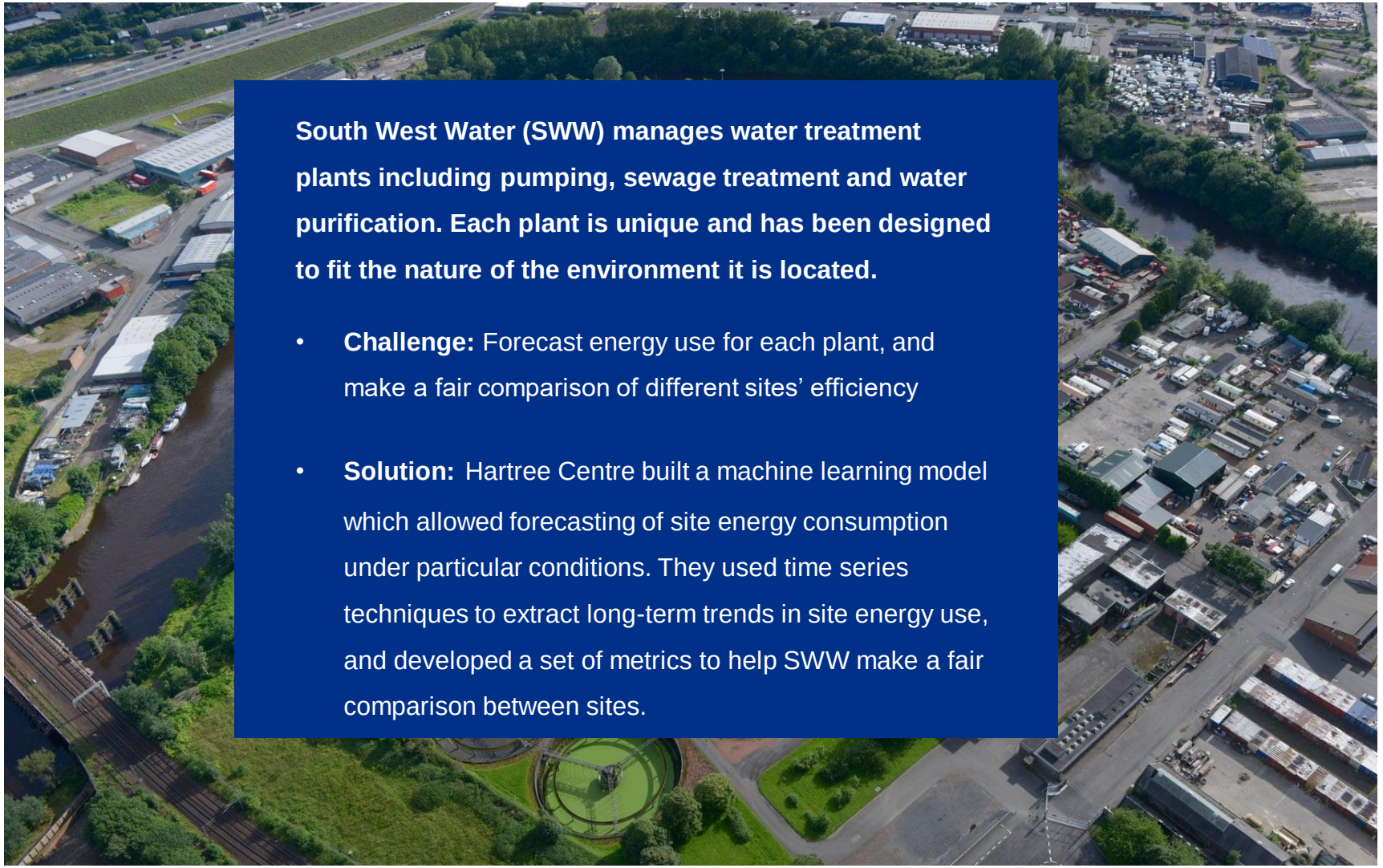
Science and Technology Facilities Council

Hartree Centre

Predictive modelling



Time Series Analysis



South West Water (SWW) manages water treatment plants including pumping, sewage treatment and water purification. Each plant is unique and has been designed to fit the nature of the environment it is located.

- **Challenge:** Forecast energy use for each plant, and make a fair comparison of different sites' efficiency

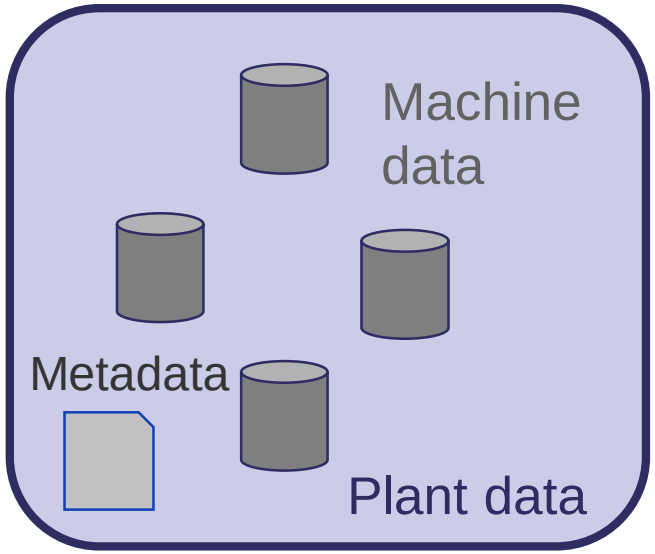
- **Solution:** Hartree Centre built a machine learning model which allowed forecasting of site energy consumption under particular conditions. They used time series techniques to extract long-term trends in site energy use, and developed a set of metrics to help SWW make a fair comparison between sites.

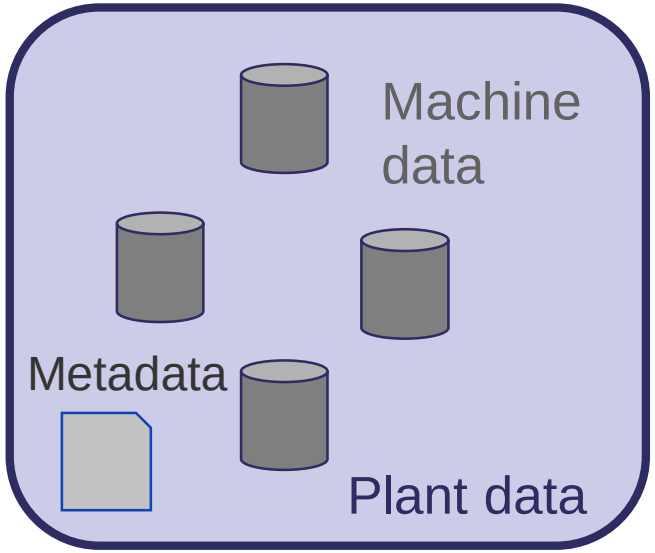
UKRI Science and Technology Facilities Council
Hartree Centre

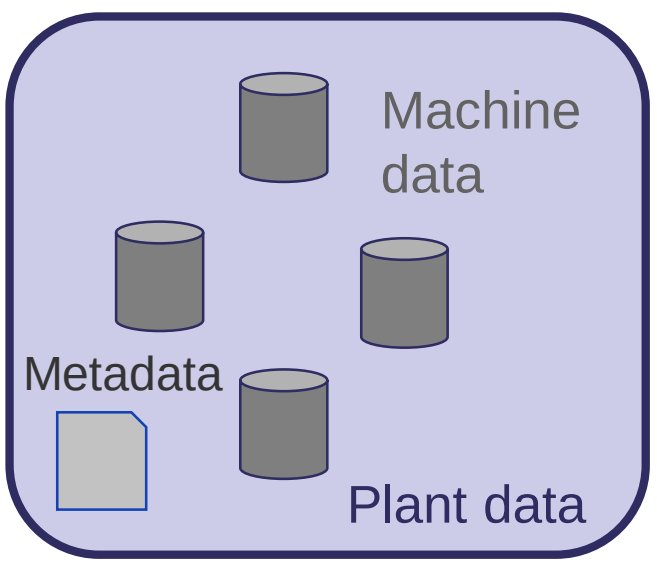
South West Water

Data fusion

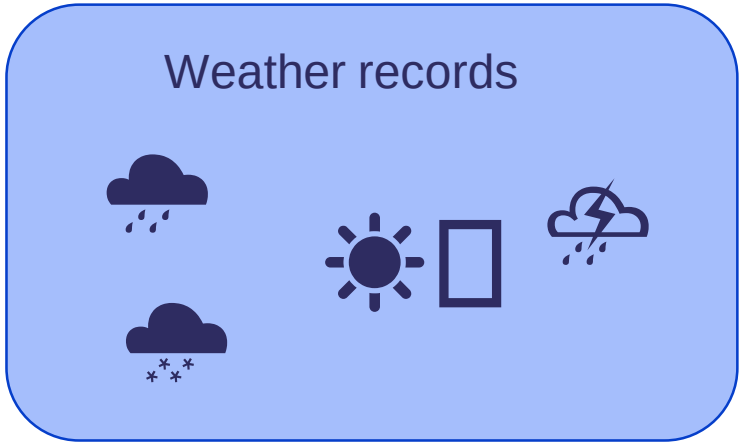
South West Water: predict energy use for waste water treatment plants







Join using date/location



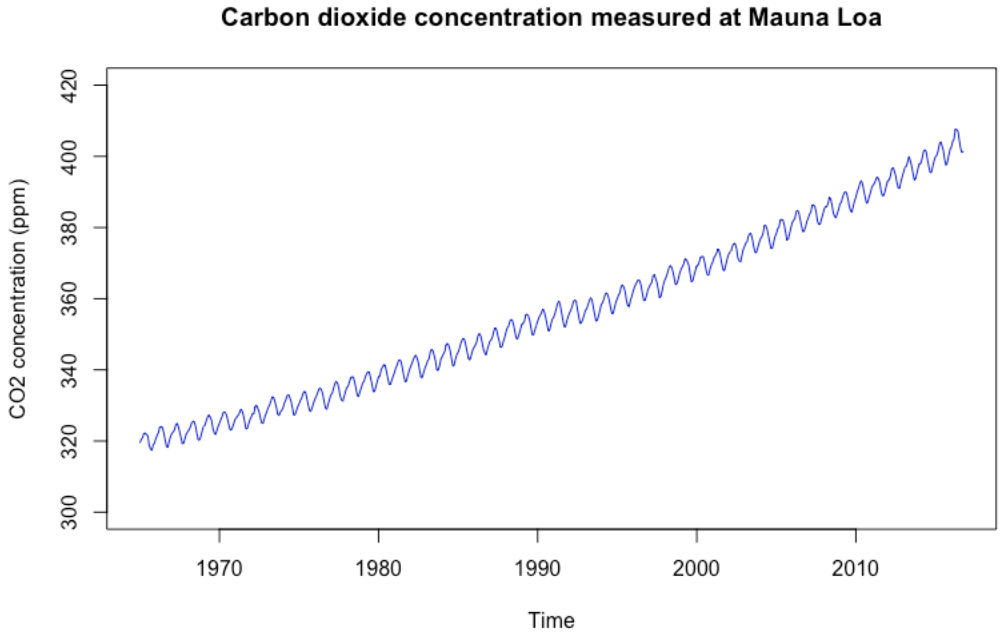
Join using location

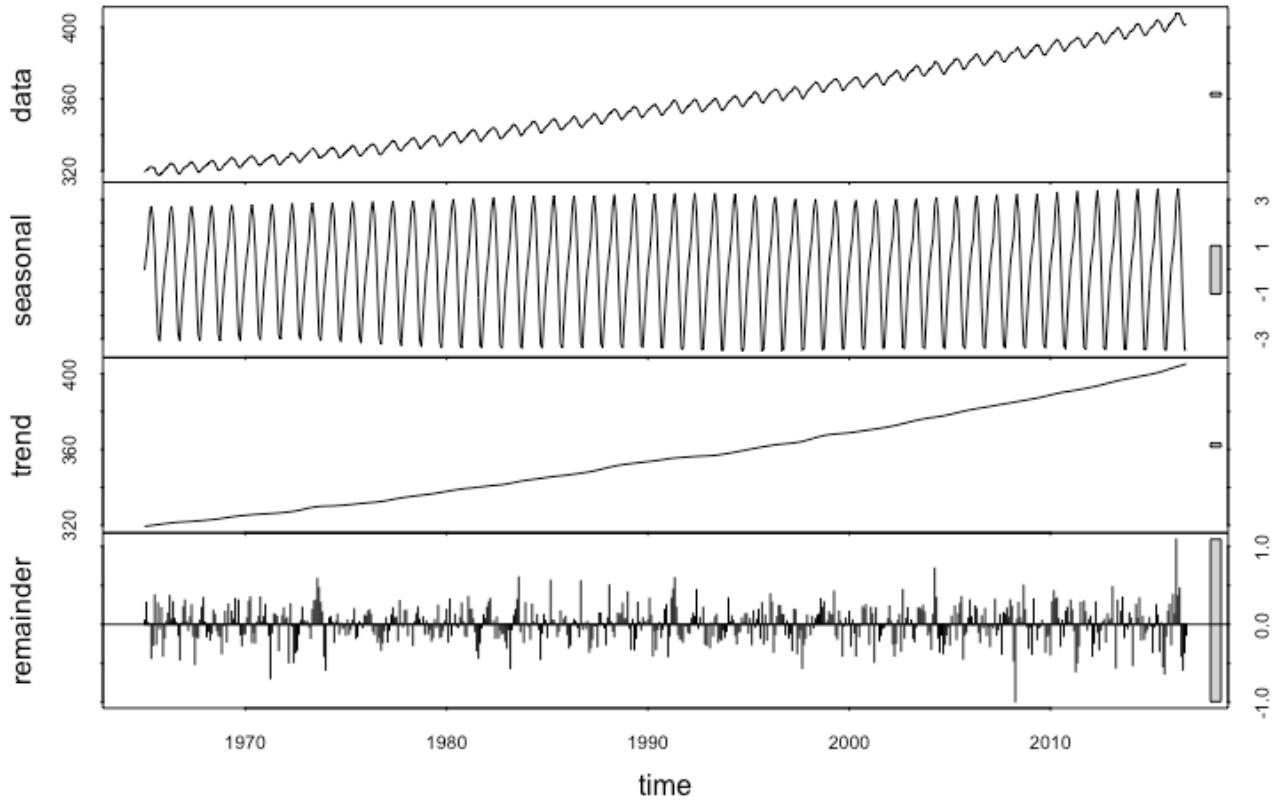
Science and Technology Facilities Council
Hartree Centre

South West Water

Time series analysis

STL Decomposition – extract seasonality and long-term trend.

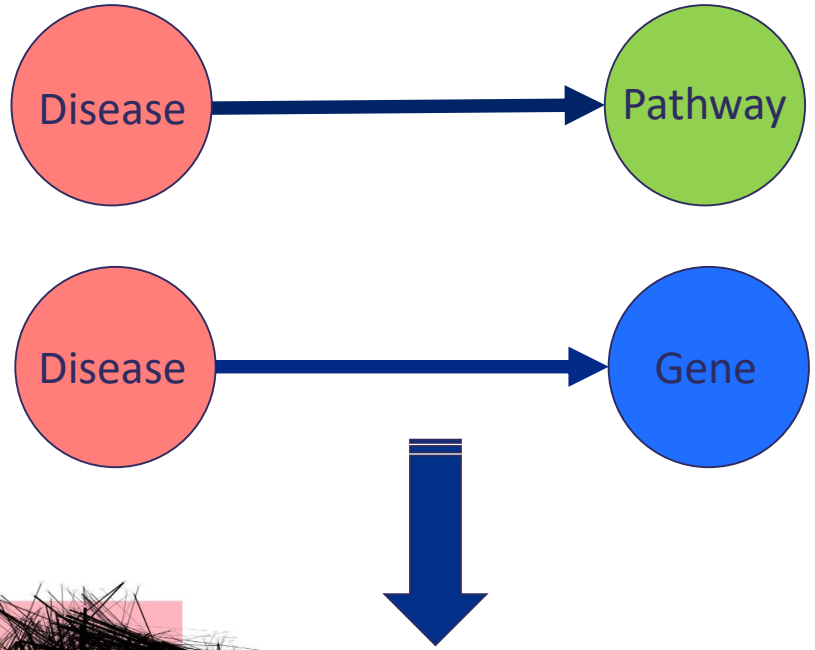




Science and Technology Facilities Council

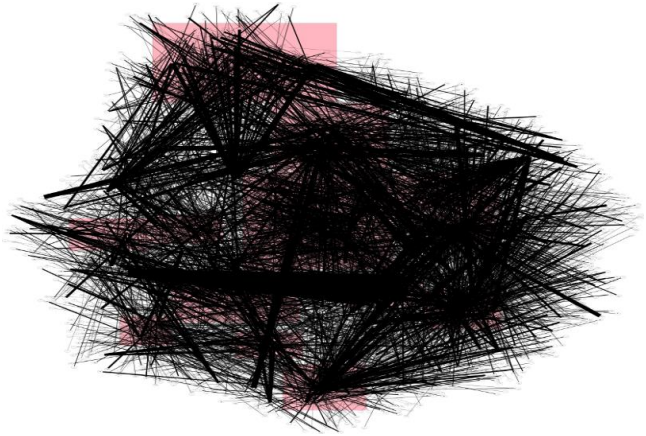
Hartree Centre

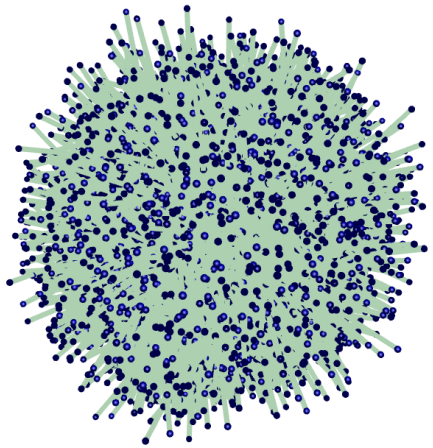
Data visualisation



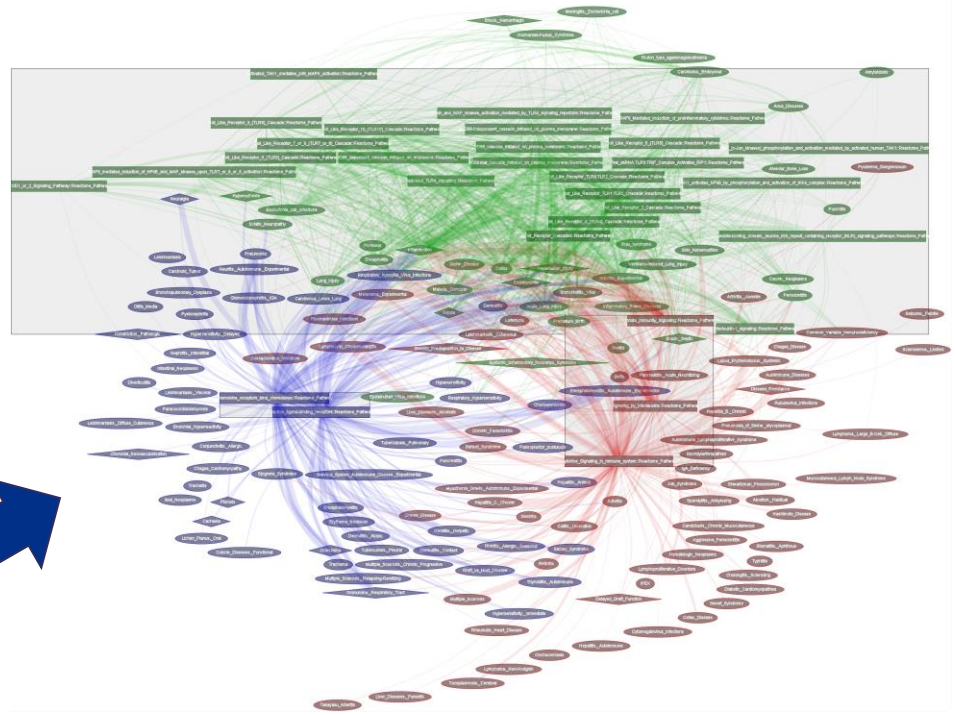
Raw network:
~ 22,000 nodes
1,000,000 edges

Data visualisation





Clusters centred on Crohn’s disease



2 known clusters, 1 new one

Science and Technology Facilities Council

Hartree Centre

Data visualisation

Whole network on our large screen

Science and Technology Facilities Council
Hartree Centre

UKRI Science and Technology Facilities Council

Hartree Centre

Questions?

Science and Technology Facilities Council

Hartree Centre

Thank You

Dr Simon Goodchild
simon.goodchild@stfc.ac.uk
sggoodc@liv.ac.uk